E-Commerce Data Analysis

Report

BY

Mishaal Amin Hajiani

DATE: 23/09/2020

Prepared For XYZ E-Commerce Store

TABLE OF CONTENTS

INTRODUCTION

This is a transnational data set which contains all the transactions occurring between 01/01/2019 and 01/01/2020 for a US-based E-Commerce store.

DATA UNDERSTANDING

- **Data Collection Duration:** ~ 1 years (2019-2020)
- **Total Rows:** ~ 186851
- **Number of Features:** 6
- **Attributes:**
  - Order Number
  - Item Title
  - No of Items Ordered
  - Price per Items
  - Order Create Date
  - Delivery Address

IMPROVING DATA QUALITY THROUGH DATA WRANGLING

The problem in this phase was to identify all notable problems related to data wrangling (data cleaning) and solve them to improve the quality of the data. It is noteworthy that data wrangling and the consequent exploratory data analysis (EDA) comprise at least 60% of data science effort. The typical practice is to clean and then convert any categorical column to numerical because almost all data science algorithms operate only on numerical data. In the following, we list the notable data quality issues that were encountered and the methods used to mitigate these identified data inconsistencies to improve the accuracy of the underlying data to drive clinical business decision

Data Wrangling Steps

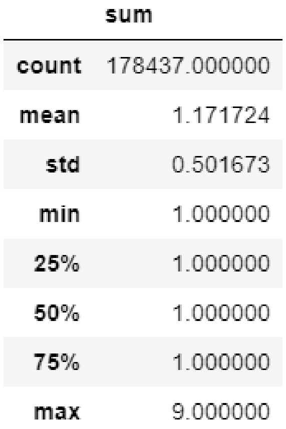
- Missing Value Analysis:
  - 545 blank rows were dropped
  - 355 headers were removed and all datasets were merged together.
- For better readability following columns were renamed:
  - 'Order Number': 'Order ID'
  - 'Item Title': 'Product',
  - 'No of Items Ordered': 'Quantity',
  - 'Quantity' : 'quantity',
  - 'Price per Item': 'Unit Price'
- Quantity column was converted into int, unit Price was converted into Float, Order date was treated as date.

- Extracted features from Order Date to get new features such as ***date, day, month, year, hour, day of week*** for further analysis.
- No Unit price was below zero which indicated that no product was given free of cost.
- ***Revenue Column*** was created using product of Unit price and Quantity.
- Cities were extracted from ***Delivery Address Column*** into a new column named as ***City***

EXPLORATORY DATA ANALYSIS (EDA)

The purpose of EDA is to understand the cleaned data, variables and the relationships between them, that helped in formulating hypotheses that could be useful when building predictive models. Here, we present some initial results.

Quantity Column Distribution



	sum
count	178437.000000
mean	1.171724
std	0.501673
min	1.000000
25%	1.000000
50%	1.000000
75%	1.000000
max	9.000000

Table 1:Deviation Column Distribution

Orders are placed in a range of [1, 9] quantity. Majority people (75%) ordered one product only.

Most bought Item:

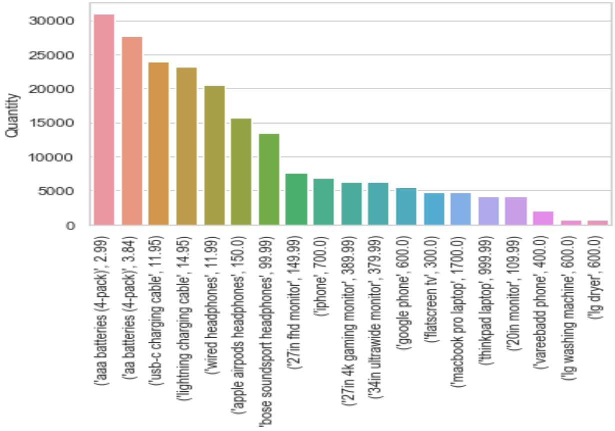
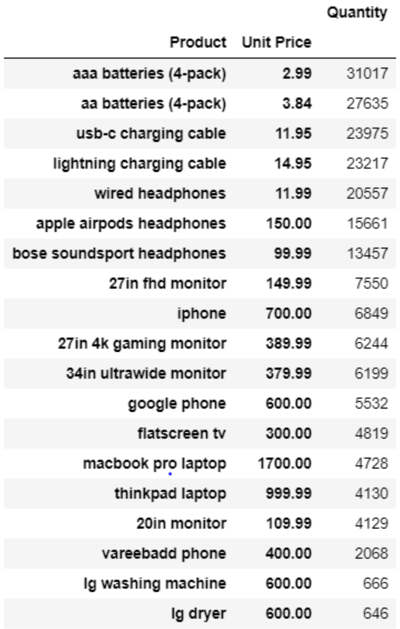


