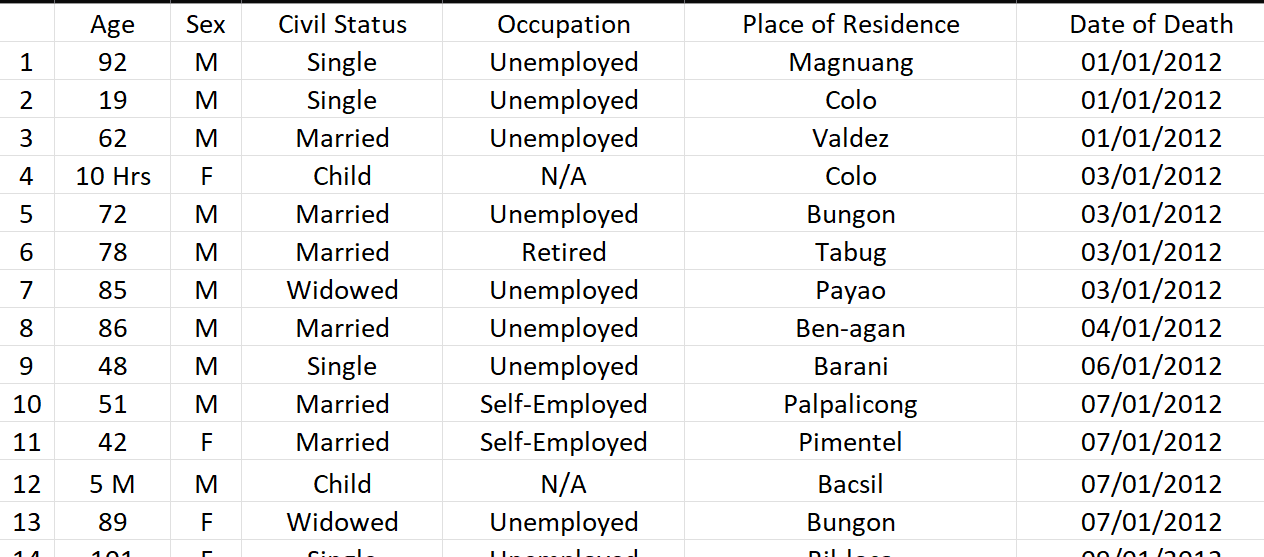
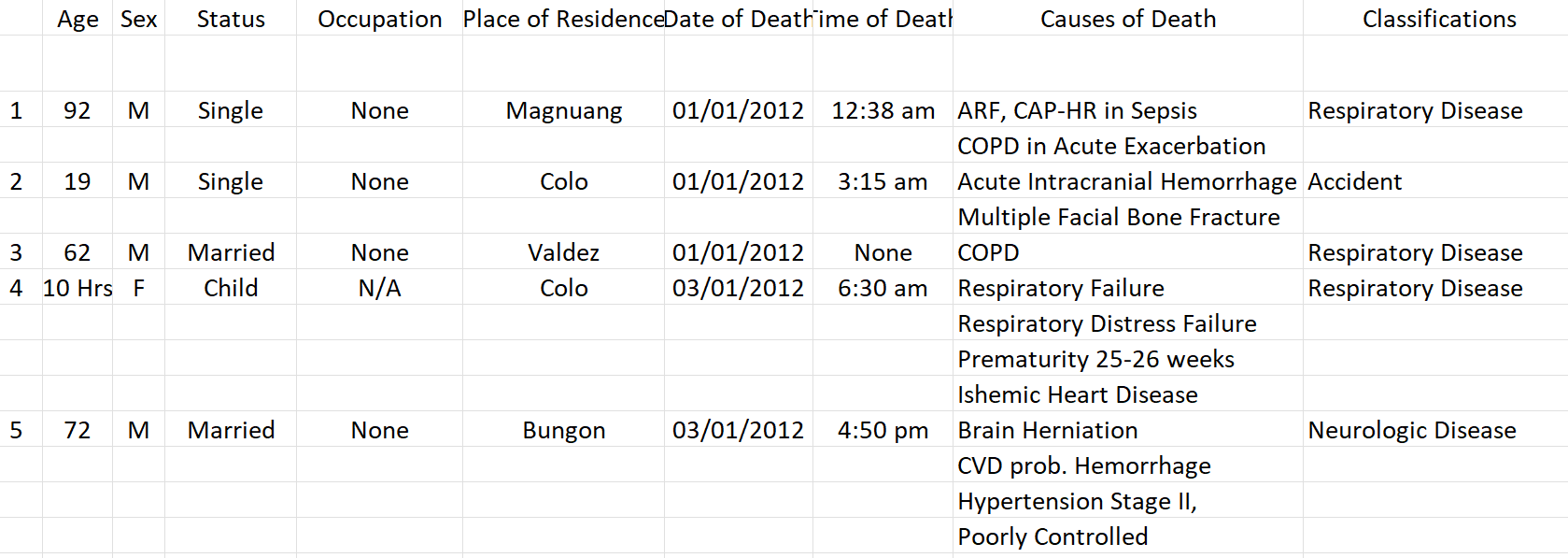
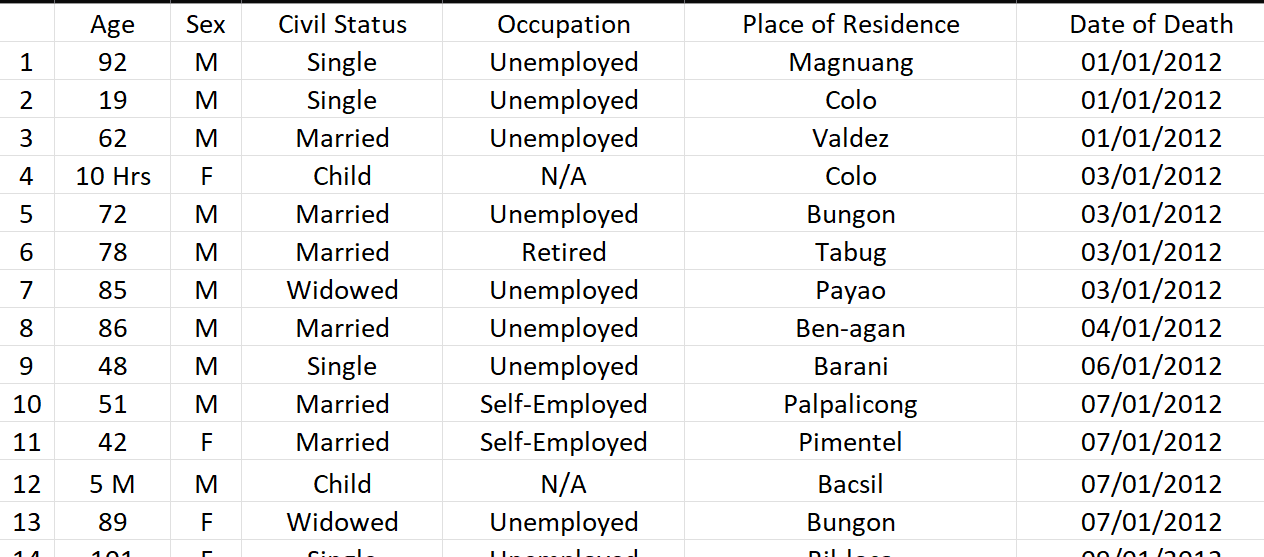
**Cleaning the Data**

