**Shannan Sharp June 8, 2021**

**Madison Lincks**

**Anthony Kerins**

**Project 2 - Technical Report**

Our project seeks to determine which college provides the greatest return on investment using median salary and starting median salary and school rank as the basis for comparison. I major component of rank is the cost of attending college and is a obvious comparison to salary.

**Extracting the data**

1. We loaded our CSV files and read them using Pandas

CSV files were pulled from <https://www.kaggle.com/wsj/college-salaries?select=salaries-by-region.csv> and <https://data.world/ian/united-states-university-rankings>

1. **Salaries by college type**
   1. In this persons research they took schools based on ivy, party, and engineering status’s to see which type of schools saw the highest earning increase since receiving job positions newly graduated from college. In their description they state “You already know your starting salary will be different based on the type of school you attend. But increased earning power shows less disparity”. In which they mean the salary will more than likely be based on the type of school you go to, in contrast to believing the higher the degree the higher the salary.
2. **Salaries by region**
   1. The research conducted by PayScale Inc. found that attending college in the Midwest leads to a lower salary both at graduation and at mid-career compared to those graduated in the Northeast and California.
3. **University Rankings**
   1. This data file aims to provide insights into more than 2000 schools to help students compare the academic quality of the highest ranked universities, and the potential value vs. cost of pursuing a degree at a top school.
4. From the 3 different data frames we then began the process of cleaning up the files into similar column headings for the transformation process

**Transforming Our Data**

1. We transformed our data into 2 new data frames
   1. The first data frame we created was the “ranking” table
      1. “name” column
      2. “rank” column
      3. “state” column
   2. The second data frame we created was the “school” table
      1. “school name” column
      2. “starting median salary” column
2. This allowed us to get a better visualization of our transformed data so we could load it into Postgres for further analysis
3. We also were able to eliminate several columns that were not needed for our analysis

**Loading Our Data**

1. To load in our data we cleaned using pandas we used PgAdmin and then we created tables and then merged the tables as well.
   1. Merge “school” table and “rankings” table:
      1. Merge using “school name” column and “name” column as common attribute
      2. Once merge complete compared to previous lists to identify no duplicates
   2. We then sorted the newly created table by greatest median salary to least for reference and comparison