Portfolio Milestone

James Y. Eakins

School of Information Studies
Syracuse University

586974381

https://github.com/jyeakins/MSADS_Portfolio

01 June 2020

Introduction

Over the courses of Data Science Program, the objectives of the program are to learn how to use the techniques and adapt new innovative tools to solve problems by using the data.

Reports and presentations were created in courses which exemplify the skills developed in the program including, but not limited to:

- IST 659: Data Admin Concepts & Database Management
- IST 687: Introduction to Data Science
- IST 707: Data Analytics
- IST 736: Text Mining

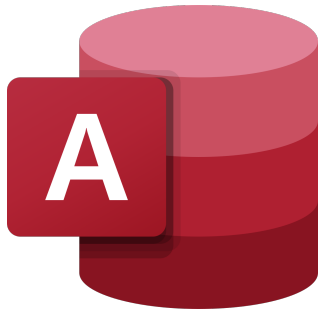
https://github.com/jyeakins/MSADS_Portfolio

Learning Objectives

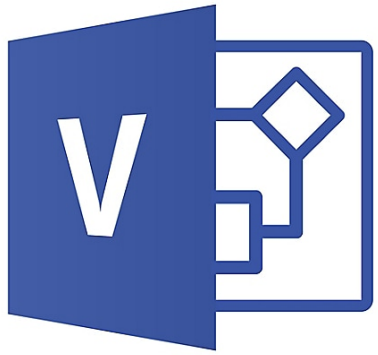
The Applied Data Science program has four major learning objectives

1. Data collection: using tools to collect and organize data
2. Data analysis: Identify patterns in the data via visualization, statistical analysis, and data mining
3. Strategy and decision: develop alternative strategies based on data
4. Implementation: develop a plan of action to implement the business decisions.

https://github.com/jyeakins/MSADS_Portfolio



Microsoft®
SQL Server™



IST 659

Data Admin Concepts & Database Management

School of Information Studies
Syracuse University

https://github.com/jyeakins/MSADS_Portfolio

IST 659: Data Admin Concepts & Database Management

Introduction

- Through the course of Database management, the database was built to give the user a list of parts, compatibility among the parts, and the total cost of the build. It also was to give the list of components to run any software that people want to use for their productivity or entertainment.

Build-A-Computer Workshop

Customer Registration

Pre-built

New Build

https://github.com/jyeakins/MSADS_Portfolio

School of Information Studies
Syracuse University

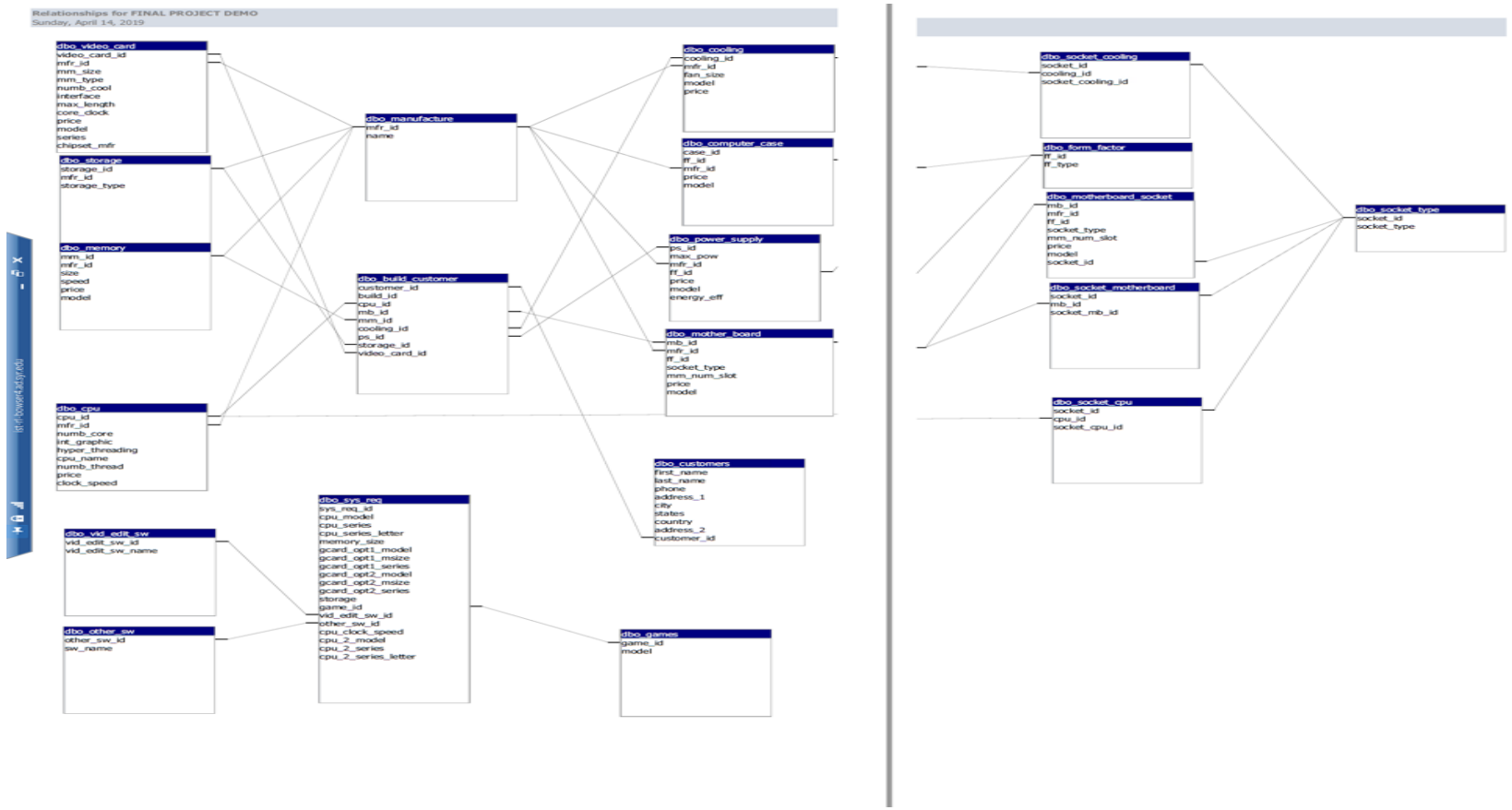
IST 659: Data Admin Concepts & Database Management

Modeling, Table Creation, and Reporting

- Microsoft Visio and Word document were used to create the conceptual and logical model to organize tables and relationships between entities.
- Microsoft SQL Server was used to create the structure of the database with the tables.
- User interactive interface, reports and summary were created by using Microsoft Access
- Data was collected from the Internet and stored in Access.

