CSE 587 Project Phase 2

Market Basket Analysis Recommendation Engine

Vivek Singh (vsingh28 | 50418786)

Sougato Bagchi (sougatob | 50411918)

Background

The motive behind our project is to recommend customers with items which are often bought together by other customers. A good recommendation becomes a win-win situation for both the customers as well as the retailers. For the retailers this results in higher revenue growth and results in customer satisfaction. From the perspective of the customers, it becomes easy for them if they are recommended with similar or related items w.r.t which they are buying. This saves their time on product research and often results in finding relevant items very easily. This ultimately leaves the customers with a good impression on the retailer buying system.

Previously in the phase 1 we have made our EDA on the dataset and shared our graphs so that it becomes easy to visualize and understand the dataset.

But to create good recommendations we need to understand our data and find appropriate patterns/co-relations. For this objective we have implemented some of the methods of which shall be discussed in the next section.

Implemented Models

Our implementations include methods like: -

- **RFM Analysis**: segment the customers based on Recency, Frequency and montary value
- **K-Means Clustering**: based on customer segmentation
- **Linear regression**: to predict the sales
- **Tree model**: to predict the sales
- **Associate Rule Mining**: to recommend the products

For all the implementations we have considered purchases from only one country i.e., UK as most of the data is from the UK with very few from other countries (shown in Project Phase 1).

1. RFM Analysis

As we are dealing with customers and their purchase data, so we have decided that we segment our customers based on 3 criteria which are: -

- Recency
- Frequency
- Monetary Value

We then mark the RFM score for each customer and then create clusters based on the RFM score. In this way we can segregate the valuable customers.

Visualizing our data: -

- ### Revenue per user

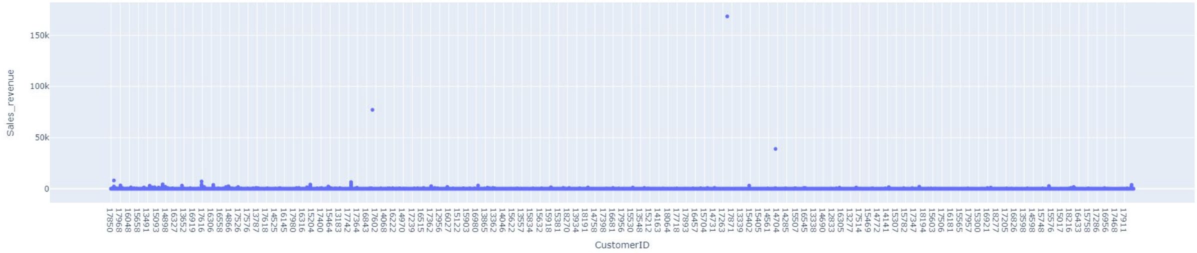


Fig 1. There are very few outliers in terms of revenue generated per user.

- ### Recency for some of the customers (Recency Analysis)

  Deals with how recent was the customer's last purchase? Customers who recently made a purchase will still have the product on their mind and are more likely to purchase or use the product again. Recency is often measured in days.

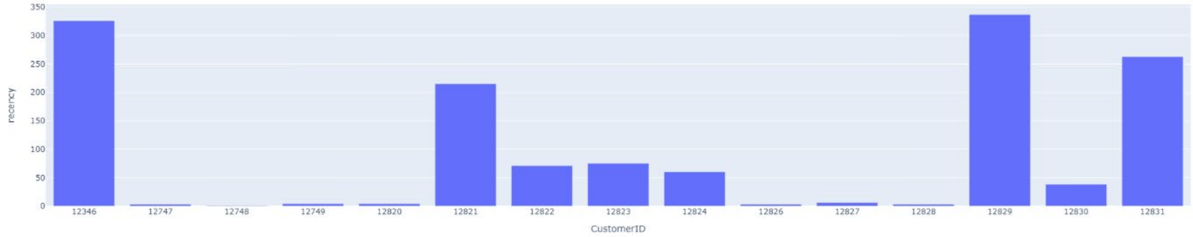


Fig 2. Recency for some of the customers.

- ### Frequency for some of the customers (Frequency Analysis)

  Deals with how often a customer makes a purchase in a given period. Customers who have made their purchase once are often more likely to do it again. Additionally, first time customers may be good targets for follow-up advertising by the retailers to convert them into more frequent customers.

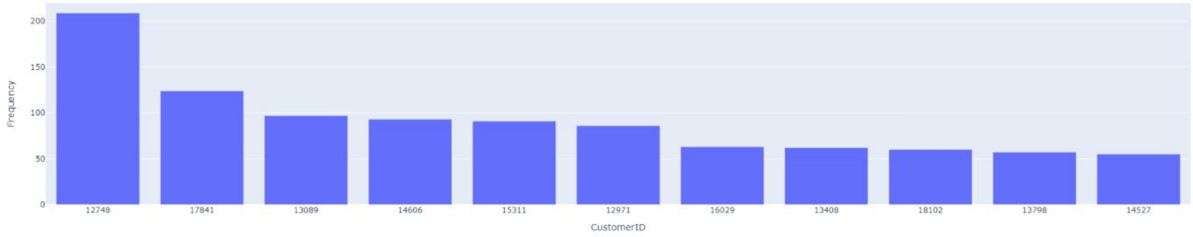


Fig 3. Frequency of purchase for some of the customers.

- Income generated for some of the customers (Monetary Analysis)
  Deals with how much money did the customer spend in a given period. Customers who spend a lot of money are more likely to spend money in the future and have a high value to a business. These customers are typically the one retails should give higher priority in retaining.

