FULL LEGAL NAME	LOCATION (COUNTRY)	EMAIL ADDRESS	MARK X FOR ANY NON-CONTRIBUTING MEMBER
Marin Yordanov Stoyanov	Bulgaria	azonealerts@gmx.com	
Jefferson Bien-Aimé	USA	jefferson.bienaime@gmail.com	
David Chege Simbarashe Kagunda			X

Statement of integrity: By typing the names of all group members in the text boxes below, you confirm that the assignment submitted is original work produced by the group (excluding any non-contributing members identified with an "X" above).

Team member 1	Marin Yordanov Stoyanov
Team member 2	Jefferson Bien-Aimé
Team member 3	

Use the box below to explain any attempts to reach out to a non-contributing member. Type (N/A) if all members contributed.
Note: You may be required to provide proof of your outreach to non-contributing members upon request.

At the very beginning of group work project 1, David Chege Simbarashe Kagunda, said that he is withdrawing from the course and this is why he did not participate in the common work.

Skewness Definition.

Skewness is a statistical term that refers to the degree of asymmetry in the distribution. It is all about the mean of this distribution. If most values are located on the left side of the distribution, then the skewness is positive (right tail). Otherwise, it is negative (left tail), but if the distribution is symmetrical then the skewness is zero. There is one special case when the skewness can be zero in both cases of distributions: symmetrical and asymmetrical. This is when the one tail is long and thin while the other is short and fat and, in this case, they cancel each other out. [1]

The formula for skewness is like this:

$$Skewness = \Sigma_{i=1}^{n} \frac{\frac{(x_i - \bar{x})^3}{n}}{SD_X^3}$$

Where:

- n is the number of data points,
- x_i is each value from the data set,
- $\bar{x}$ is the mean of the data,
- SD is the standard deviation of the data. [2]

Demonstration:

For demonstration I am using the bitcoin closing price for the period: "2020-01-01" till "2023-04-25". After plotting the histogram (fig. 1. is in the notebook) it is visible that the distribution is skewed.

Diagnosis:

There is a quick and easy visual way to find if skewness is present and it is to compare the mean, median and mode. If they all are equal, this will mean that there is no skewness (skewness =

zero). In case they are not, then if mean > median > mode, then it is positive skew. If mean< median < mode then we have a negative skew. [3]

Another way to see what the skewness is, is to use statistical methods like the one in Python called: .skew() [4] and it will automatically calculate and show the value of the skewness.

Damage:

If there is skewness in the data distribution, then the assumptions that the distribution is normal cannot be applied and used. In this case a different approach or at least some other kind of data preprocessing and transformation should be used in order to reduce skewness before analysis.

This is why I will check the skewness of bitcoins returns and later I will apply logarithm to see if the skewness will reduce.

In Figure 2. A histogram of Bitcoin's returns for the period: "2020-01-01" till "2023-04-25", it is still visible that the data is skewed. The value for Skewness is: -0.6626080861038385, so the returns are negatively skewed.

Lets try applying algorithm transformation and see what will happen to the data.

Yet again in Figure 3. A histogram of Bitcoin's logged returns for the period: "2020-01-01" till "2023-04-25 is visible that even that we have a close to the normal distribution, yet the logged bitcoin return are skewed because the parameter for skewness shows a negative skew with the following value:

Skewness: -1.1479247984414598

Kurtosis/Heteroskedasticity

Definition

Kurtosis is a statistical measure that describes the distribution of observed data around the mean. It specifically measures the tails and sharpness of the distribution of data values. The formula to compute sample kurtosis is:

$$Kurtosis = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \Sigma_{i=1}^{n} (\frac{(x_i - \bar{x})^4}{SD^4} - \frac{3(n-1)^2}{(n-2)(n-3)})$$

Where:

- n is the number of data points,
- x_i is each value from the data set,
- x̄ is the mean of the data,
- SD is the standard deviation of the data.

