Unsupervised Learning and Clustering

Introduction to Unsupervised Learning

Unsupervised learning is a type of machine learning that allows you to identify patterns in data using clustering methods like K-means and Hierarchical clustering. These methods help to gain insights and make data-driven decisions.

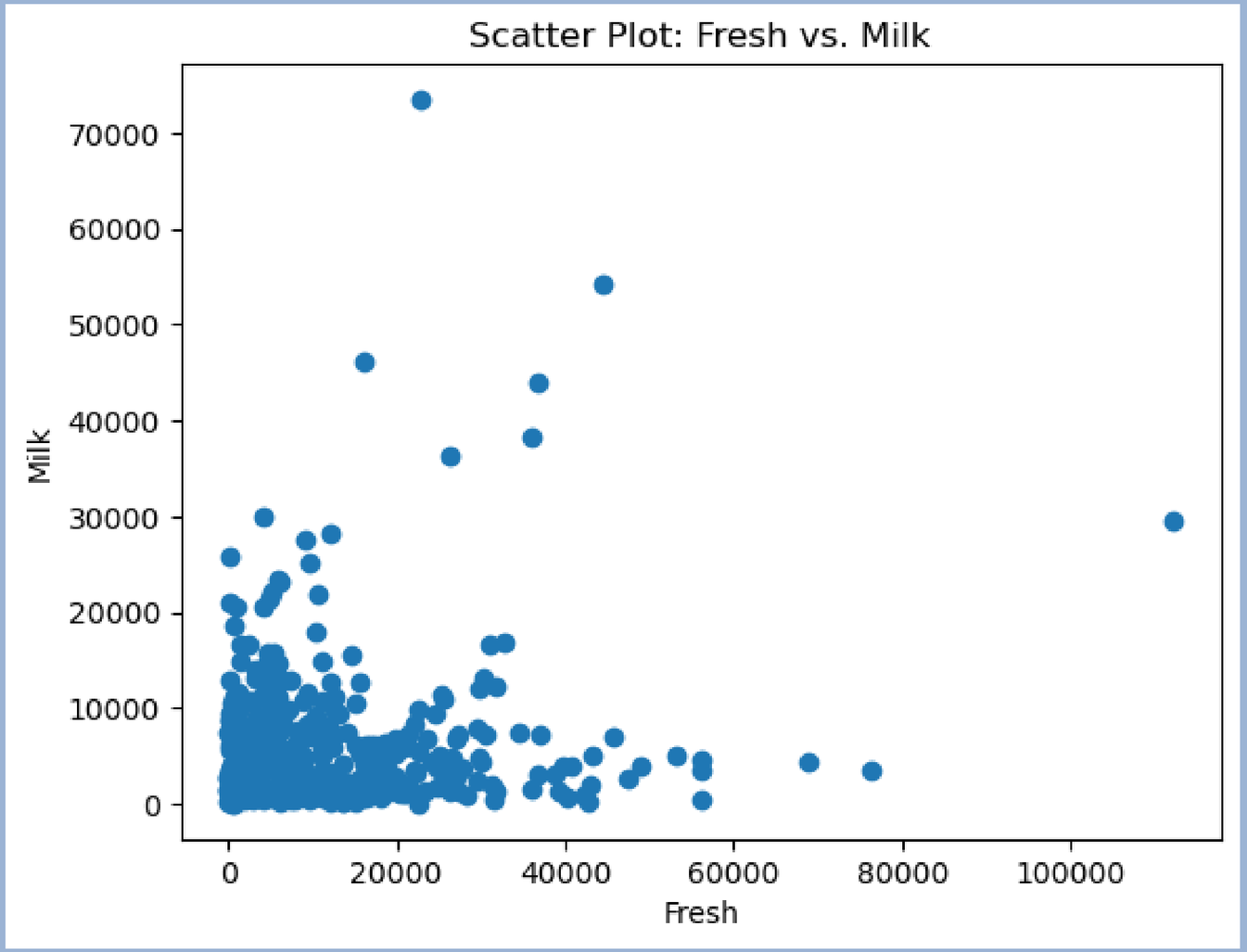
Cleaning and Preprocessing Data

Before analysis, clean data by handling missing or duplicate values, formatting inconsistent data, and removing outliers. Ensure accuracy and completeness.

