A PROJECT REPORT ON

Credit Card Customers Analysis

MES Institute of Management & Career Courses, Pune

Submitted By :2012113 Vaishnavi Rajput
2012114 Rajesh Rampure
2012115 Purva Rane
2012116 Venkatesh Ranvirkar

Mini Project INDEX

Chapter 1 : Introduction

1.1 Scope of system:

The Project is about the Analysis of the Credit Card Usage with the use of python and its various packages. And also implementing the various Machine Learning modules to predict the future behaviors of customers by using algorithms in machine learning. And also to know the behavior of the variables and to analyze attributes that indicate strong relationship with the cancellation of credit card service customers.

1.2 Module Assigned to

Module	Assigned to Member
EDA(sem-1) Model Deployment(sem-2)	Member -1 Vaishnavi Rajput
EDA(sem-1) Model Selection and Evaluation (sem-2)	Member-2 Rajesh Rampure
EDA(sem-1) Introduction (sem-2)	Member-3 Purva Rane
EDA(sem-1) Feature Engineering (sem-2)	Member -4 Venkatesh Ranvirkar

1.3 Operating Environment – Hardware and Software:

Software Specification :-

- **Operating System :-**

  All known Desktop Operating Systems(Windows 7+, Linux, Mac OS 10.07+)

- **Database :**

  MS excel

- **Browser :-**

  Chrome ver.60 or higher ,Firefox ver.70 or higher,Internet Explorer 10 or higher , Safari 10 or higher.

- **Tools and IDE :-**

  Python , Jupyter IDE,Visual Studio

- **Technologies :-**

  Python Packages

Hardware Specification :-

- **Preprocessor :-**

  Intel Core 2 Duo or AMD Athlon 64x 5600+

- **RAM :-**

  2GB or higher

- **Hard Disk :-**

  80 GB or higher

Chapter 2: Proposed System

- Provides Machine Learning Model to Predict Churning of Customers Using Python.
- And also providing web interface where user can enter the particular attributes and check the attrition of the customer from bank based on passed data in the inputs fields.

.

Chapter 3: Data Analysis

3.1 Exploratory Data Analysis :-

Exploratory Data Analysis is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods.
Here we have firstly imported various libraries to perform various operation and also visualization of data.They are as follows :-

1.NUMPY

NumPy, which stands for Numerical Python, is a library consisting of multidimensional array objects and a collection of routines for processing those arrays.
Using NumPy, mathematical and logical operations on arrays can be performed.

Operations using NumPy:
Using NumPy, a developer can perform the following operations −

Mathematical and logical operations on arrays.

Fourier transforms and routines for shape manipulation.

Operations related to linear algebra. NumPy has in-built functions for linear algebra and random number generation.

2. PANDAS

pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with “relational” or “labeled” data both easy and intuitive.
It aims to be the fundamental high-level building block for doing practical, real-world data analysis in Python.
Additionally, it has the broader goal of becoming the most powerful and flexible open source data analysis/manipulation tool available in any language.
It is already well on its way toward this goal.

FEATURES:

Fast and efficient DataFrame object with default and customized indexing.
Tools for loading data into in-memory data objects from different file formats.
Data alignment and integrated handling of missing data.
Reshaping and pivoting of date sets.
Label-based slicing, indexing and subsetting of large data sets.
Columns from a data structure can be deleted or inserted.
Group by data for aggregation and transformations.
High performance merging and joining of data.
Time Series functionality.

3. MATPLOTLIB

Matplotlib is one of the most popular Python packages used for data visualization.
It is a cross-platform library for making 2D plots from data in arrays.
Matplotlib is written in Python and makes use of NumPy, the numerical mathematics extension of Python.

%matplotlib inline turns on “inline plotting”, where plot graphics will appear in your notebook. This has important implications for interactivity: for inline plotting, commands in cells below the cell that outputs a plot will not affect the plot.

4. PLOTLY

The Plotly Python library is an interactive open-source library.
This can be a very helpful tool for data visualization and understanding the data simply and easily.
plotly graph objects are a high-level interface to plotly which are easy to use.
It can plot various types of graphs and charts like scatter plots, line charts, bar charts, box plots, histograms, pie charts, etc.

Plotly has hover tool capabilities that allow us to detect any outliers or anomalies in a large number of data points.
It is visually attractive that can be accepted by a wide range of audiences.
It allows us for the endless customization of our graphs that makes our plot more meaningful and understandable for others.

5. Scikit-learn :-

Sklearn is probably the most useful library for machine learning in Python. The sklearn library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction.Please note that sklearn is used to build machine learning models. It should not be used for reading the data, manipulating and summarizing it. There are better libraries for that (e.g. NumPy, Pandas etc.)

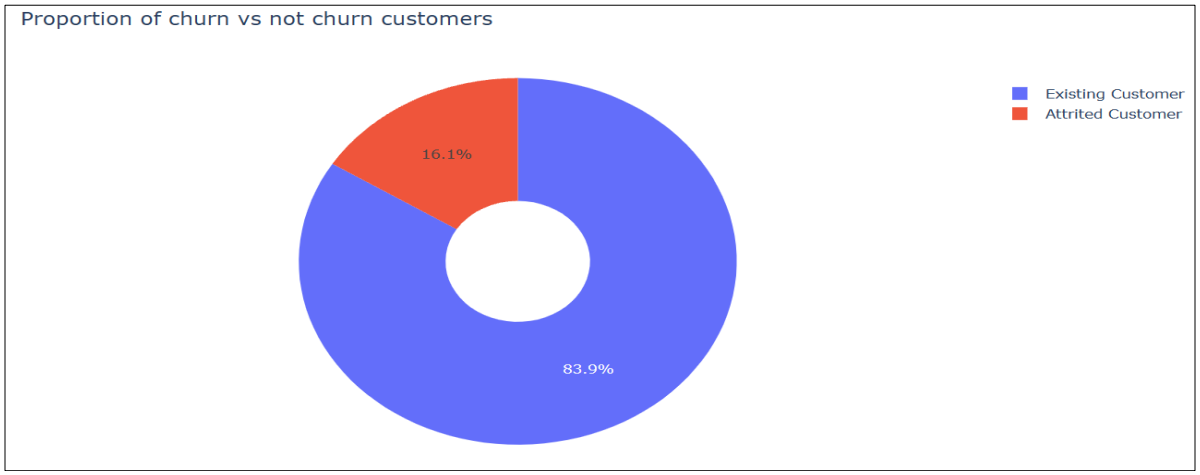
6.Pickle :-

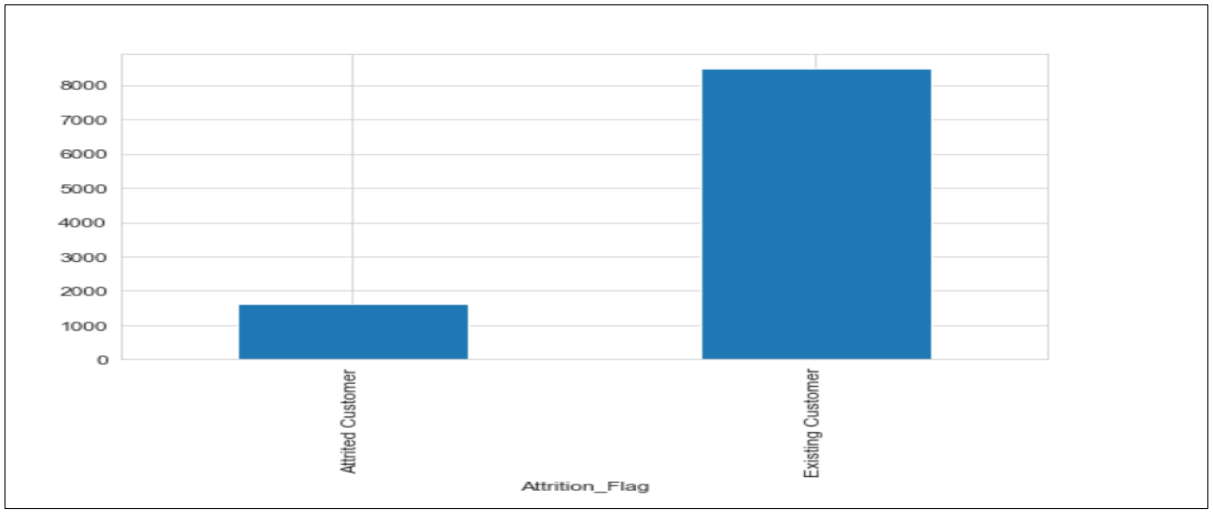
Python pickle module is used for serializing and de-serializing a Python object structure. Any object in Python can be pickled so that it can be saved on disk. What pickle does is that it “serializes” the object first before writing it to file. Pickling is a way to convert a python object (list, dict, etc.) into a character stream. The idea is that this character stream contains all the information necessary to reconstruct the object in another python script.

Visualizations on various features in the dataset :-

1. ATTRITION FLAG :

Here we are going to draw Pie chart and Bar Plot according to attrition flag means how many customers are attired and how many customers are still existing.

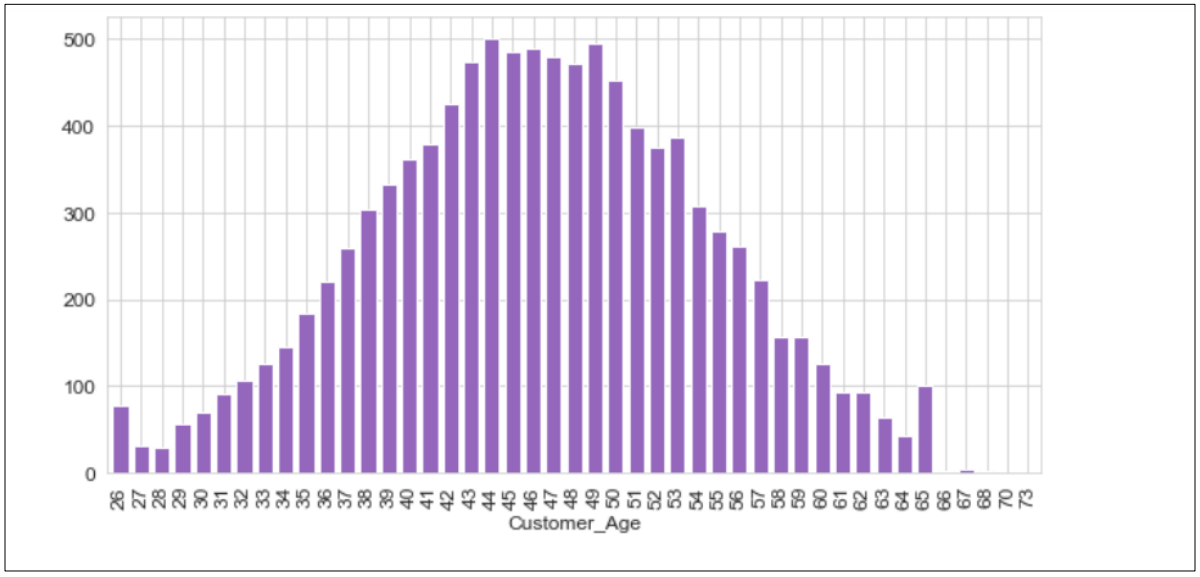




From above graph we are seeing16% of customers are attrited and 83.9% are existing customers.

2. CUSTOMER AGE :-

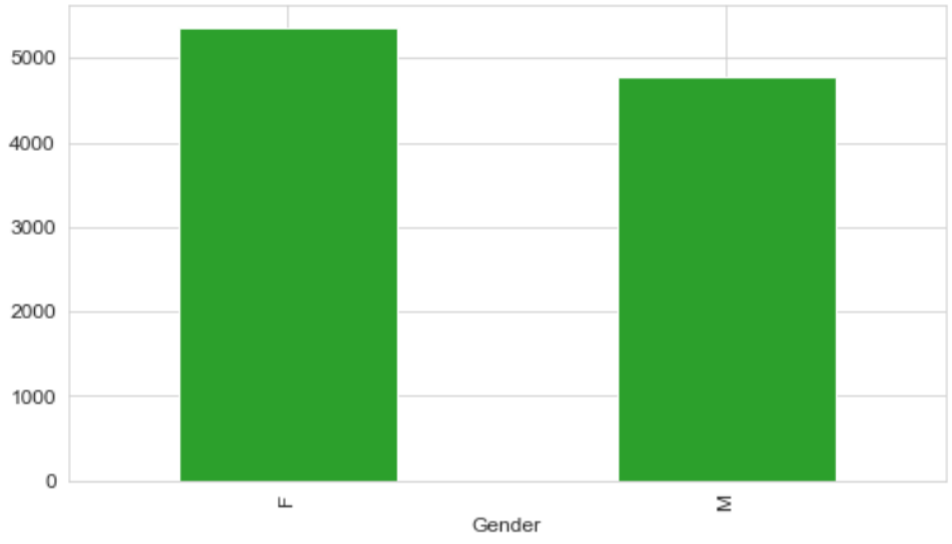
Here we are going to draw Bar Plot to see age distribution of customer.

