Stroke Events: Data Preparation and Preliminary Analysis

UNC GREENSBORO

Background & Research Goals

- According to the World Health Organization (WHO), stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths.

- We will use Exploratory and Preliminary Data Analysis to answer the research questions that are posed in next slide.

UNC GREENSBORO

Research Questions ????

- Is the myth true, males have more strokes than females?
- Does age have a direct impact on stroke?
- People having hypertension are more prone to stroke!
- Common understanding is that a person with heart disease will suffer a stroke most of the time!
- Do married people get strokes more than unmarried people?
- Do people working in the private sector have a higher chance of stroke?
- Are people living in the city at higher risk of stroke?
- Is there a relationship between glucose level or bmi and stroke?
- Are smokers at more risk of getting a stroke?

UNC GREENSBORO

Dataset and Description of Variables

❖Dataset Link: Stroke Prediction Dataset | Kaggle

❖Numerical Features:

- Age
- BMI
- Avg_glucose_level

❖Categorical Features:

- gender
- hypertension
- heart_disease
- ever_married
- work_type
- stroke(target)
- smoking_status
- Residence_type

❖Total Observations: 5110

UNC GREENSBORO

Data Cleaning and Visualizations

Count number of unique values:

gender	3
age	104
hypertension	2
heart_disease	2
ever_married	2
work_type	5
Residence_type	2
avg_glucose_level	3979
bmi	418
smoking_status	4
stroke	2

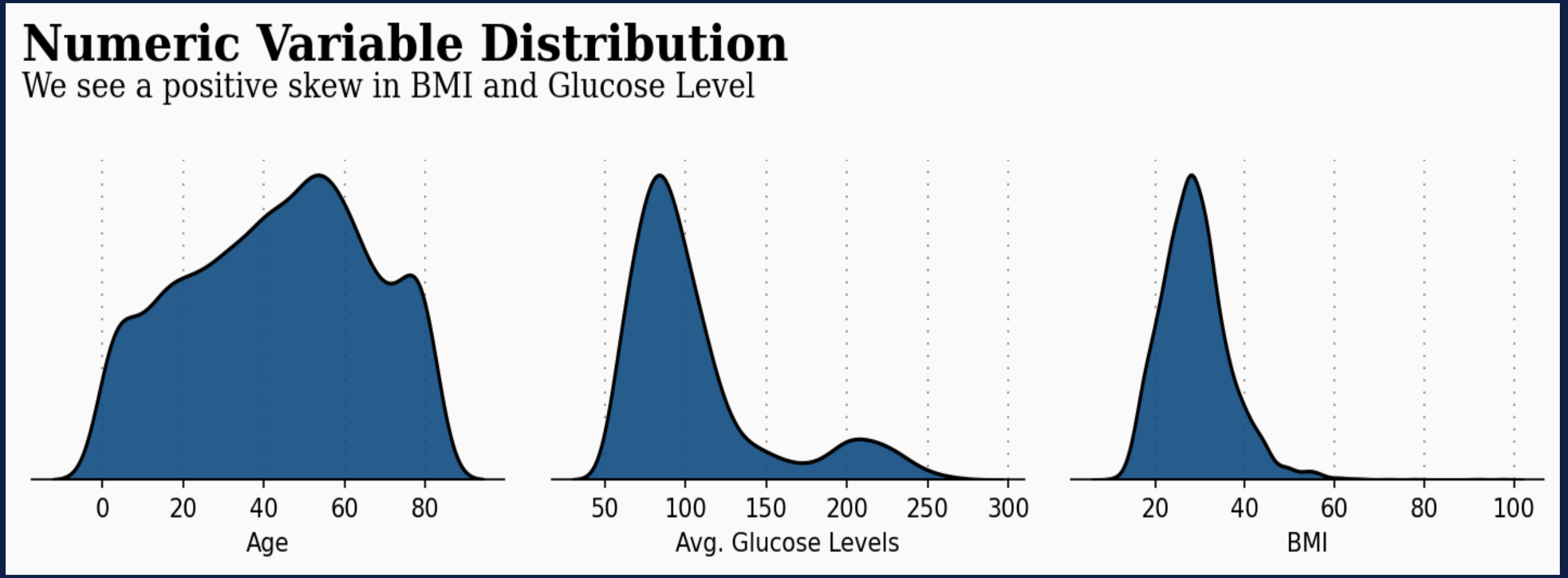
- **No duplicate rows**
- **Checked missing values:**

	Missing_Number	Missing_Percent
bmi	201	3.933464
id	0	0.000000
gender	0	0.000000
age	0	0.000000
hypertension	0	0.000000
heart_disease	0	0.000000
ever_married	0	0.000000
work_type	0	0.000000
Residence_type	0	0.000000
avg_glucose_level	0	0.000000
smoking_status	0	0.000000
stroke	0	0.000000

- **Replaced missing BMI values by median since this is more robust to outliers compared to mean.**

UNC GREENSBORO

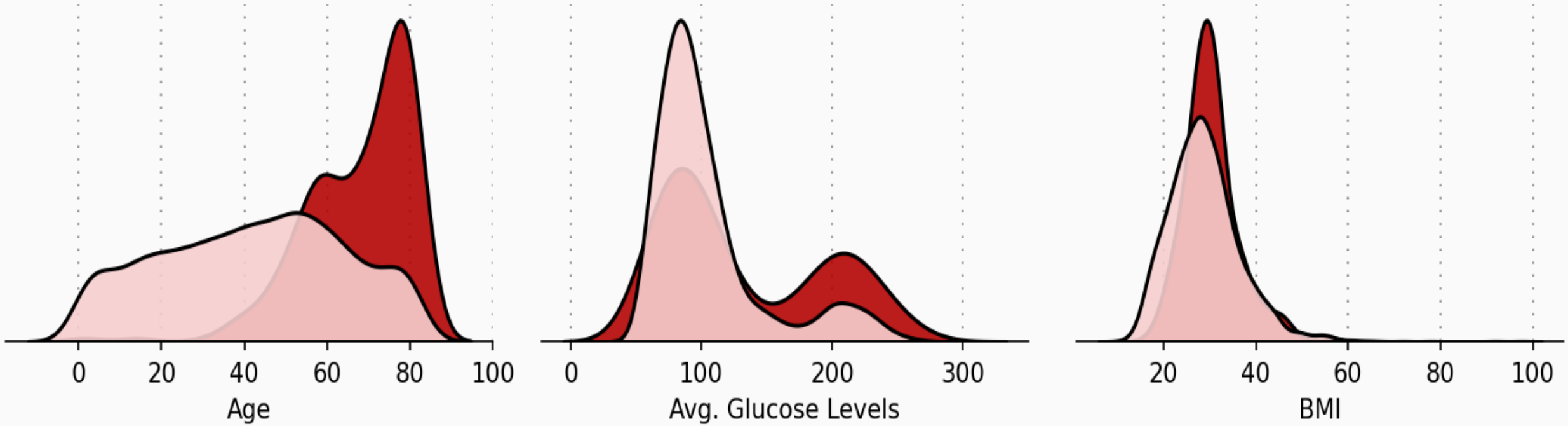
Explore Numerical Features



Explore Numerical Features

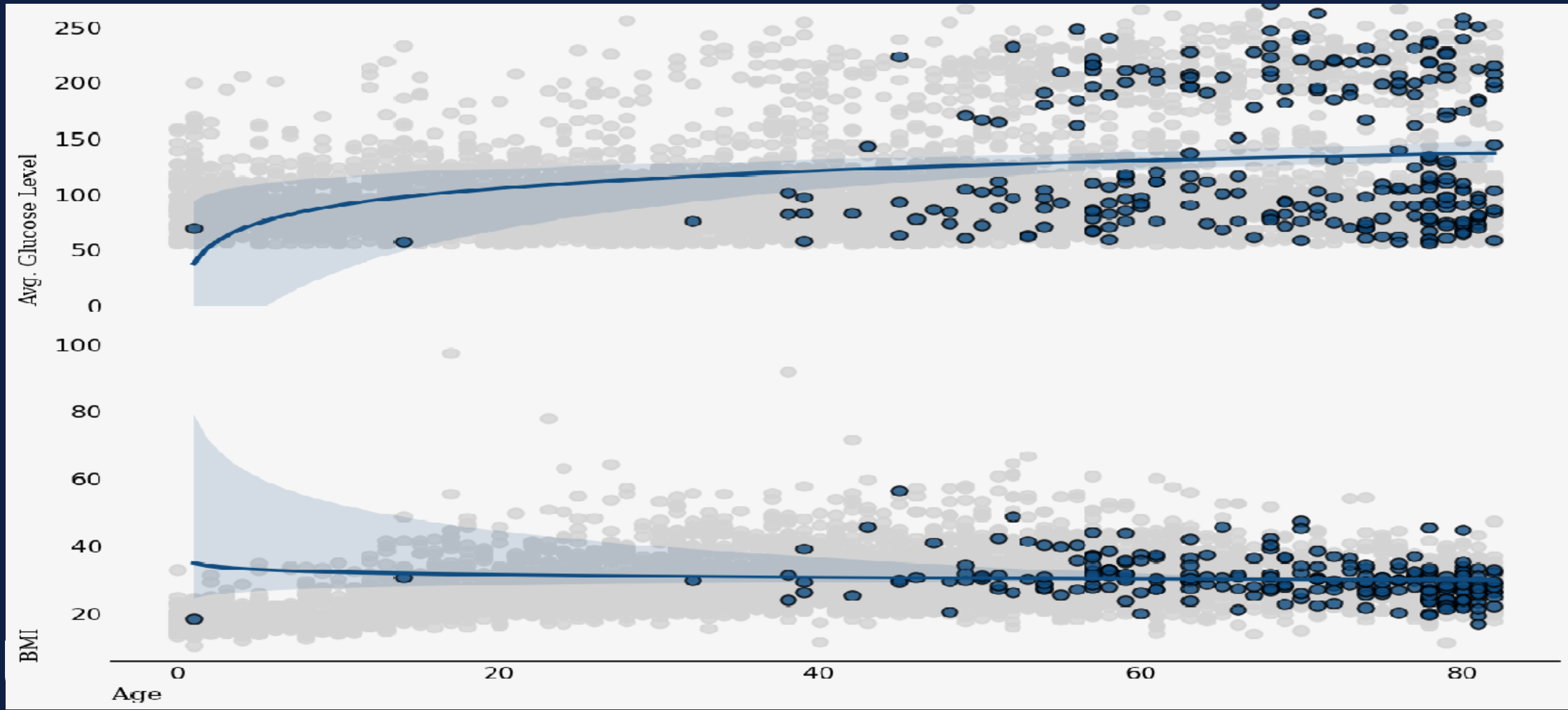
Numeric Variables by Stroke & No Stroke

Age looks to be a prominent factor

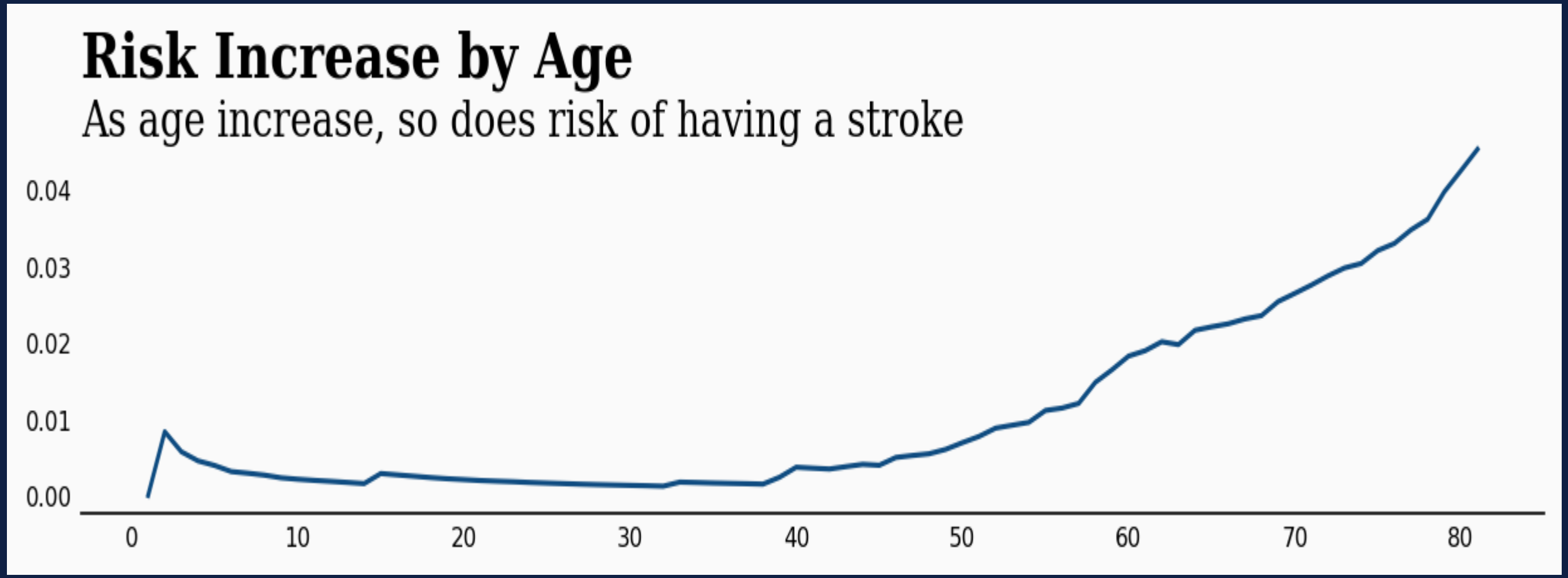


