Unsupervised Learning Project

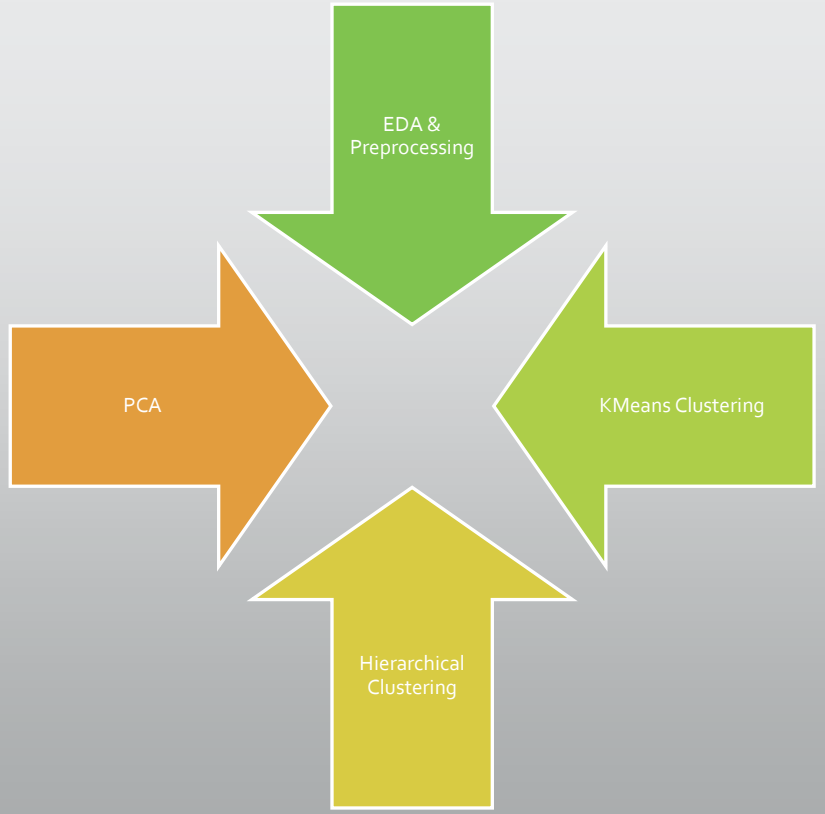
by
Cynthia Okaja
29th August 2023

LIGHTHOUSE LABS

Project Goals

- To perform a full unsupervised learning machine learning project on a "Wholesale Data" dataset

Project Execution



EDA & Preprocessing

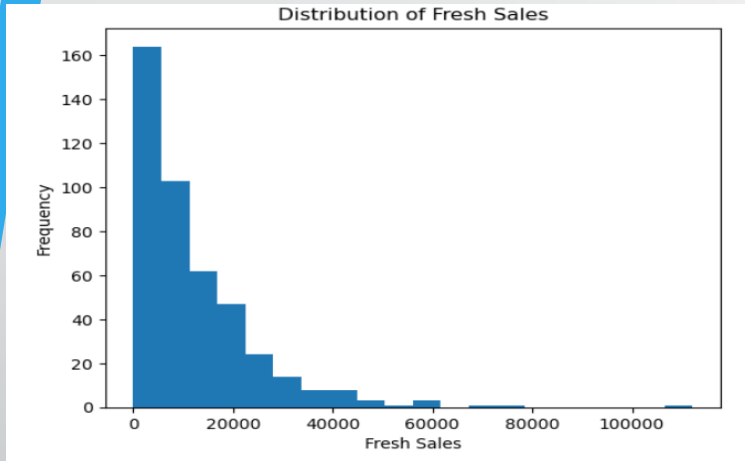
- Getting familiar with the data

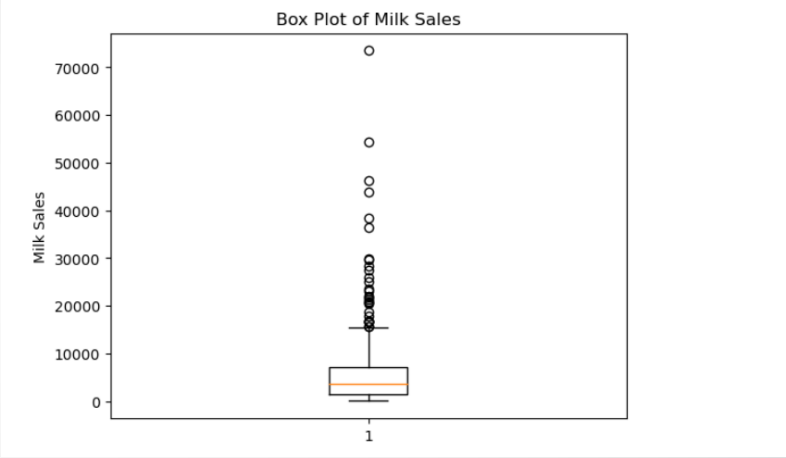
	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
count	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000
mean	1.322727	2.543182	12000.297727	5796.265909	7951.277273	3071.931818	2881.493182	1524.870455
std	0.468052	0.774272	12647.328865	7380.377175	9503.162829	4854.673333	4767.854448	2820.105937
min	1.000000	1.000000	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000
25%	1.000000	2.000000	3127.750000	1533.000000	2153.000000	742.250000	256.750000	408.250000
50%	1.000000	3.000000	8504.000000	3627.000000	4755.500000	1526.000000	816.500000	965.500000
75%	2.000000	3.000000	16933.750000	7190.250000	10655.750000	3554.250000	3922.000000	1820.250000
max	2.000000	3.000000	112151.000000	73498.000000	92780.000000	60869.000000	40827.000000	47943.000000

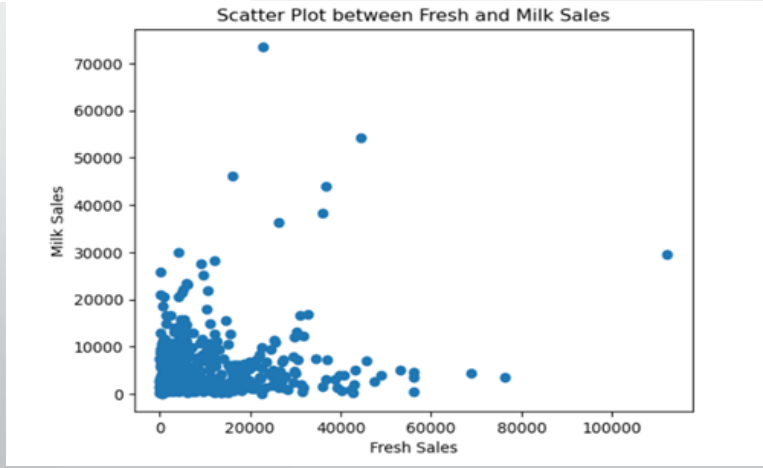
- Info

```
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Channel           440 non-null    int64
 1   Region            440 non-null    int64
 2   Fresh             440 non-null    int64
 3   Milk              440 non-null    int64
 4   Grocery           440 non-null    int64
 5   Frozen            440 non-null    int64
 6   Detergents_Paper  440 non-null    int64
 7   Delicassen        440 non-null    int64
dtypes: int64(8)
```

