Project: MTL Tladi Accountants Presentation

Situation

This document serves as the *first* executive summary for the data provided by ABC (Pty) Ltd, a tax and accounting service firm based in the US. They make some of their data publicly available on some websites.

The business presented a clear question; *we want to understand our customer base, which customers stand out and what make them stand out. Use market capitalization as a measure of a company's performance.*

Data

The data provided only has cash flows for each client up to 10 years into the past; these are cash flows from financing, operations, investing, free cash, net cash flows and cash equivalents at the end of each year. There are three additional factors which are the industry the client operates in, the market capitalization score of each client and the current price at which the client's company can be sold.

The original data has 1668 clients and 110 different industries, with 36 categories of information provided about each client.

Limitations

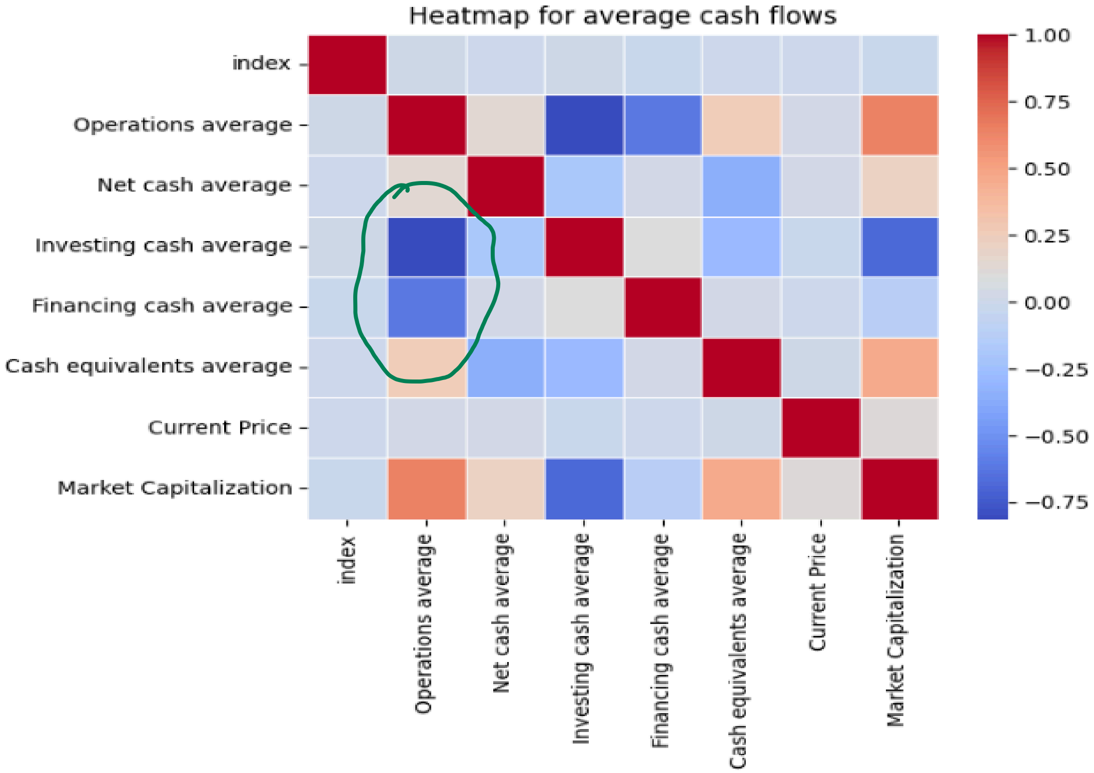
The data at hand only includes income cash flow and does not provide information on liabilities and assets, therefore the analysis is only focused on income cash flows only. There is also no information regarding the business' dealings with each client, e.g., what services they offer to each client or how much they make from the client.

Data for some cash flows is provided only for two years while some is for 6 years, causing skewness in the data. This information is crucial in making useful and more reliable insights and thus limit the scope of this project.

