TU DORTMUND – DATA SCIENCE STATISTICS REPORT

JANUARY 14, 2023

MANUEL ALEJANDRO MONTOYA SILVA

Contents

Introduction

The animal kingdom plays a huge role in the world's food chain, and humans need from animals to maintain an order within the ecosystem. There exist multiple endangered species, and to save the most of them, there are multiple sanctuaries built all around the world, in which endangered species are taken care of. In order to build spacious and suitable habitats for each specific species, it is a must to understand the needs and traits each species has. The dataset of the animal kingdom contains traits that could help build more comfortable habitats with the correct dimensions and needs for all species. A regression model will be performed in this dataset to understand how the weight and movement type influences an animal's speed, or in case it does not, what is needed in order to fully understand how species will behave in certain environments. There is an initial intuition that says that larger animals tend to be faster, yet the heaviest are not the fastest animals. Performing the analysis, it is confirmed that weight describes very little the highest speed a species can go up to in a linear way (based on correlation values and coefficient of determination), therefore other variables are needed to fit an accurate linear model, or try fitting other type of models that could describe the pattern/relationship weight has with the high speed.

Description

The following statistical report main goal, is to find what role does the weight of any species plays in the high speed it can go up to. For this, the following questions were asked:

-Does the movement type and weight determine how fast an animal can go?

-Is there any movement type group significantly faster than the others?

-Larger animals tend to have higher speeds?

In order to answer the research questions, the first step of the report is to get started with the dataset. The dataset is made of 4 columns: weight, highspeed, movement type and species. Species is the name of the animal species (categorical) with only identification purposes, weight is the species' average weight (quantitative), highspeed is the highest species' speed (quantitative) and movement type explains how the animal moves, which could be: flying, climbing, running, swimming (nominal categorical). There are a total of 159 species with no null values. Dataset entries were gathered from

different sources since they are very specific details from any species. There is a biased selection as well, due to usually high speeds are often known only for the fastest animals.

Methods and Concepts

For the following section, all methods and concepts used will be explained in a mathematical and statistical way.

Linear Regression

Linear Regression is a supervised machine learning model which tests a linear relationship between a quantitative dependent variable and one or multiple independent variables which can be either quantitative or categorical (in categorical variables, data has to be transformed in order to work in a linear regression model). It tries to find a line that describes data in the best way possible in order to predict future trends. The line is defined by a slope and an intercept which can be referred as weights. The main goal of linear regression is to find the values for the weights that fit data the best. (Codecademy, n.d.)

A simple linear regression formula is described as follows:

$$y = mx + b$$

Loss Function

In order to calculate the best values for the slope and intercept, linear regression uses a loss function which determines how well the model is performing. The most used function is the least squares error and it is described as follows:

$$\sum_{i=1}^{m} (y^i - \bar{y}^i)^2$$

In simple words, it calculates the sum of the square difference between all data points to the regression line. The idea of linear regression is to find the best slope and intercept that minimizes the loss. (Mirtaheri, 2022)

Ordinary Least Squares

OLS is an optimal method to find a linear regression's weights (coefficients), since it is defined by an enclosed formula:

$$w = (X^t X)^{-1} X^t y$$

Where X is a matrix with the training samples, X^t is the transposed matrix of X and y is a vector with all real output values for the training samples. Since computing the inverse matrix takes more time and space in memory, gradient descent is more practical for large datasets than OLS since it is an iterative process. (Mirtaheri, 2022)

Coefficient of Determination (R^2)

The coefficient of determination represents how well the data is fitted and how well predictions are going to be made. It determines how the variance of the dependent variable is explained by the independent variables used for the model. Coefficient values closer to one mean a better fit than the ones closer to zero. Usually values above 0.7 are considered a good fit. (Haider, 2016).

Pearson Correlation

Pearson Correlation coefficient is a scaled form of covariance, which measures the strength of a linear relationship with values ranging from -1 to +1 for quantitative variables. Positive relationships between variables are close to +1, while negative relationships are close to -1. It exists an association when the coefficient is 0.3 and it indicates a strong association when the coefficient is 0.6 or greater. A value of 0 means there is no association at all. (Codecademy, n.d.)

ANOVA Test

ANOVA (Analysis of Variance) is a statistical test that is used to determine differences between a quantitative and a categorical variable. The dependent variable is split into two or more categorical groups in order to find a statistical difference between groups by analyzing each group's mean using variance. (Simkus, 2022)

Natural Log Transformation

This data transformation works better with right-skewed distributions. It helps compress the range between values resulting in data being closer to a normal distribution by taking the log of all variable's values. (Codecademy, n.d.).

One-hot Encoding Transformation

Sometimes categorical data has to be transformed into a number in order to be interpreted by a machine learning model. One-hot encoding is a binary vector representation of a categorical variable values where all indexes are zero-valued except for the position of the variable's value where it is a one. (Deshpande & Kumar, 2018).

Normal Distribution

Also known as the Gaussian distribution, it is a distribution that has a symmetrical bell-shape around its mean, where most of the values are close to a central peak. It is mostly used on independent random variables. (Frost, 2020).

Skewness

In statistics, skewness is the asymmetry of a distribution compared to a normal distribution. Most values tend to be on the sides of the distribution instead of being around the central peak, resulting in a tailed-distribution. These types of distributions are a result of having multiple outliers on the extremes. (Gawali, 2022).

Histogram

Histograms are plots that display graphically the frequency distribution of a continuous variable. It is useful to identify skewness, outliers and distributions. (Laerd Statistics, n.d.).

Box plot

Boxplots are plots that display the distributions of quantitative variables across different groups of a categorical variable. It displays the variable's quartiles, maximum and minimum values and its outliers. It is useful to compare distributions and differences between a quantitative variable and a categorical variable. (Seaborn, n.d.)

Scatter plot

Scatter plots display graphic relationship between two quantitative variables represented by dots with cartesian coordinates. One variable works as the X ax, while the other variable works as the Y ax. It is useful to confirm patterns between variables whether it is a linear or non-linear relationship. (Corporate Finance Institute, n.d.)

Evaluation

The dataset has four variables (animal, weight, movement_type, highspeed) from which the *animal* variable is used only for identification purposes and it specifies the animal species. *Weight* is defined kilograms, with a mean of 1,994.11 and a standard deviation of 12,285.31 with a range of values between 0.001 to 140,000. Further details can be seen in the *figure 1* and *figure 2*:

	weight
count	159.000000
mean	1994.115969
std	12285.311105
min	0.001000
25%	2.625000
50%	22.000000
75%	310.000000
max	140000.000000

Figure 1. Weight variable's statistics

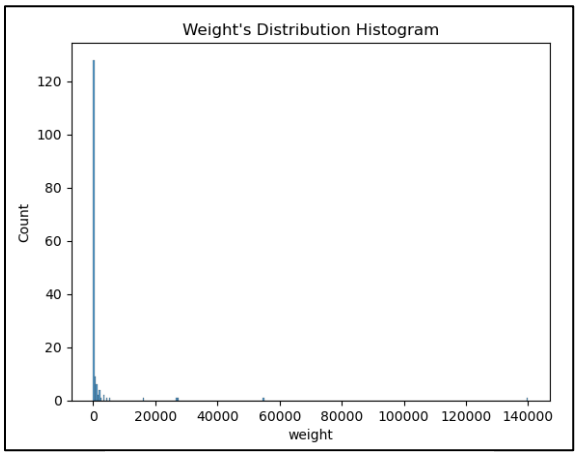

Figure 2. Weight's Distribution Histogram

By analyzing the weight's quartiles and the maximum and minimum values, there exists an obvious wide scattered range, since 75% of the data are values lower than 310 kilograms. This means there are multiple outliers that highly right skew the variable's distribution, hence the high values for the standard deviation and the mean. Therefore, the mean and the standard deviation are not representative values that could explain the typical weight that can be found in the dataset. Other more reliable robust measures towards outliers could be used in this case, the IQR and the median, which are 307.38 and 22 respectively.

