# SDA Project

## Exploratory data analysis and Modelling

# Members

Cherukuri Nikhilesh - S20180010040

Kore Nithish Kumar - S20180010086

Pulla Nagendra Babu - S20180010138

Rishabh Tripathi - S20180010147

# Methodology

1. Loading the Dataset
2. Univariate analysis
3. Checking for Null Values
4. Checking the Normality of Data
5. Exploring data trends
6. Detection and removal of influential points
7. Checking for correlation
8. Principal component analysis for feature selection
9. Factor Analysis
10. Splitting data into train and test data and Applying Multi linear regression
11. Test of Hypothesis
12. Test of assumptions(Linearity, Homoscedasticity, Normality of errors, etc.)

# About Dataset

**Independent Variables**

- Large B/P
- Large ROE
- Large S/P
- Large Return rate in last quarter
- Large Market Value
- Small Systematic risk

**Dependent Variables**

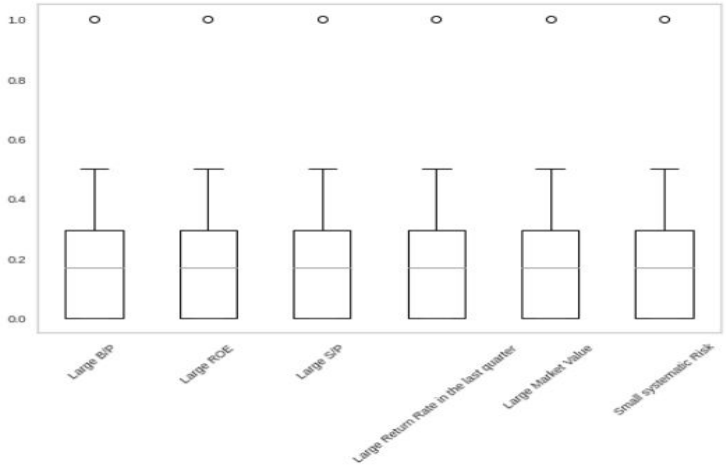
- Annual Return
- Excess Return
- Systematic Risk
- Total Risk
- Absolute Win rate
- Relative Win rate

**Given data of 4 different quarters from 1990-2010 and one with combined data of all quarters.**

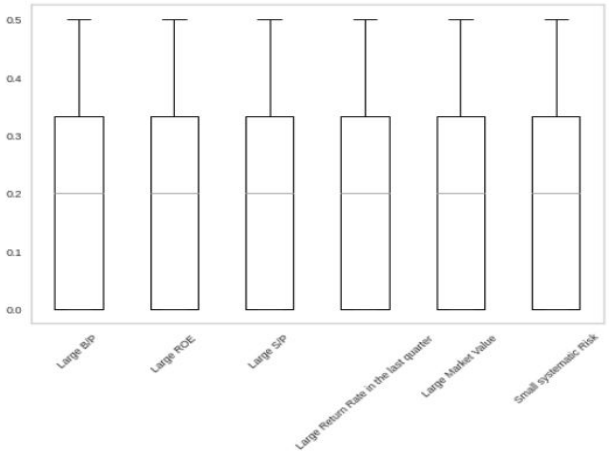
# Dataset Summary

| | Large B/P | Large ROE | Large S/P | Large Return Rate in the last quarter | Large Market Value | Small systematic Risk | Annual Return.1 | Excess Return.1 | Systematic Risk.1 | Total Risk.1 | Abs. Win Rate.1 | Rel. Win Rate.1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 63.000000 | 63.000000 | 63.000000 | 63.000000 | 63.000000 | 63.000000 | 63.000000 | 63.000000 | 63.000000 | 63.000000 | 63.000000 | 63.000000 |
| mean | 0.166619 | 0.166619 | 0.166619 | 0.166619 | 0.166619 | 0.166619 | 0.580151 | 0.576170 | 0.426494 | 0.391749 | 0.566984 | 0.547899 |
| std | 0.199304 | 0.199304 | 0.199304 | 0.199304 | 0.199304 | 0.199304 | 0.133358 | 0.137047 | 0.118178 | 0.136653 | 0.112803 | 0.159468 |
| min | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.200000 | 0.200000 | 0.200000 | 0.200000 | 0.200000 | 0.200000 |
| 25% | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.525811 | 0.519093 | 0.358600 | 0.297324 | 0.520000 | 0.411765 |
| 50% | 0.167000 | 0.167000 | 0.167000 | 0.167000 | 0.167000 | 0.167000 | 0.598516 | 0.587148 | 0.403418 | 0.368958 | 0.560000 | 0.552941 |
| 75% | 0.291500 | 0.291500 | 0.291500 | 0.291500 | 0.291500 | 0.291500 | 0.679636 | 0.669294 | 0.470571 | 0.457749 | 0.640000 | 0.694118 |
| max | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 0.800000 | 0.800000 | 0.800000 | 0.800000 | 0.800000 | 0.800000 |

# Detection of Outliers



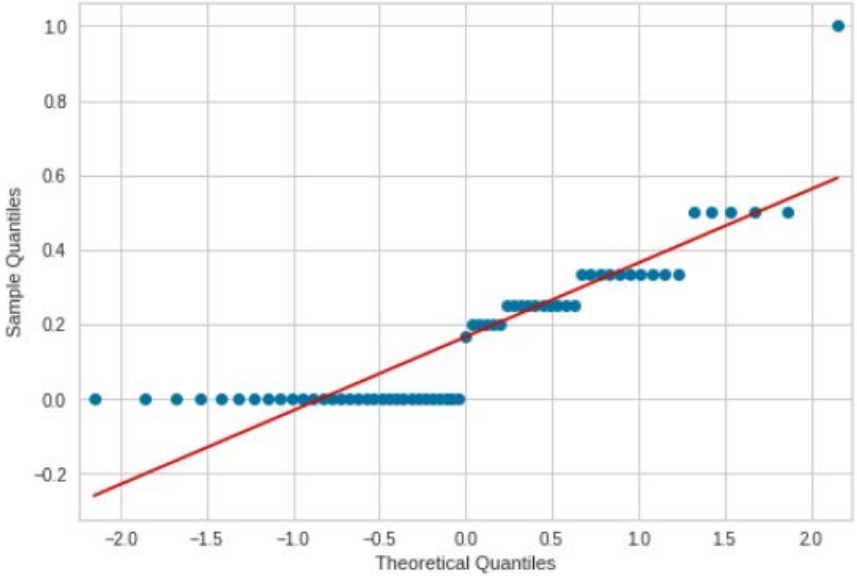

Before removing Outliers

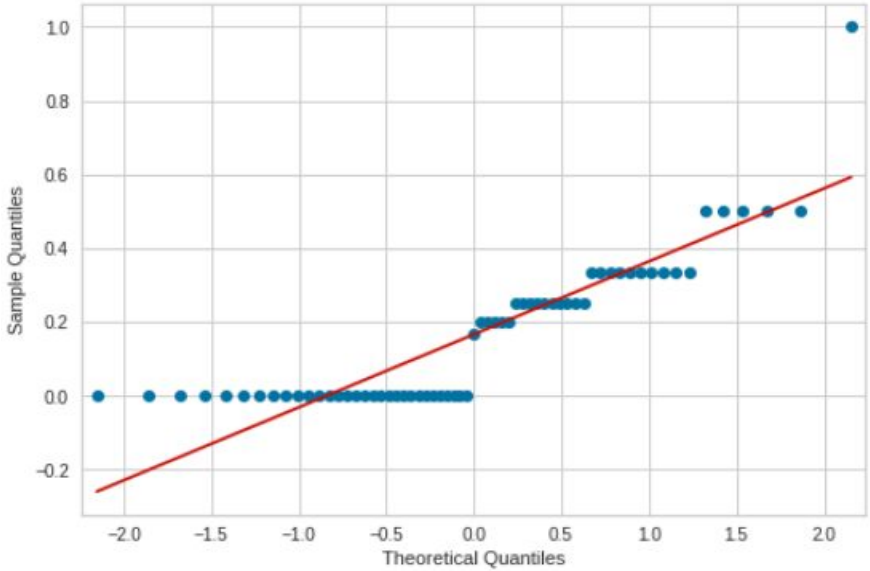



After removing Outliers

- Outliers affect the regression line badly.
- Also affect determination coefficient.

# Normality of data

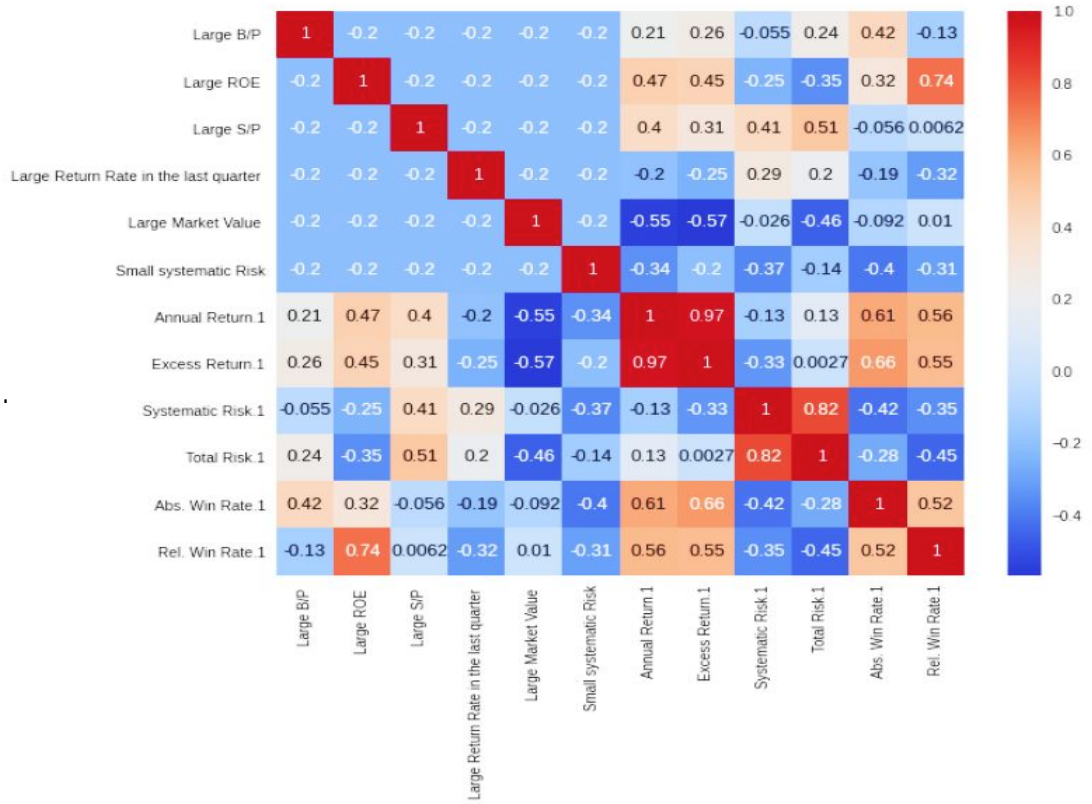


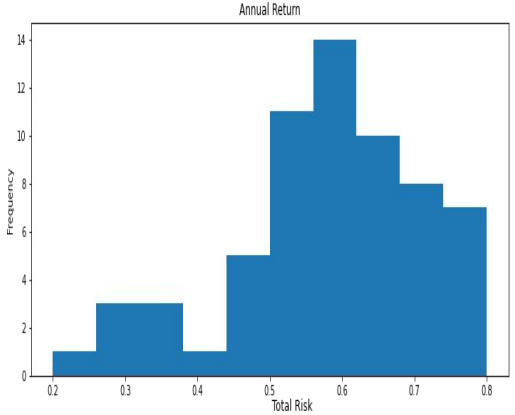

- Independent variables are not normal as they are deviating from straight line.

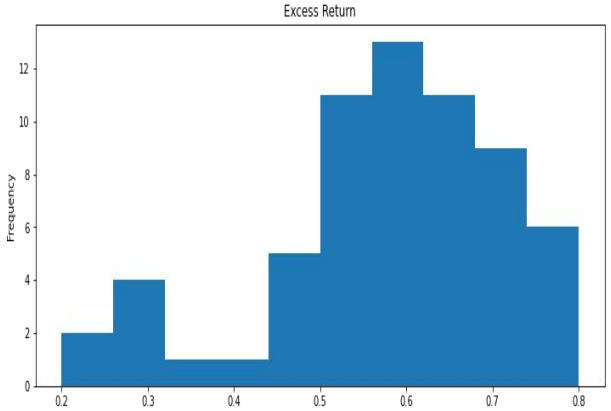
# Correlation

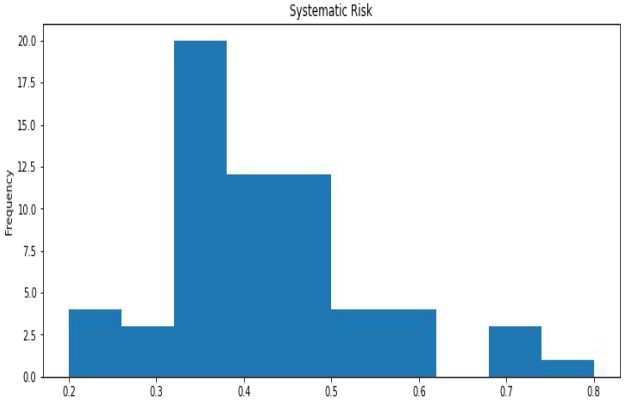
- Some of the features are correlated and some are uncorrelated.
- Values ranging from [-1, 1].

