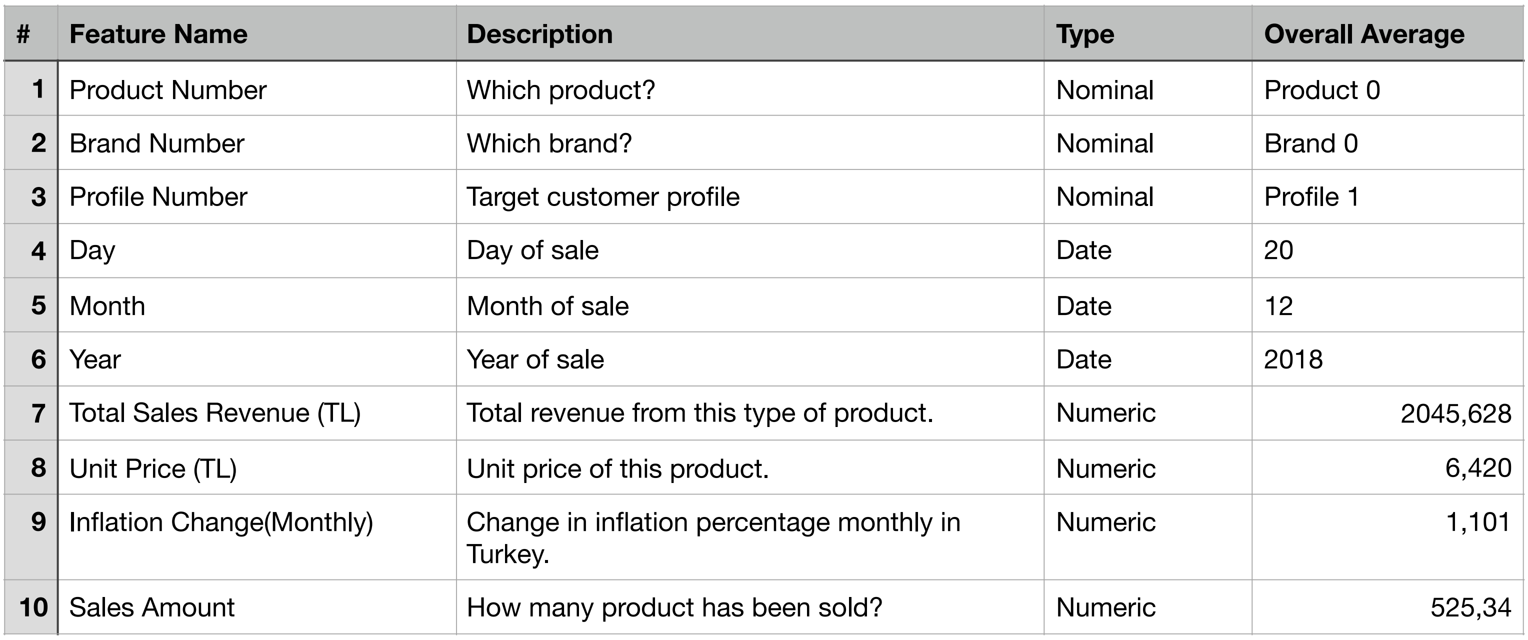
12 May 2019

Group 4

CSE 4062 Term Project

Delivery #5 - Descriptive Analytics



#	Feature Name	Description	Type	Overall Average
1	Product Number	Which product?	Nominal	Product 0
2	Brand Number	Which brand?	Nominal	Brand 0
3	Profile Number	Target customer profile	Nominal	Profile 1
4	Day	Day of sale	Date	20
5	Month	Month of sale	Date	12
6	Year	Year of sale	Date	2018
7	Total Sales Revenue (TL)	Total revenue from this type of product.	Numeric	2045,628
8	Unit Price (TL)	Unit price of this product.	Numeric	6,420
9	Inflation Change(Monthly)	Change in inflation percentage monthly in Turkey.	Numeric	1,101
10	Sales Amount	How many product has been sold?	Numeric	525,34

