Data Science Course: Capstone Project 1

eCommerce behavior in events

Capstone Mini-Project: Data Story

Dataset: Collected two months (Oct & Nov 2019) e-commerce events dataset from a large multi-category online store.

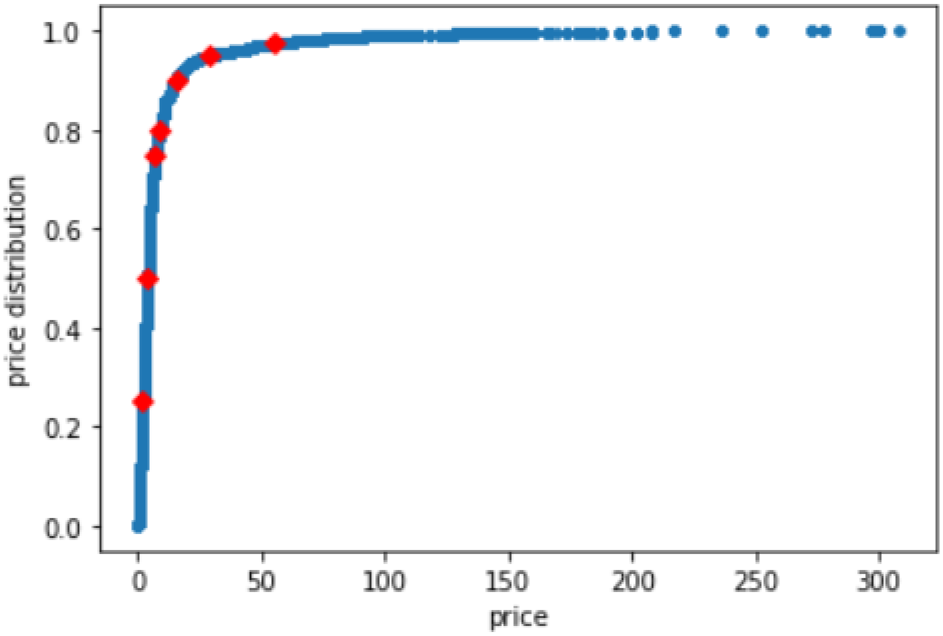
Goal: Find factors that influence buyers' purchase decision

Questions:

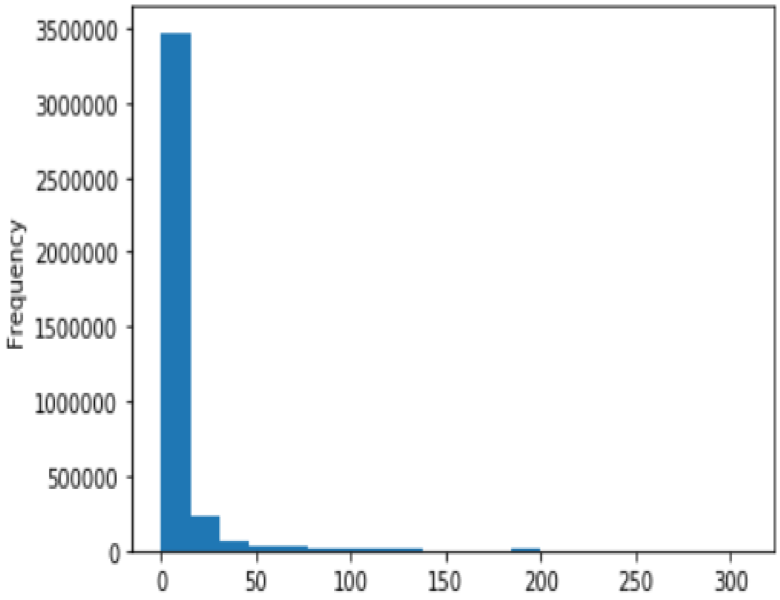
1. How does the price of a product impact buyers' decision to purchase?
2. How is sale distribution looking across different price ranges?
3. Are there certain categories of products that are being purchased more?
4. Are there certain brands of products that are being purchased more?
5. What are the top 10 products that are being purchased?
6. What are the top 5 categories of products that are being purchased?
7. What are the top 10 brands that are being purchased?
8. What are the top 10 products that are being removed from cart?
9. Find relationship between price of product and no. of views
10. What is the relationship between event time and purchase event?
    a. How is the hourly purchase trend?
    b. How is the daily purchase trend?
11. Compare Oct purchase trends with November
    a. How daily sales trends differ between Oct and Nov.?
    b. How weekly sales trends differ between Oct and Nov?
    c. How Thanksgiving affects November sales trends?

