M.S. IN APPLIED DATA SCIENCE

Abstract

A reflection on achieving the Applied Data Science Program learning goals.

Katherine Hurtado-da Silva

khurtado@syr.edu

SUID: 781774642

Introduction

Completing the Applied Data Science program was crucial in helping me gain an understanding of the different ways machine learning could be leveraged to conduct statistical analyses. The following program's learning objectives were undoubtedly reached throughout this academic journey:

- Collect, store, and access data by identifying and leveraging applicable technologies
- Create actionable insight across a range of contexts (e.g., societal, business, political) using data and the entire data science life cycle
- Apply visualization and predictive models to help generate actionable insight
- Use programming languages such as R and Python to support the generation of actionable insight
- Communicate insights gained via visualization and analytics to a broad range of audiences (including project sponsors and technical team leads
- Apply ethics in the development, use, and evaluation of data and predictive models (e.g., fairness, bias, transparency, privacy)

While gaining a more profound sense of the concepts behind traditional predictive methodologies, my focus on language analytics enhanced my problem-solving thought process. Despite the importance of every course taken in this program, four projects can mainly attest to meeting the overall criteria of the Applied Data Science program.

Influence of Political Affinities on News Reporting via Twitter[1]

Though all classes involved accessing, collecting, and storing data, IST 652, Scripting for Data Analysis, extended this skill set to semi-structured and unstructured data and associated preprocessing tasks. With Python as the primary programming language, the knowledge gained during this course served as the foundation to assess whether politically-biased reporting was evident on Twitter. Fox News, BBC, and CNN were selected for this culminating project because they are considered far-right, centrist, and far-left news outlets, respectively. Their Twitter accounts were used to determine

1 https://github.com/kdasil/IST652_ScriptingforDataAnalysis

whether these perceived political affinities were evident in each account's topic coverage and sentiment patterns. The number of tweet likes was also collected to get a picture of each news outlet's audience.

Ethically Leveraging Applicable Technologies

The most recent 1,500 tweets between April 1, 2022, and June 5, 2022, were collected into separate data frames for analysis. This date range was applied to ensure the reported topics were contemporary and because the Uvalde school shooting, a polarizing event that split right-wing gun rights supporters from democratic gun-ban advocates, was being covered during this time. Limiting the time frame effectively enabled an unbiased comparison of the reporting trends exhibited. Snscrape, a Python library, facilitated the collection, storing, and access of these news outlet tweets without any restrictions. Learning to extract information manually allowed me to get targeted information for my analysis. In addition to being able to get the tweets by the user (news outlet), I could have easily changed the filters to target general coverage of a topic by a relevant hashtag as well. Though knowing how to access data through API keys was useful, scraping was an invaluable skill that has put me in a position to circumvent API-related challenges, such as how much or what type of data I can collect.

Using Python Visualizations to Create and Communicate Actionable Insight

The current political temperament has permeated most, if not all, aspects of life, casting doubt on the credibility of once-respected news outlets. KMeans clustering and sentiment analysis were used to provide insight into the following core questions:

- How did reported topics for each news outlet differ?
- How did the sentiment and subjectivity vary according to each news outlet's tweets?
- What do the "like" counts for tweets indicate about each news outlet's audience?

After exploring topic modeling for each, word clouds were generated to reflect the essential features of their resulting clusters. Seaborn count plots were used to show the distribution of tweets by the sentiment and subjectivity label determined by TextBlob. Lastly, Seaborn box plots were used to observe the range of like counts for tweets by sentiment and subjectivity label to gauge how their audiences' reactions vary according to the content of the labeled parameters.

The analysis concluded that political inclinations associated with Fox News, BBC, and CNN could be substantiated through topic coverage and omission of details within their reporting. Figure 1 illustrates that the most prominent topic coverage was in cluster 0 for Fox, cluster 5 for BBC, and cluster 3 for CNN News.

Figure 1. Tweet Counts in Clusters by News Outlet

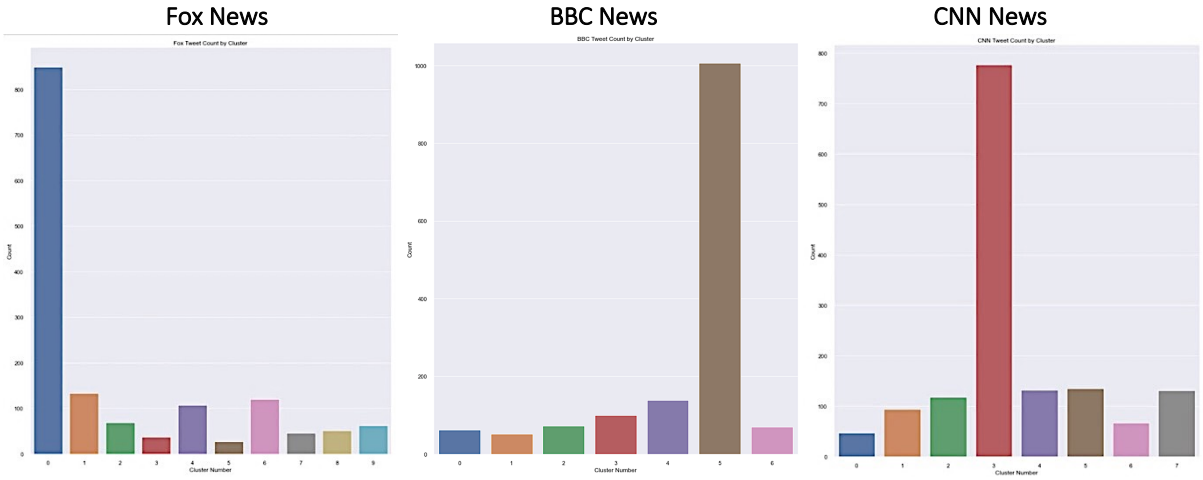
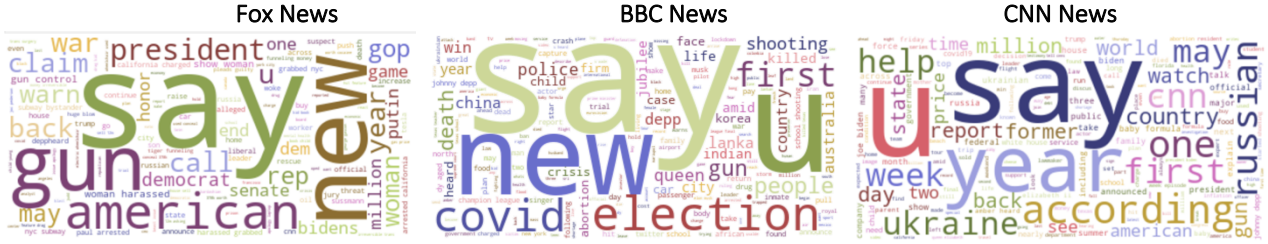


Figure 2 illustrates the word clouds for the most essential features of the tweets in these prominent clusters. The word clouds of the three news outlets reveal the Uvalde school shooting as one of the most reported top stories for BBC News. Unfortunately, this was not the case for Fox or CNN. For Fox, the bulk of reporting for this cluster was on gun regulation, American infrastructure, the Russia probe, and Biden about his role in these topics. For CNN, these topics predominantly focused on the baby formula shortage, President Biden's steps to address this issue, his visit to Asia, and the Ukrainian invasion. The differences in topics Fox and CNN chose to frequently emphasize align with the interests of their associated political parties.

