CA-1

Enterprise Database Technologies

Eddy Ekofo - X00119959

2018

# Knowing the Data

I have figured out that the data is nominal which means attributes of a variable are differentiated only by name (category) and there is no order (rank or position). The age bracket of customers is from 18 years old to 88. This means that the data is only of adults, this is what expected because of the type of dataset I am working with. Although there are missing data, but it is insignificantly small to have great impact on the overall accuracy of the dataset.

The distribution of the majority numeric values seems to follow a binomial distribution. Here is an example of age histogram:

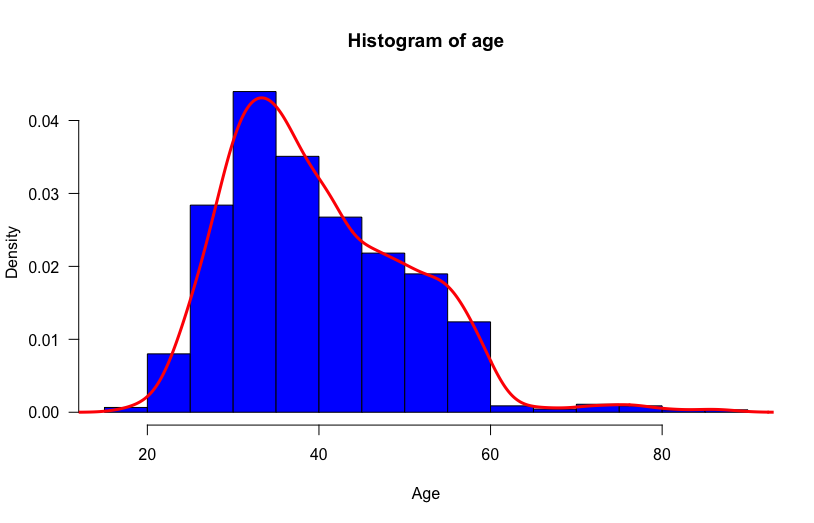
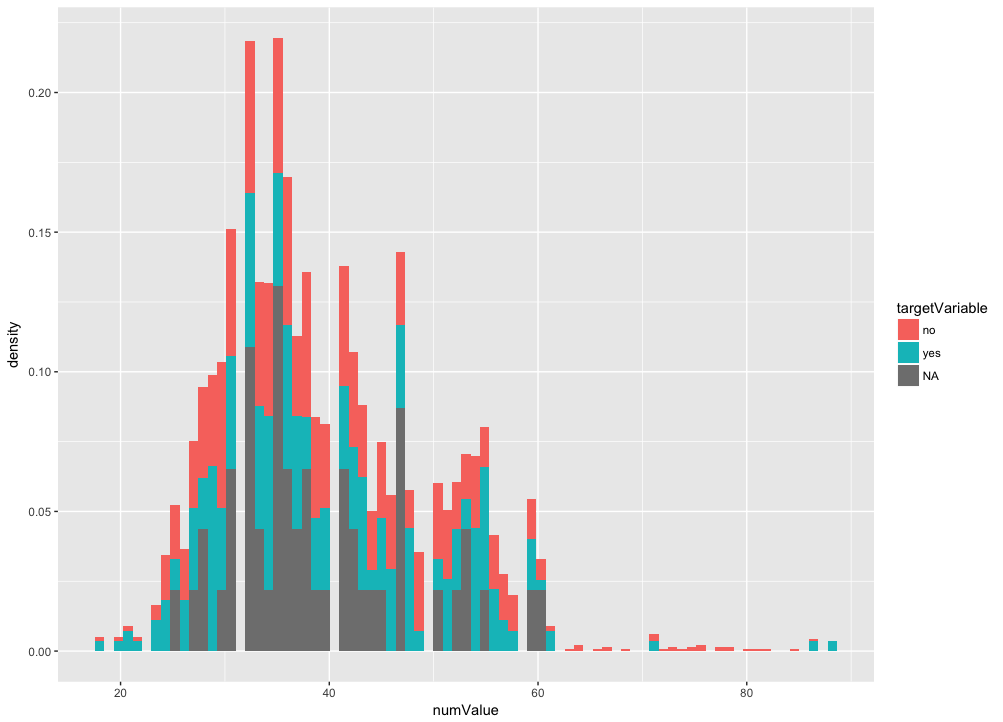


