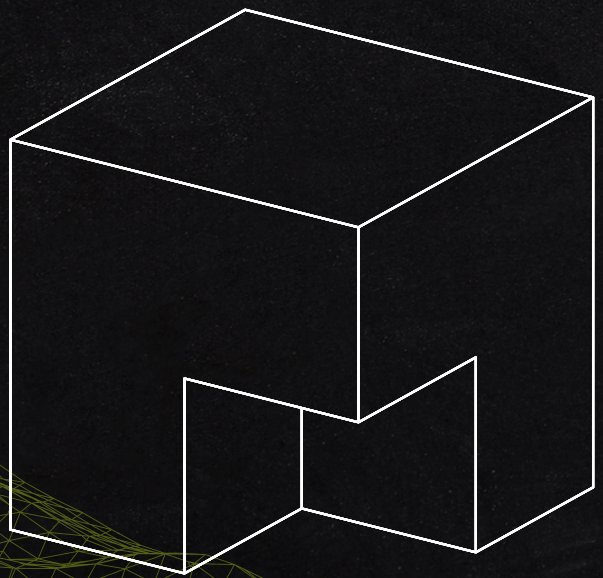
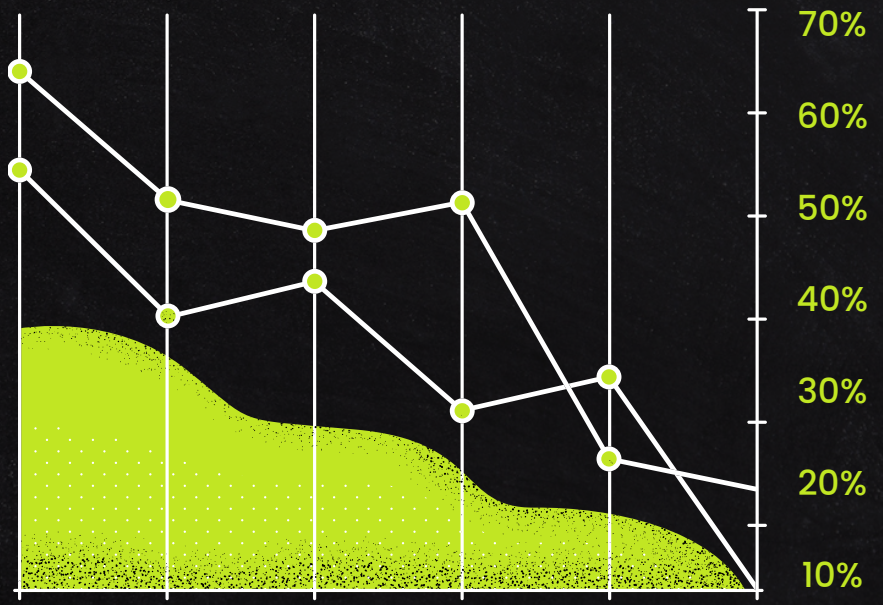
DATA ANALYSIS REPORT

Submitted by:
Ubaid Khan



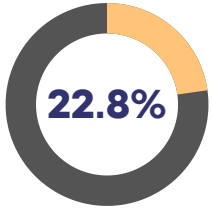


An assignment report for the recruiting team's better understanding of my work!

Kirana Club

Here, the Exploratory Data Analysis (EDA) involves using statistics and visualizations to analyze and identify trends in data sets. The primary intent of EDA is to determine whether a predictive model is a feasible analytical tool for business challenges or not.

