Black Friday- Sales Prediction

Problem Description

A retail company “ABC Private Limited” wants to understand the customer purchase behavior (specifically, purchase amount) against various products of different categories. They have shared purchase summary of various customers for selected high volume products from last month.

The data set also contains customer demographics (age, gender, marital status, city type, stay in current city), product details (product id and product category) and Total purchase amount from last month. Now, they want to build a model to predict the purchase amount of customer against various products which will help them to create personalized offer for customers against different products.

Data Set

Variable	**Definition**
User_ID	User ID
Product_ID	Product ID
Gender	Sex of User
Age	Age in bins
Occupation	Occupation (Masked)
City Category	Category of the City (A,B,C)
Stay_In_Current_City_Years	Number of years stay in current city
Marital_Status	Marital Status
Product_Category_1	Product Category (Masked)
Product_Category_2	Product may belongs to other category also (Masked)
Product_Category_3	Product may belongs to other category also (Masked)
Purchase	Purchase Amount (Target Variable)

The Data Analysis task will be covered in the following steps-

1. **Data Cleaning and Preprocessing**
2. **Exploratory Data Visualization**
3. **Statistical Analysis**
4. **Building a Model**
5. **Evaluation of the model**

Tools

- **Scikitlearn:** It is a machine learning python library. It features various classification, clustering and regression algorithms. It is designed to interoperate with the python numerical and scientific libraries NumPy and SciPy.
- **Pandas:** Software library written in python for data manipulation and analysis. Data structure called data frame is used in this project for loading and generating csv files. It helps to read and write to csv files.
- **Matplotlib:** It is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It helps in plotting the graphs.
- **Seaborn:** Seaborn is a Python visualization library based on matplotlib. It provides a high level interface for drawing attractive statistical graphics.

Detailed Analysis

1. **Firstly, we imported the necessary libraries, that we will be using in the data analysis and also set the current working directory to the path where the dataset is located.**

```
# In[1]:

import os
os.chdir("F://")
os.getcwd()
# Importing the Required Libraries

# In[2]:

import pandas as pd
import numpy as np
import matplotlib as mp
import seaborn as sns
import matplotlib.pyplot as plt
```

```
from scipy import stats
get_ipython().magic('matplotlib inline')
```

2. **Now, after the libraries and current working directory is set, load the dataset into the environment using the pandas's read_csv() command.**

```
df = pd.read_csv("blackfriday.csv")
```

3. **In the next step, we performed the initial data cleaning and preprocessing to manipulate the dataset and make it clean. We also obtained the basic characteristics of the dataset by getting the shape of the dataset (550068, 12) and summary statistics and few rows of the dataset to visualize a few rows of the data.**
4. **Now, we also checked for the missing values in our dataset and we found that two columns named Product Category 2 and Product Category 3 contains missing values which needs to be treated.**

Missing Value Treatment

Why missing values treatment is required?

Missing data in the training data set can reduce the power / fit of a model or can lead to a biased model because we have not analyzed the behavior and relationship with other variables correctly. It can lead to wrong prediction or classification.

Why my data has missing values?

We looked at the importance of treatment of missing values in a dataset. Now, let's identify the reasons for occurrence of these missing values. They may occur at two stages:

Data Extraction: It is possible that there are problems with extraction process. In such cases, we should double-check for correct data with data guardians. Some hashing procedures can also be used to make sure data extraction is correct. Errors at data extraction stage are typically easy to find and can be corrected easily as well.

Data collection: These errors occur at time of data collection and are harder to correct. They can be categorized in four types:

1)**Missing completely at random**: This is a case when the probability of missing variable is same for all observations. For example: respondents of data collection process decide that they will declare their earning after tossing a fair coin. If a head occurs, respondent declares his / her earnings & vice versa. Here each observation has equal chance of missing value.

2)**Missing at random**: This is a case when variable is missing at random and missing ratio varies for different values / level of other input variables. For example: We are collecting data for age and female has higher missing value compare to male.

3)**Missing that depends on unobserved predictors**: This is a case when the missing values are not random and are related to the unobserved input variable. For example: In a medical study, if a particular diagnostic causes discomfort, then there is higher chance of drop out from the study. This missing value is not at random unless we have included "discomfort" as an input variable for all patients.

4)**Missing that depends on the missing value itself**: This is a case when the probability of missing value is directly correlated with missing value itself. For example: People with higher or lower income are likely to provide non-response to their earning.

Which are the methods to treat missing values?

Deletion: It is of two types: List Wise Deletion and Pair Wise Deletion.

In list wise deletion, we delete observations where any of the variable is missing. Simplicity is one of the major advantage of this method, but this method reduces the power of model because it reduces the sample size.

In pair wise deletion, we perform analysis with all cases in which the variables of interest are present. Advantage of this method is, it keeps as many cases available for analysis. One of the disadvantage of this method, it uses different sample size for different variables.

Deletion methods are used when the nature of missing data is "Missing completely at random" else non-random missing values can bias the model output.

Mean/ Mode/ Median Imputation: Imputation is a method to fill in the missing values with estimated ones. The objective is to employ known relationships that can be identified in the valid values of the data set to assist in estimating the missing values. Mean / Mode / Median imputation is one of the most frequently used methods. It consists of replacing the missing data for a given attribute by the mean or median (quantitative attribute) or mode (qualitative attribute) of all known values of that variable. It can be of two types: - Generalized Imputation: In this case, we calculate the mean or median for all non-missing values of that variable then replace missing value with mean or median. Like in above table, variable "Manpower" is missing so we take average of all non-missing values of "Manpower" (28.33) and then replace missing value with it.

Similar case Imputation: In this case, we calculate average for gender "Male" (29.75) and "Female" (25) individually of non-missing values then replace the missing value based on gender. For "Male", we will replace missing values of manpower with 29.75 and for "Female" with 25.

Prediction Model: Prediction model is one of the sophisticated method for handling missing data. Here, we create a predictive model to estimate values that will substitute the missing data. In this case, we divide our data set into two sets: One set with no missing values for the variable and

another one with missing values. First data set become training data set of the model while second data set with missing values is test data set and variable with missing values is treated as target variable. Next, we create a model to predict target variable based on other attributes of the training data set and populate missing values of test data set. We can use regression, ANOVA, Logistic regression and various modeling technique to perform this. There are 2 drawbacks for this approach:

1)The model estimated values are usually more well-behaved than the true values

2)If there are no relationships with attributes in the data set and the attribute with missing values, then the model will not be precise for estimating missing values.

KNN Imputation: In this method of imputation, the missing values of an attribute are imputed using the given number of attributes that are most similar to the attribute whose values are missing. The similarity of two attributes is determined using a distance function. It is also known to have certain advantage & disadvantages.

Advantages:

k-nearest neighbor can predict both qualitative & quantitative attributes Creation of predictive model for each attribute with missing data is not required Attributes with multiple missing values can be easily treated Correlation structure of the data is taken into consideration

Disadvantage:

KNN algorithm is very time-consuming in analyzing large database. It searches through all the dataset looking for the most similar instances. Choice of k-value is very critical. Higher value of k would include attributes which are significantly different from what we need whereas lower value of k implies missing out of significant attributes.

