2020

Portfolio Milestone

PARIN PATEL

SUID: 252200445

GITHUB LINK: https://github.com/parinsights/portfolio-

Table of Contents

Introduction:

Built upon a curriculum that strongly emphasizes the direct applications of the fundamental and diverse data science principles and practices; the Applied Data Science program at Syracuse University provides students the opportunity to collect, manage, analyze, develop, and implement insights using data from a multitude of disciplines using various tools and techniques. The curriculum, through courses like Database Administration Concepts and Database Management (IST-657), Data Analytics (IST-707), Natural Language Processing (IST-664), and Quantitative Reasoning Data Science (IST-772) has provided the relevant and applicable knowledge, training, resources, and academic environment for its students to progress and impart skills that exceed the basic roles of a data scientist. Through reports and presentations, students of the School of Information Studies were able to acquire crucial experience in detecting patterns in data, developing and implementing alternative data approaches, demonstrating communication skills that focus on explaining technical outcomes to non-technical recipients, and discussing the ethical scope of data science through courses related to data privacy and policy.

Report Objective

The goal of this report is to provide sufficient evidence that the following seven learning objectives set forth by Syracuse University have been met:

1. Describe a broad overview of the major practice areas of 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 analyses.
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).

Objective 1: Describe a broad overview of the major practice areas of data science.

With skillsets in computational and inferential reasoning, data scientists are analytical experts who draw conclusions from large sets of structured and unstructured data. They utilize their industry knowledge and ability to analyze, process, model, and interpret the data to uncover solutions and crate actionable plans for organizations and companies.

To do this, data scientist are trained in skillsets far beyond just descriptive statistics, a methodology that requires analysis of past trends to understand what has happened. Instead, a data scientist must be able to also predict future trends and communicate their findings in a visually informative and understandable way. Through predictive analytics, a data scientist will use various statistical methods like forecasting, predictive modeling, machine learning, and data mining to make predictions about the future or an unknown variable. Predictive modeling most commonly will includes the use of classification models, clustering models, forecasting models, outlier models, or time series models. These models can involve the use of a variety of algorithms, based on the type of data presented and the objective of the analysis. Classification models seek to classify information into a number of classes to predict the class or labels of additional, new data. Classification algorithms include, but are not limited to logistic regression, Naive Bayes, K-Nearest Neighbor, Decision Trees, Random First, and Support Vector Machine. Clustering models seek to cluster objects into groups that either show much more similar data in the same are than those placed in another group. Common clustering algorithms include, but are not limited to K-means algorithm, fuzzy c-means (FCM) algorithm, expectation-maximization algorithms, and hierarchical clustering algorithms. Forecasting models use historical data to make informed predictions about the direction of future trends. Some common forecasting techniques include moving average, linear regression, multiple linear regression, stochastic analysis, and time series analysis. Outlier modeling helps to identify anomalies in the data, which can be a serious issue when training a machine learning algorithm or applying some statistical methodology. In addition, outlier models can give us information about localized anomalies found within the entire system. Some common methods to identify outliers are Z-Score, DBSCAN, and Isolation Forest. Finally, time series analysis encompasses methods for analyzing time series data in order to forecast future values based on the past observed ones. A few common time series forecasting is autoregression(AR), moving average (MA), autoregressive moving average (ARMA), autoregressive integrated moving average (ARIMA), and simple exponential smoothing (SES).

While the courses in the iSchool’s Master’s in Data Science curriculum taught a variety of disciplines within the data science field; all of the courses within the Primary Core Coursework established that students had a clear understanding of the core fundamentals and skillsets that would help them understand more advanced data techniques and methodologies. These courses were Introduction to Data science, Data Analytics, Data Administration Concepts and Database Administration, Data Analysis and Decision Making, Business Analytics, and Big Data Analytics. From these courses, students were able to gain a high-level of skillsets that would allow them to advance to more complex data concepts.

For example, in courses like Introduction to Data Science projects were completed to show the ability to manipulate data, connect to external data sources and mine data, visualize information, create visually appealing maps, and model data using linear modeling and association rule mining and support vector machines. These skillsets were shown through the final project that used Major League Baseball data to ultimately predict the baseball stats of a player most likely to be inducted into the Baseball Hall of Fame, based on the past data of players already inducted. This prediction utilized logistic regression models, random forest modeling, and linear modeling to predict the likelihood of players with certain statistics becoming hall of famers.

Click here to see the associated documents for this project.

Sentiment analysis reviews text data for indications of positivity, negativity, or some other form of sentiment. This value is then measured as a rating between negative one (-1) and positive one (+1). Natural language processing (NLP) is a methodology that allows researchers to analyze the sentence and language-related data using computational techniques to determine specific outcomes. It is one of the more advanced techniques taught after core fundamentals. Sentiment analysis, predictive analysis, and descriptive analysis were utilized in the final review of President Trumps Tweets. In addition, this project required mining twitter, presidential approval ratings, and historical stock market data. The goal of this project was to utilize all the skills we had learned during our tenure as students of the iSchool in order to evaluate Presidents Trumps tweets for trends using current data science techniques including NLP, Word Clouds and Sentiment analysis. In addition, we compared his sentiment analysis score of tweets to his approval ratings to see if the language he used in his tweets has an effect on the opinions of the public. Finally, we ran a comparison of his combined approval-sentiment score against the stock markets historical data for 2017 to see if the market was provoked by his tweets.

Click here to see the associated documents for this project.

Background:

All of the chosen projects show a case example of the major practice areas of Data Science. From data collection, mining, and organization to identifying patterns and developing alternative analytical strategies, to implementing business decisions and communicating, to finally managing the ethical responsibilities that arise with data management and analysis.

Objective 2: Collect and Organize Data

Learning Objective Status: Mastered 🎓🎓🎓🎓🎓

Background:

Almost all of the projects and assignments completed in the iSchool Masters in Data Science curriculum have required mastery of skills in data collection and cleaning. For example, below are two examples of my work in collecting and organizing data from first my Data Administration Concepts and Database Management course (IST-659), and secondly my Scripting for Data Analytics course (IST-652).

Reference of Student Work: Part 1

In the first course (IST-659), I successfully demonstrated skills in data management and organization by taking collected information in the form of concert venue data, musician booking information, and inventory reports to build a SQL database using Microsoft Access. The goal of this project was to increase the ease and efficiency of booking musicians by venue-owners, while simultaneously managing a website that sells related-merchandise. We started by creating a data dictionary and then established the some of the following data questions:

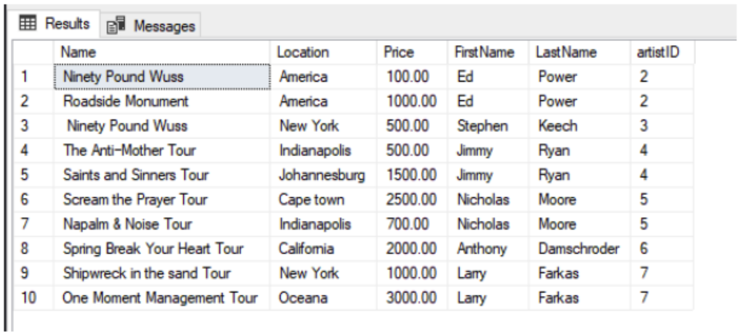
Data Questions:

1. What are the artists available to book, what is their musical genres, current location, and price per performance?
2. What artists have posters Milk Boy Art House can buy and sell? What is the quantity available, price per poster, and URL to the poster's link.
3. What are the music sales of artists, and the date they published their songs. Also wish to include their label's name and genre.
4. List of available performances ordered by the price of said performance and the type of performance given by artist.
5. List of songs with the artist and its published date.

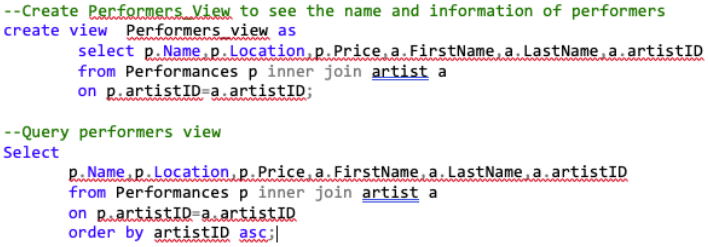
Tools & Techniques Used:

Our final synopsis required us to first map our database organization out using an entity relationship diagram, a logical model diagram, and a normalized model diagram. We then used SQL to implement and create our database by creating the tables and establishing their data types. We then used SQL to start inserting our data into the organized data tables. Finally, the last part of the project involved writing advanced queries to extract data from our organized data tables. You can see in Figure 1 the answer to our first data question, in addition to a query that will allow you to see the Performance data table sorted by artistID. The second figure (Figure 2) will show the front-end of the organized database that shows the associated revenue and costs for the venue owner.

