Introduction

This document outlines our group's collaborative efforts in developing the “Which are the most valued data science skills?” project. It includes the names of group members, the tools we will use for communication, code sharing, project documentation, and details about our data sources and database design.

Group Members

- Member 1: Inna Yedzinovich
- Member 2: Zaneta Paulusova
- Member 3: Md Asaduzzaman
- Member 4: Md. Asadul
- Member 5: Md. Simon Chowdhury

Collaboration Tools

1. Communication:
    - Tool: Whatsapp, Microsoft Outlook.
    - Purpose: Real-time messaging, video calls, and file sharing.
2. Code Sharing:
    - Tool: GitHub
    - Purpose: Version control, code collaboration, and repository management.
3. Project Documentation:
    - Tool: Google Docs
    - Purpose: Documenting project plans and meeting notes.

Data Sources

1. Source 1:
    - Description: We will conduct an in-team survey using Google Forms to name the 5 most useful data science skills based on the opinions of our teammates.
    - Location: The survey data will be collected and stored in a Google Sheets document.
    - Loading Method: We will export the survey results as a CSV file, which will be used for further analysis.
2. Source 2: [Data Science Skills Survey 2022 - By AIM and Great Learning (analyticsindiamag.com)]
    - Description: We will compare our survey results with the ‘Data Science Skill Survey 2022’ article. It does not provide the raw data of their survey so we will need to create a table from the provided information to compare or import the graph/table images from the article. To show what results of Data Science skills on the Survey 2022 with over survey.

- Location: Create a CSV file with the ending results data as raw data is not provided or do a website scraping in Rstudios
- Loading Method: Website scraping in Rstudios or exporting a CSV file that was manually created.

Database Design

Google Form questionnaire:

1. Logical Model:
    - Description: The logical model represents the structure of a database designed to store responses from a Google Form survey about data science and software engineering experience, programming languages, learning resources, areas of interest, soft skills, and valuable data science skills. The model ensures data is organized efficiently, reducing redundancy and maintaining integrity.
    - Key Entities:
        i. **Respondents**
        ii. **Experience**
        iii. **ProgrammingLanguages**
        iv. **LearningResources**
        v. **InterestAreas**
        vi. **SoftSkills**
        vii. **ValuableSkills**
2. Suggested Entity-Relationship (ER) Diagram (Tentative):
    - Description:
        i. Respondents Table
            1. **Purpose:** Stores basic information about each respondent.
            2. **Key Attributes:**
                a. respondent_id (Primary Key): Unique identifier for each respondent.
                b. first_name: The first name or nickname of the respondent.
                c. age: The age of the respondent.
        ii. Experience Table
            1. **Purpose:** Captures the respondent's experience in data science and software engineering.
            2. **Key Attributes:**
                a. experience_id (Primary Key): Unique identifier for each experience record.
                b. respondent_id (Foreign Key): Links to the respondent_id in the Respondents table.
                c. data_science_exp: Indicates if the respondent has data science experience (Boolean).

d. software_eng_exp: Indicates if the respondent has software engineering experience (Boolean).

iii. ProgrammingLanguages Table
    1. **Purpose:** Lists the programming languages used by each respondent.
    2. **Key Attributes:**
        a. language_id (Primary Key): Unique identifier for each programming language record.
        b. respondent_id (Foreign Key): Links to the respondent_id in the Respondents table.
        c. language: The programming language used by the respondent.

iv. LearningResources Table
    1. **Purpose:** Details the resources respondents use to learn new data science skills.
    2. **Key Attributes:**
        a. resource_id (Primary Key): Unique identifier for each learning resource record.
        b. respondent_id (Foreign Key): Links to the respondent_id in the Respondents table.
        c. resource: The learning resource used by the respondent.

v. InterestAreas Table
    1. **Purpose:** Specify the areas of data science respondents are interested in.
    2. **Key Attributes:**
        a. interest_id (Primary Key): Unique identifier for each interest area record.
        b. respondent_id (Foreign Key): Links to the respondent_id in the Respondents table.
        c. interest_area: The area of data science the respondent is interested in.

