2022

Brennan, Andrea

SYRACUSE UNIVERSITY

Applied Masters in Data Science

Andrea Brennan

Table of Contents

[Scope 2](#_Toc98139517)

[Fundamentals 2](#_Toc98139518)

[Describe a broad overview of the major practice areas of data science 3](#_Toc98139519)

[Collect and organize data 4](#_Toc98139520)

[Identify patterns in data via visualization, statistical analysis, and data mining 5](#_Toc98139521)

[Develop alternative strategies based on the data 7](#_Toc98139522)

[Develop a plan of action to implement the business decisions derived from the analyses 8](#_Toc98139523)

[Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization 9](#_Toc98139524)

[Synthesize the ethical dimensions of data science practice (e.g., privacy) 9](#_Toc98139525)

[Project Samples 10](#_Toc98139526)

[US Vaccination Rates (IST 772 Quantitative Reasoning) 10](#_Toc98139527)

[Summary 10](#_Toc98139528)

[Analysis & Models 11](#_Toc98139529)

[Full Project Resources 11](#_Toc98139530)

[Women’s Clothing in E-Commerce (IST 707 Data Analytics) 12](#_Toc98139531)

[Summary 12](#_Toc98139532)

[Analysis & Models 13](#_Toc98139533)

[Full Project Resources 13](#_Toc98139534)

[Image Classification Algorithm Comparison (IST 718 Big Data Analytics) 13](#_Toc98139535)

[Summary 14](#_Toc98139536)

[Skills Demonstrated 14](#_Toc98139537)

[Full Project Resources 14](#_Toc98139538)

[Kaggle Data Sets Viz-A-Thon (IST 719 Information Visualization) 15](#_Toc98139539)

[Summary 15](#_Toc98139540)

[Skills Demonstrated 15](#_Toc98139541)

[Full Project Resources 16](#_Toc98139542)

# Scope

The purpose of this document is to demonstrate the skills I have acquired throughout my time in the Applied Master’s in Data Science program at Syracuse University. The Fundamentals section provides an overview of the skills I have learned in my time here at Syracuse with insights into their importance. The Project Samples section provides an overview of 4 projects executed during my masters to demonstrate these skills.

# Fundamentals

The Syracuse University Applied Master’s in Data Science program has provided me with the knowledge and experience to turn data into meaningful information. I have gained a strong foundation of the technical skills required to perform data collection, cleaning, analysis, model generation, and interpretation of the output while also expanding my ability to deliver effective visualization and communication of the results. The core learning objectives I achieved during this program are the abilities to:

* Describe a broad overview of the major practice areas of data science.
* Collect and organize data.
* Identify patterns in data via visualization, statistical analysis, and data mining.
* Develop alternative strategies based on the data.
* Develop a plan of action to implement the business decisions derived from the analyses.
* Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization.
* Synthesize the ethical dimensions of data science practice (e.g., privacy).[[1]](#footnote-1)

## Describe a broad overview of the major practice areas of data science

The variety of courses taken throughout the Applied Data Science Master’s program provided a wealth of insight into the practice areas within Data Science. While there are many areas of focus within data science and many ways to group those specialties, the 3 major categories illustrated within the program studied here at Syracuse are data and business analysis, language analysis, and data pipelines and platforms.

**Data and Business Analysis** explores the business applications for data science. The foundation behind this aspect of data science is to understand the approaches that can be used to apply analysis of historical data as well as predictive modeling via algorithms to provide business solutions. This was an overarching theme of study throughout most courses but most notably in Business Analytics (SCM 651), Data Analysis and Decision Making (MDC 638), Introduction to Data Science (IST 687), Scripting for Data Analysis (IST 652), Data Analytics (IST 707), Quantitative Reasoning in Data Science (IST 722), Big Data Analytics (IST 718), and Marketing Analytics (MAR 653). **Language Analytics** provides the ability to parse and identify patterns in written language that can be used to provide insights into sentiment, trends, and meaningful correlations. Although there is no predictive modeling in language analytics these insights are used to provide organizations the ability to make informed decisions. This aspect of data science was most explored in Natural Language Processing (IST 664) and Applied Machine Learning (IST 707). Finally, the area of **Data Pipelines and Platforms** focuses on the ability to organize, store, and maintain the data necessary to perform data analysis and predictive modeling. These foundations were reinforced through the courses Data Administration Concepts & Database Management (IST 769) and Advanced Big Data Management (IST 769).