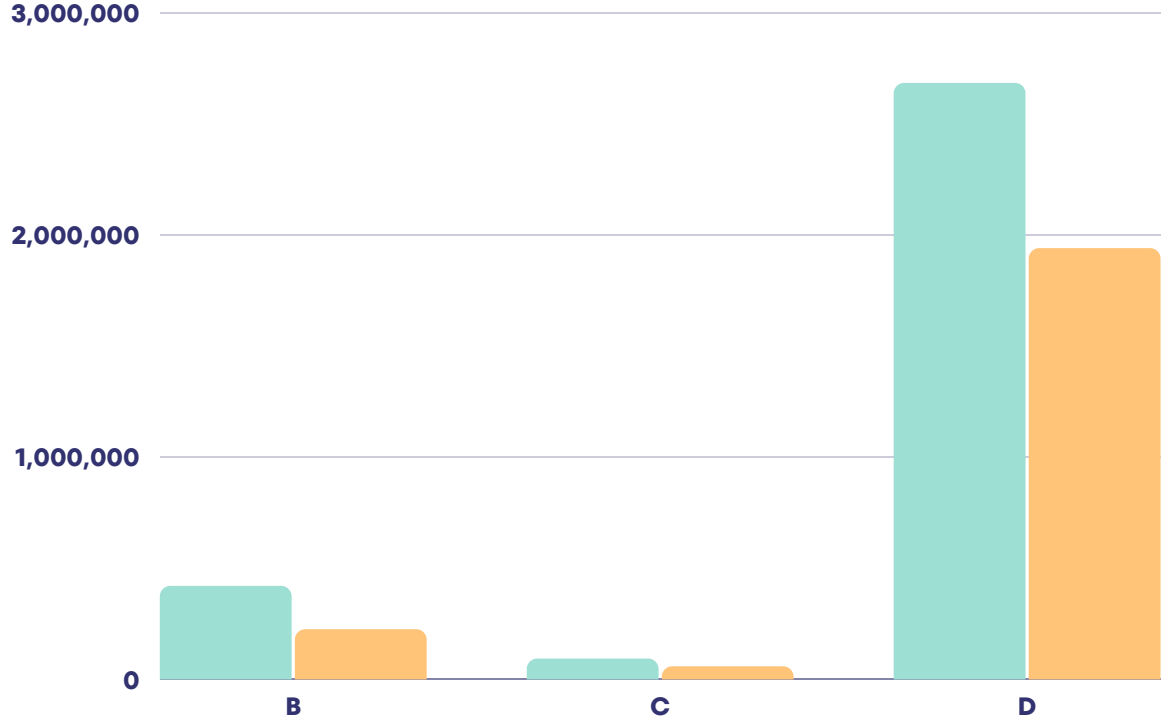
2009-2010 Dataset



22.8% missing values in the buyer or customer ID feature column.