1. What are the artists available to book, what is their musical genres, current location, and price per performance?

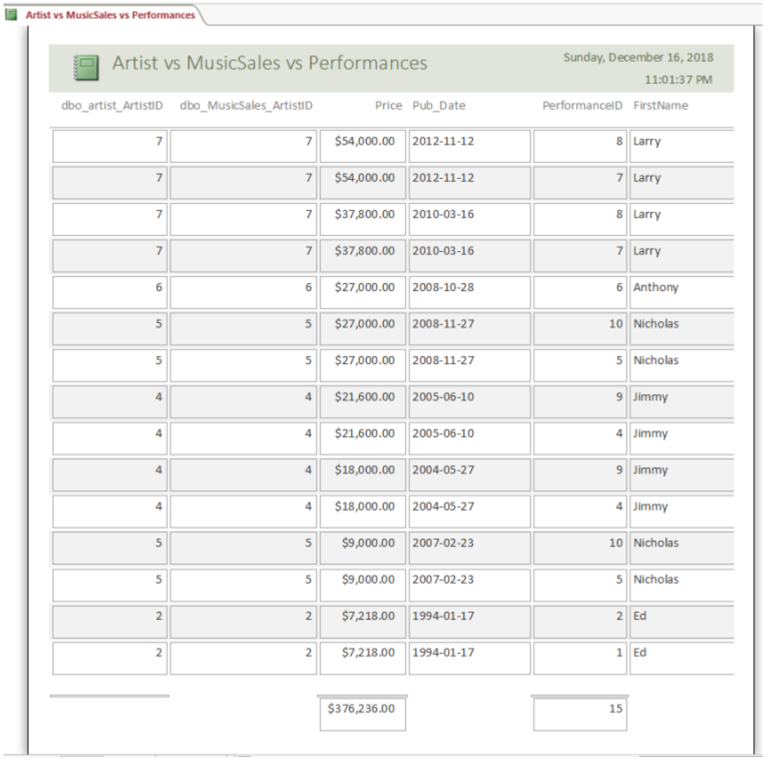


	Name	Location	Price	FirstName	LastName	artistID
1	Ninety Pound Wuss	America	100.00	Ed	Power	2
2	Roadside Monument	America	1000.00	Ed	Power	2
3	Ninety Pound Wuss	New York	500.00	Stephen	Keech	3
4	The Anti-Mother Tour	Indianapolis	500.00	Jimmy	Ryan	4
5	Saints and Sinners Tour	Johannesburg	1500.00	Jimmy	Ryan	4
6	Scream the Prayer Tour	Cape town	2500.00	Nicholas	Moore	5
7	Napalm & Noise Tour	Indianapolis	700.00	Nicholas	Moore	5
8	Spring Break Your Heart Tour	California	2000.00	Anthony	Damschroder	6
9	Shipwreck in the sand Tour	New York	1000.00	Larry	Farkas	7
10	One Moment Management Tour	Oceana	3000.00	Larry	Farkas	7



```
--Create Performers_View to see the name and information of performers
create view  Performers_view as
        select p.Name,p.Location,p.Price,a.FirstName,a.LastName,a.artistID
        from Performances p inner join artist a
        on p.artistID=a.artistID;

--Query performers view
Select
       p.Name,p.Location,p.Price,a.FirstName,a.LastName,a.artistID
       from Performances p inner join artist a
       on p.artistID=a.artistID
       order by artistID asc;
```

Figure 1: Data Organization



Artist vs MusicSales vs Performances

Sunday, December 16, 2018
11:01:37 PM

dbo_artist_ArtistID	dbo_MusicSales_ArtistID	Price	Pub_Date	PerformanceID	FirstName
7	7	$54,000.00	2012-11-12	8	Larry
7	7	$54,000.00	2012-11-12	7	Larry
7	7	$37,800.00	2010-03-16	8	Larry
7	7	$37,800.00	2010-03-16	7	Larry
6	6	$27,000.00	2008-10-28	6	Anthony
5	5	$27,000.00	2008-11-27	10	Nicholas
5	5	$27,000.00	2008-11-27	5	Nicholas
4	4	$21,600.00	2005-06-10	9	Jimmy
4	4	$21,600.00	2005-06-10	4	Jimmy
4	4	$18,000.00	2004-05-27	9	Jimmy
4	4	$18,000.00	2004-05-27	4	Jimmy
5	5	$9,000.00	2007-02-23	10	Nicholas
5	5	$9,000.00	2007-02-23	5	Nicholas
2	2	$7,218.00	1994-01-17	2	Ed
2	2	$7,218.00	1994-01-17	1	Ed
		$376,236.00		15	

Figure 2: Organization , GUI Front-End of DataBase

Click here to see the associated documents for this project.

Reference of Student Work: Part 2

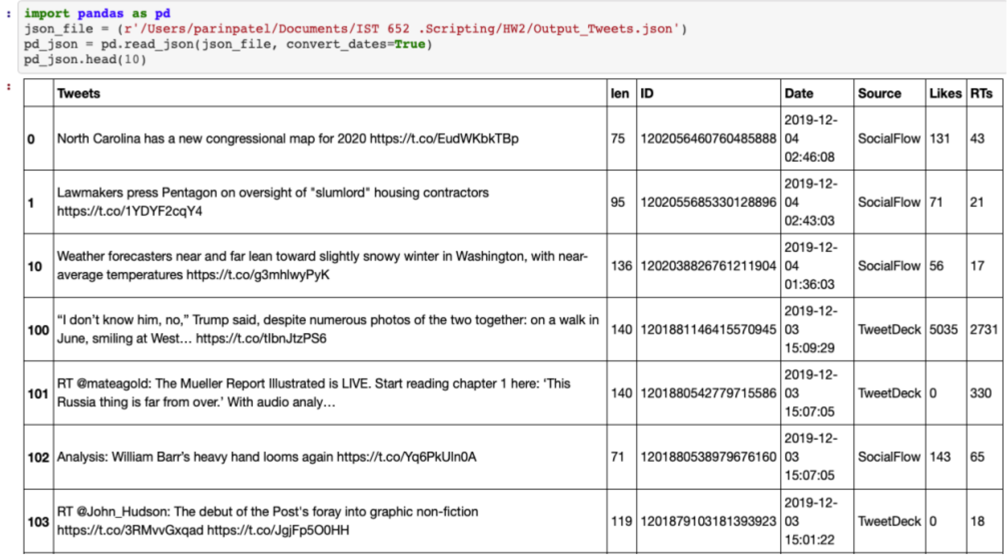
My second project related to data collection comes from my IST-652 course (Scripting for Data Analysis) where we had to write a program to that would extract, read, and explore headlines on Twitter. I chose a news source called the Washington Post, where I sought to review the headlines, they posted on Twitter to engage with users. These postings are usually free and are used to grab content users attention, and therefore, would have been a interesting source to mine and analyze.

Data Questions:

1. How will we extract this twitter information?
2. What is the average length of tweets?
3. Does the Post prefer to use a lot of characters in their tweets?
4. What is the average character count in tweets?
5. What are the most frequent words to appear?

Tools & Techniques Used:

To meet the objectives of this project we used python in Jupyter Notebook to conduct our analysis. We extracted the twitter data using a package called Tweepy. To do this, we were first required to get approval from Twitter under their "Twitter Developer" license agreement. Once we were approved, we were given a unique set of credentials' that would allow for us to extract the data. We extracted our tweets and then used the panda's data frame to process the data. After cleaning, by extracting the arrays we wanted ,we created our final working data frame. From here, we made sure to save this data frame as our "tweets output" as a JSON file, due to the data's sheer volume. As you can see in Figure 3, this JSON file was imported (due to a requirement of the assignment being to import a JSON file).



```
: import pandas as pd
  json_file = (r'/Users/parinpatel/Documents/IST 652 .Scripting/HW2/Output_Tweets.json')
  pd_json = pd.read_json(json_file, convert_dates=True)
  pd_json.head(10)
```

:

	Tweets	len	ID	Date	Source	Likes	RTs
0	North Carolina has a new congressional map for 2020 https://t.co/EudWKbkTBp	75	1202056460760485888	2019-12-04 02:46:08	SocialFlow	131	43
1	Lawmakers press Pentagon on oversight of "slumlord" housing contractors https://t.co/1YDYF2cqY4	95	1202055685330128896	2019-12-04 02:43:03	SocialFlow	71	21
10	Weather forecasters near and far lean toward slightly snowy winter in Washington, with near-average temperatures https://t.co/g3mhlwyPyK	136	1202038826761211904	2019-12-04 01:36:03	SocialFlow	56	17
100	"I don't know him, no," Trump said, despite numerous photos of the two together: on a walk in June, smiling at West… https://t.co/tlbnJtzPS6	140	1201881146415570945	2019-12-03 15:09:29	TweetDeck	5035	2731
101	RT @mateagold: The Mueller Report Illustrated is LIVE. Start reading chapter 1 here: 'This Russia thing is far from over.' With audio analy…	140	1201880542779715586	2019-12-03 15:07:05	TweetDeck	0	330
102	Analysis: William Barr's heavy hand looms again https://t.co/Yq6PkUln0A	71	1201880538979676160	2019-12-03 15:07:05	SocialFlow	143	65
103	RT @John_Hudson: The debut of the Post's foray into graphic non-fiction https://t.co/3RMvvGxqad https://t.co/JgjFp5O0HH	119	1201879103181393923	2019-12-03 15:01:22	TweetDeck	0	18

Figure 3: Collection of Tweets

Click here to see the associated documents for this project.

Conclusion of Learning Goal:

Based on the projects presented through this learning goal, I have not only demonstrated mastery of being able to collect and organize data, I have shown I can bring in complex forms of information and organize them in ways that increase meaning and usability. This is an important skill for data scientists because if you cannot increase the value of the raw information at the earliest stages of a project, then you will likely run into ethical issues down the line. This learning goal has allowed for me to take a skillset other research fields take for granted and bring value and expertise to the subject matter. In my professional work, I am tasks with creating methodologies for data collection and organization with our contractors. Being able to accurately demonstrate how quality data should be handled and maintained as it goes through the data pipeline has become a significant part of my work because I was able to show to my office the value in doing so. For example, ever since we implemented certain data standards for collection, we have seen a decrease in data quality errors.

Objective 3: Identify Patterns in Data Via: Visualization, Statistical Analysis, And Data Mining.

