Task 1: Data Cleaning and Preprocessing

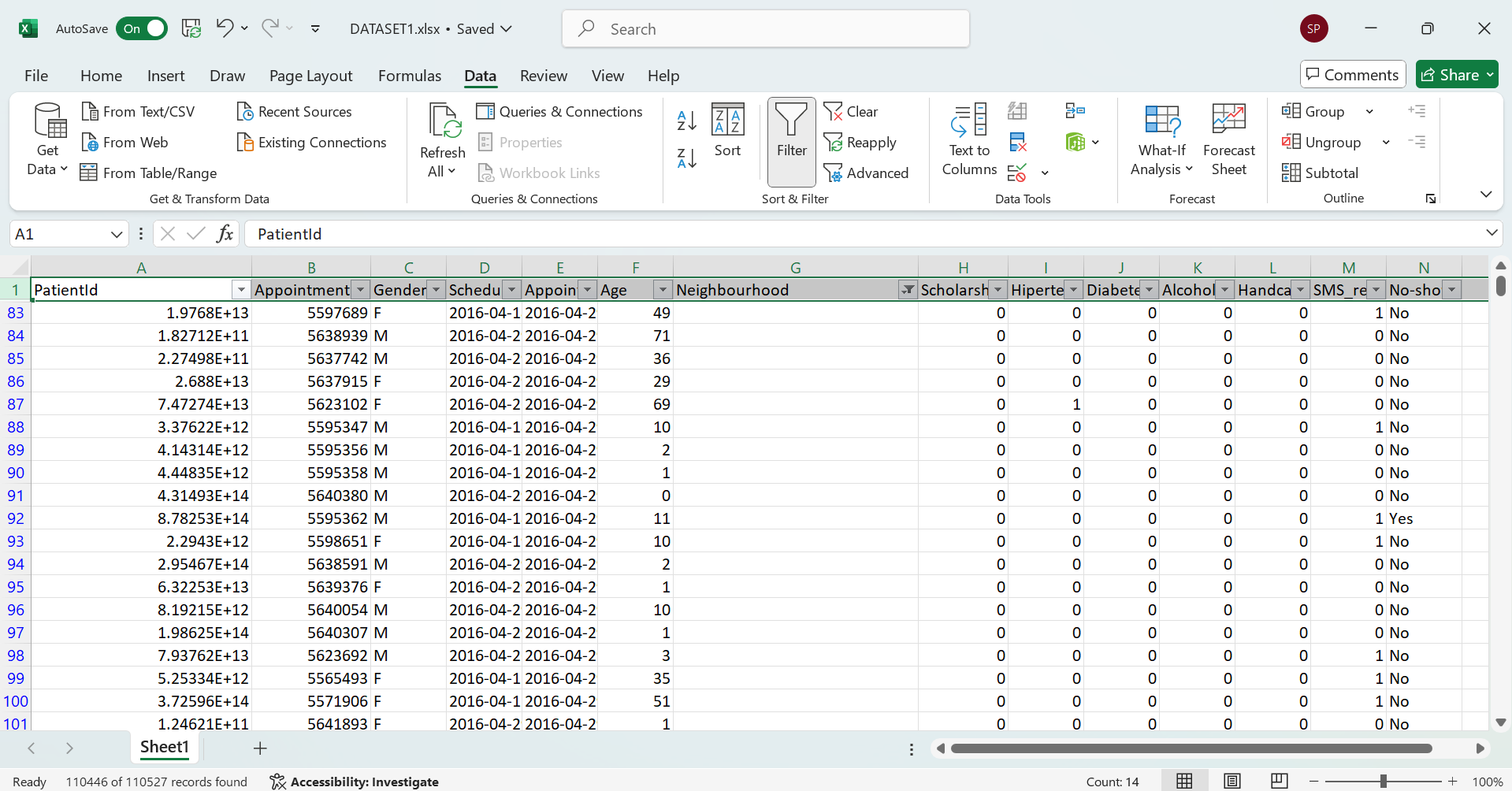
Objective: Clean and prepare a raw dataset (with nulls, duplicates, inconsistent formats).

