EDA On Stroke Data

Data

This dataset is used to predict whether a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking status. Each row in the data provides relevant information about the patient.

The Data was Taken from Kaggle

Attribute Information

1) id: unique identifier
2) gender: "Male", "Female" or "Other"
3) age: age of the patient
4) hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
5) heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
6) ever_married: "No" or "Yes"
7) work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
8) Residence_type: "Rural" or "Urban"
9) avg_glucose_level: average glucose level in blood
10) bmi: body mass index
11) smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"*
12) stroke: 1 if the patient had a stroke or 0 if not

Data Cleaning

On observing the data information, we found out that column bmi have some null values. To get rid of these we started by creating a copy of original data. And then replacing all the null values with the mean of the column.

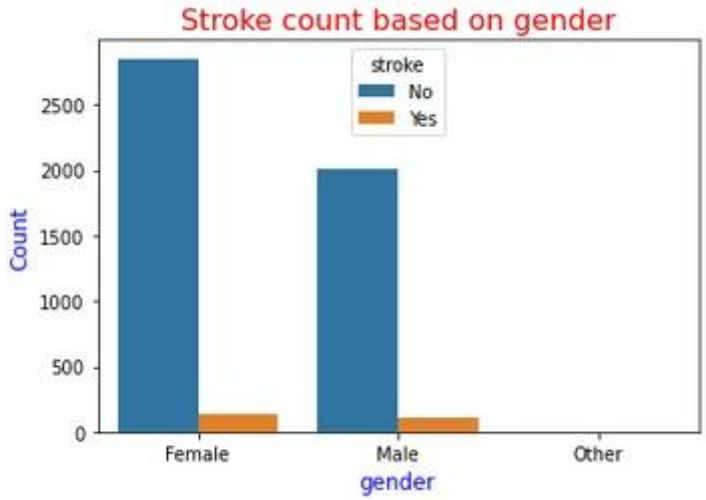
Also, to make it more appropriate for visualizing we just replace the categorical 0,1 values to No and Yes. It will help the person by making visualization more appealing.

We required to act differently on columns based on their type, so we just separated the name of categorical columns and numerical columns.

Visualizations

We started by creating bar plot for different categorical columns according to their frequencies.

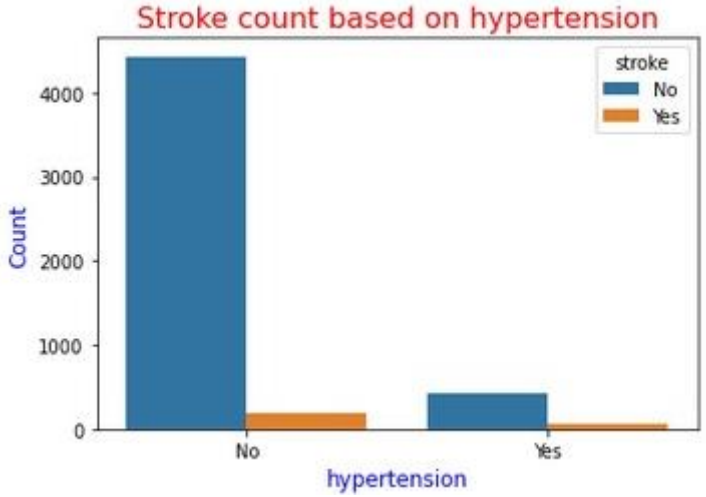
- Stroke Count based on gender



* it is around 5% for males and 4% for females those have been struggling with stroke

* this can be the important feature as 1% is a significant drop

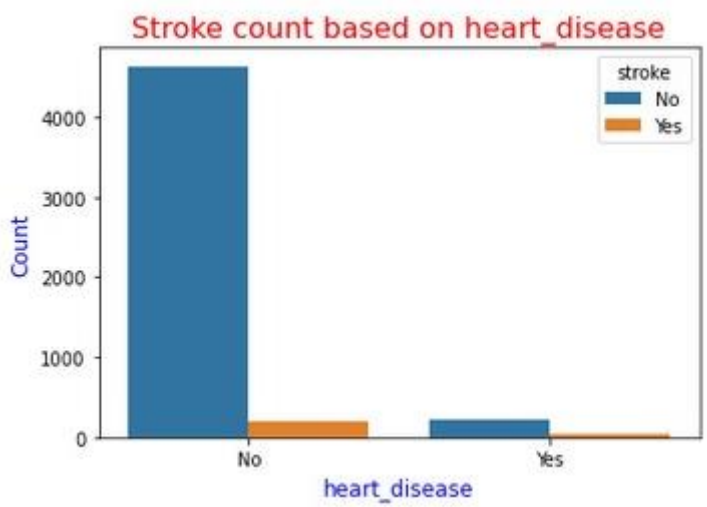
- Stroke Count based on hypertension



* It is found to be person tested with hypertension is more likely to be stroked (~ 13%)

* and this is also accepted by having some prior domain knowledge

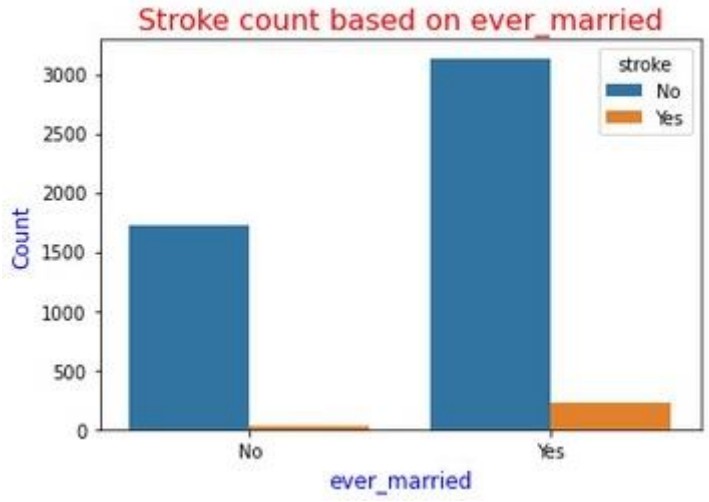
- Stroke Count based on heart_disease



* It is found to be person with heart disease are at high risk of stroke (~ 17%)

* This is also accepted by having some prior domain knowledge

- Stroke Count Based on marital Status

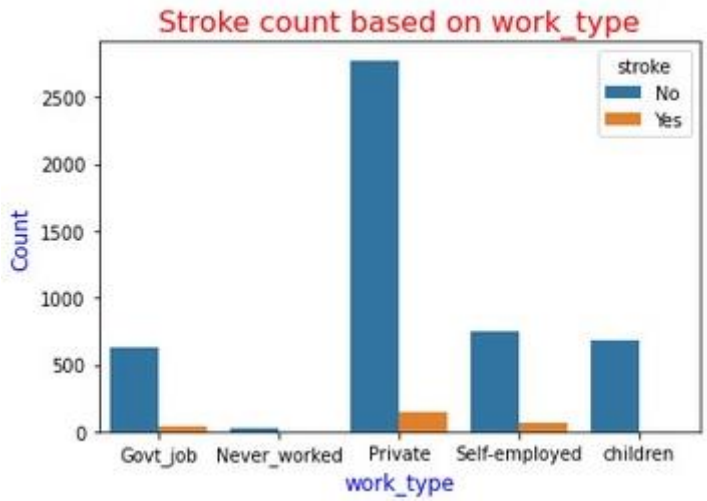


* 6% of married people have faced stroke

* also is it to be notice that we have more biased data around the married people nearly 2000 more entries married people

* This need to be investigated further

- Stroke count Based on Work Type



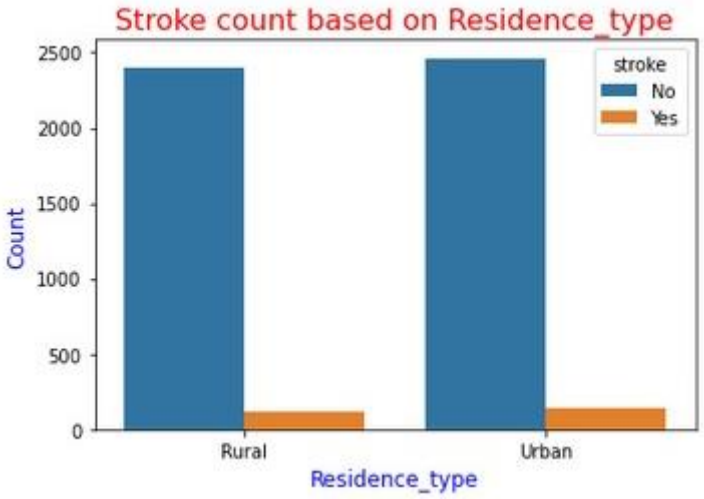
* Based on the data it is seen that there are a greater number of records for private sector jobs

* It can be possible that people having govt job are more relieved than people with private jobs but on other have people having no job don't have stroke who should probably be more worried about getting a job

