Unsupervised Learning Project

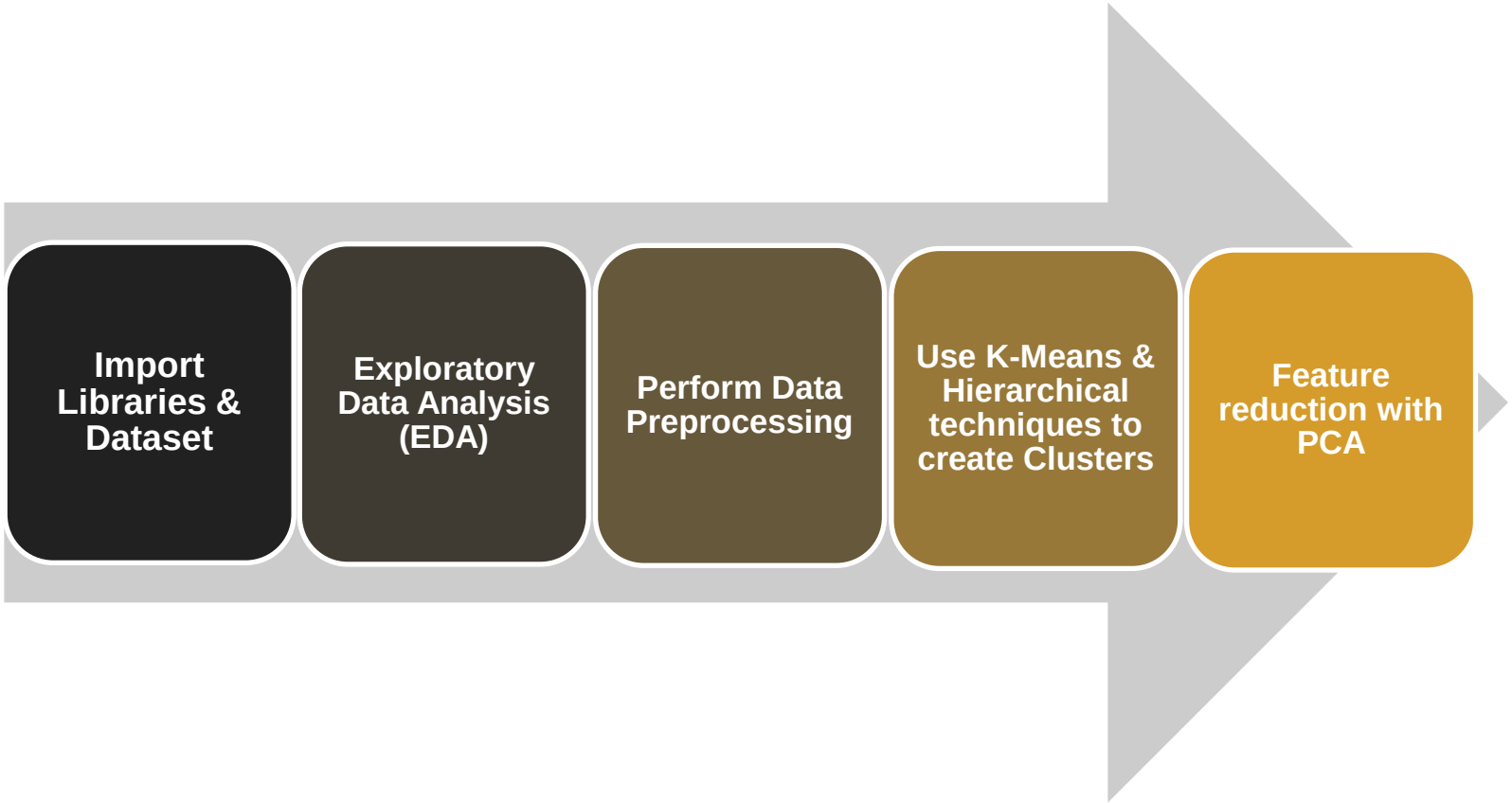
by
Abi Afolabi
28th Aug 2023

LIGHTHOUSE LABS

Project Overview & Goals

- Reinforce learning with hands-on experience
    - Apply unsupervised learning techniques to a real-world data set and use data visualization tools to communicate the insights gained from the analysis.