UNC GREENSBORO

Strokes by Age, Glucose level, and BMI

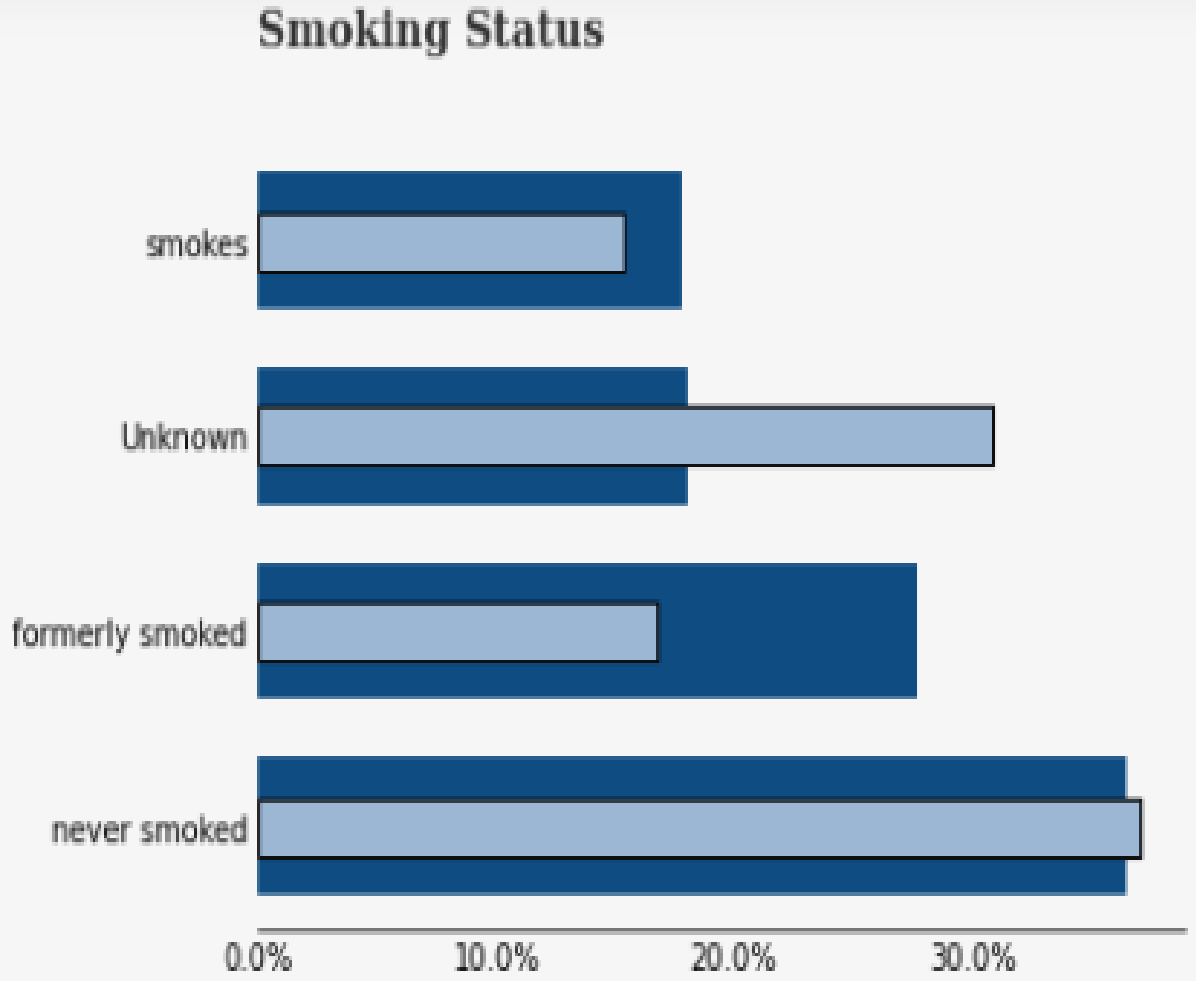


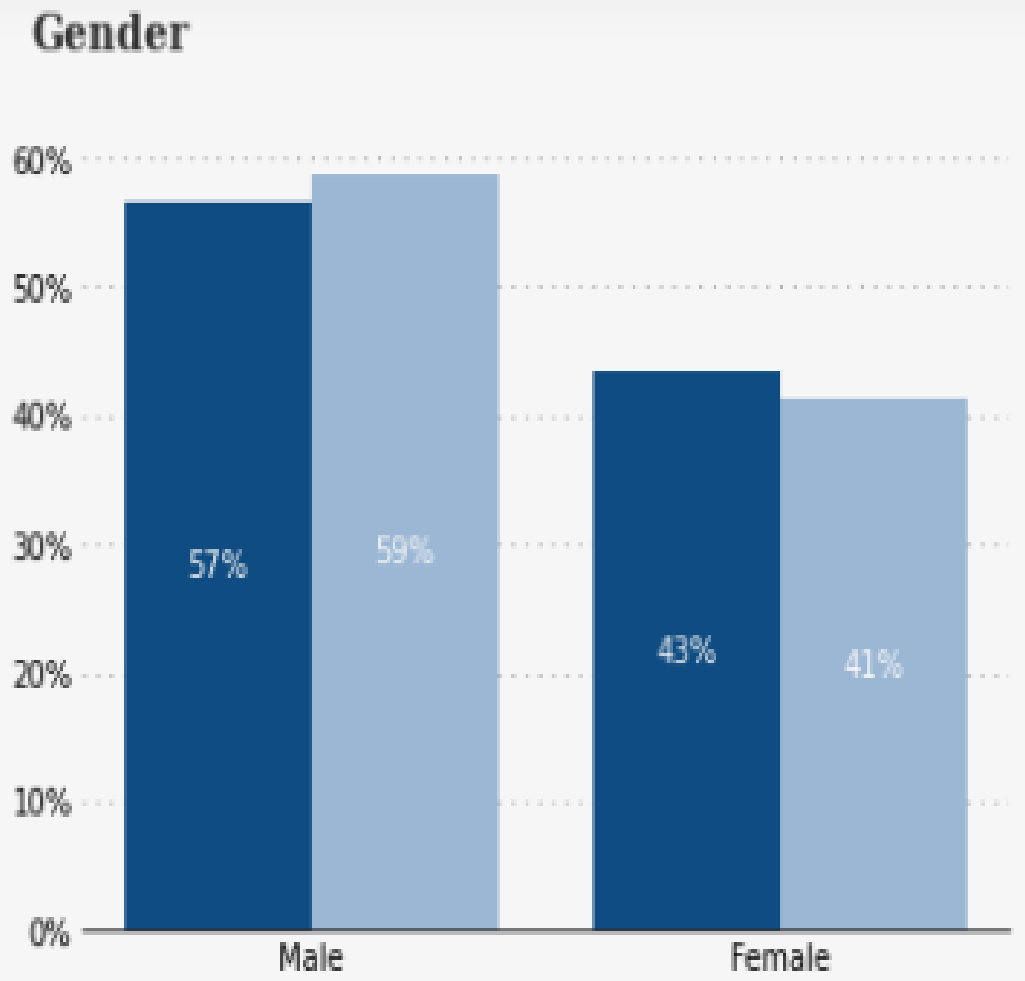
Age Versus Stroke



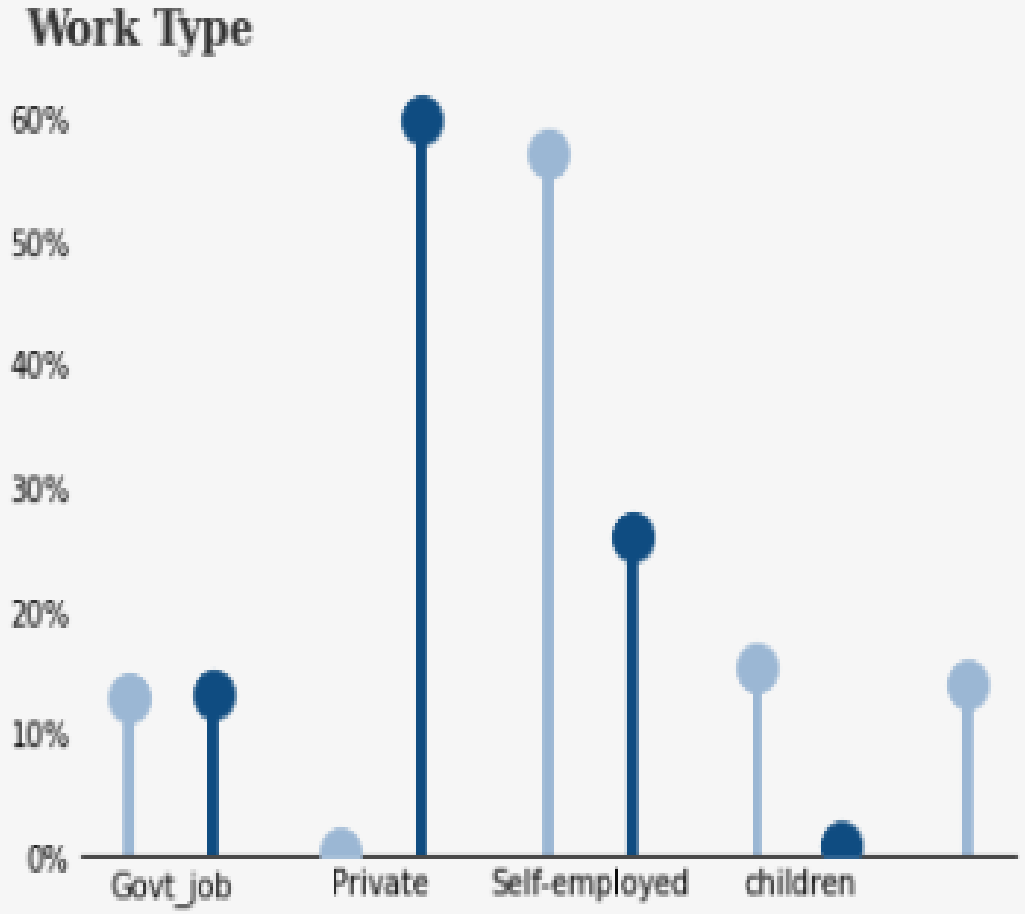
UNC GREENSBORO

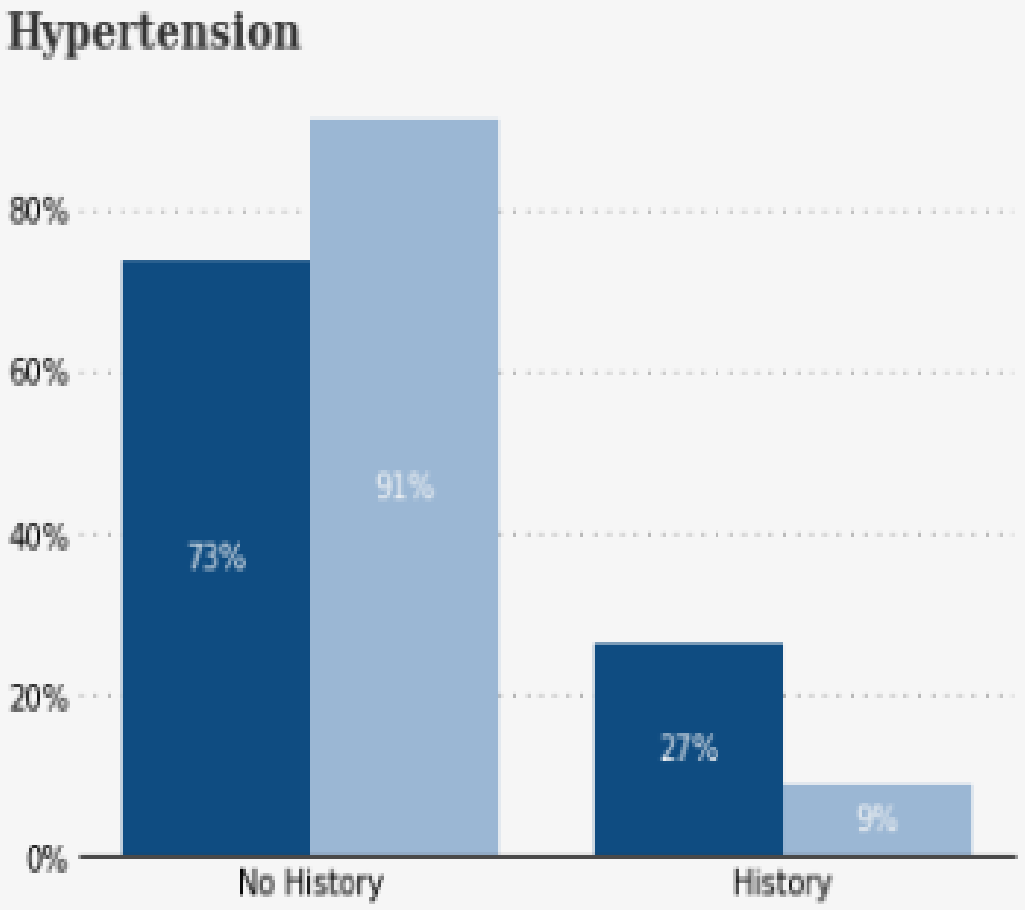