The *high-speed* is measured in km/hr. It has a mean of 52.64 and a standard deviation of 34.46 with a range of speeds between 1.5 and 195 km/hr. Further details can be seen in *figure 3* and *figure 4*:

	highspeed
count	159.000000
mean	52.640881
std	34.460258
min	1.500000
25%	30.000000
50%	48.000000
75%	70.000000
max	195.000000

Figure 3. Highspeed variable's statistics

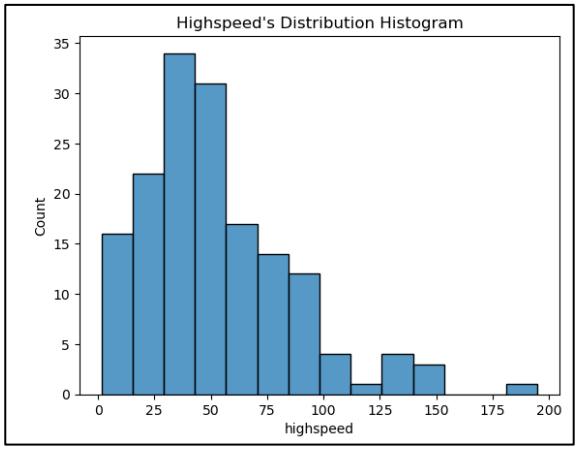

Figure 4. Highspeed Distribution Histogram

Analyzing the distribution histogram and the quartiles, there are some outliers that slightly skew the distribution to the right. Although, the mean and the standard deviation seem like a representative value of the typical high-speed that can be found within the dataset but a more accurate value could be its median, with 48 km/hr since 50% of the values are below 48 with an IQR of 40.

Movement_type categories are flying, climbing, running and swimming animals, where almost half of the high-speeds measured are for running animals. Frequencies can be seen in *Table 5*:

Table 5. Movement type categories frequencies

Movement type	Frequency	Percentage
swimming	41	26%
running	78	49%
climbing	13	8%
flying	27	17%
TOTAL	159	100%

Grouping the different species by their movement type, the weight and high speed can be analyzed in an easier and clearer way to find insights that could help understand the variables' distributions and relationships. *Weight* and *Highspeed* values grouped by movement type in *figure 6* and *figure 7*

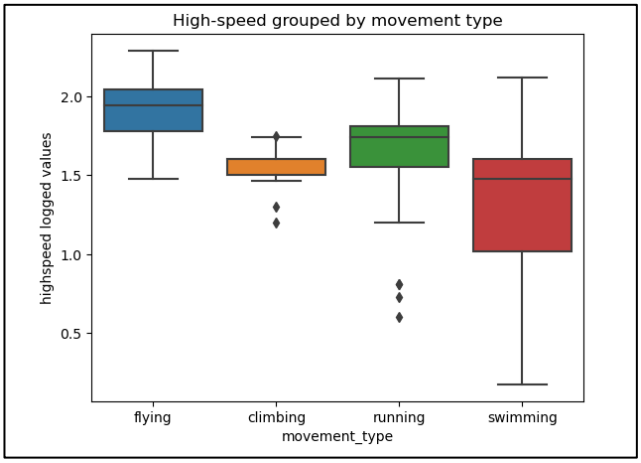

Figure 6. Highspeed grouped by movement type

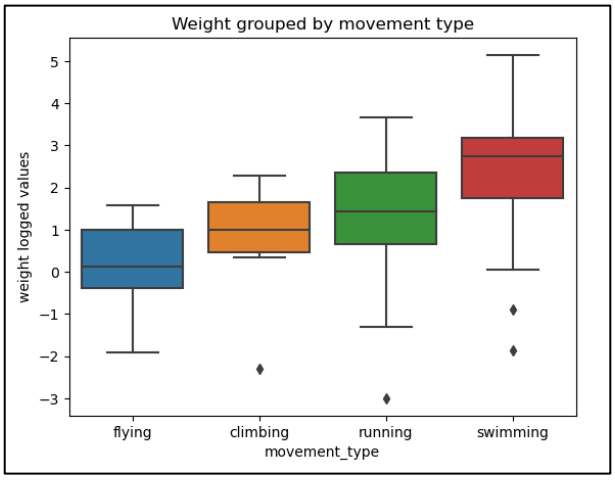

Figure 7. Weight grouped by movement type

In order to have a better look of the data, since high-speed and weight have a right skewed distribution, a natural logarithmic data transformation was applied. By visually analyzing the boxplot of the movement type against the highspeed, there exists a small relationship between these two variables, where flying animals tend to have higher speeds compared to other movement types, while swimming animals tend to have lower speeds. To confirm this relationship between high-speed and movement type, an ANOVA Test was done using *f_oneway* method from the *scipy.stats* python library (Van Rossum, 1995) with a significance threshold of 0.05, a null hypothesis of all movement types' high-speed mean is the same and an alternate hypothesis in which the means differ by movement type giving as a result an f-stat of 17.89 and a p-value of 5.04e-10. Weight variable was also grouped by their movement type for an ANOVA Test, in order to measure the relationship between weight and movement type, with an f-stat result of 19.43 and a p-value of 9.512e-11. Same tests were done without the outliers and it had a very similar result, with a final result of a significant difference between groups. For further details of which groups differ, *tukey_hsd* method was used from *statsmodel* library from python (Van Rossum, 1995) which can be seen in the appendix in *figure 10* and *figure 11*.

To measure the relationship between high-speed and weight, since they are quantitative variables, the pearson correlation is a more suitable method for this type of relationship. In python, the *pearsonr* method was used from the *scipy.stats* library (Van Rossum, 1995) giving a correlation value of -0.034. *Figure 8* displays the relationship between high-speed and weight by movement type. Additional graphs scatter plots can be seen

in the appendix for a more specific detailed view (*figure 12, figure 13, figure 14, figure 15*).

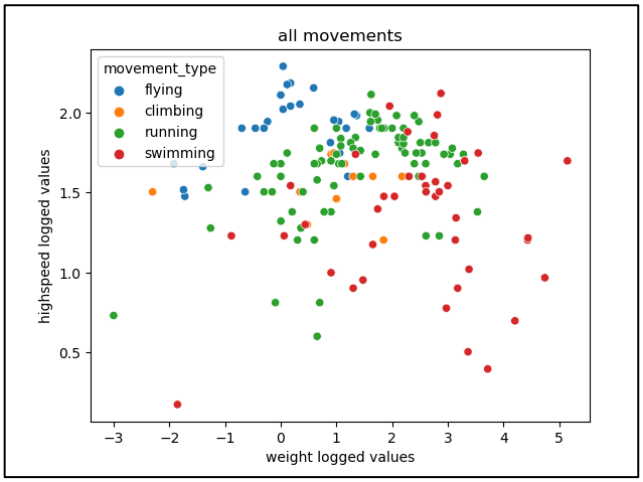

Figure 8. Weight vs Highspeed grouped by movement type

Before fitting the model, data needs to be scaled and transformed. To use categorical data such as the movement type within a linear regression model, one-hot encoding was applied in order to be fitted in the model with *pandas* method *get_dummies()*. (Van Rossum, 1995)

Table 9. Movement type variable one-hot encoded new column values

movement type	flying	running	swimming
flying	1	0	0
running	0	1	0
swimming	0	0	1
climbing	0	0	0

For the linear regression model, *LinearRegression()* class from the *sklearn* python's library was used (Van Rossum, 1995), which applies OLS method to find the model's weights. Since it is a small dataset and memory is not an issue, OLS is a better method to find accurate weights than gradient descent. Data was split into a training set and test set with an

80% to 20%. The model was fitted with the training set and the following features: weight, flying, running, swimming. The model's formula is defined as follows:

$$y = 0.05x_1 + 0.4x_2 + 0.07x_3 - 0.29x_4 + 1.47$$

Where y is the animal's calculated high speed in km/hr., x_1 is the animal's weight in kilograms, while variables x_2, x_3, x_4 represent the movement type explained in *Table 9.*

Using *.score()* method from the *LinearRegression()* class from the sklearn python's library (Van Rossum, 1995), the final model has a score of 0.29 on the test set. Analyzing the model's formula, big coefficients mean a higher influence in the dependent variable; therefore, the weight has a small influence on determining the animal's highest speed. Meanwhile for flying and swimming animals is easier to determine their speed, a confirmation of this is the movement type box plot and the ANOVA Test, in which these groups ranges are clearly different from the others, with flying animals being the fastest (+0.4 highest positive coefficient) and swimming animals being the slowest (-0.29 negative coefficient). This means higher speeds for flying animals than swimming animals.

Summary

The most important results will be presented and research questions answered.

-Is there any movement type group significantly faster than the others?

After performing an ANOVA Test and interpretating the f-stat and p-value, as well as after analyzing *Figure 6* it can be concluded that flying animals are faster than any other type of animals. Meanwhile swimming animals record the lowest speeds compared to other groups. As a note, by adding more data entries and having similar number of samples for each movement type, could result in more accurate statistical results.

-Larger animals tend to have higher speeds?

