Final Project

ML Unsupervised Learning - Wholesale Customers

- IBM Machine Learning Profession Certificate
- By Sadhana Jarag

Content

- Dataset Description
- Main Objective of Analysis
- EDA, Cleaning, Feature Engineering
- Applying various Clustering algorithms
- Machine learning analysis and findings
- Model Flaws and advance step

Data Description Section

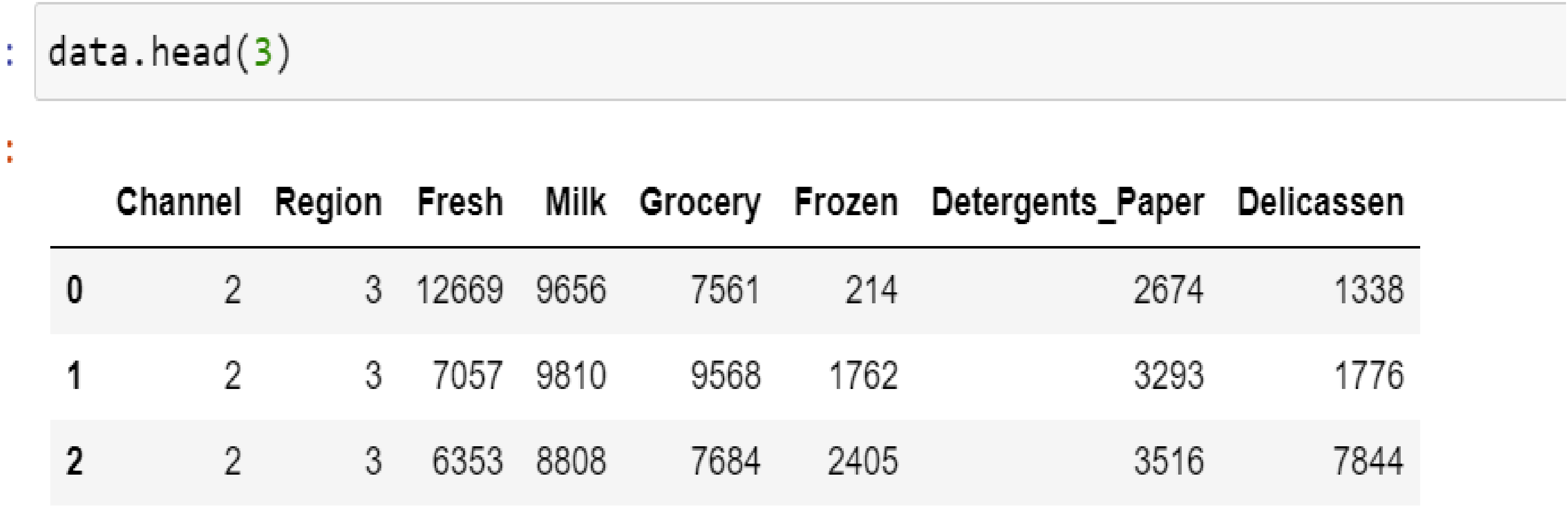
Introduction :

- The wholesale business comprises the **sale of products in bulk, and at a lower price**. Ultimately, this reduces the costs and the handling time involved.
- There are several types of wholesale businesses. A wholesale business can be in the form of merchant wholesale, brokers, or sales and distribution.
- The wholesale business model is based on buying in bulk with a significant discount from the producer/manufacturer. This way, the wholesaler will be able to sell the products to retailers on a nice margin, making a profit, thus, through markup.
- The dataset refers to clients of a wholesale distributor.
- It includes the annual spending in monetary units (m.u.) on diverse product categories**.**

Project Introduction:

- The data set refers to clients of a wholesale distributor.
- It includes the annual spending in monetary units on diverse product categories.
- It contains several products and like Fresh, Milk Grocery etc.
- They are having the different channel as well as different region for the wholesale distributor.

Dataset Description :Part 1



```
data.head(3)
```

	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

Dataset Description :Part 2

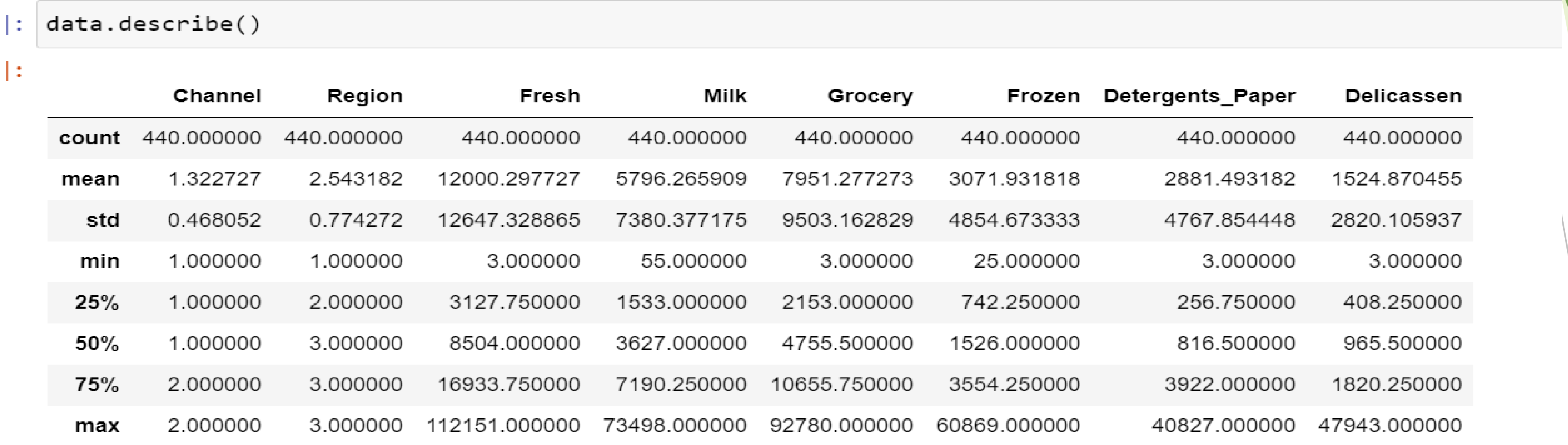
- Attribute Information:
- 1- FRESH: annual spending on fresh products (Continuous)
- 2- MILK: annual spending on milk products (Continuous)
- 3- GROCERY: annual spending on grocery products (Continuous)
- 4- FROZEN: annual spending on frozen products (Continuous)
- 5- DETERGENTS_PAPER: annual spending on detergents and paper products (Continuous)
- 6- DELICATESSEN: annual spending on and delicatessen products (Continuous)
- 7- CHANNEL: customers™ Channel - Horeca (Hotel/Restaurant/Cafe) or Retail channel (Nominal)
- 8- REGION: customers™ Region“ Lisnon, Oporto or Other (Nominal)

Dataset Description

Descriptive Statistics:

- (Minimum, Maximum, Mean, Std. Deviation)
- FRESH (3, 112151, 12000.30, 12647.329)
- MILK (55, 73498, 5796.27, 7380.377)
- GROCERY (3, 92780, 7951.28, 9503.163)
- FROZEN (25, 60869, 3071.93, 4854.673)
- DETERGENTS_PAPER (3, 40827, 2881.49, 4767.854)
- DELICATESSEN (3, 47943, 1524.87, 2820.106)
- REGION Frequency Lisbon 77 Oporto 47 Other Region 316 Total 440
- CHANNEL Frequency Horeca 298 Retail 142 Total 44

Descriptive Statistics:

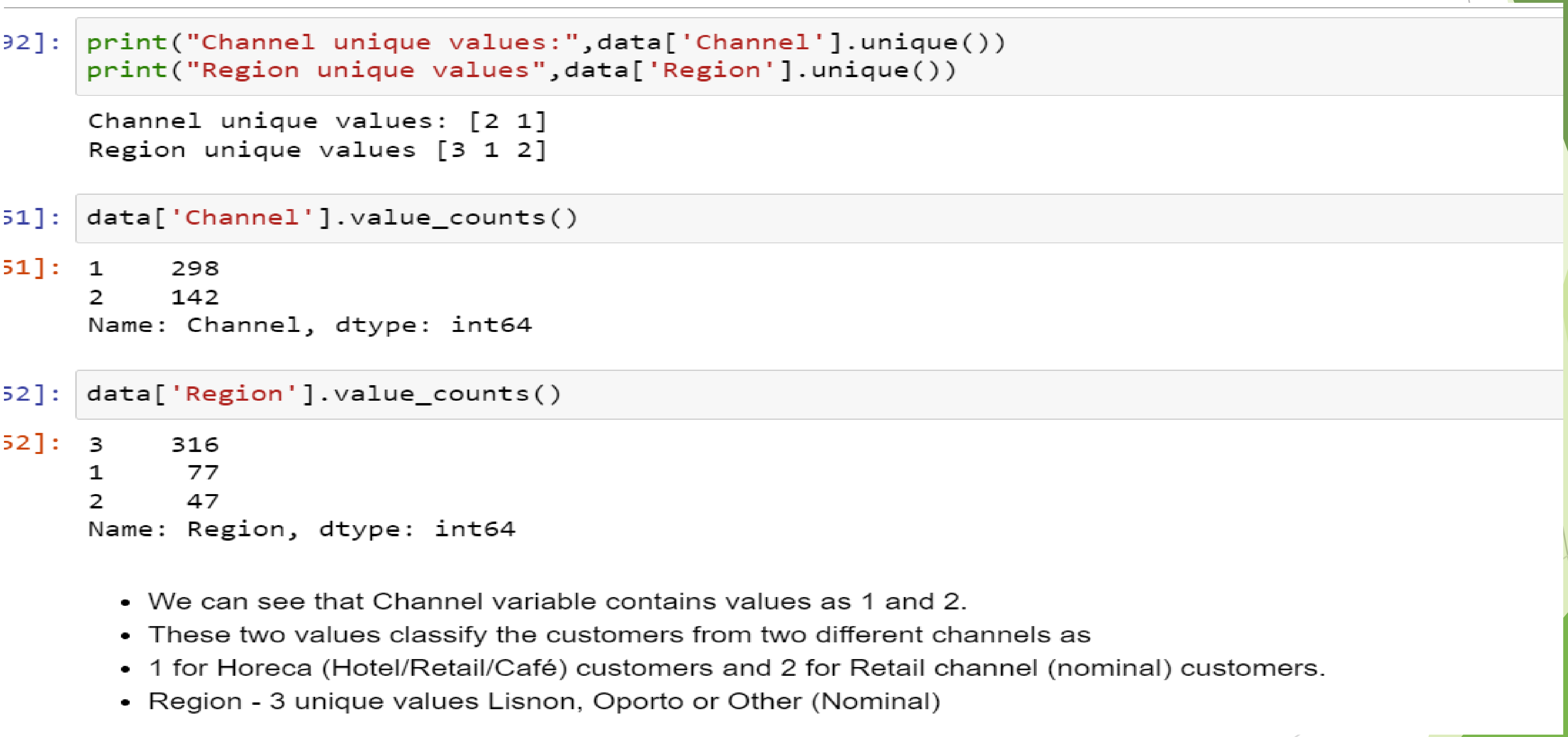


```
data.describe()
```

	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

- Insights :
- 1- Annual spending on fresh products is min 3 and max gets to 112151. that is having huge gap
- 2- Same we can fine in Grocery with value min as 3 and max as 92780
- 3-Others contineus featre are having big gap in min and max value.

Discrete value description:



```
92]: print("Channel unique values:",data['Channel'].unique())
     print("Region unique values",data['Region'].unique())

     Channel unique values: [2 1]
     Region unique values [3 1 2]
```

```
51]: data['Channel'].value_counts()

51]: 1    298
     2    142
     Name: Channel, dtype: int64
```

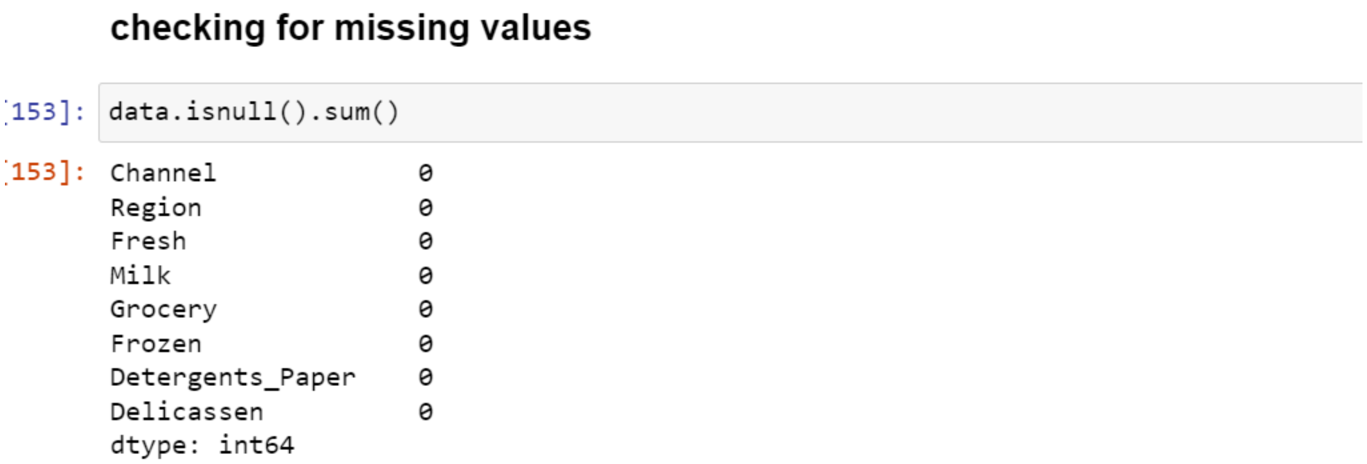
```
52]: data['Region'].value_counts()

52]: 3    316
     1     77
     2     47
     Name: Region, dtype: int64
```

- We can see that Channel variable contains values as 1 and 2.
- These two values classify the customers from two different channels as
- 1 for Horeca (Hotel/Retail/Café) customers and 2 for Retail channel (nominal) customers.
- Region - 3 unique values Lisnon, Oporto or Other (Nominal)

Main Objective of Analysis

- In this section we will explore the dataset in depth through several EDA techniques such as checking null values, outliers, skewness of the features and some visualization which can help to get know more about the data.
- We can check the correlation of each feature with respect to each other so we will get an idea about the data strength.

Checking for the missing values



checking for missing values

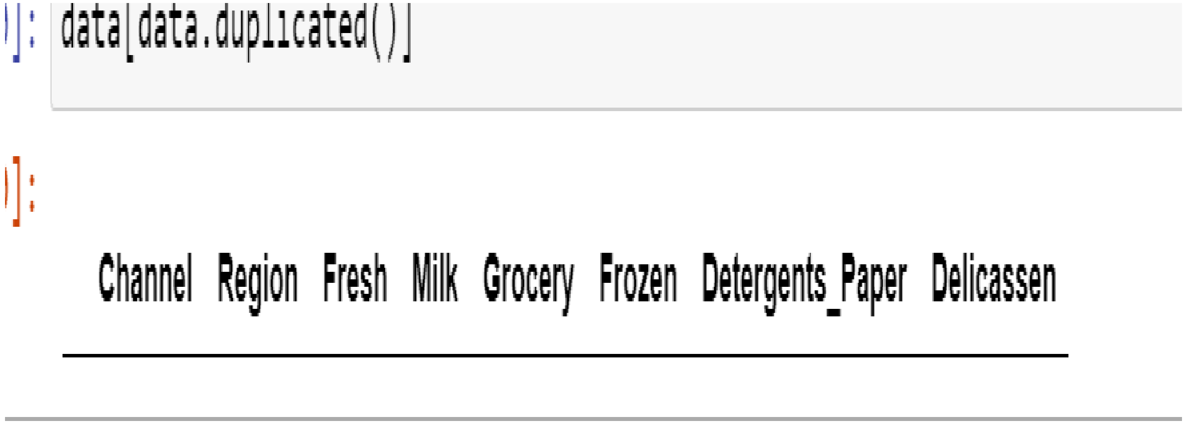
```
[153]: data.isnull().sum()
```

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

Insights : No missing value in this dataset

Checking for duplicated values

Check for duplicate Value: No duplicate Value found

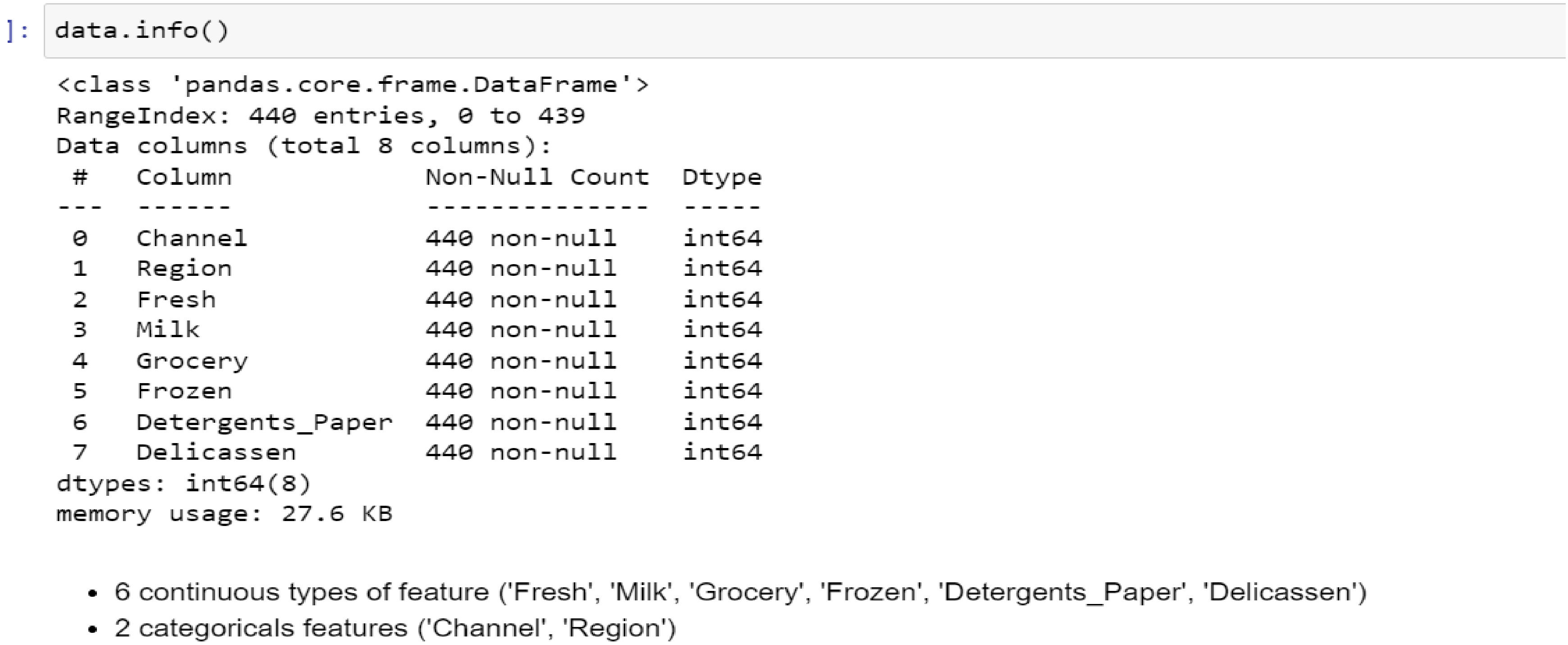


```
]: data[data.duplicated()]
```

```
]:
   Channel  Region  Fresh  Milk  Grocery  Frozen  Detergents_Paper  Delicassen
```

EDA :Exploratory Data Analysis

EDA :Exploratory Data Analysis



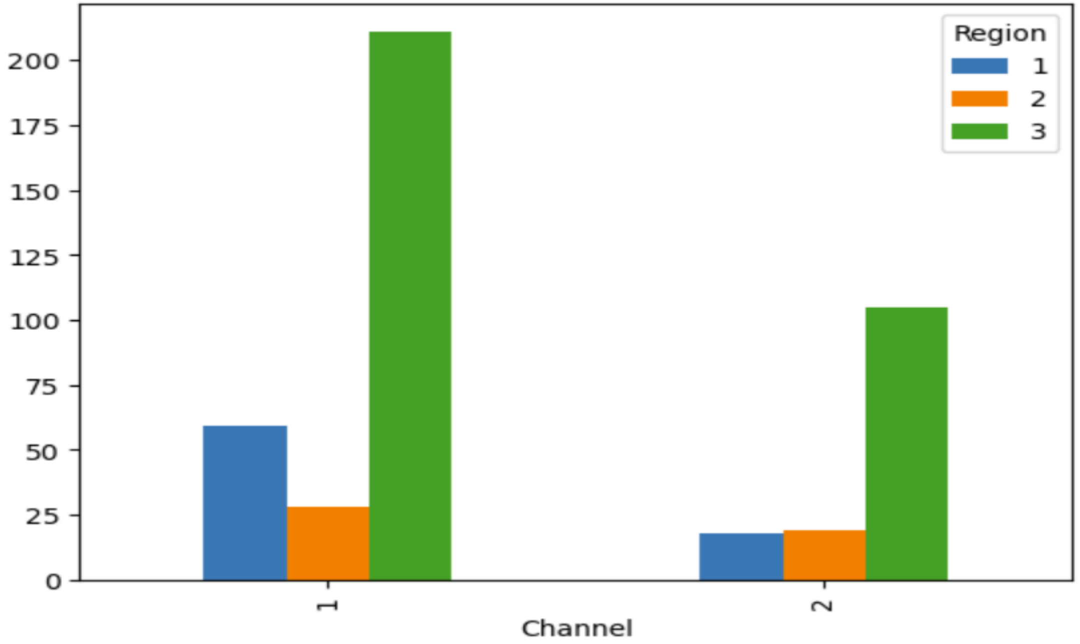
```
]: data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
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
```

- 6 continuous types of feature ('Fresh', 'Milk', 'Grocery', 'Frozen', 'Detergents_Paper', 'Delicassen')
- 2 categoricals features ('Channel', 'Region')

- All values are in integer datatypes

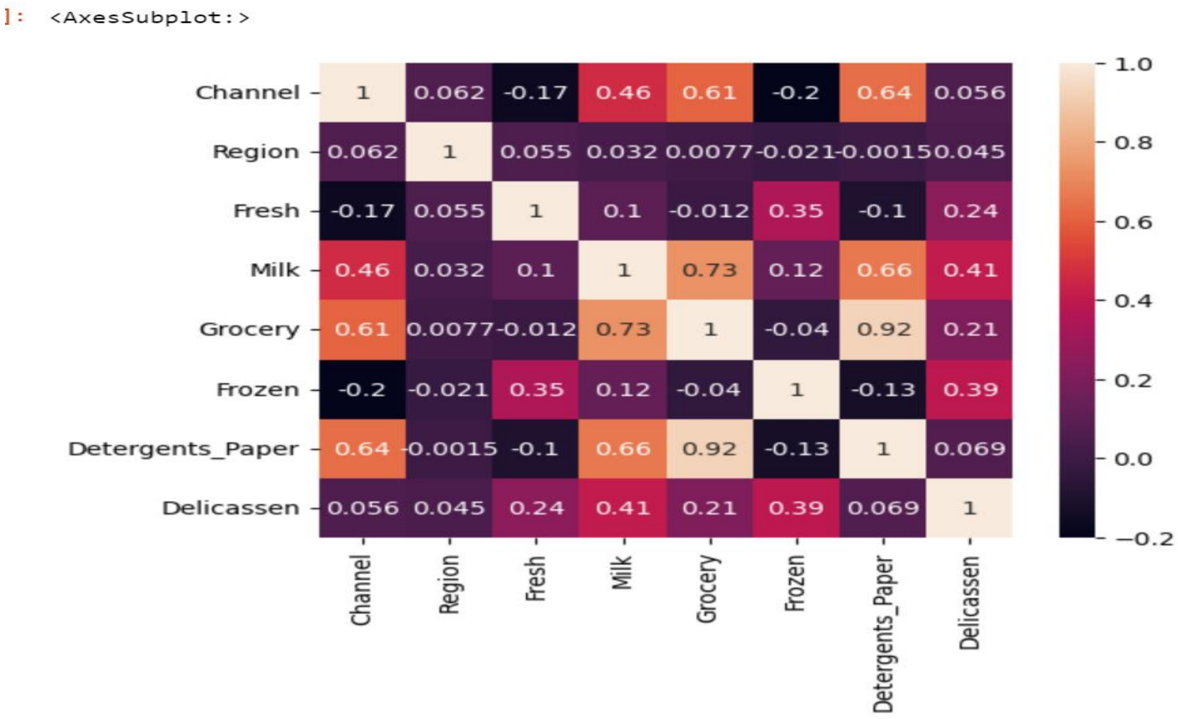
EDA :Exploratory Data Analysis



Insights : we can see channel 1 and 2 both are having max value in region 3 means with others regions

- -Insights
- We can see the channel 1 and channel 2 are in more in the Other region.

EDA :Exploratory Data Analysis



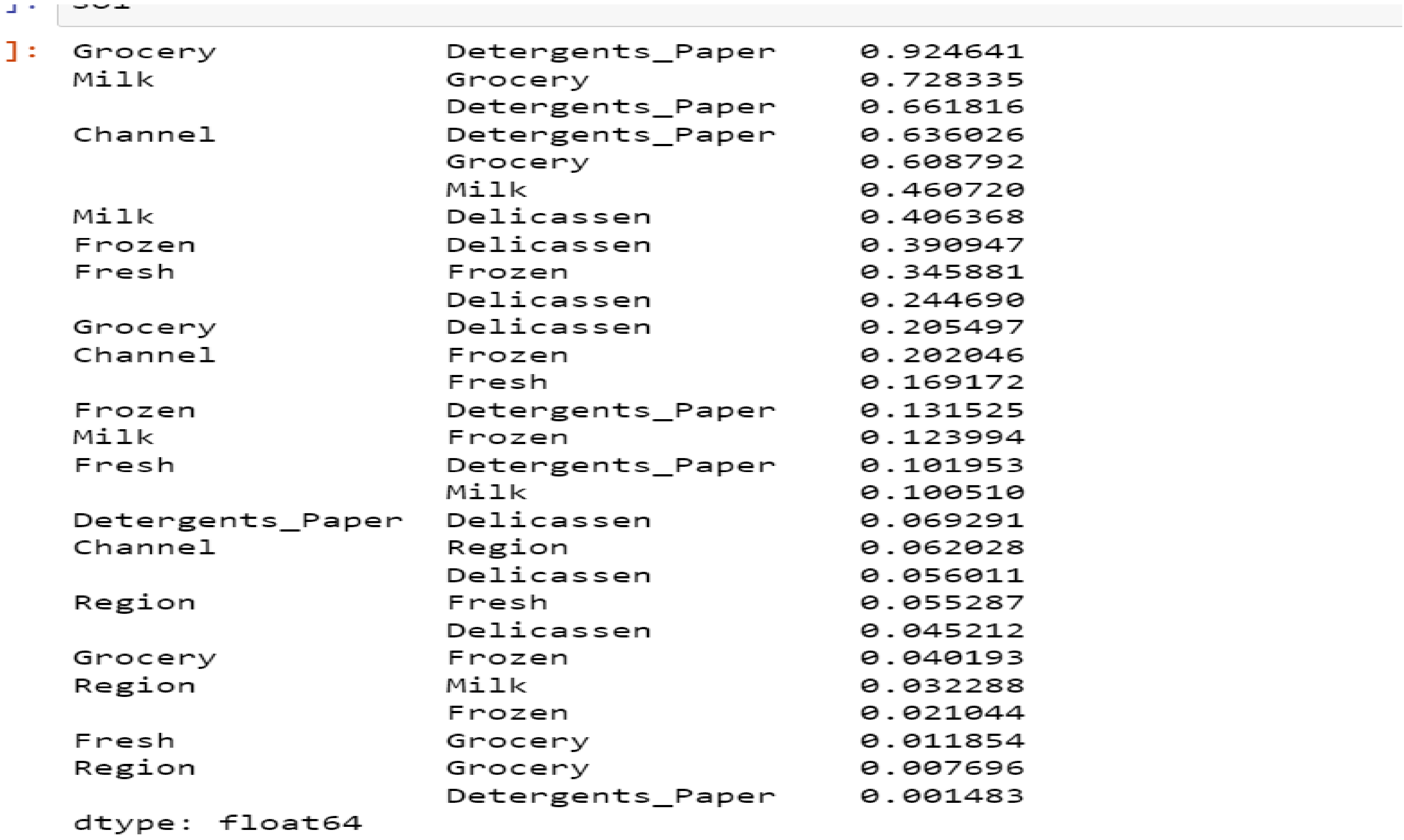
```
]: <AxesSubplot:>
```

	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
Channel	1	0.062	-0.17	0.46	0.61	-0.2	0.64	0.056
Region	0.062	1	0.055	0.032	0.0077	-0.021	-0.0015	0.045
Fresh	-0.17	0.055	1	0.1	-0.012	0.35	-0.1	0.24
Milk	0.46	0.032	0.1	1	0.73	0.12	0.66	0.41
Grocery	0.61	0.0077	-0.012	0.73	1	-0.04	0.92	0.21
Frozen	-0.2	-0.021	0.35	0.12	-0.04	1	-0.13	0.39
Detergents_Paper	0.64	-0.0015	-0.1	0.66	0.92	-0.13	1	0.069
Delicassen	0.056	0.045	0.24	0.41	0.21	0.39	0.069	1

Insights : Here we can see the correlation values with each features with another.

- Frozen and channel are more –vely correlated and Detergents_Paper and Grocery are highly correlated.

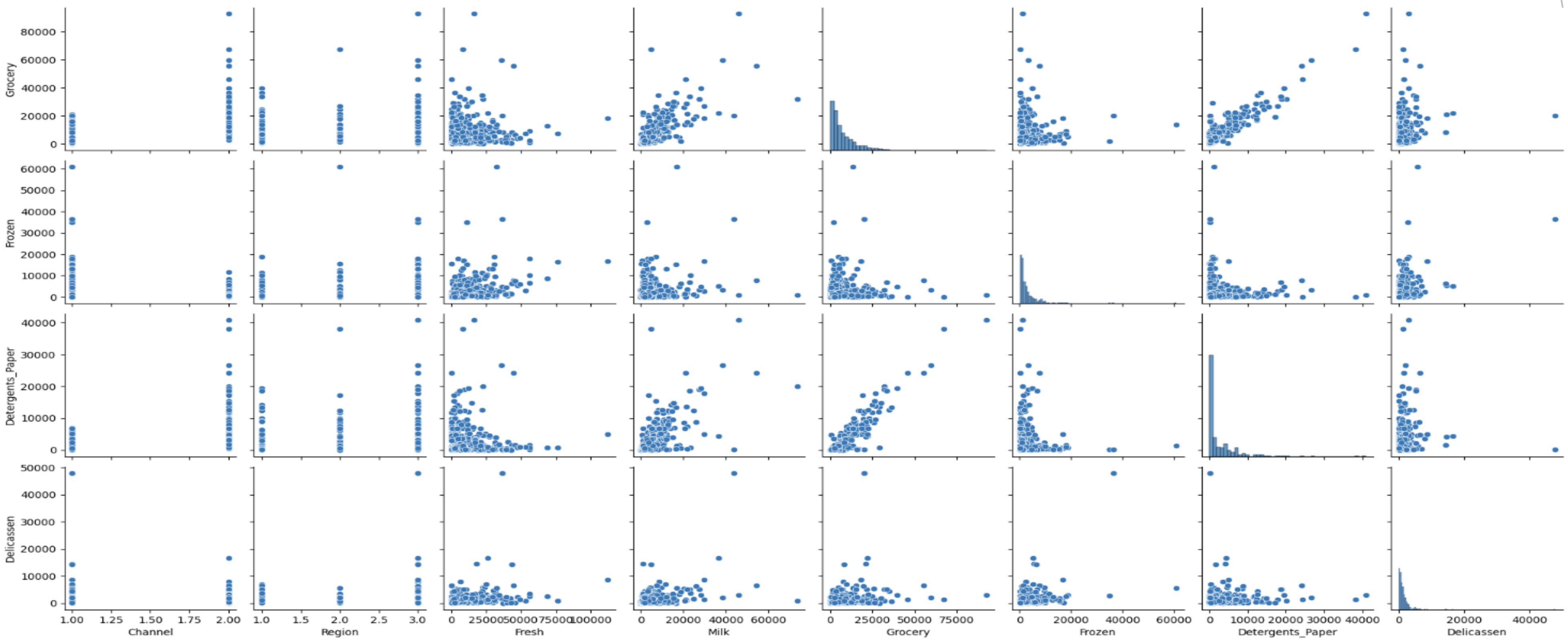
EDA :Exploratory Data Analysis



```
]: Grocery           Detergents_Paper    0.924641
   Milk              Grocery             0.728335
                     Detergents_Paper    0.661816
   Channel           Detergents_Paper    0.636026
                     Grocery             0.608792
                     Milk                0.460720
   Milk              Delicassen          0.406368
   Frozen            Delicassen          0.390947
   Fresh             Frozen              0.345881
                     Delicassen          0.244690
   Grocery           Delicassen          0.205497
   Channel           Frozen              0.202046
                     Fresh               0.169172
   Frozen            Detergents_Paper    0.131525
   Milk              Frozen              0.123994
   Fresh             Detergents_Paper    0.101953
                     Milk                0.100510
   Detergents_Paper  Delicassen          0.069291
   Channel           Region              0.062028
                     Delicassen          0.056011
   Region            Fresh               0.055287
                     Delicassen          0.045212
   Grocery           Frozen              0.040193
   Region            Milk                0.032288
                     Frozen              0.021044
   Fresh             Grocery             0.011854
   Region            Grocery             0.007696
                     Detergents_Paper    0.001483
   dtype: float64
```

