University of Malta

Master of Science in Blockchain and Distributed Ledger Technologies

DLT5002 - Research Methods

Assignment
Statistics Component

Date Submitted: TBA

Group: Marica Ciantar, Jacques Vella Critien, Julian Sciberras, Adzhar Tarakchiev, Karsten Guenther

Lecturer: Dr. Fiona Sammut

Table of Contents

Section 1 - Introduction

1.1 Why this data set?

Humans, as part of the animal kingdom, have to sleep in order to maintain the normal functioning of the body. All mammals have to sleep in some part of their daily life. As something as universal as sleep appears to be, it is actually a mysterious phenomenon. Scientists and researchers have proven the benefits for overall health and well-being, but the sleeping mechanisms and patterns vary greatly among different mammal species. In this research document, we will delve into the world of mammalian sleep using statistical analysis. We will use a combination of descriptive statistics and hypothesis testing in order to gain a general understanding and investigate the relationship between sleep and various factors such as danger, exposure, lifespan, gestation and dreaming. Throughout our analysis, we are going to gain a deeper insight into the unique ways in which sleep is regulated and expressed in the animal kingdom. All mammals, including humans, require sleep as a necessary physiological mechanism for maintaining their health and well-being. It is crucial for memory consolidation, cognitive function, and general physical and mental wellness. In the animal kingdom, different species' sleeping habits vary widely. Some creatures sleep for extended periods of time, while others only sleep for brief intervals. The main factor that influences how mammals sleep is predation. Predators are a constant threat to animals in high predation areas, thus they must be vigilant to defend themselves. These animals typically sleep less, and their naps are shorter and more sporadic. This is referred to as alertness generated by predators.

1.2 Variables

The variables being investigated for this assignment—danger, exposure, prediction, life span, gestation, and dreaming—are all in some way connected to sleep. For instance, risk and exposure have a connection to predation and can affect the quantity and consistency of sleep. Sleep patterns may also be impacted by gestational age and lifespan. Additionally, both cognitive performance and the quality of sleep are related to dreaming.

The statistical information that will be presented in this document has been conducted by a powerful software tool for statistical analysis called SPSS (Statistical Package for Social Sciences). Additionally, SPSS provides a variety of data visualization tools such as charts, graphs and maps that we have used in order to provide the statistical analysis information in this research study. The document will use exploratory data analysis such as frequency tables and dispersion of values. Furthermore, the assignment will analyse skewness and kurtosis which are the statistical mechanism which is helpful to digest the data set. The assignment will go over some graphical representations like bar graphs, histograms, pie charts and box and scatter plots. In this assignment, we will also use Binary Logistics regression analysis which allows us to view the relationship between a binary response variable and one or more

predictor variables. The research will benefit from this tool because it will give us a better understanding of how different factors affect sleep in mammals.

1.3 Aims and Objectives

The main aim is to investigate the relationship between sleep patterns in mammals and various factors mentioned in this document. The study aims to gain a deeper understanding of how these factors may influence sleep patterns and the survivability of the animal. The objectives of this study are to use descriptive statistics that review data on sleep in different mammal species and to use hypothesis testing to investigate the specific relationships between variables. Using regression coefficients and odds ratios, the paper will examine the relative importance of different predictors. This document will also examine the relationship between sleep, brain structure, body width and gestation and their effects on the lifespan of the mammals. This will give good general insight of the corelation between these variables and the aim of the study. Finally, use visualisation tools to present the results of this research in a clear and concise manner.

This data set was obtained from this link, http://lib.stat.cmu.edu/datasets/sleep, which was offered by the guidelines for this assignment.

Section 2 - Exploratory Data Analysis

In general, the analysis of a data set is viewed as aided by the use of descriptive statistics and graphical representations. Inherently, exploratory descriptive statistics support the researchers to illustrate concise and clear visualizations of data sets, based on the main characteristics of the different variables. For the purpose of this study, a number of descriptive statistics techniques were identified and incorporated into our analysis.

Frequency Tables

Through the use of frequency tables, our analysis shed insights on a variety of information, starting with the measures of location – with the most popular being the **mean** which is "a measure of central tendency [which] tried to locate the center of the data.". A complementary data found to come out of the use of frequency tables was the **median** which is the "middle observation (middle score)" and splits the data into the lower quartile, which is the 25^{th} percentile [Q_1], the median [Q_2] which is the 50^{th} percentile, and the upper quartile [Q_3] which is the 75^{th} percentile. The final measure of location is the **mode** which shows the observation which occurs most frequently.

The **dispersion of the values** was another variable studied for this assignment. This is the "extent to which the given data is spread apart; how widely scattered are the observations around the mean.". Measures of dispersion include the **range**, which is the difference between the largest and smallest observation, the **interquartile range** which is the difference between the upper and lower quartile and does not make use of the extreme values of data. Other interesting measures of dispersion are noted by the **variance** and **standard deviation** which may be viewed as "the average amount by which observations in a distribution differ from the mean". If the data point lies close to the mean, the standard deviation is small and if it is spread out over a large range of values it will be large.

Skewness and Kurtosis

Furthermore, our analysis shall also view the **skewness** which shows the asymmetry of the distribution of a set of data around its mean, and the **kurtosis** which is the measure of the tails of the distribution of a data set vis-à-vis the normal distribution. When there is an evenly balanced distribution, there is zero skewness whilst if the distribution is skewed to the left it is negatively skewed. Similarly, a positively skewed result may be expected if the distribution is observed to the right thus showing a greater number of smaller values. From a kurtosis perspective, a zero kurtosis result is expected to be indicated by a shape close to the normal distribution whilst a positive value (leptokurtosis) is characterized by heavier tails thus more observations in the extremes of the distribution. On the other hand, a negative value (platykurtosis) shows lighter tails than a normal distribution. An extremely positive kurtosis

is characterized by a value greater than 5 and results when most values are in the tails rather than the mean.

Graphical Representations

Graphical representations are integral to supporting the analysis of the data set used for this assignment. In this regard, **bar graphs** provide a representation of variables with each column showing a different level of the variable while the height shows the frequency thereof. Bar graphs can be simple hence showing only one variable, clustered thus showing two variables by grouping the bars and stacked which is another way of showing two variables.

Histograms, unlike bar charts, show the range of variable values and are multimodal when they have multiple humps, uniform when without humps, unimodal when with only one hump or bimodal with two humps. When histograms have no symmetricity, the histograms will be considered skewed. When we compare the histogram with the normal curve we will be able to envisage the skewness and kurtosis albeit a normal curve does not show such.

We also used **pie charts**, the common ones as the ones which had a shadow or were in 3d were not presenting the data in the most legible manner. Pie charts have slices, each corresponding to a different level of the variable, and the portion of each slice represents the frequency of the level.

Box plots, which also will be used, give a graphical illustration for "median, upper and lower quartiles, minimum and maximum data values [and] any outliers present in the data and it can give an insight into the spread of the data."

Finally, **scatter plots** are important to compare the "relationships between two covariates …[and] reveal outliers and unusual combinations of values in numerical data." Should it appear that there is a linear relationship between the salary variables, a regression line or line of best fit may be used to illustrate this.

The following is the set of exploratory descriptive statistics conducted on the Sleep data set. Each representation is accompanied by a brief discussion on the interpretation of the visualization.

Danger when Sleeping

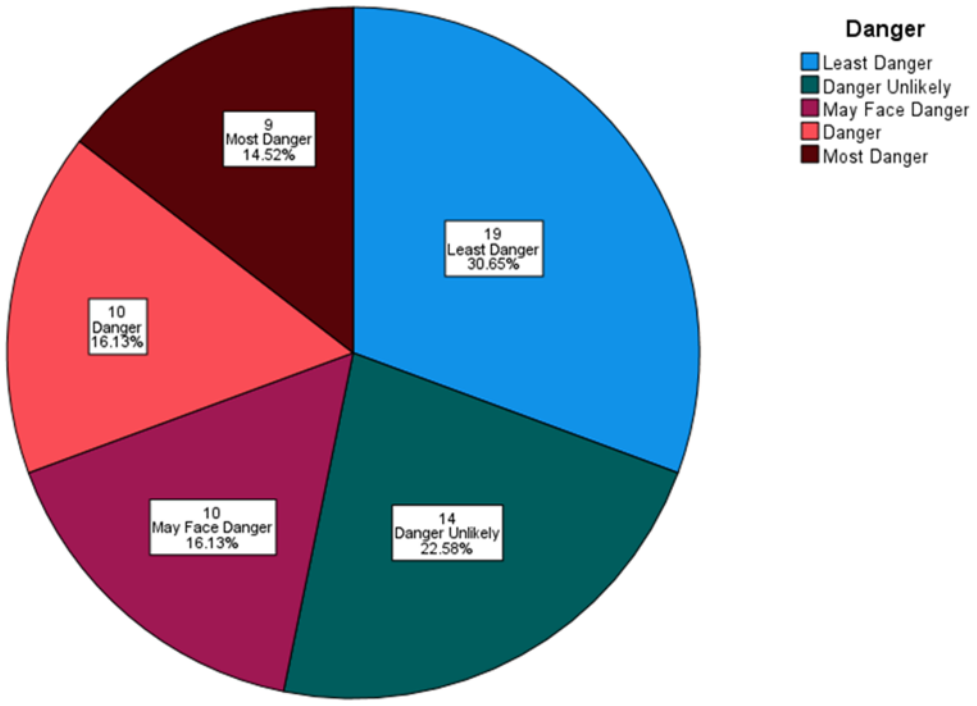


Figure 2.1

Figure 2.1 depicts the numbers of animals in the danger index, the name of the category, and the percentage of animals therein. It is apparent that there is a correlation between the number of animals and the danger faced whilst sleeping, with most animals preferring to sleep in the least dangerous areas and the least number of animals preferring to sleep in the most dangerous areas.

Danger		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Least Danger	19	30.6	30.6	30.6
	Danger Unlikely	14	22.6	22.6	53.2

May Face Danger	10	16.1	16.1	69.4
Danger	10	16.1	16.1	85.5
Most Danger	9	14.5	14.5	100.0
Total	62	100.0	100.0	

Figure 2.2

Figure 2.2 gives a similar result as we see in the pie chart (Figure 2.1), however, it also shows that we have no missing results hence, the pie chart is giving us a complete overview of the danger the animals face.

Predation Index

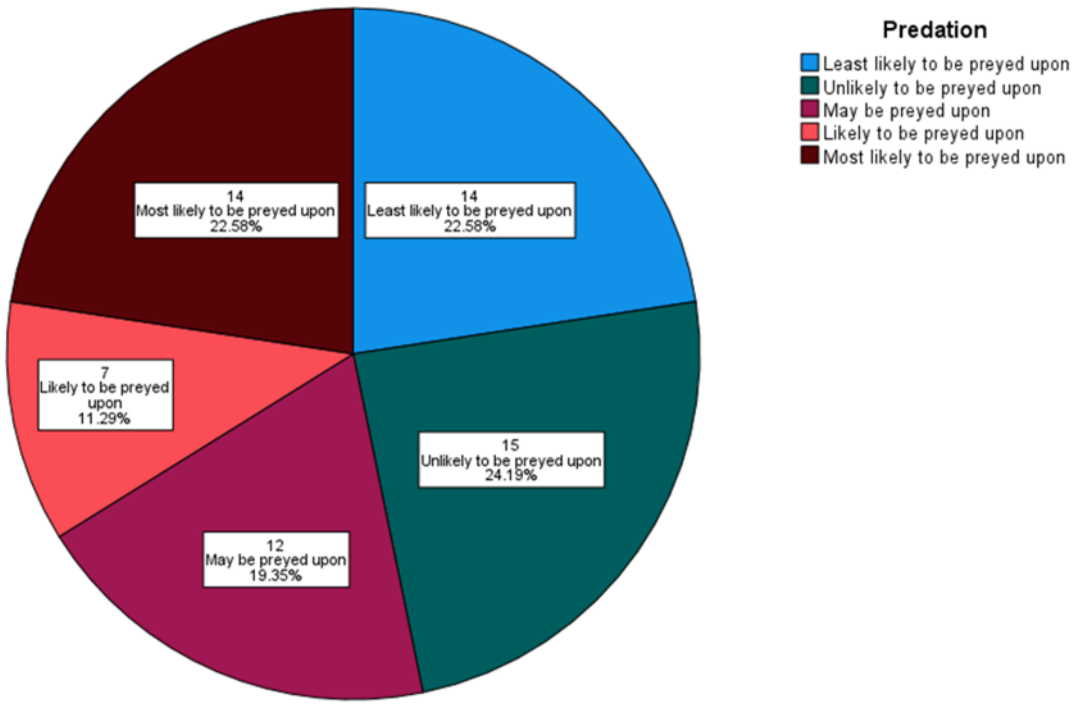


Figure 2.3

Figure 2.3 depicts the number of animals in the predation index, the name of the category, and the percentage of animals therein. It is interesting to note that the number of animals which are least and most likely to be preyed upon is the same with most animals being unlikely to be preyed upon.

Predation		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Least likely to be preyed upon	14	22.6	22.6	22.6
	Unlikely to be preyed upon	15	24.2	24.2	46.8
	May be preyed upon	12	19.4	19.4	66.1
	Likely to be preyed upon	7	11.3	11.3	77.4
	Most likely to be preyed upon	14	22.6	22.6	100.0
	Total	62	100.0	100.0	

Figure 2.4

Again, the frequency table (Figure 2.4) shows how there are no missing values in the graph thus making the pie chart accurate. It is interesting to note that most values, save for the 'likely to be preyed upon' one are quite similar in number.

Sleep Exposure

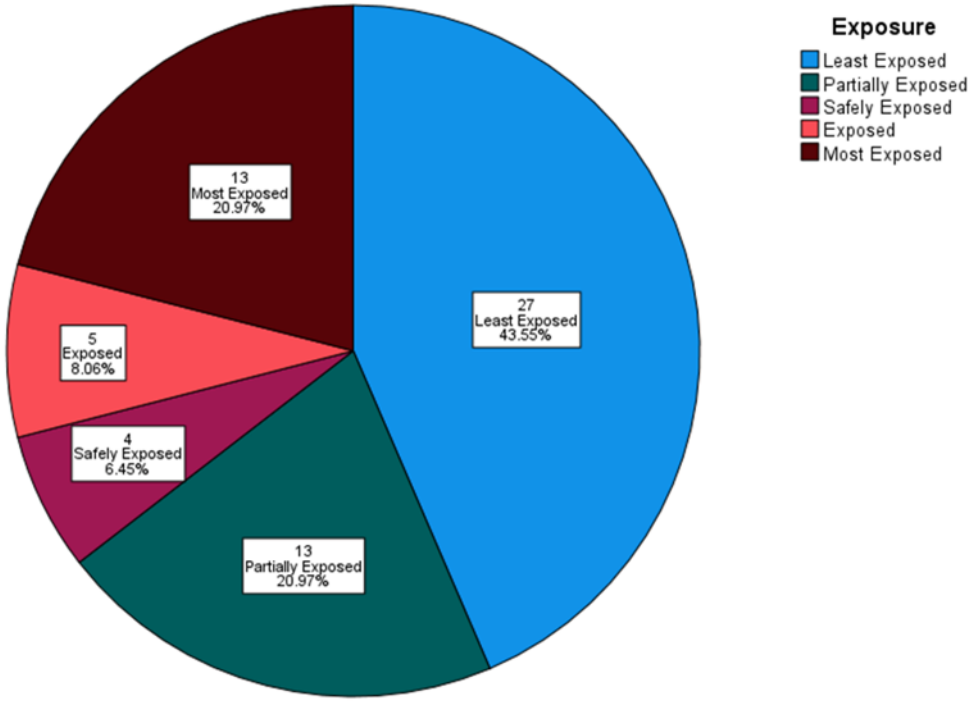


Figure 2.5

Figure 2.5 depicts the numbers of animals and their sleep exposure category, the name of the category, and the percentage of animals therein. Most animals are not exposed when sleeping whilst the same amount of animals sleeps in a ‘partially exposed’ manner and in a ‘most exposed’ manner. The remaining values, hence ‘exposed’ and ‘safely exposed’ are quite similar in number.

Exposure		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Least Exposed	27	43.5	43.5	43.5

Partially Exposed	13	21.0	21.0	64.5
Safely Exposed	4	6.5	6.5	71.0
Exposed	5	8.1	8.1	79.0
Most Exposed	13	21.0	21.0	100.0
Total	62	100.0	100.0	

Figure 2.6

Exposure, Predation and Danger

The hereunder pie chart (Figure 2.7) and clustered bar graph (Figure 2.7.1) shows how predation and exposure correspond. As one may expect, the most exposed species is the most likely to be preyed upon and there are no/little results showing that they will not be likely to be preyed upon.

In the pie chart within the 'Exposed' row, the highest predation index is 4 whilst 'least likely' and 'most likely' are equal.

In the 'Safely Exposed' pie chart, the animals experience the same amount of predation, hence, there is equal distribution throughout with 25% of the species in each slice.

For those species which are 'Partially Exposed', most animals are 'unlikely to be preyed upon' whilst the other predation values are all equal. The 'Least Exposed' cohort of animals are largely the 'least likely' to be preyed upon, followed by 2, 3, 4, and 5. The results make sense considering that the more exposure an animal has the more he is preyed upon by other animals, however, it is still important to take other factors into consideration that have not been included in the data set, such as location, time, the density of habitat and so on.

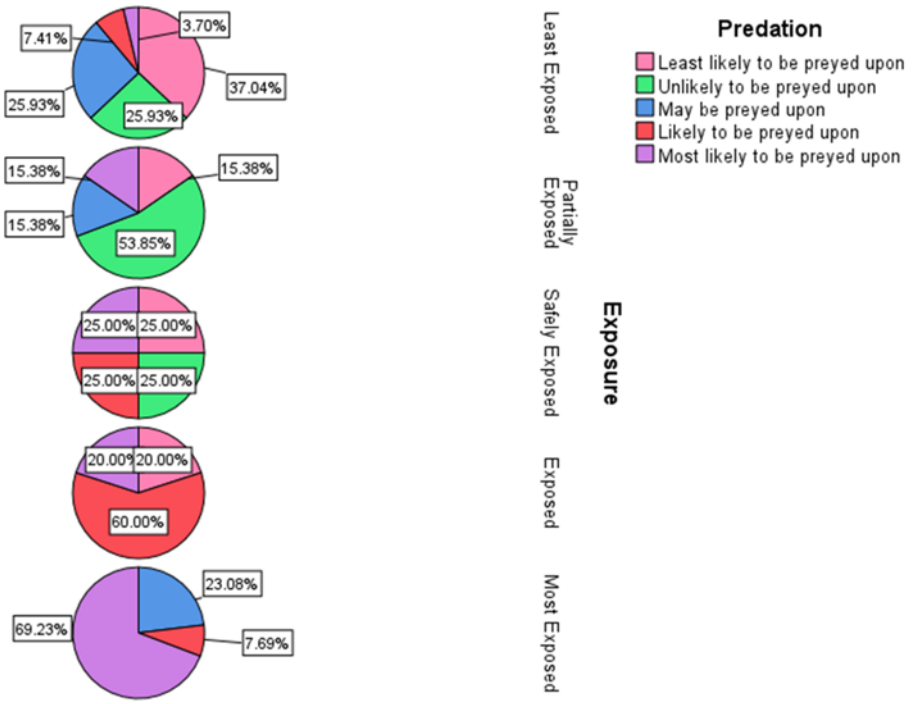


Figure 2.7

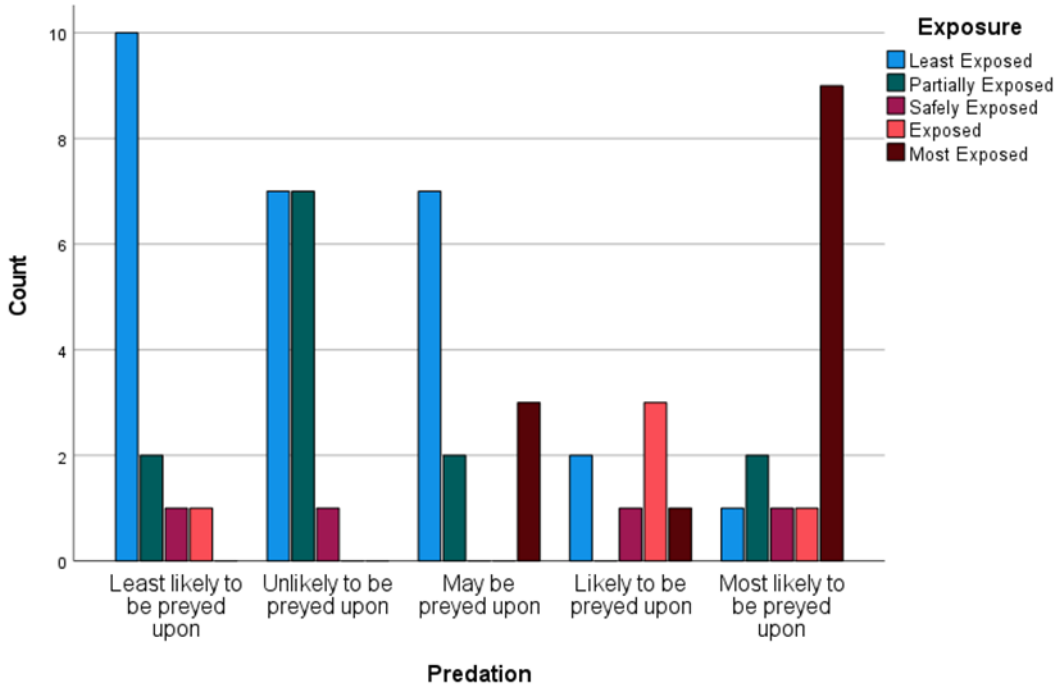


Figure 2.7.1

In a similar graphical representation to the above two figures, figures 2.8 and 2.8.2, show how danger and exposure correspond. There is proportionality between the least to most dangerous and the least to most exposed thus, again, considering the data we have available, the results do make sense.

The species which are ‘Least Exposed’ do not face any danger whereas the species which are most exposed have only recorded ‘Most Danger’, ‘Danger’, and ‘May Face Danger’.

On the other hand, the cohort for ‘Exposed’, they are largely in danger with an amount facing the least danger.

For those who are ‘Safely Exposed’ however, half of them face danger whilst the remaining half is split ¼ into those which are unlikely to face danger and those who are least in danger – thus they have equal distribution.

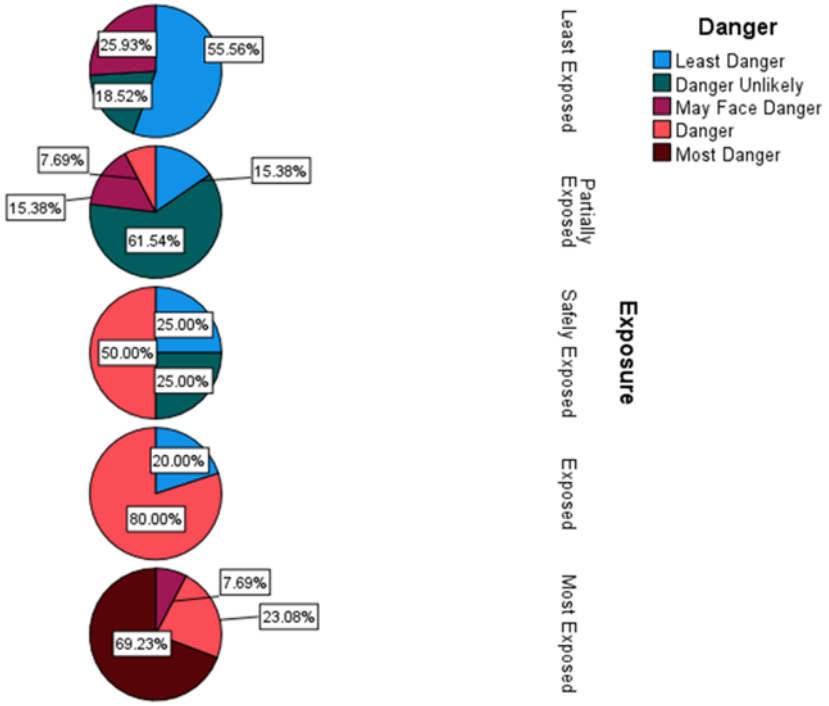


Figure 2.8

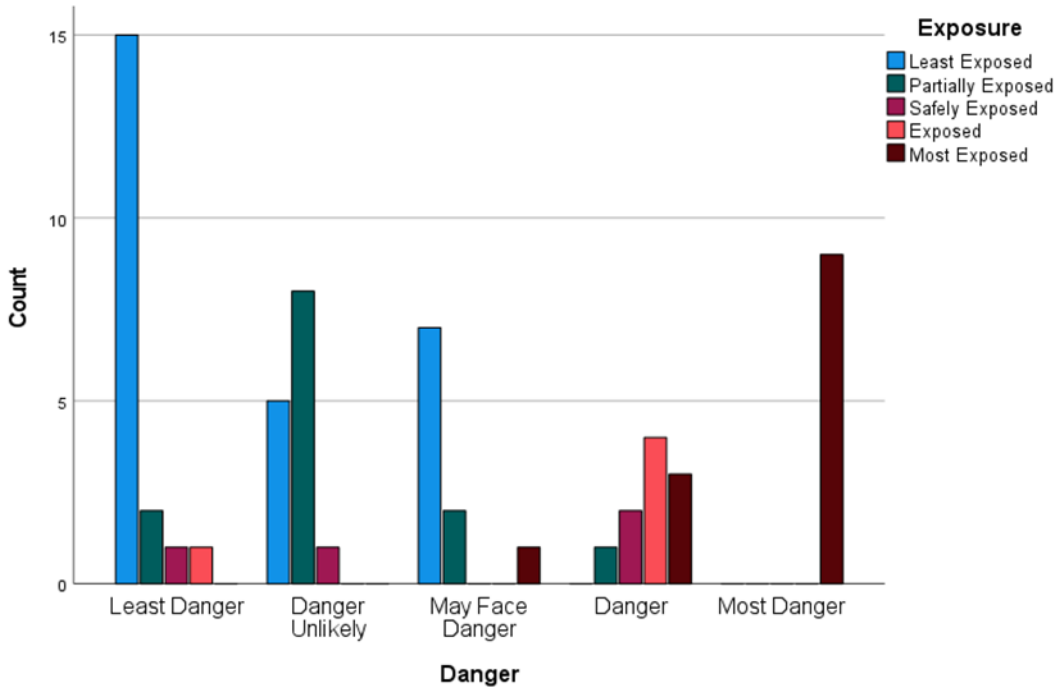


Figure 2.8.1

The following two figures, figure 2.9 and figure 2.9.1, show how danger and predation correspond. Again, there seems to be proportionality between the least to most dangerous and the least to most likely to be preyed upon, again, considering the data we have available, the results make sense.

'Least Danger' is observed to show those species are mainly 'Least likely to be preyed upon', followed by 'Unlikely to be preyed upon' and a small percentage of 'May be preyed upon'.

'Danger Unlikely' is observed to indicate species that are 'Unlikely to be preyed upon' however they 'May be preyed upon'. It is interesting to note that there are no 'Least likely to be preyed upon' species.

Species that may face danger mainly 'May be preyed upon' with equal distribution of being 'Likely to be preyed upon' and 'Most likely to be preyed upon'.

Naturally, species in the 'Most Danger' category are 100% most likely to be preyed upon.