After dealing with missing values, the next task is to deal with outliers. Often, we tend to neglect outliers while building models. This is a discouraging practice. Outliers tend to make your data skewed and reduces accuracy. Let's learn more about outlier treatment.

5. After treating the missing values, we performed the Exploratory Data Visualization to visualize the dataset.

Data Visualization was performed in two cases:

a. *Univariate Analysis*
b. *Bivariate Analysis*

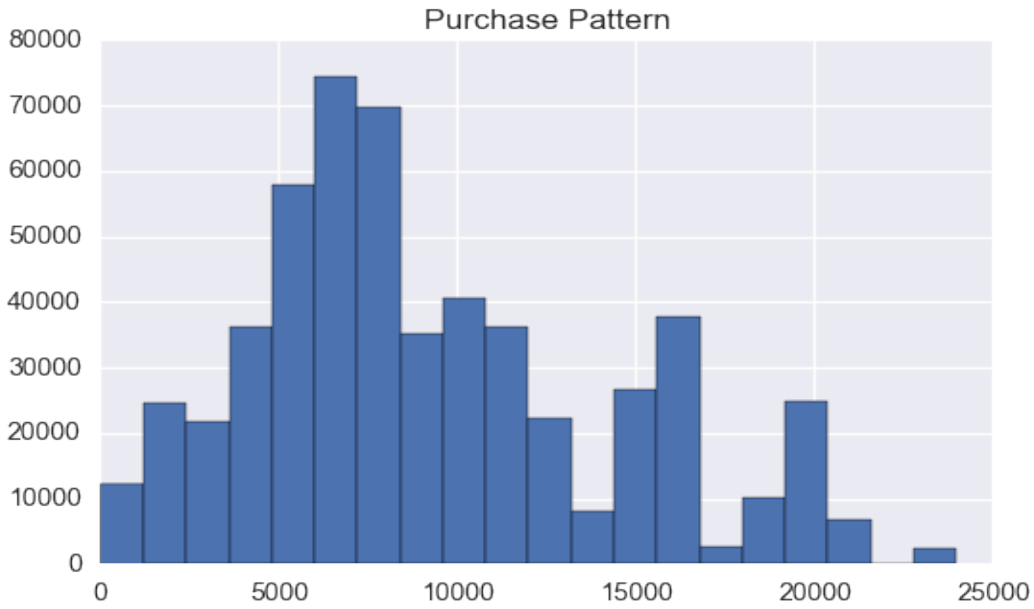
Univariate Analysis

At this stage, we explore variables one by one. Method to perform uni-variate analysis will depend on whether the variable type is categorical or continuous. Let's look at these methods and statistical measures for categorical and continuous variables individually:

Continuous Variables: In case of continuous variables, we need to understand the central tendency and spread of the variable. These are measured using various statistical metrics visualization methods like Histogram & Boxplot.

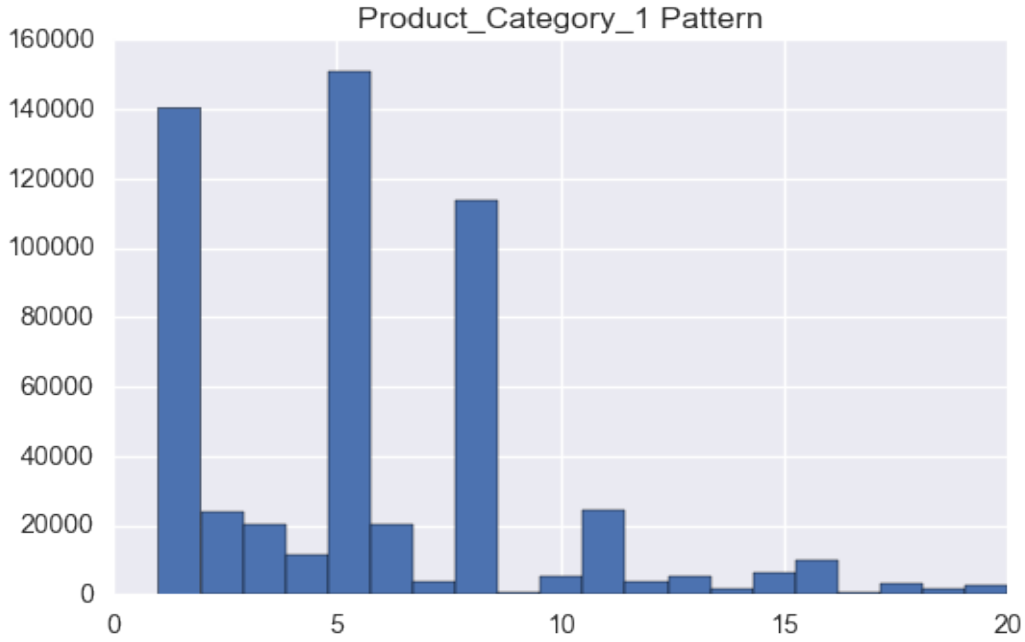
We will use Matplotlib & Seaborn package.

1. Check the distribution of dependent variable i.e. Purchase

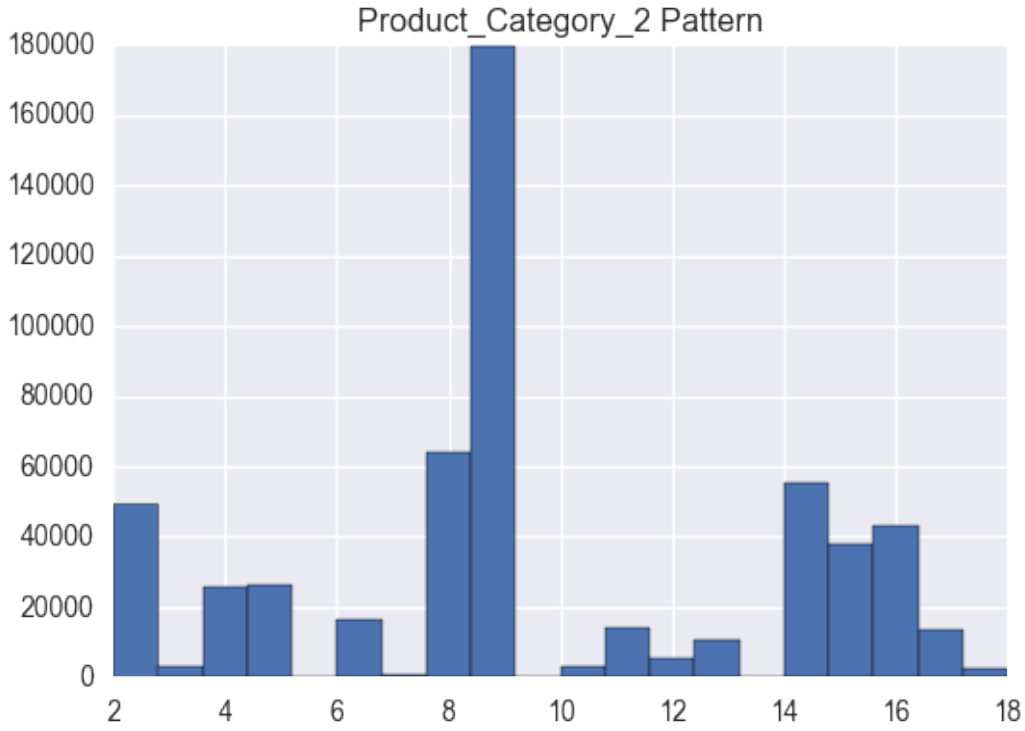


Looking at Purchase Pattern we can see that number of count is more at 6000 & 7000 and above 15000 have very less count.

