

Portfolio Milestone

Rohini Shrivastava



summer 2021

syracuse university

**TABLE OF CONTENTS**

1. **INTRODUCTION**
2. **BROAD OVERVIEW**
3. **IST652 (SCRIPTING FOR DATA ANALYTICS)**
   1. **DATA COLLECTION AND CLEANING**
   2. **FINDINGS**
      1. *Figure 1: Maps of Happiness levels in 2019 and 2021*
      2. *Figure 2: Most and Least Happiest Countries 2019 and 2021*
      3. *Figure 3: Correlation Matrix of Variables between 2019 and 2021*
      4. *Figure 4: Sentiment Analysis on Each category of tweets*
   3. **REFLECTIONS**
4. **IST736 (TEXT MINING)**
   1. **DATA COLLECTION AND CLEANING**
      1. *Figure 5: Sample set of vectorized dataframe*
   2. **FINDINGS**
      1. *Figure 6: Model Performance Accuracy for 10-Fold Validation*
      2. *Figure 7: Models with the 22 Authors*
      3. *Figure 8: Ratings and Sentiment of Each Author*
   3. **REFLECTIONS**
5. **IST687 (INTRODUCTION TO DATA SCIENCE):**
   1. **DATA COLLECTION AND CLEANING**
      1. *Figure 9: Old and New Renamed Platforms*
   2. **FINDINGS**
      1. *Figure 10: Video Games Over the Years*
      2. *Figure 11: Sales per Region vs Global*
   3. **REFLECTIONS**
6. **CONCLUSIONS**

**INTRODUCTION:**

Syracuse University’s Applied Data Science Master’s program is an interdisciplinary program that lets students learn the broad scope of data science. There are six main things students in this program were taught:

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

**BROAD OVERVIEW:**

Data science is an interdisciplinary field that uses various scientific methods, processes, and algorithms to get information from structured and unstructured data. Data science uses Artificial Intelligence, Machine Learning and Deep learning to help extract insights of data. Many organizations are using data science to refine their goods and services.

**IST652 (SCRIPTING FOR DATA ANALYSIS)**

In IST652 (Scripting for Data Analysis) taught by Dr. Debbie Landowski, I collected data through various online sources, such as Kaggle and Twitter, and learned how to script in python. The final project I did for this class was to determine what factors contribute to the happiness of countries and how covid may have affected the happiness. The word ‘happiness’ is used to categorize the mental states of a person. It is related to positive emotions like contentment to euphoria. It is hard to quantify exactly how people feel happiness and what factors cause these mood changes. However, through research and dedication there have been studies done to see how each country can try to appeal to its citizens. The questions I wanted to explore were the following:

* Who is the happiest? Who is the least happy?
* What is the strongest variable on happiness? What is the least influential variable?
* How has COVID impacted the variables associated with happiness?
* What are people’s overall sentiment in regards to COVID? What about the report itself?
* How have people’s feelings changed between 2019 and 2021?

*Data Collection and Cleaning*

Data was collected through Kaggle, World Happiness Report, and Twitter. Kaggle is a website that allows users to find/upload data sets to build models and enter competitions with other data scientists. The World Happiness Report website ranks 156 countries based on citizens responses. It uses various factors like GDP, Mortality, etc. to show the influence they may have on happiness. The last set of data was taken from Twitter using Twint. Twint is a twitter scraping tool that allows users pulls tweets without using an API. The hashtags ‘#covid’, ‘#WorldHappinessReport’, ‘#2019’, ‘#2021’ were taken from twitter.

The 2021 world happiness report had 20 columns while the 2019 report only had 9 columns. The excess columns were dropped from the 2021 dataframe, whr2021, and the remaining columns were renamed to align more closely with the foundational column names. In the 2019 dataframe, whr2019, there was an extra unnecessary “overall rank” column that was also dropped to reduce the file down to more similarities. The 2019 data did not contain regional data therefore to compare the North America and ANZ region we had to create that region by grouping those variable rows into their own named dataframe for comparison.

