MAY 2021

SUPERMARKET SALES DATA ANALYSIS

MSBA370 – DDDM

Streamlit Individual project

PREPARED BY:

Vera Al Ashkar

ABOUT THIS DATA DRIVEN DECISION DASHBOARD

Overview

I have been asked to create a tool that allows people with no coding knowledge to be able to create visuals, generate insights and make decision. What's better than a Streamlit dashboard? I created one that analyses a supermarket data that I get from Kaggle and generate insights about sales, customer behavior, payment types and many other options.

Dashboard Parts

This dashboard can be used to analyse any future data.

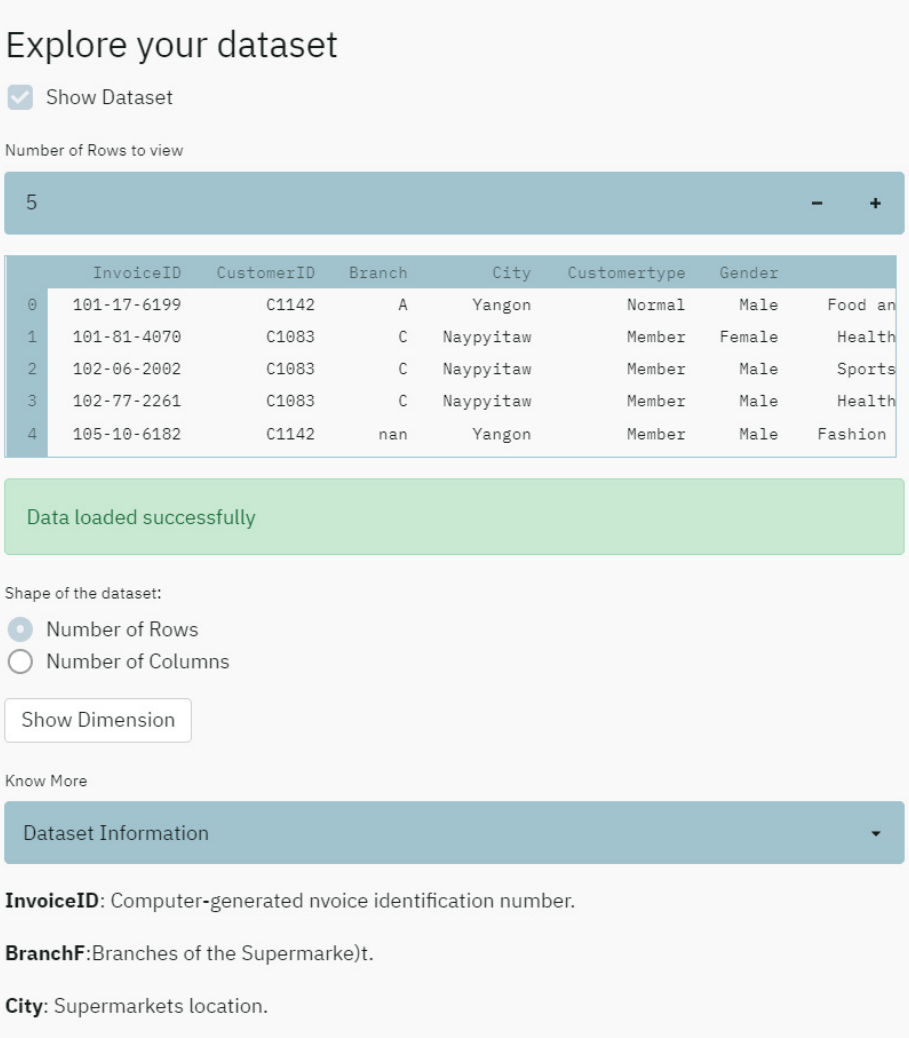
Dashboard parts are:

- Explore your data
- Create some visuals
- RFM Analysis
- Machine Learning Algorithm
- Upload your dataset

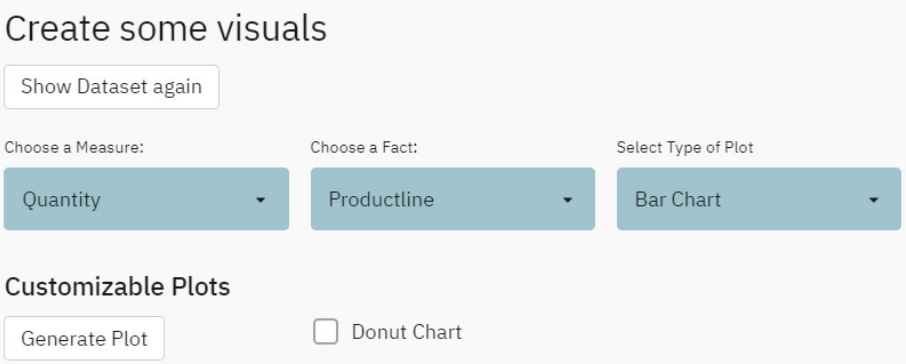
EXPLORE YOUR DATA

This part is made for the you to explore your data and take general overview of it, you will be able to:

