***Patients at High Risk of Being Readmitted to the Hospital Within 30 Days from Discharge***

**Group 3**

**Raghavendra Dasari, Han Lee, Roselle Paala, and William Buchanan**

**USF**

**ISM6136 – Data Mining**

**Prof. Daniel Conway**

**August 3, 2019**

Table of Contents

[Introduction 1](#_heading=h.30j0zll)

[Data cleaning 1](#_heading=h.1fob9te)

[Method I: Clustering 2](#_heading=h.3znysh7)

[Clustering Preprocessing 2](#_heading=h.2et92p0)

[Determining the Best Number of Clusters 2](#_heading=h.tyjcwt)

[Results of the Automatic Clusters 2](#_heading=h.3dy6vkm)

[Interpretation of the Cluster Results 4](#_heading=h.1t3h5sf)

[Method II: Comparison of four types of classification models 4](#_heading=h.4d34og8)

[Method III: Regression Model 5](#_heading=h.2s8eyo1)

[Regression Model 1 - Excel Data Analysis Toolpak 6](#_heading=h.17dp8vu)

[Regression Model 2 - Excel Data Analysis Toolpak 6](#_heading=h.3rdcrjn)

[Prediction 6](#_heading=h.26in1rg)

[Result 7](#_heading=h.lnxbz9)

[Conclusion 7](#_heading=h.35nkun2)

[References 8](#_heading=h.1ksv4uv)

# 

# 

# 

# 

# 

# Introduction

Hospitals and healthcare providers face many challenges in today’s economy and political climate. Lowering costs, providing better quality care, and staying in compliance with both state and federal regulations are at the heart of many of these challenges. Reducing the rate of readmittance of previous patients can be a huge step towards meeting the challenges faced by hospitals and healthcare providers today. By examining patient data, administrators can identify those patients at risk of returning to the hospital and being readmitted within 30 days, which is a crucial period for both the healthcare provider and the patient.

There have been several policy initiatives that have been put in place to encourage hospitals to address readmissions after discharge. The Patient Protection and Affordable Care Act of 2010 contains multiple payment reforms intended to promote hospital efforts to address and prevent readmissions within thirty days after discharge. One significant payment reform is the Hospital Readmissions Reduction Program (HRRP), which financially penalizes hospitals with above-average readmission rates for target illnesses. As nearly 20% of Medicare patients must be readmitted within 30 days of discharge, minimizing post-discharge readmittance has become a priority for the US healthcare system (Agency for Healthcare Research and Quality, 2019).

Developing predictive models on the likelihood of rehospitalization within 30 days of initial discharge can help hospital administrators to target the patients most at risk and implement care programs to reduce the possibility of readmittance. These programs will be used to attain the quality goal of providing better medical service to patients while also reducing the likelihood of the financial penalty and significant cost increases of care (American Diabetes Association, 2019).

To build these models, a dataset was used that consists of data that was collected from clinical care at 130 US hospitals between the years of 1999 and 2008 (Strack, et al., 2014). The data includes many different attributes of significance to patient care and disposition as well as demographics. These include race, gender, age range, admission type, time in hospital, number of lab tests performed, and diagnosis as well as the number of outpatient, inpatient, and emergency room visits in the year prior to the hospitalization. There is a patient number which is a unique identifier for each patient. The age attribute categorizes each patient into a 10-year range which includes the patient’s actual age such as 20-30 and 30-40. The dataset also has many diabetes medications for the patient ranging from Metformin to Insulin.

# Data cleaning

The original dataset included 50 columns which included a readmitted column indicating if the patient was readmitted to the hospital within 30 days and whether it was within 30 days or over 30 days. The objective of the predictive model is to identify which combination of attributes determine whether the patient is likely to be readmitted within 30 days or not. The following columns were identified to be the determinants: age, admission\_source\_id, time\_in\_hospital, num\_procedures, and diabetesmed. The column for patient weight has been eliminated due to having excessive null values and therefore no statistical value. The medication columns were eliminated because there are many alternatives for diabetes medications. However, the diabetesmed column was included as the fact that a patient is taking at least one medication has been determined to be of significance to readmittance.

However, the values for age were of intervals of 10 years such as [10-20). This value represented the age interval of a patient from 10 to 20 years old. In order to run a regression model, we decided to quantify this value into median age of 15. We created another column called “Med\_Age” to represent the median age that we will use to build our prediction model.

