**Masters of Science**

**in**

**Applied Data Science**

**Portfolio Milestone**

**Syracuse University**

**School of Information Studies**

**Gilbert A Guyah | SUID: 963583327 | gguyah@syr.edu**

**Table of Contents**

[Introduction 3](#_Toc73387946)

[Personal Finance Project for Database Administration (IST 659) 3](#_Toc73387947)

[Project Overview: 4](#_Toc73387948)

[Skills Obtained: 8](#_Toc73387949)

[Reflection: 9](#_Toc73387950)

[Wine Project for Applied Data Science (IST687) 9](#_Toc73387951)

[Project Overview: 10](#_Toc73387952)

[Skills Obtained: 12](#_Toc73387953)

[Reflection: 13](#_Toc73387954)

[Sarcasm Project for Text Mining IST 736 13](#_Toc73387955)

[Project Overview: 13](#_Toc73387956)

[Skills Obtained: 16](#_Toc73387957)

[Reflection: 16](#_Toc73387958)

[Men NCAA Basketball Prediction Project for Big Data Analytics IST718 17](#_Toc73387959)

[Project Overview: 17](#_Toc73387960)

[Skills Obtained: 20](#_Toc73387961)

[Reflection: 20](#_Toc73387962)

[Obesity/Overweight Project for Information Visualization IST719 21](#_Toc73387963)

[Project Overview: 21](#_Toc73387964)

[Skills Obtained: 24](#_Toc73387965)

[Reflection: 24](#_Toc73387966)

[Conclusion 24](#_Toc73387967)

[References: 26](#_Toc73387968)

# Introduction

The School of Information Studies at Syracuse University Master of Science in Applied Data Science is an interdisciplinary program in which I enrolled Fall 2019, it provided me with the skills and ability to extract data from multiple domains, conduct assessments, and develop insight using various tools and techniques such as SQL Server Management Studio, Microsoft Access, and Excel, Abode Illustrator, Tableau, Weka, Python, and R Studio. Upon successful completion of this program, myself and fellow students ought to be able to apply the learning objectives to real-world situations. The Applied Data Science Program Learning Objectives are as follow:

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

2. Collect and organize data.

3. Identify patterns in data via visualization, statistical analysis, and data mining.

4. Develop alternative strategies based on the data.

5. Develop a plan of action to implement the business decisions derived from the analyses.

6. Demonstrate communication skills regarding data and its analysis for relevant professionals in their organization.

7. Synthesize the ethical dimensions of data science practice.

Throughout this milestone summarization, program learning objectives will be demonstrated by

completed course projects.

# Personal Finance Project for Database Administration (IST 659)

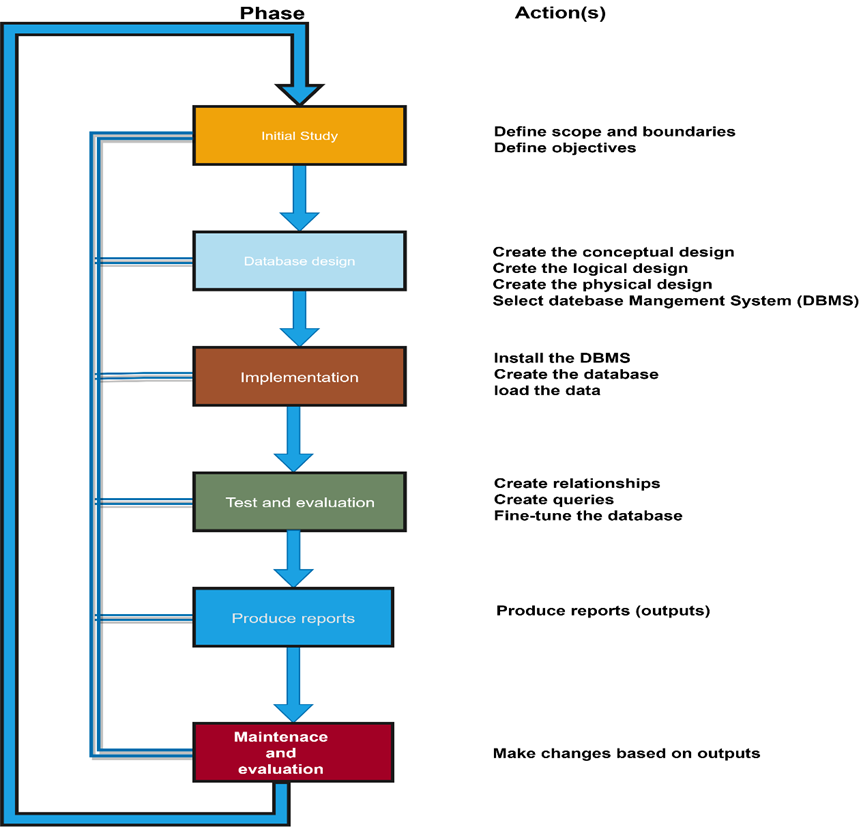
Database Administration is an introductory course in database management. The course provides students with the ability to examine data structure, database design, learn and write structured query language (SQL), database implementation, and management. The primary objectives are to create databases and database objects using SQL Server Management Studio (SSMS) and query language to solve problems. Figure 1 is the phrase approach containing the steps used as a guide through the appropriate stages.

Figure 1 Design Phase

# Project Overview:

The assignment for this course was to develop a personal database and or one that solves a business problem, personally or professionally. The database chosen to be developed was to track household spending over one year opportunity to improve financial decision marking by revealing saving options, unnecessary fees, and identify trends in spending habits.The questions this database will attempt to analyze as follow:

* What is the average dollar amount per category?
* How many times of month do I transfer funds from my other accounts and how much $$$?
* Where do I spend most of my money and on what?
* What are my monthly debits and credits?
* Where can I look for cost savings or avoidance?



Figure 2 Data Set

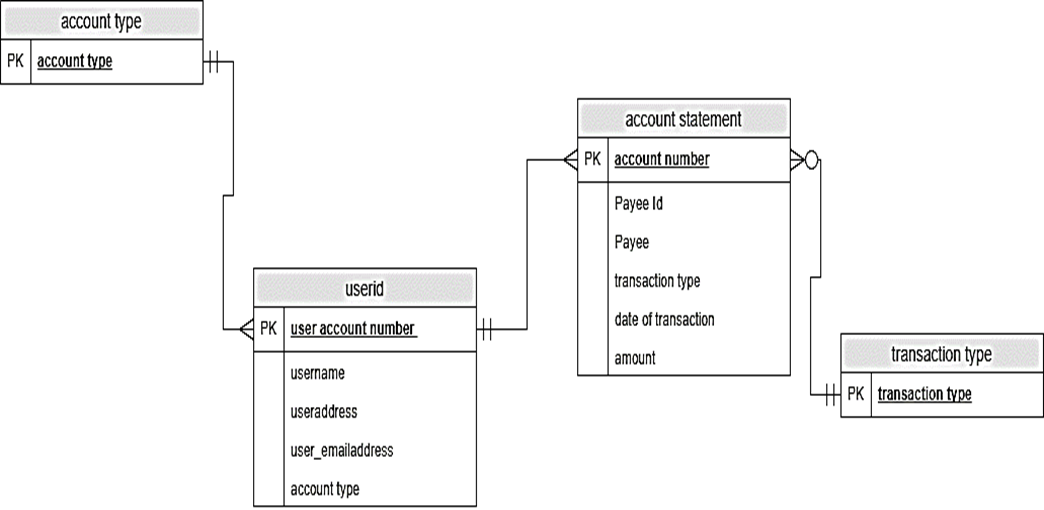
Using the knowledge obtained from the labs, data from bank statements were normalized, appropriate fields selected, a conceptual and logical model was created to organize the relationships between tables that were developed in SQL Server Management Studio (SSMS). SQL queries were used to generate reports and create functions

Figure 3 Conceptual Model

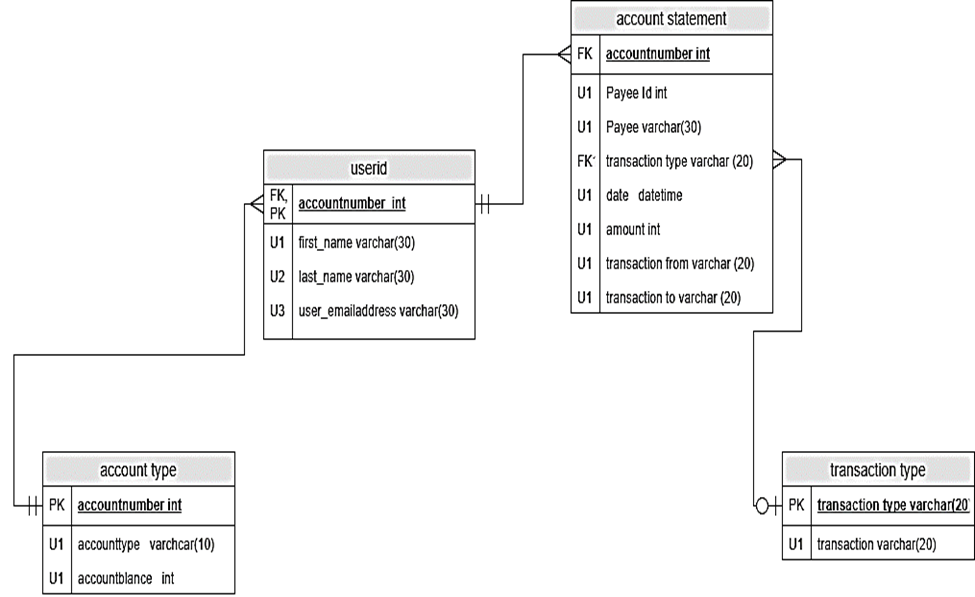


Figure 4 Logical Model

--Creating Tables for Personal Finance Database

CREATE TABLE Account Number

(

AccountNumberID begin identity primary key,

AccountType varchar (300) NOT NULL unique,

AccountOwnerFirstName varchar (300) NOT NULL,

AccountOwnerLasttName varchar (300) NOT NULL,

AccountOwnerEmailAddress varchar (300) NOT NULL,

CONSTRAINT U1\_AccountNumber UNIQUE (AccountType)

)

CREATE TABLE DepositTable (

DepositID bigint identity primary key,

DepositType varchar (300) NULL,

DepositFrom varchar (300) NULL,

DepositAmount Money NULL,

DepositDate Datetime NULL,

AccountNumberID bigint,

CONSTRAINT FK\_DepositTable FOREIGN KEY (AccountNumberID) REFERENCES AccountNumber (AccountNumberID)

)

Figure 5 Physical Model

-- Load Data in Personal Finance Database Tables

SET IDENTITY\_INSERT dbo. AccountNumber ON

INSERT INTO AccountNumber (AccountNumberID, AccountType, AccountOwnerFirstName, AccountOwnerLasttName, AccountOwnerEmailAddress)

VALUES (5638, 'CHECKING','GILBERT','GUYAH’, ‘gilbertguyah@gmail.com').

SET IDENTITY\_INSERT dbo. AccountNumber OFF

SET IDENTITY\_INSERT dbo. DepositTable ON

INSERT INTO DepositTable (DepositID, DepositType, DepositFrom, DepositAmount, DepositDate, AccountNumberID)

SELECT DepositID, DepositType, DepositFrom, DepositAmount, DepositDate, AccountNumberID FROM DepositTable$

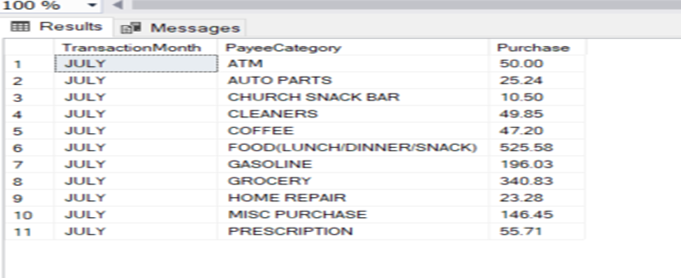


Figure 6 Output Transactional Data

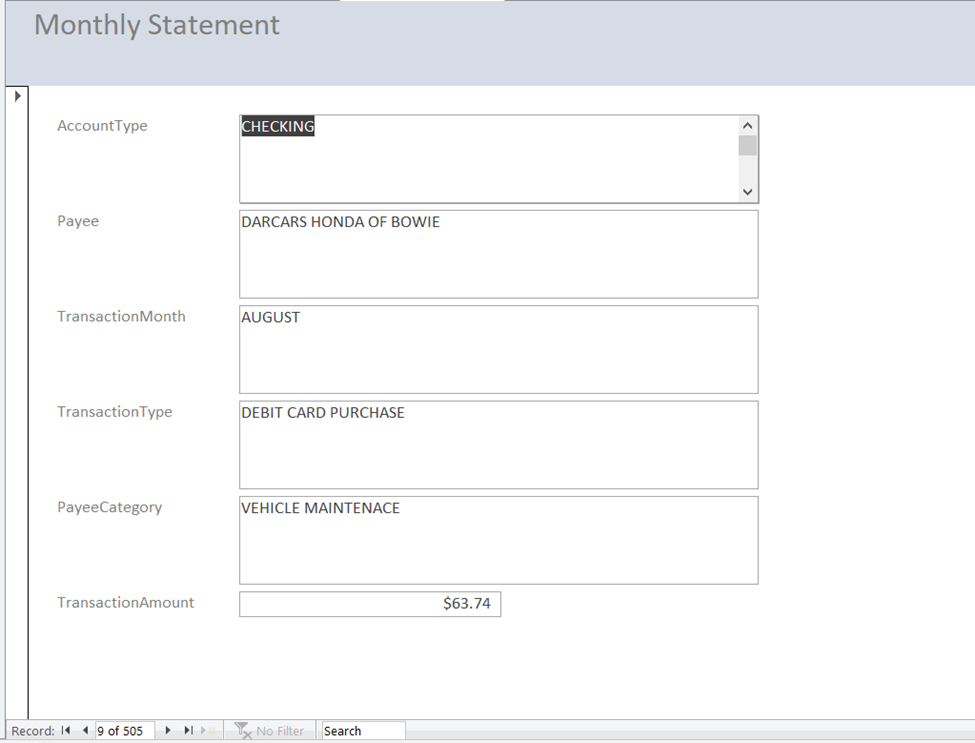
The use of Microsoft Access and R Studio was introduced as means to display outputs based on the project charter questions.

