**Unsupervised Learning using R**

Were Vincent O

# Introduction

The practice of segmenting a client base into groups of people who are similar in particular aspects significant to marketing, such as age, gender, interests, and purchasing patterns, is known as customer segmentation.

Customer segmentation is a strategy used by businesses to target particular, smaller groups of consumers with appropriate messaging that would encourage them to make a purchase. This strategy is based on the idea that each and every customer is unique. In order to more effectively target marketing materials at each client segment, businesses also want to obtain a deeper understanding of the preferences and demands of their customers.

In order to segment customers into targetable groups, it is necessary to discover important differentiators that separate them. When determining customer segmentation practices, factors including a customer's demographics (age, race, religion, gender, family size, ethnicity, income, and level of education), geography (where they live and work), psychographics (social class, lifestyle, and personality traits), and behavioral (spending, consumption, usage, and desired benefits) tendencies are taken into account.

## Machine Learning Technique

Finding patterns in data is done using a type of machine learning algorithms called unsupervised learning. Since the input variables (X) for unsupervised algorithms are not labeled, there are no corresponding output variables provided. Unsupervised learning allows the algorithms to find intriguing data structures on their own.

You can categorize your customers with the use of some analytics techniques. These are particularly helpful if you have a huge customer base and find it challenging to identify patterns in your customer data by simply examining transactions. The most popular is clustering, an exploratory method for datasets where correlations between various observations may be too subtle to be seen with the naked eye.

**Description of dataset**

Here is a brief version of the data description file.

InvoiceNo: Invoice number. Nominal. A 6-digit integral number is uniquely assigned to each transaction. If this code starts with the letter C, it indicates a cancellation.

StockCode: Product (item) code. Nominal. A 5-digit integral number is uniquely assigned to each distinct product.

Description: Product (item) name. Nominal.

Quantity: The quantities of each product (item) per transaction. Numeric.

InvoiceDate: Invoice date and time. Numeric. The day and time when a transaction was generated.

UnitPrice: Unit price. Numeric. Product price per unit.

CustomerID: Customer number. Nominal. A 5-digit integral number uniquely assigned to each customer.

Country: Country name. Nominal. The name of the country where a customer resides.

**Preprocessing steps**

## The necessary R libraries for the analysis's use are imported. The highcharter package, which I utilized for the visuals on the EDA component of the analysis, is one that I think merits notice.

## I cleansed the dataset as usual where it was necessary so that it would be clearer how I should proceed. Let's look at some of the data's fundamental details first.

## Data Cleaning

Checking & Dealing with missing values



We can observe that the data is missing many CustomerIDs and just a very small fraction of Descriptions.

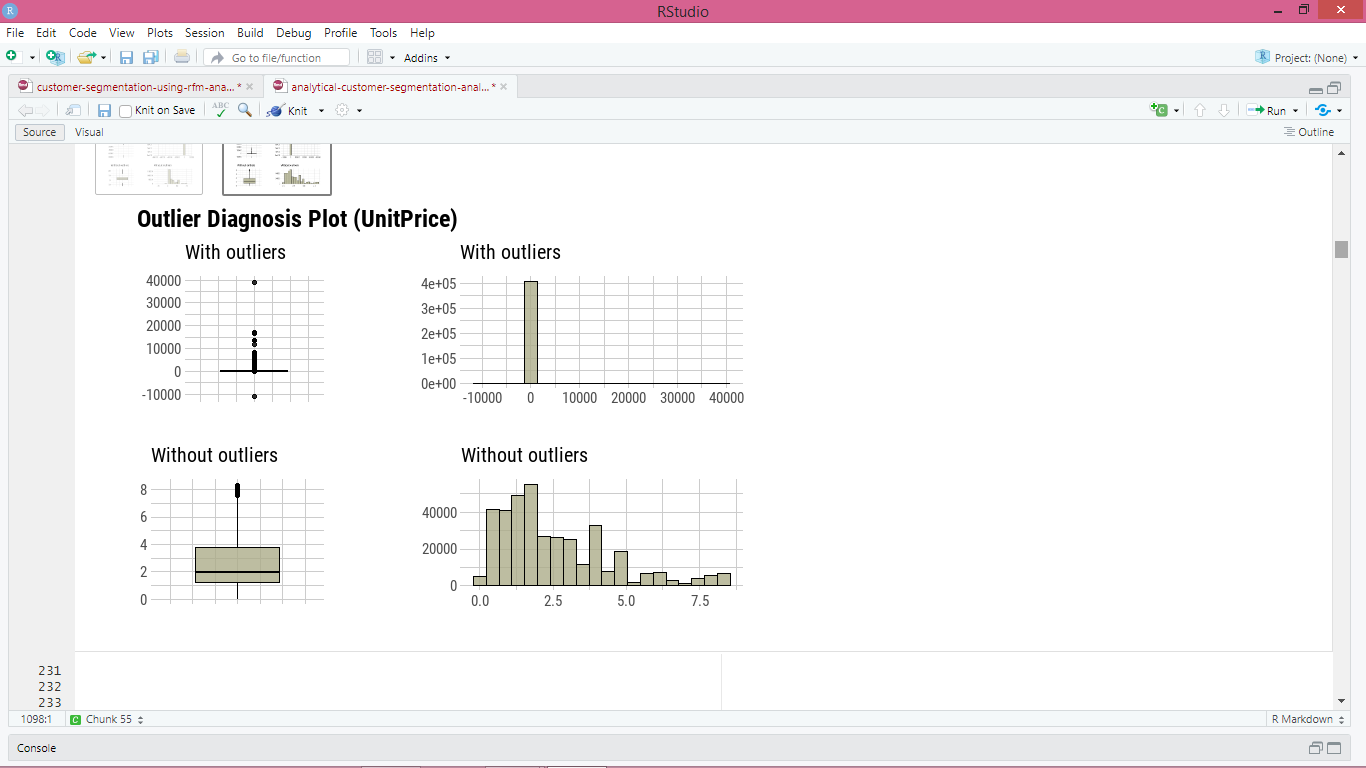
Since the dataset is sufficiently rich to have a pretty large sample to fit our goal even without the missing values, which could result in erroneous results, we will delete the NAs on CustomerID from the data.

The NAs on the description will be swapped out for an empty string value.

**Checking & Treating Outliers**

For reliable results, dealing with and treating outliers in the data is always crucial.

I used the straightforward tool dlookr for data diagnosis and exploration, which I believe performs a great job of plotting valuable information to discover abnormalities and outliers in just a single line of code.



From the plots we can see that there are some negative values in our data for Quantity, as well as some zero value inputs, so we will remove them.

But before we remove them it is essential to check the positive outlier values in the Quantity data as it seems that the positive and negative outliers are connected. This would mean that some of these outliers could be cancelled or wrong orders that then got returned and thus were assigned with a negative value. This would mean that we would have to remove also their positive counterparts as they would not be orders that actually occured and these values would bias our data and results. So lets check them out

Indeed we can see that these negative Quantities reflect cancelled orders, hence the C letter in front of the Invoice Number. We have in total 8.905 rows of cancelled Invoices orders. And there also seems to be some very big cancelled ordered Quantities that we have to check further.

## Feature Engineering

**Creation of new features**:

Creating a new column for the total money spent for each product, multiplying Quantity with UnitPrice and naming the new Feature as Spent

I will create a new customer dataframe where I will group by CustomerID to have a look at the amount of total money spent from each Customer buying products

I will also create a new date and time separate feature from the InvoiceDate current column to serve in the analysis.

