Project Report

Course: Python Machine Learning Labs

Project: Predicting sleep variables in mammals

Instructor: Christophe Bécavin

Group information:

- Amit Agarwal, DS Accelerated, amit.agarwal@edu.dsti.institute
- Archana Khandelwal, DA Accelerated, archana.khandelwal@edu.dsti.institute
- Claire Davat, DS Accelerated, claire.davat@edu.dsti.institute
- Ken Chevallier, DA Accelerated, ken.chevallier@edu.dsti.institute

Table of Contents

1. Introduction

We are a group of 4 DSTI students from the A23 cohort, all working in accelerated mode in Data Analytics or Data Science curriculum.

After the course and the different projects presentation, we chose to work on this project as it presented the opportunity to work more on data analysis (several correlations to discover, data visualization to do) and machine learning (expecting several models to try with very different results).

While there were several discussions and collaboration during the initial phases of the project, we first worked on separate notebooks to each have a good idea of the data and how to deal with it. After confronting ideas and agreeing on the decisions to take, we regrouped our solutions to make a single report.

The project has been deployed and is available on github.

2. Environment

We used Jupyter notebooks (a part of Anaconda development environment) as the main tool as seen during the class, with also some use of Excel for filtering and grouping analysis.

3. Deliverables

The project takes the input data from sleep_merged.tsv provided to us. There also exits other reference material: AllisonScience1976.pdf and savage-west-2007-a-quantitative-theoretical-framework-for-understanding-mammalian-sleep.pdf.

The file, Project_Description_ML_Sleep_2023.pdf, describes the project.

The deliverables contain the above files and also the Jupyter notebook containing the project code along with the visualizations and interpretations of the results as well as the inferences and conclusions. Henceforth, graphs related to a particular section can directly be referred to in the notebook.

4. Dataset overview and early fixes

4.1 Libraries used

We made use of the following libraries: Numpy, Pandas, Matplotlib, Seaborn, Statsmodels, Scipy, scikit-learn, yellowbrick in our notebook.

4.2 Initial observations

At dataset import, we immediately noticed the low number of rows (87) which underlines the difficulty of the project in terms of ML model performance.

Moreover, there are a lot of missing values, which will need specific work, particularly for the sleep variables, as they are the ones we want to predict.

For some attributes (BodyWt, BrainWt), the distribution is clearly unbalanced and will need adaptation.

5. Initial analysis and early fixes

We identified that there were both missing values as well as erroneous values and we dealt with them separately.

To better analyze the dataset, we added an additional column for Order+Vore (OrderVore).
We also converted the data types of columns to suitable formats (float64, categorical and string).
We also made a check for duplicate values.

6. Erroneous Values

We found erroneous values for the following

- Sleep + Awake exceeding 24h
- Dreaming + NonDreaming != TotalSleep
- Erroneous values for NonDreaming, Conservation, BodyWt and BrainWt

These were corrected accordingly.

7. Filling missing values and zero values

Due to the large number of missing values in the dataset, this activity required a lot of work and discussion on different options.

These columns contained missing values in the dataset:

Feature	# Missing
Conservation	29
NonDreaming	40
Dreaming	24
LifeSpan	33
Gestation	33
Predation	29
Exposure	29
Danger	29

We dealt with and filled in the missing values for the following by either interpolating the values, or by finding suitable values on the web or by considering mean values of related specimens:

1. Dreaming/Non-Dreaming
2. Conservation
3. Ecological attributes: Predation, Exposure, Danger
4. Gestation and LifeSpan

Apart from this, we also dealt with Dreaming/Non-dreaming having a value of zero.

8. Standardizing Data

Since the features were highly unbalanced, we opted to create two sets: one with numerical features in log scale and the other with numerical data normalized using max-min method.
Hence, we had cleaned data columns, data columns in log scale and data columns in normalized scale.

9. Adding additional columns

We added additional columns

- BrainBodyWtRatio = BrainWt/BodyWt
- DreamRatio = Dreaming/24

We have two types of sleep in the dataset: deep sleep (non-dreaming) and REM sleep (dreaming)

1. Deep Sleep or Slow-wave sleep (SWS = NonDreaming/TotalSleep)
2. Rapid eye movement sleep (REM sleep or REMS = Dreaming/TotalSleep)

We added these additional columns for all three kinds of data (standard, log and normalized)

8. Saving cleaned data with additional column

We reordered the columns and saved this cleaned data in a csv file for future reuse.

9. Exploratory Data Analysis (EDA)

After filling all missing and erroneous values, we conducted an initial analysis of the data.
We imported the previously saved cleaned data (in csv format) and created subset of the data into 3: data_std, data_log, data_norm to analyze them separately.