Main results

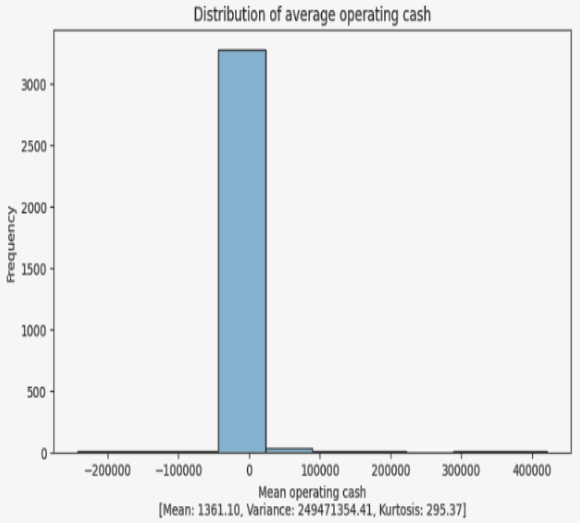
1. Heatmap

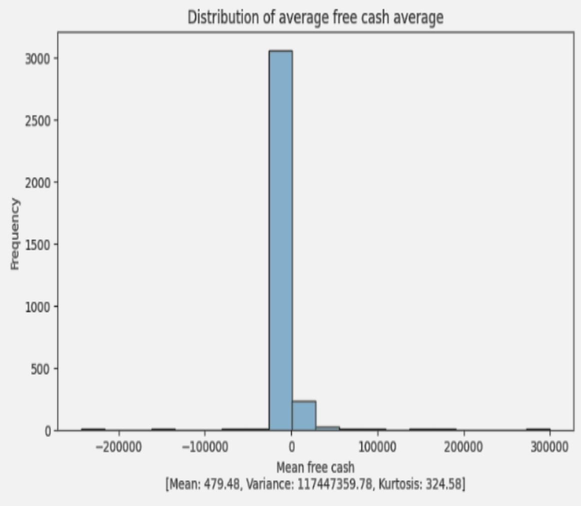
- This shows how each attribute or cash flow is related to the other types of cash flows and important relationships can be deduced from it.
- To read this, the blue values show an inverse relationship and brown shows direct. For example, the blue shade between operations average and investing cash average mean that a client with ***high*** cash flows from operations is likely to have ***low*** cash flows from investing.

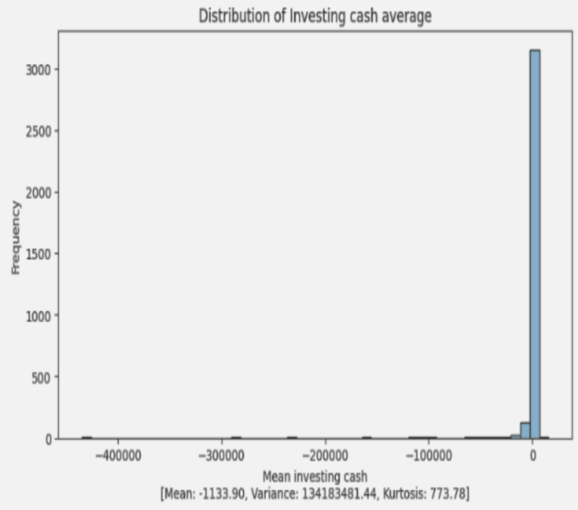


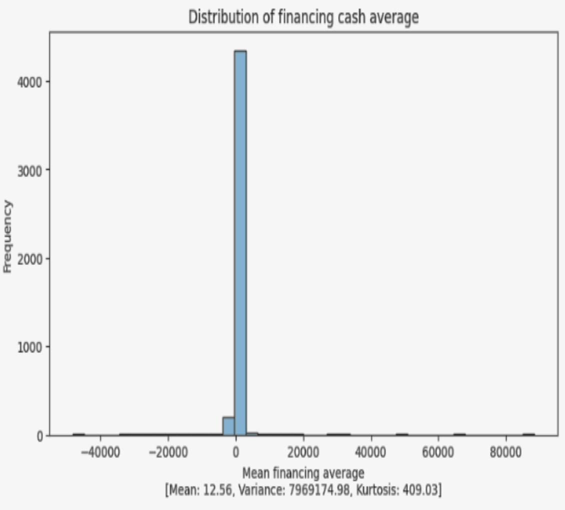
2. Distributions

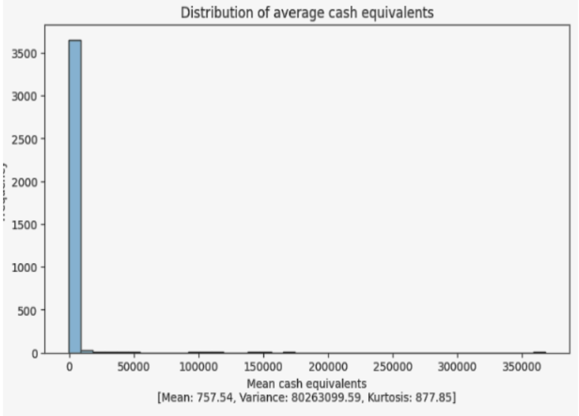
- These show how the clients are distributed around the mean for each type of cashflow. This helps locate where most of the clients fall in each category of cashflow.