Figure 1:Most bought Product

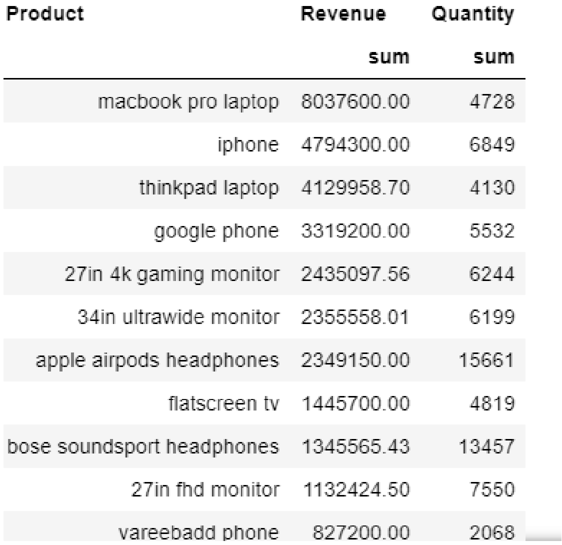


		Quantity
Product	Unit Price	
aaa batteries (4-pack)	2.99	31017
aa batteries (4-pack)	3.84	27635
usb-c charging cable	11.95	23975
lightning charging cable	14.95	23217
wired headphones	11.99	20557
apple airpods headphones	150.00	15661
bose soundsport headphones	99.99	13457
27in fhd monitor	149.99	7550
iphone	700.00	6849
27in 4k gaming monitor	389.99	6244
34in ultrawide monitor	379.99	6199
google phone	600.00	5532
flatscreen tv	300.00	4819
macbook pro laptop	1700.00	4728
thinkpad laptop	999.99	4130
20in monitor	109.99	4129
vareebadd phone	400.00	2068
lg washing machine	600.00	666
lg dryer	600.00	646

Table 2: Product Analysis Quantity wise Output

AAA batteries(4-pack) and AA batteries(4-pack) are best selling product at the price of 2.99 and 3.84 respectively. This somehow can explain as the unit price is low, people are able to afford more

Most Revenue Generated Product:



Product	Revenue	Quantity
	sum	sum
macbook pro laptop	8037600.00	4728
iphone	4794300.00	6849
thinkpad laptop	4129958.70	4130
google phone	3319200.00	5532
27in 4k gaming monitor	2435097.56	6244
34in ultrawide monitor	2355558.01	6199
apple airpods headphones	2349150.00	15661
flatscreen tv	1445700.00	4819
bose soundsport headphones	1345565.43	13457
27in fhd monitor	1132424.50	7550
vareebadd phone	827200.00	2068

Table 2: Product analysis Revenue wise

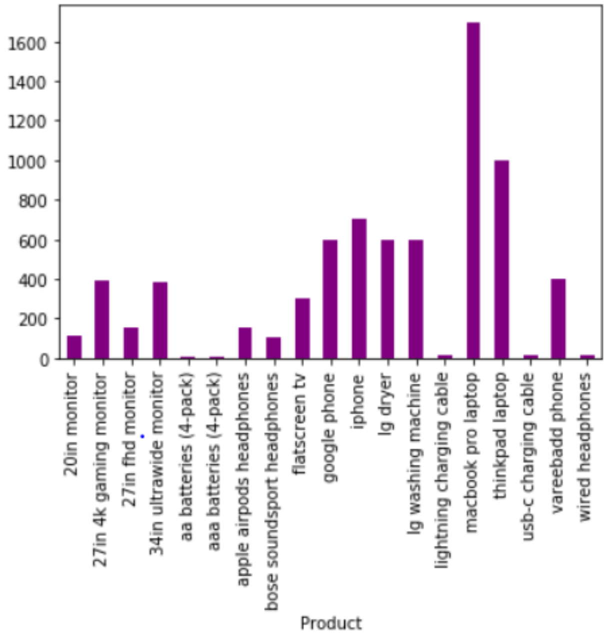
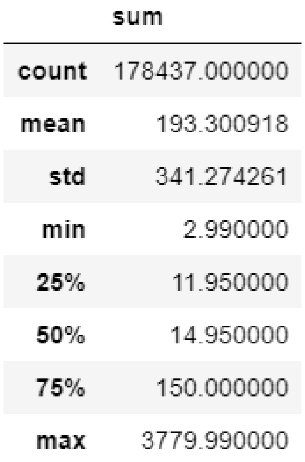


Figure 2: Price Per Item

Macbook pro laptop brings the most revenue i.e. $8037600 even though it is not the most sold product but its revenue is accumulated is the highest because the price of macbook pro is the highest as seen from the figure 4.



	sum
count	178437.000000
mean	193.300918
std	341.274261
min	2.990000
25%	11.950000
50%	14.950000
75%	150.000000
max	3779.990000

Table 4 : Transaction Statistics

75% of the people brings the revenue or spends between 2.99 to 150.

