Customer Behavior on The OLX Platform

By

Muhammad Umar Salman 20100120

Sheikh Abdul Mannan 20100261

Advisor: Dr. Fareed Zaffar

Senior Project

2020

Department of Computer Science

Syed Baber Ali School of Science & Engineering

Lahore University of Management Sciences (LUMS)

Certificate

I certify that the senior project titled **Customer Behavior on The OLX Platform** was completed under my supervision by the following students:

Muhammad Umar Salman 2020-10-0120

Sheikh Abdul Mannan 2020-10-0261

and the project deliverables meet the requirements of the program.

Signed By Dr. Fareed Zaffar

------------------------------------ Date: May 6th, 2020

Advisor (Signature)

N/A

------------------------------------ Date:

Co-advisor (if any)

Contents

Abstract

OLX is an online platform where people can advertise anything that they want to sell and other people can view these ads, contact the seller, and purchase the item. Recently, the website has allowed real estate to be advertised and sold on their platform. Tens of thousands of people navigate the property section of the website each day. Looking through logs of event-streams, we identified and observed unique event pathways through exploratory data analysis taken by a user and to solve the issue of not knowing if the purchase went through or not, we devised a metric of conversion for some customers to daily buyers. We also used pattern mining techniques such as the apriori algorithm to see which events are most commonly with each other. Furthermore, we performed both click-stream and cluster analysis to identify similar behavioral patterns that users took on the web platform in an attempt to create a classifier that can identify with high accuracy the users which are serious about purchasing property and which users are casually visiting the website without the intention of purchasing a property. We are currently working on a classification system that determines whether a user can be considered a serious buyer based on their initial few events on the platform.

1. Introduction

What is OLX?

OLX Group is a **global online marketplace**, founded in 2006 and operating in 45 countries. The OLX marketplace is a platform for buying and selling services and goods such as electronics, fashion items, furniture, household goods, cars and bikes and now also real estate in Pakistan. On OLX's online platform users usually post their ads and customers when navigate on to the platform in attempt to but whichever category it is that they need, click on one of these posted ads and if he chooses to buy it, he either clicks the chat on SMS option or can directly call the person who posted the ad with the phone number provided. The following two images are screenshots of what the online web platform looks like.

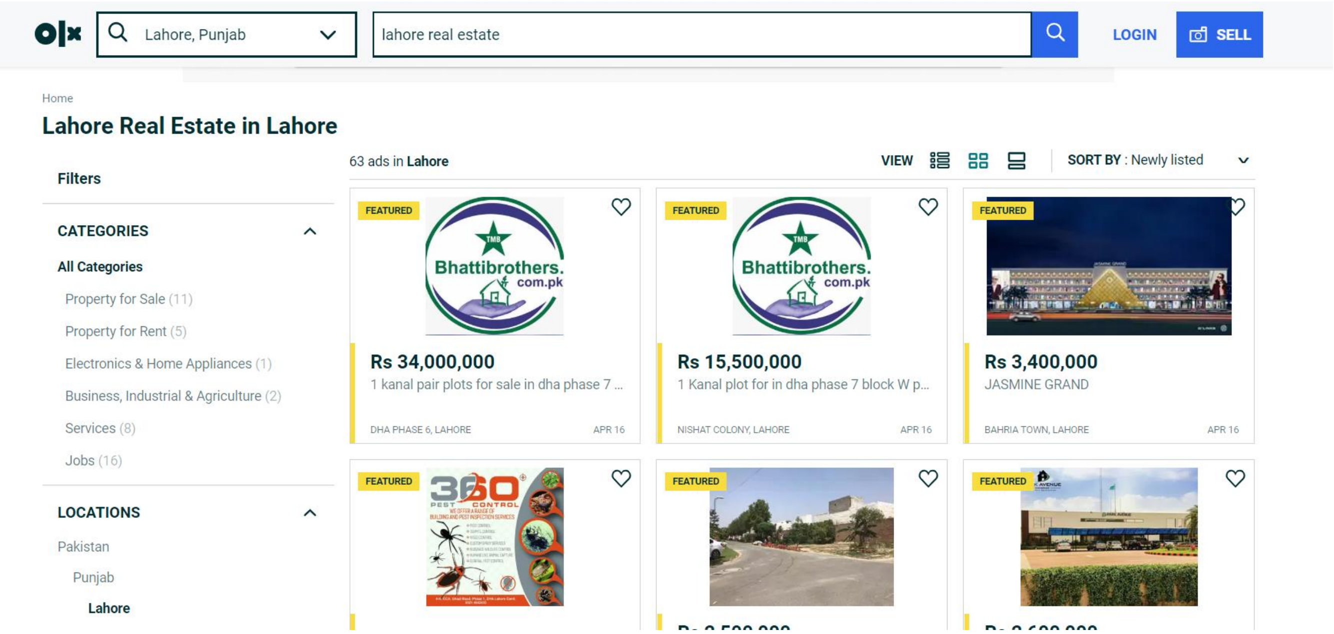
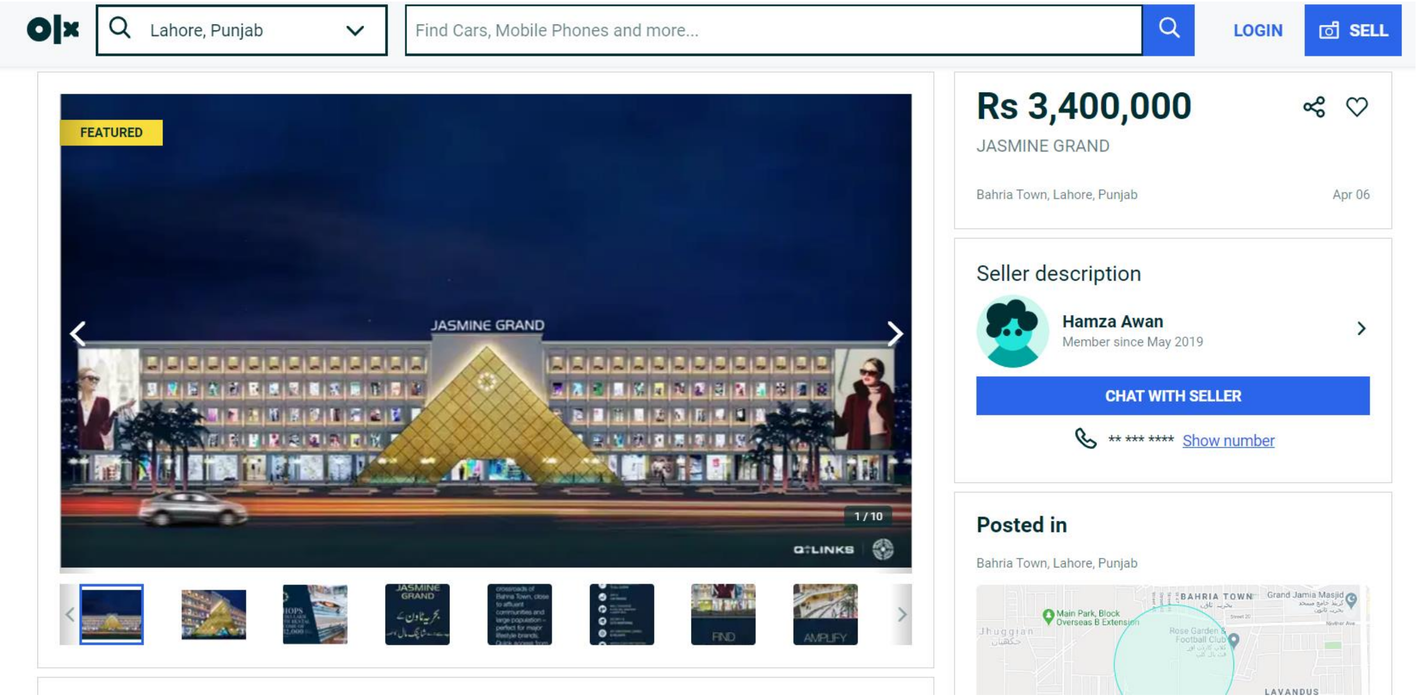


Figure 1.1.1



Rs 3,400,000

JASMINE GRAND

Bahria Town, Lahore, Punjab

Apr 06

Seller description

Hamza Awan

Member since May 2019

Posted in

Bahria Town, Lahore, Punjab

Figure 1.1.2

Thus, OLX doesn't receive information as to whether the Ad posted has been purchased or not. Even after calling up the person who posted the Ad, there are chances that the purchase doesn't go through due to various reasons. For that reasons certain assumption are made as to what events the user takes on the platform which can be considered as purchase of the Ad.

The Datasets

In this study **3 different datasets** have been used for different analysis. Each dataset has data on every event a user takes on whichever platform he uses regarding searches specific to **real estate**. The primary reason for this was due to receiving these different datasets at different instances of time. The three consist of 2 datasets which have been taken from the Web platform and 1 which has been taken from the Android phone platform. Details of each dataset has been given below.

Web

File	**Starting timestamp**	**Ending timestamp**	**Num of records**
Feb	2/1/2019 0:00	2/28/2019 23:50	1152436
March	3/1/2019 0:05	3/31/2019 23:55	589723
April May	4/1/2019 0:09	5/31/2019 23:59	575130
June	6/1/2019 0:00	6/30/2019 23:52	203157
July	7/1/2019 0:12	7/31/2019 23:39	210358
			2,730,804

Android

File	Starting timestamp	Ending timestamp	Num of records
Feb	2/1/2019 0:00	2/28/2019 23:59	4,164,681
March	3/1/2019 0:00	3/31/2019 23:59	2,800,380
April	4/1/2019 0:00	4/30/2019 23:58	1,969,188
May	5/1/2019 0:08	5/31/2019 23:59	1,492,041
June	6/1/2019 0:00	6/30/2019 23:59	1,272,847
July	7/1/2019 0:00	7/31/2019 23:59	1,156,424
August	8/1/2019 0:08	8/30/2019 23:59	903,825
			13,759,386

Web 2

Starting timestamp	Ending timestamp	Number Of Records
1/1/2019 0:00	8/31/2019 23:59	**3436444**

The difference between the Web dataset and the Web 2 dataset is that Web 2 has 31 features (columns) whereas the Web dataset has 118 features, this study initial analysis was made on Web 2 and Android data and later on the Web dataset was used.

Datasets	Shape of Dataset
Web	(2730804, 118)
Android	(13759386, 118)
Web 2	(3436444, 31)

Next, we will look into some of the important columns and see what they mean.

`meta_acceptcookies`	T/F value of whether the person has a Cookie on the OLX platform
`meta_date`	Complete Timestamp of when the event took place
`meta_invite_source`	From which site where they redirected to the OLX platform
`meta_session`	A session id is given to a user that comes on to the platform and after inactivity for 30 mins, the session ends
`meta_session_long`	The Unique id given to each user based on his account and machines IP address
`params_category_level1_id`	Category Level 1 defines a broader category of real estate later discusses in the paper
`params_category_level2_id`	Category Level 2 defines a more specific category of real estate later discusses in the paper
`params_en`	Events taken by a user
`params_filters`	Whether the user selected any filters
`params_images_count`	The count of images the Ad posted/viewed had
`params_lat`	Latitude from where the event was generated
`params_long`	Longitude from where the event was generated
`params_search_type`	Depends whether what they searched for was autocompleted or self-typed

2. Related Works

2.1. Background

Many works have been previously done on online marketplaces for different categories of work and each work has defined methodologies which they have adopted for their papers. Our primary interest was in the behavioral pattern of the customers as to which actions they take on the website platform and if we can tell if they are serious buyers or just window shoppers and if we can build profiles through **cluster analysis** as to what sequence of events may lead to a potential buy. Rather than what most papers have done citing statistics on behavioral trends taken on online platforms, we have attempted to see what metrics can help classify a user as buyer based on his/her events, used pattern mining to see which events most commonly go with each other and then we use clustering to see most common sequential events and cluster them to see some sort of behavioral trend.

We divide the related works that we found for online marketplace purchasing and selling into 3 broad categories and then a paper from which we took inspiration from for our exploratory data analysis. The following are the three categories

1. Content Quality
2. Fraud and Risks
3. User Click Behavior

Inspiration Paper: Characterizing Key Stakeholders in an Online Black-Hat Marketplace

2.2. Content Quality

1. **A framework for the selection of electronic marketplaces: a content analysis approach (2002)**
    - A content analysis of research and practitioner articles is carried out to evaluate the issues that prospective participants, seeking to purchase goods and services online, need to address in their selection process.
2. **Internet users' perception of online service quality (2003)**
    - Service quality is widely accepted as one of the key determinants of online retailers' success. This explanatory study identified 4 key dimensions of online service quality as perceived by two groups of internet users, online buyers and information searchers.
3. **Reducing consumer risk in electronic marketplaces (2018)**
    - The study focuses on how information provided by sellers about themselves (i.e., seller information) and about their products (i.e., product information) can function as risk reduction signals and how these affect a buyer's inclination to purchase. Combining signaling theory with perceived risk theory, the authors present a research model that they test using structural equation modeling with data collected in two different electronic marketplaces, including eBay.nl.

2.3. Fraud and Risks

1. **Craigslist Scams and Community Composition (2013)**
    - This paper examines the prevalence of Craigslist-based (automobile) scams across 30 American cities. Our methodology analyses historical scam data and its relationship with economic, structural, and cultural characteristics of the communities that are exposed to fraudulent

advertising. We find that Craigslist scams are not random but targeted towards specific communities.

2. **Fraud Detection in Web Advertisement (2015)**
   - In recent years fraud is major problem in online advertising. It can affect the trust, beliefs and encouragement of the customer on online marketing. In this thesis, the development of this system can be done using Naive Bayes classifier and Apriori algorithm. The system can find fraud or scam in web-based marketing and advertisement.
3. **Perceived risk and trust associated with purchasing at electronic marketplaces (2006)**
   - In this paper, we study the relationships between consumer perceptions of risk and trust and the attitude towards purchasing at a consumer-to-consumer electronic marketplace (EM). Typical for EM settings is that consumer behavior is subject to perceptions of the selling party as well as of the institutional structures of the intermediary that is operating the EM.
4. **Trust Among Strangers in the Internet Transactions: Empirical Analysis of eBay's Reputation System (2001)**
   - This paper focuses on explaining how buyers trust unknown sellers, it talks about the success of the Reputation System and Feedback System employed by eBay and examines why and how the system works. It provides information that allows buyers to distinguish between trustworthy and non-trustworthy sellers and encourages sellers to be trustworthy and discourage participation from those who aren't.

2.4. User Click Behavior

1. **Influencing the Online Consumer's behavior: the web experience (2004)**
   - The study addresses the issue of how to attract and win over the user in the highly competitive internet marketplace. It analyzes the factors affecting the consumer's behavior which include functionality, psychological and content factors. These factors have a direct and crucial influence on the consumer decision and the effect it has on the buying process.
2. **Consumer Behavior in Web-Based Commerce: An Empirical Study (2014)**
   - This study examines how certain consumers and website factors influence the online consumer experience. The study finds that perceived control and shopping enjoyment can increase the intention of new Web customers to return, but seemingly do not influence repeat customers to return. It also finds that a Web store that utilizes value added search mechanisms and presents a positively challenging experience can increase customers shopping enjoyment.
3. **Recommending or Persuading? The impact of Shopping Agent's Algorithm on User Behavior**
   - In this study we examine how the type of information that is elicited by a shopping agent for use in its recommendation algorithm may affect consumers' preference for product features and ultimately their product choice in an electronic marketplace. e. A recommendation agent both calibrates a model of a user's preference based on his/her input and uses this model to make personalized product recommendations and that through the design of their recommendation algorithms, there is a potential to influence user preferences/behavior.

