October 22nd, 2024

Stroke Dataset Analysis

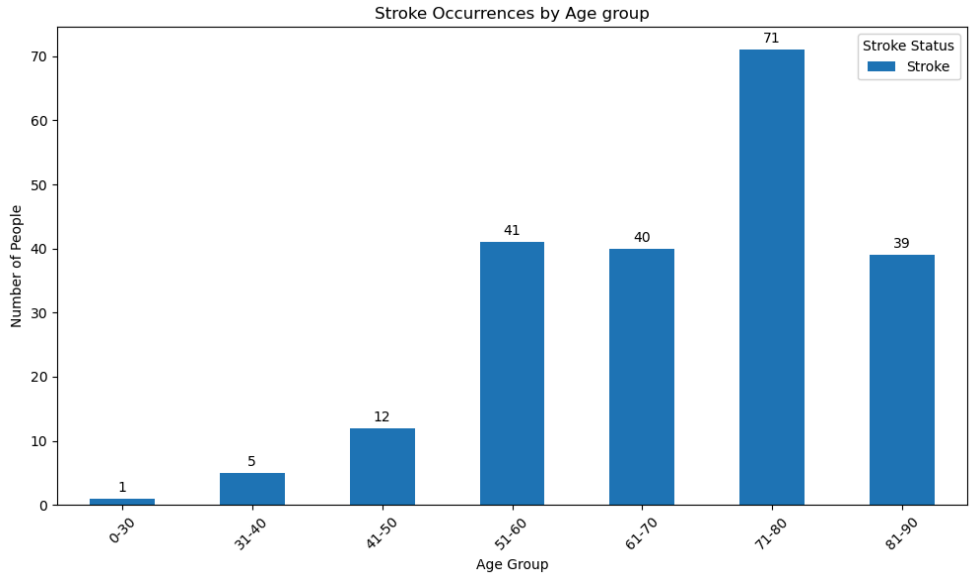
Group 1: Bhavesh Heetoo , Jana Khamis , Jiang Jun , Tala Zubi , and Paul Schaefer

Analysis:

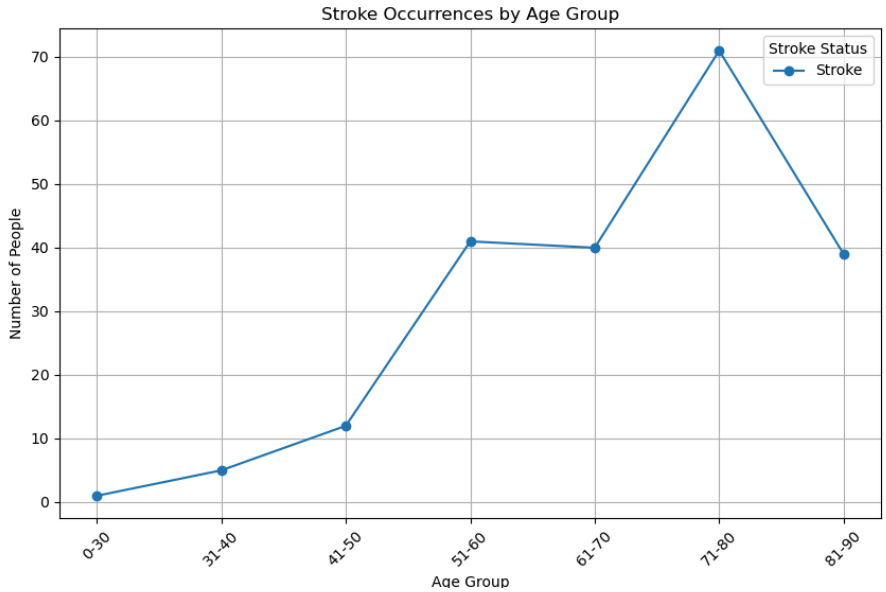
The objective of this project is to discern and investigate the relationships between various health factors and stroke incidence. The dataset utilized included a plethora of factors such as smoking status, demographic factors (age, gender, work type and residence), hypertension, and heart disease. All of the following factors have differing implications resulting in stroke risk, and we aimed to uncover these underlying connections and accurately analyze them.

Does a higher age increase the risk of stroke?

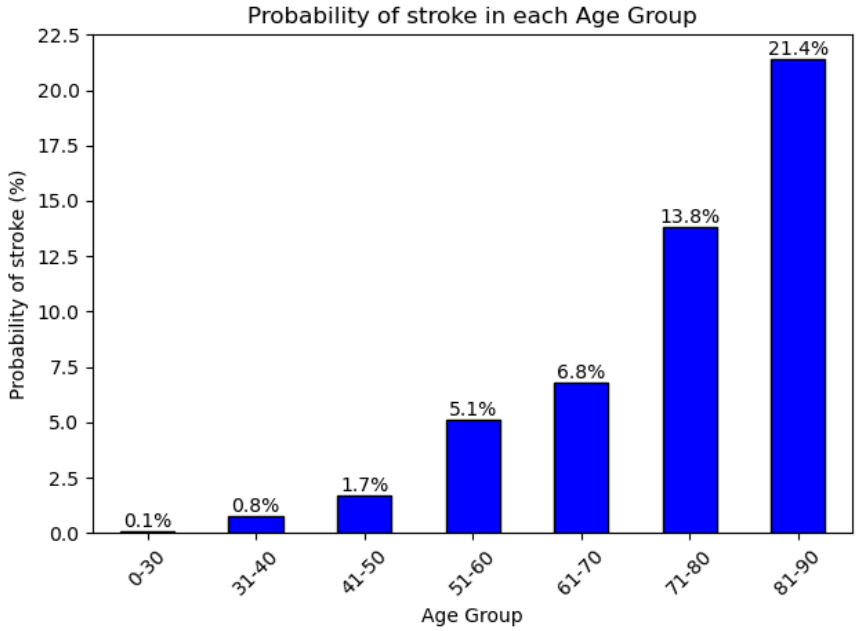
The below bar chart displays the stroke occurrence between varying age groups. From age 51-60 we see an increase of 29 stroke occurrences when compared to the age group of 41-50. There is a gradual increase in stroke occurrences with a high of 71 in ages 71-80. This graph visually confirms higher age does increase the risk of stroke occurrence.