City Wise Analysis

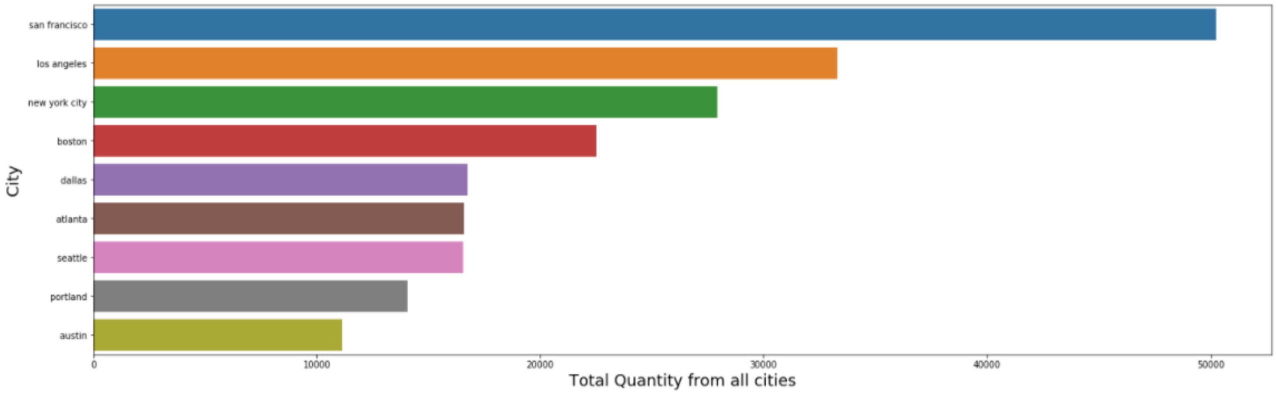


Figure 3: Total Quantity-city wise

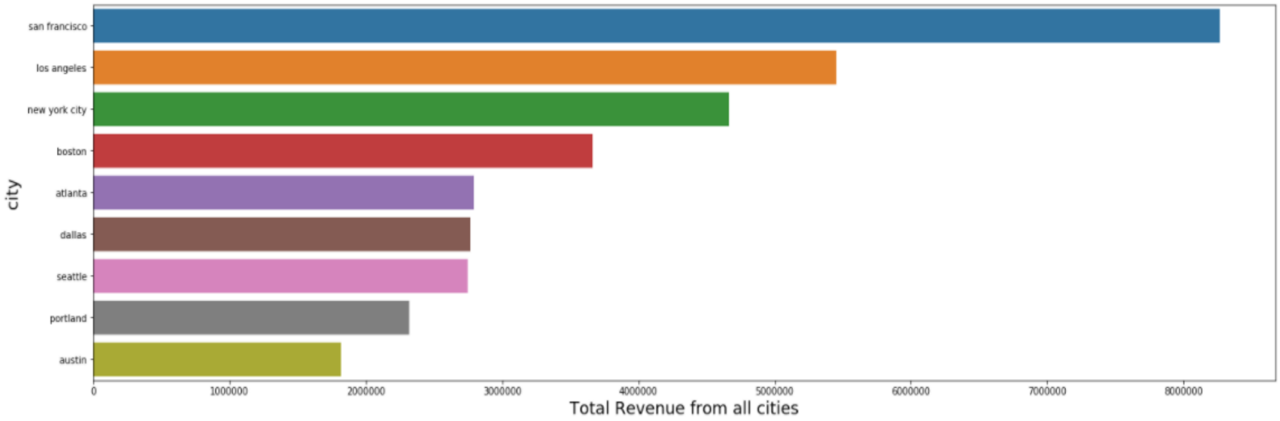


Figure 4: Total Revenue-city wise

San Francisco tops the table in terms of quantity ordered, orders placed and revenue generated.

Cities	Population	Cost of Living Index
New york	8,336,817	249
Los Angeles	3,979,576	203
Dallas	1,343,573	153
Austin	978,908	161
San Francisco	881,549	240
Seattle	753,675	194
Boston	692,600	205
Portland	654,741	185
Atlanta	546,811	158

Table 5: Deviation Count for Discharge Care Level

Above Table supports the results achieved, it is mainly because of the high CPI more people in San Francisco can bear the price of the products thus buys more.

Month Wise Analysis:

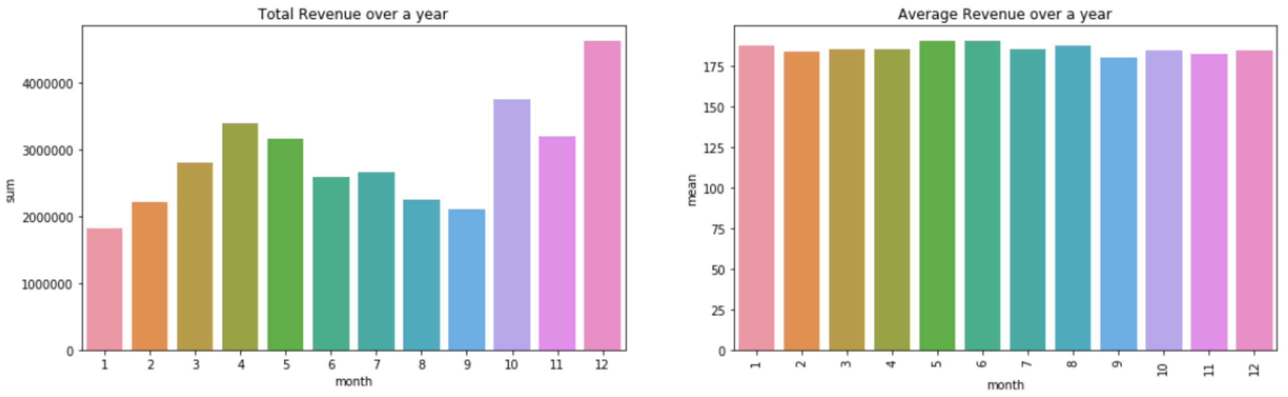
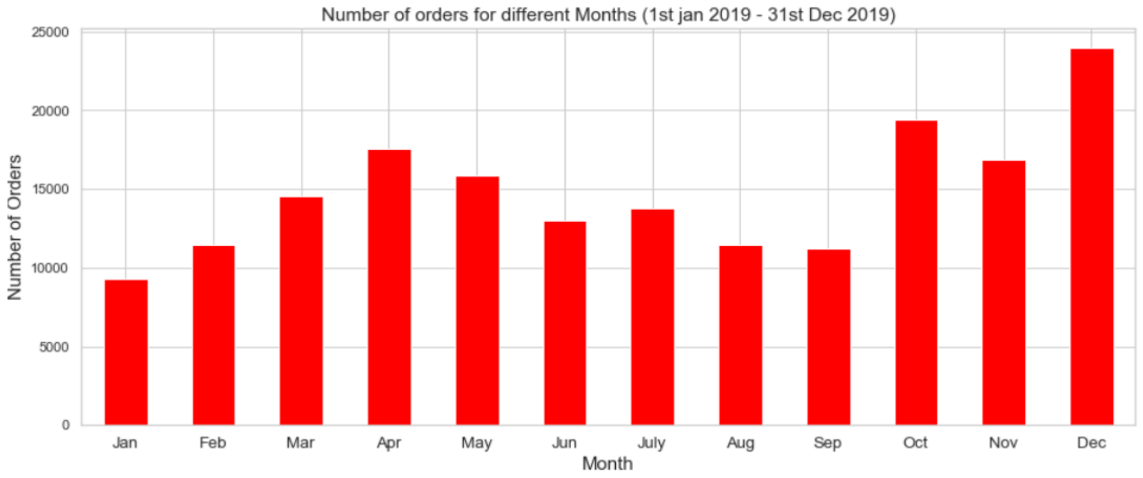


Figure 5: Month Wise revenue generated



Towards the end of the year, sales make a huge jump to over a million in the month of December with $4.6M this is mainly because Holidays and festivities like Christmas Eve, New year are observed. However, looking at average revenue diagram indicates nothing change drastically.

Time Wise Trend:

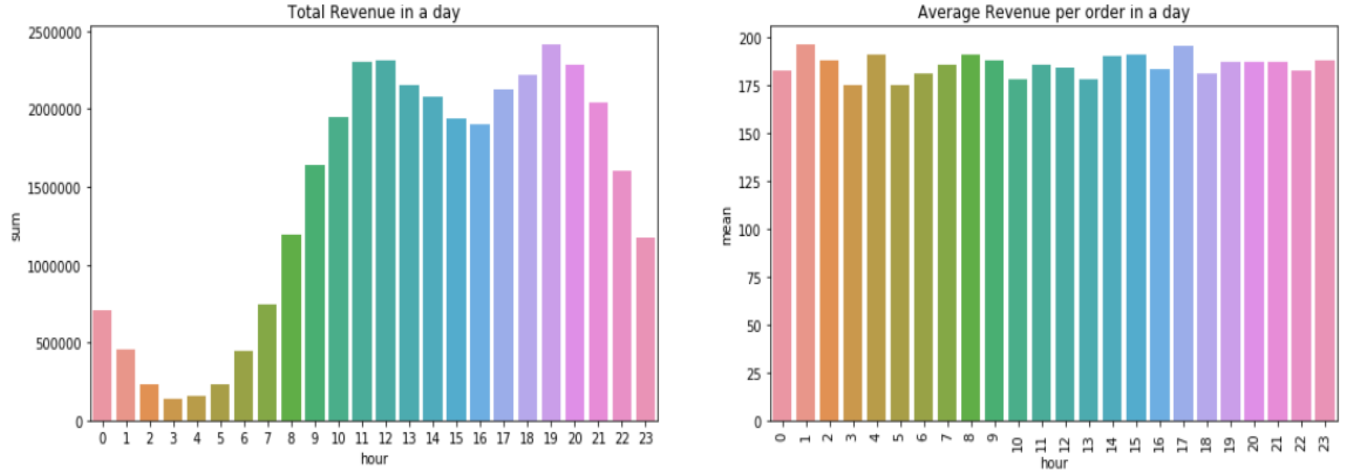


Figure 6: What time do people tend to buy our products?