2.5. Characterizing Key Stakeholders in an Online Black-Hat Marketplace

The goal of this paper was to present a detailed micro-economic analysis of a popular online black-hat marketplace, namely, SEOClerks.com. As the site provides non anonymized transaction information, the paper is set to analyze selling and buying behavior of individual users, propose a strategy to identify key users, and study their tactics as compared to other (non-key) users. The study finds that key users: (1) are mostly located in Asian countries, (2) are focused more on selling black-hat SEO services, (3) tend to list more lower priced services, and (4) sometimes buy services from other sellers and then sell at higher prices. Finally, it discusses the implications of its analysis with respect to devising effective economic and legal intervention strategies against marketplace operators and key users.

Below is a list of analysis conducted by the paper.

List Of Analysis

1. General Statistics - Summarizes the overall statistics of the SEOClerks Marketplace
2. Show the relationship between the seller join date, last login date and revenue.
3. Plots the daily registration rate of new users and the cumulative number of users on SEOClerks
4. Plots the revenue distribution for sellers on SEOClerks
5. Distribution of last login date to check activity of users
6. Plots of distribution of service prices for key and non-key users
7. Shows distribution of service volume and revenue for key and non-key users
8. Plots of distribution view count for services offered by key users and non-key users
9. View the correlation between service view count and sale volume of key and non-key users
10. Plots the distribution of the number of services listed by key and non-key users on SEOClerks
11. Plots the distribution of seller revenue for key and non-key users
12. Plots the distributions of the purchase volume by key and non-key users
13. Plots the distributions of buyer expense (the total amount of money spent by a buyer) by key and non-key users
14. Visualize the purchases made by key users from key and non-key users and vice-versa

3. Implementation and Evaluation

The implementation and evaluation phase can broadly be categorized into 4 categories. The following list shows these categories along with their subcategories.

1. Exploratory Data Analysis
2. Buyer Conversion Metric
   - Time
   - Categories
   - Price
3. Frequent Pattern Mining
4. Cluster Analysis
   - TSNE Clustering
   - K-Means Clustering on events
   - K-Means Clustering on sequential events using NLP

3.1. Exploratory Data Analysis

The data analysis was mostly done on different data sets due to the availability at different times of datasets, and sometimes a subset of the data was used due to low computational capacity.

This first analysis is done on the first month of the **Android** data, which would be **February 2019**

Users can be identified by two identifiers, **user_nk** and **session_long**. User_nk is the ID of a user that has an account on the OLX platform. This is evident in the fact since not all users have this identifier. There are 236,327 unique user_nk values in the data. Session_long is another identifier which is used as a cookie for identifying users regardless of whether the person has an OLX account or not. All users have a session_long, for every of the 22,361,655 entries in the data and there are 22,361,655 entries for the session long column. Out of these, there are 277,926 unique session_long values. This can be seen in Figure 3.1.1

Upon further inspection it was seen that certain users have multiple session longs in the range of 2-20 with far lesser users at the extreme and that majority of users had just one session long.41,559 users do not have a user_nk identifier meaning they did not make any account before using the platform.

There is also a field **'cross_domain_session_long'**. The name suggests that it can track users across domains, like a user can be identified if they are on the android and web platform as well. There are 343,330 unique values.

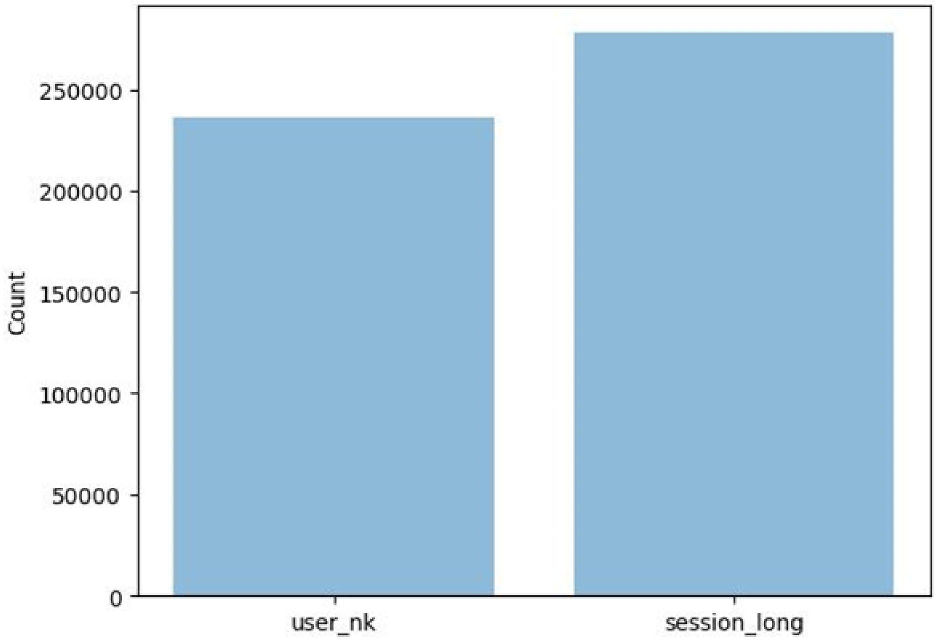


Figure 3.4.1

There are two columns which contain the **latitude** and **longitude** of users. 49,649 unique pairs of latitudes and longitudes were found, out of these 610 did not belong to Pakistan. Figure 3.1.2 below shows a heatmap of the world.

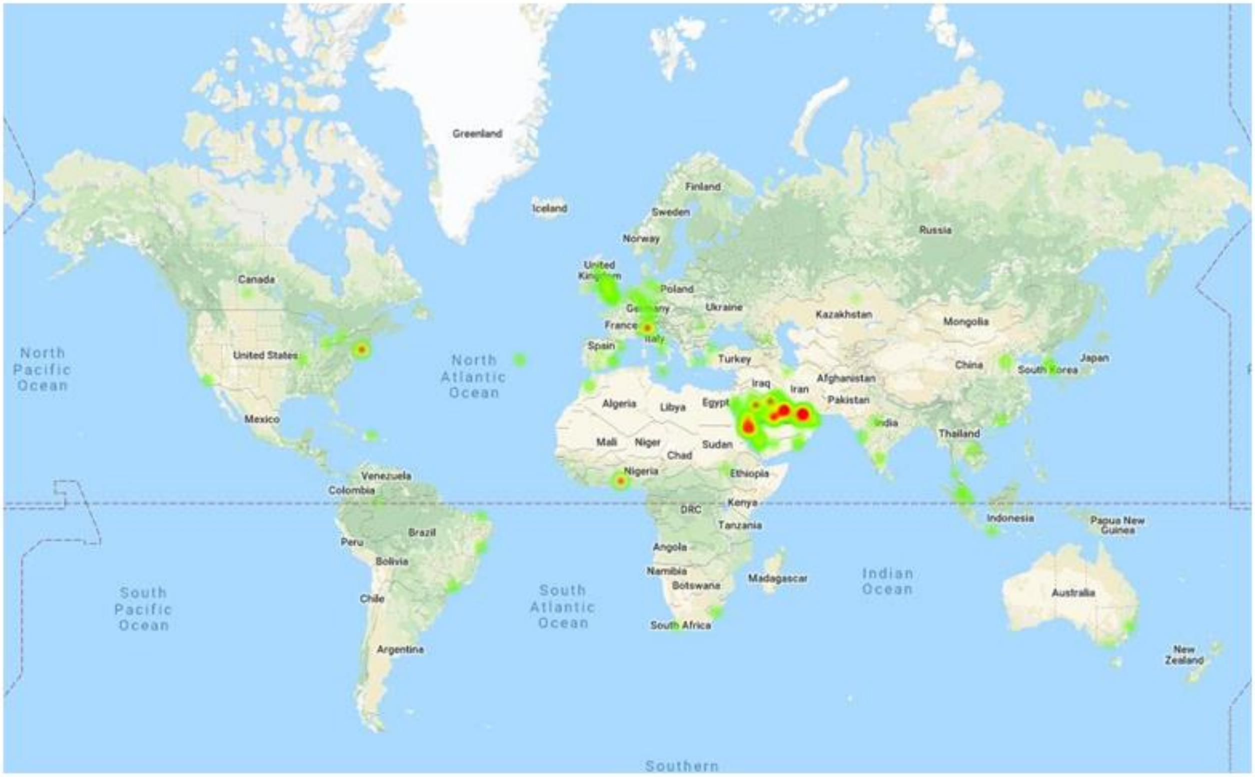


Figure 3.1.2

This image shows the location of users that are not from Pakistan. Majority of them come from the middle east, with a considerable number of users from Europe. A small number of users can also be seen from the Americas or South East Asia.

There are **two** apparent reasons for this:

1. These are valid and real users.
2. Certain Users used a VPN server due to which the location of the VPN server was used while browsing the app.

As for Pakistan the capital cities of each province are hotspots on the map, with a huge number of users spread across Punjab which is indicative of its high population density. Figure 3.1.3 shows the locations from where events were generated in Pakistan

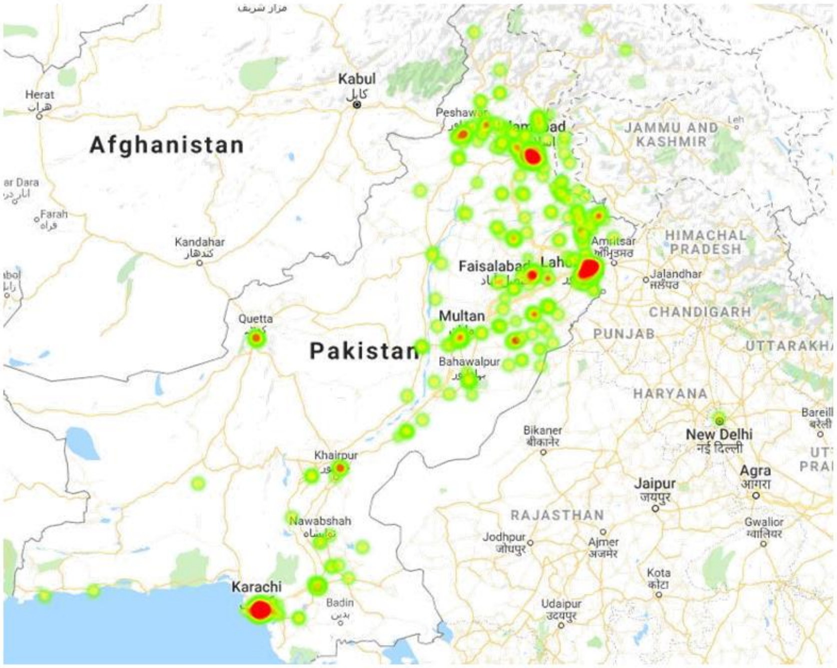


Figure 3.1.3

After having looked at these we then went on to look at user's activity for a week, this was done by looking at the users who had come for the first time on 1st Feb and we only viewed those users for the rest of the week. By this we could extract a history of **seven consecutive days** on the platform which was done by keeping a rolling period for each user. This can be seen below in Figure 3.1.4

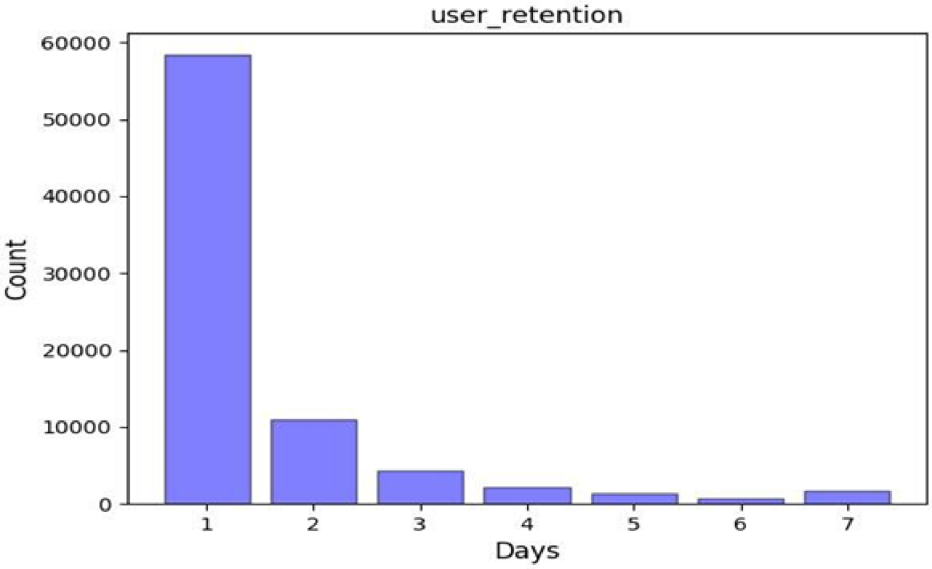


Figure 3.1.4

This graph above shows the retention of users on the platform. It shows that after the first occurrence of the user how many days within the seven-day period have they returned to the platform on.

The Figure 3.1.5 below shows the distribution of users spread out across the week.

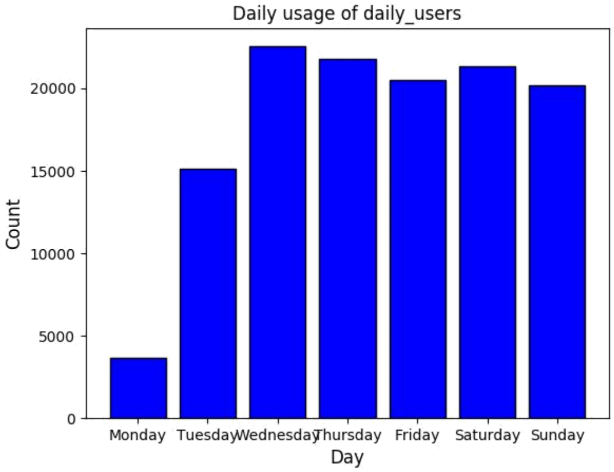
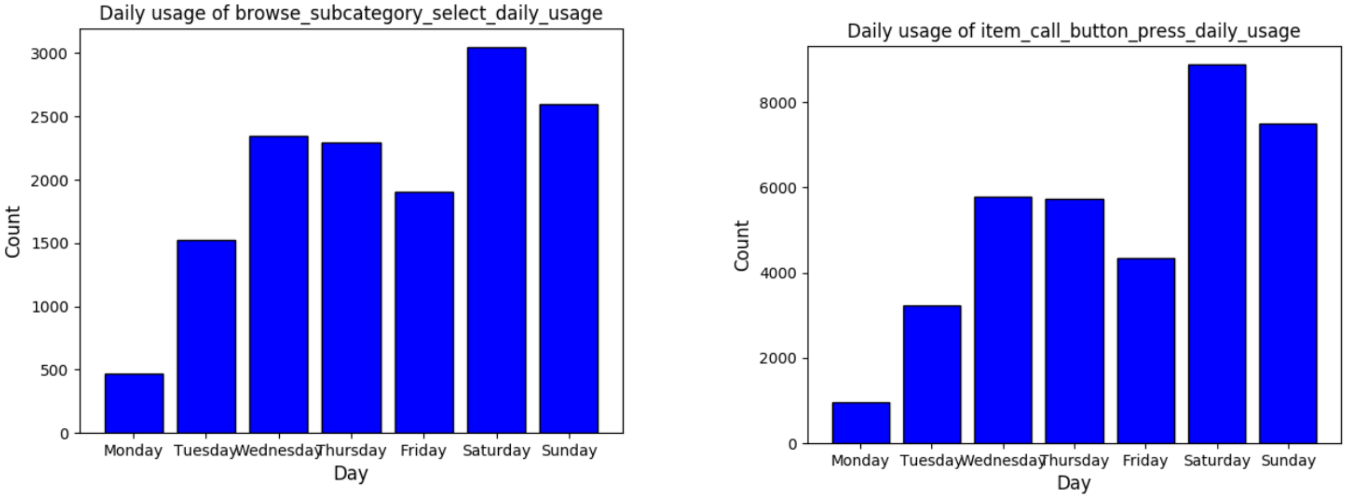
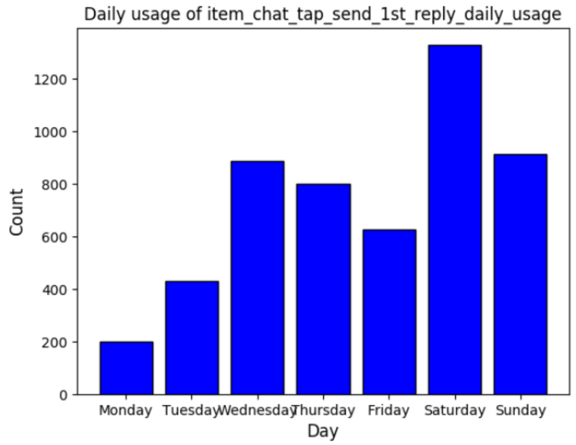
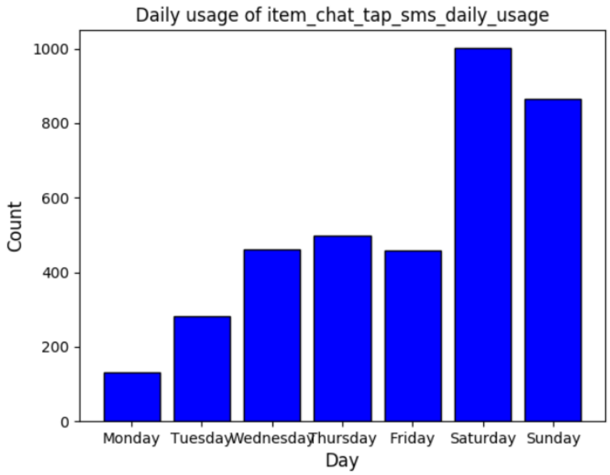