A line graph that displays the stroke occurrences distributed by age group. Stroke occurrences seem to increase significantly as people age, There is a sharp rise in strokes for people in the 51-60 age. The highest number of strokes occurs in the 71-80 age group. This suggests that older adults, particularly those in their 70s face the greatest risk.



The bar chart above illustrates the probability of stroke across different age groups in percentages. We see a low probability in younger age groups and a gradual increase to middle age with a substantial rise in older age groups. From age 61 and onwards the probability begins to rise sharply with much higher risks seen in age groups 71-80 and 81-90. The highest stroke probability is found in the 81-90 age group but it is important to note the population in this age group is lower compared to others.



The visualizations created above collectively highlight a clear trend of increasing stroke occurrences and probabilities with age. A bar chart shows a steady rise in stroke occurrences, with a sharp increase of 29 more strokes from ages 41-50 to 51-60, peaking at 71 strokes in the 71-80 age group. Similarly, a line graph confirms that stroke cases significantly rise after age 50, with the highest number of occurrences in the 71-80 group, indicating older adults, especially those in their 70s, face the greatest risk. A second bar chart illustrating stroke probability by age shows a low risk in younger groups, followed by a sharp rise after age 61. The highest probability is found in the 81-90 group, though this group has a smaller population. Overall, these visuals emphasize that stroke risk increases significantly with age, particularly for those over 60.

Hypertension and heart disease impact on stroke occurrence?

We are interested in how hypertension and heart disease impact on stroke occurrence in our dataset. In the dataset we checked individual's health status based on hypertension and heart disease and categorized them into four types: healthy individuals who don't; individuals who only have hypertension; individuals who only have heart disease; individuals who have both hypertension and heart disease.

hypertension	**heart_disease**	**Count**
0	0	4273
1	0	393
0	1	185
1	1	58

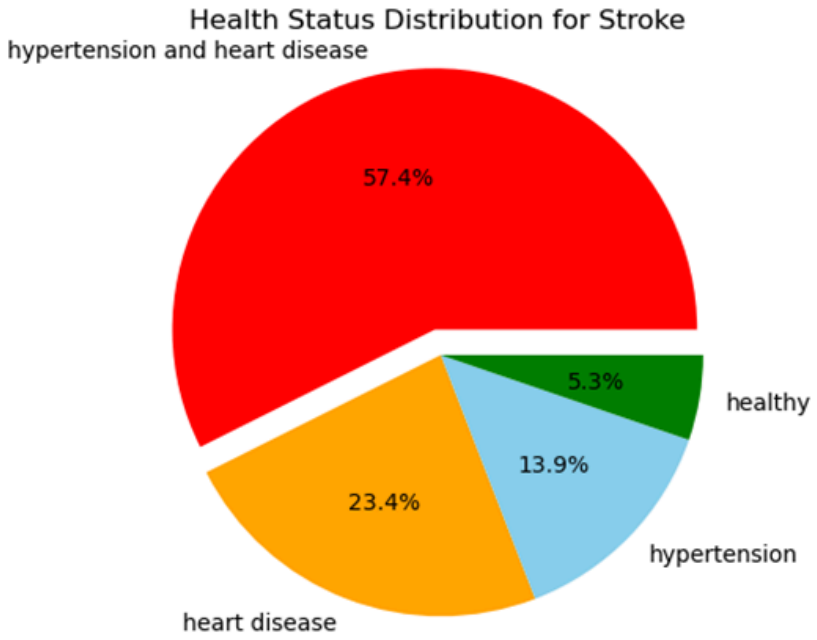
We created a pie chart to visualize the health status distribution based on the conditions of hypertension and heart disease. From the graph 87% individuals are healthy, 8% have hypertension and 3.8% have heart disease and 1.2% individuals have both diseases.

Taking a closer look at the dataset, we created a summary table to only count the stroke and no stroke for each health status.