- Here we can see the values of correlation by descending .

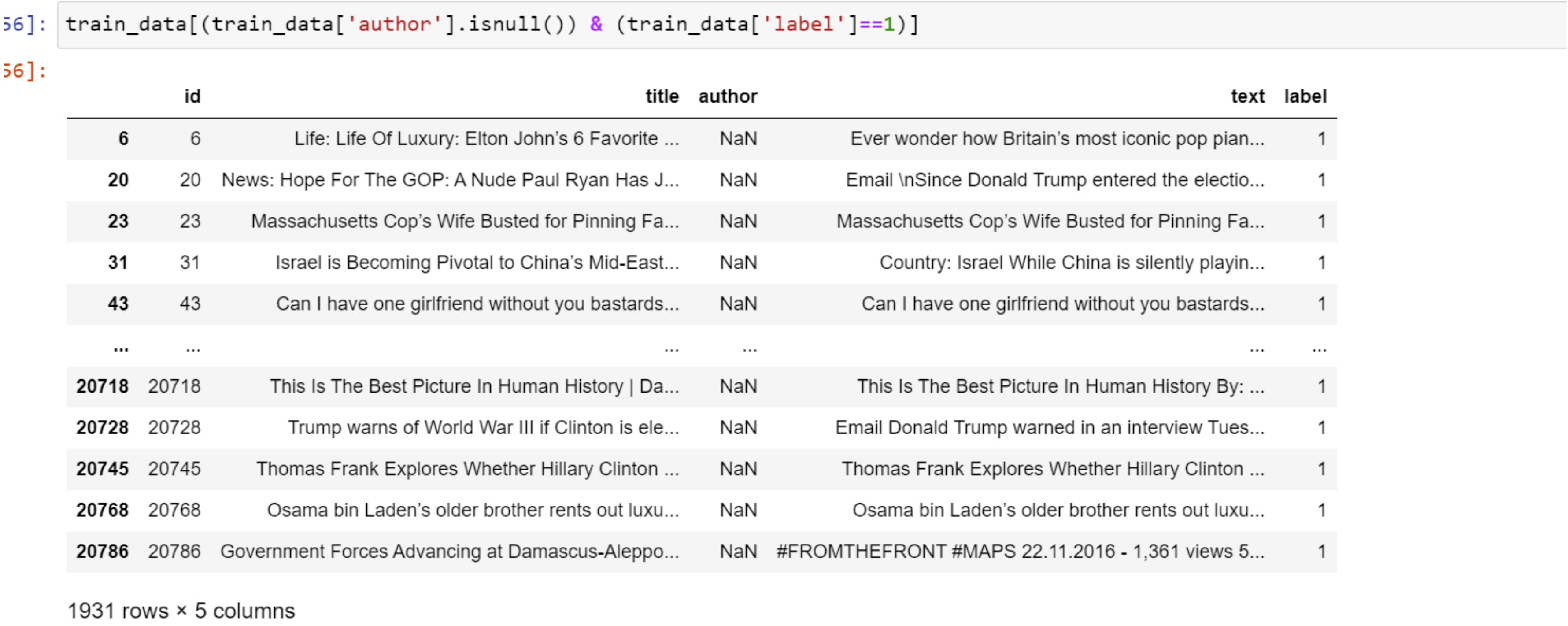
EDA :Exploratory Data Analysis



Insights :

Pairplot is one of the technique where we can see the correlation of dataset features with one another.

EDA :Exploratory Data Analysis



```
56]: train_data[(train_data['author'].isnull()) & (train_data['label']==1)]
```

56]:

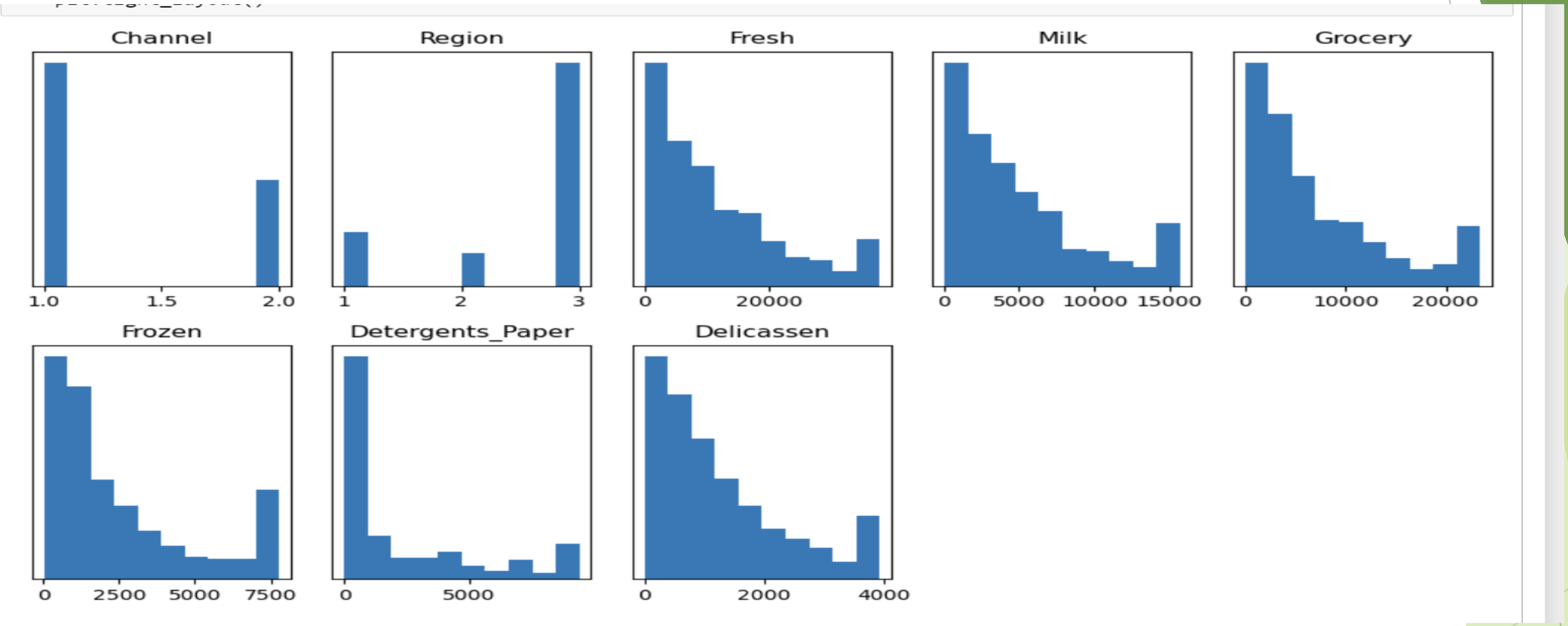
	id	title	author	text	label
6	6	Life: Life Of Luxury: Elton John’s 6 Favorite ...	NaN	Ever wonder how Britain’s most iconic pop pian...	1
20	20	News: Hope For The GOP: A Nude Paul Ryan Has J...	NaN	Email \nSince Donald Trump entered the electio...	1
23	23	Massachusetts Cop’s Wife Busted for Pinning Fa...	NaN	Massachusetts Cop’s Wife Busted for Pinning Fa...	1
31	31	Israel is Becoming Pivotal to China’s Mid-East...	NaN	Country: Israel While China is silently playin...	1
43	43	Can I have one girlfriend without you bastards...	NaN	Can I have one girlfriend without you bastards...	1
...	...	...	...	...	...
20718	20718	This Is The Best Picture In Human History | Da...	NaN	This Is The Best Picture In Human History By: ...	1
20728	20728	Trump warns of World War III if Clinton is ele...	NaN	Email Donald Trump warned in an interview Tues...	1
20745	20745	Thomas Frank Explores Whether Hillary Clinton ...	NaN	Thomas Frank Explores Whether Hillary Clinton ...	1
20768	20768	Osama bin Laden’s older brother rents out luxu...	NaN	Osama bin Laden’s older brother rents out luxu...	1
20786	20786	Government Forces Advancing at Damascus-Aleppo...	NaN	#FROMTHEFRONT #MAPS 22.11.2016 - 1,361 views 5...	1

