

[**E-retail factors for customer activation and retention: A case study from Indian e-commerce customers**](https://www.researchgate.net/publication/346412647_E-retail_factors_for_customer_activation_and_retention_An_empirical_study_from_Indian_e-commerce_customers)

Submitted by:

**AYUSH YADAV**

Under the Supervision of

**Ms. Sapna Verma**

(Mentor)

**ACKNOWLEDGMENT**

I am highly indebted to my SME Ms. Sapna Verma for her guidance and constant supervision as well as for providing necessary information regarding the project & also for her support in completing the project. Her valuable suggestions were of immense help throughout my project work.

I would like to express my gratitude towards members of Flip Robo Technologies for their kind co-operation and encouragement which help me in completion of this project.

I would like to express my special gratitude and thanks to Datatrained institute persons for giving me such attention and time. I would also like to acknowledge sklearn documentation from scikit-learn.org which helped me understand many modules necessary for completion of my project.

**INTRODUCTION**

* Business Problem Framing

The fierce competition in E-commerce has forced the companies to focus on providing customer satisfaction and gain customer loyalty. Thus, measuring the parameters which drive customer satisfaction is very important for the long-term growth of the businesses. This research work makes an attempt to identify the factors which influence the loyalty and satisfaction of customers in organized E-commerce sites. The samples are collected from users who regularly do the transactions in E-commerce sites in metro cities, majorly Delhi NCR and Mumbai. The study also identifies relative importance of factors affecting satisfaction. Attention has been given to cover all the major factors with special emphasis on the logistics part.

* Conceptual Background of the Domain Problem

Technological development, cheap and fast internet availability, nature of city life etc has led to drift consumer from traditional to internet base (online) shopping. With the movement of greater population percentage towards internet based shopping, more and more E-commerce sites are coming to the forte. Higher the number of E-commerce site (companies) in the fray higher the competition. This gives opportunity to the consumer to select the sites which provides highest customer satisfaction. While consumer behaviour in Ecommerce seems to be a complex subject, the consumer expectations are changing, challenging traditional patterns of supply of commercial websites. E-commerce has become one of the essential features in the internet era. Online shopping become the third most popular internet activity immediately following e-mail and web browsing.

* Review of Literature

As long as repeat business is important and as long as customers have chance to go somewhere else, Ecommerce sites must deliver high level of customer satisfaction to be successful, in a competitive market place that offers meaningful consumer choices. Firms that do well by their customer are rewarded by business. According to Allagui and Temessek (2005), the theoretical foundations of loyalty to a company of the internet are similar to those of traditional loyalty. It is defined as continuing relationship established between the consumer and a brand as long as repeat business is important and as customers have the chance to go somewhere else, companies must deliver high level of customer satisfaction to be successful.

* Motivation for the Problem Undertaken

I am one person who I like been productive, innovative and results oriented. Data science been an interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured, which is a continuation of some of the data analysis fields such as statistics, machine learning, data mining, and predictive analytics encompasses all of my character traits listed above thus why handling these kinds of machine learning is the perfect match for me.

What really motivate me to take this project, is the potential I usually see in the big data. From my own point of view in this data, lays brilliant business opportunity if well analysed and this will be the solution to the big problem of churning of customers, making customers loyal and retaining them.

In addition to getting a good score in this internship and perform well, another motivating factor is my love for technology and doing research. A project involving data science is very experimental and involving allot of creativity job, is the perfect venture which undoubtedly match with my hobbies.

Furthermore, my love in love and algorithms and mathematics, can be well quenched when I partake a data science project which requires someone who got good mathematics background and programming skills with deep roots in algorithms. In addition, my curiosity I also guess will rely help me become the best person to undertake this project.

Lately, I have been studying most of python libraries and framework, and the likes of numby, pandas and motplotlab with conjunction with web scraping library, beautifulsoup, selenium really empower me to undertake this task.

In conclusion, I like learning especially materials related to technology. As a matter of fact, I have solely studied two online courses, this is python and java courses from solo learn community and got certified in the two courses and also in progress with Data science course from Datatrained. I do understand that data scientist guys need to self motivated and able to learn on there own, thus, I actually see that I’m a good fit for this task.

**Analytical Problem Framing**

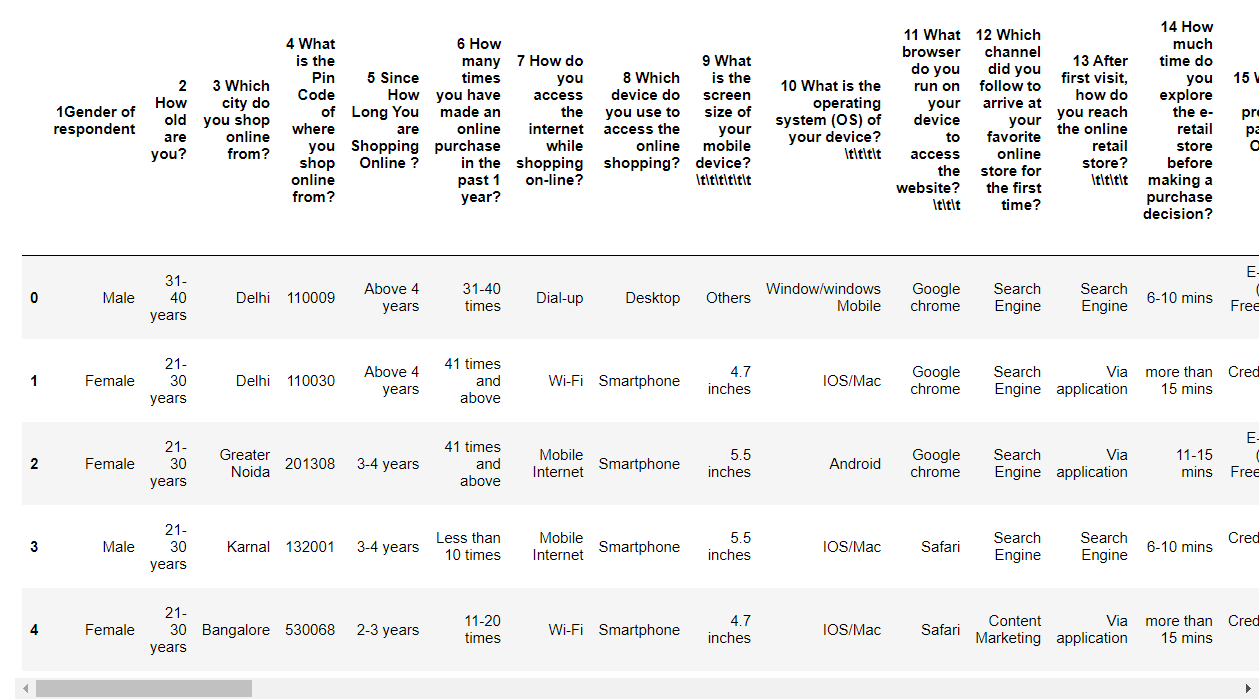
* Mathematical/ Analytical Modeling of the Problem

We will use all the dependent features such as After first visit, how do you reach the online retail store?', 'How much time do you explore the e- retail store before making a purchase decision?', 'What is your preferred payment Option?', 'How frequently do you abandon (selecting an items and leaving without making payment) your shopping cart?', 'Why did you abandon the “Bag”, “Shopping Cart”?', etc, give them to our model which will apply mathematical functions to it and predict our dependent feature.

f(independent features)-------> dependent feature

* Data Sources and their formats

The sample data is provided to us from our client database. It is hereby given to me for this exercise. In order to analyze customers and what their preffered choices are in E-retail sectors.



Describe The data is given in the comma separated value (.csv) file format. This data consists of 269 rows and 71 columns. All the columns are of object datatype except for pincode column which is of int type. There are no missing values in the dataset.

* Data Pre-processing Done

As the data was in raw form a lot of cleaning of the data was required in addition to feature engineering so that it can be analysed easily and efficiently.

First and fore-most whitespaces was removed from dataframe’s columns along with ‘\t’ characters. Ambiguity from different columns was removed to keep data consistent and to avoid confusion such as in the city column greater Noida and Noida was given which was combined as one category (Noida) later, etc.

* State the set of assumptions (if any) related to the problem under consideration

Data was categorised into different categories such as personal info, application performance, pereference of customer, Negative and positive remarks and then analysis was performed on these categories in isolation. Later on, their results were combined.