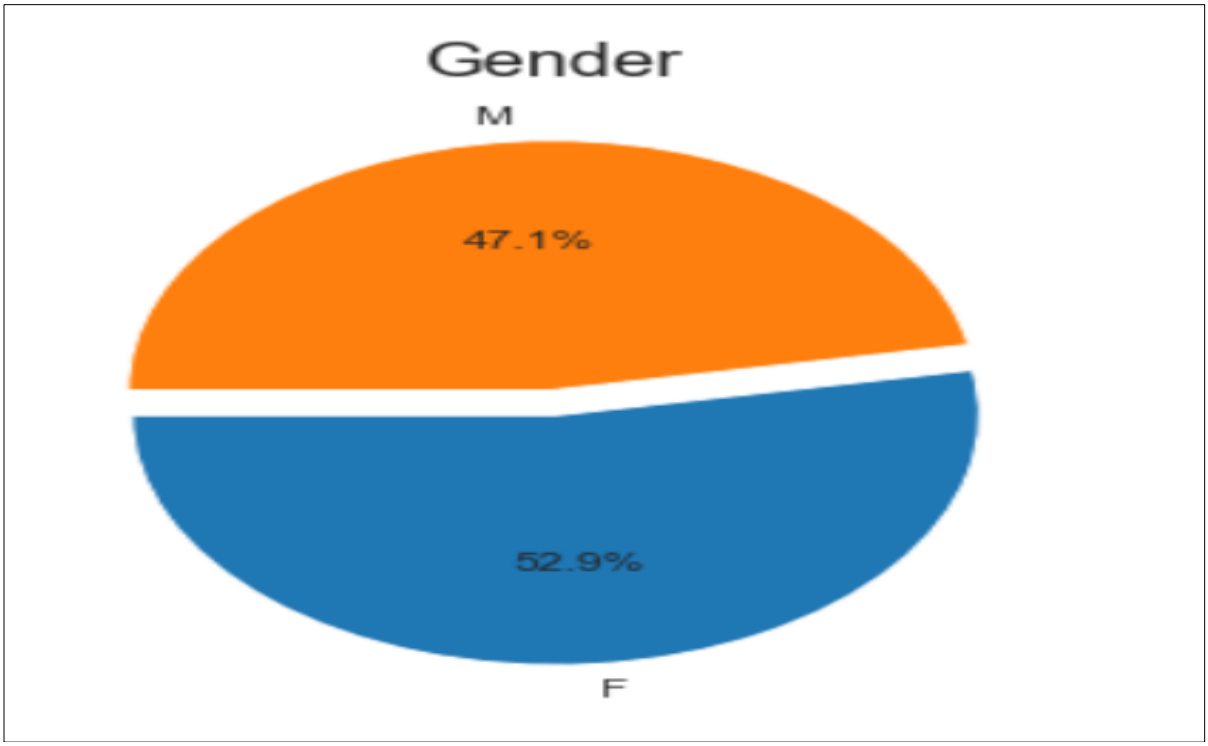


It seems like it has normal distribution.

3. Gender :-

Here we have shown the Gender distribution using Pie Chart and BoxPlot

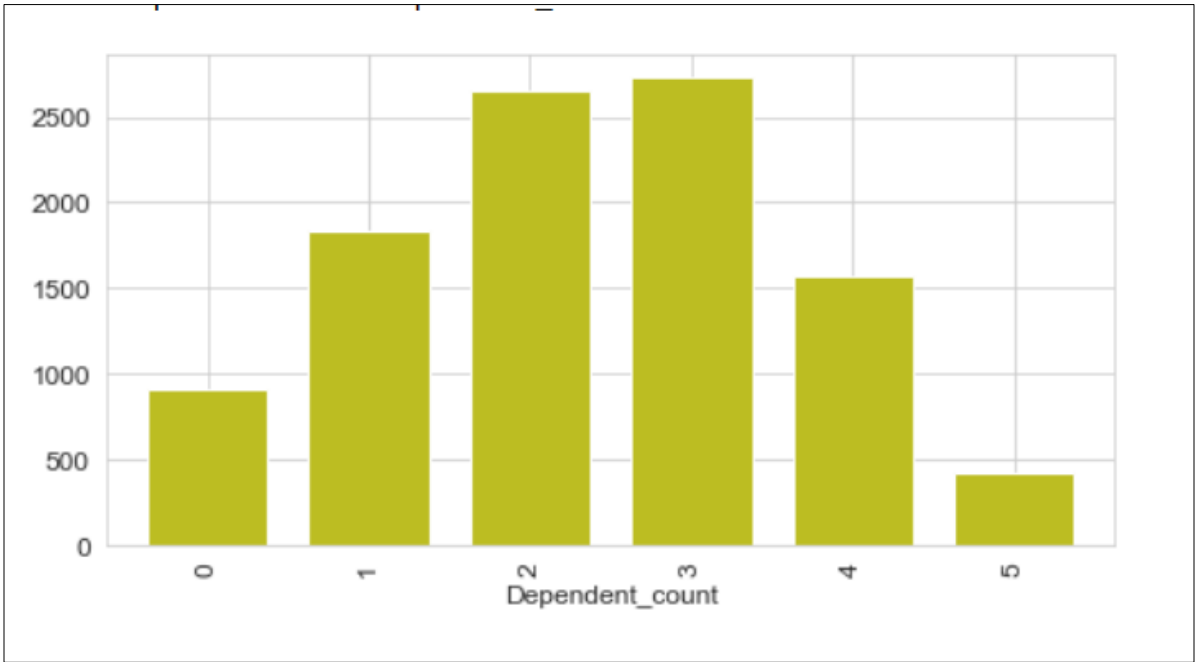




So as we see there are 47.1% Male customers and 52.9% Female customers.

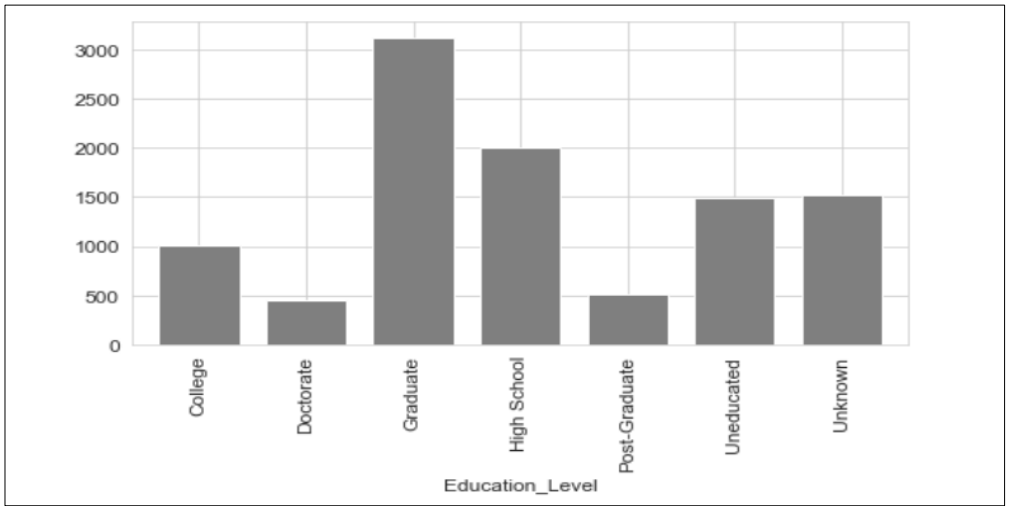
4. DEPENDENT COUNT :-

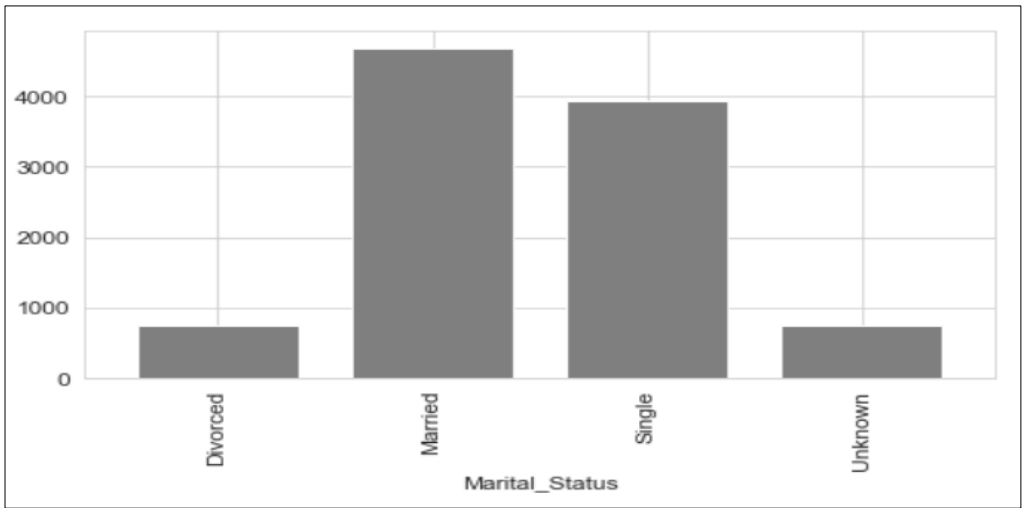
Here we are plotting the Bar Plot to see how many dependent person present on each customer.

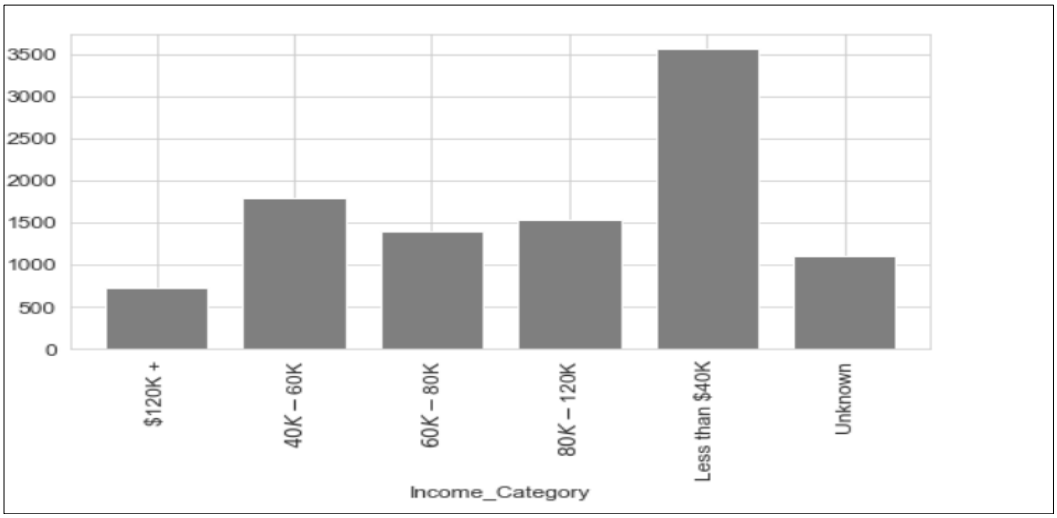


5. EDUCATION LEVEL, MARITAL STATUS, INCOME CATEGORY :-

Here we have plotted Bar Plot for EDUCATION LEVEL, MARITAL STATUS, INCOME CATEGORY.

