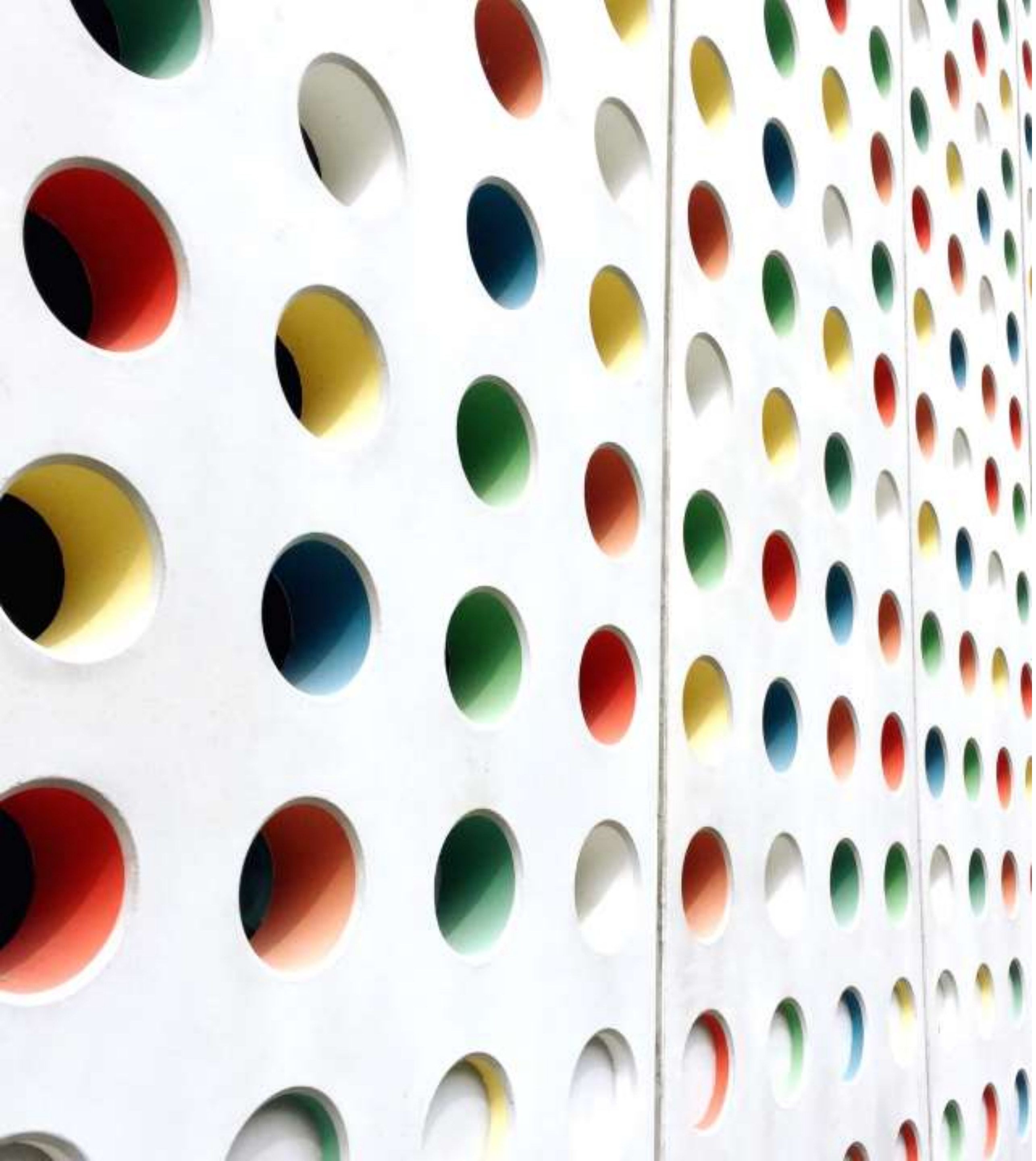


Customer Segmentation

BY:SASWAT MISHRA

About the Data

1. Customers.csv

- CustomerID: Unique identifier for each customer.
- CustomerName: Name of the customer.
- Region: Continent where the customer resides.
- SignupDate: Date when the customer signed up.

