Course: Python Machine Learning Labs
Project: Predicting sleep variables in mammals

Team Lead: Mohamed Jouhari, mohamed.jouhari@edu.dsti.institute
Team Members: Hugo Beffeyte, Kawtar Abidine, Sourena Mohit Tabatabaie, Brandt Olson

Summary

For this project, we undertook an analytical journey to see if it was possible to model mammalian sleep patterns. Utilising data engineering, science and visualisation techniques, our aim was to construct a predictive model that could explain the factors influencing mammalian total sleep duration.

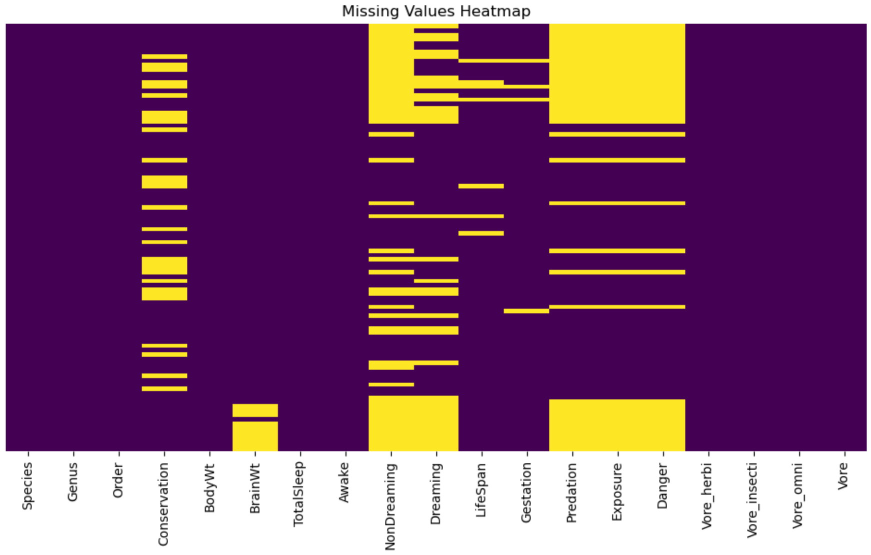
We have also modelled mammals' total dream time noting that REM dreaming data has been largely derived from captive animals where the environment is likely a factor and that observed, i.e. non-EEG collected, REM sleep data in wild animals is difficult to capture and to the exclusion of NREM dreaming[1]. So, why we have been able to produce a model, its predictive utility is debatable at present.

While this project has been an academic exercise in applied machine learning using Python; it has been insightful exercise for the team in understanding an aspect of animal behaviour none of had considered previously, and in enhancing our knowledge of mammalian biology. Food for thought as we dream of elephants and wonder if they dream of us.

1. Data Analysis and Feature Selection

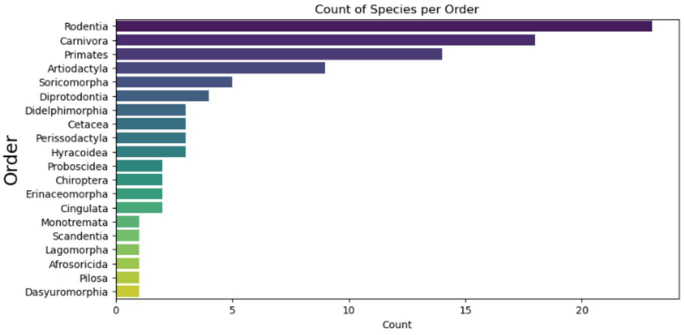
Data Processing and Cleaning

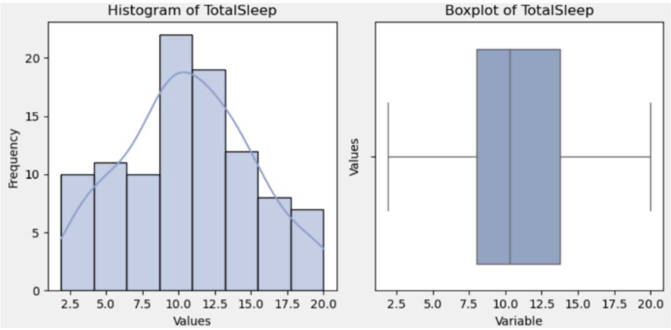
- **Dataset Overview**: The dataset provided for this project encompasses a diverse set of attributes, such as order, life span and weight, which have been considered as potentially having a relationship to mammalian sleep patterns.
- **Data Processing:** Raw data, in the form of TSV, was imported into Python using pandas to make it available for analysis.
- **Data Cleaning**: Columns with excessive missing values were identified as targets for either removal or imputation.



