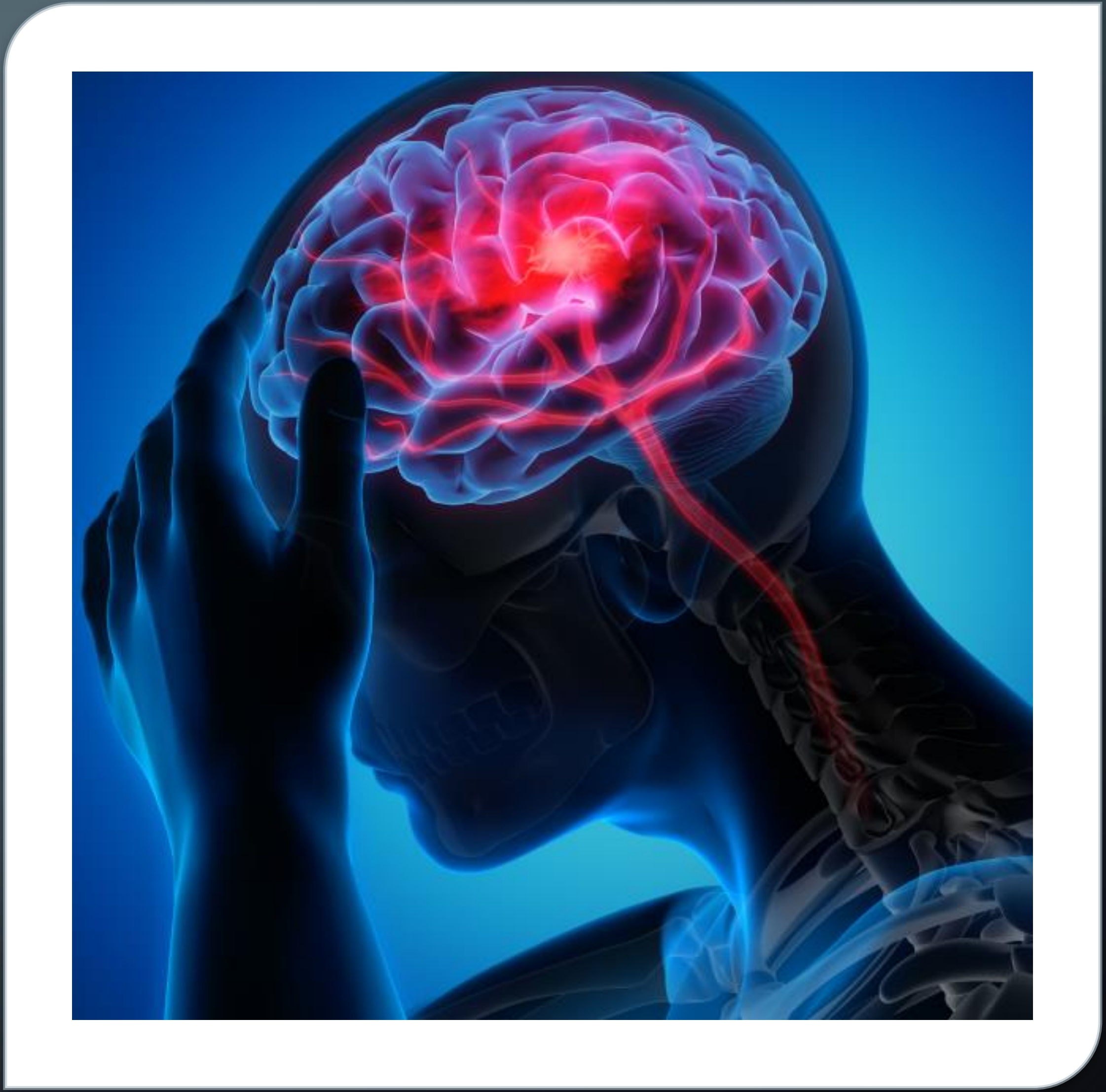
STROKE PREDICTION DATA SET EVALUATION

EXPLORATORY DATA ANALYSIS



INTRODUCTION

- Stroke is the fifth leading cause of death and disability in the United States according to the American Heart Association.

- Every 40 seconds in the US, someone experiences a stroke, and every four minutes, someone dies from it according to the CDC. A recent figure of stroke-related cost almost reached $46 billion.

- With my interest in healthcare, I wanted to explore which are important factors that contribute to stroke to help patients and health providers minimize the risk and cost.

- For this presentation I wanted to evaluate the viability of the dataset before using it for prediction.

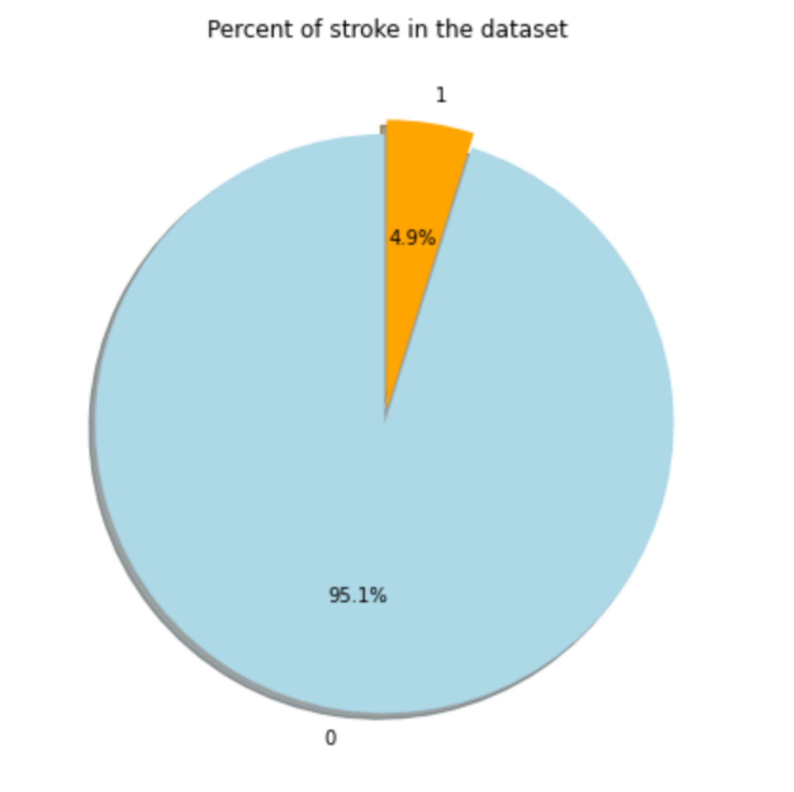
DATASET DESCRIPTION

This presentation uses the a Kaggle dataset which consists of 5110 observations:

- Three numerical features
- Eighth Categorical features

The percentage of stroke in the dataset is 4.9 %

Key observation: The overall sampling of the data set is imbalanced towards patients with no strokes.



QUESTIONS TO ASK BASED ON THE FEATURES

1) Male/Female who has more risk for strokes.

2) People of which age group are more likely to get a stroke.

3) Is hypertension a cause?

4) A person with heart disease is more likely to get a stroke?

5) Does being married play a role in causing a strokes?

