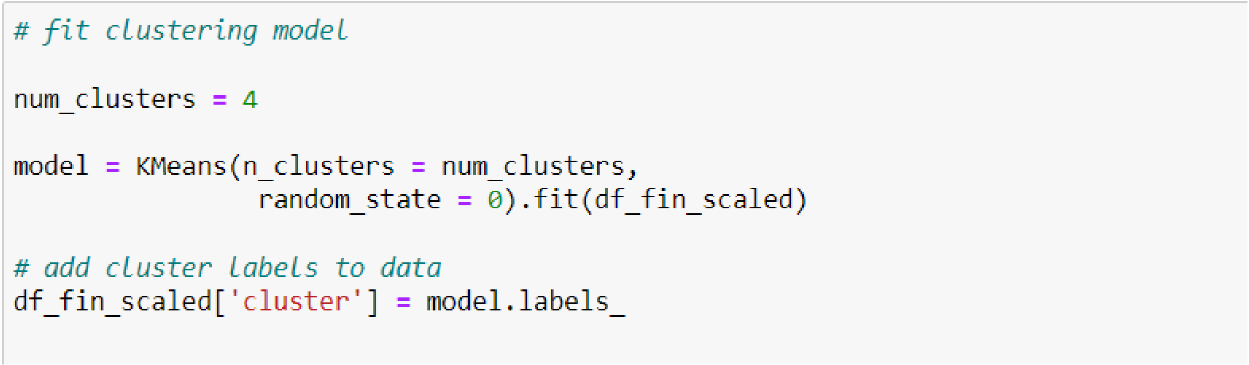
Hands-on Practice Session 2

1. Main Problem

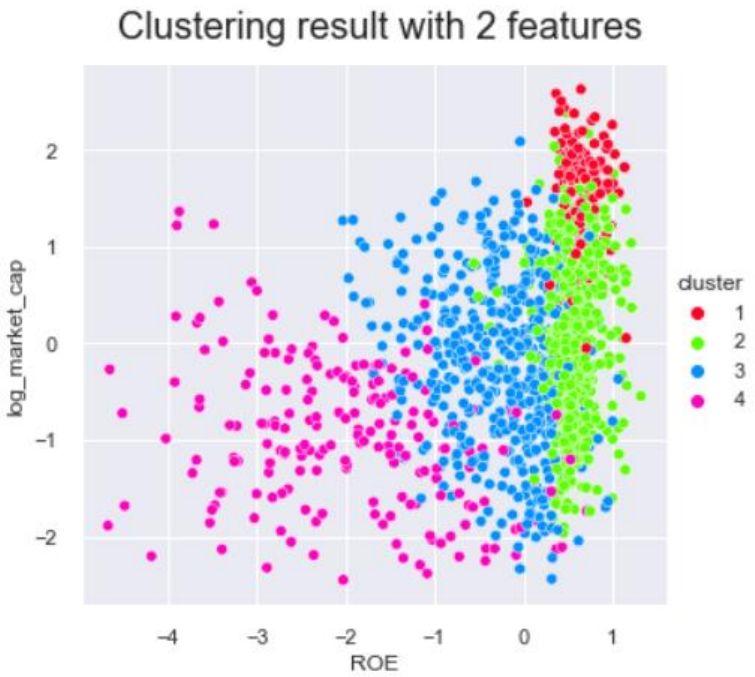
1.2 Clustering

We chose cluster number equal to 4 among the suggested choices (4, 5, and 7), and random state equal to 0.

