CSE 4062 – Introduction to Data Science and Analytics Spring 2021

Delivery #5 - Descriptive Analytics

Project Report - Group #8

Group Members:
Caner Dağdaş | 150716001 | Electrical and Electronics Engineering | canerdagdas@hotmail.com
Ceyhun Vardar | 150317022 | Industrial Engineering | vardarceyhun13@gmail.com
Büşra Gökmen | 150116027 | Computer Engineering | busragokmen67@gmail.com
Cem Güleç | 150117828 | Computer Engineering | cem.ggulecc@gmail.com
Ömer Faruk Çakı | 150117821 | Computer Engineering | omerfarukcaki@gmail.com

Project Title: Fraud Detection on Financial Data

Lecturer: Assoc. Prof. Murat Can Ganiz

04.06.2021

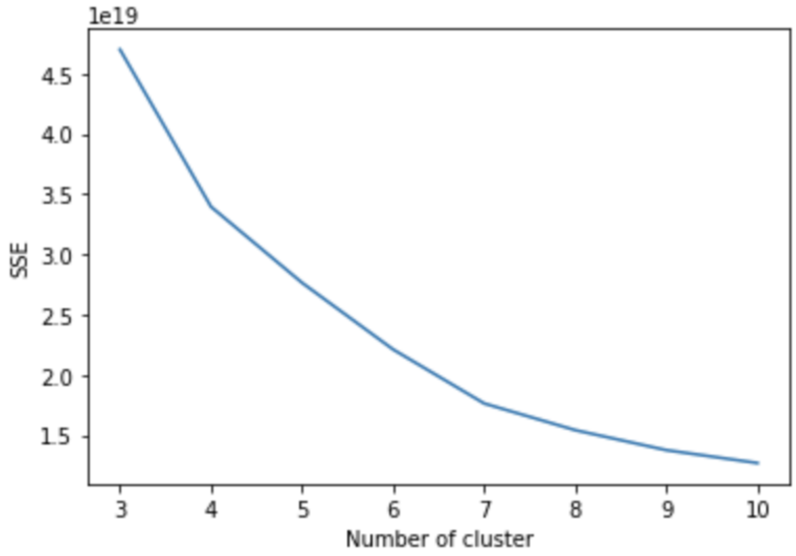


Figure 1: The Elbow Plot Showing the Optimal k=4

```
Max silhouette coeficient for df is 0.7021559575716491 for 3 cluster
```

