Session 2: Data Cleaning Activity

Handle Missing Values

Checking for missing values using df.isnull().sum(). The output shows there are no missing values found (zero missing values in all columns).

Unnamed:	0
Job Title	0
Salary Estimate	0
Job Description	0
Rating	0
Company Name	0
Location	0
Headquarters	0
Size	0
Founded	0
Type of ownership	0
Industry	0
Sector	0
Revenue	0
Competitors	0
Easy Apply	0

Handling duplicate values

To check duplicate values, using df.duplicated().sum(), We found no duplicates in this dataset.

Standardizing the columns

df.columns = df.columns.str.strip().str.lower().str.replace(' ', '_') using this code, we standardized the column names by removing whitespaces, converting column names into lowercase, replacing all spaces in column names with '_'.

Using df['rating'] = pd.to_numeric(df['rating'], errors='coerce')

df['founded'] = pd.to_datetime(df['founded'], errors='coerce')

we ensure the columns have consistent and correct data types for numerical and date/time analysis, while handling invalid formats by coercing them into missing values(NA).

Removing outliers

Using Interquartile Range (IQR) method, removing outliers.

Final summary after data cleaning

Index: 81 entries, 0 to 99

Data columns (total 16 columns):

Unnamed:	81 non-null
Job Title	81 non-null
Salary Estimate	81 non-null
Job Description	81 non-null
Rating	81 non-null
Company Name	81 non-null
Location	81 non-null
Headquarters	81 non-null
Size	81 non-null
Founded	81 non-null
Type of ownership	81 non-null
Industry	81 non-null
Sector	81 non-null
Revenue	81 non-null
Competitors	81 non-null
Easy Apply	81 non-null