# ADVANCED DATA ANALYTICS PROJECT REPORT

---

## EXPLORATORY DATA ANALYSIS

Stock-Portfolio Data Analysis


GROUP MEMBERS:-

- K Balarajaiah,S20200010085
- E Jaswanth Krishna,S20200020257

## Abstract:

This report presents the results of a study that uses a multilinear regression model to analyse a stock portfolio dataset. The purpose of the study was to identify the key variables that impact the performance of the stock portfolios and to develop a model that can be used to predict future performance.The dataset included historical data from the US stock market, and the study analysed a range of variables, including market trends, company performance, and macroeconomic factors.The results of the study show that the multilinear regression model is an effective tool for analysing the stock portfolio dataset.

## Introduction:

The data set contains totally 4 periods of 5 years .Each year contains four quarters. Hence each period contains 20 quarters comprising 5 years data.The dataset contains a total period from 1990 to 2010 (20 years[1990/9/30] to [2010/6/30] ) of 80 quarters data.The dataset has 12 attributes of totally 315 rows.6 inputs(weights) => ' Large B/P ', ' Large ROE ', ' Large S/P ',' Large Return Rate in the last quarter ', ' Large Market Value ',' Small systematic Risk'.6 outputs =>'Annual Return', 'Excess Return','Systematic Risk', 'Total Risk', 'Abs. Win Rate','Rel. Win Rate'

About Dataset Attributes:-

- **Large B/P**: This refers to a company's book-to-price ratio, which is calculated by dividing the company's book value by its market price per share.
- **Large ROE**: This stands for return on equity, which is calculated by dividing a company's net income by its shareholder equity.
- **Large S/P**: This refers to a company's sales-to-price ratio, which is calculated by dividing the company's sales per share by its market price per share.
- **Large Return Rate in the last quarter**: This refers to a company's return on investment (ROI) in the most recent quarter, which is calculated by dividing the company's net income by its total investment.
- **Large Market Value**: This refers to the total value of a company's outstanding shares of stock, calculated by multiplying its stock price by the number of shares outstanding.
- **Small systematic risk**: This refers to the risk of an investment that cannot be diversified away, such as the risk associated with macroeconomic factors or events that affect the entire market.
- **Annual Return**: This refers to the annual rate of return earned on an investment, typically expressed as a percentage.
- **Excess Return**: This is the amount by which the return on an investment exceeds the return on a benchmark, such as a stock index or bond index.
- **Systematic Risk**: This refers to the risk associated with investing in the overall market or a particular market segment, as opposed to the risk associated with a particular company or asset.

- **Total Risk**: This is the total amount of risk associated with an investment, including both systematic and unsystematic risk.
- **Abs. Win Rate**: This measures the percentage of time that an investment outperforms its benchmark.
- **Rel. Win Rate**: This is the relative outperformance of an investment compared to its benchmark, expressed as a percentage.

## Problem Statement:

The goal of this project is to do exploratory data analysis on a stock portfolio performance dataset. We performed Linear Regression on the dataset and test all the assumptions of linear regression.

## Methodology:

### 1. Data Pre processing

**a. Null values:**

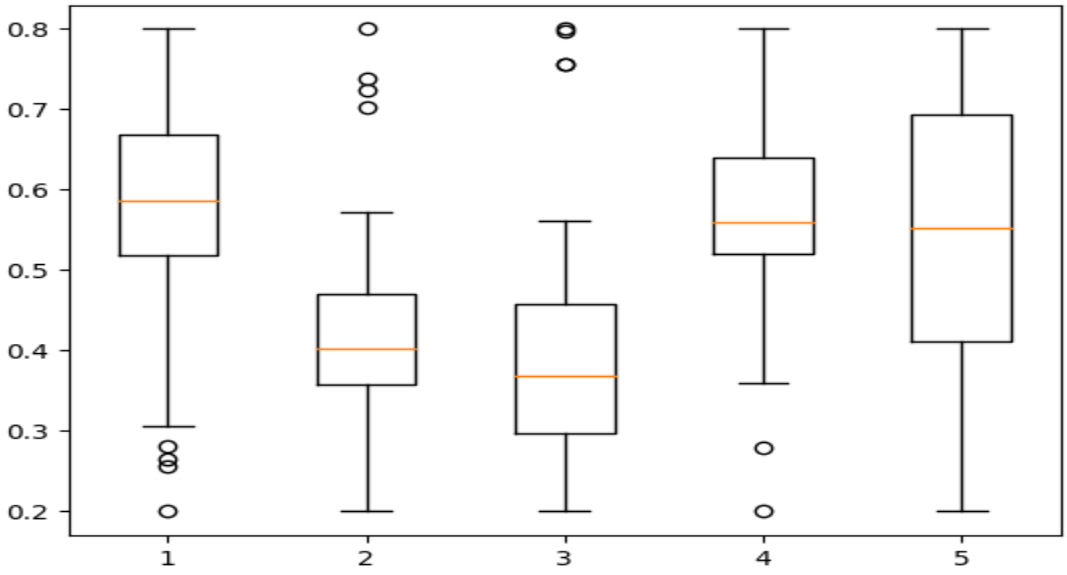
There are no null values in data.

```
Column                                  Non-Null Count
------                                  --------------
ID                                      63 non-null
 Large B/P                              63 non-null
 Large ROE                              63 non-null
 Large S/P                              63 non-null
 Large Return Rate in the last quarter  63 non-null
 Large Market Value                     63 non-null
 Small systematic Risk                  63 non-null
Annual Return                           63 non-null
Excess Return                           63 non-null
Systematic Risk                         63 non-null
Total Risk                              63 non-null
Abs. Win Rate                           63 non-null
Rel. Win Rate                           63 non-null
```

b. **Influential points:**

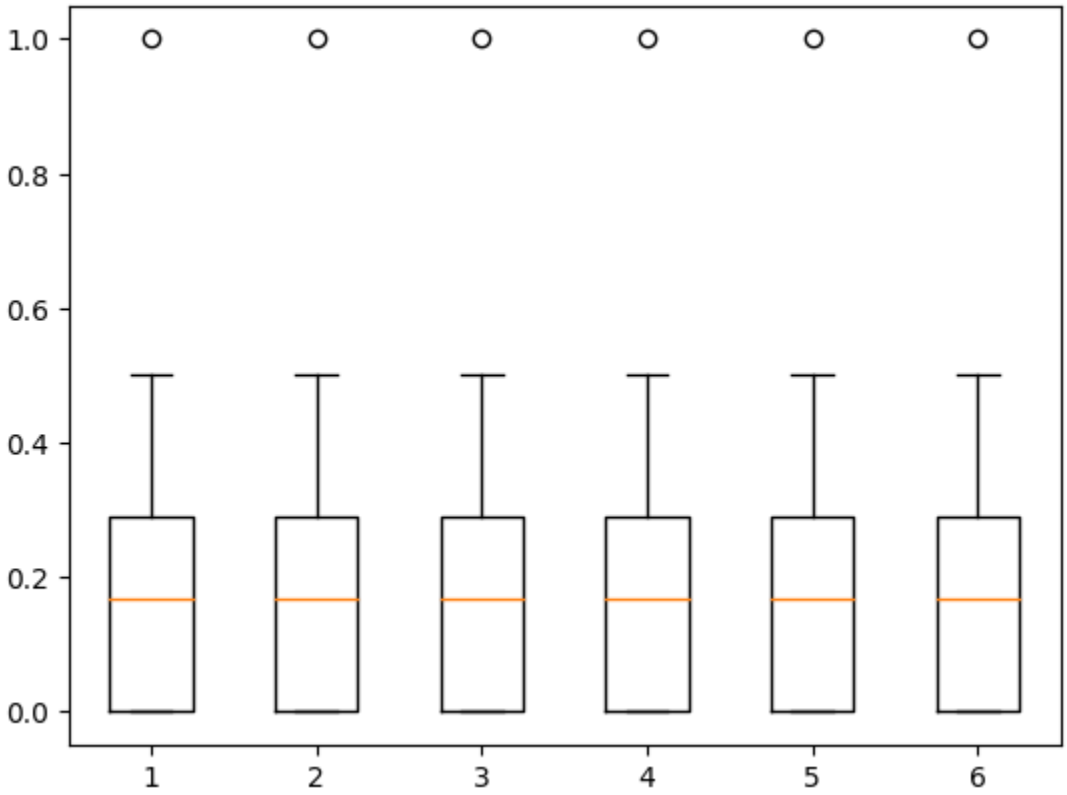
Influential points are those outliers which greatly affect the slope of the regression line. It may cause the Determination coefficient very big or sometimes too small.So it is better to remove them and then compute the regression line.The following plots show the box plots of features before and after outliers removal.

OUTLIERS IN WEIGHTS:




We haven't removed outliers because they are outliers but not.Every stock is important in terms of analysis.