The tweets pulled from twitter were placed into a pandas dataframe. Only tweets written in English were kept. Within the tweets, words that contained the hashtag were removed. Examples of this would be words like ‘COVID19’, ‘WorldHappiness’, ‘Report’, etc. Along with removing related words and stop words, any links inside the tweets were removed as well.

*Findings:*

I compared the scores between 2019 and 2021 (*Figure 1).* Through the map I was able to see that the work happiness levels looked like they increased overall. Countries in Asia and Africa are now a closer to the dark green in 2021 than compared to 2019. However, we are also able to see that more countries in Africa did not participate in the survey. Countries not participating are shows with a grey color.

A screenshot of a map

Description automatically generated with medium confidence

*Figure 1: Maps of Happiness levels in 2019 and 2021*

Finland was the happiest country in both 2019 and 2021. South Sudan was the least happy country in 2019 and Afghanistan was in 2021 (*Figure 2).* Many European countries are present in the happiest list with high life expectancy and freedom to make life choices. Graphical user interface, application

Description automatically generated

Graphical user interface, application, table

Description automatically generated

*Figure 2: Most and Least Happiest Countries 2019 and 2021*

Linear regression and a correlation matrix was used to compare the variables with each other to find the most and least influential variables on happiness. The correlation matrix was done for 2019 and 2021. It showed that GDP really had no change in effect. Social support, healthy life expectancy and generosity went from negative to positive. Freedom to make life choices and perception of corruption decreased in strength (*Figure 3). A picture containing treemap chart

Description automatically generatedFigure 3: Correlation Matrix of Variables between 2019 and 2021*

The linear regression that was ran used this correlation matrix to then go and try to predict a score based on multiple variables and compare how much weight they contribute to the equation of happiness. These models are to help further analyze which variables are worth looking at when comparing country happiness, they will not be looked at to solve for a specific country's happiness. Originally five variables were looked at: GDP, Social Support, Life Expectancy, Freedom to Make Life Choices, and Perceptions of Corruption. After running the models I found Social Support had a low p value showing its significance and the strongest correlation the Ladder score. Generosity shows very little correlation and is considered the least influential variable.

Lastly, a collection of tweets were taken through Twint and a VADER Sentiment analysis was run on them. Overall, all the categories had a more positive sentiment. The ratio between positive and negative tweets was the highest out of all the searches I made for tweets containing “World Happiness Report”. 2021 also seems to be more negative than 2019. While both categories had more positive tweets than negative, 2021 had a smaller gap between the number of each compared to 2019. 2021 also was the closest overall in the number of positive and negative tweets seen.

Chart, bar chart

Description automatically generated

*Figure 4: Sentiment Analysis on Each category of tweets*

*Reflections*

Collecting data and organizing it is important to answer questions and make predictions about trends. I was able to use various sources to gather data, such as csv files from Kaggle and the World Happiness Report website, and twitter. Through the collection of this data I was able to see the trends in countries and how happiness changed over time. The data taken from Kaggle and the World Happiness report website allowed me to see the correlation between each of the variables and run further tests to see the influence they had on happiness. Understanding what variable contribute to happiness can help leaders of the countries know what to focus on to make its citizens happy. However, happiness only means different things to different people. If leader only focus on certain aspects and ignore others it can detriment the lives of some citizens.

**IST736 (TEXT MINING):**

IST736, taught by Dr. Ami Gates, introduces concepts and methods for knowledge discovery from large amounts of text data. It also goes through the application of text mining techniques for business intelligence, digital humanities, and social behavior analysis. The final project I completed for this class centered around gothic horror authors. Writers of all genres have similar ideas or premises in their works. In the case of gothic horror, this can be based on the macabre subjects and topics of the writings and overlap is anticipated. By analyzing a subset of these authors’ works and applying different modeling techniques, there may be trends or styles that may allow a computer trained model to believe the work was written by someone else.

We asked the following questions:

1. Is it possible that the writing style of these four core authors influenced not only each other, but additional writers of the era?
2. How do the audience feel about these authors?

