**INTRODUCTION**

* Business Problem Framing

The company is looking at prospective properties to activate and retain more customers and increase the market cap against their competition. Using Data Analysis tools to predict the most valuable and prospective properties from the data increases the market capitalization.

* Conceptual Background of the Domain Problem

The main problem is found variables which impact most on the customer's minds and habits of shopping, also the features which are nowadays so important to make them shop more.

* Review of Literature

Data exploration is the first step in data analysis and typically involves summarizing the main characteristics of a data set, including its size, accuracy, initial patterns in the data, and other attributes. It is commonly conducted by data analysts using visual analytics tools, but it can also be done in more advanced statistical software, Python. Before it can analyze data collected by multiple data sources and stored in data warehouses, an organization must know how many cases are in a data set, what variables are included, how many missing values there are, and what general hypotheses the data is likely to support. An initial exploration of the data set can help answer these questions by familiarizing analysts with the data with which they are working.

* Motivation for the Problem Undertaken

Every problem of Machine learning gives us chance to enhance and develop problem-solving skills. These Problems do’s the same.

When this real-life problem of predicting the factors for increasing the sales, how to retain more no of customers and activate more no markets for increasing sales, with the help of A. I technology affordable shopping for the customers.

As Data scientists it is our role to help companies to understand the market better with older data we have for constructing the websites and product list according to that only and make profitable models.

**Analytical Problem Framing**

* Mathematical/ Analytical Modelling of the Problem

As for any basic model building, we have to understand the type of target variable, the data of the target variable is continued or classified.

Data Analysis is always the difficult part, for better understanding different kinds of bar plots, distribution plots are created with the target Column for finding the insights of the dataset we have.

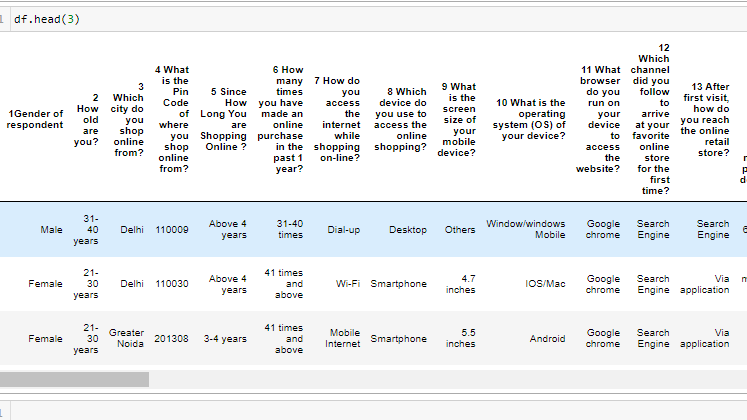
Analytical Modelling always starts with the target variable we have, and in that case, our target variable is Repeat Purchase, for that, we create some box plots with the target variable to understand which feature columns help to learn the model best and which feature columns reduce the accuracy of the model.

And after finding the relation and correlation with the target variable we understand that which feature columns have the potential to increase the sales for the companies, also focus on most related aspects for activating more no of customers.

* Data Sources and their formats

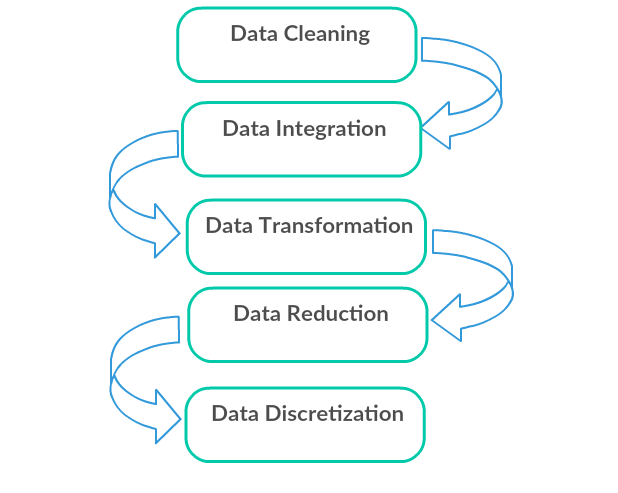
Customer satisfaction has emerged as one of the most important factors that guarantee the success of the online store; it has been posited as a key stimulant of purchase, repurchase intentions, and customer loyalty. A comprehensive review of the literature, theories, and models have been carried out 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 online customers to repeat purchase intention. The combination of both utilitarian value and hedonistic values is needed to affect the repeat purchase intention (loyalty) positively. The data is collected from Indian online shoppers. Results indicate the e-retail success factors, which are very much critical for customer satisfaction.

**Dataset looks as follows-**

****

* Data Pre-processing Done

Data pre-processing can refer to the manipulation or dropping of data before it is used to ensure or enhance performance, and is an important step in the data mining process.



1. Data Cleaning: First we clean the data which have no use in prediction like the Pincode column, then we drop the data which has a high no of missing percentages.
2. Data Integration: then we do some EDA process for finding out the meaning full insights of the data.
3. Data transformation is the process of changing the format, structure, or values of data; we use a labeled encoder for coding the object data into integer data.
4. Data Reduction: it is the process of finding the most correlated columns, and combining them because the machine does not understand which feature columns impact the most on accuracy.
5. Data discretization converts a large number of data values into smaller once, so that data evaluation and data management becomes very easy, using box plots is makes a clear understanding of the data.

Data Inputs- Logic- Output Relationships

Finding the relation between the columns of the input with the target column is always important.

**1. Repeat purchase vs. Device used.**

Observations:

1. As from the above observations, smartphones are used most.

2. Repeat purchase will depend on other features as well.

3. As the smartphones application is created sales go up.

**2. Year of shopping vs. Repeat Purchase**

Observations:

1. As from the above observations, as we see years increase the sales as the more loyal customer base is built.

2. Males year of shopping 3-4 years repeat purchase 42 times and above.

3. Females less than year shops 42 times and above.

**3. Age vs. Repeat Purchase**

Observations:

1. As from the above observations and plotting we can easily see the males less than 20 years shop 42 times.

2. Females age group 41-50 years shop 42 times and above.

* Hardware and Software Requirements and Tools Used

**Python**

Python is widely used in scientific and numeric computing:

SciPy is a collection of packages for mathematics, science, and engineering.

Pandas are data analysis and modeling libraries.

Plotly used here for plotting 3-d graphs which help to get

detailed insights.

Libraries Used for this Project include –

1. Pandas

2. NumPy

3. Matplotlib

4. Seaborn

5. Scikit Learn

6. Plotly

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

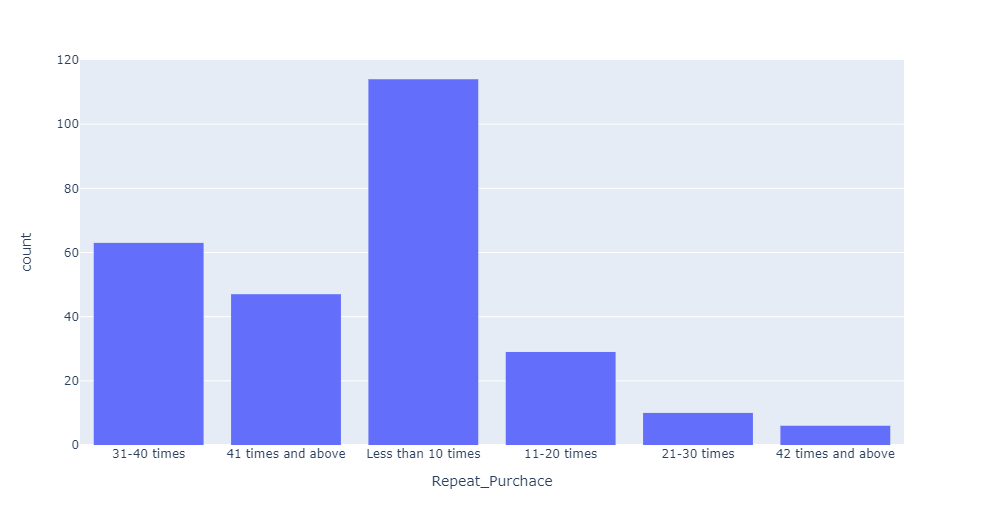
After analyzing the dataset, I observe that many of the feature columns are object types so first, we have to convert them into integer or float types so that the machine interprets the data and for that we do label encode all the features column.

After label encoding, we find that many feature columns have Nan values so we use mean and median for filling that missing data,

Then find the correlation between the columns with target columns and delete the non-related feature columns.

Data Visualization

**Bivariate analysis for better understanding which features columns is impacting more on Predicting Repeat sales.**

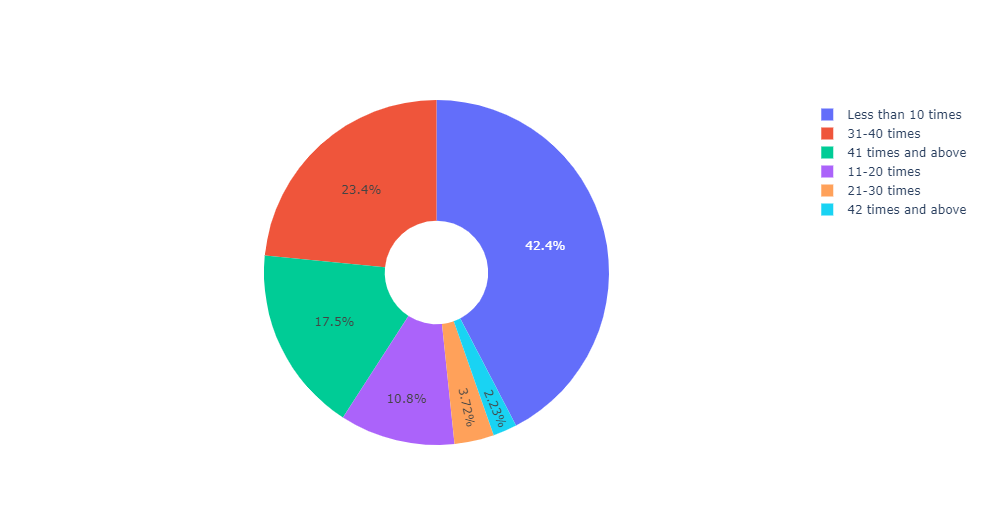
****

**1.** Data is not normalized.

2. 42 times repeat purchase is less.

3. Less than 10 times column having more counts.

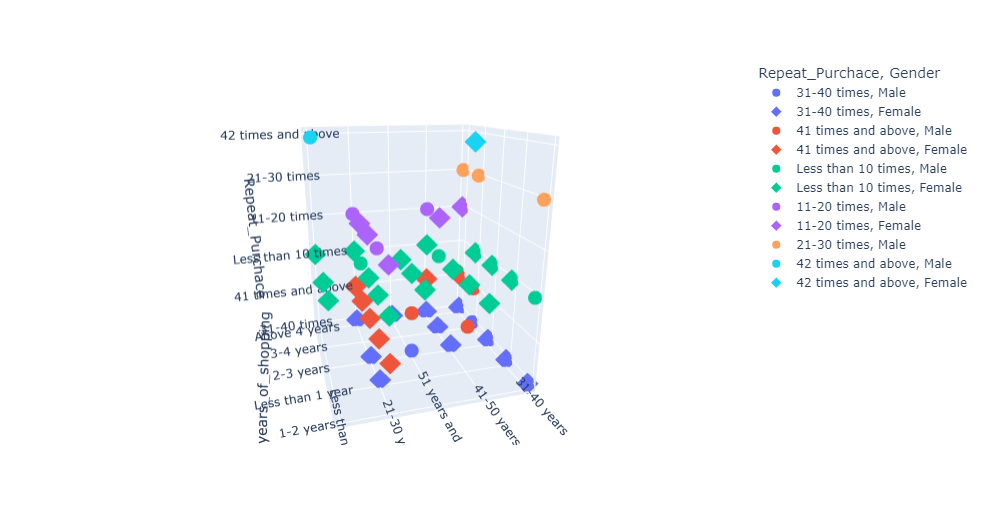
**Pie Chart**

****

1. As from the above chart we can easily analyze the Repeat sales data and their percentage group.

**Observations:**

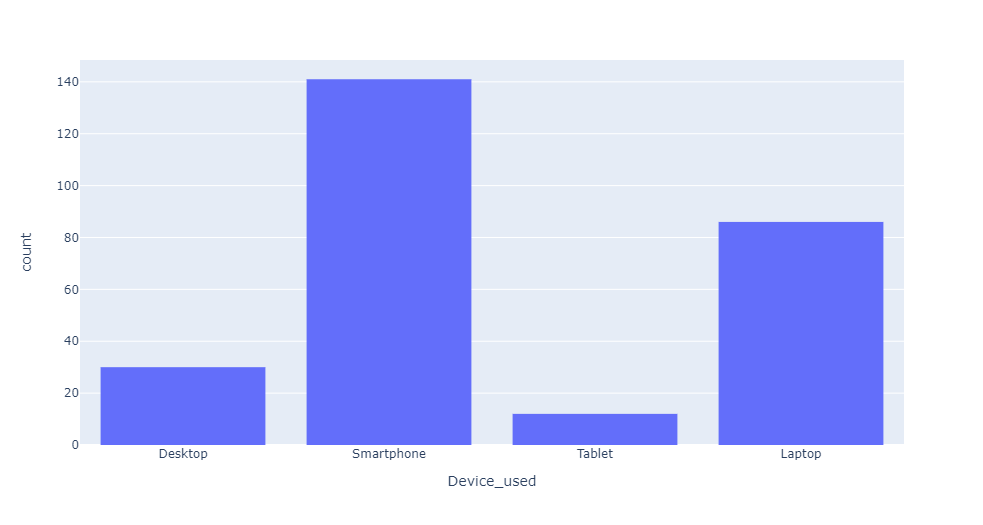
1. From the above plotting we can easily visualize which group purchase more no of times also the gender of the person as well.

