

**E-Retail Customer Activation and Retention**

Submitted by:

Amritesh Kumar

**ACKNOWLEDGMENT**

I would like to express my deepest gratitude to my SME (Subject Matter Expert) Shubam Yadav as well as Flip Robo

Technologies who gave me the opportunity to do this project on **‘E-Retail Customer Activation and Retention’** & also helping me to gain in-depth knowledge of Machine Learning and DataScience to derive insights for organizational goals or meet business needs.

Also, I have utilized a few external resources that helped me to complete this project. All the external resources that were used in creating this project are listed below:

<https://stackoverflow.com/questions>

<https://medium.com/>

<https://www.kaggle.com/>

<https://www.geeksforgeeks.org/>

<https://www.codegrepper.com/>

<https://www.analyticsvidhya.com/>

<https://towardsdatascience.com/>

<https://github.com/>

**INTRODUCTION**

Business Problem Framing

**Problem Overview**

E-retail Customer Activation and Retention

E-retail factors for customer activation and retention: A case study from Indian e-commerce customers. Customer satisfaction has emerged as one of the most important factors that guarantee the success of online store; it has been posited as a key stimulant of purchase, repurchase intentions and customer loyalty. A comprehensive review of the literature, theories and models have been carried out to propose the models for customer activation and customer retention. Five major factors that contributed to the success of an e-commerce store have been identified as: service quality, system quality, information quality, trust and net benefit. The research furthermore investigated the factors that influence the online customers repeat purchase intention. The combination of both utilitarian value and hedonistic values are needed to affect the repeat purchase intention (loyalty) positively. The data is collected from the Indian online shoppers. Results indicate the e-retail success factors, which are very much critical for customer satisfaction.

Prediction:

We have to predict E-retail factors for customer activation and retention

Conceptual Background of the Domain Problem

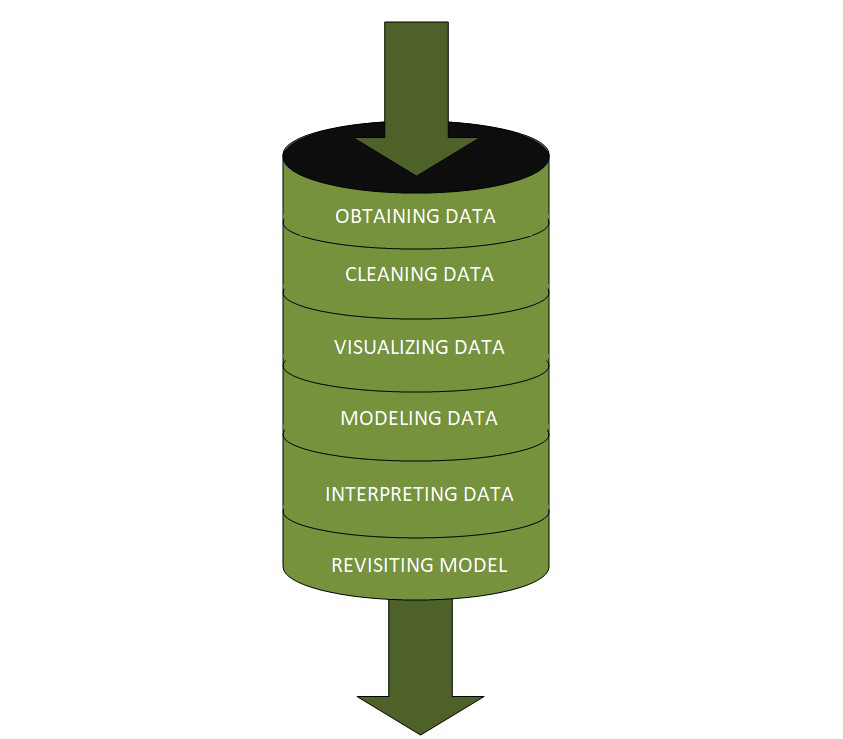
**MACHINE LEARNING AND DATA SCIENCE FOR BUSINESS:**

Machine learning is a branch of [artificial intelligence (AI)](https://www.ibm.com/cloud/learn/what-is-artificial-intelligence) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn from experience, make predictions and gradually improving its accuracy. It is an important component of the growing field of data science. Through the use of statistical methods, algorithms are trained to make classifications or predictions, uncovering key insights within data mining projects. These insights subsequently drive decision making within applications and businesses, ideally impacting key growth metrics. As big data continues to expand and grow, the market demand for data science will increase, requires to assist in the identification of the most relevant business questions and subsequently the data to answer them. Following are the ways Data science can add value to Business :

* Empowering management and officers to make better decision
* Directing actions based on trends—which in turn help to define goals
* Challenging the staff to adopt best practices and focus on issues that matter
* Identifying opportunities
* Decision making with quantifiable, data-driven evidence
* Testing these decisions
* Identification and refining of target audiences

**DATASCIENCE PIPELINE:**

The data science pipeline is a collection of connected tasks that aims at delivering an insightful data science product or service to the business organization. The responsibilities include collecting, cleaning, exploring, modeling, interpreting the data, and other processes of the launching of the product. This final product can be used for to achieve Business Goals.



**Exploratory Data Analysis:**

The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

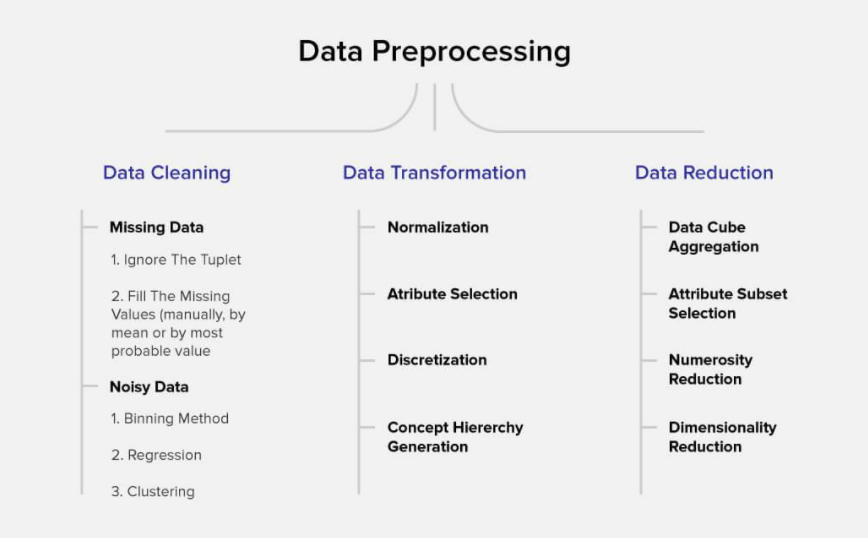
Data scientists can use exploratory analysis to ensure the results they produce are valid and applicable to any desired business outcomes and goals. EDA also helps stakeholders by confirming they are asking the right questions

**TYPES OF EXPLORATORY DATA ANALYSIS:**

* Univariate Non-graphical
* Multivariate Non-graphical
* Univariate graphical
* Multivariate graphical

**DATA PRE-PROCESSING & FEATURE ENGINEERING:**

Preprocessing simply refers to perform series of operations to transform or change data. It is transformation applied to our data before feeding it to algorithm. When creating a machine learning project, and doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data preprocessing task.



