**4. INTRODUCTION**

**4.1 STATEMENT OF THE PROBLEM**

Build models to extract customer behaviour and other trends from data. It also improves the marketing strategies of retailers which will help in developing their business.

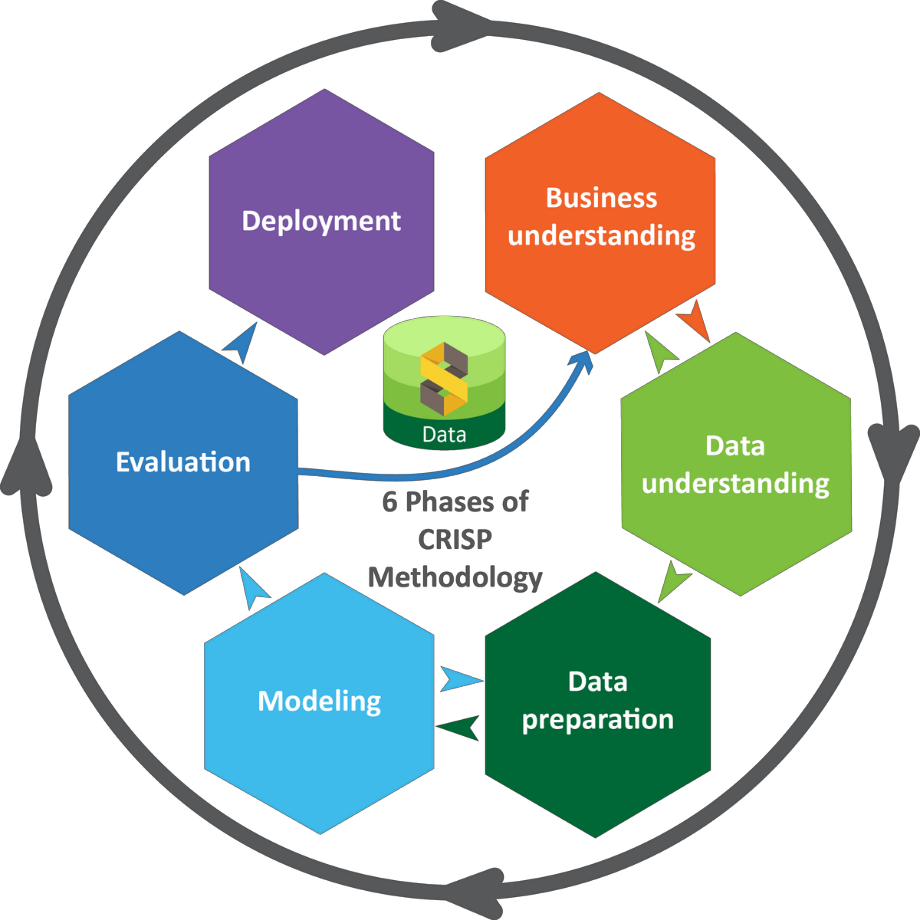
**4.2 DESCRIPTION OF DATA**

A E-Commerce site X provided real sales data transaction wise with time of the purchase, total quantity purchased along with unit price. The data set also contains customer demographics (age, gender, annual income), product details (product id, customer ratings).

**4.3 PROJECT SCOPE**

Focus of the project will be on studying the behavior of the customer. The more insights can be drawn by visualizing the data as it provides more information, and this can be done on each individual variable in data and the target variable. We can build recommender models likes ALS, Content Based Filtering for making appropriate recommendations to customers to boost the sales and other clustering algorithms like k-means for studying customers.

### SOLUTION APPROACH



CRISP-DM ( Cross Industry Standard Process for Data Mining) methodology is followed to approach the business problem using following steps.

* **Business Understanding**: Understand business problem from business perspective.
* **Data Understanding**: Collect data, perform descriptive and exploratory data analysis using univariate, bivariate and multivariate analysis to discover initial insights into the data.
* **Data Preparation**: Perform various activities to construct final dataset from raw dataset. The final dataset is used for modelling. This step includes data cleaning and data transformation.
* **Modelling**: Select and apply various modelling techniques. The model which best explains and matches the domain information and also provides good opportunity to grow sales is to be found.
* **Evaluation**: Thoroughly evaluate the model using statistical trchniques and ensure the business objective is met.
* **Deployment:** Generate insights and recommendation from the data and model that can be presented to the business.

### DATA UNDERSTANDING

There are two datasets, Sales and Customers datasets. The sales data contains 10000 rows and 9 columns, the customers data consists of 9215 rows and 5 columns. Customer information is provided for 9215 unique customers. There are no null values. As the data contains rating column which will help in building good recommender systems.The customer data can also b used for performing clustering analysis to see what kind of customers are present in our consumers and how are their purchase patterns.

|  |  |
| --- | --- |
| Variable | Definition |
| **transaction id** | **transaction id** |
| **product id** | **product id** |
| **product description** | **product description** |
| **quantity sold** | **quantity sold** |
| **transaction timestamp** | **transaction timestamp** |
| **unit price** | **unit price of the product** |
| **customer id** | **customer id** |
| **transaction country** | **transaction country** |
| **rating** | **rating** |

The data has the transaction id that is unique for each transaction. Customer id will provide the user information and this Id is different for different customers this can be used to join with customers data when needed. The product Id gives the details of the products. Unit price and quantity sold of the product in that particular transaction also given. The data also has the information of the customer’s occupation and their age.

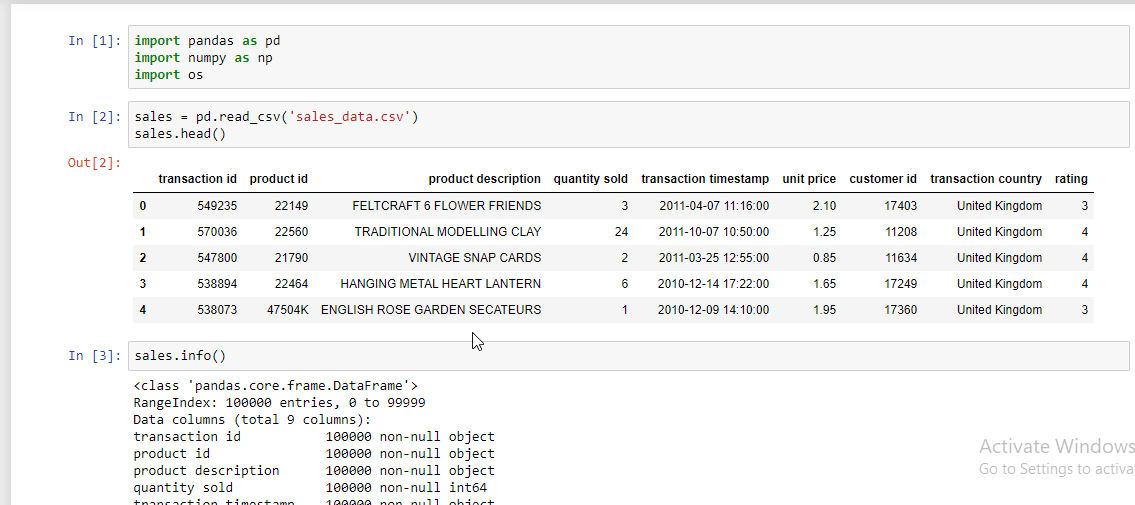
Customers.csv

|  |  |
| --- | --- |
| Variable | Definition |
| **customer id** | **customer id** |
| **gender** | **gender** |
| **age** | **age** |
| **annual income(k$)** | **annual income in thousands of dollars** |
| **score** | **Spending Score from 1 to 100** |

This data consists of customer data like their recorded stats including their spending score annual income and personal info like age and gender.

Code:

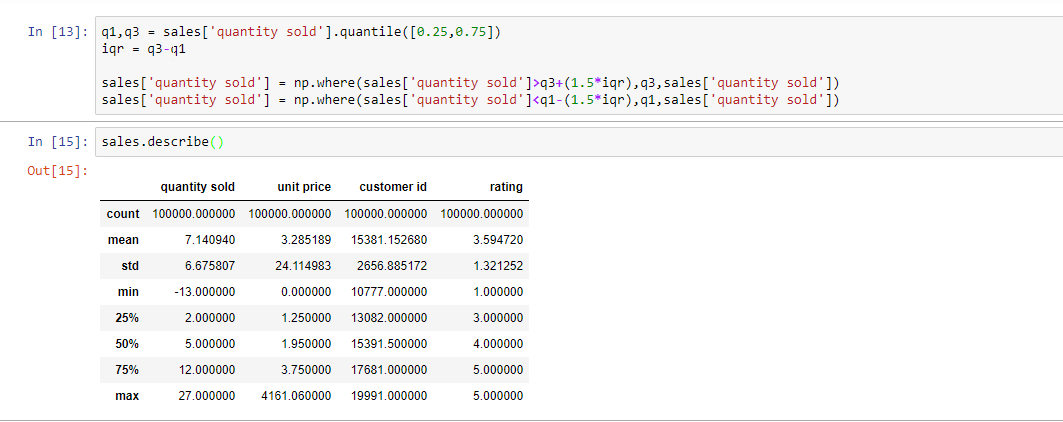
The data had no nulls so there is no much data cleaning needed



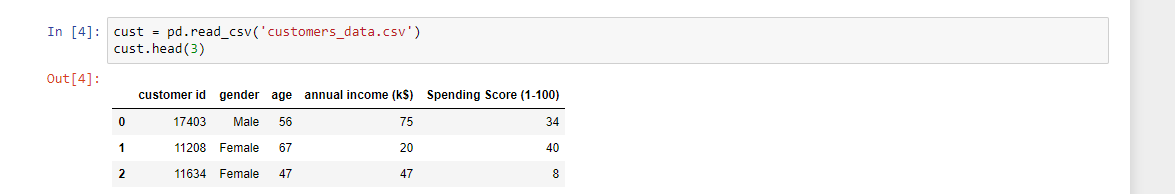
But the data had outliers in quantity sold variable

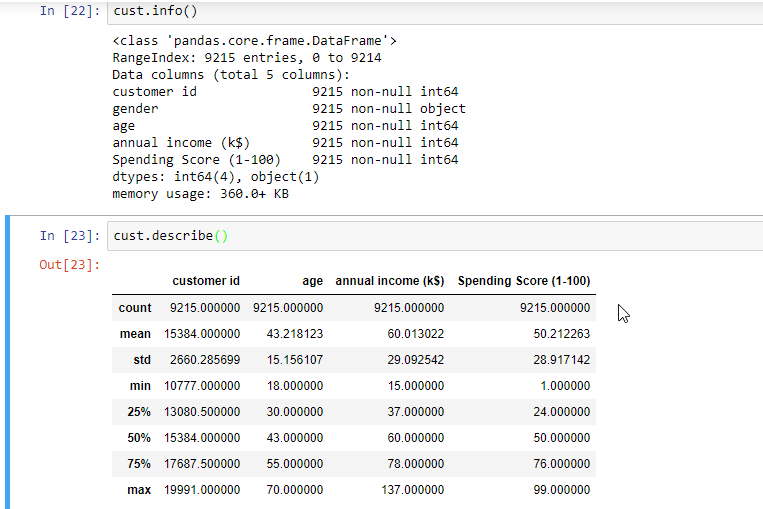


I dealt with the outliers by replacing them with upper and lower bounds.



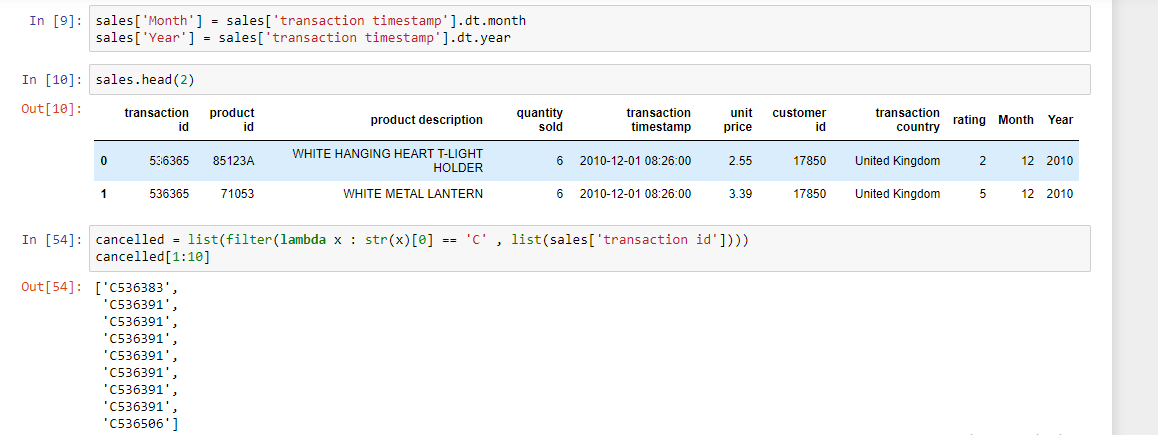
Customer data



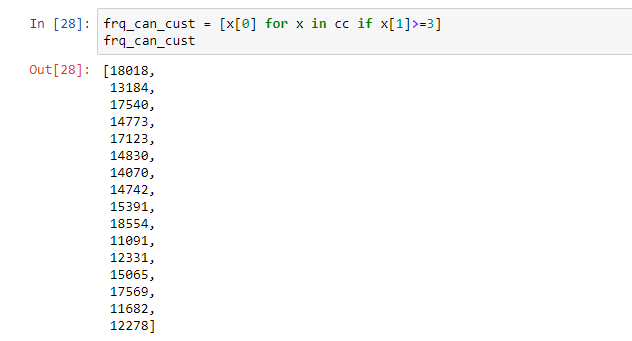
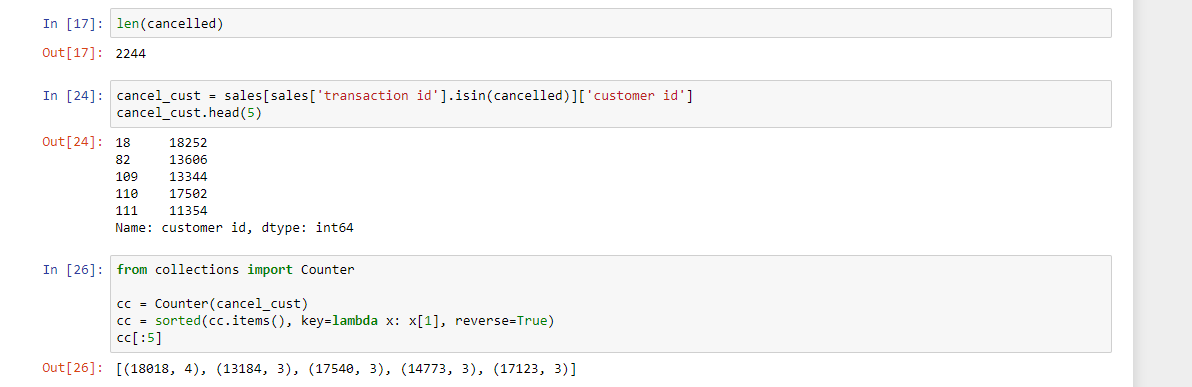


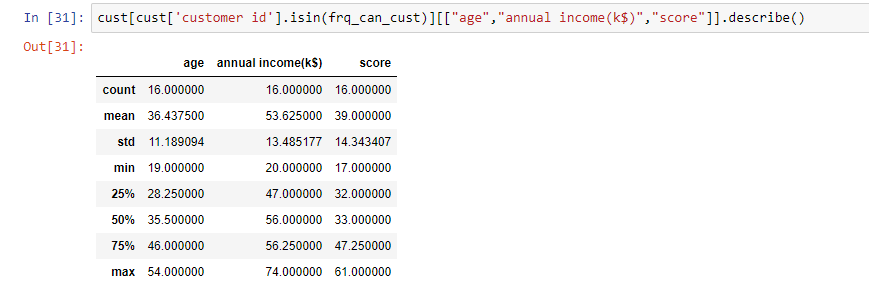
The customer data was fine without any nulls or outliers.And the number of unique customers were 9215.

The time stamp column needs to be converted to datetimetype to use date operations like extracting day/month and year of the transaction.



The transaction id starts with a C if the order is cancelled. Extracting orders that are cancelled to perform further analysis of why they were cancelled and which type of customers are cancelling the orders more and why?

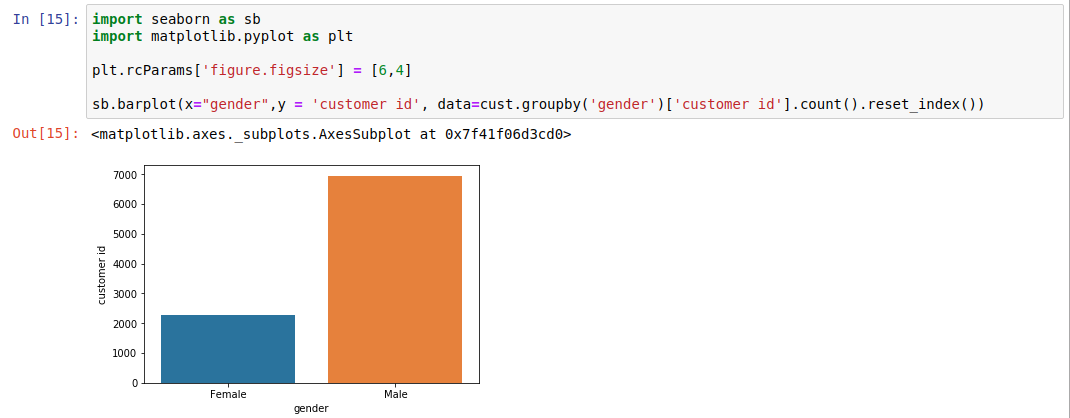




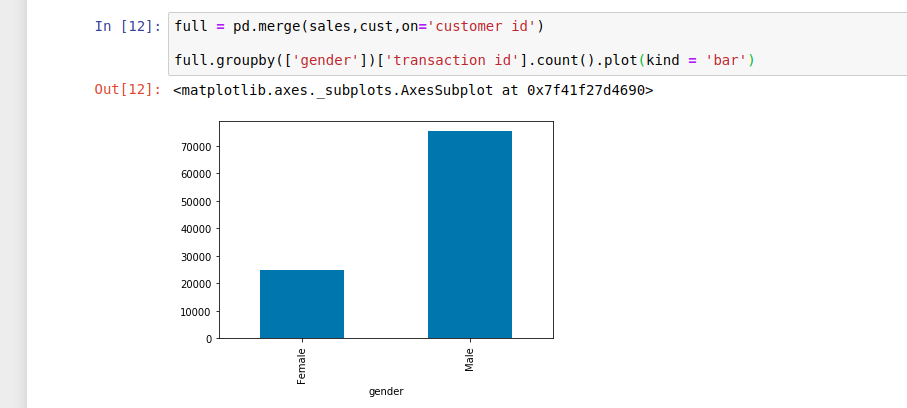
From the above statistics we can get an idea of customers who are cancelling the orders often.We can do futher analysis like the dates on which the cancellations are occurring more.

### EDA (EXPLORATORY DATA ANALYSIS)

**GENDER ANALYSIS**

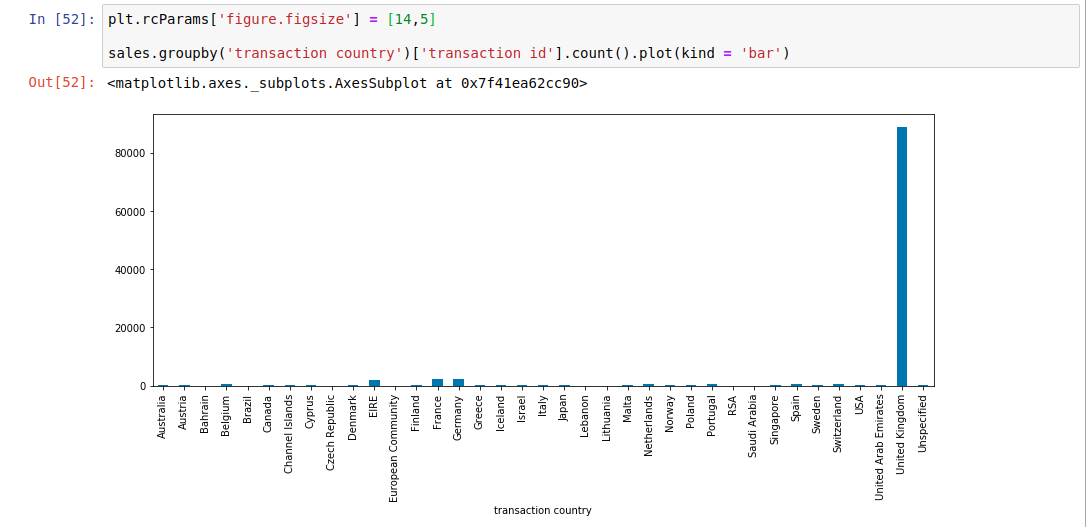
EDA done on the data:

The above figure gives the count of male and female customers in the data. The X-axis represents the type of gender. The Y-axis gives the count of the customers. From the graph we can tell that the most customers are male customers (414259) than the female customers (135809).



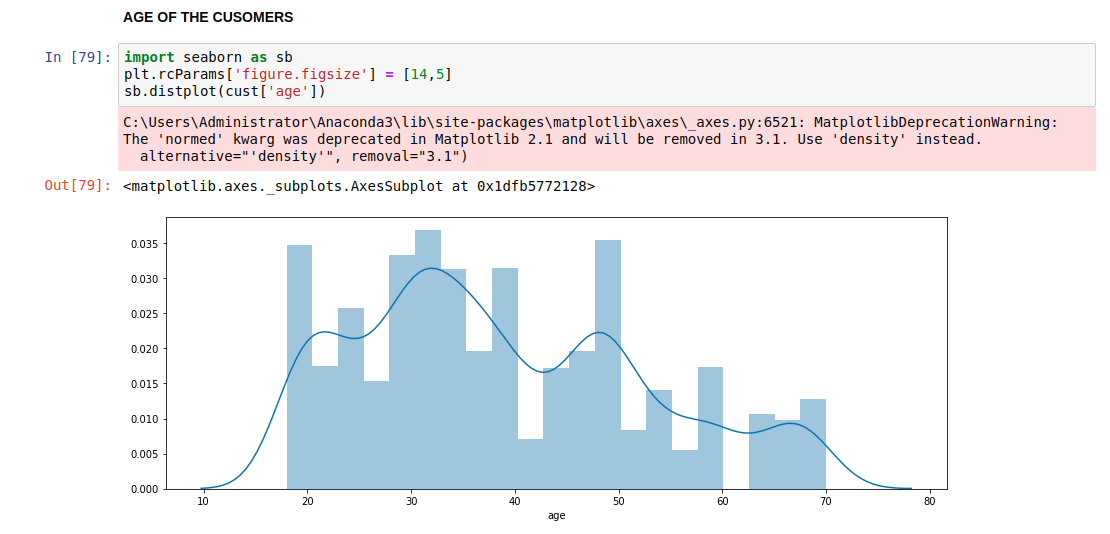
The above histogram gives the details of purchases made by male and female customers. The values in the graph represents that the Male has purchased more when compared to Female. This is because the data contains more male information.

**ANALYSIS ON PURCHASES LOCATIONS:**



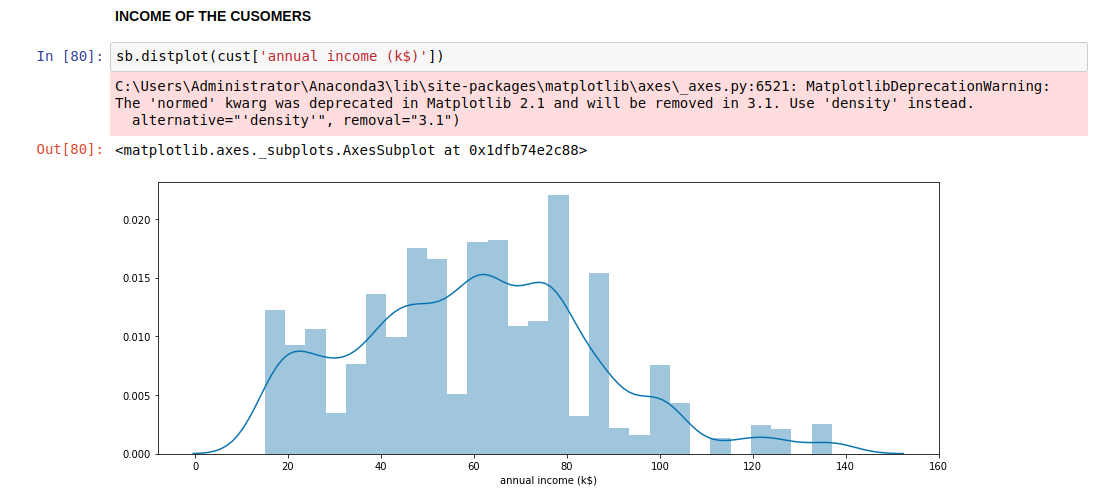
The above bargraph gives the details of purchases made by customers from different. The values in the graph represents that the Male has purchased more when compared to Female. This is because the data contains more male information.

**ANALYSIS ON PURCHASES MADE BY DIFFERENT AGE GROUPS**



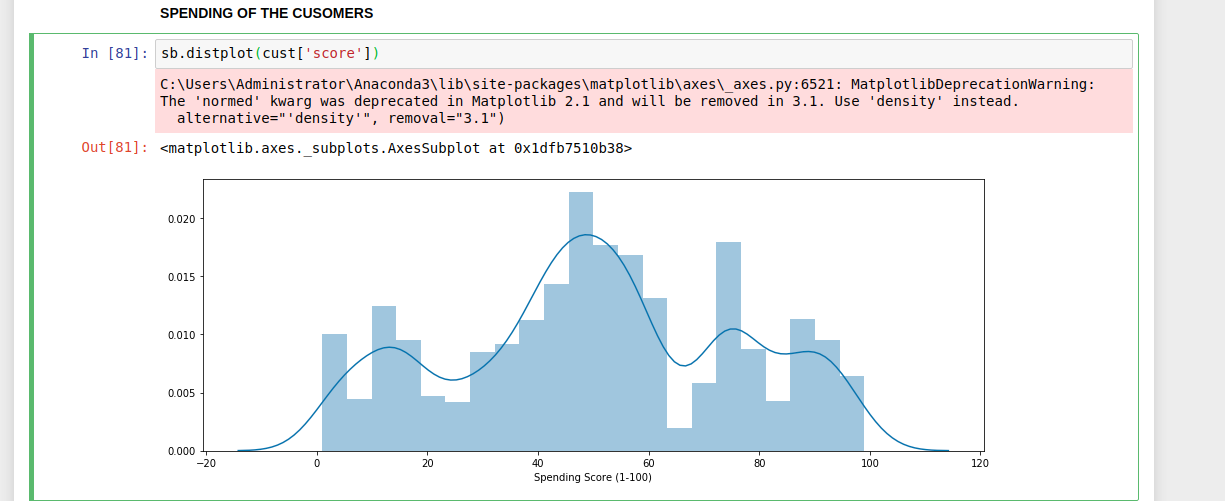
The above histogram gives the details of purchases made by different age group customers. The values in the graph represents that the age group of 26-35 years has purchased more when compared to all the other age groups.

**ANALYSIS ON INCOME OF THE CUSTOMERS**

****

Most of the customers are earning around 60-80k. So, the future plans/camppaigns should be made keeping this check.

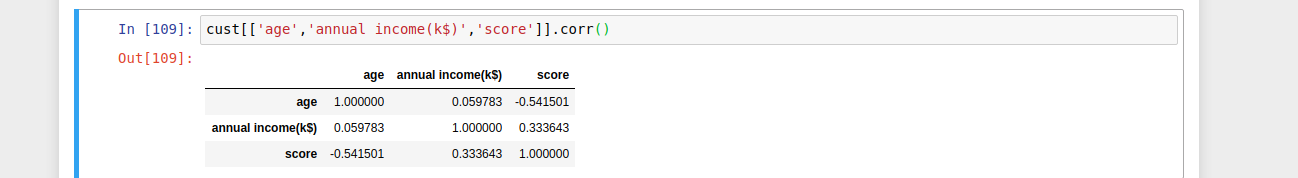
**ANALYSIS ON SPENDING OF THE CUSTOMERS**

****

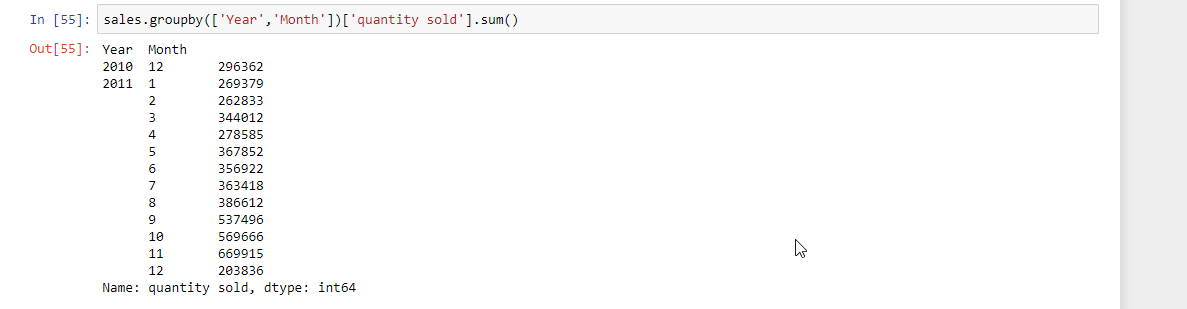
This graph shows that most people have an average spending score and we need to try and bring thiose in the lower bracket to above mean spending score.

**CHECKING THE RELATIONSHIP BETWEEN THE COLUMNS**

Correlation between different variables in customer data.

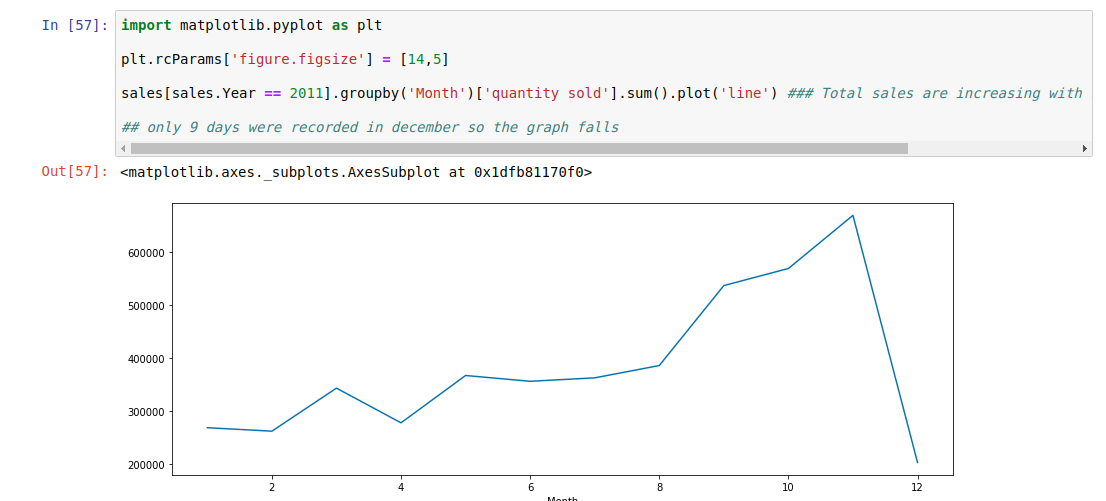
****

**ANALYSIS ON COMPANY SALES:**



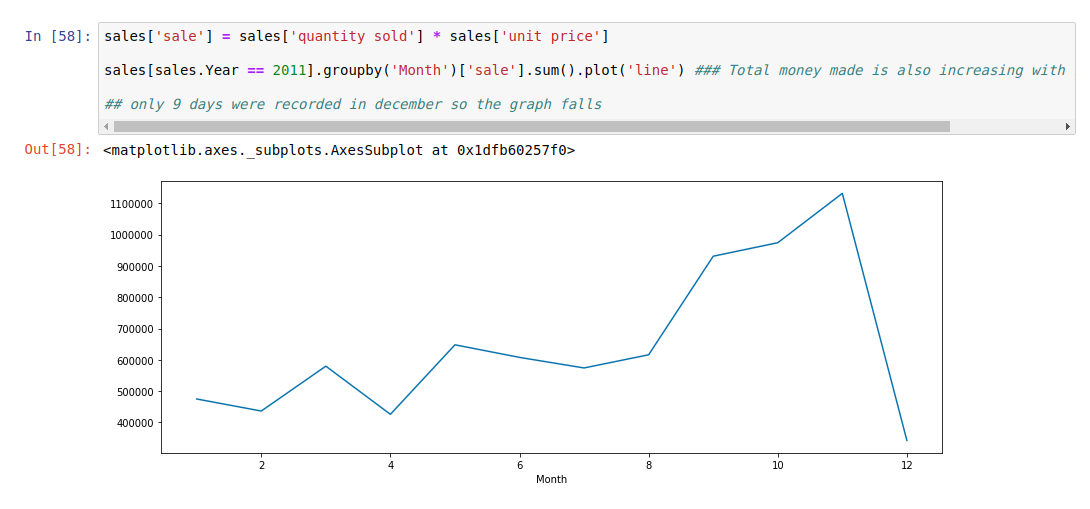
**I: Quantity sold month wise**:

With this line graph we can visualize the spikes and downfalls in the quantity sold by the company in the months of the year and we can see he spikes around second half of the year probably because of the holiday season.