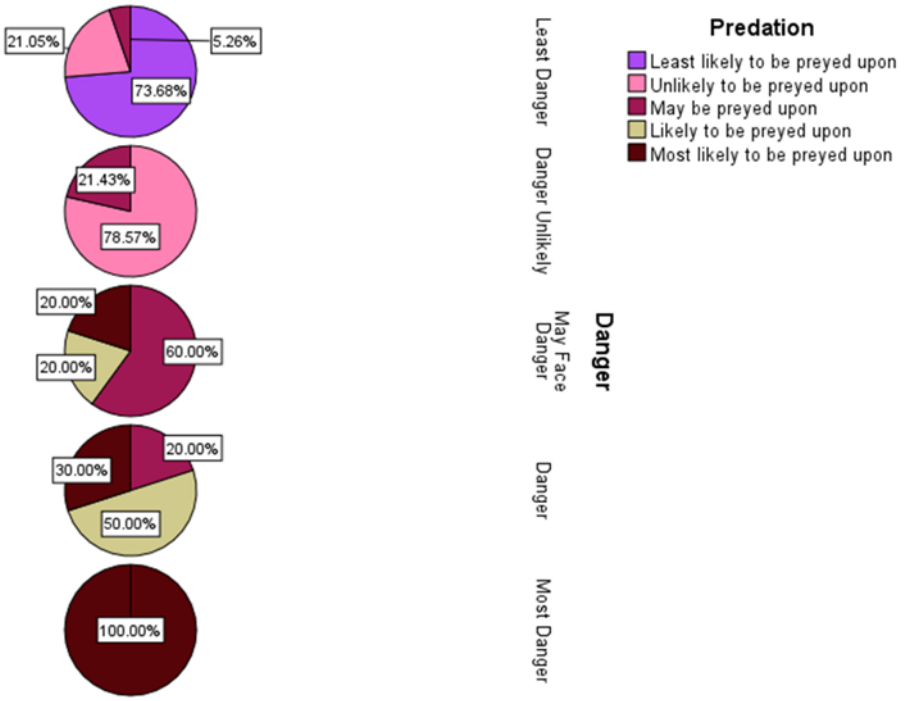


Figure 2.9

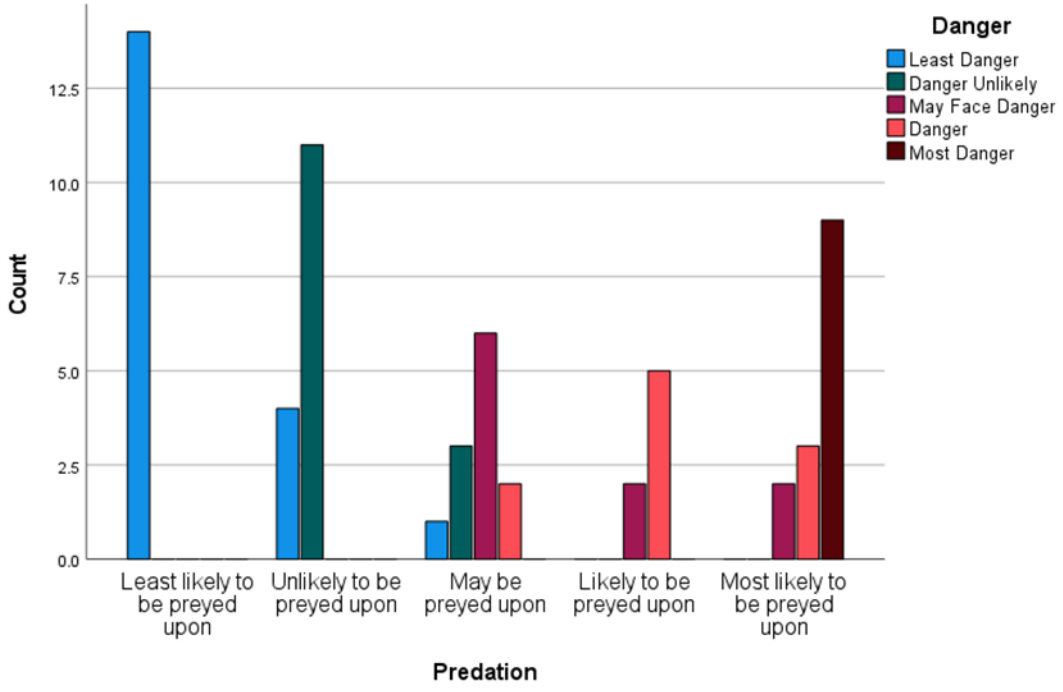


Figure 2.9.1

Non-Dreaming, Dreaming and Total Sleep

In this analysis we analyzed the variables of, dreaming, non-dreaming and total sleep in order to determine the correlation between them. According to the variables we have, when a species is sleeping it is either dreaming or non-dreaming hence it was appropriate to analyze the three variables together.

From the figures it can be seen that there are some missing values, 14 in the non-dreaming category, 12 in dreaming, and 4 in total sleep.

		Non_Dreaming	Dreaming	Total_Sleep
N	Valid	48	50	58
	Missing	14	12	4
Mean		8.673	1.972	10.533
Median		8.350	1.800	10.450
Mode		11.0	.5[a]	8.4[a]
Skewness		.298	1.453	.201
Kurtosis		-.236	2.320	-.508
Range		15.8	6.6	17.3

Minimum	2.1	.0	2.6
Maximum	17.9	6.6	19.9
Sum	416.3	98.6	610.9

a. Multiple modes exist. The smallest value is shown

Figure 2.10

The measures of location show that the values for non-dreaming and total sleep are quite similar most notably the mean, standard error thereof, median, mode and range. This is notwithstanding the fact that non-dreaming and total sleep have the most difference in their valid values. In dreaming and total sleep the figure shows that multiple modes exist and the smallest value is shown hence, this shows how the values are slightly different and are not repeated enough in the data set to take an accurate reading.

The species do not dream much, in fact there is a notable difference in the results of the measures of locations when it comes to dreaming which also has the highest number or skewness and kurtosis. Since the kurtosis for non-dreaming is negative it is expected that there will be a lighter tail for that than the other variables than those of normal distribution. On the other hand, being that skewness is positive, the distribution should be skewed to the right.

The standard deviation for non-dreaming and total sleep are similar, with the former being almost 3.7 and the latter almost 4.7 whilst for dreaming it is only almost 1.4. hence, the spread of data is larger for total sleep than the other two variables, with non-dreaming following closely behind.

The histograms, hereunder give the frequency of the species experiencing the states mentioned. Albeit they are already mentioned in a more detailed way in the above table having the histograms individual for each variable of non-dreaming, dreaming and total sleep.

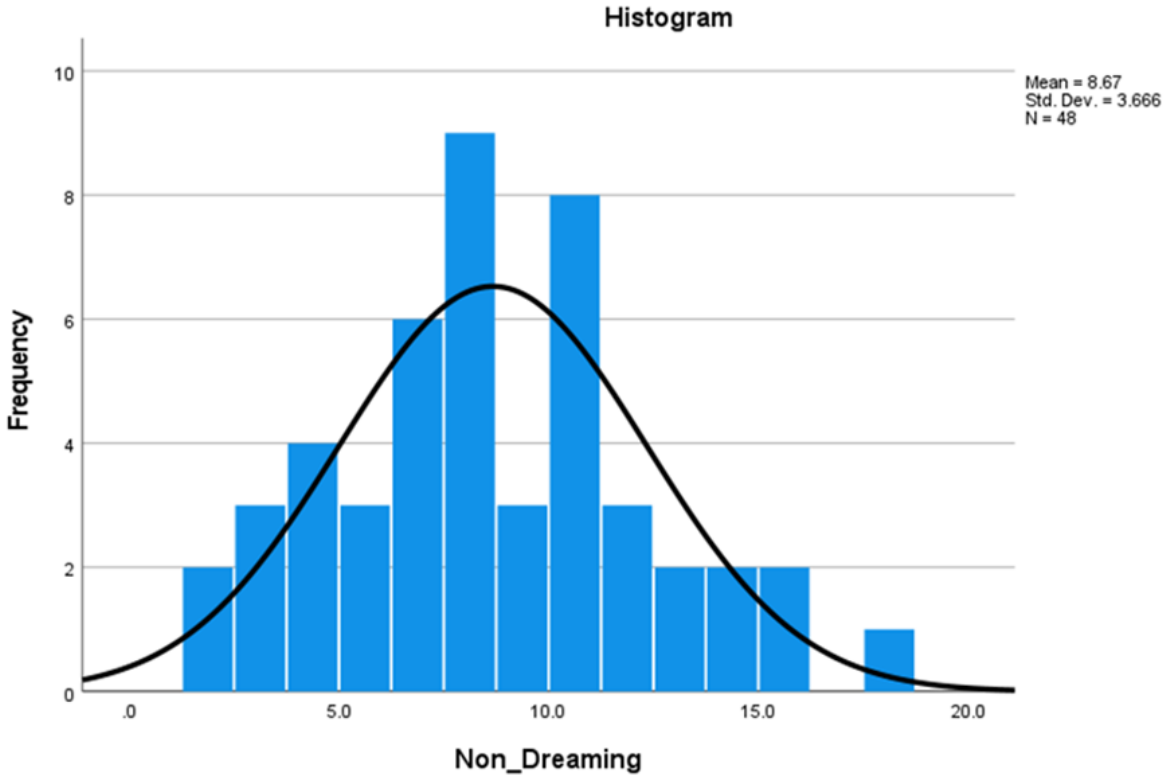

Histogram
Mean = 8.67
Std. Dev. = 3.666
N = 48
Frequency
10
8
6
4
2
0
.0
5.0
10.0
15.0
20.0
Non_Dreaming

Figure 2.11

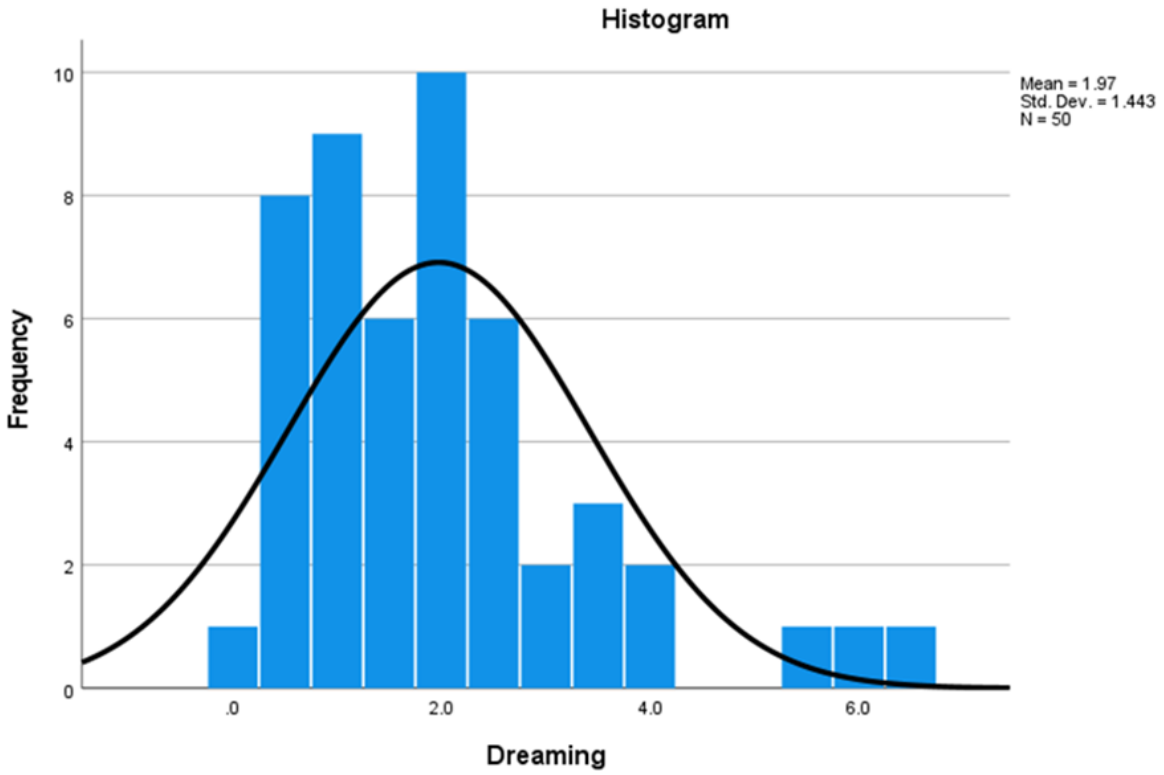


Figure 2.12

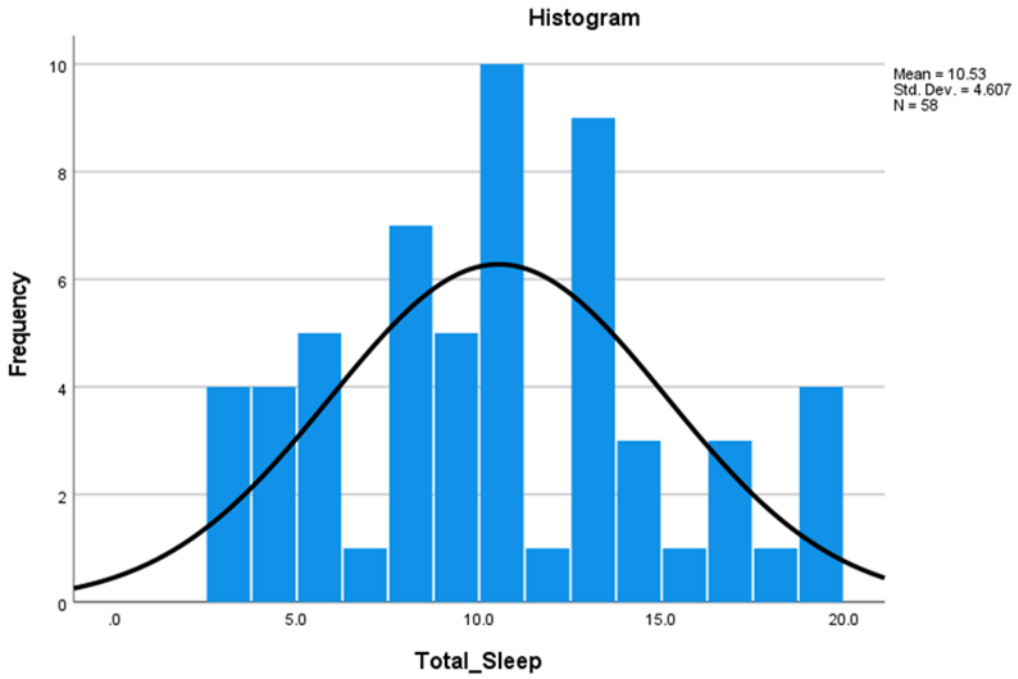


Figure 2.13

Scatter Plots

The scatter plots we will explore hereunder will serve to continue exploring the relationship between the variables.

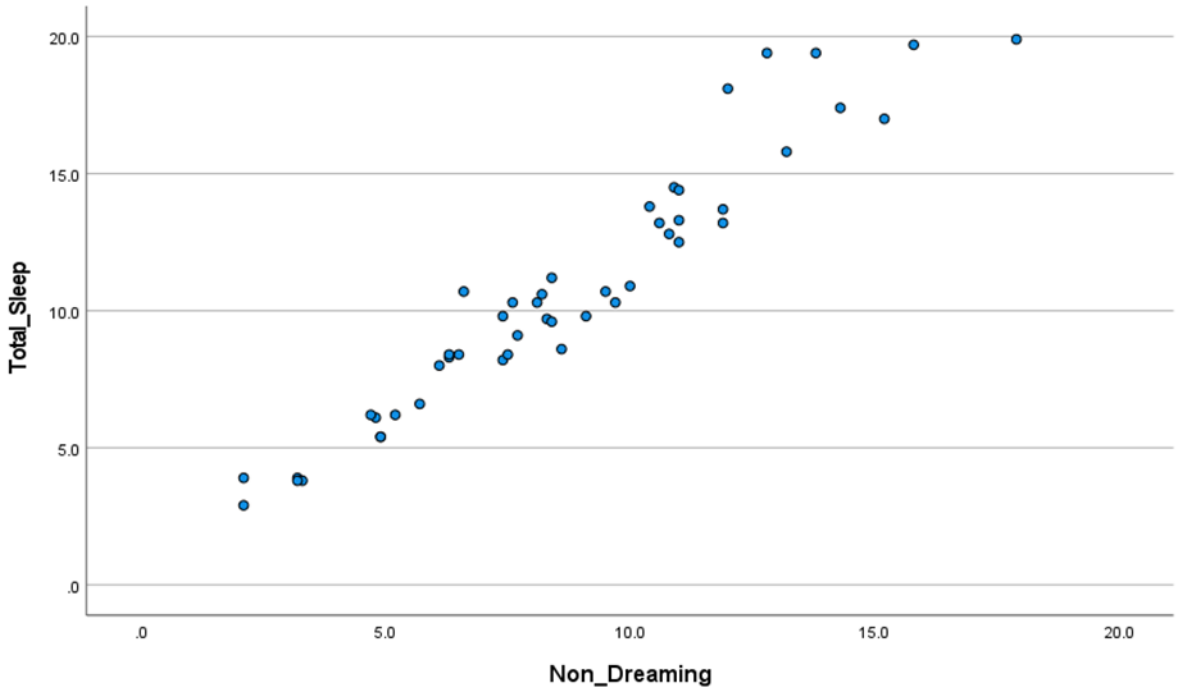


Figure 2.14

Figure 2.14 shows how there is a correspondence between the total hours of sleep and the hours of non-dreaming activity. We see that the more sleeping hours a species get the less it dreams hence, a higher amount of total hours of sleep show a high amount of non-dreaming. The animal with the most total hours of sleep and non-dreaming time records a total of around 20 hours of sleep and around 18 hours of non-dreaming time, hence it only has around 2 hours of dreaming time. On the other hand, there are two animals who have the same non-dreaming time out of 3 and 4 hours, respectively of total sleep time. According to the graph, these two animals have around half the time they are sleeping as dreaming and the other half as non-dreaming.

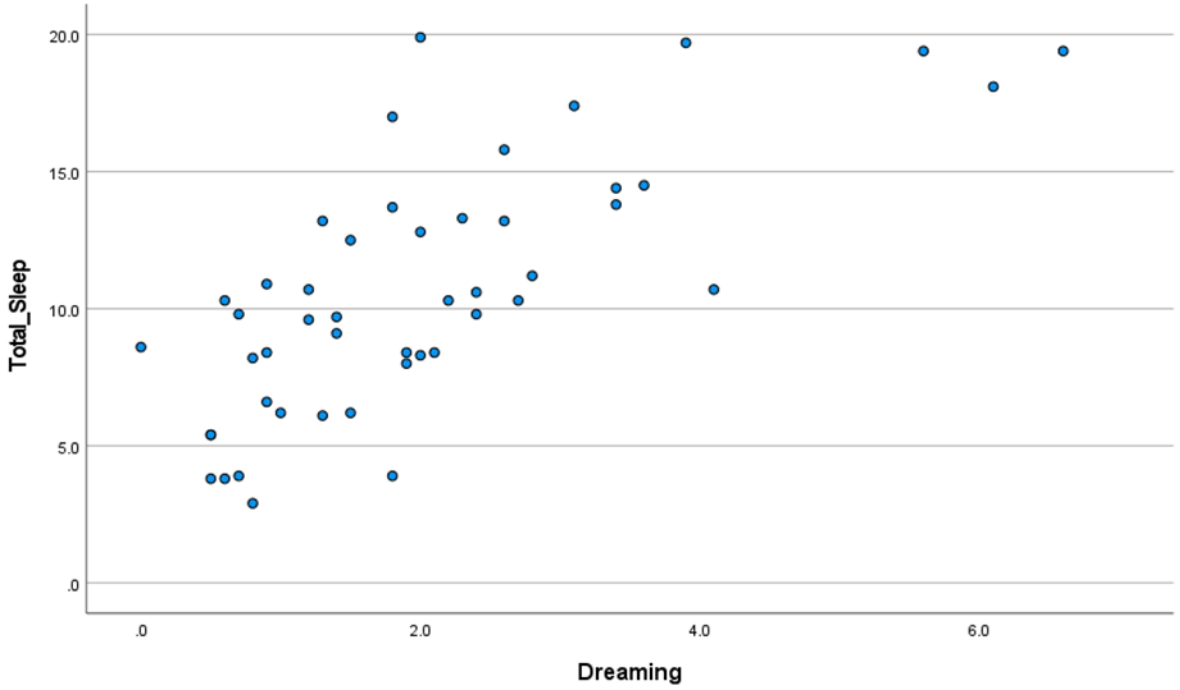


Figure 2.15

Unlike in figure 2.14, in Figure 2.15 there seems to be a bunch of species which are low on dreaming time regardless of the total hours of sleep, this makes sense when one view figure 2.14. there are 3 species which have a high number of sleeping hours (around 20 hours) and very high numbers of dreaming hours, around 6 hours. However the rest of the species have between 1 to 3 hours of dreaming hours, regardless of sleep time.

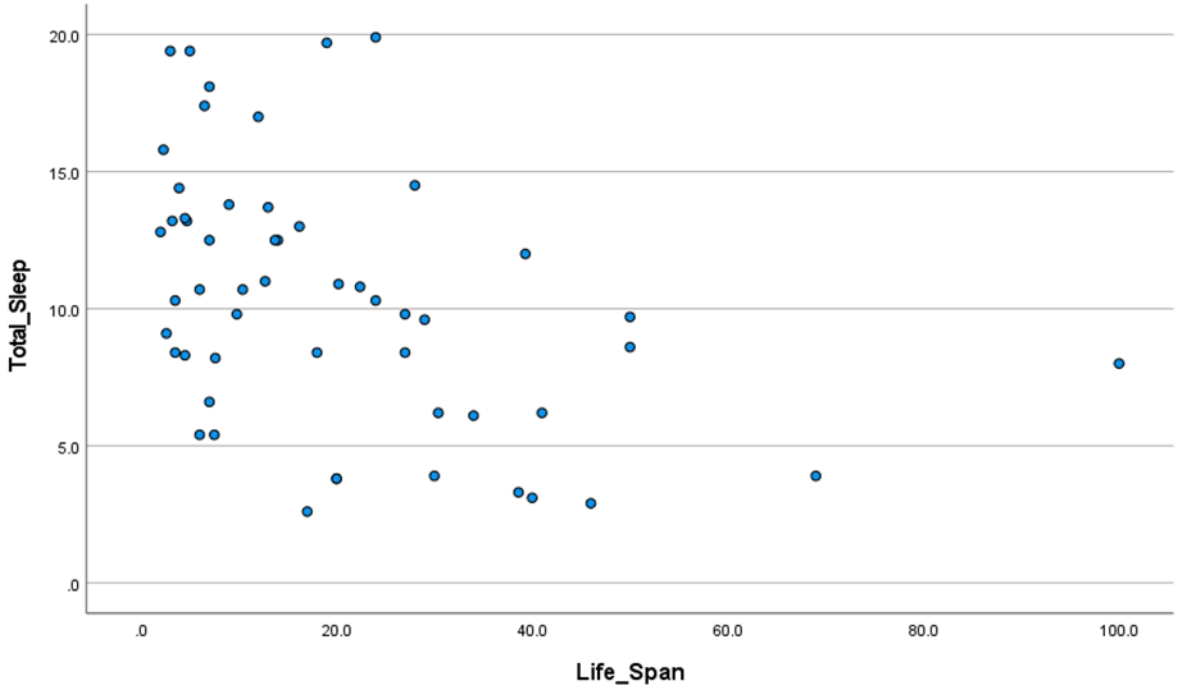


Figure 2.16

This figure 2.16 seems to suggest that a high number of sleeping hours does not equal a longer life span, with the species with the highest life span having only around 8-9 hours of sleep. On the other hand it seems that the animals which sleep most have the shortest life span, with the four species who have almost 20 hours of total sleep time having a life span of around 2 years to 30 years. On the other hand, the species with the least sleeping hours has a life span of around 20 years.

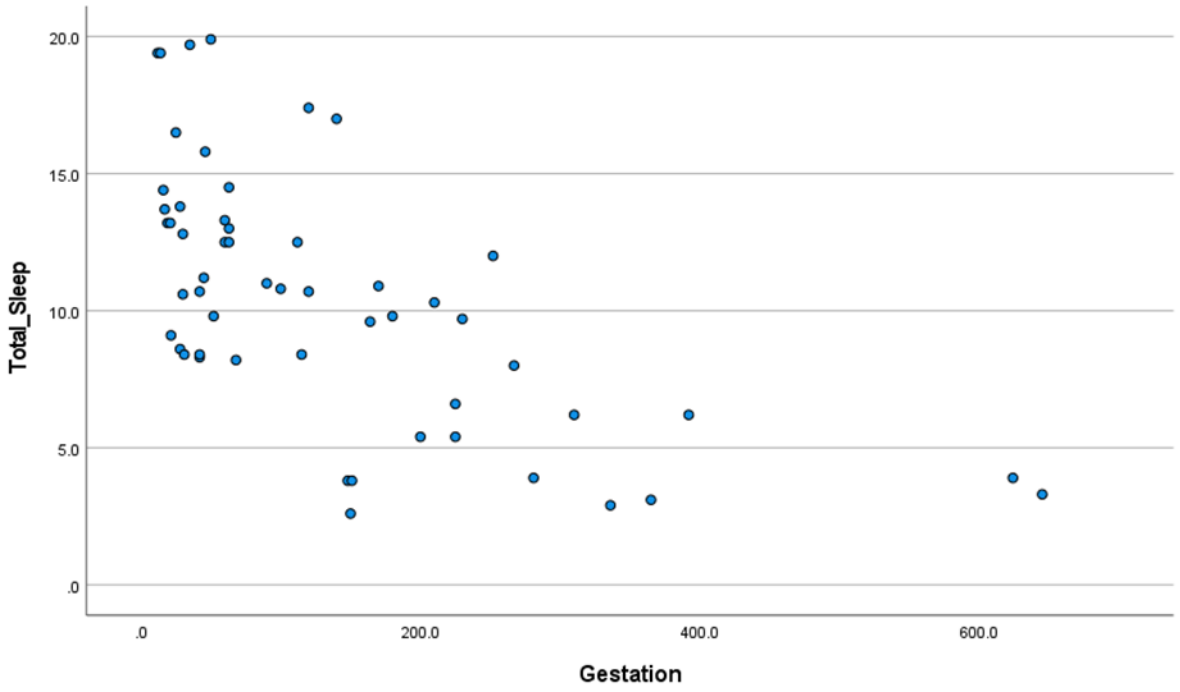


Figure 2.17

This gestation vis-à-vis total sleep and gestation time is quite similar to the graph of life span and total sleep. Most animals who have a shorter gestation time have the most number of total hours of sleep. Whilst the two species who have over 500 days of gestation time have a total sleeping time of under 5 hours. On the other hand, the animals with the least gestation time have a high amount of total sleeping hours, with 4 of them recording almost 20 hours of sleep.

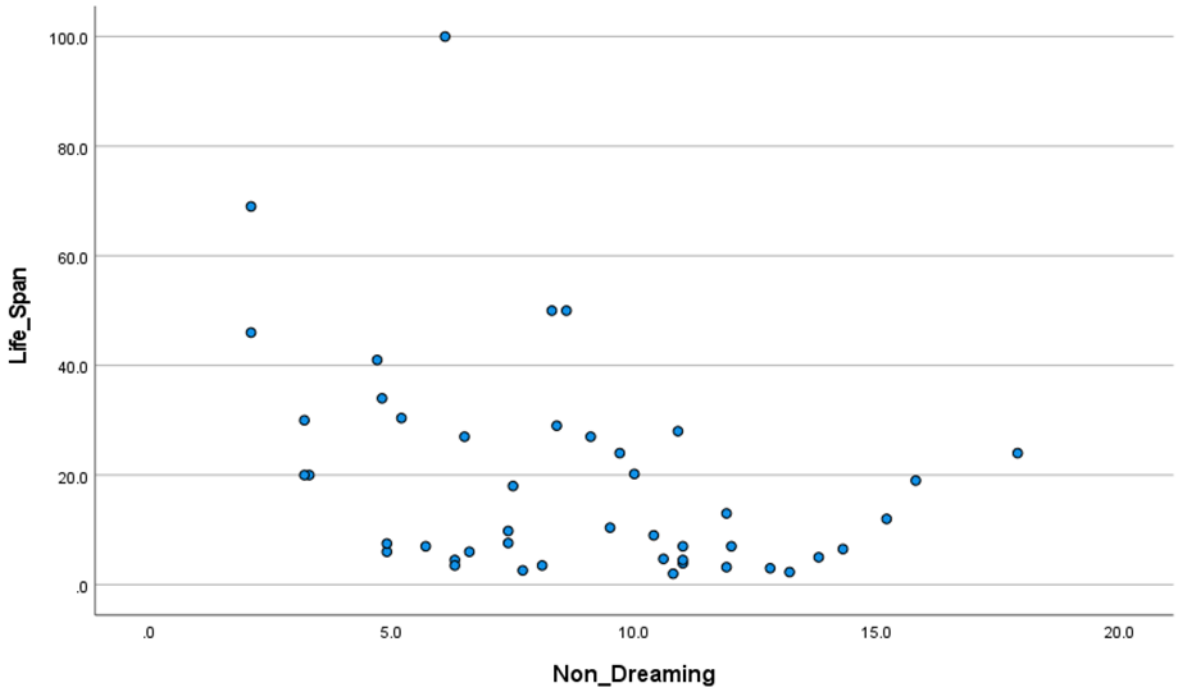


Figure 2.18

Figure 2.18 shows how that a high life span does not mean more dreaming hours. In fact, most animals show a total life span of under 40 years, with only around 6 species going over that life span, and only 1 species reaching 100 years. The longest non-dreaming time is around 18 hours and the species with that non-dreaming time has a life span of around 22 years.