Figure 1: Data not normal

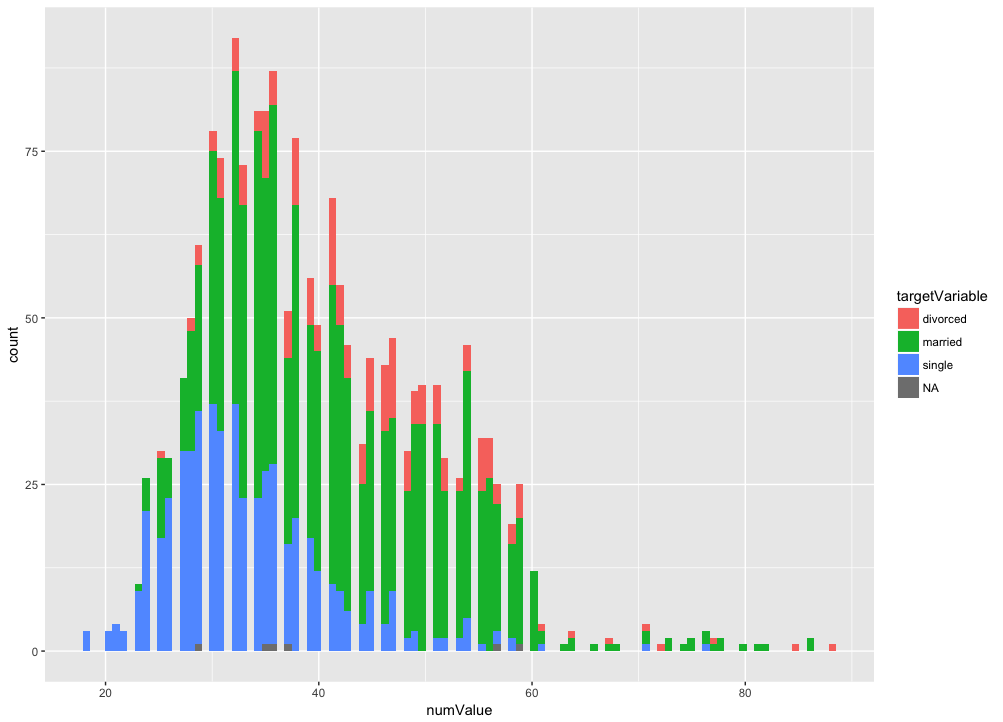
There are missing data in some of the attributes, thus data must be cleaned before any progress.

## 2

The relationship between age of the customer and if they have a loan or not is relative. The number of those with younger age. The histogram shows the number of those who have a lower age have significantly low percentage of loans compared to those who are in their 20s up to early 60s.

