The relation between the overall human health and the chances of a stroke

Abstract

Understanding the data that revolves in the healthcare industry is vital to progressing every domain of medicine. From basic diseases to the most dangerous diseases, analyzing data plays an important role in solving variant issues and improving healthcare for all ages and genders. It gives healthcare workers a better view on how to tackle and reduce the ongoing problems in societal health. One of the most common and most dangerous complications that cripple a human being is a stroke. A stroke is a sudden interruption in the blood supply to the brain that damages some parts of the brain. The effects of a stroke, depending on the damage to the brain, can go from a simple momentary disability to a severe disability lasting a lifetime. In this report, I have done different types of analysis, specifically on a dataset representing a Stroke Prediction,

1. Introduction

Since the beginning of time, the study of medicine has become an important study as it was further benefitting and improving the lives of all human beings. The study of strokes has cost billions of dollars to help at best avoid such an attack or a minimum survive this attack with the least damages to the brain. Most people cannot see or feel a stroke when it happens which makes matters worse, but one way to avoid such an incident would be to understand the underlying issues related to it. This report is an in-depth analysis of a Stroke Prediction Datasets. The following report gives a study on the relation between the age groups that have a positive result for a stroke. We will view the effects of the body mass index and the effects of the average glucose level on the chances of having a stroke. Adding to all of this, we will also give a view on the relation between smoking and the possibility of having a stroke. We will end the report by analyzing the gender gap in males and females having strokes.

2. Data and Methods

The study follows a dataset taken from Kaggle representing a "Stroke Prediction Dataset". The dataset has some challenges to configure and requires some cleaning to which we will explain every step. The dataset has 5110 rows and 12 attributes, these attributes consist of the 9 categorical variables and 3 quantitative variables. The categorical variables in the dataset are: id, gender, hypertension, heart_disease, ever_married, work_type, Residency_type, smoking_status, and stroke. The quantitative variables in the dataset are: age, avg_glucose_level, and bmi. We will primarily work with 6 of the 12 attributes for our analysis. The attributes we will use in this analysis are: gender, smoking_status, stroke, age, avg_glucose_level, and bmi. The gender attribute is a categorical type representing male and females in the study, it is identified as symmetric binary type with both outcomes equally important. The smoking_status is a categorical variable with 4 categories, formerly smoked, never smoked, smokes and Unknown. The stroke attribute is also a categorical variable to which it is made up of binary numbers, 0 for negative and 1 for positive for a stroke. The avg_glucose_level is a numerical type ranging from 55.12 to 271.74. Age is a numeric type in our dataset ranging from 0.08 to 82.00. Our last attribute, bmi, is a numerical type with NA values included.

In this dataset, I was able to configure two attributes who had missing values and invalid values. The attribute smoking_status had a category for which it was Unknown and so I decided to further analyze it with its 1544 variables and change the value Unknown to NA values. And the second attribute with such problems was bmi, it was at first a categorical variable to which I first had to change into a numeric variable and further analyze the NA values which included 201 values. I was able to determine that the missing values in both cases were random and not systematic.

To deal with the missing values or NA values I decided to replace the NA values in the BMI attribute with the mean which gave me the following graph, where you can see there was not much change compared to the first one.

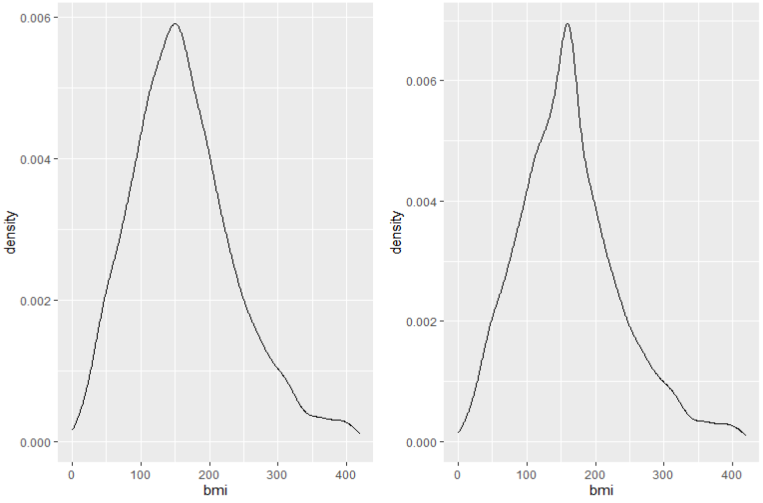

Figure 1: Representation of replacement of NA values in BMI

To deal with the missing values or NA values in the smoking_status attribute I decided to replace the missing values with the mode from the original dataset. In the following graph you will be able to see the difference it makes.