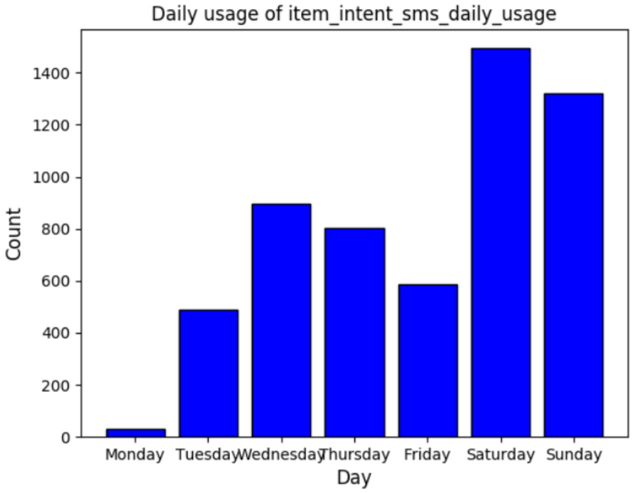
Figure 3.1.5

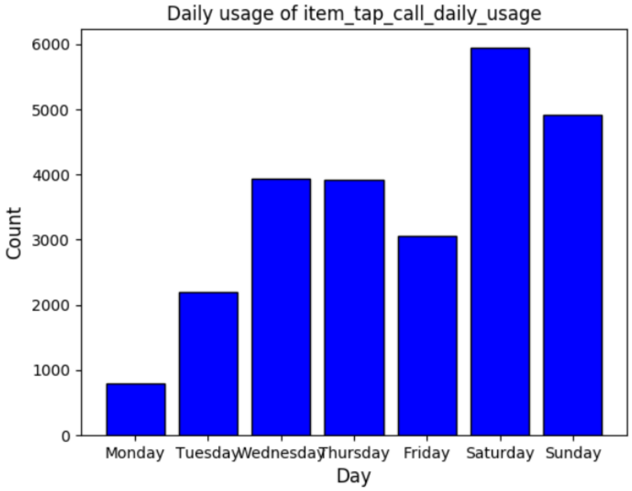
The graphs below, collectively termed as Figure 3.1.6, show the distribution of a single user's event that he performed on the OLX platform for that single week.

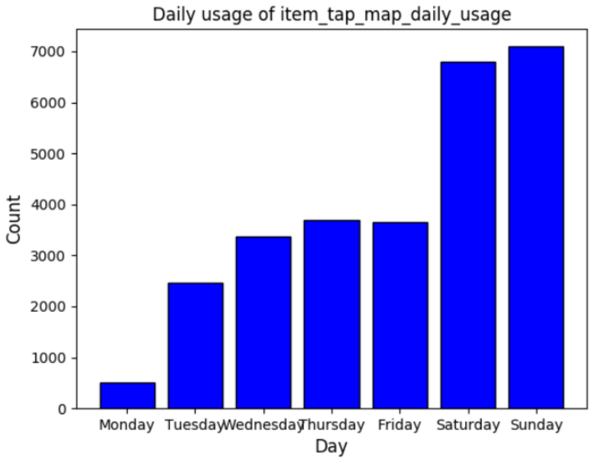


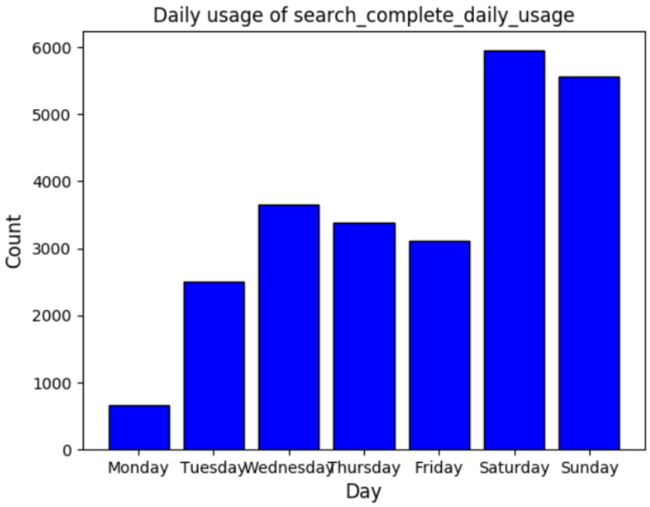

Daily usage of item_chat_tap_send_1st_reply_daily_usage
Count
0
200
400
600
800
1000
1200
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Day


Daily usage of item_chat_tap_sms_daily_usage
Count
0
200
400
600
800
1000
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Day


Daily usage of item_intent_sms_daily_usage
Count
0
200
400
600
800
1000
1200
1400
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Day


Daily usage of item_tap_call_daily_usage
Count
0
1000
2000
3000
4000
5000
6000
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Day


Daily usage of item_tap_map_daily_usage
Count
0
1000
2000
3000
4000
5000
6000
7000
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Day


Daily usage of search_complete_daily_usage
Count
0
1000
2000
3000
4000
5000
6000
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Day

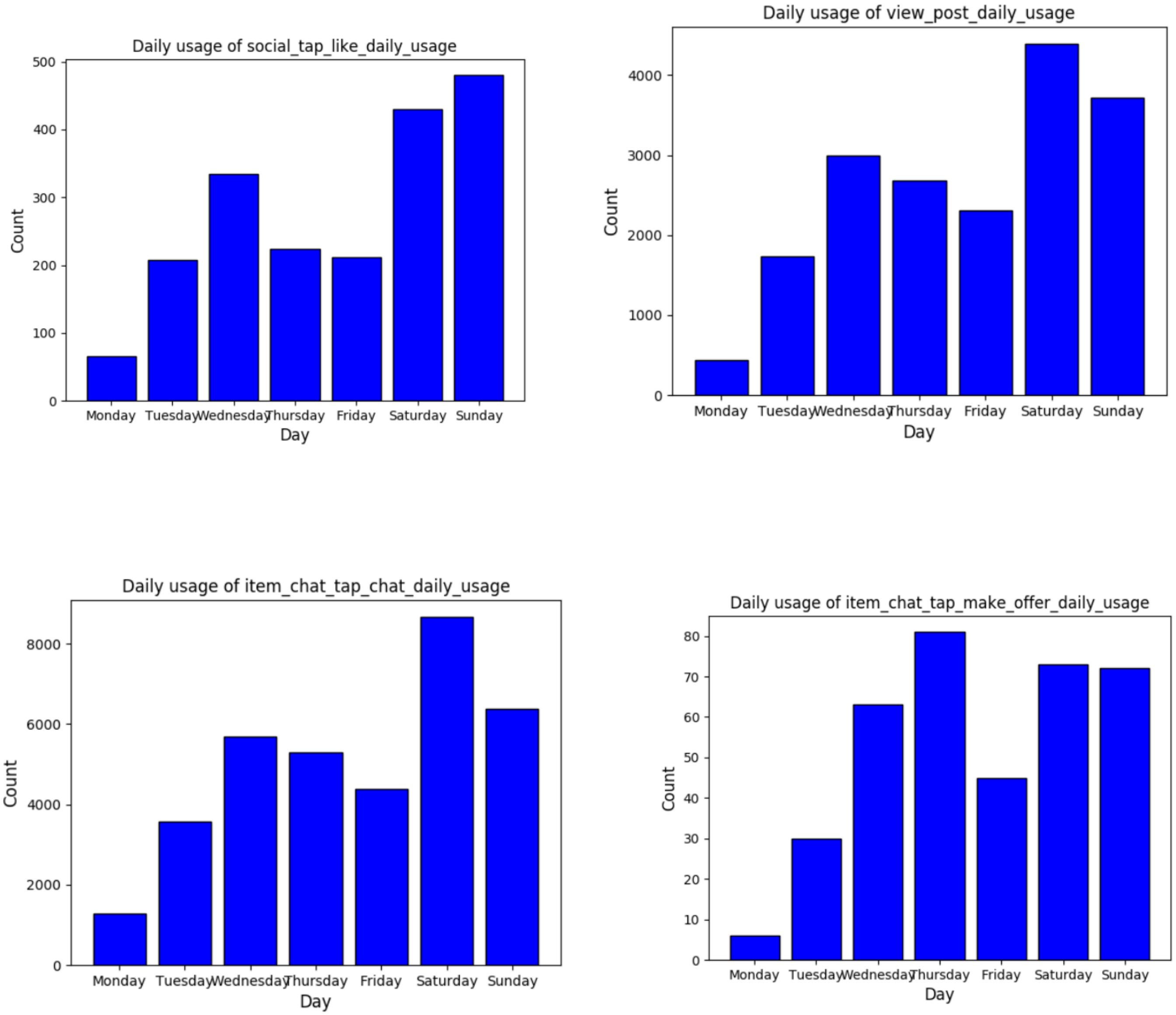


Figure 3.1.6

These events shown above such as **‘item_chat_tap_chat’**, ‘**item_tap_call’** and **‘item_call_button_press’** etc, can be later on seen are being used as events which we will classify as the purchase going through since there is no way to really tell if the purchase is made or not.

Most of these graphs have a similar pattern with Monday having the lowest count of events which can also be confirmed by the number of users that visited the site on Monday. There is an increase in the trigger of events across the week. This can be suggestive that most users are more active closer to the weekends. A concern that arises is that a month of data is not enough to extract patterns within the data.

The following two graphs shown in Figure 3.1.7 visualize the count of events triggered by over a little more than 70,000 users within one week

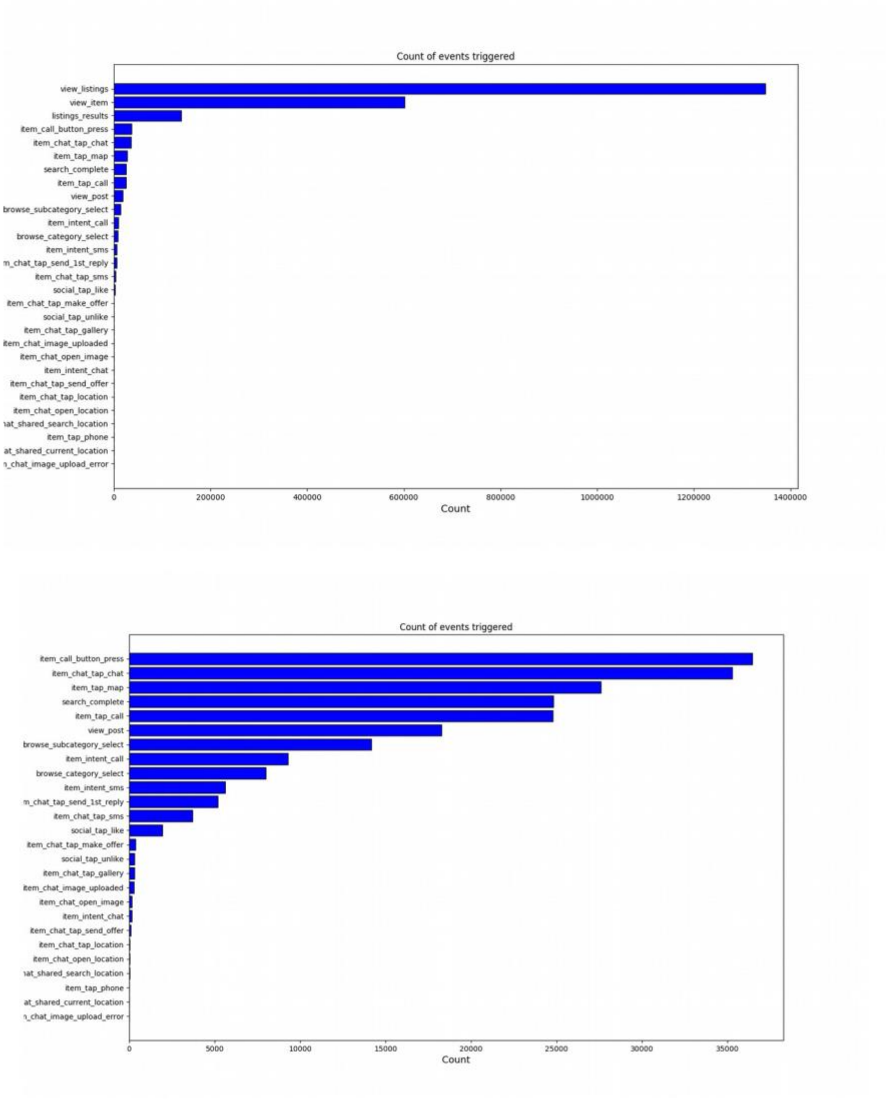


Figure 3.1.7

The 1st graph contains all possible events taken by a user. However, there is a big gap in the number of events, especially three events**, view_listings, view_item,** and **listings _result**. These events are those that are triggered for any user i.e. any time a user visits the site these events will be triggered. Thus, we removed these events to get a proper sense of the events that are not normally triggered by causal users which can be seen in the 2nd graph.

After doing the analysis on the users that came for a week, we increased the days from 7 to 10. And then we looked at the user retention, this was done on a different dataset than what was previously used. Here we used the **Web 2** dataset. The following Figure 3.1.8 shows the user retention for 10 days.

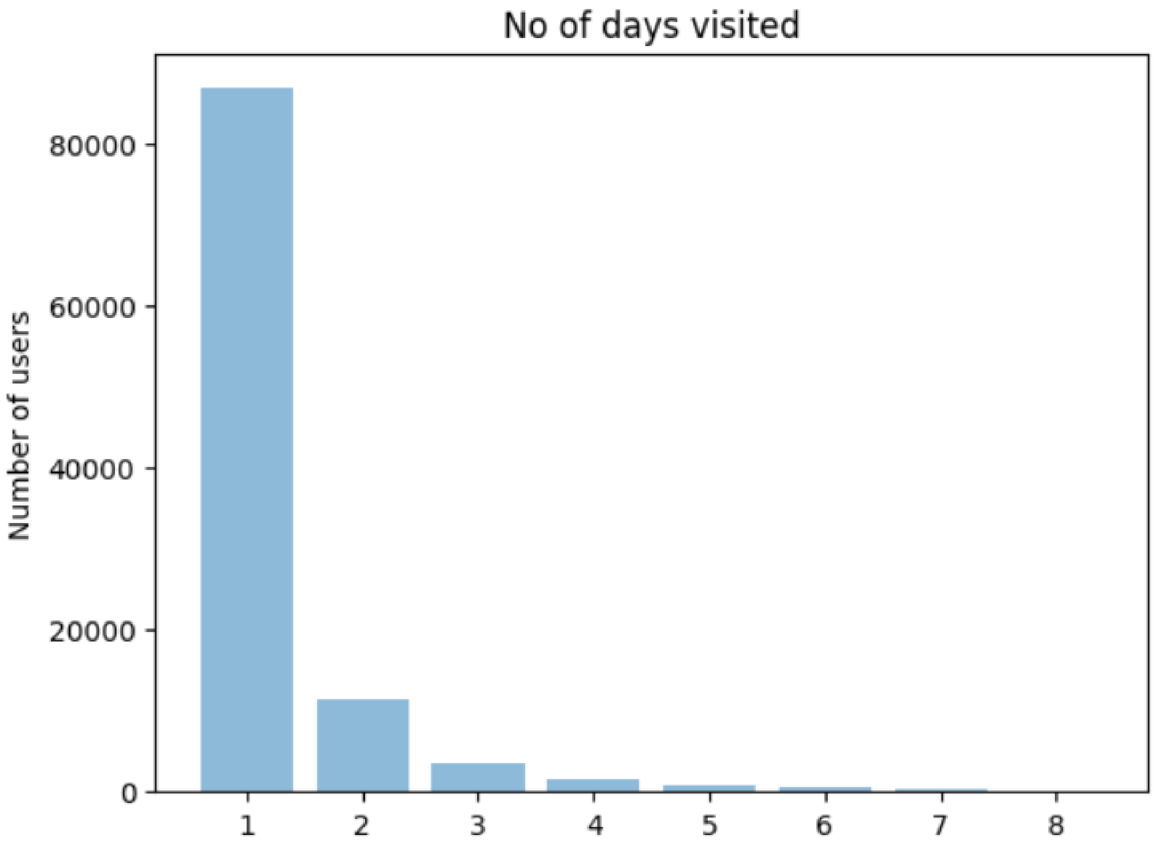


Figure 3.1.8

The diagram above shows all the users who came on the 1st Feb 2019, and then from that set of users we look which user of those users come on the second day, then third and so on. And as we can see only a handful of users came for 8 consecutive days. The count for users on successive days can be seen in the table.

Day Number	**User Count**
1	86746
2	11289
3	3575
4	1524
5	833
6	528
7	314
8	181

What we did next was for all those people who came for 8 consecutive days, we viewed the events they triggered which can be seen below in Figure 3.1.9

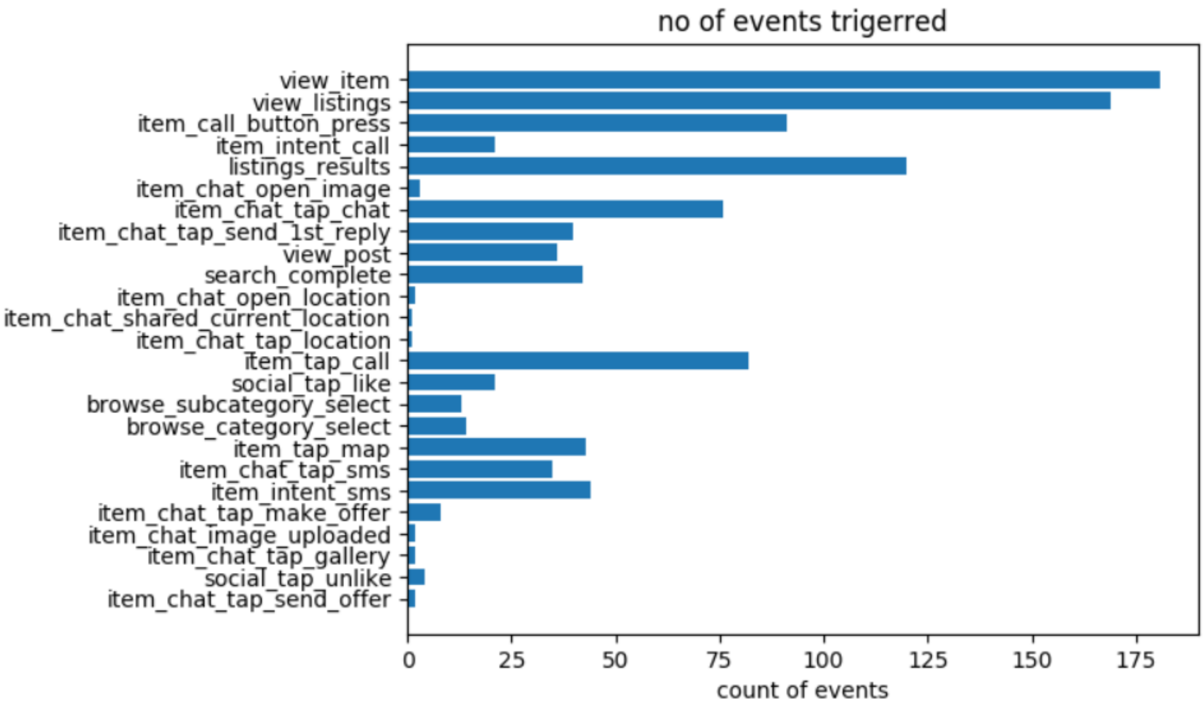
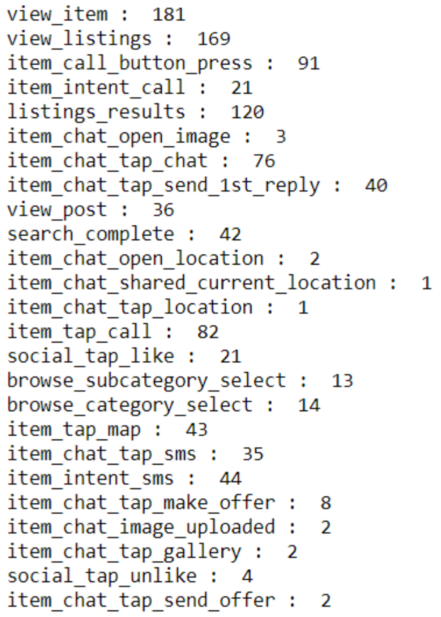


Figure 3.1.9



```
view_item :  181
view_listings :  169
item_call_button_press :  91
item_intent_call :  21
listings_results :  120
item_chat_open_image :  3
item_chat_tap_chat :  76
item_chat_tap_send_1st_reply :  40
view_post :  36
search_complete :  42
item_chat_open_location :  2
item_chat_shared_current_location :  1
item_chat_tap_location :  1
item_tap_call :  82
social_tap_like :  21
browse_subcategory_select :  13
browse_category_select :  14
item_tap_map :  43
item_chat_tap_sms :  35
item_intent_sms :  44
item_chat_tap_make_offer :  8
item_chat_image_uploaded :  2
item_chat_tap_gallery :  2
social_tap_unlike :  4
item_chat_tap_send_offer :  2
```

From the diagram above and the values in the table we can see that most users who go on the platform to view an item for consecutive days are very likely to fire one of the events which we assumed as to be a purchase event such as **‘item_intent_call’**, **‘item_call_button_press’** etc.

After this we will work on our final dataset the whole **Web** dataset. The Figure 3.1.10 shows a pie chart with users who have more than 10 sessions and those with fewer than 10 sessions.

Number of Sessions By Each User

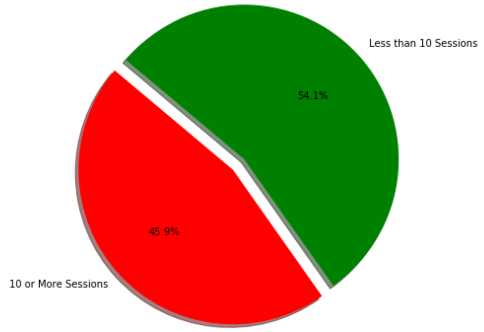


Figure 3.1.10

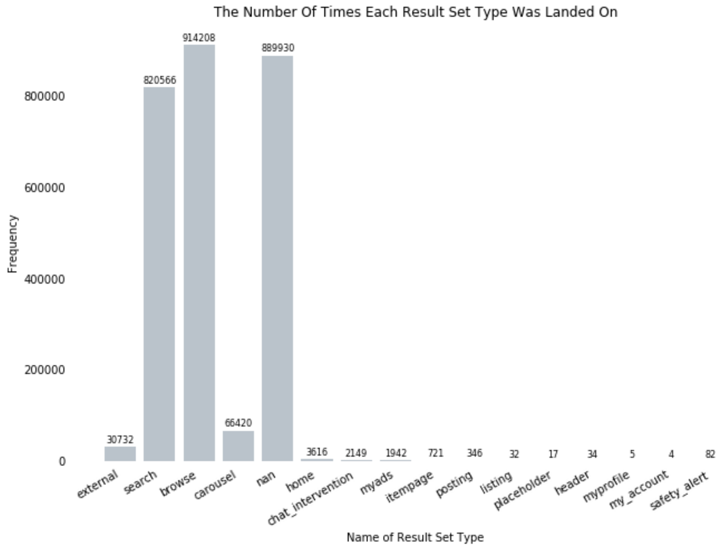


Figure 3.1.11

In Figure 3.1.11 we can see that frequency for the number of times each result set was landed on, here we can see that the search types are predominantly **search** and **browse.**

Figure 3.1.12 shows the number of times each invite source was used to and on the OLX platform, here again for the **Web** data we can see that the invite sources are predominantly **google-search**, **Facebook** and **OLX.pk** with the highest count.

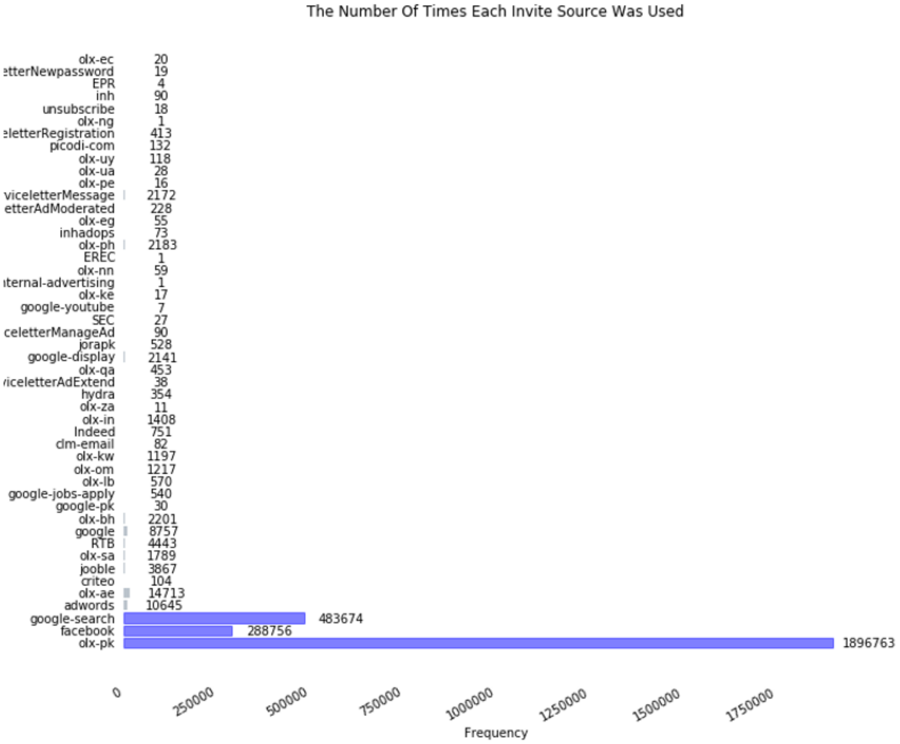


Figure 3.1.12

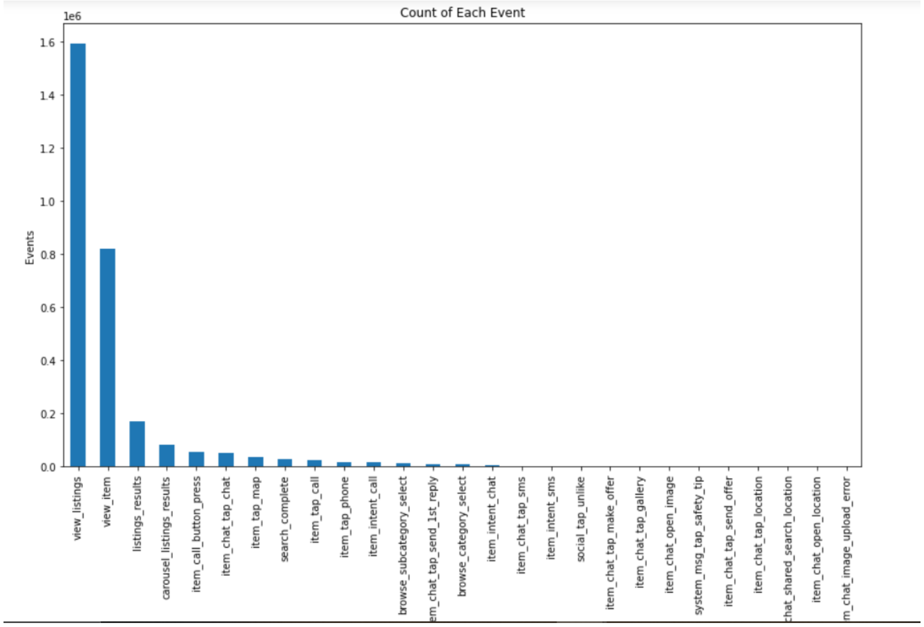


Figure 3.1.13

In Figure 3.1.13 we can see the number of events generated in total by all users for the entire time duration from 1st Feb to July 31st. We can see here that most events are **view_findings, view_item** and **listing_results.**

Figure 3.1.14 shows the zoomed in version of the count of events taken by the users excluding the 3 events which had the highest counts.

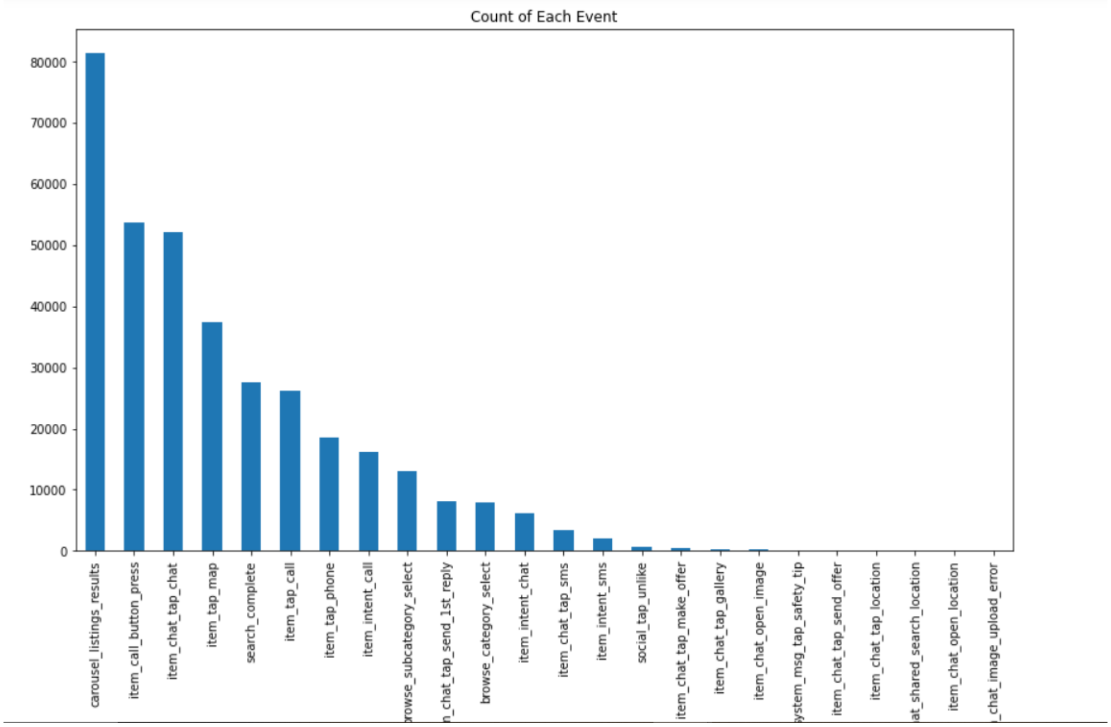
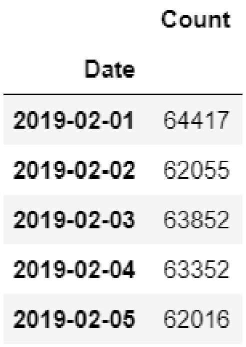


Figure 3.1.14

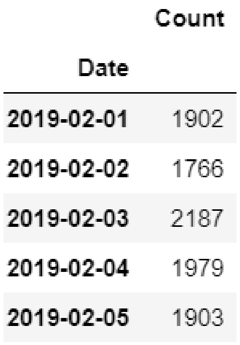
Next we look at the two graphs, the first (Figure 3.1.15) shows the counts of events triggered for each day, whereas the second (Figure 3.1.16) shows only the count of the **purchase events** triggered for each day. The table for the count of events and purchased events are shown below for each month.

Events



	Count
Date	
2019-02-01	64417
2019-02-02	62055
2019-02-03	63852
2019-02-04	63352
2019-02-05	62016

Purchase Events



	Count
Date	
2019-02-01	1902
2019-02-02	1766
2019-02-03	2187
2019-02-04	1979
2019-02-05	1903

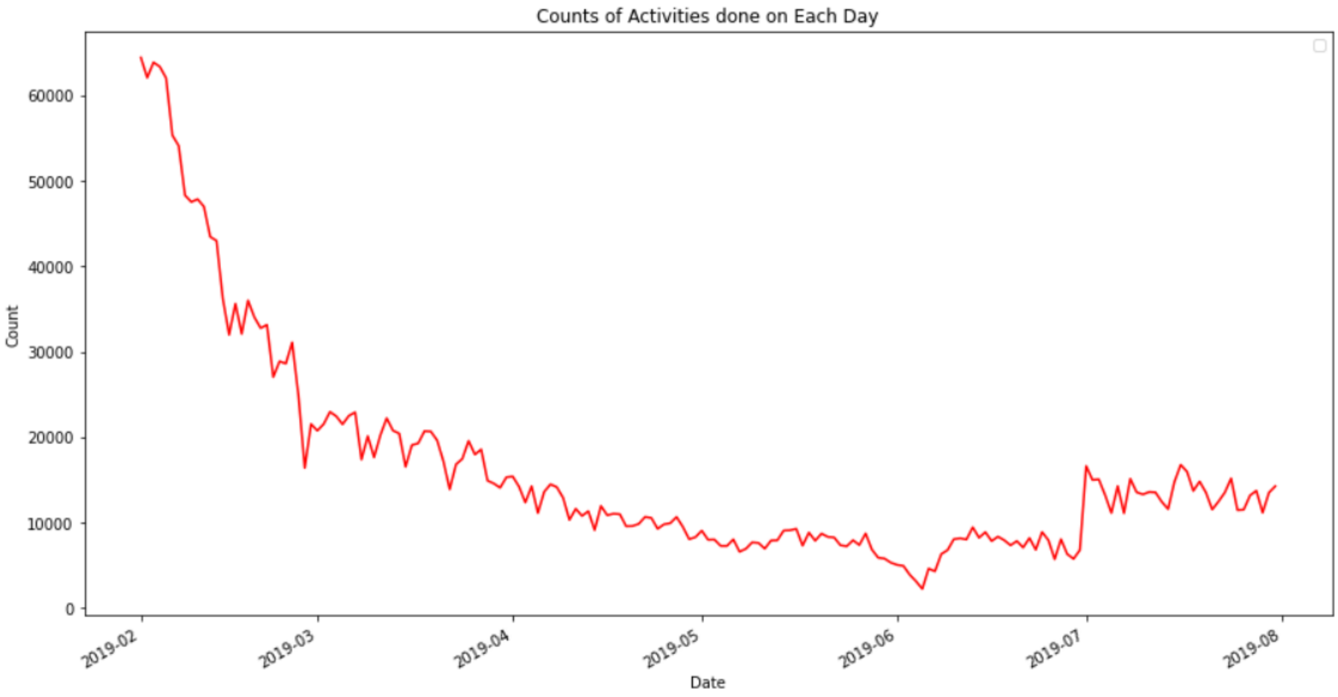


Figure 3.1.15

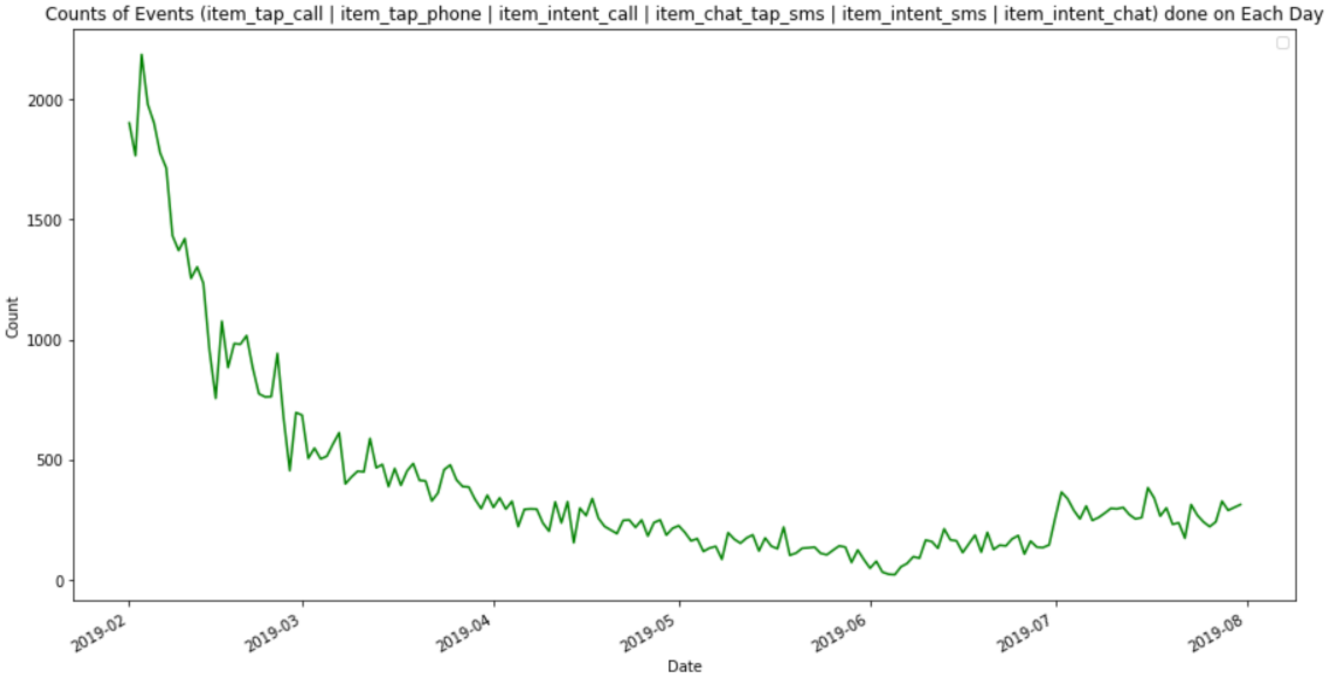


Figure 3.1.16

Figure 3.1.17 and Figure 3.1.18 show the weekly distribution of events and purchased events below.

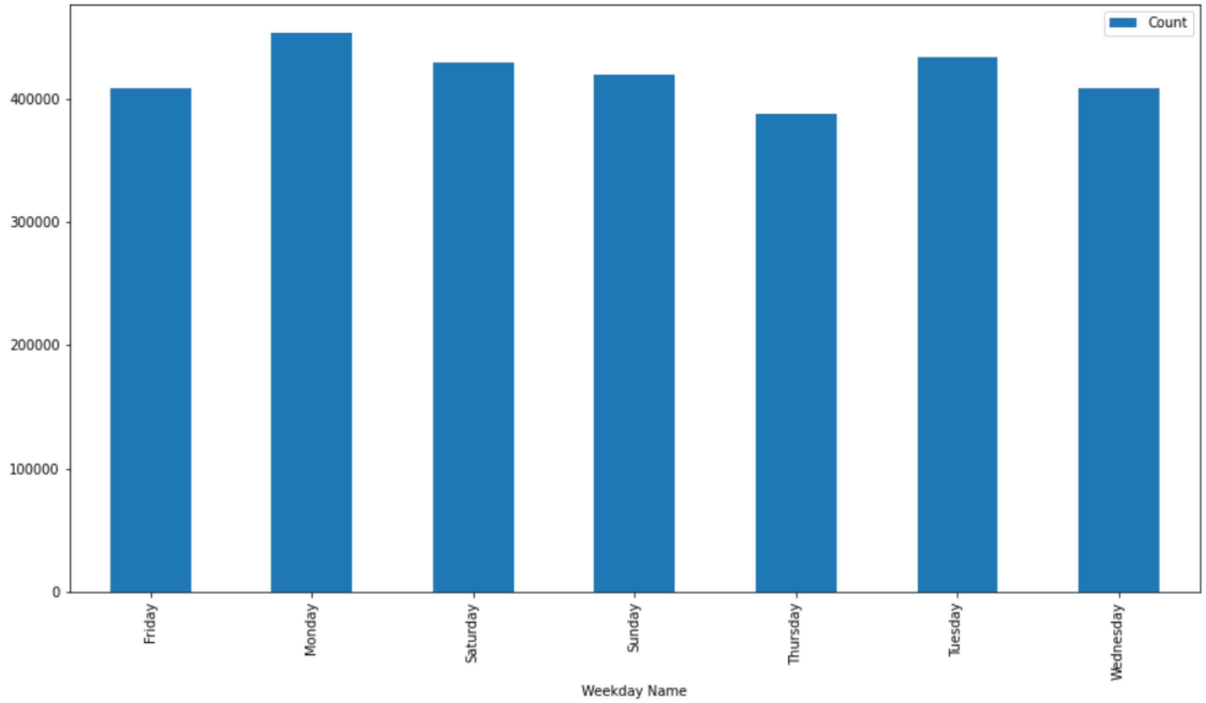


Figure 3.1.17

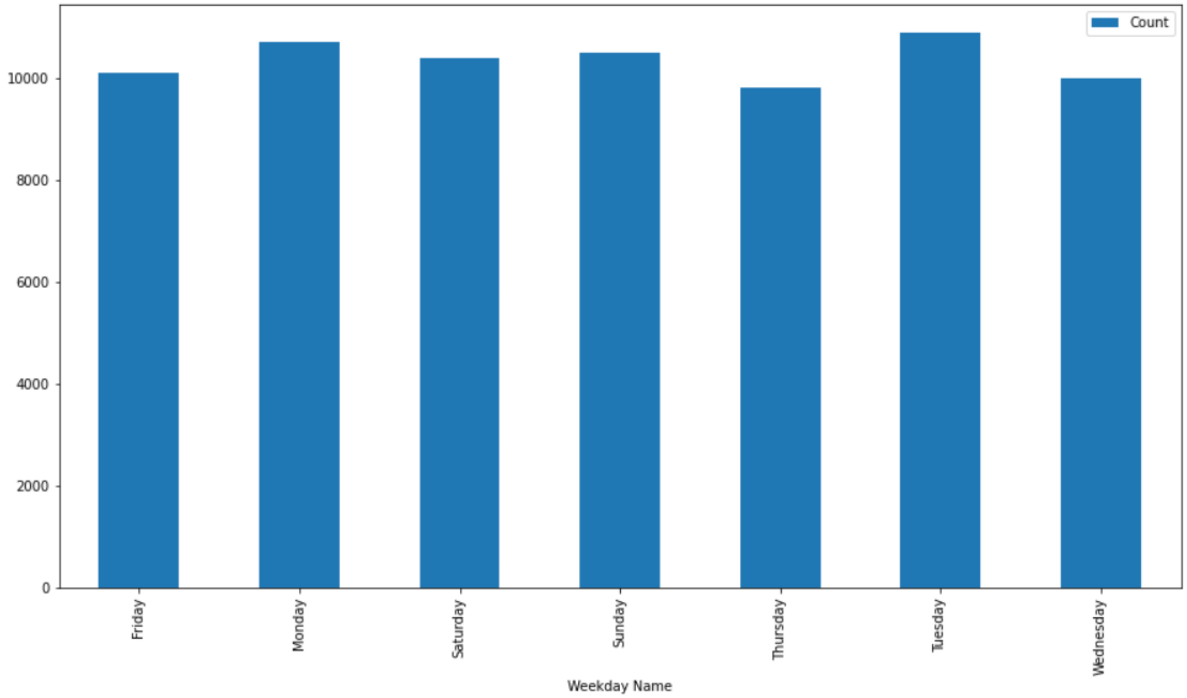


Figure 3.1.18

3.2. Buyer Conversion Metric

The end goal of this was to define a metric to determine the **Daily Unique Buyers (DUB).** To identify these buyers, we first had to define who are **Daily Active Users (DAU)** are. We define **DAU** as those who performed any action/events from a set list of actions. We term these events as **Reply Actions** or what we previously have been calling **Purchase Events.** However, these Reply Action are more than what we defined as Purchase Events. Below we have shown the list of what we consider to be Reply Actions for the Users. We then needed to through some sort of conversion figure out how we can define our DUB.

DAU * *CONVERSION* = DUB

Reply Actions =

```
['item_tap_image_pagination', 'view_item', 'item_tap_image',
        'item_chat_tap_chat', 'item_chat_tap_send_1st_reply',
        'item_tap_map', 'listing_tap_apply_sorting',
        'listing_tap_apply_visualization', 'item_chat_tap_sms',
        'item_tap_call', 'item_call_button_press',
        'posting_tap_next_step_incomplete', 'item_chat_open_location',
        'item_intent_call', 'item_tap_delete', 'item_chat_tap_make_offer',
        'item_chat_tap_copy', 'item_chat_tap_location', 'item_tap_phone',
        'social_tap_unlike', 'item_chat_tap_send_offer',
        'item_chat_shared_search_location', 'item_chat_tap_gallery',
        'system_msg_tap_safety_tip', 'item_chat_tap_camera',
        'item_intent_sms', 'item_chat_open_image', 'item_intent_chat',
        'listing_tap_select_visualization', 'help_tap_helpcenter',
        'item_chat_tap_get_direction', 'item_chat_tap_scroll',
        'trust_tap_block_user', 'item_chat_image_tap_download',
        'tap_browse_ads_near_me', 'map_location_zero_search',
        'map_location_input']
```

We did these conversions on the basis of 3 broad categories things which we considered were essential factors which lead a Daily Active User to becoming a Daily User Buyer.