Starting from 7 am, people tend to make purchase on the online store. Revenue hits the top at 7pm. Afterwards, sales gradually decrease till 10pm onwards. After that only a few of customers are left make to purchases. The second image, shows merely the same average revenue for all day. At 1 am the average is comparatively higher than the rest hours in a day. It suggests people make a huge quantity of items per transaction at this time as they got free from their work and now have ample of time to serve internet, Thus, we can add more advertisements and promotion at this hour of the day it can boost our sales.

Week Day Wise Trend:

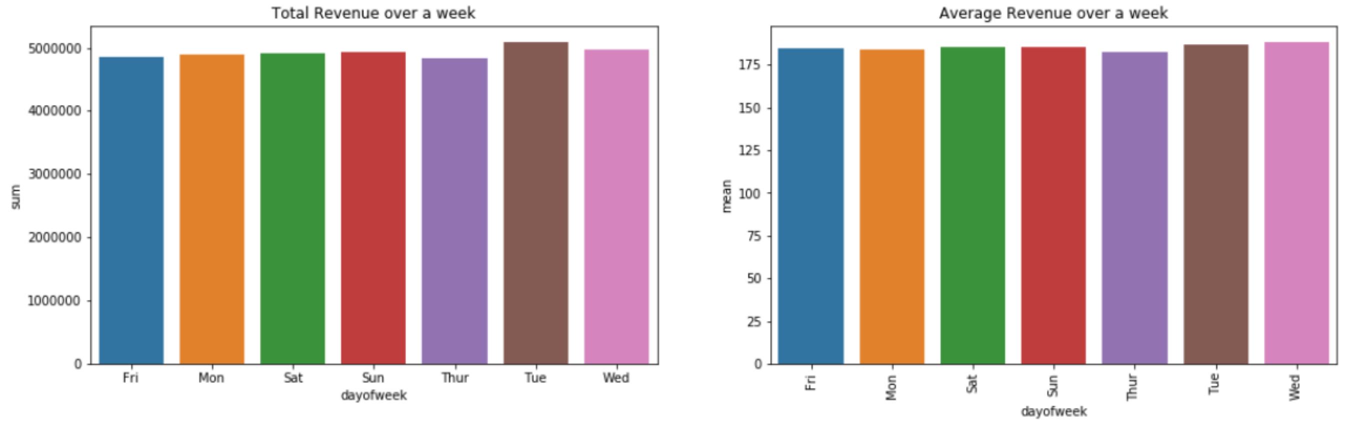
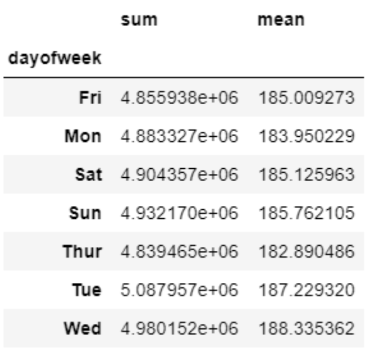


Figure 7: Week Day wise revenue Analysis



	sum	mean
dayofweek		
Fri	4.855938e+06	185.009273
Mon	4.883327e+06	183.950229
Sat	4.904357e+06	185.125963
Sun	4.932170e+06	185.762105
Thur	4.839465e+06	182.890486
Tue	5.087957e+06	187.229320
Wed	4.980152e+06	188.335362

Table 6: Day wise Output

Figure 7 and Table 6 shows that, the revenue is almost same, but in comparison Highest revenue was generated on Tuesday and Wednesday.