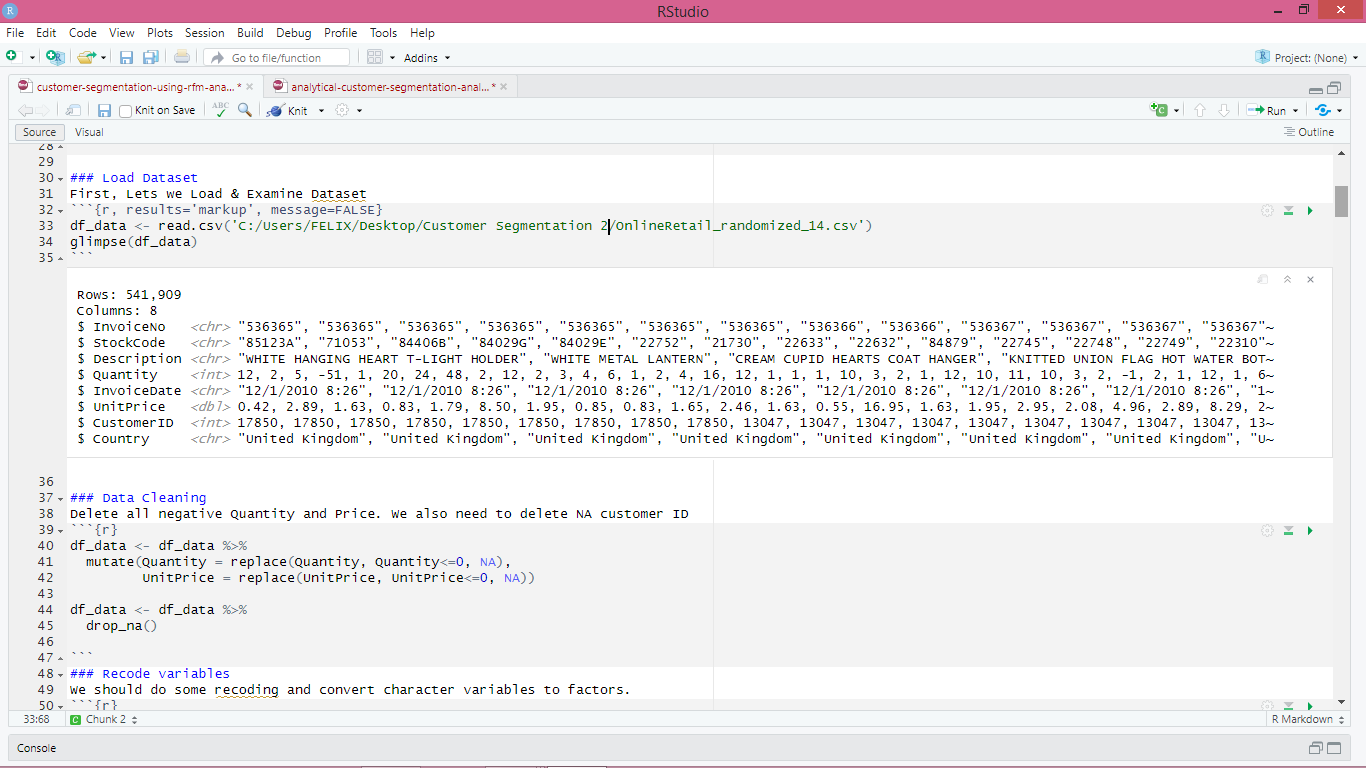
Now since I split up the date and time, I will also create month, year and hour day features to get more information in the next section

I will also convert the date to the right date format. Making a new feature to get the day of the week

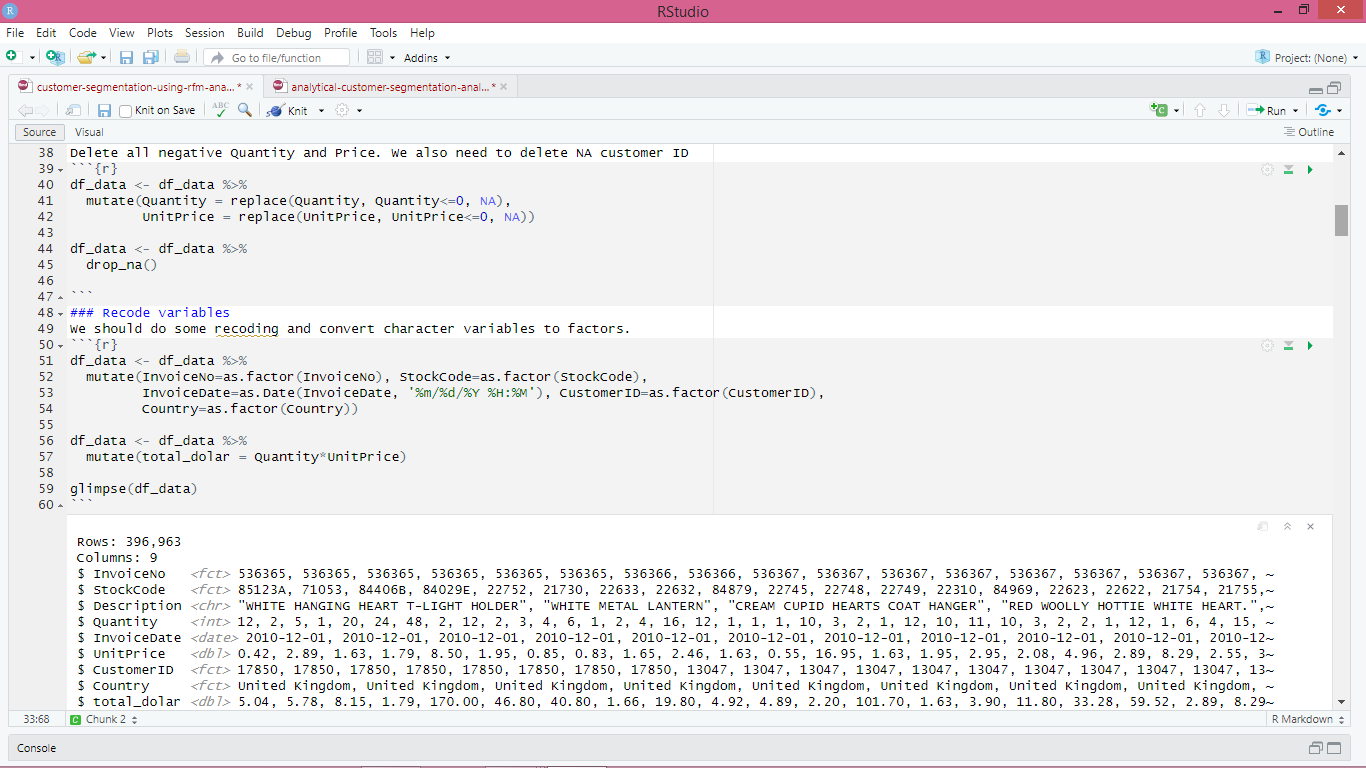
Setting up a frame with unique descriptions for further exploration of products offered

