Nick Moir & Matthew Gottfried

DS 2002

Professor Jason Williams

6 December 2024

Final Project Reflection Paper

When we embarked on this project, we did not know what we wanted to focus on. We considered looking at states that legalized sports betting and individual bankruptcies. However, we could not find ample data for us to analyze. We also considered looking at the maternal death rate in pregnancy by state after Roe v. Wade was overturned. The issue here was how recently the overturning took place, which we felt did not give us a large enough sample size. While we were discussing what path we wanted to go, we received an email notification from Chief Longo. This inspired us to think about how education affected the incarceration rate in Virginia. We decided to analyze how median household income and percent of individuals with less than a high school education influenced incarceration rates in each county. The biggest challenge we faced with data selection was finding an adequate sized dataset that would allow us to conduct accurate analysis.

Fortunately, we did not face many challenges implementing the ETL pipeline in this project. We were able to easily and efficiently upload, merge, edit, and export our files. This is because of the work that we did in project one. Our ETL pipeline operated flawlessly. When we were deciding how to analyse our data, we faced some challenges. Initially, we considered using histograms to display our data. However, we quickly realised that those did not present our data how we wished. We then realized that the best method to analyse the correlation was through an OLS regression. Another benefit of using linear regression were metrics such as pvalues and

R^2. When we were analyzing our data, the biggest challenge we faced was trying to find the best types of plots to interpret our data. After running the regression on our data, we could see the line of best fit in relation to the scatterplots. We also created a correlation heatmap to provide a different kind of visual to interpret the data.

The biggest challenge we faced was the cloud implementation. As neither of us were very familiar with Google Cloud except from our exposure in this course, we found the interface to be quite intimidating, and we had some challenges navigating it. However, we were able to persevere and properly uploaded our data using buckets and allowing public access.

Working as a team taught us many valuable lessons. The first was the importance of scheduling. Despite living together, it was very challenging to find times where both of us were available. Despite this challenge, we were able to work through it by working over break via zoom. Another challenge that we faced was that each of us had different strengths. We used this to our advantage by allocating parts of the project based on our individual strengths. We then came together to explain our work to one another. One benefit of working with someone that you already know, is that you know their personality and how they work best. We found this to be very helpful in how we divided up the project.

Although it provided many challenges, we found this project to be very enjoyable. Upon completing the project, we began to consider how we could apply this data to larger applications. In terms of improvements, we would want to look at larger datasets. In the future, ideally we would be able to get more specific data that stretches over longer periods of time so we can better analyse trends. Additionally, we would want to look at other variables that could also have an impact on incarceration rates, such as crime rates and the rank of public high schools in each county. Another thing that we would like to improve upon is being able to create a multivariate

linear regression model. Using this, we would be able to see the effect that each variable has on incarceration rate while holding all other variables constant. This could potentially remove any sort of omitted variable bias. Overall, we enjoyed this project and gained valuable skills that we can apply to other projects.