Stroke Prediction Dataset

By: Justin Blackshear

Data Exploration

The stroke dataset used in this analysis comprises 12 key attributes from 5,110 patients, including one that indicates whether a stroke occurred. The dataset contains a mix of personal demographic information—such as gender, age, marital status, employment type, and residence type—alongside medically relevant data, including body mass index (BMI), average glucose levels, smoking status, hypertension, and heart disease history. To ensure data integrity, missing BMI values were replaced with the mean BMI of the dataset. Once cleaned, the dataset provides a comprehensive foundation for exploring potential correlations with stroke occurrence.

Objective

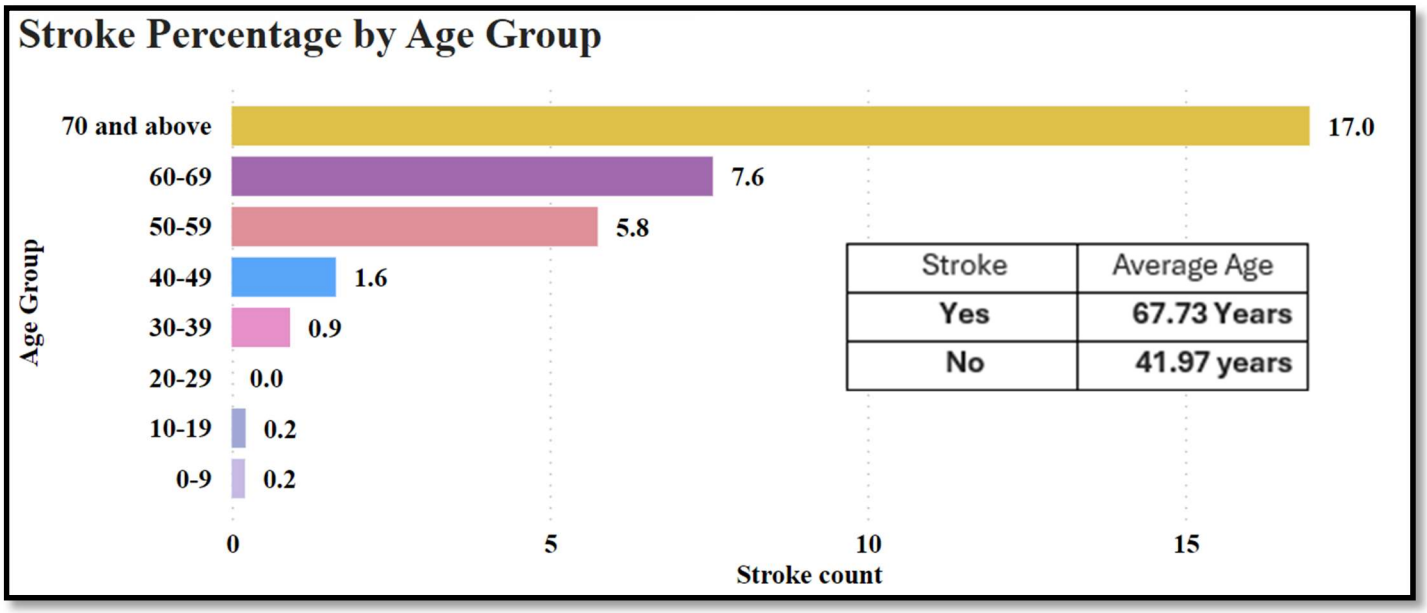
The objective of this report is to explore relationships within the dataset and identify factors that may contribute to stroke occurrence while assessing the dataset's suitability for predictive modeling.

Views were created in SQL to access and organize a subset of data to answer the following questions:

1) What is the average age of a patient who's had a stroke?
2) What is the distribution of ages for stroke victims?
3) What is the average BMI for somebody who has had a stroke?
4) How does age and BMI relate?
5) What is the stroke distribution by gender?
6) How does glucose level relate to stroke rate?
7) How does stroke rate compare between patients with heart disease and patients without heart disease?
8) How does stroke rate compare between patients with hypertension and patients without hypertension?
9) How does smoking affect stroke rate?

