MSc in Business Analytics
Academic Year 2019 – 2020
Spring Quarter

Mining Big Datasets

Assignment 1

Due Date: May 24th, 2020

Professor: Yiannis Kotidis
Assistant responsible for this assignment: Ioanna Filippidou

Vasiliki Karamesiou f2821903
Erasmia Kornelatou f2821907

Department of Management Science and Technology
Athens University of Economics and Business

1. Introduction

The purpose of this assignment is to implement a simple workflow that will assess the similarity between supermarket customers and suggest for any input customer a list of his/her 10 most similar other customers. In order to calculate the similarity between customers we will first have to compute the dissimilarity for every given attribute.

2. Dataset Description

The dataset contains demographic characteristics of supermarket customers along with a list of groceries. More specifically, the dataset includes 10000 supermarket customer profiles with the following attributes:

Customer ID: The unique id of the customer.

Age: The age of the customer.

Sex: Male-Female.

Marital Status: Married, Single, Divorced.

Education: Primary, Secondary, Tertiary.

Annual Income: The annual customer income.

Customer Rating: The rating of the supermarket from the customer (Poor, Fair, Good, Very Good, Excellent).

Persons in Household: Number of persons in the household.

Occupation: The occupation of each customer (retired, housemaid, unemployed, management, entrepreneur, blue-collar, self-employed, services, technician).

Groceries: A list of the customer groceries.

3. Implementation

Jupyter notebook was used so as to implement the solution of our problem. Python language programming was chosen with the exploit of the Pandas library.

3.1. Pre Processing

Distinguishing the type of each attribute (categorical, ordinal, numerical or set):

Column	Type
Age	numerical
Sex	categorical
Marital Status	categorical
Education	ordinal
Annual Income	numerical
Customer Rating	ordinal
Persons in Household	numerical
Occupation	categorical
Groceries	set

Figure 1: *The type of each attribute*

The attributes ‘Education’ and ‘Customer Rating’ have further stages, each with an increasing degree, so they are considered as ordinal variables. The first one has 3 stages:

1. ‘Primary’,

2. 'Secondary',

3.'Tertiary'.

The second one has 5 stages:

1. 'Poor',

2. ‘Fair’,

3. ‘Good’,

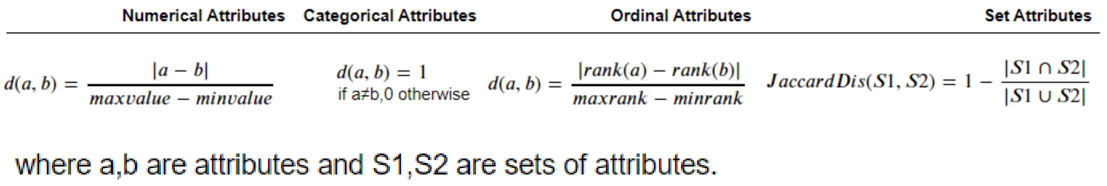
4. ‘Very_Good’,

5. ‘Excellent’

What is more, the attributes ‘Age’ and ‘Income’ have numerical missing values. There are 473 missing values in the "Age" column and 477 missing values in the "Income" column. These values will be replaced by the average value (keeping the integer part of the average) of the corresponding column.

3.2. Compute data (dis-)similarity

A pandas dataframe was used for the load of the data. The first function calculates the dissimilarity matrix between the customer given and all the other customers of the dataset for a specific attribute and it is appropriate for categorical, ordinal and numerical attributes. The second function does the same procedure but for sets of attributes using the Jaccard similarity. What is more, given a customer as input, the third function was created to calculate the average of the dissimilarity values of all the attributes of each customer.



Numerical Attributes	Categorical Attributes	Ordinal Attributes	Set Attributes
$d(a,b) = \frac{\lvert a-b \rvert}{maxvalue - minvalue}$	$d(a,b) = 1$ if a≠b,0 otherwise	$d(a,b) = \frac{\lvert rank(a) - rank(b) \rvert}{maxrank - minrank}$	$JaccardDis(S1,S2) = 1 - \frac{\lvert S1 \cap S2 \rvert}{\lvert S1 \cup S2 \rvert}$

where a,b are attributes and S1,S2 are sets of attributes.

