DS2002 Final Reflection Paper

Rachel Leach, Elizabeth Wu, Adair Hancock, Kendle Schooler, & Anisha Sapkota

Introduction

This final project for DS2002 focused on exploring the relationship between educational attainment and income. Using datasets from the U.S. Department of Education's College Scorecard and the U.S. Census Bureau's American Community Survey (ACS), we aimed to determine if obtaining a Bachelor's degree offers a meaningful return on investment. This reflection highlights the challenges we faced during data selection, ETL setup, analysis, and cloud storage, and discusses the lessons learned, skills gained, and areas for further development.

Challenges Faced

The first challenge we encountered was selecting reliable, large-scale datasets that were relevant to our research question. After considering several datasets, we chose the College Scorecard data from the U.S. Department of Education and the American Community Survey (ACS) from the U.S. Census Bureau due to the credibility and large sample size of the sources. However, while both datasets were highly valuable, merging them posed initial difficulties due to the differing formats and scales of information. The ACS data was more generalized at the state level, while the College Scorecard dataset provided detailed institutional data, requiring careful consideration in how to combine them.

Additionally, the Extract, Transform, Load (ETL) process, particularly the extraction and transformation stages, posed several technical challenges. For example, the College Scorecard dataset was too large to directly store in Github or Excel, which forced us to mount the file from Google Drive to Google Colab. This added complexity to the data loading process, especially when dealing with large CSV files. Furthermore, filtering and cleaning the data to isolate the most relevant variables as well as transforming and renaming columns required careful handling to avoid data loss or errors.

During the analysis phase, selecting the right visualization methods to represent the data was crucial. We used bar plots, scatterplots, box plots, and heatmaps, but choosing the most appropriate visual representation for the relationships we wanted to explore was not always straightforward. Additionally, while calculating correlations helped us identify key patterns, interpreting these relationships with confidence required a deep understanding of the data's limitations.

Storing the data in Google Cloud was necessary due to its size, but we faced some logistical challenges ensuring proper access control and settings. We opted for a multi-region storage configuration to maximize accessibility and set the data to public access since it was non-sensitive. Ensuring the data was accessible to everyone in the team and anyone else interested, while maintaining its integrity, required careful setup.

Lessons Learned

The project highlighted the importance of effective data cleaning and transformation. We learned that understanding the structure of the data upfront is critical to avoid errors later in the process. Using Pandas was helpful for data manipulation, but we realized that other tools like SQL databases could be beneficial for handling large datasets more efficiently in the future.

Though the members of the team had differing backgrounds and areas of expertise, the team as a whole worked well together. Each team member brought unique skills, such as programming, statistics, and visualization. However, we did experience some coordination challenges, such as finding time for group meetings and delegating tasks. Regular communication and thorough documentation of our work were essential for keeping everyone on the same page. In future projects, clearer role assignments and better use of project management tools would help better streamline workflow.

The key lesson was the importance of interpreting the data correctly. We found that while correlations were helpful in identifying patterns, understanding the causality behind these relationships was more complex due to potential confounding variables. This project revealed that deeper statistical analyses, such as hypothesis testing or regression modeling, are necessary to arrive at more robust conclusions.

Skills Gained and Areas for Further Development

As a team, we gained substantial experience cleaning and transforming large datasets, including merging two separate sources in a compatible manner. We also learned how to choose and create the most effective visualizations to communicate different data trends and relationships. Finally, the project reinforced our understanding of basic statistical methods and Python code used to summarize and analyze data.

While we performed basic correlation analysis, more advanced methods like regression analysis or machine learning would be interesting to learn and apply. This could provide deeper insights into the relationships between variables.

Conclusion

Overall, this project was a valuable learning experience that allowed our team to develop important technical and teamwork skills while addressing complex data analysis and communication challenges. We learned the importance of thorough data preparation, effective communication, and careful interpretation. These insights have not only enhanced our data analysis capabilities but also deepened our understanding of how to manage large datasets, perform statistical analysis, and work collaboratively on a major, multi-faceted project. Moving forward, we aim to further develop our skills in advanced analytics and cloud computing to tackle more complex projects.