Brian Taylor
SUID# 251968713
rtaylo11@syr.edu

Portfolio Introduction

This portfolio encapsulates some of the projects I completed during my time in the Applied Data Science program at the Syracuse University iSchool. The five highlighted projects are organized to show a progression of ability and understanding in the scientific application of data and demonstrate my ability to perform the following tasks:

1. Describe a broad overview of the major practice areas in data science.
2. Collect and organize data.
3. Identify patterns in data via visualization, statistical analysis, and data mining.
4. Develop alternative strategies based on the data.
5. Develop a plan of action to implement the business decisions derived from the analysis.
6. Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization.
7. Synthesize the ethical dimensions of data science practice (e.g., privacy).

The first project was one of my early projects in a Database Admin course. I built a database for a local youth sports program that captured important attributes of the players and coaches which originated from data on paper. This would be my only database heavy course but highlighted a basic of data science, you need data. Further, I was forced to organize the data in a way that guarded privacy as much as possible and under the constraints of the youth organization. In this project, the club is real and not a fictitious, hypothetical client.

The second project's focus was on trending YouTube videos. Data was collected, munged, transformed, statistically analyzed and used to develop multiple models to find popular videos. I developed an attribute called the popularity index (PI) which was calculated from available attributes but aimed to capture viewer involvement with the video as a measure of popularity.

Disasters in the United States was the topic for the third project in this portfolio. The purpose was to identify type, frequency and location, of previous disasters in an attempt to aid disaster response professionals to understand any patterns that might exist over the previous 64 years' worth of data. The result represents my first foray into the subsection of data visualization as a technique to convey information to a broad audience.

Building a well performing fantasy football team using text from Tweets was a well-intentioned, attempt to identify trending players. After some initial analysis, the focus of the project was to be able to predict characteristics in three attributes. Sentiment analysis was performed on the text, after extensive cleaning for processing. Multiple models were tested and displayed varying levels of performance in predicting the attributes.

Crime in Raleigh, North Carolina was the topic for the final project highlighted in this portfolio. Prevention and avoidance are both desirable for law abiding citizens. As an exercise in telling a

Brian Taylor
SUID# 251968713
rtaylo11@syr.edu

story with very few words, I analyzed open source reported crime data and developed multiple visualizations to convey to information to the view in a poster presentation theme. This project captures many of the tasks of this program of study and was one of my final projects.

The following projects attempt to represent the breadth and depth of my knowledge acquired in this program of study.

Project 1 – IST 659 – Youth Basketball Performance Evaluations

Introduction

This project was a database design and implementation exercise using data from a local youth basketball club created every season on subjective, observed player performance during tryouts and at the end of the season to answer questions about the players and coaches in the club.

The purpose of the data collection activity portion was to observe players and evaluate their subjective performance on certain pre-defined skills, as well as an "overall" category, on a 1 to 5 scale. The goal of the process is to prevent creating "stacked" teams by identifying player's skill levels share information with all coaches to assist in the seasonal draft and allow coaches unfamiliar with certain players to have a broad assessment of a player's performance.

Pre-season evaluations were collected from each coach and averaged among the coaches/evaluators. Post season assessments were done by the coach, or at least supposed to be, so ideally each player had two evaluations per season. On many occasions, the coach did not complete an assessment for all of their players.

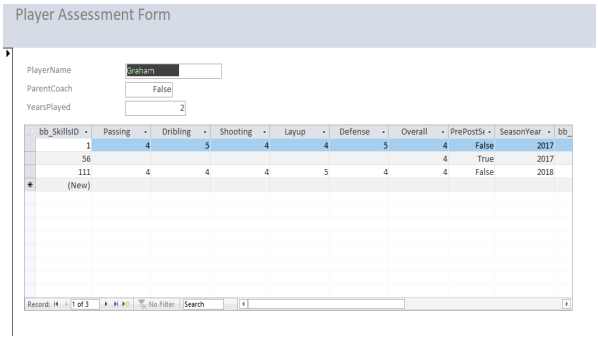
Personal privacy concerns were an issue from the beginning. Raw data was provided from every evaluator to the commissioner, however the commissioner averaged the scores and reported those scores for use. This was a way to obscure tying a coach's opinions to their names and protect coach privacy. Further, only the first name of the coach given and identification numbers were used to represent coaches in queries while only a player's first name was used to identify the player. On occasion, players with common names required the use of the first letter of their last name.

Given the age of the players, any player could have played as many as one previous season. If a player played in the previous season, their previous season's coach was not identified or linked in any way to avoid the club president's concern about personal privacy. As the players are young children, the data was pulled from the beginning of their time with the club so data is limited to as many as two pre-season evaluations and one post-season evaluation. There was one exception in that one player played up an age group before the club offers a team in that year cohort.

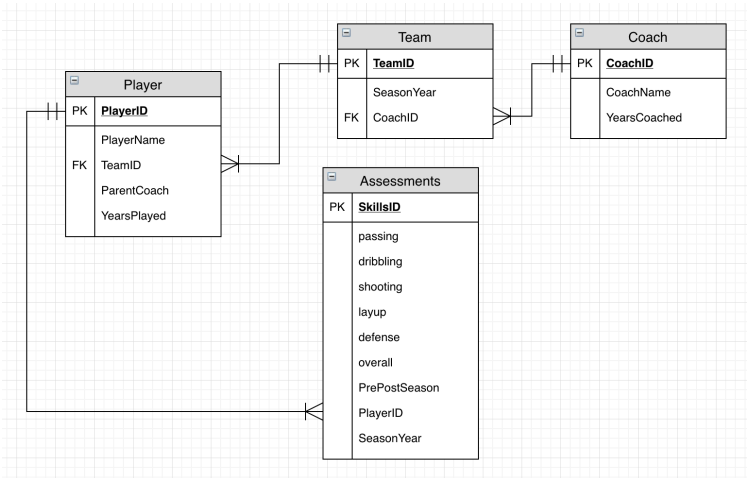
Brian Taylor
SUID# 251968713
rtaylo11@syr.edu

Methods

Data was collected from the club commissioner and manually entered through Microsoft Access into an SQL database through a player assessment form. An example of an assessment form is below.



Four database tables were used in an attempt to maintain separate records for each entity. Entities were defined as Player, Coach, Team and Assessment. Each entity possesses certain attributes shown below.



