WEEK 9 : DELIVERABLES

BANK MARKETING CAMPAIGN

‘DATA SCIENCE’

GROUP NAME: DATA SCIENCE MASTER
NAME : ABHIMANYU GANGANI
EMAIL : Agangani97@gmail.com
COUNTRY : UNITED KINGDOM
COLLEGE : ANGLIA RUSKIN UNIVERSITY
SPECIALIZATION : DATA SCIENCE

PROBLEM DESCRIPTION :

ABC Bank wants to sell it's term deposit product to customers and before launching the product they want to develop a model which help them in understanding whether a particular customer will buy their product or not.
To achieve this task they have consulted an analytics consultancy to automate the process of classification.
The Analytics company have to come up with an ML model to shortlist the customers whose chances to buy the product is higher. This will lead marketing team to target on the given lead.

BUSINESS UNDERSTANDING :

There's been a revenue decline for the ABC bank and to overcome that they want to come up with the actions needed to be taken. With analysis they came to know that customers are not depositing as frequently as before. Banks make investments from the investment made by customers to make high profits.

Banks also urges customers to buy other products such as insurance and Different kind of deposits. They want to check the customers from existing data they pursue and filter the customers having higher chances of buying any new schemes or products from the bank.

DATA CLEANSING AND TRANSFORMATION :

In this section we will use different techniques to get the final dataset that will be used for final model development and evaluation. 'Bank-additional-full' dataset will be considered for doing data cleansing.

There are various issues in the dataset as given below :

- Duplicate Entries.
- Outlier Detection – using histogram and boxplots.
- Missing value detection – Bar graph of categorical columns.

DUPLICATE OBSERVATION :

We have checked all four datasets for the duplicates and only found 12 duplicates in the bank_additional_full dataset. Rest 3 datasets consists of no duplicate values.

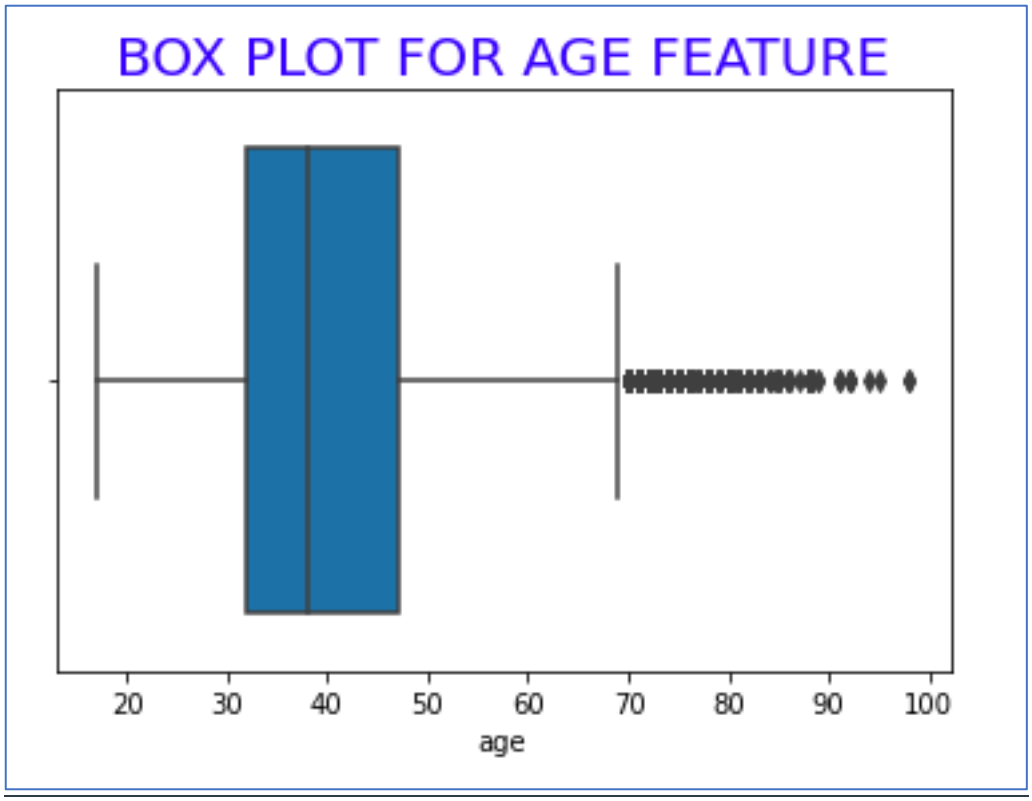
We have also dropped the duplicate values for the bank_additional_full dataset.