Data Analysis and Visualization

- **Statistical Exploration**: Our exploratory phase employed descriptive statistics to gain a preliminary understanding of the data, such as the number of species present by order.
- **Data Visualization**: We employed a variety of plots— scatter plots for initial relationship gauging, pair plots for multi-dimensional analysis, histograms and boxplots for distribution assessment, such as the distribution of total sleep values and variables.

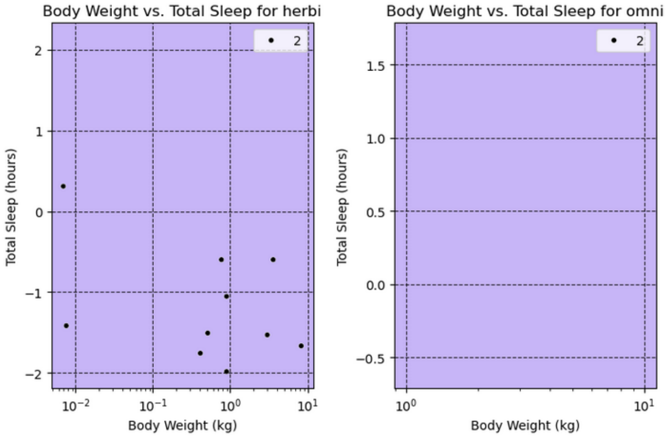




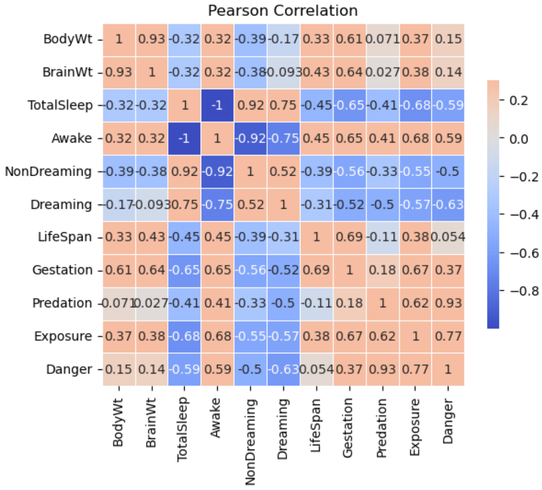
- **Insights Gained**: These visual tools, such as Pearson and Spearman Correlations, were able to identify key trends and patterns, crucial for subsequent modelling stages.

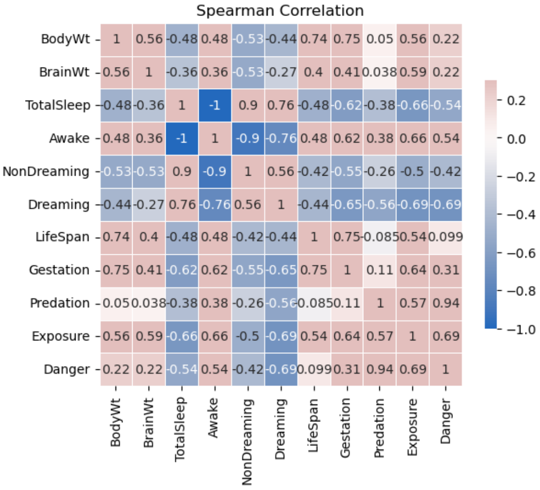
Feature Selection and Engineering

- **Innovative Feature Engineering**: We introduced new features, such as the logarithmic transformation of body weight, to better capture the complex relationships in the data.
- **Prudent Feature Pruning**: To enhance model performance, we eliminated redundant features such as awake, the inverse of total sleep, and less informative features, such as conservation, for which there isn't direct evidence linking conservation status to mammal sleep, focusing instead on variables with substantial influence on sleep patterns.

