Stroke Analysis

Vram Papyan, Leonid Sarkisyan, Elina Davtyan, Elvina Nosrati

American University of Armenia

DS 116A: Data Visualization

Professor: Gevorg Atanesyan

2023-12-03

Abstract

Stroke is an enormous public health concern, the magnitude of which can be reduced mainly by effective stroke prevention and less so by effective treatment of acute stroke. The greatest effect is likely to be achieved by a mass approach to prevention, which consists of changing lifestyle behaviors (e.g., less smoking and less intake of salt, alcohol, and fat) among the general population through public education and, most importantly, government legislation. Strokes are caused by the blockage of blood flow to the brain (Ischemic Stroke) or sudden bleeding in the brain (Hemorrhagic Stroke). Many things raise the risk of stroke. Some of these risk factors can be changed to help prevent a stroke or future strokes. The goal of the following study is to analyze possible factors that can lead to a stroke.

Research Methodology

The data was provided in "xlsx" format, and R programming language was used to analyze and create the necessary plots. An interactive dashboard was created using RShiny to apply navigation throughout the data. The research mainly consists of the creation of Bar Plots, Histograms, Scatter Plots, Violin Plots and Box Plots for visualizing our hypotheses.

Data

The following data was downloaded from Kaggle, and it will be used to visualize Heart Stroke information, which contains the following features.

Patient ID	**Patient Name**	**Age**	**Gender**
Hypertension	**Heart Disease**	**Marital Status**	**Work Type**
Residence Type	**Average Glucose Level**	**Body Mass Index (BMI)**	**Smoking Status**
Alcohol Intake	**Physical Activity**	**Stroke History**	**Family History of Stroke**
Dietary Habits	**Stress Levels**	**Blood Pressure Levels**	**Cholesterol Levels**
Symptoms	**Diagnosis**		

Hypotheses and Results

Hypothesis 1: The prevalence of heart disease varies significantly between different work types and is influenced by smoking status.

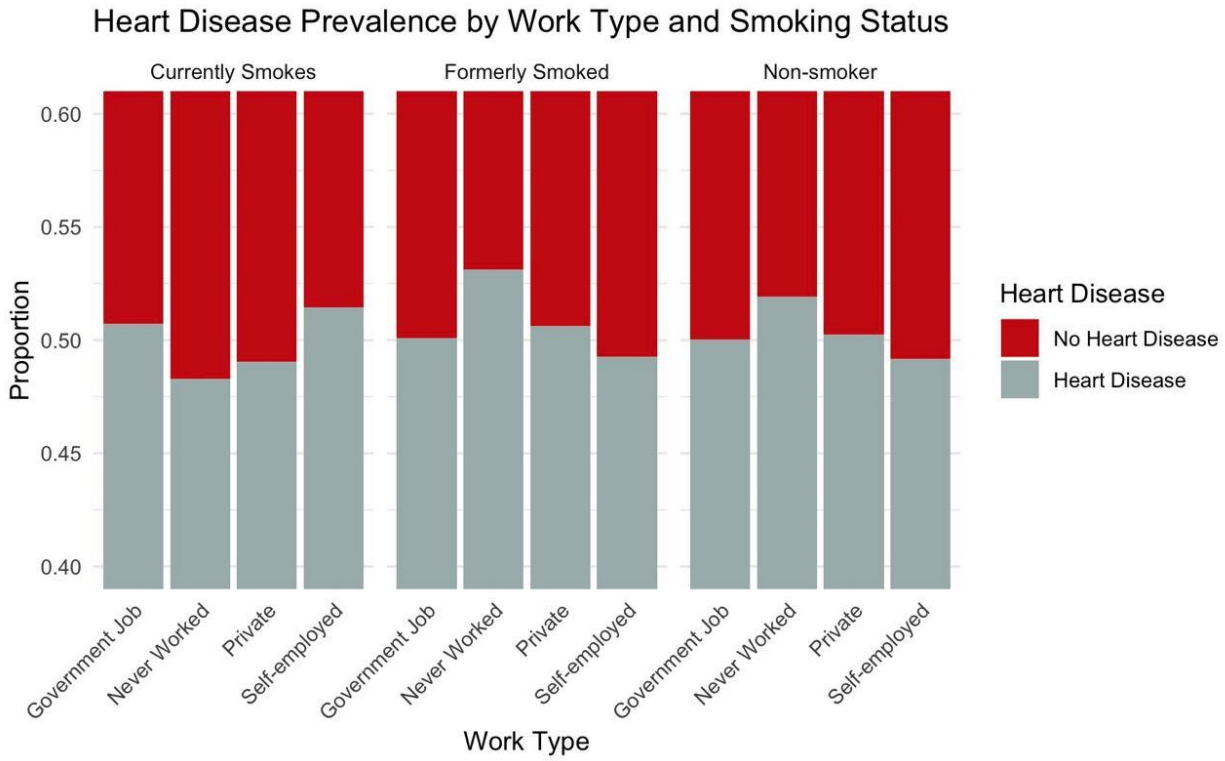

Figure 1: *Heart Disease Prevalence by Work Type and Smoking Status*

The chart portrayed in Figure 1 takes on a glace of the impact of work type and smoking status on the prevalence of heart diseases. It is divided into 3 smoking statuses: Currently Smoking, Formerly Smoked, and Non-smoker. The x-axis is the work types for each smoking status. There are four of them: Government Work, Never Worked, Private, and Self-employed. Let's look at the bar, on average the proportion for each of those categories fluctuates around 0.5(50%). Surprisingly overall Current smokers have less recorded heart disease, with the peak ratio within Self-employed over 0.52, Government Work over 0.5, and others under 0.5. Former smokers and Non-smokers have approximately the same ratio among all 4 types of work. The peak ratio is recorded to be among Never Worked with above 0.525, Government work and Private work keep at 0.5, and Self-employed under 0.5. In conclusion, most former smokers are elderly, adding the fact that age adds up to recorded heart disease it is not surprising that the ratio is higher among former smokers and for the same reason for non-smokers who are mostly older. Once again, the ages mislead of thought that current smokers have less recorded heart disease

Hypothesis 2: Average glucose levels are significantly different among patients with different dietary habits.