Figure 7 Graphical User Interface (GUI)

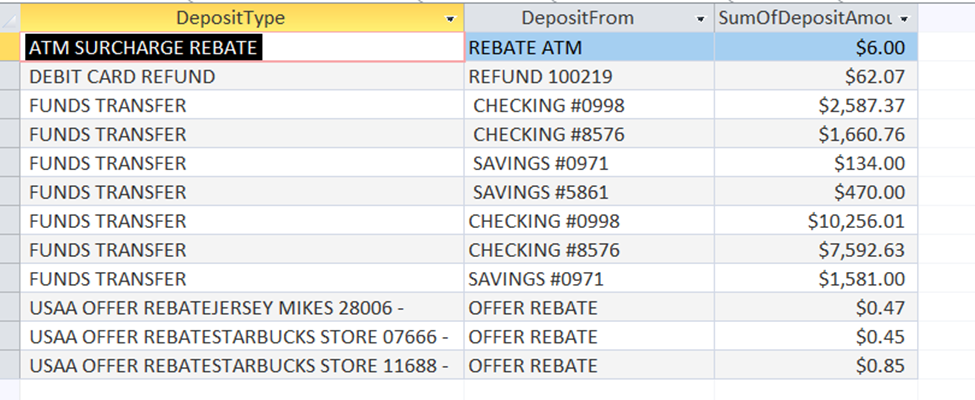


Figure 8 Microsoft Access Output

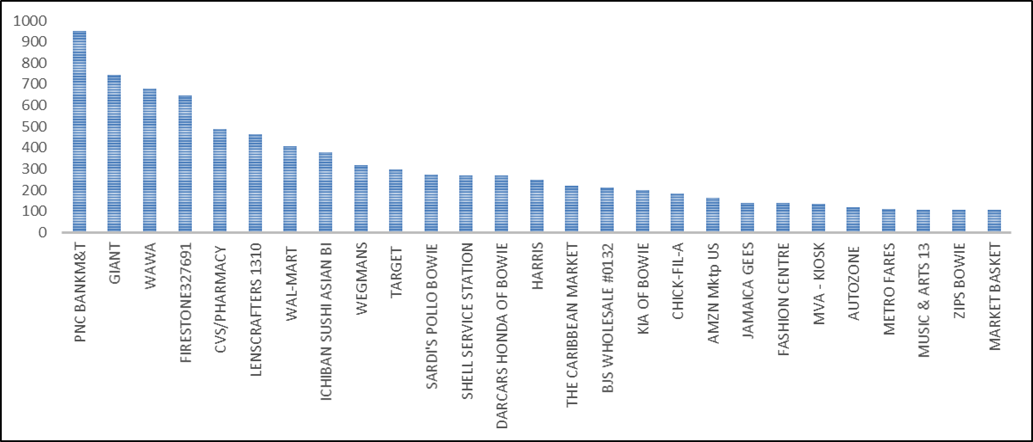


Figure 9 Payee Output

# Skills Obtained:

The skills obtained during this course have translated personally and professionally when analyzing data. Data elements that were not visible until completion of this course such as the missing values, spelling errors, and data not standardized. The use of database management systems like Access and SQL quires and understanding data security, the role of ethics in data analysis, data privacy, data backup, and recovery are currently incorporated into my data gathering and process flows.

# Reflection:

This project provides insight into how data is extracted, normalized, loaded, and transformed went developing a database to solve a problem or use other means. With no formal training on how a database was built, this was my first introductory course on the correct way to build a database and understand the terminology. Using Microsoft Excel Data Model and Pivot tables did not realize that structure of a database was embedded throughout the process, such as extracting the data from multiple sources using AdHoc quires, standardizing and normalizing the data, creating tables assigning foreign Keys and Primary Keys, developing conceptual and logical models then a star schema with nodes and various relationships: one to one, one to many, many to one, and many to many. Understanding the concept of a database, data modeling, and techniques will not struggle when developing a data model when solving a work-related or personal problem, this course has made me more effective a data analysis.

# Wine Project for Applied Data Science (IST687)

Applied Data Science course provides students with hands-on experience in data science that includes applied statistics (descriptive & Inferential), visualization, support vector machines, working geospatial data, association rule mining, text mining, connecting to external data sources, and use an open-source statistical software package R Studio to conduct machine learning. The primary objectives are computational scripting using R and R Studio, transformed data through cleaning, linking, summarization and aggregation.

# Project Overview:

The Project chosen to analyze is the wine quality data set from the UC Irvine Machine Learning Repository. Containing 1,599 Red Wine Samples, 4,898 White Wine Samples, 12 attributes 11 of which are measurable while the last attribute was not. The last attribute is titled ‘quality’ and is the median of at least 3 wine expert evaluations.

|  |  |
| --- | --- |
| * Fixed Acidity * Volatile Acidity * Citric Acid * Residual Sugar * Chlorides * Free Sulfur Dioxide | * Total Sulfur Dioxide * Density * pH * Sulfates * Alcohol * Quality \*\* |

Table 1 Wine Attributes

The goal of the project was to determine which attributes are responsible for the quality of the wine.

To be able to predict the quality of the wine based on a specific set of attributes.

Using RandomForest algorithm to obtain the significance of the variables. The top 4 variables were alcohol, density, volatile acidity, and free sulfur dioxide. Based on this information, 4 graphs were created depicting the effects these variables have on quality. The color represents the quality where red is low, orange is medium quality and green is higher quality. You can see the patterns in some of the graphs more clearly than in others

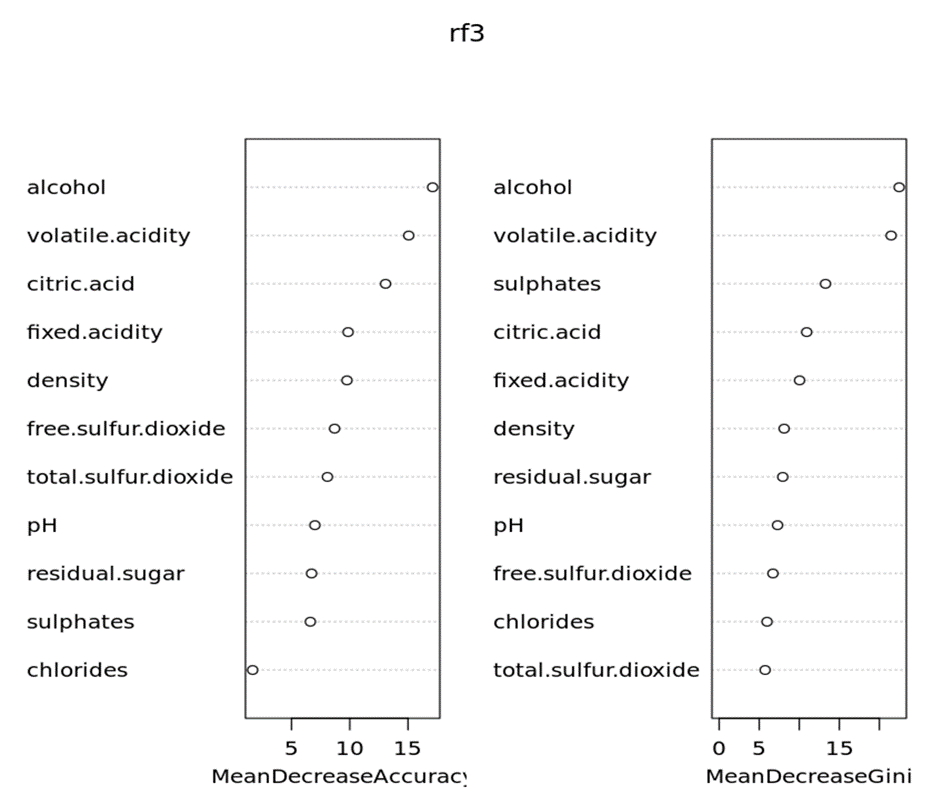


Figure 10 Random Forest Results (Red Wine)