Tools: Excel

Deliverables: Cleaned dataset + short summary of changes

Identify and handle missing values using .isnull() in Python or filters in Excel

Identify Missing Values Using Filters



Selected the blank rows using the filter function which shows only the blank spaces

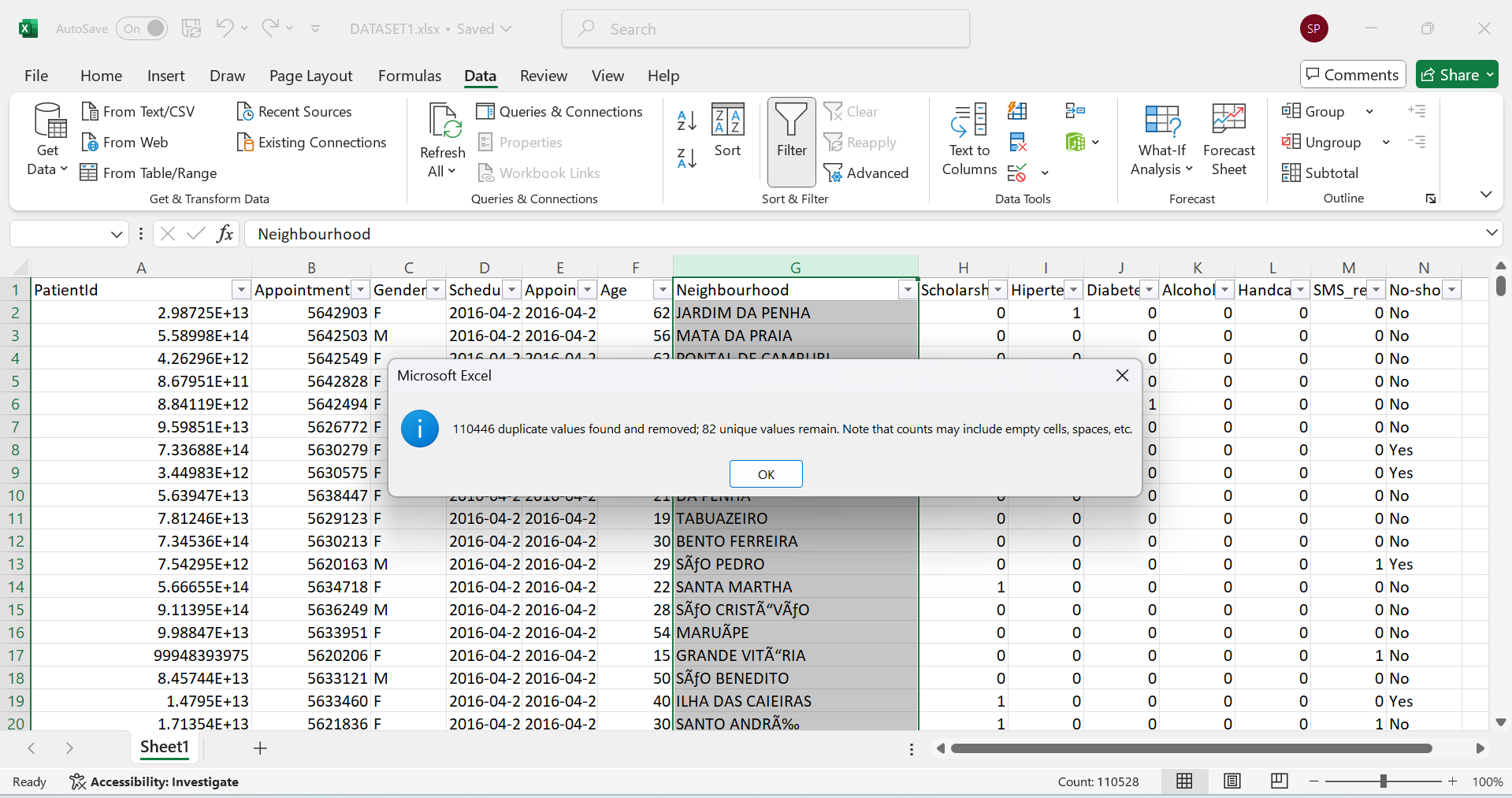
Handle Missing Values

* Highlight Missing Values (without deleting)  
  Select the entire data.
* Go to **Home → Conditional Formatting → New Rule → Format only cells that contain → Blanks**.
* Choose a color fill → Click **OK**.

