BUSINESS CASE

UK ONLINE RETAILER

CUSTOMER CLUSTERING & RECOMMENDATION SYSTEM

FRAN LLAMAS

INDEX

INTRODUCTION

The dataset found in the business case stores all the transactions made by different customers in a span of aproximately 1 year and a half.

01 Over 4000 different product categories and IDs

02 4372 different customers

03 More than 30 countries worldwide

04 Over 25000 transactions with multiple products

UNDERSTANDING THE DATA

InvoiceNo
- Categorical data type
- ID associated to each transaction

Stockcode
- Categorical data type
- ID associated to each product

Description
- Categorical data type
- Name and a small description of the product

Quantity
- Numerical data type
- Total units of product per transaction

InvoiceDate
- Datetime data type
- Exact date and hour of the transaction

UnitPrice
- Numerical data type
- Price per unit of product

CustomerID
- Categorical data type
- Unique identifier for each customer

Country
- Categorical data type
- Country of residence for each customer

PROCESS DETAILED

Cleaning & Preprocessing

- Duplicates
- Nulls
- Matching Description & ID
- Cancelled transactions
- Non-numeric product IDs
- 0 unit price

Feature Engineering

- Aggregate customer data
- Feature engineering

Correlation & Normalization

- Correlation Matrix
- Heatmap
- Multicollinearity
- Scaling with “StandardScaler”

K-means clustering

- Elbow Method
- Silhouete Method
- K-means algorithm
- Evaluation metrics

Recommendation System

- Top products for cluster
- Customer purchases
- Recommendations for each cluster
- Avoid products already purchased

FEATURE ENGINEERING

Time since last purchase (Recency)

- Difference from the most recent date and the last purchase of each customer

Number of purchases (Frequency)

- Total transactions done by each customer

Total spending (Monetary)

- Total spent by each customer in all the transactions collected in the database

Average Order Value

- Average spent per transaction by each customer

Products purchased

- The total amount of product units purchased by each customer

Unique products purchased

- The amount of unique products (how many different IDs)

Cancelled transactions frequency

- Out of the total purchases done by each customer, the percentage of cancelled ones

Favorite day & hour

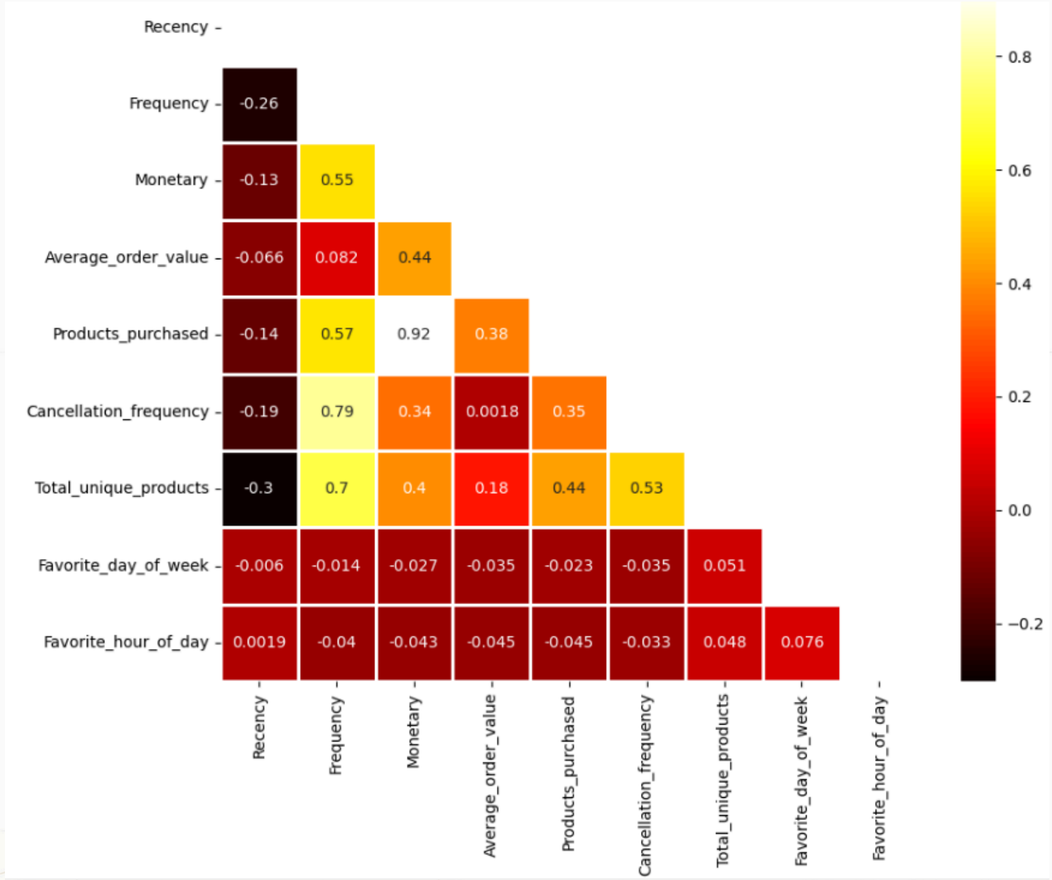
- Favorite day of the week, and favorite hour of the day to purchase by each customer

