Customer Segmentation

Rudrajit Bhattacharyya
Cohort Zanskar
AlmaBetter

Abstract:

Customer segmentation is a way to split the customers into groups based on certain characteristics that those customers share. Customer segmentation will allow marketers to better tailor their marketing efforts to various audience subsets. Our task here is to identify major customer segments on a transactional dataset which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for an UK based and registered non-store online retail. The company mainly sells unique all occasion gifts. There are 6 types of customer segmentation models present and we will be using a RFM model (Recency, Frequency and Monetary). RFM is a method often used to identify customers based on the recency of their last purchase, the total number of purchases they have made and the amount of money they have spent. This is often used to identify High Value Customers. K-Means clustering was implemented to differentiate the clusters or segments and to determine the optimal value of 'K', we used the silhouette score method and the elbow method.

Problem Statement:

In this project, our task is to identify major customer segments on a transactional data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for an UK based and registered non-store online retail. The company mainly sells unique all occasion gifts. Many customers of the company are wholesalers.

Customer segmentation is a way to split customers into groups based on certain characteristics that those customers share. Customer segmentation will allow marketers to better tailor their marketing efforts to various audience subsets. The dataset contains 541909 records of transactions with 8 features.

Data Description:

The dataset contains 541909 records and 8 features which consists of:

- **InvoiceNo:** Invoice number. Nominal, a 6 digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
- **StockCode:** Product code. Nominal, a 5 digit integral number uniquely assigned to each distinct product.
- **Description:** Product name. Nominal
- **Quantity:** The quantities of each product per transaction. Numeric.
- **InvoiceDate:** Invoice date and time. Numeric, the day and time when each transaction was generated.
- **UnitPrice:** Unit Price. Numeric, product price per unit in sterling.
- **CustomerID:** Customer number. Nominal, a 5 digit integral number uniquely assigned to each customer.
- **Country:** Country name. Nominal, the name of the country where each customer resides.

Introduction:

The objective of this project is to identify major customer segments on a transactional data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for an UK based and registered non-store online retail. The company mainly sells unique all occasion gifts.

Customer segmentation is a way to split customers into groups based on certain characteristics that those customers share. All customers share the common need of a product or service, but beyond that, there are distinct demographic differences and they tend to have additional socio-economic, lifestyle, or other behavioral differences that can be useful to the organization.

Segmentation allows marketers to better tailor their marketing efforts to various audience subsets. Those efforts can relate to both communications and product development. Specifically, segmentation helps a company:

- Create and communicate targeted marketing messages that will resonate with specific groups of customers, but not with others (who will receive messages tailored to their needs and interests, instead).
- Select the best communication channel for the segment, which might be email, social media posts, radio advertising, or another approach, depending on the segment.
- Identify ways to improve products or new product or service opportunities.
- Establish better customer relationships.
- Test pricing options.
- Focus on the most profitable customers.
- Improve customer service.
- Upsell and cross-sell other products and services.

Common customer segmentation models range from simple to very complex and can be used for a variety of business reasons. Common segmentations include:

1. **Demographic:**
   At a bare minimum, many companies identify gender to create and deliver content based on that customer segment. Similarly, parental status is another important segment and can be derived from purchase details, asking more information from customers, or acquiring the data from a 3rd party.
2. **Recency, frequency, monetary (RFM):**
   RFM is a method used often in the direct mail segmentation space where you identify customers based on the recency of their last purchase, the total number of purchases they have made (frequency) and the amount they have spent (monetary). This is often used to identify your High-Value Customers (HVCs).
3. **High-value customer (HVCs):**
   Based on an RFM segmentation, any business, regardless of sector or industry, will want to know more about where HVCs come from and what characteristics they share so you can acquire more of them.
4. **Customer status:**
   At a minimum, most companies will bucket customers into *active* and *lapsed*, which indicates when the last time a customer made a purchase or engaged with you. Typical non-luxury products consider active customers to be those who have purchased within the most recent 12 months. Lapsed customers

would those who have not made a purchase in the last 12 months. Customers may be bucketed even further based on the time period in that status, or other characteristics.

5. **Behavioral:**
   Past observed behaviors can be indicative of future actions, such as purchasing for certain occasions or events, purchasing from certain brands, or significant life events like moving, getting married, or having a baby. It's also important to consider the reasons a customer purchases your product/service and how those reasons could change throughout the year(s) as their needs change.
6. **Psychographic:**
   Psychographic customer segmentation tends to involve softer measures such as attitudes, beliefs, or even personality traits. For example, survey questions that probe how much someone agrees or disagrees with a statement are typically seeking to classify their attitudes or perspectives towards certain beliefs that are important to your brand.

There are several benefits of implementing customer segmentation including informing marketing strategy, promotional strategy, product development, budget management, and delivering relevant content to your customers or prospective customers. Let's look at each of the benefits in a bit more depth.

1. **Marketing strategy:**
   Customer segmentation can help inform your overall marketing strategy and messaging. As you learn the attributes of your best customers, how they are alike, and what is important to them, you can leverage that information in messaging, creative development, and channel selection.
2. **Promotion strategy:**
   An overall promotion strategy (i.e., our customers are deal seekers, therefore we should offer frequent deals) for sending promotions for specific segments can be made better with information from a broad customer segmentation scheme. You may find that certain cohorts of customers don't require discounts when you use certain messaging, thereby saving you from having to offer a discount for those groups at all.
3. **Budget efficiency:**
   Most companies do not have unlimited marketing budgets, so being precise

about how and where you spend is important. You could, as an example, target similar customers to segments of high value or those most likely to convert to get the most return from your marketing investment.

4. **Product development:**
   The more customers you acquire, the more you learn about what is important to them, what features they want, and which customers are the most valuable. Your company can use these insights to prioritize product features that either appeal to the most customers, those categorized as high-value customers, or other characteristics that make sense for your industry.
5. **Customers demand relevance:**
   Whether it's D2C, B2B, Millennials or GenZ; it seems that there is a study or resource on every possible group of customers stating that relevant content is important to them. These customer segments are more likely to respond, buy, and respect the brand and feel connected if provided with relevant content. By performing some level of segmentation, you can ensure that the messages you are delivering via email, on site, through digital ads, or other methods are targeted and relevant to the individual seeing it. It is almost counter-intuitive to the hyper vigilance of data privacy to use so many pieces of data in this way, but with so many marketing messages coming at people today, no one has time for something that isn't relevant to them.

Exploratory Data Analysis:

- **Data Cleaning**:

  Data cleaning is one of the most time consuming aspects of data analysis. Formatting issues, missing values, duplicated rows, spelling mistakes, and so on could all be present in a dataset. These difficulties make data analysis difficult, resulting in inappropriate results. Missing or duplicate data may exist in a dataset for a number of different reasons. Sometimes, missing or duplicate data is introduced as we perform cleaning and transformation tasks such as combining data, reindexing data, and reshaping data. Other times, it exists in the original dataset for reasons such as user input error or data storage or conversion issues.

We need to fill in the missing values because most of the machine learning models that we want to use will provide an error if we pass NaN values into it. The easiest way is to just fill them up with 0, but this can reduce our model accuracy significantly. There are many methods available for filling up missing values but for choosing the best method, we need to understand the type of missing value and its significance.