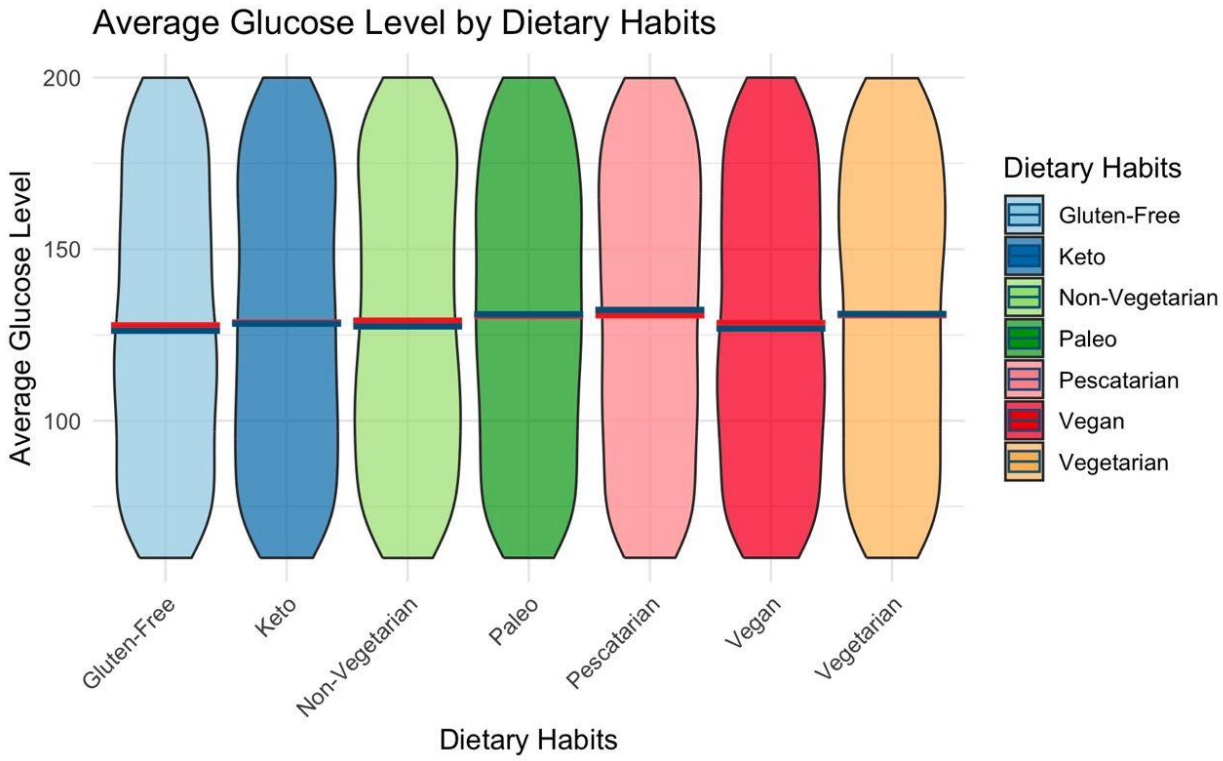

Figure 2: *Average Glucose Level by Dietary Habits*

Medically speaking glucose levels have a significant effect, so the second Hypothesis presented in Figure 2, takes into account the Glucose Levels distribution among Dietary Habits visualized each by a violin plot, where the median line is drawn in blue and the mean in red. We are looking to find any significant peaks in diet to determine the connection between Stroke to Glucose level. From the plot, we can see that overall range of each diet is the same, and the differences in mean and median are almost the same, by taking a closer look the only noticeable thing is that the Keto, Non-Vegetarian and Vegetarian have a similar distribution where the data is concentrated around 175 and 75. For another Diet the data is smooth and even from 50 to 175, and for all categories its shrinking at the start and end of the range. The records of the distributions don't have significant differences in distribution, which we conducted from the shapes and the mean median analyses. We can conclude that for this data there is no need to make any hypothesis about glucose level as the dietary analyses suggest an even distribution.

Hypothesis 3: The occurrence of different symptoms varies across age groups for stroke patients.