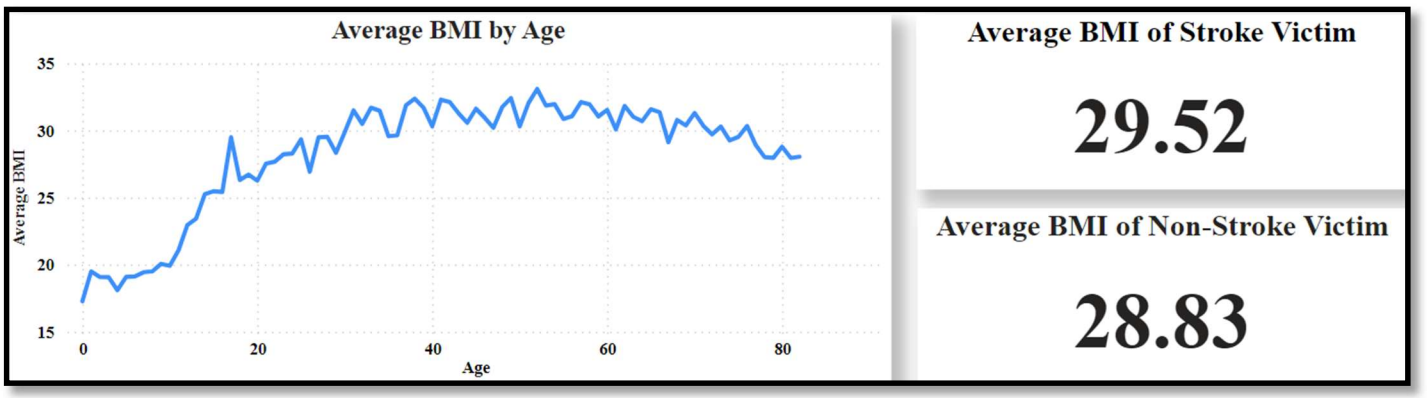
What is the average age and age distribution of patients who had a stroke?

- Patients were split into 10-year intervals (0-9, 10-19, 20-29, etc.), with the final group covering ages 70 and above. The sum of stroke occurrences was determined for each age group and then the percentage of stroke victims within each age group were found.
- As supported by the graph below, the average age of stroke victims was 67.73 years, compared to 41.97 years for non-stroke victims, further indicating an increased likelihood of stroke with age.



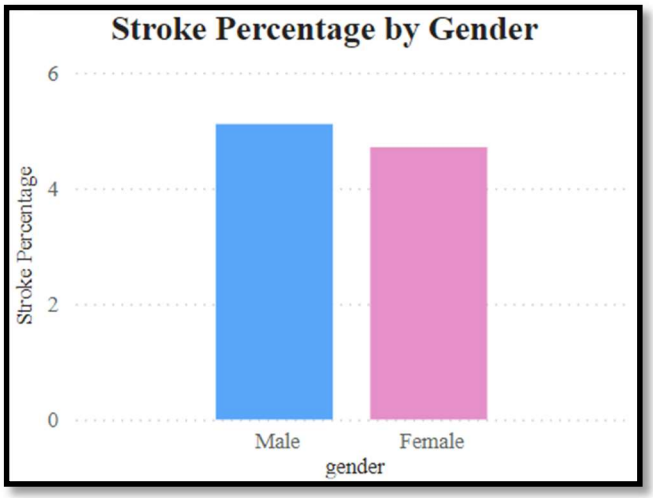
What is the average body mass index (BMI) for a stroke victim and how does BMI relate to age?

- Patients were split into 10-year intervals (0-9, 10-19, 20-29, etc.), with the final group covering ages 70 and above. The average BMI was determined for each age group and then the average BMI was found for both patients who suffered a stroke and patients who did not.
- As displayed in the graph below, the average BMI gradually increased with age until around 30, then remained steady until about 70, after which it began to steadily decline.
- The average BMI for patients who suffered a stroke and patients who did not suffer a stroke was 29.52 and 28.83 respectively. This indicates that BMI alone may not be a significant factor in determining stroke risk within the dataset.



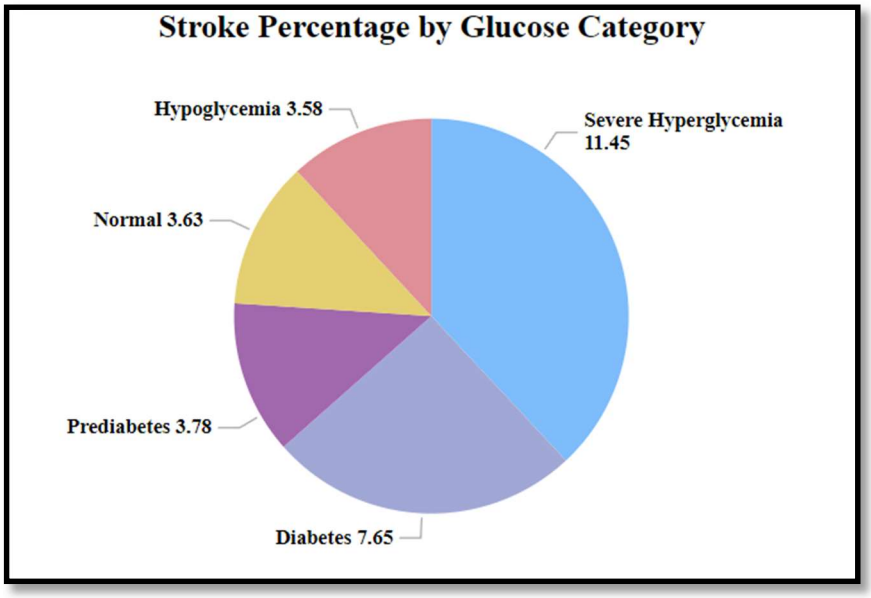
What is the stroke distribution by gender?

- Total patients within the dataset were separated by gender with 2215 total males and 2994 females. The dataset contained one value where a patient did not identify as either male or female, so that value was dropped as an outlier.
- Males suffered 108 strokes with a stroke percentage of 5.11% and females suffered 141 strokes with a stroke percentage of 4.71%.
- This indicates a minor correlation between gender and stroke occurrence as men were 8.5% more likely to suffer a stroke within the dataset.



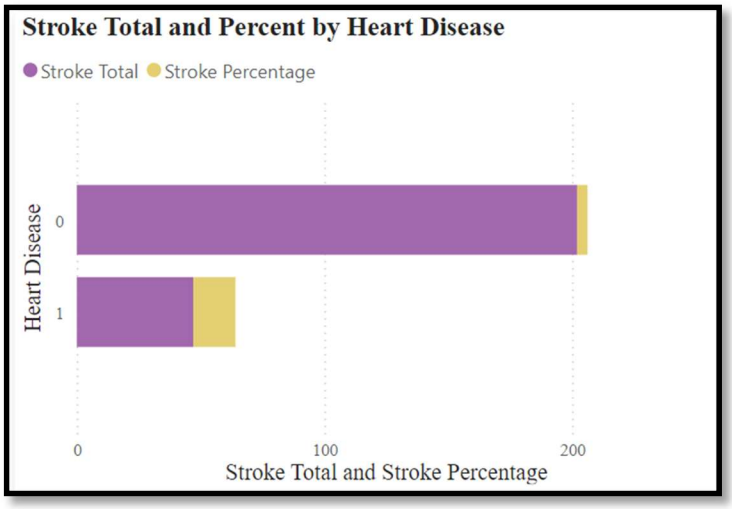
How does glucose level relate to stroke rate?

- Patients were separated based on their glucose category (Hypoglycemic, Normal, Prediabetic, Diabetic, and Sever Hyperglycemic).
- Stroke percentages were consistent for the Hypoglycemic, Normal, and Prediabetic groups (3.6% to 3.8%), with a significant increase for the Diabetic group, virtually doubling to 7.7%, followed by a sharp rise in the Sever Hyperglycemic group as the percentage rose to 11.45%.
- This indicates a strong correlation between glucose level and stroke occurrence within the dataset as somebody in the Sever Hyperglycemic group was approximately 3.15 times more likely to suffer stroke.