1. Display the dataset
2. Know more information about it
3. Check its columns and rows
4. See their dimension and read the needed info about them
5. Select and explore his preferred columns
6. Obtain data summary
7. Check and drop missing values
8. Display customer distribution by city

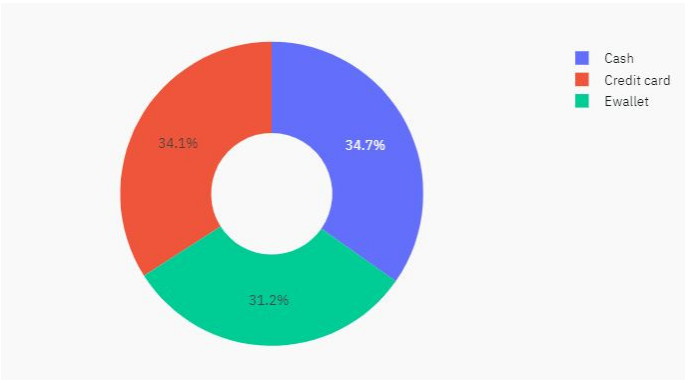


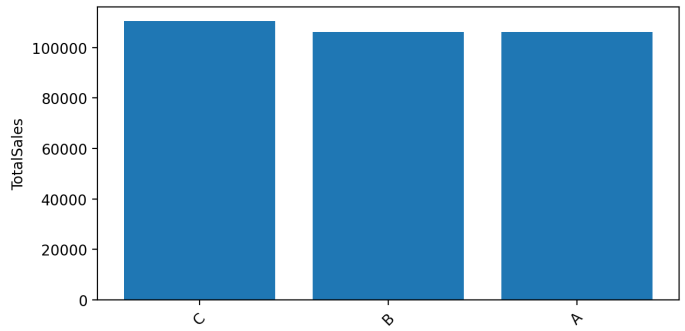
CREATE SOME VISUALS



How to understand your Data?

This part will give you the ability to customize Bar charts, Horizontal bars, and donut charts by playing with measures and facts.
This way you will get a visual summary of information that will make understanding your data much easier, you will be able to determine the relationship between two variables, analyze value and risk, and generate important insights.

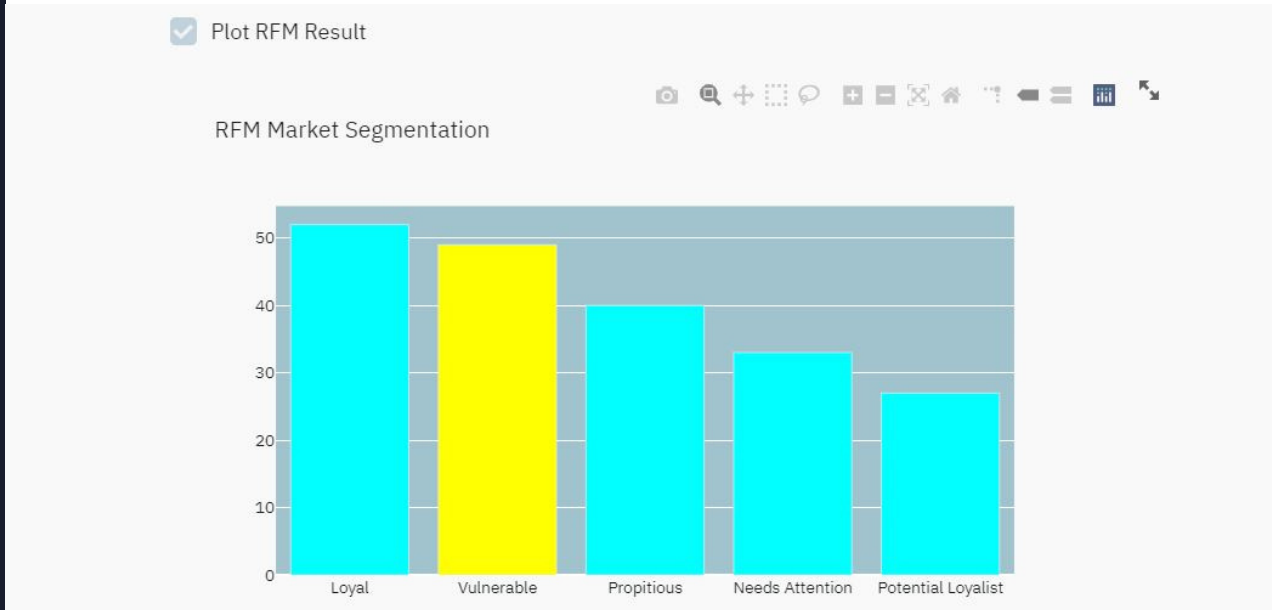




RFM ANALYSIS

RFM is a data-driven customer segmentation technique that allows you to take tactical decisions by segmenting users into homogeneous groups based on three quantitative factors:

1. **Recency:** How recently a customer has made a purchase
2. **Frequency:** How frequently does the customer makes a purchase
3. and **Monetary Value:** How much money has the customer spent



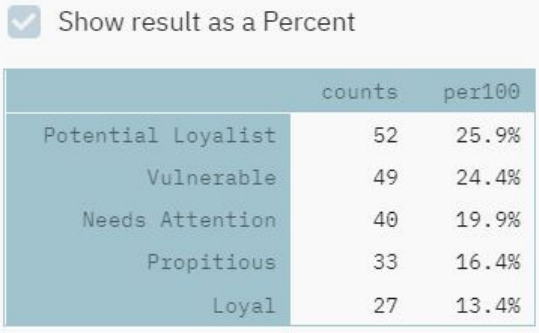
How to deal with each one of the Five Groups?

As shown in the screenshots below, after grouping your customers you will be able to display a bar chart that will give a good idea on how your customers are distributed.
This will help you knowing what to offer to each group:

- **Loyal Customers** are your champions, they are the customers who bought recently, buy most often and spend the most. They are the ones that you want to make sure they will always have the best customer service.
- **Vulnerable Customers** are at risk, who have purchased a long time ago and are required to bring them back! You might need to use some new marketing strategy with them (for example send them catchy offers, or maybe a premium shopping card)
- **Propitious Customers** are recent shoppers with above average spending. Focus on them to move them to loyal customers group.
- **Customers needing attention** are characterized with average recency ,frequency, and monetary values. These are customer that you basically lost, if you can't find a way to reconnect them back then get rid of them out of your email list, this way you will not be at risk of having a high spam email average.
- **Potential Loyalist Customers** are recent customers with above average frequency. Keep them with you, and give them offers to buy more.

What next?

This Clickbox will show you a table with grouping percentage, What can we conclude from it?
Loyal customers percentage is the highest. However, 24.4% of customer are vulnerable customers, and this risk is going to get bigger if no immediate strategy is applied



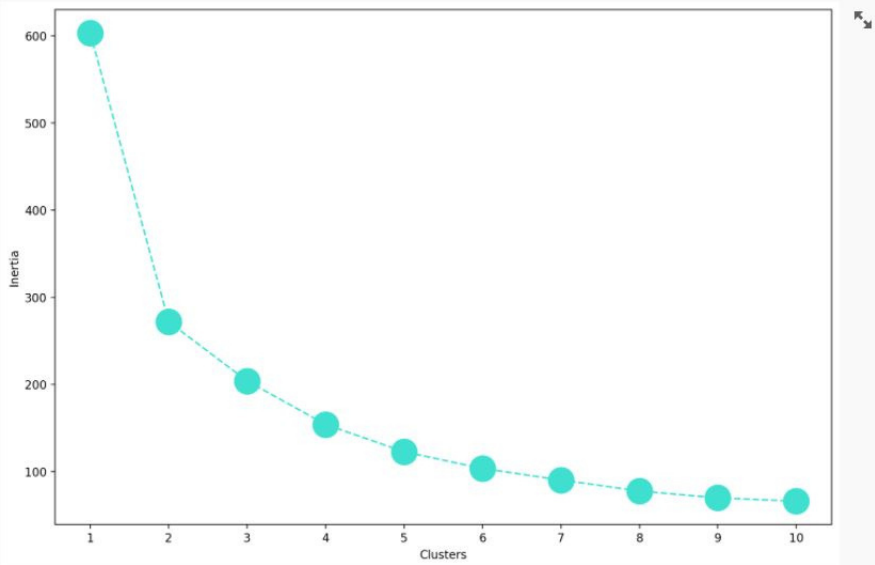
MACHINE LEARNING ALGORITHM

What is K-mean clustering and how it can be good for your business?

K-means clustering is an unsupervised machine learning that their purpose is to group observations that have similar characteristics, which means to group data points into distinct non-overlapping subgroups. We are using it to segment the supermarket customers to get a better understanding of them.

You can use cluster analysis to group data points according to the similarities between them. This practice has a widespread application in business analytics and can help you to achieve your business goals. You can use the k-means algorithm to maximise the similarity of data points within clusters and minimise the similarity of points in different clusters.

The Elbow curve is one of the most popular methods to determine this optimal value of k



Here we can conclude that the best k value is 5

The elbow curve

A fundamental step for any unsupervised algorithm is to determine the optimal number of clusters into which the data may be clustered.
The Elbow Method is one of the most popular methods to determine this optimal value of k.

