<dataantys%20task1.docx>1) What are missing values and how do you handle them?

Missing values refer to absent data. https://www.kaggle.com/competitions**Deleting:** Removing rows or columns if data loss is minimal or the column is mostly empty.

* **Imputing:** Filling them with estimated values like the mean, median (for numerical), mode (for categorical), or a new 'Unknown' category.

2) **How do you treat duplicate records?**

Duplicate records are identical rows. You treat them by **identifying and removing them**, typically keeping only one instance of the unique record. This ensures data integrity and prevents bias in analysis.

 Difference **between dropna () and fillna () in Pandas?**

* dropna() **removes** rows or columns containing missing values.
* fillna() **fills** missing values with a specified value or method, thus retaining the row/column.

3) **What is outlier treatment, and why is it important?**

Outliers are data points significantly different from others. Treatment involves identifying these extremes and deciding whether to remove, transform, or cap them. It's crucial because outliers can distort statistical analyses, negatively impact model performance, and lead to misleading insights.

4) **Explain the process of standardizing data.**

Standardizing data (numerical feature scaling) transforms features to a similar scale, typically with a mean of 0 and a standard deviation of 1 (Z-score) or a range of 0 to 1 (Min-Max scaling). This is done by fitting a scalar on training data and then transforming both training and test data. It's important for algorithms sensitive to feature magnitudes.

5) **How do you handle inconsistent data formats (e.g., date/time)?**

Handle inconsistent formats by converting them to a single, consistent type. For date/time, parse text into proper date objects (e.g., pd.to\_datetime() in Python, or Excel's "Format Cells" and specific formulas for text dates). For text, standardize case and variations (e.g., "Male" vs. "male").

6) **What are common data cleaning challenges?**

Common challenges include handling missing data, identifying and treating duplicates, inconsistent data types/formats, managing outliers, dealing with inconsistent text entries (typos, synonyms), and integrating data from multiple sources.

7) **How can you check data quality?**

Check data quality by:

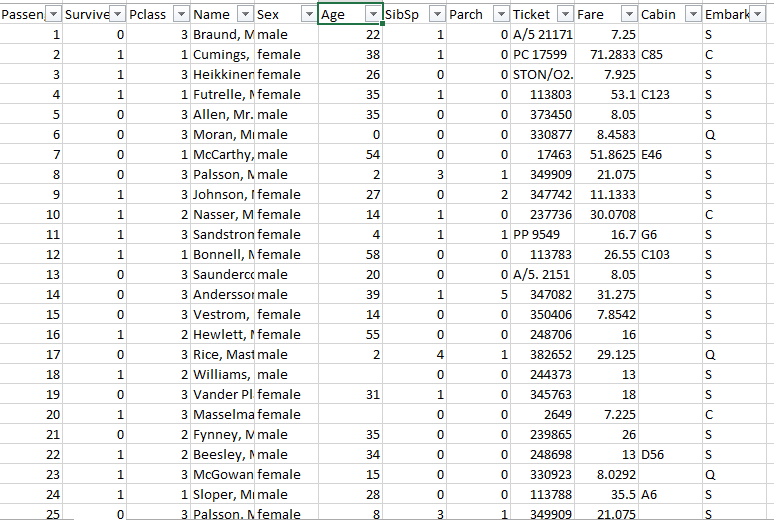
* **Descriptive Statistics:** Reviewing min, max, mean, and value counts.
* **Missing Value Counts:** Identifying columns with missing data.
* **Unique Value Counts:** Spotting inconsistencies in categorical data.
* **Data Type Inspection:** Ensuring columns have correct data types.
* **Visualizations:** Using charts (box plots, histograms) to spot outliers and distributions.
* **Logical Checks:** Verifying data against known business rules.

**Task 1: Data Cleaning and Preprocessing**

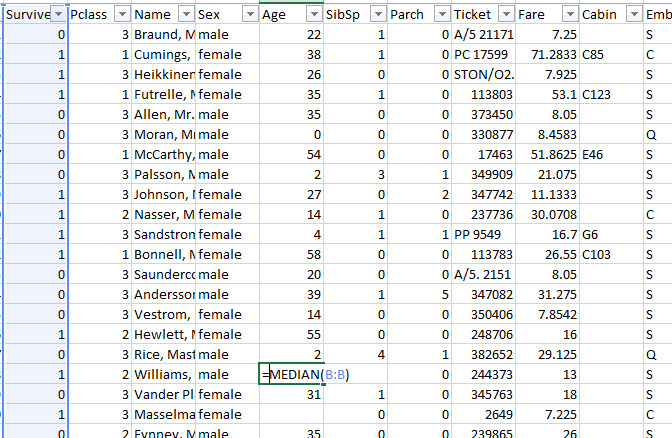
**Objective**: Clean and prepare a raw dataset (with nulls, duplicates, inconsistent formats).

