**Data 607 Spring 2023 - Project 3 Proposal**

Team Members: Taha Ahmad, Farhana Akther, Ross Boehme, Nick Climaco, Alice Ding

**Project Goal**

* Use data to answer the question, “Which are the most valued data science skills?”

**Collaboration Tools**

* We will store our data on a Microsoft Azure database in the cloud. We’ll access that data from R notebooks using a common password.
* We will analyze our data using R, backing up and sharing our notebooks as needed on Github. Using commits and pulls, we can keep each other updated on changes.
* Our primary communication methods will be Slack messaging and Zoom meetings. We also have a shared-access Google Doc where we collaborate on a checklist of tasks, outline our approach, and write out findings.

**Data Acquisition**

* Our project focuses on common data science skills requested in job postings from the NYC government.
* The City of New York's job postings are compiled into a CSV document and uploaded periodically to the [NYC Open Data website](https://data.cityofnewyork.us/City-Government/NYC-Jobs/kpav-sd4t).
* Depending on how our skills develop during our Week 7 “structured data” curriculum, we will connect that data to our Azure database either via API or CSV upload.

**Logical Model for Database**

* Based on fields available in our dataset provided above, we will create a relational database by grouping similar data fields into separate tables. The tables will be linked by keys.

Diagram

Description automatically generated

**Full Project Proposal**

The project goal’s question of "What are the most valued data science skills?" is a different question than "What skills are most valued for a data scientist?" Answering the former means looking at everything a data scientist does, and determining which is the most in demand of their skills.

Data science is an interdisciplinary field. It entails not just analyzing data but a multi-step workflow. While there are different descriptions of the workflow, a general outline is:

1. Acquire (collect and read data)
2. Prepare (clean and transform data)
3. Explore (identify labels and features)
4. Analyze (build, train, and validate models)
5. Visualize/Communicate (identify takeaways, explain insights, graph findings)

Therefore there are many people who aren't data scientists who leverage data science skills.

Our idea: Scan job postings on the City of New York official jobs site, identifying and quantifying requested skills in job descriptions which are part of the data science workflow. Our group will theorize about which skills in the data science workflow are "most valued" by determining how many jobs in the NYC government request those skills.

We can identify requests for those skills using regex to find keywords in job descriptions. For example, using the above data science workflow as a template:

1. Data Acquisition keywords: SQL, "data acquisition", databases, "cloud storage", APIs, “data entry”, etc.
2. Data Preparation keywords: Alteryx, Knime, Excel, "data engineering", "preprocessing", "data preparation", etc.
3. Data Exploration keywords: Tableau, Excel, "data analysis", "feature selection", "data engineering", “data exploration”, etc.
4. Data Analysis keywords:  Excel, "data modeling", "machine learning", “data analysis”, "artificial intelligence", etc.
5. Data Visualization/Communication: Tableau, PowerBI, "data insights", "quantitative findings", “graph creation” etc.

By understanding which types of jobs require data science skills, and what the most in-demand skills are, our group can determine which skills we should most focus on in our education, and what industries we could best help with our full stack data science knowledge.