Data pre-processing is a very vital input to machine learning models, It is to prepare the raw data & make it suitable for efficient machine learning model. These are the methods of data preprocessing and we are going to use the required ones in our project.

**FEATURE ENGINEERING:**

Feature engineering is the process of selecting, manipulating, and transforming raw data into features that can be used in supervised learning. In order to make machine learning work well on new tasks, it might be necessary to design and train better features. As you may know, a “feature” is any measurable input that can be used in a predictive model.

Feature engineering**, in simple terms, is the act of converting raw observations into desired features using statistical or machine learning approaches.** It can produce new features for both supervised and unsupervised learning, with the goal of**simplifying and speeding up data transformations**while also**enhancing model accuracy.**

**Feature Engineering Techniques for Machine Learning**

* **Imputation**
* **Handling Outliers**
* **Log Transform**
* **One-hot encoding/Label Encoding**
* **Scaling**

**Data Transformation:**

**Label Encoding:**

**As we mentioned above in library installation,** Label Encoder is used to encode labels by assigning them numbers. It is used to encode single or multiple columns. Thus, if the feature is color with values such as [‘white’, ‘red’, ‘black’, ‘blue’]., using Label Encoder may encode color string label as [0, 1, 2, 3]

**Handling Outliers:**

The most important phase in Feature Engineering is handling outliers because it ensures that our model is trained on accurate data which leads to accurate models. An outlier may occur due to the variability in the data. It may indicate an experimental error or heavy skewness in the data(heavy-tailed distribution). We have three measures of central tendency namely Mean, Median, and Mode. They help us describe the data.

Below are some of the techniques of detecting outliers

* Boxplots
* Z-score

## Variance Inflation Factor (VIF)

Variance Inflation Factors (VIFs) measure the correlation among independent variables in least squares regression models. Statisticians refer to this type of correlation as multicollinearity. Excessive multicollinearity can cause problems for regression models. The stats models package has VIF library, Let us import the package.

**SKEWNESS REMOVAL-(POWER-TRANSFORM):**

Key step prior to initiating Machine learning models, optimizing, scaling the data to provide it as a input to start the modelling.

A power transform will make the probability distribution of a variable more Gaussian. This is often described as removing a skew in the distribution, although more generally is described as stabilizing the variance of the distribution. The log transform is a specific example of a family of transformations known as power transforms. The power\_transform library present in the Sklearn. Pre-processing package.

**MINMAX SCALER:**

MinMax Scaler shrinks the data within the given range, usually of 0 to 1. It transforms data by scaling features to a given range. It scales the values to a specific value range without changing the shape of the original distribution.

Before scaling we have to train test split the data.since we have to do skewness removal and scaling only on input data.

**TRAIN TEST SPLIT:**

The scikit-learn Python machine learning library provides an implementation of the train-test split evaluation procedure via the train\_test\_split() function. The function takes a loaded dataset as input and returns the dataset split into two subsets.train\_test\_split() will split arrays data into random subsets. The ideal split is said to be 80:20 for training and testing.

Review of Literature

**ABSTRACT:**

E-retail has become the need of the hour for the modern customers nowadays. This project focuses on the key factors for customer activation & retention. Given dataset needs to be analysed in order to understand the things gone wrong / right to formulate a strategy / layout key points towards the customer activation & retention/

Motivation for the Problem Undertaken

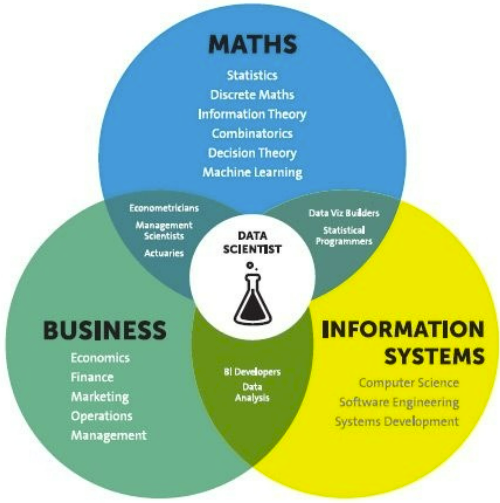
**Business Goal:**

We are required to analyse the India e-retail industry survey response dataset by bifurcating them into hedonic & Utilitarian values. Understand various influential factors customer encounters during online purchase done on a e-retail platform. Understand the customer’s choice across various platform & layout the key indices which makes them to make repeated purchases on an certain e-retail platform.

**Analytical Problem Framing**

Mathematical/ Statistical /Analytical Modeling of the Problem

Mathematics, Statistics and Analytics are three of the most important concepts of Data Science. Data Science revolves around these three fields and draws their concepts to operate on the data.we will explore its practical usages in this field.  So let’s first explore how much these three are required for data science.



**Mathematical Modelling**

Mathematical models are important, selecting the right one to answer the business question can bring tremendous value to the organization. Machine Learning is a field that focuses on computers having the ability to learn/operate without being programmed to do so.

Mathematics is playing an essential role in the latest technologies like Machine Learning, Artificial Intelligence, Data Science and Deep Learning, etc., It is because every algorithm built in the latest technologies has a mathematical function behind it and aid in identifying patterns.

The understanding of various notions of Statistics and Probability Theory are key for the implementation of such algorithms in data science. Notions include: Regression, Maximum Likelihood Estimation, the understanding of distributions (Binomial, Bernoulli, Gaussian (Normal)) and Bayes’ Theorem.

The main reason for a greater significance of mathematics is because of its various concepts like: –

· Linear Algebra

· Probability

· Calculus

· Statistics

**Linear Algebra & Calculus**

Deep learning requires us to understand linear algebra & calculus, to understand how it works, for example forward propagation, backward propagation, parameters setting etc. For linear algebra, there are matrix operations (plus, minus, times, divide), scalar product, dot product, eigen-vectors and eigenvalues.

It is a branch of  Mathematics for studying systems of equations. it can be one, two, and multi-dimensional equations. it helps us to solve numerical data or relations between two or more variables by establishing relations or equations between them. for example,

here' one basic algebraic equation:

    y = a + bx + cx2

linear-algebra has a wide range of applications such as statics and matrices calculations, linear regression equations, descriptive statistics, graphic image vectors, Fourier series, graphs, and network establishment.

machine-learning algorithms like linear regression, logistic regression uses linear algebra to solve our target variables with given inputs/attributes or feature vectors given in the data set.

**Calculus**

 Calculus is used essentially in optimization techniques. Using calculus, you can carry out mathematical modeling of artificial neural networks and also increase their accuracy and performance. For calculus, the data scientist need to understand various differentiation (to second-order derivative), integration, partial differentiation.

