CUSTOMER RETENTION ANALYSIS

Submitted by: **SANTHOSH KUMAR NAROJ**

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.

References:

1. Python Data Analysis: Perform data collection, data processing, wrangling, visualization, and model building using Python, 3rd Edition. by Avinash Navlani (Author), Armando Fandango (Author), Ivan Idris (Author)
2. Notes and classes by Datatrained Academy.

INTRODUCTION

Business Problem Framing

Large number of customers are getting attracted towards online retailing; this is because e-stores usually offer them a variety of services and products according to their preferences. Convenience, round the clock availability, flexible pricing, discounts as well as free door step delivery are some of the major benefits of shopping online. Presently, more number of online retailers are beginning to experience increase in demand for products and services.

Several e-commerce start-ups have commenced operation with innovative strategies, which differs from what was pioneered by first generation e-commerce companies. Online customer retention has become essential for the success of businesses. Keeping good relationship with customer is vital for the success of any business. Customer satisfaction, retention, loyalty and positive word of mouth are critical for the profitability as well as the success of businesses. Nevertheless, in service oriented industry. Online customer retention has become important for the business more than any time. This is especially when it comes to services. Services attempted to keep in connection with their customers and ensure their intention to visit again or to re-purchase the services. Retention of customers is less expensive than attracting new customers.

Currently, businesses are facing unprecedented challenges, customer retention has become more important than ever. Customer retention is defined as transferring the new customer to regular customers and keep good relationship with them.

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.

Analytical Problem Framing

Methodology represents a description about the framework that is undertaken. It consists of various milestones that need to be achieved in order to fulfil the objective. We have various attributes that contribute retaining customers that aids the growth of business.

The following steps represents stepwise tasks that need to be completed:

- Data Collection
- Data pre-processing
- Data Visualization
- Data Analysis

Data Sources and their formats

We have 2 Datasets of 270 entries and it has 71 variables.

Data Count 270

Data Features 71

Data Types : Object

Data Pre-processing

Data pre-processing is a process of transforming the raw, complex data into systematic understandable knowledge.

- Data sample:

	0	1	2	3	4
1Gender of respondent	Male	Female	Female	Male	Female
2 How old are you?	31-40 years	21-30 years	21-30 years	21-30 years	21-30 years
3 Which city do you shop online from?	Delhi	Delhi	Greater Noida	Karnal	Bangalore
4 What is the Pin Code of where you shop online from?	110009	110030	201308	132001	530068
5 Since How Long You are Shopping Online ?	Above 4 years	Above 4 years	3-4 years	3-4 years	2-3 years
...	...	...	...	...	...
Longer delivery period	Paytm.com	Snapdeal.com	Paytm.com	Paytm.com	Paytm.com
Change in website/Application design	Flipkart.com	Amazon.in	Paytm.com	Amazon.in, Flipkart.com	Amazon.in
Frequent disruption when moving from one page to another	Amazon.in	Myntra.com	Paytm.com	Amazon.in, Flipkart.com	Snapdeal.com
Website is as efficient as before	Amazon.in	Amazon.in, Flipkart.com	Amazon.in	Amazon.in, Flipkart.com, Paytm.com	Paytm.com
Which of the Indian online retailer would you recommend to a friend?	Flipkart.com	Amazon.in, Myntra.com	Amazon.in, Paytm.com, Myntra.com	Amazon.in, Flipkart.com	Amazon.in, Myntra.com

71 rows × 5 columns

- Data type information.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 269 entries, 0 to 268
Data columns (total 71 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Gender                    269 non-null    object
 1   Age                       269 non-null    object
 2   City                      269 non-null    object
 3   Pin code                  269 non-null    int64
 4   Since_shop_on             269 non-null    object
 5   times_onlin_pur           269 non-null    object
 6   access_internet           269 non-null    object
 7   Device                    269 non-null    object
 8   Screen_size               269 non-null    object
 9   OS                        269 non-null    object
 10  Browser                   269 non-null    object
 11  Chnl_to_website           269 non-null    object
 12  reach_store               269 non-null    object
 13  time_spent_bfore_deci     269 non-null    object
 14  Payment_opt               269 non-null    object
 15  leave_witout_payin        269 non-null    object
 16  Reason_abandon            269 non-null    object
 17  Is_cont_easy_undrstnd     269 non-null    object
 18  Info_prod_comp            269 non-null    object
 19  Comp_info_liseller        269 non-null    object
 20  relevant_info_prod        269 non-null    object
 21  ez_navi_web               269 non-null    object
 22  load_speed                269 non-null    object
 23  Usr_frd_interface         269 non-null    object
 24  payin_method              269 non-null    object
 25  Trust_on_transac          269 non-null    object
 26  empathy_assist_queries    269 non-null    object
 27  Guar_cust_privacy         269 non-null    object
 28  Response_sev_com          269 non-null    object
```

Target Variable: "Selling Which of the Indian online retailer would you recommend to a friend?" is the Target Variable; which is renamed as "Recommend". All the attributes are renamed for the convenience of Data Visualization.

```
# Changing column names as headings are more like questions and lengthy.

def new_column_names(data):

    columns = list(data.columns)
    for col in columns:
        indx = columns.index(col)

        if indx == 0:
            data.rename(columns = {'1Gender of respondent' : 'Gender'}, inplace = True)
        elif indx == 1:
            data.rename(columns = {'2 How old are you? ' : 'Age'}, inplace = True)
        elif indx == 2:
            data.rename(columns = {'3 Which city do you shop online from?': 'City'}, inplace = True)
        elif indx == 3:
            data.rename(columns = {'4 What is the Pin Code of where you shop online from?' : 'Pin code'}, inplace = True)
        elif indx == 4:
            data.rename(columns = {'5 Since How Long You are Shopping Online ?': 'Since_shop_on'}, inplace = True)
        elif indx == 5:
            data.rename(columns = {'6 How many times you have made an online purchase in the past 1 year?': 'times_onlin_pur'
        elif indx == 6:
            data.rename(columns = {'7 How do you access the internet while shopping on-line?' : 'access_internet'}, inplace =
        elif indx == 7:
            data.rename(columns = {'8 Which device do you use to access the online shopping?': 'Device'}, inplace = True)
        elif indx == 8:
            data.rename(columns = {'9 What is the screen size of your mobile device?\t\t\t\t\t\t
        elif indx == 9:
            data.rename(columns = {'10 What is the operating system (OS) of your device?\t\t\t\t
        elif indx == 10:
            data.rename(columns = {'11 What browser do you run on your device to access the website?\t\t\t
```

Feature List :

```
# New column names
```

Run this cell

```
Index(['Gender', 'Age', 'City', 'Pin code', 'Since_shop_on', 'times_onlin_pur',
       'access_internet', 'Device', 'Screen_size', 'OS', 'Browser',
       'Chnl_to_website', 'reach_store', 'time_spent_bfore_deci',
       'Payment_opt', 'leave_witout_payin', 'Reason_abandon',
       'Is_cont_easy_undrstnd', 'Info_prod_comp', 'Comp_info_liseller',
       'relevant_info_prod', 'ez_navi_web', 'load_speed', 'Usr_frd_interface',
       'payin_method', 'Trust_on_transac', 'empathy_assist_queries',
       'Guar_cust_privacy', 'Response_sev_com', 'ben_dis_shpon', 'Enjy_shop',
       'Conv_flexi', 'replace_policy_imp', 'gain_acc_loyal_prog',
       'display_qual_info_sat', 'Usr_driv_sat_app', 'Net_benf_on_usrsat',
       'Usr_sat_ntexst_trst', 'Off_widvar_prod', 'Prov_prod_info',
       'Money_savings', 'con_pat_retlr', 'sense_adven_shopon',
       'enhSoc_shop_estore', 'gratif_shop_etailor', 'fullfill_roles',
       'worth_of_money', 'onlin_reatl_shp_frm', 'easy_use_app',
       'visual_web_layout', 'variety_prod_off', 'desc_info_prods',
       'fast_loading', 'reliab_website', 'Quick_comp', 'Avail_pay_opt',
       'speed_odr_del', 'privacy_cust_info', 'security_cust_fin_info',
       'perceived_trust', 'presnce_onlin_assist', 'longer_tim_to_login',
       'longer_tim_disp_graphics', 'late_declr_price', 'lon_page_load_tim',
       'limtd_mode_pay', 'lon_delvry_period', 'chnge_in_app_des',
       'disrupt_frm_pag_mov', 'web_is_eff_bfre', 'Recommend'],
      dtype='object')
```

The data after renaming:

```
#applying the function on the dataset

df = new_column_names(df)
df.head()
```

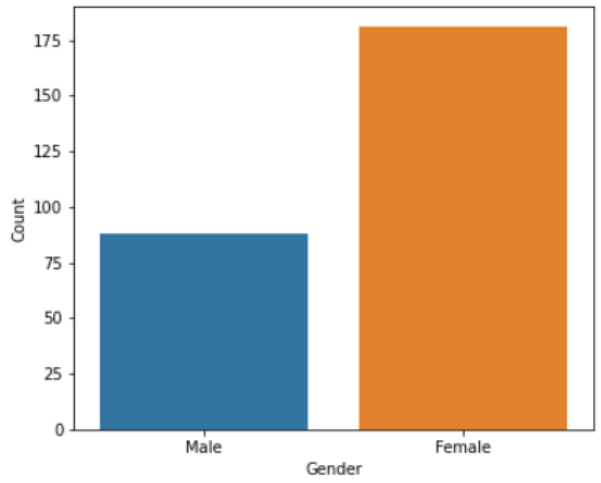
	Gender	Age	City	Pin code	Since_shop_on	times_onlin_pur	access_internet	Device	Screen_size	OS	...	longer_tim_to_login	I
0	Male	31-40 years	Delhi	110009	Above 4 years	31-40 times	Dial-up	Desktop	Others	Window/windows Mobile	...	Amazon.in	
1	Female	21-30 years	Delhi	110030	Above 4 years	41 times and above	Wi-Fi	Smartphone	4.7 inches	IOS/Mac	...	Amazon.in, Flipkart.com	
2	Female	21-30 years	Greater Noida	201308	3-4 years	41 times and above	Mobile Internet	Smartphone	5.5 inches	Android	...	Myntra.com	
3	Male	21-30 years	Karnal	132001	3-4 years	Less than 10 times	Mobile Internet	Smartphone	5.5 inches	IOS/Mac	...	Snapdeal.com	
4	Female	21-30 years	Bangalore	530068	2-3 years	11-20 times	Wi-Fi	Smartphone	4.7 inches	IOS/Mac	...	Flipkart.com, Paytm.com	

5 rows × 71 columns

Data Visualization and Analysis

1.Ratio of gender in the data.