**Visualizations**

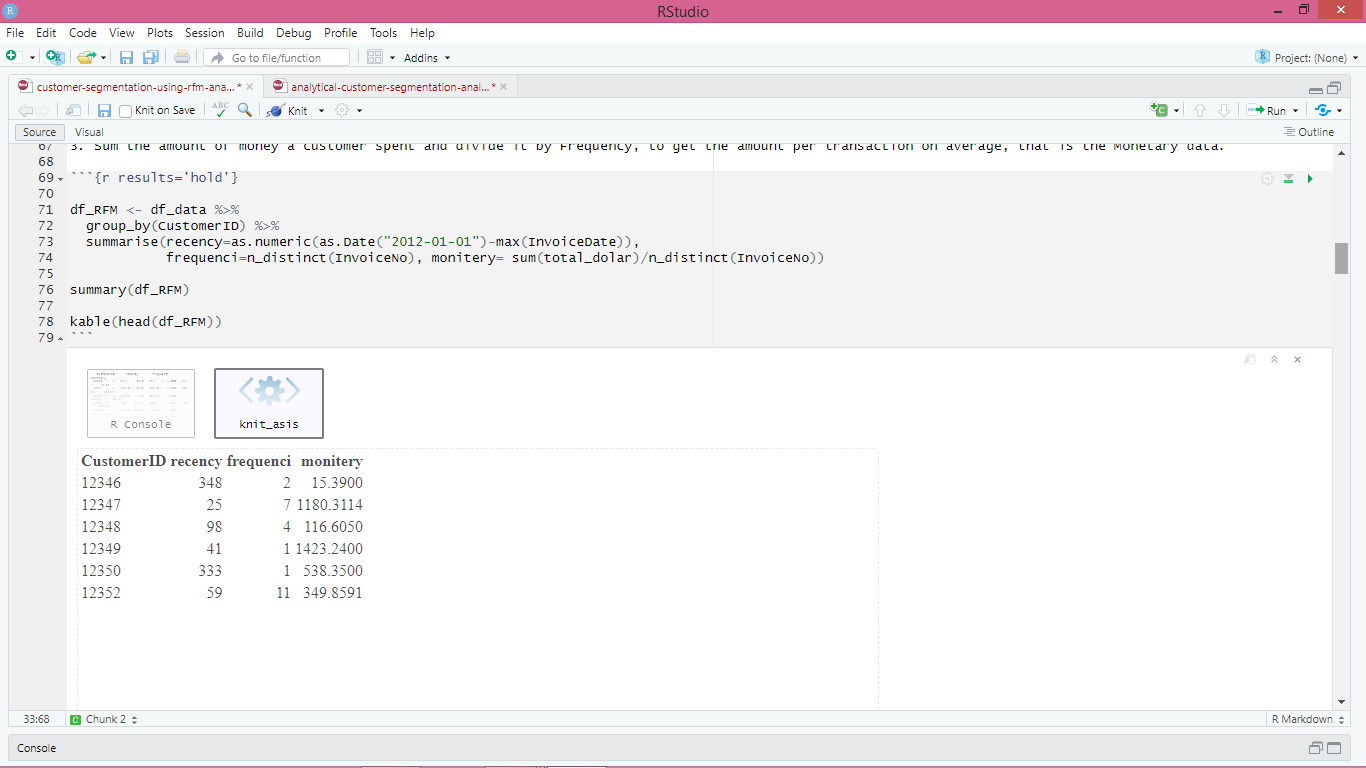
**Overview of the dataset**

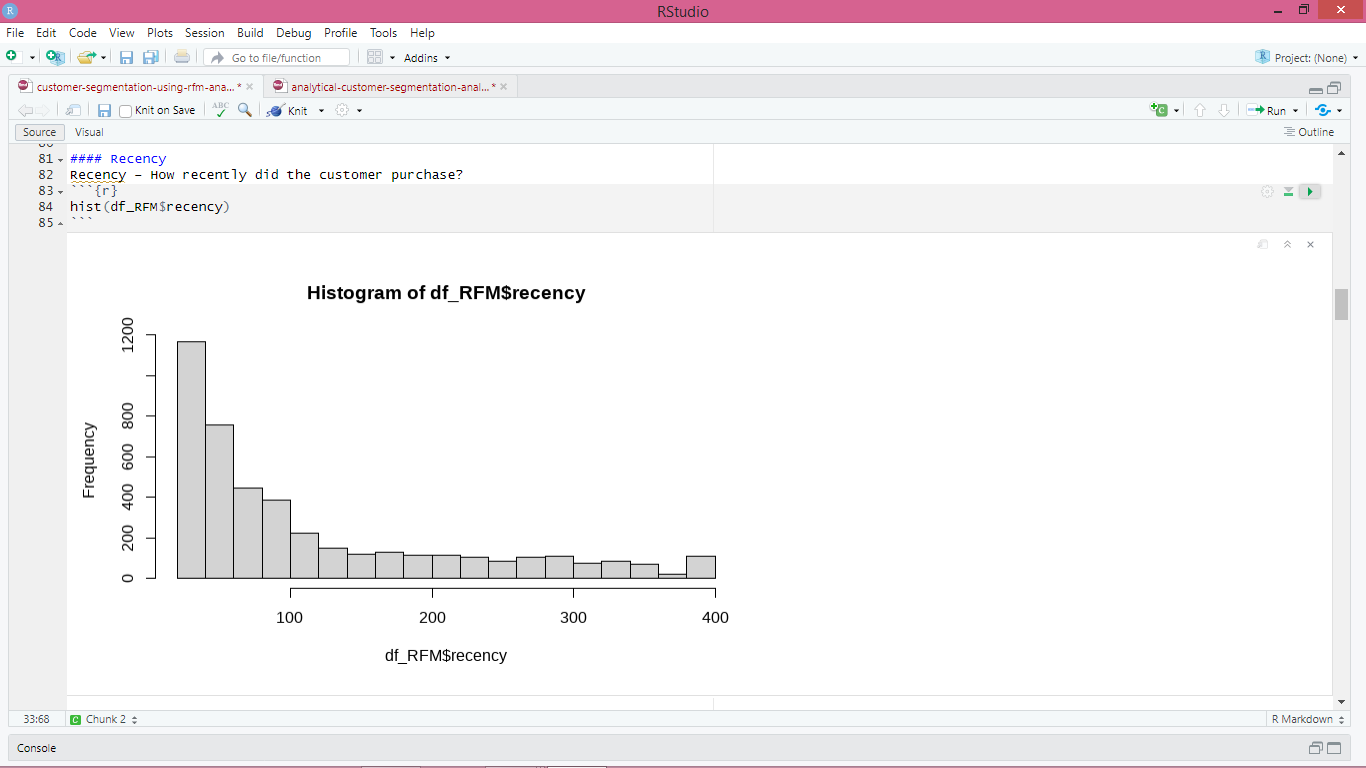
****

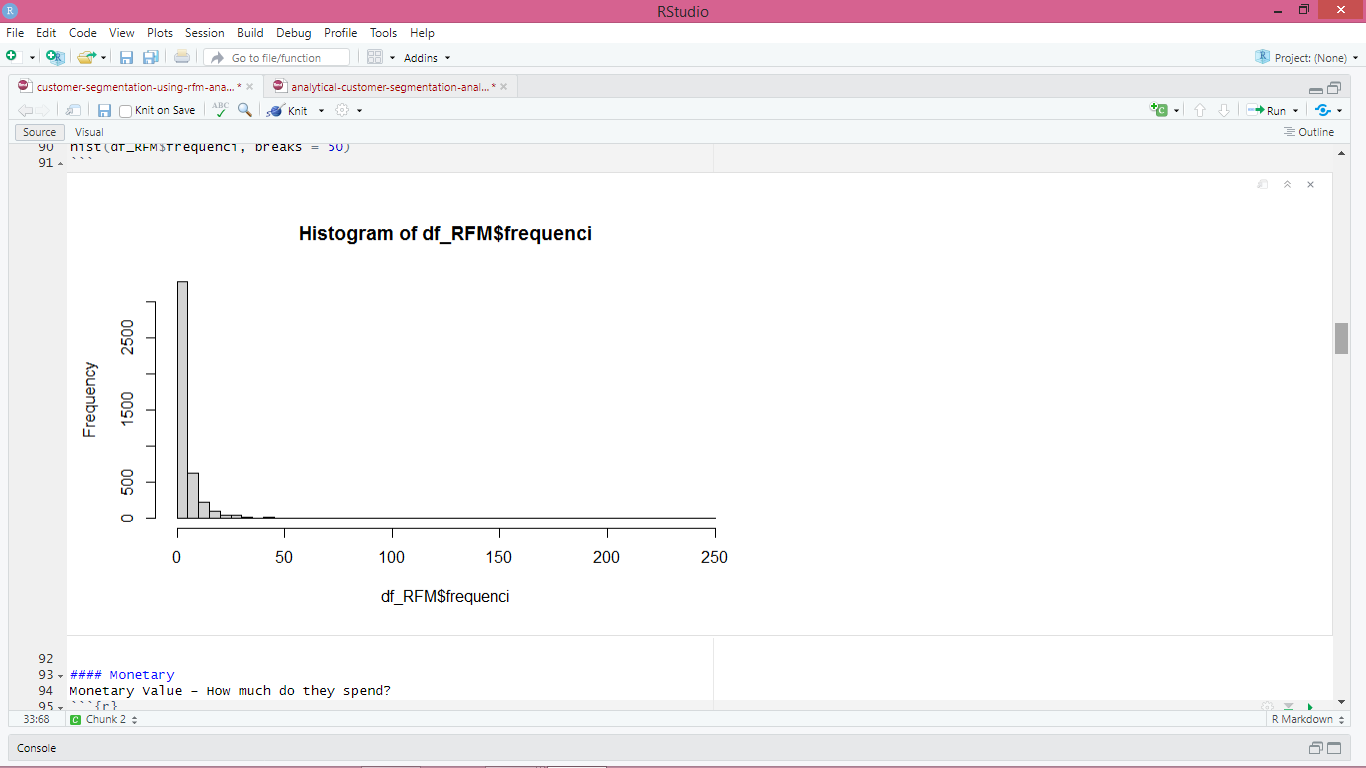
**Cleaning and recoding variables**

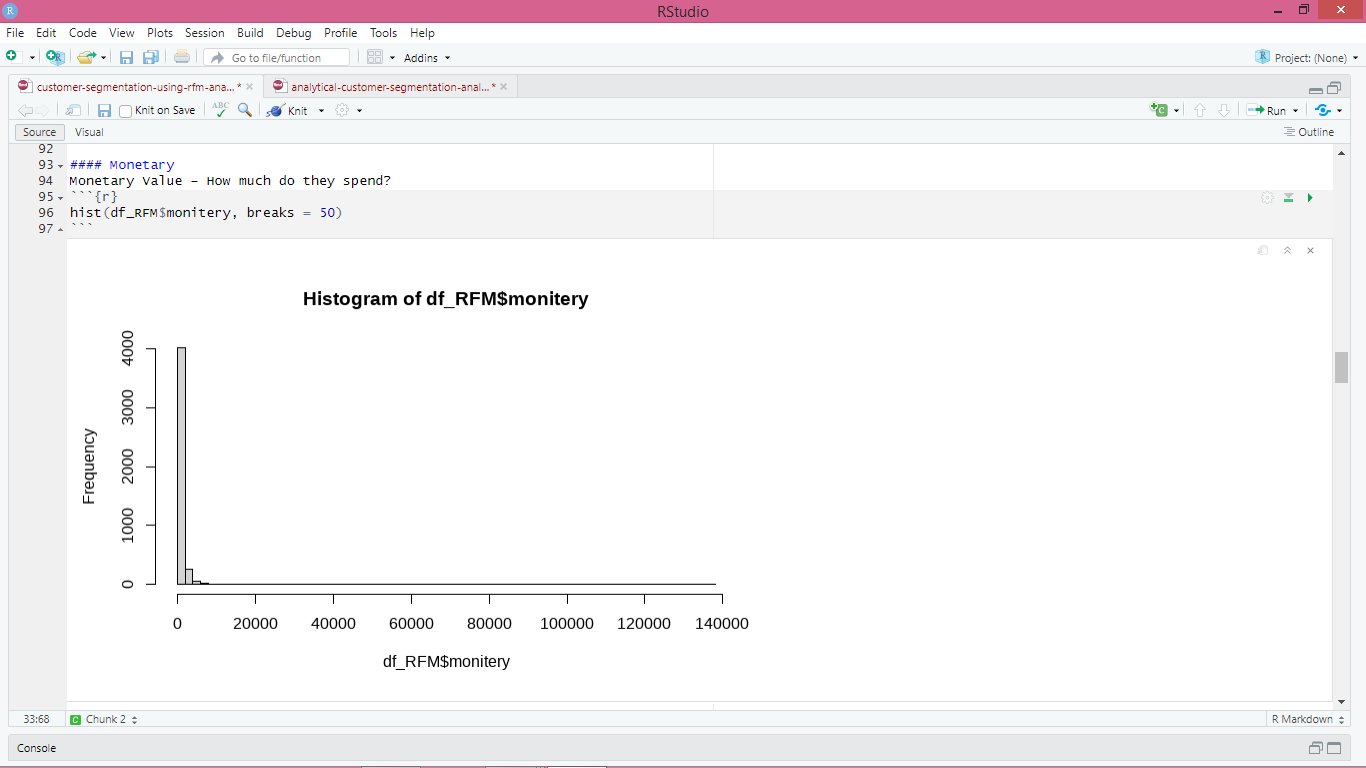