These types are being referred to the first and the second function.

$$d(a,b) = \frac{1}{9} * (d_{Age}(a,b) + d_{Sex}(a,b) + d_{MaritalStatus}(a,b) + d_{Education}(a,b) + d_{Income}(a,b) + d_{CustomerRating}(a,b) + d_{PersonsInHousehold}(a,b) + d_{Occupation}(a,b) + d_{Groceries}(a,b))$$

This type is being referred to the third function.

3.3. Nearest Neighbor (NN) search

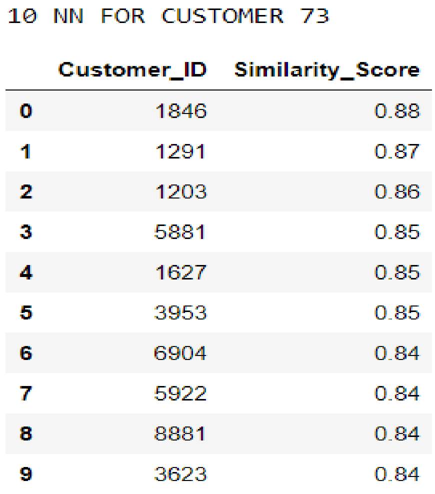
Finally, the "similarity" function was created to calculate the similarity score between the customer given and all the other customers of the dataset. When this function is called, it is called for the ten most similar customers and the similarity score is printed in a descending order.

$$similarityScore = 1 - dissimilarityScore$$

4. Results

Exploiting the above function the results for the similarity score for the requested customers are the following:

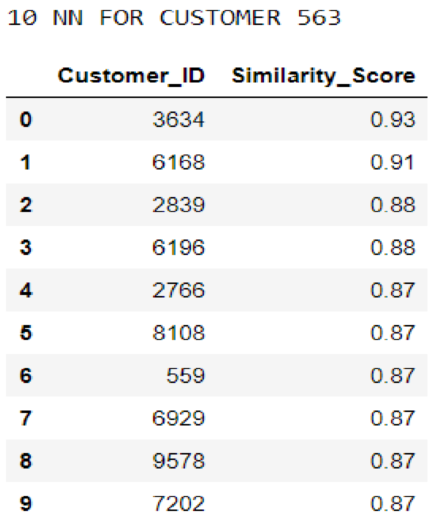
1. For Customer_ ID: 73



```
10 NN FOR CUSTOMER 73
```

	Customer_ID	Similarity_Score
0	1846	0.88
1	1291	0.87
2	1203	0.86
3	5881	0.85
4	1627	0.85
5	3953	0.85
6	6904	0.84
7	5922	0.84
8	8881	0.84
9	3623	0.84

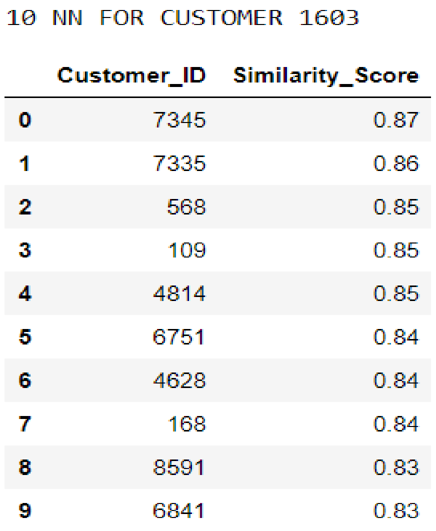
2. For Customer_ ID: 563



```
10 NN FOR CUSTOMER 563
```

	Customer_ID	Similarity_Score
0	3634	0.93
1	6168	0.91
2	2839	0.88
3	6196	0.88
4	2766	0.87
5	8108	0.87
6	559	0.87
7	6929	0.87
8	9578	0.87
9	7202	0.87

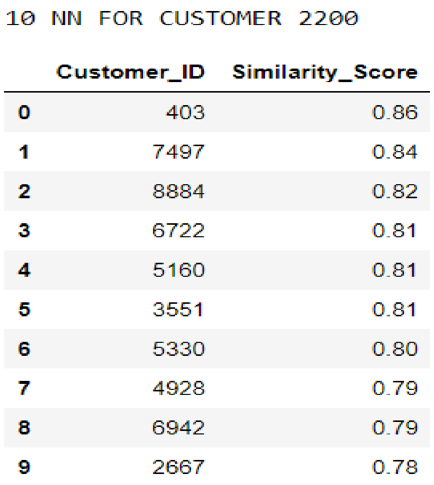
3. For Customer_ ID: 1603



```
10 NN FOR CUSTOMER 1603
```

	Customer_ID	Similarity_Score
0	7345	0.87
1	7335	0.86
2	568	0.85
3	109	0.85
4	4814	0.85
5	6751	0.84
6	4628	0.84
7	168	0.84
8	8591	0.83
9	6841	0.83

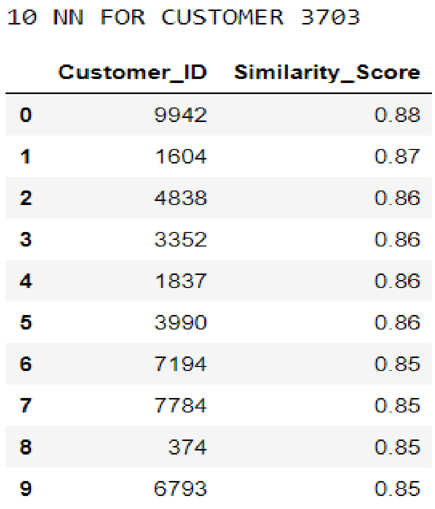
4. For Customer_ ID: 2200



```
10 NN FOR CUSTOMER 2200
```

	Customer_ID	Similarity_Score
0	403	0.86
1	7497	0.84
2	8884	0.82
3	6722	0.81
4	5160	0.81
5	3551	0.81
6	5330	0.80
7	4928	0.79
8	6942	0.79
9	2667	0.78

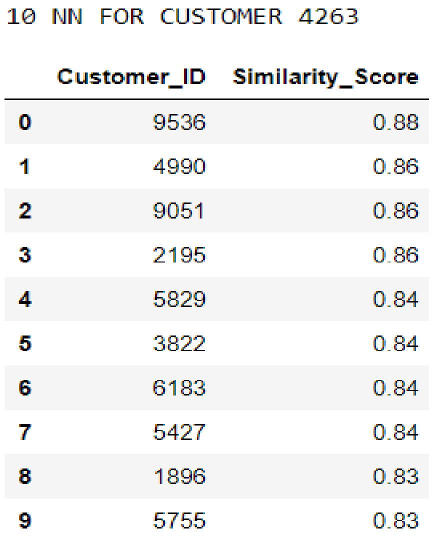
5. For Customer_ ID: 3703



```
10 NN FOR CUSTOMER 3703
```

	Customer_ID	Similarity_Score
0	9942	0.88
1	1604	0.87
2	4838	0.86
3	3352	0.86
4	1837	0.86
5	3990	0.86
6	7194	0.85
7	7784	0.85
8	374	0.85
9	6793	0.85

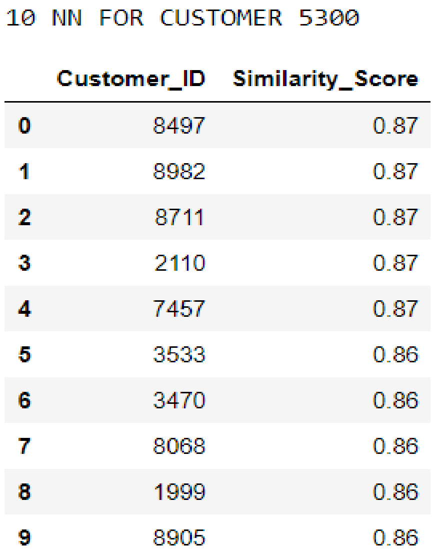
6. For Customer_ ID: 4263



```
10 NN FOR CUSTOMER 4263
```

	Customer_ID	Similarity_Score
0	9536	0.88
1	4990	0.86
2	9051	0.86
3	2195	0.86
4	5829	0.84
5	3822	0.84
6	6183	0.84
7	5427	0.84
8	1896	0.83
9	5755	0.83

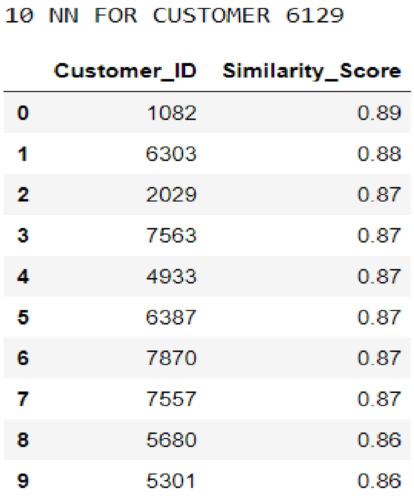
7. For Customer_ ID: 5300



```
10 NN FOR CUSTOMER 5300
```

	Customer_ID	Similarity_Score
0	8497	0.87
1	8982	0.87
2	8711	0.87
3	2110	0.87
4	7457	0.87
5	3533	0.86
6	3470	0.86
7	8068	0.86
8	1999	0.86
9	8905	0.86

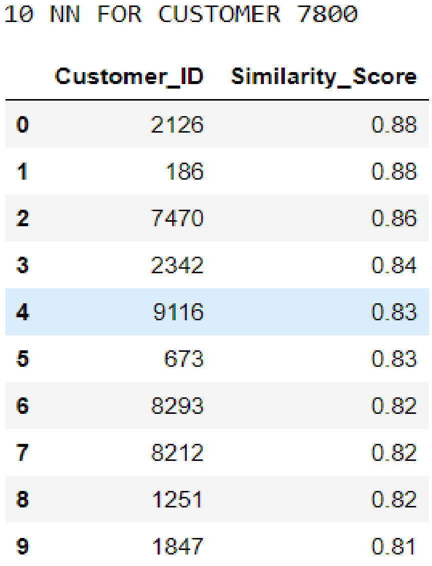
8. For Customer_ ID: 6129



```
10 NN FOR CUSTOMER 6129
```

	Customer_ID	Similarity_Score
0	1082	0.89
1	6303	0.88
2	2029	0.87
3	7563	0.87
4	4933	0.87
5	6387	0.87
6	7870	0.87
7	7557	0.87
8	5680	0.86
9	5301	0.86

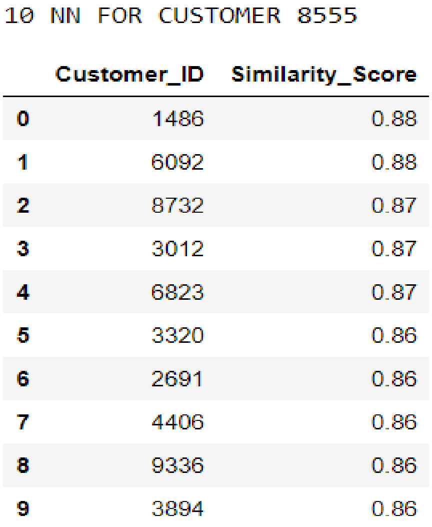
9. For Customer_ ID: 7800



```
10 NN FOR CUSTOMER 7800
```

	Customer_ID	Similarity_Score
0	2126	0.88
1	186	0.88
2	7470	0.86
3	2342	0.84
4	9116	0.83
5	673	0.83
6	8293	0.82
7	8212	0.82
8	1251	0.82
9	1847	0.81

10. For Customer_ ID: 8555



```
10 NN FOR CUSTOMER 8555
```

	Customer_ID	Similarity_Score
0	1486	0.88
1	6092	0.88
2	8732	0.87
3	3012	0.87
4	6823	0.87
5	3320	0.86
6	2691	0.86
7	4406	0.86
8	9336	0.86
9	3894	0.86