Portfolio Milestone draft

Syracuse University, September 2019

Darrell L. Nelson II

Table of Contents

[Overview/Introduction 1](#_Toc19638025)

[Describe a Broad Overview of the Major Practice areas of Data Science 1](#_Toc19638026)

[Background 2](#_Toc19638027)

[Collect & Organize Data 2](#_Toc19638028)

[Problem to be Solved 2](#_Toc19638029)

[Data Questions 2](#_Toc19638030)

[Background 3](#_Toc19638031)

[Tools & Techniques Used 3](#_Toc19638032)

[Insight Gained 3](#_Toc19638033)

[Identify Patterns in Data: Visualization, Statistical Analysis, & Data Mining 5](#_Toc19638034)

[Problem to be Solved 5](#_Toc19638035)

[Data Questions 5](#_Toc19638036)

[Background 5](#_Toc19638037)

[Tools & Techniques Used 5](#_Toc19638038)

[Insight Gained 7](#_Toc19638039)

[Develop Alternative Strategies based on the Data 9](#_Toc19638040)

[Problem to be Solved 10](#_Toc19638041)

[Data Questions 10](#_Toc19638042)

[Background 10](#_Toc19638043)

[Tools & Techniques Used 10](#_Toc19638044)

[Insight Gained 13](#_Toc19638045)

[Develop a Plan of Action to Implement the Business Decisions derived from the Analyses 14](#_Toc19638046)

[Demonstrate Communication Skills regarding Data and its Analysis 14](#_Toc19638047)

[Synthesize the Ethical Dimensions of Data Science Practice 15](#_Toc19638048)

# Overview/Introduction

This report is an attempt to provide sufficient evidence that the learning objectives set forth by Syracuse University have been met. The Applied Data Science curriculum has provided the appropriate knowledge, background, resources, and trainings to impart in me the prerequisite fundamentals of a data scientist. This report will go through all 7 learning objectives and provide detailed examples of how the objectives were met throughout the program. The explanations will be buoyed with a brief synopsis of three final team projects that highlight the appropriate learning objectives.

Learning Objectives:

1. Describe a broad overview of the **major practice areas** of data science. (Data Mining, Predictions)
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).

# Describe a Broad Overview of the Major Practice areas of Data Science

Data Science is the field of analyzing data that stretches beyond descriptive analytics. Descriptive analytics tells the story of what has happened in order to understand trends and evaluate metrics over time. It relies heavily on elementary statistics (e.g. mean, median, standard deviation, quartiles, etc.) to showcase/visualize trends.

Predictive analytics goes a step forward. While descriptive analytics is limited to past data, predictive analytics allows data scientists to predict future trends. Trends such as: forecasting, predicting the values of missing fields in a data set (supervised learning), classifying unlabeled data (unsupervised learning), and probable impact of data changes on future trends (semi-supervised learning). Sentiment analysis and credit score analysis are excellent examples of predictive analytics. Sentiment analysis is the study of text to estimate the tendency of emotions conveyed through it. The aim is to measure whether the text is indicating a positive, negative, or some other classified emotion. This is generally measured by rating a piece of text between -1 to +1, with ‘+’ side indicating a positive sentiment and vice versa. Credit score analysis is the study of past financial behavior and income growth of the individual along with economic trends to predict the likelihood they will pay their debt. These analyses produce a probability distribution and predict which category a text or person will fall into. Most models also allow the precision (% of correct labels that were predicted correctly) and recall (% of predicted labels belonging to the correct label) to be calculated.

Prescriptive analytics is the final known step of the Data Science field. Prescriptive analytics showcases viable solutions and the impact of considering a solution on future trends. It is considered the aim of any data analysis project. Prescriptive analysis is still an evolving technique and there are limited applications for it in business.

Google’s self-driving car is a perfect example of prescriptive analytics. It analyzes the environment and decides the direction to take. It decides whether to slow down or speed up, change lanes, take a long cut to avoid traffic, or sacrifice time for safety. In this way, it functions like a human driver by using data analysis at scale. There are many examples (Deep Blue, Deep Thought, and AlphaGo) of prescriptive analytical processes being used by computers to master games. By understanding the overarching goal of the game and taking in test cases, they can masterfully beat new unforeseen challengers. The most notable being Deep Blue beating chess champion Garry Kasparov in a 6-game tournament in May 1997.

## Background

All three of the selected projects provide examples of the major practice areas of Data Science. From data collection and business plan creation, developing alternative analytical strategies, identifying patterns and communicating to a multi-faceted team, and how to deal with the ethical responsibilities that arise in data science.

From a technical perspective the selected projects will demonstrate examples of supervised learning through the lens of text classification and data mining.

# Collect & Organize Data

Learning objective has been mastered. All projects and assignments undergone in this program have honed my skills in data collection and cleaning. The example below is a synopsis of my Scripting for Analytics course (IST - 652) project. Knowledge to clean text data from Twitter to extract pertinent information was taught in the IST – 652 course. In this project, the raw text files were imported from Twitter into a Python IDE and cleaned/analyzed to produce the following analysis.

## Problem to be Solved

### Data Questions

* Does social media affect how many top football recruits a collegiate school receives?
* Is there a correlation between Twitter activity in the offseason and pre-season rankings?

### Background

College athletics is one of America’s favorite pastimes. Seeing young men and women hone their craft and compete at the highest level is a soap opera that the world just can’t turn off. The NCAA, the nonprofit association that runs college athletics, takes in close to $8 billion a year. According to a Business Insider report, there are now 24 schools that make at least $100 million annually from their athletic departments. With so much media attention, hype, and money on the line schools are clamoring to find a competitive advantage. Many schools invest millions of dollars into their “amateur” sports teams to build multi-million-dollar facilities, coaches, and staff.

With such a deluge of money circulating around college athletics it is easy to imagine why schools are habitually getting caught for paying their amateur athletes. With such a huge incentive for the school/program to win, attracting talent is top priority. What if there is a way to influence top prospects to come to town without crossing the red line? The analysis laid out below investigates how social media affects preseason rankings and more importantly, recruitment.

## Tools & Techniques Used

Script was written in Python 3.7 using the Spyder IDE. A twitter developer account was required from “<https://developer.twitter.com/en.html>” to gain access to the Twitter API. Tweepy, pandas, datetime, and numpy libraries were imported. The dataset generated was a collection of Twitter data compiled by accessing the Twitter API through the python library Tweepy. Tweets were searched based on the date range of the preseason, the twitter handle of the team, and the maximum number of tweets desired. The dataset imported was a collection of 3,041 tweets that were released by the football teams in the Pacific 12 conference (Pac-12). Each tweet was imported and listed as a JSON structure. The data was manipulated with numpy and pandas dataframes. There was no preprocessing required. The semi-structured data came with the appropriate labels and datatypes to facilitate the analysis.

A function named “TwitterFunc” was created to take in a username, timeframe, and max number of tweets. For each tweet the script outputted the number of retweets and likes it had. These numbers were then put into a list and summed up, with the final output of the function being a list containing the total number of tweets found, number of total retweets, and total number of likes. Originally, a for loop was created to iteratively look up each school’s Twitter handle and find the accompanying data. However, it resulted in multiple computer crashes. Instead, each handle was looked up manually and their data was placed into an Excel spreadsheet. The spreadsheet was imported back into python with the pandas read\_excel function. Then, a new sorted table was created for each attribute. Two new columns (‘retweet per tweet’ & ‘like per tweet’) were added to the original database and sorted as well.

The exported file from Python is the dataframe pulled from the Twitter API composed of the 5 features discussed above for each team. A displacement metric was built to calculate the absolute distance between predicted rankings and actual rankings for all teams on all features. The second exported file is the displacement metric comparing recruiting rankings to Twitter features. The third exported file is the displacement metric comparing preseason rankings to Twitter features.

## Insight Gained

Based on the distance metric and data provided, the attribute that best predicted the rankings for the top 3 teams in recruiting was the retweet per tweet feature. It was also the best feature when looking at the predictive performance for all 12 schools. For preseason rankings there were two metrics tied for the best: tweets and retweet per tweet. When viewing the best predictive performance for all 12 schools the retweet per tweet was the sole champion again.