The dataset we have in our hand has 5268 duplicated records and quite a few missing values in two columns. Thus, we decided to drop these values before proceeding further.

```
[ ] # check for duplicated records
    cust_df.duplicated().sum()

    5268

[ ] # check for missing values
    cust_df.isnull().sum()

    InvoiceNo          0
    StockCode          0
    Description     1454
    Quantity           0
    InvoiceDate        0
    UnitPrice          0
    CustomerID    135080
    Country            0
    dtype: int64
```

- There are 5268 duplicated records present in the data and few missing values present in Description and CustomerID columns.

In Addition to the missing values and duplicated records we had to clean a few more features before moving into feature engineering or visualizations. The dataset contains 8872 records where the orders stand canceled and the quantity contains a negative value so we will drop these values and consider the active orders only for our segmentation purpose.

```
[ ] # check for order cancellations
    cust_df['InvoiceNo'] = cust_df['InvoiceNo'].astype('str')
    cust_df[cust_df['InvoiceNo'].str.startswith('C')]
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
141	C536379	D	Discount	-1	2010-12-01 09:41:00	27.50	14527.0	United Kingdom
154	C536383	35004C	SET OF 3 COLOURED FLYING DUCKS	-1	2010-12-01 09:49:00	4.65	15311.0	United Kingdom
235	C536391	22556	PLASTERS IN TIN CIRCUS PARADE	-12	2010-12-01 10:24:00	1.65	17548.0	United Kingdom
236	C536391	21984	PACK OF 12 PINK PAISLEY TISSUES	-24	2010-12-01 10:24:00	0.29	17548.0	United Kingdom
237	C536391	21983	PACK OF 12 BLUE PAISLEY TISSUES	-24	2010-12-01 10:24:00	0.29	17548.0	United Kingdom
...	...	...	...	...	...	...	...	...
540449	C581490	23144	ZINC T-LIGHT HOLDER STARS SMALL	-11	2011-12-09 09:57:00	0.83	14397.0	United Kingdom
541541	C581499	M	Manual	-1	2011-12-09 10:28:00	224.69	15498.0	United Kingdom
541715	C581568	21258	VICTORIAN SEWING BOX LARGE	-5	2011-12-09 11:57:00	10.95	15311.0	United Kingdom
541716	C581569	84978	HANGING HEART JAR T-LIGHT HOLDER	-1	2011-12-09 11:58:00	1.25	17315.0	United Kingdom
541717	C581569	20979	36 PENCILS TUBE RED RETROSPOT	-5	2011-12-09 11:58:00	1.25	17315.0	United Kingdom

8872 rows × 8 columns

After investigating the describe() output we found out that there are many records where the unit price is zero which cannot be true for any store or business. No store gives out items for free and hence we investigated each record and decided to consider the records where the unit price is greater than zero for our segmentation purpose.

- **Feature Engineering:**

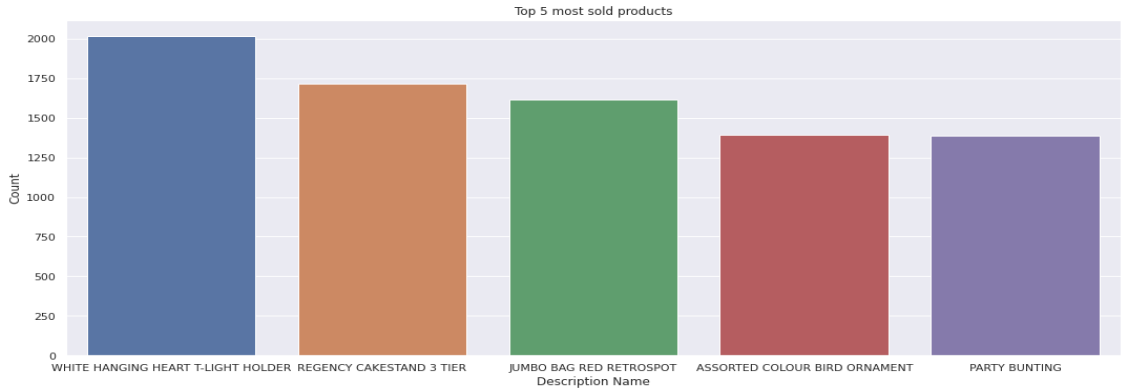
  As soon as we had a clean and ready dataset, we decided to create some new features which will help us in segmenting different customer segments on the basis of segmentation model. Feature engineering is the process of selecting, manipulating and transforming raw data into features that can be used in machine learning. In order to make machine learning work well on new tasks, it might be necessary to design and train better features.

  Feature engineering is a machine learning technique that leverages data to create new variables that aren't in the training set. It can produce new features for both supervised and unsupervised learning, with the goal of simplifying and speeding up data transformations while also enhancing model accuracy.