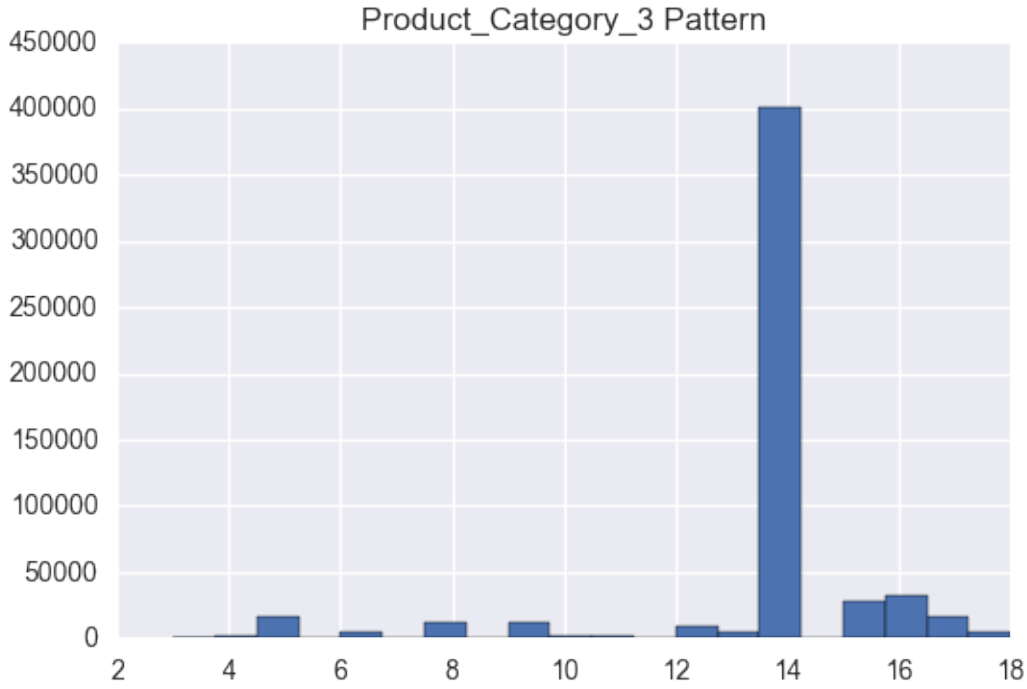
2. Check the distribution of dependent variable i.e. Product Category 1.



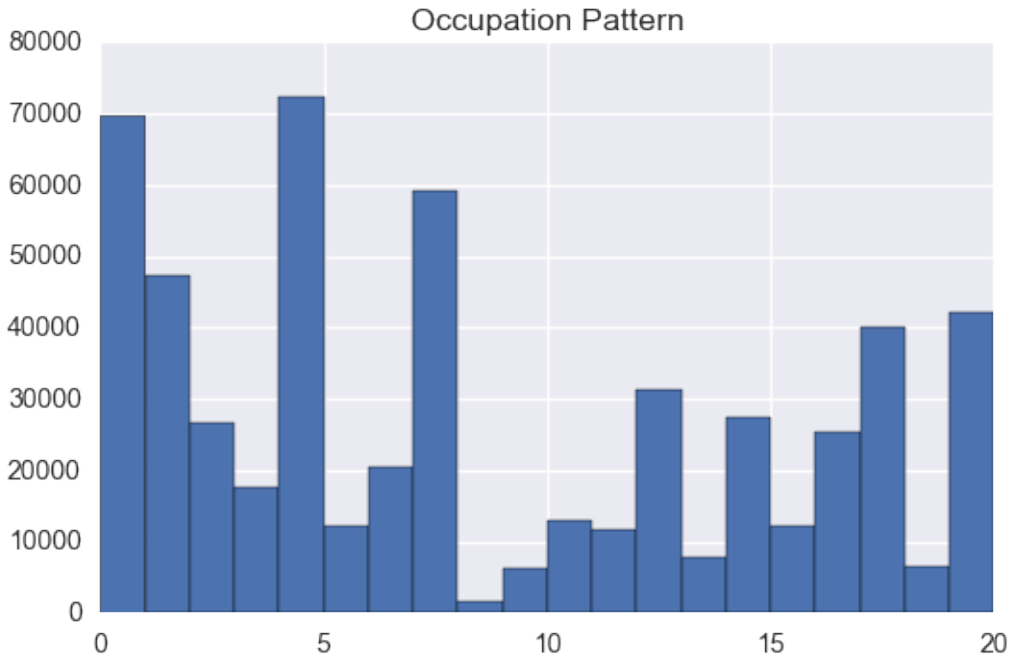
3. Check the distribution of Product_Category_2.



4. Check the distribution of Product_Category_3.



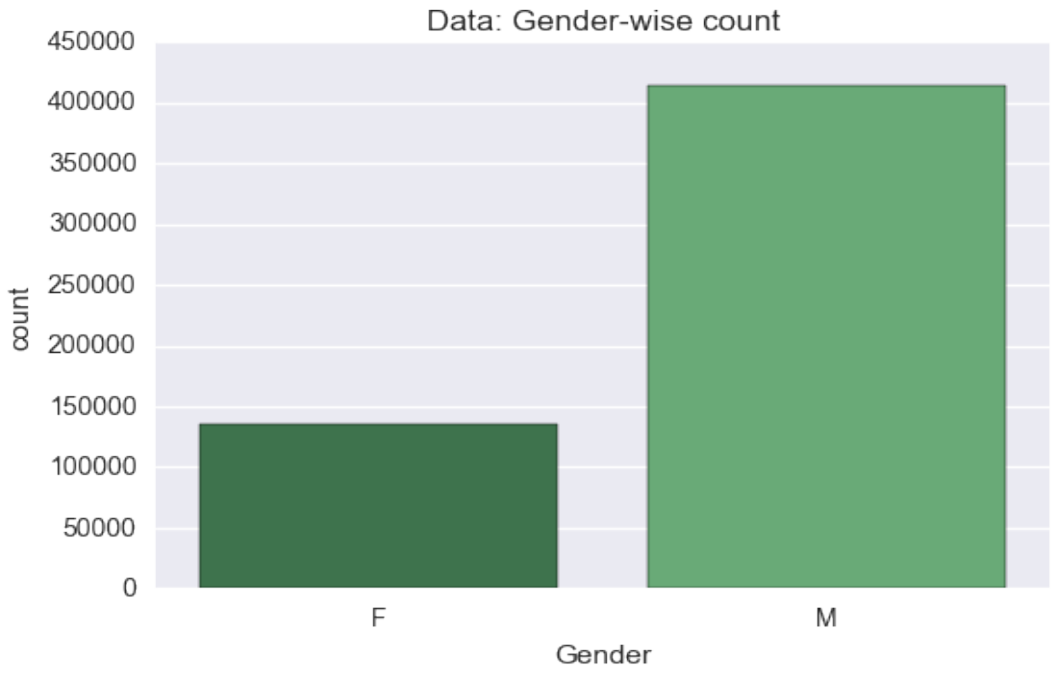
5. Check the distribution of Occupation.



