Facebook Engagement Analysis
Statistical Techniques

TABLE OF CONTENTS

Introduction

This is an analysis of a Spotify graph for research statistics and an analysis of a Facebook dataset for descriptive statistics

Question 1: Research Statistics

It is important to analyse your dataset before approaching a statistical breakdown of the data itself. There is an approach that needs to be considered when looking at categorical versus nominal data. It is also important to view the size of the data in terms of deciding a sample range for any representation of the data as a whole (Sribbr, 2022) In this analysis we look at the following graph (Figure 1)

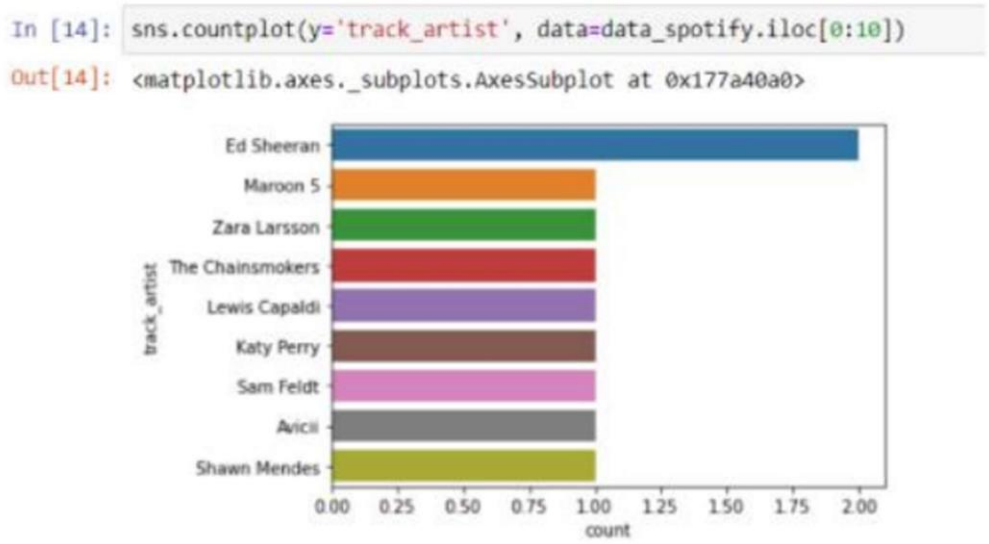

Figure 1

1.1 Symmetric Distribution Statistics

Symmetrical distribution is normal distribution in statistics. Where a symmetrical distribution exists where each side mirrors each other in Gaussian Curve. For a 'perfect' symmetrical distribution, it is accurate to say that the mean = mode = median(Zach, 2021)(Figure2) (Lehmann, 2020)

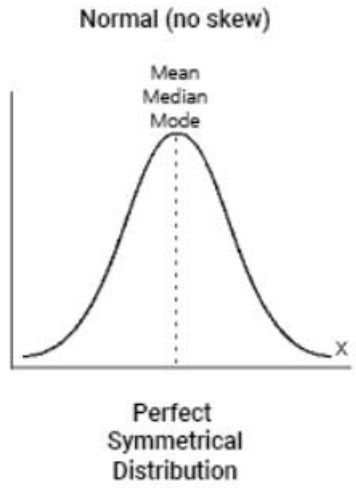

Figure 2

However, it is not always the case that this applies. While the normal distribution is symmetrical, not all symmetrical distributions are normal (Frost, 2018)

Essentially, a normal distribution is always symmetrical and the total area under the curve will always equal to 1. Normal distribution is probability distribution (Bhandari, 2020) based on central limit theorem. Symmetrical distribution works around the mean as the central point, however the distribution can still be symmetrical without it equalling the mode. A Bimodal distribution can still be symmetrical without mean = median = mode (Figure3)(Holt, 2015)

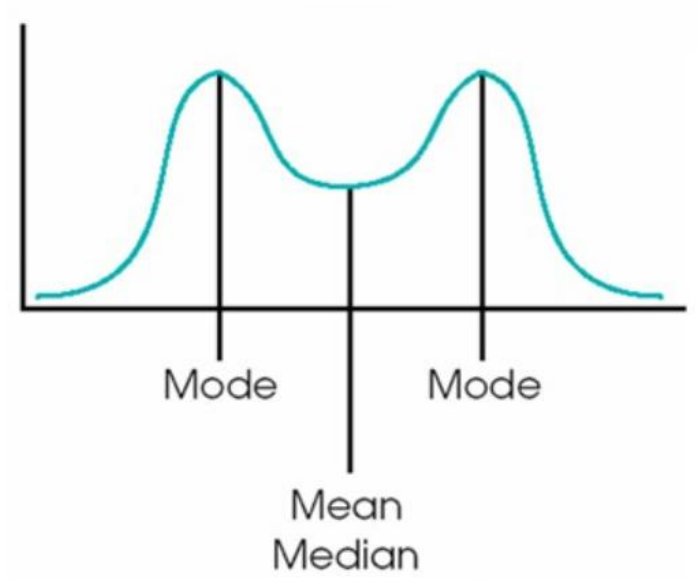

Figure 3

Below is a diagram of symmetrical distributions to illustrate further there are different types(Figure4)(Padhai Time, n.d.)

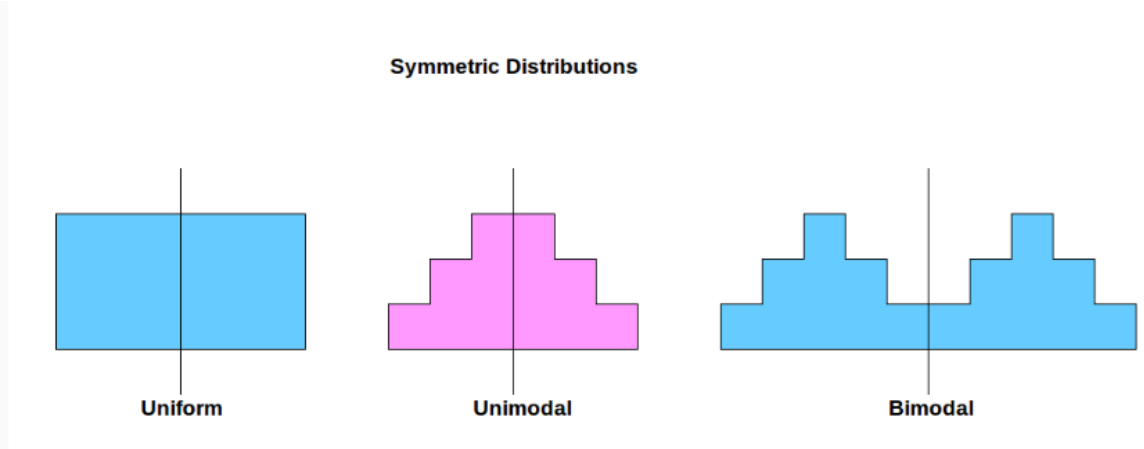

Figure 4

1.2 Can we deduce that there are only 8 different track artists in our entire dataset?

We cannot deduce that there are 8 different track artists in this dataset. If we use the function print(df[track_artist].nunique()) we can view the number of unique values within this feature in the dataset which equal 9

1.3 In the range analysed, can we deduce there is a unique mode?

The mode is the most frequently occurring value in a set of numbers. A unique mode is when the there is only one set of numbers that the highest frequency. It is fair to deduce therefore in this dataset that there most certainly is a unique mode of the track artist 'Ed Sheeran' as he has the highest count in the histogram which means that is the data value that occurs most often in the dataset (Foundation, n.d.). To establish this in python we would use mode = dataset[df].mode()

print(mode)

1.4 Is it true that in the range analysed the mode has a value of 2?

It measures of central tendency; the mode is a key indicated in the determination of the distribution of a dataset. Mode is an appropriate average in case of qualitative data i.e., the opinions or preferences of people in their choice of musician. It is a fair conclusion therefore that the mode has a value of 2, being Ed Sheeran as the track artist

1.5 How many data elements there are in the range analysed?

The python range is a built-in function. The range() function prints a list or sequence of numbers, beginning from 0 by default, and increments the return value by 1 (by default), and stops the list at a specified number given by the user. range() function is just a renamed version of the xrange() function which was used in Python2.

Question 2 Descriptive Statistics

The word "descriptive statistics" refers to data analysis that aids in describing, displaying, or summarizing data in an understandable fashion so that, for instance, patterns may appear from the data. However, descriptive statistics do not let us draw any inferences from the data we have examined or come to any conclusions about any potential hypotheses. They serve only as a means of describing our data ("Understanding Descriptive and Inferential Statistics | Laerd Statistics," n.d.)