This will **highlight all missing values**, making them easy to spot.

Remove duplicate rows using Excel’s “Remove duplicates”

Select the column where you want to remove the duplicate values and then select the “Remove duplicates” from the data tab.



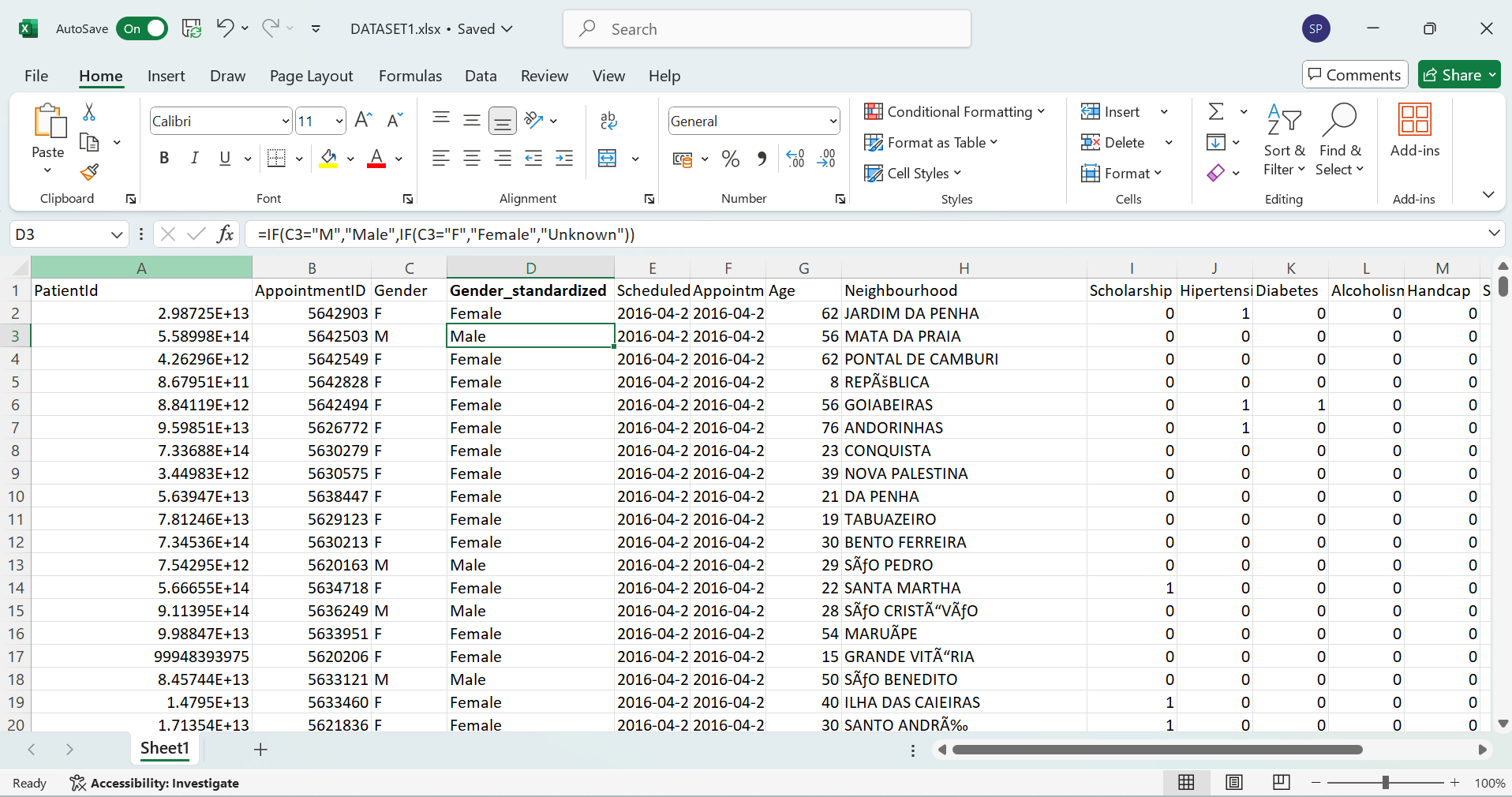
Standardize text values like gender, country names, etc

Standardize Gender

By creating a new column named “Gender Standardized”

Using the formula

=IF(C2="M","Male",IF(C2="F","Female","Unknown"))



Convert date formats to a consistent type (e.g., dd-mm-yyyy).

