# Data Mining Project

## Clustering Marketing Data

Dilara KARAKAŞ

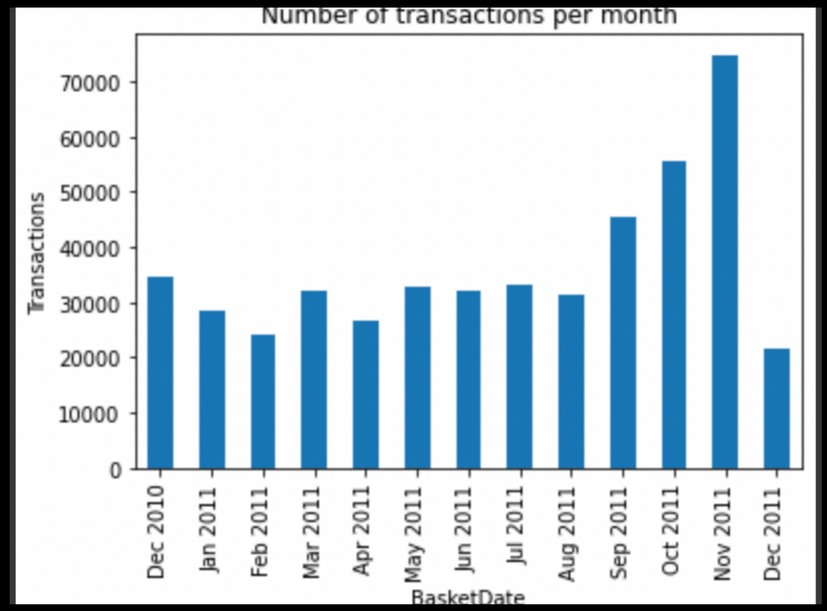
# Dimensional

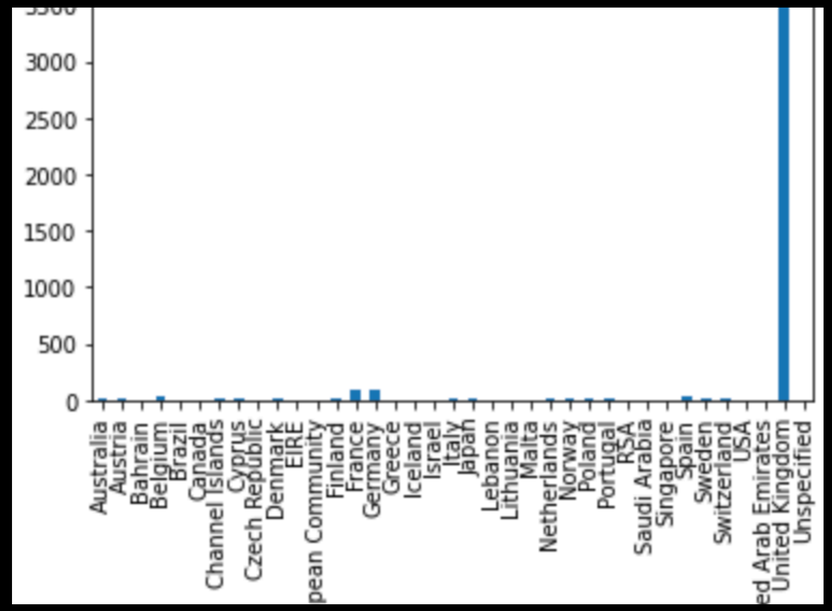
| Attribute | Type | Description |
|---|---|---|
| **BasketID** | `Numerical` | Unique alphanumeric ID assigned to a *basket*, defined as a set of transactions. |
| **BasketDate** | `Numerical` | The date on which the transaction was made. |
| **Sale** | `Numerical` | Cost of one unit of product. |
| **CustomerID** | `Numerical` | The ID of the customer. |
| **CustomerCountry** | `Categorical` | The Customer's country of origin. |
| **ProdID** | `Numerical` | The ID of the product. |
| **ProdDescr** | `Categorical` | A description of the product. |
| **Qta** | `Numerical` | The number of units bought in the transaction. |

# Data Understanding

- Data Semantics
    1. Count Null values
        - 65080 CustomerID, 753 ProdDescr
    2. Delete some transaction
        - Null Sale (1279) & Negative Qta (8095)
        - Null values
- Data distributions and statistics

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Index | BasketID | BasketDate | Sale | CustomerID | CustomerCountry | ProdID | ProdDescr | Qta |
| 2 | 0 | 536365 | 01/12/2010 08:26 | 2,55 | 178500 | United Kingdom | 85123A | WHITE HANGING HEART T-LIGHT HOLDER | 6 |
| 3 | 1 | 536365 | 01/12/2010 08:26 | 3,39 | 178500 | United Kingdom | 71053 | WHITE METAL LANTERN | 6 |
| 4 | 2 | 536365 | 01/12/2010 08:26 | 2,75 | 178500 | United Kingdom | 84406B | CREAM CUPID HEARTS COAT HANGER | 8 |
| 5 | 3 | 536365 | 01/12/2010 08:26 | 3,39 | 178500 | United Kingdom | 84029G | KNITTED UNION FLAG HOT WATER BOTTLE | 6 |
| 6 | 4 | 536365 | 01/12/2010 08:26 | 3,39 | 178500 | United Kingdom | 84029E | RED WOOLLY HOTTIE WHITE HEART. | 6 |
| 7 | 5 | 536365 | 01/12/2010 08:26 | 7,65 | 178500 | United Kingdom | 22752 | SET 7 BABUSHKA NESTING BOXES | 2 |
| 8 | 6 | 536365 | 01/12/2010 08:26 | 4,25 | 178500 | United Kingdom | 21730 | GLASS STAR FROSTED T-LIGHT HOLDER | 6 |