ERD Glossary

PlayerID – surrogate key of the player and foreign key for the PlayerAssessment
PlayerName – Name of player
TeamID – surrogate key of the team that multiple players and each coach are assigned to, foreign key to identify teams over time
Parent Coach – attribute is "Y" is the player's parent is a coach
YearsPlayed – years player played in league
CoachID – surrogate key of the coach of the team and foreign key for the Team entity
CoachName – Name of coach
YearsCoached – years coach has coached in the league
SkillsID- surrogate key of the assessment
Passing – passing skill of the player
Dribbling – dribbling skill of the player
Shooting – shooting skill of the player
Layup – layup skill of the player
Defense – defensive skill of the player
Overall – overall skill of the player
PrePostSeason – an attribute differentiating the mass pre-season assessment and the coach's post season assessment
SeasonYear- the attribute that helps distinguish when an assessment was conducted

Brian Taylor
SUID# 251968713
rtaylo11@syr.edu

Queries were used to investigate answers to questions such as “Are coaches returning season to season?” or “Is there a stacked team?”. Examples of those queries and result are shown below.

```
-- Data Question 4, Coach retention
SELECT bb_Coach.bb_CoachID
     , bb_Team.SeasonYear
     FROM bb_Team
     INNER JOIN bb_Coach ON bb_Team.bb_CoachID = bb_Coach.bb_CoachID
     GROUP BY bb_Coach.bb_CoachID, bb_Team.SeasonYear
```

100 %

Results

bb_CoachID	SeasonYear
1	2018
2	2017
2	2018
3	2017
3	2018
4	2017
4	2018
5	2017
5	2018
6	2018
7	2018
8	2017
8	2018
9	2018
11	2017
11	2018
21	2017
24	2017
26	2017
27	2017

```
-- Data Question 5, stacked team?
SELECT bb_Team.bb_TeamID
     , CAST(AVG(bb_Assessment.Overall) AS DECIMAL(4,2)) AS AvgOfOverall
     FROM ((bb_Player INNER JOIN bb_Assessment ON
          bb_Player.bb_PlayerID = bb_Assessment.bb_PlayerID)
          INNER JOIN bb_Team ON
          (bb_Team.SeasonYear = bb_Assessment.SeasonYear)
          AND (bb_Player.bb_TeamID = bb_Team.bb_TeamID))
          INNER JOIN bb_Coach ON
          bb_Team.bb_CoachID = bb_Coach.bb_CoachID
     GROUP BY bb_Team.bb_TeamID
GO
```

100 %

Results

bb_TeamID	AvgOfOverall
11	3.00
12	3.00
13	3.00
14	3.00
15	3.00
16	3.00
17	3.00
18	2.00
19	3.00
20	3.00

(10 row(s) affected)

Results

Based on the performance characteristics evaluated and if the club were to continue the process, many aspects attributes of the club could be evaluated over time. Individual player development could be tracked and evaluated and relative team strength can be monitored. For a non-profit club that requires payment to support its activities, coach and player retention is an important attribute to be aware of, particularly as children grow and either move on from the club for a higher level of competition, quit the sport all together, or in a highly transient metro area, simply move away.

It is common to have the same players on the same team year after year which could enhance one team at the expense of other teams. This separate attribute could be tracked and possibly used to attempt to make teams as equitable as possible for increased competition within the club. On the contrary, finding a cohesive and well performing team could help the club select players to invite to try out for an elite club team for competition outside of the club with the age groups best players.

Finally, this database could evolve and start to capture team and individual performance per game and season to start to objectively score player and team performance.

Conclusion

Modern Database Management by Hoffer et al. (2019), was crucial in developing my understanding of databases. Jumping back into formal education after 17 years, I had a hard time adjusting to a mostly digital class and reference library, and the hard copy, as well as this

Brian Taylor
SUID# 251968713
rtaylo11@syr.edu

project, aided my transition to the modern adult learning based model. This project was the first major deliverable in my first Data Science graduate course.

Conducting database queries is a foundational step in applying data science to many problems, especially when the data is proprietary as it is in most companies or organizations. Development of a database all the way from evaluating players as a coach to being the analyst answering the relevant club questions exposed me to the techniques and processes required to set up a usable database as well as a competence with Microsoft Access and SQL.

This project was my largest data science challenge at the time and revealed the importance of designing a database and attributing, aggregating, and querying data. I applied these concepts in my current profession while aggregating multiple data sources and extracting, loading and transforming comma separated value files for transportation analysis and presentation to senior Department of Defense Officials.

Demonstrated links (1,2,7)
Database design and management is an important area of knowledge when conducting data science as database queries are a common method for collecting data necessary to conduct data analysis techniques.

Through this project, I encountered ethical questions regarding personal information used to identify people is a real and sensitive issue for people of all ages and worked with the stakeholders to find a compromise that satisfied all parties.

I also learned that data collection from raw pen and paper to a fully operational database is a challenge that I knew nothing about. Through this project I learned how to create and ingest data, create a database, query a database and how to use these techniques to answer relevant questions for the youth basketball club. I progressed from having no knowledge of databases to a competent and applicable level of knowledge on database design, development and management.

Project 2 – IST 707 – Trending YouTube Videos

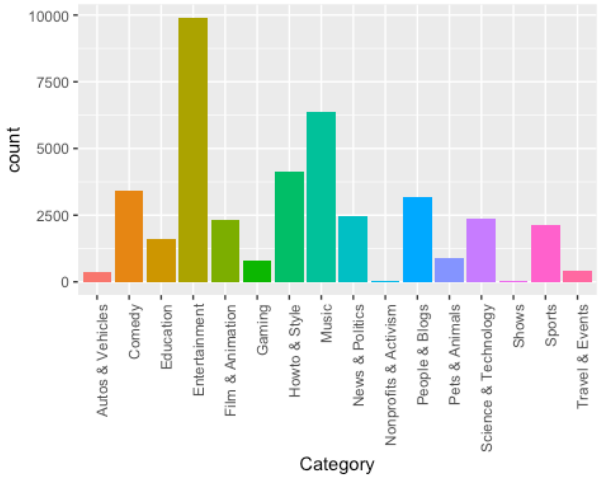
Introduction

This project utilized publicly available information with respect to YouTube video statistics to identify trending video characteristics. Identifying Social Media trends is big business. Multiple million-dollar careers were built on “going viral” which started with an online video. Online video advertising is a growing business with about 86% of marketers using video to get their messages out. I wanted to identify which type of content would generate the most exposure to YouTube viewers. As a result, two main questions were developed to drive the analysis. First, what key factors effect popular videos on YouTube? Second, how can we identify popular videos early on to maximize views of our ad in a popular video?

Brian Taylor
SUID# 251968713
rtaylo11@syr.edu

Methods

To identify key factors of trending videos, I first conducted initial data analysis of the dataset. Out of 16 attributes per video, six attributes were primarily used for further analysis. These were: Likes, Dislikes, Views, Comment_count, Category_id, and Channel_title. The top three categories for number of trending videos were Entertainment, Music and How-to & Style. The number of trending videos in each category are shown below.