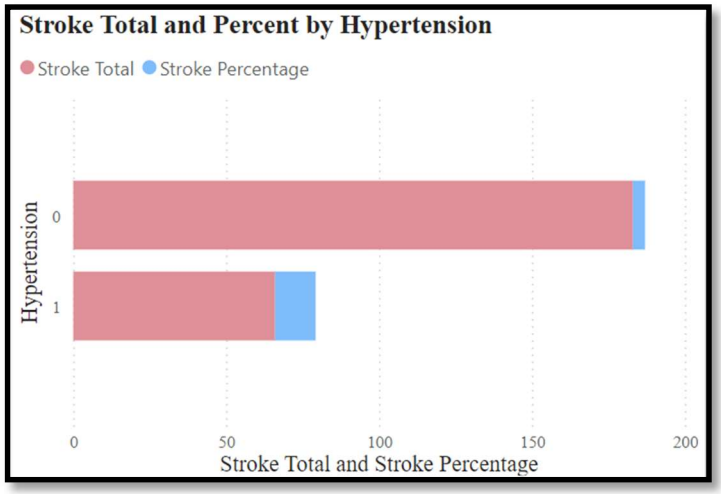
How does stroke rate compare between patients with and without heart disease?

- Patients were separated into two groups based on the presence of heart disease and then the stroke percentage of each group was determined.
- Patients with heart disease had a stroke rate of 17.03% and patients without heart disease had a stroke rate of 4.18% indicating a very strong correlation between heart disease and stroke occurrence as patients in the dataset with heart disease were approximately 4.1 times more likely to suffer a stroke.
- In the graph below the "Stroke Total" represents the number of patients in the dataset that either did or did not have heart disease and suffered a stroke. The "Stroke Percentage" represents what percent of total patients within each subgroup that suffered a stroke.



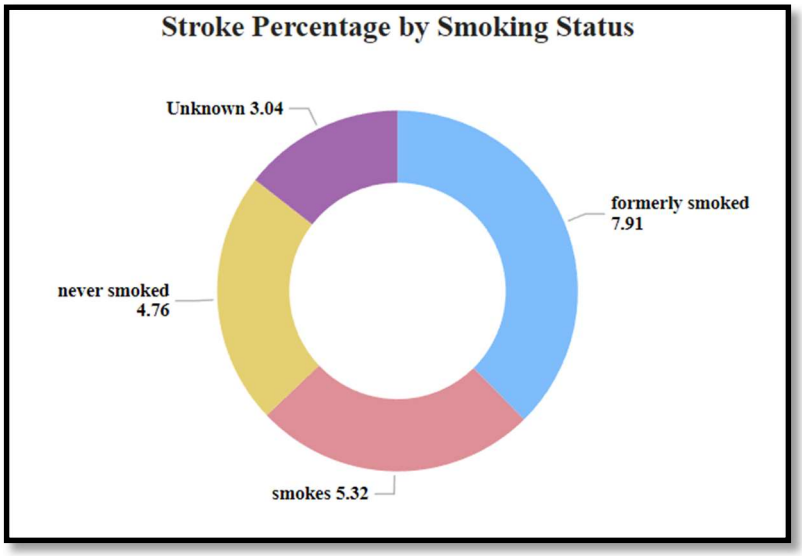
How does stroke rate compare between patients with and without hypertension?

- Patients were separated into two groups based on the presence of hypertension and the stroke percentage of each group was determined.
- Patients with hypertension had a stroke rate of 13.25% and patients without hypertension had a stroke rate of 3.97% indicating a very strong correlation between hypertension and stroke occurrence as patients within the dataset with hypertension were approximately 3.3 times more likely to suffer a stroke.
- In the graph below the "Stroke Total" represents the number of patients in the dataset that either did or did not have hypertension and suffered a stroke. The "Stroke Percentage" represents what percent of total patients within each subgroup suffered a stroke.



How does smoking affect stroke rate?

- Patients were separated into four smoking related groups (never smoked, smokes, formerly smoked, and unknown).
- Interestingly, patients who never smoked and those who currently smoke had similar stroke rates of 4.76% and 5.32%, respectively. However, those who formerly smoked had a significantly higher stroke rate of 7.91%, making them 66.18% more likely than non-smokers and 48.68% more likely than current smokers to suffer a stroke.



- A few reasons former smokers have a higher risk of stroke within the dataset are as follows:
  - Former smokers may have quit smoking due to pre-existing health conditions. Former smokers within the dataset are more likely than the other groups to have to hypertension and heart disease. Former smokers also have a higher average glucose level than the other groups. As previously illustrated, all these factors correlate with a higher rate of stroke occurrence.

