

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | |
|  | | **Project Report: Exploratory Data Analysis of Penguin Datasets** | | | | |  | |
|  |  | | | | | | |  |
|  | | | | GITHUB- LINK <https://github.com/Asadxio> LINKEDIN - LINK <https://www.linkedin.com/in/asad-ali-044240262/> Kaggle <https://www.kaggle.com/asadxio> |  | | | |
|  | | | | **1. Introduction:** This project aims to perform exploratory data analysis on two penguin datasets: "penguins\_lter.csv" and "penguins\_size.csv". The goal is to gain insights into the datasets, understand the relationships between variables, detect outliers, and draw meaningful conclusions.  **2. Data Loading:** Two datasets were loaded:   * "penguins\_lter.csv": Contains data on penguin observations, including various measurements and species information. * "penguins\_size.csv": Contains additional size-related measurements for the penguins.   **3. Data Exploration:**   * The "penguins\_lter.csv" dataset consists of X rows and Y columns. It provides insights into the penguins' observations, including their species, body measurements, and other characteristics. * The "penguins\_size.csv" dataset contains size-related measurements for the penguins, including culmen length, culmen depth, flipper length, and body mass.   **4. Handling Missing Data:**   * Null values were checked in the "penguins\_size\_df" dataset using the **isnull().sum()** command. It was observed that the dataset contained missing values. * To handle missing data, the **dropna()** command was used to remove rows with null values from the "penguins\_size\_df" dataset.   **5. Data Cleaning:**   * In the "penguins\_iter\_df" dataset, column names were assigned using the list of unique column names to improve readability and consistency.   **6. Exploratory Data Analysis:**   * Correlation analysis was performed on the "penguins\_iter\_df" dataset using a heatmap, which provided insights into the relationships between variables. * Histograms, count plots, and scatter plots were created to visualize and analyze various features of the datasets, such as "Column1", "Column2", culmen measurements, and body mass. These visualizations helped identify patterns and trends in the data.   **7. Outlier Detection:**   * Box plots were used to detect outliers in the "penguins\_iter\_df" and "penguins\_size\_df" datasets. Outliers were identified based on the distribution of data points beyond the whiskers of the box plots.   **8. Conclusion:**   * Through exploratory data analysis, valuable insights were gained into the penguin datasets. * Correlation analysis revealed relationships between variables, allowing for a better understanding of their interdependencies. * Histograms, count plots, and scatter plots provided visual representations of the data distribution and patterns. * Outliers were identified using box plots, highlighting potential data anomalies.   **9. References:**   * Pandas Documentation: <https://pandas.pydata.org/docs/> * Seaborn Documentation: <https://seaborn.pydata.org/documentation.html> * Matplotlib Documentation: <https://matplotlib.org/stable/contents.html> |  | | | |
|  | | | | Date12/07/23Mini-Project 1 |  | | | |
|  | | |  | | |  | | |



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |