Stroke Prediction

ISyE 412

Madeline Olson, Riley Nowakowski, Vedant Grover, Eric Oelschlager

Introduction

Motivation

In the realm of healthcare, understanding and promptly recognizing the early signs of a stroke is crucial for effective intervention and improved patient outcomes. According to the Stroke Awareness Foundation, “about 795,000 people suffer a stroke each year” in the United States. This analysis report delves into the intricate landscape of stroke symptoms with a primary focus on predicting stroke outcomes. Stroke, a sudden and potentially life-threatening condition, demands swift and accurate assessment to mitigate its impact. By scrutinizing the diverse array of symptoms exhibited during the onset of a stroke, this analysis aims to unravel patterns and indicators that may serve as predictive factors for the subsequent course of the condition.

About the Dataset

Due to implications over the patient's privacy, we chose a synthetically generated dataset. However, it closely models real-world health data and was designed to imitate real patients. The dataset contains 19 predictor variables which can be grouped into two main categories: lifestyle and physical factors. Lifestyle factors are considered attributes such as dietary habits, stress levels, physical activity, smoking status, marital status, etc. Physical factors included blood pressure, cholesterol, hypertension, etc. Finally, there was a binary response variable of patient stroke status: yes or no.

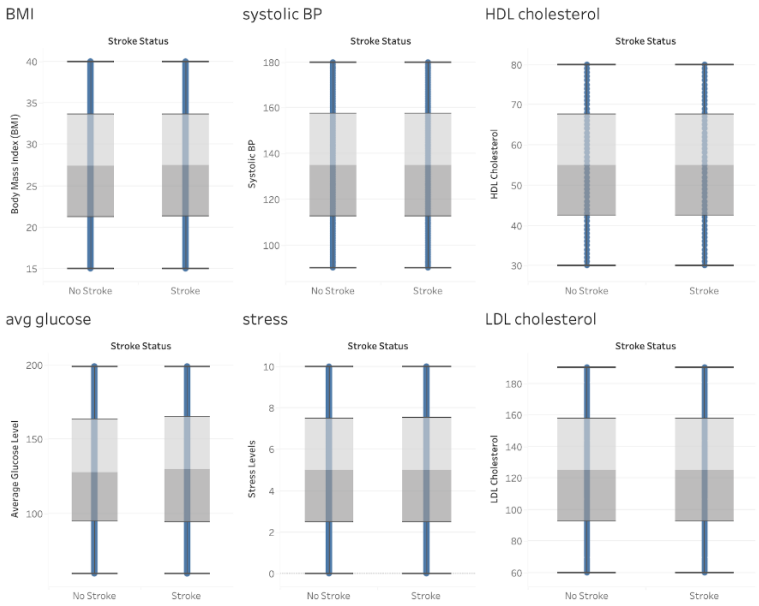
Cleaning the Data

We chose some standard methods to clean the data, first by dropping rows and columns that have missing values. There were also two columns in the dataset, Blood Pressure Levels and Cholesterol Levels, that were listed as two numbers, as part of their ‘reading’. We also decided to split these columns into two each. We made Systolic BP and Diastolic BP columns for the Blood Pressure readings, as well as HDL and LDL columns for the cholesterol readings. This made it a lot easier to analyze and visualize the effect of each factor on the possibility of stroke. We also made sure to handle and drop extreme outliers in columns such as Average Glucose Level, and we standardized the Age column. Lastly, we decided to change the Stroke ‘Diagnosis’ column from Stroke/No Stroke to 1/0, again to make analyzing and visualizing more effective.

What are the strongest predictors for predicting if a person has a stroke?

Initial Exploration

When approaching the question of determining the best predictors, there were many different graphs and visualizations considered. One of them that stood out involved examining all of the continuous predictors on a dashboard that examined a boxplot for each.



Navigating through the intricacies of the six continuous predictors presented a challenge, as discerning a standout predictor amid similar mean values for stroke and no-stroke variables proved difficult. The uniformity observed in the boxplots visually obscured potential differentiators. Furthermore, the sheer multitude of variables, encompassing both categorical and other predictors, made a direct comparison impractical. To determine significant predictors, we decided to utilize Stepwise Regression in R Studio.

