**Student Name – Sachin Shrinivas Kadam**

**Assignment Name – Understanding Raw Data**

* Review the provided dataset and describe its structure.
  + Find out the total number of records in the dataset.
  + Find out the columns in the dataset. List the column names and describe them.
  + Find the data types of each column.
* Identify common issues in raw data.
  + Find out the number of missing values in each column.
  + Find out if there are any duplicate records in the dataset.

My Analysis & Response as below

I will break down the following steps to thoroughly understand the dataset structure, number of records, columns, data types, missing values, duplicates, and any potential issues such as outliers or inconsistencies. I will also outline the Data Cleaning Recommendation and finally, conclusion.

I will attach/upload the Python code, with this document.

**Dataset Overview -**

* File Name – train.csv
* Number of Records – 891 rows of passenger
* Number of Columns - 12

**Column name with data type, and description -**

|  |  |  |
| --- | --- | --- |
| **Columns** | **Data Type** | **Data Description** |
| PassengerId | Integer | A unique identifier for each passenger. This is typically used for indexing. |
| Pclass | Integer | The passenger's class. It is a categorical variable where -  1. First Class  2. Second Class  3. Third Class |
| Name | String | The name of the passenger |
| Sex | Categorical (String) | The gender of the passenger (male or female) |
| Age | Float | The age of the passenger. Some passengers may have missing age values. |
| SibSp | Integer | The number of siblings or spouses aboard |
| Parch | Integer | The number of parents or children aboard |
| Ticket | String | The ticket number. This may have some duplicate values |
| Fare | Float | The fare paid by the passenger. There might be outliers here. |
| Cabin | String (with missing values) | The cabin the passenger stayed in. This column has a significant number of missing values. |
| Embarked | Categorical (String) | The port where the passenger boarded the Titanic. It can take the values:  C: Cherbourg, Q: Queenstown, S: Southampton  There might be missing values in this column. |
| Survived | Integer | Whether the passenger survived or not. It is a binary column where -  0 – Did not survive  1 - Survived |

**Data Issues in train.csv dataset -**

|  |  |
| --- | --- |
| **Data Issues** | **Details** |
| Missing Values | Train.csv dataset has several missing values, like below   * **Age**: This is one of the most common columns with missing values. The missing data in the Age column can be handled by imputing values (such as the mean, median, or using machine learning techniques). * **Cabin**: The Cabin column has a significant amount of missing data, with only a few passengers having recorded cabins. This could be either imputed or dropped if not relevant. * **Embarked**: This column generally has very few missing values (usually 2-3 rows in the training dataset). |
| Duplicate Values | Though there are Zero duplicate values, it is still recommended to check below things -   * **Duplicate Ticket Numbers**: Since multiple people can share the same ticket, there can be duplicate entries with the same ticket number, especially for large families or groups. * **PassengerId**: Even though PassengerId is a unique identifier, duplicate entries can occasionally exist due to errors during data collection. |
| Outliers | Train.csv has some Outliers like below   * **Fare**: There are passengers who paid extremely high fares (e.g., several thousand units of currency). These outliers may be legitimate (such as first-class passengers) or data-entry errors. This needs to be examined further. * **Age**: There may be outliers in the Age column, such as negative or extreme values. These should be checked for validity. |
| Inconsistent Data | * **Sex**: The Sex column should have consistent values (either "male" or "female"), but there might be spelling inconsistencies like "Male" or "Female" with different capitalizations. * **Embarked**: There could be some inconsistencies, such as capitalizations (e.g., 'C' vs 'c') or unexpected values if data has been manually entered or preprocessed improperly. * **Ticket**: The Ticket column may have mixed formats, containing both numerical and alphanumeric values, and these may need normalization or categorization. |

**Data Cleaning Recommendations for train.csv dataset**

|  |  |
| --- | --- |
| **Cleaning Method** | **Cleaning Recommendation** |
| Handle Missing Data | **Age**: Use imputation strategies (mean, median, or model-based imputation) to fill in missing values for Age.  **Cabin**: Since this column has too many missing values, it could be dropped or, if critical for analysis, imputed with a placeholder value like "Unknown."  **Embarked**: Missing values in Embarked can be imputed with the most frequent value (mode), usually 'S' (Southampton). |
| Remove Duplicates | Ensure there are no duplicate rows based on PassengerId, and check for duplicate Ticket values, handling them appropriately (e.g., aggregating or analyzing groups). |
| Outlier Detection | Examine the Fare column for extreme values. If these outliers are legitimate, they should be retained; if not, they may need to be handled (capped, transformed, or removed).  Check the Age column for unrealistic values (e.g., negative ages or values over 100 years) and handle them as required. |
| Standardize Data | Ensure categorical columns like Sex, Embarked are standardized, either converting all values to lowercase or correcting any anomalies. |
| Feature Engineering | Create new features like family size (SibSp + Parch), or consider categorizing ages into bins (e.g., child, adult, senior). |
| Address Inconsistent Data | Resolve any inconsistencies, such as capitalization issues, especially in categorical columns (Sex, Embarked). |

**Conclusion –** The train.csv dataset is fairly clean but does have some issues that need to be addressed for analysis and modeling. The main issues involve missing values, particularly for **Age** and **Cabin**, potential duplicates (especially in **Ticket**), and outliers in columns like **Fare** and **Age**. Ensuring proper data cleaning and feature engineering will improve model accuracy and generalization.