## Collect and organize data

There are 5 fundamental steps to collect and organize data for use.

1. **Review** – examine a data set before ingestion to evaluate initial usability and any variations that may impact ingestion of data
2. **Read** – load the raw data set for further analysis
3. **Prep** – provide meaningful labels for data columns and eliminate or correct rows of data with erroneous values preventing meaningful analysis
4. **Analyze** – leverage analysis functions and visualizations to gain insights into areas of data to be cleaned. The goals are to:
   1. Evaluate the data types and ranges of values within the data

This understanding of the intended data types is used to identify how the data will be classified and cleaned to ensure consistency.

* 1. Identify missing or NULL values

Missing or NULL values are assessed for impacts on the data and approaches for handling are considered.

* 1. Identify outliers within the data

Statistical summarization and visualization of the data is used to identify any outliers to be assessed for impact and potential cleaning.

1. **Clean** – leverage and document[[2]](#footnote-2) practices to correct misspellings in the data, assign data types, apply the chosen approach for missing or NULL values, and categorize, organize, and tokenize data as needed. A quick re-analysis of the data is often performed to ensure additional cleaning is not needed.

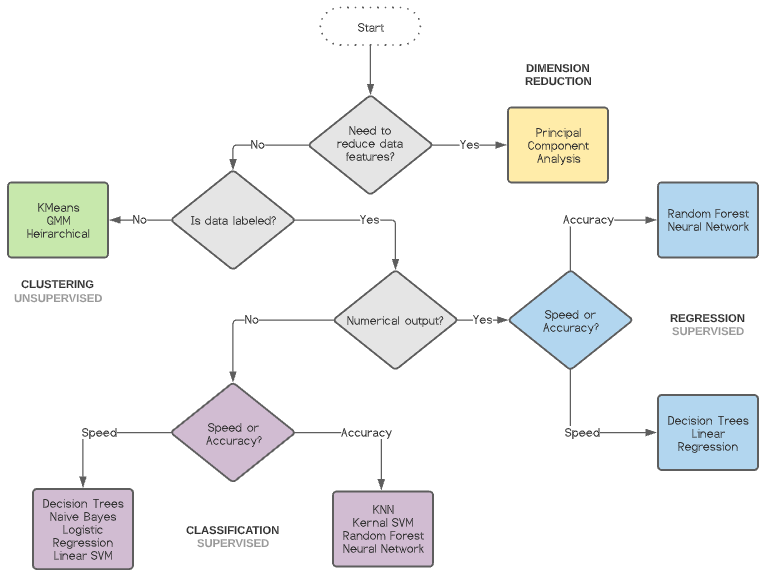
## Identify patterns in data via visualization, statistical analysis, and data mining

Though data analysis begins with questions to be answered, it is necessary to approach the process with an open mind to recognize any patterns in the data that can provide meaningful insights. These patterns within the data begin to tell a story which lead to new questions, which in turn provides a more robust analysis.

The process begins with reviewing the raw data, typically in tabular format wherever possible, and then applying statistical analysis to assess the number of observations, mean, median, mode, range, and standard deviation as applicable. The continuous data is visualized in histograms to further evaluate the shape and distribution of a measure. Correlation matrices are run against this data to identify relationships. Data points may be mapped by scatterplot to identify whether any potential linear relationships exist before additional modeling is performed. Thecategorical counts and distributions of discrete data are assessed and visualized through histograms for additional insight. Depending on the end-goal of the assessment and the values being analyzed, the categories within discrete data may be assigned integer values to allow for correlation analysis. This part of the process is searching for statistical indications of the most meaningful variables within the data for further evaluation.