Chose not to remove outliers in Fresh, Milk, Grocery, Frozen, Detergents_Paper,Delicatessen columns because were not too extreme to be possible and wanted to maintain dataset's integrity .

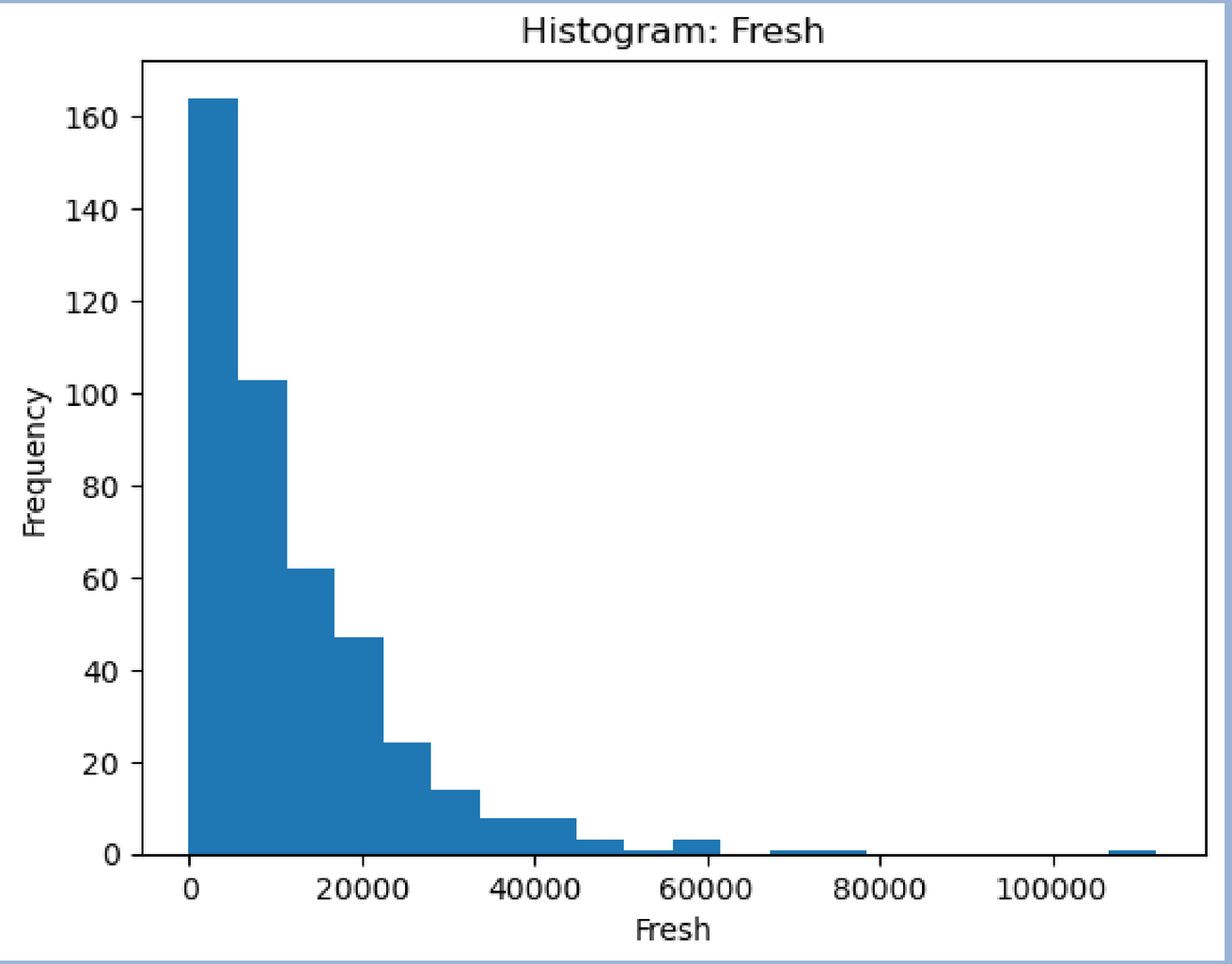
Exploratory Data Analysis (EDA)

EDA is the data exploration process to understand the data's nature, uncover underlying patterns, spot anomalies, and test statistical assumptions. It involves cleaning, analyzing, and visualizing data .