Health Status	**Stroke**	**No Stroke**
Healthy	120	4153

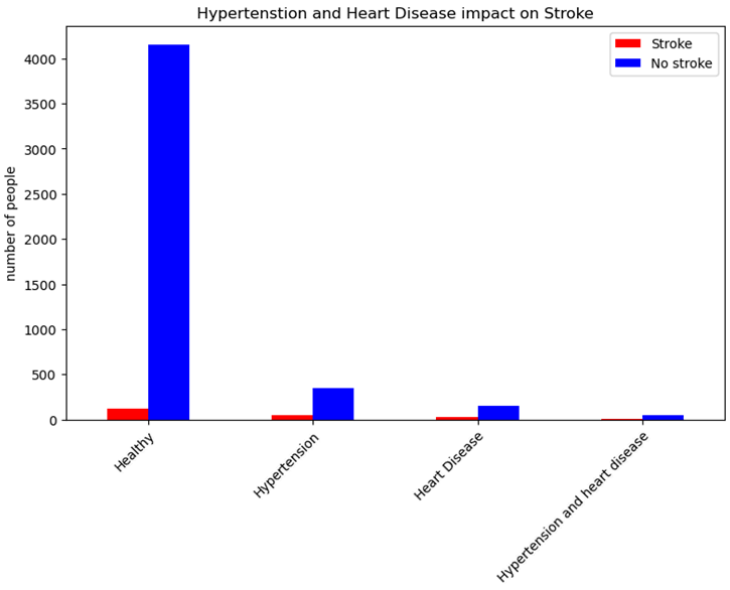
Hypertension	49	344
Heart Disease	29	156
Hypertension and heart disease	11	47

Based on the above summary table we create a pie chart to study health status for individuals who had a stroke. From the below chart, we can see more than half of people who had a stroke also suffered hypertension and heart disease (57.4% of stroke population), and only 5.3% of individuals who had stroke are healthy.

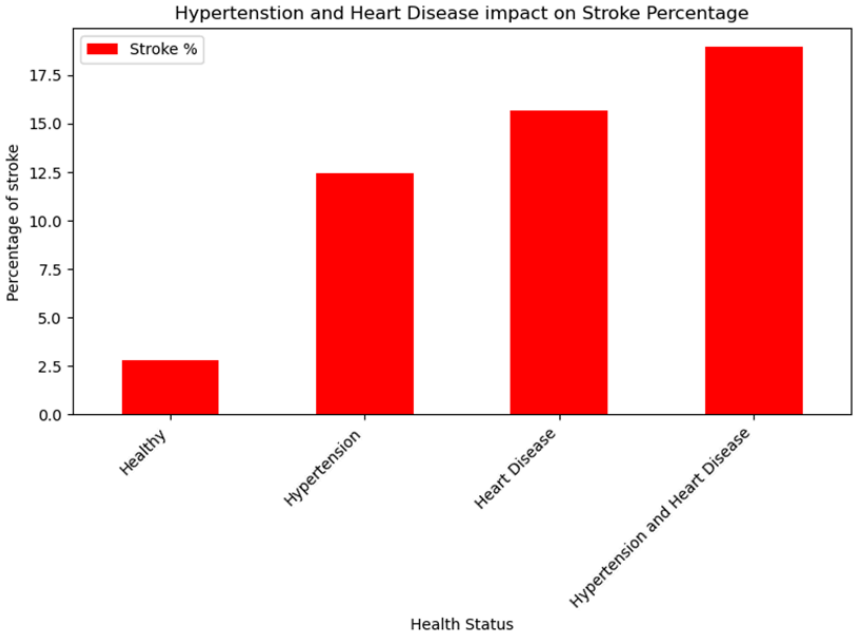


Furthermore, we wanted to see the stroke status in each health status. We tried to create a bar chart counting the number of stroke and non-stroke individuals for each health status. However, the size of each health status group was not equal, and the healthy group was much larger than the other groups. From the bar chart below, it is difficult to compare among groups. Instead, we created a bar chart based on the stroke percentage for each group.

A bar chart to visualize the impact of hypertension and heart disease on stroke.



A bar chart that shows percentage of stroke for the four health types

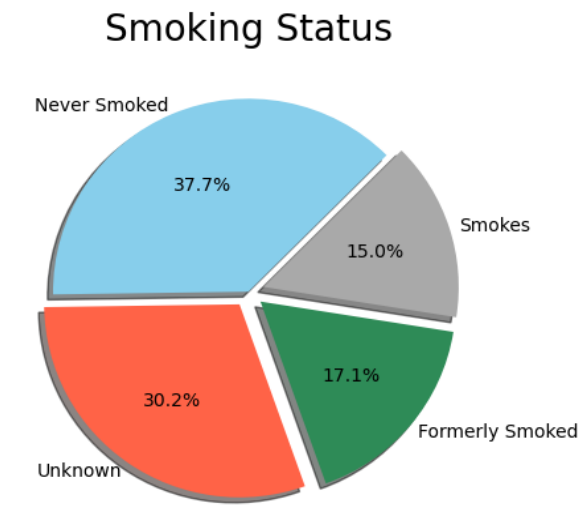


The stroke percentage in Hypertension and Heart Disease is the highest group (18.96%) followed by heart disease (15.66%). In the healthy group, only 2.8% of individuals had a stroke.