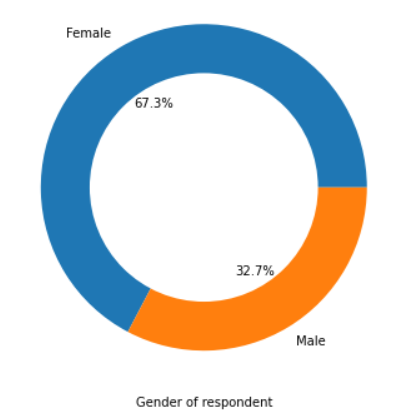
* Hardware and Software Requirements and Tools Used

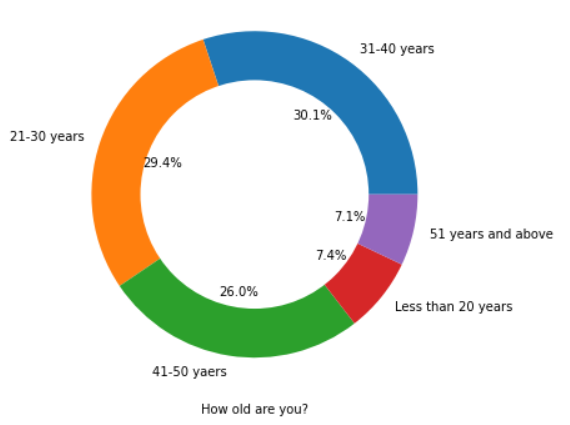
For this machine learning project I have used a hp pavilion gaming laptop, having 8gb ram, 512 gb ssd, intel core i5 9th generation processor with 4gb nvidia geforce gtx 1650 graphics card. These were all the hardware requirements. As for the software, I used python 3 programming language, along with Jupyter notebook. Libraries used during these project were pandas, which was used for various operations performed on Dataframe, numpy, which was used for various mathematical functions used during calculations, seaborn and matplotlib was used for plotting different graphs which gave a lot of information about the data and helped in performing the E.D.A. process. Finally the most important library used was sklearn. Sklearn gave various tools for preprocessing and scaling of the data, splitting the data into training and testing, for K-Fold cross validation, all the machine learning models were a boon from sklearn along with all the performation evaluating tools for the machine learning models.

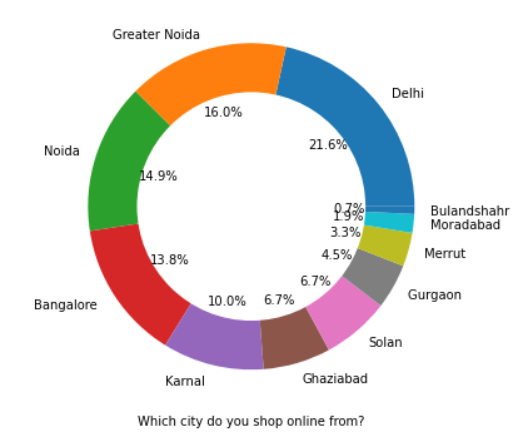
**Model/s Development and Evaluation**

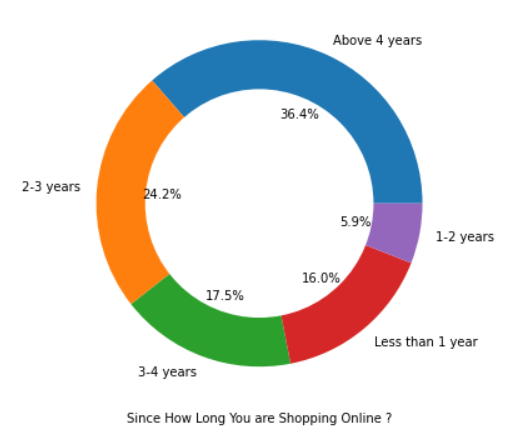
* Visualizations

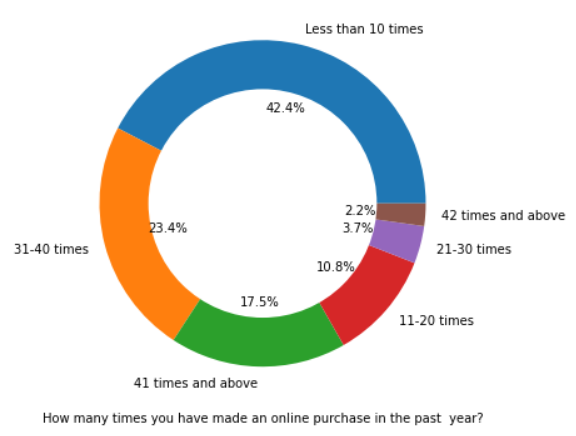
For visualization matplotlib and seaborn library was used.