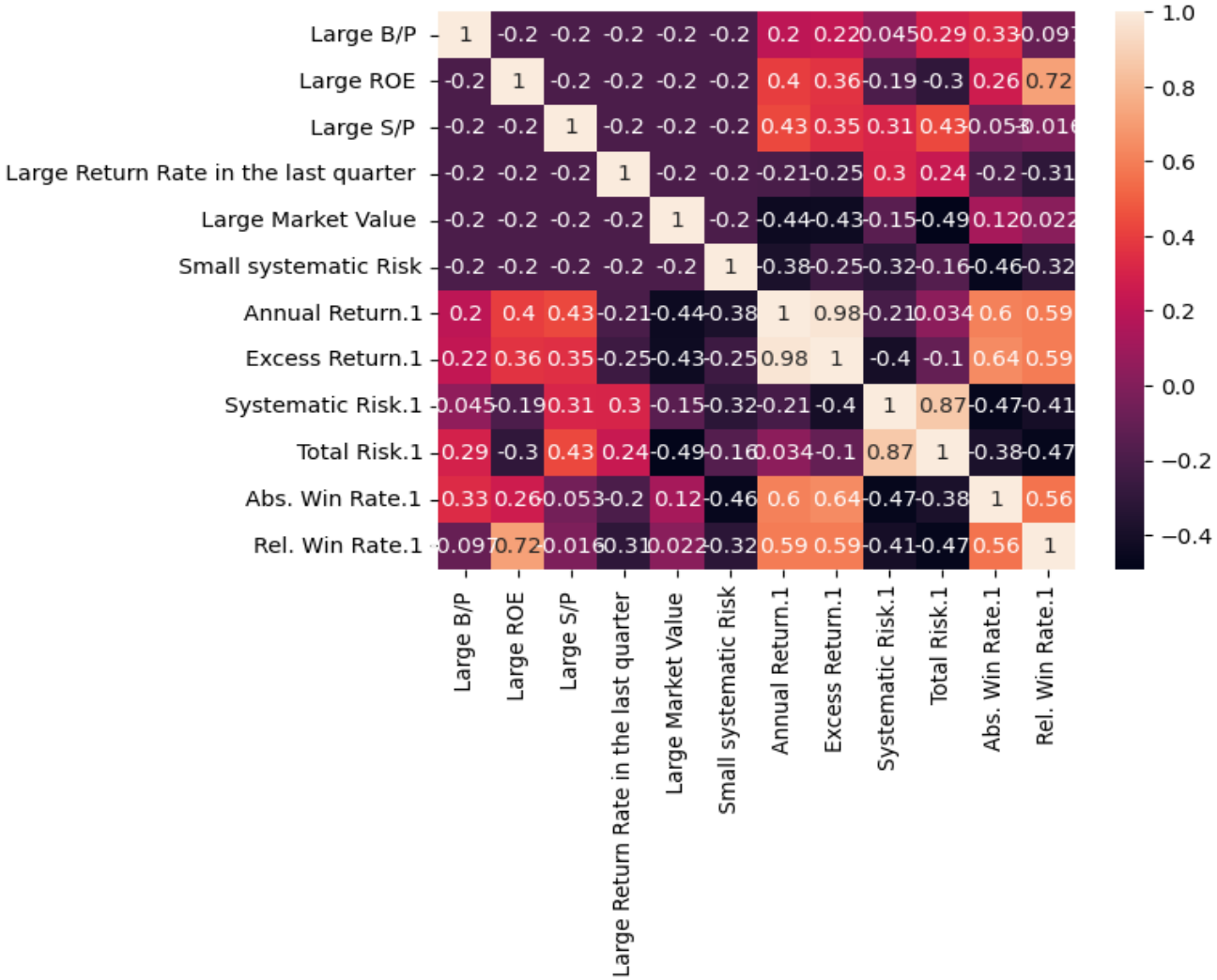
OUTLIERS IN OUTPUTS:



## 2. Exploratory Data Analysis(EDA)

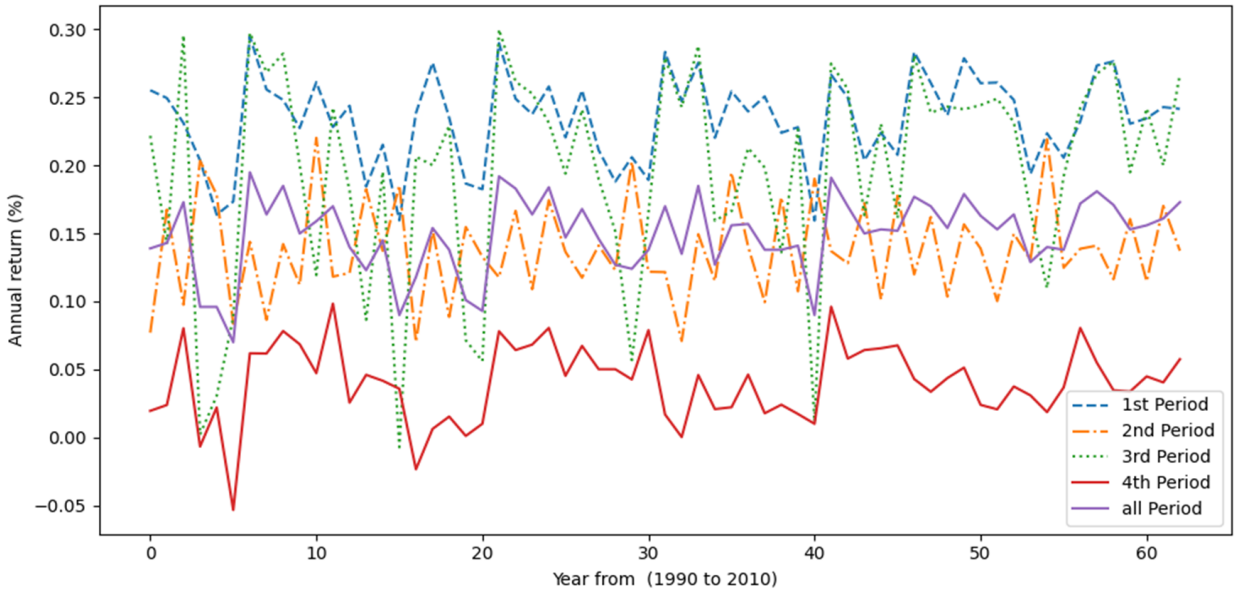
### a. Heat_Map

Here we see the correlation between all attributes in the dataset. And we can see how they are correlated. We see for weights the correlation between them is very low and equals to -0.2.So we can say that they are not correlated based on heatmap.

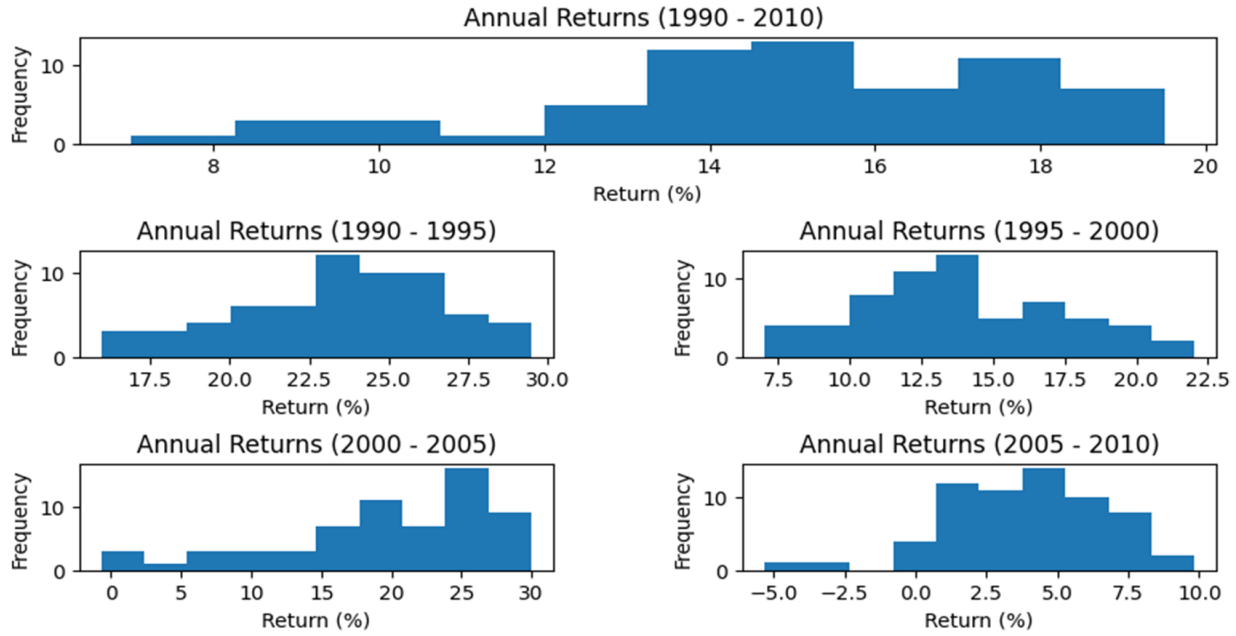


### b. Data Exploration

Analysis Annual returns of stocks of different time periods 20 years(80 quarters) , and it is compared to 4 time periods (20 quarters each).



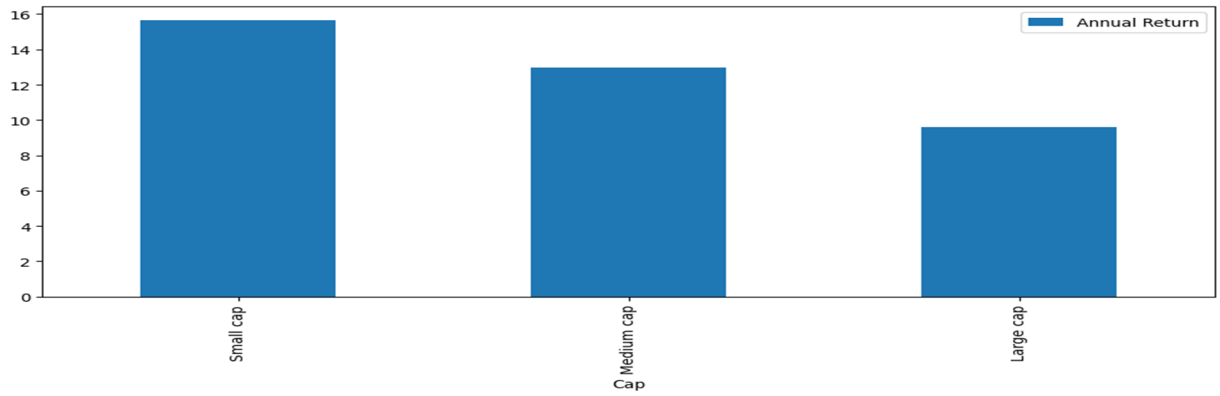

Comparing Annual return of the stocks on 20 year time period



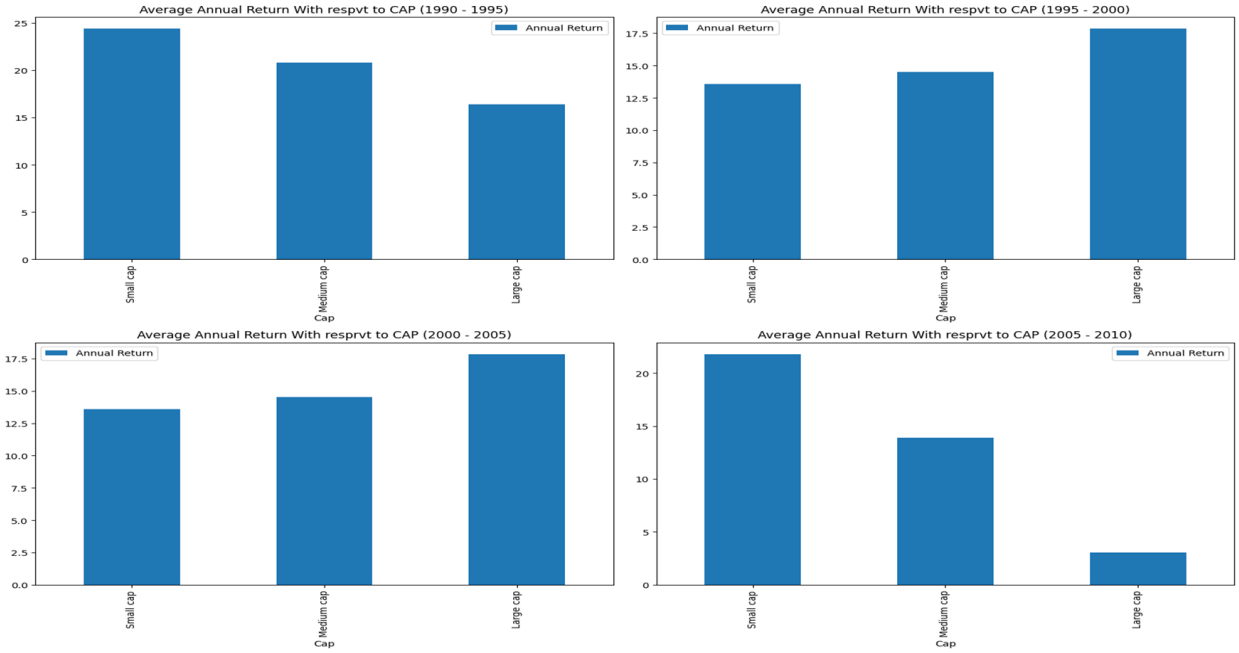
### c. Feature Engineering

We have defined three types of labels small , medium and large caps based on their market value by binning data as bins (0.0,.33,0.6,1) on overall time period.

| Large Market value | Cap |
|---|---|
| 0 - 0.33 | Small Cap |
| 0.33 - 0.6 | Medium cap |
| 0.6 - 1 | Large cap |



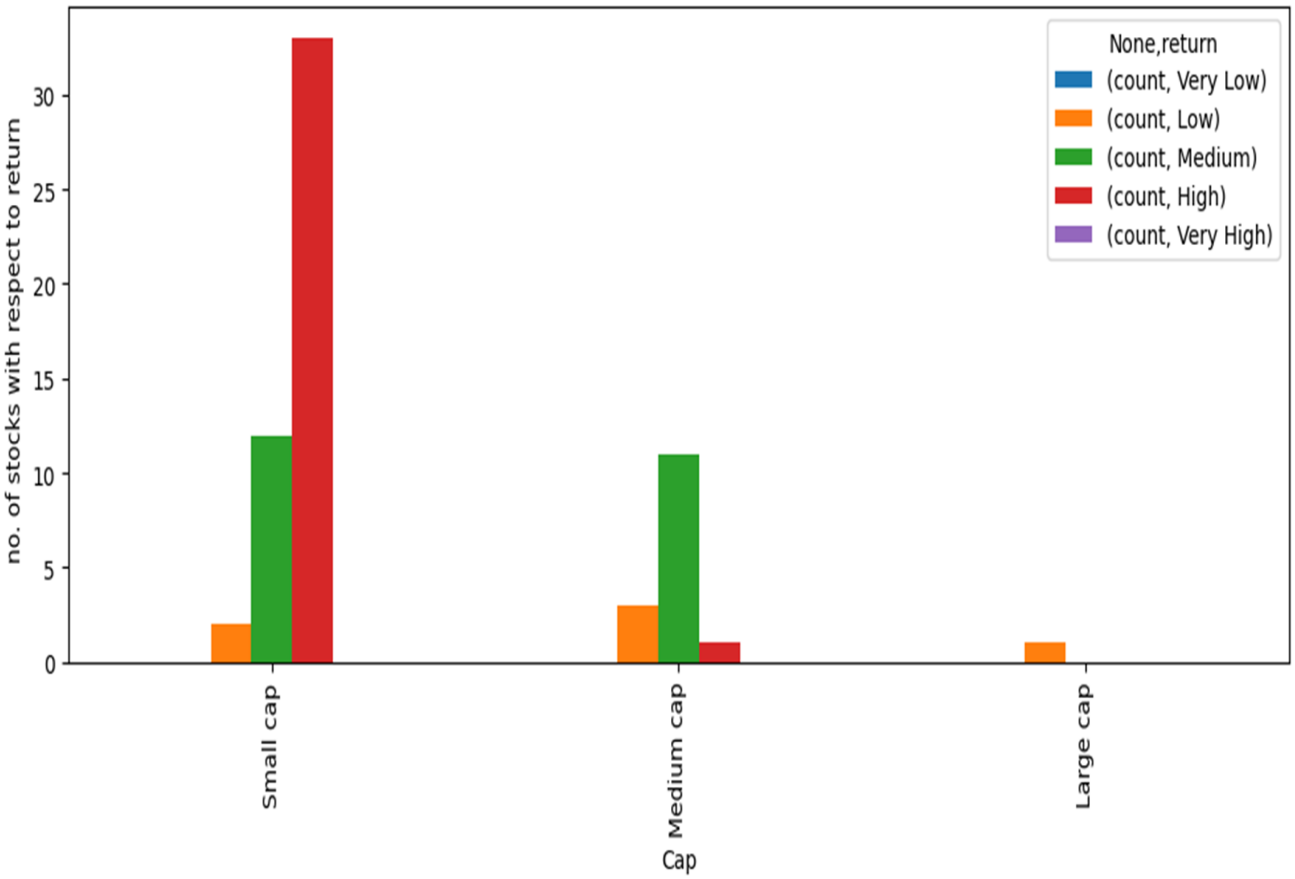

Similarly, we can see the trend of market values of small ,medium and large industries over these four periods like how it is varying.

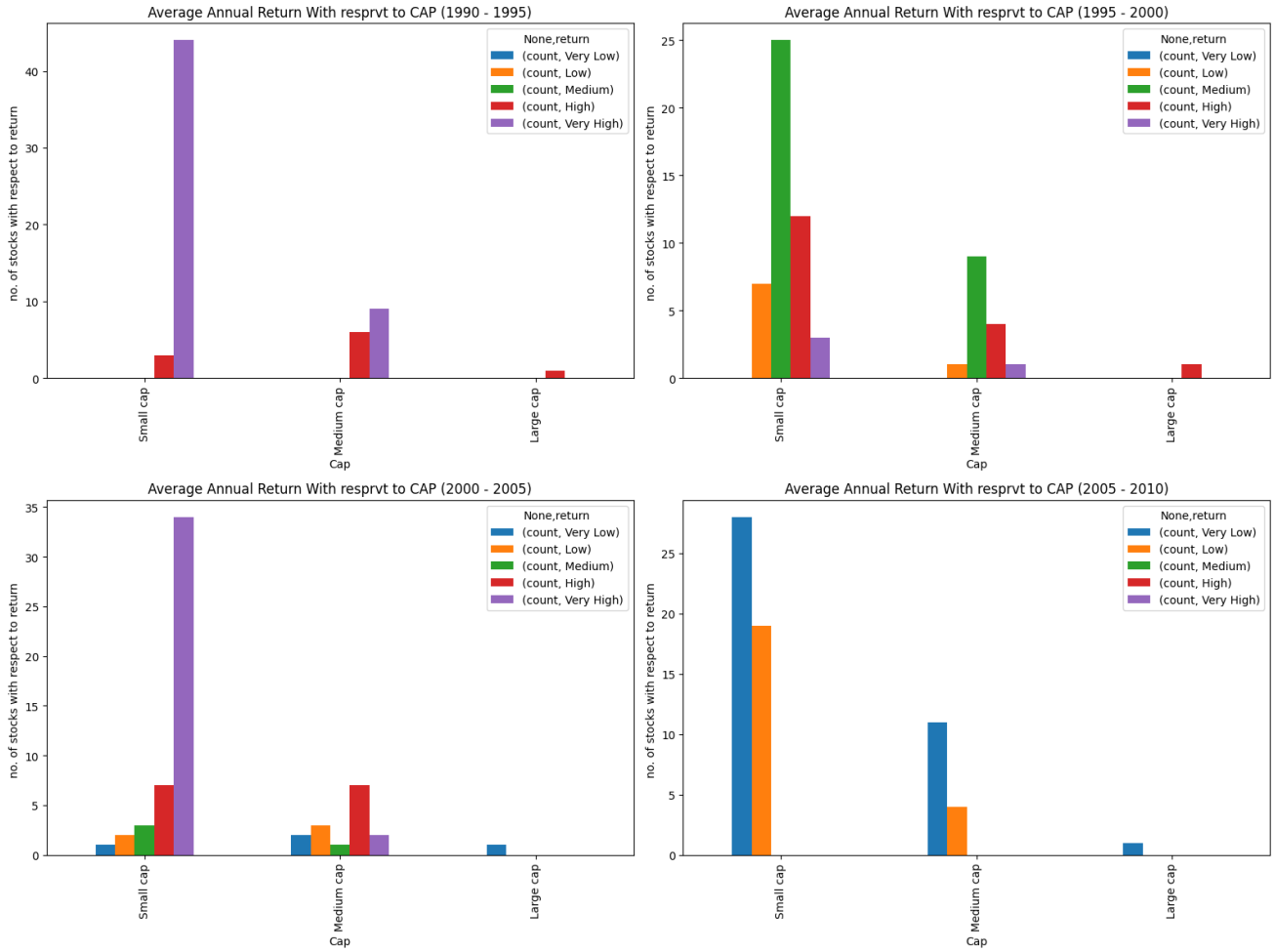


We binned annual returns of companies as 'Very Low', 'Low' , 'Medium' , 'High' , 'Very High' over all periods. We can see no. of stocks w.r.t to stocks.We binned data based on **bins = [-10,5,10,15,20,35]**