Project Execution Steps

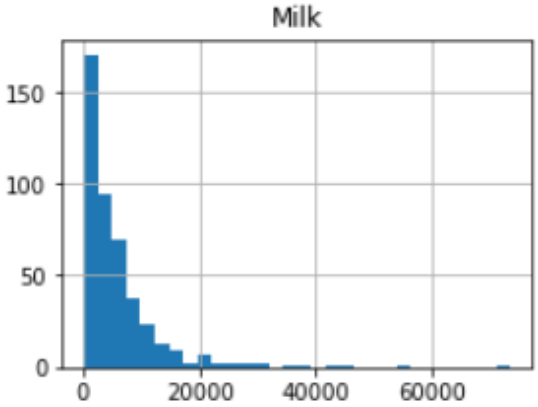


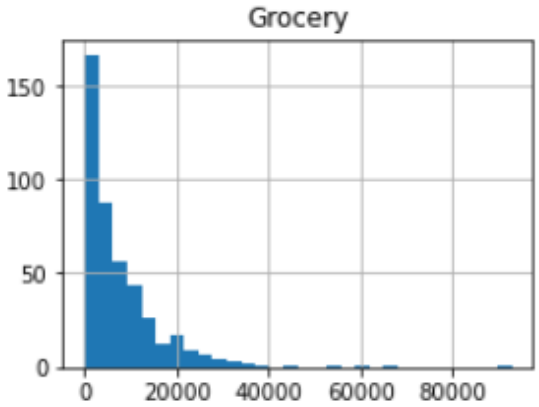
Dataset in Pandas DataFrame showing table columns and dimensions

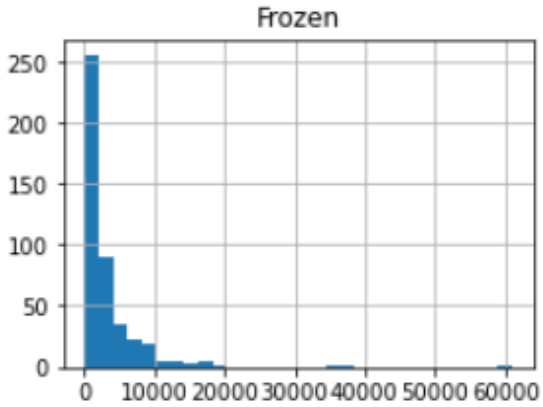
	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
0	2	3	12669	9656	7561	214	2674	1338
1	2	3	7057	9810	9568	1762	3293	1776
2	2	3	6353	8808	7684	2405	3516	7844
3	1	3	13265	1196	4221	6404	507	1788
4	2	3	22615	5410	7198	3915	1777	5185
...	...	...	...	...	...	...	...	...
435	1	3	29703	12051	16027	13135	182	2204
436	1	3	39228	1431	764	4510	93	2346
437	2	3	14531	15488	30243	437	14841	1867
438	1	3	10290	1981	2232	1038	168	2125
439	1	3	2787	1698	2510	65	477	52

440 rows × 8 columns

Discoveries from EDA – Histogram

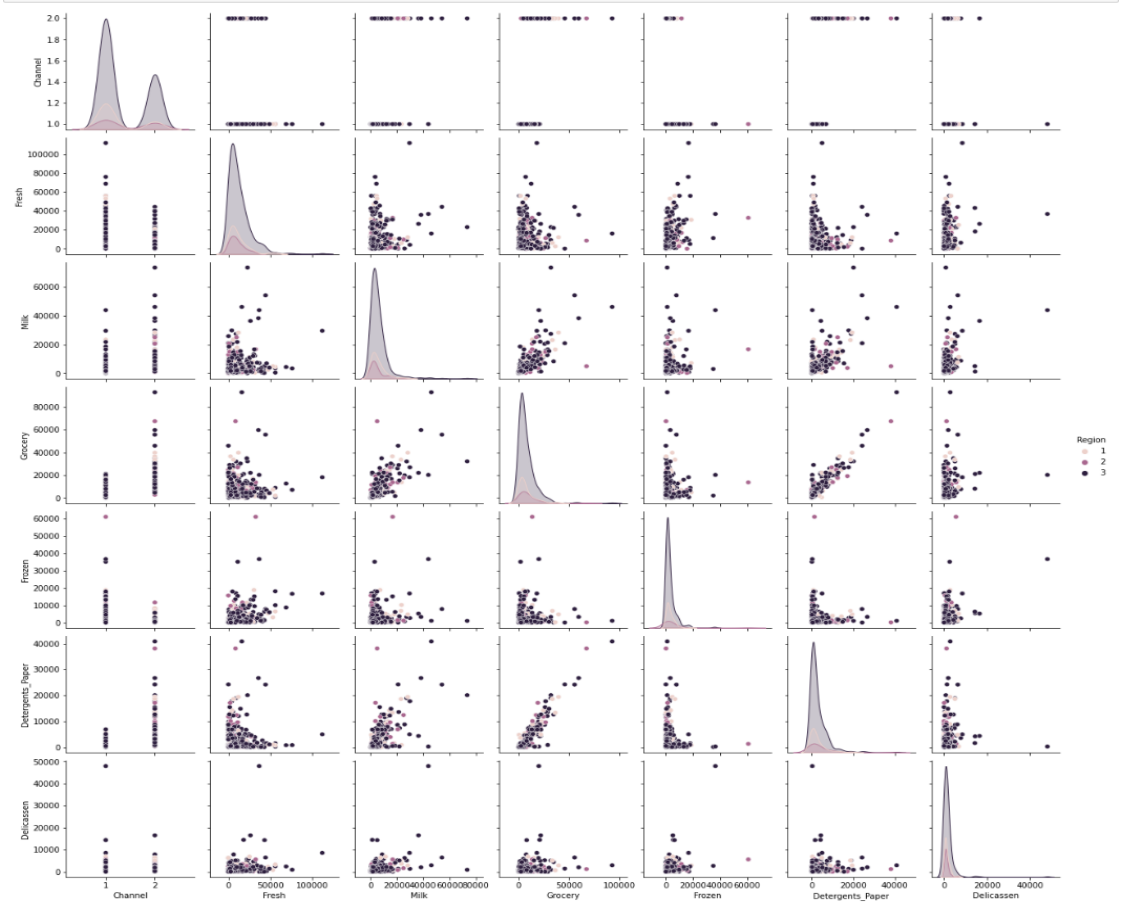






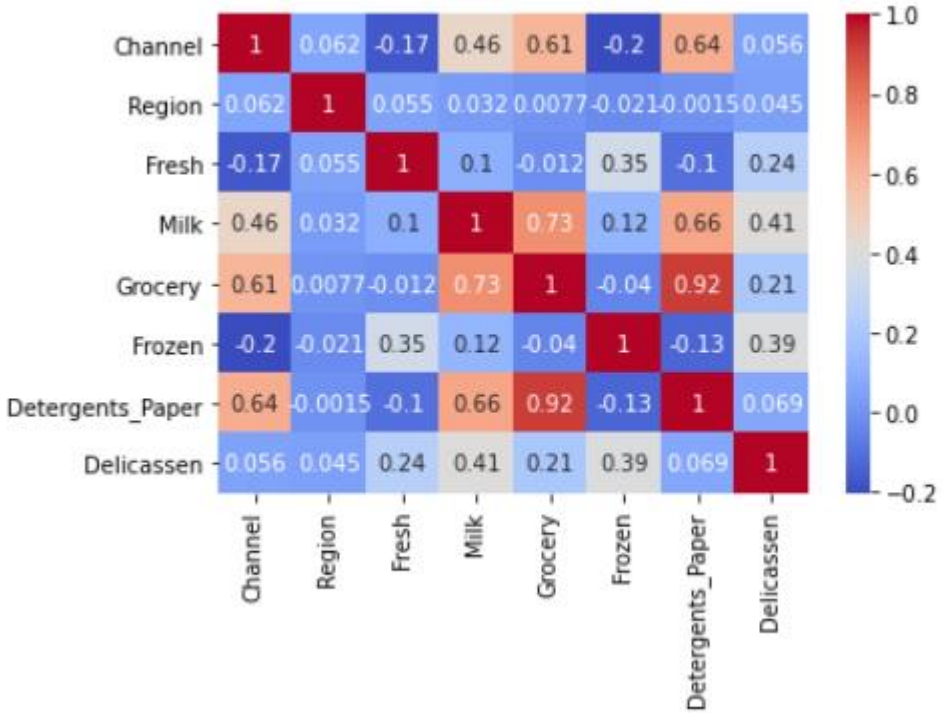
Dataset is mostly asymmetrical and right skewed

Discoveries from EDA – Pair plot



Pair plot shows most purchases occurred in region 3

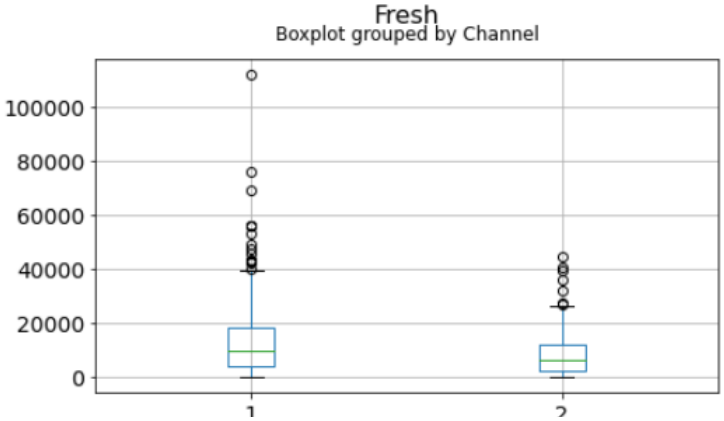
Discoveries from EDA – Correlation Matrix