Once this basic analysis of the data is complete, algorithms and modeling techniques are identified to further assess the data. The determining factors behind which algorithms are used are loosely broken out as follows. [[3]](#footnote-3)

* Is there a need to reduce the number of features in the data set?
* Is the desired outcome of the business question numerical or categorical?
* Is the data labeled?
* Is speed or accuracy paramount for algorithmic processing?



**Figure 1**

## Develop alternative strategies based on the data

Strategies are typically first developed while performing the initial inspection of the data. As algorithms are executed and tuned it often leads to additional examination or new approaches. Often results are not meaningful after the first trial of one or more algorithms, and it becomes necessary to segment the data further or perform additional experiments to analyze the data for significant results. As new information is gleaned from the data, new analysis approaches may become useful to answer the original question of the analysis. For example, when forecasting data for sales it may become prudent to add to the investigation by clustering the data to provide segmentation analysis. The quality and format of the data available can impact the ability to pursue different strategies.

## Develop a plan of action to implement the business decisions derived from the analyses

The goal of any analysis is to gain actionable insights to support strategic business decisions. Results of complex algorithms are interpreted with consideration of the primary business questions, as well as how individual factors have influenced outcomes in the past and/or may influence outcomes in the future. Plans of action are developed and recommended from these outcomes to provide value to organizations. For example, in the Women’s Clothing in E-Commerce project (IST 707) it was revealed that reviewers who submitted a 3-star rating for a purchase were historically almost equally split between recommending purchase of the product (58%) versus not recommending it (42%). This indicated reviewers in this category were the most ambivalent about recommending the product compared to other categories and therefore were the most likely to be swayed by additional interventions from incentivization programs (see IST 707 Project). The plan of action determined for this project included customer service improvements and, most importantly, contact with 3-star review customers to identify potential improvements. Such recommendations are provided with most analysis as illustrated in the projects used within this portfolio.

## Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization

It’s one thing to be able to analyze data for meaningful patterns, but without the ability to communicate those results effectively to stakeholders those results can be effectively meaningless. As a data scientist it is imperative to gather domain knowledge and successfully communicate results with stakeholders having varying levels of business acumen, statistical knowledge, and technical expertise. Such communication skills are demonstrated within each project included within this portfolio.

