Exploring Post College Salaries

Owen Morehead

December 28, 2021

Background

This dataset from The Wall Street Journal (2017) contains post graduation job salary information for over 200 of the most popular colleges in the United States. A top factor that influences many students target colleges is the salary they could recieve after graduating. Thus, analyzing this dataset, which contains the colleges school type, location, and median career salaries can provide insights to many of the millions of college sudents in the U.S.

Data Preparation and Cleaning

After obtaining the data from Kaggle there are a handful of modifications that need to be done before we can run any analysis. The two datasets we will be working with here are salaries by college and type, and salaries by college and college region. Each of these data sets share the following features:

- starting median salary
- mid-career median salary
- mid-career salaries for the 10th, 25th, 50th, 75th, and 90th percentiles

One of the first things to note is the sizes of each data table. The college by type table has dimensions (269, 8) while the colleges by region table has dimensions (320, 8). We see that there are more colleges by type than there are by region. We also note that there is only 1 college in the type data that is not in the region data and this is for Embry-Riddle Aeronautical University (ERAU). However, we can still use this college data since we know its region. On the other hand, unfortunately, there are 52 colleges that have an identified region but not a type. If one had the time they could lookup each of these colleges online and manually fill in their region, but for the purposes of this project I will discard these data since we will still have over 200 colleges to analyze.

Continuing the initial analysis, we note that for both tables, the columns which have NA values are Mid.Career.10th.Percentile.Salary and Mid.Career.90th.Percentile.Salary. Removing the mid career percentile salary data will not severely hinder any further analysis.

From here we can merge the two data frames by their school name keeping all the names in the college type data so that each college will have an applicable type and region. We also

make sure that each college name is unique as it should be. In addition we need to make sure all numeric salary data are readable by removing the *$* and *,* symbols from all the values. This can easily be achieved with the *gsub()* function.

Now we have our cleaned data as previewed below.

Table 1: College Salary Data

School Name	School Type	Region	Starting Median Salary	Mid Career Median Salary
Amherst College	Liberal Arts	Northeastern	54500	107000
Appalachian State University	State	Southern	40400	69100
Arizona State University (ASU)	State	Western	47400	84100
Arkansas State University (ASU)	State	Southern	38700	63300
Auburn University	State	Southern	45400	84700
Austin Peay State University	State	Southern	37700	59200

Visualization

Let us first visualize the frequency of our category types. Refer to Figure 1 below. Notice there are many more party schools in this dataset compared to the other school type categories. The distribution of general school location has a more even distribution, with the most school being drawn from the Northeastern region of the U.S.