Explore Categorical Features





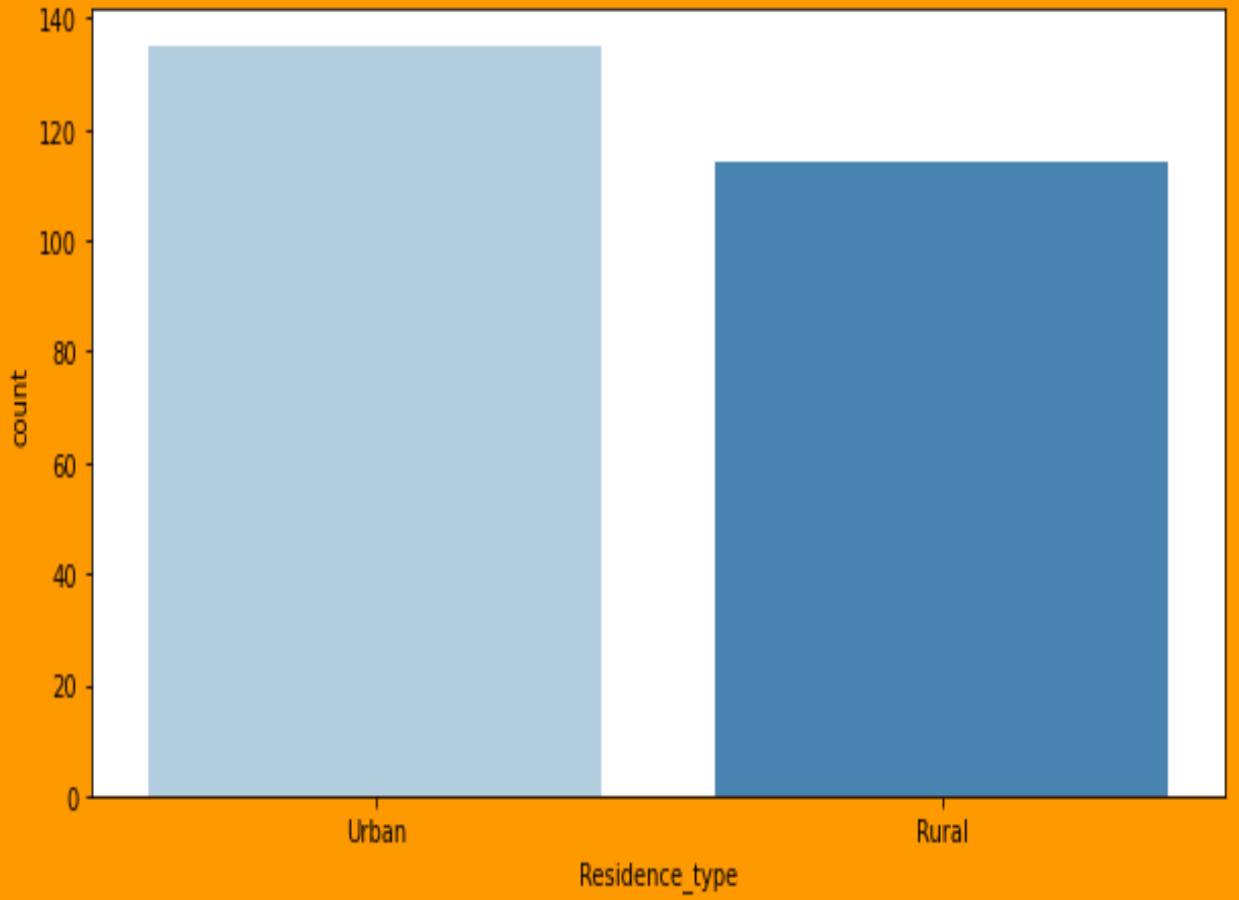
Work Type & Hypertension

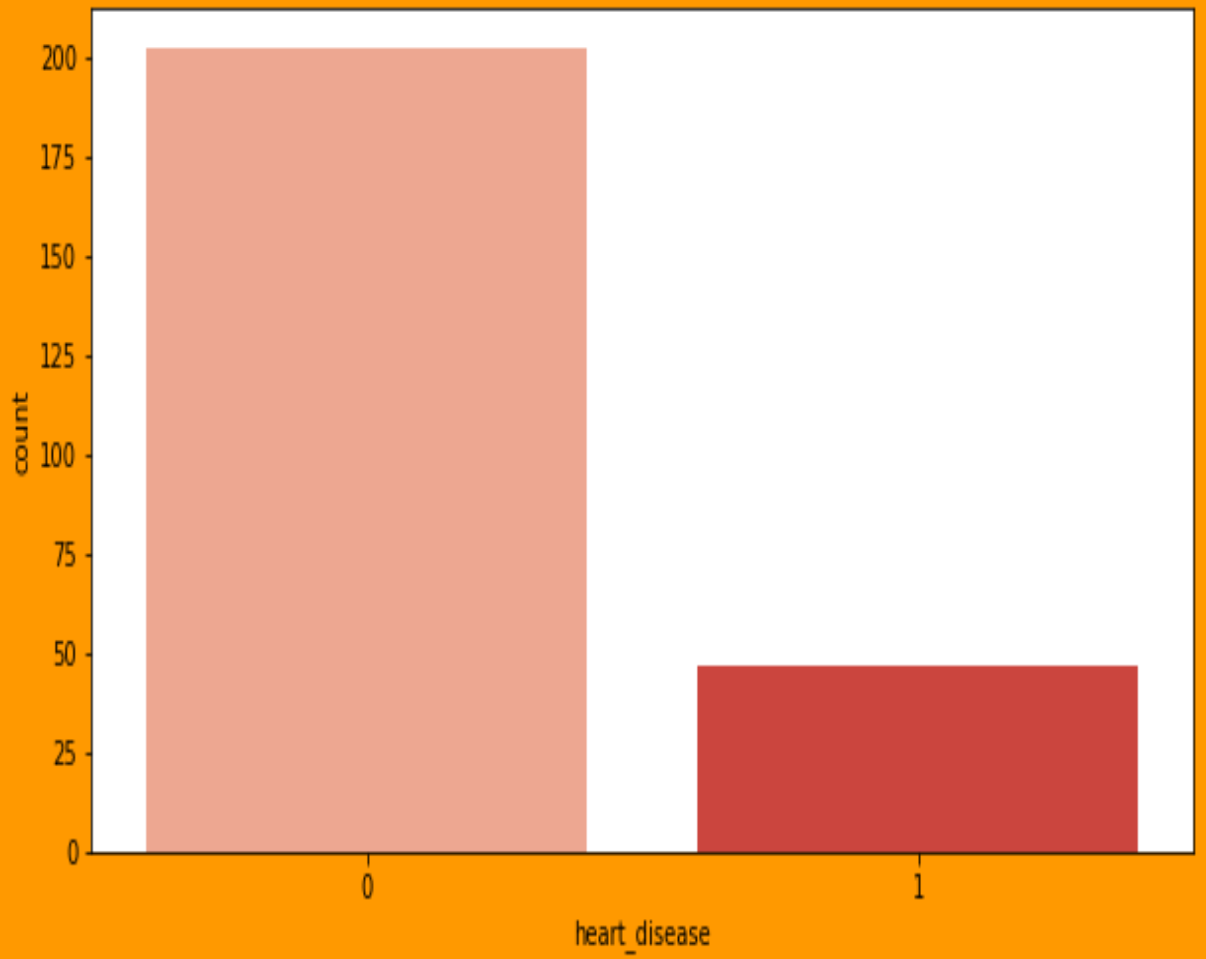




UNC GREENSBORO

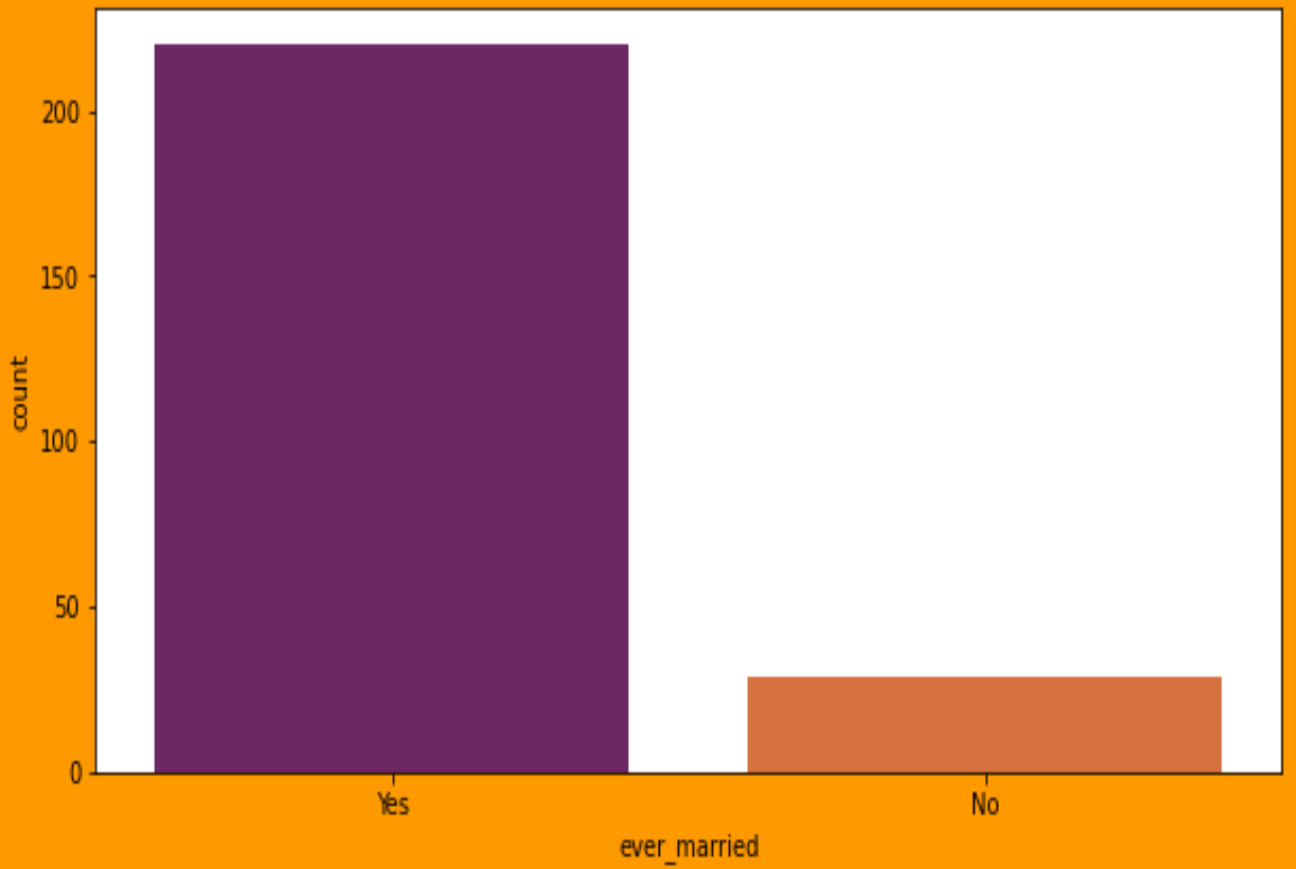
Residence Type & Heart Disease





UNC GREENSBORO

