Stroke Analysis and Prediction
Mini project 3

25 Feb 2022 by Derek Tan

Agenda

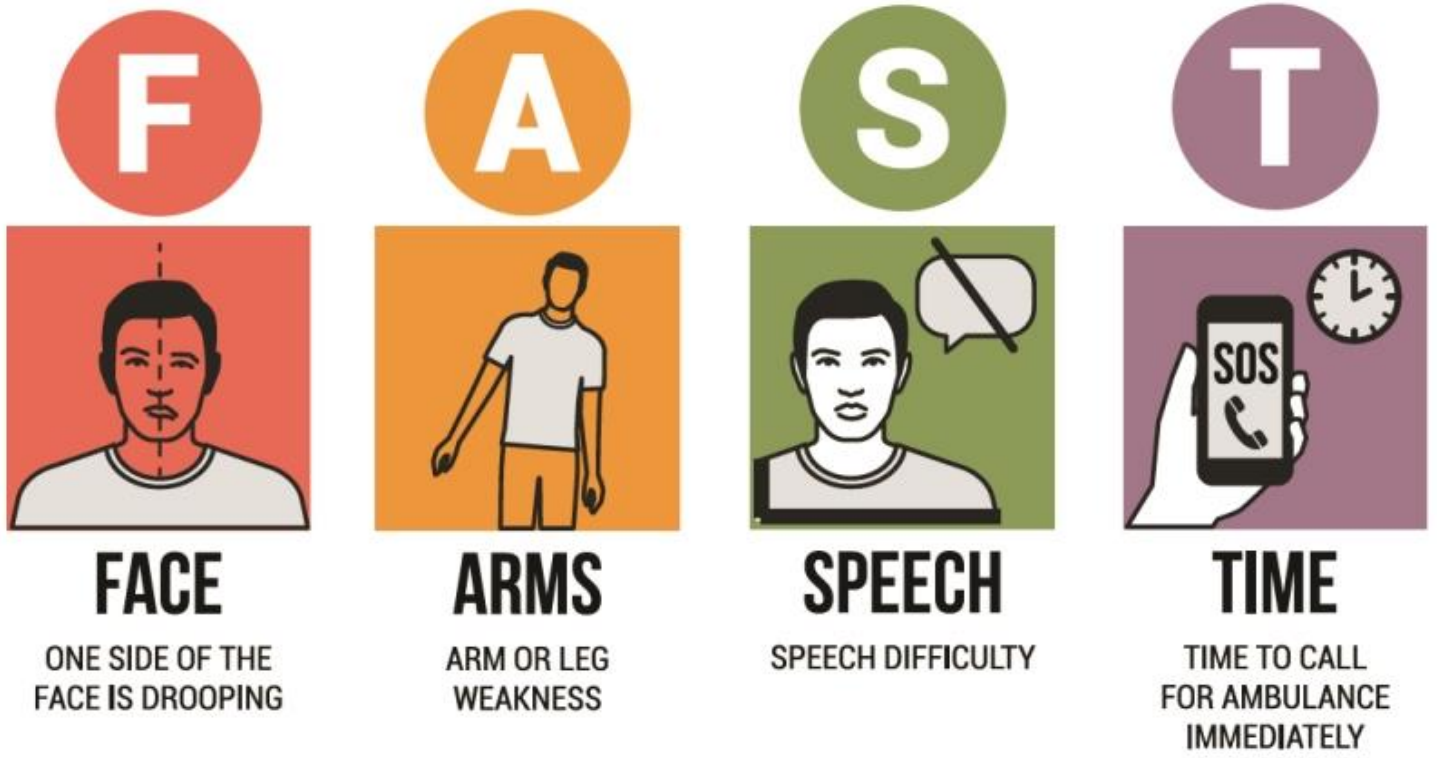
- Introduction
- Objective
- Exploratory Data Analysis
- Model
- Model Evaluation
- Conclusion
- Future Work

Introduction

What is Stroke?

Stroke is a sudden change in the blood supply to a part of the brain, sometimes causing a loss of the ability to move a particular part of the body.

How To Spot A Stroke?



Objective

Objective

Gain insights on what are the external factors that cause stroke

Build a Machine Learning Model to Predict whether a person had a stroke or not

Exploratory Data Analysis

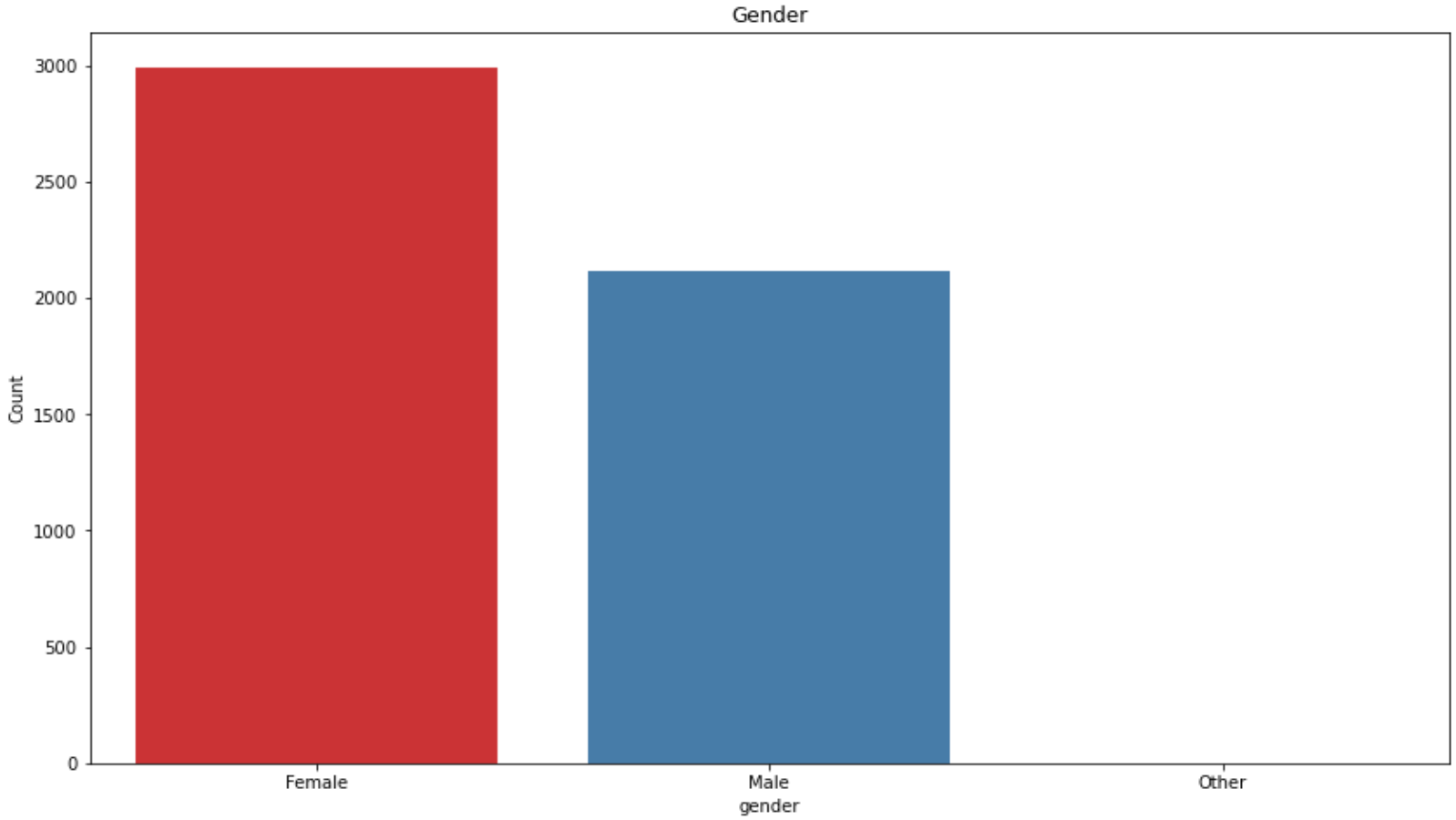
Features

- Gender
- age
- Hypertension
- heart_disease
- ever_married
- work_type
- Residence_type
- avg_glucose_level
- bmi
- smoking_status

How big is the dataset?

Total **5110** data for analysis

Gender




```
Female    2994
Male      2115
Other        1
Name: gender, dtype: int64
```

Gender

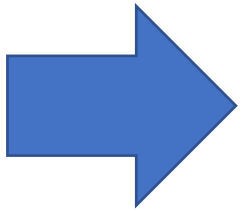
	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
3116	56156	Other	26.0	0	0	No	Private	Rural	143.33	22.4	formerly smoked	0

Feature Engineering

Stroke

```
df.stroke.value_counts()
```

```
0    4861
1     249
Name: stroke, dtype: int64
```

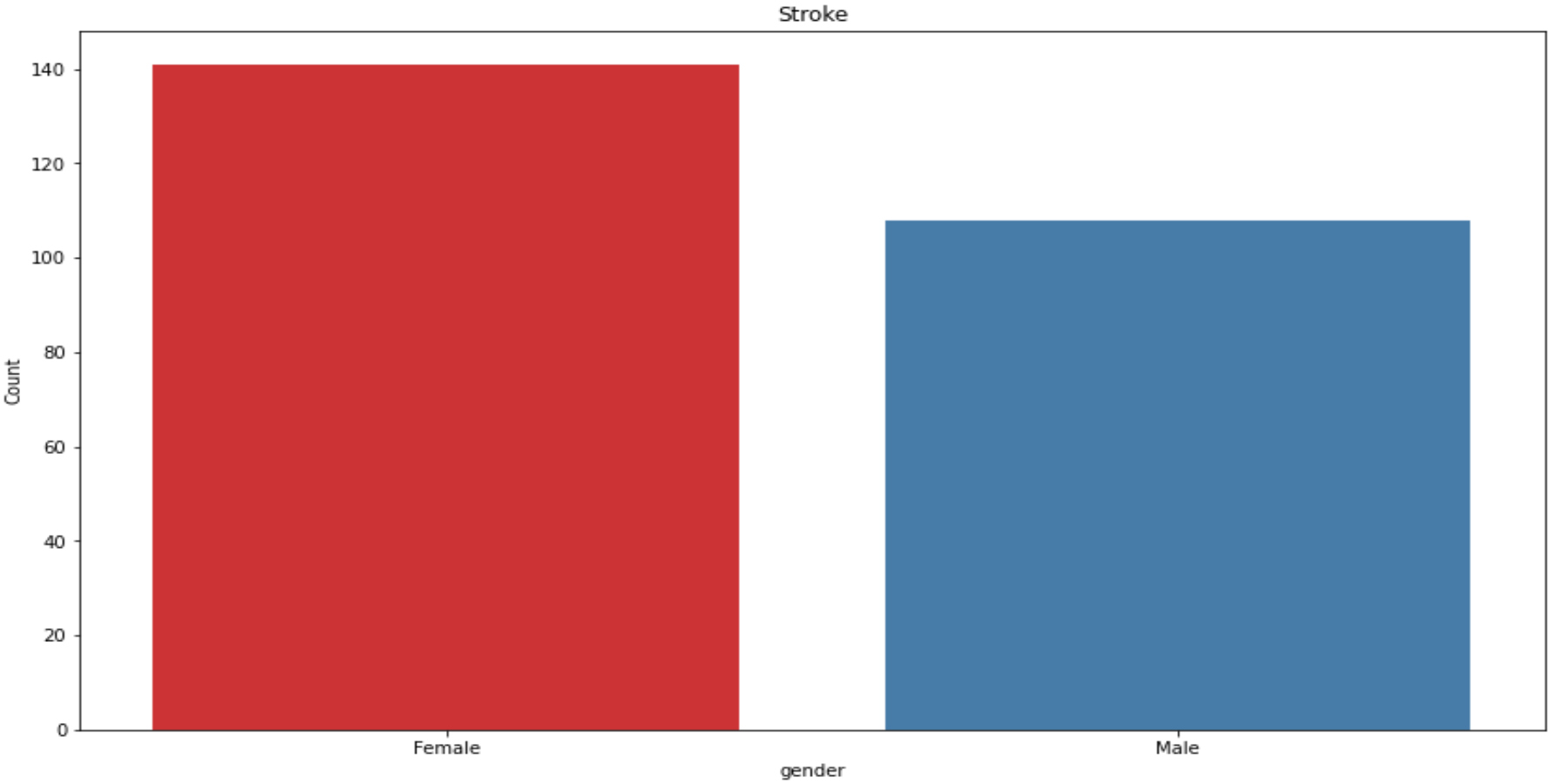


```
stroke1 = df.loc[(df['stroke'] == 1)]
```

```
stroke1.sample(5)
```