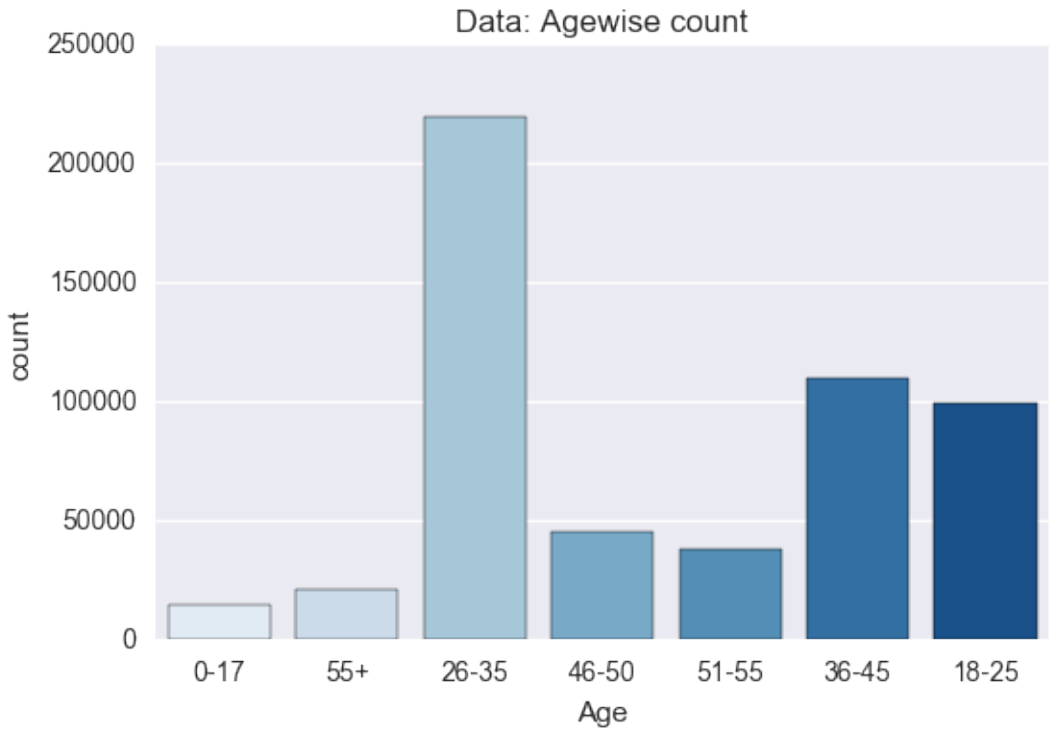
Categorical Variables: For categorical variables, we'll use frequency table to understand distribution of each category. We can also read as percentage of values under each category. It can

be measured using two metrics, Count and Count% against each category. Bar chart can be used as visualization.

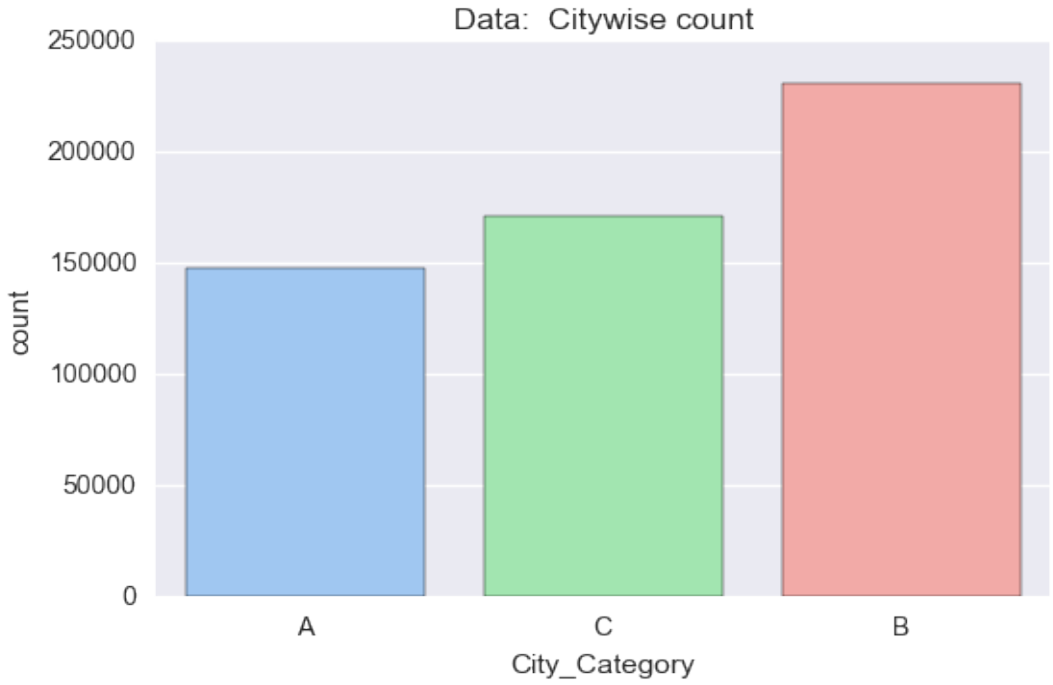
6. Number of Gender count.