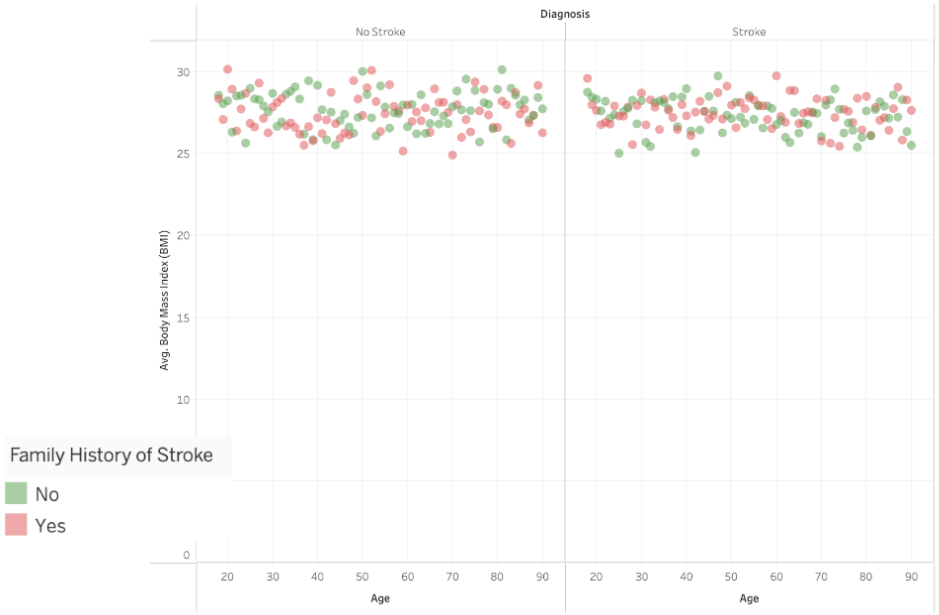
Stepwise Regression

Stepwise Regression proved to be a useful tool when identifying strong predictors for stroke likelihood. After fitting a logistic model in R Studio to predict stroke status, Stepwise Regression analysis was able to output the top three values and their corresponding P Values.

```
Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                -0.301977   0.234897  -1.286   0.1986
Family.History.of.StrokeYes -0.212561   0.114298  -1.860   0.0629 .
Age                         0.097022   0.057485   1.688   0.0915 .
Body.Mass.Index..BMI.       0.012884   0.007891   1.633   0.1025
---
```

After applying Stepwise Regression, the top three significant predictors are having a family history of stroke, age, and BMI. Of those three, the most significant would be having a family history of stroke with the smallest P Value of 0.0629. However, this P Value is not statistically significant at a confidence level of 0.05, but it is the strongest given the data set and the predictors available.

When graphing all three of these predictors, there still is not a huge variation. The data appears to be relatively uniform.

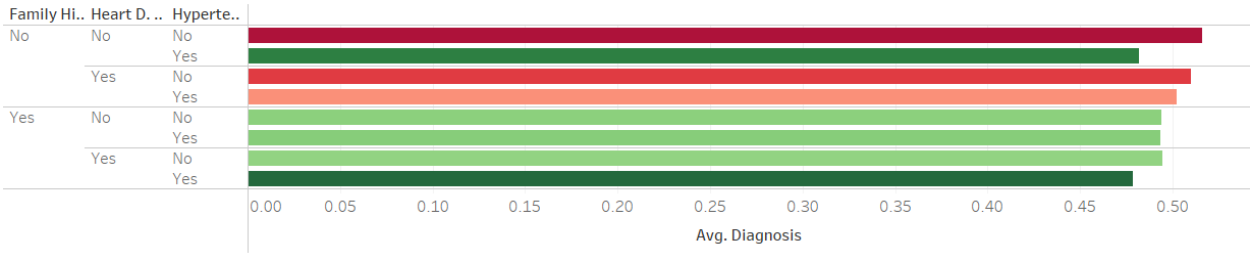


What are the Influences of Genetic, Lifestyle, and External Factors on Stroke Likelihood?

There are a variety of features in the dataset that could be summarized as genetic factors, lifestyle factors, or external factors.

Genetic Factors

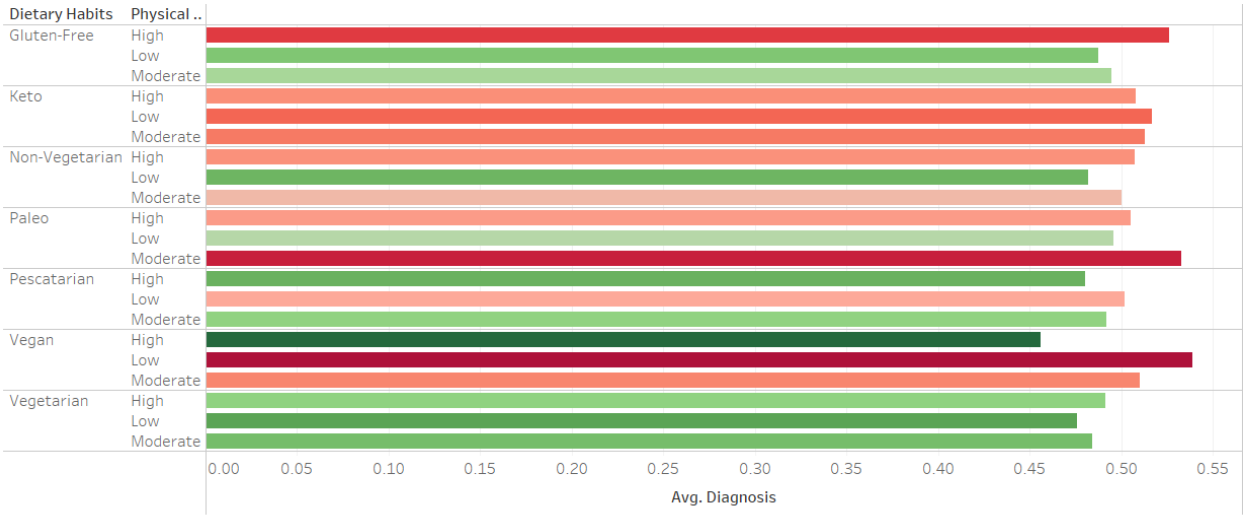
In this analysis, Genetic factors included family history of stroke, heart disease, and hypertension.



Utilizing Tableau for data visualization, the average diagnosis was plotted against every combination of these genetic factors. With a deeper red color indicating a higher average diagnosis. All of the values are relatively close, though it can be said that having heart disease increases your risk of stroke. Family history of stroke surprisingly seems to have a negative correlation to stroke. Hypertension sems to make very little impact on the likelihood of stroke.

Lifestyle Choices

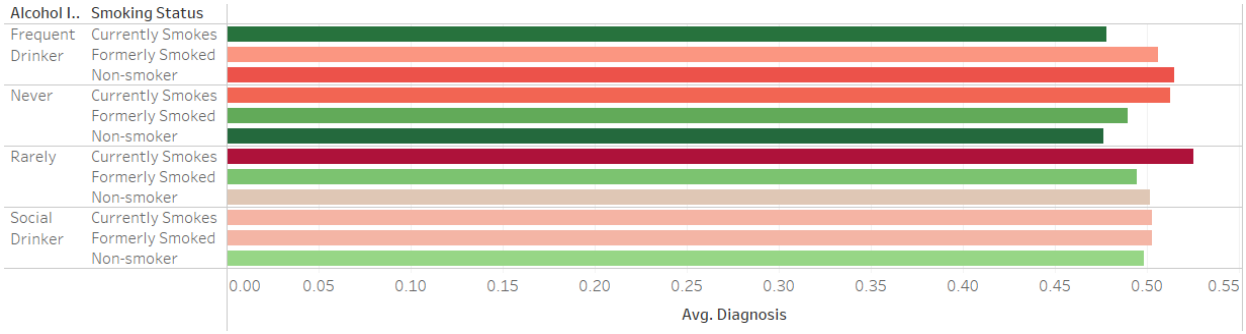
Next, lifestyle choices were analyzed. These included amount of physical activity, diet, smoking status, and alcohol intake.