1931 rows × 5 columns

- Out of 1957 we can see here 1931 values belong to label =1

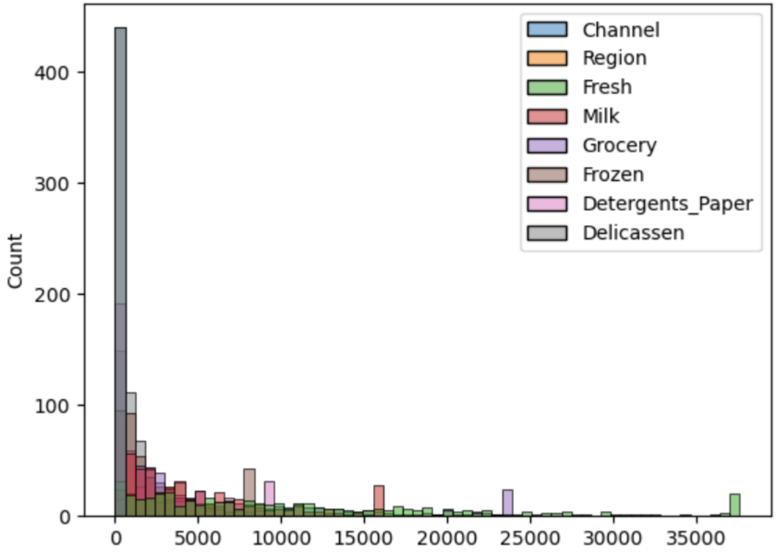
  means ‘Fake News’

EDA :Exploratory Data Analysis



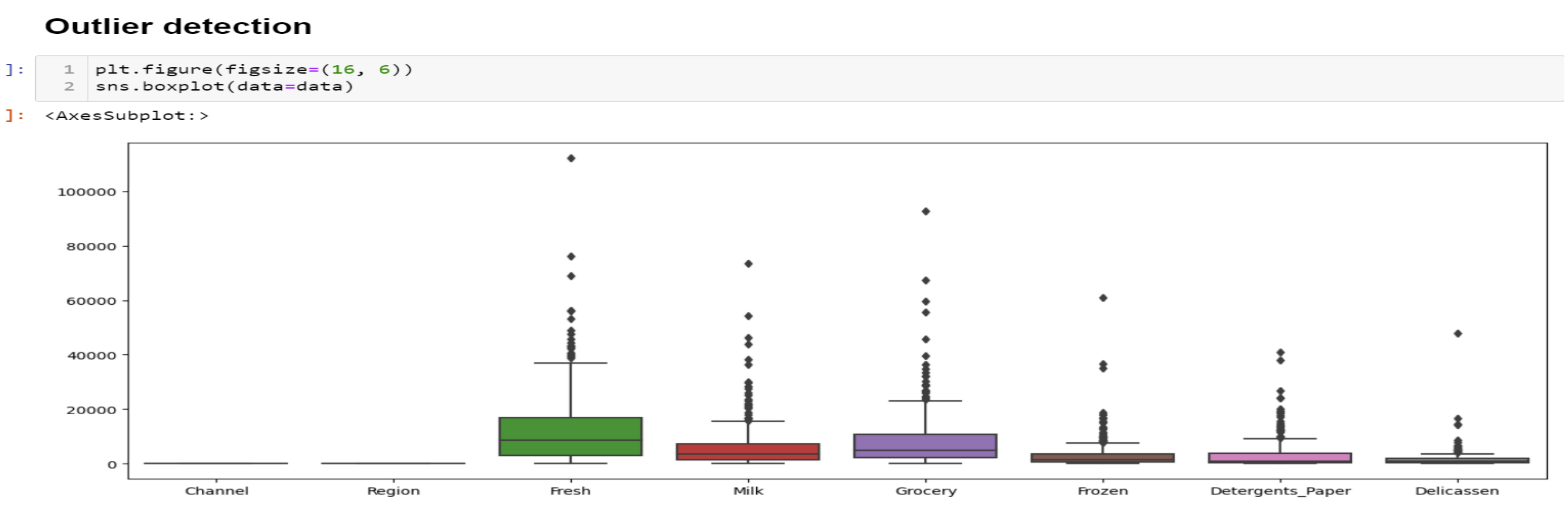
- We can see the data is mostly right skewed in the continuous features

EDA: Exploratory Data Analysis



Here we can see the distribution of the data ,mainly right skwed,we can keep it or transform in normal distribution which will impact on your model performance

EDA :Exploratory Data Analysis



Outlier detection

```
]:  plt.figure(figsize=(16, 6))
    sns.boxplot(data=data)
```

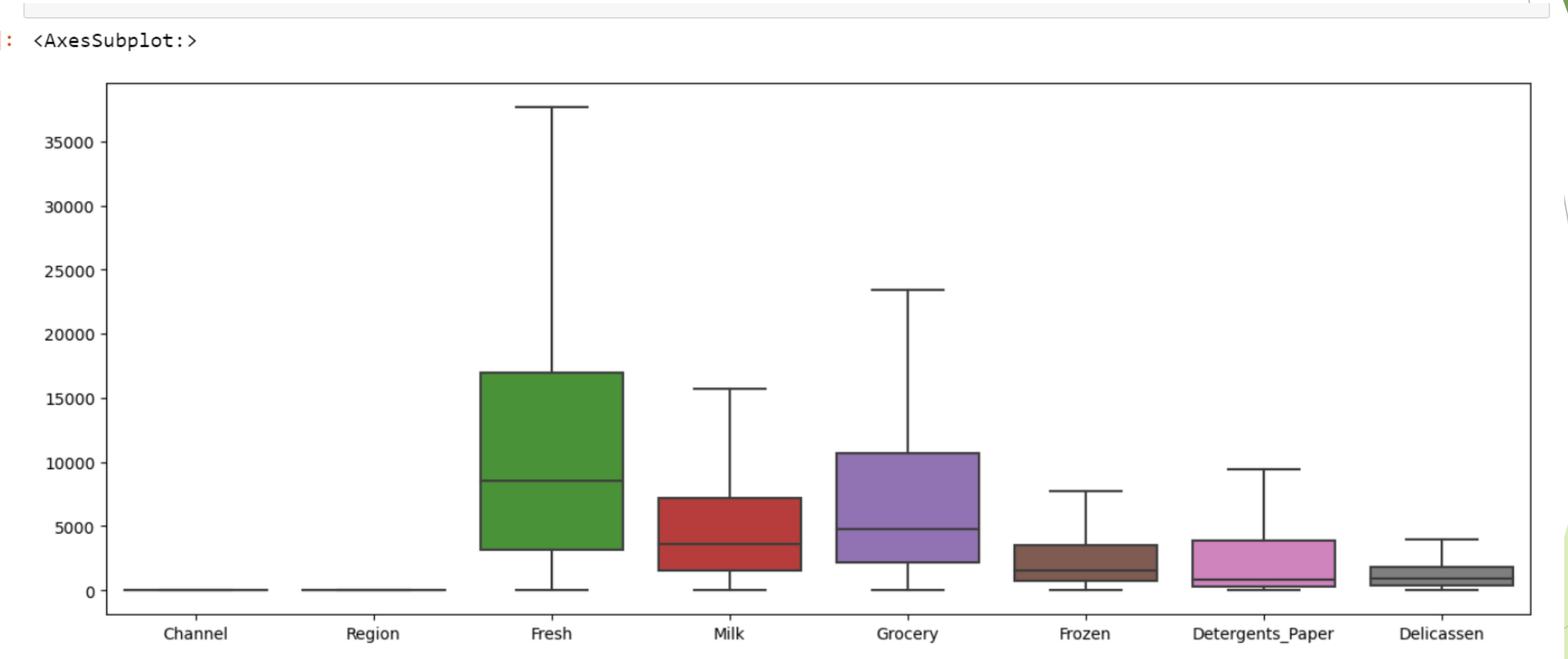
```
]:  <AxesSubplot:>
```

Insights :

Here we can see there are outliers in the given data which we need to handle.

FE : Feature Engineering

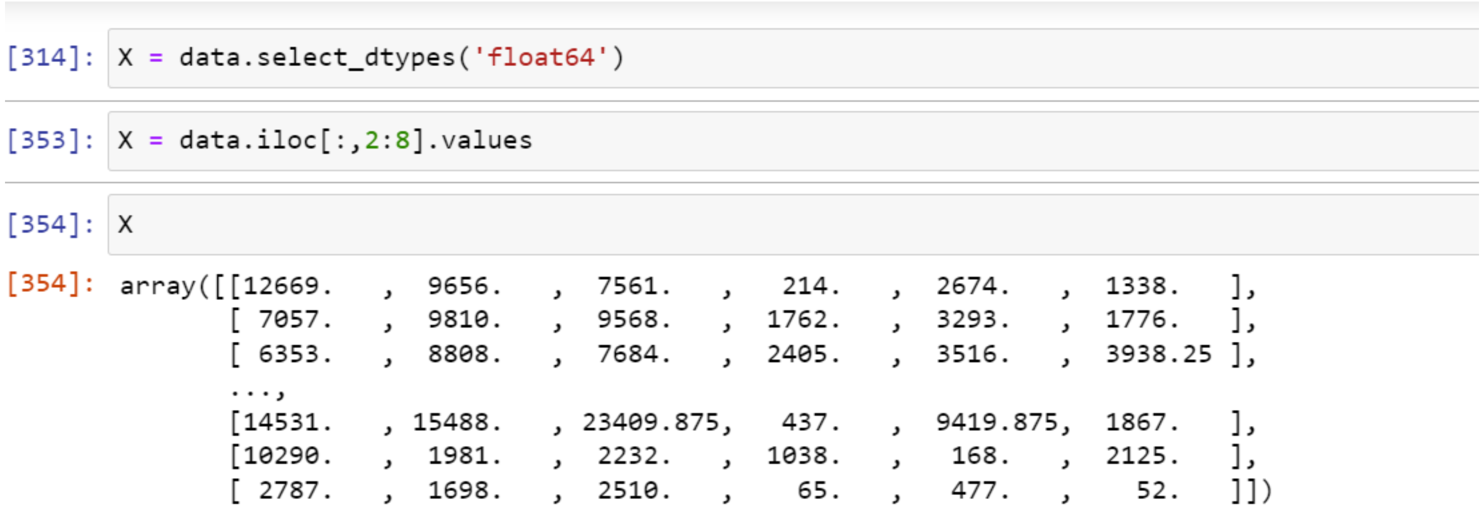
Feature Engineering: Outlier Treatment



```
: <AxesSubplot:>
```

