College Salaries

Alec Schwager, Mia Salza, Shae O'Neil, Zach Schuring

1. Introduction

Students must account for many factors when deciding where to attend college. From personal experience, proximity to home, expected tuition costs, and enrollment. Statistics and analysis can play a major role in determining how students spend their undergraduate years. One item on many students' minds is their potential earnings their degree will provide after graduation. While the "college experience" cannot necessarily be quantified, there are data points available to offer insight on the salary students can expect to earn with a degree from a given conference and institution based on historical figures.

In this project, we utilized salary data from academic institutions in the "Power 5" athletic conferences with varying characteristics in attempt to reveal the factors that greatly impact students' earnings following graduation. It should be noted that the Power 5 schools do not represent the full range of academic institutions available to students but were chosen to provide a subset of the larger dataset to be analyzed.

2. Data

This project uses two primary sources of data: Wikipedia data on Power 5 school conferences and a Kaggle dataset, "Where it Pays to Attend College".

2.1 Where is Pays to Attend College

The Kaggle dataset, "Where it Pays to Attend College,"[2] compares salaries by college, region and major.[1] The dataset is divided into three different tables formatted as CSV files: degrees that pay back, salaries by college type, and salaries by region.

The degrees that pay back table includes 50 unique college majors and the percent change between the medians for starting salary and mid-career salary. The salaries by college type table have 249 unique school names, but 269 instances every school that is classified as a party school is also classified as either a state, liberal arts, engineering, or ivy league school, duplicating some school instances but with different school types. The last table, salaries by region, includes 320 unique school names and their region: either Northeastern, Southern, Midwestern, Western, or California. However, we chose to omit the tables "degrees that pay back" and "salaries by college type".

All three tables have the following attributes: starting median salary, mid-career median salary, mid-career 10th percentile salary, mid-career 25th percentile salary, mid-career 75th percentile salary, mid-career 90th percentile salary. We cleaned the data by removing dollar signs ($) and replaced them as blank. The pre-preprocessing code is included in the "Power 5" R Script with the data frame named "all_schools."

2.2 Power 5 Conference

To identify and collect data on Power 5 conferences, we used the main Wikipedia article on teams in the Power Five conferences[1]. This page contains a section titled "Current conferences and teams" that lists all the Power 5 conferences (ACC, Big Ten, Big 12, Pac-12, and the SEC) and the member universities that belong to each conference. Each school and conference are accompanied by a hyperlink to the place's respective Wikipedia page.

Based on the table on Wikipedia we wrote a find all function to get all the conferences listed. We then scraped all the schools listed under each conference. After getting all the schools in each conference we created a data frame, of all schools and their respective conference. By vertically merging each of the conferences it created a data frame with 65 observations and 2 variables. The scraping is included in the R Script "Power5" with the data frame named "Power 5".

2.3 Integrate Power 5 Conferences and "Where it Pays to Attend College"

We then had to make sure all the school names were uniform (e.g., Arizona mixed with Arizona State) to the Kaggle dataset for merging purposes. Since our two data sets shared a column "School.Name" we were able to horizontally merge the two datasets. This data frame is included in the R Script "Power 5" named "SchoolsByConference" with 65 rows and 10 features.

Table 1 Data Dictionary (Schools By Conference)

Column	Data Type	Description
School.Name	Text	Name of School
Conference	Text	School conference ("Pac-12, SEC, Big 12, Big 10, ACC")
Starting.Median.Salary	Numeric	Starting median salary for each school
Mid.Career.Median.Salary	Numeric	Mid-career median salary for each school
Mid.Career.10th.Percentile.Salary	Numeric	10th percentile salary for each school
Mid.Career.25th.Percentile.Salary	Numeric	25th percentile salary for each school and major
Mid.Career.75th.Percentile.Salary	Numeric	75th percentile salary for each school and major
Mid.Career.90th.Percentile.Salary	Numeric	90th percentile salary for each school and major
Percent change from Starting to Mid-Career Salary	Numeric	Percentage to reflect increase in salary over time
Region	Text	Region school resides in

[1] https://en.wikipedia.org/wiki/Power_Five_conferences

[2] https://www.kaggle.com/wsj/college-salaries

3. Analysis

The goal of this project is to examine which schools and conferences have the best outcomes from college, specifically starting median salaries and career growth.

