**MA 541 Statistical Methods**

PROJECT SUPERVISOR AND PROFESSOR - HADI SAFARI

8 December 2021

AUTHOR: YASH ADMUTHE (CWID - 10469443)

# 

# Contents

1. **Section 1 - The introduction or objective of your report**
2. **Section 2 - Part 1 to Part 10**
3. **Section 3 - Discussions or Improvements**
4. **References**

# 1 Section 1 - Objective of the report

**Objective:**

The objective of this project is to perform exploratory data analysis on the given sample data using various statistical methods like Central Limit Theorem, Hypothesis testing , Visualization, etc and apply simple and multi linear regression models.

**2. Section 2 - Part 1 to Part 10**

**Part 1: Meet the data**

Mean - The mean is used to measure central tendency data.

A picture containing shape

Description automatically generated

Standard deviation: In statistics, the standard deviation is a measure of the amount of variation or dispersion of a set of values

Text

Description automatically generated with medium confidence

Data consist of four columns, EFT, Oil, Gold, and JPM with a sample size of 1000.

Mean of all the columns :

ETF Mean: 121.1529600120001

Oil Mean: 0.0010300354937470015

Gold Mean: 0.0006628360819999998

JPM Mean : 0.0005304110210000002

The standard deviation of all the columns :

ETF standard deviation : 12.563503845944297

Oil standard deviation 0.021082349463798354

Gold standard deviation 0.011283414317347945

JPM standard deviation : 0.011011052723643009

Correlation among all the columns:

Table

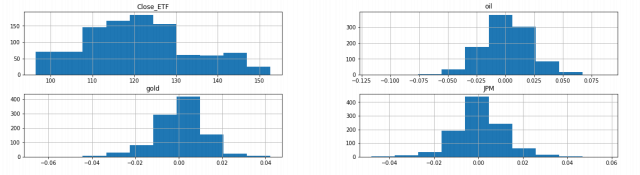
Description automatically generated

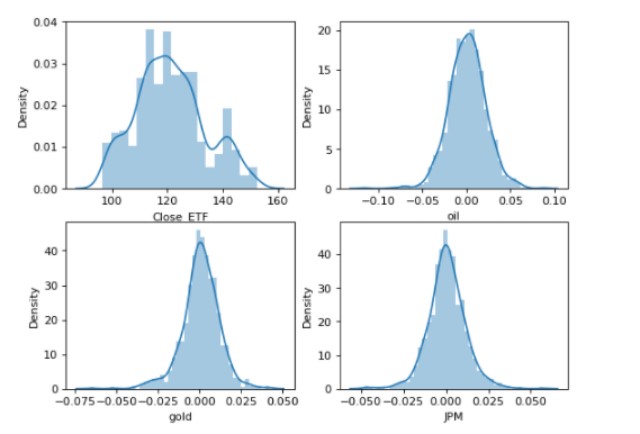
As we can see there is no strong correlation between any columns. All the columns have very low co-relation values this suggests that there is no relation between columns

**Part 2: Describe your data**

1) A histogram for each column.

The purpose of a histogram is to graphically summarize the distribution of data.



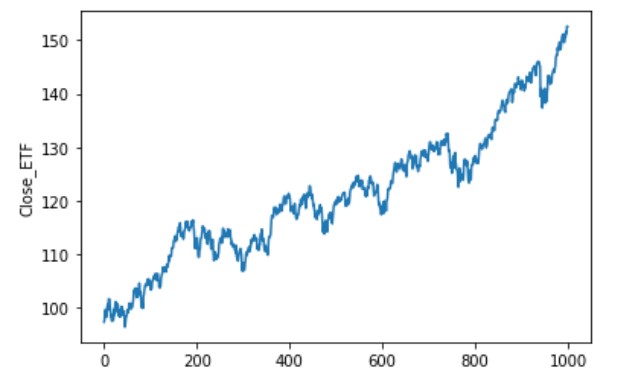


Showing histogram of each column. By histogram, it looks like each column approximately follows a normal distribution.

2) A time series plot for each column

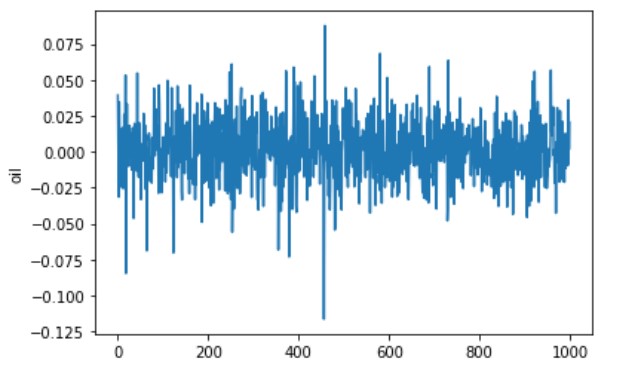
Time-series graphs can be used to visualize trends in counts or numerical values overtime or some other ordering

**ETF**



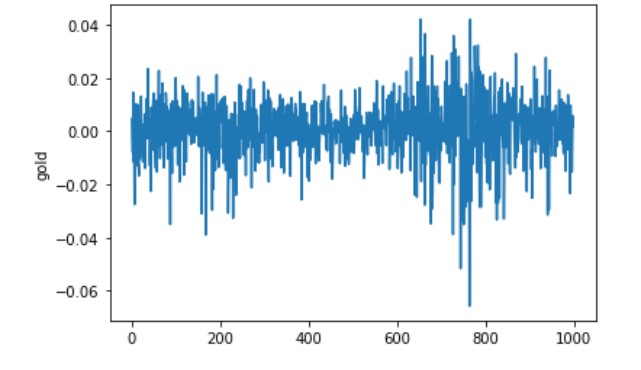
ETF graph shows an upward moving trend.

**Oil**



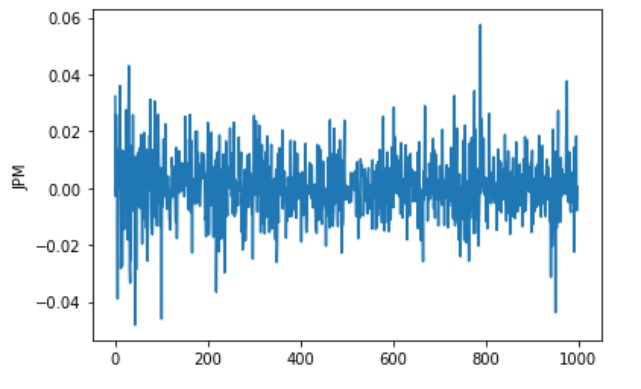
Oil graph showing random variation.

**Gold**



Gold graph showing random variation.

**JPM**



JPM graph showing random variation.

3) A time series plot for all four columns

Time series graph showing all four columns

Chart, line chart

Description automatically generated

ETF column showing upward variation and other three columns showing random variation.

4) Three scatter plots to describe the relationships between the ETF column and the OIL column; between the ETF column and the GOLD column; between the ETF column and the JPM column, respectively.

A scatter plot is used to observe and visually display the relationship between variables.

**ETF column and the OIL column**

Chart, scatter chart

Description automatically generated

By scatter plot, we can observe that there is no relation between ETF and oil.

**ETF column and the Gold column**

Chart, scatter chart

Description automatically generated

By scatter plot, we can observe that there is no relation between ETF and Gold.

**ETF column and the JPM column**

Chart, scatter chart

Description automatically generated

By scatter plot, we can observe that there is no relation between ETF and JPM.

**Part 3: What distribution does your data follow**

We will check if data is normally distributed or not using the below methods :

Graphs for Normality test

1. Histogram

A histogram is a graphical display of data using bars of different heights. A histogram is the most commonly used graph to show frequency distributions.

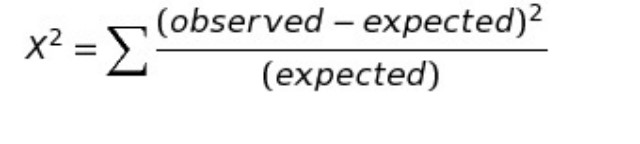
1. Quantile-Quantile Plot

Quantile-Quantile or QQ plot used to check the distribution of a data sample. A perfect match for the distribution will be shown by a line of dots on a 45-degree angle from the bottom left of the plot to the top right. Often a line is drawn on the plot to help make this expectation clear. Deviations by the dots from the list shows a deviation from the expected distribution

Statistical Tests for Normality

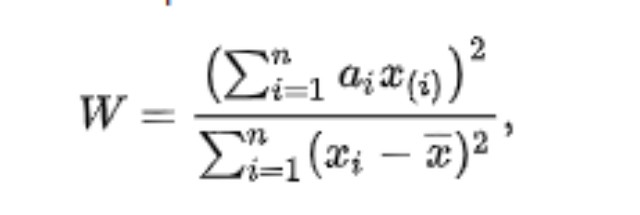
1. Chi-Square Test for Normality

The chi-square goodness of fit test can be used to test the hypothesis that data comes from a normal hypothesis



1. Shapiro-Wilk Test

The Shapiro Wilk test checks if the normal distribution model fits the observations.



xi = are the ordered random sample values ai = are constants generated from the co-variances, variance s, and means of the sample (size n) from a normally distributed sample. W = small values indicate your sample is not normally distributed

**ETF data :**

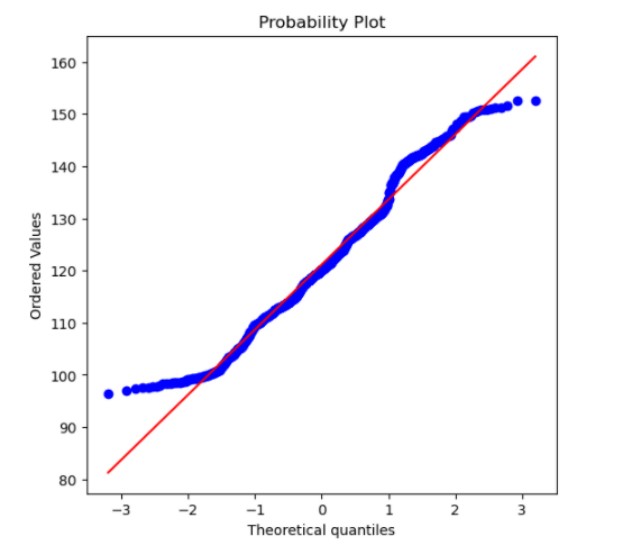
1. Histogram



The conclusion from the histogram: Does not look like a smooth normal distribution.

Looks skewed and bi-modal.

1. Quantile-Quantile Plot



We can see at many places data is deviating from the line. This shows it's s not exact normal distribute assumption motion for below test with alfa 0.05

Ho : ETF follows the normal distribution.

H1: ETF does not follow the normal distribution.

1. Chi-Square Test for Normality Calculated using Python:

Statistics=1302.829, p=0.000

The sample does not look Gaussian (reject H0)

1. Shapiro-Wilk Test

Calculated using Python:

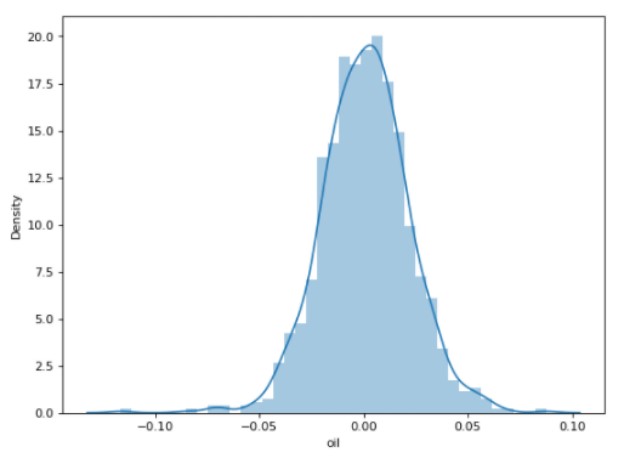
Statistics=0.980, p=0.000

The sample does not look Gaussian (reject H0)

Conclusion for ETF column: After performing the above 4 normality tests, we conclude it’s not exactly following a normal distribution.

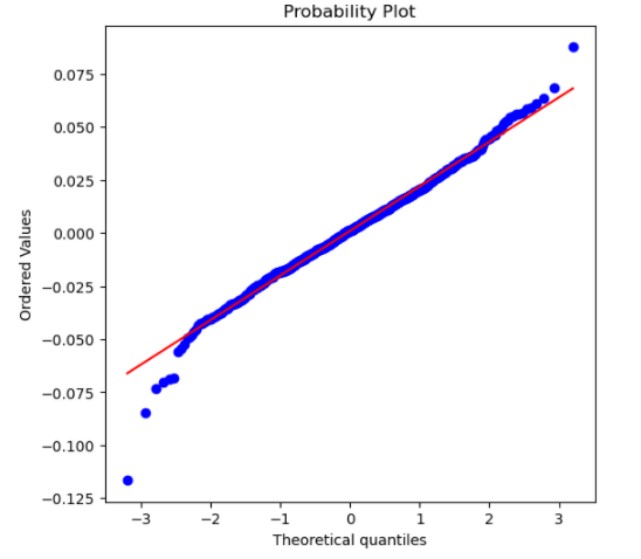
**Oil data :**

1. Histogram



The conclusion from histogram: Approximately looks like a normal distribution.

1. Quantile-Quantile Plot



Most of the line is online except few outliers. This shows it follows a normal distribution assumption for below test with alfa 0.5

Ho : Oil follows normal distribution.

H1 : Oil does not follows normal distribution.

1. Chi-Square Test for Normality

Statistics=431.505, p=1.000

Sample looks Gaussian (fail to reject H0)

1. Shapiro-Wilk Test

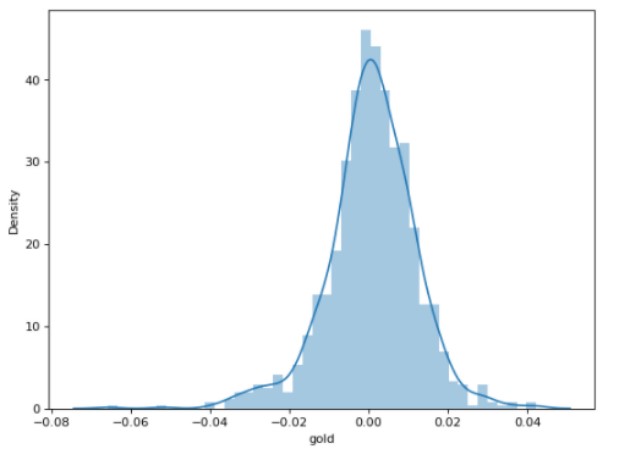
Statistics=0.989, p=0.000

Sample does not look Gaussian (reject H0)

Conclusion for Oil column : After performing above 4 normality test, we conclude its approximately following normal distribution.

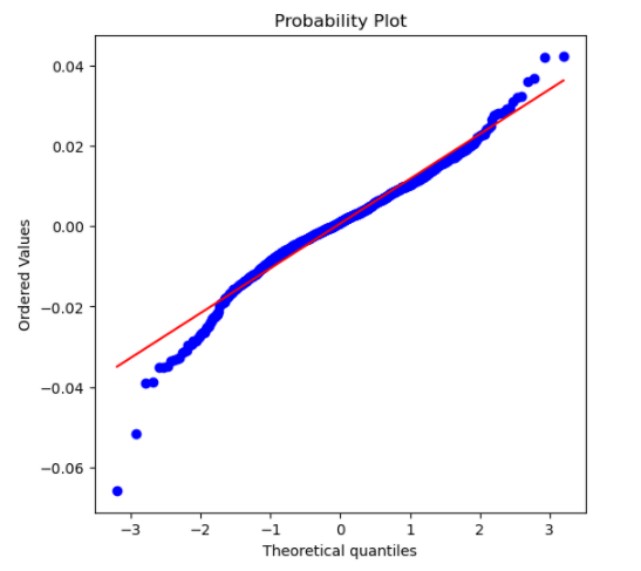
**Gold data :**

1. Histogram



Conclusion from histogram : Approximately looks like normal distribution.

1. Quantile-Quantile Plot



Conclusion from QQ Plot : Most of line is on line except few outliers. This shows its follow normal distribution.

Assumption for below test with alfa 0.05

Ho : Gold follows normal distribution.

H1 : Gold does not follows normal distribution.

1. Chi-Square Test for Normality

Statistics=192.077, p=1.000

Sample looks Gaussian (fail to reject H0)

1. Shapiro-Wilk Test

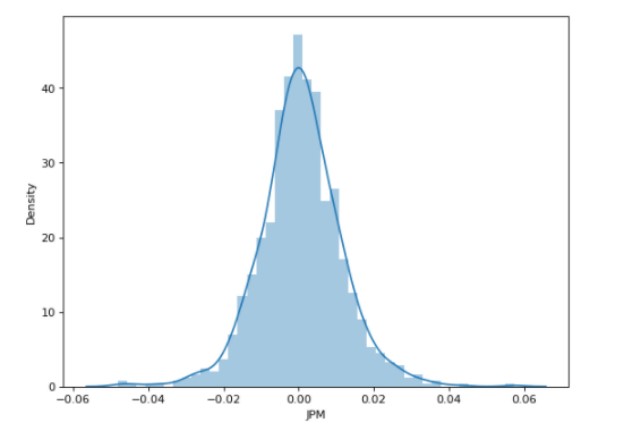
Statistics=0.969, p=0.000

Sample does not look Gaussian (reject H0)

Conclusion for Gold column : After performing above 4 normality test, we conclude its approximately following normal distribution.

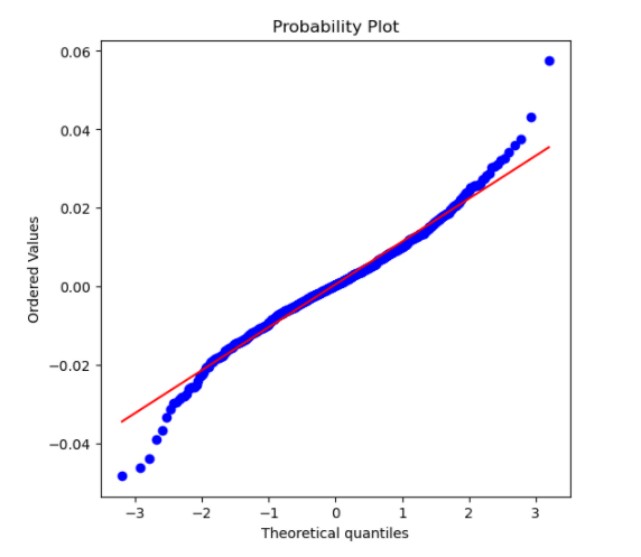
**JPM data :**

1. Histogram



Conclusion from histogram : Approximately looks like normal distribution.

2-Quantile-Quantile Plot



Conclusion from QQ Plot : Most of line is on line except few outliers. This shows its follow normal distribution.

Assumption for below test with alfa 0.05

Ho : JPM follows normal distribution.

H1 : JPM does not follows normal distribution.

1. Chi-Square Test for Normality

Statistics=228.584, p=1.000

Sample looks Gaussian (fail to reject H0)

1. Shapiro-Wilk Test

Statistics=0.980, p=0.000

Sample does not look Gaussian (reject H0)

Conclusion for JPM column : After performing above 4 normality test, we conclude its approximately following normal distribution.