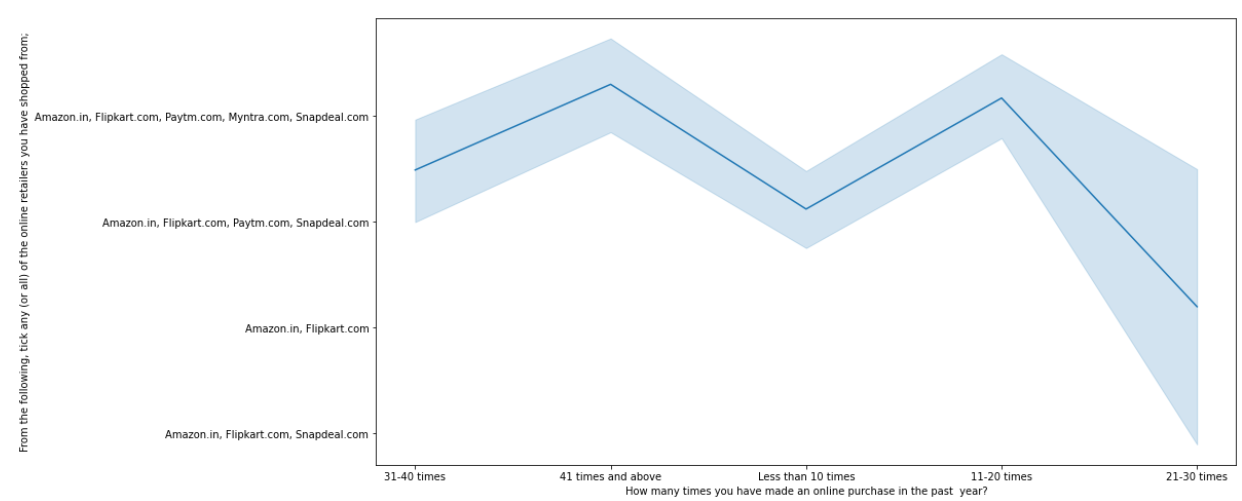




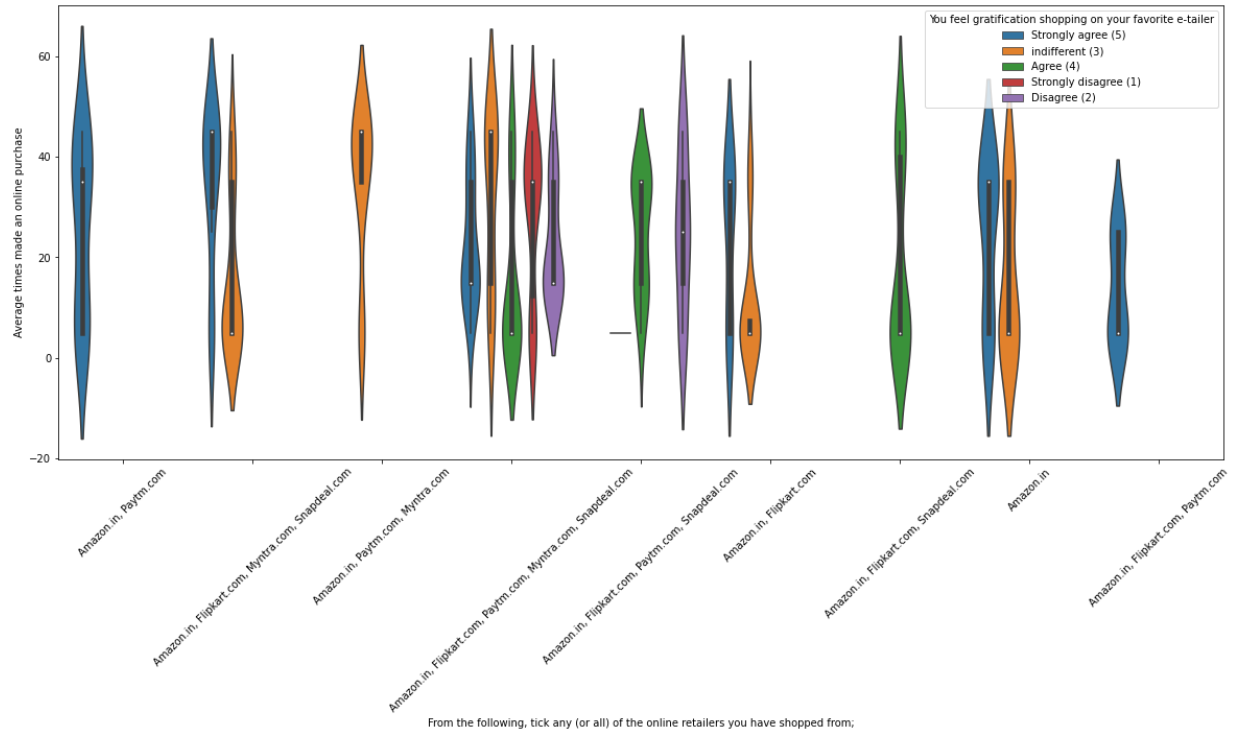




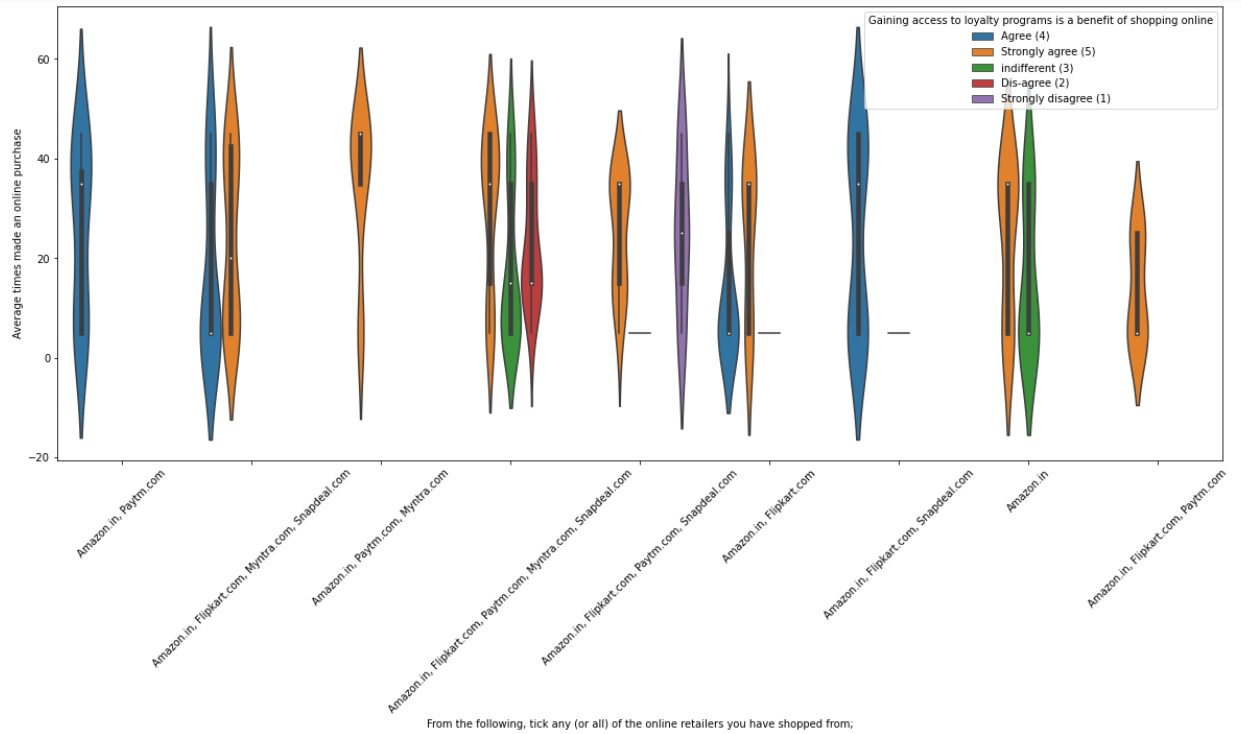
* There is double the number of women than men who have taken this survey.
* Most of the people are in their 30's followed by 20's, teenagers and senior citizen are the least in number.
* Most of the people belong from Delhi, Noida and banglore, ambiguity can also be seen as noida has two categories (noida and greater noida) which need to be handled
* Most of the people shopping online have been shopping from a long time.
* Majority of people shop online 10 times a year, amiguity can also be seen for range 42 times and above which needs to be handled



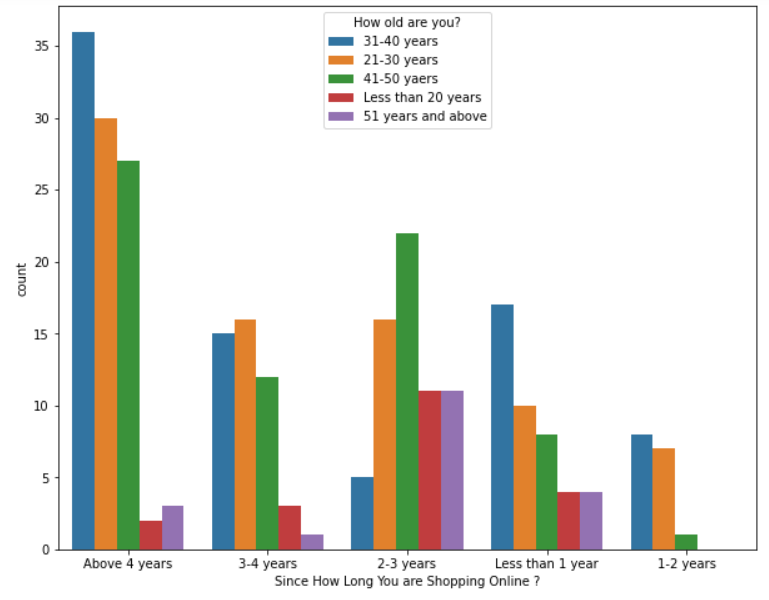
Heavy shoppers who shop more than 41 times a year shop from all the online brands, some of the people who shop for 32-40 and less than 10 times a year seem to exclude myntra. People shop from Amazon and flipkart whatever be the case.



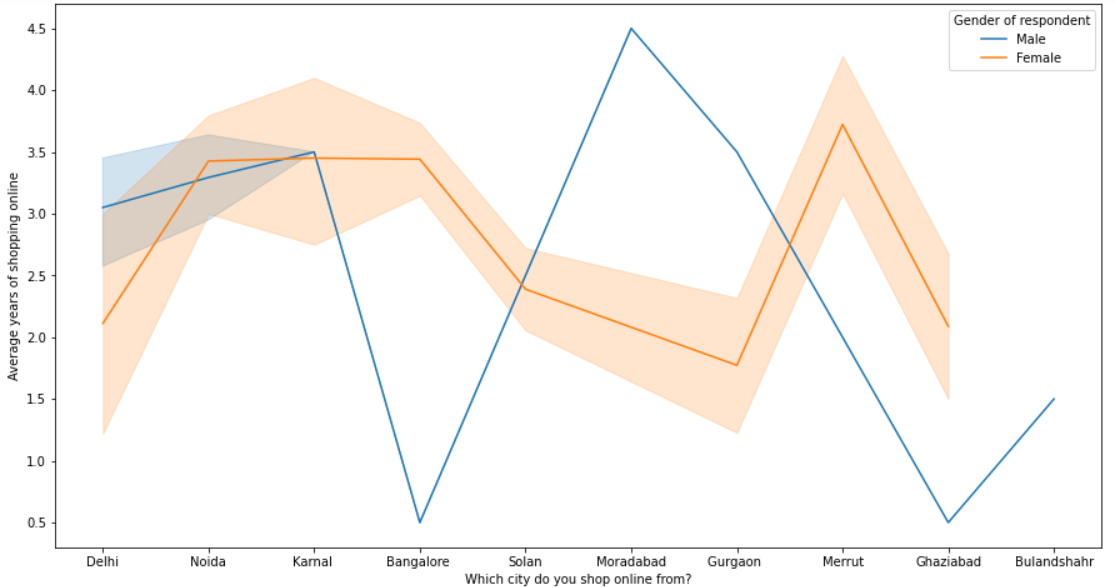
Almost all the people who have shopped from amazon, flipkart and paytm are satisfied. People who shop from a more number of online brands dosent seem to be satisfied.