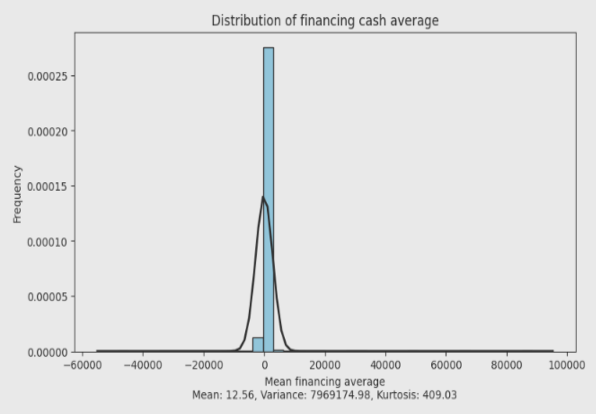






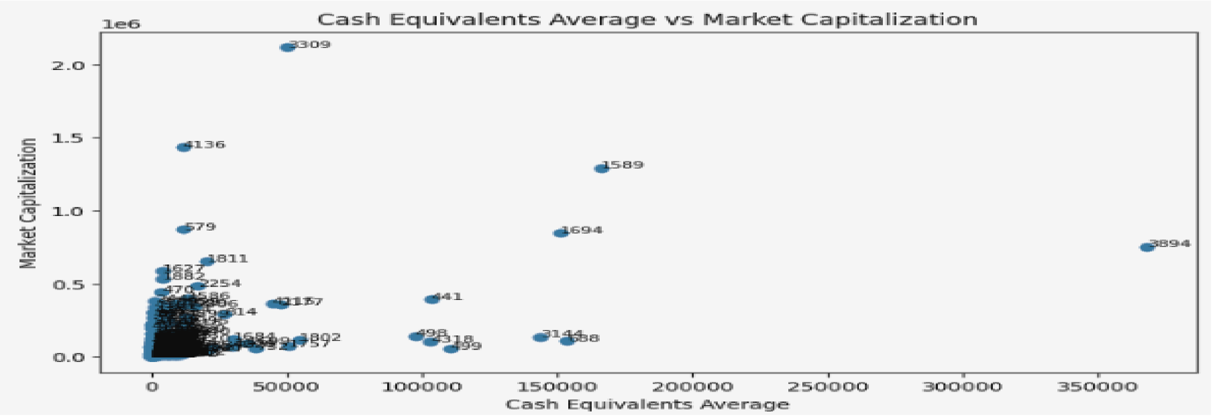


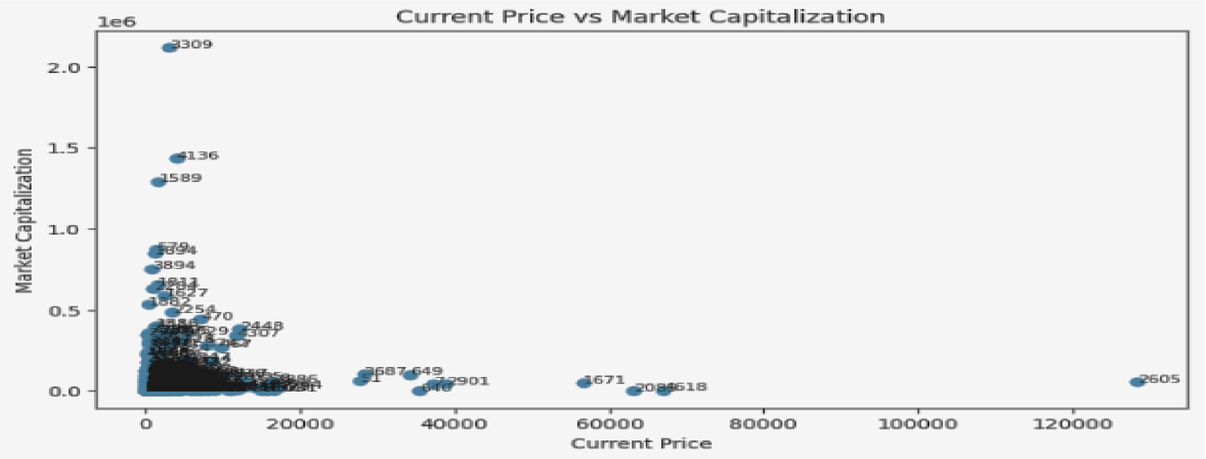


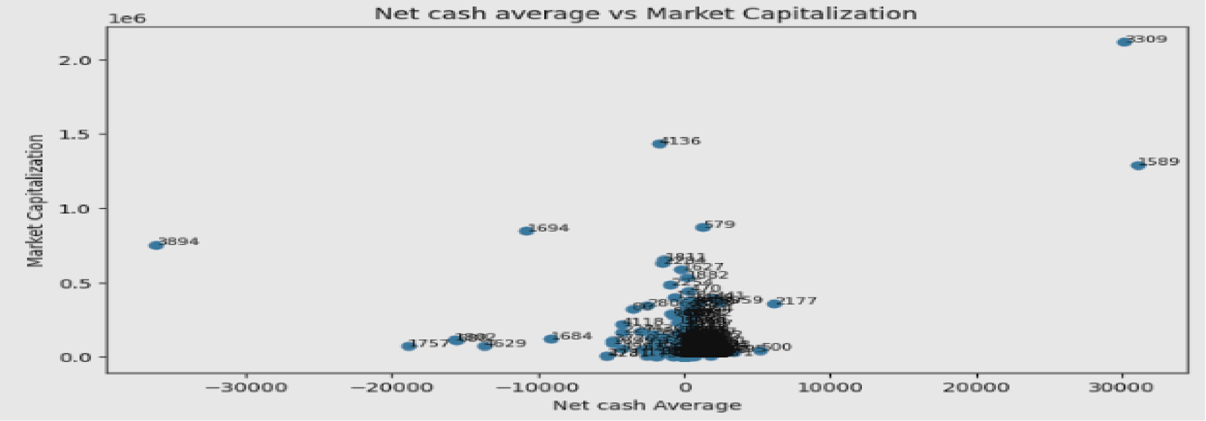


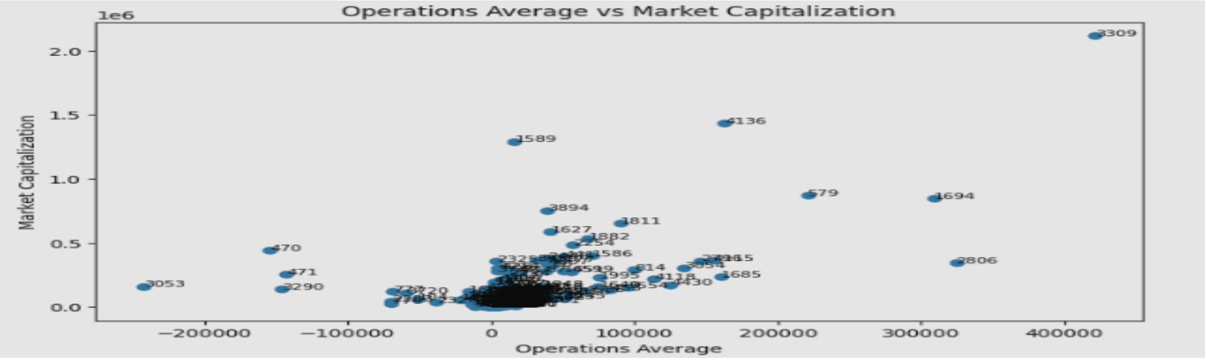
3. Scatter plots