**Part 4: Break your data into small groups and let them discuss the importance of the Central Limit Theorem**

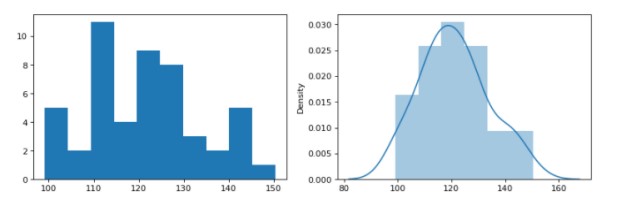
In probability theory, the central limit theorem establishes that, in many situations, when independent random variables are summed up, their properly normalized sum tends toward a normal distribution even if the original variables themselves are not normally distributed..

1. Calculate the mean and the standard deviation the population.

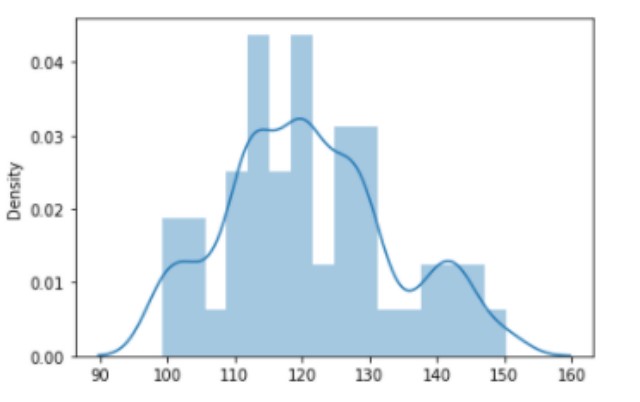
Mean of ETF column : 121.1529600120001

Standard Deviation ETF column : 12.563503845944297

1. Break the population into 50 groups sequentially and each group includes 20 values.
2. Calculate the sample mean of each group. Draw a histogram of all the sample means. Comment on the distribution of these sample means, i.e., use the histogram to assess the normality of the data consisting of these sample means.
3. Calculate the mean and the standard deviation 0f the data including these sample means. Make a comparison between mean and samples mean, between standard deviation.



If we take 50 groups sequentially 1000 times then we will get below histogram



Mean of ETF column : 121.1529600120001

Standard Deviation ETF column : 12.563503845944297

Comparison between Mean and Sample mean:

Mean of population: 121.1529600120001 and mean of samples : 121.15296001199997

Mean of population and sample mean are almost same.

sd/sqrt(n) : 1.2563503845944297 and standard deviation of samples : 12.16375686089257

1. Are the results from Items 3) and 4) consistent with the Central Limit Theorem? Why?

We didn’t get approximate normal distribution as we have used sequential sample and sample size of 20. This not consistence with Central limit theorem. Normal distribution still similar to Part 3.

1. Break the population into 10 groups sequentially and each group includes 100 values.
2. Repeat Items 3) 5).

Chart, waterfall chart

Description automatically generated

Comparison between Mean and Sample mean:

Mean of population : 121.1529600120001 and

mean of samples : 121.15296001199997 Mean of population and sample mean are almost same.

sd/sqrt(n) : 1.2563503845944297

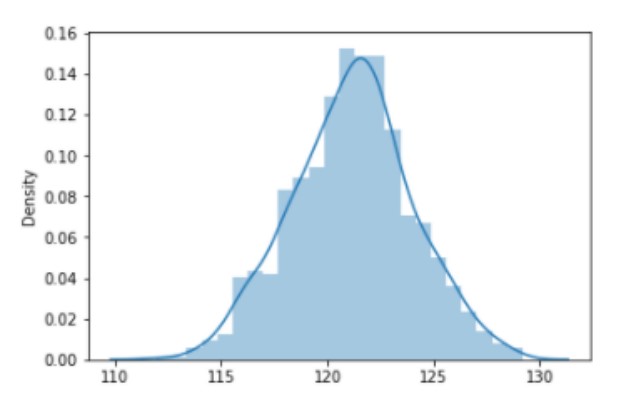
Standard deviation of samples : 12.16375686089257

Conclusion : We didn’t get approximate normal distribution as we have used sequential data. Sight improvement over 20 size sample. It not consistence with Central limit theorem.

1. Generate 50 simple random samples or groups (with replacement) from the population. The size of each sample is 20, i.e., each group includes 20 values.
2. Repeat Items 3) 5).



If we take 50 groups random sample 1000 times then we will get below histogram



Mean : 120.983180221

Standard deviation : 2.6861368907082497

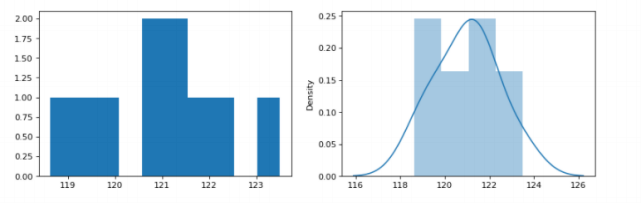
Comparison between Mean and Sample mean:

Mean of population : 121.1529600120001 and mean of samples : 120.983180221 Mean of population and sample have sight difference.

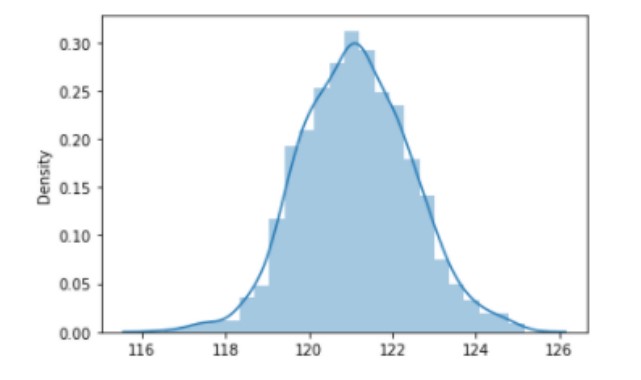
sd/sqrt(n) : 1.7767477529860964 and standard deviation of samples : 2.6861368907082497

Conclusion : We get approximate normal distribution as we have used random data sample. But still need improvement as sample size is 20. Improvement over sequential data and part 3 distribution. Its consistent with Central Limit Theorem.

1. Generate 10 simple random samples or groups (with replacement) from the population. The size of each sample is 100, i.e., each group includes 100 values.
2. Repeat Items 3) 5).



If we take 20 groups random sample 1000 times then we will get below histogram



Mean : 120.763279982

Standard deviation : 1.6665020434714146

Comparison between Mean and Sample mean:

Mean of population : 121.1529600120001 and Mean of samples : 120.763279982 Mean of population and sample have sight difference.

sd/sqrt(n) : 1.7767477529860964 and Standard deviation of samples : 1.6665020434714146

We get approximate normal distribution as we have used random data sample of 100. Improvement over sequential data, random sample with 20 size and part 3 distribution. Its consistent with Central Limit Theorem.

12) In Part 3 of the project, you have figured out the distribution of the population (the entire ETF column). Does this information have any impact on the distribution of the sample mean(s)? Explain your answer.

The central limit theorem states that if you have a population with mean and standard deviation and take su ciently large random samples from the population with replacement, then the distribution of the sample means will be approximately normally distributed.

In part 3 we observed that ETF column was not following normal distribution.

Sequential sample :

We observed that when we took 50 sample sequentially, then histogram of sample means are more or less similar to histogram of population in part 3. We observed that when we took 100 sample sequentially, then histogram of sample means are smother than 50 group sample but still similar to histogram of population in part 3.

Random sample :

We observed that when we took 50 sample randomly with replacement, then histogram of sample means better normally distributed than sequential sample and distribution in part 3.

We observed that when we took 100 sample randomly with replacement, then histogram of sample means form smother normal distribution than 50 sample randomly with replacement, sequential sample data and distribution in part 3.

**Conclusion : This concluded that as we take more and more random sample size with replacement the shape of sampling distribution approximately normal distribution.**

**Part 5: Construct a confidence interval with your data**

Confidence intervals are estimate ranges for population parameters based on the sample observations. The two main factors that decide the range of these estimates are:- 1.The sample size and 2.The population variance. Greater the sample size, greater is the amount of information/data we have to make a good guess regarding the range; which leads to a small (more accurate) range for predicting the population parameter and vice versa. Greater the population variance, greater is the deviation in values which makes it tougher to predict the population parameter with certainty. This causes the range of the intervals to increase and vice versa.

1. : Pick up one of the 10 simple random samples you generated in Step 10) of Part 4, construct an appropriate 95 confidence interval of the mean myu.

**Solution:**

For this particular question, we are supposed to take one of the 10 randomly generated samples(with 100 data points each) from step 10 of part 4 and calculate an interval with 95 percent confidence, estimating the population mean myu.

We have used the t.norm.interval() method available in the scipy library. We have made use of the normal distribution and not the student’s t distribution in order to compute the interval since we have the sample size n = 100

We have the sample mean(x-bar), the sample standard deviation(s) and the sample size(n). Now all we have to do is calculate the t-value corresponding to our sample’s degrees of freedom and significance value(95 percent).This t-value was found out to be 1.960.

After calculation, the result that we obtained was (119.4678, 124.8443).

Just to confirm the result, we also made use of the st.norm.interval() method o↵ered by the scipy library which is used to calculate the confidence interval for samples. And the results obtained were the same upto 1 decimal point.

2. : Pick up one of the 50 simple random samples you generated in Step 8) of Part 4, construct an appropriate 95 confidence interval of the mean myu.

**Solution:**

For this particular question, we are supposed to take one of the 50 randomly generated samples(with 20 data points each) from step 8 of part 4 and calculate an interval with 95 percent confidence, estimating the population mean myu.

