DATA CLEANING FOR AIR FORCE APPLICATIONS

Graduate Research Paper

Giovanna Espegio, Major, USAF

AFIT-ENS-MS-20-J-033

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENS-MS-20-J-033

DATA CLEANING FOR AIR FORCE APPLICATIONS

Graduate Research Paper

Presented to the Faculty

Department of Operational Sciences

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Operations Management

Giovanna Espegio, BS, MS

Major, USAF

June 2020

AFIT-ENS-MS-20-J-033

DATA CLEANING FOR AIR FORCE APPLICATIONS

Giovanna Espegio, BS, MS

Major, USAF

Committee Membership:

Colonel (R) Adam D. Reiman, Ph.D.
Chair

Abstract

There is more data available to the Air Force now than ever before. Leveraging this data deluge for useful applications—making use of this data in meaningful ways—has been an ongoing challenge. There are numerous commercial off the shelf programs which have the power and capability to deal with large, complex data sets; however, use of these commercial products are limited by licensing costs, restrictions on proprietary information including source code, security, and the user's familiarity and proficiency with the software. Even the widely used Microsoft Excel program, while useful, has its limits in terms of data size and speed of processing. Additionally, the Air Force faces the challenges presented by outdated data architectures, restricted visibility and data sharing capabilities, and a limited number of data science experts in the force. The purpose of this research is to answer the question: "How can the Air Force empower Airmen to leverage data as a strategic asset by facilitating the process of data cleaning to make it fit to purpose?"

To my family.

Acknowledgments

I would like to thank Dr. Adam D. Reiman for inspiring the topic of this paper and for his thoughtful and dedicated mentorship throughout this year. I would also like to thank Air Force Operational Energy (SAF/IEN) for the initial vector to get me going in the right direction. Finally, I would like to thank Ms. Pamela Bennett Bardot, Director of the General Ronald R. Fogleman Library at the USAF Expeditionary Operations School, for her excellent research support and encouragement.

Giovanna Espegio

Table of Contents

List of Figures

Page

List of Tables

Page

List of Equations

Page

DATA CLEANING FOR AIR FORCE APPLICATIONS

I. Introduction

General Issue

The Air Force is capturing and recording data at rates and in quantities like never before. As the Air Force seeks to fly, fight, and win in this *Information Age*, leveraging this informational boon in support of rapid data-driven decision-making requires data analysis of astronomically large and complex data sets at a granular level. Fortunately, in the span of a generation or two, the nation has gone from relying on the intellectual power of human computers crunching numbers in the basement of NASA Johnson Space Center to send men to the moon, to having orders of magnitude more computing power available at our fingertips.

The commercial sector has been quick to exploit these capabilities; data analysis tools such as SAS, Tableau, and even Microsoft Excel offer excellent solutions for a large variety of problem sets. However, regardless of the software tools used for data analysis, the overall process of systematically gleaning meaning from data remains the same. The field of data science, which will be defined in greater detail in following chapters, addresses this process.

After data capture, it is necessary to import or structure the data in a way which allows the analyst or researcher to access and manipulate the data effectively. Furthermore, before any bona fide analysis can take place, taking time to deliberately and systematically understand the data, accomplish preprocessing actions, and clean the data is absolutely fundamental to achieving quality analytical results. No matter how

elaborate the planned analysis or how complex the model, the old adage of "garbage in, garbage out" holds true—high quality data trumps model complexity. Dr. Monica Rogati's data science analog to Maslow's hierarchy of needs illustrates this concept well: addressing the higher level needs of belonging and self-actualization (say, machine learning algorithms or AI-powered predictive modeling) is meaningless if one has no food or shelter (solid foundation of data collection, structure, and cleansing) (Rogati, 2017).

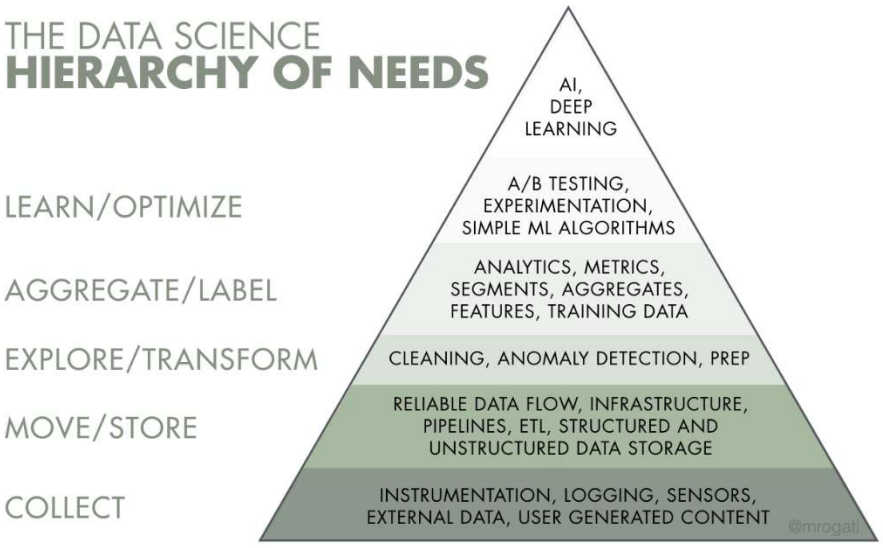


Figure 1 Dr. Monica Rogati's Data Science Hierarchy of Needs (Rogati, 2017)

Many software tools exist to facilitate the process of setting this foundation, but they are proprietary and often only available for experts (or require a certain degree of expertise to use). These are often limited from wider use within the ranks due to a variety of reasons, including product licensing fees, the number of Department of Defense (DOD) licenses available for use, and other limitations associated with the use of proprietary closed-source systems.

Problem Statement

There is a need for even the non-data scientist in the ranks to be able to leverage data at their level. Although MS Excel is the program most widely available across the Air Force, and offers a highly capable, user-friendly interface, often the data cleaning process itself is laborious and unnecessarily time-consuming in Excel. Data science expertise is limited to a select few within the service, necessitating ways to facilitate data cleaning tasks for personnel without a background in analytics.

Research Objectives and Focus

The objective of this research is to discover ways to enhance the ability of Airmen in staff and operational roles to clean data to make it fit to purpose, even for those without data science or analytics expertise. This research focuses on the data cleaning aspect of data science as it is the first critical task that must be addressed before any bona-fide data analysis can take place. By many estimates, data cleansing can take between sixty to eighty percent of the overall time spent in data analysis (Pyle, 1996; Mayo, 2019). Specifically, this research seeks to understand opportunities within the Air Force to leverage the data science capabilities of various common proprietary data analysis tools, such as Microsoft Excel, SAS, Tableau, as well as free open-source software (FOSS) options such as solutions created using the R and Python programming languages.

Research Question

How can the Air Force empower Airmen to leverage data by facilitating the process of data cleaning to make it fit to purpose?

Investigative Questions

1. What is the existing Air Force approach to data science, including data cleaning?
2. What are the gaps or limitations in the current Air Force approach to data science, including data cleaning?
3. What data science software tools are currently available for data cleaning tasks, both commercial and open source?
4. How do the different data cleaning alternatives compare?
5. What should be the future Air Force approach to data science, including the data cleaning process?

Methodology

This research used a mixed methods approach, employing a narrative review of relevant literature, semi-structured interviews with data science professionals, and systematic testing of software alternatives.

Assumptions/Limitations

While many of the observations and conclusions of this research are relevant to the state of data science in the Air Force in general, detailed analysis is limited to the aspects of data preparation and cleaning to define the scope of this research. Furthermore, while the related topics of agile software development, software acquisition, and cybersecurity are relevant to any discussion of data, any mentions of these topics are passing in nature and are not the focus of this research. Lastly, the author

supports ethical handling of data, particularly when it comes to respecting personally identifiable information, although the topic of ethics within data science are beyond the scope of this research.

Implications

The implications of this research are clear: data, particularly big data, presents an opportunity important to the future of the Air Force, and in order to take advantage of this opportunity, the Air Force must address existing gaps in the capabilities, tools, and expertise required to make the most of this opportunity. Cultivating a culture of data literacy and enabling Airmen to access and clean data is foundational to empowering Airmen to make the best use of data for Air Force applications.

II. Literature Review

Chapter Overview

There remains simple experience which, if taken as it comes, is called accident; if sought for, experiment. But this kind of experience is no better than a broom without its band, as the saying is—a mere groping, as of men in the dark, that feel all round them for the chance of finding their way, when they had much better wait for daylight, or light a candle, and then go. But the true method of experience, on the contrary, first lights the candle, and then by means of the candle shows the way; commencing as it does with experience duly ordered and digested, not bungling or erratic, and from it educing axioms, and from established axioms again new experiments.

Sir Francis Bacon, 1620
(Bacon, Hutchins, and Adler, 1952)

At the heart of the scientific method championed by the great philosopher Sir Francis Bacon is the systematic collection, organization, and understanding of observable information about one's environment, or *data*. The reader is invited to think of *data science* as one approach to ignite the candle which lights the way in Sir Francis Bacon's analogy. This chapter will introduce key definitions, describe the steps within the data science process, and delve into the details of data preparation with an emphasis on data cleansing.

Data: Definitions

In order to understand the process, one must first be familiar with the terminology used to describe elements of data science. A *datum* (plural *data*) is a measurable, observable piece of information about the universe and may come in many forms. (Although the grammatical distinction between the plural data and singular datum is well defined, in the common vernacular the term *data* is commonly used as both the plural and

the singular, so for simplicity's sake the terms will be used interchangeably where appropriate.)

In its simplest form, data is categorized as either qualitative or quantitative. Qualitative data includes information about a categorical quality of an observation which can be categorized by labels or by order in a series, while quantitative data includes information about an observed quantity or numerical measurement. The distinction may seem obvious at first inspection; however, there are some nuances which should be noted.

A common acronym used in statistics and mathematics textbooks to characterize data is N.O.I.R., which stands for Nominal, Ordinal, Interval, and Ratio data (Stevens, 1946). The first two, nominal and ordinal, are qualitative, while the last two, interval and ratio, are considered quantitative. Nominal data includes information like names, classes,

Scale	Basic Empirical Operations	Mathematical Group Structure	Permissible Statistics (invariantive)
NOMINAL	Determinaflon of equality	*Permutation group* $x' = f(x)$ $f(x)$ means any one-to-one substitution	Number of cases Mode Contingency correlation
ORDINAL	Determination of greater or less	*Isotonic group* $x' = f(x)$ $f(x)$ means any monotonic increasing function	Median Percentiles
INTERVAL	Determination of equality of intervals or differences	*General linear group* $x' = ax + b$	Mean Standard deviation Rank-order correlation Product-moment correlation
RATIO	Determination of equality of ratios	*Similarity group* $x' = ax$	Coefficient of variation

Figure 2 Types of Data (Stevens, 1946)

or labels. There are no quantities, no specified order, nor measurable distances between these nominal data; for example, aircraft name or ice cream flavor would be examples of nominal data. It is important to note that even data expressed in numbers can be

considered nominal; for example, a soccer player's jersey number is a nominal datum since it only provides identifying information but not information about any quantities relevant to that player (player #3 is not half the player #6 is, and jersey number does not normally give any information about the order players appear in). The next kind of qualitative data is ordinal data, which is information about position within a sequence. For example, the departure order for an aircraft awaiting takeoff, the finishing rank of a marathon runner, or a survey respondent's reported level of education can be considered ordinal data because it provides information about rank or order within a sequence of classes while not providing any distinct information about the distance from other observations or instances within the sequence (Stevens, 1946).

Quantitative data can be separated into two categories: interval or ratio data. Interval data provides information about an observation's position in a sequence or scale as well as the distance between other observations along that same scale. For example, temperature measured in degrees Fahrenheit is an example of interval data: 45°F is five degrees cooler than 50°F. Ratio data, such as weight, length, or altitude above sea level, provide the same level of information (position and distance on a scale) with the added distinction that there is a natural or universal meaning to a value of zero for that data, as contrasted to an arbitrary value of zero depending on how the scale is defined for interval data (Stevens, 1946).

Aside from the obvious need to understand one's data, it is necessary to be clear on the type of data because there are different descriptive statistics which are relevant and meaningful (or potentially irrelevant and meaningless) for each data type. With nominal

data, total counts for each type is the primary descriptive statistic, but beyond that, caution must be used. For example, attempting to calculate a mean for nominal data on observed aircraft types would be nonsensical. However, for ordinal data, not only can frequency counts (and thus the mode) be established for each value, but also the median. However, attempting to find the mean or standard deviation of ordinal data should be used with caution because ordinal data does not provide information about the distance or measurable difference between ranks, and thus standard arithmetic operations should be approached with caution (Stevens, 1946). For example, an aircraft who is listed sixth for takeoff will not necessarily take off in three times the amount of time that it took for the second aircraft awaiting takeoff. Quantitative data, however, can make full use of descriptive statistics, including frequency counts, mean, median, range, variance, standard deviation, and so forth. However, when considering ratios between observed values, it is important to preserve the distinction between interval and ratio data, because depending on the way that interval data is recorded, analyses which depend on ratios of observations may be misleading (Stevens, 1946).

While data can come in all shapes and sizes, there is another distinction to be made when considering implications for future analyses. In recent years, *big data* is a term whose nebulous definition has often allowed it to be bandied about as a superlative by the less informed to add intrigue or emphasis to a project, but in reality, there are a few concrete characteristics which can be used to distinguish *big data* from data at large. Generally, *big data* is "too large and/or complex to be effectively and/or efficiently

handled by traditional data-related theories, technologies, and tools" (Cao, L., 2017:4) due to three main factors: volume, velocity, and variety.

For volume, one must consider if it can be contained in a single database, such as in Microsoft Excel, with thousands or tens of thousands of lines. This is generally considered to be still within the realm of normal data, while *big data* involves "millions to billions of data points" (Hamilton and Kreuzer, 2018:5). A data set of this size could not be contained in a single worksheet in Microsoft Excel, which is limited to just over one million lines of data (1,048,576 rows to be precise), requiring any larger data sets to be broken down into multiple worksheets.

The next defining characteristic of *big data* is *velocity*, or the speed with which data points are created or captured over a finite period of time. The social media platform Twitter provides an excellent example of high velocity in the context of big data: Twitter can receive more than 500 million data points ("tweets") in a single day (Hamilton and Kreuzer 2018:5). An example of high velocity data within the Air Force would be the constant stream of high-definition video from remotely piloted aircraft (RPA) performing surveillance missions.

The final characteristic of big data is *variety*. Variety refers to the extent to which data points initially appear in different formats from various sources. Managing and incorporating data in the form in which it is created is a new challenge in contrast to the "way things were" in the past, where the corporate or governmental entity employing an analyst may have had more control over the collection methods and thus the format of what was collected and how it was stored (Hamilton and Kreuzer, 2018:5).