3.1 Mean starting salary by school's region

Which region of the United States has the highest mean starting salary? We created an aggregate function to create a data frame that lists each region and their average median starting salary. Table 2 displays the resulting summary table.

Table 2 Aggregate Function Summary

Region	AvgMedianStartingSalary
California	59425.00
Midwestern	47033.33
Northeastern	49650.00
Southern	46611.11
Western	46100.00

The summary table indicates that the California region has the highest average starting median salary. We then created an ANOVA test to determine if there was a was a relationship between the region and the starting salary. We can conclude based off the ANOVA summary that confirms that there is a significant between starting salaries and region (p=4.74e-05). The graph below (Figure 1) shows the distribution of Average Median Salary in the form of a ggplot2[3] histogram. The distribution backs up the results of ANOVA test as each region has similar results other than California.

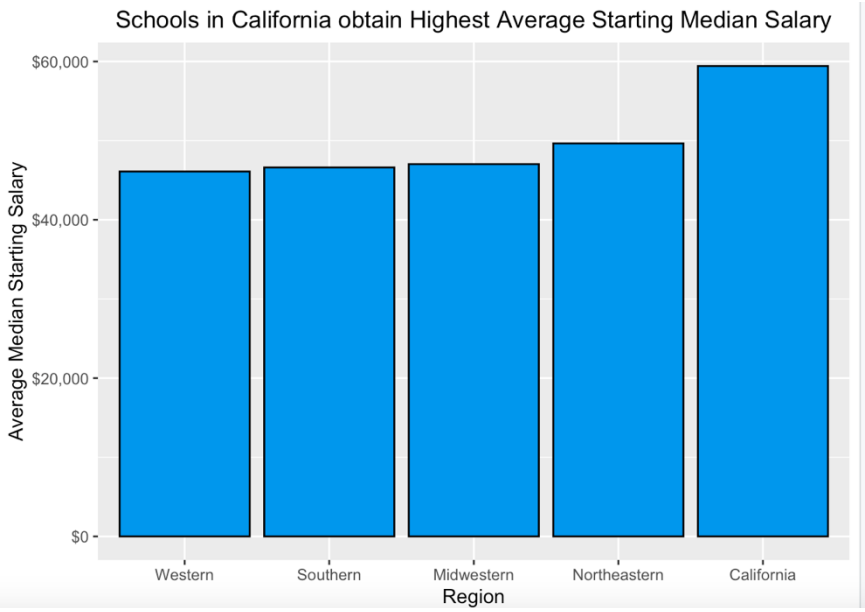

Figure 1 Distribution of Average Starting Median Salary and Region

[3] https://cran.r-project.org/web/packages/ggplot2/

3.2 Average salary growth by school's conference

Which conference in the Power 5 has the highest median salary growth? We created another aggregate function to calculate the average percentage growth in median income from the start of the career to the midway point of the career for each conference. Table 3 displays the resulting aggregate summary table.

Table 3 Aggregate Function Summary

Conference	Percent.Growth
ACC	257.22
Big 12	280.46
Big Ten	260.93
Pac-12	267.62
SEC	274.86

The summary table shows steady salary growth across all conferences, with the Big 12 and the SEC narrowly leading the way at 280.46% and 274.86% growth, respectively. Given this information, it appears that no conference has a significant advantage in terms of offering extraordinary salary growth to mid-career. To help visualize the average earnings growth across all schools in the 5 conferences, we created a stacked vertical bar chart (Figure 2) separating the average median salaries for the respective conferences and career statuses. The chart is shown below (Figure 2).

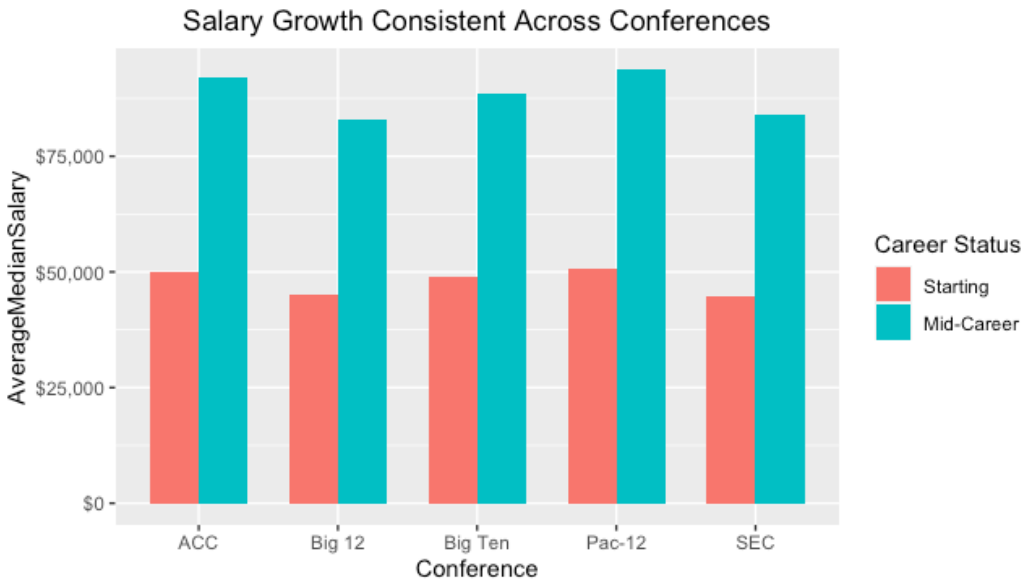

Figure 2 Salary Growth by Conference

3.3 Minimum and maximum starting salaries by school's conference

Which conference in the Power 5 has the highest and lowest median starting salary? Are there large discrepancies between the minimum and maximum starting salaries across each conference and within each conference? We created a dplyr[4] summary table to show the maximum and minimum starting median salaries for each conference in the Power 5. Table 4 displays the resulting summary table.

Table 4 Starting Salaries by Conference Summary

Conference	Maximum Starting Salary	Minimum Starting Salary
ACC	58900	42100
Big 12	49700	42400
Big Ten	52900	44700
Pac-12	70400	42200
SEC	51200	40000

The summary table shows that the minimum starting salaries are fairly even across conferences ranging from $40,000 to $44,700; however, the maximum starting salaries differ quite a bit. The Pac-12 conference has the highest maximum starting salary at $70,400 and behind it is the ACC at $58,900. The ACC and Pac-12 also have similar minimum starting salaries at $42,100 and $42,200, respectively. These conferences show the largest difference in minimum and maximum starting salaries out of all conferences in the Power 5. We displayed these results in a grouped bar chart using ggplot2, which displays the minimum and maximum starting salaries grouped by each conference. Figure 3 displays the resulting grouped bar chart.

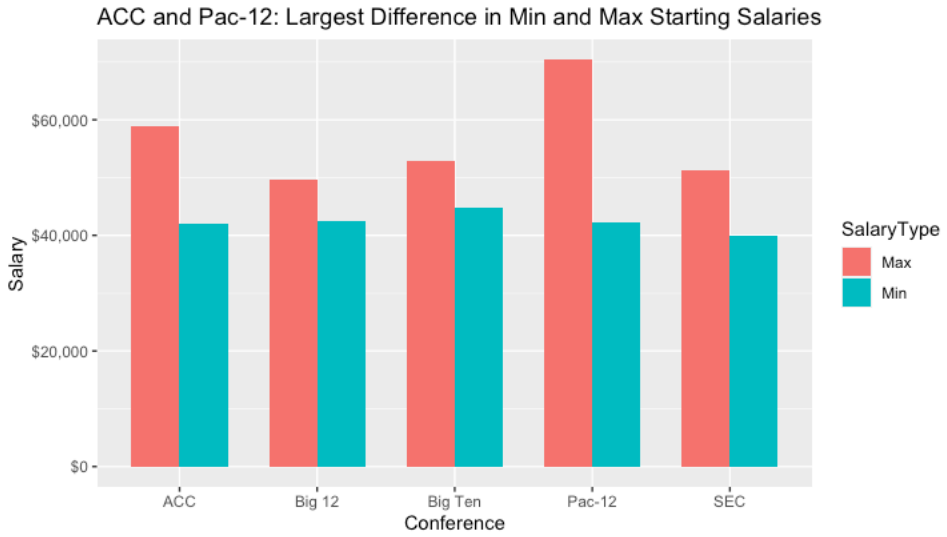

Figure 3: Minimum and Maximum Starting Median Salaries by Conference

[4] https://cran.r-project.org/web/packages/dplyr/

3.4 Starting Median Salary Fitted with Region and Conference

We thought it was necessary to see how region and conference affected starting median salary. The model produced the output below. It seems the Big Ten and Pac 12 are the only conference that value to median starting salary. All regions lower the starting median salary due to California's significance, as well as it being the feature removed by R when creating our model.

Table 5 Regression Model Output

Intercept	50,084
Big 12	-4,249
Big Ten	992
Pac 12	9,140
SEC	-5,329
Midwestern	-2,573
Northeastern	-930
Western	-13,325

When creating a regression model, it is necessary to test assumptions to validate the model accuracy. Below (Figure 4) is a Q-Q plot of the model's residuals. The graph indicates that data passes the normality test and Coefficient sizes are accurate. The indication is seen by the marks rarely deviating from the line.

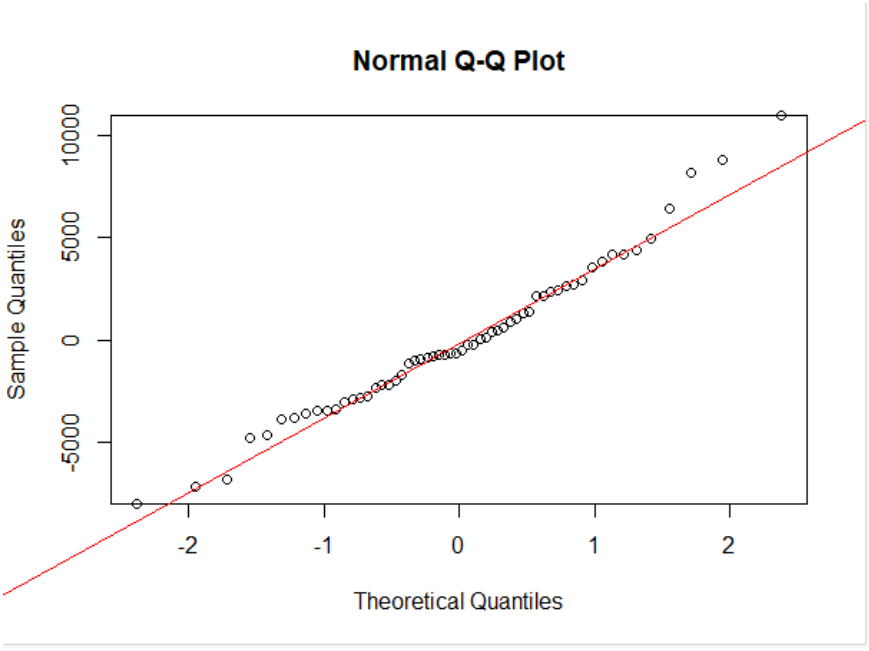

Figure 4 Quantile-Quantile Plot

Another assumption we wanted to visualize was the Homoscedasticity of the model which checks If instances have equal variances. The plot below (Figure 5) shows no true pattern except a slight skewness to the right. This test proves our regression model is appropriate for our data.