* And children are being tested positive for stroke

* Conclusively Stroke doesn't depend on Work Type

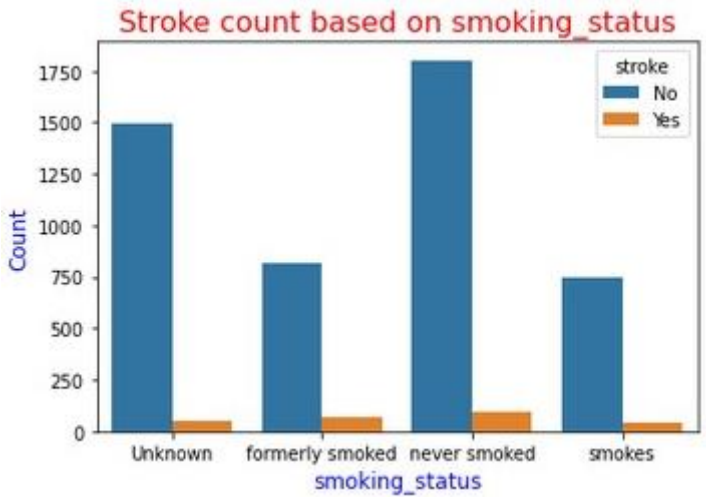
- Stroke based on Residence Type



* People living in Urban Areas have! % more Stroke than People Living in Rural areas

* Urban areas have more Pollution level and Busy life style which can have great impact on the health of a person

- Stroke Count Based on Smoking Status

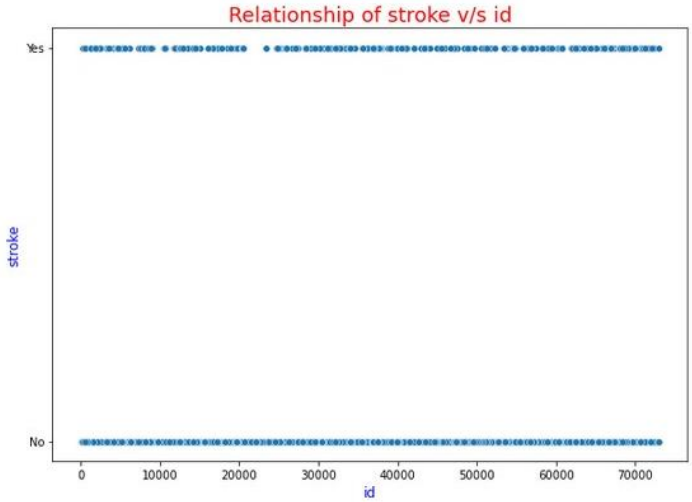


* Similarly, for people who formerly smoked or smokes have more chances to get Stroke than people who doesn't

* It is un predictable for the people with Unknown Status

Plotting Scatter plots for numerical data

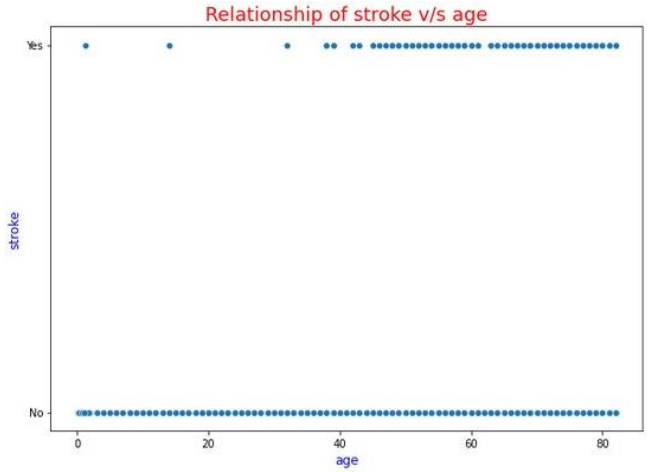
- Relationship of Stroke and id



* The feature id is totally irrelevant for prediction purpose as it adds no information to the data

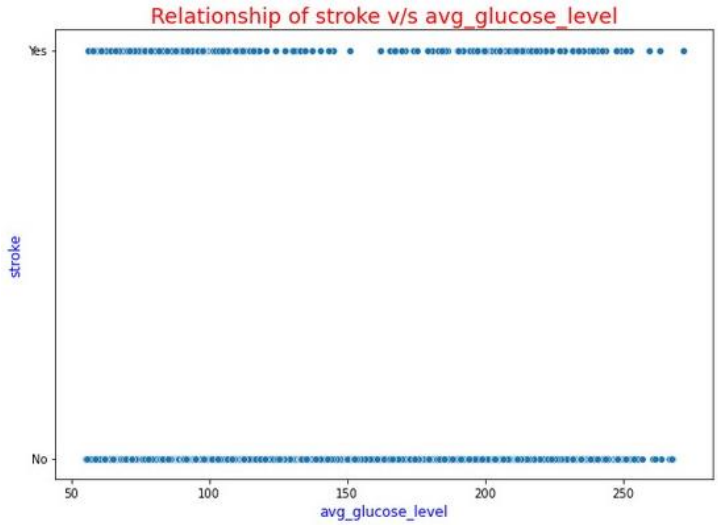
* The very person will have different id and no pattern will be formed