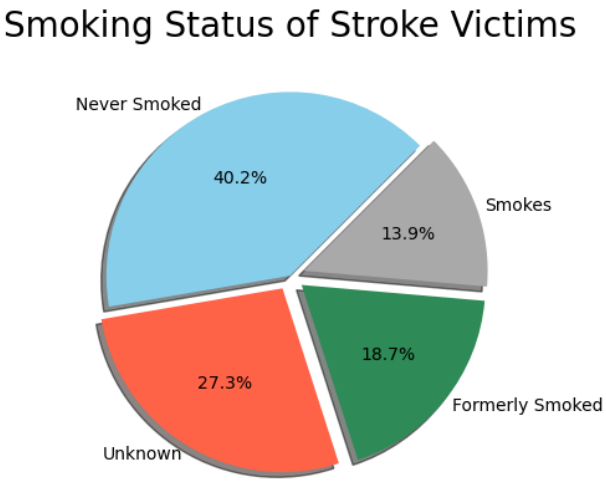
Based on the above pie charts and bar charts, chances of a stroke for healthy individuals is much lower than individuals who have either heart disease or hypertension. The chance for individuals who suffer both hypertension and heart disease to have a stroke is high.

Does smoking have an impact on the likelihood of suffering a stroke?

The smoking column had four values. Currently smoking, formerly smoking, never smoked, smoking status known. A pie chart was created to see how they spread out.

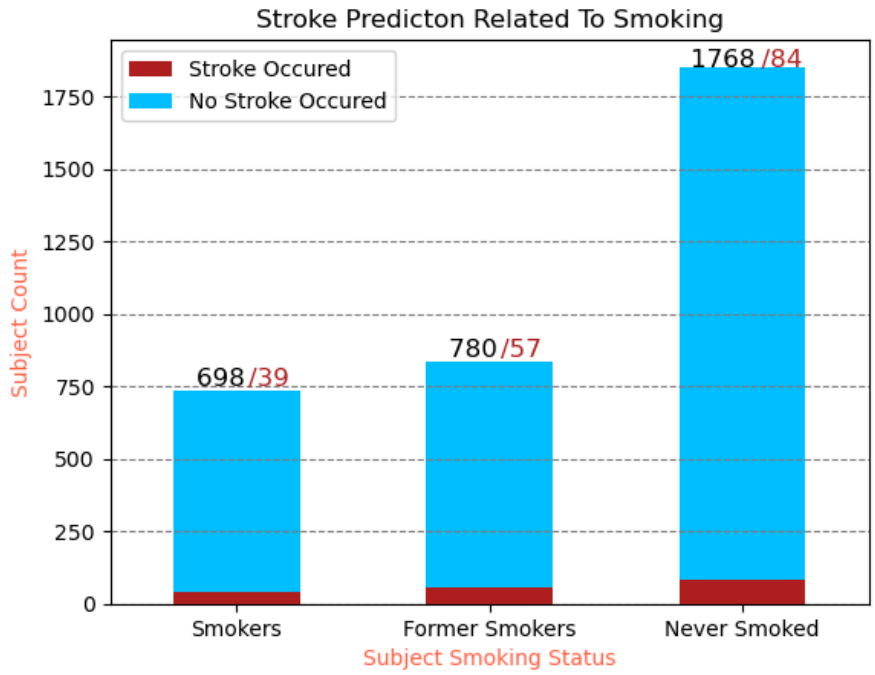


With the unbalanced distribution it was then broken down into just subjects that suffered a stroke. A new pie chart of this data was created. 37% nonsmokers, next largest is Unknown 30% with only 17% and 15% for former smokers and smoking. Most of the database either don't smoke or won't say, throwing off the database populations.

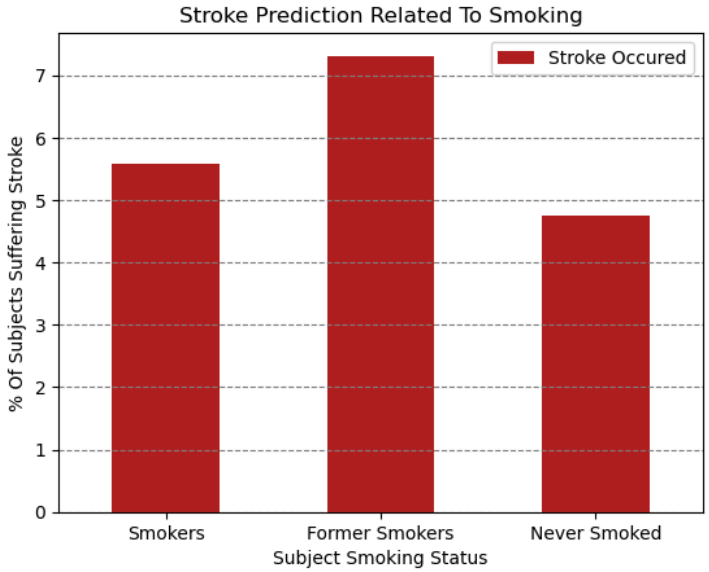


This pie chart revealed that percentages were nearly identical. It can be determined the Unknown value wasn't necessary as it provided no real value and was dropped, and still at 27% it took up a

large part of the subject count. The next step was to see how many people were included in the three remaining groups, smokers, former smokers, and nonsmokers.



The nonsmoker population was so much larger, the data could only be insightful if it was broken into a percentage for comparison, resulting in an additional bar chart. 1852 total for nonsmokers with only 737 for smokers and 837 former smokers. Beyond that, only 180 subjects suffered a stroke, so the percentage was necessary.



