**Portfolio Milestone**

Cartney Thompson

SUID: 262769286

Syracuse University

M.S. Applied Data Science

Spring 2019

<https://github.com/cartney06/Applied-Data-Science-Syracuse---Portfolio>

***Introduction***

Data science extracts insights from data by combining three areas of expertise: computer science, statistics, and field domain knowledge (<https://en.wikipedia.org/wiki/Data_science>). Prior to my entry into the Syracuse Applied Data Science (ADS) program, I felt there were areas that I was strong in; I was currently a practicing Data Scientist so I had domain expertise and I’ve been coding for several years. However, I felt my knowledge of statistics while not weak, certainly was not my strongest area. Within the ADS program at Syracuse, there are several goals the program looks to achieve to help a student become a strong data scientist. These goals are:

* 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.

In this paper, I will discuss four projects that highlighted my best work in the in connection with the goals from the ADS program.

***Project Background***

In my project “Noise Complaints in New York City” for IST 652 Scripting for Data Analysis, I analyzed where noise complaints in NYC came from and when were they likely to occur. This was a traditional exploratory data analysis project focusing on analyzing noise complaints in NYC. Analysis was done utilizing pandas within python and seaborn/matplotlib for visualizations. This project allowed me to explore different ways to analyze noise complaint data using pandas and visualization libraries that I had not used that often in the past.

The project “Predicting NYC Restaurant Sanitation Grades” for IST 565 Data Mining, I utilized R to determine which attributes about a restaurant can we use to predict the letter grade from sanitation inspections. In addition, I evaluated which model using three different algorithms; decision trees, Naïve Bayes, and support vector machines was best by measure of prediction accuracy to predict a restaurant’s inspection grade. Insights gained from the project included the pros and cons of each algorithm (Naïve Bayes was fastest; decision tree performance was slowest), and identifying the most important attributes for building the model via each attribute’s information gain (Food Protection, Vermin/Garbage, and the Food's Temperature were the strongest predictors in predicting the health inspection grade).

In IST 718 Advanced Information Analytics, my project, “Recommending Beer”, we utilized alternating least squares to build a model to recommend beers based upon a user’s profile of previously rated beers. This introduced me to building an algorithm using Spark and how much faster it is in performance vs. Python, especially when building recommendations using collaborative filtering.

Finally, for the project “Twitter and the Kaepernick Ad Campaign’s Impact on Nike” that I completed for IST 736 Text Mining, I analyzed Nike brand sentiment on twitter before the Kaepernick Nike campaign and after. Tools included crowd sourcing sentiment via Amazon Mechanical Turk, using the natural language toolkit package within python to for vocabulary vectorization, analyzing change in sentiment of a sample of tweets pre vs. post Kaepernick campaign, and key tweet topics during each timeframe utilizing topic modeling.

***Collecting and Organizing Data***

Prior to entering the ADS program, I was fairly knowledgeable in obtaining and organizing structured data. I knew how to obtain a csv file and upload it into a tool for analysis. Semi-structured and unstructured data I was less confident in my abilities. Two projects that I worked on in the program greatly enhanced my ability to obtain and organize data from an unstructured source; analyzing Nike tweets the week before and after launch of Nike’s Colin Kaepernick campaign from 2018 and analyzing user provided beer ratings for a recommendation system. Both projects provided challenges in how I obtained and organized from an unstructured data source.

The goal of the Nike project was to determine Nike brand sentiment on Twitter pre vs. post launch of Kaepernick campaign and key tweet topics during each time period. The project provided a new challenge for me of collecting, organizing, and analyzing unstructured data from Twitter. Tweets often contain symbols, urls, emoji’s, etc that can create errors and throw off formatting when uploading for analysis. In order to perform analysis on the tweets, I needed to explore alternative methods for obtaining tweets.

During the summer of 2018, I took a scripting for analysis course as part of the ADS program. In this course, I learned how to parse information from tweets and organize them into a format useful for analyzing. One method was to scrape tweets using a pre-developed package available in python. This package, called twitterscraper, allows users to scrape tweets based upon some values provided by the user. These query values included the subject and tweet dates. For this particular project, I only wanted tweets that included “Nike” or its slogan “justdoit”, found common in many tweets about Nike. As far as dates, I only pulled a week’s worth of tweets for pre-Kaepernick campaign launch and post=Kaepernick campaign launch. Each tweet along with the date of the tweet were returned in a tab delimited file which allowed for easy ingestion into a data frame for analysis in python.