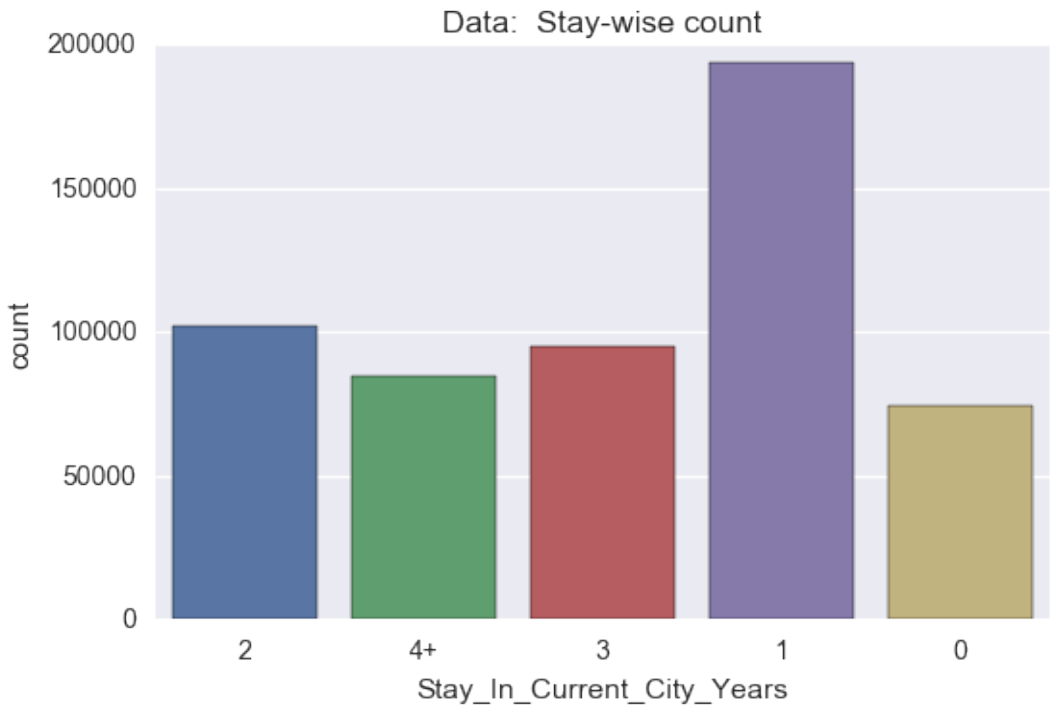
7. Number of count of Age.



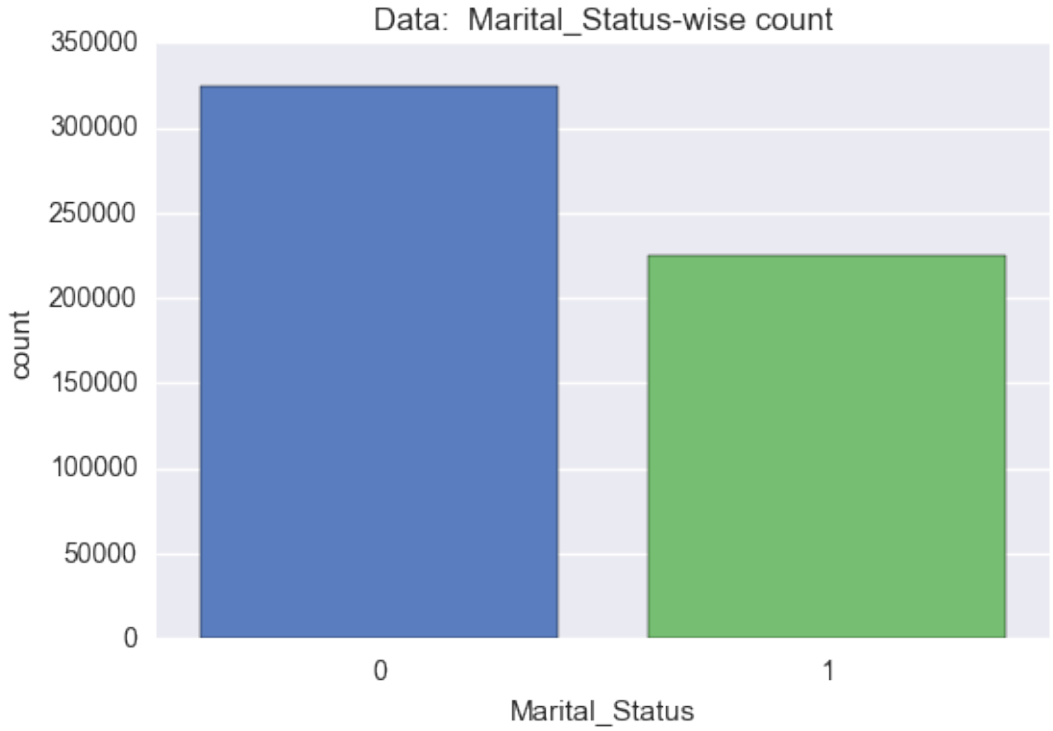
8. Number of count of City.



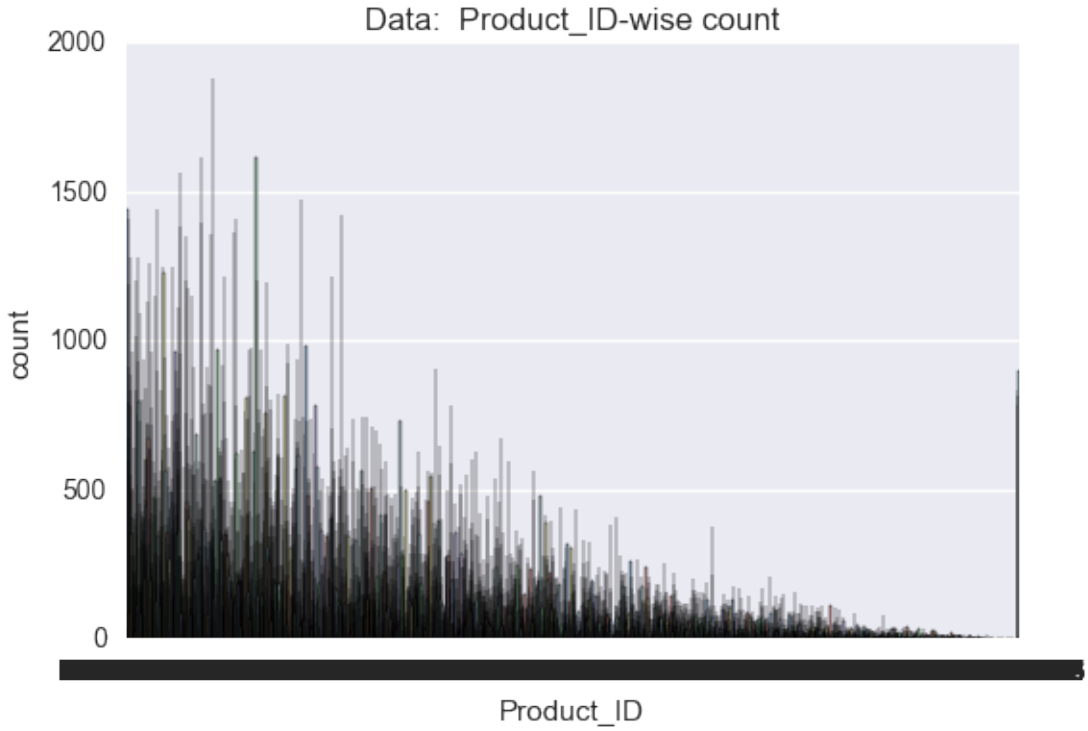
9. Number of count of stay.



10. Number of count of Marital Status.



11. Number of count of Product ID.



6. Now, we performed the Statistical Analysis to deduce the relationships among the variables.

Bi-variate Analysis

Bi-variate Analysis finds out the relationship between two variables. Here, we look for association and disassociation between variables at a pre-defined significance level. We can perform bi-variate analysis for any combination of categorical and continuous variables. The combination can be: Categorical & Categorical, Categorical & Continuous and Continuous & Continuous. Different methods are used to tackle these combinations during analysis process.