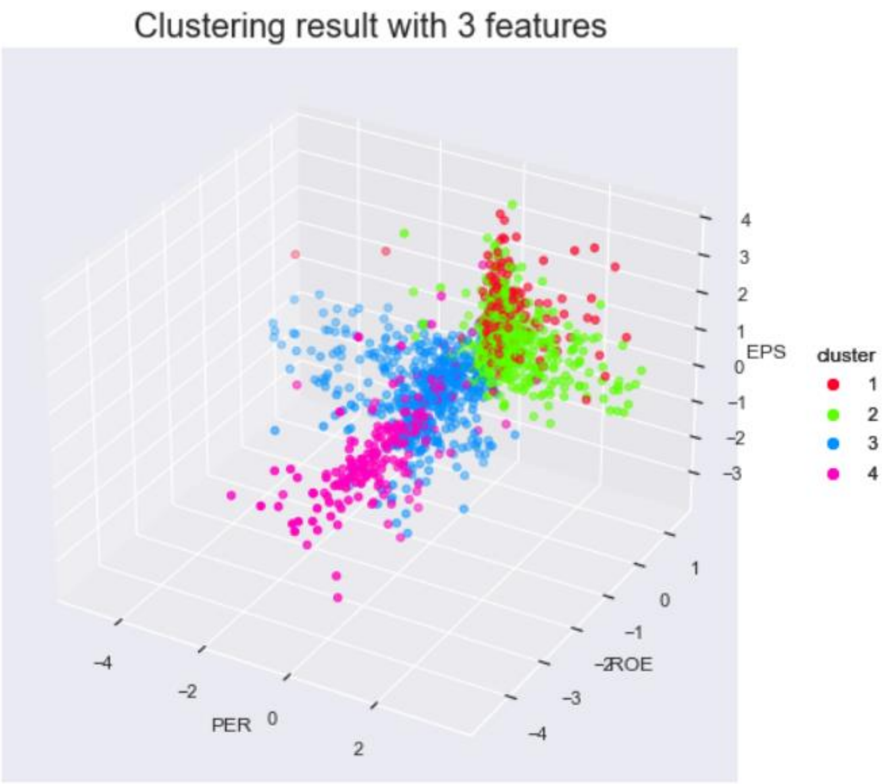


```
# fit clustering model

num_clusters = 4

model = KMeans(n_clusters = num_clusters,
                random_state = 0).fit(df_fin_scaled)

# add cluster labels to data
df_fin_scaled['cluster'] = model.labels_
```

1.3 Clustering Results



We can see that companies in the 4 clusters are well separated in terms of their market capitalization (in log form) and ROE (return on equity). From the plot, it is clear that companies having high ROE also tend to have relatively higher market capitalization than the average.