**Differential Calculus**

  Differential Calculus studies the rate at which the quantities change. Derivates are most widely used for finding the maxima and minima of the functions. Derivates are used in optimization techniques where we have to find the minima in order to minimize the error function.

**Integral Calculus**

It is the mathematical study of the accumulation of quantities and for finding the area under the curve. Integrals are further divided into definite integrals and indefinite integrals.

**Probability**

The probability theory is very much helpful for making the prediction and Estimation.With the help of statistical methods, we make estimates for the further analysis. Thus, statistical methods are largely dependent on the theory of probability.

Probability is a very important mathematical concept for data science, used in validating hypothesis, bayes theorem and interpreting outputs in machine learning.

Bases on these we try to estimate various events, and the likelihood of the outcome. sometimes we wat graphical representations of probable outcomes which we call probability density functions or density curves.

Concepts of probability help us estimate expected value from given variables, to solve confusion matrix in classification algorithms, information entropy, evidence of particular attributes in naive Bayes classification, and even in statistics for hypothesis testings.

**Statistics**

A statistical model is a mathematical representation (or mathematical model) of observed data. When data analysts apply various statistical models to the data they are investigating, they are able to understand and interpret the information more strategically.

So the areas in statistics are simple statistics like measurement of centrality, distributions and different probability distributions (Weibull, Poisson etc), Baye’s Theorem

statistics is divided into two –

* Descriptive Statistics
* Inferential Statistics

#### **Descriptive Statistics**

Descriptive Statistics or summary statistics is used for describing the data. It deals with the quantitative summarization of data. This summarization is performed through graphs or numerical representations.

### Descriptive Statistics:

1) Mean, Median, Mode

2) IQR, percentiles

3) Std deviation and Variance

4) Normal Distribution

5) Z-statistics and T-statistics

6) correlation and linear regression

**Inferential Statistics**

It is the procedure of inferring or concluding from the data. Through inferential statistics, we make a conclusion about the larger population by running several tests and deductions from the smaller sample.

### Inferential Statistics:

1) Sampling distributions

2) confidence interval

3) chi-square test

4) Advanced regression

5) ANOVA

The mathematical concepts noted above are key in understanding/implementing the following Machine Learning techniques.

* Supervised learning, including regression and classification models.
* Unsupervised learning, including clustering algorithms and association rules.

### **Regression Models**

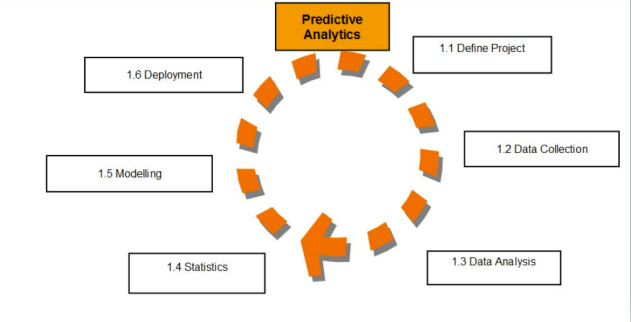
Data analysts use **regression models** to examine relationships between variables. Regression models are often used by organizations to determine which independent variables hold the most influence over dependent variables—information that can be leveraged to make essential [business decisions](https://www.northeastern.edu/graduate/blog/data-driven-decision-making/).

### **Classification Models**

**Classification** is a process in which an algorithm is used to analyze an existing data set of known points. The understanding achieved through that analysis is then leveraged as a means of appropriately classifying the data. Classification is a form of machine learning that can be particularly helpful in analyzing very large, complex sets of data to help make more accurate predictions.

**Analytical Models:**

An analytical model estimates or classifies data values by essentially drawing a line through data points. When applied to new data or records, a model can predict outcomes based on historical patterns.



.  An analytical model is quantitative in nature, and used to answer a specific question or make a specific design decision. Different analytical models are used to address different aspects of the system, such as its performance, reliability, or mass properties.Data analysis comes with the fundamental types of data analytics encounter in data science: Descriptive, Diagnostic, Predictive, and Prescriptive.

* Descriptive analytics is a statistical method that is used to search and summarize historical data in order to identify patterns or meaning.
* Descriptive analysis is often used when reviewing any past or present data. This is because raw data is difficult to consume and interpret, while the metrics offered by descriptive analysis are much more focused.
* The example of descriptive statistics or analytics is to calculate the mean, median mode, standard deviation, and similar kinds of statistical calculation on finance or sales data.
* Diagnostic analytics takes it a step further to uncover the reasoning behind certain results. Diagnostic analytics is usually performed using such techniques as data discovery, drill-down, data mining, and different type of bivariant data analysis like  correlations.etc.,
* Predictive Analytics is a **statistical method that utilizes algorithms and machine learning to identify trends in data and predict future behaviors**. Predictive Analytics can take both past and current data and offer predictions of what could happen in the future.
* Predictive models typically utilize variability in data to make the correct prediction and more variability of ingredient data that shows the relationship with what is possible to predict that united together into a prediction or valid score.
* Prescriptive analytics automatically synthesizes big data, mathematical sciences, business rules, algorithms, and machine learning to make predictions and then suggests decision options to take advantage of the predictions. Prescriptive means (optimization and simulation).

Data Sources and their formats

**Technical Requirements**:

• Data contains 269 entries each having 71 variables.

• Data set doesn’t contains Null values. We treated them using the domain knowledge and our own understanding.

• Extensive EDA has been performed to gain relationships of important variable and price.

• Data contains one numerical and all others as categorical variable. We handled them accordingly.

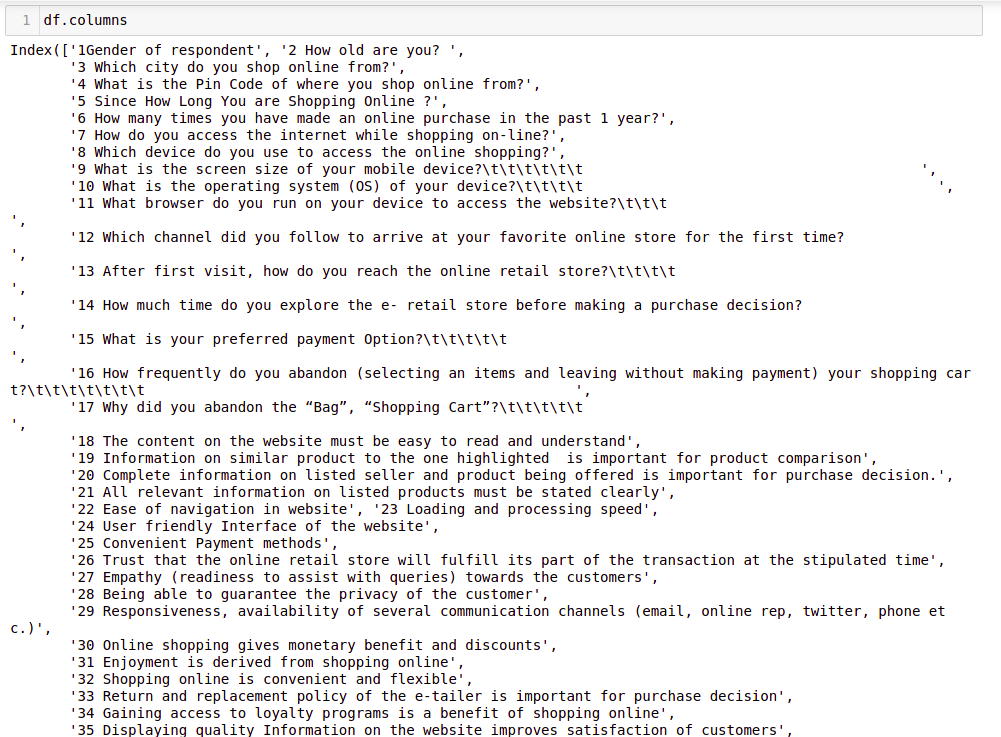
• We built Machine Learning models, applied regularization and determined the optimal values of Hyper Parameters.