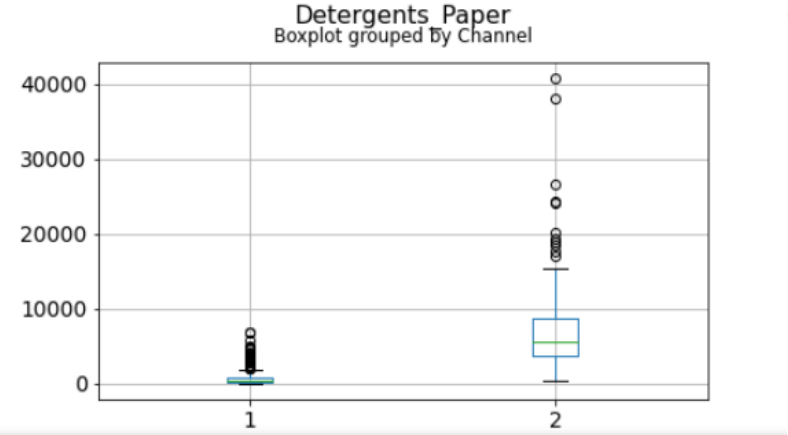


Varying levels of correlation between variables e.g.,

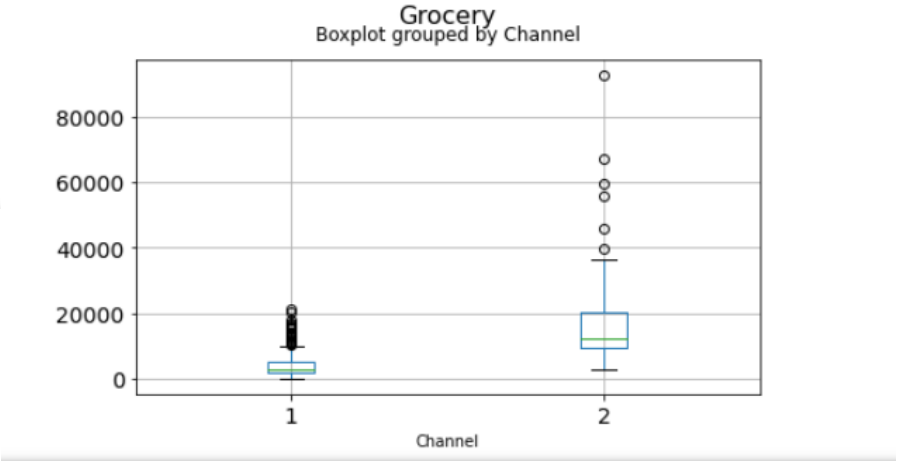
- High positive correlation between Milk & Grocery.
- Negative correlation between Frozen & Grocery

Discoveries from EDA – Box plots grouped by Channel

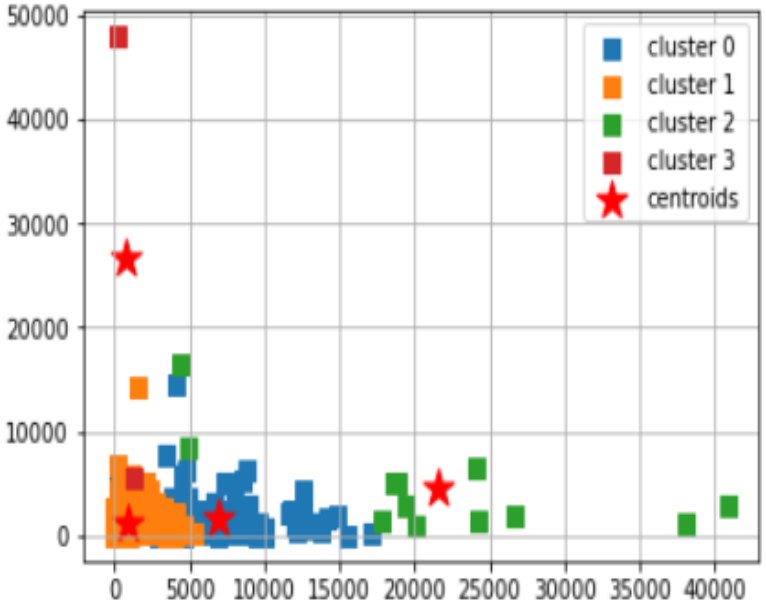




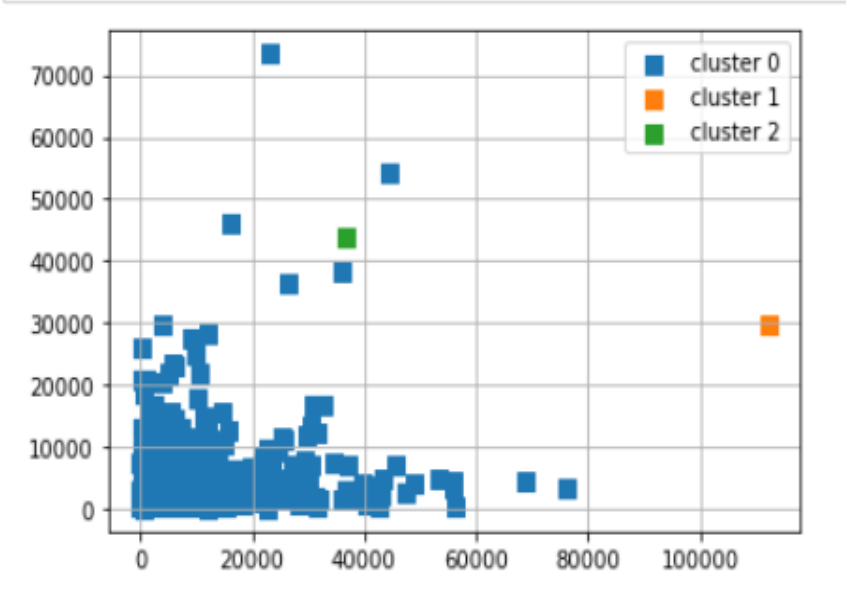
Presence of outliers in the upper quartile