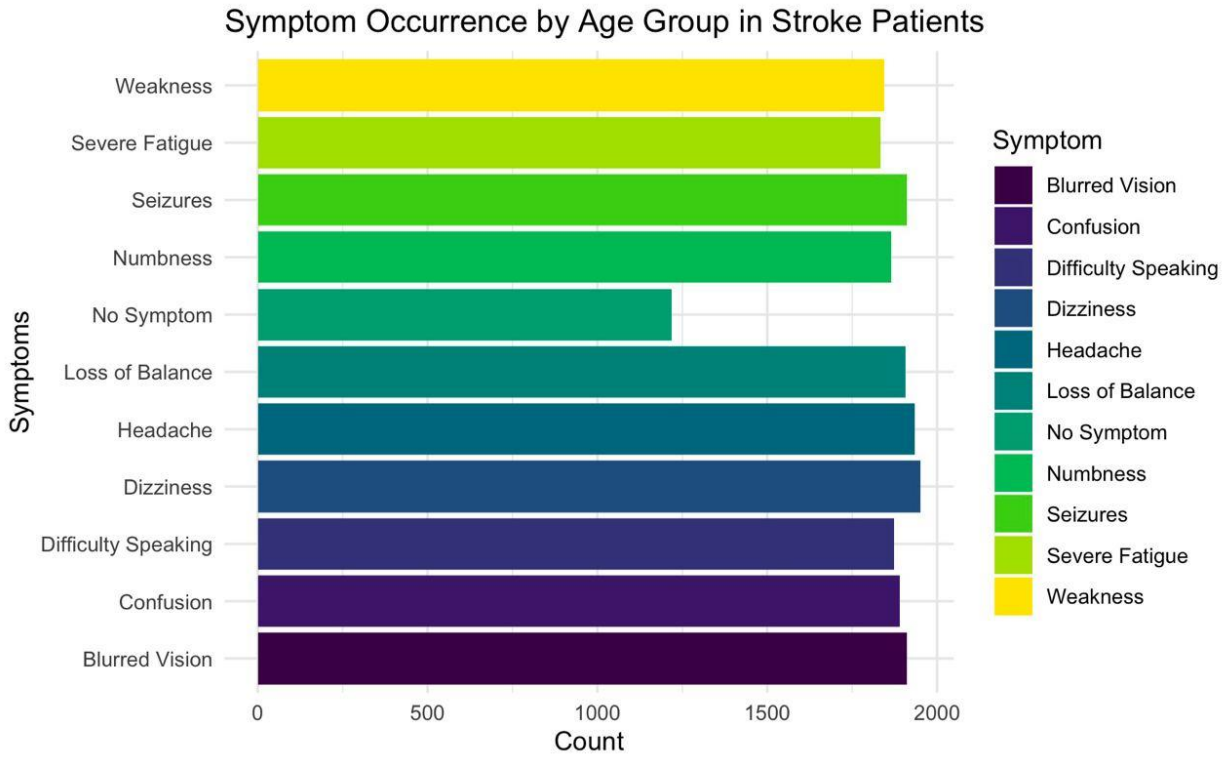

Figure 3: *Symptom Occurrence by Age Group in Stroke Patients*

The 3rd hypothesis displayed in Figure 3 suggests checking if there are differences in the symptoms occurring in different age groups. We divided the age group into 18-35, 35-60, and 60+ years old groups and plotted a flipped bar chart where the y-axis represents symptoms, and the x-axis is the count to get evidence of any patterns or changes. Overall, the pattern of counts for each group is mostly the same. A closer look at the plots identifies that for all groups No Symptoms has the lowest count, which is almost half less than the count of other symptoms, suggesting that for most of the cases Stroke will occur with Symptoms. The Elderly (60+) and Adults (35-60) have the same number for nearly the same count ratios, on the other hand, Younger People (18-35) have lower count ratios on Weakness, Dizziness, Loss of Balance, Numbness.

Hypothesis 4: Different work types have distinct distributions of stress levels.

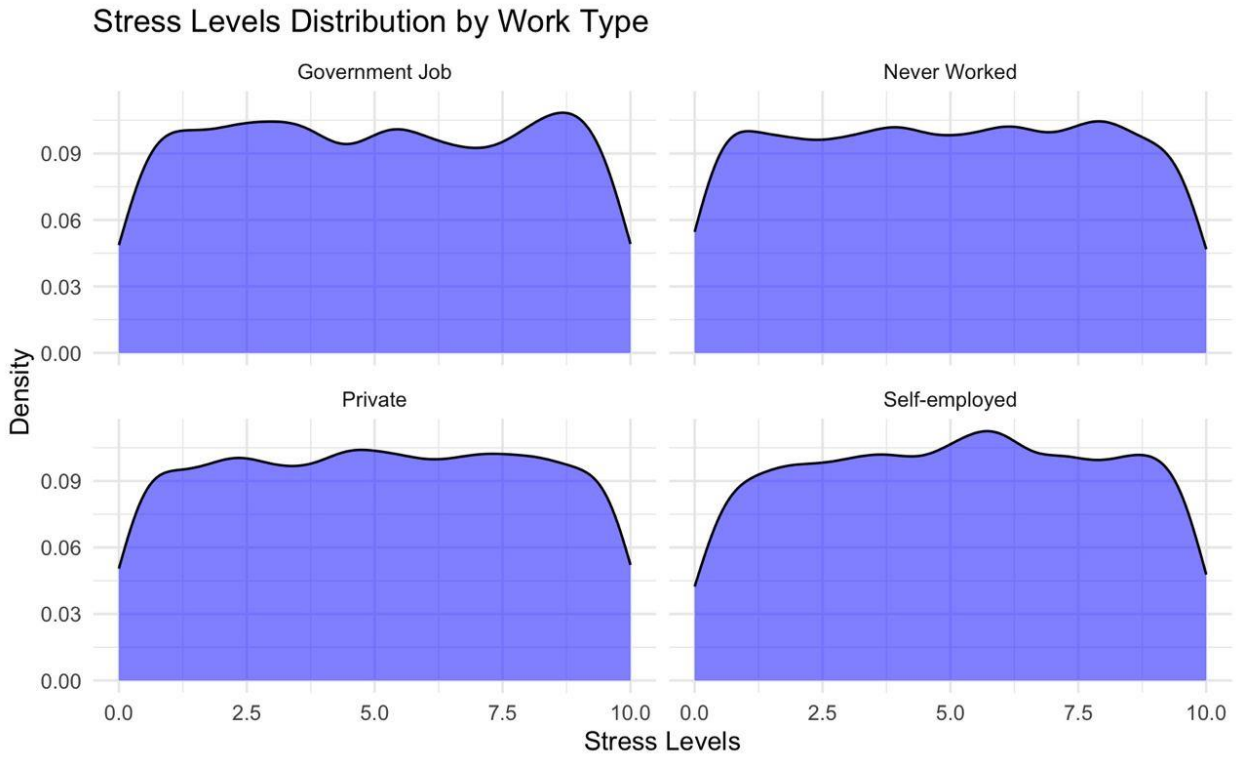

Figure 4: *Stress Level Distribution by Work Type*

Examining the density plot in Figure 4 gives us insights into the distribution of stress levels across various work types. The density plot estimate illustrates the density of stress levels, with each work type presented in a separate facet. The blue fill color with enhances the visualization by highlighting areas of higher density. From the plot, it appears that certain work types may exhibit distinct patterns in stress levels, with potential peaks or variations indicating differential stress experiences. Thus, this visualization confirms our initial hypothesis of different work types having distinct distributions of stress levels.

Hypothesis 5: The Interaction of Average Glucose Level and BMI Impacts The Stroke Risk Differently for Various Age Groups.