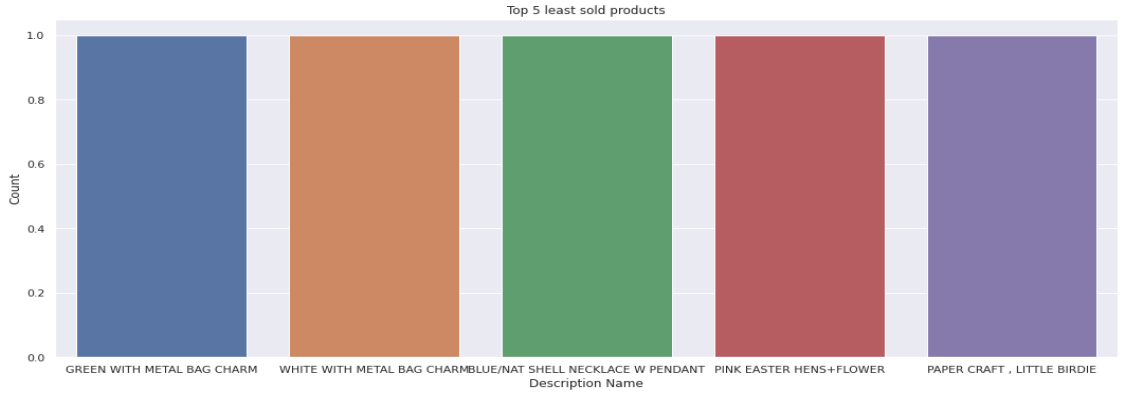
InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	day	year	month_num	day_num	hour	minute	month	TotalAmount
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom	Wednesday	2010	12	1	8	26	December	15.30
536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	Wednesday	2010	12	1	8	26	December	20.34
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom	Wednesday	2010	12	1	8	26	December	22.00
536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	Wednesday	2010	12	1	8	26	December	20.34
536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	Wednesday	2010	12	1	8	26	December	20.34

- **Data Visualization/Exploration**:

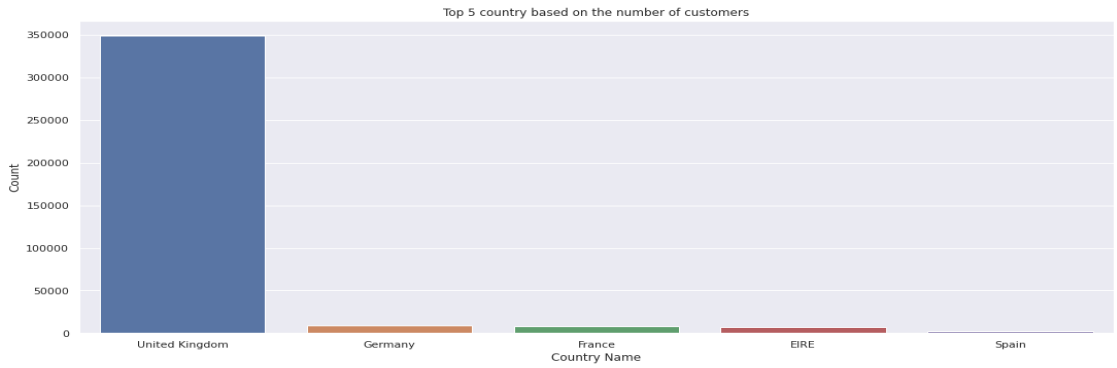
  - The top 5 most sold products are:



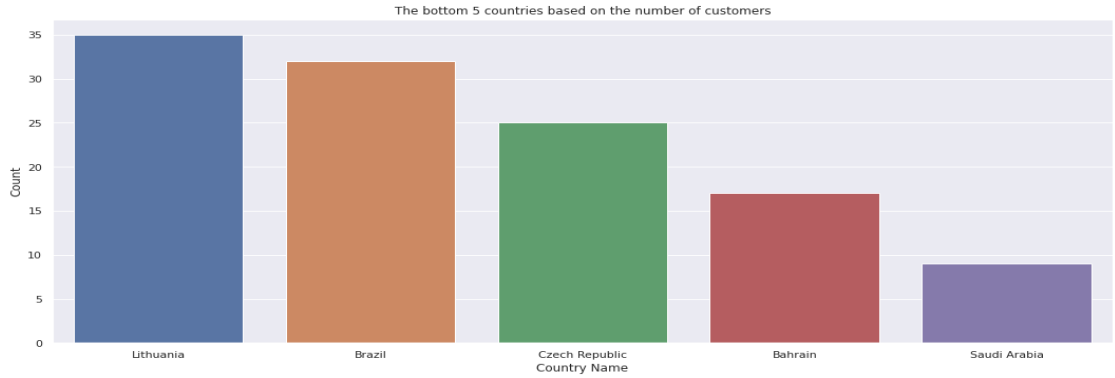
- The bottom 5 least sold products are:



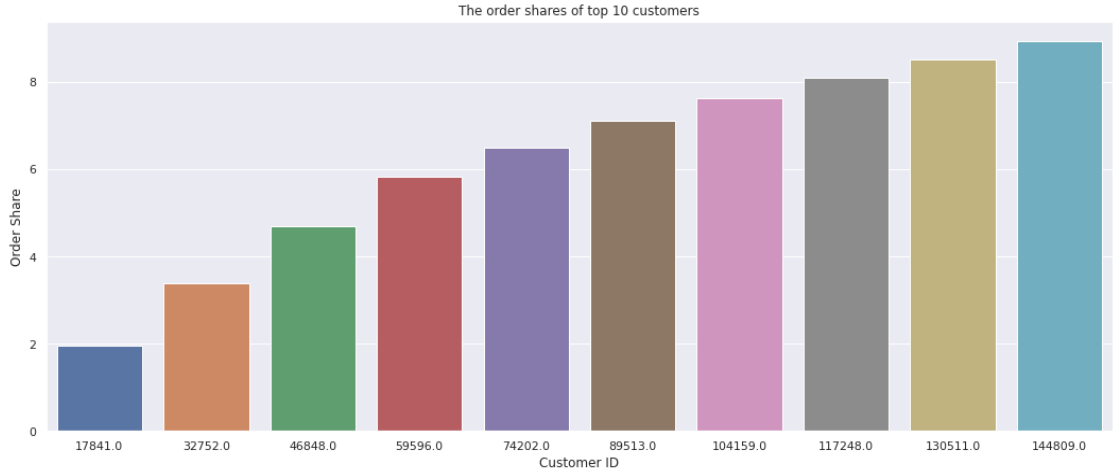
- The top 5 countries where the most customers belong or reside:



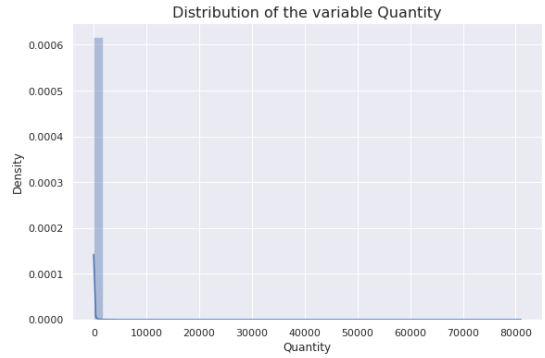
- The bottom 5 countries where the least customers belong or reside:

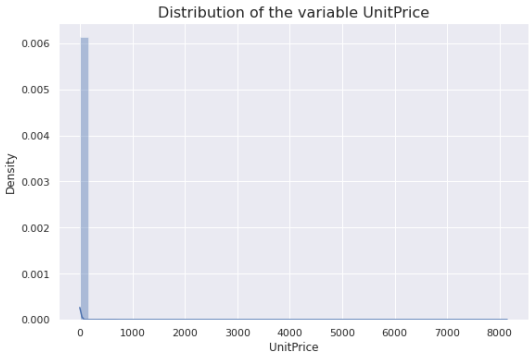


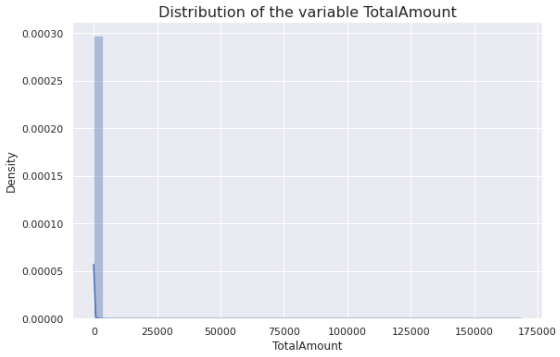
- There are 4338 unique customers but interestingly only 10 customers are responsible for approximately 9% of the orders. We can infer that these customers are wholesalers.

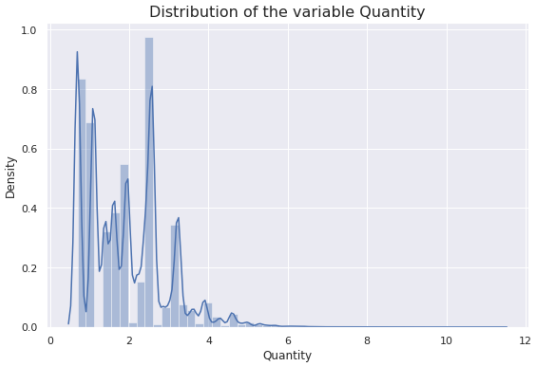


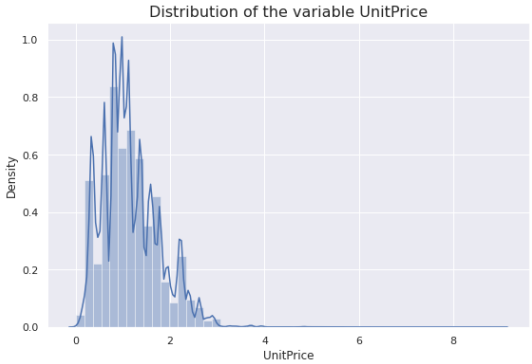
- Distribution of all the numerical features are right skewed and thus we had to apply log transformation on these features to bring them near normal.

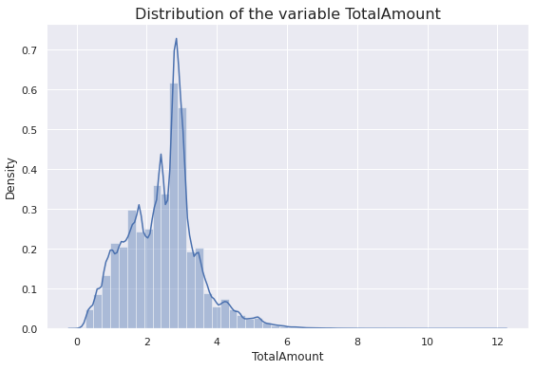




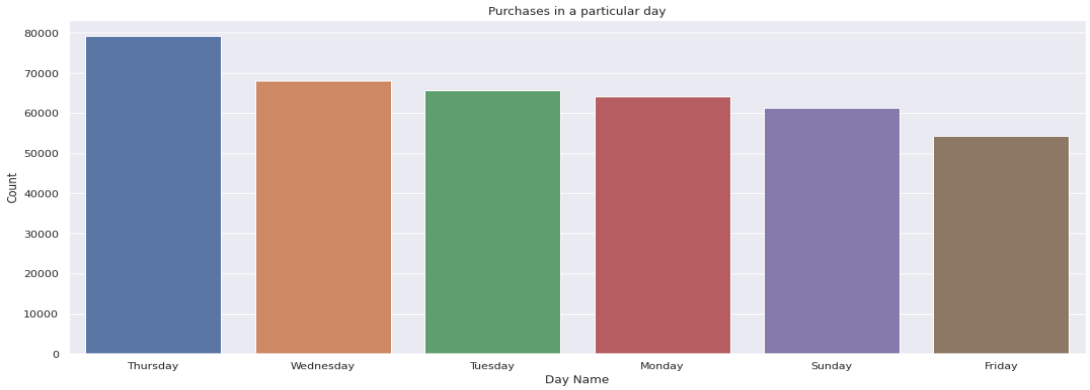




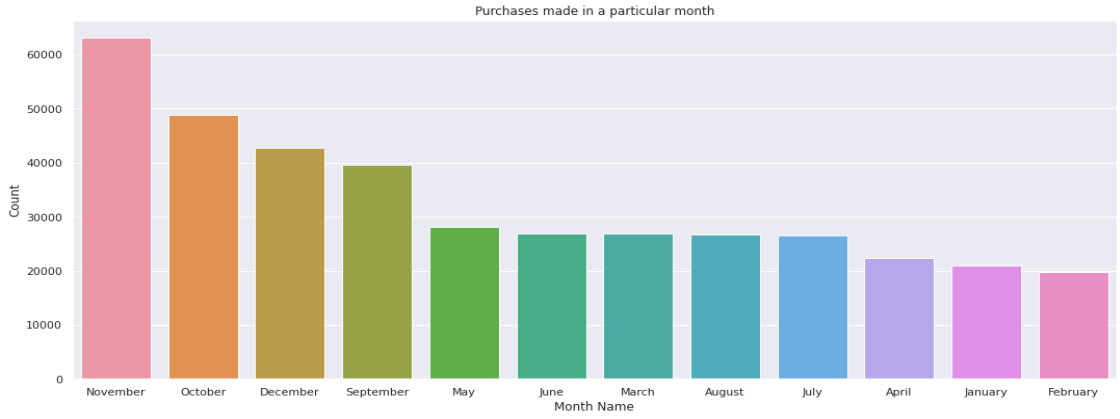




- Most of the customers had made a purchase on Thursday followed by Wednesday and the least number of purchases on Friday.



- The highest number of purchases has occurred during the festive months of October to December and the least number of purchases has occurred during the initial months of January and February.



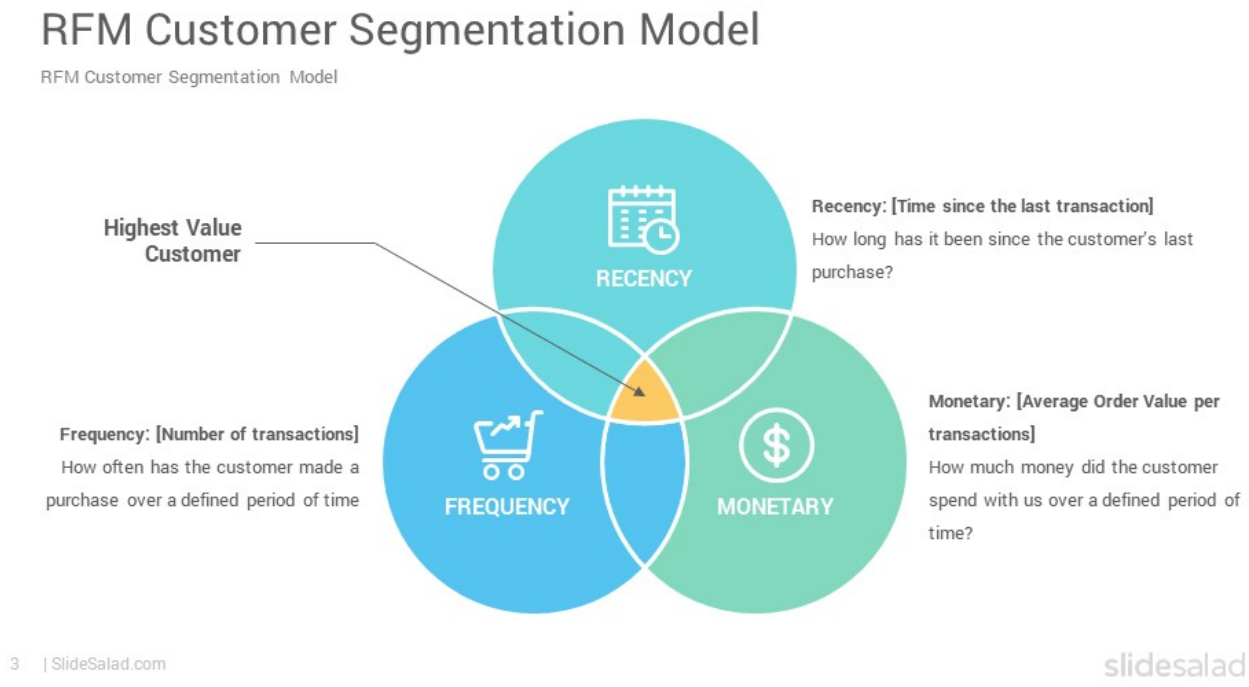
- Interestingly the time of the day in which the most number of purchases has taken place is during the afternoon and the least number of purchases during evening.

RFM Segmentation & Analysis:

Recency, frequency and monetary value is a marketing analysis tool used to identify a company's or an organization's best customers by measuring and analyzing spending habits.
The RFM model is based on three quantitative factors:

- Recency: How recently a customer has made a purchase
- Frequency: How often a customer makes a purchase
- Monetary: How much money a customer spends on purchases

RFM analysis numerically ranks a customer in each of these three categories, generally on a scale of 1 to 5. The best customers would receive a top score in every category. These three RFM factors can be used to reasonably predict how likely or unlikely it is that a customer will do business again with a firm or company. RFM analysis allows a comparison between potential customers or clients. It gives organizations a sense of how much revenue comes from repeat customers vs new customers, and which levers they can pull to try to make customers happier so they become repeat purchasers. Despite the useful information that is acquired through RFM analysis, firms must take into consideration that even the best customers will not want to be over solicited, and the lower ranking customers may be cultivated with additional marketing efforts.