While these three characteristics—volume, velocity, and variety—are typically used to define big data, there is a fourth alliterative *V* that is worth mentioning: *veracity*. There can be a lot of "noise and irrelevant data" with the volume, velocity, and variety of big data, and this necessitates a process for preparing and cleaning the data to ensure its truthfulness, otherwise known as data *veracity*. Veracity ensures that one captures what is truly important from the data, precludes selection bias, and provides awareness of big data "working hazards" (Hamilton and Kreuzer, 2018:5).

Data Science

Data science is defined as the methodical study of data (Cao, L., 2017:8). From a disciplinary perspective, data science involves a combination of many fields of study, but primarily, it has been commonly characterized as existing at the intersection of statistics, computer science, and domain expertise (O'Neil, 2014:7) with a desired outcome of increased understanding of the system or environment described in the data. This increased understanding, in the form of *data products*, which are simply any deliverables enabled by the data analysis, can also help to inform future decisions (Cao, L., 2017:8). Herein lies the overall goal of data science: the ultimate objective is to provide *insight*.

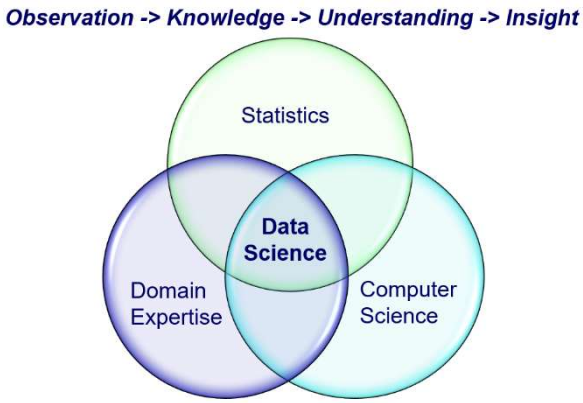


Figure 3 Data Science Venn Diagram. Adapted from Figure 1-1 in (O'Neil, 2014:7)

Data analysis involves processing data using traditional means, such as statistics and logic, in order to achieve useful insight for practical purposes (Cao, L., 2017:4). Data analytics refers to the theories, tools, and methods that lead to actionable insight from data, and includes descriptive, predictive, and prescriptive analytics (Cao, L., 2017:4). As the name implies, descriptive analytics describe or characterize the nature of the data, while predictive analytics make inferences from the data about the future. Prescriptive analytics involves providing support for one decision over another as supported by insight extracted from the data.

While the notion of "data analysis" in relation to statistics has been well established, only relatively recently has the term "data *science*" been used to refer to a distinct discipline. Although the term appeared in the title of the 1996 conference of the International Federation of Classification Societies (IFCS) (Press, 2013), the idea of data science as a distinct discipline truly gained momentum with the publishing of Dr. William S. Cleveland's position paper titled "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics" in the *International Statistical*

Review (O'Neil, 2014; Press, 2013; Cleveland, 2001). This plan called for allocating resources necessary to pivot efforts within the field of statistics to directly support the work of analysts and their ability to understand their data within the context of their field (Cleveland, 2001).

Dr. Cathy O'Neil describes how the term "data scientist" was popularized in 2008, and it was not until 2012 that Wikipedia finally published an entry on "data science" (O'Neil, 2014:8); although the component fields from which data scientists are drawn have been around for a while, the concept of data science as a field unto its own is still relatively new. No doubt this is due to the "data explosion" of the early 2000s and rapid increase in computing power. The implication here is clear: while the field of data science is relatively young, it is rapidly maturing and provides an immense opportunity for innovation.

The Data Life Cycle

It is necessary to consider the entire data analysis life cycle to provide context for understanding the importance of data cleaning within that cycle. At the most fundamental level, all analysis begins with a problem—some phenomenon of interest for which data exists or can be gathered to explore in order to gain understanding. Properly defining this problem and determining whether there is enough information available in the environment, or means to capture the data needed, sets the stage for effective analysis. For data which already exists, seeking information regarding any previous analyses may provide further insight and inform future analysis; this may also involve reviewing the metadata associated with the dataset. Collectively these tasks can be considered the *discovery* phase of the data life cycle (Bonthu and Bindu, 2018).

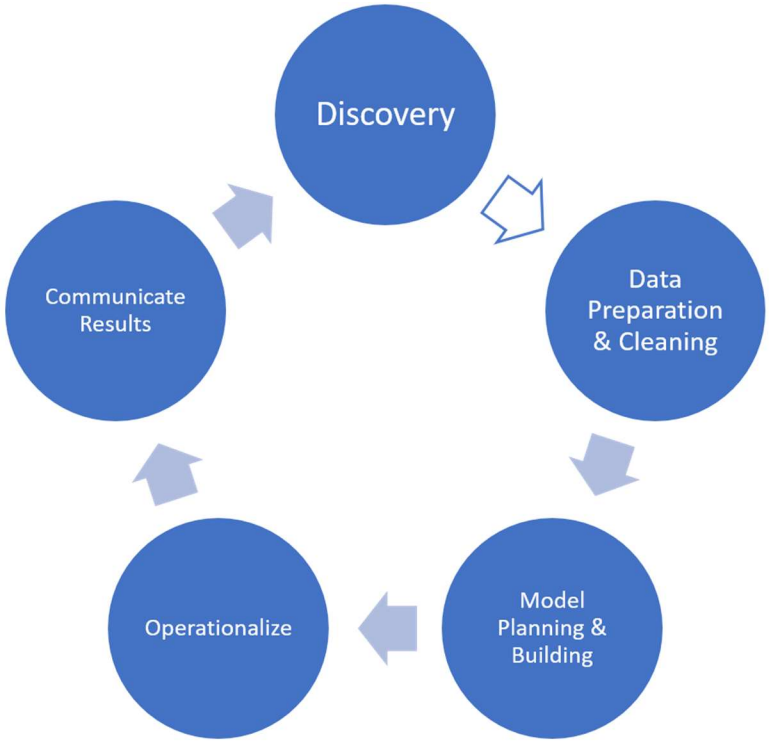


Figure 4 Data Life Cycle. Adapted from (Bonthu and Bindu, 2018)

After discovery, the data must be gathered and prepared. Depending on the nature of the systems or methods used to capture the data, it may arrive in any number of formats or degree of quality. It is during this *preparation* phase that one must extract, load, and transform the data, which is sometimes simply referred to as ELT (Bonthu and Bindu, 2018). Raw data is extracted from the source system or systems, then loaded (imported) into a target database in a useable format. Once this data resides in an electronic database, it must be transformed from raw data into the format, structure, and quality that is required for analysis (Rouse, 2020). It is during this phase that data cleaning takes place.

Once the data is prepared and cleaned, and these actions tracked and recorded in the metadata, analysis tasks such as planning and building models can take place. Once a model is checked using data sets for training and testing, then it is ready to be put into use, or *operationalized* (Bonthu and Bindu, 2018). With sufficient amounts of properly prepared, quality data, this can even include more advanced data-driven applications like machine learning algorithms and neural networks. Finally, the last stage of the data life cycle is to *communicate* the results of the analysis, identify successes or failures, and provide insights which can be used to inform the process of discovery for future analyses.

Initial Data Preparation: Gathering and Understanding the Data

Gather the Data

Before data cleaning can begin, data must be extracted from its source and imported in an appropriate format. Using the Python programming language as an example, this can be accomplished using a variety of opening modes available within the

"os" interface library. These include *w* (write), *b* (read in binary), *x* (create new file), *a* (append at the end of the file), *t* (open in text mode), or + (open for reading or writing) (Al-Jabery, Obafemi-Ajayi, Olbricht, and Wunsch, 2020:240-241). Next, the Pandas library can be used to read a variety of file types (Al-Jabery, Obafemi-Ajayi, Olbricht, and Wunsch, 2020:243), the most common of which is the comma-separated value format, or *csv*, for which the "*read_csv()* function" can be used. By default, data are separated using commas as delimiters, but this can be modified to read data using other delimiters, such as semicolons or spaces. Once the data is read, it is stored as a *data frame* object, allowing column-wise manipulation and plotting (Al-Jabery, Obafemi-Ajayi, Olbricht, and Wunsch, 2020:243). For context, a data frame is a type of n-dimensional data structure (such as an array or matrix) containing rows and columns of data, giving it a superficial similarity at a most basic level to the "spreadsheets" with which most readers will be familiar (McKinney, 2017). By comparison, a *series* is a one-dimensional array of data, as in an array of measured heights for a certain population of interest (McKinney, 2017).

Data can be read and parsed using other functions in Pandas; for example, the *read_table()* function parses data from text files into a data frame, while the *read_excel()* function enables compatibility with Microsoft Excel spreadsheets. This is useful due to the widespread use of Excel spreadsheets, providing the user with efficient options for naming sheets and omitting or skipping specific header or footer lines or rows. If extensive parsing operations are necessary for extracting data from very large data sets, the Python JavaScript Object Notation (JSON) module provides this functionality (Al-

Jabery, Obafemi-Ajayi, Olbricht, and Wunsch, 2020:243-244). A summary of common file extensions used in data science is included in the table below.

Table 1. Common File Extensions ("Python Standard Library," 2020; "IO Tools," undated)

File Extension	Name	Open in Python
.csv	Comma Separated Values	`pd.read_csv()`
.json	JavaScript Object Notation	`pd.read_json()`
.xlsx	Microsoft Excel File	`pd.read_excel()`
.sql	Structured Query Language	`pd.read_sql()`
.pickle	Python Pickle Format	`pd.read_pickle()`
.html	Hypertext Markup Language File	`pd.read_html()`
.zip	ZIP file	`zipfile.open()`
.tar	Tape Archive File	`tarfile.open()`

Understand the Data

After importing the data, it is necessary to structure and understand the data before diving into the details of data cleansing. A common data structure is the data frame, where data is structured as a two-dimensional table, even for data sets with greater than two dimensions (Goel, 2017). Columns should contain attributes, features, or variables with observed values recorded in the cells within the column. Rows should contain instances, observations, samples, or entities which have recorded values or information for some or all of the variables appearing in the columns (Wickham, 2014). Sometimes, but not always, the last column will contain an output, outcome, decision, or prediction, depending on the objective of the analysis or reason for collecting the data. Data can also take other forms, such as pictures, imagery, or even audio for machine learning applications.

Once data is imported and structured in an orderly fashion, it is necessary to characterize and understand the data by considering what information is available, the number of observations or samples, and basic descriptive characteristics. Basic information about the data such as this is considered to be *metadata*, which is simply "data about data" (Kononow, 2018). Using a library as an example, if the books on the shelves represent the data, the library catalog—a collection of index cards listing each book, author, title, date published, Dewey decimal number, and other information about each book—represents the metadata.

Metadata is not limited only to the format, structure, and descriptive statistics of the dataset (technical metadata). It also can provide context for the data, including information about how and by whom it was collected, when it was created, attribute (i.e.column) definitions or other amplifying information (business metadata), and a record of what prior manipulations or transformations have been performed on the data (operational metadata) (Gidley, 2017). This last category, operational metadata, deserves special emphasis because as one goes through the data cleaning workflow, it is very important to not only keep a copy of the original data, but also to keep track of what actions have been taken. Another metaphor is useful here: metadata can be seen as a shipping manifest, detailing what is contained within the 'cargo' (data), and allowing one to keep track of its progress as it proceeds from origin to destination (Gidley, 2017). In many cases, it is possible to include code in one's script to automatically update the metadata file to keep track of modifications.

Other characteristics to consider include the types of data, which are important because data types influence what sort of data manipulations are available and meaningful. For example, quantitative data may appear in data types such as integers, floating-point numbers, or dates; qualitative data may appear as Booleans or text strings. If a variable appears in an incorrect data type, such as a julian date appearing as a text string rather than a number, it may be necessary to convert to the correct data type in order to perform subsequent analysis. For example, in Python the *int()* function can be used to convert a string or floating-point number to an integer, or *str()* can be used to convert an integer or float into a string (Han, undated).

Furthermore, exploring preliminary visualizations such as histograms of each attribute (column values) or scatterplots between variables of interest can provide an initial idea of what patterns, distributions, and possible relationships exist. Taking time to understand one's data at the outset can allow one to identify potential issues up front, like missing values or potentially erroneous data which falls outside a reasonable range for a given variable (Rahm and Do, 2000; Al-Jabery, Obafemi-Ajayi, Olbricht, and Wunsch, 2020). Understanding the data is also important to the data cleaning process as it provides the opportunity to 'audit' the data to identify and define anomalies (Müller and Freytag, 2003). These anomalies may include duplicate data (such as two identical records), conflicting data, missing (null) values, or invalid data (data which appears outside a reasonable or possible range for the variable in question) ("Data Cleaning: In-Depth Guide," 2020). Data auditing will be discussed in more detail in the following section.

Data Cleaning

The purpose of data cleansing is, in its most basic form, to improve the quality of the data by identifying then removing or correcting anomalies in the data. Ideally, this process should be automated to the maximum extent possible rather than accomplished manually. To do this, it is necessary to create a data cleaning 'workflow', or process of operations to perform on the data to cleanse it from anomalies. This involves auditing the data, specifying the data cleaning workflow, executing the workflow, then post-processing as necessary (Müller and Freytag, 2003).

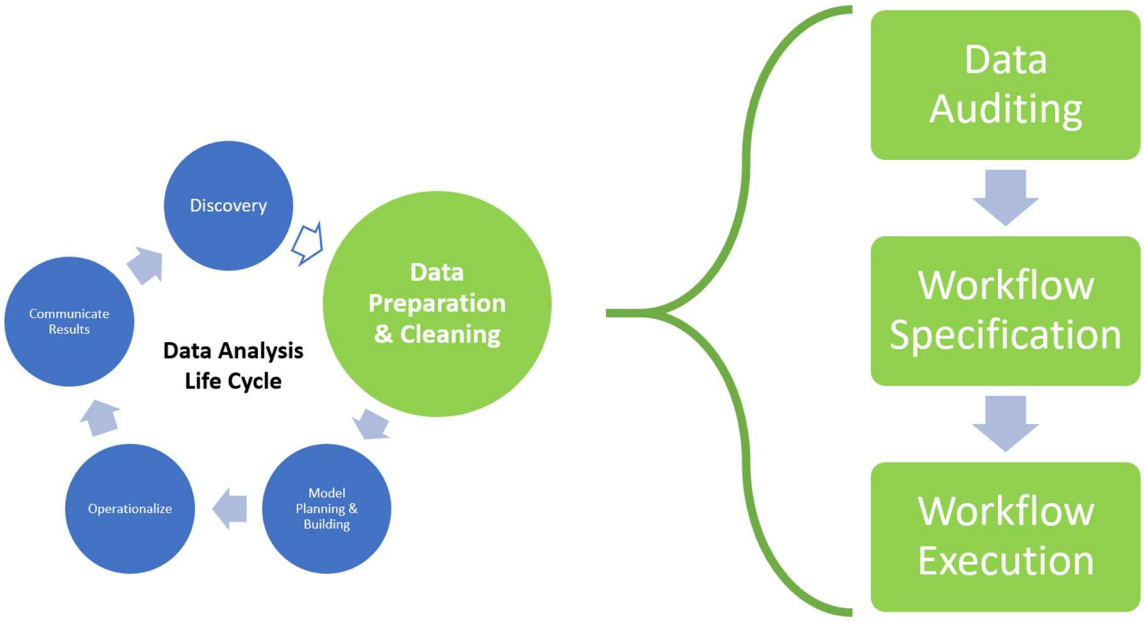


Figure 5 Data Cleaning Overview. Adapted from (Bonthu and Bindu, 2018; Müller and Freytag, 2003)

Data Cleaning Workflow

As introduced in the previous section, *Understand the Data*, auditing seeks to identify anomalies in the data. To identify such samples or variables, it may be necessary to set a threshold or set of constraints for what can be reasonably considered to be invalid

or otherwise unacceptable. A table of common constraints is included below; these can be used to establish data validity.

