Author: Bryant Le
Date: August 5, 2025
Tools: Excel, Tableau
Dataset Source: https://www.kaggle.com/datasets/shriyashjagtap/stroke-diagnosis-and-health-metrics-data/data

INTRODUCTION

This report presents an exploratory data analysis of factors related to stroke. The objective is to identify relationships among variables such as age, gender, hypertension, and others to assess the likelihood of experiencing a stroke.

DATASET OVERVIEW

Rows: 10,000
Columns: 10
Key Variables:

- Age
- Gender
- SES (Socioeconomic Status)
- Hypertension
- Heart Disease
- BMI
- Average Glucose
- Diabetes
- Smoking Status
- Stroke

DATA CLEANING

- Rounded down the age values to the nearest integer.
- Rounded the BMI and Average Glucose levels to two significant digits.

VISUAL ANALYSIS AND INTERPRETATION

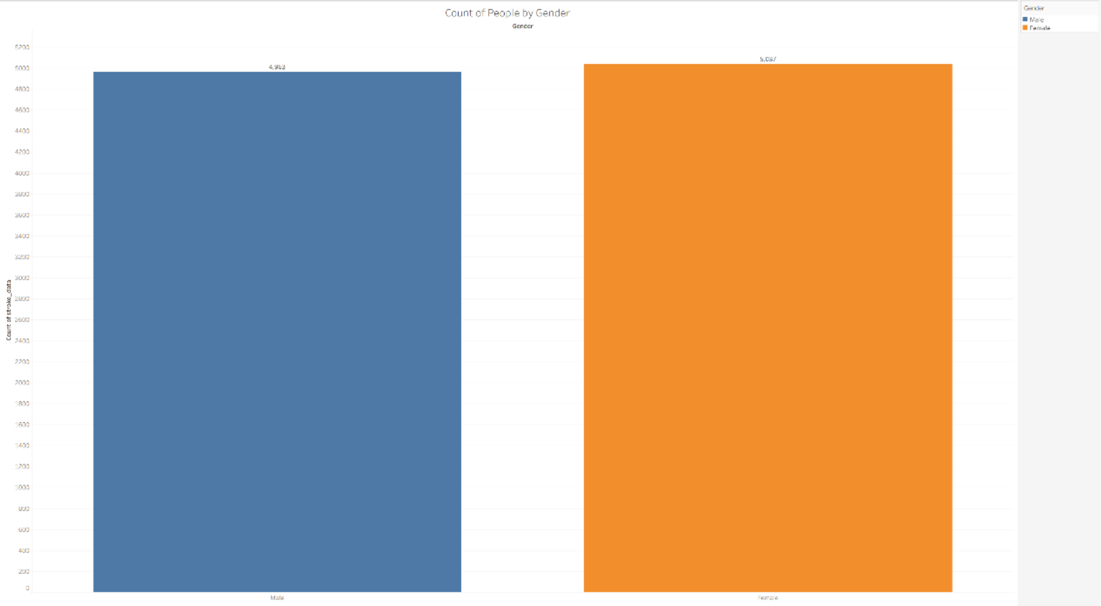

Chart 1: Count of People by Gender
Information: Females have a marginally higher presence in the dataset compared to males.
Knowledge: The ratio between both genders is almost 1:1, suggesting that bias is less likely to be a factor in future results.

