Unsupervised - Machine Learning Project

By Suresh babu Karunamoorthi

Process of Project

EDA – Exploratory Data Analysis

K-mean Cluster

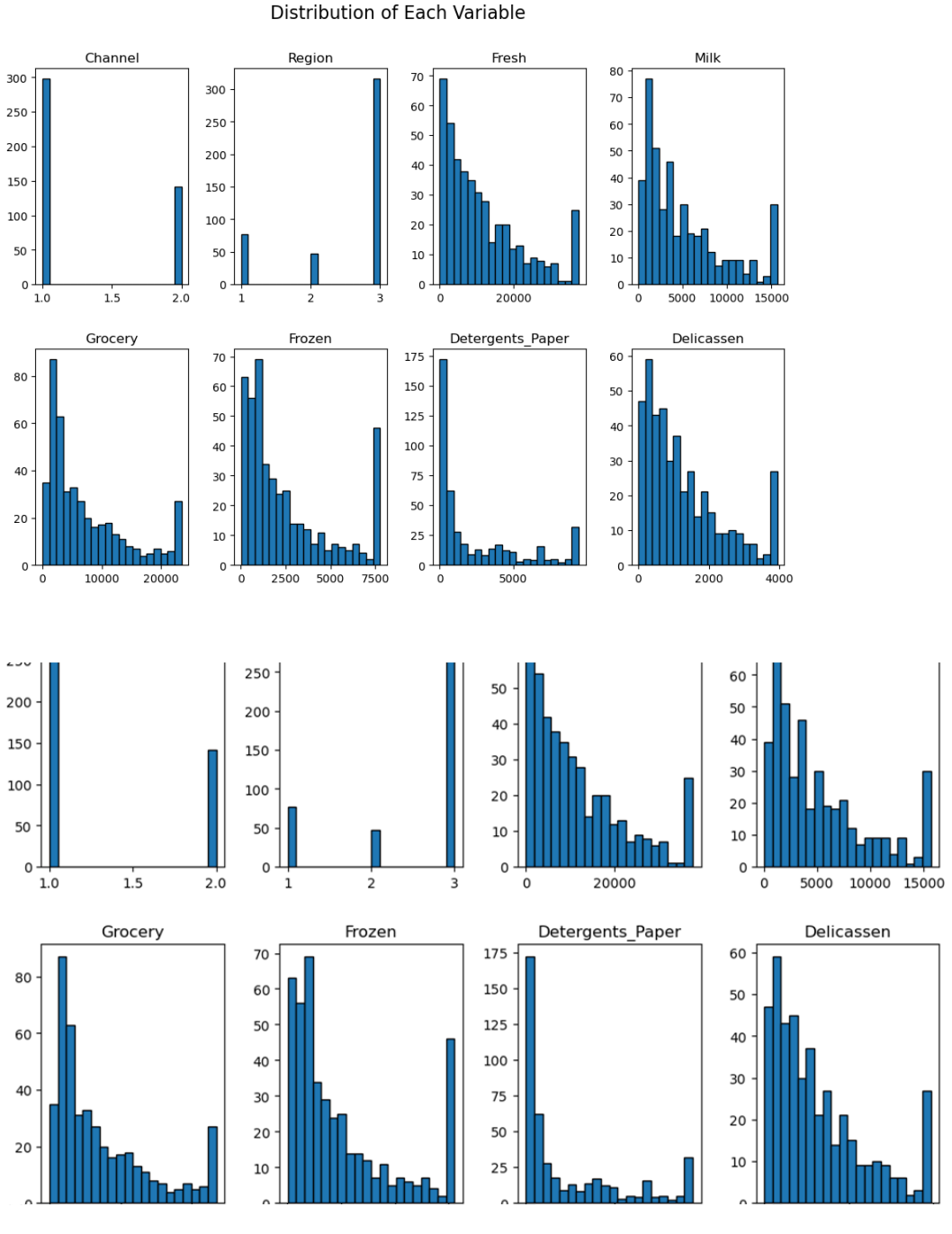
Hierarchical Cluster

PCA

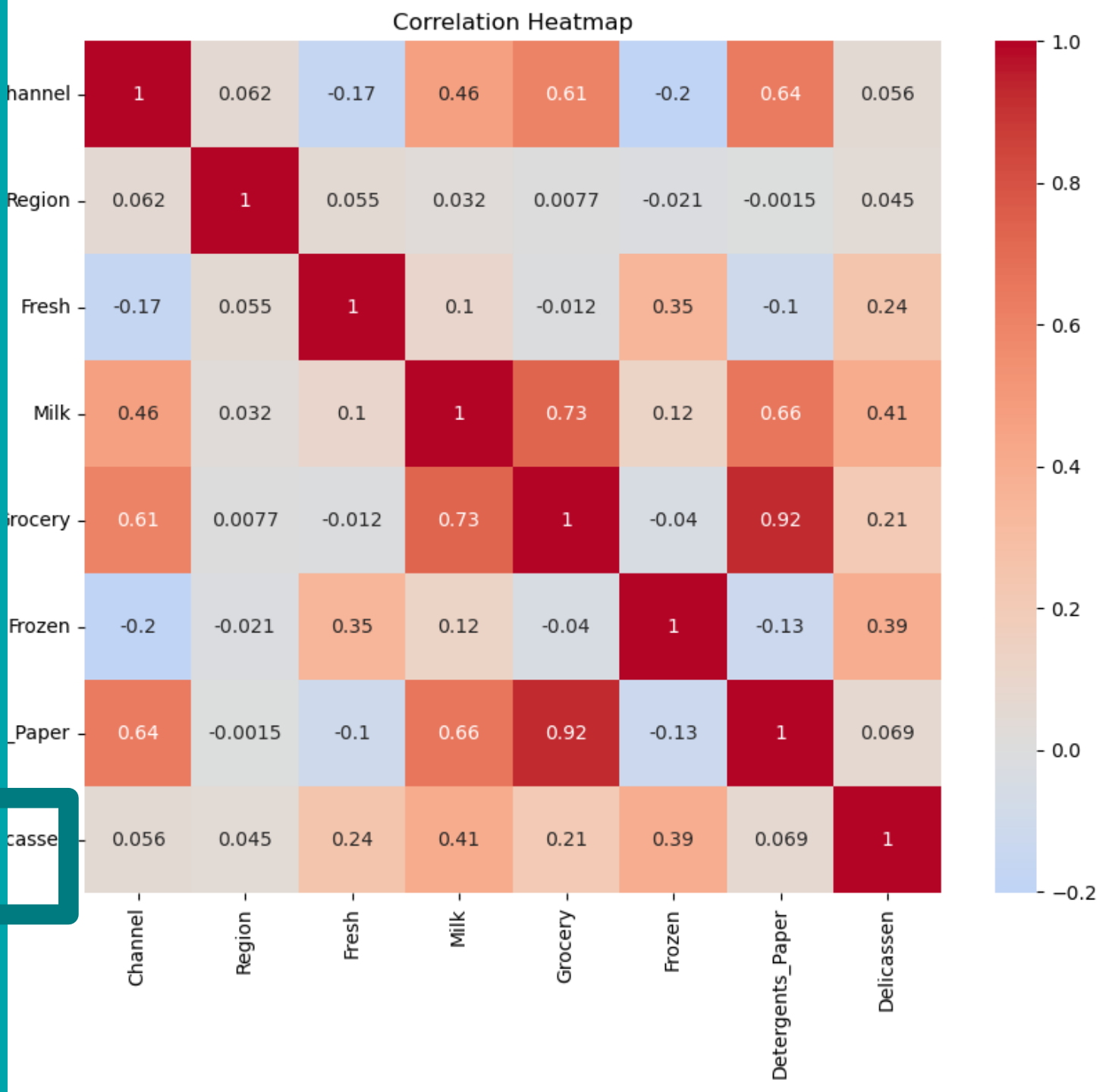
Conclusion

Exploratory Data Analysis

Distribution Of Each Variable

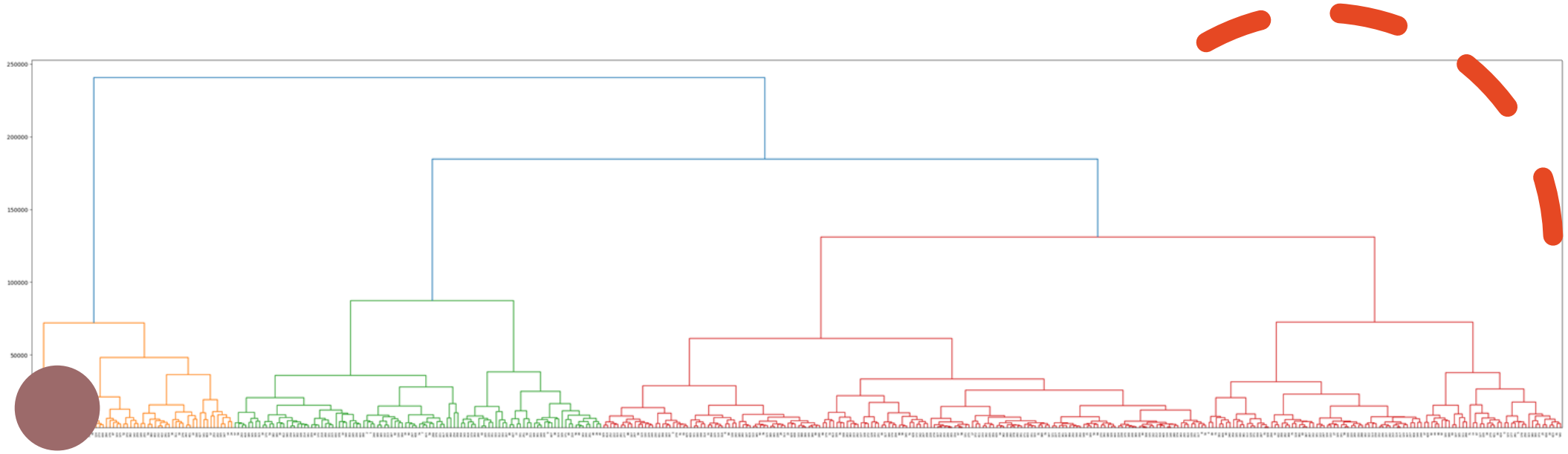


Correlation Heatmap



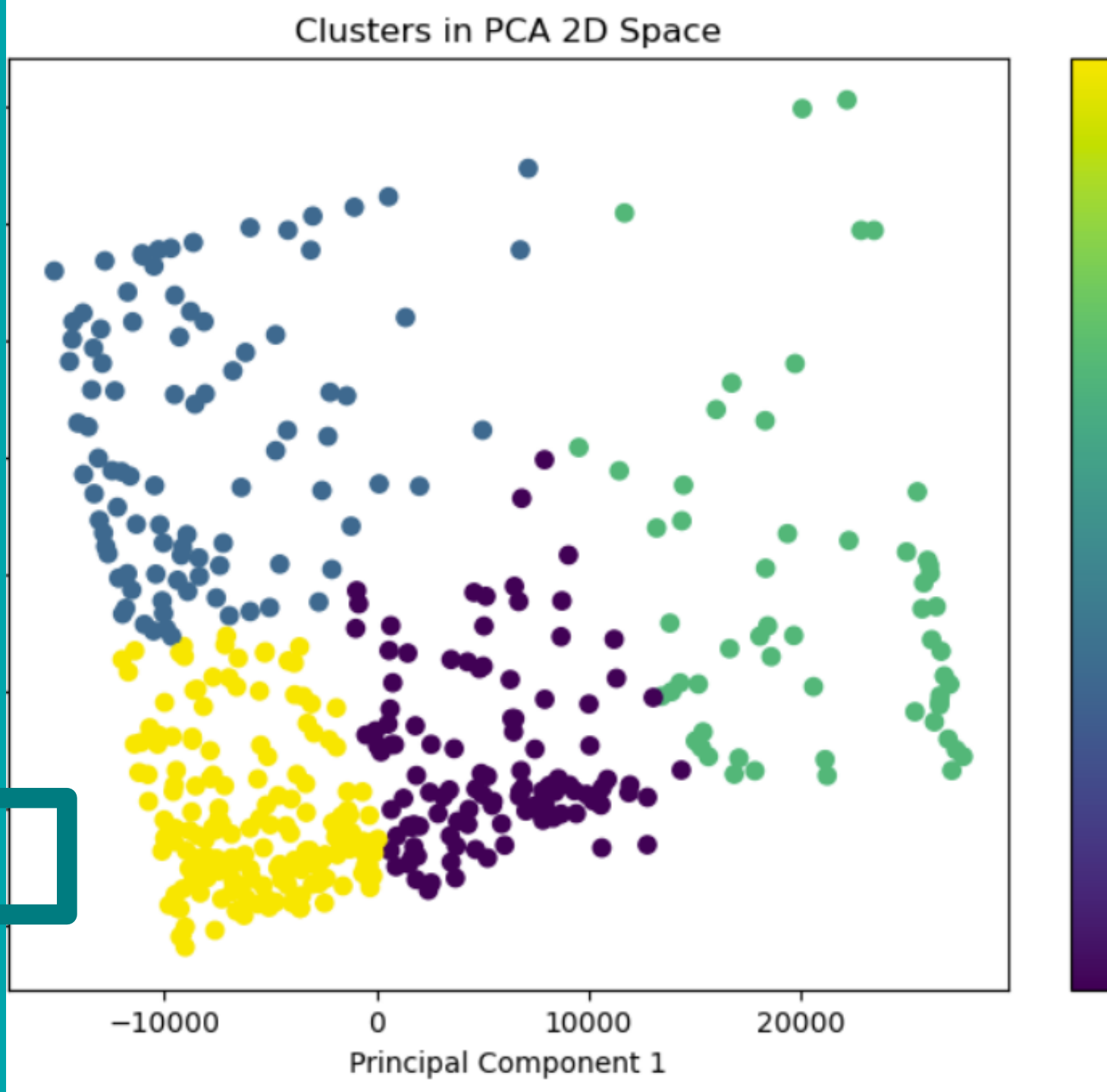
K Means Clustering

- grouped similar products together into clusters based on their attributes such as fresh, milk, grocery, frozen, detergents_paper, and delicatessen
- Used Elbow Method to determine K value
- Used k mean clustering
- As K-Means is a distance-based algorithm, the scale of the variables matters. Any variables on a large scale will have a much larger effect on the distance between the observations. So, scaling your data to have the same range is generally a good idea



Hierarchical Clustering

PCA


Clusters in PCA 2D Space
−10000
0
10000
20000
Principal Component 1

Conclusion

Cluster Analysis: The K-means clustering algorithm identified four distinct clusters within the dataset. Each cluster represents a different group of customers with distinct purchasing behaviors.

Outlier Detection: Through the exploratory data analysis, we identified and handled outliers for each feature separately. The outlier handling process involved capping the extreme values, which helped in stabilizing the clusters and improving the quality of the clustering results.

Data Dimensionality Reduction: The Principal Component Analysis (PCA) was applied to reduce the dimensionality of the dataset while preserving the most important information. The first two principal components explained around 93% of the total variance in the data.

Customer Segmentation: The clustering analysis provided valuable insights into customer segmentation, allowing the wholesale distributor to understand and target different customer groups effectively.

Thank you