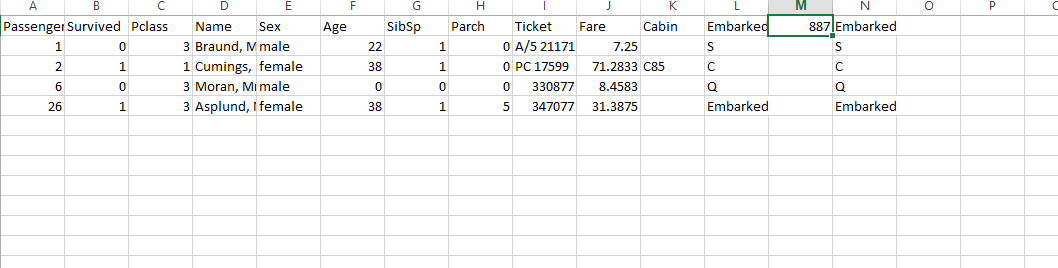
Identify and handle missing values using filters in Excel

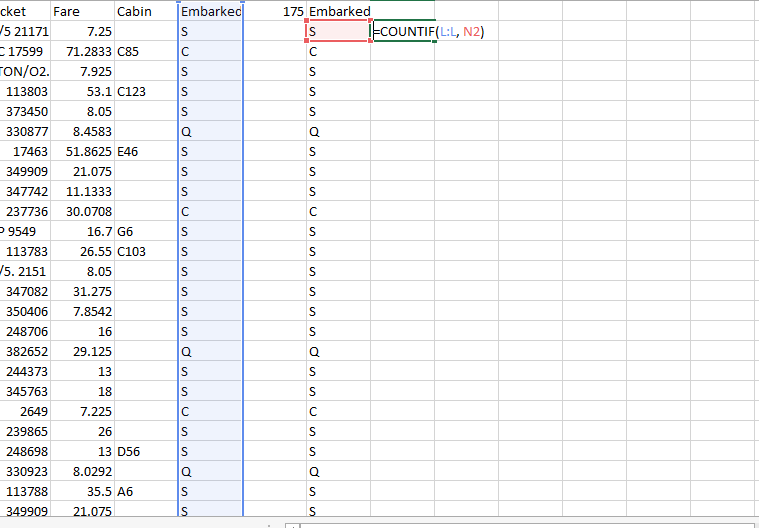
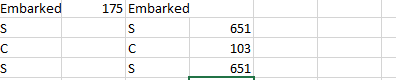
* **Excel Approach:**
  + **Identification:** In Excel, you can:
    1. Select the entire dataset.
    2. Go to the Data tab and click Filter.
    3. Click the filter arrow in the header of a column (e.g., Age).
    4. Uncheck Select All and then check Blanks to show only rows with missing values in that column.
  + **Handling:**
    1. For Age (numerical): You could calculate the median of the 'Age' column (e.g., =MEDIAN(B:B) if Age is in column B) and then manually type that median into the blank cells, or use a "Find and Replace" function (Ctrl+H) to replace blanks with the calculated median.
    2. For Embarked (categorical): You could find the most frequent value by manually inspecting or using COUNTIF functions, and then manually fill the blanks.
    3. For Cabin: You would select the entire 'Cabin' column (e.g., column K) and then right-click and choose Delete -> Entire column.

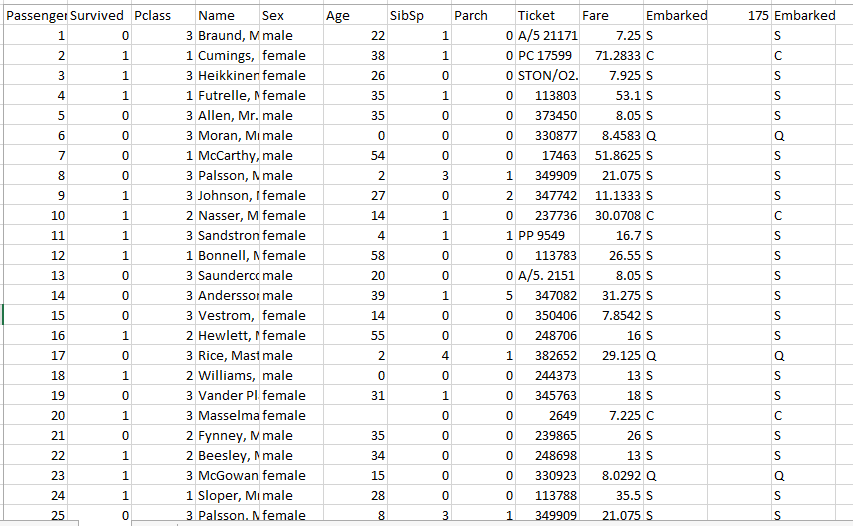


How many total columns are missing in the age columns means totally =176

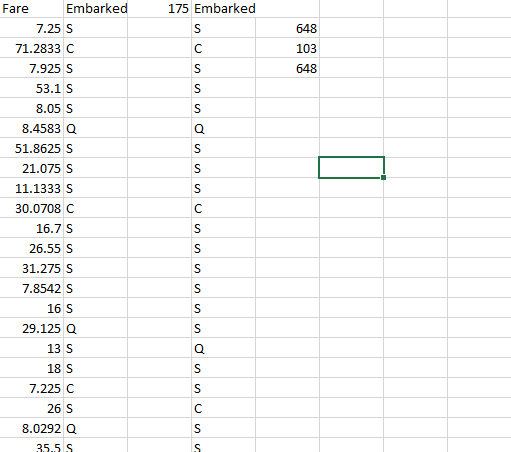




For Cabin: You would select the entire 'Cabin' column (e.g., column K) and then right-click and choose Delete -> Entire column.

2) Remove duplicate rows using .drop\_duplicates() or Excel’s “Remove Duplicates



Totally these three remove duplicate

*  **olumns to Consider for Duplicates:**
  + Below the header option, you will see a list of all your columns. By default, all columns are usually checked.
  + **To remove truly identical rows:** Keep all columns checked. This means Excel will only remove a row if *all* its values across *all* selected columns are identical to another row.
  + **To remove duplicates based on specific criteria:** If you only want to remove rows where, for example, the 'Passenger ID' and 'Name' are the same (even if other columns like 'Age' might differ), you would uncheck all columns except 'Passenger ID' and 'Name'.
* For cleaning a dataset like train.csv to ensure unique entries, it's generally best to keep **all columns checked** to remove truly identical records.