EDA

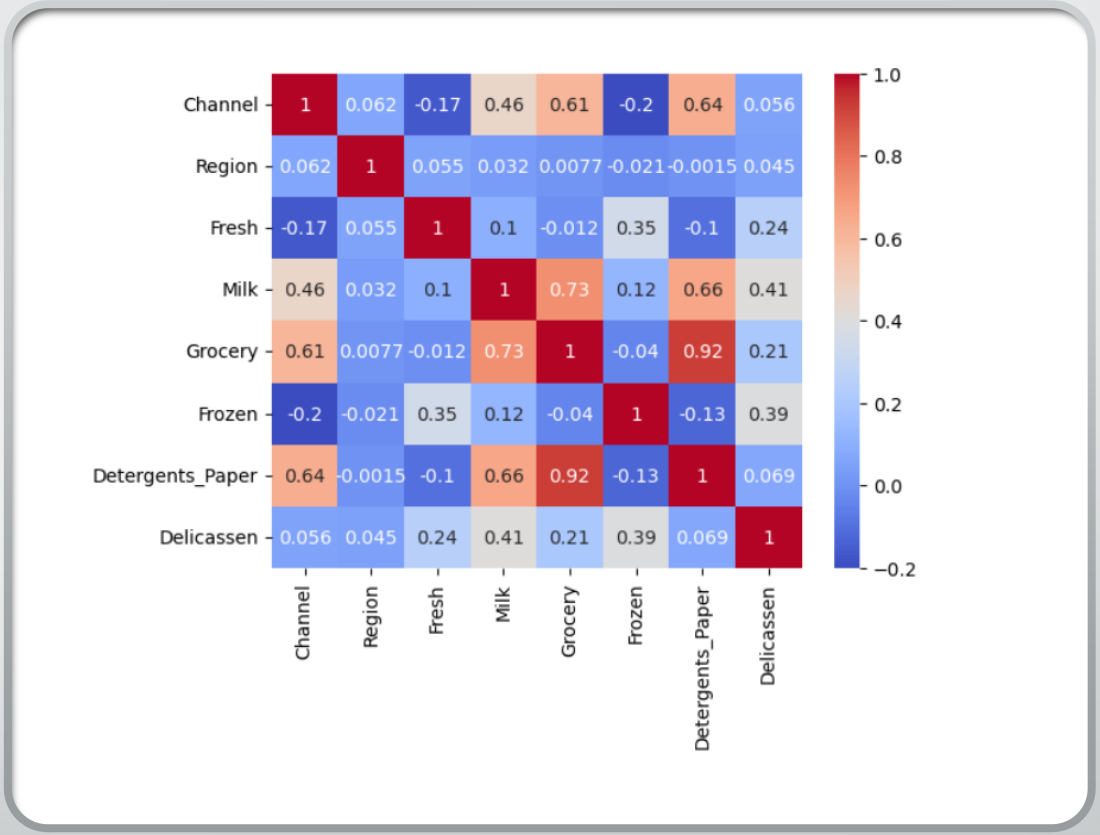

Distribution of Fresh Sales
Frequency
Fresh Sales
0
20000
40000
60000
80000
100000
0
20
40
60
80
100
120
140
160


Box Plot of Milk Sales
Milk Sales
1
0
10000
20000
30000
40000
50000
60000
70000


Scatter Plot between Fresh and Milk Sales
Milk Sales
Fresh Sales
0
10000
20000
30000
40000
50000
60000
70000
0
20000
40000
60000
80000
100000