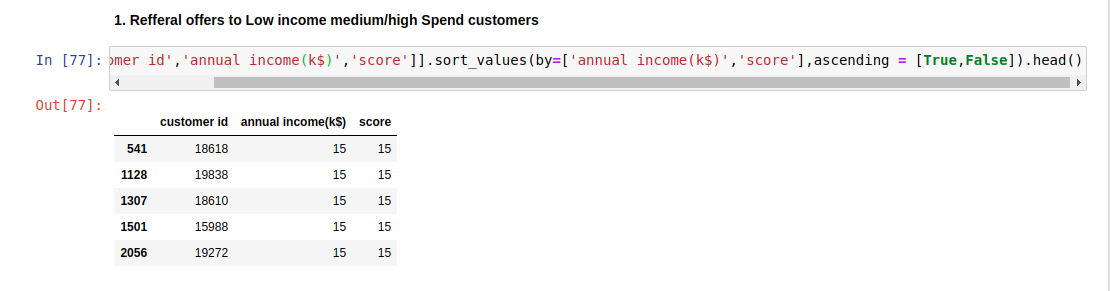
**II: Revenue generated month wise:**

Here we can see similar pattern in revenue as it is followed by number of sales this line graph we can visualize the spikes and downfalls in the total revenue generated by the company in the months of the year and we can see he spikes around second half same as above.

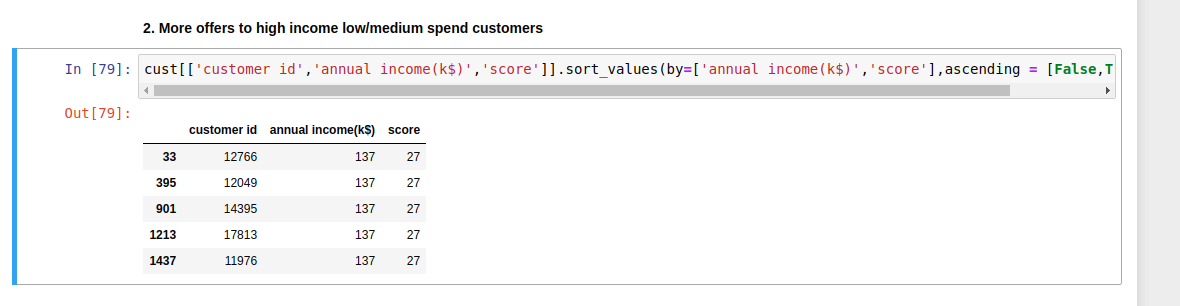


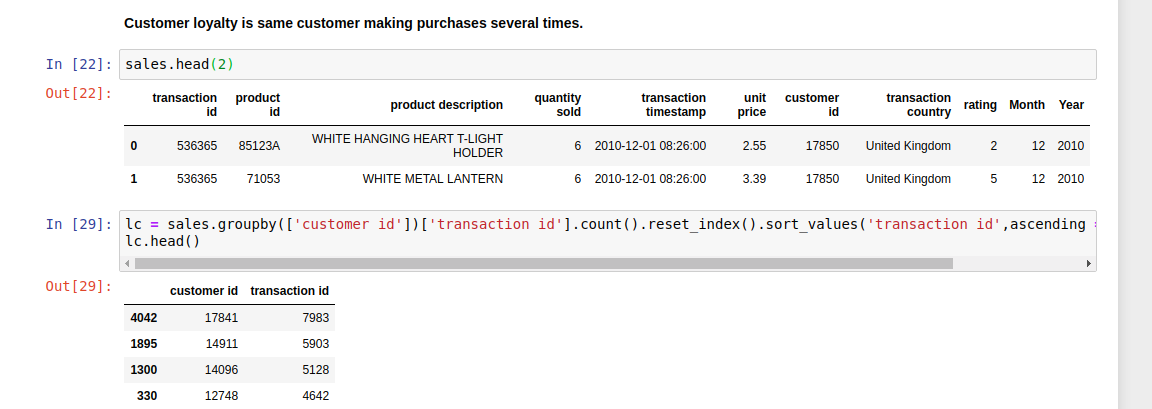
**7.2 INITIAL INSIGHTS AND TRENDS IN DATA**

**I. Types of customers:**

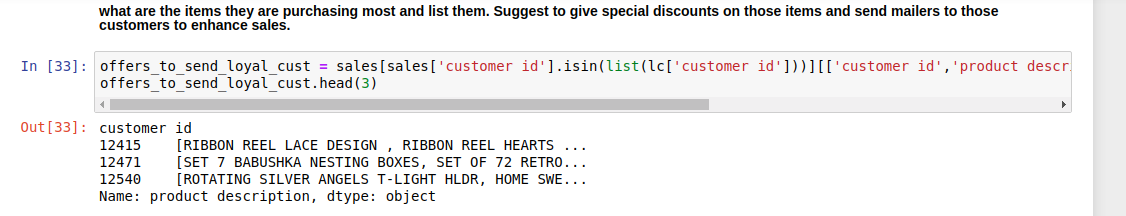


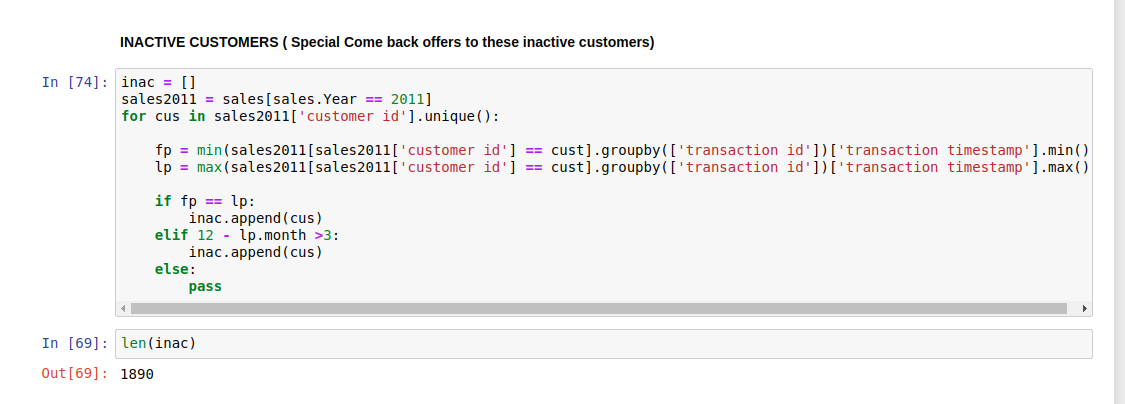
we find who are the low income high/medium spend customers we can give more refferals to these customers which will help in increase the customer base and also increase the overall revenue generated from them.

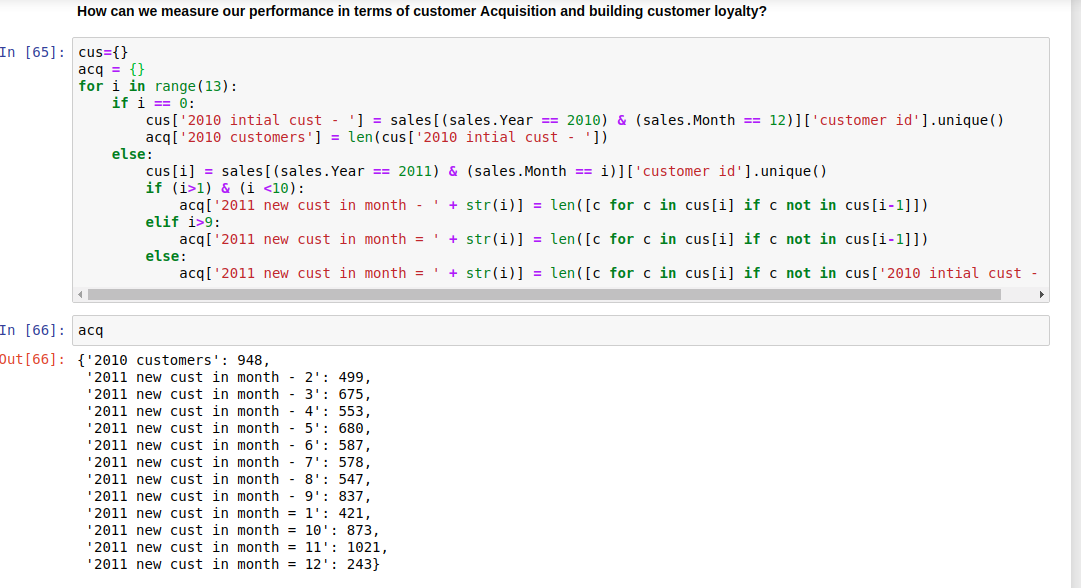
we find who are the high income low/medium spend customers we can give more special offers on costly products which have lesss sales(clearance sales) to these customers which will help in increasing their overall spend on the site and also increase their overall purchase frequency.

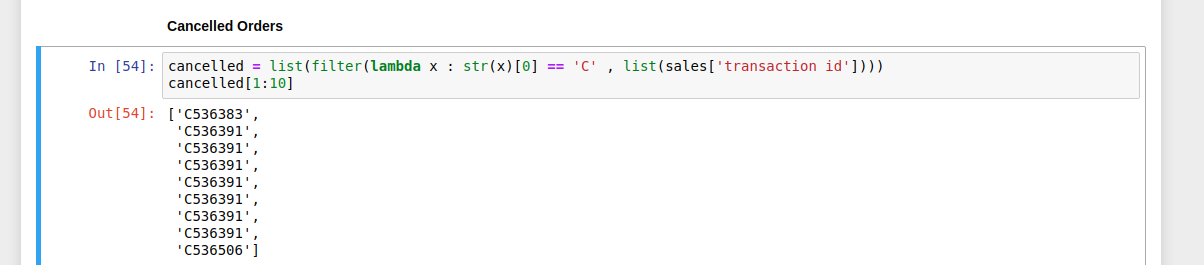


Special offers can be given to loyal customers on the products that they’re buying the most. This will keep the customers happy and make them want to stay loyal.

