Agenda

1. Research question
2. Introduction to my dataset
3. A description of the ML models: K-means algorithm
4. An overview of data analysis and results
    - 4.1. Data preparation
    - 4.2. Elbow Method
    - 4.3. K-means clustering
5. Conclusion and Discussion

Research question: How firms understand their customers?

- Understanding customers is the most important aspect of any business
- Through customer segmentation, firms gain more insight into their customers, and their strategies can be targeted to the right customer group.
- There are many ways to segment customers that firms can apply depending on the stage of business development.
- Introduction a popular method based on three key pieces of information:
  - Recency (The last period when a customer made a transaction)
  - Frequency (The frequency of a customer’s purchases)
  - Monetary (The amount of money customers spend on your business).

Introduction to my dataset: E-Commerce Dataset from Kaggle

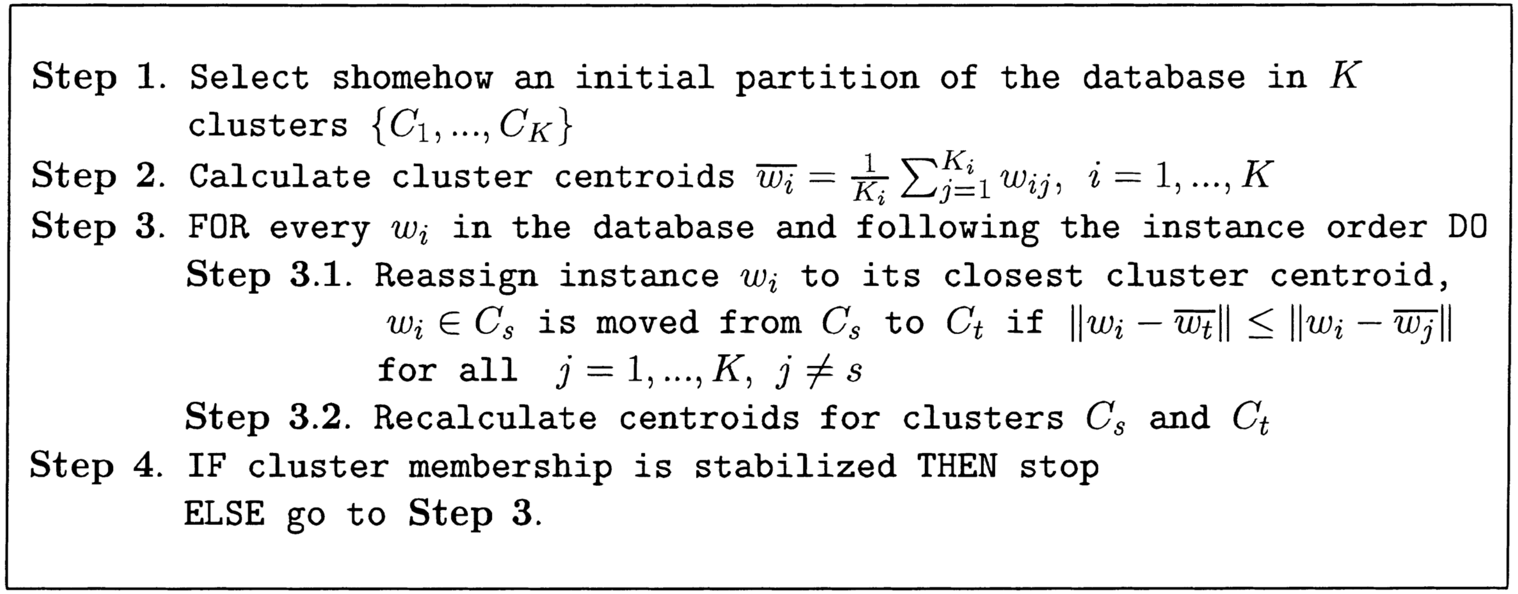
InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	United Kingdom
536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	United Kingdom
536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
…							

Table 1: E-Commerce Dataset

- **Dataset for analysis:** E-Commerce Data (click to download the dataset)
- **Introduction:** This dataset contains all actual transactions from 01/12/2010 to 09/12/2011 for a UK-based, registered non-store online retailer. The company primarily sells unique all-occasion gifts, with many of its customers being wholesalers.
  - 8 columns
  - 541,909 rows

A description of the ML models: K-means algorithm

The main K-means algorithm basically includes four small steps:



```
Step 1. Select shomehow an initial partition of the database in $K$
        clusters $\{C_1, ..., C_K\}$
Step 2. Calculate cluster centroids $\overline{w_i} = \frac{1}{K_i} \sum_{j=1}^{K_i} w_{ij},\ \ i = 1, ..., K$
Step 3. FOR every $w_i$ in the database and following the instance order DO
        Step 3.1. Reassign instance $w_i$ to its closest cluster centroid,
                  $w_i \in C_s$ is moved from $C_s$ to $C_t$ if $\|w_i - \overline{w_t}\| \leq \|w_i - \overline{w_j}\|$
                  for all  $j = 1, ..., K,\ \ j \neq s$
        Step 3.2. Recalculate centroids for clusters $C_s$ and $C_t$
Step 4. IF cluster membership is stabilized THEN stop
        ELSE go to Step 3.
```

Figure 1: The pseudo-code of the K-Means algorithm

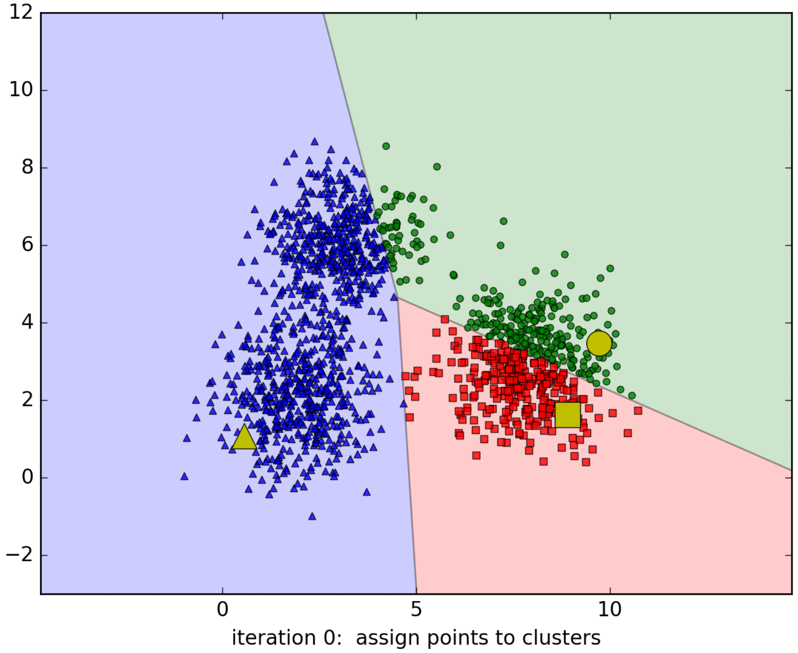
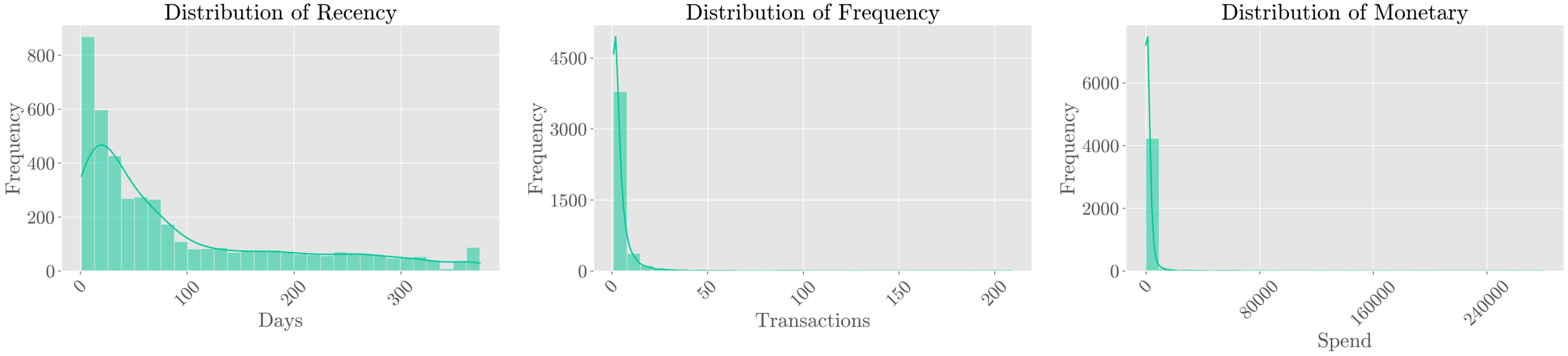


Figure 2: Visualizing K-means Clustering

An overview of data analysis and results: Data preparation

- Step 1: data cleaning and missing value handling
- Step 2: dimensionality reduction and feature engineering
- Step 3: data transformation and feature scaling (Figure 3, 4)

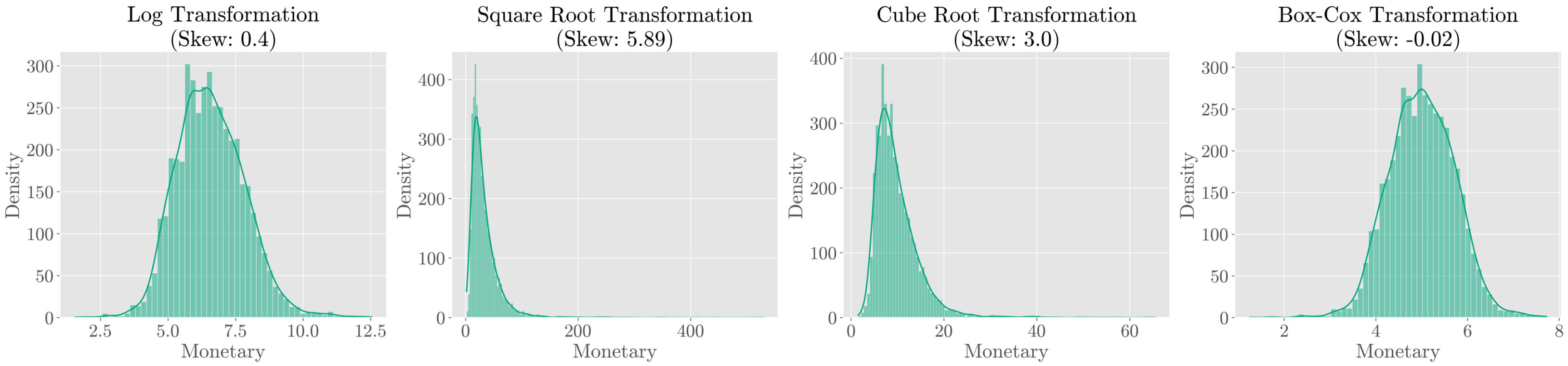


Distribution of Recency

Distribution of Frequency

Distribution of Monetary

Figure 3: The distribution of Recency, Frequency and Monetary



Log Transformation (Skew: 0.4)

Square Root Transformation (Skew: 5.89)

Cube Root Transformation (Skew: 3.0)

Box-Cox Transformation (Skew: -0.02)

Figure 4: Apply transformation for Monetary.

An overview of data analysis and results: Elbow Method

- By calculating the Sum of Squared Errors (SSE) and using the Elbow Method to plot the change in SSE, we can see that as we increase the number of k clusters, the error decreases (Fig. 5).
- From k = 3 the decrease of error becomes slower. Therefore, k = 3 is the ideal number of clusters for the K-means algorithm.