Learning Objective Status: Mastered 🎓🎓🎓🎓🎓

Background:

Almost all of the projects and assignments completed in the iSchool Masters in Data Science curriculum have required mastery of skills in visualization, statistical analysis, and data mining. For example, below is a summary of my final project for my Data Analytics course (IST- 707). In general, in IST-707, we were required to complete many projects and assignments that's displayed our knowledge of visualization, statistical analysis, and data mining. For example, our course assignments involved using all predictive models, discussed earlier in Objective 1 like Clustering and Classification; specifically, techniques like Support Vector Machines, Random Forest, Decision Trees, and Naive Bayes were utilized. In addition, each homework was accompanied by a report that was required to visually display our results and findings. In addition, all of the techniques learned during the course were then to be utilized to complete a final project, in my case, the final project was conducted to predict the success of Kickstarter Campaigns. We wanted to see how Kickstarter crowdfunding data could be used to predict the success of these campaigns in meeting their money goals.

Data Questions:

1. What demographics (if any) are most likely to succeed?
2. Which categories of product or service are more successful in funding their projects?
3. Which is the most critical factor affecting each campaign?

Tools & Techniques Used:

To meet the objective of our project we utilized R Studio and Weka to for our data mining, cleaning, preprocessing, analysis, and visualizations. The data was scraped from webrobots.io (https://webrobots.io/kickstarter-datasets/), combined, pre-processed, and created into a dataset that contains project data from 2014 to February 2019. The dataset contains 20 variables and over 192K records. Majority of the data describes various campaign characteristics such as campaign length, location, start and end dates, amount of money needed, and whether campaigns succeeded in being funded. We then created a data dictionary and decided to focus on identifying how various demographic attributes, start and end terms, monetary goals, or independent variables, affect campaign outcome, which represents the dependent variable. We then preprocessed our data to check for missing values, duplicates, data types, re-ordered the data columns, and consolidated the data to reduce the number of parameters in the predictive analysis.

This project required the use of various visualizations to demonstrate our results and findings. To do this, we utilized the ggplot2 package in R Studio. You can see in Figure 4 and 5, we started visualizations early in the report, with a descriptive analysis of our data. These two figures showcase the success and failure rates of the Kickstarter Campaigns by categories and project goals.

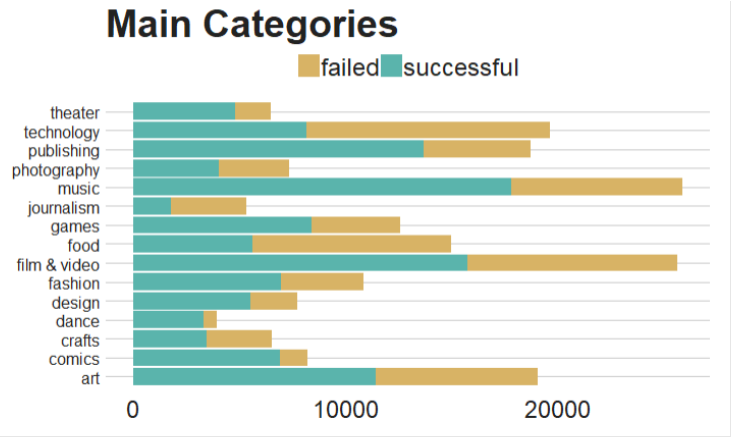


Figure 4: Main Kickstarter Categories Success vs Failure

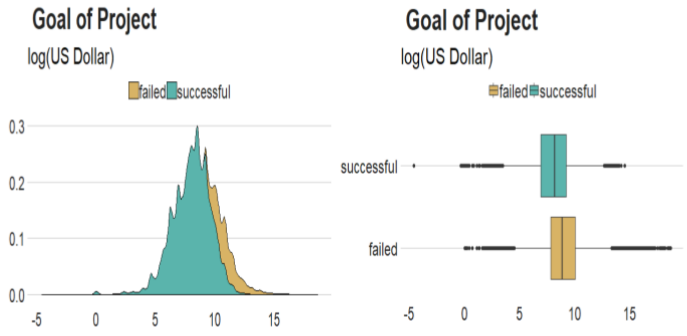
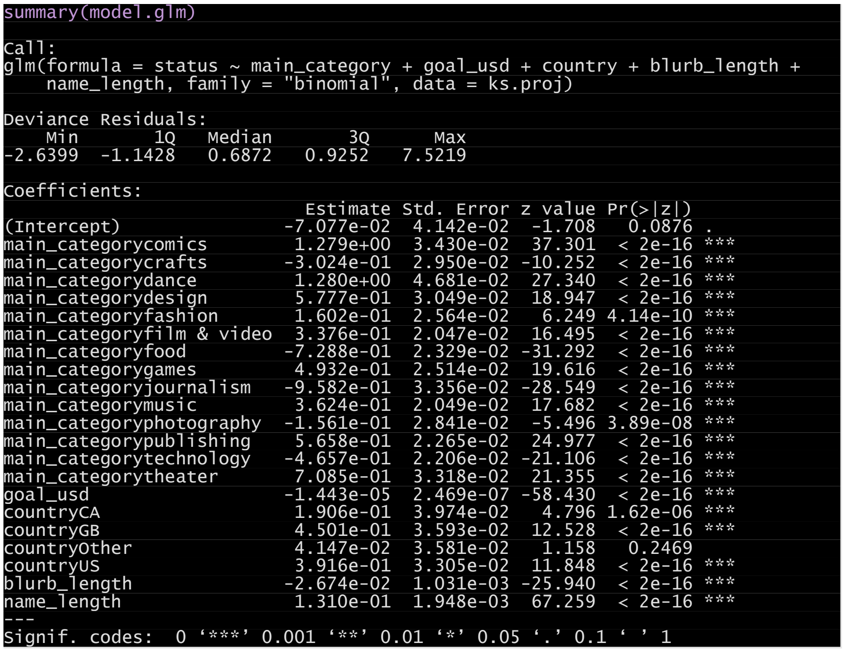


Figure 5: Kickstarter Project Money Goals: Success vs Failure

Our descriptive statistical a r data and simply describe what is or what the data shows. It is interesting to notice that majority of Kickstarter campaigns originate in the United States and in general there are more successful campaigns than failed. The most popular categories are in Music and Film & Video with higher percentage of succeeded campaigns. Majority of the successful campaigns have around 30 days duration period. Next, we moved on to showcase my predictive statistical analysis work, where predictive analysis was conducted through the use of logistic regression, classification trees, clustering, and association rules mining.

While the final report contains a more detailed account of preparation and methodology conducted for each predicative model, we will explain a few of them below to showcase my understanding and knowledge of the subject related to the program learning goal.

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous or binary such as “status” in the Kickstarter’s data. It is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. The goal of the logistic regression analysis is to estimate the log odds of an event. In addition, as seen in Figure 7, we plotted the ROC curve o optimal cut-off probability is the probability at which the logistic regression model has the least misclassification rate. Therefore, the optimal cut-off probability was plotted to find the probability with the least misclassification error. 0.54 was chosen as the optimal cut-off probability with a CV cost of 0.34. The prediction of successful or failed status is tabled below. In addition, a ROC curve was plotted to show the predictive threshold of our binary classifier model for training and validation model. Additionally, the area under the curve was tabled below to show models accuracy rate. It has a 68% accuracy, which is ok. This can be seen in Figure 8.



```
summary(model.glm)

Call:
glm(formula = status ~ main_category + goal_usd + country + blurb_length + 
    name_length, family = "binomial", data = ks.proj)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.6399  -1.1428   0.6872   0.9252   7.5219  

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)    
(Intercept)               -7.077e-02  4.142e-02  -1.708   0.0876 .  
main_categorycomics        1.279e+00  3.430e-02  37.301  < 2e-16 ***
main_categorycrafts       -3.024e-01  2.950e-02 -10.252  < 2e-16 ***
main_categorydance         1.280e+00  4.681e-02  27.340  < 2e-16 ***
main_categorydesign        5.777e-01  3.049e-02  18.947  < 2e-16 ***
main_categoryfashion       1.602e-01  2.564e-02   6.249 4.14e-10 ***
main_categoryfilm & video  3.376e-01  2.047e-02  16.495  < 2e-16 ***
main_categoryfood         -7.288e-01  2.329e-02 -31.292  < 2e-16 ***
main_categorygames         4.932e-01  2.514e-02  19.616  < 2e-16 ***
main_categoryjournalism   -9.582e-01  3.356e-02 -28.549  < 2e-16 ***
main_categorymusic         3.624e-01  2.049e-02  17.682  < 2e-16 ***
main_categoryphotography  -1.561e-01  2.841e-02  -5.496 3.89e-08 ***
main_categorypublishing    5.658e-01  2.265e-02  24.977  < 2e-16 ***
main_categorytechnology   -4.657e-01  2.206e-02 -21.106  < 2e-16 ***
main_categorytheater       7.085e-01  3.318e-02  21.355  < 2e-16 ***
goal_usd                  -1.443e-05  2.469e-07 -58.430  < 2e-16 ***
countryCA                  1.906e-01  3.974e-02   4.796 1.62e-06 ***
countryGB                  4.501e-01  3.593e-02  12.528  < 2e-16 ***
countryOther               4.147e-02  3.581e-02   1.158   0.2469    
countryUS                  3.916e-01  3.305e-02  11.848  < 2e-16 ***
blurb_length              -2.674e-02  1.031e-03 -25.940  < 2e-16 ***
name_length                1.310e-01  1.948e-03  67.259  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 6: Logistic Regression Summary Output

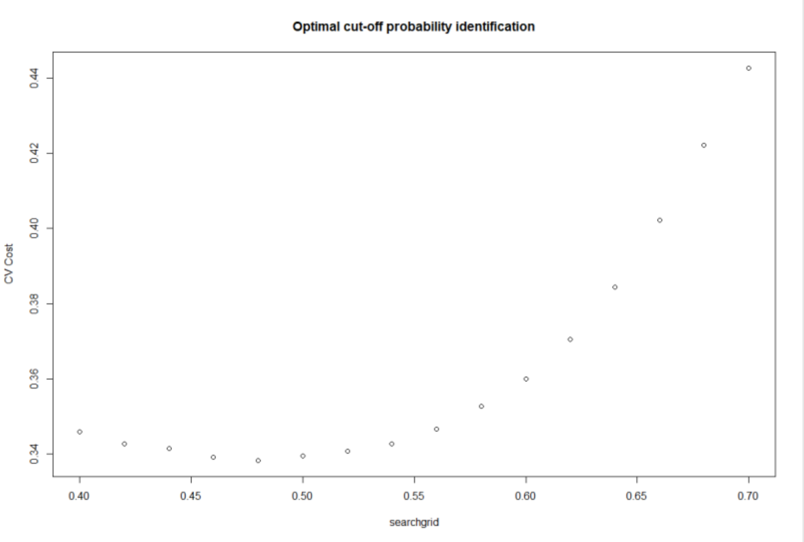
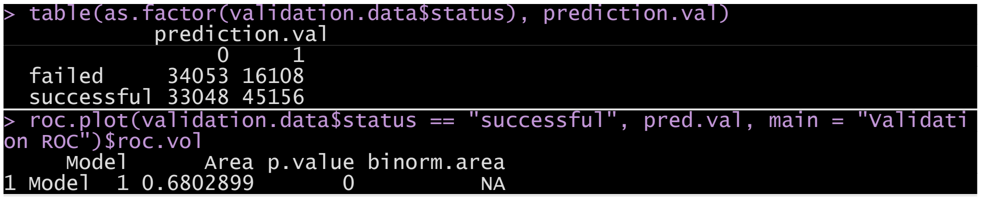


Figure 7: ROC Curve plotted to show predictive threshold of binary classifier model.