Figure 2. Word Cloud & Centroid Features for Largest Cluster



Cluster 0:
say
bidens
gun
house
woman
video
american
russia
new
trial

Cluster 5:
say
australia
election
new
covid
texas
shooting
depp
north
korea

Cluster 3:
watch
say
year
baby
president
report
cnns
ukraine
formula
russian

Further analysis of how the Uvalde school shooting was covered confirmed that this was reported at significantly lower frequencies than other topics for both Fox and CNN. Furthermore, Fox's and CNN's coverage of the Uvalde shooting was 4% and 6%, respectively. However, Figure 3 and a closer look at the actual tweets confirmed that each news outlet differed in the information they included in their reporting, particularly around the aftermath of the event. Fox's frequent use of "police," "mass," "gun," "response," "safety," and "called" centered around the investigation being conducted for the police's response to the shooter. The tweets' focus is entirely on charges related to potential gunmen (unrelated to Uvalde), police responses to Uvalde shooting, and accusations on the root of mass shootings that range from "Christian Nationalism" to "patriarchy masculinity." CNN's Tweets frequently inform details about the school shooting victims. Tweet features like "elementary," "Robb," "killed," "child," and "19" provided the context of the school shooting's victims. Fox news coverage withheld essential details about the Uvalde school shooting, providing a narrative that appears to redirect concern to how the situation was handled by the police, as opposed to who was at the receiving end of the violence. While CNN also reported on the investigation of the police's handling of the school shooting, the emphasis was on informing and reiterating that the victims were children. After revisiting the Uvalde-related Tweets for BBC news in cluster 5, it was concluded that Fox was the only outlet that omitted any mention of the victims being children or that the target was an elementary/primary school.

Figure 3. Uvalde Topic Coverage & Most Important Features

Fox News

Cluster 1:
shooting
school
texas
uvalde
police
mass
gun
response
safety
called

CNN News

Cluster 5:
school
texas
shooting
uvalde
elementary
robb
killed
child
mass
19

Figures 4, 5, and 6 illustrate the distribution and percentage of Tweets by sentiment and subjectivity. Sentiment results on the data collected indicate that BBC, the 'centrist' news outlet, has the highest percent of objective and neutral

Tweets. Compared to 'centrist' statistics with BBC, the 'far-right' Fox Tweets do not significantly vary in objectiveness or sentiment. CNN, associated with a 'far-left' political affinity, reports the most subjective and non-neutral news via Twitter. Given the topic modeling and sentiment analysis findings, it appears that CNN uses descriptive language more than Fox News, resulting in higher frequencies of tweets that may be considered non-neutral and subjective.

Figure 4. Fox News Tweets Sentiment & Subjectivity

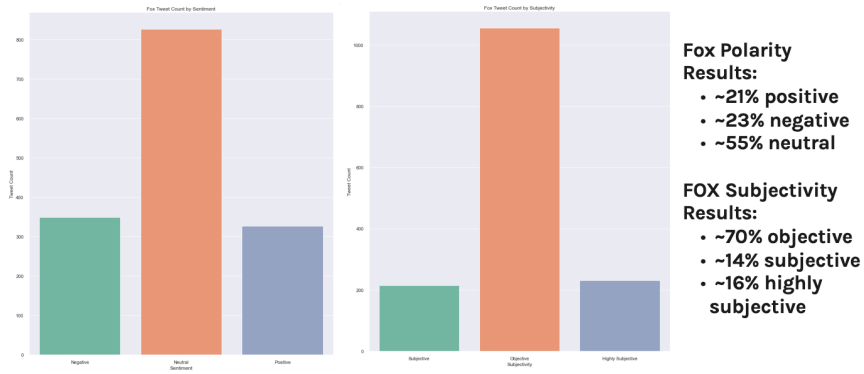


Figure 5. BBC News Tweets Sentiment & Subjectivity

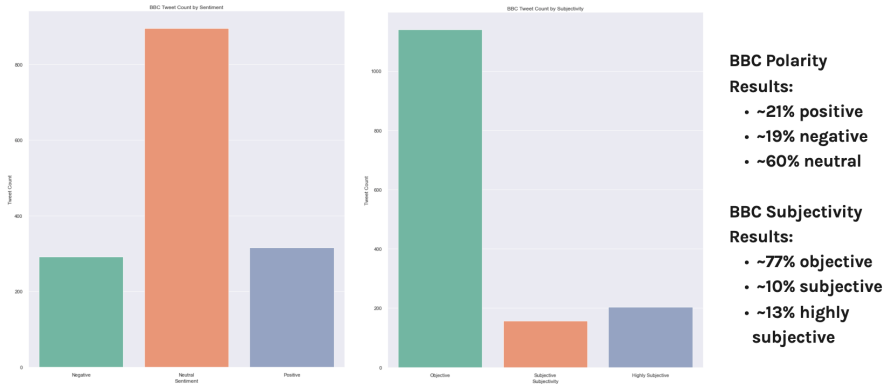
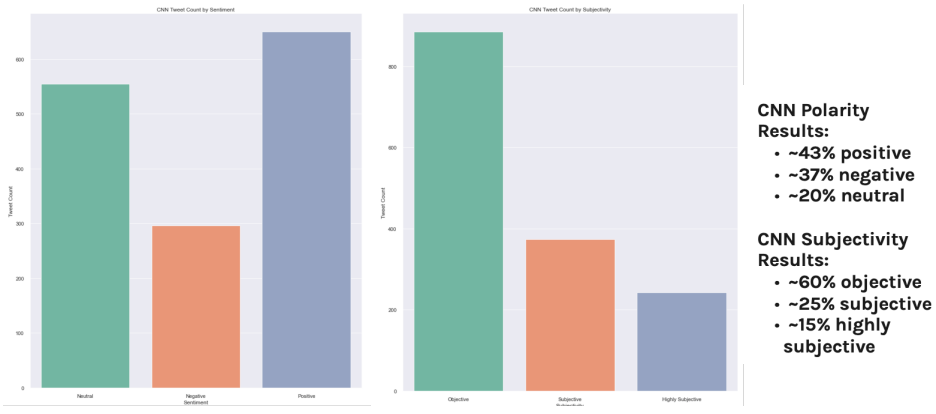


