Tessa Blankenship

TDBLANKE

[tdblanke@syr.edu](mailto:tdblanke@syr.edu)

**Project Portfolio – Applied Data Science**

Data Science includes several areas of study. The website Techterms defines “Data science is the study of data. It involves developing methods of recording, storing, and analyzing data to effectively extract useful information. The goal of data science is to gain insights and knowledge from any type of data — both structured and unstructured.” (Christensson, 2017). The Applied Data Science program has provided courses and projects that are in line with data science definition.

In order to understand and gain insights from data, there needs to be a business question or problem that is being addressed. Identifying the need is the first step in the process. Knowing the need in the business identifies the question that needs to be answered using the data. The question identifies the data that is needed. Data can come in different forms. Structured, semi-structured and/or unstructured data may need to be obtained to answer the business question. The project examples are focused on structured and unstructured datasets.

Structured data includes data that is stored in a normalized database. There are definitions in place for the data and it is structured in a way that allows for quick extraction. Unstructured data is data that is stored in text files, images, or voice recordings. It is not as easy to extract and analyze, and it will require additional processing and cleaning that may not be required with structured data.

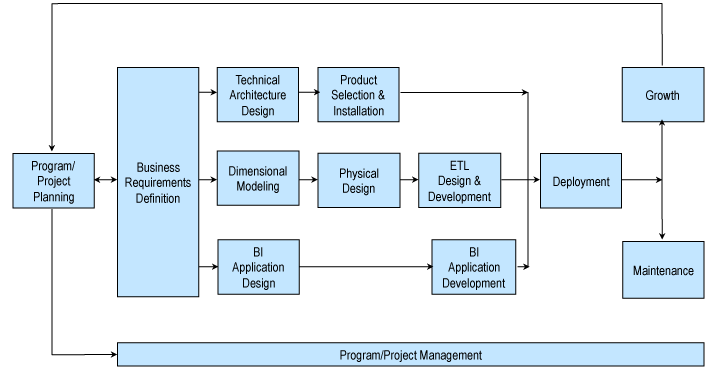
Once the data is identified it will need to be collected, organized and stored. Preparing data that can be used in an organization for continued analysis, can include the creation of a Data Warehouse. Implementing a data warehouse will centralize the company’s data in one repository. The data will be configured in a way that allows for quick and simple queries.

Project One – Data Warehousing

An example of a data warehouse implementation would include creating the ability to analyze data across different systems. If a company had two segments of business, one that ships products, and one that offers movies by mail or on demand, the data would be stored in two different systems. Implementing a data warehouse would integrate both systems into a single database.

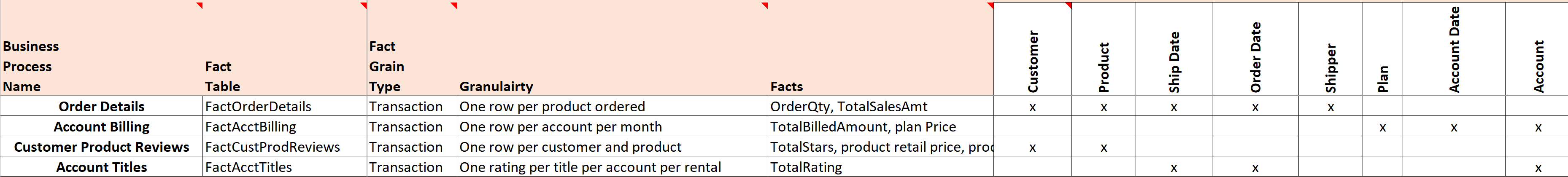
One technique used in implementing a data warehouse is the Kimball Lifecycle. The Kimball Lifecycle was developed by Ralph Kimball, Ph.D. in the early 1980’s. The Kimball Lifecycle focuses on three concepts that are fundamental to the technique: focusing on the business, creating dimensional models for reports or ad hoc querying, and developing the warehouse in iterations. (Kimball, Ross, Thornthwaite, Mundy, & Becker, 2008)

The overall process is shown in the below in the Kimball Lifecycle Milestone diagram. Much of the focus will be on the center portion of the diagram will be covered in the project example. This includes the Business Requirements Definition, Dimensional Modeling, Physical Design, ELT Design & Development, as well as the BI Application Development, and Deployment. The Technical Architecture Design and BI Application Design will not be included in the example project. The Product Selection & Installation is included in the project and consists of Microsoft SQL Server, Integration Services, Analysis Services, and PowerBI.

[](http://www.kimballgroup.com/wp-content/uploads/2012/06/kimball-core-concepts-021.png)

(Kimball DW/BI Lifecycle Methodology., n.d.)

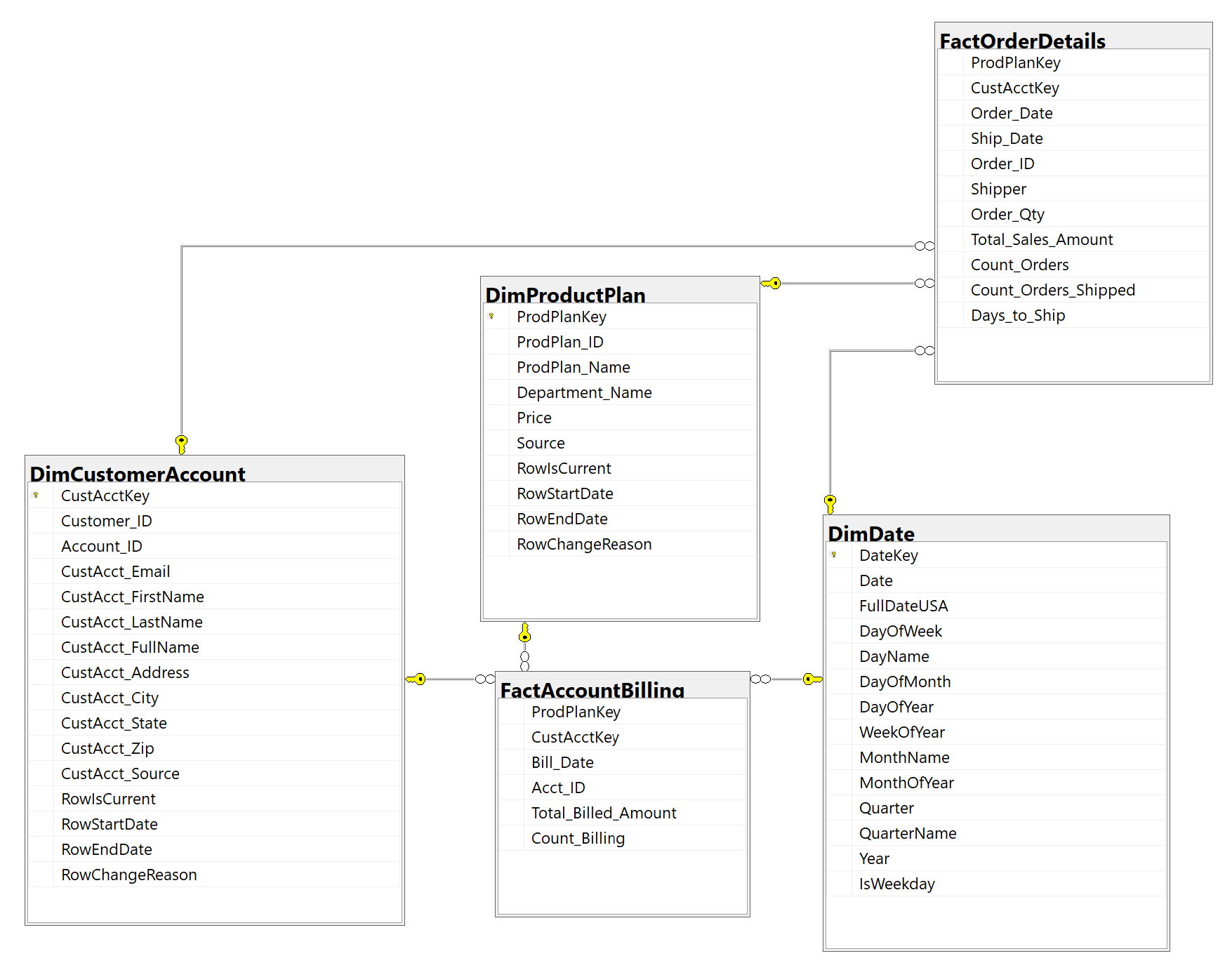
In order to create a data warehouse that would provide value for the company and “focus on the business” (Kimball, Ross, Thornthwaite, Mundy, & Becker, 2008), the key areas need to be understood. Business processes need to be identified in order to compile the fact or event tables, and the shared dimension tables that correspond. A bus matrix is used in the process. The example below is from two databases for two segments of a business. This would be the business requirement definition displayed in the Kimbell Lifecycle diagram.



It is easy to see areas where the dimensions would be similar. Customer and Account would be about the customer, and Plan and Product would be what is being ordered. The fact tables are transactional tables. That means that they are the tables where the individual sales are stored, or the individual reviews are made. To focus on the highest value add for the company, the fact tables that will be used in this example will be the Order Details and Account Billing tables. These tables show the sales amount for the company and can be used immediately for financial analysis.

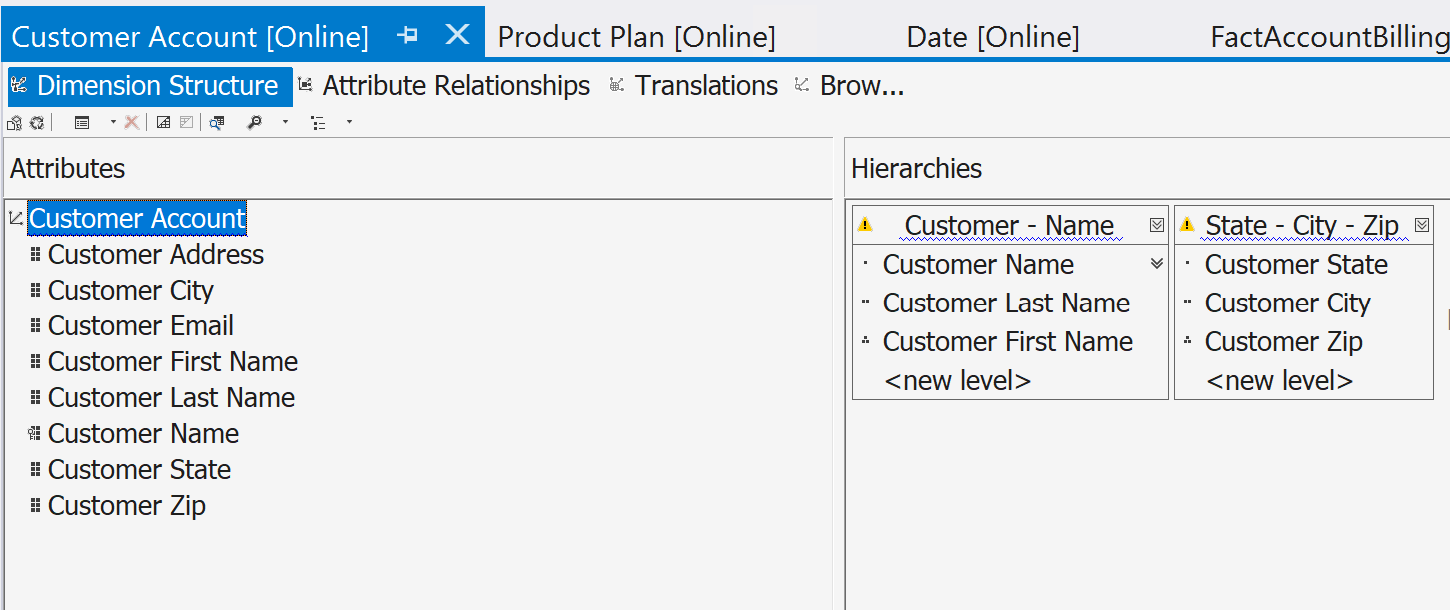
With the fact tables identified, the next step is to identify the dimension tables that will be used. This is the dimensional modeling portion of the Kimball Lifecycle diagram. In this example, the Customer and Account tables will be combined into a CustomerAccount dimension and the Product and Plan tables will be combined into a ProductPlan dimension. There will also be a Date dimension table that will be used to create data hierarchies for easy time series analysis. This data dimension table will have multiple relationships with each fact table. Each date field in each of the fact tables will be linked to the date dimension. This will allow for the date dimension to be used combine the data from both fact tables and display them in one graph.

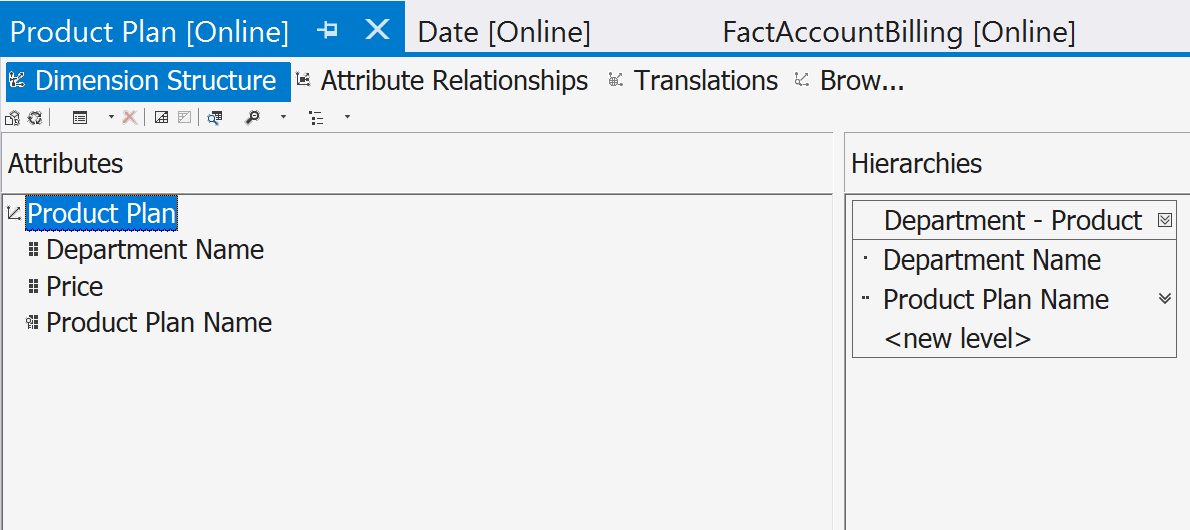
Below are the fact and dimension tables created for this example. This would be the physical design process in the Kimbell Lifecycle diagram. Note that the FactOrderDetails table is from the segment of the company that ships products out to their customers, and the FactAccountBilling fact table is from the segment of the company that offers movies by mail or on demand.

