EDA results/conclusion

Pascal Visser

2022-10-04

Exploratory data analysis results and conclusion/discussion

The Exploratory data analysis (EDA) was written about a dataset with stroke events with health and lifestyle variables. The EDA digs into the multiple variables and the contribution towards the dataset. Also are the unknowns and not available records evaluated and dealt with.

This paper is about the result of the EDA and the conclusion and discussion about these results.

1. Results

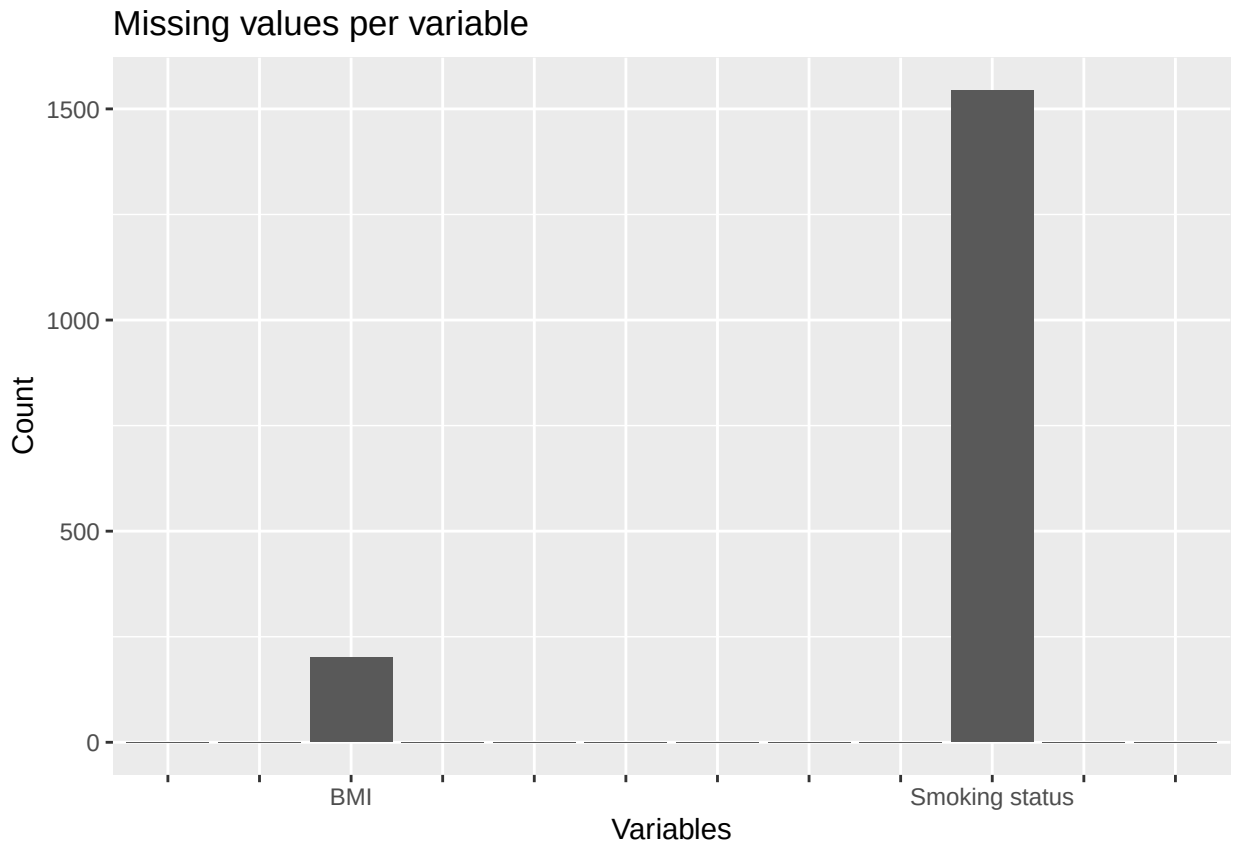
The dataset is a normal set with not to many variables. It can be divided into a classifier, health, lifestyle and id variables. In the EDA all these variables is talked over and evaluated. Most variables are visualized with a graph or table.

1.1 The missing data

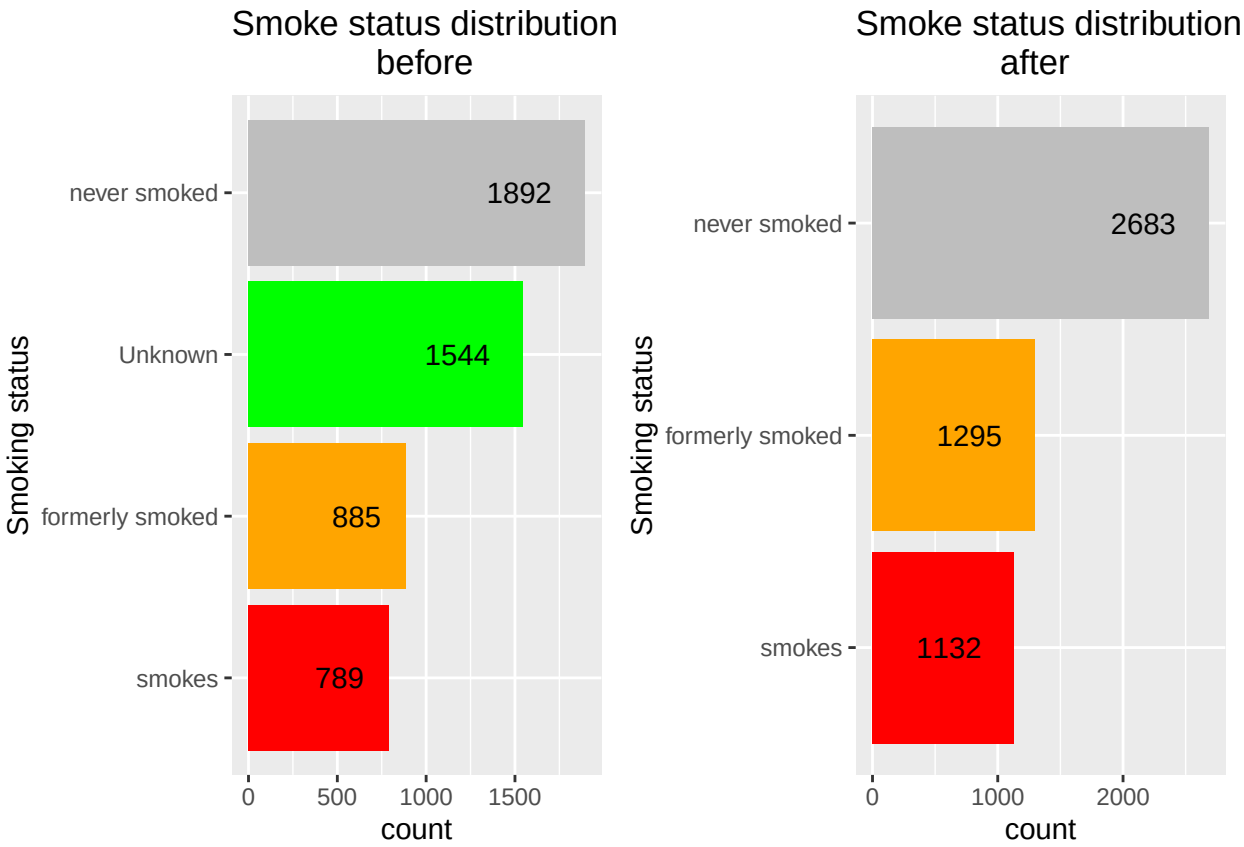
The first look at the data was good, all the variables were clear and organized. Not too many Na's or other vague records. Also there was enough data to work with:

```
## The dataset is 5110 rows long, and has 12 columns
```

The size of the data tells that it is big enough to do something with, without having concerns of getting low accuracy trough shortage of data. By looking at the not available records, BMI and smoking status are the only one with missing values. Where smoking status has the most.