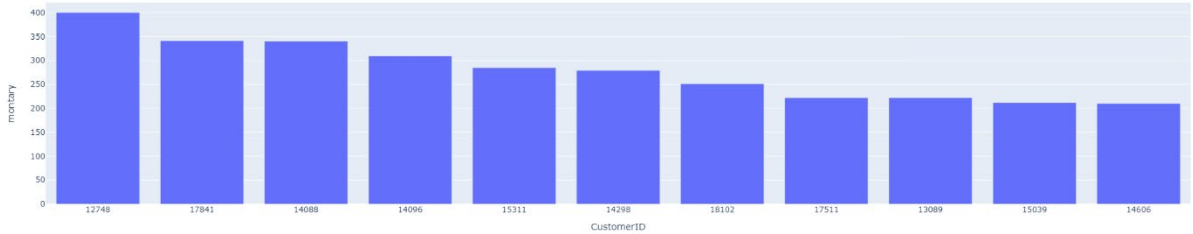
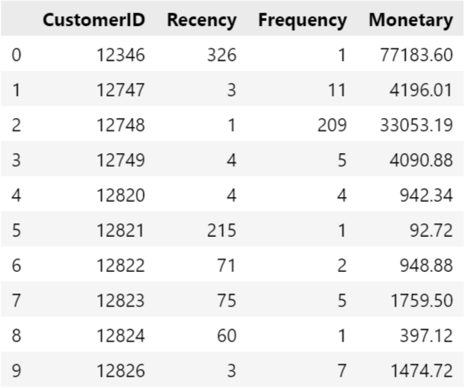


Fig 4. Amount spent by some of the customers in a given period.

RFM Table

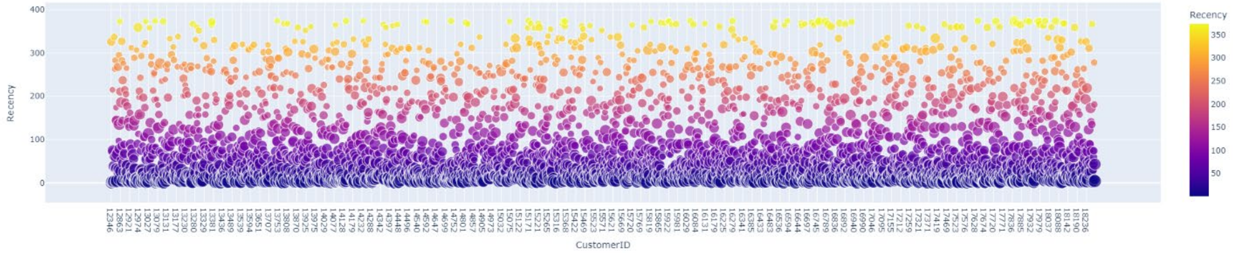


	CustomerID	Recency	Frequency	Monetary
0	12346	326	1	77183.60
1	12747	3	11	4196.01
2	12748	1	209	33053.19
3	12749	4	5	4090.88
4	12820	4	4	942.34
5	12821	215	1	92.72
6	12822	71	2	948.88
7	12823	75	5	1759.50
8	12824	60	1	397.12
9	12826	3	7	1474.72

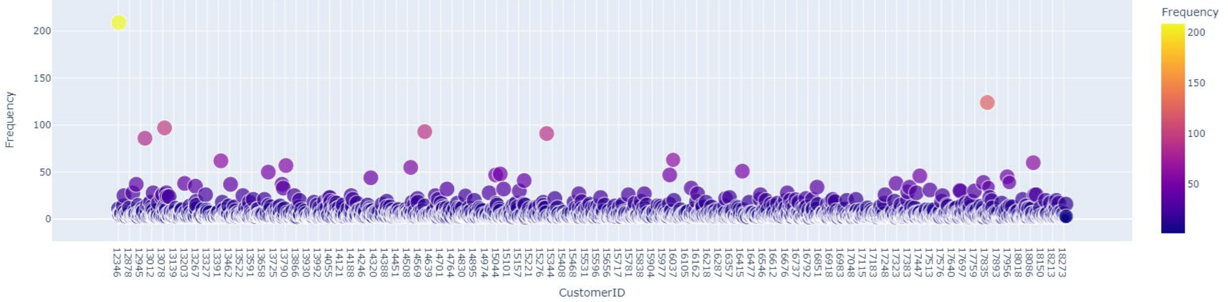
Fig 5. This is what looks like after we calculate RFM for each of the customers.

RFM score Plots

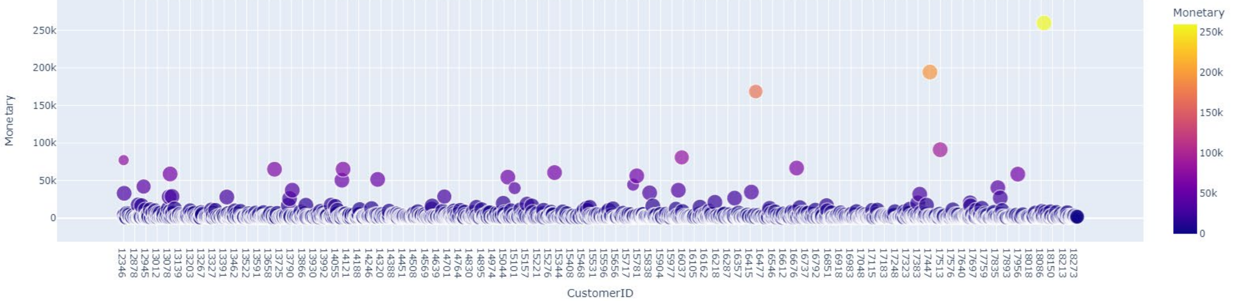
We have assigned scores for each of the customers w.r.t 3 factors (Recency, Frequency & Monetary values).



1. (Fig 6.) RFM w.r.t Recency. Here typically more the recency value, less the RFM score. Higher value of recency means that the person has made the purchase of a particular object long back.



2. (Fig 7.) RFM w.r.t Frequency. More often a customer makes a purchase, more they are valuable, so we assign them a higher RFM score.

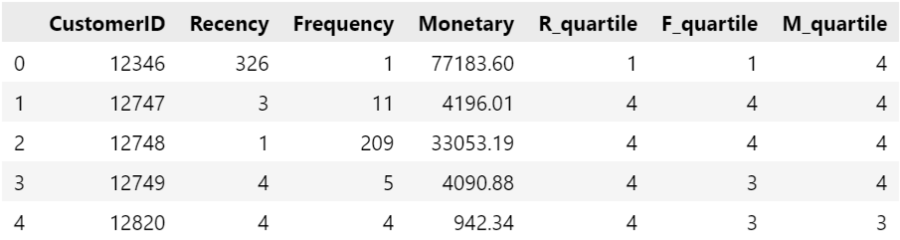


3. (Fig 8.) RFM w.r.t Monetary value. The higher the value of the purchase made by the customer, more they turn valuable to the retailers, so we assign them a higher RFM score.

Clusters in terms of Quartiles

Here each quarter signifies the avg value of Recency/Frequency/Monetary values. So now that we have got our clusters based on RFM, we need to demonstrate if a customer's RFM value sits in any of the quarter what does that really mean.

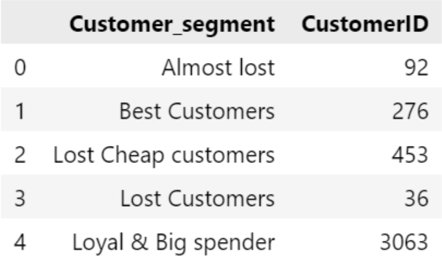
- recency value:
  - high value = bad/low valued customers (quartile 4)
  - low value = good/high valued customers (quartile 1)
- Frequency value:
  - high value = good/high valued customers (quartile 1)
  - low value = bad/low valued customers (quartile 4)
- recency value:
  - high value = good/high valued customers (quartile 1)
  - low value = bad/low valued customers (quartile 4)



	CustomerID	Recency	Frequency	Monetary	R_quartile	F_quartile	M_quartile
0	12346	326	1	77183.60	1	1	4
1	12747	3	11	4196.01	4	4	4
2	12748	1	209	33053.19	4	4	4
3	12749	4	5	4090.88	4	3	4
4	12820	4	4	942.34	4	3	3

Fig 9. Example of some of the customers & their RFM values divided into quartiles.

Now we have created meaningful clusters for the customers as per Market Basket Analysis, and which can provide valuable insight for the retailers.



	Customer_segment	CustomerID
0	Almost lost	92
1	Best Customers	276
2	Lost Cheap customers	453
3	Lost Customers	36
4	Loyal & Big spender	3063

Fig 10. These are the 5 meaningful clusters which we though would be useful with also the number of customers (CustomerID) in each of the clusters.

Segment	RFM	Description	Marketing
Almost Lost	311	Haven’t purchased for some time, but purchased frequently and spend the most	Aggressive price incentives
Best Customers	111	Bought most recently and most often, and spend the most	No price incentives, new products, and loyalty programs
Lost Cheap Customers	444	Last purchased long ago, purchased few, and spent little	Don’t spend too much trying to re-acquire
Lost Customers	411	Haven’t purchased for some time, but purchased frequently and spend the most	Aggressive price incentives
Loyal Customers & Big spenders	X1X, XX1	Buy most frequently, Spend the most	Market your most expensive products

Table 1. Key RFM segments. Here Best customer (1-1-1) means (R=1, F=1, M=1), and (X-1-X) means it only depends on the F value of 1. (Source)

The name of these clusters pretty much is self-explanatory, that these are based on how much revenue a retailer can generate from customers of each cluster. For example, cluster with no. 4 (i.e., Loyal & Big Spender) means that these people spend a decent amount of money per purchase, and they also purchase regularly and also, they must have made their last purchase very soon. These customers are supposed to be given highest priority.

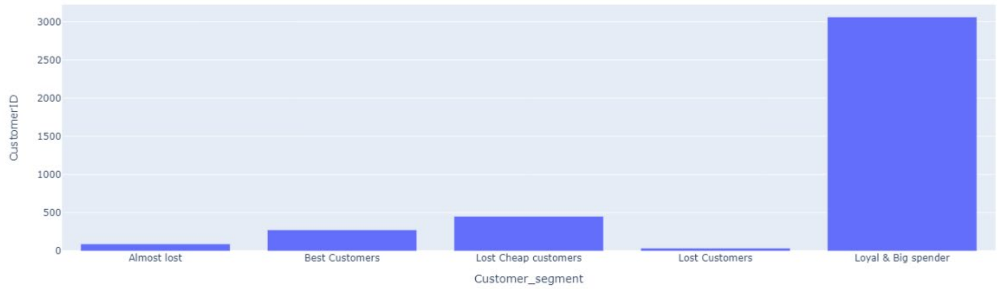


Fig 11. Now for better to visualizing this cluster info we have used this bar graph.

2. K-Means Clustering

We have also implemented K-Means on RFM data to create clusters. Here the number of clusters is selected from our Elbow Curve.

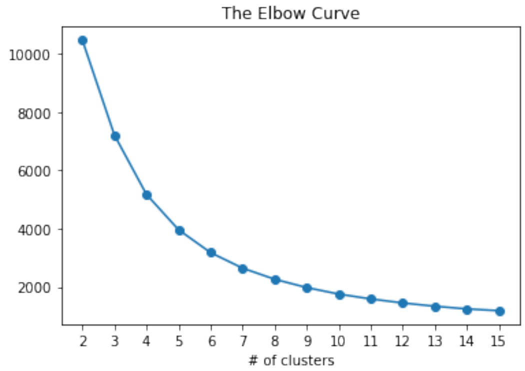
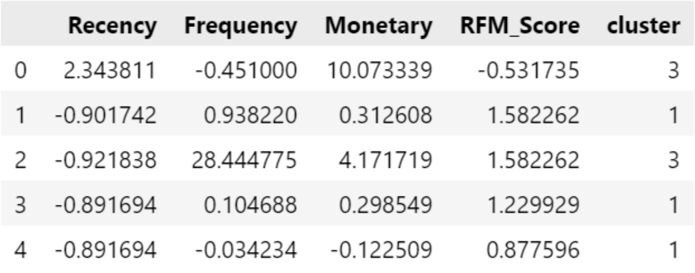


Fig 12. Here we can see that our elbow curve suggests us that pretty much after 5 no. of clusters our curve more-or-less decreases linearly. So, we have decided to use 5 clusters for our K-Means.

We have scaled the RFM values and also normalized the RFM score, so after the data processing our data looks something like the one shared in Fig 13.