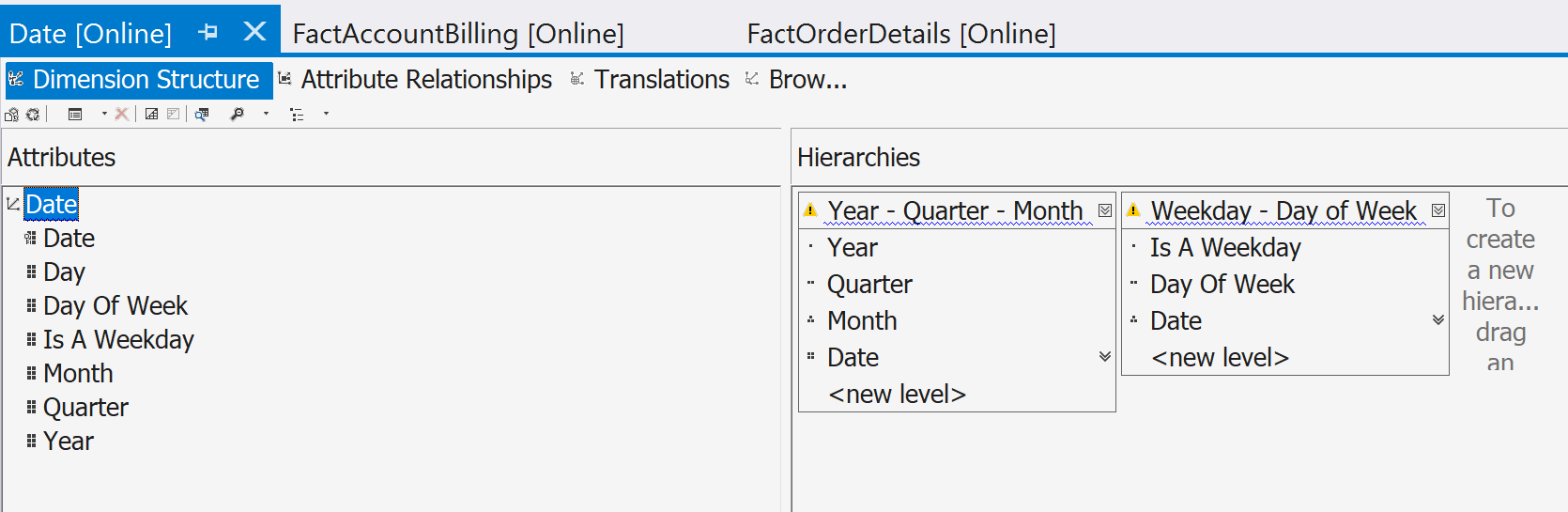


With the customer information from both company’s systems in one table, decisions must be made on how the data should be transformed to create one master customer record. These transformations will occur when the data is loaded into the new DimCustomerAccount table. The goal would be to create rules that will ensure the most accurate form of the data for each customer. An example would be to use the most recent information for a customer’s address when loading it into the table. If one segment of the company has the same customer but a newer address, then the rule would be for that newer address to be used in the DimCustomerAccount table. It could also mean that the case is changed to show the first letter of the name as a capital, when it originally was all lowercase. These rules along with the specification of which field will be used from the source database to each field in the data warehouse tables or target database. This document is referred to as a source to target.

Once the database is set up to receive the data from the original databases into the newly dimensionally modeled using SQL Server Integration Services and the source to target document, hierarchies can be created to make analysis easy and fast using SQL Server Analysis Services. This is the ETL Design and Development and BI Application Development portion of the Kimball Lifecycle diagram. SQL Server Integration Services is the ETL tool that is used in this project, and SQL Server Analysis Services is the BI tool that is used. Below are a few examples of how the hierarchies can be created in SQL Server Analysis Services.

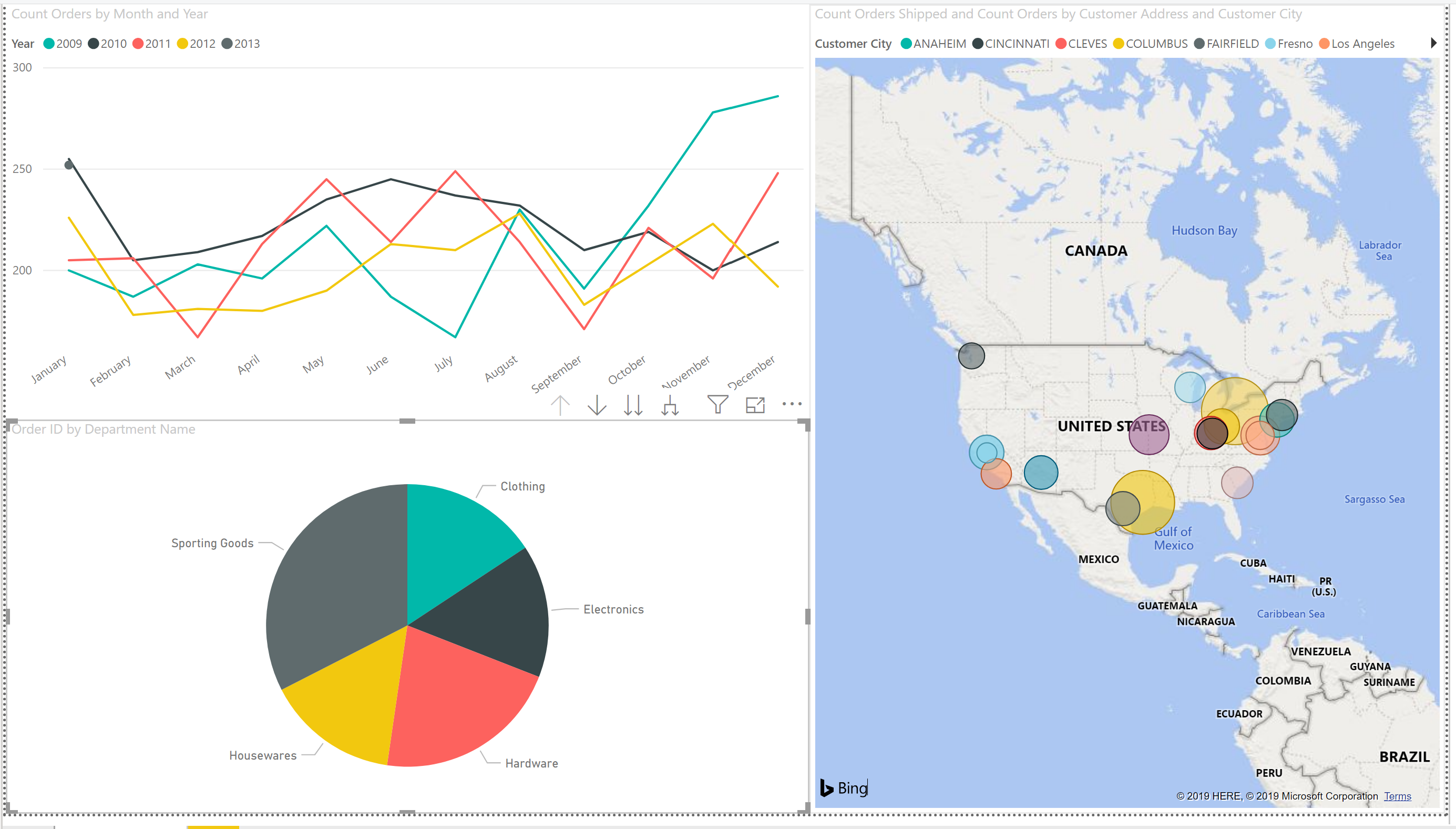


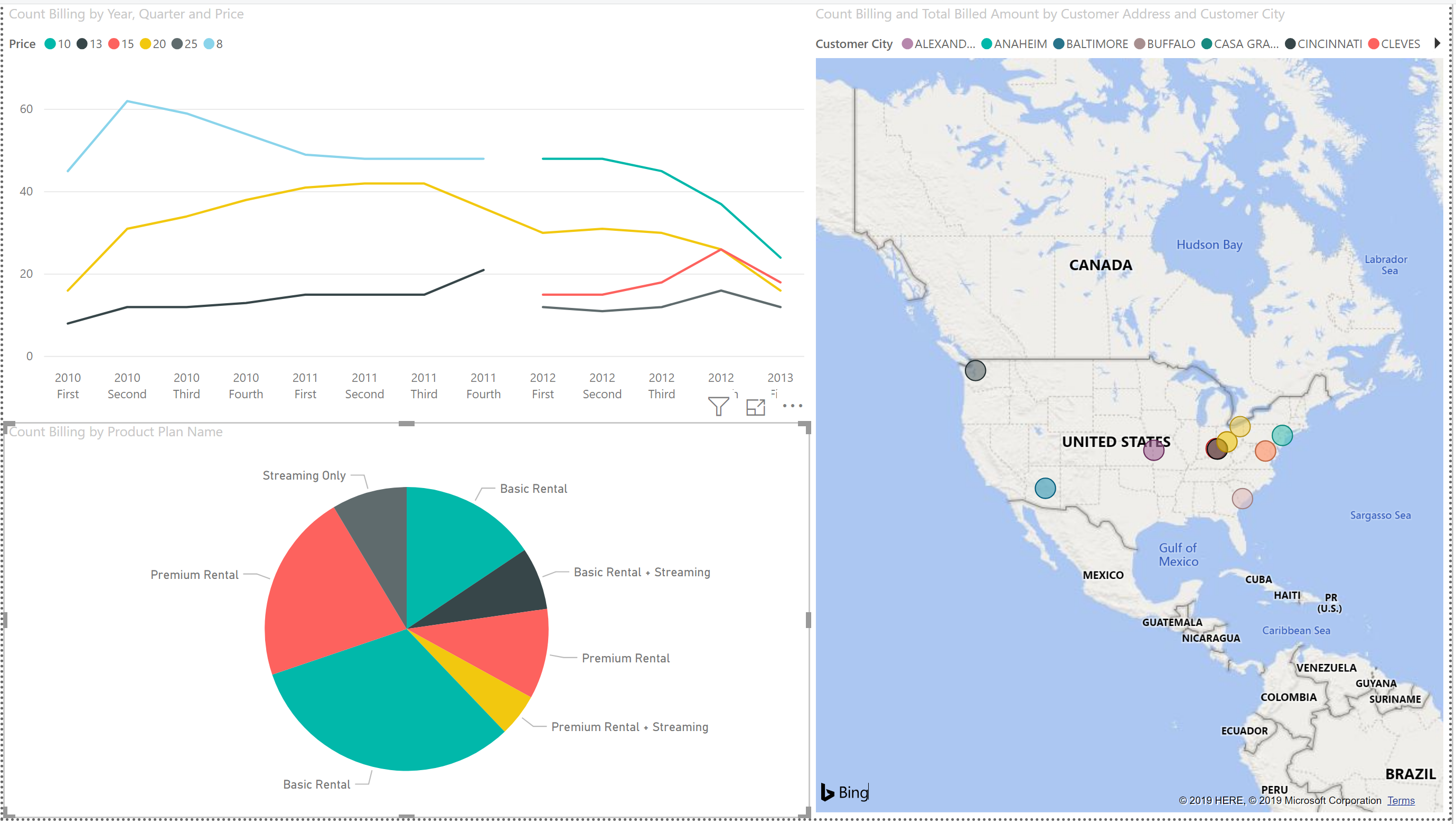




These hierarchies within a dimension and the individual dimension fields are used to create a multidimensional cube. The cube consists of measure as well. These measures are the aggregable values within the fact tables. An example of a measure would be the sum of the total sales amount. Once the dimensions and measures are defined, the cube can be built and deployed. The cube is a preprocessed object that once deployed, will pre-aggregate the measures against the identified dimensions. This pre-aggregation is how the cube makes analysis so quick. It can pre-aggregate millions of rows of data and store it for the end user to choose different combinations of fields to be displayed.

Visualization can then be created for initial analysis of the data. The example below shows reporting that can be created using PowerBI connected to the SQL Server Analysis Services cube.





Project Two - Data Analysis

Visualization and analysis of the data can also be completed using programming languages like R and Python. After the data is obtained, it can be stored like the previous example, and extracted for statistical analysis and data mining. An example of a data mining project would be the analysis of hospital readmission data. This project was analyzed using R programing language.

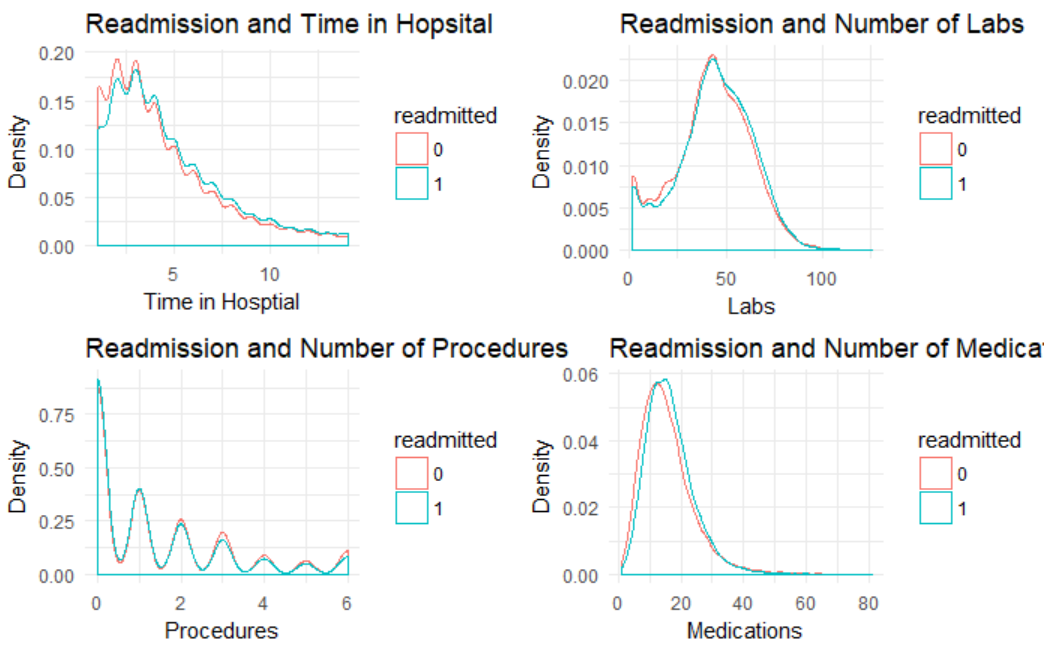
The data mining task included the examination of data found on Kaggle.com. The dataset has twenty five thousand observations and sixty five variables. The target variable is a binary categoric indicator on wither the patient was readmitted to the hospital after being released from the hospital within the last thirty days. The purpose of the analysis is to predict wither a patient is at risk for readmission so that the hospital can intervene. This will help the patient stay healthy after a hospital stay and prevent a reduction in payment that the hospital receives from Medicare if a patient is readmitted.

The dataset consists of eleven thousand four hundred and ten records categorized as readmitted, which means that the baseline for any predictive measure is 45.64%. If the model predicts the readmission with an accuracy rate greater than 45.64% then it is predicting the target variable better than chance. The dataset had been cleaned already, was structured and ready to be analyzed.

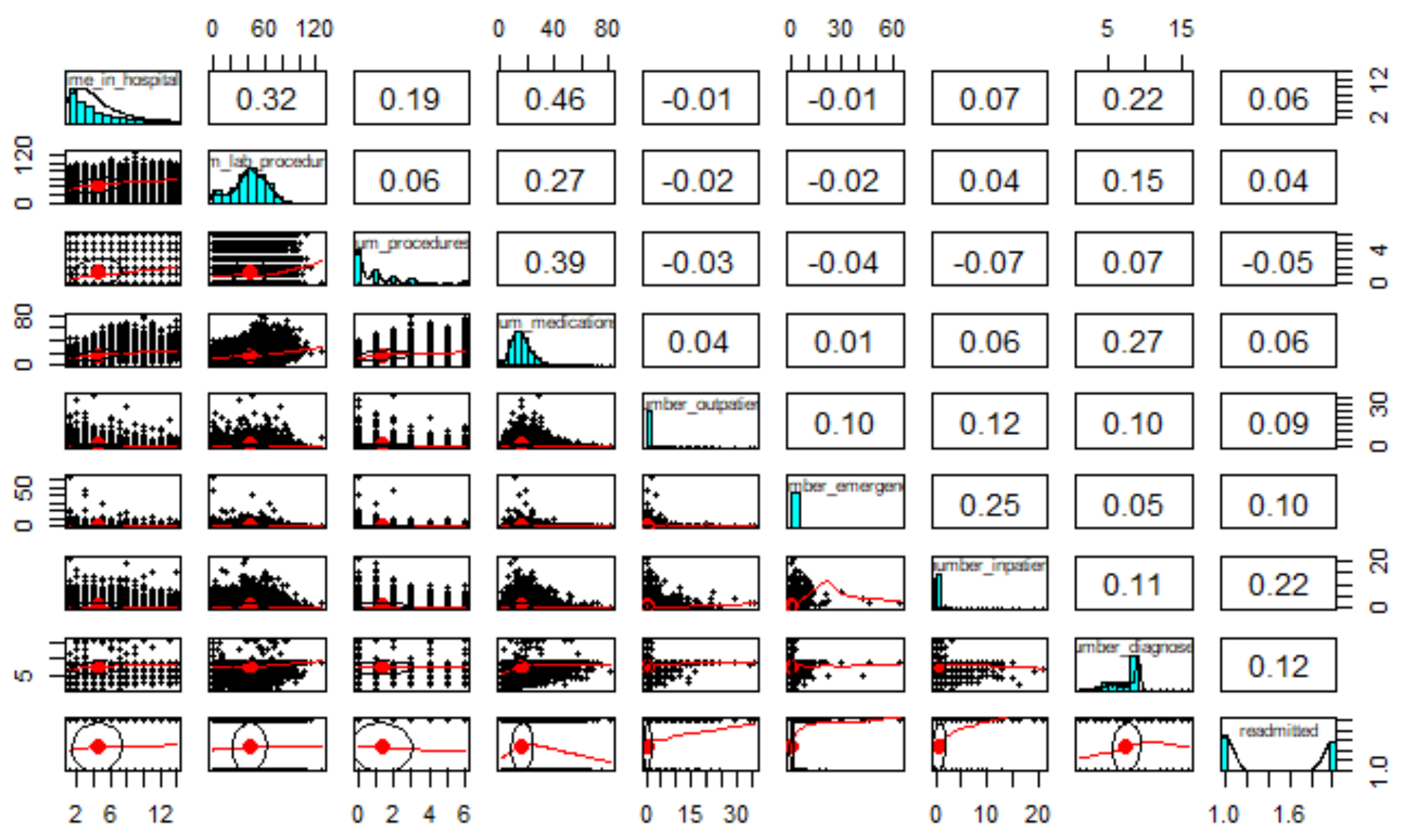
“Data mining is a technology that blends traditional data analysis methods with sophisticated algorithms for processing large volumes of data.” (Tan, Steinbach, & Kumar, 2015). Data mining is part of the Knowledge Discovery in Databases process that begins with Data Preprocessing, Data Mining, and ends with Postprocessing. Preprocessing is the process of selecting features with the highest correlation, normalizing the data, and in subsetting the dataset. This is an important step, especially when considering that the dataset has sixty five variables. Postprocessing is the final process of interpreting the patterns and visualizing the data. (Tan, Steinbach, & Kumar, 2015)

