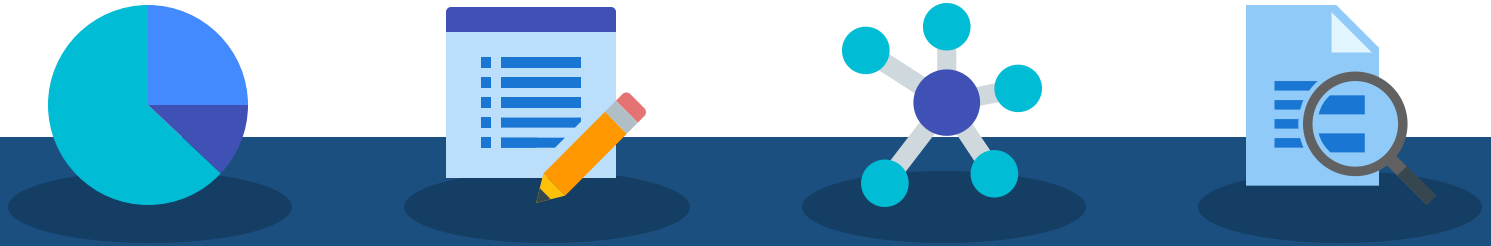
Syracuse University
School of Information Studies

Reflection on Fulfilling the Applied Data Science Learning Objectives



Katherine Hurtado-da Silva
SUID: 781774642
Email: khurtado@syr.edu

Applied Data Science Learning Objectives:

- 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)

Project Agenda:

1. Influence of Political Affinities on News Reporting via Twitter
2. Recognizing Online Toxicity
3. Predicting Attrition
4. Identification of Populations Most Vulnerable to Suicidal Behavior

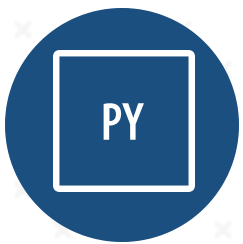
Sources

1. https://github.com/kdasil/IST652_ScriptingforDataAnalysis
2. https://github.com/kdasil/IST736_TextMining
3. https://github.com/kdasil/IST707_MachineLearning
4. https://github.com/kdasil/IST719_InformationVisualization

Influence of Political Affinities on News Reporting via Twitter

IST 652 | Scripting for Data Analysis

Course Overview: The goal of this class was to learn the tools and skills of scripting necessary to solve problems of accessing, collecting, preparing and storing data in a variety of formats and situations.



Programming Language: Python

Project Background: Skills acquired in this class enabled the collection of tweets via Snscrape and overall analysis to conclude whether politically-biased reporting was evident on Twitter.

Influence of Political Affinities on News Reporting via Twitter

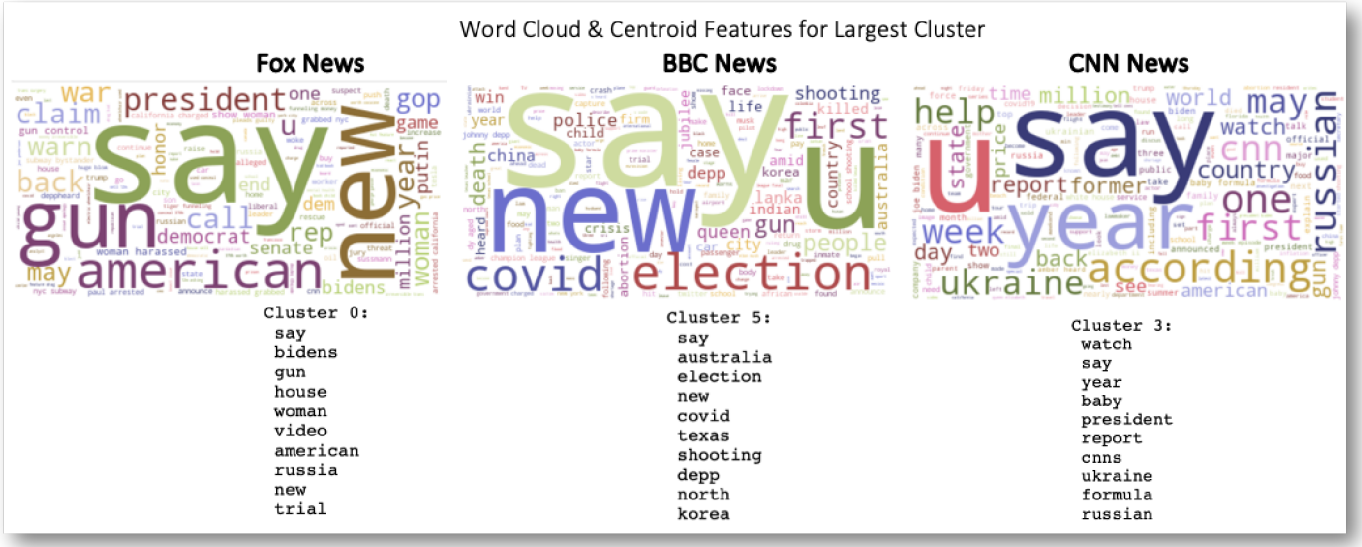
Methodology

Fox News, BBC, and CNN were selected because they are considered far-right, centrist, and far-left news outlets, respectively.

1,500 tweets between 4-1-2022 and 6-5-2022 were collected for each news outlet to ensure the Uvalde School shooting, a polarizing event that split right-wing gun rights supporters from democratic gun-ban advocates, was covered.

KMeans clustering, word clouds depicting the resulting clusters' centroid features, sentiment analysis, and bar plots showing the distribution of tweets by sentiment label were used to answer:

- How did reported topics for each news outlet differ?
- How did the sentiment and subjectivity vary according to each news outlet's tweets?



For Fox, the bulk of reporting for its largest cluster was on gun regulation, American infrastructure, the Russia probe, and Biden's role in these topics.

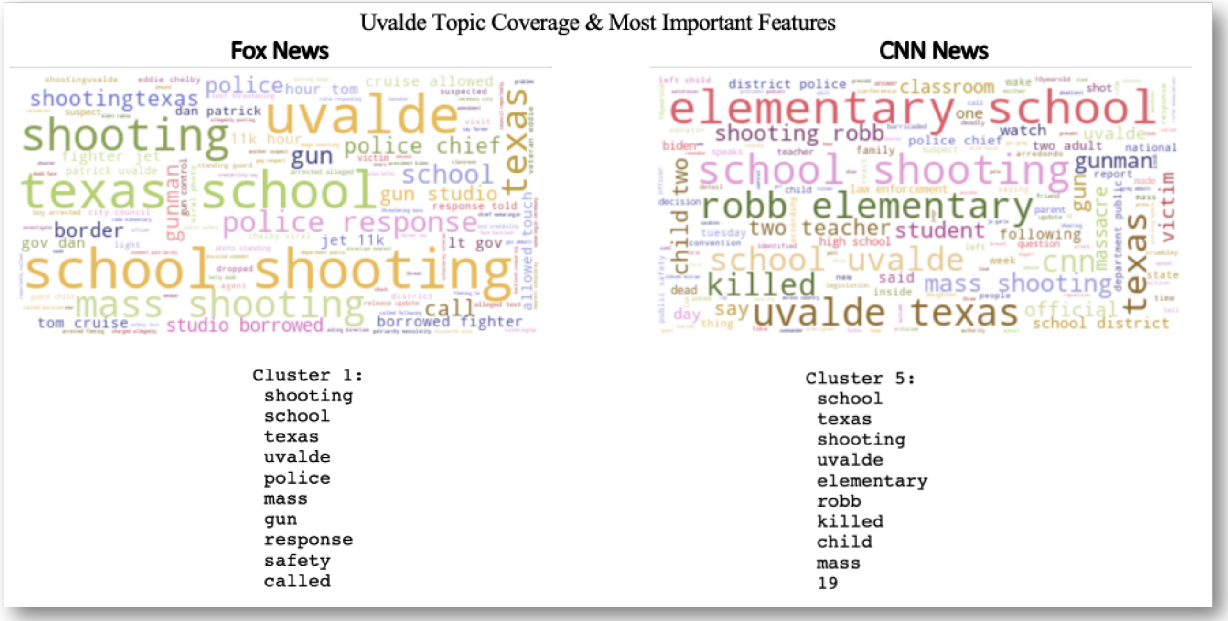
For BBC, the Australian election, Johnny Depp case, COVID outbreak in North Korea, and Uvalde school shooting were the most reported stories.

For CNN, most coverage was focused on the baby formula shortage, President Biden's steps to address this issue, his visit to Asia, and the Ukrainian invasion.

Fox and CNN most frequently covered topics align with the interests of their affiliated political parties. BBC News was the only news outlet with the Uvalde School shooting in its largest cluster, indicating Fox and CNN prioritized other matters in their Twitter account.

Influence of Political Affinities on News Reporting via Twitter

How did reported topics for each news outlet differ?