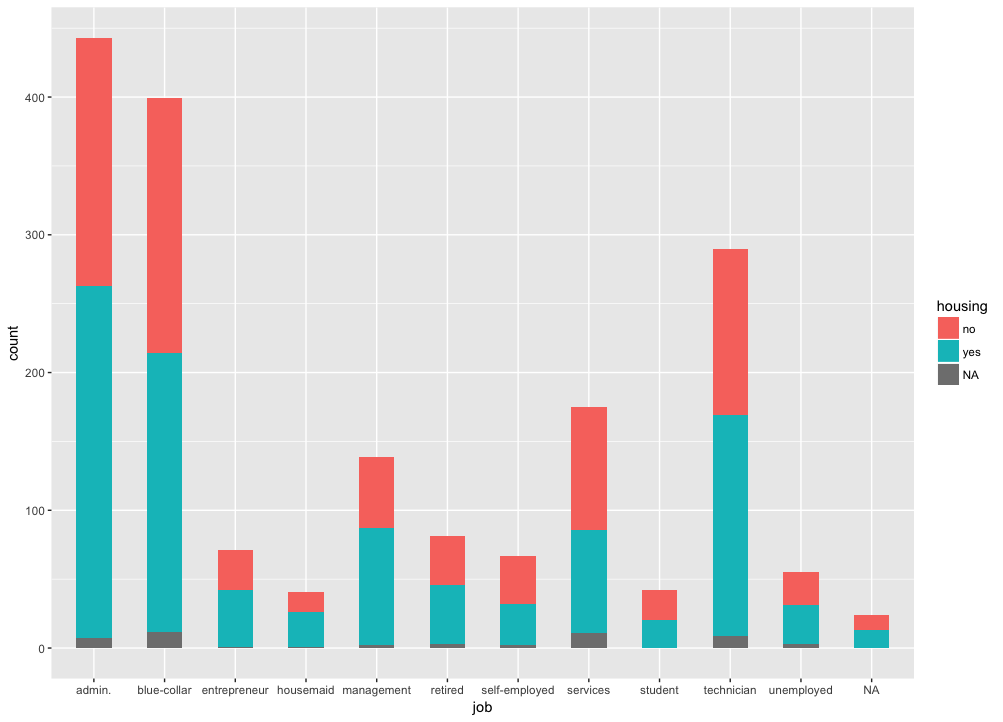


I think the age and marital graph is very interesting histogram. A classification model attempts to draw some conclusion from observed values, thus, the values shown in the histogram indicates the age in which a customer is to be most likely married. It is seeming, most customers before the age of 25 are all single. This would ideal for the classification model in a supervised machine learning algorithm approach.



## 1.3

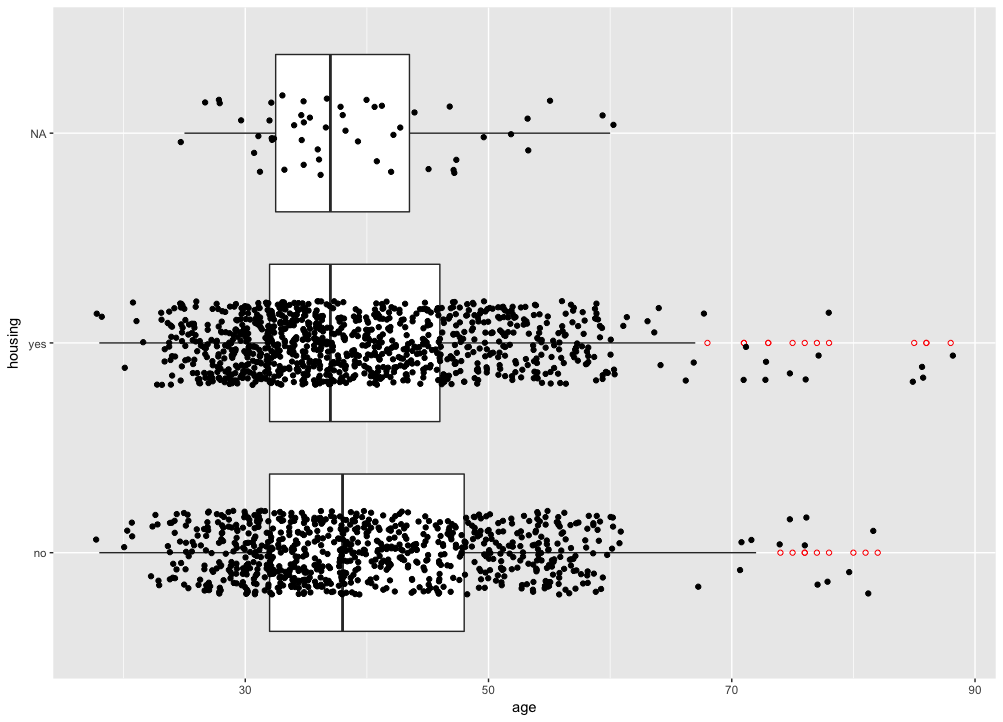
The job as predictor variable and housing (has housing loan?) as target variable. As the bar-chart below indicates, there are a certain job in which the customers are most likely to have received a loan. Those with jobs of an Admin, Blue-collar and Technician most likely will have a house loan.



Variables which would be significant for a Machine Learning classification model would be *student* and *admin*. The student is likely to be an outlier in an unsupervised machine learning approach if seeking which student has a house loan. In the other hand the variable admin has the most customers with house loans. This would be used in a linear regression classification model to test if the user has a house loan or not e.g.

p(p) = 0.01 \* (job) + 0.04 \* (housing loan or not) + 0.03 (married or not) — 0.4

## 1.4.

As the boxplot indicates there are outliers (are extreme values) in the data. The extreme data are indicated in the red. 

