Team 6

Data For Good

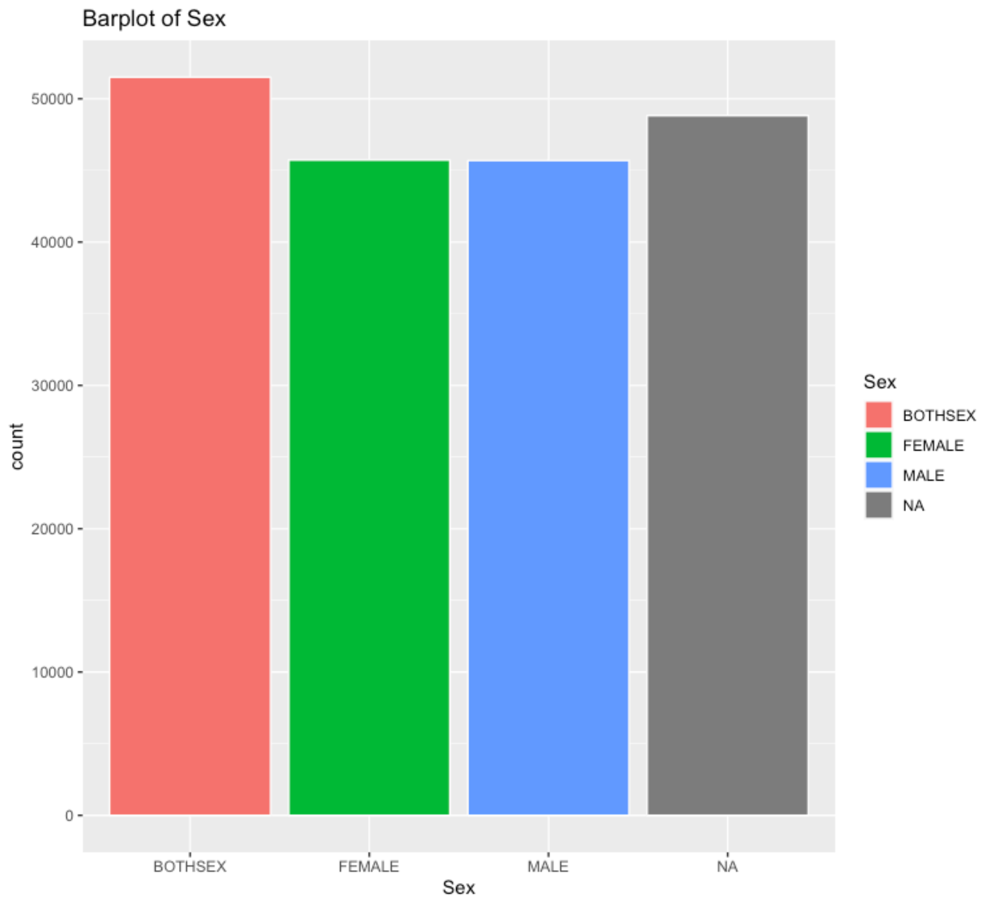
Overview of Approach

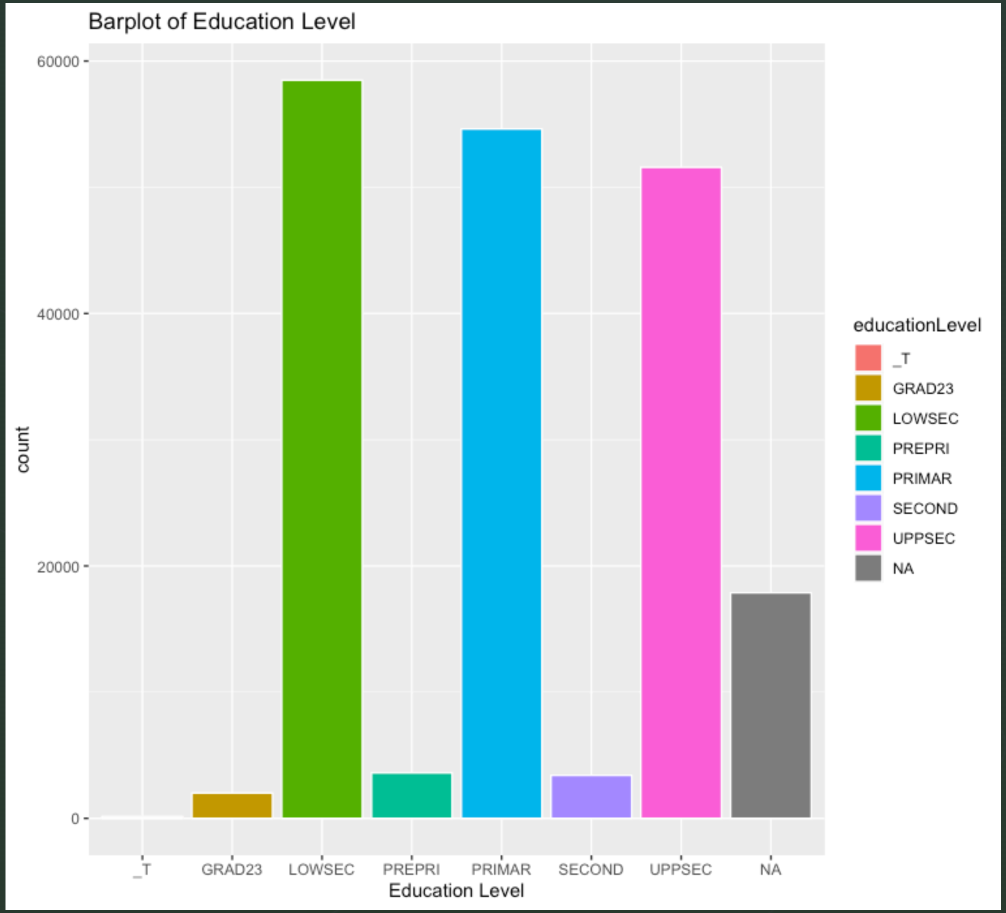
- Our approach was to first visualize the data to try to understand what we were working with.
- Established our main goals
- We prepared the data by cleansing/munging
- We each attempted more advanced visualizations and modeling autonomously, while checking in occasionally through Zoom

Insights into Findings

- We quickly discovered that valuable data were missing
- Age was completely missing, skill level was missing about 25%, and sex data was useless, as half of it was either N/A or both sexes combined
- The only data that appeared to be potentially useful were value,education level, geoAreaCode, and timePeriod
- This data proved cumbersome to work with, as most of it was categorical

Insights Into Findings (cont.)





Insights Into Findings (cont.)

- We should have run regressions on the data to find if the few useful variables were statistically significant, however, the data required extensive munging, and we ran out of time.
- We should have run different models to try to predict the value based on the geoAreaCode and the time series.
- The data also kept crashing Rstudio and Excel, making it difficult to run even basic visualizations or analysis.
- In retrospect, perhaps sampling the data at the beginning would have solved the issues with the demand on the software and some of the missing data points, while coding the discrete variables would have allowed them to be more useful for modeling.
- We would have liked to try to understand how the value could be used with the geoCode to make predictions of education level.

Showcase Working Files and Process

- We used Rstudio and Excel for our analysis. We were unable to get Jupyterhub to work well for us, so we worked locally

Problems

- We communicated through Zoom, however, our team did not collaborate well.
- We started with only three people, and eventually only one person was left participating for the last three hours of the event.
- The dataset was ambiguous and large, making the challenge more difficult. Many of the variables were difficult to understand.
- We were faced with trying to figure out new programs that weren’t working well from the beginning, as we all struggled with VPN issues, and two of the three of us didn’t have our repository files until the challenge had already started.

Recommendations

- Our mentors were very helpful, but we really needed a dedicated mentor that could work with us the whole time
- Require commitment from the group so members don’t vanish and leave the challenge
- Require collaboration, such as turning on cameras for Zoom
- Use a dataset that is less ambiguous

Conclusion

- This was a very challenging event, and we did not meet it well as a group. If we could have established a plan early on, and set benchmarks for each member to hit, we could have used our time better, and been more productive.
- Overall, it was still a great experience, and I’m looking forward to the next one!