Finding all the items that are most purchased by the loyal customers to give special offers on them to our loyal customer base.

Finding inactive customers that have been not making any purchases recently . We can give some kind of comeback offer that will pull them back and make them active again.

First n times offers can be given to New customers to get them used to our site and then prbly convert them to a long lasting loyalcustomers who can cover up the gap through life time purchases.

Cancelled orders can be used to find out what type of customers are cancelling the orders. This will also help in pinpointing the cause for cancellation. By doing this we can hopefully reduce the cancellation percentage of the products thereby reducing the costs incurred by cancellations. If any customer is repeatedlty cancelling the orders these customers can be flagged and the dispatch of order for these customers can be delayed just incase the customer decides to cancel again alternatively we can also survey such customers for the exact reason behind cancelling any order to get better idea of how to maintain and deliver products and therby reducing the chance of further cancellations.

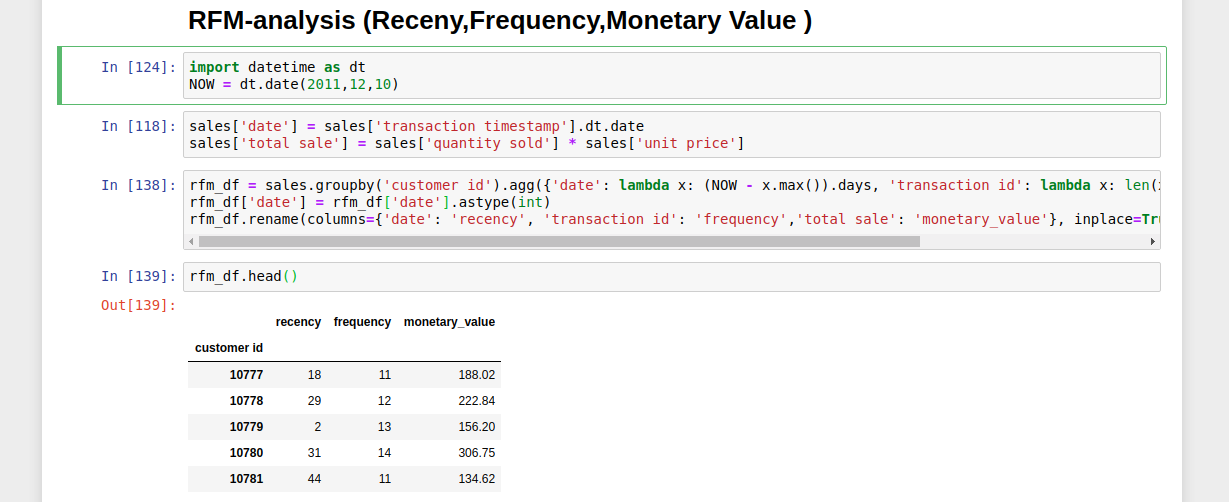
**II. RFM (recency, frequency, monetary)**

RFM (recency, frequency, monetary) analysis is a marketing technique used to determine quantitatively which customers are the best ones by examining how recently a customer has purchased (recency), how often they purchase (frequency), and how much the customer spends (monetary).

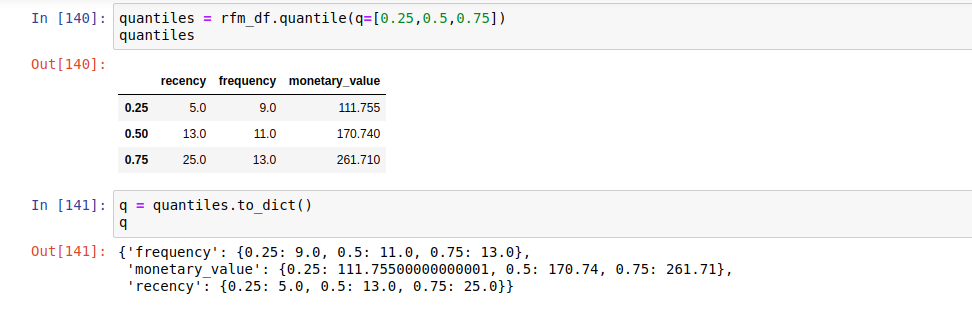
* Objective is to Calculate the RFM values for each customer (by customer id):
  + For Recency calculation, use 12/2011 as current month. So, the Recency should be how many months before he or she has made a purchase from the current date. If made a purchase in December, then the values should be 0 and so on so forth.
  + Frequency – At an average how often each customer make purchases in a month.
  + Monetary Value – Total Spend by the customers in the whole year.

**Implementation**

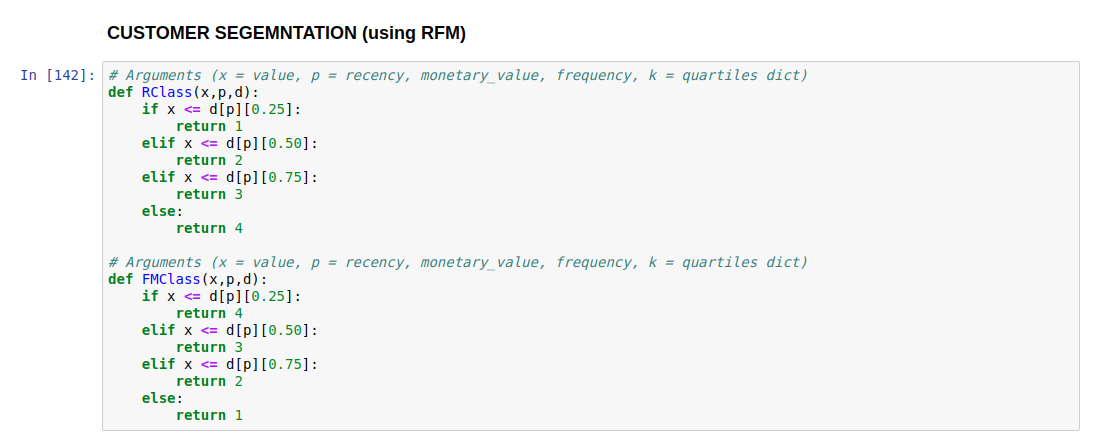
* identify top 5 customers based on frequency and monetary value. Sort them based on first frequency and then monetary value.
* identify optimal number of segments using dendogram and elbow method.
* Create final segment of customers and label the customers based on which segment they belong to.
* Explain each segment intuitively.

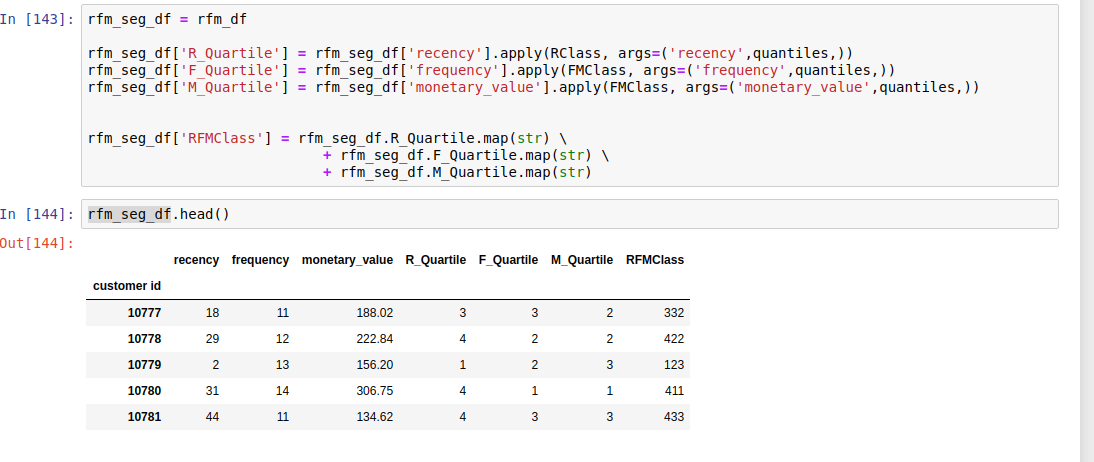


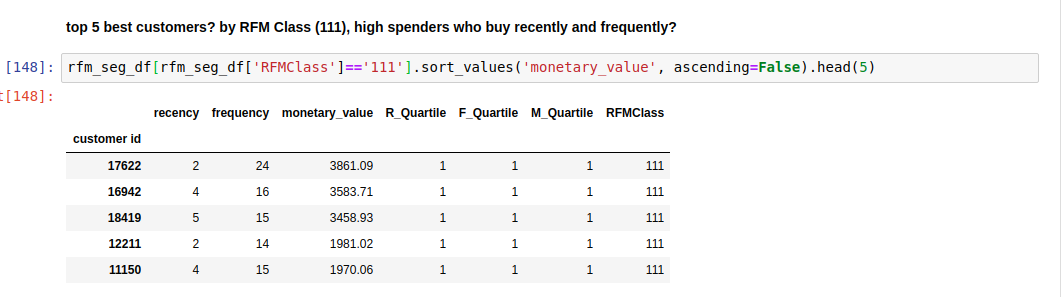
Creating a new dataframe name rfm\_df which will have the recency, frequency and monetary revenue genrated from that customer.

We can calculate the quantiiles of each of the factors which we use segment the customers.

I wrote 2 functions to classify the customers based on their derived charecterstics which will later be used to find the customers of intrest post segmentation.



After segmentation based on the classes we can see the data now is segemneted based on the RFMclass column which will give us an idea of which quantile the customer belongs to.



We can later filter the data frame to extract the customers of intrest by sorting and filtering the Dataframe.

**8. Model Building:**

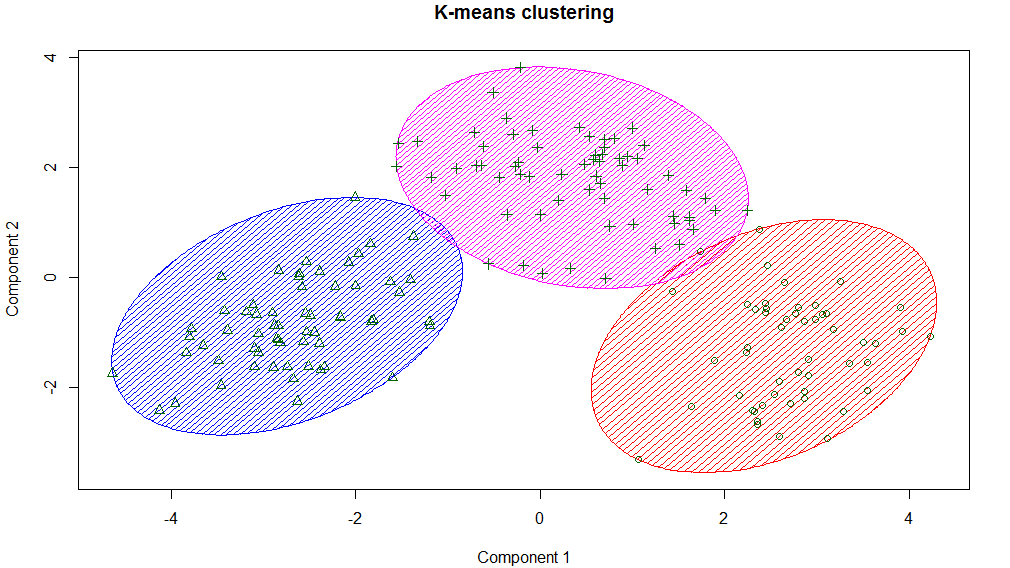
**8.1 Clustering:**

Clustering is one of the most common exploratory data analysis technique used to get an intuition about the structure of the data. It can be defined as the task of identifying subgroups in the data such that data points in the same subgroup (cluster) are very similar while data points in different clusters are very different. In other words, we try to find homogeneous subgroups within the data such that data points in each cluster are as similar as possible according to a similarity measure such as euclidean-based distance or correlation-based distance. The decision of which similarity measure to use is application-specific.

**K-Means Clustering:**algorithm is an iterative algorithm that tries to partition the dataset into Kpre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the inter-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster’s centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

The way kmeans algorithm works is as follows:

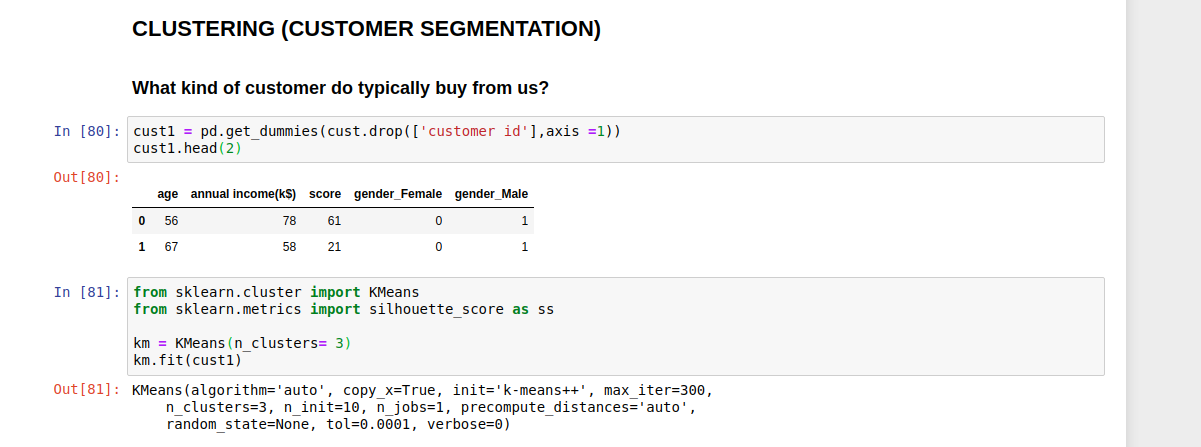
* Specify number of clusters K.
* Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
* Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn’t changing.
  + Compute the sum of the squared distance between data points and all centroids.
  + Assign each data point to the closest cluster (centroid).
  + Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

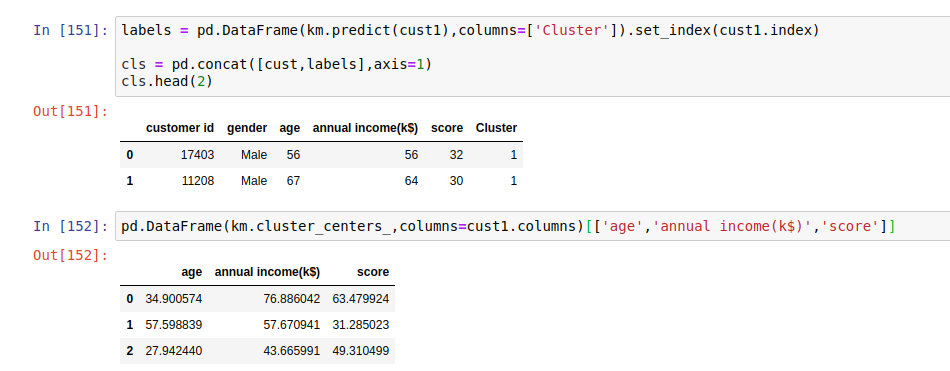


Few things to note here:

* Since clustering algorithms including kmeans use distance-based measurements to determine the similarity between data points, it’s recommended to standardize the data to have a mean of zero and a standard deviation of one since almost always the features in any dataset would have different units of measurements such as age vs income.
* Given kmeans iterative nature and the random initialization of centroids at the start of the algorithm, different initializations may lead to different clusters since kmeans algorithm may stuck in a local optimum and may not converge to global optimum. Therefore, it’s recommended to run the algorithm using different initializations of centroids and pick the results of the run that that yielded the lower sum of squared distance.
* Assignment of examples isn’t changing is the same thing as no change in within-cluster variation:

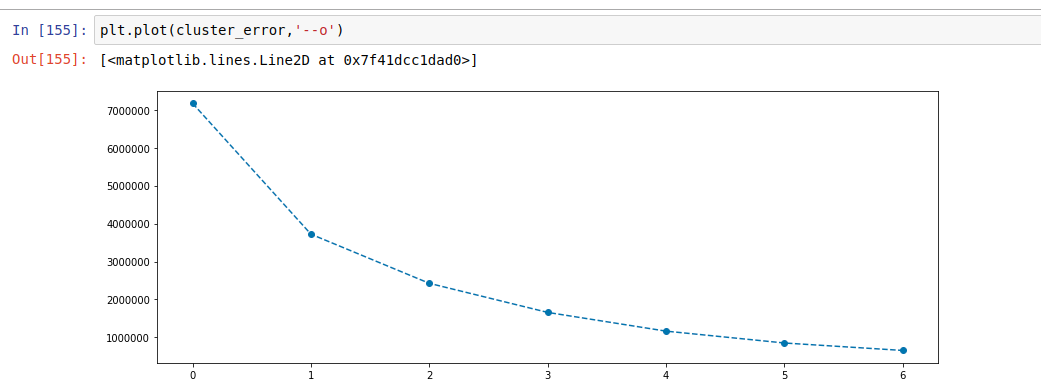
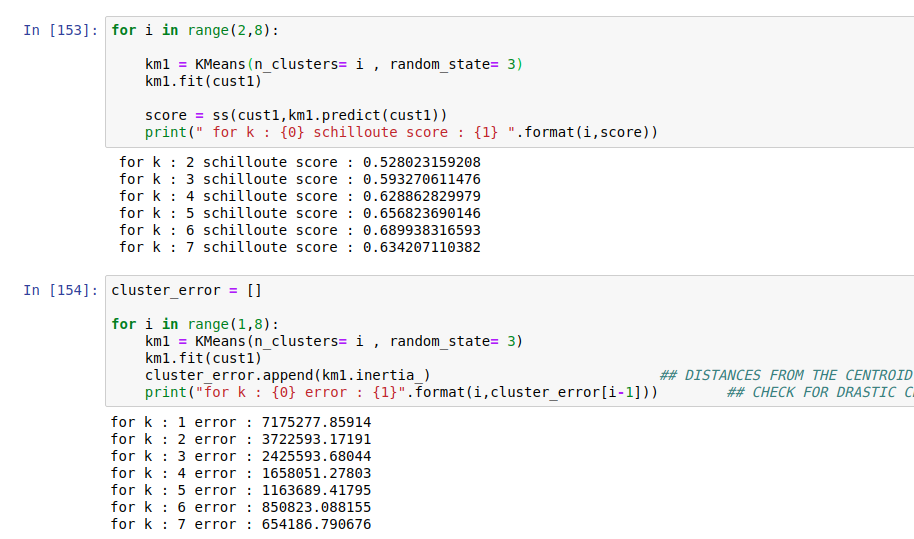
# Implementation:

Started by dummifying the data. Then applying the sklearn’s k-means algorithm on customer data to segment the customers. N\_clusters was taken as 3 for the initial run.



Initial clustering shows the centres which give an idea of what kind of customers to expect in that clusters. These customers can be labled accordingly and the customers in the same clusters can be studied further. Campaigns can be launched accordingly.

Finding the optimum number of clusters using the cluster error/schilloute score and elbow curve methods.



### 8.1.1 PCA (PRINCIPAL COMPONENT ANALYSIS)

The main idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of many variables correlated with each other, either heavily or lightly, while retaining the variation present in the dataset, up to the maximum extent. The same is done by transforming the variables to a new set of variables, which are known as the principal components (or simply, the PCs) and are orthogonal, ordered such that the retention of variation present in the original variables decreases as we move down in the order. So, in this way, the 1st principal component retains maximum variation that was present in the original components. The principal components are the eigenvectors of a covariance matrix, and hence they are orthogonal.

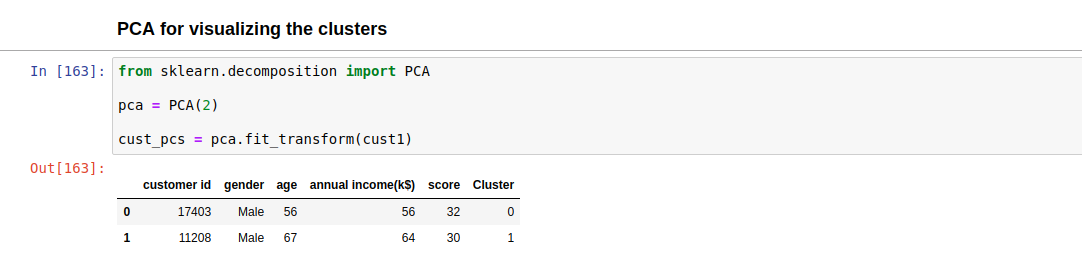
Importantly, the dataset on which PCA technique is to be used must be scaled. The results are also sensitive to the relative scaling. As a layman, it is a method of summarizing data. Imagine some wine bottles on a dining table. Each wine is described by its attributes like colour, strength, age, etc. But redundancy will arise because many of them will measure related properties.

When you’re working with a learning model, it is important to scale the features to a range which is centred around zero. This is done so that the variance of the features is in the same range. If a feature’s variance is orders of magnitude more than the variance of other features, that feature might dominate other features in the dataset, which is not something we want happening in our model.

The aim here is to achieve Gaussian with zero mean and unit variance. There are many ways of doing this, two most popular are standardisation and normalisation.

No matter which method you choose, the SciKit Learn library provides a class to easily scale our data. We can use the StandardScaler class from the library for this.

Here Primarily we are using PCA for visualizing our clusters.



We reduced the number of dimensions to 2 for making the visualization more easy. Then we colour our data points based on the lables and we can see the cluster clearly seperated.



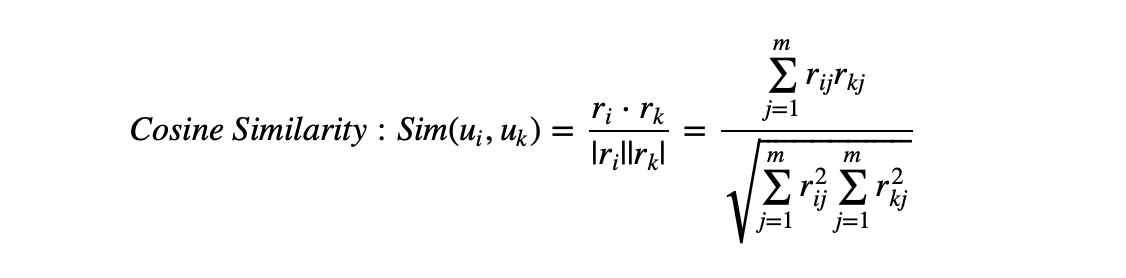
**8.2 Recommender System:**

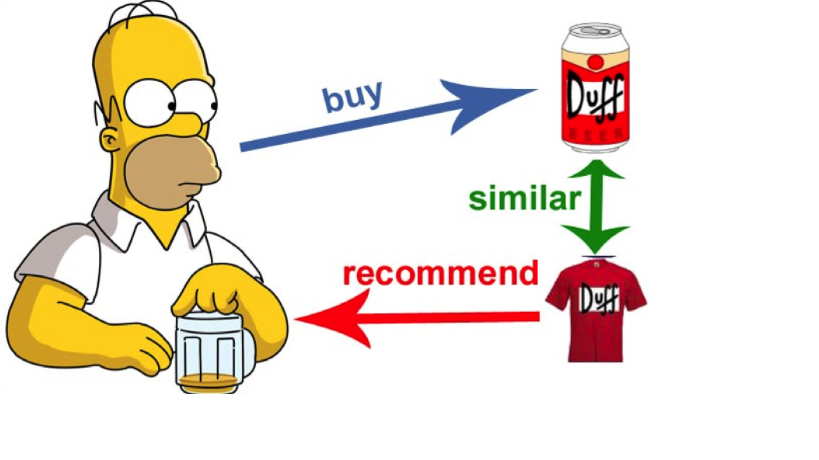
# Nearest Neighborhood Collaborative Filtering:

The standard method of Collaborative Filtering is known as Nearest Neighborhood algorithm. There are user-based CF and item-based CF. Let’s first look at User-based CF. We have an n × m matrix of ratings, with user uᵢ, i = 1, ...n and item pⱼ, j=1, …m. Now we want to predict the rating rᵢⱼ if target user i did not watch/rate an item j. The process is to calculate the similarities between target user i and all other users, select the top X similar users, and take the weighted average of ratings from these X users with similarities as weights.

While different people may have different baselines when giving ratings, some people tend to give high scores generally, some are pretty strict even though they are satisfied with items. To avoid this bias, we can subtract each user’s average rating of all items when computing weighted average, and add it back for target user, shown as below.

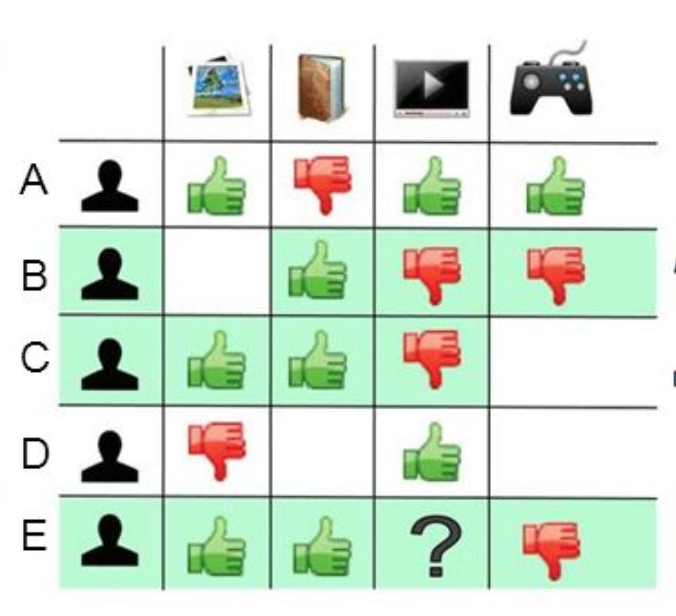
Two ways to calculate similarity are Pearson Correlation and Cosine Similarity.





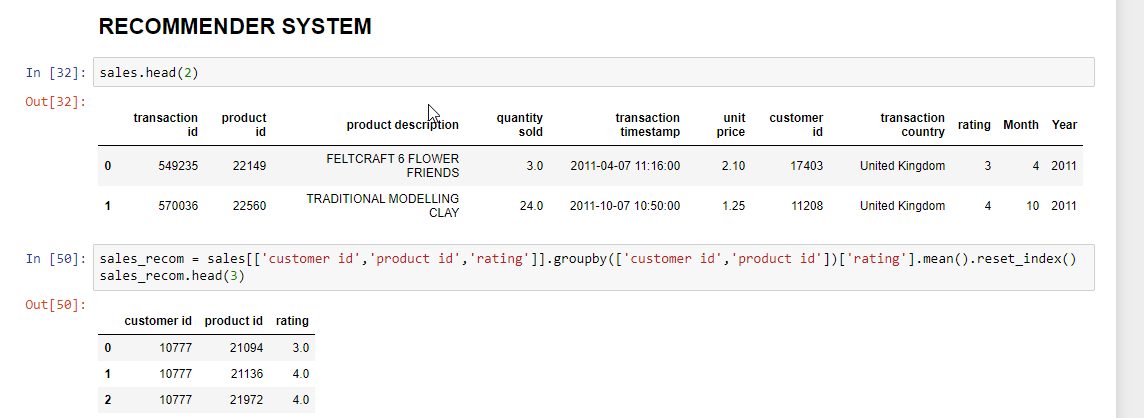
Basically, the idea is to find the most similar users to your target user (nearest neighbors) and weight their ratings of an item as the prediction of the rating of this item for target user.

Without knowing anything about items and users themselves, we think two users are similar when they give the same item similar ratings .

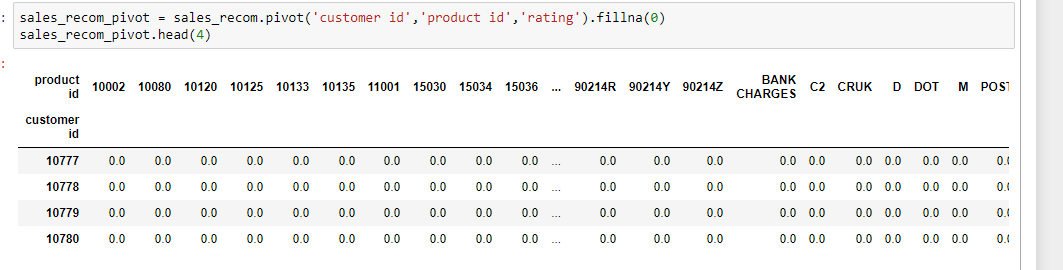


# Implementation:

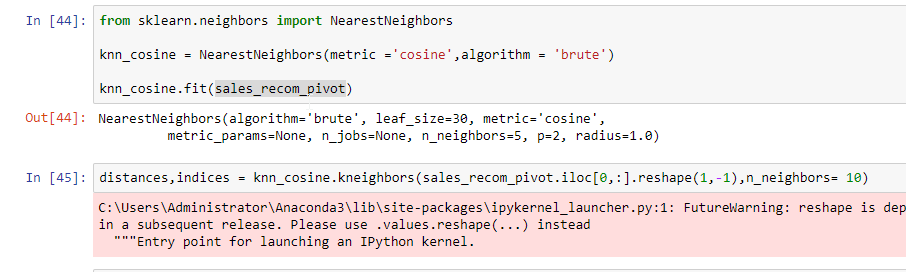
First we take out the columns that are required to build the recommender system. They are customer id, product id and rating.



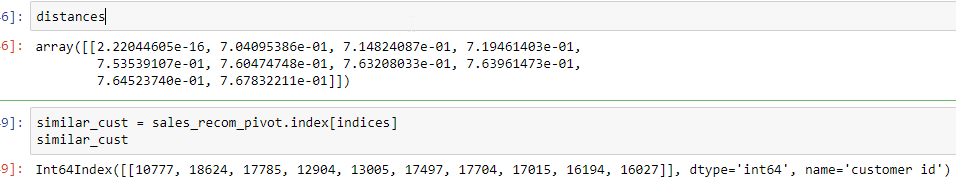
Now, We create a pivot table from these columns with rating as the value and filling na’s with 0’s this matrix is going to be very sparse.



Now, Building a user based collaborative filtering recommender system using sklearn’s Nearest Neighbors package. Here I am using cosine similarity to find out the neightbors which uses the angle between the data points as the measurement.



Recommender system building is done now to find top 10 similar customers given any customer is done by giving the model the customer of intrest’s data point and the no of neighbors needed.



# Evaluation:

This can further be validated through check the RMSE to check which recommender model/ metric (Cosine similarity, Pearsons’s correlation) is giving the best results.

# 9.Final Insights and Campaign Ideas

Through thorough study and analysis on customer behavior through the we can implement the following campaigns to drive the sales forward.