The admission\_source\_id is mapped to Emergency, Urgent, Elective, Newborn and null values. We decide to convert the admission source id values into the Level of Emergency\_Hospitalization by assigning 4 to Emergency, 3 to Urgent, 2 to Elective, 1 to Newborn and named the new column as “Level of Emergency\_Hospitalization” and deleted rows with null values. The column diabetesmed had values of true and false, true meaning the patient taking at least 1 medication to treat diabetes and false otherwise. We assigned 1 to false and 0 to true. For the column readmitted\_Quantified, we assigned 1 to “<30” (admitted within 30 days) and 0 to all the other values.

After the process of quantification of string data types and data cleaning to delete certain rows with null values, building the prediction model is started.

# Method I: Clustering

## Clustering Preprocessing

Clustering is implemented to identify the groups at risk of readmission within 30 days from initial discharge. The variables that are used in this model are the number of procedures, number of lab procedures, number of inpatient visits, number of diagnoses, number of emergency visits, and time in the hospital. The following sections will discuss the steps taken to prepare the data, train the data, and evaluate the results.

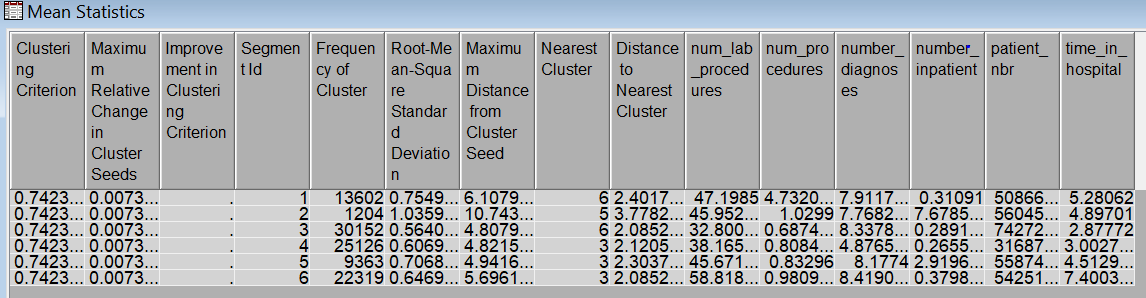
Prior to training the data, variables are evaluated. When it comes to clustering a dataset, it is important that the variables are in interval type or in numeric range. In addition, the values of each involved variables should be symmetrically distributed. This means that if there are missing values or values that are zero, these values would have to be fixed. The techniques that are used to clean the missing values and zeros are replacement and impute. After implementing these techniques, 126,892 values are left for training.

## Determining the Best Number of Clusters

Automatic Clusters and User-Defined Clusters are implemented in the dataset. To determine the best number of clusters to use in the dataset, the results of each cluster are compared. The Automatic Clusters created two clusters while the User-Defined Clusters has six clusters. The User-Defined Cluster has more defined clusters, showing many variables to interpret. In contrast, the Automatic Clusters does not provide enough information to interpret given the limited amount of segment the data is divided into. Therefore, User-Defined/K-means Cluster is used to model the dataset.

## Results of the Automatic Clusters

The data is split into six clusters, with segments two and five having the least amount of observations or frequency of clusters. The Mean Statistics and the Segment Plot results have the most relevant information to gather in this model. The columns to pay close attention in the mean statistics result are the variables selected for the clusters such as the frequency of clusters, number of lab procedures, number of procedures, number of diagnoses, number of inpatient visits, and time in hospital. The results can be interpreted by comparing the value of the number of lab procedures, number of diagnoses, number of procedures, number of inpatient visits, and time in the hospital based on their respective segments.

****

Another result to pay attention to is the Segment Plot. In the image below, there are six segments shown. The first diagram shows how the number of lab procedures are distributed. Number of lab procedures is distributed into six clusters or six segments; and each section of the bar graph illustrates the weight of a variable in the segment. For instance, the number of lab procedures shows that 37% of cluster had about 34 to 50 lab procedures done, 26% had 50 to 66 procedures done, and 15% had 17 to 34 procedures done.

****

Segment Profile Node is added in the model for a closer look of the results. Segment Profile is especially useful for models with more than three variables. The diagram above shows many histograms. Each graph compares the observation’s distribution of results with the original distribution. The number of lab procedures and time in hospital in segment 3 illustrate that out of 30,152 observation, the number of procedures, is relatively low while the number of lab procedures and number of diagnoses are mid-level. The time in hospital in segment 4 shows that out of 25,126, the time in hospital and number of procedures are relatively low. Another diagram in segment 1 illustrates that the number of procedures in observation group of 13,602 is high.

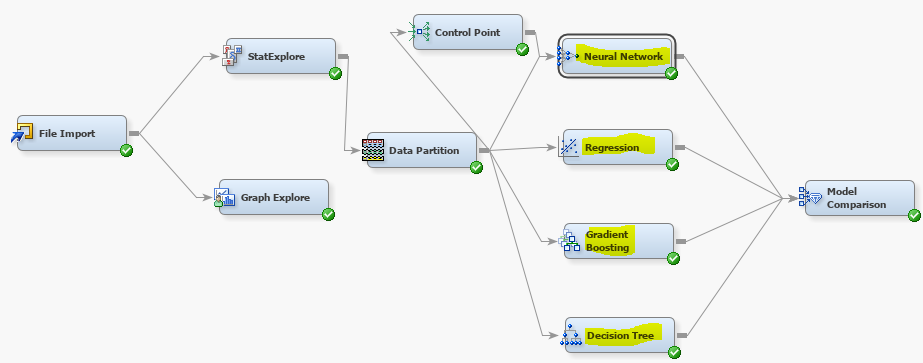
## Interpretation of the Cluster Results

It can be inferred from the results that the number of diagnoses, the number of lab procedures, and time in the hospital are one of the main variables that determine the level of risk of being readmitted. Based on the result, segment 6 has the highest risk of being admitted followed by segment 1. Segment 4 has the least risk of admittance. The remaining segments are grouped into the least risk of admittance along with segment 4.

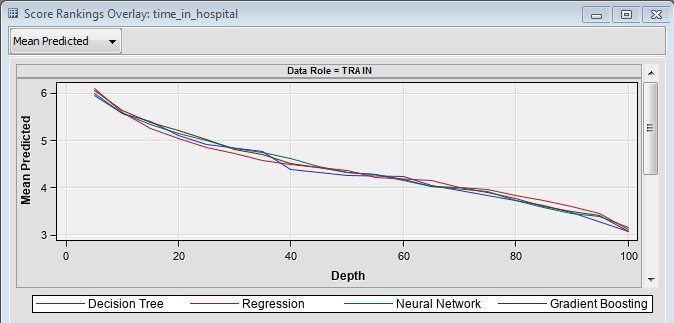
# Method II: Comparison of four types of classification models

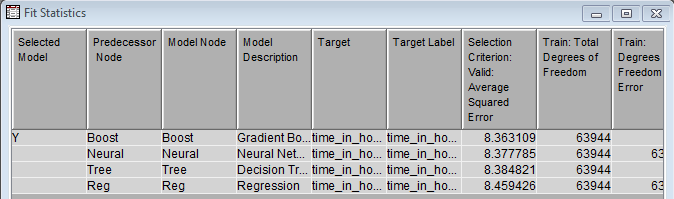
In the second method, four types of classification models have been compared to determine the best prediction of readmission. Ample time is spent on data cleaning before using the dataset for prediction. After modeling the dataset, six fields appear to have very high impact on the prediction. The target column in these prediction is readmitted\_quantified, which indicates the values of patient admitted into hospital less or more than 30 days window.

Below is a SAS Enterprise Miner diagram with four different classification models and a comparison among the four models, identifying which model provided the best prediction.

****

StatExplore and GraphExplore nodes are used to see the data representation in graphical format. Then, Data Partition node is used to divide the data into 70% training set and 30% validation set. Control Point node is used to connect various models to the dataset. Below is a comparison graph with four different prediction models, showing that pretty much all four models predicted equally, except for the regression model. The regression model did slightly better than the other three models.