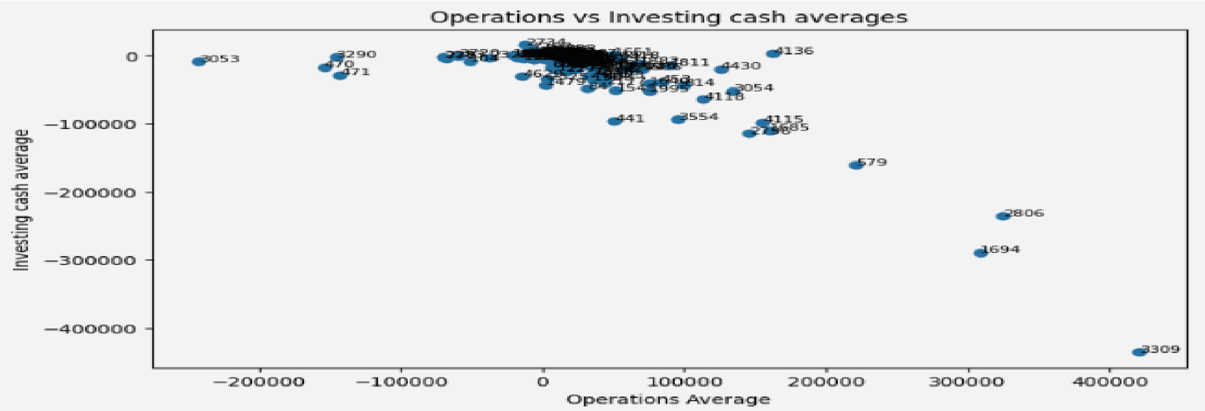
- These are great at identifying outliers since the average client will be clustered around a central point while the companies that stand out will be further away from that central point.

