|  |
| --- |
| **Project Title: Online Retail Customer Segmentation**  **(Raavi sahu)**  **Data science trainee,**  **AlmaBetter, Bangalore**    **Abstract:**  **1. Problem Statement**  This project aims to identify major customer Customer segmentation is the process of segments on a transnational (extending or dividing your customers into sub-groups going beyond national boundaries based on shared features. transnational corporations.) data set which  contains all the transactions occurring Because you use on-site data to optimize between 01/12/2010 and 09/12/2011 for a advertising off-site, segmentation happens UK-based and registered non-store online after the fact, unlike customization and retail. The company mainly sells unique all- targeting.  occasion gifts. Many customers of the Because you need to build triggers so that company are wholesalers. your consumers see the advertisements Ecommerce customer segmentation divides when they arrive, you need to do your your clients into smaller groups who share a targeting and personalization before they common interest, making it easier to come arrive.up with offers and calls to action.  **Benefits of Customer**  Personalized marketing is built on being **Segmentation in e-Commerce** relevant to customers and responding appropriately to their activities. It all starts **Higher conversion rates:** When your with recognizing and working with your marketing communications are relevant to your customers' situations and interests, customers' variances in behavior, rather than they are more inclined to buy from you.against them.  **Long-term revenue from customers:**  ***Keywords:*** Increasing the lifetime value of each client  ***Exploratory Data Analysis, Creating RFM*** as much as feasible by making your ***model, model building: K-Means clustering*** marketing contextual is a key driver of ***with different methods, DBSCAN,*** long-term success. Getting new customers is more expensive than keeping the ones ***hierarchical clustering (Dendrogram).***  you already have. Customer Lifetime  Value (CLV) can be calculated and suitable retention methods performed for each group if you know what products your consumers buy over time. |

|  |
| --- |
| **Better insight into your customer base:**  Examining the success of certain client groups can give you valuable information into the health of your customer base, help you detect trends and patterns.  Information of their purchases could a bucket and then show customized provide more information about their advertisements and also review which purchasing habits.platforms are giving us better results.  When a customer makes a purchase, there is **2. Introduction** some information that is now stored. Such information is given below:  In e-commerce, customer segmentation InvoiceNo, StockCode, Description, refers to the use of customer data to divide Quantity, InvoiceDate, UnitPrice, customers into groups that share the same CustomerID, Country. behavior and characteristics such as gender, taste or shopping patterns, interests, and more.    Segmenting the customer base helps in better understanding the customers and thus personalizing marketing and communication for each segment. This is very beneficial because people tend to respond better and be  **3.1 Data Analysis & Data** of greater value to your business when they  feel their needs and interests are being **Transformation** specifically addressed.  While importing Dataset and data Here are some of the key benefits that inspection we got to know there are customer segmentation brings to online no major inconsistencies but we had retailers that incorporate it into their to deal with missing and duplicate marketing strategies: values. Whether to extract the  InvoiceDate column at the start or to   1. Helps in Acquiring Customers extract after some visualization was 2. Improve Customer Satisfaction and little challenging. So, we created 2 retention datasets considering both visualizations and at last chose the 3. **Dataset information** best path, which was to extract the   DateTime feature after some major In the dataset, we don't have labels. So, the visualizations.  process lies in segmenting customers based  on their behavior and then finding insights from that. For e-commerce, this type of The sample size of this dataset is information helps a lot in improving 541909. That means initially we had advertisements and marking ideas. With 541909 customer records. And we customer information and their purchasing had 8 features and no label. So, we behavior, we can cluster those customers in used all features and did some |

|  |  |
| --- | --- |
| univariate and multivariate analysis which also included checking and changing the distribution of some variables.  One thing to notice is that we also had Datetime variable. Extracted it and we got many useful variables to form it.  **RFM model (Recency, Frequency,**  **Monetary value):**  Recency, frequency, monetary value is a marketing analysis tool used to identify a company's or an organization's best customers by using certain measures. The RFM model is based on three quantitative factors. Frequency: How often a customer makes a purchase. Monetary Value: How much money a customer spends on  **3.2 Machine learning Models:**  **Segmentation with K-means clustering:**    Initially, the data is subject to important stages in an analytics pipeline: exploratory data analysis, pre-processing, feature engineering, and standardization.    Then, the unsupervised classification technique, the K-means clustering algorithm, is used to determine the ideal segments of customers. Silhouette score and related cluster visualizations are used to find the optimum value of "K" (number of clusters) in the algorithm.    The observations from the results are then elaborately discussed before concluding the business perspective. | We first created the RFM model (Recency, Frequency, Monetary value). Then applied unsupervised ML models for those RM, FM, and RFM values. We applied the k-means clustering algorithm and divided data points into different clusters and then calculated silhouette score on all those clusters and by observation, we found that the silhouette score for 2 optimal clusters is highest (for RM). We then applied k-means with the elbow method to find the optimal number of clusters and then applied DBSCAN and also Dendrogram as Hierarchical clustering for RM, FM, and RFM.     1. **Dealing with null values:**   Our concern was to not miss any important information but we got null values in 2 important features. So, we dropped those null records.    After dropping our null values, we moved further to the next steps.     1. **Methodology:**   K-means clustering is a great algorithm for dividing data points into different clusters. We also used DBSCAN and Dendrogram for finding the optimal number of clusters. |

This research work allows us to have an insight into the performance of various predictions and walk through the whole process of clustering.

* **Data pre-processing and transformation**

* **Create the RFM model (Recency,**

**Frequency, Monetary value)**

* **Log transformation of R, F, M for better results.**

* **K-Means Clustering algorithm**

* **Applying K-Means with Silhouette Score Method on RM, FM, RFM simultaneously.**

* **Applying K-Means with Elbow Method on RM, FM, RFM simultaneously.**
* **Applying DBSCAN on RM, FM, RFM simultaneously.**
* **Applying Hierarchical clustering on RM, FM, RFM simultaneously.**

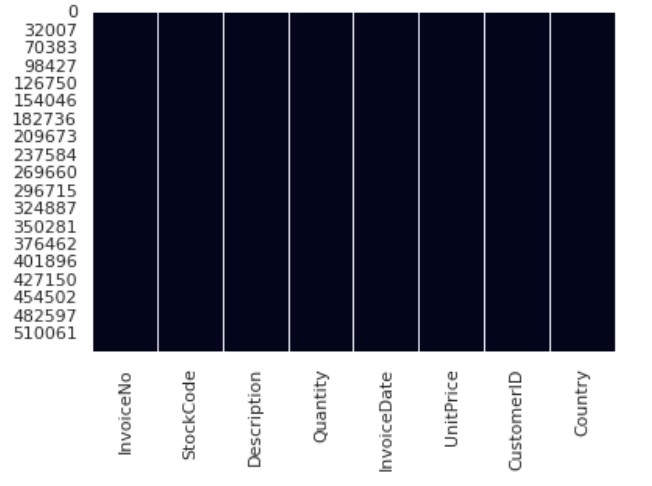
**5.1 Data pre-processing**

# and transformation

First things first, checking for missing values and making sure our dataset is complete. So, dealing with null values can be done by various methods. One way is to remove the entire feature which has null values but it is not advisable. Another way is to replace the null values with mean or median or mode. Apart from these 2 ways, one way is to delete all null values. According to the situation we choose different techniques to deal with null values.



The above plot shows we have null values in 2 features.

The above plot shows that we have dealt with null values.