2.1 Determine the most common hour to create a post on the Facebook page

This requires us to retrieve the mode of the feature 'Post_Hour'. The mode being the most common value in that feature, equal to the most common hour to create a post.

Answer: 3

2.2 Identify the average number of comments, likes, and shares and compare the result

Here I needed to determine the average (mean) of comments, likes, shares which are all separate features within the dataset. I created a bar chart to realise this outcome (Figure 5)

Answer:	Top Comments:	'Like'
	Middle Comments:	'Share'
	Lowest Comments:	'Comment'

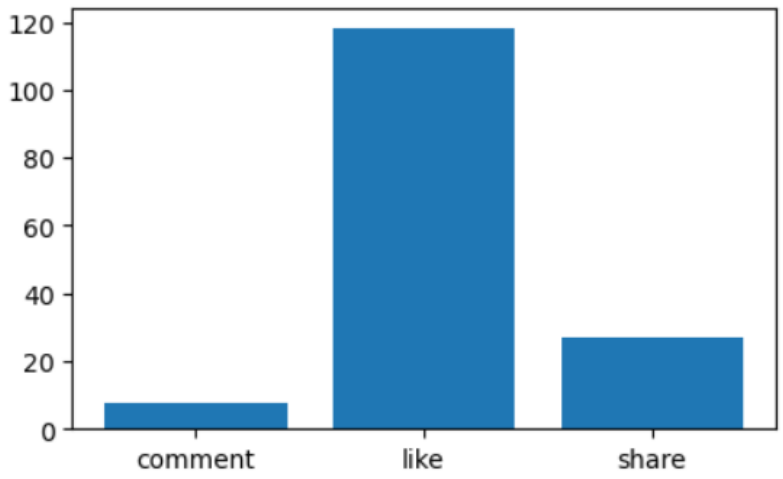

Figure 5

Answer: The average number likes exceeded the average number of comments and shares. While shares average was higher than comments. This engagement result can be helpful in finding where they need to improve their marketing as comments and shares are hugely important in terms of acquisitions of new followers

2.3 Calculate the percentage of each post type and indicate the category that has more posts

Here I calculated the counts of each value in the 'Type' Feature, divided it by the total (n) value and multiplied it by 100 to get the percentage. (Figure 6)

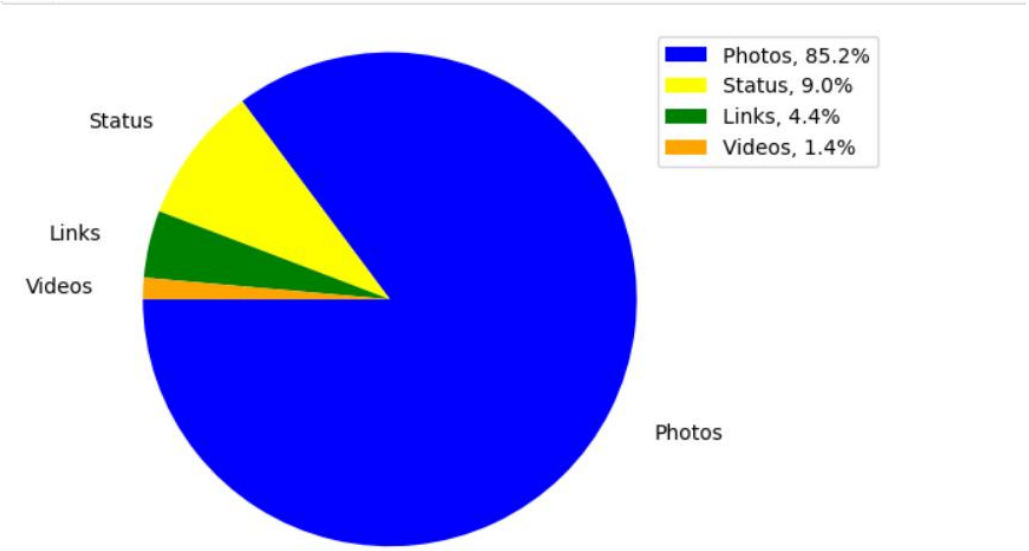

Figure 6

Answer Photos had the highest number of posts by a big margin

No 1 Posts: Photos

No 2 Posts: Status

No 3 Posts: Link

No 4 Posts: Video

2.4 Calculate the percentage of posts per month

Here I calculated the value counts and the total counts. I divided the values counts by the total counts and multiplied it by 100 to get the average percentage frequency (Figure 7)