| **Annual Return(Range)** | **Return bin** |
|---|---|
| -10 - 5 | Very Low |
| 5 - 10 | Low |
| 10 - 15 | Medium |
| 15 - 20 | High |
| 20 - 35 | Very High |

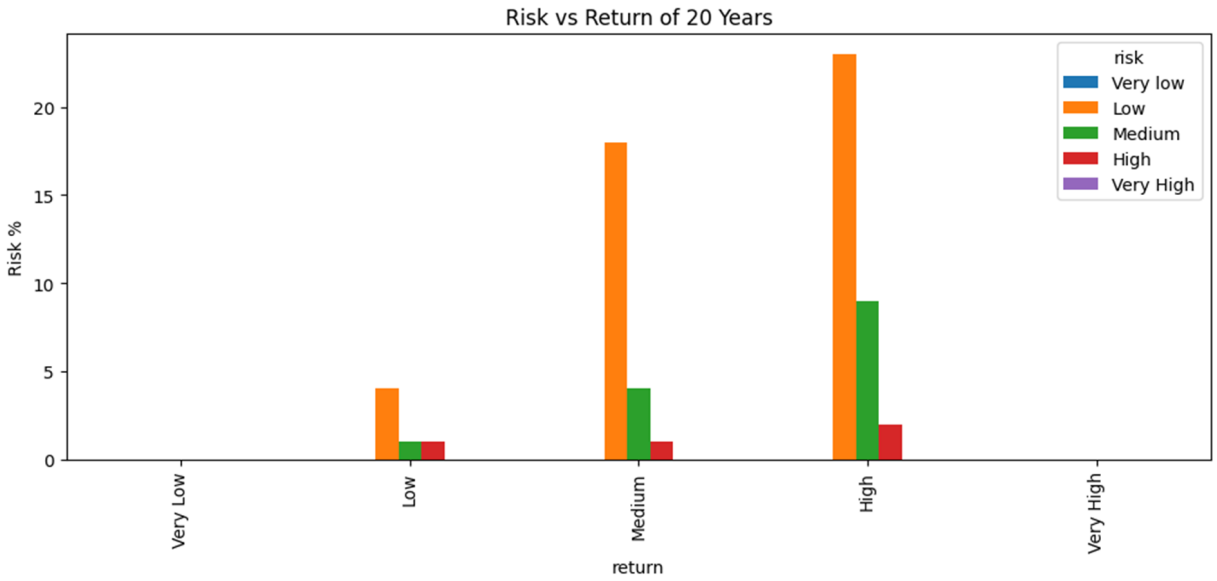
After binning the data according to annual returns , comparing them to the market cap. There are more small cap stocks giving higher returns.



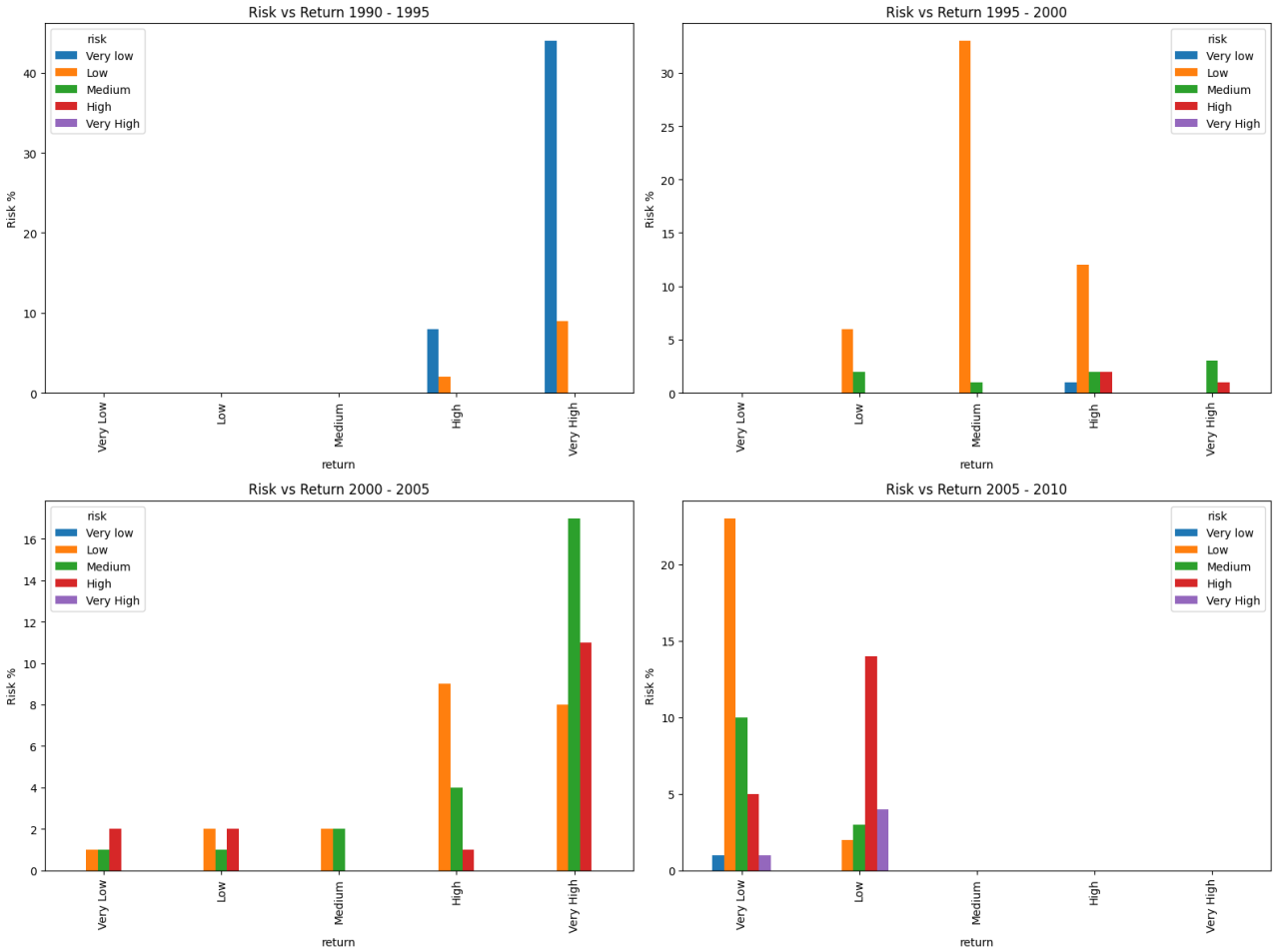



Marking the stocks according to their Risk as Very low, Low, Medium, High, Very high. The Risk of the stocks from the past 20 years are ranging up to 22 percent.

| **Total Risk %** | **Risk bin** |
|---|---|
| **0 - 8%** | **Very Low** |
| **8 - 11%** | **Low** |
| **11 - 13%** | **Medium** |
| **13 - 17%** | **High** |
| **17 - 22%** | **Very High** |

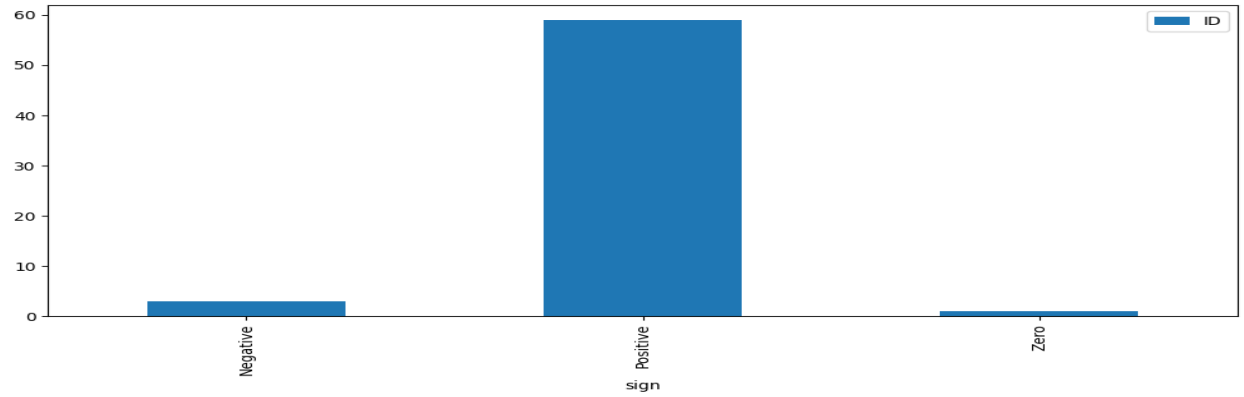


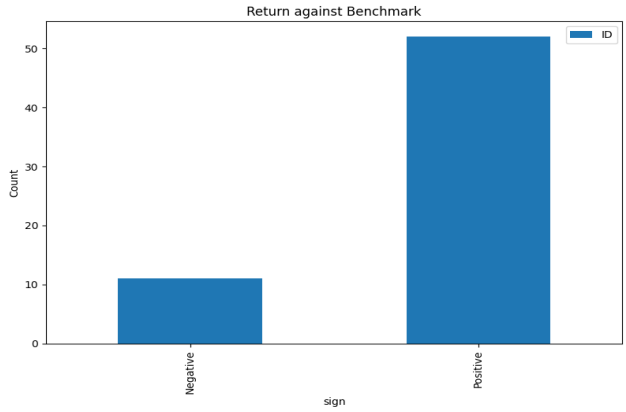

The 4 plots represent the Risk vs Return of the stocks over 20 years in 4 time frames.