	Recency	Frequency	Monetary	RFM_Score	cluster
0	2.343811	-0.451000	10.073339	-0.531735	3
1	-0.901742	0.938220	0.312608	1.582262	1
2	-0.921838	28.444775	4.171719	1.582262	3
3	-0.891694	0.104688	0.298549	1.229929	1
4	-0.891694	-0.034234	-0.122509	0.877596	1

Fig 13. Here the RFM score the mean value is 0 with standard deviation of 1.5. Due to the change of scale, we have some of the negative values in RFM too.

Clusters based on RFM scores with no. of clusters = 5

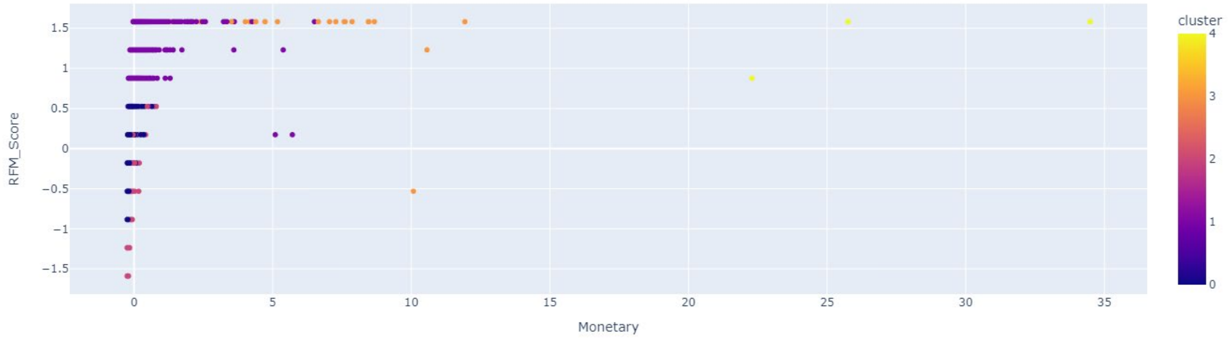


Fig 14. RFM score w.r.t Monetary value

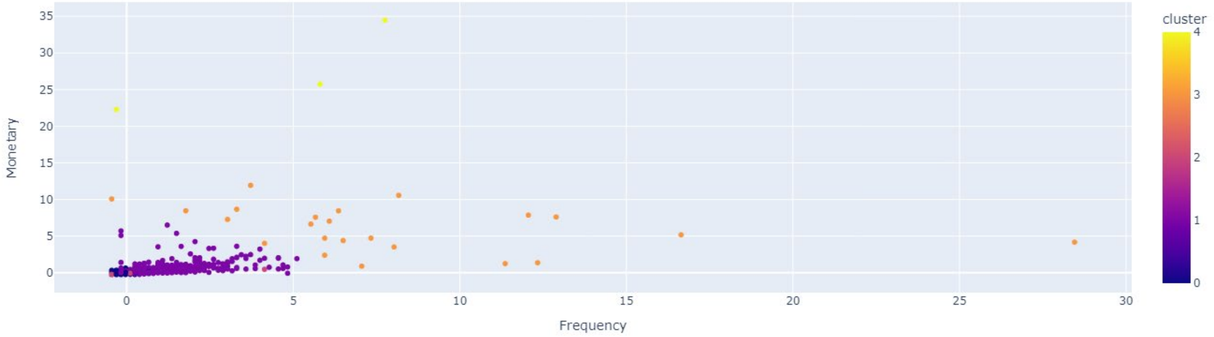


Fig 15. RFM score w.r.t Frequency

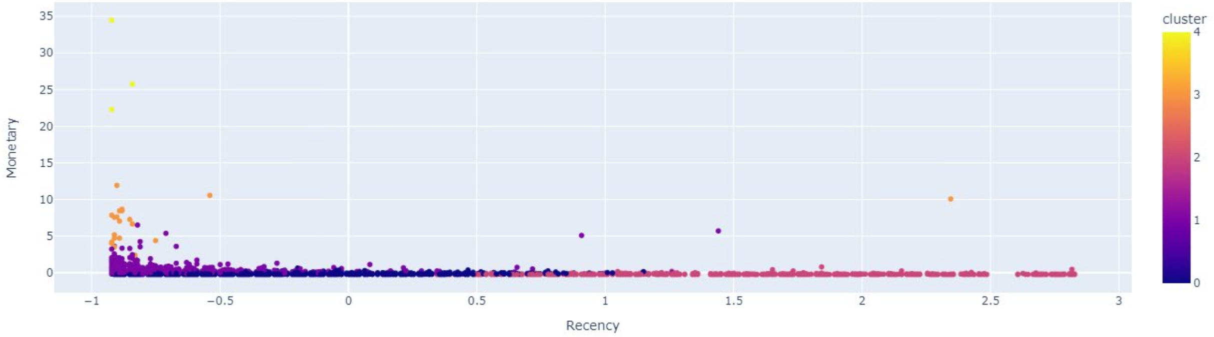


Fig 16. RFM score w.r.t Recency

3. Linear Regression

Here we are trying to predict the Sales revenue would be generate on upcoming month or on unseen data.

