In our analysis, we have decided to consider month by month and not all the months together because we want to study the behaviour of e-commerce during the selected period.

Since all the dataset have the same features and the same name for the variable, we have written one code and then we have run for each month by changing the name of the importing dataset.

Question 1)

To calculate the rate of complete funnels, we decided to group all event_time values by ‘product_id’ and ‘user_id’, drop all empty rows where at least one of the events wasn't made, count sum and divide by total number of events.

	October 2019	November 2019	December 2019	January 2020	February 2020	March 2020	April 2020
Rate	0.015	0.033	0.0325	0.0274	0.0333	0.0347	0.0271

Next is the operation that users repeat more on average: we grouped data by ‘session’ and took all average numbers of events in ‘event_type’ and then plot a bar, and took the name of event with highest number.

	October 2019	November 2019	December 2019	January 2020	February 2020	March 2020	April 2020
The operation	view	view	view	view	view	view	view

How many times, on average, a user views a product before adding it to the cart?

We made a pivot table of the times of first events, that users made, and then counted probability of user to view a product before buying

	October 2019	November 2019	December 2019	January 2020	February 2020	March 2020	April 2020
Rate	0.021	0.0465	0.049	0.043	0.041	0.049	0.0461

What’s the probability that products added once to the cart are effectively bought?

We used the same pivot from previous question, counting all rows with time of 'cart' before the 'purchase' and divided by total length

	October 2019	November 2019	December 2019	January 2020	February 2020	March 2020	April 2020
Rate	0.493	0.370	0.486	0.438	0.422	0.510	0.438

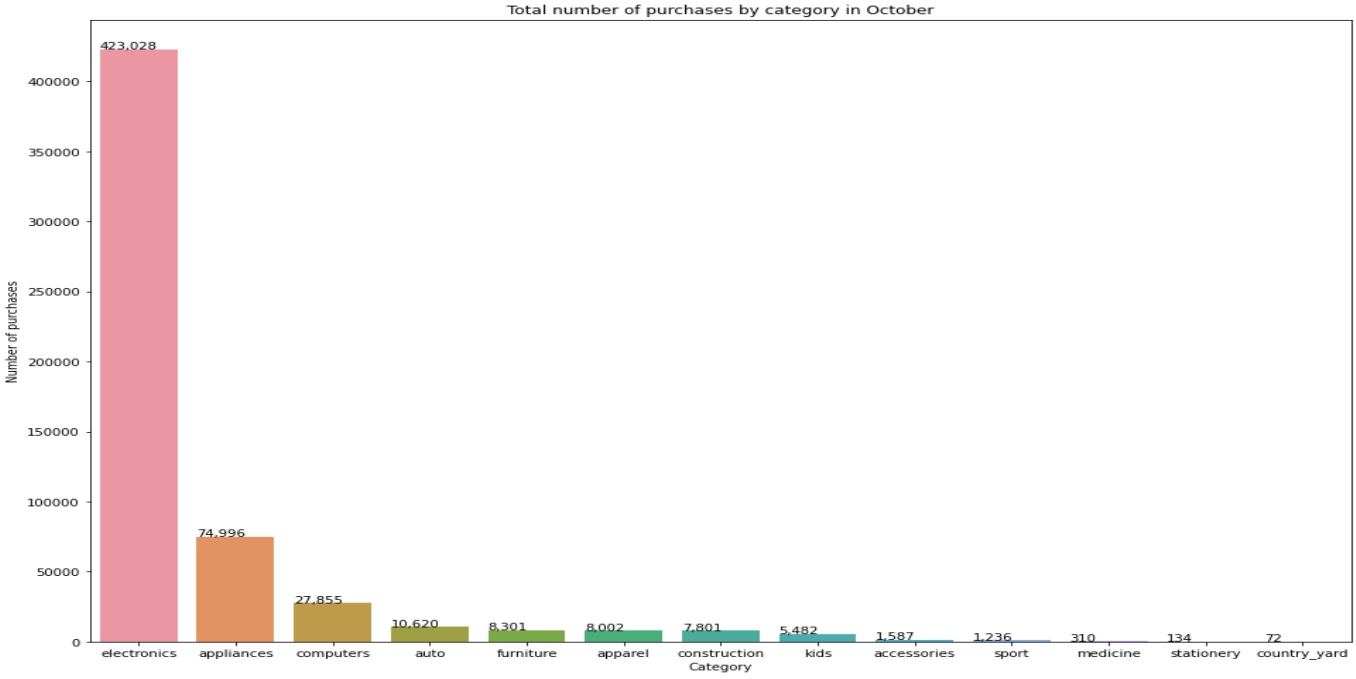
Question 2)

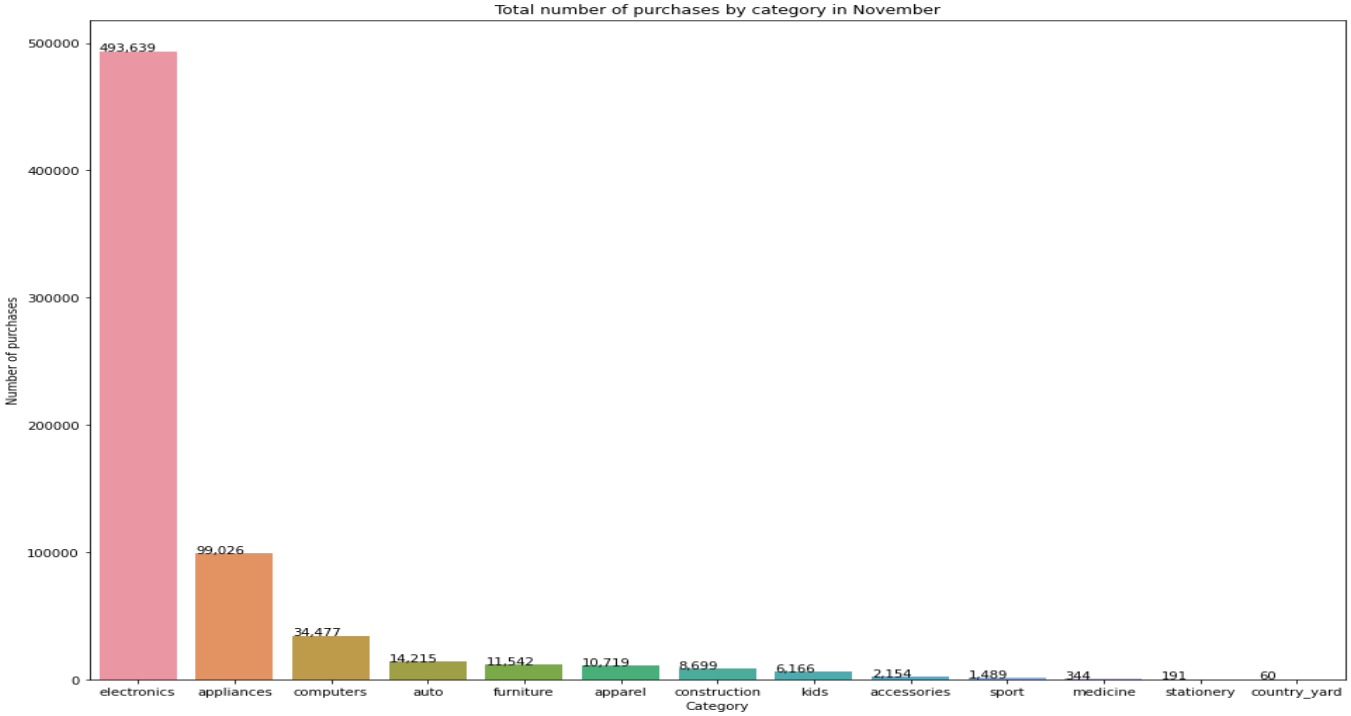
What are the categories of the most trending products overall?

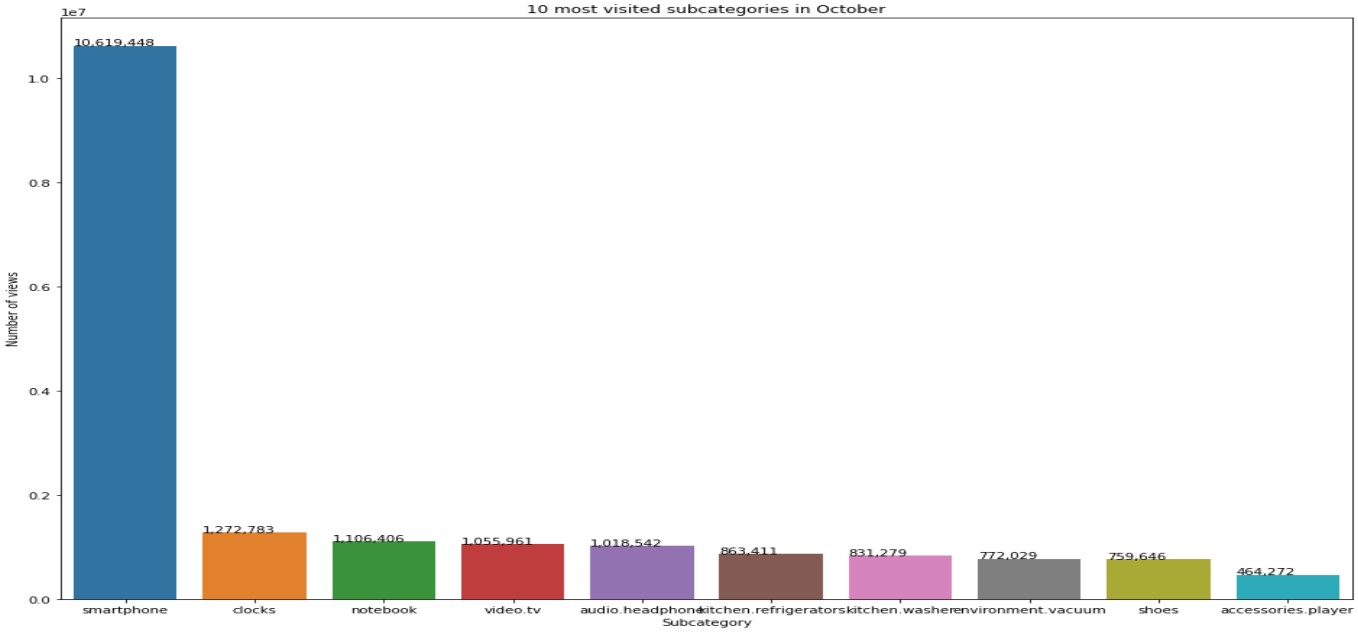
We split category_code into 'category' and 'subcategory' columns and count number of purchases of products and sort in descending order

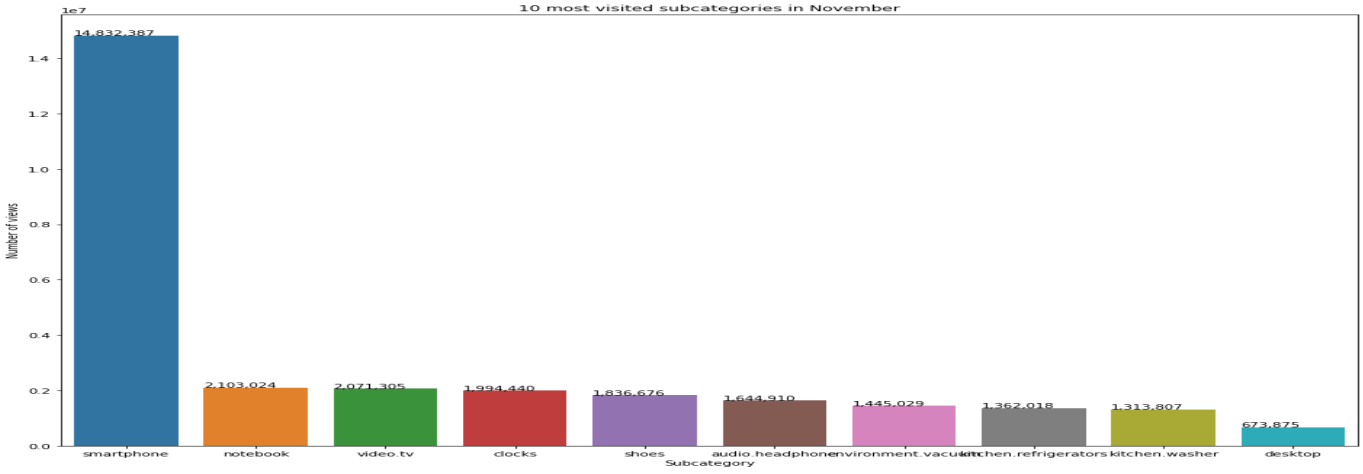
	October 2019	November 2019	December 2019	January 2020	February 2020	March 2020	April 2020
Category	electronics	electronics	construction, sport	construction, sport	construction, sport	construction, sport	construction, sport

Plots of total number of purchases and most trending subcategories:









Plots of the rest months can be found in notebooks with their names.

What are the 10 most sold products per category?

We took all names in categories with method ‘unique’ and built the code to make pivot table for every category name with top 10 most sold products in each.

