INT 353

EDA PROJECT

Final Report

Name: Ananya Datana

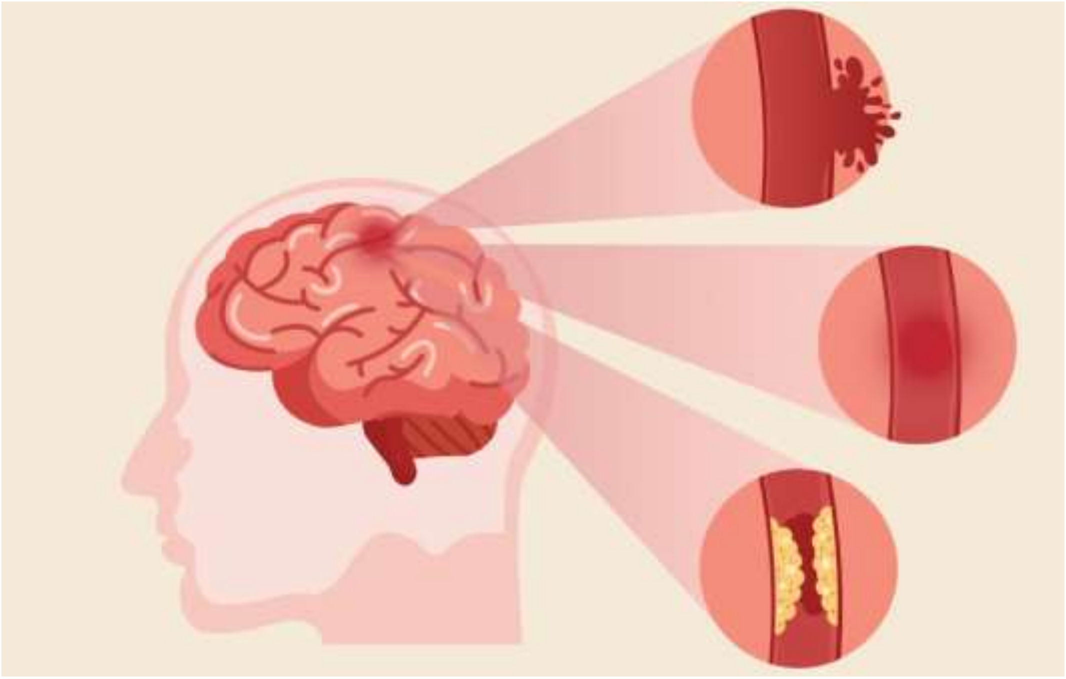
Reg. No.: 12019049

Roll No.: RK20UPB25

Course: INT 353 – EDA Project

Brain Stroke Prediction Dataset

Pretext



A stroke is a medical condition in which poor blood flow to the brain causes cell death. There are two main types of strokes: ischemic, due to lack of blood flow, and haemorrhagic, due to bleeding. Both cause parts of the brain to stop functioning properly. Signs and symptoms of a stroke may include an inability to move or feel on one side of the body, problems understanding or speaking, dizziness, or loss of vision to one side. Signs and symptoms often appear soon after the stroke has occurred. If symptoms last less than one or two hours, the stroke is a transient ischemic attack (TIA), also called a mini stroke. A haemorrhagic stroke may also be associated with a severe headache. The symptoms of a stroke can be permanent. Long-term complications may include pneumonia and loss of bladder control.

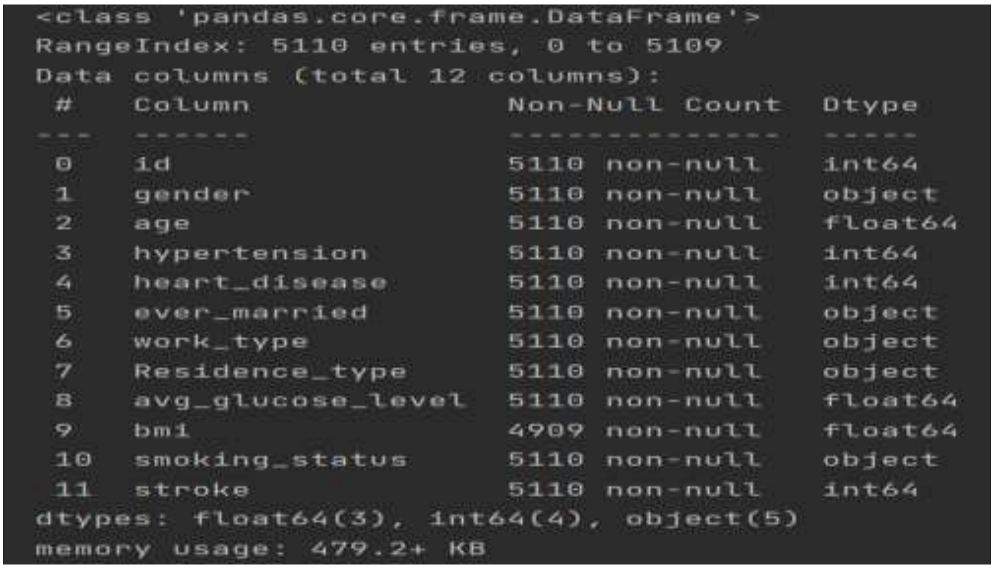
About the Dataset

The dataset contains 5110 entries as rows and 12 columns. The description about the columns is given below.

Attribute Information

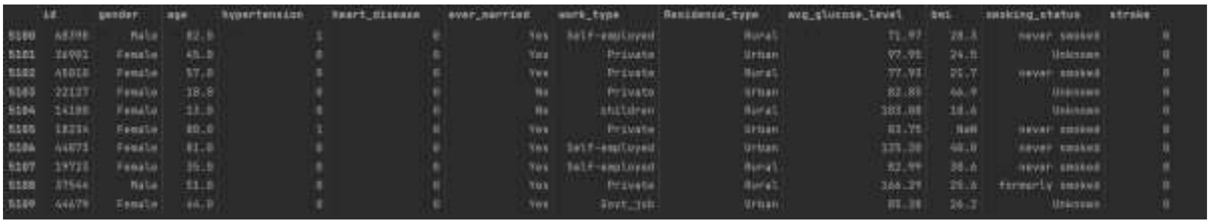
1. id: unique identifier
2. gender: “Male”, “Female” or “Other”
3. age: age of the patient
4. hypertension: 0 if the patient doesn’t have hypertension, 1 if the patient has hypertension
5. heart_disease: 0 if the patient doesn’t have any heart diseases, 1 if the patient has a heart disease
6. ever_married: “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

    *Note: “Unknown” in smoking_status means that the information is unavailable for this patient



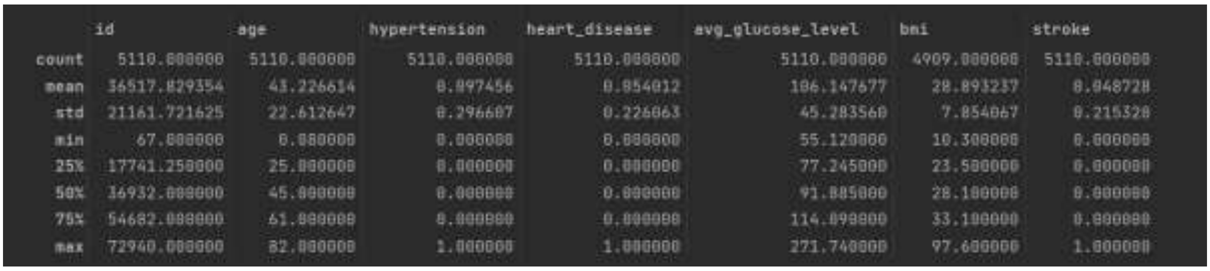
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5110 entries, 0 to 5109
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   id                 5110 non-null   int64  
 1   gender             5110 non-null   object 
 2   age                5110 non-null   float64
 3   hypertension       5110 non-null   int64  
 4   heart_disease      5110 non-null   int64  
 5   ever_married       5110 non-null   object 
 6   work_type          5110 non-null   object 
 7   Residence_type     5110 non-null   object 
 8   avg_glucose_level  5110 non-null   float64
 9   bmi                4909 non-null   float64
 10  smoking_status     5110 non-null   object 
 11  stroke             5110 non-null   int64  
dtypes: float64(3), int64(4), object(5)
memory usage: 479.2+ KB
```

A Glimpse of Dataset



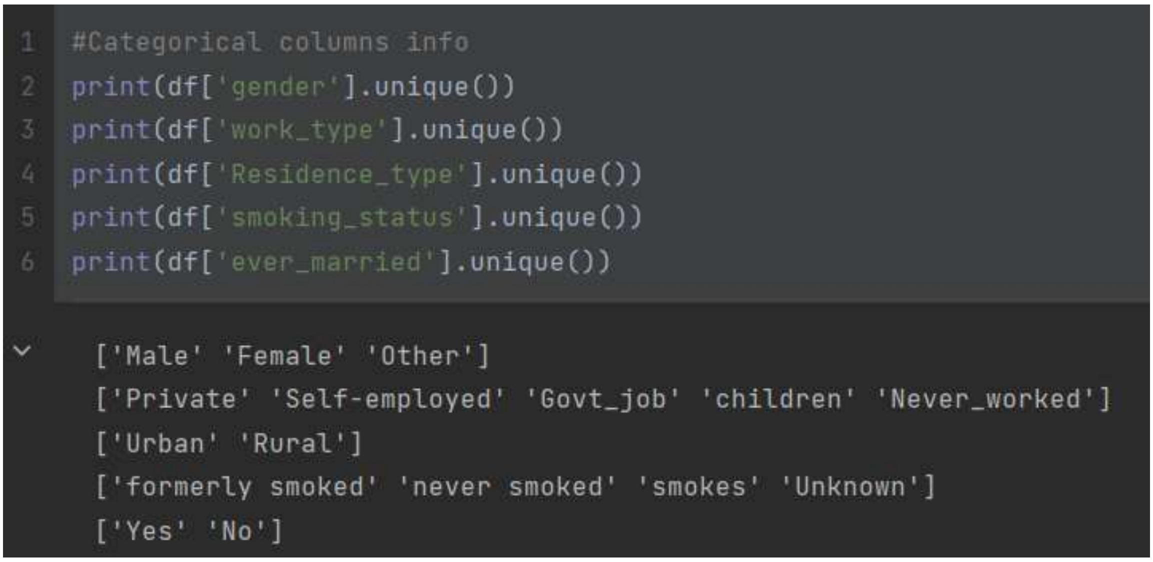
	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
5100	68398	Male	82.0	1	0	Yes	Self-employed	Rural	71.97	28.3	never smoked	0
5101	36901	Female	45.0	0	0	Yes	Private	Urban	97.95	24.5	Unknown	0
5102	45010	Female	57.0	0	0	Yes	Private	Rural	77.93	21.7	never smoked	0
5103	22127	Female	18.0	0	0	No	Private	Urban	82.85	46.9	Unknown	0
5104	14180	Female	13.0	0	0	No	children	Rural	103.08	18.6	Unknown	0
5105	18234	Female	80.0	1	0	Yes	Private	Urban	83.75	NaN	never smoked	0
5106	44873	Female	81.0	0	0	Yes	Self-employed	Urban	125.20	40.0	never smoked	0
5107	19723	Female	35.0	0	0	Yes	Self-employed	Rural	82.99	30.6	never smoked	0
5108	37544	Male	51.0	0	0	Yes	Private	Rural	166.29	25.6	formerly smoked	0
5109	44679	Female	44.0	0	0	Yes	Govt_job	Urban	85.28	26.2	Unknown	0

Summary of Dataset



	id	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke
count	5110.000000	5110.000000	5110.000000	5110.000000	5110.000000	4909.000000	5110.000000
mean	36517.829354	43.226614	0.097456	0.054012	106.147677	28.893237	0.048728
std	21161.721625	22.612647	0.296607	0.226063	45.283560	7.854067	0.215328
min	67.000000	0.080000	0.000000	0.000000	55.120000	10.300000	0.000000
25%	17741.250000	25.000000	0.000000	0.000000	77.245000	23.500000	0.000000
50%	36932.000000	45.000000	0.000000	0.000000	91.885000	28.100000	0.000000
75%	54682.000000	61.000000	0.000000	0.000000	114.090000	33.100000	0.000000
max	72940.000000	82.000000	1.000000	1.000000	271.740000	97.600000	1.000000

Categorical columns info of Dataset

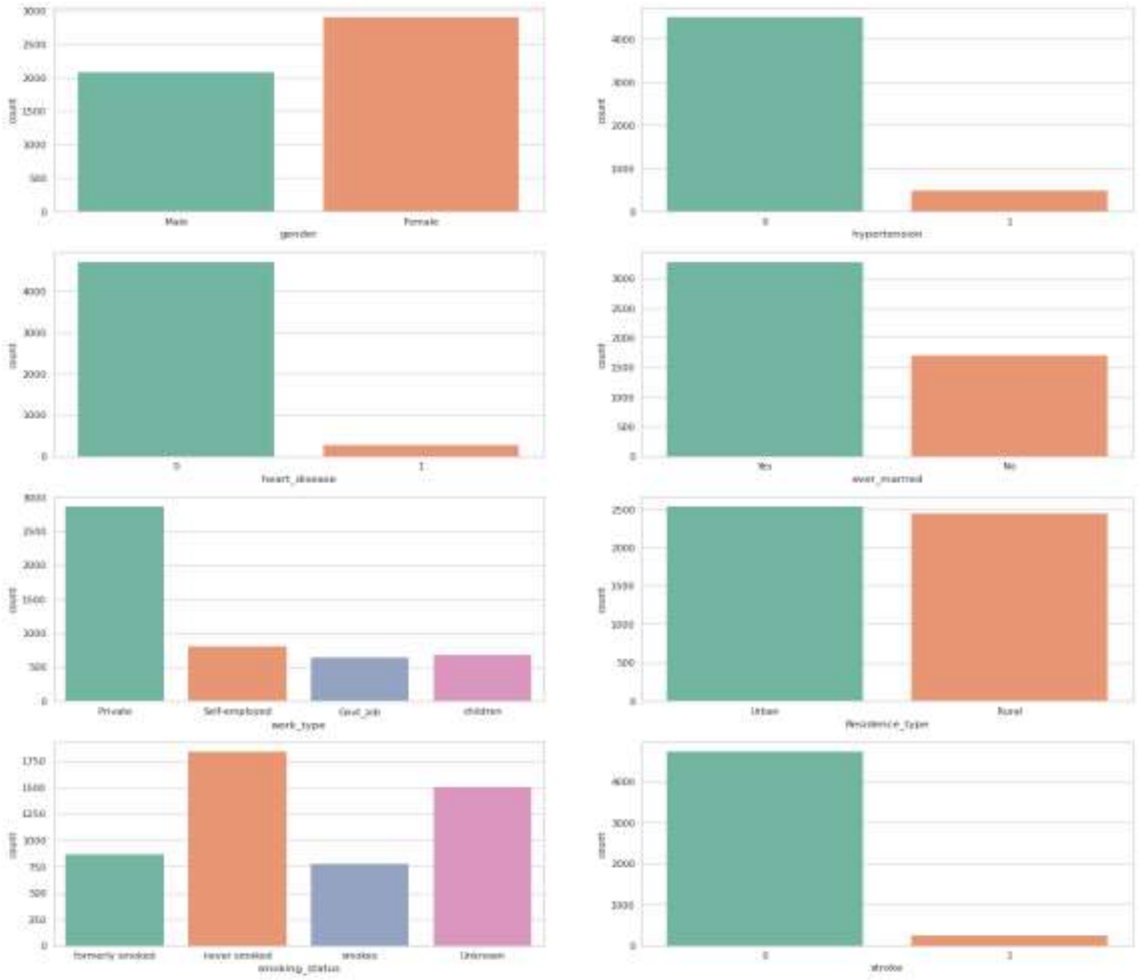


```
#Categorical columns info
print(df['gender'].unique())
print(df['work_type'].unique())
print(df['Residence_type'].unique())
print(df['smoking_status'].unique())
print(df['ever_married'].unique())
```

```
['Male' 'Female' 'Other']
['Private' 'Self-employed' 'Govt_job' 'children' 'Never_worked']
['Urban' 'Rural']
['formerly smoked' 'never smoked' 'smokes' 'Unknown']
['Yes' 'No']
```

EDA on Dataset

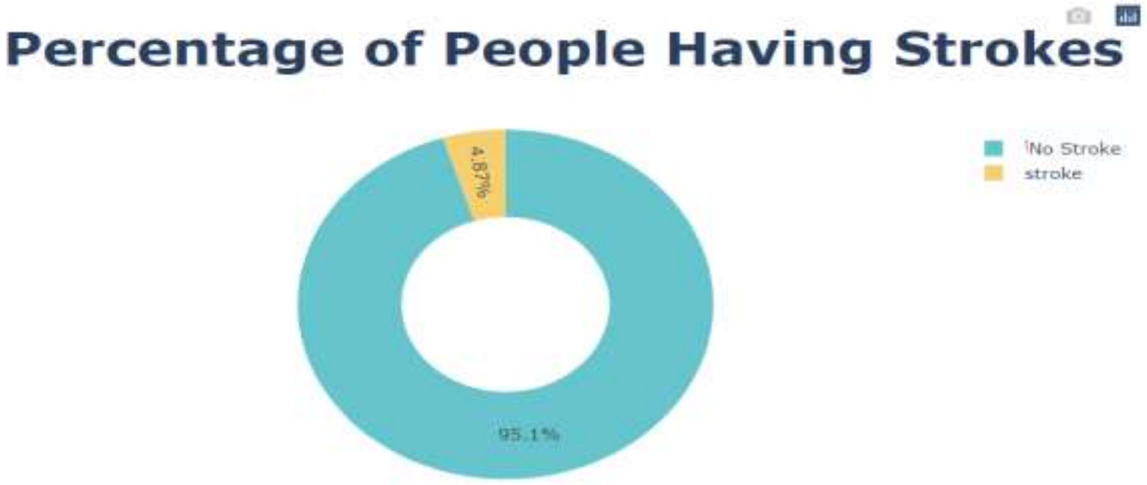
In response to requirement of knowing and understanding the data in a better way and get some insights on the data.



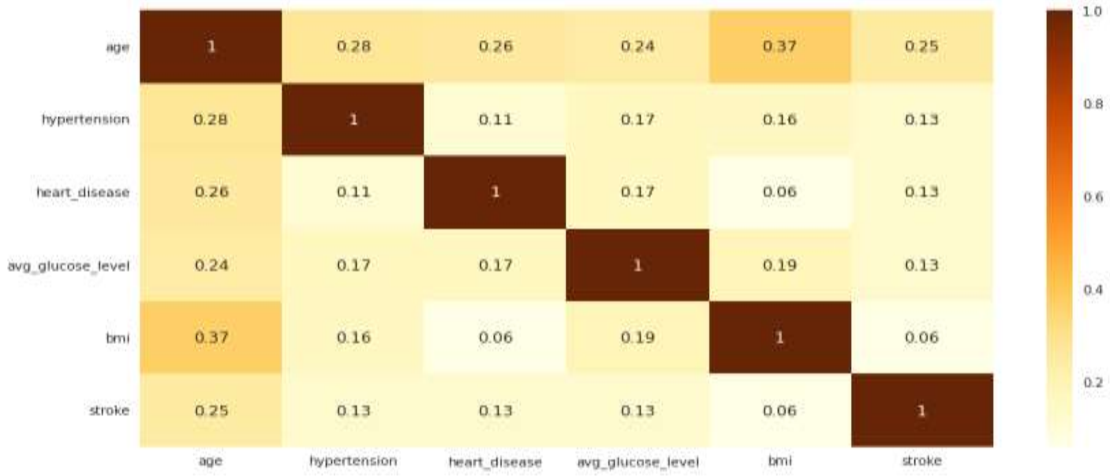
We can come to the conclusion that we have more women than men in our base, there are few people who have hypertension and heart disease, most people are married and have a private job, when we compare the type of residence the base is well balanced, a An important point is that we have a lot of data without the smoker or non-smoker information, when we look at our stroke variable we can see that it is very unbalanced.

Percentage of patients of Focus

The dataset contains data of the many people who had visited a doctor due to stroke suspicion and thus, not all of them are suffering from this disease.



Correlation between the Variables



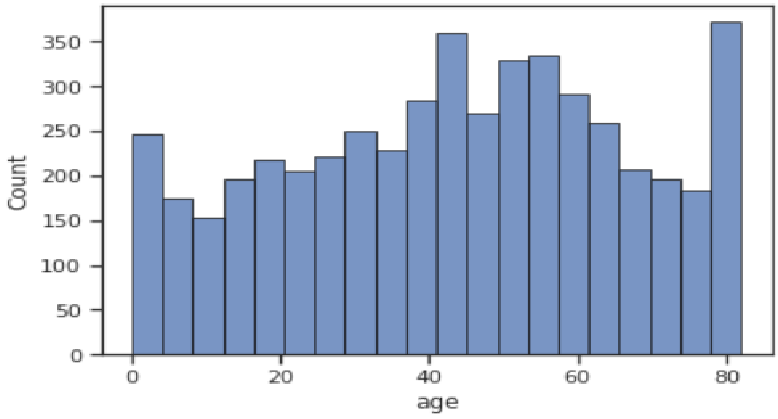
	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke
age	1	0.28	0.26	0.24	0.37	0.25
hypertension	0.28	1	0.11	0.17	0.16	0.13
heart_disease	0.26	0.11	1	0.17	0.06	0.13
avg_glucose_level	0.24	0.17	0.17	1	0.19	0.13
bmi	0.37	0.16	0.06	0.19	1	0.06
stroke	0.25	0.13	0.13	0.13	0.06	1

Checking the correlation between our variables, here we can see that we don't have a strong correlation between the variables

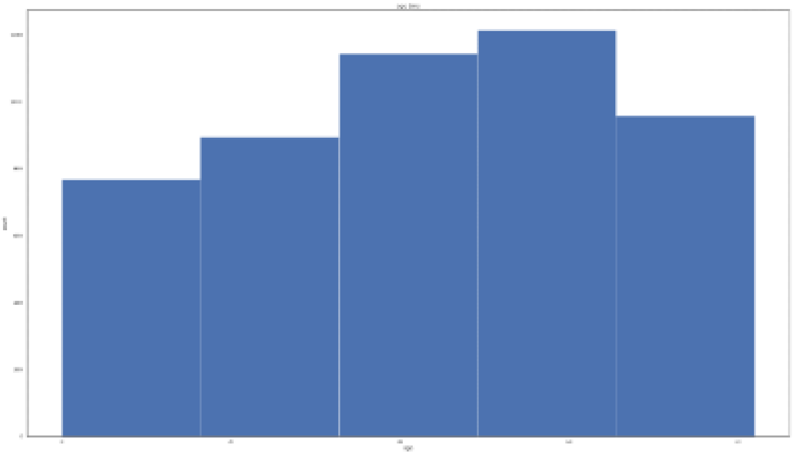
Is there any relation b/w Age and Stroke?

Age is a continuous variable so it would not be so helpful if taken as it is. Thus, the wise choice here is binning the data into categories namely: ‘Young’, ‘Young Adult’, ‘Prime’, ‘Early old’ and ‘Old’.

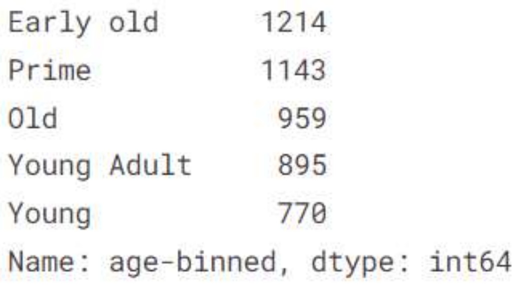
Initial intuition says that strokes will increase with age and peak would be in 40-60 age group.



Before Binning

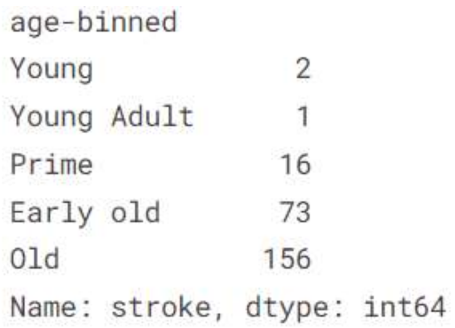


After binning



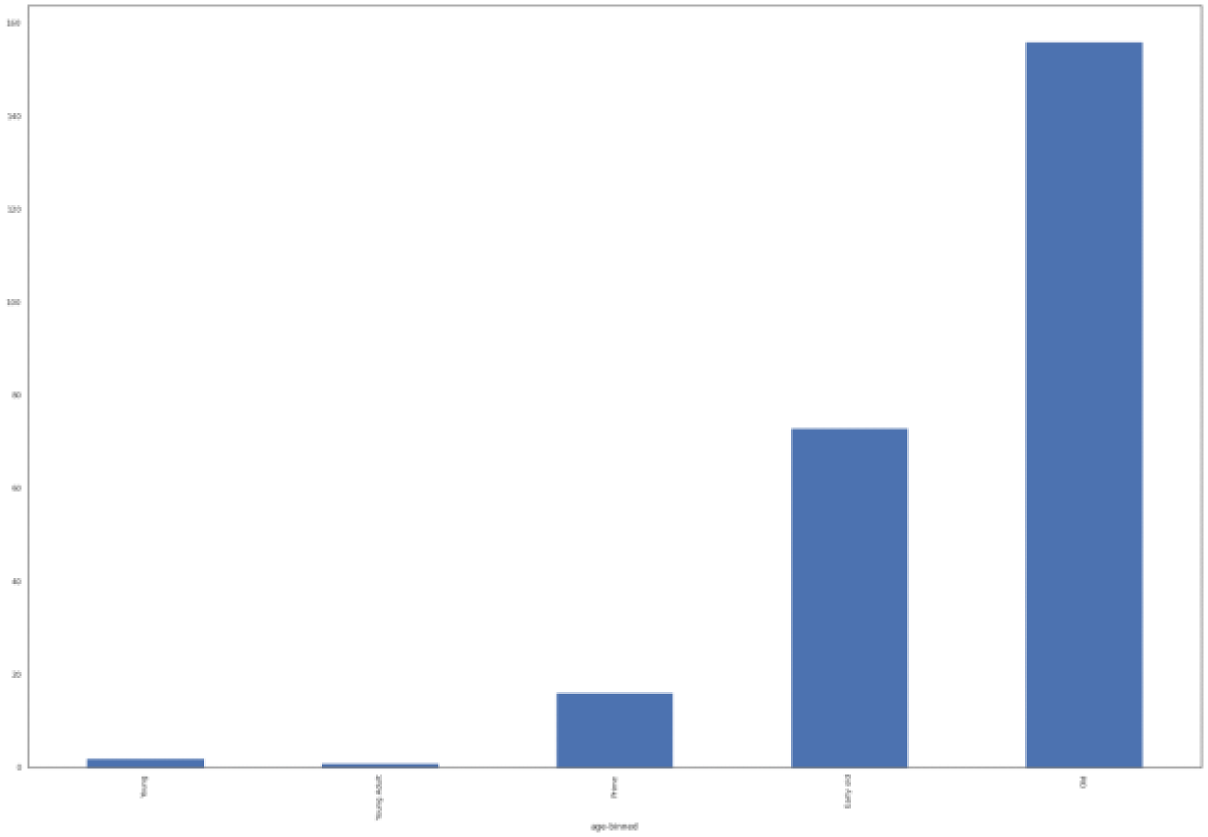
```
Early old      1214
Prime          1143
Old             959
Young Adult     895
Young           770
Name: age-binned, dtype: int64
```

Number of people in each age group



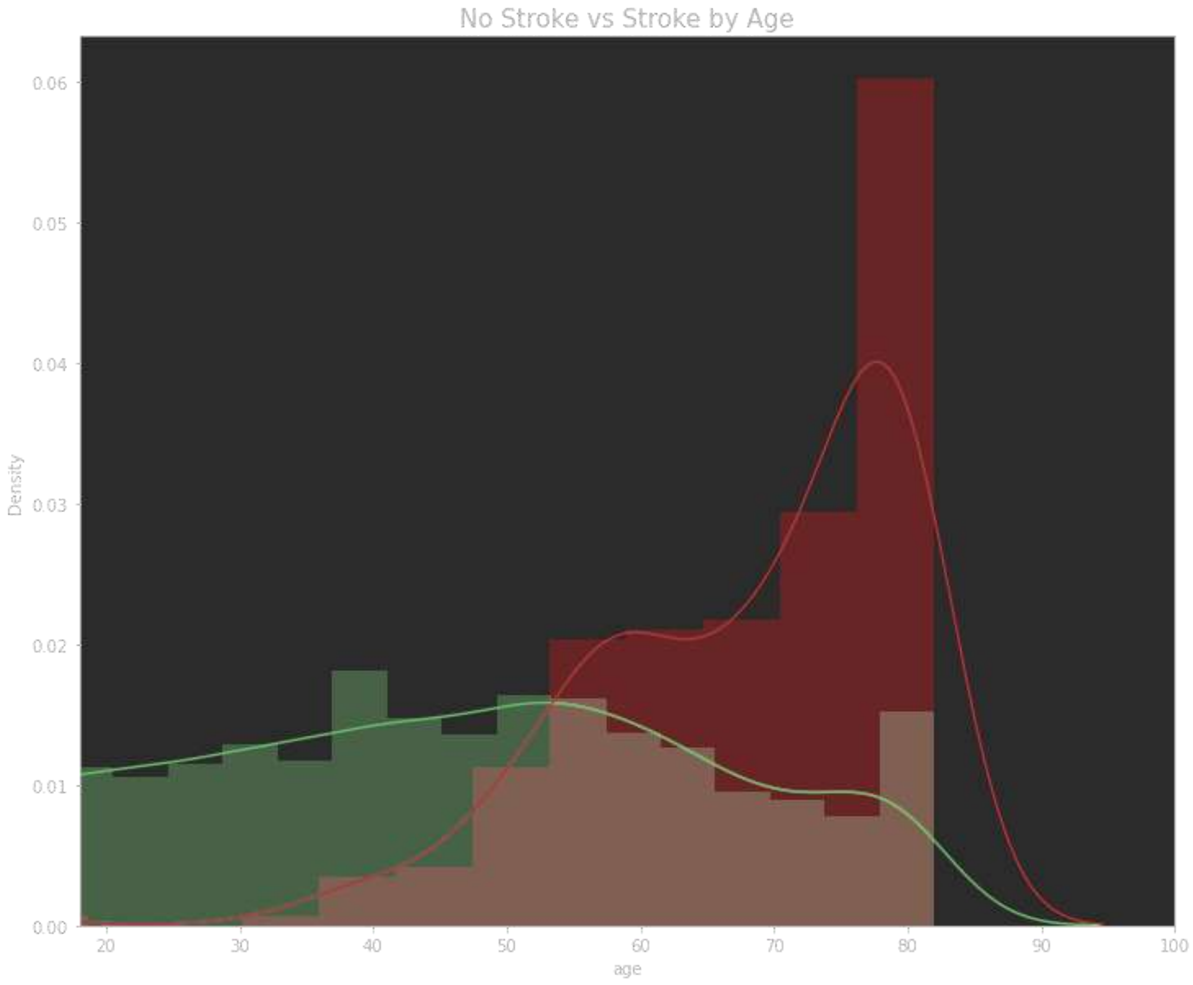
```
age-binned
Young            2
Young Adult      1
Prime           16
Early old       73
Old            156
Name: stroke, dtype: int64
```

Distribution of stroke positive patients



Distribution of stroke positive patients under binned ages

An interesting trend. People who are in old category have higher stroke counts. However, it could also be the case that because 'old' bin has higher instances, it could reflect in higher number of strokes. There is also a rising trend of more instances of stroke. Now even considering the difference between data instances between age groups, it is clear that there is a rise in trend i.e. as people get older there is a higher chance of getting a stroke.



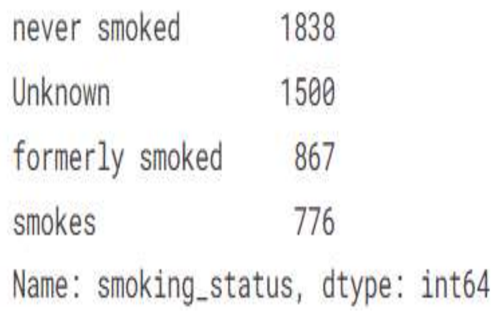
Age vs Stroke density

It’s obvious that elder people are more prone to brain strokes.

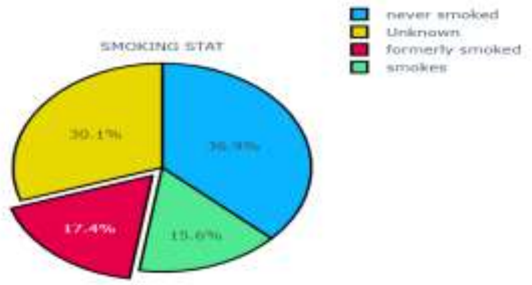
The age parameter is a little left skewed with a peak around late 70s and 80s.

When we look at our age variable, we can see that only from 40 years old people start to have Stroke, and cases below that age are very rare, the trend is that the older the person, the more likely he is to have a stroke.

Is there any relation b/w Smoking vs Stroke?

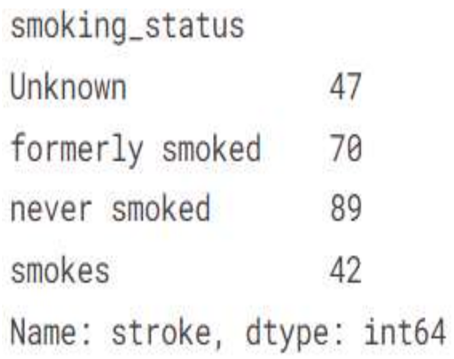


```
never smoked       1838
Unknown            1500
formerly smoked     867
smokes              776
Name: smoking_status, dtype: int64
```

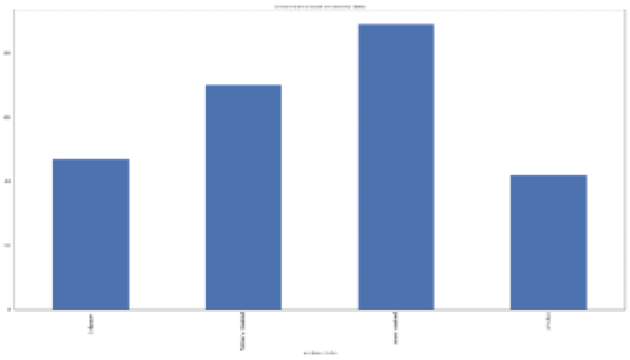


Distribution of smoking statuses in the data

Initial intuition says that active smokers will have a higher risk of stroke.



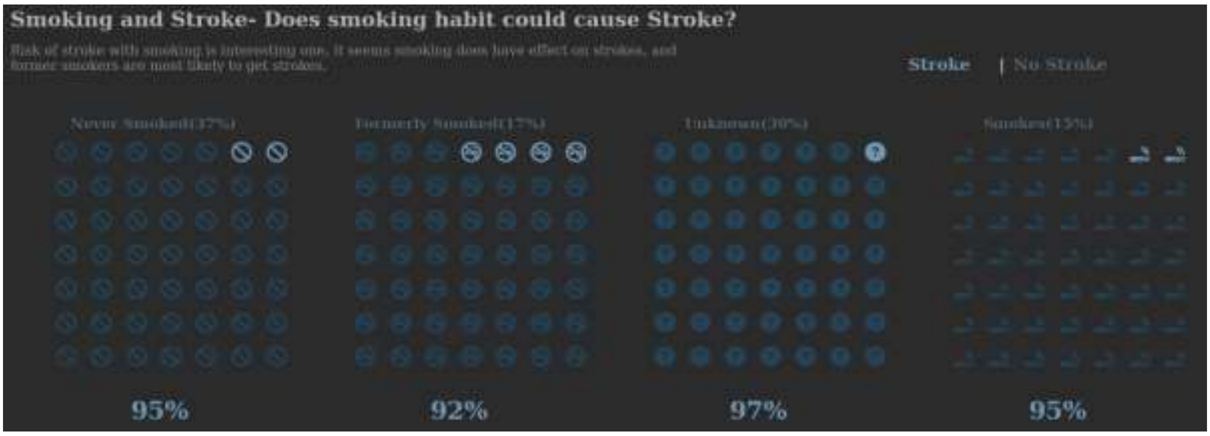
```
smoking_status
Unknown            47
formerly smoked    70
never smoked       89
smokes             42
Name: stroke, dtype: int64
```



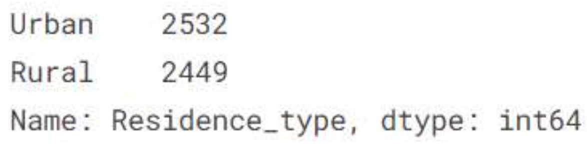
Distribution stroke positive patients vs smoking statuses in the data

People who have never smoked or used to smoke seems to have a higher count of strokes recorded compared to those who regularly smokes.

Maybe the reason is that they use smoking as a stress relieving activity.

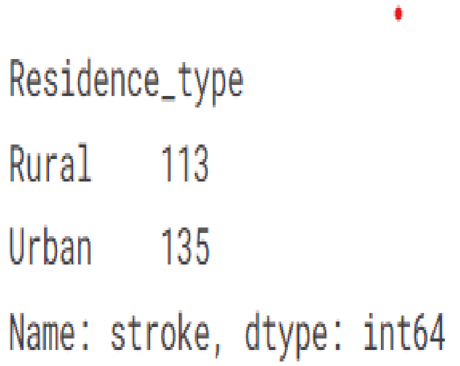


Is there any relation b/w Residence Type and Stroke?

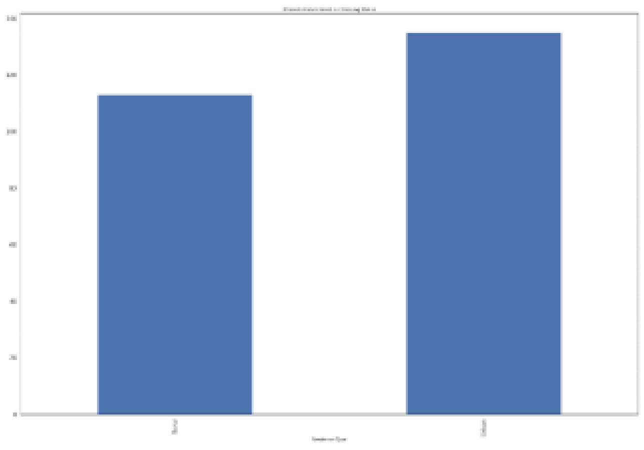


```
Urban    2532
Rural    2449
Name: Residence_type, dtype: int64
```

Distribution of residence type in the data

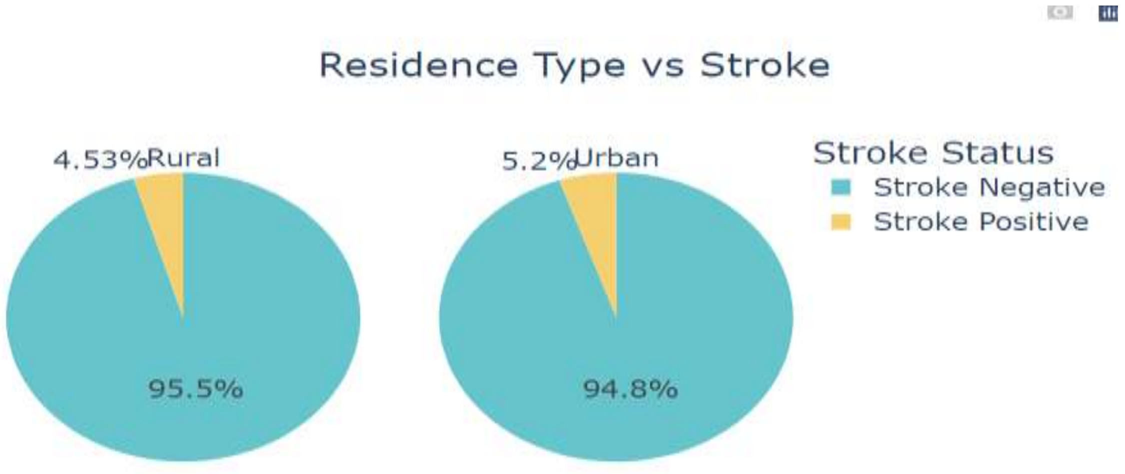


```
Residence_type
Rural    113
Urban    135
Name: stroke, dtype: int64
```

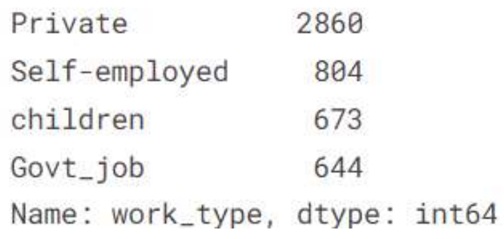


Distribution of residence type for stroke positive patients in the data

People who live in urban setting seem to have more instances of stroke. However, this lead is not by a lot. Also considering that there is a similar trend where there are more instances of urban residence type than rural, it might explain the slight lead shown in the graph above.



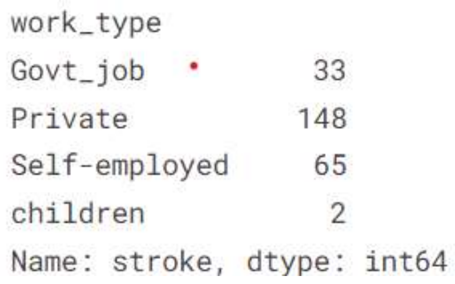
Is there any relation b/w Workplace and Stroke?



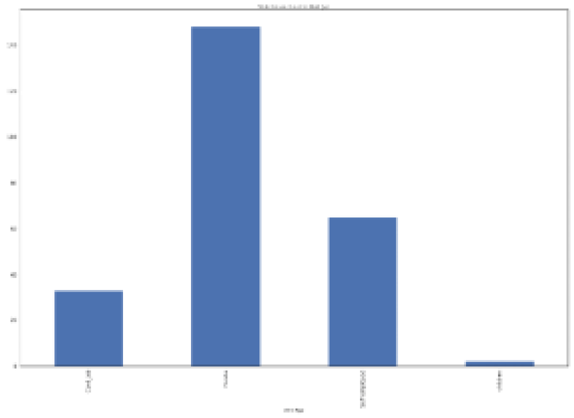
```
Private          2860
Self-employed     804
children          673
Govt_job          644
Name: work_type, dtype: int64
```

Distribution of workplace type in the data

Early intuition suggests that Private sector workers would be more prone to strokes because of high quantity of work.

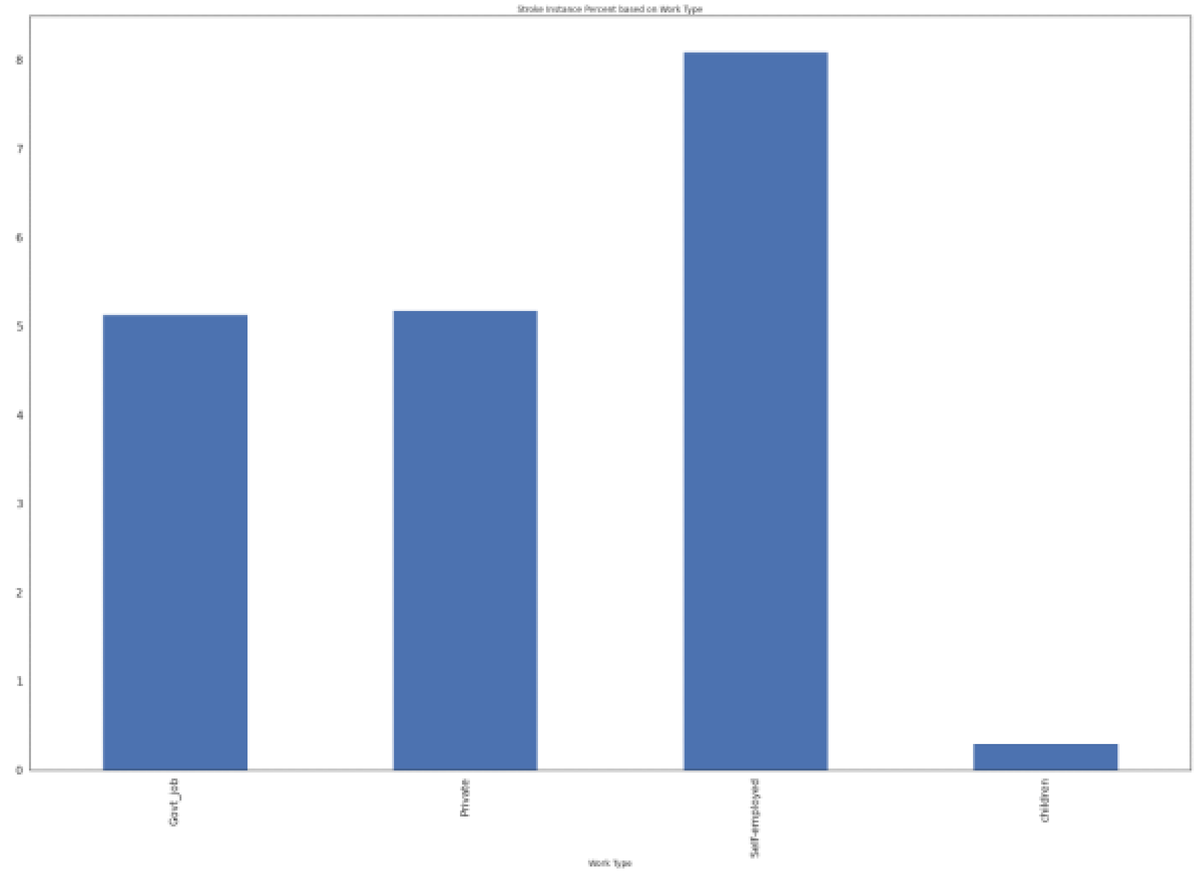


```
work_type
Govt_job            33
Private            148
Self-employed       65
children             2
Name: stroke, dtype: int64
```



Distribution of workplace type in the data for stroke positive patients

Based on work type, there is a high discrepancy between the frequency of each type of data. So taking the percent is the obvious method to move forward.



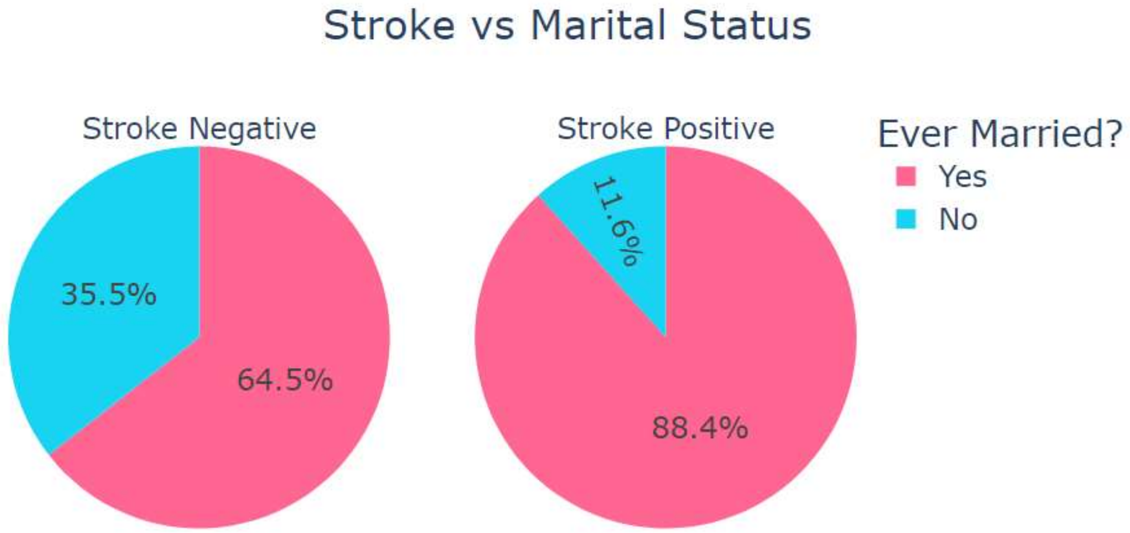
Distribution of workplace type in the data for stroke positive patients

Here we can see that people who are self-employed tend to have higher chances of getting a stroke. While private and government job employees have almost similar chances with the lowest belonging to children.

A picture is forming, every factor that is stress inducing have a higher chance of triggering a stroke. Like smoking that allegedly reduces stress, people who smoke have lesser chance to get a stroke. Similarly, self-employed individuals are more likely to have more stress than compared to people having government or private jobs. Another factor that is emerging is age - where, higher the age, higher are the chances of getting a stroke.

Is there any relation b/w Martial Status and Stroke?

Initial intuition suggests that married people will be more prone to strokes.



Risk of stroke on married people is high.

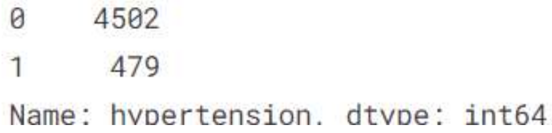
Perhaps married people have more responsibilities and are subjected to more pressure.

Married people must take care of not only themselves but also of the people dependent on them, spouses and children can be considered as valid dependents of married people.

Our initial intuition is correct.

Is there any relation b/w Hypertension and Stroke?

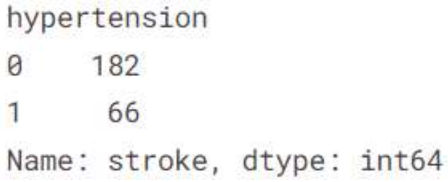
Initial intuition: Patients with hypertension are more likely to get stroke.



```
0    4502
1     479
Name: hypertension, dtype: int64
```

Distribution of hypertension in the data.

0: No *1: Yes*

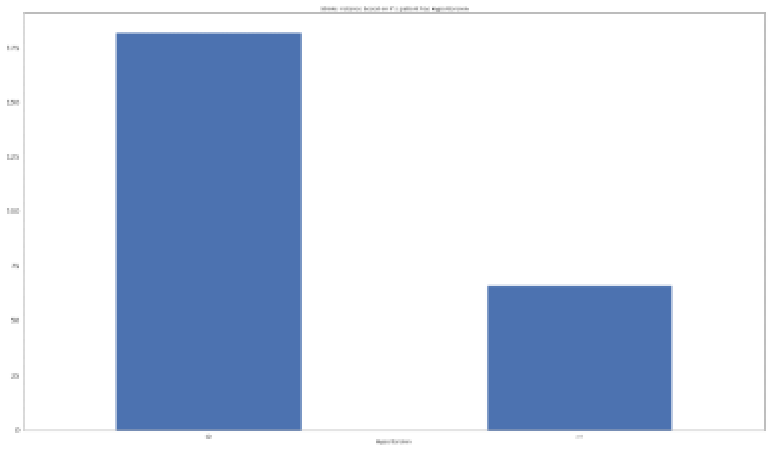


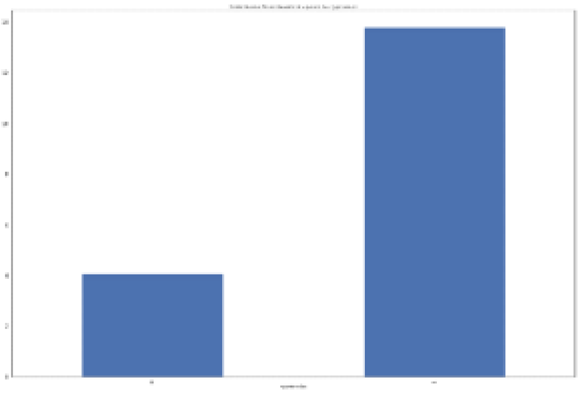
```
hypertension
0    182
1     66
Name: stroke, dtype: int64
```

Distribution of hypertension in the data in stroke positive patients.

0: No Stroke *1: Yes Stroke*

There is a huge disparity between the frequency of hypertension. So judging by taking percent seems a better option.



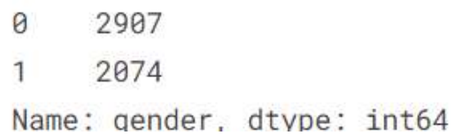


Now this makes more sense. People who have hypertension are more likely to have a stroke. 13.7% of people who have hypertension, based on the database, have suffered stroke

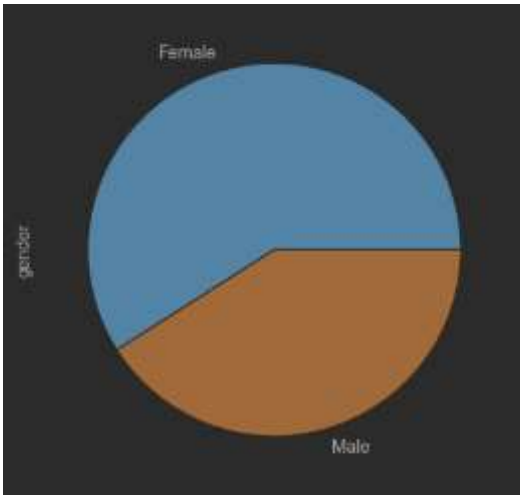
Initial intuition was correct in this case.

Is there any relation b/w Gender and Stroke?

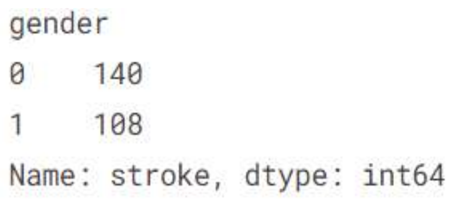
Initial intuition says that men should be more at risk of getting a stroke than women because they are seen to lead a very stressful life, more than women in general.



```
0    2907
1    2074
Name: gender, dtype: int64
```



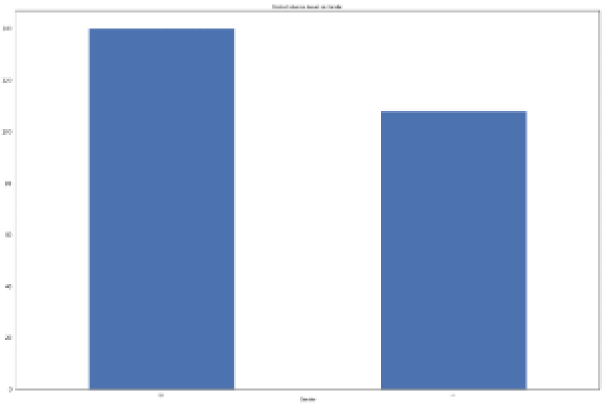
Distribution of patients in accordance to gender

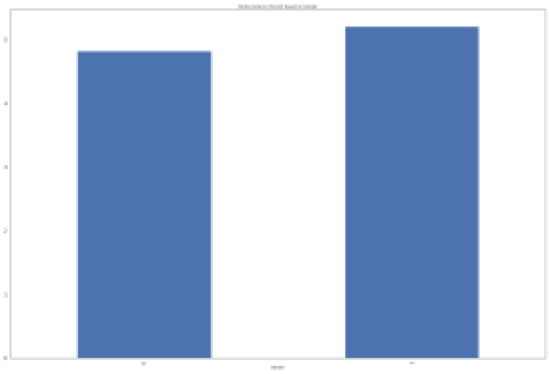


```
gender
0    140
1    108
Name: stroke, dtype: int64
```

Distribution of stroke positive patients in accordance to gender

There doesn't seem to be much of a difference. Let's take a look at percentages!



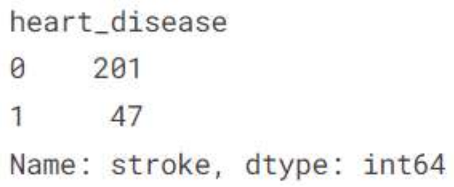


there doesn't seem to be any pattern that we can use to consider if gender is a factor for deciding the occurrence of stroke.

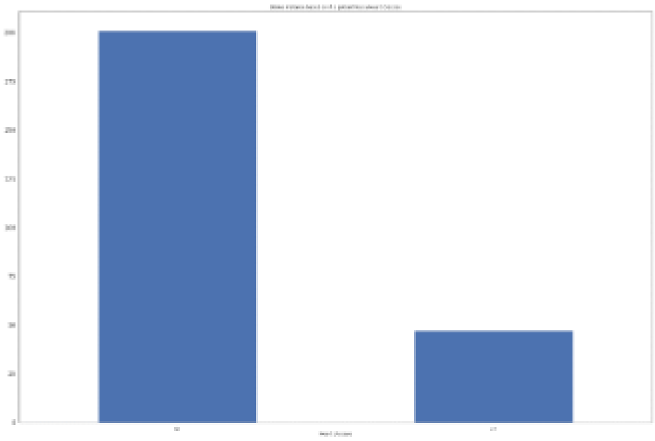
Initial intuition was correct.

Is there any relation b/w Heart Disease and Stroke?

Initial intuition says that those with heart disease are more at risk of stroke.

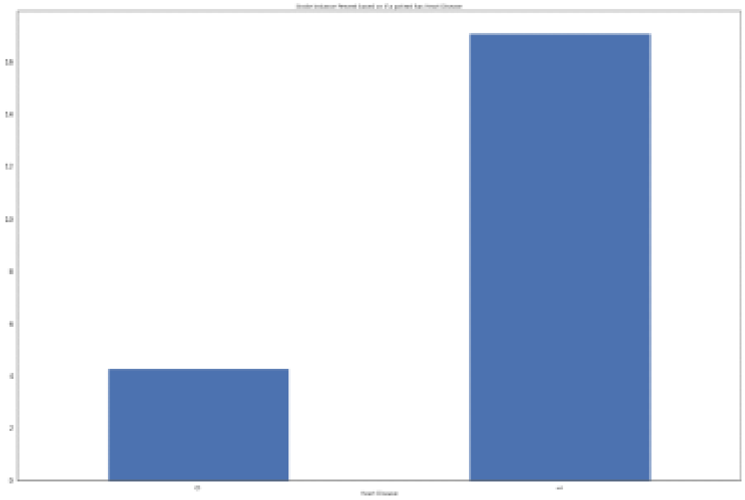


```
heart_disease
0    201
1     47
Name: stroke, dtype: int64
```



Distribution of patients in by conception of heart disease

A huge disparity between the frequency of heart disease. Thus, taking the percentage of data and working on it.

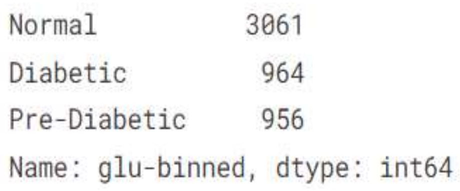


As expected, a whopping 17% of people who had heart disease suffered from a stroke, compared to only 4% of people who didn't have a heart disease. Heart disease is a factor in determining the occurrence of stroke.

Is there any relation b/w Glucose Levels and Stroke?

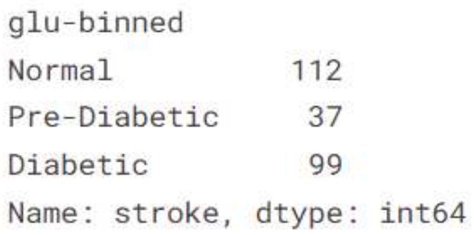
Glucose Levels are a continuous variable so it would not be so helpful if taken as it is. Thus, the wise choice here is binning the data into categories namely: ‘Normal’, ‘Pre-Diabetic and ‘Diabetic’.

Initial intuition suggests that diabetic people are more at risk of stroke.



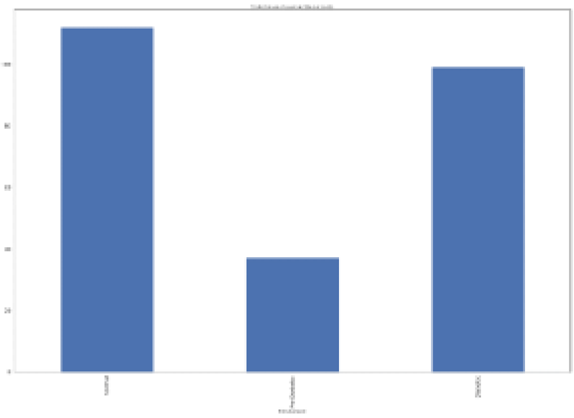
```
Normal          3061
Diabetic         964
Pre-Diabetic     956
Name: glu-binned, dtype: int64
```

Distribution of patients wrt to glucose levels(binned)