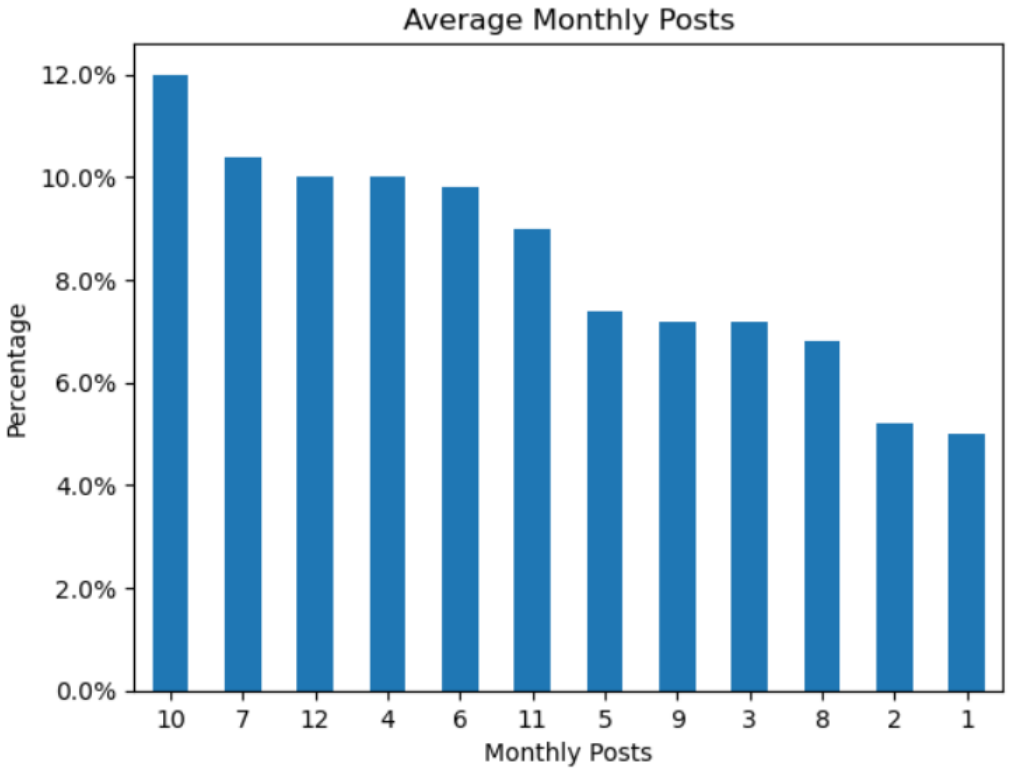

Figure 7

Answer: This is a good determining factor on the performance of a user

10 Posts 12%

7 Posts 10.4%

12 Posts 10.0%

4 Posts 10.0%

6 Posts 9.8%

11 Posts 9.0%

5 Posts 7.4%

9 Posts 7.2%

3 Posts 7.2%

8 Posts 6.8%

2 Posts 5.2%

1 Posts 5.0%

2.5 It is important to understand the audience to take actions in future. Determine whether there is a significant variation in the lifetime engaged users

Here I followed the formula for variance to determine the answer for lifetime engaged users

(Figure 5)

$$S^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

S^2 = sample variance

x_i = the value of the one observation

$\bar{x}$ = the mean value of all observations

n = the number of observations

Figure 5

I firstly used the python function .var to make sure my conclusion was correct. I determined the value minus the mean squared for each induvial observation. I then squared that and divided it by 'n' minus 1

The variance in statistics is the average squared distance between the data points and the mean (Frost, 2021). The higher the variance indicates the values are spread out from the mean and each other. The positive of this is that all values are treated the same in terms of deviation from the mean, however this can also lend weight to outliers in the data. The standard deviation is obtained by squaring the variance.

Here we can see that the variance is high, therefor we can deduct that there is significant deviation from the mean

Answer: Variance = 970752.9309619238

2.6 Calculate the number of likes reached by 50% of the sample. Identify the probability of finding up to 15 posts with that number of likes in a sample of 25 posts

I found this question a little difficult to understand. I took 50% to be the median. My logic here was that the halfway point in the number of likes. To calculate this, I took a random sample of 25 from the 'likes' feature. After determining the median value, I wanted to see how many times a number in the dataset was equal to or a above that number. Ironically the data set had 15 values that met this condition. The probability of 15 values meeting this condition out of 25 was 60%. The probability of picking 15 posts out of 25 that met this condition was 60%. So therefore, the probability that randomly picking 15 posts out of 25 that met this condition was 60% of 60.