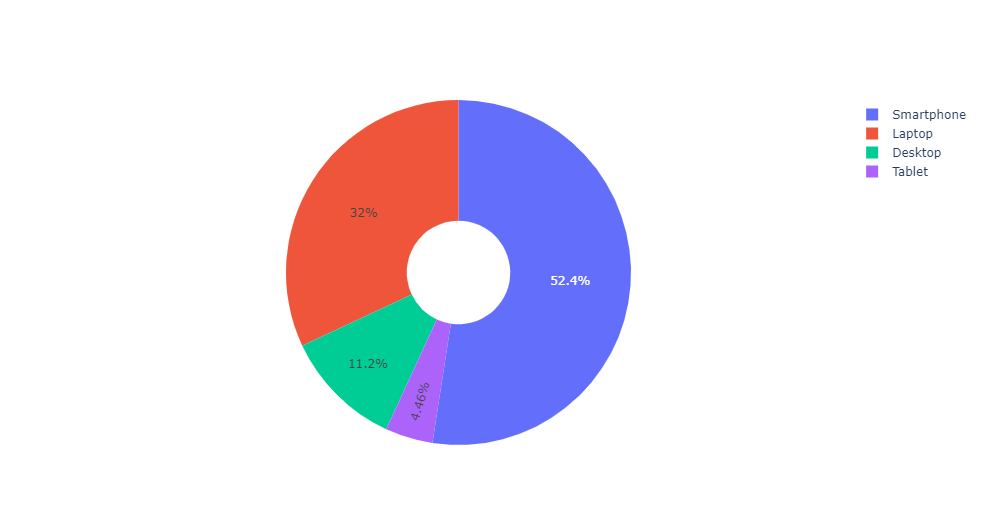


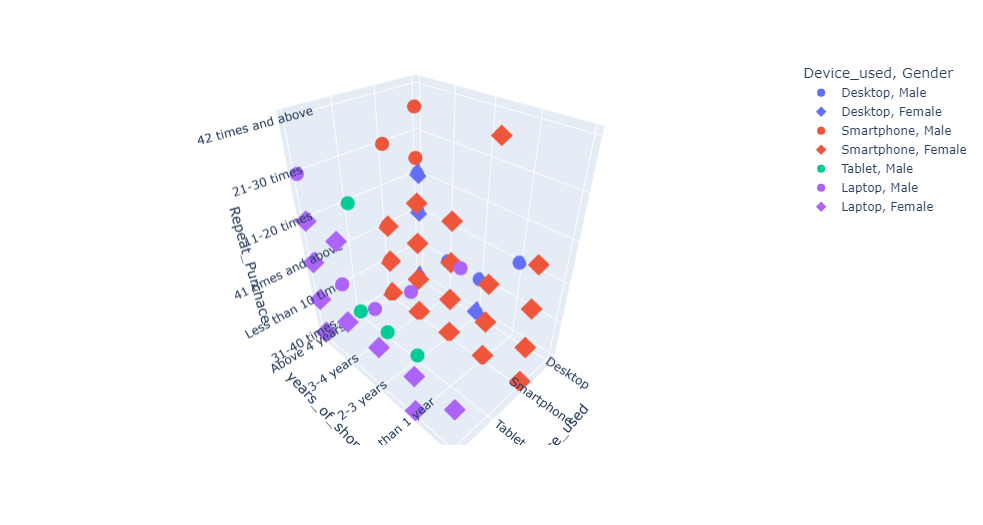
1. Males and Females have different shopping habits.

