# Data Analytics Report

## 1. Problem Statement

Objective: The dataset contains metadata about various organisms, focusing on their classification across phylum, class, order, family, genus, and species levels. The goal is to clean, enrich, and analyze the dataset to understand patterns and trends in species classification, specifically focusing on certain phyla such as Echinodermata and Mollusca.

Columns:  
- phylum (Categorical): The broadest classification group of organisms in the dataset.  
- class (Categorical): A classification that groups organisms within a phylum.  
- order (Categorical): A group of families sharing certain characteristics.  
- family (Categorical): A lower classification grouping within an order.  
- genus (Categorical): A grouping that includes one or more species.  
- species (Categorical): The specific organism name.

Dataset Cleaning Goals:  
- Remove unnecessary columns.  
- Eliminate duplicates and missing values.  
- Add additional rows for two phyla: Echinodermata and Mollusca.

## 2. Data Cleaning

- Unnecessary Columns: The following columns were deleted as they were not needed: rarespecies\_id, eol\_content\_id, eol\_page\_id, kingdom, sciName, common, file\_name.

- Duplicates: The dataset was cleaned of any duplicate entries.

- Missing Data: Rows with missing data were removed.

- Index Reset: After cleaning, the index was reset to keep a continuous order.

## 3. Exploratory Data Analysis (EDA)

Phylum Distribution:  
The number of occurrences for each phylum was analyzed to gain insights into the distribution of organisms.

Data Enrichment:  
Organisms from the phylum Echinodermata were added, including species such as Asterias rubens and Diadema setosum. Similarly, organisms from the phylum Mollusca were also added, such as Octopus vulgaris and Crassostrea gigas.

Shuffling Data:  
The dataset was shuffled to randomize the order of rows, ensuring no inherent bias based on the order of the rows during analysis.

## Visualizations

The following bar graphs were created to visualize the frequency of organisms across different classifications:  
  
1. Classes of Organisms  
2. Orders of Organisms  
3. Top 50 Families  
4. Top 30 Genera  
5. Top 30 Species

## 4. Machine Learning

Though the provided notebook did not include any machine learning models, here’s a suggestion for next steps:  
- Classification Task: Predict the class of an organism based on its family and genus.  
- Suggested Algorithms: Random Forest or Decision Trees would be appropriate for this categorical prediction.  
- Evaluation: Use classification accuracy or F1-score to evaluate model performance.

Label Encoding: To prepare categorical variables for machine learning, apply label encoding to transform categorical variables into numerical format.

## 5. Jupyter Notebook

Each section of the notebook includes the steps for:  
- Reading and cleaning the dataset.  
- Dropping columns and handling duplicates.  
- Adding new organisms to the dataset.  
- Shuffling and resetting indices for further analysis.

## 6. Presentation of Results

The report should summarize:  
- The cleaned dataset and how it was enriched.  
- The distribution of organisms across different phyla.  
- Any additional visualizations (such as bar plots of organism counts across phyla) should be added for better understanding.

## 7. Collaboration Tools

Make sure to upload the final version of the Jupyter notebook and dataset to GitHub for collaboration and version control purposes. Include a README file explaining the purpose of the project and instructions for running the notebook.