MILESTONE 5: CLUSTERING

Link to Youtube: https://youtu.be/umhR3IiHTE8

Link to GitHub: https://github.com/WQD170093/Data_Mining-MilestoneProject

Introduction

The main goal of this project is to determine the impact of haze on the Singapore's tourism stock price data. The data collected included Pollutant Standard Index (PSI) data, stock prices, number of International tourist arrival, year on year Singapore GDP Growth Rate and the moving trend of stock prices. Clustering was applied on these data in order to find the similarity among the data and group them into clusters. Hierarchical clustering (Automatic) and K-mean clustering (User-specify) were applied using SAS Enterprise Miner.

Hierarchical clustering

Hierarchical clustering was applied on the dataset and 20 clusters were created (figure 1). Too many clusters were formed and the information by segments were too detailed to be interpreted. Details were shown in mean statistics table (table 1) and examples of some segment profiles were shown in figure 2.

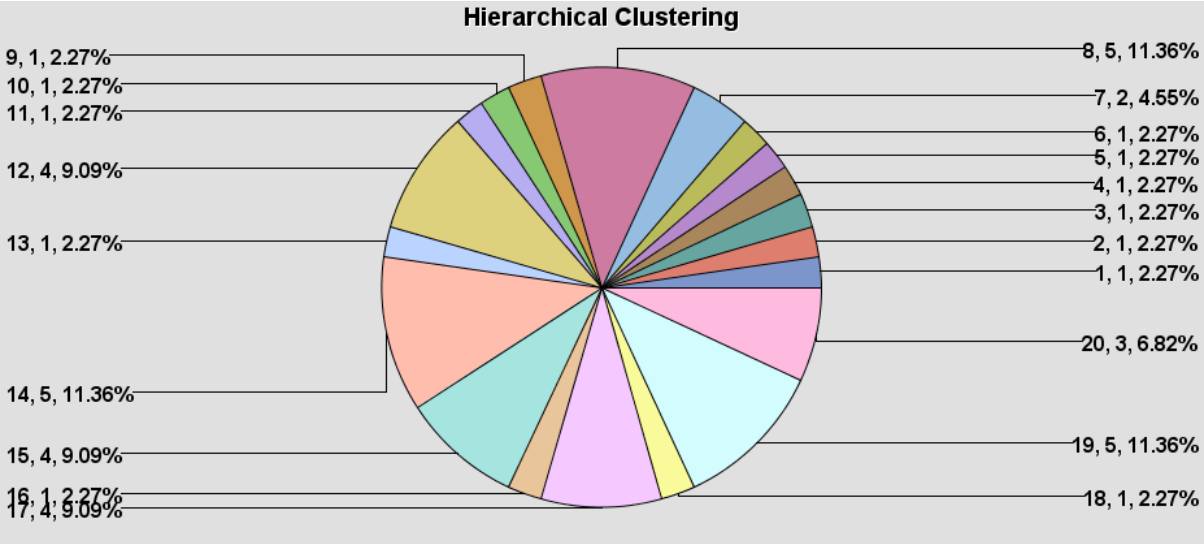

Figure 1 Twenty clusters were created using Hierarchical Clustering

Clustering Criterion	Maximum Relative Change in Cluster Seeds	Improvement in Clustering Criterion	Segment Id	Frequency of Cluster	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster	Average of PSI Avg	Average of Total International V	Average of Weighted Average Stoc	Trend	YOY GDP Growth Rate
0.1883...	0	0	1	1	.	0	2	1.7539...	-1.01589	-2.15051	-1.19746	1	-2.84
0.1883...	0	0	2	1	.	0	3	1.5024...	-1.02224	-2.147	-0.30891	1	-1.32777
0.1883...	0	0	3	1	.	0	4	1.1598...	-0.50465	-1.76196	0.6786...	1	-0.39717
0.1883...	0	0	4	1	.	0	3	1.1598...	-0.93864	-1.58956	1.1896...	1	0.5334...
0.1883...	0	0	5	1	.	0	7	1.49445	-0.91114	-1.52204	1.6673...	1	2.67382
0.1883...	0	0	6	1	.	0	5	2.0826...	-1.02641	-1.31999	1.65097	4.37E-17	3.2787...
0.1883...	0	0	7	2	0.3017...	0.47707	5	1.49445	-1.03893	-1.00912	2.6891...	1	1.7199...
0.1883...	0	0	8	5	0.3585...	1.11322	14	1.4242...	0.8562...	1.4921...	-1.01123	4.37E-17	-0.73839
0.1883...	0	0	9	1	.	0	10	1.5004...	-1.10226	-0.91948	2.1344...	4.37E-17	1.1383...
0.1883...	0	0	10	1	.	0	9	1.5004...	-0.88813	-0.74948	1.9660...	4.37E-17	-0.32737
0.1883...	0	0	11	1	.	0	19	1.2904...	-0.75539	-0.39553	0.7085...	4.37E-17	0.74282
0.1883...	0	0	12	4	0.2240...	0.6374...	15	0.6743...	0.4678...	0.8292...	-0.60715	1	-0.41462
0.1883...	0	0	13	1	.	0	14	1.6942...	2.4588...	0.4730...	-0.53792	4.37E-17	-0.23431
0.1883...	0	0	14	5	0.2492...	0.6895...	18	0.9542...	0.7845	0.2608...	-0.42638	4.37E-17	-0.33203
0.1883...	0	0	15	4	0.1531...	0.3169...	12	0.6743...	0.5401...	1.0763...	-0.09472	1	-0.05982
0.1883...	0	0	16	1	.	0	20	1.4599...	2.1908...	0.1593...	-0.41565	1	-0.56002
0.1883...	0	0	17	4	0.2572...	0.7047...	12	1.8301...	-1.043	-0.06905	-0.1733	1	-0.14707
0.1883...	0	0	18	1	.	0	14	0.9542...	-0.02579	0.1245...	-0.09404	4.37E-17	0.0216...
0.1883...	0	0	19	5	0.2440...	0.6345...	18	1.0428...	-0.99583	-0.07345	-0.3808	4.37E-17	0.1798...
0.1883...	0	0	20	3	0.2380...	0.5837...	12	1.1868...	0.8447...	-0.14487	-0.07577	1	-0.22656