There are two categories that data mining tasks fall under: predictive tasks, and descriptive tasks. Predictive tasks are used to predict the value of the target variable based on the explanatory variables in a dataset. Descriptive tasks are tasked used to find patterns in the dataset in order to identify relationships between variables. (Tan, Steinbach, & Kumar, 2015)

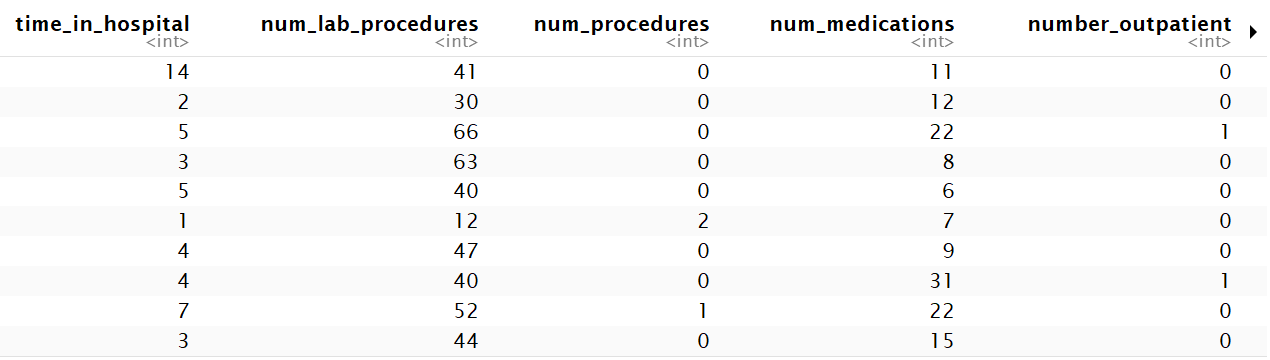
The data mining project includes both predictive and descriptive tasks and will follow the Knowledge Discovery in Databases process. Although the dataset is structured and cleaned, it will still need to be preprocessed to normalize data that will need to be normalized and features will need to be selected. The first descriptive task will be to visualize the density of a few of the explanatory variables. This task is a simple visualization to see if there is a visual correlation that can be seen based on domain knowledge of hospital readmissions. Looking at the below visualization, there is a slight correlation between the time that a patient spends in the hospital and the number of medications that they are taking.



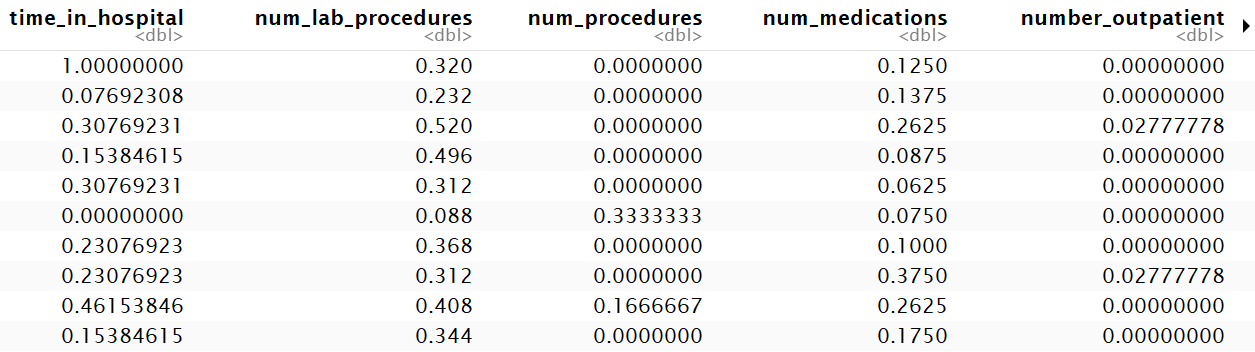
Looking at more of the fields to find correlation with the readmitted indicator, it can be seen that the number of times that the patient was admitted to the hospital, the number of diagnoses, and the number of times that they have gone to the emergency room have a slight correlation with wither they are readmitted or not.



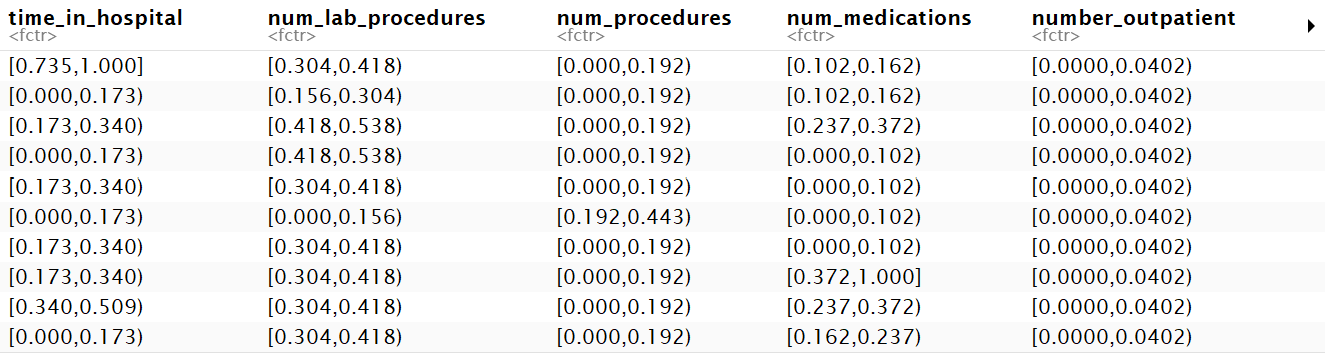
A quick visual of the actual data set shows that there is a large range of numbers that are shown in some of the columns. This can skew results by adding more weight to some of the explanatory variables. An example is the difference between the first few rows and columns of the data. A patient can send fourteen days in the hospital and have forty four lab procedures ran (first row), or a patient can spend five days and have sixty six lab procedures ran (third row).



A way to normalize the data so that the weight of the number of labs ran and the weight of the time spent in the hospital is based on the category and not skewing results across other categories is to normalize it based on each column. Each row is subtracted from the minimum value in the column then divided by the column range (maximum value of the column minus minimum value of the column). When looking at the first and third column below, the numerical values are within a zero to one range. The difference between the columns is much smaller, however, the values are conserved within each column.



Additional preprocessing is need to change number data into categorical data so that some of the algorithms can be ran and compared using both sets and in some cases, the algorithm may only except categorical data, like the Apriori algorithm used for shopping cart analysis. In order to create these categories, clusters are used. Each cluster is based on the individual column, so the clusters would just be a split in the data. In this example five clusters are applied, so the numeric data will be split into five section. Below is the normalized dataset example with the established categories.