**Which feature has the highest**

**number of unique**

**values?**

In This graph

, we can observe that

most

purchases are from the U

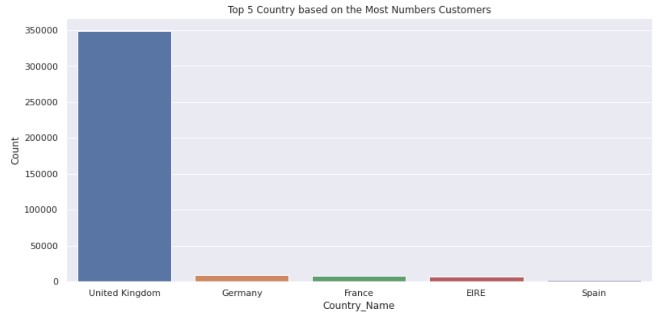
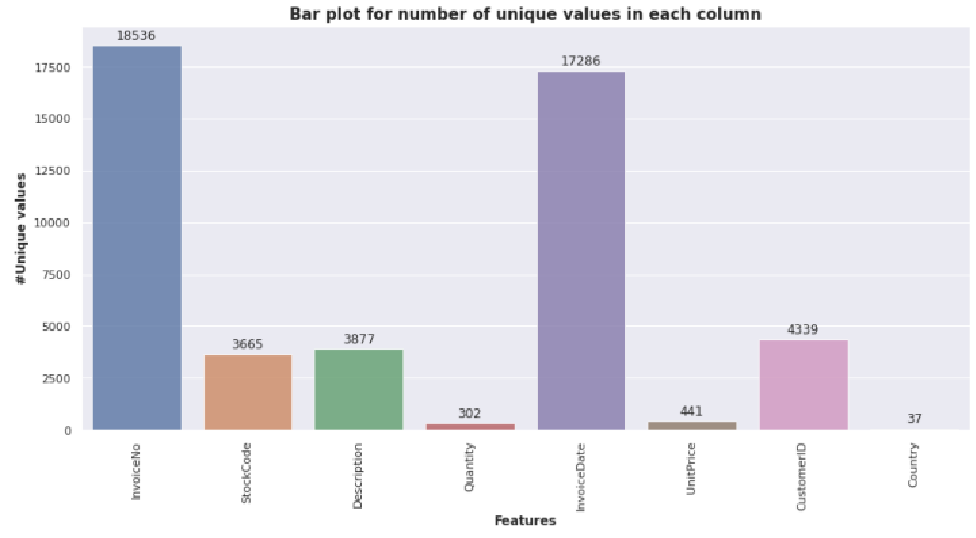
nited

K

ingdom

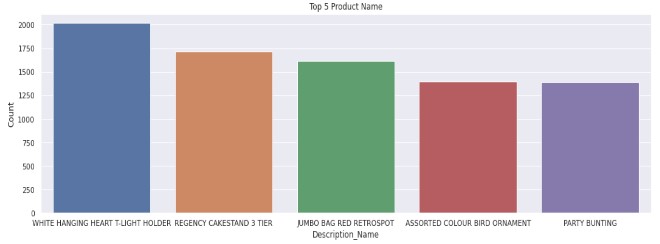
. It

is justifiable also as this is UK’s company.



The invoice number is unique for every transaction. So, it has the highest count in the dataset.

**Which are the top 5 products?**



Top products based on maximum selling are:

1. WHITE HANGING HEART T-LIGHT

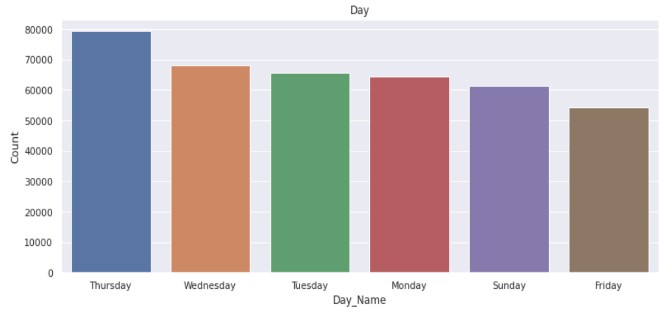
HOLDER

1. REGENCY CAKESTAND 3 TIER
2. JUMBO BAG RED RETROSPOT 4. PARTY BUNTING
3. LUNCH BAG RED RETROSPOT

**Top 5 Country based on the Most Numbers Customers?**

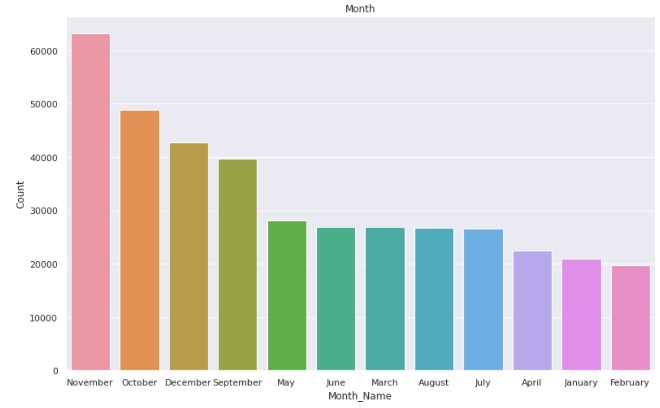
**5.2 Feature engineering:**

**Extract Invoice Date and create new features like day, month, hour, minute of transection.**



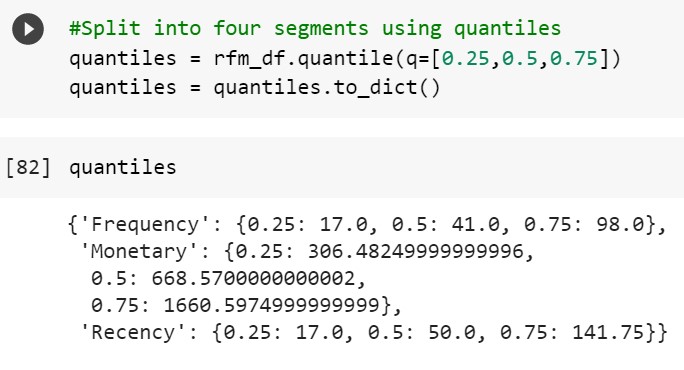
Most of the customers have purchased items on Thursday, Wednesday, Tuesday.

**Which month has the highest number of transactions?**



|  |  |
| --- | --- |
| Most of the customers have purchased items in November, October, December, and the least number of purchases in April, January, February.    **Most of the purchasing is happening in which hours?**      Most of the customers purchase in the afternoon time. The 12th hour of the day is a peak for purchasing items.    We have divided hours of the day into 3time types.   1. Morning 2. Afternoon 3. Evening     Most of the customers have purchased the items in the Afternoon, moderate numbers of customers have purchased the items in the Morning and the least numbers of customers have purchased the items in the Evening. | **5.3 Create the RFM model**  **(Recency, Frequency,**  **Monetary value)**  Recency, frequency, monetary value is a marketing analysis tool used to identify a company's or an organization's best customers by using certain measures. The RFM model is based on three quantitative factors:  Recency: How many days before the customer had made the purchase.  Frequency: How often a customer makes a purchase.  Monetary Value: How much money a customer spends on items.  **Performing RFM Segmentation and RFM Analysis Step by Step**  The first step in building an RFM model is to assign Recency, Frequency, and Monetary values to each customer. The second step is to divide the customer list into tiered groups for each of the three dimensions (R, F, and M), using Excel or another tool.    **Calculating RFM scores**  The number is typically 3 or 5. If you decide to code each RFM attribute into 3 categories, you'll end up with 27 different coding combinations ranging from a high of 333 to a low of 111. Generally speaking, the higher the RFM score, the more valuable the customer.    **Operation related to RFM model:**  1. We first created an RFM model, assigned RFM values to each customer. |

* 1. Checked RFM value distribution by using Distplot.
  2. We split RFM values into four segments using quantiles.



* 1. Add R, F, and M segment value columns in the existing data frame to show R, F, and M segment values
  2. Calculate and add the RFM score value column showing the total sum of RFM group values.
  3. Now each customer got a different RFM score. Generally, the highest RFM score means more value to the customer.
  4. We then checked the distribution of R OR F OR M separately to know more about customers purchasing behavior.