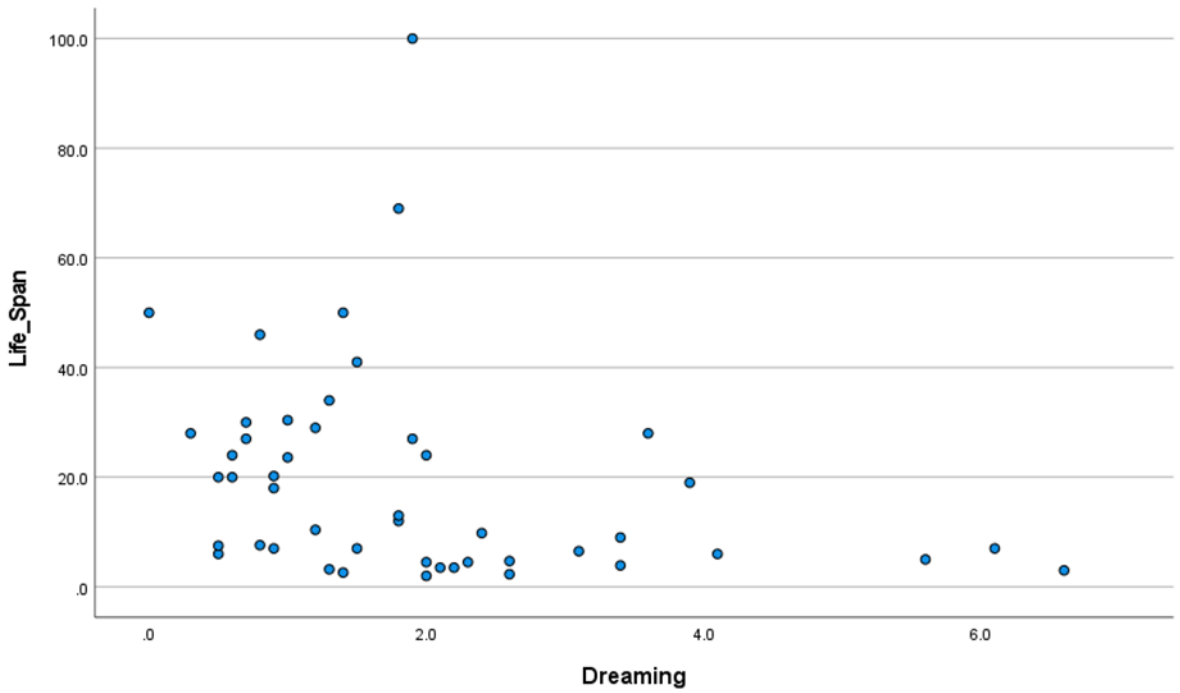


Figure 2.19

This figure complements figure 2.18 in that it shows that most animals do not dream a lot with only 3 species dreaming for around 6 hours. It seems that there is not an obvious relation between the hours of dreaming and life span. Again, most animals have a life span of under 40 years and during such time they do not dream much with most animals having less than 4 hours of total dreaming time.

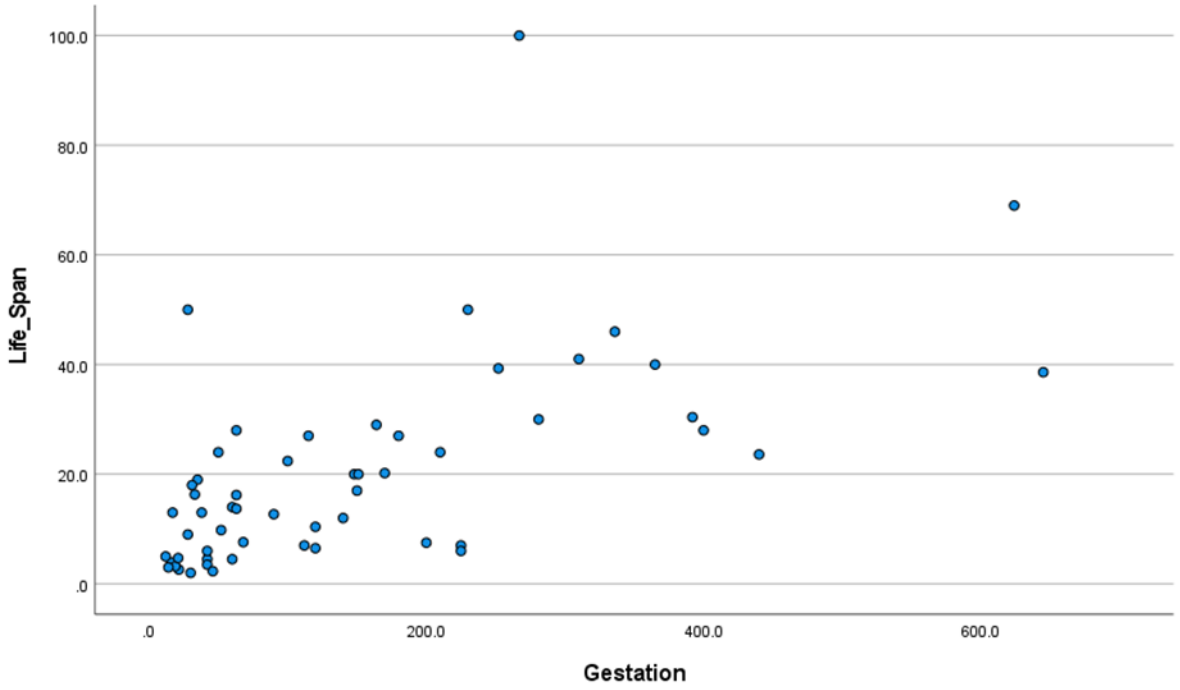


Figure 2.20

Figure 2.2 shows that most animals have a low gestation time, most under 200 days, with most animals in the under 200 days range having a gestation time of around 100 days. It appears that most animals with a short gestation time also have a short life span. However, again there is that one species which has a life span or around 100 hours with a gestation time of around 300 days. On the other hand there is another species with a gestation time of more than 600 days however its life span is just under 40 years.

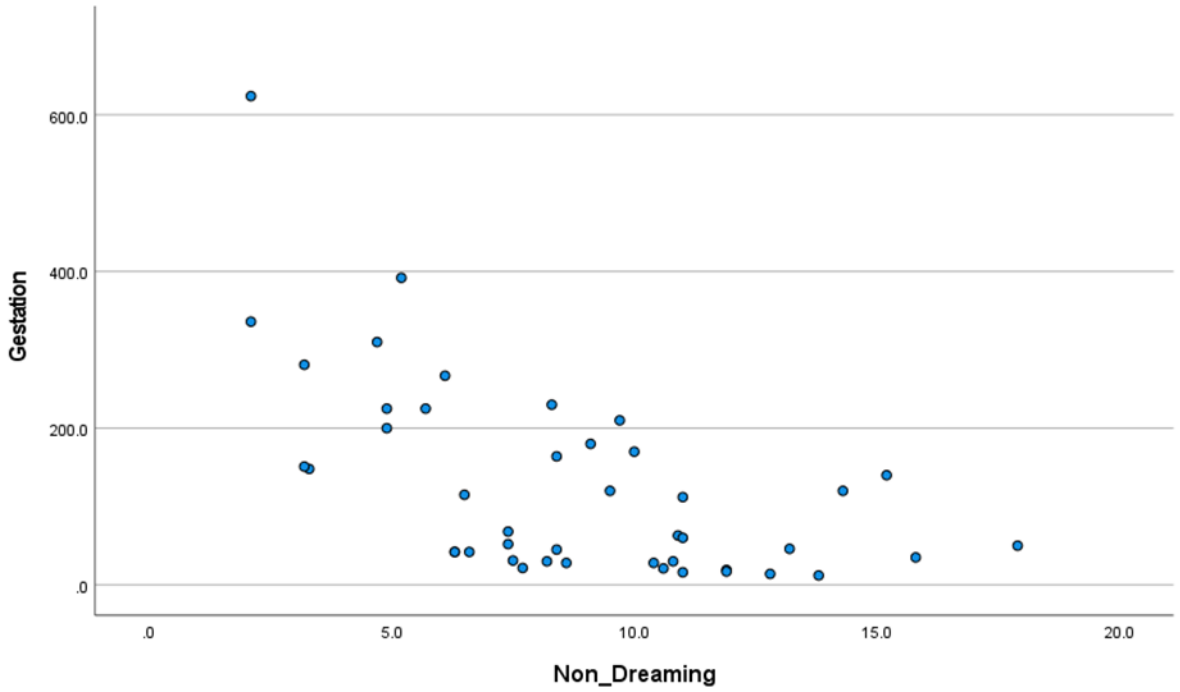


Figure 2.21

Figure 2.21 shows that most are under the 200 day gestation time. It seems that the animals with the least gestation time have around 6 to 18 hours of non-dreaming hours. It is apparent that most animals with a high gestation time have also low hours of non-dreaming time.

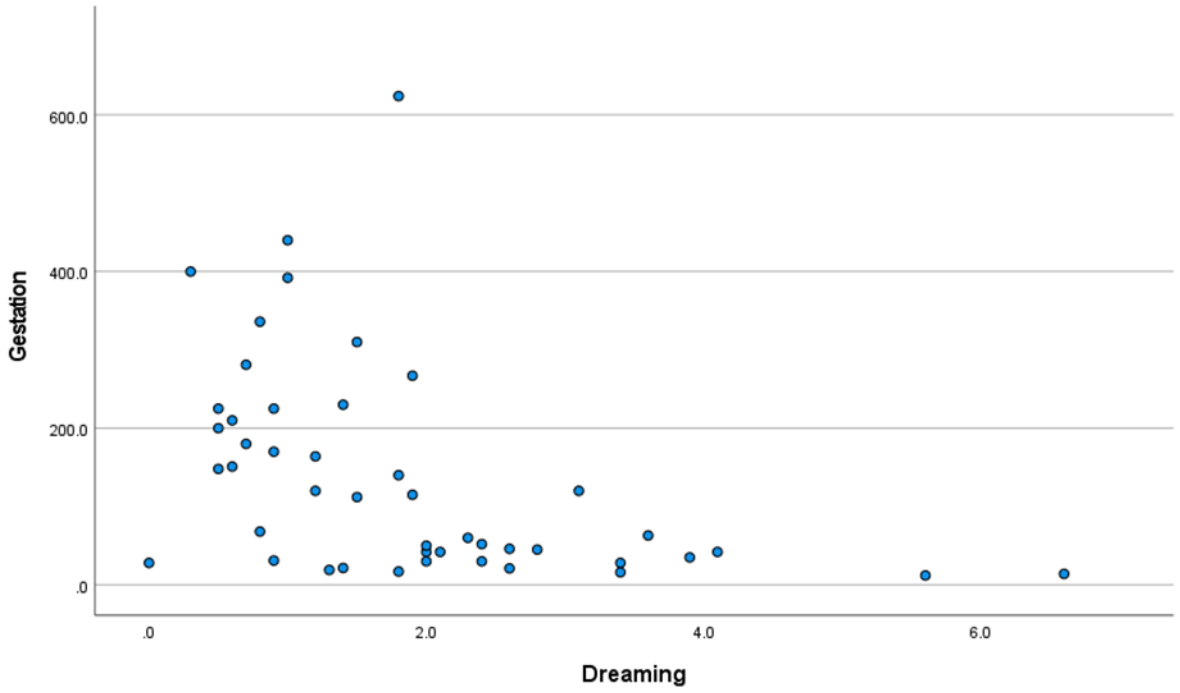


Figure 2.22

Figure 2.22 shows how most species spend little time dreaming regardless of their gestation time. Again there is that one species with a gestation time of over 600 days whose dreaming time is around 2 hours. We see that the two animals with the highest dreaming hours have a low gestation time. Hence it is apparent that species with a low gestation time have a dreaming time of around 1 to 7 hours of dreaming time.

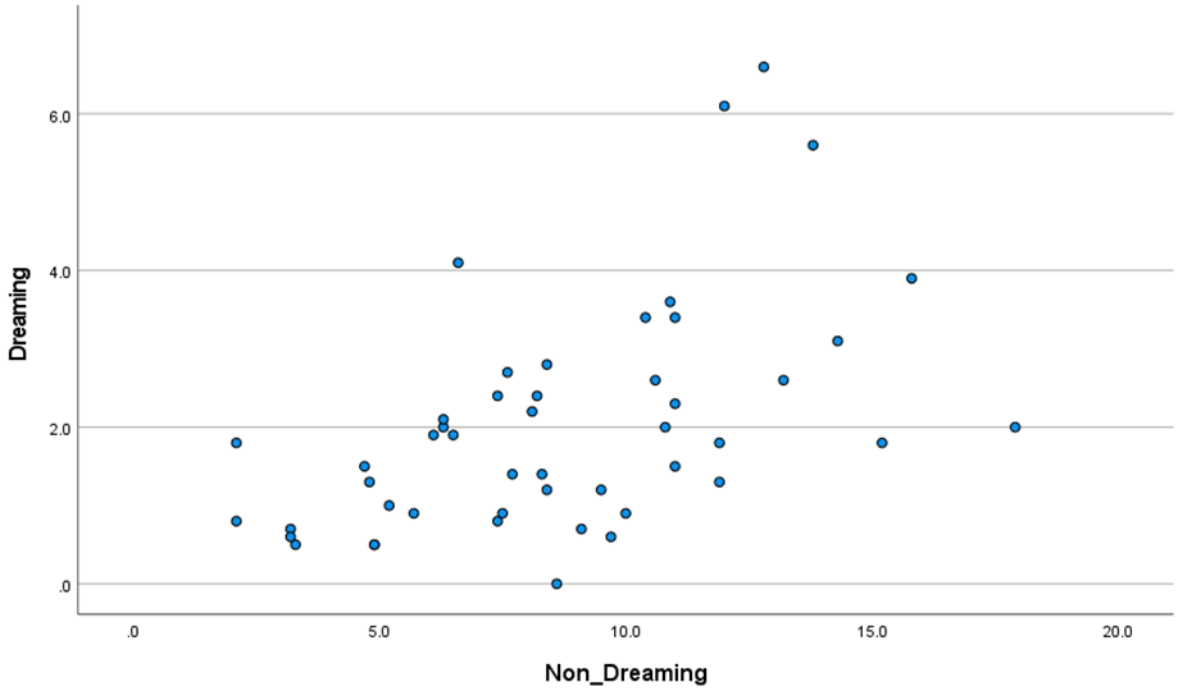


Figure 2.23

Figure 2.23 shows how many hours the species spend dreaming and non-dreaming. It is apparent that there if a creature who spends 2 hours dreaming and around 18 hours not dreaming, this is the creature with the highest non-dreaming hours. On the other hand there is a species with more than 6 hours of dreaming time and around 14 hours of non-dreaming time.

Life Span and Gestation

		Life_Span	Gestation
N	Valid	58	58
	Missing	4	4

Mean	19.878	142.353
Median	15.100	79.000
Mode	7.0	42.0[a]
Skewness	2.014	1.684
Kurtosis	5.885	2.852
Range	98.0	633.0
Minimum	2.0	12.0
Maximum	100.0	645.0
Sum	1152.9	8256.5

a. Multiple modes exist. The smallest value is shown

Figure 2.24

The above frequency table, Figure 2.24, shows that whilst we have 58 values there are 4 missing hence we do not have all the information surrounding the species and their life span and gestation time.

In addition, Figure 2.24 shows that the mean maximum life span of the species is almost 20 years with the mean gestation time being around 123 days, the median of the life span is 15 years whilst that of gestation is 79 days and for the life span there were enough results to

determine the mode, with 7 years being the most popular whilst for gestation time there were too much modes hence the smallest value of 42 is shown.

The below histograms, show the life span and gestation of the animals with a normal distribution curve imposed on the graphs. It is interesting to note that when one sees the life span and gestation histograms there are some similarities between the results.

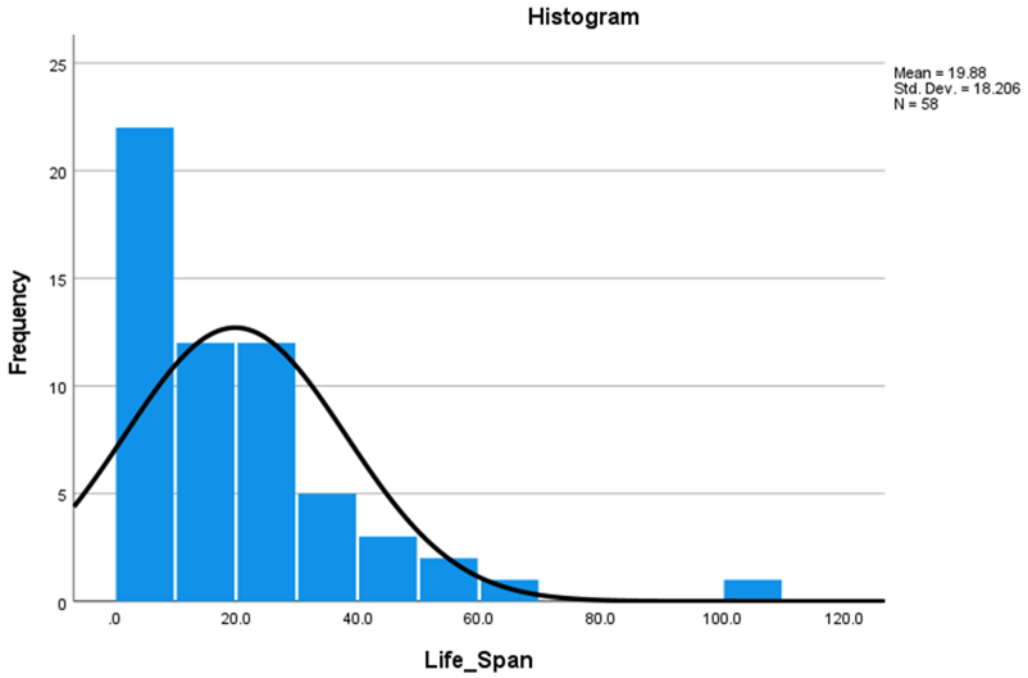


Figure 2.25

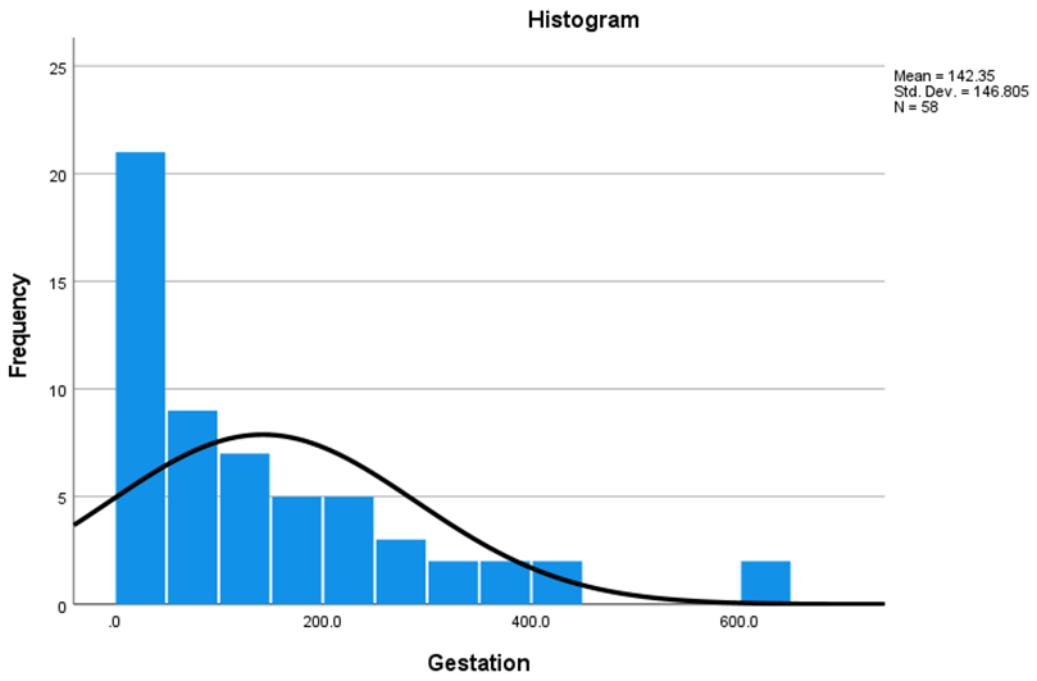


Figure 2.26

Body Weight and Brain Weight

		Body_Wt	Brain_wt
N	Valid	62	62
Mean		198.78998	283.1342
Median		3.34250	17.2500
Mode		.023[a]	1.00[a]

Skewness	6.564	5.072
Kurtosis	45.741	26.271
Range	6653.995	5711.86
Minimum	.005	.14
Maximum	6654.000	5712.00
Sum	12324.979	17554.32

a. Multiple modes exist. The smallest value is shown

Figure 2.27

The above frequency table shows the body weight and brain weight of the species. As we can see there are no missing values of body weight and brain weight hence, we have a complete picture of the results.

The mean body weight is approximately 200 kilograms whilst the mean brain weight is 283 grams. The median numbers are 3.3 for body weight and 17.3 for brain weight and due to the multiple modes and results, we have a 0.023 mode for body weight and a 1 mode for brain weight. The standard deviation of body weight and brain weight is quite similar, as are the skewness and error thereof. We see a difference in the kurtosis but again, the error thereof is the same.

As we can see from the below figure, most species have a low body and brain weight with only a few having a higher body and brain weight. It seems that when there is an increase in brain weight there also is an increase in body weight hence we do not find a large bodied species with a small brain and vice versa.

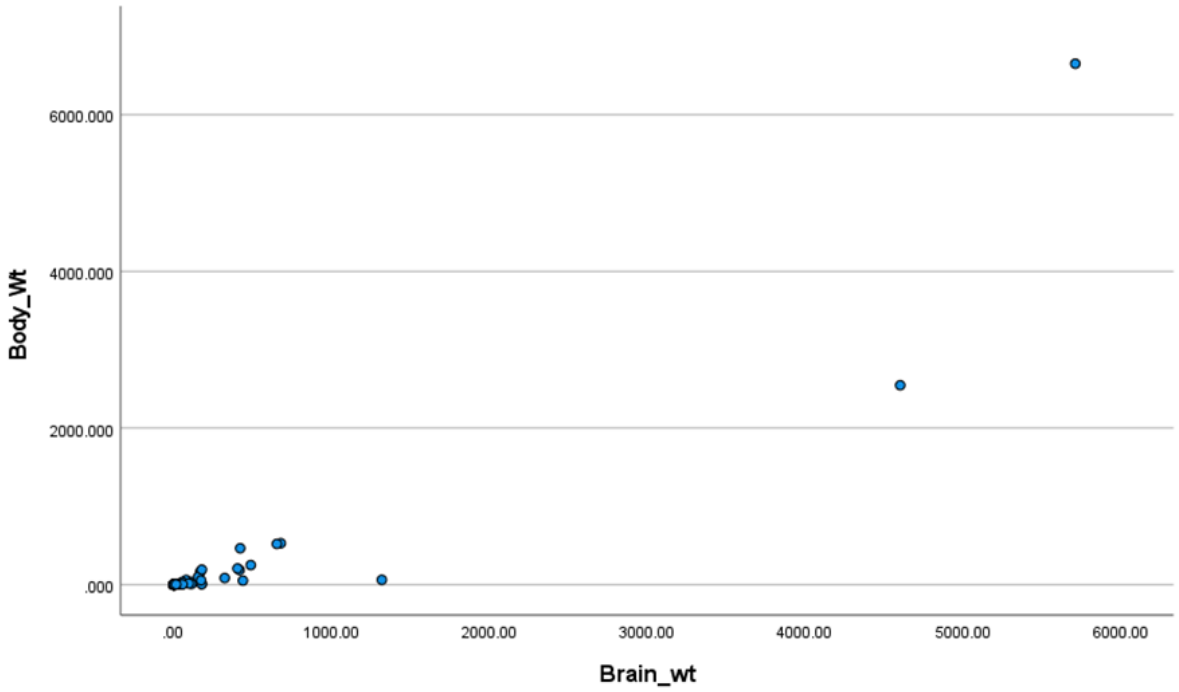

6000.000
4000.000
2000.000
.000
Body_Wt
.00
1000.00
2000.00
3000.00
4000.00
5000.00
6000.00
Brain_wt

Figure 2.28

Dreaming and Danger, Predation, Exposure

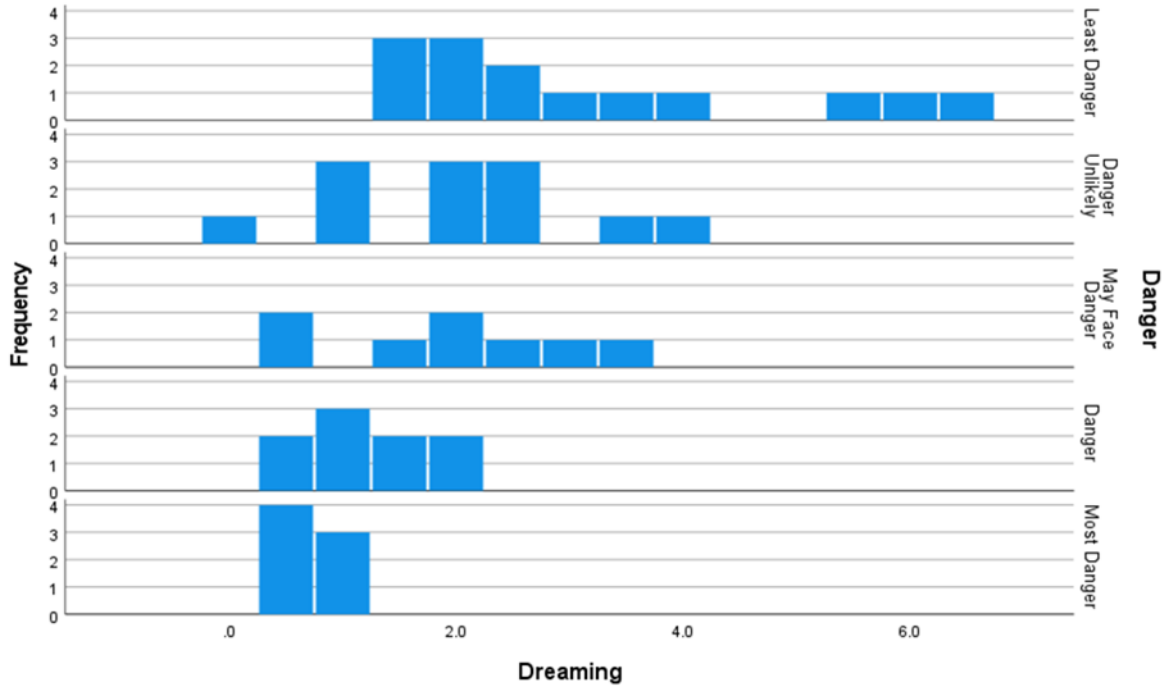


Figure 2.29

The above histogram shows the correlation between dreaming and danger. It is apparent that most animals who are not in danger are the ones which dream the most followed closely by those who are in the 'Danger Unlikely', 'May Face Danger', 'Danger', and 'Most Danger' categories. Most animals are found in the 'Least Danger' category the range of sleeping hours of which carry from 1 to 6.5. Those who have the least dreaming hours are in the 'Danger Unlikely' category whilst those who are in the 'Danger' and 'Most Danger' category show the closest dreaming hours.

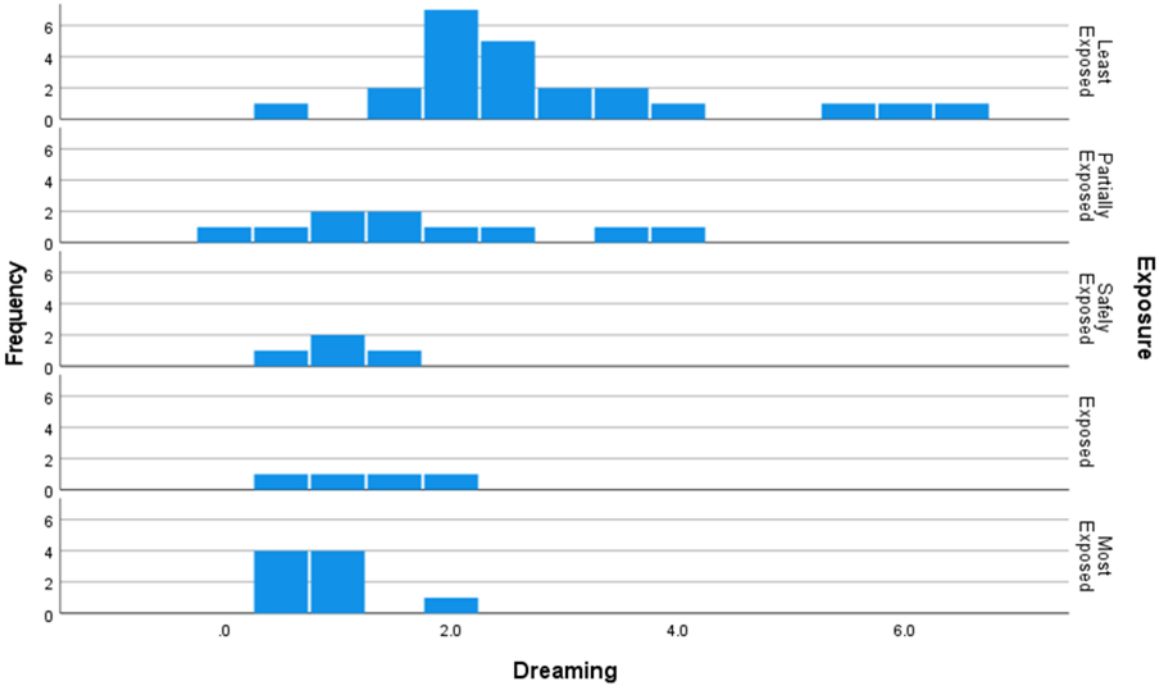


Figure 2.3

Figure 2.3 shows that animals who are the least exposed experience a lot more dreaming than those who are not exposed. This may be due to the correlation of exposure and predation the animals face as in all three tests we can see that the animals who are the safest generally have more activity whilst sleeping. The most number of animals, 22, is in the 'Least Exposed' category which also has the largest range of 'Dreaming' hours. The second largest range of sleeping is of the 'Partially Exposed' category whilst the smaller is of the 'Safely Exposed' animals. Most animals do not go over the 3 hours of dreaming, with only 8 animals doing so.

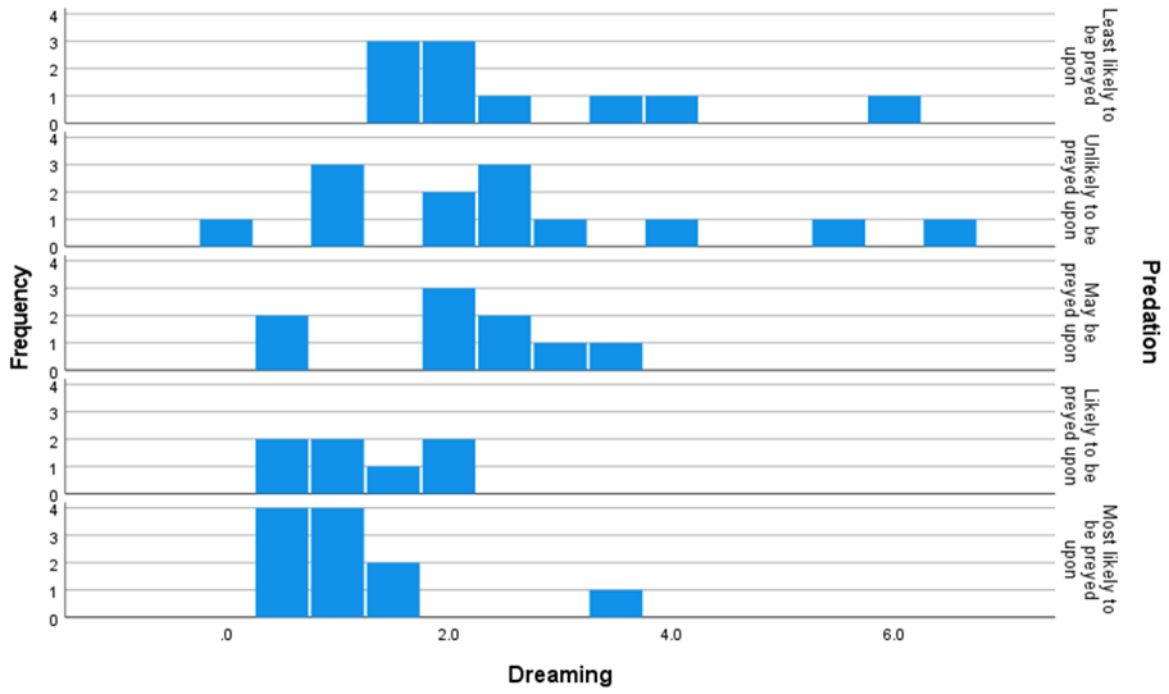


Figure 2.31

Similar results showing that predation, danger, and exposure have effects on the dreaming of the animals. Most animals are in the 'Unlikely to be preyed upon' category, such also experience the largest range of dreaming spanning from 0.5 to 6.5. The least amount of animals are in the 'Likely to be preyed upon' category and such animals also have the closest to each other number of dreaming hours.

Non-Dreaming and Danger, Predation and Exposure

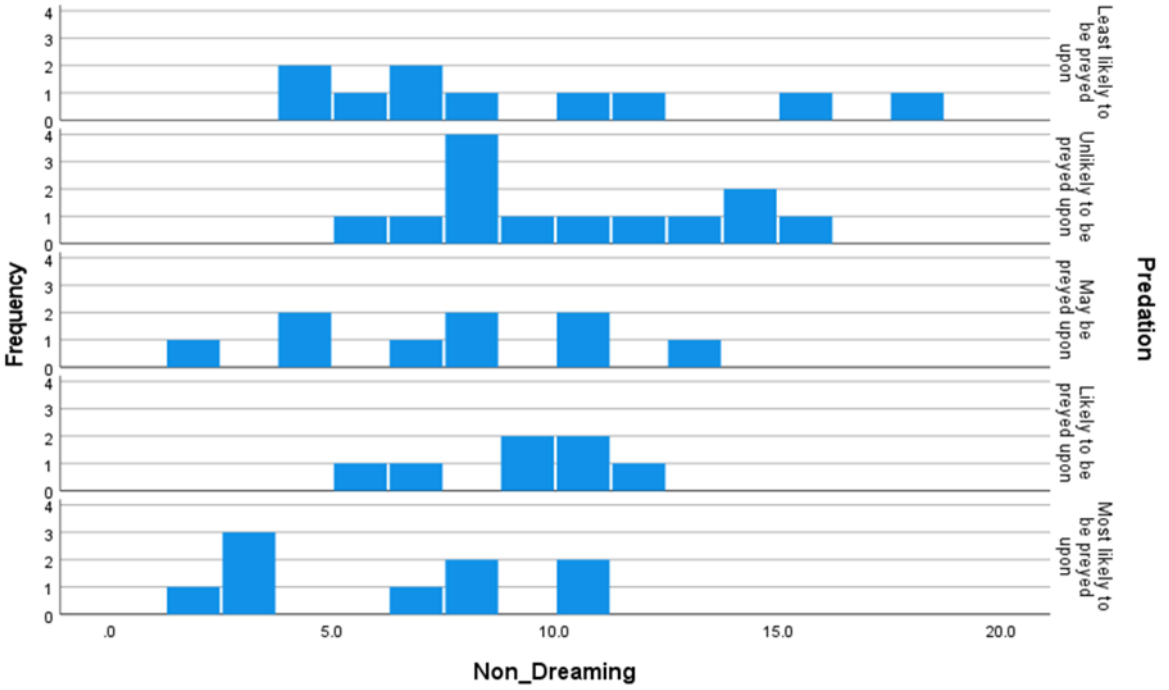


Figure 2.32

In Figure 2.32, we may observe that the largest number of species are within the 'Unlikely to be preyed upon' section which has very similar hours of non-dreaming activity. These species are followed by 'Least likely to be preyed upon' which has the most range, from around 3 to around 18. The categories of 'May be preyed upon' and 'Most likely to be preyed upon' have the same number of species in the category, 9 in each, and with similar results of non-dreaming hours. The 'Likely to be preyed upon' category has the least number of species with values ranging from around 5 to 12.

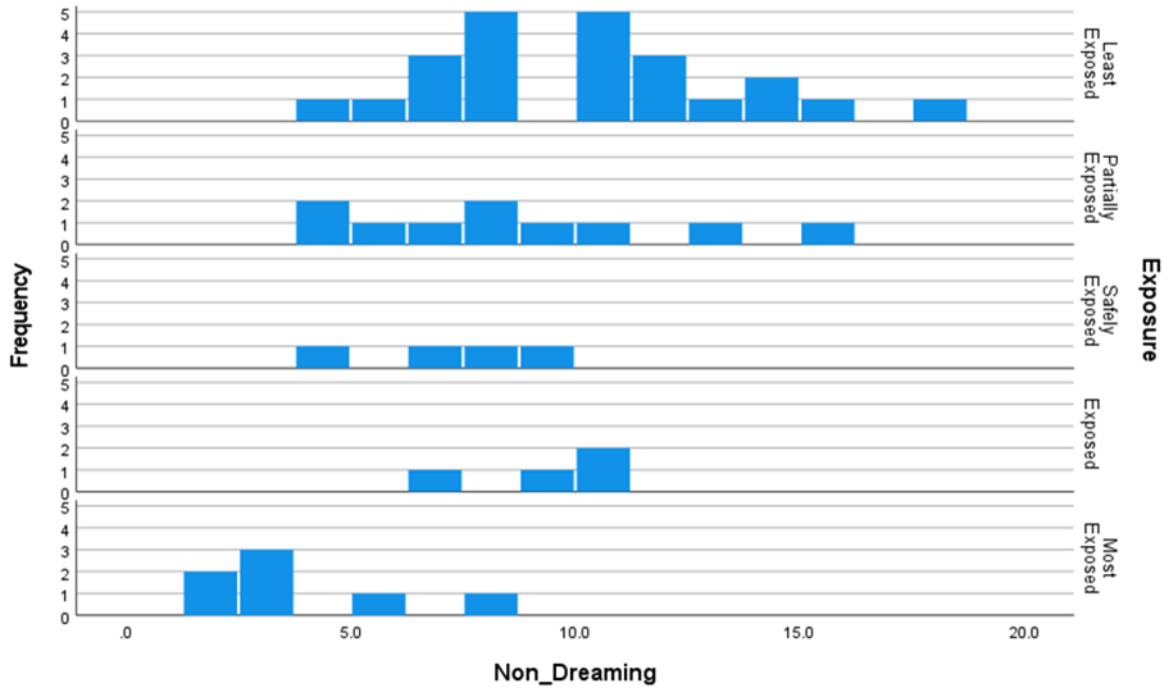


Figure 2.33

Most species in this graph, Figure 2.33, are within the 'Least Exposed' category which also has the most range of non-dreaming hours, the 'Partially Exposed' category has a similar range but with only 10 species within such. The 'Safely Exposed' and 'Exposed' categories each have 4 species in their category ranging from around 4.0 to 10.0 and around 6.0 to 12.0 respectively.

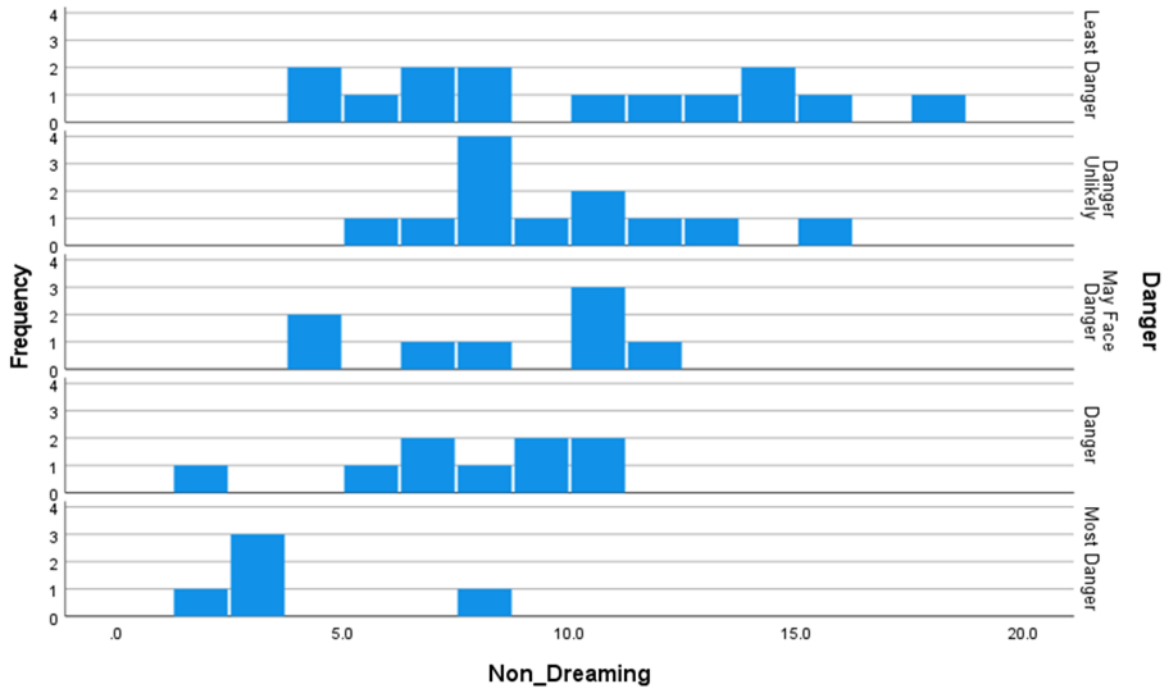


Figure 2.34

This graph, Figure 2.34 shows that most species are in the category of ‘Least Danger’ which has the most range of values and the largest non-dreaming hours followed closely by ‘Danger Unlikely’. The least amount of species is in the ‘Most Danger’ category with only 5 species and with non-dreaming hours which are less than 10 hours which marks the middle number of hours.

Total Sleep and Danger, Predation and Exposure

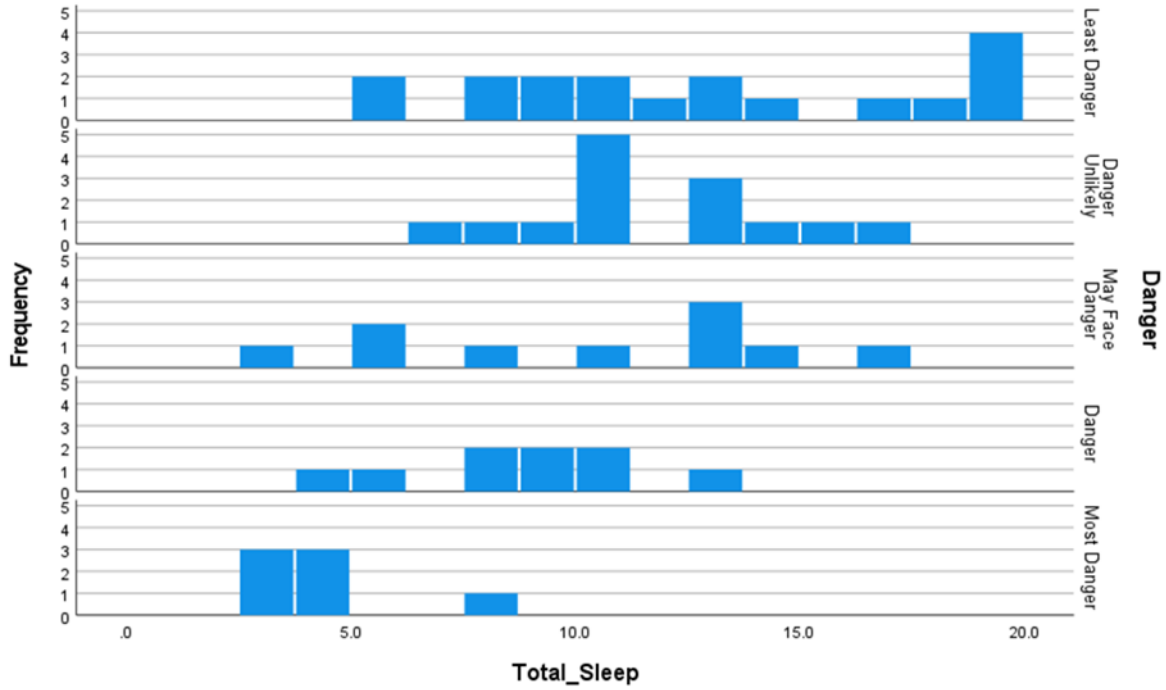


Figure 2.35

In Figure 2.35, we observe that the largest amount of species sleep in the ‘Least Danger’ of circumstances, and such species show the highest hours of total sleep in total. These are followed by the species within the ‘Danger Unlikely’ category which have quite average total hours of sleep. In the ‘May face danger’ category, we have the most spread out results of total sleep hours as opposed to species who are in the ‘Most danger’ circumstance wherein they sleep under 10 hours starting from around 2 hours to 9 hours.

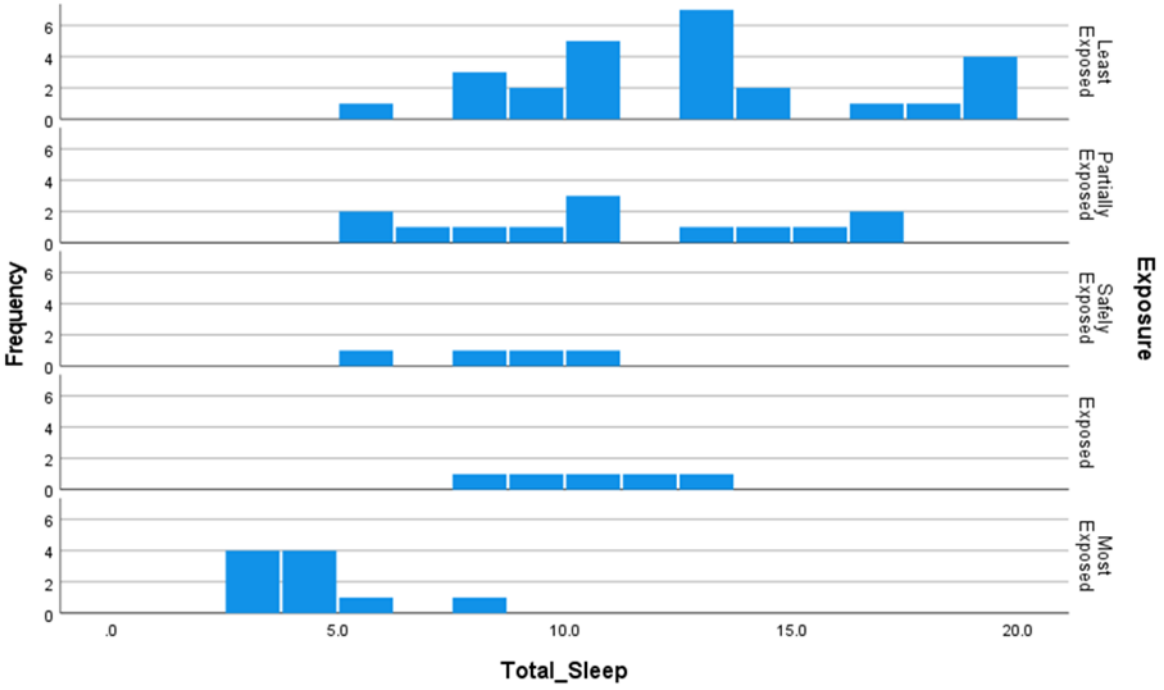


Figure 2.36

In Figure 2.36, most species are on the ‘Least Exposed’ part of the graph and such part generally records the largest total hours of sleep. Those who are ‘Partially Exposed’ start with the same minimum hours of total sleep but do not reach the same maximum, falling behind by around 2 hours. There is the least number of animals for those who sleep ‘Safely Exposed’ and ‘Exposed’ hence, the third most number of animals are in the ‘Most Exposed’ category but these all report the lowest number of sleep.

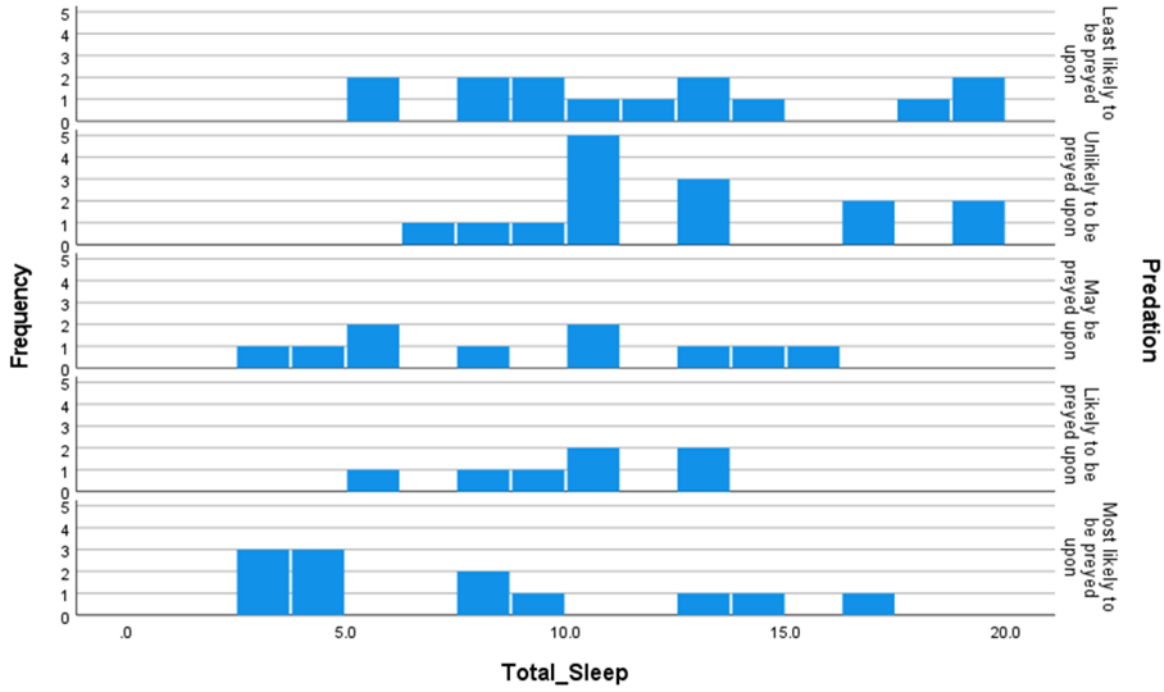


Figure 2.37

Figure 2.377 shows that the greatest number of animals are found in the 'Unlikely to be preyed upon' category, followed closely by those who are the 'Least likely to be preyed upon' and those 'Most likely to be preyed upon', the least amount of animals are found in the 'Likely to be preyed upon'. Those recording the highest number of total sleeping hours are those in the 'Least likely to be preyed upon category' which also has the most range followed closely by those 'Unlikely to be preyed upon'. The 'May be preyed upon' and 'most likely to be preyed upon' species have similar results albeit those who are 'Most likely to be preyed upon' have a species that sleeps more than those in the 'may be preyed upon category'

Life Span and Danger, Predation and Exposure

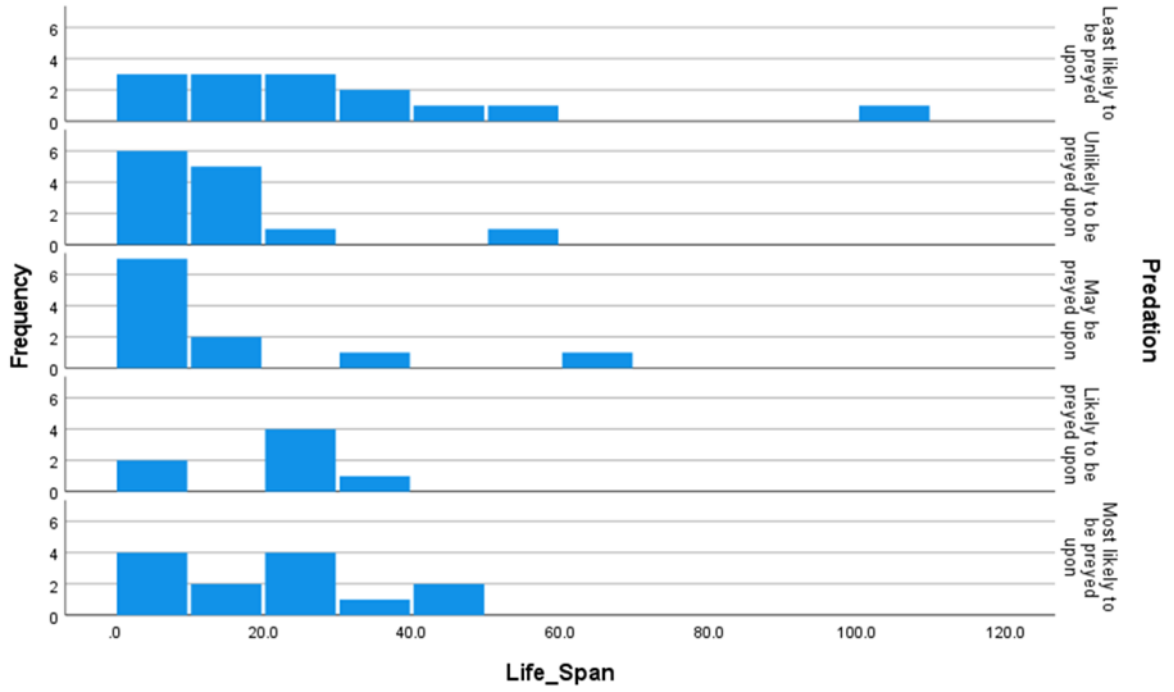


Figure 2.38

The above histogram, Figure 2.38 shows that most animals do not have a long life span, with only 1 species in the 'Least likely to be preyed upon' category living to around 100-120 years of age. Most species (22 in total) live only for around 10 years. Those who are 'Unlikely to be preyed upon' and 'May be preyed upon' show similar results but most of them also die early as well. It is interesting to note that animals in the 'Least likely to be preyed upon' category and 'May be preyed upon' show the most range whilst those who are 'Likely to be preyed upon' show the least.

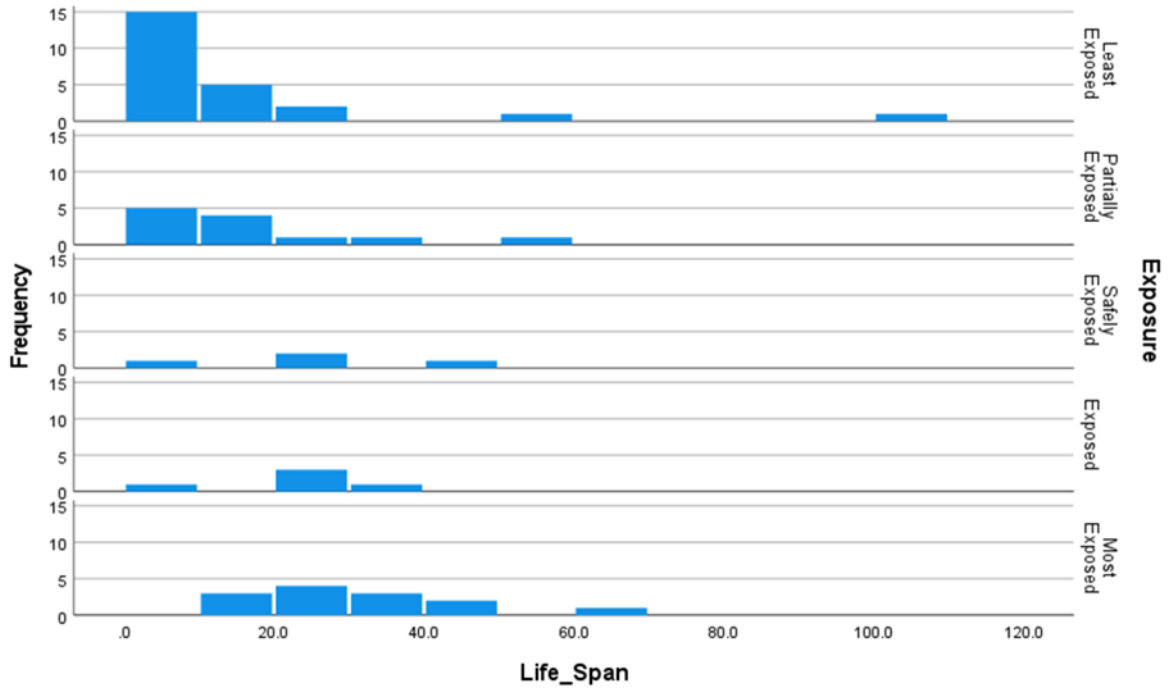


Figure 2.39

The above figure, Figure 2.39 shows that those who are the 'Least exposed' have a longer life span, with results corresponding to Figure 2.36. Again, most animals do not live must regardless of exposure, being that those who are the 'Most exposed' live from 15 years until around 70 years. The species that was least likely to be preyed upon in the above figure is also the one least exposed hence, a factor of his long life may be due to the safety he experiences.

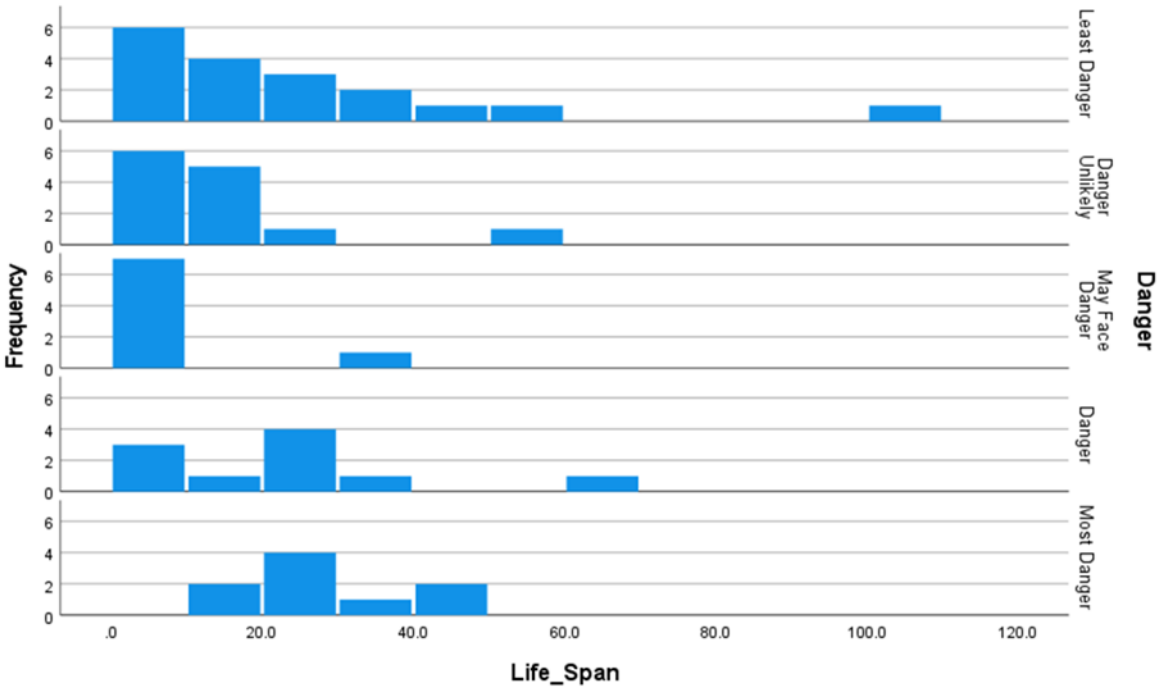


Figure 2.4

In Figure 2.4, the most number of animals are in the 'Least Danger' category which also shows the longest life span. Those who experience the 'Most Danger' vary in life span, starting from around 15 years to 50 years. The animals who 'May Face Danger' are only 8 species, with most dying within the first 10 years.

Gestation and Danger, Predation and Exposure

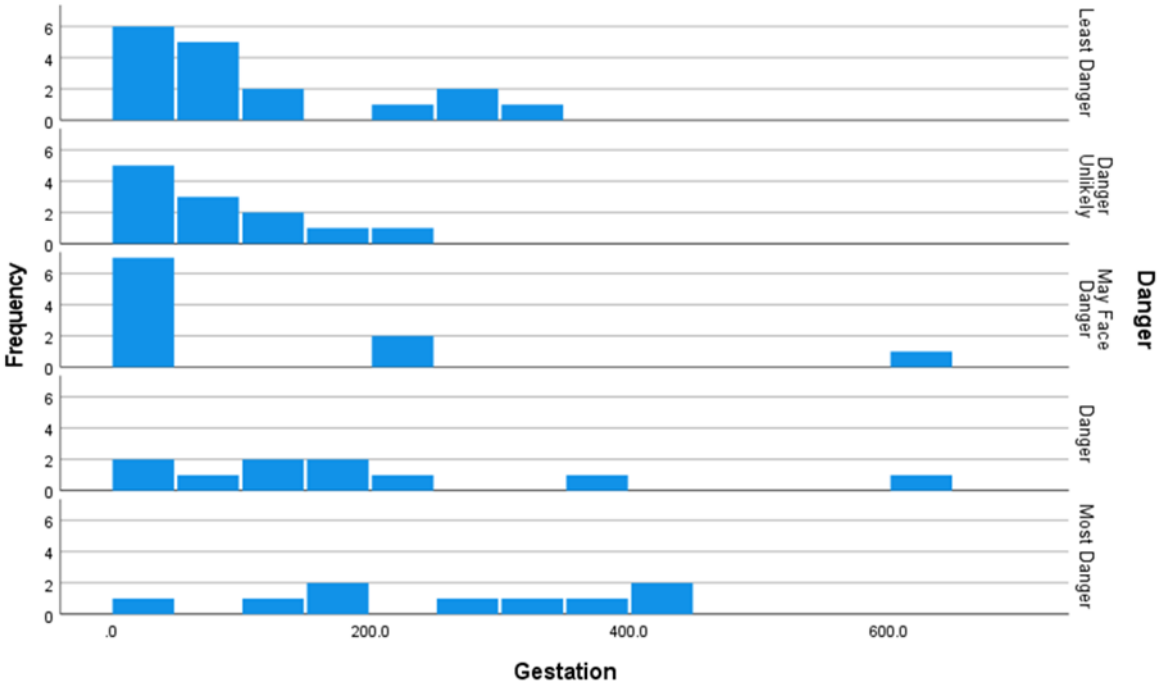


Figure 2.45

As can be observed from Figure 2.45, most animals have a short gestation time with only 2 going over the 600-day mark. The animals who sleep in the 'Least Danger', of circumstances mostly have a low to average gestation time and those who sleep in circumstances where danger is unlikely to have a gestation time that slowly decreases from the lowest point till around 200 days. 10 animals are in the 'May Face Danger' category, mostly having a low gestation time, then up to 250 days, and finally up to 600 to 650 days however, the latter two amounts are of only 3 animals, with only one animal having the largest number. The species in the categories of 'Danger' and 'Most Danger' have the most range and very similar numbers, with 10 species in one and 9 species in the other, respectively.

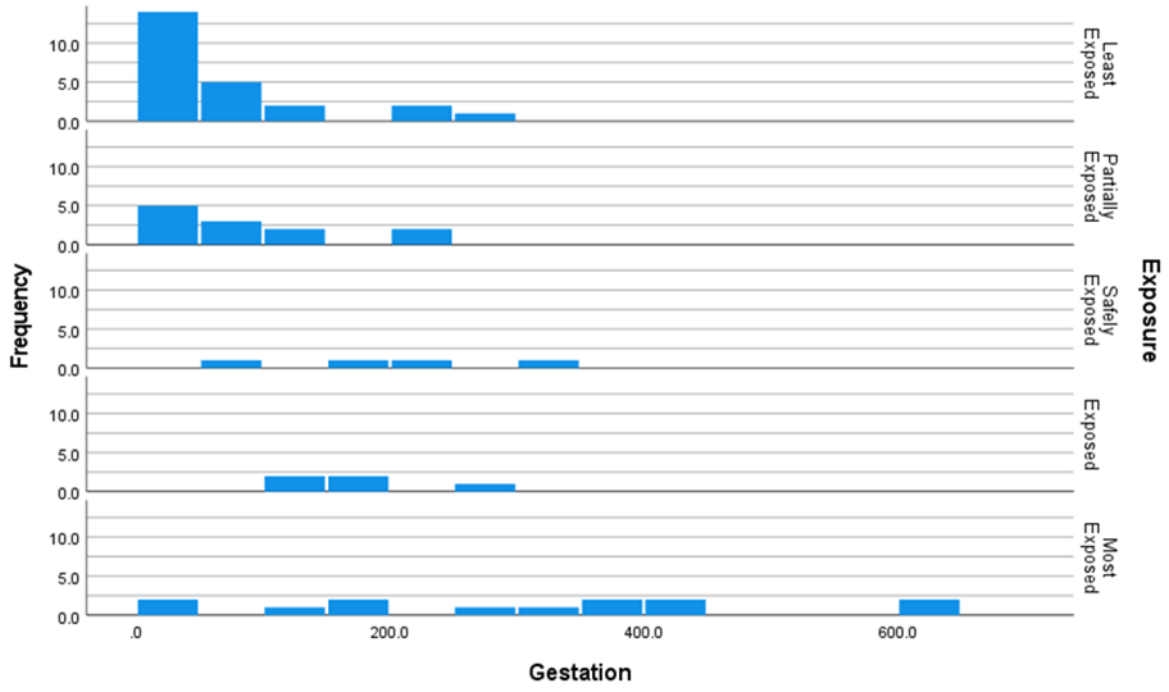


Figure 2.46

In Figure 2.46, the largest number of species can be found in the 'Least Exposed' category which mostly has the least number of gestation days. As apparent, most species have a short gestation time, with most being under 300 days but with around 7 species going over such days. The maximum gestation time is a bit more than 600 days with around 2 animals being in such a category and also being in the 'Most Exposed' category. The largest range of gestation time is found in the 'Most Exposed' category which has 13 animals and a gestation time spanning from around 1 to around 650 days.

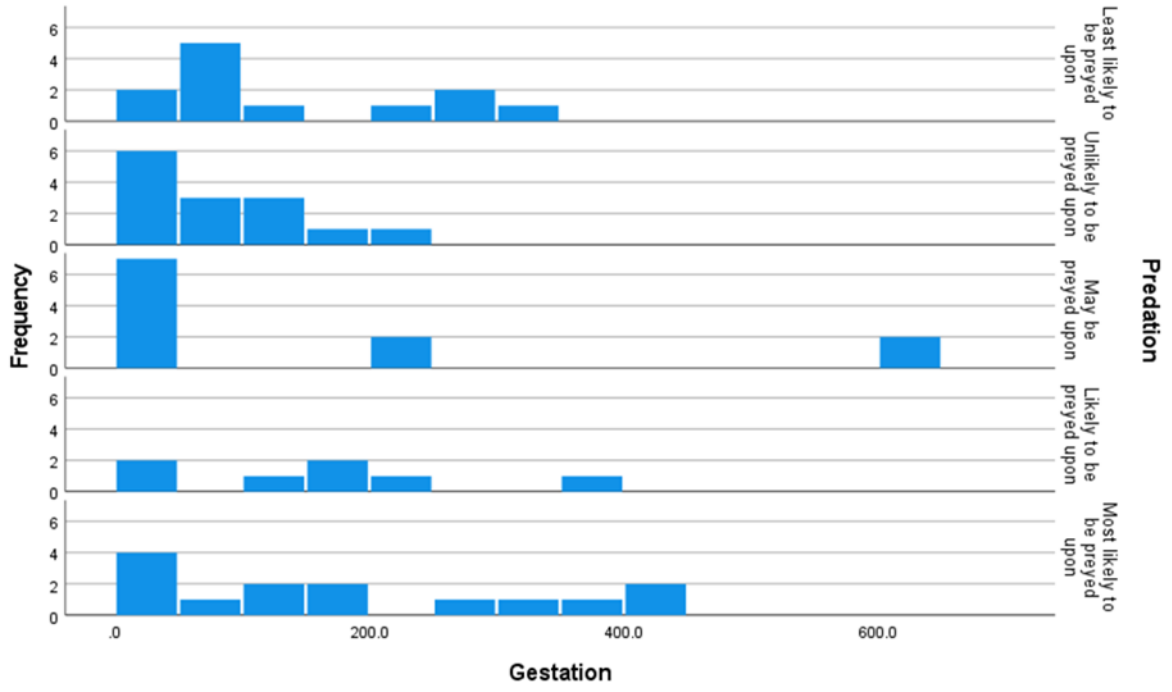


Figure 2.47

In Figure 2.47, the same number of species, 14, is found within the 'Unlikely to be preyed upon' and 'Most likely to be preyed upon' categories. Most species have less than 200-250 days of gestation time. The 2 species which have 600 and over days of gestation time are in the 'May be preyed upon' category. Following such gestation time, the largest time of gestation is between 400-450 days of gestation time which another 2 species are in.

Section 3: Statistical Analysis

In our statistical analysis, we will be conducting two independent sample t-tests; (i) a hypothesis test; and (ii) a fitting regression statistical model.

We will be conducting the following hypothesis tests:

a) Whether mean total sleep differs between the least and most preyed mammals
b) Whether mean total sleep differs between the least and most exposed mammals

The reason for these tests is to continue exploring the main objective of this study, that of understanding what affects sleep in mammals. The reasoning behind these tests is to try and understand whether mammals who tend to be most preyed upon tend to sleep more than those mammals who are least preyed upon. We think that the least a mammal is preyed on, the more it sleeps as it will feel more secure and this can be checked via Test 1. In addition, the same will be done to see whether the least exposed mammals tend to sleep more than the most exposed mammals. We think that the less exposed a mammal is, the more it will sleep as it will feel more exposed and this will be checked in Test 2.

Furthermore, we will be trying to fit a regression statistical model with 'life_span' being the dependent variable. We will be trying to use 'brain_wt', 'body_wt', 'total_sleep' and 'gestation' as independent variables.

The reasons for this are the following:

- We think that the size and weight of the brain affect the life span as the bigger the brain is, the more likely the mammal will be able to think and make decisions to save its life
- We think that the weight of the body affects the life span because we think that the more a mammal weighs, the bigger it will be and a bigger mammal can defend itself better than a small mammal. On the other hand, a mammal with a lot of weight can be overweight and unhealthy and this can also affect its life span.
- We think that the more an animal sleeps, the more alert and sharp it will be while if it does not sleep a lot, it will be more tired and less capable to defend itself. Therefore, we think that total sleep might affect the life span of the mammal.
- We think that gestation affects the development of the mammal and possibly, its size when being born. Therefore, we think that this could affect its life span because as explained in the first point, we think that the mammals size affects life span.

Therefore, we will be trying to fit a regression model to fit these variables in order to predict the life span of mammals.

3A: Testing for normality

In order to perform an independent samples t-test, we first need to test for normality.

For these tests, both the Kolmogorov-Smirnov and Shapiro-Wilk tests were conducted with SPSS, however, only the Shapiro-Wilk test was taken into consideration since it is deemed to have a better performance.

The null and alternate hypothesis for a normality test is as follows:

H_0: Variable follows a normal distribution
H_1: Variable does not follow a normal distribution

Normality Test 3A.1: Whether mean total sleep differs between the least and most preyed mammals

In order to perform this test, we need to make sure that the samples of total sleep for both the least preyed mammals and the most preyed mammals are normally distributed.

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	predation	Statistic	df	Sig.	Statistic	df	Sig.
total_sleep	1	.133	14	.200*	.923	14	.243
	2	.202	15	.099	.918	15	.178
	3	.194	10	.200*	.930	10	.443
	4	.145	7	.200*	.960	7	.818
	5	.266	12	.019	.866	12	.058

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Fig 3.1: Normality test for 3A.1

The predation values for the least and most preyed mammals are 1 and 5 respectively. As can be seen in Fig 3.1, the p-value (Sig) for both these groups are bigger than 0.05, meaning that H_0 is accepted for both variables and therefore, indicating that both follow a normal distribution.

Normality Test 3A.2: Whether mean total sleep differs between the least and most exposed mammals

In order to perform this test, we need to make sure that the samples of total sleep for both the least exposed mammals and the most preyed mammals are normally distributed.

Tests of Normality

	exposure	Kolmogorov-Smirnov[a] Statistic	df	Sig.	Shapiro-Wilk Statistic	df	Sig.
total_sleep	1	.146	26	.160	.939	26	.126
	2	.127	13	.200*	.940	13	.456
	3	.215	4	.	.949	4	.709
	4	.179	5	.200*	.958	5	.793
	5	.365	10	<.001	.764	10	.005

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Fig 3.2: Normality test for 3A.2

The exposure values for the least and most exposed mammals are 1 and 5 respectively. As can be seen in Fig 3.2, the p-value (Sig) for the least exposed mammals is bigger than 0.05, meaning that H_0 is accepted and that it follows a normal distribution. However, H_0 has to be rejected for the most exposed mammals as the p-value is smaller than 0.05, indicating that it does not follow a normal distribution. Consequently, the Mann-Whitney test has to be used instead of the parametric independent samples t-test.

3B: Test 1 - Whether mean total sleep differs between the least and most preyed mammals

For the first test, the first assumption that the two samples are normally distributed was satisfied with the Shapiro-Wilk normality test in Test 3A.1. Before applying the test we will also be testing the second assumption which needs to be satisfied, stating that total sleep data for each group has equal population variances. The latter will be testing using a Levene's Test.

For this test, we will be testing the following null and alternative hypotheses:

H0: On average, there is no significant difference in the total sleep of the least and most preyed mammals
H1: On average, there is a significant difference in the total sleep of the least and most preyed mammals

As explained above, first, we will be conducting a Levene's Test with the following null and alternative hypothesis:

H1 : The variances of the two populations (least and most preyed on) are not significantly different from each other

H0 : The variances of the two populations (least and most preyed on) are significantly different from each other

Group Statistics

	predation	N	Mean	Std. Deviation	Std. Error Mean
total_sleep	1	14	12.050	4.6023	1.2300
	5	12	7.383	4.8077	1.3879

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means							
						Significance				95% Confidence Interval of the Difference	
		F	Sig.	t	df	One-Sided p	Two-Sided p	Mean Difference	Std. Error Difference	Lower	Upper
total_sleep	Equal variances assumed	.189	.684	2.525	24	.009	.019	4.6667	1.8480	.8526	8.4808
	Equal variances not assumed			2.516	23.039	.010	.019	4.6667	1.8545	.8307	8.5026

Fig 3.3: Levene's Test and independent sample t-test

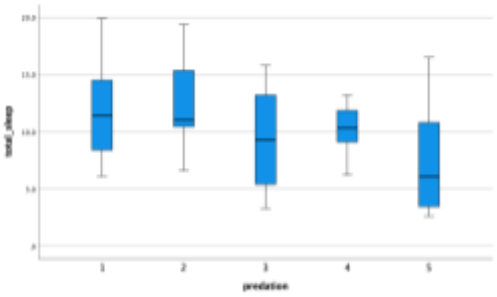


Fig 3.4: Box plot for total_sleep vs predation

As shown above, the p-value for the Lavene's test of 0.684 is greater than 0.05, meaning that we can accept H_0 and that the populations' variances are not different and equal. Consequently, both of the assumptions are satisfied and the parametric independent sample t-test can be conducted.

From the p-value in the 'Equal variances assumed' row, 0.019 is smaller 0.05 so we reject H0, meaning that there is a significant difference in the average total sleep time between mammals that are least and most preyed on. In addition, the descriptive statistics in Fig 3.3 and the box plot in Fig 3.4 shows that the mean total sleep is more for the least preyed upon mammals. This aligns with our initial thoughts and this could mean that mammals who are less attacked, sleep more as they feel safer. In addition, mammals which sleep more are more aware and alert and hence, it is more difficult for them to be preyed upon. On the other hand, animals which sleep less can be more tired and attacked more easily.

3C: Test 2 - Whether mean total sleep differs between the least and most exposed mammals

For the second test, the first assumption that the two samples are normally distributed could not be satisfied as was shown Test 3A.2. Consequently, instead of the independent sample

t-test, we will be using the Mann-Whitney test, which assesses whether the medians of two independent subgroups statistically differ from each other.

For this test, we will be testing the following null and alternative hypotheses:

H0: On average, there is no significant difference in the total sleep of the least and most exposed mammals
H1: On average, there is a significant difference in the total sleep of the least and most exposed mammals

Mann-Whitney Test

Ranks

	exposure	N	Mean Rank	Sum of Ranks
total_sleep	1	26	23.33	606.50
	5	10	5.95	59.50
	Total	36		

Test Statistics[a]

	total_sleep
Mann-Whitney U	4.500
Wilcoxon W	59.500
Z	-4.434
Asymp. Sig. (2-tailed)	<.001
Exact Sig. [2*(1-tailed Sig.)]	<.001[b]

a. Grouping Variable: exposure
b. Not corrected for ties.

Fig 3.5: Mann-Whitney Test

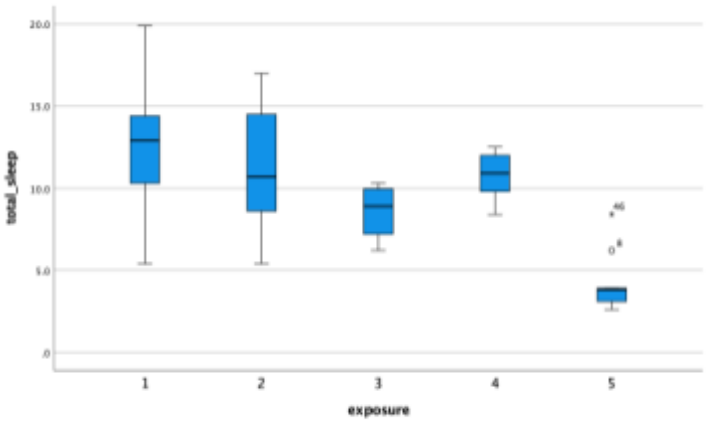


Fig 3.6: Box plot for total_sleep vs exposure

As shown above in Fig 3.5, the p-value for the Mann-Whitney test is < 0.001 which is less than 0.05, meaning that we can reject H_0, meaning that there is a significant difference in the average total sleep time between mammals that are least and most exposed. In addition, the box plot in Fig 3.6 below suggests that on average, exposed mammals tend to sleep less than mammals which sleep in a well-protected den and hence, this also aligns with our initial thoughts.

3CA - Fitting a Linear Regression Model

We will be trying to fit a linear regression model as a statistical model on this dataset. The goal shall be to establish whether there exists a relationship between "life_span" as a response variable with “brain_wt”, “body_wt”, “total_sleep”, and “gestation” as predictor variables since we believe that these variables can affect the life span of a mammal. The next step involves going through all the required assumptions to be satisfied in order for this model to be fitted.

Assumption 1: All variables are covariates

The first assumption is satisfied as all response and explanatory variables to be used (life_span, brain_wt, body_wt, total_sleep, gestation) are all covariates and qualitative variables.

Assumption 2: A linear relationship exists between the dependent variable and each of the independent variables

In order to check the linear relationship, we need to obtain correlation coefficients. First of all, we need to decide whether to use Pearson's or Spearman's coefficients. This can be done by testing all the following 3 assumptions which are required in order to use the Pearson coefficient and if one of them fails, the Spearman coefficient is used. The 3 assumptions which need to be satisfied in order to use the Pearson coefficient are the following:

a) The variables have a joint bivariate normal distribution
b) A linear relationship between the variables exist
c) No outliers in the variables

The first step is to test the join bivariate normal distribution. This has to be done for each pair of variables. To do this, we used the **mvnorm.etest** method from the **energy** package as can be seen in the script below

```
package(energy)
data <- read.csv("path_to_dir/sleep.csv")
x <- data$var1
y <- data$var2
mvnorm.etest(cbind(x, y), R=200)
```

The above test has the following null and alternative hypotheses:

H_0: The variables follow a bivariate normal distribution.
H_1: The variables do not follow a bivariate normal distribution.

The script above returns a p-value for the tests for each variable pair. The results can be found in the table below. If the p-value is greater than 0.05, it means that we accept H_0 and that the variable pair has a joint bivariate normal distribution.

	brain_wt	body_wt	total_sleep	life_span	gestation
brain_wt	NA	2.2e-16	2.2e-16	2.2e-16	2.2e-16

body_wt	2.2e-16	NA	2.2e-16	2.2e-16	2.2e-16
total_sleep	2.2e-16	2.2e-16	NA	2.2e-16	2.2e-16
life_span	2.2e-16	2.2e-16	2.2e-16	NA	2.2e-16
gestation	2.2e-16	2.2e-16	2.2e-16	2.2e-16	NA

As can be seen above, all the values are smaller than 0.05. Therefore, we have to reject H_0, meaning that the variables do not follow a bivariate normal distribution. In addition, the first assumption for the Pearson coefficient cannot be used and we have to use the Spearman coefficient. The following correlation indices were obtained for all the mentioned variable using the Spearman coefficient.

Correlations

			life_span	body_wt	brain_wt	total_sleep	gestation
Spearman's rho	life_span	Correlation Coefficient	1.000	.724**	.816**	-.457**	.673**
		Sig. (2-tailed)	.	<.001	<.001	<.001	<.001
		N	58	58	58	54	55
	body_wt	Correlation Coefficient	.724**	1.000	.953**	-.506**	.728**
		Sig. (2-tailed)	<.001	.	<.001	<.001	<.001
		N	58	62	62	58	58
	brain_wt	Correlation Coefficient	.816**	.953**	1.000	-.557**	.805**
		Sig. (2-tailed)	<.001	<.001	.	<.001	<.001
		N	58	62	62	58	58
	total_sleep	Correlation Coefficient	-.457**	-.506**	-.557**	1.000	-.657**
		Sig. (2-tailed)	<.001	<.001	<.001	.	<.001
		N	54	58	58	58	54
	gestation	Correlation Coefficient	.673**	.728**	.805**	-.657**	1.000
		Sig. (2-tailed)	<.001	<.001	<.001	<.001	.
		N	55	58	58	54	58

**. Correlation is significant at the 0.01 level (2-tailed).

Fig 3.7: Correlation indices

As can be seen in Fig 3.7, all independent variables are correlated with the dependent variable, “life_span”.

Assumption 3: No multicollinearity

There must be no multicollinearity between the predictor variables. Fig 7 can be used to assess the pairwise correlation coefficients between the independent variables. It can be seen that all the independent variables are correlated with each other. We will confirm this by the multicollinearity diagnostics in Fig 3.8 and Fig 3.9. As can be seen in the coefficients table in Fig 3.8, brain_wt and body_wt have a VIF value greater than 5, meaning that multicollinearity is present. Therefore, we will remove these variables and try again.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	10.005	7.730		1.294	.202		
	body_wt	-.024	.005	-1.221	-4.530	<.001	.123	8.152
	brain_wt	.024	.006	1.285	4.031	<.001	.088	11.400
	total_sleep	-.087	.512	-.021	-.170	.866	.563	1.777
	gestation	.061	.026	.458	2.389	.021	.242	4.125

a. Dependent Variable: life_span

Fig 3.8: Coefficients Table

In Fig 3.9, the VIF values are below 5 and the largest condition index is 8.01 which only represents weak dependencies. However, the last eigenvalue is close to 0, indicating serious or near serious dependencies. In addition, the variance proportions is greater than 0.5 which indicates that the variables are characterised by dependency as suggested by Belsley et. al (2004).

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	7.622	8.283		.920	.362		
	total_sleep	.032	.582	.008	.055	.956	.604	1.655
	gestation	.086	.019	.642	4.491	<.001	.604	1.655

a. Dependent Variable: life_span

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions		
				(Constant)	total_sleep	gestation
1	1	2.408	1.000	.01	.01	.03
	2	.554	2.085	.00	.06	.36
	3	.038	8.010	.99	.93	.61

a. Dependent Variable: life_span

Fig 3.9: Updated multicollinearity diagnostics

Consequently, since multicollinearity is present, we decided to keep only the independent variable with the highest correlation with the dependency variable, which is “brain_wt” with a correlation coefficient of 0.816 from Fig 3.8.

Fitting the model

Now, that we have tested for multicollinearity and made sure it is not present anymore, we will be fitting in the model.

The below is ANOVA table containing an analysis of variances. It can be used to test the following null and alternative hypothesis:

H_0: Model with only a constant term is a good fit for the data
H_1: Model fitted fits better

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4899.850	1	4899.850	19.608	<.001[b]
	Residual	13993.811	56	249.889		
	Total	18893.661	57			

a. Dependent Variable: life_span

b. Predictors: (Constant), brain_wt

Fig 3.15: ANOVA table

As can be seen in Fig 3.15 above, the p-value is less than 0.05, meaning that H_0 can be rejected, meaning that the fitted model is better than a model with a constant term.

The general regression model's equation is as follows:
$\mathbf{Y = \beta_0 + \beta_1 X_1 + \varepsilon}$

Where Y is equal to "life_span", β_0 is equal to the constant and β_1 is equal to "brain_wt". Using the table in Fig 16, the following hypotheses will be tested to be able to construct the general equation.

H_0: $\beta_i = 0$
H_1: $\beta_i \neq 0$

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	16.956	2.178		7.785	<.001
	brain_wt	.010	.002	.509	4.428	<.001

a. Dependent Variable: life_span

Fig 3.16: Coefficients' table

From the above Fig 3.16, it can be concluded that both H_0 can be rejected and therefore, the fitted model is as follows:

life_span = 16.956 + 0.1X_1

As can be seen in the model summary in Fig 3.17 below, the R value is 0.509. As a rule of thumb, the closer this value is to 1, the better the prediction of the model would be. Therefore a value of .509 shows that the model might not be able to predict the life span very well when using the brain weight.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.509[a]	.259	.246	15.8079	1.947

a. Predictors: (Constant), brain_wt

b. Dependent Variable: life_span

Fig 3.17: Model summary

The next steps involves checking some more assumptions, namely whether there are any independent residuals, whether there are any influential outliers, whether residuals follow a normal distribution and whether there is constant variance across observations.

Assumption 4: Independent Residuals

In order to test whether all residuals are independent of each other, the Durbin-Watson test is used. Usually, values outside the range of 1.5 and 2.5 are assumed to be a cause for concern. Therefore, since the value in Fig 3.10 is close to 2, we concluded that the assumption of having independent residuals was satisfied.

Model Summary[b]

Model	Durbin-Watson
1	1.947[a]

a. Predictors: (Constant), brain_wt

b. Dependent Variable: life_span

Fig 3.10: Durbin-Watson

Assumption 5: No influential outliers

In order to test this assumption, we have to apply outlier diagnostics after the model has been fitted. We will be looking at Leverage values, Cook's distance and stduentized residuals to find any possible outliers. The diagnostic can be performed by looking at the residual statistics in Fig 3.11 below.

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	16.957	72.159	19.878	9.2716	58
Std. Predicted Value	-.315	5.639	.000	1.000	58
Standard Error of Predicted Value	2.076	11.988	2.464	1.609	58
Adjusted Predicted Value	16.572	117.581	20.556	14.0580	58
Residual	-33.5594	70.2871	.0000	15.6686	58
Std. Residual	-2.123	4.446	.000	.991	58
Stud. Residual	-3.257	4.531	-.017	1.056	58
Deleted Residual	-78.9808	72.9864	-.6781	18.6620	58
Stud. Deleted Residual	-3.585	5.642	.000	1.166	58
Mahal. Distance	.001	31.798	.983	4.886	58
Cook's Distance	.000	7.178	.138	.942	58
Centered Leverage Value	.000	.558	.017	.086	58

a. Dependent Variable: life_span

Fig 3.11: Residual Statistics

The first diagnostic is done by observing the Leverage values. The influential point can be calculated by 2(2)/58, which results in 0.069. The maximum leverage value is 0.558, which indicates that there are potential influential points since it is greater than the 0.069. By looking at the LEV_1 values and sorting them in descending order, we have identified 2 points which are potential influential points as shown in Fig 3.12 below.

SRE_1	MAH_1	COO_1	LEV_1
-3.25682	31.79758	7.17801	.55785
.60232	20.09670	.10645	.35257
4.53090	1.12530	.39420	.01974
.28579	.15498	.00083	.00272
1.45118	.13514	.02106	.00237
-.91697	.09922	.00813	.00174
.44975	.09915	.00196	.00174

Fig 3.12: Influential points

However, only 1 of those points had a Cook's distance greater than 1. Therefore, only that point remains as a potential influential point. In addition, Fig 3.12 also shows that the same point has a studentized residual smaller than 2. Consequently, that data point is considered to be an outlier since it fails more than 2 diagnostics. Therefore that point should be removed and the model should be fitted again to see whether there is a considerable change in the resulting paarmeters.

Assumption 6: Residuals following a normal distribution

This assumption was tested by a Shapiro-Wilk test on the studentized residuals variable SRE_1. As shown in Fig 3.13 below, the test was for the hypotheses below and the resulting

p-value was smaller than 0.05 and therefore, H_0 had to be rejected. This means that this assumption was not satisfied and that the model cannot be considered valid.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Studentized Residual	.170	58	<.001	.865	58	<.001

a. Lilliefors Significance Correction

Fig 3.13: Residual normality test

H_0: Variable follows a normal distribution
H_1: Variable does not follow a normal distribution

Assumption 7: Constant variance across observations

The final assumption which needs to be satisfied is that variance is constant across observations and this is tested using a scatter plot of residual values against fitted ones. The scatter plot in Fig 3.14 below shows that there is a pattern in the relationship, indicating that residuals are not homoscedastic. Consequently, this assumption was also unsatisfied, making the model invalid.

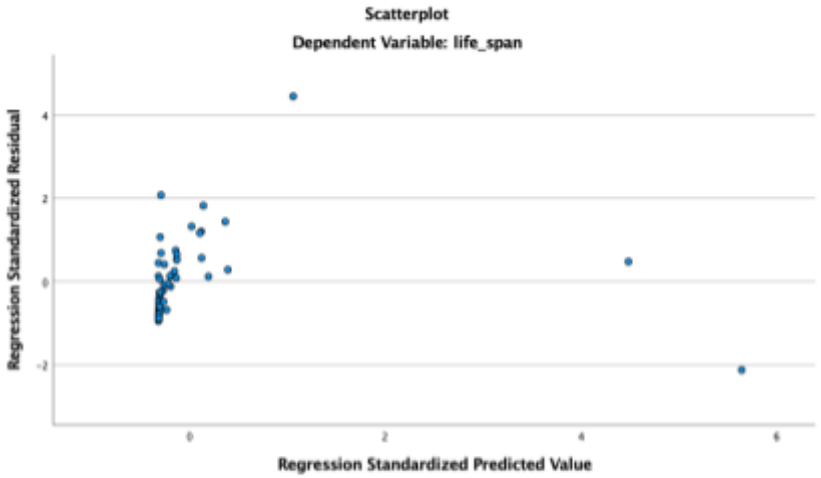
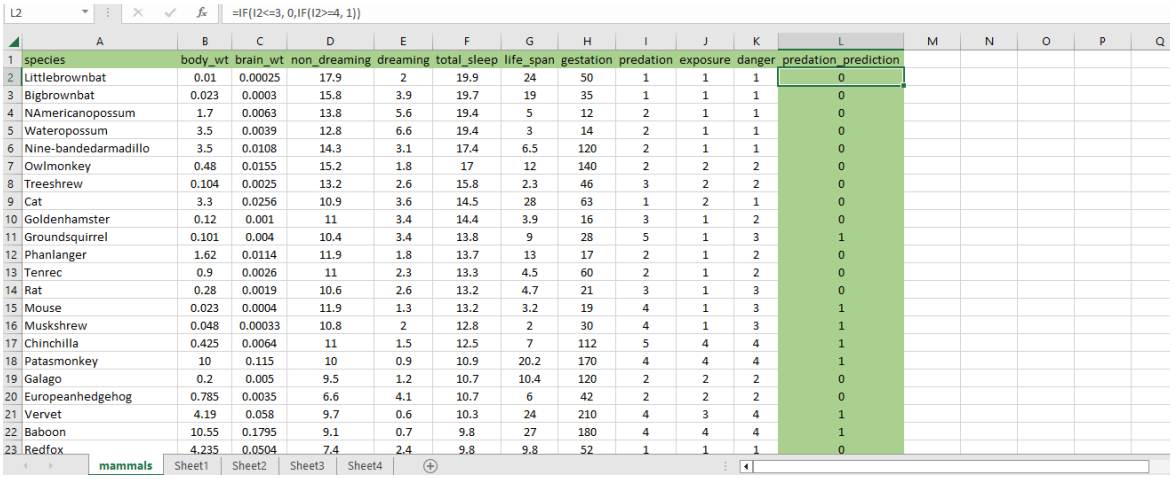


Fig 3.14: Scatter plot - life_span vs residual

Binary Logistic Regression Model

To fit a binary logistic regression we firstly have to prepare our data set for this task. A dummy variable is added to the dataset to represent factors which are of a binary nature i.e. they are either observed or not observed and so that the fitting of Regression Model is possible. A dummy variable is a binary variable that takes a value of 0 or 1 which in this case is predation_prediction.



L2 =IF(I2<=3, 0,IF(I2>=4, 1))

	species	body_wt	brain_wt	non_dreaming	dreaming	total_sleep	life_span	gestation	predation	exposure	danger	predation_prediction
2	Littlebrownbat	0.01	0.00025	17.9	2	19.9	24	50	1	1	1	0
3	Bigbrownbat	0.023	0.0003	15.8	3.9	19.7	19	35	1	1	1	0
4	NAmericanopossum	1.7	0.0063	13.8	5.6	19.4	5	12	2	1	1	0
5	Wateropossum	3.5	0.0039	12.8	6.6	19.4	3	14	2	1	1	0
6	Nine-bandedarmadillo	3.5	0.0108	14.3	3.1	17.4	6.5	120	2	1	1	0
7	Owlmonkey	0.48	0.0155	15.2	1.8	17	12	140	2	2	2	0
8	Treeshrew	0.104	0.0025	13.2	2.6	15.8	2.3	46	3	2	2	0
9	Cat	3.3	0.0256	10.9	3.6	14.5	28	63	1	2	1	0
10	Goldenhamster	0.12	0.001	11	3.4	14.4	3.9	16	3	1	2	0
11	Groundsquirrel	0.101	0.004	10.4	3.4	13.8	9	28	5	1	3	1
12	Phanlanger	1.62	0.0114	11.9	1.8	13.7	13	17	2	1	2	0
13	Tenrec	0.9	0.0026	11	2.3	13.3	4.5	60	2	1	2	0
14	Rat	0.28	0.0019	10.6	2.6	13.2	4.7	21	3	1	3	0
15	Mouse	0.023	0.0004	11.9	1.3	13.2	3.2	19	4	1	3	1
16	Muskshrew	0.048	0.00033	10.8	2	12.8	2	30	4	1	3	1
17	Chinchilla	0.425	0.0064	11	1.5	12.5	7	112	5	4	4	1
18	Patasmonkey	10	0.115	10	0.9	10.9	20.2	170	4	4	4	1
19	Galago	0.2	0.005	9.5	1.2	10.7	10.4	120	2	2	2	0
20	Europeanhedgehog	0.785	0.0035	6.6	4.1	10.7	6	42	2	2	2	0
21	Vervet	4.19	0.058	9.7	0.6	10.3	24	210	4	3	4	1
22	Baboon	10.55	0.1795	9.1	0.7	9.8	27	180	4	4	4	1
23	Redfox	4.235	0.0504	7.4	2.4	9.8	9.8	52	1	1	1	0

mammals Sheet1 Sheet2 Sheet3 Sheet4

The states of 0 and 1 are determined with the following function:

=IF(Predation_Index <=3, 0,IF((Predation_Index >= 4, 1))

This step provided us with a new variable predation_prediction which will be used for the prediction of the binary logistic regression. The next important step is to check for multicollinearity when performing a binary logistic regression.

Multicollinearity occurs when two or more predictor variables are highly correlated, which can cause unstable and inconsistent coefficient estimates. This can lead to difficulties in interpreting the results of the model and may affect the predictive power of the model.

There are several methods that can be used to test for multicollinearity, including:

Variance Inflation Factor (VIF): This measures the amount of multicollinearity in a multiple regression model. A VIF value greater than 10 is considered to indicate multicollinearity.

Tolerance: This is defined as the reciprocal of the VIF, and a value less than 0.1 indicates multicollinearity.

If multicollinearity is detected, there are several ways to address it, such as removing one of the correlated predictor variables, combining the correlated variables into a single composite variable, or using a method such as ridge regression, which is resistant to multicollinearity.

SPSS performed the linear regression and generated a collinearity diagnostics table. The table included the variance inflation factor (VIF) for each predictor variable, as well as the tolerance and condition index.

I examined the VIF values. Any values greater than 10 indicate multicollinearity. I also examined the tolerance values and looked for any values that are less than 0.1, which would also indicate multicollinearity. Since multicollinearity is detected, I removed the correlated predictor variables body_wt, brain_wt, predation, danger and kept non_dreaming, dreaming, life_span and exposure.

With this information in hand we therefore can reliably develop a binary logistic regression model using non_dreaming, dreaming, life_span and exposure variables as independent variables and predict the predation_prediction variable since this variable is depended on the other independent variable. As shown in the exploratory data analysis an life an animal is at high risk of predation if the the variables non_dreaming, dreaming, life_span and exposure are high.

Regression

```
[DataSet1]
```

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	danger, life_span, body_wt, non_dreaming, dreaming, exposure, predation, brain_wt[b]	.	Enter

a. Dependent Variable: predation_prediction

b. Tolerance = .000 limit reached.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.920[a]	.847	.810	.214

a. Predictors: (Constant), danger, life_span, body_wt, non_dreaming, dreaming, exposure, predation, brain_wt

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8.392	8	1.049	22.886	<.001[b]
	Residual	1.513	33	.046		
	Total	9.905	41			

a. Dependent Variable: predation_prediction

b. Predictors: (Constant), danger, life_span, body_wt, non_dreaming, dreaming, exposure, predation, brain_wt

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	-.801	.208		-3.841	<.001		
	body_wt	-5.133E-5	.000	-.042	-.123	.903	.039	25.335
	brain_wt	-.126	.251	-.189	-.505	.617	.033	30.184
	non_dreaming	.015	.012	.117	1.223	.230	.507	1.973
	dreaming	-.008	.037	-.022	-.211	.834	.429	2.331
	life_span	.006	.003	.246	1.823	.077	.254	3.935
	predation	.337	.086	.992	3.909	<.001	.072	13.923
	exposure	.063	.047	.196	1.356	.184	.221	4.518
	danger	-.060	.119	-.169	-.500	.620	.041	24.565

a. Dependent Variable: predation_prediction

Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics		
						Tolerance	VIF	Minimum Tolerance
1	total_sleep	.[b]	.	.	.	.000	.	.000

a. Dependent Variable: predation_prediction

b. Predictors in the Model: (Constant), danger, life_span, body_wt, non_dreaming, dreaming, exposure, predation, brain_wt

Binary logistic regression in SPSS:

SPSS generated a variety of output, including the regression coefficients, odds ratios, and model fit statistics.The outputs of a binary logistic regression model included the following:

Regression coefficients: These represent the change in the log odds of the outcome for a one unit change in the predictor variable, holding all other variables constant. The coefficients can be used to identify the relationship between each predictor variable and the outcome.

Odds ratios: These represent the change in the odds of the outcome for a one unit change in the predictor variable, holding all other variables constant. An odds ratio greater than 1 indicates that an increase in the predictor is associated with an increase in the odds of the outcome, while an odds ratio less than 1 indicates that an increase in the predictor is associated with a decrease in the odds of the outcome.
Model fit statistics: These include measures such as the null deviance, residual deviance, and AIC, which can be used to evaluate the fit of the model. A lower null deviance and residual deviance, and a lower AIC, indicate a better fit.

Classification table: This table shows the number of observations that were correctly and incorrectly classified by the model. It includes measures such as sensitivity, specificity, and overall accuracy, which can be used to evaluate the performance of the model.

To interpret the outputs of a binary logistic regression model, I first examined the regression coefficients and odds ratios to identify the relationship between each predictor variable and the outcome, then I used the model fit statistics and classification table to evaluate the fit and performance of the model.

Logistic Regression

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	42	100.0
	Missing Cases	0	.0
	Total	42	100.0
Unselected Cases		0	.0
Total		42	100.0

a. If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding

Original Value	Internal Value
0	0
1	1

Block 0: Beginning Block

Iteration History[a,b,c]

Iteration		-2 Log likelihood	Coefficients Constant
Step 0	1	55.821	-.476
	2	55.820	-.485
	3	55.820	-.486

a. Constant is included in the model.

b. Initial -2 Log Likelihood: 55.820

c. Estimation terminated at iteration number 3 because parameter estimates changed by less than .001.

Classification Table[a,b]

Observed			Predicted: predation_prediction 0	Predicted: predation_prediction 1	Percentage Correct
Step 0	predation_prediction	0	26	0	100.0
		1	16	0	.0
	Overall Percentage				61.9

a. Constant is included in the model.

b. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-.486	.318	2.335	1	.127	.615

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	dreaming	6.969	1	.008
		exposure	16.490	1	<.001
		non_dreaming	3.064	1	.080
		life_span	.064	1	.800
	Overall Statistics		20.381	4	<.001

Block 1: Method = Enter

Iteration History[a,b,c,d]

Iteration		-2 Log likelihood	Coefficients				
			Constant	dreaming	exposure	non_dreaming	life_span
Step 1	1	33.611	-2.135	-.286	.882	.070	-.025
	2	28.720	-2.866	-.468	1.416	.110	-.067
	3	27.598	-3.468	-.491	1.847	.122	-.100
	4	27.491	-3.723	-.477	2.032	.123	-.114
	5	27.490	-3.755	-.475	2.056	.123	-.116
	6	27.490	-3.755	-.475	2.056	.123	-.116

a. Method: Enter

b. Constant is included in the model.

c. Initial -2 Log Likelihood: 55.820

d. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	28.331	4	<.001
	Block	28.331	4	<.001
	Model	28.331	4	<.001

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	27.490[a]	.491	.667

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	9.678	8	.288

Contingency Table for Hosmer and Lemeshow Test

		predation_prediction = 0		predation_prediction = 1		
		Observed	Expected	Observed	Expected	Total
Step 1	1	4	3.966	0	.034	4
	2	4	3.861	0	.139	4
	3	3	3.766	1	.234	4
	4	4	3.646	0	.354	4
	5	4	3.563	0	.437	4
	6	2	3.033	2	.967	4
	7	4	2.107	0	1.893	4
	8	1	1.428	3	2.572	4
	9	0	.499	4	3.501	4
	10	0	.131	6	5.869	6

Classification Table[a]

			Predicted		
			predation_prediction		Percentage Correct
	Observed		0	1	
Step 1	predation_prediction	0	24	2	92.3
		1	3	13	81.3
	Overall Percentage				88.1

a. The cut value is .500

Overall, the binary logistic regression model appears to be performing well in terms of predicting the dependent variable "predation_prediction," which can take on either a 0 or 1 value. The model has a relatively high level of explanatory power, as indicated by the R squared and adjusted R squared values of 0.847 and 0.810, respectively. Additionally, the model is making a relatively high percentage of correct predictions for both values of the dependent variable, with 92.3% correct predictions for the value 0 and 81.3% correct predictions for the value 1.

Conclusion

The following shall serve as a conclusion and therein act as a summary, of the analysis, interpretation, and findings of the research conducted for this assignment. Similar to the structure of the assignment, the conclusion reflects on two core areas of research; (i) descriptive statistics and (ii) hypothesis testing.

Exploratory Data Analysis - Descriptive Statistics

During this chapter, our focus was on the analysis of the selected data set, Sleep.csv. First and foremost, our research underlined correlation aspects between the variables of the data set, whereupon we designed exploratory tests and apt modeling. The coupling of data with the use of descriptive statistical techniques, this chapter lent itself to a reasonable interpretation of the data points, the existing correlations, and a discussion on complementary measures, accordingly.

Over this chapter, a large multiple of tests were undertaken on various data points of the data set including, but not limited to: (i) danger, (ii) exposure, (iii) predation; (iv) life span; (v) gestation; (vi) dreaming.

Numerous graphical techniques were used to ensure a comprehensive evaluation and application of descriptive statistics. All tallied, this chapter produced thirty-nine (39) unique descriptive visualizations.

Statistical Analysis - Hypothesis Testing

From the hypothesis testing conducted, the following may be noted:

- Normality Test 3A.1 - Both least preyed mammals and most preyed mammals follow a normal distribution
- Normality Test 3A.2 - Least exposed mammals follow a normal distribution, most exposed mammals observes a p-value smaller than 0.05.
- 3B: Test 1 - Levene's Test
    - There is a significant difference in the average total sleep time between mammals that are least and most preyed upon.
    - The mean total sleep is more for the least preyed upon mammals
    - Extrapolated further, the results indicate that mammals who sleep more are more aware and alert, thus being more difficult to be preyed upon.
    - Conversely, animals that sleep less may be discussed to be more easily pursued and attacked by predators.
- 3C: Test 2 - On average, exposed mammals tend to sleep less than mammals that sleep in a well-protected habitat.

- 3CA - Linear Regression Model - Based on the assumptions taken, and the regression model formula to be **life_span = 16.956 + 0.1X_1**, it is reflected that the linear regression model may not be of optimum calibration to predict the life span when using the brain weight variable. Seven (7) assumptions were undergone during this aspect of statistical analysis.
- 3CB - Binary Logistic Regression: The binary logistic regression model that was used in this study achieved an impressive 88.1% prediction rate. This high level of accuracy suggests that the model was able to effectively learn the relationship between the predictor variables and the binary outcome, and was able to make accurate predictions on new data. It is noted that not all assumptions were satisfied.

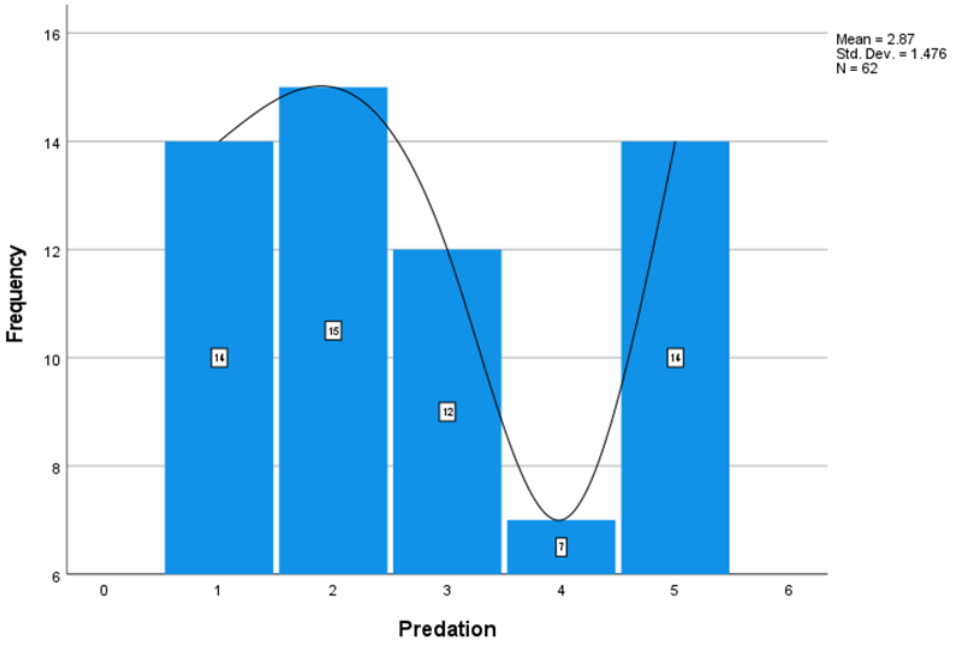
In closing, future research is recommended to be made on the identification of stronger variables to perform a more optimum calibrated linear regression model, from the available data points of the respective data set. Complementary recommendation should also be taken on the satisfaction of all assumptions when performing the linear regression model test/s.

References

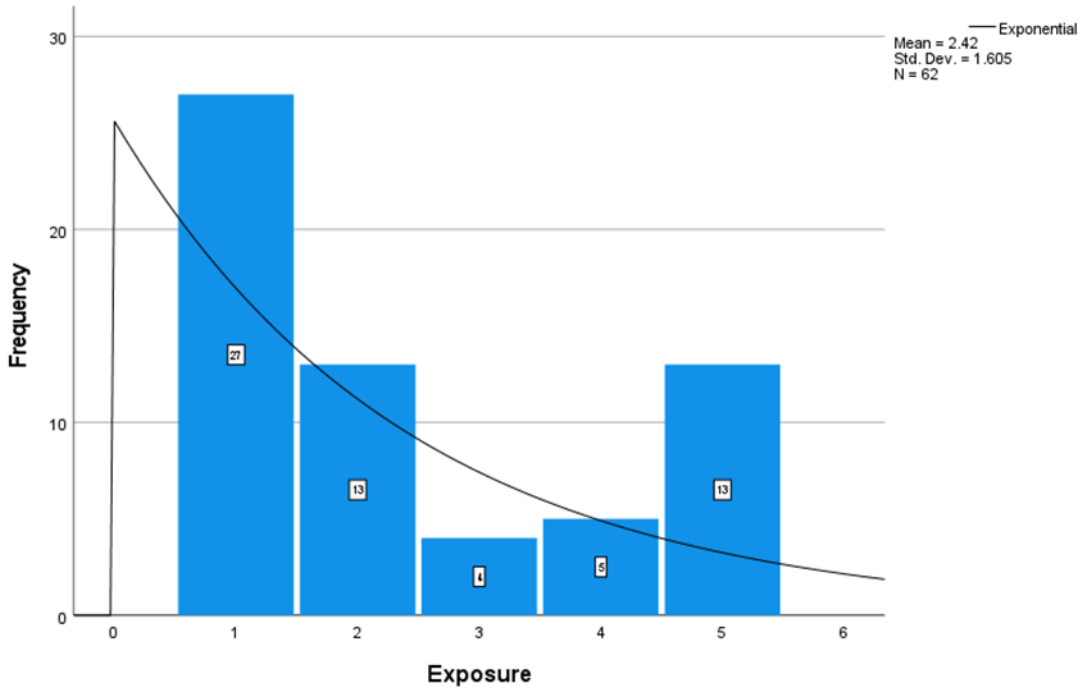
- '*Session 4 – Basic Statistics and Methods for Data Collection*', Dr. Inguanez, Dr Sammut, and Dr Suda

Appendix

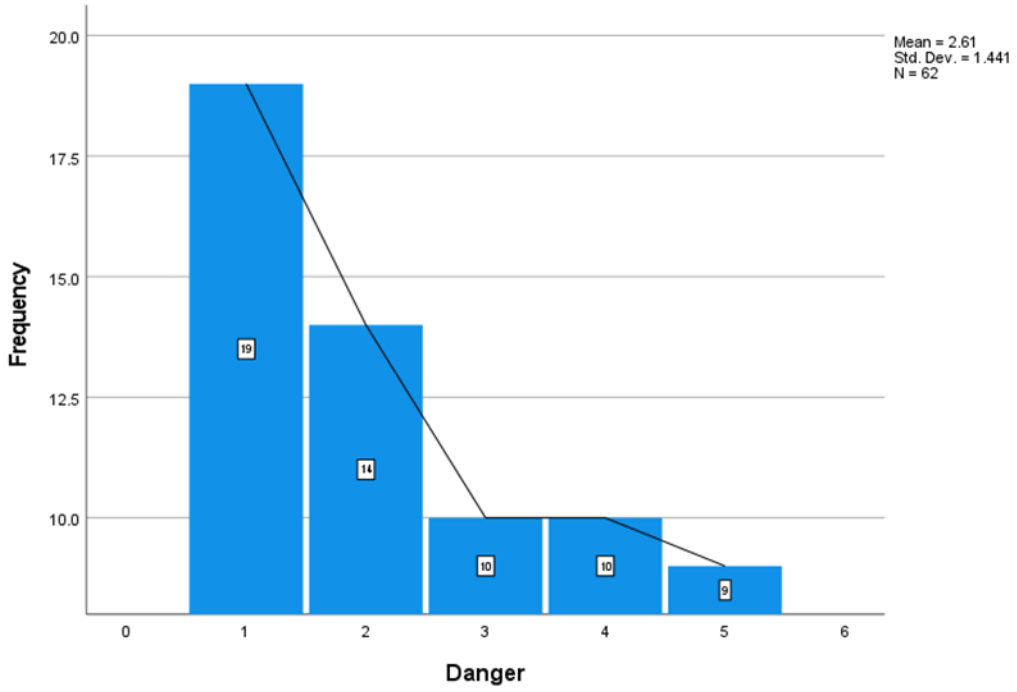
Appendix 1: Histogram of Predation with Interpolation Line



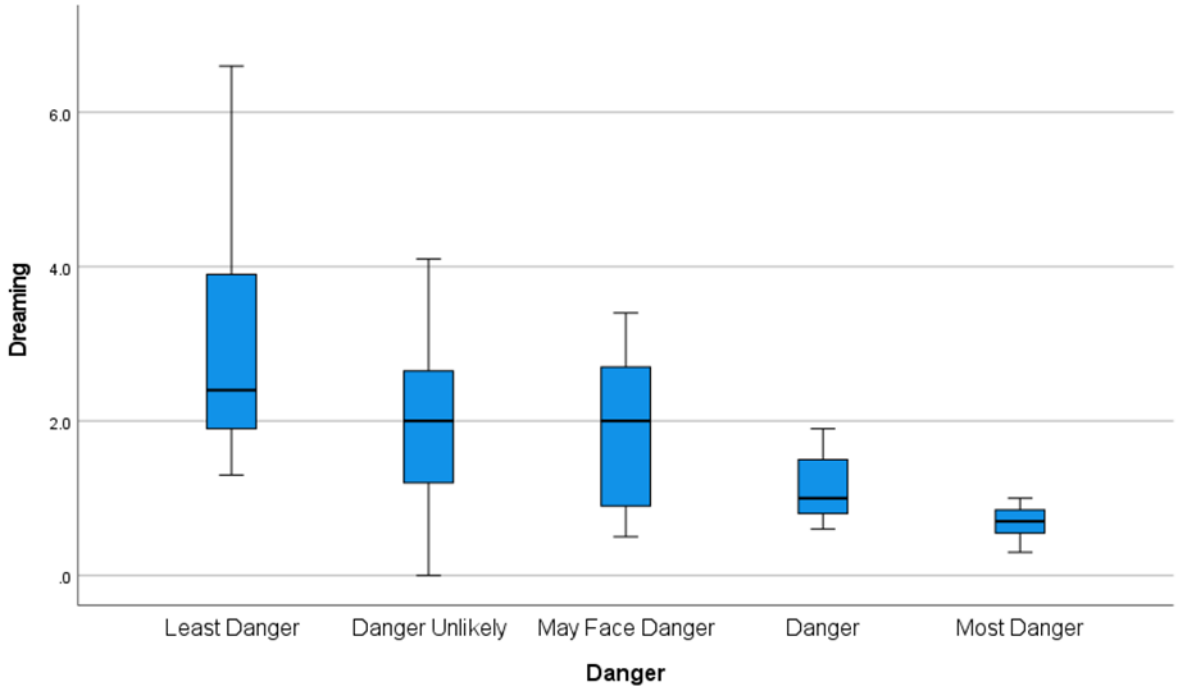
Appendix 2: Histogram of Exposure with Exponential Line



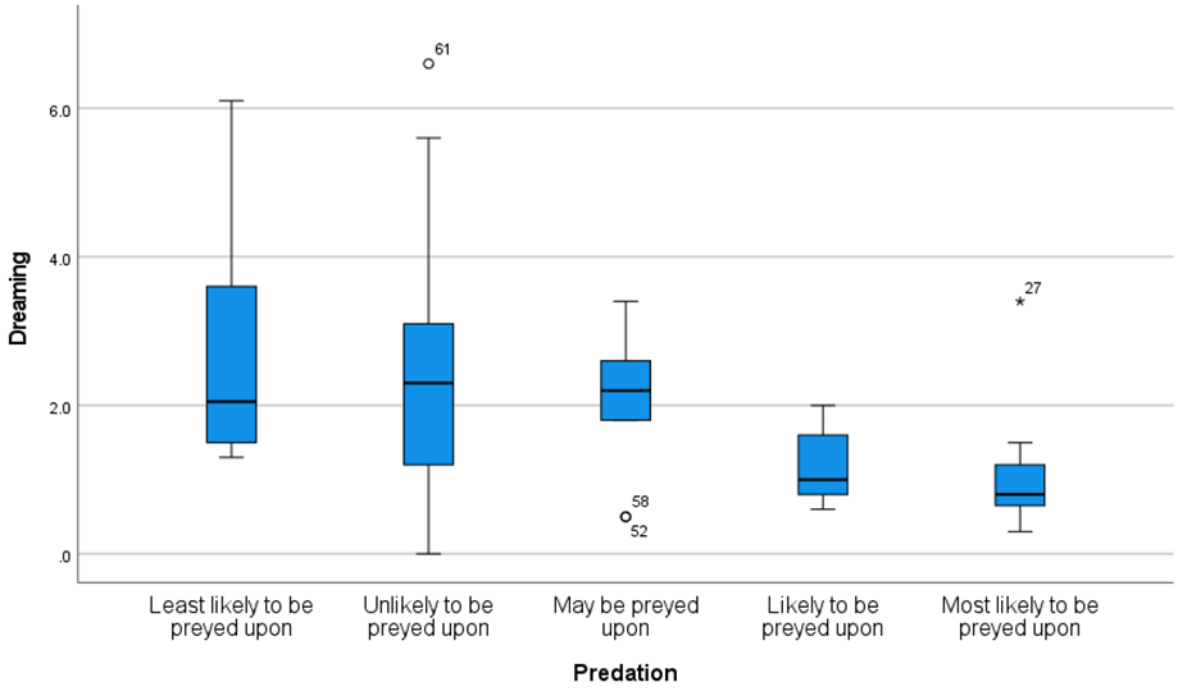
Appendix 3: Histogram of Danger with Interpolation Line



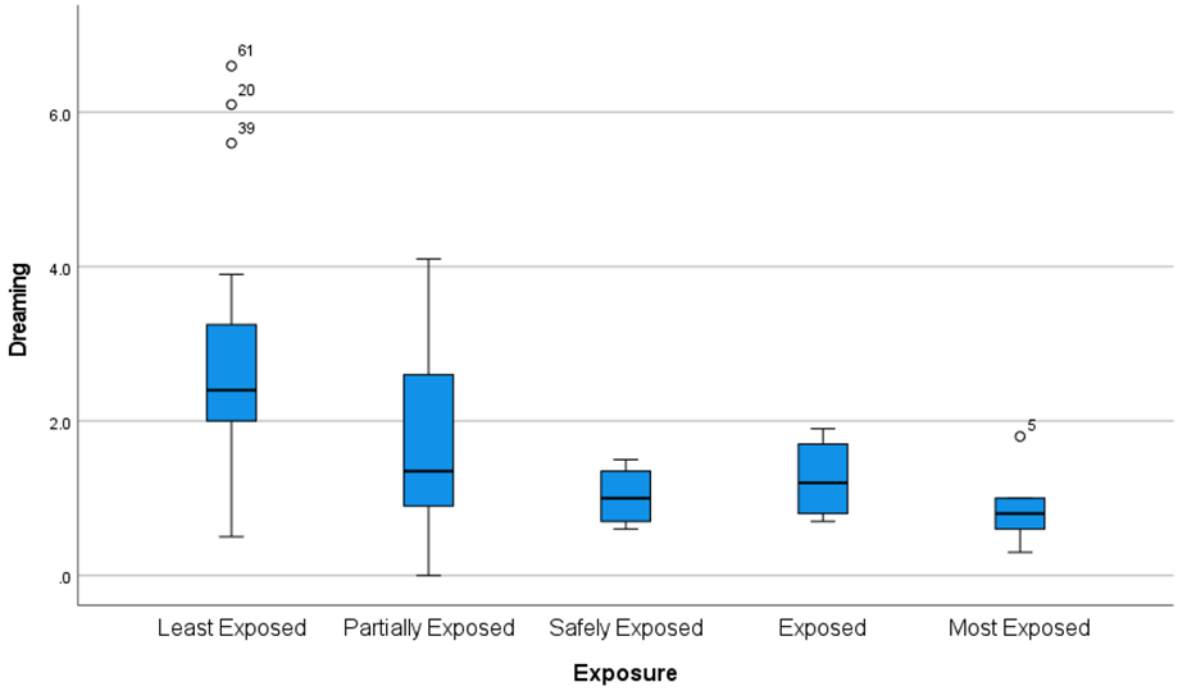
Appendix 4: Box Plot of Dreaming and Danger



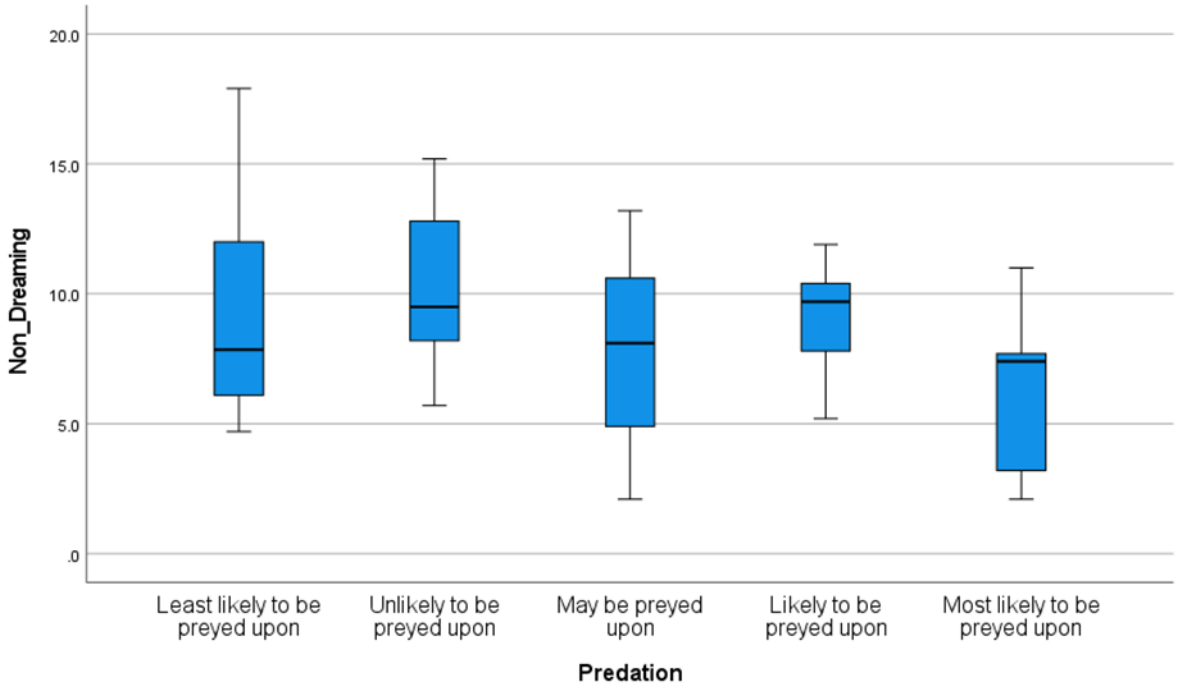
Appendix 5: Box Plot of Dreaming and Predation



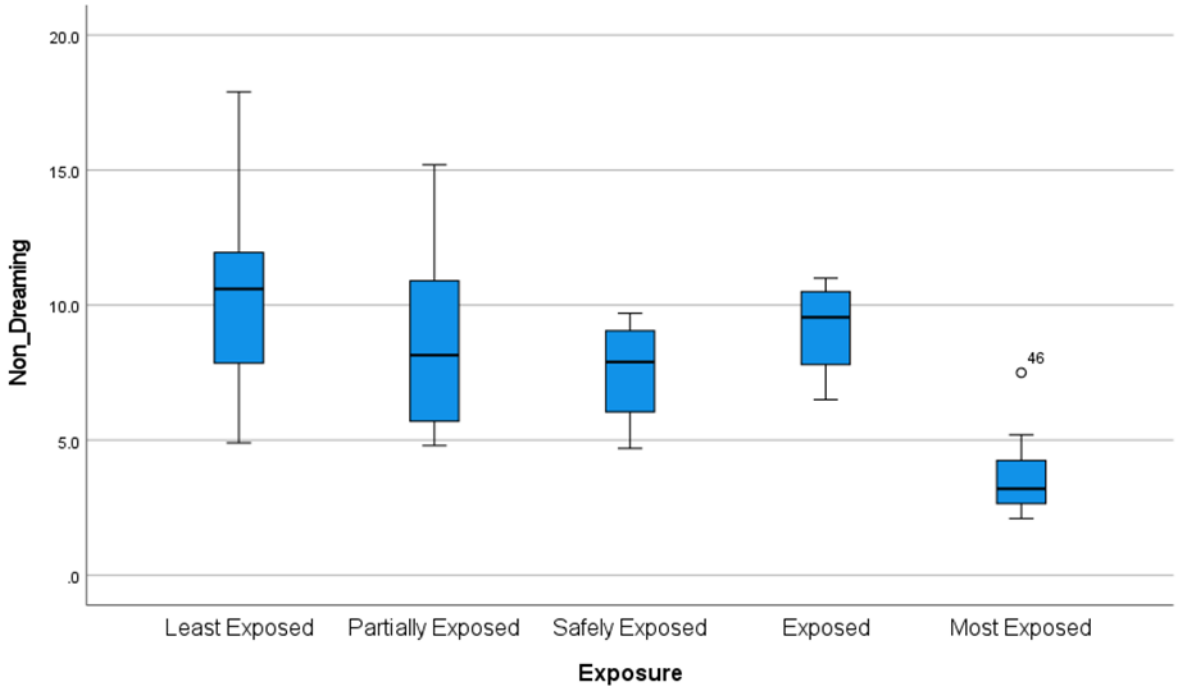
Appendix 6: Box plot of Dreaming and Exposure



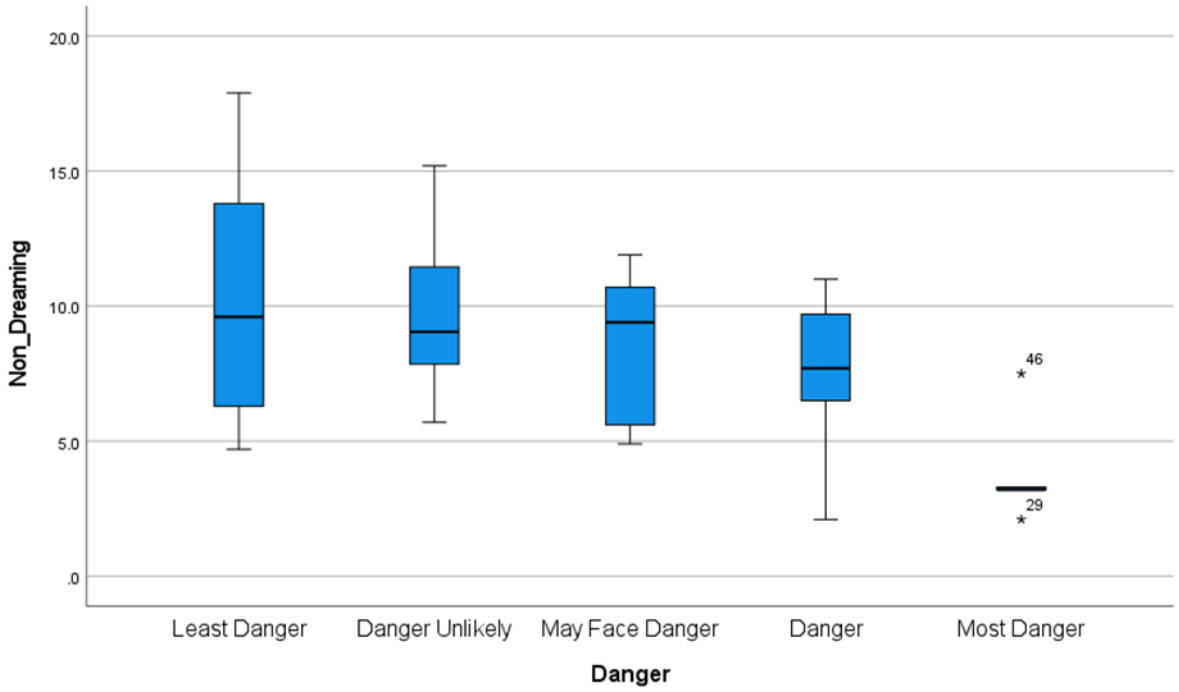
Appendix 7: Box Plot of Non Dreaming and Predation



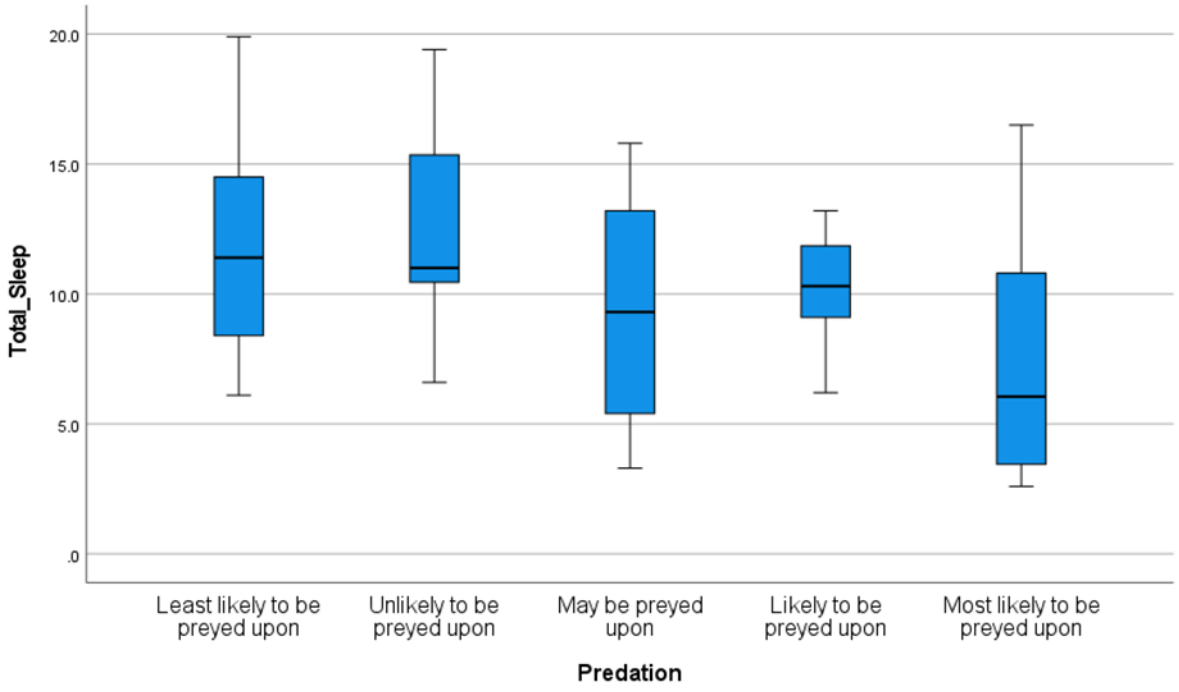
Appendix 8: Box Plot of Non Dreaming and Exposure



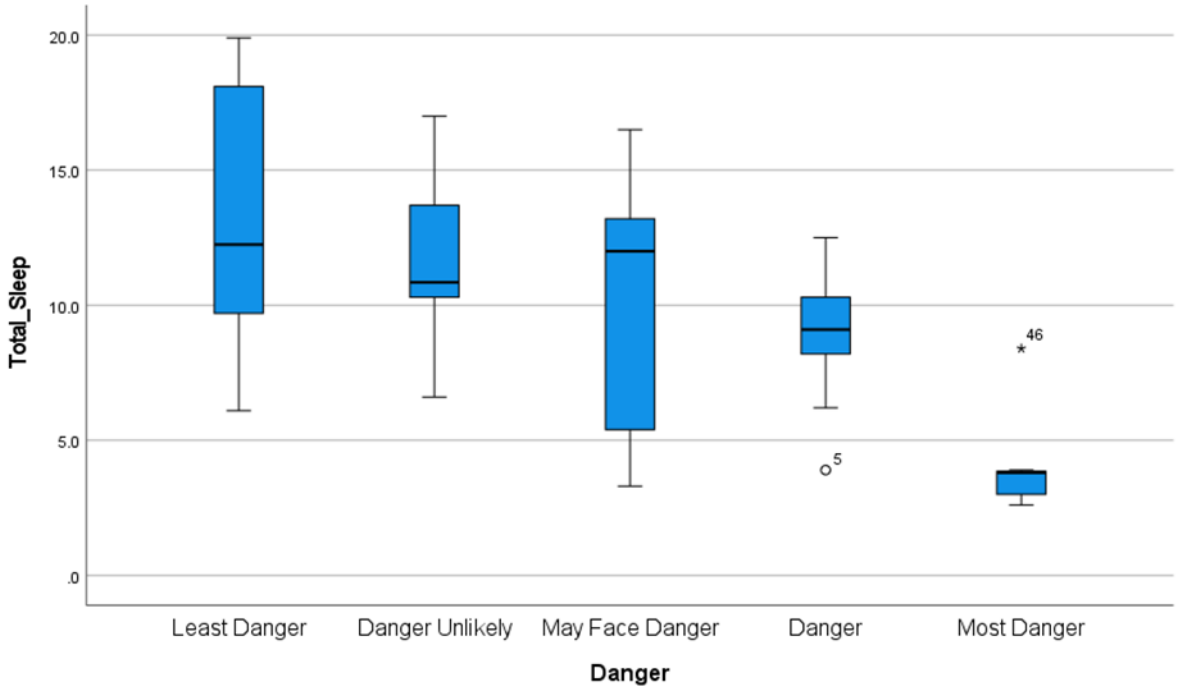
Appendix 9: Box Plot of Non Dreaming and Danger



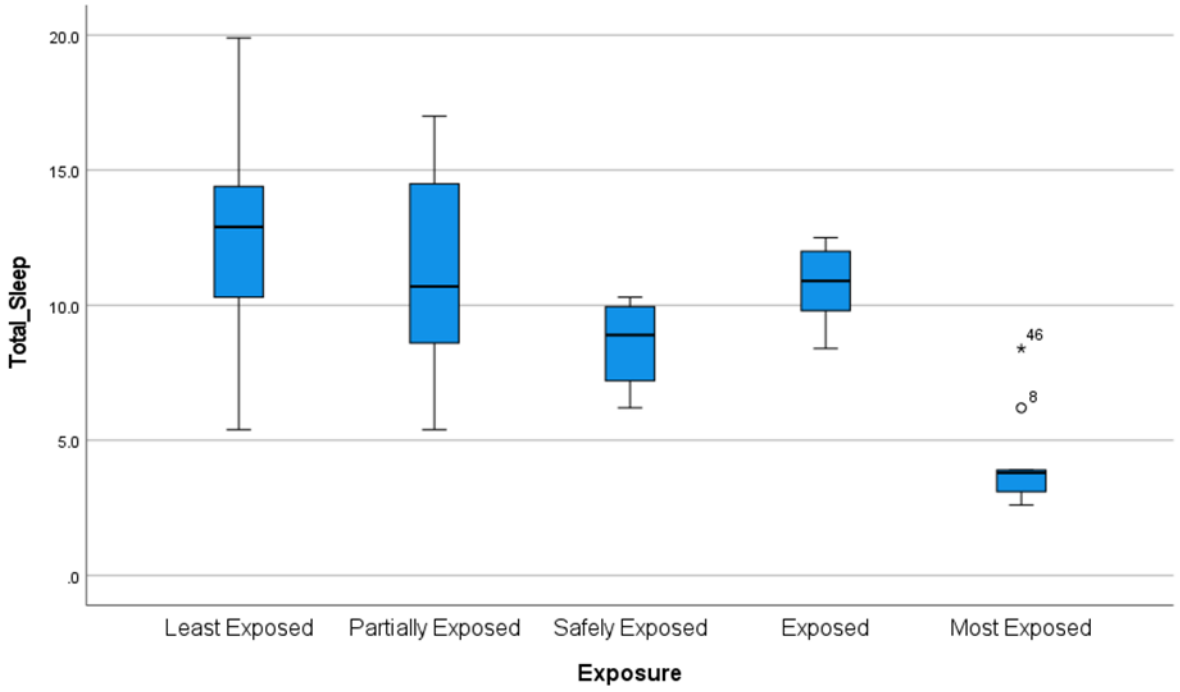
Appendix 10: Box Plot of Total Sleep and Predation



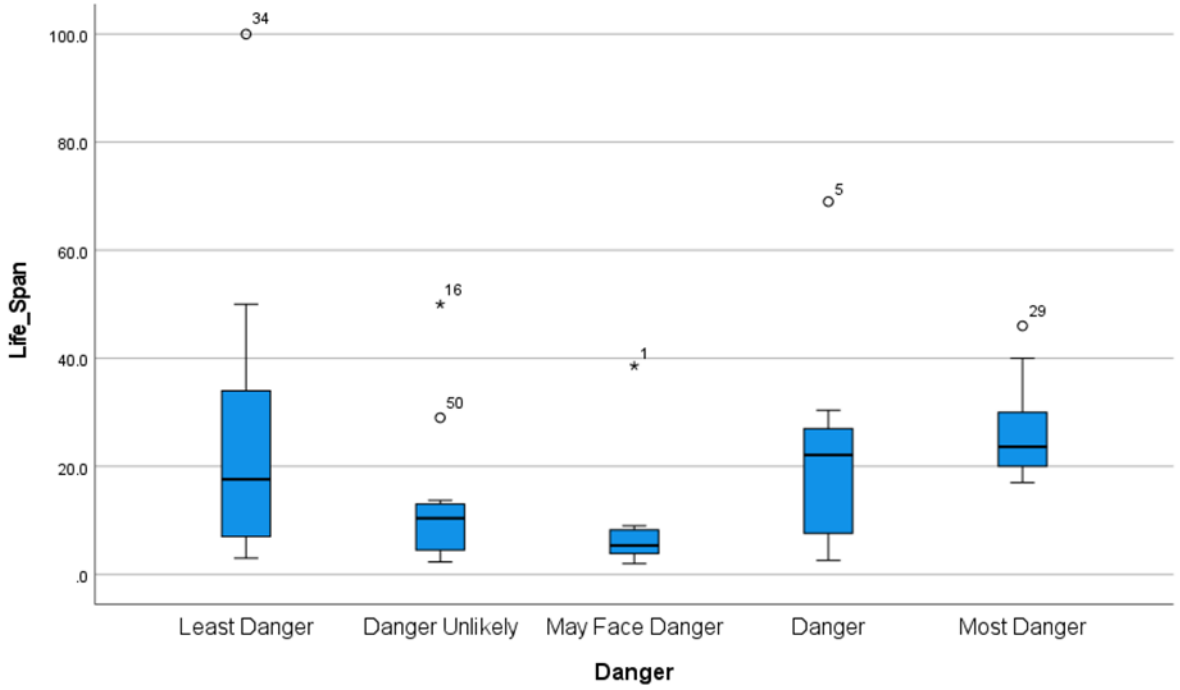
Appendix 11: Box Plot of Total Sleep and Danger



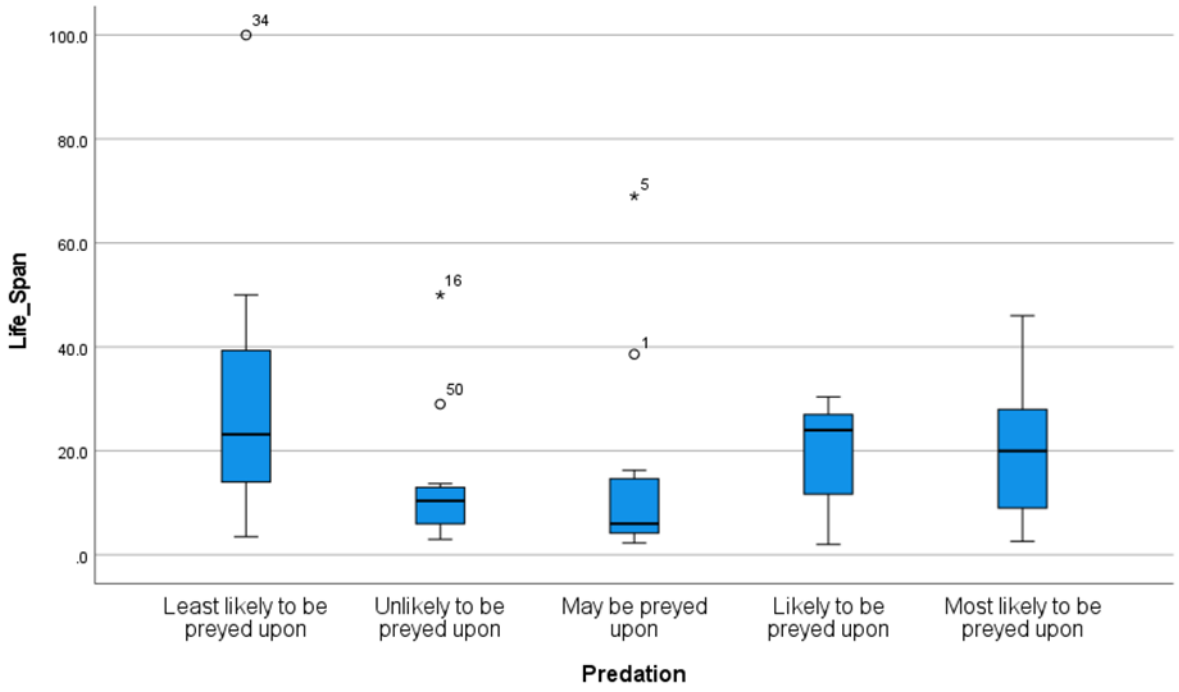
Appendix 12: Box Plot of Total sleep and Exposure



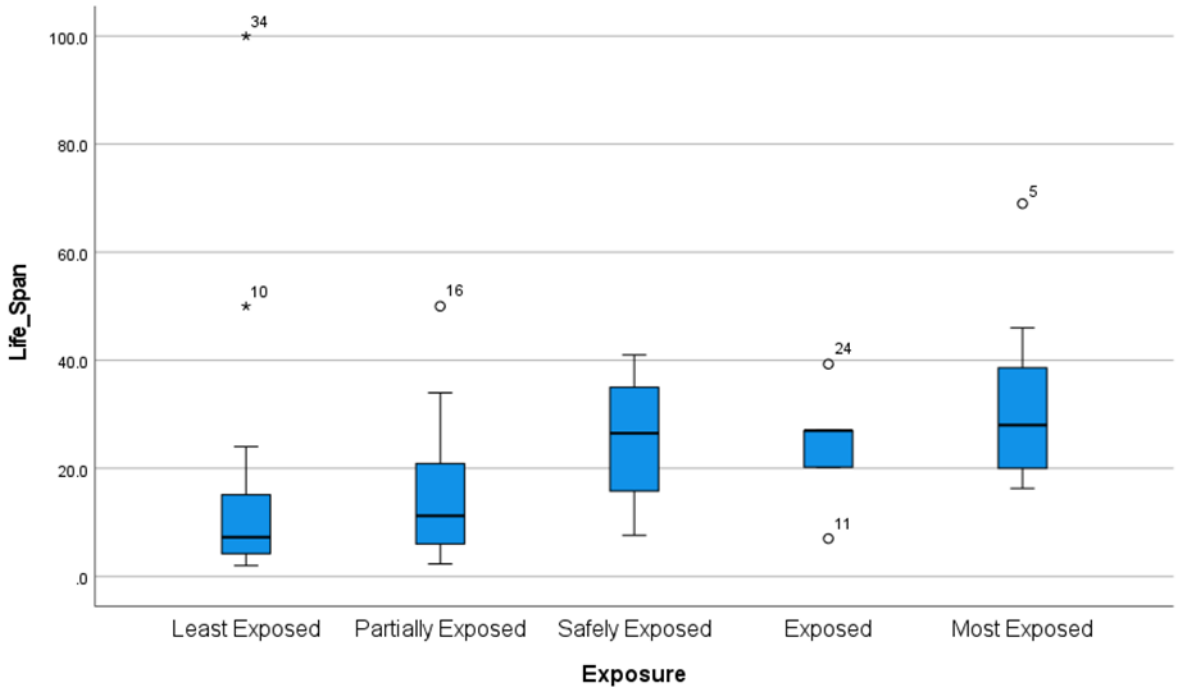
Appendix 13: Box Plot of Life Span and Danger



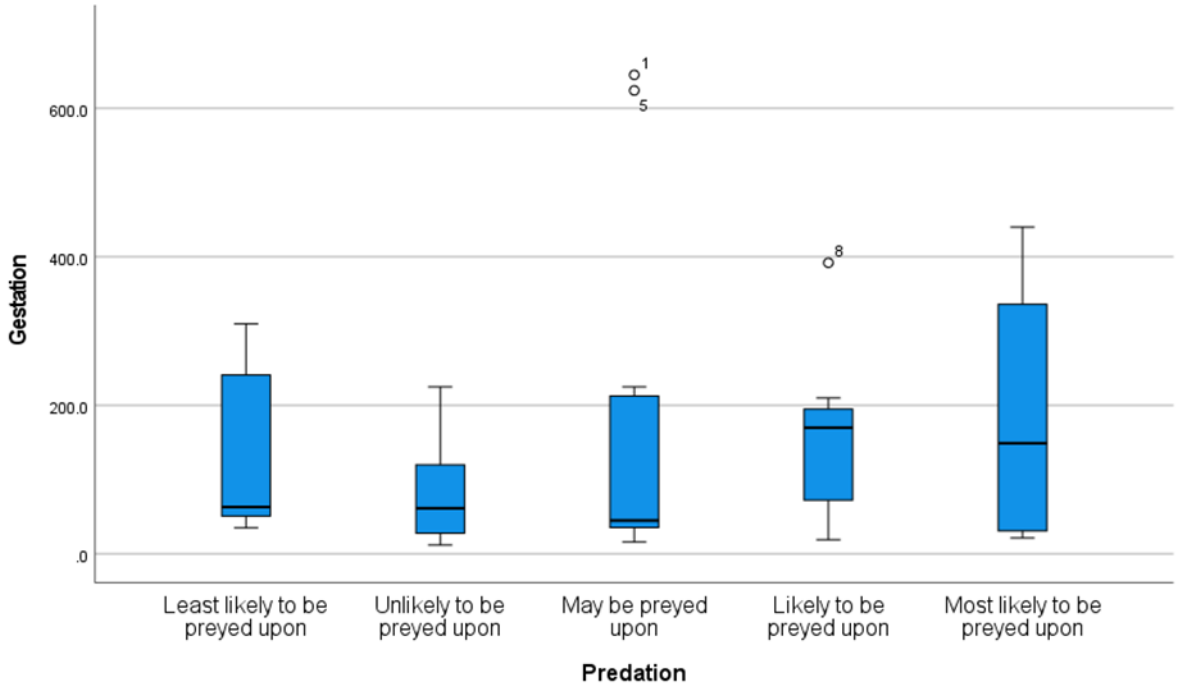
Appendix 14: Box Plot of Life Span and Predation



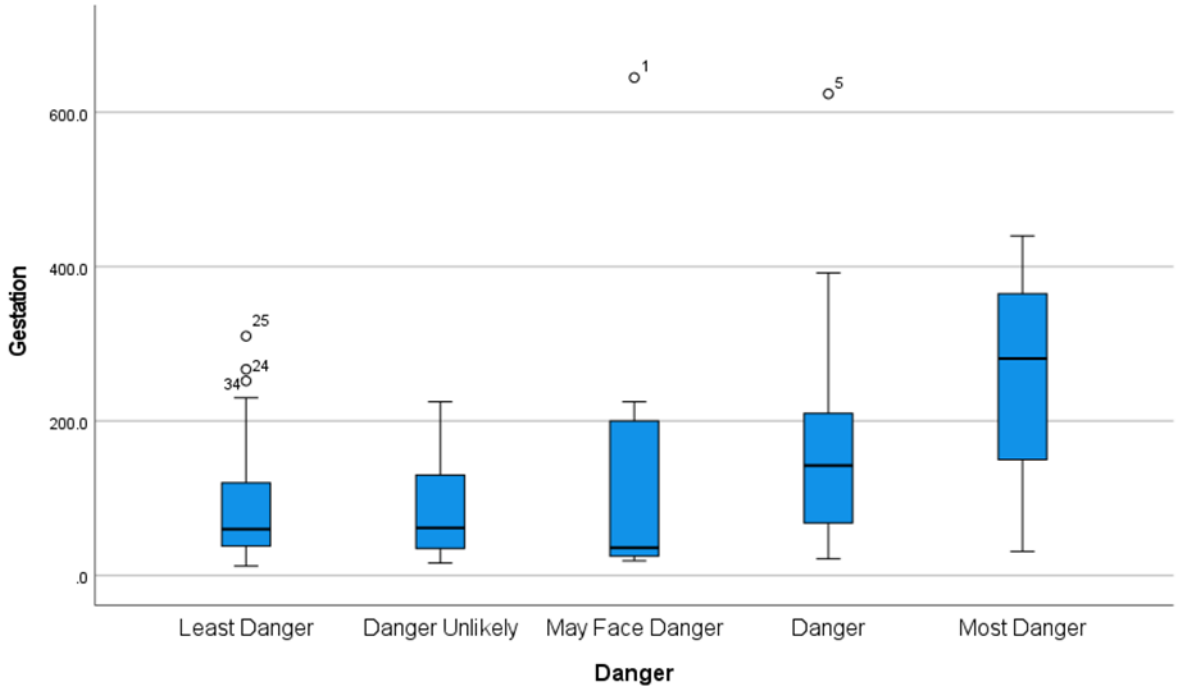
Appendix 15: Box Plot of Life Span and Exposure



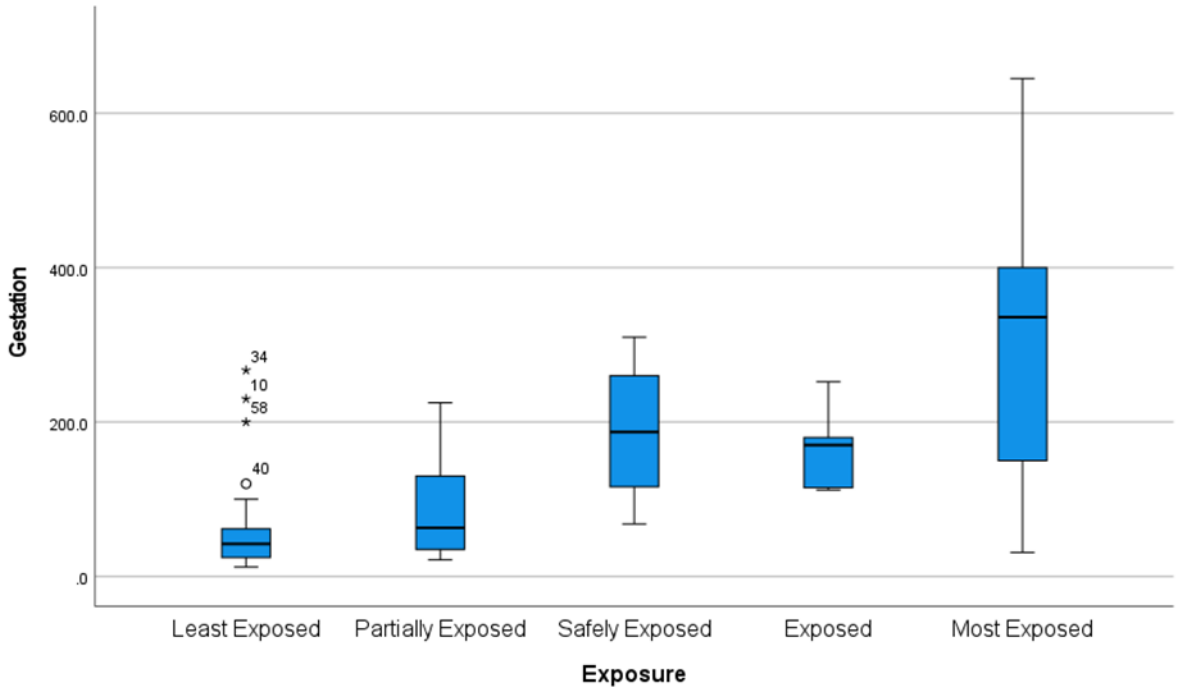
Appendix 16: Box Plot of Gestation and Predation



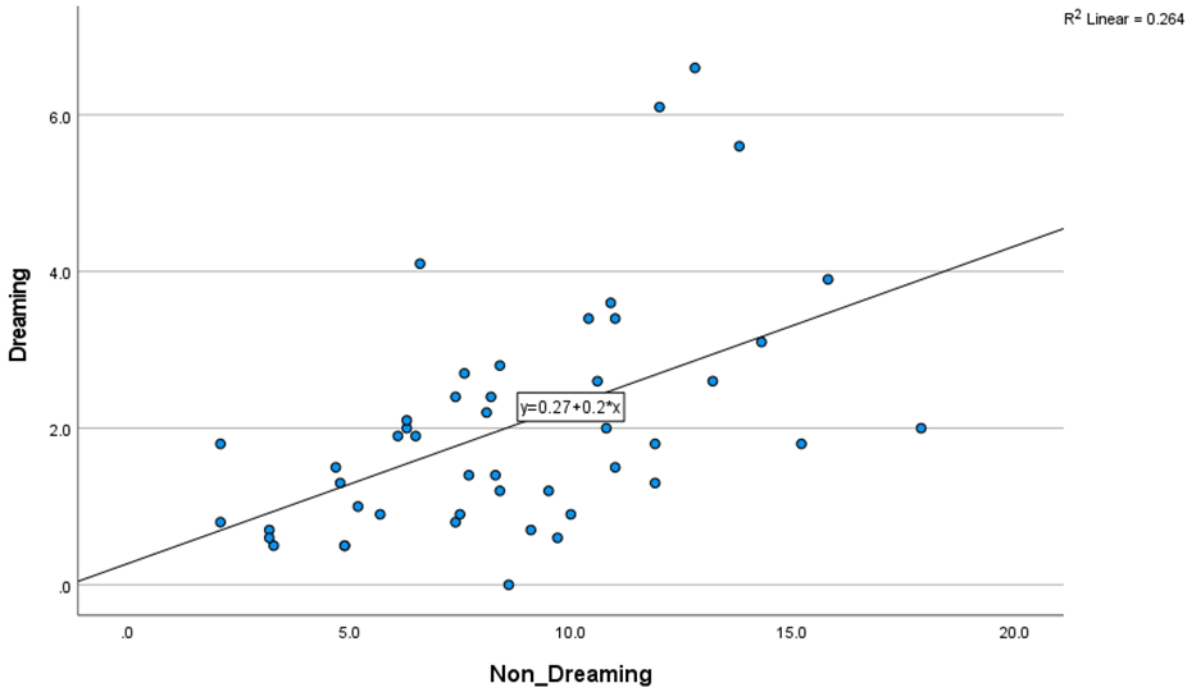
Appendix 17: Box Plot of Gestation and Danger



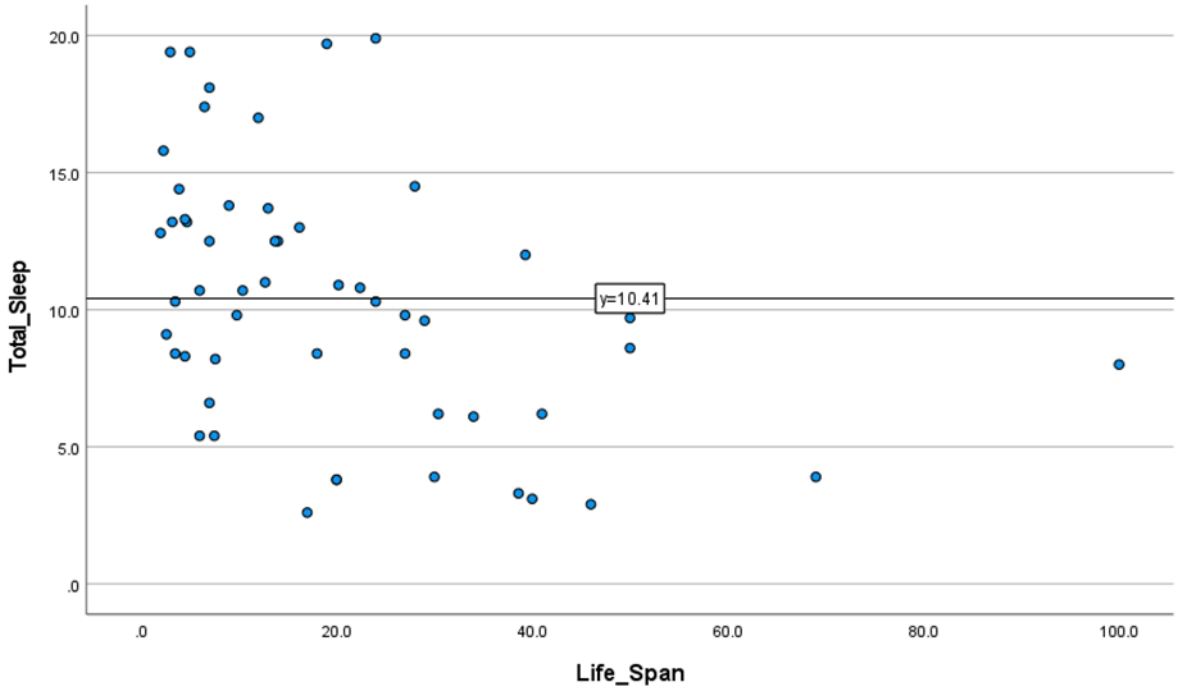
Appendix 18: Box Plot of Gestation and Exposure



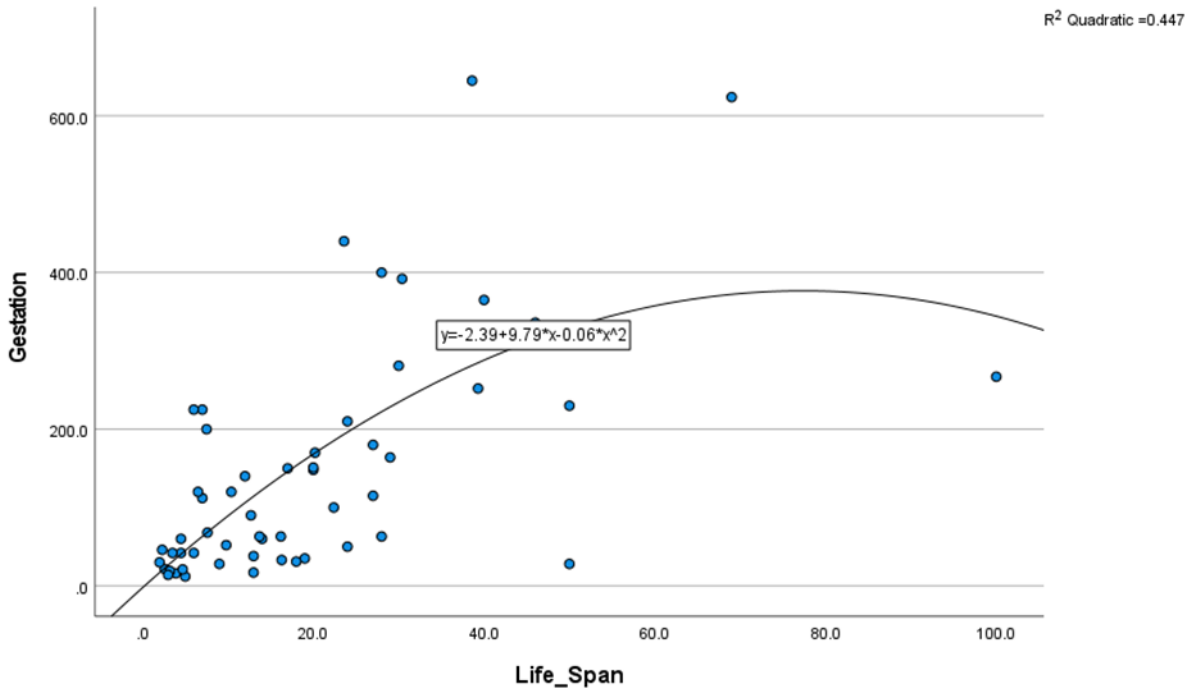
Appendix 19: Scatter Dot with Linear Line of Dreaming and Non Dreaming



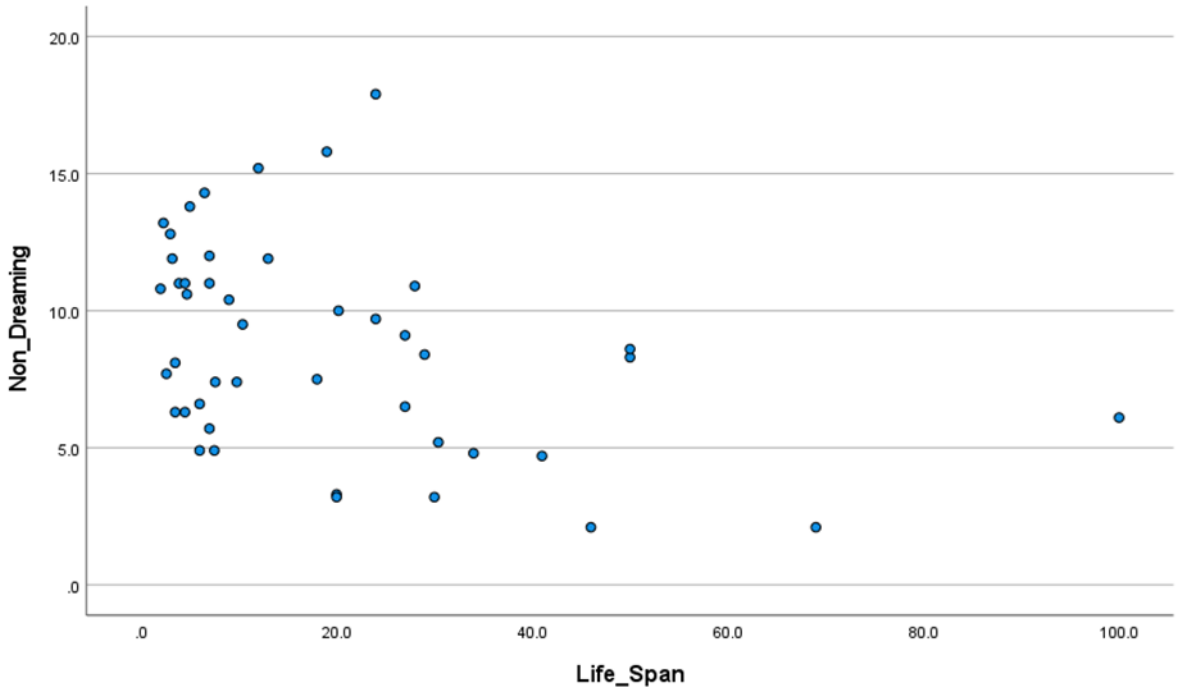
Appendix 20: Scatter Dot with Mean of Sleep Time and Life Span



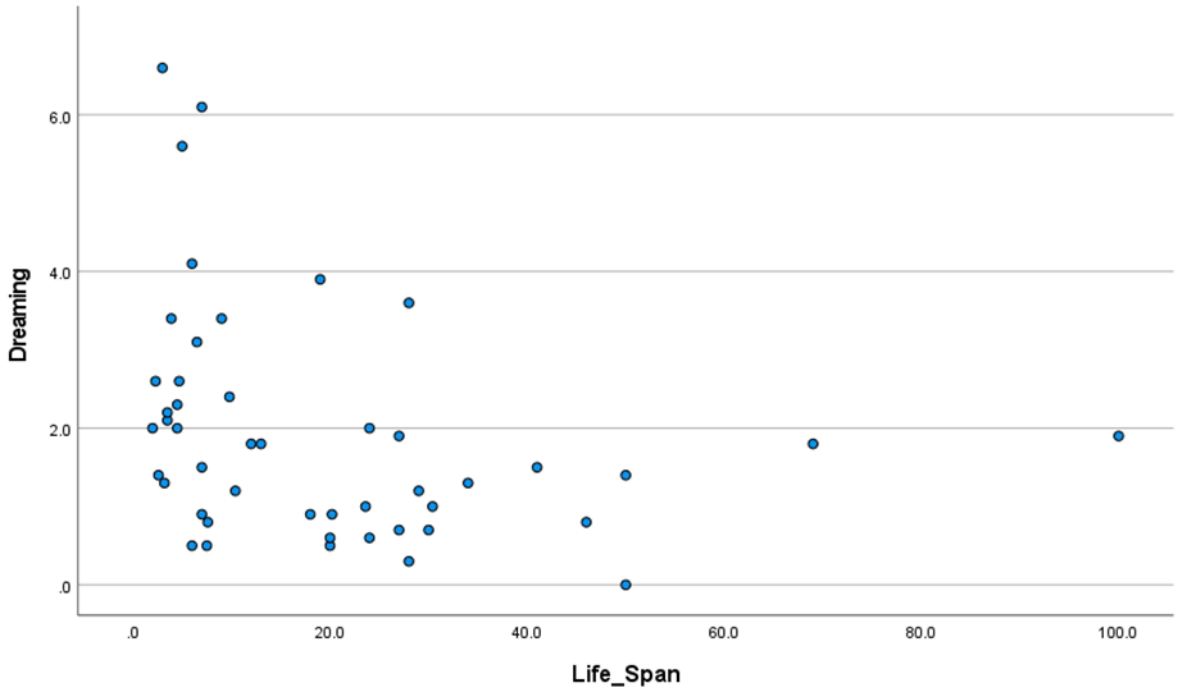
Appendix 21: Scatter Dot with Quadratic Line of Gestation and Life Span



Appendix 22: Scatter Dot with Mean of Non-Dreaming and Life Span



Appendix 23: Scatter Dot of Dreaming and Life Span



END OF APPENDIX

END OF ASSIGNMENT