# Income-Classification: Data Cleaning

## What is Data Cleaning?

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled.

If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset. But it is crucial to establish a template for your data cleaning process so you know you are doing it the right way every time.

## Steps involved in data cleaning:

- ✓ Remove duplicate or irrelevant observations
- ✓ Fix structural errors
- ✓ Filter unwanted outliers
- ✓ Handle missing data

## Data Cleaning In Dataset:

- **Dealing with Duplicates:** Duplicate observations are often encountered during data collection. In our dataset, consisting of 32,561 rows and 15 features, we identified and addressed 24 duplicated rows to ensure data integrity.

- **Handling Missing Data:**
  - We observed missing values denoted by '?' in the 'Occupation,' 'Workclass,' and 'Native-country' features. To handle this, we designated these rows as 'unknown.'
  - In **'marital-status'** feature it has three similar 'married' levels; these can be combined into one 'married' class.
  - The '**relationship'** variable has two similar values, wife and husband, which signals the same idea, of being a spouse.

- **Leading Whitespace Removal:**
  - In our dataset, 'object' type column names and string values often started with leading whitespace.
  - To enhance data consistency, we systematically removed these leading whitespaces.

- **Data Validation and Outlier Handling:**
  - **Age Column:** During data validation, we observed that the 'age' column had outliers, specifically individuals over the age of 80. It's reasonable for such outliers to exist since people tend to retire around that age.
  - Given that these outliers represented a negligible amount (264 rows) out of the 30,000+ observations, we chose to remove them as they couldn't be imputed effectively.

- **Hours-per-Week Column:** The 'hours-per-week' column displayed numerous outliers. Given the significant number of outliers and their impact on data integrity, we chose to use the Interquartile Range (IQR) method to impute outliers.
- This approach allowed us to handle outliers effectively while retaining the majority of the data.
- **'Capital-gain'** : The 'capital-gain' feature had extreme outliers and a wide range, with 1st and 3rd quartiles at 0, which needed to be retained.
    i. Extreme outliers (values >= 42,000) were capped at 42,000 to reduce their impact on the model while preserving the trend of high 'capital-gain' correlating with 'income' > 50,000. But still outliers are present.
    ii. Log transformation was initially applied to reduce the effect of outliers, but it significantly changed the distribution of the feature.
    iii. To simplify the transformation, the 'capital-gain' feature was converted into a binary class: 0 for no capital gain and 1 for having capital gain.
- **'Capital-loss'** Feature is filled with outliers. Their 1st and 3rd quartiles are also 0. We must retain these entries because we can always find a few such sections in most of the societies or countries.

➢ **Feature Encoding:**
- **Income feature:** The target feature in our dataset is a binary class classification with two categories, '<=50k' and '>50k,' represented as object type labels.
- To facilitate the deep learning process, i transformed these labels into numerical values, encoding '<=50k' as 0 and '>50k' as 1.
- This conversion enables us to use the target feature in a discrete binary classification context for our analysis.
- **Gender feature:** The 'gender' feature in our dataset is a binary class with two categories, 'Male' and 'Female.'
- To facilitate the machine learning process, we transformed these labels into numerical values, encoding 'Male' as 1 and 'Female' as 0.
- This encoding allows us to represent gender as a binary variable, which is commonly used in many classification tasks.

➢ **Data Type Casting:**
- In our dataset, the **'gender'** and **'Income'** features were originally in string format, represented with the *object data type*.
- To facilitate analysis and modeling, I have transformed these features into discrete numerical values, resulting in a change of data type to integers (*int*).
- This conversion enables us to work with these features in a more suitable format for our classification tasks.

## Conclusion:

Data cleaning is the foundation of our "income classification" project, ensuring the integrity and reliability of our dataset. Through addressing duplicate observations, managing missing data, handling outliers, promoting data consistency, and encoding key

features, we've prepared our dataset for robust analysis and modeling. Notable actions include the removal of 24 duplicate rows, strategic management of missing data using 'unknown' labels, and the systematic handling of outliers in various columns. By encoding the 'Income' and 'Gender' features, we've enabled effective binary classification. These comprehensive data cleaning efforts have prepared our dataset for the subsequent stages, securing data quality and consistency for the development of accurate predictive models in our project.