UNSUPERVISED MACHINE LEARNING FINAL PROJECT

Wholesale Customer Dataset

KAVITHA SUNDARAM

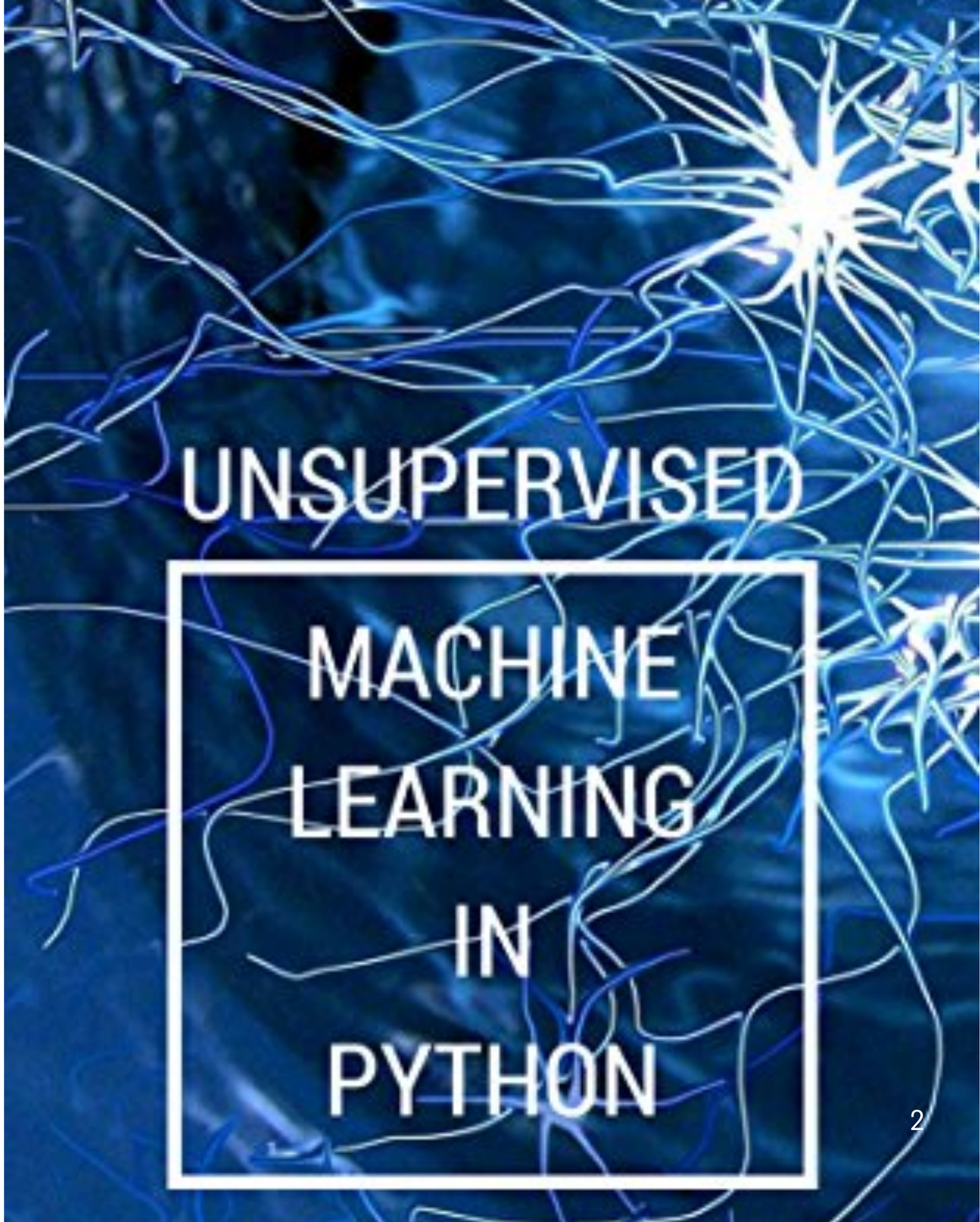
DATA SOURCE

- The data set refers to clients of a wholesale distributor. It includes the annual spending in monetary units (m.u.) on diverse product categories

