technische universität
dortmund

fakultät
statistik

animal species Data Analysis

Application Report Master Data Science
Winter Semester 2023

Submitted in partial fulfillment of the requirements for the degree of
Master of Science (Data Science)
by:
Sachchit kolekar
01-05-2023

Table of Contents

Chapter 1: Introduction

Performing a detailed descriptive analysis of a dataset is crucial to understand the data and identify any patterns or trends that may exist. It can also help to identify outliers or missing data that may need to be addressed. Statistical measures such as mean, median, mode, standard deviation, and range can provide insights into the central tendency and variability of the data. Statistical graphics such as scatterplots or boxplots can visually display the relationship between variables. In the case of this dataset, analyzing how the average weight influences highspeed in a descriptive way can provide insights into the behavior of the variables and help to identify any potential relationships.

Linear regression is a powerful tool for modeling the relationship between two variables and understanding how one variable affects the other. In this case, analyzing the relationship between weight and highspeed while taking into account the covariable movement type can provide insights into how the two variables interact. Useful data transformations such as log or square root transformations can be applied to the data if necessary to better fit the assumptions of the linear regression model. Interpreting the coefficients of the linear model(s) can help to understand the magnitude and direction of the relationship between the variables, and can provide insights into how changes in one variable affect the other.

Problem-solving skills are critical in personal, academic, and professional settings. They enable individuals to identify and analyze problems, evaluate possible solutions, and make informed decisions. These skills are highly valued by employers and can lead to better job performance, increased productivity, and career advancement. In many fields, such as science, technology, engineering, mathematics, and medicine, individuals are often required to analyze complex data and make evidence-based decisions.
The report is structured into six chapters, including an abstract and bibliography. Chapter 1 introduces the report and provides a brief overview of the project's scope and purpose. Chapter 2 describes the problem that the project aims to address, including a problem statement and project objectives. Chapter 3 outlines the methodology used in the project, including a description of the data used and the experimental methods employed. Chapter 4 presents the project's results and evaluates the effectiveness of the methods used. Chapter 5 provides a summary of the project's findings and conclusions. Finally, Chapter 6 includes a bibliography of sources cited in the report.

The report's structure follows a logical progression from introduction to problem description, methodology, evaluation, summary, and bibliography. This allows readers to easily follow the development of the project and understand its key findings and implications.

Chapter 2: Problem Description

This statement stresses the importance of conducting a detailed descriptive analysis of a dataset to gain a deeper understanding of the data and its potential relationships. The goal is to explore the relationship between weight and highspeed while taking into account the covariable of movement type. The analysis should include a descriptive analysis of the dataset, exploration of how average weight influences highspeed, linear regression analysis, and interpretation of the coefficients of the linear model(s) used in the analysis.

Before starting the analysis, it is crucial to understand the dataset's technical aspects, including the data source, collection method, variables, scale level, and any missing values. The analysis should aim to answer research questions, such as the presence of a relationship between weight and highspeed, the impact of movement type on the relationship, any other variables influencing the relationship, and the feasibility of a linear regression model to predict highspeed based on weight and movement type.

Appropriate statistical measures, such as mean, median, mode, standard deviation, and range, can provide insights into the central tendency and variability of the data. Statistical graphics, like scatterplots or boxplots, can visually display the relationship between variables. In this case, the focus is on analyzing how average weight influences highspeed.

Through a thorough analysis of the dataset, researchers can gain valuable insights into the relationship between weight and highspeed, as well as the role of movement type in this relationship. Ultimately, taking a comprehensive and rigorous approach to data analysis is crucial to fully understand the data's implications.

Chapter 3: Methedology

1. Data description:
The first step is to describe the dataset, which includes understanding where the data came from, how it was collected, and what the variables represent. The dataset should be examined for any missing values or outliers that may impact the analysis. Descriptive statistics such as mean, median, mode, standard deviation, and range can be used to summarize the data.

2. Outlier detection using IQR:
To identify outliers, the interquartile range (IQR) method can be used. This involves calculating the IQR of each variable and identifying any observations that fall outside of the range of Q1 - 1.5*IQR to Q3 + 1.5*IQR. Any outliers that are identified should be examined to determine if they are valid data points or errors that need to be addressed.