2. Which gender purchase how many times are described in this chart.

**Device Used vs Repeat Sales.**

****



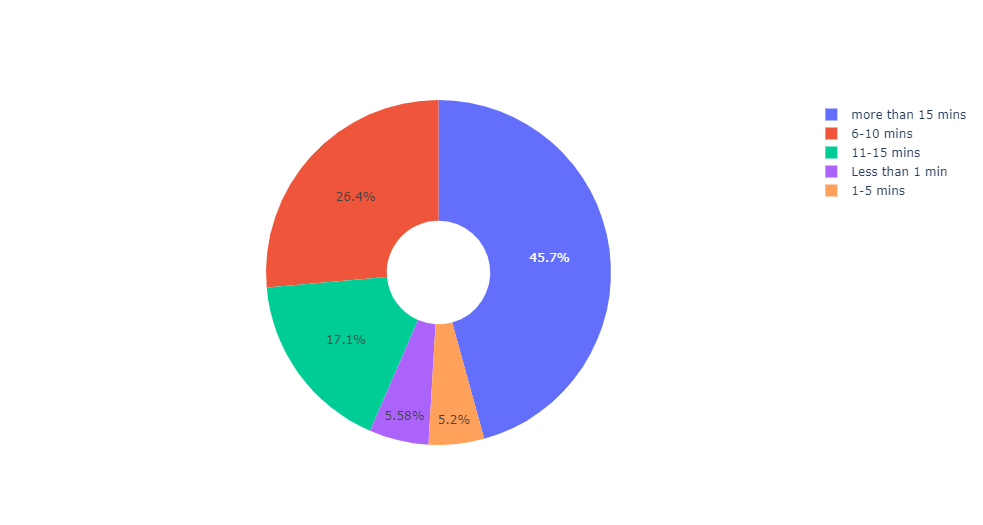


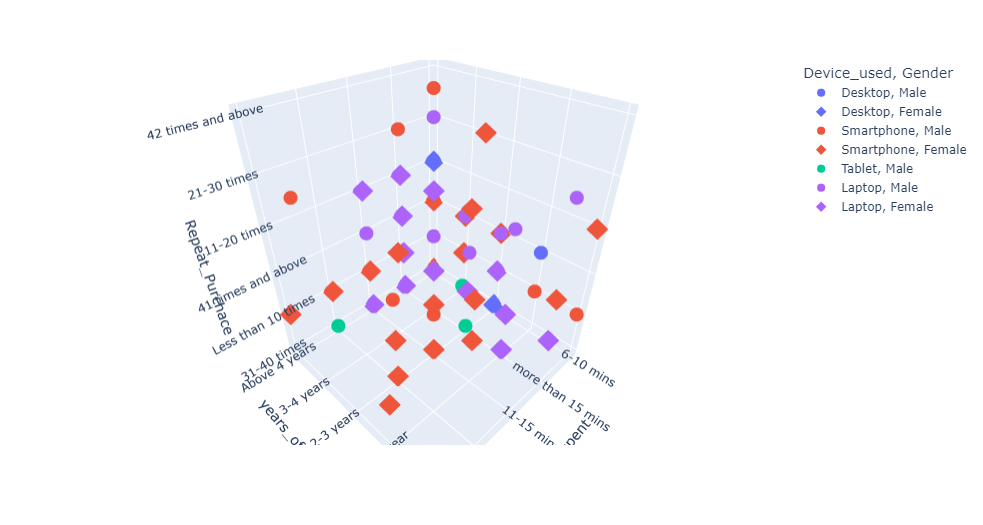
observations:

1. From the above plotting we easily look at every device used to purchase for shopping.

2. Most people used smartphones for repeat purchases.

**Time spent on Website while buying.**

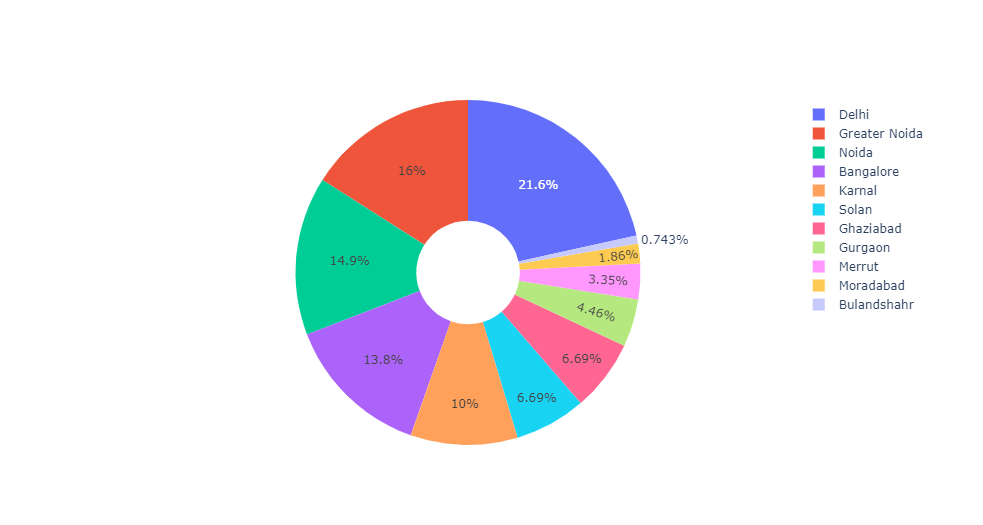




Observations:

1. Smart Phones are used maximum for shopping.

**Location**

****

Observations:

Major cities are where maximum buying happens.

