Project DATA 5100

Exploratory Data Analysis

December 2022

Overview

In this data set we have an unbalanced population, where 5% of the sample has the characteristic of interest, which is having had a stroke, compared to 95% who have not had one. As a sociodemographic summary of this sample we have the following:

- 58% of the sample are women
- The average age of the sample is 43 years.
- 65% have ever been married
- 57% work in a private company
- 50% of the sample lives in an urban residence. In addition, as a summary of characteristics associated with health, we have the following:
- 10% of the sample has hypertension
- 5% suffer from heart disease
- The average level of glucose found is 106
- The average BMI of the sample is 28%.
- 32% of the sample smokes or has ever smoked.

When doing an analysis of the variables compared with the characteristic of interest: having had a Stroke, the following are found as significant values:

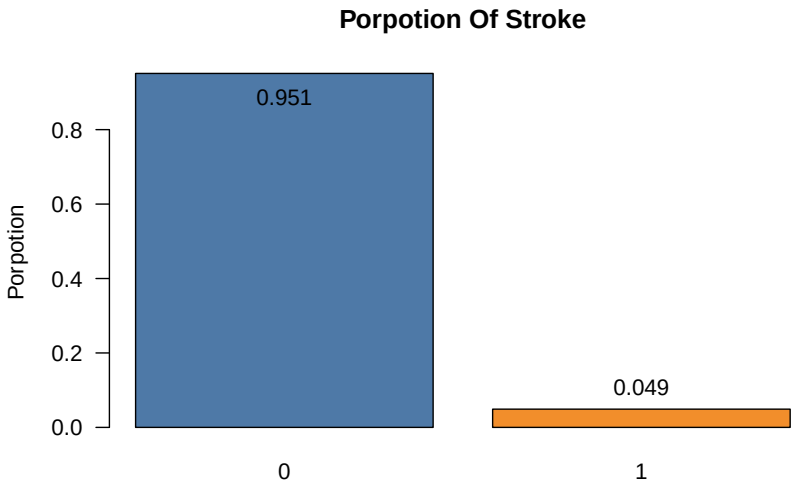
- 13% of people who have hypertension have had stroke, this value is higher than 4% of those who do not have hypertension and have had a stroke.
- 17% of people who suffer from heart disease have had a stroke, this value is higher than 5% of those who do not suffer from heart disease and have had a stroke.
- People who have had a stroke are on average 68 years old, while those who have not had a stroke are on average 42 years old.
- People who have had a stroke have an average blood glucose level of 132 while those who have not had a stroke have an average blood glucose level of 104.
- People who have suffered a stroke have an average BMI of 30.47 while those who have not suffered a stroke have an average BMI of 28.82.
- In the other variables analyzed, no significant differences were found.

From the above we can conclude that suffering hypertension and heart problems increase the risk of suffering a stroke. Likewise, strokes are directly related to older people, with a high glucose level and high BMI.

EDA

Stroke

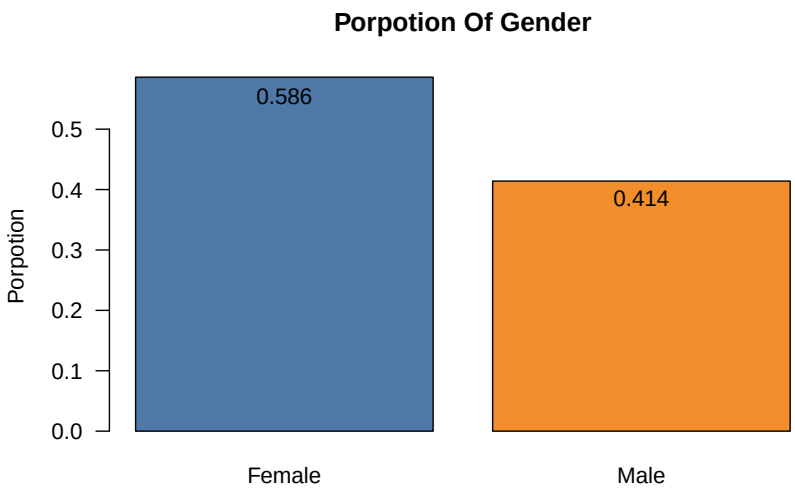
Stroke	Porpotion
0	0.951
1	0.049

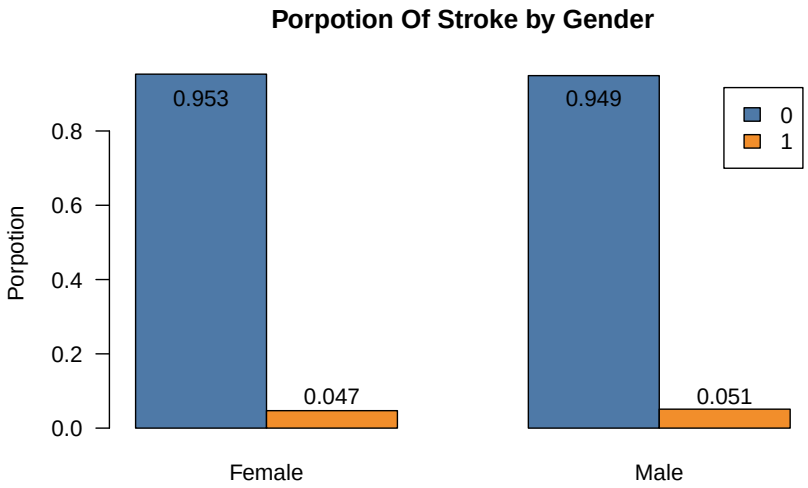


We have an unbalanced population, where 5% of the people who have suffered a stroke have suffered a stroke compared to the remaining 95% who have not suffered a stroke.

Gender

Gender	Porpotion
Female	0.586
Male	0.414





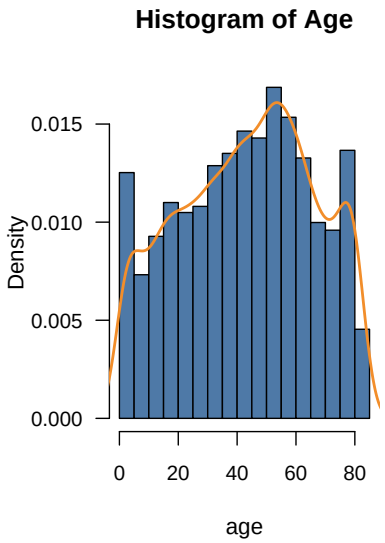
58% of the population are women, however, when analyzing by gender, the percentage of men. However, when analyzed by gender, the percentage of men and women with stroke is 5% each. stroke and the percentage of men with stroke is 5% for each gender. This indicates that it is equally likely to suffer a stroke whether male or female.