 **Click 'OK':**

* Once you've configured the columns, click the OK button in the dialog box.

 **Review the Confirmation Message:**

* Excel will process the data and then display a confirmation message. This message will tell you:
  + "X duplicate values found and removed."
  + "Y unique values remain."
* Click OK on this message.

3) Standardize text values like gender, country names, etc.

### ✅ Step 1: Open the File in Excel

1. Open the train.csv file in Excel.
2. Make sure columns like Sex, Embarked, and Name are visible.

### ✅ Step 2: ****Standardize Gender (****Sex****) Values****

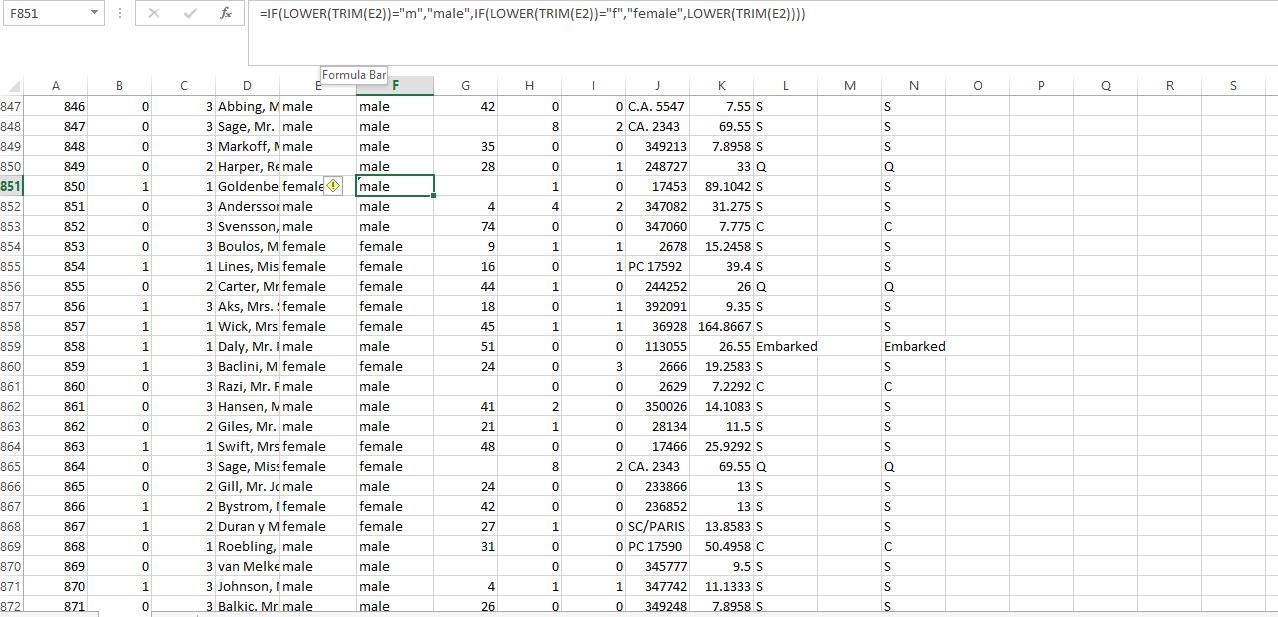
#### 🔹 Step 2.1: Create a new column called Gender\_Cleaned

In the cell next to the first Sex cell (e.g., if the first Sex is in E2, then in F2), enter this formula:

excel

CopyEdit

=LOWER(TRIM(E2))



### ✅ Step 1: Open the File in Excel

1. Open the train.csv file in Excel.
2. Make sure columns like Sex, Embarked, and Name are visible.

### ✅ Step 2: ****Standardize Gender (****Sex****) Values****

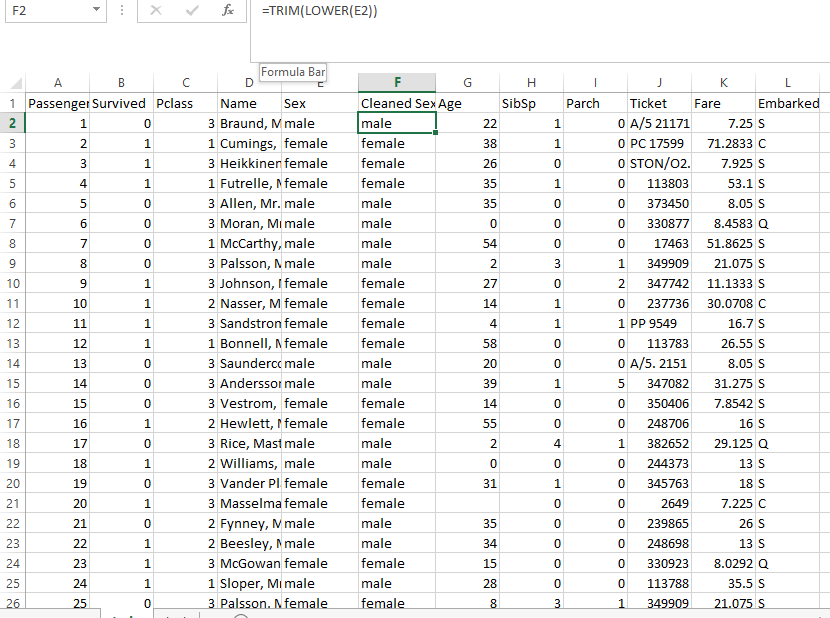
#### 🔹 Step 2.1: Create a new column called Gender Cleaned

