Portfolio Milestone

Jessica Kwon

Syracuse University – M.S. Applied Data Science

SUID – 425057708

Email – jkwon08@syr.edu

[Abstract 3](#_Toc22582028)

[Portfolio Milestone 4](#_Toc22582029)

[List of Projects 5](#_Toc22582030)

[Data Analysis on Transaction from a bakery 5](#_Toc22582031)

[Class: IST 652 – Scripting for Data Analysis 5](#_Toc22582032)

[Project Description 5](#_Toc22582033)

[Cryptocurrency Sentiment Analysis 8](#_Toc22582034)

[Class: IST 736 – Text Mining 8](#_Toc22582035)

[Project Description 8](#_Toc22582036)

[Big Mac Index Analysis 12](#_Toc22582037)

[Class: IST 719 – Information Visualization 12](#_Toc22582038)

[Project Description 12](#_Toc22582039)

[Solar Panel Marketing 14](#_Toc22582040)

[Class: MAR 653 – Marketing Analytics 14](#_Toc22582041)

[Project Description 14](#_Toc22582042)

[Conclusion 17](#_Toc22582043)

[Reference 17](#_Toc22582044)

Abstract

The goal of the portfolio required for the Applied Data Science (ADS) program at Syracuse University is to highlight and to present all the knowledge obtained throughout the program. The ADS program provides students an opportunity to collect, organize, analyze, and report meaningful insights in data from various business domains using tools and techniques that they learned. Courses such as Natural Language Processing (IST 664), Text Mining (IST 736), Scripting for Data Analysis (IST 652), Big Data Analytics (IST 718), Information Visualization (IST 719), and Marketing Analytics (MAR 653) prepare students to not only analyzing and modeling data using Python and R, but also reporting with business key metrics and actionable recommendations.

Keywords: data science, natural language processing, data analysis

Portfolio Milestone

Applied Data Science program is an interdisciplinary program that students participate more than 10 projects throughout the program to learn in a broad range of areas related to data science. According to 7 objectives of the program, students can do the following:

1. Describe a broad overview of the major practice areas of data science.
2. Collect and organize data.
3. Identify patterns in data via visualization, statistical analysis, and data mining.
4. Develop alternative strategies based on the data
5. Develop a plan of action to implement the business decisions derived from the analyses.
6. Demonstrate communication skills regarding data and its analysis for mangers, IT professionals, programmers, statisticians, and other relevant professionals in their organization.
7. Synthesize the ethical dimensions of data science practice.

# List of Projects

## Data Analysis on Transaction from a bakery

### Class: IST 652 – Scripting for Data Analysis

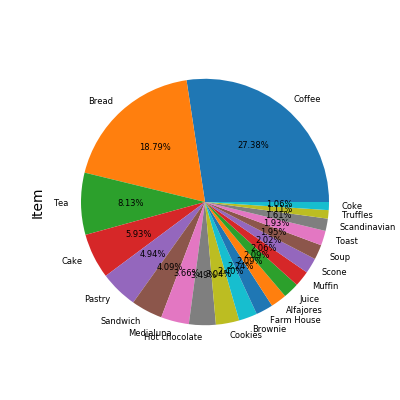
IST652 is the intro class for scripting in python that students learn to solve problems of accessing and preparing data in the forms of structured, semi-structured, and unstructured data. Throughout the course, students deal from simple spreadsheet data to HTML/XML or JSON file.

### Project Description

The project used the data called ‘Transaction from a bakery’ from Kaggle. A market basket analysis was performed on the data. Market basket analysis is one of the data science techniques often used by retailers to uncover associations between items. From the analysis, it allows retailers to identify the relationship between the items that people purchase and to understand customer behaviors on them. Python is used for data importing, pre-processing, modeling, and visualization.

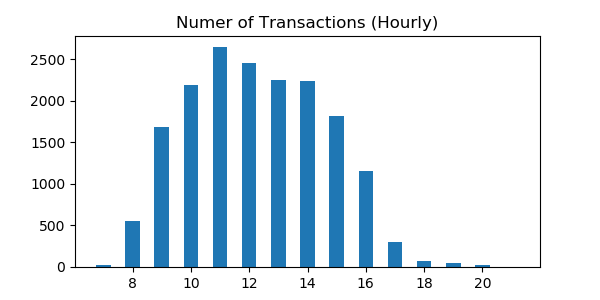
This project mainly focused on scripting in python to perform exploratory data analysis and market basket analysis. Data contains over 20000 transactions with four columns – transaction date, time, transaction number, item. A total of nine business questions are defined that could be useful information for increasing revenue, such as increasing the number of customers by offering combination deals. Based on these questions, imported data is pre-processed accordingly – any duplicate transaction and all the transactions outside of the chosen time frame are removed, date and time columns both are merged and segmented.