About the Data

3. Transactions.csv

○ TransactionID: Unique identifier for each transaction.

○ CustomerID: ID of the customer who made the transaction.

○ ProductID: ID of the product sold.

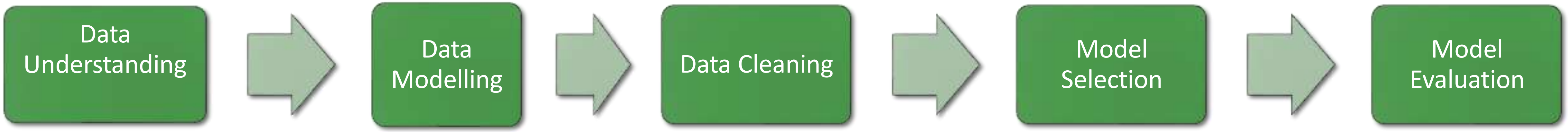
○ TransactionDate: Date of the transaction.

○ Quantity: Quantity of the product purchased.

○ TotalValue: Total value of the transaction.

○ Price: Price of the product sold.

Process



Data Understanding

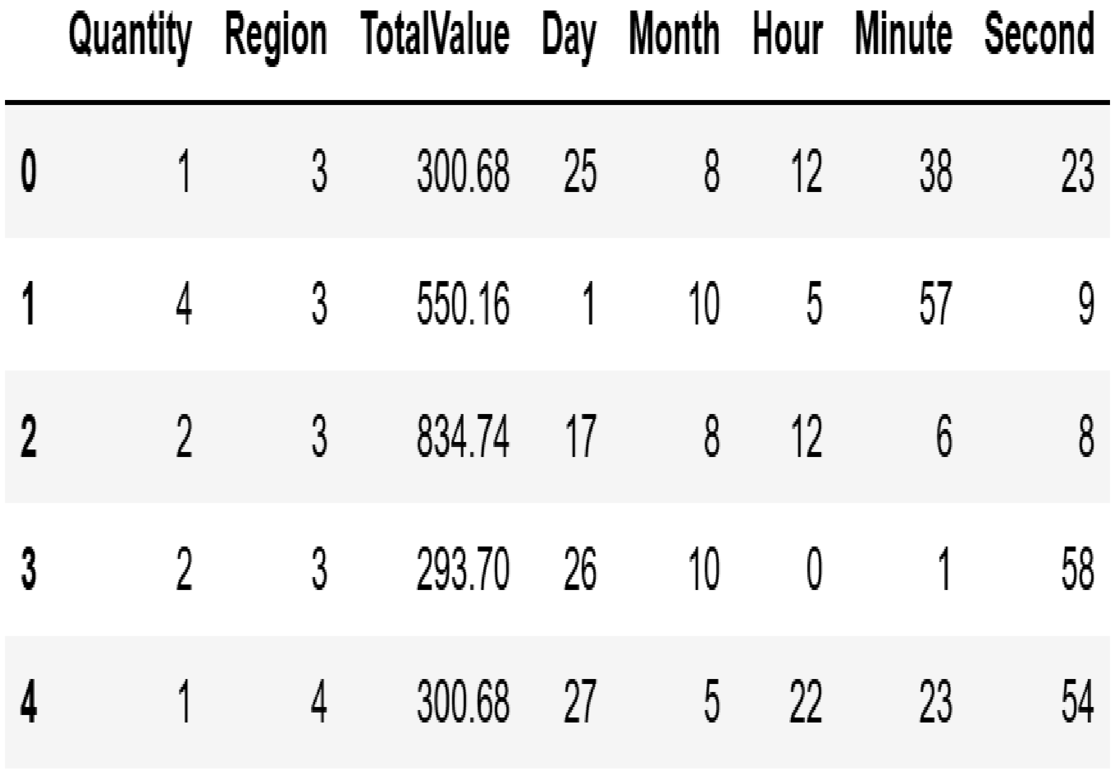
Data Modelling

Data Cleaning

Model Selection

Model Evaluation

Feature Selection

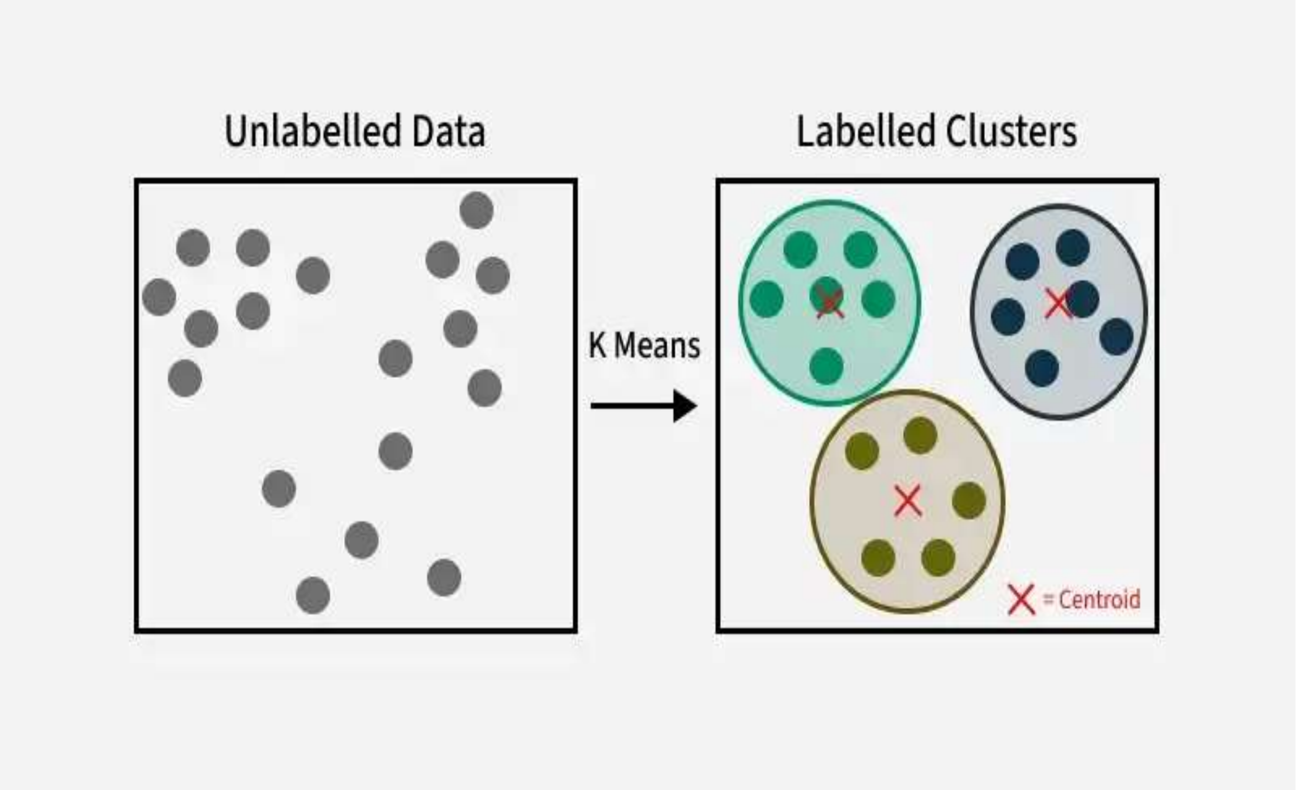


	Quantity	Region	TotalValue	Day	Month	Hour	Minute	Second
0	1	3	300.68	25	8	12	38	23
1	4	3	550.16	1	10	5	57	9
2	2	3	834.74	17	8	12	6	8
3	2	3	293.70	26	10	0	1	58
4	1	4	300.68	27	5	22	23	54

- The features that i have used, to pass further to our clustering model.