```
> table(as.factor(validation.data$status), prediction.val)
            prediction.val
                 0     1
  failed     34053 16108
  successful 33048 45156
> roc.plot(validation.data$status == "successful", pred.val, main = "Validati
on ROC")$roc.vol
     Model      Area p.value binorm.area
1 Model  1 0.6802899       0          NA
```

Figure 8: Predicting Model Accuracy Rate

We then modeled our data using classifications trees, specifically we utilized a decision tree and a random forest model classification tree. Decision tree is a supervised learning predictive model that uses a set of binary rules to determine the output value. The decision rules generated by the algorithm are visualized as a binary tree. Decision trees work by finding the variable that best splits the outcome into two groups. The recursion stops when all groups are sufficiently small. Decision tree model was built as the next predictive analysis method. The tree was built on the binary response variable “status”. The data for the tree was split based on the predictor variables. Figure 9 shows the separation between failed and successful campaigns. This is also later the same results attained for the first tuned model.

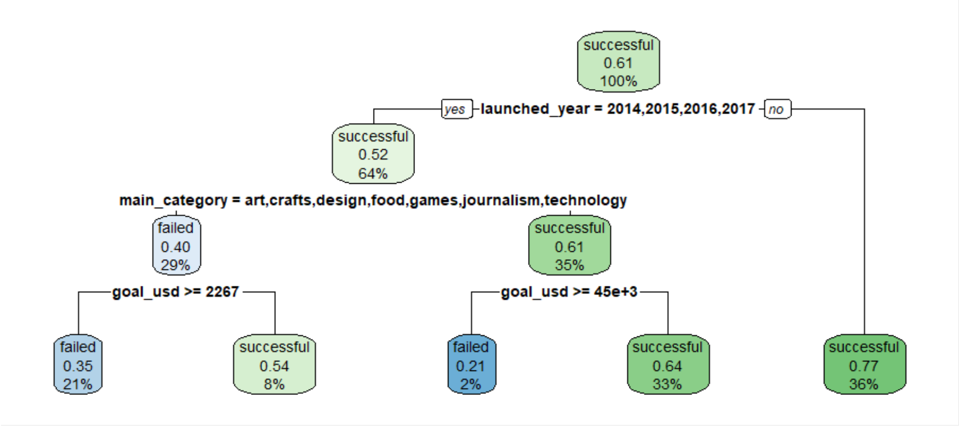
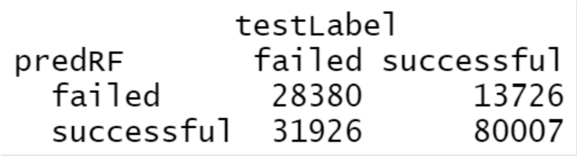


Figure 9: Classification Tree: separation between failed and successful campaigns.

We also created a Random forest model, also known as a random decision forest, combines hundreds of individual decision trees, and trains each one on a marginally different set of the observations. It then splits the nodes from each tree based on a pre-set, limited number of the features. The final predictions, as you can see from figure 10 and 11. from the random forest model are determined by averaging the projections from each individual tree. Our random forest has a constraint of 53 categories for any one variable. This caused sub-category to not be considered, resulting in the following variables: goal in US dollars, main category, duration, name length, blurb length, and country of the campaign. The dataset was shuffled by randomizing the row order using a seed of 2001.

The number of decision trees contained in the forest is 500, with a mean tree size of 711. There was a normal distribution in the tree size across the random forest. This model resulted in a 70.3% accuracy rate.



```
             testLabel
predRF        failed successful
  failed       28380      13726
  successful   31926      80007
```

Figure 10: Random Forest Success vs Failure

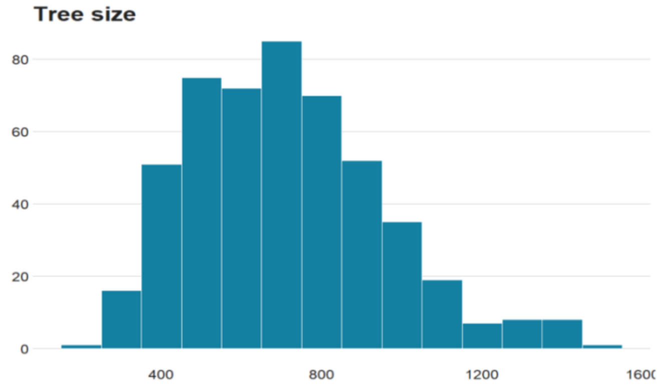


Figure 11: Random Forest Tree Size

Association analysis helps to discover patterns and interesting relationships in large sets of data. These relationships can be represented as a set of frequent rules, where features frequently occur together or correlated. The goal of associated rule mining is to find associations of items that occur together more often than it is expected from a random sampling. The algorithm is used with categorical non-numeric data. Apriori algorithm is used to run association rule mining. It uses an iterative method known as level-wise search. As you can see in figure 12, the output of Apriori algorithm was sorted by value of Lift parameter. Success of each Kickstarter campaign is primarily associated with United States and US Dollar currency as well as shorter duration and smaller amounts that are pledged.

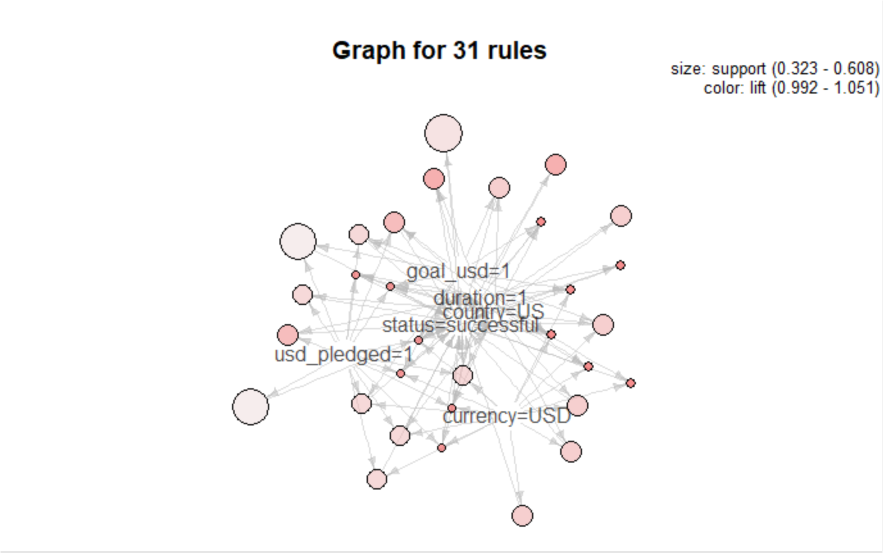


Figure 12: Apirori Algorithm Results

Conclusion of Learning Goal:

Based on the projects presented through this learning goal, I have not only demonstrated mastery of being able to identify patterns in data via visualization, statistical analysis, and data mining; I have shown that I can take large formats of data into a model after it has been ethically prepared and cleaned, and then be able to have some form of meaningful output and results from my analysis. This output is not only meaningful to me, the data scientist, but to higher level managers who would like to visually see the outputs of the tables and graphs. Mastery of this skillset not only helped train my statistical knowledge and understandings, but it helped me to better understand how to use visualization to help my analysis and presentation.

Click here to see the associated documents for this project.

Objective 4: Develop Alternative Strategies Based on The Data

Learning Objective Status: Mastered 🎓🎓🎓🎓🎓

Background:

Almost all of the projects and assignments completed in the iSchool Masters in Data Science curriculum have required mastery of skills in developing alternative strategies based on the data. For example, below is a summary of my final project for my Big Data Analytics course (IST-718). In general, in IST-718, we completed many projects and assignments that's required the ability to develop alternative strategies based on the data. For example, one our projects required predicting the recommended salary for a Syracuse Football Coach and determining what the single biggest impact on his or her salary would be. In addition, we then compared his recommended salary to if Syracuse was in the Big East or Big 10 Conferences. Additionally, we reviewed what schools were dropped from our analysis, and why. Finally, we then compared graduation rates to projected salary and then determined the accuracy of our model. However, while our initial project methodologies and objectives are clear-cut; we realized after our initial data clean up that we had to adjust our strategy to incorporate external data and fuzzy text analysis to help us create a better salary recommendation system that was based on other schools and football coaches' salaries.

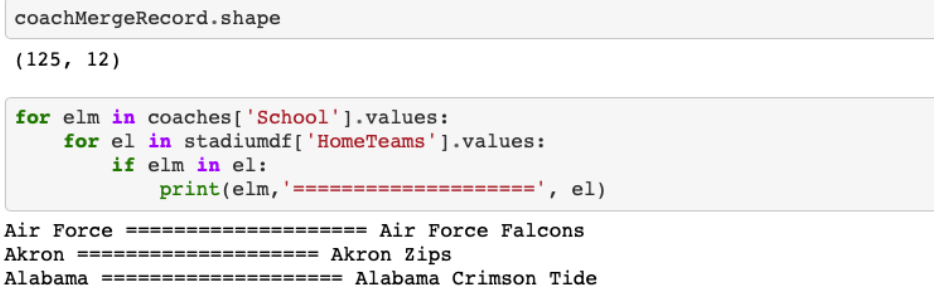
Data Questions:

1. What is the recommended salary for a Syracuse football coach?
2. What would salary be if Syracuse was still in the Big East? -Compare to Big Ten?
3. What schools were dropped from the data. - Why?
4. What effect does graduation rates have on projected salary?
5. How good is the model?
6. What is the single biggest impact on salary size?

Tools & Techniques Used:

To meet the objectives of this project we used python in Jupiter Notebook to conduct our analysis. As we mentioned earlier, we realized that in order to develop a vector for each school, we must use import coaches win/loss record data from an external source. We solved this issue by importing the coaches win and loss In addition, we needed to adjust our strategy a second time by importing a .csv file from https://www.sports-reference.com/cfb/coaches/a-index.html.

However, after we solved our first roadblock, we came across another issue with the stadium data that forced us to change our strategy. The issue was that the names of the colleges across the various datasets did not match. This would cause a major issue when merging our data sets because the university and team name are really the only ways to connect coach, stadium, graduation rate, and coach win/loss record data into a single master dataset.
Therefore, before we started to merge our files, we had to make sure our datasets could be merged based on comparable variables. So, our first check was to make sure our stadiums dataset could be merged with our coach's dataset along the lines of school and teams. We used lambda and fuzzy string matching to help us achieve this. Specifically, the levenshtein ratio and distance was calculated to find the distance between two string values. If the ratio_calc is true, then the function will compute the levenshtein distance ratio of similarity between two strings. The function would then print the matches of schools where our levenshtein distance is greater than 0.7. Figure 13 shows some of the final matched schools after we applied our fuzzy matching calculation.