```
plt.figure(figsize = (6,5))
sns.countplot(df['Gender'])
plt.xlabel('Gender')
plt.ylabel('Count')
plt.show()
```



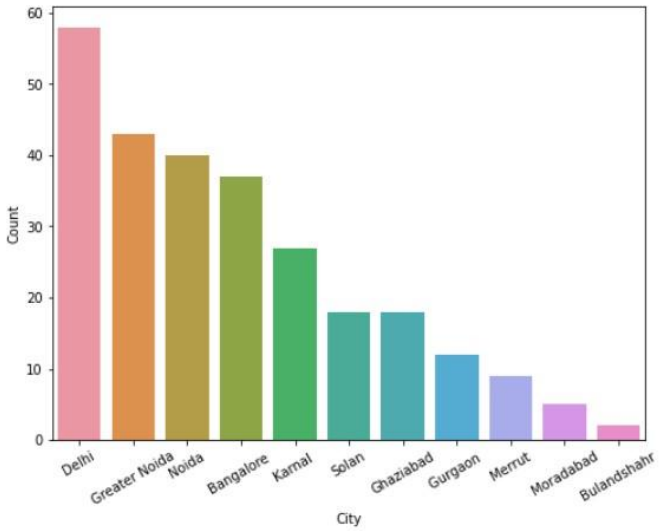
Observation:

1.Female count is more than double of male count.

2.Females shops more than males do as per the survey data.

2. Shopping Ratio as per cities

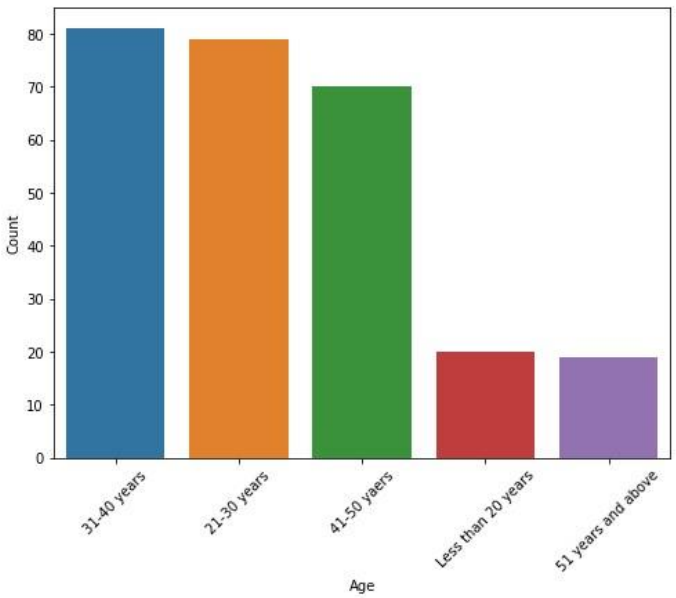
```
plt.figure(figsize = (8,6))
sns.countplot(df['City'], order = df['City'].value_counts().index)
plt.xlabel('City')
plt.ylabel('Count')
plt.xticks(rotation = 30);
```



Observations:
1.Delhi is the city from which maximum orders placed, which is due to the lifestyle of people living in the city.
2.Greater Noida, Noida, and Banglore follows in terms of online sales.
3.Least among all is Merrut, Moradabad, Bulandshire in sales as people in these places doesnt follow fast moving lifestyle as industries and MNC's are less in these area which makes people to adapt with local lifestyle than online shopping.

3. Count of customers according according to Age.

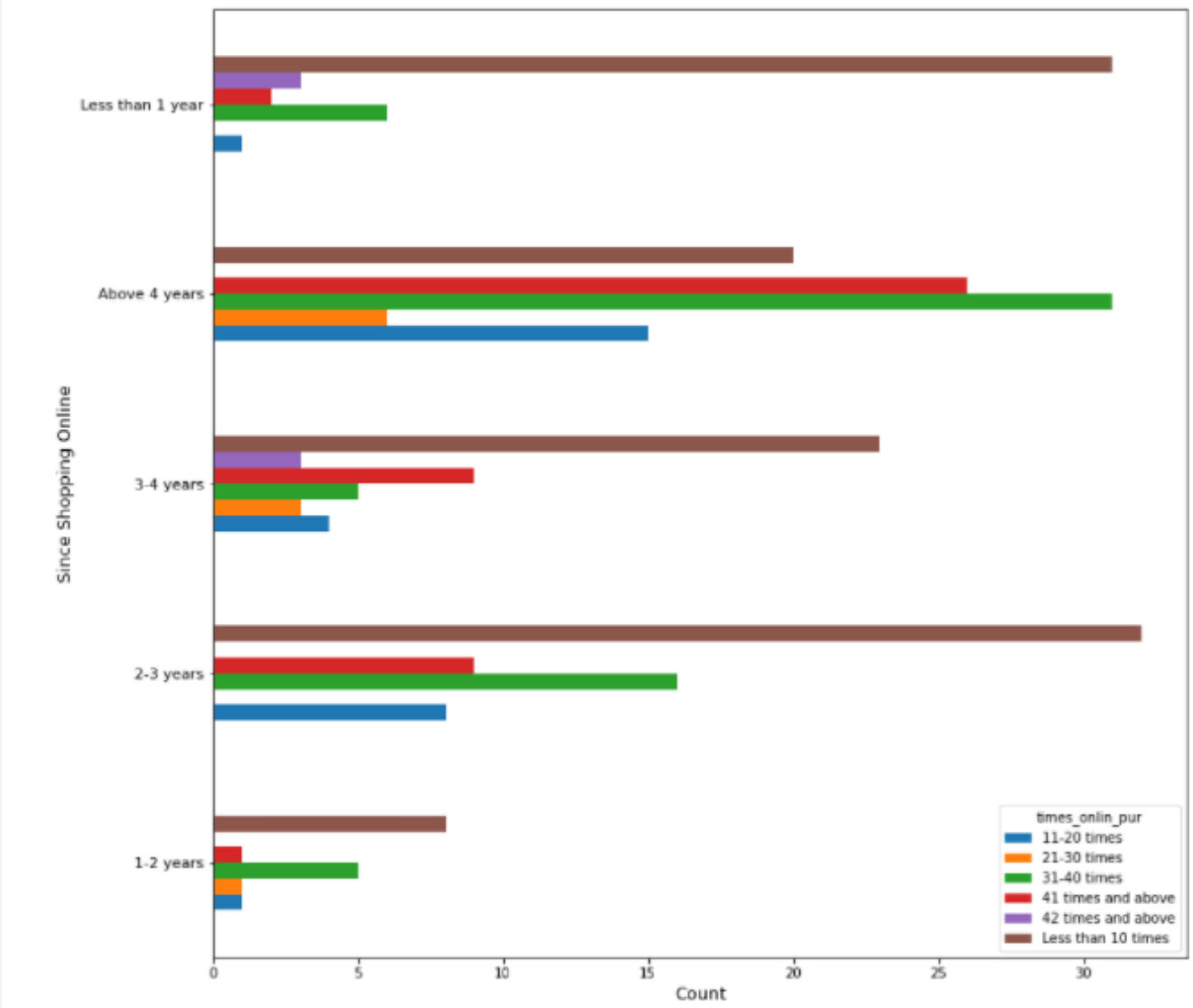
```
plt.figure(figsize = (8,6))
sns.countplot(df['Age'])
plt.xlabel('Age')
plt.ylabel('Count')
plt.xticks(rotation = 45);
```



OBSERVATION:
1.Highest number of customers are in range 31-40yrs.
2.However there is no great difference in sales from 21-50yrs, As this age range is the earning class and shops more.
3.Lowest number of customers are in the age range less than 20 and greater than 51yrs. Former category is majorly students and latter is retired majorly.

4. Comparing the time taken for Online purchase.

```
pd.crosstab(df['Since_shop_on'], df['times_onlin_pur']).plot.barh(figsize = (13, 13))
plt.xlabel('Count', fontsize = 13)
plt.ylabel('Since Shopping Online', fontsize = 13)
plt.yticks(fontsize = 11);
```



OBSERVATIONS:

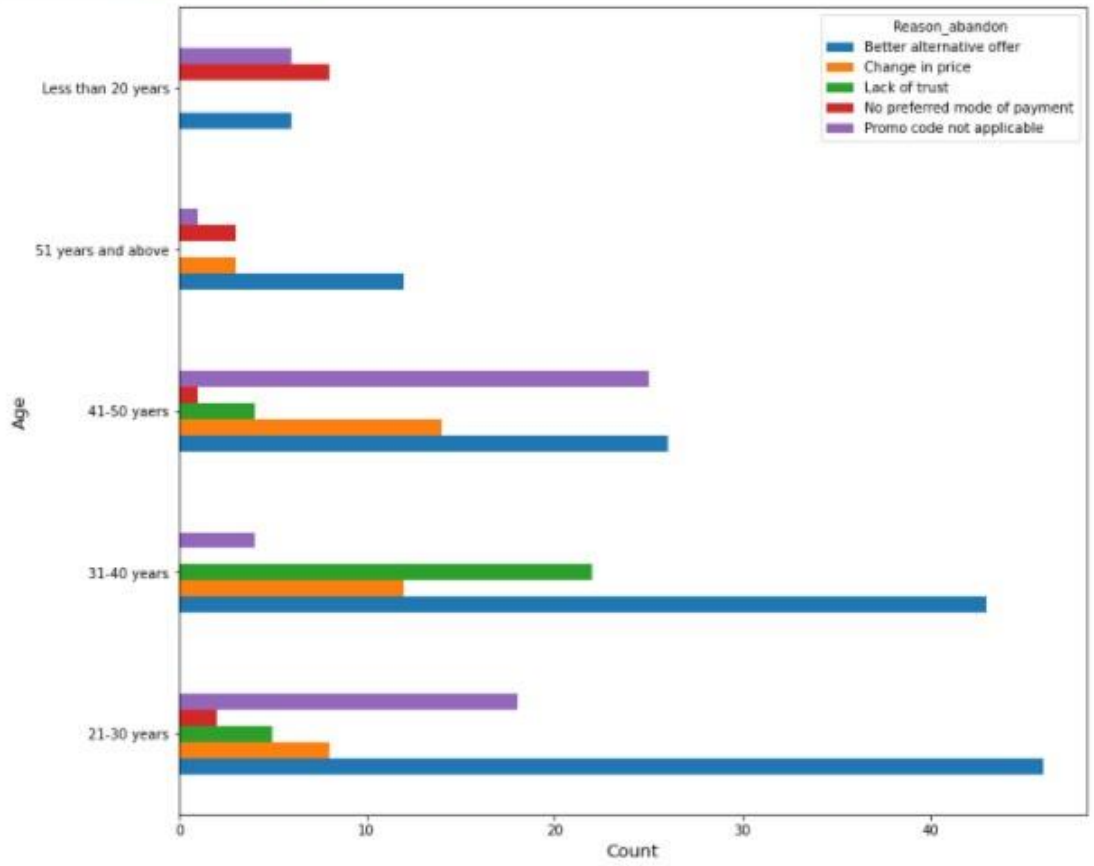
1.Customers who are purchasing more than 4yrs purchased more than 41times and above is the highest number.

2.Follows by customers who are purchasing from 3-4yrs.

3.Maximum customers who purchased less than 10 times are in the range purchasing from 2-3yrs; followed by less than 1 yr.

5.Comparison of Age and Reason for Abandoning purchase/Shopping Cart.

