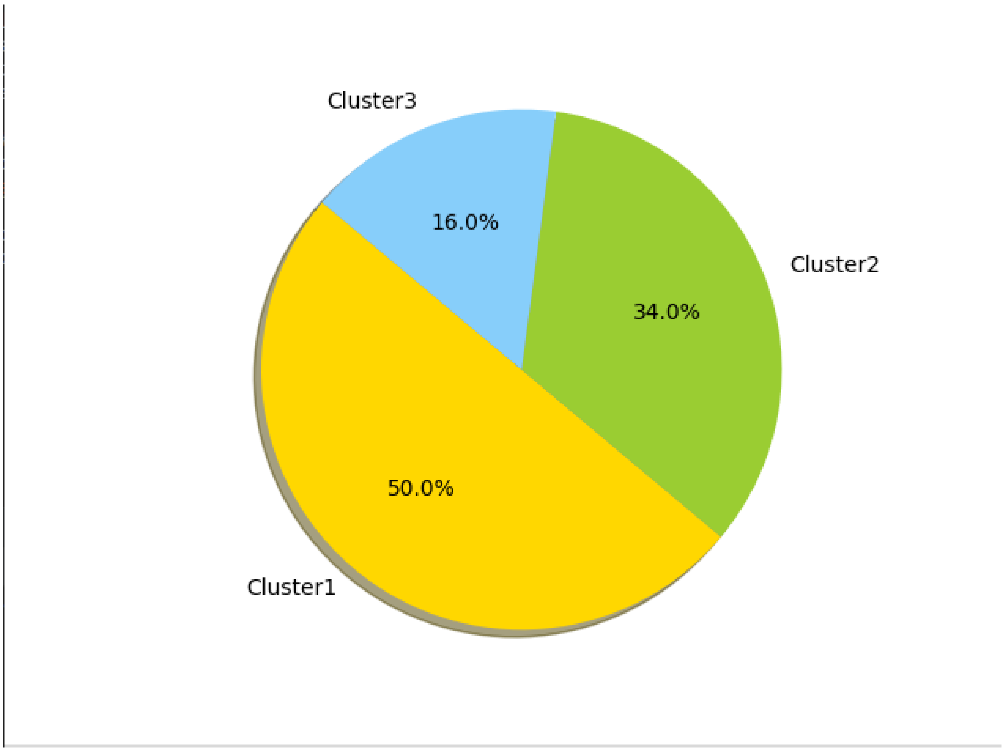
# Introduction to Data Science and Analytics

## Group 2 / Step 5

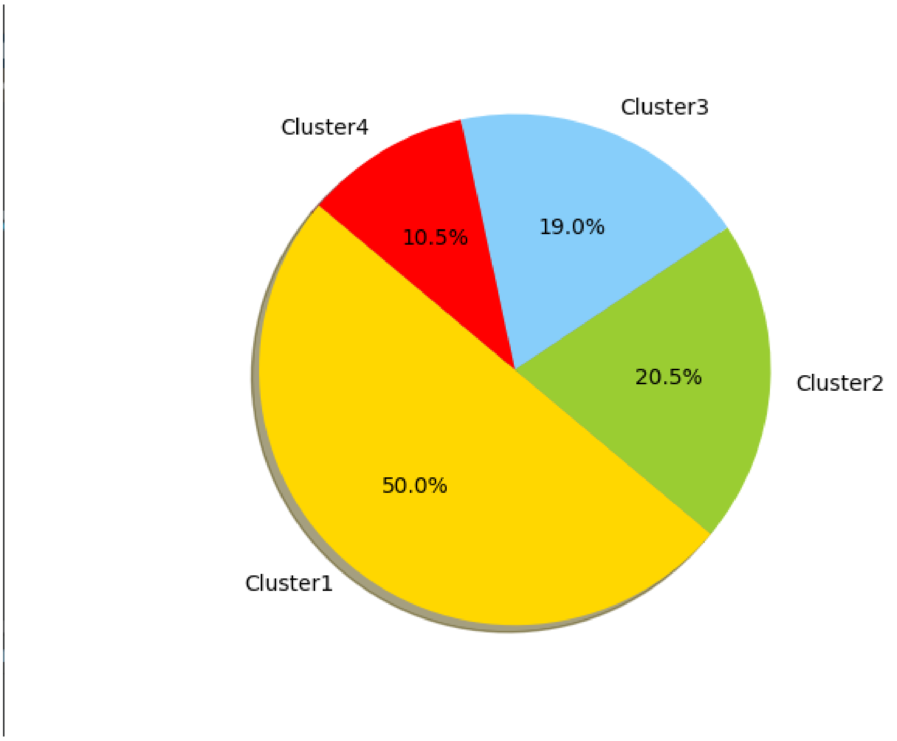
Table 1.0: Instances and their clusters by k-mean clustering algorithm, k = 3.