In all, the retweet per tweet feature performed the best in predicting rankings for recruiting and preseason. Comparing those metrics visually produced *Figures 1 & 2*.

*Figure 1. Retweet/tweet vs Recruiting Rankings*

*Figure 2. Retweet/tweet vs Pre-Season Rankings*

The *x*-axis denotes the counts of retweets per tweet (Rtw/tw) for each team, and the *y*-axis lists the teams in descending order of the recruiting/pre-season rankings with the best teams up top and the worst on bottom. Of the two graphs above, *Figure 1.* produces a clearer direct relationship which cooraberates the fact that recruiting rankings had the lowest overall sum of displacemtent. It appears recruits can be persuaded to join a team if that team is popular on social media. The pre-season rankings, which are released by the media, appear to be more immune to social media fads. Perhaps due to the media’s daily access to the teams they are able to compare and rank them without being persuaded by outside buzz.

It is strongly recommend for teams to invest in building up their social media platforms to attract more recruits. Even if the product on the field is subpar, showing teenagers that the organization is relatable and shares the same taste and culture as them could be the difference between landing that 5-star recruit and that player going to your rival.

# Identify Patterns in Data: Visualization, Statistical Analysis, & Data Mining

Learning objective has been mastered. All projects and assignments undergone in this program have honed my skills in visualization, statistical analysis, and data mining. The example below is a synopsis of my Marketing Analytics course (MAR - 653) project. Knowledge to visualize data in this method came from previous work experience outside of the program. I have ~2 years of experience in Tableau. The Introduction to Data Science (IST - 687) and Data Analytics (IST - 707) courses provided me with the necessary tools to familiarize myself with R and RStudio. These two classes (687 & 707) introduced me to ML algorithms. It is where I learned how to extract the most important features from a model to uncover clues about the process. In this project, the raw text files were imported from Kaggle into RStudio and cleaned/analyzed to produce the following analysis.

## Problem to be Solved

### Data Questions

* What are the top three parameters in predicting a country’s happiness score?
* Which continent is the most/least happy?
* Which countries have improved the most in those areas from 2015-2017?
* Which countries have fallen the most in those areas from 2015-2017?

### Background

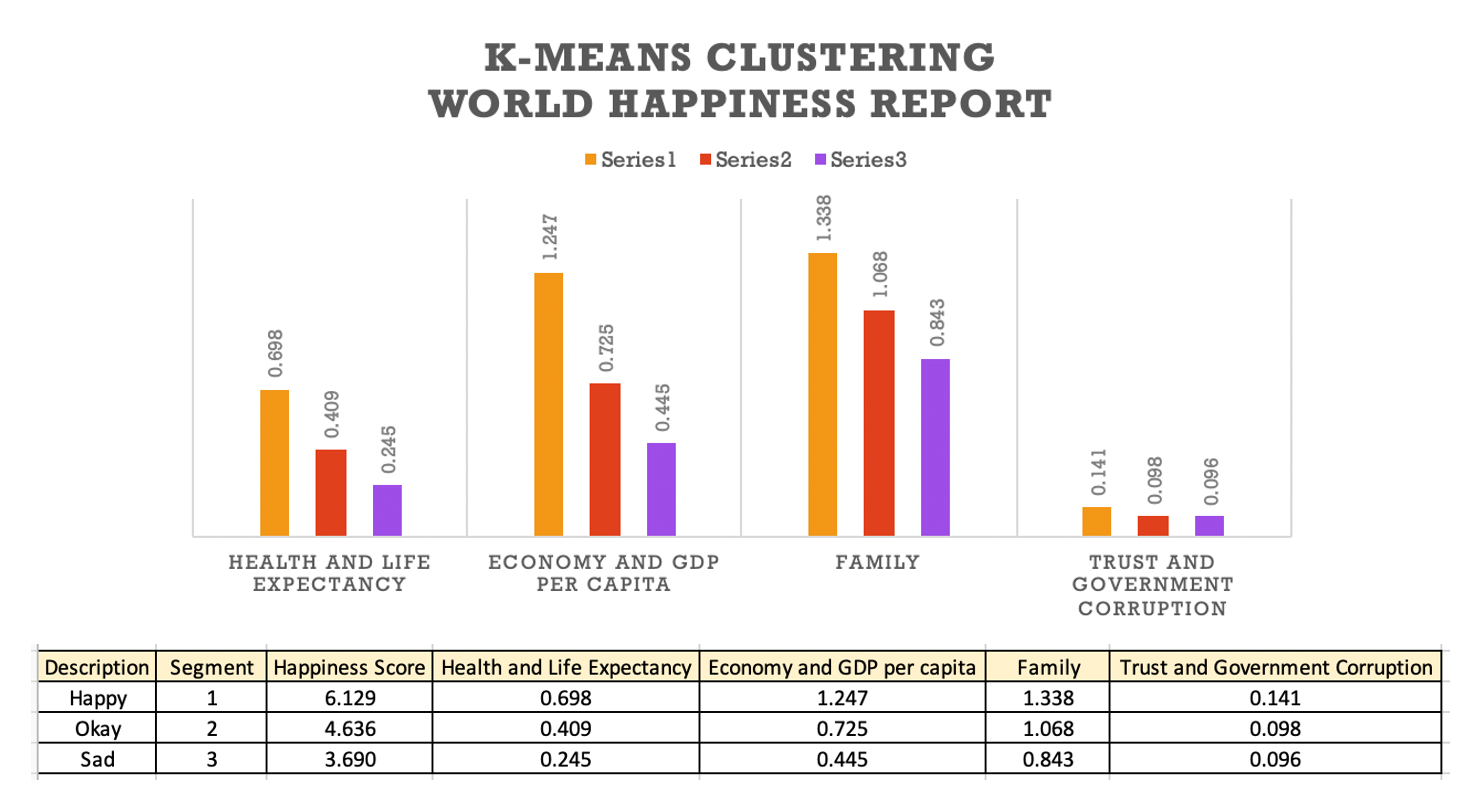
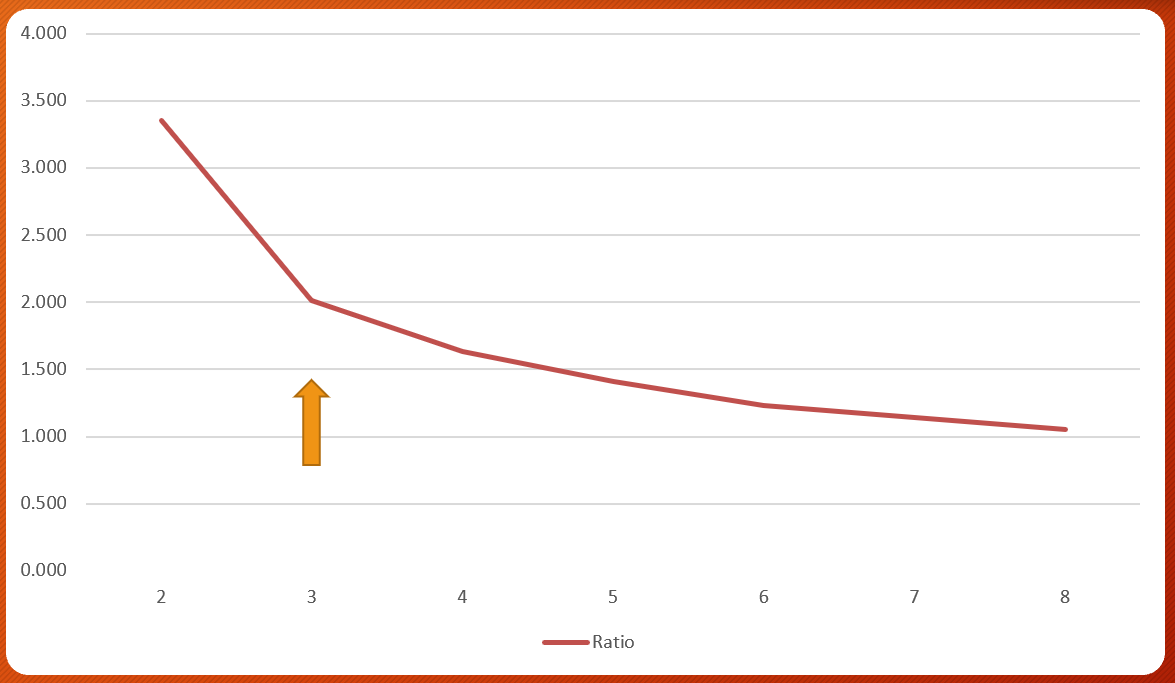
The World Happiness Report is a landmark survey of the state of global happiness that ranks 156 countries by how happy their citizens perceive themselves to be. This survey was conducted using the Cantril ladder. Respondents were asked to think of a ladder, with the best possible life being a 10 and the worst a 0. They were then asked to rate their own current lives on that scale.