Analyzing the high-speed grouped by the movement type, there is a range of weight values in which larger animals have higher speeds within their movement type, but there is also a threshold in the weight's ax, from which species start decreasing its speed. Taking the flying animals' weight-highspeed in *figure 12* as an example, the initial range of values of -2.0 to 0.0 for weight values, the speed increases in a linear way getting as high as almost 2.3, but after the 0.0 weight value threshold, speeds stay below 2.1 values with a minimum speed value of 1.6, indicating there might be a pattern between these two variables in which larger animals within its movement type tend to be faster than others, until a point in which heavier weights start decreasing the speed.

Analyzing the weight being grouped by movement type, it can be concluded that the heaviest group of animals are the swimming group, as well as the lowest ones, while the flying group are the lightest and the fastest. By these two analyses, it can be confirmed that larger animals are faster, but the heaviest animals are not the fastest ones (swimming group is the heaviest, yet not the fastest).

-Does the movement type and weight determine how fast an animal can go?

After performing a linear regression model on the dataset and doing the variables analysis, it can be concluded that the weight by itself has no influence with the speed, but adding the movement type variable helps fitting a more accurate linear model. Although species movement type and weight have a very small influence on the high speed. Weight and highspeed variables have almost a non-existent linear relationship, with a pearson correlation value of almost zero, which can be visually seen from the weight-highspeed scatter plots. While movement type, has a little more of influence within the species highspeed than the weight, since almost all groups have significant different speeds, it does not explain enough of the speed variance, hence a small coefficient of determination. By interpretating this coefficient, it can be said that almost 30% of the highspeed variance is explained by its weight and movement type. This means that other type of variables determines a species speed or at least have much more influence, such as the ecosystem in which they live and the role they play within it (some animals depend fully on their speed to survive, while others have other traits which help them for survival), or the internal bone structure (some types of bones structures could enhance speed and acceleration) or even an additional classification of light and heavy weights within each movement type group. Adding more variables which also have a strong linear relationship with highspeed may result in a more accurate linear regression model and a higher coefficient of determination. In this case, the dataset does not contain useful variables to perform an accurate linear model, therefore probably other type of non-linear models could work better with these variables.

Appendix

Figure 10 displays the tukey table for an ANOVA Test between weight and movement type where it shows which groups have a significant difference, by rejecting the null hypothesis of all means are equal, and which ones do not reject this hypothesis. In this case only climbing-flying and climbing-running have a similar mean keeping initial null hypothesis true, while the other group combinations reject the null hypothesis, stating a significant difference between those groups.

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
climbing	flying	-0.8264	0.2086	-1.9228	0.2699	False
climbing	running	0.4566	0.6158	-0.5163	1.4295	False
climbing	swimming	1.4929	0.0014	0.4592	2.5266	True
flying	running	1.2831	0.0001	0.5579	2.0082	True
flying	swimming	2.3193	0.0	1.5144	3.1242	True
running	swimming	1.0363	0.0002	0.4098	1.6627	True

Figure 10. Tukey table for weight vs movement type

Figure 11 displays the tukey table for an ANOVA Test between high-speed and movement type, where only climbing-running and climbing-swimming showed no difference between groups. While the rest of the groups rejected the null hypothesis by showing a significant difference.

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
climbing	flying	0.3636	0.0057	0.0809	0.6464	True
climbing	running	0.104	0.7041	-0.1469	0.355	False
climbing	swimming	-0.2058	0.1907	-0.4724	0.0608	False
flying	running	-0.2596	0.0024	-0.4466	-0.0726	True
flying	swimming	-0.5694	0.0	-0.777	-0.3618	True
running	swimming	-0.3098	0.0	-0.4714	-0.1483	True

Figure 11. Tukey table for high-speed vs movement type

Next set of figures display separately the weight grouped by its movement type scatter plot where a more detailed analysis can be made for weight vs high-speed in each different movement type.