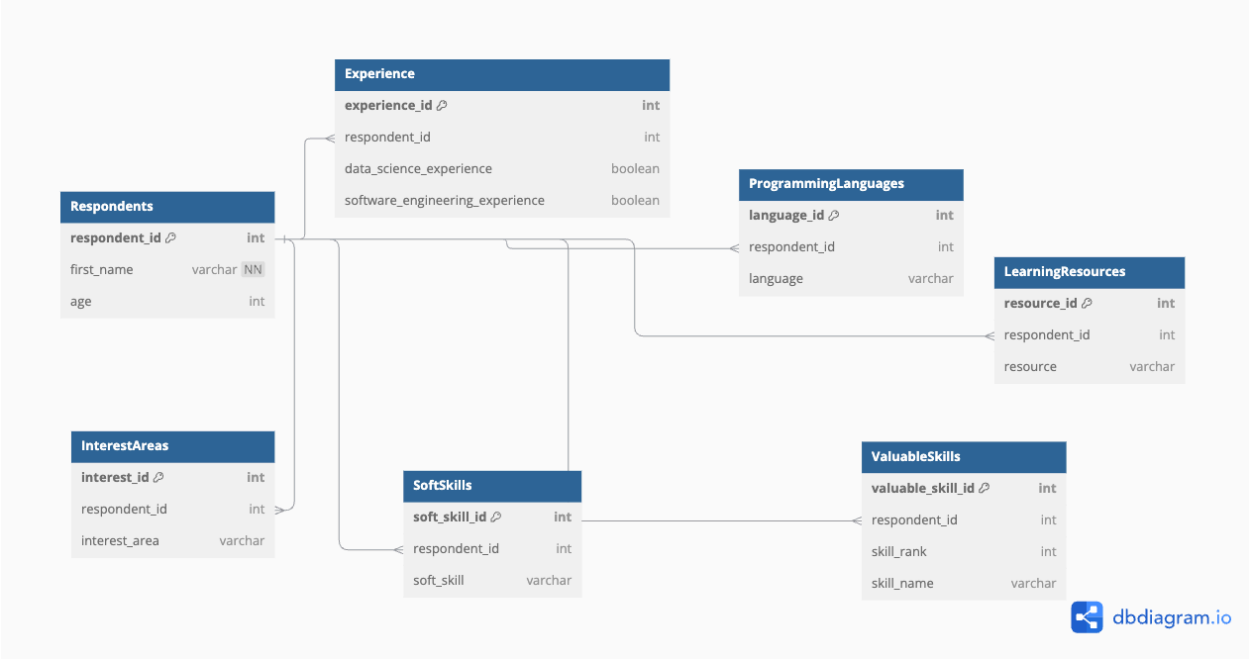
vi. SoftSkills Table
    1. **Purpose:** Identifies the soft skills respondents consider important for a data scientist.
    2. **Key Attributes:**
        a. soft_skill_id (Primary Key): Unique identifier for each soft skill record.
        b. respondent_id (Foreign Key): Links to the respondent_id in the Respondents table.

c. soft_skill: The soft skill considered important by the respondent.

vii. ValuableSkills Table

1. **Purpose:** Ranks the most valuable data science skills according to respondents.
2. **Key Attributes:**
   a. valuable_skill_id (Primary Key): Unique identifier for each valuable skill record.
   b. respondent_id (Foreign Key): Links to the respondent_id in the Respondents table.
   c. skill_rank: The rank of the skill (e.g., 1 for most valuable, 2 for second most valuable, etc.).
   d. skill_name: The name of the valuable skill.

- Suggested Diagram:



Conclusion

In conclusion, this document provides a comprehensive overview of our group's collaborative approach, the tools we will use, and the foundational elements of our project. By clearly defining our data sources and database design, we aim to ensure a smooth and efficient workflow throughout the "Which are the most valued data science skills?" project.