Last year, the U.S. was ranked 18th on the Happiness Report with a negative trend from 2015 – 2017 despite GDP rising. Pursuit of happiness is a fundamental human goal; this project aims to analyze the world happiness data to determine which key factors promote overall happiness and life satisfaction.

We believe there are specific customer segments, that would be interested in traveling to the happiest countries in the world. Based on this analysis, we will identify the happiest countries in the world that will be popular routes/destinations for the travel and tourism industry.

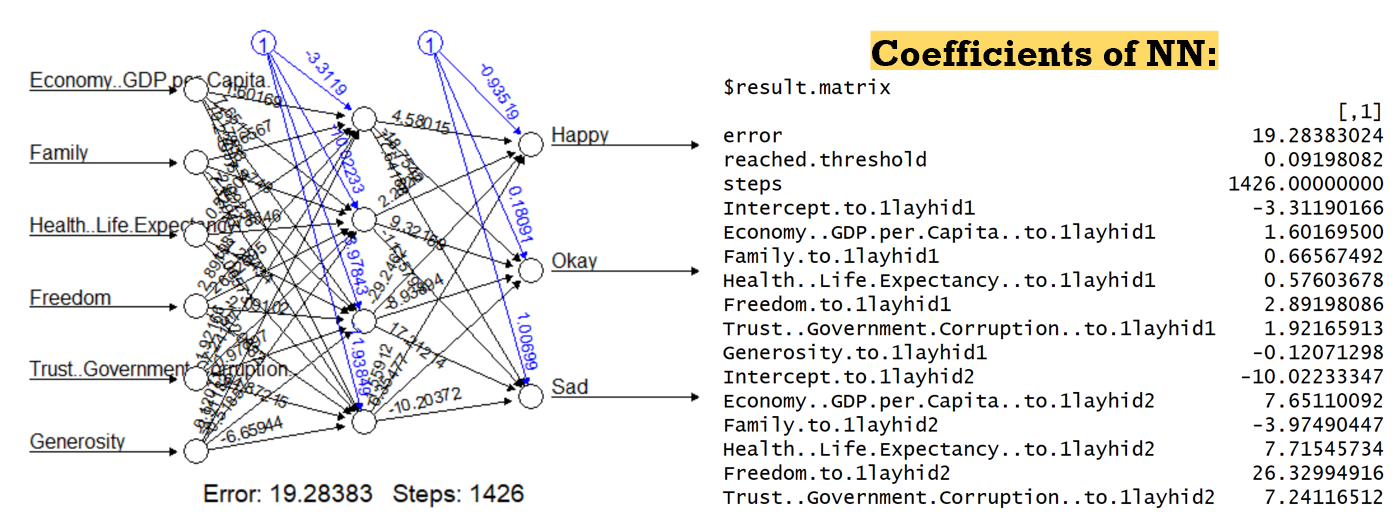
## Tools & Techniques Used

Various tools were used to generate 3 different types of predictive analyses (K-means clustering, Neural Network Classification, and Random Forest Decision Tree). The raw data was retrieved from Kaggle as three comma-separated values (CSV) files, each containing a single year from 2015 - 2017. Excel was used to combine all three CSV files. The K-means clustering algorithm was implemented on the 2017 CSV to group the most recent happiest countries. To determine the number of clusters required an elbow plot was generated using XLStat, plotting over the number of clusters. As depicted in the *Figure 3.* 3 or 4 clusters appears to be optimal. The group decided on 3 clusters (Happy, Okay, & Sad) for better computational feasibility.

 *Figure 3. Variance vs # of Clusters Table 1. 3-cluster Happiness Report*

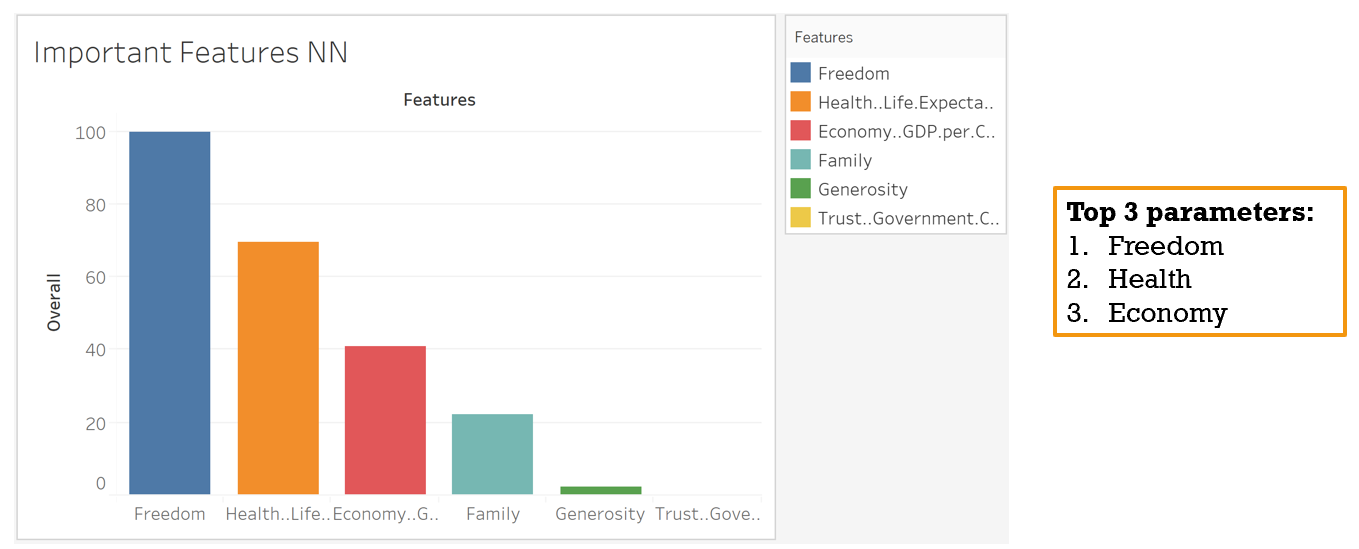
An Excel pivot table was generated to produce *Table 1.* All countries had the same relative preference for different features with family, economy, and health being the top 3 rated attributes. The Happy countries (orange) received the highest marks, then Okay countries (red), and lastly Sad (purple) countries received the fewest points.

Next, a sensitivity analysis was done on a Neural Network (NN) in RStudio to predict the most import features for determining if a country was Happy, Okay, or Sad. This algorithm utilizes logistic regression to represent nonlinear behavior and is best suited for classification problems.