- Relationship of Stroke and age



* According to the trend elder people will have high risk of getting stroked

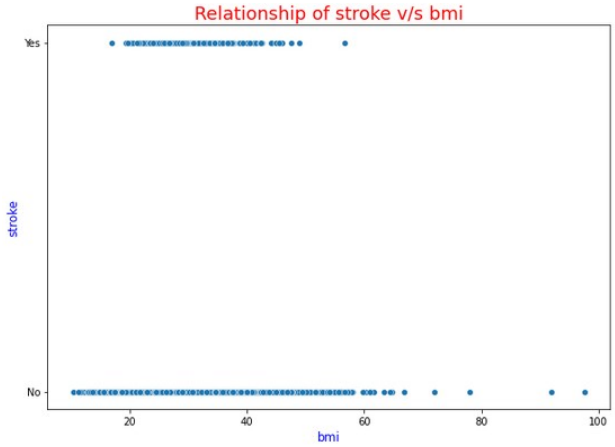
- Relationship of Stroke and average glucose level



* People having glucose level around 150 mg/dl are less prone to get stroke

* while people having more or less than this are more prone for stroke

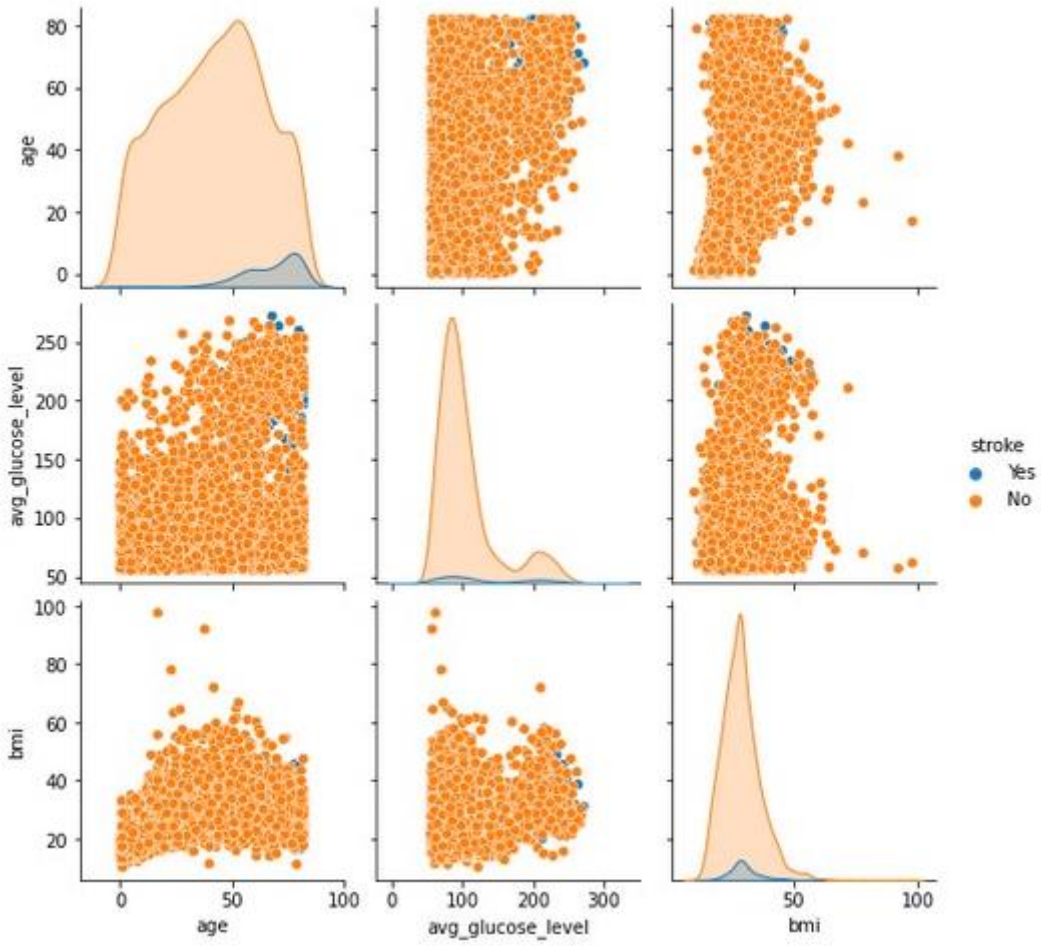
- Relationship of Stroke and average BMI