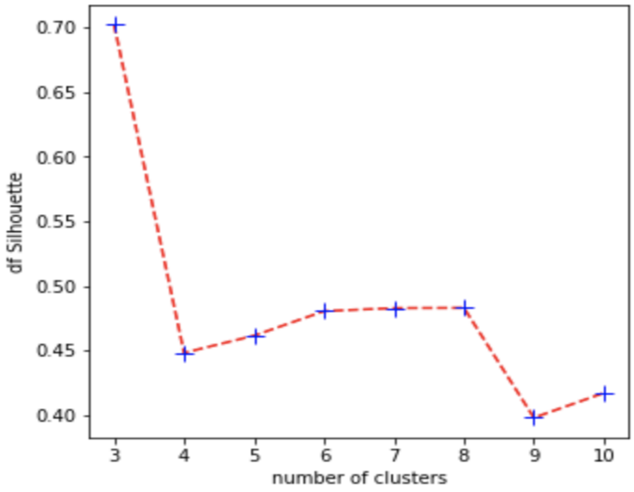


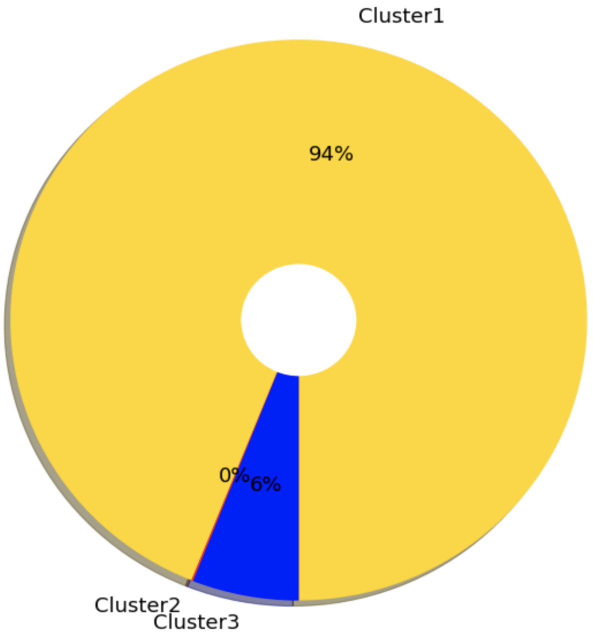
Figure 2: Silhouette coefficient calculation with different cluster numbers

1- Table listing features and their values

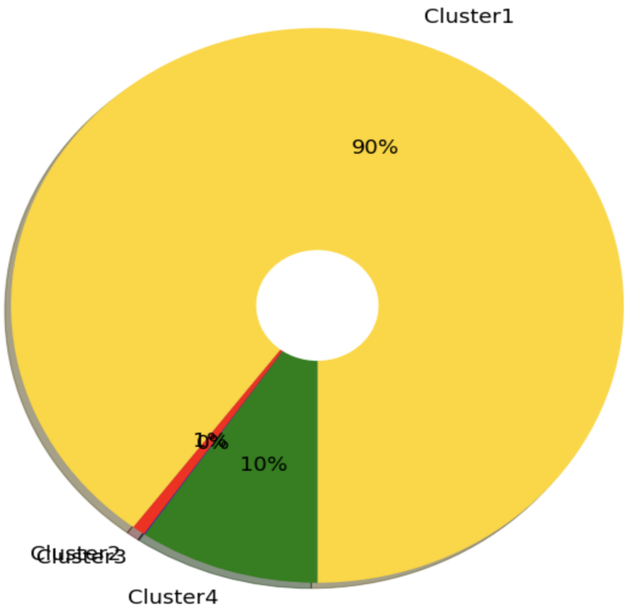
At the below table, features and their clusters by k-means algorithm (k=4) are listed.

#	*feature name*	*description*	*type*	*overall avg./ mode*	*cluster 1*	*cluster 2*	*cluster 3*	*cluster 4*
	step	maps a unit of time in the real world	numeric	*243.397*	*1117466*	*463432*	*10474462*	*142065*
	amount	*amount of the transaction in local currency*	*numeric*	*179861.9*	*2648758*	*9141*	*111705*	*805*
	nameOrig	*customer who started the transaction*	*nominal*	*C1999539787*	*692293*	*690546*	*699864*	*687706*
	oldbalanceOrg	*initial balance before the transaction*	*numeric*	*833883.1*	*2657750*	*985*	*69*	*111605*
	newbalanceOrig	*new balance after the transaction*	*numeric*	*855113.6*	*2707346*	*842*	*59*	*62162*
	nameDest	*customer who is the recipient of the transaction*	*nominal*	*C1286084959*	*689529*	*690622*	*697683*	*692575*
	oldbalanceDest	*initial balance recipient before the transaction*	*numeric*	*1100701.66*	*2488963*	*19153*	*1099*	*261194*
	newbalanceDest	*new balance recipient after the transaction*	*numeric*	*1224996.39*	*2481277*	*19568*	*1350*	*268214*