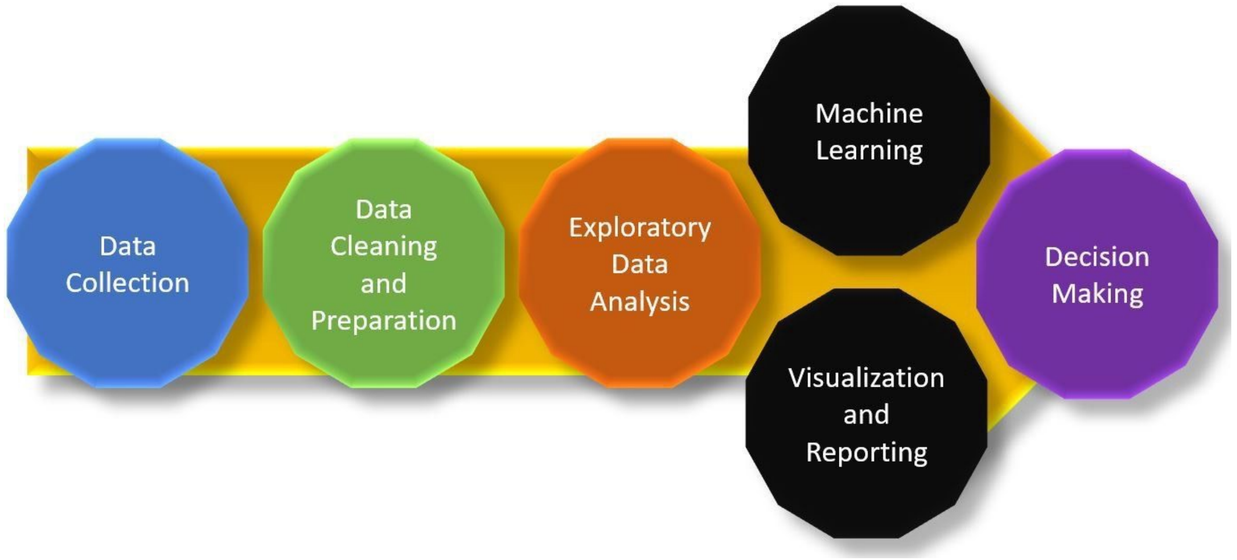
- Downloaded this wholesale customer dataset from UCI Machine Learning Repository.

- The wholesale distributor operating in different regions of Portugal has information on annual spending of several items in their stores across different regions and channels. The dataset consist of 440 large retailers annual spending on 6 different varieties of product in 3 different regions (lisbon , oporto, other) and across different sales channel (Hotel, channel)

DataSource: https://archive.ics.uci.edu/ml/datasets/wholesale+customers



CONTENTS



DESCRIPTION

- My goal is to use various clustering techniques to segment customers. Clustering is an unsupervised learning algorithm that tries to cluster data based on their similarity. Thus, there is no outcome to be predicted, and the algorithm just tries to find patterns in the data.
- Algorithms to be used, XGBoost classifier, k means clustering etc
- To predict which region and which channel will spend more and which region and channel to spend less.
- K-Means Clustering methods as Elbow,Silhoeutte, calinski_harabasz to find optimal K values to predict accuracy score in each products.

DESCRIPTION

```
RangeIndex: 440 entries, 0 to 439
Data columns (total 8 columns):
 #   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)
memory usage: 27.6 KB

None
```

	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

DESCRIPTION

	count	mean	std	min	25%	50%	75%	max
Channel	440.0	1.322727	0.468052	1.0	1.00	1.0	2.00	2.0
Region	440.0	2.543182	0.774272	1.0	2.00	3.0	3.00	3.0
Fresh	440.0	12000.297727	12647.328865	3.0	3127.75	8504.0	16933.75	112151.0
Milk	440.0	5796.265909	7380.377175	55.0	1533.00	3627.0	7190.25	73498.0
Grocery	440.0	7951.277273	9503.162829	3.0	2153.00	4755.5	10655.75	92780.0
Frozen	440.0	3071.931818	4854.673333	25.0	742.25	1526.0	3554.25	60869.0
Detergents_Paper	440.0	2881.493182	4767.854448	3.0	256.75	816.5	3922.00	40827.0
Delicassen	440.0	1524.870455	2820.105937	3.0	408.25	965.5	1820.25	47943.0

- Attribute Information:
- FRESH: annual spending (m.u.) on fresh products (Continuous);
- MILK: annual spending (m.u.) on milk products (Continuous);
- GROCERY: annual spending (m.u.)on grocery products (Continuous);
- FROZEN: annual spending (m.u.)on frozen products (Continuous)
- DETERGENTS_PAPER: annual spending (m.u.) on detergents and paper products (Continuous)
- DELICATESSEN: annual spending (m.u.)on and delicatessen products (Continuous);
- CHANNEL: customers Channel - Horeca (Hotel/Restaurant/Cafe) or Retail channel (Nominal)
- REGION: customers Region Lisnon, Oporto or Other (Nominal)
-

EXPLORATORY DATA ANALYSIS(EDA)