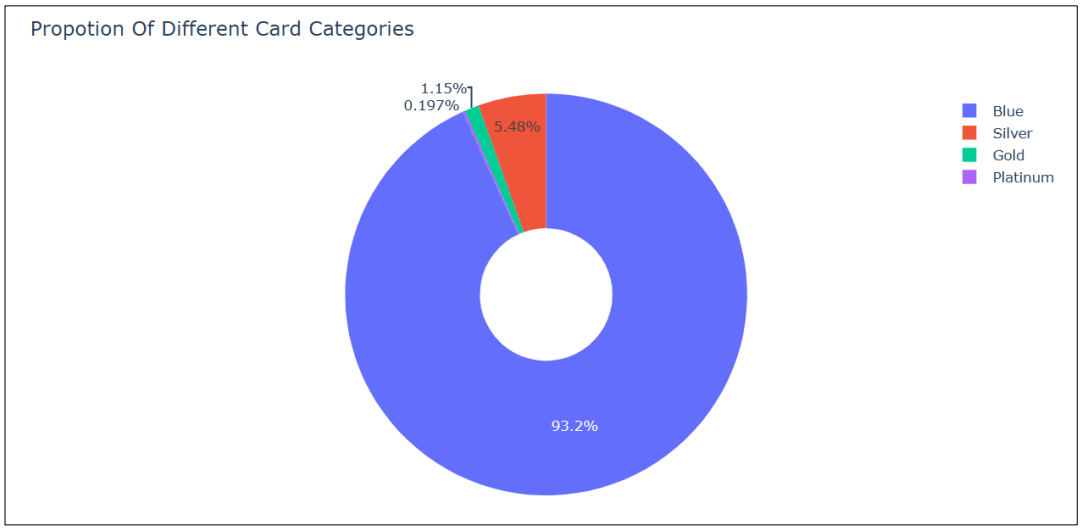






6. Card Category :-

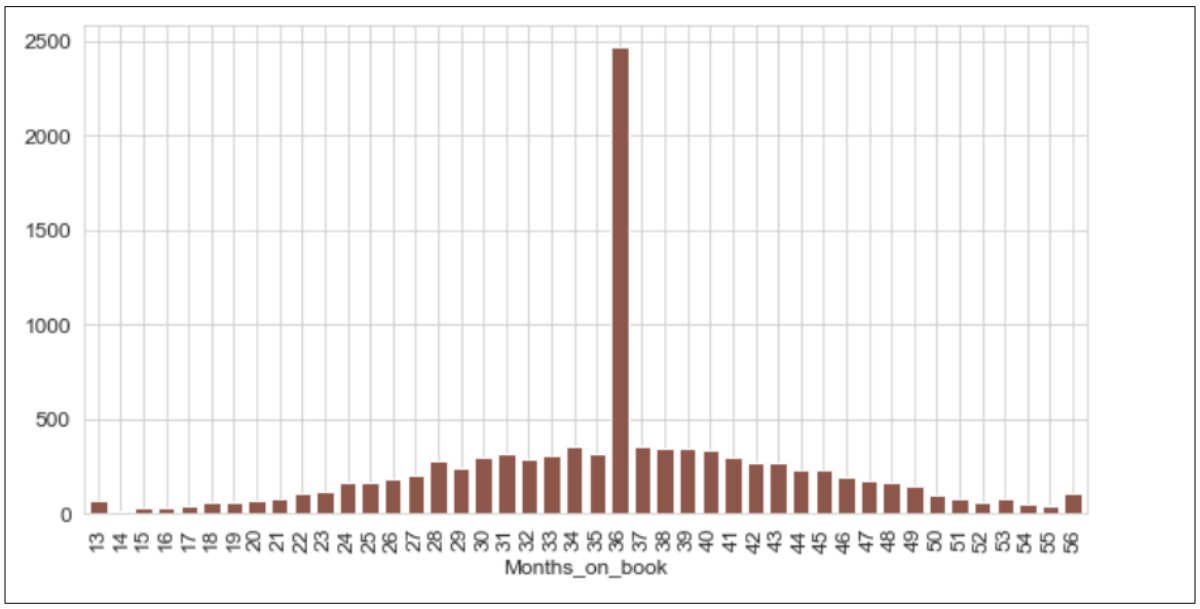
Here we have drawn Pie Chart according to the card category that is Blue Card, Platinum Card, Silver Card, Gold Card.



Blue cards are the the most popular (which is pretty obvious, they are the cheapest). But other cards are rare.

7. MONTHS ON BOOK :-

It shows the activeness of customer .We have plotted this using Bar Plot.

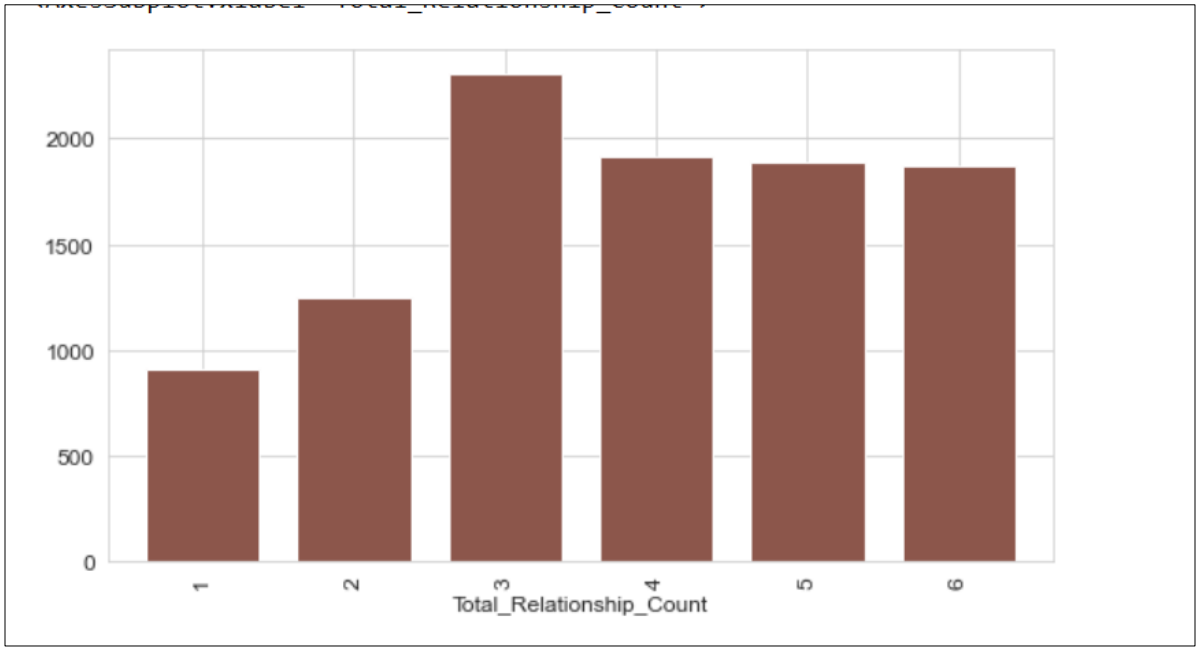


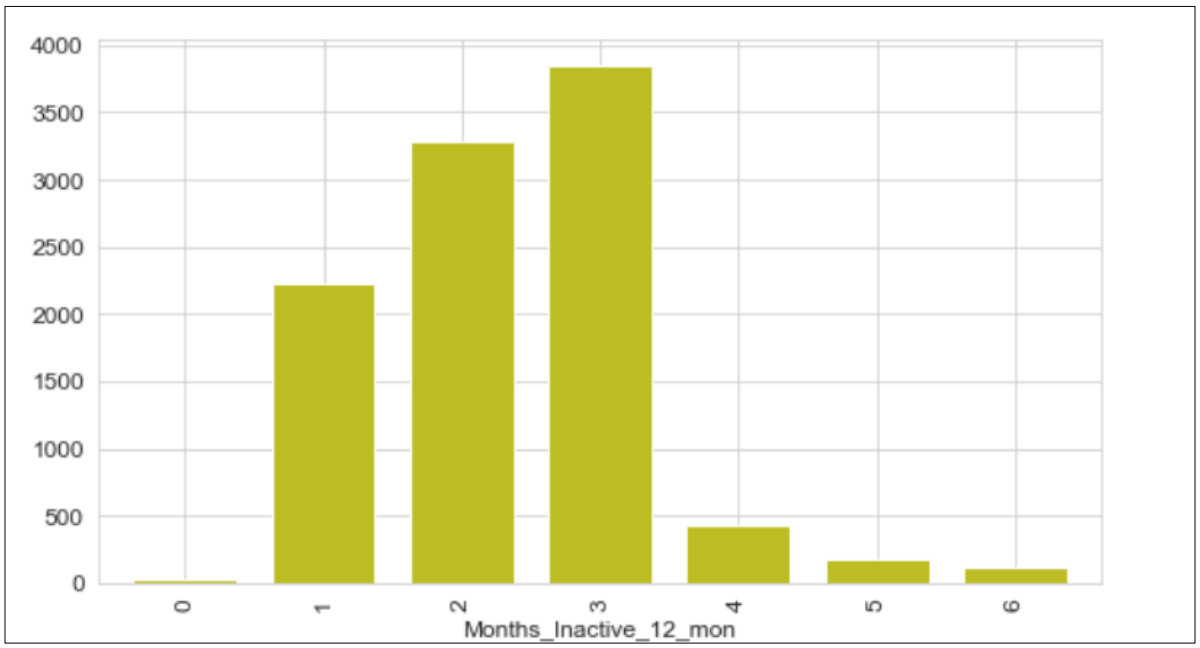
There is much more values with 36 values. Maybe there was an discount.

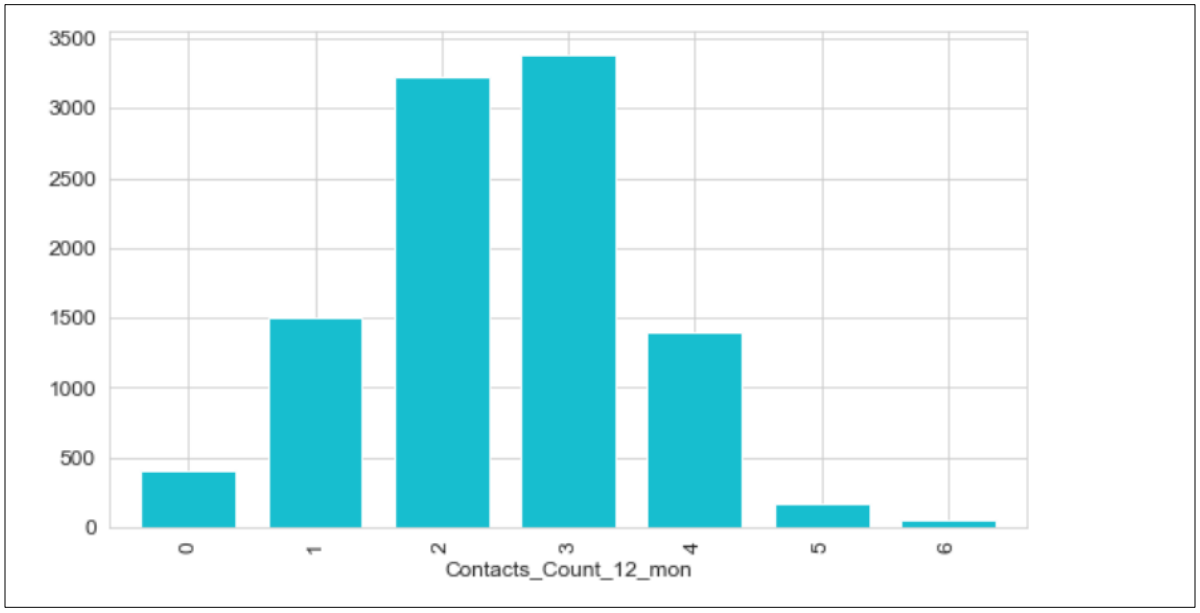
8. TOTAL RELATIONSHIP COUNT, MONTHS INACTIVE 12 MONTHS, CONTACTS COUNT 12 MONTHS :-