# Extra attributes

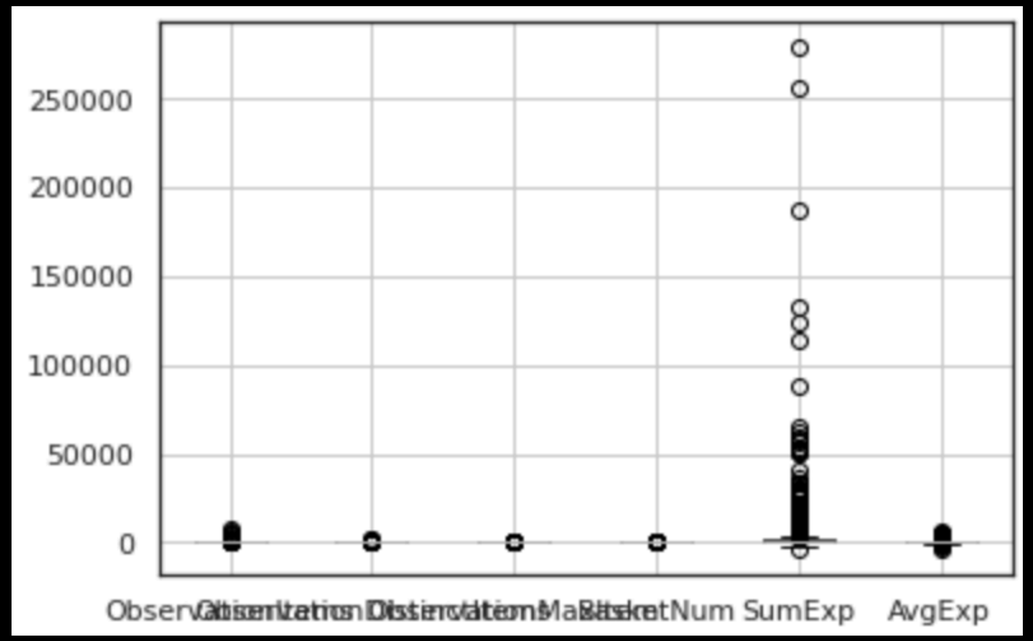
| | | |
|---|---|---|
| Total Sale | Numerical | The amount of the transaction, calculated by multiplying the Sale by the Qta. |

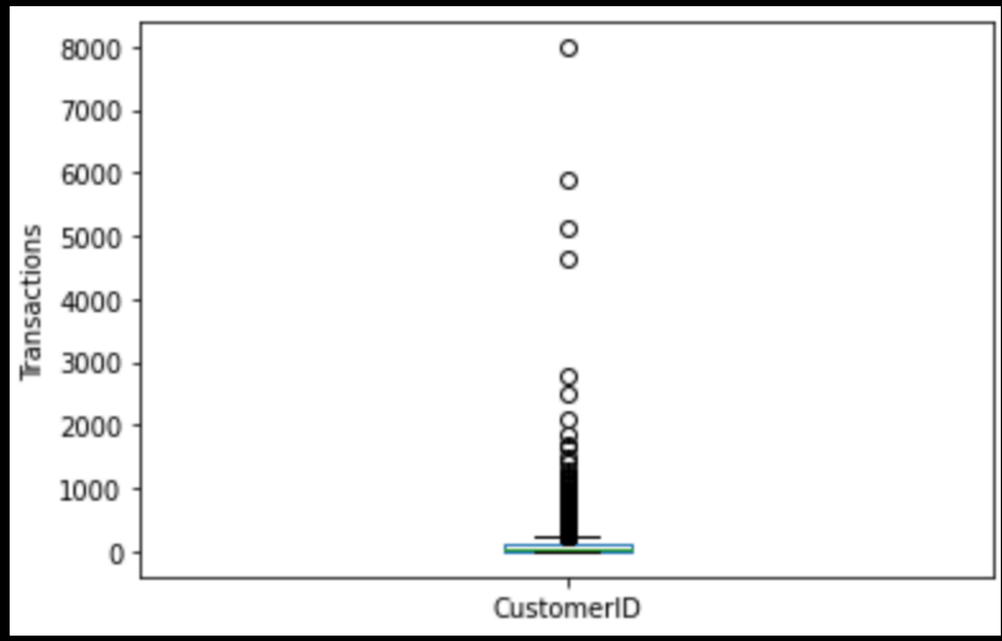
# New features

```
[ ] dataset_frame_customer.describe()
```

