
Every year, senior year students in high school need to make choices before they graduate: where to attend colleges, what type of schools to go to, and what majors to study in colleges. By analyzing college graduates' salary data, we found that the locations of colleges, majors to study and types of colleges do impact the college graduates' starting salaries and mid-career salaries down the road.

1. Datasets

   We have three datasets downloaded from Kaggle.com.

   i). Salaries for College by Type (Salaries-by-college-type.csv):

   Each row contains School Name, School Type, Starting Median Salary, Mid- Career Median Salary, Mid-Career 10$^{th}$ Percentile Salary, Mid-Career 25$^{th}$ Percentile Salary, Mid-Career 75$^{th}$ Percentile Salary, and Mid-Career 90$^{th}$ Percentile Salary.

ii). Salary Increase by Major (Degrees-that-pay-back.csv)

Each row contains Undergraduate Major, Starting Median Salary, Mid-Career Median Salary, Percent change from Starting to Mid-Career Salary, Mid-Career 10th Percentile Salary, Mid-Career 25th Percentile Salary, Mid-Career 75th Percentile Salary, and Mid-Career 90th Percentile Salary.

iii). Salaries for Colleges by Region (Salaries-by-region.csv)

Each row contains School Name, Region, Starting Median Salary, Mid-Career Median Salary, Mid-Career 10th Percentile Salary, Mid-Career 25th Percentile Salary, Mid-Career 75th Percentile Salary, and Mid-Career 90th Percentile Salary.

2. Data Wrangling

   All the salary columns in the three datasets are strings with '$' at the beginning, and the ',' between digits. We first replace all the '$' and ',' with empty spaces and then change the data type from string to numeric. In the Salaries-by-college-type dataset, we found that there are 38 null values in both Mid-Career 10th Percentile Salary and Mid-Career 90th Percentile Salary columns. We also noticed that the null values in the Mid-Career 10th Percentile Salary column and the Mid-Career 90th Percentile Salary column always occur in the same row. Since the percentage of null values in these two columns is 14% and 10th Percentile Salary and 90th Percentile Salary do not represent the majorities, so we decided to remove these two columns from the dataset. Very similar to the Salaries-by-college-type dataset, the Salaries-by-region dataset has 47 null values (15%) in both Mid-Career 10th Percentile Salary and Mid-Career 90th Percentile Salary columns, and null values in these two columns always occur at the same row. We removed these two columns from the Salaries-by-college-type dataset too.

3. Exploratory Data Analysis and Initial Findings

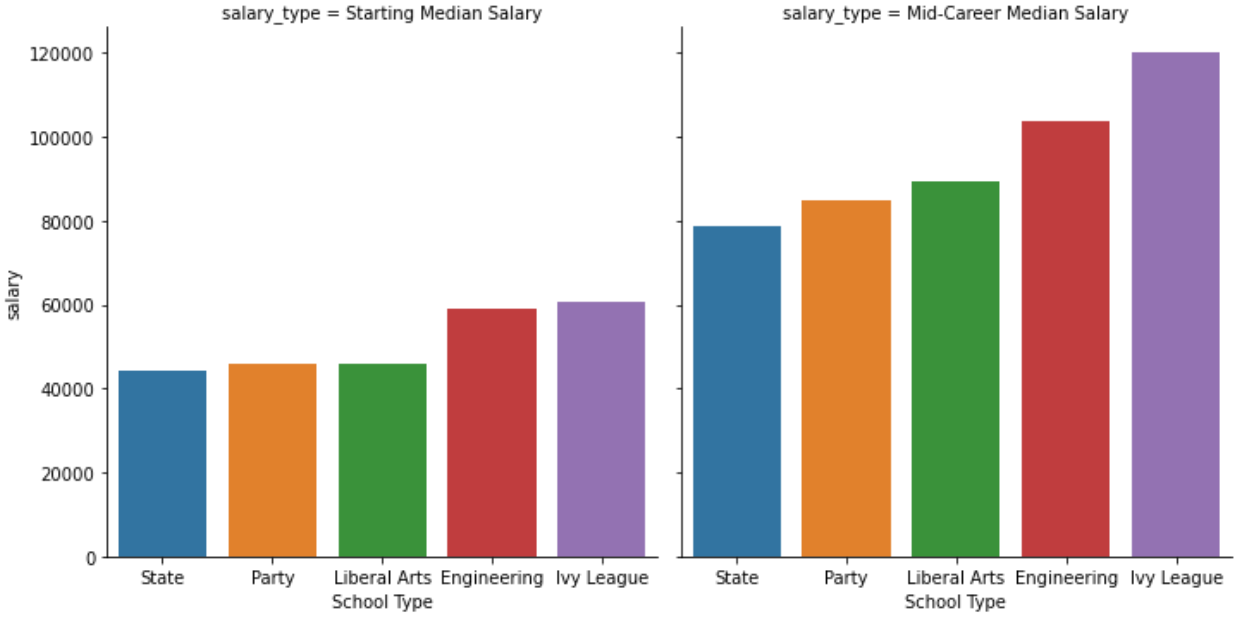
   I) Our initial finding from Salary Increased by Major is the following:

   Graduates with major Spanish have lowest Starting Median Salary $34,000
   Graduates with major Physician Assistant have highest Starting Median Salary $74,300
   Graduates with major Physician Assistant have lowest percent change from Starting to Mid-Career Salary 23.4%

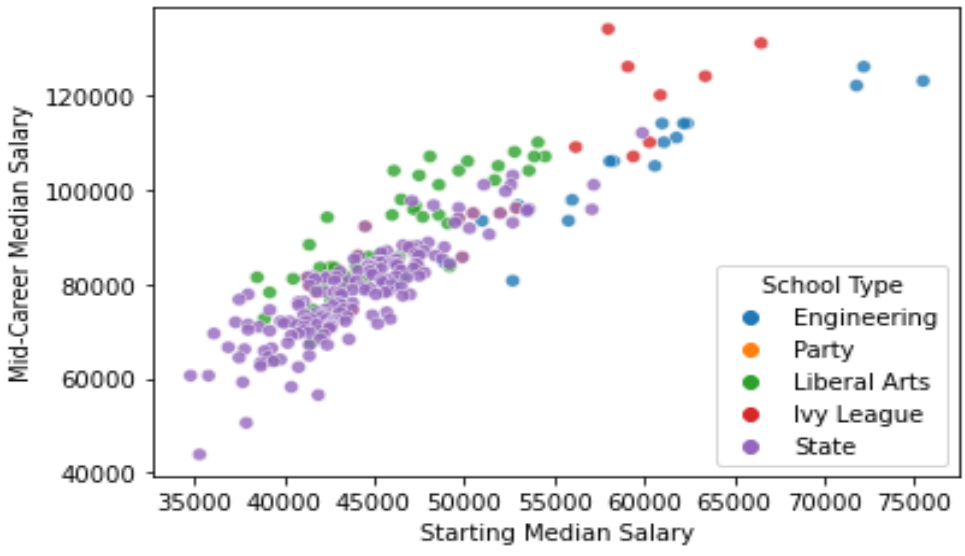
Graduates with major Math or Philosophy have highest percent change from Starting to Mid-Career Salary 103.5%
Graduates with major Education or Religion have lowest mid-career median salary $52,000
Graduates with major Chemical Engineering have highest mid-career median salary $107,000

II) From Salaries for College by Type, we have the following findings:

Graduates from Ivy League schools have the highest mean starting median salary $ 60,475 and the highest mid−career median salary $120,125. Graduates from Stat e schools have the lowest mean starting median salary $ 44,126, and the lowest mi d−career median salary $78,567

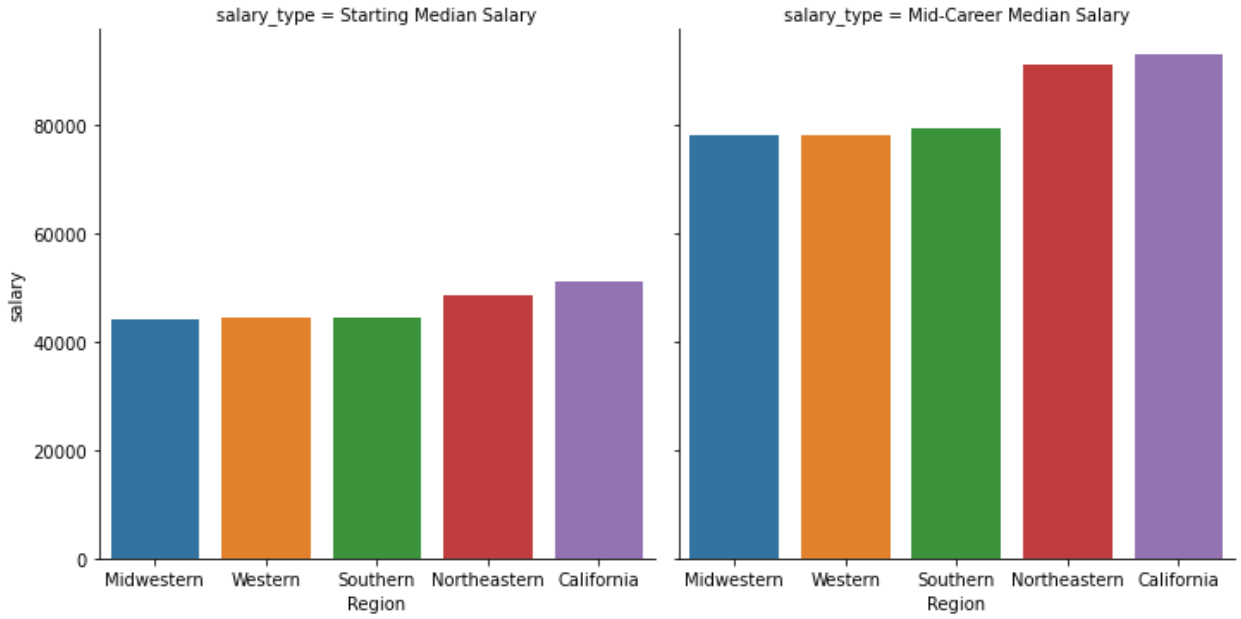



The relationship between starting median salaries and mid-career median salaries is shown in the scatter plot

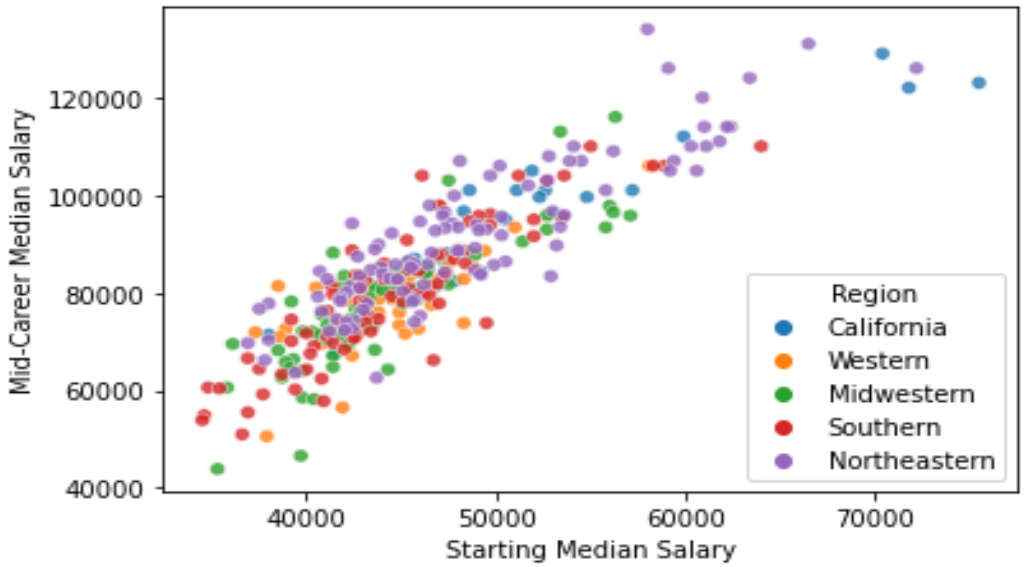


III) Our findings from Salaries by Regions are as follows:

Graduates from colleges located in California have the highest mean starting median salary $51,032. Graduates from colleges located in the Northeastern regio nhave the highest mean mid−career median salary $91,352. Graduates from colleges located in Midwest region have the lowest mean starting median salary $ 44,225, and the lowest mean mid−career median salary $78,180




The relationship between starting median salaries and mid-career median salaries is shown in the scatter plot



4. Preprocessing for Modeling
   i). Combine Salaries for Colleges by Type and Salaries for Colleges by Region based on School Names
   ii). Split combined data into training set and test set
   iii). Change categorical columns School Type and Region into dummy variables
   iv). Scale data

5. Modeling
   We used Linear Regression and Random Forest Regression to model our starting median salaries and mid-career median salaries. The metrices for these two models are the following:

| | OLS | RandomForest |
|---|---|---|
| R2_score | 0.81 | 0.71 |
| Mean absolute error | 4423.734 | 5351.23 |
| Mean squared error | 31801440 | 48819870 |
| Sqrt of MSE | 5639.276 | 6987.122 |

6. Conclusion
   By comparing the metrices for these two models, the Linear Regression model performs better than Random Forest Model.

7. Future Work
   - Do not combine Salaries for Colleges by Type and Salaries for Colleges by Region based on School Names, repeat the work on them separately to figure out how the school types and locations of colleges affect the starting salaries and mid-career salaries.
   - Calculate and add a column Percent change from Starting to Mid-Career Salary into both datasets and find out what school types and

regions have the greatest percentage salary increase10 years down the road.