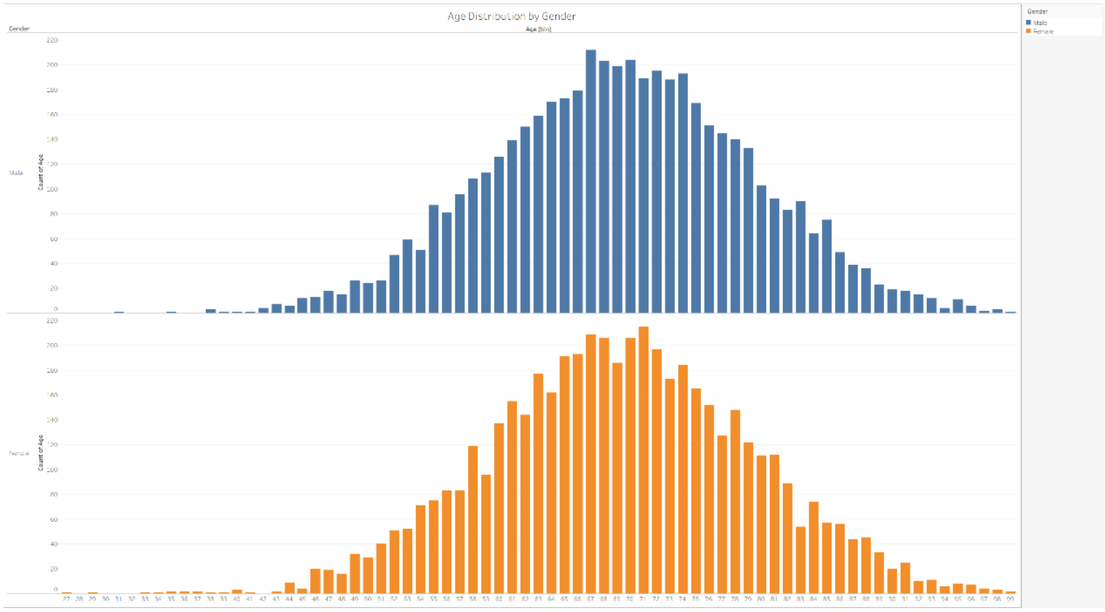

Chart 2: Age Distribution by Gender
Information: Male and female age histograms are both normal distributions with nearly identical results.
Knowledge: While there are notable differences between the male and female values, there is no significant difference in the distributions that would suggest bias or cause concern.

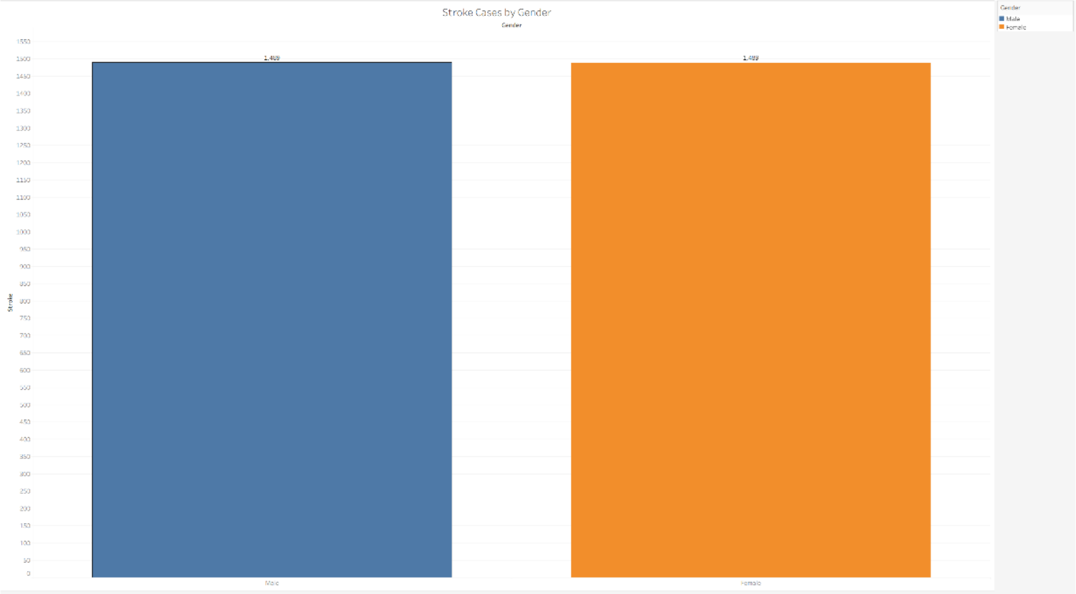

Chart 3: Stroke Cases by Gender
Information: The number of stroke cases is equal for both genders.
Knowledge: While no gender appears to have a higher risk for stroke, it would be wise to examine other datasets and/or other stroke risk factors to verify claims of whether certain genders have a higher risk of stroke.