*Figure 4. Neural Network Map and Coefficients*

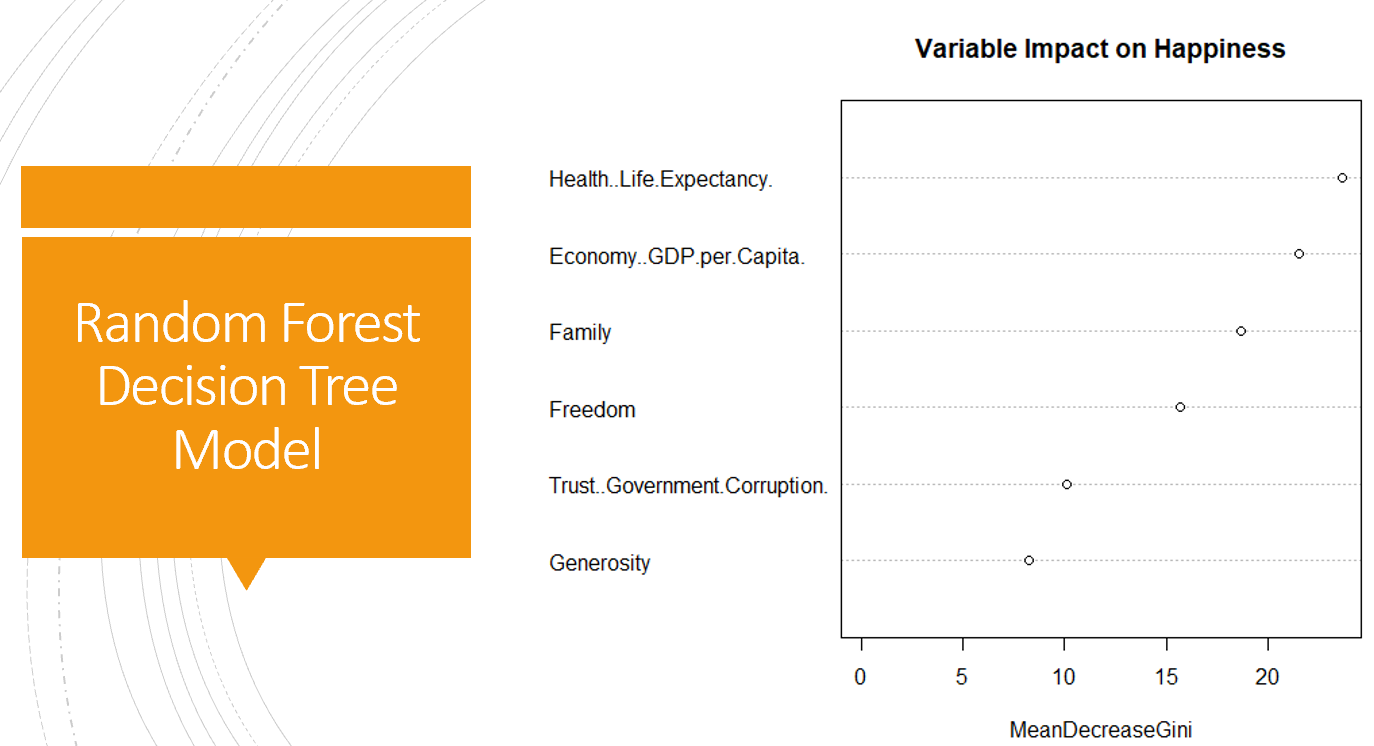
All data (3 years) was input into the algorithm and the sensitivity analysis produced in Tableau with the following results:



*Figure 5. Top Rated Features in Neural Network*

Freedom, Health, and Economy were the top 3 most important features when determining classifications in the NN.

Lastly, a Random Forest (RF) model was trained in RStudio on the same dataset with the most important features being: health, economy, and family.

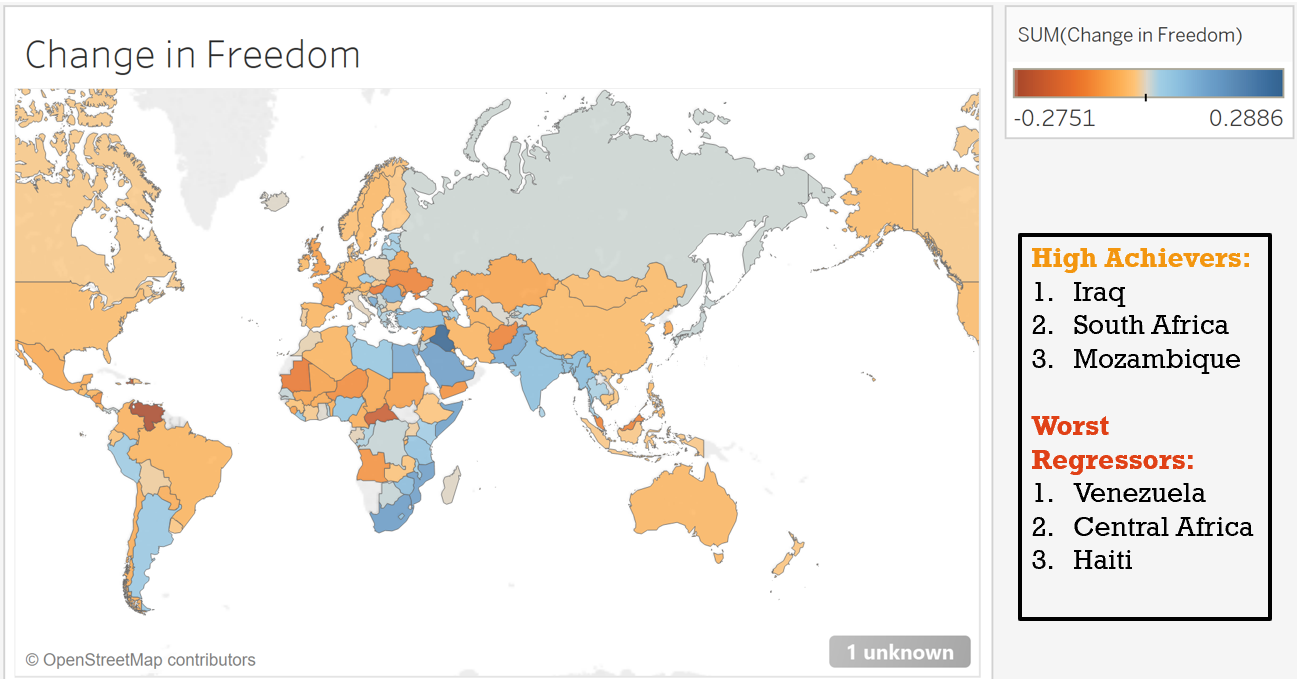


*Figure 6. Top Rated Features in Random Forest*

## Insight Gained

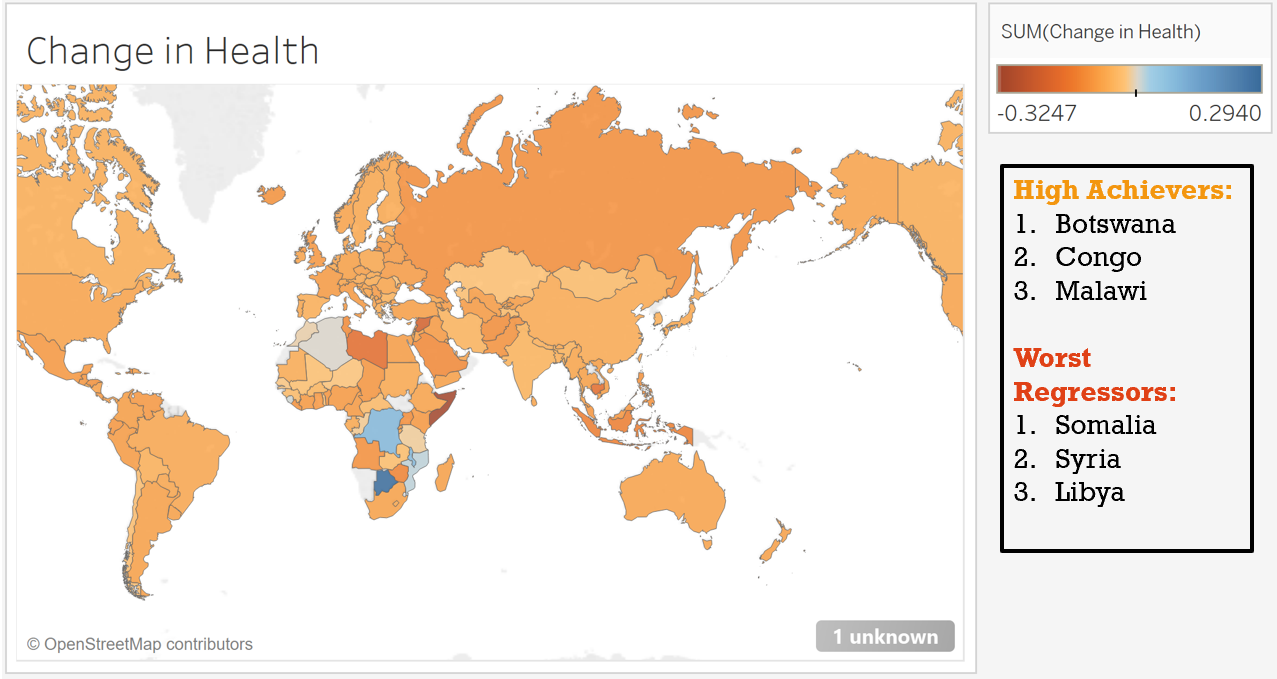
In conclusion, the top three parameters for predicting a country’s a happiness score are: family, health, and economy. The region that is the happiest in 2017 was Western Europe primarily due to the Scandinavian countries, and the least happy was Africa.