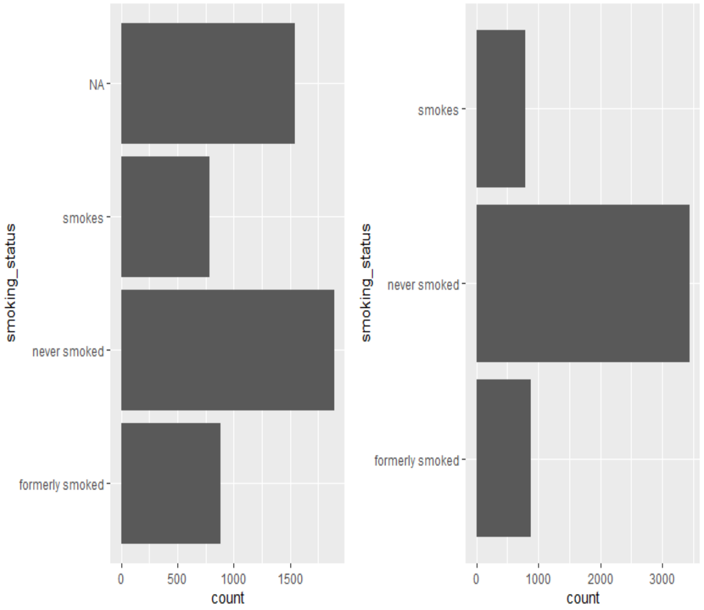

Figure 2: Representation of replacement of NA values in smoking_status

In our following analysis we will try and find the outliers that may potentially be an obstacle for our research. We will first start by using the normal distribution-based approach using the z-score method. In the case of age and bmi attributes the normal distribution methods failed as the two attributes do not have a skewed graph, but in the case of avg_glusose_levels we were able to identify some outliers based on the z-score method, but the chances that the method considered valid values as outliers is high.

The following graphs represent histogram that help us visually see some outliers as the methods for calculating outliers is flawed in our case.