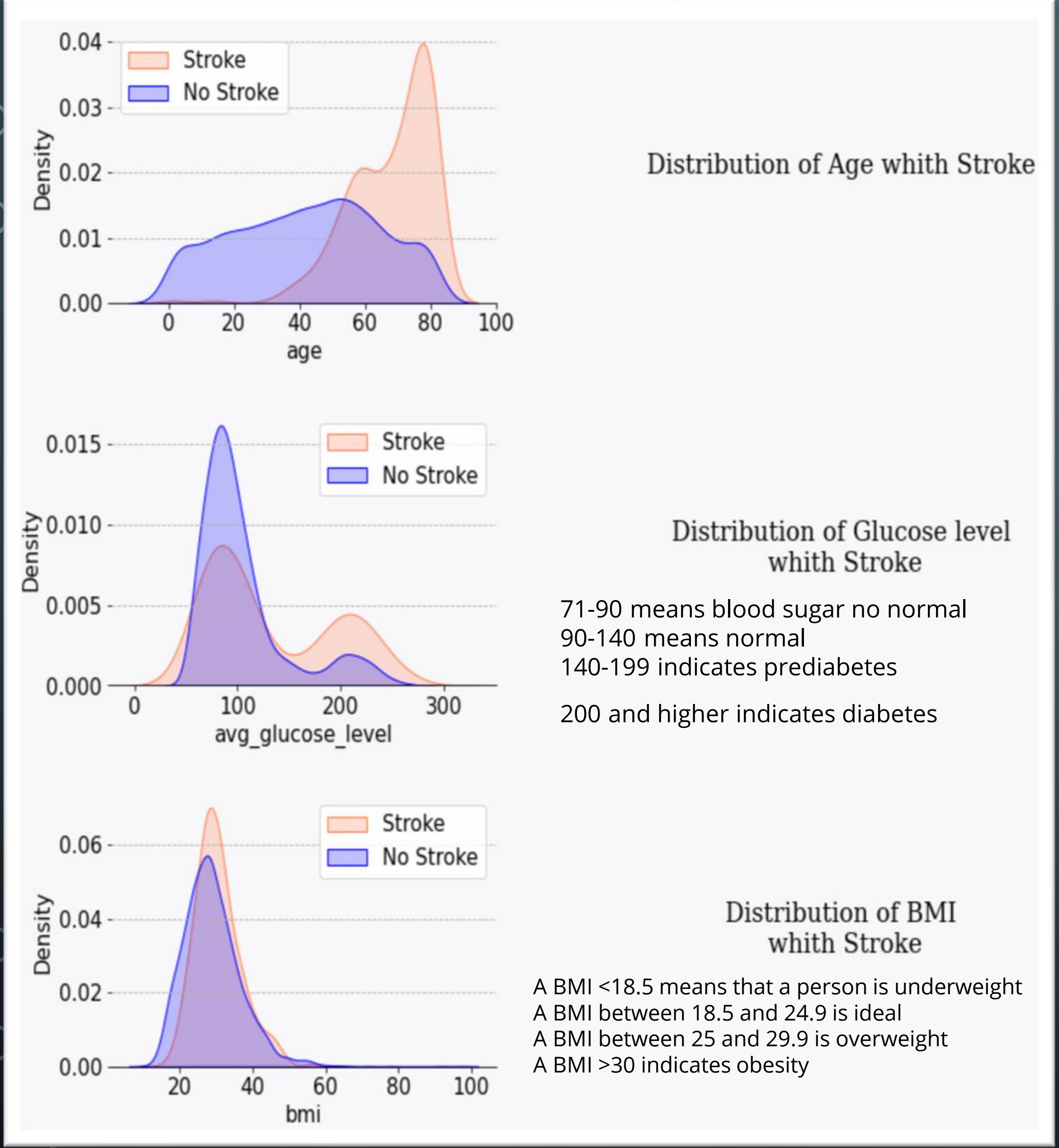
6) Does the type a job have an impact on the risk of stroke due to stress perhaps?

7) People living in urban areas have more chances of getting stroke?

8) Is the Glucose levels a factor for the risk of stroke?

9) Does the BMI have an impact in the risk of stroke?

10) Is the smoking status a high factor on the risk of stroke?



Distribution of Age whith Stroke

Distribution of Glucose level whith Stroke

71-90 means blood sugar no normal
90-140 means normal
140-199 indicates prediabetes

200 and higher indicates diabetes

Distribution of BMI whith Stroke

A BMI <18.5 means that a person is underweight
A BMI between 18.5 and 24.9 is ideal
A BMI between 25 and 29.9 is overweight
A BMI >30 indicates obesity

NUMERICAL FEATURES

- Age

Show high risk for stroke for older than 50

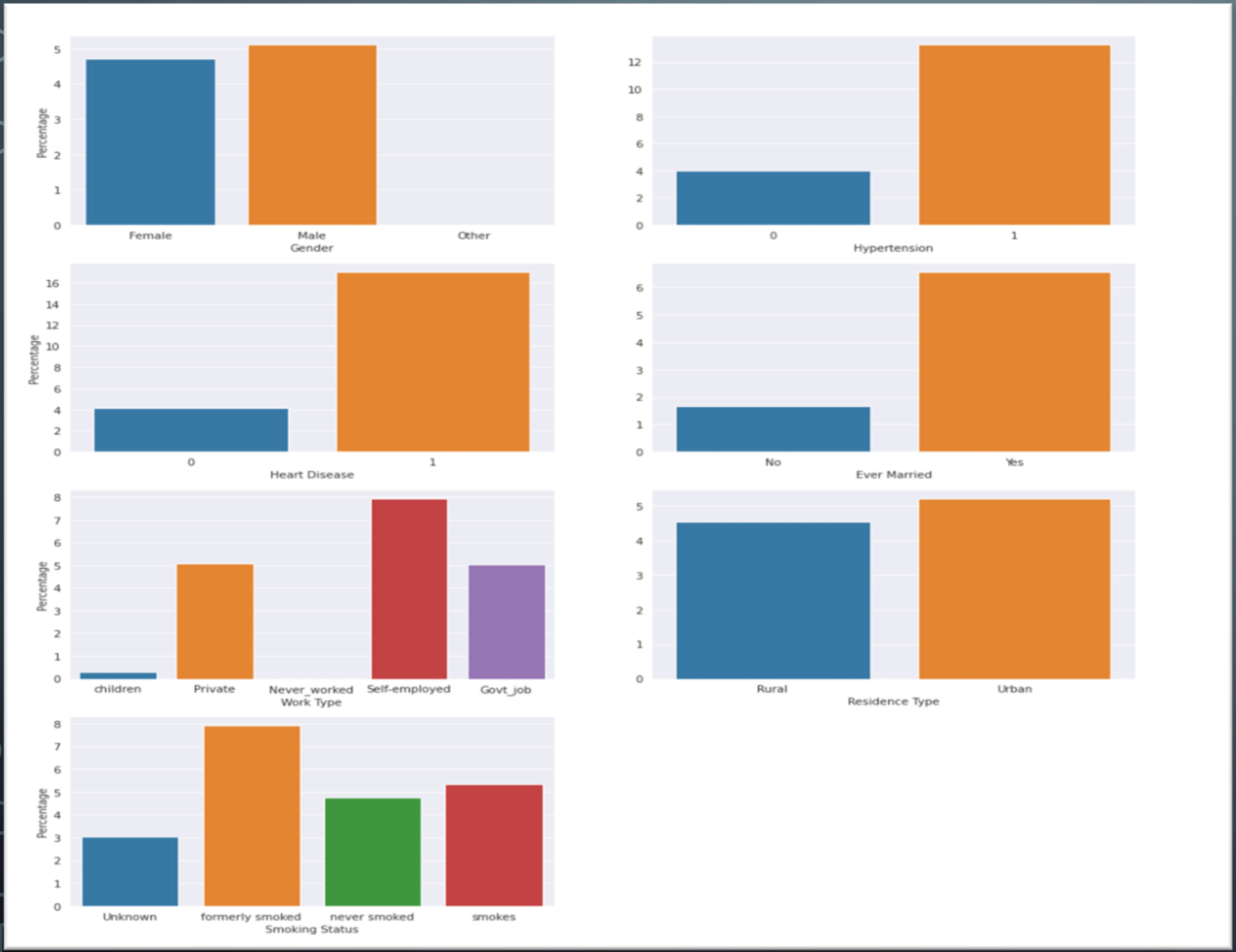
- Glucose Level

Risk for stroke for prediabetes and diabetes

- BMI

High BMI is not necessarily a significant factor?

There were 201 missing values for this feature.
Which corresponded to 16 percent of stroke data
(should not be removed)

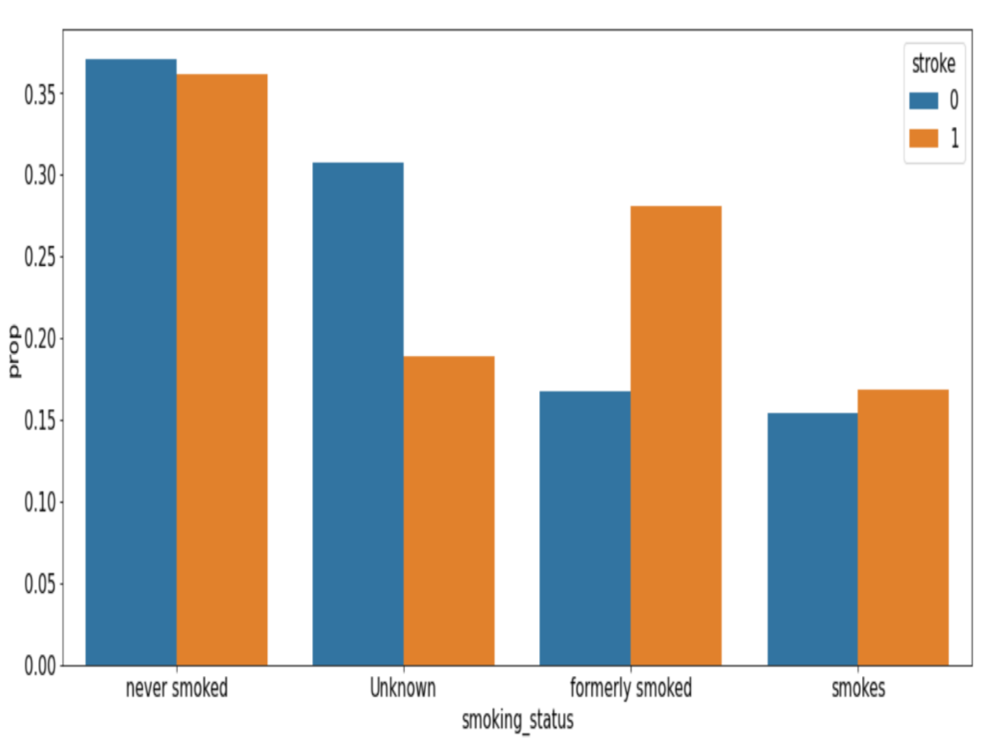


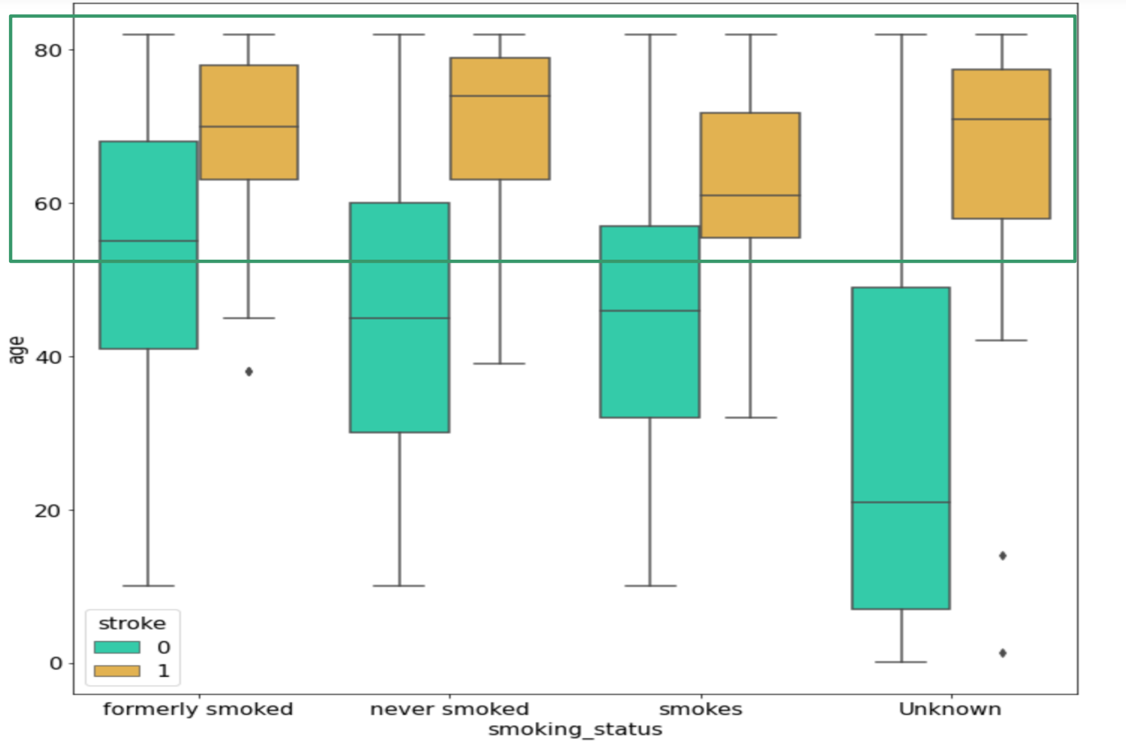
CATEGORICAL FEATURES PERCENTAGE VISUALIZATION

- Gender
- Heart Disease
- Work type
- Smoking Status
- Hypertension
- Ever married
- Residence type

SMOKING STATUS

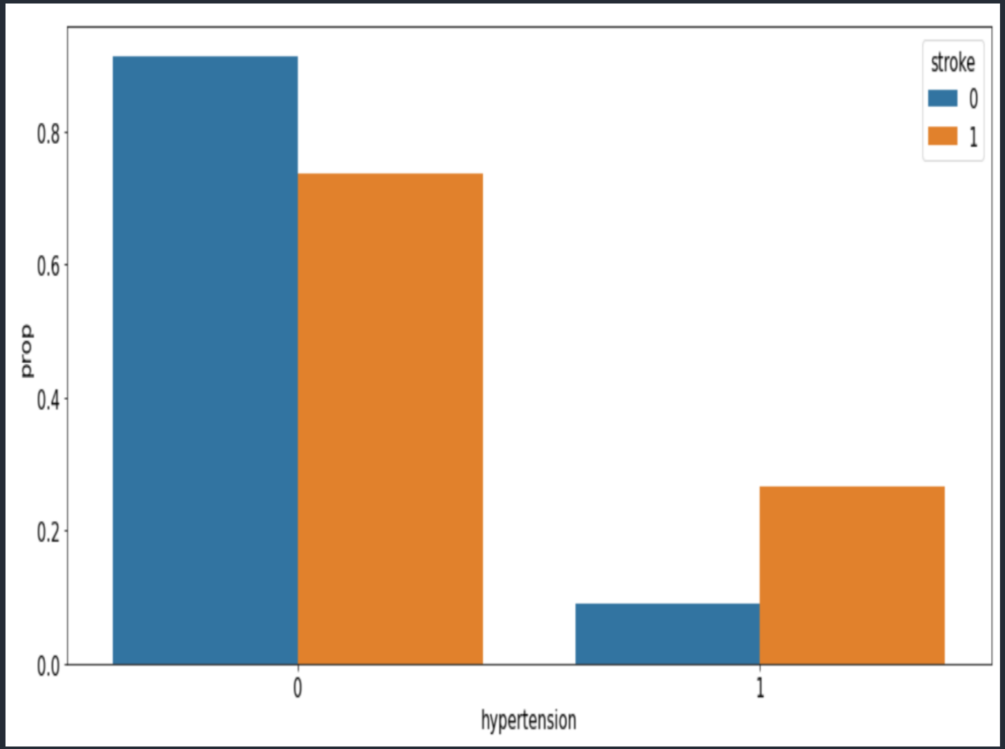
- For the people that smoke - a stroke will happen at an earlier age
- Formerly smoked - higher risk for same age range than never smoked
- Never smoked -stroke driven by age