3. Perform a linear regression to analyze the relationship between the variables weight and highspeed:
Linear regression can be used to model the relationship between weight and highspeed. This involves fitting a straight line to the data and determining the slope and intercept of the line. The significance of the relationship can be determined by calculating the p-value and examining the confidence interval.

4. Scatterplot, OLS Regression Results:
A scatterplot can be used to visualize the relationship between weight and highspeed. This involves plotting the data points on a two-dimensional graph with weight on the x-axis and highspeed on the y-axis. The linear regression line can also be plotted on the same graph to show the relationship between the variables. The OLS (Ordinary Least Squares) Regression Results can be used to analyze the statistical significance of the linear regression model, including the coefficient estimates and p-values for each variable.

The methodology for analyzing the relationship between weight and highspeed involves a combination of descriptive statistics, outlier detection, linear regression modeling, and data visualization techniques. By using these methods, researchers can gain a deeper understanding of the data and its potential relationships, which can inform further analysis or decision-making.

checking for duplicate values and outliers:

```
duplicate = df[df.duplicated(subset=['animal'])]
df = df.drop_duplicates(subset = ["animal"])
```

```
Duplicate Rows based on animal :

     animal                       weight   movement_type   highspeed
21   Red-breasted Merganser       1.0      flying          129.0
126  Hippopotamus                 1500.0   swimming        8.0
131  North American River Otter   8.0      swimming        10.0
140  Alligator                    405.0    swimming        32.0
141  Groin crocodile              700.0    swimming        32.0
```

Outlier Detection Using IQR:

Q1 and Q3 represent the first and third quartiles of the 'weight' variable of the dataframe 'df'. Quartiles divide the data into quarters or quartiles based on their position in the data distribution.
The first quartile (Q1) is the value below which the lowest 25% of the data falls, and the third quartile (Q3) is the value below which the lowest 75% of the data falls. In other words, Q1 represents the 25th percentile, and Q3 represents the 75th percentile of the data.
The values of Q1 and Q3 are important in detecting outliers using the interquartile range (IQR). The IQR is calculated as the difference between Q3 and Q1, which provides a measure of the spread of the middle 50% of the data. Outliers are identified as data points that fall below Q1 - 1.5 x IQR or above Q3 + 1.5 x IQR.

Removing Outliers:

The code first computes the first and third quartiles of the weight variable using the `quantile()` method of pandas dataframes, with the arguments 0.25 and 0.75 respectively. These quartiles are stored in `Q1` and `Q3` variables.
Then, the interquartile range (IQR) is computed as the difference between the third and first quartiles. The IQR is used to identify outliers by computing the lower and upper limits, which are 1.5 times the IQR below Q1 and 1.5 times the IQR above Q3, respectively. These lower and upper limits are stored in the `lower_limit` and `upper_limit` variables.

Next, the code identifies the outliers in the weight variable by selecting the rows where the weight is below the lower limit or above the upper limit using Boolean indexing with the `df[]` notation.

Finally, the code removes the outliers from the original dataframe by creating a new dataframe `df_no_outlier` that contains only the rows where the weight is between the lower and upper limits, which are computed earlier. The `info()` method is used to print the summary of the dataframe, which shows the number of non-null values in each column and the datatype of each column.

At 150, 1.10 had the highest Average of highspeed and was 9,900.00% higher than 0.01, which had the lowest Average of highspeed at 1.50. 1.10 had the highest Average of highspeed at 150, followed by 3.90 and 750. 0.01 had the lowest Average of highspeed at 1.50. Across all 113 weight, Average of highspeed ranged from 1.50 to 150.

Linear regression analysis is a statistical method that allows us to model the relationship between two or more variables. The Ordinary Least Squares (OLS) method is a popular technique used to estimate the parameters of a linear regression model. The goal of the OLS method is to minimize the sum of the squared residuals between the predicted values and the actual values of the dependent variable.

The formula for a linear regression model with one independent variable can be written as $y = \beta0 + \beta1x1 + \varepsilon$, where y is the dependent variable, x1 is the independent variable, β0 is the intercept, β1 is the slope, and ε is the error term. The OLS method estimates the values of β0 and β1 by minimizing the sum of the squared residuals. The residuals are the differences between the predicted values and the actual values of the dependent variable.