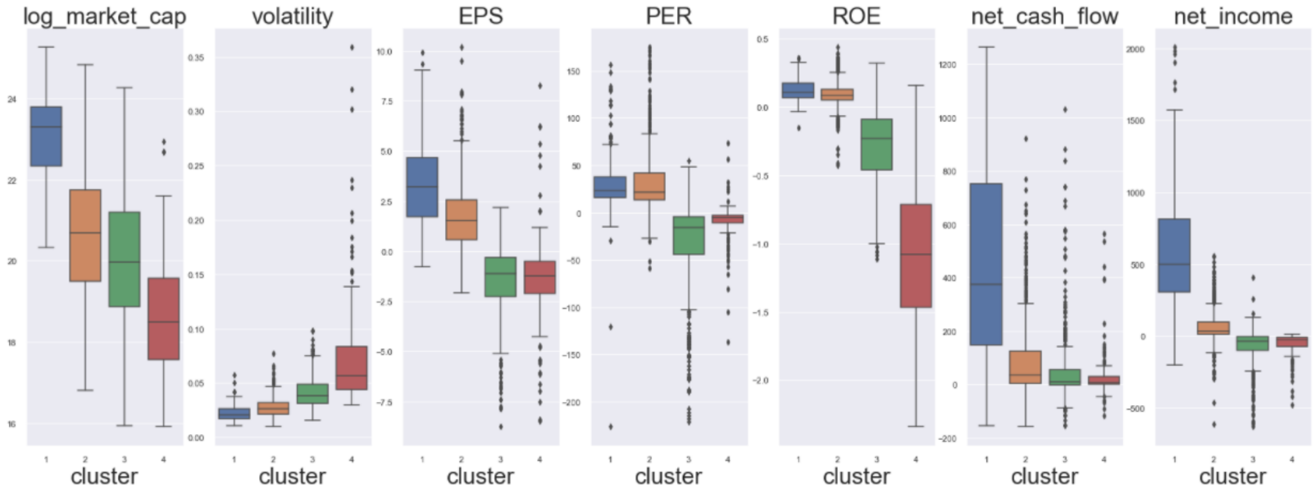


In terms of EPS, ROE and PER, companies in the clusters visualized in the above 3-dimensional graph are also well grouped. Those in cluster 1 tend to have high EPS (earnings per share), high ROE (return on equity) and average PER (price earnings ratio). Those in cluster 2 also have similar characteristics as companies in cluster 1, however they have relatively lower EPS. Firms in cluster 3 have average scores in all three financial metrics, while firms in cluster 4 tend to have less than average ROE.

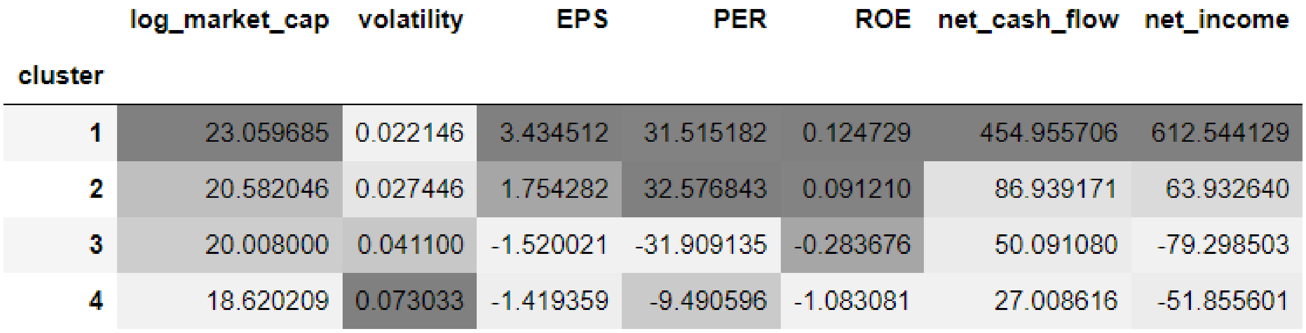
1.4 Analysis

Part A

Boxplot of Financial Properties Between Clusters



Mean Table

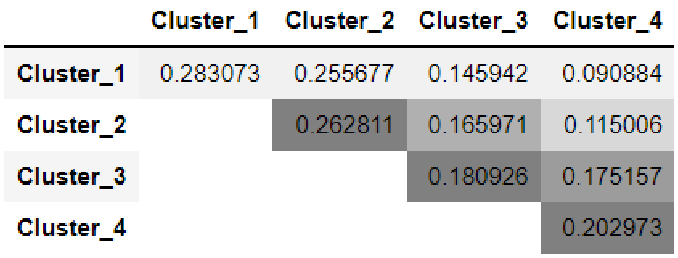


cluster	log_market_cap	volatility	EPS	PER	ROE	net_cash_flow	net_income
1	23.059685	0.022146	3.434512	31.515182	0.124729	454.955706	612.544129
2	20.582046	0.027446	1.754282	32.576843	0.091210	86.939171	63.932640
3	20.008000	0.041100	-1.520021	-31.909135	-0.283676	50.091080	-79.298503
4	18.620209	0.073033	-1.419359	-9.490596	-1.083081	27.008616	-51.855601

We can see from the boxplot and the mean table, that financial properties of firms that have highest market capitalization (cluster 1) are the most satisfactory, meaning that those firms have low volatility in their returns, high EPS, ROE, net income and net cash flow relative to firms in other clusters. Generally, the financial metrics within clusters are very similar to each other, while they show considerably high differences across clusters (their distributions in terms of mean and variance are different across clusters).