Figure



Out of 91 kinds of items at the bakery, the top 20 sold items at a bakery are shown above. (Figure 1). This result provides business owners to have a better idea of what are the items they should keep and update in the menu. Besides identifying popular items that are sold, customer behavior analysis – popular time by hourly (Figure 2), daily (Figure 3), monthly (Figure 4) – are visualized after subgrouping with segmented time variables.

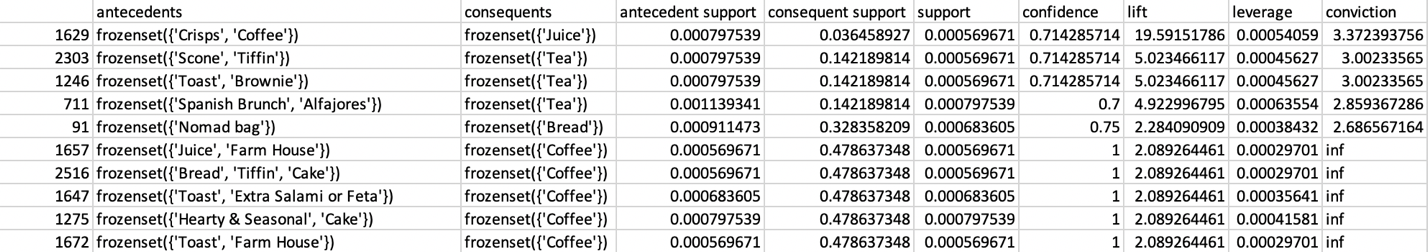
Figure



|  |  |
| --- | --- |
| Figure | Figure |

The main components of the market basket are support level, confidence level, and lift level. A support level means how frequently the itemset appears in the dataset. A confidence level means how often the rule has been found in the dataset. The lift level means the ratio of the support of the left-hand side (antecedent) of the rule co-occurring with the right-hand side, divided by the probability that the left-hand side and right-hand side co-occur if the two are independent. A lift greater than 1 suggests that the presence of the antecedent increases the chances that consequent will occur in a given transaction. The minimum support level for this project set to 0.0005 due to the size of the dataset. For a general rule, a support level for the market basket is usually around 60%. So, the interesting rule generated for the project is set to 65% and a minimum lift level to be 1.4 as well. The first 10 of the interesting rule for the baker is shown below (Figure 5).

Figure



Cryptocurrency Sentiment Analysis

### Class: IST 736 – Text Mining

IST 736 provides students to learn text mining techniques that can apply in information organization and access, business intelligence, social behavior analysis, and digital humanities. Throughout the course, students can perform natural language processing (NLP) for finding patterns in large text corpora. Students use a Python-based command-line tool called scikit-learn and a range of machine learning techniques, such as Naïve Bayes and support vector machine (SVM).

### Project Description

Cryptocurrency is an internet-based medium of exchange that uses strong cryptography to secure financial transaction by controlling the units of the medium, and by verifying the transfer of the assets (Schueffel 2017). The first introduction of the electronic payment medium was in 1983 by David Chaum, there have been many events such as bitcoin arise and price crash due to big scams. From this project, sentiment analysis of the public’s opinion on cryptocurrency was performed using machine learning techniques.

Using Twitter GET API call, tweets having ‘#cryptocurreny’ since 2013 were extracted. After the extraction of raw data, several pre-processing – removing duplicates and similar strings in excel, cleaning tweets by regex replace logic, generating sentiment scores for tweets using VADER analysis, splitting, and lemmatizing the tweets using WordNetLemmatizer() – were performed. As an exploratory data analysis (EDA), wordcloud visualization was created to snapshot the public’s common phrase and word in tweets (Figure 6).

Figure



Also, the top 10 most frequent words by dates were analyzed using tf-idf techniques. Due to the size of the tweets, comparison analysis on most frequent words counts basis versus normalized basis was limited to 11 days (Figure 7 and 8).

