The format of the paper should be:

- Summary: A one paragraph description of the question, methods, and results (about 350 words).
- Data: One to two pages discussing the data and key variables, and any challenges in reading, cleaning, and preparing them for analysis.
- Results: Two to five pages providing visualizations, statistics, a discussion of your methodology, and a presentation of your main findings.
- Conclusion: One to two pages summarizing the project, defending it from criticism, and suggesting additional work that was outside the scope of the project.
- Appendix: If you have a significant number of additional plots or table that you feel are essential to the project, you can put any amount of extra content at the end and reference it from the body of the paper.

Summary:

Data:

In order to build a model to help us predict whether or not someone will have a stroke, we looked at numerous categorical and quantitative variables. For our quantitative variables, we looked at age, BMI (body mass index), and average glucose level. The categorical variables that we chose to observe are whether or not someone has ever been married, gender, whether or not someone had heart disease, whether or not someone had hypertension, residence type, and work type. In order to prepare the original data for analysis, we split the data set into test and training set, and then we further split both datasets to separate the response variable that we are trying to predict, which is whether or not someone had a stroke (1 if the person had a stroke, 0 if the person did not have a stroke).

After splitting the data, we go into data wrangling and exploratory data analysis. We cleaned the training set first before moving on to clean the test set. The processes that we performed to clean the training set and the test set were identical as it is good practice to prepare the datasets the same way. This helps to ensure that our model built from the original test set will hopefully be more compatible when testing on the test set.

The first thing that we did in our data wrangling process was to just observe the overall shape of the predictive variables training dataset and check if there were any missing values. After we checked if there were missing values, we discovered that there were a few missing values from the BMI column. We decided to fill the missing BMI values with the average BMI of the dataset as there were not that many missing points and this is a good way to mitigate the missing values issue. We then looked at the summary statistics of the predictive variables in the training set in order to get a better grasp of the data and to see if our categorical variables were the correct data type. When we ran summary statistics and our categorical variables were outputting quantitative statistics like mean and max, we knew that the categorical variables were not encoding properly. Therefore, we turned the hypertension and heart disease column into a factor to ensure that it was a categorical variable and our models would read it as such.

We then moved on to check for outliers by making boxplots. We saw that both the average glucose level column and the BMI column had outliers, so we proceeded to windsorize them as our method for handling our outliers. Choosing how to handle our outliers was an problem initially as we were unsure if whether or not we should remove them the rows with outliers all together, but we ultimately decided to windsorize them as there were a lot of outliers and completely removing these rows gives us less data to train our model with. We then proceeded to examine our quantitative variables, age and BMI, by making histograms in order to check if we needed to transform these variables using functions like the arcsine of a column or the logarithm of a column. The histograms of these columns appear to show the distribution of the data well and spread out, so we decided that we would not need to transform these columns.

After looking at the quantitative data, we decided to check the categorical variables by ensuring that there were no data inconsistencies (ie. a female is denoted only by the phrase 'Female' and not also by 'F'). We found that there were no data entry inconsistencies with our categorical variables, so the last steps that we did to prepare our data was to one hot encode our categorical variables for our model building. We also removed the columns 'Unnamed:0' and 'id' as these variables are solely for enumerating our dataset and they serve no relevance as to predict whether or not someone had a stroke.

We want to reiterate that the process that was outlined above to prepare our training dataset was replicated to also prepare our test set for our model. This is to ensure consistency as we build and test our model.

Results:

Conclusion:

Appendix: