E-retail factors for customer activation and retention: A case study from Indian e-commerce customers

Submitted by:

Vijay Ashley Rodrigues K.

ACKNOWLEDGMENT

- This includes mentioning of all the references, research papers, data sources, professionals and other resources that helped you and guided you in completion of the project.
- I would like to thank FlipRobo Technologies for providing me this opportunity and guidance throughout the project and all the steps that are implemented.
- I have primarily referred to various articles scattered across various websites for the purpose of getting an idea on E-Commerce in general.
- I would like to thank the technical support team also for helping me out and reaching out to me on clearing all my doubts as early as possible
- I would like to thank my project SME Sajid Choudhary for providing the flexibility in time and also for giving us guidance in creating the project.
- My experience in E-Commerce also helped me start certain parts of the topic.
- The following are some of the articles I referred to in this project.

https://sumo.com/stories/ecommerce-success-stories

https://www.campaignmonitor.com/blog/email-marketing/12-effective-ecommerce-customer-retention-strategies/#:~:text=Customer%20loyalty%20and%20reward%20programs,even%20more%20with%20your%20brand.

https://www.bigcommerce.com/blog/amazon-competitors/

https://www.shopify.in/blog/best-ecommerce-platforms-india

INTRODUCTION

- ## Business Problem Framing

  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.

- ## Conceptual Background of the Domain Problem

  Describe the domain related concepts that you think will be useful for better understanding of the project.

  - This is an E-Commerce domain related problem and in general there are 4 types of domains within:
  - B2C: This refers to selling of products from a business entity to a customer i.e. and individual person.
  - B2B: When a business entity sells products or services to another business through online, it's B2B ecommerce. It could be wholesale, equipment's, manufacture to retailer etc.
  - C2B: This is where a customer or consumer give's their services to businesses. Eg: companies pay the customers to give ratings or reviews exchange for money.
  - C2C: When individual sells products or services to other individual online. Eg: Olx, Ebay where we can buy products from other customers directly.

- In our case we are referring to a B2C business model and it is the most common type of E-Commerce model.
- E-Commerce could be selling of products and could also be providing services. Eg: Web hosting, online memory management etc. But in our case, it's the physical products selling platform.

- Motivation for the Problem Undertaken

  There was a time when monopoly existed between E-Commerce giants. Often times, if you ask a person about E-Commerce, they immediately relate it to online shopping. But is it limited only to that? We know some key factors like service quality, system quality, information quality, trust and net benefit are necessary for the success of any E-commerce brand. But are they the only factors? Could there be a possibility or scope of finding some other attributes that either benefit or create a negative credibility of a brand? After all, customer satisfaction is something we cannot measure or assume directly.
  This is how I got a keen interest in understanding how and why the factors or principles that were once used in "Brick and Mortar" stores are gradually decreasing and are implemented with little or complete change for omnichannel or ecommerce domains.
  In this scenario, my objective is to determine whether the factors described as above truly live up to the mark or could there be other reasons also for the success or failure of customer satisfaction.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  Although there are many types of analytics, the dataset I worked on is of Descriptive Analytics.

  We have a dataset that narrates the past actions that has occurred with respect to the customer's purchasing satisfaction.

  This data inclines more towards all the Indian E-Commerce platform.

- Data Sources and their formats

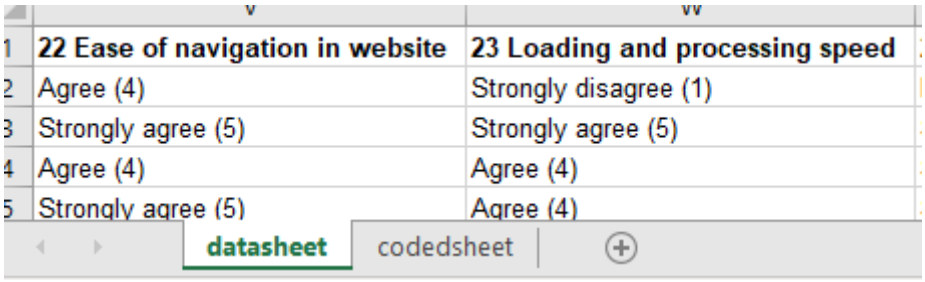
  The dataset is provided by FlipRobo and is available for academic purpose only and not for any kind of commercial activities.

  There are no details on the years during which this data was collected and is applied to and this appears to be a generic dataset.

  The dataset contains the customer purchasing preferences of Indian E-Tailers with 269 records (rows) and 71 features (columns).

  The file is in .xlsx format (Only readable by Microsoft Excel versions 2007 and onwards) with 2 tabs.

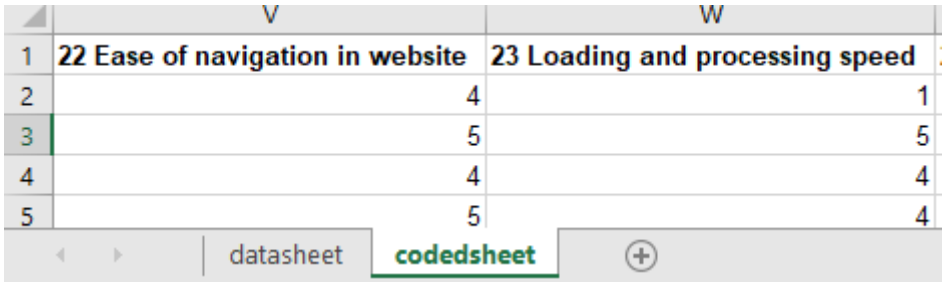
  The “datasheet” tab has all the actual information intact without any conversions or changes as shown below



	V	W
1	**22 Ease of navigation in website**	**23 Loading and processing speed**
2	Agree (4)	Strongly disagree (1)
3	Strongly agree (5)	Strongly agree (5)
4	Agree (4)	Agree (4)
5	Strongly agree (5)	Agree (4)

datasheet codedsheet

  The “codedsheet” tab is an exact replica of the former tab but couple of features are converted into numeric correspondent values as shown below



	V	W
1	**22 Ease of navigation in website**	**23 Loading and processing speed**
2	4	1
3	5	5
4	4	4
5	5	4

datasheet codedsheet

- Data Preprocessing Done

  - The dataset was is in the “Excel” format with 2 tabs when one tab is categorical in nature and the other tab was numerical to some extent.
  - I have converted this excel format to CSV in python itself for ease of data manipulation.

- For the purpose of EDA I used “datasheet” tab as the visualization I used need to have specific sub headings rather than just numerical values.
- For the purpose of encoding I have used LabelEncoder as the data was not required to be in order and I used the 2nd tab “codedsheet”
- After all the pre-processing, the data was split into x having all features and y with target feature and further into training and testing datasets for the purpose of building and predicting outputs.

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

  - Since this dataset is very small to consider a base dataset, the predictions that we make may not impact on a very large scale.
  - No features were dropped or deleted and all features were considered as the dataset is too small and I wanted to see if I can get best accuracy without the alteration.
  - From the dataset “Which of the Indian online retailer would you recommend to a friend?” is assumed to be the target variable.

- Hardware and Software Requirements and Tools Used

  Listing down the hardware and software requirements along with the tools, libraries and packages used. Describe all the software tools used along with a detailed description of tasks done with those tools.

  Hardware / Software specifications

  Windows 10 64bit

  Anaconda 2021.05

  Python version – Python 3.9.5

  Jupyter Notebook 6.4 and Google Colab

Pandas-profiling – package that performs simple EDAs for distribution of variables. This helped me give basic details about what the dataset consists of and its correlations with each other. The report is displayed using **.to_widgets()**

Sweetviz – Also an EDA package and is used in this project to generate quick visualizations that we may or may not consider throughout the project. It's then converted to an HTML file for ease of reference.

Lime – This package is an open-source ML model interpreter. Since we cannot see what really happens at ML level as it's serialized, I used this library to get a visual depiction of how the selected model works.

Pandas – This is used in the data manipulation, processing and cleaning and also to get description, stats and almost everywhere in the project.

Matplotlib and **Seaborn** - Majority of the data visualizations are plotted using Seaborn and to some extent only Matplotlib is used here.

Geopy : Used this geocoder library for getting the current location using zipcodes which then provided us latitude and longitude using **Nominatim** subsequently.

Plotly Express – This library is used for generating a web-based visualization and I have used it to map the country's location based on the output of geopy.geocoder.

LabelEncoder - I have used this **Skipy** library to convert all the non-ordered categorical data into numerical data.

train_test_split module from **sklearn.model_selection** to split the data into train and test and then used **StandardScaler** to bring the values to one level before imputing to model.

Warnings: I have used "ignore" to avoid the general errs that may occur and used "FutureWarning" to avoid errors that I got when running algorithms on Google Colab. To have a generic and efficient notebook file I used this as well.

SMOTE – I used this to check if there was an oversampling as the target column is imbalanced and this is will balance the classes of that target column.

Sciypy and **xgboost** - Used xgboost to import XGBClassifier and remaining algorithms including ensemble are part of Scipy.

Model/s Development and Evaluation

- Testing of Identified Approaches (Algorithms)

  For the model building I have considered the following algorithms for the training and testing.

  - DecisionTreeClassifier
  - XGBClassifier
  - HistGradientBoostingClassifier
  - ExtraTreeClassifier

- Run and Evaluate selected models

  - I have used a total of 4 ML algorithms to find the best and suited model.
  - I have used all 4 metrics i.e Accuracy, Precision, Recall and F1 score for all the algorithms. If you observe carefully, all show an 100% output.
  - In a typical dataset, it's practically not possible to get such percentage but since this is a very small and a perfect / clean dataset this output may have occurred.
  - I have used 1 ensemble algorithm and remaining are the regular classification algorithms.
  - But we cannot simply rely on these scores as we cannot have any scope for assumption. Hence post this I have also performed Cross Validation for all these algorithms to find the estimated performance metric when it's actually used in production.

1) DecisionTreeClassifier

In [157]:
```
from sklearn.tree import DecisionTreeClassifier

dt = DecisionTreeClassifier()
dt.fit(x_train,y_train)

y_pred = dt.predict(x_test)

print(accuracy_score(y_test, y_pred))
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

```
1.0
[[28  0  0  0  0  0  0  0]
 [ 0 24  0  0  0  0  0  0]
 [ 0  0 23  0  0  0  0  0]
 [ 0  0  0 19  0  0  0  0]
 [ 0  0  0  0 24  0  0  0]
 [ 0  0  0  0  0 23  0  0]
 [ 0  0  0  0  0  0 28  0]
 [ 0  0  0  0  0  0  0 21]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        28
           1       1.00      1.00      1.00        24
           2       1.00      1.00      1.00        23
           3       1.00      1.00      1.00        19
           4       1.00      1.00      1.00        24
           5       1.00      1.00      1.00        23
           6       1.00      1.00      1.00        28
           7       1.00      1.00      1.00        21

    accuracy                           1.00       190
   macro avg       1.00      1.00      1.00       190
weighted avg       1.00      1.00      1.00       190
```

2) XGBClassifier

In [171]:
```
from xgboost import XGBClassifier
xgb_reg = XGBClassifier(eval_metric='mlogloss')
xgb_reg.fit(x_train,y_train)

y_pred = xgb_reg.predict(x_test)

print(accuracy_score(y_test, y_pred))
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

```
1.0
[[28  0  0  0  0  0  0  0]
 [ 0 24  0  0  0  0  0  0]
 [ 0  0 23  0  0  0  0  0]
 [ 0  0  0 19  0  0  0  0]
 [ 0  0  0  0 24  0  0  0]
 [ 0  0  0  0  0 23  0  0]
 [ 0  0  0  0  0  0 28  0]
 [ 0  0  0  0  0  0  0 21]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        28
           1       1.00      1.00      1.00        24
           2       1.00      1.00      1.00        23
           3       1.00      1.00      1.00        19
           4       1.00      1.00      1.00        24
           5       1.00      1.00      1.00        23
           6       1.00      1.00      1.00        28
           7       1.00      1.00      1.00        21

    accuracy                           1.00       190
   macro avg       1.00      1.00      1.00       190
weighted avg       1.00      1.00      1.00       190
```

3) HistGradientBoostingClassifier

In [143]:
```
from sklearn.experimental import enable_hist_gradient_boosting
from sklearn.ensemble import HistGradientBoostingClassifier

hist_reg = HistGradientBoostingClassifier()
hist_reg.fit(x_train,y_train)

y_pred = hist_reg.predict(x_test)

print(accuracy_score(y_test, y_pred))
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

```
1.0
[[28  0  0  0  0  0  0  0]
 [ 0 24  0  0  0  0  0  0]
 [ 0  0 23  0  0  0  0  0]
 [ 0  0  0 19  0  0  0  0]
 [ 0  0  0  0 24  0  0  0]
 [ 0  0  0  0  0 23  0  0]
 [ 0  0  0  0  0  0 28  0]
 [ 0  0  0  0  0  0  0 21]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        28
           1       1.00      1.00      1.00        24
           2       1.00      1.00      1.00        23
           3       1.00      1.00      1.00        19
           4       1.00      1.00      1.00        24
           5       1.00      1.00      1.00        23
           6       1.00      1.00      1.00        28
           7       1.00      1.00      1.00        21

    accuracy                           1.00       190
   macro avg       1.00      1.00      1.00       190
weighted avg       1.00      1.00      1.00       190
```

4) ExtraTreeClassifier

In [149]:
```
from sklearn.tree import ExtraTreeClassifier

ext_reg = ExtraTreeClassifier()
ext_reg.fit(x_train,y_train)

y_pred = ext_reg.predict(x_test)

print(accuracy_score(y_test, y_pred))
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

```
1.0
[[28  0  0  0  0  0  0  0]
 [ 0 24  0  0  0  0  0  0]
 [ 0  0 23  0  0  0  0  0]
 [ 0  0  0 19  0  0  0  0]
 [ 0  0  0  0 24  0  0  0]
 [ 0  0  0  0  0 23  0  0]
 [ 0  0  0  0  0  0 28  0]
 [ 0  0  0  0  0  0  0 21]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        28
           1       1.00      1.00      1.00        24
           2       1.00      1.00      1.00        23
           3       1.00      1.00      1.00        19
           4       1.00      1.00      1.00        24
           5       1.00      1.00      1.00        23
           6       1.00      1.00      1.00        28
           7       1.00      1.00      1.00        21

    accuracy                           1.00       190
   macro avg       1.00      1.00      1.00       190
weighted avg       1.00      1.00      1.00       190
```

The below code shows us the cross validation performed over all the algorithms and I have used the CV values as 5.

If you observe the below scores, there is not much difference and is almost equal to the actual score we received using various algorithms.

```
In [135]: from sklearn.model_selection import cross_val_score

In [166]: scr = cross_val_score(dt, x, y, cv=5)
          print("Cross Validation score of DecisionTreeClasifier model is:", scr.mean())

          Cross Validation score of DecisionTreeClasifier model is: 0.9962962962962962

In [172]: scr = cross_val_score(xgb_reg, x, y, cv=5)
          print("Cross Validation score of XGBClassifier model is:", scr.mean())

          Cross Validation score of XGBClassifier model is: 0.9888888888888889

In [144]: scr = cross_val_score(hist_reg, x, y, cv=5)
          print("Cross Validation score of HistGradientBoostingClassifier model is:", scr.mean())

          Cross Validation score of HistGradientBoostingClassifier model is: 0.9962962962962962

In [152]: scr = cross_val_score(ext_reg, x, y, cv=5)
          print("Cross Validation score of ExtraTreeClassifier model is:", scr.mean())

          Cross Validation score of ExtraTreeClassifier model is: 0.9962962962962962
```

- From the above algorithms HistGradientBoostingClassifier, ExtraTreeClassifier and DecisionTreeClasifier have lowest difference between the accuracy score and cross validation score and are all same.
- XGBClassifier has difference value more than other algorithms hence will not consider this and I have used HistGradientBoostingClassifier in this scenario.
- Also I could not consider LogisticRegression as we have more than 2 classes in target columns

Sr.No.	Models used	Model Accuracy	Cross Validation	Difference output
1	DecisionTreeClassifier	100	0.996296296296296	99.0037037037037
2	XGBClassifier	100	0.988888888888888	99.0111111111111
3	HistGradientBoostingClassifier	100	0.996296296296296	99.0037037037037
4	ExtraTreeClassifier	100	0.996296296296296	99.0037037037037

- Let us try to tune the proposed model (HistGradientBoostingClassifier) to get better accuracy, if possible
- The "paramaters" have been selected from the skicit library and I have considered 4 paramaters

```
In [187]: parameters = {"loss":["auto", "binary_crossentropy", "categorical_crossentropy"],
                        "random_state":[50, 70, 100, 130],
                        "tol":[1e-1, 1e-5, 1e-7],
                        "max_iter":[50, 70, 100, 130]
                        }
```

- GridSearchCV is used to tune the parameters by fitting the same to the training dataset and used the best parameters after selection

```
In [189]: from sklearn.model_selection import GridSearchCV
          GCV = GridSearchCV(HistGradientBoostingClassifier(), parameters, cv=5)

In [190]: GCV.fit(x_train, y_train)

Out[190]: GridSearchCV(cv=5, estimator=HistGradientBoostingClassifier(),
                       param_grid={'loss': ['auto', 'binary_crossentropy',
                                            'categorical_crossentropy'],
                                   'max_iter': [50, 70, 100, 130],
                                   'random_state': [50, 70, 100, 130],
                                   'tol': [0.1, 1e-05, 1e-07]})

In [191]: GCV.best_params_

Out[191]: {'loss': 'auto', 'max_iter': 50, 'random_state': 50, 'tol': 0.1}
```

best_params_

- It’s observed that the model accuracy is 100% for this dataset even after Hyperparameter tuning and has not reduced.

```
In [192]: mod_hist_class = HistGradientBoostingClassifier(loss= "auto", max_iter= 50, random_state= 50, tol =0.1)

          mod_hist_class.fit(x_train,y_train)
          pred = mod_hist_class.predict(x_test)
          print(accuracy_score(y_test,pred)*100)

          100.0
```

- Key Metrics for success in solving problem under consideration

    - Using sklearn.metrics I have used accuracy_score, confusion_matrix, classification_report to check for all possible
    - I have used SMOTE technique for balance the target class so that we can have a better and unbiased accuracy.

I will proceed with SMOTE technique for Over Sampling of dataset

Although very few amount of values are imbalanced, I would still consider it to be imbalanced dataset because, the accuracy of the model after balancing the dataset was much better than the unbalanced dataset.

In [13]:
```
#Handling class imbalance problem ny oversampling the moinority class

from imblearn.over_sampling import SMOTE
SM = SMOTE()
x_over, y_over = SM.fit_resample(x,y)
```

Balanced dataset after SMOTE

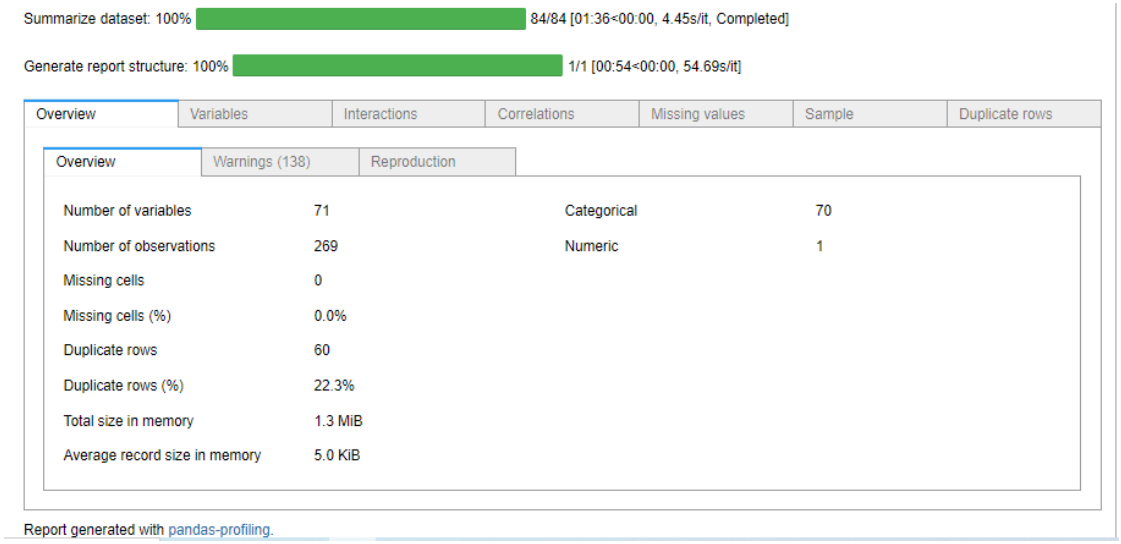
In [14]:
```
y_over.value_counts()
```

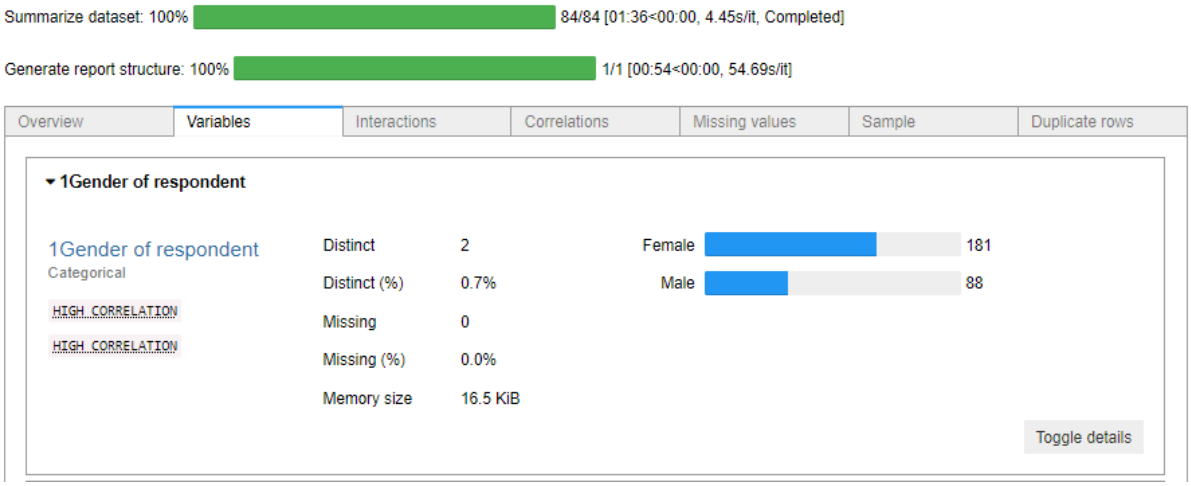
Out[14]:
```
0    79
1    79
2    79
3    79
4    79
5    79
6    79
7    79
Name: Which of the Indian online retailer would you recommend to a friend?, dtype: int64
```

- Visualizations

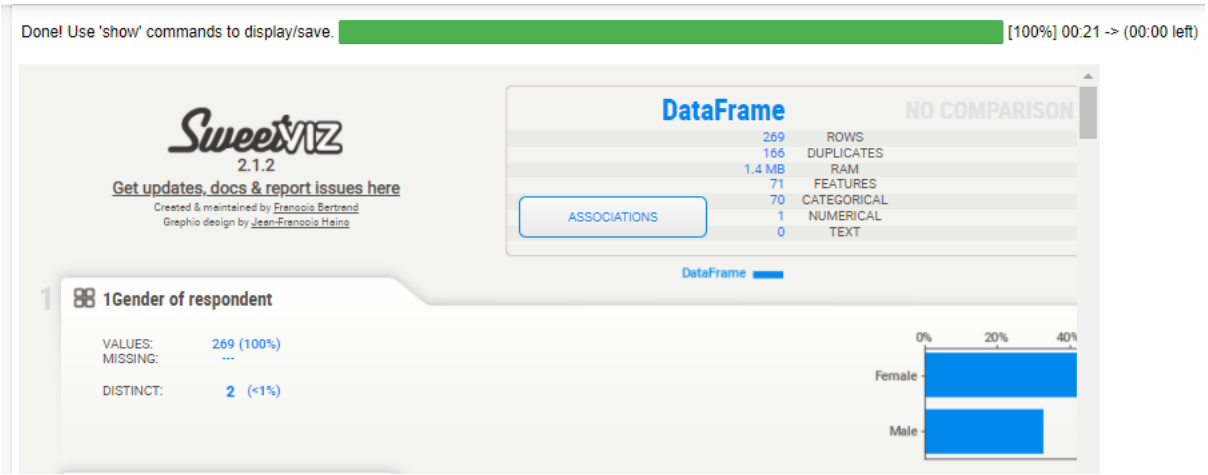
Explore the dataset with Pandas Profiling

The following library helps in analysing the dataset to give us some understanding on individual features





Exploring the dataset using Sweetviz library:

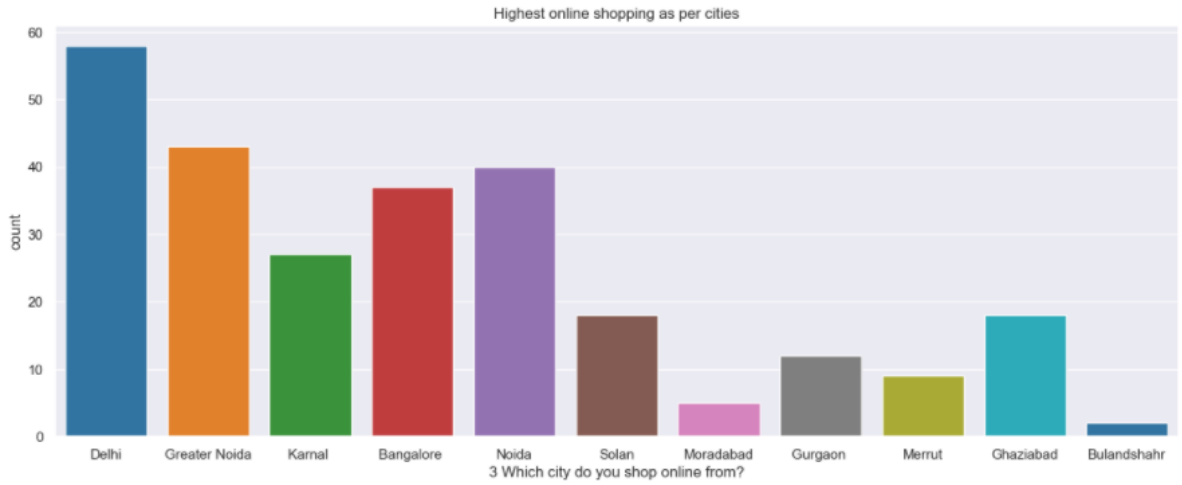


This heatmap shows us if there is any null values in the dataset. From this we can see that there are no null values

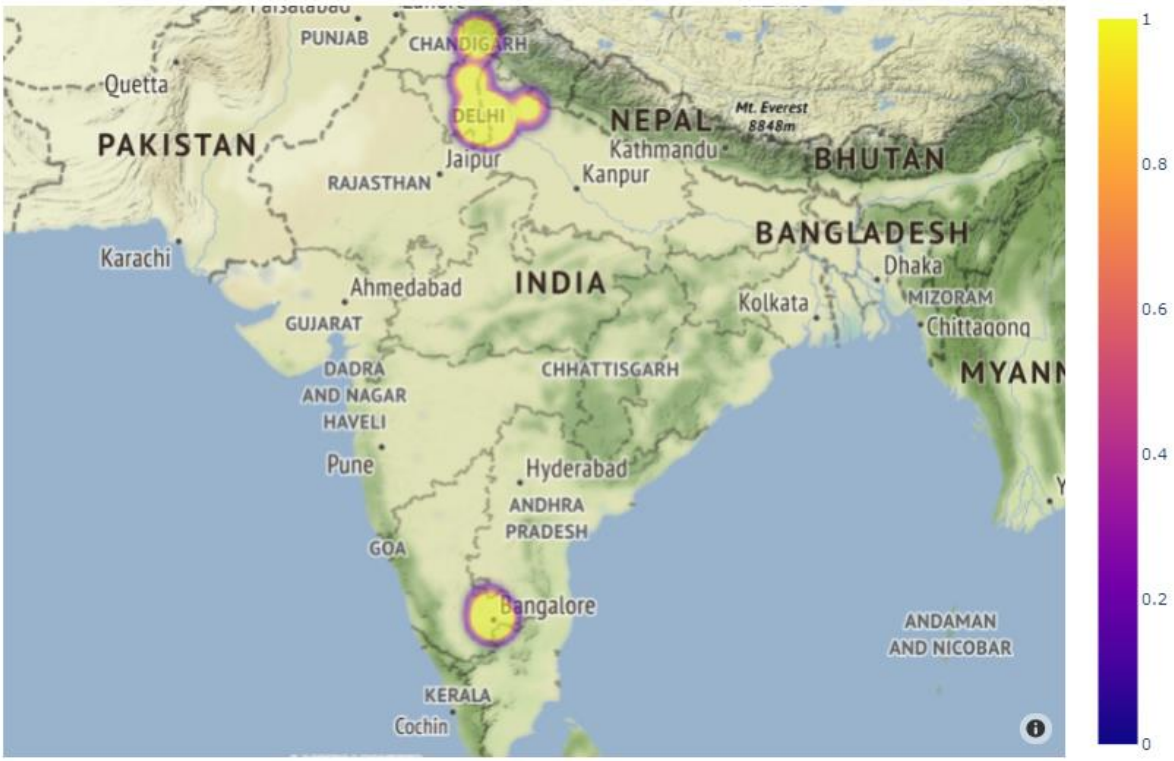
- The following dataset appears to have approximately 67% of Female respondents and approximately 32 % of Male respondents.
- But can we simply assume that Female purchase more online than Male? No we cannot. This dataset has only 269 records and it's merely difficult to conclude the majority in general.
- However we can assume that the data when this was recorded Female made more purchases than Male customers.
- Below is the percentage count followed with a plot to give us an idea of how it looks like in general.