Question 3)

In order to find the brand whose prices are higher on average, we have decided to select in the dataset only some columns, which are useful for this kind of analysis. In particular, we have chosen to use the following column: “category_code”, “brand” and “price”. Then, we group our variable by the “category_code” and the “brand”, and we compute the mean of the price for the variable. Finally, we split the variable “category_code” in “category” and “sub_category” and we arrive to find the following dataframe:

	average_price	category	sub_category	category_code	brand
0	40.401493	accessories	bag	accessories.bag	a-elita
1	48.979696	accessories	bag	accessories.bag	acer
2	15.461552	accessories	bag	accessories.bag	acron
3	100.800070	accessories	bag	accessories.bag	apple
4	14.219047	accessories	bag	accessories.bag	asus
...	...	...	...	...	...
3455	14.513684	stationery	cartrige	stationery.cartrige	sakura
3456	34.248018	stationery	cartrige	stationery.cartrige	samsung
3457	10.751290	stationery	cartrige	stationery.cartrige	superfine
3458	66.670000	stationery	cartrige	stationery.cartrige	xerox
3459	20.570000	stationery	cartrige	stationery.cartrige	xiaomi

3460 rows × 5 columns

Now, we want to write a function that asks the user a category in input and returns a plot indicating the average price of the products sold by the brand. We have decided to write three different functions in order to let the user decide which kind of variables he wants to plot among “category_code”, “category”, and “sub_category”. The following are the script of the three different function:

```
# Function for plot the sub_category
def subcategory_plot_avg_price(word):

    table2 = table.loc[table['sub_category'] == word]
    plt.figure(figsize = (15,10))
    ax = plt.subplot()
    ax.barh(table2['brand'].head(40),
              table2['average_price'].head(40),
              align = 'center', edgecolor = 'k')
    plt.axvline(table2['average_price'].mean(), c='red')
    plt.xlabel('Average Price')
    plt.ylabel('Brand')
    plt.title(f'Plot the January average price of {word}.')
    plt.show()

    return ax
```

```
# Function for plot the category
def category_plot_avg_price(word):

    table2 = table.loc[table['category'] == word]
    plt.figure(figsize = (15,10))
    ax = plt.subplot()
    ax.barh(table2['brand'].head(40),
              table2['average_price'].head(40),
              align = 'center', edgecolor = 'k')
    plt.axvline(table2['average_price'].mean(), c='red')
    plt.xlabel('Average Price')
    plt.ylabel('Brand')
    plt.title(f'Plot the January average price of {word}.')
    plt.show()

    return ax
```

```
# Function for plot the category code
def category_code_plot_avg_price(word):

    table2 = table.loc[table['category_code'] == word]
    plt.figure(figsize = (15,10))
    ax = plt.subplot()
    ax.barh(table2['brand'].head(40),
              table2['average_price'].head(40),
              align = 'center', edgecolor = 'k')
    plt.axvline(table2['average_price'].mean(), c='red')
    plt.xlabel('Average Price')
    plt.ylabel('Brand')
    plt.title(f'Plot the January average price of {word}.')
    plt.show()

    return ax
```

Finally, we want to show for each category, the brand with the highest average price. Therefore, we group our variables by “category” and then we find the maximum for each of them. The following are the result for each month after sort the price in ascending order:

October:

	category	Max_price	sub_category	category_code	brand
2	appliances	2393.170979	kitchen.refrigerators	appliances.kitchen.refrigerators	climadiff
11	sport	2382.340526	bicycle	sport.bicycle	pinarello
7	electronics	2087.455158	clocks	electronics.clocks	rado
9	kids	1801.820000	carriage	kids.carriage	hartan
4	computers	1801.820000	notebook	computers.notebook	dreammachines
5	construction	1612.069990	tools.generator	construction.tools.generator	senci
8	furniture	1546.892963	universal.light	furniture.universal.light	aldit
3	auto	1135.113469	accessories.compressor	auto.accessories.compressor	schneider
6	country_yard	983.237745	lawn_mower	country_yard.lawn_mower	viking
1	apparel	715.521637	dress	apparel.dress	weekend
0	accessories	584.043301	bag	accessories.bag	weekend
12	stationery	66.670000	cartrige	stationery.cartrige	xerox
10	medicine	51.997156	tools.tonometer	medicine.tools.tonometer	omron

November:

	category	Max_price	sub_category	category_code	brand
11	sport	2573.810000	bicycle	sport.bicycle	pinarello
7	electronics	2562.630000	audio.acoustic	electronics.audio.acoustic	fly
4	computers	2558.752243	notebook	computers.notebook	dreammachines
2	appliances	2234.804821	kitchen.washer	appliances.kitchen.washer	siemens
5	construction	2199.047736	tools.pump	construction.tools.pump	helix
8	furniture	2006.490000	bathroom.bath	furniture.bathroom.bath	jacobdelafon
9	kids	1801.820000	carriage	kids.carriage	hartan
1	apparel	1332.252174	shoes	apparel.shoes	sergiorossi
3	auto	1123.877582	accessories.compressor	auto.accessories.compressor	metabo
6	country_yard	896.938462	lawn_mower	country_yard.lawn_mower	viking
0	accessories	773.611582	bag	accessories.bag	weekend
12	stationery	66.670000	cartrige	stationery.cartrige	xerox
10	medicine	53.110095	tools.tonometer	medicine.tools.tonometer	omron

December:

	category	Max_price	sub_category	category_code	brand
7	electronics	2574.040000	video.tv	electronics.video.tv	dexp
2	appliances	2573.810000	kitchen.toster	appliances.kitchen.toster	pinarello
9	kids	2548.330000	skates	kids.skates	volta
11	sport	2522.590000	trainer	sport.trainer	nordictrack
1	apparel	2445.370000	trousers	apparel.trousers	gomeldrev
4	computers	2397.367678	peripherals.printer	computers.peripherals.printer	sony
5	construction	2344.980000	tools.screw	construction.tools.screw	yjfitness
3	auto	2268.010000	accessories.winch	auto.accessories.winch	femi
6	country_yard	2044.250000	lawn_mower	country_yard.lawn_mower	shua
8	furniture	1844.249717	living_room.sofa	furniture.living_room.sofa	trevi
0	accessories	1835.218077	bag	accessories.bag	helix
12	stationery	940.808030	cartrige	stationery.cartrige	dewalt
10	medicine	216.055650	tools.tonometer	medicine.tools.tonometer	hayali

January:

	category	Max_price	sub_category	category_code	brand
2	appliances	2573.810000	kitchen.toster	appliances.kitchen.toster	pinarello
8	furniture	2570.980000	kitchen.table	furniture.kitchen.table	florencemode
7	electronics	2561.200000	audio.acoustic	electronics.audio.acoustic	monitoraudio
9	kids	2548.330000	skates	kids.skates	volta
1	apparel	2445.370000	trousers	apparel.trousers	gomeldrev
11	sport	2372.000000	ski	sport.ski	kessler
4	computers	2360.684055	components.power_supply	computers.components.power_supply	buderus
5	construction	2344.980000	tools.screw	construction.tools.screw	yjfitness
3	auto	2280.568638	accessories.winch	auto.accessories.winch	femi
6	country_yard	2046.008387	lawn_mower	country_yard.lawn_mower	shua
0	accessories	1806.154857	bag	accessories.bag	helix
12	stationery	921.917094	cartrige	stationery.cartrige	dewalt
10	medicine	267.426757	tools.tonometer	medicine.tools.tonometer	jade

February:

	category	Max_price	sub_category	category_code	brand
2	appliances	2558.259756	kitchen.toster	appliances.kitchen.toster	pinarello
8	electronics	2540.929286	camera.photo	electronics.camera.photo	sony
9	furniture	2522.590000	kitchen.table	furniture.kitchen.table	active
6	construction	2470.850000	tools.painting	construction.tools.painting	zipp
12	sport	2372.000000	ski	sport.ski	kessler
3	auto	2290.920000	accessories.winch	auto.accessories.winch	femi
4	auto	2290.920000	kitchen.table	furniture.kitchen.table	unifur
1	apparel	2252.310000	underwear	apparel.underwear	uta
5	computers	2211.547893	peripherals.printer	computers.peripherals.printer	sony
10	kids	1914.641855	skates	kids.skates	volta
0	accessories	1789.171852	bag	accessories.bag	helix
7	country_yard	913.790000	lawn_mower	country_yard.lawn_mower	gf
13	stationery	524.031489	cartrige	stationery.cartrige	yamaha
11	medicine	264.273333	tools.tonometer	medicine.tools.tonometer	jade

March:

	category	Max_price	sub_category	category_code	brand
2	appliances	2574.040000	kitchen.coffee_machine	appliances.kitchen.coffee_machine	aeg
4	computers	2548.330000	components.power_supply	computers.components.power_supply	buderus
8	furniture	2540.269125	kitchen.table	furniture.kitchen.table	ego
7	electronics	2531.053333	tablet	electronics.tablet	wacom
5	construction	2470.850000	tools.painting	construction.tools.painting	zipp
12	sport	2372.000000	ski	sport.ski	kessler
3	auto	2290.920000	accessories.winch	auto.accessories.winch	femi
1	apparel	2252.310000	underwear	apparel.underwear	uta
9	kids	1801.820000	notebook	computers.notebook	dreammachines
10	kids	1801.820000	carriage	kids.carriage	hartan
0	accessories	1776.081765	bag	accessories.bag	helix
6	country_yard	859.740000	lawn_mower	country_yard.lawn_mower	hygge
13	stationery	472.933519	cartrige	stationery.cartrige	sony
11	medicine	236.808100	tools.tonometer	medicine.tools.tonometer	hayali

April:

	category	Max_price	sub_category	category_code	brand
8	construction	2571.360843	tools.painting	construction.tools.painting	zipp
7	computers	2558.370000	components.cooler	computers.components.cooler	cime
1	apparel	2547.530000	shorts	apparel.shorts	knief
10	electronics	2514.350000	audio.music_tools.piano	electronics.audio.music_tools.piano	fender
11	furniture	2512.170000	bathroom.toilet	furniture.bathroom.toilet	grohe
2	appliances	2393.630000	kitchen.toster	appliances.kitchen.toster	pinarello
13	sport	2372.000000	ski	sport.ski	kessler
3	auto	2290.920000	accessories.winch	auto.accessories.winch	femi
4	auto	2290.920000	toys	kids.toys	cybex
5	kids	2290.920000	accessories.winch	auto.accessories.winch	femi
6	kids	2290.920000	toys	kids.toys	cybex
0	accessories	1974.672200	bag	accessories.bag	helix
9	country_yard	859.740000	lawn_mower	country_yard.lawn_mower	hygge
14	stationery	416.113813	cartrige	stationery.cartrige	sony
12	medicine	278.000000	tools.tonometer	medicine.tools.tonometer	jade

We can note that each month the category with the highest price change; on the other hand, all months have the same category with the lowest price, which are: “medicine” and “stationary”.

Question 4)

How much does each brand earn per month?

We made a pivot table to sum the selling price per each brand and then sort it by descending order.

In all months, technological companies, such as apple,samsung,huawei,xiaomi were in the top of the list.

Write a function that given the name of a brand in input returns, for each month, its profit.

```
def profit_of_brand_for_each_month(brand_name,list_of_months):
    return [i['purchase'][brand_name] for i  in list_of_months]
```

Using the function you just created, find the top 3 brands that have suffered the biggest losses in earnings between one month and the next, specifying both the loss percentage and the 2 months (e.g., brand_1 lost 20% between march and april).

With the help of previous function we calculated the profit for each brand in both months, made Dataframe for each month and made an outer join on them, with additional column ‘loss’

Is the average price of products of different brands significantly different?

Yes, it can vary from decimals to hundreds and thousands. We can clearly see it in pivot tables, that we made

Question 5)

In order to find in what part of the day the store is most visited, we have decided to select in the dataset only some columns, which are useful for this kind of analysis. In particular, we have chosen to use the following column: “category_code”, “event_type” and “event_time”. Then, we transform the variable “event_time” in datetime using the python function “pd.to_datetime”. After that, we resample our data for different time horizons. In particular, we consider: 1 hour, 6 hour, 1 day, and 1 week starting from Monday.
In order to find in what part of the day the store is most visited, we consider the resample data with 6 hour. In this case, we split the day in 4 different range, each with length 6 hours:

- first range: from 00:00:00 to 06:00:00
- second range: from 06:00:00 to 12:00:00
- third range: from 12:00:00 to 18:00:00
- fourth range: from 18:00:00 to 00:00:00

Then, we have sorted the result in ascending order. In this way, we have found the part of the day and the day which has the highest number of visitors. In following tables are shown the top three and the last three visitors with daily range:

October 2019		
Date	Hour Range	Count
13-10-2019	06:00:00 - 12:00:00	610512
15-10-2019	12:00:00 - 18:00:00	589817
13-10-2019	12:00:00 - 18:00:00	579856
03-10-2019	18:00:00 -00:00:00	148241
07-10-2019	18:00:00 - 00:00:00	140420
02-10-2019	18:00:00 - 00:00:00	131140

November 2019		
Date	Hour Range	Count
17-11-2019	12:00:00 - 18:00:00	2779318
16-11-2019	06:00:00 - 12:00:00	2255272
17-11-2019	06:00:00 - 12:00:00	2216935
27-11-2019	18:00:00 -00:00:00	194982
25-11-2019	18:00:00 - 00:00:00	194971
26-11-2019	18:00:00 - 00:00:00	190272

December 2019		
Date	Hour Range	Count
17-12-2019	12:00:00 - 18:00:00	1107182
16-12-2019	12:00:00 - 18:00:00	1106526
18-12-2019	12:00:00 - 18:00:00	1105419
04-12-2019	18:00:00 -00:00:00	209654
03-12-2019	18:00:00 - 00:00:00	192350
31-12-2019	18:00:00 - 00:00:00	187650

January 2020		
Date	Hour Range	Count
19-01-2020	06:00:00 - 12:00:00	857471
31-01-2020	12:00:00 - 18:00:00	848071
19-01-2020	12:00:00 - 18:00:00	804141
27-01-2020	18:00:00 - 00:00:00	224517
02-01-2020	06:00:00 - 06:00:00	223629
01-01-2020	00:00:00 - 06:00:00	152842

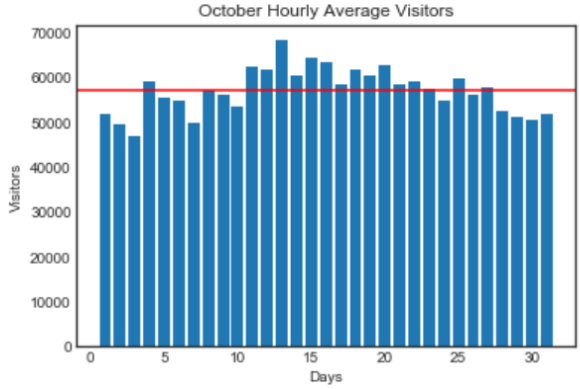
February 2020		
Date	Hour Range	Count
12-02-2020	12:00:00 - 18:00:00	1003265
12-02-2020	06:00:00 - 12:00:00	942731
16-02-2020	06:00:00 - 12:00:00	942224
27-02-2020	12:00:00 - 18:00:00	45688
28-02-2020	00:00:00 - 06:00:00	40145
27-02-2020	18:00:00 - 00:00:00	11881

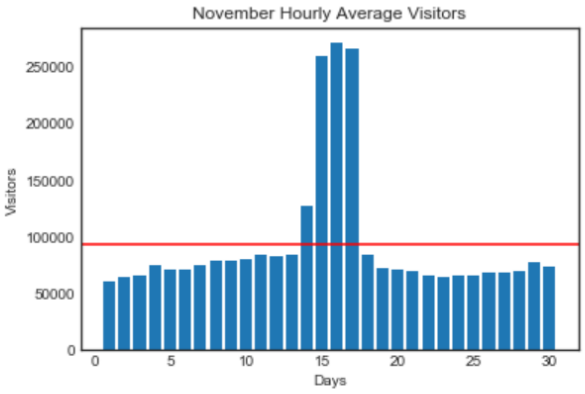
March 2020		
Date	Hour Range	Count
03-03-2020	12:00:00 - 18:00:00	1021974
04-03-2020	12:00:00 - 18:00:00	1007508
02-03-2020	12:00:00 - 18:00:00	971177
18-03-2020	18:00:00 -00:00:00	154089
17-03-2020	18:00:00 - 00:00:00	153526
29-03-2020	18:00:00 - 00:00:00	130946

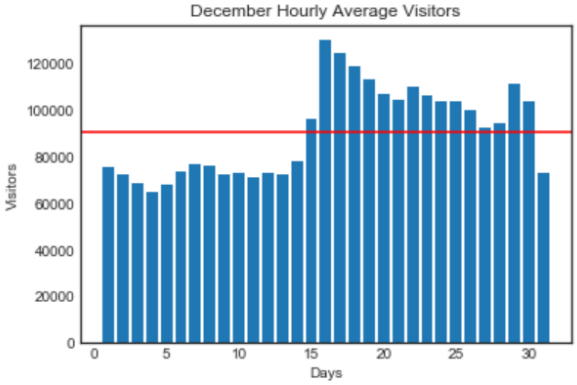
April 2020		
Date	Hour Range	Count
26-04-2020	06:00:00 - 12:00:00	979617
16-04-2020	12:00:00 - 18:00:00	979195
28-04-2020	12:00:00 - 18:00:00	957109
19-04-2020	18:00:00 -00:00:00	188772
07-04-2020	18:00:00 - 00:00:00	177749
04-04-2020	18:00:00 - 00:00:00	159828

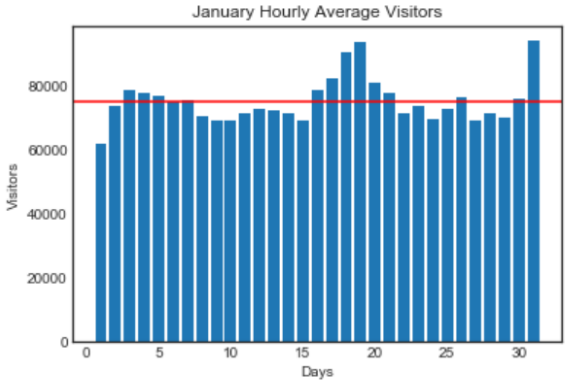
From these tables we can conclude that our shop is most visited in the second and third range than in the first and fourth range. In fact, it is easy to see that for all the tables the best three hour range is the second and the third; instead, the worst range is fourth.

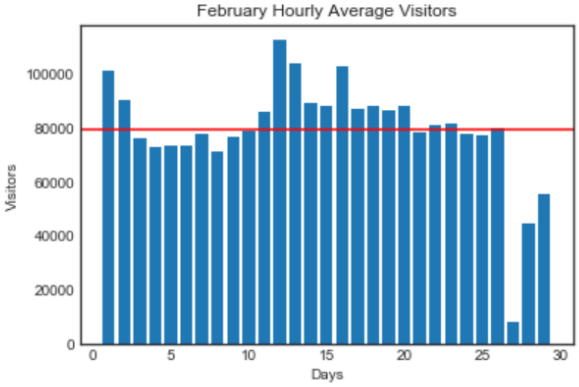
The last task consists of plotting for each day of the week the hourly average of visitors that our store has. Therefore, we consider the resample data with 1 hour and then we resample another time this data for one day and we compute the mean. The following are the plot for each month:

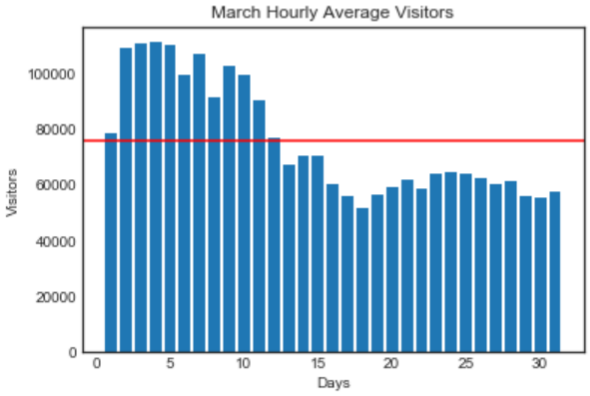


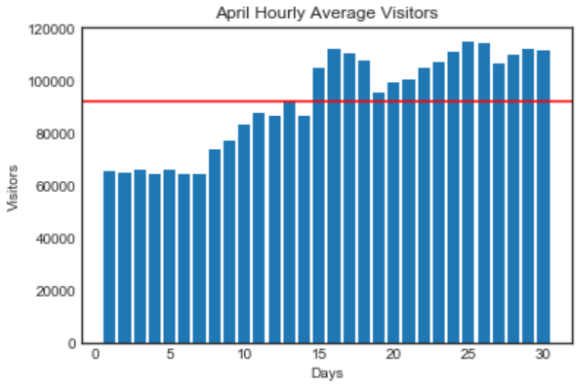

November Hourly Average Visitors
Visitors
Days
0
50000
100000
150000
200000
250000
0
5
10
15
20
25
30