Answer: 32%

2.7 Determine what is the type of post with more interactions

In this analysis I reduced the dataframe to post 'Type' and 'Total_Interactions'. I grouped them to calculated what interactions were attributed to each post type (Figure 6)

Answer No 1 Interactions: Status

No 2 Interactions: Links

No 3 Interactions: Videos

No 4 Interactions: Photos

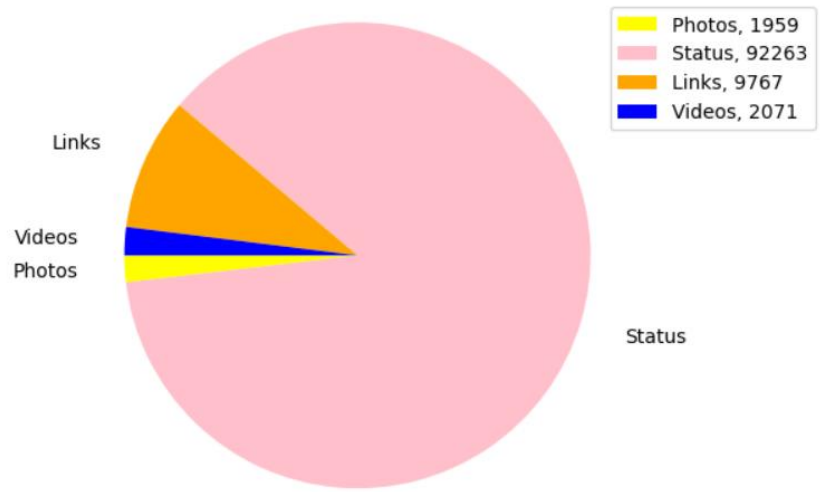

Figure 6

I then calculated the percentage of these values to represent them in a chart (Figure7)

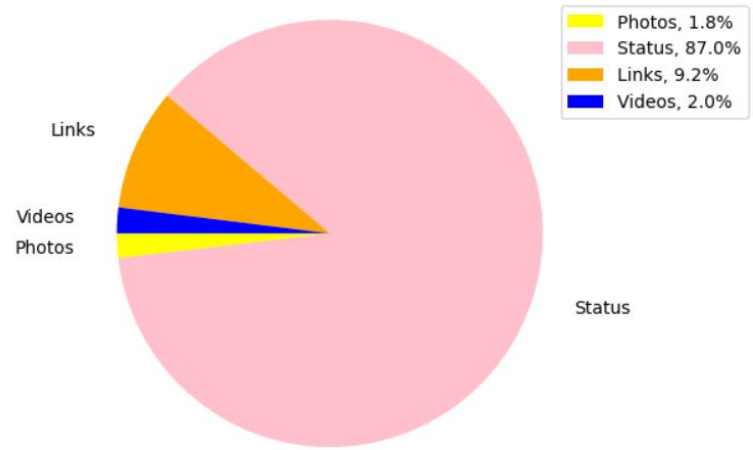

Figure 7

Answer: Status is the type with the highest Interactions

Status = 87%

Links = 9.2%

Videos = 2%

Photos = 1.8%

2.8 In average, are there more posts from consumers or consumptions?

In this case I took the average, mean, of both consumers and consumptions to determine the distribution (Figure 8)