| | ObservationItems | ObservationDistinctItems | ObservationMaxItem | BasketNum | SumExp | AvgExp |
|---|---|---|---|---|---|---|
| count | 4372.000000 | 4372.000000 | 4372.000000 | 4372.000000 | 4372.000000 | 4372.000000 |
| mean | 93.053522 | 61.211116 | 31.867795 | 5.075480 | 1898.463818 | 315.884437 |
| std | 232.471568 | 85.425119 | 31.225805 | 9.338754 | 8219.344627 | 361.237024 |
| min | 1.000000 | 1.000000 | 1.000000 | 1.000000 | -4287.630000 | -4287.630000 |
| 25% | 17.000000 | 15.000000 | 12.000000 | 1.000000 | 293.362500 | 151.991250 |
| 50% | 42.000000 | 35.000000 | 23.000000 | 3.000000 | 648.075000 | 236.987500 |
| 75% | 102.000000 | 77.000000 | 41.000000 | 5.000000 | 1611.725000 | 370.816071 |
| max | 7983.000000 | 1794.000000 | 542.000000 | 248.000000 | 279489.020000 | 6207.670000 |

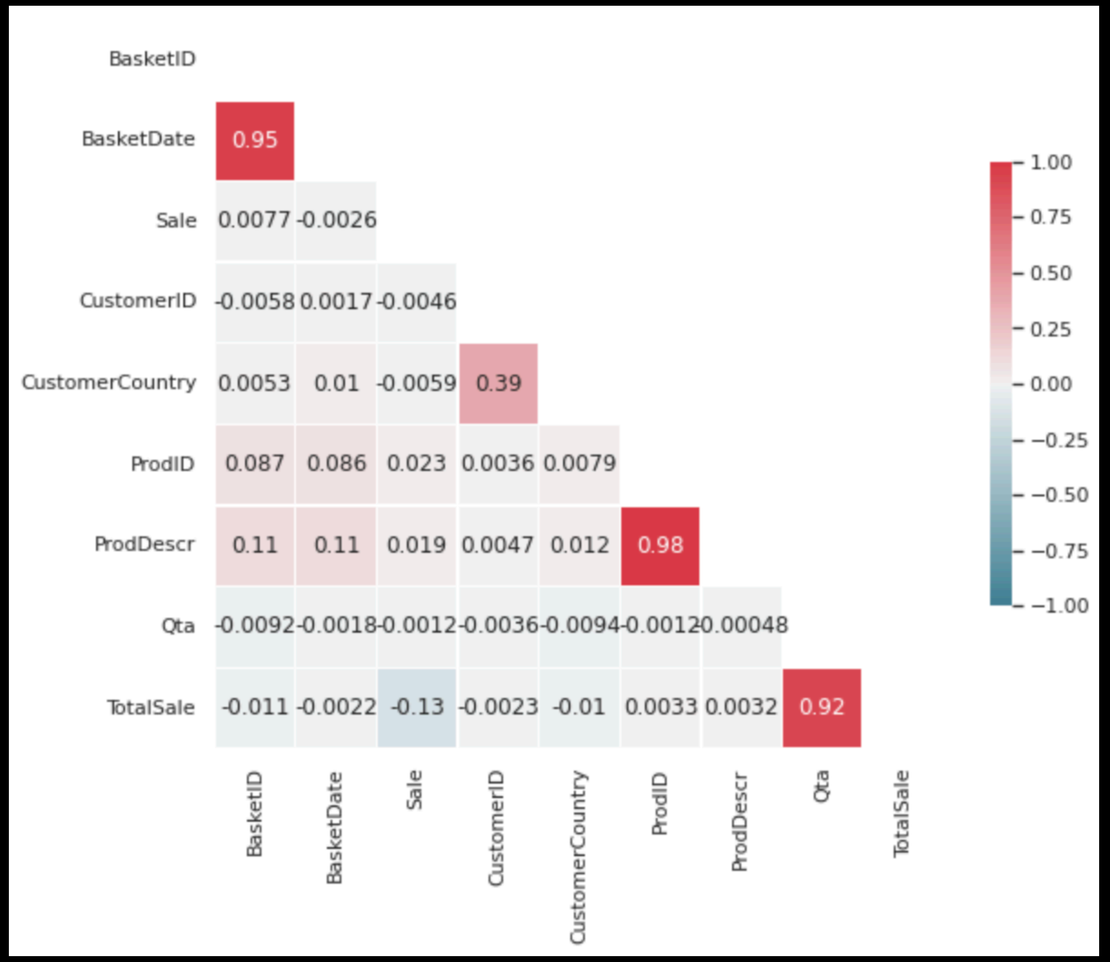
# Outliers

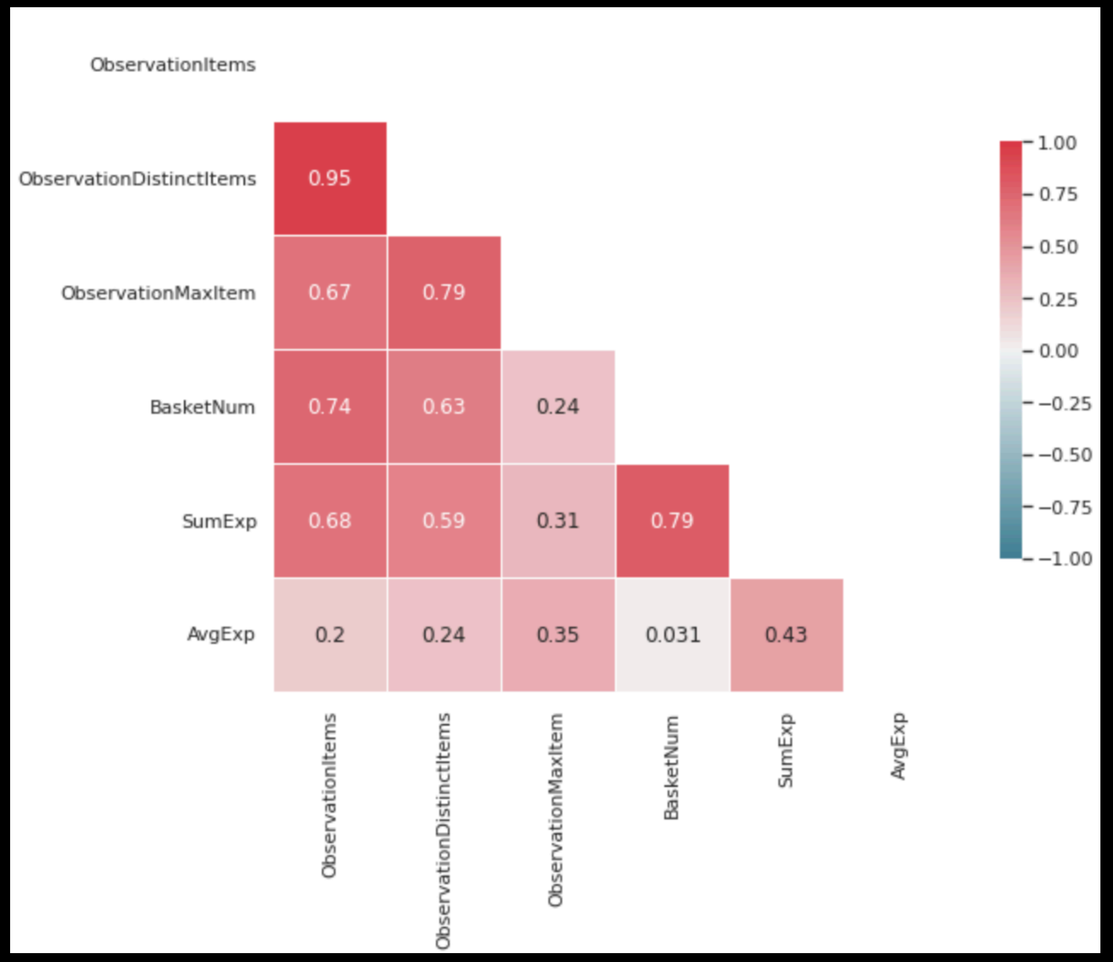



- Z Score > 3 -> Empirical Rule

# Correlation

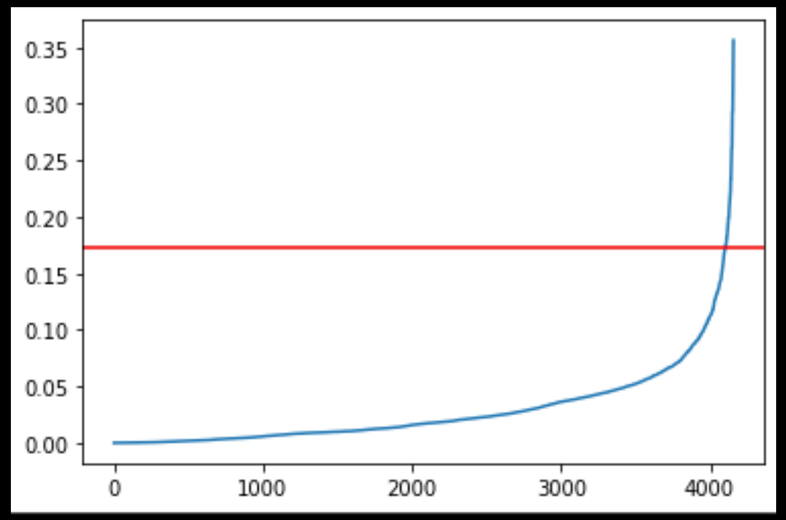



- **Pairwise Correlation**

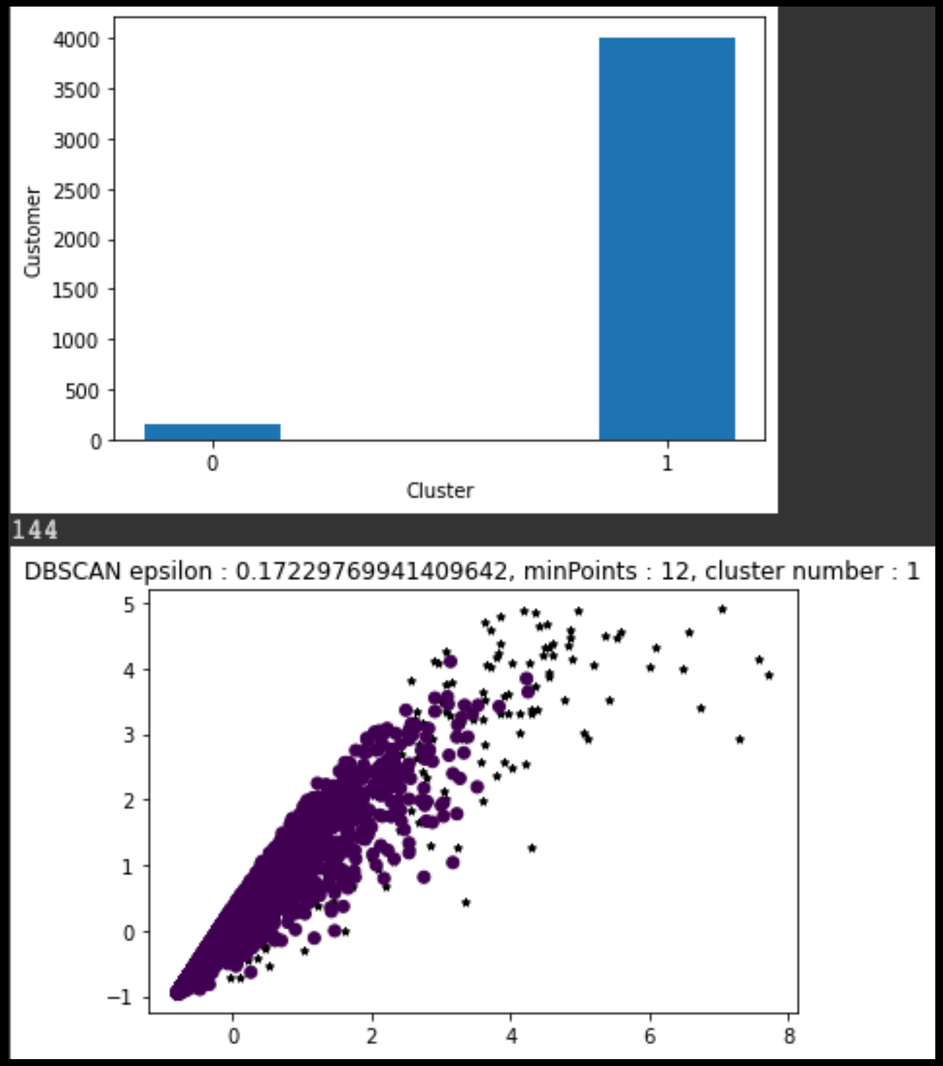
# Clustering Analysis

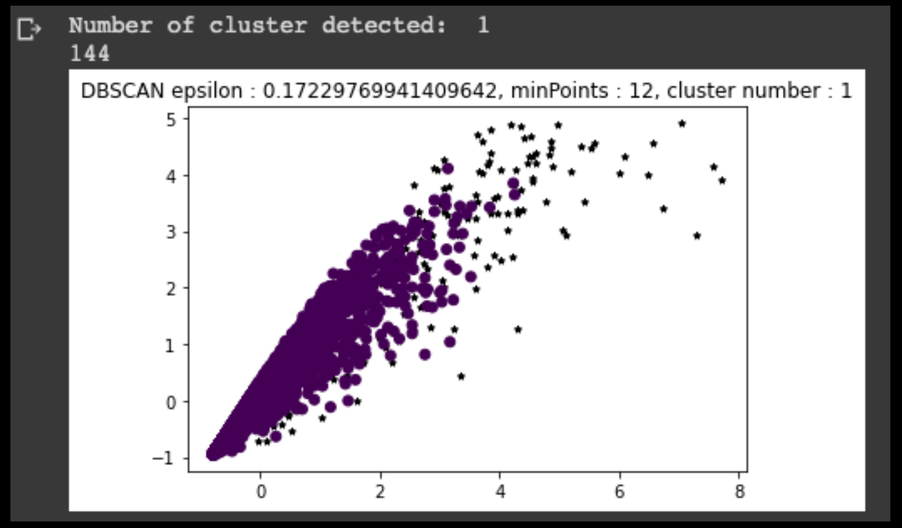
- DBSCAN
  - min_points: 2 * dataset dimensionality
  - Epsilon: Elbow Method