```
pd.crosstab(df['Age'], df['Reason_abandon']).plot.barh(figsize = (12,11))
plt.xlabel('Count', fontsize = 13)
plt.ylabel('Age', fontsize = 13);
```



OBSERVATIONS:

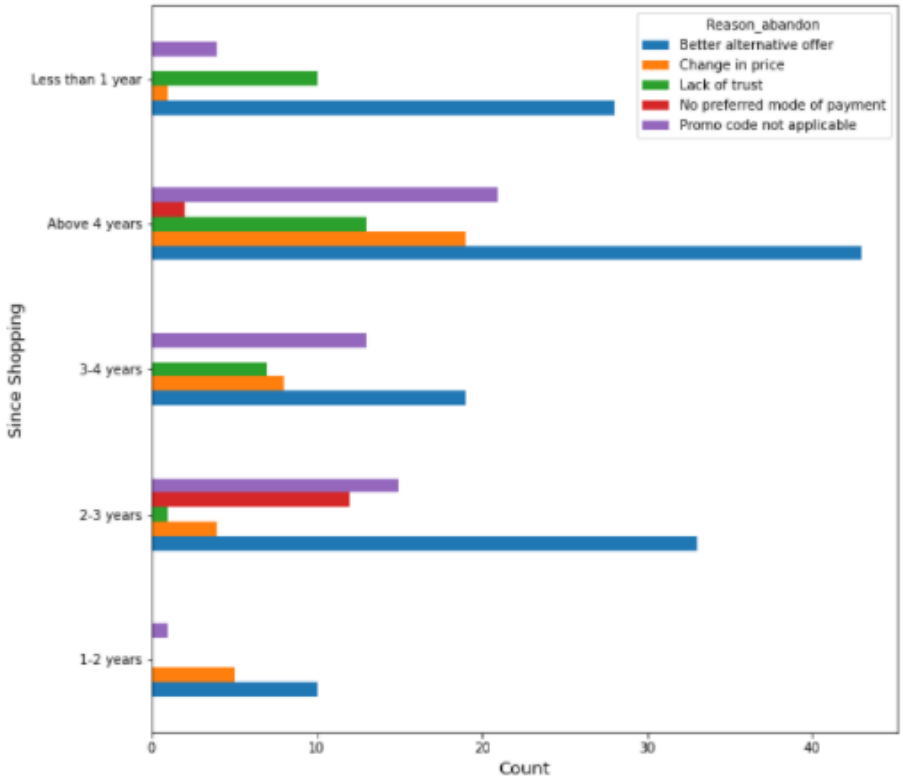
1.Better Alternative Offer is the main reason for abandoning shopping cart in almost all age categories.

2.Customers of age range 21-40yrs are the ones who browses alot online and compares prices.

3.No preffered payment mode is the least occured reason for abandoning shopping cart.

6. Comparing the relation between the Time since shopping and Reason for Abandoning.

```
pd.crosstab(df['Since_shop_on'], df['Reason_abandon']).plot.barh(figsize = (10,10))
plt.xlabel('Count', fontsize = 13)
plt.ylabel('Since Shopping', fontsize = 13);
```



OBSERVATIONS:

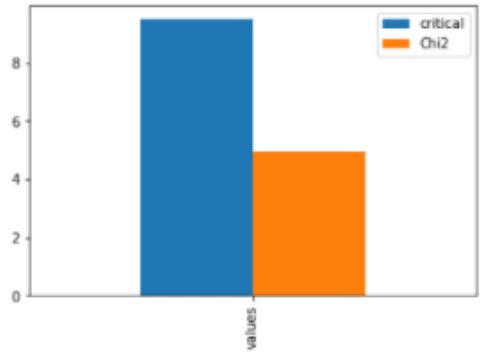
Chi2 square for checking whether these variables depend on each other.

```
pivot = pd.crosstab(df['Age'], df['Gender'])
```

```
stat, p, dof, expected = chi2_contingency(pivot)
print(f'degree of freedom is {dof}')
print(f'stats is {stat}')
print(f'p values id {p}')
```

```
degree of freedom is 4
stats is 4.964093342321093
p values id 0.29100162537226704
```

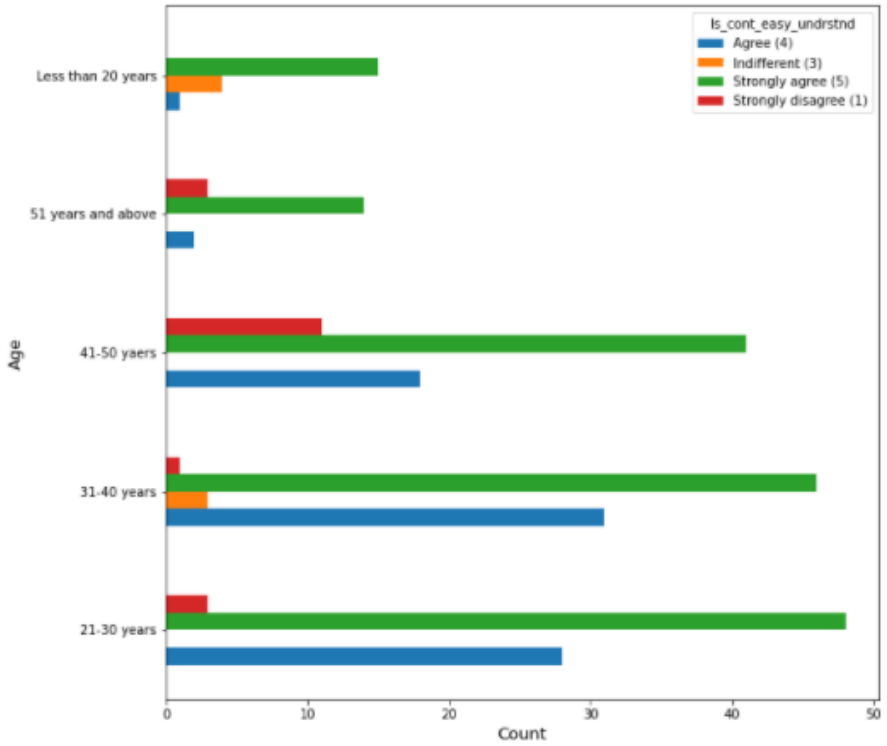
```
prob = 0.95
critical = chi2.ppf(prob, dof)
result = {'critical': critical, 'Chi2': stat}
result = pd.DataFrame(result, index = ['values'])
result.plot.bar();
```



We can see that chi2 value is less then critical value there are dependence among these variables.

Comparison of Age with Understanding the content of website