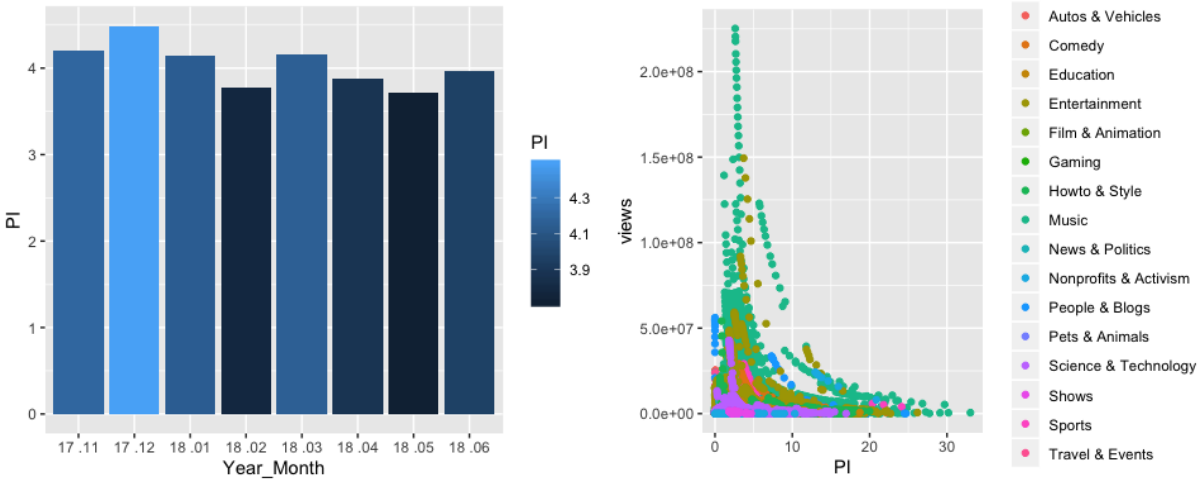


To identify popular videos, I created an evaluation variable called a popularity index, or PI. The PI was calculated as the sum of likes, dislikes and comment counts divided by the number of views.

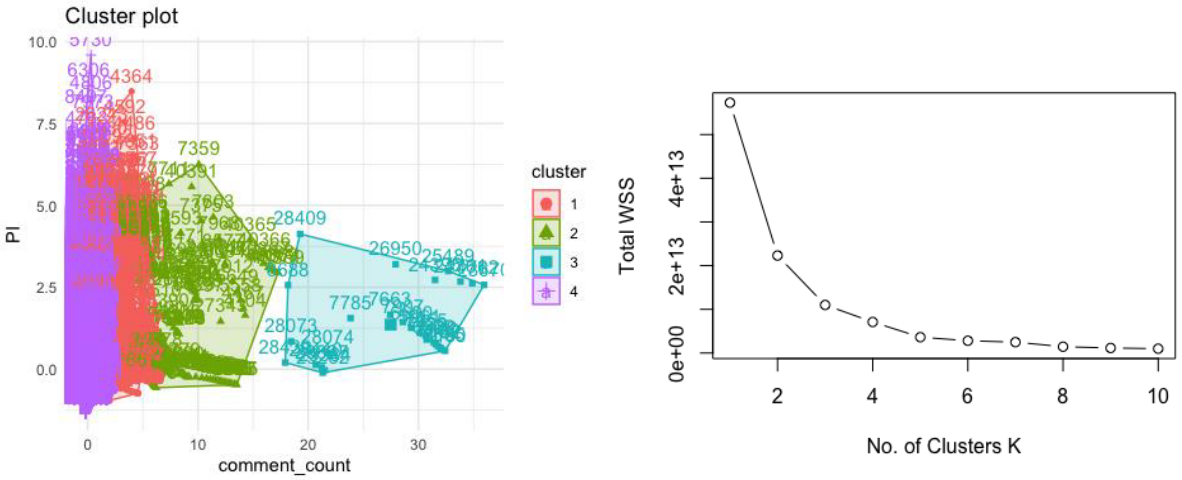
$$PI_{video} = \frac{likes_{video} + dislikes_{video} + number\ of\ comments_{video}}{number\ of\ views_{video}}$$

This PI value was defined as the metric to use in defining a popular video, to answer the second research question. A high PI indicated a high level of engagement by the viewer, normalized by the number of views. A video with a high PI was a video which advertisers could feel confident that their advertisement would reach an engaged audience. I examined PI by year and month to search for any cyclical trends and found none. The PI by month is shown below, left. Further descriptive analysis shown below plot PI with respect to views is below right on the next page.

Brian Taylor
SUID# 251968713
rtaylo11@syr.edu



To develop a predictive model, K means clustering was used on the dataset. The elbow method was used to determine K = 4, and the results of the clustering as well as the elbow are shown below.



A decision tree model was used to develop a predictive model. The predictive model used binned values of PI to predict a class of popular videos, Class 0 to Class 3, where Class 3 had the highest PI values. The decision tree model confusion matrix and summary statistics are below, left. The accuracy of 79.98% with a 95% confidence interval of (0.7705, 0.8268) and a Kappa value of 0.7234 was the best performing model of the project. A K nearest neighbors model was also created with an accuracy of 78.4%. The confusion matrix for the KNN model is below, right on the next page.

Brian Taylor
SUID# 251968713
rtaylo11@syr.edu

```
## Confusion Matrix and Statistics
##           Reference
## Prediction   0   1   2   3
##          0 203  33   9   3
##          1  23 221  30  24
##          2   0  10  69  11
##          3   0   6  13 154
## Overall Statistics
##                Accuracy : 0.7998
##                  95% CI : (0.7705, 0.8268)
##     No Information Rate : 0.3337
##     P-Value [Acc > NIR] : < 2.2e-16
##                   Kappa : 0.7234
##  Mcnemar's Test P-Value : 4.813e-06
## Statistics by Class:
##                      Class: 0 Class: 1 Class: 2 Class: 3
## Sensitivity            0.8982   0.8185  0.57025   0.8021
## Specificity            0.9228   0.8571  0.96948   0.9692
## Pos Pred Value         0.8185   0.7416  0.76667   0.8902
## Neg Pred Value         0.9590   0.9041  0.92768   0.9403
## Prevalence             0.2794   0.3337  0.14957   0.2373
## Detection Rate         0.2509   0.2732  0.08529   0.1904
## Detection Prevalence   0.3066   0.3684  0.11125   0.2138
## Balanced Accuracy      0.9105   0.8378  0.76986   0.8856
```

```
##     test_pop
## knn1   0   1   2   3
##    0 279  68   0   0
##    1  49 296  29   1
##    2   0  28  46  18
##    3   0   1  22 163
## [1] 0.784
```

Other models were developed and tested including random forest, Naïve Bayes and Support Vector Machine models utilizing a polynomial kernel as well as a radial kernel. These other models had varying degrees of accuracy and the decision tree and the K nearest neighbors were the highest performing models with accuracy of ~80% and ~78%, respectively.