```
coachMergeRecord.shape
```

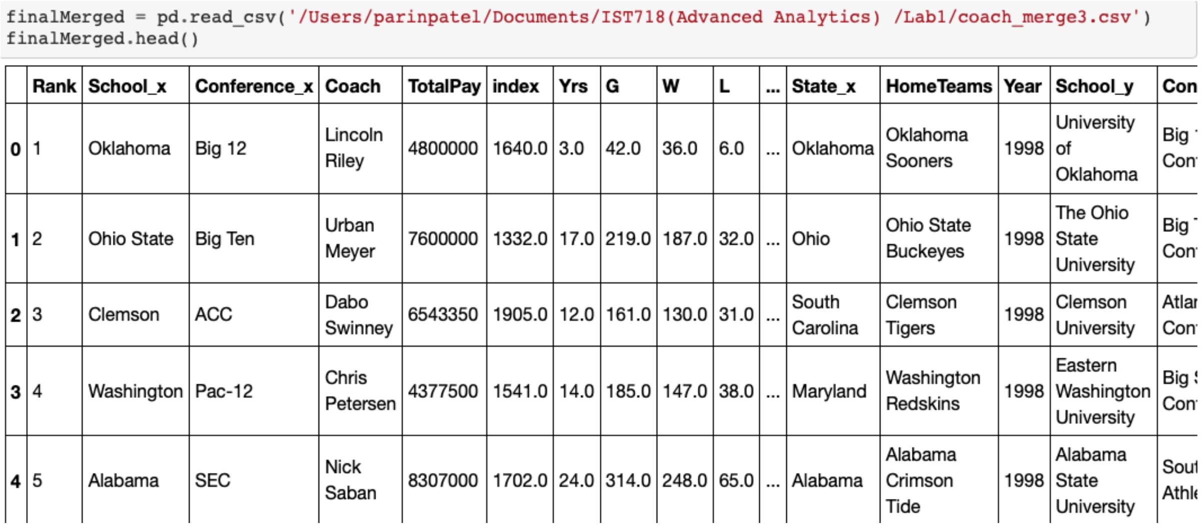
```
(125, 12)
```

```
for elm in coaches['School'].values:
    for el in stadiumdf['HomeTeams'].values:
        if elm in el:
            print(elm,'====================', el)
```

```
Air Force ==================== Air Force Falcons
Akron ==================== Akron Zips
Alabama ==================== Alabama Crimson Tide
```

Figure 13: Fuzzy Text Matching Coaches to Stadiums

When the datasets were ready to be merged, we first merged the coaches dataset with the Coaches Win and Loss Records dataset by Coaches. Finally, the Graduation Success Rate dataset was merged with the previously combined dataset by "School" to result in the final merged product. This final merged dataset was then further cleaned to remove any rows where TotalPay was either 0 or less than 0 and any rows where there was a duplicate school. Figure 14 will show a portion of the final table that merges the various tables by school.

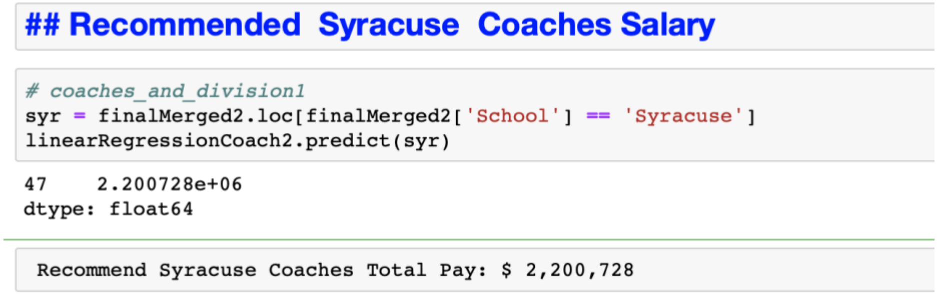


```
finalMerged = pd.read_csv('/Users/parinpatel/Documents/IST718(Advanced Analytics) /Lab1/coach_merge3.csv')
finalMerged.head()
```

	Rank	School_x	Conference_x	Coach	TotalPay	index	Yrs	G	W	L	...	State_x	HomeTeams	Year	School_y	Con
0	1	Oklahoma	Big 12	Lincoln Riley	4800000	1640.0	3.0	42.0	36.0	6.0	...	Oklahoma	Oklahoma Sooners	1998	University of Oklahoma	Big Con
1	2	Ohio State	Big Ten	Urban Meyer	7600000	1332.0	17.0	219.0	187.0	32.0	...	Ohio	Ohio State Buckeyes	1998	The Ohio State University	Big Con
2	3	Clemson	ACC	Dabo Swinney	6543350	1905.0	12.0	161.0	130.0	31.0	...	South Carolina	Clemson Tigers	1998	Clemson University	Atlar Con
3	4	Washington	Pac-12	Chris Petersen	4377500	1541.0	14.0	185.0	147.0	38.0	...	Maryland	Washington Redskins	1998	Eastern Washington University	Big Con
4	5	Alabama	SEC	Nick Saban	8307000	1702.0	24.0	314.0	248.0	65.0	...	Alabama	Alabama Crimson Tide	1998	Alabama State University	Sout Athle

Figure 14: Final Merged Dataset by Fuzzy School Matching

Because of our successful change in strategies early on , we were able to build a successful model that answered all of our data questions. For example, we were able to use linear regression and ordinary least squares (OLS) regression to predict coaches salary. For example, figure 15 shows that we predict a Syracuse football coach to make an annual salary of $2,200,728. In addition, we used the OLS model in Figure 16 to predict that a Syracuse coach would make much more (about $3.71 million) if the team was in the Big 10.



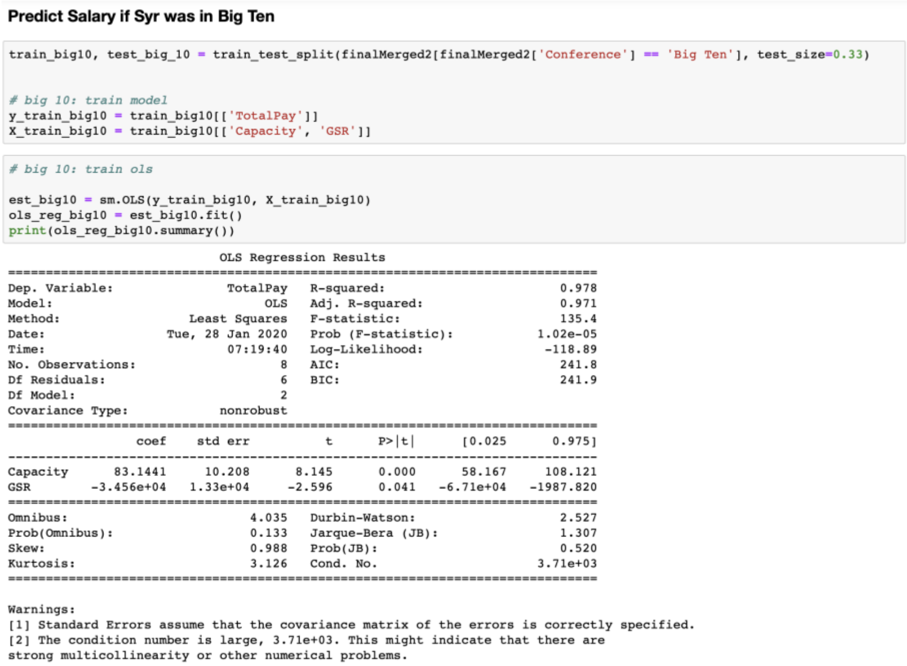
Recommended Syracuse Coaches Salary

```
# coaches_and_division1
syr = finalMerged2.loc[finalMerged2['School'] == 'Syracuse']
linearRegressionCoach2.predict(syr)
```

```
47    2.200728e+06
dtype: float64
```

```
Recommend Syracuse Coaches Total Pay: $ 2,200,728
```

Figure 15: Recommended Syr Coach Salary Linear Regression Predict



Predict Salary if Syr was in Big Ten

```
train_big10, test_big_10 = train_test_split(finalMerged2[finalMerged2['Conference'] == 'Big Ten'], test_size=0.33)


# big 10: train model
y_train_big10 = train_big10[['TotalPay']]
X_train_big10 = train_big10[['Capacity', 'GSR']]
```

```
# big 10: train ols

est_big10 = sm.OLS(y_train_big10, X_train_big10)
ols_reg_big10 = est_big10.fit()
print(ols_reg_big10.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               TotalPay   R-squared:                       0.978
Model:                            OLS   Adj. R-squared:                  0.971
Method:                 Least Squares   F-statistic:                     135.4
Date:                Tue, 28 Jan 2020   Prob (F-statistic):           1.02e-05
Time:                        07:19:40   Log-Likelihood:                -118.89
No. Observations:                   8   AIC:                             241.8
Df Residuals:                       6   BIC:                             241.9
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Capacity      83.1441     10.208      8.145      0.000      58.167     108.121
GSR        -3.456e+04   1.33e+04     -2.596      0.041   -6.71e+04   -1987.820
==============================================================================
Omnibus:                        4.035   Durbin-Watson:                   2.527
Prob(Omnibus):                  0.133   Jarque-Bera (JB):                1.307
Skew:                           0.988   Prob(JB):                        0.520
Kurtosis:                       3.126   Cond. No.                     3.71e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.71e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Figure 16: OLS Model SYR in Big 10

Click here to see the associated documents for this project.

Conclusion of Learning Goal:

Based on the projects presented in this learning goal, I have demonstrated a mastery at being able to develop alternative strategies based on the data and project objectives. Additionally, I have shown that I can be adaptable and creative when applying my knowledge of statistical methodologies. Merging the two datasets through fuzzy text matching, I had to complete an extensive amount of research. It was actually my success at implementing fuzzy text matching in this project that I felt comfortable to tackle larger issues related to text matching in my professional work. Currently, all of my major projects and research abstracts are related to solving issues related to variability in naming. In my office, we monitor all of the packaged food found in the US for nutrition and chemical additives. So, there are a lot of databases with a lot of different entry methods for brand and manufacturer name. One my most successful projects involves matching manufacturers and brand names in these databases to one another. This is just the first step in a much larger data lake project, but it has become a very important step and has made me more comfortable in actively seeking alternative strategies when working with data.

Objective 5: Develop A Plan of Action To Implement The Business Decisions Derived From The Analyses.

Learning Objective Status: Mastered 🎓🎓🎓🎓🎓

Background:

Almost all of the projects and assignments completed in the iSchool Masters in Data Science curriculum have required mastery of skills in developing alternative strategies based on the data. For example, below is a summary of one of the many projects completed for my Business Analytics course (SCM-651) that goes through optimization of product pricing. In addition, my final project for Marketing Analytics (MAR-653) required the ability to develop a plan of action to implement business decisions derived from our analysis of survey data.

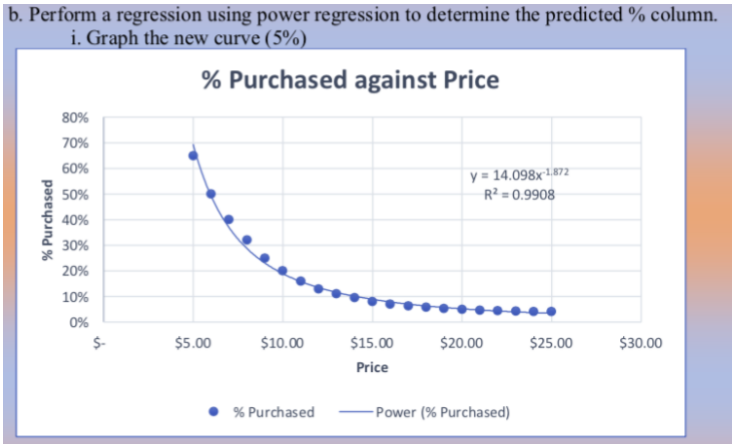
Reference of Student Work: Part 1

In the course (SCM-651), I successfully demonstrated skills related to deriving a plan of action to implement into the business decision based on analysis of data. For this assignment, we were given a situation where a bookstore wanted to price books at the most optimal rates to maximize profits. The bookstore ran a test the week before where they varied the prices on the Harry Potter 7 book to determine a demand curve. They also quantified the percent of customers who visited their website and purchased Harry Potter 7. This was all recorded in an excel spreadsheet. Now, since the author of the Harry Potter book series, J.K. Rowling, has announced a sequel, the bookstore wants to know price they should make this book at to optimize profits. We were given the following 5 assumptions prior to analysis.

Data Assumptions:

1. Assume that the demand for the book sequel will be similar to Harry Potter 7.
2. Assume that 100,000 customers will consider purchasing a book from you
3. The data is not an entirely accurate prediction of the demand, but a regression on the
4. data using a power model will give a reasonable prediction
5. Assume that you pay the publisher $5.00 for each book.

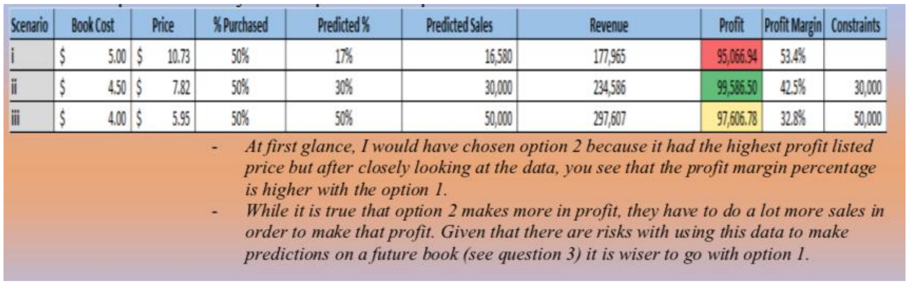
To help implement the pricing decision for the bookstore, we successful utilized regression analysis to model the percent of books purchased against price, and to determine the predicted percentage columns. As you can see in Figure 17, from our predicted column, we generated an R2 of 0.9908, which means that 99% of the change of % purchased can be explained by the change in Price using the model coefficients calculated by excel. We then went onto predict sales based on the predicted percentage equation generated from our regression analysis ($y = 14.098x^{-1.872}$). We then used the predicted sales and price of the book to generate a predicted revenue column and profit column. Figure 18 shows the final determined cost point (from the publisher) and the price (to the customer) to maximize the bookstores profits. As you can see, scenario I was the best determined option.



b. Perform a regression using power regression to determine the predicted % column.

i. Graph the new curve (5%)

Figure 17: Regression to determine predicted column



Scenario	Book Cost	Price	% Purchased	Predicted %	Predicted Sales	Revenue	Profit	Profit Margin	Constraints
i	$ 5.00	$ 10.73	50%	17%	16,580	177,965	95,066.94	53.4%	
ii	$ 4.50	$ 7.82	50%	30%	30,000	234,586	99,586.50	42.5%	30,000
iii	$ 4.00	$ 5.95	50%	50%	50,000	297,607	97,606.78	32.8%	50,000

