1. Business Objectives: Analyze and understand customer behavior for better marketing and retention

Understand different types of customers based on

* Age Group
* Category Group
* Payment Method
* Seasonal Shopping Trends
* Customer Type (low, medium, or high spenders)

Improve marketing strategies

* Target each customer segments with different approaches
* Offering the right product to the right customer
* Optimize price and discount based on customers’ spending behavior

Enhance customer satisfaction and loyalty

* Identify high-value customers and reward them
* Offering membership program for frequent shoppers
* Reduce churn by analyzing customers’ interest and improve engagement

1. I use formula function to discrete “age”, “category”, “price”

A screenshot of a computer

AI-generated content may be incorrect.I use the if function to put values in specific groups that I want.

Select column: allow me to choose which attributes I want to use, and which attribute I want to predict (like rapid miner). I ignored “invoice date” (I already extract the month from it to use for “Season”. I leave out quantity and price as these are not strong attributes for my prediction. I use “price” to generate attribute “customer type” by using if function. Attributes that I have already discrete into useful attributes, I will ignore.

Another thing is that I would want to leave the “target” column blank as I’m doing unsupervised learning to identify customer segments.

A screenshot of a computer

AI-generated content may be incorrect.Data sampler: allow me to split data. The original data include 100k samples while K means can only run up to 5000 samples, so I use it to randomly pick out 5000 samples for use.

A diagram of data sampler

AI-generated content may be incorrect.

I use distribution to display attributes

A screenshot of a computer

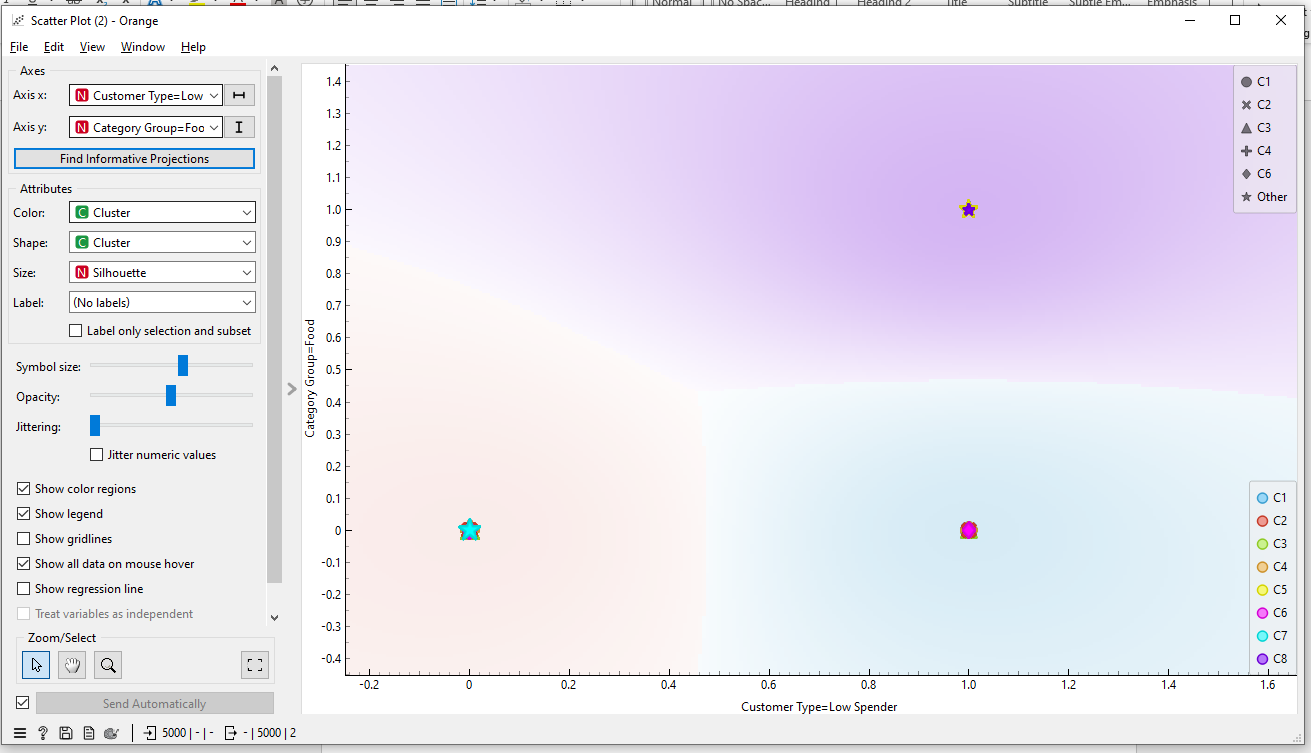
AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.A diagram of a diagram

AI-generated content may be incorrect.I use K mean and hierarchical to cluster the 5000 samples that I pick out and scatter plot to see the result. For K means, I specically choose cluster numbers from 3 to 10 and choose 8, as it is the highest.

Continuize allows me to change my data into numeric. The K mean model on the bottom show more details how each attributes behave with others, while the middle one, with no continuize give me a more overall look about the visualization.

This is the bottom scatter plot

A screenshot of a computer

AI-generated content may be incorrect.

This is the middle scatter plot

A diagram of a cluster of data

AI-generated content may be incorrect.This is my hierarchical clustering. I must connect to distances because the model requires.

A screenshot of a computer

AI-generated content may be incorrect.

1. In Orange, I noticed that there were no K medoid so I looked up and see that Orange’s K mean can run both numeric and categorical. I tried both numeric and categorical for K mean and concluded with unexpected outcomes. I tried to use outlier and see that my dataset with 5-7 attributes does not need it. DBSCAN is not as effective as K means and hierarchical clustering when running categorically. I run all attributes at first and leave out some later. I noticed that the clustering was easier to see, and I can tell more insight from it than with lots of attributes.
2. I got the dataset from Kaggle. I first found it on Data.gov, but I can’t find the link to go back. Then later I saw it again, so why not take it when it comes to me twice. First, my goal is to predict sales for Walmart, but I can’t find suitable data. Walmart offers data publicly but not much and has to pay for full dataset.
3. With the scatter plot, I can see how each category behaves. For example, Axis x (low spender) and Axis y (Shopping), would produce different color segments. I identified clusters with low spending as customer type, paying cash, purchasing entertainment products in winter as budget shoppers. My marketing strategy for them would be discount promotions.
4. At first, I did not want to identify customer segmentation, but then I found out that the data that I prep can be used more effectively with unsupervised learning. I did 2 models, and I also have predicted sales for this dataset too. The upper set of this unsupervised learning is where I regretted the most, but I would just leave it there, since I spent time and created it. I created an attribute including categories group by payment method. I want the algorithm to see if there are any connections or patterns, but it just creates more noise for the model. The outcome of it looked unrealistic. I had good time preparing the data, exploring more tools, as well as another software. I watched some videos and instruction how to use Orange. They recommended it as user friendly, but when I first tried it, I did not know where to start. If there is next time, I would want to do both sales predictions and identifying customer segments to give better business analysis. I completed my sales predictions, but when looking at other peers, I think I need to improve it more.