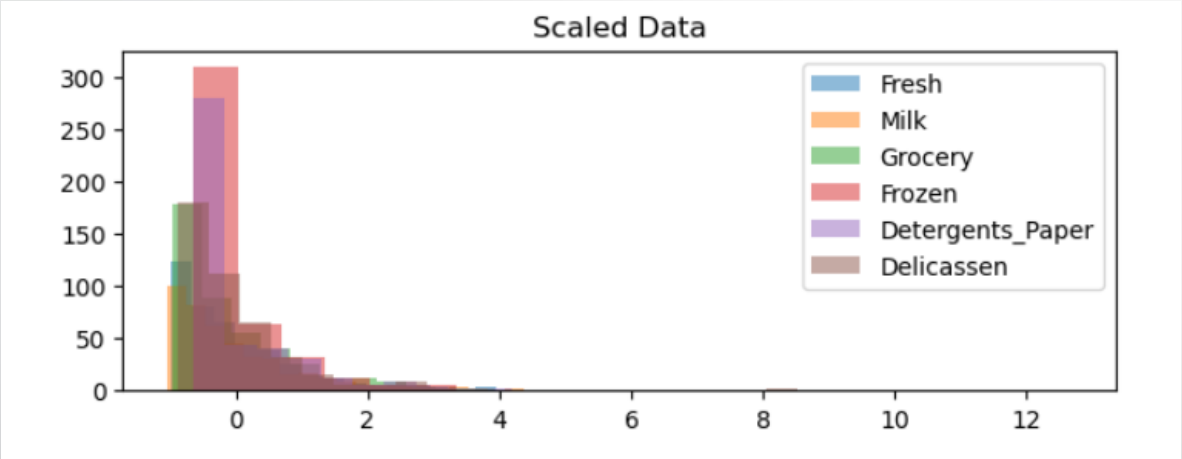
EDA - Heatmap

- Strong correlation noticed between grocery and Detergents_Paper (0.92). Some other significant correlations can be seen between channel and Detergents-paper, Grocery and Channel, Milk and Grocery, Milk and Detergents_Paper



Data Preprocessing

- Scaling Data



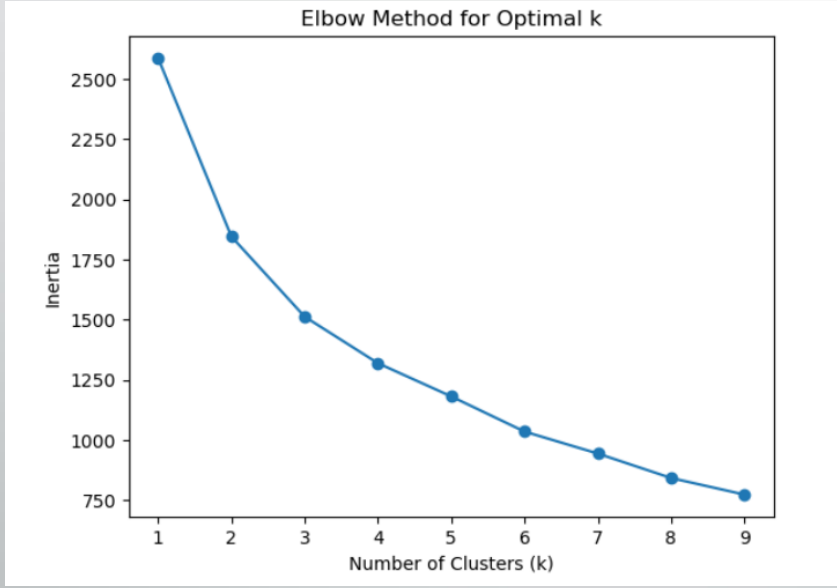
- Removed some rows with outliers

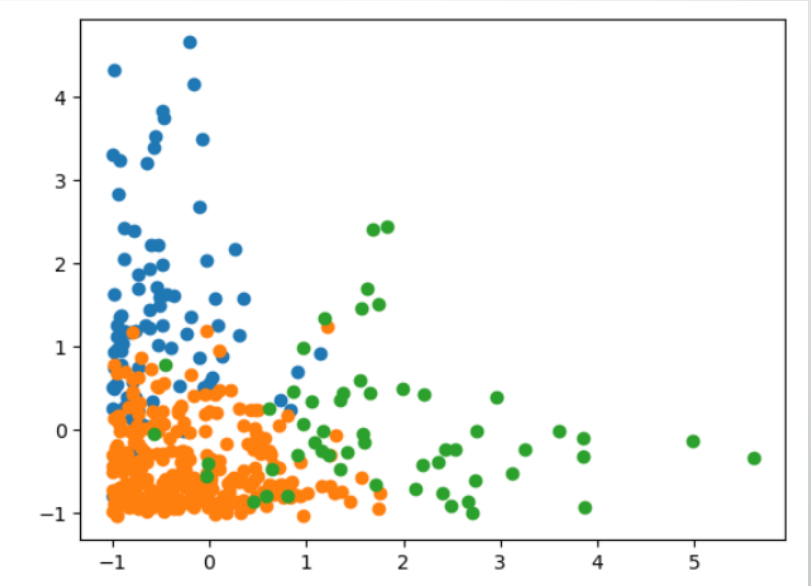
```
# Remove rows with z-scores outside the threshold
df_cleaned = df[(z_scores <= threshold) & (z_scores >= -threshold)]
```

```
df_cleaned.info()
```