```
glu-binned
Normal          112
Pre-Diabetic     37
Diabetic         99
Name: stroke, dtype: int64
```

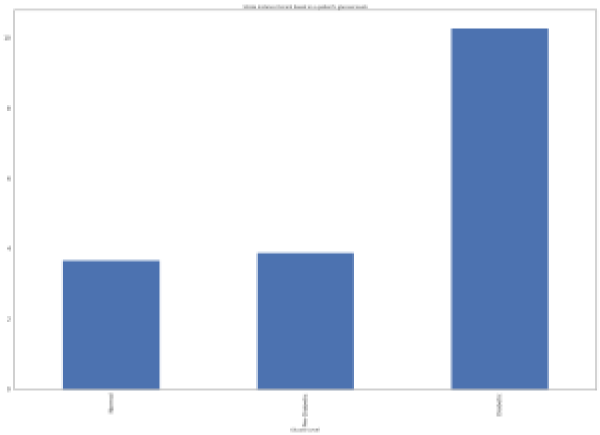
Distribution of patients wrt to glucose levels(binned) for stroke +ive



Again, a huge disparity. Percentage of average glucose level class to the rescue!

We can figure out that diabetic patients i.e. patients who have had high glucose levels had higher occurrence of stroke at 10.2%.

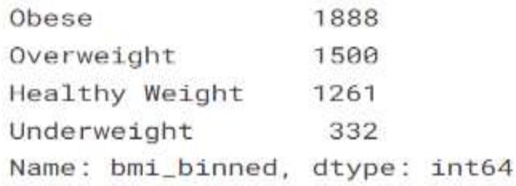
Our intuition was correct.



Is there any relation b/w BMI and Stroke?

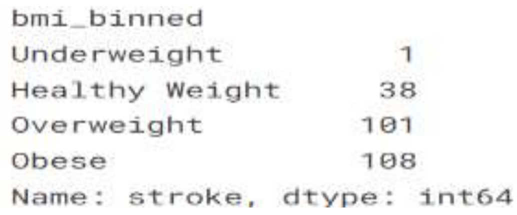
Again, BMI Stats are a continuous variable so it would not be so helpful if taken as it is. Thus, the wise choice here is binning the data into categories namely: ‘Obese’, ‘Overweight’, ‘Healthy Weight’ and ‘Underweight’.

Initial intuition suggests that heavy weighing people are more at risk of stroke.



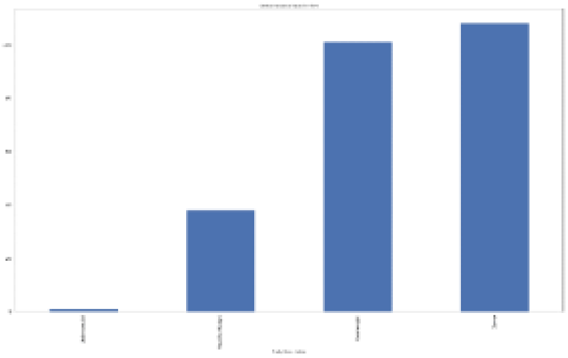
```
Obese             1888
Overweight        1500
Healthy Weight    1261
Underweight        332
Name: bmi_binned, dtype: int64
```

Distribution of patients wrt to BMI Stats(binned)

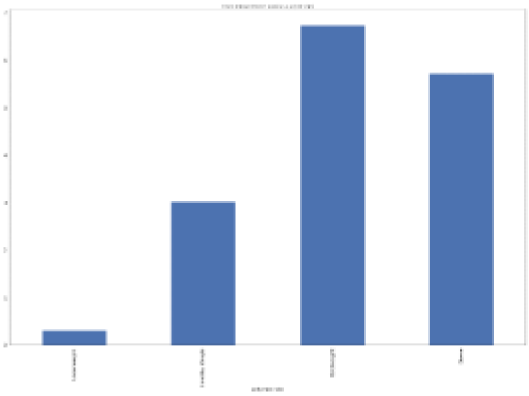