Here we are going to plot Bar Plot for :-

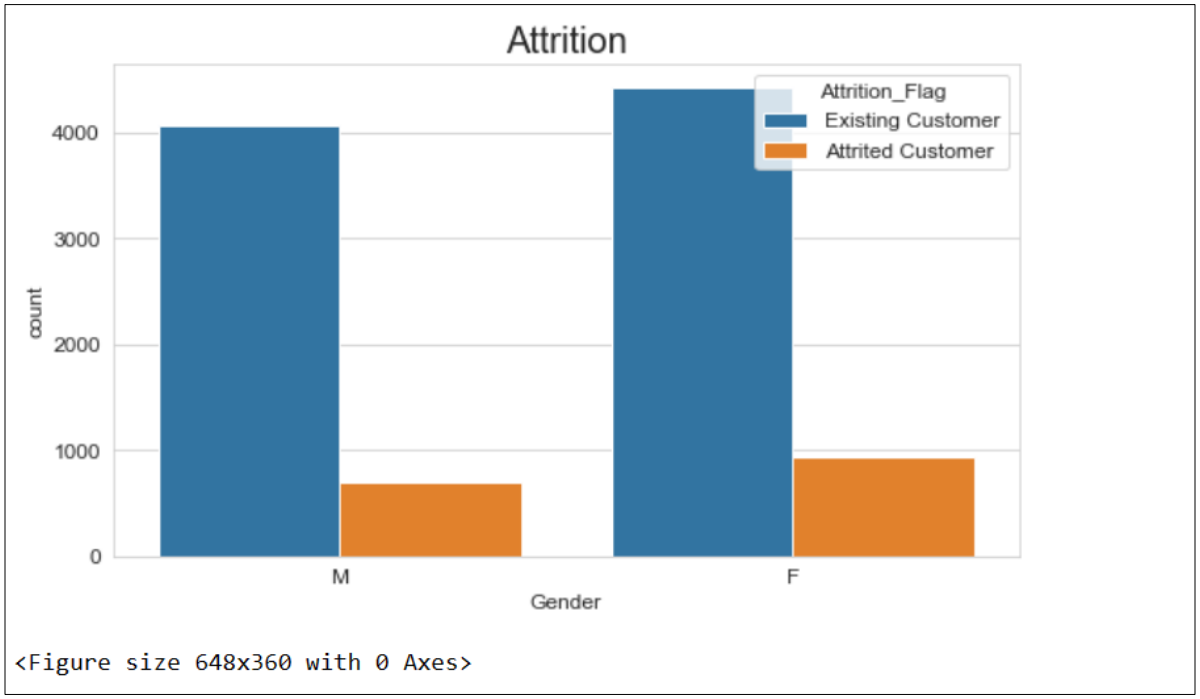
TOTAL RELATIONSHIP COUNT
MONTHS INACTIVE 12 MONTHS
CONTACTS COUNT 12 MONTHS



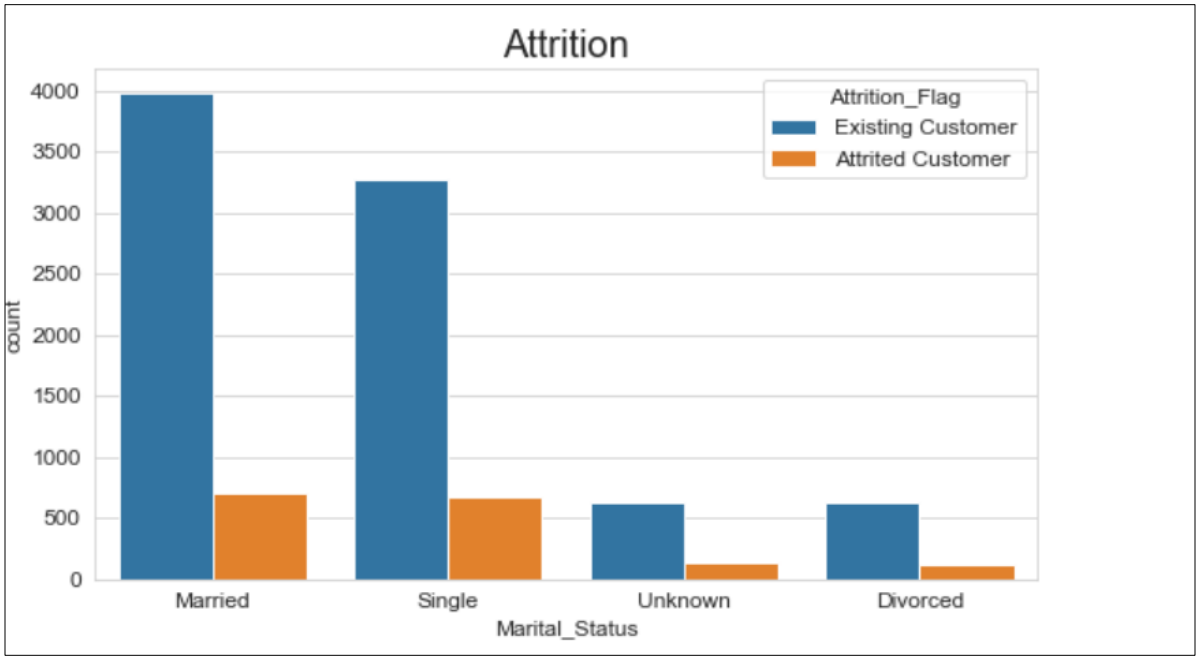




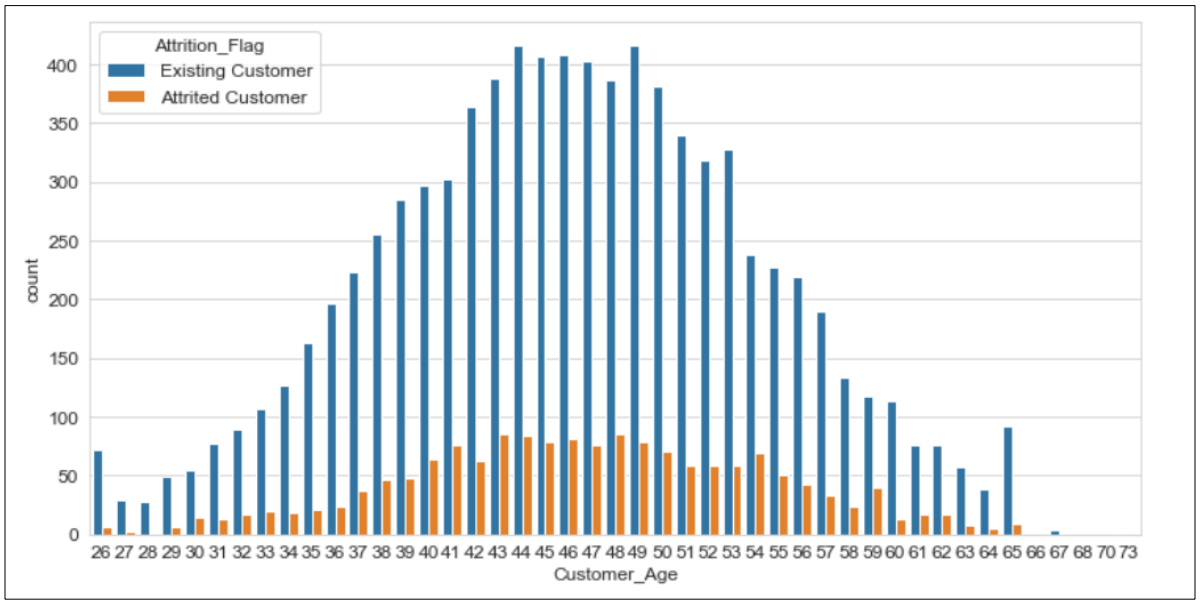
DEMONSTRATING THE CUSTOMER ATTRITION ACCORDING TO THE GENDER :



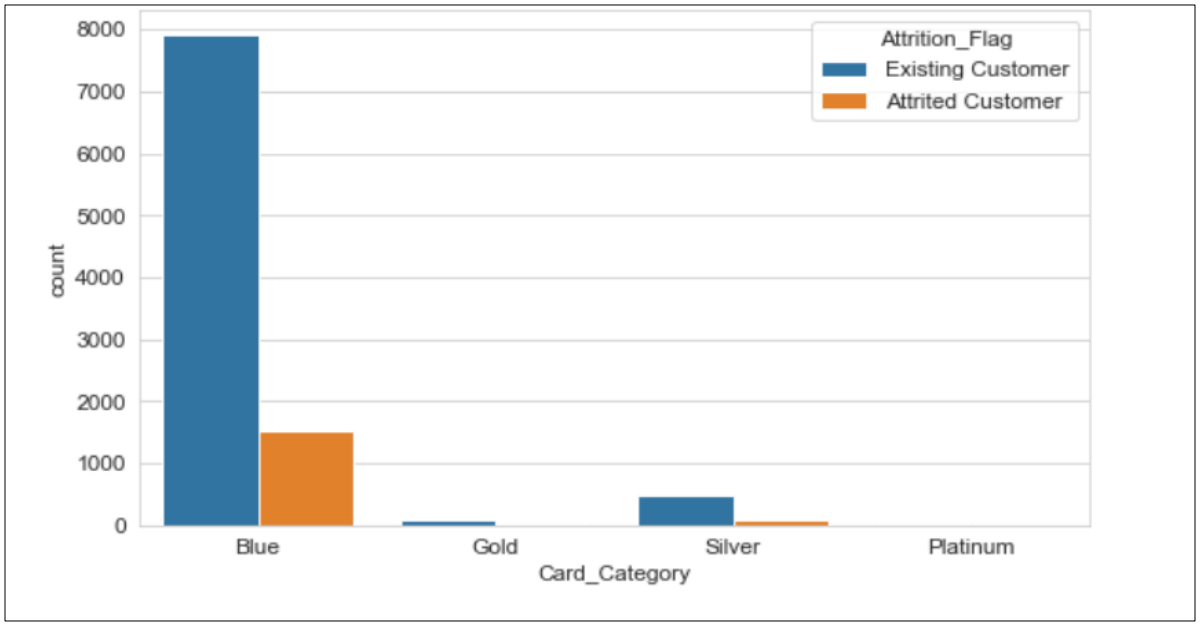
DEMONSTRATING THE CUSTOMER ATTRITION ACCORDING TO THE MARITAL STATUS :



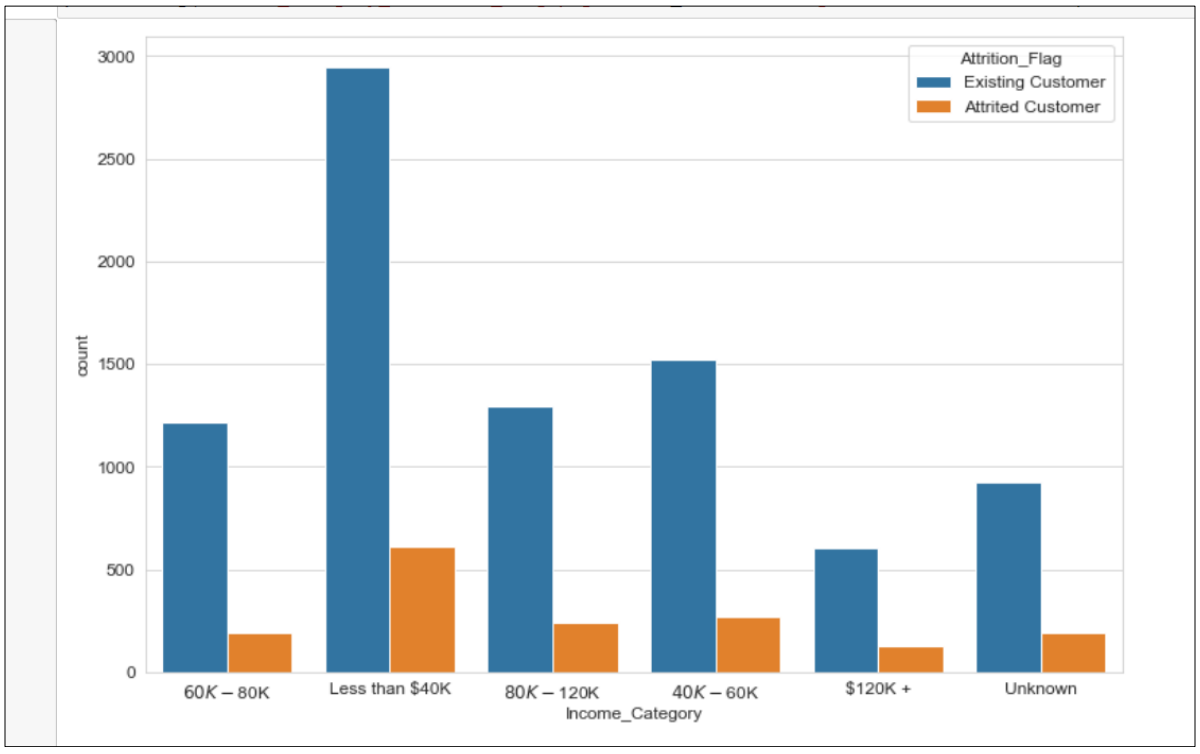
DEMONSTRATING THE CUSTOMER ATTRITION ACCORDING TO THE CUSTOMER AGE :



DEMONSTRATING THE CUSTOMER ATTRITION ACCORDING TO THE CARD CATEGORY :