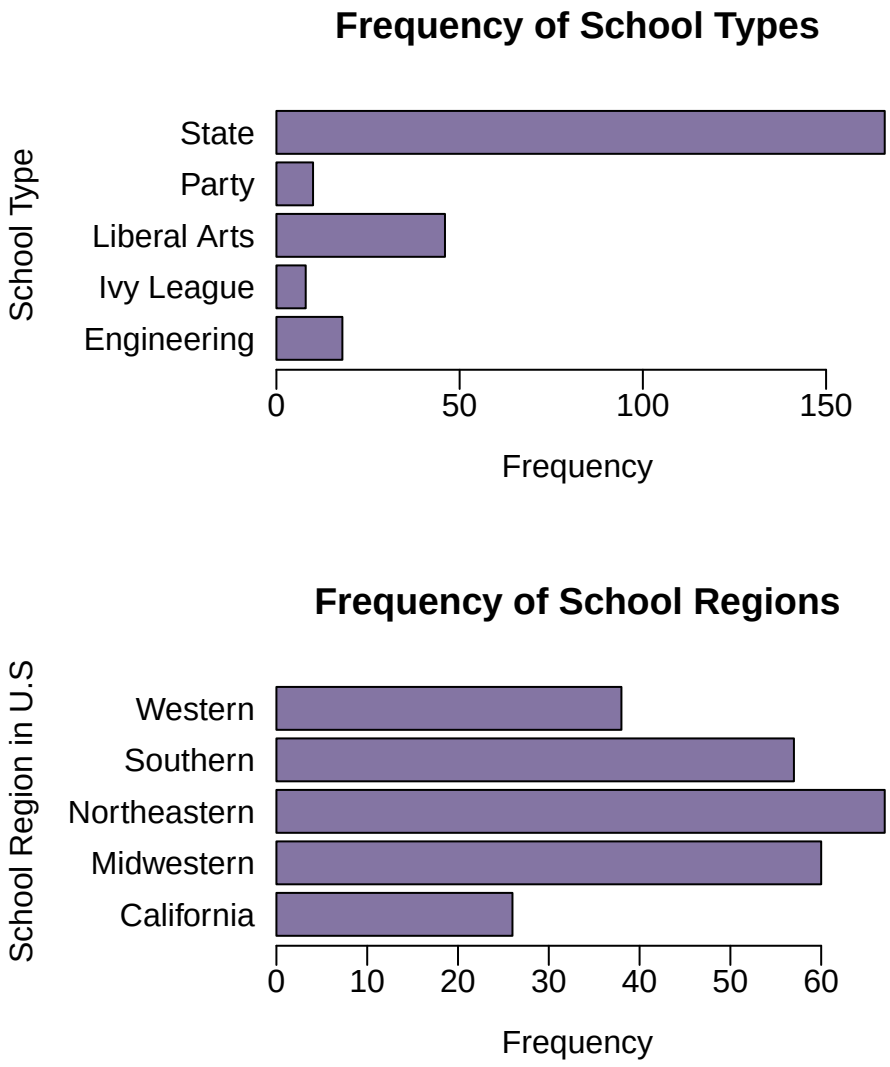

Figure 1: School Type and Region Histogram

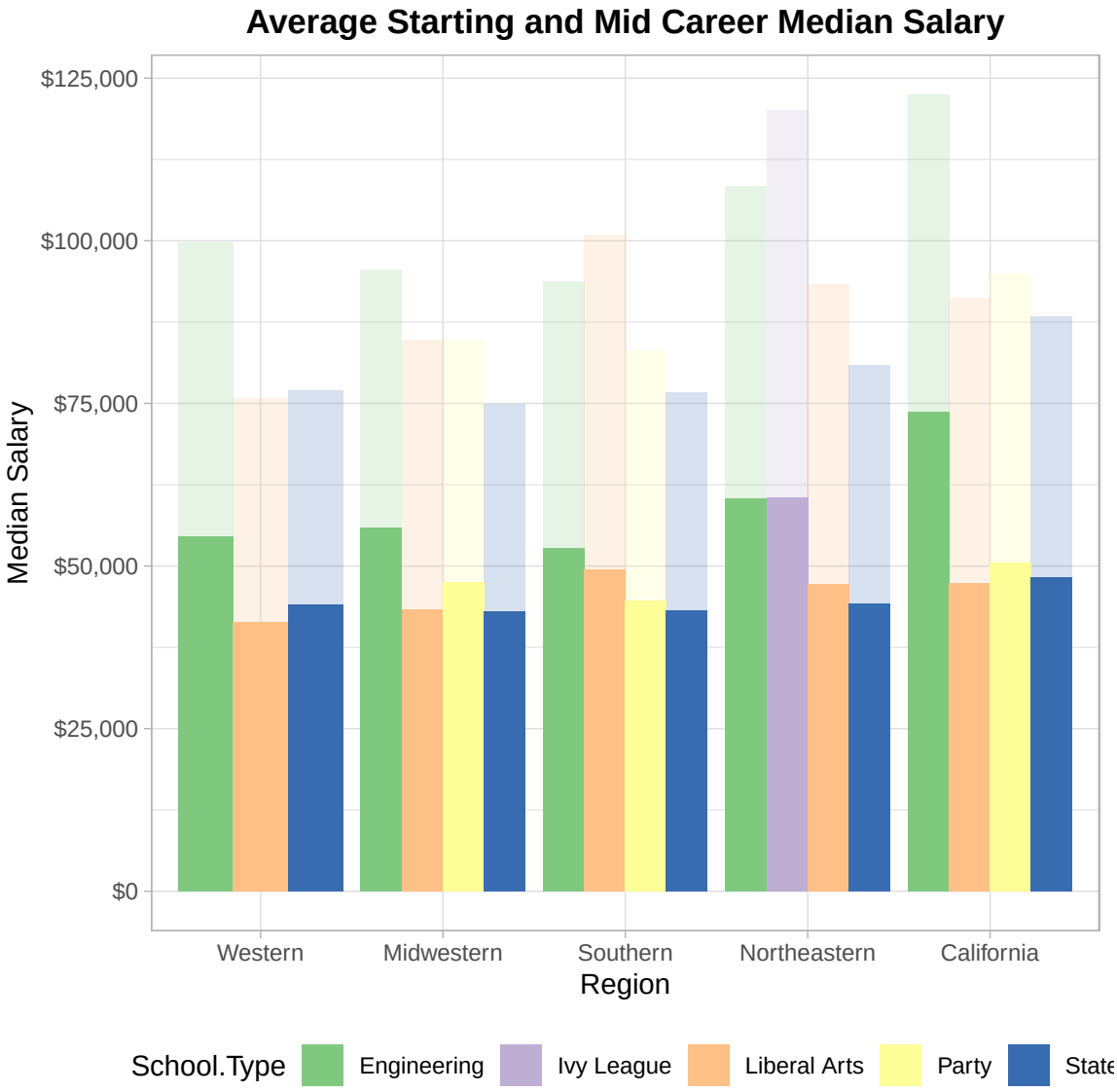

Figure 2: Averge Starting and Mid Career Median Salary by Region and School Type

A useful visualization would be to see the starting and mid-career salaries categorized by both region and school type. Refer to Figure 2 above. This also lets us easily see the change in the average median salary from starting to mid-career. (Sorry for using 'average' and 'median' next to eachother. This is beacuse our data is given as median salary values, and for this plot I am averaging these values to get the height of the bars.) We see that the better starting and mid career salaries across almost all regions are dominated by engineering schools. The only ivy league schools recorded in this data were in the northeast and they have a high starting and mid career salary as well.

Specifically, we see that for the average median starting and mid career salaries, the maximum values are
$ 73650 and $ 122500, both of which are for Engineering schools. On the other hand, the minimum values of the average median starting and mid career salaries are $ 41357 and $ 74848, and are for Liberal Arts and State school types respectively. The maximum median salaries are almost 2 times greater than the minimum median salaries which is a significant difference. At a glance this data supports the idea that in todays day and age, the college school type plays a significant role in the salary one makes after graduation, wether that be starting or mid-career.

To more easily visualize any linear relationships between variables (without interactions), we can essentially split up the barplot above (Figure 3 below). At a glance, the relationship between school type and salary seems fairly linear. So does that between school region and salary, except slightly less so.

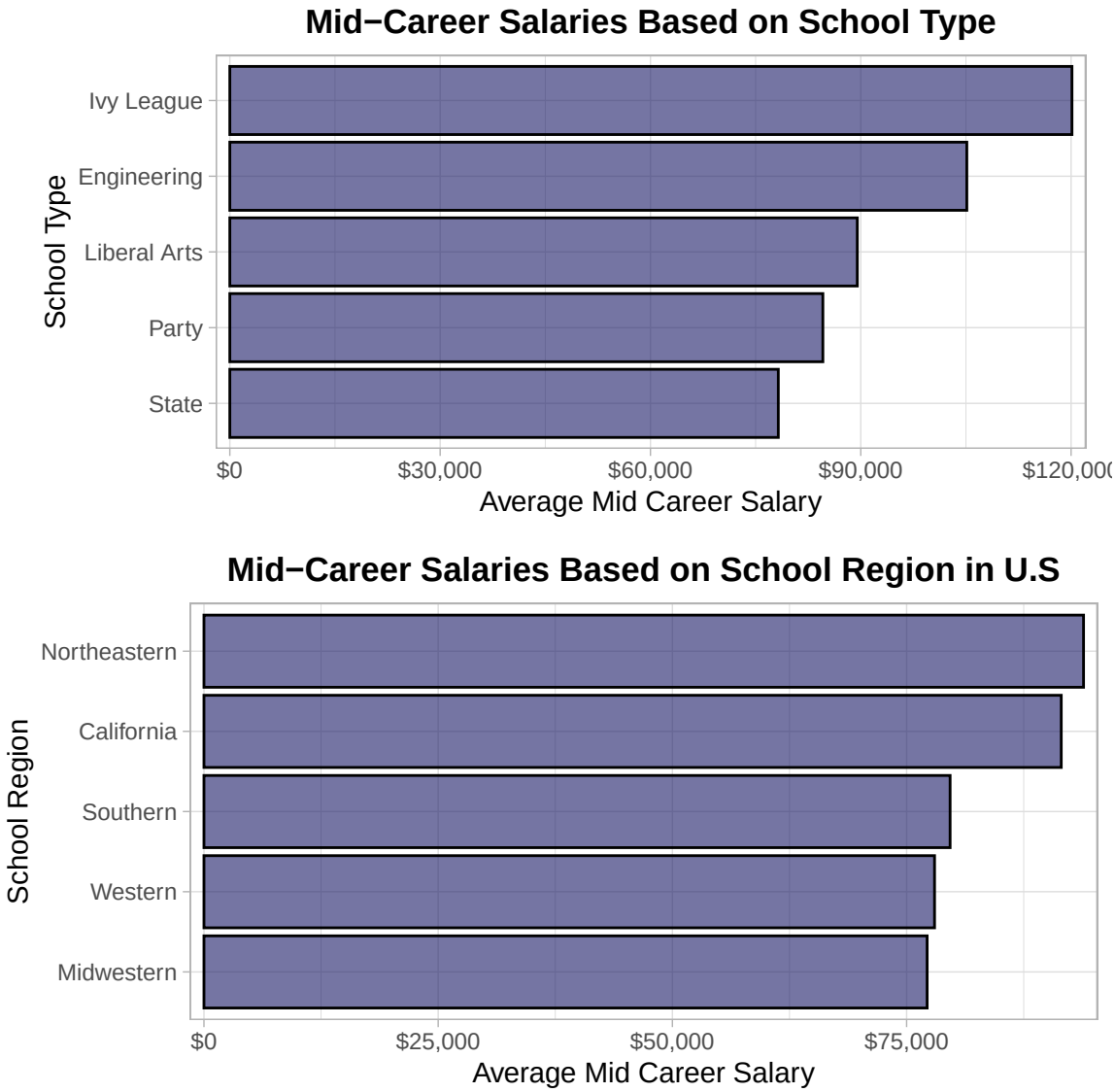

Figure 3: Mid-Career Salary by School Type and Region

We are starting to build an idea that a linear regression might fit this data well. Figure 4 below is also a visualization which would support this. We can see a fairly strong linear relationship between starting and mid-career salaries.

Between the starting and mid career salaries we find a coorelation coeficient = 0.8945. This shows that the strength of the positive linear relationship between these variables is high. Indeed it would be appropriate here to perform some linear model fitting in order to try and predict the mid career median salary from the starting career salary as well as the school type and region.