- Let's observe if there is any relationship between online purchases and the city
- We can see that the respondents in this case appear to be majorly from "Delhi" as it has highest online purchases followed by "Greater Noida" , "Noida" and "Bangalore" seem to have high traffic in online purchases

  We can observe that these are metropolitan or semi metropolitan cities at leaset, known for multiple companies and IT sector in general. A lot of outsiders come here and population is also very high.
- This could in a way may have created a pavement for more online purchases as opposed to going to Brick-and-Mortar stores

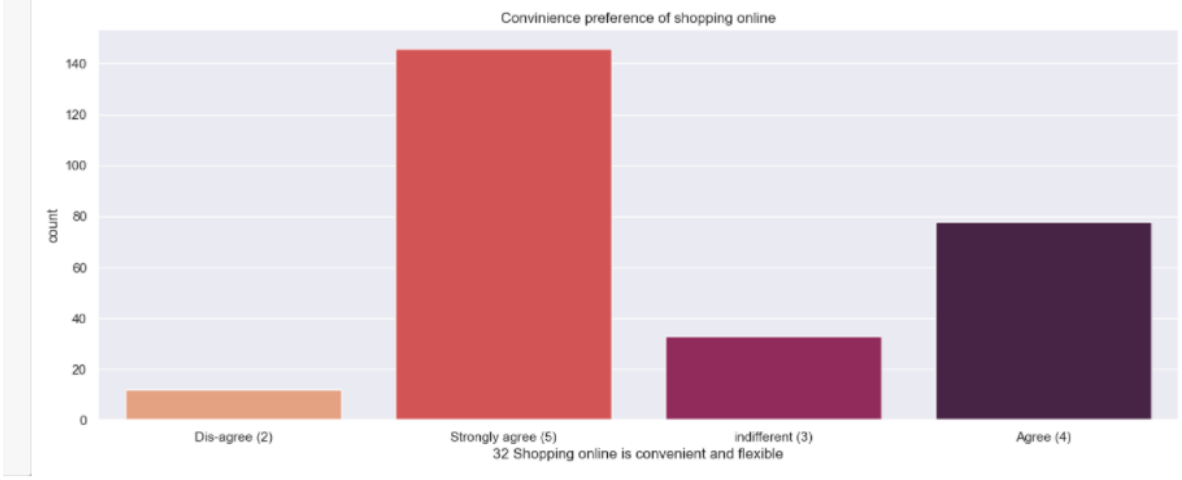


- For ease of reference, I have plotted a map of the location
- I have extracted the "latitude" and "longitude" as many as possible using the "zipcodes" from the dataset
- For this I have used Python's geopy library. Since all the zipcodes are in and around the limited cities, I have not extracted all the pincodes and only a handful of them.
- I have used "Plotly" for displaying the map using "density_mapbox"
- We can see the highlighted areas and majority of shopping have taken place in "North India" and in "South India" only Bangalore seems to be the city.

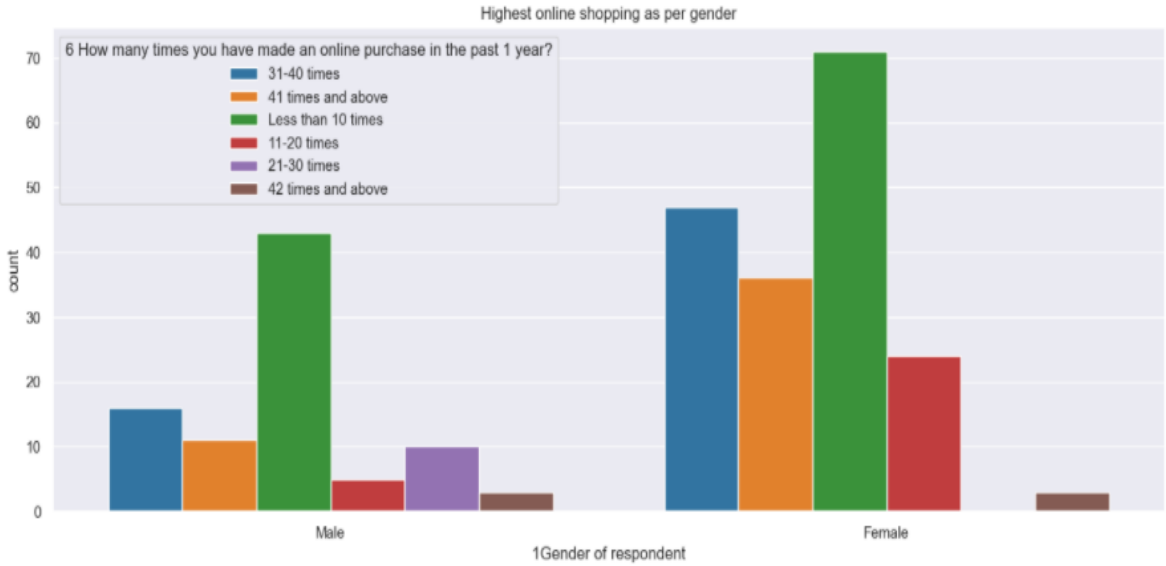


- Let's see if we can find any relationship between number of purchases made from online sources
- We can see in every case, majority of people have bought products less than 10 times atleast in 1 the last 1 year
- As the customer gives more number of years to online purchase, the quantity of purchase also increases
- We can see that customer who have been purchasing online for more than 4 years and above have bought atleast 31- 40 times

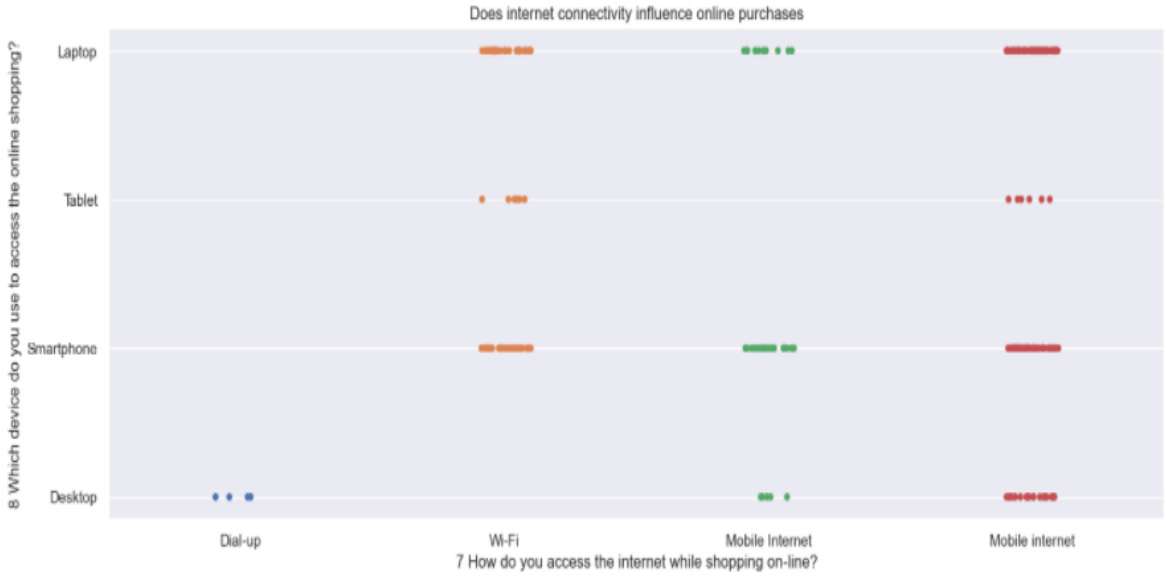
- We can see from the below plot that majority of customers prefer and are of the opinion that shopping online is convenient and flexible.
- At present, given the work conditions a person may not prefer going out. Going out for shopping also means spending more time and spending money to commute etc.
- It may not be a big deal, but imagine all you want is to buy a product that is relatively cheaper, or if products are manufacture or found in specific location, provinces etc. Travelling to that place for that product makes no sense.
- Also, the amount of time we may spend to investigate or compare physically can be reduced considerably by browsing online



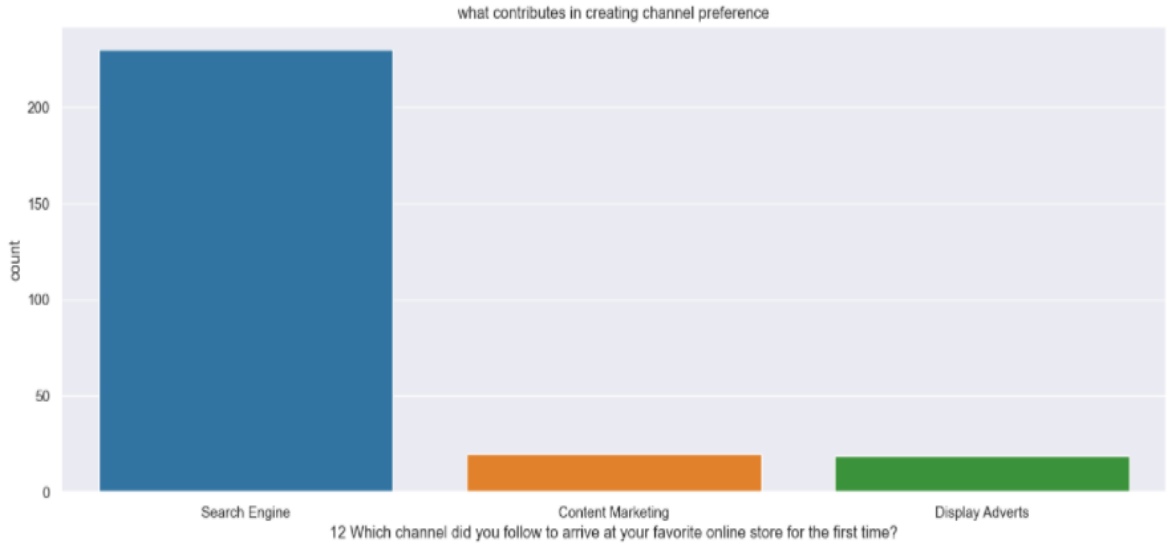
- From the previous graphs it was found that we have more Female customers and we can see very few customers have bought more than 42 times and above in a year among both the genders.
- But the rate of purchase of items atleast "31- 40 times" or "less than 10 times" is more among "Women" than the total count of "Men"



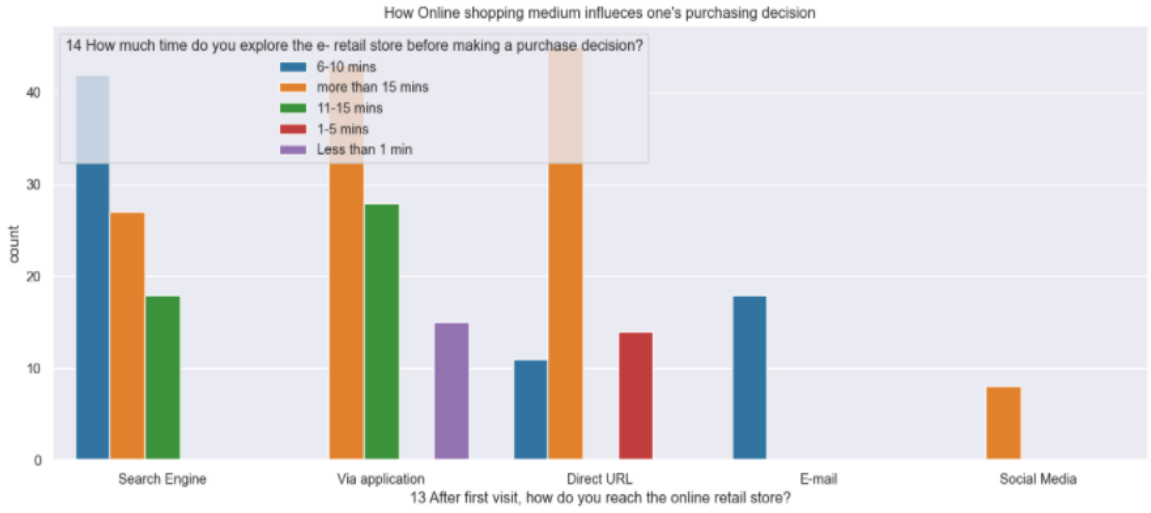
- Let's see what devices customers prefer using to make online purchases
- If we observe the points carefully, we can see for "Laptop" and "Smartphone" it’s almost full and "Smartphone" has more complete point of line.
- This indicates the customers prefer ordering / shopping online and prefer "Smartphones" followed by "Laptops". These are 2 devices that are very common to have with most of the people whether they are working class or not.
- We also know that internet services and basic smartphones in India are relatively cheaper compared to most of the countries including the already developed nations.
- The indirect supply and demand seems to boost one's needs towards online shopping, not only because one wants to , but also because one can simply do it anywhere and from any device.