Viewing which countries had the largest differences in either direction (good or bad) in the top three sectors should provide deeper granularity on what factors make a country’s population happy.



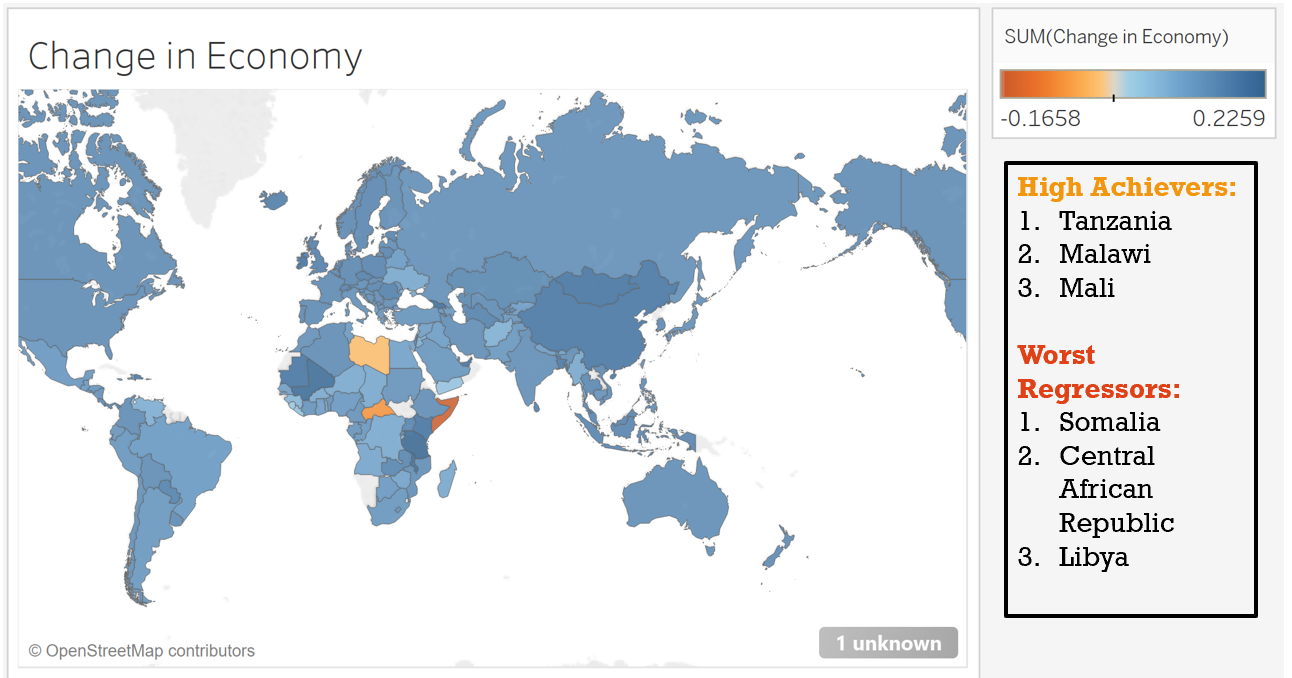
*Figure 7. Change in Freedom by Country (2015 – 2017)*

The countries with the best positive swing in freedom (High Achievers) from 2015 – 2017 were: Iraq, South Africa, and Mozambique. This aligns historically as the Iraqi war ended in December 2017. As the war began to cool and ultimately cease it left the country with a sense of freedom it has not felt since before the conflict. Venezuela, Central Africa, and Haiti had the worst negative swings in freedom (Worst Regressors). All 3 of these countries were dealing with government corruption and threats of dictatorship during this time span.



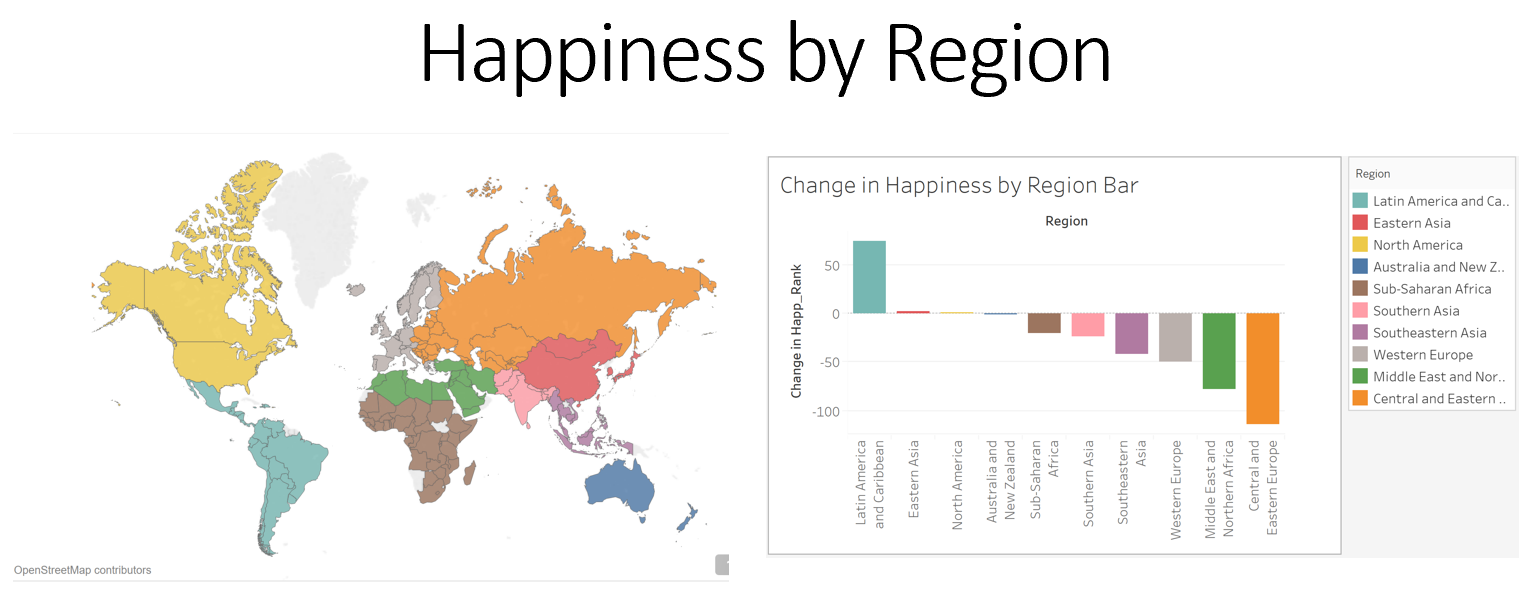
*Figure 8. Change in Health by Country (2015 – 2017)*

The High Achievers in health were African countries critically affected by the HIV/AIDS epidemic. These countries, especially Botswana because of their economic prosperity, saw major improvements to their health due to the provision of universal free antiretroviral treatment (ART) to all people living with HIV. This casted a ray of hope over these disease-stricken areas, accounting for the uptick in optimism. The Worst Regressors are countries that had a lot of political upheaval and national disturbance. The chaos caused a much higher rate of violence and rape, especially to refuges who were trying to flee those countries.



*Figure 9. Change in Economy by Country (2015 – 2017)*

Again, the biggest swings are in African countries. All 6 of these countries are highly dependent on agriculture. The High Achievers saw higher rainfall boost crop production in 2017, whereas the Worst Regressors saw a decrease in crop production and livestock due to drought.