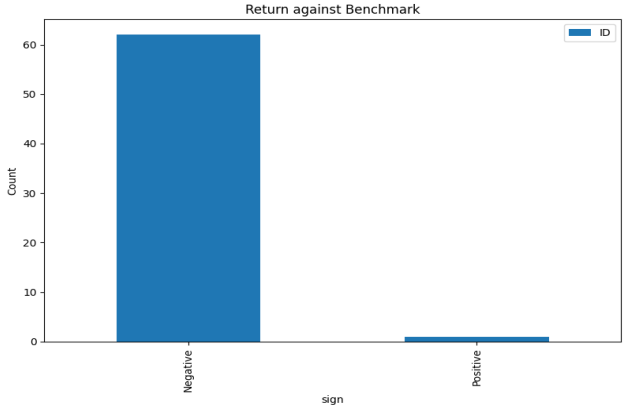


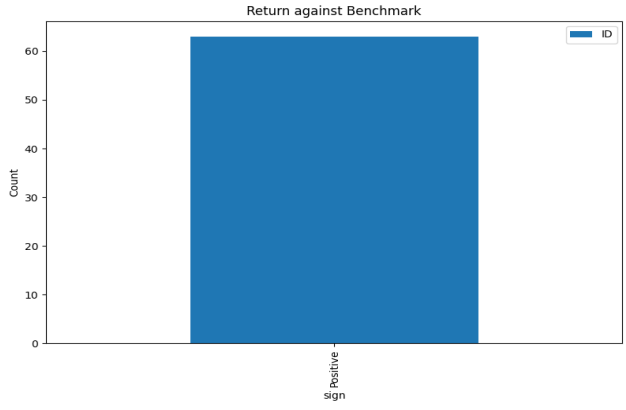
Binning the Stocks according to performance over Benchmark of that particular time period.they are named after positive, negative and zero. Positive if they gave more return than benchmark, negative if they gave less return than benchmark, zero if the return and benchmark value both are same.

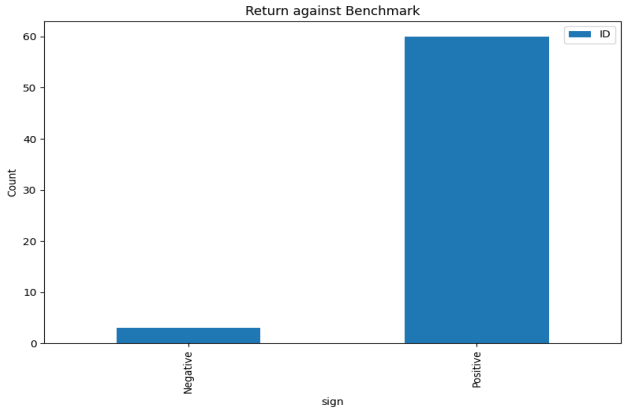
| **Excess Return** | **Sign** |
|---|---|
| >0% | Positive |
| = 0% | Zero |
| < 0% | Negative |



## 3. Principal component analysis(PCA)

PCA (Principal Component Analysis) is a statistical technique used for dimensionality reduction by transforming data into a lower-dimensional space while preserving the maximum amount of variance in the data.

### a. Bartlett's sphericity test

Bartlett's sphericity test is a statistical test used to determine if the correlation matrix underlying a PCA is significantly different from an identity matrix, indicating whether PCA is an appropriate technique for reducing the dimensionality of a dataset.

H0 : R = I
H1 : R != I
If p < 0.5 : reject H0 (PCA is required)
If p > 0.5: accept H0 (Pca is not required)

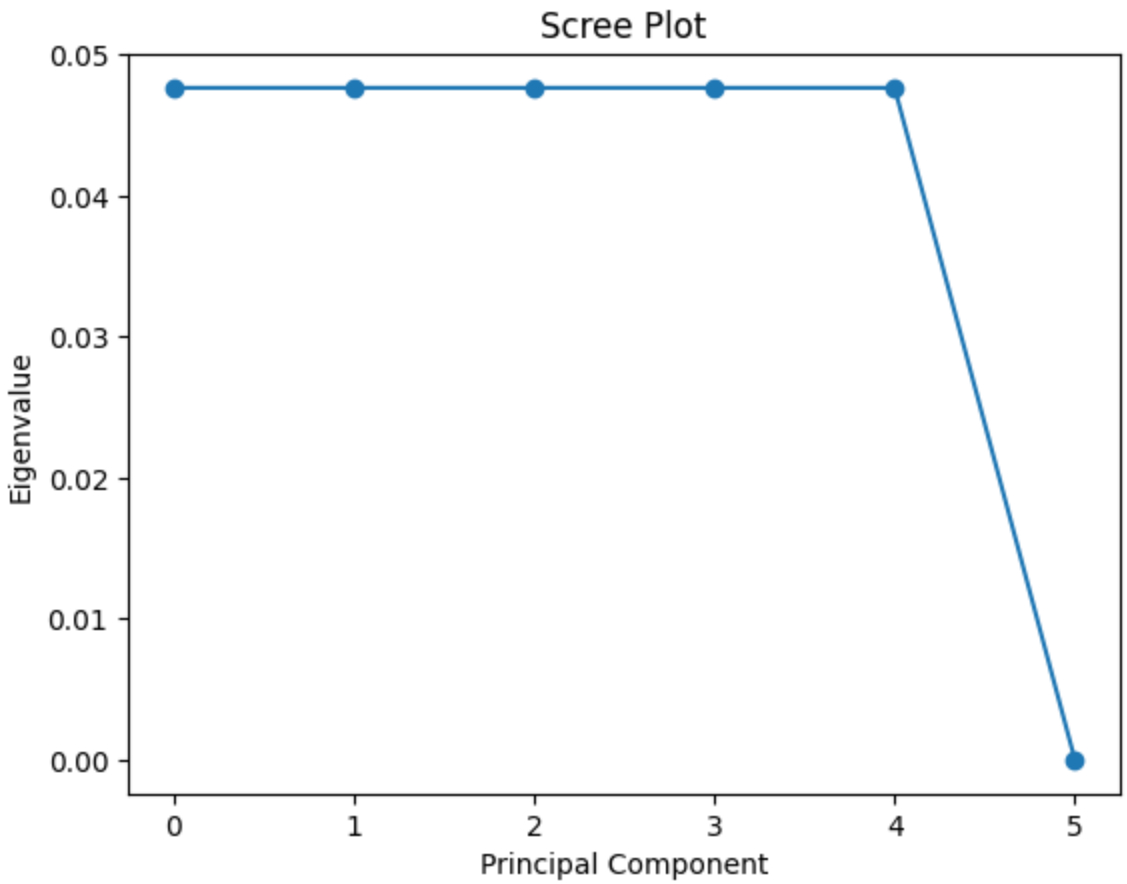
Testing weights with Bartlett's sphericity test the results are as follows:

Bartlett Sphericity Test:
Chi-square value: 749.0536602587168
P-value: 6.429146777304694e-150
PCA is required for the weights as P value is too small.

### b. Scree plot(Elbow method)

A scree plot is a simple graphical tool used to determine the optimal number of components or factors in a principal component analysis (PCA) or factor analysis. It plots the eigenvalues of each component or factor against its corresponding index or number, and the optimal number of components or factors is determined by identifying the "elbow" point on the plot where the eigenvalues level off.Finding eigenvectors and eigenvalues of the inputs(weights)

Eigenvalues = 4.76666774e-02
5.06912442e-08
4.76666774e-02
4.76666774e-02
4.76666774e-02
4.76666774e-02




### c. Feature reduction

From the scree plot we can understand that the first 5 components have 99% of variance Last component have negligible variance we can neglect that component. Therefore, principal components are reduced from 6 to 5.

## 4. Test of Assumption

### a. Multicollinearity

VIF (Variance Inflation Factor) is a measure of the degree of multicollinearity in a regression analysis. It quantifies how much the variance of the estimated regression coefficient is increased because of collinearity among the predictor variables.If VIF is greater than 10 , PCA should be done on the weights

VIF of PCA with 5 columns are as follows:

VIF of PCA1: 1.00
VIF of PCA2: 1.00
VIF of PCA3: 1.00
VIF of PCA4: 1.00
VIF of PCA5: 1.00

**b. Autocorrelation**

The Durbin-Watson test is a statistical test used to detect the presence of autocorrelation in the residuals of a regression analysis. It tests the null hypothesis that the residuals are not autocorrelated against the alternative that they follow a first-order autoregressive process.