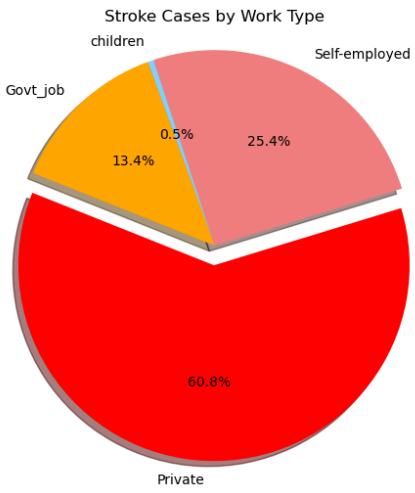
The percentage comparison reveals smoking status had minimal to no impact on suffering a stroke. Only 5.59% of smokers suffered a stroke, but that also lines up with only 4.75% of non smokers and 7.31 of former smokers.

With those percentage values the p value could be calculated, 0.760, meaning it is higher than 0.05, making the data statistically insignificant.

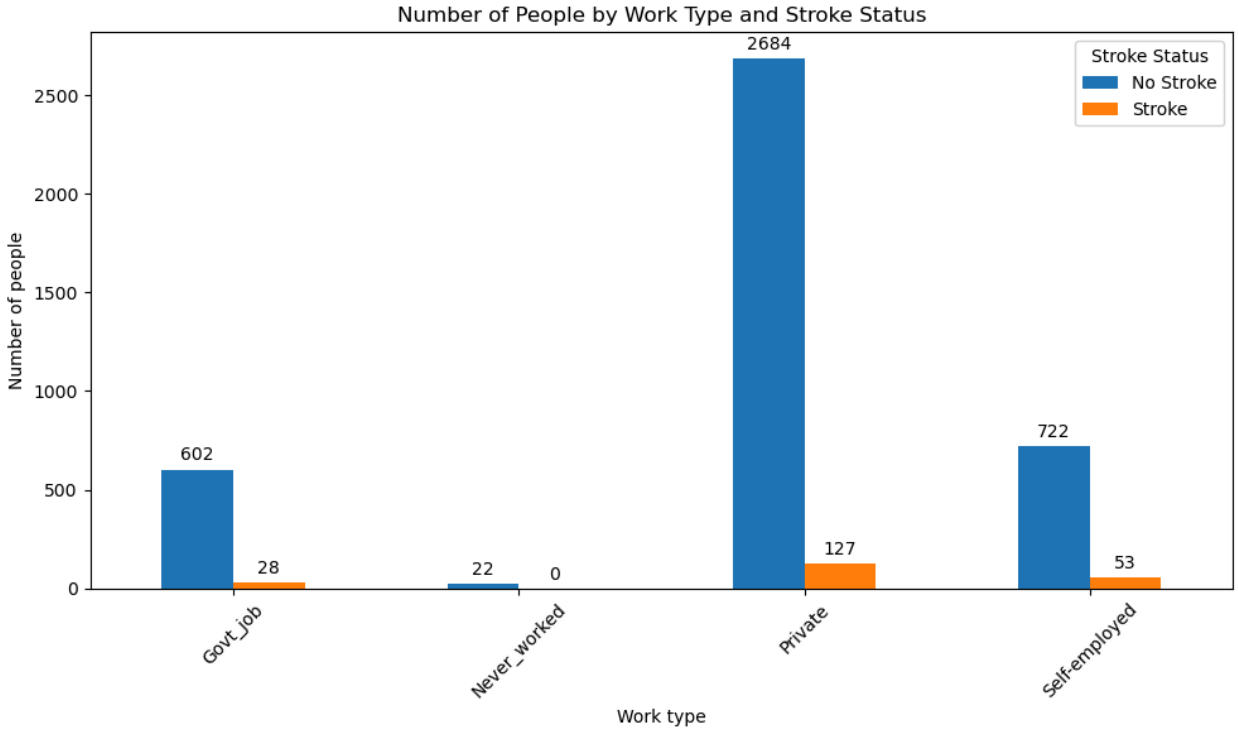
Conclusion: The Null Hypothesis, smoking status has no impact on the likelihood of a stroke.

How do different types of work (Private, Self-employed, Govt_job, Never _worked) affect stroke risk?

A pie chart to distinctly analyze the different work types and their relation to stroke. The chart suggests that the private sector jobs are more susceptible to stroke, followed by self employed jobs. The individuals with private jobs had a 60.8% possibility of stroke occurrence. The next most prominent relationship is those who are self employed with a 25.4% possibility, followed by government jobs with 13.4% and finally 0.5% in children. The results illustrate that there are nuanced factors in each job type that contribute to stroke risk. For example, it is possible that the work environment is extremely stressful in the private sector, or in a self-employed job, thus leading to higher stroke risk. Other factors that can be considered are elements like quality of healthcare provided by a job, it is possible that government jobs provide benefits that alleviate stroke implications which improves overall health. All these factors come into play in deciding which job has a greater contribution to stroke risk. The use of a pie chart is very suitable for this type of data, as it allows the reader to visually discern and comprehend the severity of the data and its difference in stroke risk. The pie chart eloquently displays the stroke occurrence allocated with each work type and allows us to note a correlation in their relationship. One can infer that the higher stress associated with a job, the more likely a stroke is to occur. This indicates that more research could be conducted for further causal relationships to be seen and that this avenue must be explored to create new public health efforts.



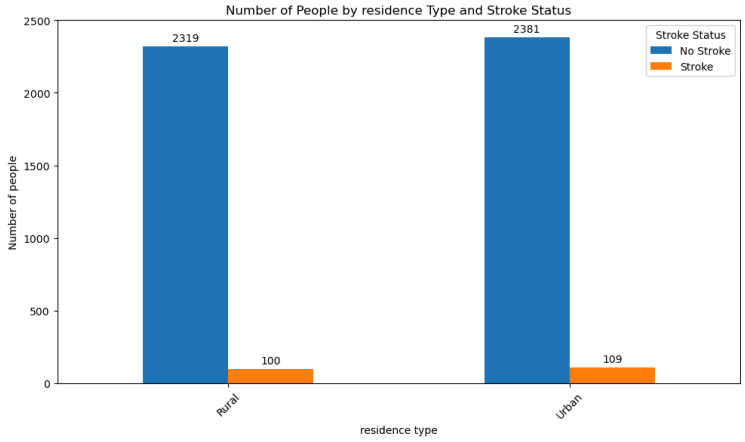
A bar graph shows the number of people in different work types, categorized by stroke status. For government jobs, only 28 out of 628 had a stroke suggesting a relatively low stroke occurrence among government employees but strokes still occur. For never worked, none experienced a stroke, this could be due to the limited sample size of 22 for this group. For the private sector, 127 out of 2648 people had experienced a stroke, this group is the largest population but still indicates that it has a notable stroke occurrence. For the self-employed, 53 out of 775 people have had a stroke suggesting that self-employed individuals face unique risk factors related to stroke when compared to other work types.



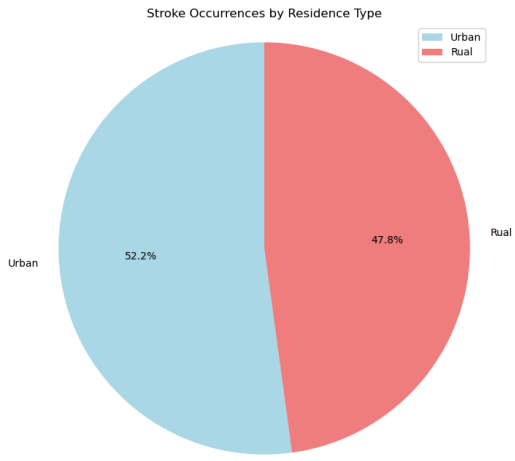
Does living in an urban or rural area affect stroke likelihood?

A bar chart to successfully compare the number of people by residence type and likelihood of stroke was used. The data illustrates that there is a difference, though minimal, between rural or urban residences in stroke occurrence. The bar graph shows that those living in urban areas had 109 stroke cases, whilst those living in rural areas only had 100 stroke cases. The graph depicts a small, but meaningful difference in the context of residence type and possibility of stroke occurrence. There are a plethora of confounding factors that could contribute to these results, for example, if people living in urban areas are more prone to greenhouse gas emissions, or if they belong to a certain age group, quality of life, or stressful living conditions, all of which could lead to stroke risk factors. The research question is imperative in deciding whether a resident type or location could impact health, however, the results displayed are not significant enough to create a definitive conclusion. While the results may be valid, more research must be conducted with a greater sample size, and better control of external variables, to avoid skewing the data.

The bar chart is a good choice for this type of data to be displayed, as we can visually comprehend the difference between rural and urban resident types and how the difference is small. More research must be conducted, to allow for better public policy efforts in improving the environment, thus improving our health and quality of life.



The Pie chart displays the stroke occurrence between people living in urban and rural areas. Urban residence types have a slightly higher proportion of stroke occurrences at 52.2% compared to rural residence types at 47.8%. The close distribution suggests that the occurrence of strokes is relatively balanced between urban and rural residence types. Potential reasons for the slight difference could be varying lifestyle factors, access to healthcare, and environmental conditions that are not captured in the dataset.

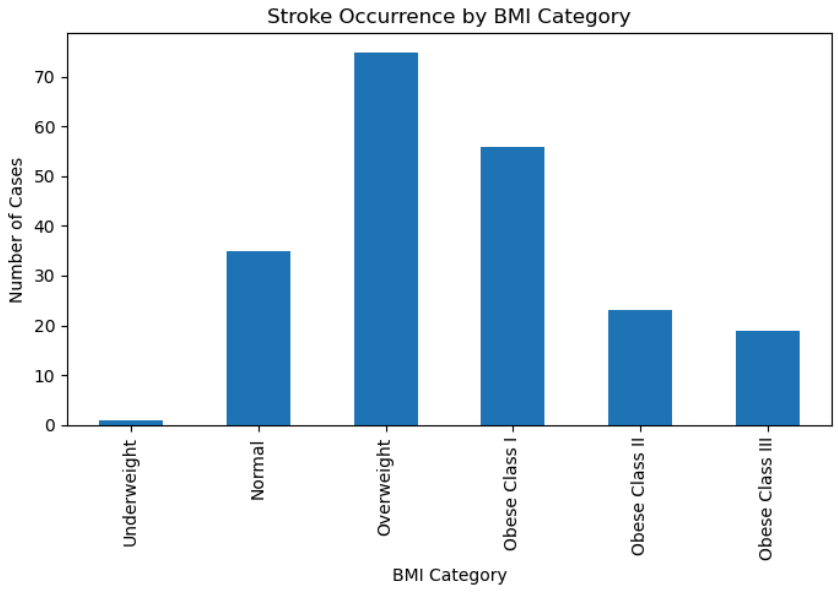


The Pie chart displays the stroke occurrence between people living in urban and rural areas. Urban residence types have a slightly higher proportion of stroke occurrences at 52.2%