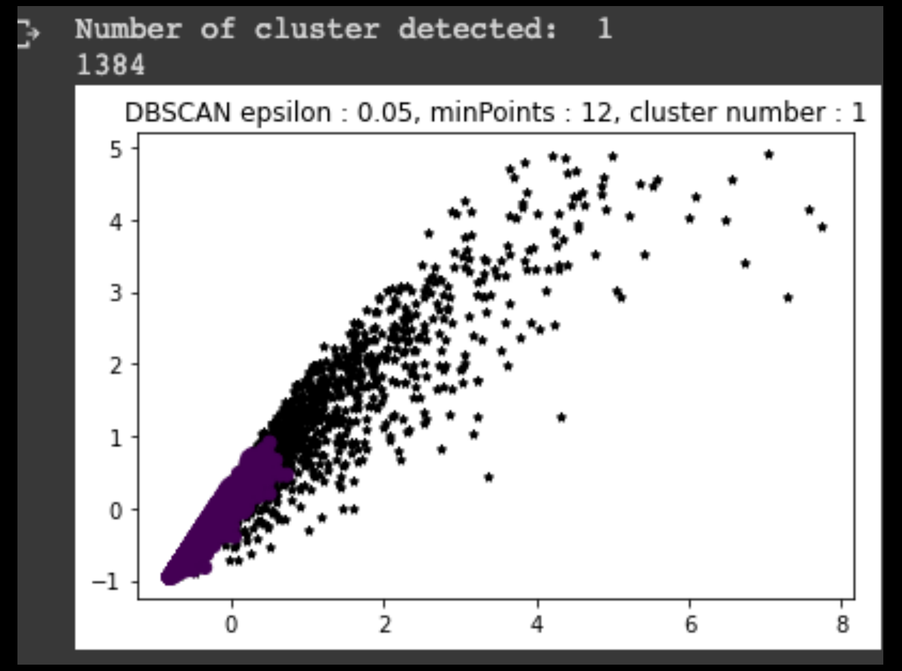



- Hierarchical
  - Average
  - Single
  - Complete
  - Ward

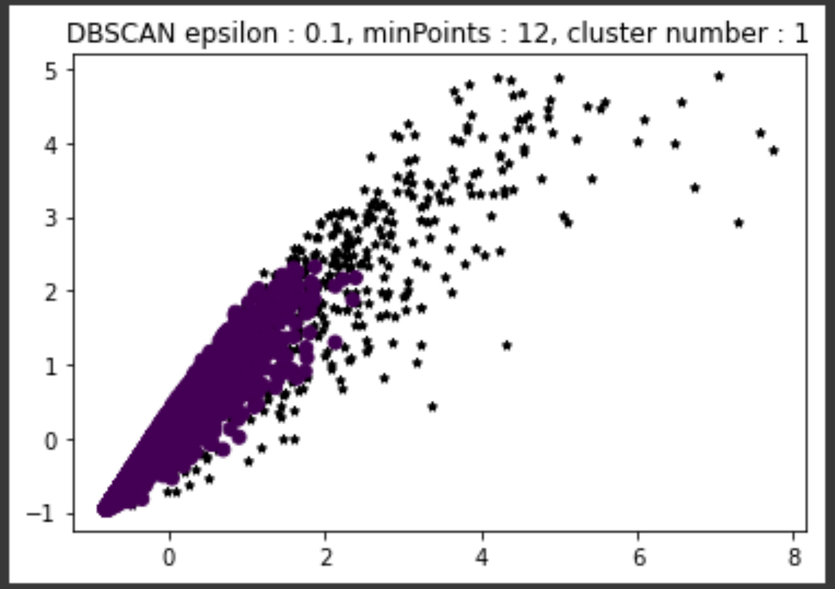
# DBSCAN Test Result



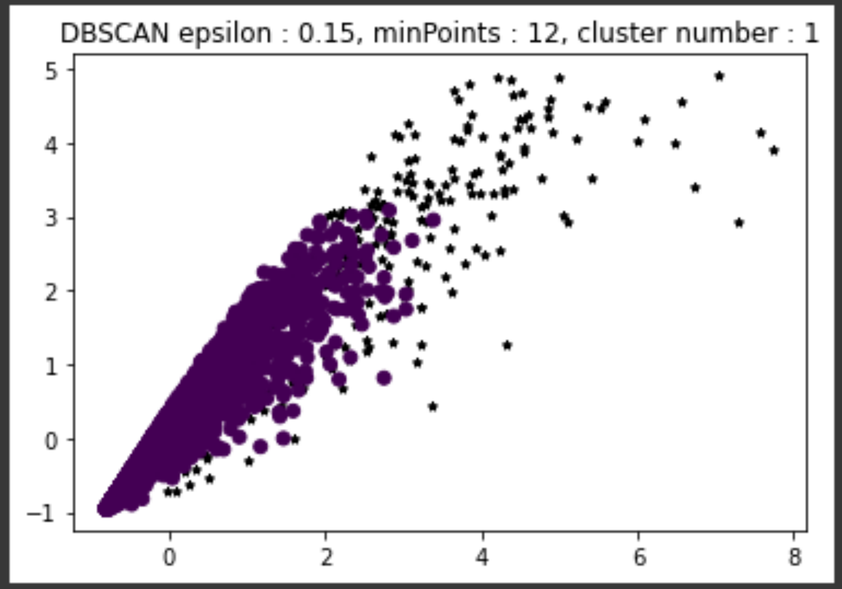