If **DW** is **two**: there is no correlation

If **DW** > **two**: there is no correlation

If **DW** < **two**: there is no correlation

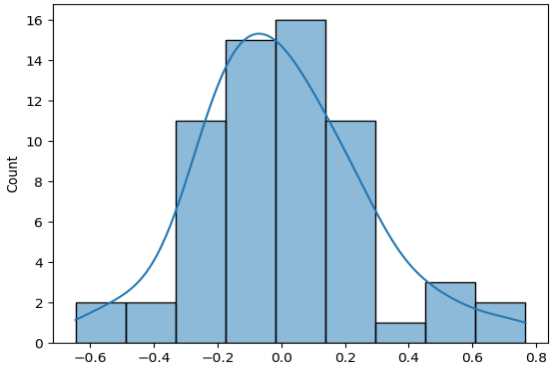
Accepted range of the DW test is approximately 2.

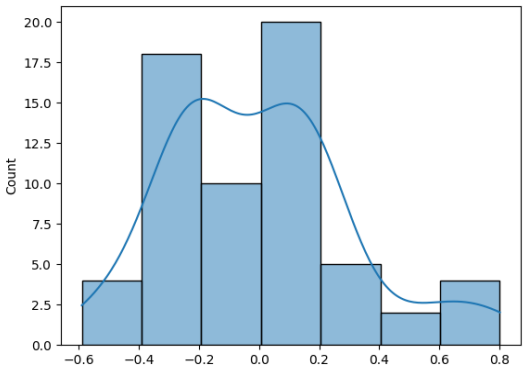
DW statistic for each column: [1.92880016 1.8952121 1.47743874 2.15861333 2.50857133]

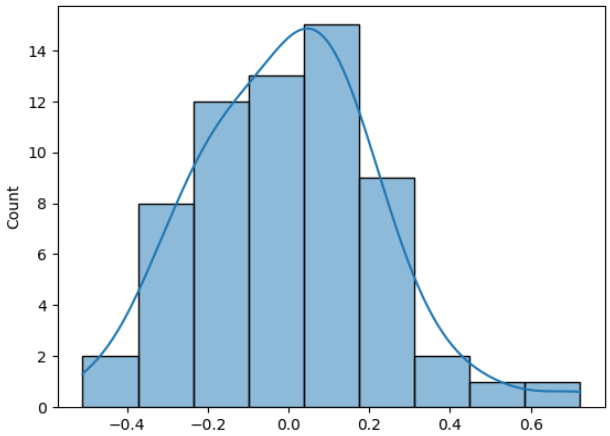
Columns having DW with 1.477 ,2.50 are dropped as they are highly correlated.

c. Normality

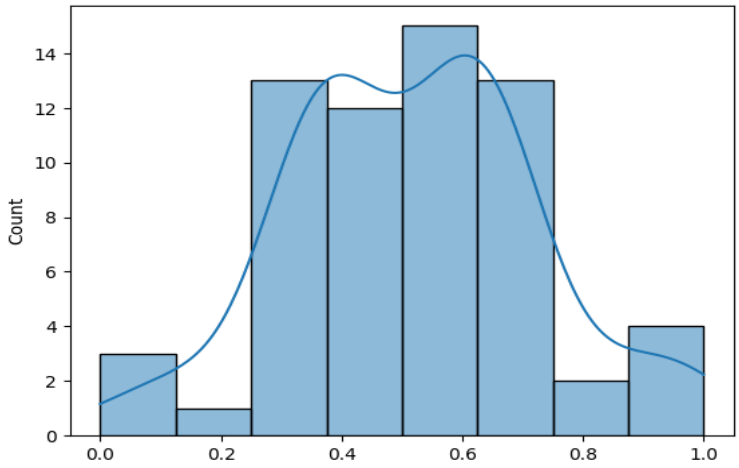
Normality can be observed using histogram of the data.and transforming the data which is not following normal distribution . Box cox is used for transforming.If the data is not normalised we cannot apply P test and F test on the data.



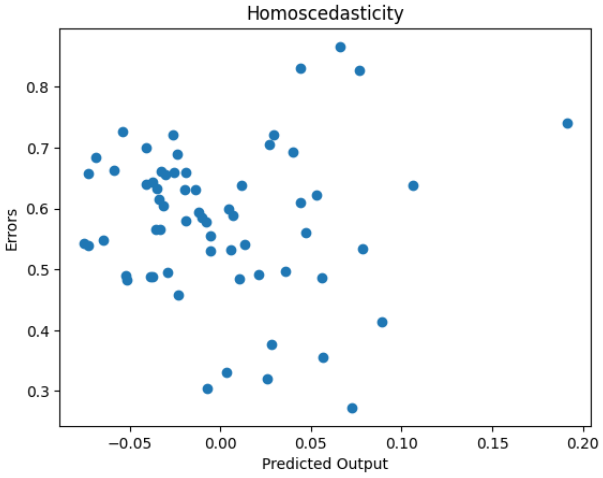



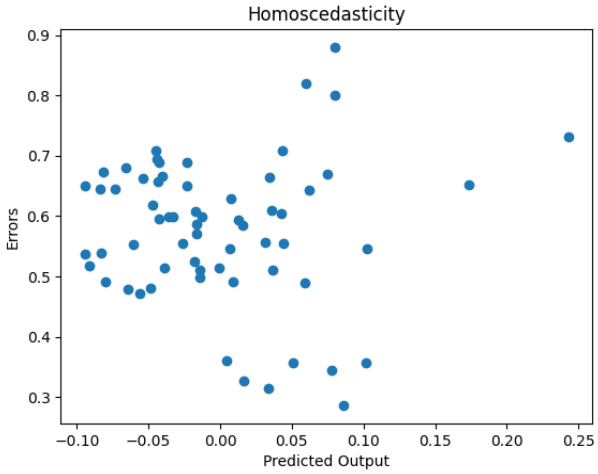
Two principal components follow normal distribution and the first component one doest follow principal components.Box cox is applied to normalise the data .Finally, weights for predicting the output , 3 principal components are needed.

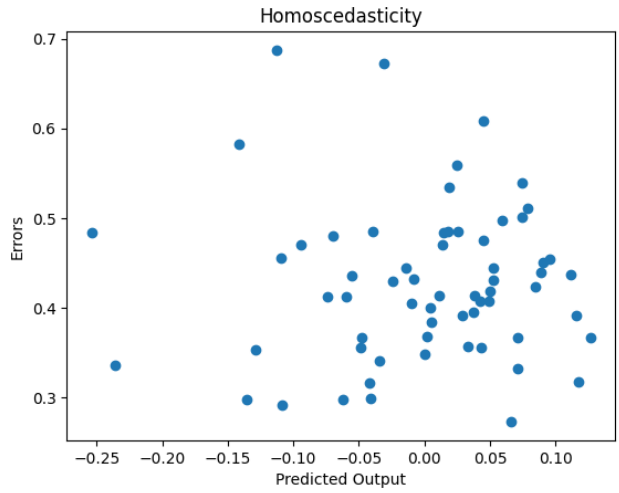


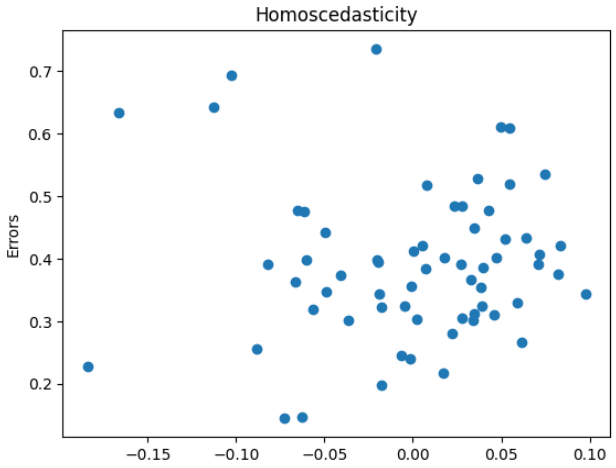

### d. Homoscedasticity

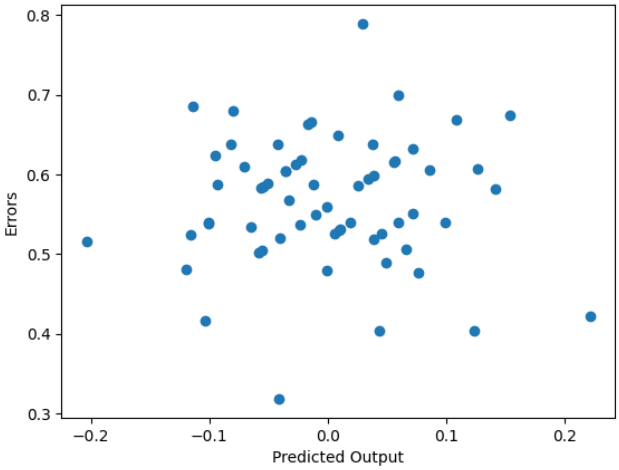
Homoscedasticity refers to the property of having equal variance in the errors or residuals of a statistical model across all levels of the predictor variables. When a model exhibits homoscedasticity, it indicates that the variability of the errors is constant and not related to the values of the independent variables.

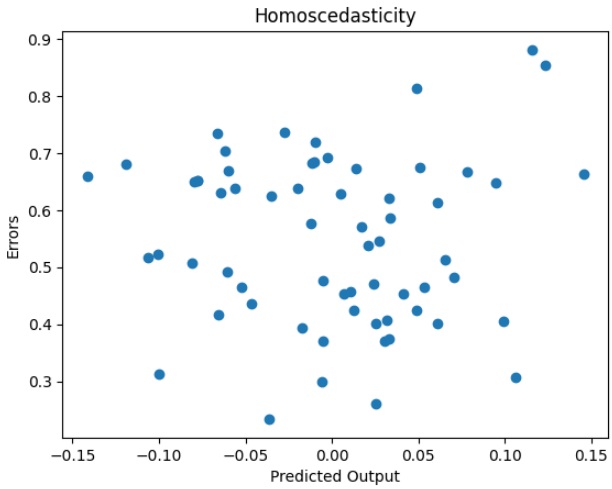





## 5. Model Ordinary Least square method (OLS)

**OLS (Ordinary Least Squares)** is a method used in linear regression to estimate the coefficients of the regression equation that minimises the sum of the squared residuals.The method assumes that the errors are normally distributed and homoscedastic, and that there is a linear relationship between the dependent variable and the independent variables.The resulting coefficients can be used to make predictions about the dependent variable.