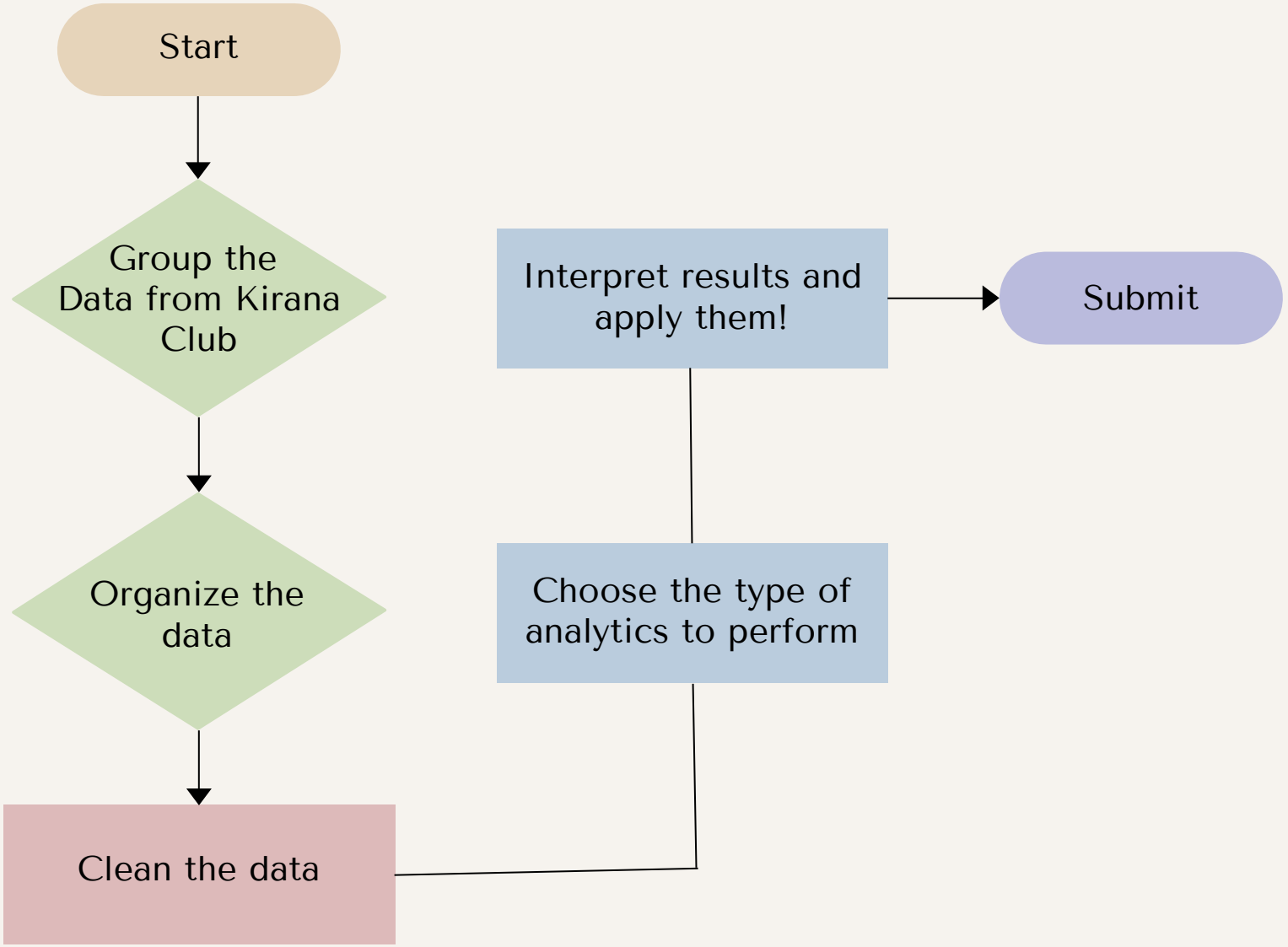
0.41% missing values in product Description feature column.

All missing values were dealt with using proper tools, techniques, and practices.

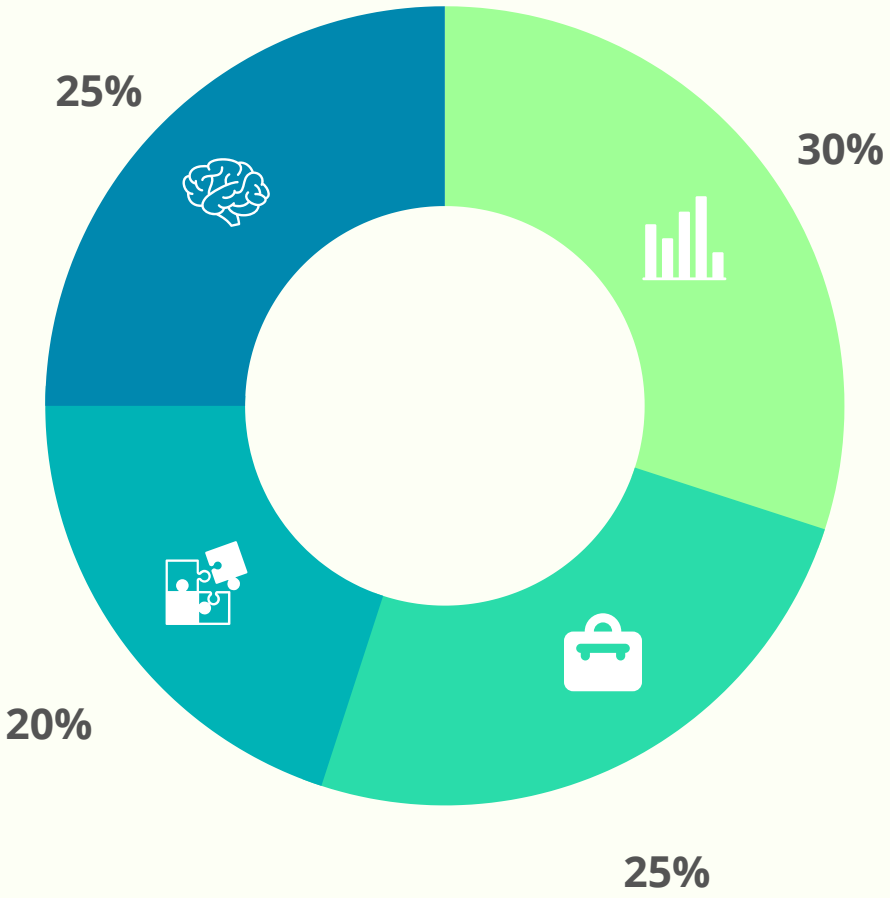


Statistical numbers from key features

PROCESS FLOWCHART



Methods & Technologies Used



Statistics

For understanding data, driving insights from it, and answering business questions.

Python Transformations

For data transformations, cleaning, organinzing, and more!

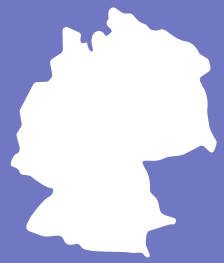
DSA

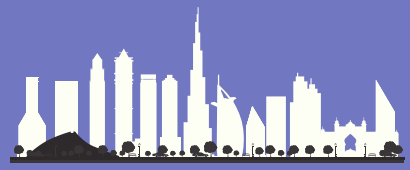
For the section-2 of the assignment.. Where fundamental knowledge of DSA was required.

Machine Learning

For the purpose of validating the conclusions made in the analysis.

CUSTOMER SEGMENTATION





CATEGORY A

1. Region Covered: UK (0.741 million buyers alone)
2. Cheap Products
3. Lesser Quantities

CATEGORY B

1. Region Covered: German buyers (17,624)
2. Cheaper Products bought
3. High product return rate, lesser loyalty

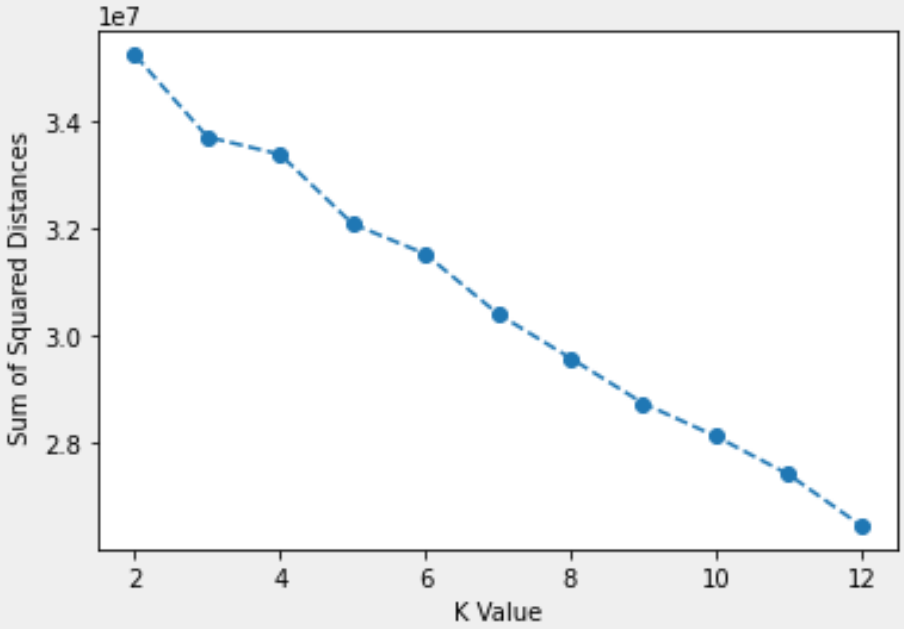
CATEGORY C

1. Region Covered: United Arab Emirates
2. Least Percent Returns
3. Intermediate Quantities
4. Cheaper Products

CATEGORY D

1. Region Covered: Almost entire Europe (65,067 buyers)
2. Most Loyal Customers with 18.95% orders from the same buyers
3. High Quantity Orders