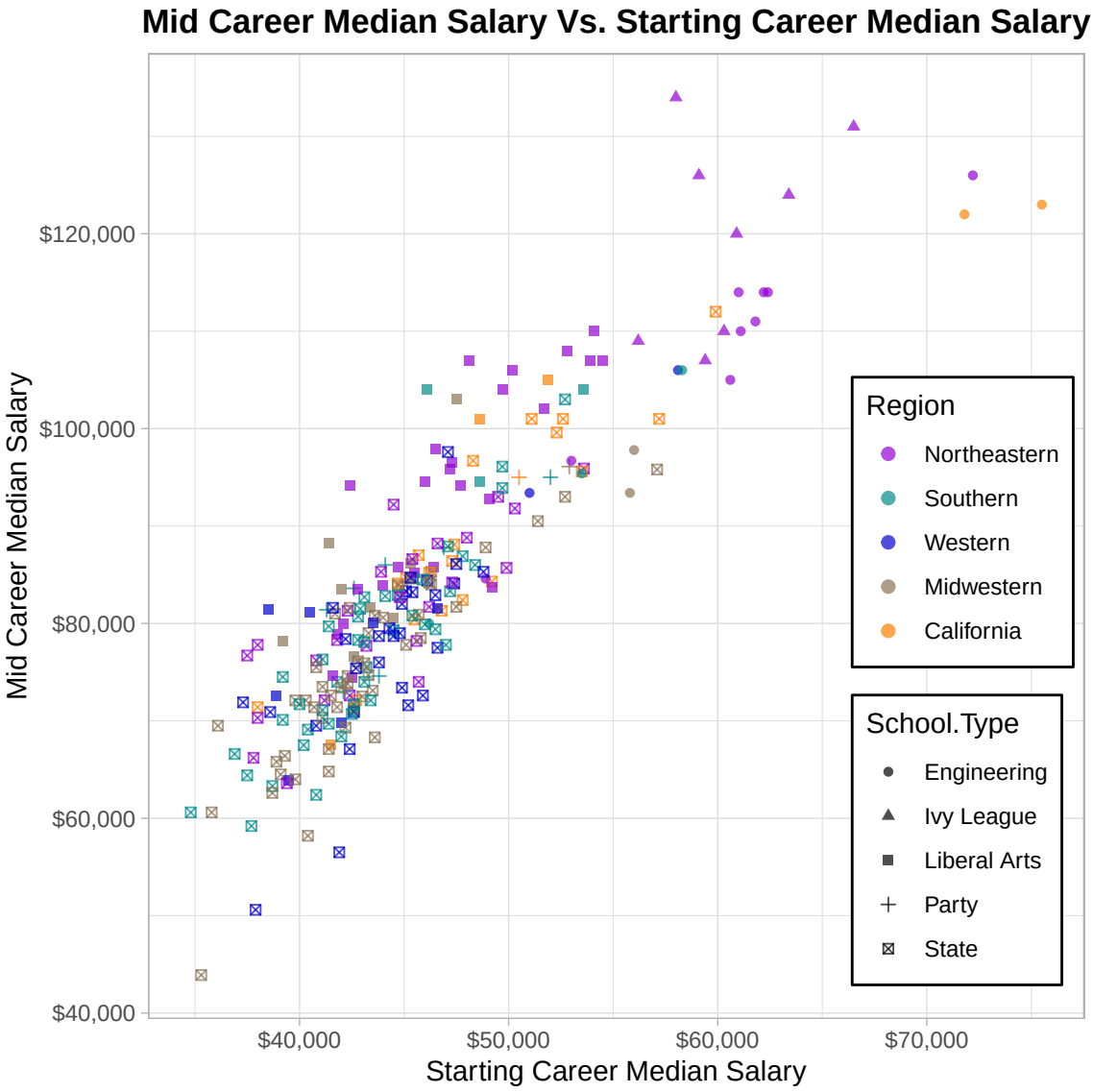

Figure 4: Mid and Starting Career Salary Comparison

Linear Model Analysis

Let us now formally ask, is there a relationship between median mid-career salary and the other three variables in the data? Our null hypothesis is that there is no relation between any of the predictors and the response. After running some linear models we should be able to either reject or fail to reject the null.

We can visualize the coviariates we will use in the model (Figure 5).

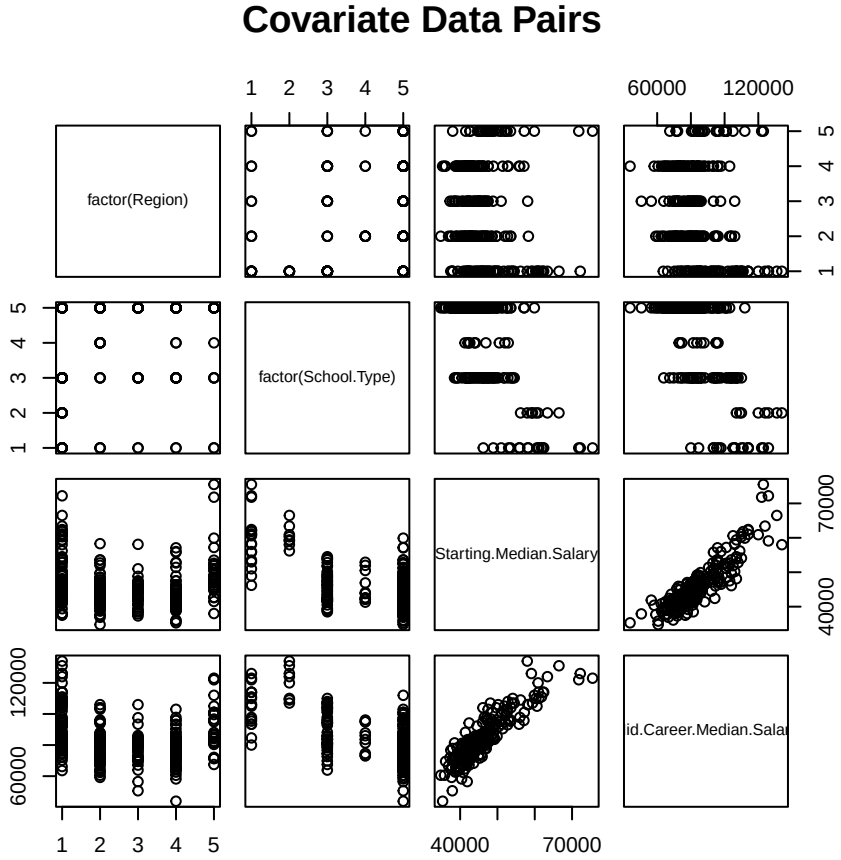

Figure 5: Model Covariates

Now we can fit some linear regression models. In modelling the mid-career median salary we can first start with all three possible coviariates while excluding any interacton terms (Table 2).

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu % Date and time: Tue, Dec 28, 2021 - 15:52:43

Table 2: Regression Results

	Dependent variable:
	Mid.Career.Median.Salary
factor(Region)Southern	−1,430.962
	p = 0.213
factor(Region)Western	−3,503.696
	p = 0.004***
factor(Region)Midwestern	−3,232.025
	p = 0.004***
factor(Region)California	−1,424.498
	p = 0.305
factor(School.Type)Ivy League	11,763.080
	p = 0.00001***
factor(School.Type)Liberal Arts	11,208.490
	p = 0.00000***
factor(School.Type)Party	6,754.526
	p = 0.008***
factor(School.Type)State	4,292.153
	p = 0.024**
Starting.Median.Salary	1.964
	p = 0.000***
Constant	−10,381.120
	p = 0.048**
Observations	248
R^2	0.865
Adjusted R^2	0.860
Residual Std. Error	5,542.006 (df = 238)
F Statistic	169.141*** (df = 9; 238)
Note:	*p<0.1; **p<0.05; ***p<0.01

We see that the p-value for the starting median salary coviarate is by far the smallest (<2.2e-16). This, alongside the high F statistic value indicates we can reject the null hypothesis. There is atleast a linear relationship between starting and mid-career salaries which is unlikely to observed simply by chance. All p-values for the school type coviarate are also within a significance level of 0.05, concluding that this coviariate is statistically significant in the model. In addition, the high value of adjusted R^2 (0.86) shwos that more than 86% of the variance in the data is being explained by the model.