https://github.com/jyeakins/MSADS_Portfolio

IST 659: Data Admin Concepts & Database Management



IST 659: Data Admin Concepts & Database Management

Build-A-Computer Workshop

Customer Registration

Pre-built

New Build

PRE - BUILT

Player Unkonwns Battle Ground

Far Cry New Dawn

Cheapest Build

$

Expensive Build

$$$

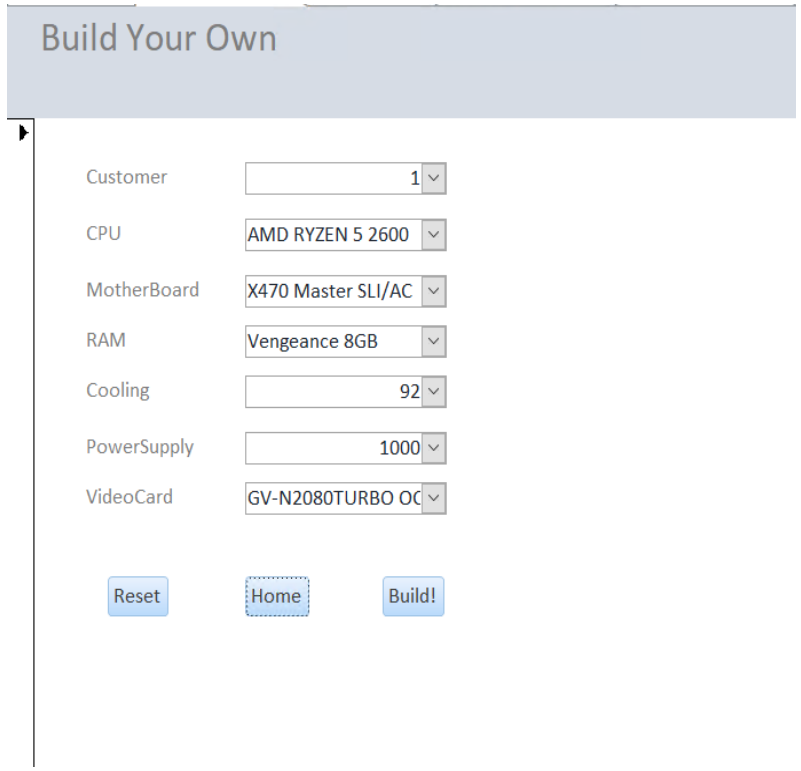
Customer Build

Home

https://github.com/jyeakins/MSADS_Portfolio

School of Information Studies
Syracuse University

IST 659: Data Admin Concepts & Database Management



https://github.com/jyeakins/MSADS_Portfolio

School of Information Studies
Syracuse University

IST 659: Data Admin Concepts & Database Management

Far Cry New Dawn

Game Name	Far Cry New Dawn

Total			526.95		
CPU	MotherBoard	RAM	PowerSupply	Cooling	VideoCard
AMD RYZ	GA-970A-DS3P	Vengeance 8GB	CORSAIR CX450	CLP0556-B	GT 710 1GD3H LP

Total			528.95		
CPU	MotherBoard	RAM	PowerSupply	Cooling	VideoCard
AMD RYZ	GA-970A-DS3P	Vengeance 8GB	CORSAIR CX450	ROCC-16003	GT 710 1GD3H LP

Total			541.95		
CPU	MotherBoard	RAM	PowerSupply	Cooling	VideoCard
AMD RYZ	GA-970A-DS3P	Vengeance 8GB	CORSAIR CX450	RR-212E-20PK-R2	GT 710 1GD3H LP

Total			546.95		
CPU	MotherBoard	RAM	PowerSupply	Cooling	VideoCard
AMD RYZ	GA-970A-DS3P	Vengeance 8GB	ROSEWILL LEPTON 600	CLP0556-B	GT 710 1GD3H LP

Total	548.95

Cheapest_build

total	262.68			
CPU	MotherBoard	VideoCard	Memory	PowerSupply
AMD FX-63	GA-970A-DS3P	GT 710 1GD3H LP	M378B5173DB0-CK0	CORSAIR CX450

Sunday, April 14, 2019 Page 1 of 1

expeinsive_build

total	2566.67			
CPU	MotherBoard	VideoCard	RAM	PowerSupply
INTEL CORI	ROG Strix Z390-E Gami	GV-N2080TURBO OC-8	Trident Z RGB DC Series	EVGA SUPERNOVA 1000

Sunday, April 14, 2019 Page 1 of 1

https://github.com/jyeakins/MSADS_Portfolio

School of Information Studies
Syracuse University

IST 659: Data Admin Concepts & Database Management

Reflection

- The development of database management solution made possible to think like an engineer and analyze the problem in many ways.
- This project allows students to sketch the blueprint of the database structure, design the interface for users and build the database to make the raw data can be useful tool.
- This project contributed in more advanced courses in the later term of Applied Data Science program to explore many different aspects such as security, database architecture and how to cross use the tools in different fields.

https://github.com/jyeakins/MSADS_Portfolio

IST 687

Introduction to Data Science

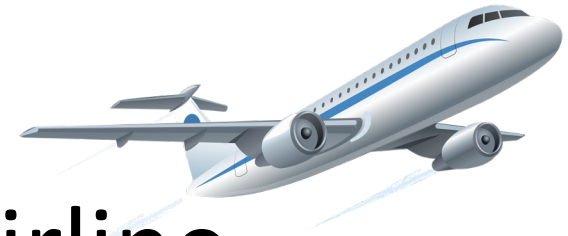
School of Information Studies
Syracuse University

https://github.com/jyeakins/MSADS_Portfolio

IST 687: Introduction to Data Science

Introduction

- Through studying of the Introduction to Data Science under Dr. Stanton, new data science techniques were introduced
- The Linear Regression Model, the Correlation Matrix, the Association Rule mining, the Support Vector Machine, the Apriori Rule Association, the Word Cloud and the visualization used to analyze the data
- R Studio is used for this project, and it was first introduction to the software and was very helpful to analyze such a big amount of data