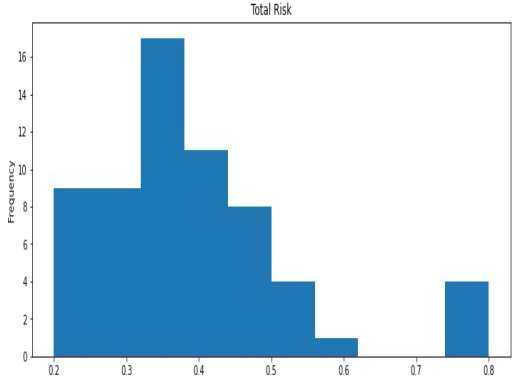


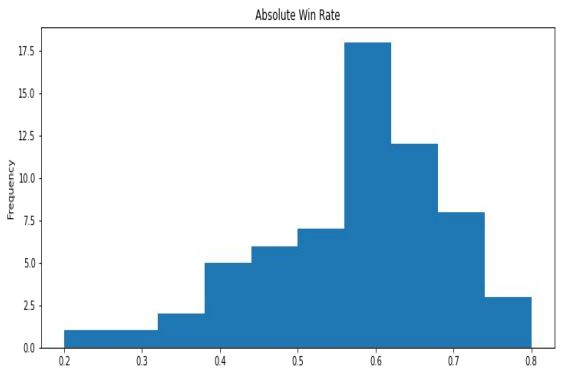
# Normality of dependent variables

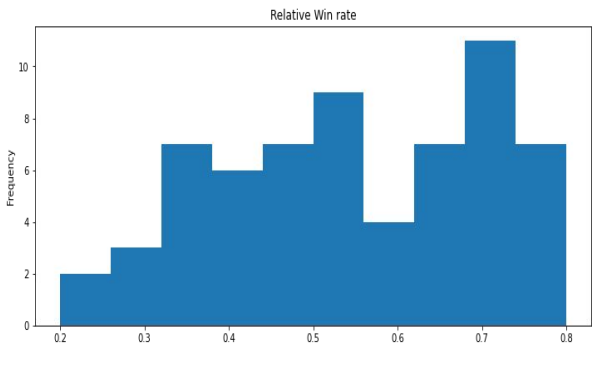


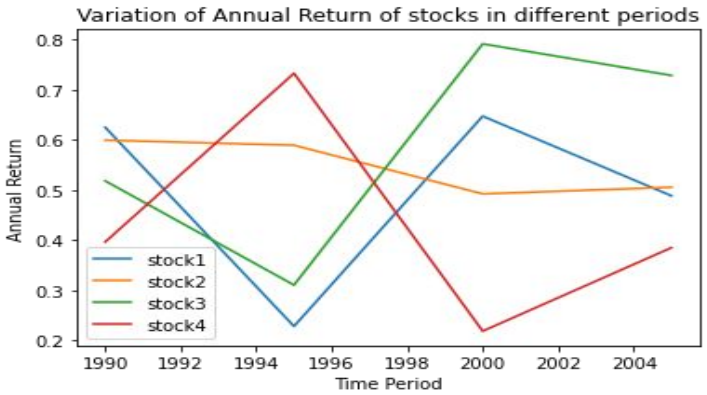



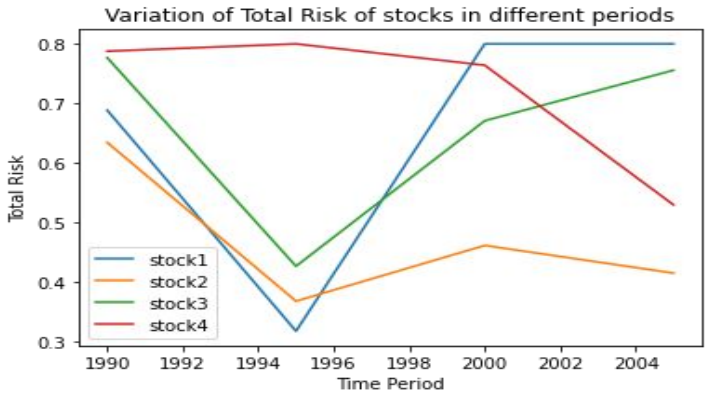



**Inference :** The above graph show the distribution of normal variables.We can say that they are Normal because they resemble bell shaped curves.

# Stocks in different periods

- Variation of Annual Return and Total Risk.

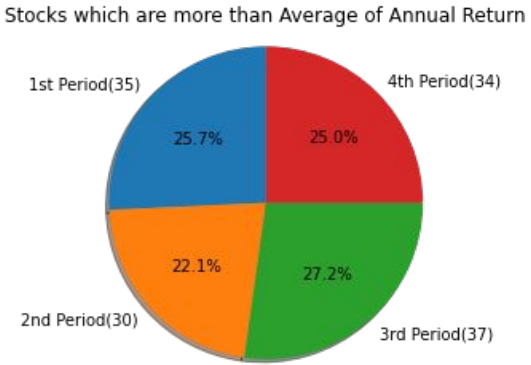


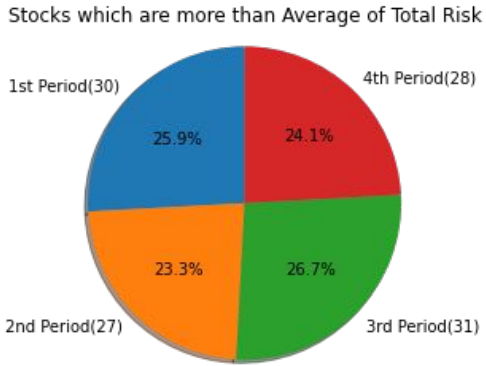




**Inference :** From period to period we can observe that values of stocks varies accordingly, stock value may increase or decrease as time passes, here for the first four stocks the change in the Annual return is shown, also total risk involved in different periods is shown.

# Performance of stocks

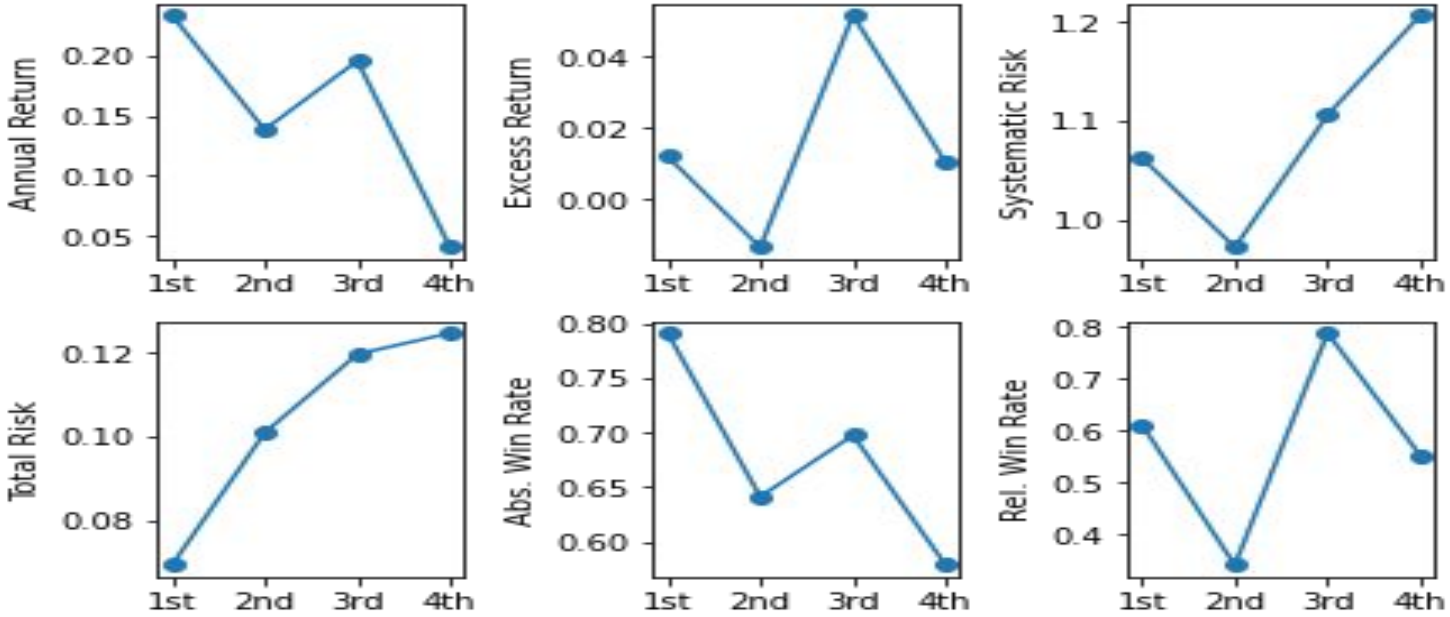
- Number of stocks performing more than average is an important statistic to classify the stocks.




**Inference :** The above graphs shows us the count of stocks which performed more than average performance in each period. We can say that in the 3rd period(2000-2005) most of the stocks performed well and gained a reasonable annual income, in 2nd period(1995-2000) less number of stocks involved in risk.

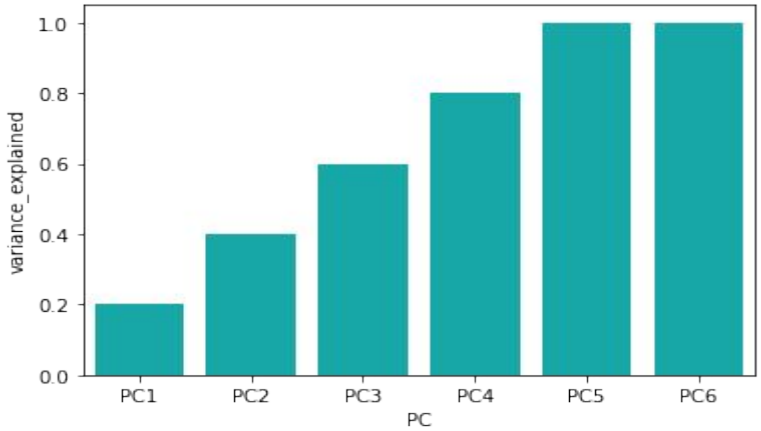
# Change in Target Variable




**Inference:** The above plots shows how the average of each target variable changes over each quarter.We can observe the trend of various attributes over different quarters.

# Principal Component Analysis

Variance Explained:  [0.19999994 0.39999988 0.59999981 0.79999975 0.99999969 1.]



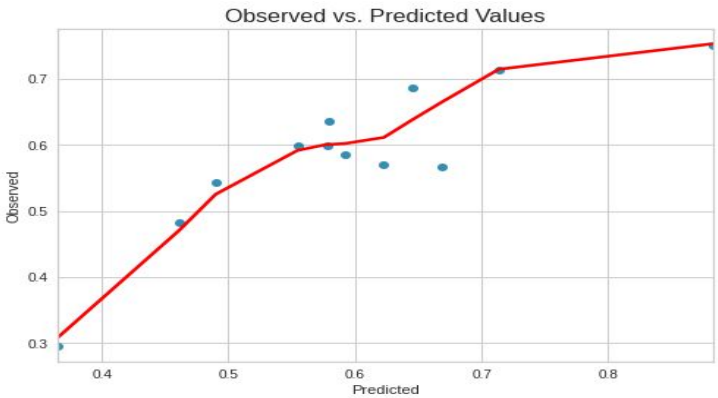

**Inference:** We can see that 100% variance of data is explained after considering the 6th principle component.Even though 6th PC explains only little variance we considered it because number of features are less and also to explain 100% variance of the data.

| | Large B/P | Large ROE | Large S/P | Large Return Rate in the last quarter | Large Market Value | Small systematic Risk |
|---|---|---|---|---|---|---|
| **PC-1** | -0.000000 | -0.280362 | -0.565721 | 0.569912 | -0.207184 | 0.483355 |
| **PC-2** | 0.912871 | -0.182574 | -0.182574 | -0.182574 | -0.182574 | -0.182574 |
| **PC-3** | -0.000000 | 0.326438 | -0.661497 | -0.080517 | 0.633831 | -0.218255 |
| **PC-4** | 0.000000 | 0.768577 | -0.159266 | 0.118150 | -0.593256 | -0.134205 |
| **PC-5** | 0.000000 | 0.155321 | -0.130445 | -0.674358 | -0.058135 | 0.707617 |
| **PC-6** | 0.408248 | 0.408248 | 0.408248 | 0.408248 | 0.408248 | 0.408248 |

PC1 - Large return rate  in last quarter
PC2 - Large B/P
PC3 - Large S/P
PC4 - Large ROE
PC5 - Small Systematic Risk
PC6 - Large Market Value

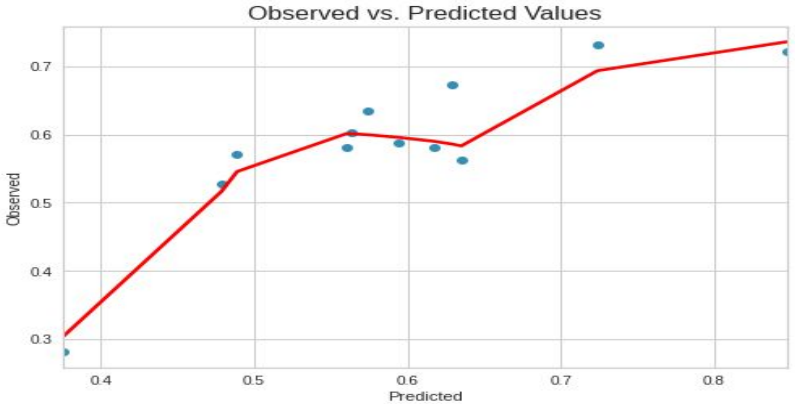
# Regression Model

Annual Return:



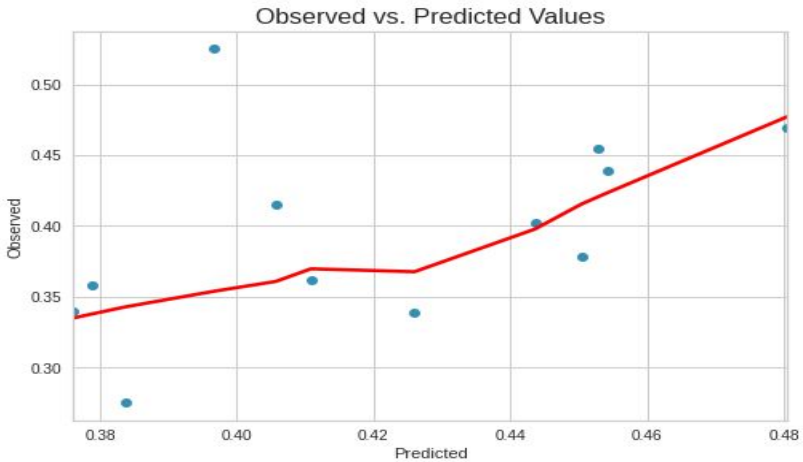

Variance Score: 0.70194
R-Squared: 0.99

Excess Return:




Variance Score: 0.671
R-Squared: 0.986

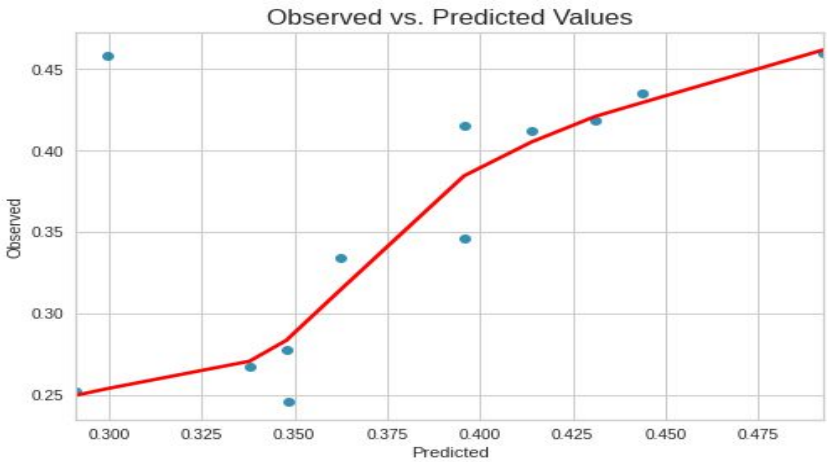
Systematic Risk:




Variance Score: 0.081

R-Squared: 0.966

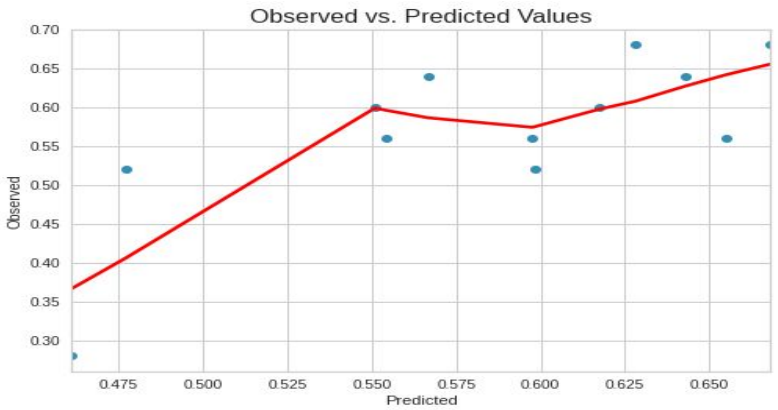
Total Risk:




Variance Score: 0.30

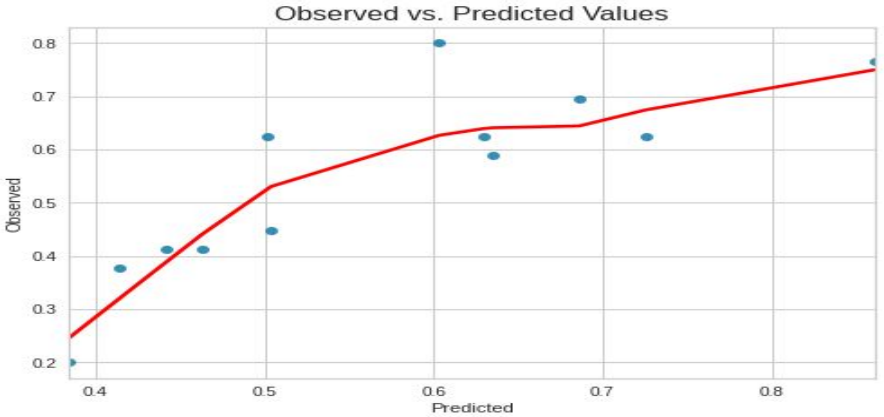
R-Squared: 0.967

Absolute Win rate:




Variance Score: 0.504
R-Squared: 0.983

Relative Win rate:




Variance Score: 0.662
R-Squared: 0.979

**Inference:** As we can see the r-square values for the models of all the dependent variable are high (above 0.96) so we can say that the error in our modeling is low,So our model performs well..