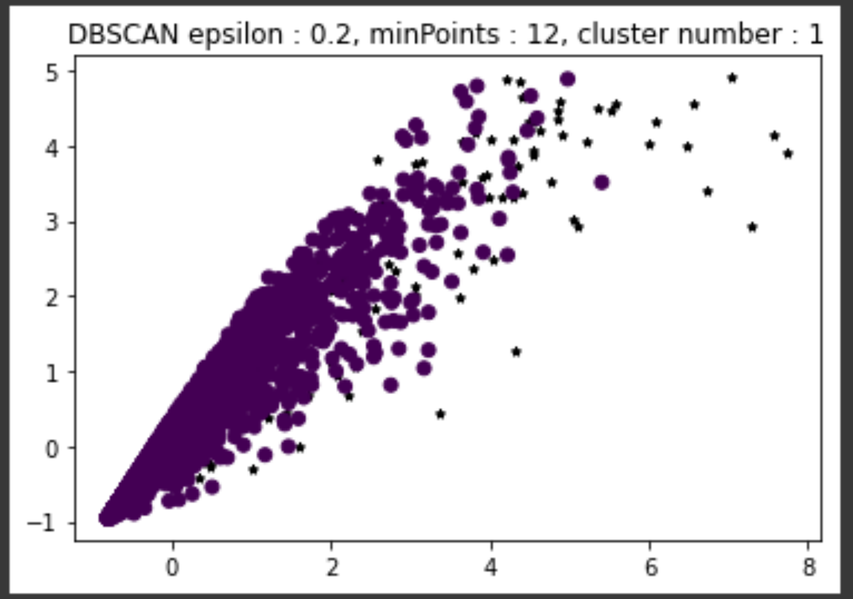
```
Number of cluster detected:  1
471
```




```
Number of cluster detected:  1
220
```

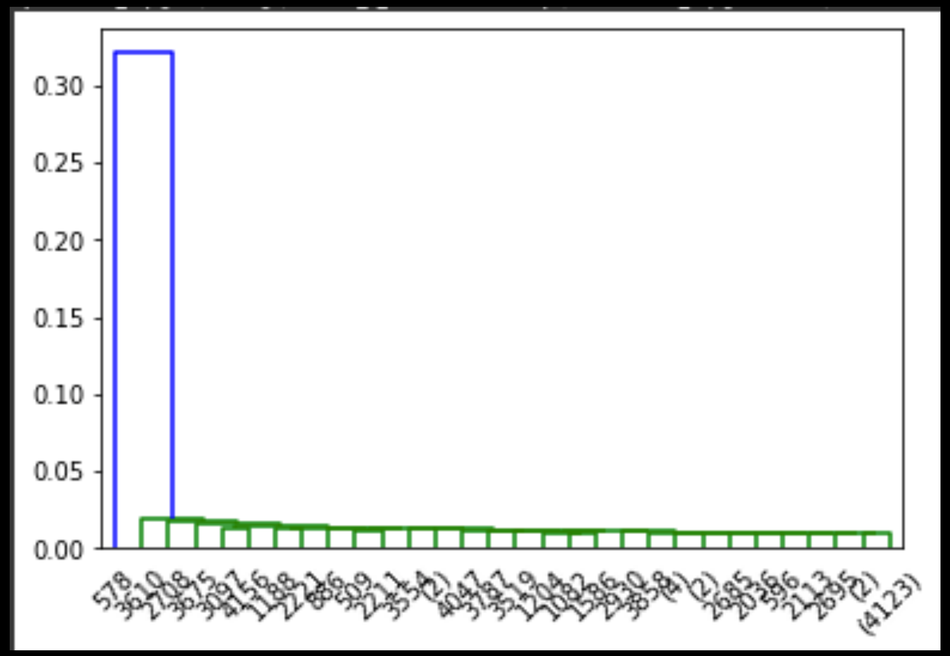



```
Number of cluster detected:  1
85
```

