# Capstone Project - Online Retail Customer Segmentation

# By Noah Christy.

Table of Contents

1. Problem Statement
2. Project Objective
3. Data Description
4. Data Pre-processing Steps and Inspiration
5. Choosing the Algorithm for the Project
6. Motivation and Reasons For Choosing the Algorithm
7. Assumptions
8. Model Evaluation and Techniques
9. Inferences from the Same
10. Future Possibilities of the Project
11. Conclusion
12. References

# Problem Statement

An online retail store is trying to understand the various customer purchase patterns for their firm, you are required to give enough evidence based insights to provide the same.

# Project Objective

# Address missing values and outliers to ensure data quality.

* Identify distinct customer segments based on their purchase patterns.
* Utilize clustering techniques to group customers with similar purchasing behavior.
* Explore clusters to understand the characteristics of each segment.
* Provide insights for customer retention strategies.
* Evaluate the effectiveness of the clustering model.
* Offer actionable recommendations based on insights.
* Analyze the importance of different features (e.g., quantity, unit price) in defining customer segments

# Data Description

**Dataset Overview:**

Provide a brief overview of the online retail dataset, including the number of records and features.

**Features Included:**

List and briefly describe the main features included in the dataset. This may include columns like 'InvoiceNo,' 'StockCode,' 'Description,' 'Quantity,' 'UnitPrice,' 'CustomerID,' 'Country,' etc.

**Description of Key Features:**

Provide a more detailed description of critical features, especially those relevant to the clustering analysis (e.g., 'Quantity,' 'UnitPrice')

**Data Types:**

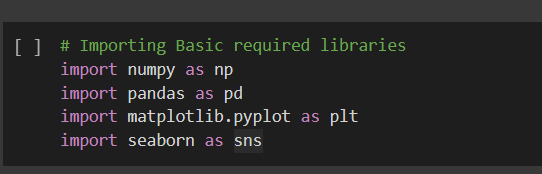
Mention the data types of each feature (e.g., numerical, categorical) to highlight the variety of information present in the dataset.

# Data Preprocessing Steps And Inspiration

The preprocessing of the data included the following steps:

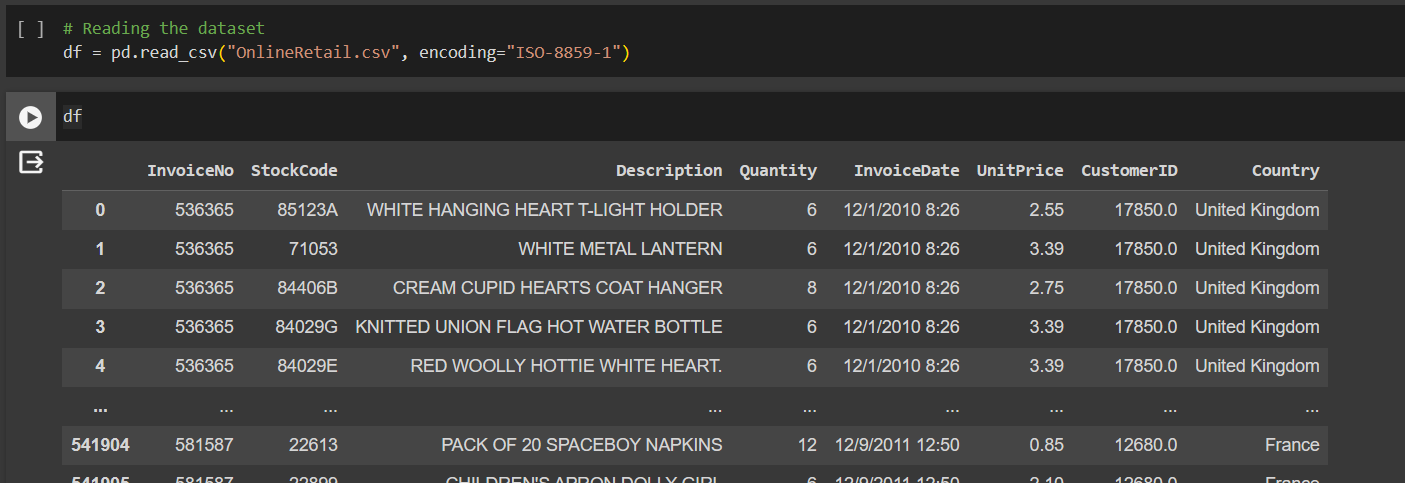
1. **Library Import:**

* Import the essential libraries for data analysis and visualization.
* The numpy library is imported as np.
* The pandas library is imported as pd.
* The matplotlib.pyplot library is imported as plt for plotting.