9.1 Data correlation analysis

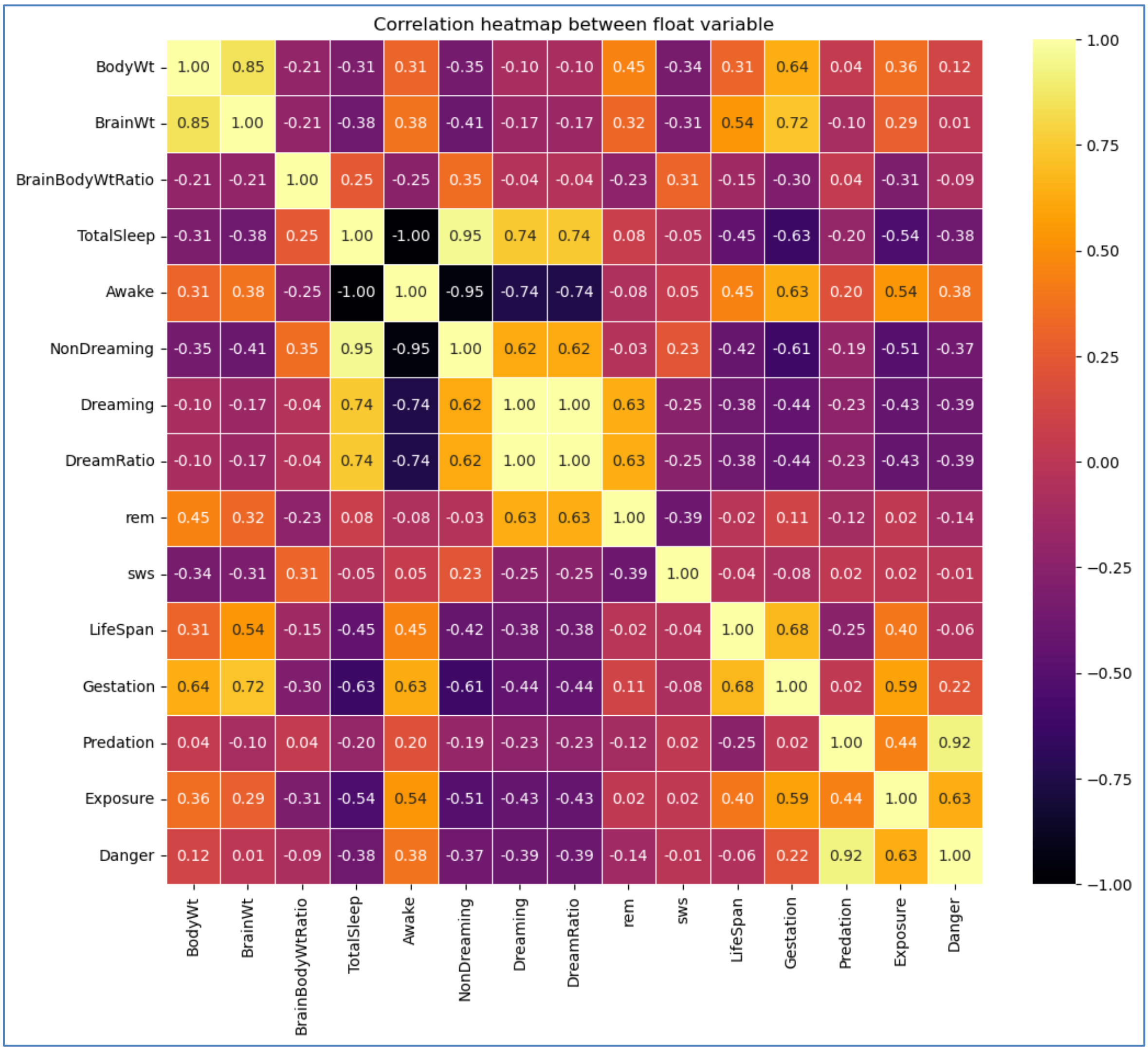
We tried to analyze relationship between several variables including,

- Relation between Danger and TotalSleep
- Relation between Predation and TotalSleep
- Relation between Exposure and TotalSleep

We also observed the distribution of TotalSleep and Dreaming.