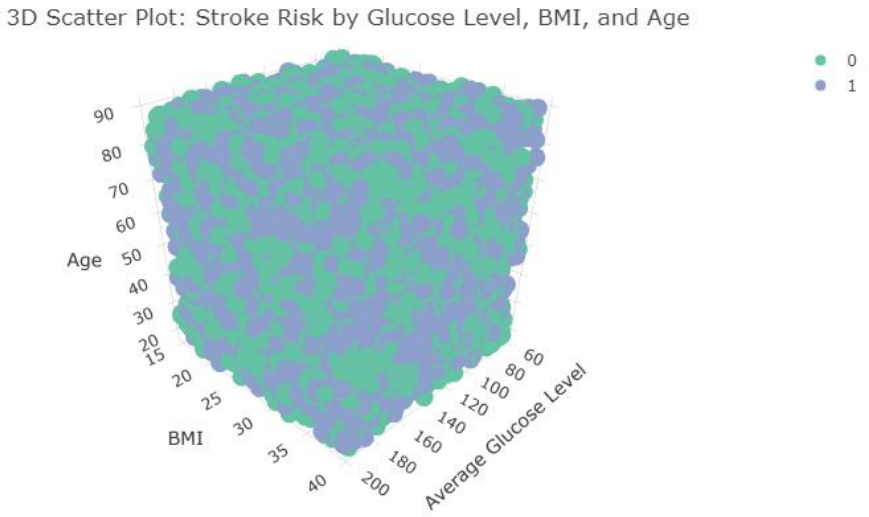

Figure 5: *Stroke Risk by Glucose Level, BMI, and Age*

The 3D Scatter Plot presented in Figure 5 consists of markers colored and symbolized based on the Stroke History column. The axes represent Average Glucose Level, Body Mass Index (BMI), and Age, and the layout includes a title, axis titles, and a legend (No Stroke history represented by "0", Stroke History represented by "1"), which make it easier for the viewer to understand. As we can see, the interaction of average glucose level and BMI impact the Stroke risk differently for various age groups.

Hypothesis 6: The Distribution of Stress Levels Across Different Age Groups Follows Distinct Patterns.

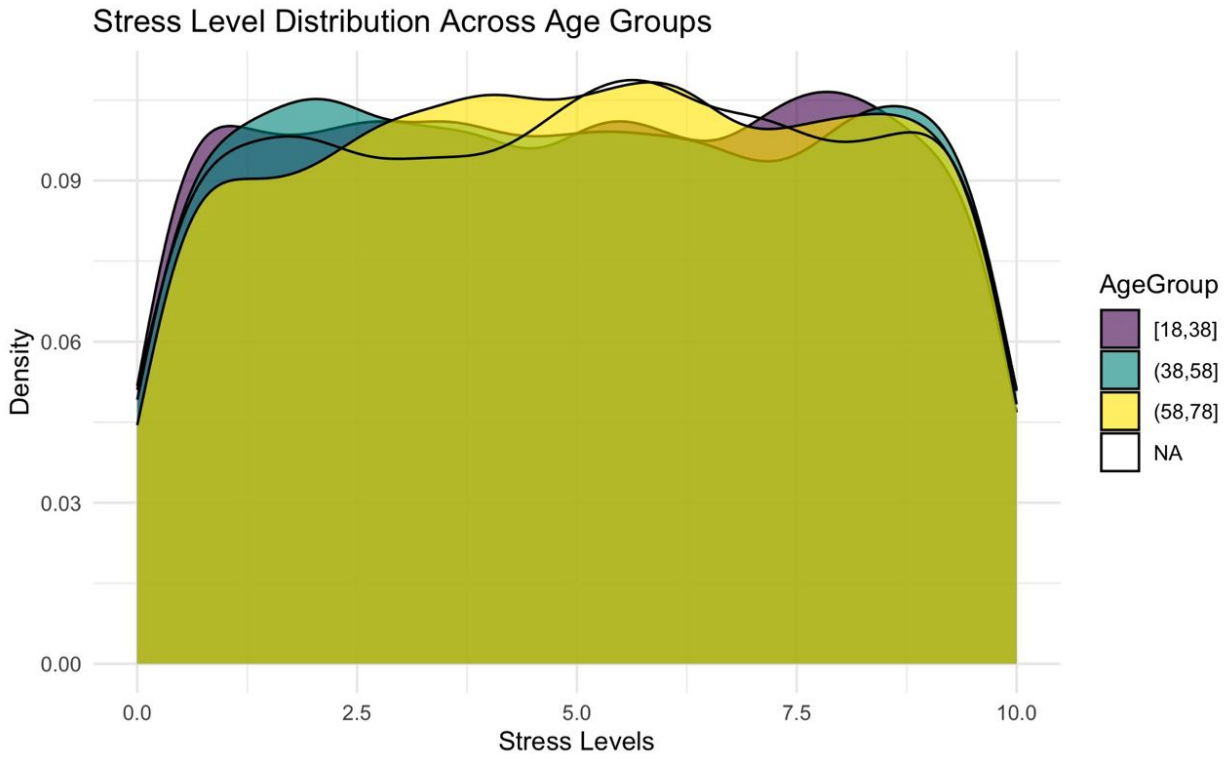

Figure 6: *Stress Level Distribution Across Age Groups*

The plot presented above (Figure 6) is a visual representation of the distribution of stress levels across different age groups (Based on 20-year Age Intervals), with a density plot showing the intensity of stress levels for each group. The transparency allows for a clearer view of overlapping areas in the density plot. As we can infer the distribution of stress levels across different age groups follows distinct patterns.

Hypothesis 7: Alcohol intake patterns differ among patients with and without a family history of stroke.