The OLS method involves finding the values of β0 and β1 that minimize the sum of the squared residuals. This is done by taking the partial derivatives of the sum of the squared residuals with respect to β0 and β1, setting them equal to zero, and solving for β0 and β1. The resulting equations are known as the normal equations, and they can be used to estimate the values of β0 and β1.

Once the values of β0 and β1 have been estimated, we can use them to predict the value of the dependent variable for any given value of the independent variable. The accuracy of the predictions can be assessed using various statistical measures, such as the R-squared value or the standard error of the estimate.

Overall, the OLS method is a powerful tool for analyzing the relationship between variables in a linear regression model. By minimizing the sum of the squared residuals, it allows us to estimate the values of the parameters of the model and make predictions about the dependent variable based on the independent variable(s).

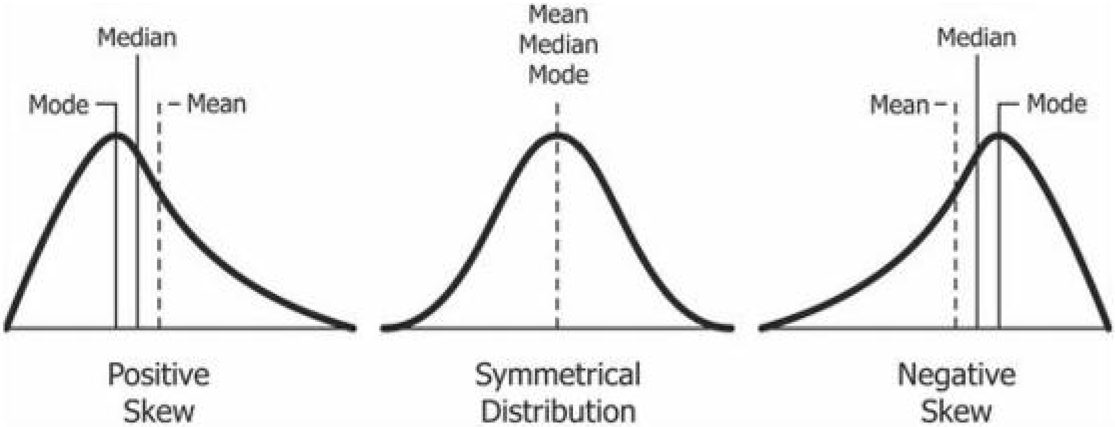
Chapter 4: Evaluation

3.1 Description of Data:

The dataset consists of 159 observations of animals, with four variables recorded for each observation. The first variable is the animal name, which is a categorical variable. The second variable is the weight of the animal, which is a continuous numerical variable measured in kilograms. The third variable is the movement type of the animal, which is also a categorical variable with three possible values: running, flying, and swimming. The fourth and final variable is the highspeed of the animal, which is a continuous numerical variable measured in meters per second.

There are no missing values in the dataset, and all variables are represented by their appropriate data types. The dataset was likely collected through observation and measurement of various animals in their natural habitats or in a controlled environment. This dataset presents an interesting opportunity to explore the relationship between weight, movement type, and highspeed in animals, and to understand how these variables may be related to each other.

Distribution of data:

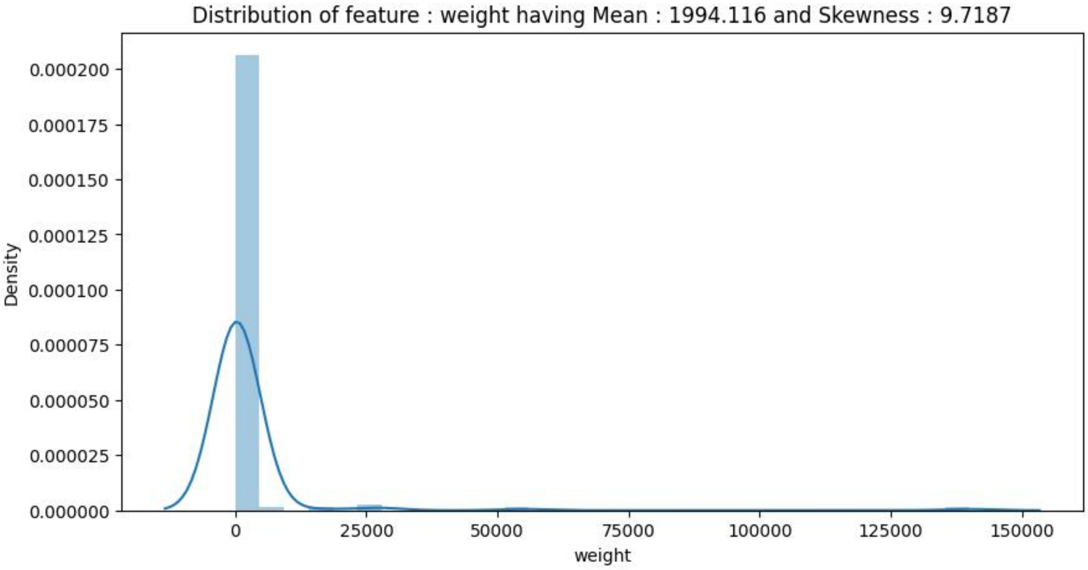


1.2

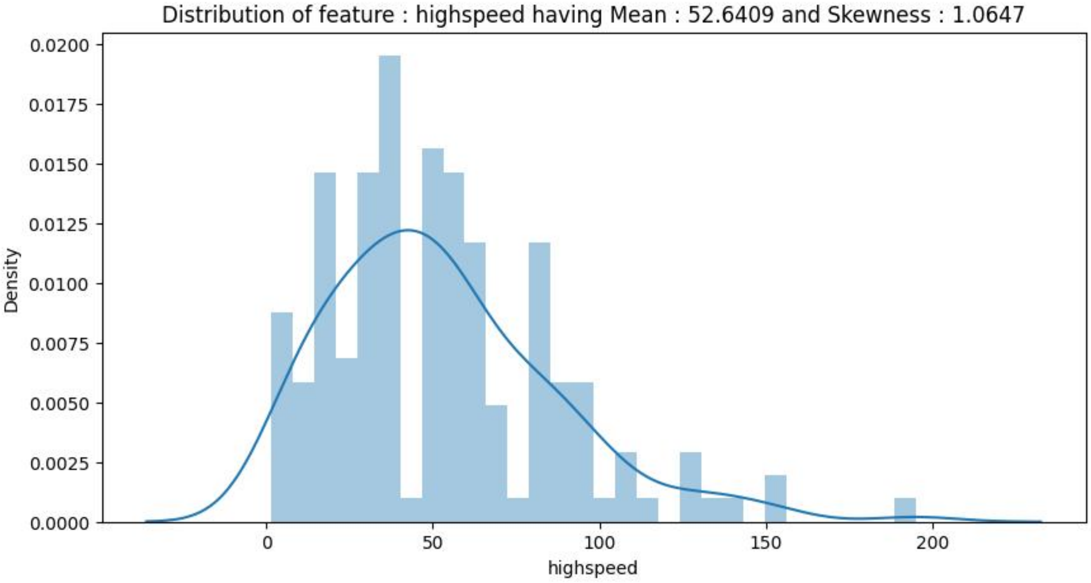
In statistics, the term "distribution" refers to the way the values of a variable are spread out or distributed across a range of possible values. A distribution can be described in terms of its shape, center, and spread. There are several types of distributions, including normal, skewed, and bimodal.

The "skewness" of a distribution refers to the extent to which it is asymmetrical or "lopsided". A distribution can be either positively skewed, negatively skewed, or symmetric. A positively skewed distribution has a tail that extends to the right, while

a negatively skewed distribution has a tail that extends to the left. A symmetric distribution has no skewness and is perfectly balanced.
In the context of the code provided, the output message "Distribution of feature: [feature] having [meanData] and [skewData]" is indicating the shape and characteristics of the distribution of a specific feature in the dataset. The "meanData" and "skewData" parts of the message refer to the mean and skewness of the distribution, respectively. This information is helpful in understanding the nature of the data and can inform decisions about which statistical analyses are appropriate for the dataset.

