ISyE 412 Final Project Report

December 2022

Stroke Occurrence and Leading Factors

Anthony Wu, Eira Puliyelil, Hema Priya Nakka, Ross Oglesby

Table of Contents

I. Introduction

With more than 200,000 US cases per year, strokes and factors that contribute to them are becoming a highly discussed concern for many. Our team strived to use explanatory variables such as hypertension, heart disease, marital status, gender, residence type, smoking status, age, average bmi, average glucose levels, and job type to see what relationships we could visualize, using tableau, about stroke occurrence. Some questions we wanted to answer include: Are there any obvious patterns present? What findings contradict what we believe? What findings reaffirm what we believe?

II. Background on Dataset

We used a stroke prediction dataset which we found on Kaggle, called Stroke Dataset Prediction (Stroke Dataset Prediction), to conduct our tableau analysis on. This dataset contains 5110 observations and 12 attributes. The 12 attributes which also serve as the column names include id, gender, age, hypertension, heart disease, ever married (marital status), work type, residence type, average glucose level, bmi, smoking status, and stroke. The id column contains a unique identifier. Gender is categorized as “male”, “female”, or “other”. Age lists the age of the patient. Hypertension is a 0 if the patient does not have hypertension and 1 if the patient does have hypertension. Heart disease is a 0 if the patient does not have heart disease and 1 if the patient does have heart disease. The ever-married column is categorized by either a “no” or “yes”. Work type is listed as either “children”, “Govt-job”, “Never_worked”, “Private”, or “self-employed”. Residence type is either “Rural” or “Urban”. Avg_glucose_level lists the average glucose level in the patients blood. BMI lists the patient's body mass index. Smoking_status is categorized as “formerly smoked”, “never smoked”, “smokes”, or “Unknown”. Stroke is a 0 if the patient did not have a stroke and 1 if the patient did have a stroke.

III. Research Questions

The overarching goal of this report was to discover what we can say about stroke occurrence using explanatory variables such as hypertension, heart disease, marital status, gender, residence type, smoking status, age, average bmi, average glucose levels, and job type.

Additionally, our team had many potential sub-research questions such as:

1. Is there a stroke occurrence difference in age?
2. Which age group has the most cases of hypertension, heart diseases or stroke?
3. Does having hypertension or heart disease increase the chance of having a stroke?
4. Will there be a difference in stroke occurrence for people with different smoking status, marriage type or job type?
5. Do people who have strokes have different glucose levels and body mass index compared to people who have no strokes?

IV. Approach

To begin our exploration of our research questions outlined above our team first imported a Kaggle Data set which contained 12 attributes and 5110 observations into the Tableau. We then set the dimension we were interested in into the columns. If the variable was numerical like age we then also create a binning variable. We then set the sum of strokes and the average of strokes in our row field so that we could see the difference in different groups. Our attribute “sum of strokes” represents the stroke occurrence and “average of stroke” shows the chance or probability of getting a stroke in a particular group. Using these attributes we then tried visualizing a plethora of different graphs in tableau to see which graphs best fit our data and to see which visualizations could provide us with interesting information in an easy to read format.

V. Results and plots

The first question we were interested in exploring was: Which age group has the highest chance of getting a stroke? To answer this question we constructed Figure 1, shown below.

Figure 1 shows that Probability of stroke rapidly increases for women after the age of 60 (~200% jump) from 2.75% to 19.27%. It also shows how the probability of stroke rapidly increases for men after the age of 70 (~100% jump) from 10.42% to 24.66%. We can then see a clear upgoing trend of chance of stroke as age increases, for both male and female.

