CCT College Dublin

Assessment Cover Page

Module Title:	Statistical Techniques for Data Analysis
Assessment Title:	Integrated CA
Lecturer Name:	Aldana Louzan
Student Full Name:	Beatriz Lobão Silva Sena
Student Number:	2022090
Assessment Due Date:	27/05/2022
Date of Submission:	27/05/2022

Declaration

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Integrated CA

Beatriz Lobão Silva Sena [1]

Aldana Louzan [2]

ABSTRACT- The following study presents based on hypothesis tests, correlation between variables, causation and liner regression. In this study, data from FIFA21 were used with the objective of analysing the data through python software.

Keywords: hypothesis tests, correlation, causation, liner regression, p-value.

Dublin

2022

[1] Student of Data Analytics at CCT college Dublin. E-mail: 2022090@student.cct.ie

[2] Professor at CCT college Dublin. E-mail: alouzan@cct.ie

INTRODUCTION

The statistics studies must know what you are describing, and the descriptive statistics are about variable. Hypothesis testing is all about the validity of making claims from a sample. Most of the time, when we want to know information about a population, we can’t get every single piece of data. In hypothesis testing, an analyst tests a statistical sample, with the goal of providing evidence on the plausibility of the null hypothesis.

Correlation a relationship between two variables. If the values of one variable depend strongly, moderate or weakly of another one, then you would say how connect the variables are, and also if the correlation is direct or conversely. Causation, the relationship between cause and effect between two variables.

A heat map is a data visualization technique that shows the correlation between two variables. The variation in color may be by hue or intensity, giving obvious visual cues to identify how strong is the relation between two variables. We can also plot the, pearson coefficient for a better analyse.

Linear regression in statistics create a linear relationship between correlated variables. The model is used to predict the value of a variable based on the value of another variable.

This CA aims to conduct research and improve knowledge in statistics. For this, it was necessary to carry out tests in python using the dataset that contains information about the players of the video game FIFA 21.

DEVELOPMENT

1. Choose one variable from your chosen dataset and perform a Hypothesis Test.

All the steps must be supported by appropriate references, statistical concepts, and calculations.

You must interpret your results, provide your own analysis and conclusion based on your Hypothesis Test.

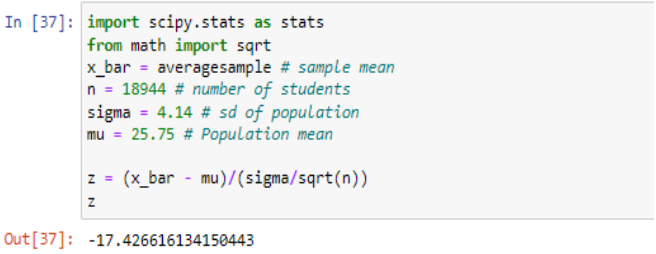
You will need to conduct research to find the parameters of the population you want to analyse.

According to National Center for Biotechnology Information the age of the soccer players ranges from 16 to 43, with an average of 25.75 ± 4.14 years.

Based on the National Center for Biotechnology it is believed that the average age of soccer players is 25.75 years old. However, the FIFA football game has a sample of 18944 soccer players from of most real football championships. The average age of players who are registered in the game is 25.2258. Because of the size of the sample, we are going to apply z test.

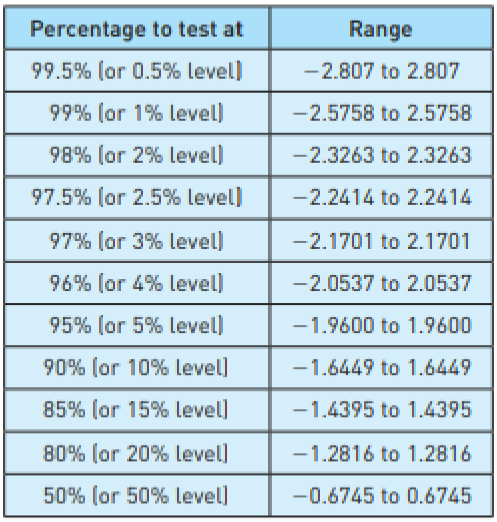
- H0: The sample is from the football players population, x_fifa_age = µ_age.
- HA: The sample is not from the football players population, x_fifa_age != µ_age.