Fox News

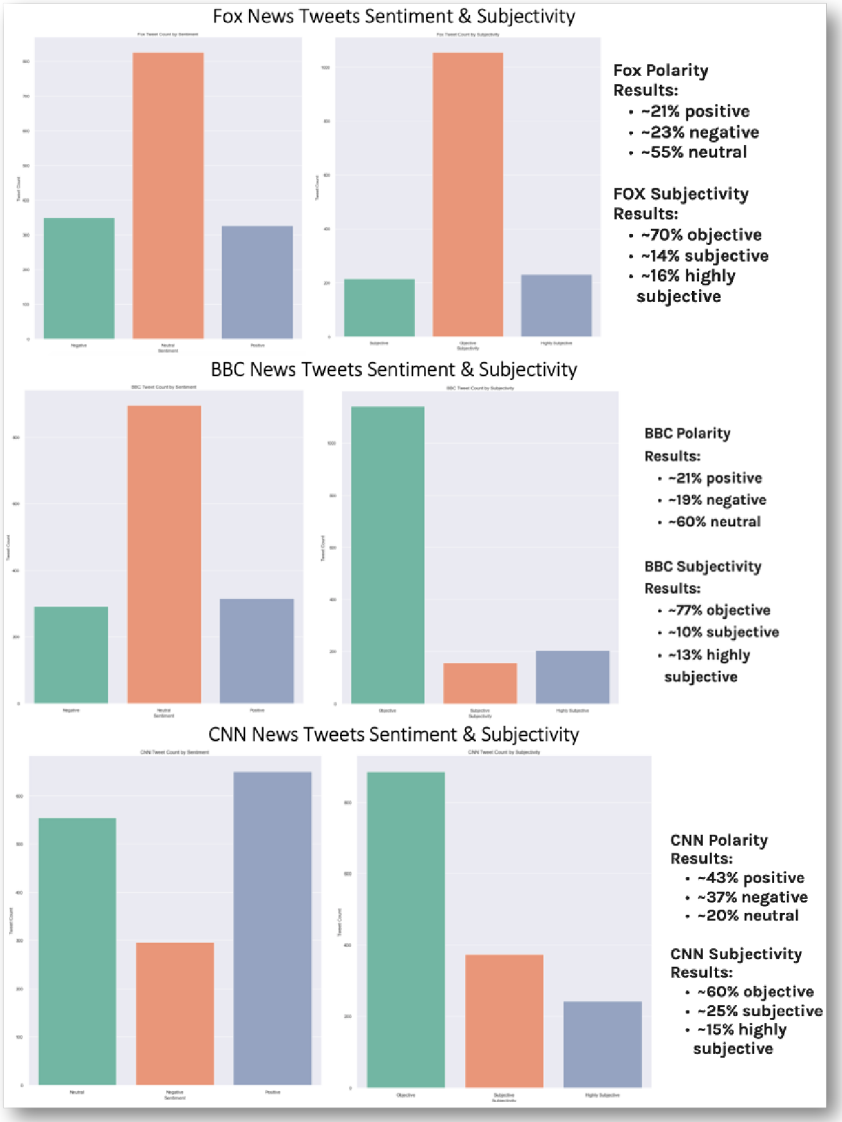
- Accounted for 4% of total reporting
- Only news outlet of the three to omit that the victims were children and that the target was an elementary/primary school.
- Coverage appeared to redirect concern to how the situation was handled by the police, as opposed to who was at the receiving end of the violence.

CNN News

- Accounted for 6% of total reporting
- Reported on the investigation of the police's handling of the school shooting
- Emphasis was on informing and reiterating that the victims were children

Influence of Political Affinities on News Reporting via Twitter

Unpacking the Uvalde School Shooting Coverage



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.

Influence of Political Affinities on News Reporting via Twitter

How did the sentiment and subjectivity vary according to each news outlet's tweets?

Influence of Political Affinities on News Reporting via Twitter

Conclusion

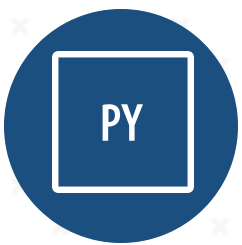
- Findings conclude Fox and CNN News selectively reported topics that favored their political affinities in higher frequencies.
- Findings show Fox omits details in reporting that do not favor their political agenda. CNN News uses highly subjective language in coverage, indicating they also engage in biased reporting.
- The degradation of Fox and CNN's reporting credibility suggests that reevaluating what has been historically viewed as reputable news sources is necessary.

Recognizing Online Toxicity

IST 736 | Text Mining

Course Overview: The goal of this class was to learn the concepts and methods necessary for:

- knowledge discovery from large amounts of text data
- the application of text mining techniques for business intelligence, digital humanities, and social behavior analysis.



Programming Language: Python

Project Background: Skills acquired in this class were used to effectively identify harmful content for businesses wanting to proactively moderate their platform's online toxicity.

