**Bank Data Logistics Regression**

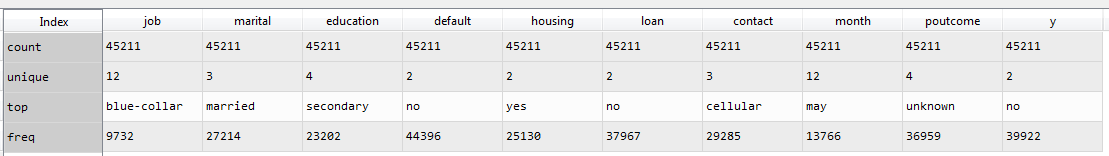
Problem Statement:

Output variable -> y

y -> Whether the client has subscribed a term deposit or not

Binomial ("yes" or "no")

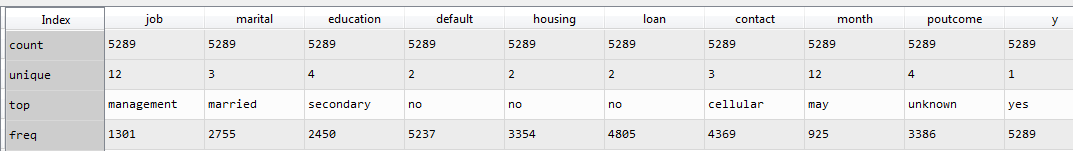
**Descriptive Statistics of the categorical variables is as follows:**



According to the above exploration we can find the following information:

* Most of the people are having blue-collared jobs, are married and have higher secondary education

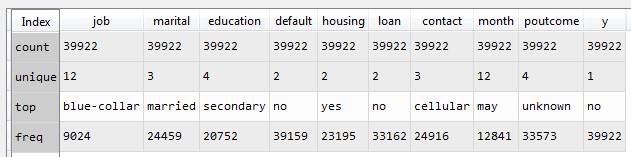
**Descriptive Stats of the categorical variables and the ones who have subscribed to the term deposit:**



According to the above exploration we can find the following information:

* Most of subscribers of term deposit are in management positions, are married and have higher secondary education

**Descriptive Stats of the categorical variables and the ones who have not subscribed to the term deposit:**



The above descriptive stats can also be considered as a sample data and the properties of this sample data is representative of the populations since majority of the people have not subscribed to the term deposit, have blue-collar jobs and have secondary jobs.

The following is the analysis of the categorical variables:

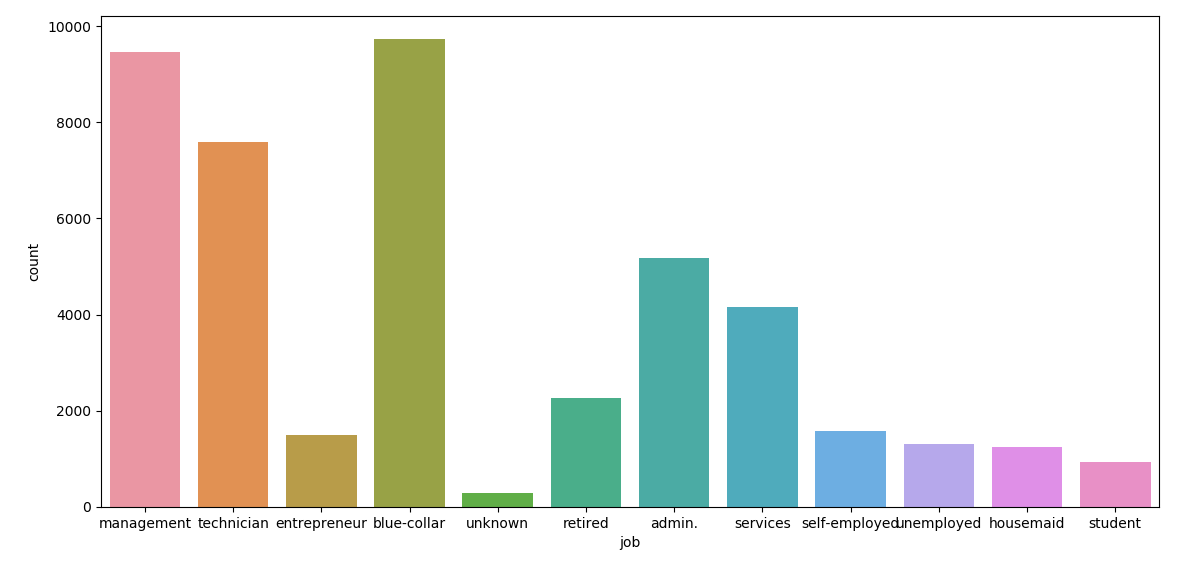
Variable **Job:**

**Unique elements:**

array(['management', 'technician', 'entrepreneur', 'blue-collar',

'unknown', 'retired', 'admin.', 'services', 'self-employed',

'unemployed', 'housemaid', 'student'], dtype=object)`

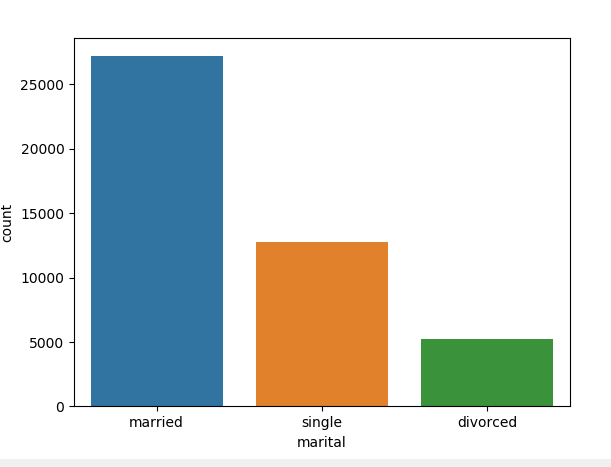


The maximum amount of jobs is from blue-collar and second highest is from the management.

**Marital:**

**Unique elements:**

array(['married', 'single', 'divorced'], dtype=object)

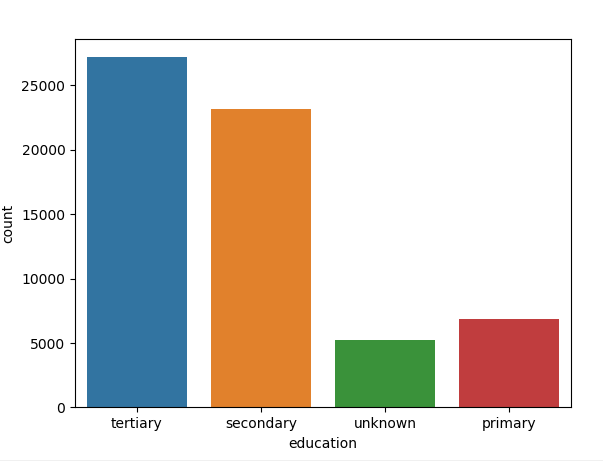


There are maximum number of married people.

**Education:**

**Unique elements:**

array(['tertiary', 'secondary', 'unknown', 'primary'], dtype=object)

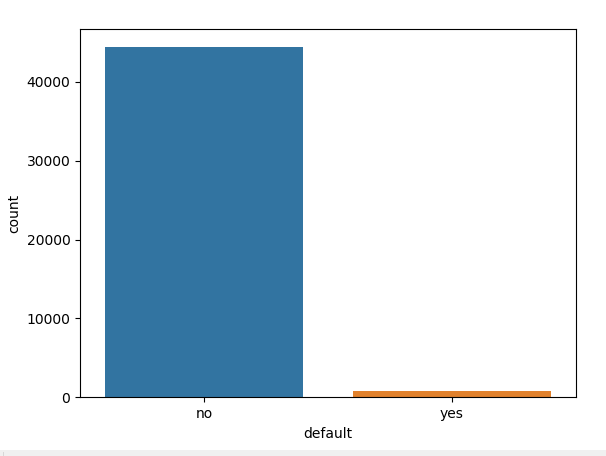


There are maximum number of tertiary educated people but most of the subscribers are from secondary educated people.

**Default:**

**Unique elements:**

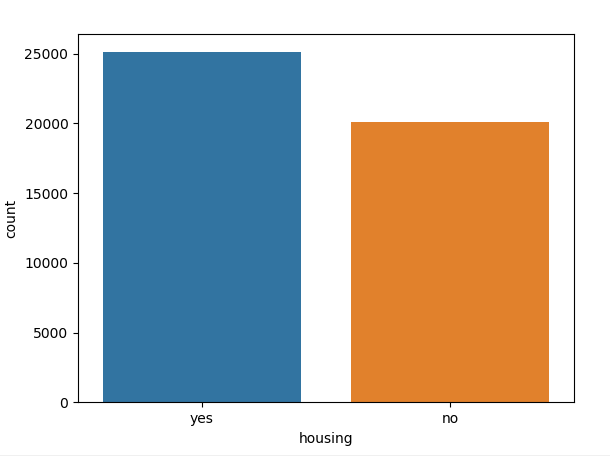
array(['no', 'yes'], dtype=object)