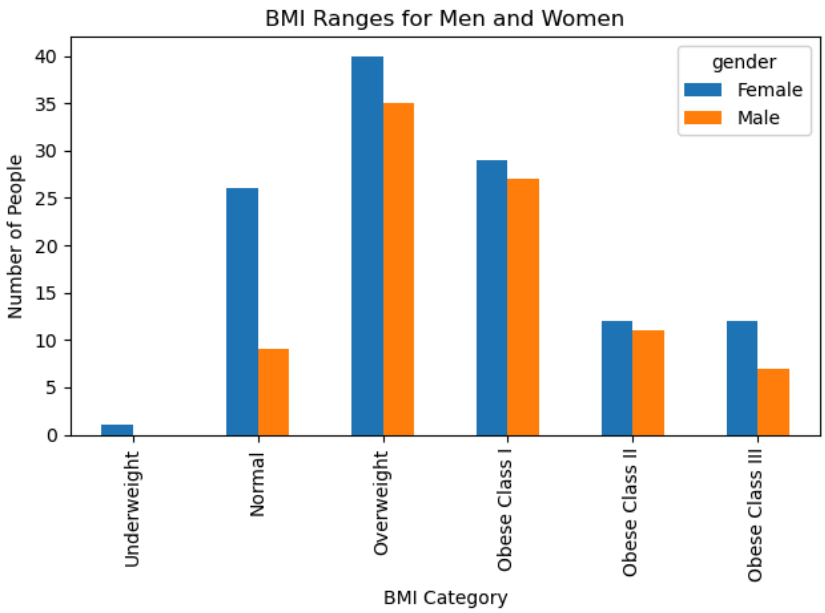
compared to rural residence types at 47.8%. The close distribution suggests that the occurrence of strokes is relatively balanced between urban and rural residence types. Potential reasons for the slight difference could be varying lifestyle factors, access to healthcare, and environmental conditions that are not captured in the dataset.

What is the relationship between BMI and stroke occurrence?

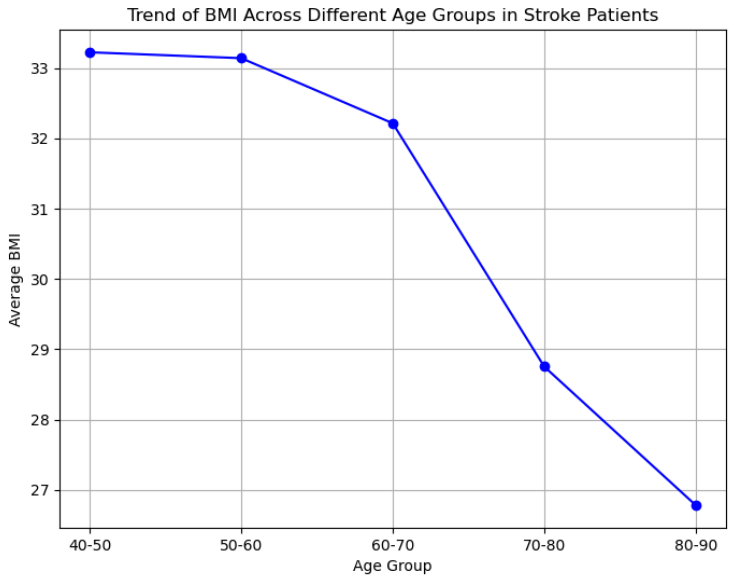
The first key finding is the association between BMI categories and stroke incidence. The statistics indicated that most stroke patients fall within the “Overweight” and “Obese Class I” BMI categories. There are 75 people in the “Overweight” BMI category group and there are 56 in the “Obese Class I” BMI category. This implies a higher BMI is a substantial risk factor for stroke, particularly in people who are “Overweight”. However, the "Obese Class II and III" show fewer stroke occurrences (23 and 19) than the "Overweight" which may seem counterintuitive. Overall, it seems that being overweight, or the early stage of obesity might increase the chance of having a stroke. However, it's essential to keep in mind that weight is only one factor, and age or lifestyle can also influence your risk of stroke.



To take a closer look at the dataset did a comparison of the BMI ranges of men and women, which showed that women are at a higher risk. Women dominate both the “Overweight” and “Obese Class I” BMI categories, with 40 women categorized as overweight compared to 35 men. However, the “Normal” BMI category shows a huge difference between men and women, where there are 26 women who are normal and 9 men. This distinction emphasizes the need for gender specific health initiatives targeted at regulating BMI and lower stroke risk, particularly in women, who are more prone to fall into higher BMI categories.