- Outlier is handled

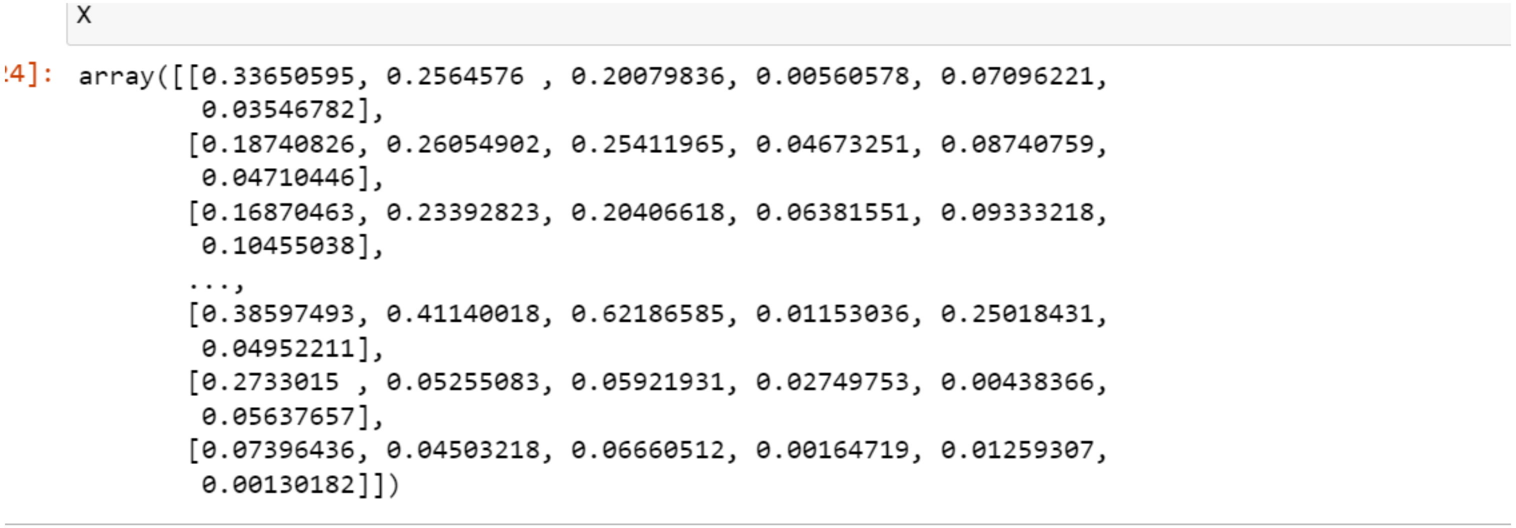
Feature Engineering :Data Transformation



```
[314]: X = data.select_dtypes('float64')

[353]: X = data.iloc[:,2:8].values

[354]: X

[354]: array([[12669.   ,  9656.   ,  7561.   ,   214.   ,  2674.   ,  1338.   ],
              [ 7057.   ,  9810.   ,  9568.   ,  1762.   ,  3293.   ,  1776.   ],
              [ 6353.   ,  8808.   ,  7684.   ,  2405.   ,  3516.   ,  3938.25 ],
              ...,
              [14531.   , 15488.   , 23409.875,   437.   ,  9419.875,  1867.   ],
              [10290.   ,  1981.   ,  2232.   ,  1038.   ,   168.   ,  2125.   ],
              [ 2787.   ,  1698.   ,  2510.   ,    65.   ,   477.   ,    52.   ]])
```

- As we need continuous numeric data for analysis so we separate out from the original dataset into array format.

Feature Engineering: Data Normalization



```
X
```

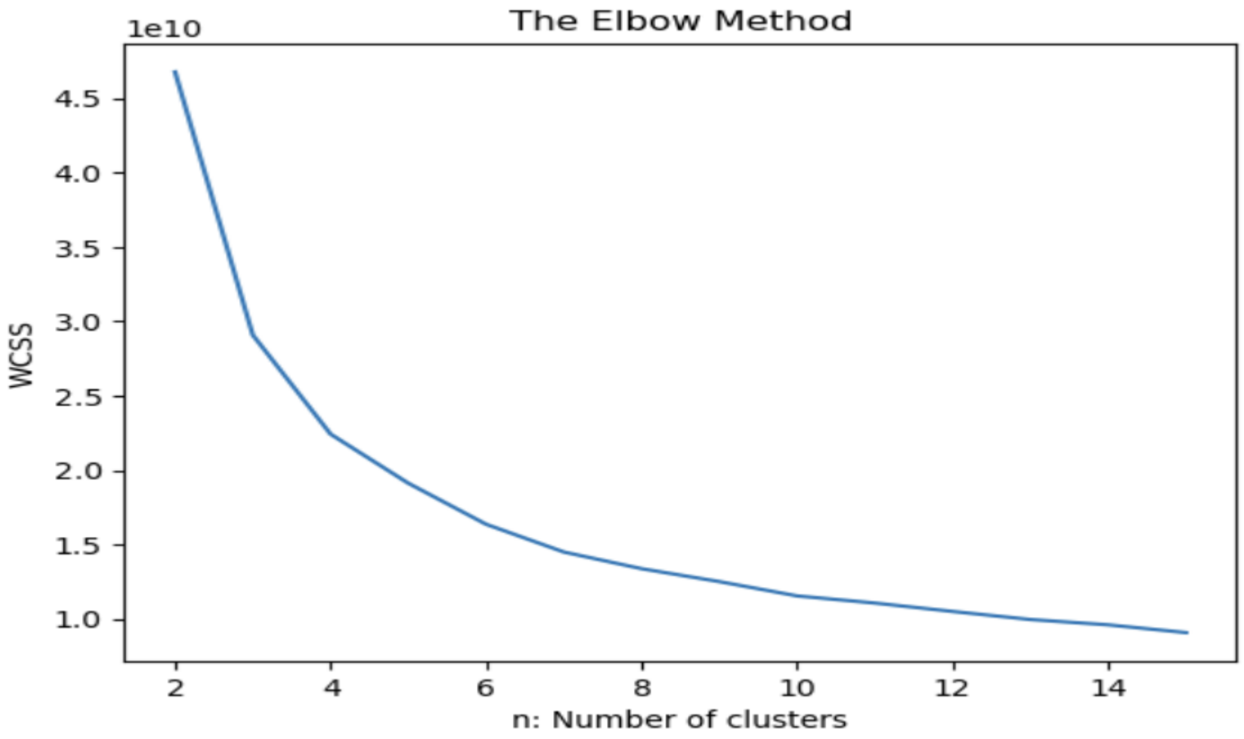
```
4]: array([[0.33650595, 0.2564576 , 0.20079836, 0.00560578, 0.07096221,
            0.03546782],
           [0.18740826, 0.26054902, 0.25411965, 0.04673251, 0.08740759,
            0.04710446],
           [0.16870463, 0.23392823, 0.20406618, 0.06381551, 0.09333218,
            0.10455038],
           ...,
           [0.38597493, 0.41140018, 0.62186585, 0.01153036, 0.25018431,
            0.04952211],
           [0.2733015 , 0.05255083, 0.05921931, 0.02749753, 0.00438366,
            0.05637657],
           [0.07396436, 0.04503218, 0.06660512, 0.00164719, 0.01259307,
            0.00130182]])
```

Data need to be on same scale for processing so normalization is applied.

Machine Learning and Analysis

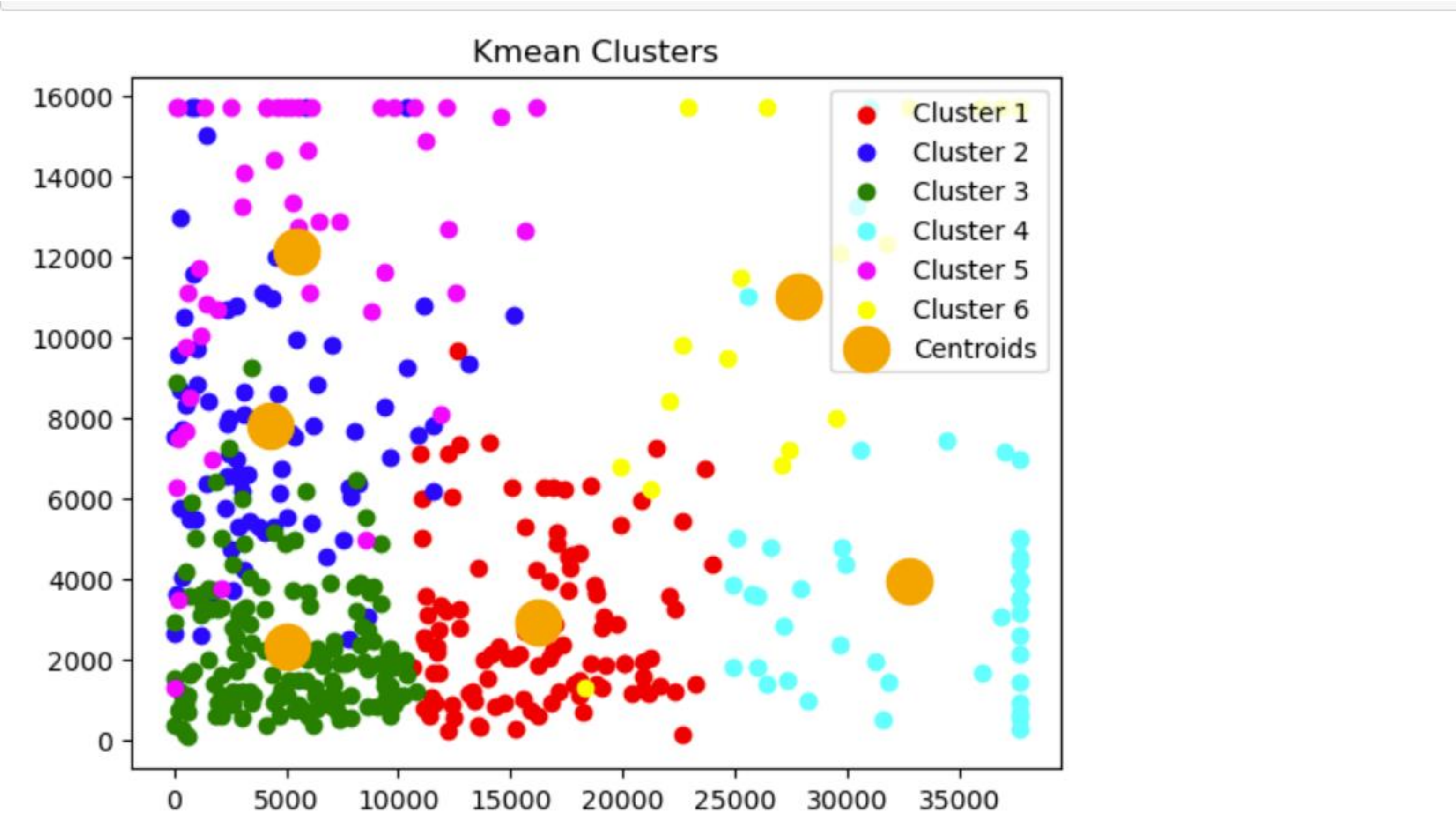
- In this following slides we will apply the four different clustering methods k-mean, Agglomerative Hierarchical clustering ,DBSCAN, Meanshift in team of finding appropriate cluster .
- Here we can see data is distributed so finding cluster is very tedious job and each algorithm have their own idea about clustering the data.