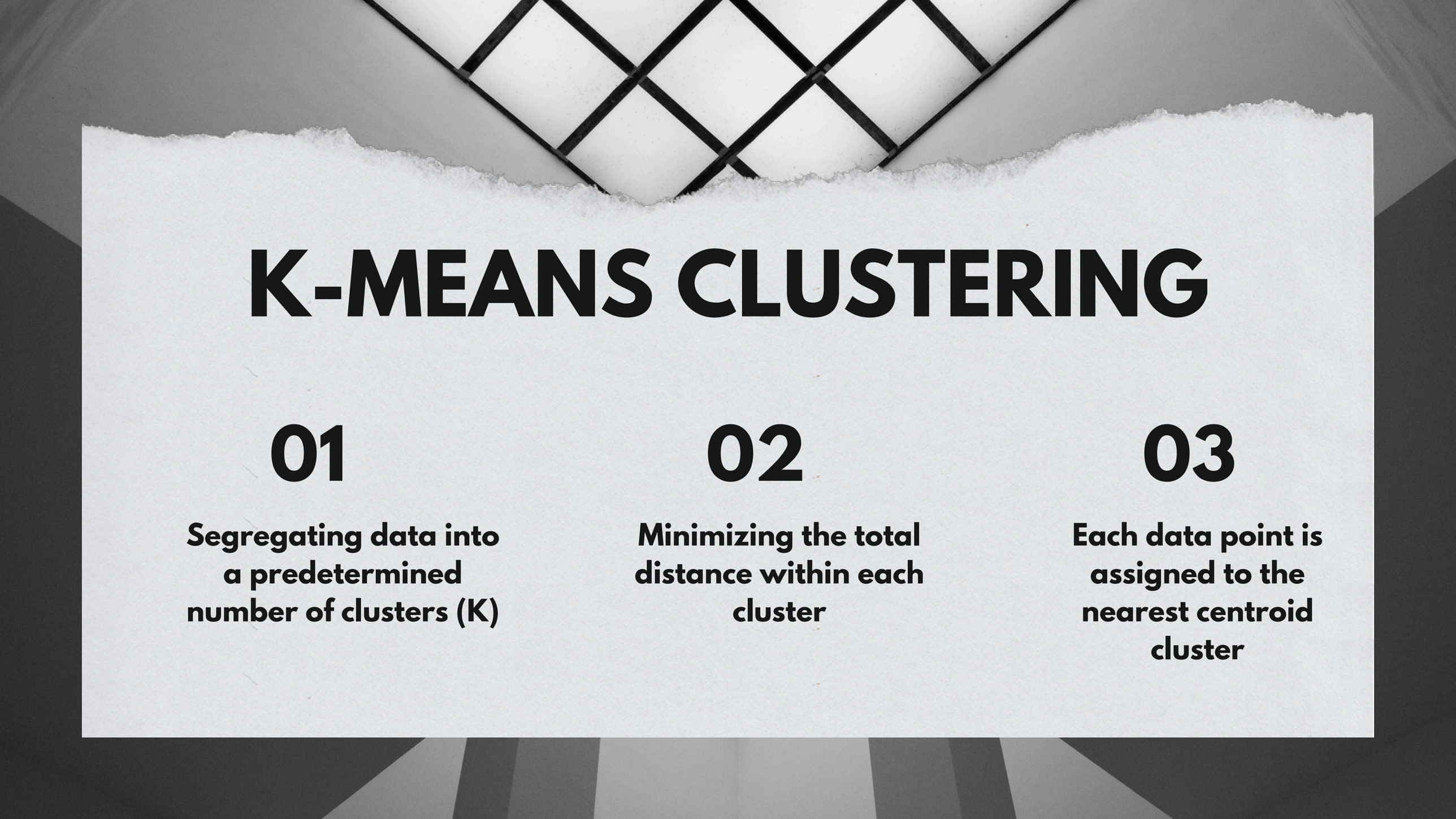
CORRELATION HEATMAP

Multicollinearity for values > 0,7

- **Total unique products - Frequency**
- **Cancellation frequency - Frequency**
- **Products purchased - Monetary**

Methods like PCA (Principal Component Analysis) help reducing the dimensionality of these pairs before scaling.





K-MEANS CLUSTERING

01

Segregating data into a predetermined number of clusters (K)