Figure 6. CNN News Tweets Sentiment & Subjectivity



The confirmation that political agendas were being promoted through selective reporting, by extension, also provided insight into how major news industries subtly perpetuated misinformation. Biased news reporting deepens people's distrust and distracts them from important issues, leading to an overall deflection of what is truly needed to solve current

concerns. The degradation of Fox and CNN's reporting credibility suggests that reevaluating what has been historically viewed as reputable news sources is necessary. Although warning notices and associated fact checks could help combat biased reporting, the findings in the analysis suggest audiences need to take the onus on ensuring they fully inform themselves through multiple means rather than blindly accepting a singular source. Failure to do so could result in severe social and political ramifications.

Preparation for Language Analytics Specialty and Skillset Reflection

This project demonstrated that I could extract and process raw text data. Data pre-processing steps such as changing the tweets to lower-case letters; removing punctuation, hashtags, mentions, special characters, and stop words; lemmatizing text; and tokenization were all used and proved to be foundational in the language analytics courses in the program. From a procedural aspect, I am confident in accessing data in various forms, cleaning it, and performing clustering or sentiment analyses. However, improving my skillset on leveraging other sentiment lexicons with negation could have enhanced this analysis. Not addressing the effects of negated tweets likely misrepresented Fox and CNN's sentiment and subjectivity labels. Similarly, getting better acquainted with other lexicon-based techniques to compare how the sentiment and subjectivity labels could vary would be beneficial for the overall fidelity of this project and future text-based tasks.

Recognizing Online Toxicity[2]

The IST 736 Text Mining course emphasized converting text data into a numerical form that could be used for unsupervised and supervised classification algorithms. With a particular focus on communicating findings to non-technical stakeholders, the culminating group project was designed as a fictional and realistic business issue. Research in various fields has undoubtedly concluded that online toxicity negatively affects its victims. With growing pressure on online platforms to accept accountability for eliminating online toxicity, the legislation that provides online platforms with immunity from civil liability based on third-party content, Section 230, is likely to be amended in the near future. Subsequently, the project's goal was to effectively identify harmful content as a consulting service for businesses wanting to proactively moderate their platform's online toxicity.

2 https://github.com/kdasil/IST736_TextMining

Ethically Leveraging Applicable Technologies

Using the Kaggle toxic comment CSV file, which contained 159,571 user comments from Wikipedia's talk page edits, multiple algorithms were used to detect varying forms of toxicity and to identify an optimal model for predicting whether new comments were toxic or not. To avoid biased results, this dataset's uneven distribution of labeled data was addressed by randomly sampling a subset with 6,360 non-toxic and 6,360 toxic comments.

Feature bias was an essential focus of the analysis. Allowing social group identifiers to be trained for identifying toxic content was explored to ensure discriminatory biases wouldn't be propagated in the models. Considering bigoted content composes most online toxicity, identifying features that should be removed to help reduce social biases without removing the essential context of the dataset was challenging. Ultimately, a list of features relating to countries, ethnicities, races, sexual orientations, and derogatory terms within these categories were curated and referred to as Social Group Features. The models were run with and without Social Group Features to compare the quality of essential features and training-to-test accuracy trends.

Using Python Visualizations and Predictive Models to Create and Communicate Actionable Insight

The scikit-learn library in Python was used to apply Latent Dirichlet Allocation (LDA), KMeans clustering, Naïve Bayes, and Support Vector Machines on the sampled dataset to provide insight on the following core questions:

- What types of toxicity exist in the dataset?
- Which words contribute the most to the identification of harmful content?
- Can we predict harmful content with high fidelity?

Given how critical it was to answer the stated core questions effectively, multiple variations of vectorizers provided by the scikit-learn library were used in the comments. The varying vectorizers and investigation of the effects of the Social Group Features were assessed by their respective predictive performances on test data.

When exploring topic modeling, LDA proved more effective than KMeans when separating non-toxic from toxic content. Additionally, it was able to detect identity hate from obscenities, providing insight into the trends within toxic language. Figure 7 visualizes the three topics LDA identified in a two-dimensional space, where the size of the topic circle is proportional to the number of words that belong in that category. This visualization effectively communicated which words defined each topic, along with their respective scope of words. Stakeholders would be able to conclude from Figure 7 that

circle 3 comprises identity hate-related language but that the scope of the words in this category is not as expansive as those in circle 1, which represents obscene terms. The ability to distinguish identity hate from obscenity allows business owners to confirm that more severe types of toxicity exist in their forums. Still, most importantly, they are positioned to target and handle it according to their company's guidelines. Since this is an interactive visualization, Figure 8 was included to see the Top-15 Relevant Terms for all topics in Figure 7.