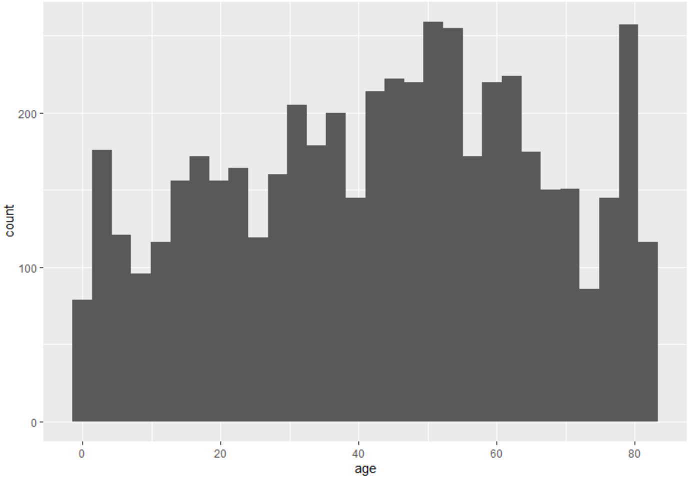

Figure 3: Age Distribution

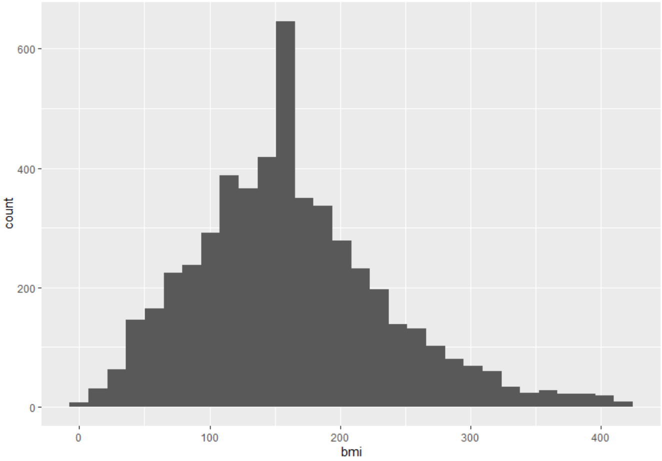

Figure 4: BMI Distribution

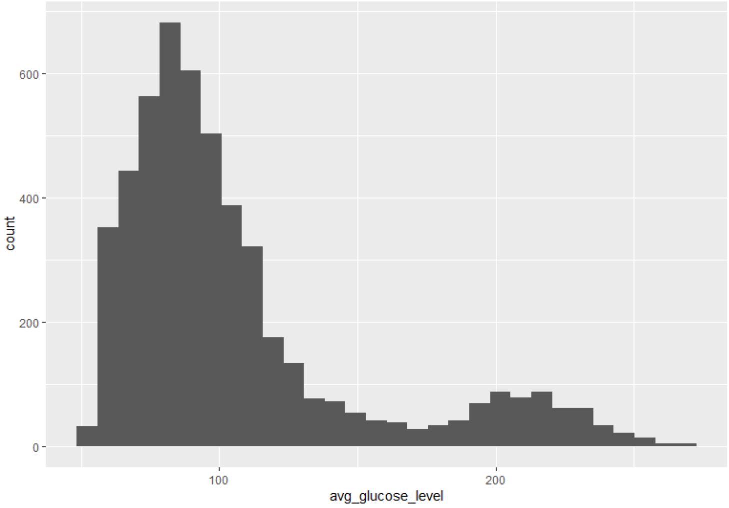

Figure 5: Average Glucose Level Distribution

In our three graphs, the graph that visually presented outlier would be the graph of the avg_glucose_level, as it was positively skewed the z-score method would potentially find these outliers. As for the age and bmi, it would be very difficult to present these outliers using the z-score method as they are approximately normally distributed. By calculating the skewness, I was able to determine how distributed each of the three variables were, age had a skewness of -0.23 (approximately normally distributed), bmi had a skewness of 0.17 (approximately normally distributed), and avg_glucose_level had a skewness of 0.94 (positively skewed). With the following information we can assume that the normal distribution method to identify outliers would only function with avg_glucose_level. With further investigation, values near to 250 avg_glucose_level would be classified as outliers.

Following more analysis, I have decided to rescale all three of my numerical attributes so that they may have an equal weight. Using min-max normalization we can view the following in which the variables were put into equal scaling points without changing the distribution.