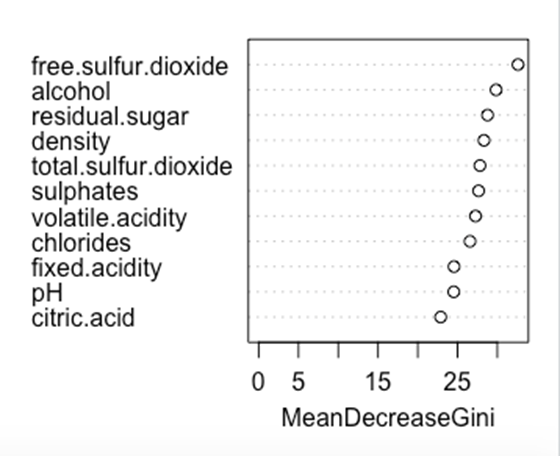
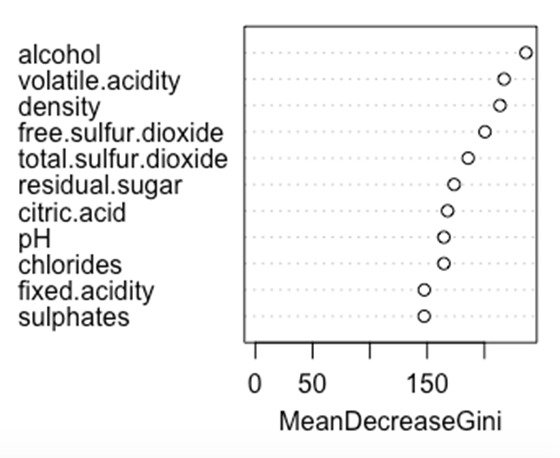


Figure 10 Random Forest Results (White Wine)

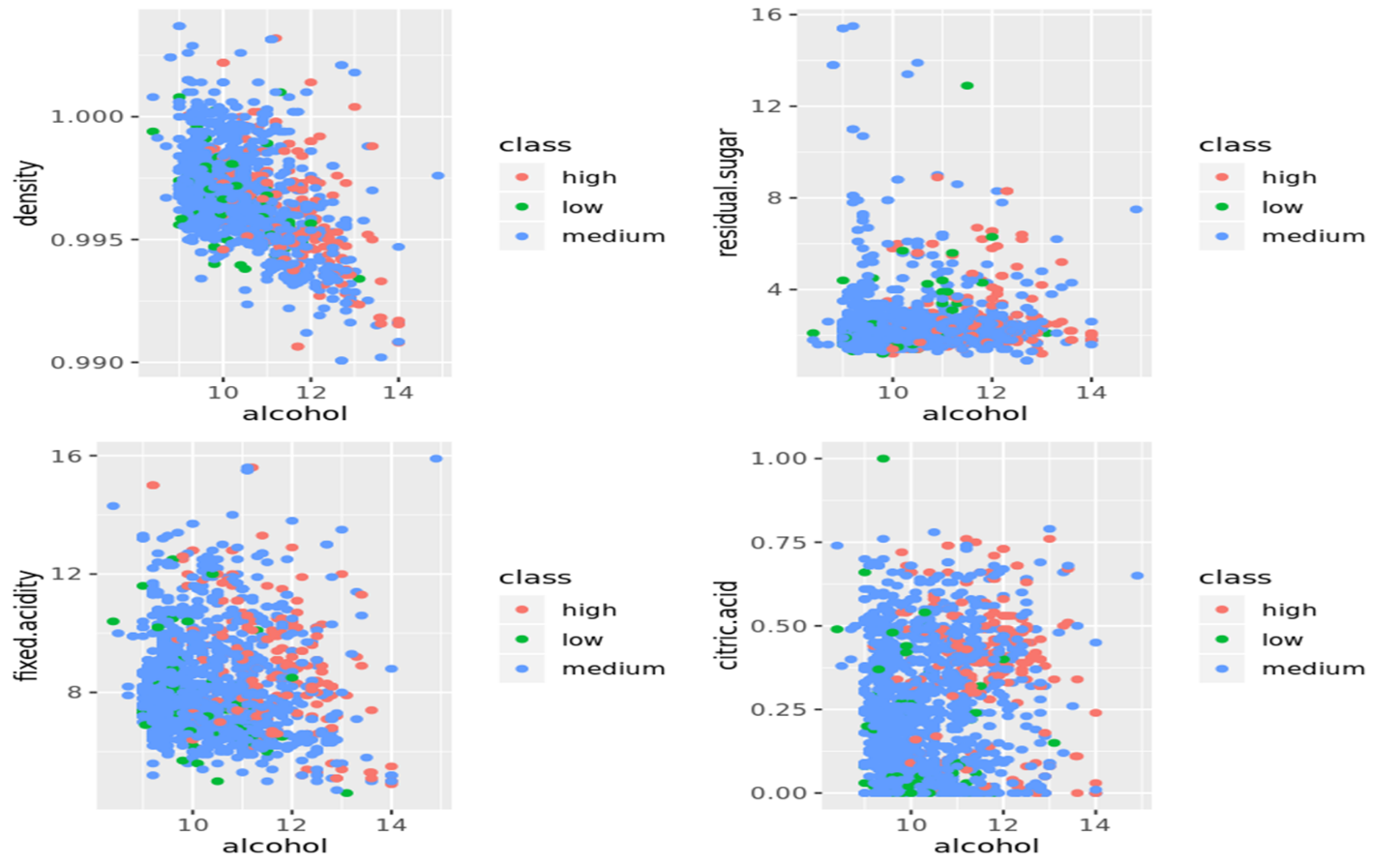


Figure 11 Quality Chart

As shown in the quality chart, lower density and higher alcohol content usually resulted in the wine being rated as higher quality. There also seems to be a pattern where lower volatile acidity results in higher to medium level qualities, but this trend isn’t as strong as the others. One reason the wines with high volatile acidity may be rated lower is that these wines tend to smell more strongly of vinegar which may result in a lower rating

# Skills Obtained:

The skills obtained conducting this project were an overview of the major practice areas in data science. Learn about winemaking and the chemical attributes of a wine, the types of grapes, its sugar content, soil, and the region in which the grapes are grown. Use of statistical analysis and machine learning algorithms such as RandomForest, Support Vector Machine, and collaboration tools such as Microsoft Teams, Modular Interactive Data Science Tool, Zoom, and Slack to communicate with team members and share project deliverables

# Reflection:

This being the first machine language course was overwhelming, ask weeks passed, completing the labs/ watching the asynchronous videos along with living class the concepts become more apparent. Using the collaboration tools and Machine learning techniques team was able to communicate effectively and share files. These skills were easily transferable to my professional life.