**Housing:**

**Unique elements:**

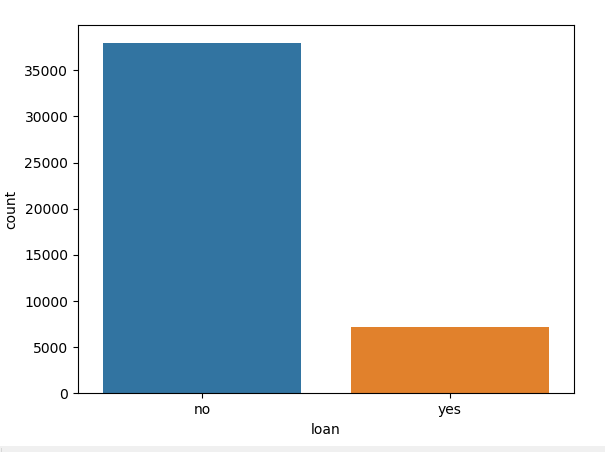
array(['yes', 'no'], dtype=object)



**Loan:**

**Unique elements:**

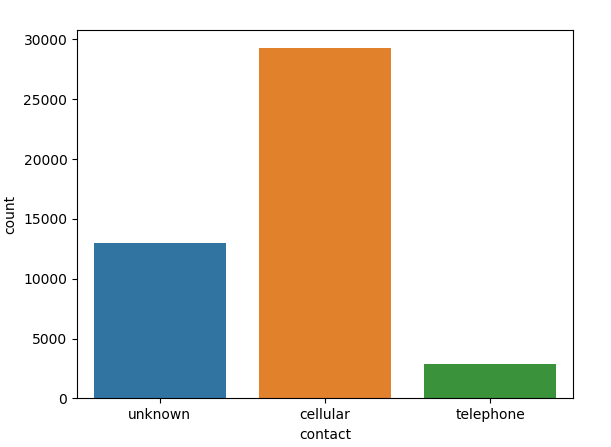
array(['no', 'yes'], dtype=object)



**Contact:**

**Unique elements:**

array(['unknown', 'cellular', 'telephone'], dtype=object)

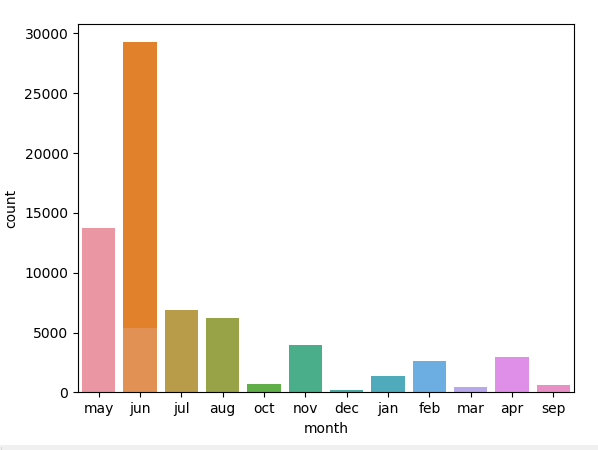


**Month:**

**Unique elements:**

array(['may', 'jun', 'jul', 'aug', 'oct', 'nov', 'dec', 'jan', 'feb',

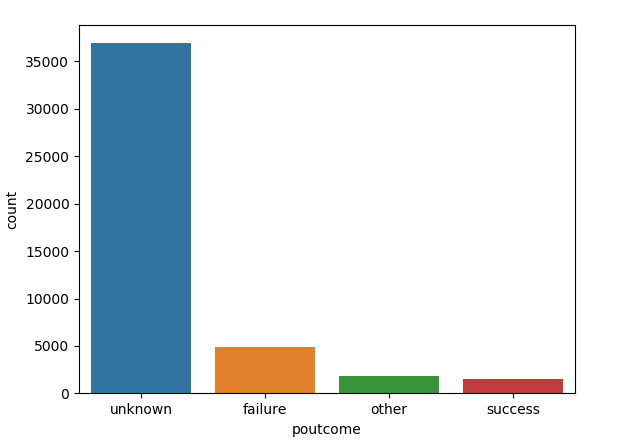
'mar', 'apr', 'sep'], dtype=object)



**Poutcome:**

**Unique elements:**

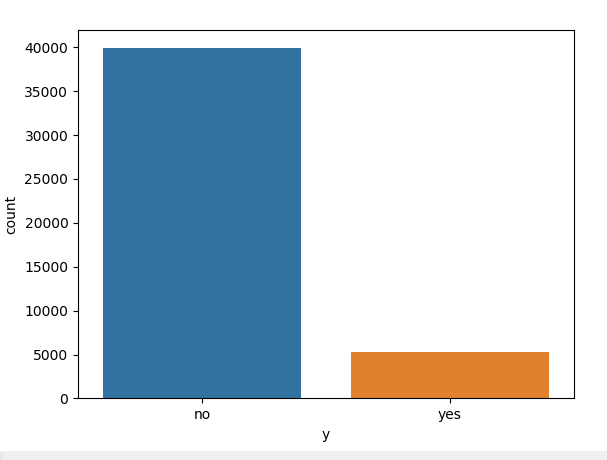
array(['unknown', 'failure', 'other', 'success'], dtype=object)



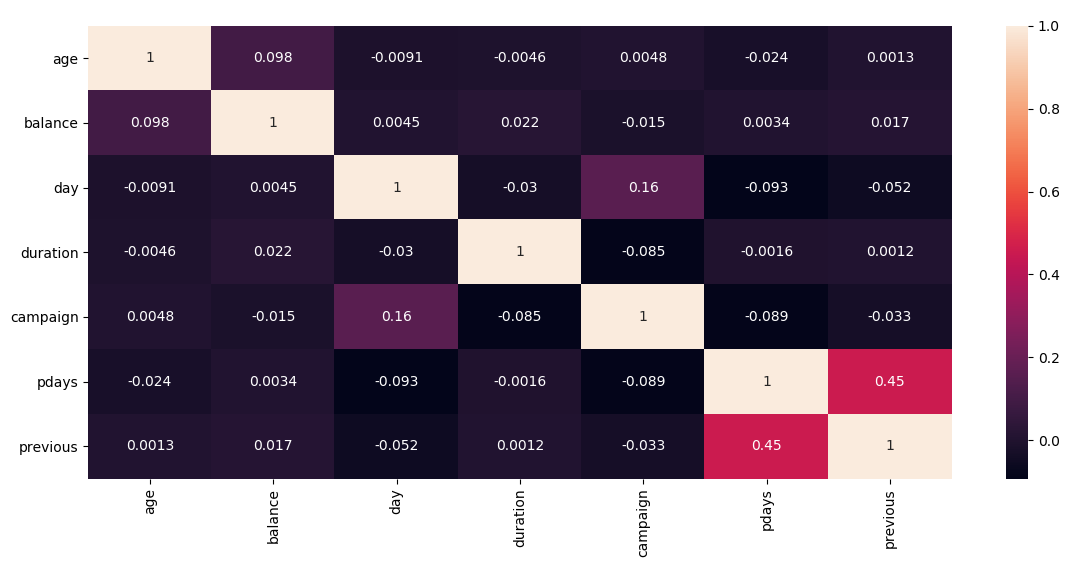
**y:**

**Unique elements:**

array(['no', 'yes'], dtype=object)

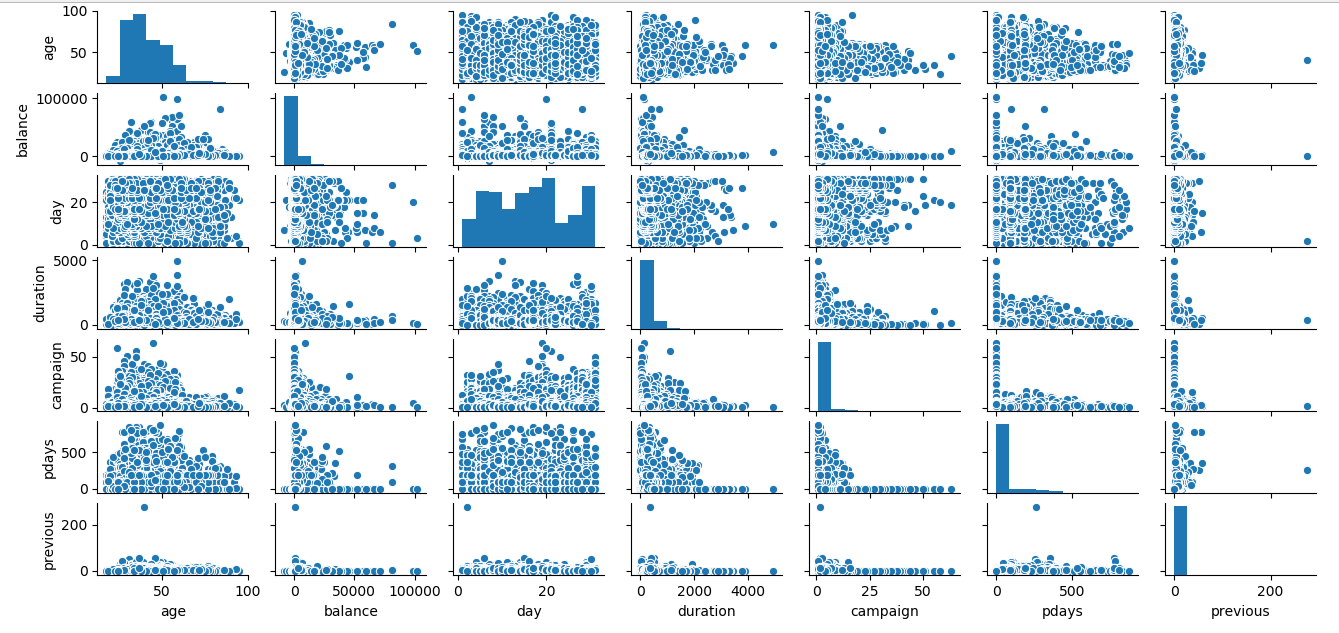


The following is the correlation of the continuous variables:



According to the above heat map, the correlation between the variables is negligible

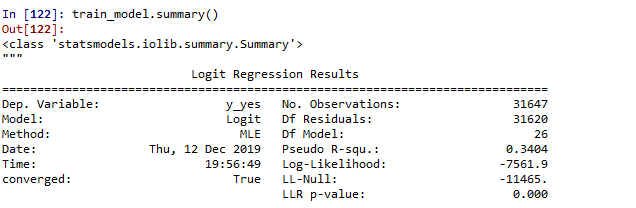
The following is the pair plot of the continuous variables:

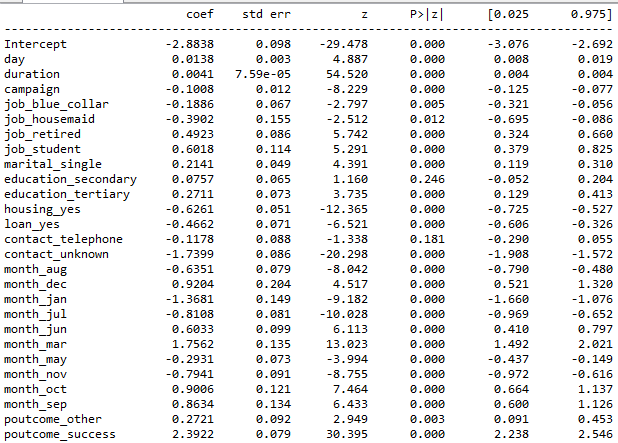


From the above pair plot, the following are the findings:

1. The histogram of age is almost normal
2. The scatterplots between the variables indicate that there is low or no correlation relationship between the variables

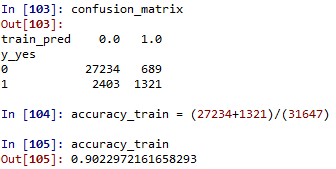
We split the data in train and test, the following is the result of the logistic regression





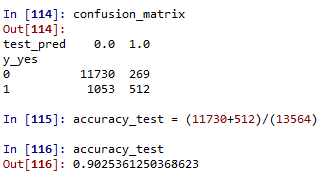
After removing the variables that are insignificant to the outcome. We split the data into test and train

**The confusion matrix of the train data is as follows:**



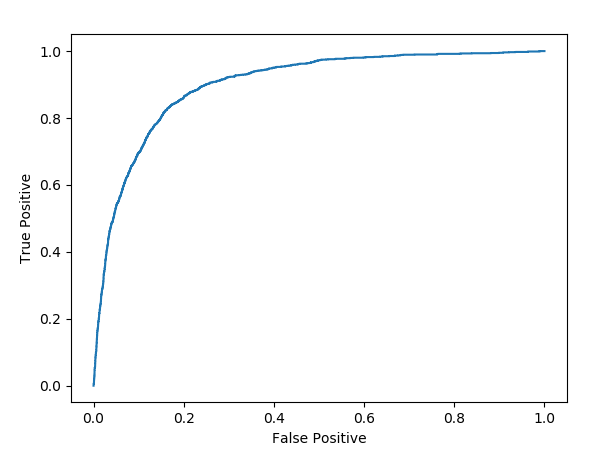
Train data provides an accuracy of 90.2%

**The confusion matrix of the test data is as follows:**



Test data provides an accuracy of 90.25%

The following is the ROC curve



The area under the curve is about 0.90068