| # | Description | Type | Overall Avg | Cluster 1 | Cluster 2 | Cluster 3 |
|---|---|---|---|---|---|---|
| 1 | The distribution of instances when k=2 | Numeric | - | 100 | 68 | 32 |

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



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



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

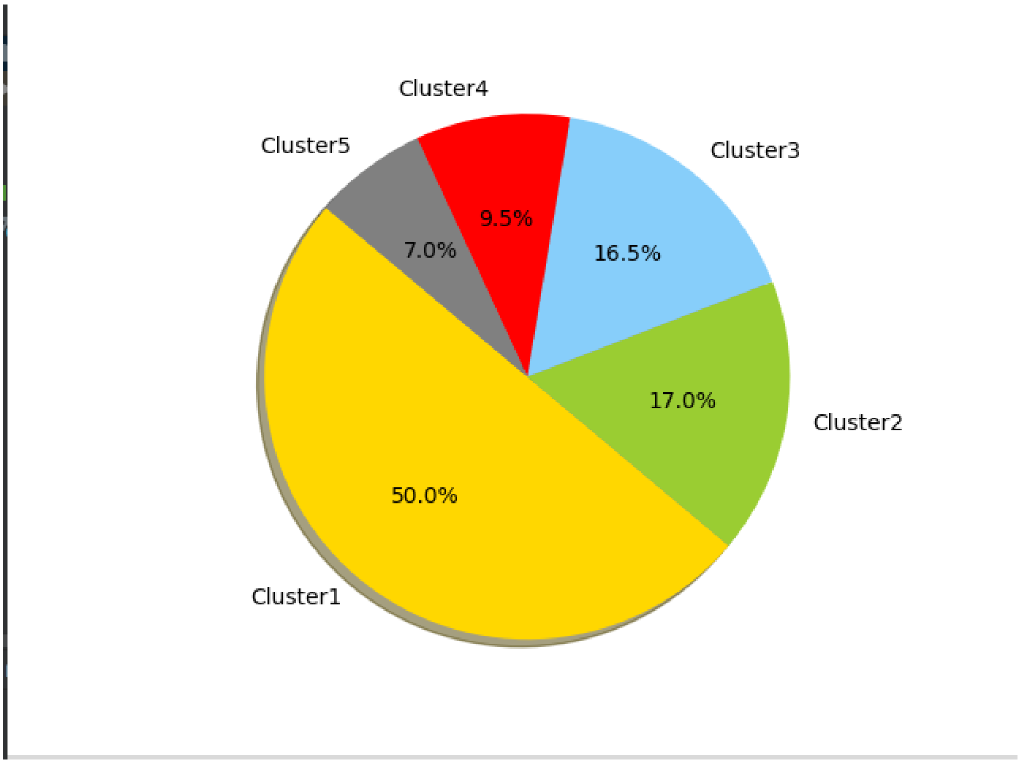


Table 2: Experiments and scores

| # | kValues | # of instances | Standard Deviation | Silhouette Score | NMI | Rand Index Score |
|---|---|---|---|---|---|---|

|  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|
| 1 | 2 | 1: 100<br>2: 100 | 0.5 | 0.385114714356 | 0.6666 | 0.7173366834170855 |
| 2 | 3 | 1:100<br>2:68<br>3:32 | 0.6305553108173779 | 0.329479895829 | 0.636009796821069 | 0.7336180904522613 |
| 3 | 4 | 1:100<br>2:41<br>3:38<br>4:21 | 1.394238143216574 | 0.324990257431 | 0.6050196792788828 | 0.7422110552763819 |
| 4 | 5 | 1:100<br>2:34<br>3:33<br>4:19<br>5:14 | 1.1643023662262306 | 0.341736552252 | 0.6086107903399828 | 0.7426633165829146 |
| 5 | 6 | 1:100<br>2:34<br>3:32<br>4:17<br>5:13<br>6:4 | 1.96 | 0.250690178793 | 0.5211188574332057 | 0.774070351758794 |

Since we can say from the table, the biggest silhouette score is occured when we use k=2 which represents the most suitable value is 2 Also, we can see that rand_index values are close to 1 which represents the clusterings on these algorithm is similar to each others.