Let's understand the possible combinations in detail:

Continuous & Continuous: While doing bi-variate analysis between two continuous variables, we should look at scatter plot. It is a nifty way to find out the relationship between two variables. The pattern of scatter plot indicates the relationship between variables. The relationship can be linear or non-linear.

Categorical & Continuous: While exploring relation between categorical and continuous variables, we can draw box plots for each level of categorical variables. If levels are small in number, it will not show the statistical significance. To look at the statistical significance, we can perform Z-test, T-test or ANOVA.

1) Z-Test/ T-Test: - Either test assess whether mean of two groups are statistically different from each other or not.

If the probability of Z is small, then the difference of two averages is more significant. The T-test is very similar to Z-test but it is used when number of observation for both categories is less than 30.

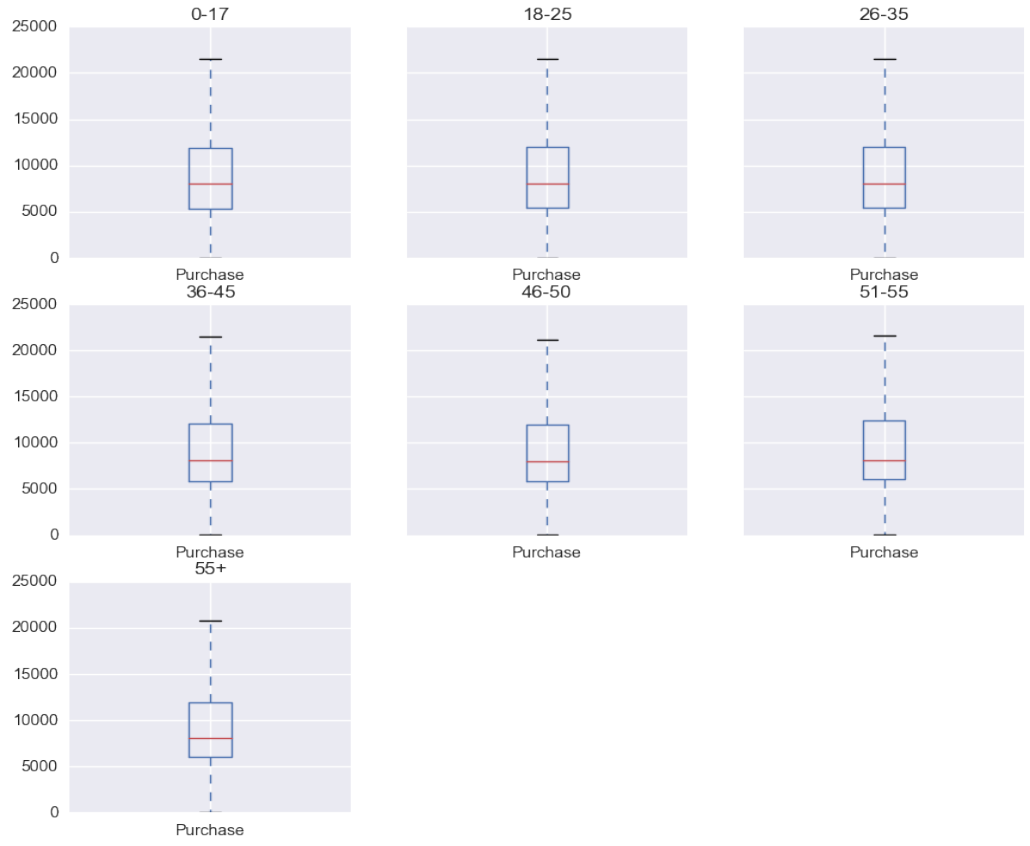
2) ANOVA: - It assesses whether the average of more than two groups is statistically different.

Example: Suppose, we want to test the effect of five different exercises. For this, we recruit 20 men and assign one type of exercise to 4 men (5 groups). Their weights are recorded after a few weeks. We need to find out whether the effect of these exercises on them is significantly different or not. This can be done by comparing the weights of the 5 groups of 4 men each.

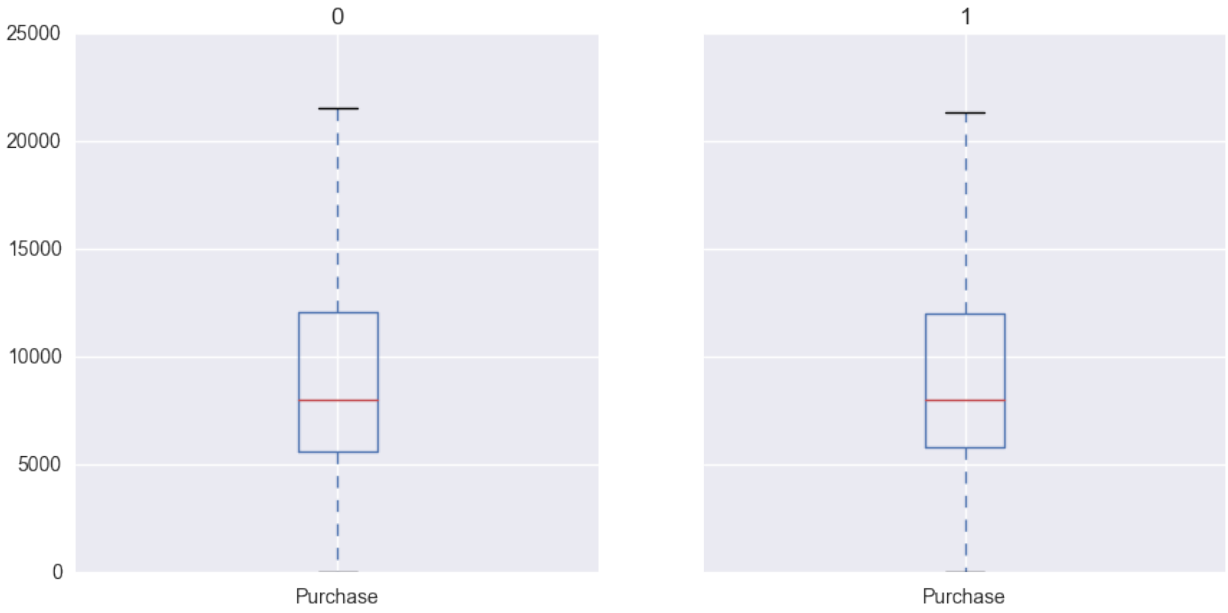
12. In this plot we plotted Purchase with gender and performed the t-test with hypothesis to check their significance.

By doing Hypothesis we can say that they have mean purchase difference i.e. they are significant.