Z-scores are a way to compare results from a test to a “normal” data or how many standard deviations away from the mean. It is a very effective method if you can describe the values in the feature space with a gaussian distribution (Parametric). Knowing a person’s age is good but might not be so when compared to the mean age. Here’s a function which calculates the Z-score:

1. z\_test < - function(x, tailed) {
2. z < -(x - mean(x)) / sd(x) # formula: z = (x – μ) / (σ / √n)
3. if (tailed == 1) return (cat('Z-score - ', z, '\np-value - ', pnorm(-abs(z)))) if (tailed == 2) return (cat('Z-score - ', z, '\np-value - ', 2 \* pnorm(-abs(z)))) if (tailed != 1 | tailed != 2) return ('can omly be one or two tailed')
4. }
5. z\_test(age, 1)

The cons of using Z-score is that the data has to be normally distributed, it wouldn’t be ideal to use with age since the data is skewed.

The IQR is another statistical method which I used to determine outliers. First, I summarised the numerical variables. From that I used the IQR built in method from R. The merits of using this method is, if the important advantage of interquartile range is that it can be used as a measure of variability if the extreme values are not being recorded exactly (as in case of open-ended class intervals in the frequency distribution).

1. summary(age)
2. IQR(age)

The demerits of IQR method is doesn’t use all the data (ignores half of the data-points, not just the outliers). Tails almost always matter in data and these aren’t included. Outliers can also sometimes matter and again these aren’t included.

## 1.5

From investigating the correlation of each numeric pairs graphically, there seems to be no correlation between any variable. I then used R’s built in correlation method (cor()) which uses the *Pearson’s Correlation Coefficient* as default. From my finding there is no correlation between the variables. Then I also used the *Spearman* and *Kendall* statistical methods, I still found no correlation. Here is the sample code which I used.

1. pairs.panels(numVals, pch = 20, main = "Banking Data by age")# select columns 1 – 4
2. plot(age, duration)# the default correlation / numVals is all the numeric values from my dataset
3. cor(numVals)
4. cor(numVals, method = "spearman")
5. cor(numVals, method = "kendall")

Here is a table using the Kendal correlation method to find out the correlation of numeric values (Rounded by 2):

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | age | duration | campaign | pdays | previous | cons.price.idx | cons.conf.idx | euribor3m | nr.employed |
| age | 1 | 0.03 | 0 | -0.01 | 0.03 | 0.01 | 0.05 | 0.02 | 0.01 |
| duration | 0.03 | 1 | -0.08 | -0.08 | 0.04 | 0.03 | 0 | -0.04 | -0.06 |
| campaign | 0 | -0.08 | 1 | 0.06 | -0.08 | 0.09 | 0 | 0.12 | 0.14 |
| pdays | -0.01 | -0.08 | 0.06 | 1 | -0.5 | 0.02 | -0.04 | 0.23 | 0.25 |
| previous | 0.03 | 0.04 | -0.08 | -0.5 | 1 | -0.2 | -0.14 | -0.39 | -0.42 |
| cons.price.idx | 0.01 | 0.03 | 0.09 | 0.02 | -0.2 | 1 | 0.19 | 0.31 | 0.32 |
| cons.conf.idx | 0.05 | 0 | 0 | -0.04 | -0.14 | 0.19 | 1 | 0.23 | 0.16 |
| euribor3m | 0.02 | -0.04 | 0.12 | 0.23 | -0.39 | 0.31 | 0.23 | 1 | 0.79 |
| nr.employed | 0.01 | -0.06 | 0.14 | 0.25 | -0.42 | 0.32 | 0.16 | 0.79 | 1 |

Here is the table of the correlation between each numerical values on the dataset. Here is the code to generate that:

1. corResult < -cor(numVals, method = "kendall")
2. corResult < -round(corResult, 2)
3. write.csv(corResult, "output.csv")

The strongest correlation I have found is nr.emploed (number of employees - quarterly indicator) & euribor3m (euribor 3 month rate - daily indicator) with a **0.78915804** strong uphill (positive) linear relationship.

The weakest, I have a few. Many numerical attributes have **0** (no linear relationship) e.g. Age & campaign.

## Cleaning and Transforming the Dataset – You may use R or Weka for this part.