# Sarcasm Project for Text Mining IST 736

Texting mining course provides students with an increased awareness of the power of large text or unstructured data, computational methods to find patterns in corpora, and used text mining technologies rooted in machine learning, natural language processing, and statistics. The primary objectives are to describe basic concepts and methods in text mining, such as document representation, information extraction, text classification and clustering, and topic modeling; how to use benchmark corpora, commercial and open-source text analysis and visualization tools to explore patterns; understand conceptually the mechanism of advanced text mining algorithms for information extraction, text classification, and clustering.

# Project Overview:

The project chosen is to understand the complexity of detecting sarcasm through the use of computers. The data was gathered from different sources including Twitter and Reddit and was analyzed.

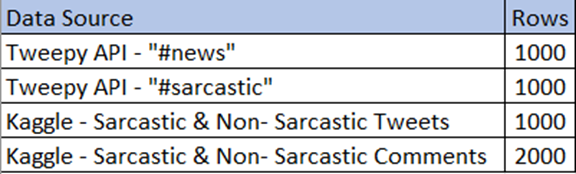


Table 2 Tweets Data Source

Both social media platforms are extremely popular with millions of users posting their opinions, reviews, and general information about any topic imaginable. The results solidify the notion that certain words are more frequently used and associated with sarcasm. For example, it was observed that words such as “love,” “obviously,” and “clearly” are usually associated with words used in the sarcastic text.



Figure 12 Word Clouds of Sarcastic Tweets

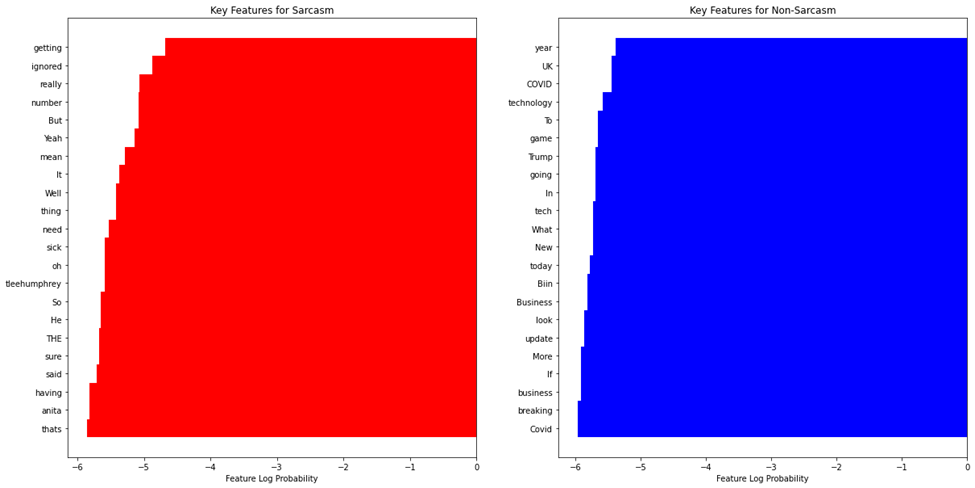


Figure 14 Key Sarcasm and Non- Sarcasm Words

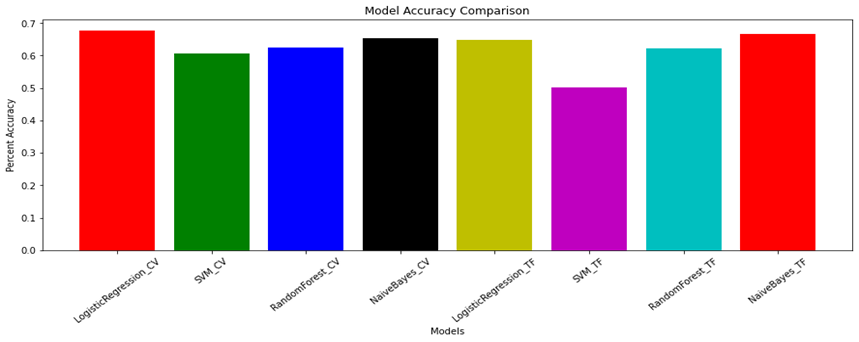


Figure 15 Model Comparison

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| model | max\_ngrams | all\_lowercase | news\_included | accuracy\_score | f1\_score |
| Naïve Bayes | 3 | yes | yes | 0.698 | 0.698 |
| Naïve Bayes | 3 | no | yes | 0.684 | 0.684 |
| Naïve Bayes | 1 | no | yes | 0.733 | 0.725 |
| Naïve Bayes | 5 | no | yes | 0.676 | 0.677 |
| Naïve Bayes | 3 | no | no | 0.669 | 0.672 |
| SVM | 3 | yes | yes | 0.701 | 0.698 |
| SVM | 3 | no | yes | 0.693 | 0.692 |
| SVM | 1 | no | yes | 0.707 | 0.706 |
| SVM | 5 | no | yes | 0.701 | 0.7 |
| SVM | 1 | no | no | 0.669 | 0.672 |

Table 3 Algorithm Results

# Skills Obtained:

The skills obtained during this course were using Python extensively to conduct text mining, the various machine learning, clustering, and classification algorithms. Communicate data and its analysis to relevant professionals with an organization

# **Reflection:**

This project provides an insight that it is much harder to detect sarcasm in tweets and or written language observations. Learned that there are many different types of sarcasm and these displayed by humans can vary. The table depicts 7 common types of sarcasm along with their description, ranging from self-deprecating, polite, obnoxious, and raging.