- Time
- Categories
- Price

3.2.1 Conversion Across Time

We looked at all the Unique Users that have come on the platform from the 1st of Feb to 31st July 2019. Figure 3.2.1.1 Below shows Count of Users who have more than a total 20+ Actions (Any) on the **Web** platform in 6 months.

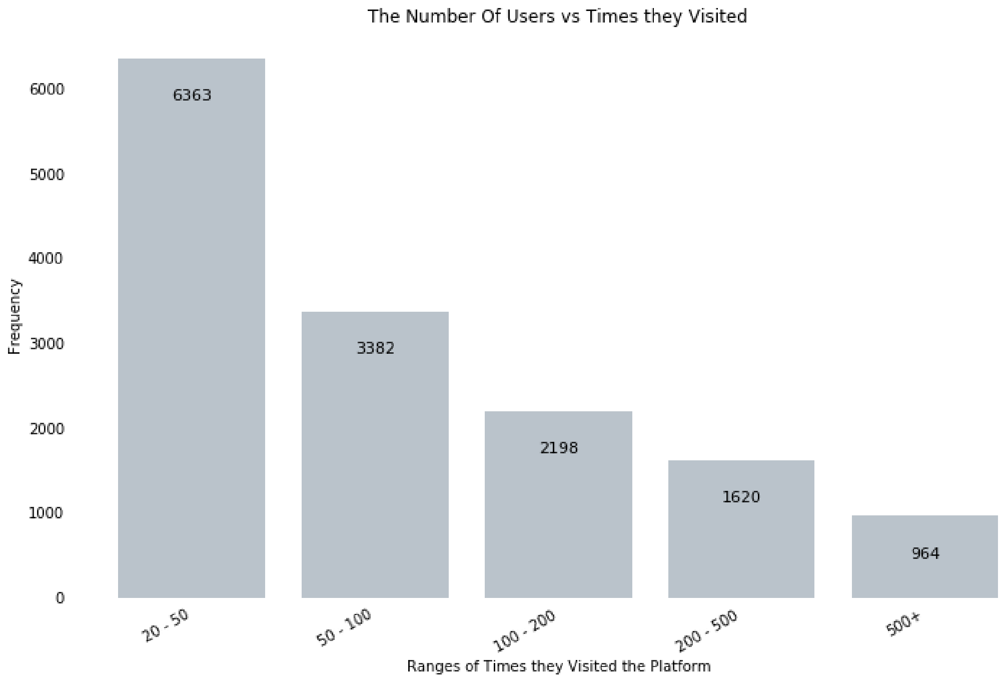


Figure 3.2.1.1

Our **conversion formula** with respect to **time** was that we looked at all Users who did **Reply Actions** and we collected the Dates on which those Users made those Reply Actions. We then looked if that User committed a Reply Action in the 1st 10 days, 2nd 10 days or 3rd 10 days of a month and we did that for all 6 months (Once or many times) and mark that period of 10 days as True or (1). From the 10-day period where he had his first True, we started observing our User and we saw all the next subsequent 10-day periods where he/she made a **Reply Action** (marked True for that 10-day period) or not till the end of July. We look at all the True (1) and all the False (0) after the **first** True and see if a user makes Reply Actions more than 50% of the time, we would classify him as a DUB. Look at the example below:

1st 10 Days, 2nd 10 Days, 3rd 10 days

February	[0, 1, 0],
March	[0, 0, 0],
April	[0, 0, 1],
May	[0, 0, 0],
June	[0, 0, 0],
July	[0, 0, 0]

Total 10 days Periods = 3*6 = 18

The 1 in February's 2nd 10 days is the First True so we will look at the Users data after that 10 Day period. So, he has 2 periods in total where he makes a Reply Action and we look at the user until the

3rd period of July so that’s a total of 17 possible periods where he could make a Reply Action. So, for this user we have 2/17 * 100 = 11.7% which is less than 50% so we won’t classify him as a **DUB**. 1st 10 Days, 2nd 10 Days, 3rd 10 days

February	[0, 0, 0],
March	[0, 0, 1],
April	[1, 0, 1],
May	[0, 1, 1],
June	[0, 1, 0],
July	[1, 1, 0]

For this User = 8/13* 100 = 61.5%. This user considered a DUB

Figure 3.2.1.2 shows a bar chart of count of Users and Actions with different constraints:

A: Count of All the Actions/Events done by All the Users

B: Count of All Reply Actions done by All User

C: Count of All Unique Users

D: Count of All Unique Users which have more than 20+ Actions (Any) in 6 months

E: Count of All Unique Users which have more than 20+ Reply Actions in 6 months

F: Count of Daily User Buyers via our Conversion

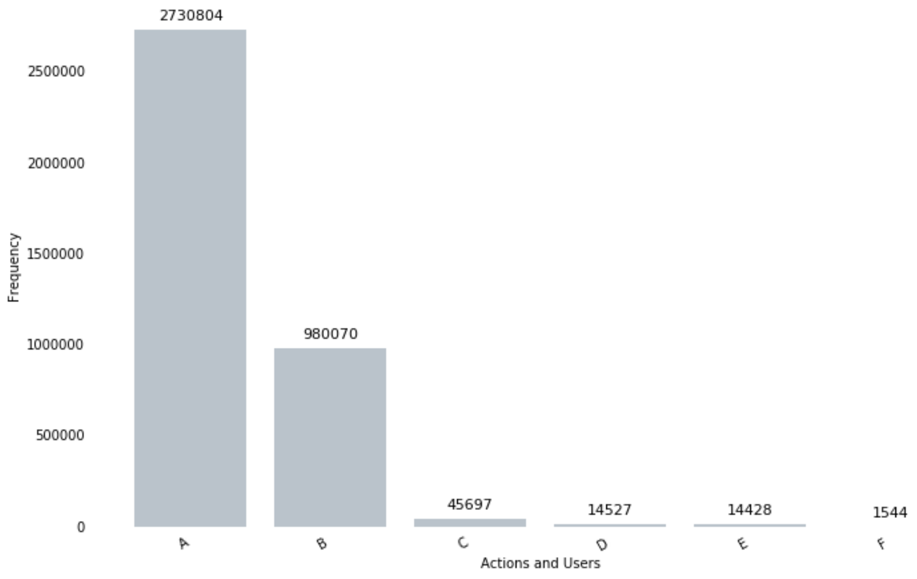


Figure 3.2.1.2

From the bar chart we can see that out of a total of 45697 unique user's only 1544 are classified as Daily Unique Buyers according to our conversion metric.

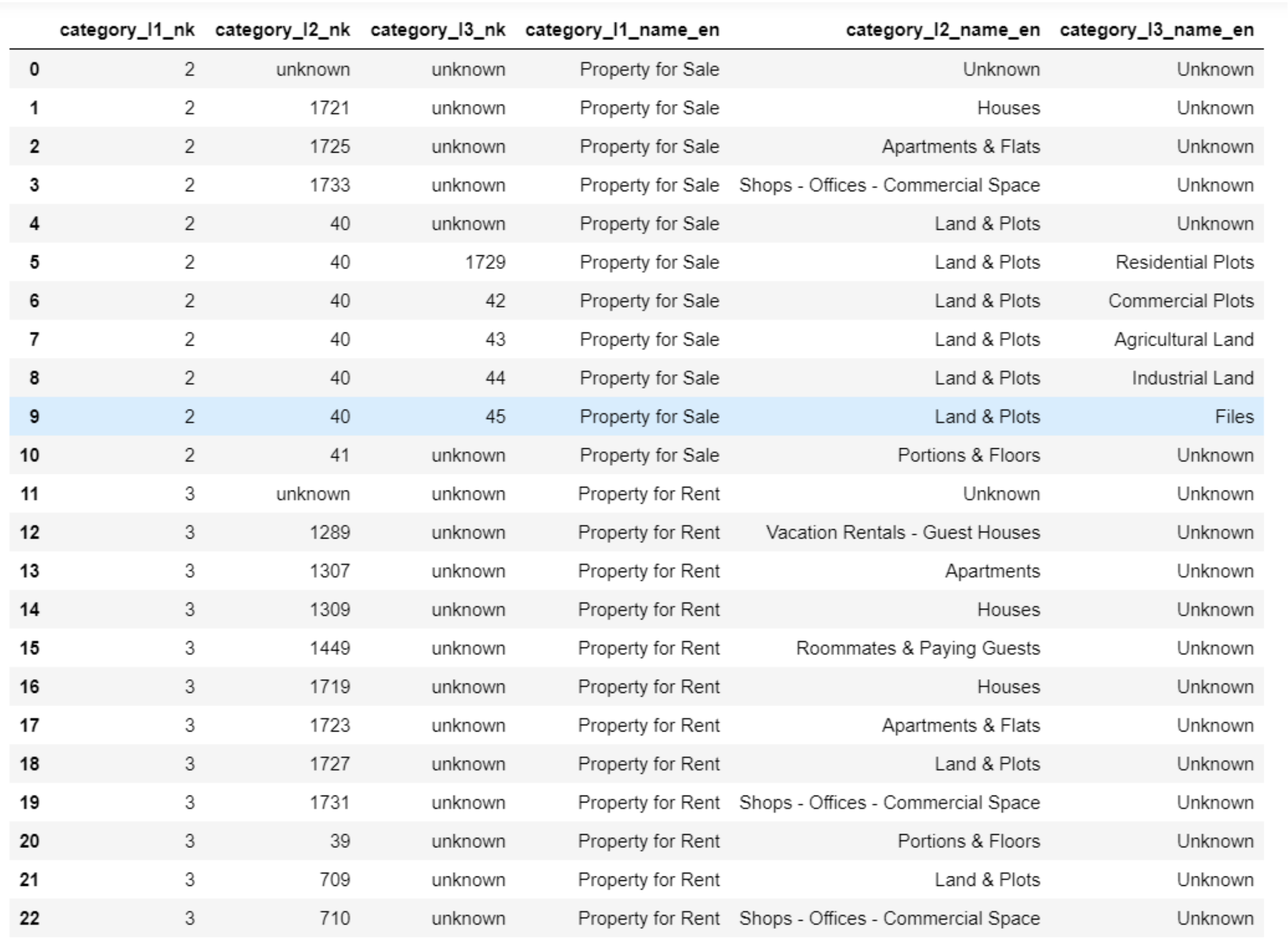
Now we will look at the other 2 categories. However, no proper metric was designed for them to convert DAU to DUB.

3.2.2 Conversion Across Categories

Here we will look at different categories, there are 3 hierarchal categories.

1. Category_l1
2. Category_l2
3. Category_l3 (Not used)

The table below shows which code number matches with which category, every category_l?_nk matches to its corresponding every category_l?_name_en.



	category_l1_nk	category_l2_nk	category_l3_nk	category_l1_name_en	category_l2_name_en	category_l3_name_en
0	2	unknown	unknown	Property for Sale	Unknown	Unknown
1	2	1721	unknown	Property for Sale	Houses	Unknown
2	2	1725	unknown	Property for Sale	Apartments & Flats	Unknown
3	2	1733	unknown	Property for Sale	Shops - Offices - Commercial Space	Unknown
4	2	40	unknown	Property for Sale	Land & Plots	Unknown
5	2	40	1729	Property for Sale	Land & Plots	Residential Plots
6	2	40	42	Property for Sale	Land & Plots	Commercial Plots
7	2	40	43	Property for Sale	Land & Plots	Agricultural Land
8	2	40	44	Property for Sale	Land & Plots	Industrial Land
9	2	40	45	Property for Sale	Land & Plots	Files
10	2	41	unknown	Property for Sale	Portions & Floors	Unknown
11	3	unknown	unknown	Property for Rent	Unknown	Unknown
12	3	1289	unknown	Property for Rent	Vacation Rentals - Guest Houses	Unknown
13	3	1307	unknown	Property for Rent	Apartments	Unknown
14	3	1309	unknown	Property for Rent	Houses	Unknown
15	3	1449	unknown	Property for Rent	Roommates & Paying Guests	Unknown
16	3	1719	unknown	Property for Rent	Houses	Unknown
17	3	1723	unknown	Property for Rent	Apartments & Flats	Unknown
18	3	1727	unknown	Property for Rent	Land & Plots	Unknown
19	3	1731	unknown	Property for Rent	Shops - Offices - Commercial Space	Unknown
20	3	39	unknown	Property for Rent	Portions & Floors	Unknown
21	3	709	unknown	Property for Rent	Land & Plots	Unknown
22	3	710	unknown	Property for Rent	Shops - Offices - Commercial Space	Unknown

So, for example here we can see that 2 refers to Property for Sale and 3 refers to Property of Rent for Category_l1.

The Figure 3.2.2.1 below shows the value counts for the different items for category_l1

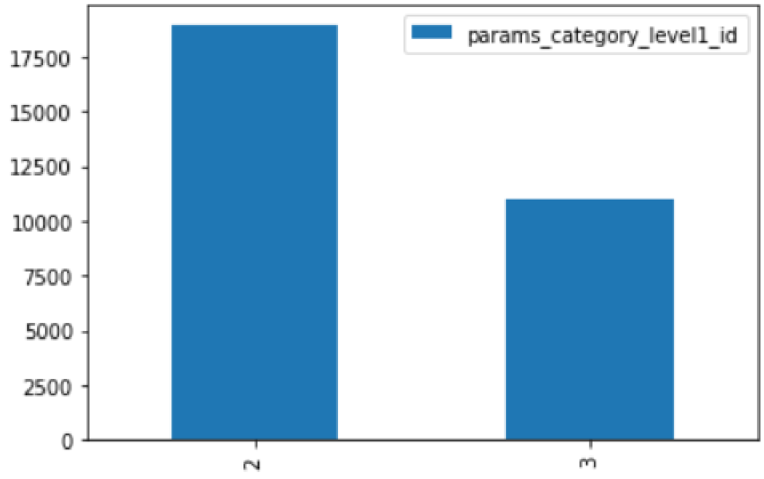


Figure 3.2.2.1

Figure 3.2.2.2 shows the different percentages and frequencies for Active Users (Those who used the reply actions) and Non-Active Users for both types of Category_l1(Property of Sale and Property of Rent)