• We found important features which affect the price positively or negatively.

The dataset is enclosed in notebook file

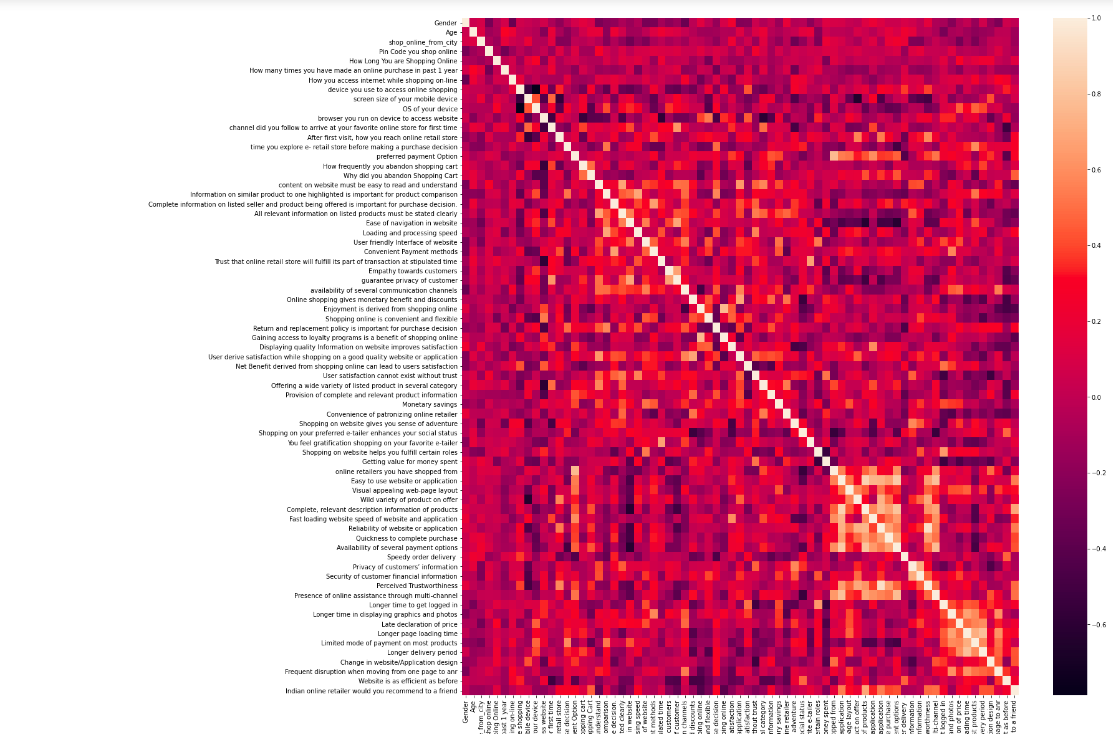
The dataset is provided to us by FlipRobo Technologies.And the dataset is in excel file format.

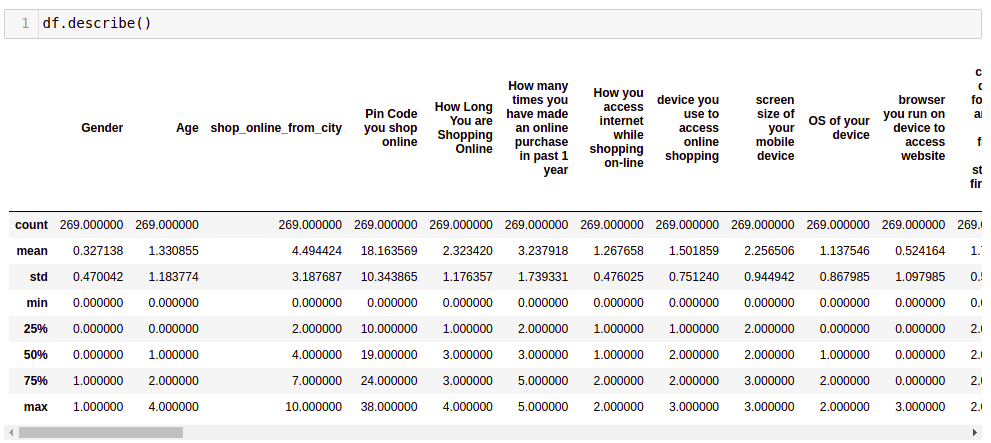
**Data Description:**



Here the columns How old are you? ,How many times you have made an online purchase in the past 1 year?,What is the Pin Code of where you shop online from?,What is the Pin Code of where you shop online from?,What is the Pin Code of where you shop online from? are categorical ordinal data type.And all other columns are categorical nominal data type.Our Target column Which of the Indian online retailer would you recommend to a friend? is the categorical nominal data type.Hence it is a Classification Problem

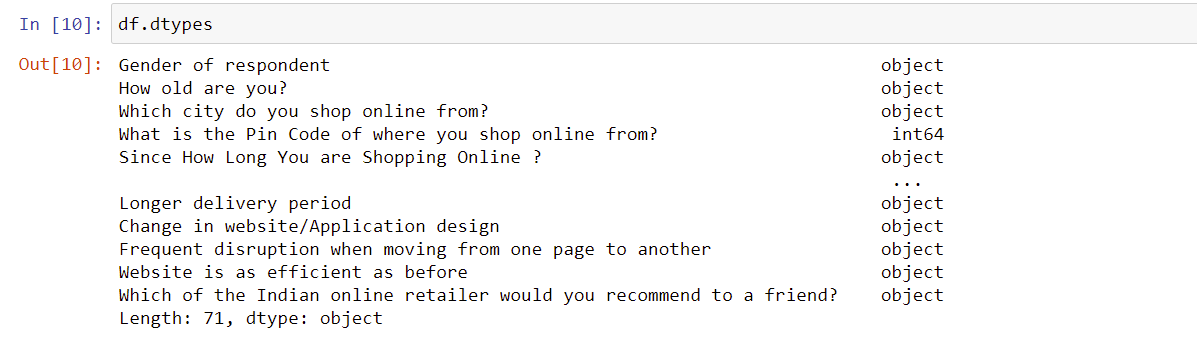
There is no null values in the dataset.