Airline Satisfaction

https://github.com/jyeakins/MSADS_Portfolio

IST 687: Introduction to Data Science

Cleaning, Analysis and Results

- The data cleaning process was the major component of this project since there are many rows and columns that may cause the faulty results on analysis
- The Linear Regression model showed the independent and dependent variables to understand the positive or negative impact on the satisfactory of the customers.
- The level of p-value explains whether the variable has the significance to the model or not. Furthermore, the linear regression model provides the adjusted R-Squared value to understand the quality of the model.

https://github.com/jyeakins/MSADS_Portfolio

IST 687: Introduction to Data Science

Introduction

- After determining the key drivers have the significance to the satisfaction, the Association Rule Mining was used to find the co-occurrences of variables. The goal of this analysis was to find a combination of variables that inevitably lead to high or low customer satisfaction.
- SVM is used to build a classification model with the customer satisfaction as the target variables by showing the confusion matrixes of the results.

IST 687: Introduction to Data Science

Reflection

- This project was the best exercise which provided the lesson that the cleaning procedure before analyzing the big data is the most important step in Data Science
- By using multiple analysis tools, it gave the opportunity to observe the satisfactory of airline customers and create the different models that lead to an insight to the bad satisfactory airlines
- In the final term of Applied Data Science program, these basic tools were used to clean and understand bigger and complex data and apply the complicated analysis.

https://github.com/jyeakins/MSADS_Portfolio

IST 707
Data Analytics

School of Information Studies
Syracuse University

https://github.com/jyeakins/MSADS_Portfolio

IST 707: Data Analytics

Introduction

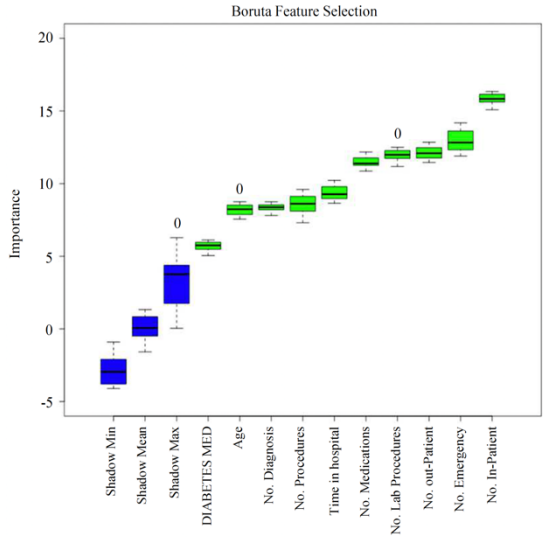
- The project for the Data Analytics course under the direction of Dr. Gates was more advanced project that involves with the mining of data, more advanced analysis tools such as regression, classification and clustering, and interpretation of results to make the story out of the numbers.
- In the Final presentation, there were four analysis techniques implemented to give insightful results about re-admission rate of hospital patients.
- The study provides with an efficient prediction model that can be deployed to a clinical scenario and help healthcare units to be prepared for the unavoidable re-admissions and provide alternative care to preventable re-admissions

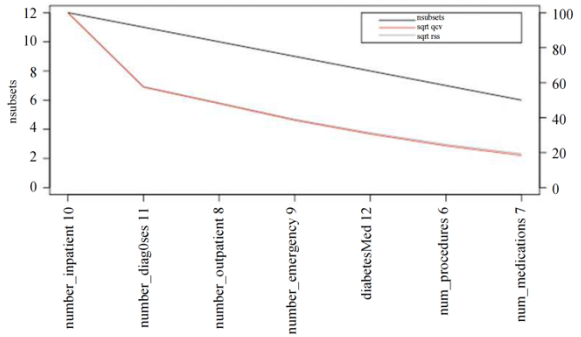
https://github.com/jyeakins/MSADS_Portfolio

IST 707: Data Analytics

Cleaning, Analysis and Results

- The data required cleaning and selecting features to eliminate unnecessary variables and focus on a few specific targets
- This study employs Boruta algorithm and stepwise regression to determine the best features within the dataset