We have used the Wilcoxon signed ranked test in r in order to get the results, since the etf feature is not perfectly normal and on top of that the sample size is less than 30.The result that we obtained is (119.620 , 132.395).Also, just for the sake of it, we also tried calculating the confidence interval using the student-t distribution method. scipy.stats.t.interval() method available in the scipy library since the size of the sample size is small. The formula for which is is the same as above, the di↵erence would be in the values of sample mean which in this case is 125.42, degrees of freedom, sample standard deviation (12.20) and the sample size(20).The result obtained using this method was (119.56, 131.28).

3. : In Part 1, you have calculated the mean myu of the population (the entire ETF column) using Excel function. Do the two intervals from 1) and 2) above include (the true value of) the mean myu? Which one is more accurate? Why?

**Solution:**

The population mean (121.1529) is included in the confidence intervals calculated by both the samples. The confidence interval calculated in step 1 is obviously more accurate since the sample size in that case is larger(5 times) than the sample size from step 2.

**Part 6: Form a hypothesis and test it with your data**

Hypothesis testing is a procedure to take an inferential decision regarding a business problem with help of statistics. Every hypothesis problem consists of two parts:- 1. Null hypothesis and 2. Alternate hypothesis. In case of Z-test, the null hypothesis is rejected when the z-statistic lies on the rejection region, which is determined by the significance level and the type of tail (two-tailed, left-tailed or right-tailed).

1. : Use the same sample you picked up in Step 1) of Part 5 to test Ho: myu=100 and Ha:myu!=100 at the significance level 0.05. What’s your conclusion?

**Solution:**

Since in this case, the population standard deviation is known and the number of samples is big enough, we will make use of the z-test. We get the value of z as 17.63

Now corresponding to the significance value given(0.05), we need to find the z-critical value for it, which is obtained from the z-table Therefore the rejection region for this two-tailed test is R=z:—z—¿1.96 We can also confirm this using the p-value approach, the p-value is coming out to be 0 in this case.

Hence, we reject the null hypothesis in both the methods.

1. : Use the same sample you picked up in Step 2) of Part 5 to test Ho: myu=100 and Ha:myu!=100 at the significance level 0.05. What’s your conclusion?

**Solution:**

Since the size of the data-sample in question is less and the distribution is not perfectly normal, we decided to go with the Wilcoxon signed ranked test in r. The test type was given as ”two-sided” and the hypothesis mean was taken as 100.Since the p-value came out to be so less, we reject the null hypothesis.

1. : Use the same sample you picked up in Step 2) of Part 5 to test Ho:sigma = 15 vs Ha:sigma != 15 at the significance level 0.05. What’s your conclusion?

**Solution:**

We will be using the chi-square technique in order to solve this part since it is used to test if the variance of a population is equal to a specified value.

Where (N-1) is degrees of freedom,s is the sample standard deviation and sigma0 is the hypothetical standard deviation

Using the formula we get,

T = 12.569

Now, the critical values for significance level 0.05 for a two tailed chi-square test is given by *x*2 (*↵*/2,N-1) = 8.907 *x*2 (1-*↵*/2,N-1) = 32.852

Hence critical region: Reject H0 if T *<*8*.*907*orT*¿32*.*852

Since the T value we received is greater than 8.907, we failed to reject the null hypothesis.

1. : Use the same sample you picked up in Step 2) of Part 5 to test Ho:sigma = 15 vs Ha:sigma ¡ 15 at the significance level 0.05. What’s your conclusion?

**Solution:**

The entire procedure from above will be the same for this part, the only di↵erence will be the critical value.

Now, the critical value for significance level 0.05 for a lower tailed chi-square test is given by ⇤⇤ 2(*↵/*2*,N* 1) = 12*.*569

Hence critical region: Reject H0 if *T <* 10*.*117

Since the T value we received is less than 10.117, we failed to reject the null hypothesis.

**Important Note**: All of tests performed until now were performed using random samples with a random-state taken as 100, but it was noted that for Part 6 questions 3 and 4, taking a random state of 0 bought some interesting changes to the results.

The test statistic value for this sample was coming out to be 6.96 which is below 8.907 and 10.117, hence the null hypothesis was getting rejected in this case. Hence it should be noted that the null hypothesis is getting rejected in some of the samples.

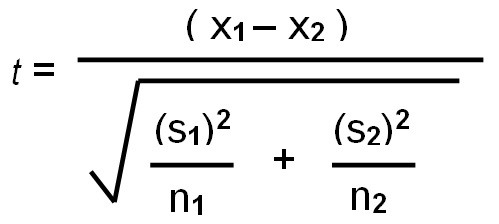
**Part 7: Compare your data with a different data set**

1. : Consider the entire Gold column as a random sample from the first population, and the entire Oil column as a random sample from the second population. Assuming these two samples be drawn independently, form a hypothesis and test it to see if the Gold and Oil have equal means in the significance level 0.05.

**Solution:**

For this question, we need to compare the means of two independent features and we decided to do it using the two tailed independent t-test. The method ttestind is used in order to conduct this test and is taken from the scipy.stats module.

The formula for t-statistic is given by:-



The result of the operation was obtained as follows:- Statistics=0.485, p=0.6274695258

Since the p-value is greater than 0.05, we accept the null hypothesis.

1. : Subtract the entire Gold column from the entire Oil column and generate a sample of di↵erences. Consider this sample as a random sample from the target population of di↵erences between Gold and Oil. Form a hypothesis and test it to see if the Gold and Oil have equal means in the significance level 0.05.

**Solution:**

For this question, we need to compare the means of two independent features(oil and gold) using the paired samples t-test. The method ttestrel is used in order to conduct this test and is taken from the scipy.stats module. And we have also solved it manually using the formula.



The result of the operation was obtained as follows:- statistic=0.54133, pvalue=0.5884

Since the p-value is greater than 0.05, we accepted the null hypothesis.

1. : Consider the entire Gold column as a random sample from the first population, and the entire Oil column as a random sample from the second population. Assuming these two samples be drawn independently, form a hypothesis and test it to see if the Gold and Oil have equal standard deviations in the significance level 0.05.

**Solution:**

Since we are required to form a hypothesis regarding the variance of two groups, we are should be using the F-test.

Following the formula we get the f-value = 12.187

After this step, we need to calculate the degrees of freedom for both the samples. They come out to be 998 for both the samples.

Then we find out the critical values: Fl and Fu corresponding to the significance level and the degrees of freedom. They come out to be Fl=0.883 and FU = 1.132.

Since the calculated value is higher than the table value, we reject the null hypothesis.

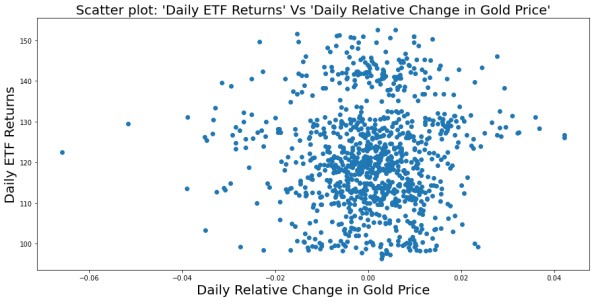
**Part 8: Fitting the line to the data**

(Consider the data including the ETF column and Gold column only.)

Using Python,

1. Draw a scatter plot of ETF (Y) vs. Gold (X). Is there any linear relationship between them which can be observed from the scatter plot?

A Scatter/Dot plot of Daily ETF returns (Y) Vs. Daily relative change in Gold price (X) is as follows:



Observations:

* + - On the x axis the data for ”Daily Relative Change of Gold price” does not start from 0. On the y axis the values for the variable ”Daily ETF returns” does not start from 0 either. Starting points of the graph are on x axis:-0.065804741 on y axis: 96.419998
    - By looking at the dot/scatter plot it can be said that there is linear relationship between ”Daily ETF returns” and ”Daily Relative Change of Gold price”.
    - But just by looking at the above plot it is hard to say magnitude and direction of co-relation between both of them. (That we can find out using Coe cient of Correlation(r) as follows.

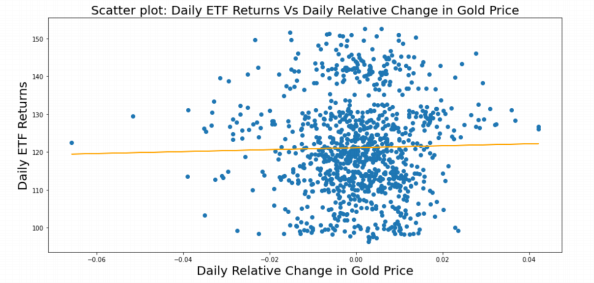
1. Calculate the coe cient of correlation between ETF and Gold and interpret it.
   * Correlation coe cients are used to measure how strong a relationship is between two variables. Returns a value between -1 and 1. • Where 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other.
   * -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other.
   * Zero means that for every increase, there isn’t a positive or negative increase. The two just aren’t related.
   * The absolute value of the correlation coe cient gives us the relationship strength. The larger the number, the stronger the relationship.
   * However the baseline may di↵er case to case to decide strong or weak relation. For example, | *.*75| = *.*75, which has a stronger relationship than *.*65.
   * There are several types of correlation coe cient, but the most popular is Pearson’s. Pearson’s correlation (R) is a correlation coe cient commonly used in linear regression. Pearson’s correlation coefficient(Pearson Product Moment Correlation (PPMC)) formula is as follows:

Interpretation of the value of r that we got:

* + The |*r*| value for we got is 0.023.
  + It is positive. Indicates there is a positive co-relation between two variables.
  + If we take *r*0*.*5 as a strong linear relationship then in that case the relationship between between these two variables will be **weak positive linear relationship**
  + That is for every positive increase in one variable, there is a positive increase of a fixed proportion in the other.