02

Minimizing the total distance within each cluster

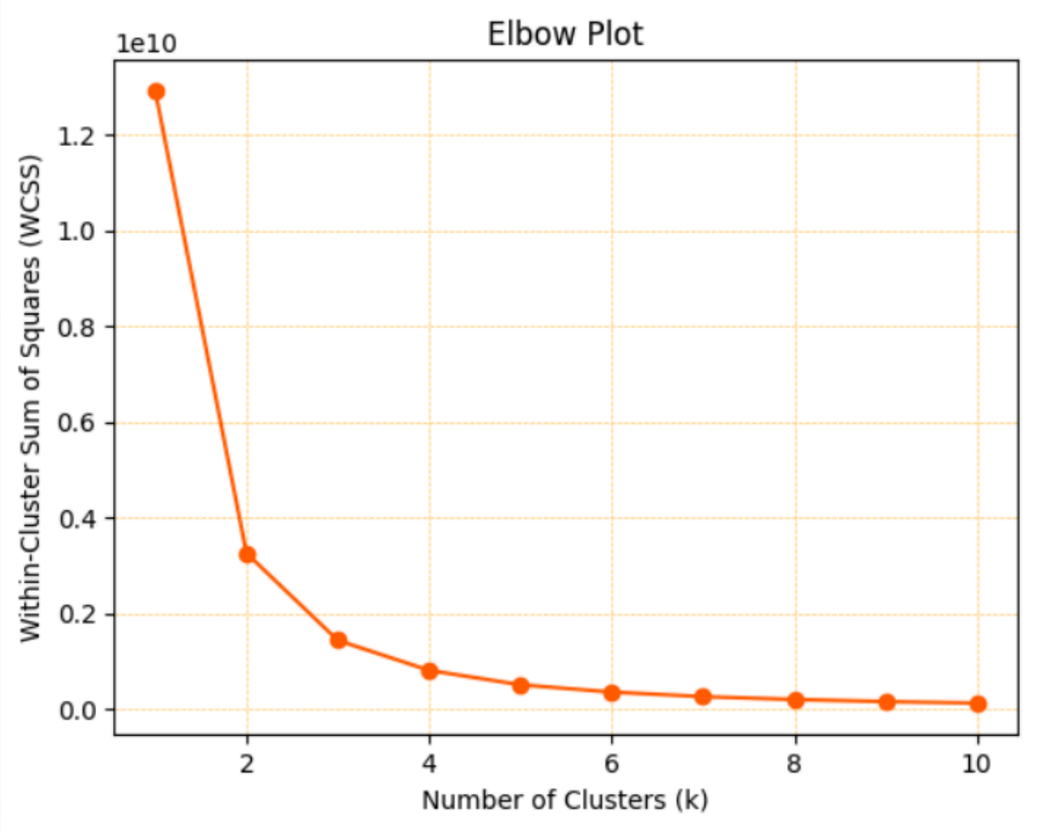
03

Each data point is assigned to the nearest centroid cluster

ELBOW METHOD

Process for defining the optimal number of clusters

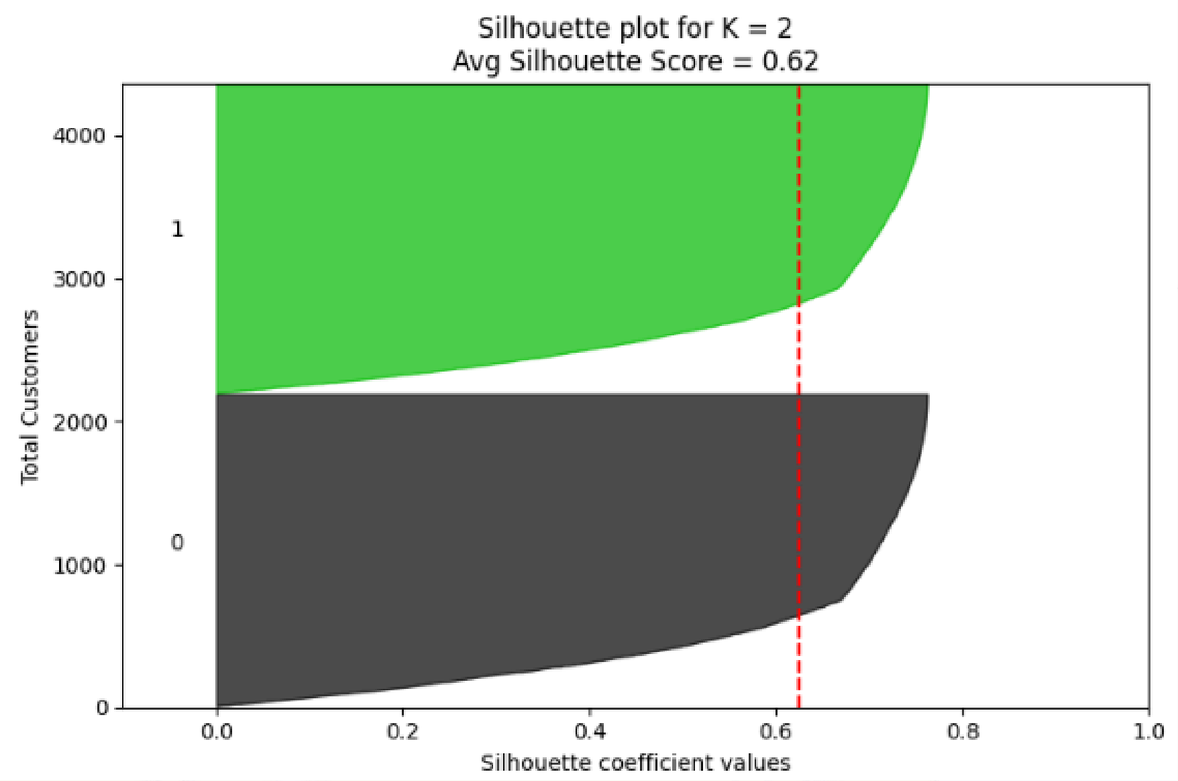
- **Name given by its shape**
- **2 is the suggested “optimal” k**



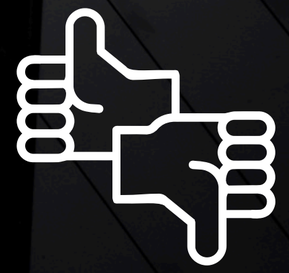
SILHOUETTE METHOD

Process for defining the optimal number of clusters

- **Scores range from -1 to 1**
  - **1 : data well clustered**
  - **-1 : data not distinctly separated from other clusters**
- **0,62 sil. score**
- **2 is the suggested “optimal” k**



RECOMMENDATION SYSTEM



INPUTS

- Top purchased products for each cluster
- All purchases done by each different customer
- All scaled data from the feature engineering process

OUTPUTS

- New database
- 3 product recommendations for each customer
- Alternative recommendations in case of already purchased products

THANK YOU ALL FOR YOUR ATTENTION