We can generate a couple more models through backwards selection. Let us eliminate the region coviariate to see if this model will fit better. The California and Southern regions are not statistically significant variables so it is worth comparing a model without the region predictor. We can also consider the only coviarate to be the starting median salary since there was the strongest linear relationship between this predictor and the response. We can also ask if there are interactions between any coviarates which might enhance our model. The null hypothesis here is that the extra coefficients in the model with interaction terms are all equal to zero, i.e., interaction terms do not enhance our linear model.

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu % Date and time: Tue, Dec 28, 2021 - 15:52:44

Table 3: Regression Comparison Results

	Dependent variable:		
	Mid.Career.Median.Salary		
	(1)	(2)	(3)
factor(Region)Southern	−1,430.962	−1,466.951	−14,026.060
	p = 0.213	p = 0.196	p = 0.117
factor(Region)Western	−3,503.696	−3,245.215	−20,245.380
	p = 0.004***	p = 0.009***	p = 0.075*
factor(Region)Midwestern	−3,232.025	−2,964.864	−7,172.159
	p = 0.004***	p = 0.008***	p = 0.403
factor(Region)California	−1,424.498	−1,142.184	8,614.791
	p = 0.305	p = 0.412	p = 0.299
factor(School.Type)Ivy League	11,763.080	23,020.350	12,201.600
	p = 0.00001***	p = 0.568	p = 0.00001***
factor(School.Type)Liberal Arts	11,208.490	−27,843.630	11,148.210
	p = 0.00000***	p = 0.040**	p = 0.00000***
factor(School.Type)Party	6,754.526	−12,622.820	6,560.800
	p = 0.008***	p = 0.558	p = 0.010***
factor(School.Type)State	4,292.153	−25,945.190	4,223.389
	p = 0.024**	p = 0.020**	p = 0.027**
Starting.Median.Salary	1.964	1.526	1.924
	p = 0.000***	p = 0.000***	p = 0.000***
factor(School.Type)Ivy League:Starting.Median.Salary		−0.177	
		p = 0.790	
factor(School.Type)Liberal Arts:Starting.Median.Salary		0.722	
		p = 0.006***	
factor(School.Type)Party:Starting.Median.Salary		0.294	
		p = 0.511	
factor(School.Type)State:Starting.Median.Salary		0.533	
		p = 0.007***	
factor(Region)Southern:Starting.Median.Salary			0.282
			p = 0.149
factor(Region)Western:Starting.Median.Salary			0.377
			p = 0.136
factor(Region)Midwestern:Starting.Median.Salary			0.087
			p = 0.647
factor(Region)California:Starting.Median.Salary			−0.198
			p = 0.234
Constant	−10,381.120	15,521.990	−8,439.436
	p = 0.048**	p = 0.133	p = 0.222
Observations	248	248	248
R^2	0.865	0.871	0.869
Adjusted R^2	0.860	0.864	0.862
Residual Std. Error	5,542.006 (df = 238)	5,462.666 (df = 234)	5,497.190 (df = 234)
F Statistic	169.141*** (df = 9; 238)	121.367*** (df = 13; 234)	119.622*** (df = 13; 234)
Note:			*p<0.1; **p<0.05; ***p<0.01

In Table 3 we can see the regression results for the model with no interaction terms, the model with interaction between starting median salary and school type, and the model with interaction between starting median salary and school region. Looking at the p-values, we see that the interaction between starting median salary and school region is not statistically significant. On the other hand there is statistical significance in some of the interactions between the starting median salary and the schools type. The p-values for the interaction of starting salary with Liberal Arts and State school types are below the 0.05 threshold.

Lets take a look at a AIC table for these models (Table 4). In short, the AIC is a method for evaluating how well a model fits the data it was generated from. The smaller the AIC value the better the models fit. We see that all the AIC values are within a narrow range of eachother, but the lowest value is for the model which includes all three covariates as well as the interaction between the median starting salary and the schools type.