```
bmi_binned
Underweight          1
Healthy Weight      38
Overweight         101
Obese              108
Name: stroke, dtype: int64
```

Distribution of stroke positive patients wrt to BMI Stats(binned)



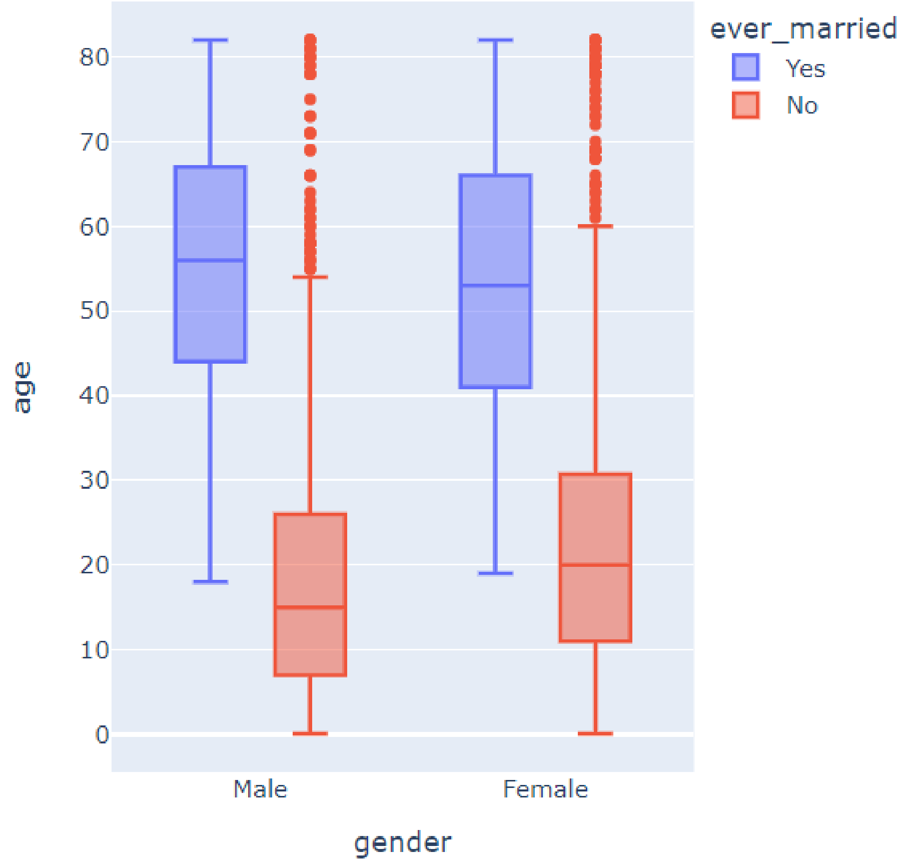
The disparity does not seem to be that big. It seems that the pattern the above graph is showing is quite faithful. But just to be on the safer side, let's take the percent.



As expected the frequency graph gave an almost faithful result.

Initial intuition seems correct.

Is there any relation b/w Marriage, Age and Gender?

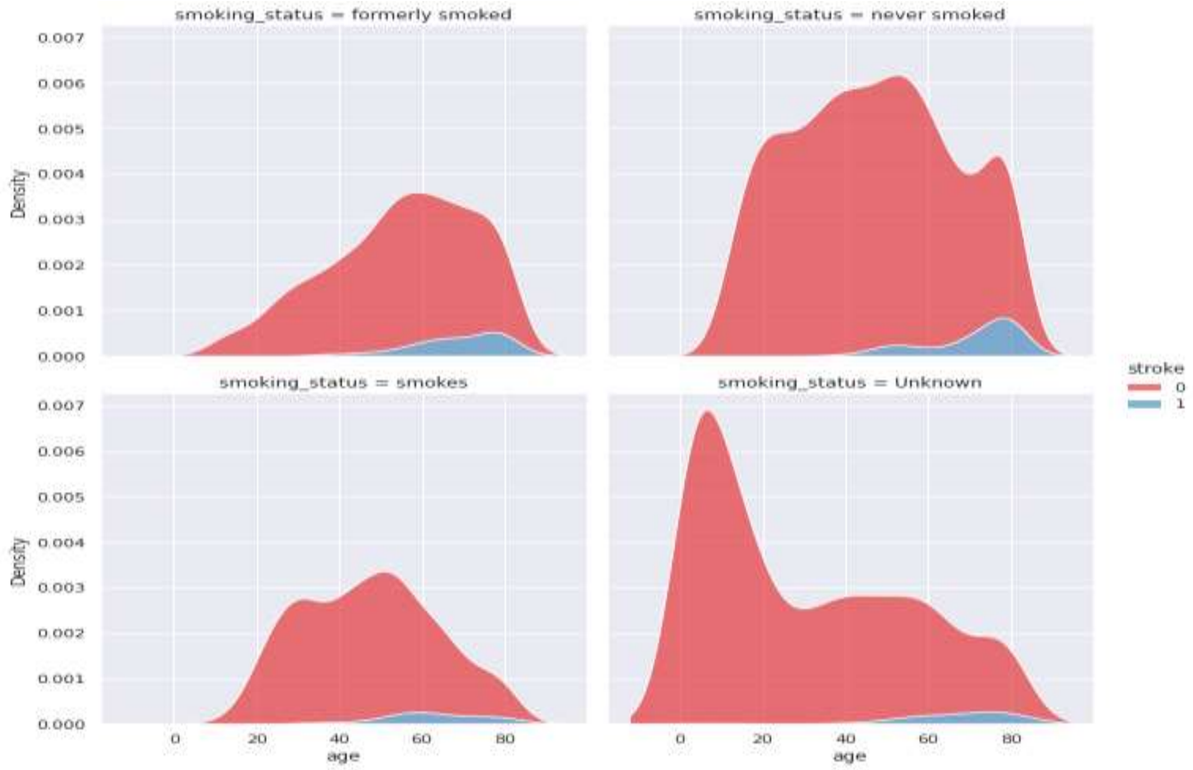


We can see that in our data we have abundance of married men and women at the age bracket of 50-60 years of age followed by young adults at the age group 18-22.

This is important insight as we can use it at later stages to draw out better inferences from the graphs.

Is there any relation b/w Smoking, Age and Stroke?

Initial intuition says that those who are non – smokers and older, will have more chances for getting a stroke. This multivariate assumption is taken with regards to our previous findings.



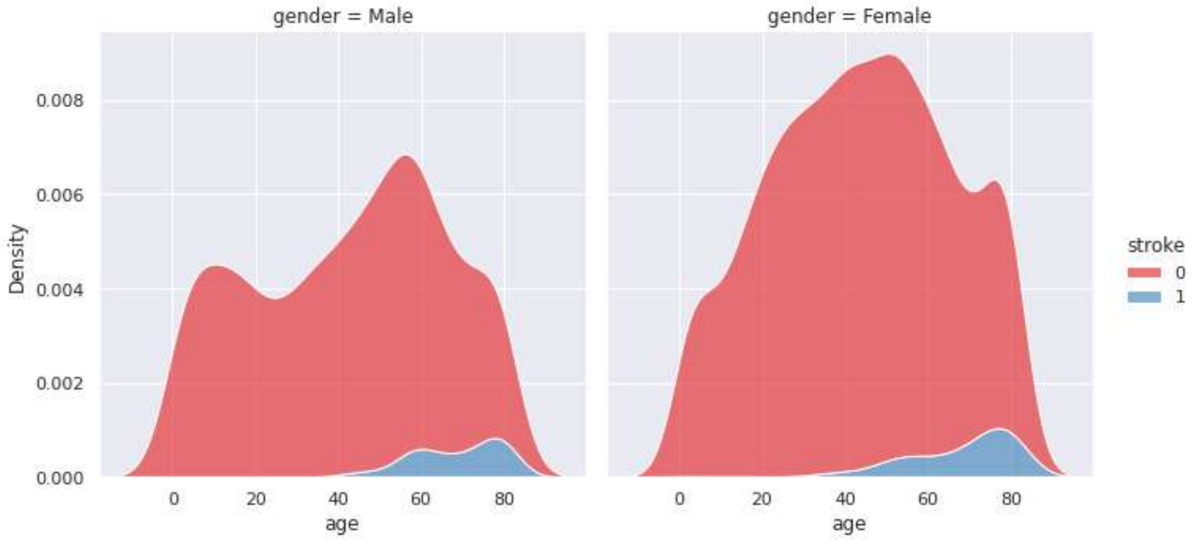
As we can see in the above graphs, strokes are more common in old people who were non – smokers. This proves our initial hypothesis correct.

Moreover, formal smokers were also at a higher risk of getting a stroke.

Finally, the smokers seem to be pretty safe from strokes, but it does not mean that smoking is good, it does a lot more damage to the body and should not be practiced.

Is there any relation b/w Gender, Age and Stroke?

Initial intuition suggests that middle aged men would be more susceptible to strokes. And men in general would be more prone to strokes. This multivariate assumption is taken with regards to our previous findings.



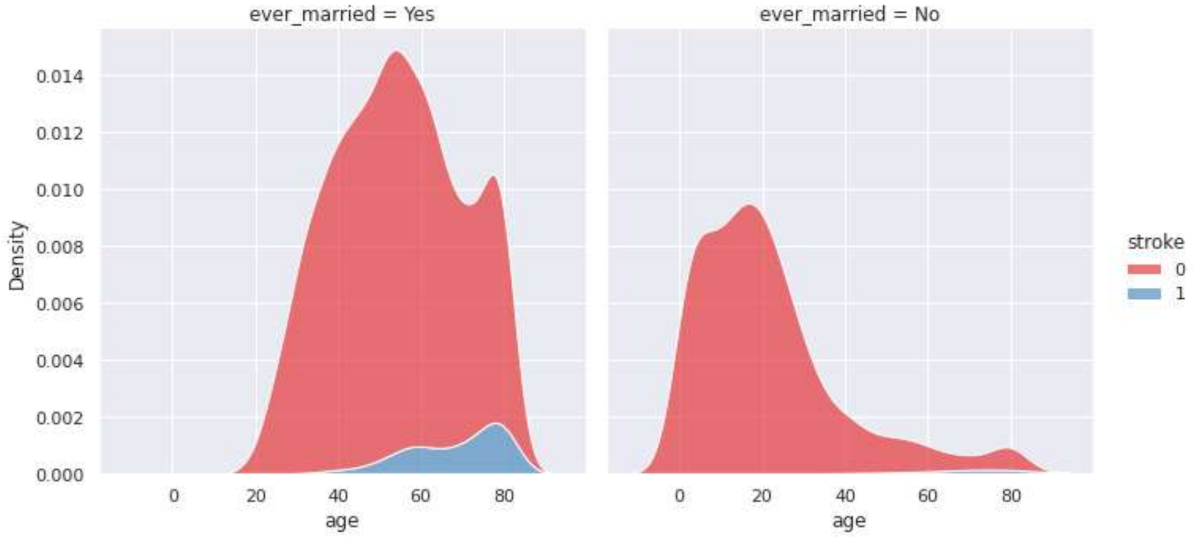
As evident from the above graph, older women are more common patients of stroke and in general women are more prone to having a stroke.

The graphs for both genders are pretty much similar in middle ages (40-60) but there is a rise as we look at old patients who are females.

These findings thoroughly disapprove of our initial intuitions and presents new insight in front of us.

Is there any relation b/w Marriage, Age and Stroke?

Initial intuition suggests that married people would be more susceptible to strokes. This multivariate assumption is taken with regards to our previous findings.



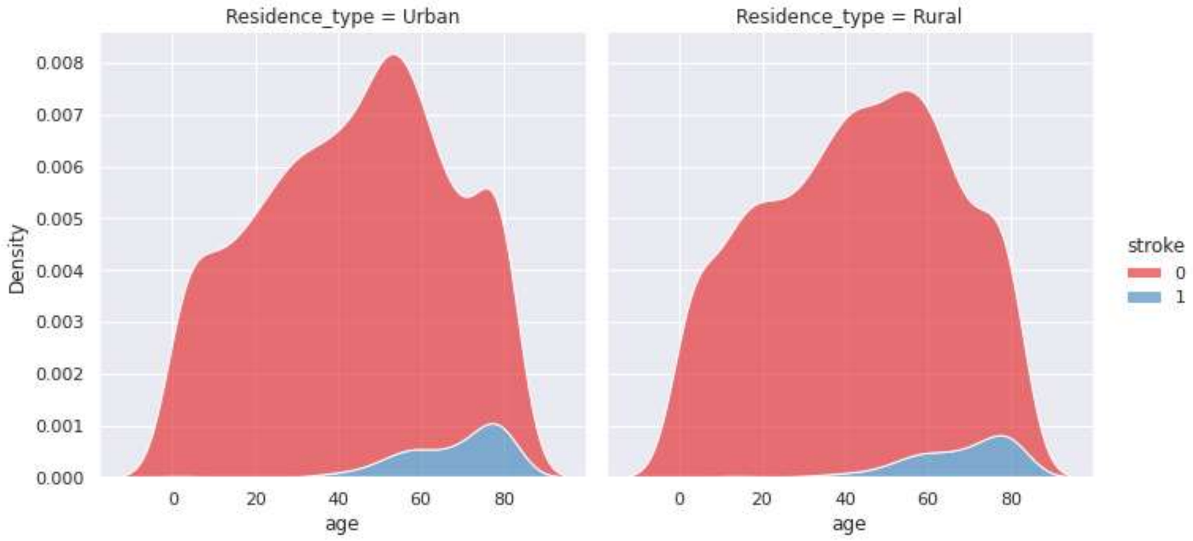
At first glance, the graphs give a very comical result. It says that unmarried people are at an almost ‘zero’ risk of stroke and that married people are very common patients of stroke.