December Hourly Average Visitors
Visitors
Days
0
20000
40000
60000
80000
100000
120000
0
5
10
15
20
25
30


January Hourly Average Visitors
Visitors
Days
0
20000
40000
60000
80000
0
5
10
15
20
25
30


February Hourly Average Visitors
Visitors
Days
0
20000
40000
60000
80000
100000
0
5
10
15
20
25
30





From all the plots, we can see how much is volatile the hourly average of visitors is from October 2019 to April 2020. In fact, all the months have different patterns from the previous one. It is interesting note that in December 2019 we start to have a peak of visitors in our ecommerce from the 16th and the visitors remain more or less stable until the 19th.

Question 6)

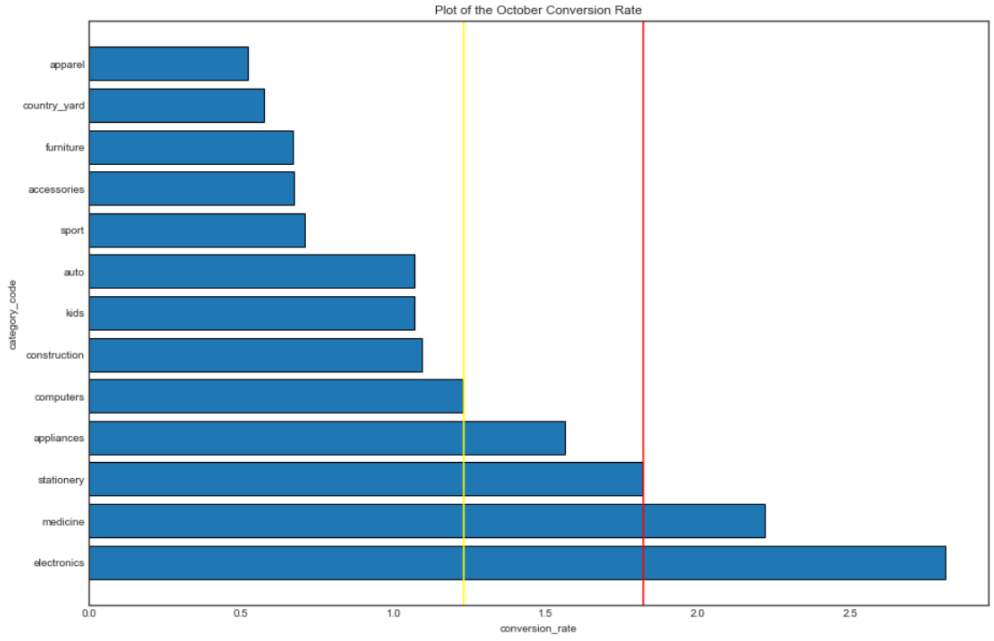
The conversion rate of a product is given by the number of times a product has been bought over the number of times it has been visited. In order to find the overall conversion rate for our store, we have decided to select in the dataset only some columns, which are useful for this kind of analysis. In particular, we have chosen to use the following column: “event_type”, “product_id”, “category_id” and “category_code”.

Now, we count the number of different events in “event_type” which could be: purchase, view and cart. The following table, show the overall conversion rate for each month:

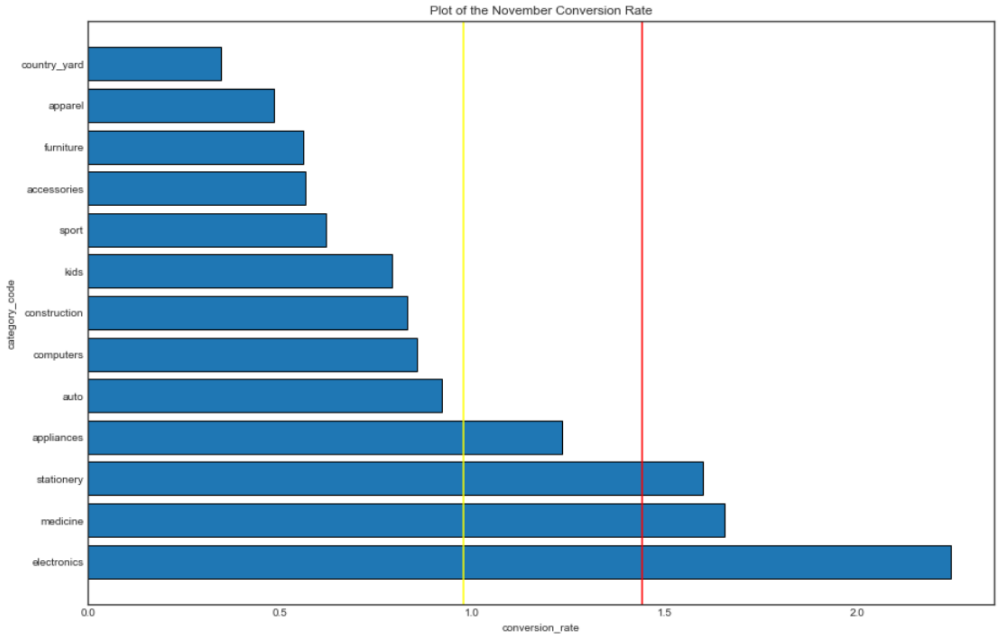
October 2019	November 2019	December 2019	January 2020	February 2020	March 2020	April 2020
1.82 %	1.44 %	1.84 %	1.59 %	2.34 %	1.96 %	1.55 %

The average conversion rate among all the months is equal to: 1.79 %.