Table 2. Data Validity Constraints. Adapted from (Müller and Freytag, 2003)

Constraint Type	Description
Mandatory Constraint	Certain columns cannot be empty
Data-Type Constraint	Values in a column must be of a certain data type
Range Constraint	Values must fall between defined minimum and maximum values
Foreign-Key Constraint	A set of values in a column are defined in the column of another table containing unique values
Unique constraints	Field must be unique in a dataset (Example: disallow duplicate SSNs)

A threshold for the maximum number of acceptable missing values in a single sample should also be set. If a sample does not meet the initial quality threshold, it can be removed. Missing value handling will be discussed in further detail in the next section.

It is also useful to identify outliers at this point, although a domain expert should be consulted before discarding outliers as erroneous since they may hold valuable information about exceptional circumstances. This can be done by setting a range constraint outside of which a value can be considered an outlier. Although there are several ways to identify possible outliers, including visually, a common statistical definition that can be used to define an observation x as a potential outlier is included below (McClave, Benson, and Sincich, 2014:91-93).

$$x < Q_L - 1.5(IQR) \ \ or \ \ x > Q_U + 1.5(IQR) \quad (1)$$

Where

$$Q_L = Lower\ quartile$$

$$Q_U = Upper\ quartile$$

$$IQR = Interquartile\ range$$

Once data auditing is complete, the data cleaning workflow is specified and then executed. In this context, the workflow refers to the specific sequence of operations that need to be performed on the data. For example, a decision must be made on how to handle the anomalies identified by the constraints set (Müller and Freytag, 2003). This may include removing all anomalous data or separating certain cases, like outliers, into a separate file for individual consideration. Missing values have several options for handling, which will be discussed in the following section. Regardless of the choice of action, a careful documentation of actions taken during cleansing, preferably recorded in the metadata associated with the data frame, is essential to keep track of what has been done to the data, as well as making one's actions reproducible by future researchers.

Missing Value (Null) Handling

Missing values, also known as null values, exist in almost all real data sets; for example, values may be missing due to sensor failure, a subject's reluctance to answer certain survey questions, or simple data entry errors. Dealing with these missing values appropriately is key to preserving useful observation data without introducing bias or

corrupting the integrity of the analysis. This is especially important in the case of machine learning or artificial intelligence applications where missing values can cause these analysis approaches to fail. It is not always appropriate to simply delete observations which include missing values, only basing one's analysis on complete observations; instead, the way missing values are treated depends on the character of the observed dataset. How missing values are handled informs the shape of the final cleansed dataset; therefore, the assumptions and algorithm selected for imputation, if used, must be appropriate for the nature of the data and the reasons behind the missing values (Al-Jabery, Obafemi-Ajayi, Olbricht, and Wunsch, 2020:8-9).

There are three primary ways to handle missing values: removal algorithms, utilization algorithms, and imputation methods (Al-Jabery, Obafemi-Ajayi, Olbricht, and Wunsch, 2020:9). As its name suggests, a removal algorithm selectively chooses to disregard any observation with missing values. Although simple to implement, there are some risks to this method. Simply removing observations with missing values can degrade the data and waste otherwise valuable data. Furthermore, since the sample of fully observed cases may not be representative of the full dataset, it may also lead to biased results or a loss of precision (Al-Jabery, Obafemi-Ajayi, Olbricht, and Wunsch, 2020:9).

Once these risks are carefully considered, there are several methods to set up removal algorithms. The default in many data processing software packages is a complete-case or list-wise deletion method, which is the easiest to implement since this method simply deletes any observations with missing values. Alternatives include only

removing observations with "many" missing values as determined by having a number of missing values above a certain predetermined threshold. Instead of removing samples or observations, it is also possible to remove features or variables which themselves have many missing values; however, it is important in all cases for the code to keep track of which features or variables were removed in order to consult with domain experts to prevent the elimination of important features (Al-Jabery, Obafemi-Ajayi, Olbricht, and Wunsch, 2020).

The next method, the weighted-complete case method, takes this a step further by reweighting a sample with missing values within a certain subset in order to make the resultant recorded data more representative of the entire dataset. Frequently used in surveys, an example of this method would be reweighting the data from an underrepresented group of respondents to result in a more accurate representation of the entire population. This would be a valid method to ensure that the resulting data is representative of the population of interest (Al-Jabery, Obafemi-Ajayi, Olbricht, and Wunsch, 2020:9).

The next way to handle missing values is with utilization methods. These methods use data from both completely and partially observed cases rather than simply omitting the partially observed cases (Al-Jabery, Obafemi-Ajayi, Olbricht, and Wunsch, 2020:10). This method is better than complete-case methods for correcting for bias when there are missing data. When utilizing this method, especially if imputation will also be used later in the process, it is important to consider the nature of the data which is missing in order to determine if the data is missing *completely at random* or simply

missing *at random*. This may seem like a trivial difference at first glance, but there are important implications depending on which category the missing data fall. First, data which is missing *completely at random* means that missing observations can be considered a random subset of all observations, meaning that the missing values and the set of all observations of that category have similar distributions. However, values which are missing simply *at random* may have systematic differences in the distributions between the missing and the observed values which can be explained by other observed values (Al-Jabery, Obafemi-Ajayi, Olbricht, and Wunsch, 2020).

The last way to handle missing values is through imputation. Imputation simply fills in missing values with plausible data, such as replacing missing values with the mean of a continuous variable, or by replacing with one of the repeated values for a discrete variable either at random or based of the probability distribution for that particular category (Al-Jabery, Obafemi-Ajayi, Olbricht, and Wunsch, 2020:12). There are many other techniques for imputing values, including both single and multiple imputation methods. For a detailed description of existing imputation methods, see (Al-Jabery, Obafemi-Ajayi, Olbricht, and Wunsch, 2020: Ch 2) and (Cokluk and Kayri, 2011).

Data Cleaning Summary

Together, these preparatory tasks are referred to as data cleansing; an overview of these tasks is included in the figure below. Even for small data sets where these tasks can be accomplished by hand, data cleansing is a tedious and time-intensive process. Yet, this is not something that can be rushed or skipped: cleansing and preparing data to make

it fit for purpose is absolutely fundamental to achieving valid, high quality, and meaningful results during one's analysis.

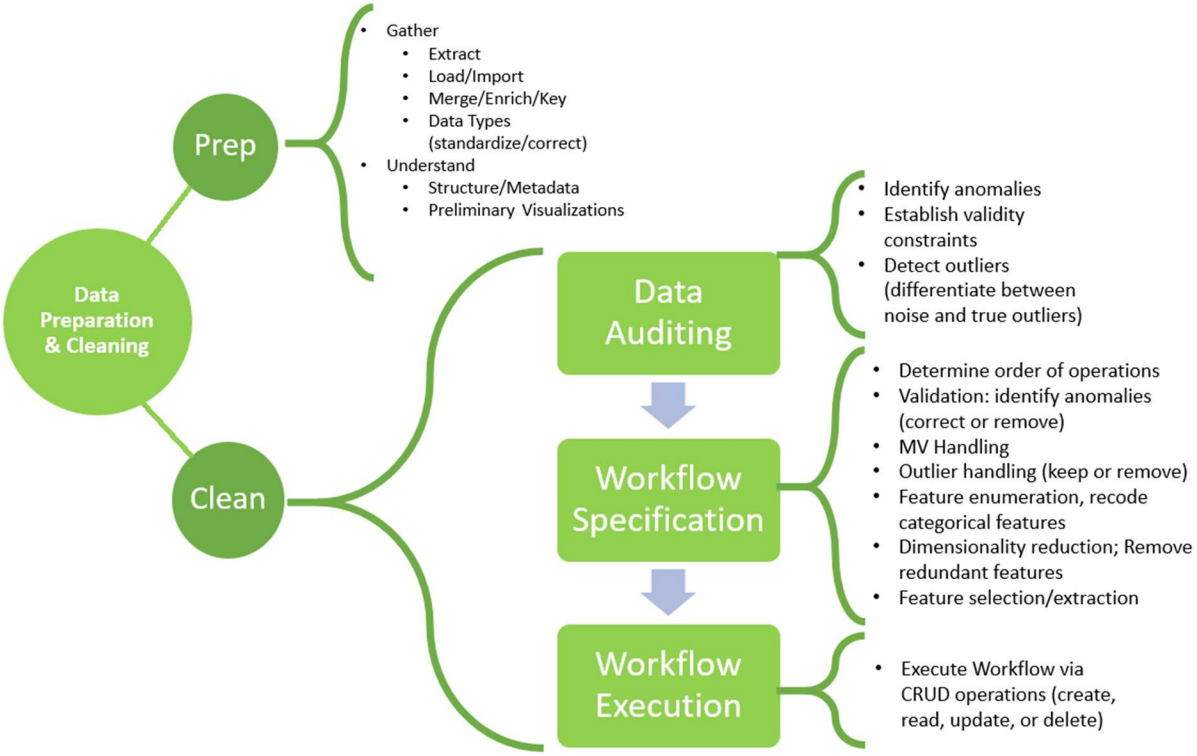


Figure 6 Data Preparation and Cleansing Process Summary

The Air Force Approach

To understand the current state of the Air Force's approach to data science, it is necessary to start at the beginning, as it were, with the overarching strategic understanding of the role which data plays in USAF operations. Being part of the Department of Defense, and thus the federal government, the recent evolution in the ways the Air Force sees (and seeks to exploit) data is part of a larger shift in mindset within the federal government and subordinate agencies to move towards transparency and

enhanced access to federal data collections. As such, the 2020 Federal Data Strategy Action Plan provides an excellent place to begin.

The Federal Data Strategy (FDS) was created in response to goals set in the President's Management Agenda (PMA) released in March 2018. Specifically, the PMA identified a new cross-agency priority (CAP) goal of "leveraging data as a strategic asset" ("Federal Data Strategy," 2020:3) in order to support ongoing efforts to increase government transparency and accountability, enable evidenced-based policy making, and enhance government operational efficacy and modernization. These goals were expressed in one form or another in various executive orders, acts of Congress, and official memoranda over the last decade, notably the May 2013 Executive Order titled "Making Open and Machine Readable the New Default for Government Information" (Executive Order No. 13642, 2013). This Executive Order effectively created the Open Data Policy, the principles of which were more specifically institutionalized in the Office of Management and Budget (OMB) Memorandum M-13-13 as directive guidance for all "executive departments and agencies" (Burwell, VanRoekel, Park, and Mancini, 2013), including, of course, the Department of Defense and thus the U.S. Air Force. Memorandum M-13-13 declared data ("information") to be "a valuable national resource and a strategic asset" (Burwell, VanRoekel, Park, and Mancini, 2013:1), the proper management of which will streamline government operations as well as "help fuel entrepreneurship, innovation, and scientific discovery" through open, accessible data sets made public. As examples, the Memorandum mentioned how the government's decision to make weather information and Global Positioning System (GPS) data available to the

public decades ago propelled great advances in navigational systems, weather predictions and reporting, and even new farming technologies (Burwell, VanRoekel, Park, and Mancini, 2013:1). The launch of *Data.gov* in 2009 was an incipient step in the trend toward open, accessible, machine-readable government data made public. Although it was launched with only 47 data sets initially (Kim, 2019), this database has grown to over 211,000 available data sets ("Data Catalog," 2020) available mostly in the non-proprietary CSV and JSON formats.

In January 2019, Congress signed into law the "*Foundations for Evidence-Based Policymaking Act of 2018*" which contained the OPEN (Open, Public, Electronic, and Necessary) Government Data Act (United States Congress, 2019). This legislation made *Data.gov* a statutory requirement rather than a matter of policy (Kim, 2019), as well as mandating the development of "an online repository of tools, best practices, and schema standards to facilitate the adoption of open data practices across the Federal Government" (United States Congress, 2019). This set into motion a cascade of initiatives by federal agencies to modernize data operations and capitalize on the opportunities. The 2020 Federal Data Strategy formalized this legislation by articulating cross-agency policy (CAP) goals and a series of enumerated actions required to achieve those goals.

The goals and actions consolidated within the Federal Data Strategy provides additional direction to data-related strategy in the *USAF Strategic Master Plan*. Published in 2015, the Master Plan described the Air Force's strategic approach for the next twenty years (Department, 2015). Although it is clear that the five years which have

elapsed since its publishing have only increased the prominence of the role which data occupies, there are some patterns which have persisted through the years, namely data science (or references to data analytic-type activities) has fallen primarily under the ISR, information operations, and information management umbrellas. However, it is clear that this way of thinking—relegating data science capabilities, including data cleansing, to the realm of intelligence—is quickly becoming obsolete for obvious reasons. Fortunately, Air Force and Department of Defense leadership recognized this as evidenced by recent initiatives, which will be discussed later in this paper.

The Air Force Strategic Master Plan provides the motivation behind why critical data skills and automated processing tasks are important, and provides context for how and where data science can play a role; however, to see if there was any set cohesive process for the AF's approach to data science and specifically data cleaning, it was necessary to dive further into AF publications, starting with the Air Force Doctrine Document (AFDD) 1-2: *Air Force Supplement to the DoD Dictionary of Military and Associated Terms* (2012). Searching for every mention of the term *data* or *information* in a data context led to several other relevant doctrine documents which would provide additional clarity on the historical role of data and information in AF doctrine.