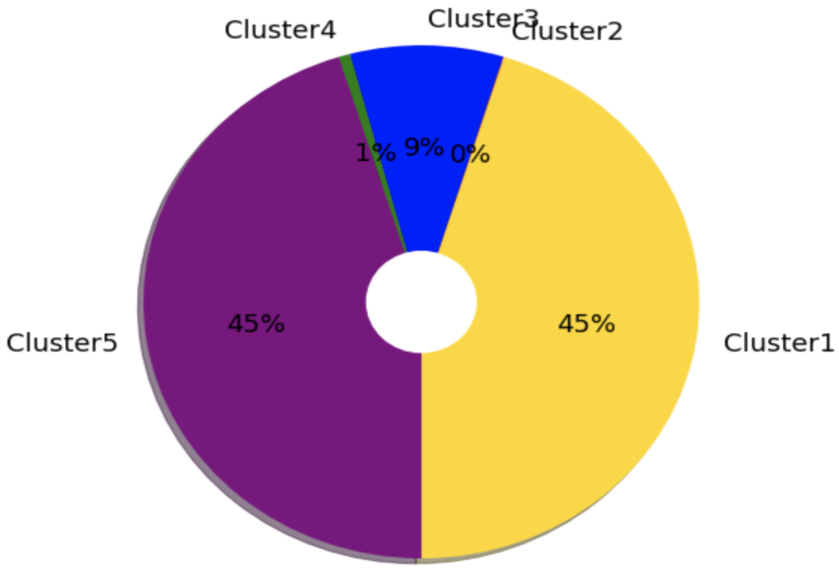
2- Pie charts showing the instance distributions of each cluster as percentages



Graph 1: Pie chart of instance distribution when k=3



Graph 2: Pie chart of instance distribution when k=4



Graph 3: Pie chart of instance distribution when k=5

3- Table for evaluation of clustering experiments

At the below table, instances and their clusters by k-means algorithm with different k values are listed.

Experiment	*# of clusters*	*avg. number of instances in clusters*	*std.dev.*	*SSE*	*NMI*	*Silhouette Value*	*RI*
1	*k=3*	*0:2602389* *1:2863* *2:165157*	*0.47*	*4.7*	*0.00035*	0.70	*-0.0032*
2	*k=4*	*0: 2486468* *1: 19187* *2: 1182* *3: 263572*	*0.88*	*3.39*	*0.000177*	0.44	*-0.00209*
3	*k=5*	*0: 1252392* *1: 1180* *2: 246299* *3: 18738* *4: 1251800*	*1.90*	*2.76*	*6.614*	0.46	*-0.000292*
4	*k=6*	*0: 1158884* *1: 1039* *2: 11616* *3: 352473* *4: 86594* *5: 1159803*	*2.30*	*2.21*	*0.00011*	0.48	*-0.000414*
5	*k=7*	*0: 1139539* *1: 16878* *2: 506* *3: 366760* *4: 2439* *5: 1139916* *6: 104371*	*2.36*	*1.76*	*0.000127*	0.48	*-0.000440*
6	*k=8*	*0: 18931* *1: 1132563* *2: 371895* *3: 920* *4: 216* *5: 1132195* *6: 110247* *7: 3442*	*1.54*	*1.96*	*0.000131*	0.48	*-0.000447*

7	k=9	0: 1074103 1: 885 2: 1074747 3: 3010 4: 59599 5: 210 6: 14768 7: 390814 8: 152273	2.66	1.37	0.000153	0.39	-0.000486
8	k=10	0: 390756 1: 3783 2: 1070432 3: 15489 4: 498 5: 155246 6: 1069712 7: 63065 8: 1329 9: 99	2.33	1.26	0.0001615	0.41	-0.000493

4. Our Inferences and Results

In this section, we analyzed our dataset with K-Means clustering. We are testing the clustering algorithm with different k(3,10) values then we draw a plot for k(3,10) values to SSE values. While using the K-means clustering algorithm, we used the Elbow method to select an optimum K number. Elbow the method is very simple and common; experiments with different values of k and takes the sum of squared errors. The plot has an elbow-like shape. where the break is the shape of an elbow, this place indicates the optimum k to be selected. The higher the value of k, the closer the cluster centers are to each other. After a point, the development of the model will decrease and will be the most optimum value for the elbow point and the k value. According to the plot, the optimum k value is 4 according to the k-means clustering model. We obtained 8 experiments by changing the number of clusters between 3 and 10. We saw that each time we increased the number of clusters, the number of standard deviations increased, but the SSE value decreased.