Utilizing Tableau for data visualization, the average diagnosis was first plotted against every combination of diet and exercise. With a deeper red color indicating a higher average diagnosis. From this data a few things stick out. Firstly, having low physical activity generally seems to

increase your likelihood of stroke. Secondly, unbalanced diets increased the likelihood of stroke. Diets such as keto and vegan had some of the highest likelihoods of stroke, while also being some of the most unbalanced in terms of nutrition.

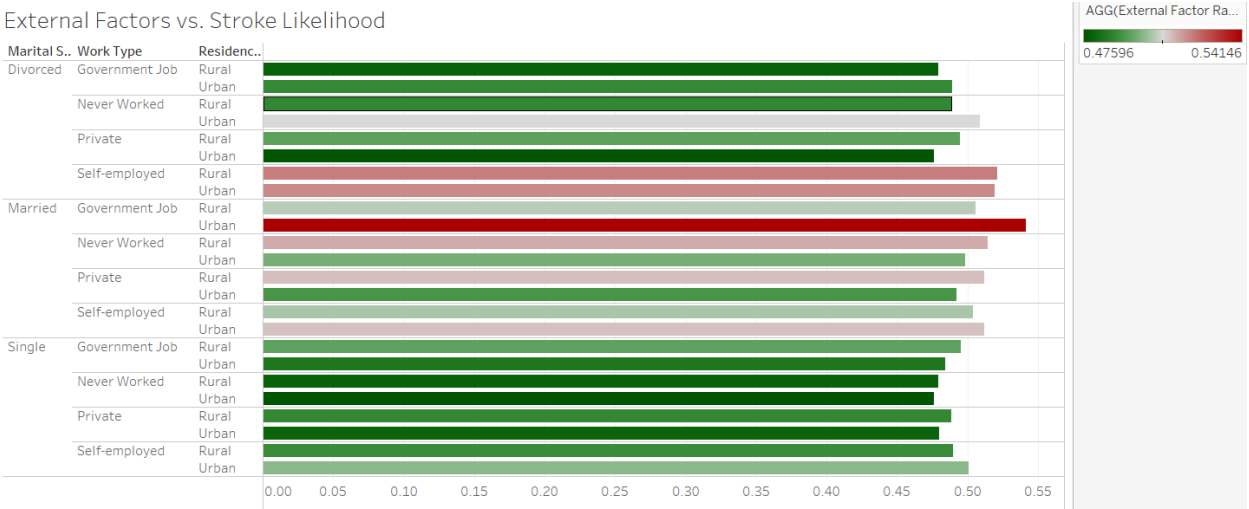
Secondly, lifestyle choices regarding smoking and alcohol were analyzed in the same way.



This analysis reveals that smoking generally increases the likelihood of stroke more than drinking. Current smokers had the highest likelihood of stroke among people in most of the alcohol consumption brackets. Rarely drinking and currently smoking poses the highest risk, perhaps because smokers who do not drink may smoke more than smokers that do drink.

External Factors

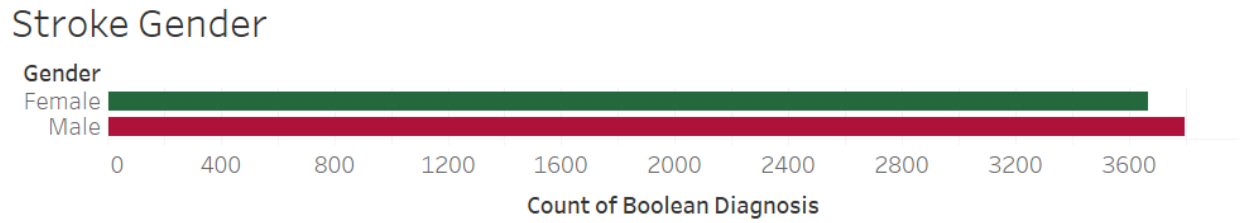
Finally External factors were analyzed. These include marriage status, work type, and home type.