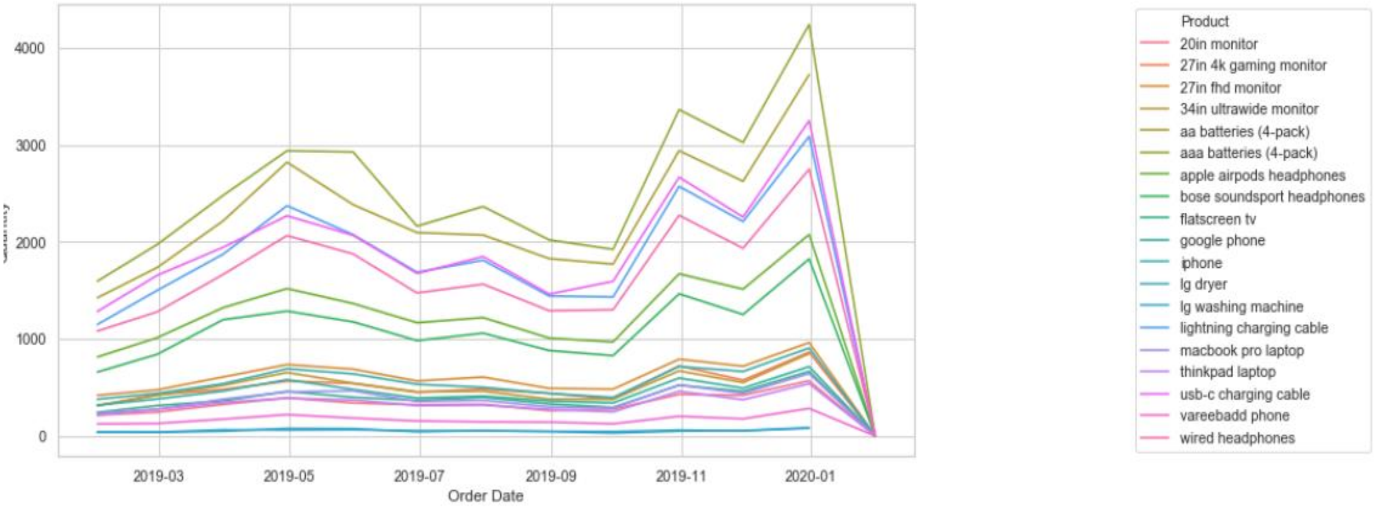


Figure 8: Product Trend line

Figure 8 is an other way to show the trend of product through out the year, this visual clearly show AAA batteries(4-pack) is most sold product also, this visual shows how towards the end of year sales increases and witness a downfall at the beginning of the year.

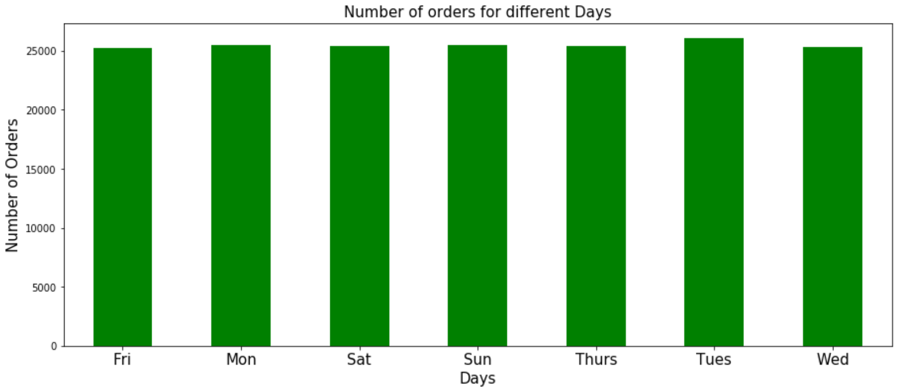
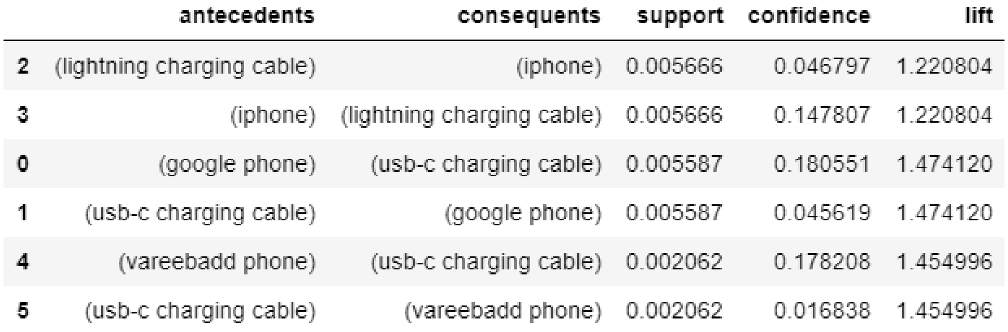


Figure 8: No. Orders placed Day wise

Over all there is no major difference in number of orders placed each day, but Comparison highest number of orders were placed on Tuesdays i.e. 26063 orders.

Market Basket Analysis:

Association rule mining is a technique to identify underlying relations between different items. This can helps us to improve the marketing strategies.



	antecedents	consequents	support	confidence	lift
2	(lightning charging cable)	(iphone)	0.005666	0.046797	1.220804
3	(iphone)	(lightning charging cable)	0.005666	0.147807	1.220804
0	(google phone)	(usb-c charging cable)	0.005587	0.180551	1.474120
1	(usb-c charging cable)	(google phone)	0.005587	0.045619	1.474120
4	(vareebadd phone)	(usb-c charging cable)	0.002062	0.178208	1.454996
5	(usb-c charging cable)	(vareebadd phone)	0.002062	0.016838	1.454996

Table 6:market basket analysis for all cities

Apriori limitations: A large number of possibilities in a single basket among an even larger number of baskets results in having a "sparse matrix", full of 0s which causes the support of basket occurrences to drop drastically. Like in the case above majority of the people purchased one product. The output achieved has its top support of 0.051 (5%). Lift basically tells us that the likelihood of buying a Lightning charging cable and iPhone together is 1.22 times more than the likelihood of just buying the lightning cable. This relation makes sense as a phone requires a charging cable. A Lift of 1 means there is no association between products A and B. Lift of greater than 1 means products A and B are more likely to be bought together. Finally, Lift of less than 1 refers to the case where two products are unlikely to be bought together. So, we can recommend the users while they are buying or searching for phone to make it easier for them avoid wasting time searching for a charger, later. We can also offer package of both, which can attract more users and help them improve their shopping experience and lastly we can introduce some discounts on it. This way sales of both items will increase.