It looks like that married people are candidates for stroke but in fact the stroke cases are strongly correlated with Age. Majority of unmarried people are less than 40 years old and therefore with low risk of stroke.

So, our intuition was comically truthful, if we just take the numbers as the only argument.

Is there any relation b/w Residence, Age and Stroke?

Initial intuition suggests that people who live in urban areas would be more susceptible to strokes. This multivariate assumption is taken with regards to our previous findings.

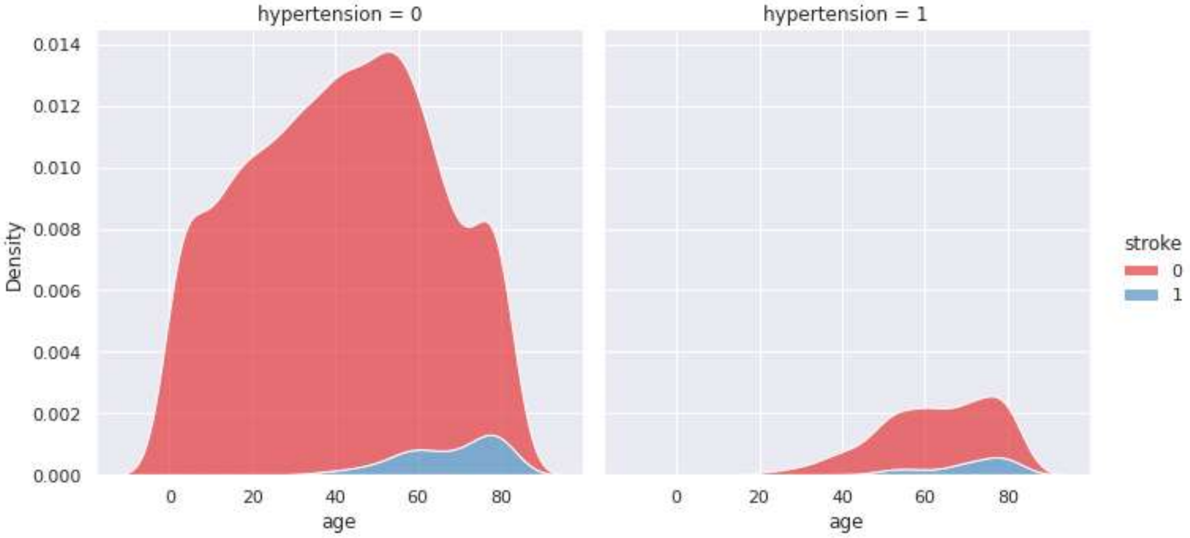


Our initial intuition seems to be correct as people in urban areas are more prone to strokes.

The main reason would possibly be the stressful and unrestful life that people in cities have to live. The workplace grind may be a very important factor in this insight.

Is there any relation b/w Hypertension, Age and Stroke?

Initial intuition suggests that old patients who have hypertension would be more susceptible to brain strokes. As both are diseases related to brain, thus it biases our reasoning. This multivariate assumption is taken with regards to our previous findings.



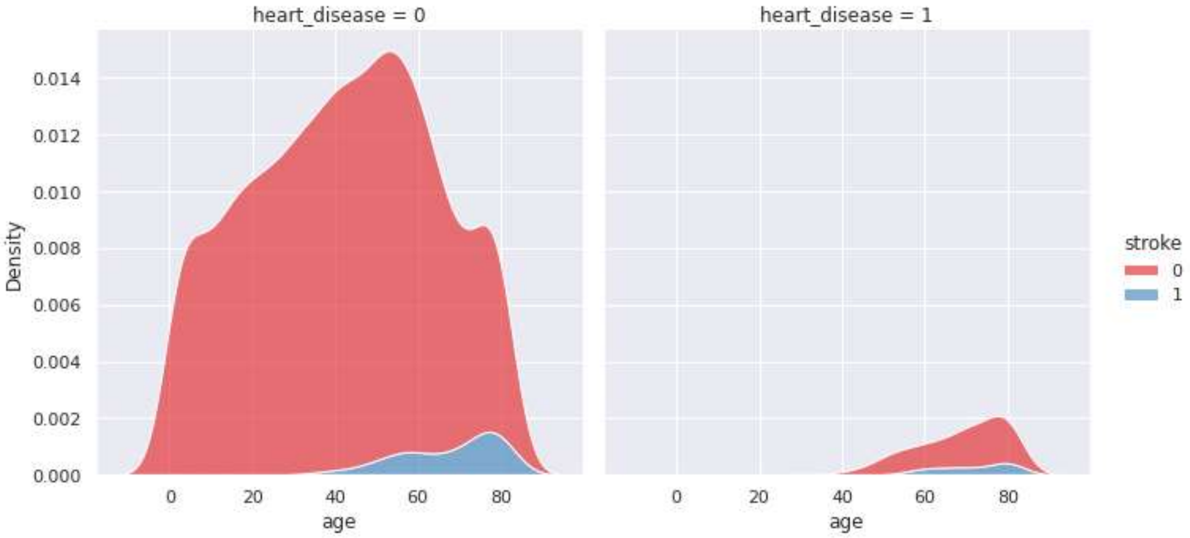
By seeing the graph, we can say that our initial intuition has been proved wrong and hypertension patients are comparatively better off than stroke patients.

But we should also not miss the point that there is a very vast difference in the number of patients who suffer from hypertension and the number of patients who don’t.

So, our hypothesis is wrong.

Is there any relation b/w Heart Disease, Age and Stroke?

Initial intuition suggests that old patients who have hypertension would be more susceptible to brain strokes. As the reason for stroke is that the brain does not get enough oxygen supply which is directly related to lack of pumping of the required volume of blood by the heart. This multivariate assumption is taken with regards to our previous findings.

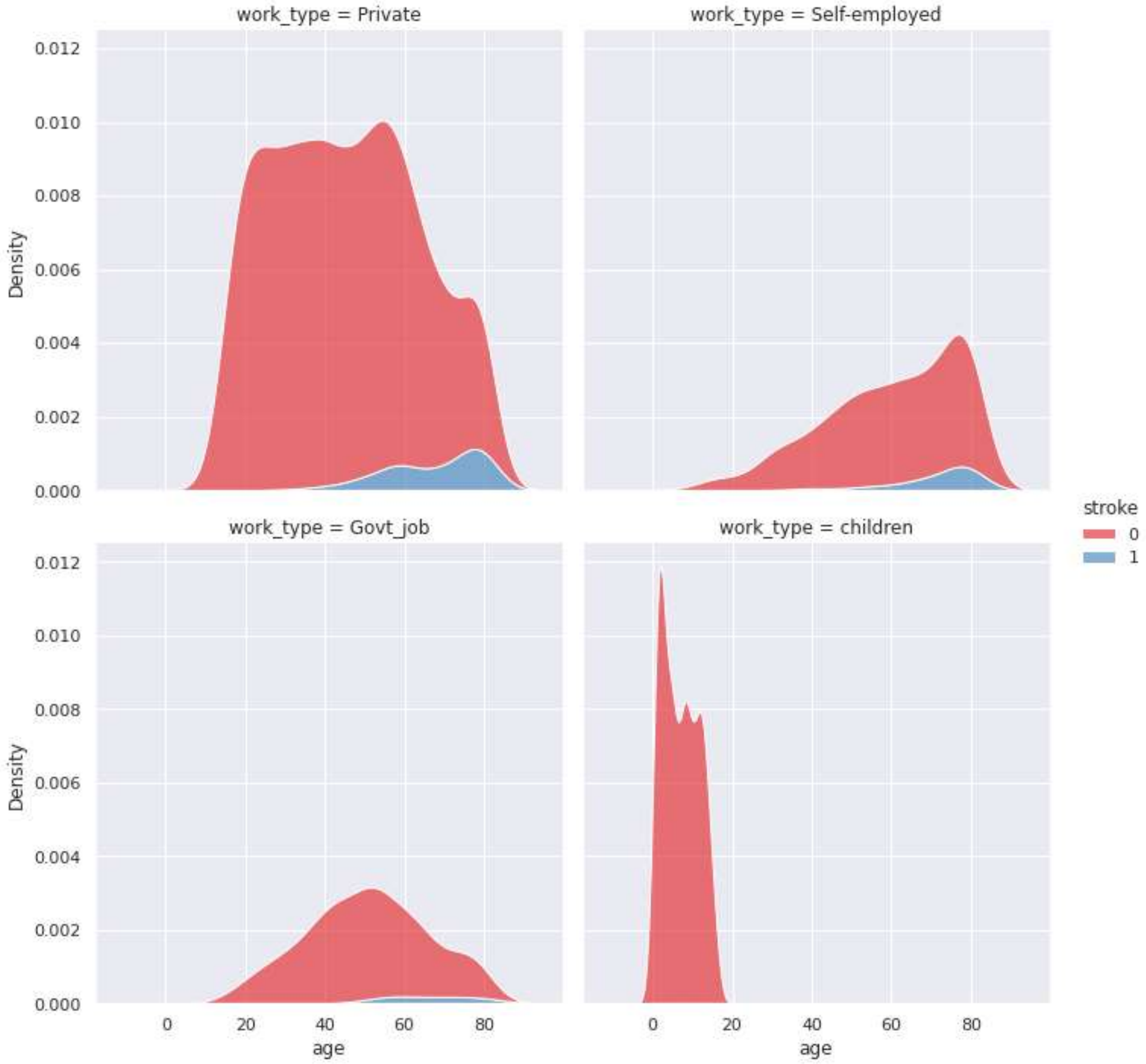


The inference from this graph is similar to the previous one.

At first glance it ay seem that our intuition was correct but as we can see that the there is again a discrepancy in the numbers of patients suffering from heart disease and those who don’t, results are opposite to our intuition.

Is there any relation b/w Job Type, Age and Stroke?

Our initial intuition says that older people in private sector and self-employed people are more likely to have a stroke.



By seeing the graph we can safely say that our initial hypothesis is true.

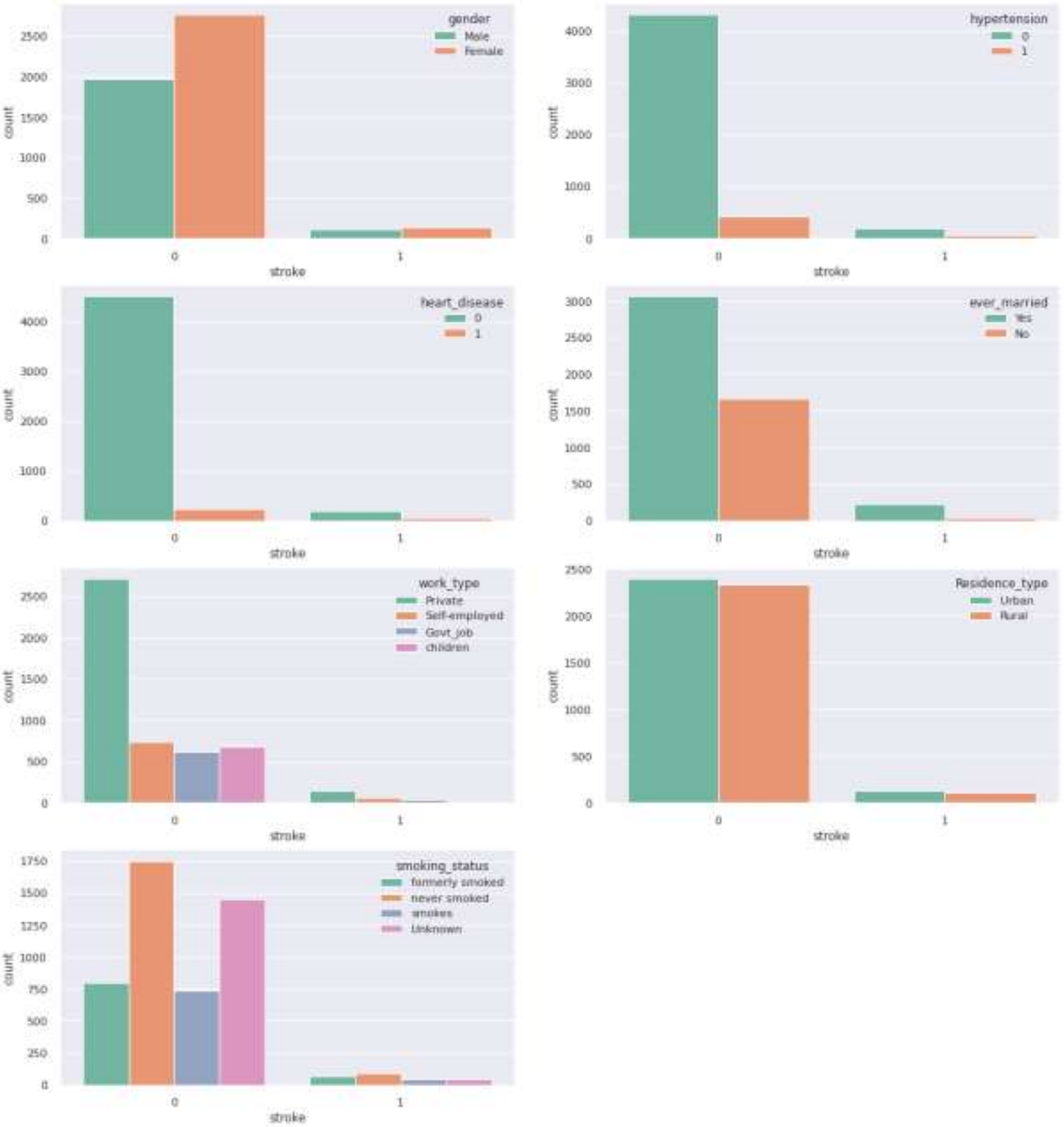
Old people in public sector are most susceptible to a stroke followed by old self-employed patients.

Government employees are at a very low risk of stroke and children are at zero risk.

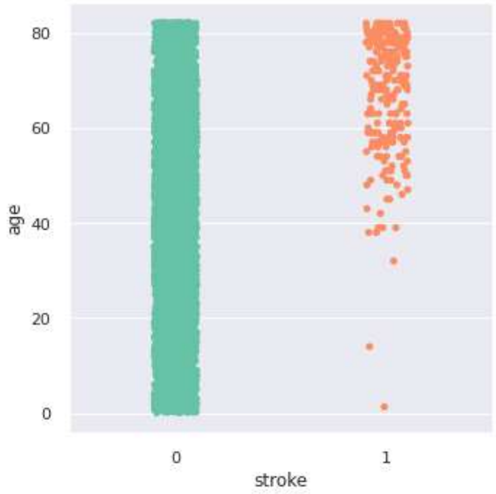
Thus we can safely conclude that work place stress plays are very important role in causing brain strokes.

Conclusion

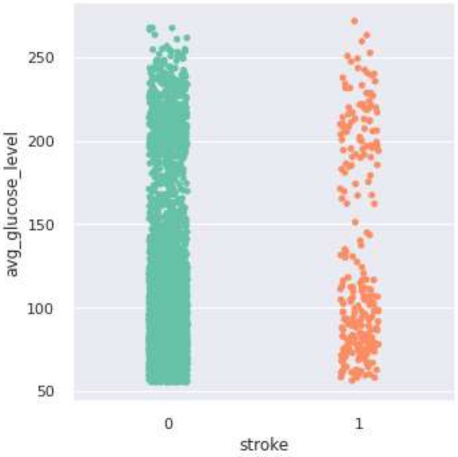
- The following trends in stats are clear by the analysis on the dataset:



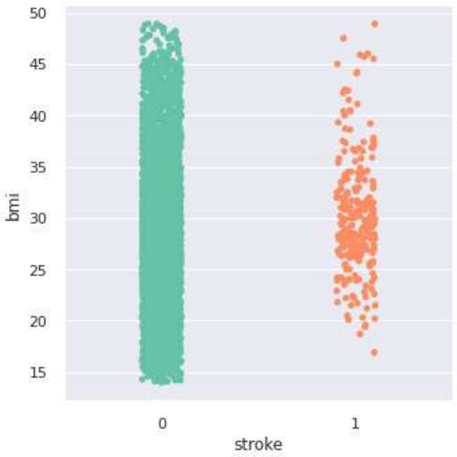
- When we look at our age variable, we can see that only from 40 years old people start to have Stroke, and cases below that age are very rare, the trend is that the older the person, the more likely he is to have a stroke.



- When we look at our Glucose variable we can see that it is well distributed.



- When we look at the BMI we can see that it's more common for people to have a stroke when the BMI is on average, not too high and not too low.



- More stress leads to higher risk of stroke as seen through the distribution in workplace, marital status and hypertension.

- Diseases like diabetes, heart diseases and hypertension increase the chance of stroke.

- Gender and residence don’t play a significant role in predicting stroke.

- Showing results against common intuition, active smokers are at a lesser risk of stroke than non-smokers.

- Our data has an abundance of married men and women at the age bracket of 50-60 years of age followed by young adults at the age group 18-22.

- Older non-smokers are at the greatest risk of stroke, followed by older formal smokers. As interesting as it may seem, smokers were the safest from strokes.

- Old women are the most susceptible to strokes. Moreover, women in general are more common patients of stroke. Meanwhile men seem to be safer than women in this regard.

- Just going by the stats and numbers, we can see that married people are more probable to get a stroke and unmarried people are at zero risk of stroke. But there is a twist that most unmarried people were children or young adults.

- Urban life is more stressful and as the person ripens with age, they are more likely to get a brain stroke. Rural people are a bit better off in this regard.

- Our intuition that people with a major disease namely heart disease and hypertension would be more likely to have a stroke was proved wrong by the analysis.

- Over all, the most important inference and pattern from this analysis is that age is one of the most important factor in deciding that someone would have a stroke or not. The patterns show that people with more age are more likely to get a stroke attack than the younger people across all other variables.

- It was seen across all graphs that young people had less stroke patients and old people had more.

REFERENCES AND LINKS

Dataset :-

https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

GitHub Project link :-

https://github.com/ananyaaD/EDA-Project

Thank You