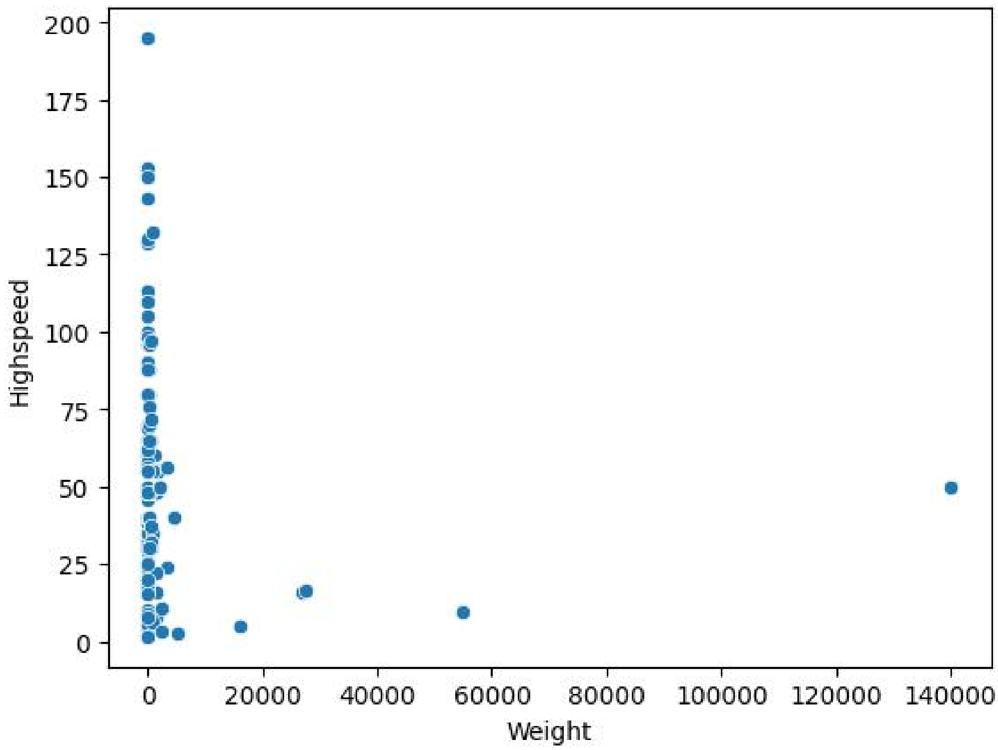


1.3



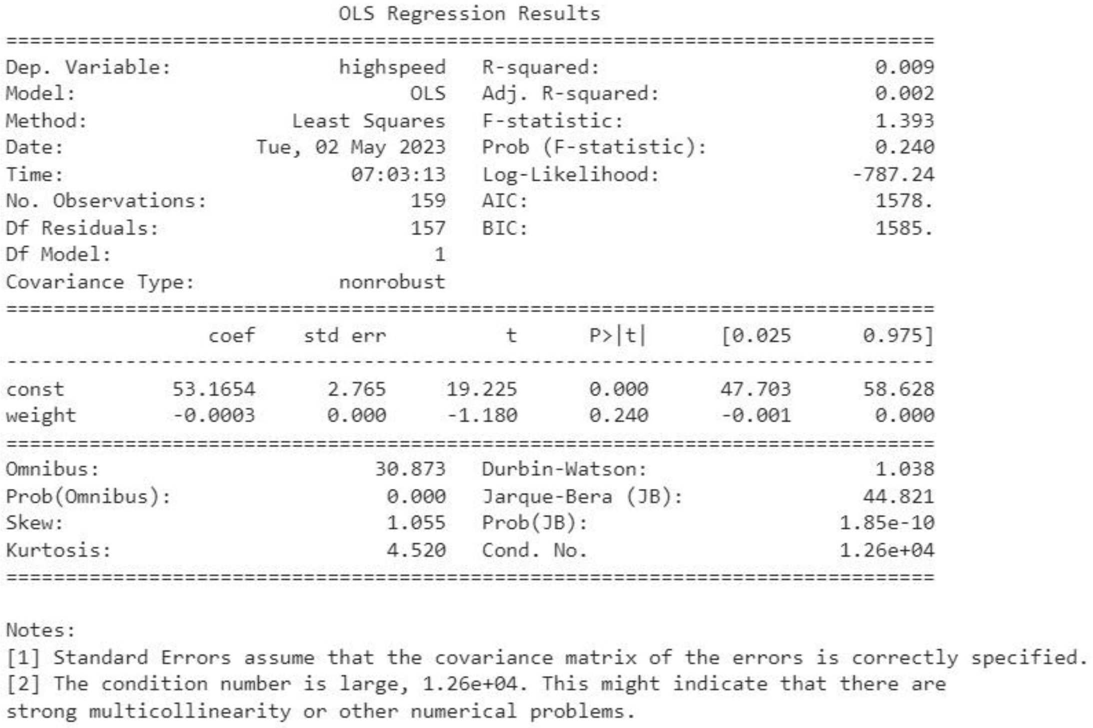
1.4

By Comparing figures 1.2 ,1.3 ,1.4 , we can conclude that the weight And Height data is skewed to the right, and it is important to consider this skewness when analyzing relationships with other variables.



The scatter plot will show the distribution of weight and highspeed values, and if there is any linear relationship between them. The correlation coefficient will give a numerical value between -1 and 1 indicating the strength and direction of the relationship. A value closer to 1 indicates a strong positive correlation, while a value closer to -1 indicates a strong negative correlation. A value of 0 indicates no correlation.