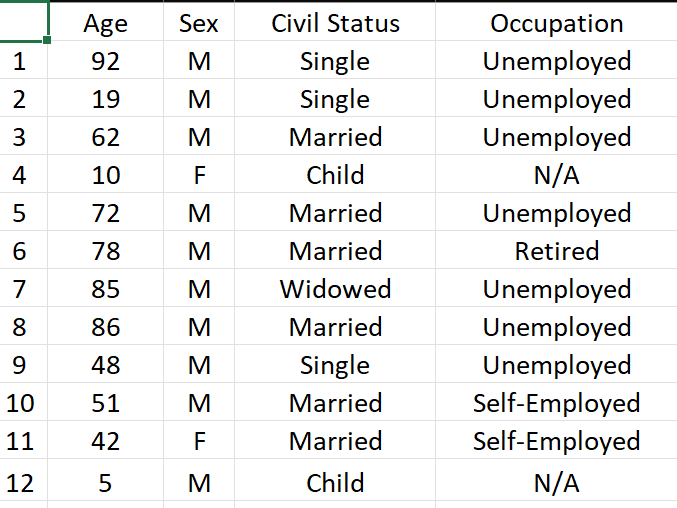
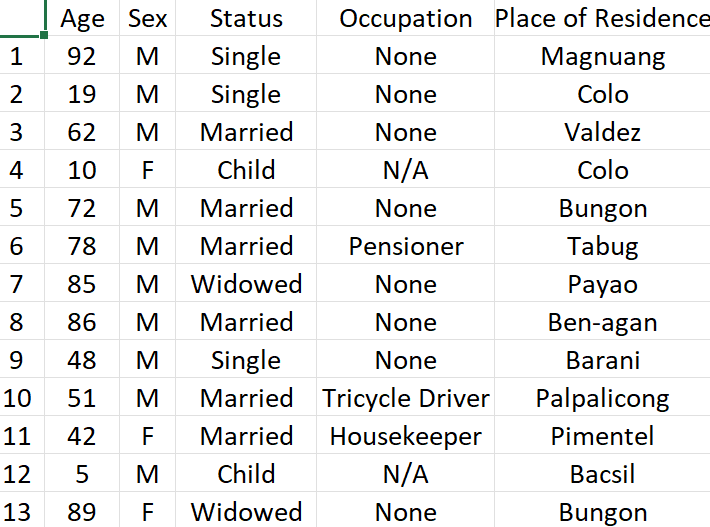
****

First, we carefully examined the dataset and removed all rows that were completely blank or lacked any data. This process ensured that only relevant and meaningful information was retained for analysis.

****



We noticed that the age column contained entries with letters like "Hrs," "D," and "M" alongside the numbers. To clean the data, we removed these letters, leaving only the numeric values for consistency and clarity.

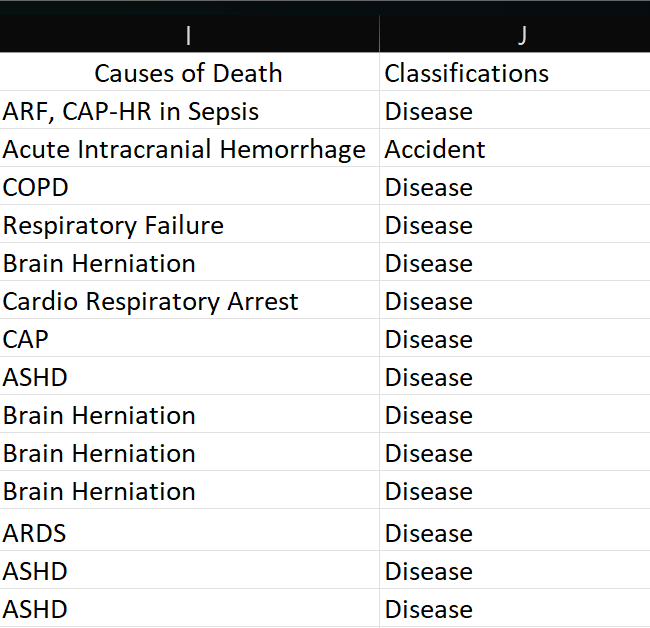
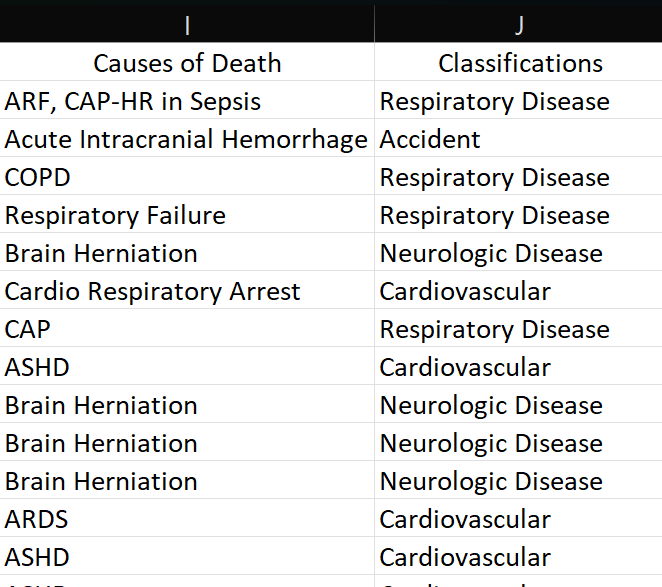


We also organized the data by classifying the occupations into specific categories.

1. None – Unemployed, Housewife and more.
2. Employed – Government Employee, Security Guard, Seaman, etc..
3. Self-Employed – Tricycle Driver, Housekeeper etc..
4. Retired – Retired Police, Ret., Retired Employee, Pensioner, etc..



We updated the date format in the dataset, changing it from **MM/DD/YYYY** to **DD/MM/YYYY**. This ensured consistency and aligned the format with the preferred standard for our analysis.



x`

We also grouped different diseases into a single category based on their similarities or common characteristics.

* Respiratory Disease
* Neurologic Disease
* Cardiovascular

will be classified as **Disease**.





**Cleansed Data**

**Raw Data**