Model 1 : K-Mean Clustering

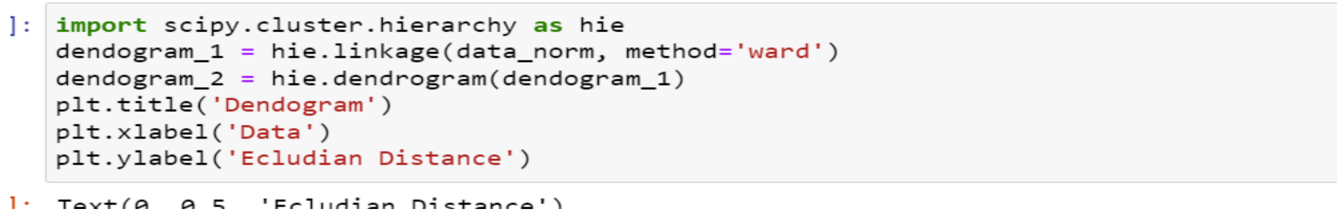


- Here we can apply K-Mean clustering and with the help of Elbow method we can cluster the data.
- Here we can see 6 will be the approximate number we can get for clustering
- We can also use WCSS method to get correct value of K

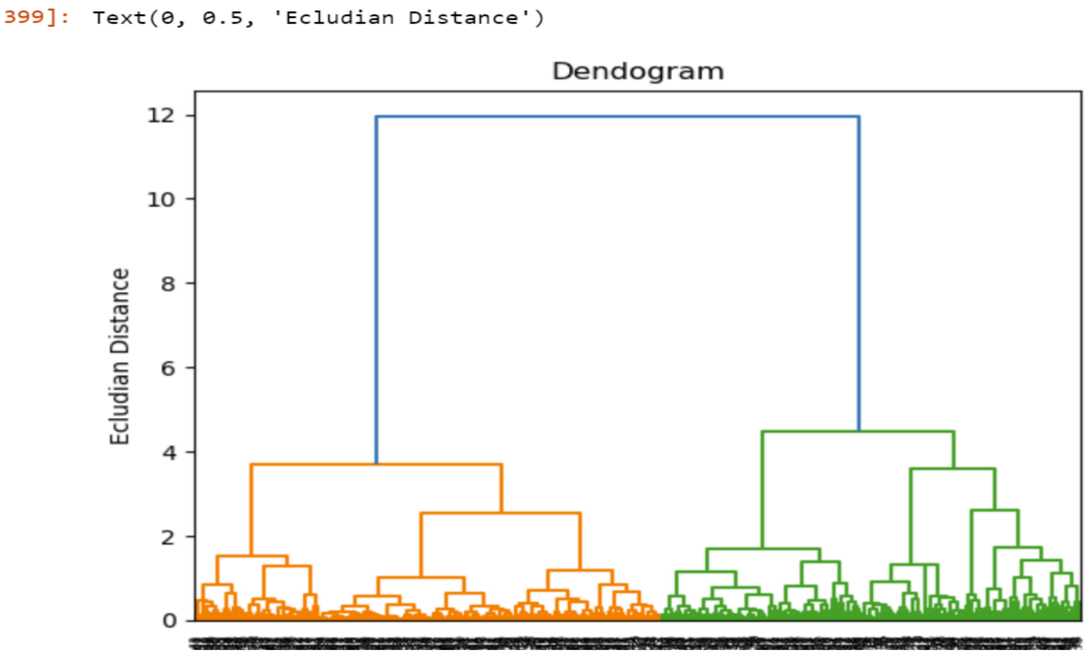
Model 1 : Evaluation K Mean



Model 2 :Hierarchical Clustering



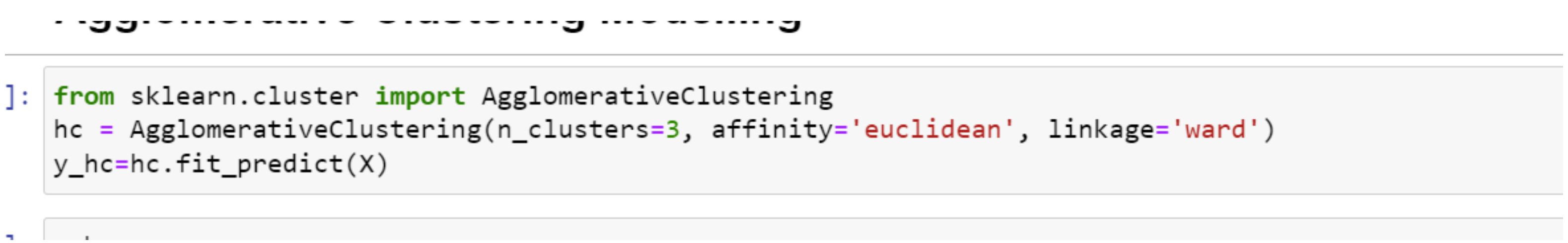
```
]: import scipy.cluster.hierarchy as hie
   dendogram_1 = hie.linkage(data_norm, method='ward')
   dendogram_2 = hie.dendrogram(dendogram_1)
   plt.title('Dendogram')
   plt.xlabel('Data')
   plt.ylabel('Ecludian Distance')

]: Text(0, 0.5, 'Ecludian Distance')
```



```
399]: Text(0, 0.5, 'Ecludian Distance')
```

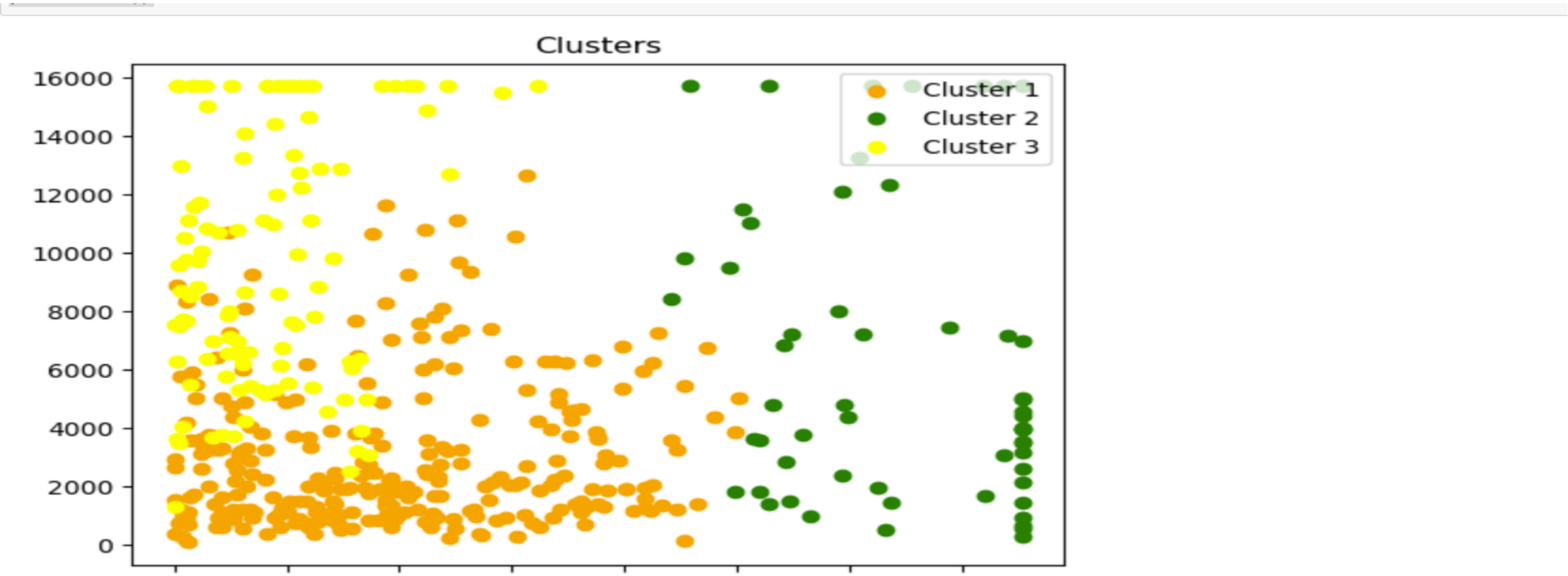
- Here we can see we form so many cluster but here by using distance we can get new cluster.

Model 2 : Visualization of cluster-Agglomerative Clustering

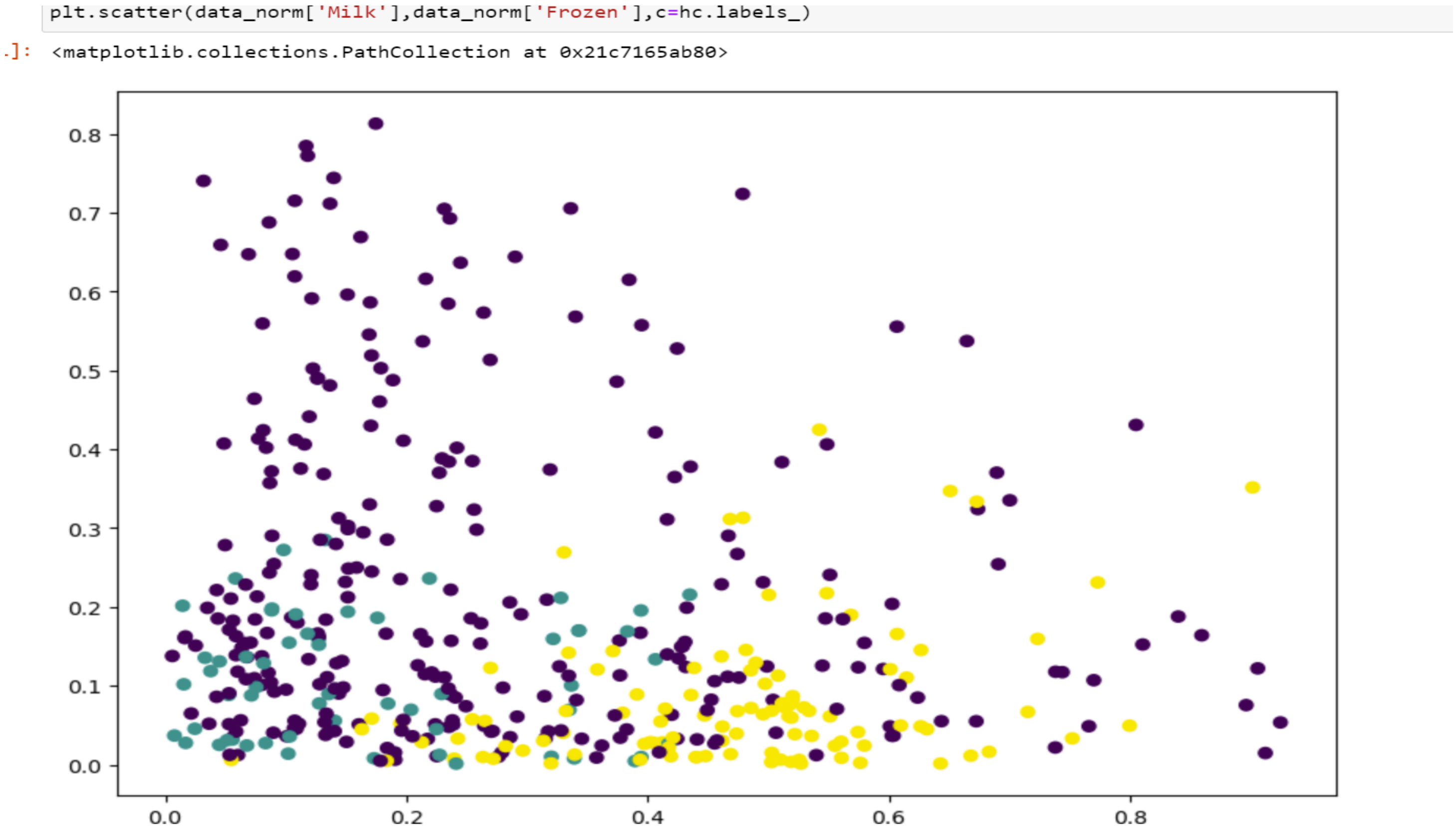


Agglomerative Clustering Modelling

```
]: from sklearn.cluster import AgglomerativeClustering
   hc = AgglomerativeClustering(n_clusters=3, affinity='euclidean', linkage='ward')
   y_hc=hc.fit_predict(X)
```