****

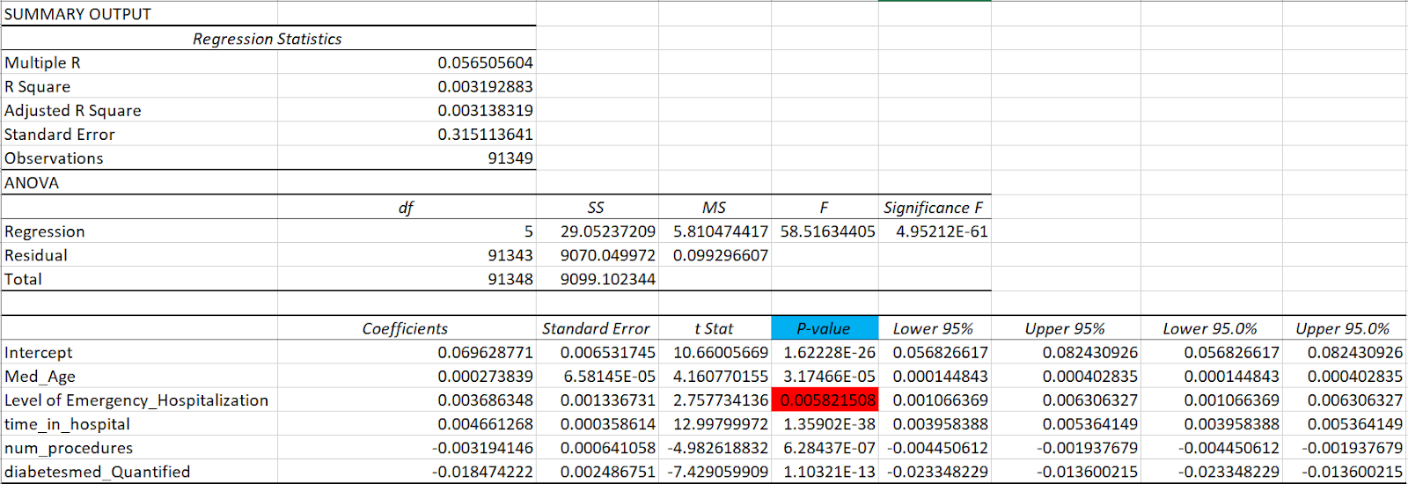
****

In the above table, the average squared error looks almost similar for all models except for the regression model. Again, the regression model performed slightly better than the other three models.

# Method III: Regression Model

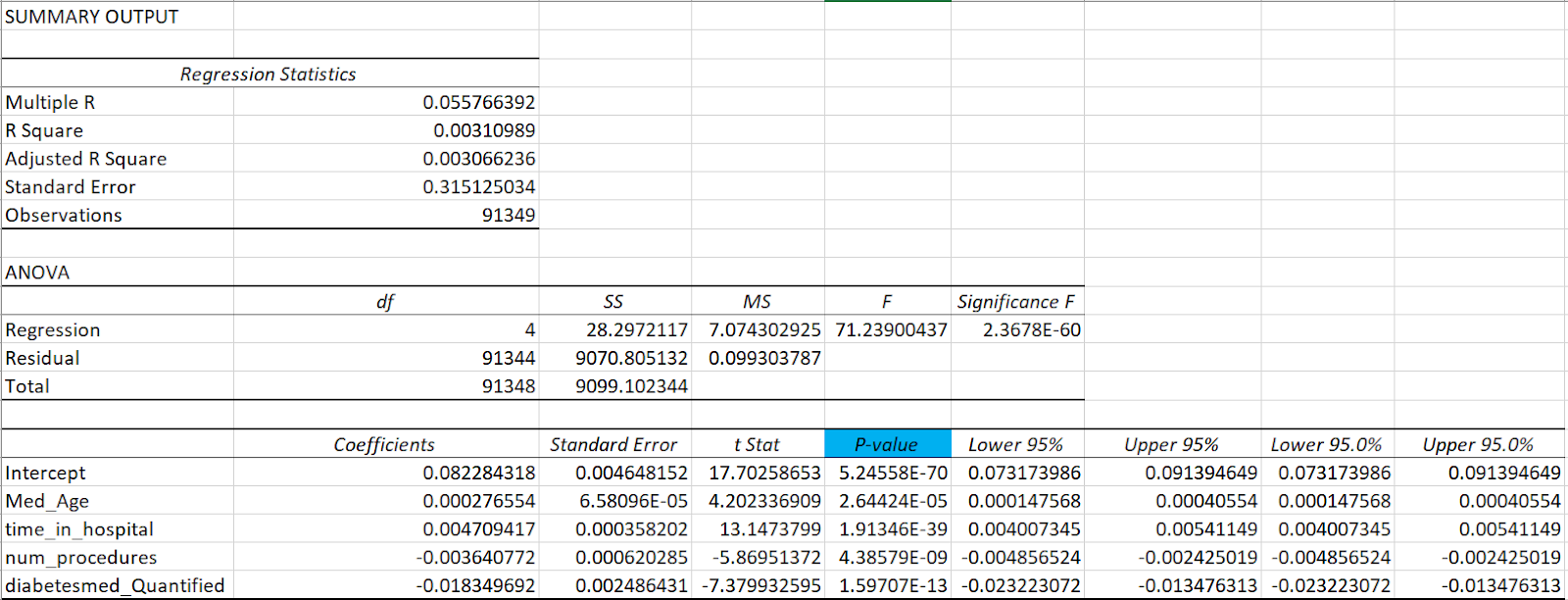
After the quantification of the string data types, the quantified independent variable attributes Med\_Age, Level of Emergency\_Hospitalization, time\_in\_hospital, num\_procedures, diabetesmed\_Quantified were taken into consideration to predict the dependent variable column readmitted\_Quantified. The Excel Data Analysis Toolpak was used to build the regression model.