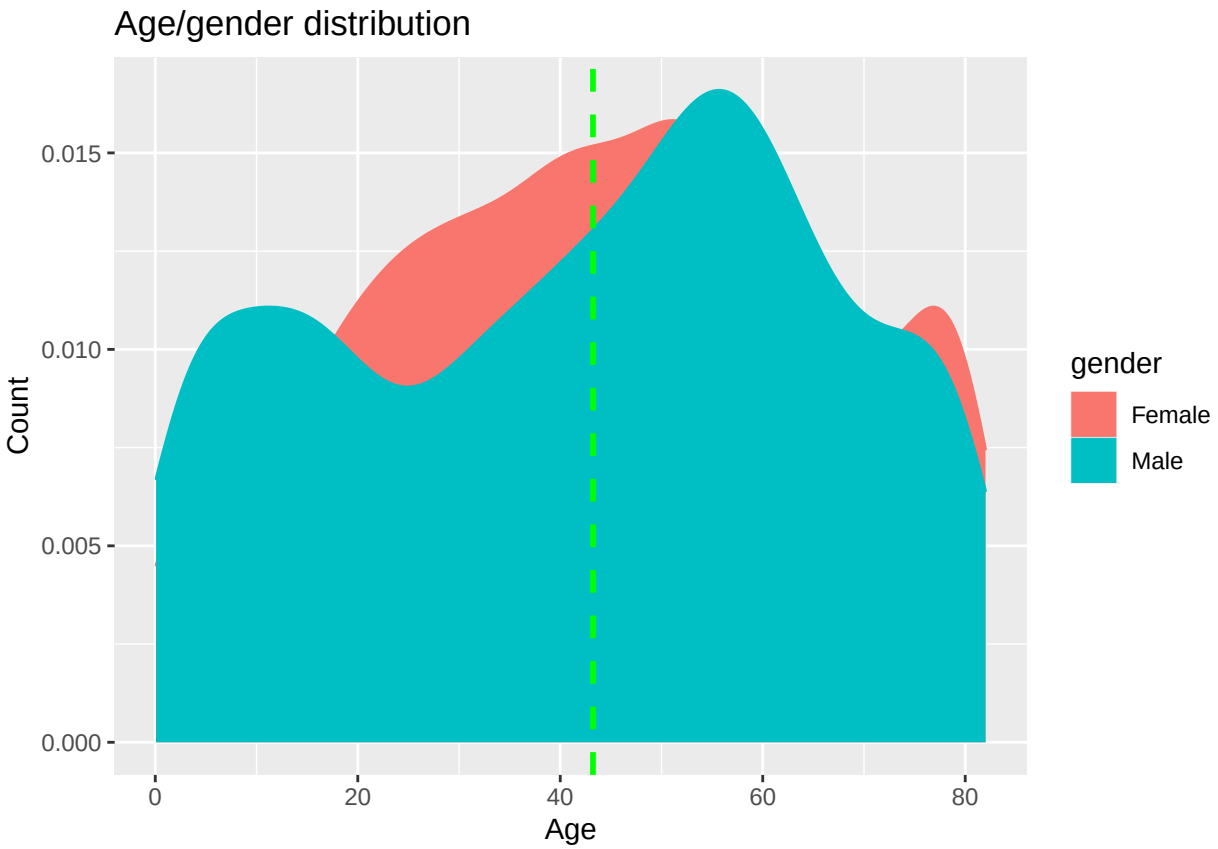
Of the 5110 records in Smoking status, ~1500 records have a unknown status. This unknown status cause problems because is it not a good label, it is incomplete. By calculating the weight of the other labels in the variable, the unknown label is replaced with one of the other three labels based on the occurrence in the variable. Which looks like this:



Now the unknown is evenly replaced over the dataset without removing records. This was the biggest concern with the missing values. For BMI the Na's are simply replace with the mean of the column. But the size and impact of the BMI missing values is a lot smaller.

1.2 Test subjects distribution

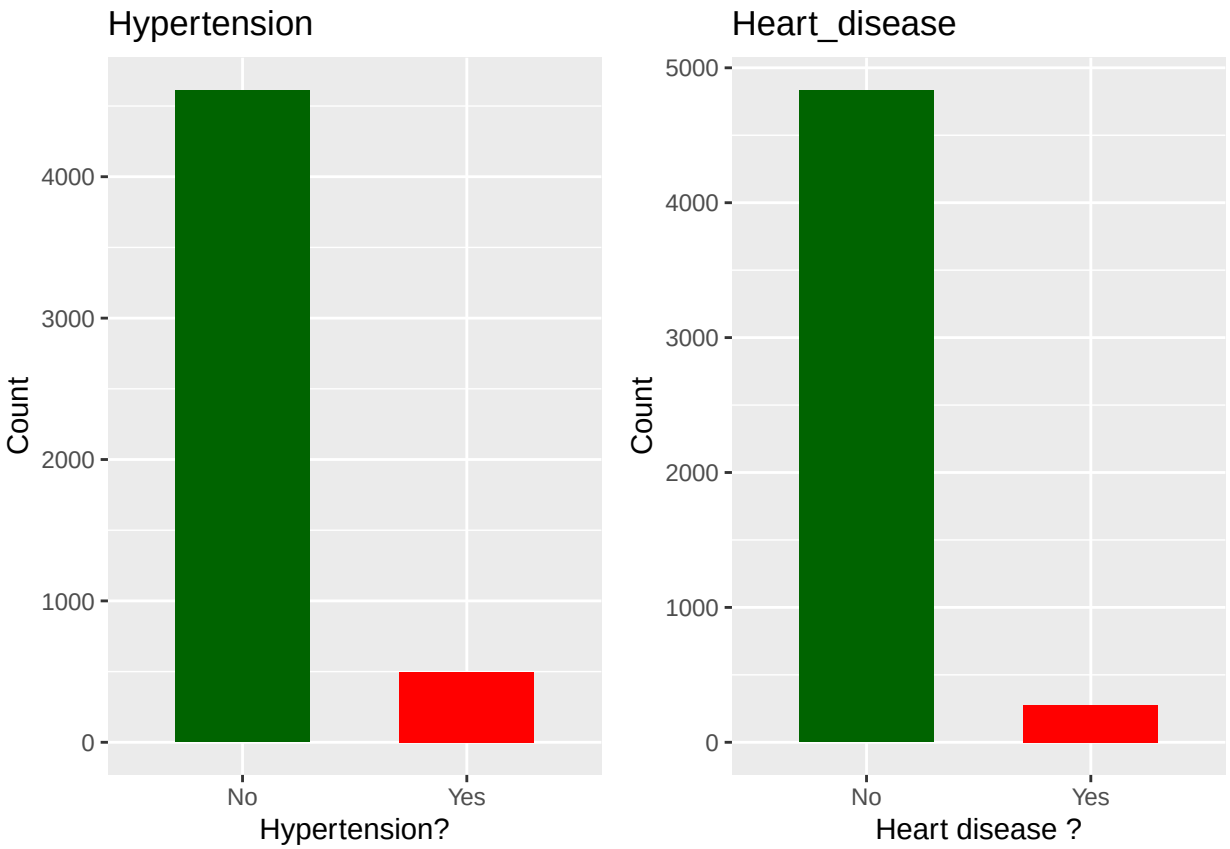
The research is about stroke events by people, so you want a evenly distributed group of gender and ages. Ages is the most important, because with many elderly people the chance of stroke and/or cardiovascular diseases is increased. This increase in cardiovascular diseases can raise the stroke events unwanted. The plot below shows the distrubution of the age and genders:



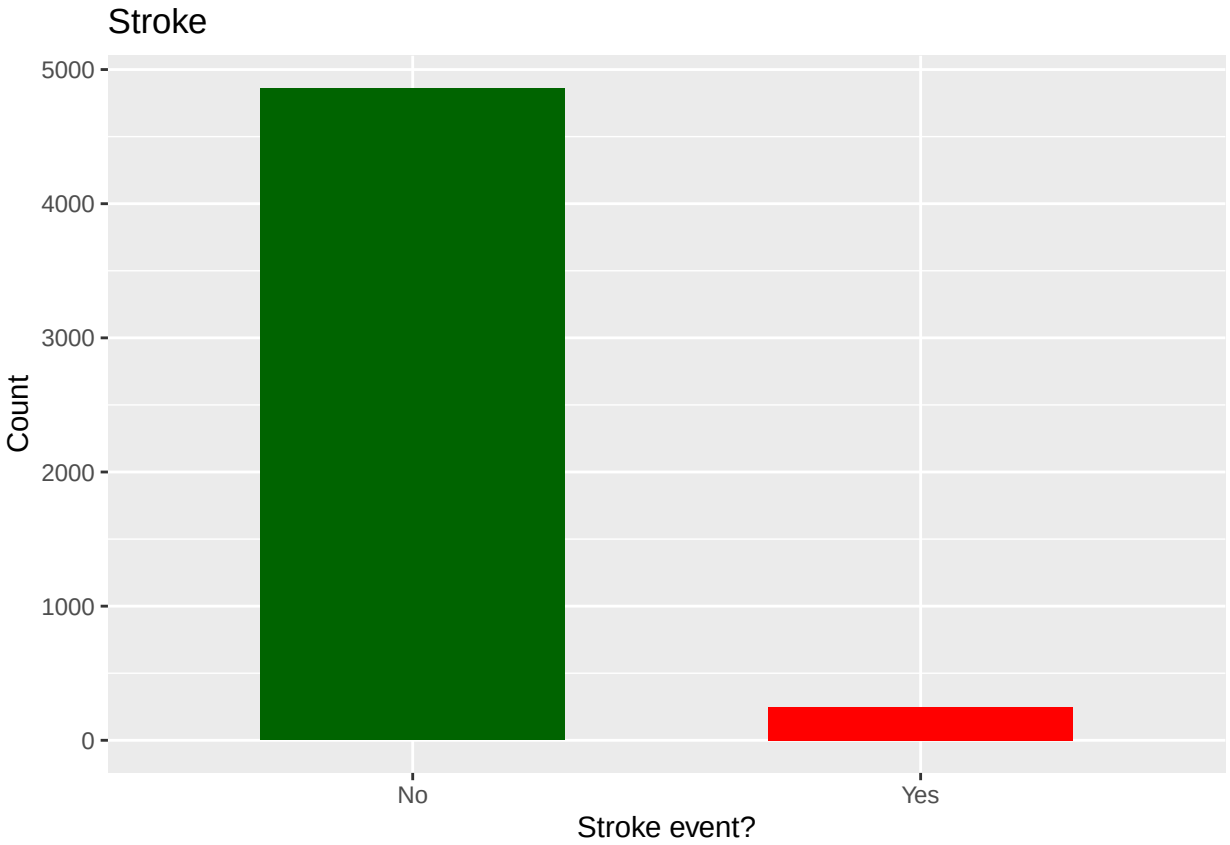
Looking at the graph, The mean is nicely settled in the middle, meaning that we have a diverse group of all ages. Also the gender is fairly spread around the age groups. Only the female group is overrepresented in the ages of 20 - 40.

1.3 Health variables

viewing the health variables, Hypertension and heart disease, the imbalance imminently draws attention. The records with 'No' are heavily represented inside the variable. This causes a huge skewed column of data.



This is a point that need to be discussed later on. Looking at the classifier, the same problem occurs:



Even the classifier is fairly skewed, The huge imbalance will cause the machine learning algorithms to lean toward the 'no' because of the overrepresentation inside the variable. Going with the 'no' gives a 95%+ accuracy. It is the biggest discussion point of the data.

2. Conclusion/Discussion

The data is in a good condition with several variables to work with for machine learning. The removal of the Na's made the quality of the data better, mainly trough the conversion of the term 'unknown' by smoking status. Also the casting of types helps al lot, by creating a simple but challenging dataset, But the data is still having one big flaw.

The plots shows that the distribution of certain variables skewed is, namely: stroke, hypertension and heart_disease. These variable have a very big imbalance inside. This imbalance can cause problems with machine learning because saying 'No' at these variables can give you a 95%+ accuracy. This imbalance need to be tackled before moving on into machine learning. By doing this, the dataset can be workable for machine learning.