Model 2 : Agglomerative Clustering Visualization

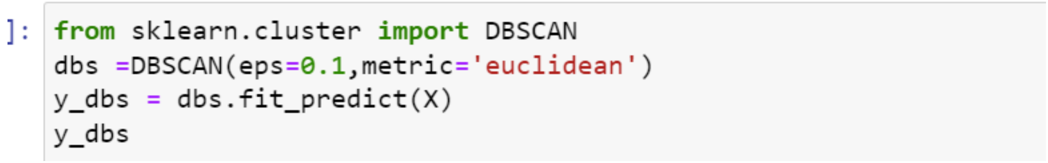


```
plt.scatter(data_norm['Milk'],data_norm['Frozen'],c=hc.labels_)
```

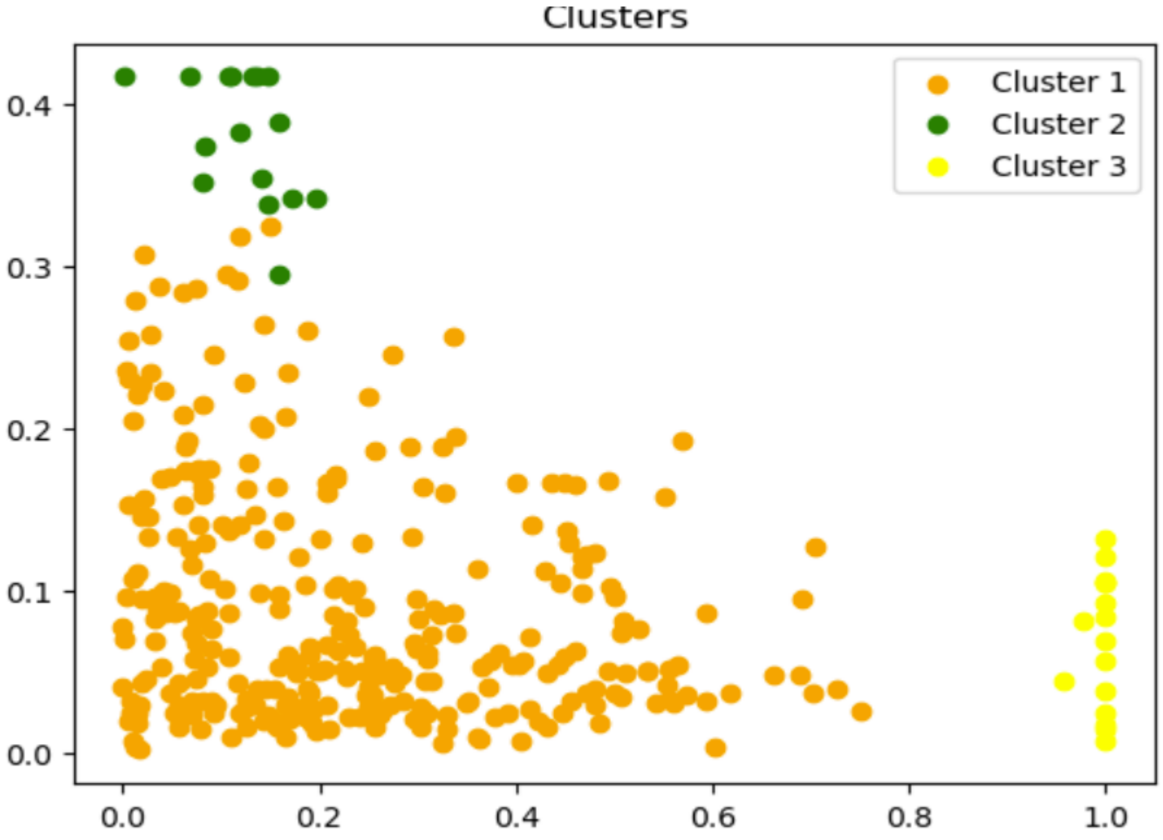
```
.]: <matplotlib.collections.PathCollection at 0x21c7165ab80>
```

- ➢ Here we can see the how values are performed in specific cluster here with help of agglomerative clustering we can see how data of feature milk and frozen get scattered in different clusters

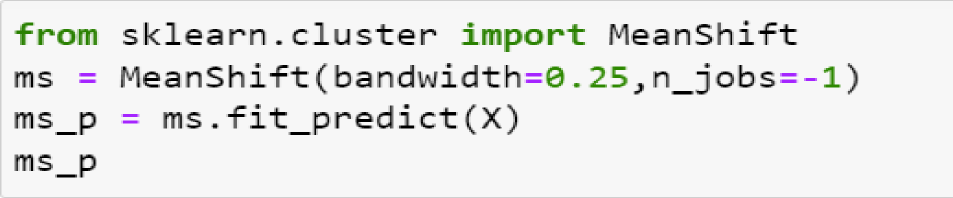
Model 3 :DBSCAN Algorithm



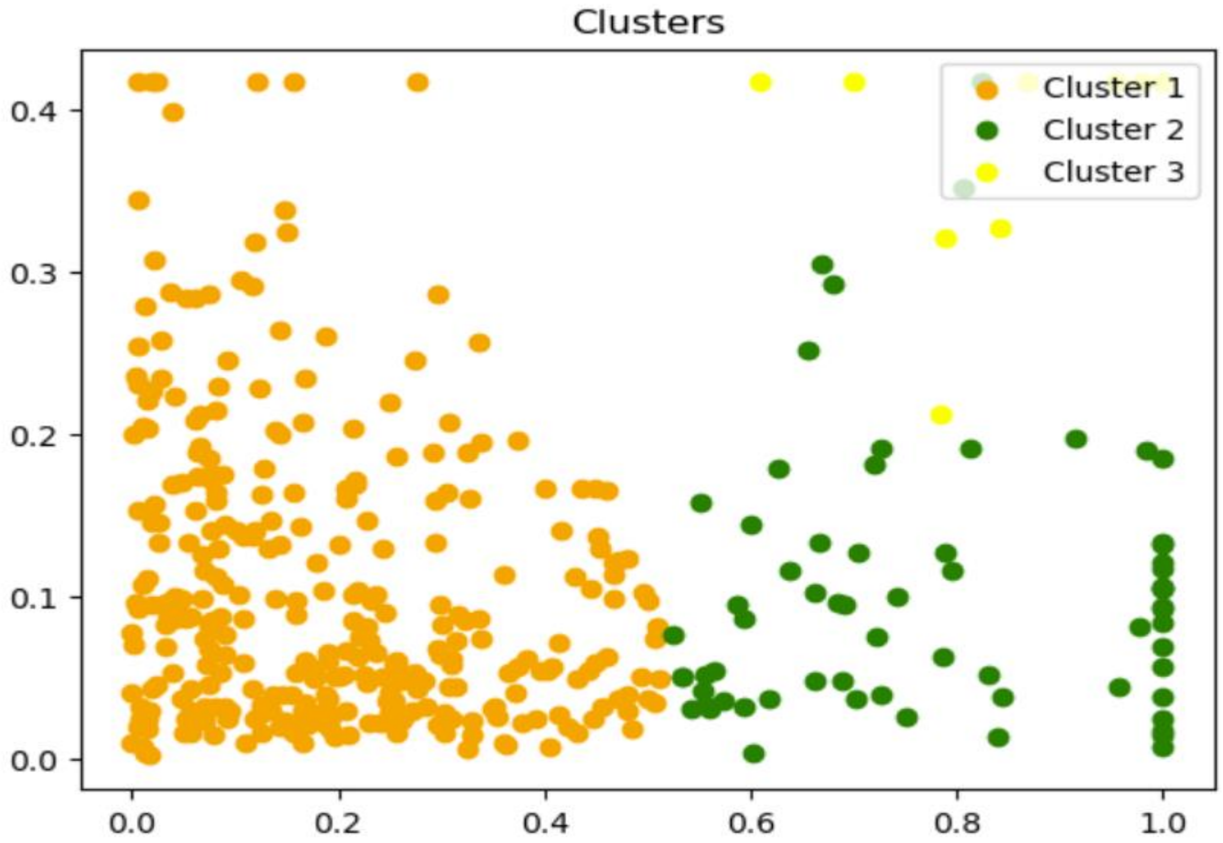
```
]: from sklearn.cluster import DBSCAN
   dbs =DBSCAN(eps=0.1,metric='euclidean')
   y_dbs = dbs.fit_predict(X)
   y_dbs
```



Model 4 :Mean Shift Algorithm



```
from sklearn.cluster import MeanShift
ms = MeanShift(bandwidth=0.25,n_jobs=-1)
ms_p = ms.fit_predict(X)
ms_p
```



Model Comparison, Model Flaws ,Advanced Step and further Suggestion

- **Here is the analysis**
- Here we can see we have 3 region and 2 channel we can use channel as output parameter to check the result as in region we can see there is region others which means we can get again other clusters in that region.
- Here we can see by using K mean clustering we can get 6 as cluster number by using given data and other clustering algorithm behave differently with respect to there bandwidth and distance criteria.
- As we can see DBSCAN is giving good result w.r.t to clustering.
- Mean shift Algorithm is working well if we minimized the bandwidth then only its works fine or else it combine all datapoints in single clusters.
- Here we can do so many permutation combination of different parameters of each of algorithms and check the accuracy of the dataset and then we can finalized the correct clustering algorithm
- - You can fine the code at : https://github.com/sadhanajarag/IBM-Certificate-Unsupervised-Machine-Algorithms/blob/main/unsupervised%20Machine%20Algo.ipynb

Thank You!!!!

Unsupervised Machine Learning

By Sadhana Jarag