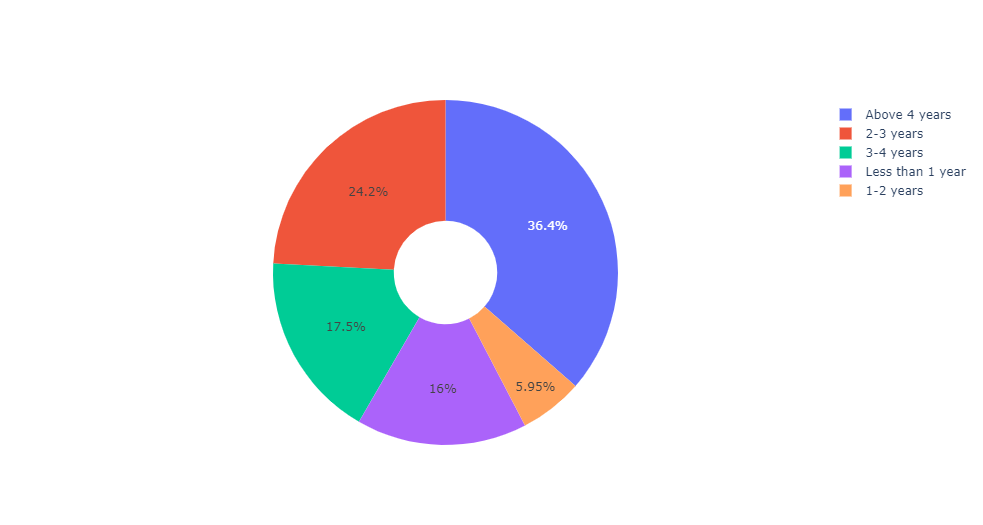
1. Delhi

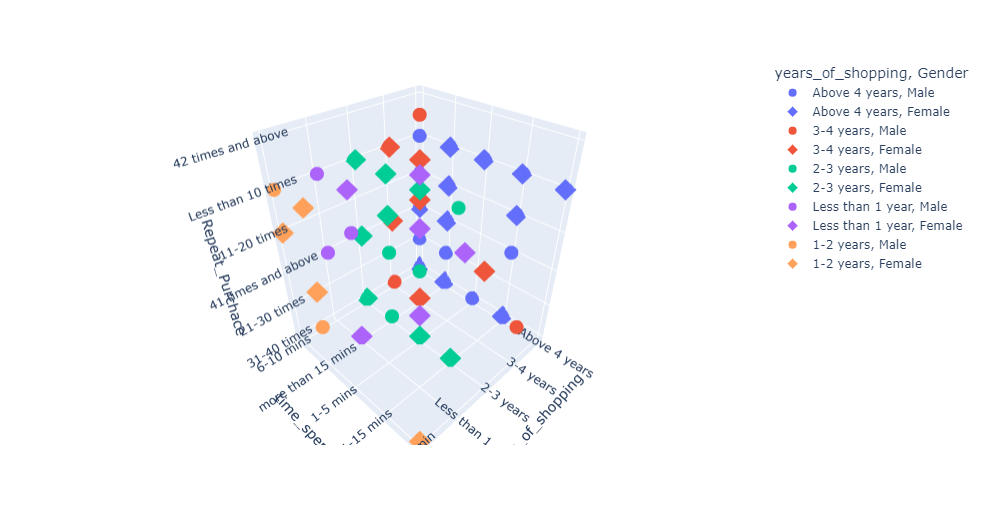
2. greater Noida

3. Noida

4. Banglore

**Year's of Shopping**

****



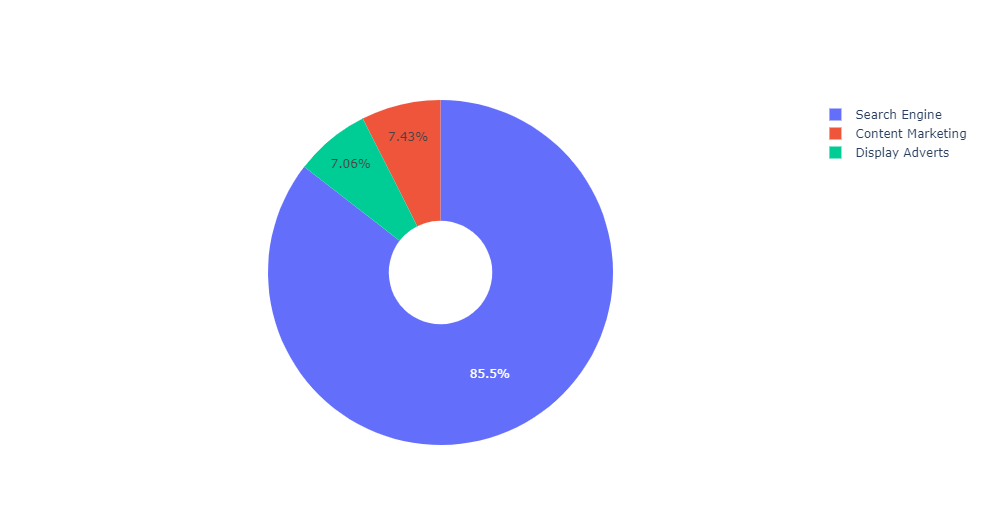
Observations:

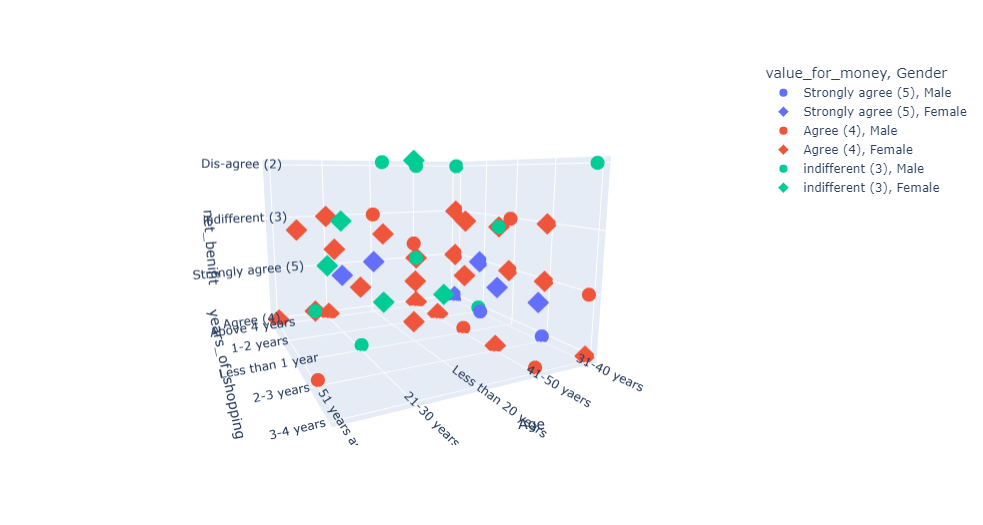
1. As from the above observations, as we see years increase the sales as the more loyal customer base is built.