Data Insights

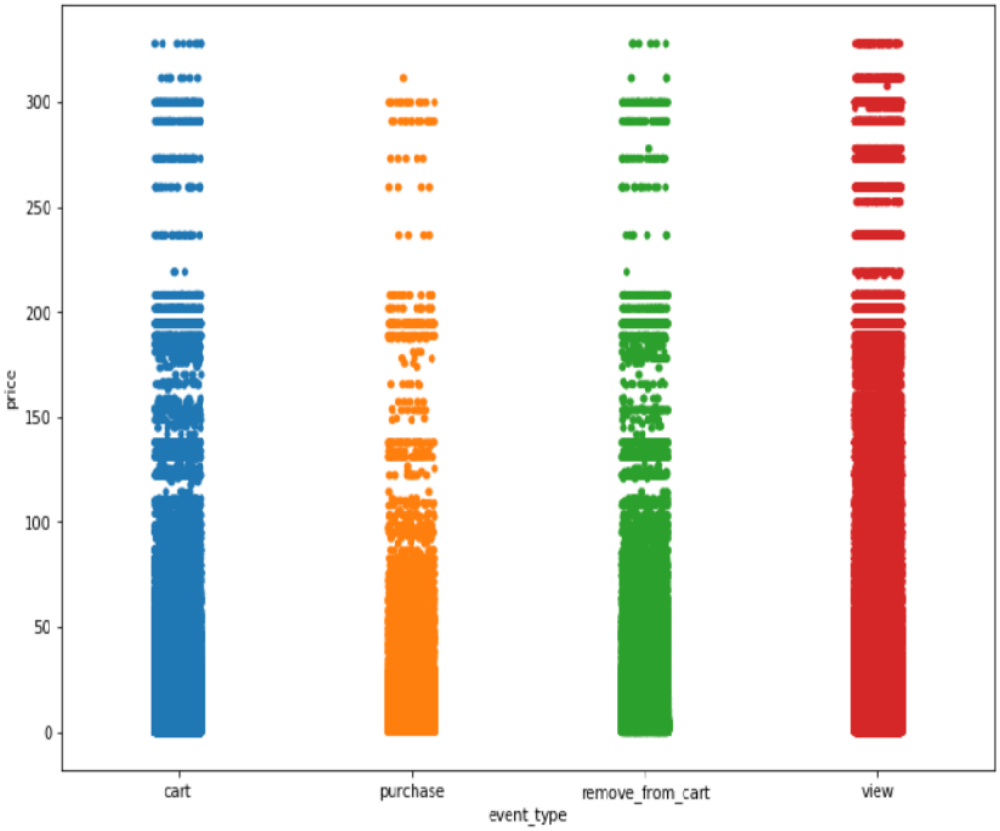
1. The dataset shows products ranging in price from 0 - 307.6$. However, plotting the frequency distribution and Empirical Cumulative Distribution function (ECDF) for 'price' across entire purchases made in Oct. and Nov. reveals that 97.5 percentile of purchases is being made for products whose price is less than or equal to 55$.