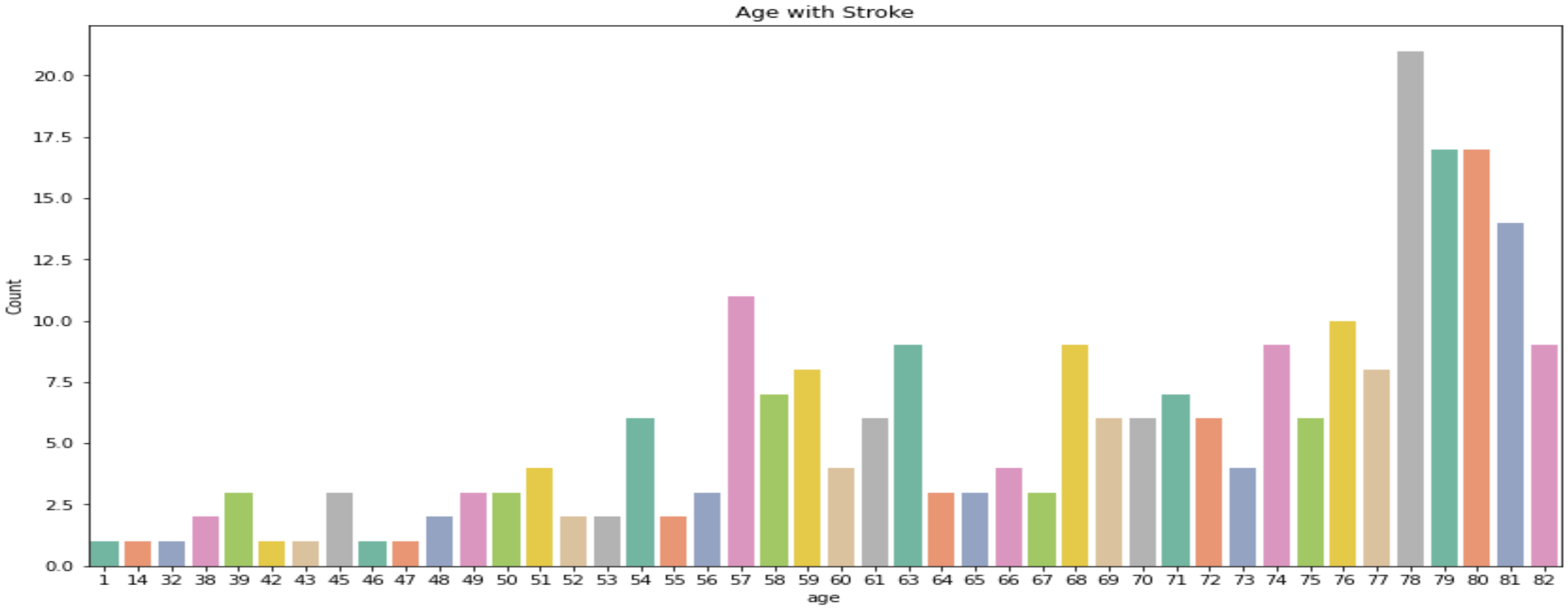
	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
164	3512	Female	70.0	1	0	Yes	Self-employed	Urban	89.13	34.2	formerly smoked	1
73	50784	Male	63.0	0	0	Yes	Private	Rural	228.56	27.4	never smoked	1
177	36841	Male	78.0	1	0	Yes	Self-employed	Rural	56.11	25.5	formerly smoked	1
123	44033	Male	56.0	1	0	Yes	Private	Rural	249.31	35.8	never smoked	1
52	59190	Female	79.0	0	1	Yes	Private	Rural	127.29	27.7	never smoked	1

Gender(Stroke)

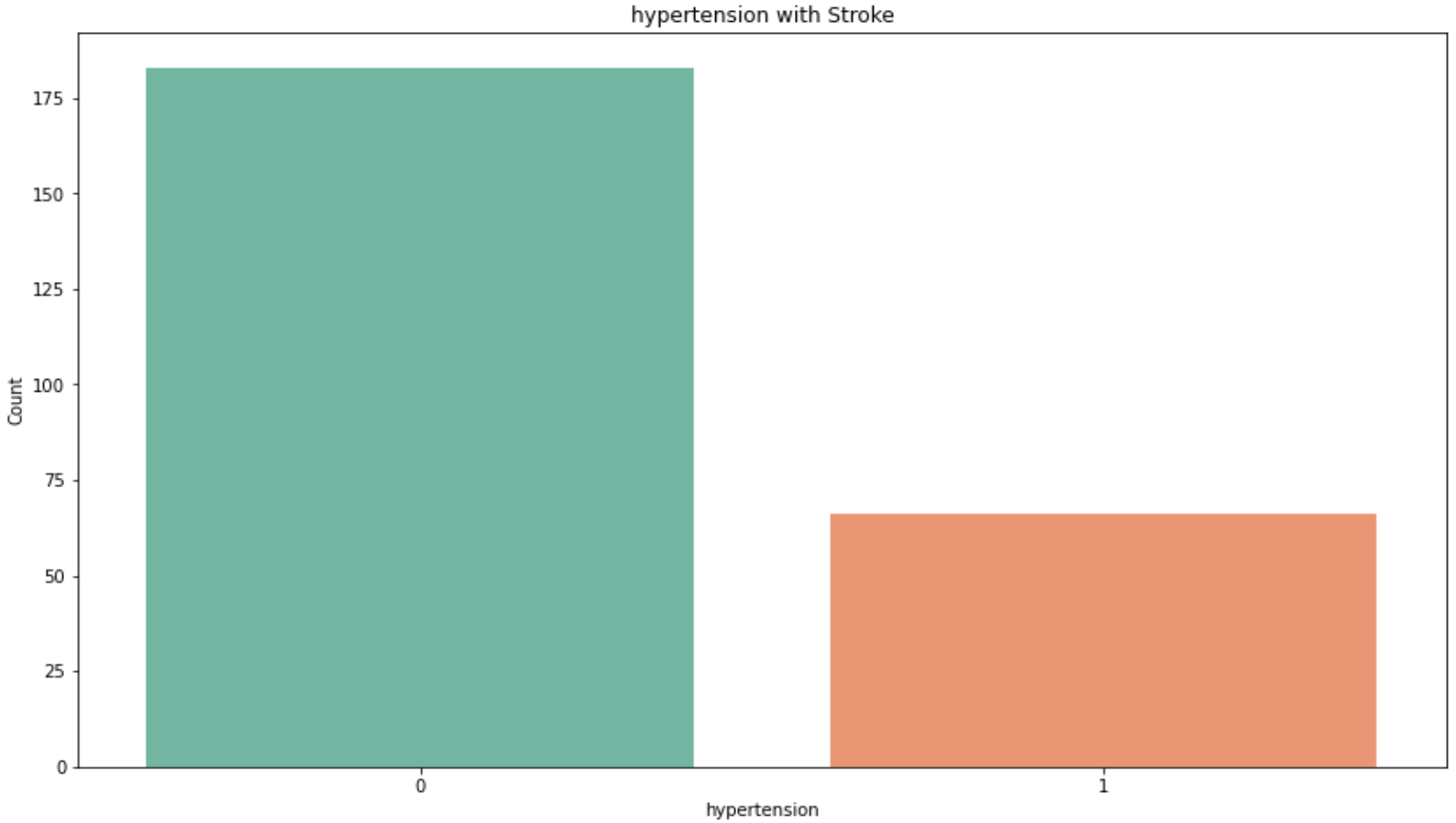



```
Female    141
Male      108
Name: gender, dtype: int64
```

Age with Stroke

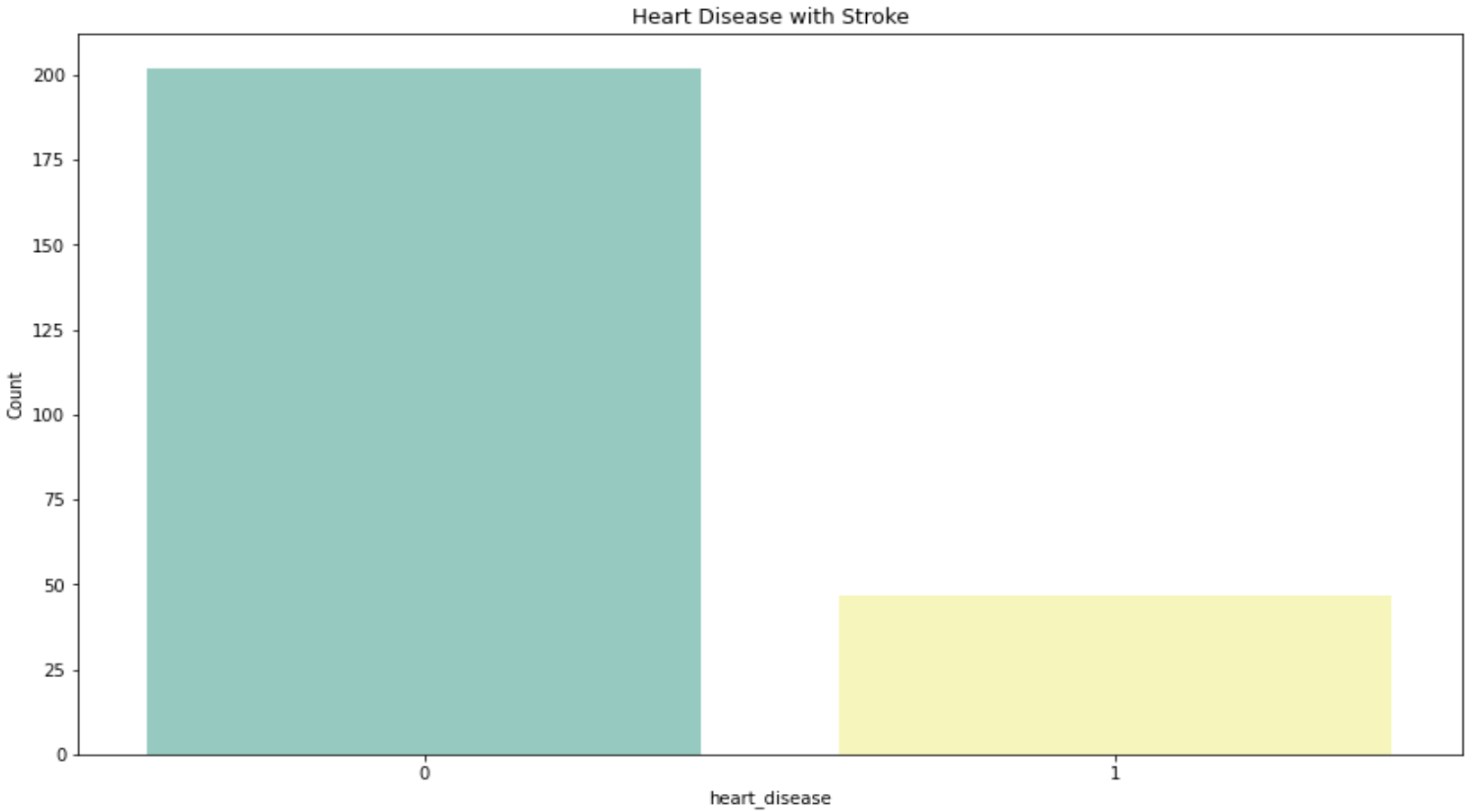


Hypertension with Stroke



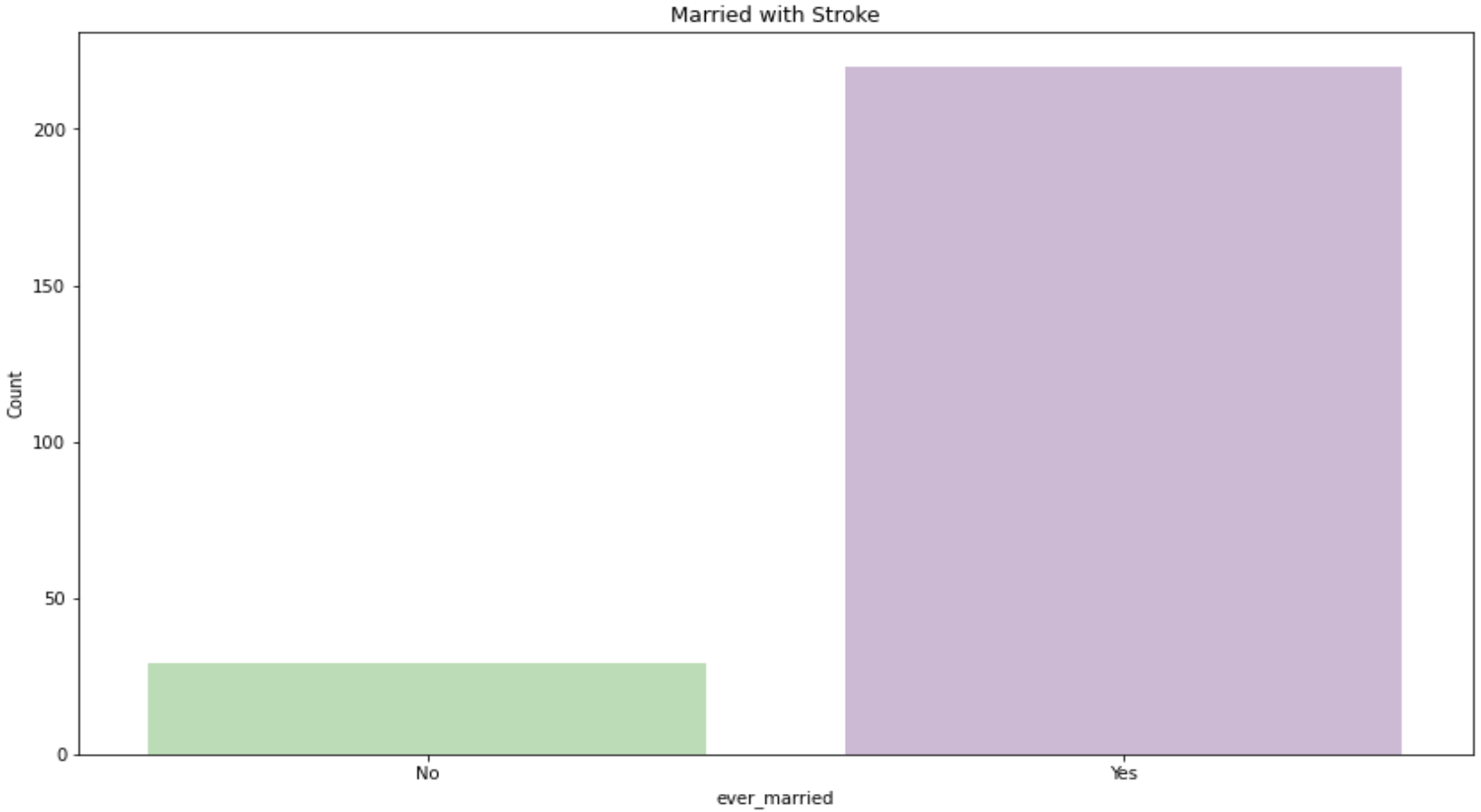

```
: 0    183
  1     66
  Name: hypertension, dtype: int64
```

Heart Disease with Stroke



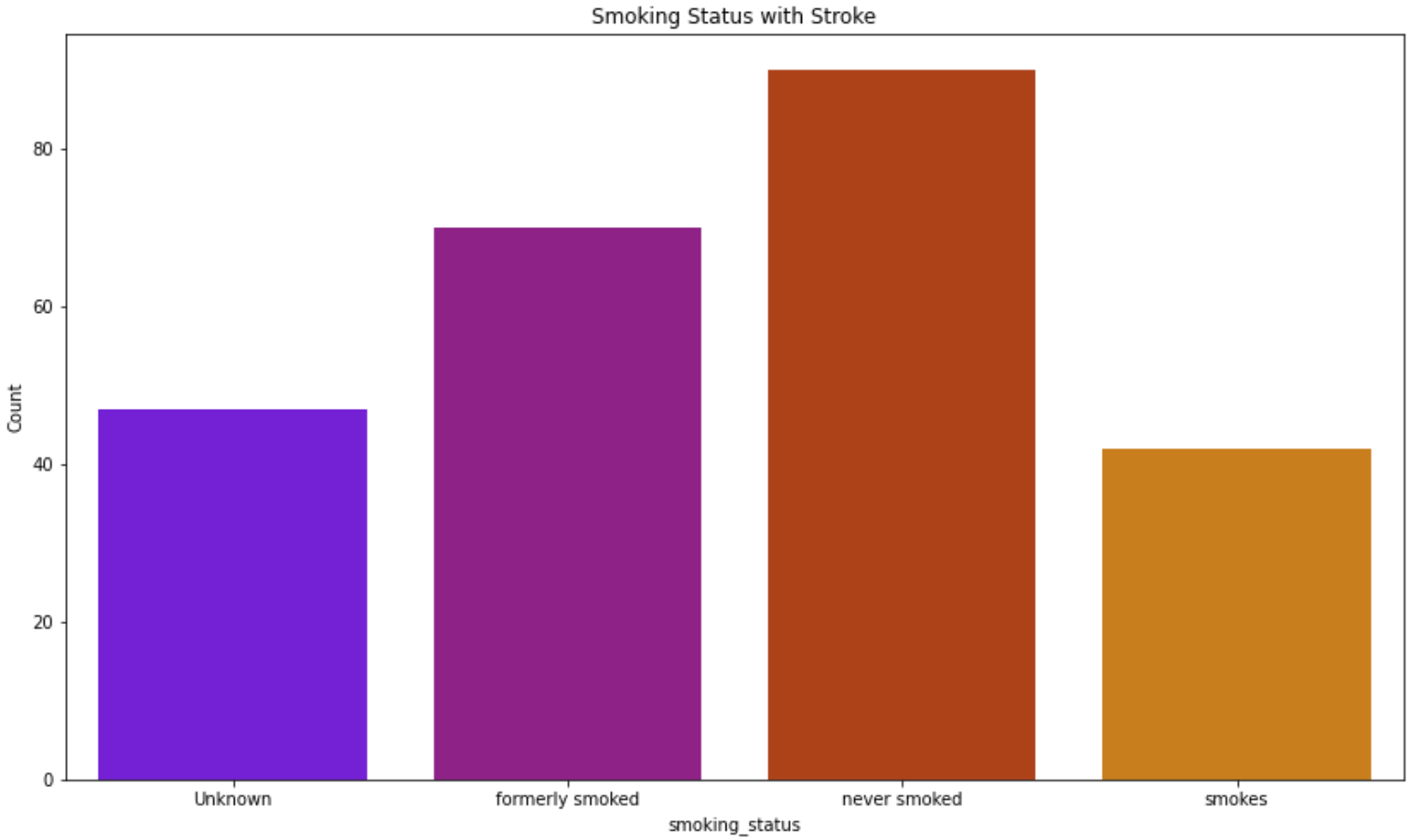

```
0    202
1     47
Name: heart_disease, dtype: int64
```