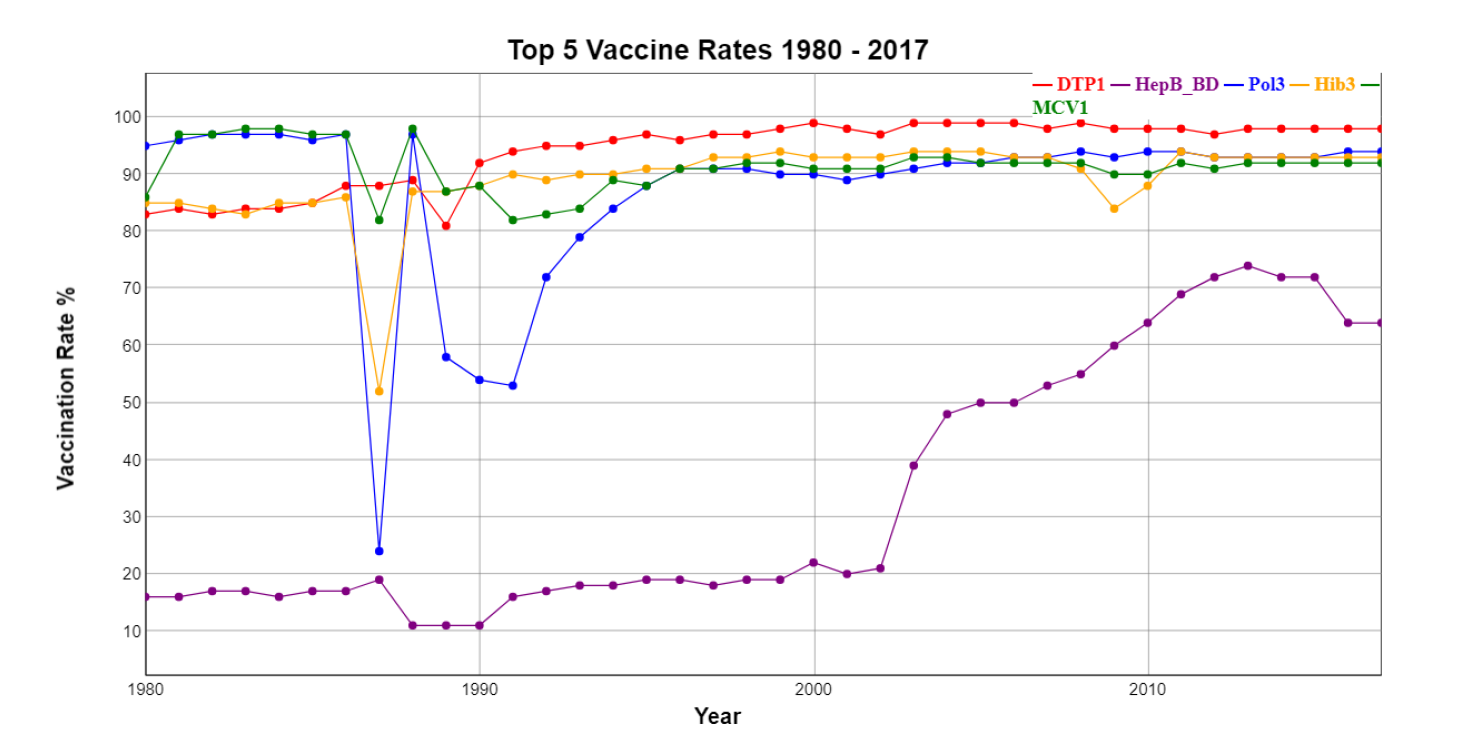
## Synthesize the ethical dimensions of data science practice (e.g., privacy)

It is imperative data is collected, protected, analyzed, and presented in an ethical manner to ensure the integrity of each effort and its results. When collecting data, it is important to do so in a way that is as neutral and unbiased as possible while also protecting personal information. Respecting data privacy is paramount. It is also essential to document all data cleaning practices to accurately present any changes to the original data that may result in unintentional bias. Likewise, practices used in tuning the models must be documented for full transparency. Most importantly it is our responsibility to ensure the stories told by our results are represented accurately without bias, including negative findings and providing confidence levels to indicate room for error.

# Project Samples

## US Vaccination Rates (IST 772 Quantitative Reasoning)

US Vaccination Rates



### Summary

This final exam project analyzed 3 data sets related to vaccination rates in the United States:

* **US Vaccines** – provided vaccination rates for diphtheria/pertussis/tetanus (DPT1), hepatitis B (HepB\_BD), polio (Pol3), influenza (Hib3), and measles (MCV1).
* **California Kindergarten Vaccination Reporting Data** – a list of California kindergartens and whether they reported vaccination data to the state in 2013
* **Districts** – a sample of California school districts from the 2013 data collection along with specific numbers and percentages for each district

The goal was to perform quantitative analysis to identify correlations and predictors of vaccination rates and factors that impact vaccination rates in the United States.

### Analysis & Models

The following analysis and modeling were performed:

* Changepoint analysis of vaccination rates
* Change in variability analysis for average US vaccination rates
* Mean Change Point analysis of US vaccination rates
* Comparison of public and private school vaccination reporting
* Frequentist and Bayesian Analysis of Correlation between vaccines
* Generalized Linear Modeling and Bayesian Estimation of Logistic Regression to predict efficacy of district reporting

### Full Project Resources

The final paper with code is attached to the end of this document. Additional code and other documentation can be found here:

<https://github.com/anbradsh/ABrennan_Portfolio/tree/main/IST%20772%20Quantitative%20Reasoning>

## Women’s Clothing in E-Commerce (IST 707 Data Analytics)

Women's E-Commerce Clothing Reviews

### Summary

The Women’s E-Commerce Clothing Reviews project examined anonymized online women’s clothing customer review data from a sample of retailers across the industry. The goal was to utilize various data analytics techniques to understand how an online retailer could increase sales and market share via the following methods:

1. Text analysis of positive and negative reviews

2. Age range correlations

3. Customer perception of other customer reviews

4. Rating correlations

### Analysis & Models

The following analysis and modeling were performed:

* Sentiment analysis of reviews by rating
* Association Rule Mining of rating associations, positive feedback counts, age groups, product types, and recommendations
* K-Means Clustering by rating
* Support Vector Machine analysis for recommendations
* Naive Bayes Analysis for recommendations
* Random Forest Analysis for recommendations
* Decision Tree Analysis of influences to ratings, positive feedback, and age group purchases

### Full Project Resources

The final paper, project code, and other documentation can be found here:

<https://github.com/anbradsh/ABrennan_Portfolio/tree/main/IST%20707%20Data%20Analytics>

## Image Classification Algorithm Comparison (IST 718 Big Data Analytics)



### Summary

The Image Classification project examined the accuracy and performance of various classification approaches for the Fashion MNIST image data set. The core of the effort was to demonstrate the pros and cons of the selected approaches to understand the nuances of selecting the best algorithm for a job.

### Skills Demonstrated

The following models were used in the Image Classification Algorithm Comparison effort:

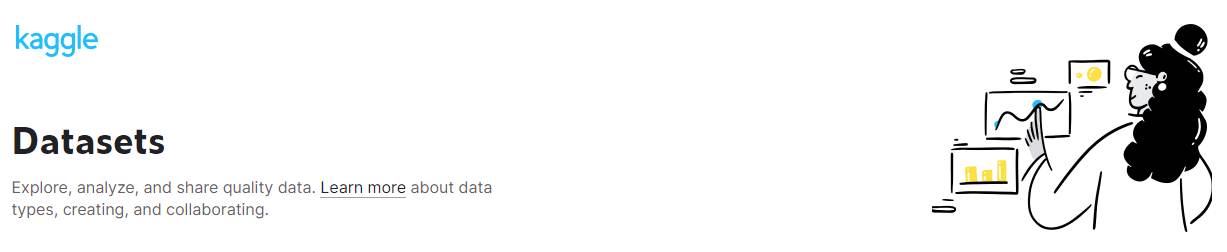
* Neural Networks
* Decision Trees
* Random Forest
* KNN
* Naïve Bayes

### Full Project Resources

The code and output of this effort can be accessed here:

<https://github.com/anbradsh/ABrennan_Portfolio/tree/main/IST%20718%20Big%20Data%20Analytics>

## Kaggle Data Sets Viz-A-Thon (IST 719 Information Visualization)



### Summary

The goal of the Viz-A-Thon effort was to demonstrate the ability to effectively visualize data analysis targeted for various levels of distance and engagement. For this effort Analysis was performed on a data set comprised of meta-data and usage metrics for data sets available on Kaggle.com in June 2016. The primary question to answer was ‘What are the characteristics of data sets most leveraged by the users?’

### Skills Demonstrated

The following visualization techniques were performed:

* Utilization of sizes to target visualizations for 3 distances – across the room, within a few feet, and up close
* Utilization of columnar layout and intentional violation of structure to draw the idea to most important visualizations
* Leveraging of white space to allow the eye to rest
* Using colors to guide the eye and infer relationships
* Meaningful usage of a Sankey graph, word cloud, bar chart, correlation matrix, scatterplots, and density graphs

### Full Project Resources

The final poster, code, and other documentation can be found here:

<https://github.com/anbradsh/ABrennan_Portfolio/tree/main/IST%20719%20Information%20Visualization>

1. <https://ischool.syr.edu/academics/applied-data-science-masters-degree/> [↑](#footnote-ref-1)
2. It is essential to document cleaning practices to provide full transparency into any manipulation of data. This is necessary for the practice of ethical data science. [↑](#footnote-ref-2)
3. <https://www.kdnuggets.com/2020/05/guide-choose-right-machine-learning-algorithm.html> [↑](#footnote-ref-3)