AFDD 3-13 *Information Operations* (2011) was mentioned in association with data several times, as well as AFPD 10-7 *Information Operations* (2014), and AFDD 33-3 *Information Management* (2006). These doctrine and policy documents are primarily focused on intelligence operations, such as intelligence preparation of the battlespace, without getting into the details of how data analysis specifically should take place.

However, these documents proved useful to provide leads to further references relevant to the Air Force's or DOD's approach to data science.

AFPD 33-3 *Information Management* (2006) provides a definition of information in the context of Air Force doctrine. Here, information has two meanings: (1) unprocessed data, or (2) meaning that a "person assigns to that data" (AFPD 33-3, 2006:2). The policy document describes the view that information is an "asset to be managed", involving "multiple disciplines traditionally associated with [information management], data management, records management, multimedia management, documents management, workflow management, and publications/forms management" (AFPD 33-3, 2006:7).

Perhaps one of the most useful documents on the topic is a whitepaper published by the Deputy Chief of Staff for ISR titled "Data Science and USAF ISR Enterprise" (Otto, Kimminau, and Ray, 2016). This document describes in depth how the Intelligence, Surveillance, and Reconnaissance (ISR) community understands the importance of leveraging data, particularly big data, and data science in support of our twenty-year strategy as defined in the Air Force Strategic Master Plan. This document also briefly describes the data science process, including depicting where in the process data cleaning must take place.

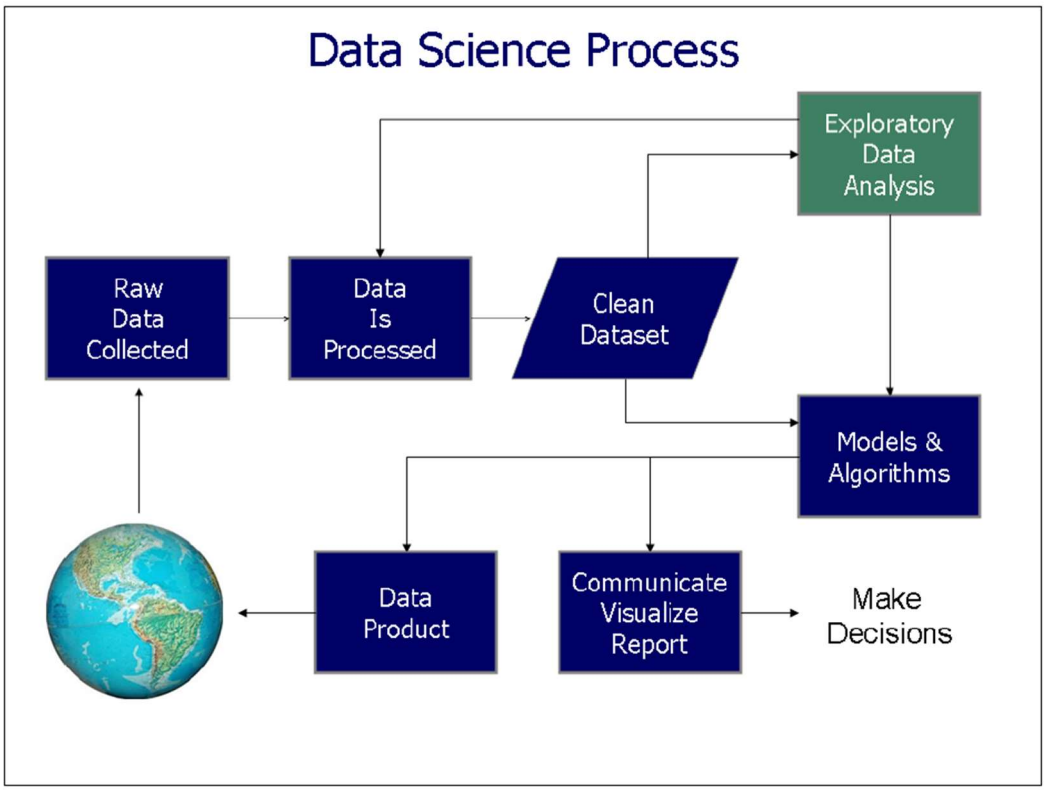


Figure 7 Data Science Process (Whitepaper 2016, p.7)

What the researcher found is that while there is not a cohesive single framework for data processing specifically, there has been a lot of recent movement to support strategic efforts to address the data science gap. At the confluence of these efforts is the Air Force Chief Data Office (CDO), designated in late 2017, which launched an initiative to provide an Air Force "data platform" called VAULT, which stands for Visible, Accessible, Understandable, Linked, and Trustworthy (Vidrine, 2019). VAULT hosts a number of data science tools, including several well-suited to data cleansing. There are also open source tools available by default, including an online Python interpreter and other applications including Apache Zeppelin, Kylo, and Hue. Licensed software, such as Axon Data Catalog, Databricks, Trifacta, and Tableau, are available by request.

VAULT is still quite new, although its capabilities have been demonstrated via a series of initial use cases which met with some significant successes (N. Shields, personal communication, May 14, 2020). The implications will be discussed and analyzed in the following chapter.

Summary

The review of the existing literature provided a theoretical framework for understanding data, data science, and the steps of data preparation and cleaning. It also provided a summary of the federal, Department of Defense, and Air Force guidance and publications relevant to the Air Force's strategic approach to data. The next chapter will address the methodology used to address the research question.

III. Methodology

Chapter Overview

This research employed a mixed methods approach, including a narrative review (Paré, 2017) of existing literature within the field of data science, semi-structured interviews with data science professionals, and systematic testing of software alternatives. This methodology also included a review of program documentation for data analysis software tools, as well as governmental publications and press releases within the Air Force and Department of Defense in order to address the investigative questions and ultimately the research question. This chapter will describe the researcher's approach to uncovering the answer to each investigative question.

Narrative Literature Review

In order to provide context for understanding the Air Force's approach to data cleansing, it was first necessary to explore the evolution of the field of data science itself. To this end, the researcher conducted an initial thematic search for resources on data science and data analytics, focusing on recent sources (within the last five years) either in peer-reviewed journals or textbooks. An initial review of these sources provided references to earlier seminal works, such as (Pyle, 1999) and (Cleveland, 2001) which provided an even more fundamental look at the evolution of data science as a field.

Certain resources, such as (Al Jabery, 2020) were written for biomedical research applications, but nonetheless provided a clear description of the steps of data cleaning and approaches using both open-source and commercially licensed data analytics software. This provided a valuable basis for comparison with later forays into

Department of Defense and Air Force publications which covered topics relevant to data analysis and information management.

After conducting an initial review of the literature and tracking down commonly cited sources, it was necessary to move on to the next phase, which was addressing the investigative questions concerning the Air Force's approach to data science and existing gaps in the process. Understanding that data science within the Air Force is part of larger federal data ecosystem, the researcher began with the Federal Data Strategy, then searched for relevant Department of Defense Directives or Instructions (DODD, DODI). To understand the Air Force's approach specifically, the researcher conducted an extensive search of Air Force Doctrine Documents (AFDD) and Policy Documents (AFPD) to search for references to 'data' or 'information' while noting any relevant cross references in other Air Force publications. The researcher also conducted a keyword search of Air Force E-Publishing for any published instructions (AFIs) or handbooks (AFHs).

The researcher also conducted a search of the Defense Technical Information Center (DTIC) to find any relevant papers or publications. One such publication, a whitepaper published in 2016 by the office of the Deputy Chief of Staff of Intelligence, Surveillance, and Reconnaissance titled "Data Science and the USAF ISR Enterprise" proved to be particularly useful, not only for the information presented within, but also for the wealth of cross references which offered further insight into the topic of data science in the Air Force (Otto, Kimminau, and Ray, 2016). Additionally, the researcher conducted a search of Air Force press releases containing references to data, analytics, or

data science, leading to a wealth of information on the Air Force Chief Data Office and the recent launch of the VAULT Data Platform. Together with slide shows and handouts published by the Chief Data Office on their SharePoint site, this provided a good understanding of the Air Force's recent efforts and initiatives.

The researcher also conducted a semi-structured interview with a data scientist currently employed in the Chief Data Office as well as the Deputy Director of Strategy. This provided an opportunity to clarify details regarding access to and the structure of the VAULT Data Platform, the target audience for the platform, and data cleaning tools available. It also provided an idea of what has been accomplished in prior use cases, as well as potential additional capabilities and tools which may be made available within the data platform in the future.

Lastly, a systematic test of two software alternatives was conducted in order to offer insight into the limitations and opportunities of available data science tools. It was also necessary to review program documentation for commonly used data science platforms for comparison purposes. The tools considered are described in detail in the analysis section.

Summary

This research employed mixed methods, including a narrative review of the literature and a semi-structured interview, to address the investigative questions and ultimately the research question. In the next chapter, the paper will analyze the results of this review and answer the investigative questions.

IV. Analysis and Results

Chapter Overview

This chapter presents the research results by addressing each of the investigative questions directly. The implications of these results will be discussed in the following chapter, which will also make recommendations to achieve the objective of empowering Airmen to leverage data as implied by the research question.

Investigative Question 1

What is the existing Air Force approach to data science, including data cleaning?

A major paradigm shift is in progress with the creation of the Air Force Chief Data Office and the 2019 launch of the VAULT Air Force Data Platform, and thus the Air Force is leading the way in terms of data capabilities in the DOD. In fact, the Department of the Air Force has a service-level implementation plan developed and ready to release once the Department of Defense publishes the *DOD Data Strategy*, which is in the final stages of coordination and is expected to be published shortly (R. Synakowski, personal communication, May 18, 2020). Prior to this, the Air Force's approach to data and information, as indicated by references in Air Force Publications to 'data' and 'information' in an analytic or scientific context, was heavily focused in the Intelligence, Surveillance, and Reconnaissance (ISR) community, although it is clear that all major commands and communities within the Air Force stand to benefit from a cohesive approach to leveraging data as a strategic asset. The establishment of the Chief Data Office and its recent initiatives is evidence of a much-needed revolution in the way the Air Force views and approaches data.

The prevailing approach to data cleansing has primarily been up to data owners and organization analysts, if assigned, to determine the right approach for a given data cleansing task. Due to the nature of the data cleansing process, there is not so much of a need for an Air-Force-specific approach as there is for a set of guidelines to ensure that data cleansing for Air Force applications is accomplished in a manner consistent with industry best practices (N. Shields, personal communication, May 14, 2020) in order to preserve the validity and nuances of each dataset.

Additionally, with respect to the tools used for data cleaning, and the accessibility of both these tools and data upon which to use them, there is a great opportunity for improvement of the current prevailing process. In an interview with Valerie Insinna of Defense News in late 2018, Lieutenant General VeraLinn Jamieson, Deputy Chief of Staff for ISR and Cyber Effects Operations, discussed how Airmen are limited by the tools made available to them to handle data, resulting in a very manpower-intensive ISR enterprise. She described how Airmen are limited to handling and sharing data regarding sensing capability outcomes "via PowerPoint…constructed using Excel spreadsheets to look at the data, identify what is the data, and try to then manually layer the data in this construct" (Insinna, 2018). Taking a manual approach to handling large amounts of data is time consuming and takes Airmen away from higher-level tasks requiring skilled human interpretation (Department, 2015:47). This is especially frustrating given the existence and proliferation of highly capable data science tools, both commercially licensed and open source, to facilitate data wrangling tasks.

These problems are hardly isolated to the intelligence community. Across the Air Force, Airmen in staff and operational roles by default have only had access to basic Microsoft Office tools, such as Excel, on the Air Force network, along with whatever proprietary database software interface happens to be authorized and in use for their particular community or operation (such as GTIMS and GDSS II for managing mobility flight crews and missions, for example). Even these proprietary software tools have their limitations as part of the legacy "waterfall" software acquisition process (Eddins, 2018). In order to manipulate and visualize the data, but limited by the tools to which they have access, Airmen have found creative albeit manpower-intensive solutions.

An example of this is a sight familiar to many in the flying community, the author included, is the scheduling puck board. In October 2016, members of the Defense Innovation Board (DIB) and Defense Innovation Unit Experimental (DIUx) visiting the Combined Air Operations Center (CAOC) noticed a group of Airmen huddled around such a board, shuffling pucks and labeling missions on a magnetic whiteboard while

Figure 8 Refueling Operations Whiteboard. US Air Force Photo (Eddins, 2018)

entering mission data on an Excel spreadsheet to organize the day's missions. The author can personally attest to the fact that this decidedly old-fashioned approach is also in fact quite common in mobility and training squadrons throughout the Air Force. In this particular example observed by the DIB and DIUx within the CAOC, this work-around was necessary because the tools that the Airmen had to plan their missions were practically "useless, dating back to Desert Storm" (Eddins, 2018) while the replacement software acquisition process had been "hopelessly bogged down…for years" (Eddins, 2018).

Noticing an opportunity, the managing director of DIUx made a few calls, leading to a team of Air Force coders and industry software developers to eventually develop a tool called JIGSAW, turning an "eight-hour task for six people…into one that a single person could accomplish in three hours" (Eddins, 2018). This success eventually led to the creation of the Kessel Run agile-software development program which teams up Airmen coders with software experts from industry to create rapid solutions for the Air Force (Eddins, 2018). Although the topic of agile software development as a whole is beyond the scope of this paper, it is relevant to the discussion of the current Air Force limitations in software tools available to accomplish data-related tasks in the sense that it illustrates the need to supply Airmen with appropriate tools for the task or risk expending unnecessary effort and manpower to accomplish the same work manually and much less efficiently.

Even with the right tools, Michael Conlin, DoD Chief Data Officer, acknowledges that "we spend 80 percent of our resources cleaning and standardizing the data to get it

ready for analysis" (Mayo, 2019). Thus, a logical starting point to address the issue of enabling Airmen to effectively leverage data as a strategic asset is first addressing the tools and processes used in the fundamental first steps of data access, standardization, preparation, and cleaning.

Investigative Question 2

What are the gaps or limitations in the current Air Force approach to data science, including data cleaning?

There are three primary gaps in the current Air Force approach to data: lack of data visibility and access, limited analytics capability, and challenges arising from data proliferation in stove-piped systems (Sirota, 2019; Vidrine, 2019). The most salient challenges present with respect to data cleaning in particular are a limited number of data science experts within the Air Force and barriers to access for both the data itself as well as the appropriate tools to prepare and clean the data. A 2016 whitepaper published by the office of the Deputy Chief of Staff of ISR addressed the issue of limited data science expertise within an ISR context, although many of the observations are relevant to the Air Force in general. First, the whitepaper is careful to make a distinction between true data science, involving a high degree of computer science and mathematical proficiency, and the work of data analysts, which the paper characterizes as focusing on determining the usefulness of data to customers, identifying where gaps exists, and using data to make projections about the future (Otto, Kimminau, and Ray, 2016:4-5). It goes on to point out the existing gap in expertise: "data science as a distinctive Air Force career field does not exist" (Otto, Kimminau, and Ray, 2016:5). The closest relevant career fields in the Air

Force include Operations Research Analysts (Air Force Specialty Code (AFSC) 61A and the civilian 1515 series), and those in the Intelligence career field (14N and 1NX AFSCs). However qualified these may be in the area of analytics, the fact remains that a gap exists between these capabilities and those required for data science tasks, particularly in the realm of big data and data cleaning and preparation. Fortunately, there has been an increase in interest by the U.S. government in recent years to hire data scientists as "an acknowledgment of the limitation of assigning non-technical personnel to perform data science work" (Otto, Kimminau, and Ray, 2016:6).