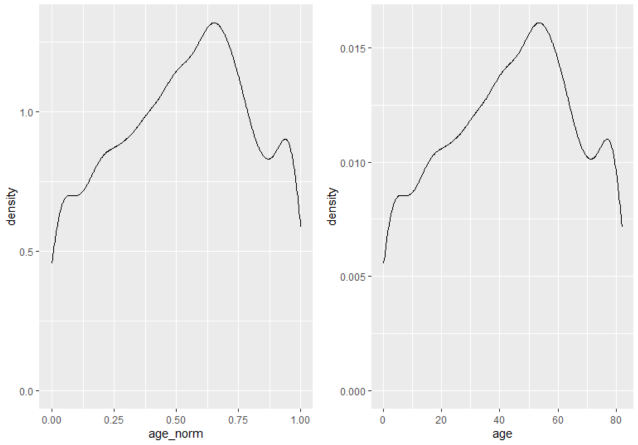

Figure 6: Age Min-Max Normalization Representation

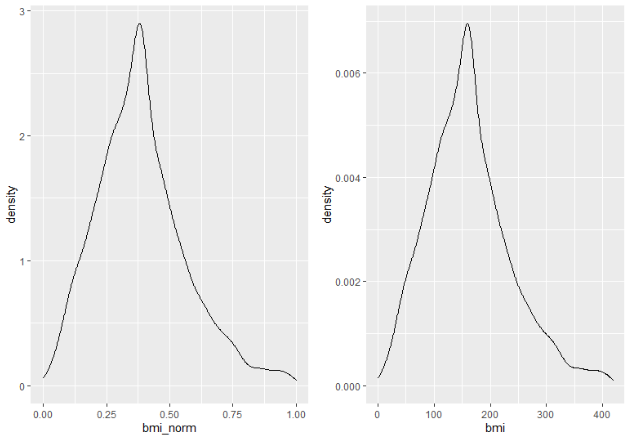

Figure 7: BMI Min-Max Normalization Representation

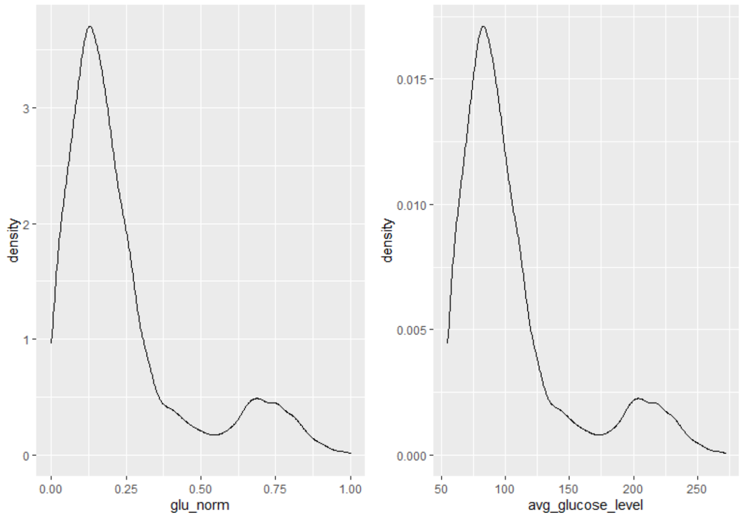

Figure 8: Average Glucose Level Min-Max Normalization Representation

In this part we will specifically work with the avg_glucose_level attribute as it is the only one that is perfectly skewed for our next method of analyzing the variable. We are going to observe the three transformations for normality (square root, inverse square root, and natural log) and decide which one of the three has the most symmetric distribution.

In our case, the most symmetric distribution would be the natural log transformation. Although, it would be very hard to decern its symmetry from that of the inverse square root.

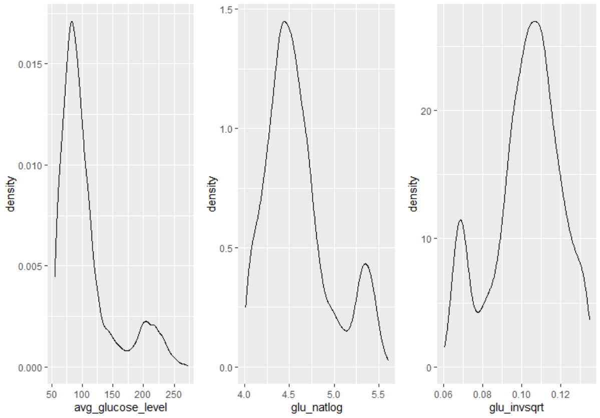

Figure 9: Average Glucose Levels Representation of Inverse Square Root and Natural Log

In this part of our work, we will see use equal binning to convert the numerical attribute, which is bmi, into a categorical variable with 3 distinct categories. Depending on the equal binning method, I specifically introduced the 3 distinct categories in the bmi attribute, these attributes are Low, Mid, and High.

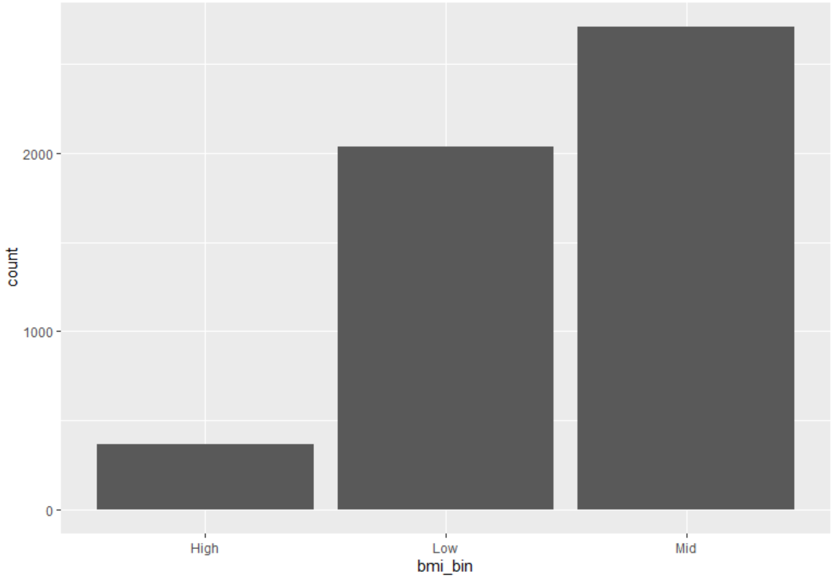

Figure 10: BMI Equal Width Binning in Three Categories

3. Results

To understand the relation between the age and the possibility of a stroke we will visualize the following on a graph.