Figure 7. Intertopic Distance Map and Top 30 Words for Topic Group 3

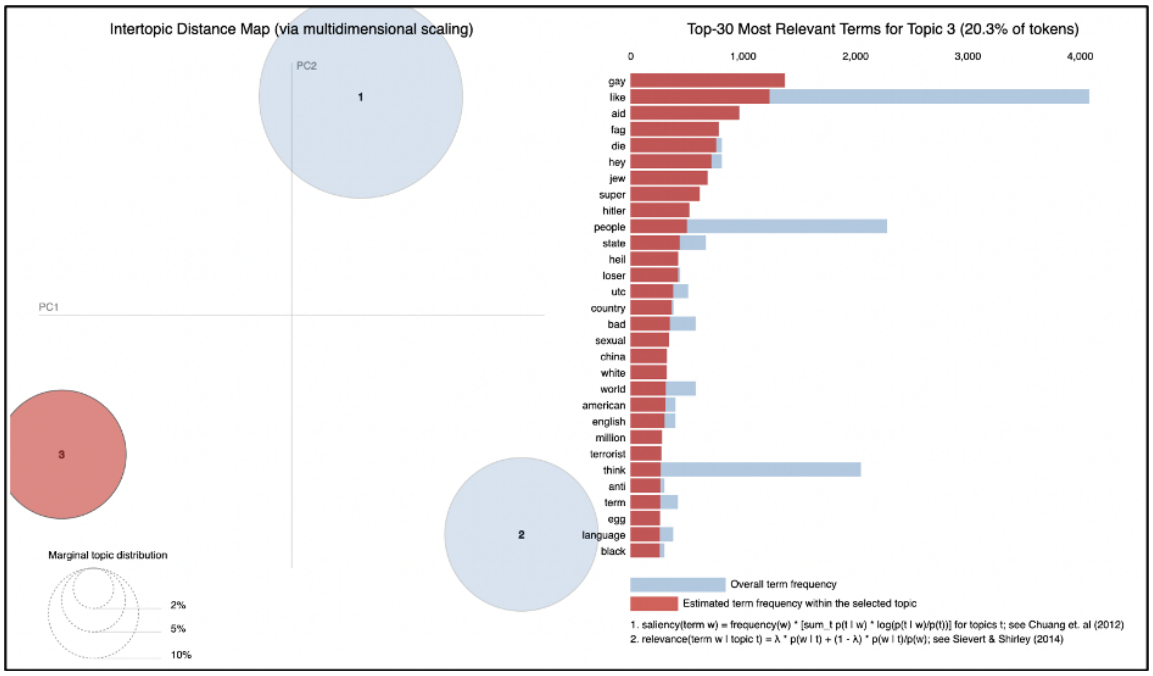
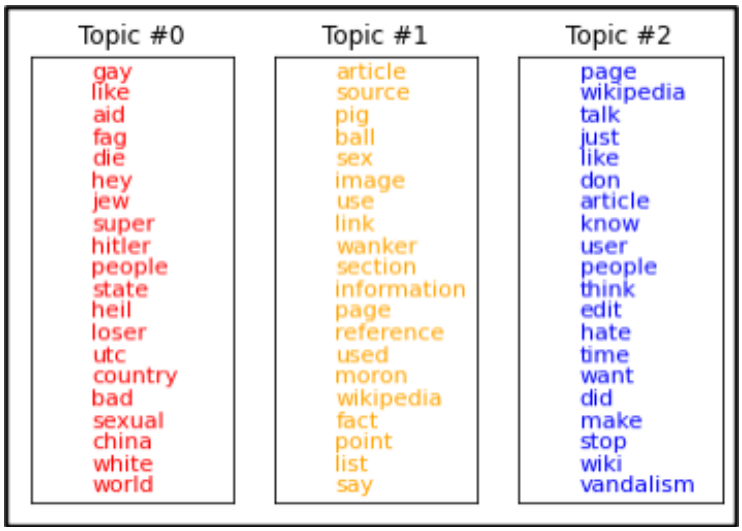


Figure 8. Top 30 Words by Topic Group



Multinomial Naïve Bayes resulted in better predictive performance over Support Vector Machines. Figure 8 and 9 illustrate the mean accuracy over five folds of cross-validation for all seven variations on the training and testing datasets. These plots indicate that the highest predictive performance was attributed to CV5 when it included the Social Group Features and CV1 when it excluded it. CV5 and CV1 refer to the vectorizer variations used when searching for an optimal model. The removal of Social Group Features resulted in a slight predictive decline, where CV1 was the only model to demonstrate higher predictive accuracy than its training performance. However, Figure 10 shows that the rest of the predictive results stemming from removing Social Group Features were below its training performance, an undesirable trend usually associated with overfitting.

Figure 9: Naïve Bayes Results Summary- with Social Group Features

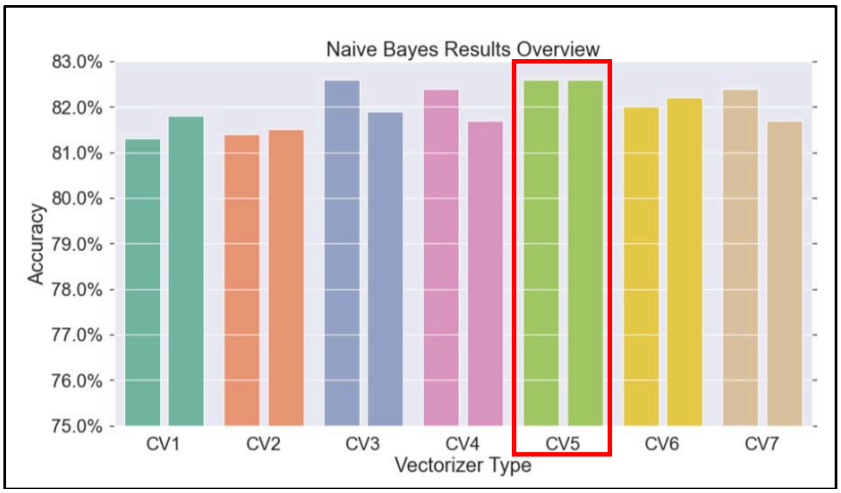
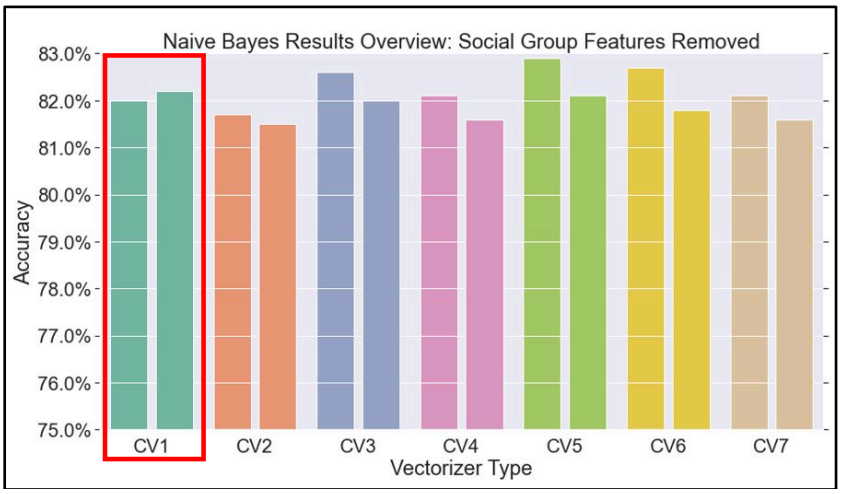


Figure 10: Naïve Bayes Results Summary- without Social Group Features



After five folds of cross-validation, the aggregated confusion matrices were included in Figures 11 and 12 for the training and testing datasets to investigate further whether the Social Group Features should be removed. Due to the slight decrease in predictive performance between CV5 and CV1, they were evaluating the types of errors each made was necessary before making conclusions. With a 10% false positive rate, both models demonstrated a propensity towards

identifying comments as toxic when they were not. However, Figure 11 and 12 shows that the false negative rates varied by 1% in the train and test set, making Naïve Bayes less likely to miss toxic comments when the Social Group features remained in the dataset. Due to the ramifications of having harmful content circulating online platforms, CV5, the model with the lowest false negative rate, was identified as the best predictive model for identifying toxicity in future online content.

Figure 11: Naïve Bayes Predictive Performance Comparison-with and without Social Group Identifiers

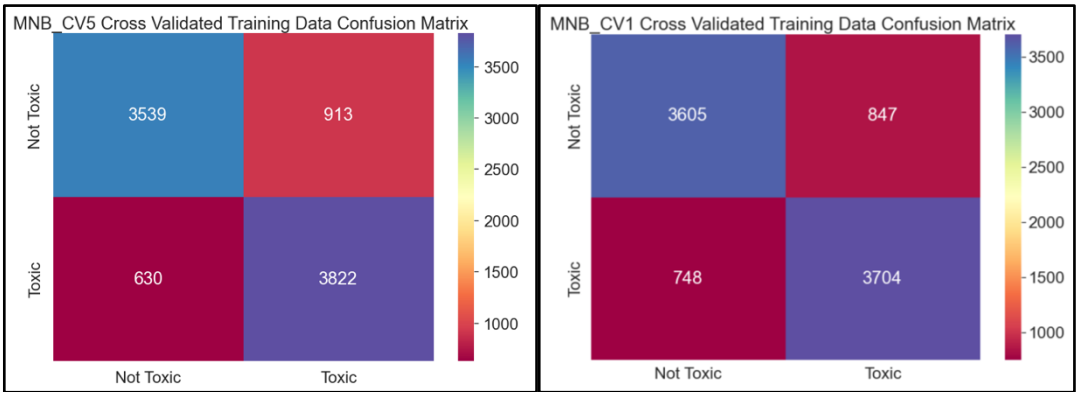
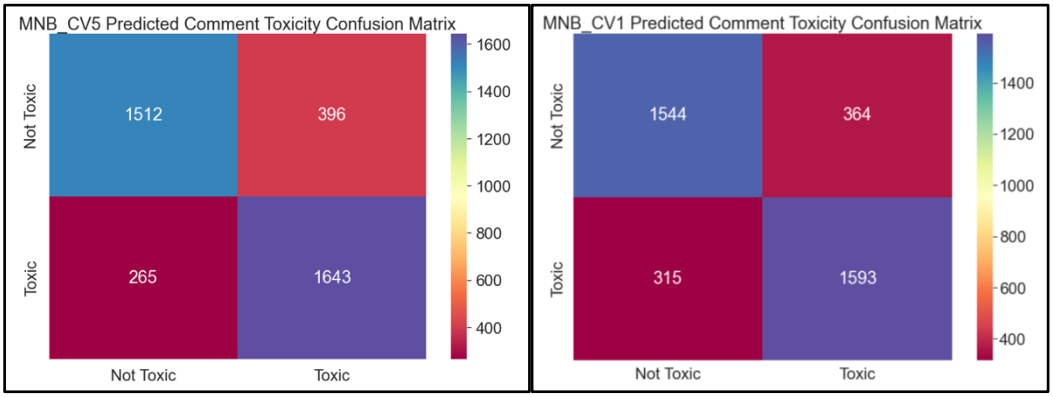


Figure 12: Naïve Bayes Predictive Performance Comparison-with and without Social Group Identifiers



The use of machine learning to distinguish and predict different types of harmful content has political, social, and health implications. Censorship, as a means to stifle online toxicity, is a contentiously debated topic. Often interpreted as a violation of free speech, citizens and policymakers have struggled to synthesize this right into its digital equivalent. Online toxicity is nuanced and cannot be dealt with monolithically. With LDA's ability to unearth the types of toxicity existing in forums and Naïve Bayes predictive performance, online platforms would be able to take appropriate actions against the different tiers of toxicity in a way that balances free speech and content moderation. Aside from being able to regulate online toxicity constitutionally, victims of online harassment would be spared from the emotional distress and trauma that has been statistically proven to emerge from it.

Preparation for Language Analytics Specialty and Skillset Reflection

As a core requirement in the Language Analytics track, all aspects of this course prepared me for topic modeling and predictive modeling using text. Considering what aspects of language could lead to biases in predictive models was the most challenging aspect of text mining, particularly in this project. In efforts to mitigate bias, manually removing derogatory and identity-based features proved to be a double-edged sword. This analysis concluded that efforts to prevent bias by removing the list of social identifiers diluted the quality of the components used to predict harmful content. It also led to overfitting. Handling bias in a dataset and model was a challenging task that would benefit from having legal, linguistic, and mental health subject matter experts curating a list of common words in subsets of harmful content that are not under the provisions of free speech. These subject matter experts would be able to offer insight into what constitutes accurate indicators of illegal toxicity, despite the different contexts and personal perspectives on toxicity.

Predicting Attrition[3]

The IST 707 Applied Machine Learning course introduced me to data mining methods needed to find and substantiate patterns relevant to real-world issues. Using R as the primary programming language, classification, clustering, and association rule mining algorithms were taught. The objective of the final project was to use these skills to solve a fundamental data mining problem. The Great Resignation is a term that refers to the unprecedented 4.4 million people quitting their jobs in the wake of the COVID-19 pandemic. The Business Journal reported that "retaining existing talent" was rated as the top management challenge by 79% of CEOs and that 67 % of CEOs said "attracting qualified talent" was a top concern[4]. Therefore, the goal of the analysis was to provide leaders with the information needed to optimize employee retention by effectively predicting attrition.