price distribution
1.0
0.8
0.6
0.4
0.2
0.0
0
50
100
150
200
250
300
price

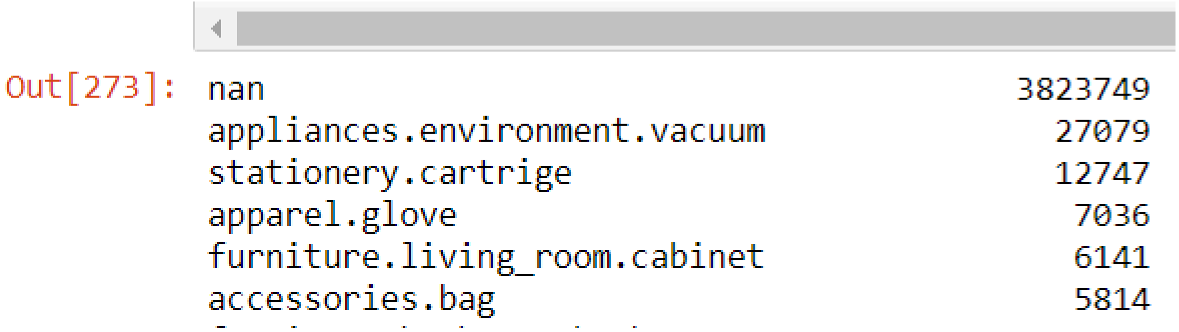


Frequency
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
0
50
100
150
200
250
300



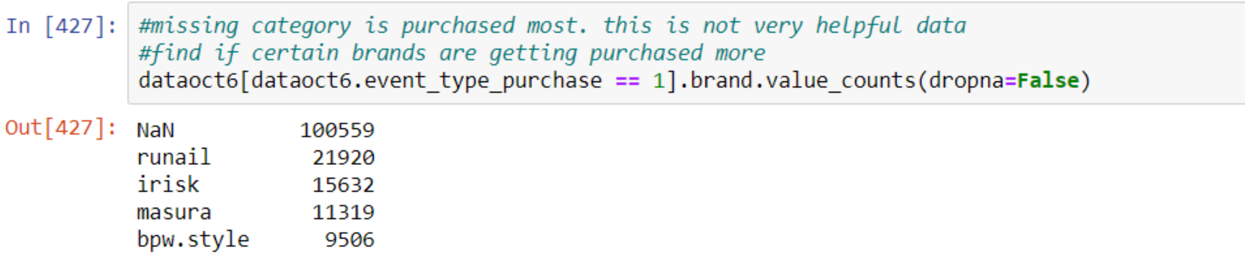
2. Products with price > 55$ make only 10% of the total sales in Oct. Products with price > 54$ make 9% of the sales in Nov. So, 90% of the sales is happening for products less than or equal to 55$

3. There are 3.8 millions of data records (out of approx. 4 million) with missing category code. So, the biggest category with maximum sales is the 'missing' category. This is not very helpful information. The second category that is purchased most is 'stationery.cartridge'.



```
Out[273]: nan                                     3823749
          appliances.environment.vacuum             27079
          stationery.cartrige                       12747
          apparel.glove                              7036
          furniture.living_room.cabinet              6141
          accessories.bag                            5814
```