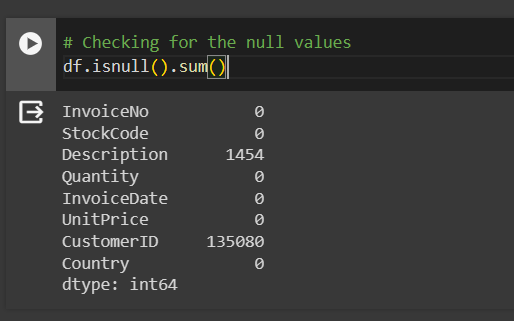
1. **Reading the Dataset:**

* The code reads a dataset from a CSV file named "OnlineRetail.csv" using the pd.read\_csv function from the Pandas library.
* The encoding="ISO-8859-1" parameter specifies the character encoding to be used while reading the file. This encoding is commonly used for datasets with international characters.

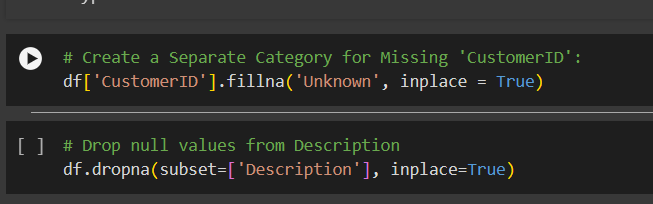


1. **Checking for Null Values:**

* The code snippet uses df.isnull().sum() to check for the number of null values in each column of the DataFrame.



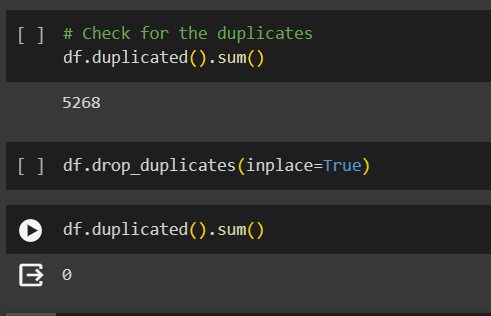
* The code snippet fills the missing values in the 'CustomerID' column with the string 'Unknown' using the fillna method.



* The code removes rows with null values specifically in the 'Description' column

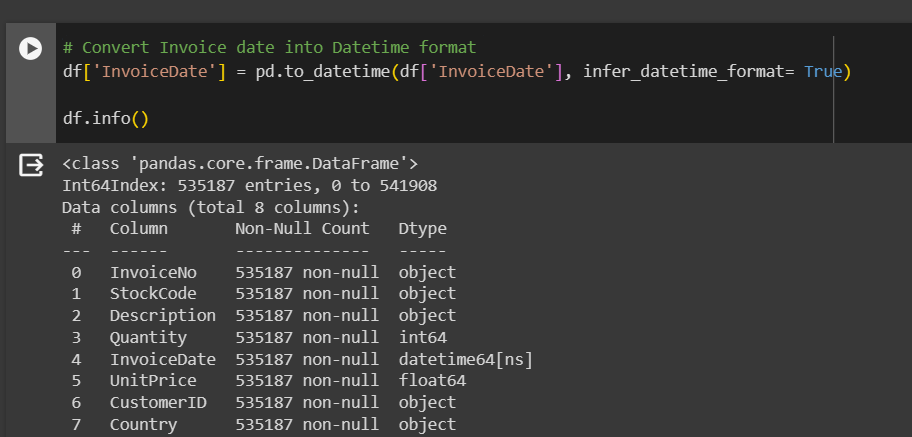
1. **Checking for Duplicates:**

* It provides a count of the duplicated rows, which is useful for identifying and addressing potential redundancy in the data.