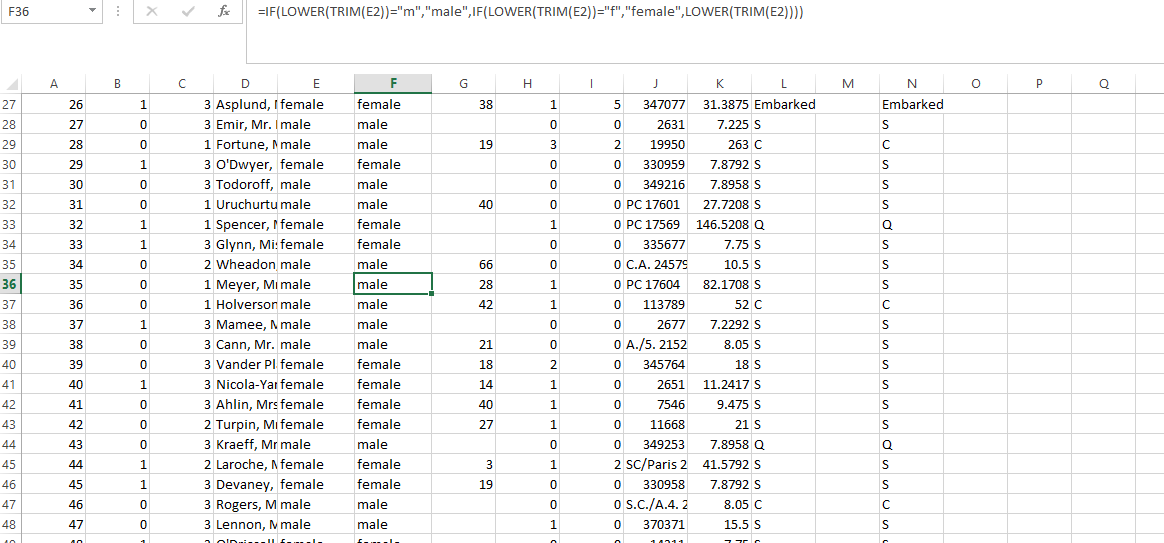
In the cell next to the first Sex cell (e.g., if the first Sex is in E2, then in F2), enter this formula:

excel

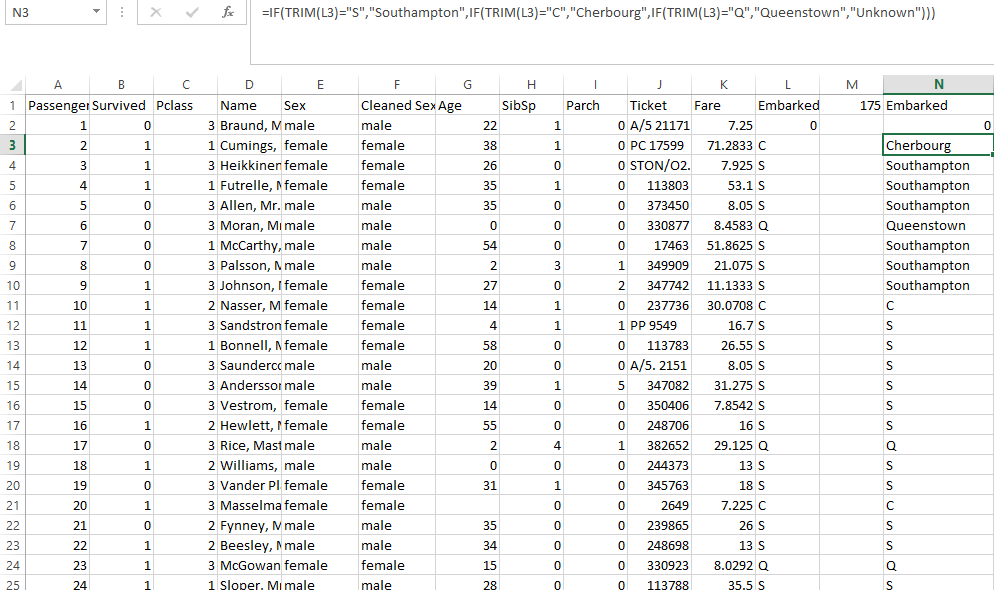
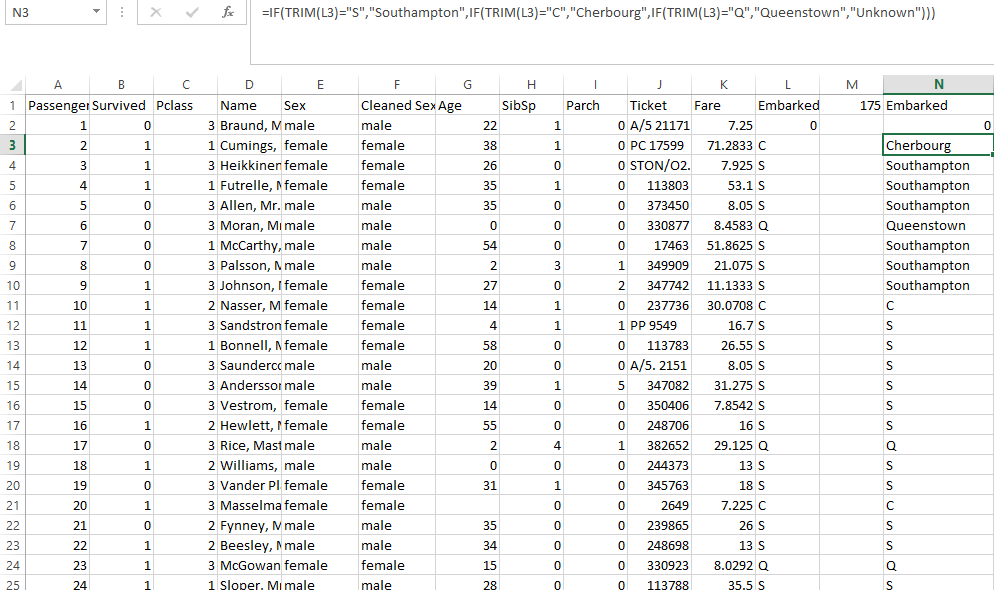
CopyEdit