1. Fit a regression line (or least squares line, best fitting line) to the scatterplot. What are the intercept and slope of this line? How to interpret them?

The Regression Line to the Scatter plot is as follows:



Interpretation of Slope and intercept of this line:

* + For the linear relationship between ETF(the daily ETF return) and Gold(the daily relative change in the gold price) we got the slope: 25.604 and intercept: 121.135.
  + The slop value indicates: the rate of change in ”daily ETF return” per unit change of ”daily relative change in Gold price” is by 25.6043 • On the other hand intercept value shows that the value of the daily ETF return is 121.136 if the daily relative change in the gold price is 0. (if it is in the form y = mx + c where x = ’daily relative change in the gold price’ and y is ’daily ETF return’)

1. Conduct a two-tailed t-test with *H*0 : 1 = 0. What is the P-value of the test? Is the linear relationship between ETF (Y) and Gold (X) significant at the significance level 0.01? Why or why not?

About two-tailed t-test

Hypothesis testing is an essential procedure in statistics. A hypothesis test evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data.

Two sampled T-test: The Independent Samples t Test or 2-sample ttest compares the means of two independent groups in order to determine whether there is statistical evidence that the associated population means are significantly di↵erent. The Independent Samples t Test is a parametric test. This test is also known as: Independent t Test.

The model behind linear regression:

When we are examining the relationship between a quantitative outcome and a single quantitative explanatory variable, simple linear regression is the most commonly considered analysis method.

We postulate a linear relationship between the population mean of the outcome and the value of the explanatory variable. If we let Y be some outcome, and x be some explanatory variable, then we can express the structural model using the equation

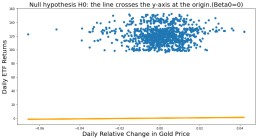
*E*(*Y* |*x*) = 0 + 1*x or y*ˆ = ˆ0 + ˆ1*x*

we are ultimately always interested in drawing conclusions about the population not the particular sample we observed. In the simple regression setting, we are often interested in learning about the population intercept 0 and the population slope 1 . As we know, confidence intervals and hypothesis tests are two related, but di↵erent, ways of learning about the values of population parameters.

Steps to conduct a two-tailed t-test with *H*0 : 1 = 0:

An *↵ level* hypothesis test for slope parameter 1. We follow standard hypothesis test procedures in conducting a hypothesis test for the slope 1.

1. First, Specify the null and alternative hypotheses: Null hypothesis*H*0 : 1 = 0, Alternative hypothesis *HA* : 1 = 06 . If we put 1 = 0 we get following plot



1. Third, use the resulting test statistic (t score) to calculate the Pvalue. As always, the P-value is the answer to the question ”how likely is it that we’d get a test statistic t\* as extreme as we did if the null hypothesis were true?” The P-value is determined by referring to a t-distribution with n-2 degrees of freedom.
2. Finally, we make a decision:
   * If the P-value is smaller than the significance level *↵*, we reject the null hypothesis in favor of the alternative. We conclude ”there is su cient evidence at the *↵* level to conclude that there is a relationship in the population between the predictor x and response y.”
   * If the P-value is larger than the significance level *↵*, we fail to reject the null hypothesis. We conclude ”there is not enough evidence at the *↵* level to conclude that there is a relationship in the population between the predictor x and response y.”Now using above mentioned steps let’s verify the Null hypothesis. For the Linear model with input variable is X = ”Daily Relative Change in Gold Price” along with constant and Y = ”Daily ETF Returns” we got the following model report in Python:

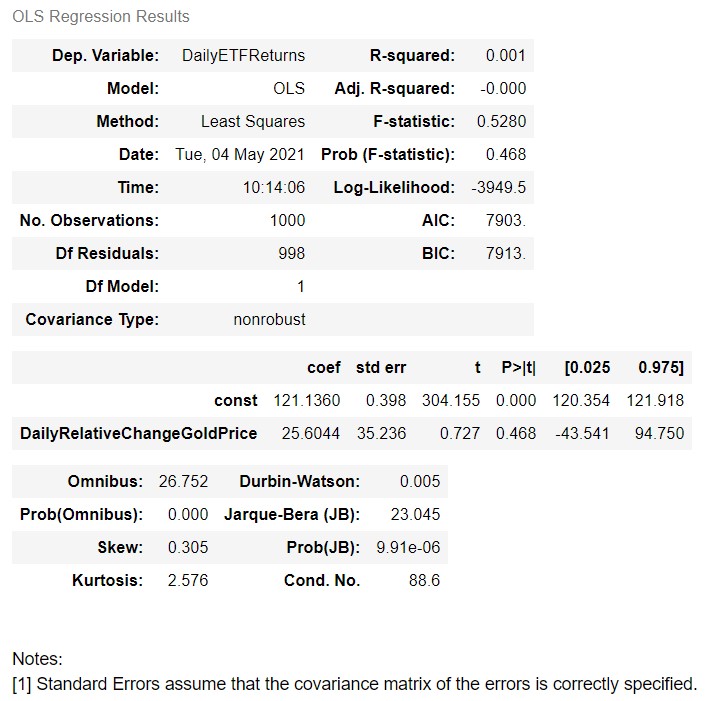
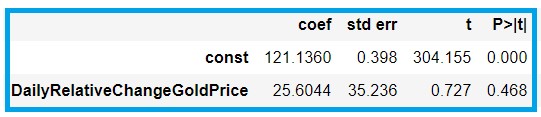


Figure 2: Model report

In which the Coe ficients Table is as follows:



We can see the coe cient of the intercept, or the constant as they’ve named it in our case. Hence we have

Where 0 is y-intercept and 1 is the slope in the regression line.

By default, the test statistic is calculated assuming the user wants to test that the slope is 0. Dividing the estimated coe cient for Daily Relative Change in Gold Price = 25.6044 by the estimated standard error for the Daily Relative Change in Gold Price = 35.236, that model reports printed the test statistic T is

Which is matching with the given t value for this variable in the report too.

Calculate the critical region: Consider the = 0*.*01 and from the report

table we got the

*Degree of freedom* = 898

So using the following table let’s derive the critical value:From the standard t distribution table - With significance Level = 1% ( two sided ) we have: Critical value as: 2.58.

Drawing conclusions about slope parameter 1 = 0 using T-statistics and critical region values:

* Since Critical region: (1*,* 2*.*581][[2*.*581*,*1) and t-score as 0.727 therefore the t-score does not belongs to the critical region, so we accept *H*0 and reject *Ha* .
* That is the we accept: H0: ’Daily Relative Change in Gold Price’ does not a↵ect ’Daily ETF Returns’ and reject that H1: that is ’Daily Relative Change in Gold Price’ does a↵ect ’Daily ETF Returns’.

p-value significance level *↵*:

By default, the P-value is calculated assuming the alternative hypothesis is a ”two-tailed, not-equal-to” hypothesis. From the report we have

.

*degrees of freedom*(*n* 2) = 898

*P* = 0*.*468 ( *to three decimal places*)

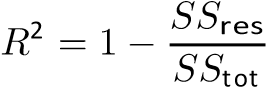
Upon calculating the probability that a t-random variable with n-2 degrees of freedom would be larger than —t=0.727—, and multiplying the probability by 2, we have

2 ⇥ *p* = 2 ⇥ 0*.*468 = 0*.*936 *> ↵* = 0*.*01

Drawing conclusions about slope parameter 1 = 0 using T-statistics and critical region values:

* Because the P-value is so big (greater than *↵*), we can accept the null hypothesis *H*0 and conclude that 1 = 0.
* Therefore, there is suficient evidence, at the *↵* = 0*.*01 level, to conclude that there is a no relationship in the population between ”Daily Relative Change in Gold Price” and ”Daily ETF Returns”.

1. Suppose that you use the coeficient of determination to assess the quality of this fitting. Is it a good model? Why or why not?
   * The COD is the proportion of the variance in the dependent variable that is predictable from the independent variable(s)



* + It gives you an idea of how many data points fall within the results of the line formed by the regression equation. The higher the coe cient, the higher percentage of points the line passes through when the data points and line are plotted. For example if the coe cient is 0.80, then 80% of the points should fall within the regression line. • Values of 1 or 0 would indicate the regression line represents all or none of the data, respectively. A higher coe cient is an indicator of a better goodness of fit for the observations. The best possible score is 1 which is obtained when the predicted values are the same as the actual values. *R*2 score of baseline model is 0.
  + The COD can be negative, although this usually means that your model is a poor fit for your data. It can also become negative if you didn’t set an intercept. During the worse cases, *R*2 score are negative.

*SStot* = X (*yi y*¯)2

*SSres* = X (*yi f*¯*i* )2 = X *e*2*i*

Although we can consider a value for the comparison let’s say 0.5 and can say if the *R*2 is greater that this then it is best or good model. But generally it is considered that if *R*2 value is near to 1 then it is good model.

For the linear model we got

*R*2 = 0*.*0005287962431211879

Therefore we can say that since the *R*2 = 0*.*0005 value is very small compared to 0.5 hence model is bad.

1. What are the assumptions you made for this model fitting?

Four assumptions are being maid:

(a) Assumption of Linearity check:

With assumption linearity or to seen the outcome variable is linearly related to any of the predictors ( by straight line ) since the predictors combined e↵ect (in order to sum up their linear relation by straight line) will help to make prediction of dependent variable.

Diagnosing linearity using visual inspection method can be di cult since there is no basis (or well defined standards ) to determine what is considered linear and and what is not.

* + - Method 1: Using co-relation matrix

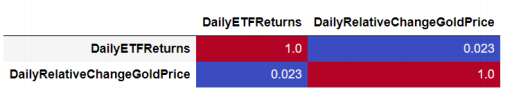
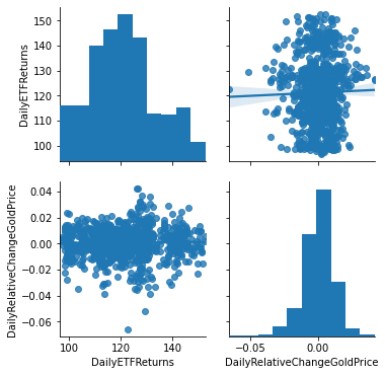


Figure 3: correlation of coe cient matrix for ETF and Gold

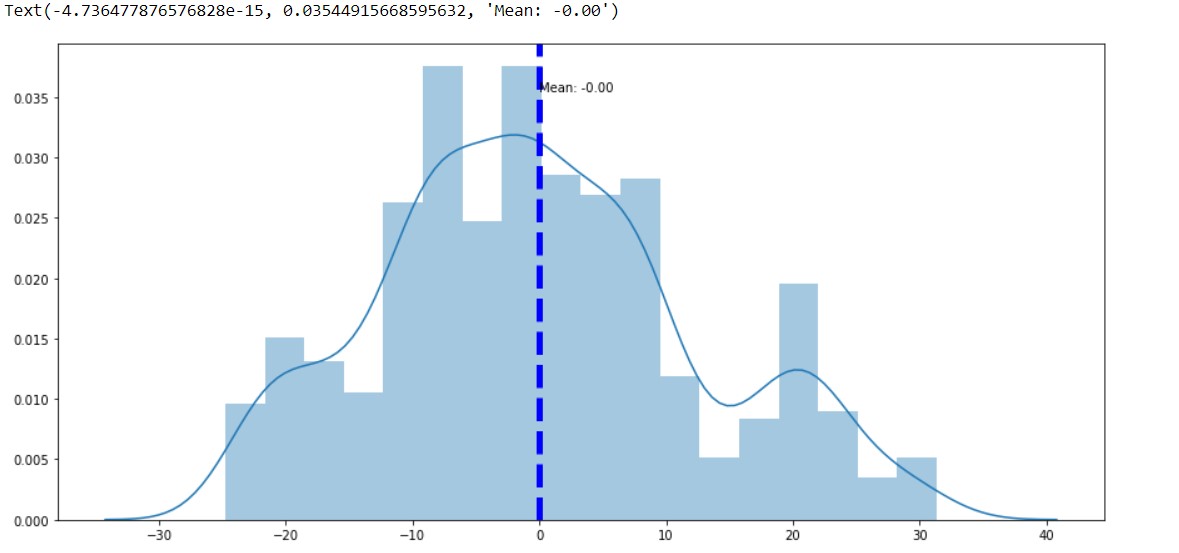
Dependent variables and the independent variable are having weak liner co-relation. But the the co-relation is positive ¿ 0. So we can say that the independent variable shows positive weak co-relation with dependent one.

* + - Method 2: Using Matrix scatter plots and prediction line The first assumption we check is linearity. We can visually check this by fitting ordinary least squares (OLS) on a data.



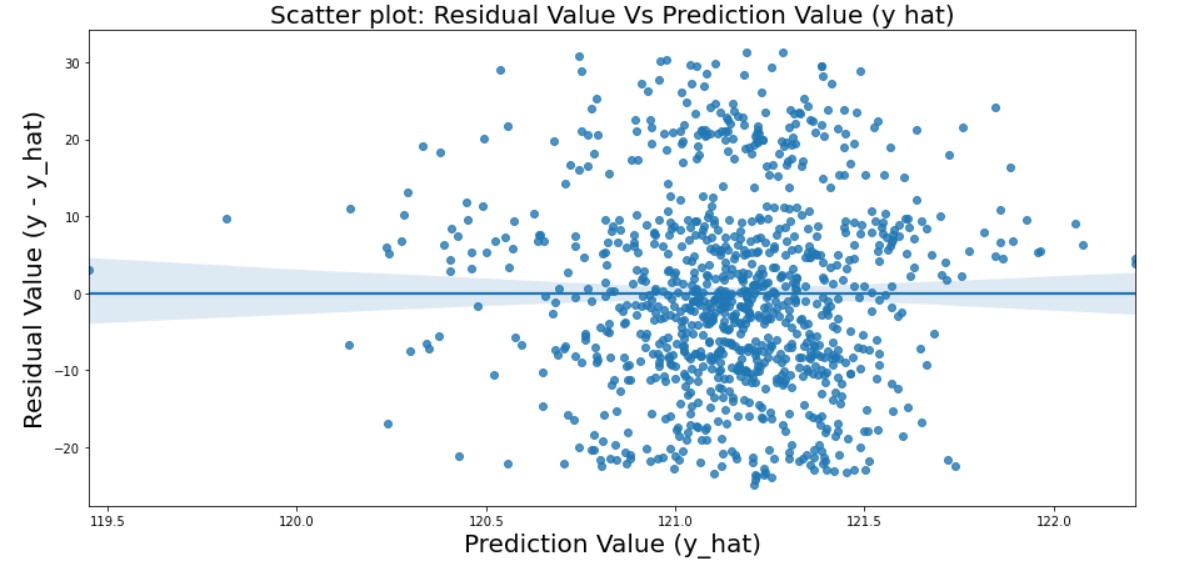
Observations/Linearity Diagnosis: Check scatter plot between Dependent variable and the other independent variables. With assumption of linearity we are looking to see if the outcome variable is linearly related to each of the independent variable

We draw a histogram of the residuals, and then examine the normality of the residuals. If the residuals are not skewed, that means that the assumption is satisfied. It can be checked the predictor not so linearly related to dependent variable that is ETF. It is also cleared from the co-relation table too.



Even though is slightly skewed, but it is not hugely deviated from being a normal distribution. We can say that this distribution satisfies the normality assumption.

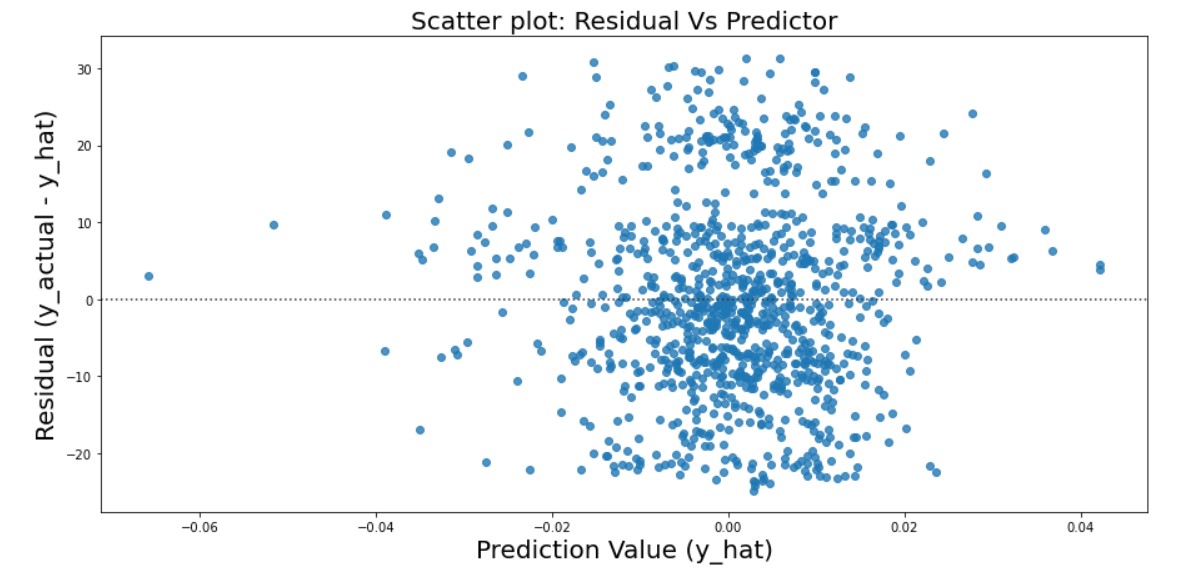
1. Independent Errors assumption



The assumption of IE is that successive residuals should be independent that means there is no patterns to the residuals. The residuals aren’t highly corelated there is no long runs of positive or negative residuals

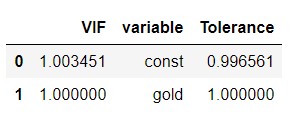
When the successive residuals are correlated we refer to this condition as ’auto-corelation’ and that frequently occurs when the data are collected over a period of time. In our case There is no specific pattern that is it does not include or indicate autocorrelation. Hence Independent Errors assumption is not violated from the graph

1. Assumption of Homoscedasticity:



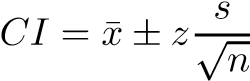
From the previous explanation as we know that the dots are scattered the way that they are it indicates the data meet the assumption of the errors being normally distributed also because the dots are scattered it indicates that the variances of the residuals are constant (d) Multi Co-llinearity:

This occurs when two or more variables very closely linearly related as co-linearity increases so do the standard errors of the be codified coe cient the beta coe cient so if you think back to what the standard error represents the big standard errors for the beta coe cients mean that the beta coe cients vary more across the samples in essence our betas become less trustworthy it also makes it di cult to assess the individual importance of a predictor if the predictors are highly co-related and each accounts for similar variants in the outcome then how do we know how can we tell which of the two variables is important we really cant the model could include either interchangeably one way we can check for molecularity is to scan a correlation matrix for high coe cients those that are above say 0.8 or 0.9 in another words what we are looking at the correlation matrix to check for multi col-linearity we want to have low correlation coe cients between each of the independent variables and each of the other independent variable you do not really need to include the dependent variable in the correlation matrix to check this assumption though.



