DATA 607 Project 3 - pt 1

Team Silver Fox

Team Members

- Dan Rosenfeld
- Magnus Skonberg
- Mustafa Telab
- Josef Waples

Collaboration Tools

- **Slack:** regular, written communication.
- **Google Meet:** 1-2x/wk meetups.
- **Github:** code sharing and collaboration.
- **Google Docs:** collaborative written project documentation.

Data Source(s)

We identified our sources of data as the following (cited APA-style below):

1. Kaggle. (2018). **2018 Kaggle ML & DS Survey** [Data file]. Retrieved from https://www.kaggle.com/kaggle/kaggle-survey-2018?select=multipleChoiceResponses.csv
2. Jeff Hale. (2018). **The Most in Demand Skills for Data Scientists** [Data file]. Retrieved from https://www.kaggle.com/discdiver/the-most-in-demand-skills-for-data-scientists/log?select=ds_job_listing_software.csv

For more background on Data Source #1:

- The **2018 Kaggle ML & DS Survey** dataset had (3) .csv's: the Survey Schema, Free Form Responses, and Multiple Choice Responses.
- *The Survey Schema* contained questions and explanations of response exemptions. While we do not plan on analyzing this .csv, it was useful pre-screening questions for pertinence.
- *The Free Form Responses* contained text answers submitted by survey takers when their response was not covered by multiple choice options. If we deem this set valuable / unique for purposes of analysis, we will draw on it. Otherwise, it may be disregarded as well.
- *The Multiple Choice Responses* contain categorical variables submitted by respondents. This is where our "gold" lies and likely where the brunt of our data tidying, transforming, and analysis efforts will go.

Data Loading

Our data will be loaded in the following manner:

1. **Pre-screen:** after viewing the initial Kaggle **2018 Kaggle ML & DS Survey** multiple choice response dataset and seeing that we would not be able to access this dataset via Github due to its size (39+ MB), we deemed "pre-screening" essential. We reviewed all 50 questions from the SurveySchema.csv file and kept only those applicable to the question at hand. Once this dataset had been "pre-screened" it was uploaded to Github along with Jeff Hale's general and software skill csv files (for later comparison).
2. **Acquire data via reading .csv's:** get the URL, read each .csv from Github (in its raw form), and put the data into tabular form before moving on to tidying and transforming.

Entity-Relationship Diagram

To document the design of our database and provide a logical model for its normalization, we created two separate ERD diagrams. To access these diagrams, **please click-through the ERDs listed below:**

- 2018 Kaggle ML & DS Survey - ERD
- Most in Demand Skills - ERD

Explanation of Approach

The Kaggle ML & DS Survey dataset is messier and will require more tidying and transforming. Our plan is to build a normalized table of questions and answers, and then "link" tables for questions and answers that relate to areas of interest (ie. general skills v software skills v "value filter").

After gaining insight into software v general skills within this dataset, we can then utilize a "value filter" (via income, education, or experience level) to compare highest ranking skills for the general survey set against those filtered for higher income, experience, education or some combination of these variables.

At this point, we then transition to comparing data science skills deemed as valuable in Kaggle's 2018 survey with the most in-demand skills for data scientists across multiple job sites. This "pro level" dataset, prepared by Jeff Hale, was pulled in to paint the picture of contrast (if there was any) between the skills data scientists deem as most valuable (ie. those doing the work) vs. those that employers hire for (ie. hiring managers).

Through the transformation, analysis, and "value filtering" of our Kaggle ML & DS Survey dataset through its comparison to / pulling from a "pro level" dataset, our analyses and visualizations should provide multiple siftings and thus more valuable of insights regarding the question at hand:

> What are the most valued data science skills?