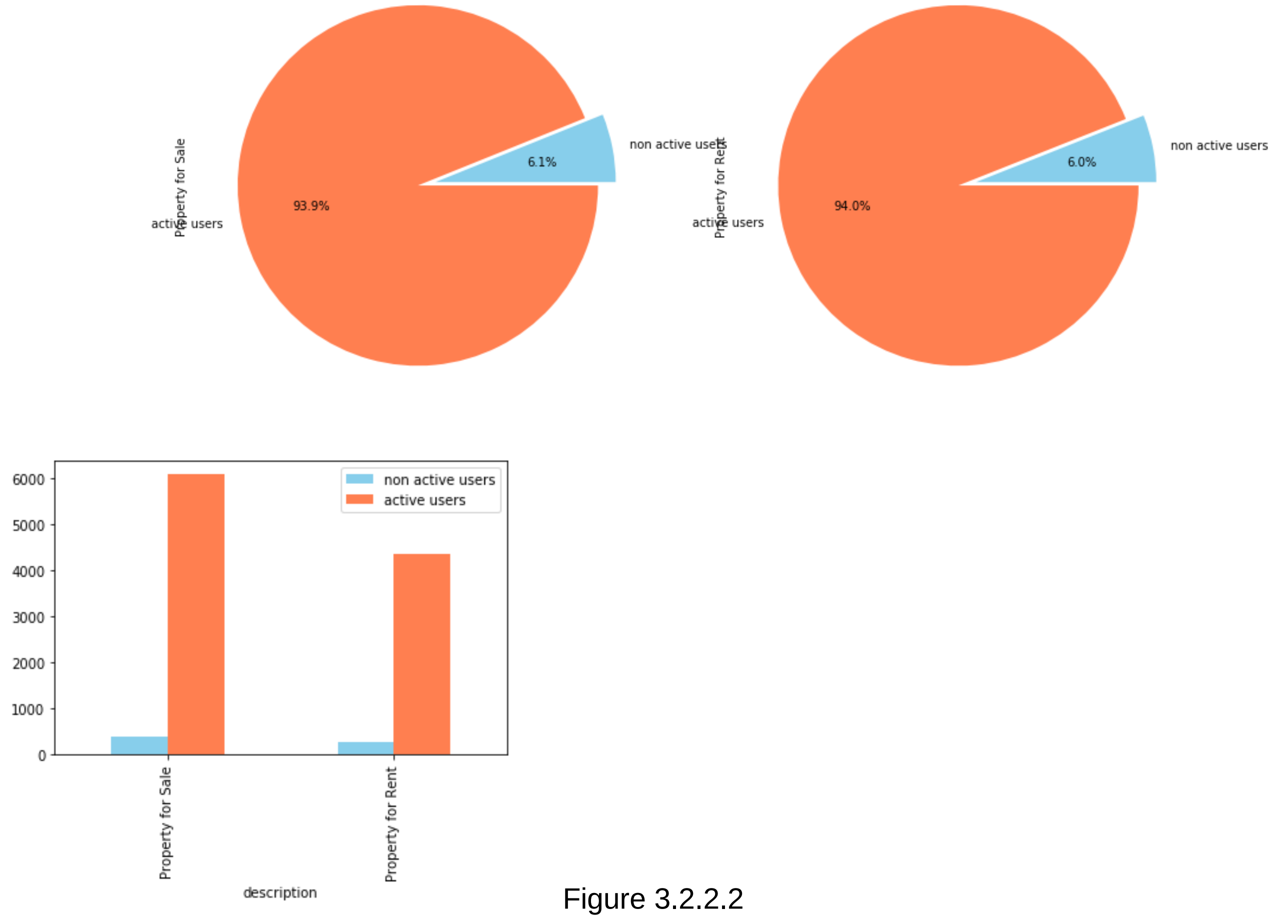


Figure 3.2.2.2

As for Category 2, go one hierarchy deeper for both broad categories of l1. Here they type of items are also such that for category_l1 there were only two types. Here for category_l2 there are 15 types. Figure 3.2.2.3 shows which category_l1 they are in and which category_l2 they are in and shows the frequency between Active and Non-Active Users for each combination. Each code can be matched by the table in the start.

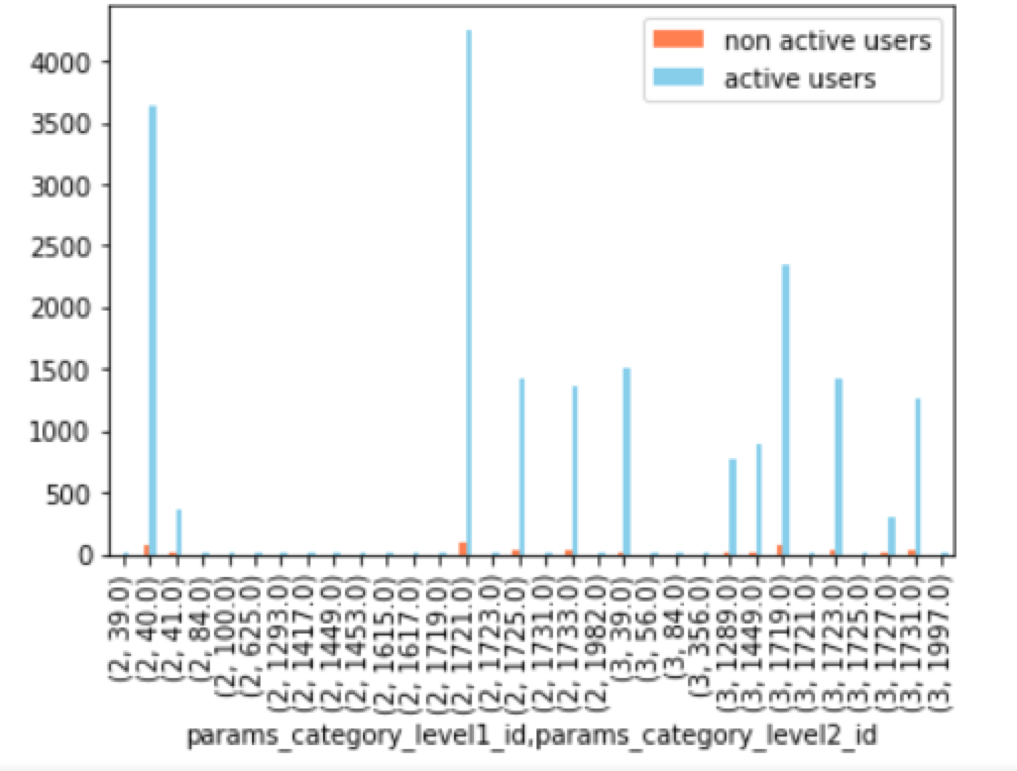


Figure 3.2.2.3

3.2.3 Conversion Across Prices

Here too it can e seen that no proper metric was properly formed to convert DAU to DUB. Different prices for each item were given which allowed us to only view what price ranges are more frequently used over others. Figure 3.2.3.4 shows the frequencies of different price ranges that were selected by Active Users.

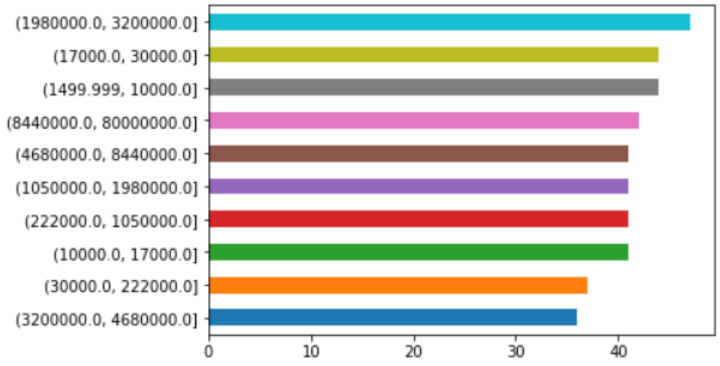


Figure 3.2.3.4

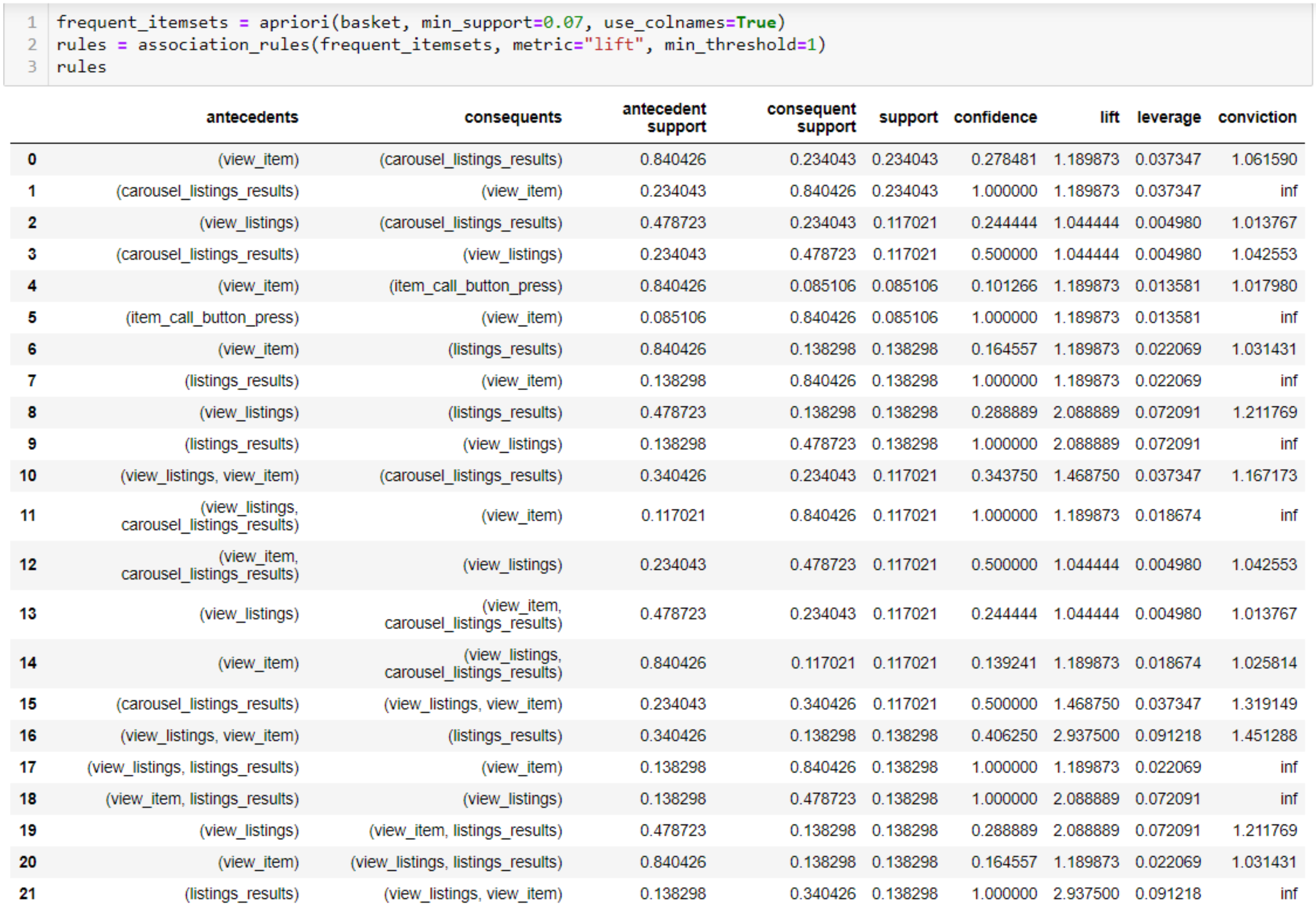
3.3. Frequent Pattern Mining

Here we will use the Apriori Algorithm. Apriori is an algorithm for frequent item set mining and association rule learning over datasets. It proceeds by identifying the frequent individual items in the dataset and extending them to larger and larger item sets if those item sets appear sufficiently often in the dataset. Here we will see what frequent events that are associated with each other. To understand the results of our pattern mining analysis we must look at a few metrics which can help us understand our results.

- **Support(A)** = (Transactions containing (A))/(Total Transactions)
- **Confidence(A→B)** = (Transactions containing both (A and B))/(Transactions containing A)
- **Lift (A -> B)** refers to the increase in the ratio of sale of B when A is sold.
  Lift (A –> B) can be calculated by dividing Confidence (A -> B) divided by Support(B).

- IF Lift (B, C) = 1, B and C are independent
- IF Lift > 1 Positively Correlated
- IF Lift < 1 Negatively Correlated, which means Lift(B, ~C) are positively correlated

Since we are talking about pattern mining in terms of events generated together in a session then if Lift (A->B) = 3.33 then it is basically telling us that the likelihood of Event A and Event B together is 3.33 times more than the likelihood of just Event B. The Figure 3.3.1 shows the results of our Apriori Algorithm on events taken by a user in a single session.



```
frequent_itemsets = apriori(basket, min_support=0.07, use_colnames=True)
rules = association_rules(frequent_itemsets, metric="lift", min_threshold=1)
rules
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(view_item)	(carousel_listings_results)	0.840426	0.234043	0.234043	0.278481	1.189873	0.037347	1.061590
1	(carousel_listings_results)	(view_item)	0.234043	0.840426	0.234043	1.000000	1.189873	0.037347	inf
2	(view_listings)	(carousel_listings_results)	0.478723	0.234043	0.117021	0.244444	1.044444	0.004980	1.013767
3	(carousel_listings_results)	(view_listings)	0.234043	0.478723	0.117021	0.500000	1.044444	0.004980	1.042553
4	(view_item)	(item_call_button_press)	0.840426	0.085106	0.085106	0.101266	1.189873	0.013581	1.017980
5	(item_call_button_press)	(view_item)	0.085106	0.840426	0.085106	1.000000	1.189873	0.013581	inf
6	(view_item)	(listings_results)	0.840426	0.138298	0.138298	0.164557	1.189873	0.022069	1.031431
7	(listings_results)	(view_item)	0.138298	0.840426	0.138298	1.000000	1.189873	0.022069	inf
8	(view_listings)	(listings_results)	0.478723	0.138298	0.138298	0.288889	2.088889	0.072091	1.211769
9	(listings_results)	(view_listings)	0.138298	0.478723	0.138298	1.000000	2.088889	0.072091	inf
10	(view_listings, view_item)	(carousel_listings_results)	0.340426	0.234043	0.117021	0.343750	1.468750	0.037347	1.167173
11	(view_listings, carousel_listings_results)	(view_item)	0.117021	0.840426	0.117021	1.000000	1.189873	0.018674	inf
12	(view_item, carousel_listings_results)	(view_listings)	0.234043	0.478723	0.117021	0.500000	1.044444	0.004980	1.042553
13	(view_listings)	(view_item, carousel_listings_results)	0.478723	0.234043	0.117021	0.244444	1.044444	0.004980	1.013767
14	(view_item)	(view_listings, carousel_listings_results)	0.840426	0.117021	0.117021	0.139241	1.189873	0.018674	1.025814
15	(carousel_listings_results)	(view_listings, view_item)	0.234043	0.340426	0.117021	0.500000	1.468750	0.037347	1.319149
16	(view_listings, view_item)	(listings_results)	0.340426	0.138298	0.138298	0.406250	2.937500	0.091218	1.451288
17	(view_listings, listings_results)	(view_item)	0.138298	0.840426	0.138298	1.000000	1.189873	0.022069	inf
18	(view_item, listings_results)	(view_listings)	0.138298	0.478723	0.138298	1.000000	2.088889	0.072091	inf
19	(view_listings)	(view_item, listings_results)	0.478723	0.138298	0.138298	0.288889	2.088889	0.072091	1.211769
20	(view_item)	(view_listings, listings_results)	0.840426	0.138298	0.138298	0.164557	1.189873	0.022069	1.031431
21	(listings_results)	(view_listings, view_item)	0.138298	0.340426	0.138298	1.000000	2.937500	0.091218	inf

3.4. Cluster Analysis

Finally we do cluster analysis to see whether we can find any meaningful structure within the data that can tell us about similar behavioral patterns between customers, this is an attempt to find out whether certain sequence of certain events are more likely to purchase events which could help us tell whether these are serious buyer or window-shoppers.

Here we use 3 different types of clustering techniques:

1. TSNE-Clustering on Numerical Attributes
2. K-Means Clustering on one-hot encoded sessions
3. K-Means Clustering on sequential events using NLP

3.4.1 TSNE-Clustering

In this we first process our feature set on which we want to perform the clustering. The feature set that was created for each user contained: number of events for the user, number of distinct sessions (on the same device or on different devices), number of times a particular search filter was used (there were 11 different filters available), and the number of times a particular event was logged for the user. The algorithms that were used were t-distributed Stochastic Neighbor Embedding (TSNE) clustering. Figure 3.4.1 shows the clusters formed by this technique.

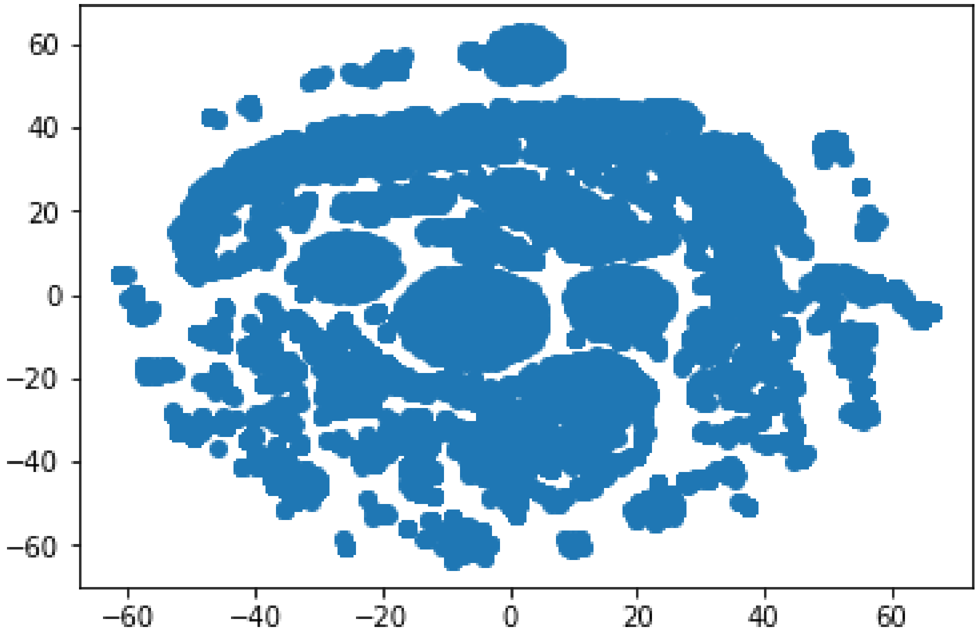


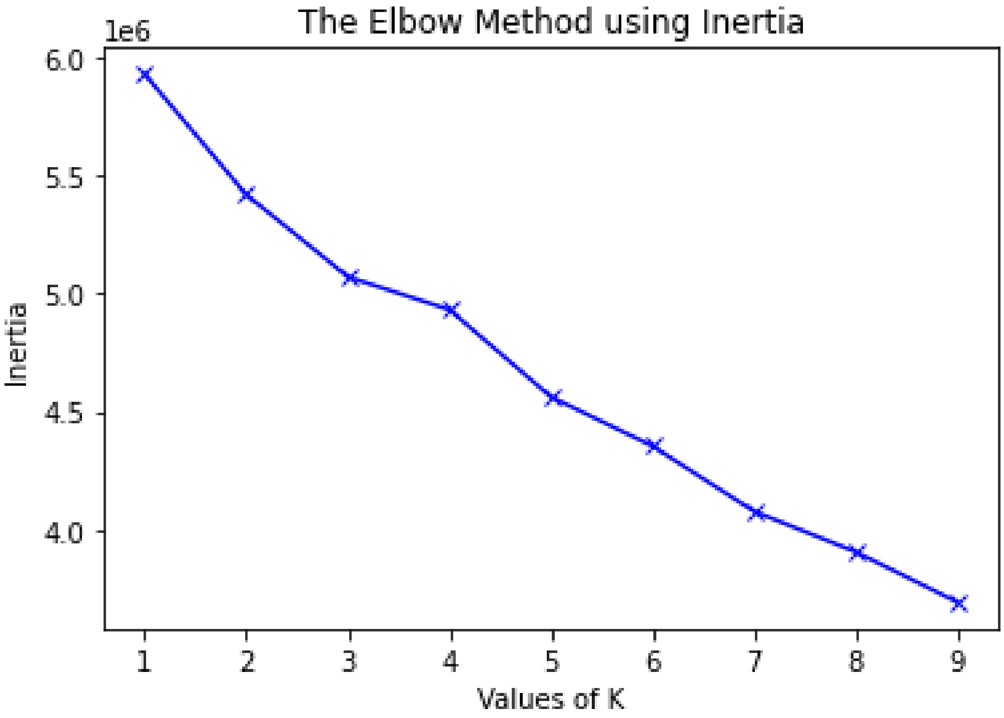
Figure 3.4.1

As we can see that the clusters formed are very random, and don’t seem to give any meaningful results. There are few bigger clusters in the center and on the outskirts are very small outlier clusters. And from looking at the clusters there was no generic pattern that could be observed.

3.4.2 K-Means Clustering on Events

Due to the fact that our last approach failed and that we got no meaningful results, we tried to adopt another approach, this approach was to take all the events a user takes in a session and one-hot encode those events, thus we get a sparse matrix of 1s and 0s where 1s indicate whether the event took place and 0s that the event was generated by the user.

We used the elbow method along with K-Means clustering to find the optimal number of clusters. However, no dip could be seen indicating that the clustering would not give any meaningful results. The following 2 diagrams (Elbow Method & Clusters) are shown below in Figure 3.4.2.1



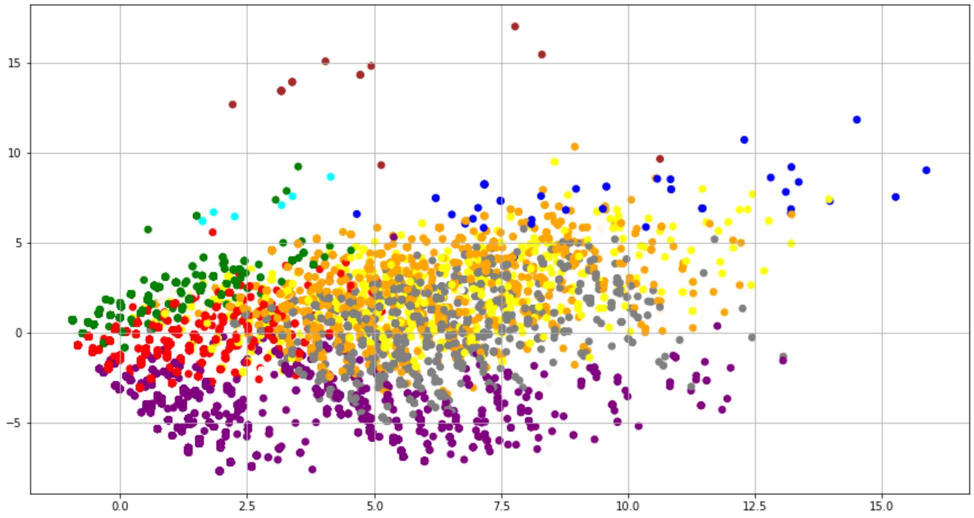


Figure 3.4.2.1

As can be seen that the clusters didn't seem to give any valuable information on what events users take during a session which could lead him to be a potential buyer. So, we move to our third method.

3.4.3 K-Means Clustering on Sequential Events Using NLP

In the previous clustering technique, we did not make use of the sequence of events that the users generated. Here using Natural Language Processing techniques, we will try and attempt to do a more meaningful clustering by looking at the events as sequence rather than Boolean values. In this clustering we made embeddings of the sessions using TF-IDF which showed best results among all the vectorizers. After incorporating TF-IDF, we experimented with N-grams of various lengths and noted that our overall results further improved by this. We used N-grams of length 2 to 6 as the sessions of length greater than 6 are outliers and most probably occurrences of OLX scraping however that's not the concerned problem in these subsets of experiments.

After the above vectorization, we tried various clustering algorithms and observed that the algorithms with high space complexity failed to give results under our constrained resources. The best performant was the K-Means algorithm. By this clustering, we observed three dominant clusters. Below we reflect on each cluster separately.

The first cluster was consisting of 30% of the total data and mostly contained "fb, listings, view" events. It shows that the user visited the site through Facebook and viewed the listings and some products and did not purchase anything.

The second cluster consisted of the 55% of the total data and mostly comprised of "source, listings, view, …., listings, view". This cluster is same as the first one with the exception that user viewed multiple listings/products and entered the site through any other external source than Facebook. Also, there were no purchases in this cluster as well.

Finally, the third cluster consisted of 15% of total events and contained all the sessions with purchases. However, there was no apparent pattern among the sessions which could easily describe the purchases. Regardless, we aim to further investigate this cluster and come up with extracted patterns which could describe the pattern of events potentially leading to purchases.

CLUSTERING OF SERIOUS BUYERS

Due to getting much better results using NLP techniques we continued with that approach where we separately looked at clusters of serious buys and non-serious buyers. We cluster user sessions in which events indicating a sale is present. 27,144 such sessions were extracted out of 186,379 and processed further. The events of interest (EoIs) or what we previously have been referring to as purchase events are the following:

1. Item_call_button_press
2. Item_chat_open_location
3. Item_chat_shared_search_location
4. Item_chat_tap_chat
5. Item_chat_tap_make_offer
6. Item_chat_tap_send_1st_reply
7. Item_chat_tap_send_offer
8. Item_chat_tap_sms
9. Item_intent_call
10. Item_intent_chat
11. Item_intent_sms
12. Item_tap_call
13. Item_tap_phone

Within these sessions we create sequences of events of size 3. We use these sequences to create a corpus to pass to our tf-idf vectorizer. By doing so, we can ascertain which sequences of events are more repetitive/important when a buyer comes and buys something on the OLX platform. Using this tf-idf matrix, we cluster these sessions using K-means clustering with 3 centers. The results are shown below in Figure 3.4.3

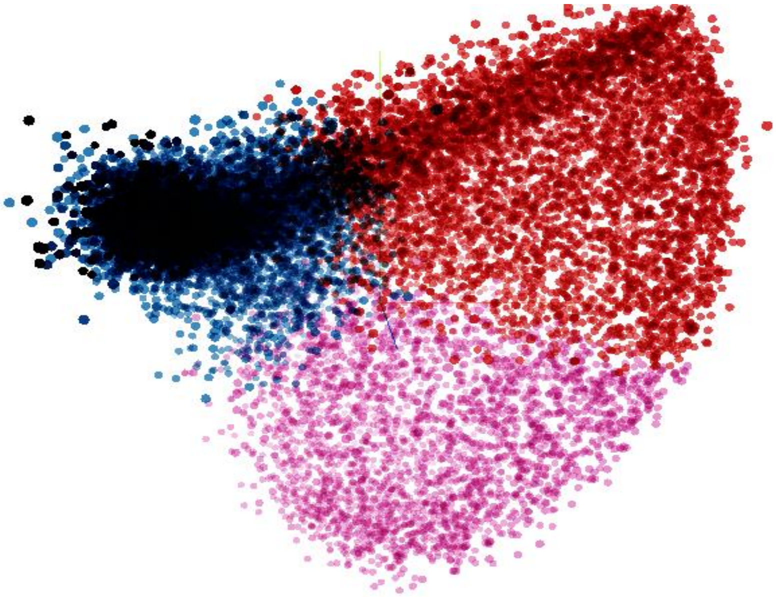


Figure 3.4.3

This graph shows 3 apparent and distinguishable clusters, with one cluster having a quite dense center. The sizes of the cluster are as follows:

Blue (Cluster 0)	Red (Cluster 1)	Purple (Cluster 2)
17,355	6,651	3,138

CLUSTER 0

Cluster 0 contains approximately 64% of our sessions of interest. The sessions included within this cluster show greater amounts of events indicating a user's event to buy the listed product. On average, after every two events there is an EoI present within a session. The sessions in cluster 0 also include other events that are not usually seen in the other two clusters e.g. carousel_listings_results. The following are some example sessions from the cluster:

1. "Google-search","external","view_item","item_chat_tap_chat","item_intent_chat","item_call_button_press","item_tap_call","item_tap_phone","view_item"

2. "Facebook","external","view_item","item_call_button_press","item_tap_call","item_call_button_press", "item_tap_call"
3. "Olx-pk","search","view_item","item_call_button_press","item_intent_call","item_tap_call","item_tap_phone","view_item","item_call_button_press","item_tap_call","view_item","item_call_button_press","item_tap_call","view_item","item_call_button_press","item_tap_call","view_item","item_call_button_press","item_tap_call","item_tap_phone","view_item","item_call_button_press","item_tap_call","view_item"

CLUSTER 1 & 2

The other two clusters have much fewer EoIs than cluster 0. Both clusters mainly have sessions that indicate a user browses the platform most of the time with on average one or two EoIs per session. There are two main differences between cluster 1 and 2:

1. Cluster 2 has sessions that are slightly longer sessions than those in Cluster 1.
2. Cluster 2 has much fewer EoIs than Cluster 1.

Some example sessions from cluster 1:

1. "Olx-pk","search","view_listings","view_item","view_listings","view_item","view_listings","view_item","item_intent_call","item_call_button_press","item_intent_call","view_item","view_listings","view_item","view_listings","view_item","item_call_button_press","item_intent_call","item_tap_call","view_listings","view_item","view_listings","view_item","view_listings","item_call_button_press","item_tap_call","view_item"
2. "Olx-pk","browse","view_listings","view_item","view_listings","view_item","view_listings","view_item","item_call_button_press","item_intent_call"
3. "olx-pk","browse","view_listings","view_item","view_listings","view_item","view_listings","view_item","view_listings","view_item","item_intent_call","item_call_button_press","item_chat_tap_sms","item_tap_call", "item_call_button_press","view_item"

The number of EoIs decrease from Cluster 0 to 1 to 2. From moving from the clusters in the order mentioned, one can see that the length of a session on average. The EoIs within a session decrease considerably between clusters with specific events increasing just as much. These events indicate that a user is browsing listings on the platform. Clusters 1 and 2 indicate a user browses on the platform on shows an interest to buy on one or two listings. Cluster 0 indicates that a user may behave this way for two reasons:

1. The user continues to show interest in only one listing by repeatedly performing events on that listing.
2. The user shows interest on multiple listings within one session along with performing EoIs for each of those listings.

CLUSTERING OF NON-SERIOUS BUYERS

The number of sessions present in this part of the visualization are 159,235, accounting for about 85% of the entire data. There are two apparent clusters after applying k-means clustering. Unfortunately, the visualization of the two clusters are not visible enough.

The size of the two clusters are as follows:

Cluster 0(C0)	Cluster 1(C1)
60,216	99,019

The main difference between the two clusters is basically the length of sessions in both. Cluster 0 contains sessions of length 3-5. Cluster 1 contains sessions longer than that of C0.

One possible explanation for C0 is that users click on an ad and view one or two listings max and terminate the session. While users in C1, browse on the platform for much longer but do not trigger our EoIs in any circumstance.

4. Conclusion

From our EDA and Cluster analysis, we have identified key features and sequential events that indicate the seriousness of a buyer on the platform. Building on this, we are currently working on converting our problem into a classification problem i.e. how soon can you tell that a user on the platform is willing to buy a product? given the initial sequences of events a person takes. For this problem, we are looking into two main ideas, clickstream analysis through Hidden Markov Models and Sequential Pattern Mining. **Hidden Markov Models (HMMs)** are a class of probabilistic graphical models that allow us to predict a sequence of unknown (hidden) variables from a set of observed variables. With this we can look at the first few observable events a person takes and from that determine how likely is he to purchase an event. **Sequential Pattern Mining** on the other hand is a topic of data mining concerned with finding statistically relevant patterns between data examples where the values are delivered in a sequence. With this we can perform algorithms such as BIDE and TSP to figure out what sequential patterns among events are most frequent and determine what sequential combination leads to customers purchasing an item. This sequential approach through HMMs and Sequential Pattern Mining are only possible due to the insight provided by the cluster analysis conducted on sequential events.