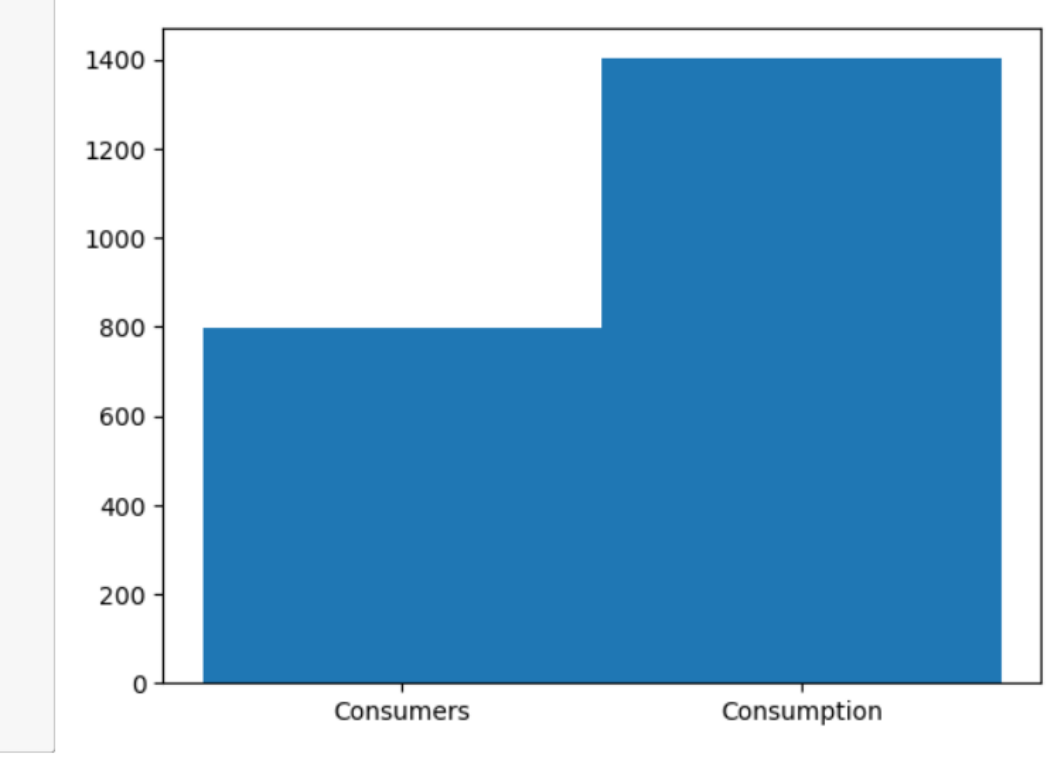

Figure 8

Answer: Engagement by Consumption is on average higher than Consumers

Consumers =799

Consumption = 1401

Executive Summary

Statistics play an important part for Facebook in determining the different factors involved in both consumers and consumption. There is varying data in that consumption versus consumers are notably different. For the marketers they must consider various elements in terms of the statistical data provided. For instance, status updates are the most prevalent in terms of engagement, and this needs to be considered when looking at average hours posted and the variance of the engaged users. Statistics plays a vital role to marry different elements together to conclude the most affluent selling point or in Facebook context, what targeted marketing is most appropriate based on uses, hours engaged, content, mutual engagement and profiling, which is not included in this analysis

Analytics in Facebook is also extremely important from an individual business perspective. From these analytics, one can determine the most engagement from particular posts in terms of promotions, events, marketing type, leads, and followers.

Bibliography

Bhandari, P., 2020. Normal Distribution | Examples, Formulas, & Uses [WWW Document]. Scribbr. URL https://www.scribbr.com/statistics/normal-distribution/ (accessed 12.2.22).

Foundation, C.-12, n.d. Histograms | CK-12 Foundation [WWW Document]. URL https://flexbooks.ck12.org/cbook/ck-12-basic-probability-and-statistics-concepts/section/7.8/primary/lesson/histograms-bsc-pst/ (accessed 12.3.22).

Frost, J., 2021. Variance: Definition, Formulas & Calculations. Statistics By Jim. URL https://statisticsbyjim.com/basics/variance/ (accessed 12.11.22).

Frost, J., 2018. Normal Distribution in Statistics. Statistics By Jim. URL http://statisticsbyjim.com/basics/normal-distribution/ (accessed 12.2.22).

Holt, R., 2015. Elementary Statistics for the Social Sciences (UC:CSU) - 3 units - ppt download [WWW Document]. URL https://slideplayer.com/slide/6380352/ (accessed 12.11.22).

Lehmann, J., 2020. Statistics and Excel: Evaluating Normality.

Padhai Time, n.d. Distribution Shapes | Padhai Time [WWW Document]. URL https://padhaitime.com/Statistics/Distribution-Shapes (accessed 12.11.22).

Sribbr, 2022. Statistics Archieven [WWW Document]. Scribbr. URL https://www.scribbr.com/category/statistics/ (accessed 12.11.22).

Understanding Descriptive and Inferential Statistics | Laerd Statistics [WWW Document], n.d. URL https://statistics.laerd.com/statistical-guides/descriptive-inferential-statistics.php (accessed 12.11.22).

Zach, 2021. Symmetric Distribution: Definition + Examples. Statology. URL https://www.statology.org/symmetric-distribution/ (accessed 12.2.22).