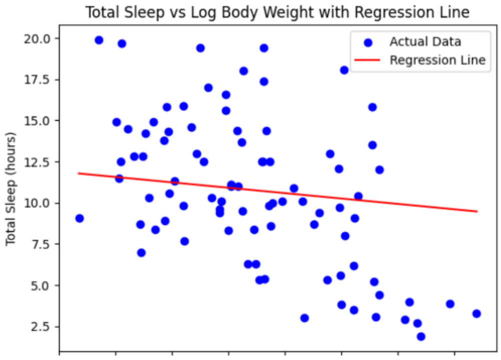


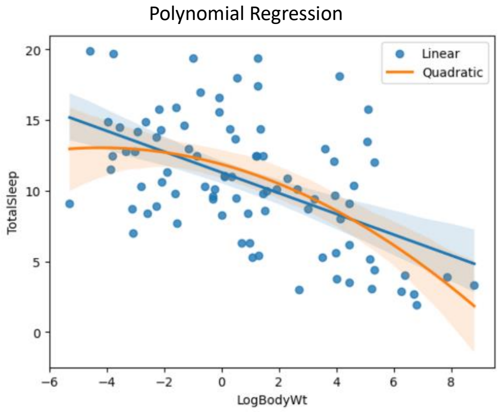
- **Rationale for Choices**: Utilising Pearson and Spearman Correlations along with the Analysis of Variance (ANOVA) statistical method was used in addition to the Pearson and Spearman Correlations for Total Sleep we have shown moderate to strong evidence of significant correlation with TotalSleep, such as 'BodyWt', 'BrainWt', 'NonDreaming', 'LifeSpan', 'Gestation', 'Exposure', and 'Danger'. For Dreaming we have shown moderate to strong evidence of significant correlation with Dreaming, such as 'BodyWt', 'BrainWt', 'NonDreaming', 'LifeSpan', 'Gestation', 'Predation', 'Exposure', 'Danger', and 'Vore_herb





2. Model Training and Evaluation





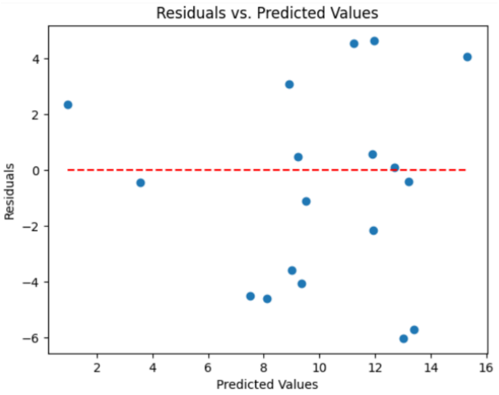
Model Training Approach

- **Simple to More Complex**: Starting with Simpler regression we then moved onto decision trees and then more complex models, such as Random Forests and Gradient Boosting.
- **Linear Regression Model**: With the limited data available the most applicable model and the cornerstone of our analysis was a linear regression model.
- **Exploratory Polynomial Regression**: To ensure thoroughness, we also explored polynomial regression, allowing us to investigate potential non-linear relationships without overly complicating the model.

Robust Model Evaluation

- **Rigorous Evaluation Metrics**: All model's efficacy was scrutinized using Mean Squared Error (MSE) and R-squared metrics.
- **Interpretation of Results**: The Total Sleep model exhibited a moderate R-squared of .4930 value, signifying a respectable fit. The close alignment of linear and polynomial regression lines in our residual analysis further endorsed the linear model's appropriateness. For Total Dreaming an R-squared value of .3144 suggests low predicative value, however, we did note when using Radom Forest that the score increased to .6525, but as noted earlier the data for dreaming is questionable.

- **Evaluation of Results**: A residuals analysis was conducted to further confirm the validity of the regression model.



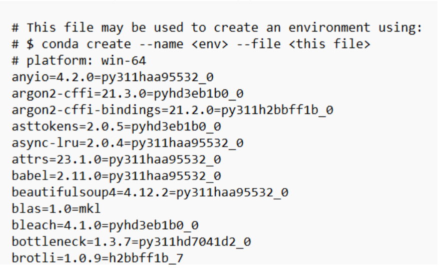
3. Conclusion

This project has provided great insight for the team with regards to the potential of data science to extract academic insights from complex biological data. Our initial 'guess' was that the thermodynamics and associated metabolic rate for large animals vs small animals would be at play, that sleeping as a means of energy conversation could be a factor. While we did not have the data to confirm or not this 'thermodynamic guess', the linear regression model shed surprising light on the link between mammal size and the total sleep time in mammals, offering a window into their ecological adaptations and survival strategies.