1. Check for trend and seasonality:
    a. *On daily basis –*

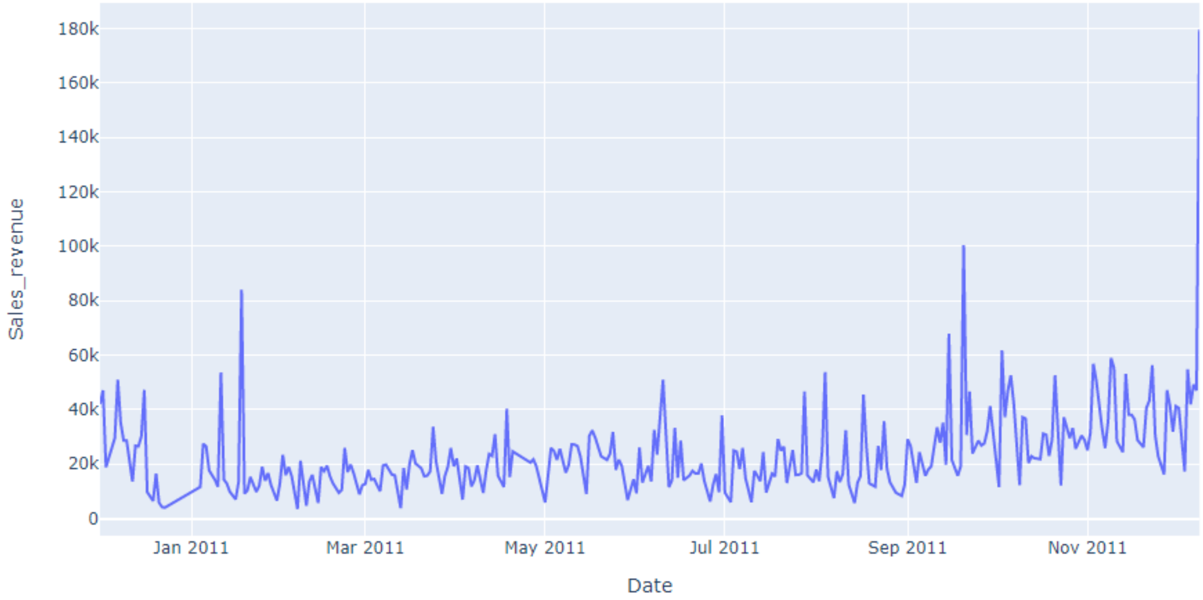


Fig 17. Trend and seasonality on daily basis

    b. *Monthly basis –*

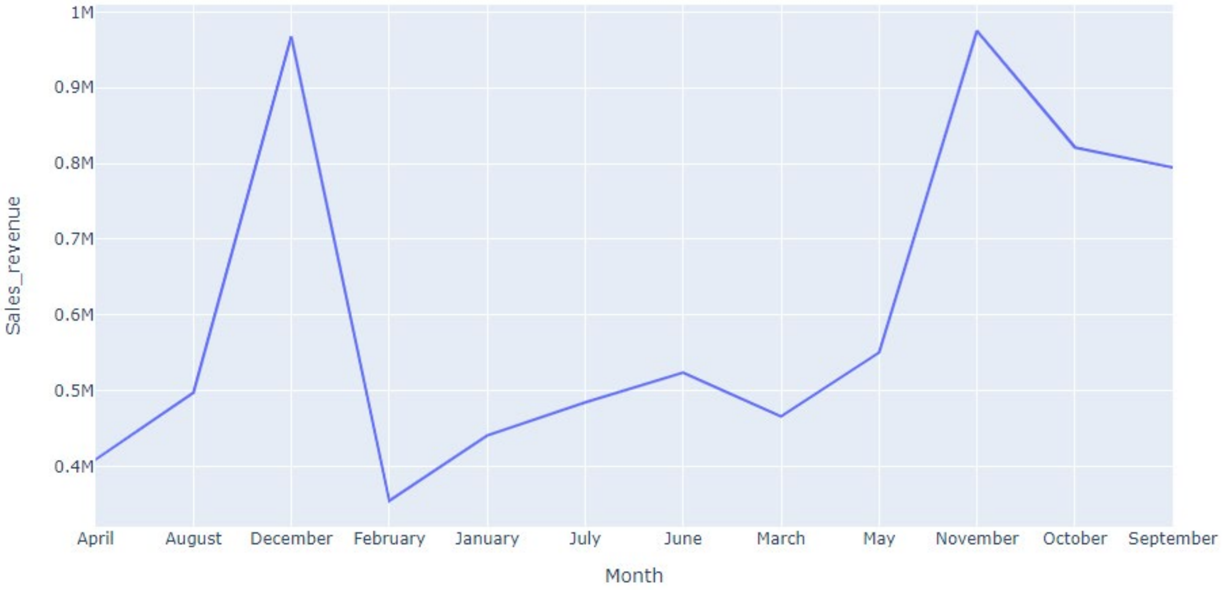


Fig 18. Trend and seasonality on a monthly basis

Observations: Since the we cannot draw a clear picture for trend and seasonal pattern in data. We have used statistic methods like rolling mean and rolling standard deviation to understand it better.

2. Statistical method:

   a. Rolling mean and standard deviation:

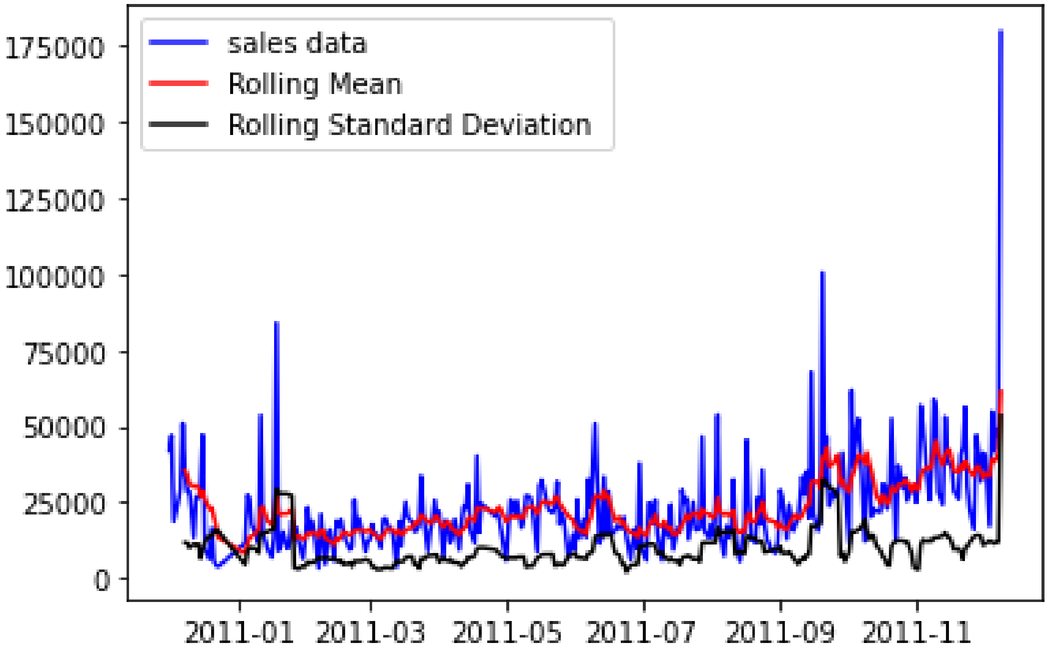
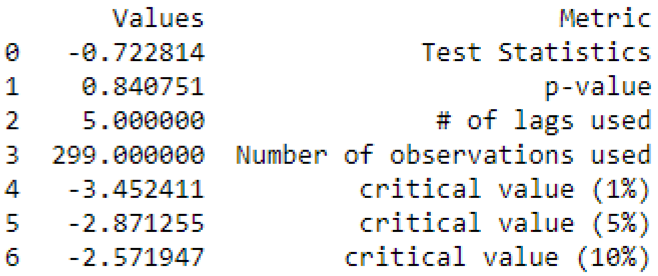