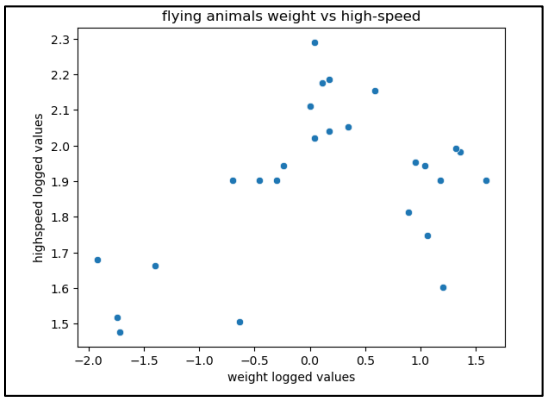

Figure 12. Scatter plot for flying animals of weight vs high-speed

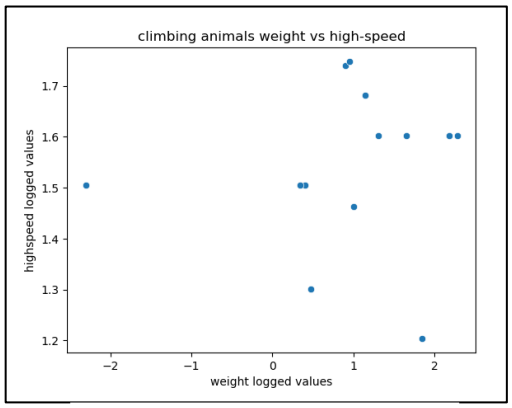

Figure 13. Scatter plot for climbing animals of weight vs high-speed

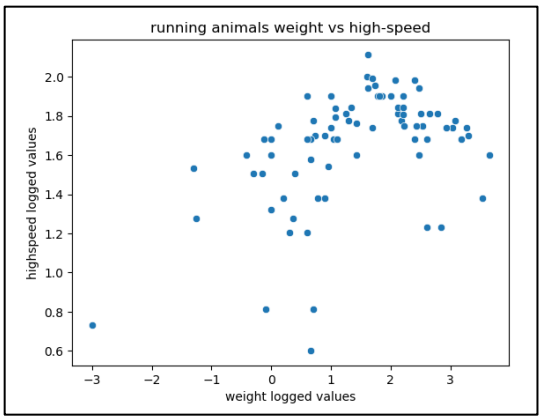

Figure 14. Scatter plot for running animals of weight vs high-speed

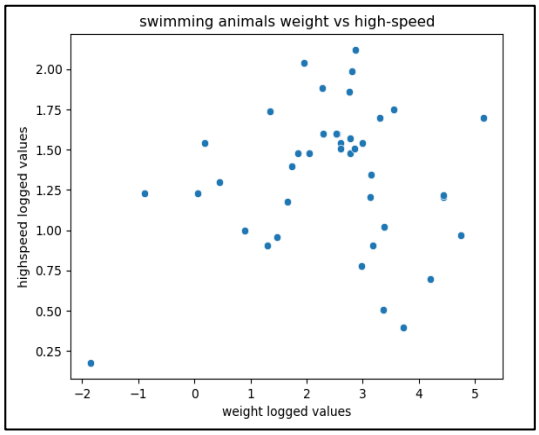

Figure 15. Scatter plot for swimming animals of weight vs high-speed

Bibliography

Deshpande, A., & Kumar, M. (2018). *Artificial Intelligence for Big Data: Complete guide to automating big data solutions using artificial intelligence techniques*. Packt Publishing.

Frost, J. (2020). *Introduction to statistics: An intuitive guide for analyzing data and unlocking discoveries*. Statistics by Jim Publishing.

Gawali, S. (2022, November 30). *Skewness and kurtosis: Shape of data: Skewness and kurtosis*. Analytics Vidhya. Retrieved January 2, 2023, from https://www.analyticsvidhya.com/blog/2021/05/shape-of-data-skewness-and-kurtosis/

Haider, M. (2016). *Getting started with data science: Making sense of data with analytics*. *Google*. IBM Press.

Histograms. Histograms - Understanding the properties of histograms, what they show, and when and how to use them | Laerd Statistics. (n.d.). Retrieved January 2, 2023, from https://statistics.laerd.com/statistical-guides/understanding-histograms.php

Mirtaheri, S. L. (2022). *Machine learning: Theory to applications*. CRC Press.

Seaborn.boxplot#. seaborn.boxplot - seaborn 0.12.2 documentation. (n.d.). Retrieved January 9, 2023, from https://seaborn.pydata.org/generated/seaborn.boxplot.html

Simkus, J. (2022). *What is ANOVA (Analysis Of Variance)*. Simply Psychology. www.simplypsychology.org/anova.html

Scatter plot. Corporate Finance Institute. (2022, December 19). Retrieved January 2, 2023, from https://corporatefinanceinstitute.com/resources/data-science/scatter-plot/

Van Rossum, G., & Drake Jr, F. L. (1995). *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam.