Project Proposal

1. **Dataset description:**
   1. Dataset: Tech job positions and salaries from glassdoor.com

Source: Kaggle.com

Link: [Jobs](https://www.kaggle.com/datasets/thedevastator/jobs-dataset-from-glassdoor/?select=salary_data_cleaned.csv) [Dataset](https://www.kaggle.com/datasets/thedevastator/jobs-dataset-from-glassdoor/?select=salary_data_cleaned.csv) [from](https://www.kaggle.com/datasets/thedevastator/jobs-dataset-from-glassdoor/?select=salary_data_cleaned.csv) [Glassdoor](https://www.kaggle.com/datasets/thedevastator/jobs-dataset-from-glassdoor/?select=salary_data_cleaned.csv)

* 1. As Computer Science students, we are interested in how much we can make in the future. We want to explore how much data scientists earn compared to other professions using the Glassdoor dataset. Additionally, we aim to understand the factors that influence these salaries and make predictions about future salaries.
  2. File **salary\_data\_cleaned.csv** has 28 columns and 742 observations on this data, which means there are 28 variables.

**Job Title**: The job title (string).

**Salary Estimate**: Estimated salary (string).

**Job Description**: Description of the job (string).

**Rating**: Company rating (numeric).

**Company Name**: The name of the company along with its rating (string).

**Location**: The location of the company (string).

**Headquarters**: The headquarters of the company (string).

**Size**: The number of employees in the company (string).

**Founded**: The year the company was founded (numeric).

**Type Of Ownership**: The type of ownership of the company (string).

**Industry**: The industry the company is focused on (string).

**Sector**: The sector the company operates in (string).

**Revenue**: The company's revenue (string).

**Competitors**: Names of competitors (string) or -1 if there are none.

**hourly**: A binary value indicating if the company pays hourly wages.

**employer\_provided**: A binary value indicating if the salary is employer-provided.

**min\_salary**: The minimum salary offered (numeric).

**max\_salary**: The maximum salary offered (numeric).

**avg\_salary**: The average salary offered (numeric).

**company\_txt**: The name of the company (string).

**job\_state**: The state where the job is located (string).

**same\_state**: A binary indicator of whether the job is in the same state as the person viewing the job (string).

**age**: The age of the person viewing the job (numeric).

**python\_yn**: A binary indicator of whether the person knows Python (string).

**R\_yn**: A binary indicator of whether the person knows R (string).

**spark**: A binary indicator of whether the person knows Spark (string).

**aws**: A binary indicator of whether the person knows AWS (string).

**excel**: A binary indicator of whether the person knows Excel (string).

1. **Data questions:**
   1. The response variables**: min\_salary, max\_salary, avg\_salary**

The potential predictors**: Job Title, Rating, Location, Size, Type Of Ownership, Industry, Revenue, job\_state, python\_yn, R\_yn, spark, aws, excel**

* 1. We will prioritize prediction to achieve our final goal.
  2. Is it possible to create a predictive model that accurately estimates job salaries based on job title, location, industry, company size, and specific skills? Additionally, can we explore the key factors influencing job-related decisions and understand the relationships between these factors for potential inference?

1. **Models and methods:**
   1. There are 2 models that we think it can be applied to our project:
2. ***Linear Regression:***
   * + This model is good to predict numeric outcomes.
     + Advantages:
3. Provides interpretable coefficients to understand predictor impacts
4. Easy to implement and computationally efficient
5. Suitable for large datasets
6. ***Random Forest:***
   * + Purpose: Captures complex relationships in the data and handles both numeric and categorical predictors.
     + Advantages: Highly robust and versatile, suitable for large datasets, less prone to overfitting due to averaging over multiple trees and can handle missing values and maintain accuracy.
   1. The methods we may use for these two models include: data wrangling, model training, model evaluation with k-fold cross-validation, model comparison, interpretation, model selection.