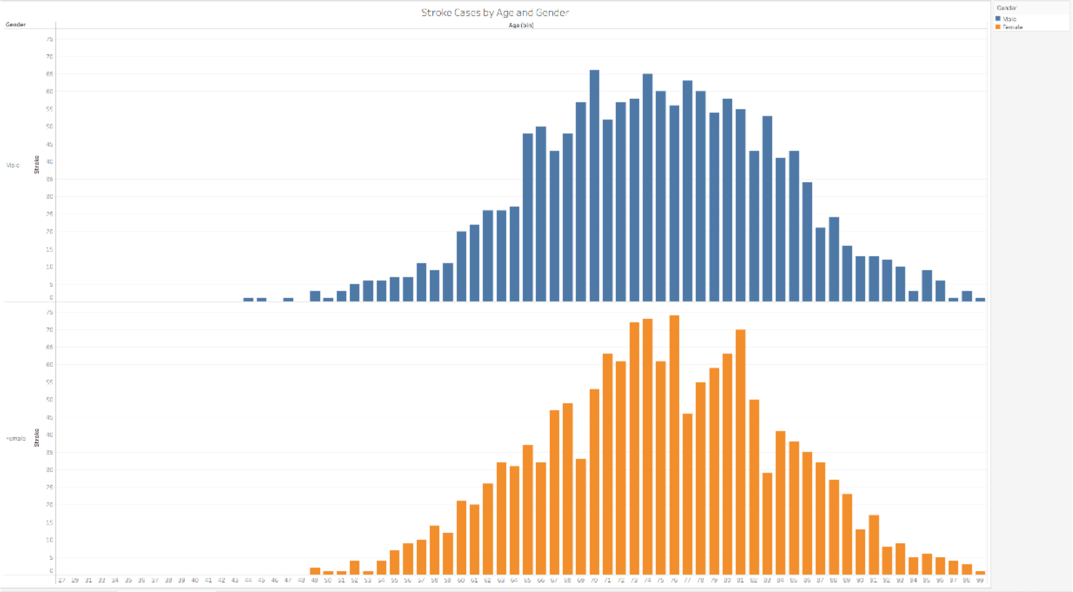

Chart 4: Stroke Cases by Age and Gender
Information: Females have a higher stroke count for individuals aged 73, 74, and 76, compared to males.
Knowledge: This is sufficient evidence to suggest that females in this age range have a higher risk of stroke, but further research on their lifestyles will be necessary.

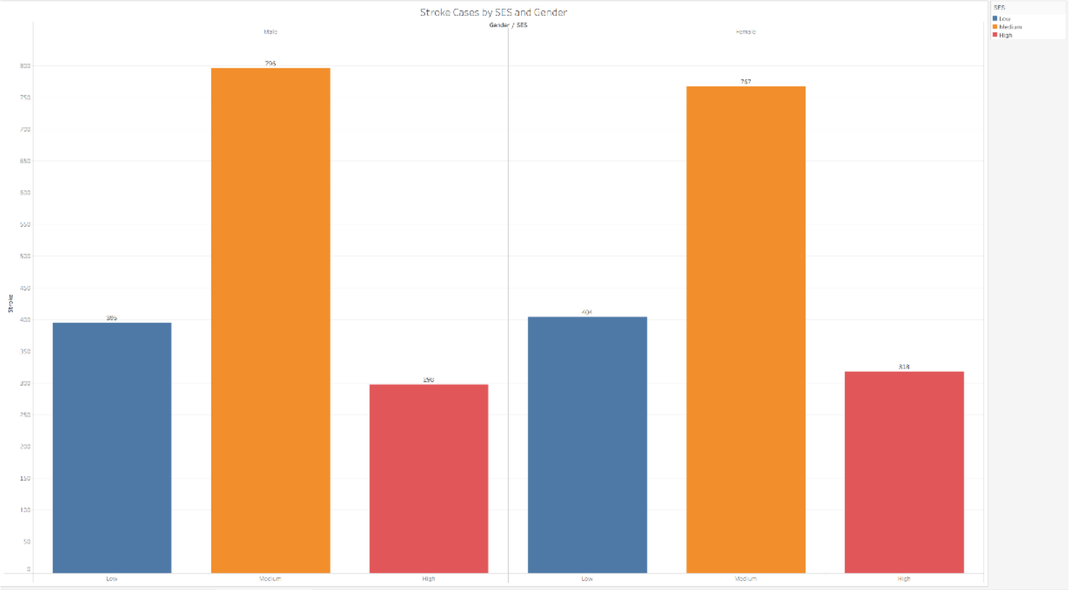

Chart 5: Stroke Cases by SES and Gender

Information: People in the medium SES group have a significantly higher rate of stroke cases compared to low and high SES groups.