**R square :** R-squared, also known as the coefficient of determination, is a statistical measure that represents the proportion of variance in the dependent variable that is explained by the independent variable(s) in a regression model. It is a value between 0 and 1, with higher values indicating a better fit of the model to the data.

**Equation of the regression model :**

Output = W1 * PCA1 + W2* PCA2 + W3 * PCA3 +Constant

**Annual Return**

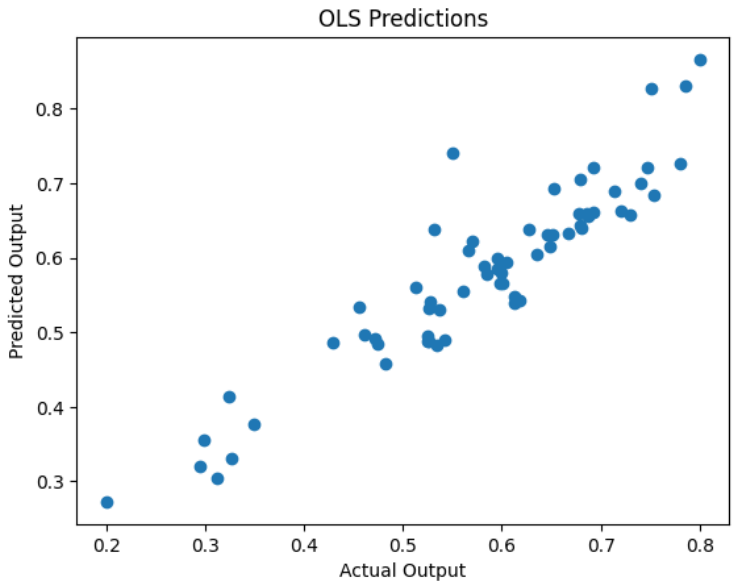
OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Annual Return.1 | R-squared: | 0.858 |
| Model: | OLS | Adj. R-squared: | 0.851 |
| Method: | Least Squares | F-statistic: | 118.6 |
| Date: | Mon, 03 Apr 2023 | Prob (F-statistic): | 5.94e-25 |
| Time: | 11:22:27 | Log-Likelihood: | 99.473 |
| No. Observations: | 63 | AIC: | -190.9 |
| Df Residuals: | 59 | BIC: | -182.4 |
| Df Model: | 3 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 0.8532 | 0.018 | 48.576 | 0.000 | 0.818 | 0.888 |
| x1 | -0.5253 | 0.031 | -16.732 | 0.000 | -0.588 | -0.462 |
| x2 | 0.2166 | 0.024 | 8.846 | 0.000 | 0.168 | 0.266 |
| x3 | 0.0097 | 0.029 | 0.329 | 0.743 | -0.049 | 0.068 |

| | | | |
|---|---|---|---|
| Omnibus: | 16.448 | Durbin-Watson: | 1.232 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 20.666 |
| Skew: | -1.072 | Prob(JB): | 3.25e-05 |
| Kurtosis: | 4.810 | Cond. No. | 6.19 |

R square = 0.858
Weights for the input variables = [-0.5253 , 0.2166 , 0.0097]
Constant = [0.8532]



**Excess Return**

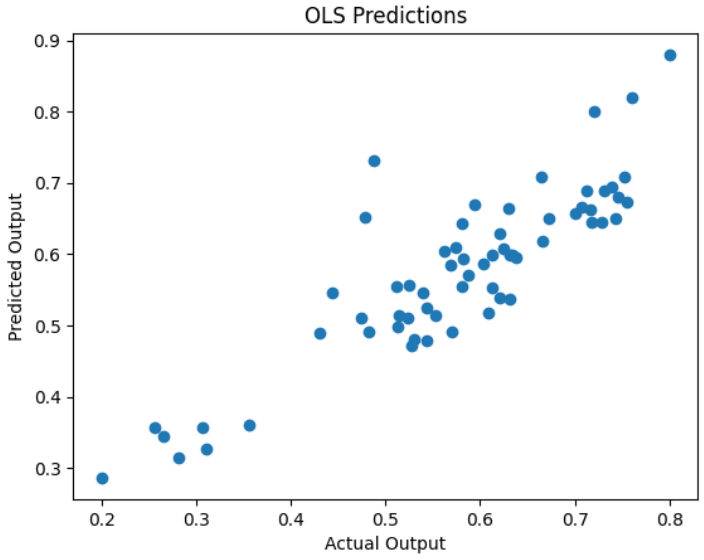
OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Excess Return.1 | R-squared: | 0.770 |
| Model: | OLS | Adj. R-squared: | 0.759 |
| Method: | Least Squares | F-statistic: | 65.98 |
| Date: | Mon, 03 Apr 2023 | Prob (F-statistic): | 7.72e-19 |
| Time: | 11:22:27 | Log-Likelihood: | 82.666 |
| No. Observations: | 63 | AIC: | -157.3 |
| Df Residuals: | 59 | BIC: | -148.8 |
| Df Model: | 3 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 0.8529 | 0.023 | 37.188 | 0.000 | 0.807 | 0.899 |
| x1 | -0.5324 | 0.041 | -12.986 | 0.000 | -0.614 | -0.450 |
| x2 | 0.1686 | 0.032 | 5.272 | 0.000 | 0.105 | 0.233 |
| x3 | 0.0693 | 0.038 | 1.806 | 0.076 | -0.007 | 0.146 |

| | | | |
|---|---|---|---|
| Omnibus: | 16.095 | Durbin-Watson: | 0.889 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 19.981 |
| Skew: | -1.058 | Prob(JB): | 4.58e-05 |
| Kurtosis: | 4.771 | Cond. No. | 6.19 |

R square = 0.770
Weights for the input variables = [-0.5324,0.1686,0.0693]
Constant = [0.8529]



**Systematic Risk**

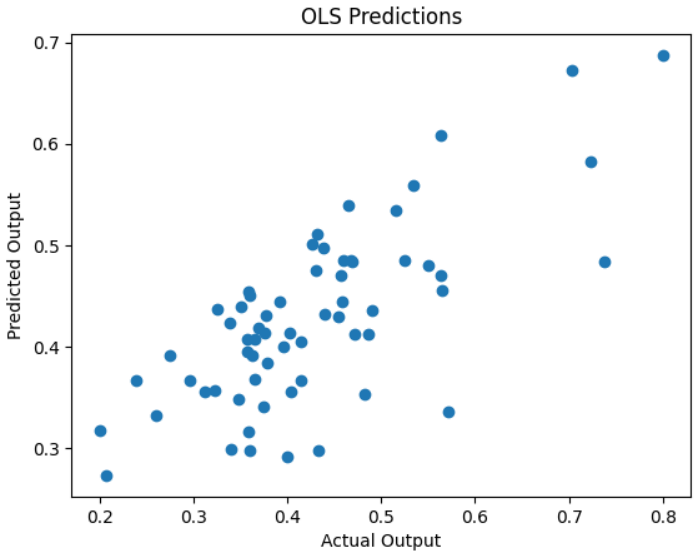
OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Systematic Risk.1 | R-squared: | 0.535 |
| Model: | OLS | Adj. R-squared: | 0.512 |
| Method: | Least Squares | F-statistic: | 22.67 |
| Date: | Mon, 03 Apr 2023 | Prob (F-statistic): | 6.92e-10 |
| Time: | 11:22:27 | Log-Likelihood: | 69.804 |
| No. Observations: | 63 | AIC: | -131.6 |
| Df Residuals: | 59 | BIC: | -123.0 |
| Df Model: | 3 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 0.3227 | 0.028 | 11.474 | 0.000 | 0.266 | 0.379 |
| x1 | 0.1996 | 0.050 | 3.970 | 0.000 | 0.099 | 0.300 |
| x2 | 0.2590 | 0.039 | 6.605 | 0.000 | 0.181 | 0.337 |
| x3 | -0.1374 | 0.047 | -2.920 | 0.005 | -0.232 | -0.043 |

| | | | |
|---|---|---|---|
| Omnibus: | 11.661 | Durbin-Watson: | 1.007 |
| Prob(Omnibus): | 0.003 | Jarque-Bera (JB): | 11.874 |
| Skew: | 0.950 | Prob(JB): | 0.00264 |
| Kurtosis: | 3.955 | Cond. No. | 6.19 |

R square = 0.535
Weights for the input variables = [0.1996,0.2590,-0.1374]
Constant = [0.3227]