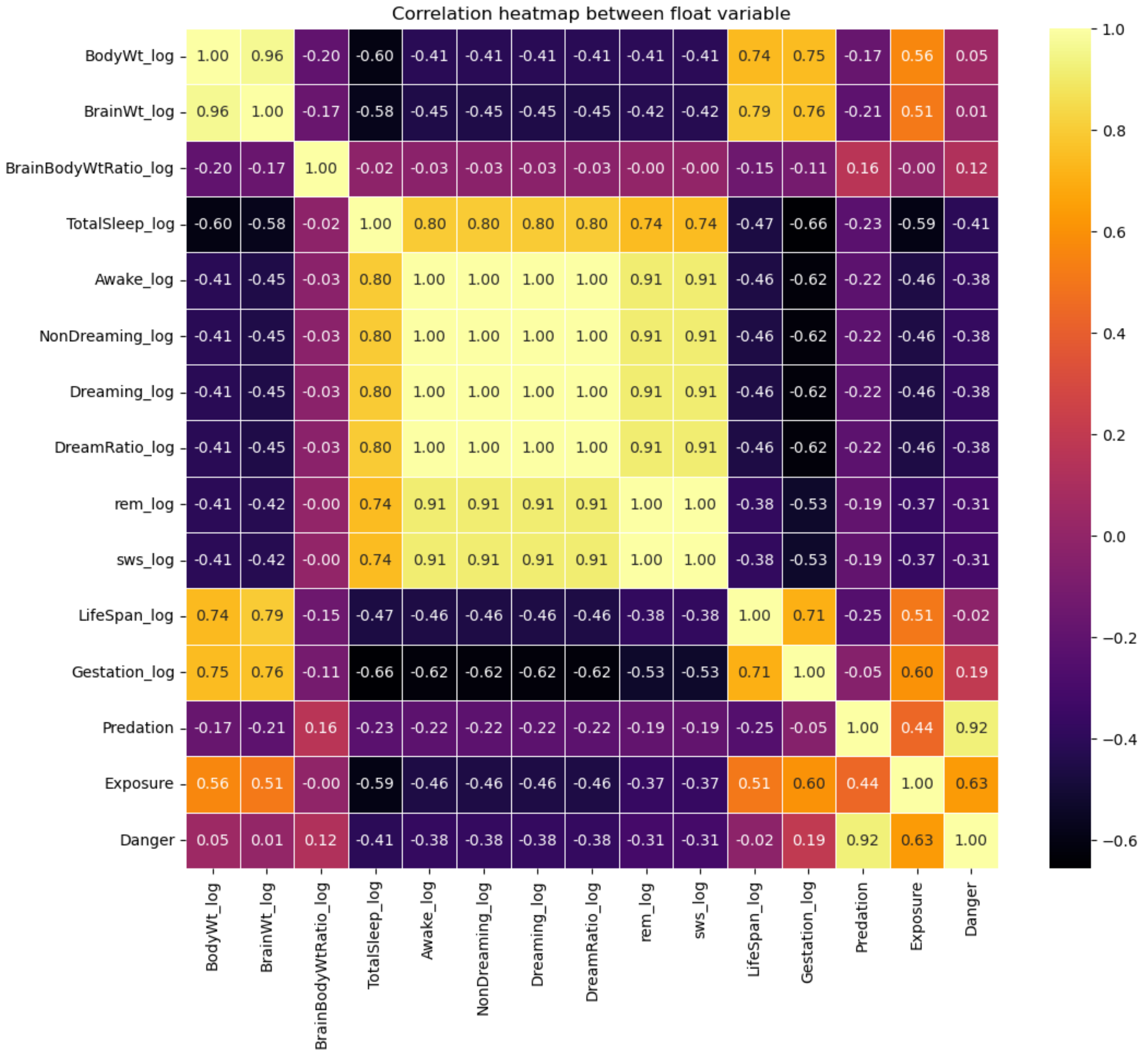
We also analyzed the correlation between the features using a heatmap.