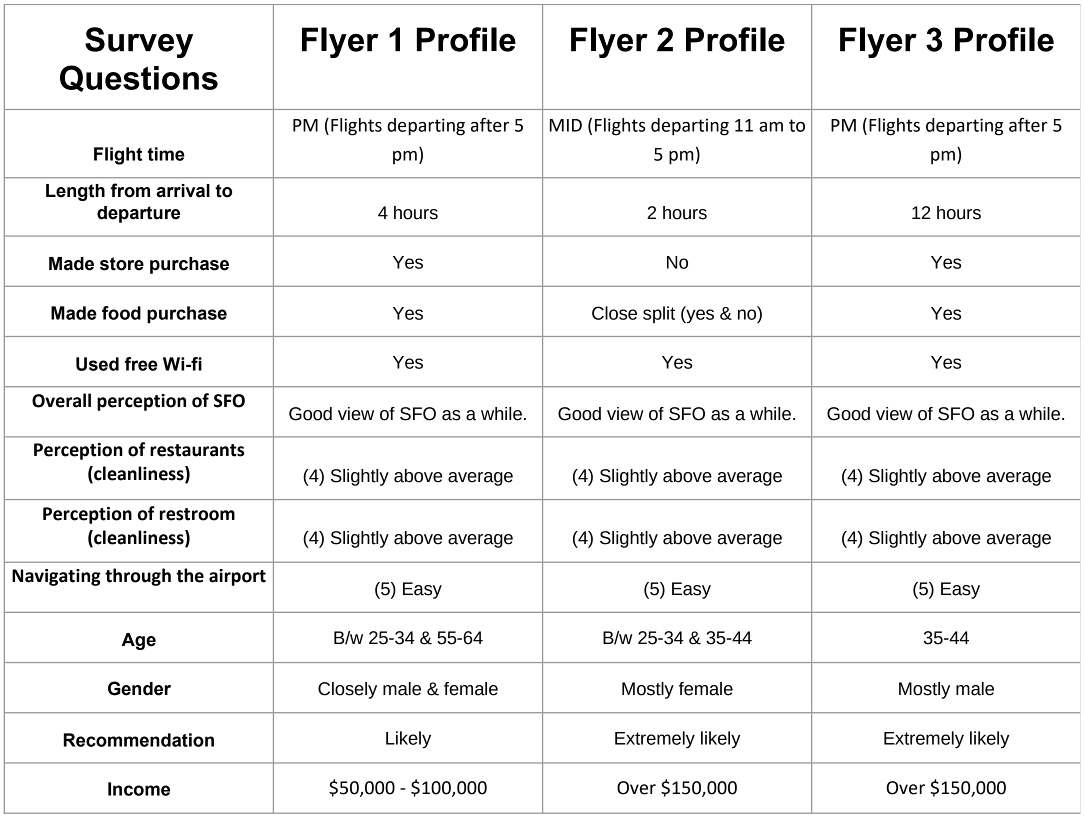
- *At first glance, I would have chosen option 2 because it had the highest profit listed price but after closely looking at the data, you see that the profit margin percentage is higher with the option 1.*
- *While it is true that option 2 makes more in profit, they have to do a lot more sales in order to make that profit. Given that there are risks with using this data to make predictions on a future book (see question 3) it is wiser to go with option 1.*

Figure 18: Comparison of top 3 pricing scenarios.

Click here to see the associated documents for this project.

Reference of Student Work: Part 2

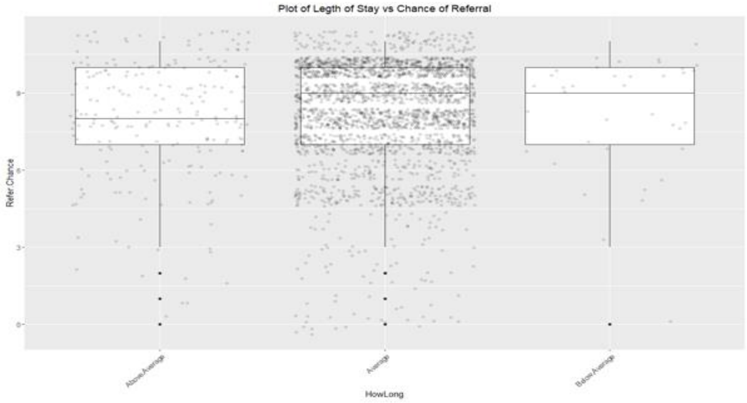
In the course (MAR-653), I successfully demonstrated skills related to deriving a plan of action to implement into the business decision based on analysis of data. For our final project, we analyzed survey data from San Francisco International Airport where customers were asked to rate their experience at the airport with nineteen (19) different categories from cleanliness to travel style and spending behavior at the airport, we identified lapses in restaurant and store spending. By identifying where customer satisfaction were the lowest, we could identify and provide recommendations for improvement. Some of the analysis techniques included k-means clustering ,ordinal logit regression, and logistic regression. For example, by using K-means clustering, we were able to profile flyers to gain a deeper insight into the types of people of people visiting and not visiting stores and restaurants at SFO. Based on our model, results you can see in Figure 19, we identified three (3) flyer profiles. This would not only serve as a means to understand SFO flyers but would become essential for our recommendations to store and restaurant owner. For example, the Flyer 3 profile are older people who are wealthy. They are more likely to spend more on stores and restaurants/bars due to their disposable income. Additionally, their longer stay at the airport makes them susceptible to impulse buys.



Survey Questions	Flyer 1 Profile	Flyer 2 Profile	Flyer 3 Profile
Flight time	PM (Flights departing after 5 pm)	MID (Flights departing 11 am to 5 pm)	PM (Flights departing after 5 pm)
Length from arrival to departure	4 hours	2 hours	12 hours
Made store purchase	Yes	No	Yes
Made food purchase	Yes	Close split (yes & no)	Yes
Used free Wi-fi	Yes	Yes	Yes
Overall perception of SFO	Good view of SFO as a while.	Good view of SFO as a while.	Good view of SFO as a while.
Perception of restaurants (cleanliness)	(4) Slightly above average	(4) Slightly above average	(4) Slightly above average
Perception of restroom (cleanliness)	(4) Slightly above average	(4) Slightly above average	(4) Slightly above average
Navigating through the airport	(5) Easy	(5) Easy	(5) Easy
Age	B/w 25-34 & 55-64	B/w 25-34 & 35-44	35-44
Gender	Closely male & female	Mostly female	Mostly male
Recommendation	Likely	Extremely likely	Extremely likely
Income	$50,000 - $100,000	Over $150,000	Over $150,000

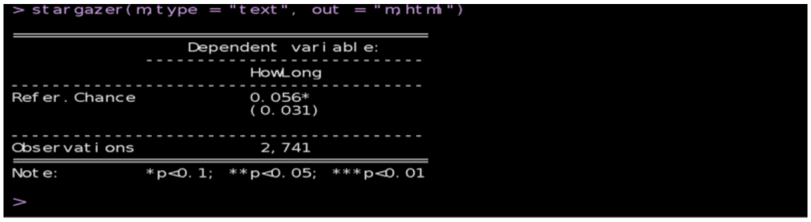
Figure 19: K-Means Clustering Results Flyer Profiles

We then were able to create two different ordinal logit models to assess if referral ratings were affected by passenger time spent at the airport and if restaurant ratings have an effect on referrals. Figure 20 and 21, shows the results of our first analysis that Referral Chance increases one unit, it is 0.056 times more likely to be in a higher category. The odds of moving to a higher category in the outcome variable is -94% more likely when Referral Chance moves one unit. Since Referral Chance is not statistically significant of time at the airport, the longer a customer stays at the airport is not related to a better approval rating. We can take this information to the restaurants and shops to help tune marketing strategies that are not focused on taking the time to attract customers



New Code	
Below 52 Minutes	Below Average
53-283 Minutes	Average
Above 284 Minutes	Above Average

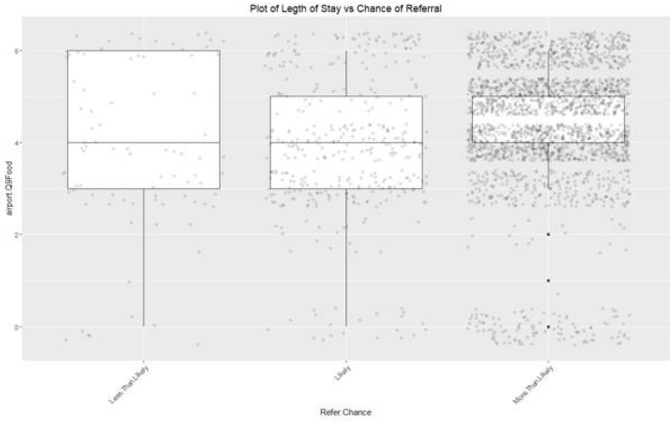
Figure 20: Results of Time at Airport vs Referral



```
> stargazer(m,type = "text", out = "m.html")

===========================================
                    Dependent variable:
                 --------------------------
                          HowLong
-------------------------------------------
Refer.Chance              0.056*
                         (0.031)

-------------------------------------------
Observations               2,741
===========================================
Note:           *p<0.1; **p<0.05; ***p<0.01

>
```

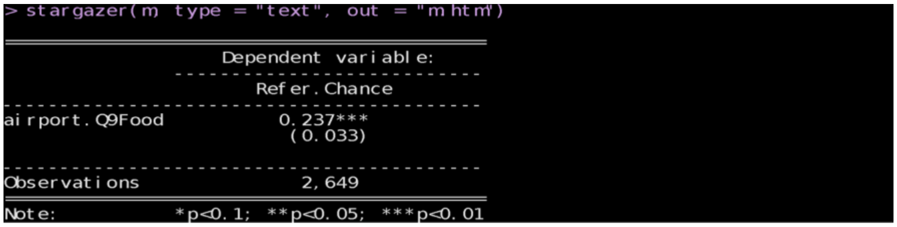
Figure 21: Time at Airport vs Referral R Output Ordinal Logit Regression

For the second model we re-coded passenger referral ratings to narrow our spread. Similar to the first, we grouped the referral into three categories. As you can see from Figure 22 and Figure 23, when a passenger's chance of referral increases by one unit, it is 0.237 times more likely to be in a higher category. The odds of moving to a higher category in the outcome variable is 76.3% more likely when Referral Chance moves one unit. Also, Referral Chance is statistically significant to a passengers rating of the restaurant. Therefore, a passenger experience at SFO's restaurants does affect the outcome of their referral.