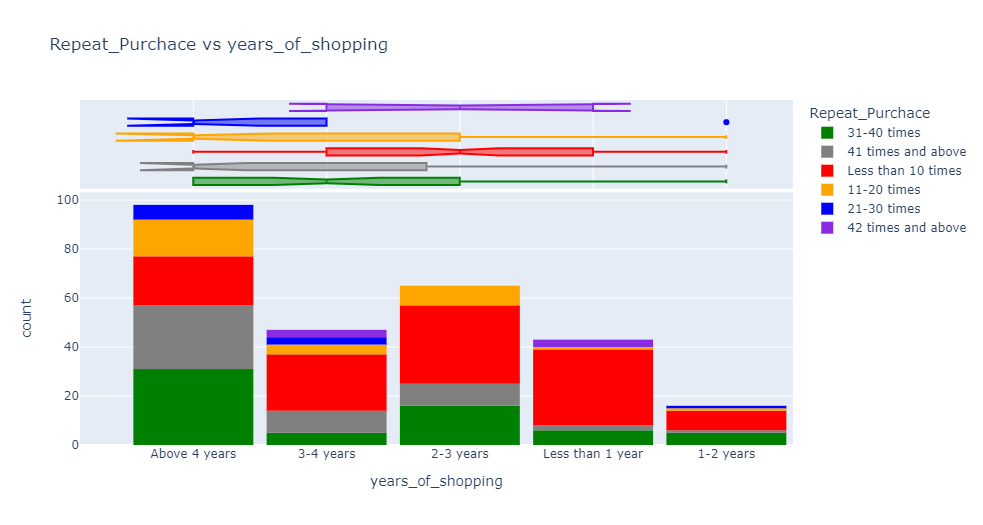
2. Males year of shopping 3-4 years repeat purchase 42 times and above.