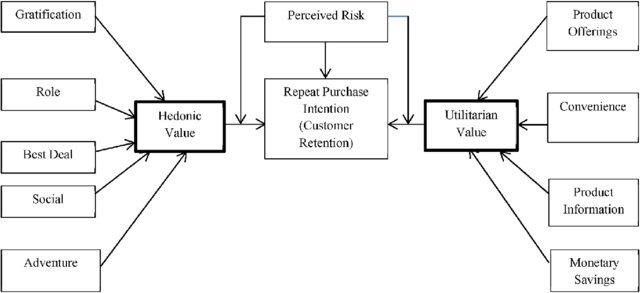
**DATA ACQUISITION**



**FEATURE DESCRIPTION:**

Five major factors that contributed to the success of an e-commerce store have been identified as: service quality, system quality, information quality, trust and net benefit

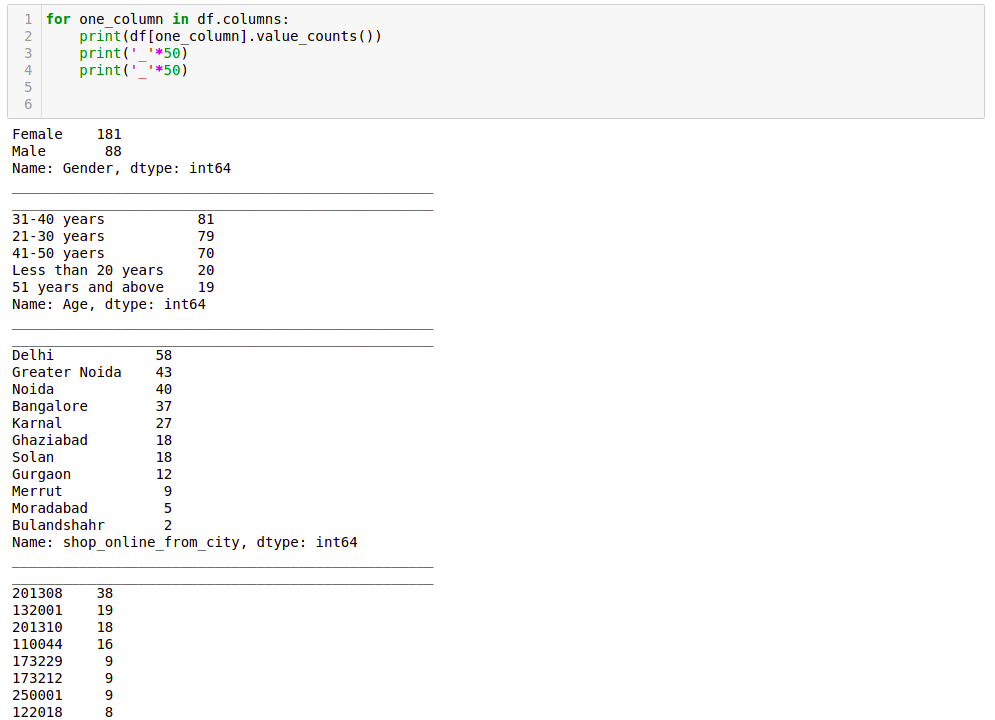


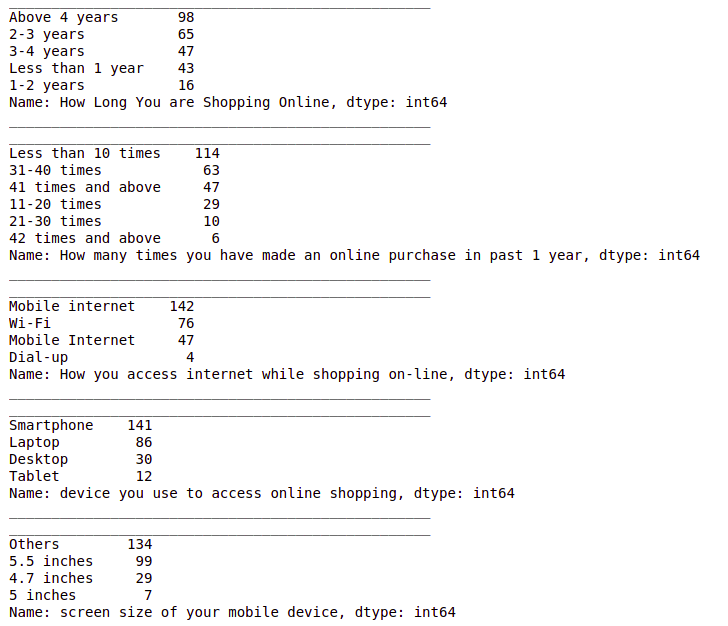


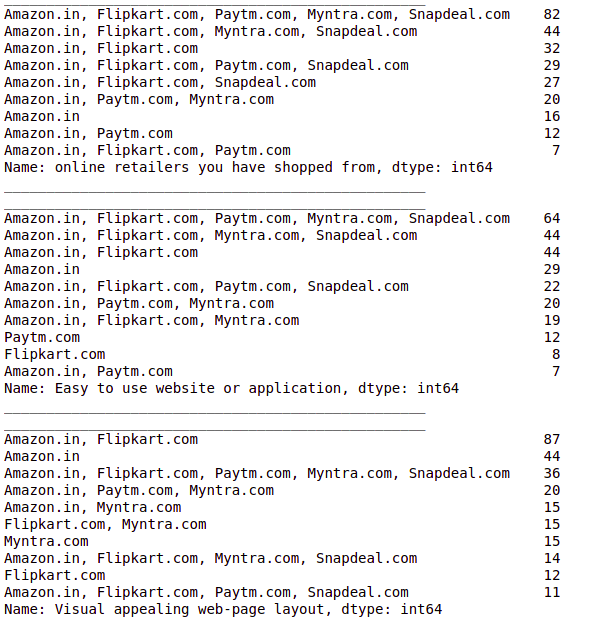
**The left side are Hedonic values and the right side corner values are Utilitarian values.And these values we found in the dataset as attributes.These Attributes contributes as main impactful factors for the Customer Satisfaction and Retention.So our analysis also proved this with more further investigation.And also Found furthermore impactful factors for the E-Retail Business Enhancement.**