Figure . Top 10 most frequent words by dates (Count)

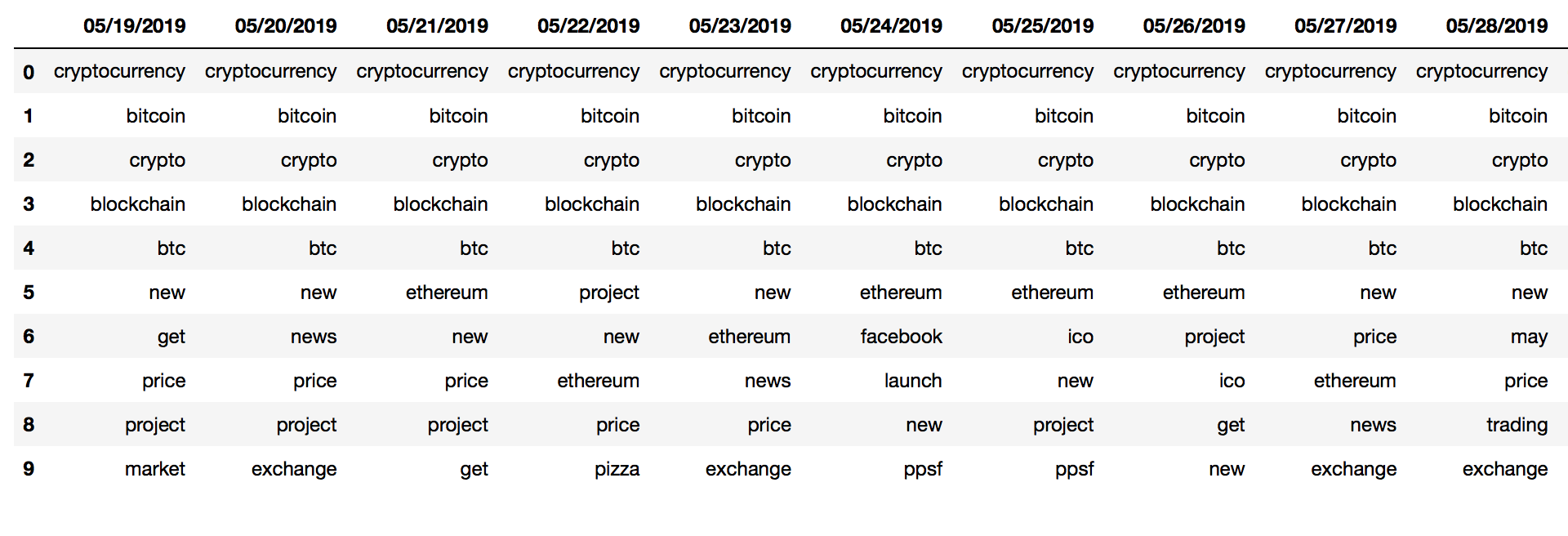
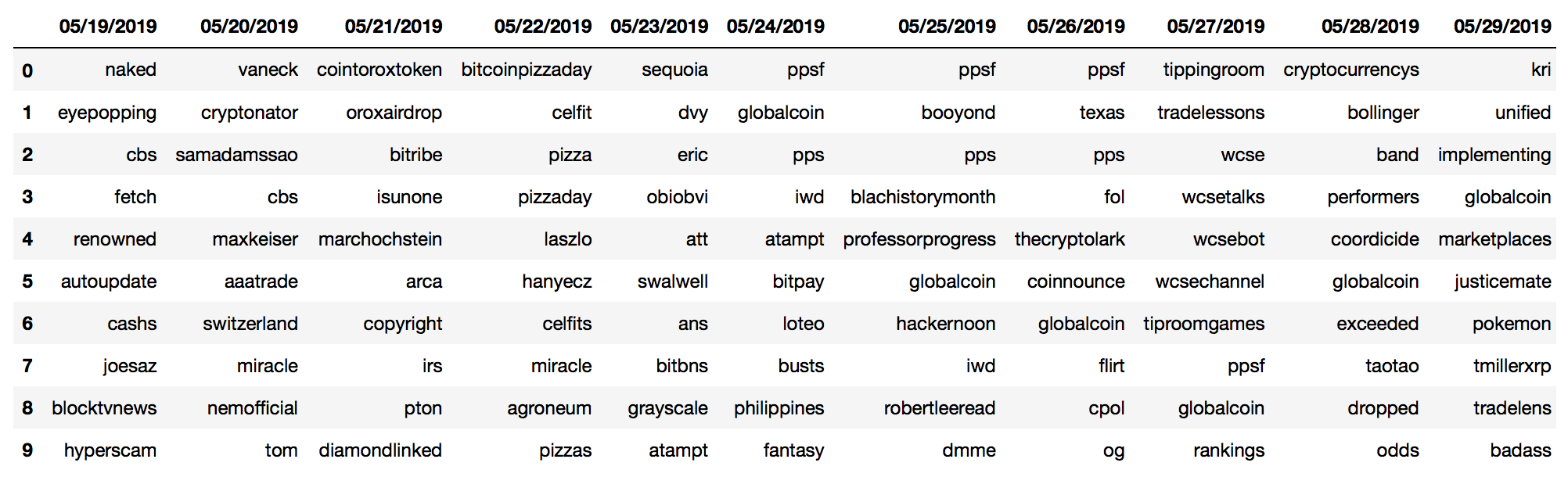