MODELING

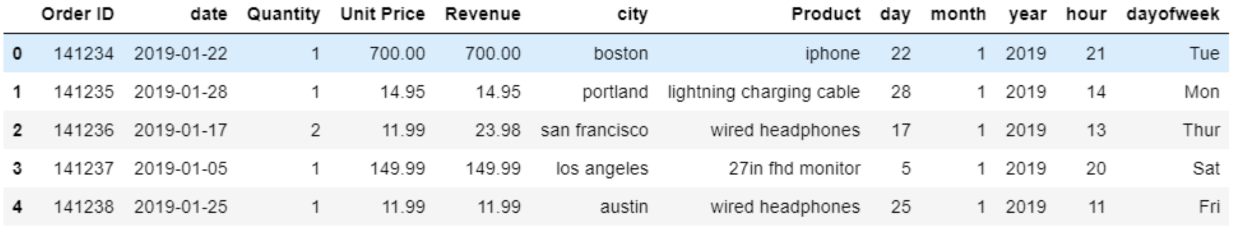
Goal here is to predict which features have most influence on Revenue to boost up marketing campaigns.

Possible features we thought will be important to predict Revenue:

- Different seasons
- Public Holiday in US
- Promotions
- Weekend
- School Holiday
- StartOfMonth_Initial10Days
- MidOfMonth
- EndOFMonth_last10Days
- Last 5 days/ Last 3 days Revenue (if dealing with total sales of the day)
- Lag sales((if dealing with total sales of the day)
- Month
- Quarter
- Days of Week
- First six month/last six month?

In Order to apply any regression model on the given dataset, following changes were made:

- New dataframe was created on the basis of unique Order ID to get the total revenue generate by each placed order.



	Order ID	date	Quantity	Unit Price	Revenue	city	Product	day	month	year	hour	dayofweek
0	141234	2019-01-22	1	700.00	700.00	boston	iphone	22	1	2019	21	Tue
1	141235	2019-01-28	1	14.95	14.95	portland	lightning charging cable	28	1	2019	14	Mon
2	141236	2019-01-17	2	11.99	23.98	san francisco	wired headphones	17	1	2019	13	Thur
3	141237	2019-01-05	1	149.99	149.99	los angeles	27in fhd monitor	5	1	2019	20	Sat
4	141238	2019-01-25	1	11.99	11.99	austin	wired headphones	25	1	2019	11	Fri

Figure 9: Final look of the data

On the dataframe showed in Figure 9, few changes were done:

- Order ID, date, Quantity, year, day columns were dropped from the dataset. Date was deleted after extracting important features from it, Order ID was just a unique no. representing each transaction and thus was meaning less for the data now, Quantity variable had skewed values as majority of people ordered one item and very few people purchased more than 3 items as shown below in Figure 10 making data imbalance. Lastly, year column was dropped as the dataset used belongs to 2019, redundant value.