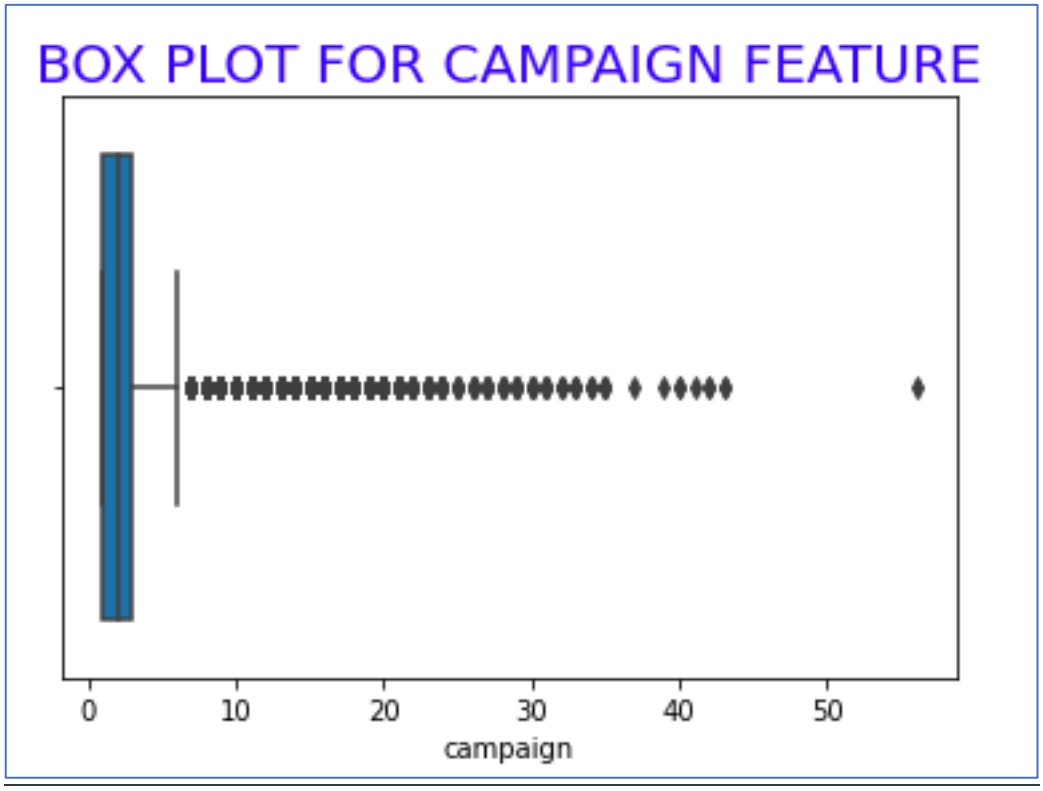
```
In [15]: #Checking the count of duplicates in bank_add_full dataset
         print(f'There are {bank_add_full.duplicated().sum()} duplicates in bank_addition_full.')
         bank_add_full.drop_duplicates(inplace=True, keep= 'first')

         There are 12 duplicates in bank_addition_full.

In [16]: #Checking the count of duplicates in bank_add_full dataset
         print(f'There are {bank_full.duplicated().sum()} duplicates in bank_addition_full.')

         There are 0 duplicates in bank_addition_full.

In [17]: #Checking the count of duplicates in bank_add_full dataset
         print(f'There are {bank.duplicated().sum()} duplicates in bank_addition_full.')

         print(f'There are {bank_add.duplicated().sum()} duplicates in bank_addition_full.')

         There are 0 duplicates in bank_addition_full.
         There are 0 duplicates in bank_addition_full.
```