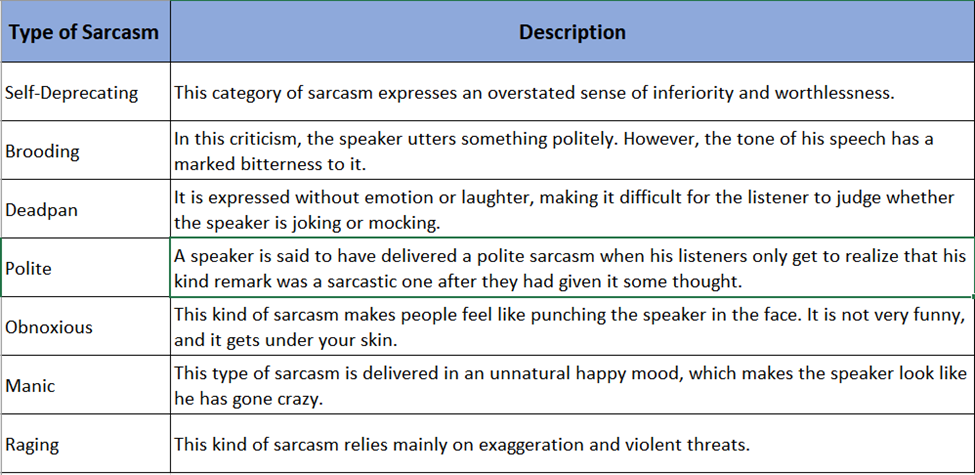


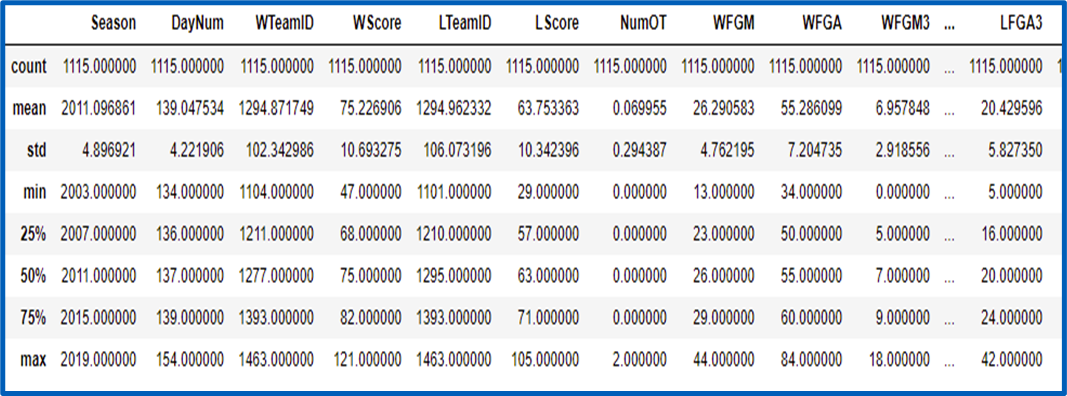
Table 3 Types of Sarcasm

Obtaining a Twitter account and requesting for Twitter API, could see how tweets can be misinterpreted. However, in a society where 280- character Tweets and Emojis are interchangeably used to convey moods, and emotions can improve sarcasm detection across the digital spectrum.

# Men NCAA Basketball Prediction Project for Big Data Analytics IST718

Big Data Analytics course provides students with the ability to obtain data and explain data structures and data elements. Explore data by analyzing using qualitative techniques including descriptive statistics, summarization, and visualizations. The course objective is to perform data manipulation in Python, Model relationships between data using the appropriate analytical methodologies matched to the information and the needs of clients and users, and communicate the results in a meaningful way.

# Project Overview:

The project chosen was to Use historical data to accurately predict the winner of every possible matchup in the NCAA Division I Basketball tournament. Each season there are thousands of NCAA basketball games played between Division I men's teams, culminating in March Madness, the 68-team national championship that starts in the middle of March. As per Forbes.com, this year March Madness gambling is expected to break the $8.5 Billion record. The dataset contains 357 Division-I Teams, Team names, conferences, rankings, location, Tournament seeds since 1984-85 season, Team stats for all regular season, conference tournament, and NCAA® tournament games since 2002-2003 season and Kenneth Massey historical data