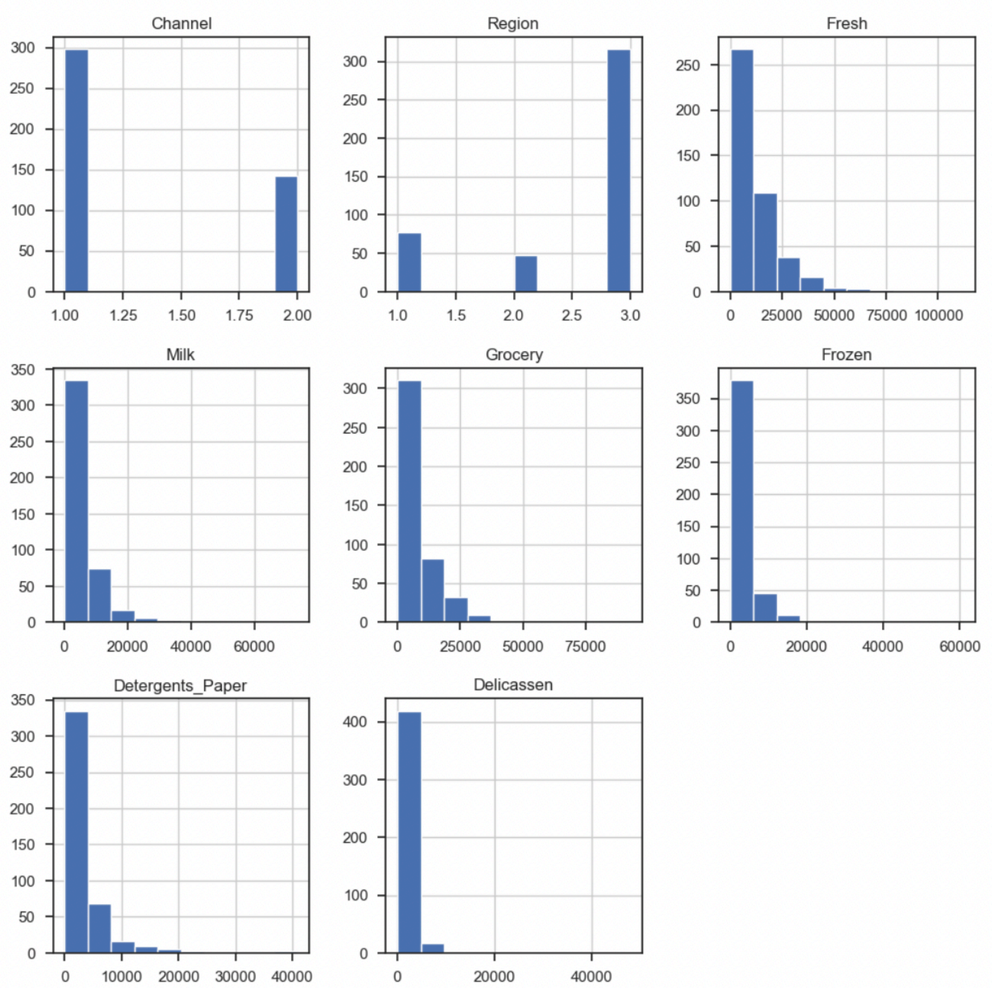
- Lets Check null values and data types of all variables for model analysis.

```
Channel             int64
Region              int64
Fresh               int64
Milk                int64
Grocery             int64
Frozen              int64
Detergents_Paper    int64
Delicassen          int64
dtype: object
```

```
Channel             0
Region              0
Fresh               0
Milk                0
Grocery             0
Frozen              0
Detergents_Paper    0
Delicassen          0
dtype: int64
```

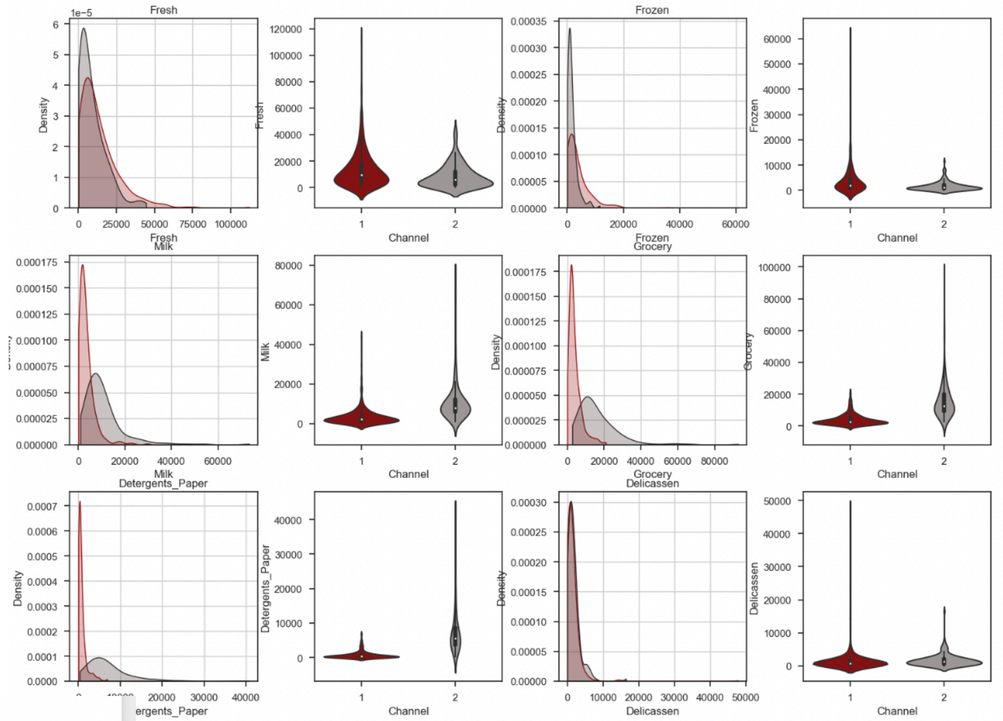
EXPLORATORY DATA ANALYSIS(EDA)