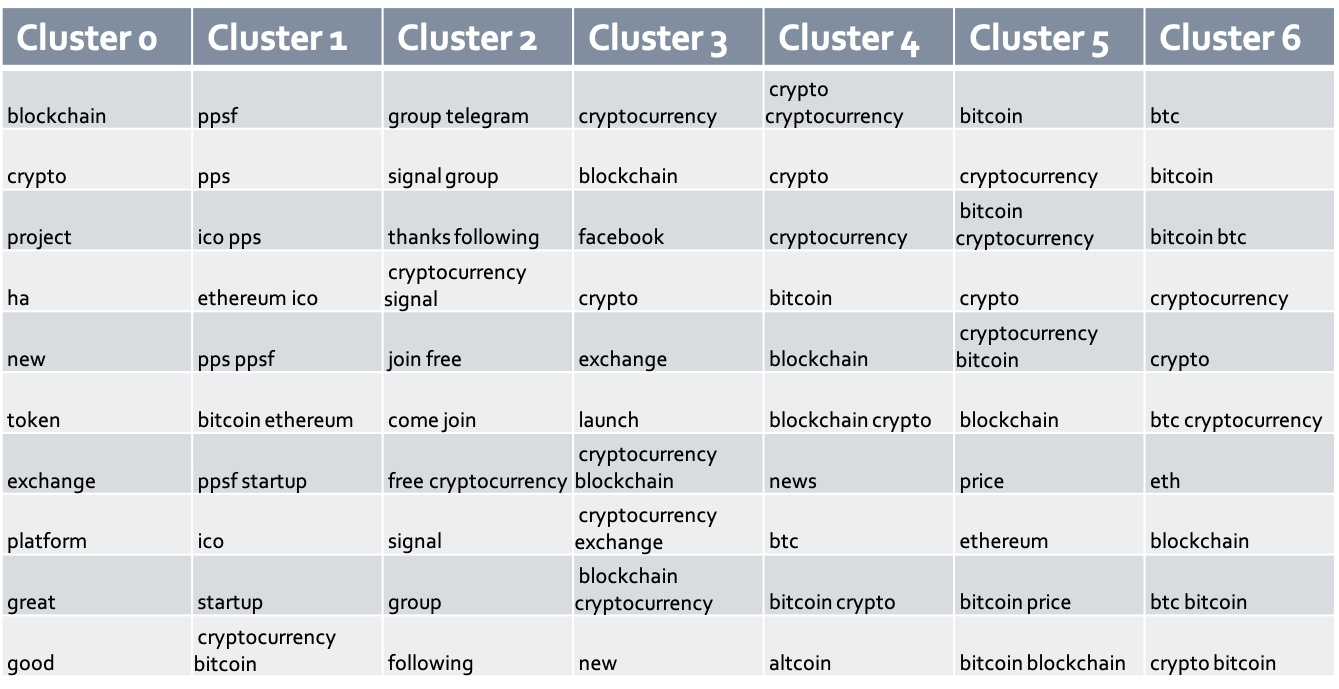
Figure Top 10 most frequent words by dates (Normalized using Tf-idf)



Not only by the dates, but top terms per cluster were also analyzed (Figure 11). Because of the uncertainty of the number of groups in public sentiments, a simple unsupervised machine learning algorithm, K-means, was performed to select the optimal number of clusters. The result of K-means clustering analysis, 7 clusters would be ideal clusters to categorize the collected tweets even though cluster 0 contains more than half of the extracted tweets (Figure 9 and 10).