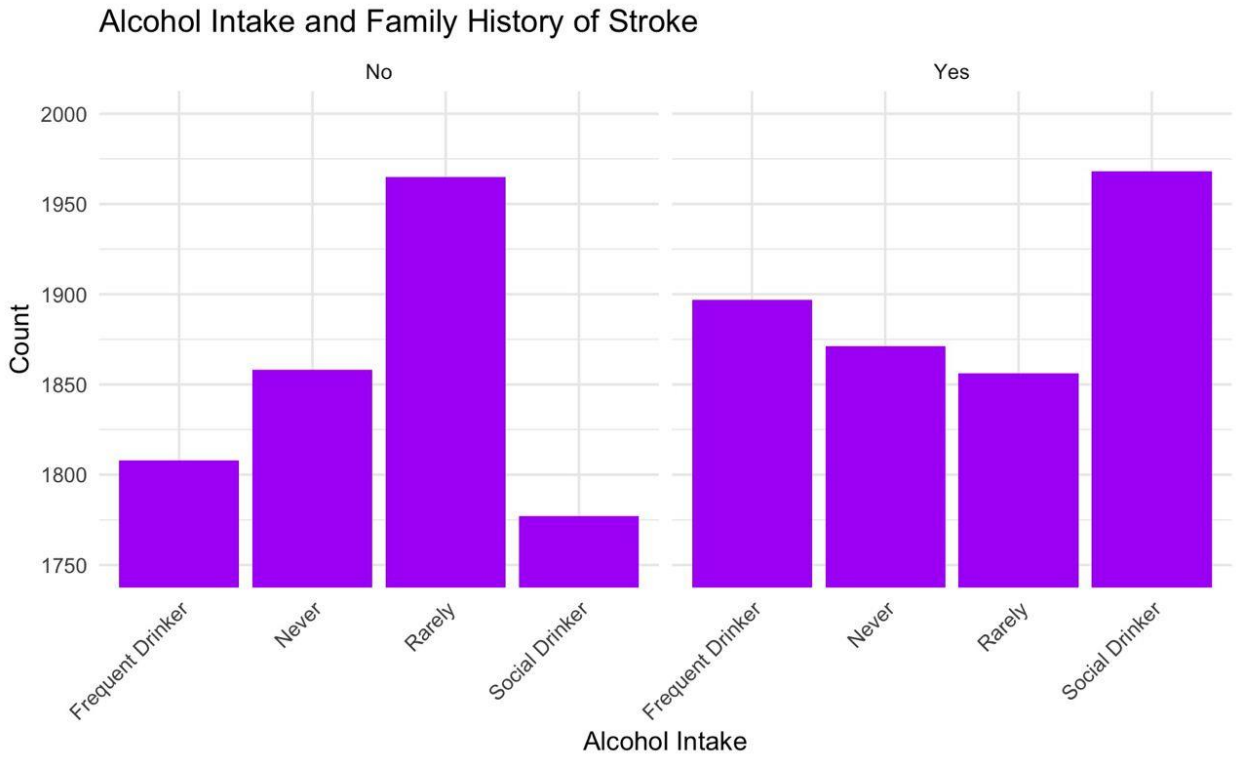

Figure 7: *Alcohol Intake and Family History of Stroke*

The plot presented in Figure 7 is a comparative visualization of two types of families. On the left we can see a visual representation of families with no history of stroke, while on the right a representation of families who do have a history of stroke. As we can see families with no history of stroke (Left Hand Side) typically tend to consume less alcohol in general. On the other hand, families who do have a history of stroke (Right Hand Side) tend to drink more. The plot's emphasis on a stroke count range improves the visualization's clarity.

Hypothesis 8: This hypothesis suggests that the relationship between a patient's Age and their Average Glucose Level may vary depending on their Gender.