Results

With respect to the original data questions, I was able to identify key factors that contribute to popular videos as well as developed a predictive model that was 80% accurate. Those key factors for popularity were primarily the category of content and surprisingly, the channel posting the video.

The resultant model could prove useful for directing advertising traffic for certain channels as well as assist YouTube in determining advertising rates on certain channels as well as categories of videos.

Conclusion

Introduction to Data Mining by Tan et al. (2005) provided the details and reference material to better understand classification, cluster analysis and decision trees, which were absolutely necessary for this project. In my professional capacity, I had worked with cluster analysis but lacked the academic underpinnings to fully understand the work I was required to review.

This project was my first exposure to JSON files and associated transformation techniques. This project was also one of the first times I started with data in an unfamiliar format and an idea, and ended up providing descriptive statistics, visualization and multiple predictive models in an attempt to find the most accurate model. It proved to be a significant challenge at the time, but upon reflection, I felt proud that this project was the first time I was able to pull almost all aspects of data science together into one, cohesive project.

Brian Taylor
SUID# 251968713
rtaylo11@syr.edu

This is an important project to me in order to highlight the basics of the work which I expect to conduct in my next career as a data scientist. It is a reminder of how far I have come from being limited to just using the Office suite of programs to utilizing open source platforms and languages to custom code scripts to investigate specific questions and turn data into knowledge and hopefully, informed decisions. It is also a reminder that despite not knowing how to accomplish a task, with enough research and study, I can learn how to manipulate data, regardless of the structure and format. Further, as JSON data is prolific, my confidence in working with that format increased greatly.

Demonstrated Link (2, 3, 4, 5)
By utilizing a composite variable and further binning that variable into a category, I was able to develop an alternative strategy to working with the data.

Through data visualization, I was able to identify the popularly viewed categories of videos as well as find the elbow for clustering. By examining the summary statistics of the decision tree model, I was able to identify the performance of model with respect to accuracy and Kappa values.

The ingestion and manipulation of JSON files was a novel experience at the time, and time consuming, but demonstrated my ability to collect and organize data in a new format.

As the questions to answer for this project had business applications in the field of advertising, the development of a model as part of an advertising plan is integral to implementation of business decisions derived from analysis.

Project 3 – IST 652 - U.S. National Disasters over the last 64 years

Introduction – problem
This project focused on the history of natural disasters using a dataset that contained every federal emergency and disaster from 1953 to 2017. Data attributes included the duration, location, time, duration, and type of the emergency or disaster as well as available assistance programs. The intent was to answer specific questions to better inform national emergency response coordinators about historical disaster events. This information was intended to be used to better prepare for the type and frequency of disasters by state.
The analysis questions to be answered were:

- What declaration type is most common?
- What state has the highest number of declarations?
- What are the major disasters in the top five disaster prone states?
- Which regions of the United States have had the greatest number of hazard events declarations over the last 64, 7 and 3-year periods?
    - Are there regions within the United States which have dramatically less counts?
        - If so, where are they?

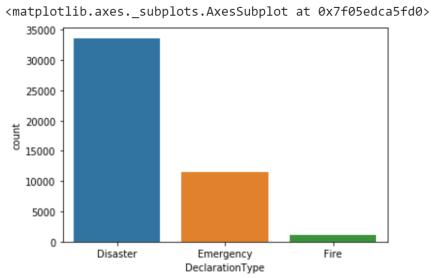
Brian Taylor
SUID# 251968713
rtaylo11@syr.edu

- Are there any visual correlations among these maps individually and/or in relation to each other?
  - If so, what are they?

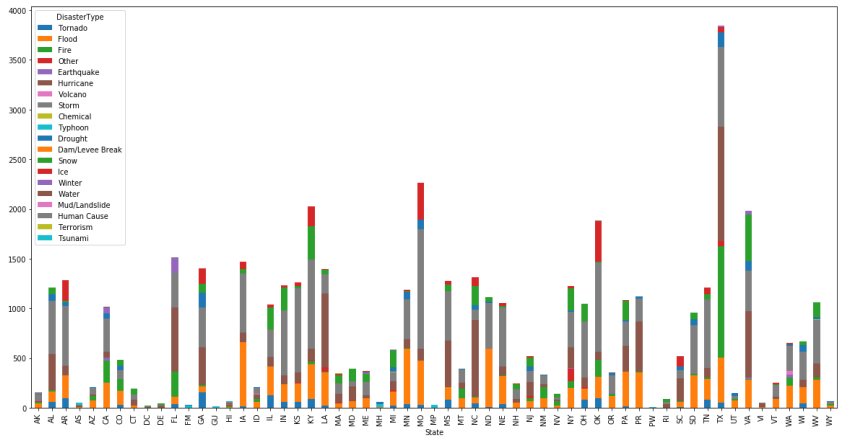
Methods - tools & techniques used

Common data munging was performed on the Kaggle dataset (https://www.kaggle.com/fema/federal-disasters) was performed to change data set column names, change the type of data in the data frame such as to datetime or categorical. The primary method of analysis was data visualization. DeclarationType and DisasterType, categorical attributes, were primarily used for the visualizations, as well as location information were used to analyze what types of disasters, how many and where they occurred.

To answer the first question, “What declaration type is most common?”, the following chart was created.

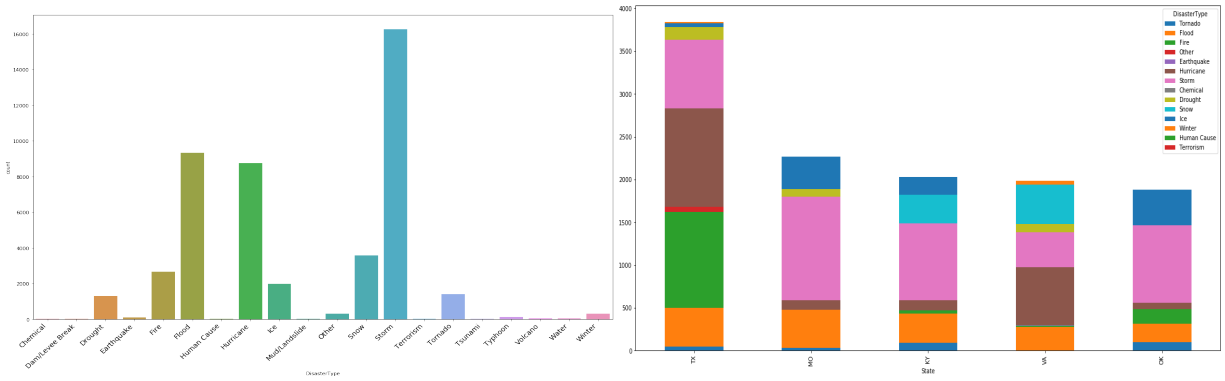


The following chart displays the declarations by state and answers “What state has the highest number of declarations?”. Texas has the greatest number of declarations.

