**Data Science Bootcamp**

**Project 2**

**Group members: Rich Kirschenheiter, Arielle Eagan, and Timothy Kelly**

***“Current Data Science Salaries in the United States”***

***Proposal***

*Chosen topic and rationale:*

As future Data Science and STEM professionals, we are curious about the job market and earning potential in new roles available after we finish this Bootcamp.

Through Kaggle, we found a dataset of US-based Data Science and STEM Salaries available through Kaggle, which provides 62,000 salary records from top companies throughout the country. This dataset, which was provided in csv format, contains information such as company name, job title, location (city and state), compensation (total yearly compensation as well as base salary, bonus, stock grants), and demographics including education level, gender, and race. Looking at the dataset’s source, we found that the data was scraped from levels.fyi. Using this source, we found an API in JSON format that allowed us to pull more updated data, running only about 6 days behind current present day.

We plan to create a thick server, which means that most of our time will be dedicated to exploratory data analysis (EDA) and data cleaning.

*Planned steps:*

* We will first do exploratory data analysis (EDA)
* We will then clean our original dataset in python
  + We will filter the data to create a data subset that includes:
    - US-based positions (remove international)
    - Individual contributors (remove managers)
    - Subset of positions (ex: Data Scientists)
    - Create position/salary levels
      * Entry/associate
      * ?
      * ?
      * Senior
    - Possibly create geographic regions (northeast, Midwest, west coast)
    - Possibly group by companies and compare across companies (Google, Apple, etc.)
* We will export our cleaned subset dataset to our website using mongo db.
* We will create 1 main visualization with drop down options for the position/salary levels.

*Data and Metadata:*

* Data Science and STEM Salaries data set:
  + Original csv link via Kaggle: <https://www.kaggle.com/jackogozaly/data-science-and-stem-salaries>
  + Metadata:

Graphical user interface, text, application, email

Description automatically generated

* + Data source: <https://www.levels.fyi/?compare=Google,Facebook,Microsoft&track=Software%20Engineer>
  + API in JSON format: <https://www.levels.fyi/js/salaryData.json>

*Screenshots of relevant inspiring visualizations:*

Map

Description automatically generated

<https://www.washingtonpost.com/technology/2019/07/23/ever-wonder-how-much-tech-workers-get-paid-your-town-this-map-might-have-answer/>

*Sketch of final design:*

For our visualization, we plan to create an interactive map of the US like the one above.

Display: individual contributor role in data science

* + Drop down options: 5 options
    - Drop down
      * Across all levels
      * Level 1
      * Level 2
      * Level 3
      * Level 4

Planned text on site:

* + Insight that we found looking at data
  + How we decided to group levels
  + How people should interpret what they are looking at

*Link to Github:*

<https://github.com/ArielleEagan/Project-2.git>