Nash Delcamp

[Applied Data Science Portfolio Milestone](https://github.com/ndelcamp/Applied-Data-Science-Portfolio/tree/master/IST%20736%20-%20Presidential%20Speeches)

7/10/19

As a student in the Applied Data Science Graduate Program at Syracuse University, I have refined several techniques relating to data science. These include the collection and organization of data, the identification of patterns through visualization, analysis, and data mining, development of action plans based on data, and the communication of the results of analyses. These techniques can be summarized as the abilities to effectively ask questions about data, perform analyses based on those questions, and communicate the results to an audience.

Visualizations were used in nearly every assignment to better understand the data and find patterns for initial analyses. These visualizations served as a great starting point to begin analysis. Several methods of analysis have been learned in detail. These methods include descriptive, prescriptive, and predictive models to better understand the data and make recommendations based on it.

Based on the patterns noted and the results of analyses, different types of action plans were developed. These included recommendations to businesses, governments, and the self. The recommendation requires an advanced level of communication for the party of interest to fully understand what actions to make based on the analyses conducted. These recommendations were made in both written form and in the form of vocal presentations.

One project that demonstrated these skills was in Text Mining: [IST 736 – Presidential Speeches](https://github.com/ndelcamp/Applied-Data-Science-Portfolio/tree/master/IST%20736%20-%20Presidential%20Speeches). One skill developed and demonstrated to more efficiently and reliably collect and organize data was web scraping. In this project, information and speeches of each president from The Miller Center Presidential Speeches were scraped to build both a dataset of demographic data of presidents and a corpus of speeches for text mining analysis. The text of each speech was transformed to vectorized word counts using TFIDF vectorization which enabled its use in modeling. Further exploratory work was done to better understand how presidential sentiment has increased over time and K-means clustering – with interpretation – was used to cluster the presidents based on their speeches. Additionally, metadata such as political party, term length, and time period were modeled using these speeches. Machine learning methods including Multinomial Naïve Bayes, Support Vector Machines, Neural Networks, and Random Forests were used to model each of these metadata variables. These models were all tuned and compared to one another to select the best model for each use case. It was demonstrated in this project that the vocabulary used by presidents is not strongly related to their political party or term length but is characteristic of the time period in which the speech was given.

Another project that serves as proof of these abilities was in Big Data Analytics: [IST 718 – Suicide Rates](https://github.com/ndelcamp/Applied-Data-Science-Portfolio/tree/master/IST%20718%20-%20Suicide%20Rates). Here, with two team members, data was pulled and merged from multiple sources. Patterns between variables were visualized with pair plots and correlation plots and well as time series plots of the suicide rates of many different demographics including age, gender, and location. Again K-means clustering was used to gain greater insight into some of the patterns in the data. These clusters were also plotted in many unique ways, including on the world map and on a ternary plot. Suicide rates were modeled using Multiple Linear Regression and a Random Forest model. The recommendation made to all countries was to immediately implement suicide prevention strategies, focusing on elderly males and to further research the relationship between gun ownership rates, internet usage, and suicide rates.

A third project that demonstrates proficiency in these topics was in Marketing Analytics: [MAR 653 – Student Alcohol Consumption](https://github.com/ndelcamp/Applied-Data-Science-Portfolio/tree/master/MAR%20653%20-%20Student%20Alcohol%20Consumption). In this project, relationships between student drinking habits were explored and modeled with a goal of preventing underage drinking and promoting healthy habits. Through descriptive statistics grouped by different variables, different factors were discovered to better influence the modeling approach. K-means clusters were created which illustrated four distinct groups of the combinations of poor/good grades paired with low/high alcohol consumption. These data were modeled with linear and logistic regression models to predict student alcohol consumption. The linear models performed poorly, but both a logistic model and ordinal logistic model fit the data well in predicting whether a student “over-drank.” While parents and school districts would likely prefer students to drink less alcohol, a large portion of students were discovered who achieved good grades while consuming a lot of alcohol.

The ability to obtain, transform, visualize, analyze, and model data as well as effectively communicate the results of these methods has been developed greatly throughout this program.