Heteroscedasticity refers to the circumstance in which the variability of a variable is unequal across the range of values of a second variable that predicts it.

Description:

Kurtosis measures the "tailedness" of the distribution of data: high kurtosis means heavy tails and a sharp peak (leptokurtic); low kurtosis indicates light tails and a flat peak (platykurtic); zero kurtosis indicates normal distribution Heteroscedasticity, occurs when the variance or standard deviation of the error term varies over different levels of the independent variables. The spread of the predicted values differ at different values of the independent variables.

Demonstration:

To demonstrate both Kurtosis and Heteroscedasticity, I am using the returns of the bitcoin for the past 5 years. That demonstration will be discussed in the python notebook as well as the diagram. As you may see in the diagram. Kurtosis is really high Kurtosis of the returns 8.94377055909077 indicating the data has kurtosis.

Diagnostic

Kurtosis can be seen by using the formal above or by using. Galton skewness (also known as Bowley's skewness Heteroscedasticity can be diagnosed by statistical tests like the Breusch-Pagan where the p-value is less than 0.05.

Damage:

High kurtosis can result in a higher error rate and may imply potential outliers which can affect the results of statistical analysis. For Heteroscedasticity: It can lead to inefficient parameter estimates and it can make the model's standard errors, t-statistics, and F-statistics unreliable.

Sensitivity to outliers

In statistics, an outlier is an observation that is significantly different from other values in a sample. According to Hawkins (1980), an outlier is an observation that deviates so much from other observations that it raises suspicions of being generated by a different mechanism. One way to identify outliers is by using the Interquartile Range (IQR) method. With this method, a data point is considered an outlier if it falls below Q1 - 1.5*IQR or above Q3 + 1.5*IQR, where Q1 is the first quartile (25th percentile), Q3 is the third quartile (75th percentile), and IQR is the interquartile range (Q3 - Q1) [14]. These data points are considered too far from the central values to be reasonable.

Description:

Outliers are extreme values that deviate significantly from other observations in the data. They might represent a significant finding in the data, an error in measurement, or variability in the data.

Demonstration:

For demonstration we will use the bitcoin closing price for the period and the diagram below gives more explanation

Diagnosis

To diagnose outliers, a number of statistical techniques can be employed. Graphical methods like scatter plots and box plots, statistical methods like the z-score method or the IQR method can be used.

Damage:

Outliers can skew statistical measures and data distributions, leading to misleading results. For example, they can significantly affect the mean of the data and distort the standard deviation and variance.

Directions

Skewness

Real-world data sets are often skewed, meaning they have an asymmetrical curve on a graph with a "tail" on either the left or right side. This can be either a negative skew, where the tail is on the negative side of the graph, or a positive skew, where the tail is on the positive side of the graph. In contrast, a normal distribution has a skew value of zero and is symmetrical, with nearly symmetrical data also having a skew value near zero. However, skewed data can present challenges for statistical models, as outliers can cause skew and negatively impact the model's performance.

1. First of all it is always good to do initial exploratory data analysis by plotting the data. Scatterplots are usually used to see if there are some linear relations between the data. Also a histogram can be plotted so that we could see the distribution of data. QQ plots

can be applied to see if the data has a normal distribution. Boxplots can be applied to see how the outliers (if any) are located compared to the mean. [10]

2. For further and more precise analysis there should be applied some additional statistical tests for finding what is the data distribution like Kolmogorov-Smirnov test [11] or Shapiro-Wilk test for normality [12]. Also there is a very useful and easy to be applied Python method that shows you exactly what the skewness value is and this is the .skew() method.
3. When the data is close to a normal distribution, then specific transformations can be made so that the distribution will be closer to a normal one. These are the exponential transformation, the power transformation, the box-cox transformation. The most used transformation (in my opinion) is the logarithm transformation (the log) transforms the data very closely to a normal distribution. [13]

Kurtosis/Heteroskedasticity

Usually Kurtosis is a suggestion about your data (i.e., potential outliers or lack thereof) and taking appropriate action. We can apply either log transformation, square root transformation or inverse transformation. The goal would be to achieve a more normal distribution. To do this, you will use python code in the Jupyter notebook to illustrate this.