Data Analysis and Preprocessing Done

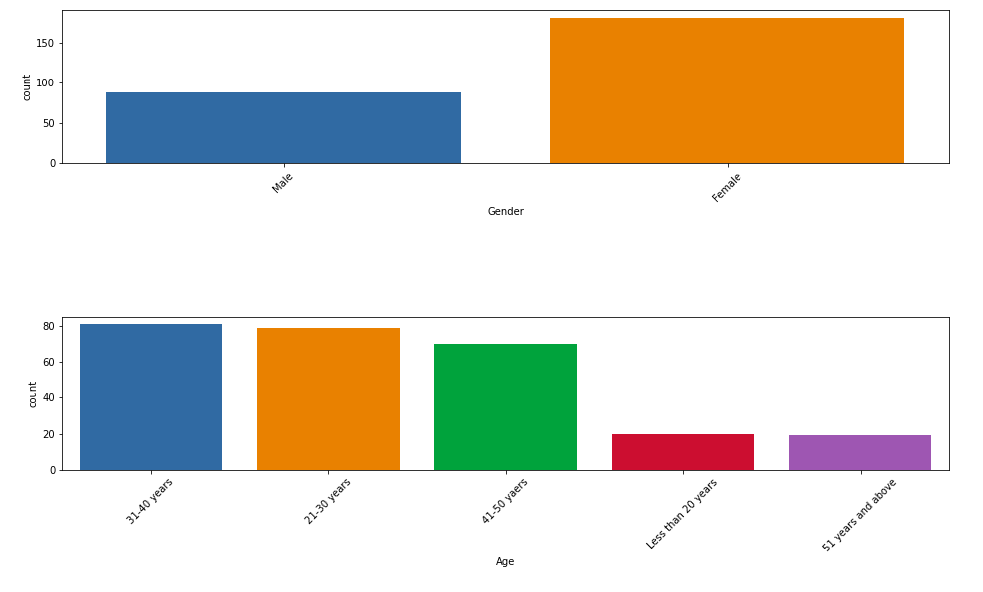
There are no null values in the dataset.

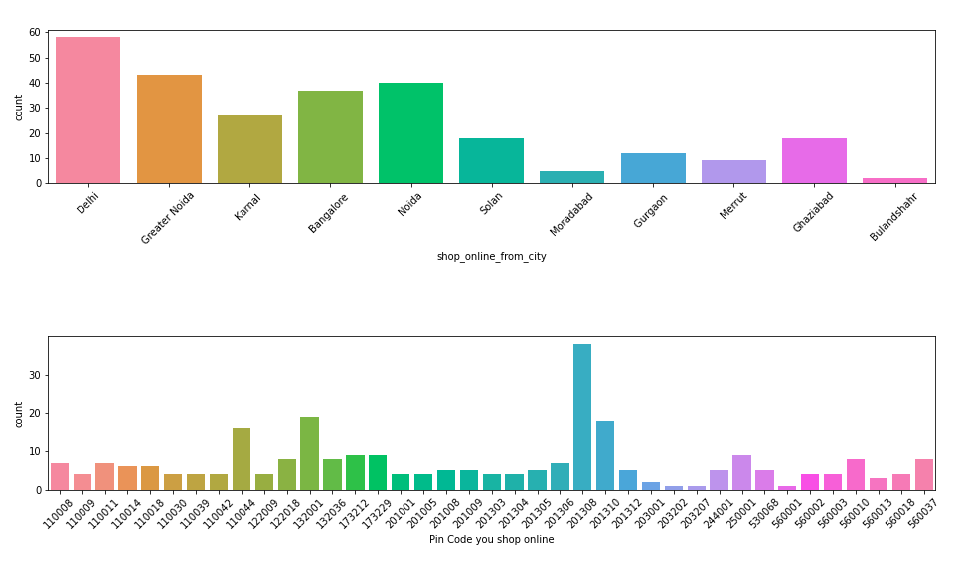






The target column values got segregated for analysis and prediction requirement





Since we have no null values and we segregated the target column.so data is clean now.

Data Inputs - Logic- Output Relationships

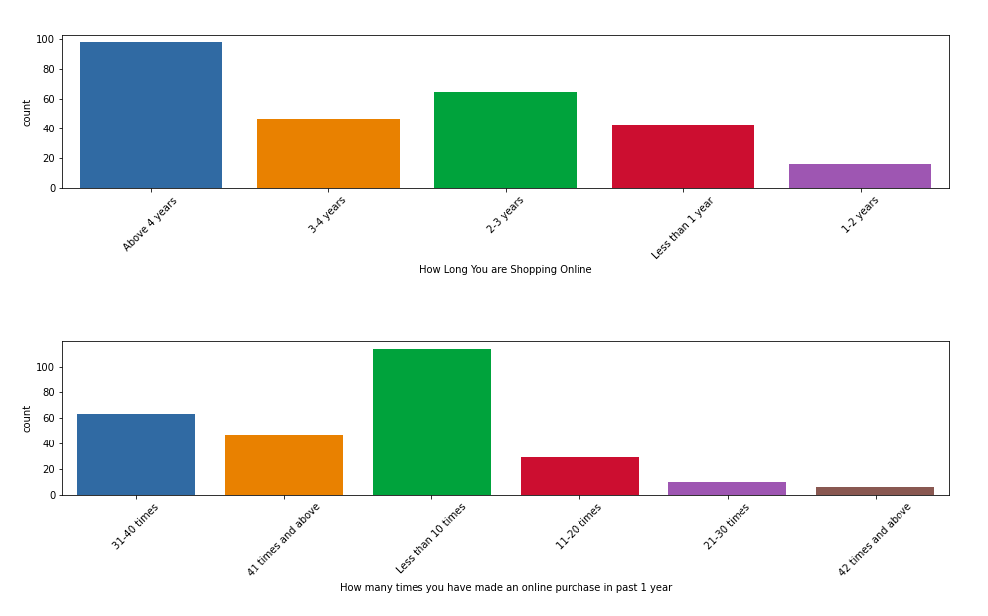
In Classification, the output variable must be **a discrete value**.

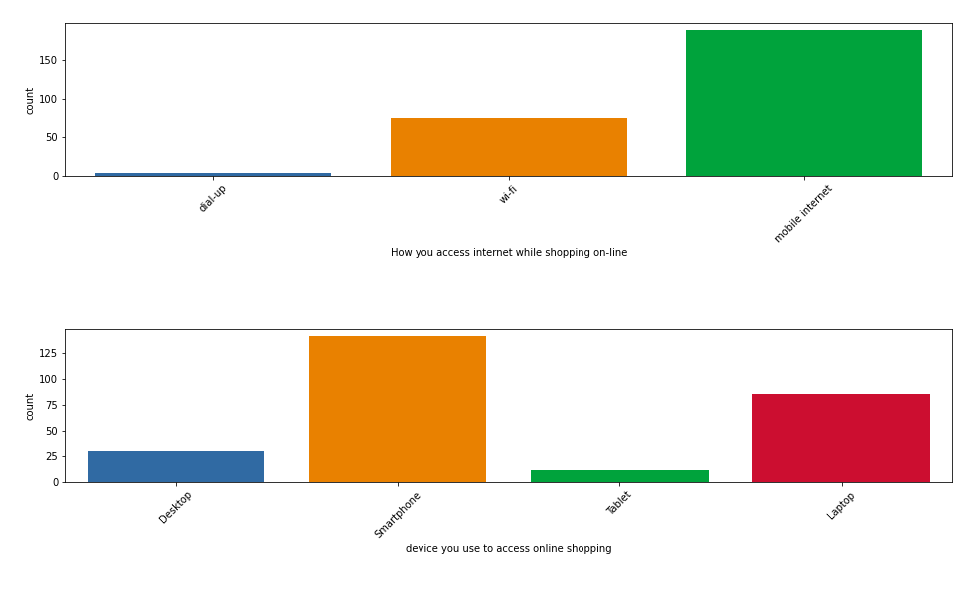
In classification, **inputs are divided into two or more classes**, and the learner must produce a model that assigns unseen inputs to one (or multi-label classification) or more of these classes. This is typically tackled in a supervised way.