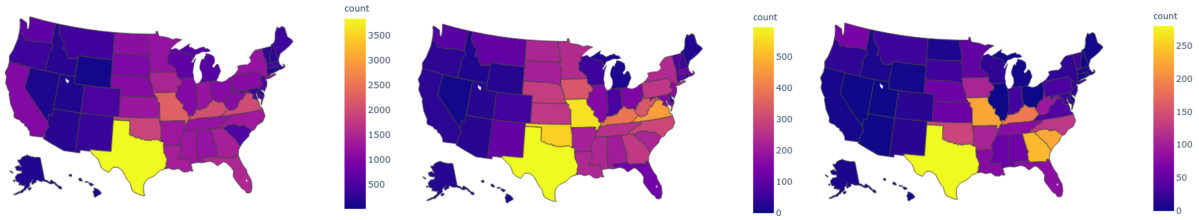


The charts below (on the next page) show the major disasters in the top 5 most disaster-prone states of Texas, Missouri, Kentucky, Virginia and Oklahoma.

Brian Taylor
SUID# 251968713
rtaylo11@syr.edu



And a comparison of 64, 7- and 3-year hazard events by state are shown below, from left to right.



Results - level of knowledge
This project focused primarily on an examination of historical data to identify states that are affected by natural disasters. I used graphical techniques to display the historical data to show where and what those disasters are and possibly expose some trends in the frequency of natural disasters.

Surprisingly, Texas was shown to have the most disaster declarations despite an initial assumption that California might have the most based on news reporting on wildfires. This analysis helped dispel the myth and showed that the middle (latitudinally) of the relatively wet portion of the country, and a large state that is a common target for hurricanes, should be the focus areas for frequency of water related disasters.

This information could be used to preposition disaster response and relief resources specific to the state or region's proclivity for certain disasters.
It would behoove Texans to prepare for hurricanes, floods, storms and fires. Missourians would do well to prepare for storms, tornados and floods. Kentuckians need to prepare for those same events, but also consider snow. Virginia is also affected by hurricanes and snow, while Oklahoma's primary disaster makers are tornados and ice. All of this information is actionable for preparing personal and government resources, for public service announcements and could be used to preposition disaster response and relief resources specific to the state or region's proclivity for certain disasters.

Brian Taylor
SUID# 251968713
rtaylo11@syr.edu

Conclusion - insight gained and how. How this links to work outside, demonstrates present strength and challenge, leverage into life-long learning, reference to key conceptual works and how that influenced *Python for Everybody* (PY4E - Python for Everybody, 2021), is a staple in any new coder's library of resources. It was a staple for the last three years as I constantly switched between programing languages depending on the course and remains my go to resource when I forget some of the basics.

Since before this course, I have believed that a well-made data visualization could be more concise and provide broader and deeper understanding of a topic vice just written words. I am a visual learner and this was my first project where data visualization was used as the primary method to relay answers to study questions. The experience on this project steered me towards taking a course on data visualization and leading me down a path focused on data visualization as a primary means of helping decision makers make better decisions.

Demonstrated Link (2, 3, 5, 6)
Kaggle provided the comma separated value file for this project. By importing, transforming and filtering, I was able to collect and organize the data into a usable form for the graphics used in this project.

This entire project depended on the use of data visualization to explore the data. Through multiple graphics, I was able to show where, what and how often a disaster causes a declaration of a disaster.

The goal of the project was to inform disaster response planning professionals of where and what types of disasters they should be ready to respond to. These professionals are responsible for managing constrained assets during a disaster response. This analysis should be viewed as the first step in determining the allocation of resources for both prevention and response. Many plans could be developed from this basic knowledge from this project.

Project 4 – IST 736 Text Mining for the Ultimate Fantasy Football Team

Introduction

Fantasy Football has become big business. Although many times friends compete for bragging rights, weekly leagues and legacy leagues exist where collectively large amounts of money changes hands based on a team manager's ability to choose their players wisely either for just the week or multiple seasons.

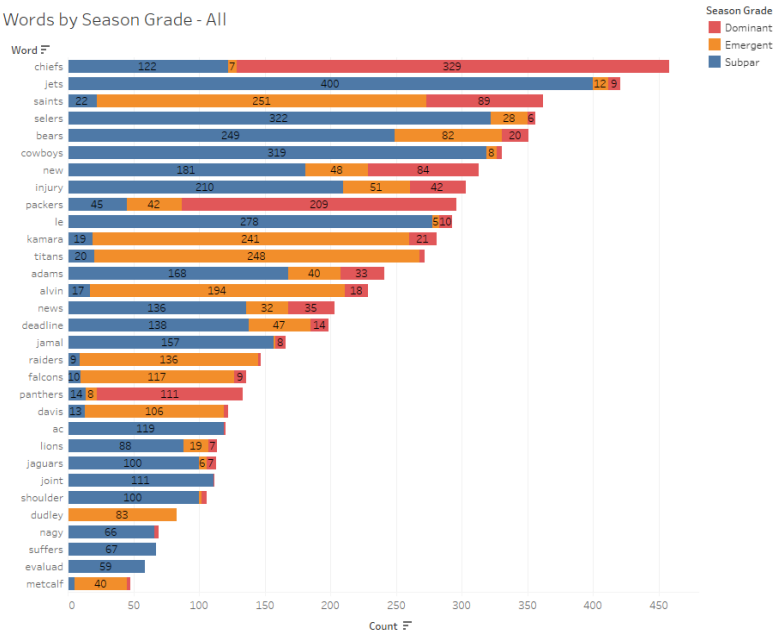
The initial idea for this project was to text mine tweets in an attempt to identify injuries or emerging players over the course of a season in the course of compiling a wining week-to-week team. After further initial data analysis, I realized that accomplishing that task was not possible given the content of the tweets. I pivoted and chose to use tweets to identify the name of the player, their position, and one of three categories of season performance. The positions of

Brian Taylor
SUID# 251968713
rtaylo11@syr.edu

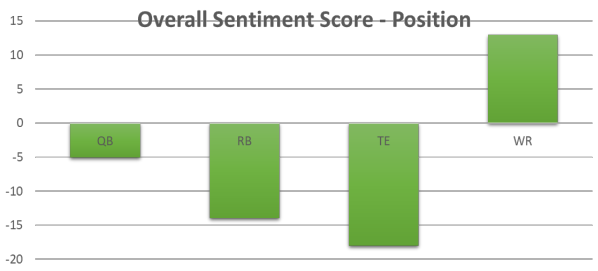
Quarterback, Wider Receiver, Tight End and Running Back were the position categories. Subpar, Emergent and Dominant were the categories for season performance and a total of 26 NFL players were chosen with 6 to 7 players in each position.