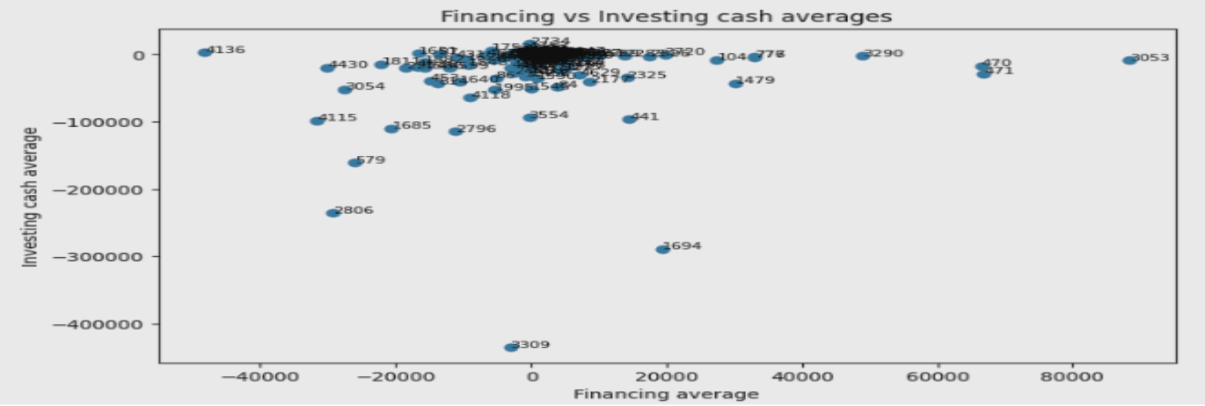








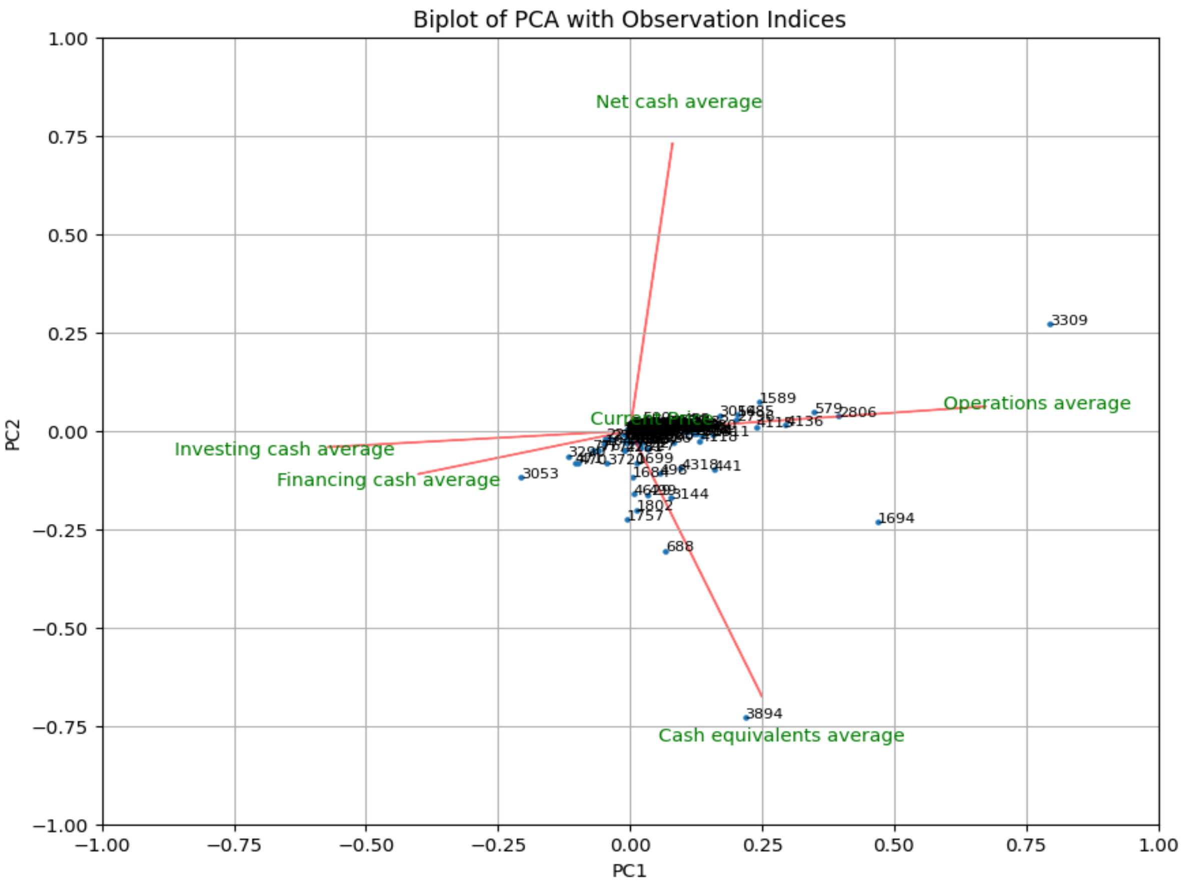

Operations vs Investing cash averages
Investing cash average
Operations Average
0
−100000
−200000
−300000
−400000
−200000
−100000
0
100000
200000
300000
400000
3053
3290
470
471
2734
1651
3811
4430
4136
3054
4118
441
3554
4115
1685
579
2806
1694
3309


Financing vs Investing cash averages
Investing cash average
Financing average
0
−100000
−200000
−300000
−400000
−40000
−20000
0
20000
40000
60000
80000
4136
4430
1811
1651
3054
4118
2734
2325
1479
3290
470
471
3053
4115
1685
2796
3554
441
579
2806
1694
3309

4. Model 1: PCA biplot

- This plot is made from an *svd* model and also creates a 2D-plot of the data. The first two components explain 69% of the variation in the data.

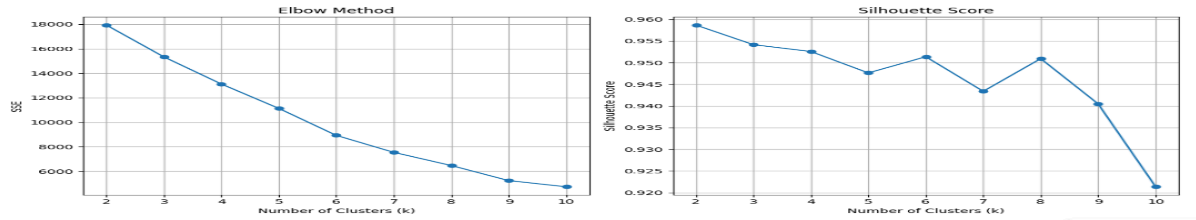
- The d-matrix from the ***svd***: [0.4612 ; 0.2262 0.1893 0.0809 0.0400 0.0022]

