




M.S., Applied Data Science

# PROGRAM IN REVIEW

Michelle Mak

June 2019  
Syracuse University



## Overview

As a complete newcomer to the world of data science, my goal upon entering this program was simply to be able to use raw data to find insights that can help make better, more educated business decisions. However, with the completion of each course, both the incredible capabilities of data analysis and the inherent complexities that come with the gathering and interpreting of data became more and more apparent.

The Applied Data Science program at Syracuse gave me a good foundation of how to model, visualize, and interpret data that will be useful in any number of fields. This [portfolio](#) will demonstrate the achievements of the overarching learning objectives laid out by the program. The learning objectives are listed below, along with how this program has helped me achieve these goals.

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

Data science methods can be applied to almost any industry, as long as relevant data exists. The over-arching goal of this field is to turn data into information. Data scientists collect, clean, and organize large amounts of data in order to discover insights that may not have otherwise been found if the data remained in its original form. Using data models to find significant variables and data visualization methods, Data Scientists should be able to take their findings and re-interpret them for stakeholders. The “Key Works and Projects” in the section below will demonstrate this broad overview of major practice areas.

### *2. Collect and organize data.*

Collecting and organizing data for data science projects consists first of identifying relevant data to make insights about the hypothesis or observation. Getting the right context to frame the question is the most important part of data gathering because data can be manipulated to say almost anything. Therefore, it is a key responsibility of a data scientist to try and achieve the most honest answer to a data question.

Additionally, a common message in each class, whether it was a course about business or analytic methods, was that data is everywhere, but data is dirty. Thus, proper data cleaning, processing, and organizing prior to analysis usually takes the most time and effort of any project to ensure that the data being analyzed is prepared correctly. For example, how to take care of missing data, outliers, etc.

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

There are infinite varieties of ways in which one can visualize the same data. However, some visualization and modeling methods are better suited to demonstrate specific insights. It is, therefore, a data scientist's job to identify the appropriate and most efficient method to communicate insights to stakeholders. For example, a time series plot is best for showing change over time and, perhaps, forecasting. A statistical regression is best for comparing the significance of variables and trends. Box and whisker plots are useful for displaying data significant markers in the data range as well as identifying outliers.

### *4. Develop alternative strategies based on the data.*

Sometimes, where the data is hypothesized to go and where it actually goes are two very different things. As a data scientist, it is important to be able to identify this change in observation and investigate accordingly to find the best answer. Having a breadth of knowledge about the many data methods and strategies out there will allow for this necessary pivot. The Spotify project in the “Key Works and Projects” section demonstrates

this by using multiple data methods to answer the same question – however, each method produced very different results.

*5. Develop a plan of action to implement the business decisions derived from the analyses.*

After analyzing the data and finding insights, it is key to be able to develop a plan of action. This can be in the form of a marketing strategy or an implementation plan, using the data to steer business actions. The most important part of developing a plan of action is being able to justify decisions with data analysis. Secondly, it would also be wise to verify this through A/B testing.

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

As not all team members and departments are trained in data methods, the ability to communicate results so that the information is digestible and easily understandable is pivotal. While other data scientists might be interested in the methods that were used, other stakeholders just want to know why. Understanding the scope of each audience's knowledge will be helpful when choosing what types of information is relevant enough to share and which should be omitted.

Throughout this program, each project and most labs required a written paper that meticulously documented the pre-processing and analysis methods that were used. These papers are more useful for sharing with fellow professionals, programmers, and statisticians who are wondering how the results were achieved. A powerpoint presentation also accompanied most projects, which just communicated the significant results to less involved parties.

*7. Synthesize the ethical dimensions of data science practice.*

Since big data is a relatively young industry, the policies and rules for it are just being discovered. One key point of everyone's concern is the matter of privacy. Data is inevitably tied to an individual person's life, so the security of this information has led to much controversy. The principles of ethical data science first and foremost value the privacy of the customer.

However, since the world of information security is also just emerging, a lot of data has been and still is mishandled. Thus, important policies must be in place to protect users, keep companies accountable, and ensure that data is handled ethically.

## Key Works & Projects

The work listed below was completed throughout the course of this program and showcase the achievement of each of the learning objectives.

### Arrival Delay Analysis | IST 687

**Summary:** This group project analyzes all recorded airline delays in the year of 2018. The investigation is conducted entirely from the perspective of the carrier American Airlines with the goal of identifying specific locations that exhibit patterns of heavy delays. The analysis would find solutions for based on the locations identified and also compare the American Airlines' delay patterns with competition airlines to form a baseline for improvement.

**Learning Applications:** This was the first true data analysis project in which the data had to be obtained (via Kaggle), scrubbed, and preprocessed for various types of modeling using R markdown. In addition, the team developed a “plan of action” using a Kanban board on Trello to stay organized and manage roles.

In order to reach our objectives, we use various exploratory data analysis techniques to analyze the trends in the data. This involved a large amount of ggplot2 and visualizations ranging from line charts, histograms, bar charts, and maps. Additionally, we use linear regression, SVM models, and logistic regression to predict the length of a delay and to understand the causes.

The results, which include flight path insights and suggested solutions, are both documented in a research paper and presented in class.

**Learning Objectives:** 1, 2, 3, and 6.

### Spotify: Global Music Taste Analysis | IST 565

**Summary:** This project seeks to discover whether there is a connection between culture and music, more specifically rhythm and tune. Using the Spotify API to gather data, my partner and I created a data set containing the Top 50 songs (as of September 12, 2018) from ten select countries around the world. We selected the countries of the United States, Germany, Hong Kong, Norway, New Zealand, Singapore, Australia, Iceland, United Kingdom, Portugal, and Mexico to encompass a diverse taste in music and span the globe. These countries were also selected based on the availability of their “Top 50 Songs” playlists on Spotify.

This investigation mainly aims to discover whether songs can be classified by country based on its audio features, using methods such as clustering, K-Nearest Neighbor, support vector machines, naïve bayes, decision tree, and random forest.

**Learning Applications:** The data collection for this project goes beyond just finding a prepared dataset online since it required the use of an API. The investigation is conducted entirely on R markdown. The dataset has to be organized and reorganized a number of ways in order to answer exploratory questions and to apply all of the classification models, which have different preprocessing requirements.

We explore the data using a number of techniques, including summary statistics of song traits, text mining song titles, and other forms of visualization. The use of multiple classification methods help to compare and verify results, as well as demonstrate how some methods are better than others to answer specific questions.

**Learning Objectives:** 1, 2, 3, and 6.

### Off-Season of Gifting: Market Analysis and Business Strategy | Mar 653

**Summary:** This investigation leverages sales and customer data from an unnamed multi-channel company to develop a marketing strategy to take advantage of business opportunities during off-season months in the Spring.

We use tools in XLSTAT, calculate customer life time value, and perform a clustering analysis to separate customers based on the seasonality of their spending habits. These methods break up the customers into Christmas shoppers and Non-Christmas shoppers, which should reveal useful profiling traits such as channel choice, potential touch points, and gifting patterns.

Using the data derived from the cluster analysis, customer profiles will be created to offer the company a clear and succinct overview of the right segments to target. We present the findings of this investigation to build and present a potential marketing strategy for the company.

**Learning Applications:** The identity of the company used in the investigation remains private, which keeps in line with big data ethics codes. All customer information is also withheld. But even without these details, we are still able to achieve some great insights with the available data.

Mainly, the cluster analysis reveals two main clusters that can be used as the basis to create segments of Christmas versus non-Christmas shoppers. The data also provides the variables needed to calculate the average customer life time value. And finally, a logistic regression sheds light on future channels in which it might be worth investing to further develop the customer base. These insights are summarized in a marketing strategy/business plan and presented to the class and professor.

**Learning Objectives:** 1, 3, 4, 5, 6, and 7

### Data Science Job Prospects | IST 652

**Summary:** This is an investigation into the job market for Data Scientists in anticipation of my impending graduation. Using data scrubbed from Indeed.com and Zillow Research, I pinpoint optimal locations for potential jobs as well as define the most valuable skills to learn or highlight on my résumé.

This investigation applies data visualization and text mining to answer questions and help make decisions regarding what classes to take next and where to potentially move.

**Learning Applications:** The data gathered for this project has to be merged and transformed in order to answer key questions posed in the exploratory analysis. Visualization techniques included using groupby functions and word clouds. Although the findings in this study are limited, a number of alternate strategies are proposed for future investigation.

**Learning Objectives:** 1, 2, 3, 4, 5, and 6

### Coaches' Salaries | IST 718

**Summary:** Using coaches' salary data provided by the professor, this case study recommends a base salary for the next head football coach at Syracuse University. The investigation uses supplemental data, like graduation rate of football players, stadium size of each school, in order to make an educated decision.

**Learning Applications:** This lab requires students to build a data frame, fit a regression model and find relevant predictors, and think creatively about how to organize and re-organize the data. The lab also introduces hypothetical scenarios to see if changing the setting would also alter the outcome, which prompts more creative data manipulation to answer the question.

**Learning Objectives:** 1, 2, 3, 4, 5, and 6

### Zip Code Investment | IST 718

**Summary:** This case study investigates how we can predict three zip codes that provide the best investment opportunity for the Syracuse Real Estate Investment Trust (SREIT). Using historical data of housing prices, it will calculate the compound average growth rate of every zip code from 1997 - 2018 in order to identify the zip codes with the best CAGR. After the data has been down sampled, a time series regression is performed using FBProphet to predict future house price trends.

**Learning Applications:** This lab combines data modeling with some financial analytics. The data from Zillow is used to create a brand-new variable (CAGR) that is used to as a technique to down sample the large dataset. After performing the initial analysis and running the time series through FBProphet, the RMSE is also calculated and used as a variable to make the decision. This lab is unique in that it requires the student to use a new package and relies on Google Colab notebooks as a means of completing the modeling more efficiently, thus forcing students to branch out and look at the latest and greatest software.

**Learning Objectives:** 1, 2, 3, 4, 5, and 6

### Law & Weather | IST 718

**Summary:** This is an investigation into the possibility of using weather data to predict crime incidents in two different cities. Crime and weather data are gathered for the city of San Diego, which is used as a "control city" since the temperatures in San Diego are relatively stable. Likewise, the data in New York is analyzed as the variant in order to truly see how weather may correlate with the number of crime incidents throughout the year.

**Learning Applications:** This was a very interesting study since it does A/B test some variables in a way. Weather data is gathered through the National Oceanic and Atmospheric Administration's REST API, and crime data from government agencies. The data requires

quite a bit of cleaning and manipulation in order to create a merged dataset and further wrangling in order to fit the rigid requirements of FBProphet, the method for time series modeling. Data visualizations such as bar graphs and heat maps are used initially to establish correlation between weather and crime.

**Learning Objectives:** 1, 2, 3, 4, 5, and 6

Github: [https://github.com/michmak/syr\\_portfolio](https://github.com/michmak/syr_portfolio)