**CUNY 607 - Project 3 Design**

**Purpose**: To determine the most valued data science skills

**Group Assigned-A**:

* Avery Davidowitz
* Jennifer Abinette

*\*Originally assigned 3 other members, but were either already in another group or elected to drop this project*

**Approach**: We will answer what data science skills are most valued by analyzing Job Listings for Data Scientist positions in the United States (U.S.). We created a list of 53 skills based on a Google Search of the most valued data scientist skills that will be searched for in each job description. The skills most frequently referenced are the ones that will be considered the most valued for a Data Scientist. We will further analyze how many skills are included per job listing and the affects (if any) of job location population. We want to explore if job listings for larger cities reference more skills than smaller cities.

**Project Documentation**:

* Stored in Github <https://github.com/JAbinette/CUNY-607-Project-3-Data-Science-Skills>
* Communication via email and Slack

**Data Sources:**

* U.S. Data Scientist Job Listings – data set retrieved from Kaggle.com <https://www.kaggle.com/datasets/sl6149/data-scientist-job-market-in-the-us?resource=download&select=alldata.csv>
  + 6,953 data scientist job listings scraped from Indeed.com in 2018 and posted for use on Kaggle.com. We filtered out positions that did not include ‘data’ in the position name leaving 2,079 before importing into R.
* U.S. Population by City – data set retrieved from census.gov

<https://www.census.gov/data/tables/time-series/demo/popest/2020s-total-cities-and-towns.html>

* U.S. State Abbreviations – data set retrieved from usps.gov <https://about.usps.com/who/profile/history/state-abbreviations.htm>
* List of Data Science Skills – data set created based on 1st page of Google Search results of Data Scientist Skills

<https://github.com/JAbinette/CUNY-607-Project-3-Data-Science-Skills/blob/main/Google%20Search%20-%20Data%20Scientist%20Skills.png>

**Methodology**:

* Identify datasets to use for project (Jen)
  + See list above
* Tidy/Transform as needed in R to create data frames for import into database (Jen)
  + All data imported into R is stored in Github Repository <https://github.com/JAbinette/CUNY-607-Project-3-Data-Science-Skills>
  + Job Listing sample was further adjusted in R to only include positions with Data Science or Data Scientist (1,444)
  + Population data included numerous aggregate rows (e.g., State total, County total) and rows created for census purposes. U.S. Census website included additional information for how to remove these instances by using following criteria:
    - PLACE variable cannot be zero determined from information posted on <https://www.census.gov/programs-surveys/geography/guidance/geo-identifiers.html>
    - Functional Status variable (FUNCSTAT) cannot be inactive per <https://www.census.gov/library/reference/code-lists/functional-status-codes.html#:~:text=The%20functional%20status%20(FUNCSTAT)%20code,each%20code%20is%20valid%20for.>
* Build database in Google Cloud and Import into normalized tables (Avery)
  + 4 tables: Location, Skills, Job Listings, & Job Listing Skills
* Conduct analyses in R including: (Avery)
  + Identifying the skills most frequently listed – Bar Chart
  + Identifying the number of skills listed per job description - Histogram
  + Effects of job location on skills listed - Scatterplot

**Findings:**

Our analysis indicates that the most frequently mentioned and therefore most valued data science skills include R, Python, Machine Learning, Statistics, Mathematics and Programming. On average, a job listing referenced around eight of the 53 data science skills we searched, but, overall, there did not seem to be a relationship between location population and the quantity of skills listed in the job listing.