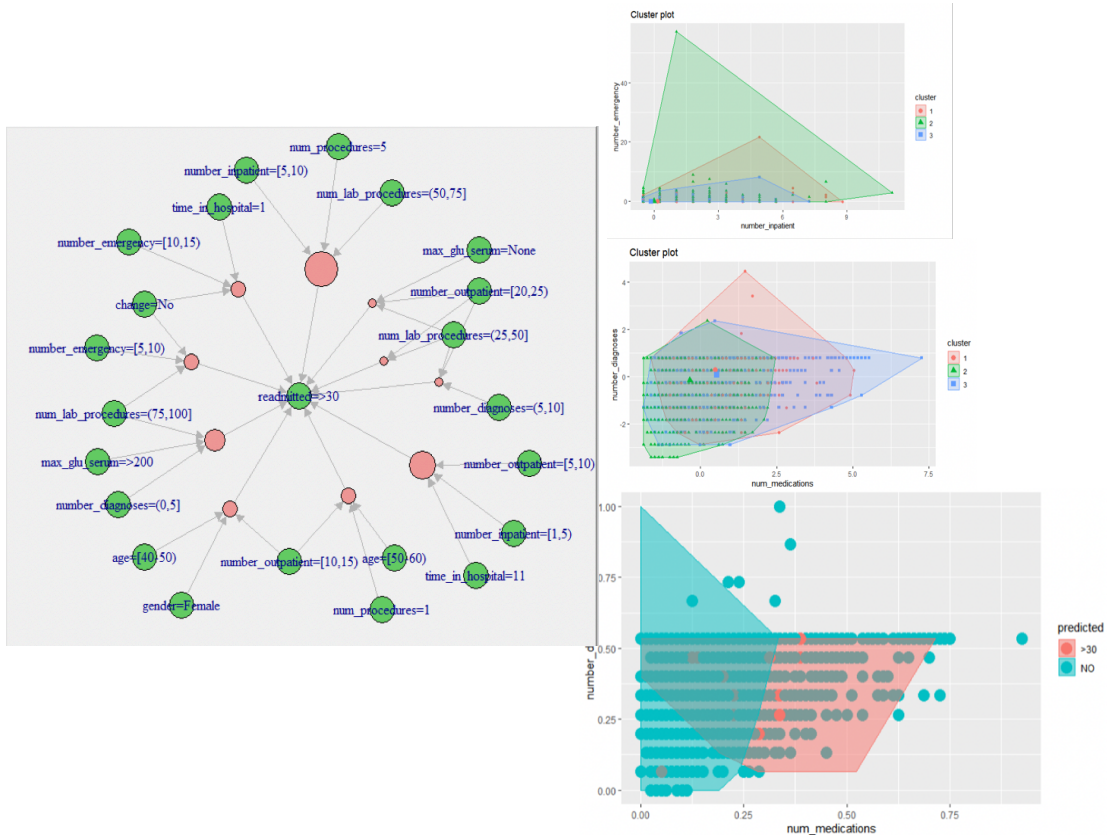
https://github.com/jyeakins/MSADS_Portfolio

School of Information Studies
Syracuse University

IST 707: Data Analytics

Cleaning, Analysis and Results

- SVM generated an accuracy of about sixty-seven percent where are random forest was found to highly over-fitting (Due to the size of dataset)
- Association rules and data visualization (Fig.9) was found to be the two most useful methods to understand the underlying factors causing hospital re-admissions



https://github.com/jyeakins/MSADS_Portfolio

School of Information Studies
Syracuse University

IST 707: Data Analytics

Cleaning, Analysis and Results

- This study implements a predictive analytical approach to identify patients prone to readmission and thus, systematically reduce the number of avoidable re-admissions mainly caused by patient non-compliance to medication instruction or early discharge from hospital
- The novelty of this method is to directly incorporate patients' history of re-admissions into modeling framework along with other demographic and clinical characteristics.

https://github.com/jyeakins/MSADS_Portfolio

IST 707: Data Analytics

Reflection

- Understandably, there are instances where techniques or tools sometimes cause unpredicted results as SVM and Random Forest shown from this project.
- The visualization is a great tool to show the results of the analysis, but it also helps to pick the key features, or see where to focus on when it comes to large datasets.
- This project successfully met all four learning goals where it required students to collect the data from outside of sources, analyze the data, find different strategies and decide which tools to use when there are obstacles in the way, and finally provide insightful conclusion to readers to implement the solutions to the problems which can be found in the real world.

https://github.com/jyeakins/MSADS_Portfolio

IST 736
Text Mining

School of Information Studies
Syracuse University

https://github.com/jyeakins/MSADS_Portfolio

IST 736: Text Mining

Introduction

- The project for IST 736 course was launched under direction of Dr. Gates, and Text mining from tweets, analyze the sentiments and text analysis techniques were introduced to understand how to treat the unstructured data from the scratch.
- In the final presentation, the tweets about recent tragedy of Kobe Bryant was collected to analyze public sentiments toward Kobe Bryan and his daughter, Gigi.
- With the history of Kobe Bryan's unpleasant incidents and the status as heroic player at Los Angeles Lakers, this study was to give insightful recommendation to commercial companies on whether they should feature Kobe Bryant for the commercial campaign.

https://github.com/jyeakins/MSADS_Portfolio

IST 736: Text Mining

Collecting, Cleaning, Analysis and Results

- This project required immediate data collection from Twitter since the limitation on how far the user can collect the data from the past
- By using the Python's Tweepy package, 35804 tweets were collected from the public
- The data was cleansed to have only raw text from the tweets since the emojis and other non-English characters can ruin the analysis.

https://github.com/jyeakins/MSADS_Portfolio

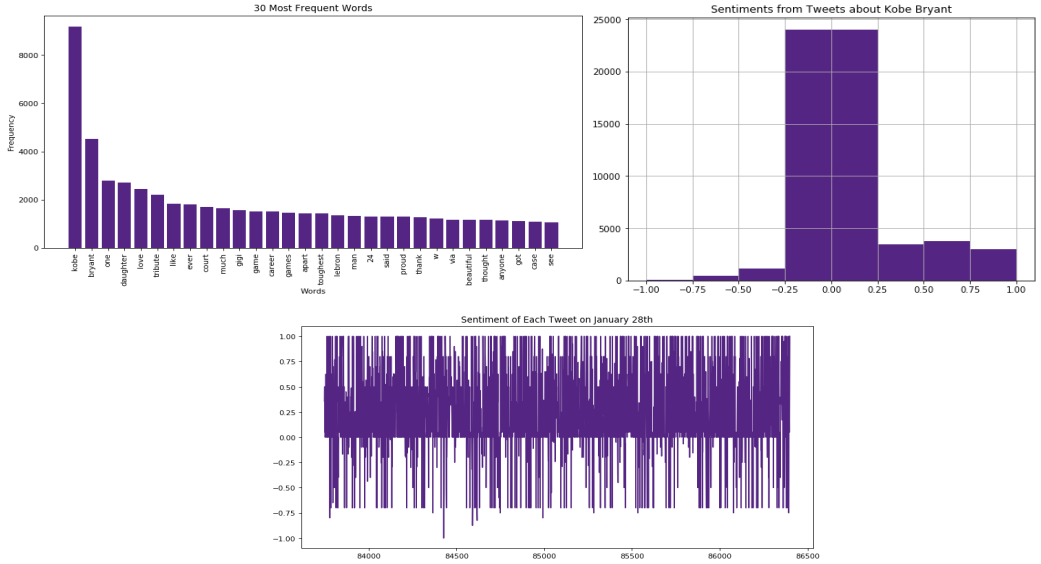
IST 736: Text Mining

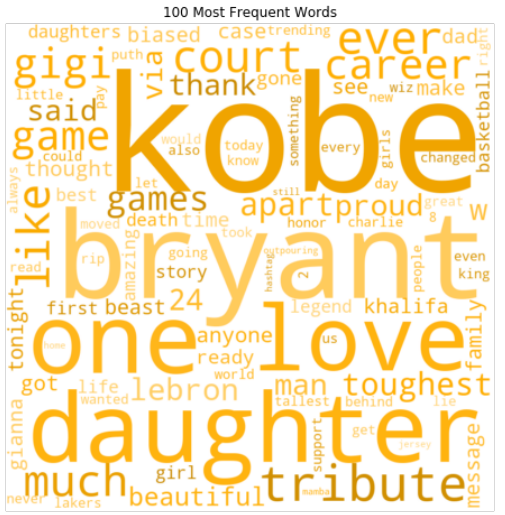
Collecting, Cleaning, Analysis and Results

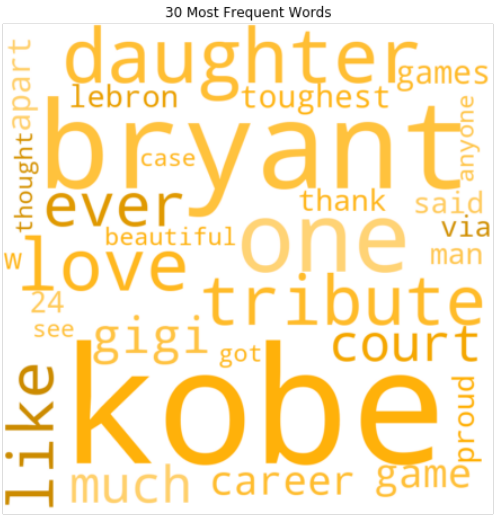
- The first exploratory analysis implemented the statistical analysis and the Word Cloud to show that there are more positive sentiment tweets were found than the negative comments and the topics that were mentioned the most among the Tweets.

https://github.com/jyeakins/MSADS_Portfolio

IST 736: Text Mining







https://github.com/jyeakins/MSADS_Portfolio

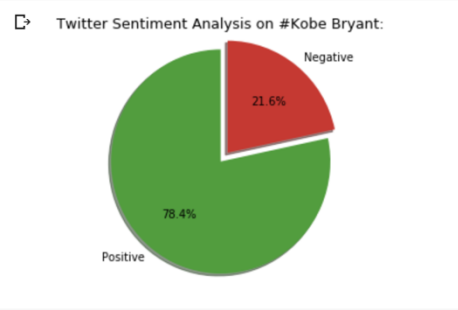
School of Information Studies
Syracuse University

IST 736: Text Mining

- The sentiment analysis (Fig.13) showed the statistics of sentiments of tweets with more accurate results, and the SVM (Fig.14) was used to predict based on training data, whether a tweet had a negative or positive or neutral sentiment.
- It also categorized and sorted the words that had an extremely negative sentiment as well as very positive sentiment.
- Lastly, in order to identify the possible location for campaigns around the country, this study has identified the tweets location with positive sentiments.

https://github.com/jyeakins/MSADS_Portfolio

IST 736: Text Mining



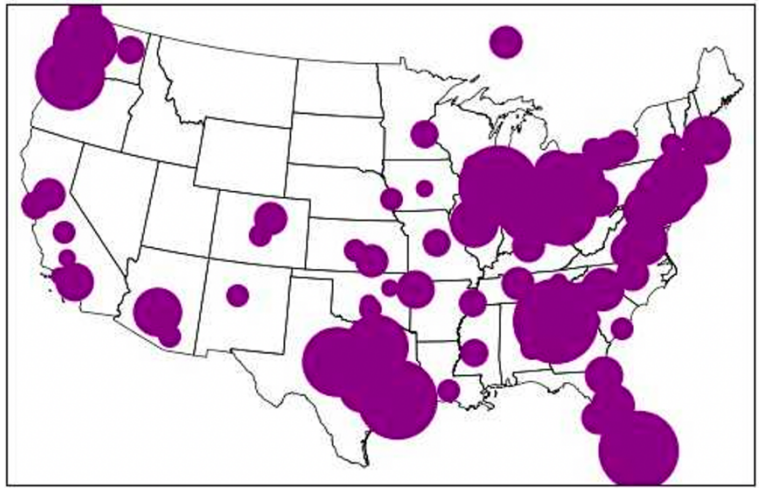

```
[ ]  1 #Number of positive reviews and Negative reviews
     2 df_pos = df[df["sentiment"] > 0.0]
     3 df_neg = df[df["sentiment"] < 0.0]
     4 print("Positive Tweets:", len(df_pos))
     5 print("Negative Tweets:", len(df_neg))
```

```
Positive Tweets: 15509
Negative Tweets: 4285
```