Heatmap using standard cleaned data



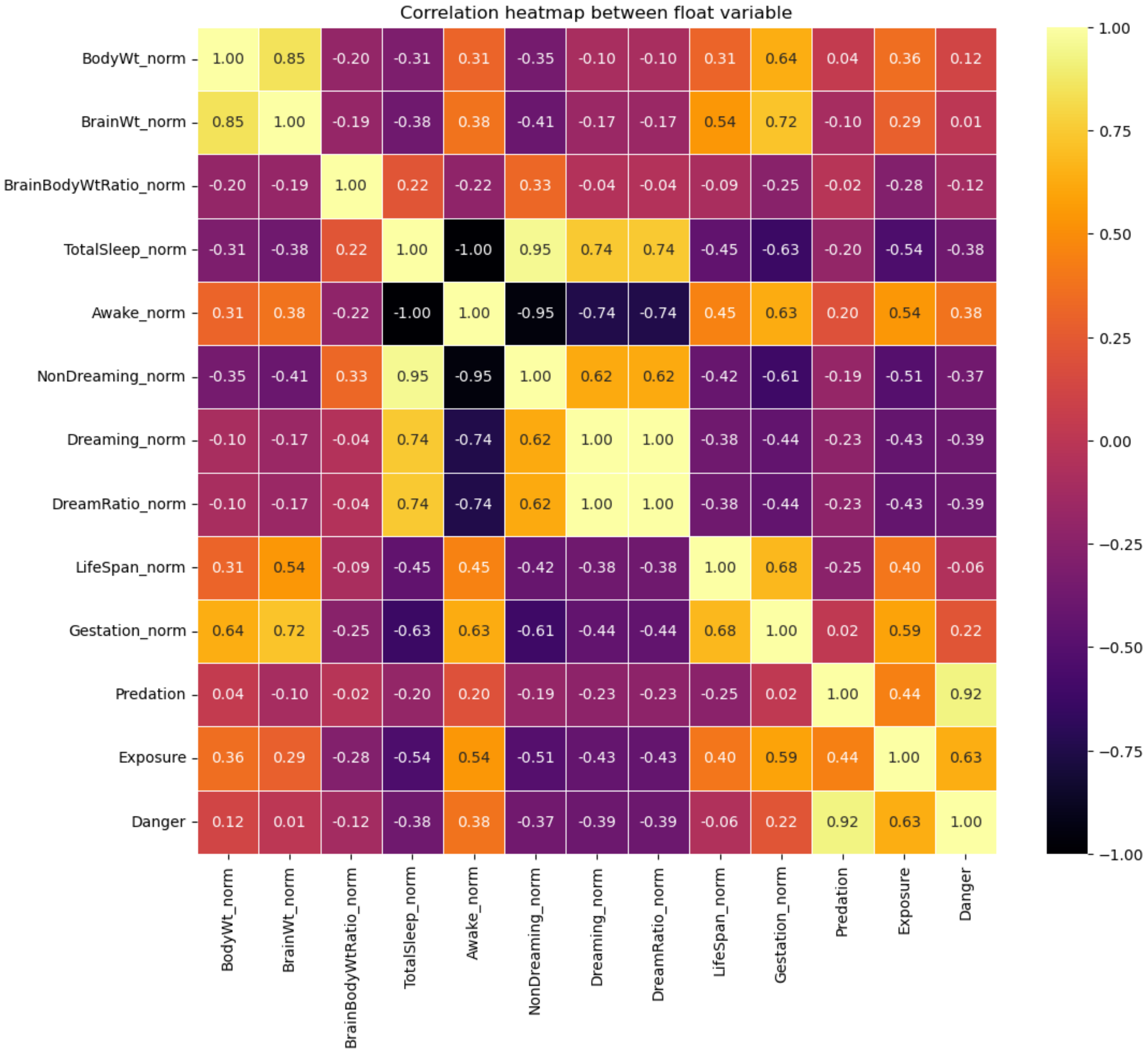
	BodyWt	BrainWt	BrainBodyWtRatio	TotalSleep	Awake	NonDreaming	Dreaming	DreamRatio	rem	sws	LifeSpan	Gestation	Predation	Exposure	Danger
BodyWt	1.00	0.85	-0.21	-0.31	0.31	-0.35	-0.10	-0.10	0.45	-0.34	0.31	0.64	0.04	0.36	0.12
BrainWt	0.85	1.00	-0.21	-0.38	0.38	-0.41	-0.17	-0.17	0.32	-0.31	0.54	0.72	-0.10	0.29	0.01
BrainBodyWtRatio	-0.21	-0.21	1.00	0.25	-0.25	0.35	-0.04	-0.04	-0.23	0.31	-0.15	-0.30	0.04	-0.31	-0.09
TotalSleep	-0.31	-0.38	0.25	1.00	-1.00	0.95	0.74	0.74	0.08	-0.05	-0.45	-0.63	-0.20	-0.54	-0.38
Awake	0.31	0.38	-0.25	-1.00	1.00	-0.95	-0.74	-0.74	-0.08	0.05	0.45	0.63	0.20	0.54	0.38
NonDreaming	-0.35	-0.41	0.35	0.95	-0.95	1.00	0.62	0.62	-0.03	0.23	-0.42	-0.61	-0.19	-0.51	-0.37
Dreaming	-0.10	-0.17	-0.04	0.74	-0.74	0.62	1.00	1.00	0.63	-0.25	-0.38	-0.44	-0.23	-0.43	-0.39
DreamRatio	-0.10	-0.17	-0.04	0.74	-0.74	0.62	1.00	1.00	0.63	-0.25	-0.38	-0.44	-0.23	-0.43	-0.39
rem	0.45	0.32	-0.23	0.08	-0.08	-0.03	0.63	0.63	1.00	-0.39	-0.02	0.11	-0.12	0.02	-0.14
sws	-0.34	-0.31	0.31	-0.05	0.05	0.23	-0.25	-0.25	-0.39	1.00	-0.04	-0.08	0.02	0.02	-0.01
LifeSpan	0.31	0.54	-0.15	-0.45	0.45	-0.42	-0.38	-0.38	-0.02	-0.04	1.00	0.68	-0.25	0.40	-0.06
Gestation	0.64	0.72	-0.30	-0.63	0.63	-0.61	-0.44	-0.44	0.11	-0.08	0.68	1.00	0.02	0.59	0.22
Predation	0.04	-0.10	0.04	-0.20	0.20	-0.19	-0.23	-0.23	-0.12	0.02	-0.25	0.02	1.00	0.44	0.92
Exposure	0.36	0.29	-0.31	-0.54	0.54	-0.51	-0.43	-0.43	0.02	0.02	0.40	0.59	0.44	1.00	0.63
Danger	0.12	0.01	-0.09	-0.38	0.38	-0.37	-0.39	-0.39	-0.14	-0.01	-0.06	0.22	0.92	0.63	1.00