Hence it can be seen that for the gold the assumption is not violated if compared with constant.

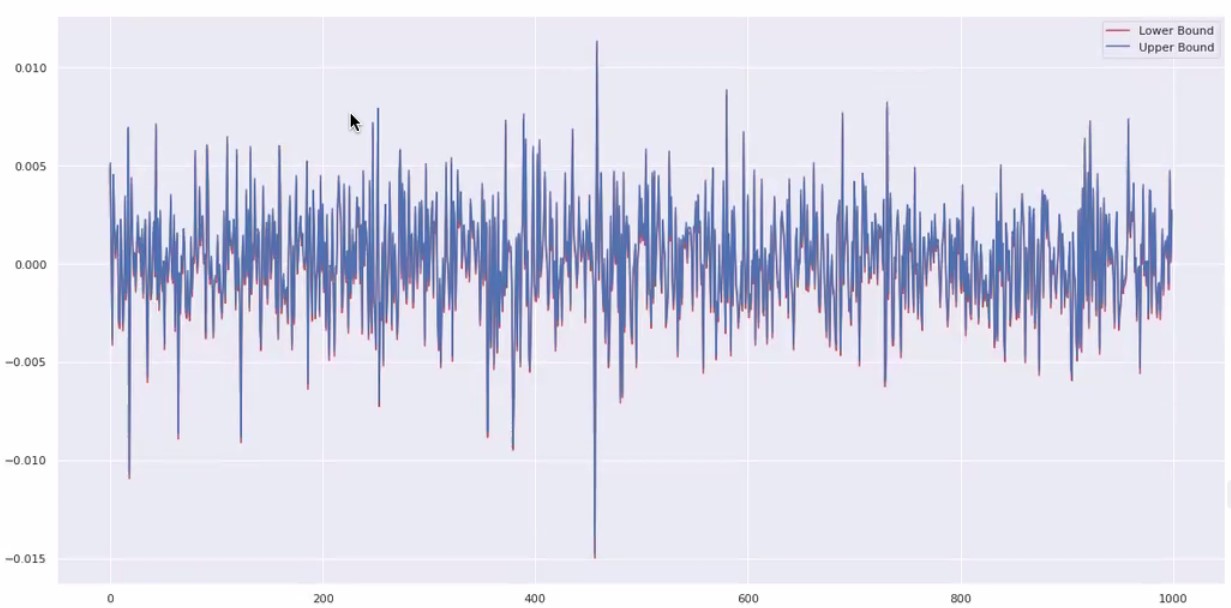
7. Given the daily relative change in the gold price is 0.005127. Calculate the 99% confidence interval of the mean daily ETF return, and the 99% prediction interval of the individual daily ETF return.



Where CI: confidence interval, ¯*x* is sample mean, z is confidence level value, s is sample standard deviation, n sample size.

Confidence interval for *CloseE TF* is (118*.*57220037473246*,*123*.*73371964926773)

For alpha: 0.99



Also the 99% prediction interval of the individual daily ETF return Interpretation:

There is 99% chance that the confidence interval of (array([118.57220037, -2.5800968 ]), array([123.73371965, 2.58142247])) contains the true mean.

**Part 9: Does your model predict?**

Consider the data including the ETF, Gold and Oil column. Using any software, fit a multiple linear regression model to the data with the ETF variable as the response. Evaluate your model with adjusted R square.Multiple regression is like linear regression, but with more than one independent value, meaning that we try to predict a value based on two or more variables. Before fiting a model for the data let’s check the how the variables are co-related:

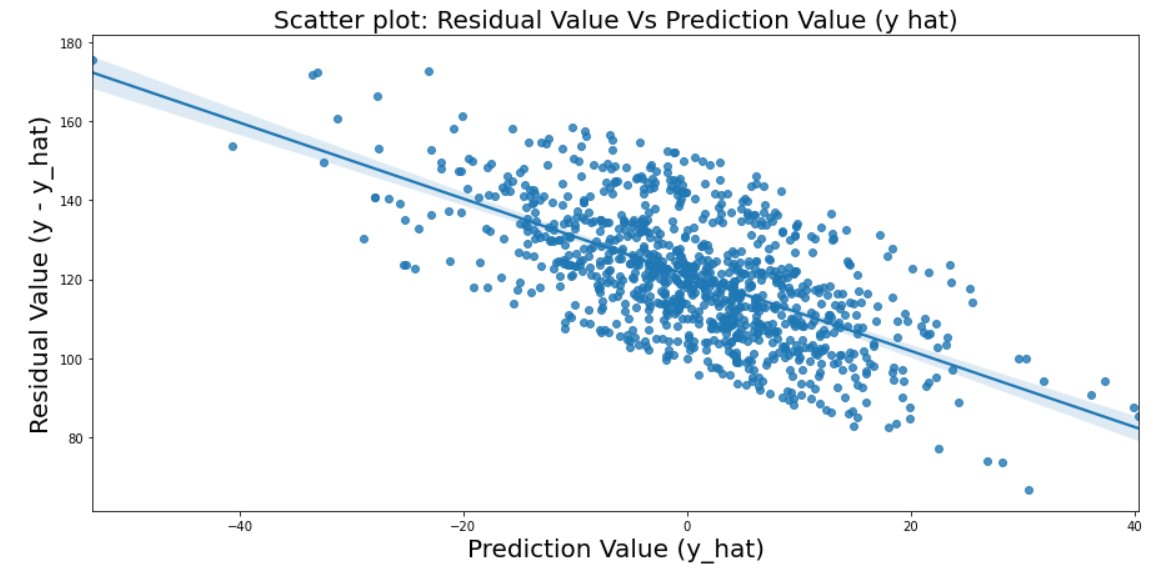


As it can be seen that the variables are weakly co-related to each other as follows:

* The magnitude shows that it is ¡ 0.5 hence are weakly correlated
* The Oil is negatively co-related with the ETF and on the other hand gold is positively.

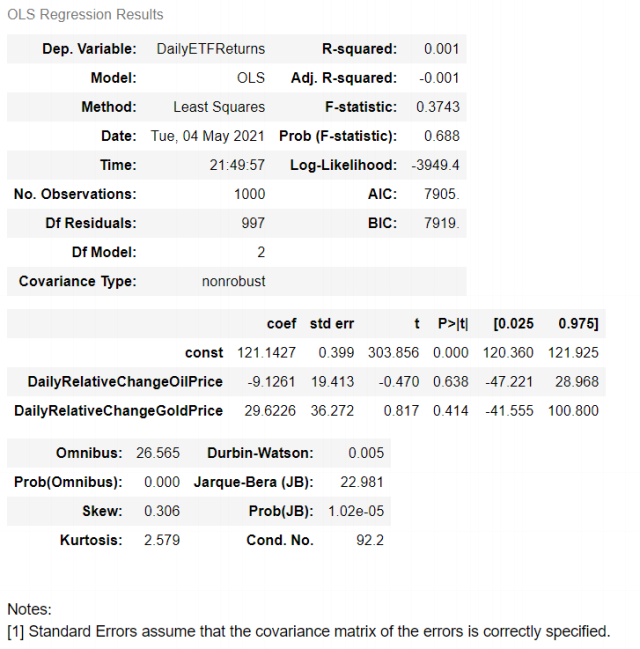
If we draw a graph for the Predict Vs Residual To Check Linearity we get

following plot



From the graph it is hard to tell anything about linearity assumption since there are more than one independent variables hence using 2d the Predict Vs Residual will not help.

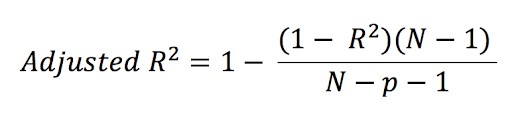
After model is fit we got the summary of the model as follows:



From the report the Value of the *R*2 is as follows:

*R*2 = 0*.*001

**Evaluate your model with adjusted** *R*2 If we recall in Part 8 the we got the *R*2 as 0.000 and after adding oil as the new feature for model training, we get the new value of adjusted r- square as -0.001. Now, in order to interpret this change, let’s take a look at the formula for adjusted r-square which is given by:-



where, the p term in the denominator acts as a penalty term which penalizes the model performance score(adjusted r-square) if the newly introduced feature is not helping in improving the model performance because it is located in the denominator of the equation. If however, the model performance increases due to the newly introduced variable, the value of r-square(located in the denominator) will compensate for the penalization and would improve the overall score.

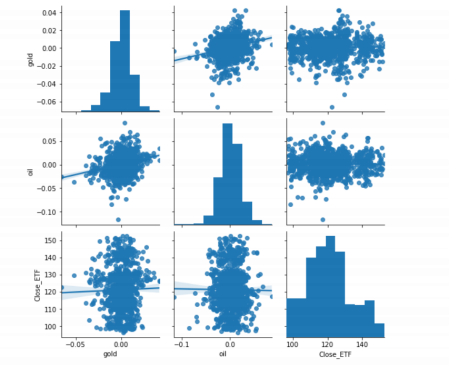
It shows that the newly added feature(oil) is not adding any value to the model performance.

**Part 10: Checking residuals and model selection**

Calculate the residuals of the model fitting you did in Part 9.

Check the four assumptions made for the error terms of the multiple regression model using these residuals (mean 0; constant variance; normality; and the independence). Plot the residuals to check these assumptions. For example, draw a Normal Probability Plot to check the normality assumption; draw a scatter plot of Residuals vs. Fitted Values to check the constant variance assumption and the independence assumption; and so on. You may refer to the following link https://www.youtube.com/watch?v=4zQkJw73U6I for some hints. In your project report, all the relevant plots and at least one paragraph of summary of checking the four assumptions using those plots must be included.

