# Understanding, Cleaning and Visualizing Data Through Pandas

## Data Cleaning and Preprocessing

1. **Detailed Parsing of Salary Information**:
   - **Remove Currency Symbols**: Remove ”$” or other currency symbols from ”Salary Estimate” to simplify the salary parsing process.
   - **Salary Type Extraction**: Create a column that identifies whether the salary is presented as ”per year,” ”per hour,” etc., based on common indicators in ”Salary Estimate.”

2. **Handle Special Characters in Text Columns**:
   - **Text Standardization**: Remove or replace special characters (e.g., ”\n”, ”\t”, and other symbols) from columns like ”Job Description” and ”Company Name” for consistency.
   - **Lowercasing**: Convert all text columns (e.g., ”Job Title,” ”Company Name,” ”Industry”) to lowercase to standardize the format and avoid case sensitivity issues during analysis.

3. **Company Name Refinement**:
   - **Remove Rating from Company Name**: Clean up ”Company Name” by removing any appended rating values (e.g., ”Company Name\n3.8”).
   - **Unique Company Identifier**: Generate a unique identifier for each company based on the cleaned ”Company Name” to help with comparisons.

4. **Standardizing Location Data**:
   - **Expand State Abbreviations**: For the ”Location” column, expand state abbreviations (e.g., ”CA” → ”California”) to ensure uniformity.
   - **Location Format Consistency**: Check for any inconsistent formats in ”Location” and ”Headquarters” and standardize to ”City, State” format.
   - **Distance Calculation**: If city and state information are available, create approximate distance calculations between ”Location” and ”Headquarters” to identify remote positions.

5. **Revenue Column Parsing and Cleaning**:
   - **Extract Revenue Range**: For each value in ”Revenue,” extract minimum and maximum revenue values to separate columns.
   - **Currency Standardization**: Ensure all revenue values are in the same currency (e.g., USD) and handle conversions if any foreign currencies are detected.
   - **Fill Missing Values in Revenue**: Fill missing values in the ”Revenue” column.

6. **Missing Data Handling**:
    - **Fill Missing Values in Categorical Columns**: Fill missing values in columns such as ”Industry” and ”Sector.”
    - **Fill Missing Values in Rating**: Fill missing values in ”Rating.”

7. **Cleaning and Extracting Data from ”Job Description”**:
    - **Length Calculation**: Create a column to calculate the word count or character length of each ”Job Description” to help filter lengthy descriptions.
    - **Remote Work Flag**: Identify if the job allows remote work by checking for keywords (like ”remote,” ”telecommute”) in the ”Job Description”.
    - **Experience Level Parsing**: Add columns to identify required experience levels (e.g., ”entry level,” ”5+ years”) by parsing ”Job Description.”

8. **Additional Column Standardization**:
    - **Numeric Conversion of Founding Year**: Ensure ”Founded” is a numeric column, and set a minimum year threshold (e.g., 1800) to filter out any erroneous entries.
    - **Size Categorization**: For ”Size,” create size categories (e.g., ”small,” ”medium,” ”large”) based on ranges in employee count.

9. **Outlier Detection and Handling**:
    - **Salary Outliers**: Detect outliers in the salary range and flag unusually high or low salary estimates for each job title.
    - **Rating Outliers**: Identify extreme ratings (e.g., below 2.0 or above 4.5) and flag them to verify their accuracy.
    - **Founded Year Outliers**: Flag companies with founding years outside a realistic range as potential data entry errors.

10. **Column Optimization and Data Type Conversion**:
    - **Convert Numerical Columns**: Ensure columns like ”Rating” and ”Founded” are in a suitable numeric format (`int64` or `float64`) and convert others as appropriate.
    - **Memory Optimization**: Downcast numeric columns (e.g., ”Rating,” ”Founded”) to the smallest integer or float type without data loss to optimize memory usage.

---

## Exploratory Data Analysis (EDA)

1. **Salary and Rating Analysis**:

   - **Salary Range Analysis**: Calculate the median, mean, and standard deviation of salary ranges across different job titles, industries, and sectors.
   - **Top and Bottom Companies by Salary**: Identify companies that offer the highest and lowest average salary ranges.
   - **Rating Distribution Analysis**: Calculate the proportion of companies with ratings in different ranges (e.g., <3.0, 3.0-4.0, >4.0).

2. **Sector and Industry Analysis**:

   - **Sector Diversity**: Count the number of unique industries within each sector, and analyze if certain sectors have a higher concentration of industries.
   - **Top Industries for Data Science**: List the top 5 industries with the most job postings for data science roles and their average salaries and ratings.
   - **Revenue by Sector and Industry**: Calculate the average revenue for each sector and industry to identify the most profitable areas.

3. **Location and Company Insights**:

   - **Top Locations by Job Count**: Identify the top 10 locations (cities and states) with the highest job counts and analyze their average salaries and ratings.
   - **Headquarters vs. Job Location**: Compare ratings for jobs at a company's headquarters vs. remote offices.
   - **Company Size and Location Correlation**: Explore whether certain locations are more likely to host larger or smaller companies.