Married with Stroke



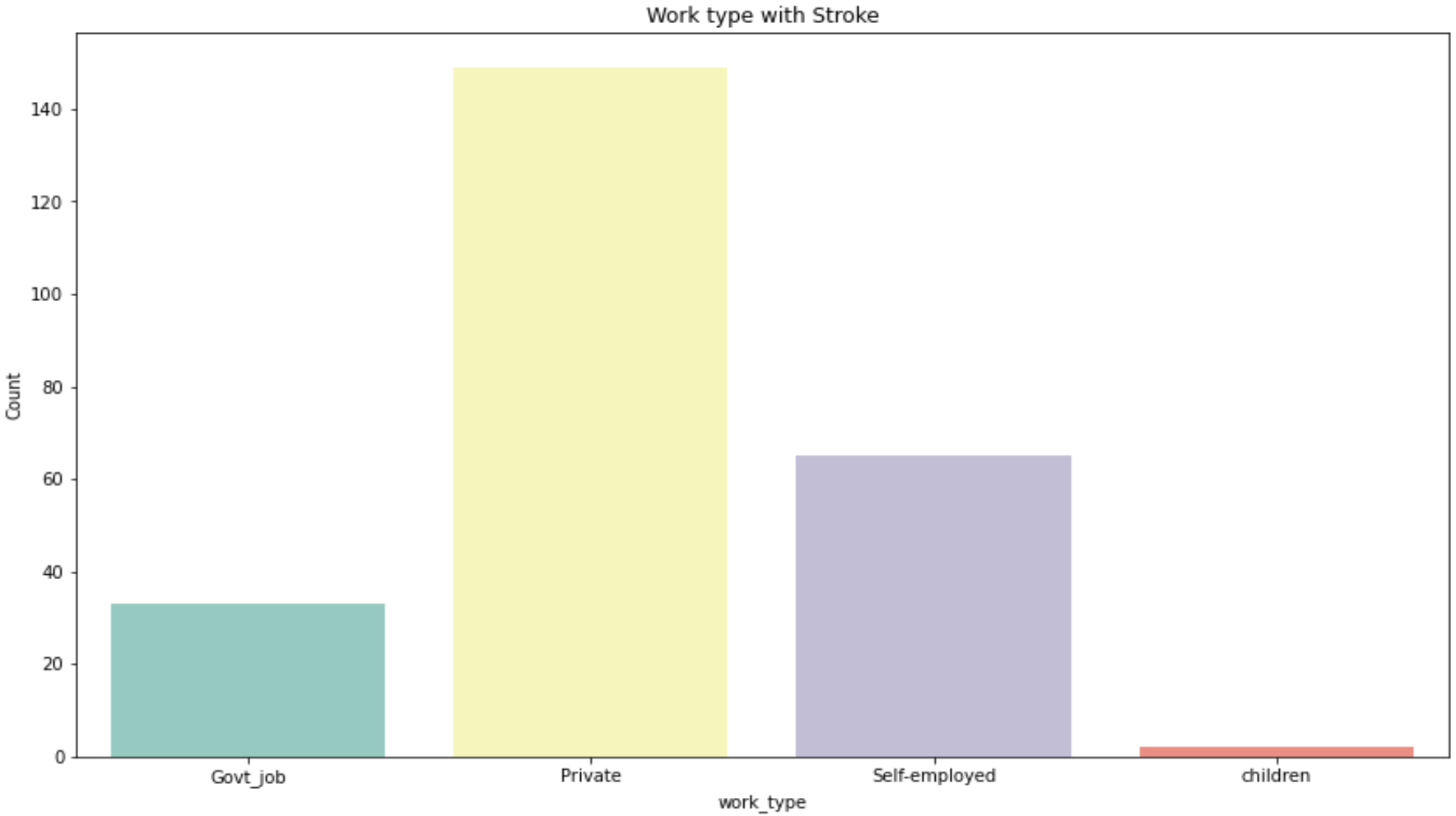

```
Yes    220
No      29
Name: ever_married, dtype: int64
```