=LOWER(TRIM(E2))

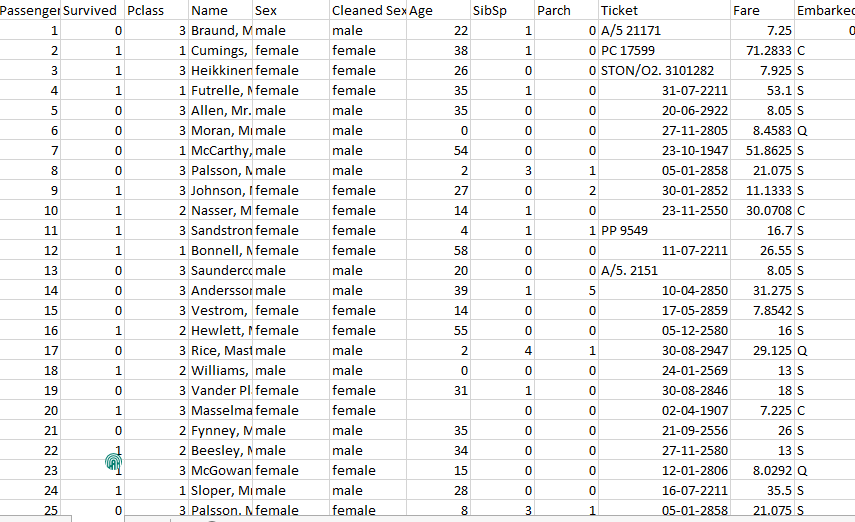


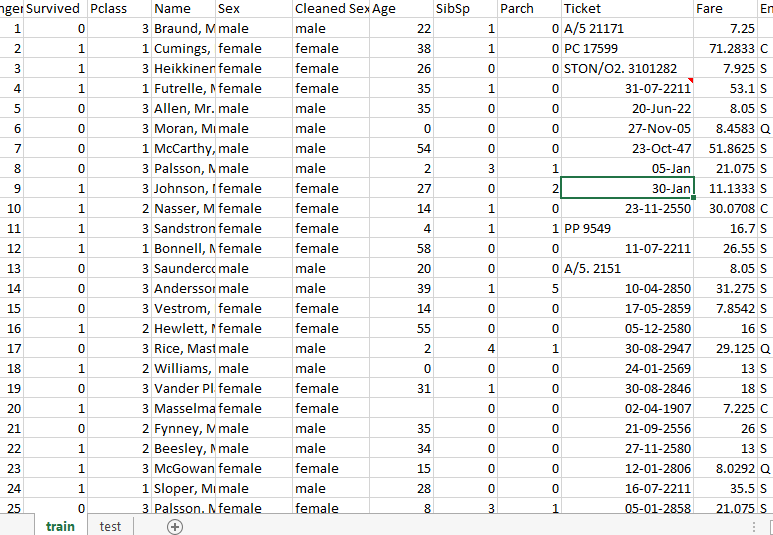


### ✅ Step 3: ****Standardize Embarked Values****



4) Convert date formats to a consistent type (e.g., dd-mm-yyyy). excel format this dataset:-

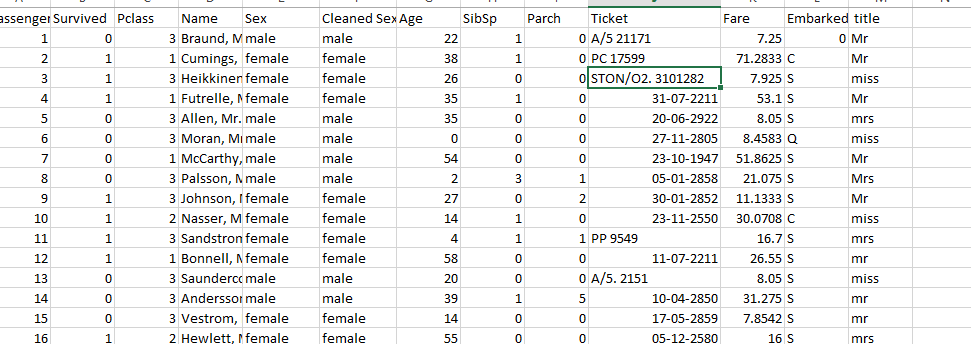




Select the data column and then to select the entire column that contains your dates. This ensures all entries in that column will be affected by the formatting change. Open 'Format Cells' Dialog Box  **ight-click** anywhere within the selected column (**Column J**).

 From the context menu that appears, select Format Cells...  **Navigate to the 'Number' Tab and Choose 'Date':**

* In the 'Format Cells' dialog box that opens, make sure the Number tab is selected at the top.
* In the Category list on the left side, select Date.