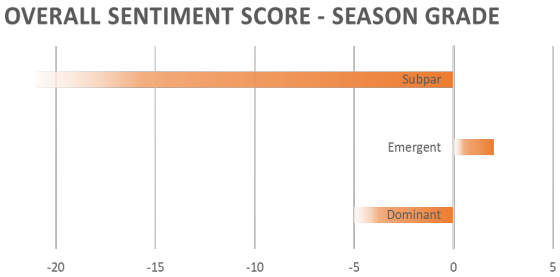
Methods

Sentiment analysis was performed on over 12,000 tweets after removing stop words, emojis, punctuation and other non-informative characters. A frequency of words identifying the season performance are shown below.



Sentiment analysis for each position and season performance are shown below.





I then build five models to predict the three identified attributes including Bernoulli Naïve Bayes, multinomial Naïve Bayes, multinomial Naïve Bayes with Term Frequency Inverse Document Frequency (TF-IDF), linear and polynomial Support Vector Machine. Further, I conducted latent Dirichlet allocation (LDA) for 10, 7, 4 and 3 clusters.

Brian Taylor
SUID# 251968713
rtaylo11@syr.edu

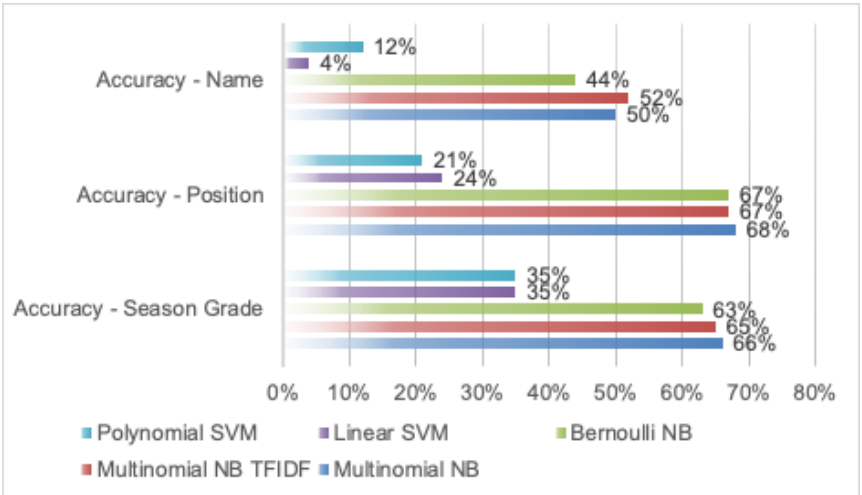
Results

The pivot in focus is explained by the LDA analysis in that the topics identified were just too broad to make predicting player success on a fantasy football team realistic. The resultant clusters are shown below.

Topic #0	Topic #1	Topic #2
trade	yards	week
like	nfl	pts
just	td	te
jets	week	nfl
le	patriots	saints
steelers	fantasy	cowboys
injury	game	wr
better	good	kamara
game	gt	brees
nfl	vs	johnson

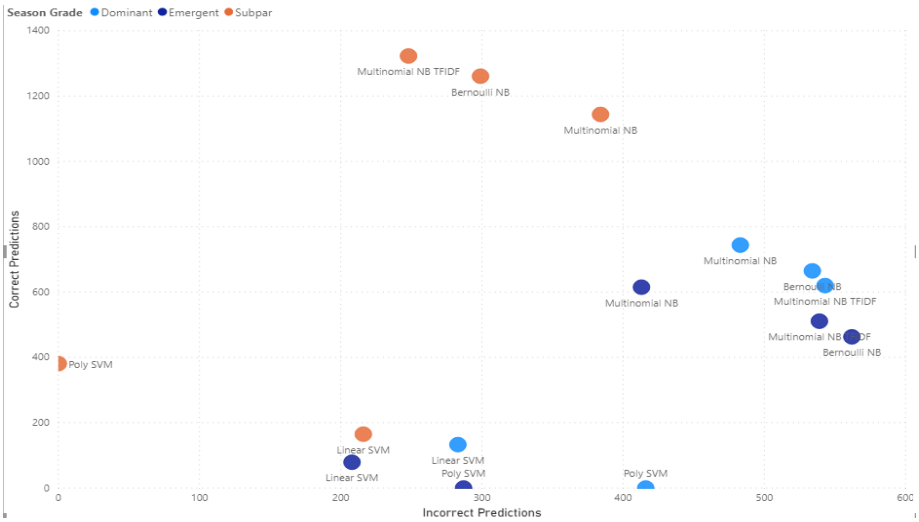
Topic 0 appears to be about injuries and how that affects games, Topic 1 seems to be the fantasy football/predictors topic, and the last one, Topic 3 seems to be about the Saints and some of their games they played. None of this information was helpful in determining fantasy football composition.

A comparison of the NB and SVM models on the three attributes of player name, season grade and position showed that the NB models handily outperformed the SVM models. A comparison is shown below.



A breakout of the model performance specifically on season grade is shown below on the next page.

Brian Taylor
SUID# 251968713
rtaylo11@syr.edu



The polynomial SVM model was an interesting observation in that the performance for subpar season grades was perfect, at the expense of the other two season grade categories. If I had to choose one model for all three attributes to attempt to perform better than a coin flip, I would have chosen the multinomial Naïve Bayes model. Text mining was able to produce identify three different categories of performance for a player, their name, and their positions. Although not perfect, the models achieved useful results.

Conclusion

Predictive text mining is a difficult field of study within data science. *Fundamental of Predictive Text* Mining by Weiss et al. (2010) was a strong resource for this project as well as the class. I had minimal understanding of text mining at the onset of the class. Achieving the goals in the final project were daunting. Pivoting from the initial intent of the project to predicting categories in certain attributes was a difficult choice, but I realized that the goal was a bit too challenging given the situation.

I learned that data used for training models should be balanced in all aspects possible. The season grade category was not balanced for this project and likely affected model performance. Additionally, thesis do not always pan out. Discussion of injuries on Twitter did not product the topic clusters I was expecting. Text mining in a particular field likely demand a specific sentiment analyzer to achieve better accuracy.

Demonstrated Link (1, 2, 4)

Text mining is an increasingly important topic within data science and technology in general. Finding insights in the massive amount of written word data can be very difficult, but rewarding endeavor.

Cleaning text data is time consuming and requires subject matter expertise to perform well. Compare to other data formats, text is dirty.

Brian Taylor
SUID# 251968713
rtaylo11@syr.edu

Finally, changing a course of action is a difficult decision in the middle of analysis, but highlights the importance or exploratory data analysis in preparing a plan of action to gain knowledge from analysis.

Project 5 – IST 719 Crime Visualization

Introduction

This data visualization project explored crime in a metropolitan city, specifically, Raleigh, North Carolina. The premise of the project is that when people move, one of the questions often asked is about crime in a particular area. As I was preparing to move my family, I chose to research the crime in the city to where I was moving.

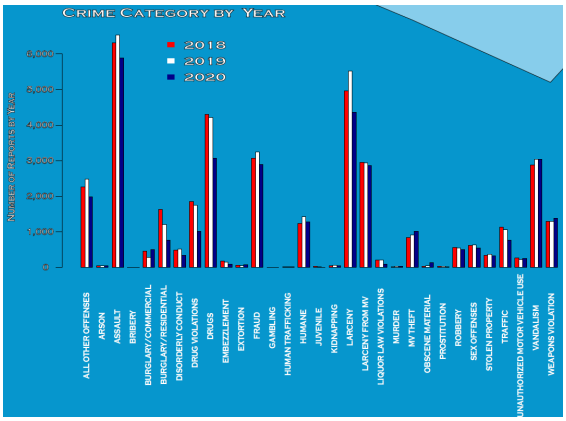
Further, local town municipalities could use this information for funding allocation, the Chamber of Commerce could use the analysis to propose infrastructure improvements and outreach programs could focus on specific areas to reduce certain types of crime in a particular area.

The three major data analysis questions were simply: Where? What? and When? Data was collected from https://data.wakegov.com/ which had attributes such as such as crime category, crime description, block address, district, date, time and location in latitude and longitude.

Methods

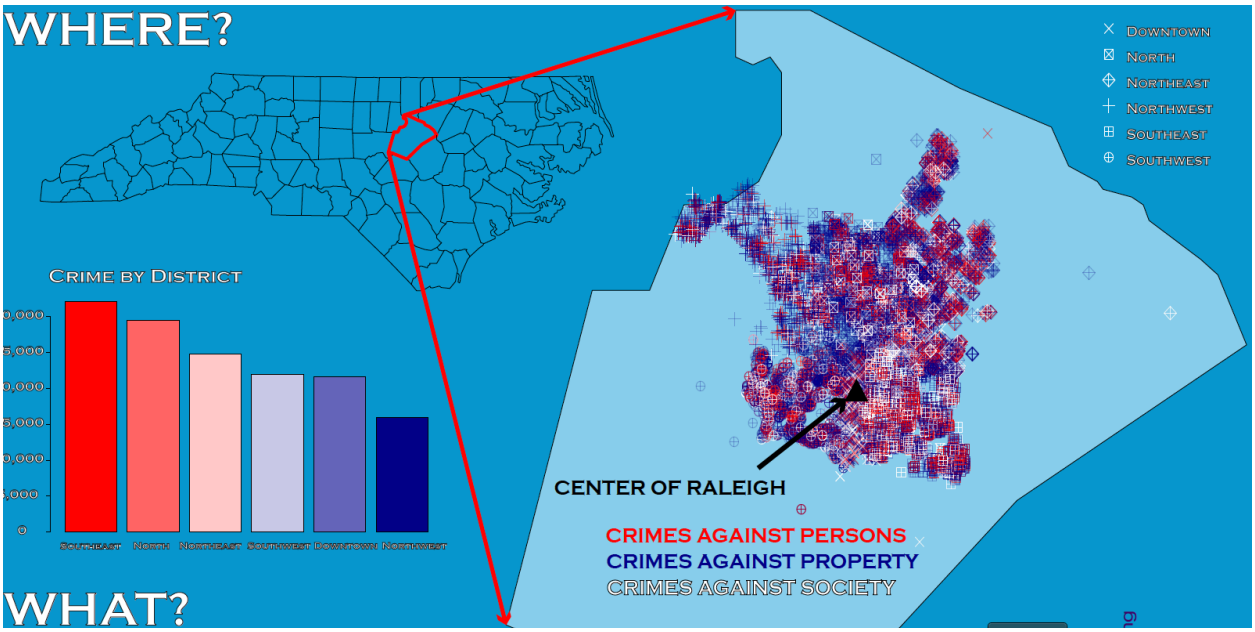
After obtaining the comma separated value file from the website above, I munged the data and filtered out all years except between 2017 and 2020. For the georeferenced graphic, only 2020 data was used for clarity of the graphic and relevance to conveying the results.

Basic R graphics were used, primarily bar plots, to convey the number of reported crimes in each district, the type of crime over time by category, reported crime by the time of day and the day of the week. The plots were then manipulated in Adobe Illustrator and positioned on the poster. The crime by type and year graphic is shown below.

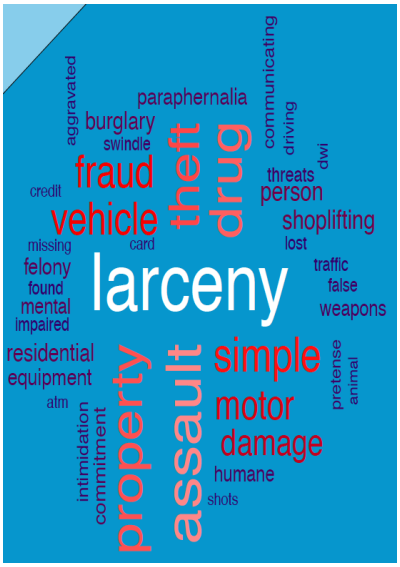


Brian Taylor
SUID# 251968713
rtaylo11@syr.edu

Multiple map packages were used for the georeferenced graphic. The shape of the crime icon pertains to the location, specifically the district, where the crime was reported. The color of the crime icon represented a category if crime such as a crime against a person, or against property, or against society. This graphic is below.