Heatmap using log data



	BodyWt_log	BrainWt_log	BrainBodyWtRatio_log	TotalSleep_log	Awake_log	NonDreaming_log	Dreaming_log	DreamRatio_log	rem_log	sws_log	LifeSpan_log	Gestation_log	Predation	Exposure	Danger
BodyWt_log	1.00	0.96	-0.20	-0.60	-0.41	-0.41	-0.41	-0.41	-0.41	-0.41	0.74	0.75	-0.17	0.56	0.05
BrainWt_log	0.96	1.00	-0.17	-0.58	-0.45	-0.45	-0.45	-0.45	-0.42	-0.42	0.79	0.76	-0.21	0.51	0.01
BrainBodyWtRatio_log	-0.20	-0.17	1.00	-0.02	-0.03	-0.03	-0.03	-0.03	-0.00	-0.00	-0.15	-0.11	0.16	-0.00	0.12
TotalSleep_log	-0.60	-0.58	-0.02	1.00	0.80	0.80	0.80	0.80	0.74	0.74	-0.47	-0.66	-0.23	-0.59	-0.41
Awake_log	-0.41	-0.45	-0.03	0.80	1.00	1.00	1.00	1.00	0.91	0.91	-0.46	-0.62	-0.22	-0.46	-0.38
NonDreaming_log	-0.41	-0.45	-0.03	0.80	1.00	1.00	1.00	1.00	0.91	0.91	-0.46	-0.62	-0.22	-0.46	-0.38
Dreaming_log	-0.41	-0.45	-0.03	0.80	1.00	1.00	1.00	1.00	0.91	0.91	-0.46	-0.62	-0.22	-0.46	-0.38
DreamRatio_log	-0.41	-0.45	-0.03	0.80	1.00	1.00	1.00	1.00	0.91	0.91	-0.46	-0.62	-0.22	-0.46	-0.38
rem_log	-0.41	-0.42	-0.00	0.74	0.91	0.91	0.91	0.91	1.00	1.00	-0.38	-0.53	-0.19	-0.37	-0.31
sws_log	-0.41	-0.42	-0.00	0.74	0.91	0.91	0.91	0.91	1.00	1.00	-0.38	-0.53	-0.19	-0.37	-0.31
LifeSpan_log	0.74	0.79	-0.15	-0.47	-0.46	-0.46	-0.46	-0.46	-0.38	-0.38	1.00	0.71	-0.25	0.51	-0.02
Gestation_log	0.75	0.76	-0.11	-0.66	-0.62	-0.62	-0.62	-0.62	-0.53	-0.53	0.71	1.00	-0.05	0.60	0.19
Predation	-0.17	-0.21	0.16	-0.23	-0.22	-0.22	-0.22	-0.22	-0.19	-0.19	-0.25	-0.05	1.00	0.44	0.92
Exposure	0.56	0.51	-0.00	-0.59	-0.46	-0.46	-0.46	-0.46	-0.37	-0.37	0.51	0.60	0.44	1.00	0.63
Danger	0.05	0.01	0.12	-0.41	-0.38	-0.38	-0.38	-0.38	-0.31	-0.31	-0.02	0.19	0.92	0.63	1.00

Heatmap using normalized data



9.2 Addressing questions from description

Several questions were already provided in the project description.