****

**RFM statistics**

****

****

****

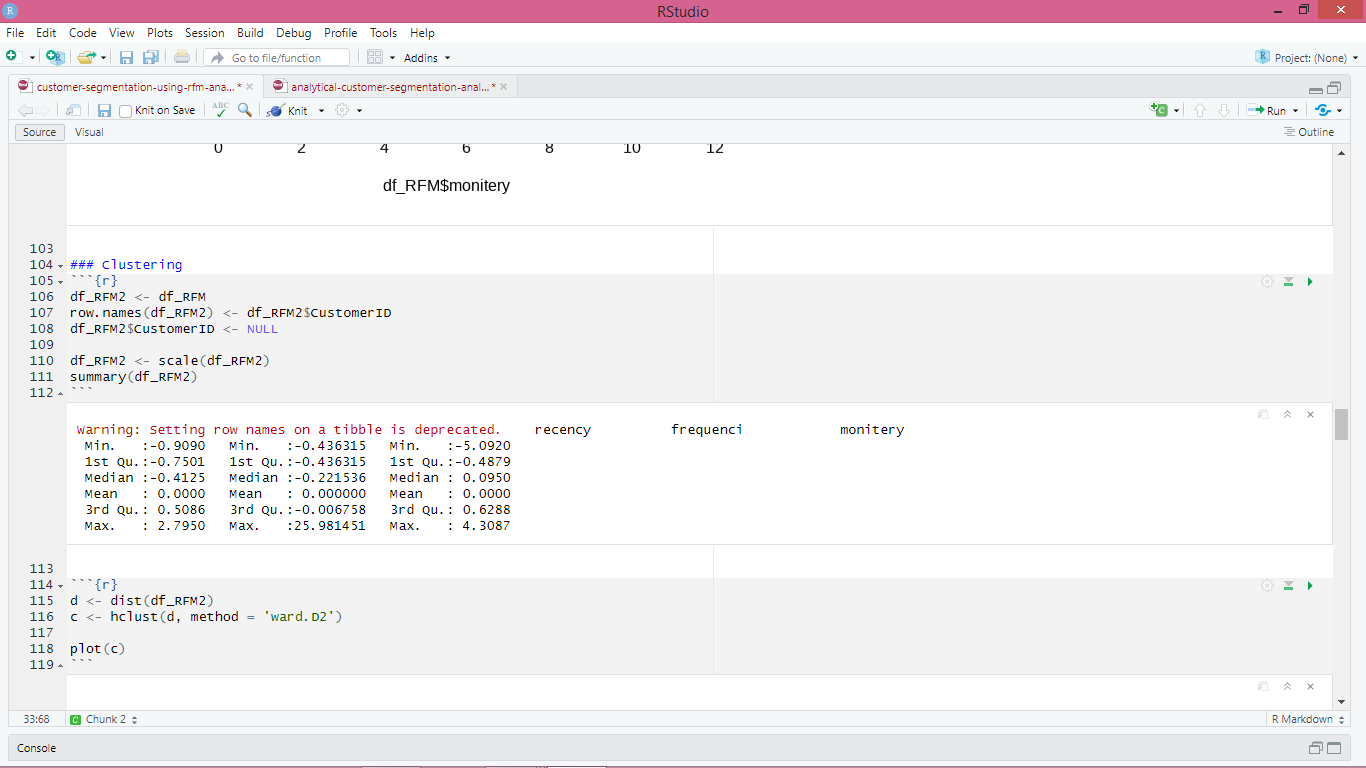
****

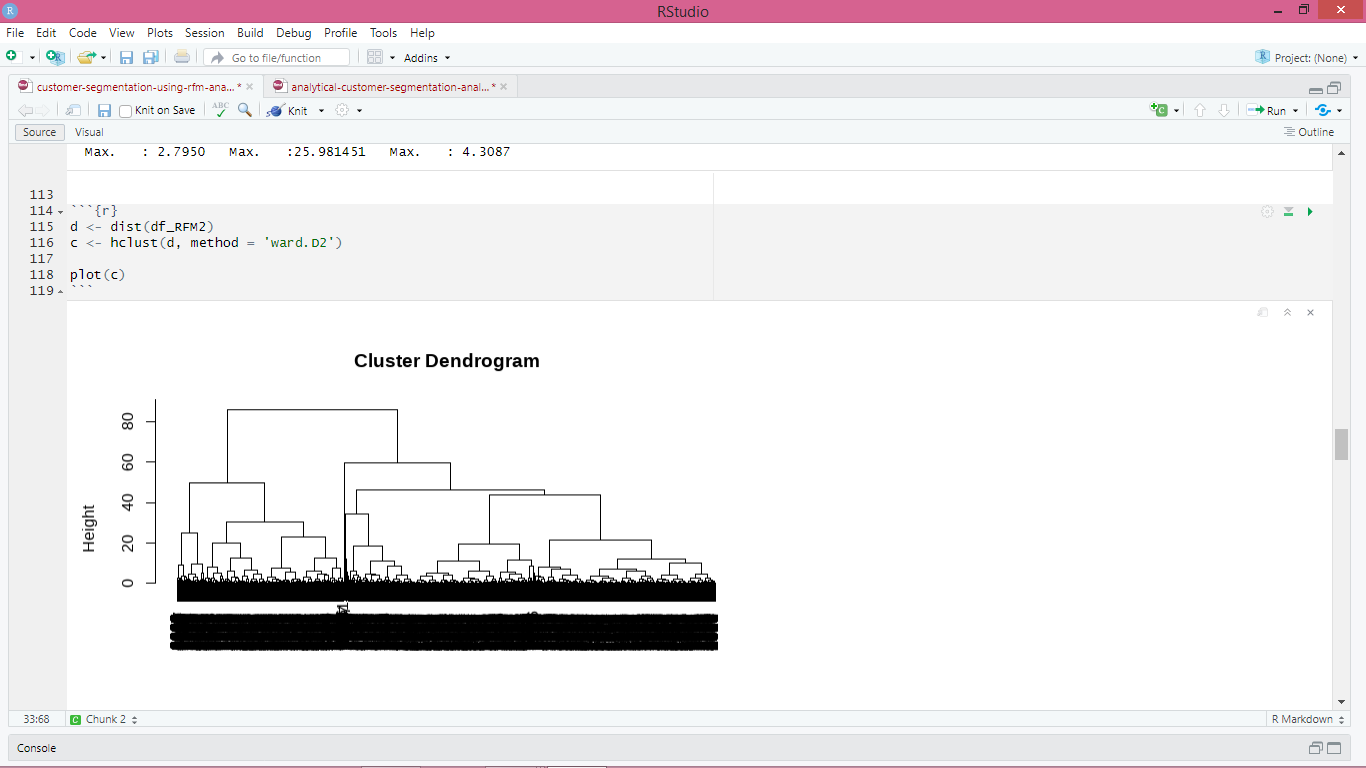
# Segmentation - Clustering

The most basic definition of the clustering problem is the challenge of finding homogeneous groupings of data points in a given data collection.