Changing the format into 10 characters of date hence using the below formula to extract the date and time

=LEFT(D2,10)

Creating a new column to each of the date column using the above formula   
Now select the new column

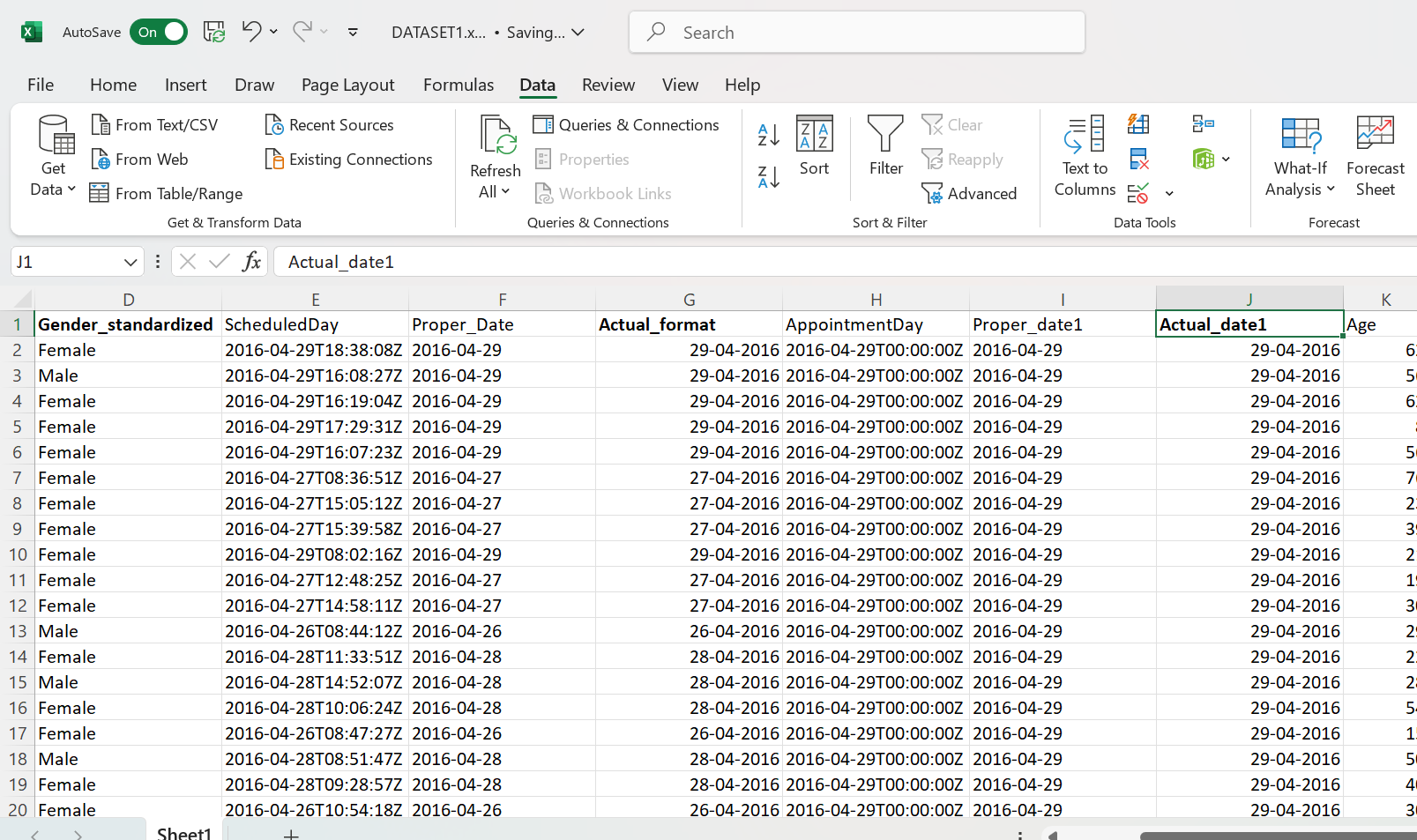
Go to **Data → Text to Columns → Finish** (this converts text to real date format).