Do large mammals have better sleep?

We see that TotalSleep has the following characteristics:

mean	10.61
std	4.47
min	1.9
max	19.9
25%	8.15
50%	10.3
75%	13.75

We decided to introduce two new features to help us analyze the TotalSleep:
SleepBodyWtRatio = TotalSleep / BodyWt
SleepBrainWtRatio = TotalSleep / BrainWt

We used the OLS (from statsmodel) to determine the features that contributed most to TotalSleep by examining the coefficient and the p-value of the features.
We used Pearsons's correlation for Gestation and TotalSleep which gave a value of -0.627.
BodyWt was the 2nd feature that explained TotalSleep in the OLS results.

Conclusions:

- The main variable to explain TotalSleep is Gestation : TotalSleep is inversely correlated to Gestation.
- Larger mammal does not have a better sleep, it's the contradictory.
- BodyWt is the 2nd variable to explain Gestation, is moderately correlated but when transformed with log we can see a better correlation

Do predators have better sleep?

At first glance, we do not see that predators have longer sleep as Pearsons's correlation was -0.196.
Danger had a higher negative Pearsons correlation of -0.545.
In general, specimens with lowest Exposure (Danger = 1) have the highest TotalSleep.

Conclusions:

- The trends vary inside each of the pins in the Prediction categories.
- So, we conclude that it is not necessary for predators to have longer sleep.
- As Exposure is more correlated than Predation for explain TotalSleep, Exposure seems a more important variable.

Who has the longer dreams?

At first glance by checking the top 10, longer dreams seem to be an attribute of smaller animals. Except for the giant armadillo, all the animals from the Top 10 dreamers weigh less than 4 kg.
We observe several outliers in Predation category 1&2 but in general, specimens in Predation = 2 have longer dreams.
Exposure gave a good tendency towards Dreaming. We see that Exposure = 1 has highest dreaming.
In general, we saw that insecti has higher dreaming while herbi has the least.
For the categorical attributes, Conservation, Genus, and Order, no clear correlation with TotalSleep was found.

Conclusions:

- Dreaming decreases with exposure (moderately inversely correlated).
- By vore, we saw that insecti has higher dreaming while herbi has the least.
- Dreaming is highly correlated with TotalSleep.
- REM increases with an increase in dreaming.

10. Motivation for model selection

The following were our considerations for selecting the predictive models

- Since our target variable is a numerical variable, we tried to identity regression techniques.
- We also took into consideration the fact that we had very limited amount of data.
- Even though we had several variables that were dependent on each other (eg: BodyWt and BrainWt or Danger and Exposure), we decided not to create a composite variable using PCA.

Despite having selected these models, some were not at all adopted to our scenario (for example: Lasso).

11. Prediction Algorithms

We have used the following models for the prediction of TotalSleep:

1. Linear Regression
   - Easy to implement
   - Provides a linear relation between the features and target
   - Does not capture non-linear relationships between variables
   - Sensible to outliers
   - Does not work very well with variables showing collinearity

2. Multiple Linear Regression (MLR)
   - Captures complex relationships between features and target
   - It factors in multi- collinearity
   - Can be affected by not so useful features.
   - Prone to overfitting

3. Polynomial Regression

Can capture non-linear relationships
Does not work well on small data sets

Prone to overfitting (especially on small data sets)

4. Ridge Regression
    - Works well with small data set
    - Well suited for features displaying multi- collinearity
    - Needs regularization parameters which needs to be adjusted
    - Difficult to interpret the results

5. Kernel Ridge Regression
    - Prone to overfitting (especially on small data sets)
    - Uses l2-norm regularization
    - Needs more time to compute
    - Difficult to interpret the results

6. Lasso Regression
    - Integrate capacity to select the main features
    - Less sensitive to outliers
    - Unstable with highly correlated features
    - Difficult to interpret the results

7. Random Forest Model
    - Works well with both categorical and continuous values
    - Works well with small data set
    - robust to outliers
    - Higher training time due to its complexity
    - Easy to interpret results

8. Support Vector Regression
    - Works well with continuous values
    - Works well with small data set
    - supports both linear and non-linear regressions
    - robust to outliers
    - Uses l2-norm regularization

12. Models training and evaluation

12.1 Training set selection

As the dataset is small, we decided to use a training set selection corresponding to 70% of the original dataset, so that we maximize the model's training ability. Also the computation time for training the model was not a concern for us.