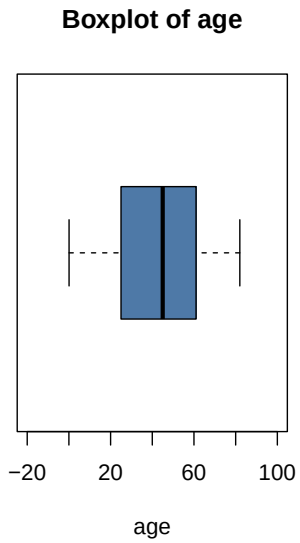
Age

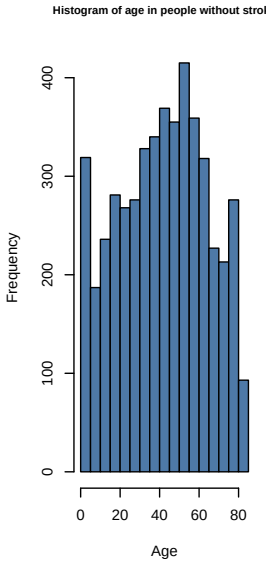
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.08   25.00   45.00   43.23   61.00   82.00
```

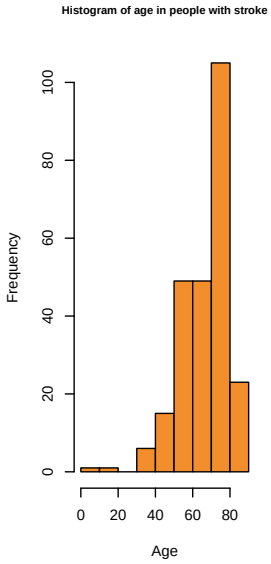
Table 3: Summary Age

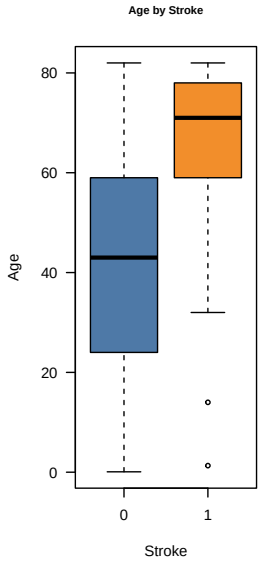
mean	sd
43.22999	22.61358









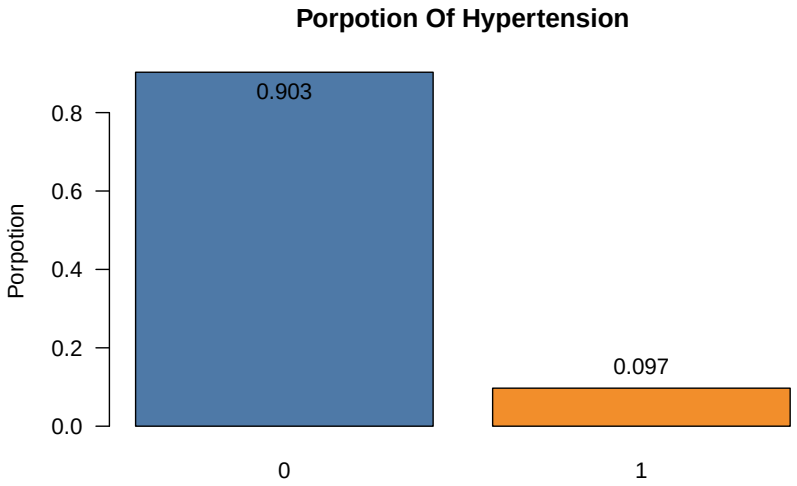


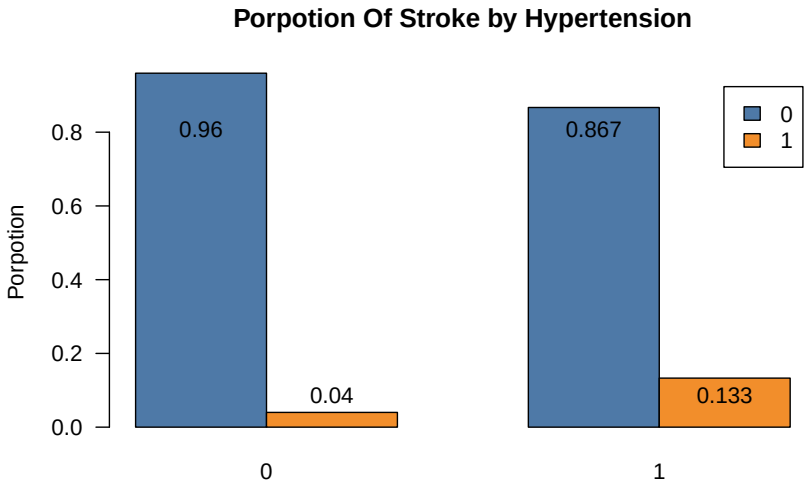