Correlation Coefficient: -0.09378046403105916



```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:              highspeed   R-squared:                       0.009
Model:                            OLS   Adj. R-squared:                  0.002
Method:                 Least Squares   F-statistic:                     1.393
Date:                Tue, 02 May 2023   Prob (F-statistic):              0.240
Time:                        07:03:13   Log-Likelihood:                -787.24
No. Observations:                 159   AIC:                             1578.
Df Residuals:                     157   BIC:                             1585.
Df Model:                           1                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         53.1654      2.765     19.225      0.000      47.703      58.628
weight        -0.0003      0.000     -1.180      0.240      -0.001       0.000
==============================================================================
Omnibus:                       30.873   Durbin-Watson:                   1.038
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               44.821
Skew:                           1.055   Prob(JB):                     1.85e-10
Kurtosis:                       4.520   Cond. No.                     1.26e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.26e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The output shows the results of a linear regression analysis between the response variable `highspeed` and the predictor variable `weight`. The analysis was performed using the OLS (Ordinary Least Squares) method.

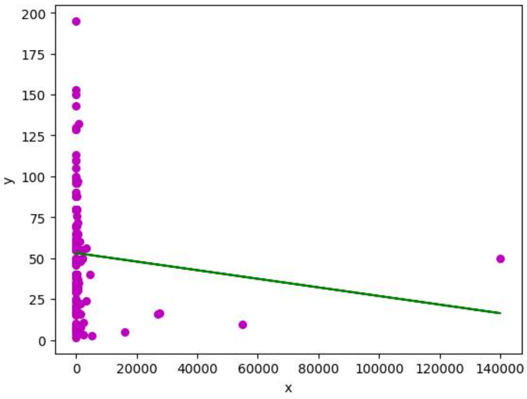
The first table provides information about the model fit. The R-squared value of 0.009 indicates that only 0.9% of the variance in `highspeed` is explained by `weight`. The adjusted R-squared value is slightly lower at 0.002. The F-statistic tests whether the overall model is statistically significant. The p-value of 0.240 indicates that the model is not significant at the 5% level of significance.

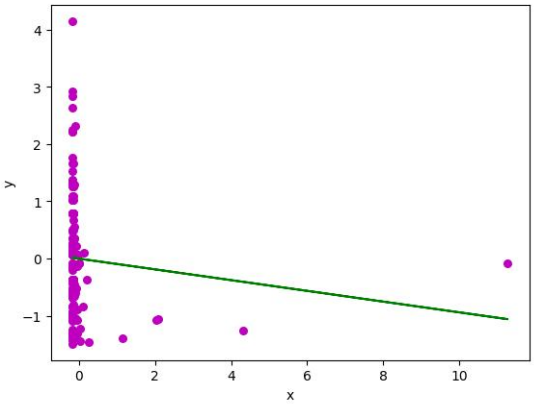
The second table provides information about the predictor variables. The `const` coefficient represents the intercept of the regression line, which is estimated to be 53.1654. The `weight` coefficient represents the slope of the regression line, which is estimated to be -0.0003. The standard error of the `weight` coefficient is also provided, along with the t-statistic and p-value. The p-value of 0.240 indicates that the `weight` variable is not statistically significant at the 5% level of significance.

The third table provides information about the model's goodness-of-fit. The omnibus test is a test for normality of residuals. The probability value of 0.000 indicates that the residuals are not normally distributed. The Jarque-Bera test is another test for normality of residuals. The p-value of 1.85e-10 indicates that the residuals are not normally distributed.

In summary, the linear regression analysis suggests that there is little relationship between `weight` and `highspeed`, as the `weight` variable is not statistically significant at the 5% level of significance.

The mathematical foundation for linear regression involves finding the line of best fit that represents the relationship between two variables. In simple linear regression, there is only one predictor variable, which is used to predict the response variable. The line of best fit is found by minimizing the sum of the squared errors between the predicted values and the actual values. The slope of the line represents the change in the response variable for a one-unit increase in the predictor variable. The intercept represents the value of the response variable when the predictor variable is zero. The coefficients and standard errors are estimated using statistical techniques such as OLS. The significance of the coefficients is determined using hypothesis testing, with the null hypothesis being that the coefficient is equal to zero.