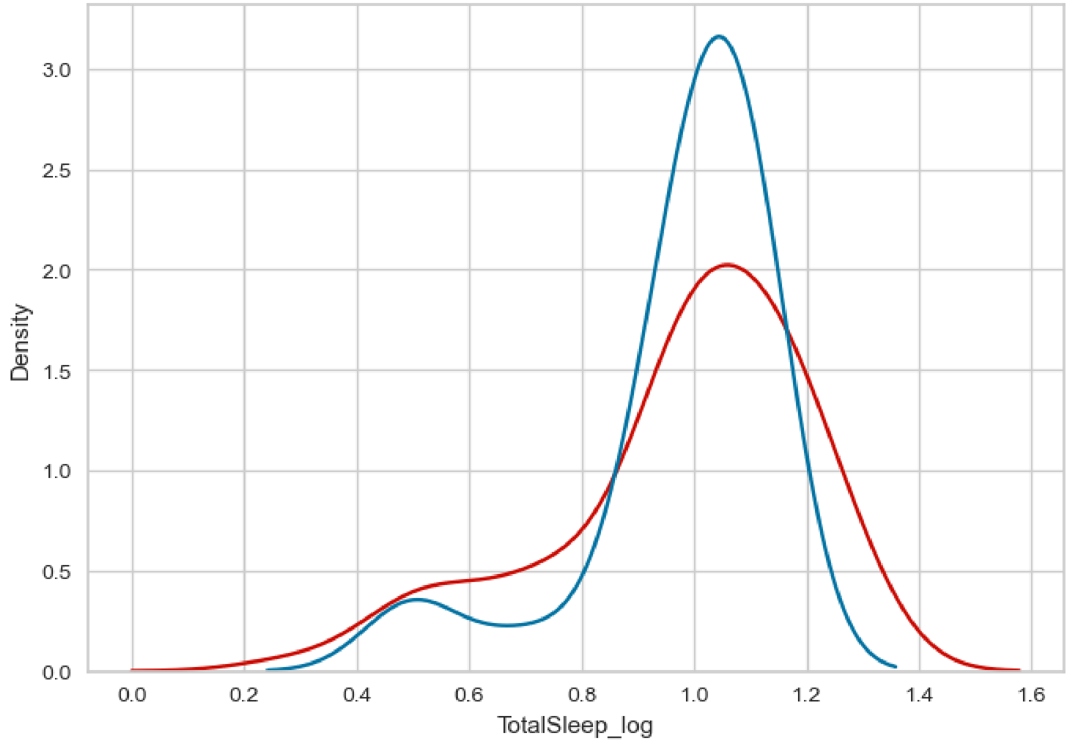
We then considered many different models for the prediction, among which the most meaningful ones are presented next.

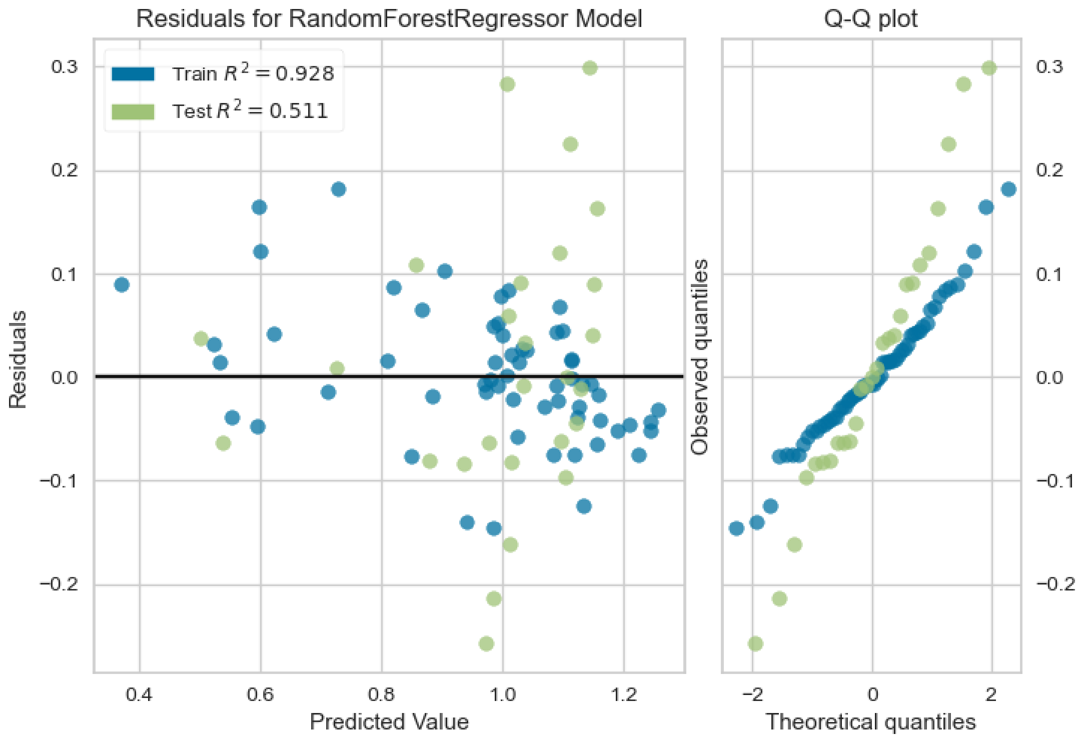
Again, to ease the reusability of the models, we defined each model with its corresponding function, and with its scoring methods.