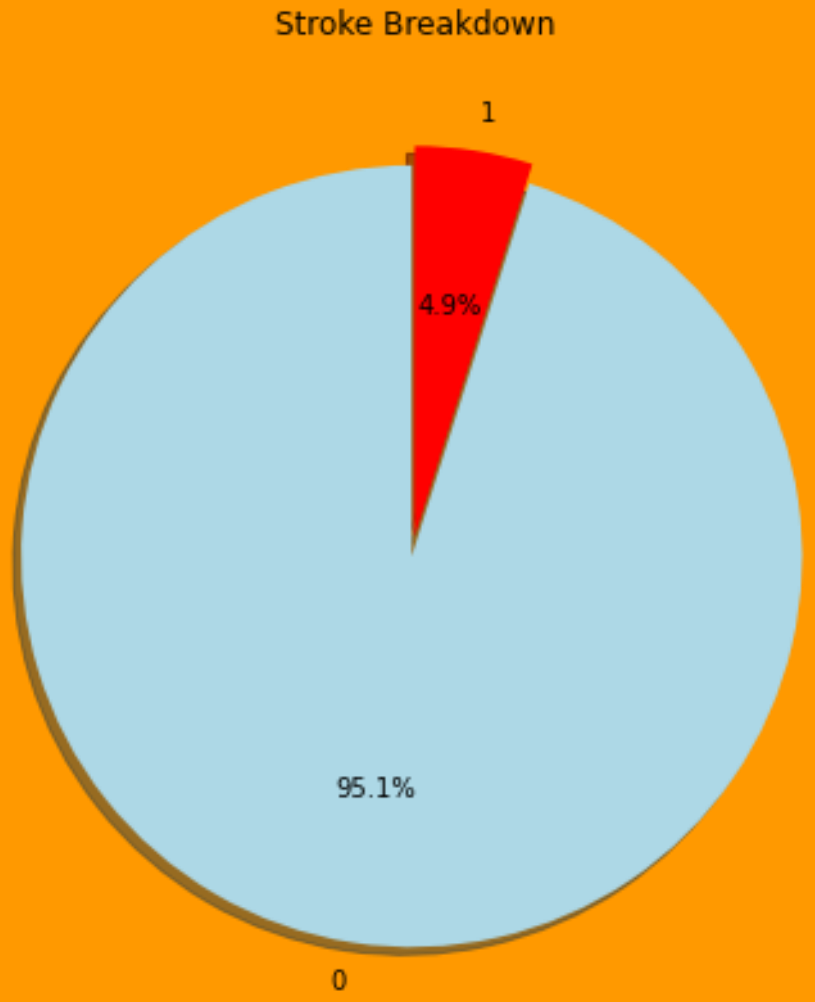
Stroke vs Marital Status

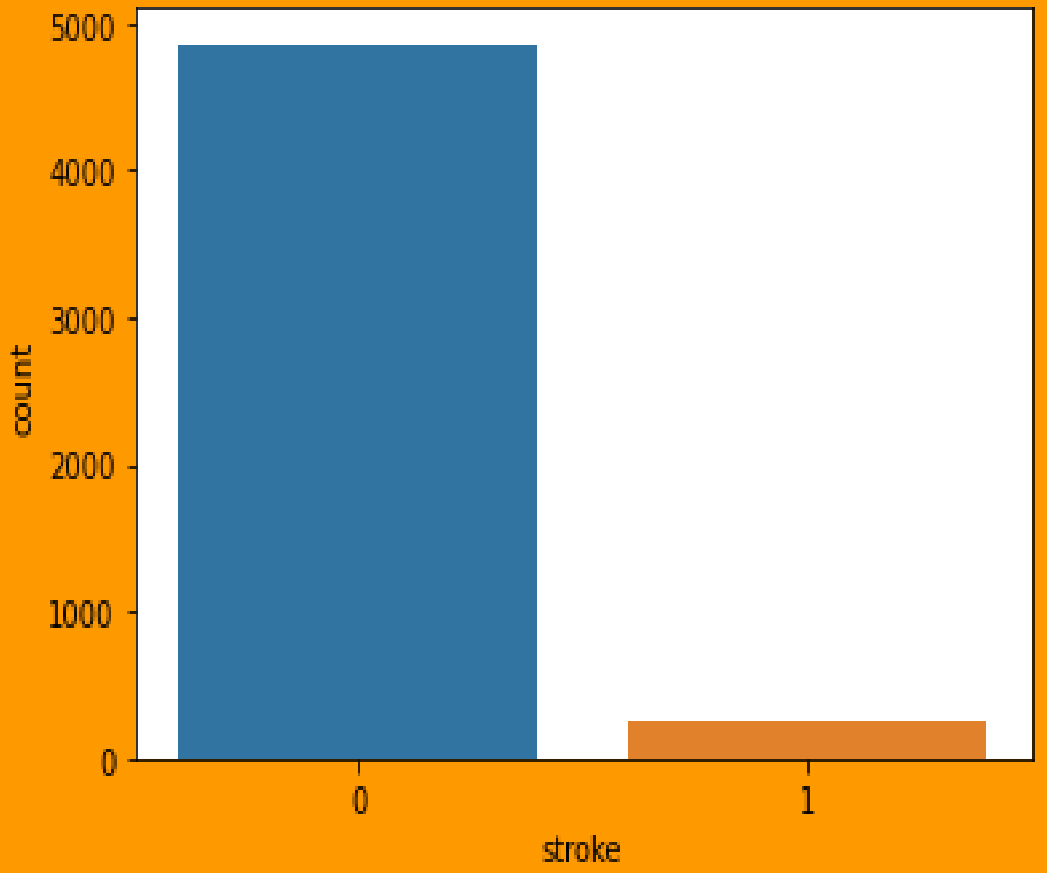


That was expected, wasn't it 👀

UNC GREENSBORO

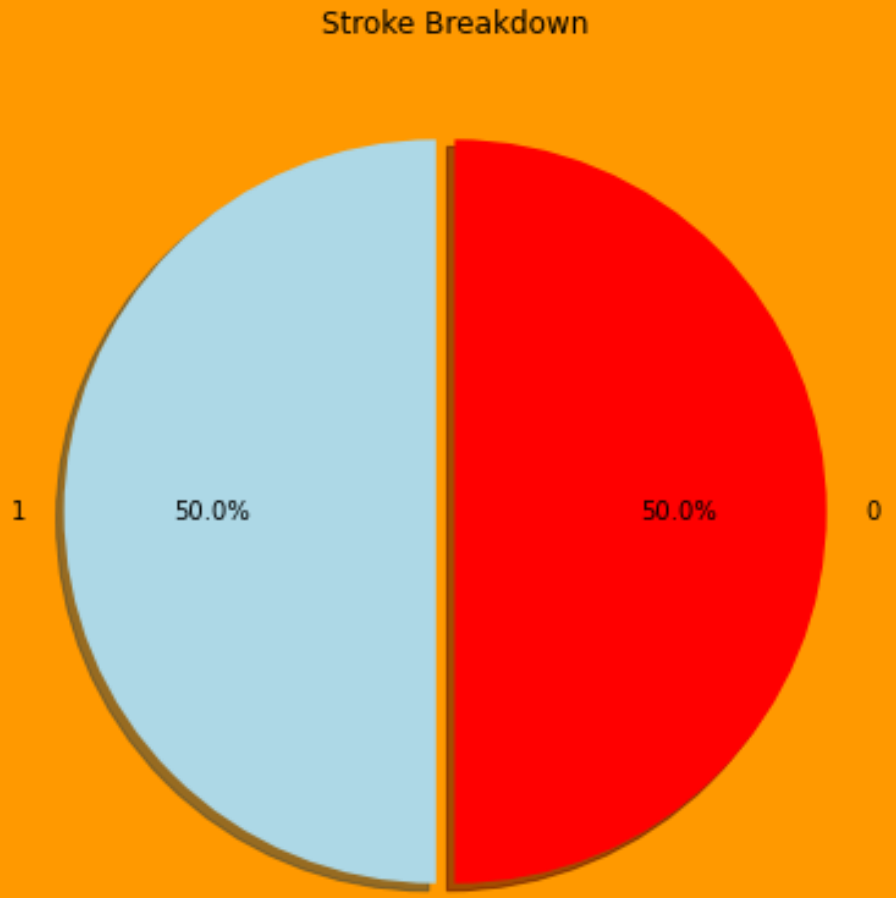
Stroke Breakdown

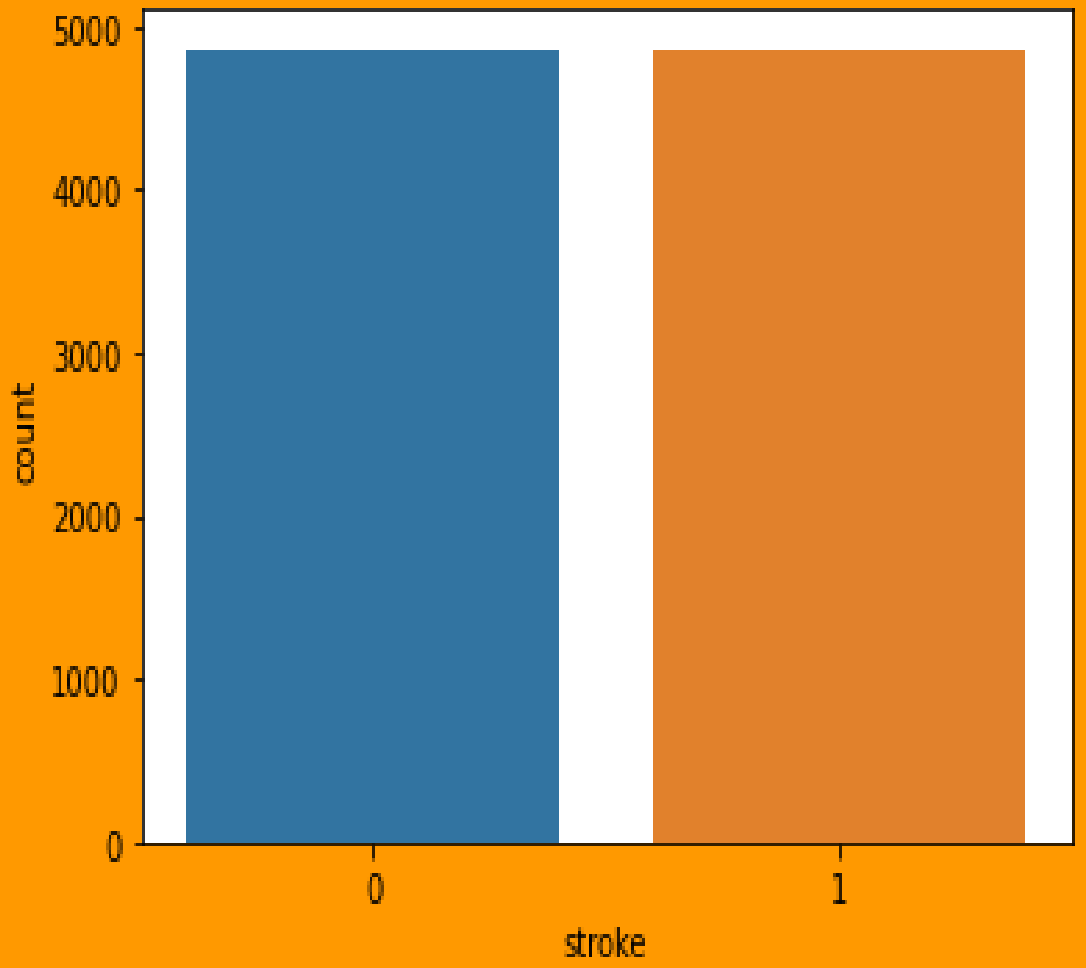




UNC GREENSBORO

