Group 3: John McCarroll, Willow Rose, Killian Jakstis
CSCI 620 - Section 1
Professor Rajendra Raj
October 21st, 2024

Group Assignment 2

Section 1 - Summary:

In this assignment we were asked to take a small education dataset from the 2022 United States Census, store it in a disk-based database, and implement our previous analysis in a big data compatible way. In the initial phase of the project, group members individually explored approaches to storing the data. In our solution, we decided to use a relational database due to the fact that the original data was already structured into tables. Specifically, we used SQLite due to its ease of configuration and the availability of libraries in R and Python to easily connect to and query the database. Initially two approaches were taken, one in Python and one in R. Data was successfully loaded using scripts i and stored on the disk, however ultimately the R script was chosen due to its data cleaning capabilities.

R was used to clean the data before loading it into the database. Data was loaded into a dataframe from an excel file for cleaning. Since each row of the dataset containing real data began with a \t character, all lines not beginning with a \t were removed from the data set to remove extraneous information. Additionally, all columns other than the column containing group names were converted to numeric data. Finally, \t characters were removed from the group names and columns were renamed to remove spaces and be more legible. After cleaning, data was loaded into a SQLite database using the RSQlite library.

Visualizations were created in either Python or R based on each group member's experience. Figures 1 and 2 were created in Python using matplotlib, Figure 3 was created in R using ggplot2. Data was retrieved from the database in chunks using SQL statements before being used to create the visualizations. The visualizations are recreations of the those produced for assignment 1.

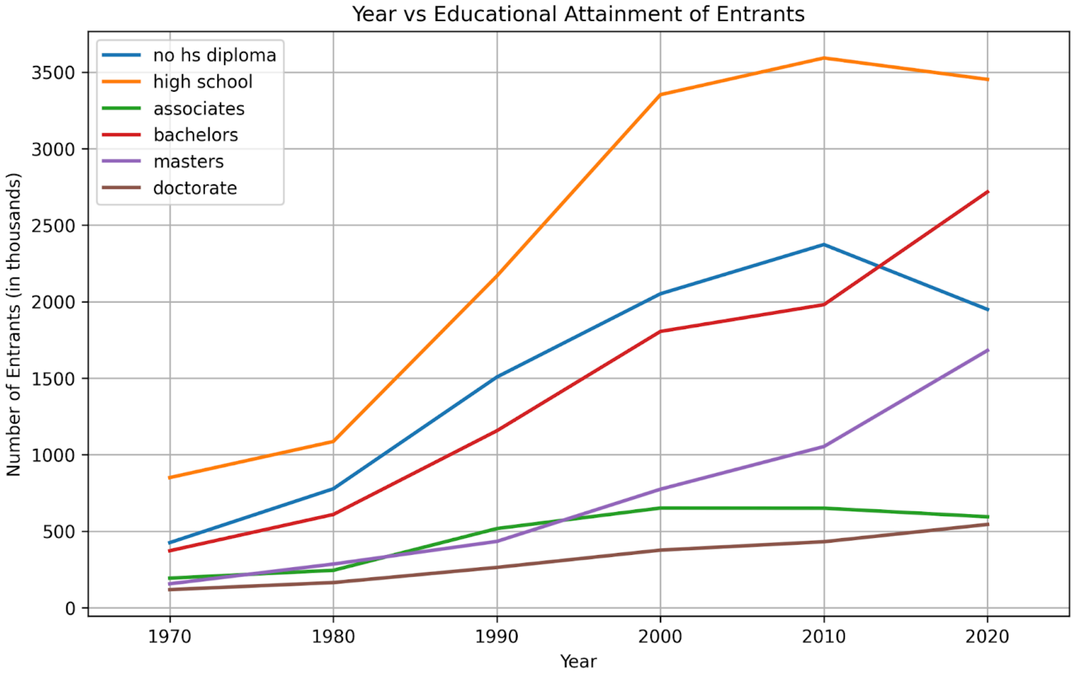

Fig 1. Graph of Educational Attainment vs Year of Entry

The first figure is a line graph of the educational attainment of immigrants by year of entry into the United States. It should be noted the “Number of Entrants” displayed on the Y axis is in the thousands (1 = 1000 entrants). There is a general trend of total entrants increasing each decade, since the 1970s. Interestingly, the number of entrants with college degrees (Bachelors, Masters, and Doctorates) increase relatively linearly by year of entry in the dataset. Most entrants attained a High School diploma, across all years of entry in the dataset.