1. Assumption of Linearity check:

Chart, scatter chart

Description automatically generated

Table

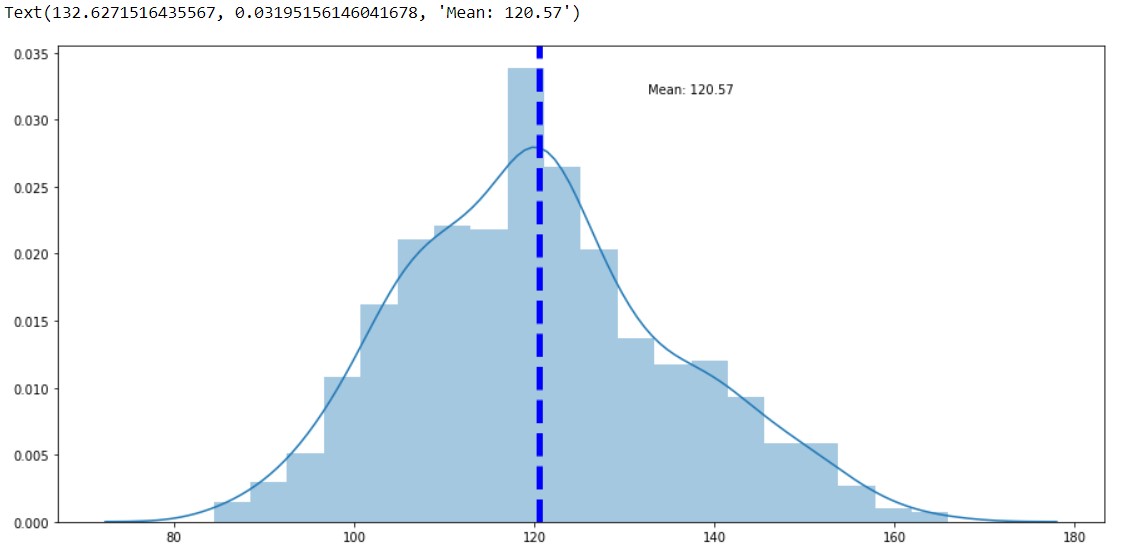
Description automatically generated

From the above two plots and the residual table we can see there is some linear relationship exists between dependent and independent variables although it is not much in terms of magnitude of co-relation but we can say there are weak linearly relationships. Oil and ETF are having weak negative linear co-relation.

Also if we check the VIF values for the gold and oil it is:

*oil* : 1*.*059952*, gold* : 1*.*059952

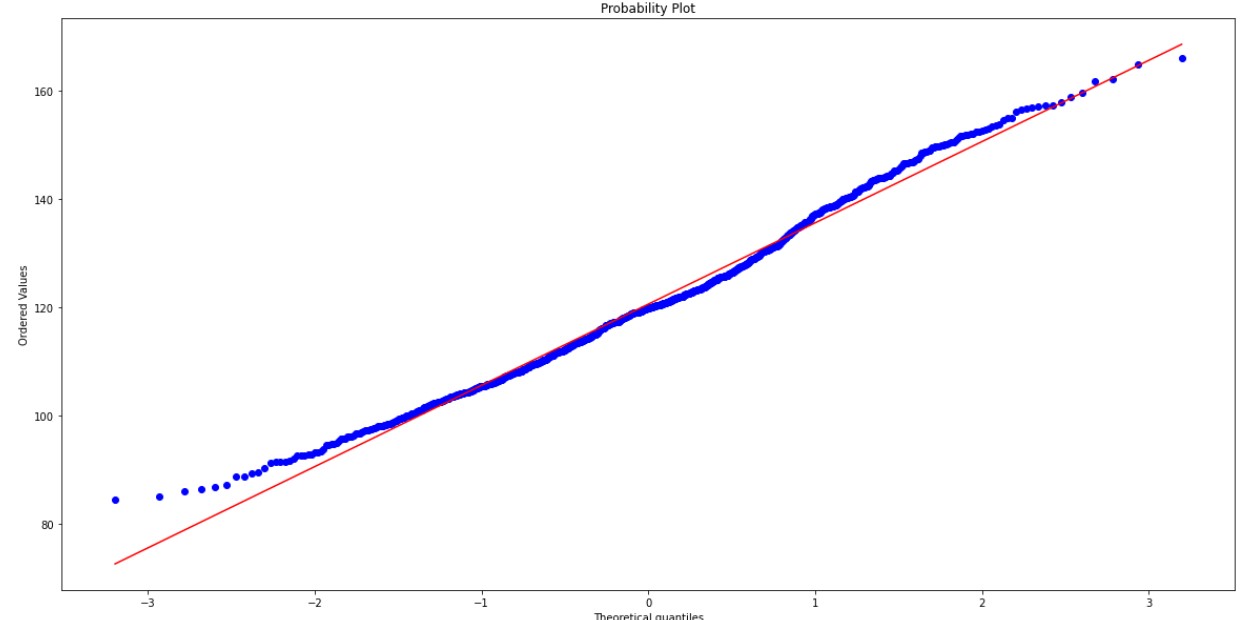
Following is the histogram of residuals:



If the residuals are not skewed, that means that the assumption is satisfied. From the above plot it can be seen that the histogram is not skewed one and the data is distributed across the mean so that the Linearity assumption is not violated.

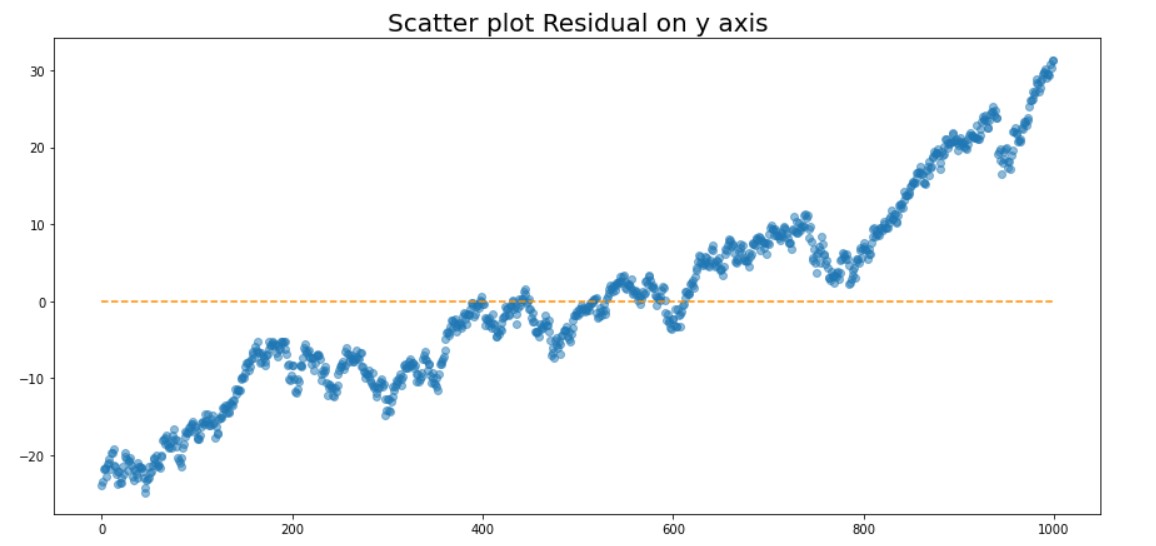
2. Independent Errors assumption:

From the above QQ plot it can be see that the plot is having fat tails, and the data is not Gaussian at the tails. So it seems it is violating the assumption of independent errors.



We thus have to be careful if we implement a t-test to test whether elements of are non-zero, as our test statistic requires that is Gaussian. Gaussian errors would imply a Gaussian , but since we don’t have Gaussian residuals, we don’t have Gaussian errors, and we need to see whether the central limit theorem can tell us that is approximately Gaussian. We must investigate this carefully before concluding so. There are three common ways to deal with normality violations other than invoking the central limit theorem: 1) transform either your features or your response 2) fit a GLM instead of doing linear regression 3) bootstrapping.

Assumption of homoscedasticity:



From the above Scatter plot of Residuals in case of multi-linear model we can see it is not distributed across the horizontal line passing through 0 and it is following some patter indicating Assumption of Homoscedasticity is violated.

Assumption of Multi-Collinearity:

Text

Description automatically generated

**3 Section 3 - Discussions or Improvements**

1. Various more data pre-processing techniques can be used for more better insights of data.

2. More models can be tried out for better performance

3. We can try including JPM column along with other columns for the multi-regression model.

**4 References**

1. Mathematical Statistics and Data Analysis (Third Edition) – John A. Rice.
2. An Introduction to Statistical Learning – Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani.
3. Statistics -OpenStack
4. <https://towardsdatascience.com/normality-tests-in-python-31e04aa4f411>
5. <https://medium.com/analytics-vidhya/central-limit-theorem-and-machine-learning-part-1-af3b65dc9d32>
6. https://github.com/sharmasw/Data-Science-with-python/blob/master/
7. <https://sixsigmastudyguide.com/1-sample-wilcoxon-non-parametric-hypothesis-test/>
8. <https://www.youtube.com/watch?v=0Pd3dc1GcHc>
9. <https://www.youtube.com/watch?v=8aaIdXENNJI>
10. [*https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/f-test/*](https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/f-test/)
11. [*https://online.stat.psu.edu/stat501/node/644*](https://online.stat.psu.edu/stat501/node/644)
12. [*https://jeffmacaluso.github.io/post/LinearRegressionAssumptions/*](https://jeffmacaluso.github.io/post/LinearRegressionAssumptions/)
13. <https://365datascience.com/tutorials/statistics-tutorials/null-hypothesis/>