In machine learning, classification refers to **a predictive modeling problem where a class label is predicted for a given example of input data**. A classification model attempts to draw some conclusion from observed values. Given one or more inputs a classification model will try to predict the value of one or more outcomes. The output variables are often called **labels or categories**. ... A classification can have real-valued or discrete input variables. A problem with two classes is often called a two-class or binary classification problem. A problem with more than two classes is often called a multi-class classification problem. A classification algorithm, in general, is a function that weighs the input features so that the output **separates one class into positive values and the other into negative values**. Classification is a data mining function that **assigns items in a collection to target categories or classes**. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks.

Classification analysis is a data analysis task within data-mining, **that identifies and assigns categories to a collection of data to allow for more accurate analysis**. ... Classification analysis can be used to question, make a decision, or predict behavior through the use of an algorithm.

DATA PREPROCESSING AND FEATURE ENGINEERING





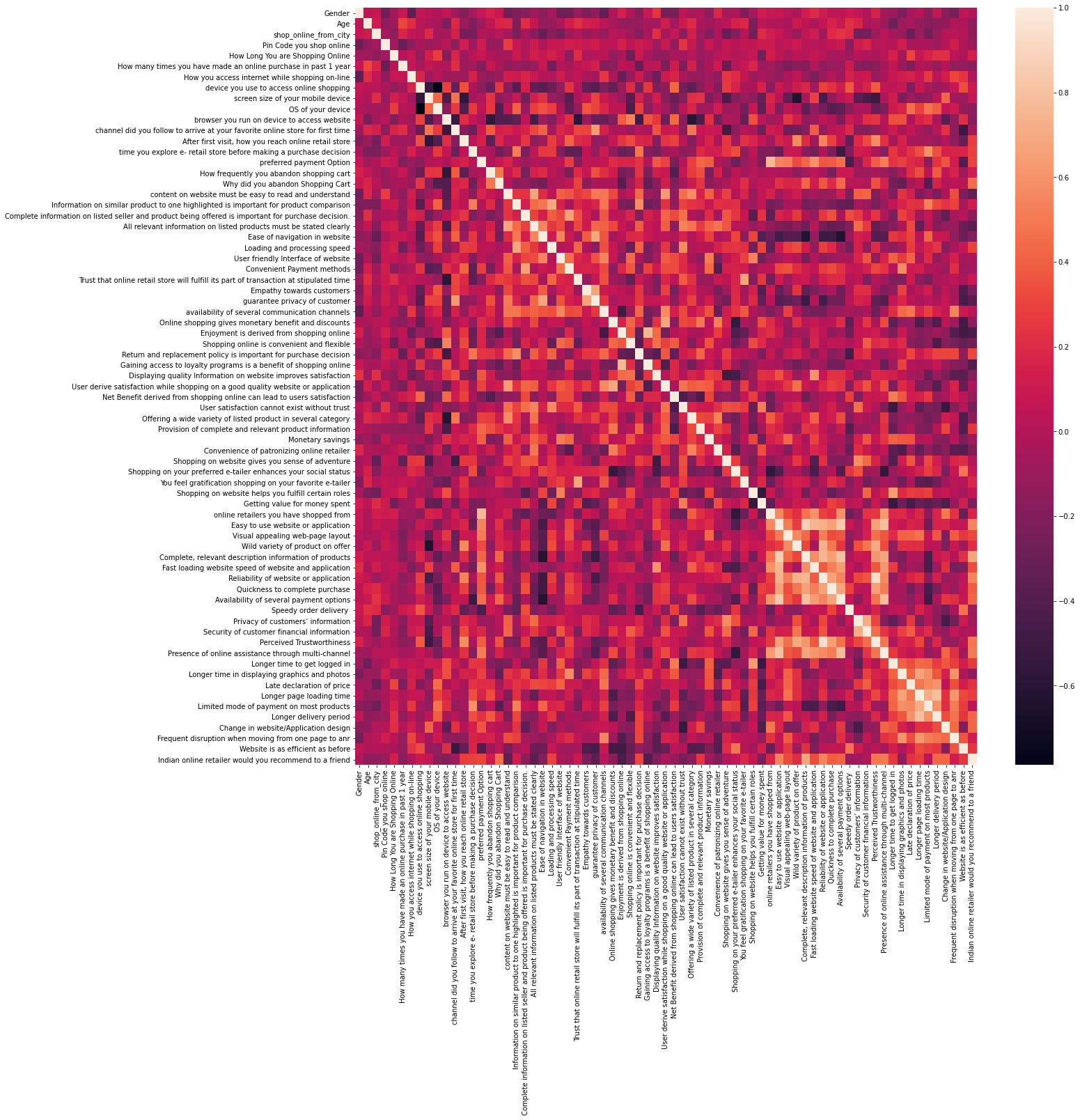
**Correlation with Heatmap:**

The correlation coefficient is a statistical measure of the strength of the relationship between the relative movements of two variables. The values range between -1.0 and 1.0. A calculated number greater than 1.0 or less than -1.0 means that there was an error in the correlation measurement. A correlation of -1.0 shows a perfect [negative correlation](https://www.investopedia.com/terms/n/negative-correlation.asp), while a correlation of 1.0 shows a perfect [positive correlation](https://www.investopedia.com/terms/p/positive-correlation.asp). A correlation of 0.0 shows no linear relationship between the movement of the two variables.Correlation statistics can be used in finance and investing. Pearson correlation is the one most commonly used in statistics. This measures the strength and direction of a linear relationship between two variables.

It can also be defined as the measure of dependence between two different variables. If there are multiple variables and the goal is to find correlation between all of these variables and store them using appropriate data structure, the **matrix data structure**is used. Such matrix is called as **correlation matrix.**

Correlation heatmap is graphical representation of **correlation matrix**representing correlation between different variables.