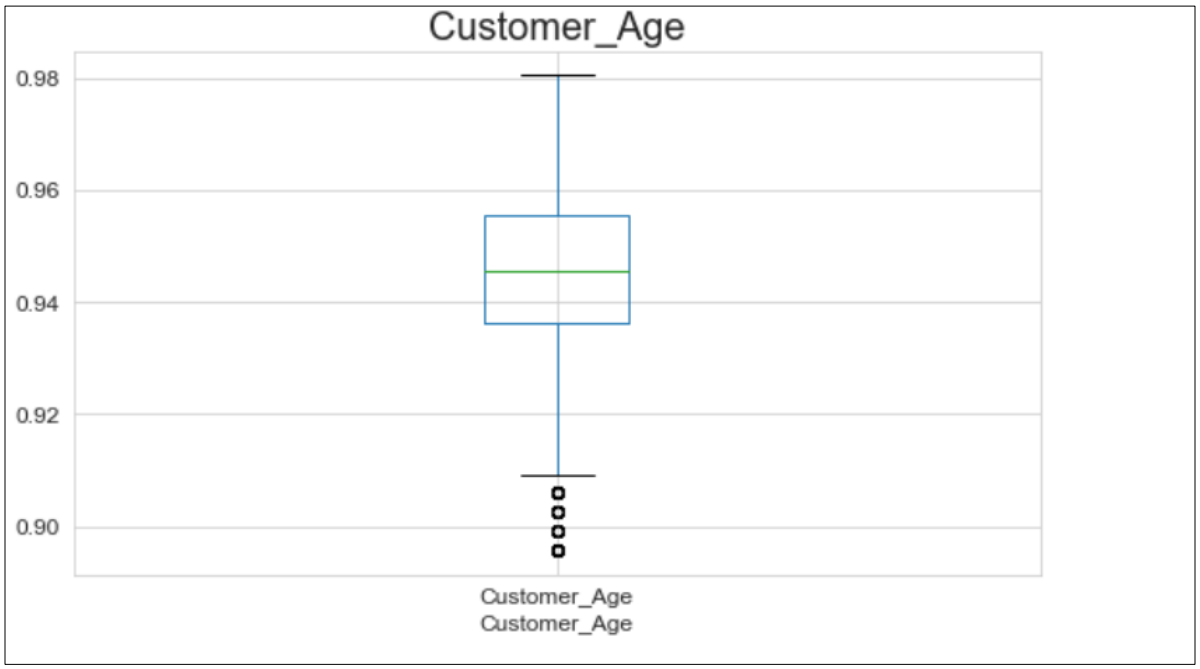


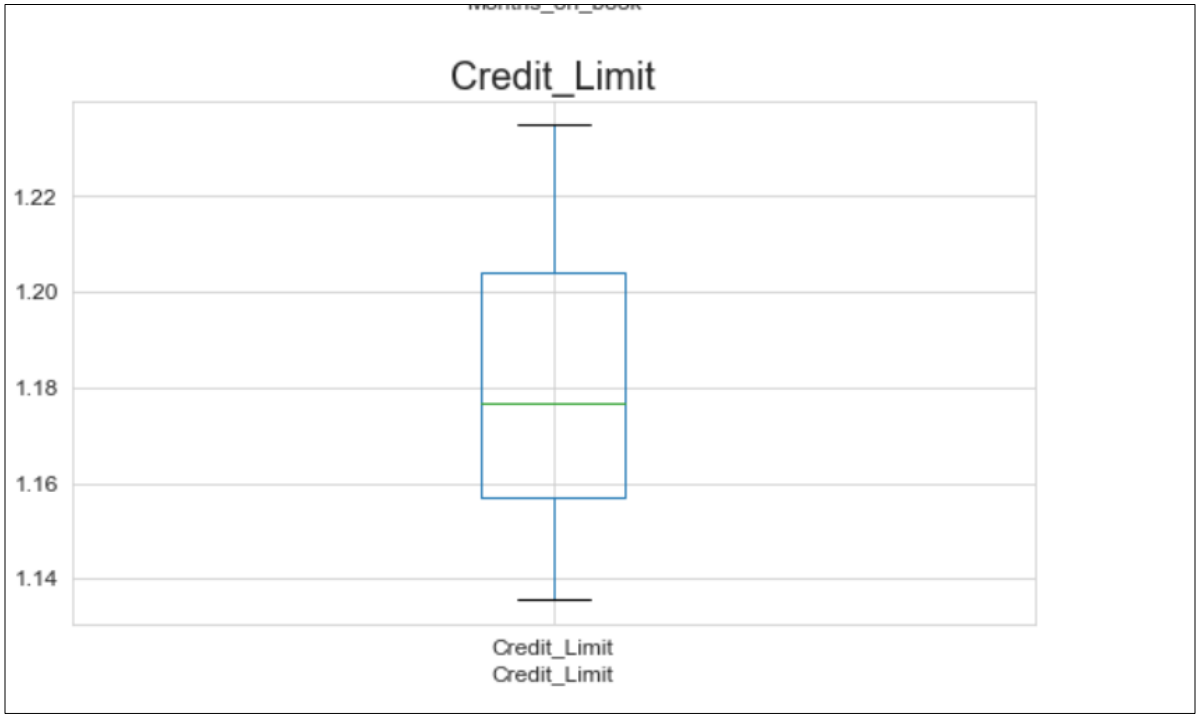
DEMONSTRATING THE CUSTOMER ATTRITION ACCORDING TO THE INCOME CATEGORY :



Outliers :-

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.





3.2. Feature Engineering :-

The main aim of this stage is to know about the data .Exploring data on different levels seeing the different parameters in the data and mark it down which will help in future analysis.

In Feature Engineering we will analyze to find out the below stuff :-

1. Missing Values
2. All the numerical Variables.
3. Distribution of the numerical variables.
4. Outliers.
5. Categorical variables.
6. Encoding of categorical variables.

1. Missing Values :-

The first step comes when we get the data is to handle missing values in data. For that we have to see our data first after seeing all the data we are able to find the missing values in the data using the python in-build library called as Pandas's function known as df.info() which helps us to know the missing values or null values in our dataset.

By knowing that we can easily handle the missing values according to the data and according to the amount of missing values present in the dataset. There are many ways to replace the missing values such as replace with mean,meadian,mode and by particular values. As our Bank churners dataset does not contain any missing values ,so we don not have to worry about our dataset missing values.

2. All the numerical variables :-

Here all the numerical variables are get categorized means now we are getting all the columns in the dataset whose values are in the form of numerical data that is in number format. By using Pandas library we can perform this action of collecting all numerical columns data from dataset. By checking datatypes we can perform this operation by checking the datatypes of all the columns and classify by its types means in numerical and in categorical .Here there are 14 columns whose datatype is numerical they are as follows :-

Total number of numerical features: 14
['Customer_Age', 'Dependent_count', 'Months_on_book',
'Total_Relationship_Count', 'Months_Inactive_12_mon', 'Contacts_Count_12_mon',
'Credit_Limit', 'Total_Revolving_Bal', 'Avg_Open_To_Buy',
'Total_Amt_Chng_Q4_Q1', 'Total_Trans_Amt', 'Total_Trans_Ct',
'Total_Ct_Chng_Q4_Q1', 'Avg_Utilization_Ratio']

3. Distribution of the numerical variables :-

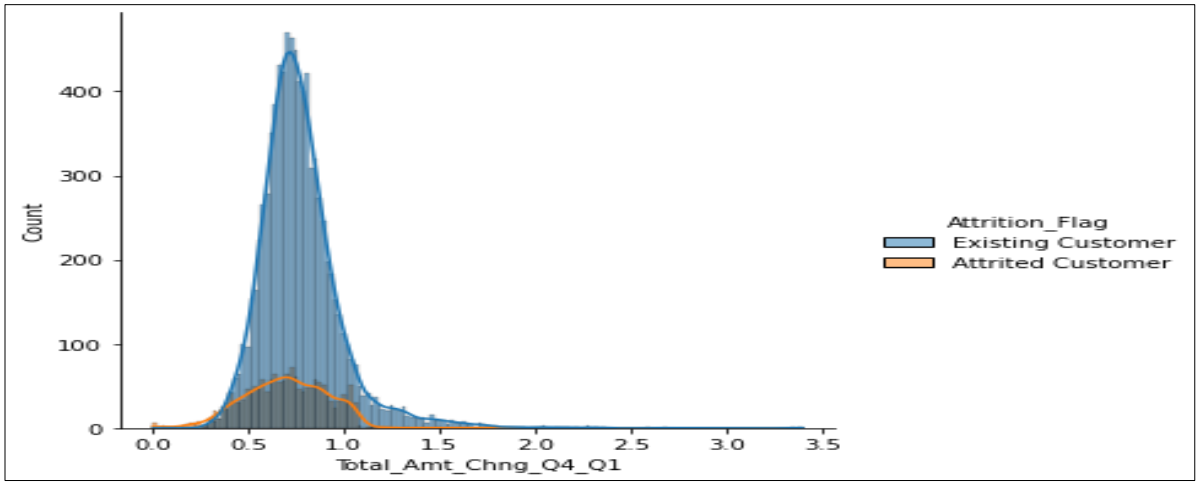
Now we have the numerical variables list which are present in our dataset. Now the further classification in numerical variable is divided into 2 categories they are :

- continuous variables :-

  A continuous variable is defined as a variable which can take an uncountable set of values or infinite set of values. For instance, if a variable over a non-empty range of the real numbers is continuous, then it can take on any value in that range. In our dataset this are continuous variables :-

  ['Customer_Age','Months_on_book','Credit_Limit','Total_Revolving_Bal','Avg_Open_To_Buy','Total_Amt_Chng_Q4_Q1','Total_Trans_Amt','Total_Trans_Ct','Total_Ct_Chng_Q4_Q1','Avg_Utilization_Ratio']

  Here the graphical representation of the continuous variable using python inbuilt libraray called seaborn according to the attired and existing customers :-

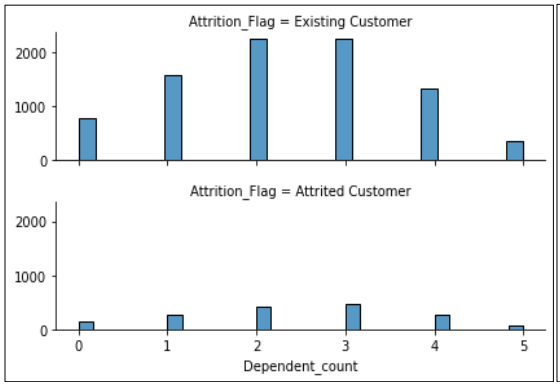


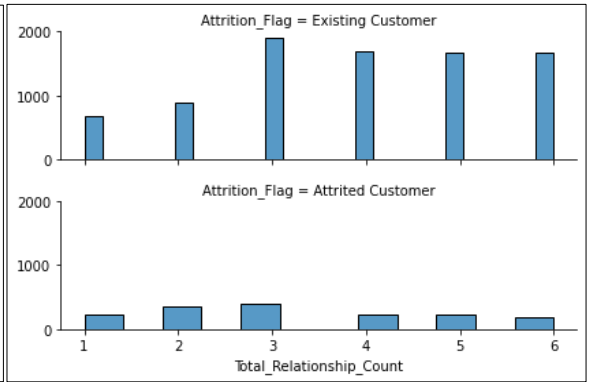
- discrete variables :-

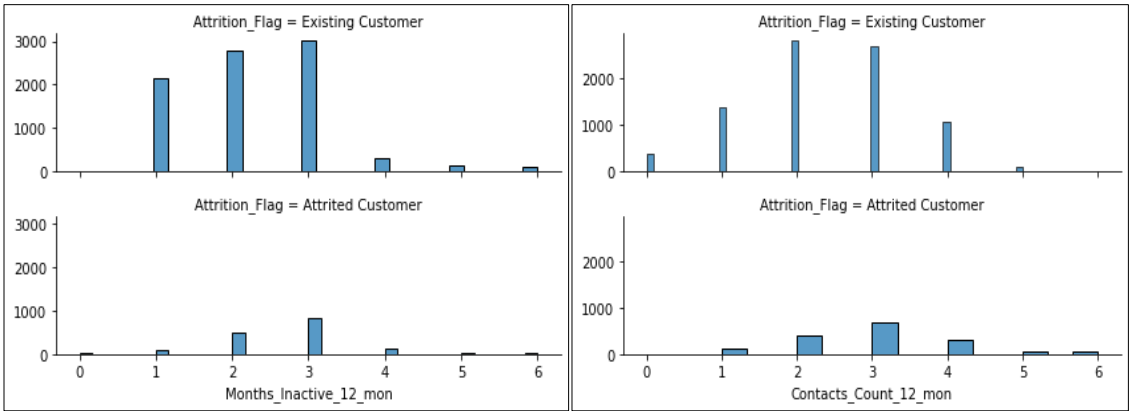
  A discrete variable is a type of statistical variable that can assume only fixed number of distinct values and lacks an inherent order. Also known as a categorical variable, because it has separate, invisible categories.In our dataset following are the discrete variables :-

  ['Dependent_count','Total_Relationship_Count','Months_Inactive_12_mon', 'Contacts_Count_12_mon']