**Total Risk:**

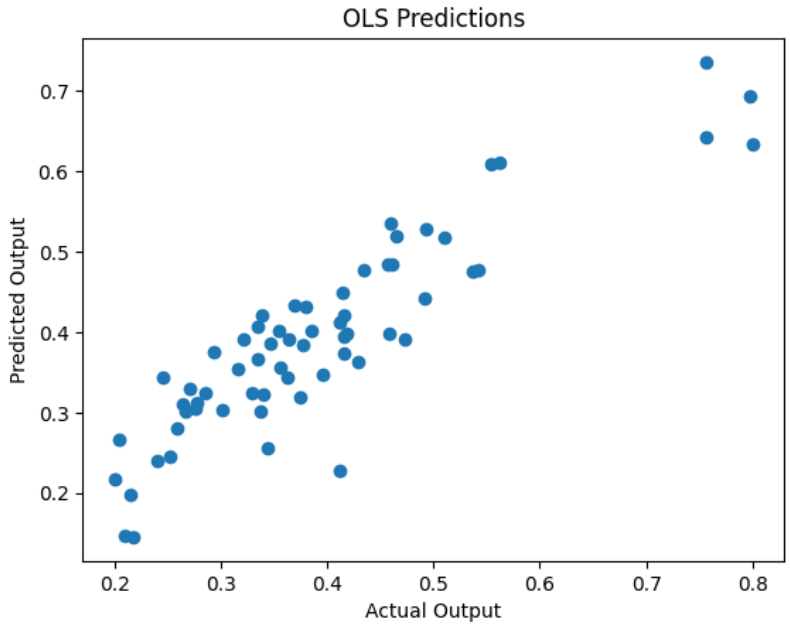
OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Total Risk.1 | R-squared: | 0.810 |
| Model: | OLS | Adj. R-squared: | 0.800 |
| Method: | Least Squares | F-statistic: | 83.61 |
| Date: | Mon, 03 Apr 2023 | Prob (F-statistic): | 3.16e-21 |
| Time: | 11:22:27 | Log-Likelihood: | 88.744 |
| No. Observations: | 63 | AIC: | -169.5 |
| Df Residuals: | 59 | BIC: | -160.9 |
| Df Model: | 3 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 0.2992 | 0.021 | 14.365 | 0.000 | 0.257 | 0.341 |
| x1 | 0.1781 | 0.037 | 4.785 | 0.000 | 0.104 | 0.253 |
| x2 | 0.4372 | 0.029 | 15.057 | 0.000 | 0.379 | 0.495 |
| x3 | -0.0064 | 0.035 | -0.183 | 0.856 | -0.076 | 0.063 |

| | | | |
|---|---|---|---|
| Omnibus: | 9.651 | Durbin-Watson: | 1.368 |
| Prob(Omnibus): | 0.008 | Jarque-Bera (JB): | 9.301 |
| Skew: | 0.887 | Prob(JB): | 0.00955 |
| Kurtosis: | 3.631 | Cond. No. | 6.19 |

R square = 0.810
Weights for the input variables = [0.1781,0.0.4372,-0.0064]
Constant = [0.2992]



**Abs. Win Rate**

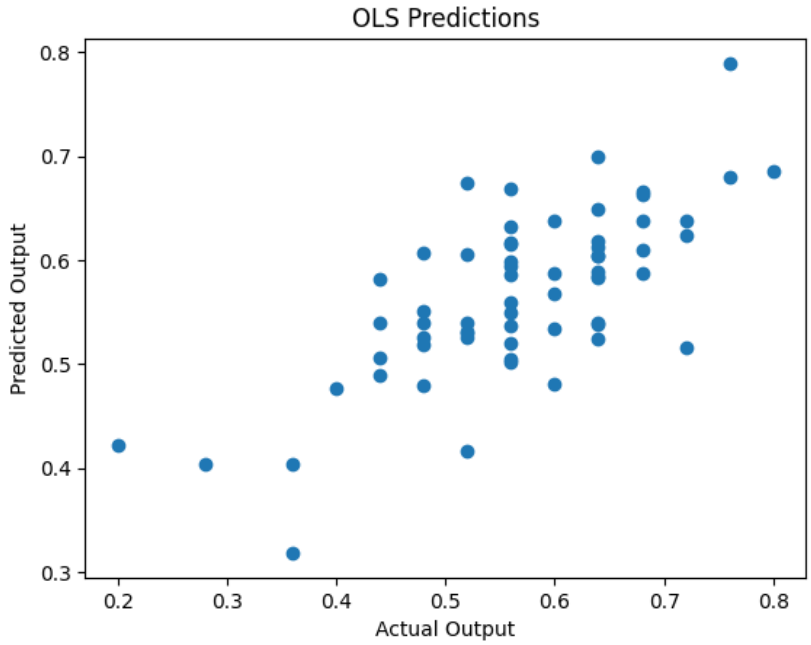
OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Abs. Win Rate.1 | R-squared: | 0.516 |
| Model: | OLS | Adj. R-squared: | 0.492 |
| Method: | Least Squares | F-statistic: | 21.01 |
| Date: | Mon, 03 Apr 2023 | Prob (F-statistic): | 2.22e-09 |
| Time: | 11:22:27 | Log-Likelihood: | 71.472 |
| No. Observations: | 63 | AIC: | -134.9 |
| Df Residuals: | 59 | BIC: | -126.4 |
| Df Model: | 3 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 0.7542 | 0.027 | 27.531 | 0.000 | 0.699 | 0.809 |
| x1 | -0.3601 | 0.049 | -7.355 | 0.000 | -0.458 | -0.262 |
| x2 | -0.0326 | 0.038 | -0.852 | 0.397 | -0.109 | 0.044 |
| x3 | 0.1352 | 0.046 | 2.950 | 0.005 | 0.044 | 0.227 |

| | | | |
|---|---|---|---|
| Omnibus: | 0.932 | Durbin-Watson: | 1.830 |
| Prob(Omnibus): | 0.627 | Jarque-Bera (JB): | 0.498 |
| Skew: | -0.204 | Prob(JB): | 0.780 |
| Kurtosis: | 3.150 | Cond. No. | 6.19 |

R square = 0.516
Weights for the input variables = [-0.3601,-0.0326,0.0064]
Constant = [0.7542]



**Rel. Win Rate**

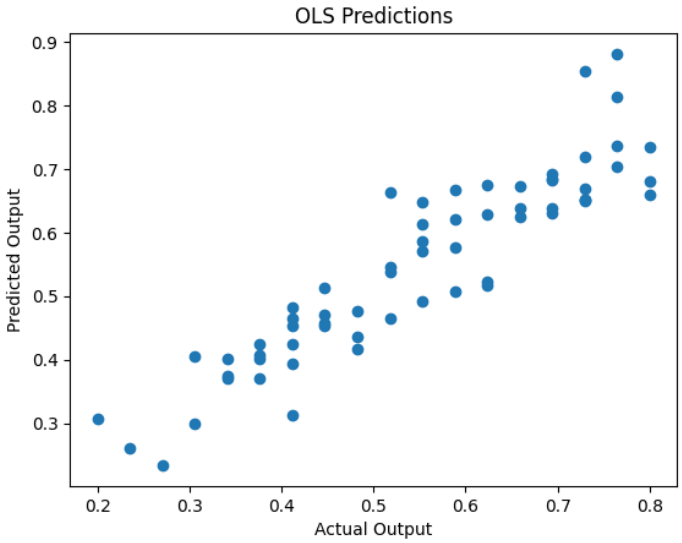
OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Rel. Win Rate.1 | R-squared: | 0.841 |
| Model: | OLS | Adj. R-squared: | 0.833 |
| Method: | Least Squares | F-statistic: | 104.1 |
| Date: | Mon, 03 Apr 2023 | Prob (F-statistic): | 1.55e-23 |
| Time: | 11:22:27 | Log-Likelihood: | 84.717 |
| No. Observations: | 63 | AIC: | -161.4 |
| Df Residuals: | 59 | BIC: | -152.9 |
| Df Model: | 3 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 0.8892 | 0.022 | 40.052 | 0.000 | 0.845 | 0.934 |
| x1 | -0.6565 | 0.040 | -16.545 | 0.000 | -0.736 | -0.577 |
| x2 | -0.1632 | 0.031 | -5.274 | 0.000 | -0.225 | -0.101 |
| x3 | -0.1045 | 0.037 | -2.813 | 0.007 | -0.179 | -0.030 |

| | | | |
|---|---|---|---|
| Omnibus: | 0.466 | Durbin-Watson: | 1.946 |
| Prob(Omnibus): | 0.792 | Jarque-Bera (JB): | 0.601 |
| Skew: | -0.012 | Prob(JB): | 0.741 |
| Kurtosis: | 2.522 | Cond. No. | 6.19 |

R square = 0.810
Weights for the input variables = [-0.6565,-0.1632,-0.1045]
Constant = [0.8892]



## Results:

| Output variable | R Square |
|---|---|
| Annual Return | 0.858 |
| Total Risk | 0.810 |
| Rel. Win Rate | 0.810 |
| Excess Return | 0.770 |
| Systematic Risk | 0.535 |
| Abs. Win Rate | 0.516 |

Annual return , Total Risk , Relative Win rate , Excess returns have given a good fit of regression . Systematic Risk and Absolute win rate have an average fit of the regression.

## Conclusion:

PCA (Principal Component Analysis) is a technique used to reduce the dimensionality of a dataset by identifying the most important features or components. In regression modelling, PCA can be used to reduce the number of independent variables and simplify the model.ln your scenario, the PCA analysis reduced the number of independent variables from 6 to 5. However, during the test of assumptions, it was found that there was correlation amongst the principal components.

Therefore, the PCA was further reduced to 3 components.Based on this, you have built 6 models for 6 output variables using the 3 principal components. This means that each model includes the 3 most important components that were identified through the PCA analysis.lt's important to note that while PCA can be a useful technique for reducing the number of variables in a regression model, it's important to carefully consider the number of components to include and to validate the results of the analysis. Additionally, it's important to ensure that the final model meets the assumptions of linear regression, such as normality , Autocorrelation,Multicollinearity and homoscedasticity.

Ordinary Least Squares (OLS) modelling provided a good fit for predicting the return, risk, and the win rate of stocks. OLS Modelling is a powerful statistical technique that can help investors make informed decisions by providing insights into the relationship between the independent variables (such as stock prices and market trends) and the dependent variables (such as return, risk, and win rate).This can provide valuable information for investors looking to make informed decisions about which stocks to invest in and when.

By using OLS modelling to predict the outputs of stocks, investors can choose the best stocks to gain maximum return at a safer to reward ratio. This means that they can identify stocks with a lower risk profile while still achieving a satisfactory level of return.