Using Visualizations and Predictive Models in R to Create and Communicate Actionable Insight

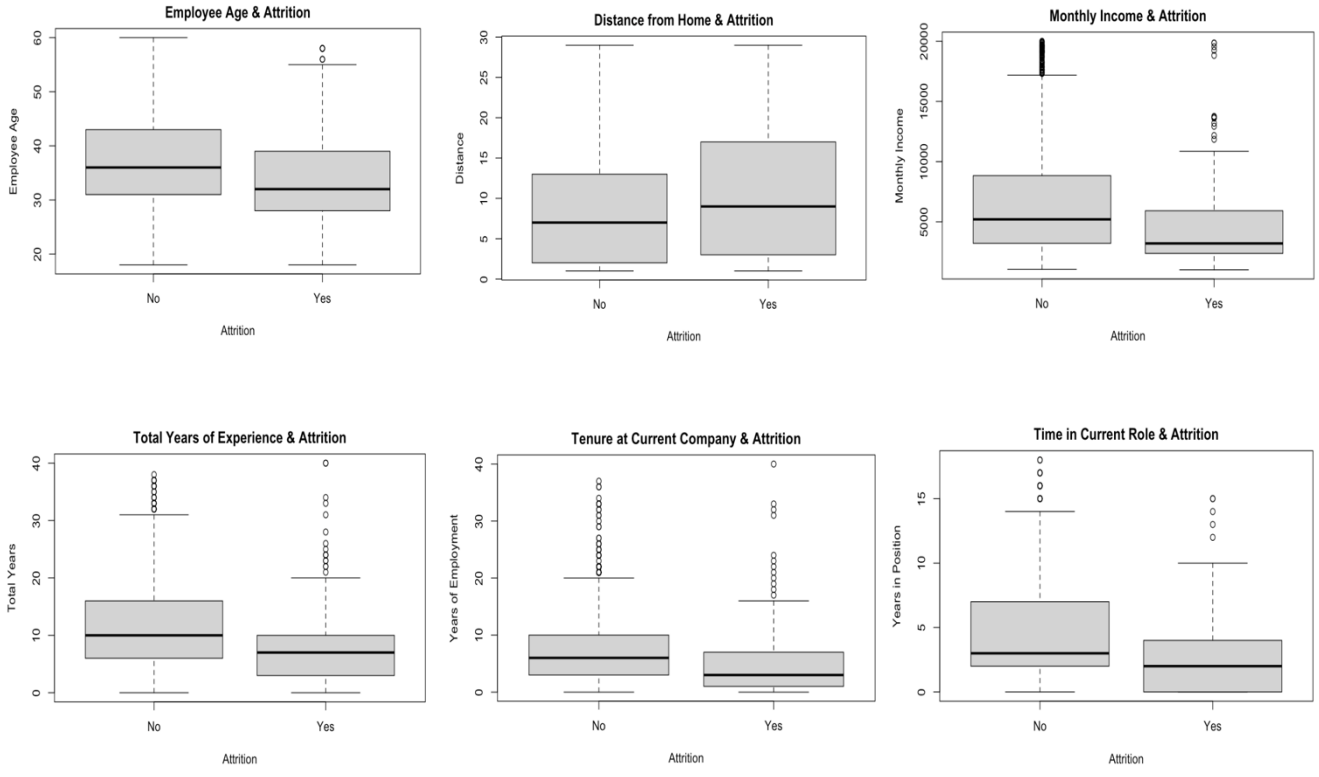
Box and bar plots were used to determine whether factors attributing to attrition could be identified. For example, Figure 13 statistically represents the profile of individuals by their employment status. The findings indicate that about 50%

[3] https://github.com/kdasil/IST707_MachineLearning

[4] https://www.bizjournals.com/bizjournals/news/2019/11/11/the-new-ceo-battleground-retaining-talent.html

of employees who leave are about 30 years old or younger, tend to live further away from work, earn less than $4,000 a month, and leave within the first three years at the company as a whole and within their current role. 75% of former employees earned around $6,000 or less, while only 50% of the remaining employees were in that pay range. Though tenure in the company and their current role were likely to happen within the first three years, the high number of outliers exceeding the 18-year range points to a different challenge. Generally speaking, a leader could decide to address the demographic of employees on the lower-earning scale and with shorter tenures by providing performance-based bonuses, ensuring salaries are competitive with the industry or improving other benefits that would offset their desire to go to another higher-paying company. A review of the current promotional system or available leadership positions would need to be evaluated to adequately address the small group of outliers choosing to quit after offering 18 years or more of service. Surveying the needs of this small yet essential demographic would better guide the development of a plan of action to retain their talent, which is not easily replaceable.