4. Ensuring Project Reproducibility

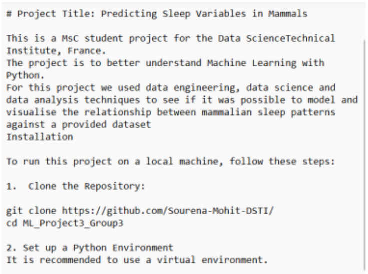
- **Comprehensive Requirements File**: A requirements.txt file to ensures anyone can replicate our analysis can be found on GitHub.
- **Detailed README File**: An accompanying README file, detailing the setup and execution process, can also be found on Git Hub.

Requirements Snippet



```
# This file may be used to create an environment using:
# $ conda create --name <env> --file <this file>
# platform: win-64
anyio=4.2.0=py311haa95532_0
argon2-cffi=21.3.0=pyhd3eb1b0_0
argon2-cffi-bindings=21.2.0=py311h2bbff1b_0
asttokens=2.0.5=pyhd3eb1b0_0
async-lru=2.0.4=py311haa95532_0
attrs=23.1.0=py311haa95532_0
babel=2.11.0=py311haa95532_0
beautifulsoup4=4.12.2=py311haa95532_0
blas=1.0=mkl
bleach=4.1.0=pyhd3eb1b0_0
bottleneck=1.3.7=py311hd7041d2_0
brotli=1.0.9=h2bbff1b_7
```

Readme Snippet



```
# Project Title: Predicting Sleep Variables in Mammals

This is a MsC student project for the Data ScienceTechnical
Institute, France.
The project is to better understand Machine Learning with
Python.
For this project we used data engineering, data science and
data analysis techniques to see if it was possible to model and
visualise the relationship between mammalian sleep patterns
against a provided dataset
Installation

To run this project on a local machine, follow these steps:

1.  Clone the Repository:

git clone https://github.com/Sourena-Mohit-DSTI/
cd ML_Project3_Group3

2. Set up a Python Environment
It is recommended to use a virtual environment.
```

5. Project Hosting and Deployment

- **GitHub Repository**: The project's home on GitHub not only ensures wide accessibility but also fosters collaborative improvement and review.
- **Structured for Accessibility**: The repository's architecture is designed for ease of navigation, ensuring stakeholders can effortlessly access and understand our work.
- **GitHub Repository Link:** https://github.com/Sourena-Mohit-DSTI/ML_Project3_Group3/tree/main

Conclusion

Our journey into the domain of mammalian sleep patterns highlighted for us the intersection of data and biology. In applying data engineering, science and visualisation we were able to understand the type of insights which can be gleaned and how they can contribute to progressing scientific discovery.

[**End of Report**]

Citations

1. Rattenborg NC, de la Iglesia HO, Kempenaers B, Lesku JA, Meerlo P, Scriba MF. 2017 Sleep research goes wild: new methods and approaches to investigate the ecology, evolution and functions of sleep. Phil. Trans. R. Soc. B 372: 20160251. http://dx.doi.org/10.1098/rstb.2016.0251

Sources

https://www.kaggle.com/datasets/volkandl/sleep-in-mammals?select=mammals.csv
https://www.kaggle.com/datasets/mathurinache/sleep-dataset
https://www.kaggle.com/datasets/aguado/dcase-2023-task-5
https://www.kaggle.com/datasets/pereyrax/msleep-mammals-sleep-dataset
https://www.openml.org/search?type=data&sort=runs&id=205&status=active
https://figshare.com/articles/dataset/Mammals_Dataset/1565651
https://commons.datacite.org/doi.org/10.5061/dryad.rv15dv45r
https://www.openintro.org/data/index.php?data=mammals
http://www.statsci.org/data/general/sleep.html
https://www.pnas.org/doi/abs/10.1073/pnas.0610080104
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5613192/
https://rstudio-pubs-static.s3.amazonaws.com/374930_2e95cac7cb6d42c9b1bae838c457de39.html
https://www.researchgate.net/publication/6790848_A_Phylogenetic_Analysis_of_Sleep_Architecture_in_Mammals_The_Integration_of_Anatomy_Physiology_and_Ecology
https://developers.google.com/machine-learning/clustering/prepare-data