Fig 19. Trend and seasonality on a monthly basis

   b. AdFuller test:

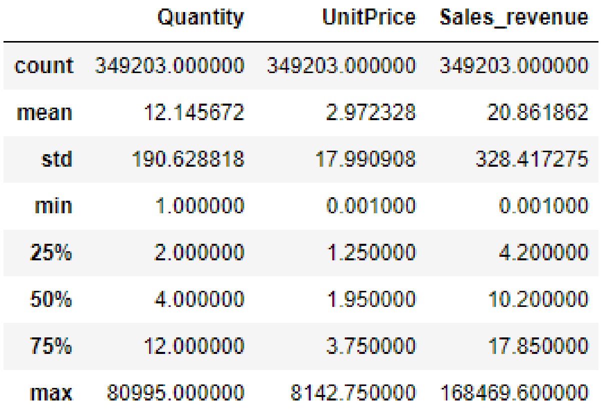


```
         Values                    Metric
0     -0.722814           Test Statistics
1      0.840751                   p-value
2      5.000000            # of lags used
3    299.000000  Number of observations used
4     -3.452411        critical value (1%)
5     -2.871255        critical value (5%)
6     -2.571947       critical value (10%)
```

Observation: From rolling mean and standard deviation we observed that the mean and standard deviation moves closely to each other which states that the data is stationary in nature. To state our hypothesis, we used the Adfuller test, and we got p-value greater than 0.5, hence we accept our null hypothesis.

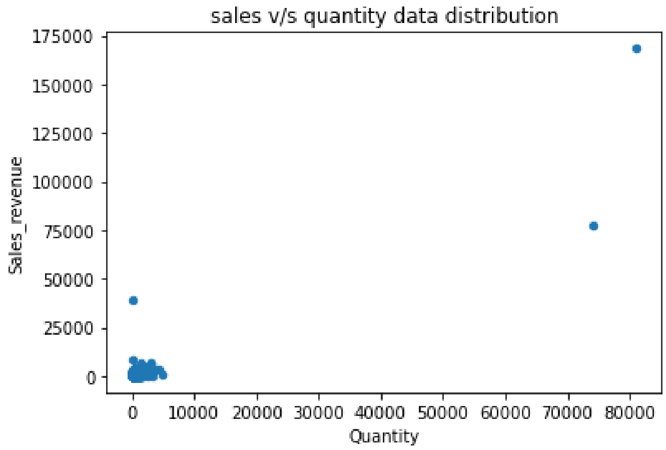
We have used attributes: “Quantity” and “UnitPrice” to predict the sales revenue.

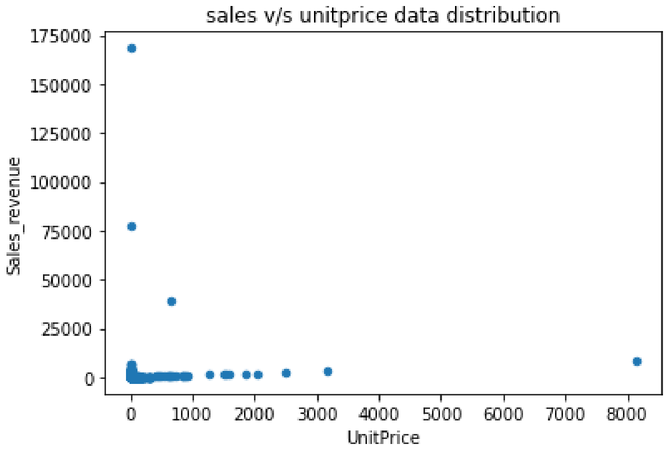
Statistical distribution of data:



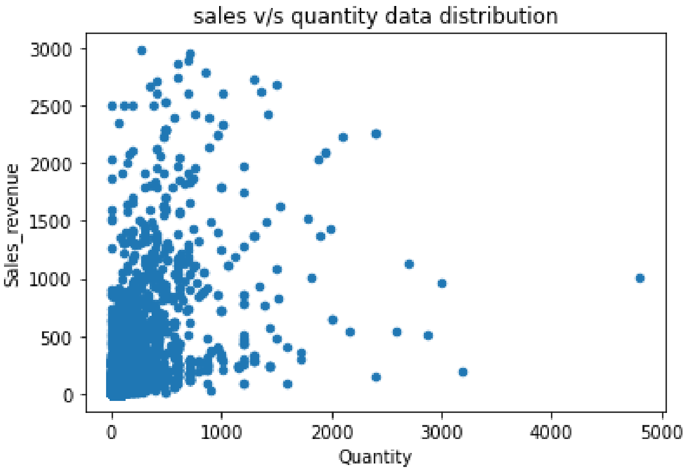
	Quantity	UnitPrice	Sales_revenue
count	349203.000000	349203.000000	349203.000000
mean	12.145672	2.972328	20.861862
std	190.628818	17.990908	328.417275
min	1.000000	0.001000	0.001000
25%	2.000000	1.250000	4.200000
50%	4.000000	1.950000	10.200000
75%	12.000000	3.750000	17.850000
max	80995.000000	8142.750000	168469.600000