Knowledge: While the rate of stroke cases in medium SES individuals is concerning, it should supersede the presence of stroke cases in low and high SES individuals. Further analysis will be needed to determine if medium SES individuals have higher rates of other health-related issues.

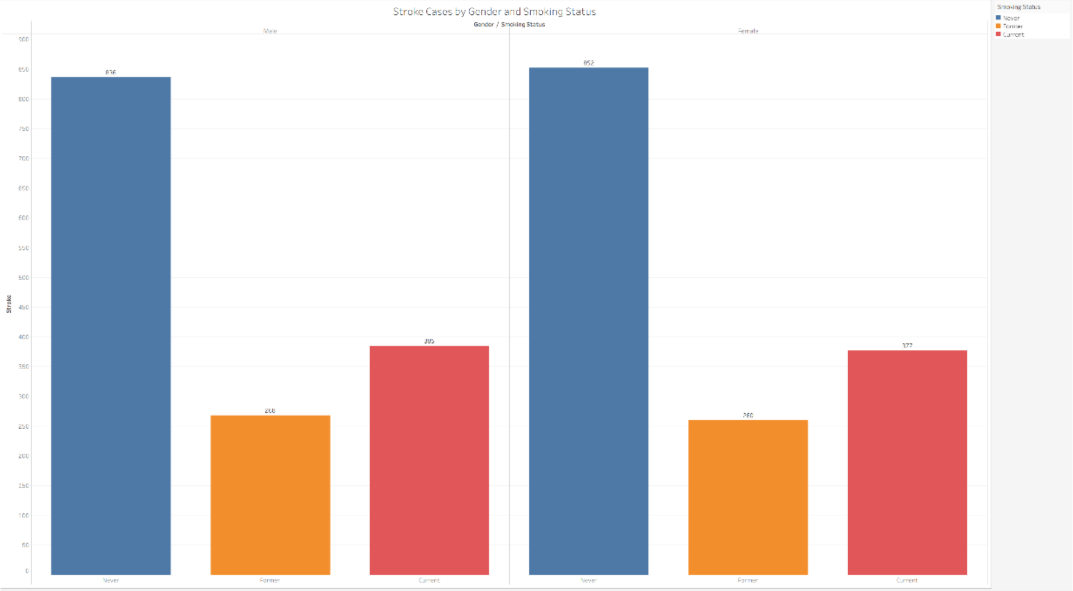

Chart 6: Stroke Cases by Gender and Smoking Status

Information: Most participants claim to have never smoked in their lives.

Knowledge: The data suggests that smoking is most likely not to be a significant factor in whether a person has a stroke.