1. **Model Building:**

**K-Means Clustering**

**DBSCAN**

**Hierarchical Clustering**

**DBSCAN -** Density-Based Spatial

Clustering of Applications with Noise. Finds core samples of high density and expands clusters from them. Good for data that contains clusters of similar density.

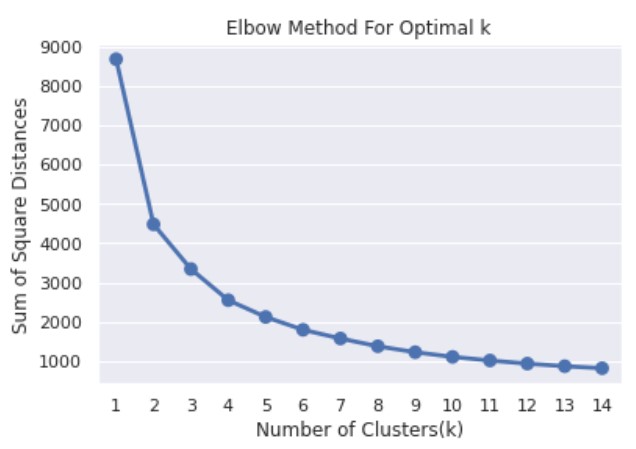
**6.1 Calculating models on**

## Recency and Monetary

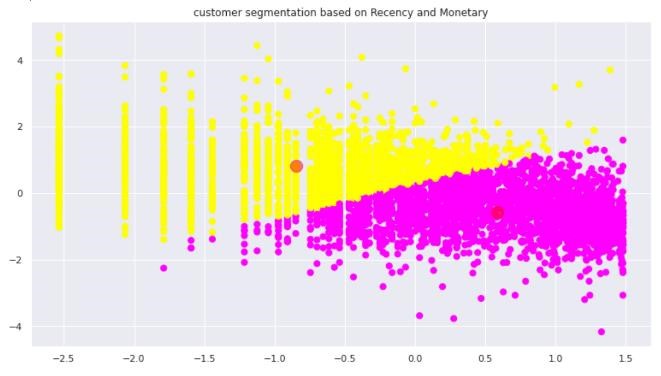
**K-Means with silhouette score**

For n\_clusters = 2, silhouette score is highest.

### K-Means with Elbow Method on Recency and Monetary

 With the elbow method also we are getting 2 as the optimal number of clusters.

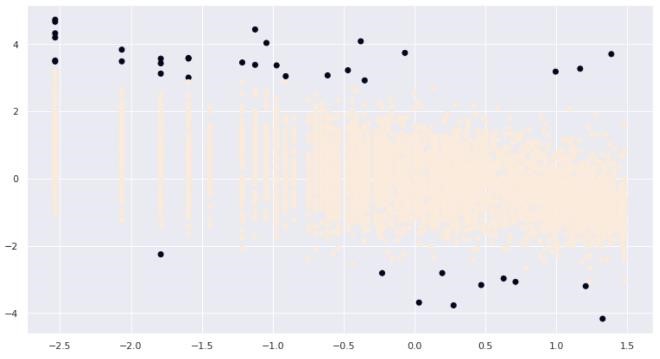
#### Let’s visualize where the optimal number of clusters = 2 using this plot



As a validation, we visualized taking 2 as the optimal number of clusters. And data points are well separated. This plot tells us

that we can divide data points into 2 clusters. Also, we see that Customers are well separated when we cluster them by Recency and Monetary.