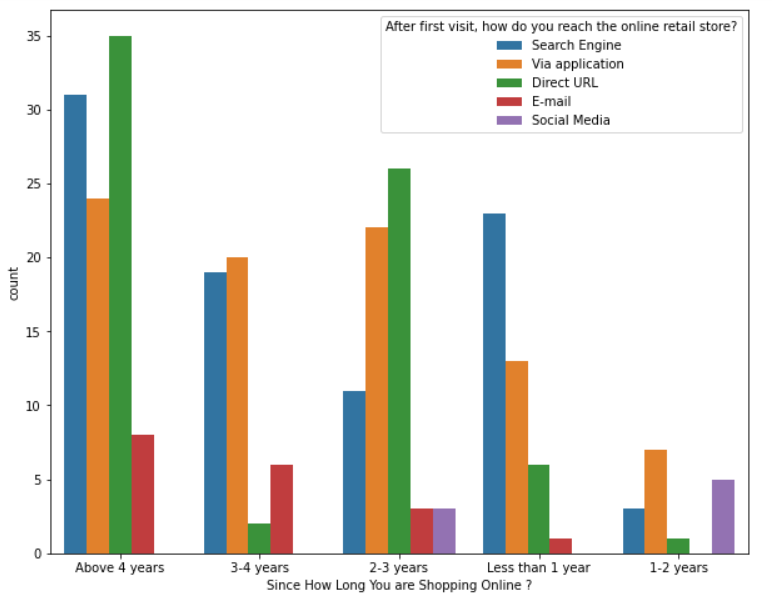
People shopping from amazon and paytm are getting benefits from the loyalty points, flipkart and sanpdeal also seem to give such benefits but people who shop from almost everywhere disagree with this statement too



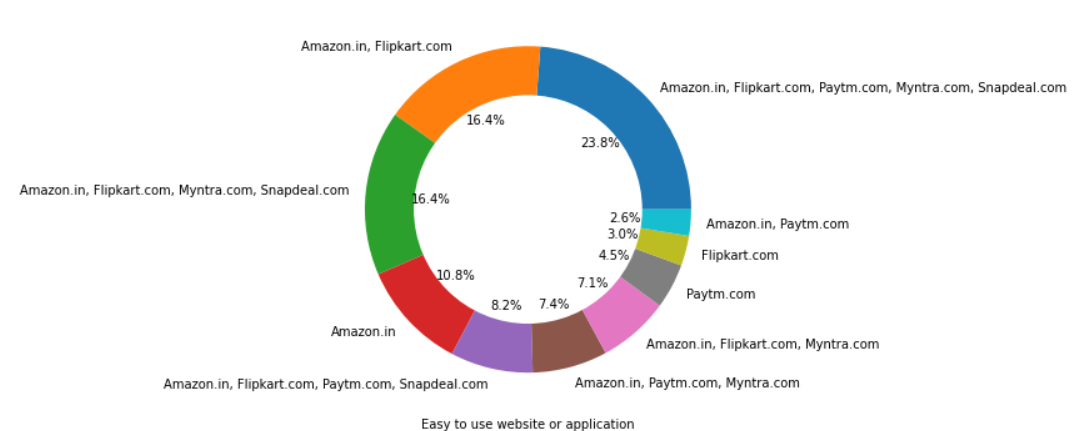
Highest number of people have been shopping online for above 4 years except for the age group below 20 years and above 50 years. People who are shopping online for 1-2 years does not include teenagers and elder people.

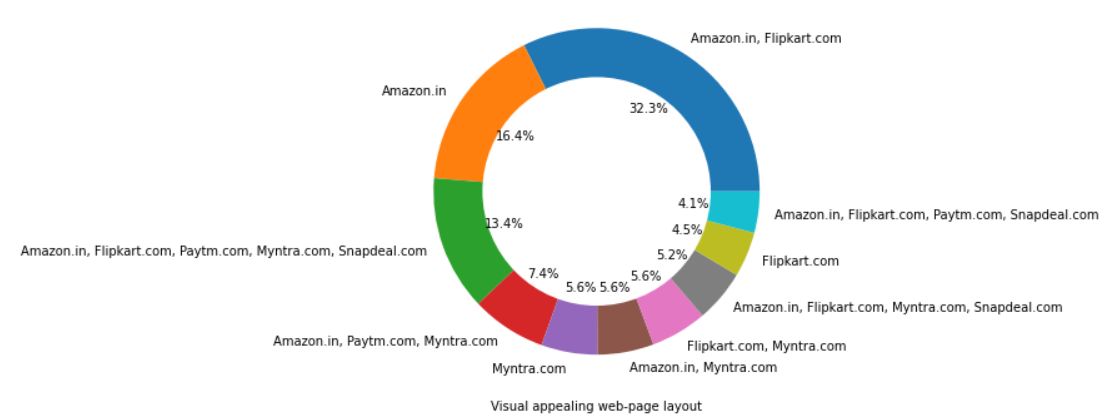


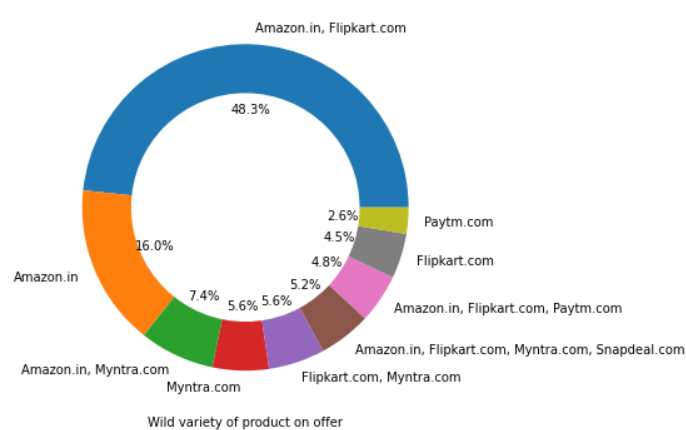
In lines, we can see that density of female customers is more than male. Men living in banglore and ghaziabad shop have shopped online for less than 1 year. Highest number of men shopping online belong from delhi and noida, while men from moradabad have been shopping online for the longest. Women from meerut and noida have shopped the longest.

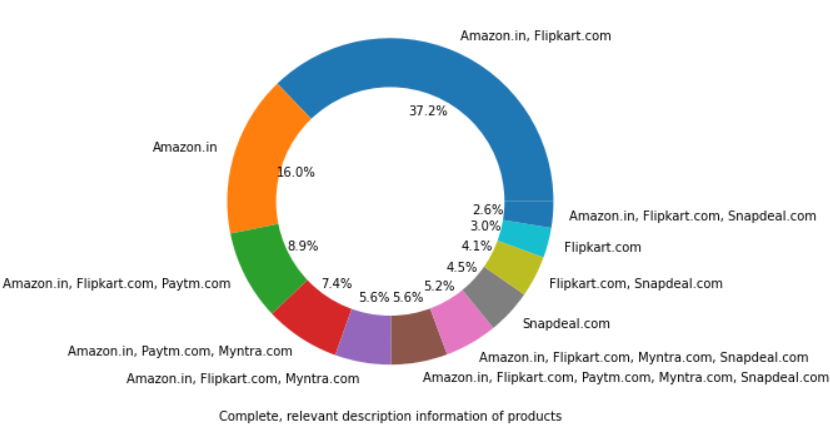


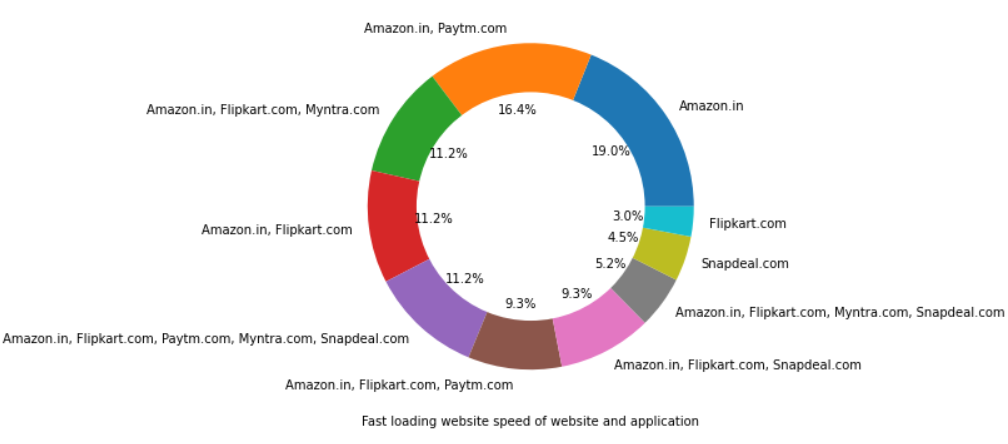
Even though people who are shopping online for more than 3 years donot use the application rather use search engine and direct url's in large number which indicates that online brands should update all their platforms rather than just application.