Figure 1: Data description

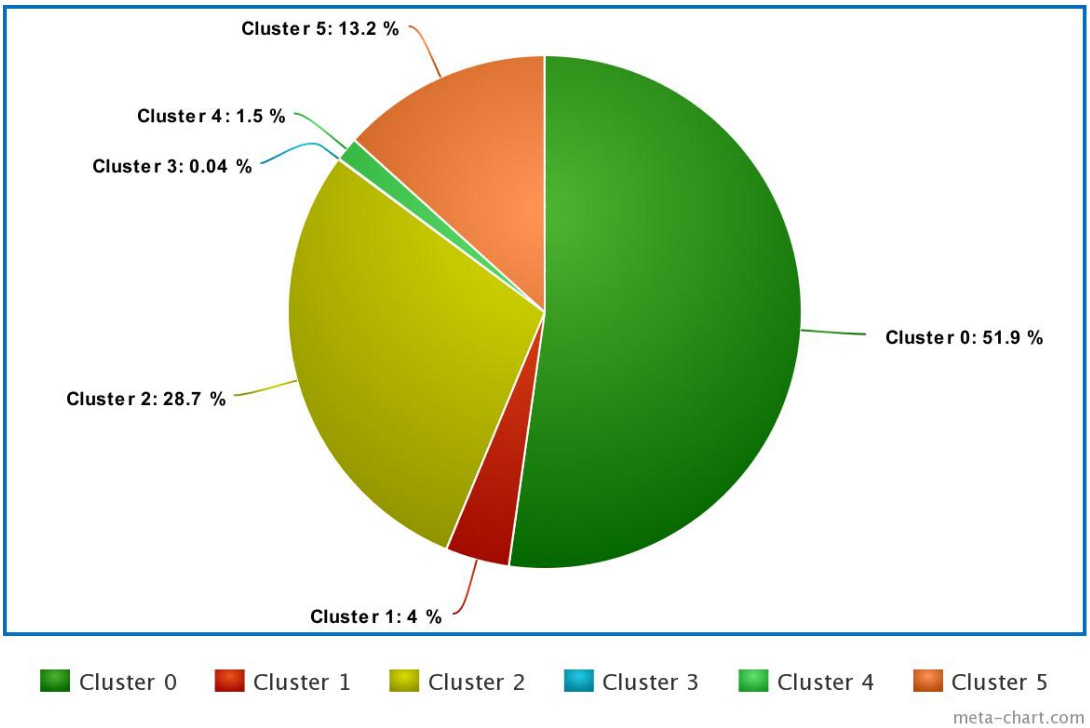
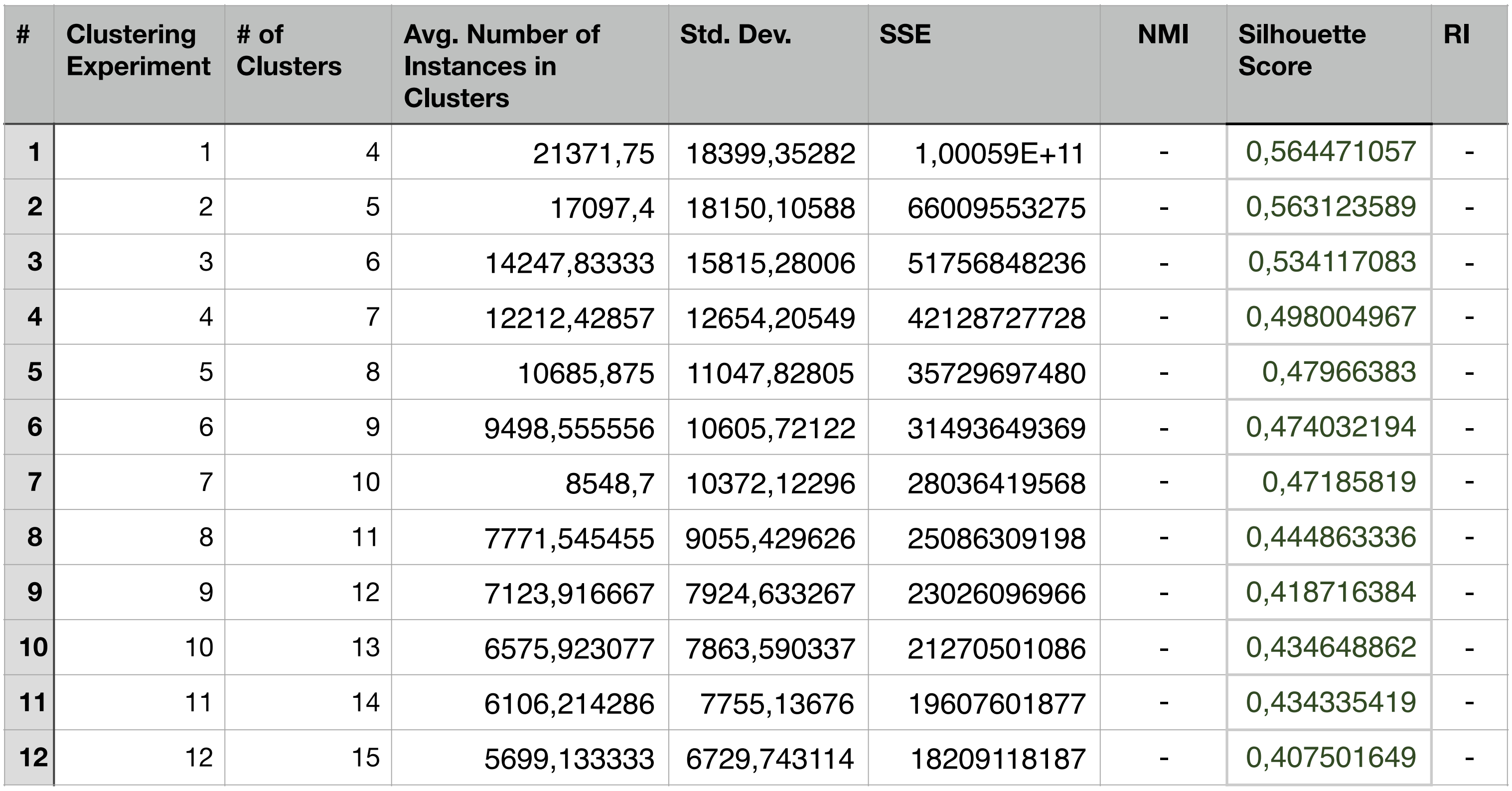