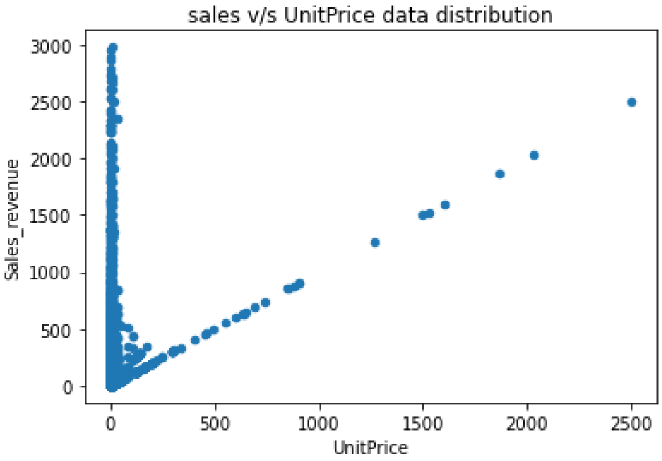
We can see there are some outliers in our dataset like I below charts. Here we can observe that few outliers are making our data very sparse in nature.



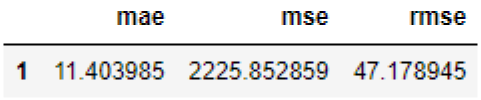


After removing the outliers, we have less sparse data than before:





Result: In order to evaluate our model, we have used metrices like Mean absolute error, mean square error, root mean square error and we got the results on our test dataset.



	mae	mse	rmse
1	11.403985	2225.852859	47.178945

Our results shows that the model is performing very poor on test datasets, which means it is unable to fit the data correctly due to high variance and sparseness in data.

4. Tree Base Model:

A sample tree-structured classifier can have three types of nodes.

- The Root Node is the initial node which represents the entire sample and may get split further into further nodes.
- The Interior Nodes represent the features of a data set, and the branches represent the decision rules.
- The Leaf Nodes represent the outcome.

The final output (let's say leaf node A) which is supposed to be the predicted value is the average of all the values of the dependent variables in that particular leaf node (A). Through multiple iterations the tree is able to predict a proper value for each leaf node by changing the value of yes/no.

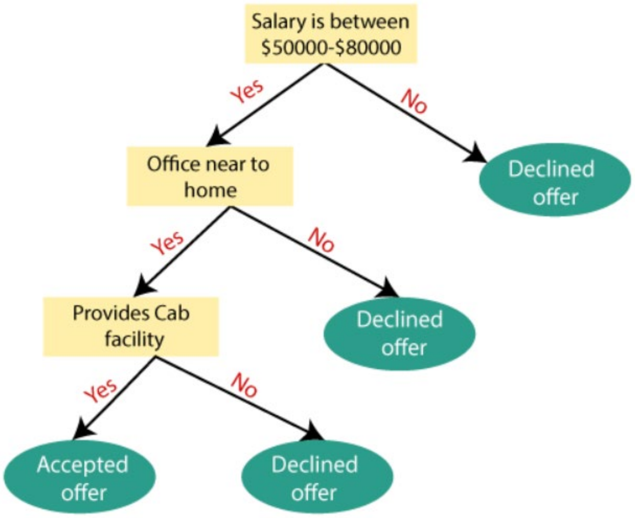
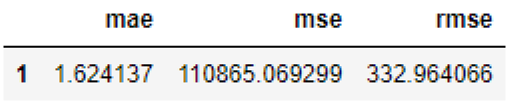


Fig 20. An example for Decision Tree (Source)

Our Implementation: We have tried the decision tree regressor method to predict the y variable "sales revenue" w.r.t our x variables "quantity and UnitPrice".

Result: In order to evaluate our model, we have used metrices like Mean absolute error, mean square error & root mean square error and we got the results on our test dataset.



	mae	mse	rmse
1	1.624137	110865.069299	332.964066

Due to sparseness and high variance, we couldn't achieve higher accuracy. We faced high Mean Square Error & Root Mean Square Error.

Conclusion: For both of our predictive models the results are poor which leads us to the conclusion that predictive models are not suitable for our dataset.

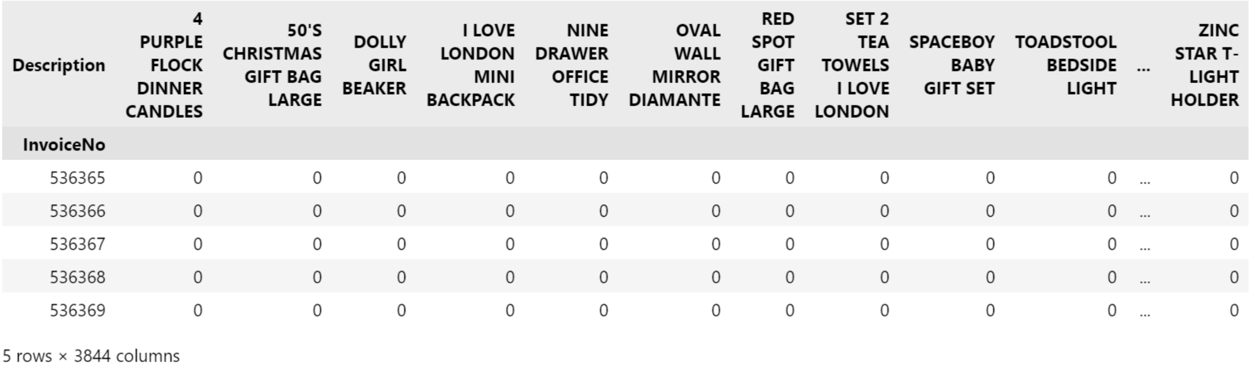
5. Associative Rule Mining

It's a procedure which helps in finding the frequently occurring patterns, correlations, or associations from datasets found in various databases. These are basically the "if/then" statements which helps in discovering relations between seemingly independent attributes. Some of the useful application can be in the field of Healthcare, macro/micro-economics etc.