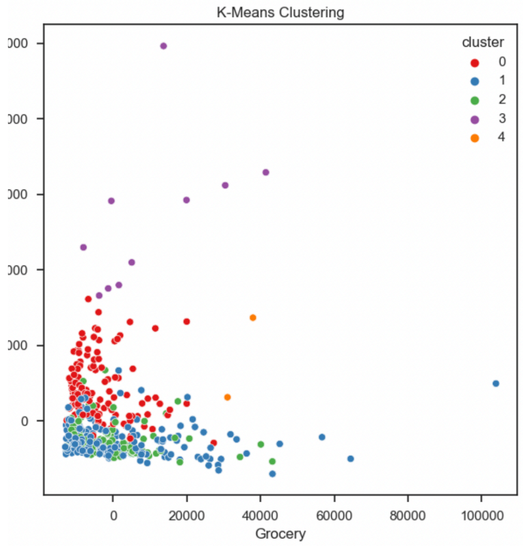
- After analysing above histograms, we can easily divide our variables into
- categorical(channel, region)
- numerical(fresh, frozen, milk, delicatessen, detergents_paper, grocery)



DATA PREPROCESSING

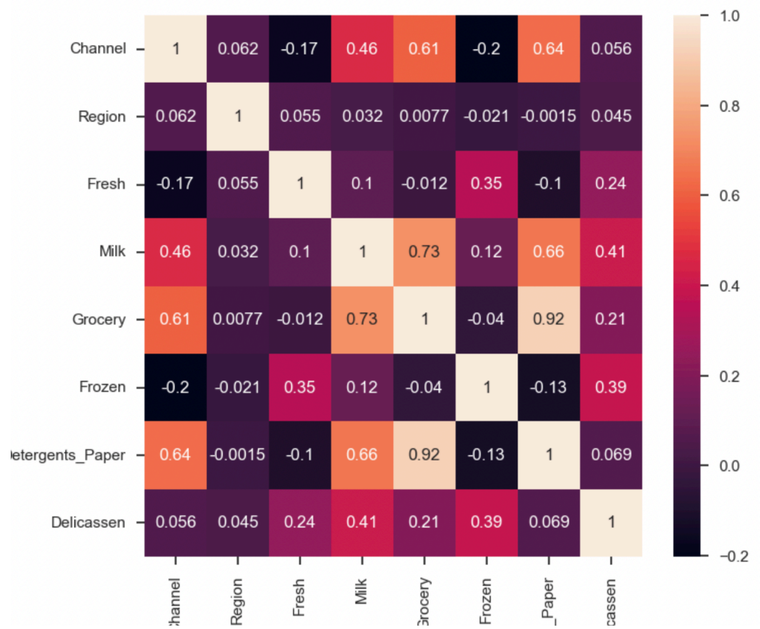
- Look at the structure of **Grocery** and **Detergents_Paper**, both are having difference in violin plots.
- Lets analyse more of categorical datatypes.
- There is **strong correlation** between 'grocery' and 'detergents_paper' and customers who buy grocery along with detergents_paper spend more money in this two products.





DATA PREPROCESSING

- 1. Most of the varuables are uncorrelated. as you can see **Grocery** and **Detergents_Paper** are positively correlated.
- 2. **92%** strong correlation between grocery and detergents products.
- Milk and Fresh share a weak correlation.

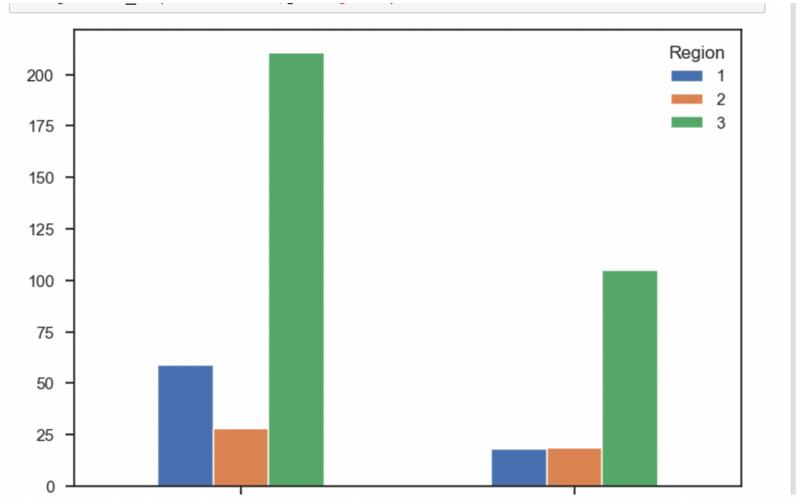


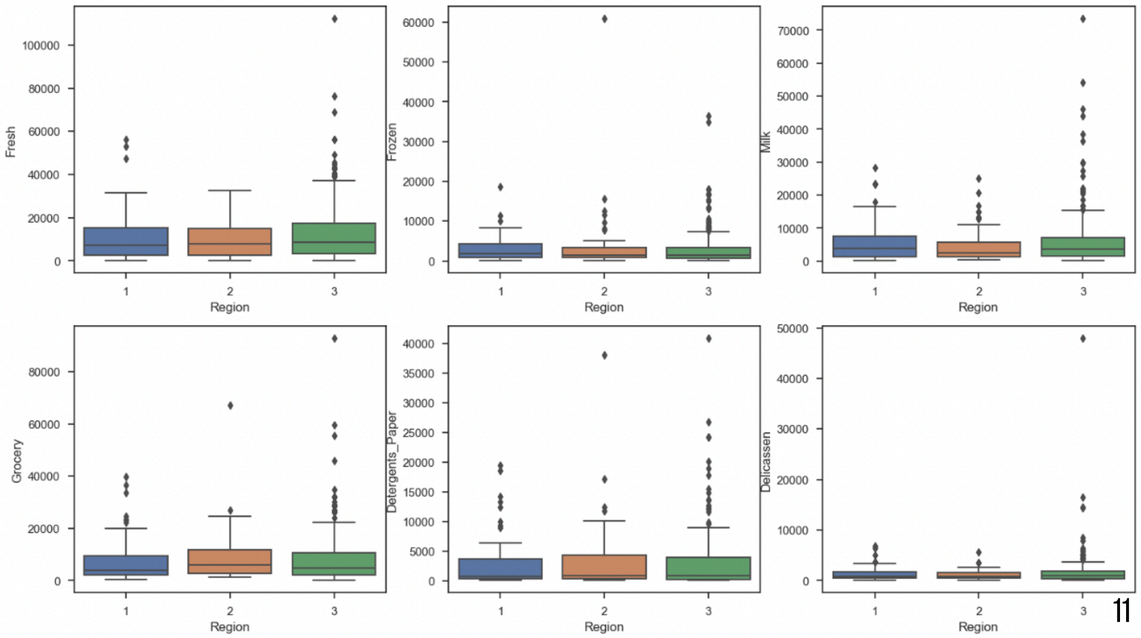
6]:

	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicasse
Channel	1.000000	0.062028	-0.169172	0.460720	0.608792	-0.202046	0.636026	0.05601
Region	0.062028	1.000000	0.055287	0.032288	0.007696	-0.021044	-0.001483	0.04521
Fresh	-0.169172	0.055287	1.000000	0.100510	-0.011854	0.345881	-0.101953	0.24469
Milk	0.460720	0.032288	0.100510	1.000000	0.728335	0.123994	0.661816	0.40636
Grocery	0.608792	0.007696	-0.011854	0.728335	1.000000	-0.040193	0.924641	0.20549
Frozen	-0.202046	-0.021044	0.345881	0.123994	-0.040193	1.000000	-0.131525	0.39094
ts_Paper	0.636026	-0.001483	-0.101953	0.661816	0.924641	-0.131525	1.000000	0.06929
licassen	0.056011	0.045212	0.244690	0.406368	0.205497	0.390947	0.069291	1.00000

DATA PREPROCESSING

- From this categorical plot, we can define highest spending channel = 1 and Lowest spending channel = 2.
- Highest spending Region = 3 and Lowest spending Region = 2
- There are some outliers in the data.





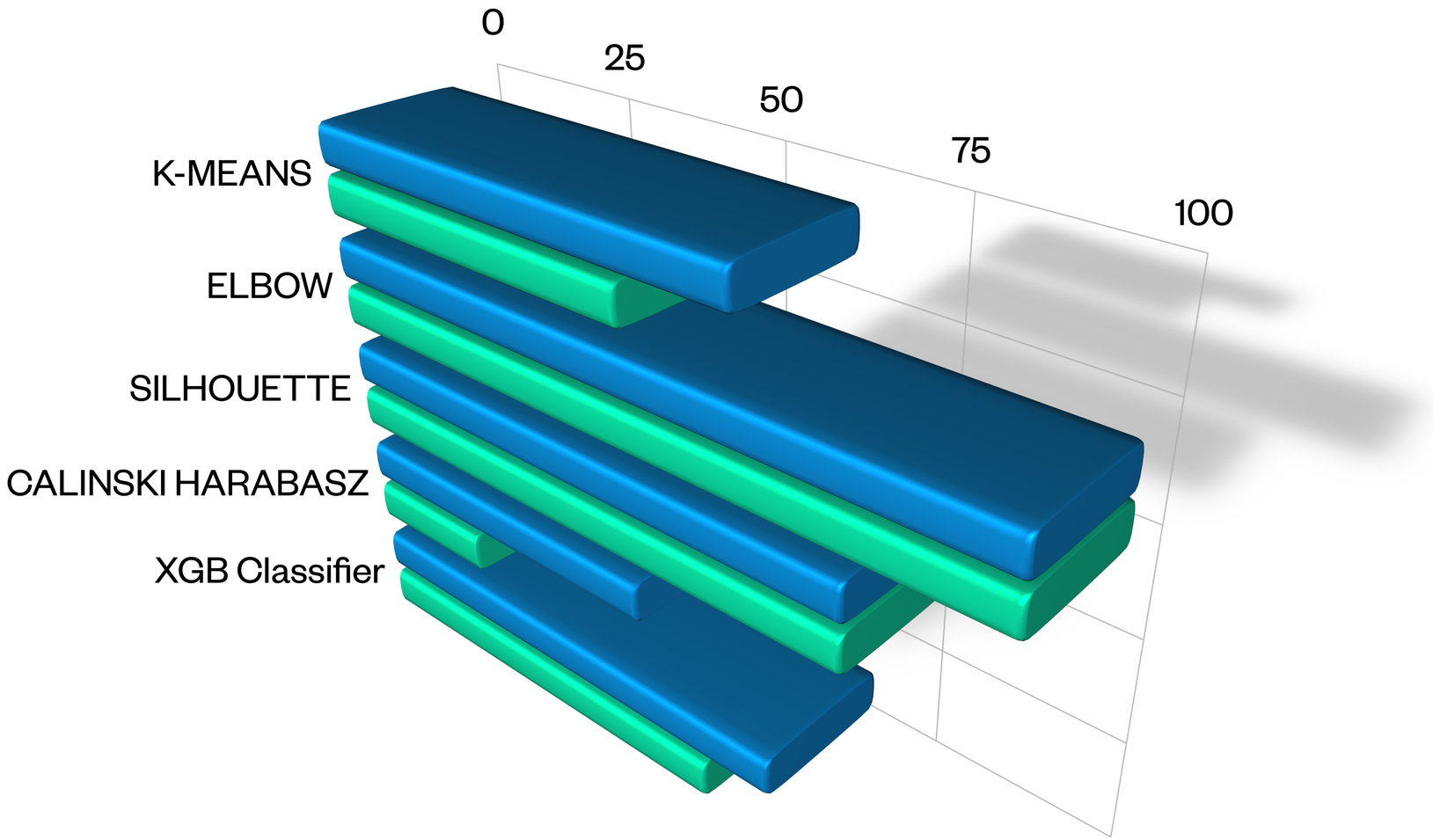
We have some outliers in the data.