```
## $`0`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.08   24.00   43.00   41.97   59.00   82.00
##
## $`1`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.32   59.00   71.00   67.73   78.00   82.00
```

The average age is 43 years, 75% of the people in the database are equal to or younger than 61 years old When analyzing the ages by stroke, it is identified that the average age of people who have not suffered a stroke is 41 years old, while those who have suffered a stroke are on average 67 years old, i.e. the older the person is, the greater the probability of suffering a stroke.

Hypertension

Hypertension	Porpotion
0	0.903
1	0.097

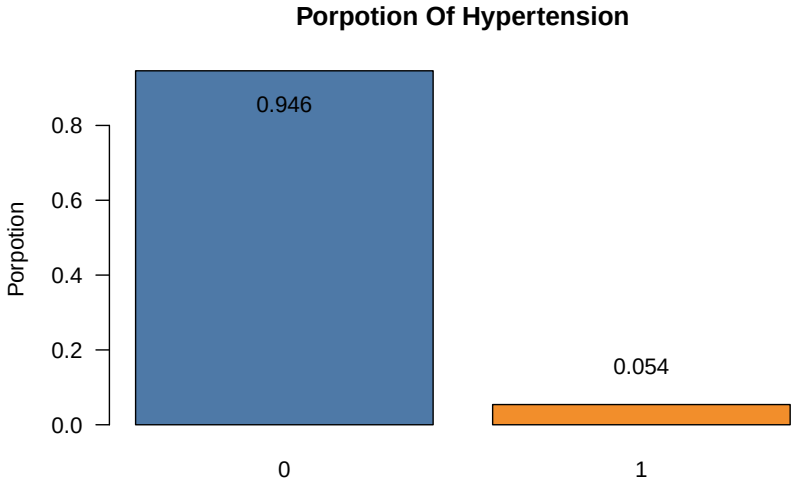


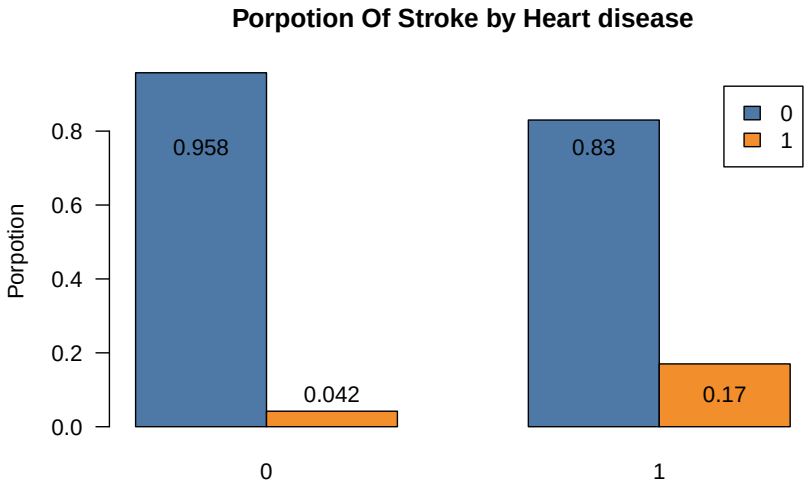


10% of the people in the database had suffered from hypertension. When reviewing the differences by stroke, we found that 4% of the people who do not suffer from hypertension have had a stroke compared to 13% of the people who do have hypertension who have suffered a stroke, this difference in proportions suggests that if a person has hypertension, he/she is more likely to suffer a stroke.