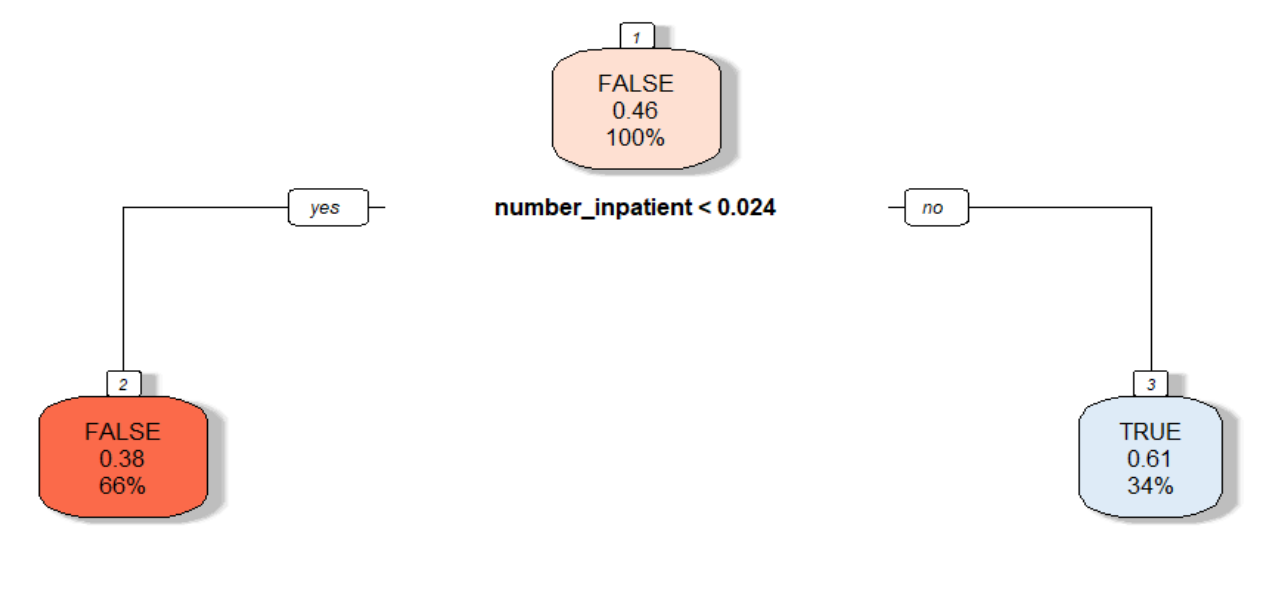


Now the data set is ready to process through a different data mining algorithms. The first algorithms that will be ran will be part of the preprocessing step. This step will use unsupervised algorithms for feature selection and data subsetting. The purpose is to find the best predictors so that they can be used in supervised algorithms to make a predictive model. Decision trees, and association rules will be used to selecting the explanatory variables. Similar to the correlation matrix, this process will help identify variables that have a higher predictive value for predicting the target variable.

A random forest algorithm was ran using all of the explanatory variables with the readmitted indicator as the target variable. Random forest is a decision tree algorithm that selects variables randomly to be at the top of the tree, then takes the average of the tree performance as the accuracy rate. The accuracy rate for this model was 62.33%, which is better than the baseline accuracy rate of 45.64%. The highest ranked predictors for readmission were the number of lab procedures the patient had, the number of medications the patient was on, the number of days the patient was in the hospital, the number of times they were admitted, the number of diagnoses, number of times the patient come in for outpatient services, gender, the number of times the patient went to the emergency room and the number of medical procedures that they had.



A traditional decision tree algorithm was ran using all of the explanatory variables. The traditional decision tree calculates the best predictor of the target variable and uses it at the top of the tree. The findings for this tree show that the number of times that a patent has been admitted is the top predictor for wither they are readmitted. The normalized value of 0.024 equates to less than 2. This means that if the patient was admitted more than one time then they are likely to be readmitted. Which means that the number of admissions may include the readmission.



An association rule algorithm Apriori was ran on all of the explanatory variables for the dataset. This algorithm looks for the most frequent associations that go with the readmission of a patient. This is commonly used to identify products that are purchased together. In the context of this project it will be used as an unsupervised predictive task in the preprocessing step to find common variables that will be used in conjunction with the previous unsupervised algorithms, to select the best explanatory variables. The top five rules showed the most common values found with readmission are: Caucasian, Females, with inpatient normalized categorical value .336 - .631, diagnoses value .483 - .643, number of procedures value 0 - .192, and not on the following medications glimepiride, repaglinide, or glucose serum.



The preprocessing has been completed. The common explanatory variables that showed promise are number of procedures, number of diagnoses, number of lab procedures, number of medications, time in the hospital, number of times to the emergency room, number of inpatient admissions and gender. The data mining algorithms that will be used are supervised models. This means that require the dataset to be split into a training and testing set. The dataset is split using eighty percent as the training set and twenty percent as the testing set. The dataset is grouped so that the training set and the testing set both have the same percentage of readmitted cases.

This phase of the Knowledge Discovery in Databases process is data mining. The algorithms that will be used are supervised predictive algorithms and consist of: Linear Support Vector Machine (SVM), K Nearest Neighbors (KNN), and Naïve Bayes. The test dataset will allow for the accuracy to be measured; this is how the models will be compared.

A linear SVM algorithm utilizes the training dataset to calculate the maximum margin of linear separation between the data that show the redamation indicator as true and the data that shows the readmission indicator as false (Tan, Steinbach, & Kumar, 2015). Then when the test data set is provided to the SVM model, the predicted values will be based on the training dataset and the algorithm that provided the maximum margin of separation. This model had an accuracy rate of 60.94%, which is higher than the 45.64% baseline.

The KNN algorithm takes the training dataset and calculates the Euclidean Distance between pairs and creates clusters where the data has the lowest distance (Tan, Steinbach, & Kumar, 2015). Similar to the SVM algorithm, the model that is created based on the training data is ran against the test data to predict wither the patient was readmitted or not. This model has an accuracy rate of 58.44%, this is higher than the baseline, but the SVM model preformed slightly better.

The last predictive task will be the Naïve Bayes classifier algorithm. Naïve Bayes posterior probability based on the explanatory variables to determine the probability of the target variable value (Tan, Steinbach, & Kumar, 2015). The probability of the patient being readmitted is formulated based on the training dataset. The calculation is based on each of the variables. The training data is used to create the model that is used to predict if a patient is going to be readmitted or not. This model preformed with an accuracy rate of 61.24%, which makes it the best preforming model.

The analysis shows that the probability of a patient being readmitted to hospital can be predicted with an accuracy rate of 61.24%. The presentation of the results would be considered the last step in the Knowledge Discovery in Databases process, postprocessing. The conclusion for this project example is that some additional data may be able to improve model. There may be social-economic factors, wither the patient adheres to the medication that they are prescribed, and wither or not the patient maintains regular doctor visits.

Project Three – Text Mining

The final project example will review the analysis of unstructured data. For this example, the Python programing language will be used for most of the text mining analysis, and SentiStrength (SentiStrength, n.d.) will be used for sentiment analysis. The hypothesis for the project is to determine if conservative or liberal bias in news articles can be determined using text mining algorithms and methods. The dataset was obtained from Kaggle, a data science website that allows for datasets and competitions to be posted. The dataset for this project includes articles published between June and August 2019, and between October 2019 and January 2020. The new sources used are: ABC News, Al Jazeera English, BBC News, Breitbart News, CBS News, CNN, Daily Mail, Fox News, Huffington Post, Independent, Metro, Mirror, NBC News, Reuters, The Hill, The New York Times, and The Verge.

The text data must first be transformed from unstructured data to structured data in order to be analyzed. The dataset comes in a csv file with the source, author, title, description, url the article was retrieved from, requested date, published date and in some cases the content. The source shows the news agency that published the article, this can be used as a label. The description column looks like a summary of the article. This field has data in all of the rows and provides more information than the headline alone.

With the objective of the project to see if conservative or liberal bias can be identified based on the news article, and in this case a summary of the news article, the dataset needs to include a label that identifies if the article is considered conservative or liberal. This can be done using Amazon Mechanical Turks. A random same of article descriptions was taken from the dataset and assigned to five different workers per article to find the consciences for where each news agency leans. This includes an option for neutral. Below are the results:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| News Agency | Liberal | Neutral | Conservative | Results |
| ABC News | 29 | 49 | 27 | Neutral |
| Al Jazeera English | 33 | 42 | 25 | Neutral |
| BBC News | 27 | 38 | 35 | Neutral |
| Breitbart News | 37 | 29 | 34 | Liberal |
| CBS News | 44 | 27 | 29 | Liberal |
| CNN | 31 | 37 | 37 | Conservative/Neutral |
| Daily Mail | 25 | 42 | 33 | Neutral |
| Fox News | 32 | 35 | 28 | Neutral |
| Huffington Post | 19 | 6 | 20 | Conservative |
| Independent | 35 | 31 | 29 | Liberal |
| Metro | 8 | 25 | 12 | Neutral |
| Mirror | 21 | 59 | 20 | Neutral |
| NBC News | 35 | 44 | 21 | Neutral |
| Reuters | 22 | 47 | 31 | Neutral |
| The Hill | 34 | 41 | 25 | Neutral |
| The New York Times | 24 | 17 | 19 | Liberal |
| The Verge | 19 | 47 | 34 | Neutral |
|  |  |  |  |  |
| Totals | 475 | 616 | 459 |  |

There were several news agencies where the results are unexpected. Specifically, Breitbart News showing as Liberal and CNN leaning on the Conservative side. The average Kappa was calculated for each of the News Agencies to determine how well the workers agreed with each other on the label. “Kappa measures the percentage of data values in the main diagonal of the table and then adjusts these values for the amount of agreement that could be expected due to chance alone.” (Simon, n.d.). Below are the average Kappa percentages for each of the News Agencies:

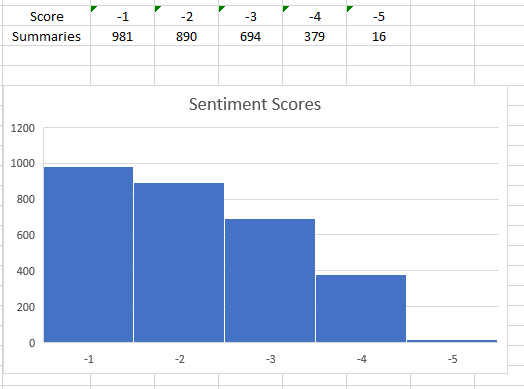
|  |  |
| --- | --- |
| **News Agency** | **Average of Kappa** |
| ABC News | 47% |
| Al Jazeera English | 33% |
| BBC News | 27% |
| Breitbart News | 0% |
| CBS News | 27% |
| CNN | 30% |
| Daily Mail | 0% |
| Fox News | 0% |
| Huffington Post | 42% |
| Independent | 33% |
| Metro | 18% |
| Metro US | 27% |
| Mirror | 21% |
| NBC News | 35% |
| Reuters | 47% |
| The Hill | 41% |
| The Huffington Post | 34% |
| The New York Times | 40% |
| The Verge | 47% |
| **Total Average:** | **28%** |

Based on the Kappa scores, there wasn’t any indication that the labels can be used. The Kappa score should be at least 60% or more to indicate that there is Good or Very good agreement (Simon, n.d.). The highest agreement percentage was 47%, for The Verge, Reuters, and ABC News. The only label that will be available for analysis is the news agency name. Additional analysis can still take place on the data set to see if there are similarities between news agencies during the news dataset timeframe.

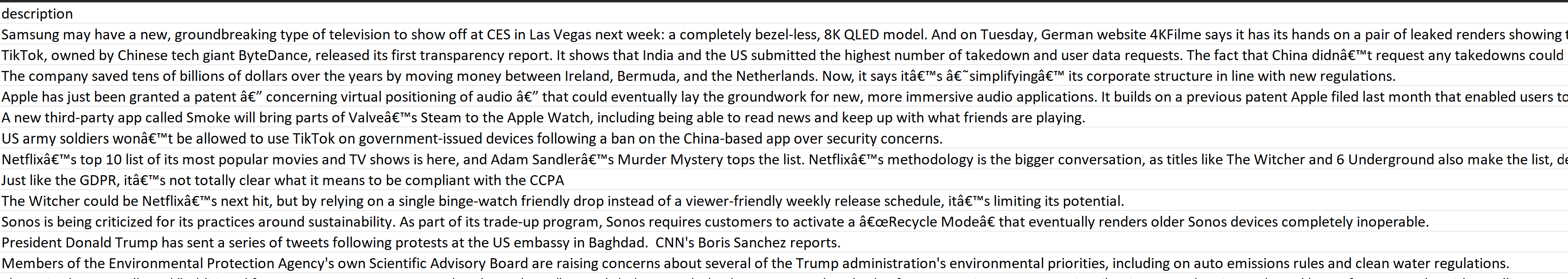
Amazon Mechanical Turks allows additional parameters to be set to help rule out bias. One parameter that may have helped in this task would have been selecting only those registered as a Democrat or as a Republican to complete the tasks and comparing the results. This would allow for personal bias to be identified. This is an area where the study can be improved on in retrospect.

Sentiment analysis is completed on the description field to identify if there are any news agencies that are more negative or positive during the news timeframe. Below are the results for each news agency based on the number of articles for each sentiment.

The results indicate that all the news agencies report negative stories. The range used in the results are 1 as not positive to 5 as extremely positive, and -1 not negative to -5 extremely negative (SentiStrength, n.d.). Most of the articles for each news agency are in the -1 to -2 range, so they may be closer to neutral.

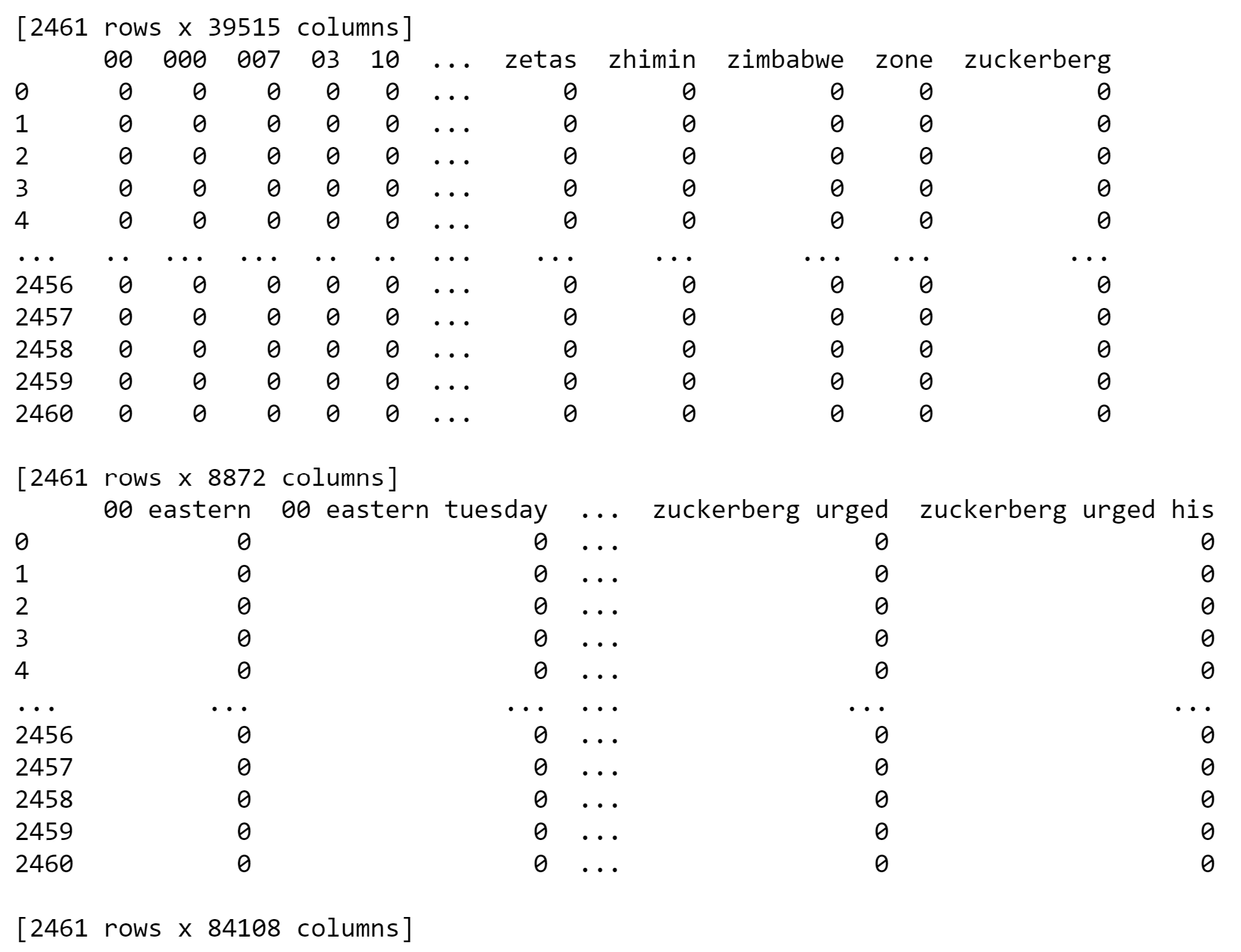
**

The text mining techniques using Python that will be applied to the dataset are: Word Clouds, and K-Means Clusters. In order to run these methods, some data cleaning and vectorization will need to take place. The description column is a string of data that has irregular characters. Both the title and content column have the same issue as well. These characters will be removed using regular expressions in Python. Below is an example of the original text.

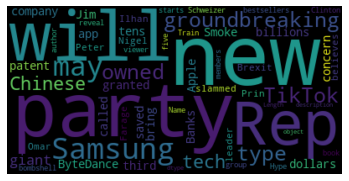


The text in the all three fields will also be tokenized. Tokenization is the process of separating the string of text into individual words or characters using the white space between the characters in the document or string (Weiss, Indurkhya, & Zhang, 2010). The tokenization will be competed using a unigram and bigram. Unigrams are tokenized with each individual word, while bigrams are the combination of two words as they appear next to each other in the text.

The dataset is then vectorized so that it is transformed from a list of words into the count of each word in the document. This is done for the description, title and content columns with the label set as the news source. Below is an example of the transformed dataset for both unigrams, bigrams and trigrams.

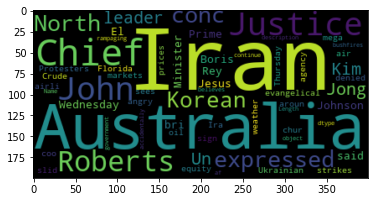


The results from the Word Cloud show the most frequently used words by making the larger in font in the graphic. Below is a Word Cloud that shows the frequent words from all of the text in the description column.

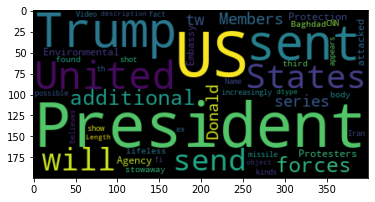


Looking at the Word Cloud for Reuters, CNN and Fox News, then comparing them back to the full Word Cloud, there are stories that are not being reported across all news agencies. There is a difference in what is being reported.

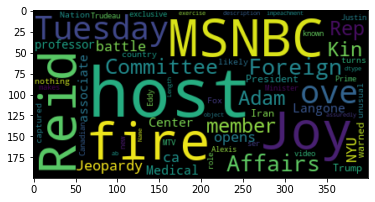
Reuters – Word Cloud



CNN – Word Cloud

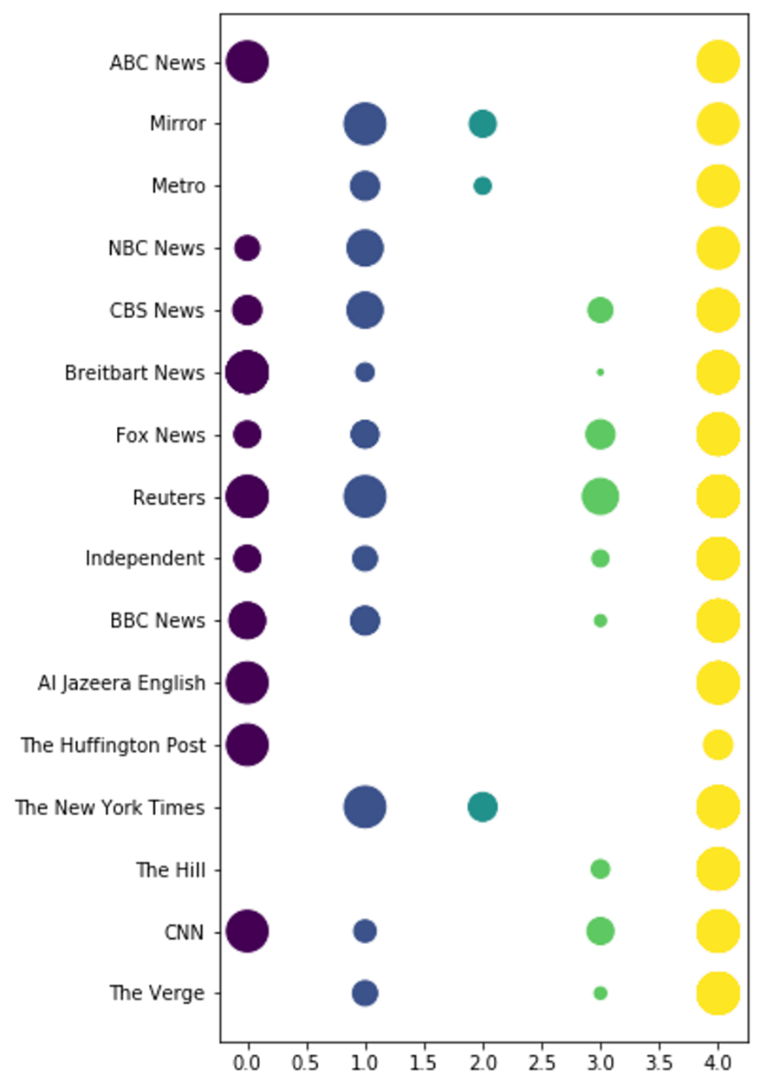


Fox News – Word Cloud



Lastly a K-means cluster analysis will be performed on the dataset to see if the differences in what is reported show news agencies that are more closely aligned, this may indicate that a group of news agencies have the same bias. K-means is an unsupervised machine learning algorithm that creates clusters based on the number of clusters that are chosen, the mean is the cluster centroid (Weiss, Indurkhya, & Zhang, 2010).

The elbow method ran on the content bigram and trigram data to find the optimal k value of 5. The clusters show that there is a base cluster that spans across all the news agencies. This may indicate that the news sources had the same core information. There were some groups that formed but most of the news sources were in multiple clusters. However, cluster 2 includes only three news sources, Mirror, Metro, and The New York Times. These three are all newspapers. The Mirror and Metro newspapers are both based in the United Kingdom, while The New York Times is in the United States.



The clusters indicate that the primary type of delivery may need to be considered when comparing the sources. The expectation was that there would be distinct clusters, especially given that some of the news sources are not United States based. In retrospect, comparing sentiment based on matching stories may be a better way to find bias. If the same story is positive at one news source, but negative when provided by another source, that might indicate bias. Further analysis would need to be completed to know for sure.

Data Science is a large field that requires skill sets that span across collection, storage, and analysis of data. The output of this work is the presentation of the information gleaned from the data. Understanding the methods used to obtain this valuable information is crucial for a Data Scientist. The Data Scientist must know how the different algorithms work in order to apply them and present the findings. The projects reviewed cover how to store data in a data warehouse utilizing SQL Server, Integration, and Analysis Services; and how the analyze structured and non-structured data sets using R and Python programing languages.

Bibliography

Christensson, P. (2017, August 17). *Data Science Definition*. Retrieved 2020, Mar 17, from <https://techterms.com>

Kimball DW/BI Lifecycle Methodology. (n.d.). Retrieved April 17, 2020, from <https://www.kimballgroup.com/data-warehouse-business-intelligence-resources/kimball-techniques/dw-bi-lifecycle-method/>

Kimball, R., Ross, M., Thornthwaite, W., Mundy, J., & Becker, B. (2008). The Data Warehouse Lifecycle Toolkit: Practical Techniques for Building Data Warehouse and Business Intelligence Systems. Indianapolis: Wiley.

SentiStrength. (n.d.). Retrieved April 25, 2020, from http://sentistrength.wlv.ac.uk/

Simon, S. (n.d.). StATS: What is a Kappa coefficient? (Cohen's Kappa). Retrieved April 25, 2020, from http://www.pmean.com/definitions/kappa.htm

Tan, P.-N., Steinbach, M., & Kumar, V. (2015). Introduction to Data Mining. Dorling Kindersley: Pearson.

Weiss, S. M., Indurkhya, N., & Zhang, T. (2010). Fundamentals of Predictive Text Mining. London: Springer London.