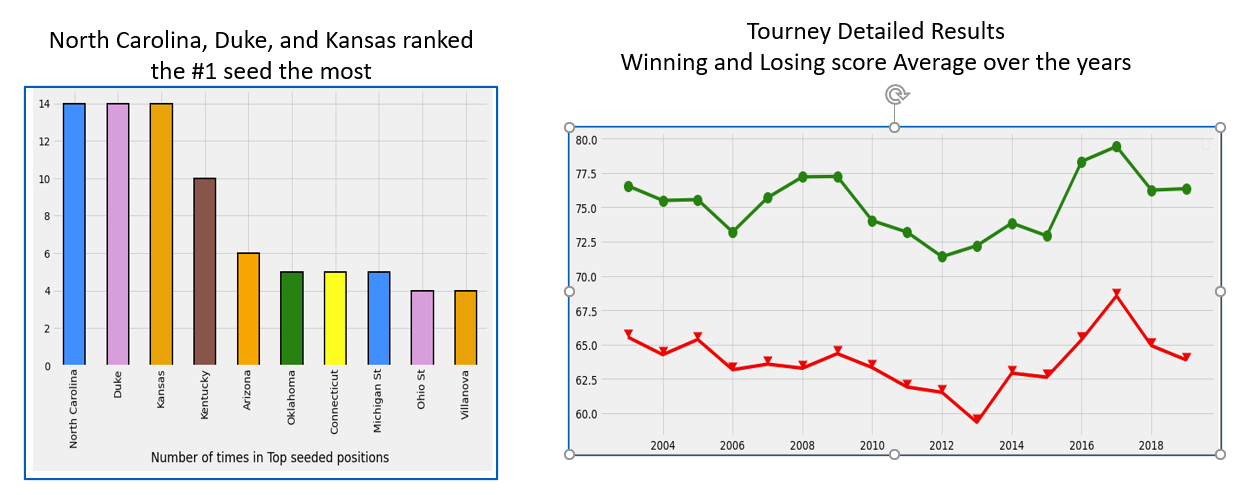


Figure 17 Top Seed Teams Figure 16 Winner/Losing Scores

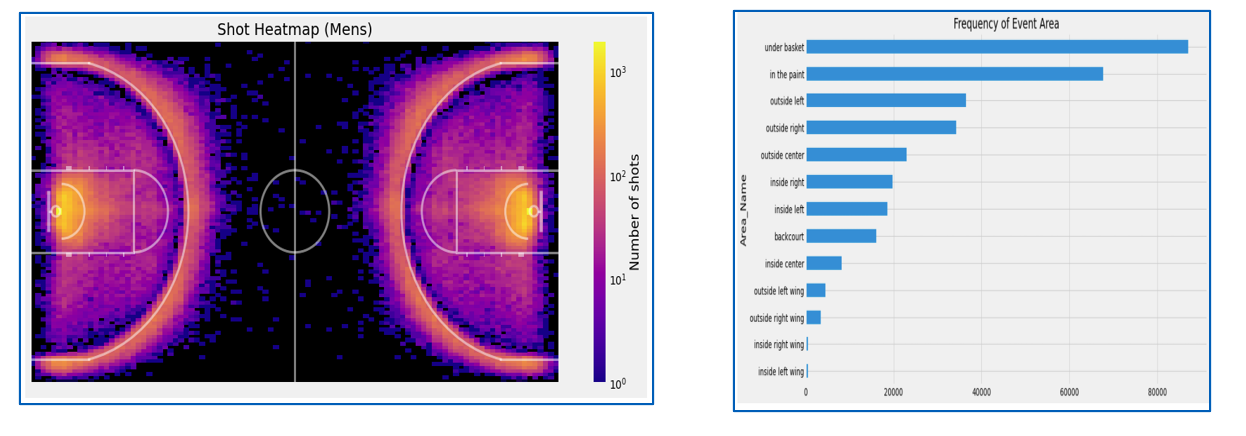


Figure 18 Heat Map Frequency of Shots Taken

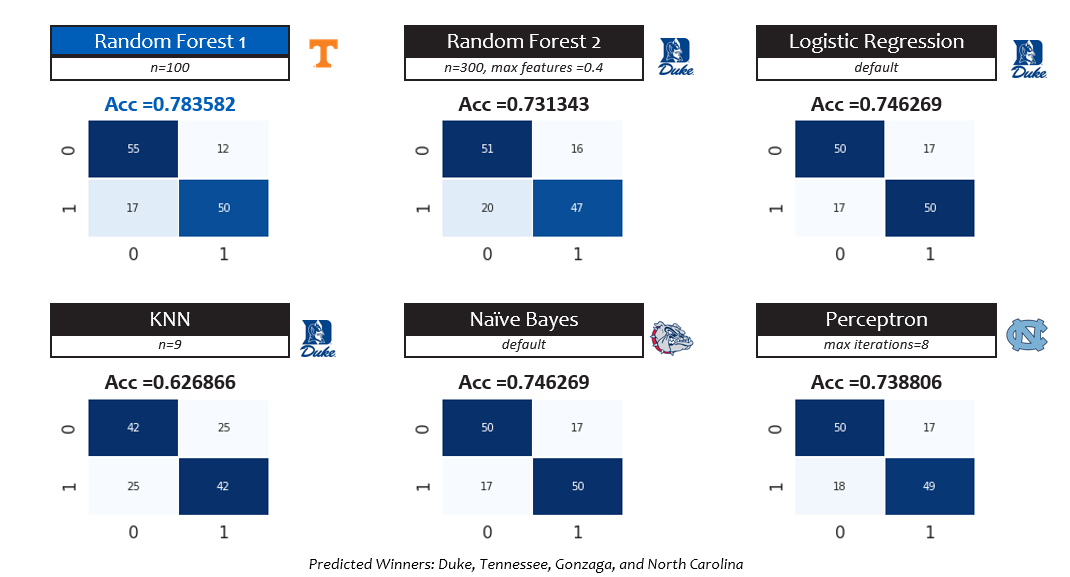
Several modeling techniques were used in the prediction such as Random Forest, Logistic Regression, KNN, Naïve Bayes, Perceptron. 

Figure 19 Algorithm Predictions

The models were tested with the 2019 tournament. The RandomForest model had the best accuracy in predicting the winner of the most games in the tournament.

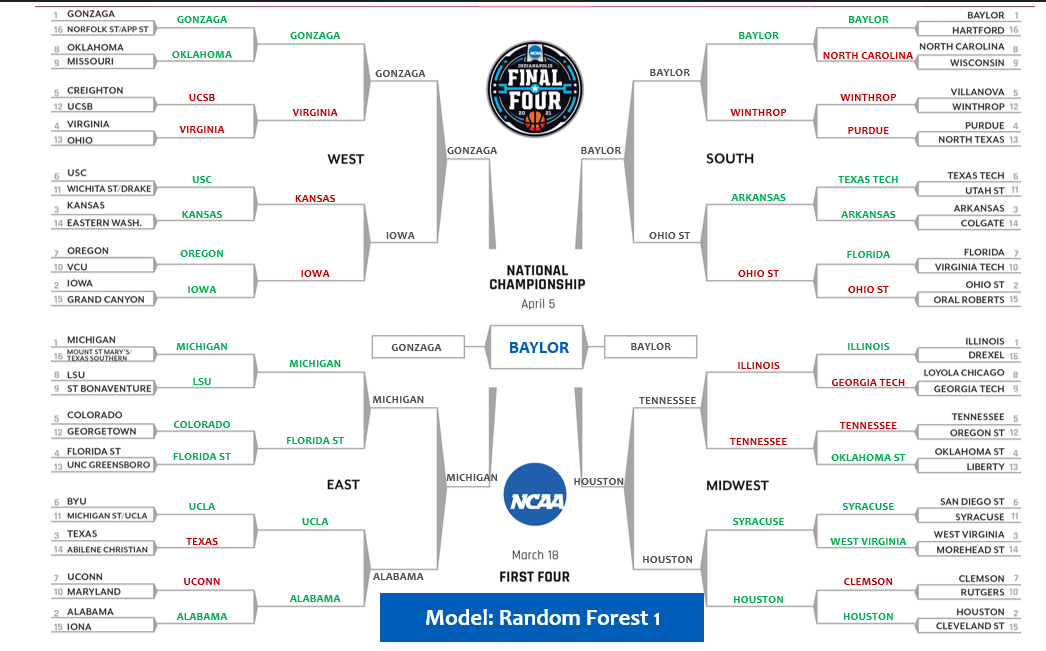


Figure 20 NCAA Final Four Bracket Prediction

# Skills Obtained:

How to communicate the results in a meaningful way. Select an applicable analytical methodology for real problems in areas such as business, science, and engineering. Python, Spark, and TensorFlow. Use Google Colab and NVidia GPU acceleration analysis

# Reflection:

This was the most challenging course I have taken within the data science curriculum. It requires more than 100 percent of your time and attention. It a compilation of all the statical analysis courses taken, not to mention the advanced machine learning algorithms and scripting. The home works were challenging one felt they were on-the-job research for data to depict an accurate prediction and report. This course provides me the best insight on what a data scientist and analysis is required to perform at an organization.

# Obesity/Overweight Project for Information Visualization IST719

Information Visualization provides students with introductory skills and techniques through the R programming language and Adobe Illustrator. These skills include using data cleaning techniques, controlling the R graphics environment, developing custom plots, visually exploring data, using design concepts to visually communicate the story in the data, and discussing issues related to the ethics of data visualization. Conceptual themes will be presented alongside technical aspects of data visualization. The course learning objectives are using R to prepare basic data cleaning, and preparation on a wide range of datasets, identify stories within the datasets through exploration, generate rich visual artifacts that communicate data stories. Identify the optimal type of visualization to minimize viewer cognitive overload and maximize image interpretability

# Project Overview:

The assignment for this course was to developed an Adobe poster of a real-world problem, the chosen visualized was the increase in obesity in America based on age, gender, race, and education. The data set for this project came from the Behavioral Risk Factor Surveillance System (BRFSS) Prevalence Data (2011 to present) published Centers for Disease Control and Prevention.

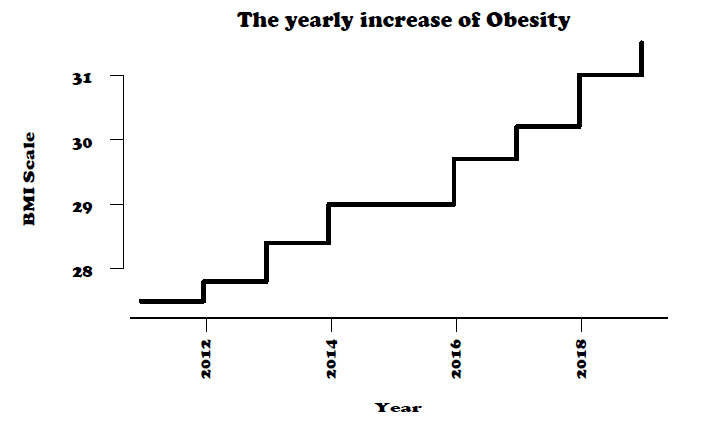
Since the national adult obesity rate has increased by 26 percent since 2008 and is a major health issue resulting in numerous diseases, specifically increased risk of certain types of cancer, coronary artery disease, type 2 diabetes, stroke, as well as significant increases in early mortality and economic cost. The project questions as shown below was used to analyze the data to understand the effect of obesity in our society:

Figure 21 Obesity Yearly Increase

* What is the BMI between being obese and being overweight by adults/male & female?