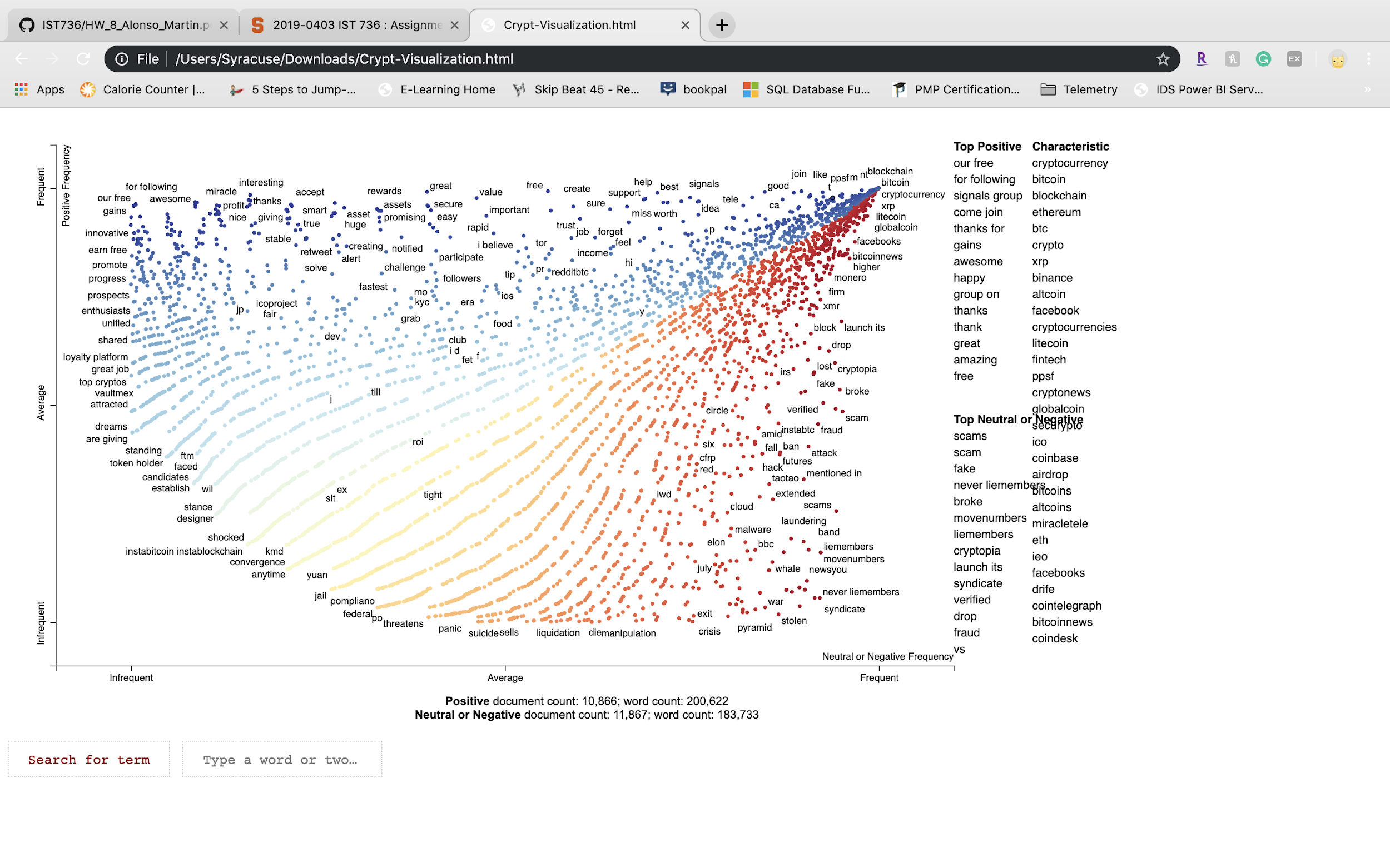
|  |  |
| --- | --- |
| Figure . Elbow graphs of K-means | Figure . Tweet counts by sentiment label and cluster types |

Figure . Top 10 terms for 7 clusters



After the EDA, sentiment prediction models were built using the support vector machine (SVM). For SVM modeling, five steps were performed. First, data were vectorized into unigram/bigrams, removed the English stop-words, then set to lower case in tf-idf setting. Then, sentiment labels were added to the vectorized data frame. Due to the unbalanced result of sentiment label group – the negative sentiment was approximately 10% of the data – sentiment labels were recoded as positive and not positive. Data were split into training, and testing sets (60/40) then iterated to test SVM parameters. After fitting the model on the test data set, the SVM model has 86.2% accuracy with 86% precision. Each sentiment has many featureset vocabularies. The featureset graph was generated using the SVM prediction model (Figure 12).

Figure . Featureset Word by sentiment group



## Big Mac Index Analysis

### Class: IST 719 ­– Information Visualization

From the IST 719 course, students learn various data visualization techniques to design data visualization. Students are introduced to R and Adobe Illustrator to perform simple data cleaning to visualizing data. Systems used often in the class are maps, rworldmap, ggplot2, wordcloud2, and RColorBrewer.