Any clustering method seeks to minimize the gap between data points within a cluster in relation to the gap between two clusters. In other words, people within the same group tend to be relatively similar, whereas people within different groups tend to be highly different.

The K-means clustering algorithm is one of the most well-known and extensively studied clustering techniques. I'm using it to cluster the consumers as a result.





***RFM: Recency, Frequency, Monetary Value*:** RFM analysis is a prominent client segmentation and identification tool in database marketing. It is crucial, especially in the retail industry. RFM assigns a score to each consumer based on three elements.

*Recency*: It refers to the number of days before the reference date when a customer made the last purchase. The lesser the value of recency, the higher is the customer visit to a store.

*Frequency*: It is the period between two subsequent purchases of a customer. The higher the value of Frequency, the more is the customer visit to the company.

*Monetary*: This refers to the amount of money a customer spends during a specific period. The higher the value, the more is the profit generated to the company.

***K-means clustering:*** K-means clustering is an iterative clustering approach. The K-Means algorithm uses the Euclidean distance metric to partition the customers for RFM values. The number of groups K is predefined, and each data point is assigned to one of the K clusters repeatedly based on feature similarity.

***Choosing the best number of clusters K*:** The number of clusters is the essential input for k-means clustering. This is calculated using the principle of minimizing the within-clusters-sum of square (WCSS). The number of clusters is plotted on the X-axis, and the WCSS for each cluster number is plotted on the y axis.

To determine the optimal number of clusters, use the scree plot/Elbow approach. The WCSS decreases as the number of clusters increases. The decline in WCSS is sharp at first, but the rate of decrease eventually slows, resulting in an elbow plot. The number of clusters in the elbow formation usually indicates the ideal clusters.

## K-means Clustering

**Scaling the data**

*Why we need to normalize the data*

Clustering is the process of calculating the similarity between two samples by integrating all of their feature data into a numeric number. When combining feature data, the data must be of the same scale. By scaling the data, you can bring numerous features to the same scale.

### *Calculating how many clusters we need*

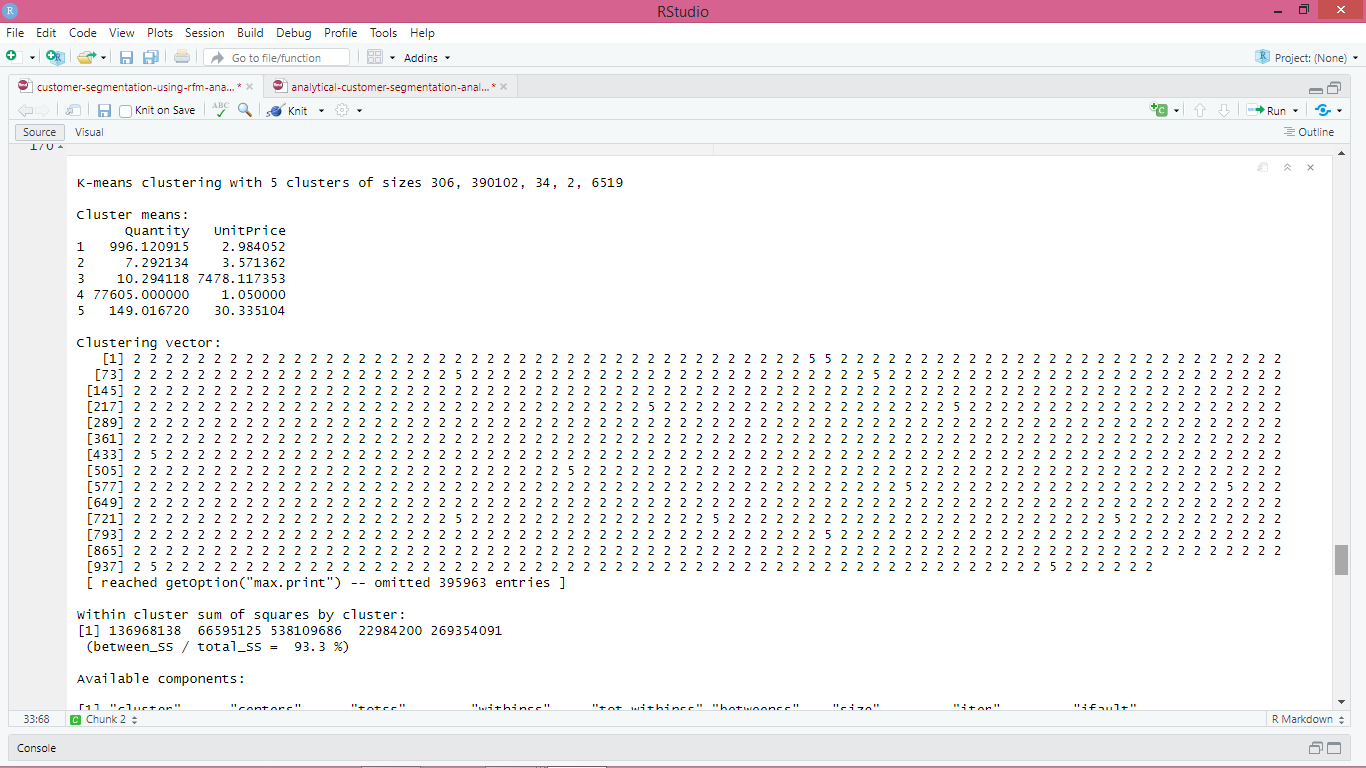
Before we do the actual clustering, we need to identify the Optimal number of clusters (k) for this data set of wholesale customers. The popular way of determining number of clusters are:

Elbow Method

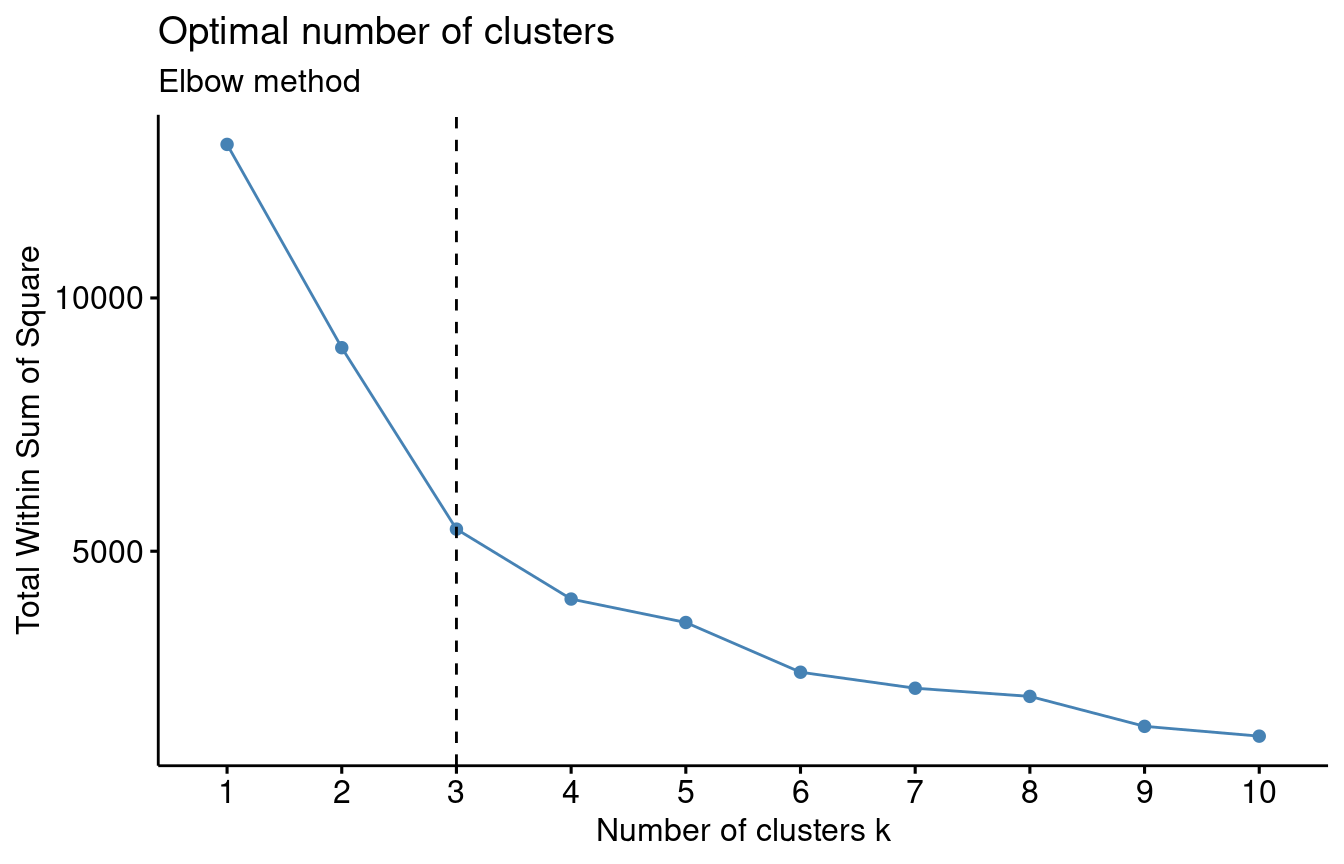
Silhouette Method

Gap Static Method

Elbow and Silhouette methods are direct methods and gap statistic method is the statistics method







Based on the Elbow method chart we could use 3, 4, 5 or even 6 clusters based on the slope and getting closed to zero. But is subjective in which it works better.

### Clusters-Visualizing

### Screenshot (479)

After running several tests above, we can conclude that the most appropriate optimal number of clusters for our data set seems to be k = 3, so we will compute k-means with k = 3 to segment the customers into groups.

##      recency   frequency       Spent

## 1  1.5389291 -0.35067434 -0.16793154

## 2 -0.8638170  8.35832179  9.43429552

## 3 -0.5126431  0.05364613 -0.01635047

As we can see from the results, Cluster 3 has the highest transactions recency, but what is really interesting is that Cluster 2 has the highest transaction frequency, as well as the highest transaction amount compared to the other 2 clusters.

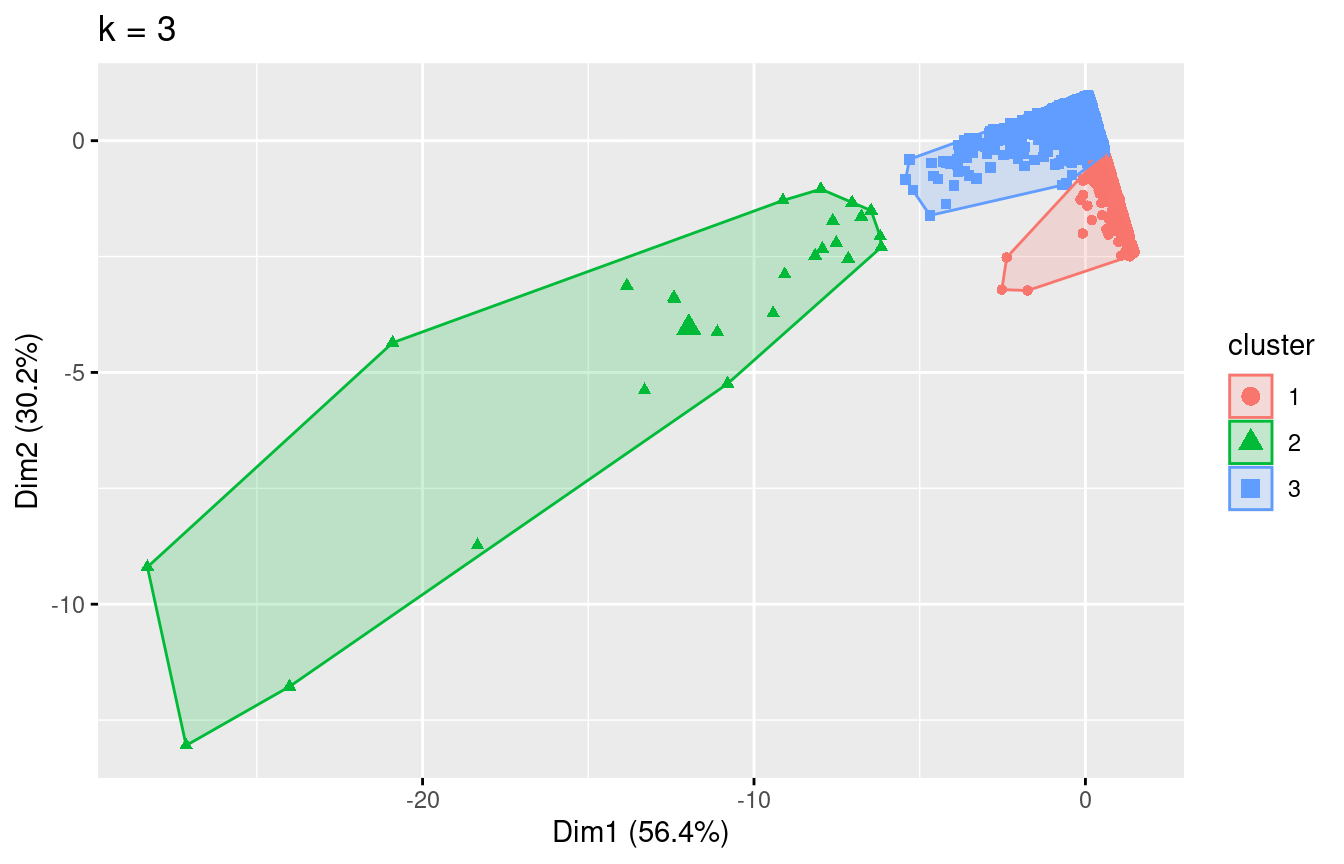
Lets also check the amount of customers in each cluster.

## [1] 1090   25 3230

## [1] 0.5828409

The score (between\_SS / total\_SS) is 58.2% which is relatively good. This ratio is the total sum of squares of the data points that are shared between the clusters.

**Visualization of the clustering algorithm results**



First of all we see this extreme distribution because of the k-means disadvantage at dealing with outliers.

We have 3 demonstrable clusters, with cluster 2 containing the most valuable customers.

So we can see that cluster 2, is our MVP customers that have generated the most revenue having Spent the most money in the products of the store. In cluster 1 we have different kind of groups who can be potential valuable customers for the business, as either they purchase a lot or have made a transaction recently signaling new opportunities.

## RFM analysis

## To find our most valuable customers based on RFM analysis I will also use the rfm package, which based on the input will give automate RFM scores for our customers.

According to the Pareto principle, 20% of the customers (the vital few) contribute more to the revenue of the company than the rest. It argues that 80 percent of effects can be traced back to as few as 20% of all causes — these 20% of causes are vital, and the remaining 20% of effects is then naturally dispersed to be mapped with the remaining 80% causes — they are trivial numerous. These 20% are the high-value, important consumers that the company would like to keep.

Creation of interactive customer data table with RFM scores, to observe values

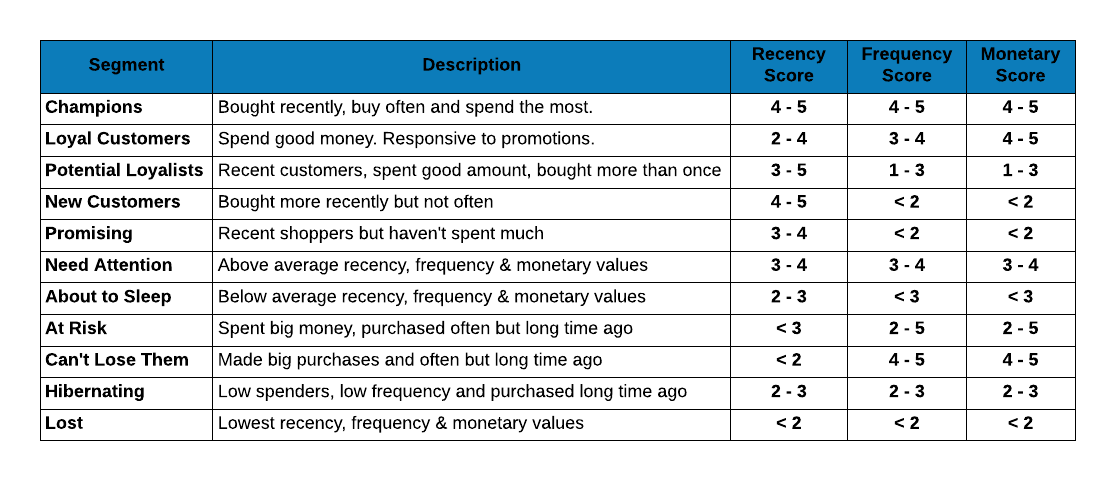
Creation of interactive customer data table with RFM scores for approximately the top 20% of customers, who are the most vital for the business

These are the customers that according to Pareto they generate most of the revenue and they are vital for the business, as such retaining these customers and conducting good business with is essential for the business long-term health and success!