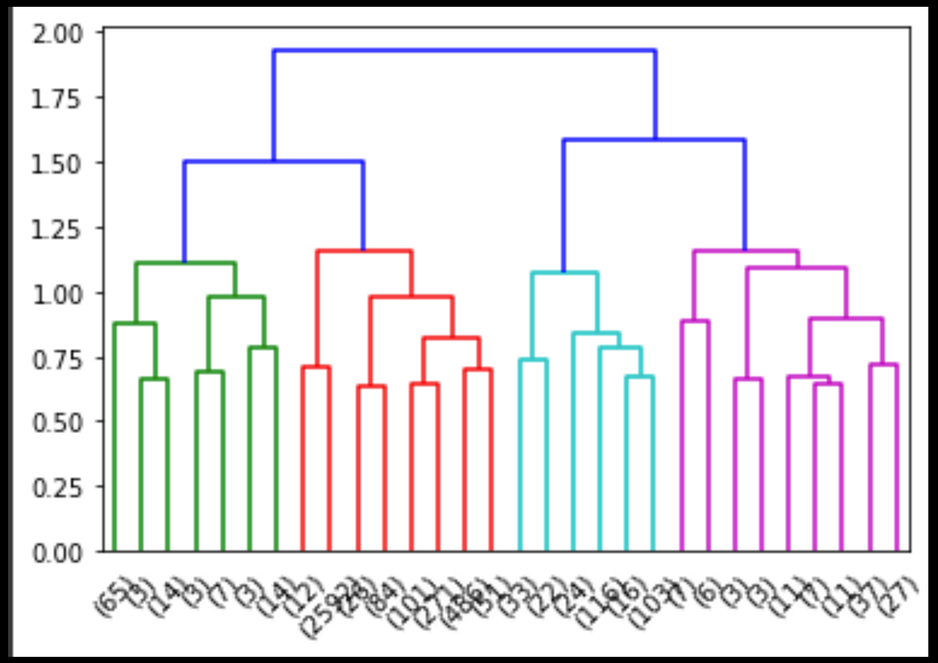


# Hierarchical Test Result

- Single

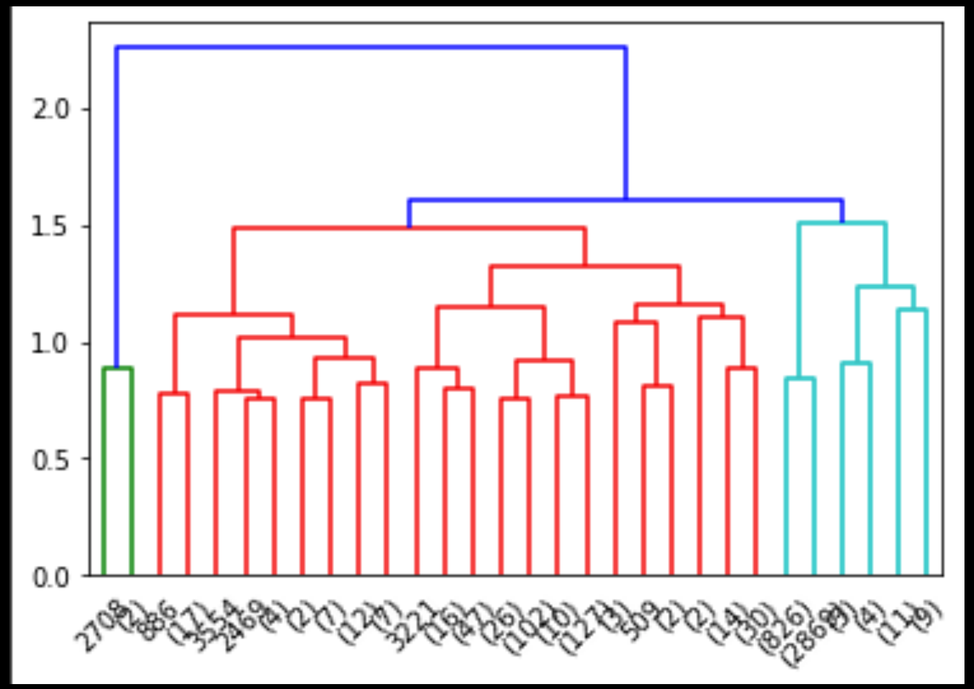



- Complete

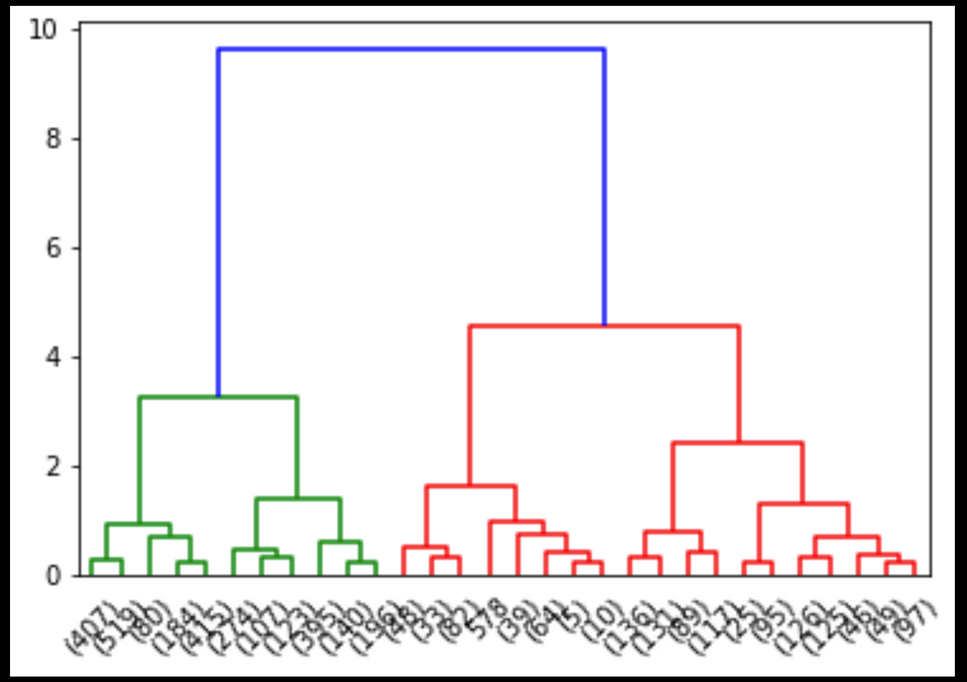


# Hierarchical Test Result

- Average




- Ward



# Conclusion

- DBSCAN , no matter what the values of the initial parameters, I couldn’t obtain a good clustering because it always produced a huge cluster with the majority of the points and another negligible one.
- The Hierarchical approach produced very unbalanced clusters with some linking methods (eg Single, Ward). Instead, when using other combinations (like the Euclidean Distance Exact Method), it produced a better distribution for my data.
-

Thanks for listening :)