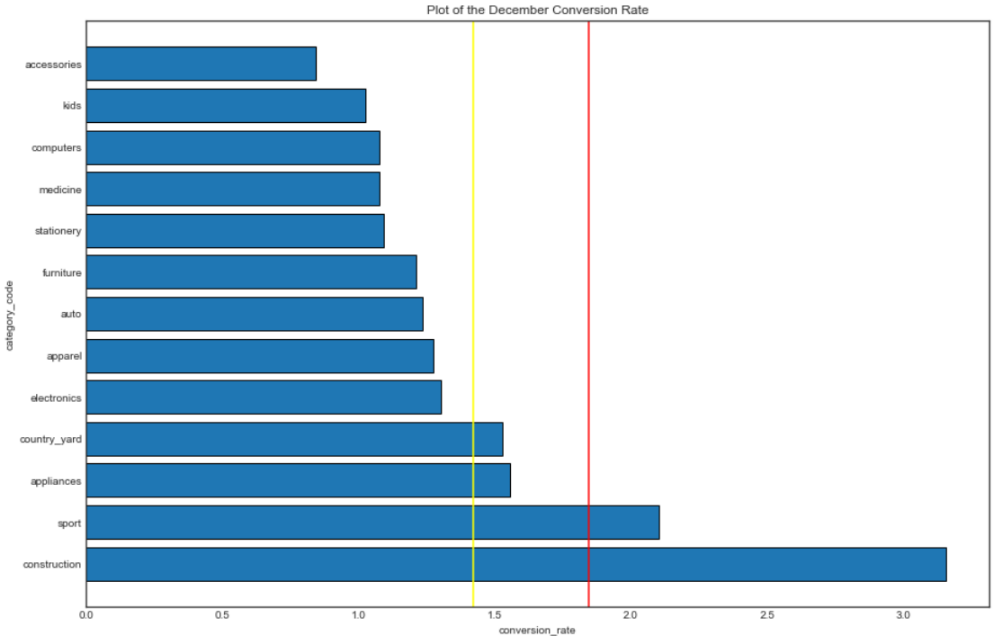
Then, we need to compute for each category the conversion rate and then plot in ascending order. In order to compute this, we perform the following step: first, we group our dataset by the “cateogy_code” and we count the “event_type”. Second, we have created the column for the purchase items and the column for the view items. Then, we have merged all together in a new dataframe called “Merge”. Finally, we have split the variable “category_code” in “category” and “sub_category” and we have computed the conversion rate for the variable “category” but before, we have deleted the column in excess. We find the following result:



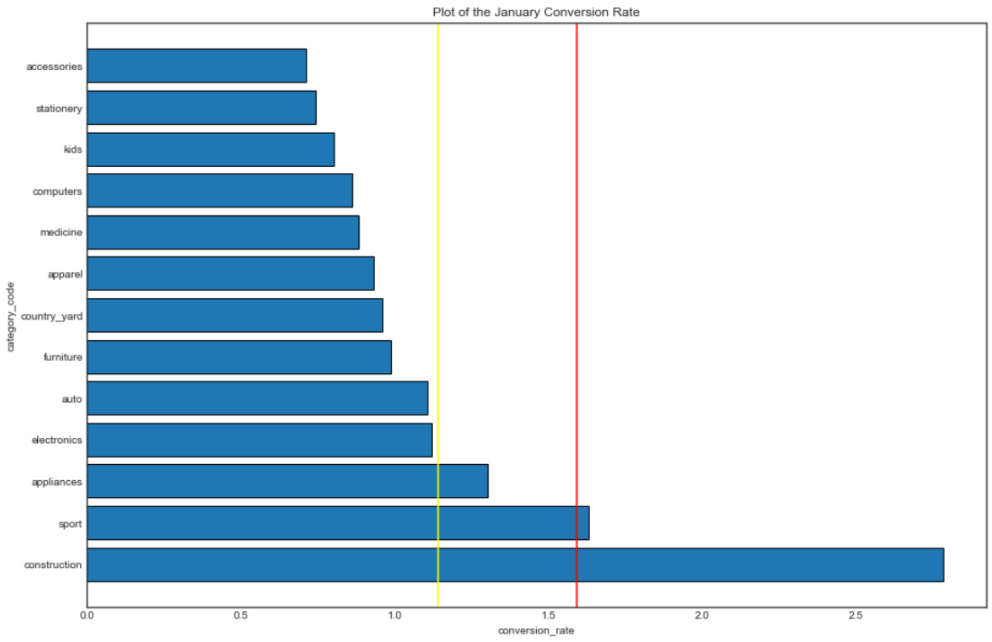
	category	conversion_rate
7	electronics	2.812987
10	medicine	2.218406
12	stationery	1.815718
2	appliances	1.562437
4	computers	1.227672
5	construction	1.094585
9	kids	1.069690
3	auto	1.069357
11	sport	0.709025
0	accessories	0.672167
8	furniture	0.671356
6	country_yard	0.573797
1	apparel	0.522268



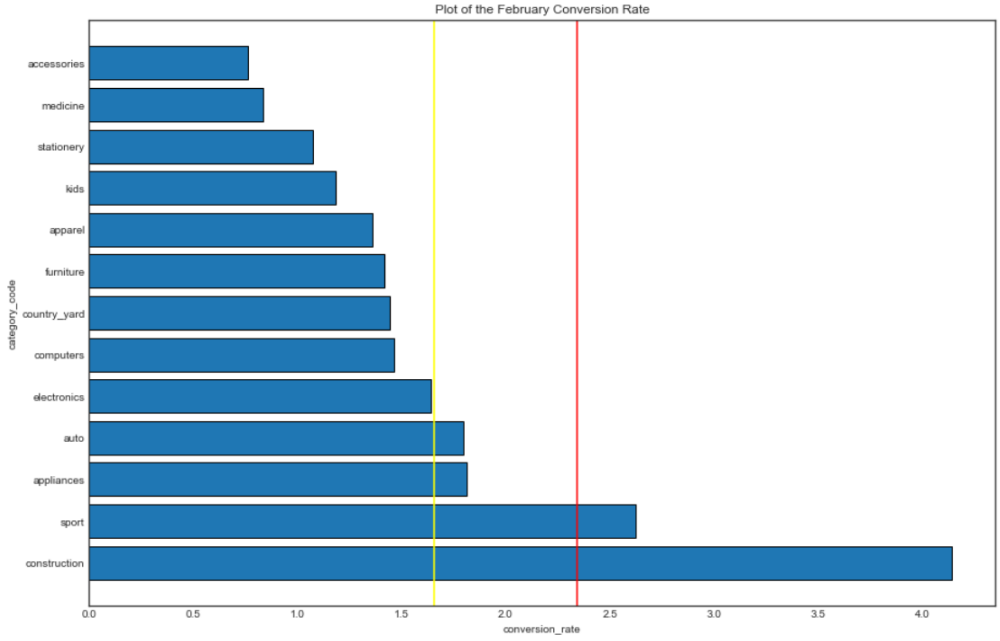
	category	conversion_rate
7	electronics	2.245019
10	medicine	1.656714
12	stationery	1.599263
2	appliances	1.231976
3	auto	0.920656
4	computers	0.855711
5	construction	0.830793
9	kids	0.791003
11	sport	0.618451
0	accessories	0.566286
8	furniture	0.560618
1	apparel	0.483303
6	country_yard	0.346881