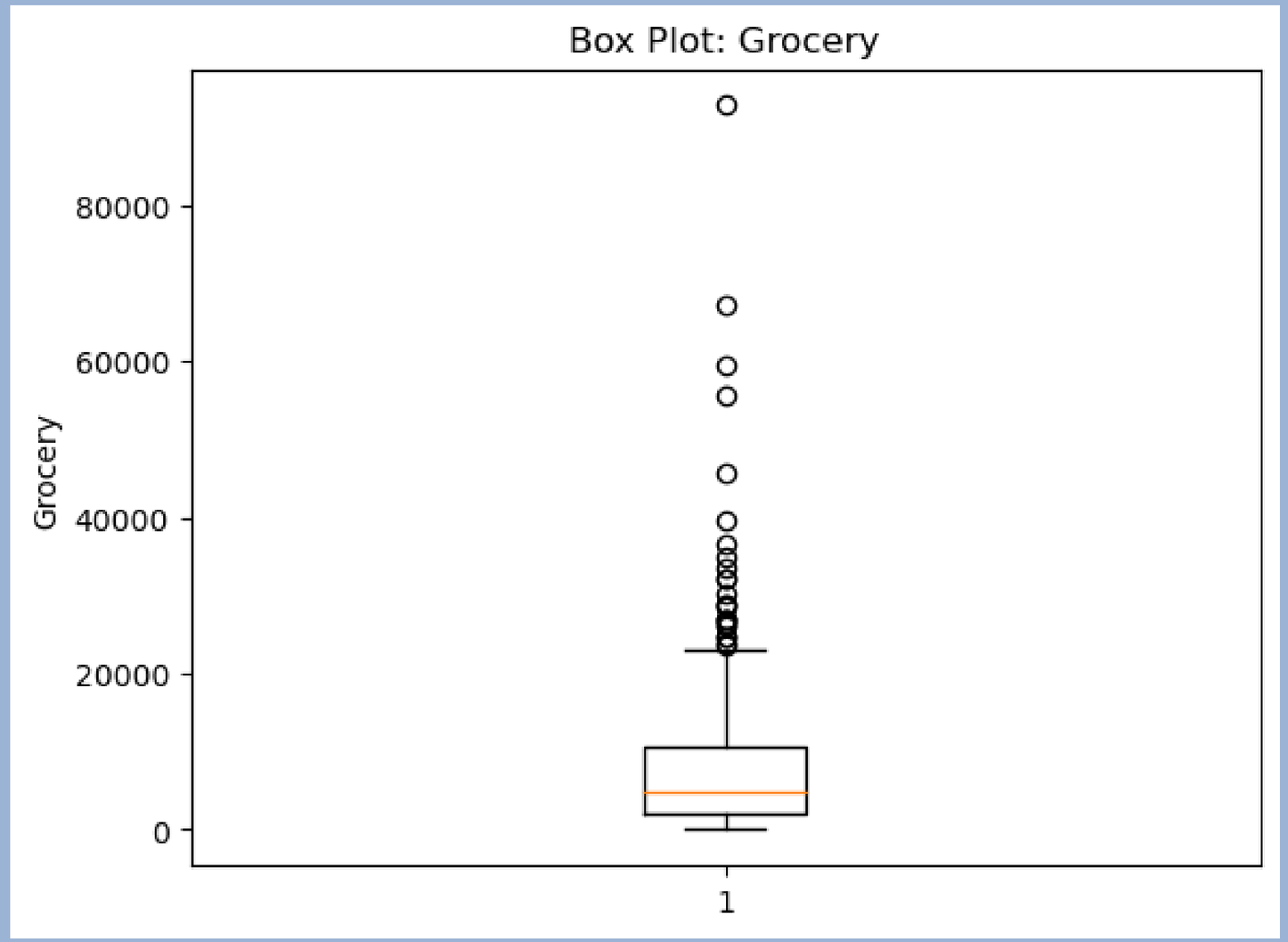
Exploratory Data Analysis (EDA)

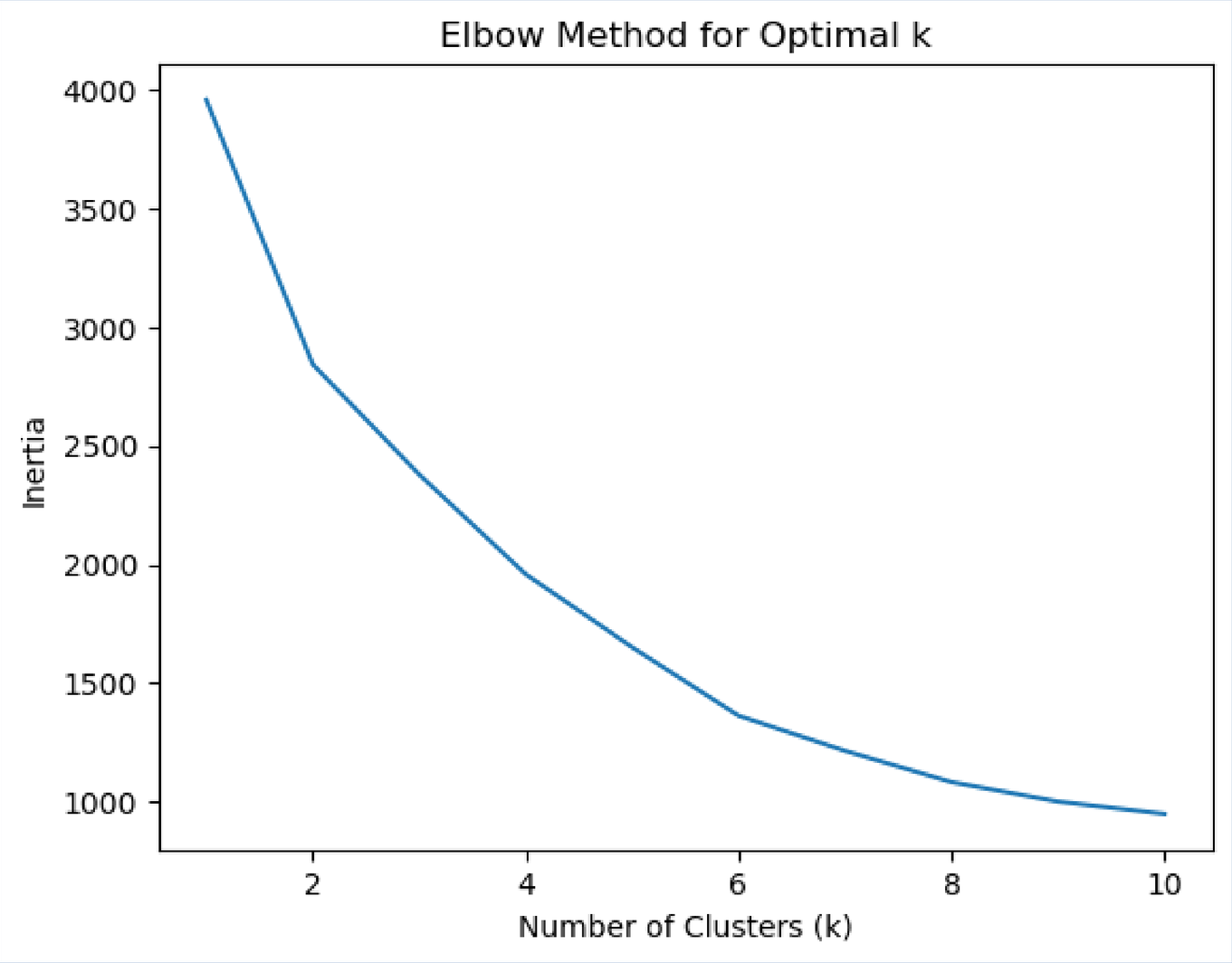
Count of purchases at different pricepoints from the fresh category



Exploratory Data Analysis (EDA)

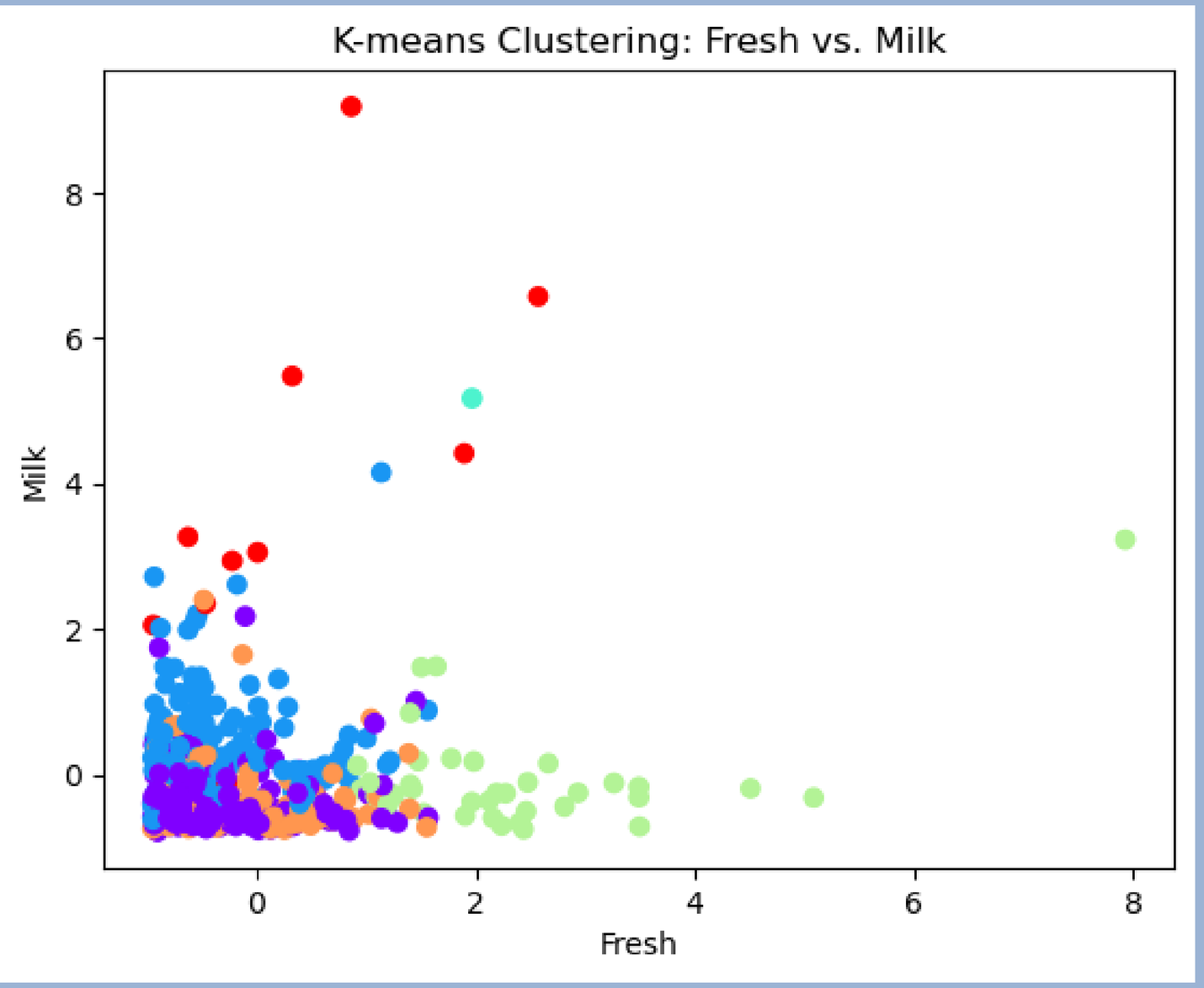
Checking for outliers in the grocery category did not remove because possible for one customer account ie.family or institution to spend these amounts on groceries





K-means Clustering

K-means clustering is a machine learning technique that groups data points into clusters based on similarity. Elbow method shows the most extreme drop in inertia when k=6