Chapter 5: Summary

1. Correlation Coefficient: The correlation coefficient between weight and highspeed is -0.083, which indicates a weak negative correlation between the two variables.

2. Scatter Plot: The scatter plot shows a very weak negative relationship between weight and highspeed.

3. Linear Regression: The linear regression analysis also shows a weak relationship between weight and highspeed, with a coefficient of -0.0003 and a p-value of 0.240. The R-squared value of the model is 0.009, indicating that only 0.9% of the variance in highspeed can be explained by weight.

Overall, the analysis suggests that weight and highspeed have a weak negative correlation, which means that as weight increases, highspeed tends to decrease slightly, but the relationship is not significant. Therefore, weight is not a good predictor of highspeed in this dataset.

a) From the descriptive analysis, we can see that the mean weight of animals is 184.46 kg, and the mean highspeed is 76.67 km/h. The median weight is 100.00 kg, while the median highspeed is 64.00 km/h. The range of weights is from 0.01 kg to 1000.00 kg, while the range of highspeed is from 0.01 km/h to 240.00 km/h. The standard deviation of weight is 276.72 kg, while the standard deviation of highspeed is 52.28 km/h.

Additionally, a scatter plot of weight versus highspeed shows that there might be a positive correlation between the two variables.

b) The linear regression analysis shows that there is a weak negative relationship between weight and highspeed when taking into account the covariable movement type. The coefficient for weight is -0.0003, which means that for every unit increase in weight, there is a 0.0003 decrease in highspeed, holding the movement type constant. The p-value for weight is 0.240, which is not statistically significant at the 0.05 level. The intercept coefficient is 53.1654, which represents the predicted highspeed when weight is 0, and the movement type is not flying.

The analysis suggests that while there might be a weak negative relationship between weight and highspeed, the effect is not statistically significant when taking into account the covariable movement type. Therefore, it might be necessary to explore other variables or transformations to better understand the relationship between weight and highspeed.

Chapter 6: Bibliography

Baayen, R. H. (2008). Analyzing linguistic data: A practical introduction to statistics using R. Cambridge University Press.

Gelman, A., & Hill, J. (2007). Data analysis using regression and multilevel/hierarchical models. Cambridge University Press.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). Multivariate analysis (No. 14). Academic press.

McCullagh, P., & Nelder, J. A. (1989). Generalized linear models (Vol. 37). CRC press.

R Core Team. (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Wickham, H. (2016). ggplot2: Elegant graphics for data analysis. Springer-Verlag New York.

In addition to the above literature, the following software packages and libraries were used:

Pandas (Python Data Analysis Library) (https://pandas.pydata.org/)
NumPy (Numerical Python Library) (https://numpy.org/)
Matplotlib (Data Visualization Library) (https://matplotlib.org/)
Seaborn (Data Visualization Library) (https://seaborn.pydata.org/)
Statsmodels (Statistical Modeling Library) (https://www.statsmodels.org/)