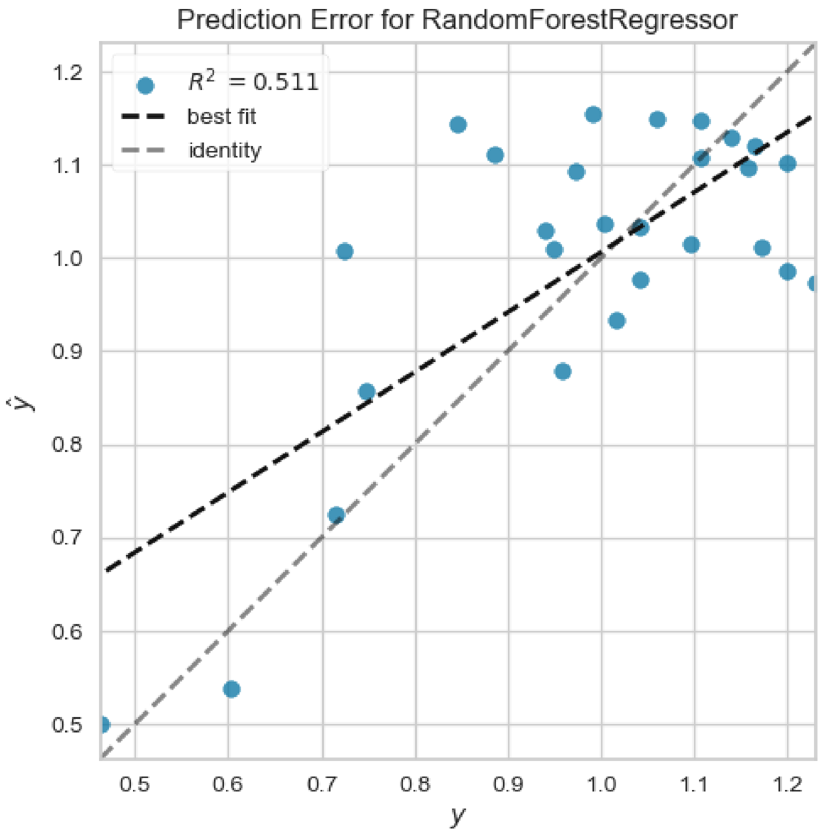
We trained and tested the various models on each of the three types of datasets: standard cleaned, cleaned in log-scale, cleaned, and normalized.

13. Results

We find that Random Forest using the log-scaled data performed better than other models with an accuracy score of 92.21% and mean_squared_error of 0.016 and mean_absolute_error of 0.100.







14. Further study

- We overly cleaned the data and used logic like replacing missing data with the mean values of OrderVore etc or other logic to fill in missing values. This may not be very reliable as the size of each category might or might not have been sufficient.
- Also, we can try to use PCA to create a composite feature out of features showing strong collinearity.

15. Conclusion

This project enabled us to work on a very complete example of data analysis and machine learning with Python.
From data cleaning and visualization to model training and feature refinement, we learned a lot of useful techniques to try to predict a mammal's sleep quantity or quality.
The fact that this topic is "unsolved" with a lot of data still to be discovered made it even more interesting.
Overall, our results indicate that the sleep variables should be predicted with a non-linear regression model.

16. References

- [R1] Allison T, Cicchetti DV. Sleep in mammals: ecological and constitutional correlates. **Science. 1976** Nov 12;194(4266):732-4. doi: 10.1126/science.982039. PMID: 982039.
- [R2] Savage VM, West GB. A quantitative, theoretical framework for understanding mammalian sleep. **Proc Natl Acad Sci U S A. 2007** Jan 16;104(3):1051-6. doi: 10.1073/pnas.0610080104. Epub 2007 Jan 10. PMID: 17215372; PMCID: PMC1783362.