### Project Description

The Big Mac index helps understand currency valuations and is so useful for economists and especially currency traders. The Big Mac index has been published annually by the Economist since 1986 and is rated as a clear indicator of a country’s purchasing power. Big Mac index data for the project was from statista website – URL is the following (<https://www.statista.com/statistics/274326/big-mac-index-global-prices-for-a-big-mac/>). The original data for this index analysis included 1162 rows with 19 attributes. Before pre-processing the data, business questions and the targeted audience were defined. Three questions to answer are the following:

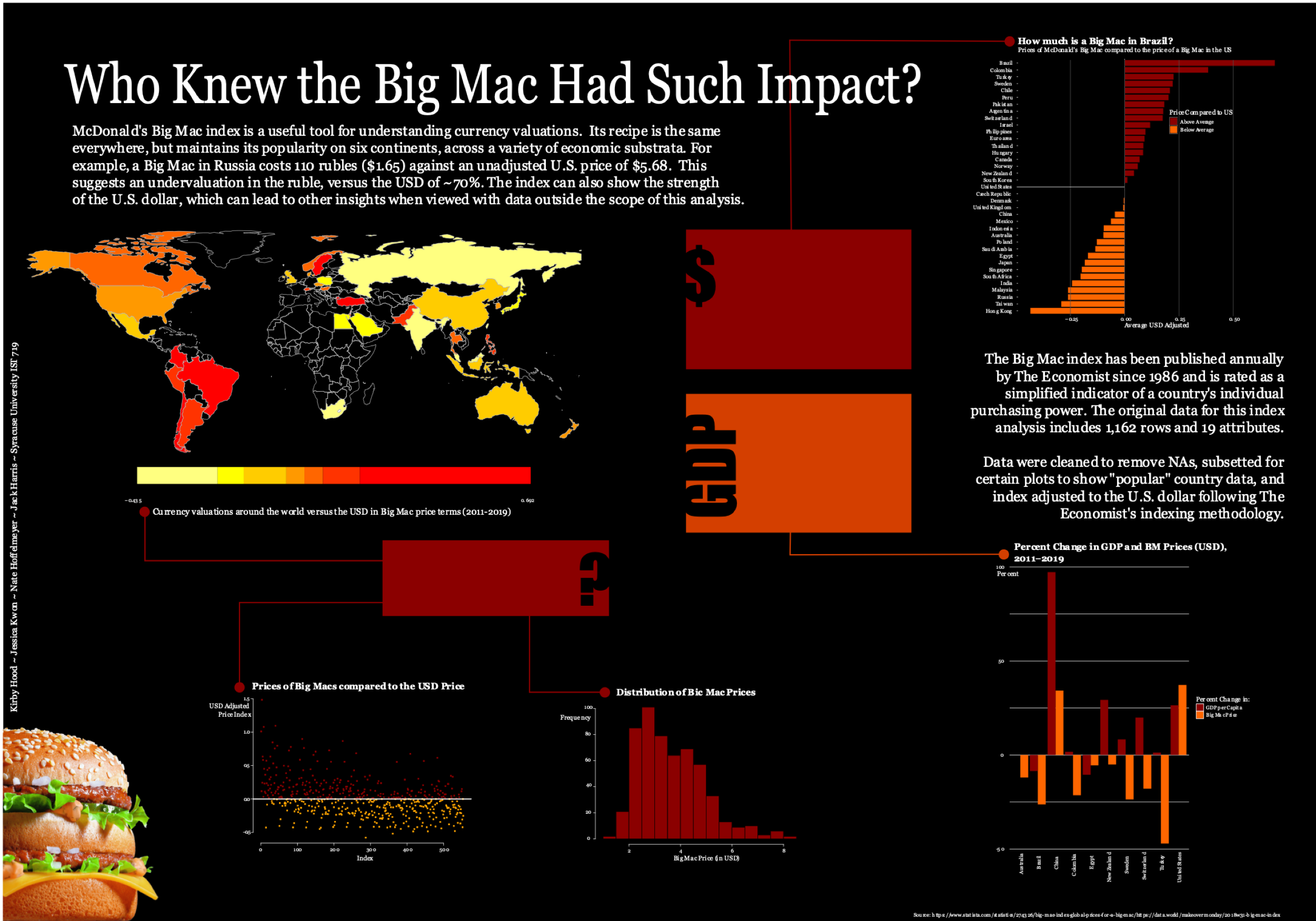
1. Whether or not the Big Mac prices can be a proxy for currency inflation;
2. If there is a relationship/trend between a country’s economic status and the Big Mac prices adjusted to the U.S. dollar; and
3. If there is a correlation between the GDP of highlighted countries and Big Mac prices.

Data were cleaned to remove N.A.s, subset for specific plots to show ‘popular’ country data, and index adjusted to the U.S. dollar following the Economists indexing methodology. In order to answer the first question, three graphs – scatter plot, distribution bar graph, and world map – were generated. joinCountryData2Map() function was used to join the Big Mac Index data with country reference data to a map. The generated world map showed currency valuations around the world versus the USD in Big Mac Price term (2011-2019). USD adjusted field was used for the graph. The darker the color is, the more over valued the big mac price is. According to the world map, Big mac prices in North America and the South America continents are generally over-valued and were under-valued in the Asia and Australia continents compared to U.S. big mac price (Figure 13).

For the second question, the diverging bar chart was generated using ggplot(). The graph is not only showing average adjusted big mac price in USD but also showing the binomial groups of above/below average big mac price. For the last question to see if there is any correlation with GDP, key countries – Australia, Brazil, China, Colombia, Egypt, New Zealand, Sweden, Switzerland, Turkey, United States – were chosen based on the answers from the second question above. GDP per Capita and Big Mac Price fields were a subset. Percent changes in GDP per capita and in big mac prices were calculated for the third graph using two-year points of 2011 and 2019.

After analyzing data, and a big mac price in certain countries has an undervaluation or an overvaluation in their currency. For example, a Big Mac in Russia cost 110 rubles($1.65) against an unadjusted U.S. price of $5.68. This example suggests an undervaluation in the ruble, versus the USD of -80%. The index can also show the strength of the U.S. dollar, which can lead to other insights when viewed with data outside the scope of this analysis.

Figure



## Solar Panel Marketing

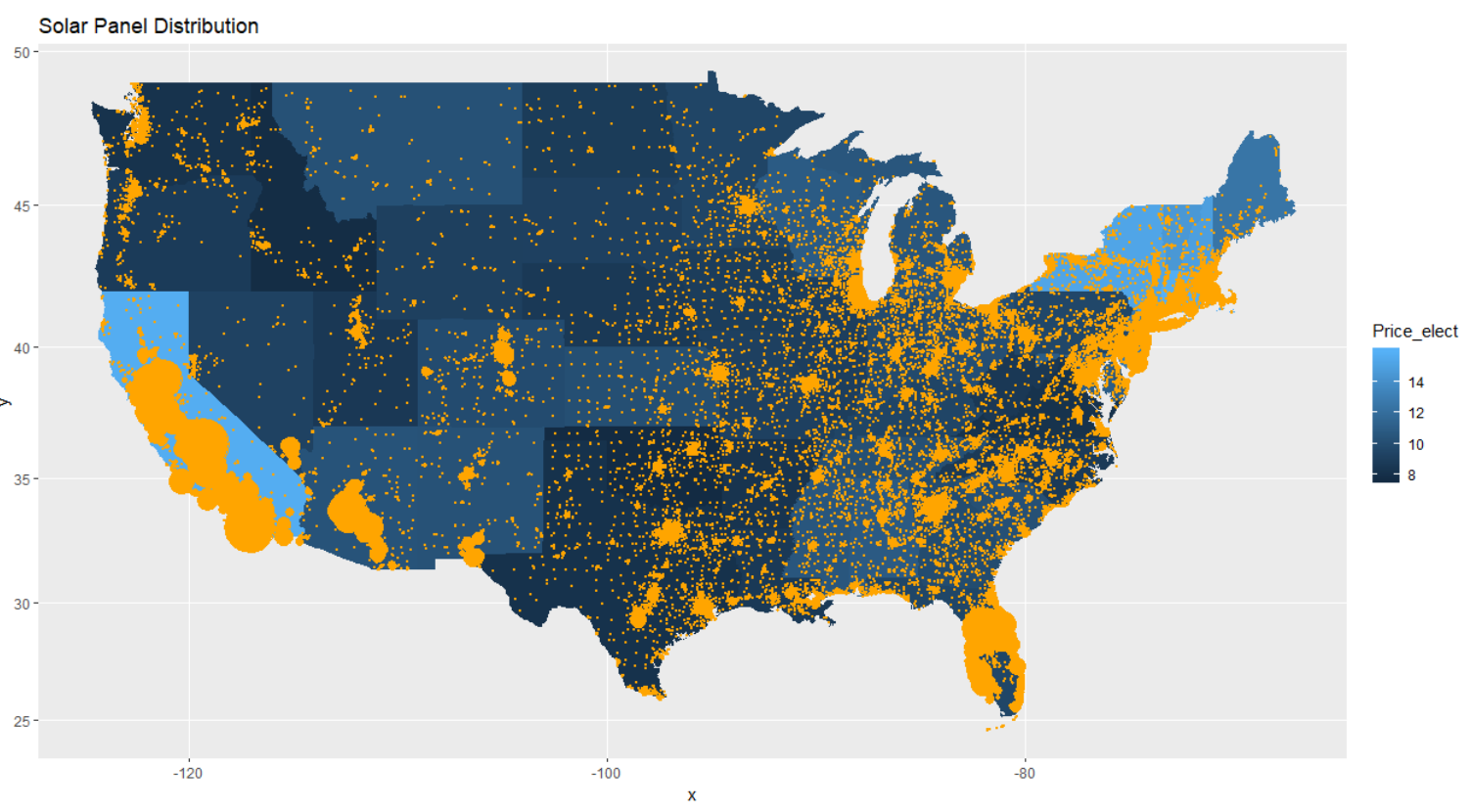
### Class: MAR 653 – Marketing Analytics

