# Customer Segmentation

**Wholesale customers Data Set**

Presenter: Fan Wang

Feb 5, 2020

# Project Introduction

## Dataset:

- wholesale customer dataset from UCI Machine Learning Repository
- The dataset refers to clients of a wholesale distributor. It includes the annual spending in monetary units (m.u.) on diverse product categories.

## Goal:

- Apply clustering techniques to identify segments that are relevant for certain business activities, such as rolling out a marketing campaign.

| Attribute | Description | Datatype |
|---|---|---|
| FRESH | annual spending (m.u.) on fresh products | Continuous |
| MILK | annual spending (m.u.) on milk products | Continuous |
| GROCERY | annual spending (m.u.)on grocery products | Continuous |
| FROZEN | annual spending (m.u.)on frozen products | Continuous |
| DETERGENTS_PAPER | annual spending (m.u.) on detergents and paper products | Continuous |
| DELICATESSEN: | annual spending (m.u.)on and delicatessen products | Continuous |
| CHANNEL | Channel - Horeca (Hotel/Restaurant/Cafe or Retail channel | Nominal |
| REGION | Regions - Lisnon, Oporto or Other (Nominal) | Nominal |

Attribute Information

# Processing Technique

## Domain Background

- Market segmentation: The overall aim of segmentation is to identify high yield segments – that is, those segments that are likely to be the most profitable or that have growth potential – so that these can be selected for special attention (i.e. become target markets). [Source: Wikipedia]

## Technique:

- Unsupervised Learning
- K-means Clustering: Identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.

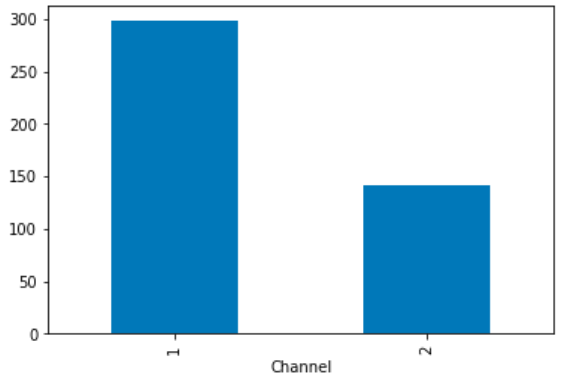
## How it works:

- Starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster.
- Performs iterative (repetitive) calculations to optimize the positions of the centroids.

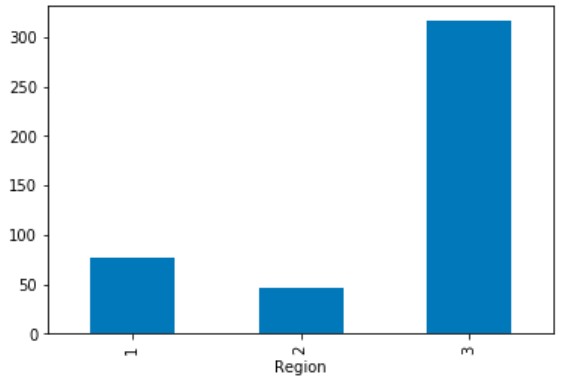
# Data Prepressing(categorical data)

What is the Problem with Categorical Data?

- Some algorithms can work with categorical data directly.
- For example, a decision tree can be learned directly from categorical data with no data transform required (this depends on the specific implementation)
- Many machine learning algorithms cannot operate on label data directly. They require all input variables and output variables to be numeric.




Channel: Horeca, Retail



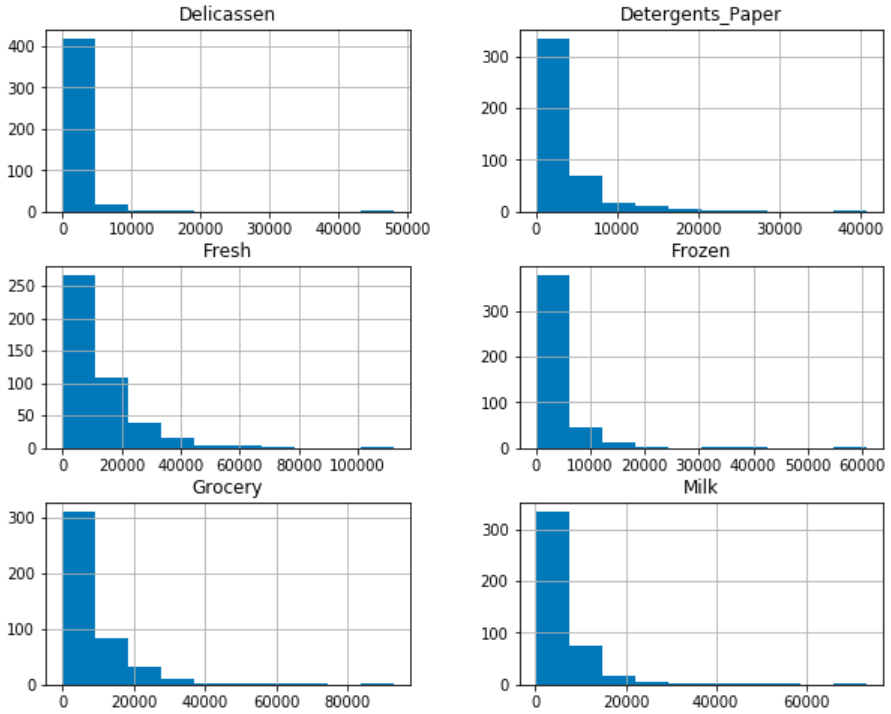

Region: Lisbon, Oporto, Others

Rename the attribute (manually one-hot encoding)

# Data Prepressing(continues data)

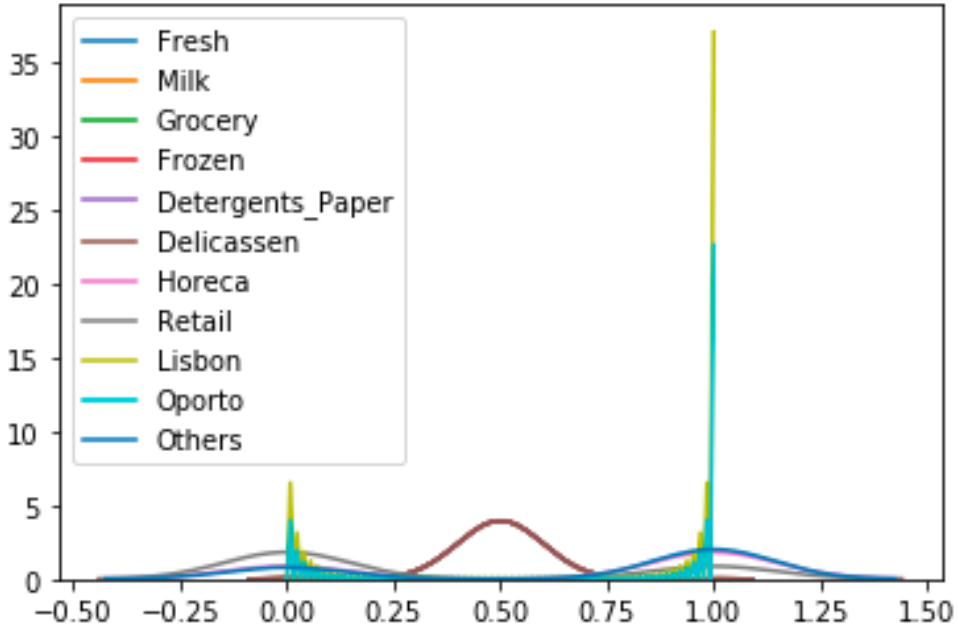
## Handling data before import into the model

- All continues variables are normally distributed;
- All variables are on the same scale. $Z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$