Part B

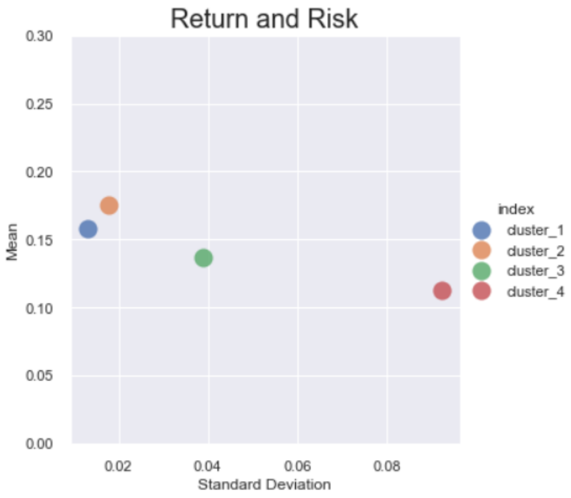
Correlation Table



	Cluster_1	Cluster_2	Cluster_3	Cluster_4
Cluster_1	0.283073	0.255677	0.145942	0.090884
Cluster_2		0.262811	0.165971	0.115006
Cluster_3			0.180926	0.175157
Cluster_4				0.202973

We can see that within cluster correlations tend to be higher than between cluster correlations small values, which is a desirable quality.

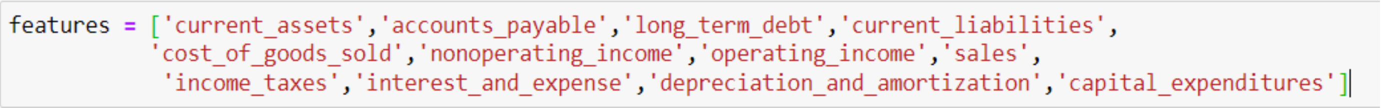
Scatterplot



As can be seen from the scatterplot, the return and risk factors of companies located in different clusters show some differences, although risk and return characteristics of firms in cluster 1 and 2 are quite close to each other.

2. Extension Problem

2.1 Selecting variables and pre-processing data

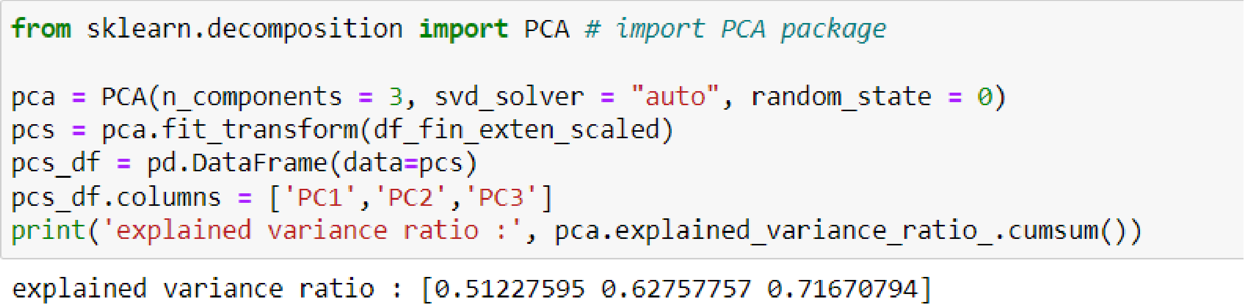


```
features = ['current_assets','accounts_payable','long_term_debt','current_liabilities',
            'cost_of_goods_sold','nonoperating_income','operating_income','sales',
             'income_taxes','interest_and_expense','depreciation_and_amortization','capital_expenditures']
```

We chose the above features for our analysis. In the code file, we tried to see the distribution of each feature, and we found that all features, except features such as **nonoperating_income**, **operating_income** and **income_taxes**, are highly skewed to the right, and thus we applied log transformation on those features. We also removed outliers from all chosen variables.

2.2 Clustering

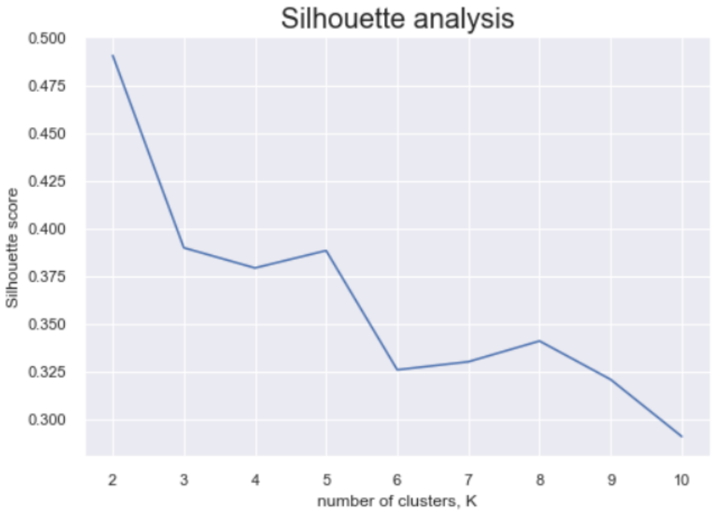
PCA



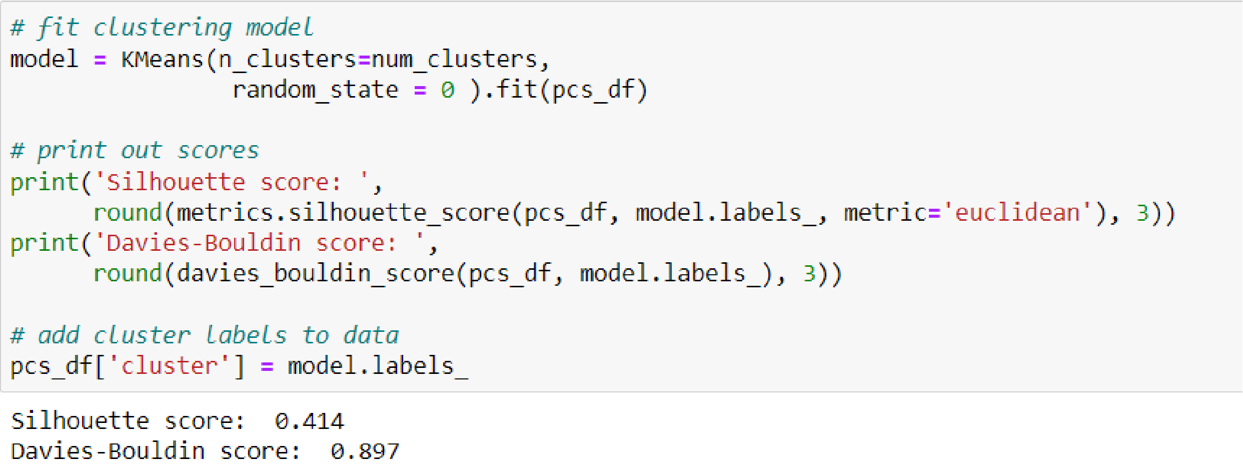
```
from sklearn.decomposition import PCA # import PCA package

pca = PCA(n_components = 3, svd_solver = "auto", random_state = 0)
pcs = pca.fit_transform(df_fin_exten_scaled)
pcs_df = pd.DataFrame(data=pcs)
pcs_df.columns = ['PC1','PC2','PC3']
print('explained variance ratio :', pca.explained_variance_ratio_.cumsum())
```

```
explained variance ratio : [0.51227595 0.62757757 0.71670794]
```

We reduced our data using PCA with 3 principal components.