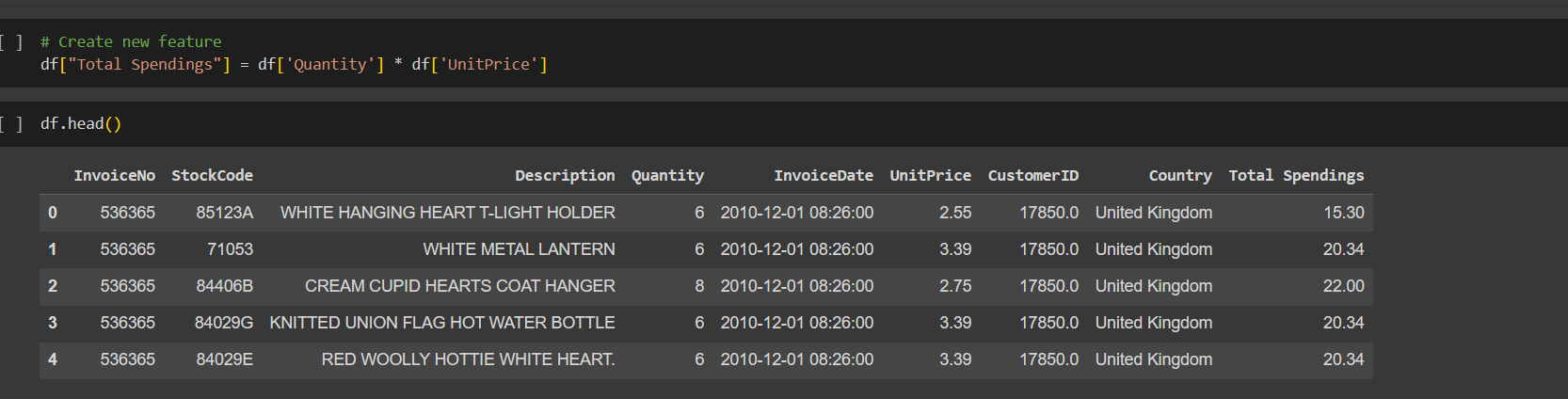
1. **Converting 'InvoiceDate' to Datetime Format:**

* The conversion to datetime format enables better handling and manipulation of date-related operations.
* The code snippet uses pd.to\_datetime() to convert the 'InvoiceDate' column to datetime format
* The infer\_datetime\_format=True parameter allows Pandas to automatically infer the datetime format, which can be more efficient for large datasets.



1. **Creating a New Feature:**

* The code snippet creates a new feature named "Total Spendings" in the DataFrame.
* The feature is derived from the product of the 'Quantity' and 'UnitPrice' columns.
* The new feature, "Total Spendings," represents the total amount spent on each transaction, considering the quantity of items purchased and their unit prices.



# Choosing the Algorithm For the Project

Choosing the K-means clustering algorithm for a project is often based on several factors, and here are some reasons why K-means might be a suitable choice for your project:

1. **Simplicity and Efficiency.**

* K-means is a simple and computationally efficient clustering algorithm, making it suitable for large datasets and projects with limited computational resources.

1. **Scalability.**

* K-means can handle a large number of data points and features, making it scalable for datasets with varying sizes and dimensions.

1. **Suitable for Numerical data.**

* K-means is well-suited for numerical data and continuous variables. It calculates cluster centroids based on the mean of data points in each cluster**.**

1. **Quantifiable Metrics (Inertia).**

* K-means optimization is driven by minimizing the sum of squared distances (inertia) between data points and their assigned cluster centroids. This provides a quantifiable metric to assess the quality of clustering.

1. **Interpretability.**

* The results of K-means are relatively easy to interpret. The algorithm partitions data into clusters, and each cluster is represented by a centroid, making it straightforward to understand the grouping.

# Assumptions

The following assumptions were made in order to create the model for XYZ project.

1. **Numeric Data.**

* K-means is designed for numerical data. It may not perform well with categorical or non-numeric data. Preprocessing steps, such as one-hot encoding or appropriate feature engineering, may be necessary for categorical features

1. **Optimal Number of Clusters (K).**

* The choice of the optimal number of clusters (K) is an important assumption. If an incorrect K value is chosen, the clustering results may not reflect the true underlying structure of the data.

# Model Evaluation and Technique

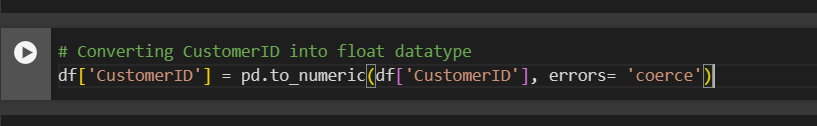
The following techniques and steps were involved in the evaluation of the model.

1. **Data Preparation for Clustering.**

* The code snippet indicates that data preparation steps are being performed in preparation for clustering analysis.
* Data preparation is crucial to ensure the dataset is in a suitable format for the clustering algorithm.

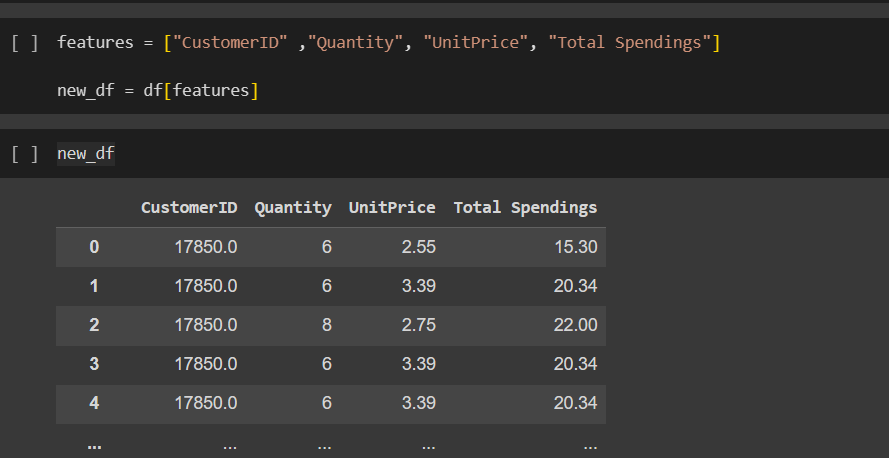
1. **Conversion of 'CustomerID' to Float Datatype.**

* The code uses pd.to\_numeric() to convert the 'CustomerID' column to a float datatype.
* The errors='coerce' parameter is used to handle any non-numeric values by converting them to NaN (Not a Number).



1. **Feature Selection.**

* The code snippet selects a subset of features from the original DataFrame for further analysis.



# **Standardizing the Data.**

# 

# The code snippet standardizes the selected features in the DataFrame new\_df using the StandardScaler from scikit-learn.

# Standardization transforms the data to have a mean of 0 and a standard deviation of 1, ensuring that all features are on a similar scale.

# An instance of the StandardScaler is created and assigned to the variable scaler.

# The fit\_transform method is used to scale the selected features in new\_df. The scaled data is stored in the variable df\_scaled

# The scaled data is then converted back to a DataFrame (df\_df1) with column names ["CustomerID", "Quantity", "UnitPrice", "Total Spendings"].

# **KMeans Clustering.**

# The code snippet uses the KMeans clustering algorithm from scikit-learn for clustering analysis.

# 

# The variable clusters is defined as a list ranging from 2 to 7, representing the potential number of clusters to explore.

# The Elbow Curve method is employed to find the optimal number of clusters for the dataset. This involves running KMeans for a range of cluster numbers and plotting the sum of squared distances (SSD) against the number of clusters.

# The ssd list is initialized to store the sum of squared distances for each corresponding number of clusters.

# A loop iterates over each number of clusters in the specified range.

# For each iteration, a KMeans model is created and fitted to the standardized data (df\_df1). The SSD for each clustering configuration is calculated and appended to the ssd list.

# 

# **Elbow Curve Visualization.**

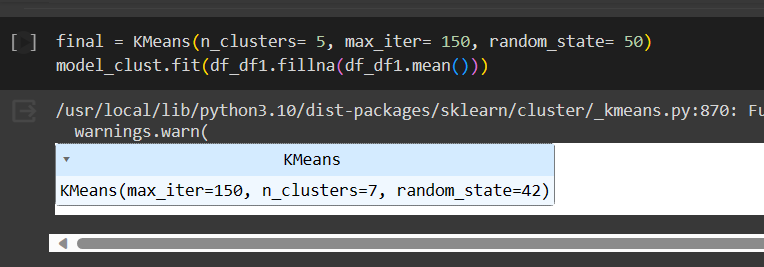
# The code snippet uses plt.plot(clusters, ssd) to create a line plot to visualize the Elbow Curve.