*Data Collection and Cleaning:*

Using Project Gutenberg (<http://gutenberg.org>), we looked up works by four authors: Edgar Allen Poe, Bram Stoker, M.R. James, and Algernon Blackwood. 65 books were taken from the website with each other having between 15 and 17 works to create a balanced dataset. Each work was downloaded in plain text format and placed into an individual text file. I wanted to test the four authors against others in their genre as well to see how much influence there was, so an additional 22 authors’ works were imported as well. From GoodReads 15 reviews were taken for each author.

After getting all the works, each text file was cleaned up. All words were lowercased, punctuation and special characters were removed, and the Gutenburg information were removed as well. Once the text had been cleaned up, all stop words were removed and the remaining words were vectorized and put into a dataframe. Table

Description automatically generated  
*Figure 5: Sample set of vectorized dataframe*

*Findings:*

Five different models were run to predict the author of each of the 65 works. The following models were used with the vectors created with *CountVectorizer and TfidVectorizer:*

* Multinomial Naïve Bayes Classifier
* Bernoulli Naïve Bayes Classifier
* Support Vector Machines (SVM) with a Linear kernel, cost value 1
* Support Vector Machines (SVM) with the Radial Basis Function (RBF) kernel, cost value 100
* Support Vector Machines (SVM) with the Polynomial kernel, cost value 100

A training set with 75% of the data and a test set with the remaining 25% were created and used. A 10-fold cross validation was run for each of the models (*Figure 6).* From the table the best method was Support Vector Machines with a Linear kernel and cost value of 1. The mean accuracy was 90% using *CountVectorizer* and 93.33% using *TfidfVectorizer*. The least affective was the Support Vecor Machine with Polynomial kernel and cost value of 100.

Table

Description automatically generated

*Figure 6: Model Performance Accuracy for 10-Fold Validation*

We ran the model again, but this time against the other 22 authors to see who’s style matched the most. There was also one work from each of the original four authors. The Multinomial NB model predicted the four original authors perfectly. The model also shows that many of the other 22 authors used similar vocabulary to that of Blackwood and Stoker (*Figure 7).* The Bernoulli NB model was not as accurate and predicted most of the works to be written by Poe.

Chart, calendar, bar chart

Description automatically generated  
*Figure 7: Models with the 22 Authors*

We also looked at GoodReads reviews of the author to determine how readers felt. The fifteen reviews were sorted by rating and sentiment. The ratings were between 1 and 5, with 5 being the highest. Blackwood and James were rated the highest.

It was important to see how readers feel about these authors. Fifteen reviews were found for each of the original authors and reviewed for ratings and sentiment.

As indicated on the left - the ratings were widely spread from 1 to 5 (5 being the highest); however, Blackwood and James were clearly higher rated overall than the other two authors. Then the sentiment for each review was found using VADER. The sentiment was then plotted with green being positive and red being negative. As shown in the GoodReads rating, Blackwood is at the top with Poe and James coming next. As expected, as the rating number got larger, there was more positive sentiment in the review (*Figure 8).*

![Chart, bar chart

Description automatically generated]() ![Chart, scatter chart

Description automatically generated]()  
*Figure 8: Ratings and Sentiment of Each Author*

*Reflections:*

Text comes in as unstructured data and through organizing it, we can transform it into structured data for easier analysis. In this case I used text data collected to classify other works and run sentiment on the authors. I was able to find patterns in the data by seeing most used words after the vectorization of the text. This lets me see how authors may affect each other with their writing styles. Through the modeling techniques used, I was able to see that computers can accurately classify documents. I was also running sentiment on the authors and their works. Businesses can use this kind of sentiment analysis to see how customers feel about their products.