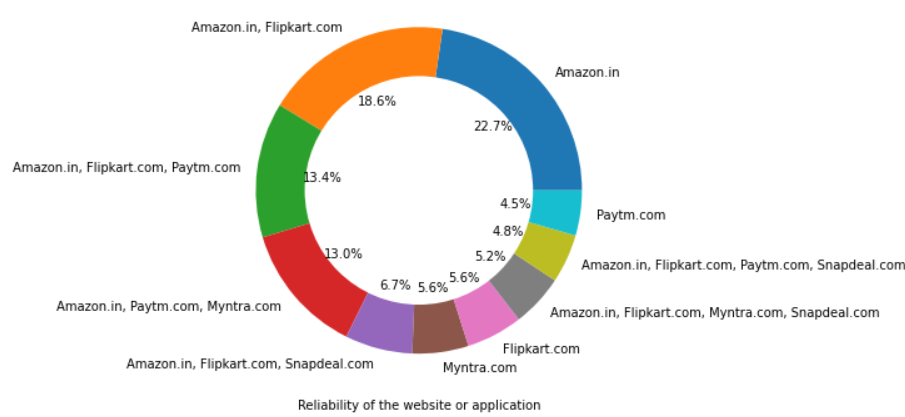


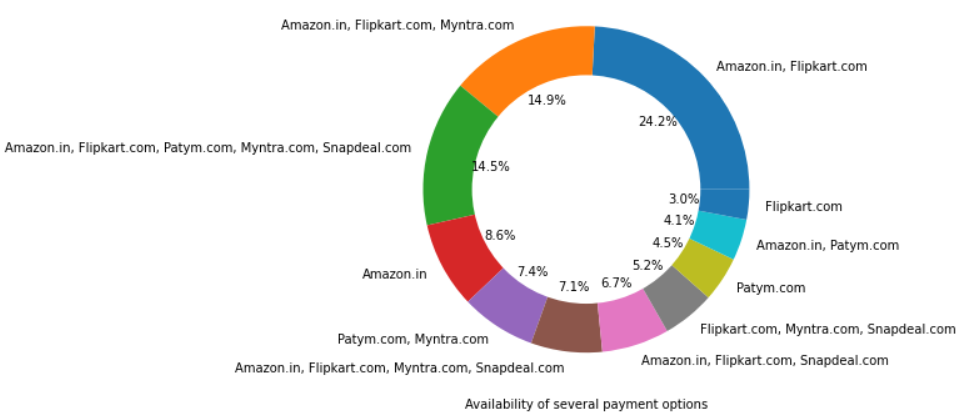


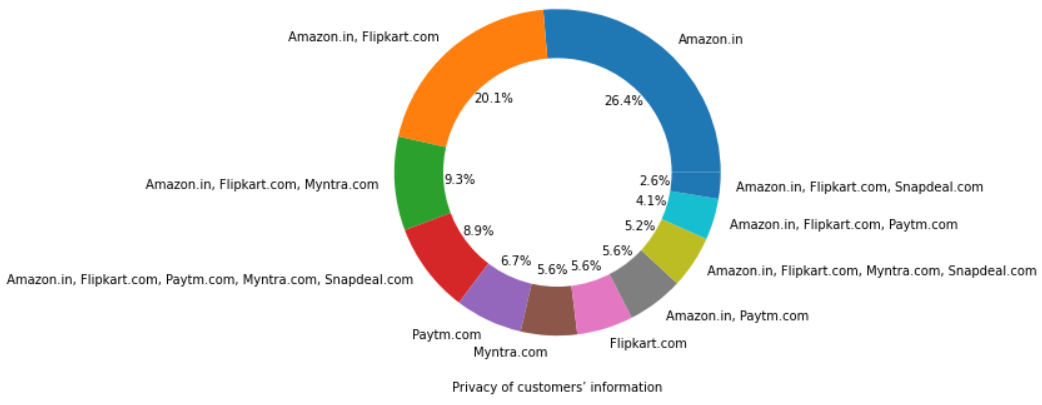


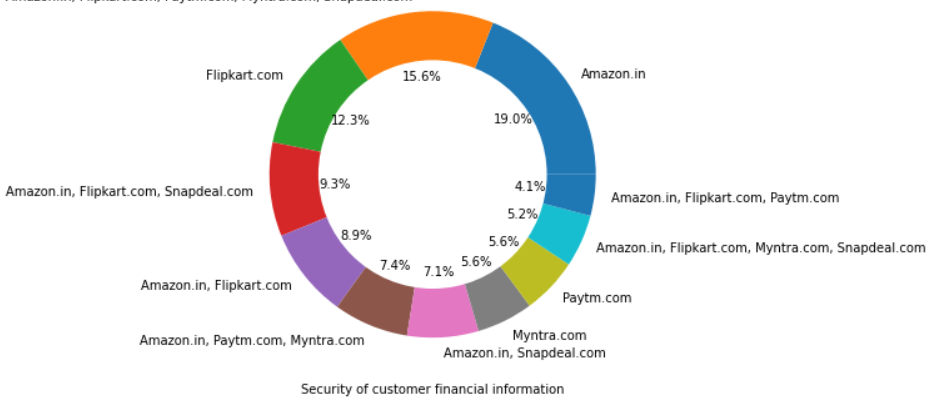


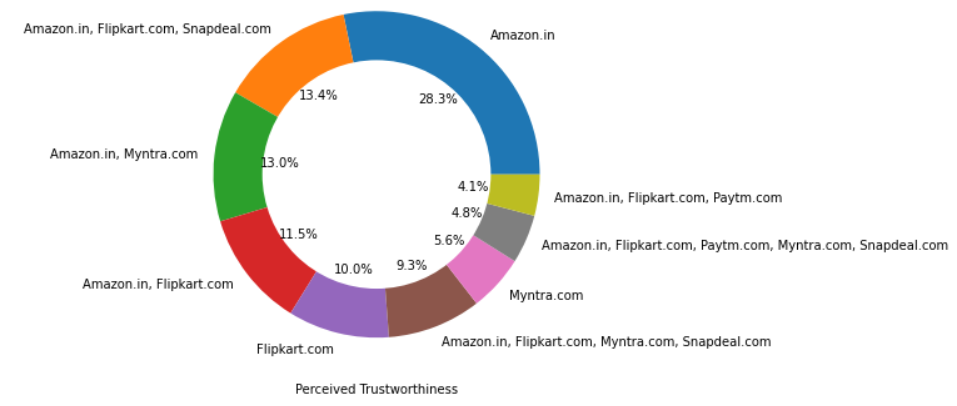


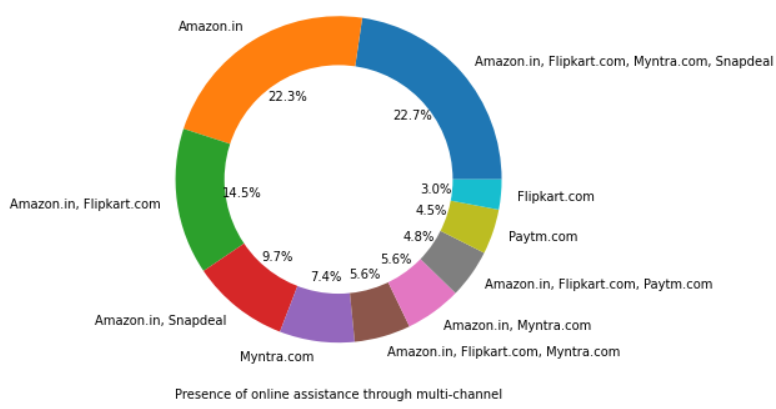




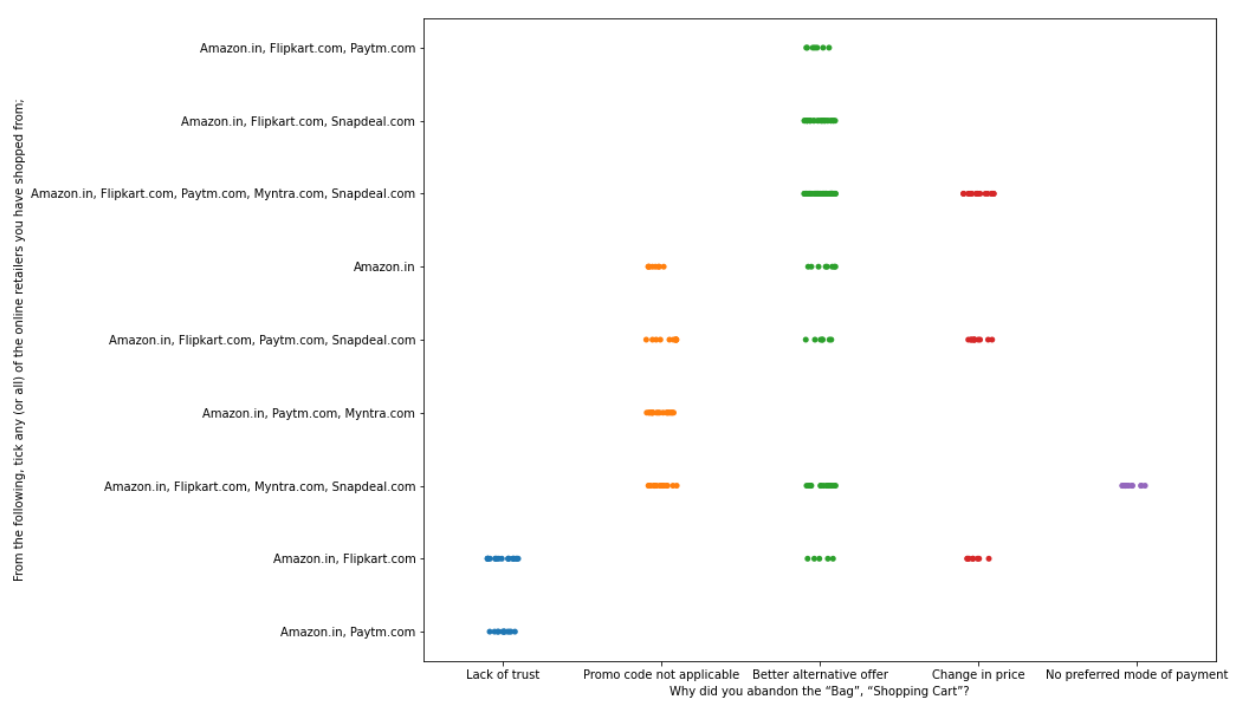




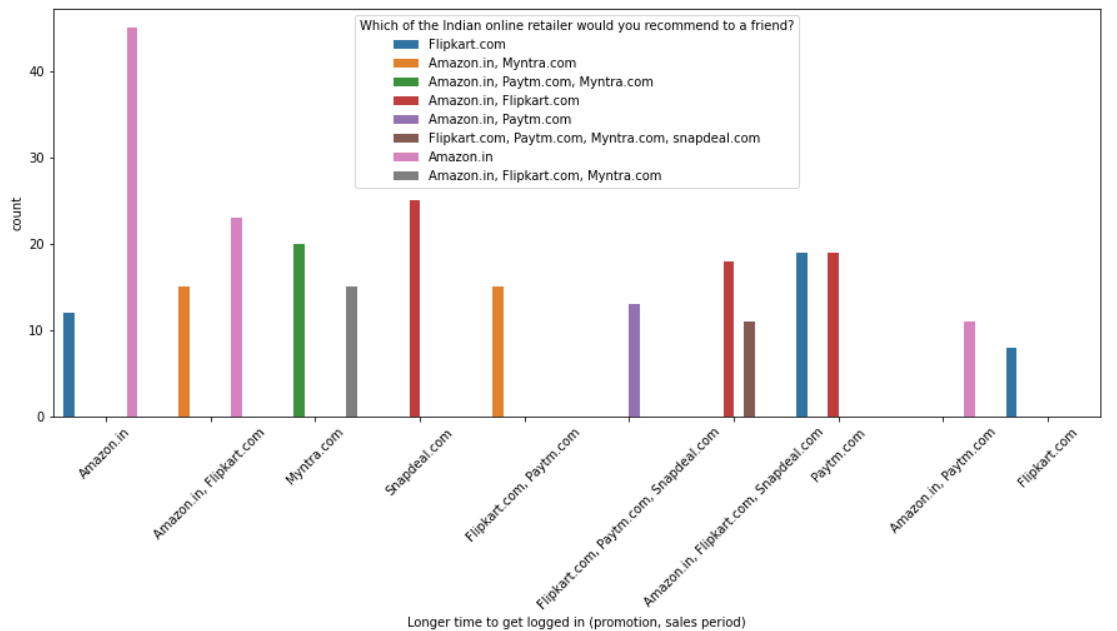