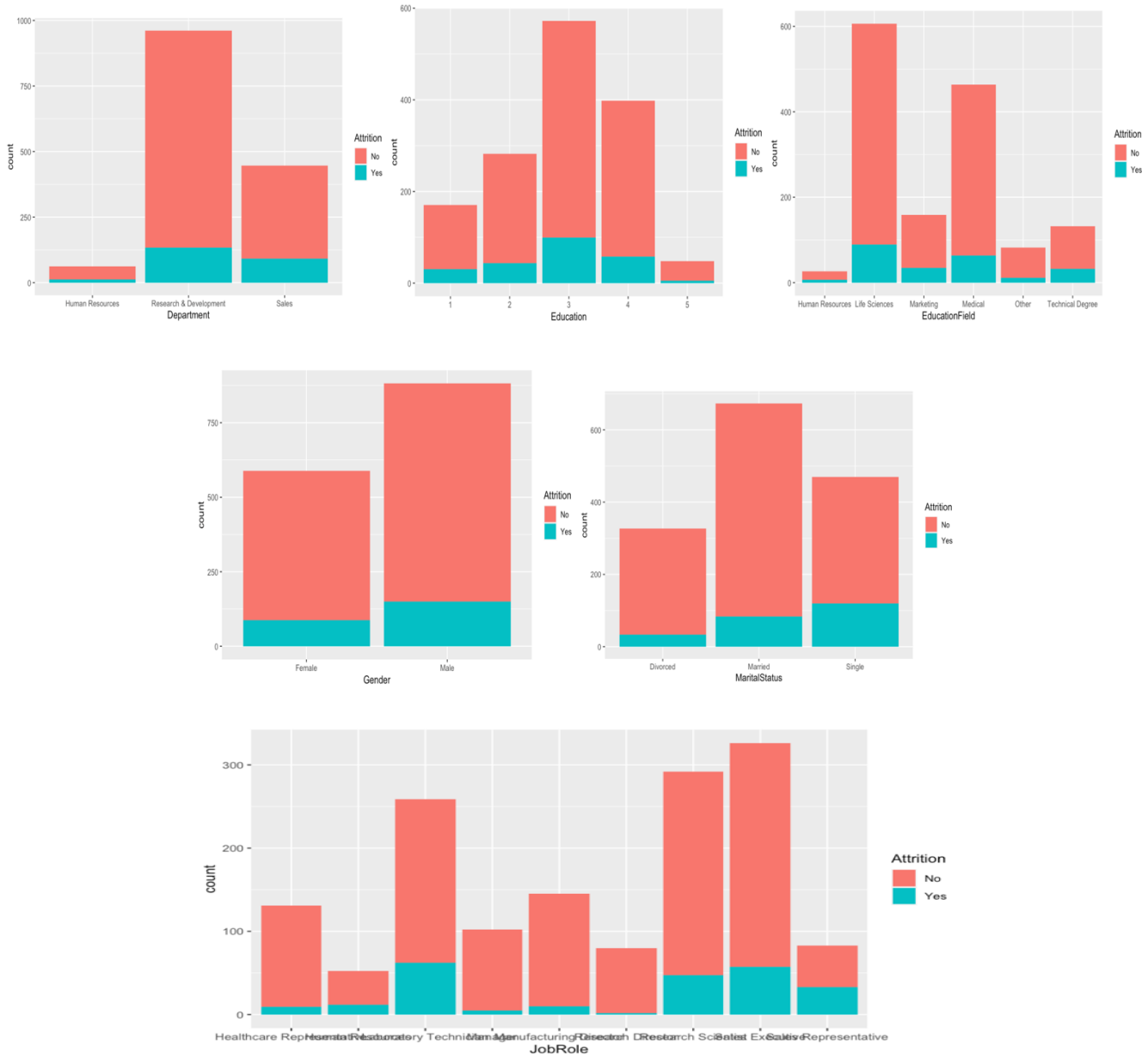
Figure 13: Employee Attrition Trends



The bar plots in Figure 14 grouped the categorical attributes that could specify a more detailed profile of the employees leaving the company. Proportionally speaking, the Sales department had the highest level of attrition, but with a

10% attrition being the industry-accepted healthy marker, the Research and Development department also needs improvement. Additionally, observed trends revealed that the exiting employees are more likely to be males, have Bachelor's degrees, be in the Life Sciences educational field, and be single. Within job roles, the largest group of exiting employees were Laboratory Technicians. These trends support the notion that the company has challenges retaining young employees and is more conducive to established family-makers. Schedule and pay in the Sales, Research, and Development departments and the Laboratory Technician role should be explored further to make sense of what contributes to their high attrition.

Figure 14: Employee Attrition Trends by Categorical Attributes



In preparation for the predictive models, the data set was divided into two groups with a set seed of 10, where 80% was randomly sampled into the train set and 20% into the test set. The set seed was necessary to produce the same random sample and to ensure the results were replicable. Furthermore, the data was trained using the K-Fold Cross-Validation to determine how reliable the two selected algorithms were. The first classification algorithm used was Random Forest, an ensemble learning technique that creates many trees on the subset of the data and combines all of the outputs. This technique is known for reducing overfitting and variance and improving accuracy.

Furthermore, Random Forest does not require normalization and can handle outliers easily, which was the case with this dataset. Due to the algorithm's stability, it was used with the data set as it currently was, even though the exploratory data analysis revealed that further data transformation would be needed to address outliers. Lastly, the data set was composed of categorical and continuous attributes, variables with which Random Forest works well. The second algorithm used was Support Vector Machines, another supervised machine learning model that uses a decision boundary as a line or hyperplane that best separates the data points. Anything that falls within each side of the line or plane is classified as its group. The best hyperplane is determined by its ability to maximize the margins from both groups. Although Support Vector Machines tend to have high accuracies, it is not known for handling large data sets and does not handle non-sparse data well, which is the case with this data.

Considering the advantages and disadvantages of each, both algorithms resulted in promising accuracy measures. Figure 15 includes the statistical fidelity of the predictive models, where both accuracies differ by a slight hundredth. However, with a Kappa value of.483, Support Vector Machines is undeniably the better model. In this context, Kappa values were used to evaluate the accuracy found in both algorithms, where values less than .4 are considered poor, and values between .4 and .75 are considered good. Random Forest had a poor Kappa value, meaning its classification performance was as good as randomly assigning the attrition label to the test set, hence the conclusion that Support Vector Machines was the superior model.

Figure 15: Predictive Model Results on Test Set

```
confusionMatrix(svm, HRNew_test$Attrition, positive = "Yes")

Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  252  24
       Yes   3  15

               Accuracy : 0.9082
                 95% CI : (0.8692, 0.9386)
    No Information Rate : 0.8673
    P-Value [Acc > NIR] : 0.0201964

                  Kappa : 0.483

 Mcnemar's Test P-Value : 0.0001186

            Sensitivity : 0.38462
            Specificity : 0.98824
         Pos Pred Value : 0.83333
         Neg Pred Value : 0.91304
             Prevalence : 0.13265
         Detection Rate : 0.05102
   Detection Prevalence : 0.06122
      Balanced Accuracy : 0.68643

       'Positive' Class : Yes
```

```
confusionMatrix(random_forest, HRNew_test$Attrition, positive = "Yes")

Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  253  26
       Yes   6   9

               Accuracy : 0.8912
                 95% CI : (0.8498, 0.9244)
    No Information Rate : 0.881
    P-Value [Acc > NIR] : 0.3329890

                  Kappa : 0.3108

 Mcnemar's Test P-Value : 0.0007829

            Sensitivity : 0.25714
            Specificity : 0.97683
         Pos Pred Value : 0.60000
         Neg Pred Value : 0.90681
             Prevalence : 0.11905
         Detection Rate : 0.03061
   Detection Prevalence : 0.05102
      Balanced Accuracy : 0.61699

       'Positive' Class : Yes
```

The findings in this analysis concluded that identifying incentives relevant to individuals under the age of 30, single, and in Sales or Research and Development within the first three years of employment is crucial in retaining new talent. To a lesser degree, career opportunities for employees with employment tenures beyond 18 years are also necessary, as finding replacements with comparable experience was doubtful. An effective predictive model was found and could be applied by company leaders to flag potential resignations based on the attributes that defined their leaving demographic. These efforts holistically placed leaders in a position to deter turnover expenses, lost knowledge, and productivity roadblocks, thus promoting organizational growth.

Identification of Populations Most Vulnerable to Suicidal Behavior[5]

The IST 719 Information Visualization course used R programming to create visual artifacts to share findings with various audiences. Color, size, and appropriate plots were strategically used to explore and communicate trends geared toward augmenting understanding. Discussions on our ethical responsibility to convey accurate and relevant data were emphasized. Now more than ever, data deception is rampant. Learning how visualizations have been distorted to manipulate audiences according to agendas was valuable, not only to avoid falling prey to these unethical practices but also to know how to inspect work from colleagues effectively.