DATA PREPROCESSING

Channel1	Channel2
Highest Spending = Fresh	Highest Spending = Grocery
Lowest Spending = Detergents_paper	Lowest Spending = Frozen

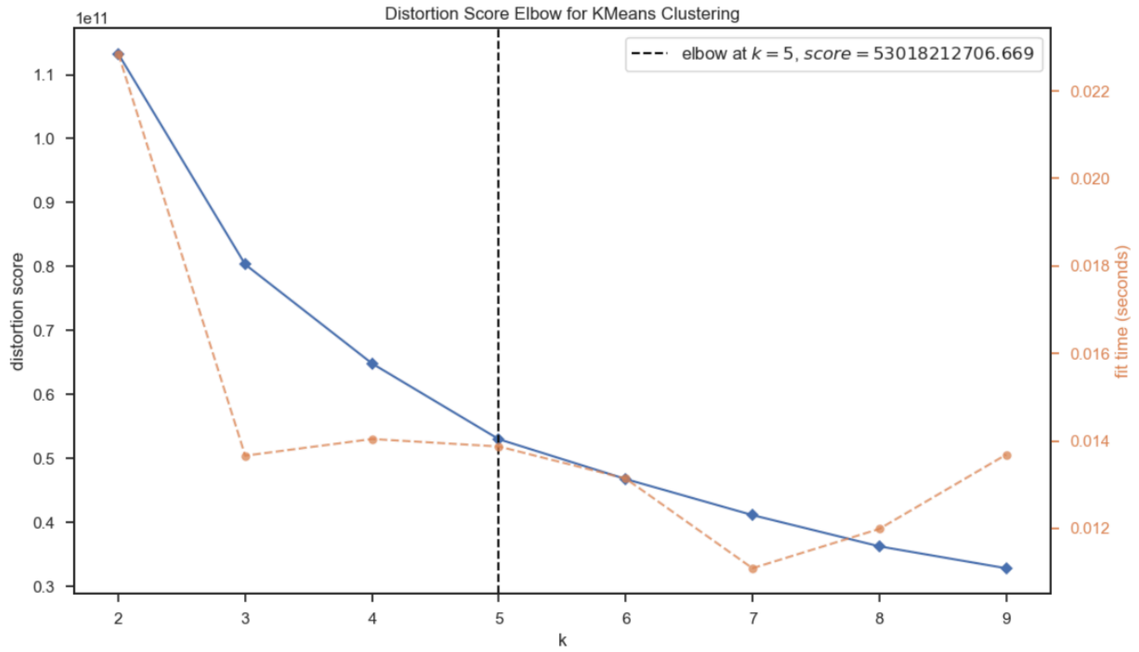
Region1	Region2	Region3
Highest Spending = Fresh	Highest Spending = Fresh	Highest Spending = Fresh
Lowest Spending = Delicassen	Lowest Spending = Delicassen	Lowest Spending = Delicassen