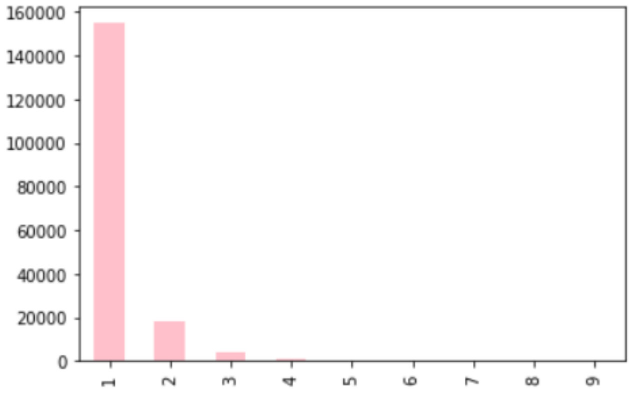
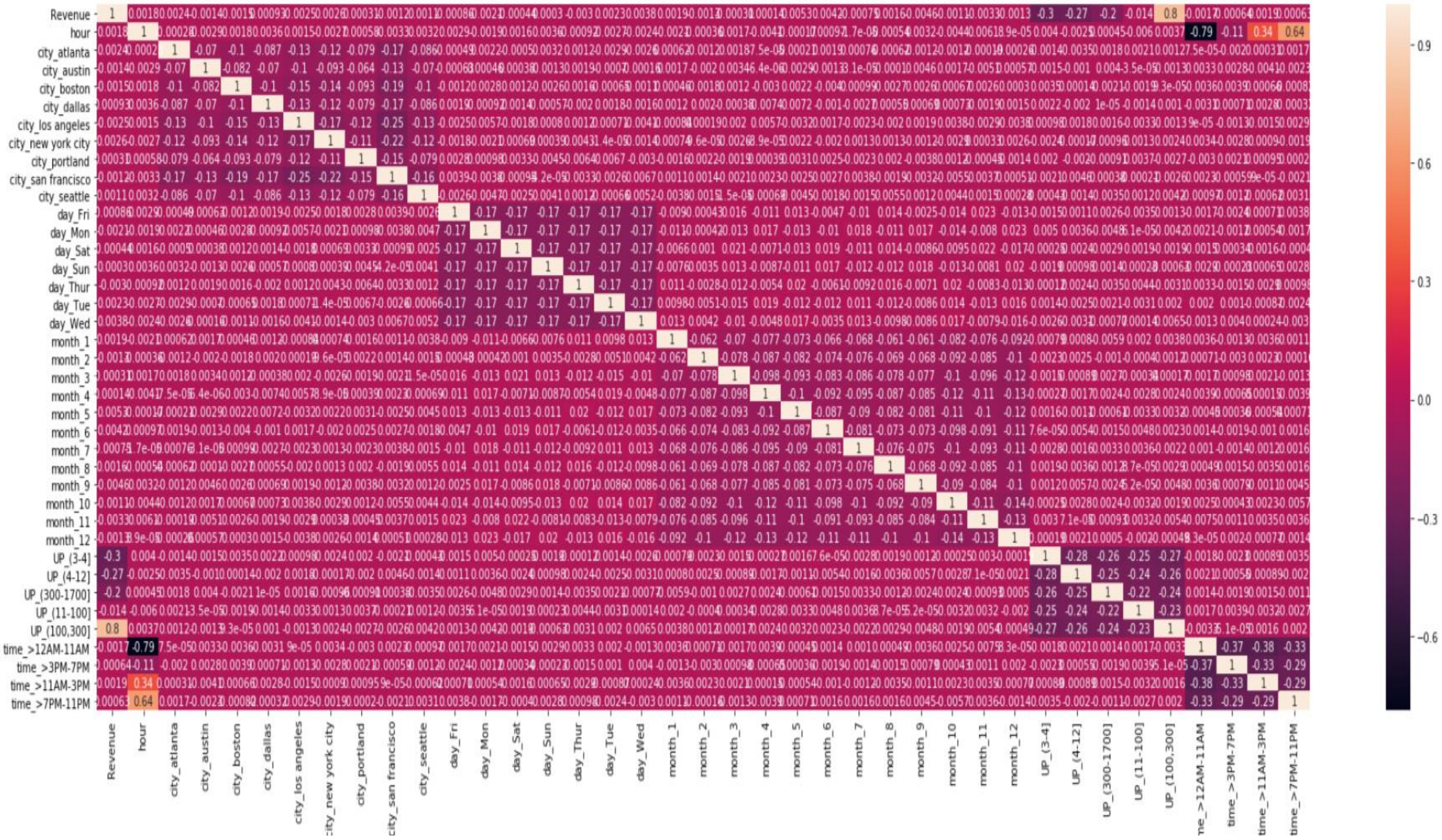


Figure 10: Quantity Purchased per Order

- Dummy variables were created for city, month, dayofweek and product. These there columns were treated as categorical columns. If we are using year 2020 data as testing data, in that case we have to remove the month variable as year 2020 data only consists month of January data.
- Discretization was done for hour and Unit price, here mean unit price is used. This unit price tells us about average spending.
  Binning For hour: ['>12AM-11AM','>3PM-7PM','>11AM-3PM','>7PM-11PM']
  Binning For Unit Price: ['(3-4]','(4-12]','(300-1700]','(11-100]','(100,300]']

Correlation Analysis

Correlation is an important but weaker alternative to determine the relevant features for predicting the target variable. Below is the correlation matrix of the well-known Pearson correlation coefficient, measuring the strength of the association between the two variables. Darker color i.e. r=1 defines positively strong correlation while dark red colour i.e. r=-1 indicates the negatively strong correlation. There is some amount of correlation in the data but, correlation across these X variables are not highly correlated i.e close to 1 or -1. If the measure was let say greater than 90%, we should worry about multicollinearity.



Data set splitting

Training Set = 60%

Cross Validation Set = 20%

Testing Set = 20%

PERFORMANCE ANALYSIS OF REGRESSION ALGORITHMS

- All Regression algorithms will be applied and their respective performances on training, cross validation and testing set will be analyzed.
- Applied Feature engineering: features will be added iteratively. We’ll check the correlation of each feature using the ANOVA table ,R^2 value, Standard Errors and P-values.