**Introduction**

**Guiding Questions**

1. How was salary affected by each feature?

The columns are univariate and numerical, for the univariate data, transfer them to categorical to see the effects; for numerical data, make different plots to analyze.

1. Which feature affects salary the most, experience level or job title?

Use EDA to compare and analyze these two factors by plotting

1. Find a combination of features that can earn the highest salary?

After finishing all analysis of this dataset, we can get a general understanding about different people’s salary, thus we can find some combinations that can earn a high salary than others.

**Dataset**

This dataset was found on Kaggle, it is an online community platform for data scientists and other data workers and provides almost 80,000 datasets including many areas such as computer science, education, and internet, all the datasets are open to public.

There are 11 columns and 607 rows in this csv format dataset, it’s a survey form on the salary of data science practitioners, questions including annual salary (in USD), experience level: entry, junior, senior or expert; the type of employment: is the responder work in full-time or part-time; remote ratio: the overall amount of work done remotely and the company size: the average number of people that worked for the company during the year: S means there are less than 50 employees, M means 50 to 250 employees and L means more than 250 employees; job title: like data scientists, data analyst, data engineer and machine learning engineer.

The data was ordered by the year the salary was paid in ascending order from 2020 to 2022.

**Tasks**

1. There is a column called “salary” which represented the total gross salary paid different currency should be deleted since they are already be converted to USD
2. Group all numeric and categorical data, 先逐个分析，最后再按组分析。
3. Numerical data can use pie chart, categorical data can use bar chart.
4. For “Job Title”, and “company location” we can apply NLP to find the frequency for each title and plot a WordCloud to visualize