# RFM And K-means for Customer Segmentation

Capstone Project - The Battle of Neighborhoods (Week 2) - Report

# Introduction

I may have a convenience store after several years, then I will have many customers. I can use RFM (Recency, Frequency, Monetary Value) analysis to find out the customer structure.

What is RFM Analysis? RFM (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 recently a customer has made a purchase.
- Frequency: How often a customer makes a purchase.
- Monetary Value: How much money a customer spends on purchases.

After I have a RFM data, I will split the numbers into segments, then I can use K-means to find out how many customer classes I have, and what to do with them.

# Problem

- Who are the best customers?
- Who are your loyal customers?
- Which customer is losing interest?
- Which customer have lost risk?
- Who are the lost customers?
- Does the store still run well?
- What should I do if it not well?

# Data

```
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    541909 non-null  object
 1   StockCode    541909 non-null  object
 2   Description  540455 non-null  object
 3   Quantity     541909 non-null  int64
 4   InvoiceDate  541909 non-null  object
 5   UnitPrice    541909 non-null  float64
 6   CustomerID   406829 non-null  float64
 7   Country      541909 non-null  object
```

The data have "InvoiceNo", "StockCode", "Description", "Quantity", "InvoiceDate", "UnitPrice", "CustomerID", "Country" in it, we need CustomerID to collect "Frequency". Quantity and UnitPrice to calculate the total amount which needed by "Monetary". InvoiceDate should be transformed into datatime for calculating "Recency".

# Data clean

- We don't need "InvoiceNo","StockCode","Country","Description", drop them from dataset.
- There is 135080 rows null data in CustomerID, the row should be deleted from the dataset.
- Find 5227 duplicated data and delete the rows from the dataset.
- Quantity and UnitPrice should not be less than 1 and should not be negative number.
- Format InvoiceDate, from string to datatime.

# Data processing

| CustomerID | R | F | M |
|---|---|---|---|
| 12346 | 325 | 1 | 77183.60 |
| 12347 | 2 | 182 | 4310.00 |
| 12348 | 75 | 31 | 1797.24 |
| 12349 | 18 | 73 | 1757.55 |
| 12350 | 310 | 17 | 334.40 |

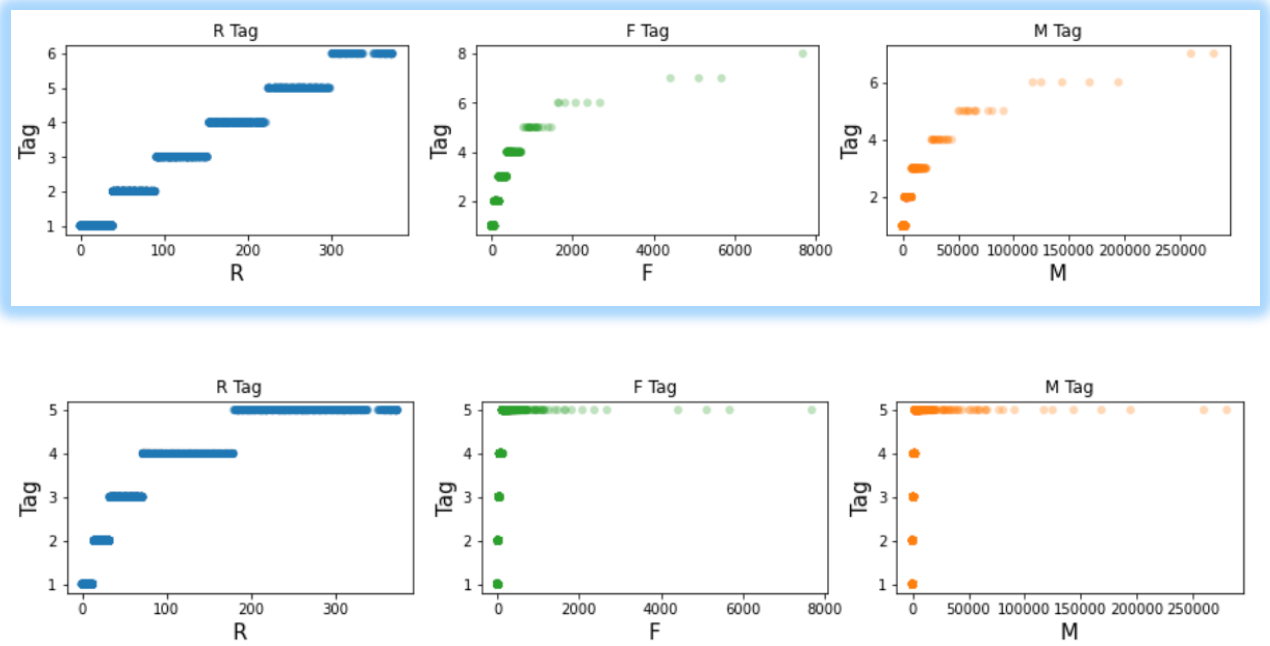
Find min and max date in InvoiceDate, use the next day as collectTime. Use collectTime and InvoiceDate to calculate the Recency (How recently a customer has made a purchase), named R.

Group by CustomerID and count rows number, it is Frequency (How often a customer makes a purchase), named F.

We use Quantity* UnitPrice as total amount, Group by CustomerID and sum total amount as Monetary Value (How much money a customer spends on purchases), named M.

Join R, F, M table by CustomerID and we got RFM data.

# Data segmentation




Results for Jenks Natural Breaks and use Percentile (20%, 40%, 60%, 80%) split

For 1D data, we use Percentile (20%, 40%, 60%, 80%) or Jenks Natural Breaks to split the data, then chose a better group

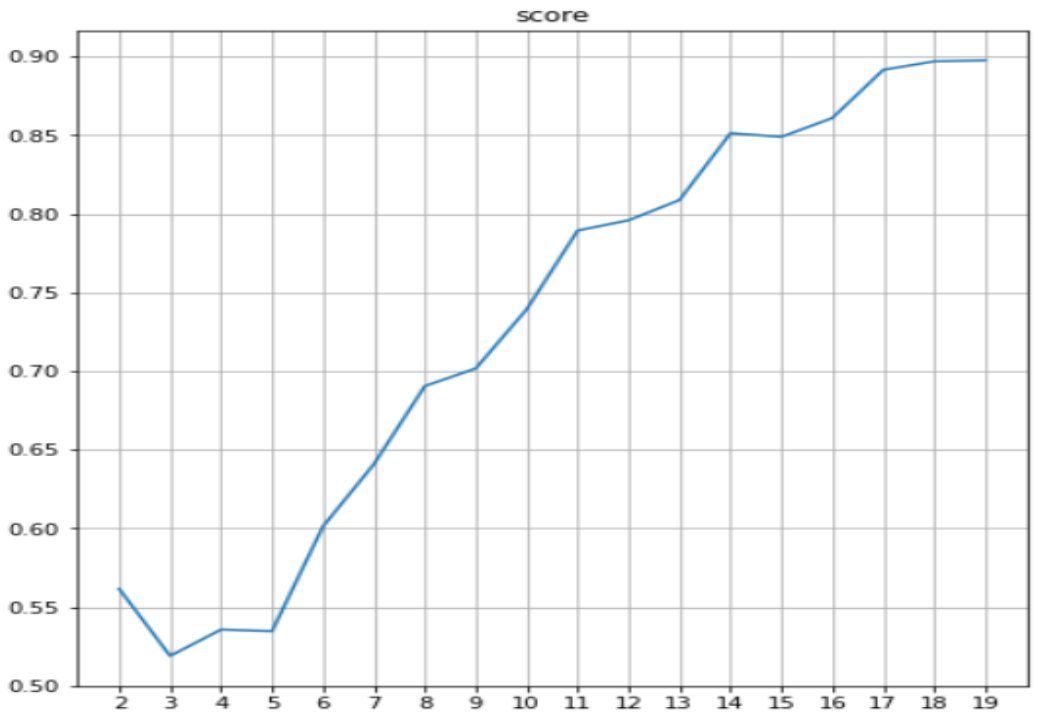
The Jenks Natural Breaks Classification (or Optimization) system is a data classification method designed to optimize the arrangement of a set of values into "natural" classes. A Natural class is the most optimal class range found "naturally" in a data set. A class range is composed of items with similar characteristics that form a "natural" group within a data set.

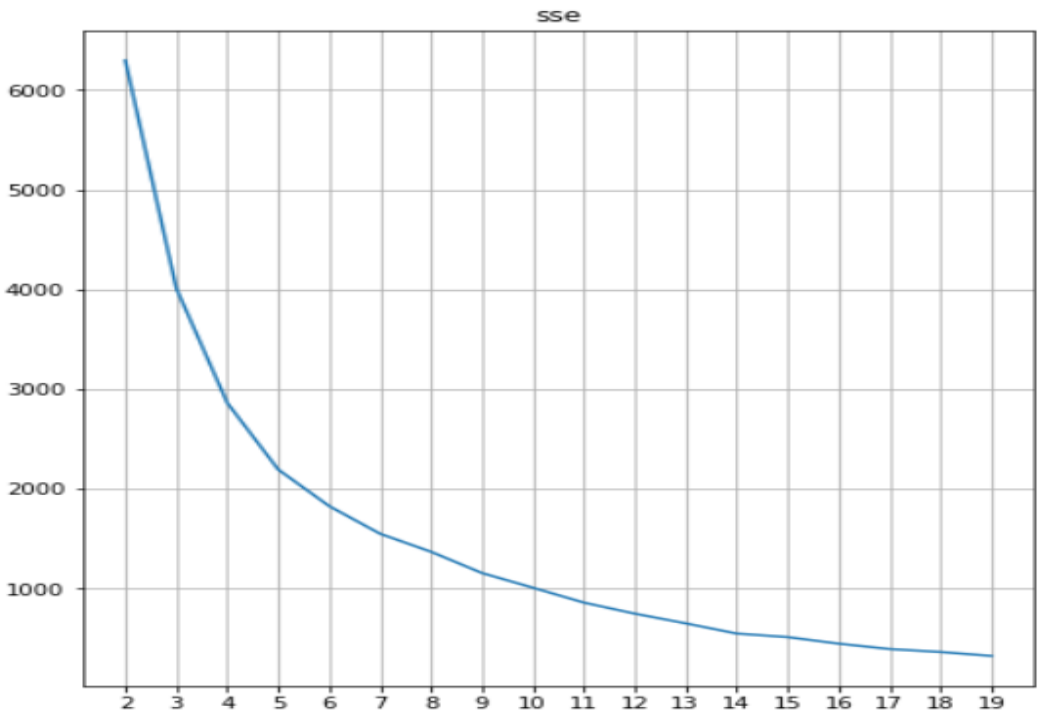
This classification method seeks to minimize the average deviation from the class mean while maximizing the deviation from the means of the other groups. The method reduces the variance within classes and maximizes the variance between classes. It is also known as the goodness of variance fit (GVF), which equals the subtraction of SDCM (sum of squared deviations for class means) from SDAM (sum of squared deviations for array mean).

# RFM segments

| type | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| R | [min,3528.0] | (3528.0,3579.0] | (3579.0,3642.0] | (3642.0,3712.0] | (3712.0,3789.0] | (3789.0,max] | | |
| F | [min,68.0] | (68.0,182.0] | (182.0,378.0] | (378.0,740.0] | (740.0,1477.0] | (1477.0,2677.0] | (2677.0,5670.0] | (5670.0,max] |
| M | [min,2392.83] | (2392.83,8347.20] | (8347.20,21429.39] | (21429.39,44534.3] | (44534.3,91062.38] | (91062.38,194390.79] | (194390.79,max] | |