  Here the Graphical representation of discrete variables using python in-built library called seaborn according to attired and existing customers :-





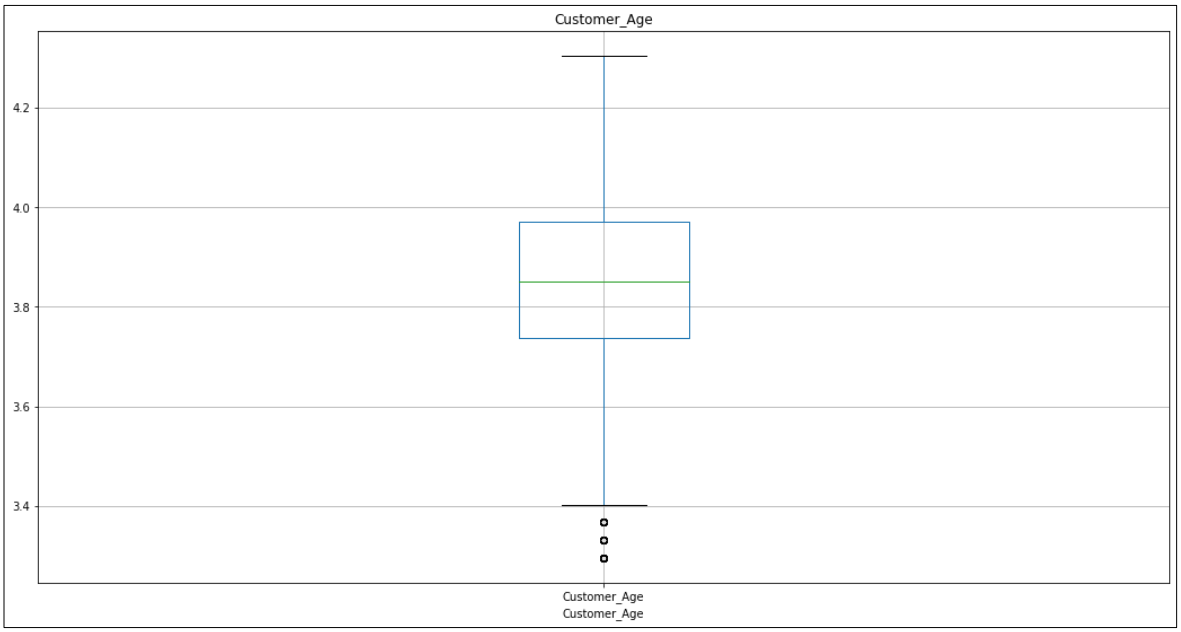


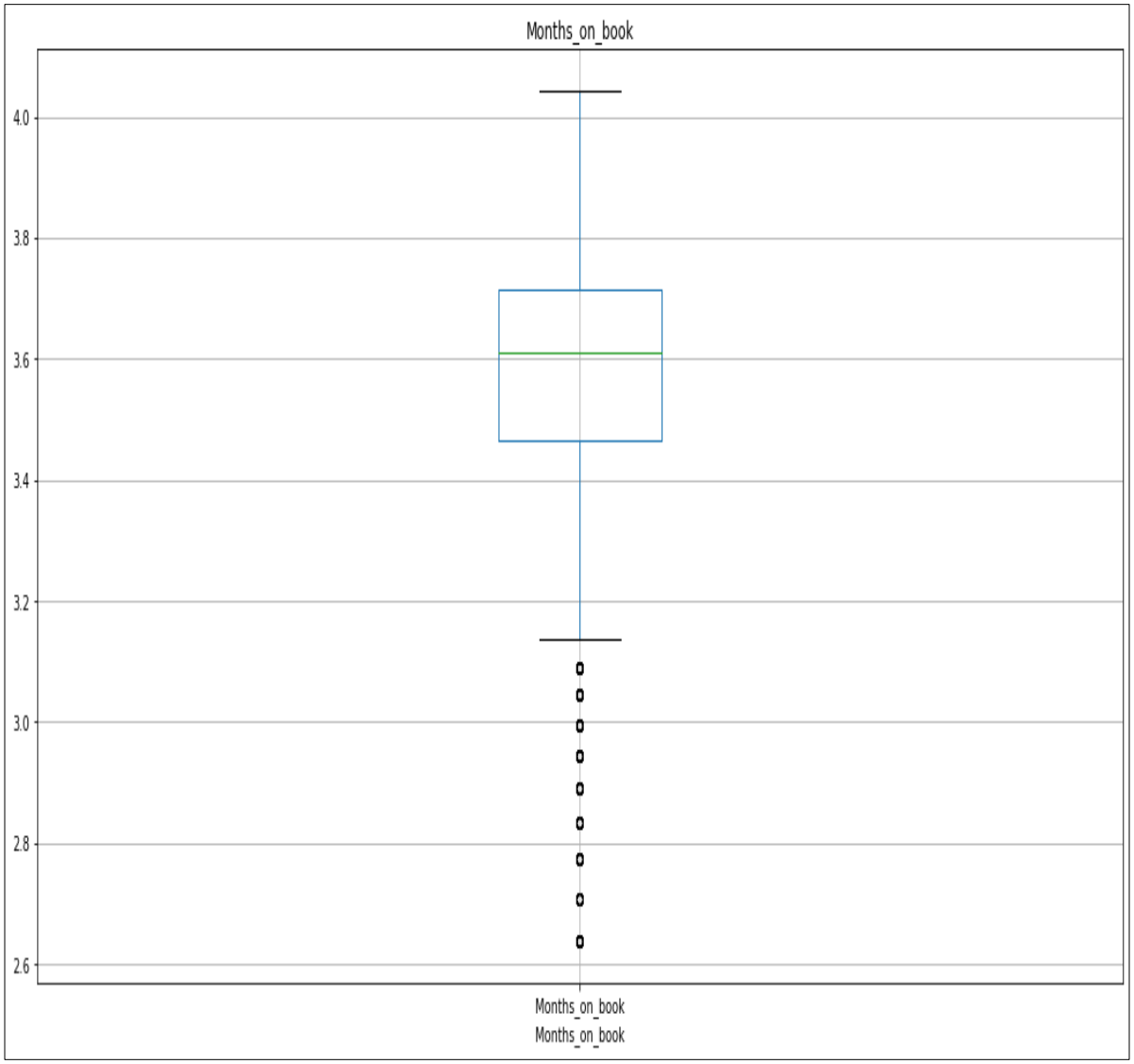
4. Outliers :-

In simple terms, an outlier is an extremely high or extremely low data point relative to the nearest data point and the rest of the neighbouring co-existing values in a data graph or dataset you're working with. Outliers are extreme values that stand out greatly from the overall pattern of values in a dataset or graph.

In our dataset as we can see in above graphs there are outliers are present which we are going to see through graph we have used box plot to plot outliers in our dataset using python library matplotlib .

Here are some outliers shown :-




Months_on_book
4.0
3.8
3.6
3.4
3.2
3.0
2.8
2.6
Months_on_book
Months_on_book

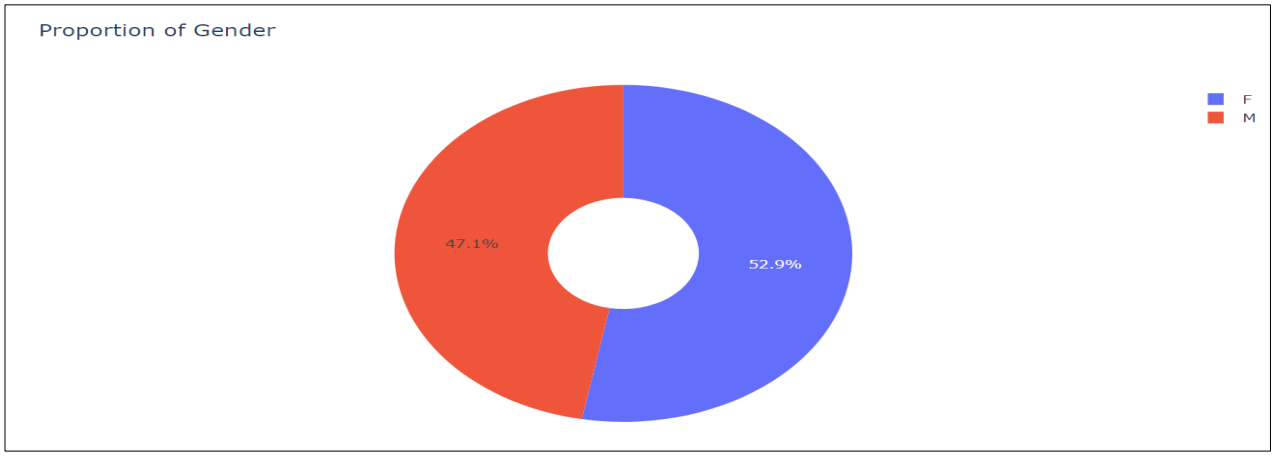
5. Categorical Variables :-

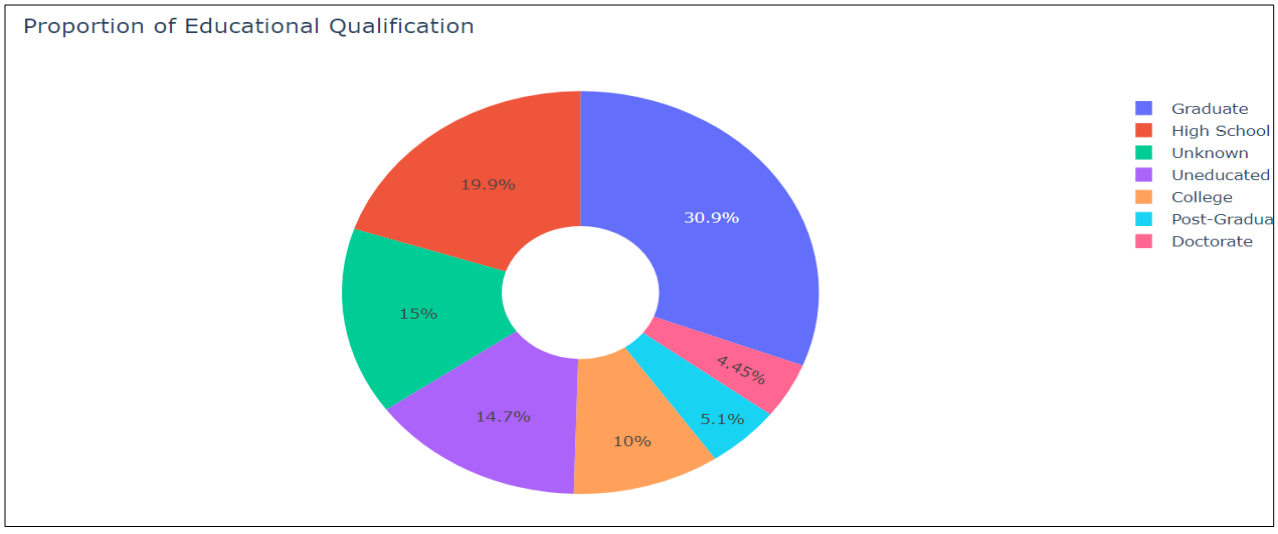
The categorical data consists of categorical variables which represent the characteristics such as a person's gender, hometown etc**.** Categorical measurements are expressed in terms of natural language descriptions, but not in terms of numbers.