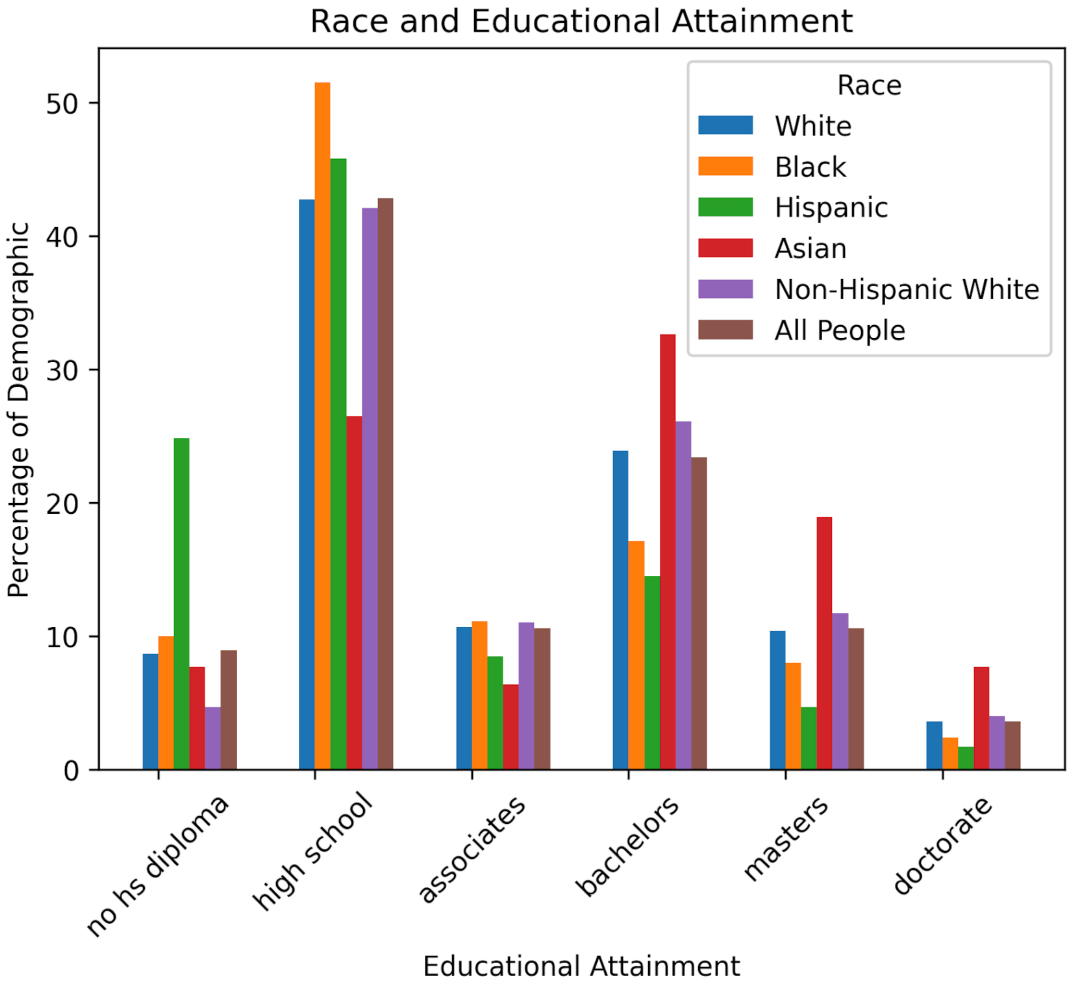

Fig 2. Bar Chart of Educational Attainment by Race

Figure 2 compares rates of educational attainment across several races. In higher education attainment, White, Non-Hispanic White, and Asian groups met or exceeded the general population rate for bachelor's, master's, and doctoral degrees. Notably, the Asian demographic achieved master's and doctoral level degrees at approximately twice the rate of the general population and had the highest rates of higher education attainment for bachelor's, master's and doctoral degrees. The Hispanic demographic stuck out for having substantially higher rates of individuals not having achieved a high school diploma (or equivalent) and of individuals who have not attained a higher education degree.