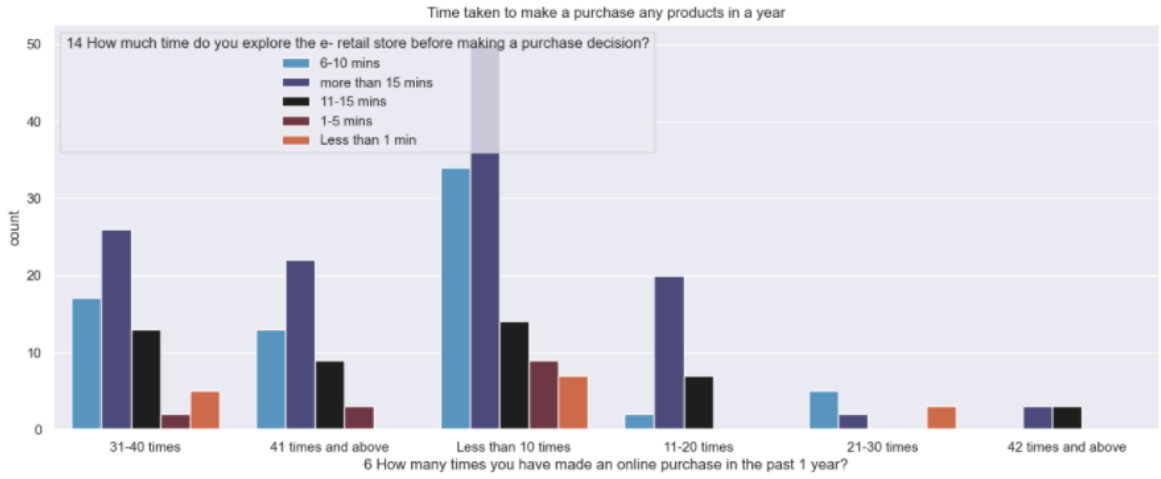
- Let's see if there is any relationship between channel and online store.
- We can see majority of people got to know about a specif store through search engines. These days it's quite common to search for an item online before buying products and the moment we do search, we get tonns of ads.
- The system understands our needs and in a way acts as a recommendation systems by targetting specific portions of that product by indirectly placing it in the search feeds.
- If we consider the other categories "Content Marketing" and "Display Adverts", we can see the count if very less and these categories could also be influencing customers through online medium



- From the below plot we can understand that a person shopping online would prefer browsing through "Search Engine" as mentioned earlier, "Via application" or through "Direct URL"
- We can see if a person purchase " within 1-5 mins" they may prefer going to specific online store directly. They know the services provided, they may have some kind of bias or inclination towards that E-commerce store and end up buying products impulsively or simply by trust.
- But majority of spend more than 15 mins before they make any purchases online as they spend time on exploring the channel, or may compare prices etc.
- We can also see that when it comes to "Search Engine" a perso may spend somewhere between " 6 - 10 mins". When we do not know what we are looking for and if we have a limited or irrelevant information, we tend to search through search engine like Google, Bing etc. which then directs us to relevant E-commerce stores

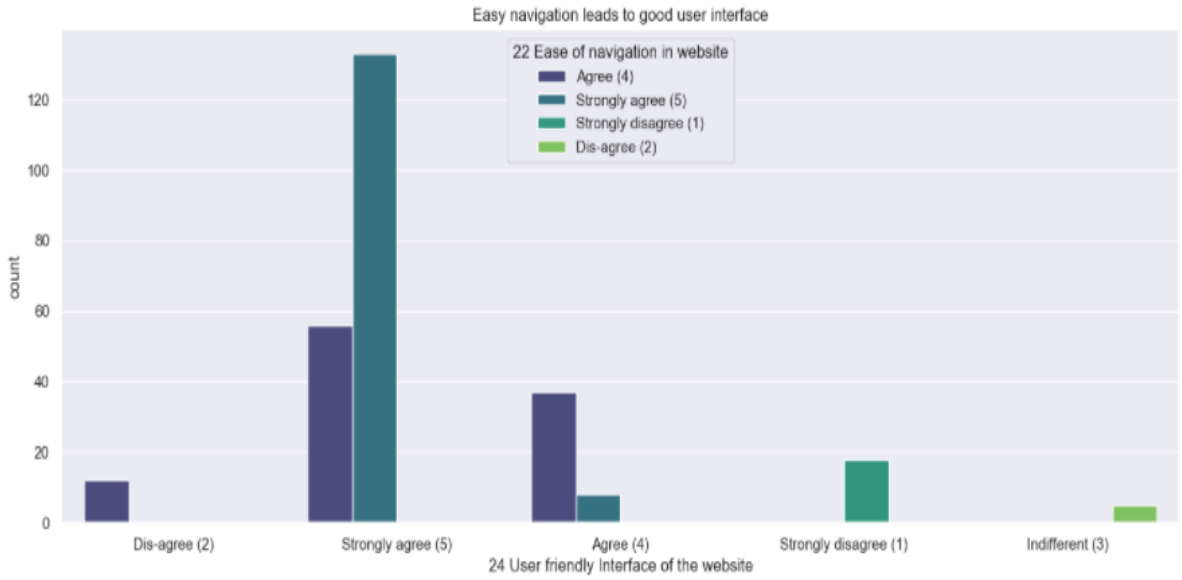


- Let's observe if a person shops online frequently or not
- We can see that a persons likely to spend more time if they seldom bought anything online. We can see from the below plot that a person may spend "more than 15 mins" if they have bought less than 10 products in a year.
- As and when a person's purchases increase the time spent on looking or comparing is less. The customer is familiar with the brand, they may have already done a proper research before shopping online.

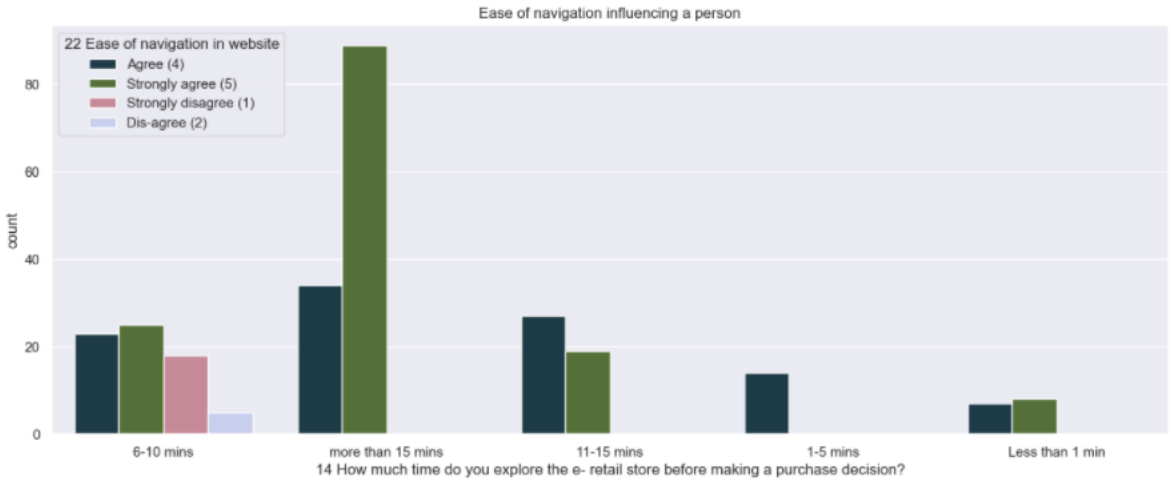


- Let's observe how often a person decides not to make a purchase right before making payment.
- We can see that majority of customers don't do this and only "Sometimes" end up not shopping online. This could be due to decline in Credit/Debit cards, no availability of EMI or availability of EMI on specific Banks cards only etc.
- We can also consider "Phonepe" or "GooglePay" as options as they are directly linked to Debit cards and although seem like E-Wallets they are much different.
- It's also a typical mindset amongst most Indians when shopping online they don't prefer online payment for expensive items. There could be scenarios where there is no cash on delivery or even chances where a person's online transaction is cancelled due to daily limit