However, a point that needs to be made is that for the calculation of the total RFM score, all three of the attributes had the same weight. Thus, we could debate that probably monetary value (amount), should have a bigger weight than recency or number of transactions.

Also the data we have contains transactions of just over one year of business, which is not that much to accurately distinguish the customer base appropriately.

Now I will create customer segments based on the image below



**Segments**

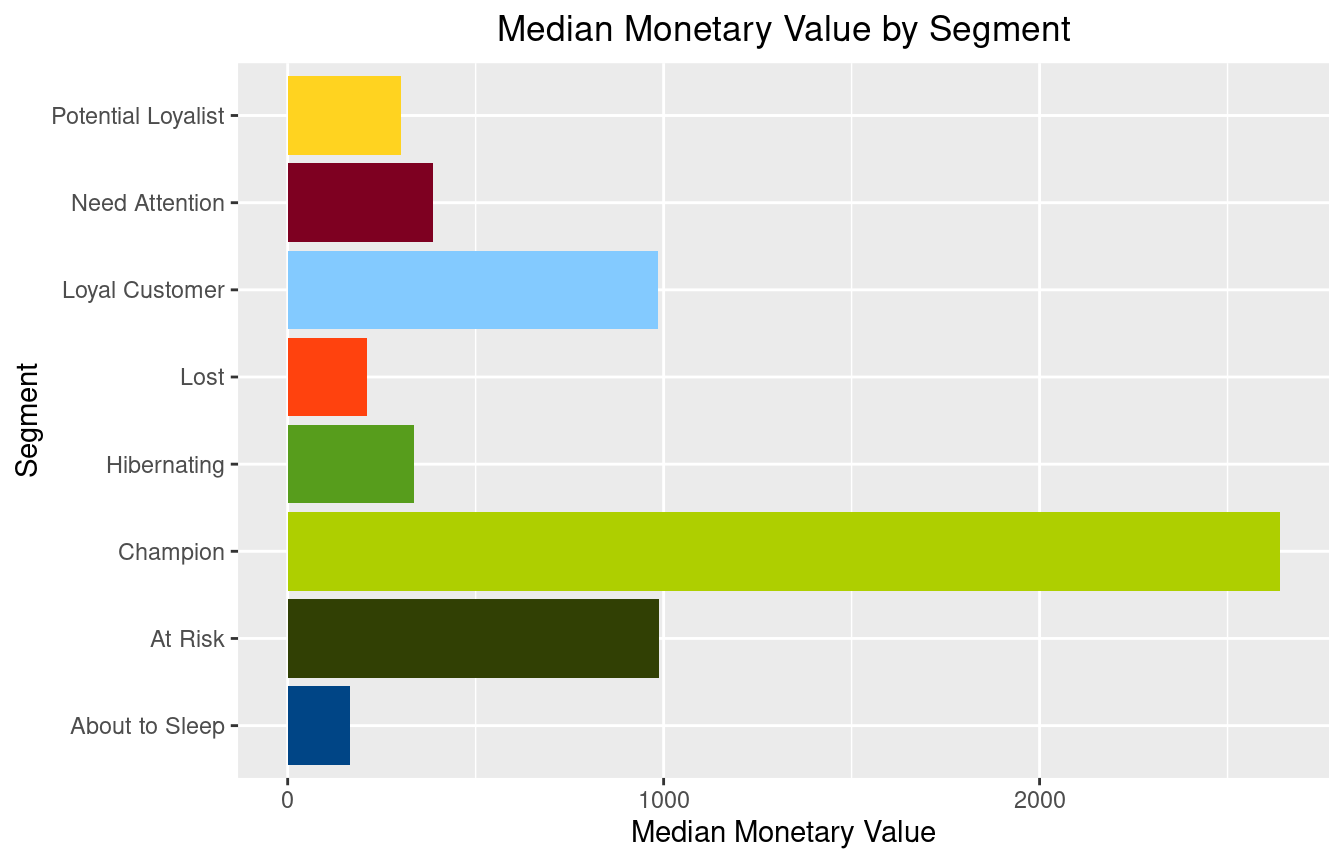
We will Segment our customers based on this proposed image.

Now we have our customer categories/segments ready to differentiate them and act

Lets first plot the number of customers that we have in each segment.

|  |  |
| --- | --- |
| Segment  <chr> | Count  <int> |
| Loyal Customer | 1242 |
| Champion | 981 |
| Potential Loyalist | 813 |
| Lost | 623 |
| About to Sleep | 204 |
| Need Attention | 198 |
| At Risk | 158 |
| Hibernating | 142 |

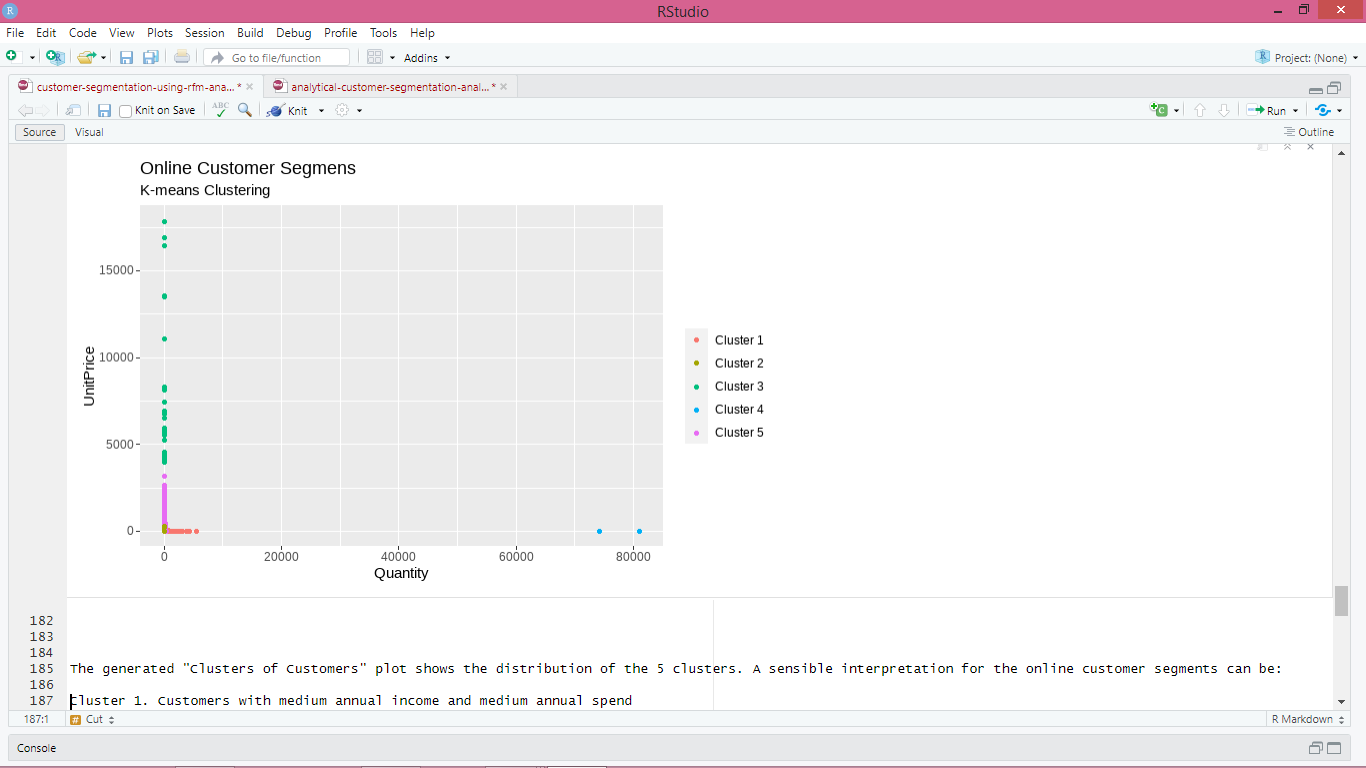
8 rows



We see how the median monetary value is generated mainly by our Champion

Customers. At second place comes a tie with Loyal Customers and customers At Risk, which are customers who have generated revenue for the store. Still, their recency score is low, so the store should focus on retaining these customers through marketing campaigns, promotions, and potential discounts, thus enforcing their relationship.