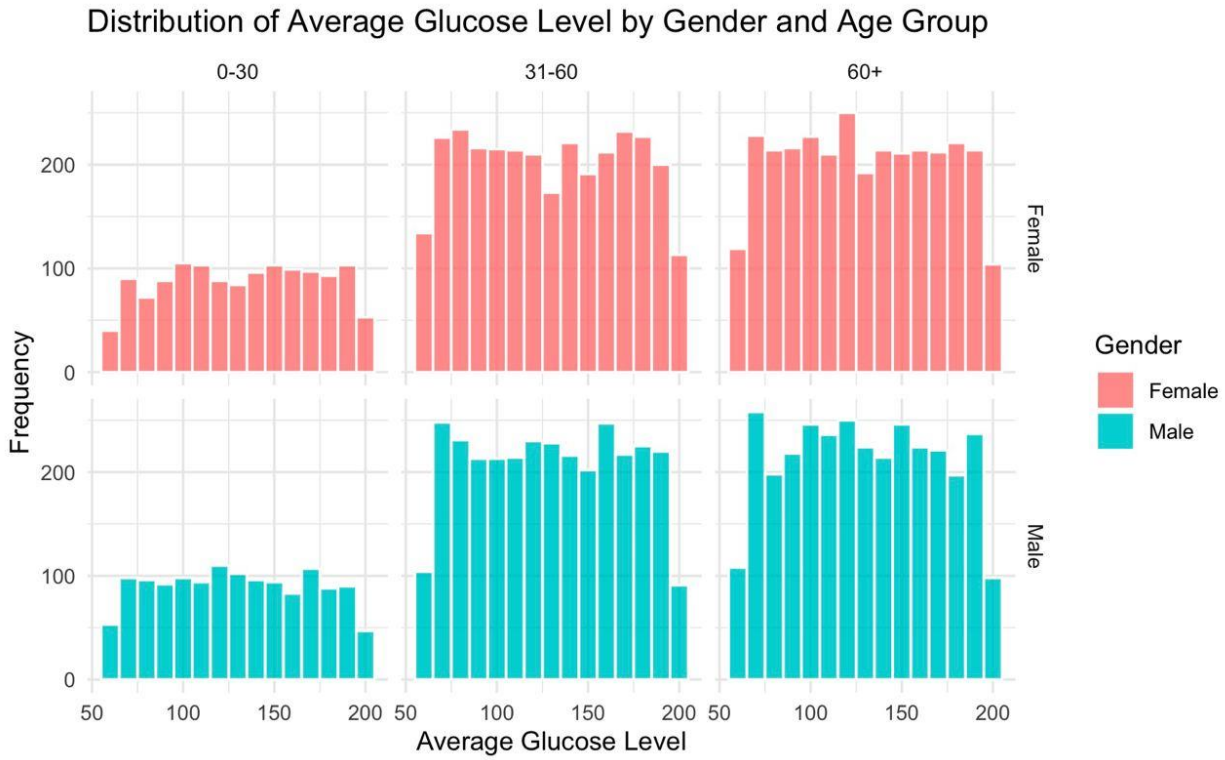

Figure 8: *Distribution of Average Glucose Level by Gender and Age*

Upon analysis of our plots presented above (Figure 8), the results show a remarkable consistency in the measured variable between male and female patients. In terms of age, an apparent trend reveals that patients who are 31 years of age and older tend to have higher average glucose levels. This proves the first hypothesis—that a patient's age and average glucose levels are correlated—to be true. Moreover, it is significant that the correlation between age and mean blood glucose levels could vary depending on the patient's gender, suggesting that gender could have an impact on this relationship.

Hypothesis 9: Average BMI varies significantly between age groups and is affected by physical activity levels.

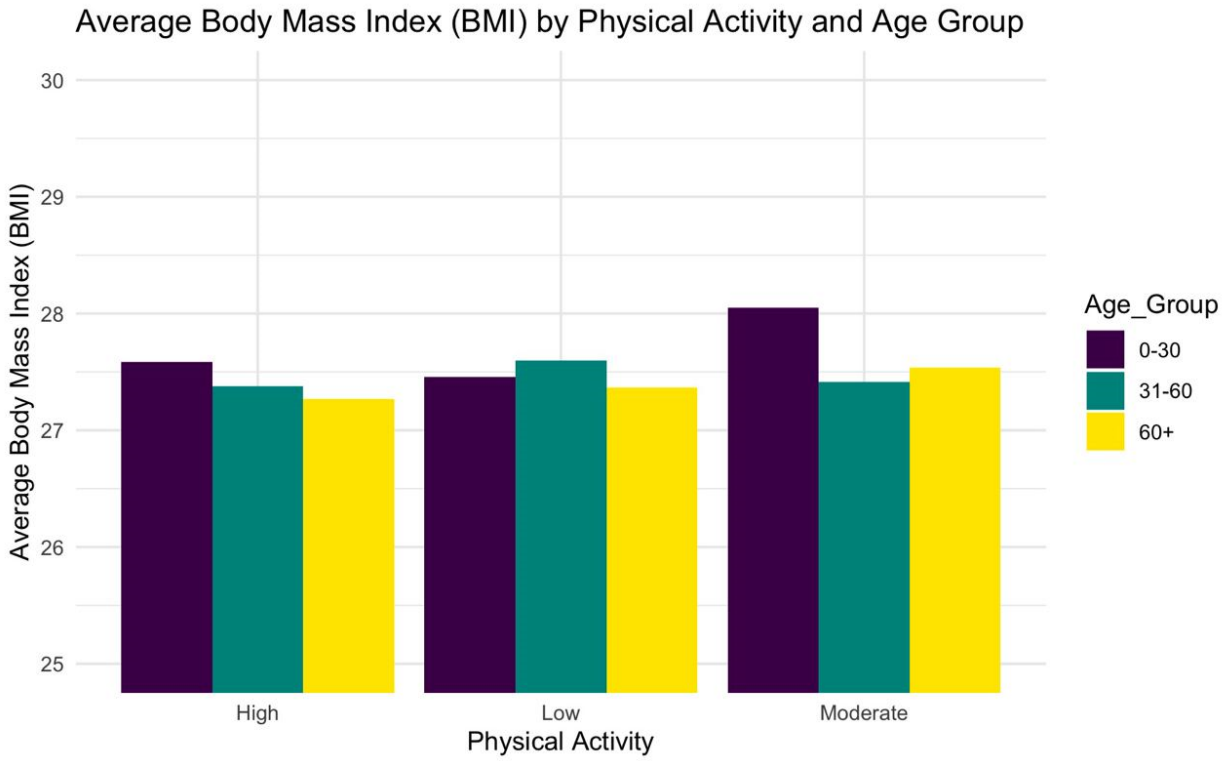

Figure 9: *Average Body Mass Index (BMI) by Physical Activity and Age Group*

By looking at the side-by-side bar chart presented in Figure 9, we can say that generally, there is a trend suggesting that higher levels of physical activity are associated with lower average BMI in all age groups. Also, the average BMI for each physical activity category increases with age. The youngest age group (0–30) has the lowest average BMI, while the oldest age group (60+) regardless of their level of physical activity, has the highest average BMI. Interestingly, for the age group 31–60, moderate physical activity seems to be associated with a higher average BMI than low physical activity. Talking about the hypothesis, the graph does show that the average BMI increases with age. The 0–30 age group has the lowest average BMI, the 31–60 group has a higher average BMI, and the 60+ group has the highest average BMI across all levels of physical activity. On the contrary, the graph suggests that higher levels of physical activity are associated with lower average BMI in the 0–30 and 60+ age groups. However, in the 31–60 age group, the average BMI is slightly higher for those with moderate physical activity compared to those with low physical activity, which does not completely align with the hypothesis.

Hypothesis 10: There is a differing prevalence of strokes between individuals residing in urban and rural areas.