Recognizing Online Toxicity

Methodology

Dataset: Kaggle toxic comment CSV file, which contained 159,571 user comments from Wikipedia's talk page edits

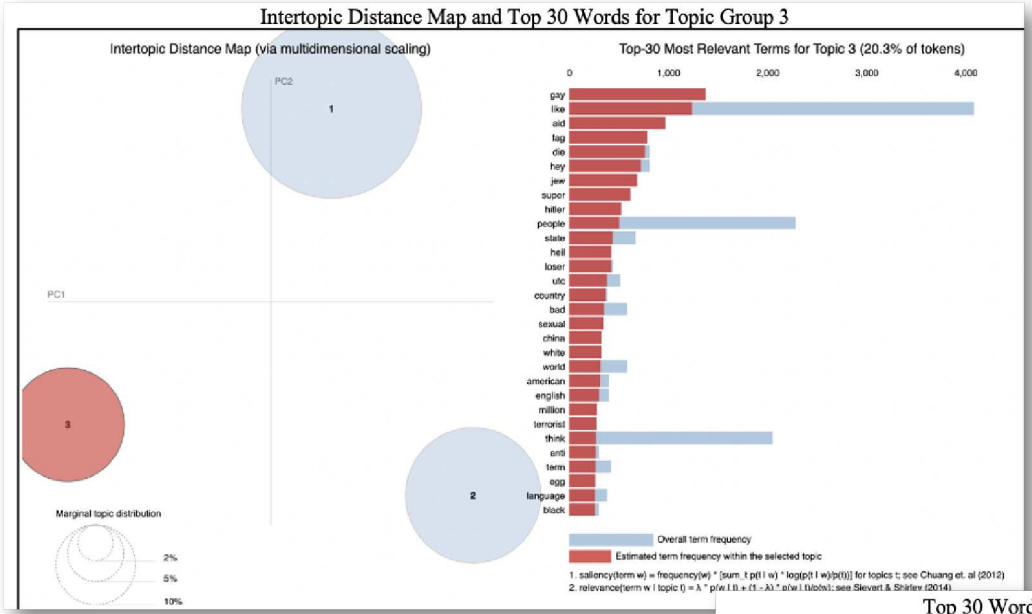
Combatting feature bias in models: a list of features relating to countries, ethnicities, races, sexual orientations, and derogatory terms within these categories were curated and referred to as Social Group Features. The models were run with and without Social Group Features to compare the quality of essential features and training-to-test accuracy trends.

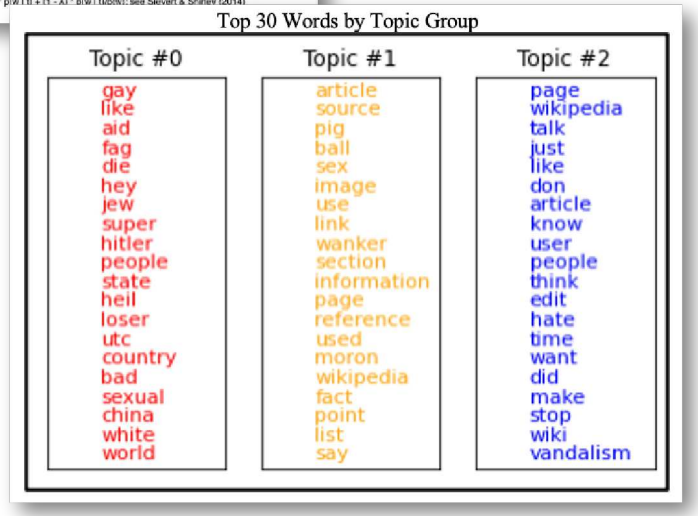
The scikit-learn library in Python was used to apply Latent Dirichlet Allocation (LDA), KMeans clustering, Naïve Bayes, and Support Vector Machines on the sampled dataset to provide insight on the following core questions:

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

Five-fold cross validation over seven vectorizer variations were applied to training and test sets as a measure against overfitting

Recognizing Online Toxicity





LDA identified three topics in the contents of the dataset. Aside from non-toxic comments, it was able to distinguish identity hate from obscenity.

What types of toxicity exist in the dataset?

Multinomial Naïve Bayes resulted in better predictive performance over Support Vector Machines.

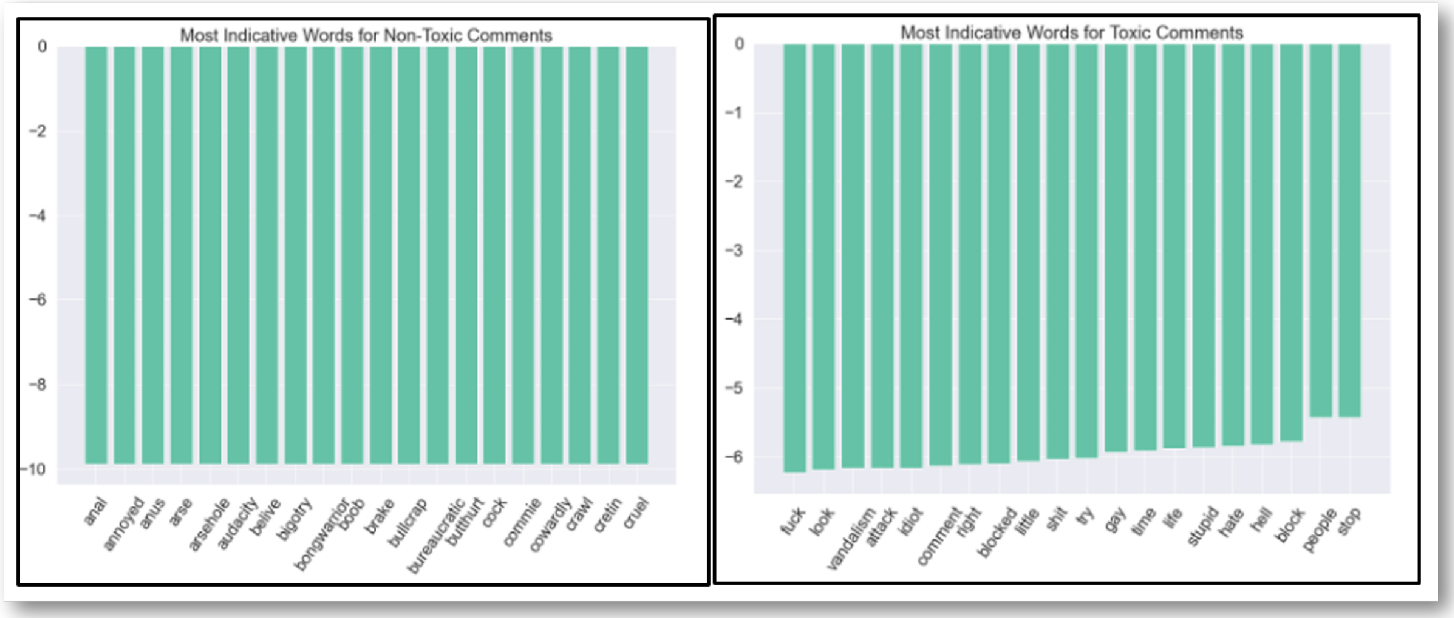
The results overview indicate that the highest predictive performance was attributed to CV5 when it included the Social Group Features and CV1 when it excluded it. CV5 and CV1 refer to the vectorizer variation used when searching for an optimal model.

The removal of Social Group Features resulted in a slight predictive decline. The rest of the predictive results stemming from removing Social Group Features were below its training performance, an undesirable trend usually associated with overfitting.

Recognizing Online Toxicity

Can we predict harmful content with high fidelity?

Recognizing Online Toxicity



The features “anal”, “annoyed”, “anus”, “arse”, “arsehole”, “audacity”, “belive”, “bigotry” and “bongwarrior” were mostly used to classify a comment as non-toxic.

The features “fuck”, “look”, “vandalism”, “attack”, “idiot”, “comment”, “right” and “blocked” were mostly used to predict toxic comments.

Which words contribute the most to the identification of harmful content?

Logarithmic probabilities were used to generate the top 20 features used to differentiate non-toxic from toxic comments. The plots illustrate these features for the best model, Naive Bayes with vectorizer CV5 and Social Group Features.

Recognizing Online Toxicity

- Findings show LDA was able to unearth the types of toxicity existing in the comments, obscenities and identity-hate speech.
- The removal of social group identifiers did not improve the predictive performance and led to overfitting.
- Naive Bayes was able to predict toxic comments with a mean 82.6% fidelity.
- All of these tools can help online platforms take appropriate actions against different tiers of toxicity, in a way that balances free speech with content moderation, while sparing online harassment victims from the emotional distress that has been statistically proven to emerge from it.

Conclusion

Predicting Attrition

IST 707 | Applied Machine Learning

Course Overview: The goal of this class was to learn the popular data mining methods for extracting knowledge from data and to become familiarized with real-world applications.

Programming Language: R

Project Background: The data mining skills acquired in this class was used to provide employers with information needed to optimize employee retention and predict attrition within their existing workforce.

Predicting Attrition

Methodology

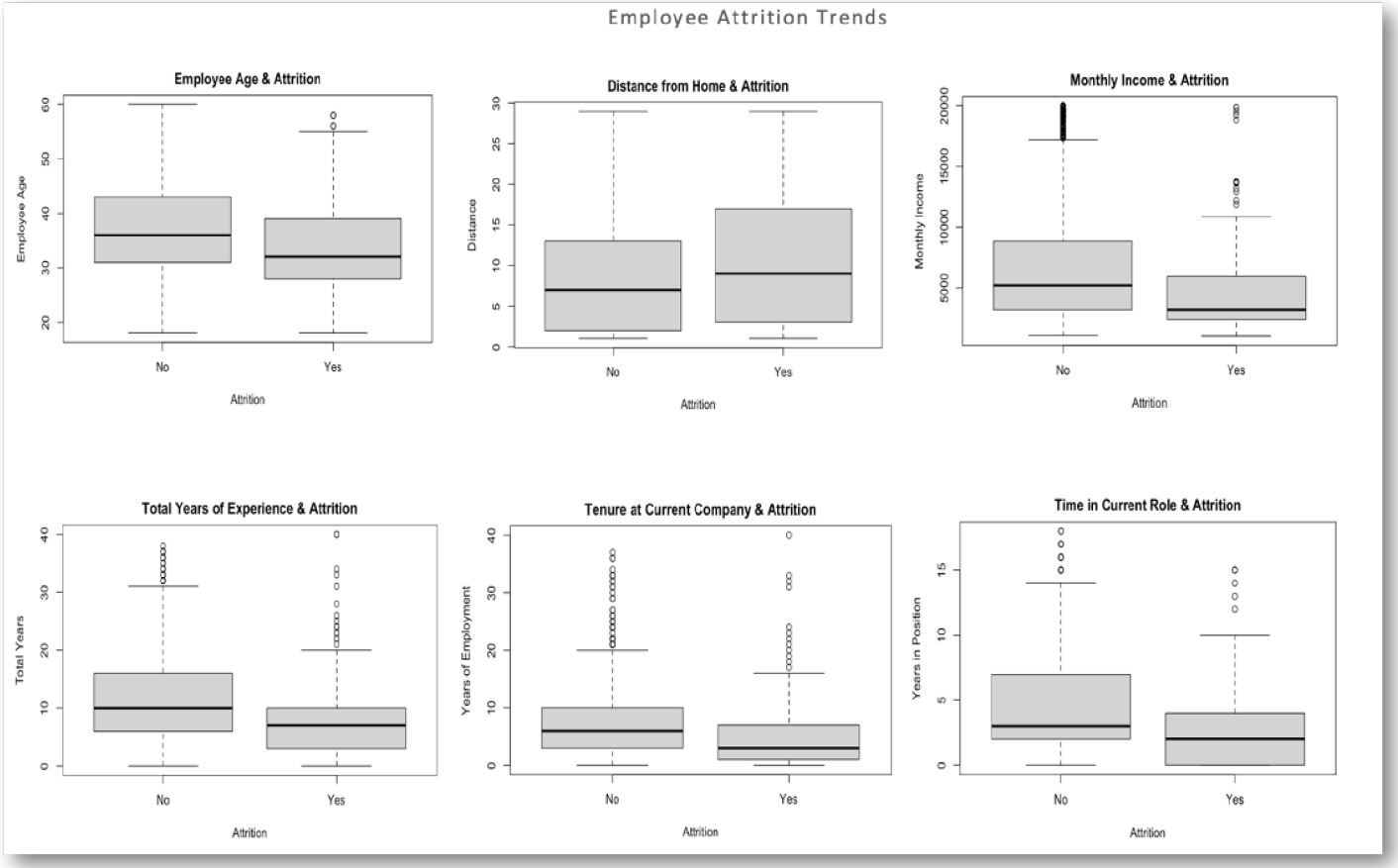
Dataset: IBM HR data that includes 1,470 employee records with 35 attributes

Box and count plots were generated to identify trends in population of people leaving the company

Two supervised learning algorithms were applied, Random Forest and Support Vector Machines to predict resignations.

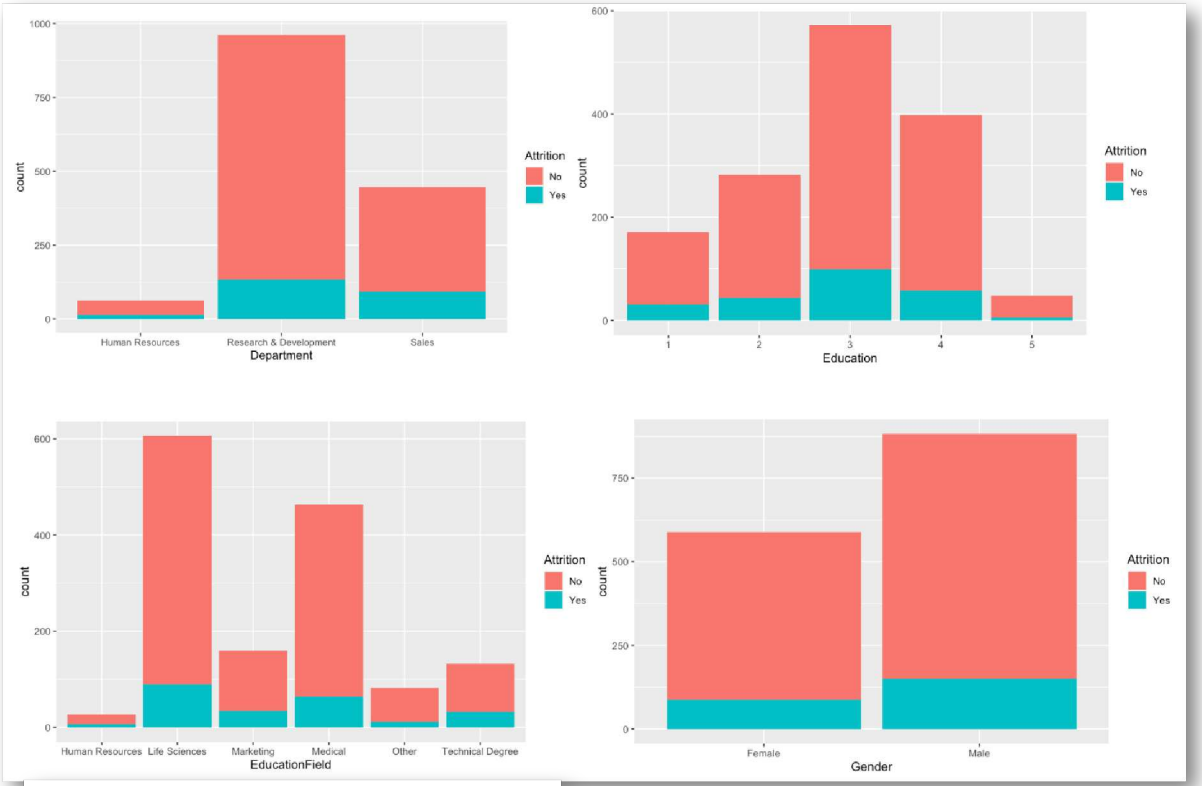
Six-fold cross validation was applied to the training and test sets as a measure against overfitting

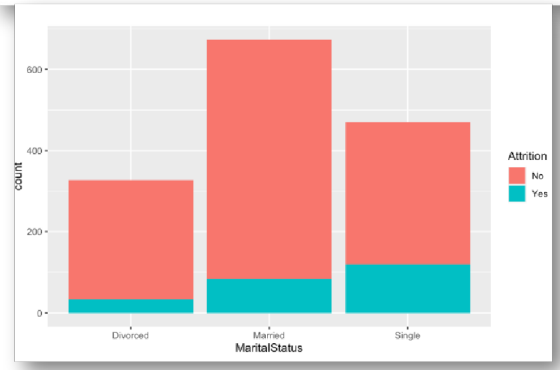
Predicting Attrition



Trends in Resigning Population

The findings indicate that about 50% 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.





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.

Predicting Attrition

Trends in Resigning Population

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
```

Despite both algorithms resulting in promising accuracy measures, Support Vector Machines was the superior model due to its Kappa value of.483, meaning its classification performance was not the same as randomly assigning the attrition label to the test set.

Predicting Attrition

Algorithmic Performance

Predicting Attrition

Conclusion

- Findings conclude 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
- 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 can help leaders promote organizational growth by deterring turnover and its associated expenses, lost knowledge, and productivity roadblocks.

Identification of Populations Most Vulnerable to Suicidal Behavior

IST 719 | Information Visualization

Course Overview: The goal of this class was to learn data cleaning techniques, how to control R graphics environment, develop custom plots, visually explore data, use design concepts to visually communicate the story in the data, and discuss issues related to the ethics of data visualization.

Programming Language: R

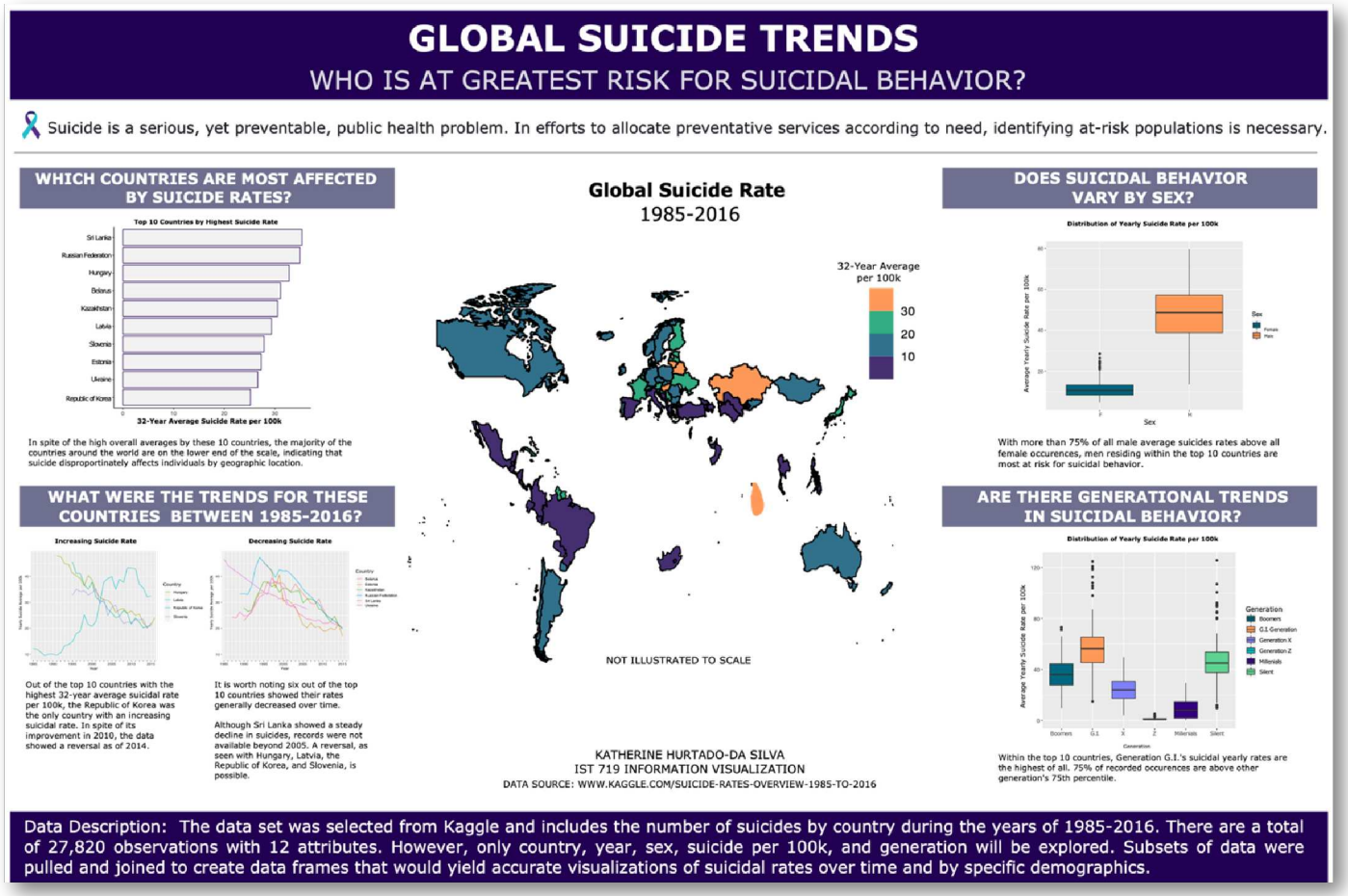
Project Background: Skills acquired in this class enabled the generation of varying visualizations necessary to convey which demographics were more susceptible to suicidal behavior from 1985-2016.

Identification of Populations Most Vulnerable to Suicidal Behavior

Methodology

Dataset: Sourced from Kaggle, where the number of suicides by country was provided from 1985-2016. A total of 27,820 observations with 12 attributes was included. 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.

Bar plots, line plots, box plots, and a heat map map were used to communicate which communities were most vulnerable and in most need of interventions.



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. Males and individuals born during the G.I. generation were most vulnerable to suicide.

Identification of Populations Most Vulnerable to Suicidal Behavior

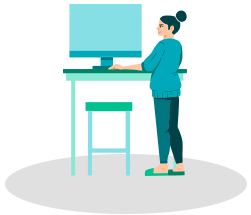
Identification of Populations Most Vulnerable to Suicidal Behavior

Conclusion

Organizations wanting to target the most vulnerable demographics by geography 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, with The Republic of Korea being most critical due to its continuous increase in yearly suicides.

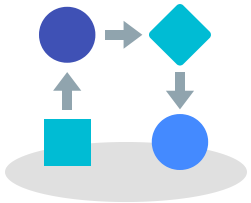
Males and individuals born during the G.I. generation were most vulnerable to suicide, meaning that resources could be distributed according to this demographic if focusing unilaterally on a single location was impossible.

Final Takeaways



Am able to complete various forms of analyses using Python, R, Excel, and SQL programming

Am able to develop and communicate findings from visualizations and predictive models that generate actionable insight



Acquired data wrangling and mining skills necessary to perform analytical tasks with numerical, categorical and text data,

Will engage in life-long learning to improve natural language processing skills and bias mitigation in the development and deployment of algorithms.

Thank You