ALGORITHM USED FOR CLUSTERING



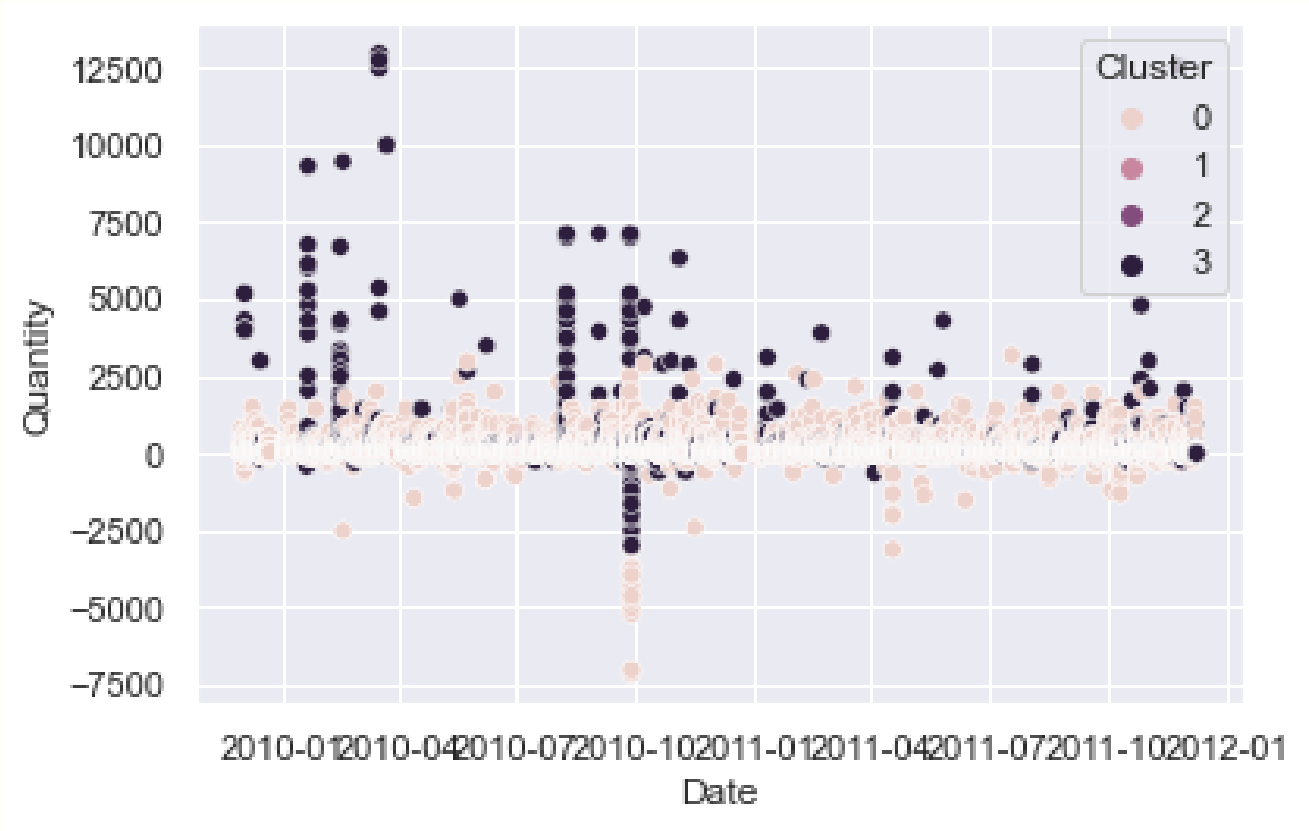
What is K-means?

k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

Seasonality found in data

Seasonality is a characteristic of a time series in which the data experiences regular and predictable changes that recur every calendar year.

This graph indicates seasonality around Cluster-3 or Category-D of our buyers, i.e. Europeans. Number of orders hiked in the months of January, April, July, & October in the year 2010.



Conclusions

Here, we can see the difference between the 'mean' of Price per item in a purchase between different categories of A, B, C, D. Stating that even though 'A' has 0.7+ million orders, their average price per order is still low. Inversely similar conclusion can be made for D, who has 10 times lesser orders than A but double the average price per order.