## Regression Model 1 - Excel Data Analysis Toolpak

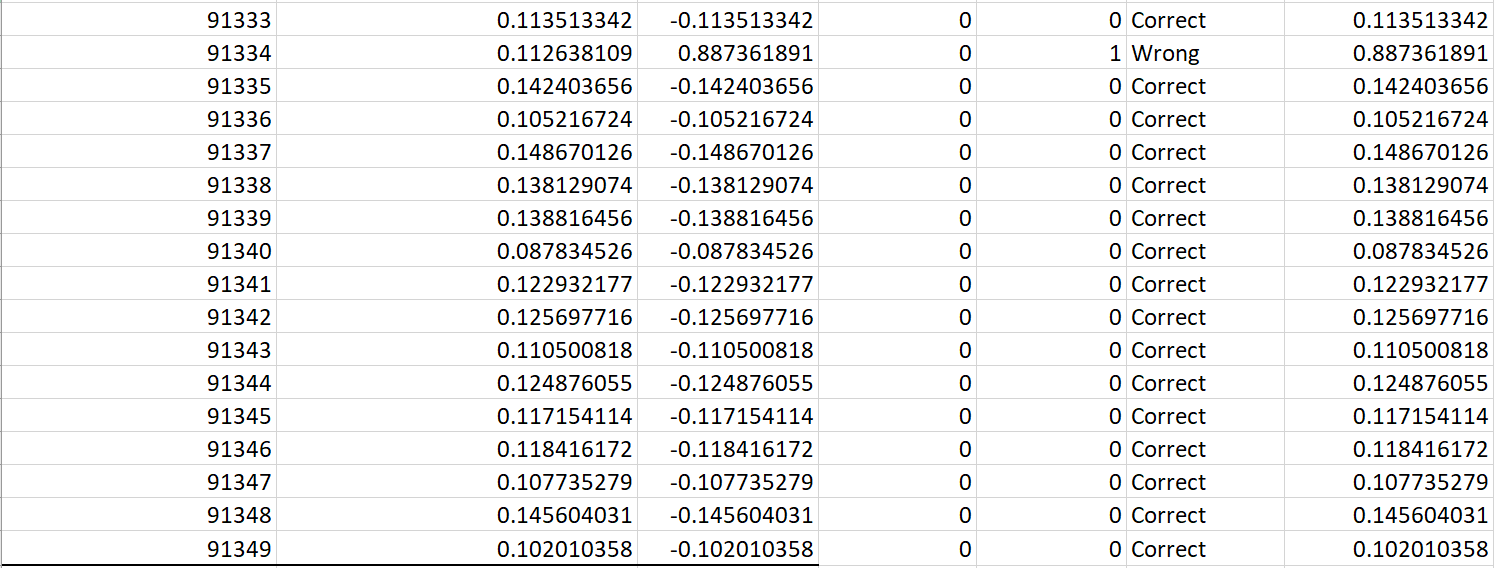
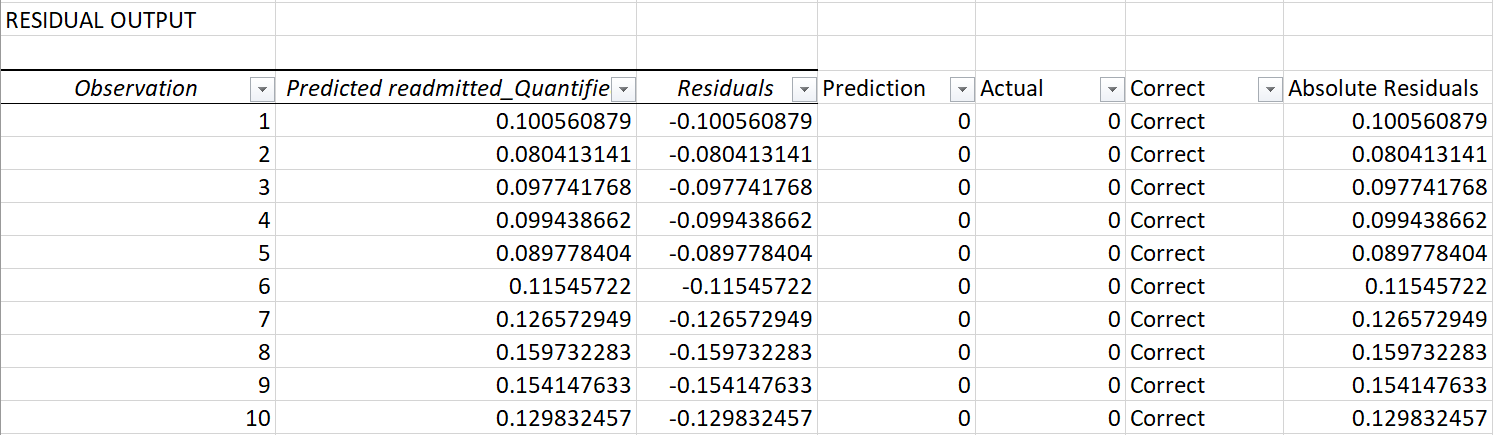