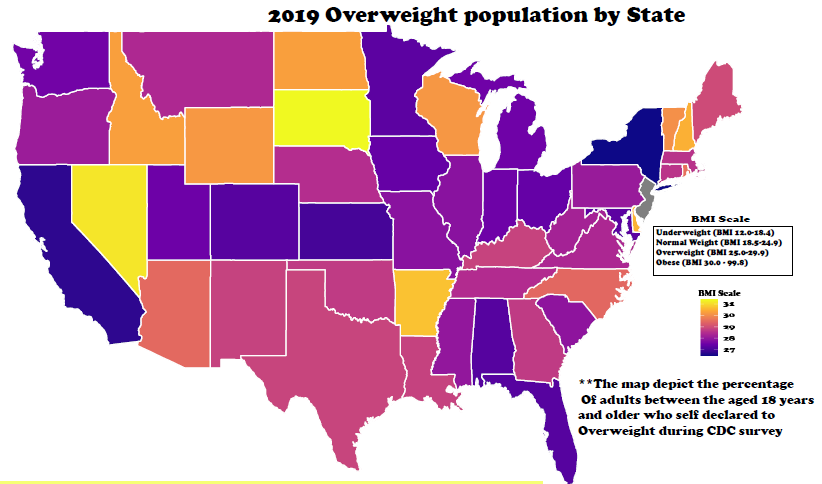


Figure 22 Overweight US Population by States

The highest rate of obesity between Gender, Race

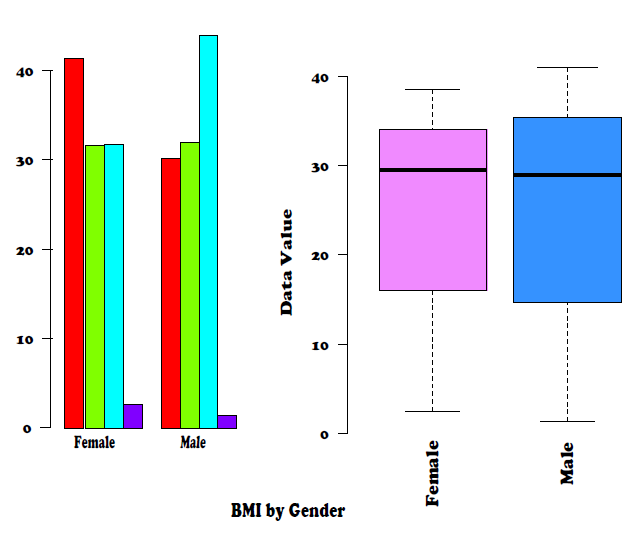


Figure 23 BMI by Gender

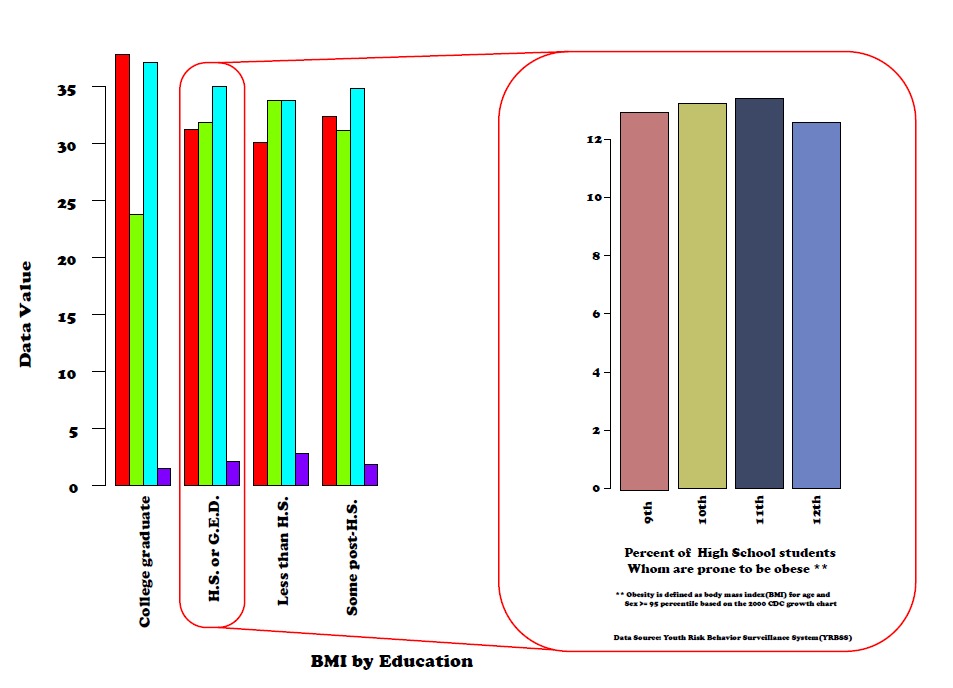
What groups have the highest obesity by education 

Figure 24 BMI by Education levels

# Skills Obtained:

The skills obtained during the course are using Adobe Illustrator to merge R data visualizations, design elements, and context cues into a single artifact, identify patterns in data via visualization, statistical analysis, and data mining and develop a plan of action to implement the business decisions derived from the analyses.

# Reflection:

Using adobe illustrator to manipulate R visualization was very challenging since my only experience using adobe was to view pdf files, redact PPI within forms and or letters. In addition, this course taught me how to use colors, layouts, charts, and fonts size to capture the audience's attention during the presentation. After reviewing the CDC data on obesity, the impacts it has on our society and further understanding the Body Mass Index (BMI) chart by age and gender, allowed me to evaluate and conduct some lifestyle changes about food and exercise to ensure a healthy lifestyle.

# Conclusion

The portfolios chosen have demonstrated the learning objectives and the major practice areas in data science. The collection and organization of data using application programming interfaces and web scrapping to be analyzed using data mining and statistical techniques such as liner /logistics regression. To implement the business decisions derived from the analyses, understanding the patterns using clustering techniques, help analysis develop alternative strategies, eliminate basis, and mitigate the ethical dimensions practices in data science. As data scientists not investigating this pattern can result in decisions being to impact certain groups or part society, more even depict the wrong message for a person / corporate gains. i.e., corporation embellishing numbers or shown distorted charts.

Throughout the program, communication skills were developed and demonstrated in project reports and presentations regarding data and the analysis preform to ensure relevant professionals an organization can make effective decisions

Using the methods of learning in this program, a graduate can easily incorporate the skills acquired to confront real-world problems in corporate settings, develop and explain solutions to stakeholders.

# References:

Chin, R. (2011, November 14). The Science of Sarcasm? Yeah, Right. Retrieved December 09, 2020, from https://www.smithsonianmag.com/science-nature/the-science-of-sarcasm-yeah-right-25038/

Geoffrey Hinton Quotes. (2020). Retrieved December 09, 2020, from https://www.brainyquote.com/quotes/geoffrey\_hinton\_875276

Guyah, G.A., (nod), (2021) Guyah-Projects Retrieved from https://github.com/gilguyah/Guyah- Projects.

Hoffer, J.A, Ramesh, V., & Topi, J. (2016). Modern database management (12th ed.). New York: Pearson

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, An Introduction to Statistical Learning with Applications in R, Springer, 2013.

Kaggle Data Dictionary Retrieved from https://www.kaggle.com/c/ncaam-march-mania-2021/data

Merriam-Webster. (n.d.). Sarcasm. In Merriam-Webster.com dictionary. Retrieved December 9, 2020, from https://www.merriam-webster.com/dictionary/sarcasm

Miller, Thomas W., Modeling Techniques in Predictive Analytics with Python and R, Pearson, 2015

Watt, A. Chapter-13-database-development-process. Retrieved 24 October 2019. https://opentextbc.ca/dbdesign01/chapter/chapter-13-database-development-pocess