Table 1 Mean Statics table of Hierarchical Clustering

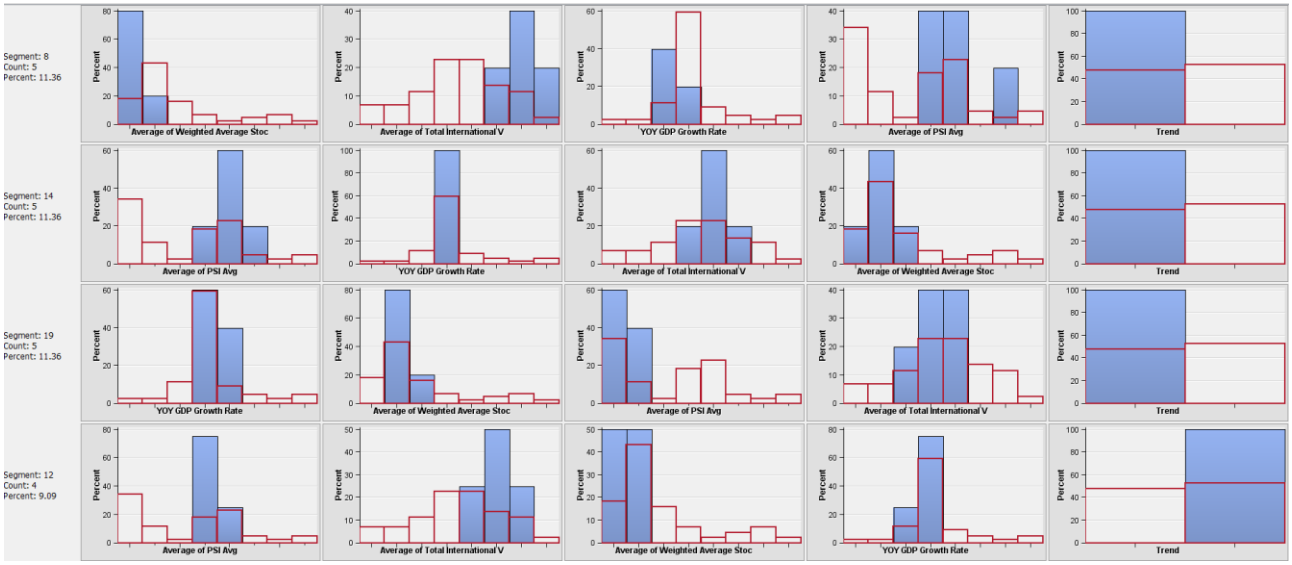

Figure 2 Examples of some segment profiles using Hierarchical Clustering

K-Mean Clustering

K-mean clustering with two clusters were applied on the dataset. Number of clusters was decided as two because we would like to check what's the pattern of the other variables when the weighted average stock prices were increased and what's the situation that caused the stock prices to decline. Figure 3 showed that 34.09% of the data belong to cluster 1 while 65.91% of the data belong to cluster 2. Table 2 and figure 4 indicated the segment profile of each cluster. Segment profile of cluster 1 revealed that the number of tourist arrival was lower than average and the PSI readings were low. However, the tourisms' stock prices were higher (the graph skewed to the right) and there were more increasing trend than the decreasing trend in the trend variable profile. The GDP growth rate was either in extremely high value or extremely low value in cluster 1. On the other hand, in cluster 2, the tourisms' stock prices were lower (the graph skewed to the left) and there were more decreasing trend. PSI data and number of tourist arrival were higher than average level. GDP growth rate was centralized at the average level.