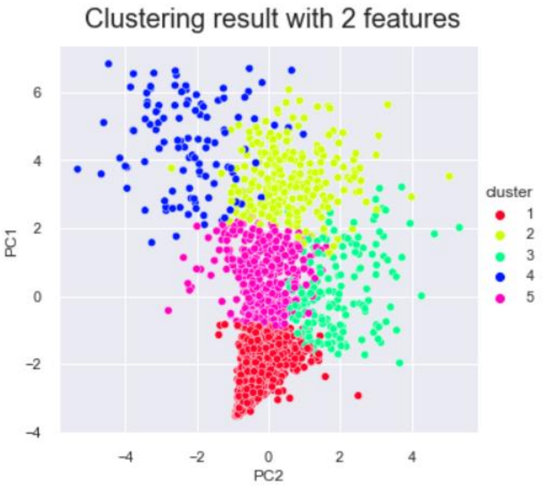
We used Silhouette clustering metric for our data set. Based on the graph, we decided to choose 5 as the number of clusters in our dataset since it had highest Silhouette score between other possible choices such k=3 and k = 7.



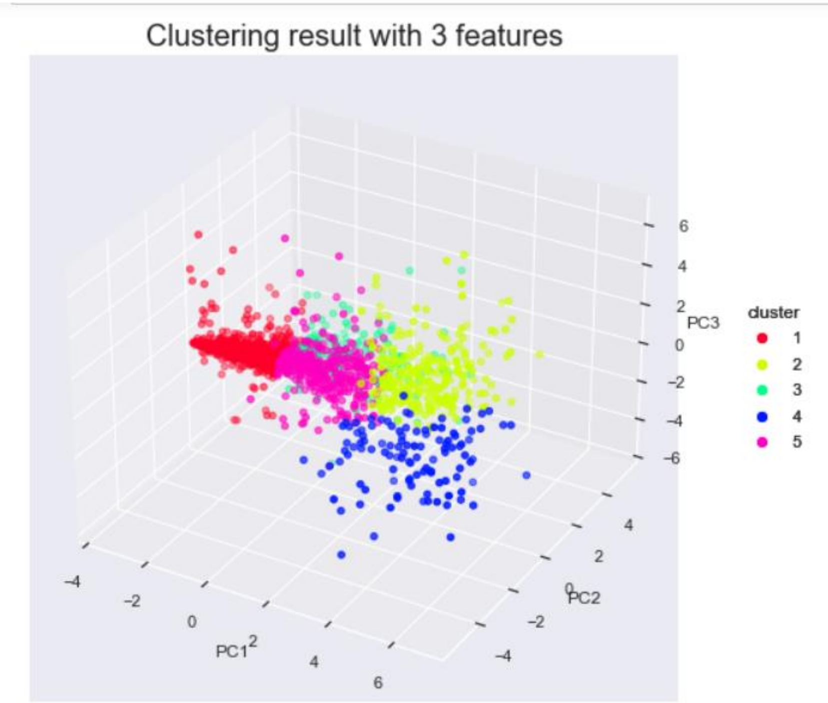
```
# fit clustering model
model = KMeans(n_clusters=num_clusters,
                random_state = 0 ).fit(pcs_df)

# print out scores
print('Silhouette score: ',
      round(metrics.silhouette_score(pcs_df, model.labels_, metric='euclidean'), 3))
print('Davies-Bouldin score: ',
      round(davies_bouldin_score(pcs_df, model.labels_), 3))

# add cluster labels to data
pcs_df['cluster'] = model.labels_
```

```
Silhouette score:  0.414
Davies-Bouldin score:  0.897
```