```
pd.crosstab(df['Age'], df['Is_cont_easy_undrstnd']).plot.barh(figsize = (10,10))
plt.xlabel('Count', fontsize = 13)
plt.ylabel('Age', fontsize = 13);
```

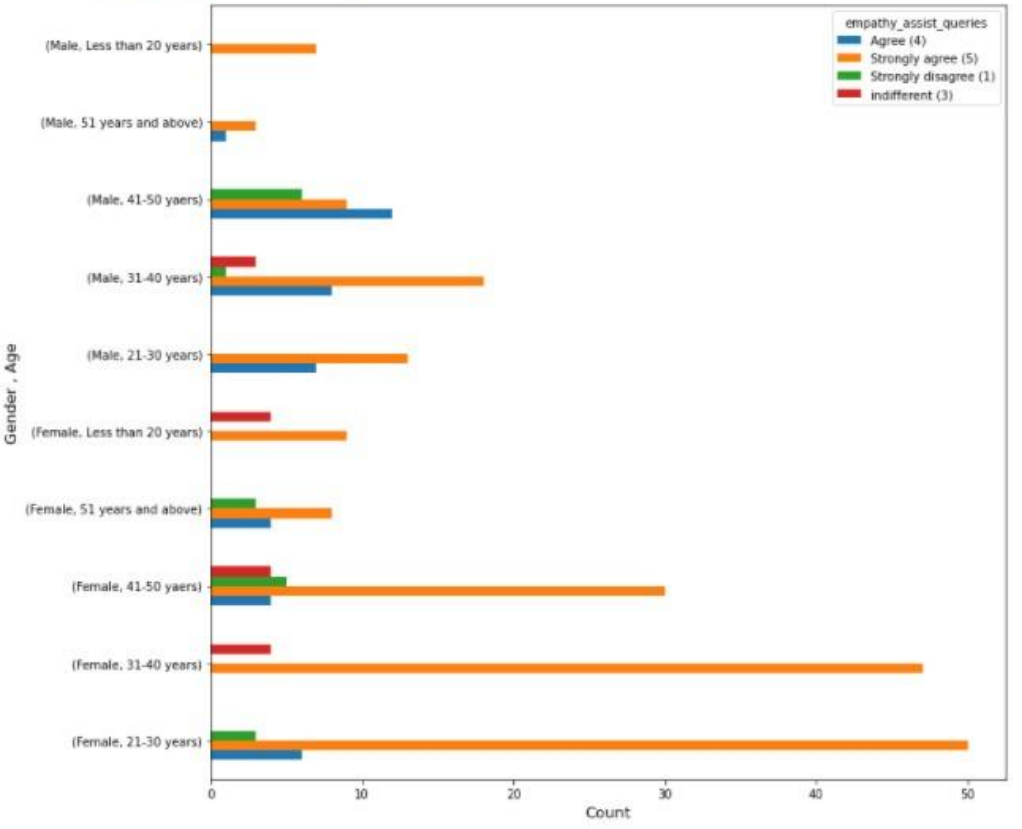


OBSERVATION:

1.people with age 21-30 years are strongly agreeing for content should be easy to understand.

Comparison on Age, Gender and Empathy on Assisting Querries.

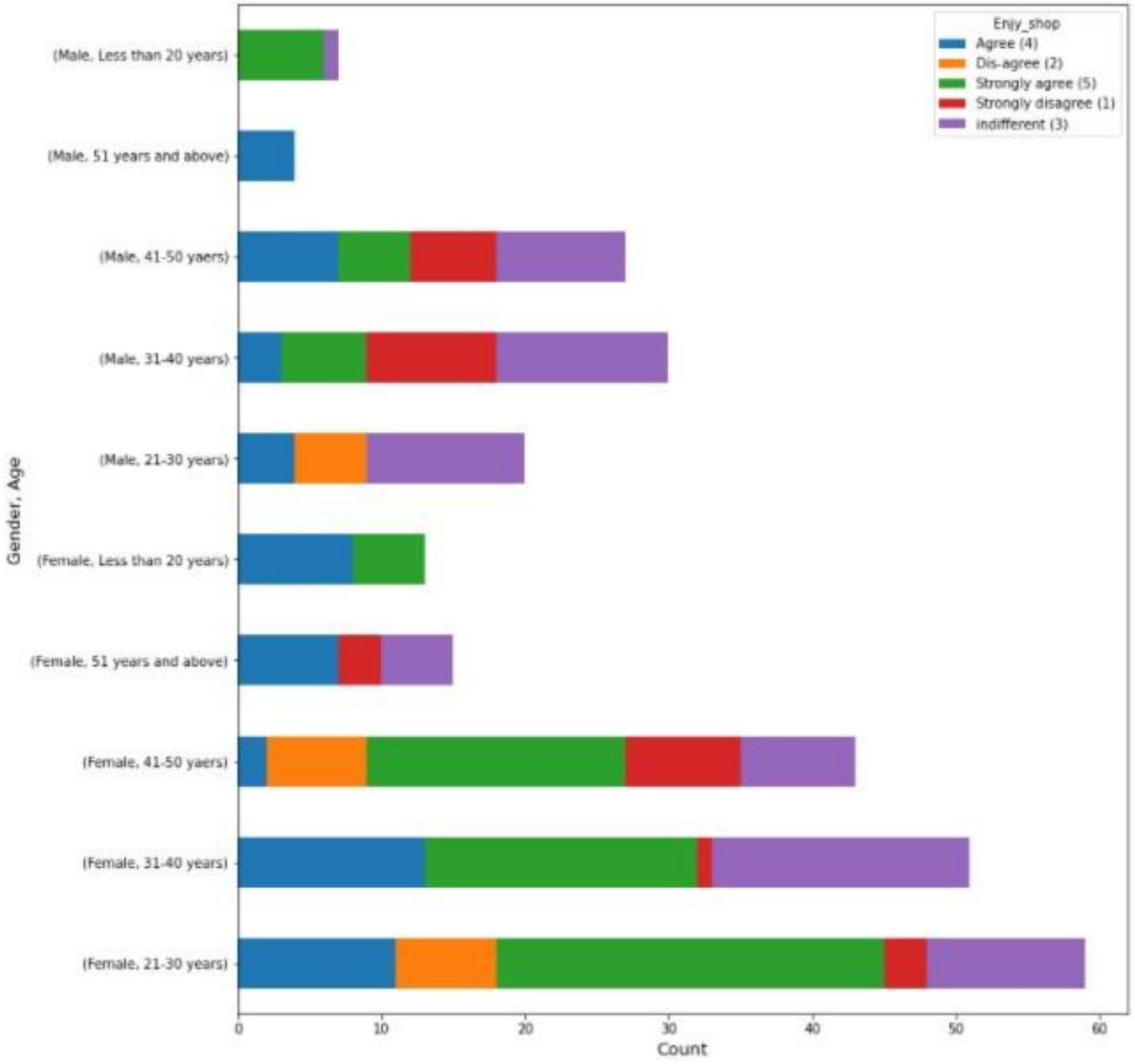
```
pd.crosstab([df['Gender'], df['Age']], df['empathy_assist_queries']).plot.barh(figsize = (12,12))
plt.xlabel('Count', fontsize = 13)
plt.ylabel('Gender , Age', fontsize = 13);
```



Female with age 21-30 are strongly agree for Empathy (readiness to assist with queries) towards the customers.

Female with age 21-30 are strongly agree for Empathy (readiness to assist with queries) towards the customers.

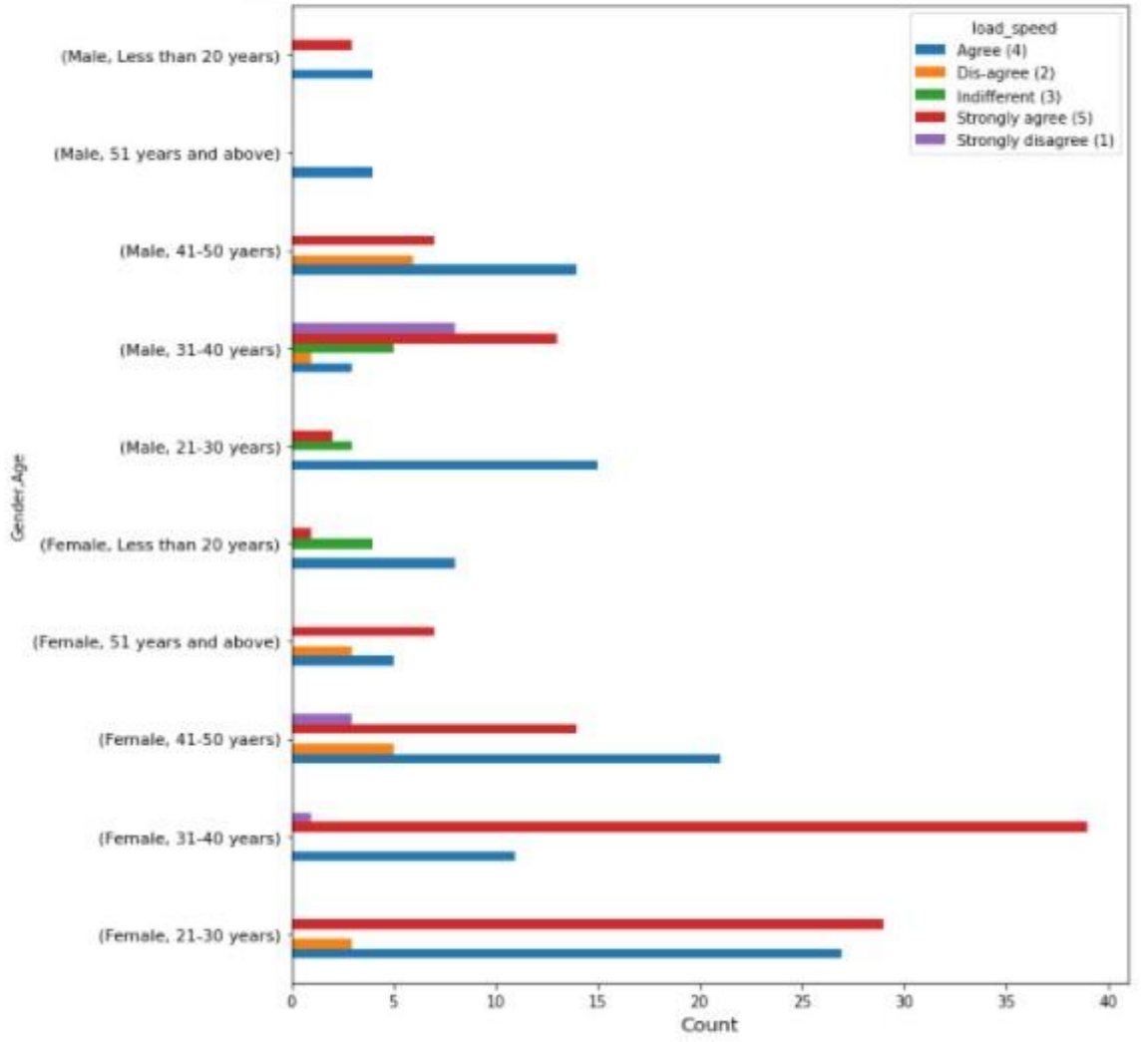
```
pd.crosstab([df['Gender'], df['Age']], df['Enjy_shop']).plot.barh(figsize = (12,14), stacked= True)
plt.xlabel('Count', fontsize = 13)
plt.ylabel('Gender, Age', fontsize = 13);
```



Female with age 21-30 are strongly agreeing that enjoying shopping.

Comparison of Gender, Age and Internet loading speed.

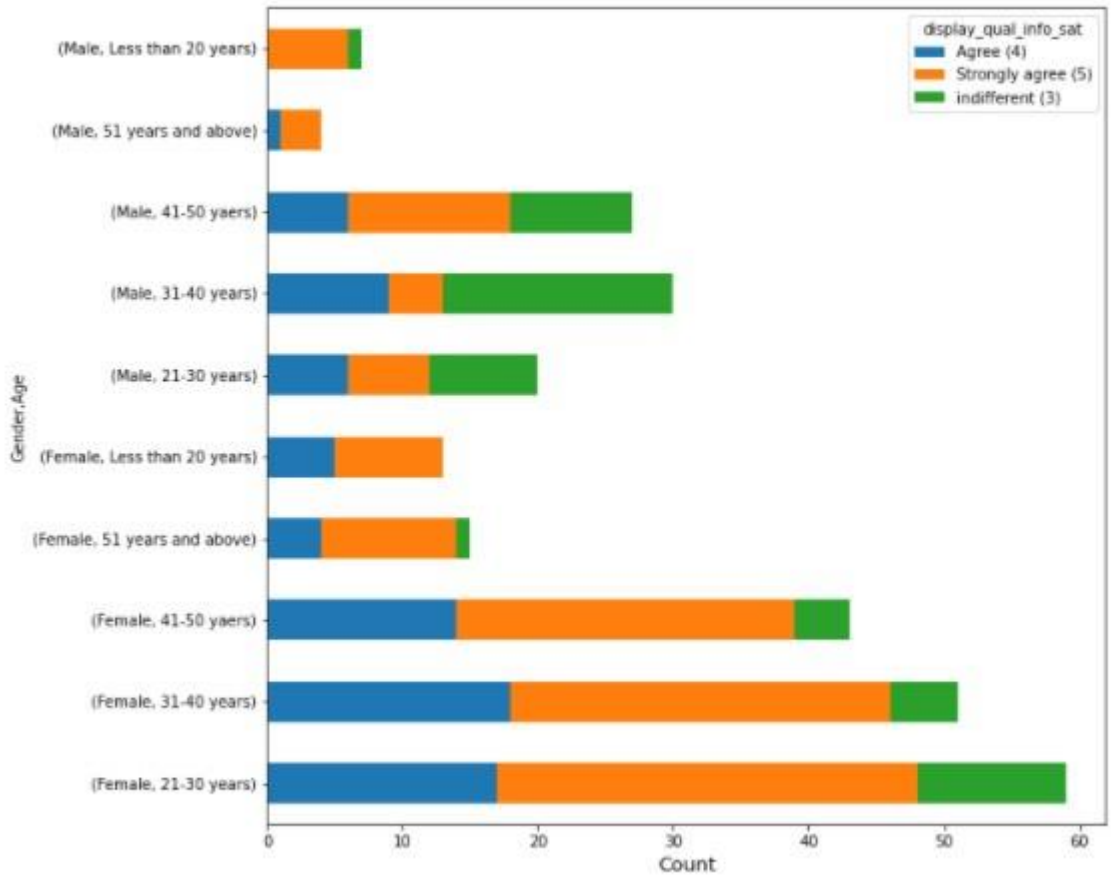
```
pd.crosstab([df['Gender'],df['Age']], df['load_speed']).plot.barh(figsize = (10,12))
plt.xlabel('Count', fontsize = 13)
plt.yticks(fontsize = 11);
```



Female with age 31-40 are strongly agreeing that load speed as a important factor.

Comparison of Gender, Age and Display quality.

```
pd.crosstab([df['Gender'], df['Age']], df['display_qual_info_sat']).plot.barh(figsize = (10,10), stacked = True)
plt.xlabel('Count', fontsize = 13);
```



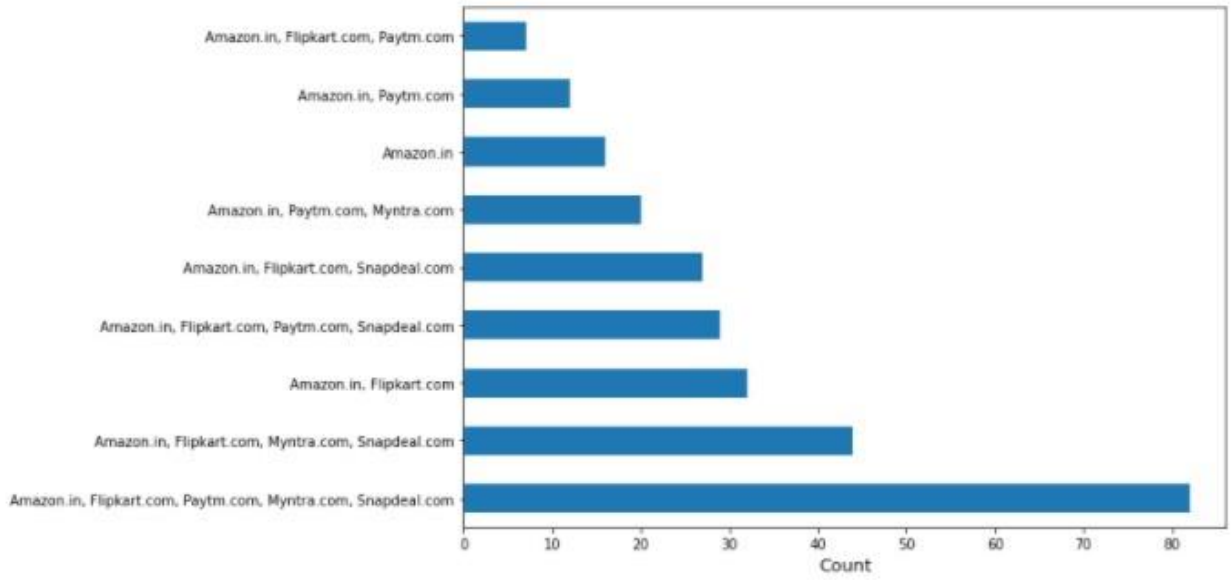
Female under age 21-30 are agreeing more for displaying quality.

```
pd.crosstab([df['Gender'],df['Age']], df['onlin_reatl_shp_frm'], margins = True)
```

	onlin_reatl_shp_frm	Amazon.in	Amazon.in, Flipkart.com	Amazon.in, Flipkart.com, Myntra.com, Snapdeal.com	Amazon.in, Flipkart.com, Paytm.com	Amazon.in, Flipkart.com, Paytm.com, Myntra.com, Snapdeal.com	Amazon.in, Flipkart.com, Paytm.com, Snapdeal.com	Amazon.in, Flipkart.com, Snapdeal.com	Amazon.in, Paytm.com	Amazon.in, Paytm.com, Myntra.com
Gender	Age									
Female	21-30 years	8	8	10	0	17	1	7	0	8
	31-40 years	0	15	3	0	25	0	4	1	3
	41-50 yaers	5	1	4	0	10	4	11	3	5
	51 years and above	0	0	7	0	4	4	0	0	0
	Less than 20 years	0	0	8	4	1	0	0	0	0
Male	21-30 years	0	3	1	0	7	4	5	0	0
	31-40 years	0	5	1	3	7	6	0	8	0
	41-50 yaers	3	0	4	0	10	9	0	0	1
	51 years and above	0	0	3	0	0	1	0	0	0
	Less than 20 years	0	0	3	0	1	0	0	0	3
All		16	32	44	7	82	29	27	12	20

Count of Retail store in Datasets.

```
df['onlin_reatl_shp_frm'].value_counts().plot(kind = 'barh', figsize = (10,7))
plt.xlabel('Count', fontsize = 13);
```



surprisingly people are using most of platforms for online purchase instead depending on just one site.

Variable independency of device and screen size.

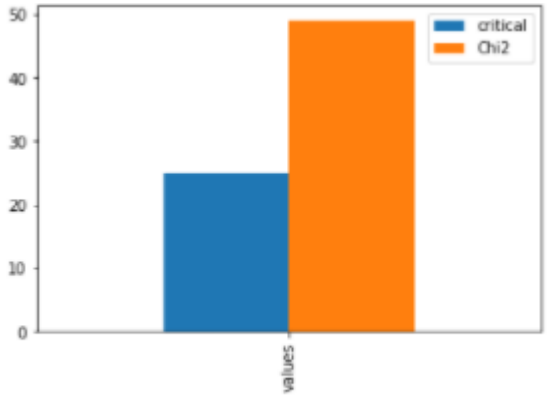
```
device_tim = pd.crosstab(df['Device'], df['times_onlin_pur'])
device_tim
```

times_onlin_pur	11-20 times	21-30 times	31-40 times	41 times and above	42 times and above	Less than 10 times
Device						
Desktop	1	0	13	7	0	9
Laptop	12	6	14	10	0	44
Smartphone	12	4	28	30	6	61
Tablet	4	0	8	0	0	0

```
stat, p, dof, expected = chi2_contingency(device_tim)
print(f'degree of freedom is {dof}')
print(f'stats is {stat}')
print(f'p values id {p}')
```

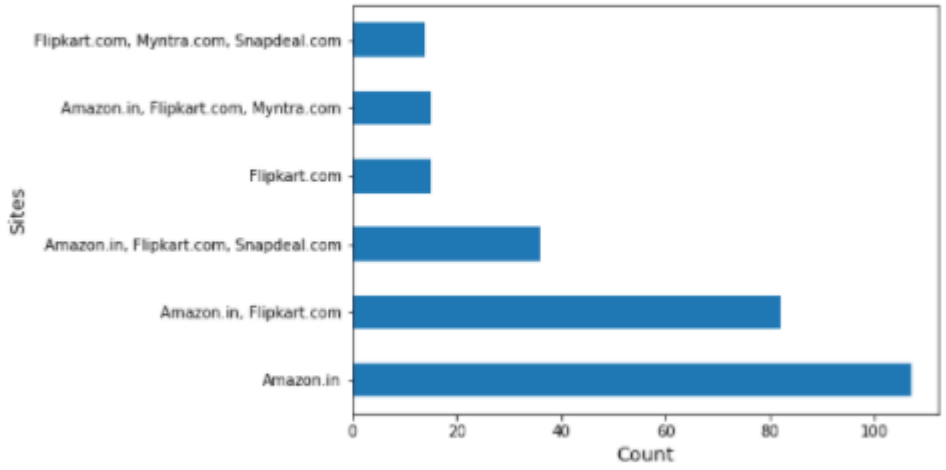
```
degree of freedom is 15
stats is 49.038004229846884
p values id 1.7272959701139484e-05
```

```
prob = 0.95
critical = chi2.ppf(prob, dof)
result = {'critical': critical, 'Chi2': stat}
result = pd.DataFrame(result, index = ['values'])
result.plot.bar();
```



As the chi2 value is more then critical value there is a significance dependence between device and time of online purchase.

```
# Site which delivers fastly

df['speed_odr_del'].value_counts().plot.barh(figsize = (7,5))
plt.xlabel('Count', fontsize = 13)
plt.ylabel('Sites', fontsize = 13);
```



Amazon is leading for speed order delivery

Plotting Sites and Delivery Speed.

```
shops = pd.crosstab(df['onlin_reatl_shp_frm'], df['speed_odr_del'])
shops
```

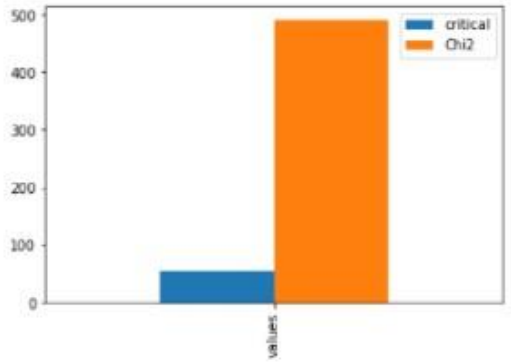
speed_odr_del onlin_reatl_shp_frm	Amazon.in	Amazon.in, Flipkart.com	Amazon.in, Flipkart.com, Myntra.com	Amazon.in, Flipkart.com, Snapdeal.com	Flipkart.com	Flipkart.com, Myntra.com, Snapdeal.com
Amazon.in	16	0	0	0	0	0
Amazon.in, Flipkart.com	5	19	0	0	8	0
Amazon.in, Flipkart.com, Myntra.com, Snapdeal.com	0	15	15	0	0	14
Amazon.in, Flipkart.com, Paytm.com	0	0	0	0	7	0
Amazon.in, Flipkart.com, Paytm.com, Myntra.com, Snapdeal.com	46	11	0	25	0	0
Amazon.in, Flipkart.com, Paytm.com, Snapdeal.com	0	18	0	11	0	0
Amazon.in, Flipkart.com, Snapdeal.com	8	19	0	0	0	0
Amazon.in, Paytm.com	12	0	0	0	0	0
Amazon.in, Paytm.com, Myntra.com	20	0	0	0	0	0

lets see if people selecting sites to purchase as an impact on delivery speed

```
stat, p, dof, expected = chi2_contingency(shops)
print(f'degree of freedom is {dof}')
print(f'stats is {stat}')
print(f'p values id {p}')
```

```
degree of freedom is 40
stats is 490.4318950886271
p values id 7.165785301206301e-79
```

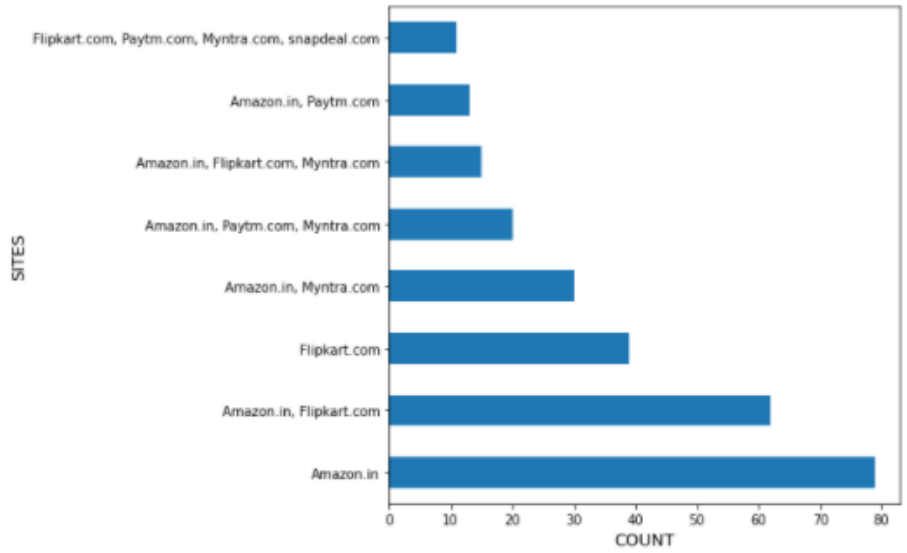
```
prob = 0.95
critical = chi2.ppf(prob, dof)
result = {'critical': critical, 'Chi2': stat}
result = pd.DataFrame(result, index = ['values'])
result.plot.bar();
```



As the chi2 value is more then critical value there is a significance dependence between choosing site and time for login

LETS SEE WHICH SITES ARE MOSTLY RECOMMENDED

```
df['Recommend'].value_counts().plot.barh(figsize = (7,7))
plt.xlabel('COUNT', fontsize = 13)
plt.ylabel('SITES', fontsize = 13);
```



Amazon is highly recommended and followed by flipkart.

```
recom = pd.crosstab(df['speed_odr_del'], df['Recommend'])
recom
```

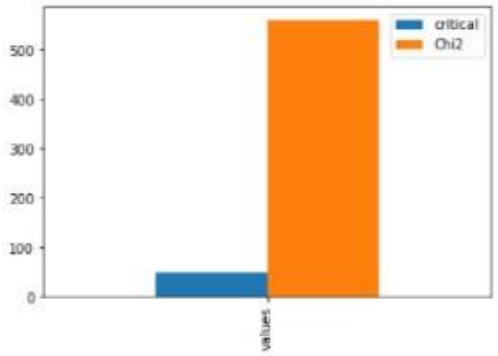
Recommend speed_odr_del	Amazon.in	Amazon.in, Flipkart.com	Amazon.in, Flipkart.com, Myntra.com	Amazon.in, Myntra.com	Amazon.in, Paytm.com	Amazon.in, Paytm.com, Myntra.com	Flipkart.com	Flipkart.com, Paytm.com, Myntra.com, snapdeal.com
Amazon.in	47	0	0	15	13	20	12	0
Amazon.in, Flipkart.com	11	37	0	15	0	0	19	0
Amazon.in, Flipkart.com, Myntra.com	0	0	15	0	0	0	0	0
Amazon.in, Flipkart.com, Snapdeal.com	0	25	0	0	0	0	0	11
Flipkart.com	7	0	0	0	0	0	8	0
Flipkart.com, Myntra.com, Snapdeal.com	14	0	0	0	0	0	0	0

CHECKING THEIR DEPENDENCY

```
stat, p, dof, expected = chi2_contingency(recom)
print(f'degree of freedom is {dof}')
print(f'stats is {stat}')
print(f'p values id {p}')
```

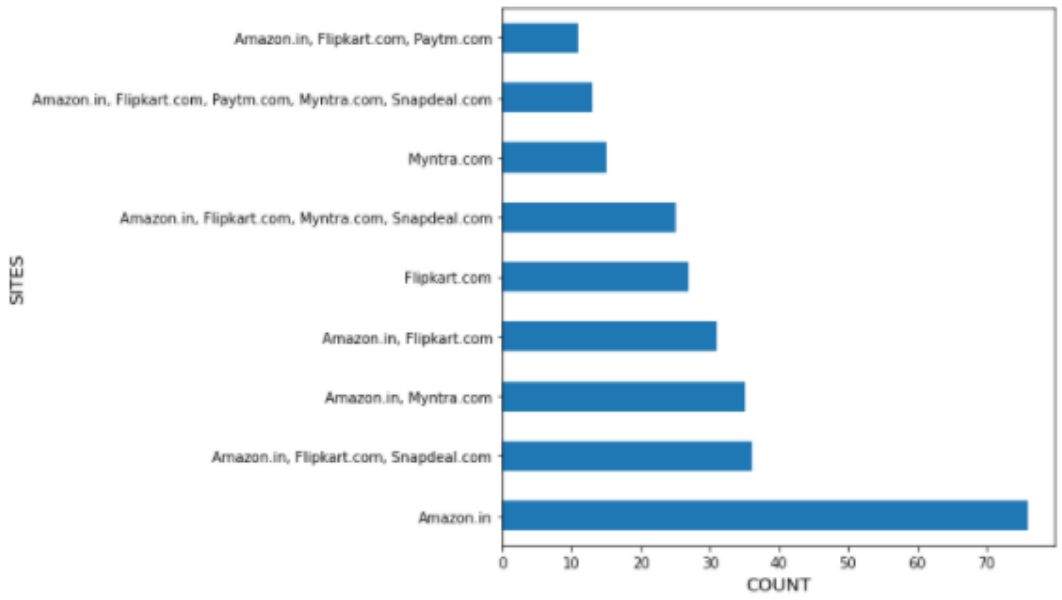
```
degree of freedom is 35
stats is 559.5687910398661
p values id 9.065945953119557e-96
```

```
prob = 0.95
critical = chi2.ppf(prob, dof)
result = {'critical': critical, 'Chi2': stat}
result = pd.DataFrame(result, index = ['values'])
result.plot.bar();
```



As the chi2 value is more then critical value there is a significance dependence between recommending site and speed delivery of order.

```
df['perceived_trust'].value_counts().plot.barh(figsize = (7,7))
plt.xlabel('COUNT', fontsize = 13)
plt.ylabel('SITES', fontsize = 13);
```



AMAZON has received more amount of trust compare to other sites.

CHECKING FOR DEPENDENCY FOR RECOMMENDATION AND TRUST

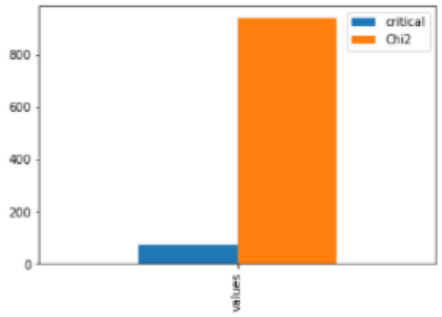
```
trust = pd.crosstab(df['perceived_trust'],df['Recommend'])
trust
```

Recommend	Amazon.in	Amazon.in, Flipkart.com	Amazon.in, Flipkart.com, Myntra.com	Amazon.in, Myntra.com	Amazon.in, Paytm.com	Amazon.in, Paytm.com, Myntra.com	Flipkart.com	Flipkart.com, Paytm.com, Myntra.com, snapdeal.com
perceived_trust								
Amazon.in	42	19	15	0	0	0	0	0
Amazon.in, Flipkart.com	5	7	0	0	0	0	19	0
Amazon.in, Flipkart.com, Myntra.com, Snapdeal.com	14	11	0	0	0	0	0	0
Amazon.in, Flipkart.com, Paytm.com	11	0	0	0	0	0	0	0
Amazon.in, Flipkart.com, Paytm.com, Myntra.com, Snapdeal.com	0	0	0	0	13	0	0	0
Amazon.in, Flipkart.com, Snapdeal.com	0	25	0	0	0	0	0	11
Amazon.in, Myntra.com	0	0	0	15	0	20	0	0
Flipkart.com	7	0	0	0	0	0	20	0
Myntra.com	0	0	0	15	0	0	0	0

```
stat, p, dof, expected = chi2_contingency(trust)
print(f'degree of freedom is {dof}')
print(f'stats is {stat}')
print(f'p values id {p}')
```

```
degree of freedom is 56
stats is 939.550970270752
p values id 1.2847037251980906e-160
```

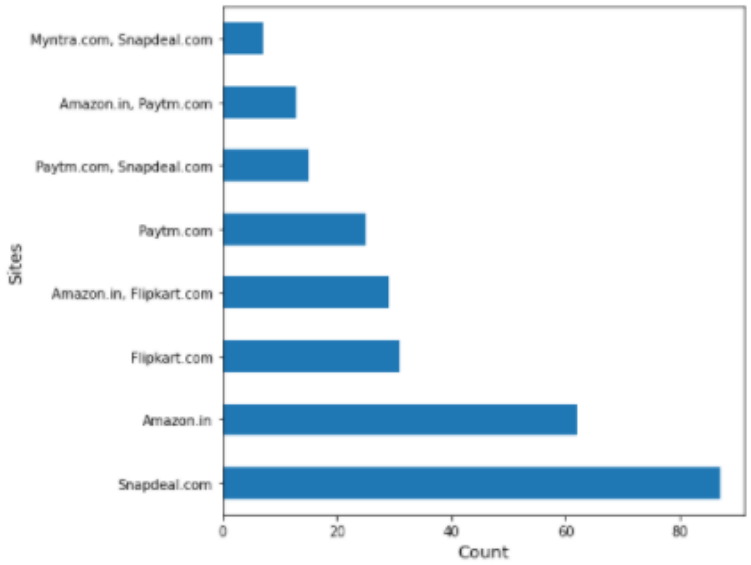
```
prob = 0.95
critical = chi2.ppf(prob, dof)
result = {'critical': critical, 'Chi2': stat}
result = pd.DataFrame(result, index = ['values'])
result.plot.bar();
```



As the chi2 value is more then critical value there is a significance dependence between recommending site and perceived trust.

Visualizing which site has less limited mode of payment.