3. Females less than year shops 42 times and above.

**Favorite online store**

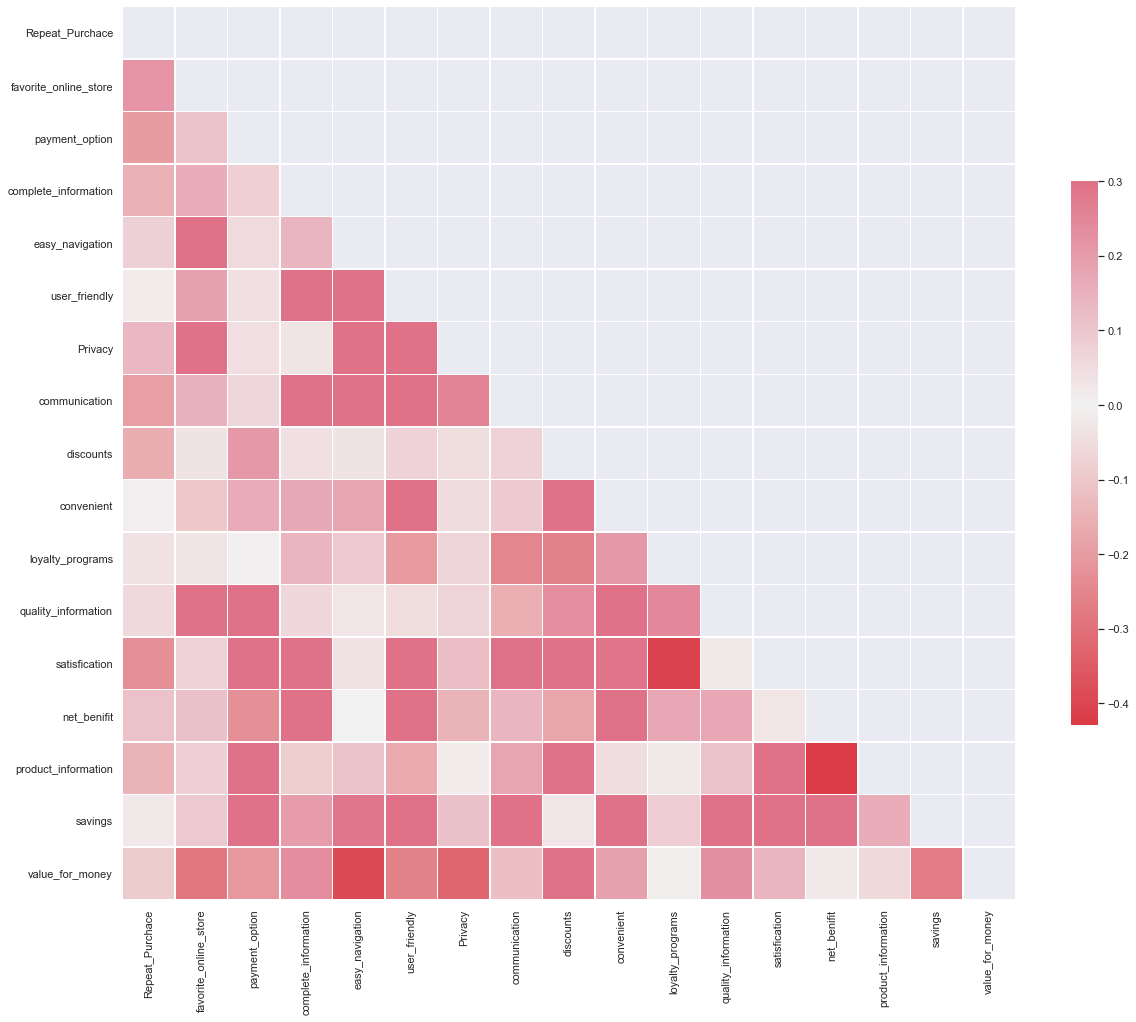
****

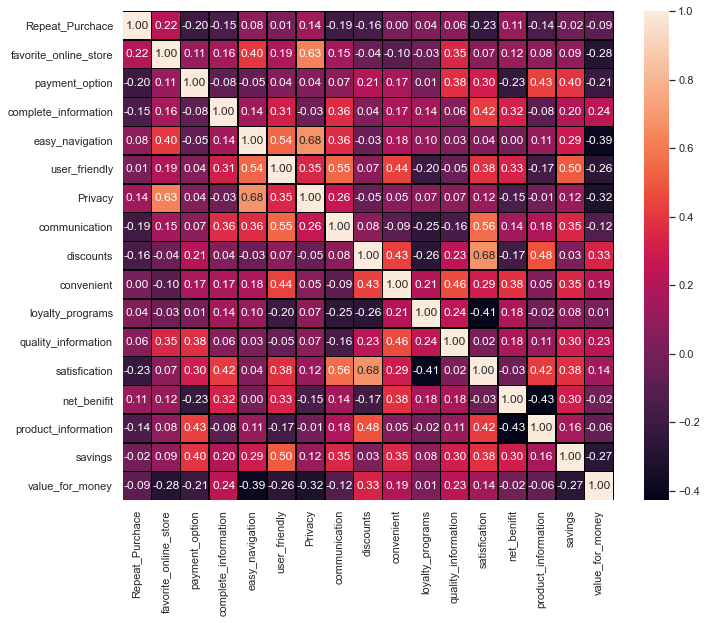




Years of shopping increase the sales.

**CORRELATION BETWEEN THE COLUMNS:**





**Observations:**

1. From the above heat map we observe that

favorite online store,

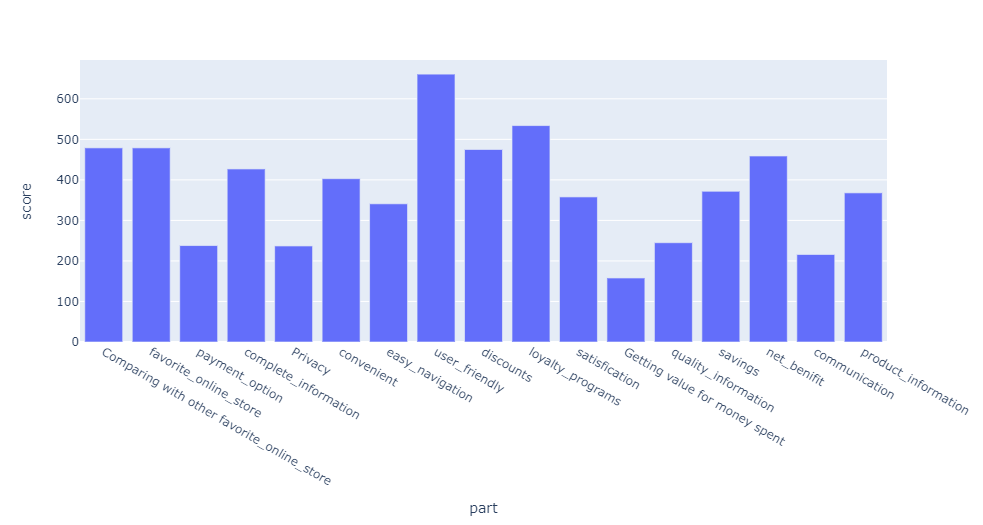
easy navigation,

privacy,

loyalty programs,

quality information and net benefits

makes the customer purchase on a repeat note.



* **Which part satisfies customers the most?**
  + 1. Those websites are user-friendly having more repeat customers.
  + User-friendly scores 661 points which are best from the rest of the options.
* **Which part satisfies customers the least?**
  + Value for money, those websites that do not give value for money options are not retaining the customers.
  + Value for money scores 158 points.

**CONCLUSION**

* Key Findings and Conclusions of the Study
* 1. When customers find the website user-friendly they visit more and more, which makes them loyal customers.
* 2. When they find what they get is not value for money they don't buy and leave the cart empty.
* 3. More time spent on a website makes them buy more.
* 4. Most user buys in less than 10 minutes which means what they want they find easily which means product information is a must-have feature.
* 5. Loyalty programs and Discounts makes them repeat purchase from the website, which means that website gives timely discounts have more customers.
* 6. Net benefit is also a must-have feature it increases the repeat purchase.

**Thank you**