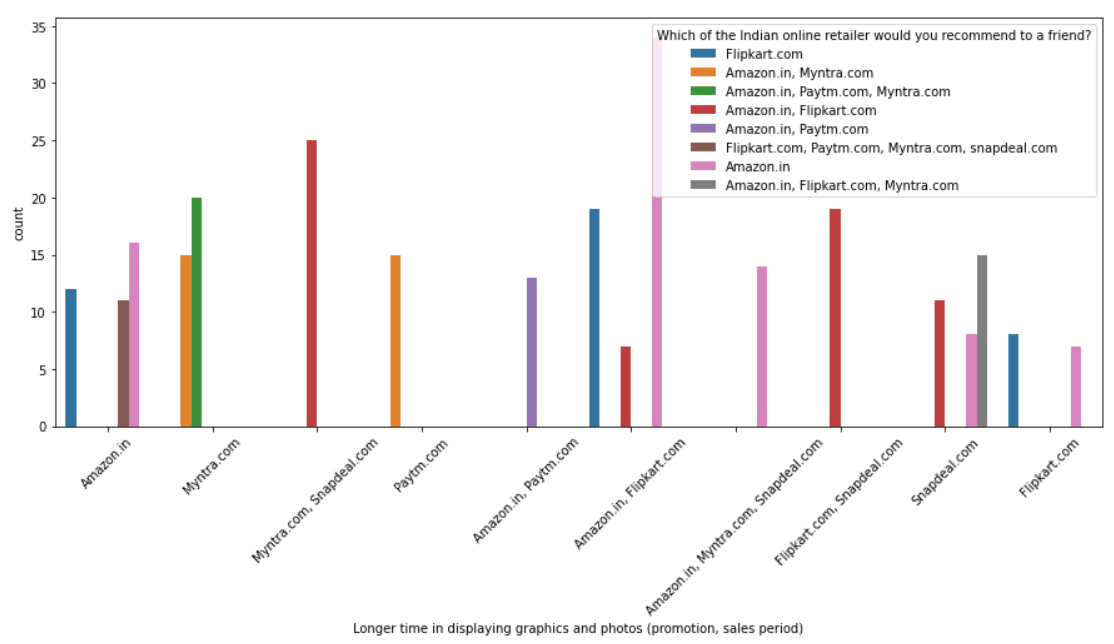


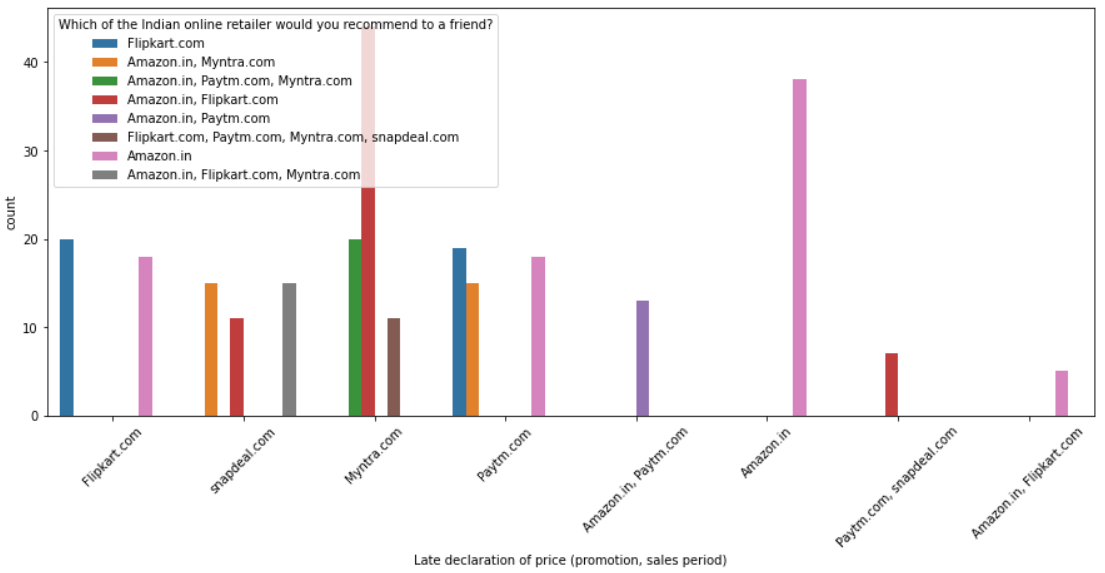
Even though people who are shopping online for more than 3 years donot use the application rather use search engine and direct url's in large number which indicates that online brands should update all their platforms rather than just application.