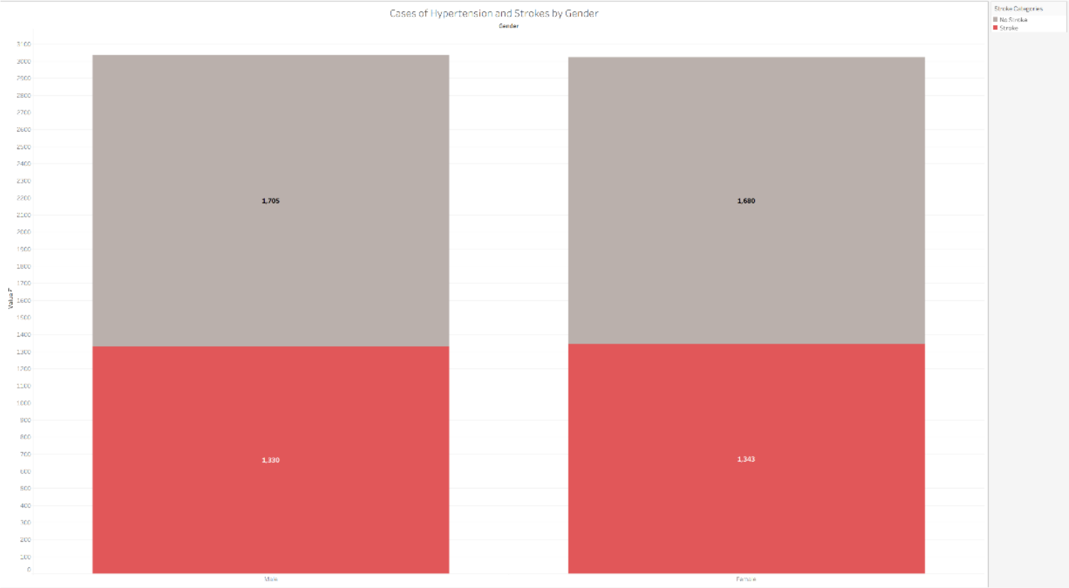

Chart 7: Cases of Hypertension and Strokes by Gender

Information: The ratios for no-stroke cases and stroke cases are nearly equal for both genders.

Knowledge: While hypertension is likely to contribute to a stroke, it is unknown if individuals had hypertension prior to a stroke or developed hypertension after a stroke.

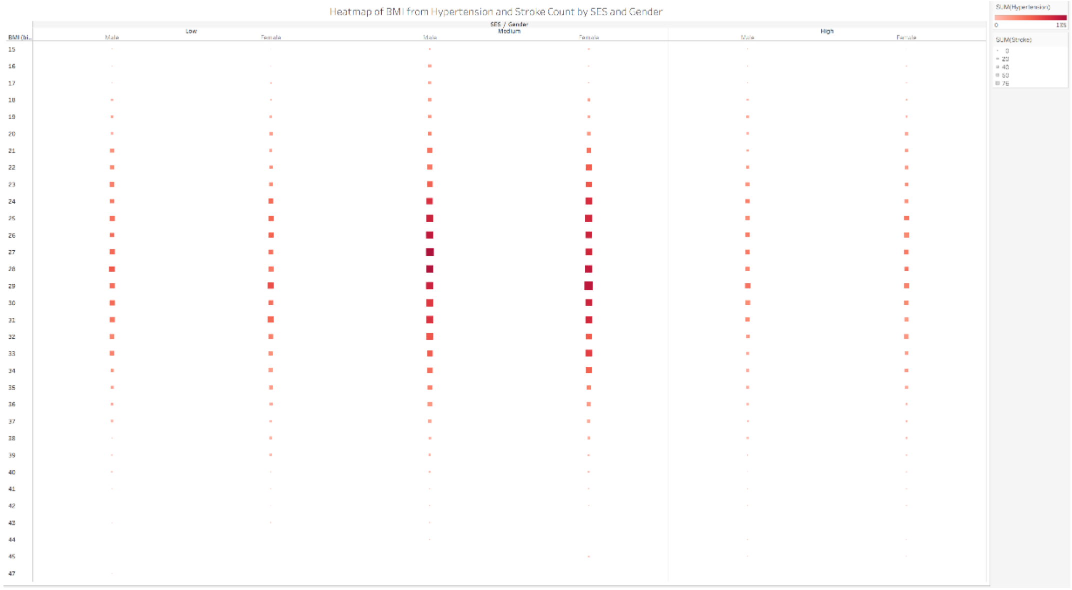

Chart 8: Heatmap of BMI from Hypertension and Stroke Count by SES and Gender

Information: Medium SES individuals with a BMI of 23-33 have reported high cases for hypertension and stroke.

Knowledge: It is possible that medium SES individuals have a higher representation than low and high SES individuals in the dataset.

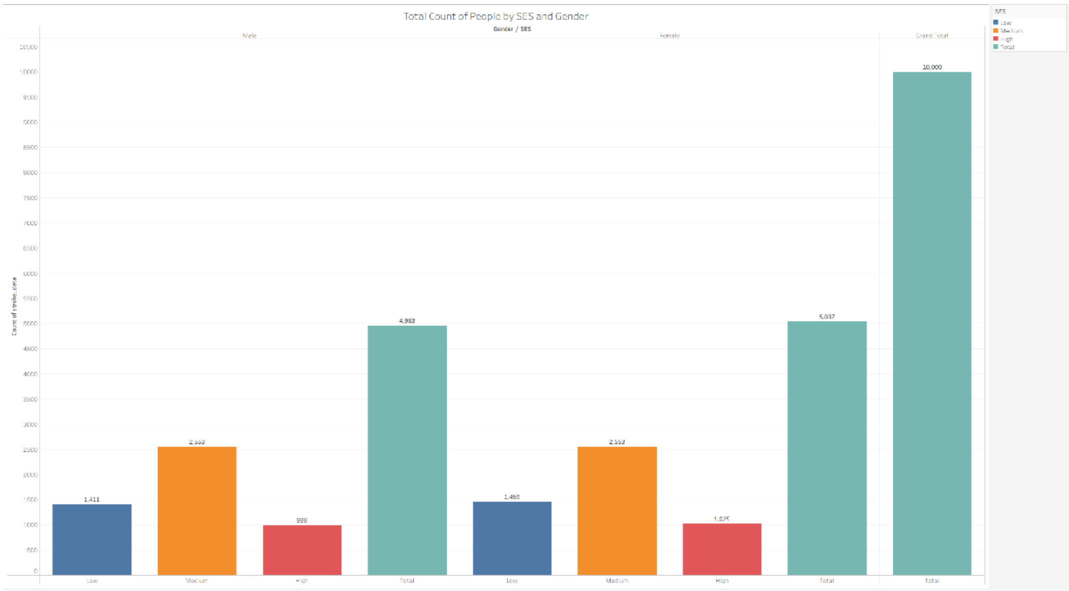

Chart 9: Total Count of People by SES and Gender
Information: Medium SES individuals have the most presence in the dataset.
Knowledge: This likely explains the higher representation of medium SES individuals compared to the others, potentially skewing the results.

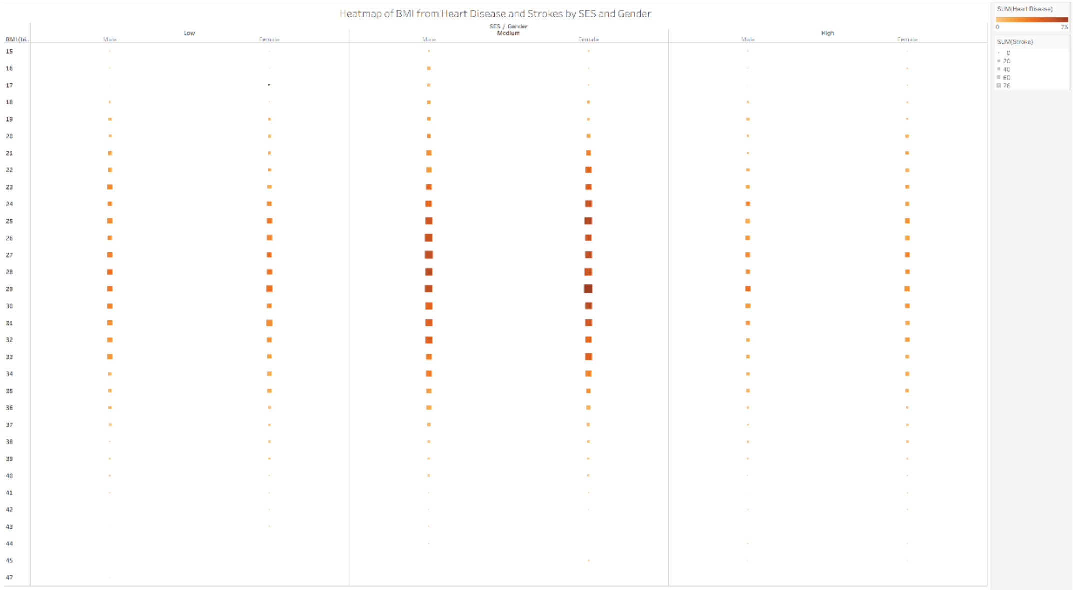

Chart 10: Heatmap of BMI from Heart Disease and Strokes by SES and Gender
Information: Medium SES individuals with a BMI of 23-33 have reported high cases of heart disease and stroke.
Knowledge: Since medium SES individuals have a higher representation in the dataset, it would be difficult to interpret how other SES groups are affected by stroke-related factors.