Smoking Status	Hypertension	Heart Disease	Average Glucose
Formerly Smoked	13.56%	8.70%	112.90
Smokes	11.91%	7.73%	108.02
Never Smoked	12.26%	4.76%	107.56
Unknown	3.37%	3.11%	99.60

  - Former smokers within the dataset are approximately eight years older (on average) than patients that either never smoked or currently smoke. It's also been illustrated that age has a strong correlation with stroke occurrence.

Average Age	Smoking Status
54.93	Formerly Smoked
47.10	Smokes
46.70	Never Smoked
30.23	Unknown

- Given this information, despite displaying a correlation between formerly smoking and stroke rate, the data does not support a causal relationship between smoking status and stroke risk.

Is the dataset suitable for a predictive model?

- While several attributes within the views created for the dataset show a correlation to stroke rate, the overall linear relationships (shown below) within the dataset are not strong enough to construct a predictive model.

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
id	1.000000	0.002511	0.003538	0.003550	-0.001296	0.013690	-0.015757	-0.001403	0.001092	0.002977	0.014074	0.006388
gender	0.002511	1.000000	-0.028202	0.020994	0.085447	-0.031005	0.056422	-0.006738	0.055180	-0.026102	-0.062581	0.008929
age	0.003538	-0.028202	1.000000	0.276398	0.263796	0.679125	-0.361642	0.014180	0.238171	0.325956	0.265199	0.245257
hypertension	0.003550	0.020994	0.276398	1.000000	0.108306	0.164243	-0.051761	-0.007913	0.174474	0.160205	0.111038	0.127904
heart_disease	-0.001296	0.085447	0.263796	0.108306	1.000000	0.114644	-0.028023	0.003092	0.161857	0.038916	0.048460	0.134914
ever_married	0.013690	-0.031005	0.679125	0.164243	0.114644	1.000000	-0.352722	0.006261	0.155068	0.335711	0.259647	0.108340
work_type	-0.015757	0.056422	-0.361642	-0.051761	-0.028023	-0.352722	1.000000	-0.007316	-0.050513	-0.299450	-0.305927	-0.032316
Residence_type	-0.001403	-0.006738	0.014180	-0.007913	0.003092	0.006261	-0.007316	1.000000	-0.004946	-0.000119	0.008237	0.015458
avg_glucose_level	0.001092	0.055180	0.238171	0.174474	0.161857	0.155068	-0.050513	-0.004946	1.000000	0.168767	0.063437	0.131945
bmi	0.002977	-0.026102	0.325956	0.160205	0.038916	0.335711	-0.299450	-0.000119	0.168767	1.000000	0.219149	0.038971
smoking_status	0.014074	-0.062581	0.265199	0.111038	0.048460	0.259647	-0.305927	0.008237	0.063437	0.219149	1.000000	0.028123
stroke	0.006388	0.008929	0.245257	0.127904	0.134914	0.108340	-0.032316	0.015458	0.131945	0.038971	0.028123	1.000000

- There is also a significant imbalance in data given that there is a much larger sample of non-stroke patients compared to stroke patients. This would lead to a biased predictive model that performs poorly for the stroke patients.

Conclusion

The analysis of this dataset reveals several factors that significantly correlate to stroke rate. The most prominent of these factors is aging. A patient in their 60s was approximately six times more likely than patient in their 30s to have a stroke and patients in their 70s were over twice as likely as patients in their 60s to have a stroke. Stroke rate also significantly correlated with other health factors such as the presence of hypertension, heart disease, and high glucose levels.

Other factors in the dataset such as gender, BMI, and smoking status showed possible correlations with stroke rate, but such correlations should be taken with caution. The correlations with gender and BMI were minimal and could be the result of underlying factors. Smoking status had identifiable underlying factors uncovered within the dataset such as former smokers being older and more likely to suffer from hypertension, heart disease, and high glucose levels.

Despite uncovering several factors that correlate with stroke rate, the dataset proved to be unsuitable for a predictive model due to a lack of linear correlation within the data set and a strong imbalance between non-stroke patients and stroke patients.

Ultimately, this analysis underscores the importance of taking a multifaceted approach to stroke prevention, particularly as individuals age. Future studies could explore the interactions of these risk factors in greater detail and assess the impact of lifestyle interventions on stroke prevention.