OUTLIER DETECTION:

Outliers are the values which lie at above 3 standard deviation distance from the other Values in normal distribution.
It might occur due to improper collection of the data. . Outliers can disturb our analysis by changing the mean on normal distribution graph.

There are few techniques involved in outlier treatment. One of them is totally dropping the outliers from the dataset. This however would lead to loss of important data which can help in giving realistic predictions from the model. The outliers can be detected by creating univariate plots like histogram or box plot of the numeric features. For example, a box plot of the ‘age’ and ‘campaign’ features is visualized.


BOX PLOT FOR AGE FEATURE
20
30
40
50
60
70
80
90
100
age


BOX PLOT FOR CAMPAIGN FEATURE
0
10
20
30
40
50
campaign

The data points outside the whiskers of the box indicate there are a significant number of outliers in this distribution.

On a detailed examination, we looked at the maximum values of the 'age' and 'campaign' variables.

```
In [41]: #OUTLIER TREATMENT

         # On observation, features 'age' and 'campaing' shows outlier in their distribution.
         # 'Duration' is not being used in the analysis so ignoring it.
         bank_add_full[['age','campaign']].describe()
```

Out[41]:

	age	campaign
count	41176.00000	41176.000000
mean	40.02380	2.567879
std	10.42068	2.770318
min	17.00000	1.000000
25%	32.00000	1.000000
50%	38.00000	2.000000
75%	47.00000	3.000000
max	98.00000	56.000000

```
In [ ]: #Age feature is having maximum value as 98 which doesn't looks like as unrealistic so we will not treat outliers
        #for age variable
        #Campaign feature is having maximum value as 56 which is also realistic so we will not treat the outlier
```

The maximum value for 'age' variable is 98 and that of 'campaign' variable is 56 and both are significant values.

Since model is needed to be generalized to reflect the real world data we are not going to remove these outliers as these seems to be realistic value .

MISSING VALUE DETECTION:

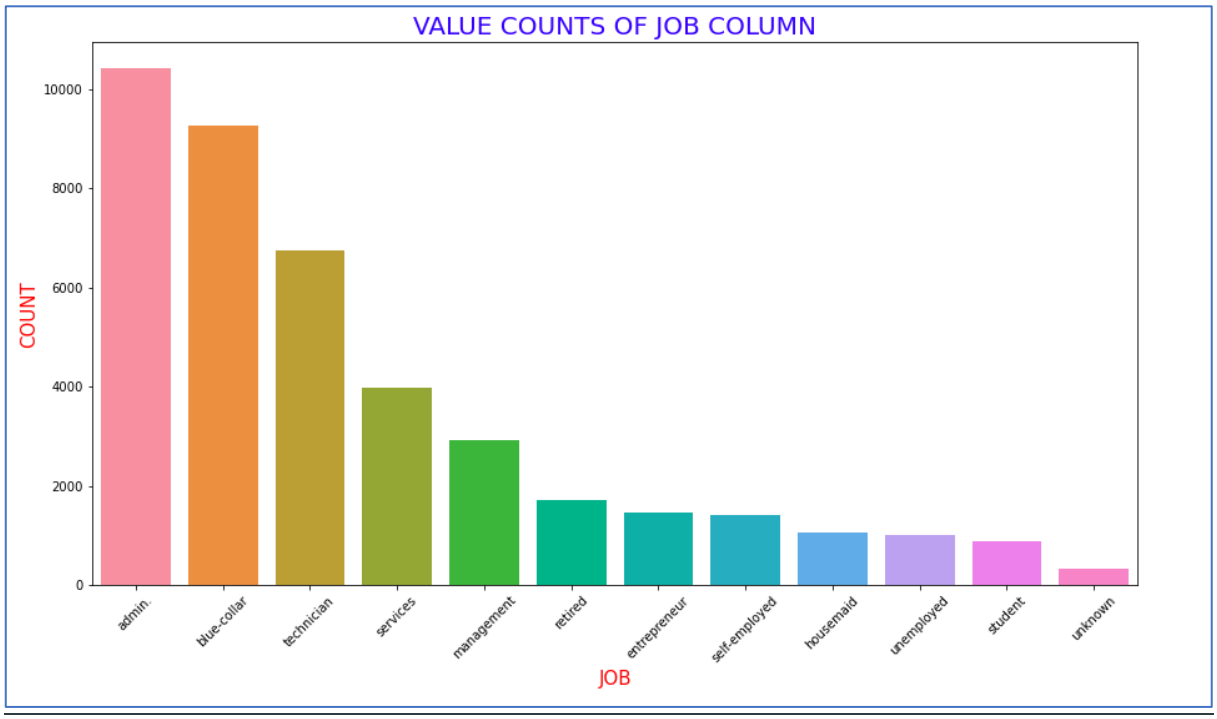
To check the missing values in the dataset we have done univariate analysis of the categorical and numerical features of the dataset.

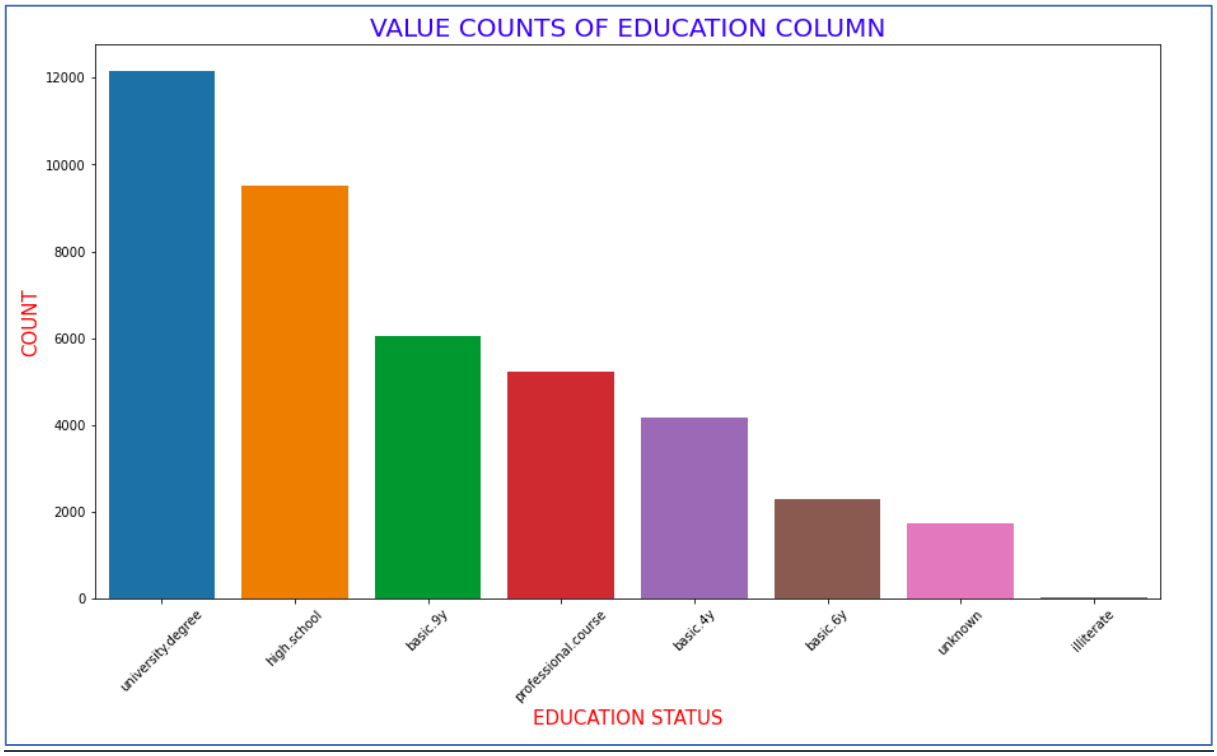
- **IMPUTATION OF CATEGORICAL FEATURES :**