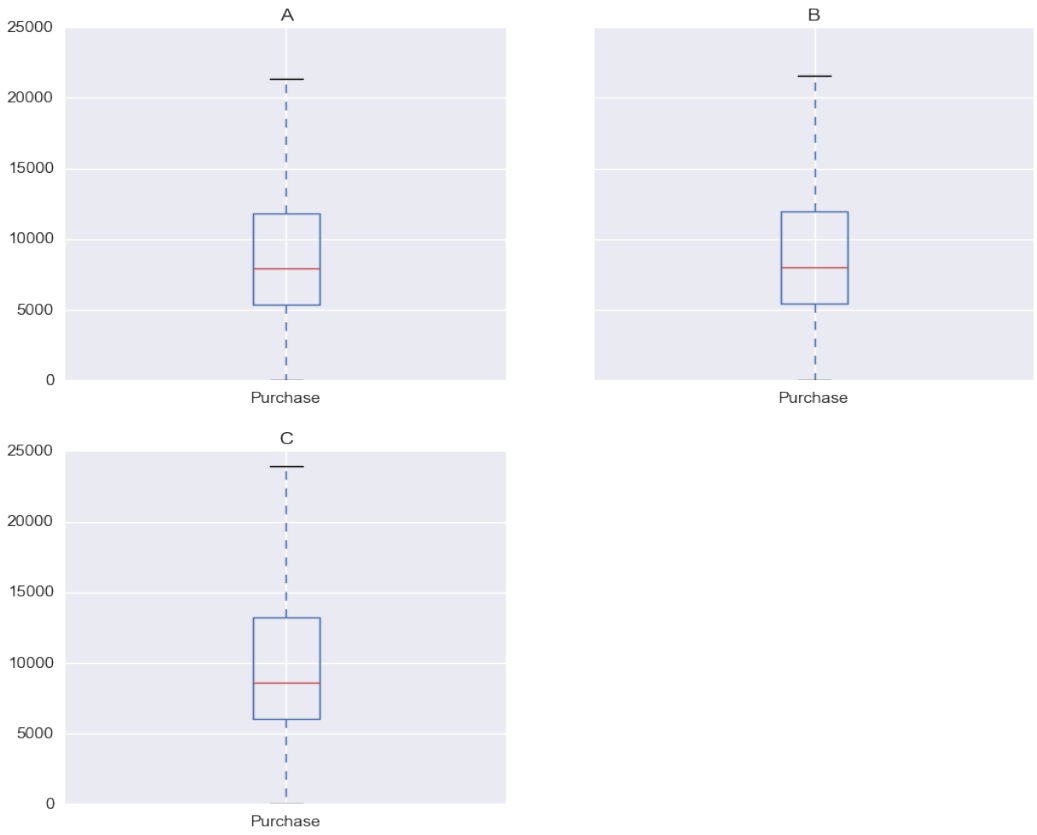
13. Similarly, we performed hypothesis testing for Age with Purchase.



14. Purchase and Marital Status.



15. Purchase and City Category.



And, with all these results and hypothesis, we reached the conclusion that all variables are significant in determining the Purchase label.

Categorical & Categorical: To find the relationship between two categorical variables, we can use following methods:

Two-way table: We can start analyzing the relationship by creating a two-way table of count and count%. The rows represent the category of one variable and the columns represent the categories of the other variable. We show count or count% of observations available in each combination of row and column categories.

Stacked Column Chart: This method is more of a visual form of Two-way table.

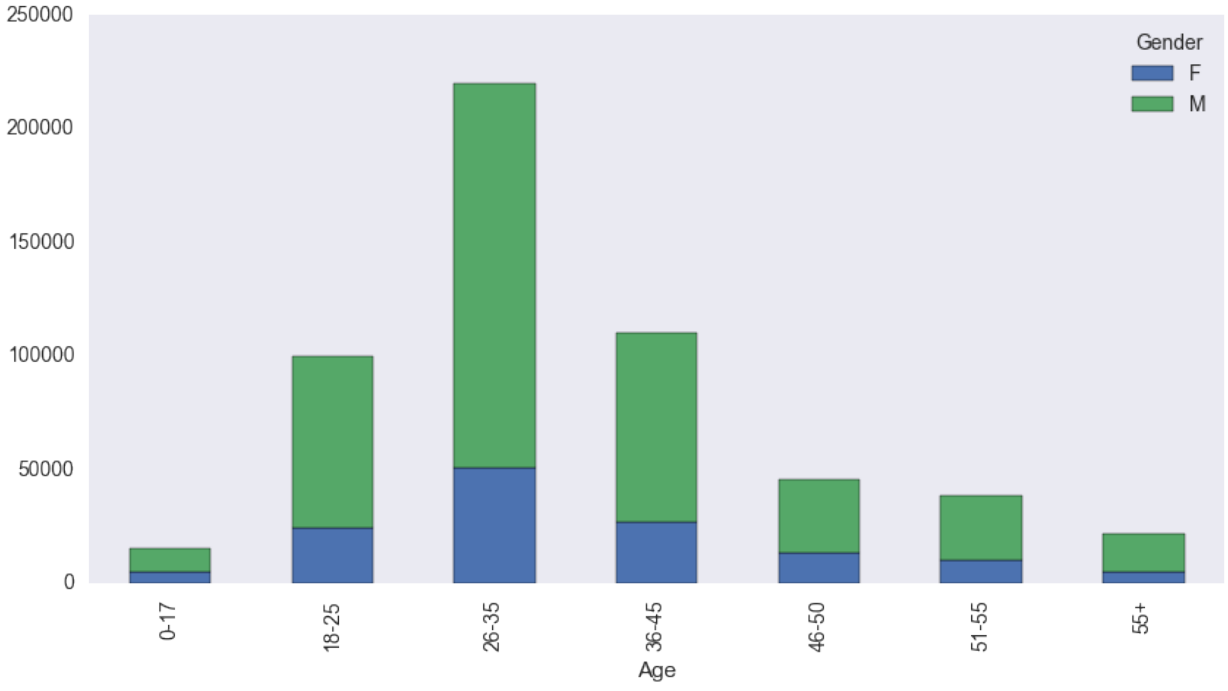
Chi-Square Test: This test is used to derive the statistical significance of relationship between the variables. Also, it tests whether the evidence in the sample is strong enough to generalize that the relationship for a larger population as well. Chi-square is based on the difference between the expected and observed frequencies in one or more categories in the two-way table. It returns probability for the computed chi-square distribution with the degree of freedom.

