Team 2 - Project 3: Part 1

Team Members - Eric Lehmphul, William Aiken, Bikram Barua,
Nnaemeka Newman Okereafor, Catherine Cho

10/10/2021

Team Collaboration

Our team decided to use Slack as the main method of written communication, Zoom as a means to carry out team meetups, and Asana as a way to assign tasks to individuals and set deadlines.

Code and documents will be shared through two github repositories:

- https://github.com/baruab/msdsrepo/tree/main/Project_3_607
- https://github.com/baruab/Team2_Project_3_607

We may shift to a single github repository, but for the moment collaboration is occurring through both github repositories. Currently, https://github.com/baruab/msdsrepo/tree/main/Project_3_607 holds the data sources for the project and https://github.com/baruab/Team2_Project_3_607 is being used for team collaboration.

Timeline (10/3 - 10/10)

- *10/4* - Team communication initiated on Slack.
- *10/5* - Initial Zoom meetup to discuss how the team would collaborate. Created tasks in Asana to visualize what the team needs to accomplish. Created Github repository to allow for the team to share all necessary code and documents.
- *10/6* - Zoom meetup to determine the dataset we plan to use for Project 3
- *10/7* - Zoom meetup to assign tasks that were created in earlier meetups. Created the Entity-Relationship diagram for the chosen data.
- *10/8* - 2 Zoom meetups: one to discuss the plan for analysis with the data and another to review Entity-Relationship diagram and database design.
- *10/10* - Finalize and submit write up for Part 1 of Project 3.

Data Source

Our data source is from the '*2018 Kaggle Machine Learning & Data Science Survey*' which we retrieved from Kaggle.com (https://www.kaggle.com/kaggle/kaggle-survey-2018). This data source contains three separate datasets:

- SurveySchema.csv - Provides the questions that were asked to all participants within the survey.
- freeFormResponses.csv - Contains responses to questions that required a written answer.
- multipleChoiceResponses.csv - Contains responses to questions that required a multiple choice answer.

After some time exploring the data, the team decided to focus our analysis on the multipleChoiceResponses.csv file.

Loading the Data

We stored the data into the Github repository to allow us to easily access the data in r from any machine. To load the data run the following lines of code:

```
## survey schema

survey.schema <- read.csv("https://raw.githubusercontent.com/baruab/msdsrepo/main/Project_3_607/kaggle-

## freeform responses

free.form <- read.csv("https://raw.githubusercontent.com/baruab/msdsrepo/main/Project_3_607/kaggle-surv


## multiple choice

multiple.choice <- read.csv("https://raw.githubusercontent.com/baruab/msdsrepo/main/Project_3_607/kaggl
```

Entity Relationship Diagram

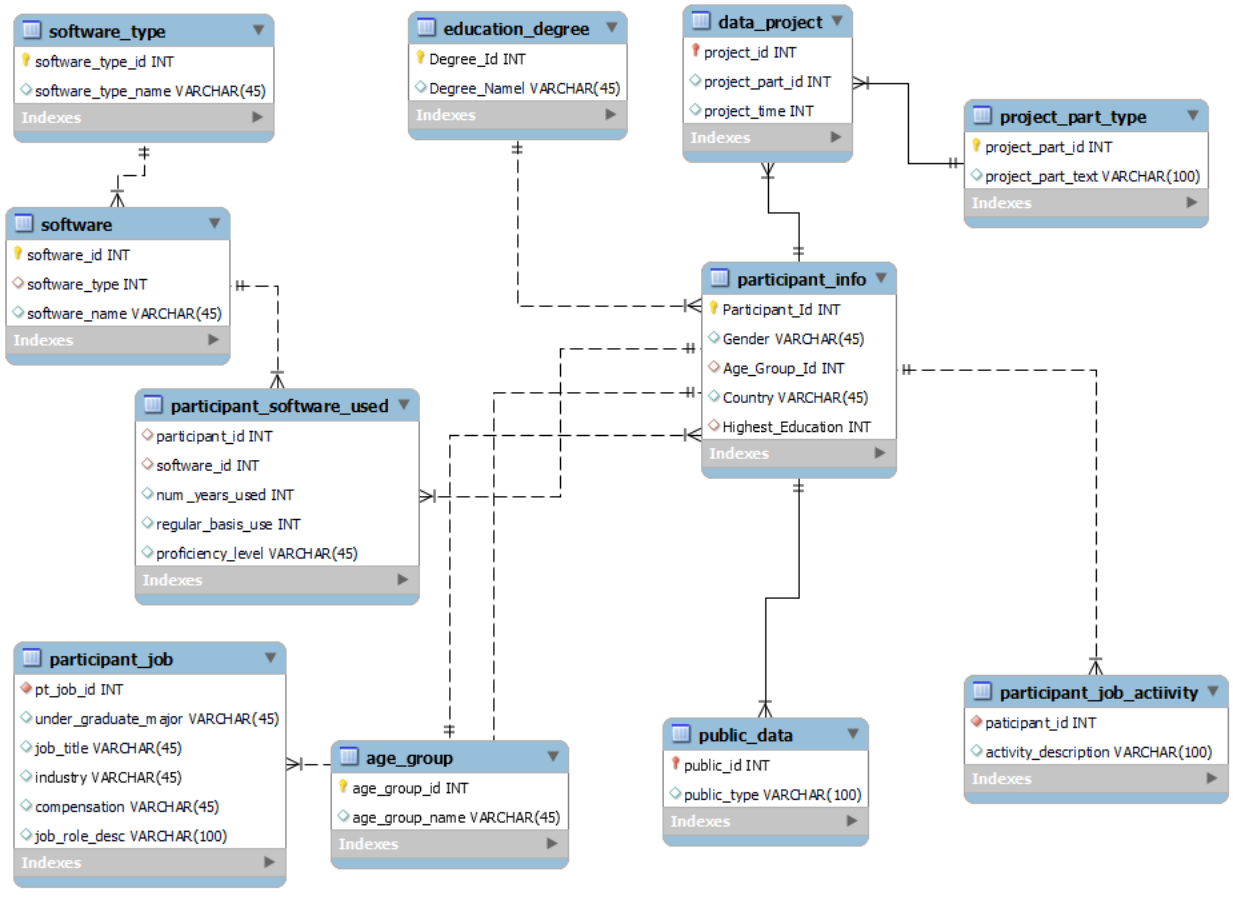

Figure 1: Project 3: Entity Relationship Diagram