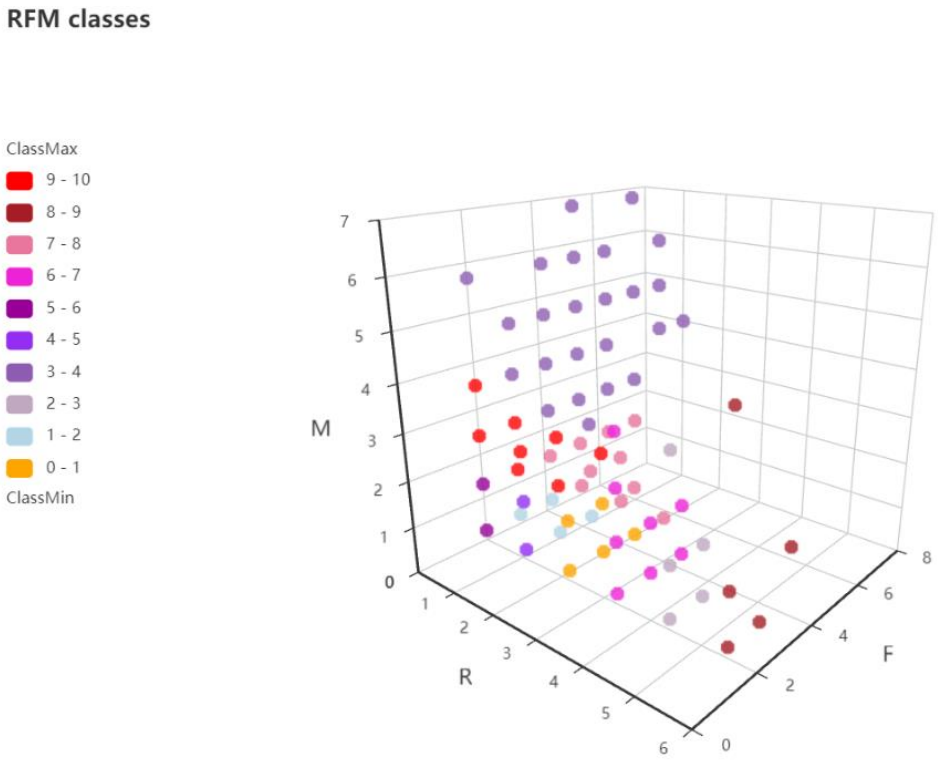
# Find Best K for K-means

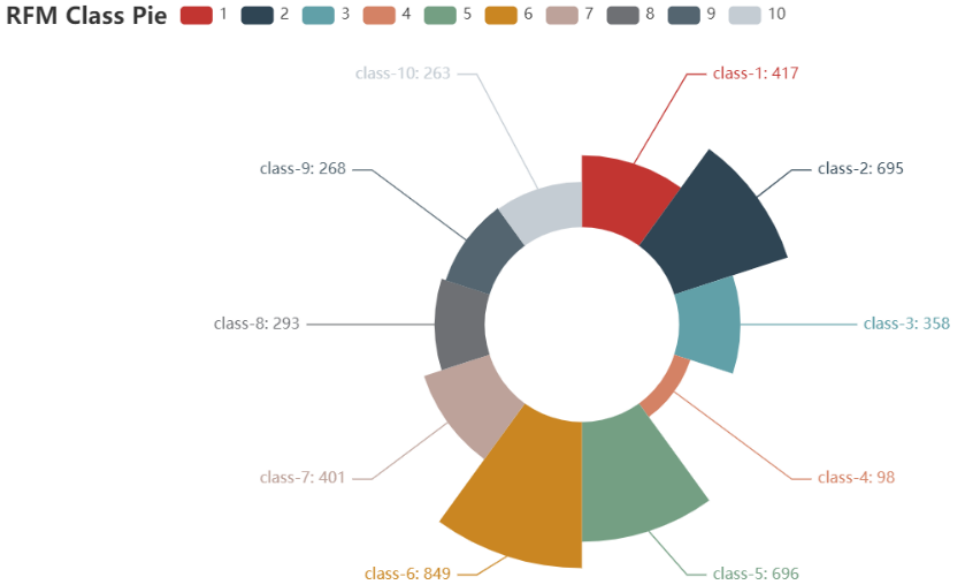



We run 20 times K-means and compare silhouette score and sse, the best K is 10.

# Classify by K-means




| R | F | M |
|---|---|---|
| 1.000000 | 1.000000 | 1.038869 |
| 5.428115 | 1.055911 | 1.023962 |
| 1.081081 | 3.260442 | 1.825553 |
| 2.000000 | 1.000000 | 1.021552 |
| 4.000000 | 1.132832 | 1.032581 |
| 2.280702 | 2.108187 | 1.251462 |
| 1.000000 | 2.000000 | 1.000000 |
| 1.000000 | 4.242424 | 3.863636 |
| 3.000000 | 1.000000 | 1.018182 |
| 1.019802 | 1.970297 | 2.163366 |