- As K Means is a distance based algorithm, I have further done scaling using the MinMax Scaler on some columns.

Model Used : K Means Clustering

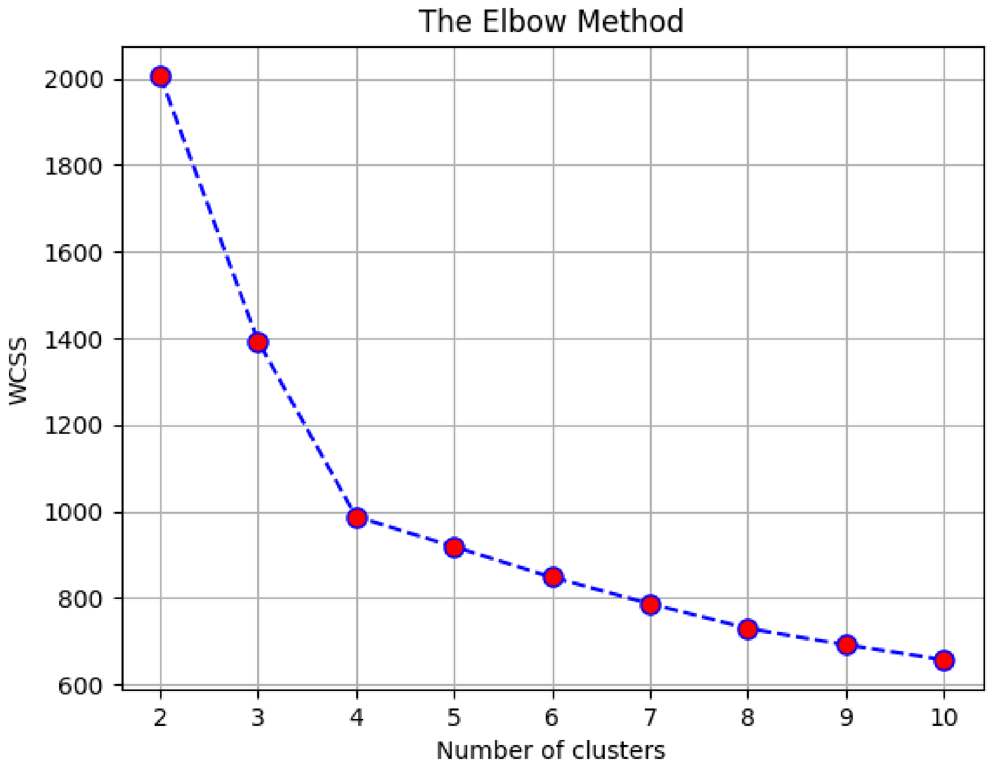


K-means clustering is an unsupervised machine learning algorithm used for partitioning data into **K clusters**. It minimizes intra-cluster variance by iteratively assigning data points to the nearest cluster centroid and recalculating centroids based on the mean of assigned points. The algorithm follows these steps:

1.Select **K** initial centroids.
2.Assign each data point to the nearest centroid.
3.Recalculate centroids as the mean of assigned points.
4.Repeat steps 2-3 until centroids stabilize.

It is efficient but sensitive to the initial choice of K and outliers. Applications include image segmentation, anomaly detection, and customer segmentation.