#### Applying DBSCAN on Recency and Monetary



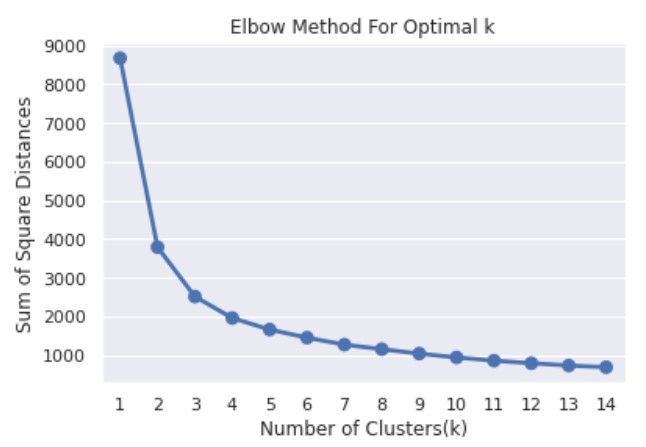
We can see that black dots are 1 cluster and brown colors are another cluster. In DBSCAN, we get the optimal number of clusters as output. So, we got 2 as optimal numbers of output.

### 6.2 Calculating models on Frequency and Monetary

#### K-Means with silhouette score

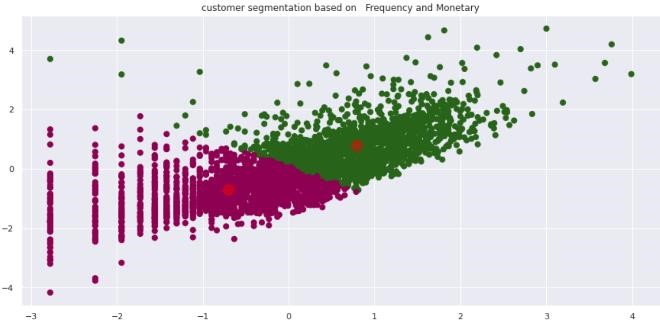
For n\_clusters = 2, silhouette score is 0.47, which is highest.

#### K-Means with Elbow Method on Frequency and Monetary



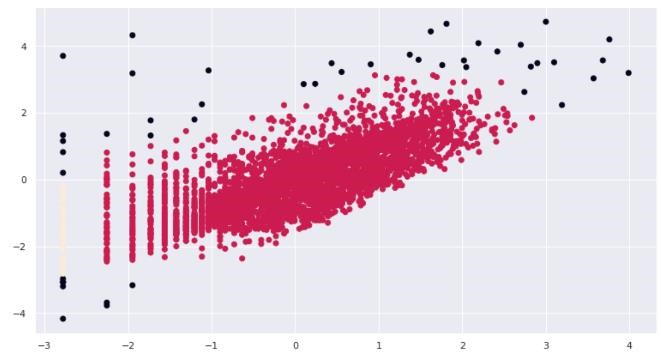
From the plot above, 2 looks like the optimal k for the cluster. Now let's visualize for 2 clusters.

##### Let's visualize where the optimal number of clusters = 2 using this plot



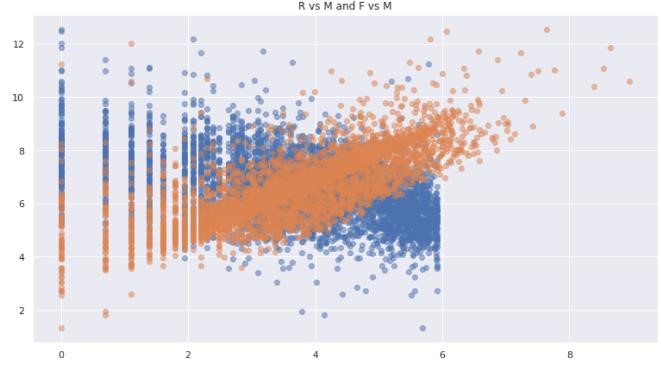
we see that Customers are well separated when we cluster them by Frequency and Monetary.

##### Applying DBSCAN to Method on Frequency and Monetary



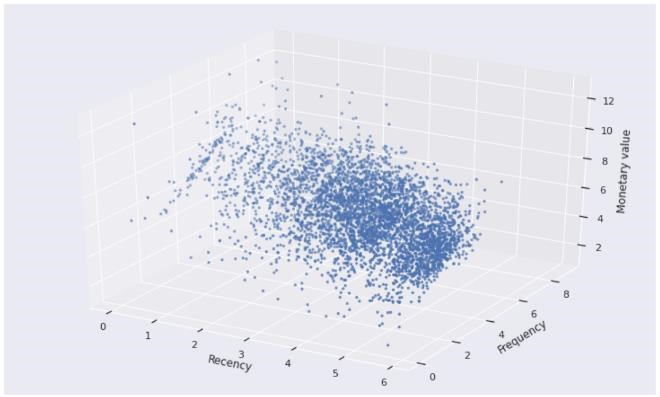
we see that Customers are well separated when we cluster them by Frequency and Monetary.

##### The plot of R vs M and F vs M



We plotted R vs M and F vs M, to find the relation between the 2 distributions. But could not see any relation.

##### 3D visualization of Recency, Frequency, and Monetary



To find the relation between Recency, Frequency, and Monetary but we found no relation between them.

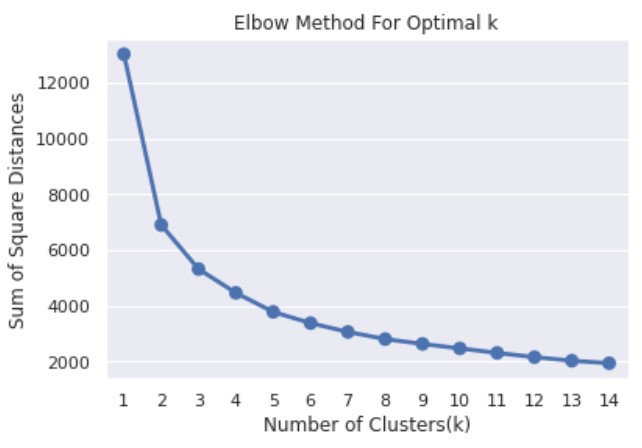
### 6.3 Calculating models on

**Recency, Frequency, and Monetary**

#### Applying Silhouette Method on Recency, Frequency, and Monetary

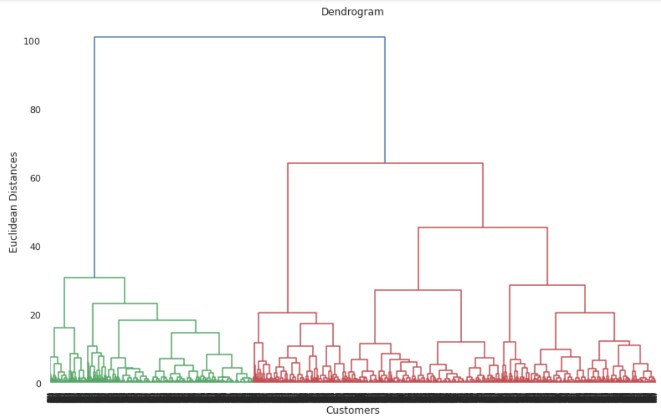
We found the number of clusters = 2 as the optimal number of clusters. Because silhouette score for 2 clusters is maximum.

#### Applying Elbow Method on Recency, Frequency, and Monetary



This method also gives information about the optimal number of clusters. According to it, 2 is the optimal number of clusters.

#### Using the Dendrogram to find the optimal number of clusters



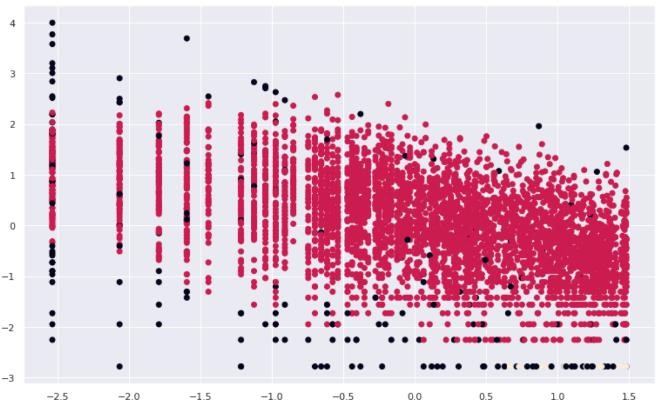
The number of clusters will be the number of vertical lines which are being intersected by the line drawn using the threshold 90. By doing so we got the optimal number of clusters = 2.

