Intern Id: 180010 **Intern Name: Bibek Kr. Sah**

Project Details & Instructions

Wholesale Customer Segmentation

GitHub Link:

https://github.com/bbksa/Wholesale-Customer-Segmentation.git

Problem Statement:

The aim of this problem is to segment the clients of a wholesale distributor based on their annual spending on diverse product categories, like milk, grocery, region, etc.

Dataset Description:

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

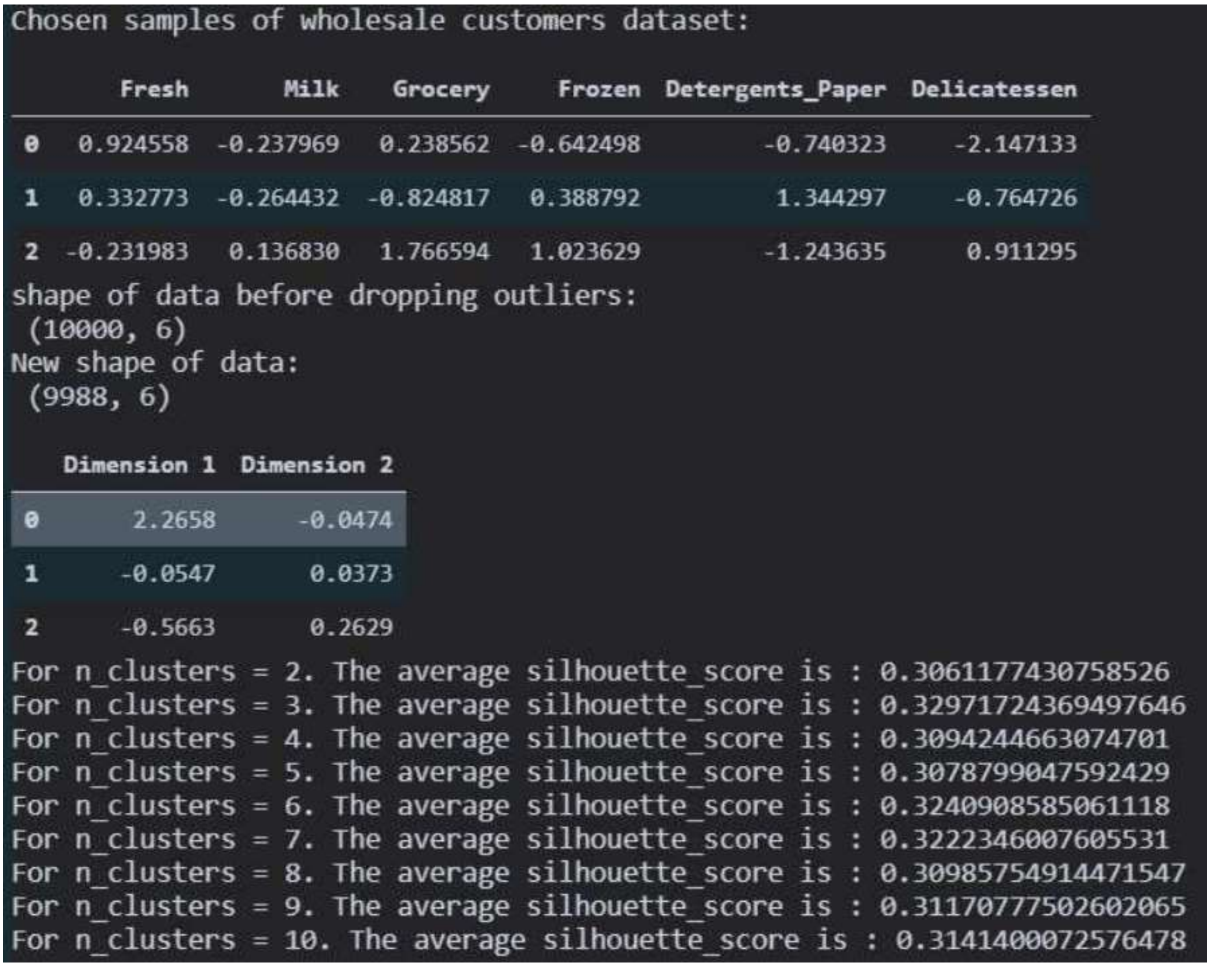
Attribute Information:

1) FRESH: annual spending (m.u.) on fresh products (Continuous);
2) MILK: annual spending (m.u.) on milk products (Continuous);
3) GROCERY: annual spending (m.u.)on grocery products (Continuous);
4) FROZEN: annual spending (m.u.)on frozen products (Continuous)
5) DETERGENTS_PAPER: annual spending (m.u.) on detergents and paper products (Continuous)
6) DELICATESSEN: annual spending (m.u.)on and delicatessen products (Continuous);
7) CHANNEL: customersâ€™ Channel - Horeca (Hotel/Restaurant/CafÃ©) or Retail channel (Nominal)
8) REGION: customersâ€™ Region â€“ Lisnon, Oporto or Other (Nominal)

There are multiple product categories – Fresh, Milk, Grocery, etc. The values represent the number of units purchased by each client for each product. The goal of this project is to make clusters from this data that can segment similar clients together.

Conclusion and Implications:

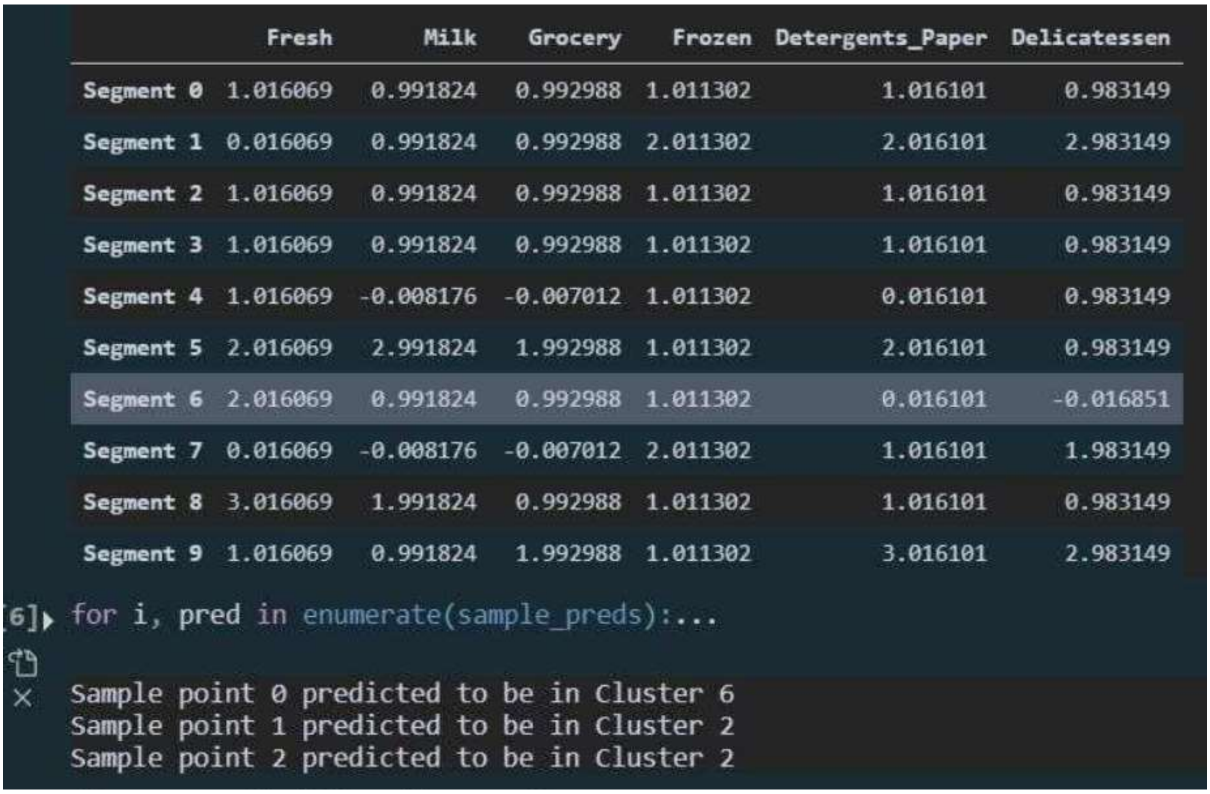
1. **How to use this knowledge?**
2. **How can the wholesale distributor use the customer segments to determine which customers, if any, would react positively to the change in delivery service?**
3. **How can the wholesale distributor label the new customers using only their estimated product spending and the customer segment data?**



```
Chosen samples of wholesale customers dataset:

       Fresh      Milk   Grocery    Frozen  Detergents_Paper  Delicatessen
0   0.924558 -0.237969  0.238562 -0.642498         -0.740323     -2.147133
1   0.332773 -0.264432 -0.824817  0.388792          1.344297     -0.764726
2  -0.231983  0.136830  1.766594  1.023629         -1.243635      0.911295
shape of data before dropping outliers:
 (10000, 6)
New shape of data:
 (9988, 6)

   Dimension 1  Dimension 2
0       2.2658      -0.0474
1      -0.0547       0.0373
2      -0.5663       0.2629
For n_clusters = 2. The average silhouette_score is : 0.3061177430758526
For n_clusters = 3. The average silhouette_score is : 0.32971724369497646
For n_clusters = 4. The average silhouette_score is : 0.3094244663074701
For n_clusters = 5. The average silhouette_score is : 0.3078799047592429
For n_clusters = 6. The average silhouette_score is : 0.3240908585061118
For n_clusters = 7. The average silhouette_score is : 0.3222346007605531
For n_clusters = 8. The average silhouette_score is : 0.30985754914471547
For n_clusters = 9. The average silhouette_score is : 0.31170777502602065
For n_clusters = 10. The average silhouette_score is : 0.3141400072576478
```