2.3 Clustering results



Clustering result with 2 features



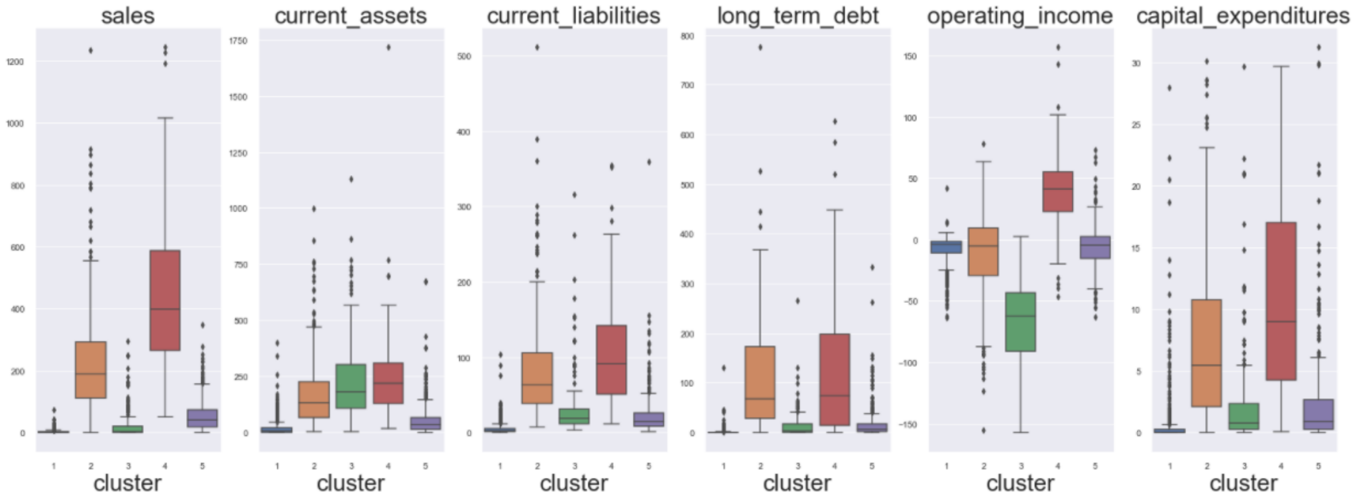
Clustering result with 3 features

We can see from graphs above that clusters in both 2 dimensions and 3 dimensions are separated clearly.

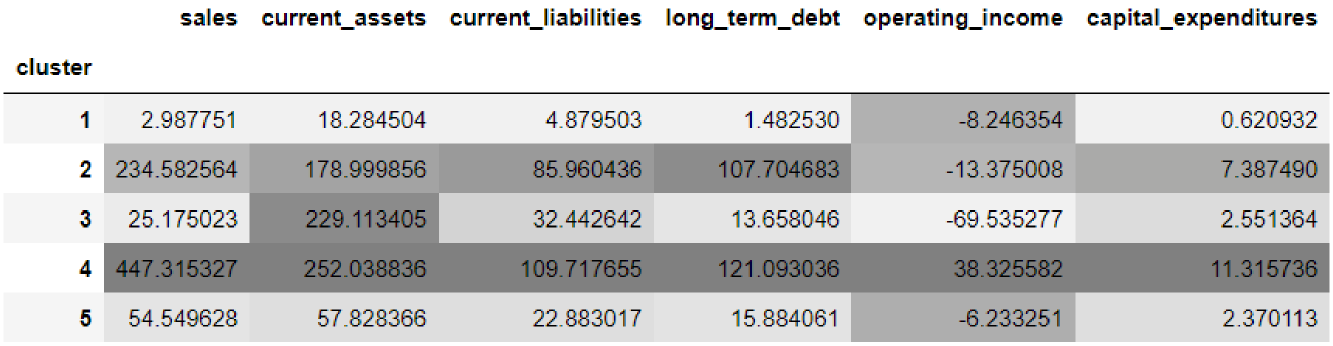
2.4 Analysis

A.