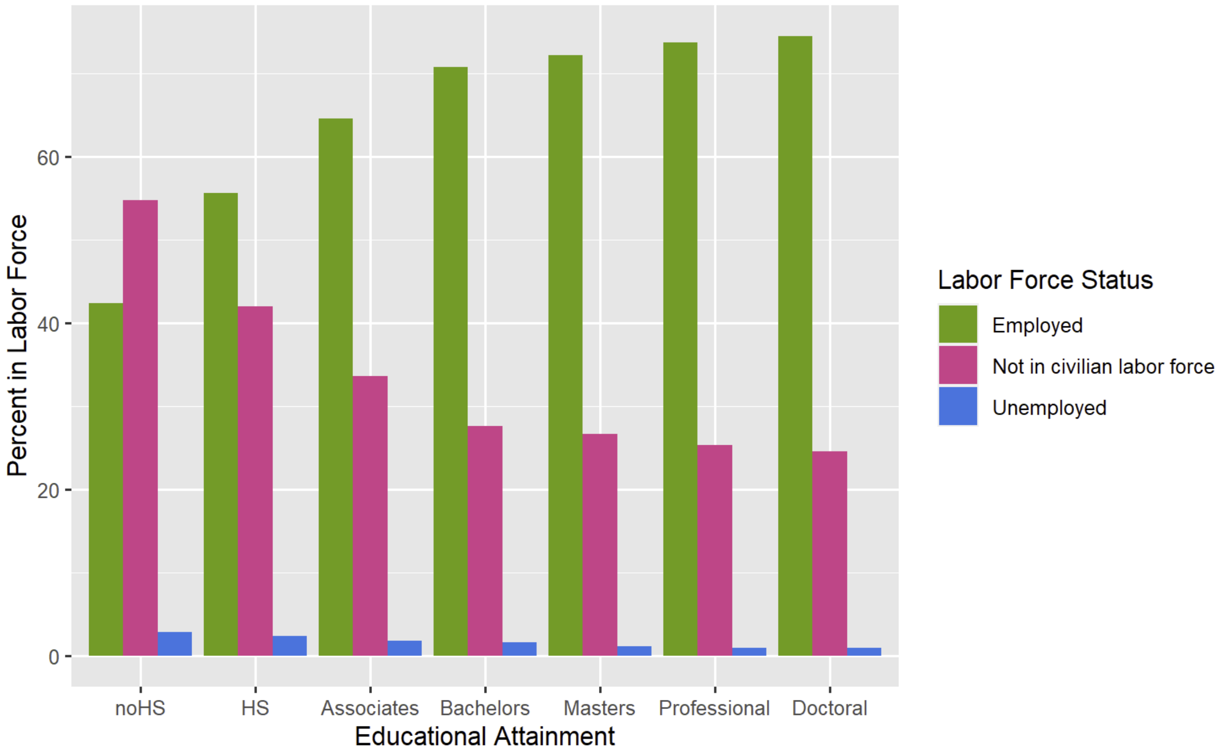

Fig 3. Percentage in Labor Force by Educational Attainment

The third figure is a bar chart of the Percentage in the Labor Force of each group of Educational Attainment. This chart visualizes the allocation of Labor Force Status across each Educational Attainment group. The “not in civilian labor force group” includes individuals who are unemployed but have not recently sought employment. There is a clear trend of higher rates of employment for groups of higher educational attainment. Doctoral graduates had the highest percentage of employment, at about 74% employed. Inversely, there was a trend such that the percentage of individuals who are unemployed or not in the civilian labor force decreases with greater educational attainment. Notably, the “No high school diploma” group was the only group for which there were proportionally more members who were not in the civilian labor force than there were employed members, and the group had the highest percent of unemployed individuals.