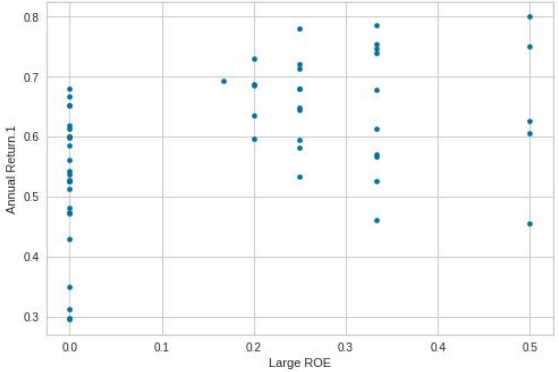
# Ols regression results

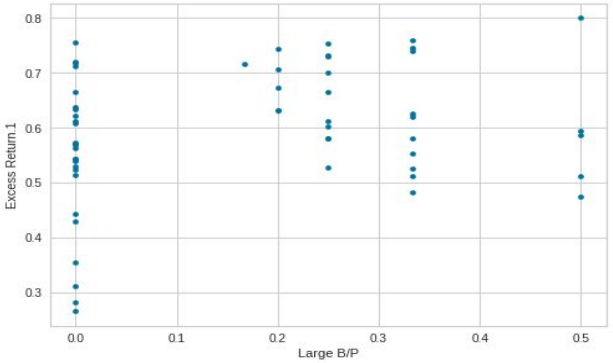
- **$R^2$-values :**
  - Y1 - 0.9900
  - Y2 - 0.986
  - Y3 - 0.966
  - Y4 - 0.967
  - Y5 - 0.983
  - Y6 - 0.979
- **Adj - $R^2$-values :**
  - Y1 - 0.988
  - Y2 - 0.984
  - Y3 - 0.961
  - Y4 - 0.962
  - Y5 - 0.980
  - Y6 - 0.975
- **DW - values**
  - Y1 - 2.099
  - Y2 - 2.016
  - Y3 - 2.020
  - Y4 - 2.038
  - Y5 - 1.902
  - Y6 - 1.945

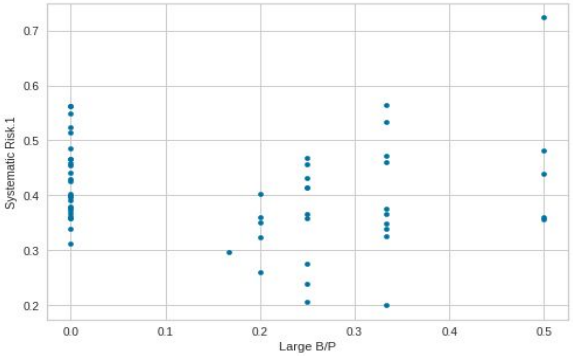
- Here, $R^2$-values are high, so we can use our data directly for the prediction.
- For the 1st four dependent variables Durbin-Watson values are >2, so they are negatively correlated, where as for 5th and 6th dependent variables DW-values are <2, so they are positively correlated .