```
df['limtd_mode_pay'].value_counts().plot.barh(figsize = (7,7))
plt.xlabel('Count', fontsize = 13)
plt.ylabel('Sites',fontsize = 13);
```



Its clearly visualized that snapdeal has limited mode of payment.

checking for dependencies of each limited mode of payment with recommending.

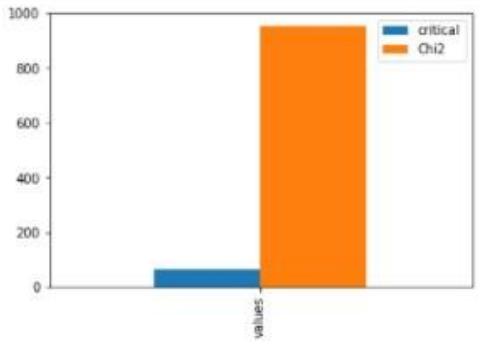
```
payment = pd.crosstab(df['limtd_mode_pay'], df['Recommend'])
payment
```

Recommend	Amazon.in	Amazon.in, Flipkart.com	Amazon.in, Flipkart.com, Myntra.com	Amazon.in, Myntra.com	Amazon.in, Paytm.com	Amazon.in, Paytm.com, Myntra.com	Flipkart.com	Flipkart.com, Paytm.com, Myntra.com, snapdeal.com
limtd_mode_pay								
Amazon.in	30	0	0	0	0	20	12	0
Amazon.in, Flipkart.com	29	0	0	0	0	0	0	0
Amazon.in, Paytm.com	0	0	0	0	13	0	0	0
Flipkart.com	12	0	0	0	0	0	8	11
Myntra.com, Snapdeal.com	0	7	0	0	0	0	0	0
Paytm.com	0	25	0	0	0	0	0	0
Paytm.com, Snapdeal.com	0	0	15	0	0	0	0	0
Snapdeal.com	8	30	0	30	0	0	19	0

```
stat, p, dof, expected = chi2_contingency(payment)
print(f'degree of freedom is {dof}')
print(f'stats is {stat}')
print(f'p values id {p}')
```

```
degree of freedom is 49
stats is 953.0835699613103
p values id 7.897334953205374e-168
```

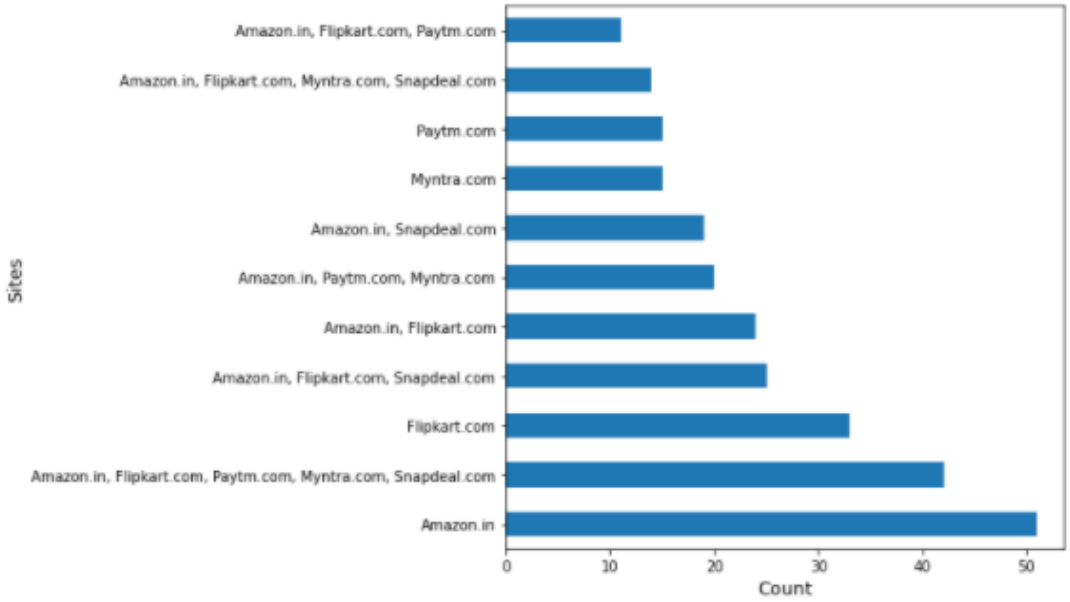
```
prob = 0.95
critical = chi2.ppf(prob, dof)

result = {'critical': critical, 'Chi2': stat}

result = pd.DataFrame(result, index = ['values'])
result.plot.bar();
```



As the chi2 value is more then critical value there is a significance dependence between recommending site and limited mode of payment.

VISUALIZING THE SECURITY ON CUSTOMER INFO

```
df['security_cust_fin_info'].value_counts().plot.barh(figsize = (7,7))
plt.xlabel('Count', fontsize = 13)
plt.ylabel('Sites', fontsize = 13);
```



1. Amazon is leading here for security payment
2. Also all other sites have got more or less equal votes

CHECKING FOR DEPENDENCY OF ON EACH VARIABLES

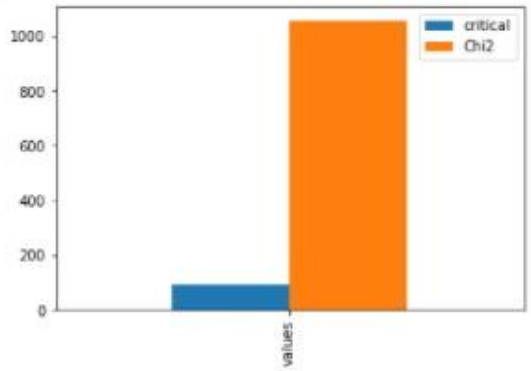
```
info = pd.crosstab(df['security_cust_fin_info'], df['Recommend'])
info
```

security_cust_fin_info \ Recommend	Amazon.in	Amazon.in, Flipkart.com	Amazon.in, Flipkart.com, Myntra.com	Amazon.in, Myntra.com	Amazon.in, Paytm.com	Amazon.in, Paytm.com, Myntra.com	Flipkart.com	Flipkart.com, Paytm.com, Myntra.com, snapdeal.com
Amazon.in	24	0	15	0	0	0	12	0
Amazon.in, Flipkart.com	5	0	0	0	0	0	19	0
Amazon.in, Flipkart.com, Myntra.com, Snapdeal.com	14	0	0	0	0	0	0	0
Amazon.in, Flipkart.com, Paytm.com	11	0	0	0	0	0	0	0
Amazon.in, Flipkart.com, Paytm.com, Myntra.com, Snapdeal.com	0	18	0	0	13	0	0	11
Amazon.in, Flipkart.com, Snapdeal.com	0	25	0	0	0	0	0	0
Amazon.in, Paytm.com, Myntra.com	0	0	0	0	0	20	0	0
Amazon.in, Snapdeal.com	0	19	0	0	0	0	0	0
Flipkart.com	25	0	0	0	0	0	8	0
Myntra.com	0	0	0	15	0	0	0	0
Paytm.com	0	0	0	15	0	0	0	0

```
stat, p, dof, expected = chi2_contingency(info)
print(f'degree of freedom is {dof}')
print(f'stats is {stat}')
print(f'p values id {p}')
```

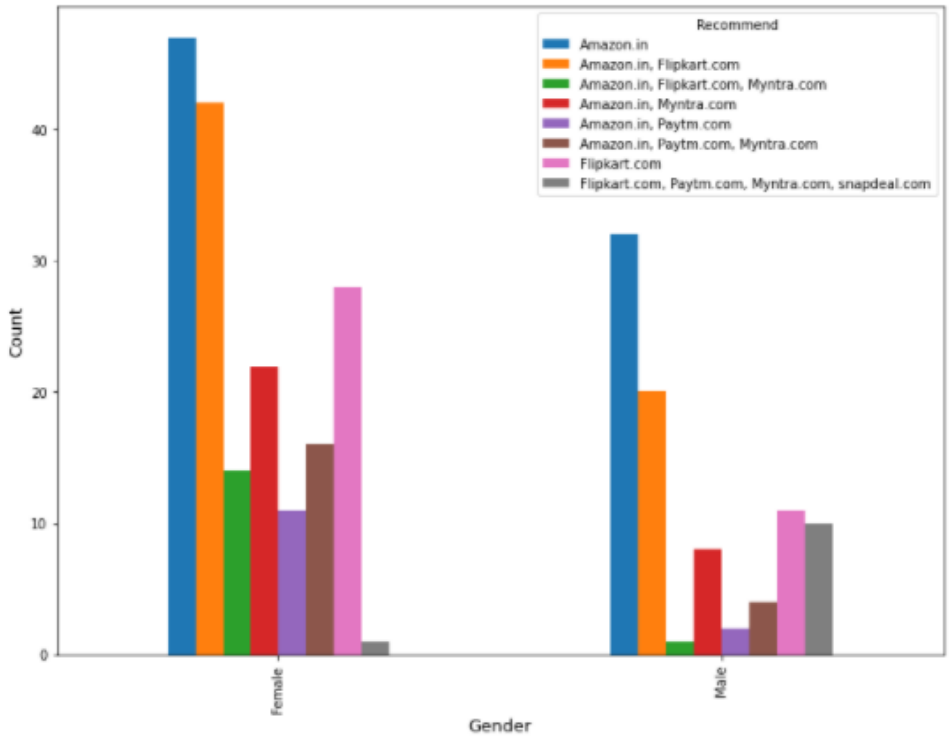
```
degree of freedom is 70
stats is 1054.4266384574735
p values id 1.3817450273358843e-175
```

