Stroke Predictions Data Analysis

...

Background on Data

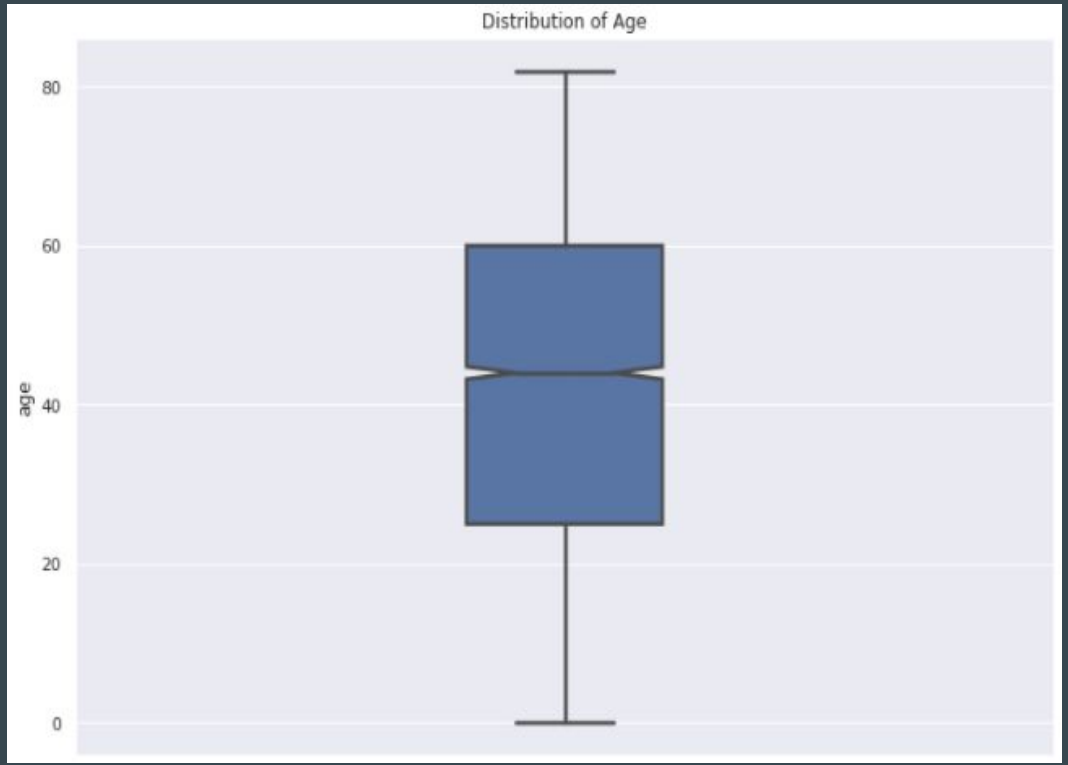
- 12 features and 5510 observations.
- Our features consist of Id, Gender, Age, Hypertension, Heart Disease, Ever Married, Work Type, Residence Type, Average Glucose Level, Bmi, Smoking Status, and Stroke.
- Our data consist of patients from the ages starting as early as infancy all the way to 80 years of age

EDA

- Dropped Id column just unique a identifier.
- There were no duplicates in our data.
- There were 201 null values for our “bmi” column. I created a heat map to see if there were any strong correlations between our bmi and other features.
- Ended up dropping the 201 rows which was 3.9% of our data. My original plan was too further inspect the data to maybe fill in those missing bmi values, but there wasn’t any accurate way to fill in the values.
- No syntax errors in our category columns.
- No inconsistent or impossible values.

Age Group

- For age we have no outliers, but it is interesting how we have patients who are newborn babies as well as the elderly 80 years of age.
- After doing further research on the features of our data, newborns can also have hypertension, heart disease, and even a stroke.

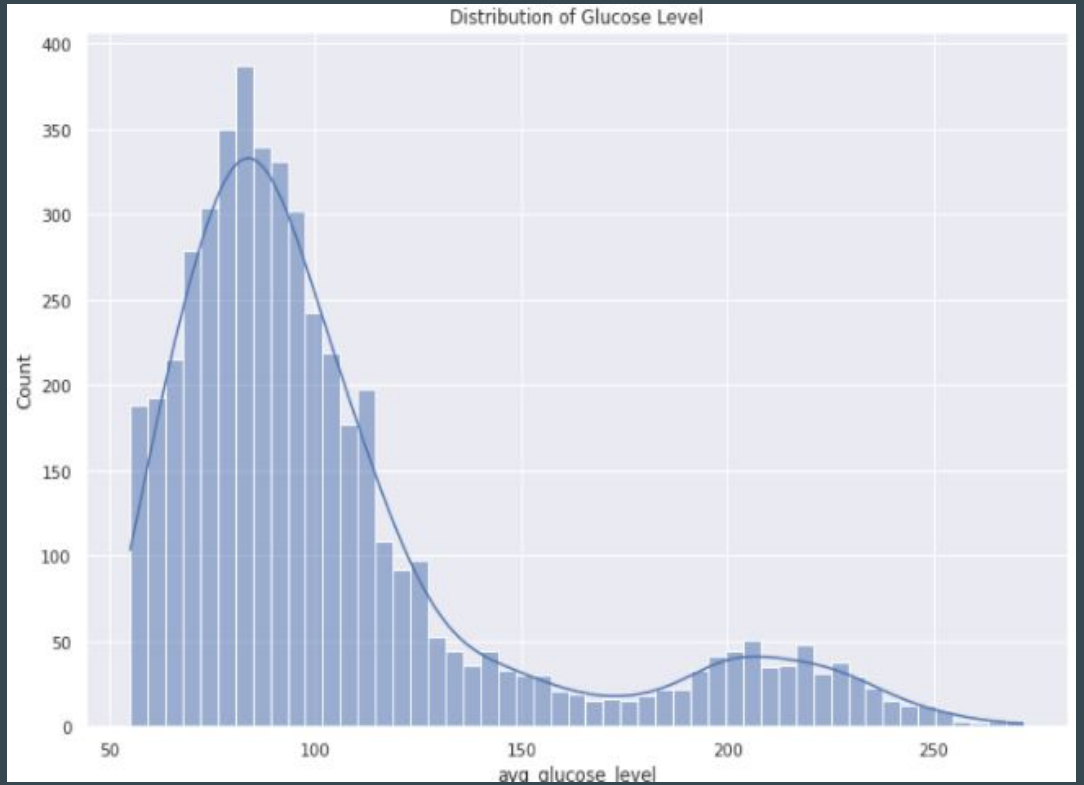


BMI

- We have a very broad range of bmi, also many outliers.
- Our average is around 25-30. Ranging from 10 all the way to 97. Which are not impossible numbers for bmi.

Distribution of Glucose Levels

- From the distribution plot we can see that most fall under the range of 60-110 glucose levels. We do have some outliers.
- Since there are some high values for glucose levels we can maybe link this to hypertension, heart disease, and stroke.



Unique Challenges

- For starters, I believe BMI might not be the best metric to predict stroke. This being because BMI is an inaccurate measure of body fat content, because it does not take into account muscle mass, bone density, or racial and sex difference.
- Since our data covers infancy all the way to senior citizens many of our entries for smoking status are unknown and this was due to the fact that newborns can't answer that question. So our unknowns aren't nulls they were intentional.
- One problem that I can foresee is that our data is highly imbalance with more people not having a stroke, so this is going to be an issue in the model development phase.
- With that being said a new question arises from my observations, not if we can predict someone being at risk for a stroke but if the features in our data are sufficient enough for us to be able to predict someone being at risk for a stroke.