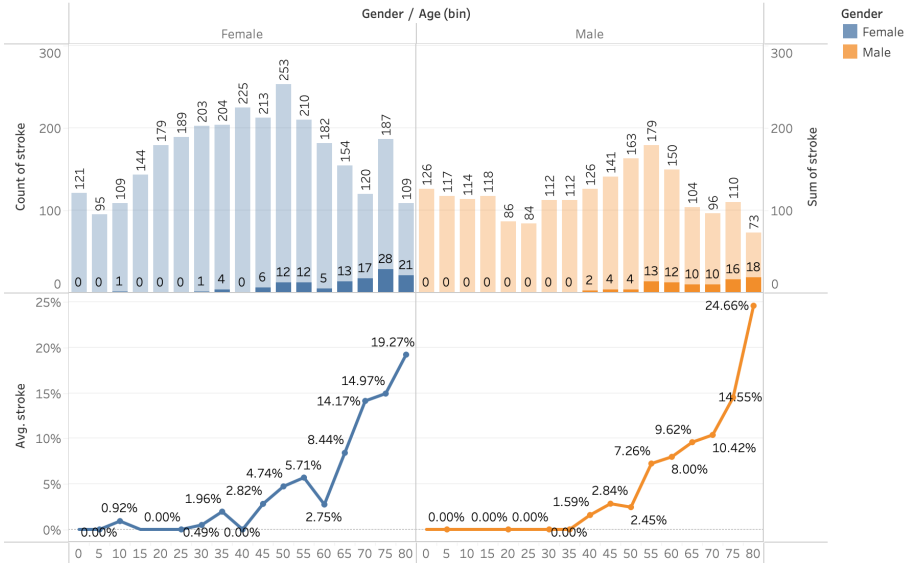

Figure 1: Gender, Age vs Stroke Occurrence

The second question we were interested in exploring was: Which age group has the highest occurrence of hypertension, heart disease and stroke?

In Figure 2 and Figure 3 below, we can observe that 75 is the most common age for getting Hypertension and Stroke. This data is consistent with what we would

assume to be true as an individual's likelihood of getting disease increases as they get older. We were also able to observe that 75 is the most common age for getting Heart Disease and Stroke. Additionally, comparing Figure 2 and Figure 3 side by side we can see that both Hypertension vs Age and Heart Disease vs Age follow a very similar distribution. This is again consistent with what we would assume to be true .

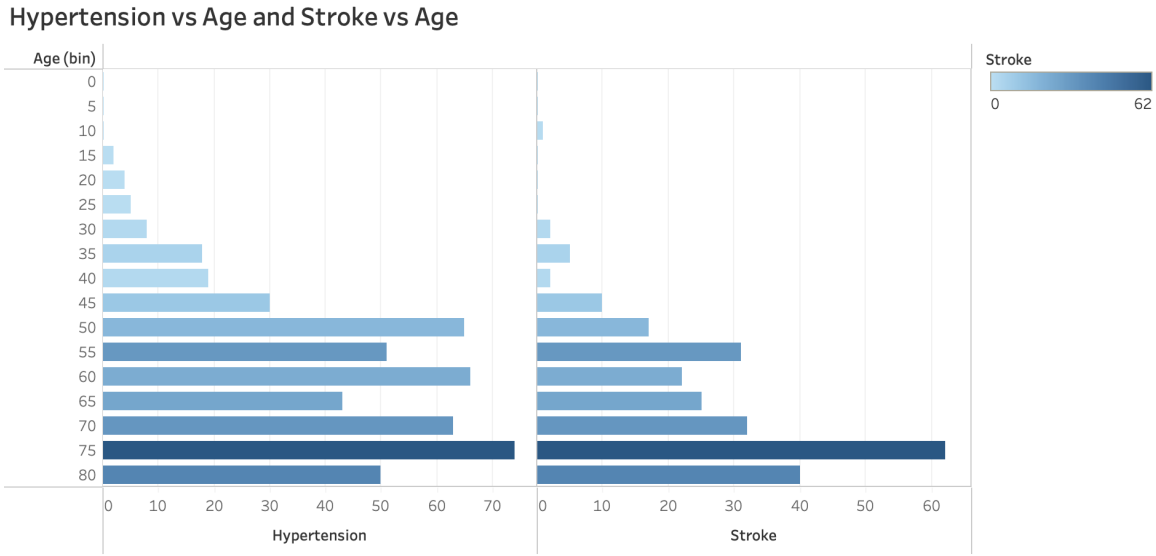

Figure 2: Hypertension, stroke vs age

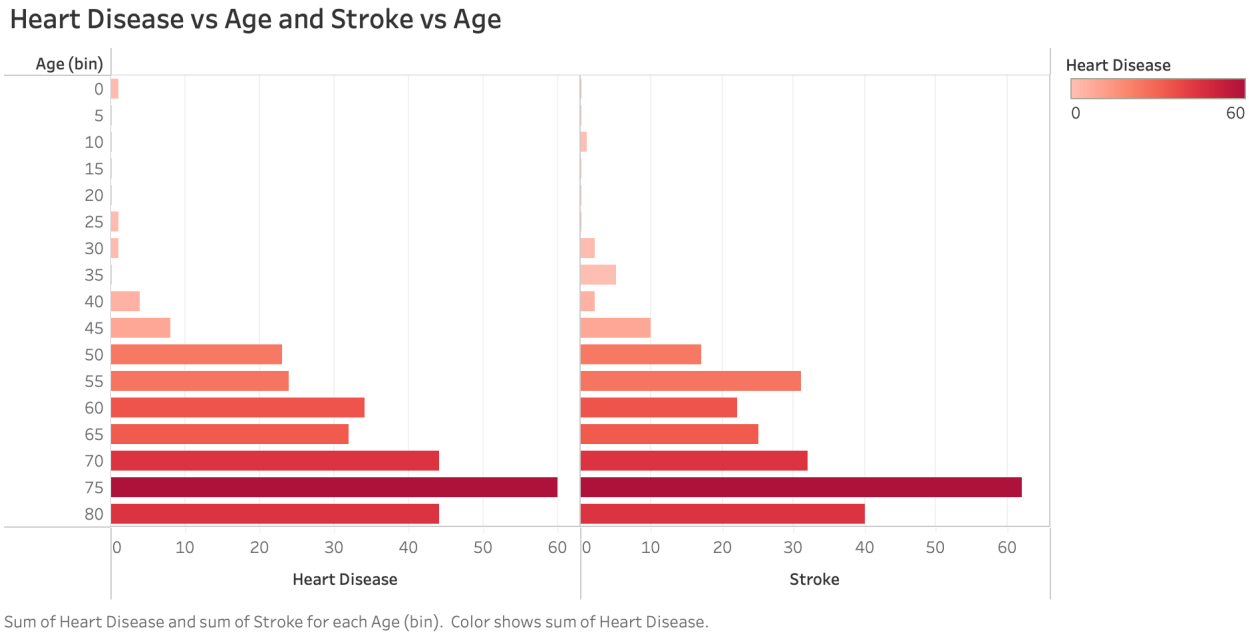

Figure 3: Heart Disease, stroke vs age

Another question we were interested in was: Does having hypertension or heart disease increase one's chance of having a stroke?

As Figure 4 shows below, there is a 350% increase in average stroke if heart disease is present (from 3.62% to 16.46%). Moreover, the probability of stroke quadruples if the patient has hypertension (3.34% to 13.30%). In both cases, there is a significant difference in the chance of having a stroke given that a patient had heart disease or hypertension in the past.

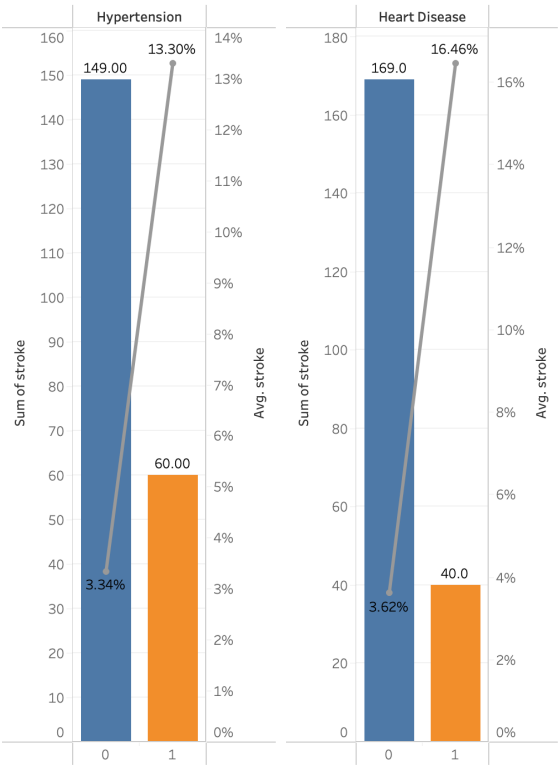

Figure 4: Past Disease and Stroke

The next question we wanted to address was: Is smoking status a predictive variable for strokes?

Figure 5 shows that females who never smoked are the largest group who have the highest occurrence of stroke, but that might be due to the fact that we have a big proportion of females who never smoked in our sample. Males who formerly smoked had the highest stroke percentage. It was interesting to find that the average stroke was less in the "females who smokes" group compared to the "females who never smoked" group. However, discarding gender, the average stroke is higher for people who formerly smoked compared to people who never smoked. Therefore, we can conclude that smoking status does have an effect on the chance of getting a stroke.

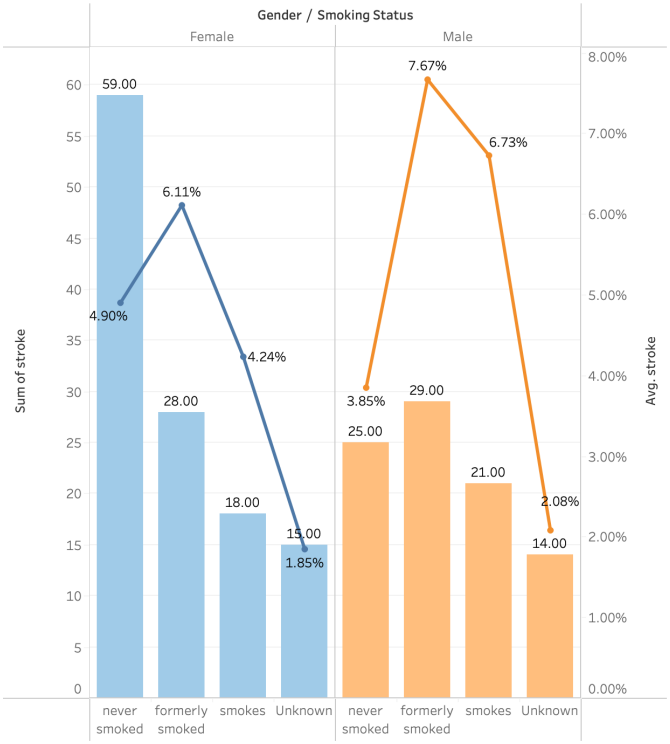

Figure 5: Smoking Status and Gender vs Stoke

Another question we strived to answer was: What is the difference in the chance of getting a stroke for married and unmarried groups, and do these results differ by gender?

In Figure 6, the shallow bar represents the count of the group, while the deep bar represents cases of strokes in that group. We observed a significant increase in chance of stroke for both females and males after marriage, 5.38% for female and 6.45% for male if they are married.

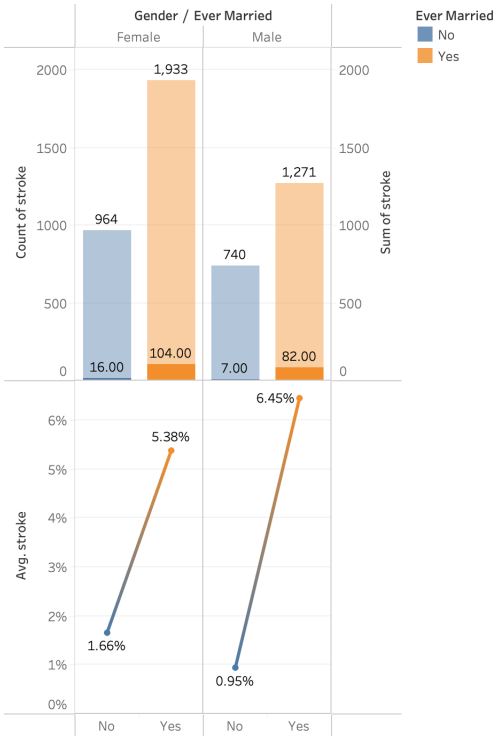

Figure 6: Marriage and Gender vs Stroke

Another question we wanted to investigate was: Do occurrences of strokes differ in groups with different job types?

Figure 7, shown below, shows that there does appear to be a relationship between job type and stroke occurrence. More specifically, we can see that those who do not work (represented by the light blue small circle in the center of the figure) or have government jobs experienced less stroke occurrences compared to those who worked in the private sector or were self-employed. However, in terms of chance of having a stroke, the self-employed job type has the highest chance, as the graph shows that the color of self employed is deepest. This was an interesting finding for our team. Our team hypothesized that this may have something to do with job related stress increasing the likelihood of one's chance of having a stroke. However, to confirm this an additional data visualization, as was performed in Figure 2 and Figure 3, will have to be performed using Stress and Stroke Occurrence.

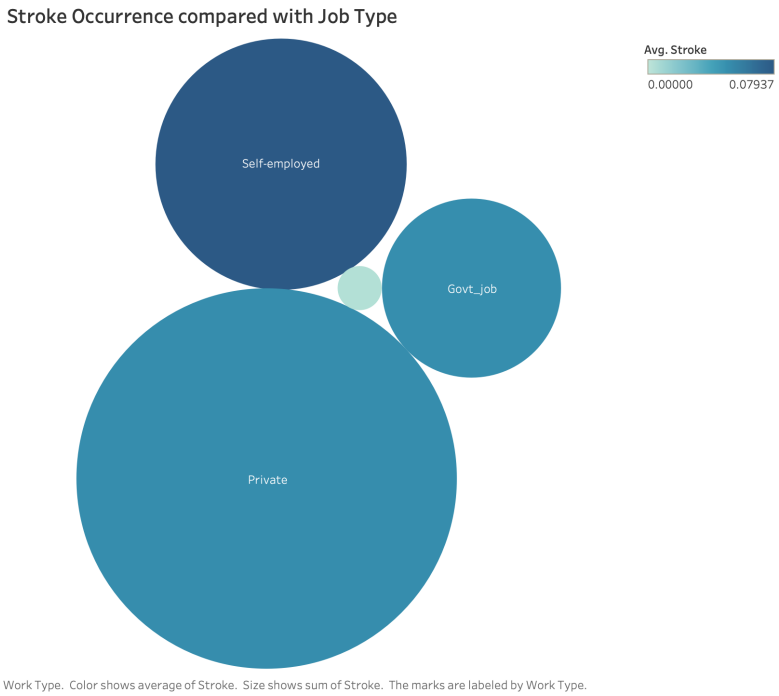

Figure 7: Stroke Occurrence compared with Job Type

Next, our team wanted to look into the following: Do people who have strokes have different glucose levels and body mass index compared to people who have no strokes?

We can clearly see from Figure 8 that BMI was not a key contributing factor to the occurrence of a stroke. Average BMI was slightly higher for cases where a stroke had occurred, but not significantly enough to consider it a key contributing factor(5.4%). We used this to determine that we would not be focusing on BMI to predict whether a stroke occurs or not.

In comparison to BMI, we can see Average Glucose Levels had a much higher correlation with the occurrence of a stroke. From Figure 8 below, we can see that the average glucose levels in patients where strokes had occurred were 22.7% higher compared to patients where no strokes had occurred.

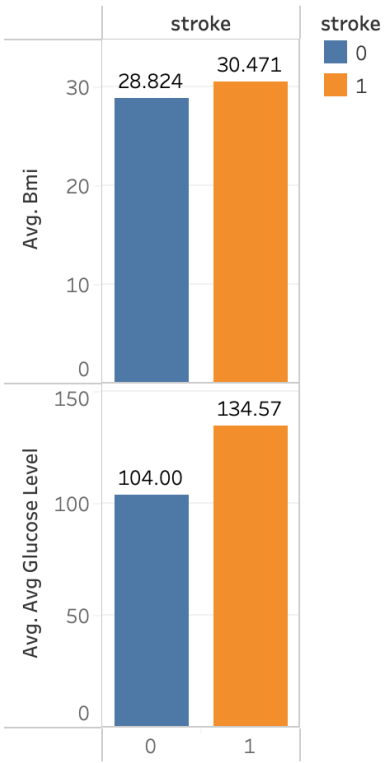

Figure 8: Stroke type related to Avg. BMI and Avg. glucose level

VI. Dashboard

The following dashboard contains our major key findings which we found to be pivotal in addressing our research questions. Our goal with this dashboard was to create one space where we could easily summarize our findings and present our most relevant data.