```
prob = 0.95
critical = chi2.ppf(prob, dof)
result = {'critical': critical, 'Chi2': stat}
result = pd.DataFrame(result, index = ['values'])
result.plot.bar();
```



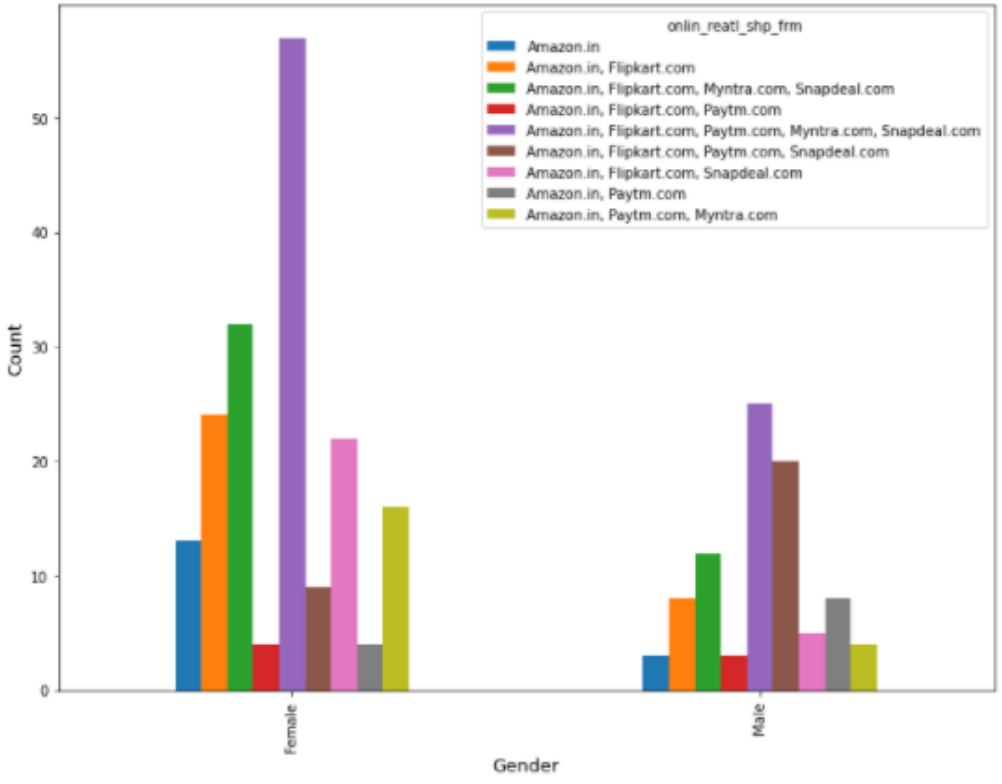
As the chi2 value is more then critical value there is a significance dependence between recommending site and security on customer info.

```
pd.crosstab(df['Gender'], df['Recommend']).plot.bar(figsize =(12,9))
plt.xlabel('Gender', fontsize = 13)
plt.ylabel('Count', fontsize = 13);
```



In both cases all are recommending amazon the most.

```
pd.crosstab(df['Gender'], df['onlin_reatl_shp_frm']).plot.bar(figsize =(12,9))
plt.xlabel('Gender', fontsize = 13)
plt.ylabel('Count', fontsize = 13);
```



people are using all the websites for online purchase though they recommend amazon the most.

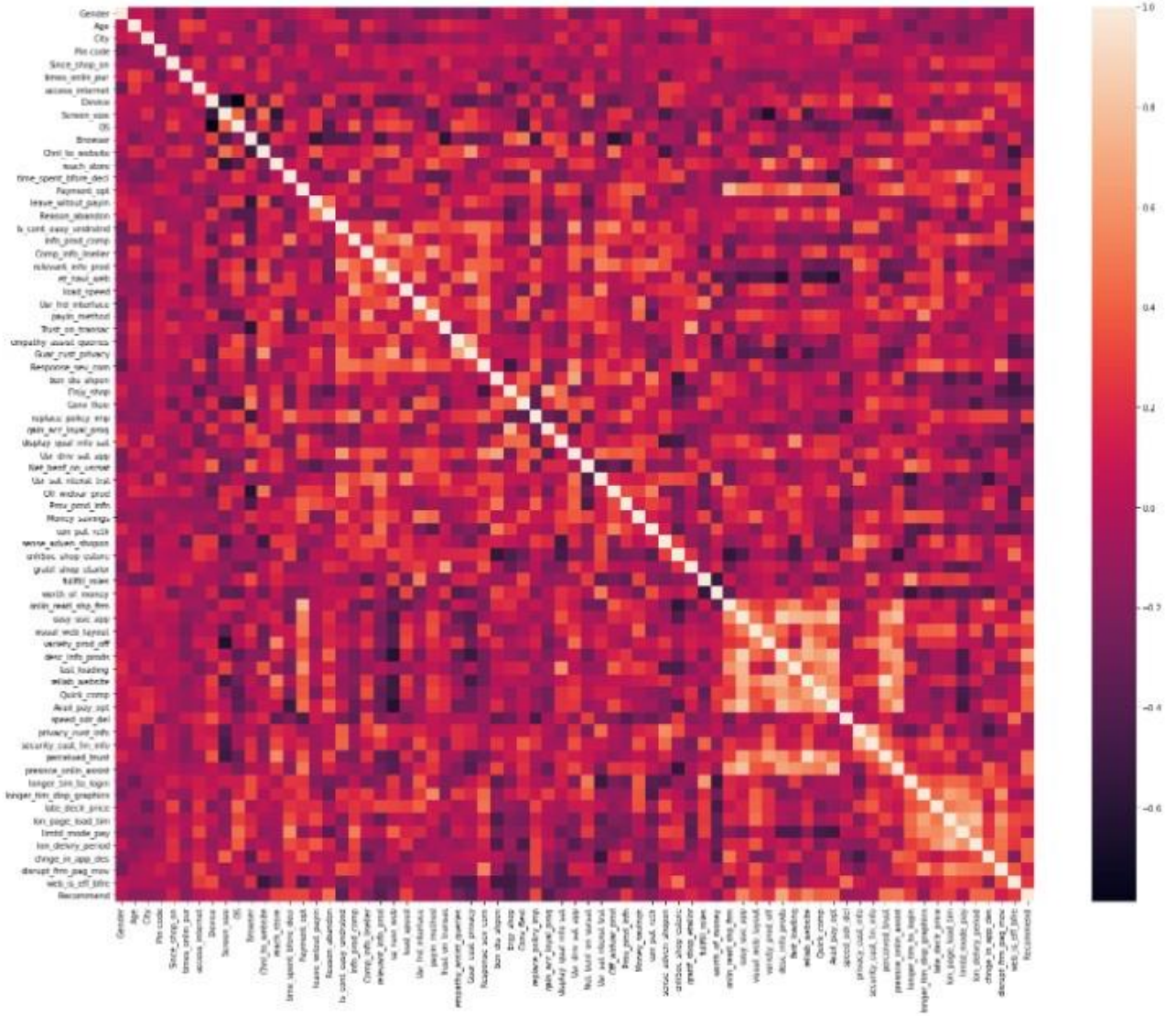
```
le = LabelEncoder()
df_label = df.copy()
```

```
for col in df_label.columns:
    df_label[col] = le.fit_transform(df_label[col])
```

```
df_label.head()
```

	Gender	Age	City	Pin code	Since_shop_on	times_onlin_pur	access_internet	Device	Screen_size	OS	...	longer_tim_to_login	longer_tim_disp_graphics
0	1	1	2	1	3	2	0	0	3	2	...	0	0
1	0	0	2	5	3	3	3	2	0	1	...	1	6
2	0	0	4	23	2	3	1	2	2	0	...	7	6
3	1	0	6	11	2	5	1	2	2	1	...	9	7
4	0	0	0	31	1	0	3	2	0	1	...	5	8

```
plt.figure(figsize = (25,20))
corr = df_label.corr()
sns.heatmap(corr);
```



We can see that there are many variables which have high and low correlation with other independent variables. We can also use correlation matrix to check the relation/ dependencies between the independent and target variables. We can remove one of the highly correlated independent variable to get rid of the problem of multicollinearity.

```
corr = df_label.corr().unstack()
c = corr.sort_values(kind = 'quicksort')
c[c.values > 0.7].drop_duplicates()
```

```
53]: security_cust_fin_info  privacy_cust_info      0.709004
     limtd_mode_pay          lon_page_load_tim      0.712607
     variety_prod_off        reliab_website         0.730907
     easy_use_app            reliab_website         0.731459
                             fast_loading           0.747567
     desc_info_prods         reliab_website         0.751957
     Enjy_shop               gain_acc_loyal_prog    0.761451
     onlin_reatl_shp_frm     Payment_opt            0.767374
     presnce_onlin_assist    Avail_pay_opt          0.769993
                             easy_use_app           0.791192
     lon_delvry_period       lon_page_load_tim      0.797480
     desc_info_prods         Avail_pay_opt          0.801683
     easy_use_app            Avail_pay_opt          0.834385
     desc_info_prods         easy_use_app           0.863354
     reliab_website          perceived_trust        0.933661
     Gender                  Gender                 1.000000
     dtype: float64
```

The top features which are correlated are shown above.

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.

1. Lenovo flex corei5 laptop

2. Jupyter Notebook :The Jupyter Notebook is an interactive environment for running code in the browser. It is a great tool for exploratory data analysis and is widely used by data scientists.

3. MS PowerPoint: For preparing presentation of project.

4. MS word: For preparing report

5. Pandas: is a software library written for the Python programming language for data manipulation and analysis.

6. Numpy: is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices.

7. Matplotlib: is a plotting library for the Python programming language and its numerical mathematics extension

8. NumPy.Seaborn: is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

9. Sklearn: Scikit-learn (formerly scikits. learn and also known as sklearn) is a free software machine learning library for the Python programming language.

10. LabelEncoder: In label encoding in Python, we replace the categorical value with a numeric value between 0 and the number of classes minus 1. If the categorical variable value contains 5 distinct classes, we use (0, 1, 2, 3, and 4).

CONCLUSION

In this paper we have evaluated the factors influencing sales and retention of Online e-commerce. We have filtered down the major features contributing sales from the 71 attributes. The findings showed that responsiveness, customer satisfaction, ease of use, and attitude are critical for the online customer retention. In addition, the findings showed that online trust is important for the customer and retailers must establish a trusting relationship with the customers. Also it is noticed the online shoppers prefer Amazon.com most over other shopping sites in terms of trust, mode of payment, security and speed of delivery.