Boxplot



Mean Table

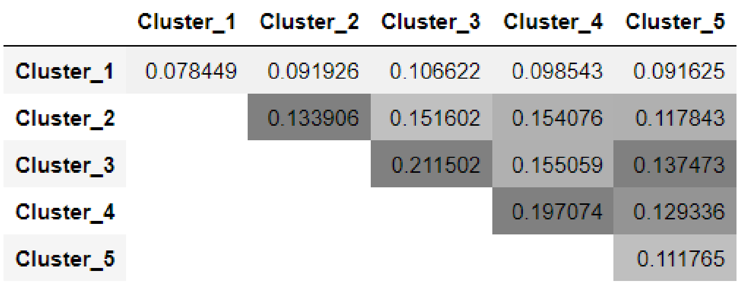


cluster	sales	current_assets	current_liabilities	long_term_debt	operating_income	capital_expenditures
1	2.987751	18.284504	4.879503	1.482530	-8.246354	0.620932
2	234.582564	178.999856	85.960436	107.704683	-13.375008	7.387490
3	25.175023	229.113405	32.442642	13.658046	-69.535277	2.551364
4	447.315327	252.038836	109.717655	121.093036	38.325582	11.315736
5	54.549628	57.828366	22.883017	15.884061	-6.233251	2.370113

By looking at the boxplot and the mean table, we can point out two things: 1) firms in cluster 4 have high mean values in all chosen features relative to other clusters, 2) only firms in cluster 3 seems to have negative mean value in terms of operating income, which might mean that those companies are performing poorly. Overall, firms in different clusters tend not to have similar mean values among the chosen features, which is what we desire.

B.

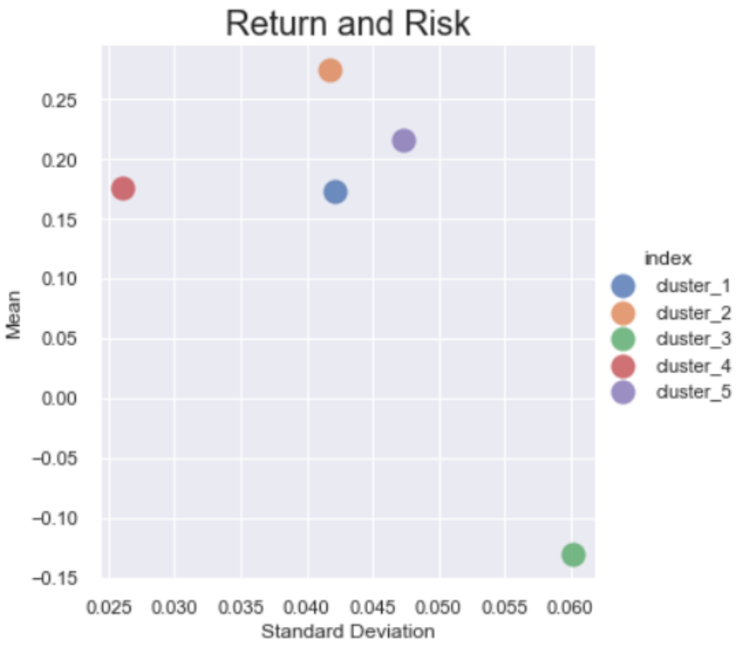
Correlation Table



	Cluster_1	Cluster_2	Cluster_3	Cluster_4	Cluster_5
Cluster_1	0.078449	0.091926	0.106622	0.098543	0.091625
Cluster_2		0.133906	0.151602	0.154076	0.117843
Cluster_3			0.211502	0.155059	0.137473
Cluster_4				0.197074	0.129336
Cluster_5					0.111765

Correlations of stock returns within clusters seems to be pretty small (except for cluster 3 and cluster 4), and the difference between within-cluster correlations and among-cluster correlations is also quite low.

Scatterplot



Firms in cluster 2 have the highest mean return compared to firms in other clusters, while firms in cluster 4 have the smallest mean standard deviation. A surprising result is that firms in cluster 3 have a negative mean return, which is quite high. So, we might avoid having those firms in our portfolio.