#### Visualizing the clusters (two dimensions only)



By applying a different clustering algorithm to our dataset. We get the optimal number of clusters is equal to 2. We can visualize it here.

#### Applying DBSCAN to Recency, Frequency, and Monetary

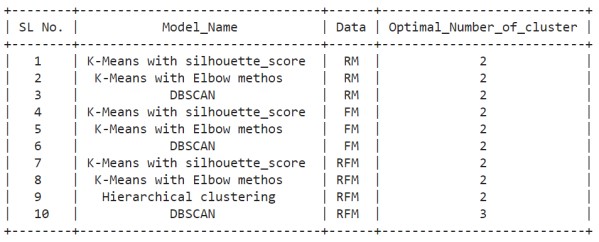


We see that customers are well separated when we cluster them by Recency,

Frequency, and Monetary.

The optimal number of clusters is different here, which is 3.

### 7. Summary



By applying a different clustering algorithm to our dataset. we get the optimal number of

clusters is equal to 2

#### 8. Conclusion

We investigated the dataset, checked null values, duplicate values. Then getting Dataset ready for exploratory data analysis. Then we did some feature engineering, we extracted the DateTime column for getting some more useful columns. We also created the Total amount as a new variable. Thus we prepared our data for models. After that, we created an RFM model (Recency, Frequency, Monetary value). We did a log transformation of R, F, M for better results. Now comes the model-building part, we used the K-Means clustering algorithm to

|  |  |  |
| --- | --- | --- |
| find the optimal number of clusters that can separate customers based on their  purchasing behavior. We then applied the KMeans algorithm with the Silhouette Score Method on RM, FM, RFM simultaneously. Where we got the optimal number of clusters = 2. We then visualized our results with scatterplots. Then moving forward we applied K-Means with Elbow Method on RM, FM, RFM simultaneously.  Next, we applied DBSCAN on RM, FM, and RFM where we got the optimal number of clusters as 2, 2, 3 simultaneously.  After that, we applied Dendrogram to find the optimal number of clusters and found that the optimal number of clusters is equal to 2.      **9. REFERENCES**   1. [https://towardsdatascience.com/custome r-segmentation-in-online-retail1fc707a6f9e6](https://towardsdatascience.com/customer-segmentation-in-online-retail-1fc707a6f9e6)      1. [https://www.shopify.in/encyclopedia/cu stomer-segmentation](https://www.shopify.in/encyclopedia/customer-segmentation)      1. [https://www.intercom.com/blog/cust omer-segmentation/](https://www.intercom.com/blog/customer-segmentation/)      1. [https://www.business2community.co m/customer-experience/4-types-ofcustomer-segmentation-allmarketers-should-know-02120397/](https://www.business2community.com/customer-experience/4-types-of-customer-segmentation-all-marketers-should-know-02120397/)      1. [https://en.wikipedia.org/wiki/Market](https://en.wikipedia.org/wiki/Market_segmentation/)   [\_segmentation/](https://en.wikipedia.org/wiki/Market_segmentation/)     1. [https://hbr.org/2014/07/what-youneed-to-know-about-segmentation/](https://hbr.org/2014/07/what-you-need-to-know-about-segmentation/)      1. [https://www.anblicks.com/blog/bene fits-of-customer-segmentation-ecommerce/](https://www.anblicks.com/blog/benefits-of-customer-segmentation-e-commerce/) |  | 1. [https://looker.com/blog/creatingactionable-customer-segmentationmodels/](https://looker.com/blog/creating-actionable-customer-segmentation-models/)      1. Moutinho, L., "Segmentation,   Targeting, Positioning and Strategic Marketing," Chapter 5 in Strategic Management in Tourism, Moutinho, L. (ed), CAB International, 2000, pp. 121–166 |