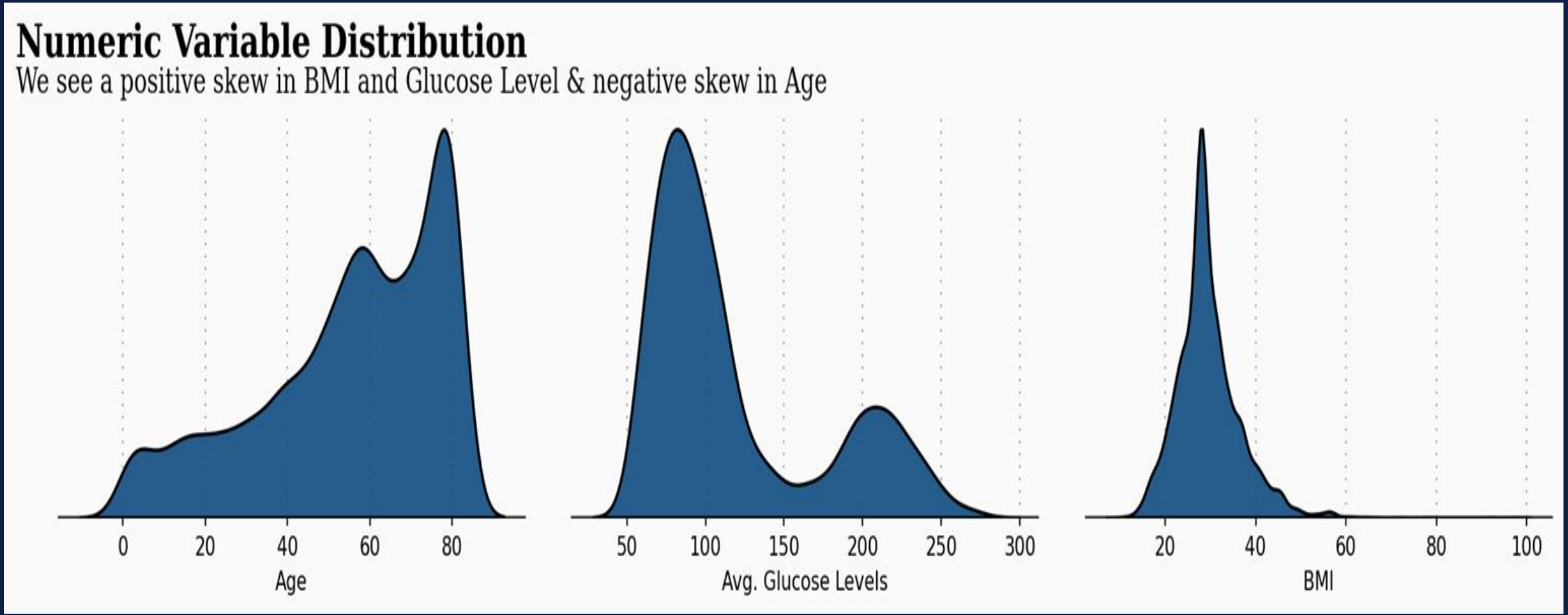
Stroke Breakdown after Upsampling





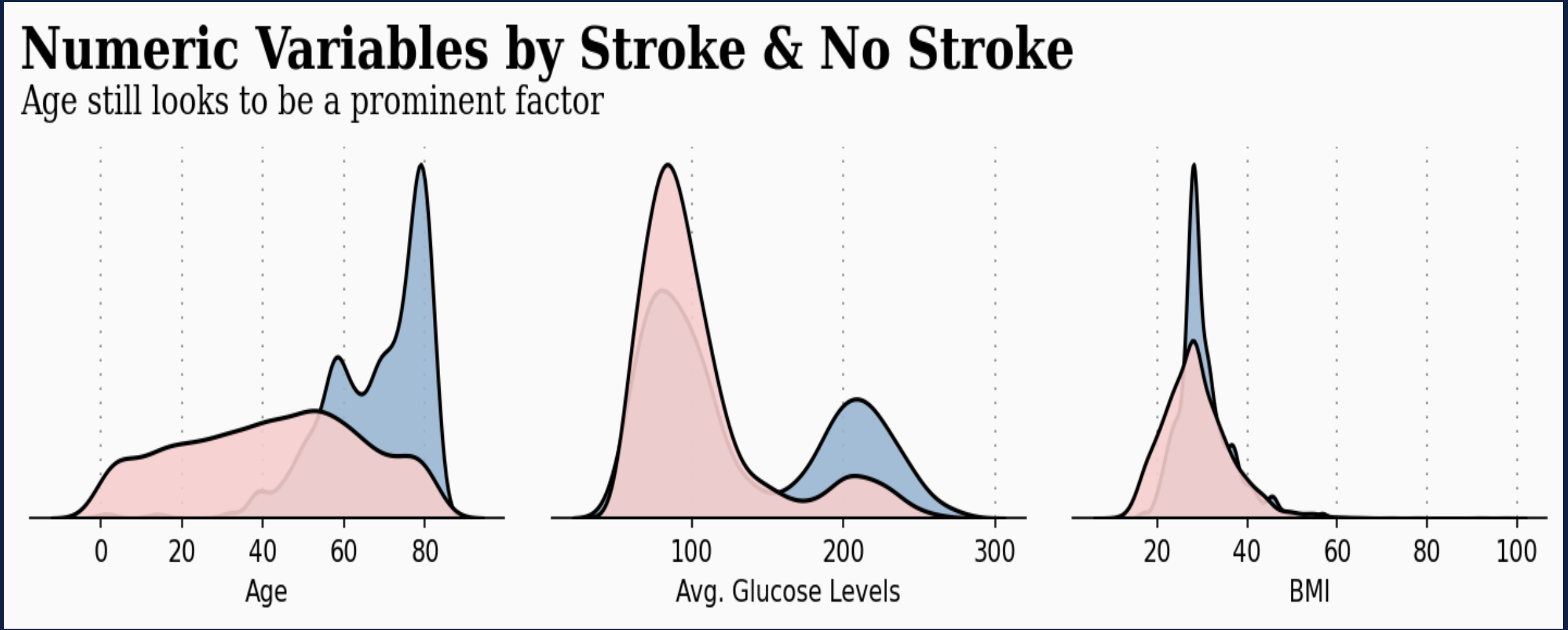
UNC GREENSBORO

Explore Numerical Features(upsampled)

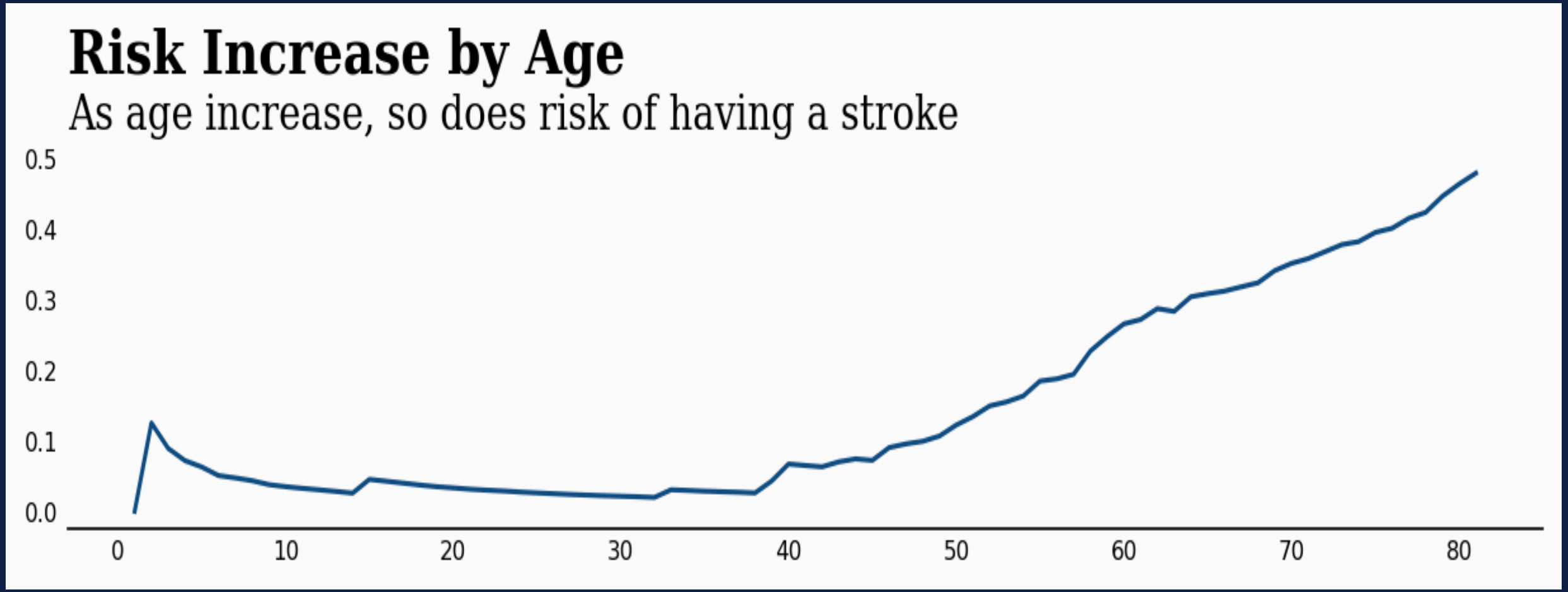


UNC GREENSBORO

Explore Numerical Features(Upsampled)