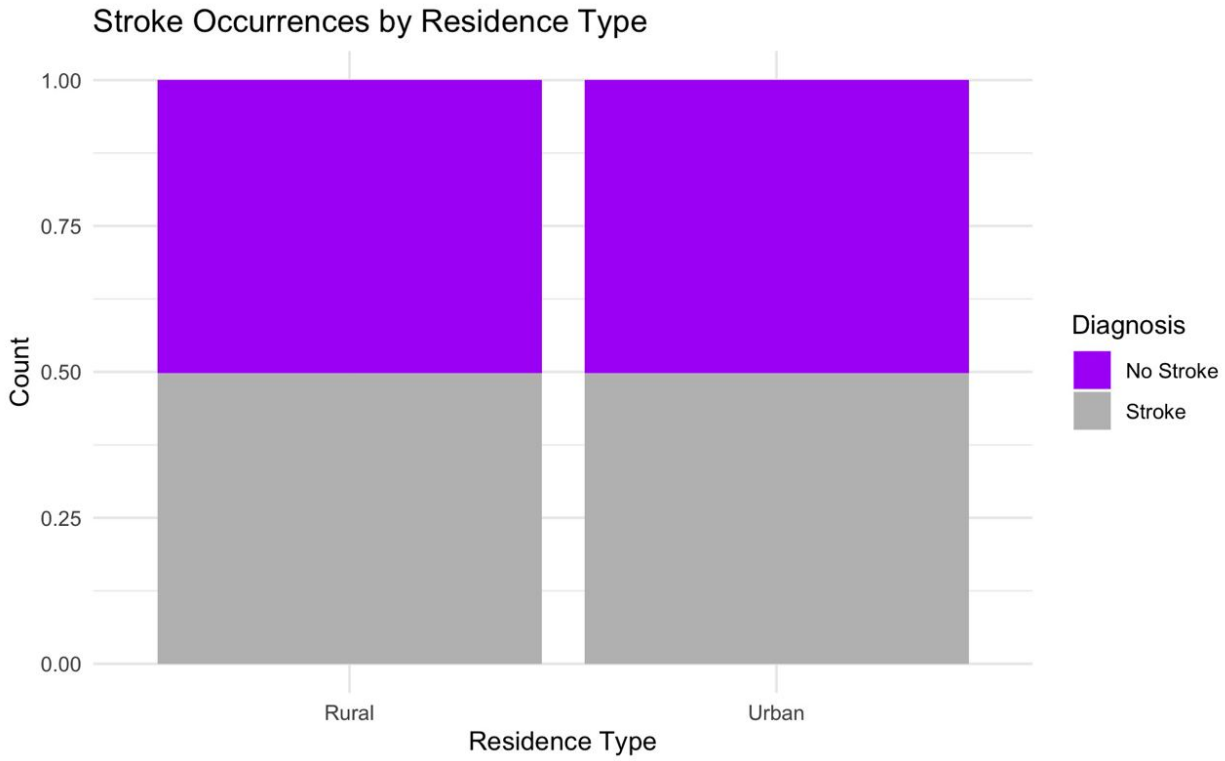

Figure 10: *Stroke Occurrences by Residence Type*

Upon examining the bar plot (Figure 10), it becomes apparent that there's no noticeable difference in stroke occurrence between people living in urban and rural areas. This finding contradicts our initial assumption that the living area might be connected to the likelihood of having a stroke. Therefore, based on this data, we must discard the idea that one's residential area is linked to the probability of experiencing a stroke.

Hypothesis 11: There is a correlation between marital status and the incidence of strokes.

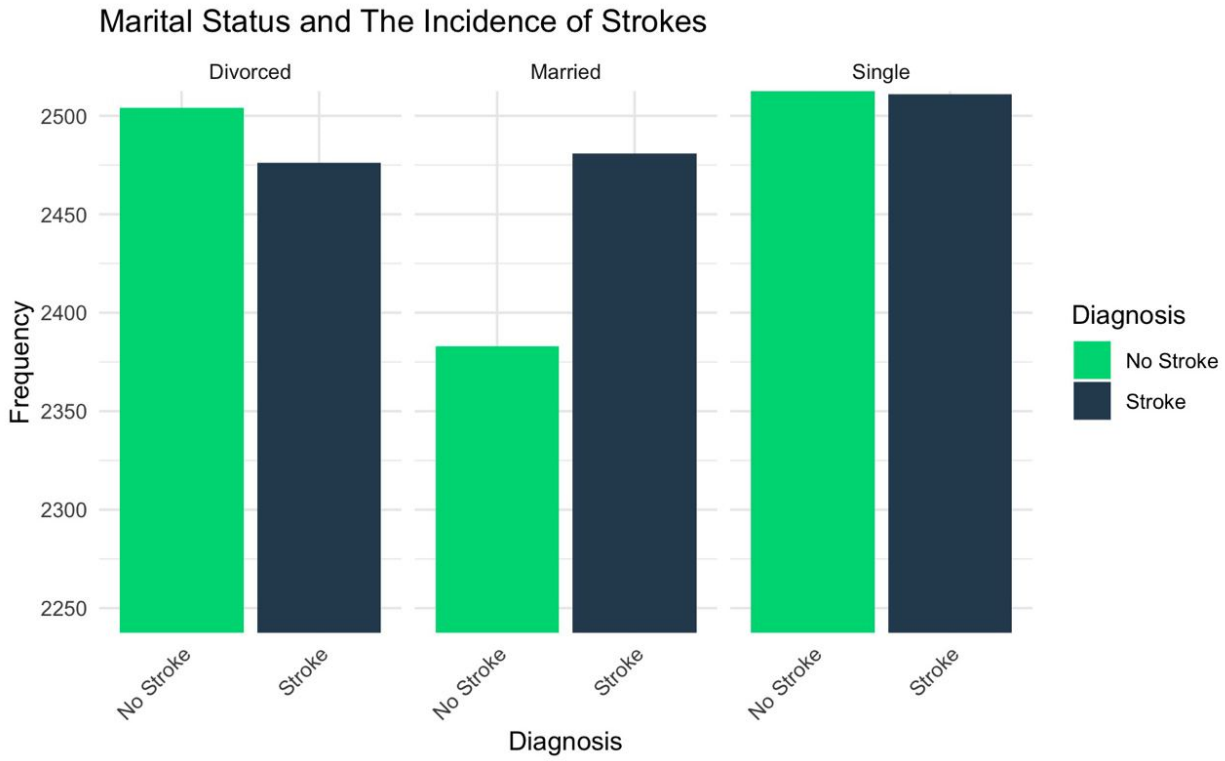

Figure 11: *Marital Status and The Incidence of Strokes*

Upon examining the side-by-side bar charts (Figure 11), patterns regarding the incidence of strokes among different marital statuses become apparent. In particular, the data suggests that those who are categorized as single had a similar risk of suffering a stroke. But there is a small difference when you compare the married and divorced groups; the married population has a somewhat higher risk of strokes than the divorced population does. Thus, in this instance, we can conclude that there is a relationship between the frequency of strokes and marital status.