Hearth disease

Hearth disease	Porpotion
0	0.946
1	0.054

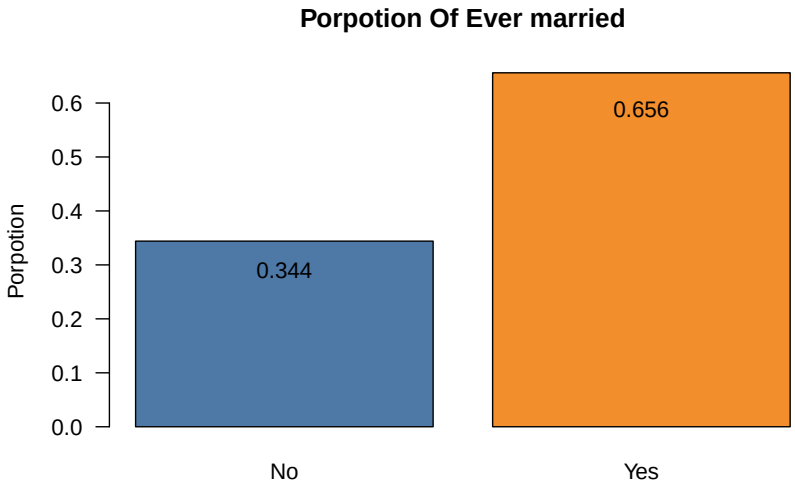


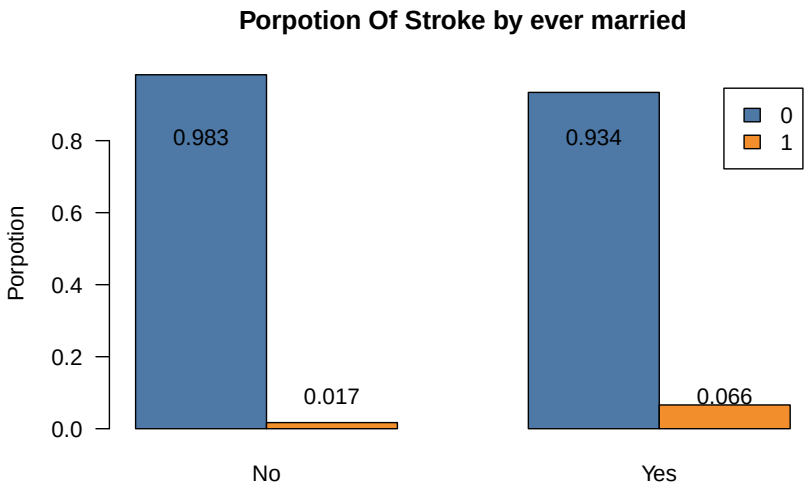


5% of the sample suffers from heart disease. When reviewing the differences by stroke, we see that 17% of the people with heart people with heart disease have had a stroke, compared to 5% of those without heart disease. 5% of those without heart disease. This indicates that there are differences between these two groups. and therefore if a person has heart problems, he or she is more likely to have a stroke

Ever married

Ever married	Porpotion
No	0.344
Yes	0.656

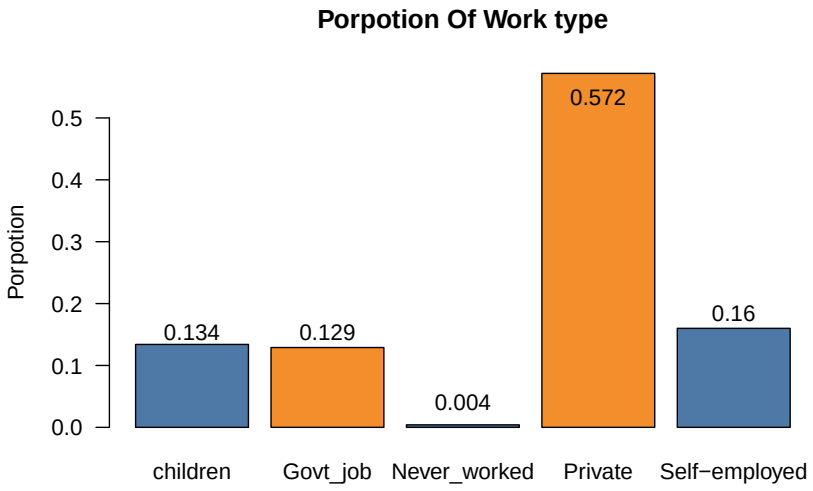


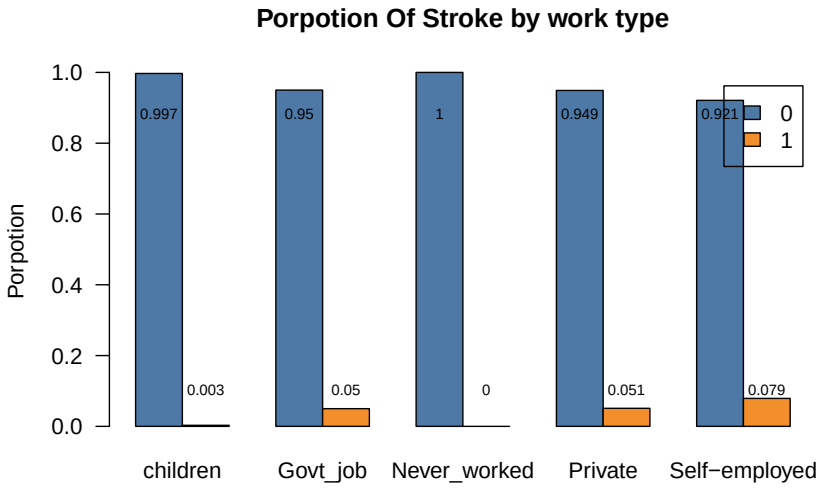


65% of the sample had ever been married. Of those who are married, 6% have suffered a stroke compared to 1% of those who are not married. 1% of those who are not married, although the difference is not significant