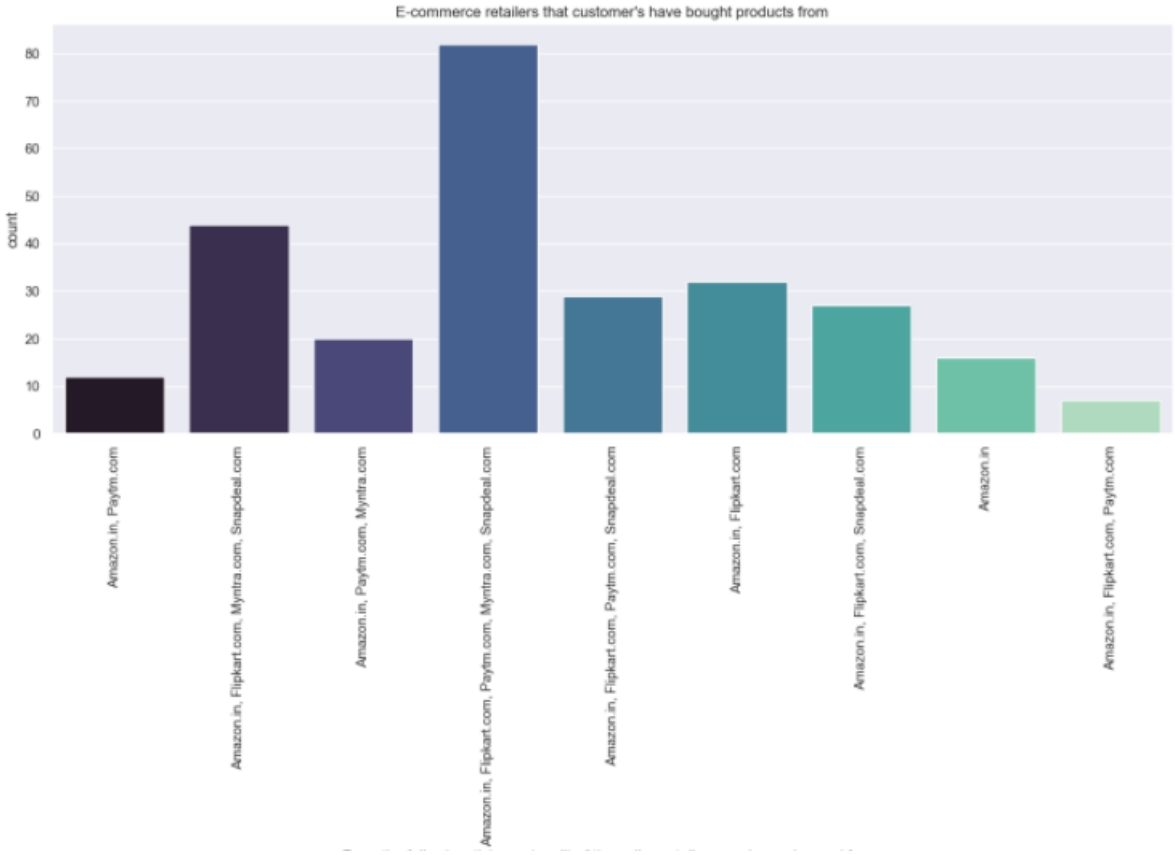
- Let's see how a person responds to the usage of website.
- We can see a person prefers a very good user interface that in turn helps in ease of navigation.
- Today we have multiple E-Commerce channels but still we pick only handful of them for most of our shopping needs.
- A good-looking website with proper navigation that can include selection, comparison, etc till the payment mode will help in creating a good user interface.



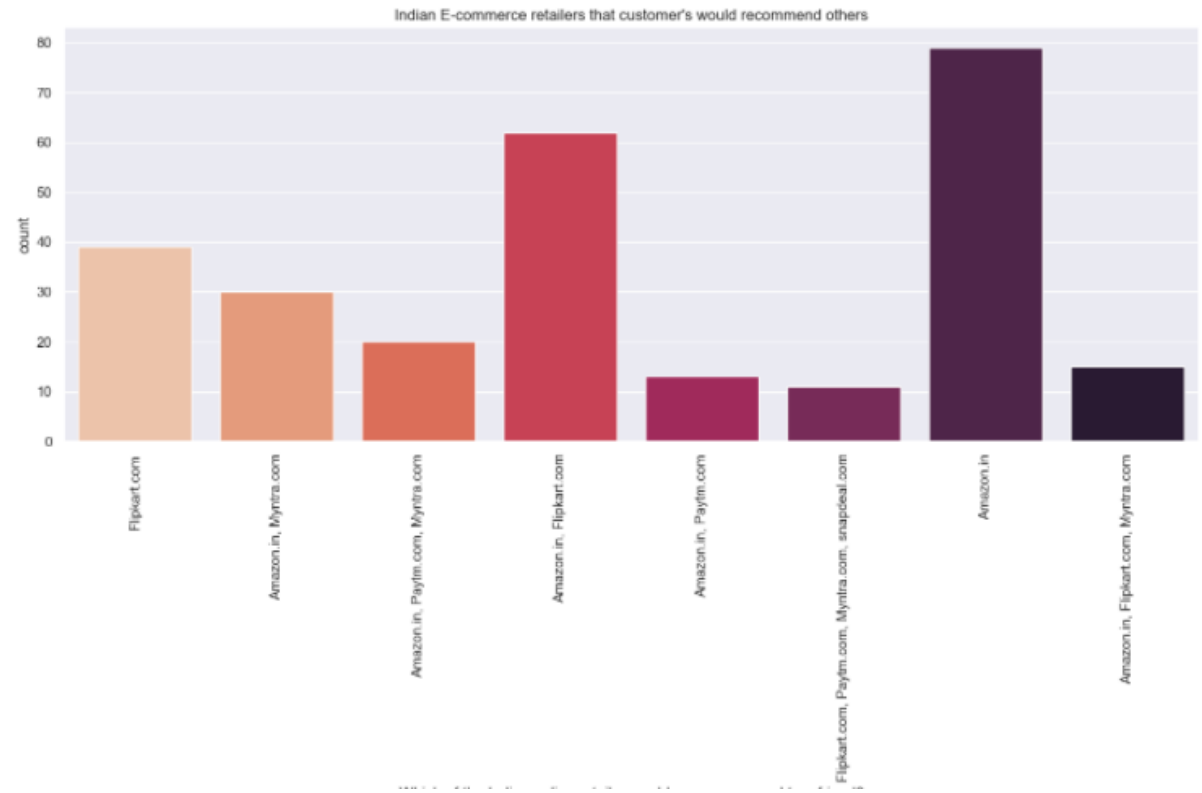
- We can see that if there is ease of navigation, there are good amount chances for a person ending up buying product.
- This is an understatement and based on this alone we cannot determine the outcome of purchasing behaviour but it surely keeps you occupied for longer time to browse some more products.
- If the user interface and navigation are good, we can even find similar products quite easily and this reduces the person's option of moving out of the cart by increasing the pool of products to select from.



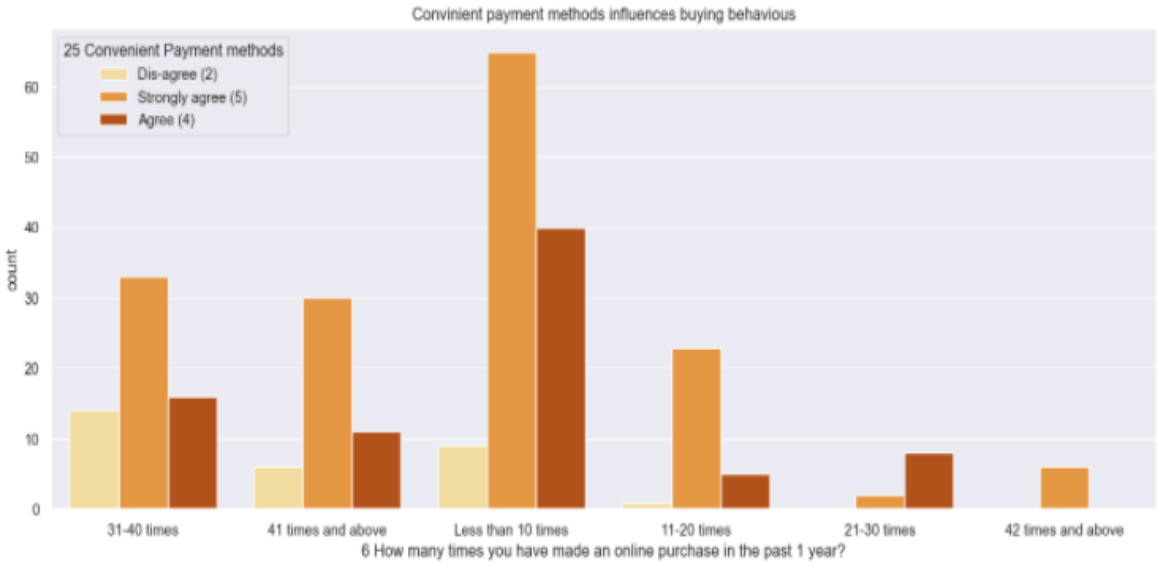
- We can see from the below plot the preference of E-Commerce retailers the customer prefer in general.
- We can see that Amazon, Flipkart, Paytm, Myntra and Snapdeal are the most preferred online retailers. Customers have bought items from these retailers at least once.



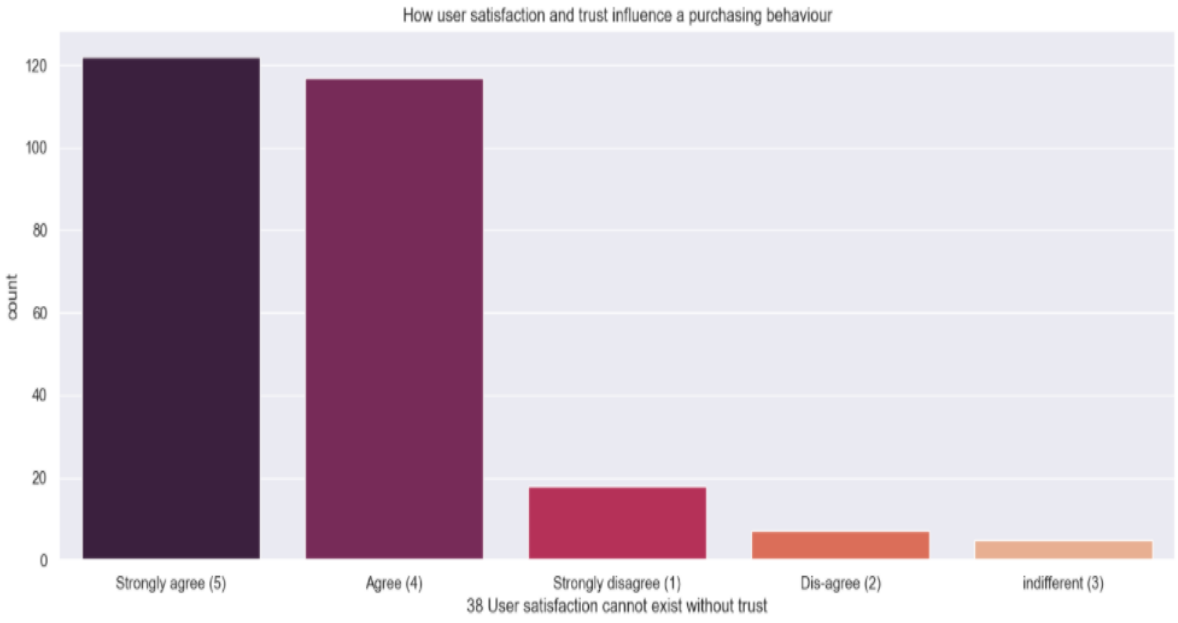
- We can observe from the below plot that majority of Indians prefer Amazon.in over other E-Commerce retailers.
- We know that Flipkart is an Indian brand and still people prefer Amazon. We can say factors like good user interface, feasible price, good quality etc. comes into place even when selecting E-Commerce platform.
- Good purchasing experience and work of mouth can play an important role. More than any means if advert work of mouth has more power and Amazon may have achieved that so far
- Flipkart seems to be catching up with Amazon and other E-Commerce brands have an average recommendation



