**Data Cleaning Report**

**1. Introduction and Objectives**

The data cleaning process was implemented to ensure that the survey data collected in the Excel file is accurate, consistent, and ready for further analysis and integration into our diagnostic application. The primary objectives were to eliminate inconsistencies, address missing values, correct data types, and standardize text and numeric fields. This process is fundamental to achieve reliable insights and robust model training later in our project.

1. **Survey Data**

* **Issues to clean:**
* Incomplete responses
* Duplicate entries
* Invalid or inconsistent answers
* Irrelevant or blank “Other (please specify)” responses

1. **Interview Data (from mechanics)**

* **Issues to clean:**
* Transcription errors
* Background noise or irrelevant conversation in audio recordings
* Incomplete or inconsistent responses
* Mislabeled or missing metadata (mechanic name, location, experience level)

1. Image Data (Dashboard warning lights photos)

* **Issues to clean:**
* Blurry, dark, or unclear images
* Incorrectly labeled or misclassified images
* Duplicates or irrelevant photos (e.g. car interiors, speedometers)
* Low-resolution images that can't be used for model training

1. Audio Data (Engine sounds)

* **Issues to clean:**
* Recordings with background noise (talking, traffic, wind)
* Incomplete or cut-off recordings
* Mislabeling (wrong issue type tagged to a sound)
* Duplicate or overlapping sounds
* Silent or low-volume audio clips
* **Why Clean This Data?**
* **To ensure:**
* Accurate AI model training (for sound recognition and computer vision)
* Reliable analytics
* Correct and meaningful insights from surveys/interviews
* Consistency in app diagnosis output

**2. Tools and Technologies**

* **Python & Pandas:** For reading, manipulating, and processing the data.
* **NumPy:** For numerical computations and operations on arrays.
* **Excel File Handling:** Using Pandas’ read\_excel() method.
* **Additional Libraries:** Built-in Python functionality for string operations and outlier detection.

**3. Methodology**

The cleaning process was organized into the following key steps:

**3.1. Data Loading and Exploration**

* **Loading:** The survey data was loaded from an Excel file using pd.read\_excel().
* **Exploration:** The dataset was examined using head(), info(), and describe() methods to understand its structure, identify columns, and assess the overall quality.
* **Missing Values Overview:** A summary of missing values for each column was generated to plan subsequent handling strategies.

**3.2. Handling Missing Values**

* **Strategy Selection:** Two strategies were considered—removing rows with missing values and imputing missing values. For this dataset, imputation was chosen.
* **Numeric Columns:** Missing numeric values (e.g., in the "Rating" column) were replaced with the column mean.
* **Categorical Columns:** Missing text values were filled with a placeholder (e.g., "Unknown").

**3.3. Removing Duplicate Records**

* The dataset was scanned for duplicate rows using duplicated() and all duplicates were removed using drop\_duplicates(). This step ensured that analyses were not biased by repeat submissions.

**3.4. Data Type Corrections**

* Columns were inspected for correct data types:
  + **"Rating" Column:** Converted from a possible string/object type to numeric, ensuring proper calculations and statistical analyses.
  + **"Timestamp" Column:** Converted to datetime format to facilitate time-based analyses.

**3.5. Trimming and Text Standardization**

* All text fields were processed to remove extra whitespace and standardized (converted to lowercase) to achieve consistency across categorical data.

**3.6. Outlier Detection and Removal**

* Outliers in numeric data (such as in the "Rating" column) were identified using a statistical threshold of three standard deviations from the mean. Rows considered to be outliers were removed to reduce skew and improve the quality of subsequent analyses.

**3.7. Saving the Clean Data**

* The final, cleaned dataset was saved to a new CSV file, ensuring that the results of the cleaning process were persistently stored for future use.

**4. Results and Observations**

* **Data Size:** The cleaned dataset maintained a large majority of records after handling missing values by imputing rather than dropping rows.
* **Improved Consistency:** With duplicates removed, corrected data types, and standardized text fields, the dataset now presents a uniform format that supports accurate analysis.
* **Outliers:** The removal of extreme values in numeric fields further contributed to a more stable and normally distributed dataset for model training.
* **Documentation:** All steps were documented within the processing script, ensuring the reproducibility of the cleaning process.

**5. Conclusion**

The data cleaning process has successfully transformed the raw survey data into a reliable and analysis-ready format. The systematic approach comprising data exploration, missing value imputation, duplicate removal, data type correction, text standardization, and outlier management has enhanced the integrity of the dataset. This refined dataset now serves as a solid foundation for further analytical tasks and for training machine learning models in our car diagnostic application.