The purpose of this investigation was to give an insight to high school students about future careers. Our three graphs show information provided by indeed.com and outside sources such as the Bureau of Labor Statistics that demonstrates where in the United States are jobs, whether having a high school and college education played a role in qualifying for certain jobs, and which industries have high valued companies. This may influence their future choices concerning pursuing higher education, where they go to college and what they major in.

When introducing the idea about finding a job to high schoolers, our group tried to define what makes a good company. Some companies had a score of 0, so we developed a **new scoring system (`new_score`)**. For nonzero `avgoverallcompanyrating`, we kept such score in `new_score`; otherwise we used our prediction model to calculate the value of the company (`new_score`) and filled in the blanks. Such prediction model is fed by nonzero `new_score`; it contains four major variables: salary, company size, number of applicants, and popularity (number of clicks of that company's web page).

However, when testing the model, we found out that the correlation between `clicks` and `applies` was one—which, in real world data, is unlikely. For fitting our multiple regression model we have to take one of them away. Thus, in our final model, we have the following three variables: **`employeecount`, `applies`, and `Salary`.**

We then further cleaned our date. Since the database doesn't include salary, we searched on "Bureau of Labor and Statistics" and added it into our table. By generalizing the 59 types of `normtitlecategory` into 11 larger categories we got the median wage for each category. With this method, we gave each individual company (each row) a simulated salary based on what type of job they are offering. We then divided the `employeecount` into three major groups, instead of eight. In this way, we reduced the chance that by having too specific categories, `employeecount` has too much influence on the regression model.

When thinking of the aimed audience, we wanted to aim this towards high schoolers who were looking at jobs and wanted to show the number of jobs correlated with the number of institutions in each state. By doing so, we used the variables pertaining to this idea, including admin1, JobCount, state, long, and lat. In addition to this, we imported other data regarding the number of institutions and center coordinates of each state.

We first "summarised()" all of the jobs in each state, grouping by admin1. This way, we would have one job count per state, which was not the case prior to summarising. After doing so, we joined two tables; one of which included the total job count per state and one which included the coordinates to create the map of the United States (from map_data("states")). We also joined another table that combined the table that was previously described, and one that included the center coordinates of each state. By combining these two tables, we were able to create a data graphic that included the map of the United States, with color correlating to the number of jobs and size of points correlating to the number of institutions.

For high school students, pursuing a higher education is a decision to make. Does going to college give them a higher chance of getting a better job? Thus, we used an index that shows the company values and the education requirements that companies have for their job offers. The result in our box plot shows that companies that require high school education have the highest company value, which is slightly higher than that of companies that require a higher education.

In our final graph, we wanted to analyze the relationship between the careers and companies found in indeed.com. We decided to categorize eleven general industries based on the careers provided in the data frame to incorporate our company value index. Using these two variables, we were able to demonstrate the overall value of companies found in each industry.