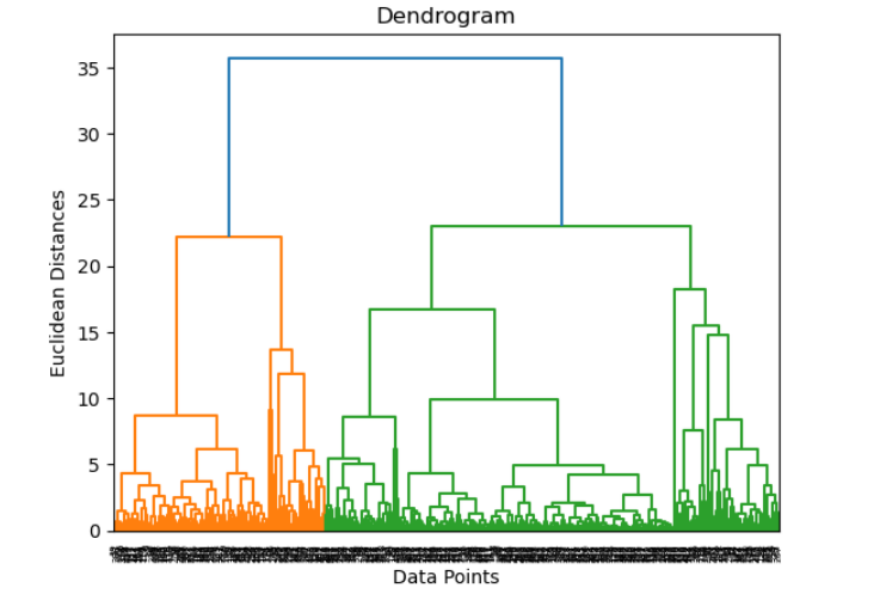
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 431 entries, 0 to 439
Data columns (total 8 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Channel           431 non-null    int64
 1   Region            431 non-null    int64
 2   Fresh             431 non-null    int64
 3   Milk              431 non-null    int64
 4   Grocery           431 non-null    int64
 5   Frozen            431 non-null    int64
 6   Detergents_Paper  431 non-null    int64
 7   Delicassen        431 non-null    int64
```

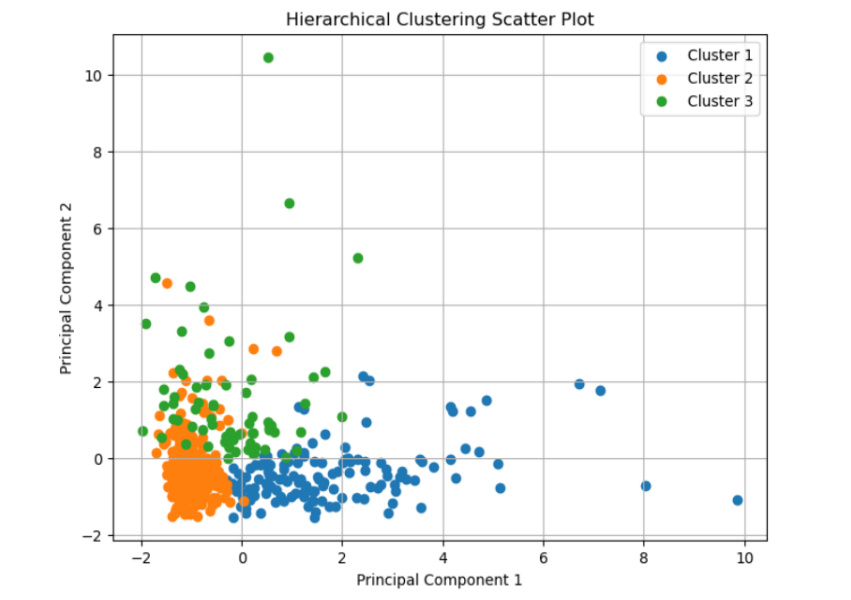
KMeans Clustering



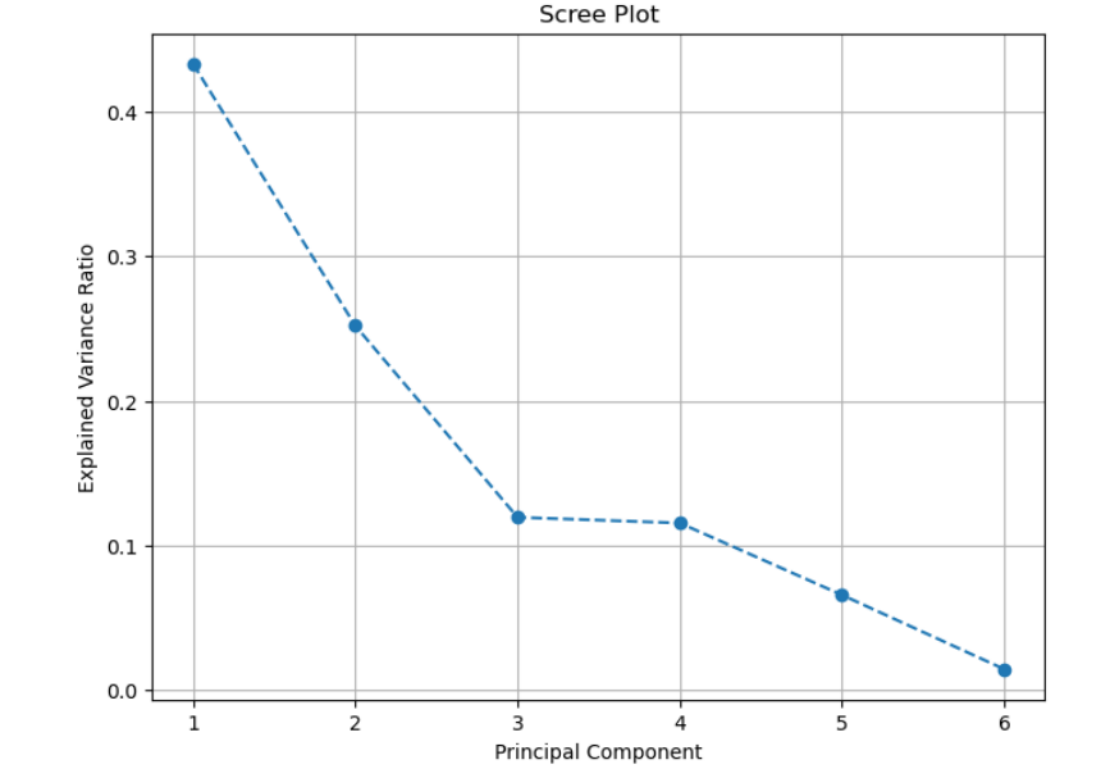


Hierarchical Clustering





PCA


Scree Plot
Explained Variance Ratio
0.0
0.1
0.2
0.3
0.4
1
2
3
4
5
6
Principal Component

Conclusion

- Using data visualization to understand correlation in the dataset, a strong correlation was noticed between grocery and detergents_paper (0.92).
- PCA was applied to reduce the dimensionality of the dataset while preserving essential information. The first few principal components captured most of the variance in the data, indicating that a reduced set of features can explain most of the dataset's variability.
- Through clustering analysis, distinct customer segments within dataset were identified. These segments provided valuable insights into different customer behaviors, such as high-spending customers, low-spending customers, and medium-spending customers.
- They were noticeable overlap in the clusters especially within the K-Means, this can be improved with further analysis