**Student Name – Sachin Shrinivas Kadam**

**Assignment Name – Data Cleaning Techniques**

* Apply data cleaning methods such as handling missing values using mean, median and mode imputation.
  + If the column with missing values will be retained, what is your strategy to impute the missing values? Justify your decision.

My Analysis & Response as below

* Understand the train.csv dataset
  + **PassengerId**: Unique identifier for each passenger (no missing values).
  + **Pclass**: Class of the passenger (no missing values).
  + **Name**: Name of the passenger (no missing values).
  + **Sex**: Gender of the passenger (no missing values).
  + **Age**: Age of the passenger (may have missing values).
  + **SibSp**: Number of siblings or spouses aboard the Titanic (no missing values).
  + **Parch**: Number of parents or children aboard the Titanic (no missing values).
  + **Ticket**: Ticket number (no missing values).
  + **Fare**: Fare for the ticket (no missing values).
  + **Cabin**: Cabin number (may have missing values).
  + **Embarked**: Port of embarkation (may have missing values).
* Imputing Missing Values - The strategy for imputing missing values should depend on the **nature** of the missing data (e.g., numerical vs. categorical) and the **importance** of the column for modeling.
  + Age (Numerical Column) - The **Age** column is numerical and may/often has missing values. The strategy for imputing Age should balance accuracy with simplicity. As per my analysis, below methods can be considered.
    - **Impute with the median**: The **median** is often a safe choice for imputing numerical features because it is less sensitive to outliers compared to the mean. The distribution of ages in the train.csv dataset is somewhat skewed, and using the median ensures that the imputed values do not disproportionately impact the model.
      * **Justification**: The median is robust and will prevent the imputed values from being unduly influenced by outliers or skewed distributions.
    - **Impute with the mean**: If you believe the data is reasonably symmetric (or you check the distribution), using the **mean** might also work.
    - **Impute using other variables**: If available, using a **more advanced strategy** like multiple imputation or predictive modeling (e.g., using a regression model to predict age based on features like Pclass, Sex, etc.) could be considered. However, this approach is more complex and computationally expensive, and it may not always provide a significant advantage.
  + Cabin (Categorical Column) - The **Cabin** column may/often has a lot of missing values, which are typically sparse or encoded as "U" for unknown. Rather than imputing missing values for the **Cabin** column, it might be more effective to create a new feature that indicates whether a passenger has a cabin or not.
    - **Impute with "Unknown" or "No Cabin"**: Given that the number of missing cabin values is quite large, we can treat this as a **binary feature** indicating whether the passenger had a cabin (1 = Yes, 0 = No).
      * **Justification**: This allows the model to recognize the absence of data as a distinct pattern rather than filling in random values, which might add noise.
  + Embarked (Categorical Column) - The **Embarked** column indicates the port of embarkation and typically has very few missing values. For a small number of missing values, you can impute using the **mode** (most frequent value) of the column.
    - **Impute with the mode**: The most frequent embarkation port (often "S") can be used to fill the missing values in this case.
      * **Justification**: Since the "Embarked" column likely has only a few missing values, using the mode will preserve the original distribution and prevent introducing bias.
  + Other Columns (SibSp, Parch, Fare, Ticket, Name, Sex, Pclass, PassengerId)
    - These columns generally do not contain missing values and do not need imputation. However, if any missing values do appear, they should be handled based on the type of column:
      * **SibSp**, **Parch**, **Pclass**: If any missing values appear, you can impute with the **mode** (for categorical variables) or **mean/median** (for numerical variables).
      * **Name**, **Ticket**, **Sex**: These columns are important but unlikely to have missing values. If they do, consider imputing with the mode (for categorical) or a placeholder like "Unknown."
* Summary, and Justification of Imputation Strategy
  + Summary - **Age**: Impute with the **median** age of the dataset (robust to outliers).
  + Justification - **Median for Age**: This method is robust to outliers and helps prevent distorting the distribution of the data, making it a safe option for numerical columns with skewed distributions like age.
  + Summary - **Cabin**: Create a new binary column to indicate whether a passenger had a cabin or not. For missing values, impute with "Unknown."
  + Justification - **Cabin**: The missing values in the Cabin column are numerous, and it’s unlikely to be useful to fill in actual cabin numbers. Creating a binary indicator provides valuable information about whether a passenger had a cabin, which might have an impact on survival rates.
  + Summary - **Embarked**: Impute missing values with the **mode** (most frequent value) of the column.
  + Justification - **Embarked**: Since there are only a few missing values, imputation with the most frequent value (mode) makes sense to preserve the existing distribution without adding noise.
  + Summary - **Other Columns**: For any other missing values in the dataset (if they occur), impute using appropriate strategies such as the **mode** (for categorical variables) or **median/mean** (for numerical variables).