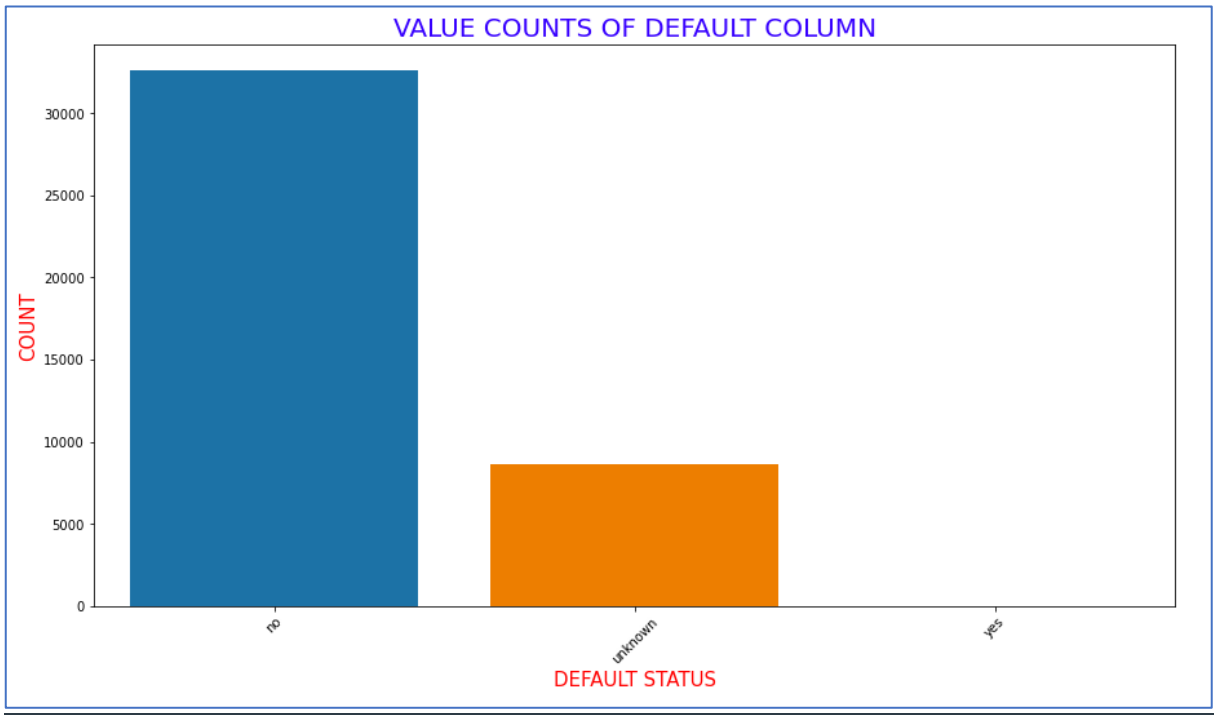
After having a look on value counts of different categorical variables we found that 'Unknown' category is present for many features.
We have different methods to impute the missing data from which one way is to drop the data which is having the feature variable as unknown