Then, the P-value of each of the independent variable columns is analyzed. The P-value for Level of Emergency\_Hospitalization was greater than 0.005 as P-values for other columns were close to 0. The Level of Emergency\_Hospitalization column is taken off the regression model and rebuilt in our model.

## Regression Model 2 - Excel Data Analysis Toolpak

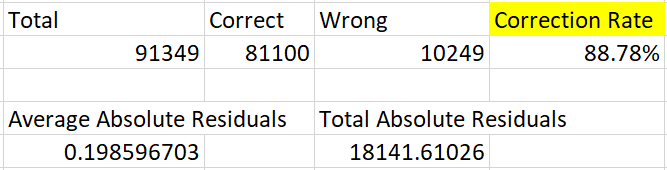


## Prediction

The P-values of each independent variable columns were close to 0. The residual output displayed the observation number and predicted readmitted value (1 for Readmission within 30 days of initial discharge, 0 for otherwise) for each observation of a patient. The calculated value of predicted readmitted value is rounded to the nearest integer to get either 0 or 1 to evaluate the prediction. 

## Result

The prediction of 81,100 patients’ likelihood of rehospitalization within 30 days of initial discharge were correct, 10,249, otherwise. The correction rate for prediction was 88.78%.



# Conclusion

In conclusion, our predictive and clustering models have shown it is possible to identify, with a strong degree of accuracy, whether a patient is at risk of being readmitted to the hospital with-in a thirty-day period after being discharged. This can be accomplished using various attributes such as age, the admission reason, a patient’s time in the hospital, the number of procedures done, and whether they are taking any medication such as diabetes medication. Hospitals and health care providers can use these attributes to target at-risk patients to make sure they receive the proper care to prevent readmission within the 30-day time frame. By doing so it will help them to reduce costs, provide better care, and to avoid any financial penalties.

# References

Agency for Healthcare Research and Quality. (2019, January). *Readmissions and Adverse Events After Discharge*. Retrieved from Patient Safety Network: https://psnet.ahrq.gov/primers/primer/11/Readmissions-and-Adverse-Events-After-Discharge

American Diabetes Association. (2019, January 30). *The Cost of Diabetes*. Retrieved from Diabetes.org: http://www.diabetes.org/advocacy/news-events/cost-of-diabetes.html

Strack, B., DeShazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., & Clore, J. N. (2014). Diabetes 130-US hospitals for years 1999-2008 Data Set. *BioMed Research International, 2014*(Article ID 781670), 11. Retrieved from UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008