Stat 220 Lab 1: Analyzing Data Scientist Salaries

In this assignment, you will explore and analyze a dataset containing salary information for 581 data scientists. The dataset can be accessed at the following URL: `https://richardson.byu.edu/220/ds_salary.csv`. Additionally, descriptions of the data variables can be found here: `https://huggingface.co/datasets/hugginglearners/data-science-job-salaries`. Please note that the dataset provided has been slightly modified from its original form.

Task Overview

Your objective is to analyze how data scientist salaries vary in relation to three key factors: experience, company size, and job title. To achieve this, you will create plots and tables that facilitate a clear understanding of these relationships.

Instructions

1. **General Salary Data:**
    - Start by plotting and describing the general shape of the salary data. This initial step sets the foundation for your analysis.
2. **Effect of Experience on Salary:**
    - Investigate how experience influences data scientist salaries. Create visualizations or tables to illustrate any trends or patterns.
3. **Impact of Company Size on Salary:**
    - Explore how company size affects data scientist salaries. Generate plots or tables to visualize this relationship.
4. **Job Title Analysis:**
    - Job titles can vary significantly. For this project, select a keyword or phrase to distinguish between different job titles. For instance, you can differentiate between job titles containing the word "Analyst" and those that do not. To achieve this, you may find the `str.contains` function in Pandas helpful. Refer to this resource for examples: `https://www.geeksforgeeks.org/python-pandas-series-str-contains`. Explore various job titles to decide on the keyword or phrase you want to focus on.
5. **Interactions Between Relationships:**

- Investigate potential interactions between the three factors (experience, company size, and job title). Create a two-way table that displays essential statistics at the intersection of two variables. For instance, construct a table with job titles on the left, company sizes at the top, and each cell representing the mean salary for each job title/company size combination.

6. **Conclusion:**

   - Summarize your findings by listing all significant relationships you have discovered during your analysis.

Deliverables

Your assignment report should consist of six sections, each addressing one of the above tasks. While these sections need not be overly detailed, they should be concise and clearly explain the purpose of your analysis. They should also be clearly labeled. One easy way to do this is to use the text boxes in a Python notebook. Here is guide for formatting text in a Python notebook: `https://www.earthdatascience.org/courses/intro-to-earth-data-science/file-formats/use-text-files/format-text-with-markdown-jupyter-notebook/`. For example, you can use headers, i.e. "# Header 1" to organize each section. If you do this, ensure that each section includes relevant plots and/or tables to support your findings.