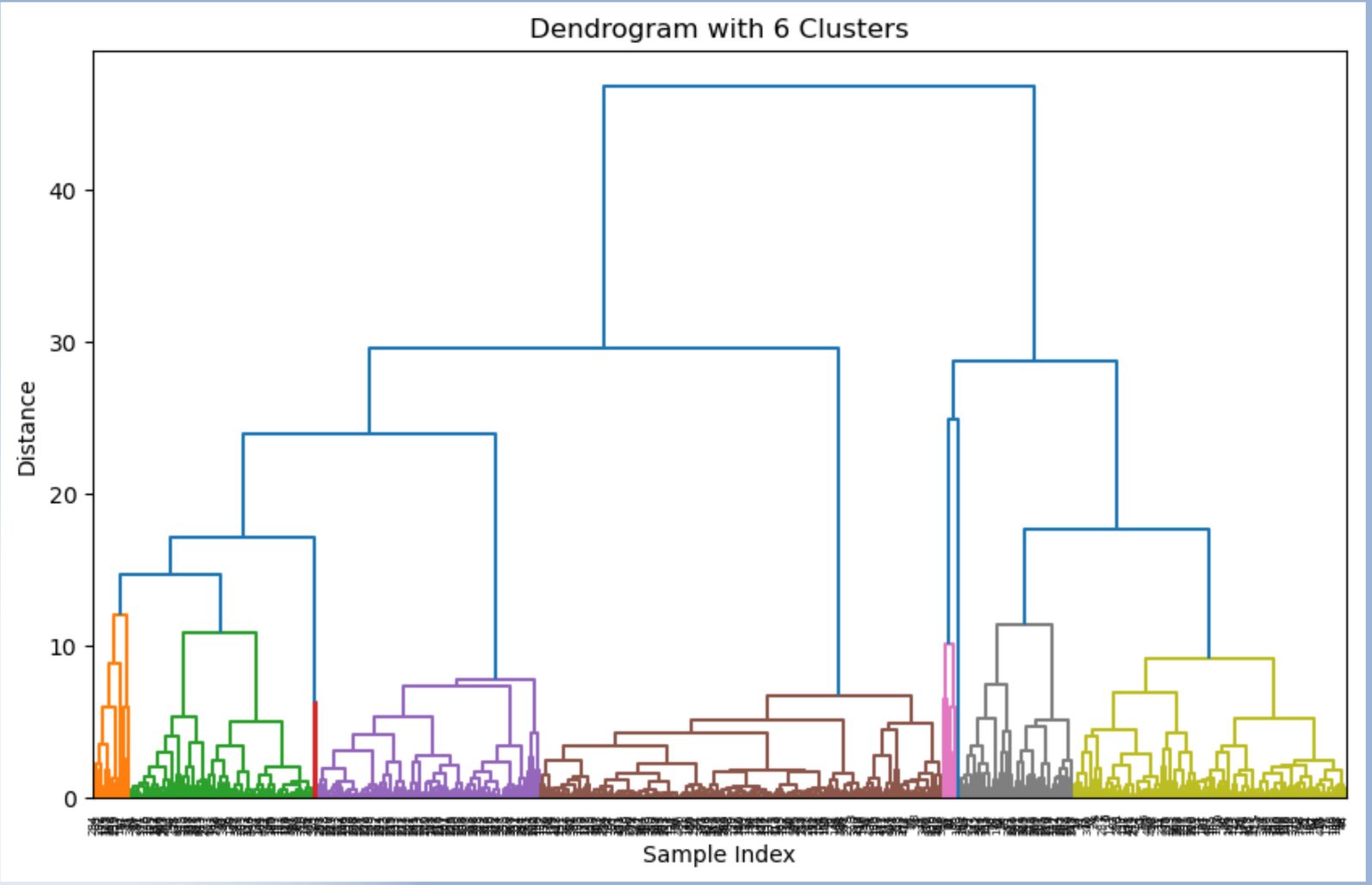


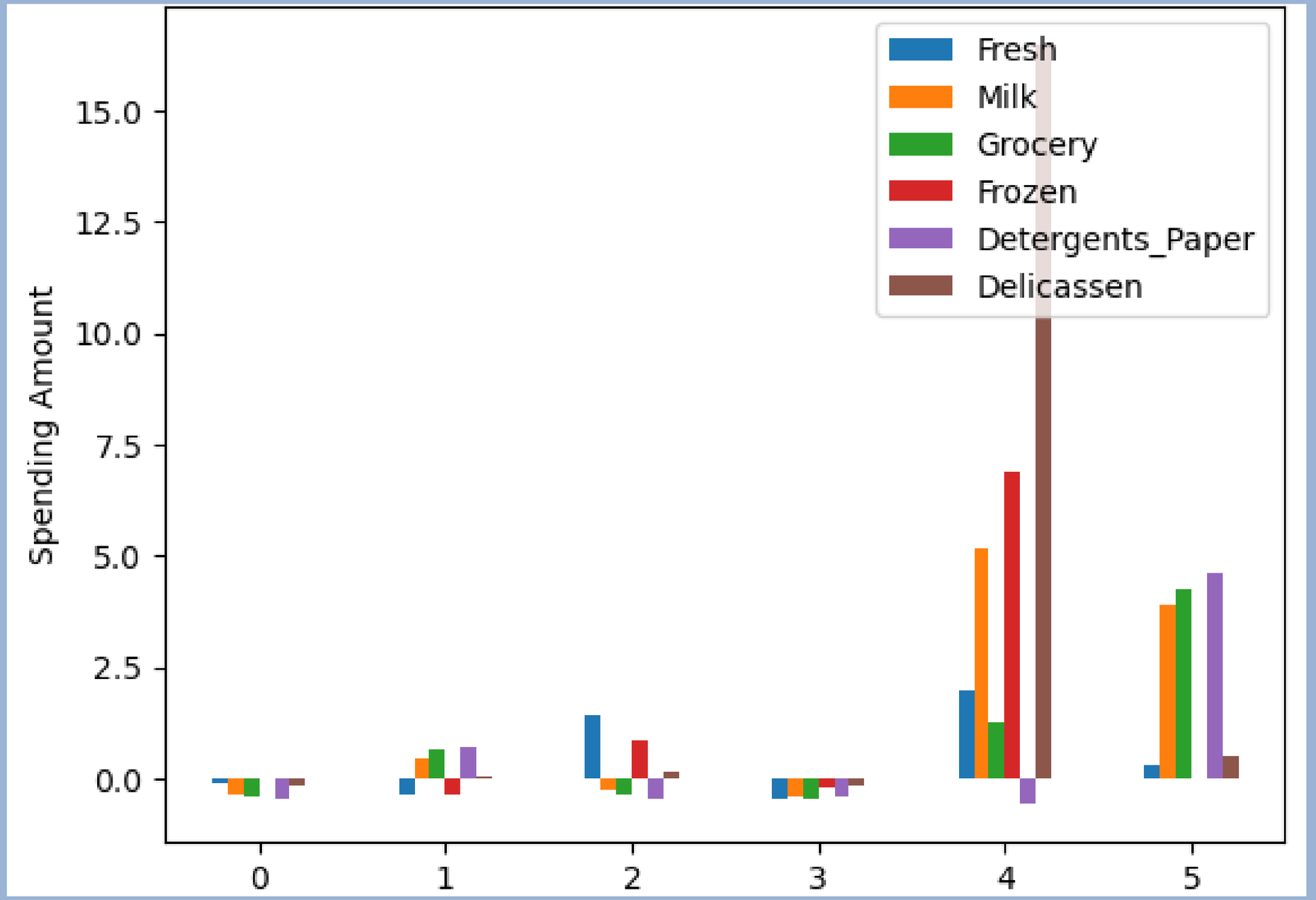
K-means Clustering

Scatter plot of Milk vs Fresh

Hierarchical Clustering

Hierarchical clustering is a type of unsupervised learning that helps to discover patterns, groups, and structures in unlabeled data. It produces a cluster tree diagram that shows the relationships between data points.




Fresh
Milk
Grocery
Frozen
Detergents_Paper
Delicassen
Spending Amount
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0
1
2
3
4
5

Thank you