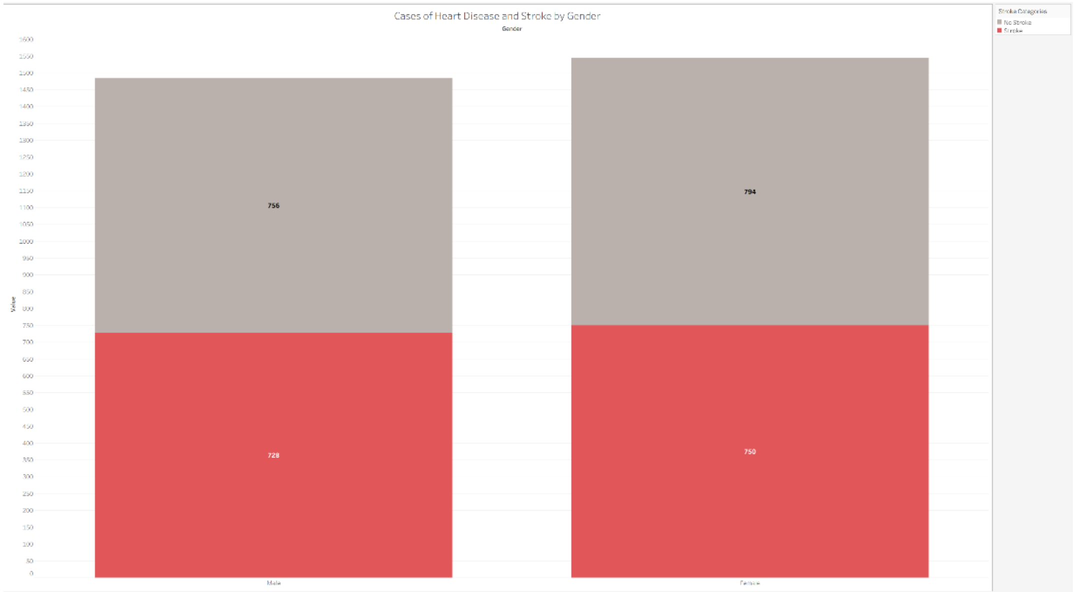

Chart 11: Cases of Heart Disease and Stroke by Gender
Information: Women have slightly more cases of heart disease than men.
Knowledge: Despite women having more cases of heart disease, the ratios for males and females are nearly identical. Therefore, it is unlikely that women will contract heart disease more than men based on this result.