classify cluster center

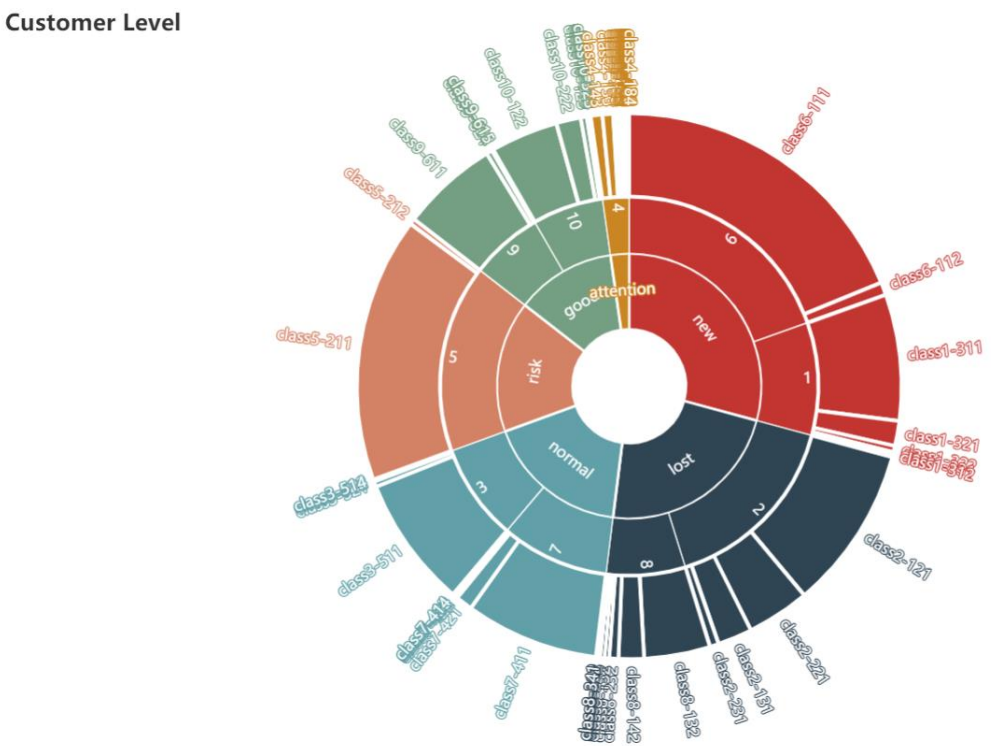
Classes

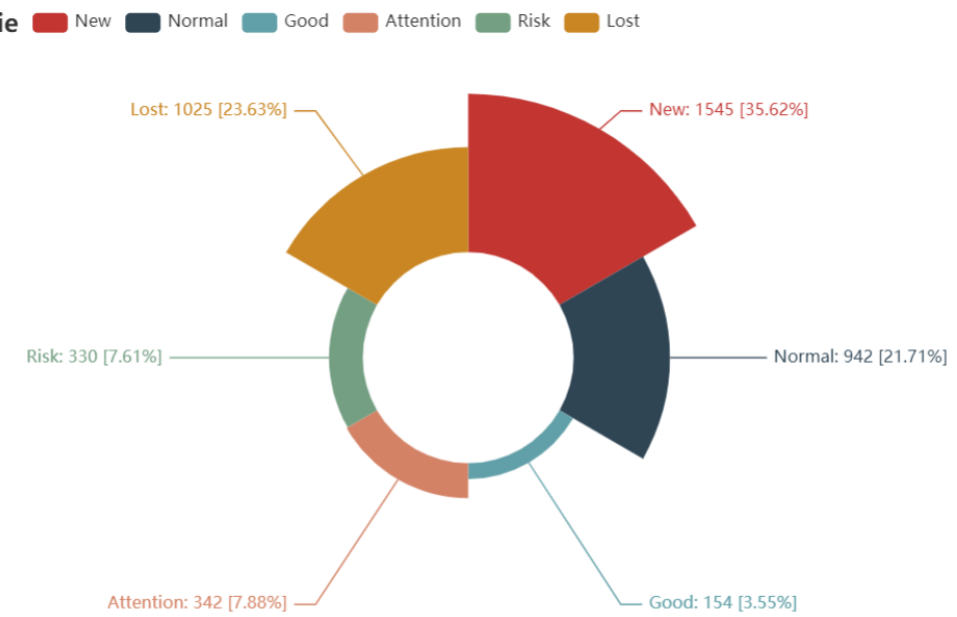
Proportion

Cluster centers

We created K-means model. K is 10, score is 0.749867, sse is 1002.771267.

# Conclusion




Lost customers are 23.63%. Risk customers are 7.61%. Attention customers are 7.88%. Normal customers are 21.71%. New customers are 33.62%. Good customers are 3.55%.

Lack of customer stickiness: total of level risk and attention is more then 15%, need to do preference analysis for them. Some promotion activities should be useful.

Level new is 33.62%, member points and gifts may helpful.

The goods structure is not fit consume preference: level good is much smaller than level normal. Level lost is about 1/4, need more promotion activities and advertising to bring them back.