Probability of 0: It indicates that both categorical variable is dependent

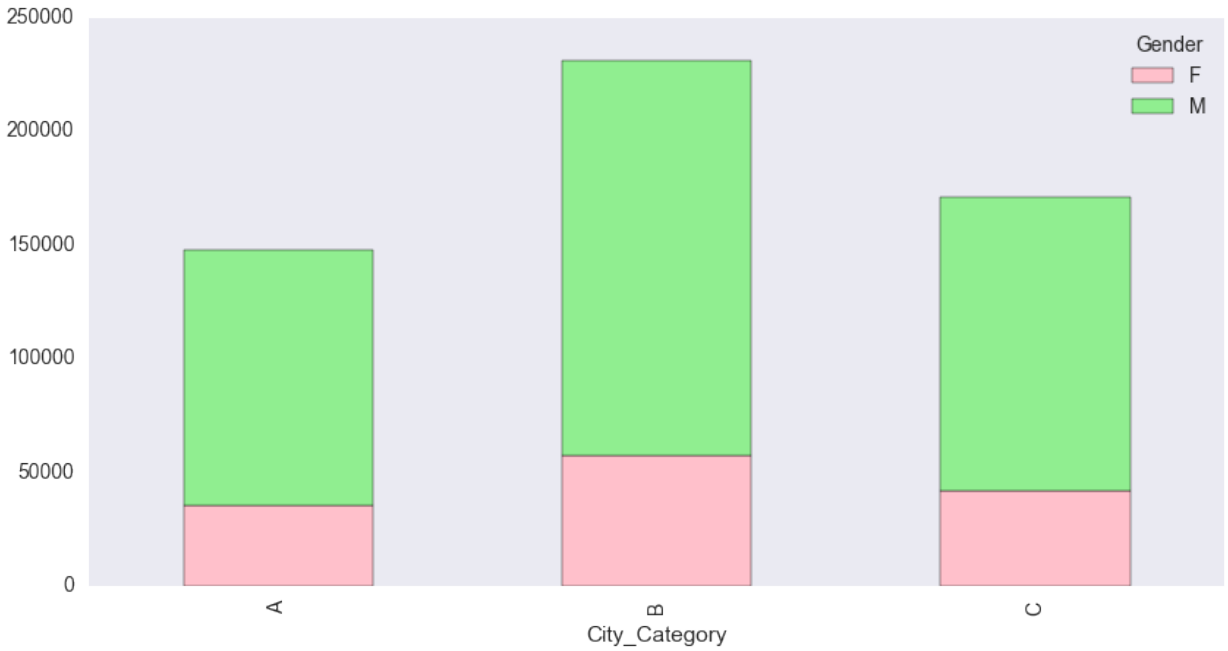
Probability of 1: It shows that both variables are independent.

Probability less than 0.05: It indicates that the relationship between the variables is significant at 95% confidence.

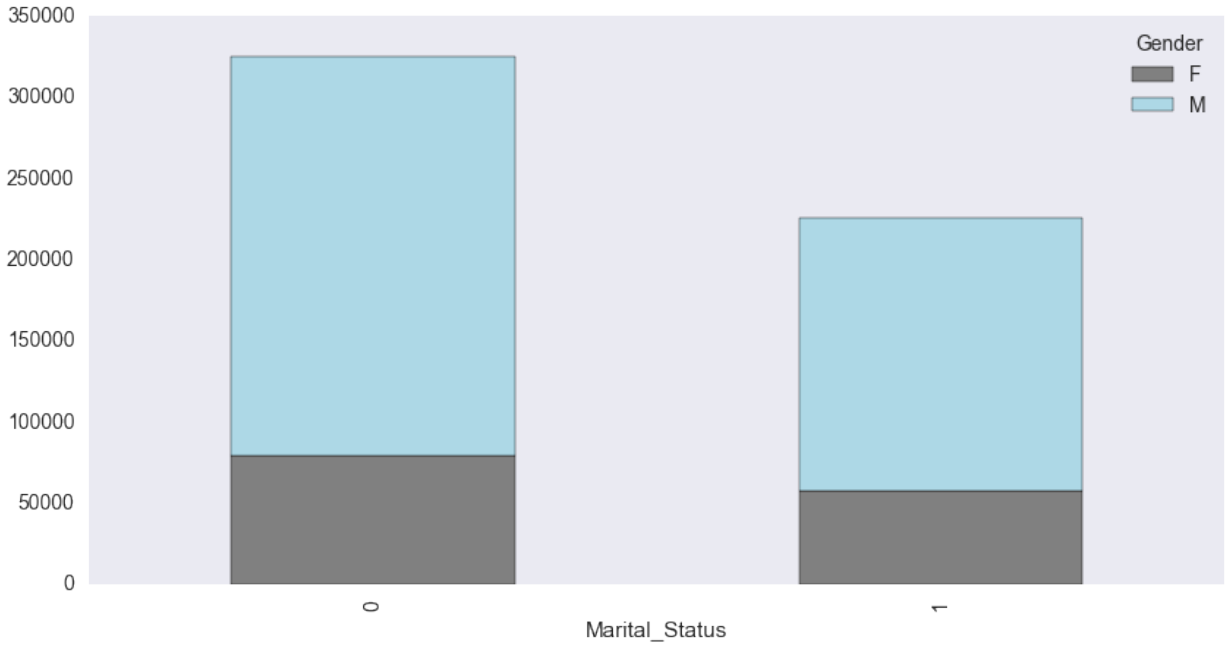
16. Age and Gender.



17. City Category and Gender.



18. Marital Status and Gender.



We used the Chi Squared Test on all these three categorical variable relations and found that these are all related or dependent on each other.

7. In the step, we performed Variable Transformation.

Variable Transformation

What are the common methods of Variable Transformation?

There are various methods used to transform variables. As discussed, some of them include square root, cube root, logarithmic, binning, reciprocal and many others. Let's look at these methods in detail by highlighting the pros and cons of these transformation methods.

Logarithm: Log of a variable is a common transformation method used to change the shape of distribution of the variable on a distribution plot. It is generally used for reducing right skewness of variables. Though, it can't be applied to zero or negative values as well.

Square / Cube root: The square and cube root of a variable has a sound effect on variable distribution. However, it is not as significant as logarithmic transformation. Cube root has its own advantage. It can be applied to negative values including zero. Square root can be applied to positive values including zero.

Binning: It is used to categorize variables. It is performed on original values, percentile or frequency. Decision of categorization technique is based on business understanding. For example, we can categorize income in three categories, namely: High, Average and Low. We can also perform co-variate binning which depends on the value of more than one variables.

8. We used the label encoding and Dummy Variable Characteristics on the dataset. Now, the dataset contains these additional columns as Dummy features.

Additional Columns:

- City_Category_0
- City_Category_1
- City_Category_2
- Age_0
- Age_1
- Age_2
- Age_3
- Age_4
- Age_5
- Age_6
- Stay_In_Current_City_Years_0
- Stay_In_Current_City_Years_1
- Stay_In_Current_City_Years_2
- Stay_In_Current_City_Years_3
- Stay_In_Current_City_Years_4

9. **In the next step we performed the splitting to convert the data into train and test dataset with test size to 0.3.**

10. **The next step was to build the model on the train dataset and evaluate the model using the test dataset based on the RMSE (root mean squared error), which must be minimized to get the best and accurate model.**

11. **We used three different Regression Model;**

a. *Linear Regression Model*
b. *Decision Tree Regression Model*
c. *Radom Forest Regression Model*

- *The RMSE in case of Linear Regression Model was 4679.21*
- *The RMSE in case of Decision Tree Regression Model was 2907.40*
- *The RMSE in case of Random Forest Regression Model was 2895.84*

So, here we can conclude that, Random Forest Regression Model provided the minimum RMSE a nd hence it is the best model.