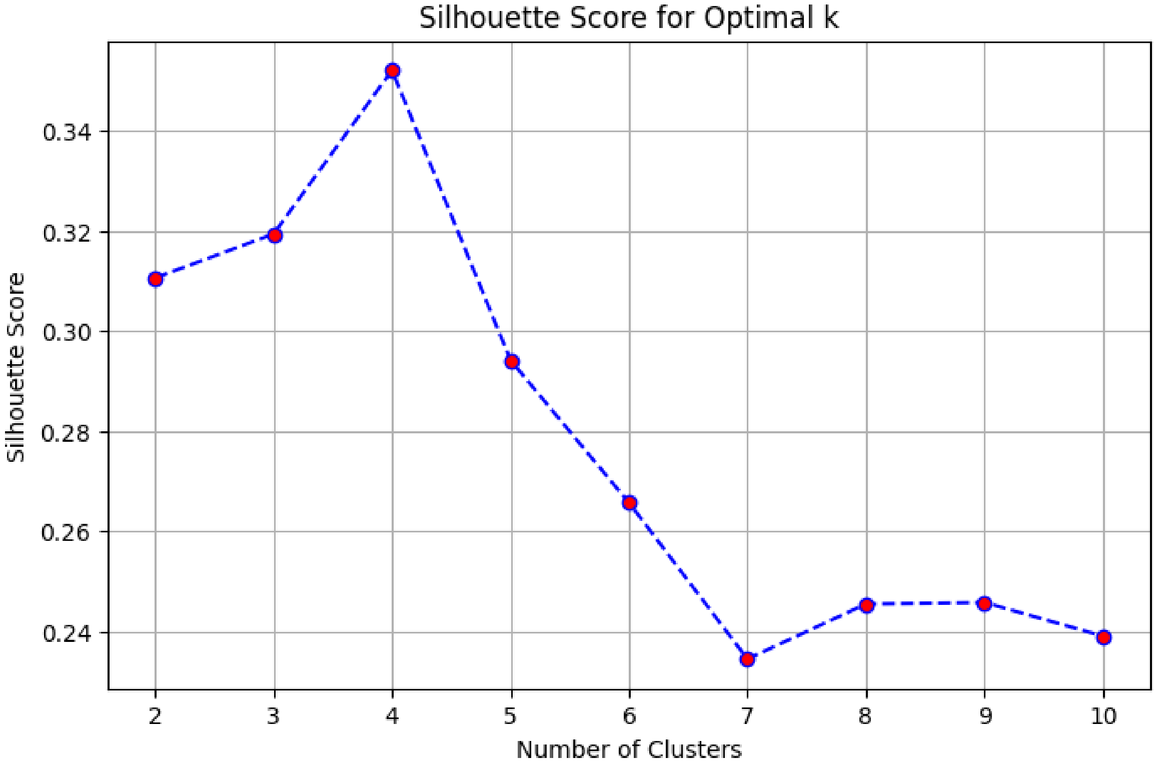
Model Evaluation



- Firstly we have used the “Elbow Method” to evaluate our model.

- By looking at the model, we can determine that the elbow point is most likely 3 or 4.

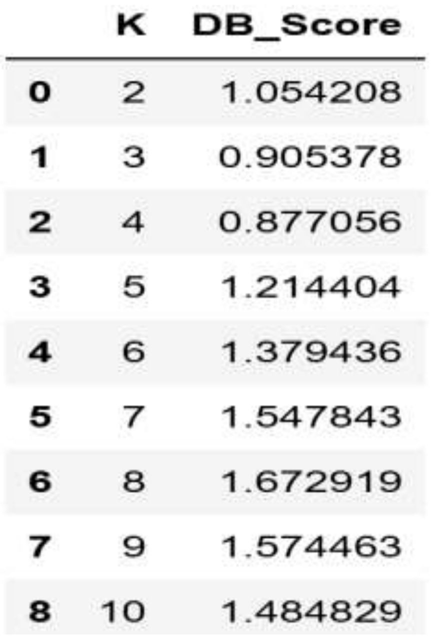
Model Evaluation



- The second evaluating metric that we have used is “Silhouette Score”.