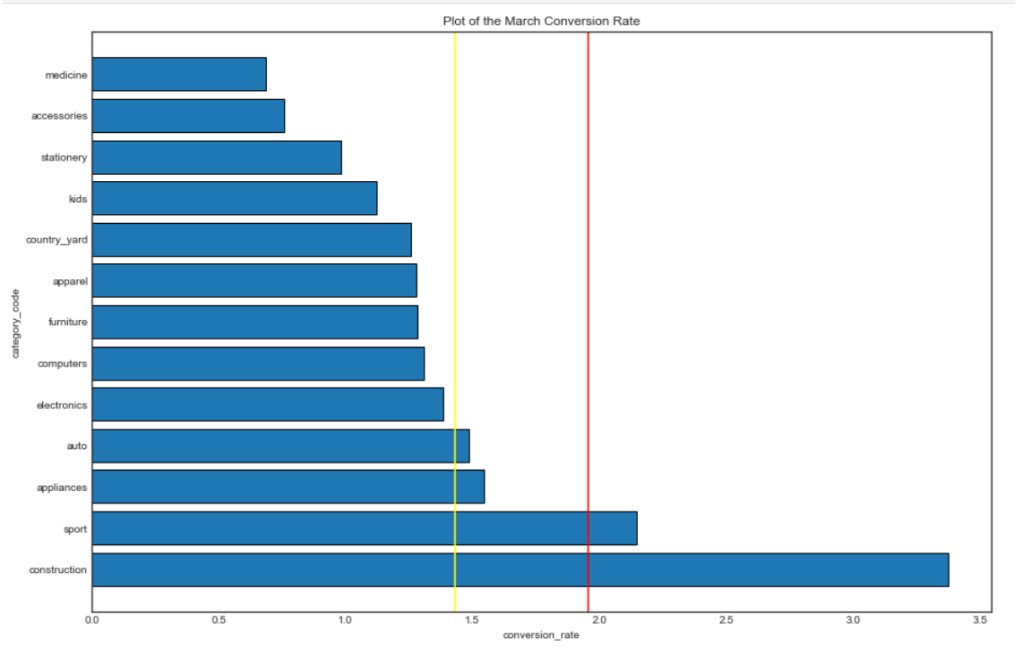
	category	conversion_rate
5	construction	3.156990
11	sport	2.104632
2	appliances	1.557255
6	country_yard	1.527898
7	electronics	1.301229
1	apparel	1.275024
3	auto	1.235583
8	furniture	1.210272
12	stationery	1.094190
10	medicine	1.076850
4	computers	1.075873
9	kids	1.026402
0	accessories	0.843021



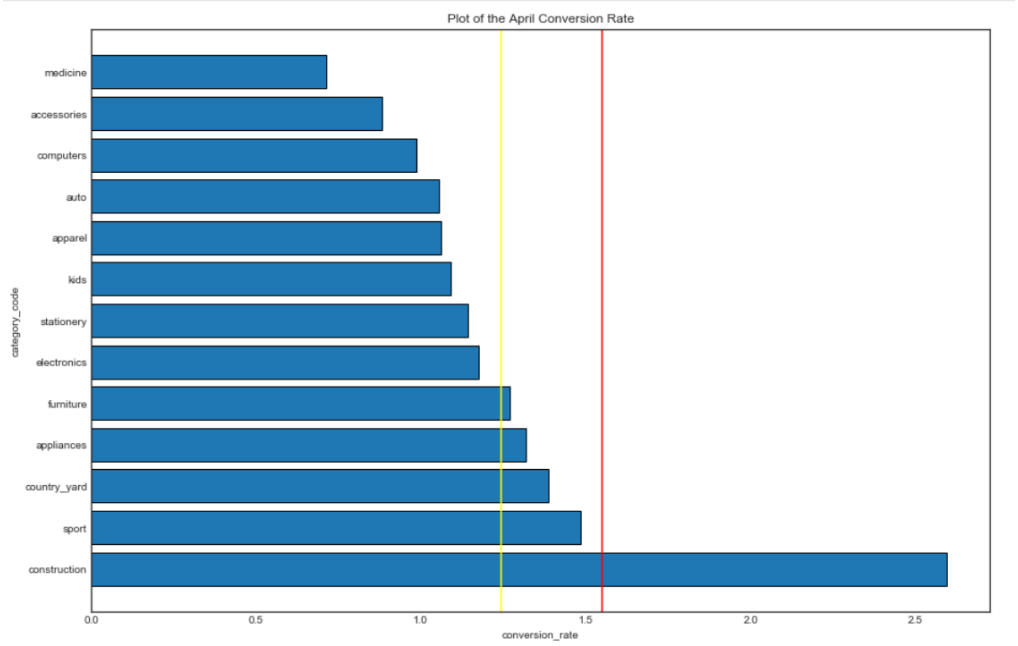
	category	conversion_rate
5	construction	2.783280
11	sport	1.629048
2	appliances	1.302445
7	electronics	1.121931
3	auto	1.108210
8	furniture	0.987573
6	country_yard	0.961586
1	apparel	0.932511
10	medicine	0.882174
4	computers	0.862653
9	kids	0.802345
12	stationery	0.744528
0	accessories	0.712642



	category	conversion_rate
5	construction	4.140509
11	sport	2.622622
2	appliances	1.812138
3	auto	1.799226
7	electronics	1.640478
4	computers	1.466623
6	country_yard	1.441042
8	furniture	1.416217
1	apparel	1.361450
9	kids	1.184395
12	stationery	1.073592
10	medicine	0.837877
0	accessories	0.762740



	category	conversion_rate
5	construction	3.376754
11	sport	2.148312
2	appliances	1.545115
3	auto	1.485204
7	electronics	1.386553
4	computers	1.309799
8	furniture	1.284267
1	apparel	1.278051
6	country_yard	1.257974
9	kids	1.123657
12	stationery	0.980810
0	accessories	0.759853
10	medicine	0.684943



	category	conversion_rate
5	construction	2.595080
11	sport	1.484078
6	country_yard	1.388107
2	appliances	1.317417
8	furniture	1.268618
7	electronics	1.176032
12	stationery	1.143619
9	kids	1.090400
1	apparel	1.061660
3	auto	1.054717
4	computers	0.985444
0	accessories	0.881798
10	medicine	0.713340

The red line represents the overall conversion rate for each month. On the other hand, the yellow line represents the mean among all the conversion rates of each category for each month.

Question 7)

Our thought was to take shop's top 20% of customers and see how much profit they bring. To do that, we grouped the data by user_id and summed all the price when user was purchasing. After sorting and deleting rows that were empty, we calculated the revenue they bring, multiplied by 100 and divided by total revenue. In all months results were near 70%, which partially proves our assumption, and with the growth of market the percentage will eventually grow to 80%.