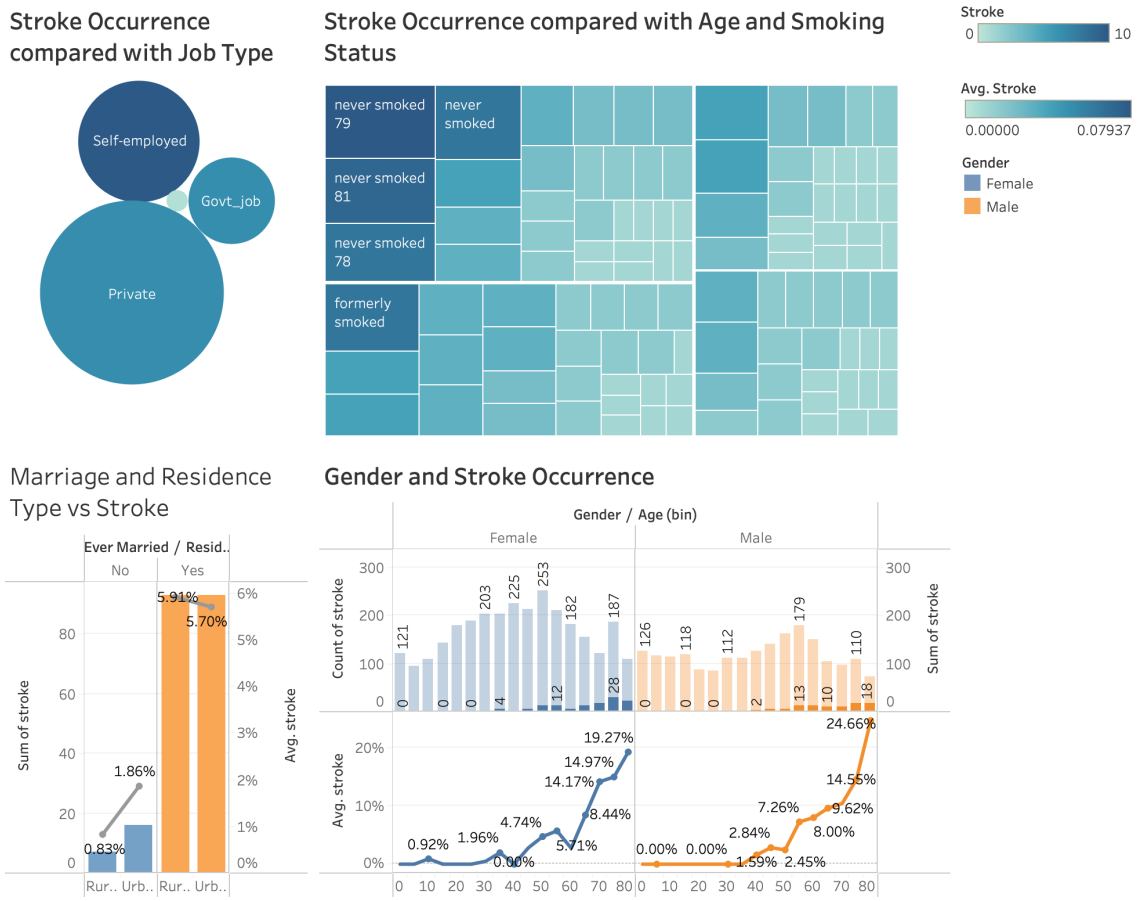

Figure 9: Final Dashboard

VII. Takeaways and Summary

One key takeaway from our data visualizations was that heart disease, hypertension and stroke are highly correlated. This was something that we expected given our prior knowledge of how the chance of getting disease increases with age. Additionally, we found that irrespective of gender, age plays a predominant role in one's likelihood of getting a stroke following a similar pattern of hypertension & stroke. Groups with age higher than 70 face a higher risk of strokes. In terms of average stroke, males who formerly smoked have the highest percentage of strokes. Generally, people who formerly smoked had a higher chance of getting strokes. We also observed a higher occurrence of stroke in non-smoking women, however this could also be due to a large sample of non-smoking women who happened to get a stroke. We also determined that gender is just a classification but not a key distinguishing factor when it comes to stroke detection. Another major takeaway was that people who have a self-employed job or work in the private sector have the highest chance of getting a stroke. Additionally, a significant increase of probability of stroke is observed in married persons, both females and males. Another takeaway was that high average glucose level is one of the possible symptoms of future stroke and can be used to foresee the occurrence of stroke, while body mass index doesn't show a significant increase in the group who has stroke compared to the group who has no stroke.