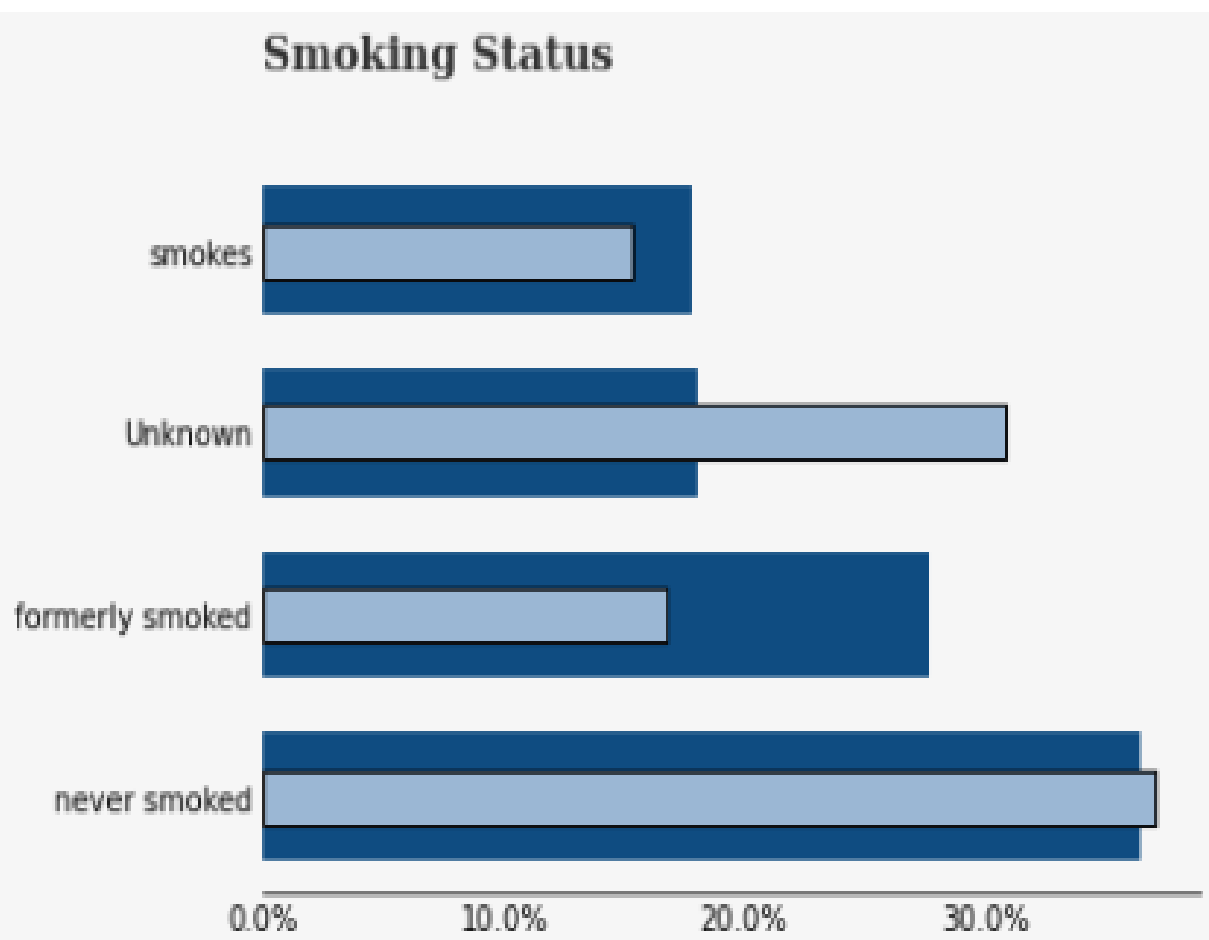


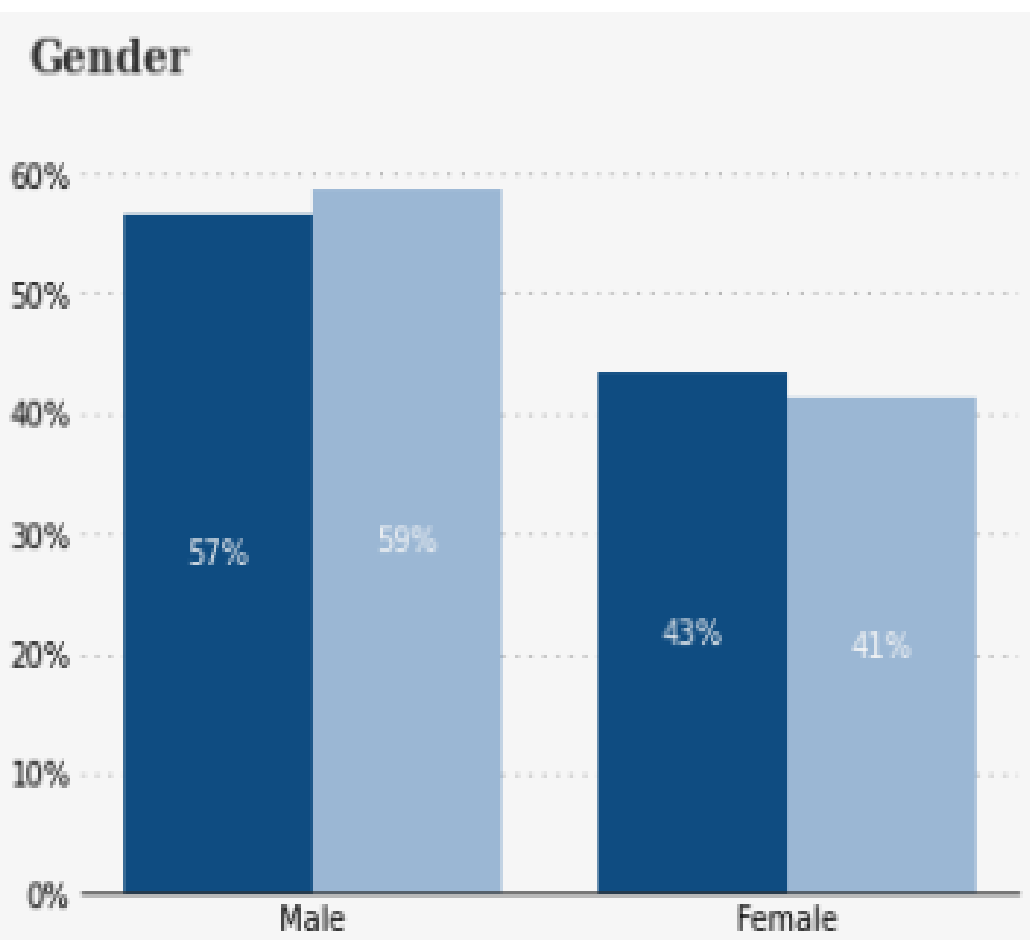
Age Versus Stroke(upsampled)



UNC GREENSBORO

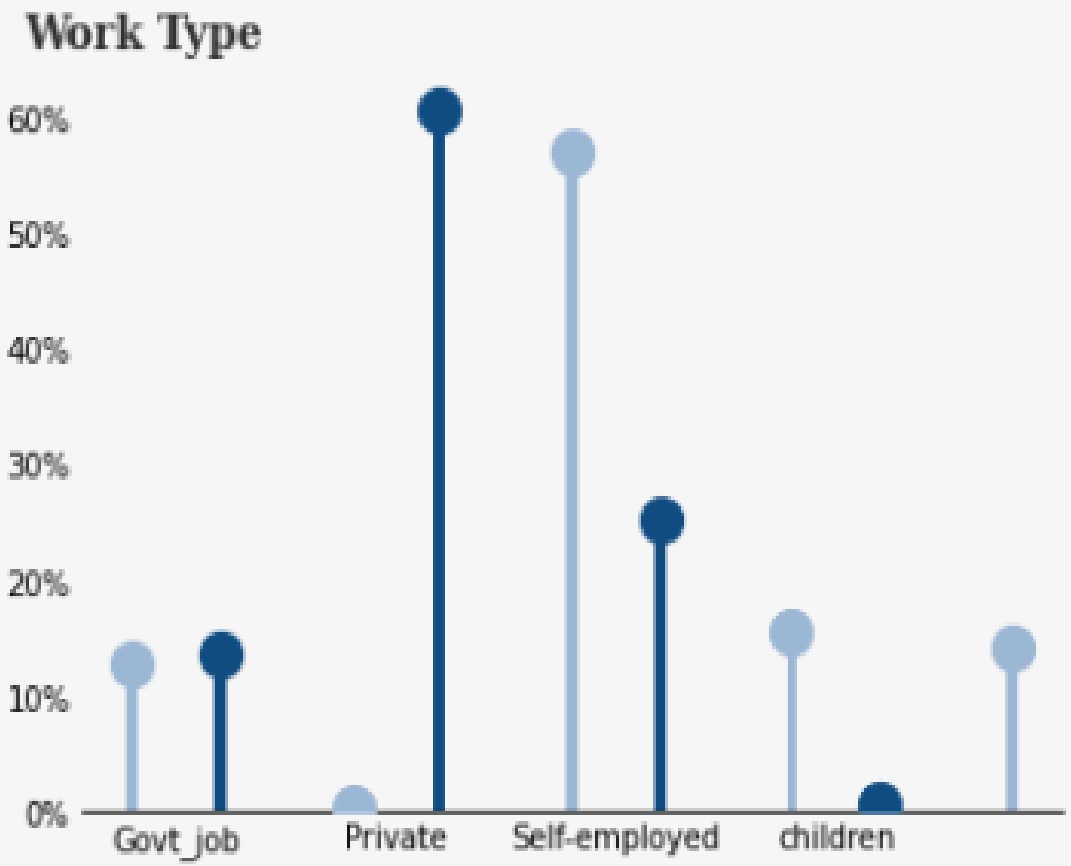
Explore Categorical Features (Upsampled)

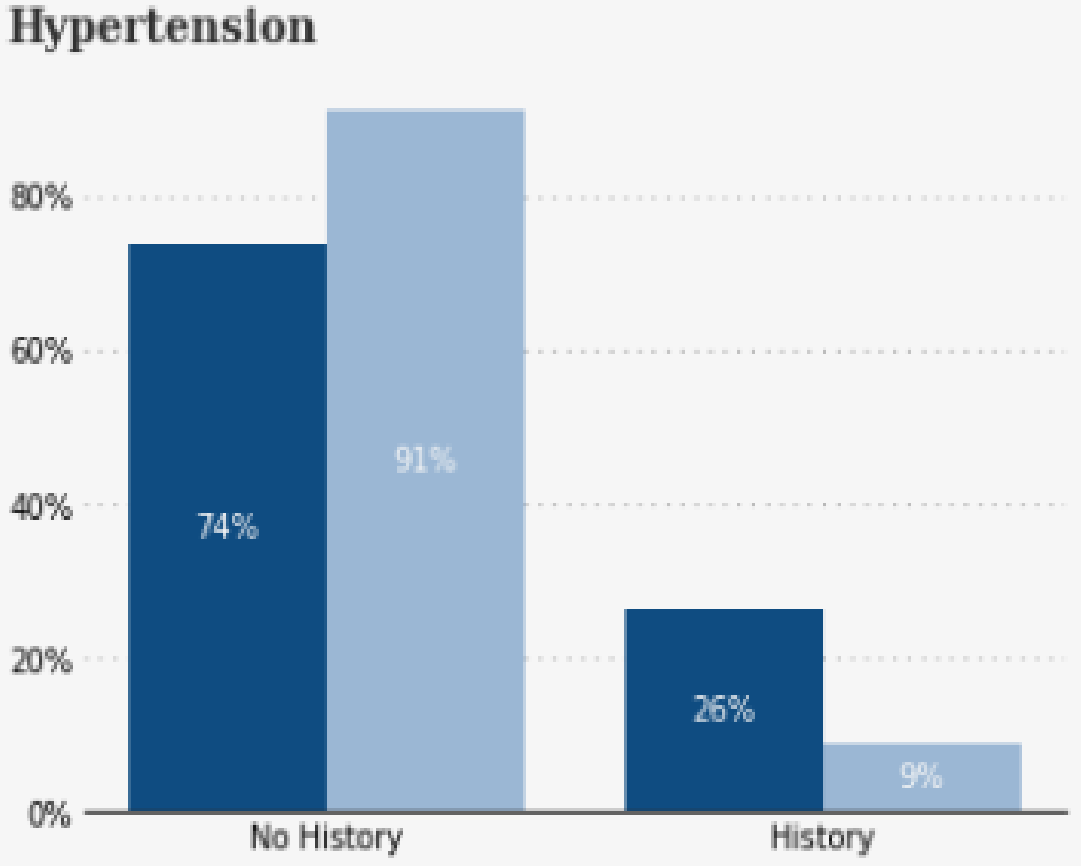




UNC GREENSBORO

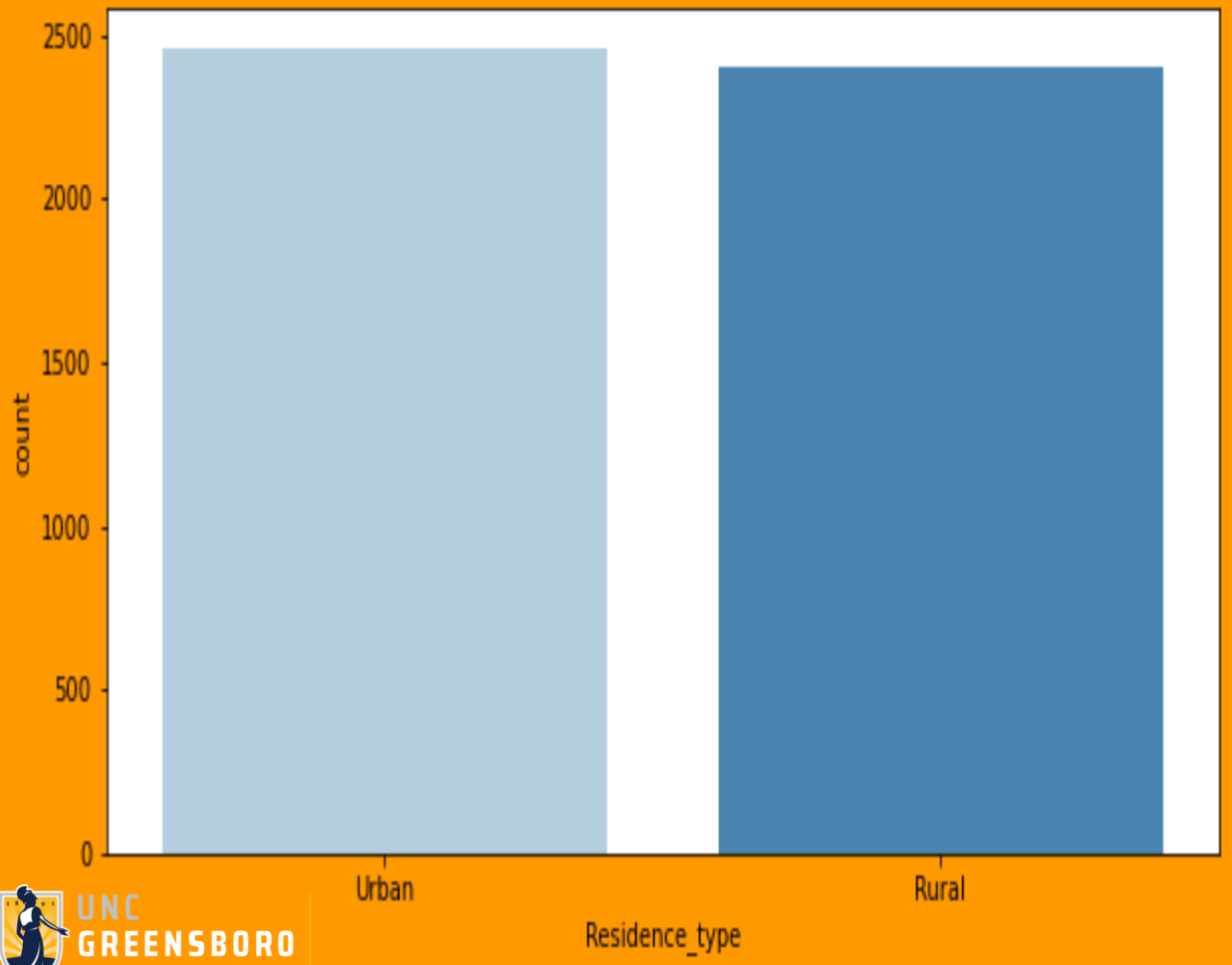
Work Type & Hypertension (upsampled)

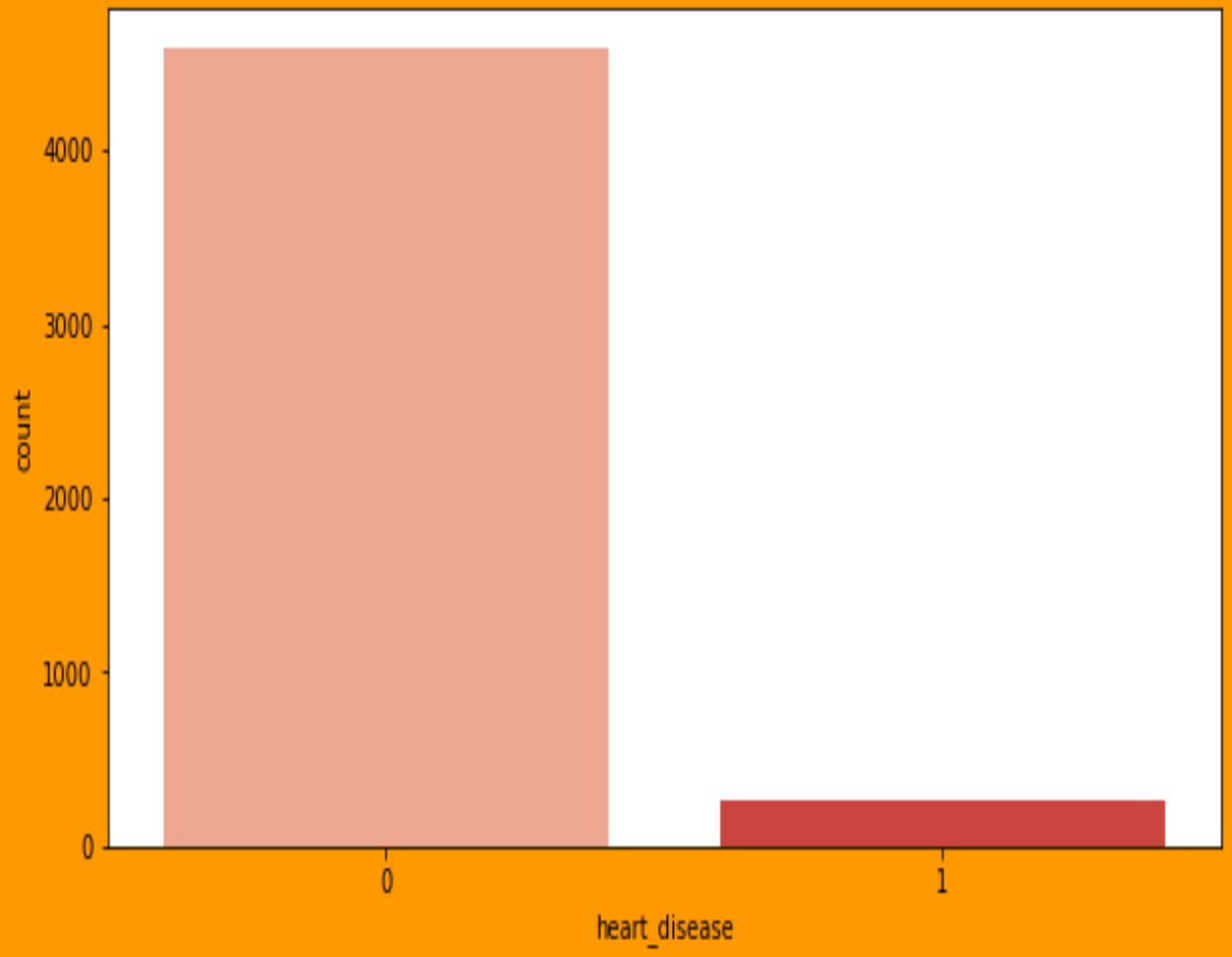




UNC GREENSBORO

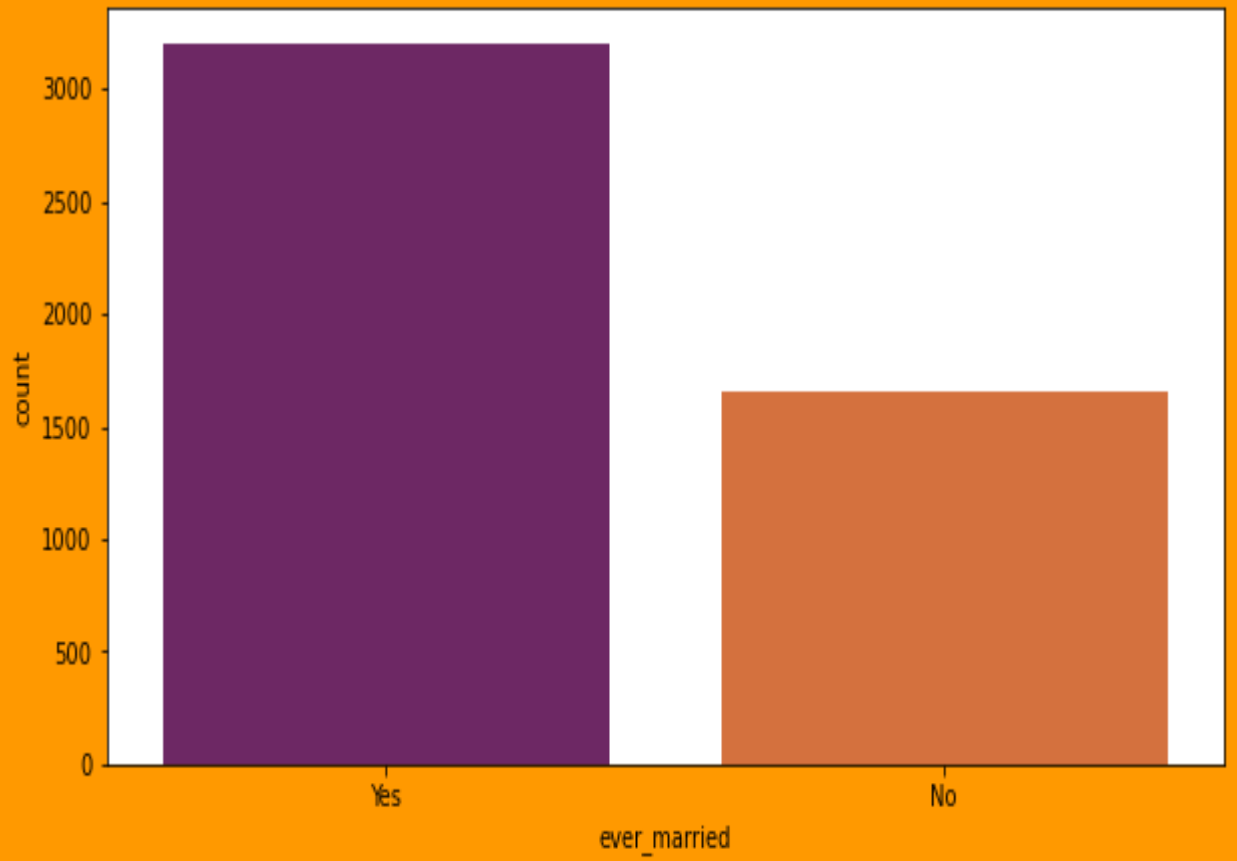
Residence Type & Heart Disease (Upsampled)





UNC GREENSBORO

Stroke vs Marital Status(upsampled)



UNC GREENSBORO

Conclusions!

- Age is a big factor in stroke patients ---the older you get the more you are at risk of stroke.
- Being a smoker or a former smoker does not relatively increase the risk of having a stroke.
- Location(Rural vs. Urban) does not appear to be significant factor.
- Most of the people who have had a stroke do not have any heart disease, but that does not prevent it being an influential factor.
- More than 25% of stroke cases, they had hypertension.
- Male are more likely to get stroke compared to female!!

Thank You!!

UNC GREENSBORO