The next challenge, barriers to data visibility and accessibility, is a legacy of the ways and systems in which the Air Force collects and stores raw data. Historically, the military and defense industry have used proprietary stove-piped data storage systems that were only designed to "perform specific functions or support single mission areas" (Emerson, 2019), lacking connections to any larger, enterprise-level data architecture. The Air Force Chief Data Office refers to these as "data jails" ("Public Release Announcement," 2019). Dr. Jon Kimminau, U.S. Air Force Analysis Mission Technical Advisor, describes this issue as an inability to understand the data already in possession since it is captured from "varying sensors, compiled in separate databases, and not accessible…by any single application" (Kimminau, 2015). Even as the amount of data collected by and made available to the Air Force increases, legacy barriers to data visibility limit the ability of Airmen at all levels to effectively leverage the existing data. As discussed in the previous section, Airmen are also limited by the tools with which they have to clean, prepare, analyze, and otherwise use the data that they have.

Fortunately, the recent establishment of the Air Force VAULT Data Platform seeks to address these issues head-on, as will be discussed later in this chapter.

Investigative Question 3

What data science software tools are currently available for data cleaning tasks, both commercial and open source?

There are many commercially licensed software tools that are available to tackle data preparation and cleaning tasks. Some of the more common commercial-off-the-shelf (COTS) tools include Microsoft Excel, Tableau, Trifacta, Informatica Axon, Paxata, Statistical Analysis System (SAS), IBM Statistical Package for Social Sciences (SPSS), and Stata (Cao, L., 2017; "Top 21," 2020). There are also many free open source software (FOSS) options as well, allowing individuals with coding experience to build tailor-made solutions to meet their dataset's needs. Popular programming languages for data science applications include R, Python, SQL, and Scala (Cao, L., 2017; "TIOBE Index," 2020).

Investigative Question 4

How do the different data cleaning alternatives compare?

Although not particularly well suited to large-scale automated data cleaning efforts, the prevalence of and widespread familiarity with Microsoft Excel deserves mention. It allows straightforward viewing and manipulation of data in a two-dimensional tabular format, as well as manual data cleaning actions and exploratory visualizations (Moeschlin, 2018). With the inclusion of the "Get & Transform" set of

features in the Data tab (formerly available as the Power Query Add-In), it is even easier to import data from a variety of sources and to transform the data during initial preparation (“Get & Transform in Excel,” undated). This set of features even allows a user to retrieve data from Apache Hadoop, a powerful big-data framework with a distributed data storage system allowing access to data sets too large for any one machine (Carlson, 2019). However, Excel is limited in the size of data sets it can consolidate into one worksheet, with each worksheet able to contain up to a maximum of 1,048,576 rows and 16,384 columns, and up to 32,767 characters per cell ("Excel specifications and limits," undated). Between this limitation and the necessity for data to be contained on the machine running the software (no distributed data storage post-import), Excel is not suited for big data analytics, even with the Hadoop tie-in. Lastly, because calculation occurs for all cells referenced in a formula when it is run for an array ("Guidelines and examples,” undated), which is computationally inefficient for large data sets, this can vastly reduce the speed with which an array can be manipulated compared to some of the more capable analytics tools available.

Tableau is a commonly used visualization tool used for business intelligence applications in the private sector and used by analysts and academics in the Air Force in limited numbers. It also has a variety of data cleaning functions as well as a

“recommendations” function which offers suggested cleaning operations for the user to choose ("Clean and Shape Data," undated).

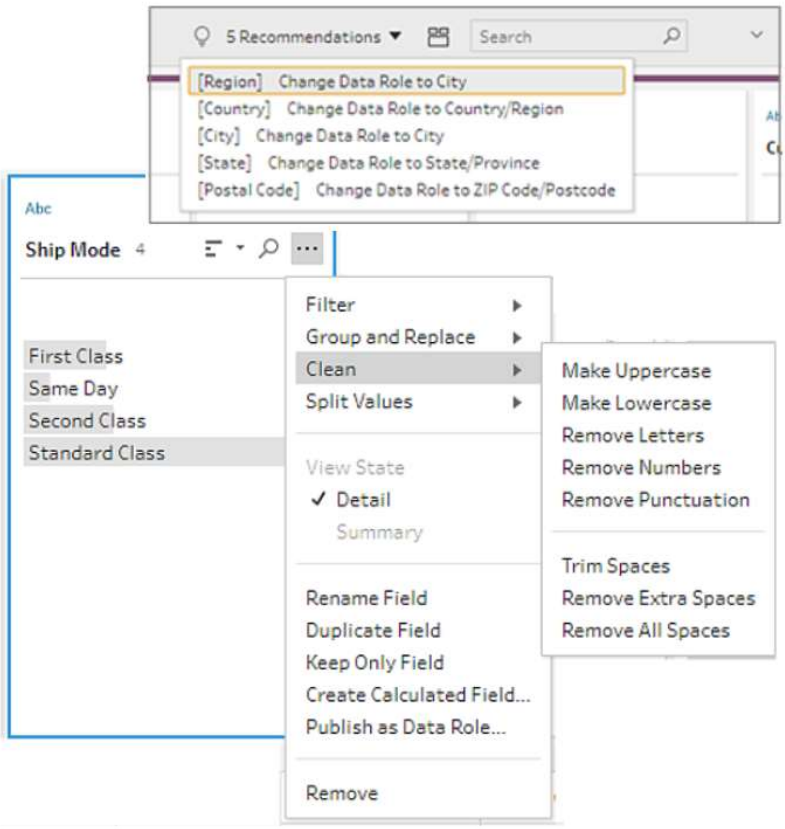


Figure 9 Example of Available Tableau Cleaning Options ("Clean and Shape Data," undated)

Tableau also allows the user to pause data updates while cleaning to save processing time ("Clean and Shape Data," undated), an advantage over Microsoft Excel’s resource-intensive approach. Furthermore, Tableau is capable of handling data of any size by supporting native data connectors to sources (“Tableau and Big Data,” undated). This allows Tableau to interface with big data frameworks including Hadoop, allowing users to work with large distributed data sets using parallel data processing (Carlson, 2019), and Apache Spark, a fast and flexible computational analytics engine which offers

data processing speeds up to one hundred times faster than Hadoop (spark.apache.org). However, despite these capabilities, Tableau was created first and foremost as a data analysis and visualization tool (Carlson, 2019), while the data preparation toolset was added relatively recently in 2018 (Van Loon, 2018). Lastly, although Tableau has no explicitly defined row or column limitations, there are known issues when importing large CSV files, limiting the number of columns which can be imported to 255, with a maximum of 255 characters per field ("Limitations to Data," 2019). This is further limited by the fact that Tableau allows users to create tables to view only up to a maximum of 50 columns ("Tableau 2019.4," undated).

Another commercial option for data cleaning is the relatively new data wrangling tool, Trifacta, founded in 2012, which originated from the Stanford Data Wrangler prototype created in 2008 by computer science professors who wanted to address the time consuming process of data preparation in order to provide refined data easily fed into downstream analytical packages (Black, 2016). Trifacta is a self-service data wrangling software platform for non-technical users specifically designed to ease the process of data preparation, including structuring, exploring, and cleaning the data. It also provides big-data functionality via compatibility with Hadoop, as well as offering the user suggestions for cleaning operations ("Trifacta," undated).

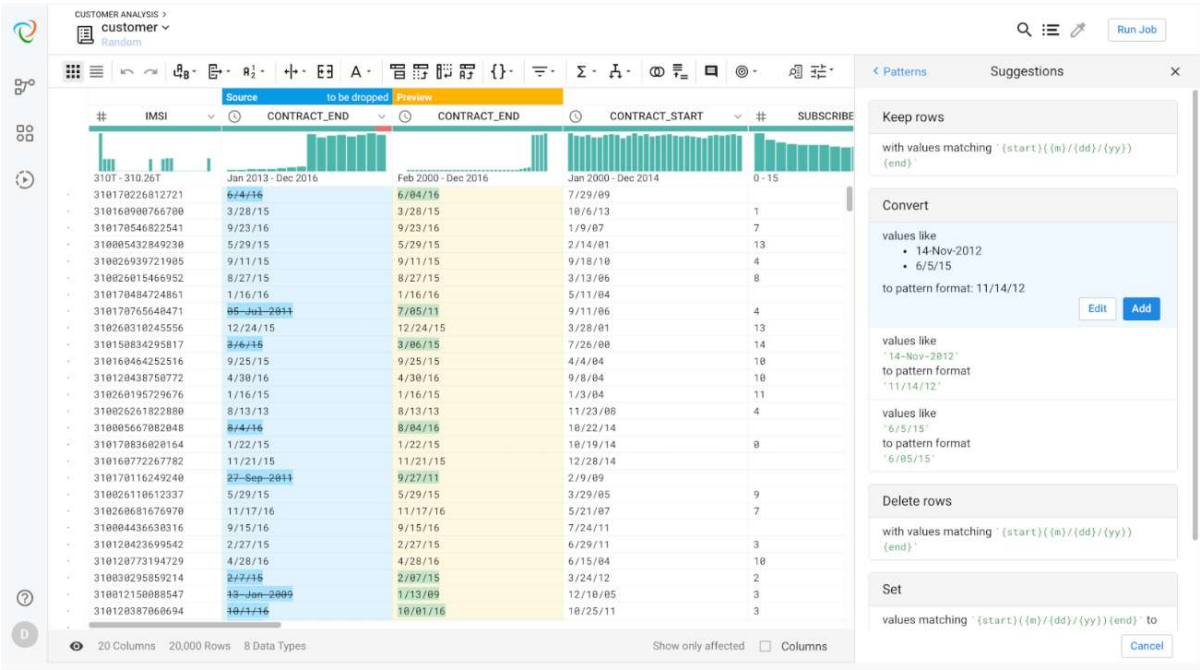


Figure 10 Example of Cleaning Operations in Trifacta ("Trifacta," undated)

Trifacta has three core components: interactive profiling, which allows users to understand the data's shape, size, structure, and outliers; predictive transformation, which provides tailored recommendations to perform data cleaning operations; and intelligent execution, which compiles the user's data cleaning steps and actions to Apache Spark to "[leverage] the power of the computing environment to do the work" across the entire cluster (Black, 2016). This makes Trifacta particularly well-suited to cleaning big data, with no limits on data that can be accessed or viewed through the application, an advantage over Excel and Tableau.

Lastly, SAS is a well-established, highly capable data science software suite specifically designed for statistical modeling, well suited to performing data cleaning operations on data sets of all sizes. Developed in the 1960s ("SAS Analytics Software," undated), it has been the historical industry standard for decades (Shah, 2016) and is favored by large companies for its reliability and company-backed support ("Essential Data Science," 2019). Additionally, it is "extremely efficient at sequential data access" (Shah, 2016) and offers a wide selection of statistical functions in an accessible graphical user interface, making it a capable and fast tool for cleaning large data sets. However, with even the most basic licensing package starting at $8,700 per user per year, (Dinsmore, 2014), there is ample justification to consider equally capable tools using open source software.

While the data cleaning capabilities offered by Tableau, Trifacta, SAS, and other commercial software make data cleaning and preparation tasks accessible to users without hard coding or data science expertise, they come at a cost. Additionally, with closed-source proprietary software, the lack of access to source code means that there are less opportunities for customized data cleansing algorithms. For this reason, open-source programming languages such as Python and R are often preferred by data scientists and other data experts who are able to leverage the power of these platforms directly (N. Shields, personal communication, May 14, 2020) by coding their own algorithms to best suit their objectives.

Both R and Python are free, open source, high-level programming languages with large, active development communities and a plethora of data science libraries and

modules which offer the same analytic functionality as any of the commercial products and more (Dinsmore, 2014; Shah, 2016; Carlson, 2019). Python libraries such as Pandas, SciKit-Learn, NumPy, SciPy, Matplotlib, and Seaborn offer a wide variety of data science capabilities, from data preparation and cleaning to data analysis, visualization, and machine learning tasks (Al-Jabery, Obafemi-Ajayi, Olbricht, and Wunsch, 2020). Tkinter ("Tkinter," 2019) and WxPython ("Welcome to WXPython," 2020) provide the capability to build graphical user interfaces, allowing even those without coding experience to tap into the data science capabilities Python has to offer. R offers similar data cleaning capabilities through its TidyVerse library of packages but has lagged in popularity compared to Python ("TIOBE Index," 2020) due to its reputation as having a steeper learning curve (Carlson, 2019) as compared to Python's simpler, more intuitive syntax.

Python versus Excel: A Five-Operation Comparison

To provide a quantitative comparison between available options for data cleaning, the researcher chose to compare the default tool most widely used across the Air Force, Microsoft Excel, with the most popular open-source option, Python. To accomplish this, the researcher tested both options for five basic operations useful during data preparation and cleaning: importing/loading the data set; auditing the data via preliminary visualizations, both single-attribute and for all attributes; missing value (null) handling; and exporting the data for downstream analysis. The data set used for testing was a 64.9-megabyte comma separated values (CSV) file containing historical GDSS II data from

2012. This file contained 202,550 rows and 46 columns of data, providing a sufficiently large data set to provide ample challenge for the tools to be tested for the selected operations while not overwhelming the processing power of a single computer.

The test was conducted on a Lenovo IdeaPad S340 with 8 gigabytes of RAM and an Intel Core i7-8565U quad-core processor with a 1.8GHz base clock speed. Each operation was executed individually and asynchronously with no programs running or open other than Python (via Spyder integrated development environment (IDE) launched from Anaconda Navigator), Microsoft Excel, and Windows Task Manager (to ensure no unexpected background processes would skew the results). The researcher used Dr. Reiman's Python Data Science Application (DSA) as an interface to accomplish each operation. Times were measured for each operation using a digital stopwatch, and similar functions were tested in alternating fashion between Excel and Python to preclude bias due to user reaction time or external conditions. All tests were performed while battery was at 100% and plugged in to a power supply. For operations that could be accomplished in one step, measurement began at the moment the operation command was executed and does not include the time necessary for the user to navigate to and select the desired operation after selecting the data.

The first test conducted was importing the file into both Python and Excel, which was executed and timed thirty times each (sample size, $n = 30$). It took longer to load the file in the Python Data Science Application (average of 29.6 seconds to open) than it did in Excel (average of 7.77 seconds to open), although this was the only test where Python was slower than Excel. Results are summarized in the table below.

Table 3. Import

Test: Open File (seconds)		
	Python	**Excel**
Avg	29.57433	7.765667
Std Dev	0.559302	0.17224
n	30	30

The second test conducted involved creating a histogram of a single attribute (column of data) as would be done during data auditing to gain preliminary understanding of the shape of the variable's distribution. For this test, a single histogram for "Departure Latitude" was created; once data was selected, this was a two-click operation for Python using Dr. Reiman's Data Science Application (DSA) and three

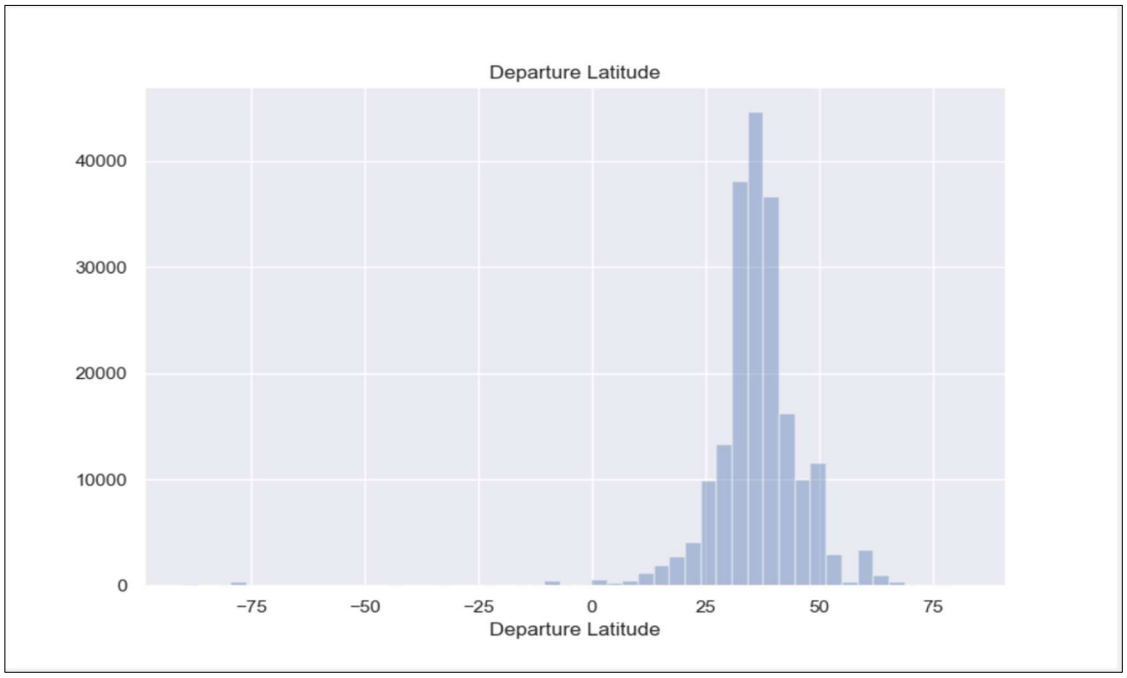


Figure 11 Example of Histogram Created using Python's Seaborn

clicks for Excel. Creating a single histogram was nearly instantaneous (<1 second) for both Python and Excel; examples and a summary table are included below.

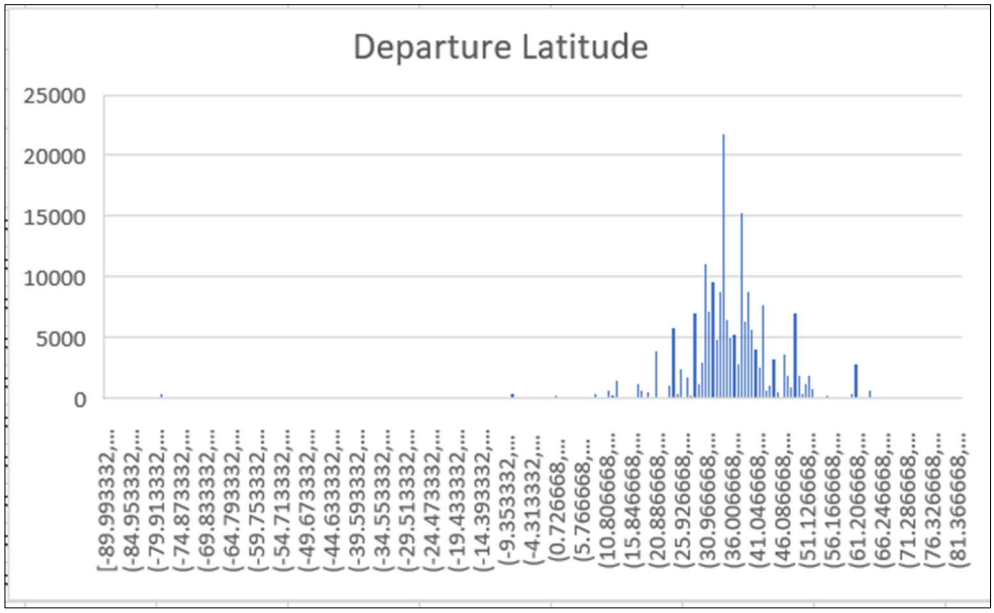


Figure 12 Example of Histogram Created using Excel

Table 4. Visualize Distribution (Single Histogram)

Test: Single Histogram (seconds)		
	Python	Excel
Avg	0.524333	0.764333
Std Dev	0.054689	0.129367
n	30	30

The third test involved visualizing the distributions of all the columns via histograms or bar graph as appropriate for the data type, as would be useful during data auditing to understand the shape of the data. This operation took an average of 5.5 seconds with Python, but this was a much more involved process in Excel. Visualizing all distributions was a two-click operation for Python DSA, but in Excel each chart must

be created individually for each column in Excel (46 columns for this dataset). This required creating a pivot table (~10.6 seconds), then a frequency table (~5.49 seconds), then the bar graph itself. This last step, creating the bar graph itself from the pivot table, a four-click operation, was measured 46 times to create graphs for all columns. Additionally, in order to view all charts together at the end, it was necessary to allocate time (10 seconds) to copy and paste each chart into a separate location since using a pivot table meant only one chart was displayed at a time. This ten-second provision for copying each chart was more efficient than creating new pivot and frequency tables for each chart, which would have taken even more time. Even so, it took 794.4 seconds (over 13 minutes) to accomplish in Excel what it took Python 5.5 seconds to do. In practice, this operation would take even longer since this does not include the user navigating to create different kinds of charts, whether a frequency count for discrete data or bar graph for discrete, which would add even more time.

Table 5. Visualize All Distributions (46 Attributes)

Test: All Columns (seconds)		
	Python	**Excel**
Avg	5.491667	794.4
Std Dev	0.183399	N/A
n	30	1

The fourth operation tested was missing value (null) handling; specifically, the listwise-deletion cleaning operation, where null values for a particular column are identified and then the entire observation (row) is removed. For this operation, the researcher selected the "Primary Delay Hours" column to remove missing values, as one

would do if one were interested in performing further analysis only on missions which experienced delays.

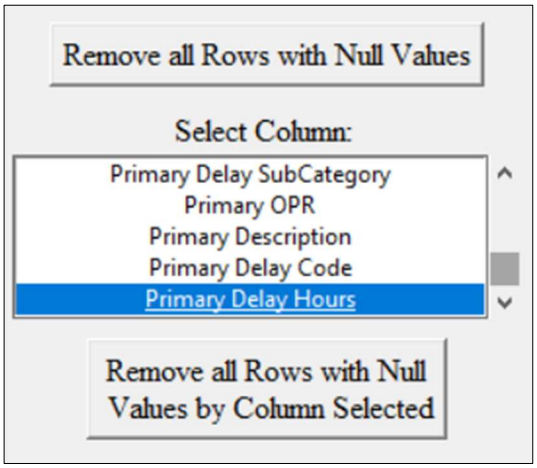


Figure 13 Example of Listwise Deletion using Python DSA

It took Python less than a second to perform the operation, removing a total of 142,935 null values, but it took Excel an average of five minutes and 42 seconds to accomplish the same task by first selecting blank cells using the “Go To - Special" function, then deleting rows containing blank fields in the column of interest. In fact, Excel warns the user prior to executing that the operation will take an extended period of time.

Table 6. Missing Value Handling (Listwise Deletion of 142,935 Rows)

Test: Remove Missing Values (seconds)		
	Python	**Excel**
Avg	0.776	342.66
Std Dev	0.177285	4.252464
n	5	5

The last operation tested was exporting the data for downstream analysis. Once data set manipulations were complete, the 64.9 MB file was saved as a CSV. Both programs were able to complete this task relatively quickly, with Python slightly faster at an average of 1.4 seconds compared to Excel's 3.1 seconds.

Table 7. Export Test

Test: Export		
	Python	**Excel**
Avg	1.391333	3.131333
Std Dev	0.375383	0.213336
n	15	15

In summary, although it took Python approximately 20 seconds longer to load the file, it outperformed Excel in every other test, significantly so in the simultaneous visualization of the distributions of all columns as well as in missing value handling via listwise deletion of rows containing nulls for the variable of interest. Even with this data set of moderate size, Python's advantage is evident; considering these results, as well as Excel's row and column limits, it is clear that Python is much better suited to tackle data cleaning tasks, especially for large data sets. A summary of the results is included below.

Table 8. Summary of Test Results

SUMMARY		
	Python	**Excel**
Open	29.57	7.77
Visualize Distribution (Single Attribute)	0.52	0.76
Visualize Distributions (All Columns)	5.49	794.40
Remove Nulls	0.78	342.66
Export	1.39	3.13
	(all times in seconds)	

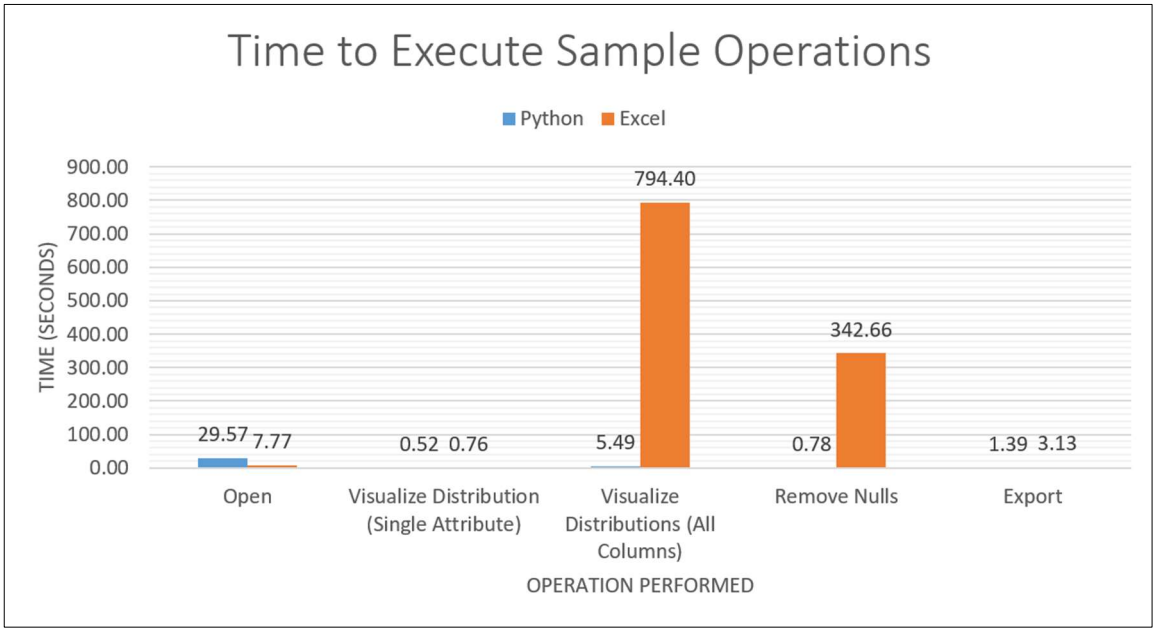


Figure 14 Test Results: Time to Complete Data Preparation & Cleaning Operations

Investigative Question 5

What should be the future Air Force approach to data science, including the data cleaning process?

The future of data science in the Air Force is bright: with the establishment of the Air Force Chief Data Office and launch of the VAULT Data Platform, there is a great opportunity to address the gaps previously identified, both with respect to data cleaning

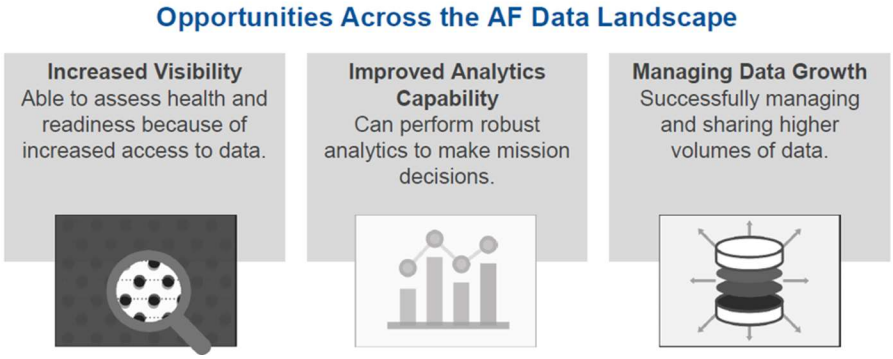


Figure 15 Data Opportunities (Vidrine, 2019)

and also to the strategic approach to data and data science in general. The Chief Data Office has recognized and outlined three opportunities for the future of data in the Air Force: increased visibility, improved analytics capability, and effectively managed data growth (Vidrine, 2019).

In order to achieve the goal of increased data visibility in the future, the Air Force needs to transition from legacy data storage systems which actively restrict access and prevent sharing of relevant data within the force to an open architecture which makes data "secure, visible, accessible, understandable, linked, and trusted" (SVAULT) ("Data Services Reference Architecture," 2019). These principles form the basis for the creation of the Air Force VAULT Data Platform, which will enable sharing of data across platforms and functional communities via access to a secure cloud-based storage layer called the *data lake* ("VAULT Data Platform Capabilities," 2019).

An improved analytics capability in the near future is possible with improved access to data via the VAULT's data lake, as well as access to the proper tools necessary to accomplish data cleaning and analysis tasks. As discussed previously, several COTS and FOSS tools exist to empower Airmen to prepare and clean data to make it fit for further analysis, but access on Air Force networks was limited. With access to the VAULT Data Platform, users will have access to a variety of cloud-based data science tools, both by default (for FOSS tools) and by request (for COTS incurring licensing costs). Tools useful for data cleaning and preparation which are currently available include COTS tools such as Tableau and Informatica Axon, as well as access to FOSS

such as Apache Zeppelin, a browser-based notebook which allows a user to code in both Python and R ("VAULT Data Platform Capabilities," 2019).

In addition to offering access to data and the tools with which to prepare and clean the data, the future Air Force approach should address the issue of expertise. Not only will the Air Force continue to need to employ trained and qualified data scientists in key roles (Otto, Kimminau, and Ray, 2016), but also the Air Force needs to improve the level of data literacy within the service through education and training. Recommendations to address this issue will be offered in the next chapter.

Lastly, a successful future approach to data will involve effective management of the growth of data to ensure that it remains accessible and visible between functional communities through effective sharing. While the VAULT's data lake can provide a repository for such data, it is not at this point intended to be a comprehensive source for all Air Force data, nor to take authority away from data owners, creators, or stewards; rather, it offers a conducive environment for the sharing and leveraging of data within and between organizations (N. Shields, personal communication, May 14, 2020; R. Synakowski, personal communication, May 18, 2020). Thus, a framework for effective data sharing and careful management of metadata will be necessary to achieve this goal in the future ("Data Services Reference Architecture," 2019).

Research Question

How can the Air Force empower Airmen to leverage data by facilitating the process of data cleaning to make it fit to purpose?

The Air Force can empower Airmen to leverage their data by recognizing the current identified gaps and addressing them; specifically, this can be achieved by providing Airmen with the data, the tools, and the skills necessary to manage, understand, prepare, and clean their data. Creating a culture of data literacy starts with a cohesive data strategy (Insinna, 2018). The evidence suggests that with the creation of the Air Force Chief Data Office and the roll-out of the VAULT Data Platform, the Air Force is making great strides in addressing the capability gaps in how Airmen at all levels approach data science related activities, including data preparation and cleaning. With this in mind, there are several recommendations which the researcher will propose in the next chapter to further enhance the ability of Airmen in staff and operational roles to clean data to make it fit to purpose, even for those with limited data science or analytics expertise.

Summary

This chapter addressed each of the investigative questions, as well as the research question. In the next chapter, the author will identify further opportunities and offer recommendations to address the goal implied by the research question: empowering Airmen to leverage data.

V. Conclusions and Recommendations

Chapter Overview

This chapter provides recommendations to achieve the objective of enhancing the ability of Airmen in staff and operational roles to leverage data, starting by cleaning and preparation task to make it fit to purpose, even for those Airmen without data science or analytics expertise.

Conclusions

With the return of great power competition as outlined in the 2018 National Defense Strategy, American dominance in air, space, and cyberspace cannot be assumed to continue while maintaining the status quo ("Digital Air Force," 2019). The processes and tools used in past eras to deal with information are being rapidly replaced with a new paradigm, one in which *big data* and worldwide interconnectedness are transforming society and the global economy. The Air Force must modernize its information architecture and transform the way it leverages data in order to maintain its strategic advantage ("Data Services Reference Architecture," 2019).

To this end, the Air Force has made great strides within the last year to address some of its most significant gaps in data visibility and has laid the framework to address the issue of analytics capabilities. Furthermore, with the establishment of the Chief Data Office, the Air Force has acknowledged the value that data science brings to the fight and the necessity of having a strategic focal point for policies, procedures, structure, and guidance to ensure a cohesive, unified approach to managing the explosion of data available today.

Significance of Research

The significance of the role of data in Air Force operations of the future cannot be overstated. The opportunities of this new era are clear. This research provides context and motivation for the necessity of empowering Airmen to tackle data analysis tasks by equipping them with the tools and knowledge necessary to understand the data they have and to prepare it to make it suitable to purpose. It also documents justification and support for the Air Force Chief Data Office's recent initiatives, and it offers recommendations for future actions and possible partnerships to cultivate a culture of data literacy throughout the force.

Recommendations for Action

Recommendation 1

In order to support the VAULT Data Platform roll-out and achieve the aforementioned strategic objectives, it is necessary to raise awareness among Airmen of the capabilities and opportunities it provides. Particularly among younger Airmen, who were raised in the digital generation, there are many who joined the service already having some level of coding and analysis capability (Insinna, 2018), so introducing this segment of the force to the capabilities offered by the VAULT Data Platform can fuel and enable grassroots-level innovation. Even what Airmen at the tactical level may lack in coding or analytics skills, they more than make up for in domain expertise: no one understands the nuances of operational data better than the operators themselves. Thus, there is value and opportunity in enabling them to leverage their domain expertise by introducing them to the tools and resources available to them. An ideal opportunity for

this would be during professional military education (PME) such as Squadron Officer School for company grade officers, who will spread their knowledge back in their operational squadrons, and intermediate developmental education for field grade officers, who will be better positioned to leverage data in their future staff or command roles. Whether this introduction entails merely an awareness of the capabilities and an opportunity to gain access, or a more involved course or seminar on analytics, presents an opportunity for future research in coordination with Air University.

Recommendation 2

The Air Force Institute of Technology has recognized the need for data analytics training and has offered analytics workshops since 2017 (Geiger, 2017). Recently, AFIT has developed a graduate certificate program in Data Science, as well as offering data analytics courses to qualified candidates ("Data Science Certificate," undated) from a variety of commands and backgrounds. There is an opportunity for partnership between the Chief Data Office and AFIT's data science program to educate and prepare Airmen to harness the power of data analytics, in addition to providing students access to the VAULT Platform's growing data lake and software capabilities.

Recommendation 3

The Data Service Reference Architecture (DSRA) recommends prioritizing FOSS tools to minimize the support burden ("Data Services Reference Architecture," 2019:10). The VAULT Data Platform both provides a Python coding environment as well as the

opportunity for tenants to promote their coded algorithms to production in the Air Force Common Computing Environment (CCE) to share with other users ("Air Force VAULT," undated). This provides an opportunity to create and provide front-end data science applications to make data cleaning more accessible to Airmen without data science backgrounds. An example of a tool which could serve this purpose is an application under development by Dr. Adam Reiman at AFIT which uses the Tkinter library in Python to create a user interface to provide a framework to allow a user to perform common data cleaning actions without necessarily being familiar with the code itself. This tool covers each step of the data science process from gathering and cleaning to modeling and exporting the data.

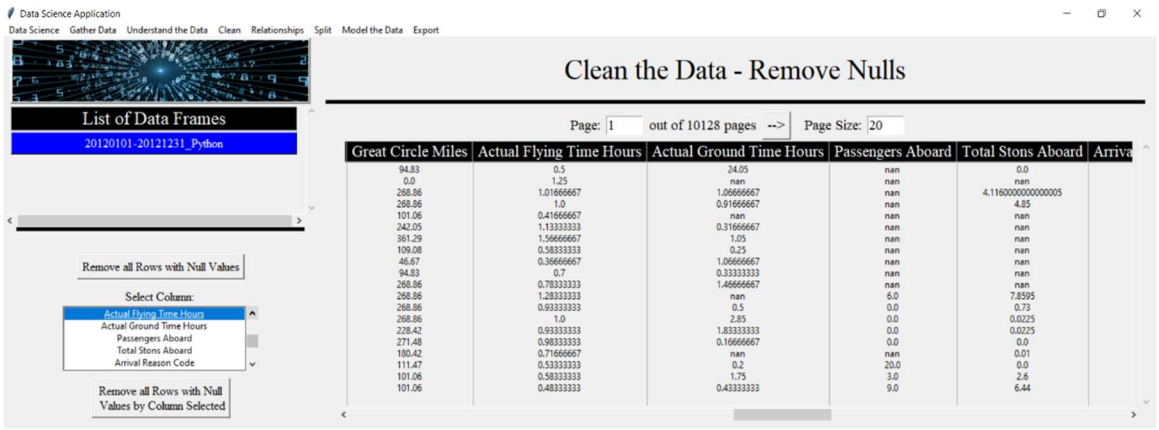


Figure 16 Dr. Adam Reiman's Data Science Application

Recommendations for Future Research

This research focused on the initial enabling steps of the data lifecycle, namely data preparation and cleaning, although many of the observations, conclusions, and recommendations are relevant to downstream analytics tasks and data science in general. Future research could address how the Air Force could approach machine learning,

neural networks, or artificial intelligence tasks. Another opportunity for future research could involve an Air Force Data Platform use case specific to a supply chain management scenario. An additional future research opportunity could involve an exploration of options for increasing data literacy in the Air Force through analytics training and education.

Summary

The Air Force Strategic Master Plan states that "true advances will come with rapid, accurate, shared situational awareness" (Department, 2015:47), and there is no greater opportunity today to achieve this than by revolutionizing the tactics, techniques, and procedures used to share, manage, and employ data within the Air Force, Department of Defense, and beyond. In order to leverage data as a strategic asset, the Air Force needs to support initiatives to redefine its data storage paradigm, leverage the expertise of data science professionals, increase data literacy throughout the service, and empower Airmen with the right tools to build the foundation of all analysis: clean, quality data.

DATA CLEANING FOR AIR FORCE APPLICATIONS

Maj Giovanna Espegio
Advisor: Dr. Adam Reiman, Col (Ret.), PhD
Advanced Study of Air Mobility (ENS)
Air Force Institute of Technology

Abstract

There is more data available to the Air Force now than ever before. Leveraging this data deluge for useful applications—making use of this data in meaningful ways—has been an ongoing challenge. There are numerous commercial off the shelf programs which have the power and capability to deal with large, complex data sets; however, use of these commercial products are limited by licensing costs, restrictions on proprietary information including source code, security, and the user's familiarity and proficiency with the software. Even the widely used Microsoft Excel program, while useful, has its limits in terms of data size and speed of processing. Additionally, the Air Force faces the challenges presented by outdated data architectures, restricted visibility and data sharing capabilities, and a limited number of data science experts in the force. The purpose of this research is to answer the question: "How can the Air Force empower Airmen to leverage data as a strategic asset by facilitating the process of data cleaning to make it fit to purpose?"

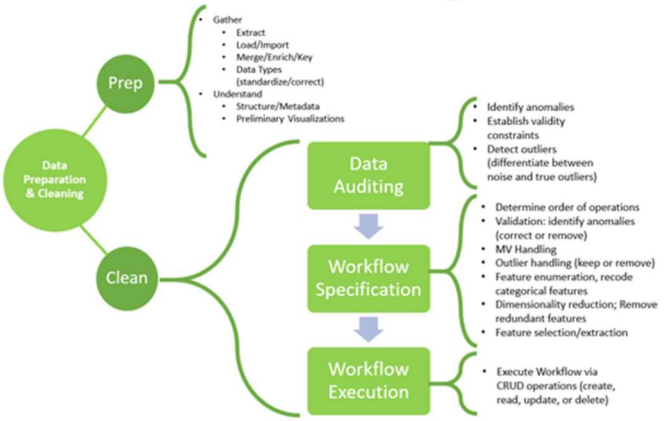
Methodology

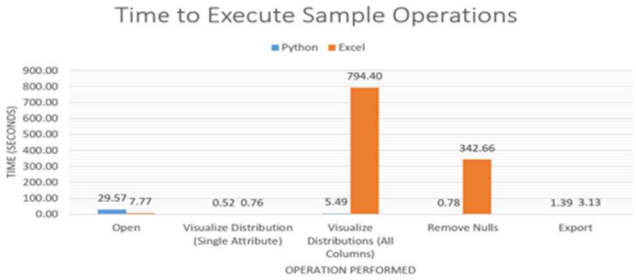
This research used a mixed methods approach, employing a narrative review of relevant literature, semi-structured interviews with data science professionals, and systematic testing of software alternatives.

Collaboration

SAF/IEN

AIR FORCE INSTITUTE OF TECHNOLOGY





Research Question

How can the Air Force empower Airmen to leverage data by facilitating the process of data cleaning to make it fit to purpose?

Significance

The significance of the role of data in Air Force operations of the future cannot be overstated. The opportunities of this new era are clear. This research provides context and motivation for the necessity of empowering Airmen to tackle data analysis tasks by equipping them with the tools and knowledge necessary to understand the data they have and to prepare it to make it suitable to purpose. It also documents justification and support for the Air Force Chief Data Office's recent initiatives, and it offers recommendations for future actions and possible partnerships to cultivate a culture of data literacy throughout the force.

Research Objective

The objective of this research is to discover ways to enhance the ability of Airmen in staff and operational roles to clean data to make it fit to purpose, even for those without data science or analytics expertise. This research focuses on the data cleaning aspect of data science as it is the first critical task that must be addressed before any bona-fide data analysis can take place. By many estimates, data cleansing can take between sixty to eighty percent of the overall time spent in data analysis (Pyle, 1996; Mayo, 2019). Specifically, this research seeks to understand opportunities within the Air Force to leverage the data science capabilities of various common proprietary data analysis tools, such as Microsoft Excel, SAS, Tableau, as well as free open-source software (FOSS) options such as solutions created using the R and Python programming languages.

Bibliography

"Air Force VAULT Data Platform: A modern data environment for the Digital Air Force Enterprise," *SAF/CO SharePoint*. https://org2.eis.af.mil/sites/13007/Pages/Data-Platform.aspx. Accessed 27 January 2020.

Al-Jabery, K. K., Obafemi-Ajayi, T., Olbricht, G. R., Wunsch II, D.C. *Computational Learning Approaches to Data Analytics in Biomedical Applications*. Academic Press: 2020. https://doi.org/10.1016/B978-0-12-814482-4.00009-7 Accessed 24 January 2020.

Bacon, F., Hutchins, R., & Adler, M. *Novum Organum, Great Books Of The Western World Vol. 9.* Encyclopedia Britannica, Inc.: 1952.

Black, Doug. "Data Wrangling 'Decoder Ring' Homogenizes Polyglot Data Lakes," *Enterprise AI: Advanced Computing in the Age of AI*. https://www.enterpriseai.news/2016/02/11/trifactas-data-wrangling-decoder-ring-homogenizes-polyglot-data-lakes/. 11 February 2016. Accessed 17 May 2020.

Bonthu, S., and Bindu, K. H. "Review of Leading Data Analytics Tools," *International Journal of Engineering & Technology,* 7:10-15 (26 August 2018).

Burwell, Sylvia M., VanRoekel, Steven, Park, Todd, and Mancini, Dominic J., "Open Data Policy—Managing Information as an Asset," Executive Office of the President, Office of Management and Budget. Memorandum for the Heads of Executive Departments and Agencies. M-13-13. Washington, D.C., 9 May 2013.

Cao, Longbing. "Data Science: A Comprehensive Overview." *ACM Computing Surveys*, *50*(3), 1–42, 2017. https://doi.org/10.1145/3076253 Accessed 24 January 2020.