Hypothesis 12: Sedentary work environments, such as government jobs have higher risk of stroke than other work types.

Figure 12: *Relationship Between Work Type and Age by Diagnosis*

Examining the boxplot presented in Figure 11 provides an insight about the relationship between age, work type, and stroke diagnosis. The age distribution across various work kinds, broken down by stroke diagnosis status, is plotted. Interestingly, the data suggests that age distributions differ between different types of work and are affected by the incidence of strokes. These patterns are highlighted by the use of distinguishing colors, orange for "Stroke" and light blue for "No Stroke." It is clear from examining the distribution and core patterns within each work type that some occupational categories are linked to either a higher or lower risk of strokes. Thus, we can deduce a link between occupational characteristics and the observed patterns in age distributions among diagnoses and work types.

Hypothesis 13: The younger you are, the more likely you are to be a current smoker.

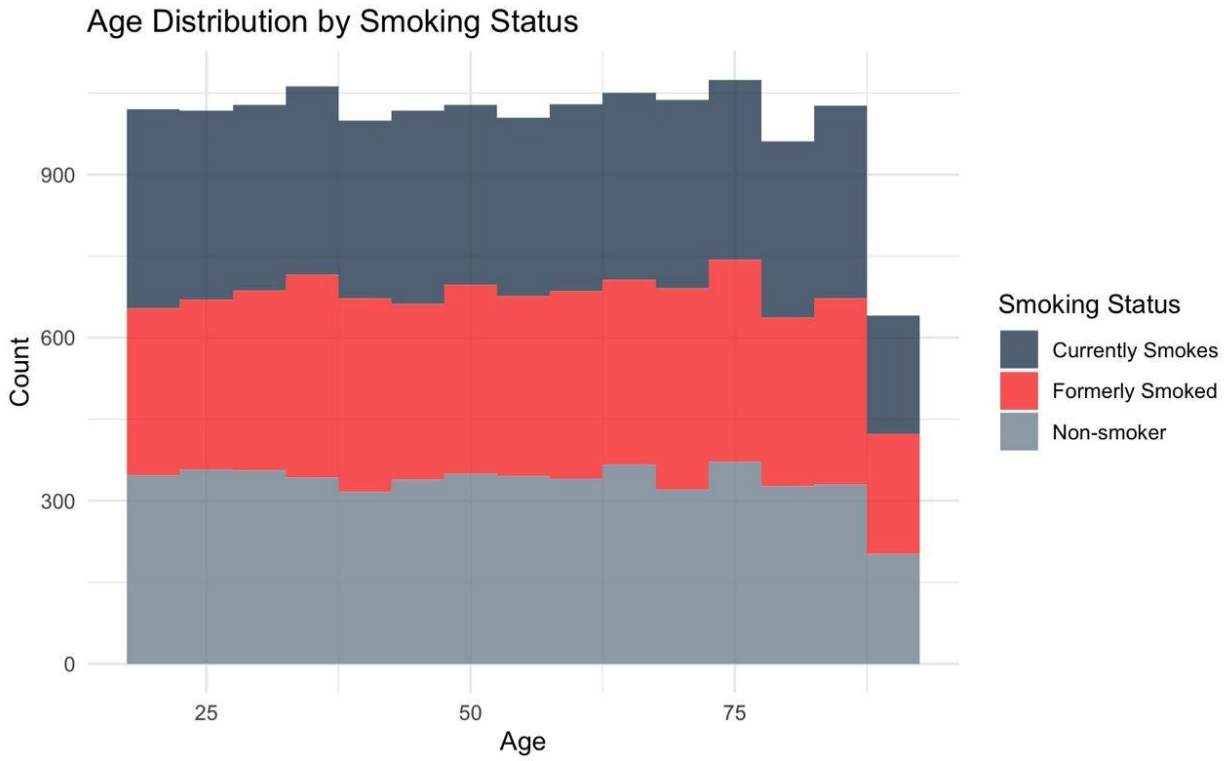

Figure 13: *Age Distribution by Smoking Status*

Analyzing the histogram visualization in Figure 13, gives us distinct patterns in the age distribution based on smoking status. The plot depicts the frequency distribution of ages, categorized by different smoking statuses: non-smoker, formerly smoked, and currently smokes. Consisting of a bin width of 5, the histogram effectively showcases the density of age groups within each smoking category. The colors, comprising shades of gray for non-smokers, deep red for former smokers, and dark blue for current smokers, enhances the visual contrast. As depicted, there appears to be a concentration of younger individuals among current smokers, while former smokers and non-smokers exhibit a broader age distribution. Consequently, these observations prove our initial hypothesis mentioned above.

Conclusion

To sum up, taking into account all factors mentioned above, we can conclude that most of our initial hypotheses were true. Most notably Hypotheses Number 1, 4, 5, 6, 7, 8, 11, 12 and 13. Closely monitoring Heart Disease and Stroke is a very important topic, which with efficient visualizations can save many lives. A good suggestion for the future would be to include more information in the data such as diabetes or blood cholesterol levels, which can significantly enhance the diagnosis and visualization part of things.

References

Andrewg, (2022), “Stroke Prediction”, Kaggle, *https://www.kaggle.com/code/andreiguliaev/stroke-prediction-with-r-and-tidymodels/input*

American Stroke Association, (2020), “The Connection Between Diabetes and Stroke”, American Heart Association, *https://www.stroke.org/-/media/Stroke-Files/Lets-Talk-About-Stroke/Prevention/Lets-Talk-About-the-Connection-Between-Diabetes-and-Stroke.pdf*