# The x-axis represents the number of clusters, and the y-axis represents the corresponding sum of squared distances (SSD).

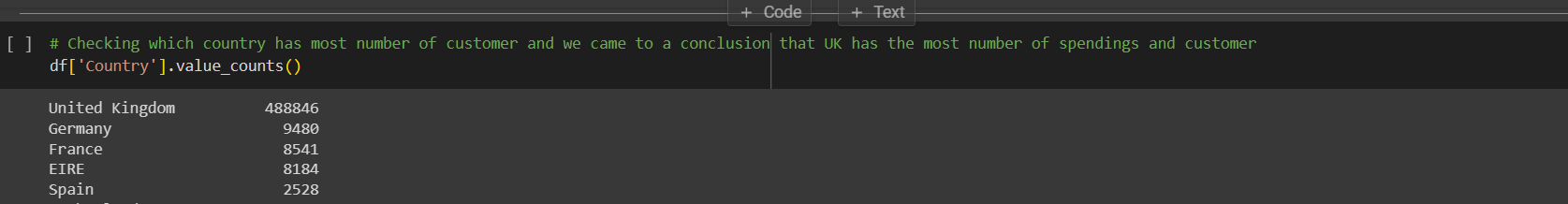
# 

# Inferences from the Project

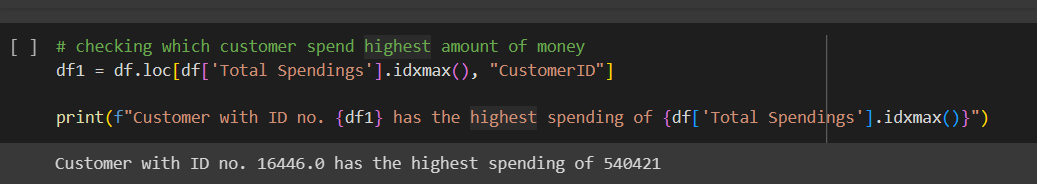
* From the Elbow curve we choose the ideal cluster number to be 5.



* Concluding that the UK has the Most Customers.



* Identifying the Customer with the Highest Spending.



# Future Possibilities

1. **Customer Segmentation Refinement.**

* Conduct a more detailed customer segmentation analysis based on the clustering results. Explore additional features or alternative clustering algorithms to create more refined and meaningful segments.

1. **Predictive Modeling.**

* Utilize clustering results as features for predictive modeling tasks, such as predicting customer churn, lifetime value, or future purchase behavior. Machine learning models can be trained using the clustered data.

1. **Product Recommendation Systems.**

* Develop personalized product recommendation systems based on the clustering results. Recommend products that are popular within specific clusters, enhancing the customer shopping experience.

1. **Interactive Visualization.**

* Create interactive visualizations to allow stakeholders to explore and interpret the clustering results more effectively. Use tools like dashboards to provide a user-friendly interface.

1. **Integration with E-commerce Platforms.**

* If applicable, integrate clustering insights directly into the e-commerce platform to automate personalized recommendations and promotions for users in real-time.

# Conclusion

In conclusion, the K-means clustering project on the online retail dataset has yielded valuable insights into customer behavior and spending patterns. Through segmentation, customers were categorized into distinct clusters, enabling targeted marketing strategies and personalized interactions. The analysis highlighted the United Kingdom as the country with the highest customer count and spendings, providing geographical insights for business decisions. Additionally, the project successfully identified the customer with the highest spending, presenting opportunities for tailored engagement and loyalty programs. Feature engineering, including the creation of "Total Spendings," contributed to a more comprehensive analysis. Data preprocessing steps, such as handling missing values and standardization, were crucial in ensuring the dataset's readiness for clustering analysis. Future possibilities include refining customer segmentation, dynamic clustering, and integrating external data for a more nuanced understanding of customer behavior. Overall, the project provides a foundation for ongoing strategic decision-making and further exploration in the online retail domain.

# References

# Kaggle Website

# Analytics Vidhya website

# Intellipaat resources

# Chat gpt