4. **Experience Level Analysis**:

   - **Job Title Levels**: Calculate the average rating and salary for each level (e.g., ”Junior,” ”Mid-level,” ”Senior”).
   - **Experience Level and Location**: Analyze if certain experience levels (e.g., entry-level, senior) are more commonly available in specific locations or industries.
   - **Experience Level by Sector**: Analyze the distribution of experience levels by sector and identify if certain sectors focus more on senior or junior roles.

5. **Tricky EDA**:

   - **Founding Year vs. Salary**: Analyze if older companies tend to offer higher or lower salaries compared to newer companies.
   - **Remote Work and Rating**: Compare the ratings of jobs marked as remote vs. non-remote to see if remote positions correlate with higher ratings.
   - **Competitors and Salary**: Check if companies with multiple listed competitors tend to offer higher salaries on average.

### Data Transformation

1. **Advanced Keyword Tagging**:

   - **Add Skill Flags**: Create binary columns to flag if certain skills (e.g., ”Python,” ”SQL,” ”machine learning”) are mentioned in the ”Job Description.”
   - **Identify Benefits**: Add columns for common benefits (e.g., ”healthcare,” ”retirement plan”) based on keywords found in the ”Job Description.”
   - **Remote Job Tagging**: Add a column to indicate if the job allows remote work based on keywords in ”Job Description” or ”Location.”

2. **Time-Based Transformation**:

   - **Convert Founded Year to Age**: Calculate each company’s age by subtracting the founding year from the current year.
   - **Establish Company Age Categories**: Categorize companies into age groups (e.g., ”New” for <10 years, ”Established” for 10-50 years, ”Legacy” for >50 years).
   - **Estimate Job Posting Age**: If date information is available, calculate how long ago each job posting was listed.

3. **Categorization and Grouping**:

   - **Sector and Industry Mapping**: Create a mapping dictionary to reclassify ”Industry” values into broader ”Sector” categories if needed.
   - **Location Grouping**: Group locations into regions (e.g., Northeast, Midwest) for U.S. data or other relevant geographical groupings.
   - **Company Size Bucketing**: Create buckets for company size (e.g., ”Small” for <500 employees, ”Medium” for 500-5000, ”Large” for >5000) and add a new column.

4. **Advanced Revenue and Salary Conversion**:

   - **Calculate Revenue per Employee**: Create a column for revenue per employee by dividing the revenue by the employee size estimate.
   - **Salary Normalization**: Normalize salary ranges (min and max) to a single yearly estimate for comparison, especially for hourly roles.
   - **Create Salary Range Buckets**: Classify salary ranges into brackets (e.g., ”<$50k,” ”$50k-$100k,” ”>$100k”) to enable clearer comparisons.

5. **Competitor Data Transformation**:

   - **Competitor Count**: Add a column indicating the number of competitors each company lists.

- **Frequent Competitors**: Add a flag for companies that list commonly seen competitors and analyze if these companies have different rating or salary trends.
- **Common Competitor Networks**: Create clusters for companies with the same competitors, allowing analysis of companies within specific competitive networks.

---

## Comparative Analysis

1. **Sector vs. Industry Comparison**:
   - **Rating Comparison**: Compare the average rating across industries within each sector to see if certain industries outperform others within the same sector.
   - **Salary Range by Industry**: Compare the median salary ranges for each industry within top sectors.
   - **Revenue vs. Sector**: Analyze if certain sectors tend to have higher revenue brackets compared to others.
2. **Ownership Type Analysis**:
   - **Public vs. Private Salary Comparison**: Compare average salaries between public and private companies.
   - **Rating Distribution by Ownership**: Compare rating distributions across different ownership types to see if public or nonprofit organizations have higher ratings.
   - **Company Size by Ownership**: Analyze if public or private companies tend to be larger or smaller in terms of employee size.
3. **Company Age vs. Rating**:
   - **Compare Ratings by Age Group**: Analyze average ratings of companies in each age group to see if older companies have higher ratings.
   - **Rating Stability Over Time**: Examine if companies founded before 2000 have more consistent (less variable) ratings than newer companies.
   - **Sector Popularity by Age**: Compare the prevalence of different sectors within each age group, observing which sectors are popular among older vs. newer companies.
4. **Location-Based Comparison**:
   - **Salary by Region**: Compare average salary ranges across regions to see if certain regions offer higher pay for data science jobs.
   - **Remote vs. Onsite Roles**: Compare salary ranges, job counts, and ratings for remote vs. onsite positions.

  - **Job Count by State**: Compare the total job count and average salary for each state to identify states with the highest job availability.

5. **Tricky Comparative Analysis**:
   - **Competitor Impact on Salary**: Compare salary ranges for companies with and without listed competitors to analyze if competition impacts salaries.
   - **Industry Longevity Analysis**: Compare average founding years by industry to see if certain industries have more established companies.
   - **Rating and Salary Disparity**: Analyze if companies with higher ratings offer salaries that deviate more from industry or sector norms.

---

### Advanced Data Insights and Parsing

1. **Competitor Relationships and Analysis**:
   - Identify and count the most frequently listed competitors across all companies. Create a column to show how many times each competitor is mentioned.
   - For companies with multiple competitors listed, calculate the average rating difference between them and their competitors. Are highly-rated companies more likely to list high-profile competitors?
   - Create a column indicating whether a company shares competitors with any top-rated companies (Rating 4), and analyze if this has any effect on its own rating.
2. **Industry and Sector Dynamics**:
   - Determine the sectors with the most diversity in industries and analyze if this correlates with higher salary estimates or job counts.
   - Find the industries with the highest average ratings and analyze which specific sectors they belong to. Does belonging to a specific sector influence the industry's average rating?
   - Identify industries with the highest number of job postings and check if these align with high salary estimates. Is there a trend between industry popularity and salary?
3. **Company Size and Growth Patterns**:
   - Group companies by ”Size” and analyze the average founding year for each size group. Are larger companies generally older, or is there a trend of newer, fast-growing companies in the dataset?
   - Create an ”Employee Growth Index” by dividing the number of employees (from ”Size”) by the number of years since founding. Compare this index across sectors to identify fast-growing sectors.

   - For each company, calculate the ratio of the company size to the number of competitors listed. Does a larger company tend to have more competitors?

4. **Revenue and Financial Analysis**:

   - For companies with available revenue data, calculate the revenue per employee (using ”Revenue” and ”Size”) and analyze which sectors have the highest revenue per employee.
   - Determine the average revenue for each sector and examine if high-revenue sectors have higher-rated companies on average.
   - Identify companies with the widest revenue ranges and examine if these companies have more job postings or higher ratings.

5. **Salary Insights and Trends**:

   - For each job title, calculate the average salary range width (difference between max and min salary) to see which roles have the most salary variability.
   - Identify the top 5 job titles with the highest average salary ranges and analyze if they are concentrated in certain sectors or locations.
   - Compare median salary ranges between private and public companies to determine if there’s a significant difference based on ownership type.

6. **Job Description Keywords and Popularity**:

   - Identify the top 10 keywords in ”Job Description” across different sectors. Are certain keywords unique to specific sectors?
   - Parse common industry-specific terms from ”Job Description” (e.g., ”fintech” for finance, ”sustainability” for environmental) and analyze if jobs mentioning these terms have higher average ratings.
   - Check the percentage of job descriptions mentioning benefits like ”remote,” ”flexible hours,” or ”bonus,” and analyze if jobs mentioning these perks have higher average ratings.

7. **Competitor and Rating Correlation**:

   - Calculate the average rating for companies that list specific high-profile competitors (e.g., Google, Microsoft) and compare it to the ratings of companies without those competitors.
   - Create a ”Competitor Count Impact Score” by examining if companies with a higher number of competitors listed tend to have higher or lower ratings.
   - Analyze if companies with unique competitors (listed only once) have distinct characteristics in ratings or salary compared to those with common competitors.

### Visualization

1. **Basic Visualizations**:
   - Plot the distribution of company sizes (in terms of employee numbers) and see how it aligns with ”Type of Ownership” (e.g., private, public, nonprofit).
   - Create a bar chart showing the average rating for each ”Type of Ownership” to see if ownership types correlate with employee satisfaction.
2. **Location-Based Visualizations**:
   - Generate a bar chart to compare the number of job postings across the top 10 cities, highlighting popular cities for data science roles.
   - Create a map plot to show the geographic distribution of job postings by state, using color intensity to represent job density.
3. **Salary and Rating Analysis**:
   - Create a scatter plot showing the relationship between average salary range and company ratings, and add a trend line to analyze if higher-rated companies tend to offer higher salaries.
   - Plot a histogram of salary ranges for different sectors, allowing comparison of sector-based salary distributions.
   - Generate a heatmap to visualize the correlation matrix between numerical columns (e.g., ratings, salary estimates, founding year) to identify potential relationships.
4. **Tricky Visualizations**:
   - Plot a line graph of the average rating of companies over time by ”Founded Year” to see if there’s a trend in ratings for companies founded in different decades.
   - Create a box plot showing the salary range distributions by industry, highlighting which industries have the widest salary range.
   - Generate a stacked bar chart of job titles within each sector to observe which titles are most common in different industries.
5. **Competitor and Popularity Insights**:
   - Create a network graph to show connections between companies and their competitors, with node sizes based on the company’s average rating.
   - Plot a bar chart showing the top 10 most common competitors listed, along with the average rating of companies that have these competitors.
6. **Company and Industry Focused Visualizations**:
   - Create a stacked area chart to show how job postings by company size (”Size”) are distributed over time (based on ”Founded Year”).

- Plot a donut chart representing the distribution of job postings by ”Type of Ownership” to visualize how job availability varies by ownership type.
- Generate a radar chart (spider chart) comparing key metrics (e.g., average rating, median salary, company size) across top sectors to easily compare sector strengths.