Now that I had collected the necessary tweets, I needed to organize them to be able to analyze sentiment and develop topic models for pre and post launch of Nike Kaepernick campaign. Part of this process included stripping out the symbols, urls, and picture links to have data in a structured format. Working on the Nike project allowed me to collect and organize a different type of data set I have not been used to analyzing.

For the beer recommendation project, beer ratings came in an unstructured text file format that needed to be parsed before analysis. Data for the recommendation project came from Beer Advocate – a beer rating website that rates beers, beer stores, and bars (<https://en.wikipedia.org/wiki/Beer_rating#BeerAdvocate>). The data consisted of user submitted beer reviews spanning several years in plain text format. In order to analyze the data, the file needed to be parsed into a data frame. In my typical day-to-day work, data is usually in a clean format where a function to parse columns is not required. However, this project required I utilize what I learned from my scripting for data analysis course; write a function in python to ingest the beer data, strip out trailing whitespace, and parse columns by a colon.

While collecting and organizing unstructured data from social media was new for me, collecting and organizing structured data was not. This had been a daily practice of mine as a Data Scientist. However, I was accustomed to data in an environment controlled by my employer; highly structured and neatly organized by data engineers responsible for formatting the data. The Nike Kaepernick campaign and the beer recommender project allowed me to obtain data from unfamiliar sources and in unfamiliar formats to organize for analysis.

***Data Analysis***  
 Having spent more than a decade in analytical roles, I most certainly felt my level of knowledge was strong in this area prior to starting the ADS program. However, working on projects such as analyzing noise complaints, predicting restaurant sanitation grades in New York City, and analyzing Nike brand sentiment on Twitter allowed me to test those analytical abilities in a different way.

To no one’s surprise, New York City is noisy. Between the honking horns of yellow cabs along the avenues, to the loud music blaring from bars and clubs, we are somewhat immune to the noise as New Yorkers. However, there are times when the noise is too loud and is a distraction from whatever task we may be performing. When this happens, New Yorkers file noise complaints to the local authorities. With the “Noise Complaints in New York City” project, I wanted to explore and answer key questions with regards to noise complaints: Do noise complaints occur most often at certain times? Are noise complaints more likely to come from specific areas of the city? What are the characteristics of where the most (top 10) noise complaints occur?

In order to answer these questions, I limited myself strictly to utilizing pandas in python for analysis. As mention in the project’s background, this was a project for the Scripting for Data Analysis course, so python was the language of choice. Up until this point, all classes as part of the ADS program centered around analysis performed in R. Utilizing python to perform the analysis was a nice change of pace for me, as most analysis at my place of work was also performed in R.

The most difficult task of the project was exploring the data. There are many ways to analyze the noise complaints including when the occur, where they occur, the type of noise complaint, etc. This project allowed me utilize data visualizations to tell a cohesive story and answer my research questions. Data was analyzed using Pandas data frames within python and matplotlib/seaborn to produce visualizations. The analysis resulted in understanding that most noise complaints occurred in the late hours, Thursday – Sunday morning and Manhattan residents were most likely to file noise complaints.

Sticking with the New York city theme, with the Predicting NYC Restaurant Sanitation Grades project, I wanted to answer the question of can we predict the sanitation letter grades based upon location of the restaurant, the cuisine it offers, and violation codes it incurs during inspection? This particular project allowed me to collect and organize data and identify patterns in the data via visualizations and data mining.

After obtaining and organizing the data, the data set was now down to three features and predictor variable for analysis; Borough location of the restaurant, violation code, restaurant cuisine, and the letter grade (the predictor variable). Exploratory analysis included understanding the grade distribution of the data set, grade distribution by borough, and percentage of inspection violations by category. Close to two-thirds of inspections resulted in an A grade, one in four resulted in a B grade, and one in ten resulted in a grade of C, the lowest letter grade a restaurant can receive. The grade distribution did not vary much by borough as evidenced by the marginal differences by borough. Lastly, more than half of all inspection violations were from facility maintenance, meaning general building cleanliness or sanitation issues.

