

## Data Cleaning & Manipulation

### ***Data Cleaning Definition:***

Data cleaning is the process of identifying and fixing inaccurate, incomplete, duplicated, or poorly formatted data to improve its quality.

#### Examples:

1. Removing Duplicates  
`df.drop_duplicates()`
2. Dropping Columns  
`df.drop(columns = "Not_Useful_Column")`
3. Strip  
`df["Last Name"].str.lstrip("...")`
4. Standardizing ID  
`df["Identity Document"].apply(lambda x: x[0:3] + '-' + x[3:6] + '-' + x[6:10])`
5. Splitting Column  
`df[["Street_Address", "State", "Zip_Code"]] = df["Address"].str.split(',', 2, expand = True)`
6. Replace Content  
`df["Paying Customer"].str.replace('Yes', 'Y')`

### ***Data Manipulation Definition:***

Data manipulation is the process of organizing, restructuring, and transforming data to make it more useful for analysis.

#### Examples:

1. Sorting Data  
`df.sort_values(by="Sales", ascending=False)`
2. Merging Two DataFrames  
`df_merged = pd.merge(df_orders, df_customers, on="Customer_ID", how="inner")`
3. Pivoting Data  
`df_pivot = df.pivot_table(index="Region", columns="Product", values="Sales", aggfunc="sum")`
4. Filtering Rows  
`df_filtered = df[df["Sales"] > 5000]`
5. Aggregating Data  
`df.groupby("Category")["Sales"].sum()`