First Part of Project 3

Instructions

Create a short document, with the names of group members. You should briefly describe your collaboration tool(s) you'll use as a group, including for communication, code sharing, and project documentation. You should have identified your data sources, where the data can be found, and how to load it. And you should have created at least a logical model for your normalized database and produced an Entity-Relationship (ER) diagram documenting your database design.

Team Members

- Jerald Melukkaran
- Md. Tanzil Ehsan
- Lawrence Yu
- Isaias Soto
- Woodelyne Durosier
- Sheriann McLarty
- Zahid Chowdhury

Effective Collaboration Tools

Is essential in the field of data science projects. For seamless communication, version control and documentation we will adapt the following tools:

Communication

- Slack & Zoom

  These platforms will allow real time discussion, updated and coordination between the team. This approach will provide structured conversation, document sharing as well as improving teamwork collaboration.

Code Sharing

- GitHub
  GitHub will provide the backbone needed for version control, coding association, and maintained of repository. According to Chacon and Straub (2019), the team uses GitHub to guarantee code integrity and traceability, which promotes an environment of organized software development.
- Slack for quick snippets

Project Documentation

- Draw.io – ER diagram
- RStudio – data processing
  - Include sql dummy password for sharing
- Python – Retrieving data
- MySQL – data storage and encryption
- Google Docs
  Documenting work done such as for Project 3 Part 1

Data Sources and Management

- **Job Postings:**

  Screen through platform such as LinkedIn, Glassdoor, and Indeed. This insight will highlight employers' expectation and skills demand. Furthermore, these platforms will expose the team to vast API interactions which will provide a boost to extract structured data featuring job descriptions and needed skills sets for the data field. There are a few ways to get the data and it may depend on the specific job board. Web scraping is one possibility. Another is checking through sites to see if they have downloadable data sets or APIs.

- **Online Courses and Certifications:**
  Platforms like Coursera, edX, and Udacity analyze course enrollments, syllabi, and certification trends to provide information on in-demand skills. To identify the most taught and sought-after skills, web scraping and API-based extraction approaches will be used as needed (Patterson, 2022).

- **Kaggle – Surveys**

  We'll be using a dataset of surveys hosted on Kaggle.com

- **JobSpy**

  Using this open source project to scrape data from glassdoor, indeed, etc. to get text descriptions of jobs into a dataframe for a more digestible format. Searching for specific words, finding matches to see frequencies of skill mentions. While this is done in Python, dataset analysis will be done in R as required.

Data Storing and Loading

- **Data Processing Tools**
  Python, R are ideal to extract, transform load operations. These tools allow for more efficient data processing, ensuring consistency and scalability when dealing with large datasets. We'll also be using pycharm to get data into MySQL for team use.

- **Data Cleaning and Transformation**
  To increase data quality, preprocessing actions such as resolving missing values, eliminating duplicates, and normalizing features will be undertaken. This ensures that the dataset is prepared for analytic investigation.

- **Development and Analytical Tools**
  - **RStudio**
  - **RMarkdown(Rmd)File**
- **MySQL**
  The data will be stored using MySQL for data efficiency and integrity. From here we will also be loading our data for tidying and analysis in R.

Reference:

Chacon, S., & Straub, B. (2019). *Pro Git*. Apress.

Kaggle. (2023). *Kaggle developer survey 2023 results*. Retrieved from https://www.kaggle.com/surveys

Patterson, J. (2022). *The state of online education and professional certifications in data science*. Coursera Research.

Williams, K., & Cowley, P. (2021). *Virtual teamwork in data-driven projects: Best practices and challenges*. Oxford University Press.