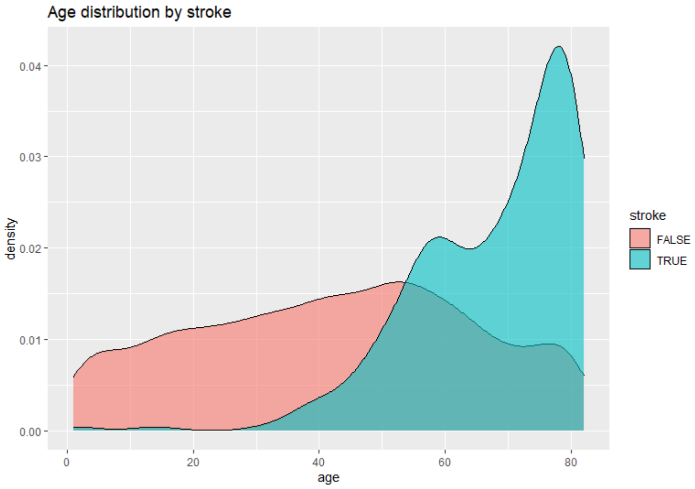

Figure 11: Age Distribution by Stroke

In figure 11, we can visibly determine that strokes are likely to occur within a population of age 40 and above. This can be affiliated with the human health that usually deteriorates after the age of 50. The graph also shows that there are some people above the age of 40 that do not have a positive sign for a stroke, this can be related to the lifestyle choices of these people.

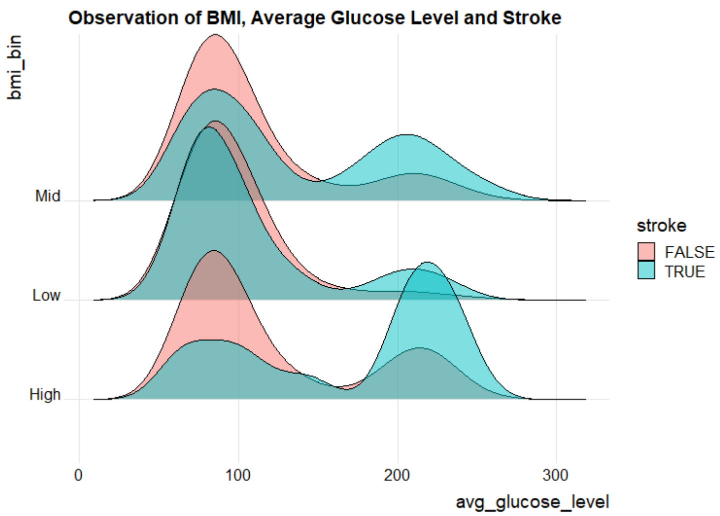

Figure 12: Observation of BMI, Average Glucose Level, and Stroke

In figure 12, we can observe the relation between the body mass index, the average glucose levels, and the chances of the person having a stroke. The graph can be interpreted in different ways, for people who have a low level bmi and a low average glucose level it can be said that chances of a stroke are both positive and negative. This result may be due to some outliers in the avg_glucose_level variable. For a mid-level bmi and a high-level average glucose, it can be said that we have numerous cases of strokes. For our last interpretation, a high-level bmi and a high-level average glucose it can be stated that people tend to have strokes.