Now format it:

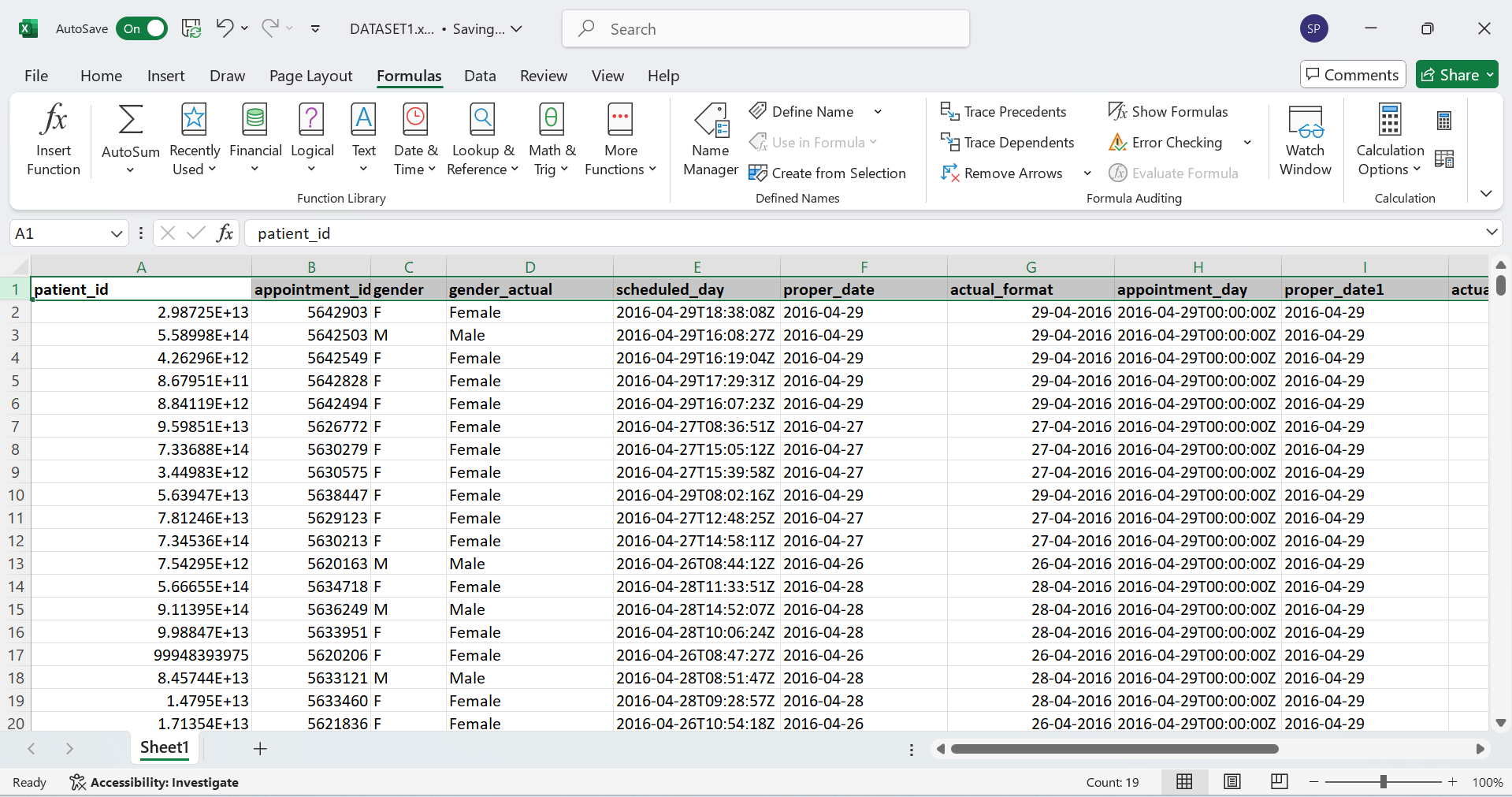
* Go to **Home → Number → Short Date**
* Then right-click → **Format Cells → Custom →** enter

dd-mm-yyyy

Now the dates shows consistently as 29-04-2016.



Rename column headers to be clean and uniform (e.g., lowercase, no spaces).



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

Change the age value to integer using the =INT(K2) column