OUR DATASET

Here we have created recommendation of products for the customers as a Market Basket Analysis problem. We think that the most important thing for these kinds of data is creating recommendations which are generated by the retailers (or entities having huge data). In case of our dataset its product recommendations but for other datasets there can be other recommendations too.

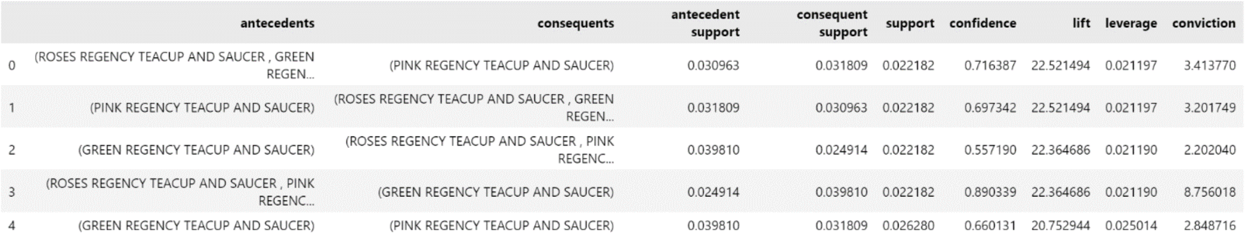
For this we have mapped our data according to InvoiceNo vs all the product items. This is because we are trying to find the items bought in a single order.



Description	4 PURPLE FLOCK DINNER CANDLES	50'S CHRISTMAS GIFT BAG LARGE	DOLLY GIRL BEAKER	I LOVE LONDON MINI BACKPACK	NINE DRAWER OFFICE TIDY	OVAL WALL MIRROR DIAMANTE	RED SPOT GIFT BAG LARGE	SET 2 TEA TOWELS I LOVE LONDON	SPACEBOY BABY GIFT SET	TOADSTOOL BEDSIDE LIGHT	...	ZINC STAR T-LIGHT HOLDER
InvoiceNo												
536365	0	0	0	0	0	0	0	0	0	0	...	0
536366	0	0	0	0	0	0	0	0	0	0	...	0
536367	0	0	0	0	0	0	0	0	0	0	...	0
536368	0	0	0	0	0	0	0	0	0	0	...	0
536369	0	0	0	0	0	0	0	0	0	0	...	0

5 rows × 3844 columns

Fig17. This is how or data looks like after the re-arrangement. As we are trying to find association between the products, we have filtered out the rows which have less than 2 products.



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(ROSES REGENCY TEACUP AND SAUCER , GREEN REGEN...	(PINK REGENCY TEACUP AND SAUCER)	0.030963	0.031809	0.022182	0.716387	22.521494	0.021197	3.413770
1	(PINK REGENCY TEACUP AND SAUCER)	(ROSES REGENCY TEACUP AND SAUCER , GREEN REGEN...	0.031809	0.030963	0.022182	0.697342	22.521494	0.021197	3.201749
2	(GREEN REGENCY TEACUP AND SAUCER)	(ROSES REGENCY TEACUP AND SAUCER , PINK REGENC...	0.039810	0.024914	0.022182	0.557190	22.364686	0.021190	2.202040
3	(ROSES REGENCY TEACUP AND SAUCER , PINK REGENC...	(GREEN REGENCY TEACUP AND SAUCER)	0.024914	0.039810	0.022182	0.890339	22.364686	0.021190	8.756018
4	(GREEN REGENCY TEACUP AND SAUCER)	(PINK REGENCY TEACUP AND SAUCER)	0.039810	0.031809	0.026280	0.660131	20.752944	0.025014	2.848716

Fig18. Some of the rules with their occurrence information.

In case of rule mining, we have some different terminologies. Let us explain these in detail in the following section.

- **Antecedents & Consequents**: - These are the 2 main components where an "antecedent" signifies "if" and a "consequent" signifies "then". The pillar of this algorithm is based on if-then statements. Antecedent is something that is found in the data (this can be a single item, or it can also be a combination of different items) and with that consequent is an item (can also be a combination of more than 1 item) that is found in combination with the antecedent.

  A simple example would be: If he has graduated from UB, he is 80% likely to land in a job (of starting salary $80,000). In my example my consequent is a combination of 2 items.
- **Support**: - It tells us about the probability of the item/relationship's occurrence.
- **Confidence**: - Probability of the relationships having been found to be **true**.
  `Confidence = Support (X, Y)/Support(X)`
- **Lift**: - It is the ratio of the joint probability of two items x and y, divided by the product of their probabilities.
  `Lift = P (X, Y)/[P(X)P(Y)]`
- **Conviction**: - It compares the probability that X appears without Y if they were independent with the actual frequency of the appearance of X without Y.
  `Conviction(X→Y) = (1 - Support(Y)) / (1 - Confidence(X→Y))`
- **Leverage**: - Leverage measures the difference of X and Y appearing together in the data set and what would be expected if X and Y where statistically dependent.
  `leverage (X -> Y) = P (X and Y) - (P(X)P(Y))`

Some of the recommendations that we have found are: -

- `{'ROSES REGENCY TEACUP AND SAUCER ', 'GREEN REGENCY TEACUP AND SAUCER'} => {'PINK REGENCY TEACUP AND SAUCER'}`
- `{'PINK REGENCY TEACUP AND SAUCER'} => {'ROSES REGENCY TEACUP AND SAUCER ', 'GREEN REGENCY TEACUP AND SAUCER'}`
- `({'GREEN REGENCY TEACUP AND SAUCER'}) => {'ROSES REGENCY TEACUP AND SAUCER ', 'PINK REGENCY TEACUP AND SAUCER'}`
- `({'ROSES REGENCY TEACUP AND SAUCER ', 'PINK REGENCY TEACUP AND SAUCER'}) => {'GREEN REGENCY TEACUP AND SAUCER'}`

These rules are in decreasing order w.r.t **lift** and these mean that when item like **'ROSES REGENCY TEACUP AND SAUCER**’ & **'GREEN REGENCY TEACUP AND SAUCER**' are bought together there is a high chance that the customer will buy '**PINK REGENCY TEACUP AND SAUCER**'