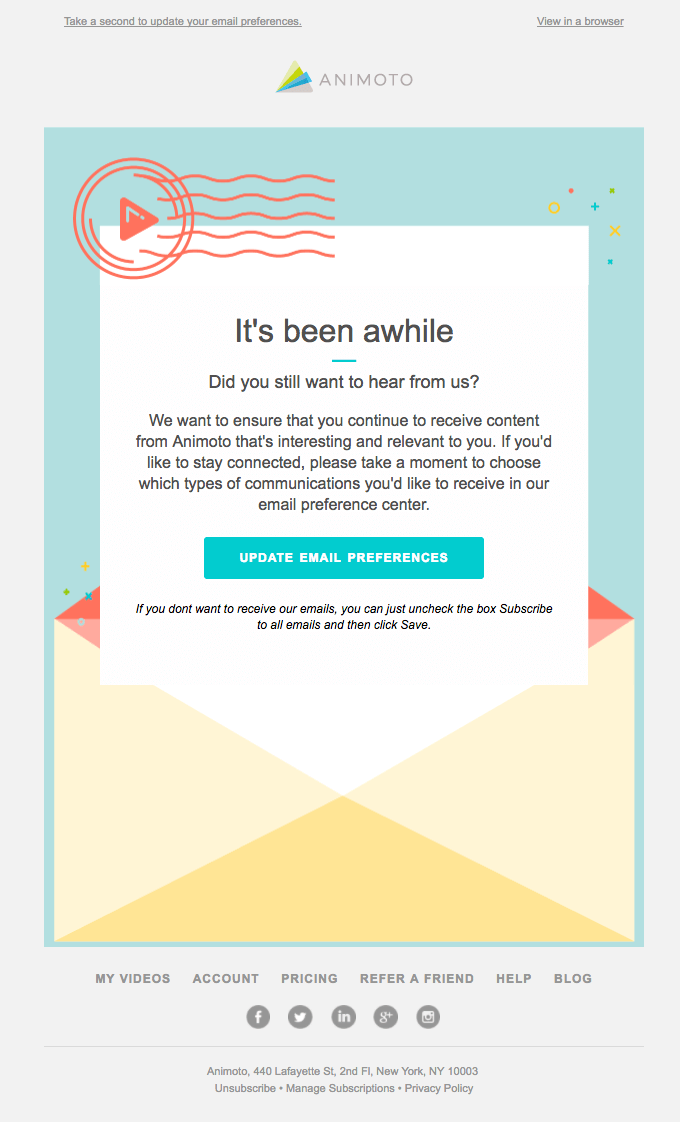
# 9.1 Reactivation:

Reactivation (or re-engagement) is a technique used by marketers to reach out to those people who have previously expressed interest, purchased, or otherwise engaged with their company but have since “gone dark” or disengaged. The goal is to encourage them to become active again.

Typically, the most common (although not used enough, in my opinion) method for reactivation is via email.

Sending a short series of engaging emails over a period of time with a message showing that you miss and care about winning back the customer’s business. Perhaps end the engagement series with an incentive, like a discount mentioned in the subject line. Of note: subject lines that included a discount in the form of an exact dollar amount were nearly twice as successful as subject lines that included a discount in the form of a percentage.

Re-engagement emails can reduce list churn rates and boost incremental engagement. In terms of deliverability, re-engaging your inactive subscribers can enhance or at least help you maintain inbox placement rates for your emails.

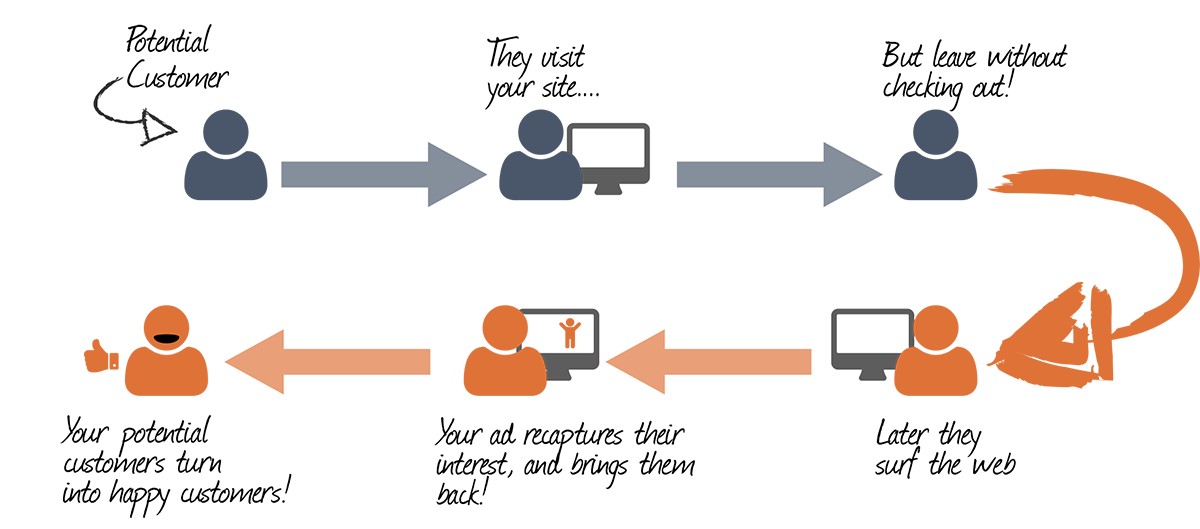


# 9.2 Retargeting:

Retargeting, also known as remarketing, is a form of online advertising that can help you keep your brand in front of bounced traffic after they leave your website. For most websites, only 2% of web traffic converts on the first visit. Retargeting is a tool designed to help companies reach the 98% of users who don’t convert right away.

Retargeting is a cookie-based technology that uses simple Javascript code to anonymously ‘follow’ your audience all over the Web.

Here’s how it works: you place a small, unobtrusive piece of code on your website (this code is sometimes referred to as a pixel). The code, or pixel, is unnoticeable to your site visitors and won’t affect your site’s performance. Every time a new visitor comes to your site, the code drops an anonymous browser cookie. Later, when your cookied visitors browse the Web, the cookie will let your retargeting provider know when to serve ads, ensuring that your ads are served to only to people who have previously visited your site.

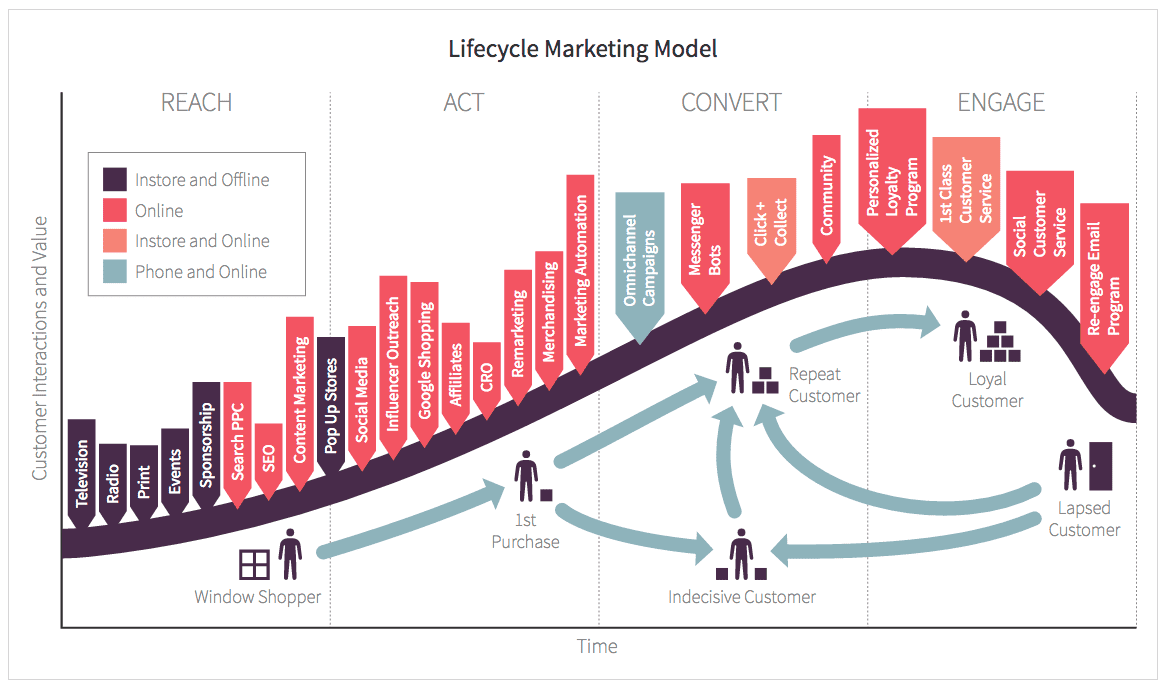


Retargeting works best in conjunction with inbound and outbound marketing or demand generation. Strategies involving content marketing, AdWords, and targeted display are great for driving traffic, but they don’t help with conversion optimization. Conversely, retargeting can help increase conversions, but it can’t drive people to your site. Your best chance of success is using one or more tools to drive traffic and retargeting to get the most out of that traffic.

# 9.3 Life Cycle:

Renewed focus by marketers on prospect and customer engagement, conversion and retention has led to the appearance of a new phrase in the digital marketing lexicon: **Customer lifecycle marketing**, or sometimes just **lifecycle marketing**.

Life time revenue generated from loyal customers.



# 9.4 Cross selling:

Cross selling is the process of selling a different product or service to a customer to increase the value of a sale For example, consider a fast-food employee asking if you want fries with your burger.

Bundling of products that are frequently bought together or some products that can be bundled based on the season of the market.



### 10. LIMITATION AND FUTURE ENHANCEMENT

### 10.1 LIMITATIONS:

The first disadvantage of marketing in general is the cost. Adverting and marketing costs money. If you don’t do the proper research then you might end up throwing money away. Wasting marketing efforts by targeting the wrong audience using an inappropriate medium would be a serious and costly mistake. So it is important to do your research beforehand and keep your costs to a minimum.

As well as the financial cost, marketing your business will require investment of time. Researching the appropriate marketing strategy, designing and writing the adverts, getting them published, dealing with any response. It’s important to spend time keeping track of how successful or not your marketing campaign is. A potential disadvantage of marketing here is the risk of time wasted for an unsuccessful campaign.

Research shows that people in general have to see a piece of information between 3 and 30 times before it sinks in. So the obvious disadvantage of marketing here is the fact that our marketing campaign will need to be ongoing and consistent. Increasing costs and time spent on it. This is where drip marketing comes in.

### 10.2 MODEL ENHANCEMENT

### The future enhancement can be done by using hyperparameter tuning and cross validation techniques to reduce the RMSE values. Each model has certain parameters that can tuned to get better results. Machine learning involves predicting and classifying data and to do so, you employ various machine learning models according to the dataset. Machine learning models are parameterized so that their behavior can be tuned for a given problem. These models can have many parameters and finding the best combination of parameters can be treated as a search problem.

### HYPERPARAMETER TUNING

The best way to do a deeper customer segmentation and trends extractions is by tuning our models to our liking for example using aglormative clustering and deciding the geight of the dendogram for clustering. This provides far better flexibility and deeper insights on the customers.

Same with Recommender System we can use different metrics like correlation, cosine similarity to

choose the nearest neighbors in products or customers depending on the task at hand.

### 11.CONCLUSION

The Insights drawn from data can be implemented through necessary campaigns. This will help boost the sales and improve customer engagement and loyalty. Machine learning Models like the recommender systems can be deployed on the website to improve sales and further provide customer engagement by providing a variety of data driven accurate choices. Clustering can also help with customer study and analysis through which we can better understand our customer’s mindsets and target accordingly .Overall Marketing campaigns will have a good impact in boosting the sales and driving revenue if implemented consistently.

### 12. REFERENCES

<https://towardsdatascience.com/intro-to-recommender-system-collaborative-filtering-> [64a238194a26](https://towardsdatascience.com/intro-to-recommender-system-collaborative-filtering-64a238194a26)

<https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a>

<https://github.com/nikadeap/RFM-Analysis>