*Figure 10. Changes in Happiness by Region (2015 – 2017)*

Overall, the biggest swings in Happiness from 2015 – 2017 were in Latin America and Central/Eastern Europe. Latin America showed the highest improvement in happiness, despite being threatened by multiple dictatorships, proves that happiness is a mindset and not about personal possessions.

# Develop Alternative Strategies based on the Data

Assuming alternative strategies means being forced to dive into a different method of analysis due to subpar performance of the initial process, there is one project that truly exemplifies this. In the Big Data Analytics (IST - 718) course’s final project my group decided to predict which team would win the Super Bowl based on regular season performance. A myriad of classifiers were used, but they all gave nonsensical results. The reasoning for this behavior was due to the constrictive nature of football playoff outcomes. Only 1 team can be the Super Bowl champion, only 1 Super Bowl loser, only 2 teams can lose in their respective conference championships, etcetera. These added constraints were never addressed in the initial program, however the ability to search the internet for troubleshooting and model adaptations is highly stressed throughout this program. Armored with previous exposure of model and function building from my time in this program, an entirely new method of analysis came to fruition. This different strategy created a model that was not only more accurate but realistic as well.

## Problem to be Solved

### Data Questions

1. Can we effectively predict the winner of the Super Bowl?
2. What combination of skills, performance, and coaching maximizes the chances of winning the big game?

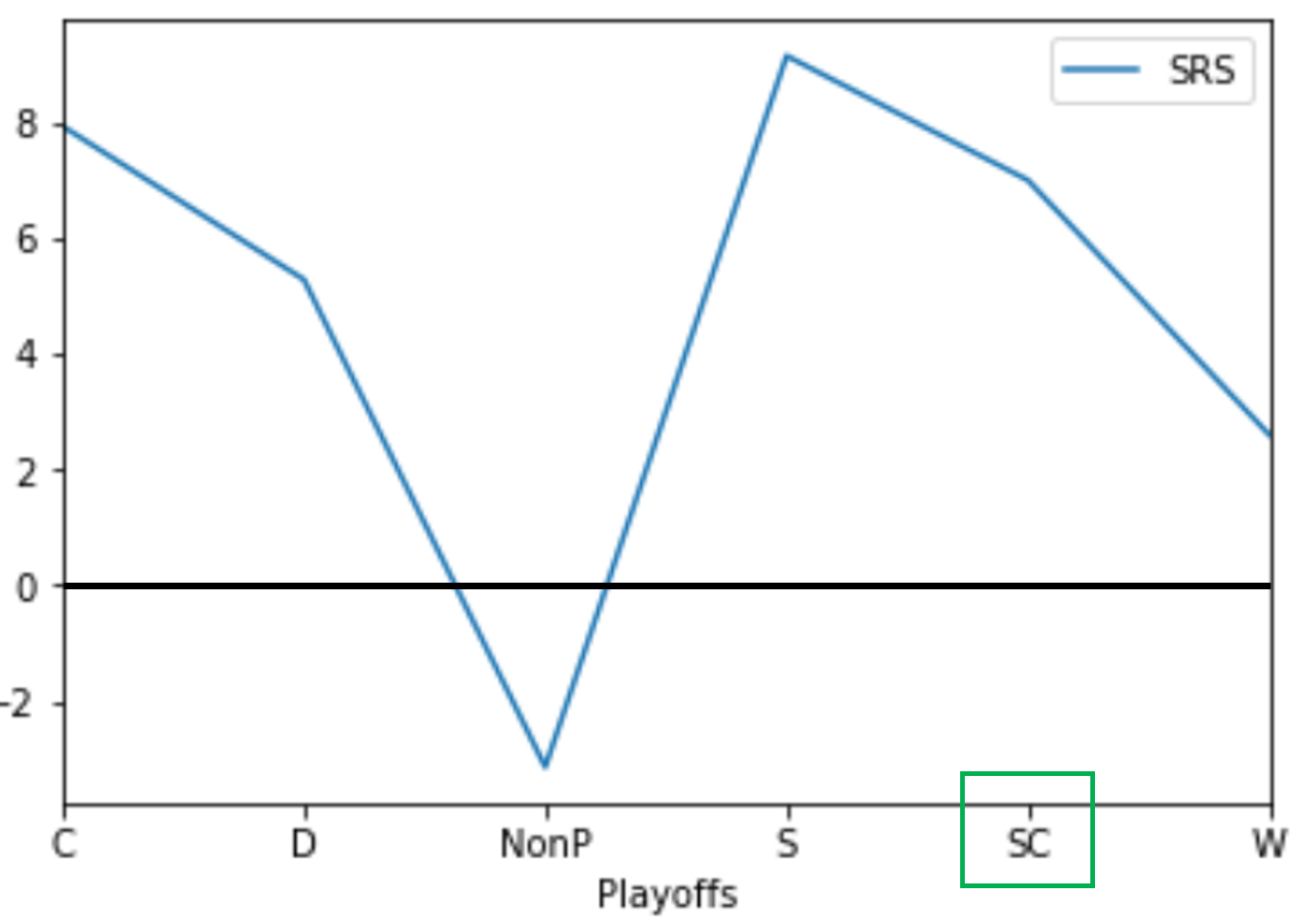
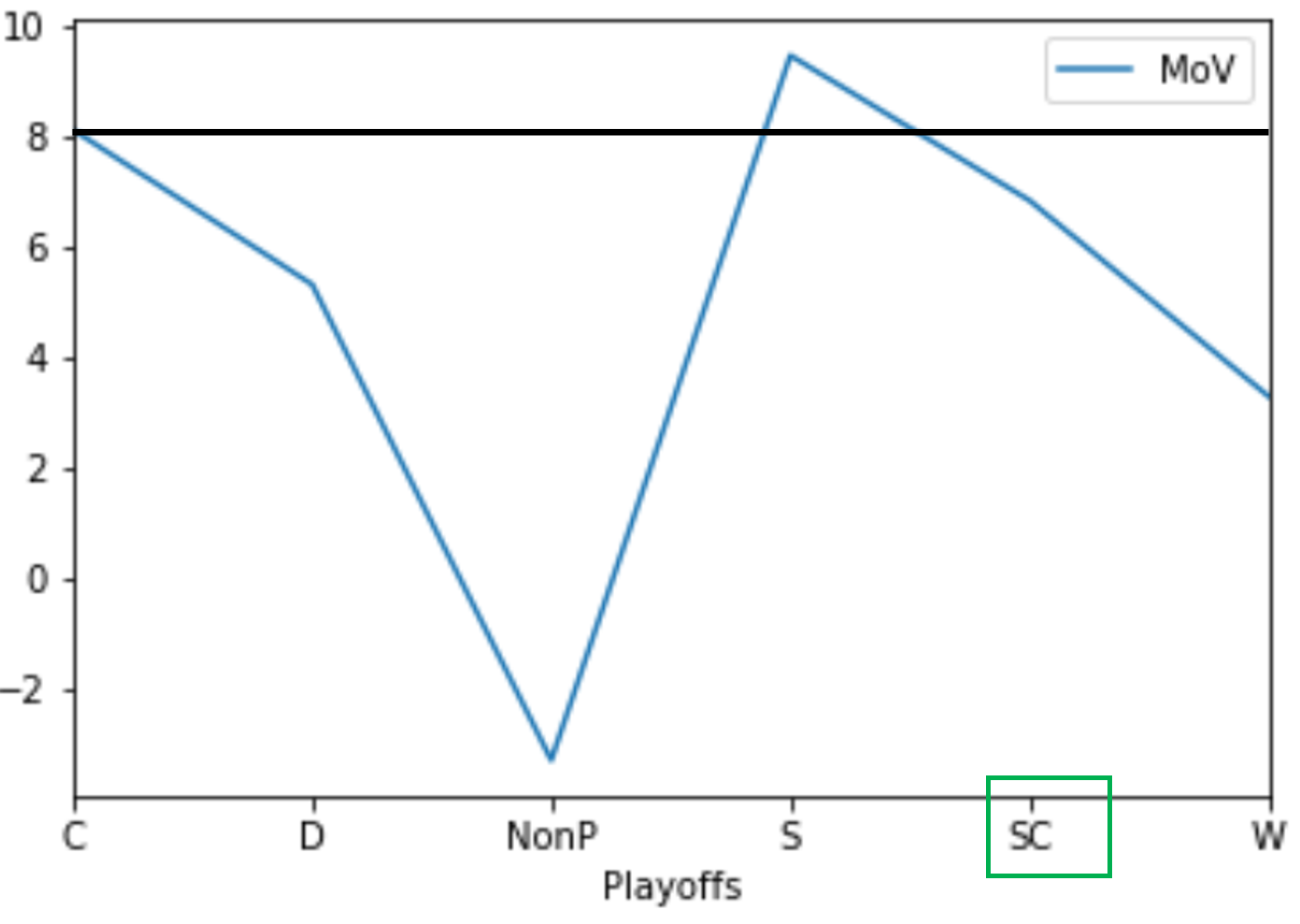
### Background