 

5) Rename column headers to be clean and uniform (e.g., lowercase, no spaces) in excel method:-

**Locate Your Column Headers:**

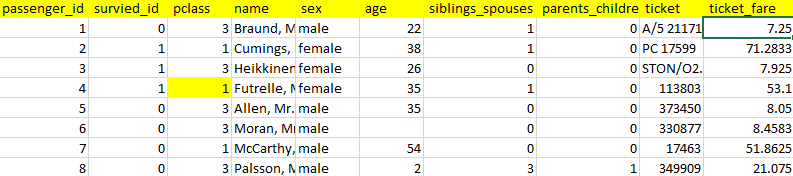
* **Action:** Your column headers are typically in the first row of your dataset (e.g., Row 1).

 Select **the First Header to Rename:**

* **Action:** Click directly on the cell containing the first column header you want to change (e.g., cell A1 which contains PassengerId). This will make the cell active and allow you to edit its content.

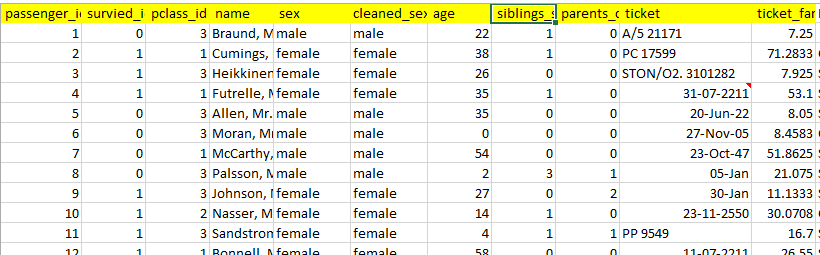
 Edit **the Cell Content:**

* **Method A (Direct Edit):**
  + **Action:** Double-click on the cell (A1). The cursor will appear inside the cell, allowing you to type and delete.
  + **Action:** Manually type the new, clean header name (e.g., passenger\_id).
  + **Action:** Press Enter on your keyboard to finalize the change.



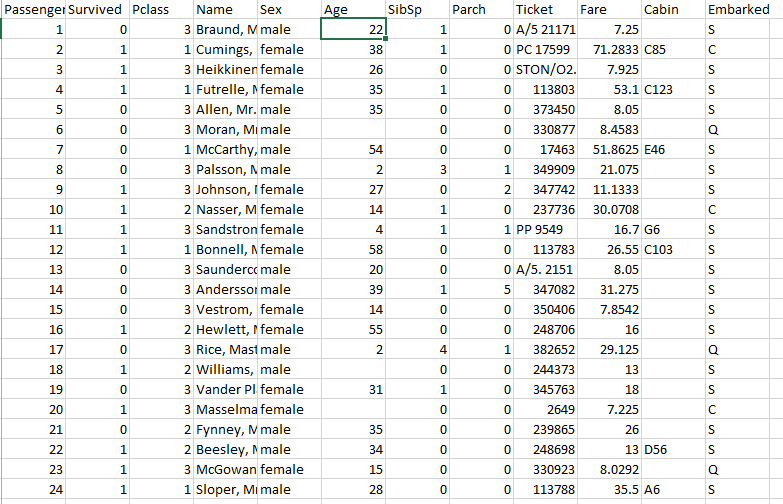
**Method B (Using Formula Bar):**

* **Action:** Click once on the cell (A1) to select it.
* **Action:** Look at the "Formula Bar" at the top of Excel (the long bar where formulas appear, usually starting with fx). The content of cell A1 (PassengerId)ill be displayed there.
* **Action:** Click inside the Formula Bar and manually edit the text to your new clean header name (e.g.,  passenger\_id).
*  **Action:** Press Enter on your keyboard or click the green checkmark next to the Formula Bar to finalize the change



**6)** **Check and fix data types (e.g., age should be int, date as date time).**

**Check Current Data Types/Format (Visual Inspection & 'Format Cells')**



**Visual Inspection:**

Look at the alignment of the data in your columns. By default, numbers and dates in Excel usually align to the **right** of the cell. Text usually aligns to the **left**. If your numbers or dates are aligned to the left, it often means Excel is treating them as text. Look for green triangles in the top-left corner of cells. This often indicates a "Number Stored as Text" error.

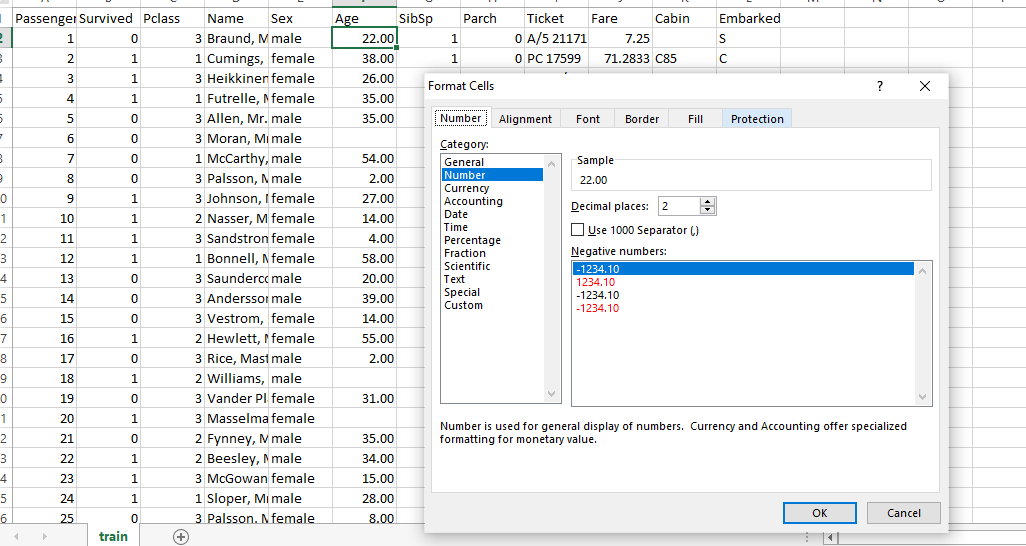
**Using 'Format Cells' for Confirmation:-**

 Select the entire column you want to check (e.g., **Column F** for 'Age' in train.csv).