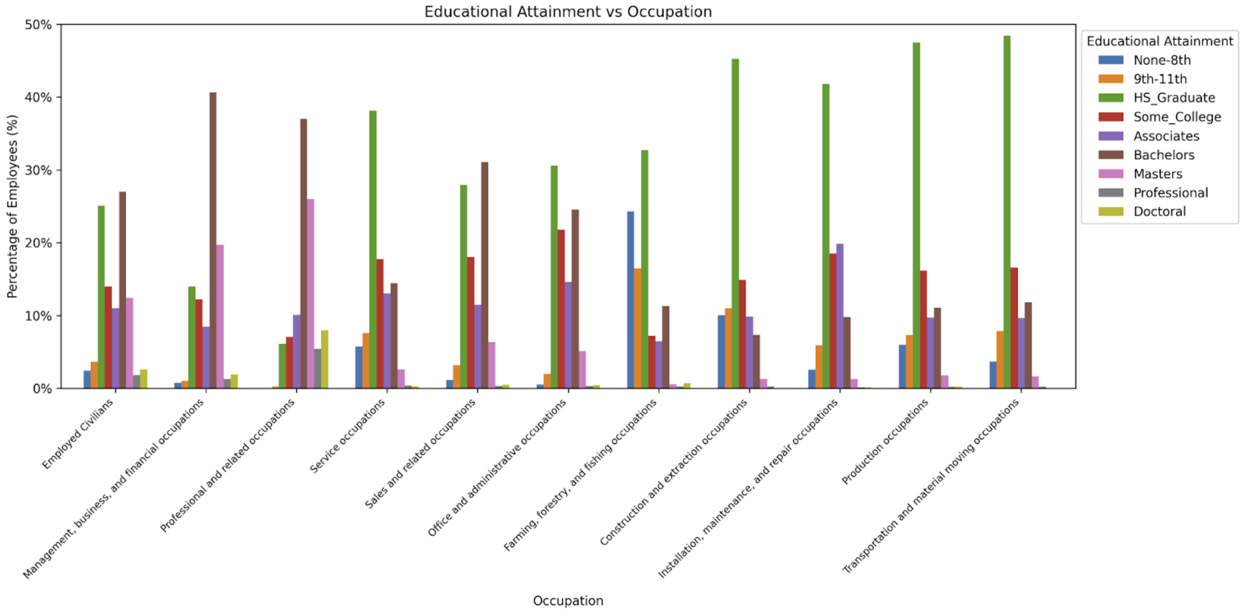

Fig 4. Educational Attainment vs Occupation

The fourth figure is a bar chart of Educational Attainment vs Type of Occupation. The Y axis displays the percentage of employees within that type of occupation. The first group on the X axis, “Employed Civilians”, contains the entire civilian labor force. All other occupation types are subsets of the Employed Civilians group. The most common level of educational attainment is a Bachelor’s Degree and the second most common is a High School diploma amongst Employed Civilians. The group with the higher proportion of Doctoral Degrees was the “Professional and related occupations”, raising questions about what kinds of job roles are included in that vague category. Interestingly, construction, maintenance, production, transportation, service, and sales occupations all had disproportionately high levels of High School graduates compared to the entire civilian labor force.

Section 2 - Lessons Learned

Through this assignment, we learned the importance of data engineering in the data analysis process as a whole. The original data format was structured very poorly for carrying out analysis using SQL given the numerous rows that had been added for readability of the excel file. As a result, we decided to parse the data upfront before loading it into our SQLite database in our final solution. Even though it took some initial overhead of restructuring the table schema, we realized that it would make carrying out our analysis much more straightforward.

We also learned the value of undergoing proper data cleaning before performing data analysis. The use of different data types present within the same columns would have made parsing through the data during analysis time-consuming and inefficient. To address this, we substitute values that did not align with the expected data type of a given column. For instance, percentages rounding to 0 were denoted 'Z' in the original data, so we opted to replace such instances with 0 to prevent type mismatches during our analysis.

In terms of data analysis, we explored the approach of processing large amounts of data in blocks, rather than retrieving all relevant records within a single query. We saw how this method of taking chunks of data took additional analysis and visualization code, seeing as we had to run several iterations of our query and build up visualizations iteratively. While the actual data volume was low, approaching analysis and visualization in this manner gave us a better understanding of how we could carry out analysis on large databases in the future.