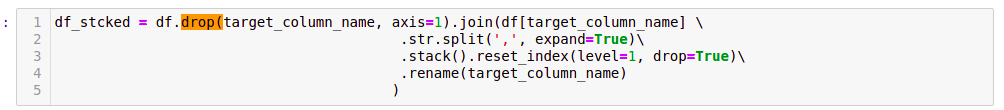
**For to do feature selection and make feature ready for the model building.we check correlation of variables using heatmap.And describe method for the census data set.**



**Correlation model:**

Graph depicts clearly the positive and negative correlation of each variables with target column, justifies the outcome outlined in Multivariate analysis, that higher the education higher the gain & vice-versa

Some of the vif values are very high and since its infinity we use PCA principle Component Analysis Technique to make column reduction.



Dropped this column since it has zero correlation with the target.

Hardware and Software Requirements and Tools Used

**HARDWARE & Software Tools, Libraries and Packages Used:**

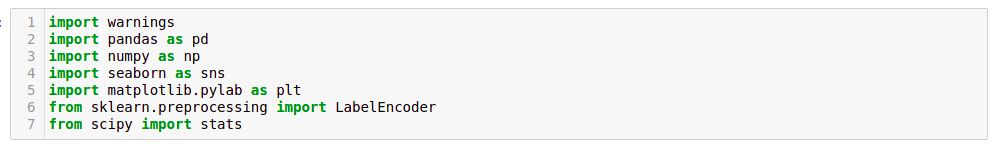
Hardware :Intel i7,RAM 16GB used.

Software: Jupyter Notebook (Anaconda 3)

Language: Python

Libraries:

1. Pandas
2. Numpy
3. Matplotlib
4. Seaborn
5. Sklean
6. Scipy
7. Statsmodels
8. Pip-Package install Manager



|  |  |  |
| --- | --- | --- |
| **Category** | **Tool** | **Function** |
| Data loading and analysis | Import pandas as pd | Pandas is a Python library that is used for faster data analysis, data cleaning and data pre-processing. Pandas is built on top of numpy. So, numpy gets some superpower with pandas. It offers data structures and operations for manipulating numerical tables and time series. |
| Import numpy as np | NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays.It has Quantile method too for removing outliers. It is the fundamental package for scientific computing with Python |
| Data visualization | Import matplotlib.pyplot as plt | Matplotlib is a plotting library used for data visualization. |
| Import seaborn as sns | Seaborn is also a plotting library. It is more advanced than matplotlib but works with matplotlib |
| Scikit Learn Preprocessing Libraries | Sklearn.preprocessing | Package provides several common utility functions and transformer classes to change raw feature vectors into a representation that is more suitable for the downstream estimators.Has power transformer to remove skewness. In general, learning algorithms benefit from standardization of the data set. If some outliers are present in the set, robust scalers or transformers are more appropriate.  It has MinMaxScaler to scale the data. |
| Sklearn.preprocessing import LabelEncoder | Label Encoding in Python can be implemented using the Sklearn Library. Sklearn furnishes a very effective method for encoding the categories of categorical features into numeric values. Label encoder encodes labels with credit between 0 and n-1 classes where n is the number of diverse labels. |
| Import statistics | Import statsmodels.api as sm | From scipy import stats This module provides functions for calculating mathematical statistics of numeric (Real-valued) data. This library provides a number of common functions and types useful in statistics. It focus on high performance, numerical robustness, and use of good algorithms |

Variance Inflation Factors (VIFs) measure the correlation among independent variables in least squares regression models. Statisticians refer to this type of correlation as multicollinearity. Excessive multicollinearity can cause problems for regression models. The stats models package has VIF library and we can import this library.

The scikit-learn Python machine learning library provides an implementation of the train-test split evaluation procedure via the train\_test\_split() function. The function takes a loaded dataset as input and returns the dataset split into two subsets.train\_test\_split() will split arrays data into random subsets. The ideal split is said to be 80:20 for training and testing.

**The most commonly used Performance metrics for classification problem are as follows,**

* Accuracy.
* Confusion Matrix.
* Precision, Recall, and F1 score.
* ROC AUC.
* Log-loss.

Grid search is used as an approach to hyper-parameter tuning that will methodically build and evaluate a model for each combination of algorithm parameters specified in a grid. GridSearchCV helps us combine an estimator with a grid search preamble to tune hyper-parameters.

Cross-validation is a technique in which we train our model using the subset of the data-set and then evaluate using the complementary subset of the data-set.

The three steps involved in cross-validation are as follows :

1. Reserve some portion of sample data-set.
2. Using the rest data-set train the model.
3. Test the model using the reserve portion of the data-set.

The most important use of PCA is to **represent a multivariate data table as smaller set of variables** (summary indices) in order to observe trends, jumps, clusters and outliers. This overview may uncover the relationships between observations and variables, and among the variables.

Also import all the required algorithms for classification purpose below.

**Model/s Development and Evaluation**

Identification of possible problem-solving approaches (methods)

since these are all categorical columns and of discrete values.Hence no need to care much about outliers and skewness.Because it applies to numerical data.

Inferences:

1. There is 71 column and 269 rows

2. In the this Female 181 Male 88

3. Female do more shopping than man

4. Age group of 21-30 and 31-40 do more shopping than others

5. Most of the shopping are done from Delhi itself

6. 201308 Pin code do more shopping then others

7. Most people are shopping form 4 years or more

8. Most people made shopping less than 10 times in year

9. Most of people use mobile internet for shopping

10. Most of the people order from smartphone

11. Most people use chrome browser for shopping

12. Most of people search in search engine to reach the shopping site

13. Most of people explore more than 15 minutes before purchase

14. Most of people order from credit/debit card while paying

15. reason of leaving the cart is because they find other good option

16. Most people Strongly Agree that "Return and replacement" should be there

17. Some People feel lack of trust in flipkart and some didn't found any preferred mode of payment in amazon

**CONCLUSION**

Key Findings and Conclusions of the Study

Customer Satisfaction and Retention can be improved by concentrating on the above mentioned hedonic and utilitarian values. And it can be attained by concentrating on the following impactful factors such as

Getting value for money spent Loading and processing speed User satisfaction cannot exist without trust The Convenience of patronizing the online retailer Privacy of customers’ information Quickness to complete purchase You feel gratification shopping on your favorite e-tailer Longer time in displaying graphics and photos (promotion, sales period)', Late declaration of price (promotion, sales period) Monetary savings Frequent disruption when moving from one page to another

Five major factors that contributed to the success of an e-commerce store have been identified as: service quality, system quality, information quality, trust and net benefit

Limitations of this work and Scope for Future Work

The main limitation of random forest is that **a large number of trees can make the algorithm too slow and ineffective for real-time predictions**. In general, these algorithms are fast to train, but quite slow to create predictions once they are trained.

For **very large data sets, the size of the trees can take up a lot of memory**. It can tend to overfit, so you should tune the hyperparameters.

Here we used PCA to project the data into a new space where the 'new features' will be orthogonal to each other. We then, trained the model with the new features, but we found that the performance is the same.so we must You simply rotate original decision boundary to overcome this limitation in future and try with the same model.