**IST687 (INTRODUCTION TO DATA SCIENCE):**

IST687 is taught by Dr. Mohammed Syed and teaches students how to collect, process, transform, analyze, and communicate about data through R. For the final project in this class I investigated video game sales to see how patterns and trends have changed over time. We can use this data to see what games, regions, genres have the greatest rate of return on investment. The video game industry is a very profitable business, especially with the current events occurring. It is projected is that over a billion people around the globe play videogames in one form or another. I looked at the questions below to explore the data:

1. What console released the most games overall?
2. What year had the most games released?
3. What genre of game was the highest in sales overall?
4. Which region has the most sales?
5. Which publisher published the most games?
6. What is the trend of global video game sales over the years?
7. What are the average sales per game released in a given year?
8. Can we create a model that predicts global sales by utilizing only one of the three regions?

*Data Collection and Cleaning:*

I acquired our dataset, Video Game Sales, by Gregory Smith on Kaggle. The dataset has 16,599 rows and 11 columns for a total of 182,589 datapoints*.* After importing the data into R we removed all the N/A. The “Rank” column was redundant as we can use the sales to rank each row, so that column was dropped. Various platforms were very specific, but only for a few records. It was decided to lump them with other platforms that were similar (*Figure 9)*

Table

Description automatically generated  
*Figure 9: Old and New Renamed Platforms*

Once everything was cleaned up and placed into a dataframe, I began the analysis to answer my questions.

*Findings:*

By sorting the dataframe, I was easily able to answer the first few questions. The Playstation 2 released the most games, and game sales peaked in 2009. Action games dominated in sales for each genre. Most games sold were in North America. Electronic Arts (EA) published the most games.

Average video game sales have gone down since the late 80s/early 90s, however we have seen an increase in overall video game sales and amount of games released. This may be showing that the market has become oversaturated. Video game sales peaked in 2008- which follows trend of the number of games sold by year as well (*Figure 10).*

Timeline

Description automatically generated  
*Figure 10: Video Games Over the Years*

I used linear modeling to see if I was able to predict with region was the best predictor for global sales. Three separate models were created for North America, Europe, and Japan. We assessed the model and fit to see if there was anything unusual. Overall, North America was the best predictor of global sales with the highest R-squared value at 88.60%. Japan was the worst predictor at 37.75% (*Figure 11).*

Graphical user interface, chart

Description automatically generated   
*Figure 11: Sales per Region vs Global*

*Reflections:*

The video game industry is worth 90 billion dollars as of 2020, so it is important for companies to understand their audience. Collecting the data from Kaggle was a good start to understanding the market. I visualized the data to see the changes and trends in video games through out the year. With statistical tests I was able to see how global sales can be affected by regional sales. This can help in making business decisions, such as where to market their game more. For business leaders it is helpful to know what types of games in high demand so that they can cater to the market.

**CONCLUSION:**

The collection of IST687 (INTRODUCTION TO DATA SCIENCE), IST736 (TEXT MINING), IST652 (SCRIPTING FOR DATA ANALYSIS) helped in successfully implementing the learning objectives. In IST652 I was able to collect data from different sources and create various visualizations to show what was learned through the dataset. Linear regression and sentiment analysis was run to understand the factors of happiness. The results were communicated on what country leaders should focus on. Ethically however, it is not a good idea to focus only on certain aspects that was looked on in the report as happiness is different to different people. IST687 focused on using Kaggle to get video game data to determine trends. Regression was also done for this to predict sales on a global scale. Business leaders can use this data to focus their advertising efforts. IST736 gathered books from the Project Gutenburg to analyze text data through mining. Different statistical tests such as sentiment analysis and SVM. The analysis was used to match writing styles between authors. Text prediction can be used for understanding customer reactions to products.

Being in an age where data is everywhere means that it is important to fully understand how it works. Syracuse University’s M.S. in Applied Data Science program teaches it’s students different aspects of Data Science. Through the curriculum students can organize, identify patterns, make business decisions, and understand the ethics behind data. The data science field has grown tremendously within the past few years, and it is important to recognize how data can be used to answer a wide range of questions. Through this program I was able to be learn and grow my abilities to be a better Data Scientist.

**CITATIONS:**

*What is data science?* Oracle. (n.d.). https://www.oracle.com/data-science/what-is-data-science/.