Table 4: AIC Results

	df	AIC
mod_reduced	11	4991.162
mod_Sal_Type	7	4995.898
mod_Sal	3	5072.060
mod_full_joint_Sal_Type	15	4987.806
mod_full_joint_Sal_Reg	15	4990.931
mod_full_joint_Type_Reg	21	4992.338
mod_full_joint_all	38	5002.917

As additional confirmation we can use the two-way ANOVA to perform a F-test between nested models (Table 5). Again, the null hypothesis is that the extra coefficients in the model with interactions between the starting median salary and the schools type are all equal to 0, i.e., we favor the reduced model. Note, we already have an idea if which model fits better based on the AIC values above. In confirmation with the AIC results, we see that the p-value associated with the F statistic is approximately 0.03, meaning we can indeed reject the null hypothesis at a significance level of 0.05. In other words, adding the interaction between starting median salary and school type to the linear model leads to a significantly improved fit over the model with no interaction terms.

Table 5: F-Test Between Nested Models

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
238	7309891363	NA	NA	NA	NA
234	6982728602	4	327162761	2.740909	0.0293971

Let us analyze the regression diagnostics for the model with the best AIC, that is the one including all covariates as well as the interactions between the starting median salary and the school type. These diagnostic plots provide checks for heteroscedasticity, normality, and

influential observations (Figure 6).

We do not see any distinctive patterns in the Residuals vs Fitted plot. If there was, say, any sort of non-linear relationship in this plot, that would be indicitive that this relationship was not explained by our linear model and was left out in the residuals.

The Normal Q-Q plot shows that the residuals follow a straight line as desired.

The Scale-Location plot alllows us to check the assumption of heteroscedasticity (equal variance). We see that the residuals are spread out fairly equally along the ranges of predictors which is a good sign.

Lastly the Residuals vs Leverage plot helps us to find influential cases if any. There is a slight left skew and a few outliers, but there are no cases with high Cook's distance scores (i.e., points that fall outside the labeled Cook's distance). This would confirm that there are no outlying cases which are possibly influential against the regression line.

The plot of Cook's distance for our model observations in Figure 7 is another way to observe this. However, if we are being more precise we should consider observations with a Cook's distance over 4/n (n being the total number of data points) to be possible influential outliers. We can see that there are a handful of observations which lie above this threshold (dotted line in plot), which is important to note.