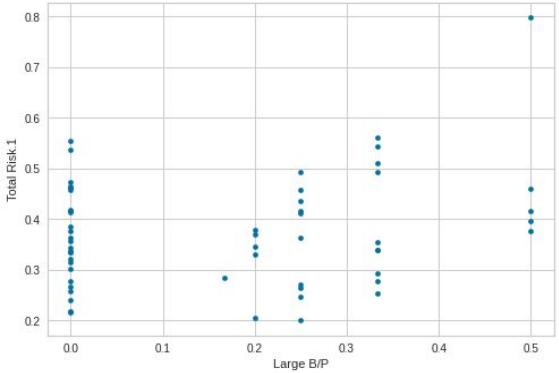
# TEST OF ASSUMPTIONS

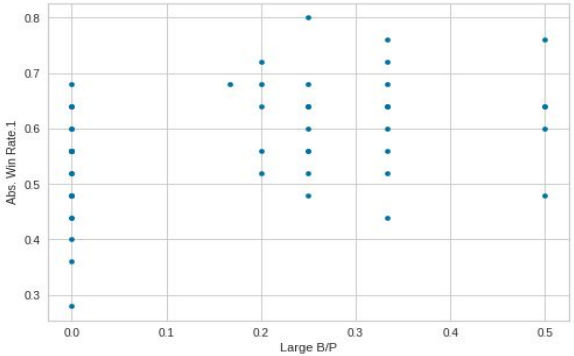
# Linearity

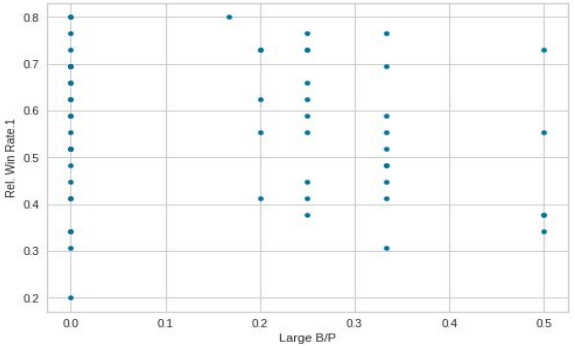

Annual Return.1
Large ROE


Excess Return.1
Large B/P


Systematic Risk.1
Large B/P

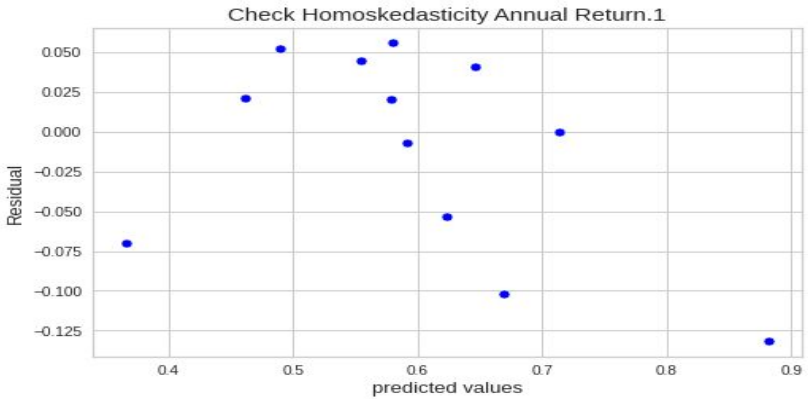

Total Risk.1
Large B/P

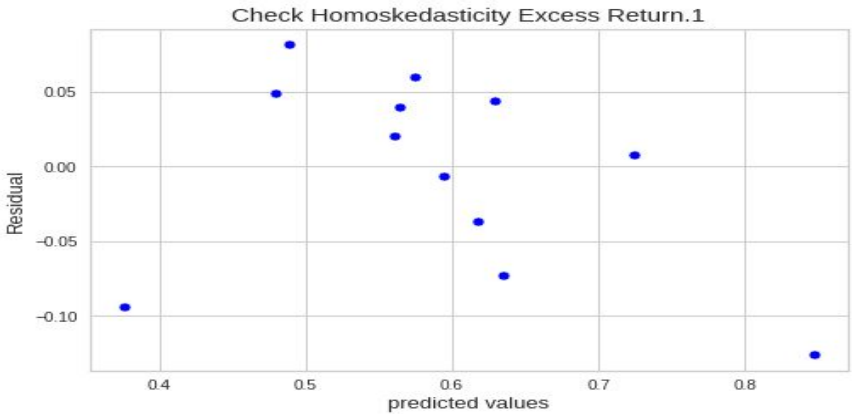

Abs. Win Rate.1
Large B/P

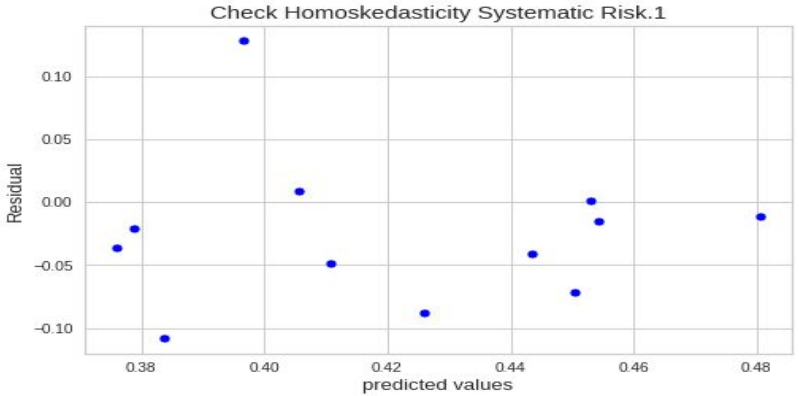

Rel. Win Rate.1
Large B/P

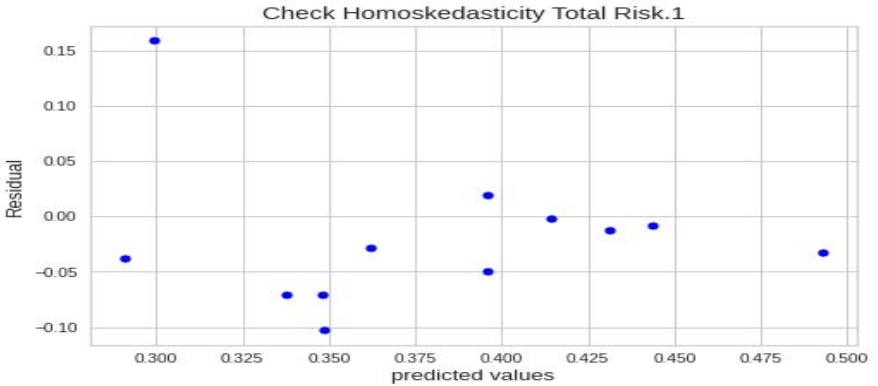
**Inference** : The above graph show the plot between independent variable and dependent variable.By observing the graphs we can say that all the points are scattered which are not completely linear.So we can say that linearity condition is not completely satisfied.

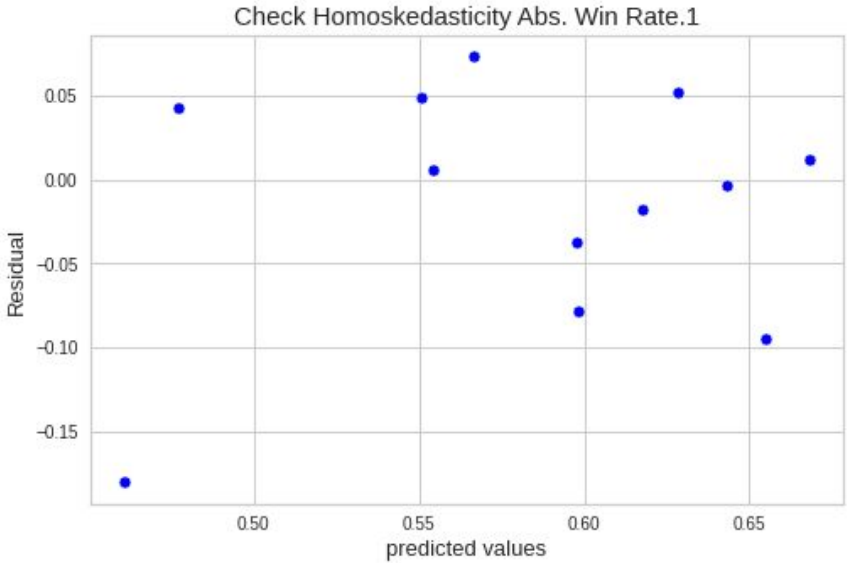
# Homoscedasticity

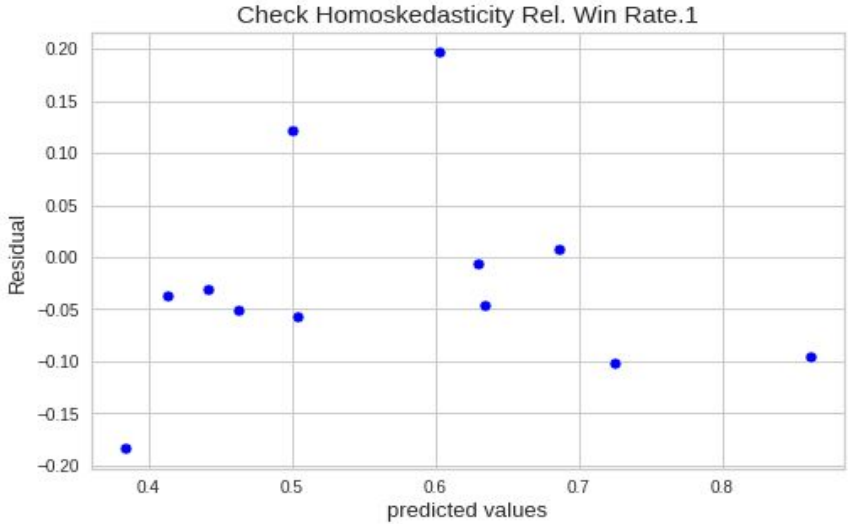


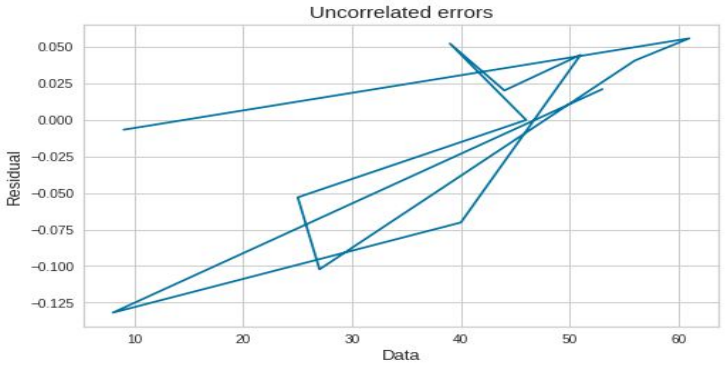


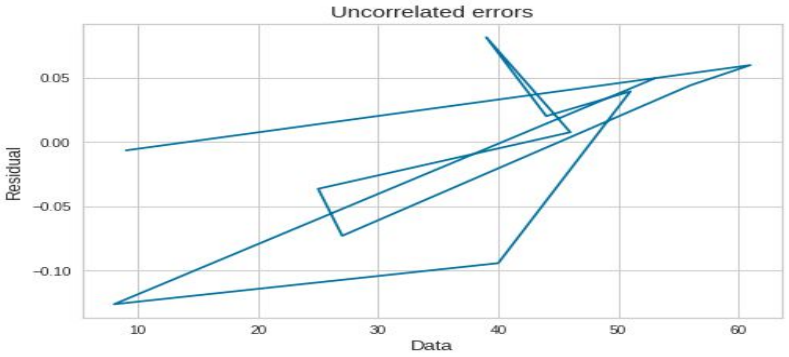

**Inference:** The above graph shows the plot between Residuals and predicted values. The points are random with no funnel shape which confirms that Homoscedasticity is true.

# Uncorrelated Errors

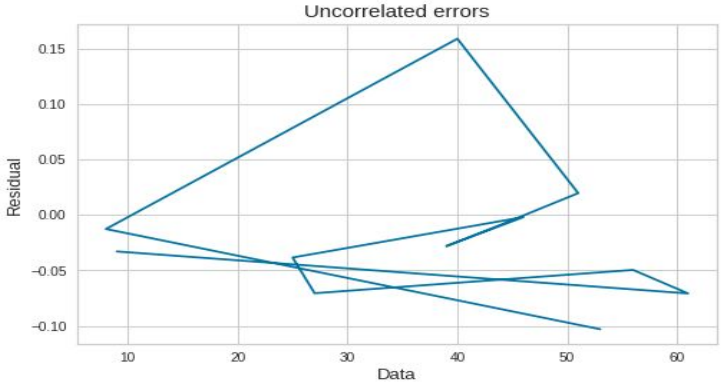
## Annual Return



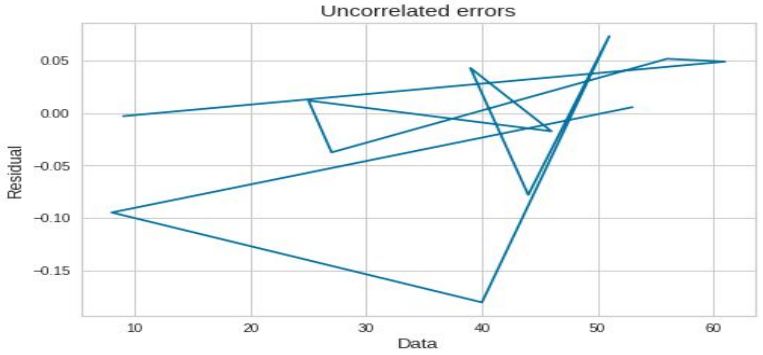

## Excess Return




## Total Risk




## Abs. Win rate

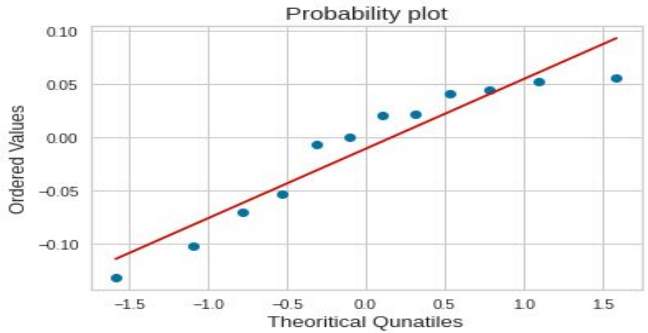


**Inference :** For the above plots, we can see some correlation or pattern between the errors, which means there are correlation errors.

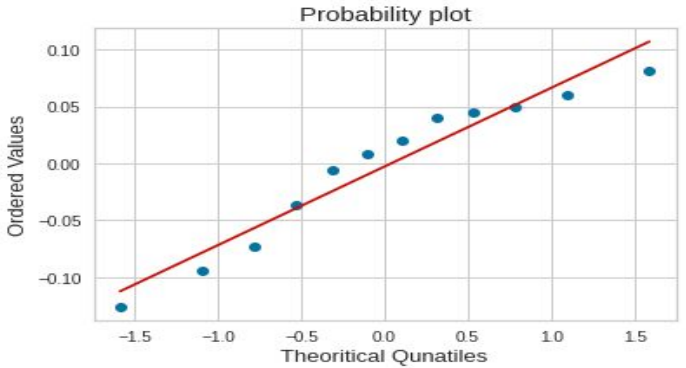
- We can also check this correlations by performing durbin watson test.
  - If Dw = 2, we can say there is no correlation
  - If Dw < 2, we can say the errors are positively correlated.
  - If Dw > 2, we can say the errors are negatively correlated.
- In the given dataset 4 dependent variables are positively correlated and 2 dependent variables are negatively correlated

# Normality of error terms

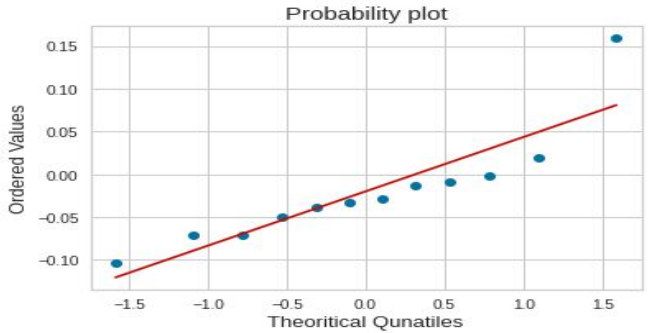
Annual Return



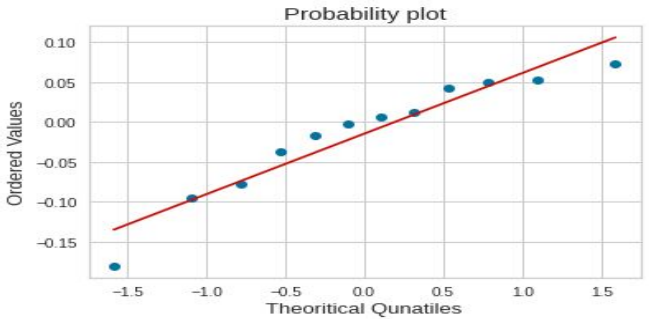

Excess return




Total Risk




Abs. Win Rate



- The above plots are between theoretical quantiles of standard normal variables and ordered values of sample quantiles.

- **Inference :** If we observe the plots for the errors, we can conclude that errors follow normal distribution, because the plot shows the fluctuation around the line and there is not much deviation, from this inference we can say that graphs are linear.

# Conclusion

After performing Linear Regression and checking the test of assumptions we can say that the goodness of model judged based on R-squared value is High.We can achieve higher r-square and less error if Linearity condition between dependent and independent variables holds.

# THANK YOU