- From this plot, we can see how people agree on the importance of convenient payment method.
- This also implies availability to have cash on delivery options which most of the Indians prefer when it comes to expensive purchases.
- Having multiple and credible options of payments adds confidence in customers and could also lead to more purchases.
- We can see that people "Strongly agree" in most of the cases

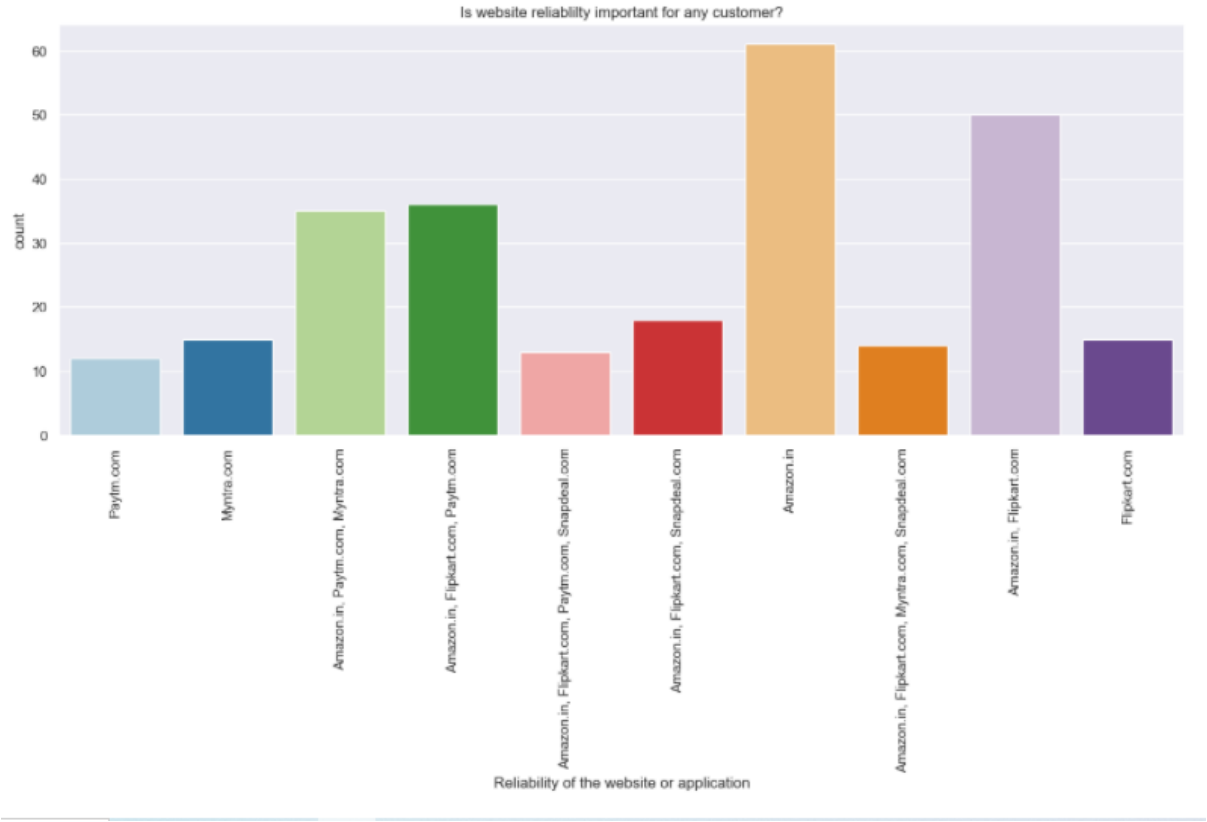


- This plot clearly states that trust is very much important if a person makes purchases online.
- If we observe specifically, many of the E-Commerce retailers provide ratings of products, rating of retailers, positive and negative feedbacks, reviews etc along with product for our ease of reference.
- Each factor describes differently to individual customer. Eg: for someone having 4.5 or 5 start rating means it's good quality, for someone having long term warranty is important over other attributes.
- Since E-Commerce gathers crowed with various mentality and thoughts, trust is very much important in building a valueadded relationship amongst each other

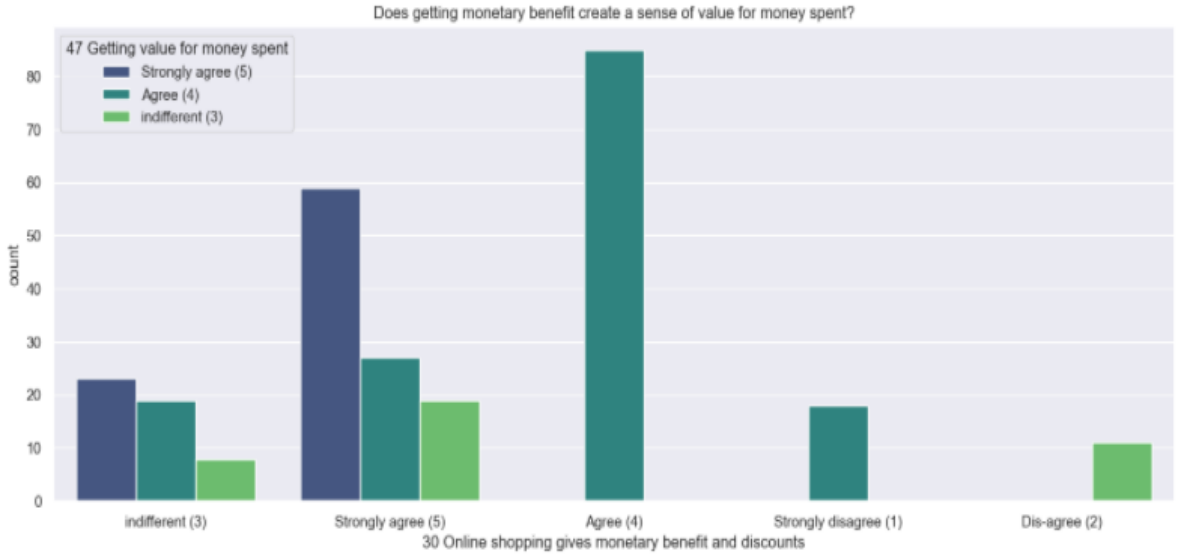


- People seem to prefer websites that are highly reliable. This could include how fast it loads, how soon a graphic or certain feature is loaded, how flexible is the website that can be toggled on both Laptops and Smartphones.
- This also explains the past user experience and no matter how good it was, if there is a very bad experience, the customer is lost for good and may not prefer that channel next time.

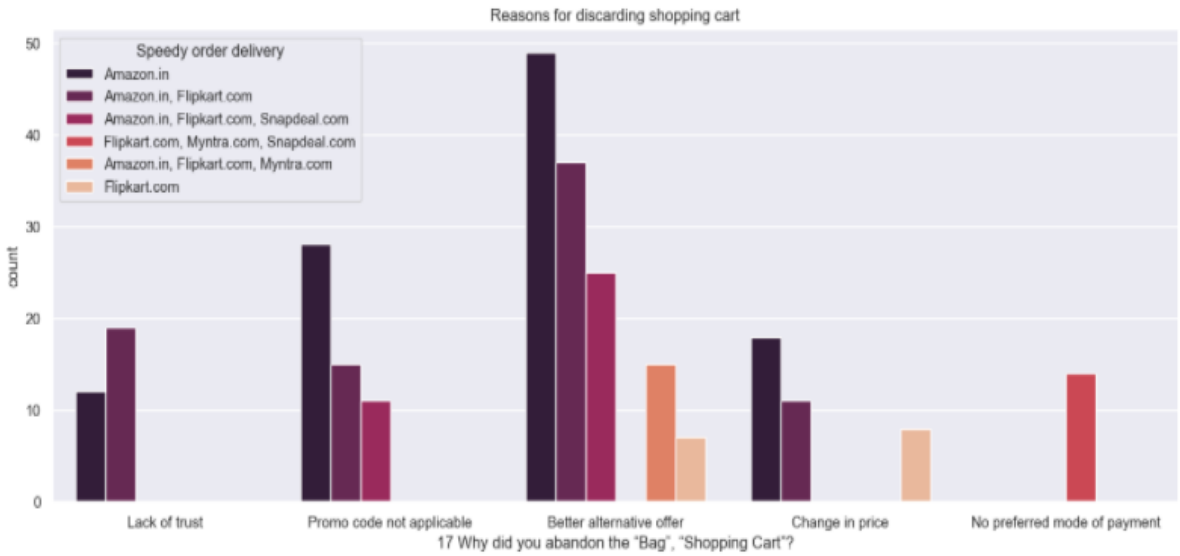
- In this case we can see Amazon has good customer base followed by Flipkart and are have more reliable and stable platform / website compared to others. In short, these are trusted openly compared to other E-Commerce brands.