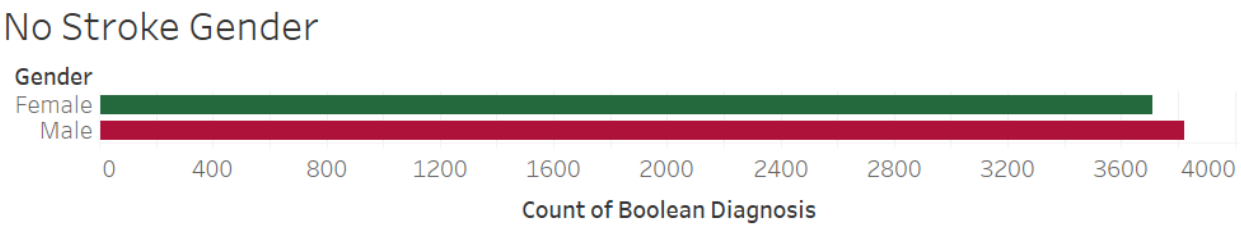


These results were visualized similarly to the other factors. The average diagnosis does not vary much. That being said, being divorced and self employed seems to correlate to higher stroke risk. This could be due to factors not shown in this visual such as stress level. The home type does not appear to have any impact on stroke risk.

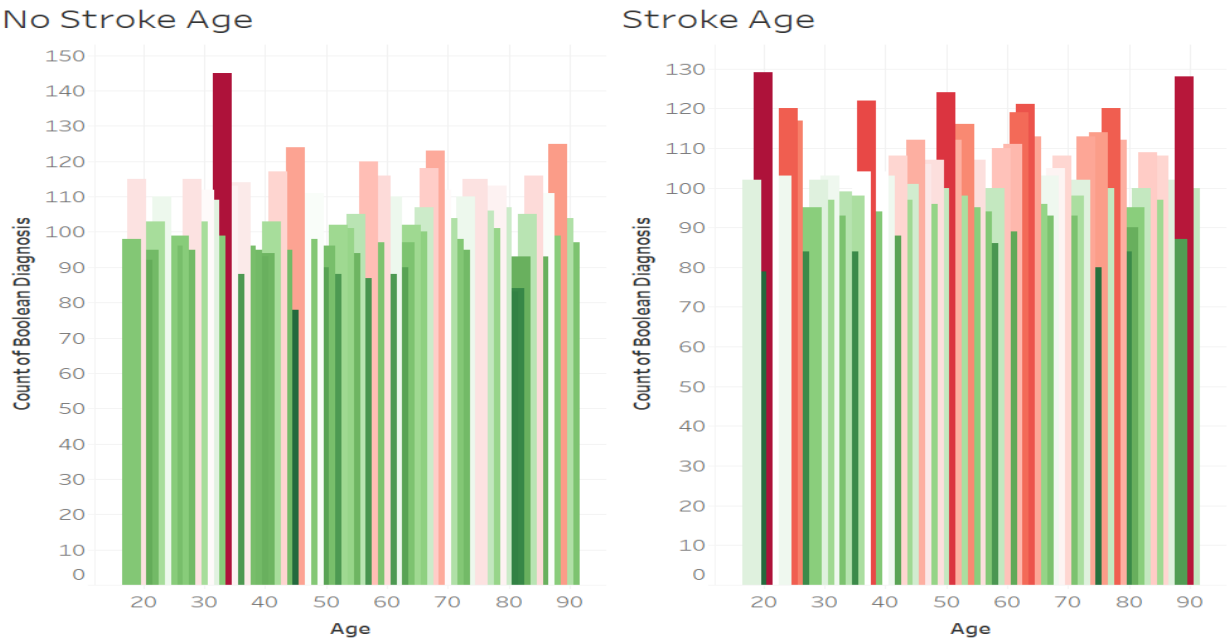
Do demographics such as gender and age play a role in who is diagnosed with a stroke and who is not?

Charts were created in Tableau to determine if these demographics had an influence on stroke diagnosis. The first chart that was created included gender and stroke diagnosis. This showed the overall number of people with each stroke diagnosis and what their gender was. There showed to be little to no change in diagnosis whether male or female. One conclusion that could be made is that males tend to come in with stroke-like symptoms more than females, as there were more males in this study than females.



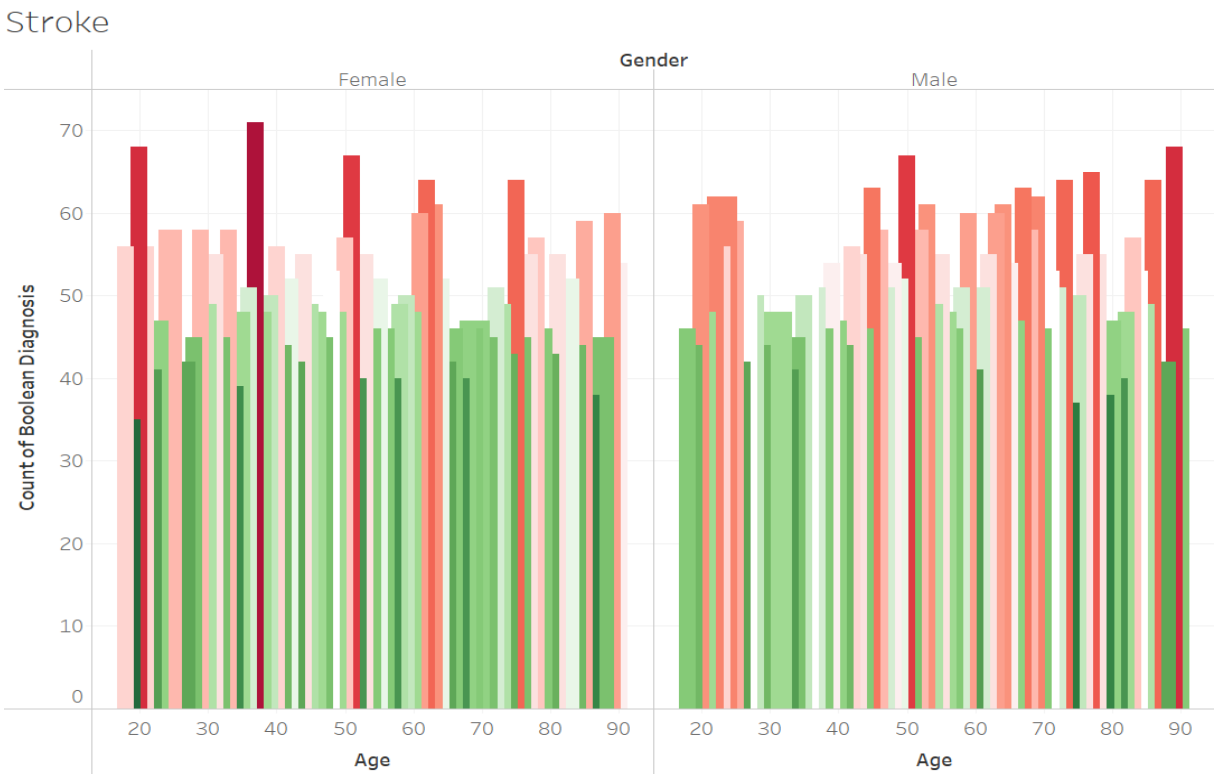


The next step was to look at Age and determine if how old a person was played a large factor in whether they would be diagnosed with a stroke or not. This bar chart showed that people are more likely to be diagnosed with a stroke in early adulthood and when they are above 70 years old.



How do age and gender together have an influence on stroke probability?

Gender and age did not have a huge influence on predicting whether or not a person was diagnosed with a stroke. To take this a step further the next question was to see if together they have an effect on stroke probability. To see this, two bar charts were created, one with males who had strokes and their age and a second with females who had strokes and their age.