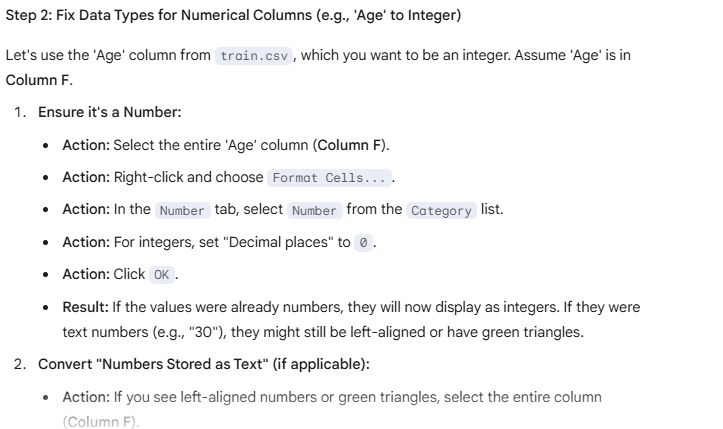
 Action**:** Right-click anywhere in the selected column and choose Format Cells....

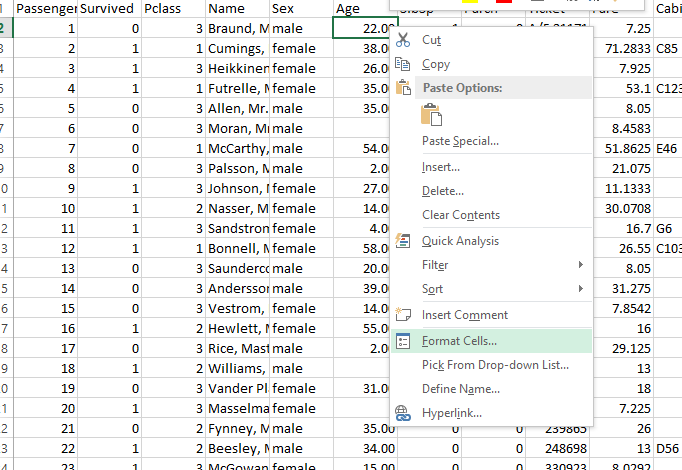
 **Action:** In the Format Cells dialog box, go to the Number tab.

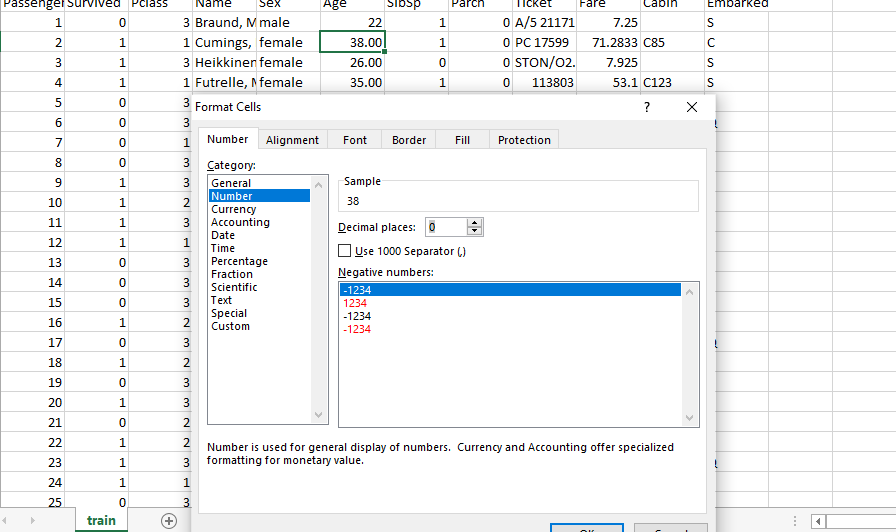
 Result**:** Look at what Category is currently selected (e.g., General, Number, Text, Date). This tells you Excel's current interpretation.