Normalize

scale



# Model Implementation(comparison)

## Find the best K parameters

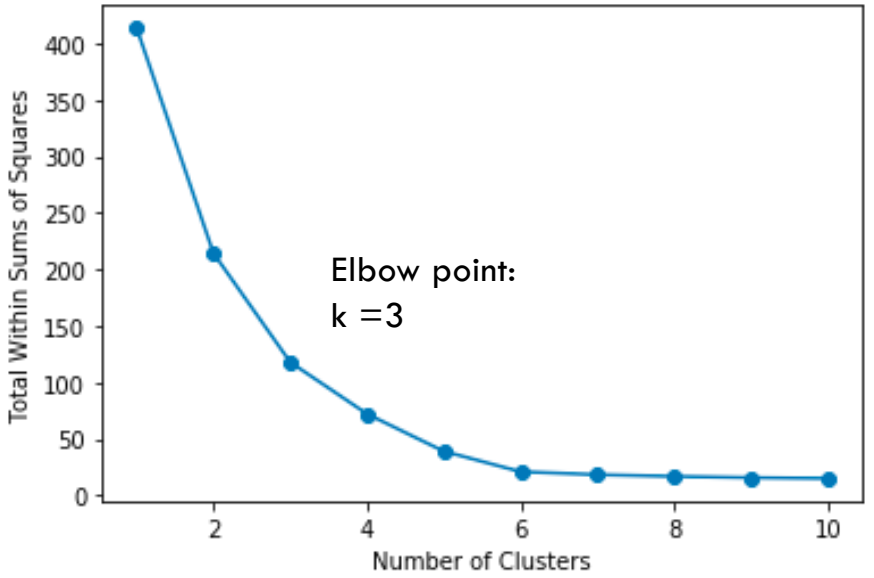
- Trial and error
- Elbow method: Sums of Square Errors(SSE)

$$\sum_{i=1}^{n}(y_i - \bar{y})^2$$

| | cluster | Fresh |
|---|---|---|
| **0** | 0 | 59 |
| **1** | 1 | 211 |
| **2** | 2 | 105 |
| **3** | 3 | 28 |
| **4** | 4 | 37 |

| | cluster_3 | Fresh |
|---|---|---|
| **0** | 0 | 87 |
| **1** | 1 | 211 |
| **2** | 2 | 142 |

Optimal K ?




## Things should consider:

- Cost
- Time
- Marketing Campaign Strategy

# Conclusion

Suppose we use 3 clusters:

| | cluster_3 | Fresh | Milk | Grocery | Frozen | Detergents_Paper | Delicassen | Horeca | Retail | Lisbon | Oporto | Others |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 12499.402299 | 3366.218391 | 4145.011494 | 3969.804598 | 799.965517 | 1167.781609 | 1.0 | 0.0 | 0.678161 | 0.321839 | 0.000000 |
| 1 | 1 | 13878.052133 | 3486.981043 | 3886.734597 | 3656.900474 | 786.682464 | 1518.284360 | 1.0 | 0.0 | 0.000000 | 0.000000 | 1.000000 |
| 2 | 2 | 8904.323944 | 10716.500000 | 16322.852113 | 1652.612676 | 7269.507042 | 1753.436620 | 0.0 | 1.0 | 0.126761 | 0.133803 | 0.739437 |

Conclusion:

- Cluster 0 : Customers are most from Hotel/Restaurant/Café Channel, they are from Lisbon and Oporto. Then tend to spend more on Fresh and Frozen products.
- Cluster 1 : Customers are mostly from Hotel/Restaurant/Café Channel, they are from Region Others. These customers tend to spend more Fresh, Frozen and Delicassen products.
- Cluster 2 : Customers are mostly from Retail channel. These customers tend to spend more on Milk, Grocery and Detergents_Paper products.

- Strategy: Giving discounted pricing based promotional campaigns for the target group to retain them.

# Future work

Other Models:

- K-methods
- GMM
- Hierarchical clustering

Questions?