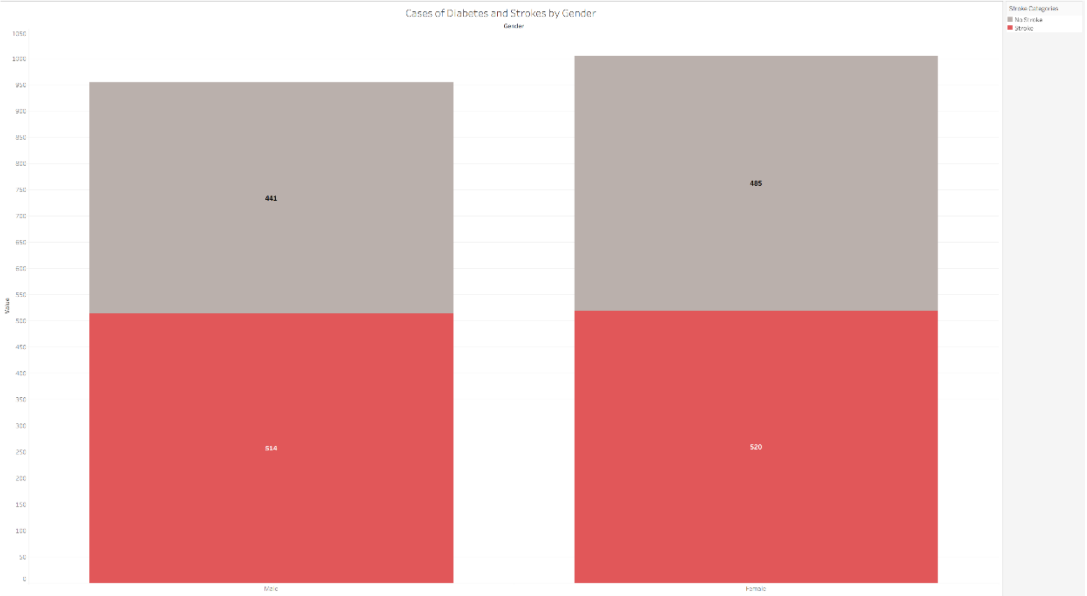

Chart 12: Cases of Diabetes and Strokes by Gender
Information: Women have slightly more cases of diabetes than men.
Knowledge: Similarly to the previous visualization, the ratios for males and females are nearly identical. Therefore, it is unlikely that women will contract diabetes more than men based on this result.

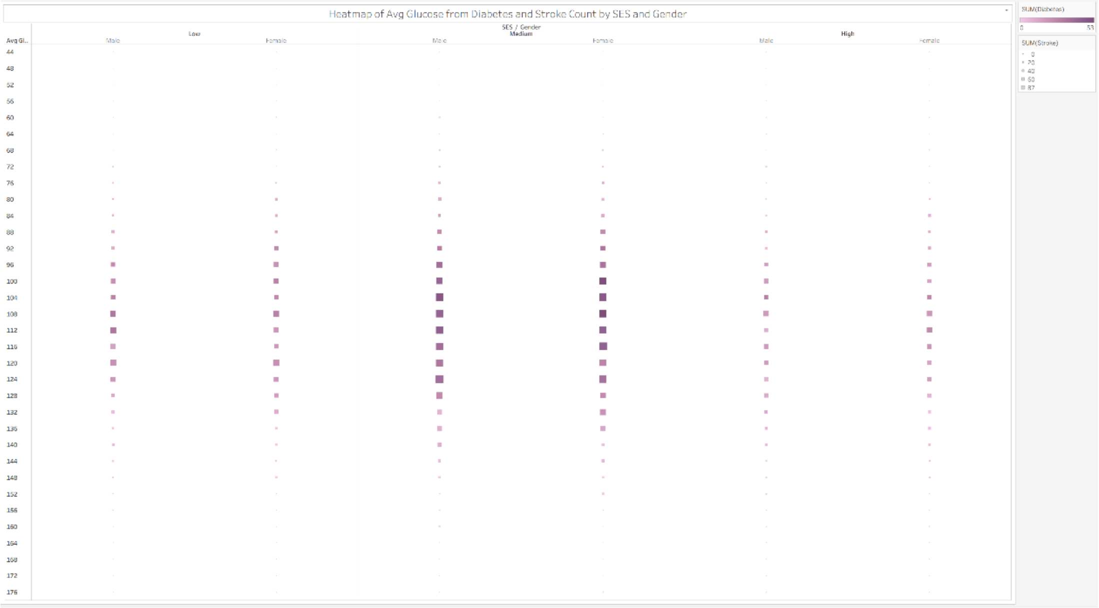

Chart 13: Heatmap of Avg Glucose from Diabetes and Stroke Count by SES and Gender
Information: Medium SES individuals in the 92-124 range have the highest representation of diabetes and stroke cases.
Knowledge: Most individuals in the medium SES group are classified as pre-diabetic, so it is possible that diabetes may be a contributing factor to having a stroke.

CONCLUSION

It is possible that hypertension and heart disease are risk factors for having a stroke; however, it is difficult to draw this conclusion given that medium SES individuals are overrepresented in this dataset. This suggests that the dataset may not be reliable for deriving a causation.

LIMITATIONS

Medium SES individuals exceed the number of low and high SES individuals combined, so these visualizations can be interpreted as more biased towards the former. Therefore, these visualizations should not be representative of the entire population.

FOLLOW-UP PLANS

- Examine low- and high-SES individuals to determine if the results are like medium-SES individuals.
- Research datasets that focus on low- and high-income earners if unbiased datasets are unavailable.
- Replicate visualizations to determine if male and female results remain consistent if using other datasets.