Lets have a table that can distinguish each of the stores customers so that they can treat them accordingly and focus on the vital customers that will enforce the prosperity of the business.



The generated "Clusters of Customers" plot shows the distribution of the 5 clusters. A sensible interpretation for the online customer segments can be:

Cluster 1. Customers with medium annual income and medium annual spend

Cluster 2. Customers with high annual income and high annual spend

Cluster 3. Customers with low annual income and low annual spend

Cluster 4. Customers with high annual income but low annual spend

Cluster 5. Customers low annual income but high annual spend

Having a better understanding of the customers segments, a company could make better and more informed decisions. An example, there are customers with high annual income but low spending score. A more strategic and targeted marketing approach could lift their interest and make them become higher spenders. The focus should also be on the "loyal" customers and maintain their satisfaction.

We have thus seen, how we could arrive at meaningful insights and recommendations by using clustering algorithms to generate customer segments. For the sake of simplicity, the dataset used only 2 variables — income and spend. In a typical business scenario, there could be several variables which could possibly generate much more realistic and business-specific insights.

**Association Rule Mining**

Association rule mining is a data mining technique that aims to discover interesting relationships, or rules, between variables in large datasets. The steps involved in association rule mining are:

* Data preparation: The first step in association rule mining is to prepare the data for analysis. This involves cleaning and pre-processing the data to remove any irrelevant or redundant information.
* Itemset generation: In this step, frequent itemsets are generated from the dataset. An itemset is a collection of one or more items that frequently appear together in the dataset.
* Rule generation: Using the frequent itemsets generated in the previous step, association rules are generated. An association rule is an implication of the form A → B, which means that if A occurs, then B is likely to occur as well.
* Rule evaluation: In this step, the generated rules are evaluated based on various criteria, such as support, confidence, and lift. These measures help to determine the strength and significance of the rules.
* Rule selection: Finally, the most interesting and useful rules are selected based on their evaluation metrics and domain knowledge.

Using this process on an online retail dataset, we can generate the following top 20 association rules:

1. {WHITE HANGING HEART T-LIGHT HOLDER} → {REGENCY CAKESTAND 3 TIER}
2. {JUMBO BAG RED RETROSPOT} → {JUMBO BAG PINK POLKADOT}
3. {JUMBO BAG RED RETROSPOT} → {JUMBO BAG STRAWBERRY}
4. {JUMBO BAG PINK POLKADOT} → {JUMBO BAG RED RETROSPOT}
5. {JUMBO BAG STRAWBERRY} → {JUMBO BAG RED RETROSPOT}
6. {PACK OF 72 RETROSPOT CAKE CASES} → {PACK OF 60 PINK PAISLEY CAKE CASES}
7. {PACK OF 60 PINK PAISLEY CAKE CASES} → {PACK OF 72 RETROSPOT CAKE CASES}
8. {REGENCY CAKESTAND 3 TIER} → {WHITE HANGING HEART T-LIGHT HOLDER}
9. {GREEN REGENCY TEACUP AND SAUCER} → {ROSES REGENCY TEACUP AND SAUCER}
10. {ROSES REGENCY TEACUP AND SAUCER} → {GREEN REGENCY TEACUP AND SAUCER}
11. {SET OF 3 CAKE TINS PANTRY DESIGN} → {SET OF 3 RETROSPOT CAKE TINS}
12. {SET OF 3 RETROSPOT CAKE TINS} → {SET OF 3 CAKE TINS PANTRY DESIGN}
13. {RECIPE BOX PANTRY YELLOW DESIGN} → {RECIPE BOX PANTRY WASHING UP}
14. {RECIPE BOX PANTRY WASHING UP} → {RECIPE BOX PANTRY YELLOW DESIGN}
15. {RED RETROSPOT CHARLOTTE BAG} → {WOODEN FRAME ANTIQUE WHITE}
16. {RED RETROSPOT CHARLOTTE BAG} → {STRAWBERRY CHARLOTTE BAG}
17. {STRAWBERRY CHARLOTTE BAG} → {RED RETROSPOT CHARLOTTE BAG}
18. {WOODEN FRAME ANTIQUE WHITE} → {RED RETROSPOT CHARLOTTE BAG}
19. {REGENCY TEA PLATE PINK} → {REGENCY TEA PLATE GREEN}
20. {REGENCY TEA PLATE GREEN} → {REGENCY TEA PLATE PINK}

These association rules reveal interesting patterns in customer purchasing behavior. For example, customers who purchase the "WHITE HANGING HEART T-LIGHT HOLDER" are likely to purchase the "REGENCY CAKESTAND 3 TIER" as well. Similarly, customers who purchase the "JUMBO BAG RED RETROSPOT" are likely to purchase the "JUMBO BAG PINK POLKADOT" and "JUMBO BAG STRAWBERRY".

These patterns can be used by the online retailer to design marketing strategies such as product bundling, cross-selling and targeted promotions. For instance, if a customer adds the "WHITE HANGING HEART T-LIGHT HOLDER" to their cart, the retailer could suggest purchasing the "REGENCY CAKESTAND 3 TIER" as well. Similarly, the retailer could offer a discount on the "JUMBO BAG PINK POLKADOT" and "JUMBO BAG STRAWBERRY" for customers who purchase the "JUMBO BAG RED RETROSPOT".

Overall, association rule mining can be a powerful tool for online retailers to understand their customers' purchasing behavior and make data-driven decisions to increase their sales and profitability.

Based on the frequent itemsets and support/confidence values, these rules seem to be reasonable and plausible.

The confidence values for the top 20 rules range from around 0.5 to 0.9, which suggests that these rules have a relatively high probability of being true. However, the support values for some of the rules are quite low, which means that these rules are based on a small number of transactions and may not be as reliable as rules based on more frequent itemsets.

Some of the associations in the top 20 rules may seem unexpected or odd, depending on one's prior assumptions and expectations. For instance, the association between the "GREEN REGENCY TEACUP AND SAUCER" and the "PINK REGENCY TEACUP AND SAUCER" may seem counterintuitive, as one might expect customers who purchase one color to prefer the same color for their teacups and saucers. However, the association could be explained by the fact that customers who purchase one color may be more likely to be interested in the entire "Regency" collection, regardless of the specific color of the teacup and saucer.

# Conclusion

In this study, I looked at the online retail data to try to identify key business drivers and segment the client base into actionable information for improved customer relationship management.

More specifically, using the k-means approach, I was able to isolate three key customer clusters, one of which contained 25 of the company's most valuable MVP customers.

In order to motivate actions, I applied an RFM analysis and divided the clients into more focused groups. CRM procedures can be referenced from the interactive datatable above.

Other key differentiators for more in depth division of customers into appropriate groups, other data that could be used include:

Demographics (age, race, religion, gender, family size, ethnicity, income, education level)

Psychographic (social class, lifestyle and personality characteristics)

Geography (more specific locations)

**K-means clustering algorithm & Disadvantages:**

K-means clustering is a simple and efficient method. Additionally, it can handle large data sets with ease. On the other hand, there are a few drawbacks to the k-means approach. The disadvantage of K-means clustering is that it requires us to predetermine the number of groups. K-means also has the drawback of being sensitive to outliers, which might lead to inconsistent findings if the data is rearranged.

RFM analysis is comparable. In the end, a company's ideology and culture would determine the method it would choose to segment its consumer base.

Additionally, the distinctive nature of the goods or services they offer would have a significant impact on the strategy they would like to employ, as well as how they would classify their clients and where they would like to concentrate their efforts in order to build stronger client relationships and, ultimately, a lucrative future.

**References**

Usama, M., Qadir, J., Raza, A., Arif, H., Yau, K. L. A., Elkhatib, Y., ... & Al-Fuqaha, A. (2019). Unsupervised machine learning for networking: Techniques, applications and research challenges. *IEEE access*, *7*, 65579-65615.

Hahne, F., Huber, W., Gentleman, R., Falcon, S., Gentleman, R., & Carey, V. J. (2008). Unsupervised machine learning. *Bioconductor case studies*, 137-157.

Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2020). A systematic review on supervised and unsupervised machine learning algorithms for data science. *Supervised and unsupervised learning for data science*, 3-21.

Kassambara, A. (2017). *Practical guide to cluster analysis in R: Unsupervised machine learning* (Vol. 1). Sthda.

Bhavsar, H., & Ganatra, A. (2012). A comparative study of training algorithms for supervised machine learning. *International Journal of Soft Computing and Engineering (IJSCE)*, *2*(4), 2231-2307.