	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Segment 0	1.016069	0.991824	0.992988	1.011302	1.016101	0.983149
Segment 1	0.016069	0.991824	0.992988	2.011302	2.016101	2.983149
Segment 2	1.016069	0.991824	0.992988	1.011302	1.016101	0.983149
Segment 3	1.016069	0.991824	0.992988	1.011302	1.016101	0.983149
Segment 4	1.016069	-0.008176	-0.007012	1.011302	0.016101	0.983149
Segment 5	2.016069	2.991824	1.992988	1.011302	2.016101	0.983149
Segment 6	2.016069	0.991824	0.992988	1.011302	0.016101	-0.016851
Segment 7	0.016069	-0.008176	-0.007012	2.011302	1.016101	1.983149
Segment 8	3.016069	1.991824	0.992988	1.011302	1.016101	0.983149
Segment 9	1.016069	0.991824	1.992988	1.011302	3.016101	2.983149

```
6] for i, pred in enumerate(sample_preds):...
```

```
Sample point 0 predicted to be in Cluster 6
Sample point 1 predicted to be in Cluster 2
Sample point 2 predicted to be in Cluster 2
```

Impact on Segment 0

- Intuitively, the impact on Segment 0's customers should be minimal.
- This is because their products are mainly non-perishable products from "Grocery" to "Detergents Paper".
- However, this situation is complicated as this segment has high spending on "Milk" products which is perishable.
- But with advances in preservation, most "Milk" products last more than a week these days.

Impact on Segment 1

- One would surmise that Segment 1's customers would have a substantial impact by the change in delivery service.
- This is because their products are highly perishable such as "Fresh" products including fruits, vegetables, seafood and meat.
- We can formalize the impact by running an experiment to determine which group of customers would have the greatest impact.

1. Randomly sample 4 groups where we sample 2 groups from each cluster.
    - Group 0a, 0b would be the group experiencing the change and the control group respectively for cluster 0.
    - Group 1a, 1b would be the group experiencing the change and the control group respectively for cluster 1.

2. We will change the schedules for group 0a and 1a keeping the schedules for 0b and 1b unchanged.
3. We will have 2 metrics.
   - We will conduct customer satisfaction survey for all groups.
   - We will cross-reference their satisfaction level with their spending.
4. Clients experiencing a negative impact would have a low satisfaction level and a decreased or similar spending. And clients experiencing a positive impact would have a high satisfaction level and an increased or similar spending.
   - We can investigate anomalies where clients display contradictory signals like expressing a low satisfaction level and increasing spending, and vice versa.