**Data Analysis & Preprocessing Report**

**By Kritika ,enrol.no - 23116047**

**Objective:**

Analyzed and cleaned a **Smartwatch Health Dataset**, performed **Exploratory Data Analysis (EDA)**, handled **outliers & missing values**, and prepared the data for further analysis.

**🔹 Step 1: Data Exploration**

**Loaded & inspected the dataset**

* Checked the data types and structure using df.info() and df.head().
* Found missing values and typos in categorical columns.

**Identified categorical and numerical features**

* "Activity Level" and "Stress Level" were categorical.
* Other columns were numerical (e.g., "Heart Rate (BPM)", "Step Count").

**🔹 Step 2: Data Cleaning**

**Handled Missing Values**

* Filled missing values in **numerical columns** using the **mean**.
* Filled missing values in **categorical columns** ("Activity Level") using the **mode**.

**Fixed Typos & Inconsistencies**

* Found typos in "Activity Level" ("Actve" → "Active", "Seddentary" → "Sedentary").
* Replaced incorrect values using .replace().
* Did required encoding so that it twas easy to analyze the data.

**Converted Data Types**

* Converted **numeric values stored as strings** into actual numbers.
* Ensured "Activity Level" and "Stress Level" were stored as categorical (object type).

**🔹 Step 3: Outlier Detection & Handling**

**Identified Outliers** using box plots and the IQR method.  
 **Removed extreme outliers** only from "Step Count", keeping the dataset balanced.  
 **Replaced outlier values with the nearest valid boundary** (IQR bounds).

**🔹 Step 4: Exploratory Data Analysis (EDA)**

**Visualized Distributions**

* Used **histograms** to check the distributions of numerical variables.
* Used **box plots** to confirm outlier handling.

**Checked Relationships Between Features**

* **Scatter Plots**: Showed relationships (e.g., "Blood Oxygen Level (%)" vs "Step Count").
* **Correlation Heatmap**: Identified weak or strong correlations between numerical features.

**Analyzed Categorical-Numerical Interactions**

* **Bar Plots, Violin Plots, & Box Plots**: Compared "Activity Level" & "Stress Level" with numerical features.

**Multivariate Analysis**

* **Pair Plots**: Showed scatter plots of all numerical features together.
* **Heatmaps**: Displayed category-wise aggregated statistics.
* **Grouped Comparisons**: Used **bar plots** to compare "Heart Rate" across "Activity Level" and "Stress Level".