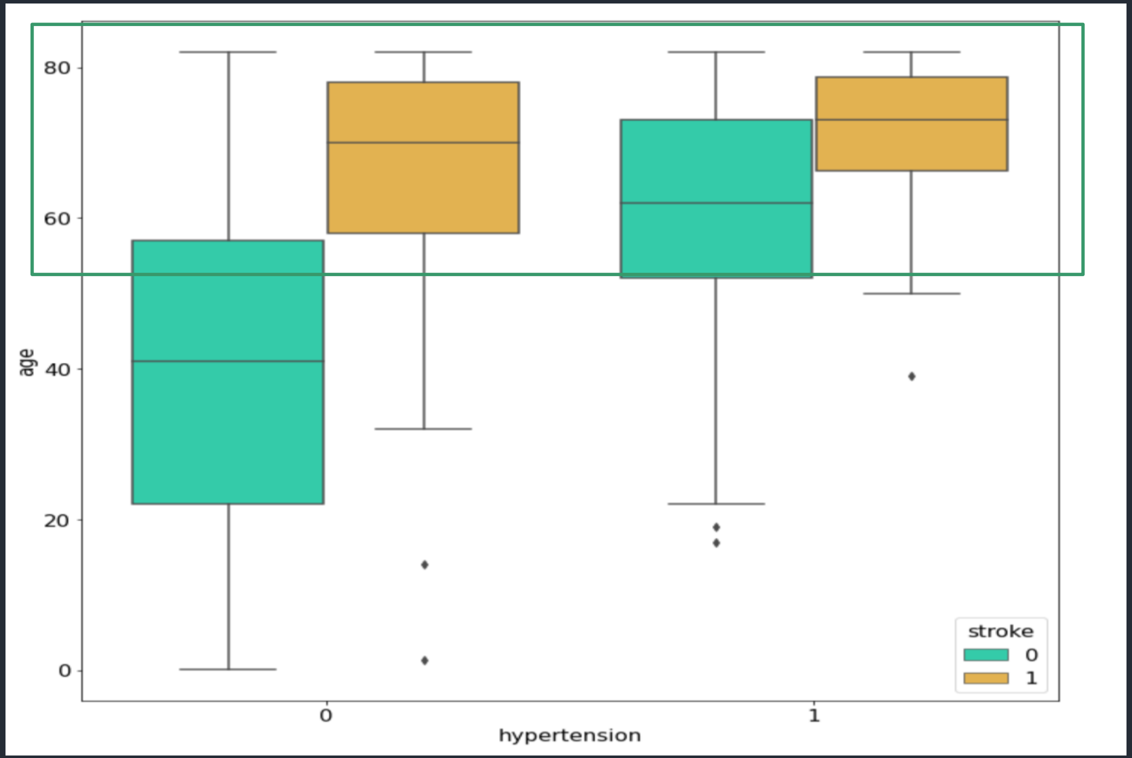




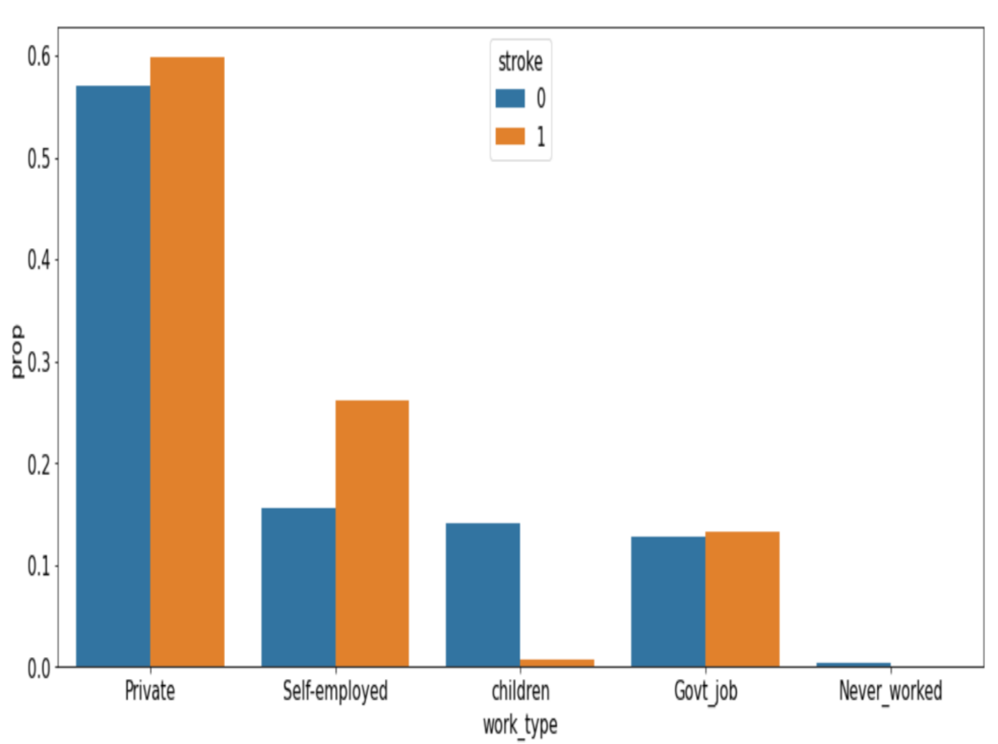
HYPERTENSION

- Hypertension occurs at older age
- Stroke might be driven by both age and hypertension (collinearity ?)



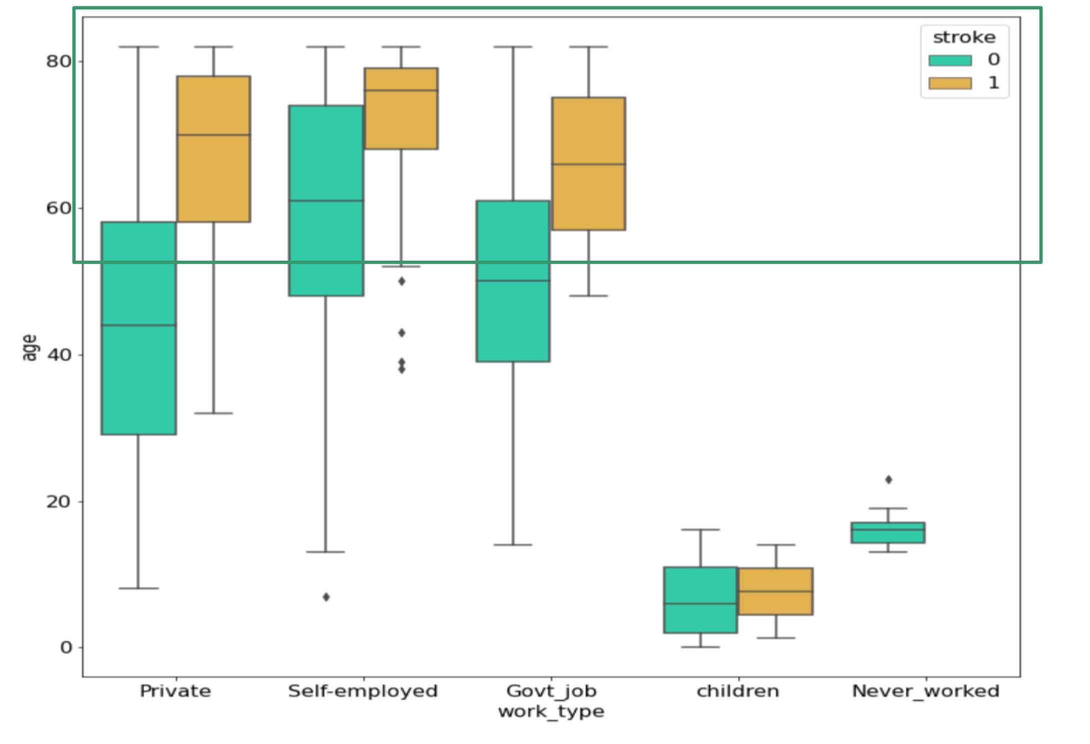


WORK TYPE



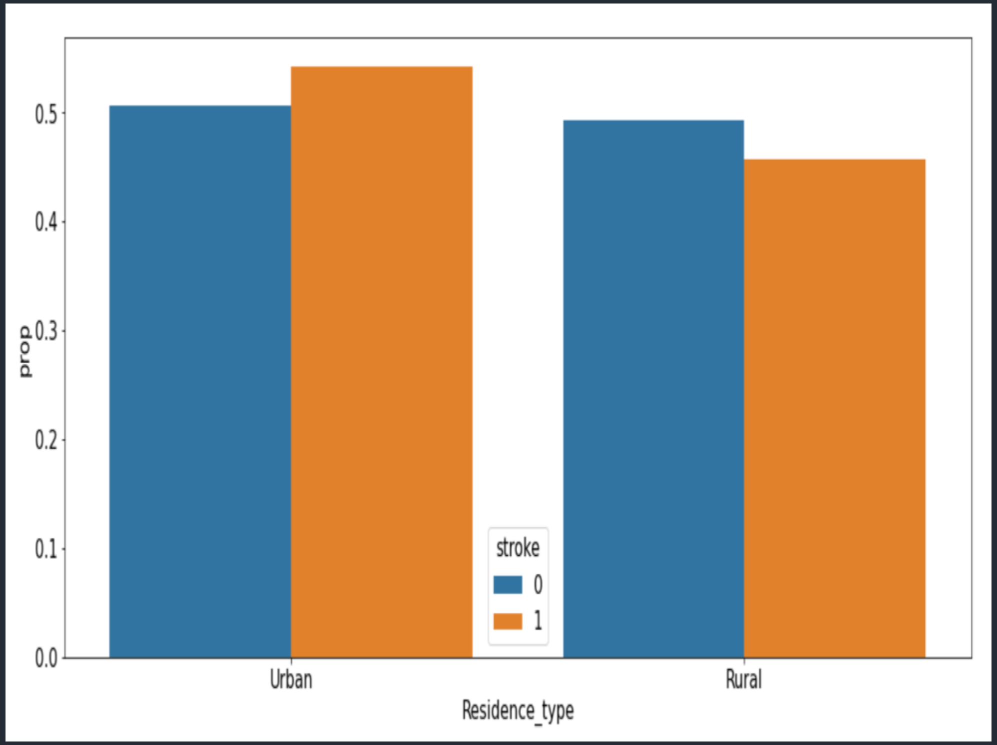
Work type

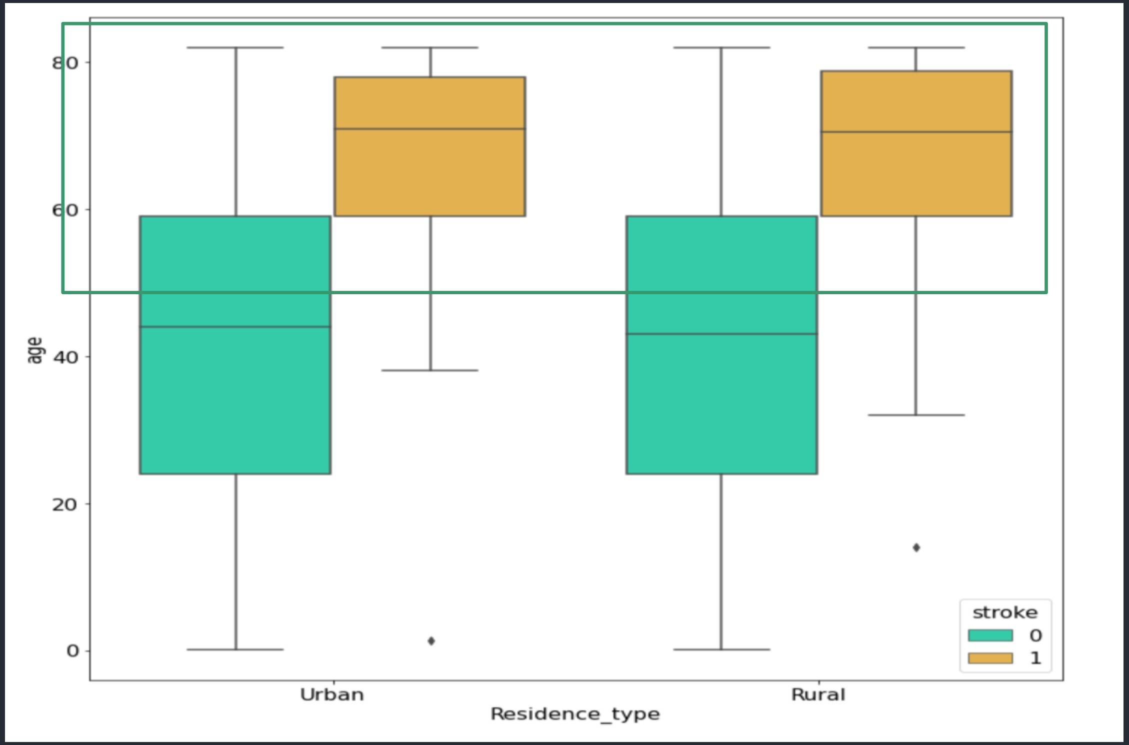
- The risk of stroke appears to be age related, not related to work type



RESIDENCE TYPE

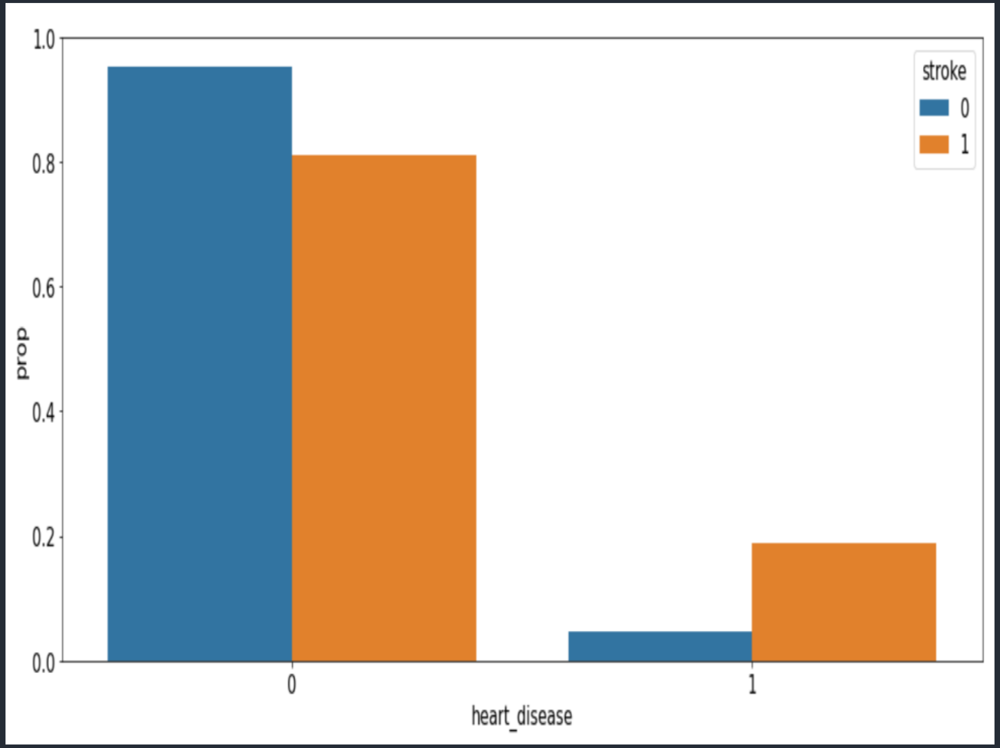
- Stroke in this case is only driven by age Urban or Rural has no impact in the risk of stroke

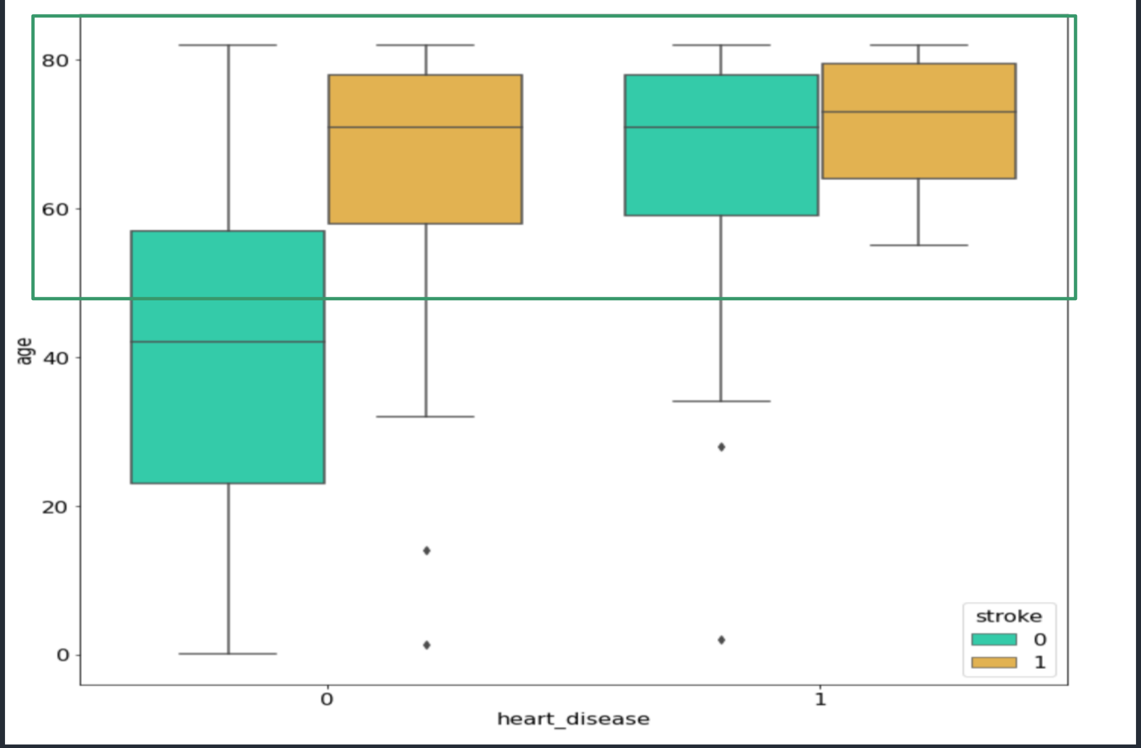




HEART DISEASE

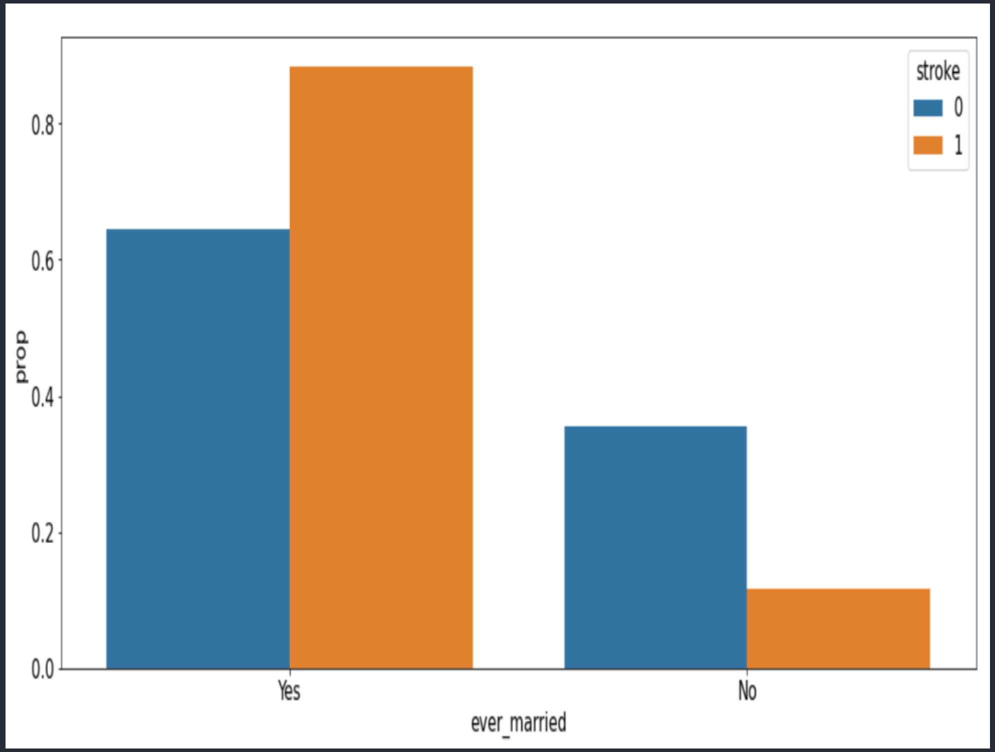
Heart disease and old age higher risk of stroke

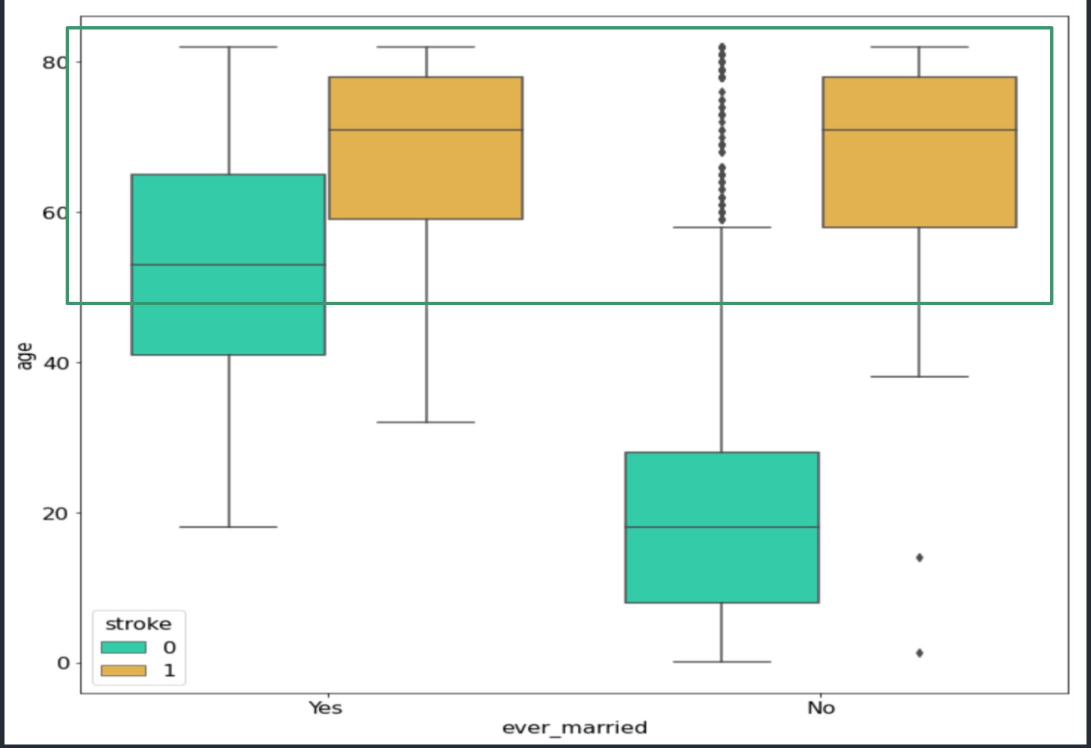




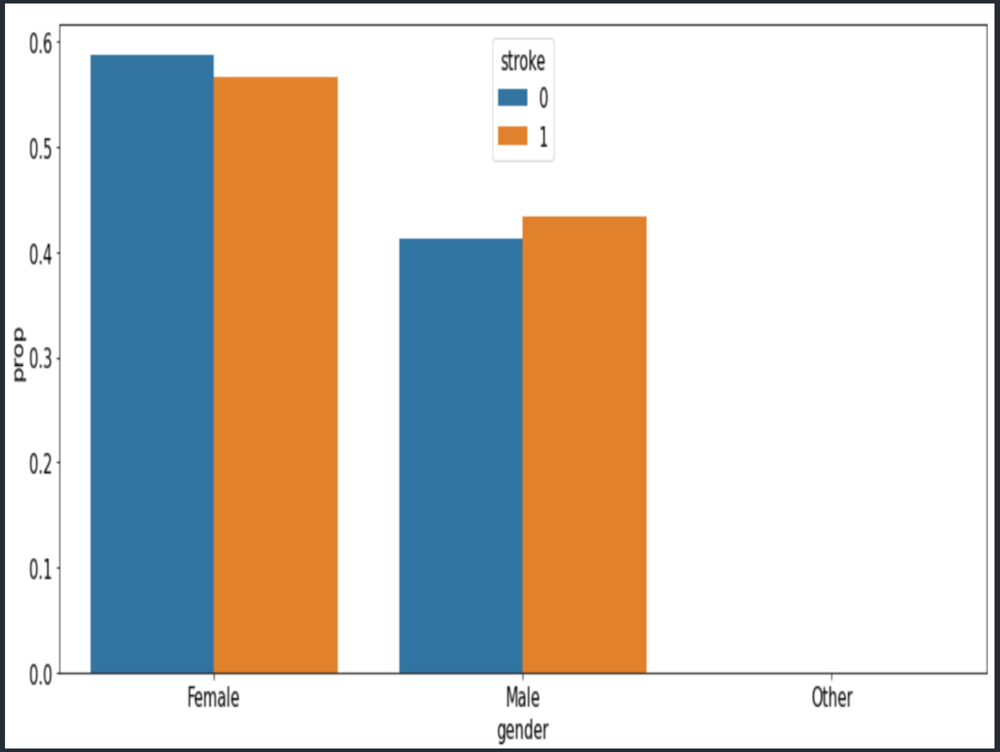
MARRIAGE STATUS

Marriage status does not show to be a factor on the risk of stroke.

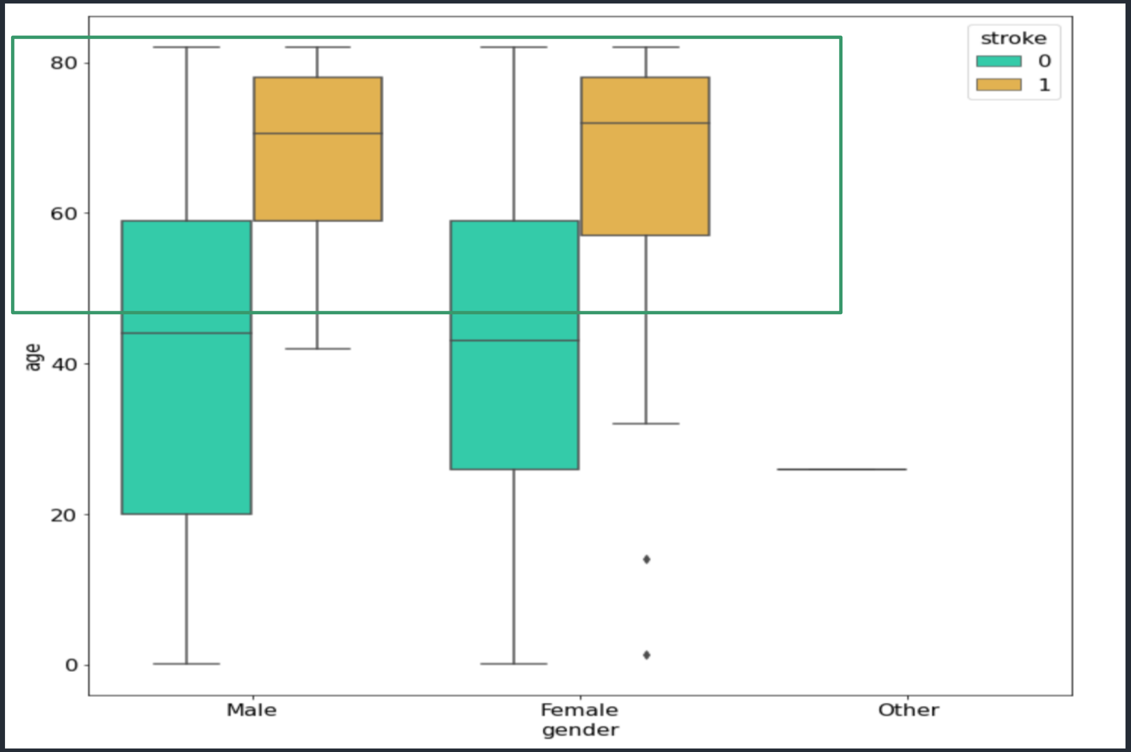


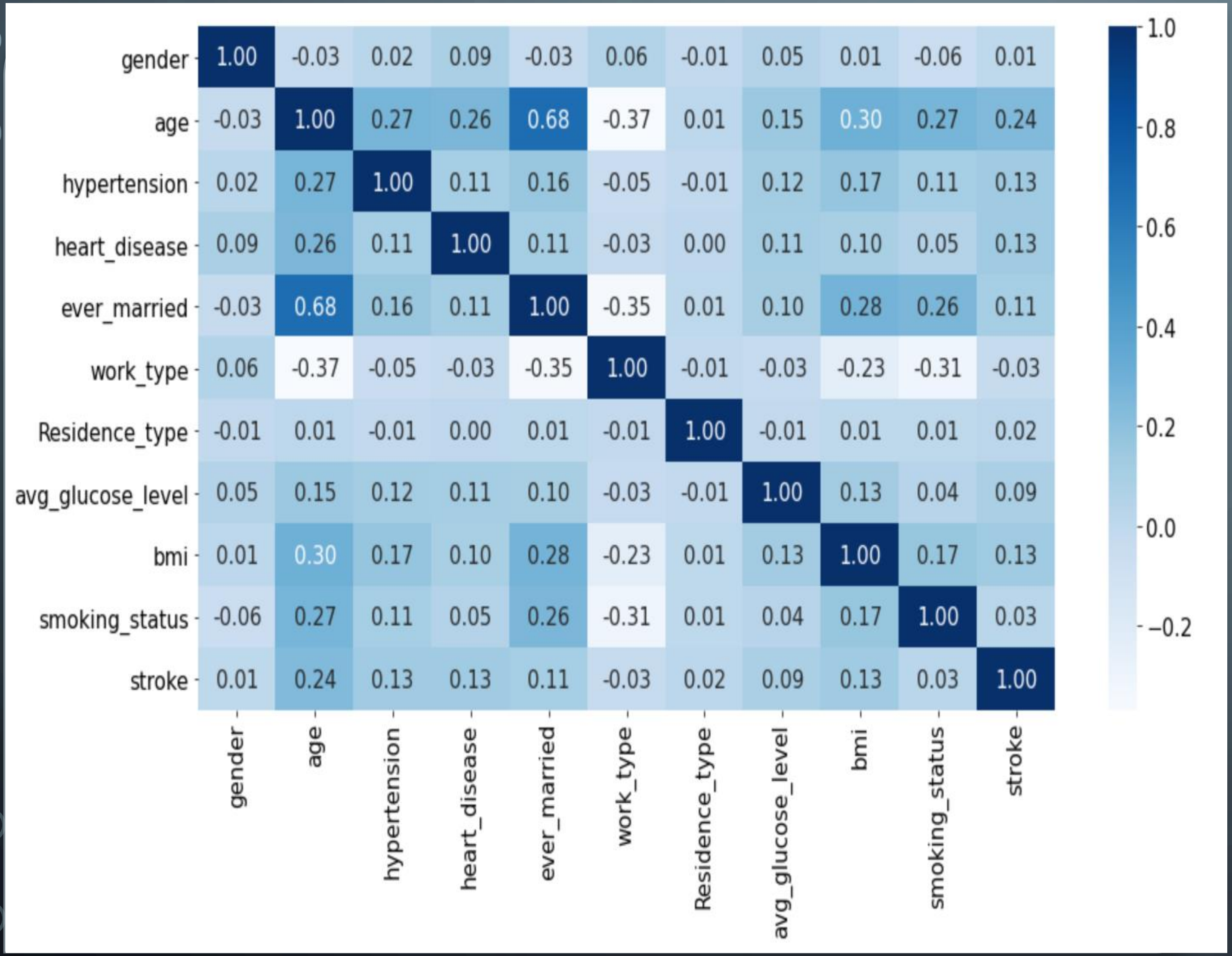


GENDER



Gender is not a strong factor to determine risk for stroke.

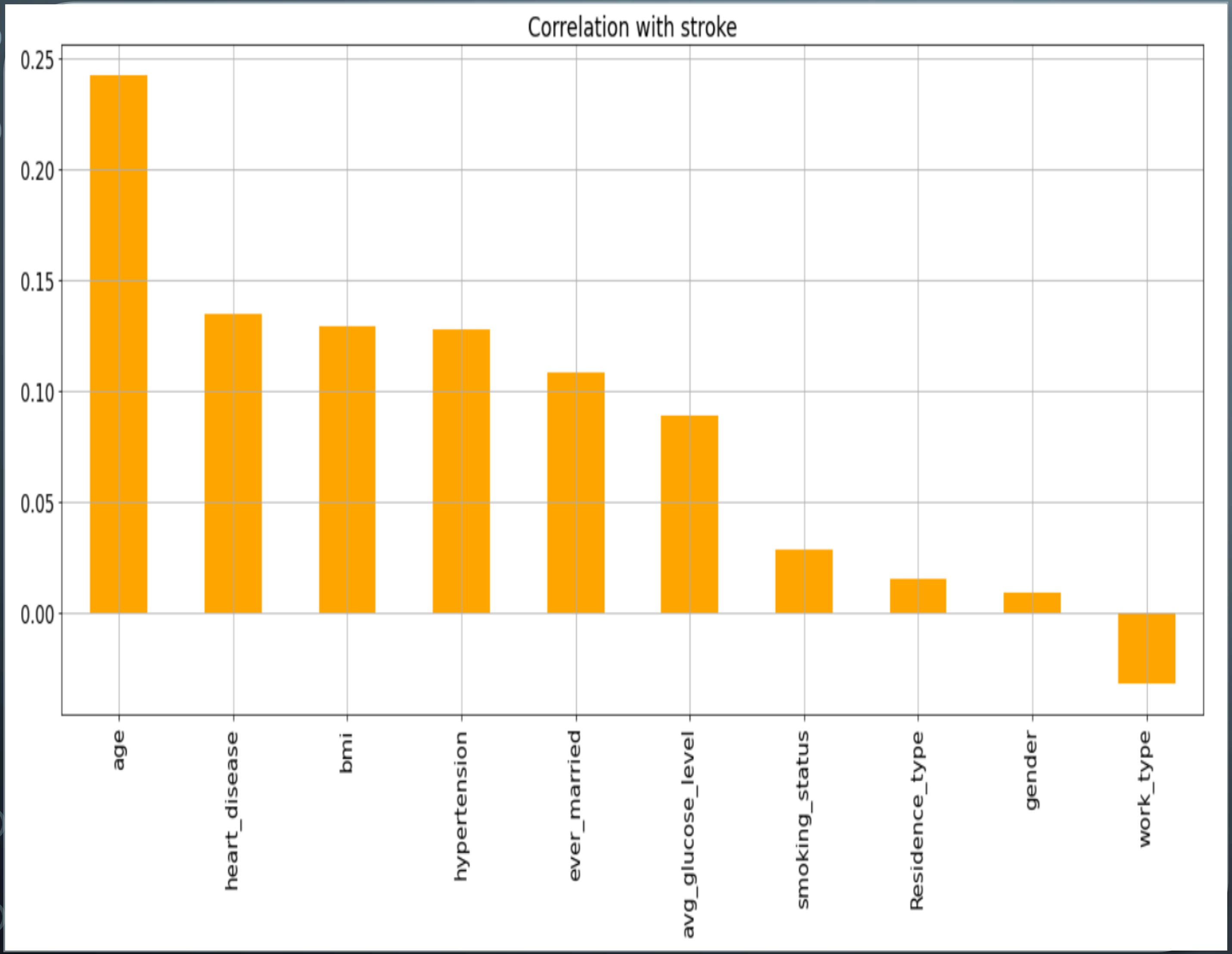




	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
gender	1.00	-0.03	0.02	0.09	-0.03	0.06	-0.01	0.05	0.01	-0.06	0.01
age	-0.03	1.00	0.27	0.26	0.68	-0.37	0.01	0.15	0.30	0.27	0.24
hypertension	0.02	0.27	1.00	0.11	0.16	-0.05	-0.01	0.12	0.17	0.11	0.13
heart_disease	0.09	0.26	0.11	1.00	0.11	-0.03	0.00	0.11	0.10	0.05	0.13
ever_married	-0.03	0.68	0.16	0.11	1.00	-0.35	0.01	0.10	0.28	0.26	0.11
work_type	0.06	-0.37	-0.05	-0.03	-0.35	1.00	-0.01	-0.03	-0.23	-0.31	-0.03
Residence_type	-0.01	0.01	-0.01	0.00	0.01	-0.01	1.00	-0.01	0.01	0.01	0.02
avg_glucose_level	0.05	0.15	0.12	0.11	0.10	-0.03	-0.01	1.00	0.13	0.04	0.09
bmi	0.01	0.30	0.17	0.10	0.28	-0.23	0.01	0.13	1.00	0.17	0.13
smoking_status	-0.06	0.27	0.11	0.05	0.26	-0.31	0.01	0.04	0.17	1.00	0.03
stroke	0.01	0.24	0.13	0.13	0.11	-0.03	0.02	0.09	0.13	0.03	1.00

CORRELATION MATRIX

- No strong correlation between any of the features except age and ever_married



CORRELATION WITH THE TARGET

This visualization shows that age is the main factor for stroke.

Smoking doesn’t show a high correlation which is troublesome as it is well known that smoking is a strong risk factor for stroke

CONCLUSION

- Age is a strong driving factor for the risk of stroke.
- BMI was not highly correlated to risk of Stroke which is surprising?
- Diabetes is a risk factor for stroke.
- Smoking status is a factor to determine the risk of having a stroke.
- Attributes like Heart disease, hypertension, ever married and employment status pointed more to be age related risk for stroke .
- Residence type has no correlation with the risk of stroke.
- Gender does not appear to be a strong factor for stroke in this dataset although according to several studies Women suffer about 55000 more strokes each year than men.

CONCLUSION (CONT.)

Data set concerns:

- The data set is imbalanced – only 4.9 percent belongs to stroke.
- Some of the features such work type are ambiguous (children is this working with children?), is this related to stress?

Future Work

- Balance the data.
- Perhaps limit the age range for the study .
- Find a Better dataset?