Hailed as America’s most popular and wealthiest professional sports league, the National Football League (NFL) provides a source of income, prestige, and pride to their communities. The NFL generated over $15 billion in revenue during the previous season (2018 – 2019), including over $450 million in ad revenue from the Super Bowl alone. The winner of the final showdown also receives a bonus of over $5 million. With so much money flowing into this sport there is a ton of incentive for each organization to put the best product on the field. But what exactly is that? What combination of skills, performance, and coaching maximizes the chances of winning a Super Bowl? What is the best winning formula?

My team will seek to crack the code of what constitutes a Super Bowl winner. Armed with this knowledge any team can go from worst to first and take home a bigger slice of the NFL revenue pie while basking in all the glory and immortality a Super Bowl championship brings to their city.

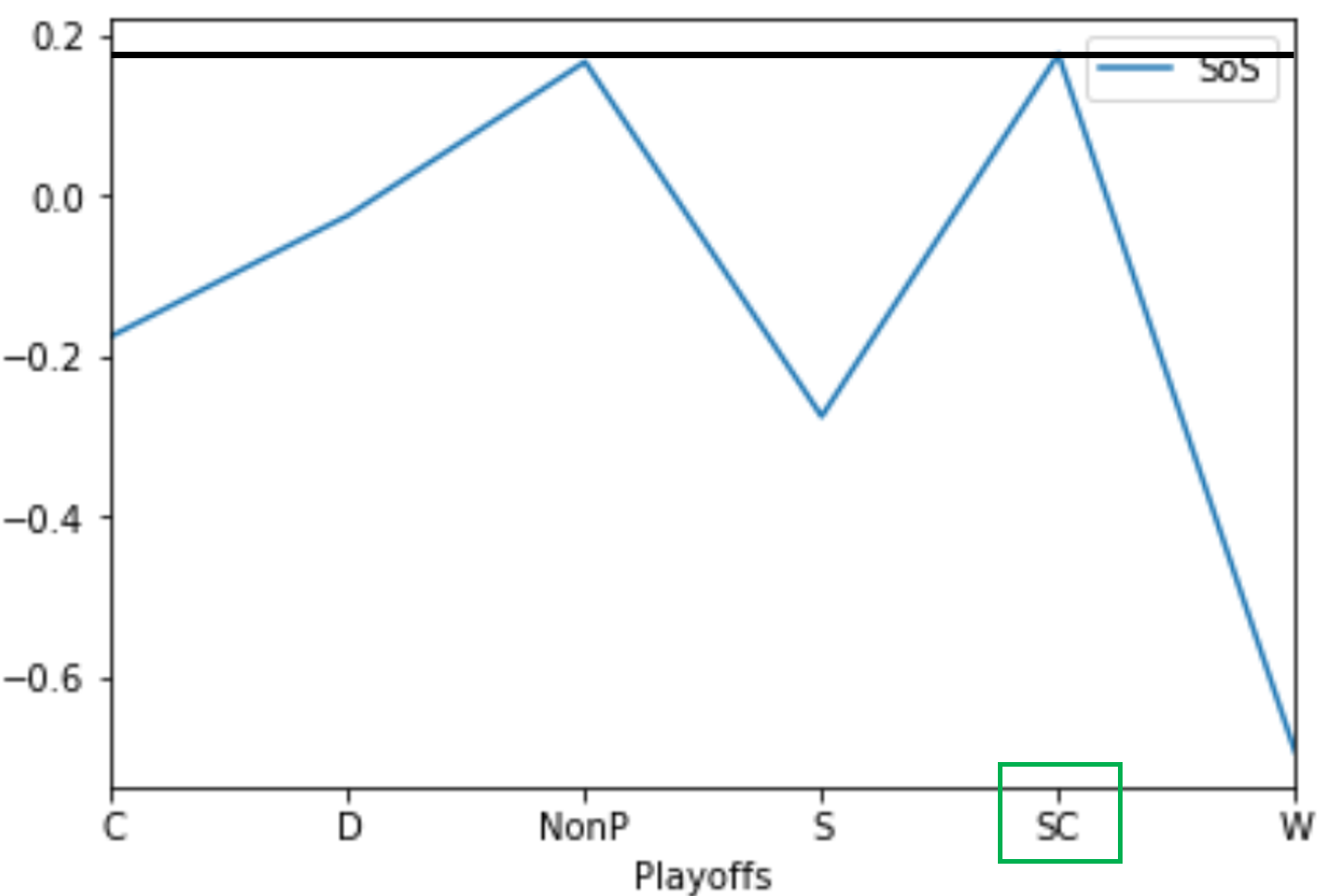
## Tools & Techniques Used

Over 30 offensive and defensive metrics from all 32 NFL teams over the past 9 years were collected to determine if there is a winning formula for surviving the regular season, playoffs, and ultimately winning the big game. The offensive/defensive metrics were taken from Pro Football Reference (www.pro-football-reference.com) and exported to a CSV. All final team rankings over the past 9 years were taken from ESPN ([www.espn.com](http://www.espn.com)). Team names and year were used as the primary key to join the two datasets. Models were developed in Python using a Jupyter IDE. The models were trained to predict the Super Bowl winner and the playoff results of all teams based off their regular season statistics. The team with the best overall statistics was initially hypothesized to be most likely to win the Super Bowl. However, there was a myriad of interesting statistical anomalies which went against that conventional way of thinking. The most interesting facts were centered around the conclusion that Super Bowl champions (SC) were not the most statistically dominant in the regular season. In fact, the most statistically dominant teams lost in the Super Bowl (S) (*Figure 11 & 12*).

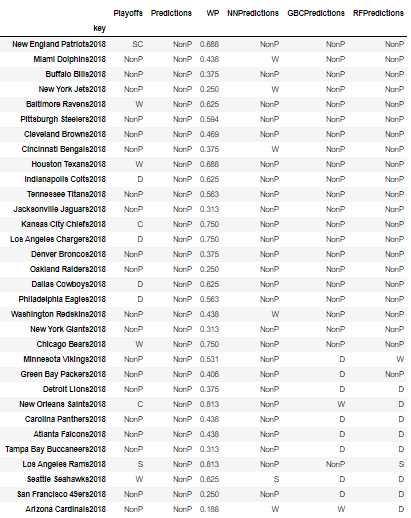
*Figure 11. Team’s Rating vs Final Playoff Standings Figure 12. Margin of Victory vs Final Playoff Standings*

There was one category the Super Bowl champions did dominant, Strength of Schedule (SoS). As shown in *Figure 11.* the eventual Super Bowl champion usually had the most difficult regular season schedule, comparable to teams that did not make the playoffs. The team that won it all played more games against playoff teams in the regular season, which prepared them for the higher level of performance demanded to excel in the playoffs. The nonplayoff teams are probably not in the playoffs because facing those high-quality teams resulted in too many losses. Therefore, having exposure to high quality teams in the regular season enhances a team’s probability of beating quality teams in the playoffs. However, they must be careful to not bite off more than they can chew, or they can ruin their title run before it starts.



*Figure 13. Strength of Shedule vs Final Playoff Standings*

A myriad of unique models were built to find the best predictor: Support Vector Machines, Neural Network MLP Classifiers, Gradient Boosting Classifiers, and Random Forests. Initial results are shown below:



*Table 2. Model Predictions*

The initial models gave horrendous results overall. Some performed better than others, but they all were highly inaccurate and unrealistic. This is the first time in the Applied Data Science program the model could not classify any team into any desired class. There must be limitations!! In normal classification problems there is nothing wrong with a model predicting every observation to the same class, but here, there are a finite number of examples that can be assigned to each class. Only one team can be predicted as the Super Bowl winner, only one runner up, etcetera. This limitation created the need to think of prediction in a different way.

In the next iteration of models, the GBC and MLP Neural Network were selected due to their high performance in the initial phase. A loop was constructed to iterate through every team in the 2018 season and predict the probabilities of each team receiving each classification. If the classification had not been taken by another team, the team was given that prediction. If the classification had been taken, the teams 2nd highest class probability was used. If the 2nd highest classification was taken, their 3rd highest probability was used, and this process repeated for all classifications.

This iterative process produced a model that would properly classify teams in a manner similar to how the actual NFL playoffs unfold. This resulted in models that were more realistic and had vastly improved accuracies.



*Table 3. Model Predictions (Part 2)*

Nineteen of the twenty non-playoff teams were correctly classified in the GBC model (*Table 3.*). Only a few of the playoff teams were correctly placed, however the classifications were never more than +/- 1 round of the playoffs off. The only exception was the New England Patriots who statistically were not a tremendous team in 2018 and were predicted to lose in the Wild Card, but won the Super Bowl. The most difficult teams for the models to place were between the Wild Card and Divisional rounds of the playoffs. Concluding that the teams losing in the first two rounds of the playoffs are very similar statistically.

## Insight Gained

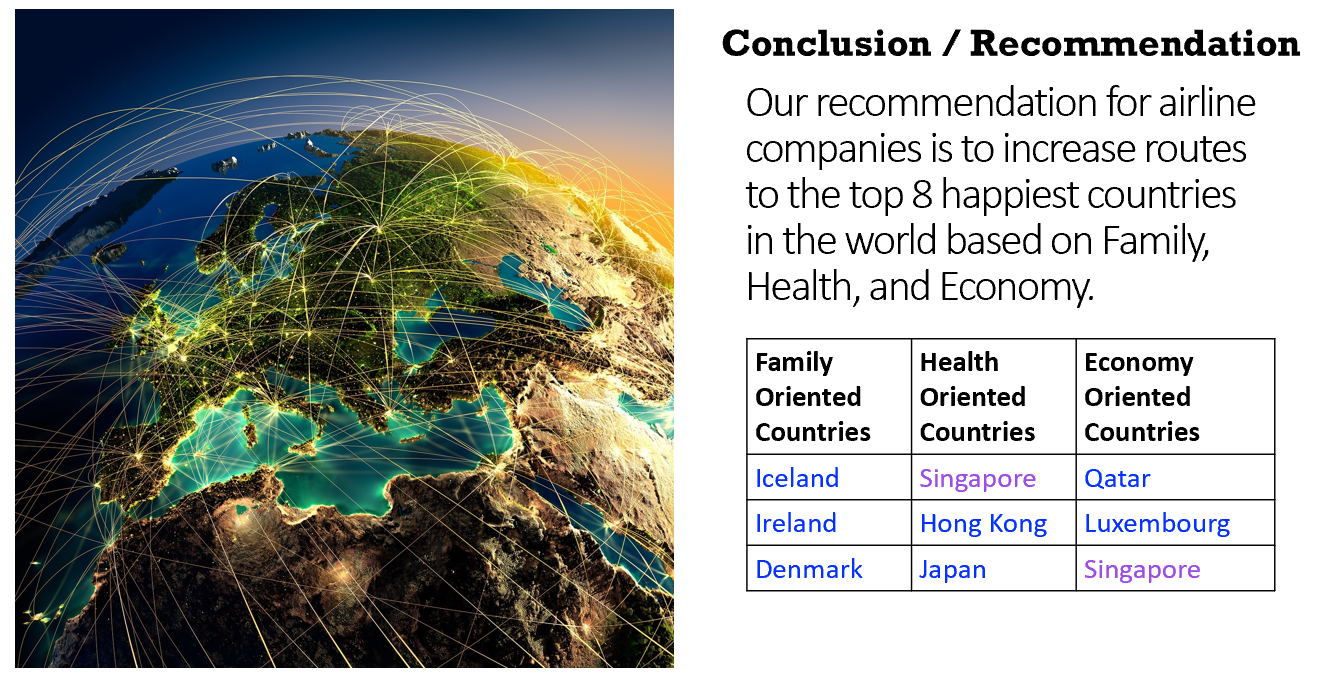
This model can be used to project a team’s playoff success. Teams certainly need to believe that they can win every game, but this analysis can give a breakdown of how the playoffs will unfold which is beneficial in determining their most difficult matchups. Teams can also look at their statistical profile and see where the model predicts they will finish, allowing the team to create various points of emphasis in their game-plans to overcome their deficiencies.

# Develop a Plan of Action to Implement the Business Decisions derived from the Analyses

Learning objective has been mastered. All projects and assignments undergone in this program have honed my skills in understanding what data means and how to develop a business plan of action to take advantage of those results. An example already given was from my Scripting for Analytics class. Once the Retweet per Tweet attribute had been identified as the best predictor to influence recruiting rankings, the next step was for the school (business) to build up their social media infrastructure/platform to better relate to the younger population. This Plan of Action was specifically targeted to increase the social media buzz of a school to affect how many top football recruits their team received.

# Demonstrate Communication Skills regarding Data and its Analysis

Learning objective has been mastered. All projects and assignments undergone in this program have honed my skills in communication regarding data analysis. An example already given was from my Marketing Analytics class. Once the top attributes that constituted a happy country were established, the original dataset was used to discover which countries ranked highest in those areas. From there, it was recommended that airline companies increase routes to the top 8 happiest countries and develop a marketing campaign centered around experiencing the happiest places on Earth.



*Table 4. Top Destinations for Happiness*

This learning objective was also demonstrated when viewing which countries made the biggest leaps (good or bad) in happiness ranking. The data was analyzed, and I was able to effectively communicate solid reasons as to why those particular countries experienced an increase/decrease in their happiness.

# Synthesize the E**thical Dimensions** of Data Science Practice

I had much less exposure to this learning objective throughout this program. Very few times were the ethical dilemmas of data science brought up in the classes I was exposed to. Although I was not directly challenged by project requirements to talk about ethical dilemmas an example of this can be seen in my Scripting for Analytics project. As previously mentioned, this project looked at finding a way to increase the number of top football recruits to go to a school. The results proved that having an active social media feed, specifically on Twitter, that generated a lot of retweets per tweet (your followers shared your content on their pages) was a good indicator of final recruiting rankings.

A question can be posed about the moral ethics of this solution. Is manipulating kids to pick a particular school crossing the moral boundary? Is this moving toward the slippery slope of behavioral mapping and control? Where a company can induce any targeted customer into buying their product if they have enough information on them. Is this stripping away free will, or at least the perception of free will? My personal answer to this is yes. By understanding how humans make decisions, companies are able to highjack the decision-making pathways in the brain to make their products seem more desirable. Given the amount of personal data a company can collect on a subject (purchase data from social media, email, and app subscriptions) they can tailor an optimized marketing campaign. As data becomes more and more available, companies will begin to do this with more accuracy and veracity. I believe the day will come when we have to decide between the ease of the latest computer amenities and our own free will. While this feels like a post-singularity proclamation, it is not. Companies like Costco and Target are already using big data to build purchasing profiles of their customers. They group customer behavior based on region and understand what they like to purchase, what they look for, and what they are attracted to based on their impulsive buys. The store then orients the shopping experience in such a way where customers have to travel through a gauntlet of their strongest temptations in order to purchase what they initially came in for. Is this an invasion of privacy and personal information, or strategic marketing? Is there a difference? These questions are ones that cannot be taken lightly and with so few rules in place to monitor how personal information is used by companies this method of targeted marketing will only increase.