	precision	recall	f1-score	support
0	0.93	0.88	0.90	1294
1	0.95	0.98	0.96	4762
2	0.97	0.96	0.96	4686
accuracy			0.96	10742
macro avg	0.95	0.94	0.94	10742
weighted avg	0.96	0.96	0.96	10742

```
Very negative words
(1.6870859064292206, 'grief')
(1.6957342902477337, 'dead')
(1.6994033892559843, 'sick')
(1.7019223883214432, 'mad')
(1.7037626643486024, 'mean')
(1.7246460579721141, 'fuck')
(1.7412120522691208, 'dusty')
(1.7577548784878836, 'fucking')
(1.7868544144883602, 'killed')
(1.8112370221309277, 'failure')
(1.8120197064684497, 'shocking')
(1.8127894196564907, 'tough')
(1.8741911413365382, 'hard')
(1.9367371595480731, 'bad')
(1.9416706649852078, 'sorry')
(1.9457149847401977, 'taller')
(2.0155722515033796, 'disgusting')
(2.0763351566671871, 'crazy')
(2.0888265573505382, 'devastating')
(2.0917534631129984, 'insane')
(2.2450293953803895, 'painful')
(2.3029912049725231, 'awful')
(2.3990339411417086, 'terrible')
(2.4401947726343221, 'hate')
(2.4847292013398308, 'tragic')

positive words
(-2.7535460316172085, 'best')
(-2.4460983823176541, 'proud')
(-2.3158083552172397, 'special')
(-1.9402401876719395, '2000')
(-1.8857058617048823, 'gifted')
(-1.8262404212499677, 'wome')
(-1.7745417397809198, 'easily')
(-1.7206964405006897, 'great')
(-1.7149111144335394, 'kid')
(-1.7081164373677342, 'alets')
```



IST 736: Text Mining

Reflection

- From collecting the data by writing scripts in Python to analyzing the data by choosing proper tools and techniques, it reminded that how challenge a project can be if the instruction or guidelines were not existed.
- It was great opportunity to learn various private policies and how data need to be handled if the data was collected from the public available source
- One thing that brought the attention was, even if it is a public source, the data does not represent the whole population which could lead to false assumption or conclusion.

https://github.com/jyeakins/MSADS_Portfolio

Learning Objectives

- This portfolio is the best example of the successful student who acquired all learning objectives throughout the courses by completing the challenging projects.
- The data collection was performed in many different ways such as writing a script to automate the process or download from a website, the database created, stored the data and managed to give tools for users, the collected data was analyzed by using data analysis and statistical tools and techniques such as Regression, Support Vector Machine, Random Forest, Classification, Clustering, or Bernoulli.
- There were charts, graphs, and other visualization accompanied to show the analysis in a graphical way to represents the results from the analysis.

https://github.com/jyeakins/MSADS_Portfolio

Learning Objectives

- The presentations of the project helped to learn how to compose slides that easy to read for people who does not involved in the project.
- The communication skills were also developed to use easy terms and give deliverables in a non-technical manner so that others can understand what the problem was to solve and how to implement the solution
- There were many other aspects learned throughout the projects, such as how to manage the timeline, how to be a team player, what to consider when collect the data, and etc.

https://github.com/jyeakins/MSADS_Portfolio

Conclusion

- From Applied Data Science program, the learning objectives challenged students to adapt new ways to view the issues around the world.
- By having perspective of data scientists and the methods to solve the problem, it gives the opportunity to orchestrate the skills that learned from the program; the data collection, analysis, strategy and make decisions and implementation.
- The data scientists from Syracuse University will have proper skills and techniques to solve a variety of problems in the organization.

https://github.com/jyeakins/MSADS_Portfolio

Thank you

Template Retrieved From:
https://styleguide.ischool.syr.edu

School of Information Studies
Syracuse University

References:

- Eakins, J. Y. (2019) IST 659: Data Admin Concepts & Database Management. Retrieved from https://github.com/jyeakins/MSADS_Portfolio/tree/master/IST659
- Eakins, J. Y. (2019) IST 687: Introduction to Data Science. Retrieved from
- https://github.com/jyeakins/MSADS_Portfolio/tree/master/IST687
- Eakins, J. Y. (2020) IST 707: Data Analytics
- https://github.com/jyeakins/MSADS_Portfolio/tree/master/IST707
- Eakins, J. Y. (2020) IST 736: Text Mining
- https://github.com/jyeakins/MSADS_Portfolio/tree/master/IST736
- Nilsson, R., 2007. Consistent feature selection for pattern recognition in polynomial time. J. Mach. Learn. Res.