Finally, I utilized multiple packages, such as wordcloud, to develop a word cloud of the top 40 words used in the crime description. I first removed common stop words and punctuation in the corpus and assigned a color palette to the words so that the most frequent word was the lightest color. The word cloud is below.



Brian Taylor
SUID# 251968713
rtaylo11@syr.edu

Results

The results of the project were able to inform any interested party about the time of day, day of week, type of crime, location of crime, crime category, and in which district crime occurred with minimal use of words. The interested party would know where Raleigh is located in North Carolina and even the shape of the county where Raleigh is located.

Depending on the interests of the client, they could focus attention to certain aspects such as their own district or the popular times for crime to occur and tailor crime prevention or response efforts to those areas at those times. In the case of a future resident of the suburbs of Raleigh, that person could weigh the costs and benefits of residing in a particular location in or around the city.

In general, there is more crime in the southeast district than the northwest. Most crime, with the exception of motor vehicle theft, vandalism and weapons violations have decreased in 2020 compared to 2018 and 2019. Larceny and assault are the two most common crimes. Finally, crimes around Raleigh tend to occur during normal working hours, dinner time, and during the work week.

Conclusion

This project was one of the more enjoyable and challenging efforts of the program. Constructing a poster sized data visualization is something which I had done before, but that was with paper, scissors, glue and markers. Constructing the entire project on a computer to the quality level required was quite the challenge.

Yau's (2011) *Visualize this: The Flowing Data Guide to Design, Visualization, and Statistics* greatly increased my understanding of the art of visualization. I relied on that resource heavily when designing the layout of my poster.

Creating a good visualization is not an easy task. Many decisions are made regarding layout, color palette, what style of font, the size of images and more. Although I have been creating slide shows for work for the past 20 years which integrated graphics, there were always business rules which drove or constrained them. This project allowed me to work outside of the rules I had been following, and the understanding of how to use the tools allowed me to produce a high resolution, quality, digital poster.

A good poster can tell its own story to anyone who views it. They can convey so much information in a way the viewer is free to interpret and even generate different opinions or hypotheses pertaining to the data displayed. A good visualization gets the viewer as involved as they want to be, which can drive intricate and nuanced conversations that can drive the next project. There is a reason academia has used poster presentations to interact with students and faculty for so long.

Brian Taylor
SUID# 251968713
rtaylo11@syr.edu

Demonstrated Link (1, 2, 3, 5, 6, 7)

Data visualization allows the use many data science practice areas. Predictive analytics, text mining, big data analysis, and descriptive statistics results can all be displayed graphically, often aiding the understanding of a topic intuitively.

Although publicly available, the data set from the Raleigh Police Department contained some attributes that either didn't pertain to this study or were redundant. Filtering those attributes out of the dataset reduced its size. Further, not every report contained information for all attributes and null values were present.

Data visualization was used heavily in this project for analysis of the topic. Six main charts were used to convey information on three years of reported crime in the capital city of the state.

The initial plan of action was for use in the decision-making process of moving to an area. The application of this information, although not explicit, allows a potential relocator to determine where they would prefer to establish residency if crime avoidance is a priority. As a city manager or council member, this information could be used to support or refute requests for limited resources in certain districts in order to spend more wisely.

"A picture is worth a thousand words" is a time worn cliché, but true. Depending on a viewer's professional or personal background, they take away a somewhat tailored understanding of the topic, depending on the level of engagement. Some of the art of a poster presentation is that font size and attention-grabbing graphics need to be considered based on the proximity of the viewer.

Data ethics applies to the use of data. Using data to improve a situation is ideal. Using data for malice is not. At least two viewpoints were demonstrated in this project: the person relocating to the area, and a government official responsible for the allocation of limited resources. While one viewer is looking to avoid trouble, the other is looking to find it and apply resources to reduce it. In the end, the two viewers want the same result, a safer community. Using data to improve lives is an ethically good purpose.

Portfolio Conclusion

This portfolio is my representation of five projects performed for the Applied Data Science graduate program which best demonstrate my understanding and performance of the following tasks:

1. Describe a broad overview of the major practice areas in data science.
2. Collect and organize data.
3. Identify patterns in data via visualization, statistical analysis, and data mining.
4. Develop alternative strategies based on the data.
5. Develop a plan of action to implement the business decisions derived from the analysis.

Brian Taylor
SUID# 251968713
rtaylo11@syr.edu

6. Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization.
7. Synthesize the ethical dimensions of data science practice (e.g., privacy).

In the first project with youth basketball players in an athletic club, I created a database from scratch and performed queries of that database in order to answer some basic business question that the club might have covering player performance, retention and ensuring a relatively fair level of competition between teams. In dealing with personal information, I learned the importance of safeguarding sensitive information to protect the privacy of those included in the study.

When investigating a business question related to marketing, the focus of the plan of action became identifying trending or popular YouTube videos to maximize exposure to an engaged audience. I developed an attribute which combined multiple attributes from the videos as a way to gauge popularity for the marketing team to implement. Through modeling, I was able to achieve high accuracy in predicting popular videos and the characteristics of popular videos.

Performing data visualization on disaster response data was my first deep foray into the data visualization field of study. Evaluating historical data to provide insights for disaster planners can affect the plans of those important professionals. Being able to explain what the visuals mean and propose plausible reasons for the frequency of disasters in certain areas by type can help tailor prevent and response packages to be as responsive as possible and preserve as much life and property as possible.

Pulling apart text for meaning and application proved challenging, but was a blessing in disguise. The identification of the importance of exploratory data analysis is a fundamental I will take with me for the rest of my professional life. I now view adaptation to what the data is exposing, rather than plunging blindly ahead into possibly wasted analysis, is a part of the art of data science.

Fully incorporating visualization and the ability to convey a meaningful story to a broad span of professionals was a challenging and rewarding task. Pictures can transcend language barriers and tell all of the stories of the many different data analysis techniques studied in this program and can greatly assist in the ultimate goal of data science, by turning data into information and knowledge which can be applied for decision making and hopefully improvement of every area of focus.

Brian Taylor
SUID# 251968713
rtaylo11@syr.edu

References:

Hoffer, J. A., Ramesh, V., & Topi, H. (2019). *Modern Database Management*. Pearson Education.

PY4E - Python for Everybody. (2021). PY4E. https://www.py4e.com/book

Tan, P., Steinbach, M., & Kumar, V. (2005). *Introduction to Data Mining* (1st ed.). Pearson.

Weiss, S. M., Indurkhya, N., & Zhang, T. (2010). *Fundamentals of Predictive Text Mining (Texts in Computer Science)* (2010th ed.). Springer.

Yau, N. (2011). *Visualize This: The FlowingData Guide to Design, Visualization, and Statistics* (1st ed.). Wiley.