1. Motivations & Your Questions

- **Motivations:**

Data is everywhere today, and data Science is a field that manages, researches, analyzes data, and thus provides meaningful information to non-professionals. It's one of the hottest professions of the 21st century. As a result, data scientists are in high demand and receive competitive salaries. However, the salaries for data scientists may vary widely due to a range of factors. Understanding these variations is crucial for aspiring data scientists. For students planning their career path, knowing which factors most influence salary can help them make informed decisions about their skill development and education. In this project, we aim to analyze the factors, such as the level of formal education, company type, and the number of proficient languages, that affect the salary of a data scientist. By identifying these patterns, we hope to provide valuable insights and practical reference for those in need.

- **Questions :**
    1. **Background: Does studying a master degree have a great impact on salaries?**
    2. **Programming Language: What kind of programming languages do software engineers have to code if they want to earn a decent salary?**
    3. **Location: The work location has a significant impact on employees' average salaries.**
    4. **How to choose between startup or big companies?**
    5. **Is there a significant difference between a person's salary before and after the COVID-19 pandemic?**

2. The plan for data collection

We are utilizing an existing dataset on Kaggle, which was collected in the form of a survey containing 34 questions. From this dataset, we will extract several comprehensive and relevant fields, such as salary, educational background, and programming languages, to aid in our analysis. However, since some people may not want to share their personal information, such as salary, some fields may be incomplete or inconsistent, limiting their usability in our study. Additionally, there is a gender imbalance among survey participants, with 3,212 females and 16,138 males. This imbalance may introduce biases into the analysis results, which will need to be considered in our interpretation and conclusions.

- **link for dataset:**
  https://www.kaggle.com/code/kailex/education-languages-and-salary/input

- **contents :**

  Duration (integer):
  The time taken by the user to complete the survey, measured in seconds. This can be used to filter out responses from users who completed the survey too quickly.

  highest level of formal education(string):
  The highest level of formal education that the user has attained.

  country reside(string):
  The country where the user was living at the time they completed the survey.

  yearly compensation(integer):
  The user's annual income, measured in USD.

  regularly used programming language(integer):
  In the original data, there are 13 columns representing 10 major languages used in data science and “None”, “Other”, “Other(number)”, where the users will fill in -1 or the number of other languages they use in the “Other(number)” column. We sum them up and use an integer to represent the number of languages a user uses.

3. The data analysis methods that you want to apply

- Plan to perform descriptive analysis:

Given the data, we first plot the data(e.g. a bar plot, with y-axis representing the salary) to provide a preliminary visualization of the distribution we aim to explore further. Also, we will calculate the descriptive statistics including mean, median, and variance. This will help in understanding central tendencies and variability in the data, providing statistics for further analyses, such as z-test or t-test.

-Hypothesis Formation

Based on the questions we proposed the following hypothesis will be test:

1. **Relation between salary and education level**

   (Pooled t-test)

   - **H0 : Obtaining a higher level of education, such as master's degree, is not essential for data scientists/software engineers to secure a higher salary**
   - **H1 : Obtaining a higher level of education, such as master's degree, is essential for data scientists/software engineers to secure a higher salary**

2. **Relation between number of familiar programming language and salary**

   (Welch's t-test)

   - **H0 : Learning more programming languages cannot lead to higher salary for data scientists/software engineers.**
   - **H1 : Learning more programming languages can lead to higher salary for data scientists/software engineers.**

3. **Relation between a person's salary before and after the COVID-19 pandemic**

   (Paired t-test)

   - **H0 : There is no significant difference between a person's salary before and after the COVID-19 pandemic.**
   - **H1 : There is a significant difference between a person's salary before and after the COVID-19 pandemic.**

4. Expected results and conclusions

With the help of statistical methods, we aim to understand the interrelationships between salary, highest educational attainment, and programming language used. We expect to identify patterns and correlations that reveal how these factors influence each other. For instance, we anticipate analyzing whether higher educational qualifications lead to higher salaries or if proficiency in certain programming languages correlates with better compensation. This analysis will provide insights into the factors contributing to career success in the tech industry, enabling data-driven conclusions about the impact of education and technical skills on income levels.