* We can't really tell if this is a good predictor or not

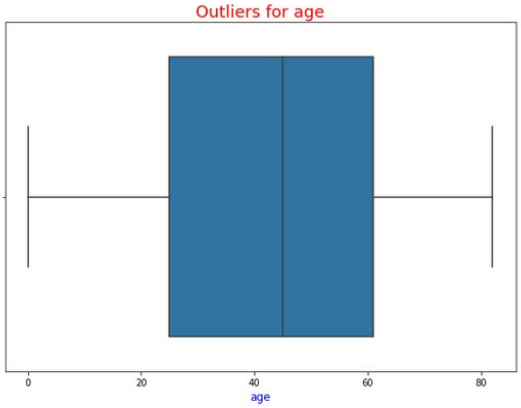
* Further investigation is needed

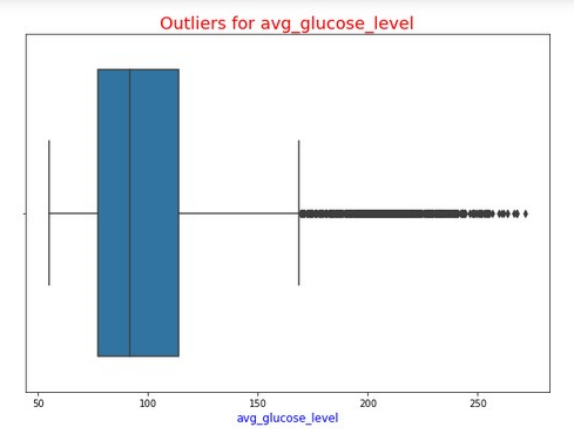
Further plotting we can see bmi and avg_glucose_level works together to decide the impact

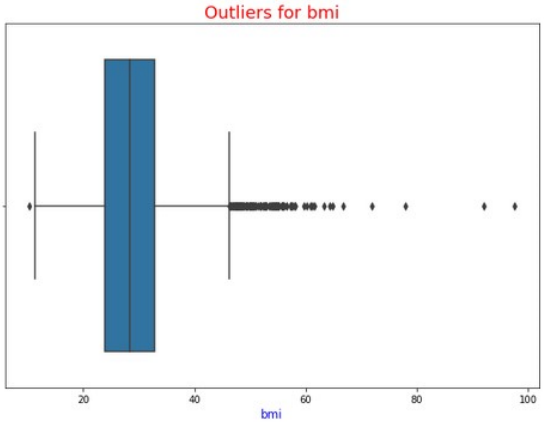


We can see that as the glucose level increase but the bmi of person remain same there are chances that the person will still get a stroke.

Now we want to take the numbers of outliers we have in our data







From this we can see clearly the Column bmi and average glucose level have a ton of outliers we can handle this with suitable techniques like replacing it with mean or median or dropping in in this case we are leaving these outliers as it is.

In hope that an extreme behaviour target variable is highly dependent on the pattern of these features

Hypothesis

Creation of Hypothesis

The hypothesis is:

- For Marital Status

- Null Hypothesis: Married people are more prone to stroke
- Alternative Hypothesis: All people have same chances of getting stroked, Marital status doesn't matter
- Significance Level: 0.01

- For BMI

- Null Hypothesis : People having BMI between 20 and 50 with high avg glucose level > 200 are more prone to stroke
- Alternative Hypothesis: All people have same chances of getting stroked, BMI doesn't matter
- Significance Level: 0.01

- For Avg Glucose Level

- Null Hypothesis: People having avg glucose level between 140 and 160 are less prone to stroke
- Alternative Hypothesis: All people have same chances of getting stroked
- Significance Level: 0.01

Testing

On testing the Hypothesis for Marital Status, we found out that the resulting p_value of the test is high

And hence we accept the null hypothesis and conclude that marital status can affect the chances of getting stroke.

```
stat_score,p_value=ttest_ind(married_people,all_people)


if p_value>0.01:
    print('accept the hypothesis null hypothesis')
else:
    print('Rejetc Null hypothesis')
```

```
accept the hypothesis null hypothesis
```

We can perform the similar tests for other hypothesis and check whether to accept or reject the null hypothesis.

Conclusion

The data is fairly less to make prediction based on the field of study. And the proportion of the data is not in right amount. More samples are Negative.

Although the visualization made it fairly easy to analyse the data and get familiar with the trend.

References

See the full code here