Data Science Job Project

**Data and Introduction:**

Data scientist has been named the “sexiest job of the 21st century” (<https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century>), but what exactly does that mean? As students majoring in BAIS, many of us are looking to careers post-graduation and wonder about working in this field. Specifically, what are the benefits of working in this Data Science field and what does the field look like in terms of working arrangements, job opportunities, and salary. In this project, we apply data-driven analysis using Python to visualize and analyze key metrics in the field to gain insight. Our Python analysis and visualization may prove useful for students and other individuals interested in data science careers. This project uses data gathered from a [2023 Kaggle dataset](https://www.kaggle.com/datasets/ruchi798/data-science-job-salaries?resource=download)) of the data science industry and salaries within the industry. The dataset contains 607 rows of unique data across 12 different fields.

**Functions used during analysis:**

Our first function, processData(), processes all of our csv data in the dataframe to clean it up before we perform our analyses. The function takes our dataframe as input to clean the data, and it outputs a print statement stating the data had been processed. The function cleans the dataframe through code of dropping columns, replacing data values, changing datatypes, and renaming columns.

Our second function, getTopJobTitles(), retrieves the top 5 most common job titles in the data science field. The function uses the count of the top 5 job titles in the sample and their titles as input. This function outputs a print statement listing the top 5 data science jobs with their respective popularity and a bar chart displaying the top job titles and their count. We found the top 5 job titles by using the values\_count() method which found the count of all job titles in the sample and filtered it down to only the top 5 job titles.

Our third function, getRemoteWork(), evaluates the working arrangements of jobs in the data science industry. This function uses the percentages of each working arrangement and their type as input. The function outputs a print statement listing the percentage of people working remotely, hybrid, or in-person and outputs a pie chart of the working arrangements and their percentages. We found the percentage of each working arrangement by utilizing the values\_count() method to find the amount of people working remotely, hybrid, and in-person and then divided these values by the total amount of people in the sample to provide the percentages of each working arrangement

Our fourth function, getSalary(), evaluates the overall salaries for data science jobs through a histogram. The function uses the values of the salary column in the dataframe as input. The function outputs a print statement listing the median and average salary values and outputs a histogram showing the frequency of salaries.

Our fifth function, getSalaryByExperience(), evaluates the mean salary of people in the data science industry by experience level. The function uses the mean salary by experience level as input, and it outputs a print statement listing the mean salary per experience level and outputs a bar graph showing the mean salary by experience level. We found the mean salary by experience by grouping the dataframe by experience level and finding the mean salary value for each group.

Our sixth function, getSalaryByCompany(), evaluates the mean salary of data science jobs by company size. The function uses the mean salary by company size as input. The function outputs a print statement listing the mean salary per company size and outputs a bar graph showing the mean salary by company size. To find the mean salary by company size we grouped the dataframe by company size and found the mean salary for each group.