1. Load the data
2. Calculate the initial kurtosis
3. Transformation of the data
4. Calculate the kurtosis of the data
5. Compare the result
6. The appropriate result would be to choose the transformation that brings the kurtosis closest to 3.

Heteroscedasticity: There could be multiple ways for handling heteroskedasticity One can use robust standard errors, transform the dependent variable (e.g., log transformation), or use other modeling approaches like weighted least squares regression that take heteroscedasticity into account.

We can use the following steps to solve heteroskedasticity :

1. We will need to perform tests such as Breusch-Pagan or White Test to confirm the presence of heteroskedasticity. We will have the **Null Hypothesis (H_0) to indicate that** Homoscedasticity is present, We will also have **Alternative Hypothesis where** Heteroscedasticity is present. If the p-value is less than 0.05, then we reject the null hypothesis and conclude that heteroscedasticity is present.
2. You will need also do transformation. For example we can use Log Transforamtion. We can use other transformation like square root, or Box-Cox transformation to stabilize the variance.
3. The next step would be to use the weighted least square or Generalized Least Squares (GLS) instead of OLS. It incorporates weights to account for heteroskedasticity and take into account the variance of the observations.

When conducting a statistical test for heteroscedasticity in a regression model, we compare the p-value of the test to a pre-determined significance level (such as $\alpha = .05$). If the p-value is less than this significance level, we reject the null hypothesis and conclude that there is evidence of heteroscedasticity in the model.

Sensitivity to outliers

Outliers can be managed through various ways. They can be excluded, transformed, or replaced through imputation. The choice depends on the nature of the data and the purpose of the analysis. One thing, you should be careful about is the fact that outliers are real data points and removing them may biased the analysis. What should we do to address the issue is the following :

1. You draw the boxplot of the data with the outlier
2. As mentionned in the definition, we can use the lower and upper quantiles as well as the IQR to determine the lower bound and the the upper bound. With the both of them, you can determine the outliers and remove them from data to have a clean data.
3. Next you will draw a boxplot that shows the data without the outliers.
4. Another method to use is to use the Zscore to dete

REFERENCES:

1. Illowsky, Barbara; Dean, Susan; "2.6 Skewness and the Mean, Median, and Mode - Statistics"; 27 March 2020
2. Stan Brown; "Measures of Shape: Skewness and Kurtosis"; Oak Road Systems; 2016
3. A.W.L. Pubudu Thilan; "Applied Statistics I: Chapter 5: Measures of skewness"
4. Python documentation; scipy.stats.skew; https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.skew.html
5. Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. Econometrica: Journal of the Econometric Society, 1287-1294.
6. Balanda, K. P., & MacGillivray, H. L. (1988). Kurtosis: a critical review. The American Statistician, 42(2), 111-119.
7. Hawkins, D. M. (1980). Identification of outliers. Chapman and Hall.
8. DeCarlo, L. T. (1997). On the Meaning and Use of Kurtosis. Psychological Methods, 2(3), 292–307.
9. Aggarwal, C. C. (2016). Outlier Analysis (2nd ed.). Springer.
10. Jim Frost, How to Identify the Distribution of Your Data, How to Identify the Distribution of Your Data - Statistics By Jim
11. Stephens, M. A. (1974). "EDF Statistics for Goodness of Fit and Some Comparisons". Journal of the American Statistical Association.
12. Shapiro, S. S.; Wilk, M. B. (1965). "An analysis of variance test for normality (complete samples)". Biometrika
13. Dario Radečić, Top 3 Methods for Handling Skewed Data, 04.01.2020
14. Igor Crha, Jana Zakova, Martin Huser, Pavel Ventruba, Eva Lousova & Michal Pohanka, Digital holographic microscopy in human sperm imaging