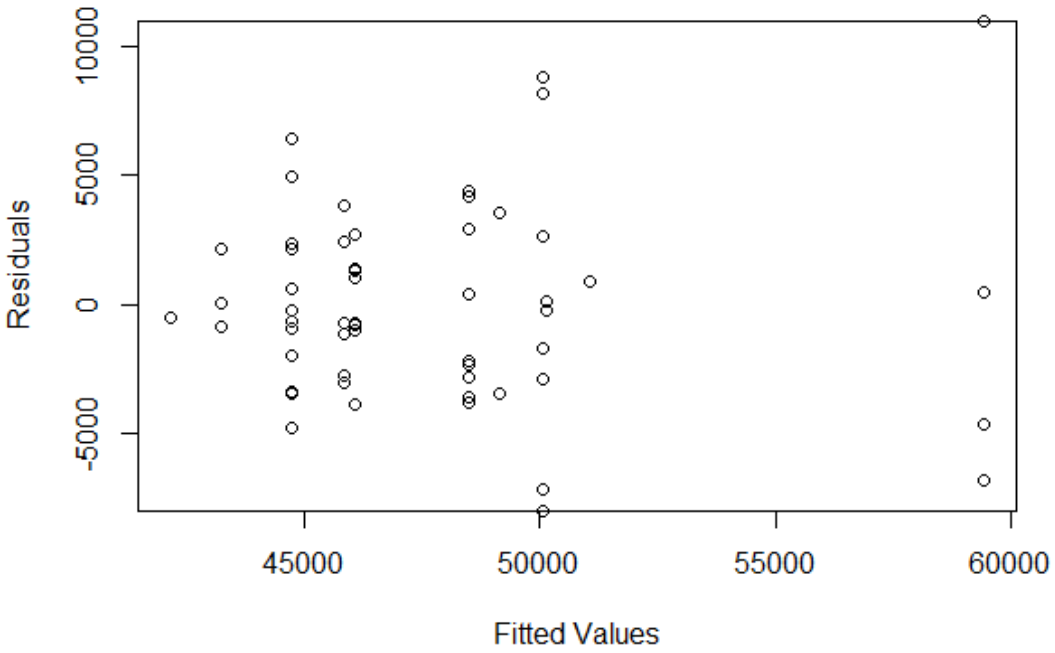

Figure 5 Homoscedasticity Scatter Plot

4. Conclusion

In this project, we incorporated Kaggle Data for all different types of schools with various information on them as well as Power 5 conferences scraped from Wikipedia. Specifically, we performed 4 analyses considering the relationship between starting median salaries and schools conference, the average growth of salaries based on school's conference, and the maximum and minimum starting salaries considering school's conference. We used summary tables, visualization methods, ANOVA tests, and aggregate functions to make assumptions based off evidence about salaries and Power 5 conferences. First, we reject the null hypothesis that there is not a significant difference between average median salaries across regions; therefore, we can conclude that salaries between regions do differ. Moreover, the Big 12 has the highest overall growth of salaries. The Pac-12 has the highest max starting salary, while the minimum starting salaries are all relatively equal. Overall, based off of our regression modeling, different conferences and regions produce different outcomes on salaries, with the exception of median salaries.

Throughout this project we encountered several limitations that are important to note. The first one being that the schools that belong to Power 5 conferences to do not represent all colleges in the United States. A second limitation on top of that is that there are many more schools in each region than we had listed. Finally, some regions have more prestigious schools versus others which can contribute to higher salaries, or outliers in the data. For example, the California region has UCLA, University of California at Berkley, USC, and Stanford. Future work could include all colleges in the United States with adequate information.