MAR 653 is provided by Whitman Business School that focuses on developing marketing strategies and resource allocation decisions driven by quantitative analysis. Students learn about market segmentation, market response models, customer profitability, product recommendation systems, churn predictions, media attribution models, and resource allocation throughout the course.

### Project Description

With the growing attention to global warming, the widespread transition to renewable energy like solar panels became popular. For this group project, the analysis was performed as a business owner's perspective in the solar panels market. The data set is found on Kaggle (<https://www.kaggle.com/tunguz/deep-solar-dataset>). The dataset has 72000 observations and 169 attributes. The goal of the analysis is to identify regions and populous that would have a high propensity to invest in solar energy. As is evident, some states with higher energy costs show a more extensive solar panel distribution, and some do not (Figure 14).

Figure . Distribution of Current customers of solar panel by price of electricity



In order to find potential customers with a high propensity for solar panel installation, significant variables correlated with owning solar panels need to be identified. These variables then were narrowed again to the attributes of existing solar panels consumers. From the initial exploratory data analysis, seven variables – high electricity consumption, large populations with an education, high income areas, higher overall earth temperature, high population older than 45, areas with an equal economic standing, people who prefer alternate means of transportation—were found as a driver variable to solar panel sales.

Using unsupervised learning, population data were grouped into four segments (Figure 15). Of the four segments, consumers in one cluster ware almost split 50/50 on owning solar panels vs. not. This cluster was used for psychographics and demographics in determining potential customers acquired. Another cluster had a population with 92% solar panel customers. This cluster was used for determining what the current customer segmentation looks like (Figure 16). From this clusters, the current consumers have 12 strongly correlated attributes, which are: high average household income; eco-friendly; the age of 45-54; love rebates; education; have a mortgage; expensive electricity price; high sales tax area; high Gini index; warn weather; high rate of employment; and democrats.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Figure . Elbow graph of k-means clustering analysis | Table . Population break by cluster   |  |  |  | | --- | --- | --- | |  | 0 | 1 | | 1 | 2940 | 5870 | | 2 | 233 | 1741 | | 3 | 948 | 10356 | | 4 | 5339 | 17665 | |

Using these correlated attributes from clustering analysis and sales driver attributes, the potential customers were defined with eight aspects – areas open to incentives, people with more land, fuel-based heating, frost days, lower-income, massive industrial area, driving alone, and low sales tax. After analyzing potential customer segmentation, the chosen segments to be targeted are the customer who are older, and who choose environmentally friendly options for transportation. Then, the location to be targeted are catered towards areas with large industries that would generate more solar power from 'frost days.' Current customers and potential customers both are rebate friendly that the option of working with local authorities on residential rebates was also highly recommended from the analysis.

# Conclusion

Completing the applied data science program from Syracuse results differently based on the student’s own goal and direction of the program. No matter what their focus is, all students in the program can achieve seven objectives of the program.

All four projects taught different practice areas of data science. ‘Data analysis on a transaction from a bakery’ introduced the application of data mining and statistical analysis. ‘Cryptocurrency sentiment analysis’ taught machine learning and cognitive computer development. ‘Big mac index analysis’ was for data visualization and presentation. Lastly, the ‘solar panel marketing’ project was more about market-related data analytics. All these projects used various ways to collect and preprocess the data. Statistical analysis and visualization often used to identify patterns and answers for business questions. Typically, alternatives strategies and multiple recommendations were made to provide better solutions in terms of practicality and ethicality.

# Reference

Schueffel, Patrick (2017). [*The Concise Fintech Compendium*](http://www.heg-fr.ch/EN/School-of-Management/Communication-and-Events/events/Pages/EventViewer.aspx?Event=patrick-schuffel.aspx). Fribourg: School of Management Fribourg/Switzerland.

Chaum, David (1983). *Blind Signatures for Untraceable payments*. University of California Santa Barbara, CA