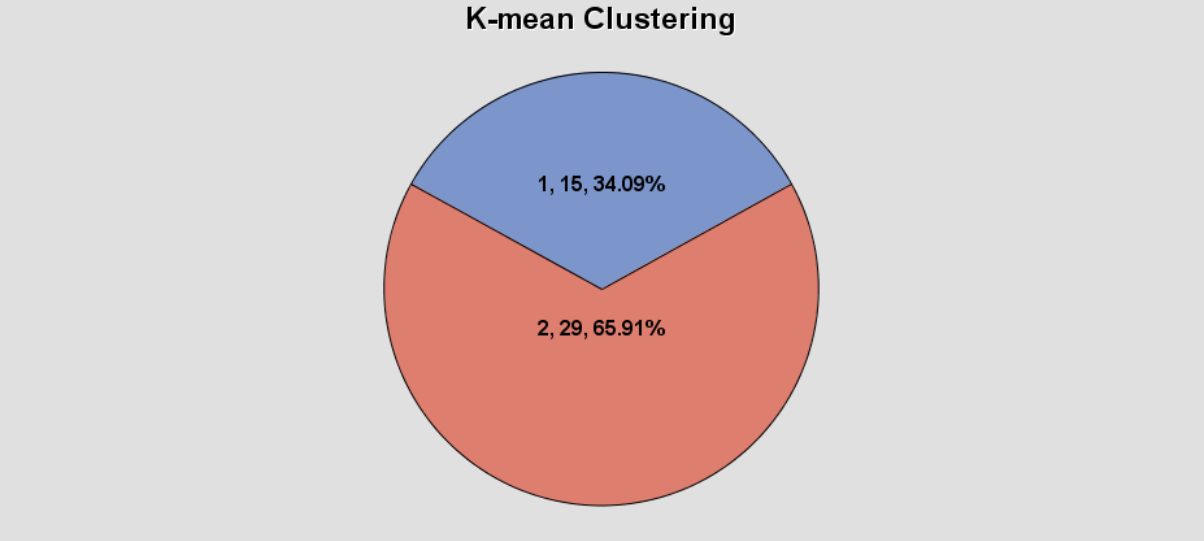

Figure 3 Two clusters were created using K-mean clustering

Clustering Criterion	Maximum Relative Change in Cluster Seeds	Improvement in Clustering Criterion	Segment Id	Frequency of Cluster	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster	Average of PSI Avg	Average of Total International V	Average of Weighted Average Stoc	Trend	YOY GDP Growth Rate
0.8028...	0	0	1	15	1.0356...	4.1239...	2	2.7088...	-1.01173	-1.02027	0.8856...	0.6666...	0.4698...
0.8028...	0	0	2	29	0.6835...	2.5211...	1	2.7088...	0.52331	0.5465...	-0.45809	0.4482...	-0.26102

Table 2 Mean Statics table of K-mean Clustering

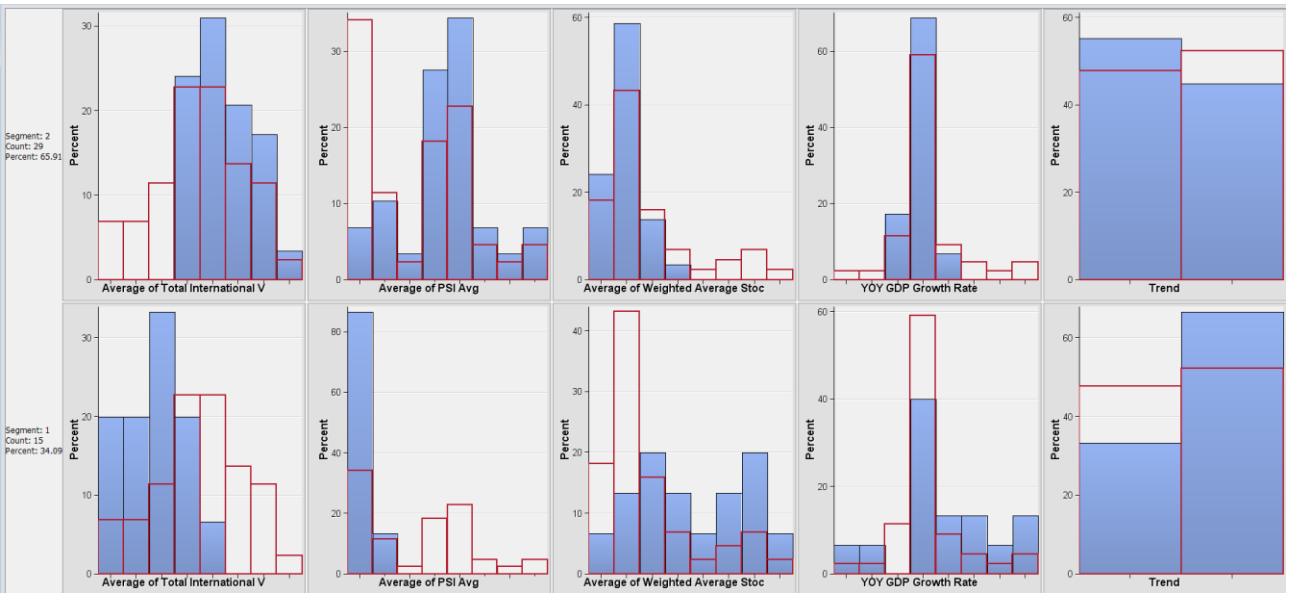

Figure 4 Segment profiles using K-mean Clustering