MODEL CLUSTERING



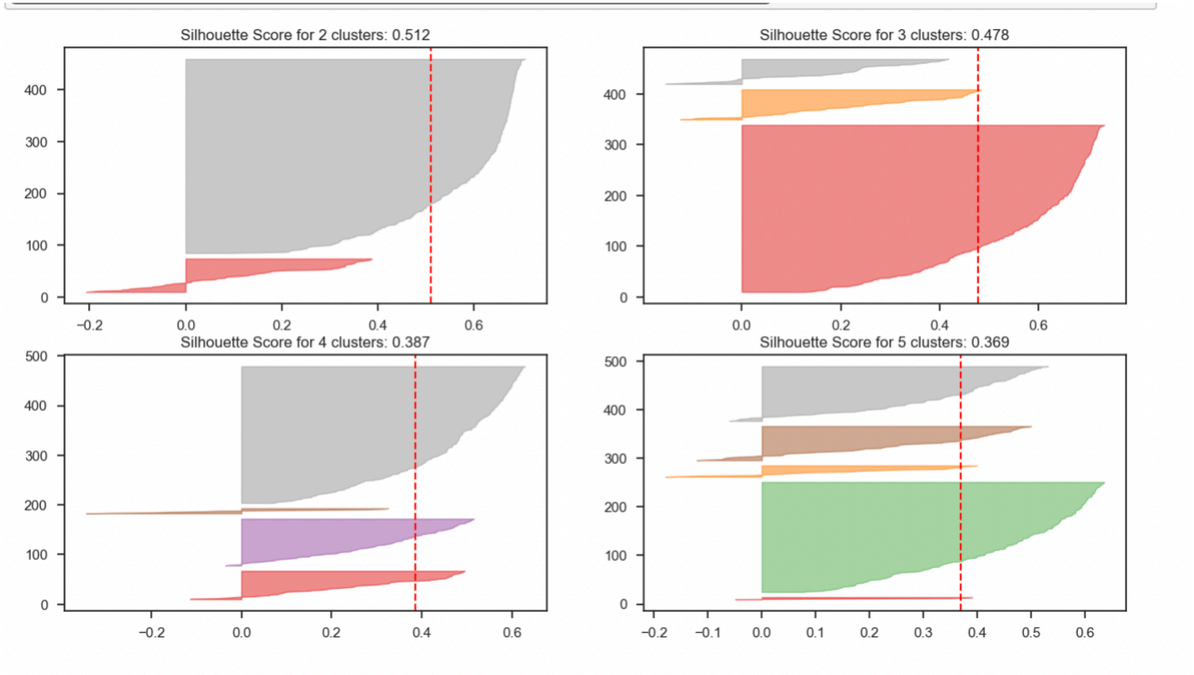
K-MEANS CLUSTERING

- We have initialized two clusters and pay attention the initialization is not random here.
- We got an inertia value of almost 2149. Now, let's see how we can use the elbow method to determine the optimum number of clusters in Python.
- As you can see from above plot, elbow visulaizer is clearly showing cluster k =5 of score 0.530.To find optimal number of clusters, elbow and silhouette methods are used.



K-MEANS CLUSTERING

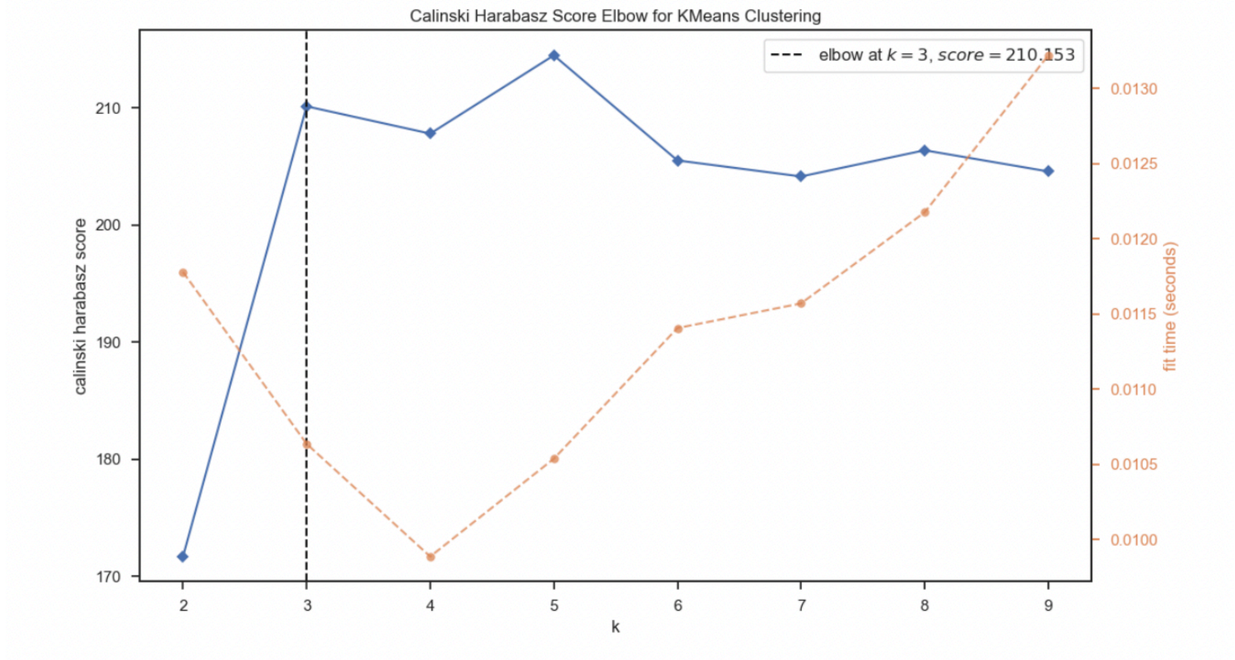
- As you can see from above plot, SILHOUETTE visulaizer is clearly showing cluster k =2 of score 0.512.



SILHOUETTE K-MEANS

K-MEANS CLUSTERING

- As you can see from above plot, CALINSKI HARABASZ visulaizer is clearly showing cluster k =3 of score 0.210.

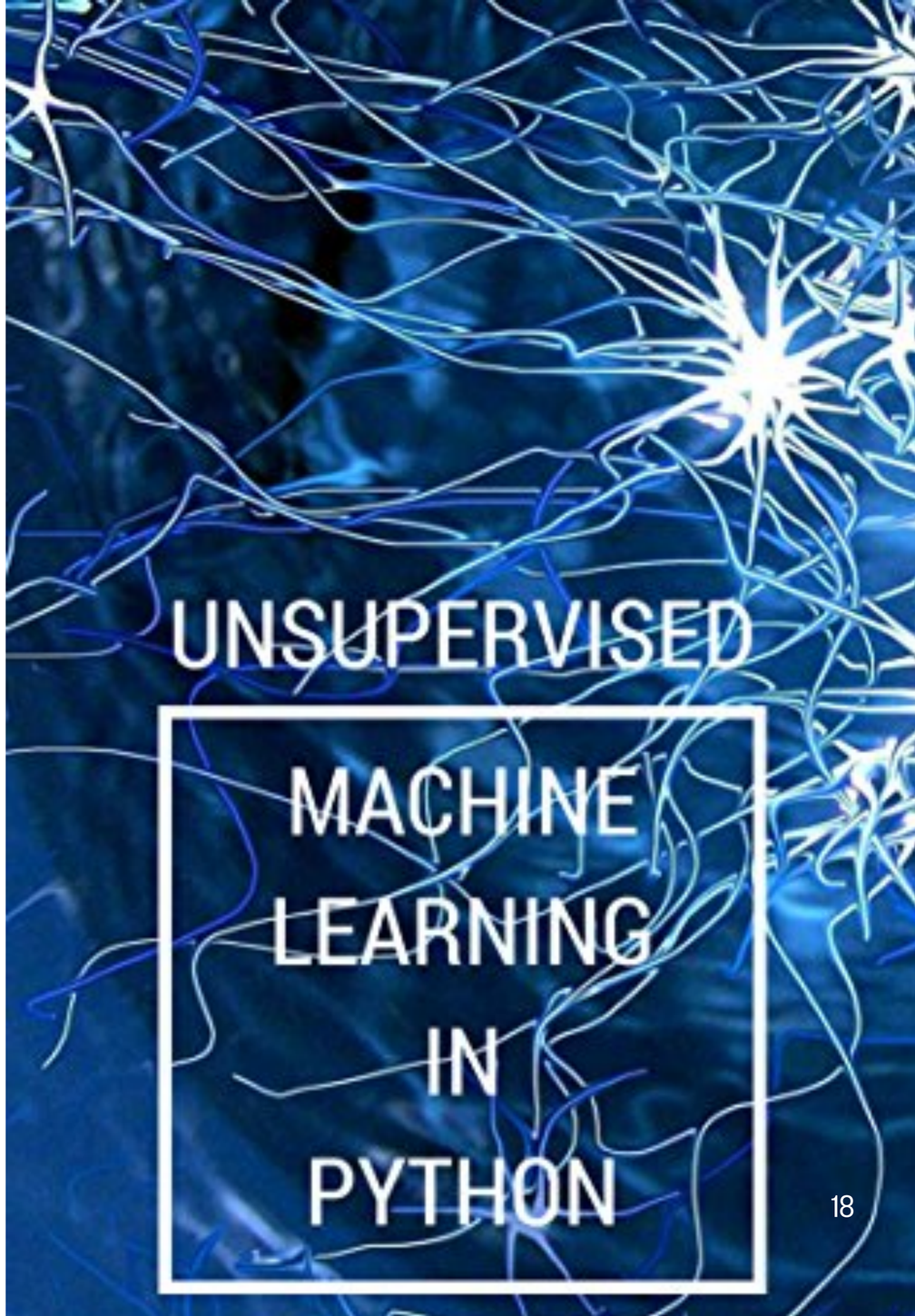


RESULT AND ANALYSIS

- optimizing values using **elbow** method, K=5, score as **0.530**
- Sihoutte method, k = 2 , score as 0.512
- calinski_harabasz method, k = 3, score as 0.221
- Accuracy with XGB Classifier = 97.50

CONCLUSION

- Best overall model seems to be the K-Means Elbow Method with K = 5 optimized value with score = 0.530 on the sampled dataset, that delivers the best results in terms of accuracy .
- I have used almost all Clustering K-means algorithm models to predict the score of each customers features according to the feature provided with dataset.
- I also want to look into feature selection for logistic regression algorithms.
- There are some other clustering models as PCA, principle component analysis and find centroid and
- optimize along with dataset untill we find the center point and NMF non- matrix factorization methods to find optimal value in data points in future.
- So, the customers who bought grocery along with detergents paper spends more money than other products like frozen
- Fresh and frozen products are bought least . We can make discounts for grocery and detergents products for higher sales in wholesale customers.



REFERENCES

kaggle references:
https://datagy.io/seaborn-catplot/ \
https://dev.to/thalesbruno/subplotting-with-matplotlib-and-seaborn-5ei8 \
https://www.statology.org/coefficient-of-variation-in-python/
#:~:text=CV%20%3D%20%CF%83%20%2F%20%CE%BC,%CE%BC%3A%20The%20mean%20of%20dataset\
https://www.kaggle.com/code/farhanmd29/unsupervised-learning/notebook\
https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/ \
https://towardsdatascience.com/cheat-sheet-to-implementing-7-methods-for-selecting-optimal-number-of-clusters-in-python-898241e1d6ad#:~:text=Calculating%20gap%20statistic%20in%20python,with%20varying%20number%20of%20clusters.

GITHUB REPOSITORY LINK

https://github.com/kavishant87/UnSupervised_Final_5510_Project

UNSUPERVISED MACHINE LEARNING FINAL PROJECT

THANKS FOR WATCHING...