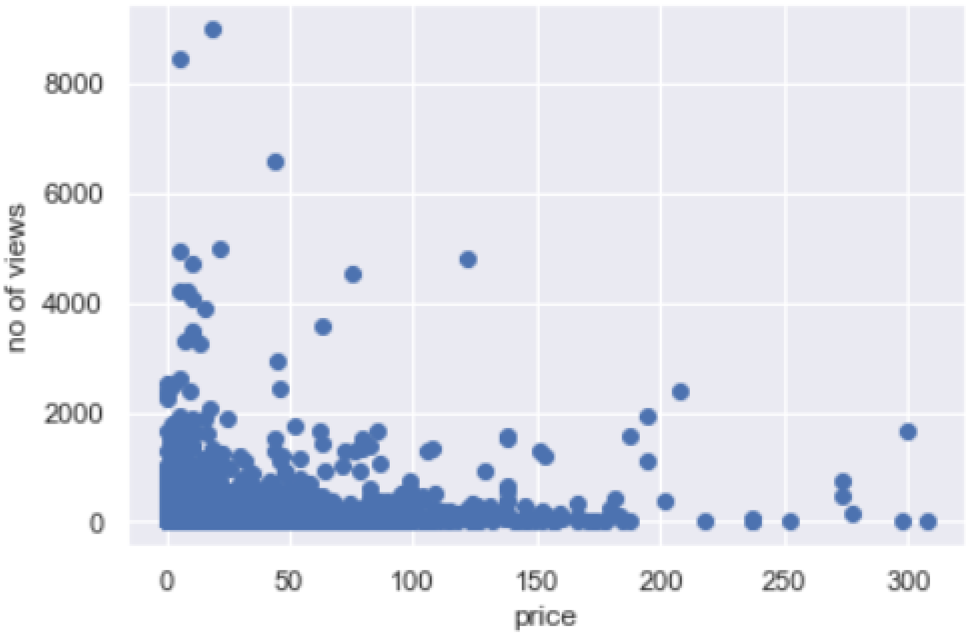
4. There are about 1 million records (out of approx. 4 million) with missing brand value. The top most brand that users purchase is the ‘missing’ brand. This is again not very helpful. The second brand is ‘runail’.



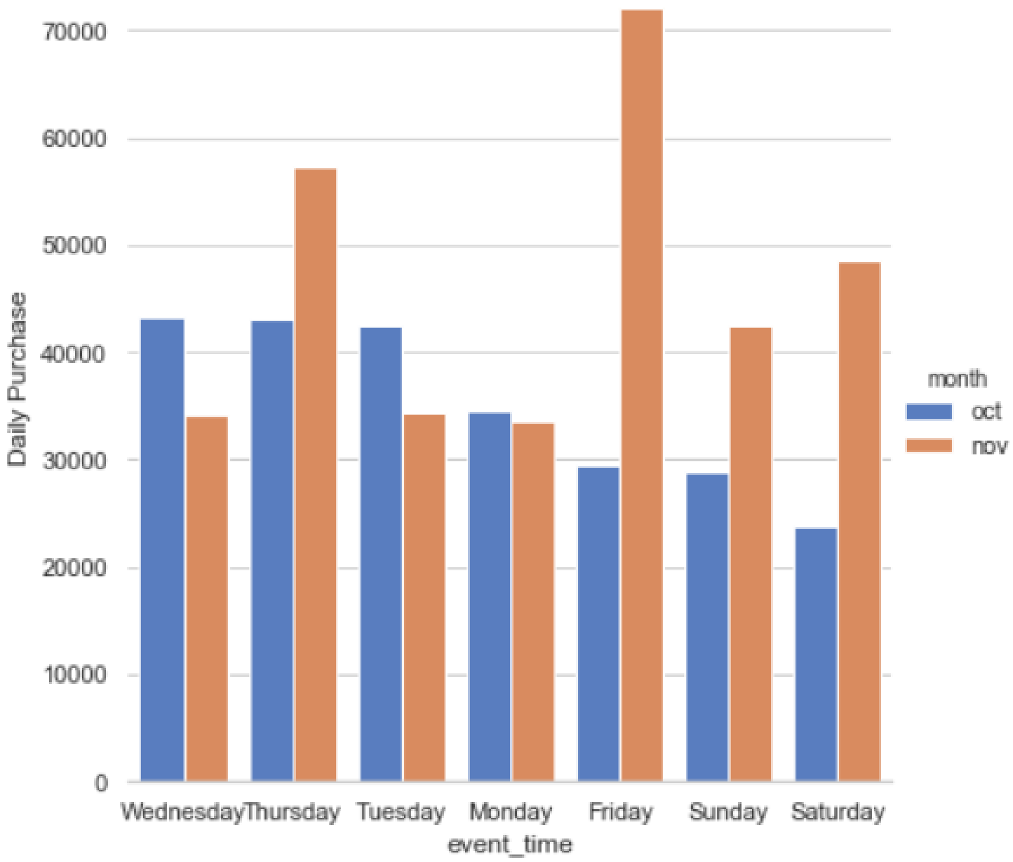
```
In [427]: #missing category is purchased most. this is not very helpful data
          #find if certain brands are getting purchased more
          dataoct6[dataoct6.event_type_purchase == 1].brand.value_counts(dropna=False)

Out[427]: NaN          100559
          runail        21920
          irisk         15632
          masura        11319
          bpw.style      9506
```