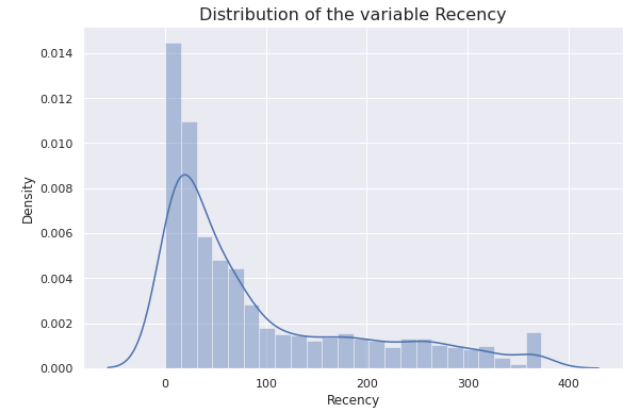
Performing RFM segmentation, step by step:

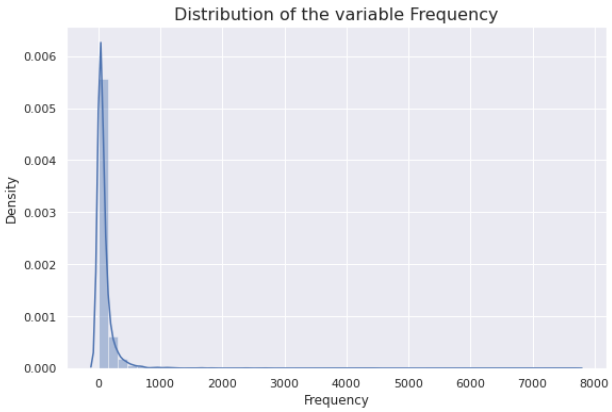
- **Step 1:** The first step in building an RFM model is to assign Recency, Frequency and Monetary values to each customer. The raw data for doing this, which should be readily available in the company's CRM or transactional databases, can be compiled in an Excel spreadsheet or database.
- **Step 2:** The second step is to divide the customer list into tiered groups for each of the three dimensions (R,F,M) using Excel or another tool. Unless using specialized software, it's recommended to divide the customers into four tiers for each dimension, such that each customer will be assigned to one tier in each dimension.
- **Step 3:** The third step is to select groups of customers to whom specific types of communications will be sent, based on the RFM segments in which they appear.
- **Step 4:** The fourth step actually goes beyond the RFM segmentation itself, crafting specific messaging that is tailored for each customer group. By focusing on the behavioral patterns of particular groups, RFM marketing allows marketers to communicate with customers in a much more effective manner.

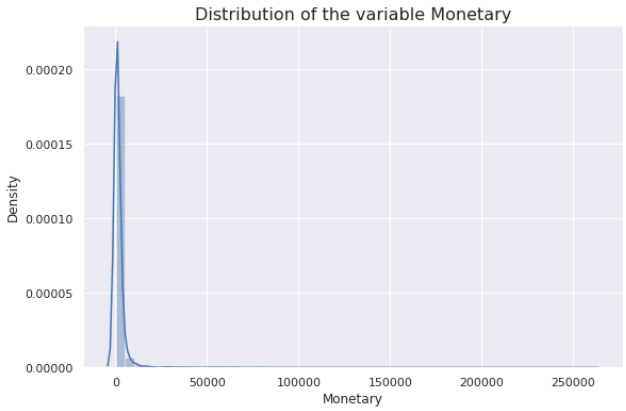
We calculated the RFM scores and then the RFM rank for each customer as shown below:

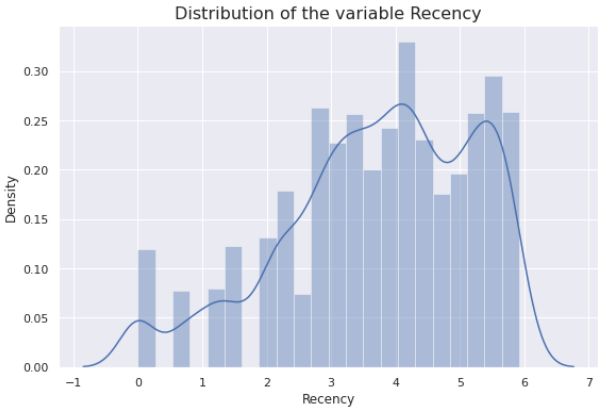
	CustomerID	Recency	Frequency	Monetary	R	F	M
0	12346.0	325	1	77183.60	4	4	1
1	12747.0	2	103	4196.01	1	1	1
2	12748.0	0	4412	33053.19	1	1	1
3	12749.0	3	199	4090.88	1	1	1
4	12820.0	3	59	942.34	1	2	2

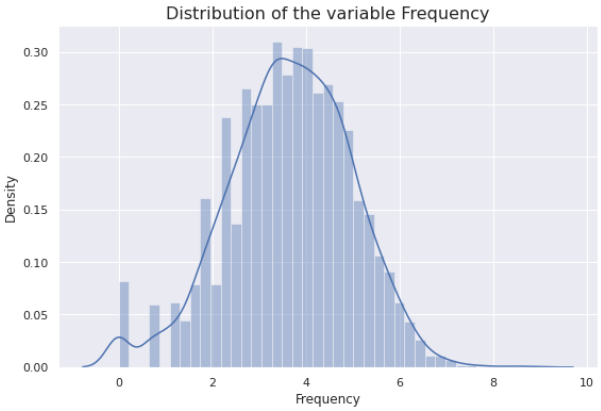
The distribution of Recency, Frequency and Monetary is again right skewed and log transformation was applied to bring their distribution near normal.

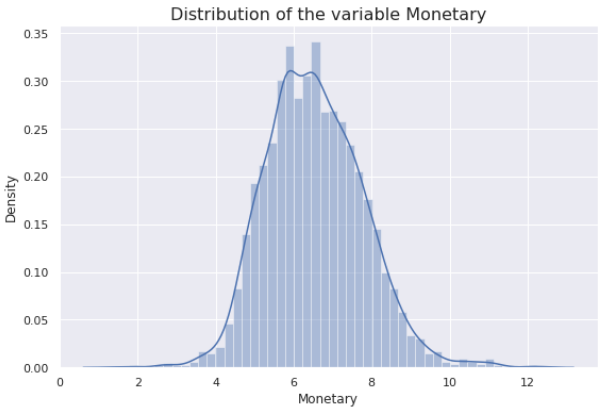












Clustering Models:

Clustering can be considered the most important unsupervised learning problem, so as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be “the process of organizing objects into groups whose members are similar in some way”. A cluster is therefore a collection of objects which are similar between them and are dissimilar to the objects belonging to other clusters.

- **K-Means Clustering:**

  The K-means clustering algorithm computes centroids and repeats until the optimal centroid is found. It accomplishes this using a simple conception of what the optimal clustering looks like:
  - The cluster center is the arithmetic mean of all the points belonging to the cluster.

- Each point is closer to its own cluster center than to other cluster centers.

These two assumptions are the basis of the K-means model. It is presumptively known how many clusters there are. It is also known as the flat clustering algorithm. The number of clusters found from data by the method is denoted by the letter 'K' in K-means.

In this method, data points are assigned to clusters in such a way that the sum of the squared distances between the data points and the centroid is as small as possible. It is essential to note that reduced diversity within clusters leads to more identical data points within the same cluster. K-means implements the Expectation-Maximization strategy to solve the problem. The expectation step is used to assign data points to the nearest cluster, and the maximization step is used to compute the centroid of each cluster.

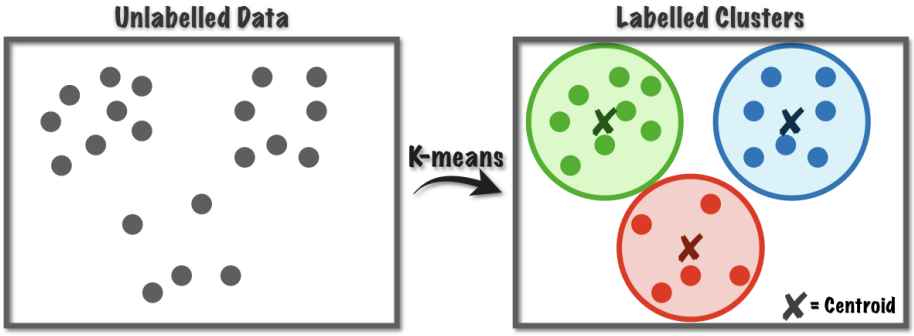
Algorithm 1 *k*-means algorithm

1: Specify the number k of clusters to assign.
2: Randomly initialize k centroids.
3: **repeat**
4: **expectation:** Assign each point to its closest centroid.
5: **maximization:** Compute the new centroid (mean) of each cluster.
6: **until** The centroid positions do not change.

The quality of the cluster assignments is determined by computing the sum of squared error (SSE) after the centroids converge, or match the previous iteration's assignment. The SSE is defined as the sum of the squared euclidean distances of each point to its closest centroid. Since this is a measure of error, the objective of K-means is to try to minimize this value.

There are a few issues to be aware of when using the expectation-maximization algorithm:

- The globally optimal result may not be achieved
- The number of clusters must be selected beforehand
- K-means is limited to linear cluster boundaries
- K-means can be slow for large number of samples

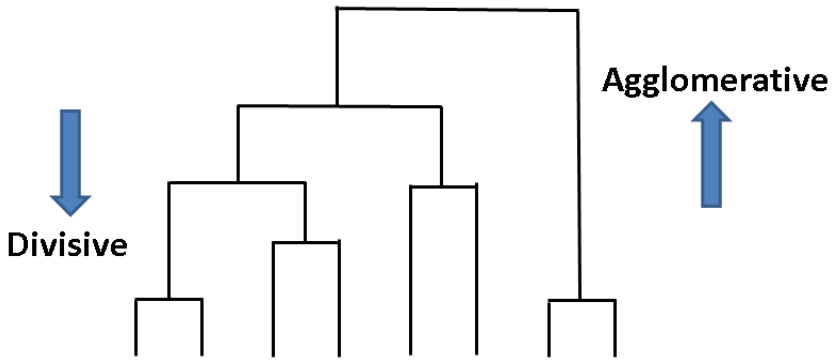


- **Hierarchical Clustering:**

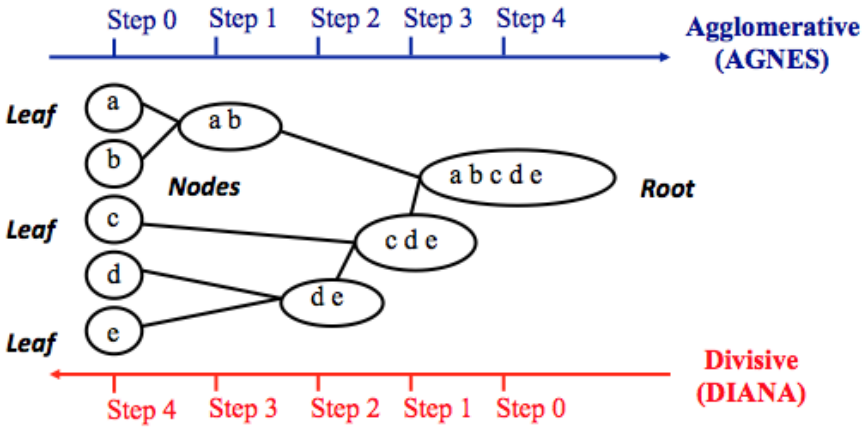
  Hierarchical clustering determines cluster assignments by building a hierarchy. This is implemented by either a bottom-up or a top-down approach:

  - Agglomerative clustering is the bottom-up approach. It merges the two points that are the most similar until all points have been merged into a single cluster.
  - Divisive clustering is the top-down approach. It starts with all points as one cluster and splits the least similar clusters at each step until only single data points remain.

  These methods produce a tree-based hierarchy of points called a dendrogram. Similar to partitional clustering, in hierarchical clustering the number of clusters (*k*) is often predetermined by the user. Clusters are assigned by cutting the dendrogram at a specified depth that results in *k* groups of smaller dendrograms.



Unlike many partitional clustering techniques, hierarchical clustering is a deterministic process, meaning cluster assignments won't change when you run an algorithm twice on the same input data.



The strengths of hierarchical clustering methods include the following:

- They often reveal the finer details about the relationships between data objects.
- They provide an interpretable dendrogram.

The weaknesses of hierarchical clustering methods include the following:

- They're computationally expensive with respect to algorithm complexity.
- They're sensitive to noise and outliers.

Evaluation Metrics:

- **Silhouette Method:**

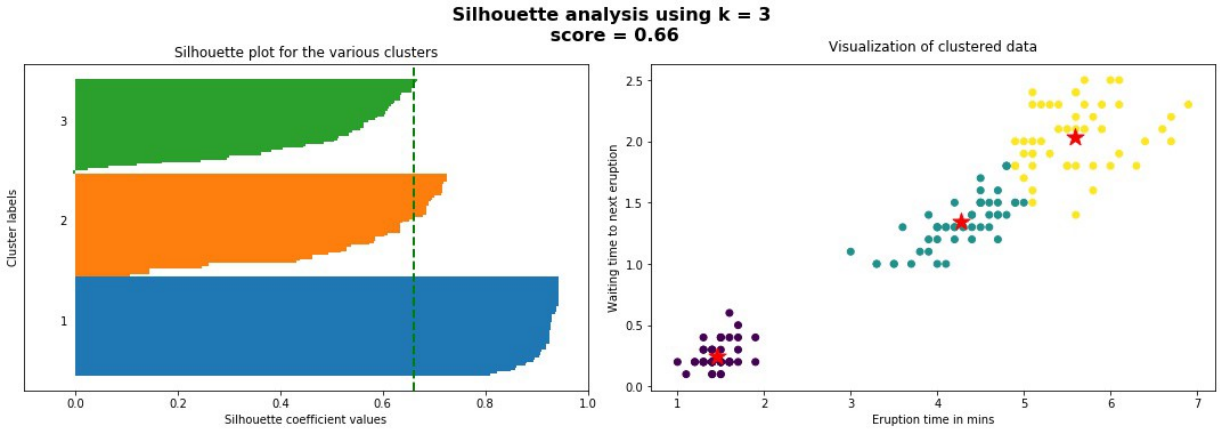
Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K-means in terms of how well samples are clustered with other samples that are similar to each other. The silhouette score is calculated for each sample of different clusters. To calculate the silhouette score for each

observation/data point, the following distances need to be found out for each observation belonging to all the clusters:

- Mean distance between the observation and all other data points in the same cluster. This distance can also be called a mean intra-cluster distance. The mean distance is denoted by a.
- Mean distance between the observation and all other data points of the next nearest cluster. This distance can also be called a mean nearest-cluster distance. The mean distance is denoted by b.

$$Silhouette - score = \frac{b_i - a_i}{\max(bi, a_i)}$$

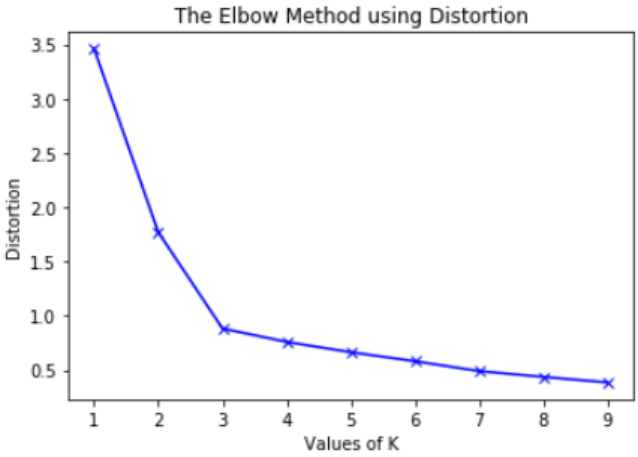
The value of silhouette score varies from -1 to 1. If the score is 1, the cluster is more dense and well separated than other clusters. A value near 0 represents overlapping clusters with samples very close to the decision boundary of the neighboring clusters. A negative score [-1,0] indicates that the samples might have been assigned to the wrong clusters.



- **Elbow Method:**

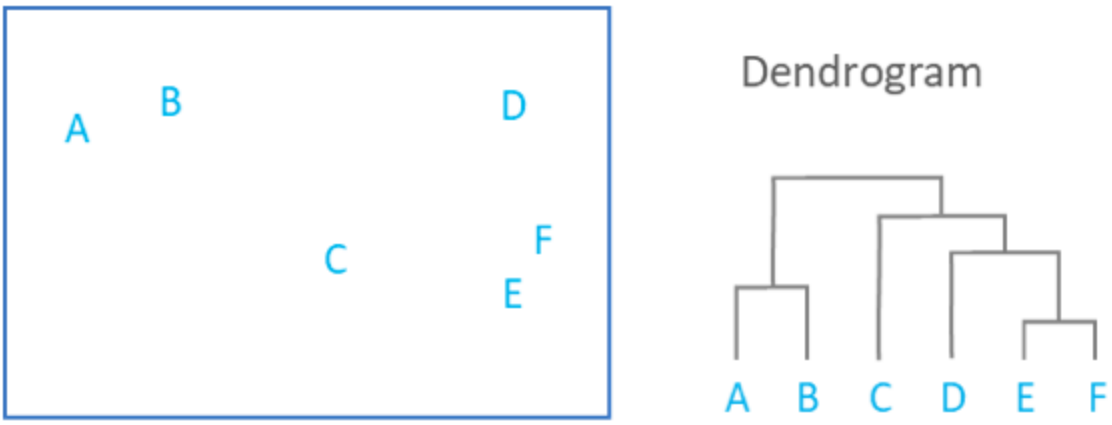
The KElbowVisualizer implements the elbow method to help data scientists select the optimal number of clusters by fitting the model with a range of values for K. If the line chart resembles an arm, then the elbow is a good indication that the underlying model fits best at that point. In the visualizer the elbow will be

annotated with a dashed line. By default, the scoring parameter metric is set to distortion, which computes the sum of squared distances from each point to its assigned center. However, two other metrics can also be used with the KElbowVisualizer - silhouette and calinski_harabasz. The silhouette score calculates the mean silhouette coefficient of all samples, while the calinski_harabasz score computes the ratio of dispersion between and within clusters.



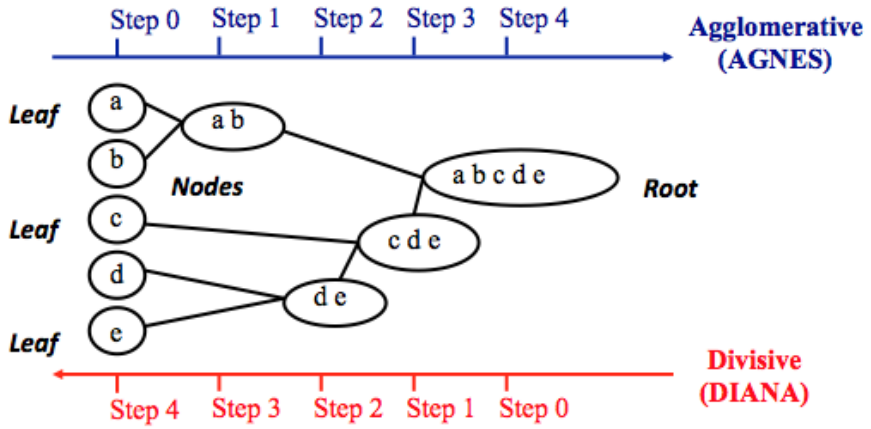
- **Dendrogram:**

  A dendrogram is a diagram that shows the hierarchical relationship between objects. It is most commonly created as an output from hierarchical clustering. The main use of a dendrogram is to work out the best way to allocate objects to clusters. The dendrogram below shows the hierarchical clustering of six observations shown on the scatterplot to the left.



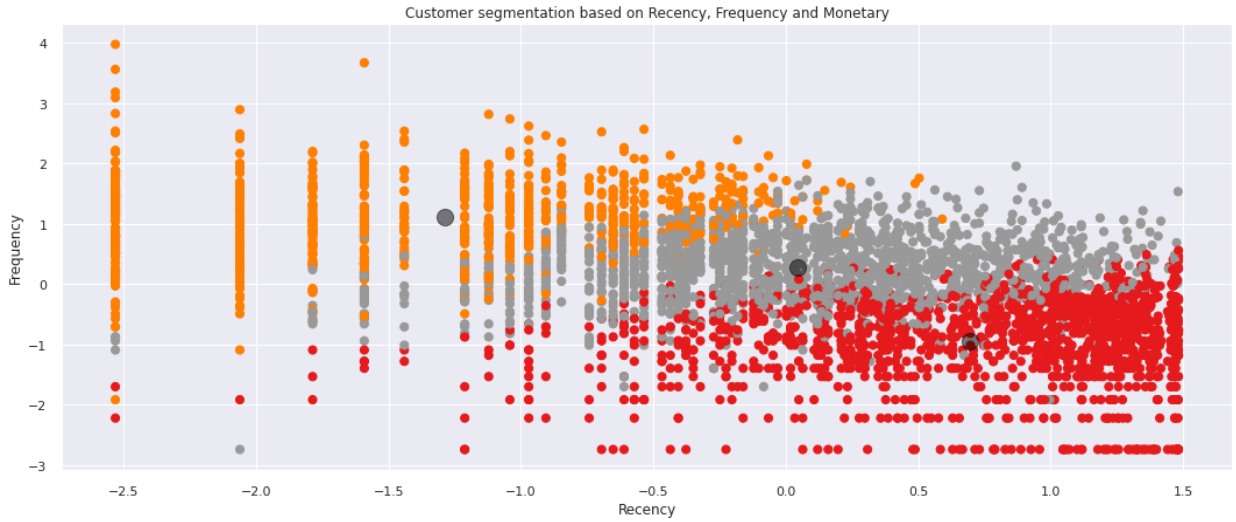
The key to interpreting a dendrogram is to focus on the height at which any two objects are joined together. In the example above, we can see that E and F are most similar, as the height of the link that joins them together is the smallest. The next two most similar objects are A and B. In the dendrogram above, the height of the dendrogram indicates the order in which the clusters were joined. It shows us that there is a big difference between clusters of A and B vs that of C, D, E and F.

It is important to appreciate that the dendrogram is a summary of the distance matrix, and as occurs with most summaries, information is lost. For example, the dendrogram suggests that C and D are much closer to each other than is C to B, but the original data shows us that this is not true. To use some jargon, a dendrogram is only accurate when data satisfies the ultrametric tree inequality, and this is unlikely for any real world data.



Final Model:

K-means clustering was applied on RFM data as the final model and to determine the best 'K' or the optimal number of clusters, we used silhouette score and elbow method. The best value of 'K' was found out to be 3 and the final three clusters or customer segments were formed as follows:



- **Loyal Customers (Orange Points):** These are the most frequent buyers with significant monetary contribution.
- **Casual Customers (Gray Points):** These buyers are not so frequent or recent but contribute moderately in terms of money.
- **New Customers (Red Points):** These are very recent buyers or the new customers who haven't yet contributed much in terms of money.

Thus, we have identified 3 major customer segments from our transactional dataset of a non-store online retail company.

Additional Model Comparison:

In Addition to the K-means clustering model implemented on RFM data, we have implemented K-means separately on Recency, Monetary and Frequency, Monetary data

along with hierarchical clustering on RFM data to compare the performances of each clustering method.

We have again used the silhouette score method and the elbow method for determining the best 'K' value for K-means clustering and a dendrogram for hierarchical clustering. A dendrogram is a diagram that shows the hierarchical relationship between objects. It is most commonly created as an output from hierarchical clustering. The main use of a dendrogram is to work out the best way to allocate objects to clusters.

SL No.	Model Name	Data	Optimal Number of Clusters	Cluster Interpretation
1	K-Means	RM	4	High valued customers, Moderate valued customers, Casual customers, New customers
2	K-Means	FM	3	New customers, Casual customers, Loyal high valued customers
3	K-Means	RFM	3	Loyal customers, Casual customers, New customers
4	Hierarchical clustering	RFM	2	Loyal customers, New customers

We can observe from the different clustering methods that the optimal number of clusters for each method is around 2 or 3. Thus, we can consider 3 as the optimal number of clusters in our final model of K-means on RFM data and identify the different customer segments.

Challenges:

- There were many duplicated records and missing values present.
- There were canceled orders present in the transactional dataset and along with that few records had zero as unit price which isn't possible in real life.
- No feature followed a normal distribution.
- Features had to be created to calculate RFM scores.
- Finding the optimal number of clusters through silhouette method or elbow method.
- Making sense of the clusters formed or derived.

Conclusion:

We have reached the end of our customer segmentation task where we had to identify major customer segments on a transactional dataset. Customer segmentation is a way to

split customers into groups based on certain characteristics that those customers share. Customer segmentation will allow marketers to better tailor their marketing efforts to various audience subsets.

The dataset contained 541909 records with 8 features. The dataset had a good number of duplicated records and missing values which was dropped before moving on to the visualizations. We discovered a few insights from EDA and created some new features before implementing clustering models. Feature engineering is a machine learning technique that leverages data to create new variables that aren't in the training set with the goal of simplifying and speeding up data transformations while also enhancing model accuracy.