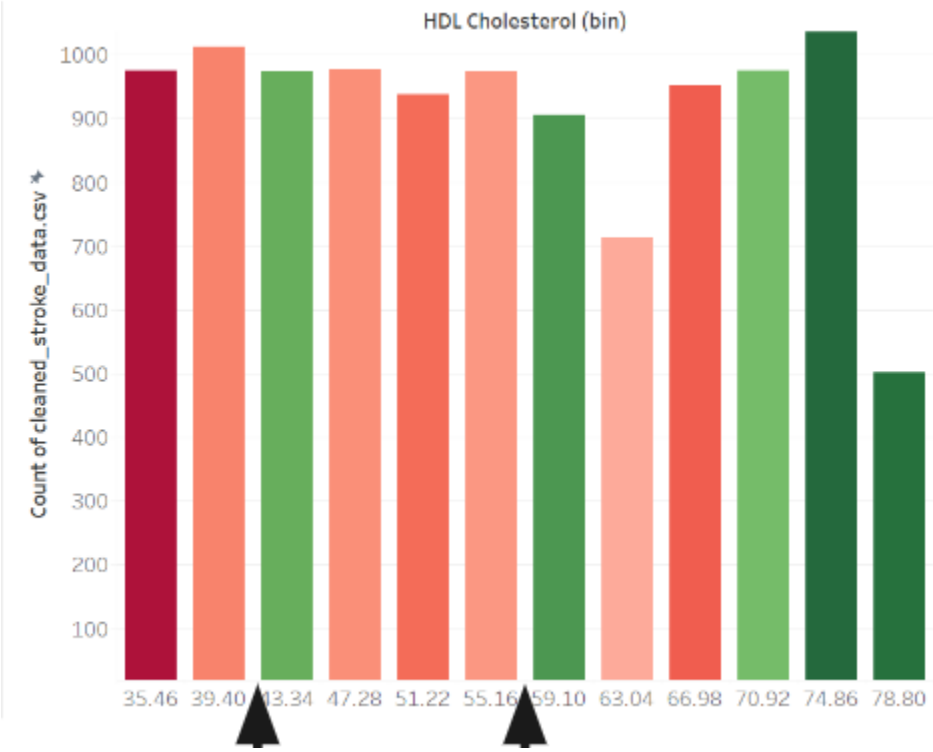
These charts show that females tend to have more strokes earlier in life than males, specifically before the age of 40. In contrast, males tend to have more strokes than females later in life mainly over the age of 70. This is useful information and can be used to help people know when they are at an elevated level of stroke possibility.

Is there any turning point of the clinical data, such as BMI, Blood Pressure, Cholesterol, stress level, etc., that represents a strong increase in strokes?

After analyzing the cleaned dataset thoroughly, we were able to find a couple factors in the dataset that presented turning points in the likelihood of getting a stroke. Using Tableau, we created graphs and visualized all the different factors against the number of stroke diagnoses in each range. The darker the red means more patients had strokes in this range, and the darker the green meaning less.

HDL Cholesterol Levels

Unhealthily low levels of HDL Cholesterol, which stands for High-density Lipoprotein, proved to negatively affect stroke risk. HDL is the lower of the two numbers in a cholesterol reading. Through previous medical research, it has been said a desirable range of HDL is 60 mg/dl+, below that presents some level of risk. However, below 40 mg/dl, is where there is very serious risk, and this was backed with the dataset, as shown in this graph, made in Tableau:

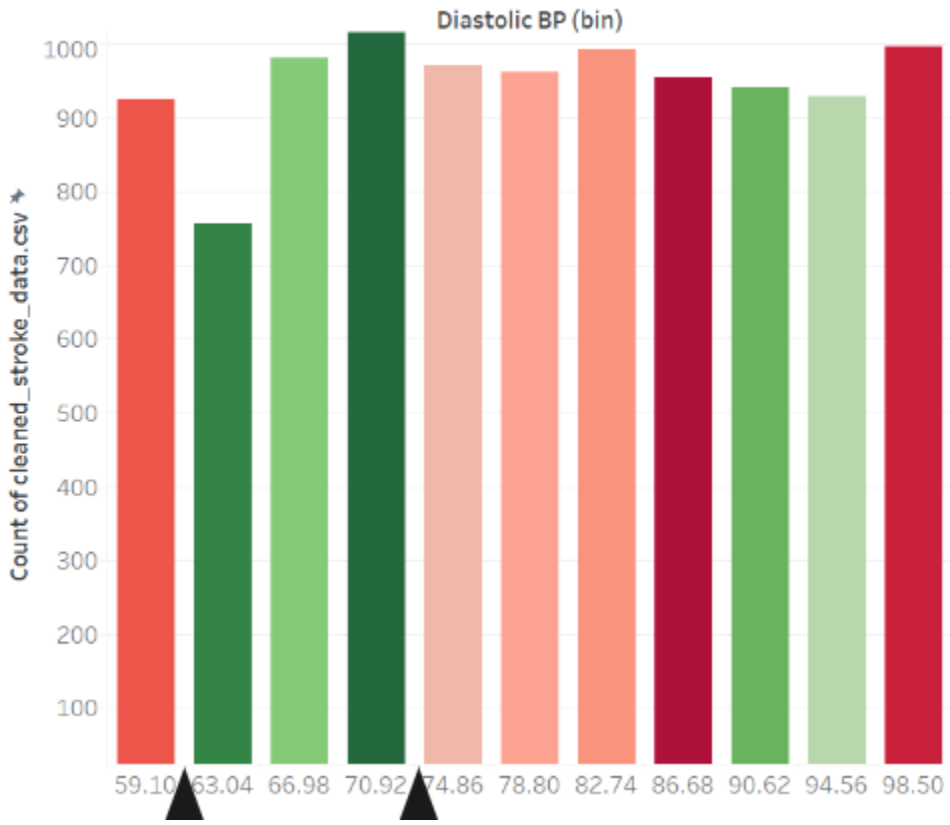


As indicated above, below the first arrow, is all darker red which means those patients with HDL levels below 40 had the highest number of stroke diagnosis. As indicated by the second arrow, below 60 is also at some level of risk, with slightly brighter red bars. Overall, above 60 mg/dl is considered healthy, and lower chance of stroke, but of course there are some anomalies which is why a couple of those bars are slightly red, but in general we can consider 40 and 60 mg/dl the turning points in HDL data. Lower HDL Cholesterol correlates with a number of long term factors, including lack of physical activity, chain smoking, alcohol abuse, Type II Diabetes, and more, showing how a more unhealthy lifestyle would increase a person's chances of stroke.

Diastolic Blood Pressure Levels

A healthy range of Diastolic Blood Pressure is said to be between 60 and 80 mm Hg. Diastolic is the lower of the two numbers in a Blood Pressure reading. There are different stages of 'High BP' above 80, and below 60 is also considered unhealthy, although it is less common. We can consider these two the 'turning points' in Diastolic and overall Blood Pressure readings that

present a higher number of strokes recorded. As we can see in the graph created in Tableau, indicated by the two arrows, is the healthy range of 60 to 80 mm Hg Diastolic BP:



These 3 bars between the arrows are all green, meaning lowest number of strokes recorded in these ranges. There is a clear and strong relationship between a healthy Diastolic BP and a lower chance of having a stroke, and outside of a couple anomalies, there is also a direct relationship between an unhealthy Diastolic BP and a higher chance of having a stroke, as indicated by majority of the bars outside the 'healthy' range being red. With this data visualization, we can see that 60 and 80 are the two turning points for Diastolic BP. Similarly to HDL Cholesterol, an unhealthy lifestyle can cause higher and abnormally low Diastolic levels. These factors include high stress levels, obesity, long term smoking, lack of physical activity, and more.

Conclusion

Our data analysis yielded several conclusions regarding factors that increase likelihood of stroke.

Firstly we performed stepwise regression and saw that age, BMI, and family history are significant for logistic regression models. While age and BMI are positively correlated to stroke likelihood, family history is negatively correlated with stroke likelihood.

In terms of genetic factors, having heart disease increases risk of stroke. Furthermore, lifestyle choices such as an unbalanced diet, lack of exercise, and smoking increase risk of stroke.

In our analysis of the effects of gender and age on stroke risk we found that females tend to have more strokes than men under the age of 40, while males tend to have more strokes when they are over 50.

In our analysis of cholesterol and blood pressure we found that low HDL increases risk of stroke and high and low diastolic blood pressure levels increase risk of stroke