What we want to do is REJECT our null hypothesis, instead of trying to prove our alternative hypothesis. That's the way significance testing works. We are going to apply a two-tailed test because we don't care if the sample mean is greater than or less than the population mean. Then, we specify a significance (alpha) level. Usually, statistical significance is associated with an alpha level of α = 0.05 or smaller. Next, we use a z table to look up the critical z value that corresponds to this α level.



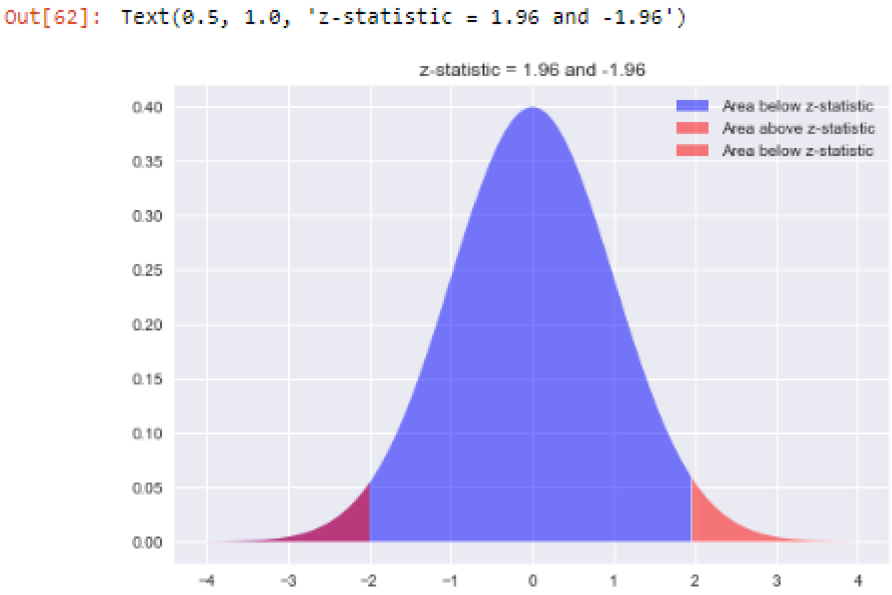
```
In [37]: import scipy.stats as stats
         from math import sqrt
         x_bar = averagesample # sample mean
         n = 18944 # number of students
         sigma = 4.14 # sd of population
         mu = 25.75 # Population mean

         z = (x_bar - mu)/(sigma/sqrt(n))
         z
Out[37]: -17.426616134150443
```

Figure 1: Z stats



Percentage to test at	Range
99.5% (or 0.5% level)	−2.807 to 2.807
99% (or 1% level)	−2.5758 to 2.5758
98% (or 2% level)	−2.3263 to 2.3263
97.5% (or 2.5% level)	−2.2414 to 2.2414
97% (or 3% level)	−2.1701 to 2.1701
96% (or 4% level)	−2.0537 to 2.0537
95% (or 5% level)	−1.9600 to 1.9600
90% (or 10% level)	−1.6449 to 1.6449
85% (or 15% level)	−1.4395 to 1.4395
80% (or 20% level)	−1.2816 to 1.2816
50% (or 50% level)	−0.6745 to 0.6745

Figure 2: Two Tailed Z value



```
Out[62]:  Text(0.5, 1.0, 'z-statistic = 1.96 and -1.96')
```

Figure 3: Non reject and reject areas.

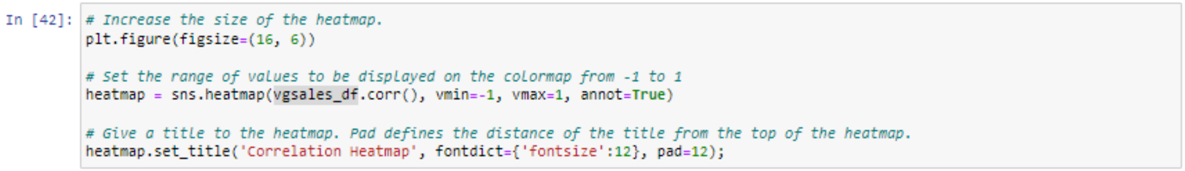
The critical values are at 95% level of confidence -1.9600 to 1.9600 as we can see on the table below. Even z_stat= -17.42 is less than z_critical, because is a two tailed test, is out of the non-rejection area, so we REJECT the null hypothesis and accept the alternative. So, we cannot say that the average age of the players that are at FIFA football game is equal to the real life with 95% level of confidence. In Figure 3, it is possible by the graphical visualization that the value of the statistical z is within the rejection of the null hypothesis.

2. Carry out a correlation analysis between 2 variables. Interpret your results and check if the correlation implies causation.

Provide a short explanation and conclusion based on your findings.

The most common measure of correlation is the Pearson correlation coefficient is a measure of the strength of a linear association between two variables. It can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; directly proportional. A value less than 0 indicates a negative association; that is, inversely proportional. 1 an -1 are the most strong relation between variables, normally we can see that when it is the correlation of the variable with itself.

The variables I chose to analyse the correlation are: Quality and market value, seen in the dataset as: "**Overall**" and "**value_eur**", respectively. As we can see in the heatmap the scale between -1 and 1, which is the maximum scale of correlation between variables. Despite the strongest correlation between variables being market value("value_eur") and player's salary("wage_eur"), overall and value_eur were chosen because for the analyst it makes more sense to analyse the correlation between the quality and value of the player in the market than the market value and the player’s salary, since the two are related to money and obviously are part of the same logic, one is not depended of the other. As we can see the correlation value 0.63 demonstrates a **moderate and positive correlation**. This means that the better the quality of the player the higher the sale price of him in the football market.



```
In [42]: # Increase the size of the heatmap.
         plt.figure(figsize=(16, 6))

         # Set the range of values to be displayed on the colormap from -1 to 1
         heatmap = sns.heatmap(vgsales_df.corr(), vmin=-1, vmax=1, annot=True)

         # Give a title to the heatmap. Pad defines the distance of the title from the top of the heatmap.
         heatmap.set_title('Correlation Heatmap', fontdict={'fontsize':12}, pad=12);
```

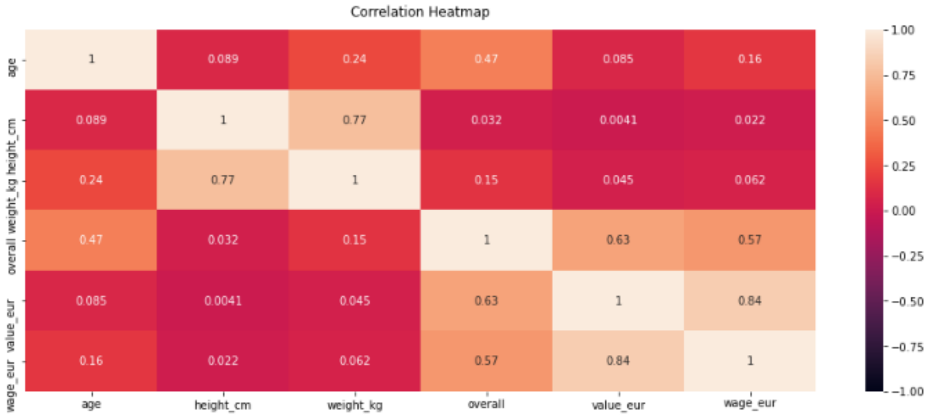
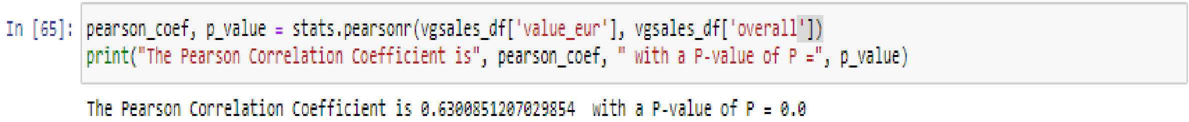


Figure 4: Heatmap before data preparation



```
In [65]: pearson_coef, p_value = stats.pearsonr(vgsales_df['value_eur'], vgsales_df['overall'])
         print("The Pearson Correlation Coefficient is", pearson_coef, " with a P-value of P =", p_value)

         The Pearson Correlation Coefficient is 0.6300851207029854  with a P-value of P = 0.0
```

Figure 5: Pearson coefficient and p-value before data preparation

Since the p-value is < 0.001, the correlation between market value(‘value_eur’) and Quality of the player(‘overall’) is statistically significant, although the linear relationship isn’t extremely strong (~0.63).

- Correlation: a measure of the association between variables.

- Causation: the relationship between cause and effect between two variables.

A confounding variable is an unmeasured third variable that influences both the supposed cause and the supposed effect.

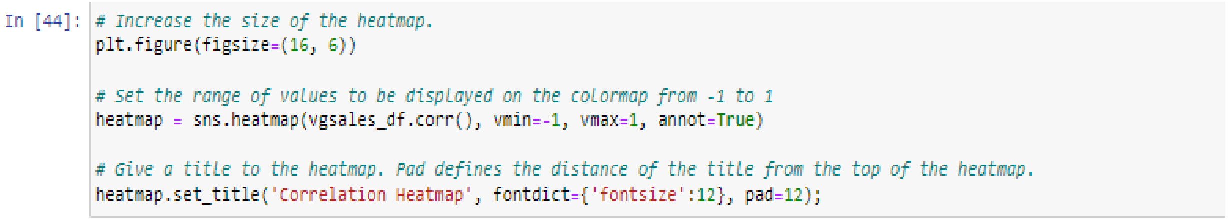
The third variable problem means that a confounding variable affects both variables to make them seem causally related when they are not.

The directionality problem is when two variables correlate and might actually have a causal relationship, but it’s impossible to conclude which variable causes changes in the other.

In our dataset of FIFA it is clear that correlation does not influence the cause due to a third variation. For example, the age of the player directly influences the value of his market, since, players retire very young, and hiring an older player may imply a sooner retirement. To exemplify this, I will take as an example L. Messi the player with the highest overall, that is, the most qualified in the market, within the parameters considered by FIFA. It is 33 years old and its market value is 67,500,000 euros. Kylian Mbappé is only the sixth best player, but the fact that he is only 21 years old (when the input data was collected) makes him much more expensive than L. Messi 105,500,000 euros. Kylian Mbappé still has plenty of career time ahead while Messi only a few years.

In order to increase the correlation between the variables, two data preparations were applied. First of all, I set the age at 23 years, once it became clear, by the career time that remains to the player that this variable interferes and a lot in market value. In addition, I dropped the outliers, because very qualified players and with very strong names in the football market like Messi, Cristiano Ronaldo and Neymar not only worth the quality of them, when these players are bought by a club, the club gains more visibility, sells many products with the name of these players, closes more advantageous sponsorships, increase in the number of socio fans who give profit to the team. Therefore, these market values differ greatly from a player with the same quality but that do not bring these financial returns.

After this data processing we obtained the following heatmap:



```
In [44]: # Increase the size of the heatmap.
         plt.figure(figsize=(16, 6))

         # Set the range of values to be displayed on the colormap from -1 to 1
         heatmap = sns.heatmap(vgsales_df.corr(), vmin=-1, vmax=1, annot=True)

         # Give a title to the heatmap. Pad defines the distance of the title from the top of the heatmap.
         heatmap.set_title('Correlation Heatmap', fontdict={'fontsize':12}, pad=12);
```

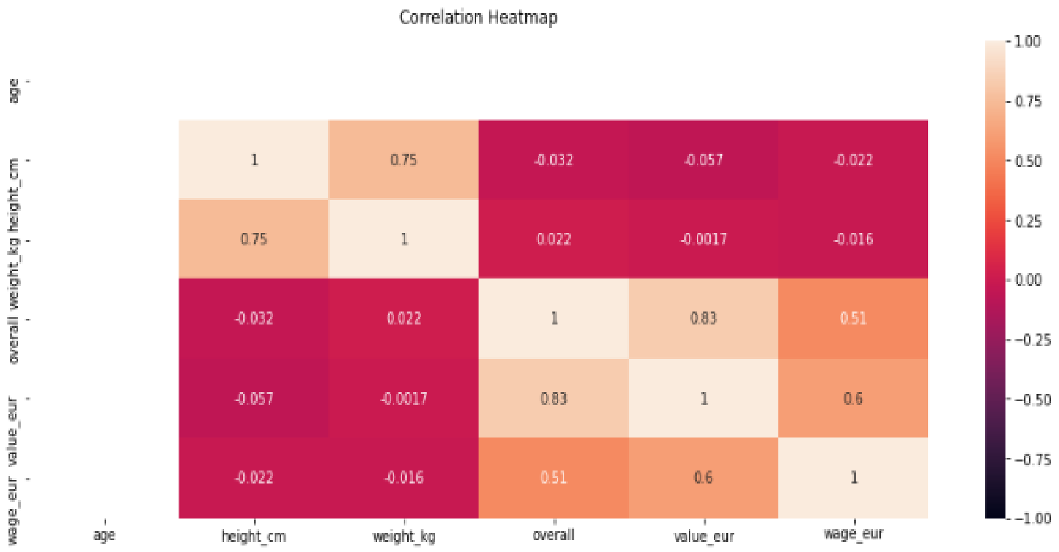


Figure 6: Heatmap after data preparation.

Since the p-value is < 0.001, the correlation between market value('value_eur') and Quality of the player('overall') is statistically significant, in addition, the correlation is now **strong and positive** (~0.83).

In conclusion for this section, based on the above analysis it was clear that that correlation does not imply in the causation, and in the example above, the third variable age affects the causation, after defining an age and removing the outliers the correlation increased considerably.

3. Using the same 2 variables in the second section, build a linear regression model that allows you to predict information about those variables. Interpret your results, and provide a short explanation and conclusion based on your findings.

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. A linear regression line has an equation of the form ***Y* = *a* + *bX***, where ***X*** is the explanatory variable and ***Y*** is the dependent variable. The slope of the line is ***b***, and ***a*** is the intercept (the value of ***y*** when ***x*** = 0).

If we have two variables in a linear correlation, even though the two variables that we are analysing is not one dependent and another one independent we can use statistical linear regression to predict those variables. Because we can have variables that have correlation, but that doesn't mean that one variable causes the another one, as we already describe in a previous second section.

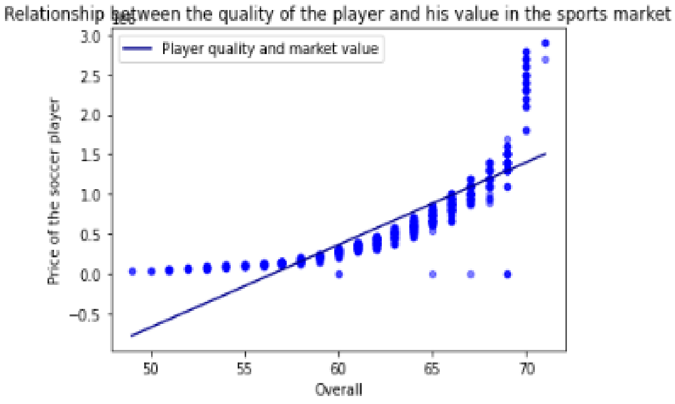


Figure 7: Graph of linear regression model and data points.

y=-5860124.68+103720.38*x

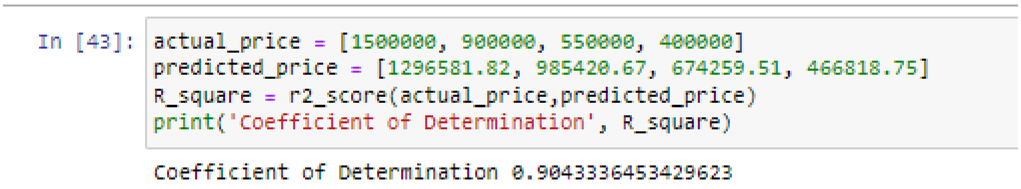
Figure 8: Regression line model.

The linear function seen above (y=103720.38x-5860124.68) is the function that will predict the player's market values based on overall. Even considering the line the best fit to predict the player's market value still to errors, we can see that because there is a distance between the points, which are dice, and the line. For every 1 point, the overall amount (x) increases, we would expect the value in the football market amount to increase by 103720.38 €.

As we can see this relationship generates a growing line, proven by the positive value of "a" (103720.38). This was also expected, because, as seen in the second session the relationship between the variables chosen is positive, that is, when the player's quality increases, his market value also increases. The value of “B” represents the player's market value if he does not present any quality (overall =0). Since it is not possible to have a player with no skills, the lowest overall seen among players is 47. This makes us even though we know that there is no negative market value, the model considers overall values too low as negative, since the smallest data that the model has received has overall 47.

The centroid is "best fits" the data, it makes sense that the line passes through the means. In our dataset, the centroid is (63.09, 683526.63). The centroid is one of the most important point in the linear regression model, the point tell us where the line will cross.

R-squared (R^2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. To test the accurancy of our model I tested 4 players with 23 years and different overall. The overall tested are close to most players. The players used to test the model were: Rodri TarÃn, J. Simpson, Henrique Trevisan and Pedro LÃ³pez.



```
In [43]: actual_price = [1500000, 900000, 550000, 400000]
         predicted_price = [1296581.82, 985420.67, 674259.51, 466818.75]
         R_square = r2_score(actual_price,predicted_price)
         print('Coefficient of Determination', R_square)

         Coefficient of Determination 0.9043336453429623
```

Figure 8: R^2 of the regression line model.

Considering the market values that are in the dataset and predicting their values according to the linear function generated in the model the result was that the "r" square was 0.9043. That’s means, approximately 90% of the observed variation can be explained by the model's inputs. A high R-squared is between 85% and 100%. So, that’s a very good proportion that is represented by the model.

CONCLUSION

The study verified whether the average age of football players registered in the FIFA football game is equal to the average age of real-life players. After the hypothesis test using two tailed z test, due to the sample size, and the alternative hypothesis is not equal, it was defined based on the statistical z and z critical the rejection of the null hypothesis. That is, the actual average age and age in the game are different.

In the second session initially, we had a positive and moderate correlation between the variables "overall" and "value_eur", which shows the relationship between the player's quality and their market value. However, after defining that the player's age greatly affected his market value, a single age (23) was defined. In addition, outliers who are players far above the general media and who bring a financial benefit beyond football games, have been cut. After this data processing, the correlation between "overall" and "value_eur" became strong and positive.

In the third session, the statistical model of linear regression was used to define the equation that would predict the player's market value based on its quality. It is worth remembering that the model was developed after data preparation (age = 23 and no outliers). The equation defined in the model was: "y=103720.38x-5860124.68". After testing this equation with four players with different "overall" the value of R^2 = 0.90 was obtained. This indicates that the model predicts well the player's market value based on the quality.

For the next steps, it would be ideal to consider the position of the player. The existing positions are: forward, midfield, defending, and goalkeeping positions. And it is popular knowledge that forward, midfield usually has better market value than defending, and goalkeeping position.

REFERENCES

[1]www.kaggle.com. (n.d.). *FIFA 21 complete player dataset*. [online] Available at: https://www.kaggle.com/stefanoleone992/fifa-21-complete-player-dataset?msclkid=a9b5a7daa78511ecaf60535548ab5ea4 [Accessed 25 May. 2022].

[2] Lakin, S. (2011). *How to use statistics*. Harlow: Prentice Hall[Accessed 27 May 2022].

[3] Kalén, A., Rey, E., de Rellán-Guerra, A.S. and Lago-Peñas, C. (2019). Are Soccer Players Older Now Than Before? Aging Trends and Market Value in the Last Three Decades of the UEFA Champions League. *Frontiers in Psychology*, 10. doi:10.3389/fpsyg.2019.00076[Accessed 27 May 2022].

[4] Rowntree, D. (2000). *Statistics without tears : an introduction for non-mathematicians*. Harmondsworth: Penguin[Accessed 27 May 2022].

[5] www.ibm.com. (n.d.). *About Linear Regression | IBM*. [online] Available at: https://www.ibm.com/topics/linear-regression[Accessed 27 May 2022].

[6] Majaski, C. (2021). *How Hypothesis Testing Works*. [online] Investopedia. Available at: https://www.investopedia.com/terms/h/hypothesistesting.asp#:~:text=Hypothesis%20testing%20is%20an%20act[Accessed 27 May 2022].

[7] Zach (2021). *How to Perform One Sample & Two Sample Z-Tests in Python*. [online] Statology. Available at: https://www.statology.org/z-test-python/[Accessed 27 May 2022].

[8] YALE (2019). *Linear Regression*. [online] Yale.edu. Available at: http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm[Accessed 27 May 2022].

[9] Hayes, A. (2020). *R-Squared*. [online] Investopedia. Available at: https://www.investopedia.com/terms/r/r-squared.asp#:~:text=R%2Dsquared%20(R2) [Accessed 27 May 2022].

[10] Laerd Statistics (2018). *Pearson Product-Moment Correlation - When you should run this test, the range of values the coefficient can take and how to measure strength of association.* [online] Laerd.com. Available at: https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php[Accessed 27 May 2022].

[11] Scribbr. (2021). *Correlation vs. Causation | Differences, Designs & Examples*. [online] Available at: https://www.scribbr.com/methodology/correlation-vs-causation/#:~:text=A%20correlation%20is%20a%20statistical%20indicator%20of%20the [Accessed 27 May 2022].

[12] www.calculators.org. (n.d.). *Z Critical Value Calculator*. [online] Available at: https://www.calculators.org/math/z-critical-value.php [Accessed 27 May 2022].

[13] Thomas, L. (2020). *Confounding Variables* | *Definition, Examples and Controls*. [online] Scribbr. Available at: https://www.scribbr.com/methodology/confounding-variables/[Accessed 27 May 2022].