Let us now summarize the insights generated from EDA:

- White hanging heart T-light holder was the most sold product and Green with metal bag charm was one of the least sold products.
- UK had the most number of customers which is pretty obvious because it is an UK based online retail company.
- Saudi Arabia followed by Bahrain had the least number of customers.
- There are 4338 unique customers and only 10 customers had an order share of approx 9% which implies that these customers could be wholesalers.
- The distribution of all the variables were heavily right skewed and log transformation was applied to bring them close to a normal distribution.
- Most of the customers had made a purchase on Thursday followed by Wednesday and the least number of purchases was made on Friday.
- The most purchases were made during the festive months of October to December and the least number of purchases was made during the initial months of January and February.
- Most of the people had made their purchases during the afternoon period and very less number of purchases during evening.

There are 6 types of customer segmentation models and we have decided to use the Recency, Frequency and Monetary method for segmenting our data. RFM is a method often used in the direct mail segmentation space where we identify customers based on the recency of their last purchase, the total number of purchases they have made and the amount they have spent. This is often used to identify high value customers. RFM analysis numerically ranks a customer in each of these three categories generally on a scale of 1 to

5. The best customer would receive a top score in every category. RFM analysis allows a comparison between potential customers or clients. It gives organizations a sense of how much revenue comes from repeat customers vs new customers.

After the RFM analysis was done, we decided to implement K-means clustering to identify different customer segments in our data. In this method data points are assigned to clusters in such a way that the sum of the squared distances between the data points and the centroid is as small as possible. We have used the silhouette score method and elbow method to help us determine the best possible 'K' or the optimal number of clusters. We found out that the optimal number of clusters is 3 for the K-means model implemented on RFM data.

The final three clusters or customer segments were formed are as follows:

- **Loyal Customers (Orange Points):** These are the most frequent buyers with significant monetary contribution.
- **Casual Customers (Gray Points):** These buyers are not so frequent or recent but contribute moderately in terms of money.
- **New Customers (Red Points):** These are very recent buyers or the new customers who haven't yet contributed much in terms of money.

Thus, we have identified 3 major customer segments from our transactional dataset of a non-store online retail company. We have additionally used K-means clustering on Recency, Monetary and Frequency, Monetary along with hierarchical clustering on RFM data to compare the clustering performances of each method.

References:

- Analytics Vidhya
- Statistics How To
- Machine Learning Mastery
- Scikit-learn
- Shopify Blog
- Towards Data Science
- freecodecamp