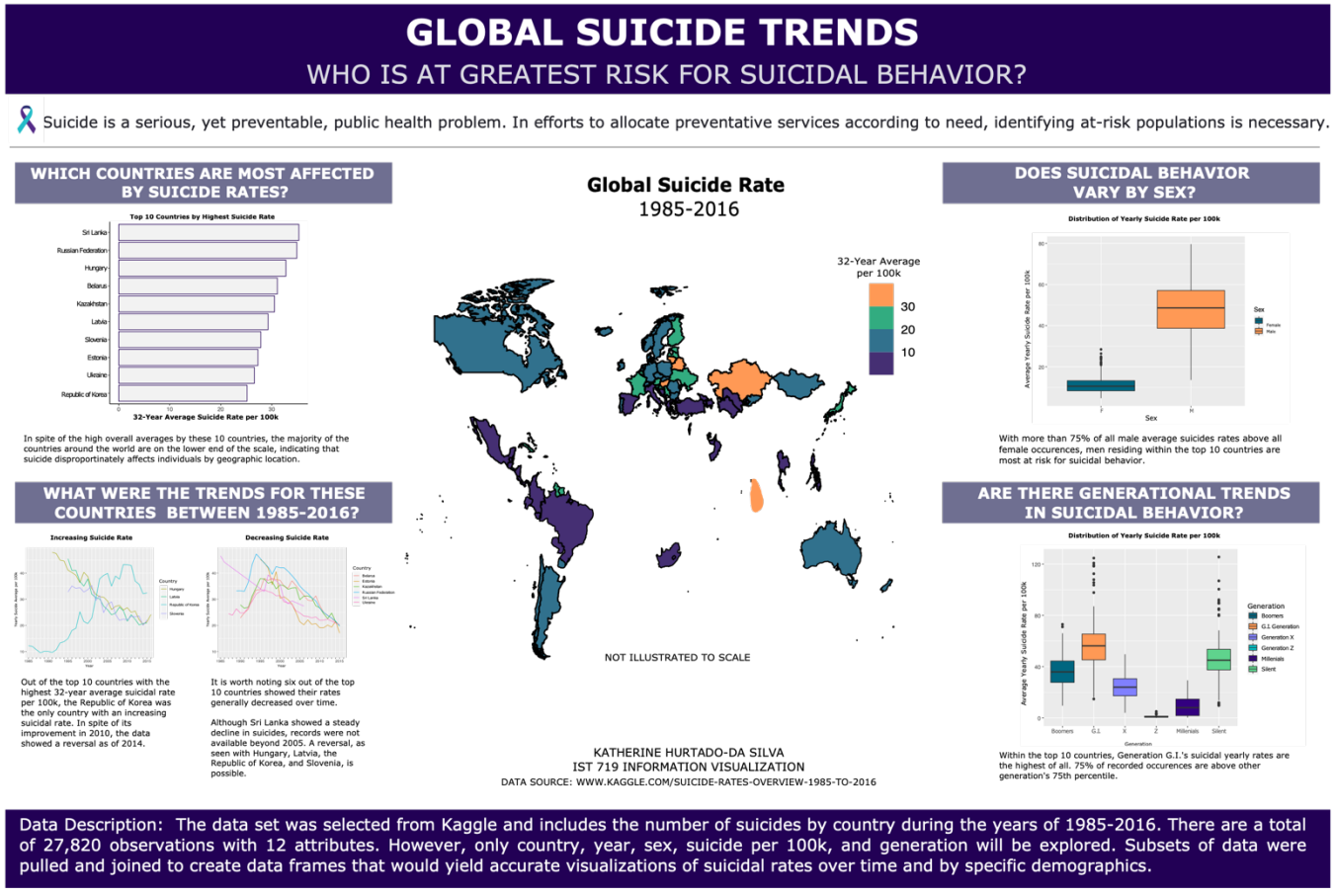
Using Visualizations in R to Create Actionable Insight

The final poster project leveraged skills developed throughout the class. The data was selected from Kaggle and included the number of suicides by country from 1985-2016. There were a total of 27,820 observations with 12 attributes.

[5] https://github.com/kdasil/IST719_InformationVisualization

However, only country, year, sex, suicide number, suicide per 100k, Human Development Index (HDI), Gross Domestic Product (GDP) per capita, and generation group was explored. The omissions were due to the attributes not adding insight or being aggregated attributes within the data set. For example, suicide number and population were omitted because the crude rate per 100k represented both values. The GDP for the year's removal was also due to the GDP per capita being based on this value. Figure 16 includes the final project poster, where bar, line, and box plots were used to communicate which demographic communities were adversely affected and in most need of interventions. Organizations wanting to target the most vulnerable demographics would be able to conclude that Sri Lanka, the Russian Federation, Hungary, Belarus, Kazakhstan, Latvia, Slovenia, Estonia, Ukraine, and the Republic of Korea were the global leaders in the suicide rate. In contrast, the Republic of Korea was most critical considering its continuous increase in yearly suicides. Aside from geographic location, males and individuals born during the G.I. generation were most vulnerable to suicide, meaning that resources could be distributed according to this demographic if focusing unilaterally on a single location was impossible.

Figure 16: Providing Insight on Global Suicide Trends



Closure

The Applied Data Science program taught me how to do predictive, descriptive, prescriptive, and diagnostic analytics. With a focus on Python, R, Excel, and SQL programming throughout the program, I am equipped with more tools to execute analytical tasks. Being familiar with these different programming languages has also allowed me to adapt more quickly to industry or future employer standards. Expanding the types of information I can work with, such as categorical data and corpora, has contributed to me being more competent in answering complex questions currently relevant to industries. Gathering raw data despite site limits is another advantage that will be leveraged in future tasking involving real-time content.

As previously discussed, my language analytics skills will improve with more familiarity with sentiment lexicons. Being acquainted with ongoing research studies in this field will also serve as the basis for enhancing the methodologies I can currently employ in text analytics. Bias in models is complicated and challenging to consider, especially when taking into account the different fields machine learning could be applied. This is another example of how remaining up-to-date on this field will be critical in propagating ethical models. Though biased datasets are typically the point of conversation when discussing ethical algorithms, there are instances where bias is introduced in the aftermath, meaning the decisions and actions made with the model's data is what taint the overall efforts. Given the fast-moving evolution of machine learning, it is hard to fathom how a data scientist would not engage in life-long learning in this field. I intend to stay well-informed of new developments and their associated ethical ramifications.