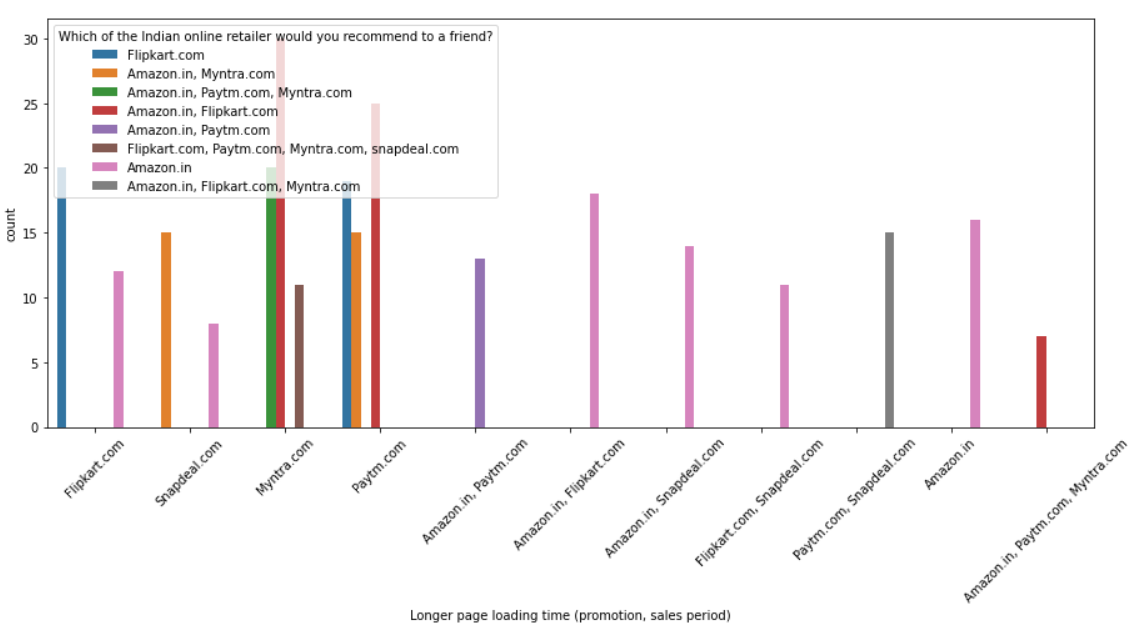


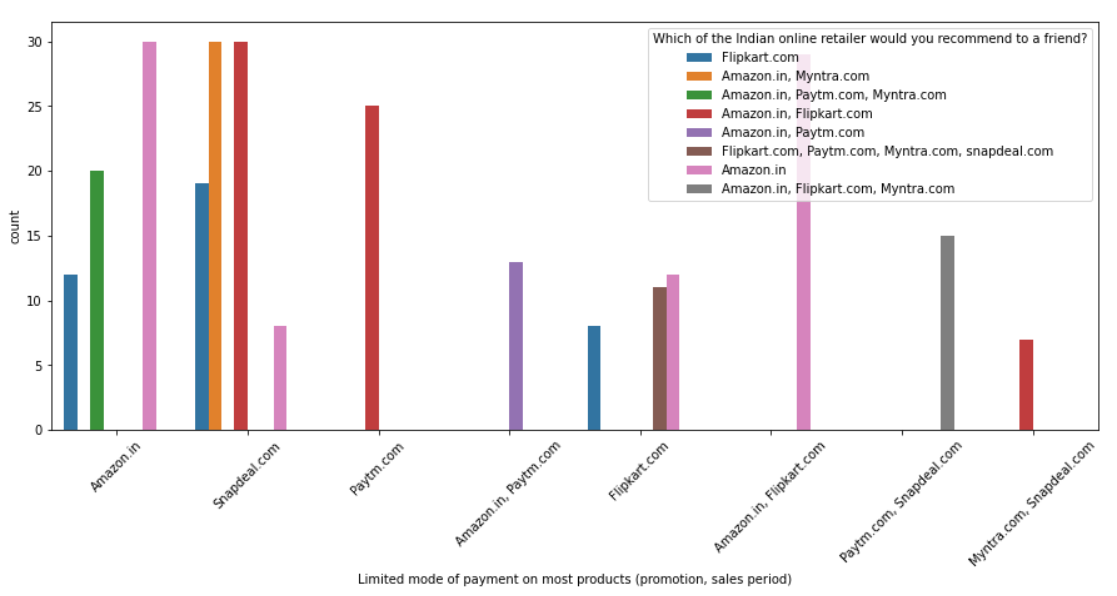
We can clearly see that most of the time people abandon the bag is beacuse they get a better alternative offer or promo code not applicable. There is also lack of trust seen in amazon, flipkart and paytm by some people.

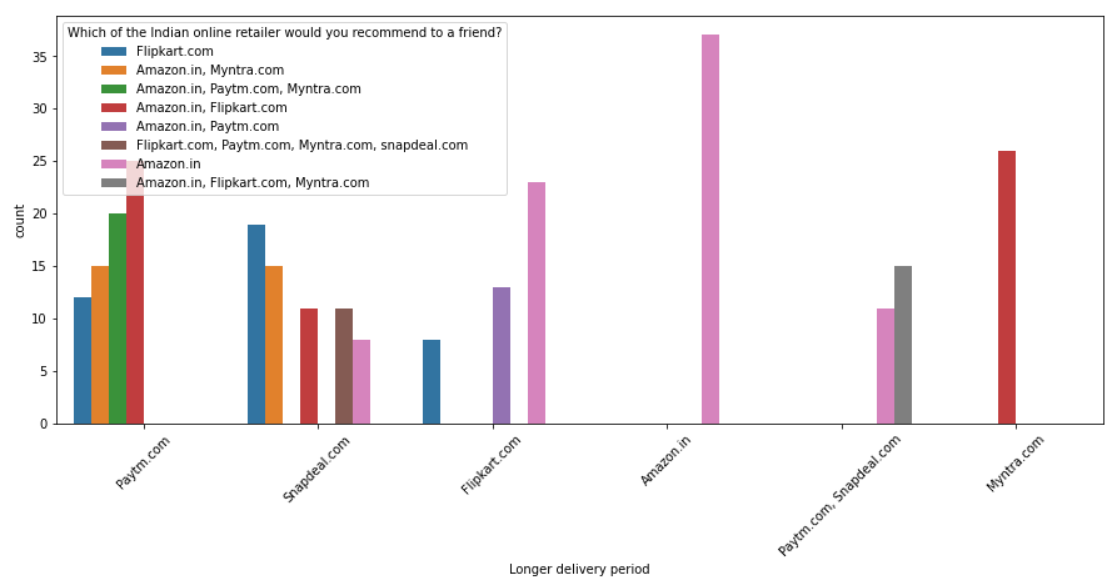


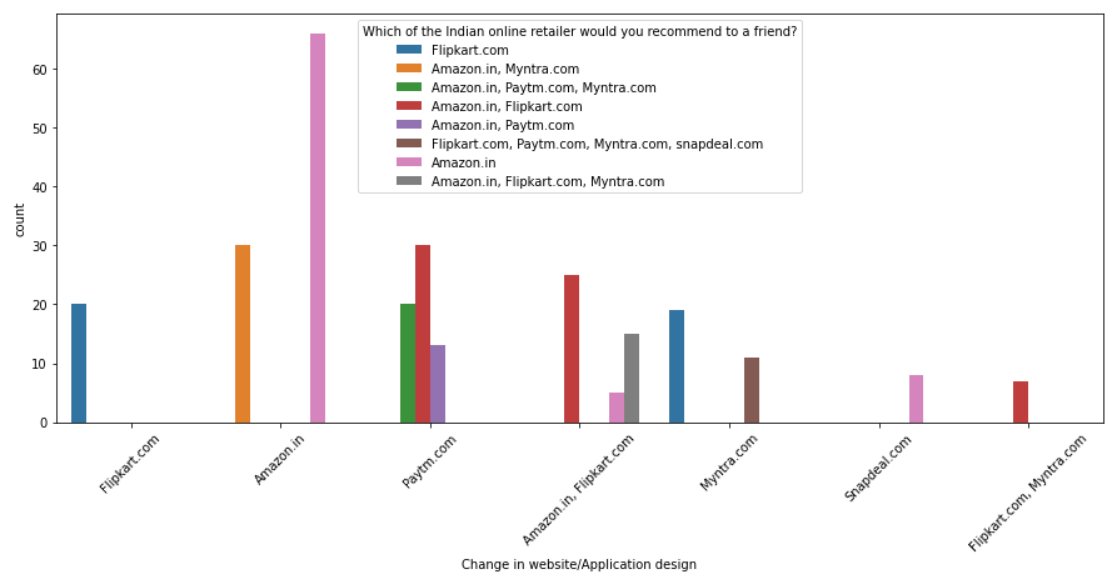


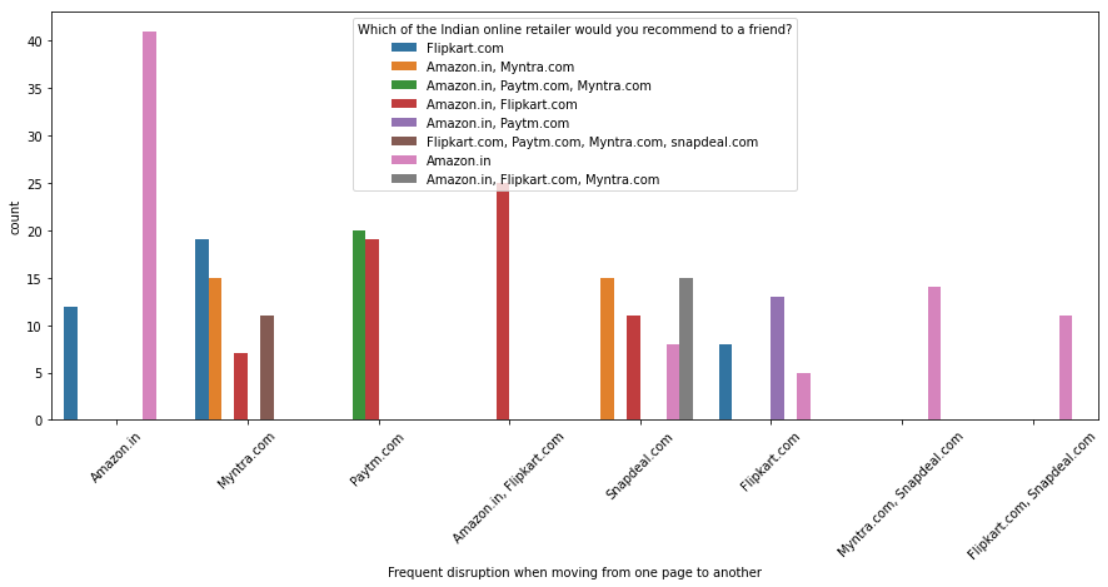












Customers seem to be more loyal to amazon, flipkart and paytm as even though many of them have given negative remarks about them still they would recommend these platforms to their friend

* Interpretation of the Results

Intention of Repeat purchase: If loyalty is defined as an uninterrupted sequence many purchases of the same brand, according to the marketing literature, a minimum of three to four successive repeat purchases are enough to talk about loyalty. Behaviour on the internet is much less stable over time. Indeed, the cyber consumer has more alternatives to choose from, the cost of change is relatively small and information about the sites is available at low cost. The consumer is no longer a mere purchaser of the product or user of the site, but a true partner with whom we can cooperate and be willing to make sacrifices in the short run. Online Retailing: A 2012 global online survey by Nelsen suggests that about 875 million consumers across the world prefer to shop online. The report confirms that 85% internet users shopped online. According to Google India, out of more than 100 million internet users, half make online purchases and the number is growing ever year. On line retail in India is projected to grow to 200 billion dollars by 2025. Brand image: Brand image is the soul of the product or service. It translates into customers, makes them believe in the products and helps them to make purchase decisions. The brand image affects consumer cognition of the product or service quality. This is why enterprise work hard on it and maintain their brand image.

Brand satisfaction: Satisfaction is necessary but not sufficient component of loyalty, satisfaction is a prior to loyalty. The notion of satisfaction is considered here as an indirect source of loyalty. Brand Satisfaction refers to transactional measures focusing on a discrete incident construct resulting from a service of transactions. It is an antecedent of trust, with increase in satisfaction leading to increase in trust, which in turn leads to loyalty.

Loyalty: A deeply held commitment to reuse the services of a company is called loyalty. The consumer brand relationship is critical to the building of loyalty. It is important for several reasons. First, it results in higher sales volume. Second, companies with loyal customers don’t have to spend as much money on marketing the product. This type of loyalty can be a great asset to the firm: customers are willing to pay higher prices, they may cost less to serve, and can bring new customers to the firm. Businesses have to exert significant effort to facilitate loyalty. Companies need to convince potential customers that their product has a significant advantage over other products to justify consistent purchases.

Satisfaction: Research has shown that consumer behaviour is difficult to predict. It studies features of individual consumers such as demographics and behavioural variables in an attempt to understand people’s wants. It also tries to assess influences on the consumer from groups such as family, friends, reference groups and society in general. The state of satisfaction depends on a number of both psychological and physical variables which correlate with satisfaction behaviours such as other products against which the customer can compare the products to shopping sites which match the predominant motives of customer have higher patronage and different product categories have different motives. Once the customer is satisfied with one e-retailer or brand he doesn’t generally change to another one. Satisfaction has generally been presented as an emotional state arising from the non confirmation of positive or negative initial expectations for the experience of possession or consumption

**CONCLUSION**

* Key Findings and Conclusions of the Study

The cost of the product, the reliability of the E-commerce company and the return policies all play an equally important role in deciding the buying behaviour of online customers. The cost is an important factor as it was the basic criteria used by online retailers to attract customers. The reliability of the E-commerce company is also important, as it is even required in offline retail. It is important because customers are paying online, so they need to be sure of security of the online transaction. The return policies are important because in online retail customer does not get to feel the product. Thus, he wants to be sure that it will be possible to return the product if he does not like it in real. Whereas, the logistics factor, which included Cash on delivery option, One day delivery and the quality of packaging plays a secondary role in this process though these are Must-be-quality as per KANO model. This is so because these all does not interfere with the real product and people believe that this is the basic value that E-commerce websites provide. 2. All the websites were not equally preferred by online customers. Amazon was the most preferred followed by Flipkart. This can be explained easily by previous result that we got. These two companies are most trusted in the industry and hence, have a huge reliability. Also, the sellers listed on these websites are generally from Tier 1 cities as compared to Snapdeal and PayTM which have more sellers from tier 2 and 3 cities. Also, these websites have the most lenient return policies as compared to others and also the time required to process a return is low for these.

* Learning Outcomes of the Study in respect of Data Science

Many things were picked up by me during the course of the project. In visualizations, I came to learn about the different ways I could implement different types of graphs along with there parameters. How could I make graphs more appealing and informative at the same time. I also learned which type of graph is suited to describe categorical variable, which is suited for nominal, ordinal and numerical types of features.

While in data cleaning process I learnt that removing outliers blindly can lead to loss of important data which can ultimately lead to lowering of eficiency of the model.

The biggest challenged I faced was to decide which method to use for the same work, like for when balancing the dataset should I go for over sampling or under sampling, or for scaling should I use Min max scaler or Standard scaler. After doing some research I found out there is no best method, you have to keep experimenting by applying permutation and combinations.

* Limitations of this work and Scope for Future Work

Although this study yields significant results, it has some limitations. Notably, the use of only the accepted loan applications is an example. The study may be expanded by considering the rejected loans, without the need of balancing the dataset. That would help to obtain more persuasive results. More complex analysis can be performed using feature-selection techniques to assign weights on features by their importance in the future studies.It would also be interesting to make a study concerning what metrics are the most relevant for this type of the problem. As mentioned previously, in this project the main metric all evaluations were analyzed by was accuracy score and mean of k-fold validation score, because the aim was to achieve a trade-off between the bias and the variance.

An assumption can also be made that if there is data available for longer time span as well as broader geography of clients, there is an interest to implement marco economical variables, which in turn might open some new insights about factors impacting default of a customer as well as what machine learning methods are more suitable for this type of a problem.