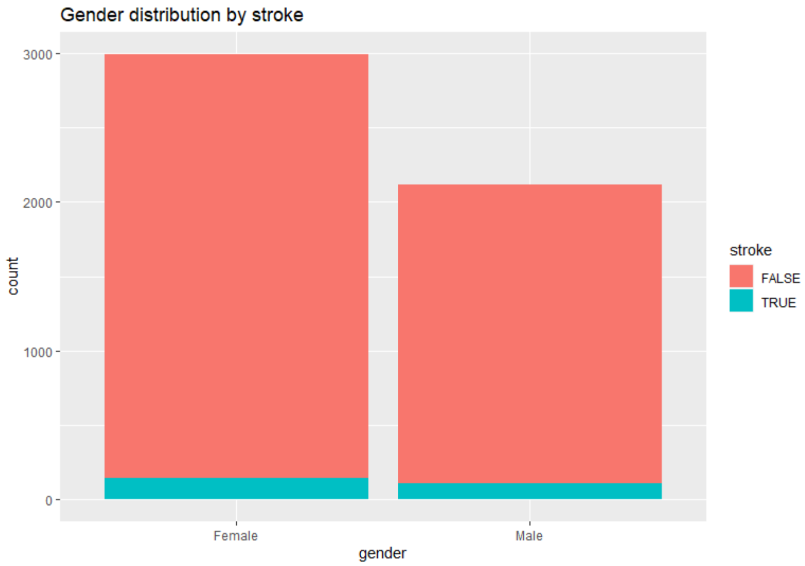

Figure 13: Gender Distribution by Stroke

In figure 13, we can notice that there is a difference in the number of females and males in our dataset, it can be estimated that there are approximately 3000 females compared to approximately 2100 males. We can easily detect that gender does not play an important role in whether a stroke affects a man or a woman. In our study, females who had a positive return for a stroke are approximately in the hundreds and just a few above males, but we can safely assume that they are approximately equal.

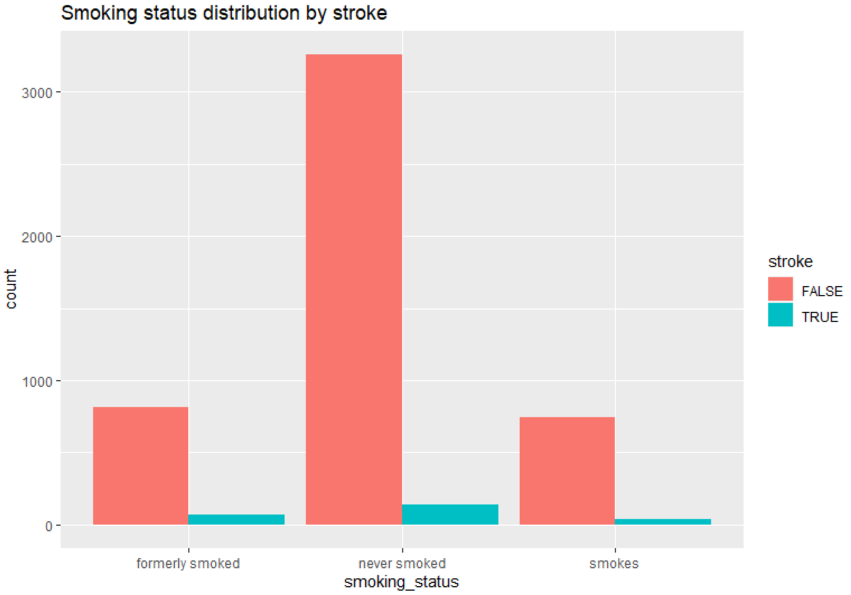

Figure 14: Smoking Status Distribution by Stroke

In figure 14, we can directly notice that the numbers of people with a false return in the possibility of a stroke, who have never smoked is drastically higher, with approximately 3400 people, than that of people who formerly smoked, with approximately 800 people, and that of people who currently smoke, with approximately 700 people. After further analysis, we can determine that the chances of a stroke does not depend on whether a person smokes, formerly smoked, or not, the results in each category for a positive return of a stroke is very low. Viewers might think that the chances of a stroke in people who never smoked is higher than that of the two other categories, but that is due to the missing values being replaced by the mode, in this case “never smoked”.

4. Conclusion

In summary, I used different methods to clean and transform the data so that I may be able to visualize the findings. I used density plots and bar charts to visualize the problem I am trying to solve. Using a density plot we were able to determine that people with ages above 40 were predicted to have a stroke. Further, we were able to determine that people with a high-level bmi and a high-level average glucose had a tendence of having a stroke. Furthermore, analyzing the gender difference in males and females having strokes, we observed that gender did not play an important role in the prediction of a stroke as the number of males and females with a positive prediction of a stroke tended to be equal. And finally, we were able to show that the smoking status of a person did not weigh heavily in our prediction of strokes.

In discussion, I chose not to analyze all the variables as I would like to come back and study this dataset in another assignment. Since, Assignment 3 will rely on partitioning and predictive models I decided to use this dataset to practice on. Most of my analysis are limited due to the lack of domain knowledge.

5. References

Fedesoriano. (2021, January 26). Stroke prediction dataset. Retrieved March 15, 2021, from https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

The internet stroke center. (n.d.). Retrieved March 15, 2021, from http://www.strokecenter.org/patients/about-stroke/what-is-a-stroke/