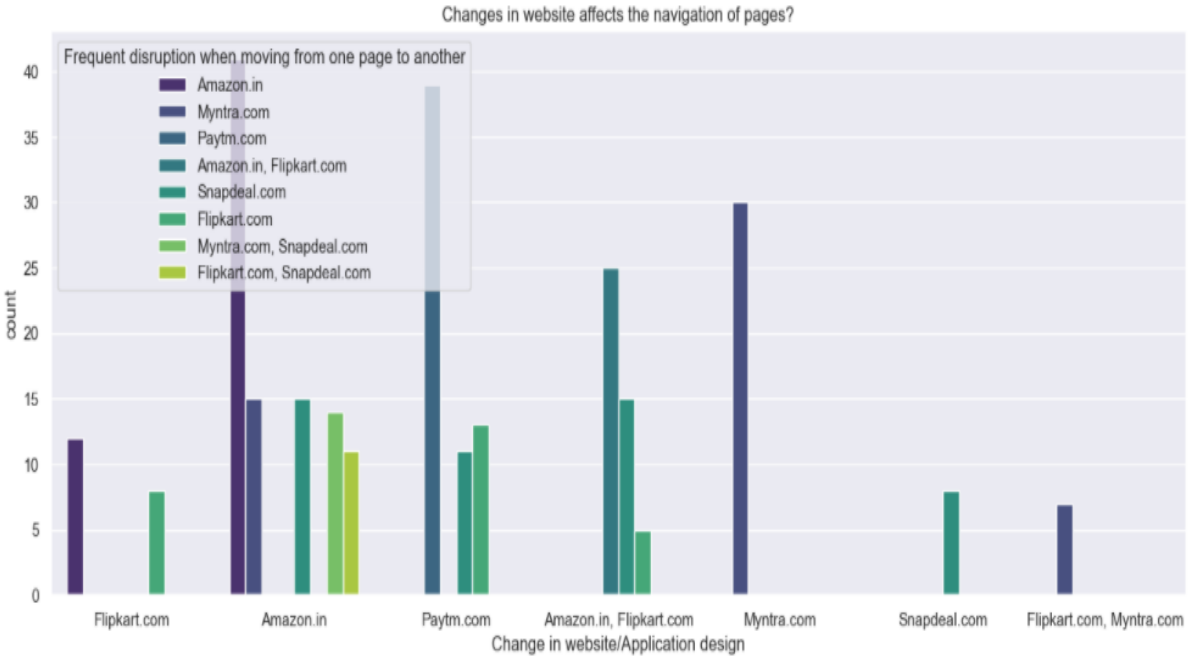
- People may also prefer purchasing online not just for discounts but also feel value for money to some extent.
- Value for money is a satisfaction a person feels or adheres to when purchasing any product. No matter how important that product could be, if a person's financial budget doesn't fit or if they feel it’s not that worth in terms of monetary value that person is likely to drop the Cart.
- By giving good discounts, providing EMI options and payment flexibility there is a good chance that that person could end up being a potential customer.
- Value for money and monetary benefit could also be seen since delivery charges and packaging would be free, once can order from anywhere and get it delivered it anywhere which intern saves money that we end up spending if we go to a Brick and Mortar store.



- From the below plot, Offers, goodies, slight savings etc. highly influence a person's purchasing behaviour
- We can see that Amazon appears to have such perks more than other websites. We can see majority of people do abandon for either "Better alternative offer" or if "Promo code isn't applicable" compared to other reasons.
- This plot can also suggest us the discounts the product would offer, other additional items that can come along with the actual product etc.
- Such discounts are very common if you try purchasing electronic goods like Laptops, Smartphones or Smart TV's etc. as different retailer sell same or similar product and location from where the product is manufacture or shipped from can increase or decrease price to some extent.



- The below plot show's how a website change or shutdown affects the smooth purchasing.
- Majority of people have noticed such incidents happening on Amazon followed by Paytm and Myntra. if disruptions happen especially during payment, one may end up paying accidently and some issues may arise.
- It is also not appealing to see constant changes as a person get's used to certain designs and is perceived that such interfaces are user friendly. Rather than sudden changes, such changes could be done on certain occasions to see the reaction or get feedback of customers and implement on a longer run if it's proven beneficial.
- Since people use different devices that may or may not support certain browsers, changes made on website should be as efficient as possible so that anyone can access on any devices with less fluctuations as possible.
- Also if changes are really required, it would be wise for customers to be informed in prior either through emails or popup ads that are common these days.

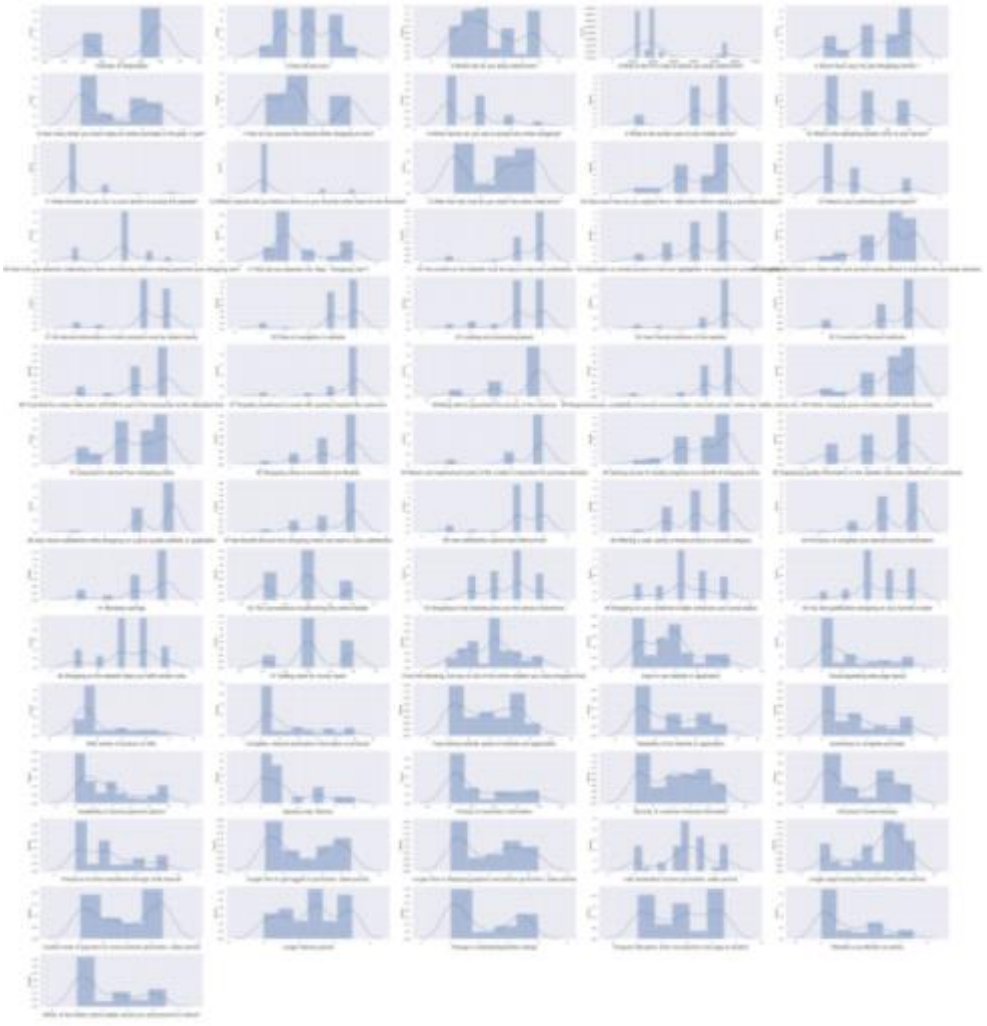


- We can see that quality of information and detailed information is very much important. The point of not going to stores physical to examine the product, is an expectation that products displayed online have all the information

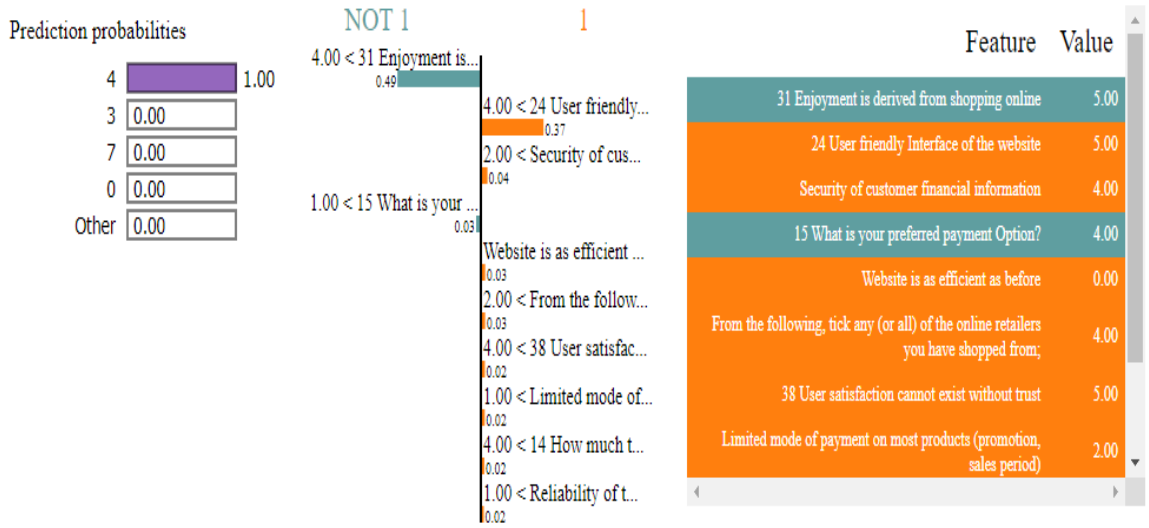
- For some, colour of the product is important, for some quality and durability, for some price is important etc and having detailed information increase the purchasing decision.
- Users get a sense of purchasing satisfaction also because of good quality of website.

- We can see that return and return policy is highly influencing a person's purchasing choice.
- Amazon's seems to have good policies when it comes to this. Since we buy a product without knowing the attributes physically, there would be some sort of fear back in our minds psychologically whether the product we see is same as it looks like.
- Having return options can expand our horizon in getting better products as we know we can always exchange if the product doesn't meet our expectations.
- This in turn increases the loyalty of customers.

- Since all the columns are actually categorical in nature, there is no need to deduce any understanding from distribution as it doesn't apply.
- Although we have pincodes as Int, these are also not the typical continuous variables and hence distribution is not considered for this as well.
- This is how I would check the distribution and then proceed to remove outliers using either IQR or Z-score method to reduces skewness if the dataset was continuous features.



- A handful of data is selected from train dataset.
- In the following figure, we can see what features are falling under



CONCLUSION

- Learning Outcomes of the Study in respect of Data Science

  - With the help of this dataset, I was able to work on multiple algorithms, work on encoding techniques and also got to know the limitation of this dataset.
  - I also learnt how to use Plotly and Shapash libraries which I did not use back then for other projects and utilized the same for understanding the problem with different set of perspectives.
  - I got to know the importance of visualizations and how beneficial it is to conclusions and helped me in debunking random assumptions.

- Limitations of this work and Scope for Future Work

  - If this dataset had more records, we could generalize the results and make assumptions in general.
  - We can have specific details like the years of the data, how old this data is. We can predict or try to get outcome if the data is diverse and not limited to few features only.
  - Since E-Commerce is a dynamic domain, changes in information affects the business every now and then. If we have a data that is 5 years or 10 years older than current year, it is not really useful to predict as a lot of changes may have happened and the output could be wrong.
  - Also because this dataset had no missing or null values we ended up getting a clean dataset which is generally not practical when it comes to real business environment.
  - The dataset has more amount of Female population and less number of Male population and the assumptions may be biased to some extent.