Carlson, Esther. "What Technology Tools Can Help Data Scientists?" *Dis.co.* https://dis.co/blog/what-technology-tools-can-help-data-scientists/. 13 November 2019. Accessed 16 May 2020.

"Clean and Shape Data," *Tableau Prep Help, Tableau.* https://help.tableau.com/current/prep/en-us/prep_clean.htm. Accessed 16 May 2020.

Cleveland, William S. "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics," International Statistical Review, 2001.

Cokluk, Omay, and Kayri, Murat. "The Effects of Methods of Imputation for Missing Values on the Validity and Reliability of Scales," *Educational Sciences: Theory & Practice,* 11(1): 303-309 (Winter 2011).

"Data Catalog," *DATA.GOV*, https://catalog.data.gov/dataset. 2020. Accessed 4 May 2020.

"Data Cleaning: In-Depth Guide." *AI Multiple*. 2 May 2020. https://research.aimultiple.com/data-cleaning/ Accessed 7 May 2020.

"Data Science Certificate," *Air Force Institute of Technology.* https://www.afit.edu/EN/programs.cfm?a=view&D=60. Accessed 16 May 2020.

"Data Services Reference Architecture," *Air Force Chief Data Office (SAF/CO)*. https://org2.eis.af.mil/sites/13007/Documents/Service_Reference_Architecture_42519.pdf. March 2019. Accessed 13 May 2020.

Department of the Air Force. *Air Force Supplement to the DoD Dictionary of Military and Associated Terms.* AFDD 1-2. Washington: HQ USAF, 2012.

Department of the Air Force. *Information Management.* AFDD 33-3. Washington: HQ USAF, 2006.

Department of the Air Force. *Information Operations.* AFDD 3-13. Washington: HQ USAF, 2011.

Department of the Air Force. *Information Operations.* AFPD 10-7. Washington: HQ USAF, 2014.

Department of the Air Force. *USAF Strategic Master Plan.* Washington: HQ USAF, 2015.

"The Digital Air Force," *United States Air Force White Paper.* https://www.af.mil/Portals/1/documents/2019%20SAF%20story%20attachments/USAF%20White%20Paper_Digital%20Air%20Force_Final.pdf?ver=2019-07-09-181813-390&timestamp=1562710801965. July 2019. Accessed 14 May 2020.

Dinsmore, Thomas W. "SAS Versus R Part Two," *ML/AI Machine Learning/Artificial Intelligence.* https://thomaswdinsmore.com/2014/12/15/sas-versus-r-part-two/. 15 December 2014. Accessed 7 May 2020.

Eddins, J.M. Jr. "CHANGING THE STORY: Kessel Run Supports Warfighters While Cracking the Code of Agile Software Development and Acquisition for the Air Force," *Airman Magazine*, https://airman.dodlive.mil/2018/11/05/changing-the-story/. 5 November 2018. Accessed 7 May 2020.

Emerson, Noah. "AFSPC Rolls Out New Enterprise Data Strategy," *Air Force Space Command Public Affairs.* https://www.afspc.af.mil/DesktopModules/ArticleCS/Print.aspx?PortalId=3&ModuleId=599&Article=1927415. 6 August 2019. Accessed 13 May 2020.

"Essential Data Science Ingredients: 14 Most Used Data Science Tools for 2019," *Data Flair,* https://data-flair.training/blogs/data-science-tools/. 23 May 2019. Accessed 7 May 2020.

"Excel Specifications and Limits," *Microsoft Office*, https://support.office.com/en-us/article/excel-specifications-and-limits-1672b34d-7043-467e-8e27-269d656771c3. Accessed 29 April 2020.

Executive Order No. 13642, *78 Federal Register 28111*, 14 May 2013.

"Federal Data Strategy: 2020 Action Plan," *President's Management Agenda*. https://strategy.data.gov/. Accessed 4 May 2020.

Geiger, Stacey. "Data Analytics Can Help Air Force Become More Effective," *88th Air Base Wing Public Affairs,* https://www.maxwell.af.mil/DesktopModules/ArticleCS/Print.aspx?PortalId=62&ModuleId=24565&Article=1180030. 11 May 2017. Accessed 7 May 2020.

"Get & Transform in Excel," *Microsoft Office*. https://support.office.com/en-us/article/get-transform-in-excel-881c63c6-37c5-4ca2-b616-59e18d75b4de. Accessed 7 May 2020.

Goel, Aman. "What is an R Data Frame?" *Magoosh*. https://magoosh.com/data-science/what-is-an-r-data-frame/. 15 December 2017. Accessed 19 May 2020.

Gidley, Scott. "What Is Metadata and Why Is It Critical in Today's Data Environment?" *DZone: A Devada Media Property*. https://dzone.com/articles/what-is-metadata-and-why-is-it-critical-in-todays. 20 February 2017. Accessed 18 May 2020.

"Guidelines and examples of array formulas," *Microsoft Office*, https://support.office.com/en-us/article/guidelines-and-examples-of-array-formulas-7d94a64e-3ff3-4686-9372-ecfd5caa57c7. Accessed 16 May 2020.

Hamilton, S. P., & Kreuzer, M. P. "The Big Data Imperative: Air Force Intelligence for the Information Age." *Air & Space Power Journal*, *32*(1), 4–20, 2018. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=asn&AN=128199511&site=ehost-live Accessed 24 January 2020.

Han, Na-Rae. "Data Types and Conversion." University of Pittsburgh. https://www.pitt.edu/~naraehan/python3/data_types_conversion.html

Insinna, Valerie. "How the Air Force data strategy is evolving," C4ISRNET. https://www.c4isrnet.com/intel-geoint/isr/2018/08/30/how-the-air-force-data-strategy-is-evolving/. 30 August 2018. Accessed 7 May 2020.

"IO Tools," *Pandas*. https://pandas.pydata.org/pandas-docs/stable/user_guide/io.html. Accessed 21 May 2020.

Kim, Hyon. "Data.gov at Ten and the OPEN Government Data Act," DATA.GOV, https://www.data.gov/meta/data-gov-at-ten-and-the-open-government-data-act/. 31 May 2019. Accessed 4 May 2020.

Kimminau, Jon A. "Five Examples of Big Data Analytics and the Future of ISR." *JFQ: Joint Force Quarterly*, (77), 30–31, 2015. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=asn&AN=102583697&site=ehost-live Accessed 24 January 2020.

Kononow, Piotr. "Data Terminology: What is Metadata?" *Dataedo.* https://dataedo.com/kb/data-glossary/what-is-metadata. 16 September 2018. Accessed 18 May 2020.

"Limitations to Data and File Sizes with Jet-based Data Sources," 2019. Tableau. https://kb.tableau.com/articles/Issue/limitations-to-data-and-file-sizes-with-jet-based-data-sources?_ga=2.72133078.45852630.1590100664-1131258224.1590100664. Accessed 17 May 2020.

Mayo, Josh. "DoD Working on Joint Data Strategy," MeriTalk.com. https://www.meritalk.com/articles/dod-working-on-joint-data-strategy/. 2019. Accessed 7 May 2020.

McClave, James T., Benson, P. George, Sincich, Terry. *Statistics for Business and Economics*. Pearson, 2014.

McKinney, W. *Python for Data Analysis*. O'Reilly Media. 2nd Kindle Edition, 2017.

Moeschlin, Fredrik. "Excel for Data Science?" *Towards Data Science.* https://towardsdatascience.com/excel-for-data-science-a82247670d7a. 26 June 2018. Accessed 16 May 2020.

Müller, Heiko, and Freytag, Johann-Christoph. "Problems, Methods, and Challenges in Comprehensive Data Cleansing," *Humboldt-Universität zu Berlin*. 2003.

O'Neil, Cathy and Rachel Schutt. Doing Data Science: Straight Talk from the Frontline. O'Reilly Media, Inc: 2014.

Otto, Robert P., Kimminau, Jon A., Ray, Brian J. *Data Science and the USAF ISR Enterprise.* Deputy Chief of Staff, Intelligence, Surveillance and Reconnaissance, https://defenseinnovationmarketplace.dtic.mil/wp-content/uploads/airforce/Data_Science_and_the_USAF_ISR_Enterprise-_White_Paper.pdf. March 2016. Accessed 3 May 2020.

Paré G, Kitsiou S. “Chapter 9: Methods for Literature Reviews,” In: Lau F, Kuziemsky C, (editors). *Handbook of eHealth Evaluation: An Evidence-based Approach*, University of Victoria; 27 February 2017. Available from: https://www.ncbi.nlm.nih.gov/books/NBK481583/

Press, Gil. “A Very Short History of Data Science,” *Forbes.* https://www.forbes.com. 28 May 2013. Accessed 13 May 2020.

"Public Release Announcement of the Air Force Data Services Reference Architecture," Air Force Chief Data Office. https://www.af.mil/DesktopModules/ArticleCS/Print.aspx?PortalId=1&ModuleId=850&Article=1842061. 8 May 2019. Accessed 16 May 2020.

Pyle, D. *Data Preparation for Data Mining (1st Ed.)*. Morgan Kaufmann, 1999.

"The Python Standard Library," *Python*, https://docs.python.org/3/library/. 18 May 2020. Accessed 21 May 2020.

Rahm, Erhard, and Do, Hong H. “Data Cleaning: Problems and Current Approaches,” *Data Engineering*, *IEEE Computer Society*, 23(4): 3-13 (December 2000).

Rogati, Monica. “The AI Hierarchy of Needs.” *Medium, HackerNoon.com,* 1 Aug 2017, medium.com/hackernoon/the-ai-hierarchy-of-needs-18f111fcc007. Accessed 18 April 2020.

Rouse, Margaret. “Extract, Load, Transform (ELT),” *Tech Target*. https://searchdatamanagement.techtarget.com/definition/Extract-Load-Transform-ELT. January 2020. Accessed 18 May 2020.

“SAS Analytics Software & Solutions,” *SAS*, https://www.sas.com/en_us/home.html. Accessed 7 May 2020.

Shah, Aatash. “R, Python or SAS: Which one should you learn first?” *Data Science Central: The Online Resource for Big Data Practitioners*. https://www.datasciencecentral.com/profiles/blogs/r-python-or-sas-which-one-should-you-learn-first. 1 November 2016. Accessed 17 May 2020.

Shields, Nicholas S. Data Scientist, SAF/CO Air Force Chief Data Office. Telephone interview. 14 May 2020.

Sirota, Sara. “Air Force Holding Industry Day for VAULT Data Support Contract,” Inside Defense. https://insidedefense.com/insider/air-force-holding-industry-day-vault-data-support-contract. 28 October 2019. Accessed 4 May 2020.

Stevens, S. S. “On the Theory of Scales of Measurement,” *Science, New Series, American Association for the Advancement of Science,* 103(2684): 677-680 (7 June 1946).

Synakowski, Ron J. Deputy Director, Strategy, SAF/CO Air Force Chief Data Office. Telephone interview. 18 May 2020.

"Tableau 2019.4," undated. Tableau. https://www.tableau.com/2019-4-features

“Tableau and Big Data: An Overview,” Tableau. https://www.tableau.com/learn/whitepapers/tableau-big-data-overview. Accessed 17 May 2020.

“TIOBE Index for April 2020,” *TIOBE*. https://www.tiobe.com/tiobe-index/. April 2020. Accessed 26 April 2020.

"Tkinter," *Python.org*. https://wiki.python.org/moin/Tkinter. 6 Dec 2019. Accessed 14 January 2020.

“Top 21 Self Service Data Preparation Software,” *Predictive Analytics Today*. https://www.predictiveanalyticstoday.com/data-preparation-tools-and-platforms/. 2020. Accessed 16 May 2020.

"Trifacta," *Trifacta*. https://www.trifacta.com/. Accessed 14 May 2020.

United States Congress. *Foundations for Evidence-Based Policymaking Act of 2018*. Public Law No. 115—435, 115th Congress, 1st Session. Washington: GPO, 2019.

Van Loon, Monica. “Tableau’s New Data Prep Tool: 10 Killer Features,” *SENTURUS*. https://senturus.com/blog/10-features-tableau-data-prep-maestro/. 2018. Accessed 17 May 2020.

“VAULT Data Platform Capabilities,” *Air Force Chief Data Office*. https://org2.eis.af.mil/sites/13007/Pages/Data-Platform.aspx. 2019. Accessed 13 May 2020.

Vidrine, Eileen. “Air Force Chief Data Office Overview,” HQ USAF. https://org2.eis.af.mil/sites/13007/Documents/CDO%20Overview_FINAL.11.27.19.pdf. 27 November 2019. Accessed 10 May 2020.

“Welcome to WXPython,” *WxPython*. https://wxpython.org/. 24 April 2020. Accessed 7 May 2020.

Wickham, Hadley. “Tidy Data.” *Journal of Statistical Software*, 59(10) (August 2014).

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From – To)*
05-06-2020	Graduate Research Paper	JUL 2019 – JUN 2020

4. TITLE AND SUBTITLE
DATA CLEANING FOR AIR FORCE APPLICATIONS

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Espegio, Giovanna, Major, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)
Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/ENS)
2950 Hobson Way, Building 640
WPAFB OH 45433-8865

8. PERFORMING ORGANIZATION REPORT NUMBER
AFIT-ENS-MS-20-J-033

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Jordan Eccles, PhD, SAF/IEN Current Operations CTR
Office of the Deputy Assistant Secretary of the Air Force, Operational Energy
Pentagon
Washington DC 20318
202-802-5861. jordan.eccles.ctr@us.af.mil

10. SPONSOR/MONITOR'S ACRONYM(S)
SAF/IEN

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution Statement A. Approved For Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
There is more data available to the Air Force now than ever before. Leveraging this data deluge for useful applications—making use of this data in meaningful ways—has been an ongoing challenge. There are numerous commercial off the shelf programs which have the power and capability to deal with large, complex data sets; however, use of these commercial products are limited by licensing costs, restrictions on proprietary information including source code, security, and the user's familiarity and proficiency with the software. Even the widely used Microsoft Excel program, while useful, has its limits in terms of data size and speed of processing. Additionally, the Air Force faces the challenges presented by outdated data architectures, restricted visibility and data sharing capabilities, and a limited number of data science experts in the force. The purpose of this research is to answer the question: "How can the Air Force empower Airmen to leverage data as a strategic asset by facilitating the process of data cleaning to make it fit to purpose?"

15. SUBJECT TERMS
Data, data science, data cleaning, data cleansing, data preparation, Python, analytics, big data, VAULT

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Reiman, Adam D., Col (R), Ph.D., AFIT/ENS
U	U	U	UU	83	19b. TELEPHONE NUMBER *(Include area code)* (609) 975-2782 x #### (Reiman.Adam@gmail.com)

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18