- The silhouette score too indicates that the ideal no. of clusters is likely to be 4.

Model Evaluation

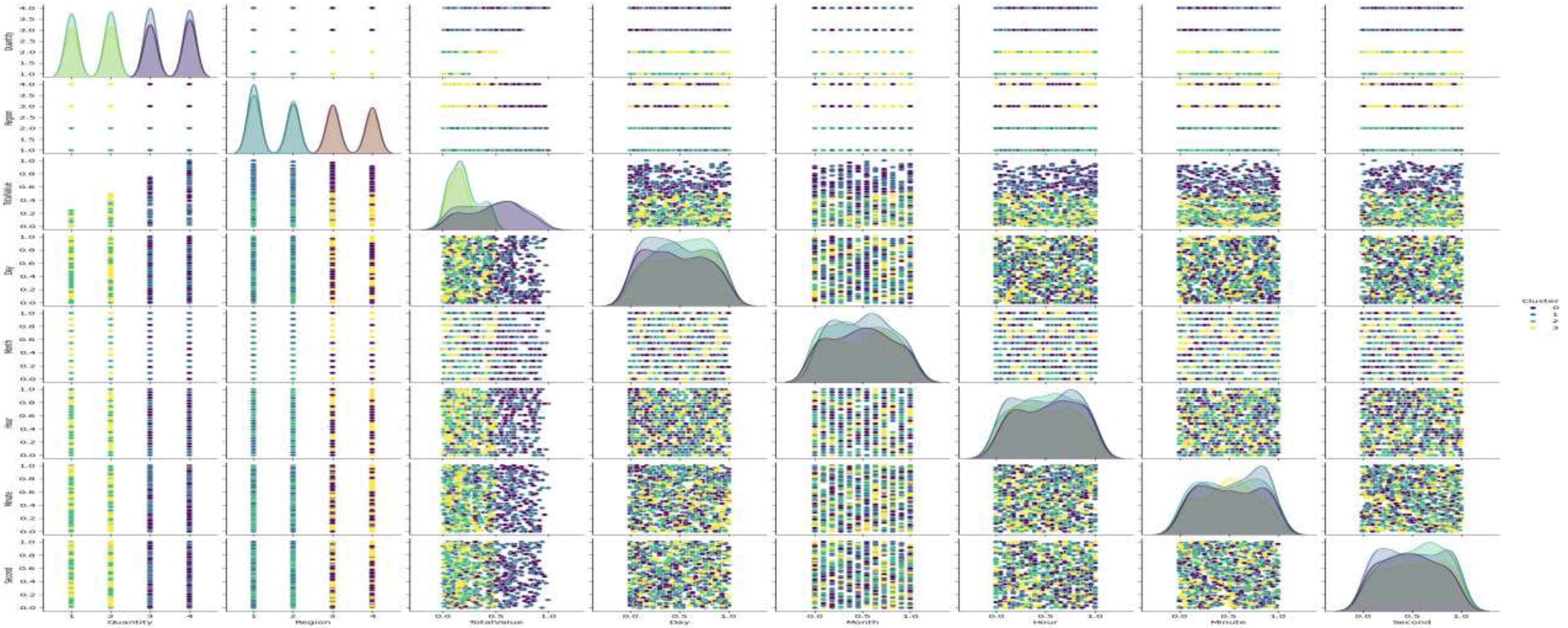


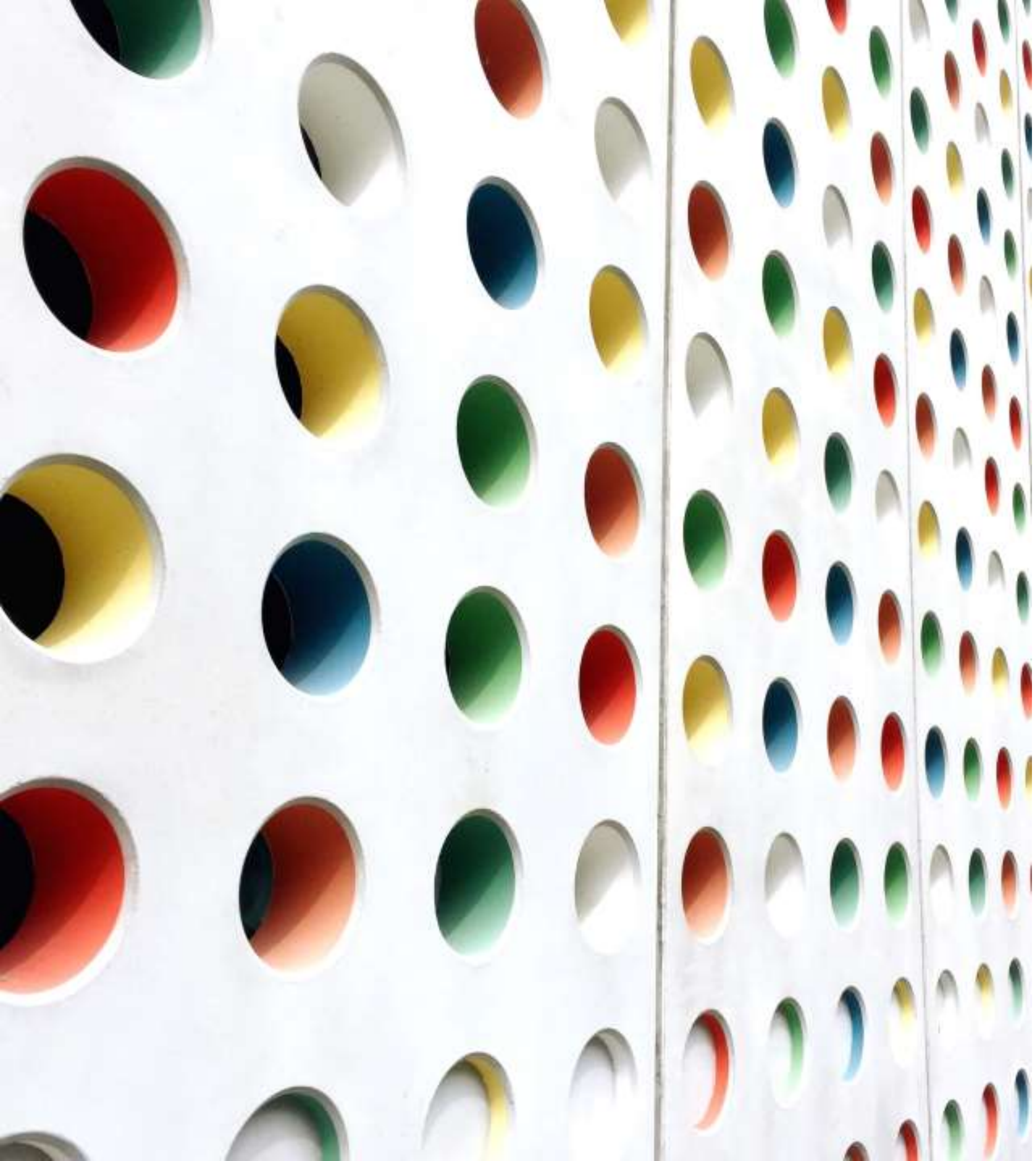
	K	DB_Score
0	2	1.054208
1	3	0.905378
2	4	0.877056
3	5	1.214404
4	6	1.379436
5	7	1.547843
6	8	1.672919
7	9	1.574463
8	10	1.484829

- The third evaluating metric that we have used is “Davies Bouldin Score”.

- The lowest DB_Score is for K=4, hence we’ll settle for 4 no. of clusters.

Clusters visualization using pairplot





Thank You

ANY QUESTIONS?