Figure 2: Pie chart for instance distribution on each cluster



#	Clustering Experiment	# of Clusters	Avg. Number of Instances in Clusters	Std. Dev.	SSE	NMI	Silhouette Score	RI
1	1	4	21371,75	18399,35282	1,00059E+11	-	0,564471057	-
2	2	5	17097,4	18150,10588	66009553275	-	0,563123589	-
3	3	6	14247,83333	15815,28006	51756848236	-	0,534117083	-
4	4	7	12212,42857	12654,20549	42128727728	-	0,498004967	-
5	5	8	10685,875	11047,82805	35729697480	-	0,47966383	-
6	6	9	9498,555556	10605,72122	31493649369	-	0,474032194	-
7	7	10	8548,7	10372,12296	28036419568	-	0,47185819	-
8	8	11	7771,545455	9055,429626	25086309198	-	0,444863336	-
9	9	12	7123,916667	7924,633267	23026096966	-	0,418716384	-
10	10	13	6575,923077	7863,590337	21270501086	-	0,434648862	-
11	11	14	6106,214286	7755,13676	19607601877	-	0,434335419	-
12	12	15	5699,133333	6729,743114	18209118187	-	0,407501649	-

Figure 3: Clustering metrics results for k-means algorithm

For this delivery, data is updated as follows:

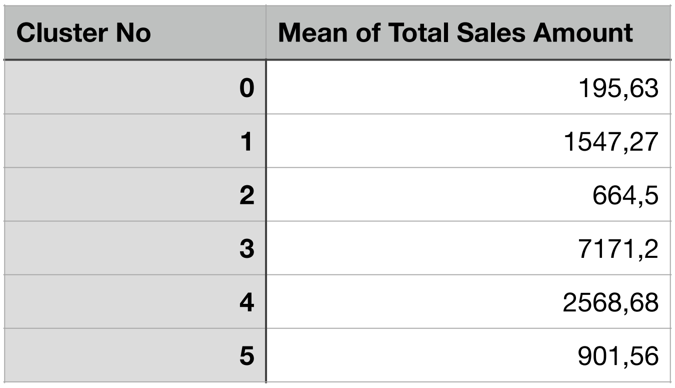
1- Monthly inflation changing rate is added.

2- Date column is split into 3 different columns as day, month and year.

3- Product Number, Brand Number and Profile Number columns were label encoded since k-means is not valid for discrete data.

4- Combined 3 Excel sheets into 1 Excel sheet (2016, 2017, 2018).

K-Means algorithm is used for clustering. All features were used for clustering experiments. k interval chosen as [4, 16) for experiment and optimal k value is obtained from least consecutive decrease of sum of squared errors (SSE) and optimal k was found as k=6. So, there will be clusters with id's 0,1,...,4,5 and with k=6, algorithm clustered well according to mean of total sales amount. Differences between mean of total sales amount for each cluster can be seen clearly in Figure 4.



Cluster No	Mean of Total Sales Amount
0	195,63
1	1547,27
2	664,5
3	7171,2
4	2568,68
5	901,56

Figure 4: Mean of total sales amount for each cluster

Since the data have no defined true labels for clustering purposes, rand index (RI) and normalized mutual information (NMI) cannot be calculated for Figure 3.

References:

- https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
- https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html
- https://pandas.pydata.org/