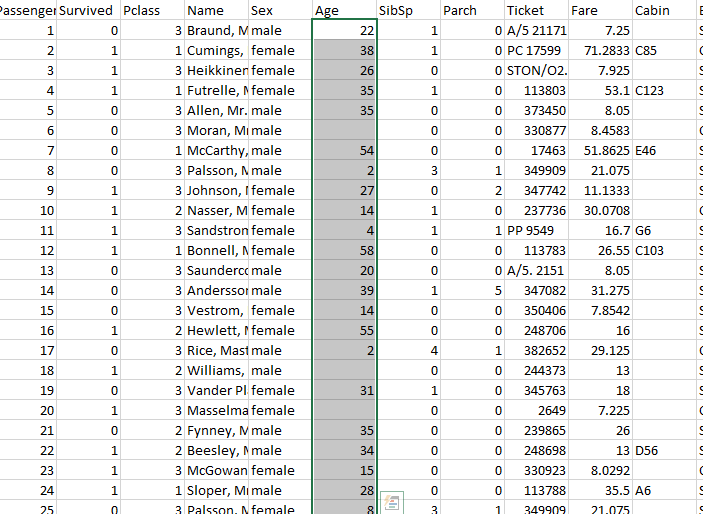
****

Fix Data Types for Numerical Columns (e.g., 'Age' to Integer):-

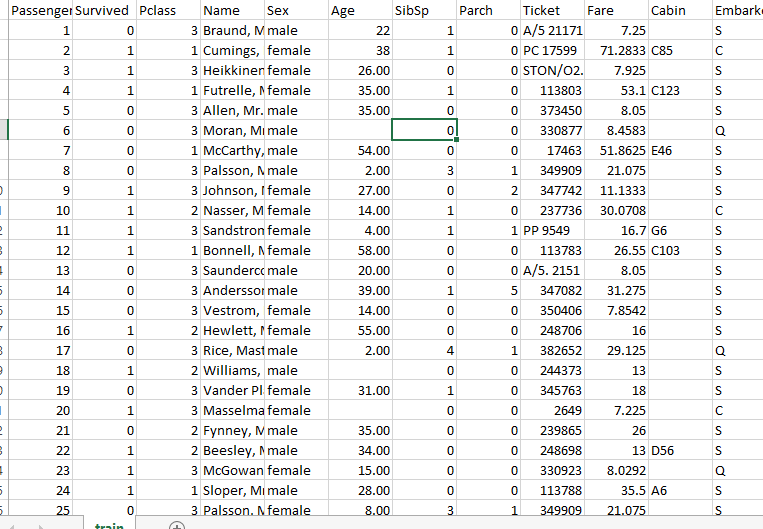
****

****





**Convert "Numbers Stored as Text" (if applicable):**



 Action**:** If you see left-aligned numbers or green triangles after Step 2.1, select the entire column (**Column F**).

 **Action:** Look for a small warning icon (a diamond with an exclamation mark) that appears when a cell with a green triangle is selected, or look for the Data tab on the ribbon.

 Method **A (Error Check Option - Quick Fix for many cells):**

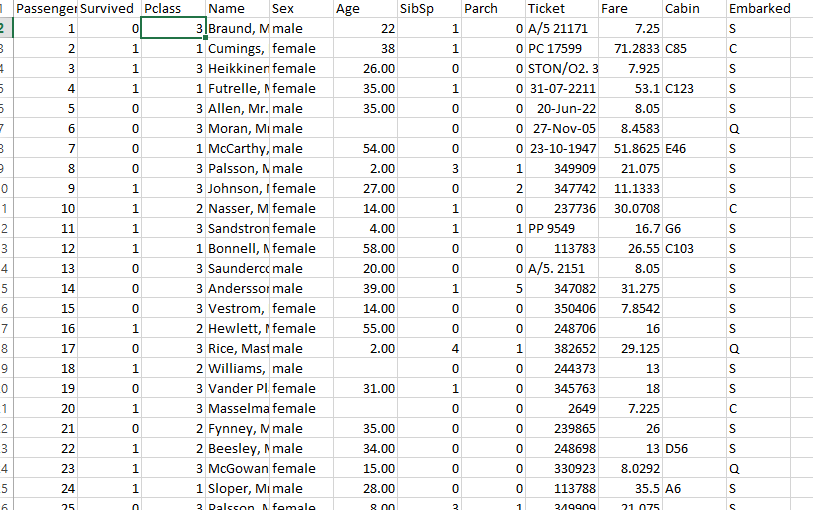
* **Action:** Click the warning icon (it usually appears near the selected cell or at the top of the column).
* **Action:** From the options that appear, choose Convert to Number.

 **Method B (Text to Columns - More Robust for mixed data or complex text-to-number conversions):**

* **Action:** Go to the Data tab on the Excel ribbon.
* **Action:** In the Data Tools group, click Text to Columns.
* **Action:** In Step 1 of the wizard, choose Delimited and click Next. (Unless your numbers have a fixed width, Delimited is generally fine even if there are no actual delimiters).
* **Action:** In Step 2, ensure no delimiters are selected (unless your numbers have unexpected delimiters). Click Next.
* **Action:** In Step 3, under Column data format, select General (Excel will try to convert it to the most appropriate type, usually Number) or explicitly Number.
* **Action:** Click Finish.
* **Result:** This forces Excel to re-evaluate the cells, converting text-formatted numbers into actual numbers. They should now right-align.

Fix Data Types for Date/Time Columns:-

Convert Text Dates to Actual Dates (Crucial First Step if they are text):



(**Convert Text Dates to Actual Dates (Crucial First Step if they are text):**

* **Action:** Select the entire column containing your dates (e.g., **Column G**).
* **Action:** Go to the Data tab on the Excel ribbon.
* **Action:** In the Data Tools group, click Text to Columns.
* **Action:** In Step 1 of the wizard, choose Delimited and click Next.
* **Action:** In Step 2, ensure no delimiters are selected (unless your dates have unexpected delimiters). Click Next.
* **Action:** In Step 3, under Column data format, select Date.
* **Action:** From the dropdown menu next to Date, select the **original format** of your text dates. This is very important! For example:
  + If your text dates are "01-01-2023" (Day-Month-Year), choose DMY.
  + If they are "Jan 1, 23" (Month-Day-Year), choose MDY.
  + If they are "2023/01/01" Year/Month/Day), choose YMD.
* **Action:** Click Finish.
* **Result:** Excel will attempt to convert the text entries into actual date values. If successful, they will typically right-align. Dates that couldn't be converted might remain left-aligned or show an error.

 Apply **Consistent Date Display Format (After they are actual dates):**

* **Action:** With the column of now-actual dates selected (**Column G**).
* **Action:** Right-click and choose Format Cells....
* **Action:** In the Number tab, select Date from the Category list.
* **Action:** Choose your desired consistent display format from the Type list (e.g., dd-mm-yyyy, 14-03-2012).
* **Action:** Click OK.