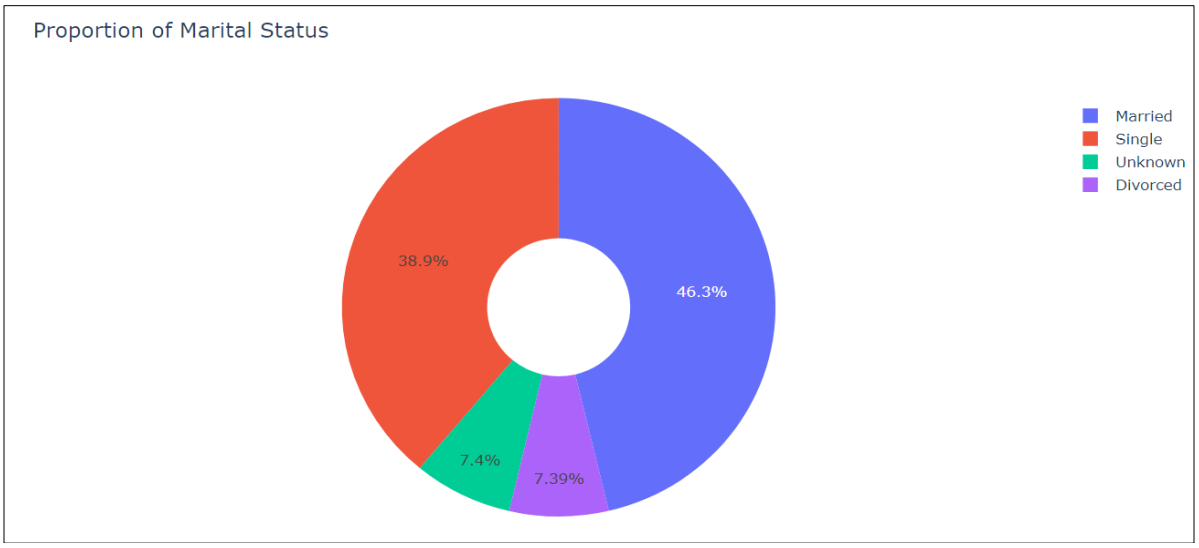
In our dataset we have following categorical columns :-

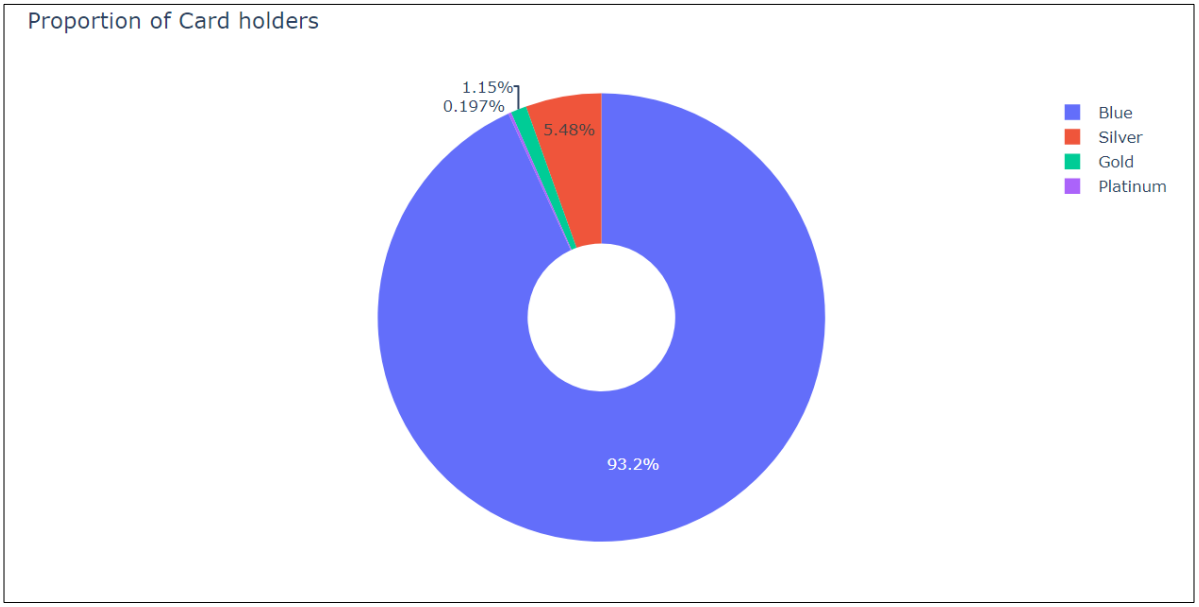
['Gender','Education_Level','Marital_Status','Income_Category','Card_Category']

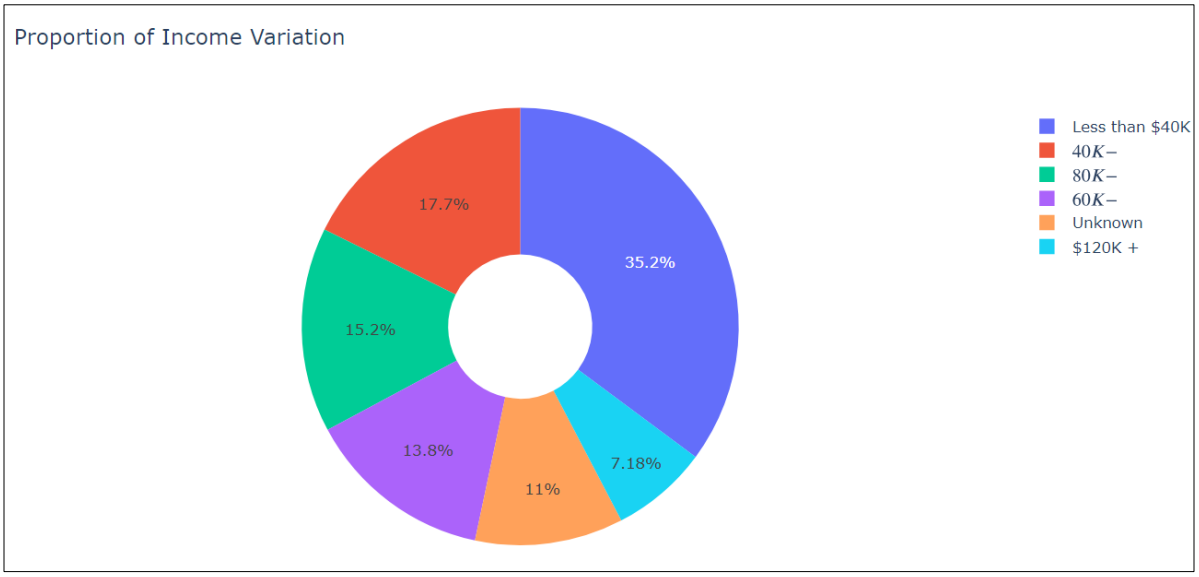
Categorial variable contains very less number of different categories so we will try to plot them in see the relation.
We are showing this using pythons in-built libraray called as plotly.express :-






Proportion of Marital Status
Married
Single
Unknown
Divorced
46.3%
38.9%
7.4%
7.39%


Proportion of Card holders
Blue
Silver
Gold
Platinum
1.15%
0.197%
5.48%
93.2%


Proportion of Income Variation
Less than $40K
40K–
80K–
60K–
Unknown
$120K +
35.2%
17.7%
15.2%
13.8%
11%
7.18%

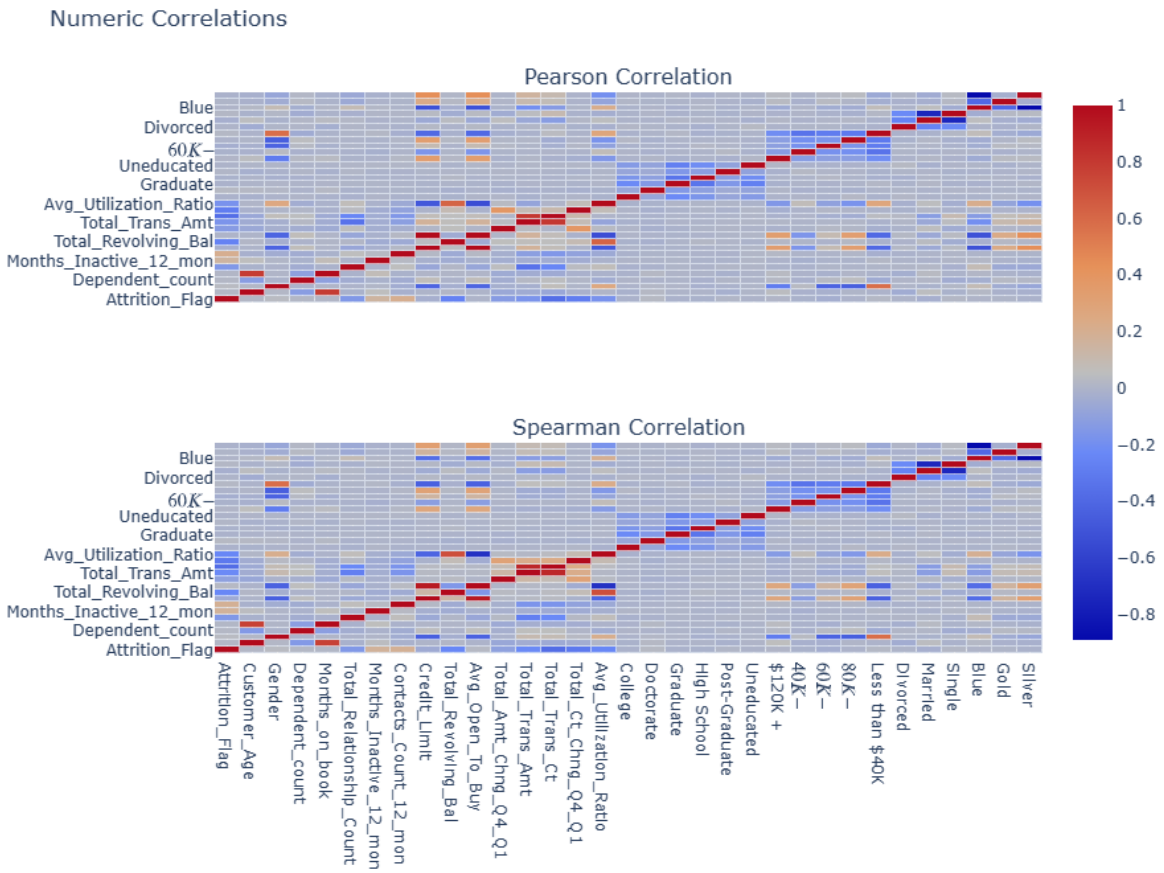
6. Encoding of categorical variables :-

Encoding categorical data is a process of converting categorical data into integer format so that the data with converted categorical values can be provided to the models to give and improve the predictions.There are different types of encoding present in machine learning. For our dataset we have used one hot encoding.

One-Hot Encoding :-

In One-Hot Encoding, each category of any categorical variable gets a new variable. It maps each category with binary numbers (0 or 1). This type of encoding is used when the data is nominal. Newly created binary features can be considered dummy variables. After one hot encoding, the number of dummy variables depends on the number of categories presented in the data.

After performing One – Hot encoding our all dataset values are now in the form of numerical datatypes.



As we know that our data imbalanced that is data is not properly distributed means our data have 16.1% Attried Customers and 83.9% Existing Customers . It will be very hard to train our machine learning model with this type of data because the data is not equally distributed we perform Data Upsampling on data using SMOTE(SMOTE stands for **Synthetic Minority Over-Sampling Technique**. SMOTE is performing the same basic task as basic resampling (creating new data points for the minority class) but instead of simply duplicating observations, it creates new observations along the lines of a randomly chosen point and its nearest neighbors.

Now our dataset is balanced lets see it :-

3.3Model Building:-

Principal Component Analysis of One Hot Encoded Data :-

We will use principal component analysis to reduce the dimensionality of the one-hot encoded categorical variables losing some of the variances, but simultaneously, using a couple of principal components instead of tens of one-hot encoded features will help me construct a better model.

So there are the features which are get reduced by the PCA they are :

X_features = ['Total_Trans_Ct','PC-3','PC-1','PC-0','PC-2','Total_Relationship_Count']

Target=['Churn']

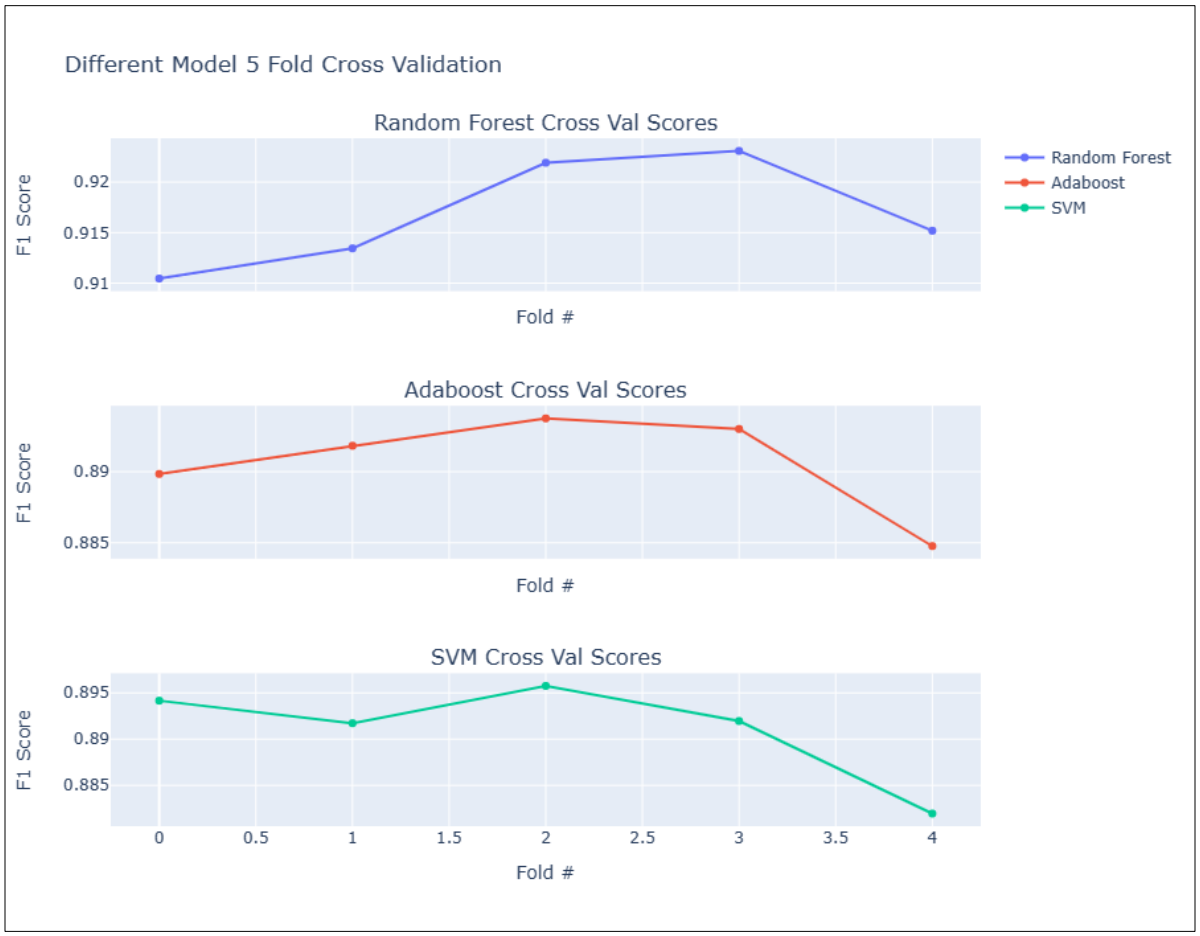
So now this are our model parameter which we have selected for model evaluation.

Here in model evaluation we have used the sklearn.model_selection.train_test_split. The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model.

It is a fast and easy procedure to perform, the results of which allow you to compare the performance of machine learning algorithms for your predictive modeling problem. Although simple to use and interpret, there are times when the procedure should not be used, such as when you have a small dataset and situations where additional configuration is required, such as when it is used for classification and the dataset is not balanced.

Cross Validation: -

Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample.The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation.



3.4 Model Evaluation :-

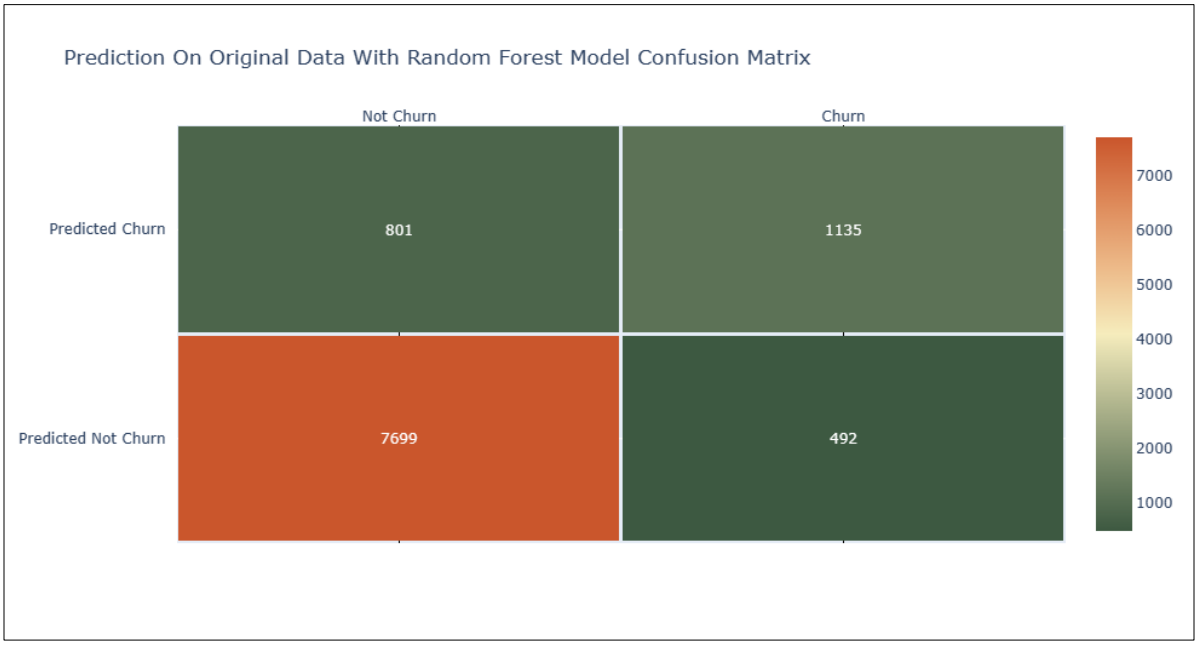
Model Evaluation is the subsidiary part of the model development process. It is the phase that is decided whether the model performs better. Therefore, it is critical to consider the model outcomes according to every possible evaluation method. Applying different methods can provide different perspectives.
Using sklearn.model_selection.train_test_split we trained our model and get the accuracy score according to various machine learning algorithm.

Model Results On Test Data

Model	F1 Score On Test Data
Random Forest	0.91
AdaBoost	0.88
SVM	0.89

As we see Random Forest algorithm gives us best accuracy. Now we will predict the on data using Confusion Matrix and Random Forest algorithm.

Confusion Matrix :-
A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix.



3.5 Model Deployment :-

Deployment is the method by which you integrate a machine learning model into an existing production environment to make practical business decisions based on data.

1. Feature selection for model deployment :-

   We have selected features using Mutual information technique for feature selection. Mutual information technique : This method combines both feature–feature mutual information and feature–class mutual information to find an optimal subset of features to minimize redundancy and to maximize relevance among features.

From above feature selection technique we get the top 5 feature are related to our target variable. They are as follows:-
1.Total_Amt_Chng_Q4_Q1

2.Total_Trans_Amt

3.Total_Trans_Ct

4.Total_Ct_Chng_Q4_Q1

5.Avg_Utilization_Ratio

2. Model Selection :-

   As we have selected 5 important features so now we will split the data into test and train using sklearn.model_selection.train_test_split , using Random Forest algorithm

   Now our model is ready to predict so now for deployment we have to pack our model using pickle .

   Pickle file - Pickle is a module in Python used for serializing and de-serializing Python objects. This converts Python objects like lists, dictionaries, etc. into byte streams (zeroes and ones).

   Now our Pickle file is created now we are going to deploy this model using streamlit on Heroku.

   Streamlit :-

   Streamlit is an open-source python framework for building web apps for Machine Learning and Data Science. We can instantly develop web apps and deploy them easily using Streamlit. Streamlit allows you to write an app the same way you write a python code. Streamlit makes it seamless to work on the interactive loop of coding and viewing results in the web app.

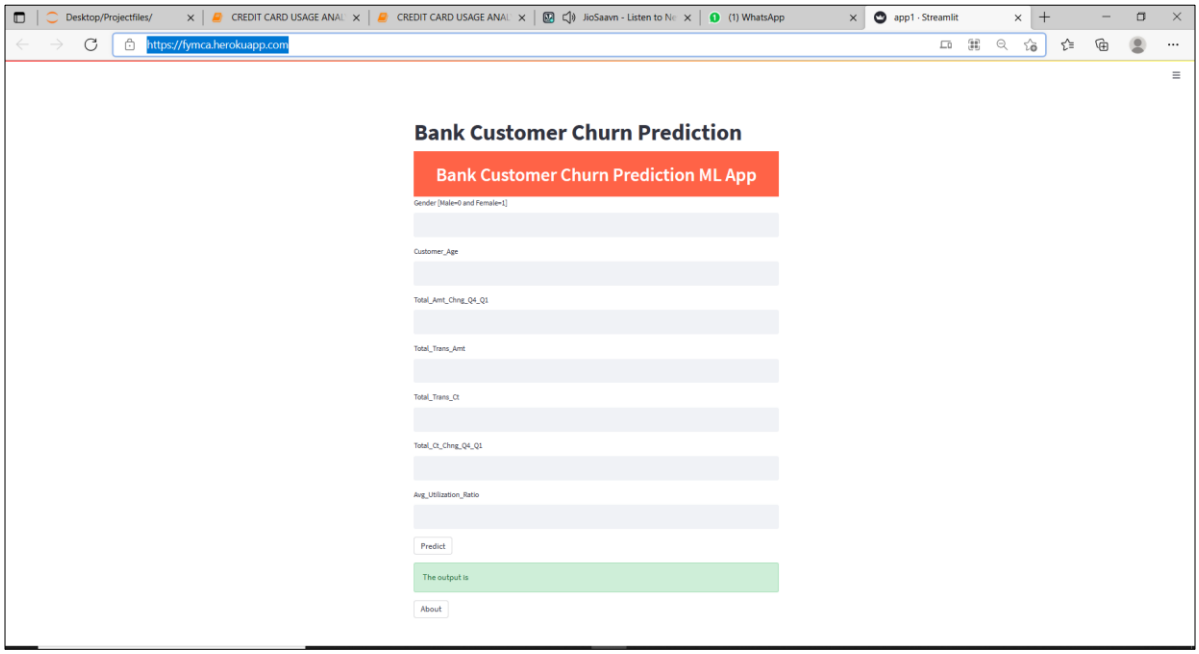
   We have created App.py file using streamlit library for designing our model and adding our pickle model in our application.

Heroku :-

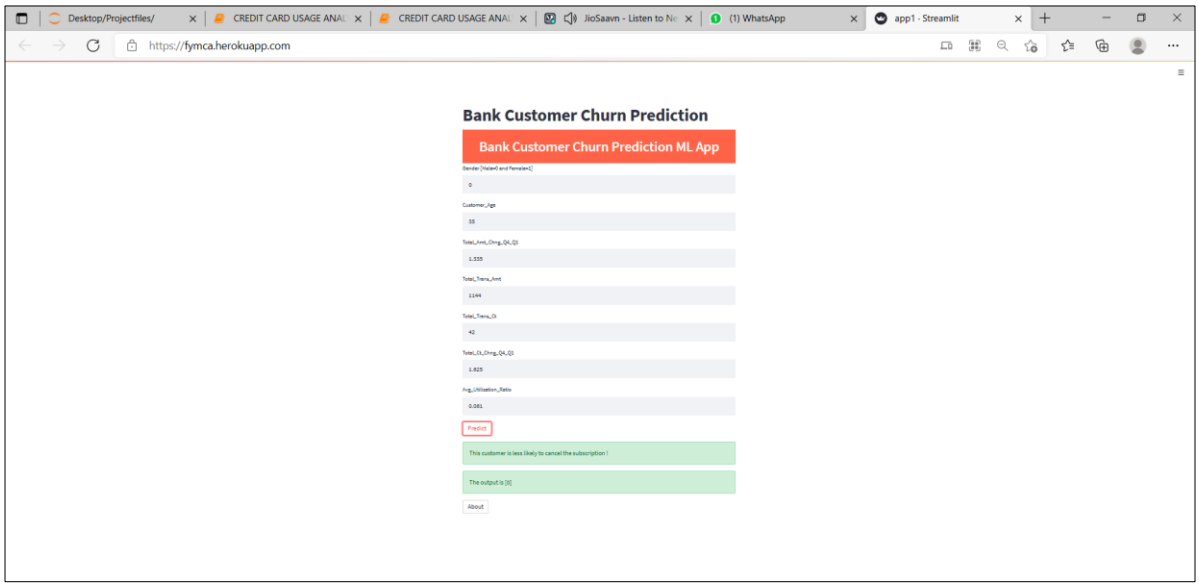
Heroku is a container-based cloud Platform as a Service (PaaS). Developers use Heroku to deploy, manage, and scale modern apps. Our platform is elegant, flexible, and easy to use, offering developers the simplest path to getting their apps to market.

4.Output Report :-

Deployed model on Heroku



Model is predicting correctly when we enter values it is shown in below snapshot



5. **Bibliography :-**

1. Towards Data Science
2. Analytics Vidhya - Learn Machine learning, artificial intelligence, business analytics, data science, big data, data visualizations tools and techniques. | Analytics Vidhya
3. Machine Learning, Data Science, Big Data, Analytics, AI - KDnuggets
4. Feature selection tutorial | Types of Feature Selection Methods (prwatech.in)
5. Machine Learning Mastery