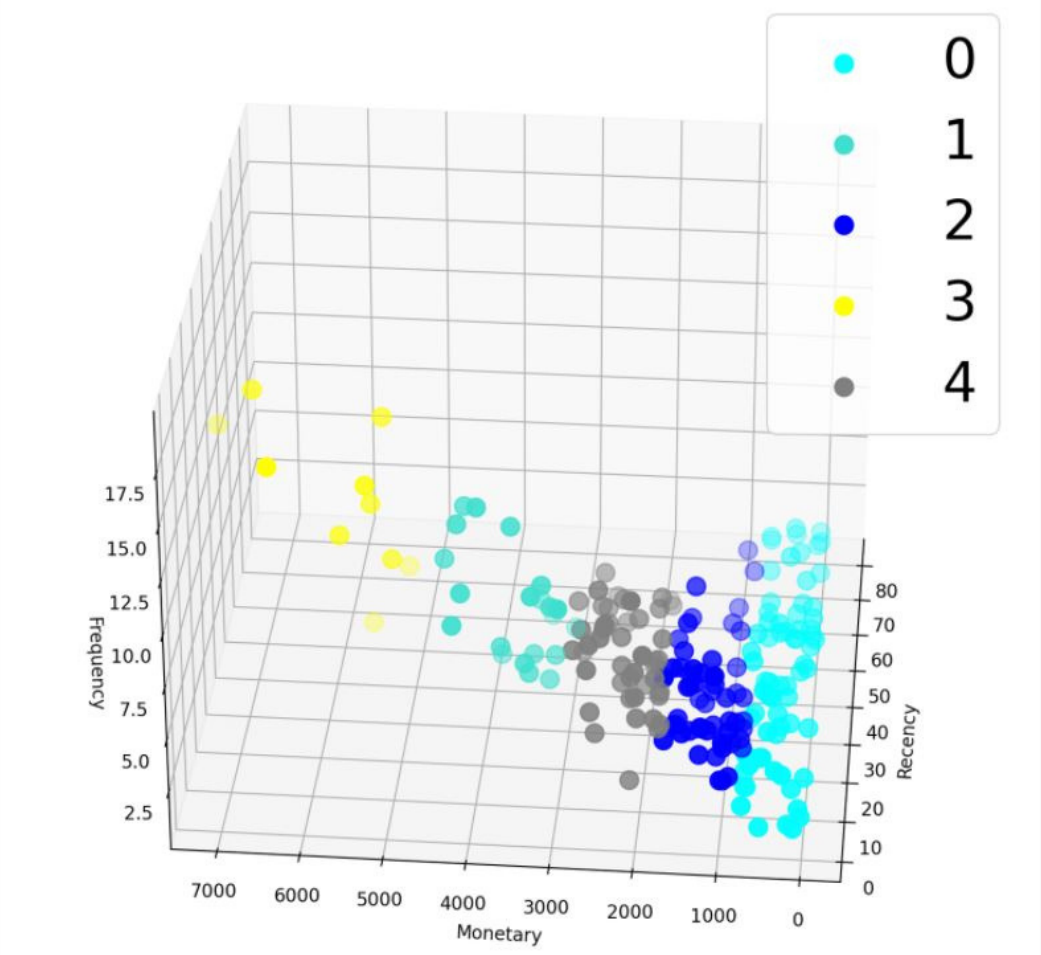
SELECTING THE NUMBER OF CLUSTERS WITH SILHOUETTE ANALYSIS

How can this help?

Silhouette analysis is used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters. The value of the silhouette score range lies between -1 to 1.

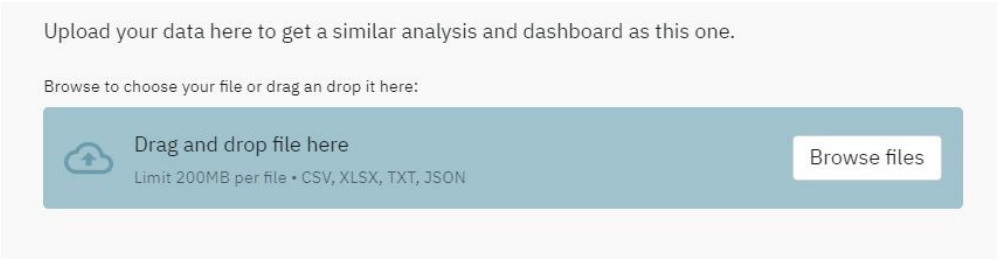
Using the silhouette, you can ensure that the optimal value of k is the same as the one obtained from the elbow curve.

Getting 5 as k value in both method ensure that we chose the optimal k value.



Upload your dataset

This option will let readers upload their dataset so we can analyze it and generate a dashboard for them similar to this one.



Using this report you can Turn your ideas into awesome apps, digital publications and interactive presentations without writing a single line of code.