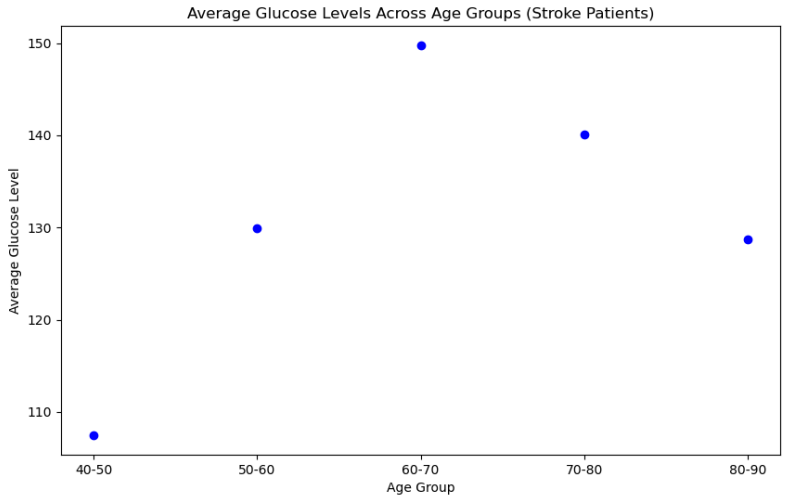


An important trend discovered in the data is decrease in BMI with age among stroke patients. A line chart of average BMI across groups demonstrates that younger stroke patients (aged 40-50) had a higher BMIs (around 33) than those aged 80-90, whose BMI falls to less than 27. This data implies that greater BMI in young people may lead to an increased risk of stroke. It also suggests that early intervention to manage BMI in younger people may have a major influence on stroke prevention in this population. Therefore, this suggests that variables such as health related changes or the BMI (may be in the underweight or normal categories) in older age may lead to reduced BMI in stroke patients.

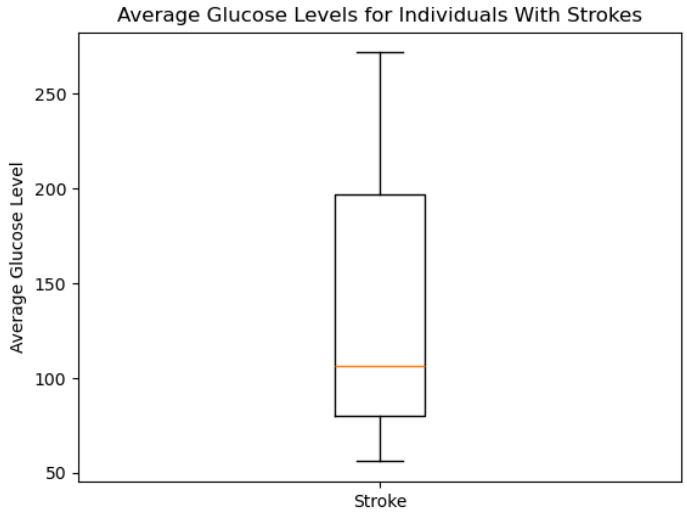


How do glucose levels vary across patients with stroke?

The examination of glucose levels across age groups provides additional support for these fundings. The scatter figure shows that average glucose levels in stroke patients increase with age, from around 129 mg/dl in the 50-60 age group to roughly 140 mg/dl in the 70-80 age group. This increase in glucose levels is a significant risk factor for stroke, especially in older patients, because increased glucose is usually associated with diabetes and other metabolic disorders. As a result, glucose monitoring and treatment are critical for lowering stroke risk in elderly population.



The box plot representing average glucose levels in stroke patients shows substantial diversity in glucose levels across this population. The median glucose level is slightly above 100, meaning that half of the stroke patients have glucose levels below this threshold and the other half have higher levels. The interquartile range (IQR), which ranges from above 100 to 200, indicates that 50% of stroke patients keep their glucose levels within this range. However, the whiskers range from about 60 to 250, demonstrating the wide range of glucose levels in stroke patients. This range demonstrates significant variation in glucose control, with some people having relatively low levels and others having excessive glucose levels. The data show a possible link between increased glucose levels and stroke incidence, as a considerable proportion of patients have raised glucose levels. The picture clearly supports the concept that glucose plays a significant role in stroke outcomes, and more statistical analysis might investigate whether there are specific glucose thresholds that increase the risk of stroke, understanding these trends may help improve the glucose control regimen for stroke prevention and therapy.



Finally, this study emphasizes the relevance of BMI and glucose level control in lowering stroke risk. The findings, backed by visualization such as bar charts, line charts, and box plots, give compelling evidence that public health initiatives should prioritize weight management, gender, specific treatments, and glucose control as major stroke prevention methods. Early treatments aimed at younger persons with higher BMIs, as well as continuous glucose monitoring in older adults, have the potential to significantly reduce the incidence of strokes, resulting in better overall health outcomes.

Conclusion and Limitations:

In conclusion topics like these are paramount for further public health initiatives and this research is necessary to improve overall health and well being. Developments in computer science and data analytics are essential in improving our understanding and technology in analyzing the status quo, uncovering and assessing trends, and creating innovative solutions. In terms of our specific studies, our most significant conclusions are that hypertension and heart disease, age and BMI are the most influential factors for stroke risk, therefore more efforts are necessary to alleviate this burden. A few notable limitations would be that this was an observational analysis, therefore in order to benefit from this data, more experimental studies must be conducted, and we must increase our sample size, and account for any confounding variables in future studies. Thank you for listening to our presentation.