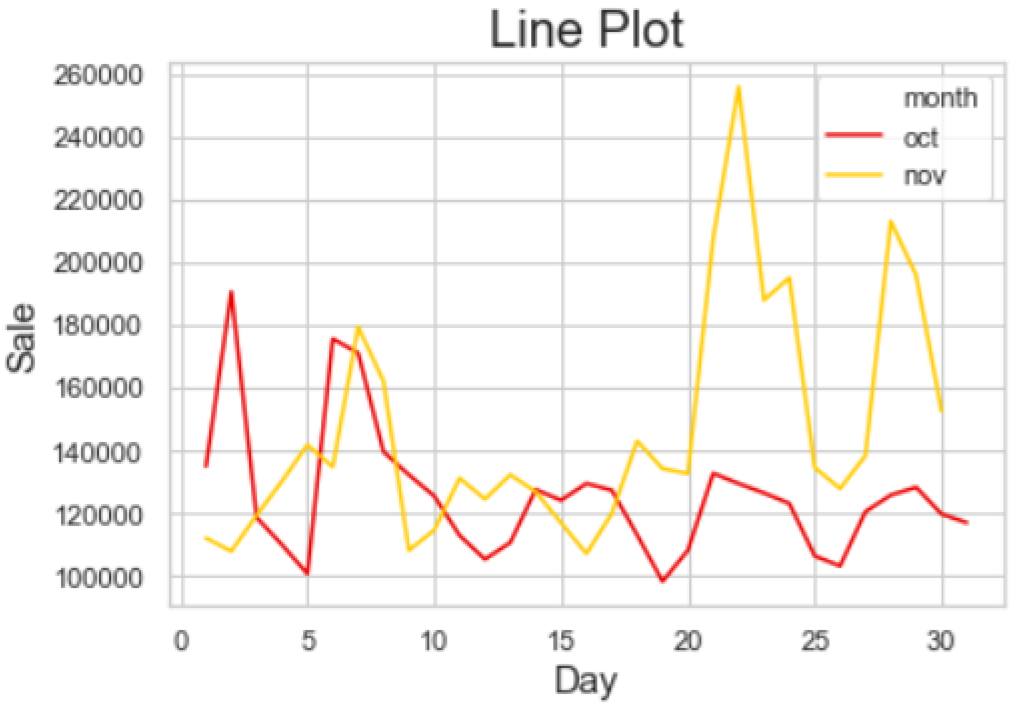
5. Product_id 5854897 is getting purchased the most in Oct. 5809910 is purchased most in Nov. 5700037 is the second most purchased product in Oct, while 5854897 is second in Nov.
6. 5809910 is the product that is getting removed from cart as well the most in Nov. 5809912 is the second product that is getting removed from cart the most in Nov.
7. 5700037 is the product that is getting removed from cart the most in Oct. It looks like the products that are getting purchased most are the ones that also get removed from cart the most.
8. Product with highest remove_from_cart to purchase ratio is TBD
9. The scatter plot between ‘price’ and ‘event_type_view’ shows that products with price < 50 are viewed more. Products in price range 50 and 190 have moderate views but price greater than 190 have handful views.



10. In Oct, maximum sales happen on Wednesday, followed by Thursday and Tuesday. November, however, Friday tops the sales due to Thanksgiving sale, followed by Thursday and Saturday.



11. October daily sale trend shows there is more sale happening in the first two weeks of the month. In Nov., however, the first 3 weeks are low on sale and peaks in the final week due to Thanksgiving.


Line Plot
Sale
260000
240000
220000
200000
180000
160000
140000
120000
100000
0
5
10
15
20
25
30
Day
month
oct
nov