Smoking Status With Stroke



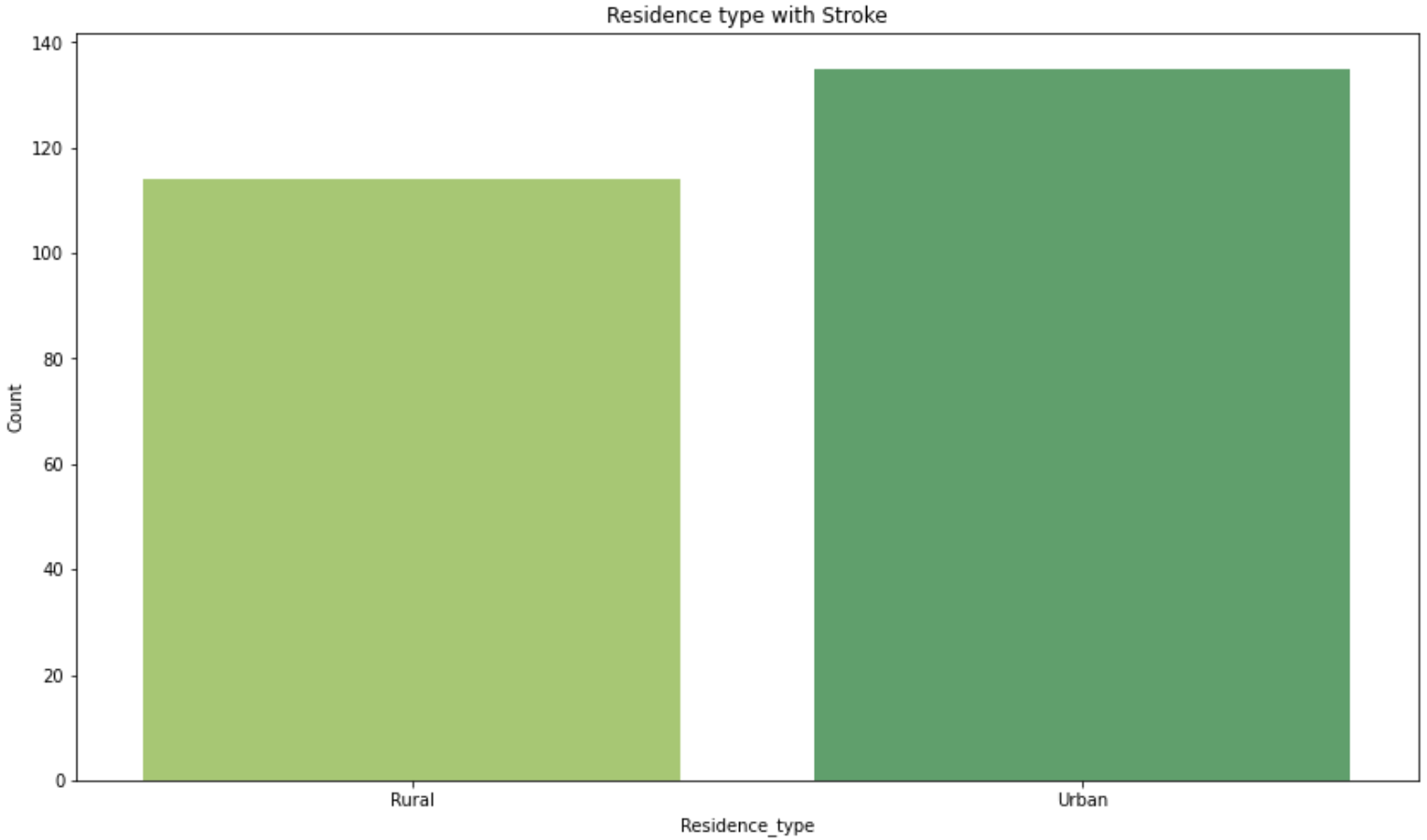

```
never smoked       90
formerly smoked    70
Unknown            47
smokes             42
Name: smoking_status, dtype: int64
```

Work Type with Stroke

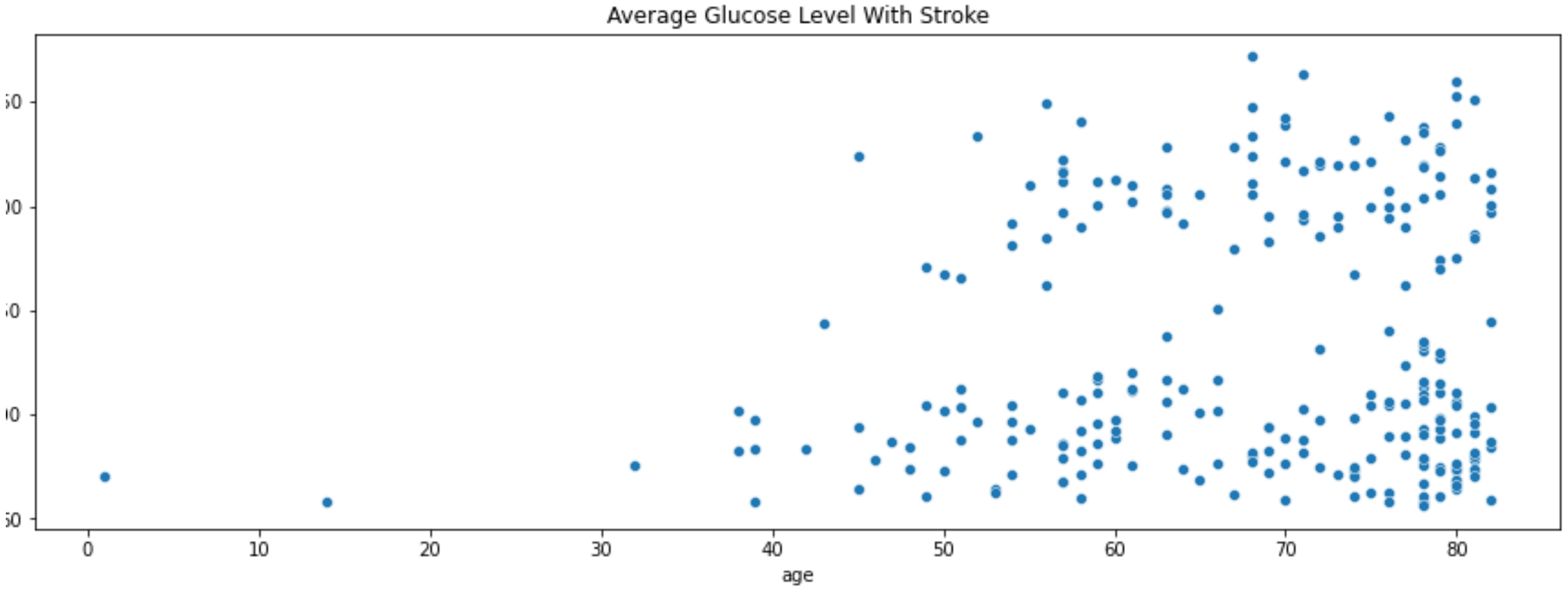



```
Private          149
Self-employed     65
Govt_job          33
children           2
Name: work_type, dtype: int64
```

Residence Type With Stroke




```
Urban    135
Rural    114
Name: Residence_type, dtype: int64
```



Average Glucose Level With Stroke

```
: print('Total Non Diabeties Patients with stroke :' , blood140less['avg_glucose_level'].count())
```

```
Total Non Diabeties Patients with stroke : 156
```