Work type

Work type	Porpotion
children	0.134
Govt_job	0.129
Never_worked	0.004
Private	0.572
Self-employed	0.160

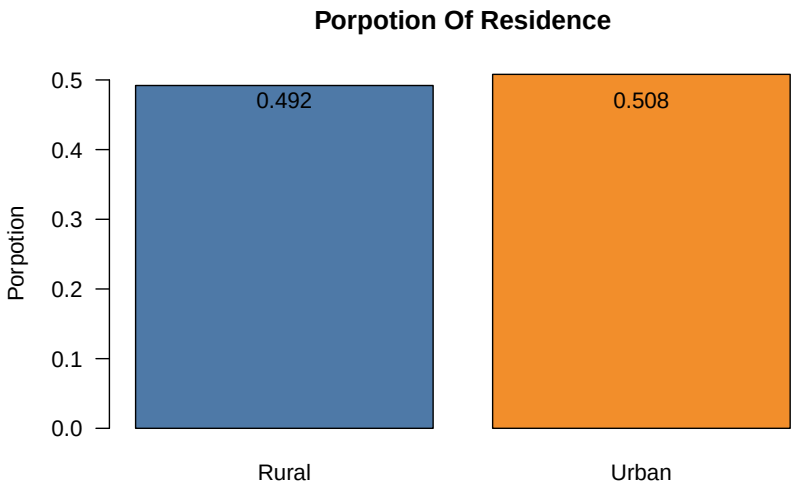


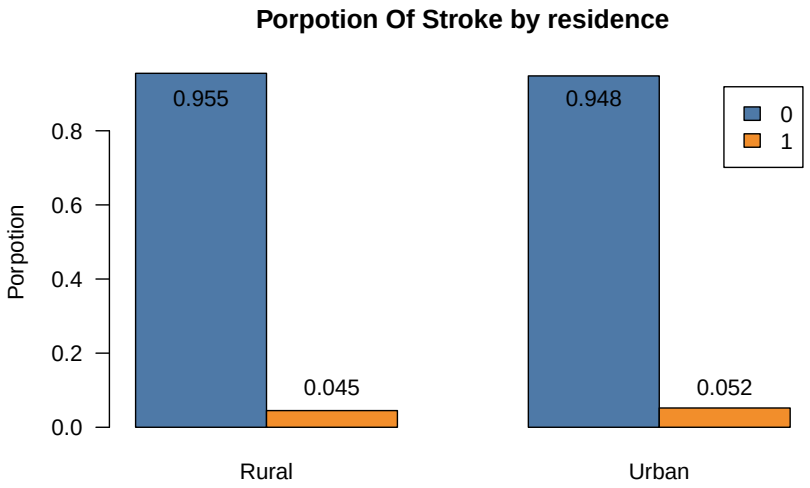


57% of the sample works in a private company. No differences were identified in the percentage of stroke by profession, i.e. the profession does not influence the strike. i.e. the profession does not influence the occurrence of a stroke.

Residence type

Residence	Porpotion
Rural	0.492
Urban	0.508





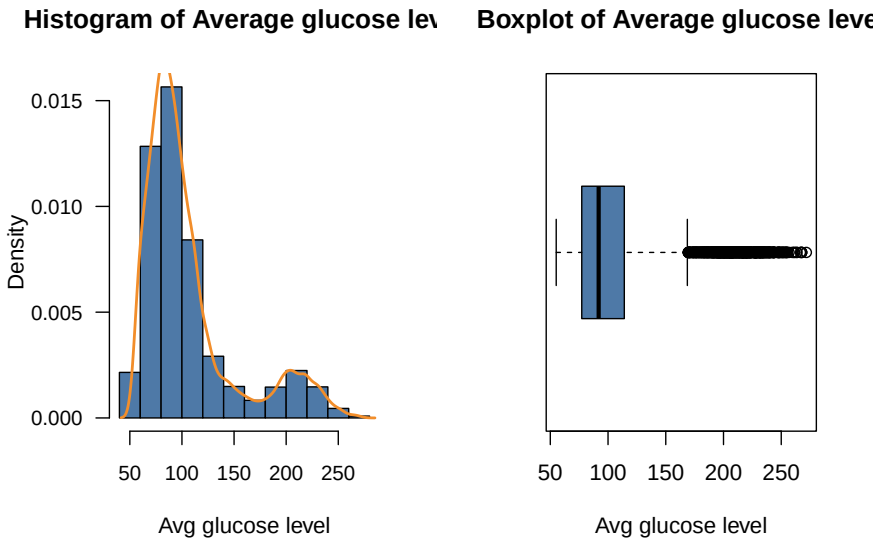
The sample is very equitable with respect to type of residence, 50% live in rural areas and 50% in urban areas. When comparing by stroke, no differences are found, i.e., the fact that a person that the fact that a person has a stroke does not depend on his or her place of residence. place of residence.

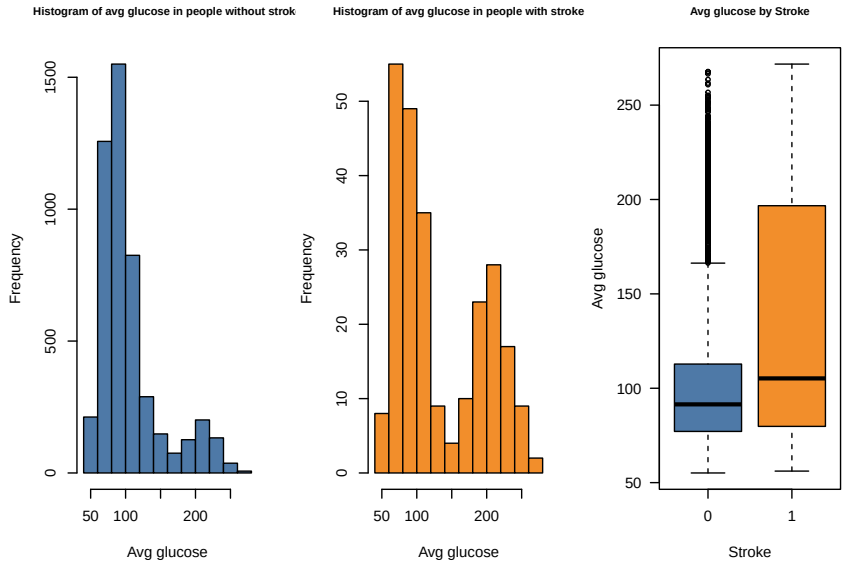
Average glucose level

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   55.12   77.24   91.88  106.14  114.09  271.74
```