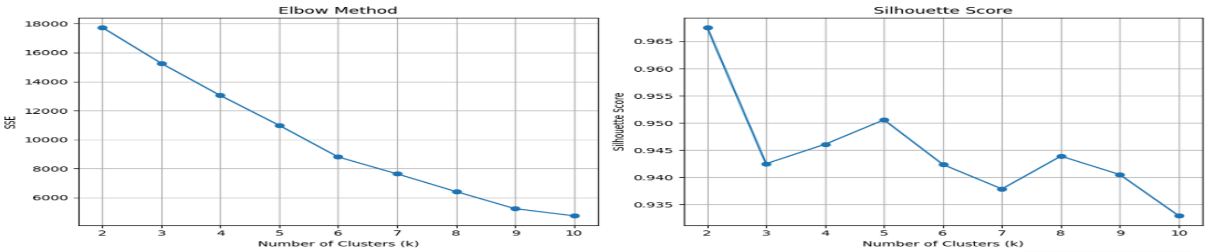


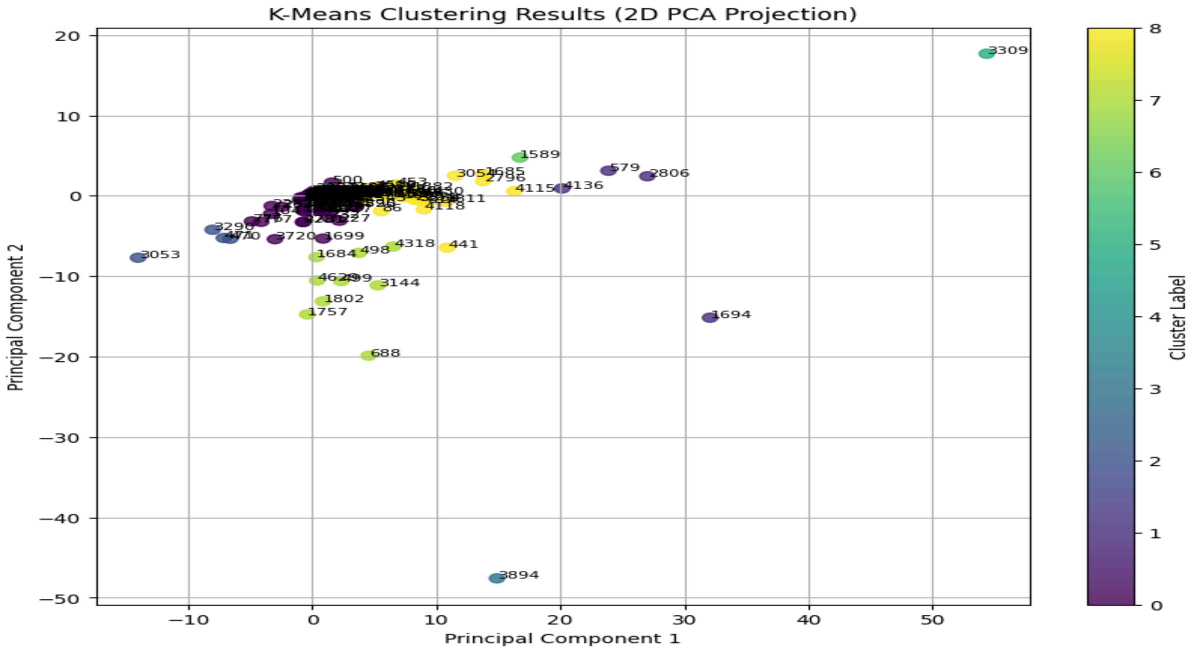
-

5. K-means Clustering Plot

- This is an unsupervised model which groups the data into multiple groups depending on how many groups the user desires. This model is build up from pca biplot. The plot provided uses 10 groups.

