Emily Phillips
DSC 540 – Data Preparation
Professor Williams
Nov. 19, 2021

DSC540 Final Project Summary

This final project for our Data Preparation course has been insightful, applicable, and intriguing! We were able to work with three different sources of data: a CSV file, a website, and an API. Each of my sources pertained to different pieces of my problem statement of how universities contribute to their students' post-grad salaries. Therefore, the final dataset was a nice combination of all the factors and measures, related to schools & earnings.

For each of the sources, I had to read in the data, convert it into a Data Frame for easier usage with Python, and then perform various data transformation and/or cleansing steps to prepare the data for the final step of merging and visualizing. For these steps that I performed on the three different sources, I mostly focused on making sure the data was readable (i.e., clear column names, wide vs. long formatting), identifying outliers, finding, and removing duplicates, and fixing casing or inconsistent values. It took continuous overviews of the data to identify areas where cleansing or transformation was needed, and then, I had to determine the best strategy for handling them.

I think the greatest area of thought for me was identifying and handling missing data. I did not want to immediately remove missing values, as I believed that most of them held value to investigating how one's university impacts their earnings post-graduation. Therefore, I took the strategy of removing missing data if it was the majority of a row or a column. If that was not the case, then I decided on various methods of filling, whether it was by some numerical aggregation, the most frequent value or with forward and backward fills.

However, as with any task of data preparation, there are ethical concerns with removing and manipulating data. Ideally, we would like to keep the source in a state as similar as to how

we receive it. By making any changes, there are increased risks of losing valuable data and introducing bias. It can be easy to formulate the data in a way that perfectly aligns with one's business problem or question. Therefore, for my data preparation work with this project, I tried my best to keep the data as similar to how it came from the source. I mostly removed or filled data that was causing skew or errors with groupings and inconsistent values. I didn't want to lose the original structure and intention of the data, and I think this can be a very thin line that is easily crossed with data preparation!

In the end, the visualizations were a nice way for me to wrap-up my data and to gain insights around my problem statement. This visualization step required data understanding, and five visualizations just scratched the surface of what I believe this whole dataset can convey about the statement!