Table 9: Summary Glucose

mean	sd
106.1404	45.285






```
## $`0`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   55.12   77.12   91.47  104.79  112.80  267.76
##
## $`1`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   56.11   79.79  105.22  132.54  196.71  271.74
```

On average, the people in the sample had a glucose level of 106. The distribution associated with the average glucose level is positively skewed and bimodal. positively skewed and bimodal. When reviewed by stroke, the distributions retain their bimodality, however in the case of stroke, high glucose levels are more frequent. In addition to this, we found that on average a person with stroke has a avg glucose level of 132 compared to 104 for a person without stroke. This indicates that on average people who have suffered stroke have a higher avg glucose level.

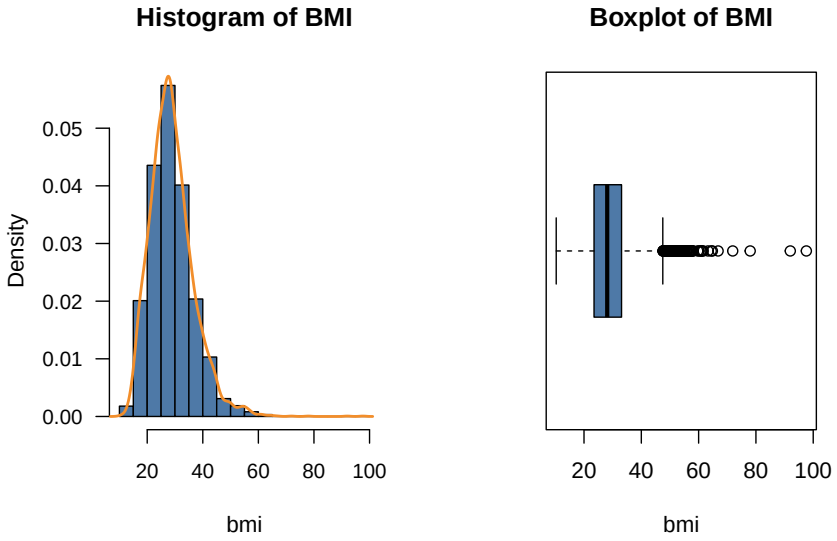
Bmi

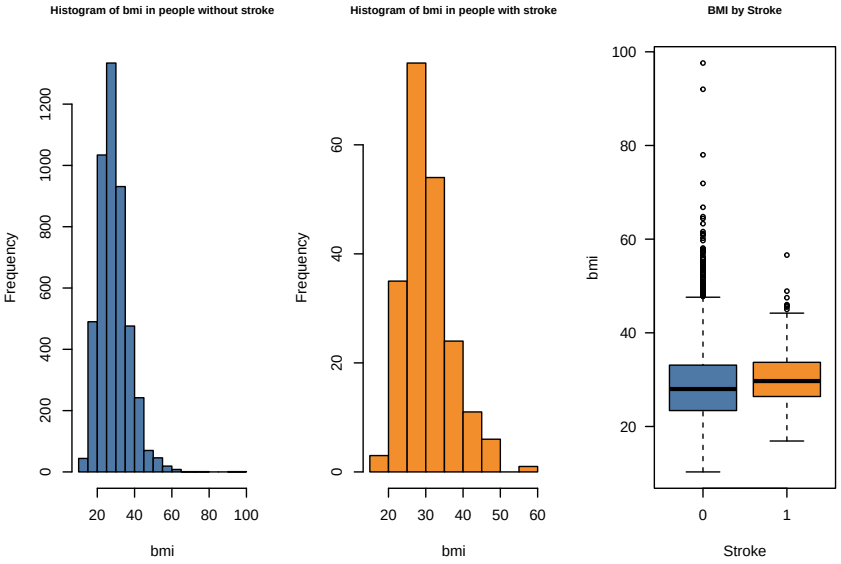
```
## Warning: NAs introducidos por coerción
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   10.30   23.50   28.10   28.89   33.10   97.60     201
```

Table 10: Summary BMI

mean	sd
28.89456	7.85432



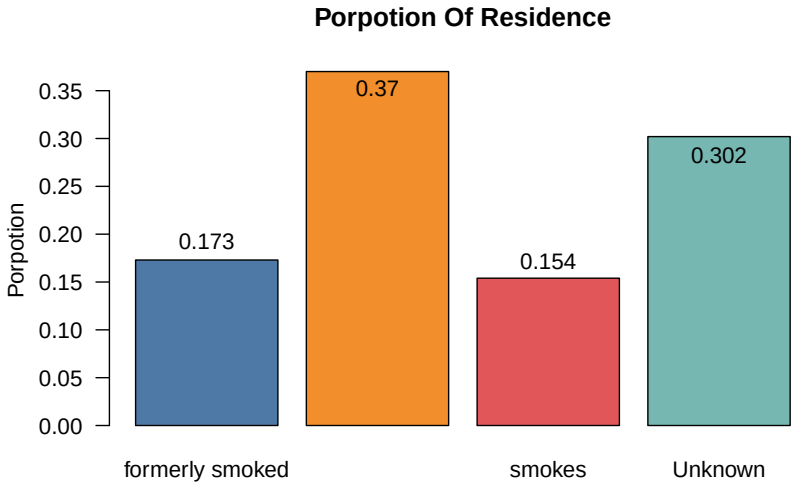


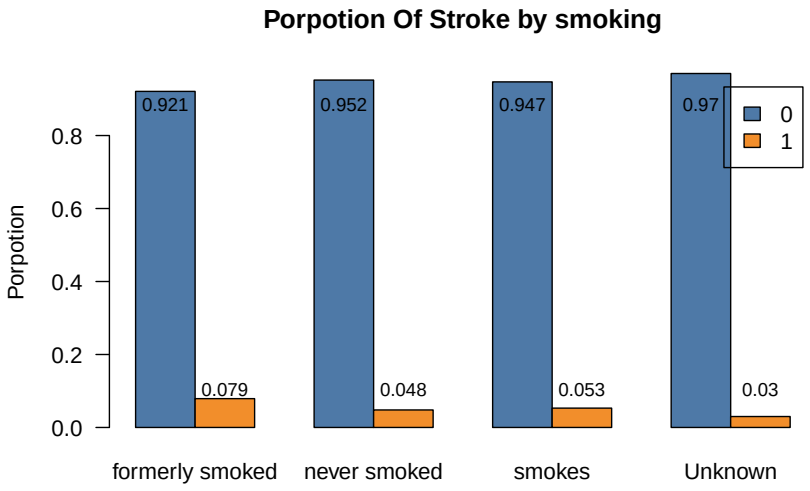

```
## $`0`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   10.30   23.40   28.00   28.82   33.10   97.60     161
##
## $`1`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   16.90   26.40   29.70   30.47   33.70   56.60      40
```

The distribution of bmi is positively skewed, with mean at 28. 75% of the people have a bmi lower than 33.10 and a total of 201 null values. When reviewed by stroke, we find that on average a person with stroke has a bmi of 30.47, higher than the 28.82 average of a person who has not suffered stroke. In other words, the data show that if a person has a higher bmi, he or she is more likely to have a stroke

Smoking status

Smoking	Porpotion
formerly smoked	0.173
never smoked	0.370
smokes	0.154
Unknown	0.302





Seventeen percent of the sample previously smoked, 15% smoke and 37% have never smoked. The proportion of people with stroke in each of the smoking status categories are very similar, i.e., there are no obvious differences between the two categories. The proportions of people with stroke in each of the smoking status categories are very similar, i.e. there are no obvious differences to suggest that smoking increases the risk of strok