Final step was to see how accurately we can predict restaurant grades utilizing three different algorithms; decision trees, Naïve Bayes, and support vector machines. This project let know first-hand some key differences between algorithms we studied throughout the term. For predicting restaurant grades, decision tress was actually the most accurate (using 3-fold cross validation on our training set) although it took the longest to run (Naïve Bayes was actually the fastest algorithm to run but was least accurate). Accuracy to predict test data from Decision trees was around 83%, 80% for SVM, and 74% for Naïve Bayes. Out of variables used to predict inspection grades, based on information gain food protection, vermin/garbage, and food temperature violations were the strongest predictors in predicting the health inspection grade. This information could be extremely valuable to restaurants to help them focus on areas such as food protection, vermin, and food temperature once inspection time comes.

Lastly, analyzing text data was most certainly a weak point for me coming into the program. Although I am most certainly better now, I will continue to analyze text data to to sharpen my skills. The Twitter and the Kaepernick Ad Campaign’s Impact on Nike was my first real deep dive into text mining and just from completing the project, I was able to understand how to vectorize words and its importance to building a model used for sentiment, the importance of crowd sourcing data to help determine attributes for text that you cannot do on your own due to timing or inherit bias, and how to topic model and its importance in distinguishing key topics pre Kaepernick campaign and post Kaepernick campaign.

***Alternative Strategies from the Data and Plan of Action***

Sometimes what we see in our data provides us with alternative methods for analyzing it or building models that differs from our original plan. While completing the Nike Kaepernick campaign analysis, I originally was going to cluster tweets based on upon frequencies of words appearing in tweets to distinguish between key topics. I noticed that many tweets were grouped together despite having different sentiment scores and meanings. It was difficult to distinguish topics associated with each cluster because often times many of the same top words were in each cluster.

Topic modeling provided better modeling of the key topics found in the collection of tweets about the Kaepernick campaign. There were clear distinctions between tweets that contained the words “hate”, “burn”, or “superhero” for example. Tweets that included the word hate often times were negative tweets about the Nike Kaepernick campaign, tweets that included the word burn often time were about burning Nike apparel, a negative action, and superhero tend to include tweets about the Serena Williams campaign that ran the week before the Kaepernick campaign.

***Communication Skills***

For each project completed at the end of course, I typically had to deliver three items; the code used to generate the analysis, a written report that highlights methods and findings, and a presentation given to classmates and the professor summarizing key points about the project.

Providing the code allows for not only feedback on how the analysis was done, but provides others a clear, written document to traceback how I came to the conclusions I did. For all four of my projects, my code is commented throughout to explain why I wrote a specific block of code. Certainly, from the first project I completed in this portfolio, Noise Complaints in NYC, to my last, Nike Kaepernick Campaign, I have greatly improved in how I comment on my code and where I comment.

In my career, the most important aspect in communicating report findings is the actual report others will see. Often times the written report is what goes on the company’s wiki page or becomes published. Being able to communicate analysis findings to others is a strength of mine and all four projects highlighted this. One project report that I especially proud of is Predicting Restaurant Grades in NYC. Here, I utilized real world examples to communicate the different machine learning concepts. For example, when we think of a decision tree, think of it as making decisions to achieve the most reasonable outcome. For restaurant grading, a store that had a food temperature violation, food protection violation, and any other food related violation, was destined for a C grade. When I mentioned that an algorithm is accurate, it is important to tell a non-technical audience that definition of accuracy. The percentage of times the algorithm makes a correct prediction for that restaurant’s sanitation grade on a set of data.

***Ethical Dimensions of Data Practice***

As a data scientist at a global organization, privacy was a major concern for us. More specifically, General Data Protection Regulation (GDPR) compliance was top of mind for the organization, especially for anyone that handled sensitive user information like myself. During the program, I consciously tried to steer clear of any sort of analysis that included sensitive data. This included medical data, personally identifiable information, or any other sensitive information like financials. However, when obtaining data from the Nike Kaepernick Campaign project, part of the tweet scrape did include a user’s twitter name and their id.

When scraping tweets from twitter, it did not make sense to include such information as a user name or id as that could be potentially personally identifiable information. Due to my experience from my occupation, I certainly felt making decisions on what could potentially be sensitive information for this project was a strong area for me. I understood that a user’s name on twitter or their twitter id was of no use to me for analysis so it did not make sense to pull that information.