Overall, we took away that data engineering is a necessary precursor for making data analysis a manageable and fruitful process, and that processing subsets of data individually can serve as a technique for analyzing datasets that are larger than main memory.

Section 3 - Individual Contributions:

Summary Table:

Group Member	Total Posts	Average Daily Posts	Max Daily Posts	Min Daily Posts	Total Files
jtm5356	48	2.285714286	17	0	1
ejb3831	32	1.523809524	9	0	6
kjj6427	33	1.571428571	11	0	1

Table 1. Discussion Post Summary by Contributor

Frequency Chart:

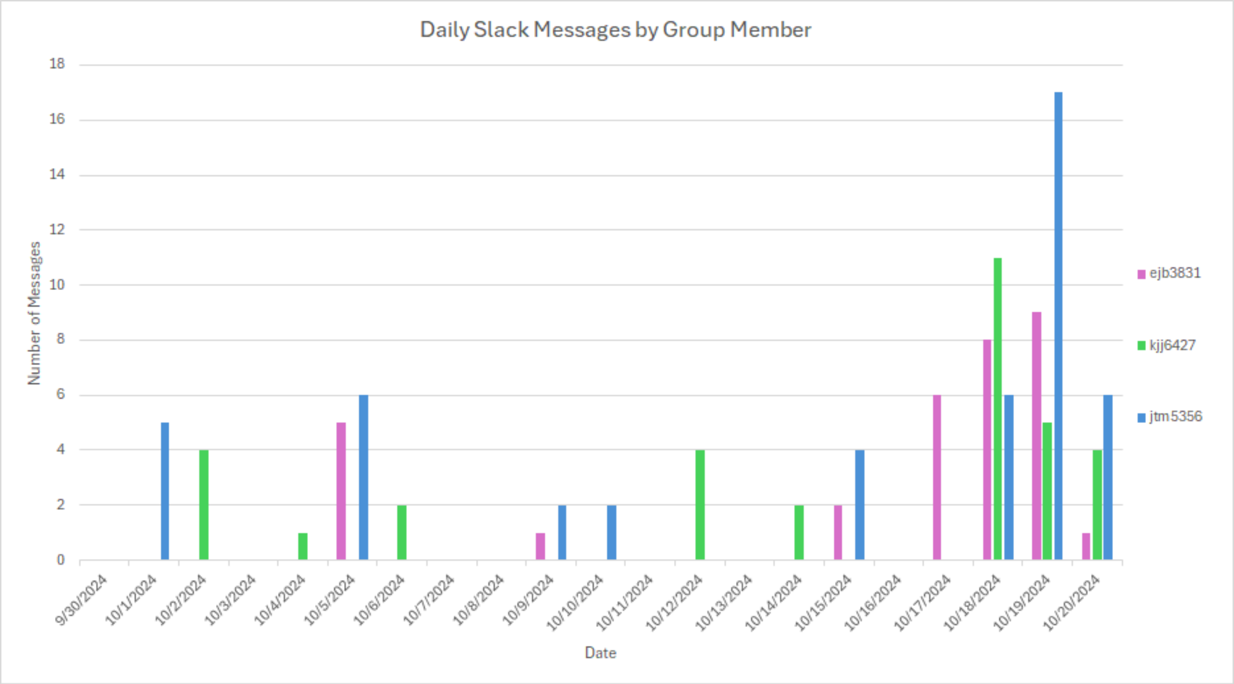

Fig 4. Number of Posts by Date

Coding contributions:

Willow Rose (ejb3831) coded the R scripts used to clean the raw data and load it into the SQLite database (GA2_Load_Data.Rmd) and to create Figure 3 (GA2_Load_Data.Rmd). John McCarroll (jtm5356) coded the initial python load_db.py script and coded the visualize_occupation.py script which generates Figure 4. Kilian Jakstis (kjj6427) wrote code in the initial python load_db.py script to clean table 2's raw data, as well as the year_of_entry_viz.py and race_attainment_visualization.py files that create Figures 1 and 2 respectively.