New Code Referral	
Below Rating of 4	Less Than Likely to Refer
Rating 4-6	Likely to Refer
Rating Over 6	More Than Likely To Refer

Figure 22: Referral vs Restaurant Rating



```
> stargazer(m type = "text", out = "m htm")

==========================================
                    Dependent variable:
                ---------------------------
                      Refer.Chance
------------------------------------------
airport.Q9Food          0.237***
                        (0.033)

------------------------------------------
Observations             2,649
==========================================
Note:          *p<0.1; **p<0.05; ***p<0.01
```

Figure 23: Ordinal Logit Regression Referral vs Restaurant Rating

We finally then conducted a normal logistic regression model to isolate which variables have an effect on the likelihood of making a purchase at the restaurants. Dummy variables that examined day of the week, time of flight and gender were introduced to measure if they had an effect. In addition, we found the total time spent in the airport and respondent age showed statistical significance. In addition, we found that as you spend more time in the airport, you're less likely to make a purchase.

Based on our analysis, we were able to develop a plan of action for SFO to attract more customers to increase restaurant and shopping experience. Figure 24 visually shows these recommendations.

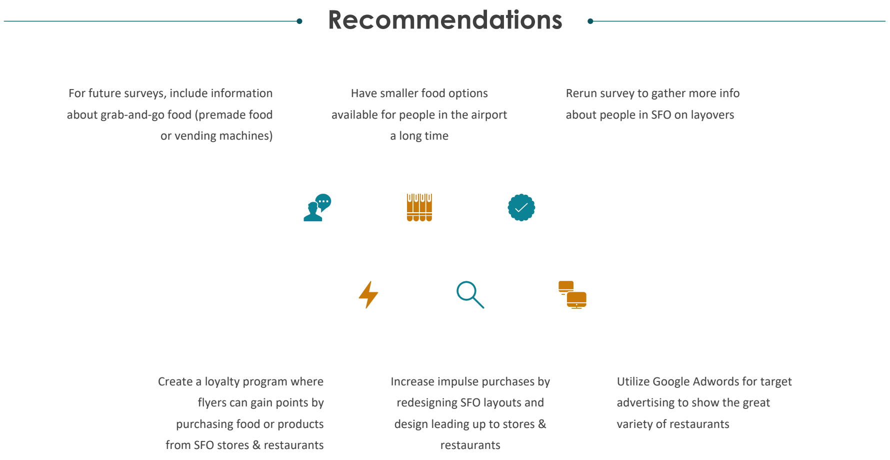

Recommendations
For future surveys, include information about grab-and-go food (premade food or vending machines)
Have smaller food options available for people in the airport a long time
Rerun survey to gather more info about people in SFO on layovers
Create a loyalty program where flyers can gain points by purchasing food or products from SFO stores & restaurants
Increase impulse purchases by redesigning SFO layouts and design leading up to stores & restaurants
Utilize Google Adwords for target advertising to show the great variety of restaurants

Figure 24 Recommendations for SFO

Click here to see the associated documents for this project.

Conclusion of Learning Goal:

Based on the projects presented in this learning goal, I have demonstrated a mastery at skills related to deriving a plan of action to implement into the business decision based on analysis of data. From the above two examples, you can see how I was able to use the conclusions derived from my analysis to help recommend optimal prices of products or recommendations based on consumer surveys. This objective has taught me a wide variety of action-based skills that I am able to take to my professional career. For example, I am able to take responses and feedback from our contractors to help create better future programs

Objective 6: Demonstrate Communication Skills Regarding Data and Its Analysis for Managers, IT Professionals, Programmers, Statisticians, And Other Relevant Professionals In Their Organization.

Learning Objective Status: Mastered 🎓🎓🎓🎓🎓

Background:

Almost all of the projects and assignments completed in the iSchool's Masters in Data Science curriculum and those highlighted in the earlier objectives have required mastery of skills in communications regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization. For almost every one of the courses, the ability to effectively communicate the results of a technical analysis to a non-technical audience was key to excelling and passing the course. As a student progresses through the data science program at Syracuse, they will find that the ability to communicate their findings through presentations, reports, and group discussion is a crucial requirement and will serve as a long-term benefit in the student's professional life.

For example, below is my final report for my Big Data Analytics course (IST-718). The goal of the project was to complete an examination of President Trumps Twitter posts using sentiment analysis and machine learning techniques like cosine similarity to see the trends found in tweets. The analysis sought to answer the following questions:

Data Questions:

1. Does Tweeting impact policy?
2. Can we map and trend Presidential Tweets ?
3. Can we predict Presidential re-electability by tweet behavior?
4. Does Presidential tweeting affect the stock market?

Tools & Techniques Used:

Sentiment analysis reviews text data for indications of positivity, negativity, or some other form of sentiment. This value is then measured as a rating between negative one (-1) and positive one (+1). Natural language processing (NLP) is a methodology that allows researchers to analyze the sentence and language-related data using computational techniques to determine specific outcomes. It is one of the more advanced techniques taught after core fundamentals. Sentiment analysis, predictive analysis, and descriptive analysis were utilized in the final review of President Trumps Tweets. In addition, this project required mining twitter, presidential approval ratings, and historical stock market data. The goal of this project was to utilize all the skills we had learned during our tenure as students of the iSchool in order to evaluate Presidents Trumps tweets for trends using current data science techniques including NLP, Word Clouds and Sentiment analysis. In addition, we compared his sentiment analysis score of tweets to his approval ratings to see if the language he used in his tweets has an effect on the opinions of the public. Finally, we ran a comparison of his combined approval-sentiment score against the stock markets historical data for 2017 to see if the market was provoked by his tweets. Since one of the objectives of this project was to communicate to a high-level target audience of the President, Directors/Policy Managers, and Voters; our findings were communicated in three ways: first as a written report, secondly as a take-away PowerPoint, third as a limited-PowerPoint discussion with classmates. The final discussion involved only a few key slides, but mostly was a discussion of our findings and feedback amongst team members. As you will be able to see in the projects folder, the PowerPoint was kept brief with only key figures and tables. The report contains the most detailed version of our findings and processes.

Conclusion of Learning Goal:

As I have mentioned extensively throughout this objective and paper, the ability to communicate technical findings in a clear, apprehensible manner is a very important life skillset. It is what sets data scientists apart from each other but is also necessary for professional and personal growth. In my personal life, the ability to communicate and understand the importance of communicating my findings to those with different backgrounds has been a very useful skillset. It is something that I work on daily and will continue to work on. It is also why I enjoy the research profession, since the ability to communicate is not just limited to PowerPoints and discussion here. For my work I have to draft a large volume of papers and abstracts, and having this skill makes me a much stronger candidate.

Click here to see the associated documents for this project.

Objective 7: Synthesize the Ethical Dimensions of Data Science Practice.

Learning Objective Status: Mastered 🎓🎓🎓🎓🎓

Background:

As students of data science, when we come across a standard data science question or problem, our responses are often finite. We have a limited number of options, due to certain data restrictions. We often know how to proceed because we are trained, through formal education or experience, what to do next. When it comes to considering the ethical dimensions of data science, and putting these considerations into practice, we often come across a much larger dilemma, often asking ourselves, “What do I do?”

Training to become an ethical data scientist is difficult to pin-point, because often the right and wrong path is not pre-defined. In a field as pioneering as data science, ethics are unfortunately second thought to result-generation. In my previous professional work, in the food data industry, I found this to be a common case – companies wanting to generate "big data" without actually understanding the basics of data standardization and quality management. Thereby, when it came to sell their accumulated datasets, they ran into large quality and consistency errors. I believe that through my education at the iSchool, I was able to become a more ethical data scientist first because I was taught how to be a data scientist. My professors taught me the consequence of having bad data, and how the results are skewed when we over-tune or under-tune our data. They also taught me to ethically manage outliers, and not just simply discard them. For example, in my Data Analytics course (IST-707), I discussed earlier how my final project involved predicting the success of Kickstarter campaigns. However, in this same course, we were taught to utilize the same data, but with different models, to understand how different models often yielded different outputs. Therefore, we were taught to identify the best-case methods when working with certain datasets. In this case, we were trying to predict the disputed author of the Federalist Papers using first clustering algorithms and then secondly through classification algorithms. In courses like IST-707, we were able to understand how misuse of certain methodologies yields in incorrect results. Through this experience, I am able to identify poor quality data models at work and give feedback to my peers on how they can better their findings. It also allows me to better analyze and review research papers prior to implanting their findings in my projects.

Finally, through courses like Information Policy (IST-618), I was able to research and synthesize the findings of data privacy and security. For one my papers, I wrote a reflective analysis of privacy and security in Australia. The reflection reviewed privacy laws of Australia and how General Data Protection Regulation (GDPR) has affected Australian businesses and legislation. In addition, the review is discussed the advantages and disadvantages of Australia's collection and use of its citizens and non-citizens personal information. Finally, the reflection had a risk analysis of the way Australian government uses surveillance on its people, and a recommendation for future policy to protect certain populations. Reflections like this are a great way for technical students, like me, who usually end up coding in all their projects or assignments to take a step-back and consider greater ethical topics.

Conclusion of Learning Goal:

As I have mentioned extensively throughout this objective and paper, when it comes to considering the ethical dimensions of data science, the path is not always so clear-cut. Therefore, being an ethical data scientist is something that can be instilled in students through training and understanding of the consequences of poorly handled data. Through my professional experience, where I work with mostly chemists and toxicologists, I have come to believe that the data environment should be handled similar to a laboratory. For example, similar to a lab environment, we should clearly write out our methodologies and results. It is also just as important to state what did not work, as it is to mention what worked when working in a data environment. A data scientist undergoes a wide variety of tasks, and there should be a clear log of what was completed in case the someone else wishes to reproduce their findings.

Click here to see the associated documents for this project.