Team 5 Project Overview
Project Name: Data Science Skills
Date: 3/10/2023

Table of Contents:

- Overview
- Team Members
- Collaboration Tools
- Data Collection
- Data Cleaning
- Data Storage
    - Database Design
    - Entity Relationship Diagram
- Data Ingestion
- Analysis

Overview:

For this project, we will be collecting job postings related to various Data Science openings across the US and using this information to create a database of skills desired by employers. This methodology is driven by the belief that the most important data science skills are those that will result in an individual being hired as such. As a result, we believe that understanding what employers are looking for when hiring for a Data Scientist is a strong proxy for identifying these important skills needed in Data Science.

Team Members:

- Genesis Middleton
- Joe Garcia
- Kory Martin
- Pei-Ming Chen

Collaboration Tools:

	Tool(s)	Description
Communication Tools	Zoom, Slack, Email	- For Slack we created a workspace that allows us to communicate and sync up in real-time and to have a place for dropping updates - Zoom is our primary channel for meeting up in a virtual face to face manner - Email is a way that we are able to communicate to discuss things that are bit more long-form and are not time-sensitive
Project Documentation	Google Drive, Google Docs, Google Sheets, Google Slides, Google Draw	- We decided to use Google Drive and its various productivity tools to develop documents related to the various aspects of the project. Additionally, we are able to upload different files and documents to Google Drive that we are sharing across the team

		- Google Sheets was used to develop a master project tracking document as well as a staging area for our initial data collection - Google Docs was used to share our project documentation - Google Slides was used to collaborate on the presentation deck that we will be using for our project report-out - Google Draw was used to develop our Entity Relationship Model
Code Collaboration	GitHub, AWS Database, Google Drive	- GitHub is being used to collaborate on our code - Google Drive is also used to store the raw code files - AWS Database was where we stored our data in the cloud

Data Collection:
For this project, we have decided that we will begin by collecting our raw data from the following job boards:

- LinkedIn
- Glassdoor
- Indeed
- Simply Hired
- Monster

Initially, our team members will be responsible for visiting the job boards listed above and copying the data for 25 job postings each (for a total of 125 postings) related to Data Science positions. We will be collecting the posting for positions at different seniority levels. For these job postings we will be collecting the following information into a shared Google Sheet:

- Job Board
- Post URL
- Company
- Location
- Position/Title
- Industry
- Skills
- Seniority Level
- Job Type (i.e. Remote, Hybrid, On-Location)
- Minimum Years Experience

Data Cleaning:
Once the data has been stored in our shared Google Sheet, we will import the data into R to perform the necessary cleaning and preprocessing so that we can import the data into our database. One of the biggest challenges we anticipate is breaking down the raw skills data and converting it into more discreet skills listings.

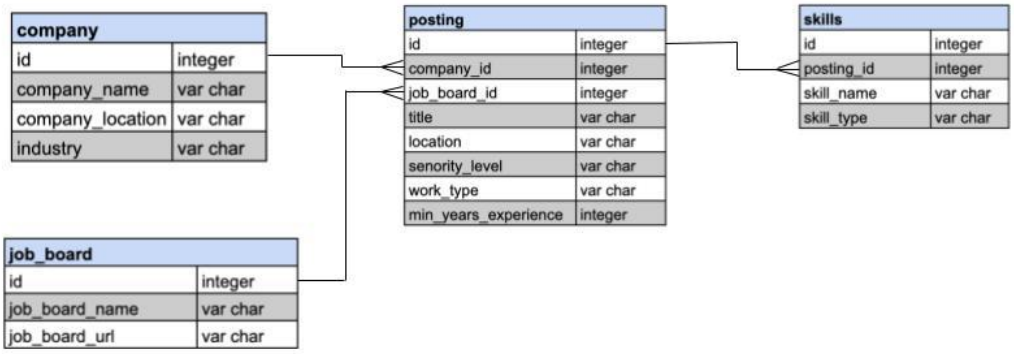
Data Storage:

- Our plan is to store the data in a PostgresSQL database hosted on AWS
-

Database Design:
For the database we've decided on the following database tables:

Table Name	Description
job_board	This is the table of the job boards that we are using to collect our job postings
company	This is a table of the various companies that are hiring for the jobs that we collect in the job boards
posting	This is a data table of information related to the actual job and position that we collected
skills	This will be a table of all the skills that we collect from across the various job postings

Entity Relationship Diagram:



Data Ingestion:
Once we've cleaned the raw data in R, we will create data frames based on the data tables listed above. We will then create a connection to our AWS hosted database and ingest the data into the appropriate tables. We chose to do it this way so that we are only ingesting cleaned and structured data into our structured database.

Analysis:

Finally, using our cleaned data, we will be able to analyze the data in R to answer the question of “What are the most valued data science skills”.

Some of the questions we will attempt to answer through the data are:

- What skills are the most commonly listed across all jobs and positions?
- What skills are the most commonly listed across jobs based on seniority level?
- What are the top skills across different skill types (e.g. Hard Skills, Soft Skills)?