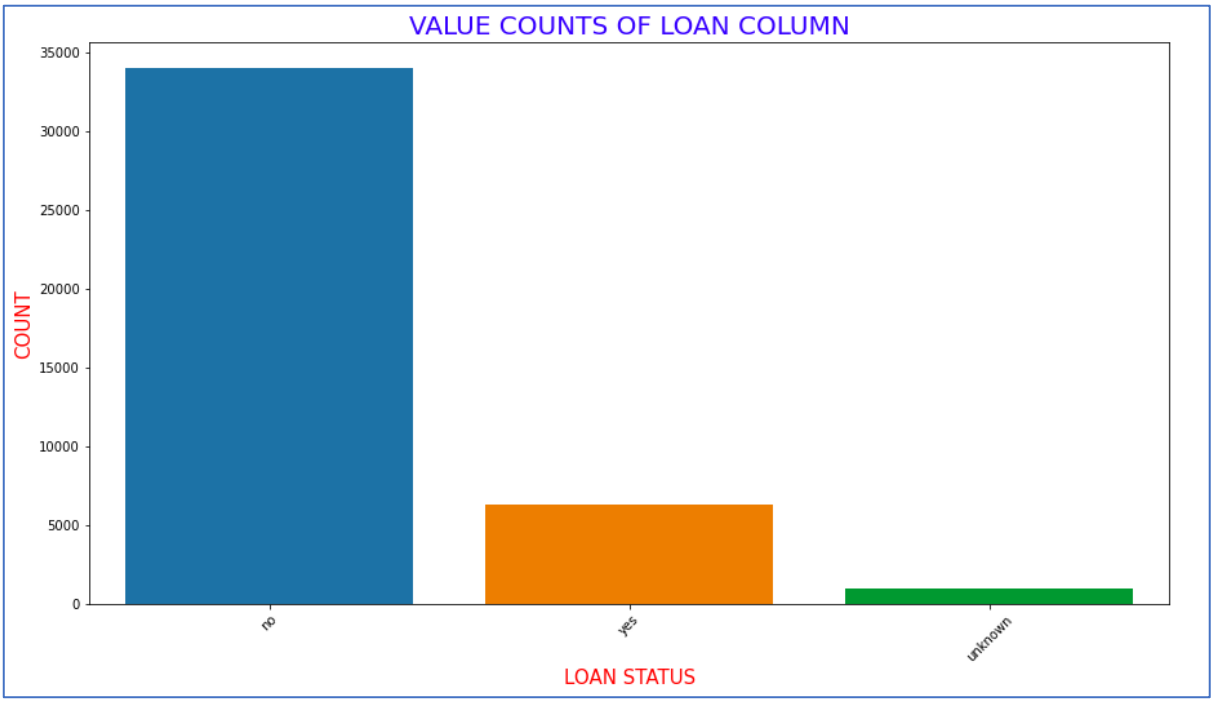
but this will affect the purpose of developing accurate and realistic prediction model.

Another way of imputing is to replace the mode or most occurring value of that particular feature with the missing or unknown value. This doesn't assure us to have the correct value for the missing value but majority of them will be addressed correctly and can help us improve the accuracy of the model. The problem with this imputation is it might create biasness in the data.
Features having 'Unknown' categories in the dataset are 'education', 'loan', job', 'marital', 'default'.
'Unknown' seems to be significant for 'education', 'job', 'loan', 'default'.




VALUE COUNTS OF EDUCATION COLUMN
COUNT
0
2000
4000
6000
8000
10000
12000
university.degree
high.school
basic.9y
professional.course
basic.4y
basic.6y
unknown
illiterate
EDUCATION STATUS


VALUE COUNTS OF DEFAULT COLUMN
COUNT
0
5000
10000
15000
20000
25000
30000
no
unknown
yes
DEFAULT STATUS



The other way to address the 'unknown' values in the categorical columns is by considering them as another category of the features. For example, on looking at the bar plot of 'default' feature, there are approximately 9000 clients with a response of 'unknown'. This means these clients do not want the bank to know their actual default status. Therefore this value can be a good addition in analysing the customer behaviour.

GITHUB LINK

https://github.com/AbhimanyuGangani/Week_7_Bank_Marketing/tree/main/Weel_9_bank_marketing