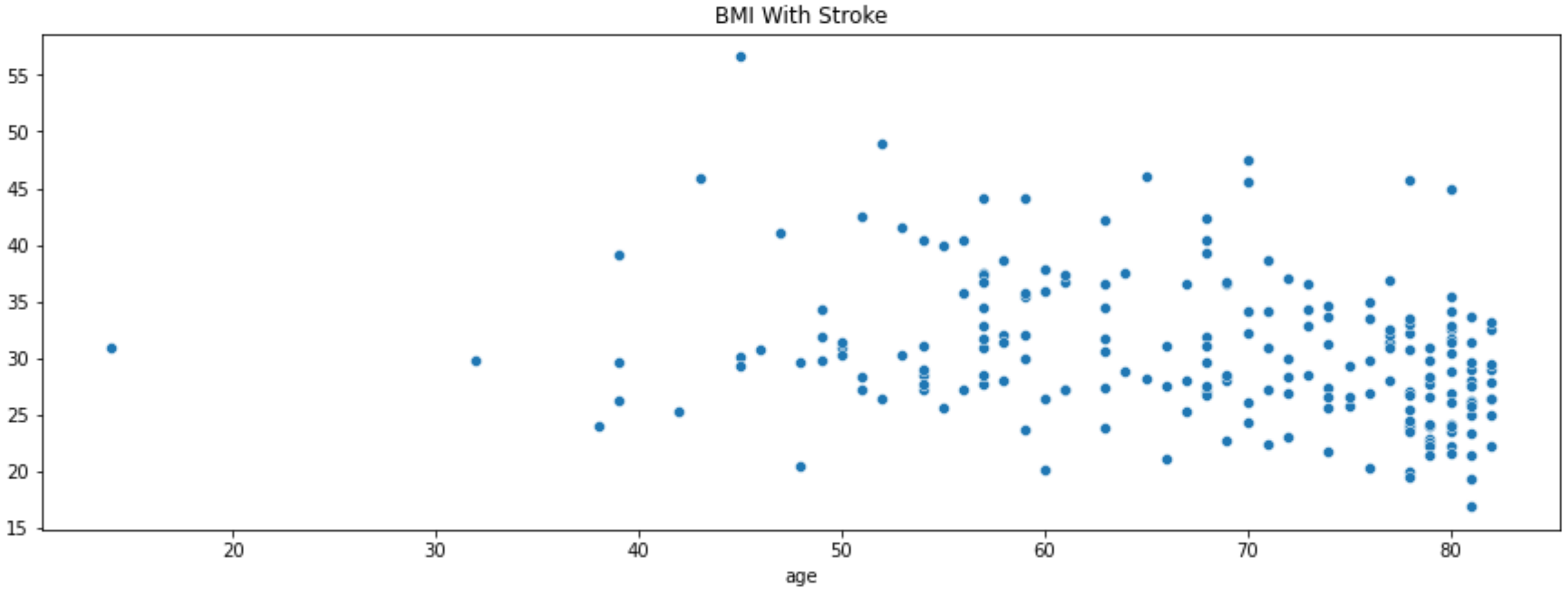
```
: print('Total Prediabetes Patients with stroke : ', blood140199['avg_glucose_level'].count())
```

```
Total Prediabetes Patients with stroke :  37
```

```
: print('Total Diabetes Patients with stroke : ', blood200over['avg_glucose_level'].count())
```

```
Total Diabetes Patients with stroke :  59
```

Average Glucose Level With Stroke



BMI with Stroke

```
print('Total Underweight Patients with stroke : ', underweight['bmi'].count())
```

```
Total Underweight Patients with stroke :  59
```

```
print('Total healthy Patients with stroke : ', healthy['bmi'].count())
```

```
Total healthy Patients with stroke :  35
```

```
print('Total overweight Patients with stroke : ', overweight['bmi'].count())
```

```
Total overweight Patients with stroke :  75
```

```
print('Total obese Patients with stroke : ', obese['bmi'].count())
```

```
Total obese Patients with stroke :  96
```

BMI with Stroke

Data cleaning

	Total	Percent
bmi	201	0.039335
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

Data Preprocessing

- 5 features: GENDER, EVER_MARRIED, WORK_TYPE, RESIDENCE_TYPE, SMOKING_STATUS (Convert from categorical to numeric data)

```
from sklearn.preprocessing import LabelEncoder
enc=LabelEncoder()
```

```
gender=enc.fit_transform(df['gender'])
smoking_status=enc.fit_transform(df['smoking_status'])
work_type=enc.fit_transform(df['work_type'])
Residence_type=enc.fit_transform(df['Residence_type'])
ever_married=enc.fit_transform(df['ever_married'])
```

```
df['ever_married']=ever_married
df['Residence_type']=Residence_type
df['smoking_status']=smoking_status
df['gender']=gender
df['work_type']=work_type
```

Model

Model

- Decision Tree
- Decision Tree with Bagging
- Random Forest
- Random Forest With Adaboost
- XG Boost

Solving imbalance problem

```
Before OverSampling, the shape of train_x: (4088, 10)
Before OverSampling, the shape of train_y: (4088,)
Before OverSampling, counts of label 1: 199
Before OverSampling, counts of label 0: 3889

After OverSampling, the shape of train_x: (7778, 10)
After OverSampling, the shape of train_y: (7778,)
After OverSampling, counts of label 1: 3889
After OverSampling, counts of label 0: 3889
```

Model Evaluation

Model Evaluation

	Model	Accuracy	Precision	Recall	ROC_AUC
0	Decision Tree	1.000000	1.000000	1.000000	1.000000
1	Decision Tree With Bagging	1.000000	1.000000	1.000000	1.000000
2	Random Forest	1.000000	1.000000	1.000000	1.000000
3	Random Forest With AdaBoost	1.000000	1.000000	1.000000	1.000000
4	XG Boost	0.989843	0.982278	0.997686	0.989843

Conclusion

Older patients are more likely to suffer a stroke than younger ones

Unmarried reduces the risk of stroke

Working as private has the highest number of stroke cases

Healthy BMI decrease the risk of stroke

Future Work

GET MORE DATA

DEPLOYMENT

Thank you!

Reference

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