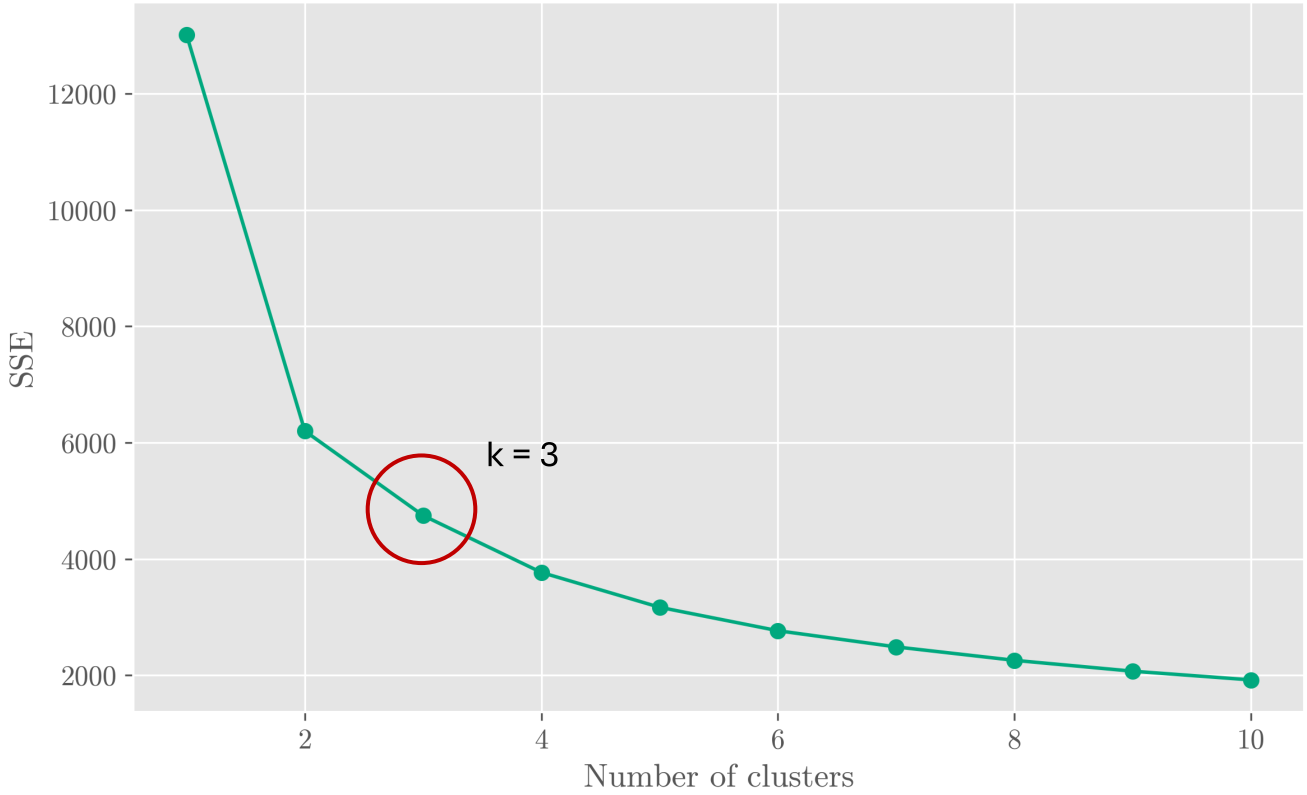


Figure 5: Elbow method for optimal number of k clusters.

An overview of data analysis and results: K-means clustering

Run	Cluster 0	Cluster 1	Cluster 2	Number of Points	SSE
1	(0.71, -1.04, -0.89)	(-1.05, 1.23, 1.15)	(-0.04, 0.25, 0.14)	(1649, 1072, 1617)	4745.72
2	(-1.04, 1.23, 1.15)	(0.71, -1.04, -0.89)	(-0.04, 0.25, 0.14)	(1072, 1649, 1617)	4745.73
3	(-0.03, 0.25, 0.14)	(-1.04, 1.22, 1.15)	(0.71, -1.05, -0.89)	(1614, 1077, 1647)	4745.75
4	(0.71, -1.05, -0.89)	(-1.05, 1.23, 1.15)	(-0.03, 0.25, 0.14)	(1649, 1074, 1615)	4745.71
5	(0.71, -1.04, -0.89)	(-1.05, 1.23, 1.15)	(-0.04, 0.25, 0.14)	(1649, 1072, 1617)	4745.72
...					

Table 2: K-means clustering result

- K-means algorithm achieved the initial objective of the analysis. The data points representing customers with the same characteristics are grouped into the same cluster.
- The results of each trial are not the same. The difference in initial centroids creates a difference in final centroids, and the number of points in each cluster changes with each run.
- The movement of points from one cluster to another is sometimes as high as 30%

Conclusion and Discussion

- K-Means is a powerful tool for customer segmentation. Enables businesses to tailor strategies for different customer groups.
- Real-world data that was not immediately suitable for applying the machine learning algorithm.
- The K-means algorithm still has several limitations: strict requirements for input data and the number of clusters k; results that depend on the initial centroids.