Result of Clustering – K-Means & Hierarchical

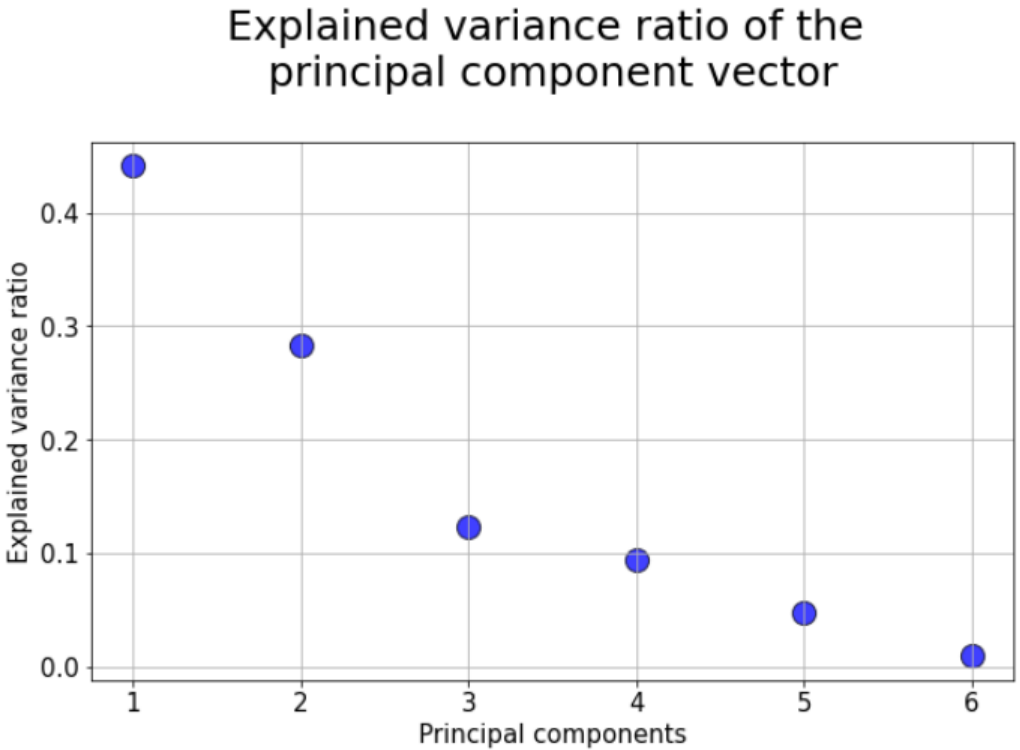


K-Means



Hierarchical

Dimensionality Reduction with PCA



Conclusion

- The dataset is asymmetrical and right skewed. Correlations amongst the features varies. There are strong positive as well as negative correlations.
- Most sales are from Region 3. Purchases from channel 1 is ~double of channel 2.
- The clustering in this dataset is not distinct. There are areas of ovelaps and the inter-cluster distance is low.
- All the six numerical features can be reduced to 2 or 3 components that will be representative of the features

https://github.com/AbiAfolabi/ml-project-unsupervised-learning