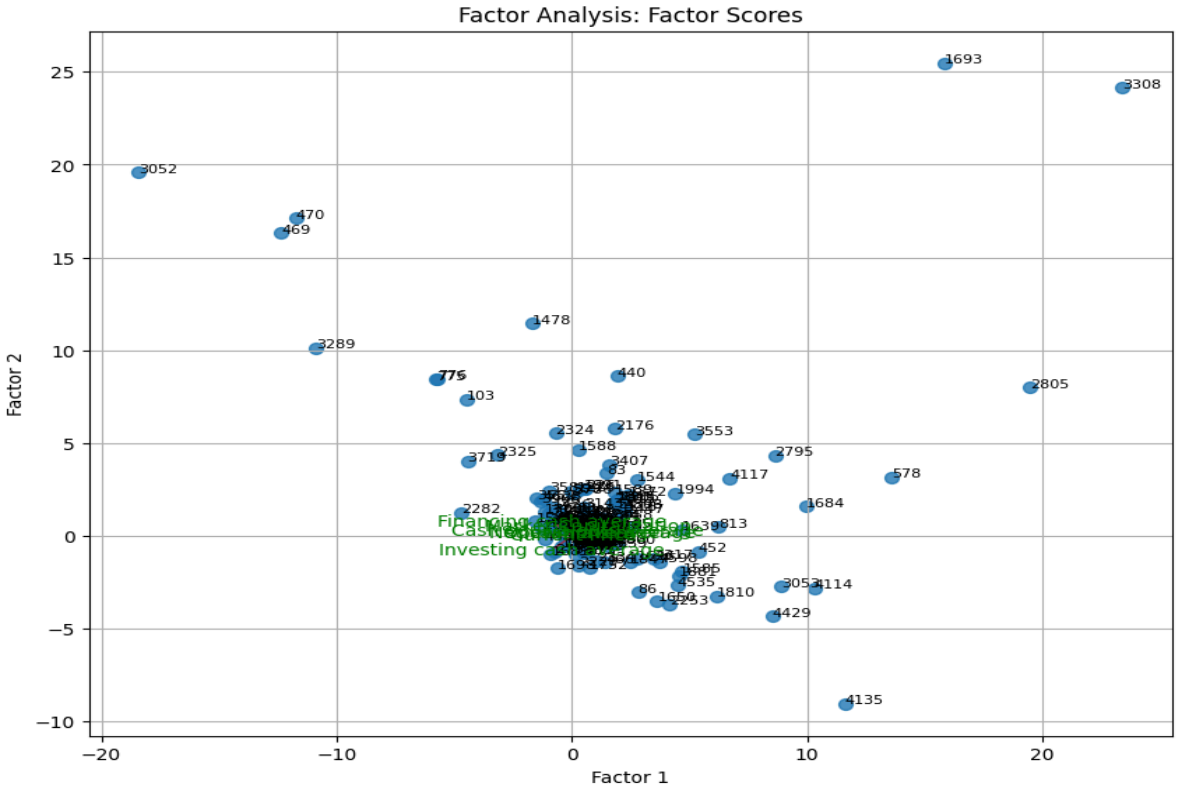






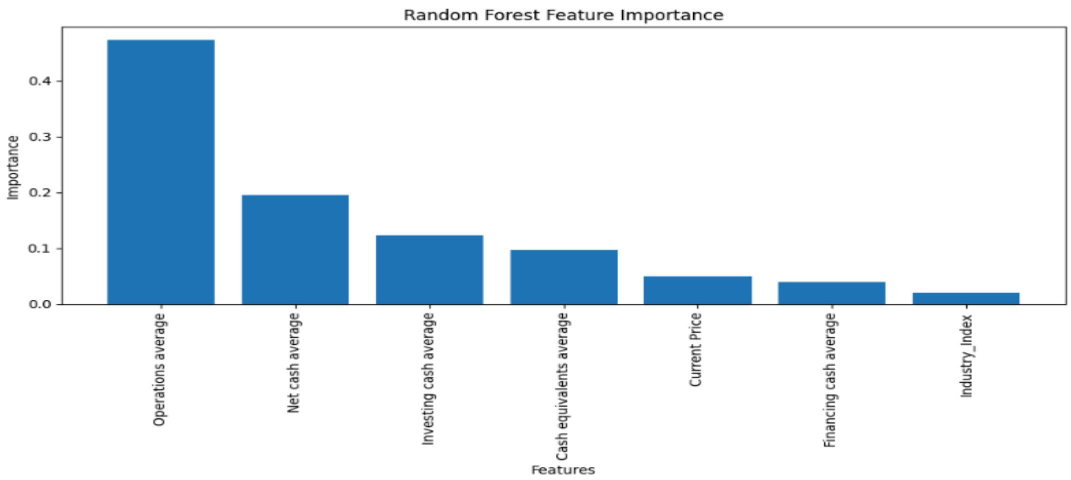
6. Factor Analysis biplot

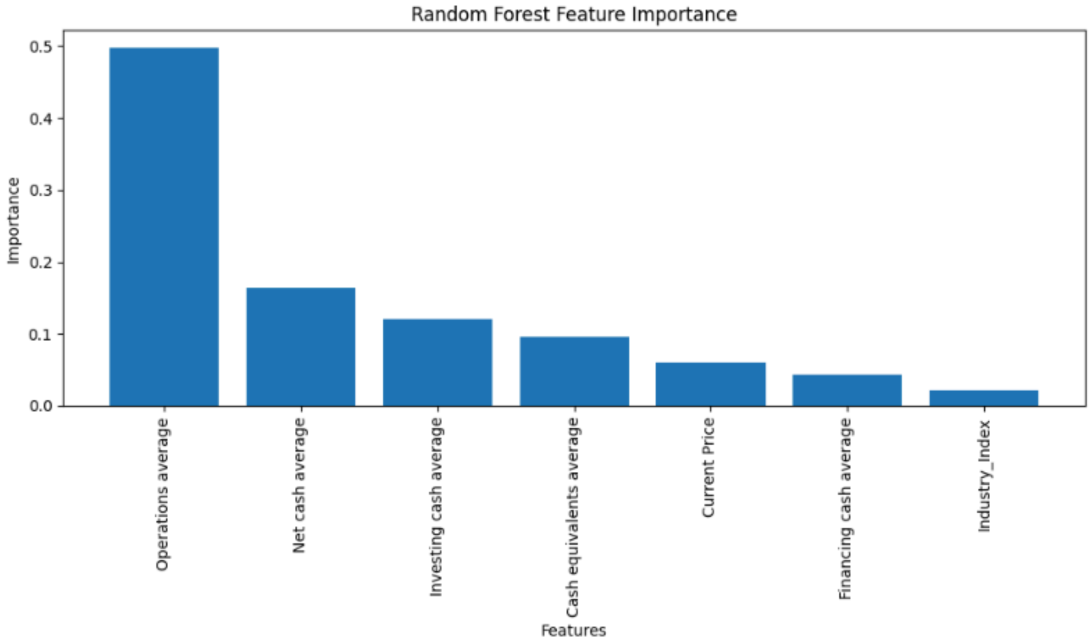
- Another plot produced from ***svd*** which reduces the dimension of the data and produces a 2D plot which can show which clients deviate from average.

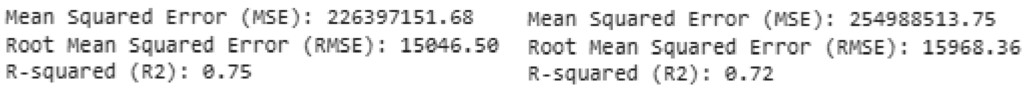


7. Random Forest Feature importance

- Useful for rating percentage of *mse* explained by each variable. The plots are from two models with different parameters.





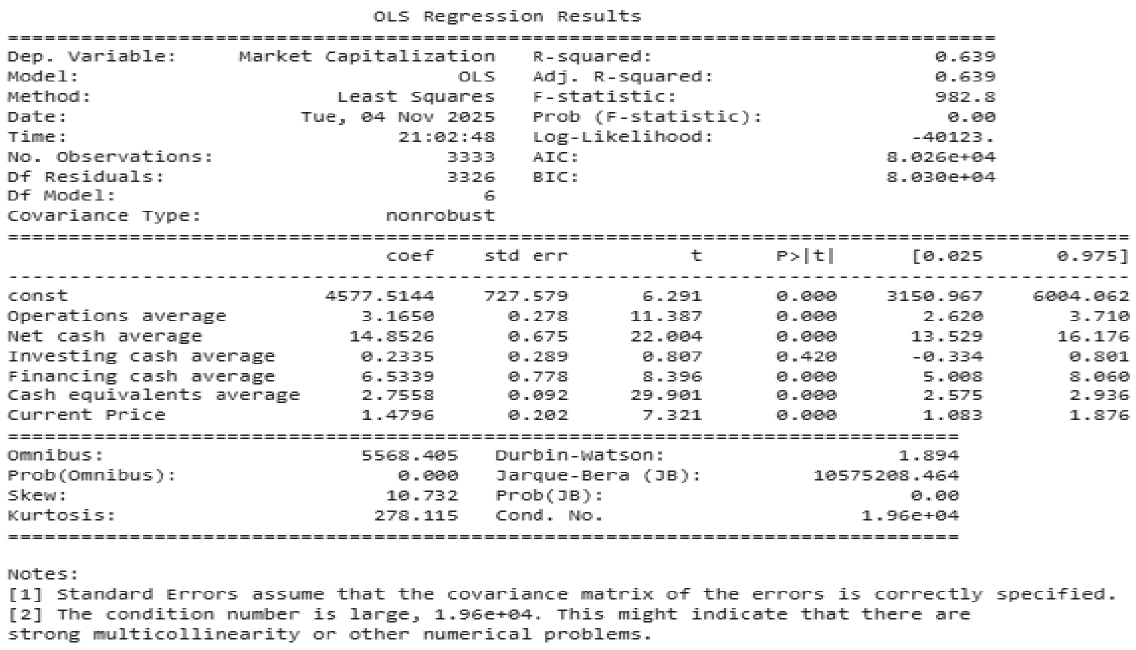


```
Mean Squared Error (MSE): 226397151.68
Root Mean Squared Error (RMSE): 15046.50
R-squared (R2): 0.75
```

```
Mean Squared Error (MSE): 254988513.75
Root Mean Squared Error (RMSE): 15968.36
R-squared (R2): 0.72
```

8. Multiple Linear Regression

- Model diagnostics or summary is useful in establishing relationships and in predicting if a client will have high or low market capitalization. This model is powerful in inference, determining how client attributes are related to market capitalization.



```
                              OLS Regression Results
==================================================================================
Dep. Variable:     Market Capitalization   R-squared:                        0.639
Model:                               OLS   Adj. R-squared:                   0.639
Method:                    Least Squares   F-statistic:                      982.8
Date:                   Tue, 04 Nov 2025   Prob (F-statistic):                0.00
Time:                           21:02:48   Log-Likelihood:                 -40123.
No. Observations:                   3333   AIC:                          8.026e+04
Df Residuals:                       3326   BIC:                          8.030e+04
Df Model:                              6
Covariance Type:               nonrobust
=============================================================================================
                                coef    std err          t      P>|t|      [0.025       0.975]
---------------------------------------------------------------------------------------------
const                      4577.5144    727.579      6.291      0.000    3150.967     6004.062
Operations average            3.1650      0.278     11.387      0.000       2.620        3.710
Net cash average             14.8526      0.675     22.004      0.000      13.529       16.176
Investing cash average        0.2335      0.289      0.807      0.420      -0.334        0.801
Financing cash average        6.5339      0.778      8.396      0.000       5.008        8.060
Cash equivalents average      2.7558      0.092     29.901      0.000       2.575        2.936
Current Price                 1.4796      0.202      7.321      0.000       1.083        1.876
==============================================================================
Omnibus:                     5568.405   Durbin-Watson:                   1.894
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         10575208.464
Skew:                          10.732   Prob(JB):                         0.00
Kurtosis:                     278.115   Cond. No.                     1.96e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.96e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

9. Hypothesis testing

- One requires at least one undergraduate statistics but the results will be clearly communicated when the tests are done.