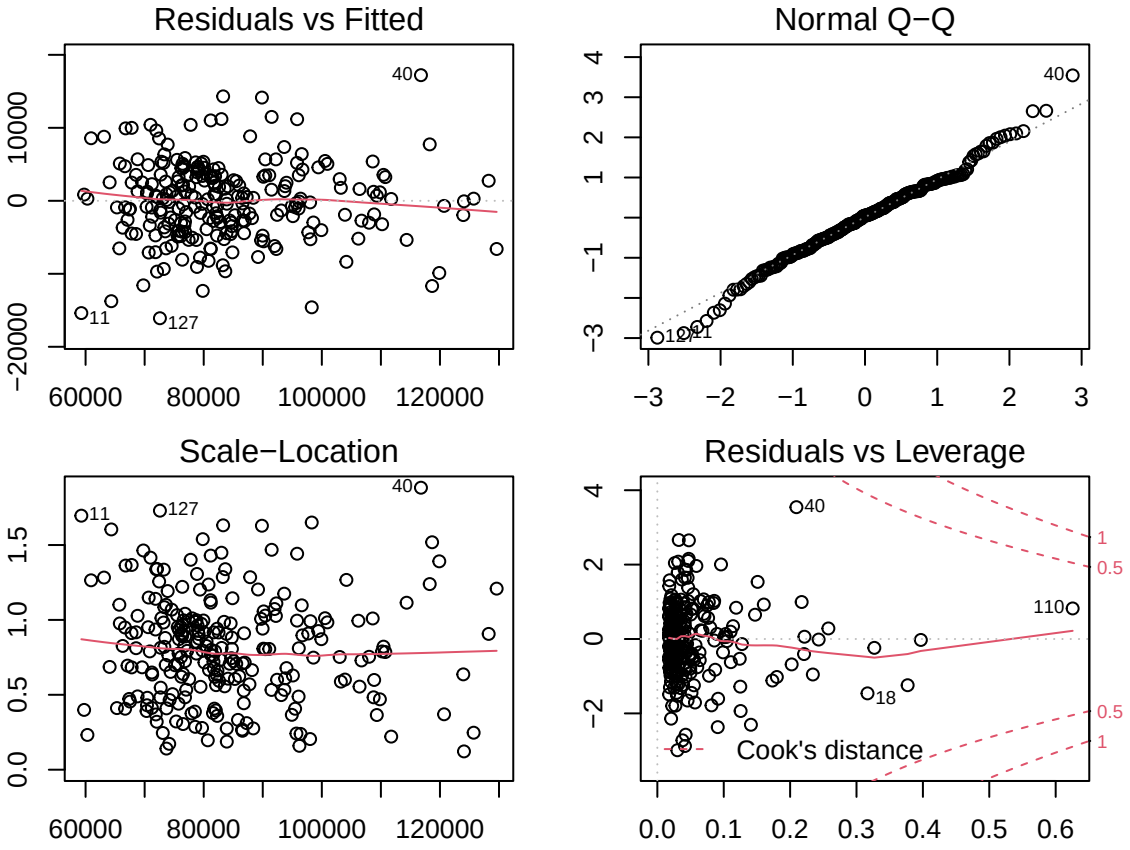

Figure 6: Regression Diagnostics

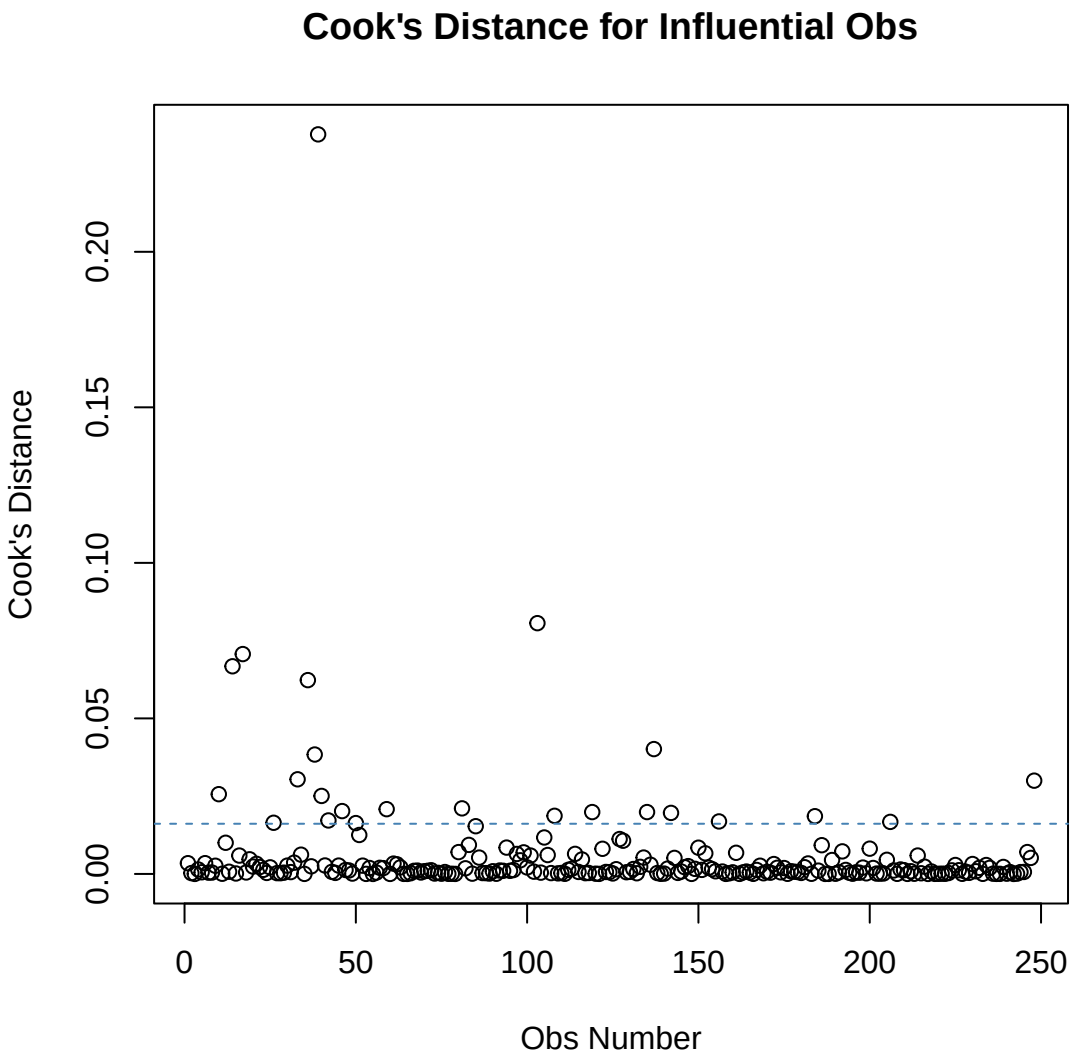

Figure 7: Cooks Distance

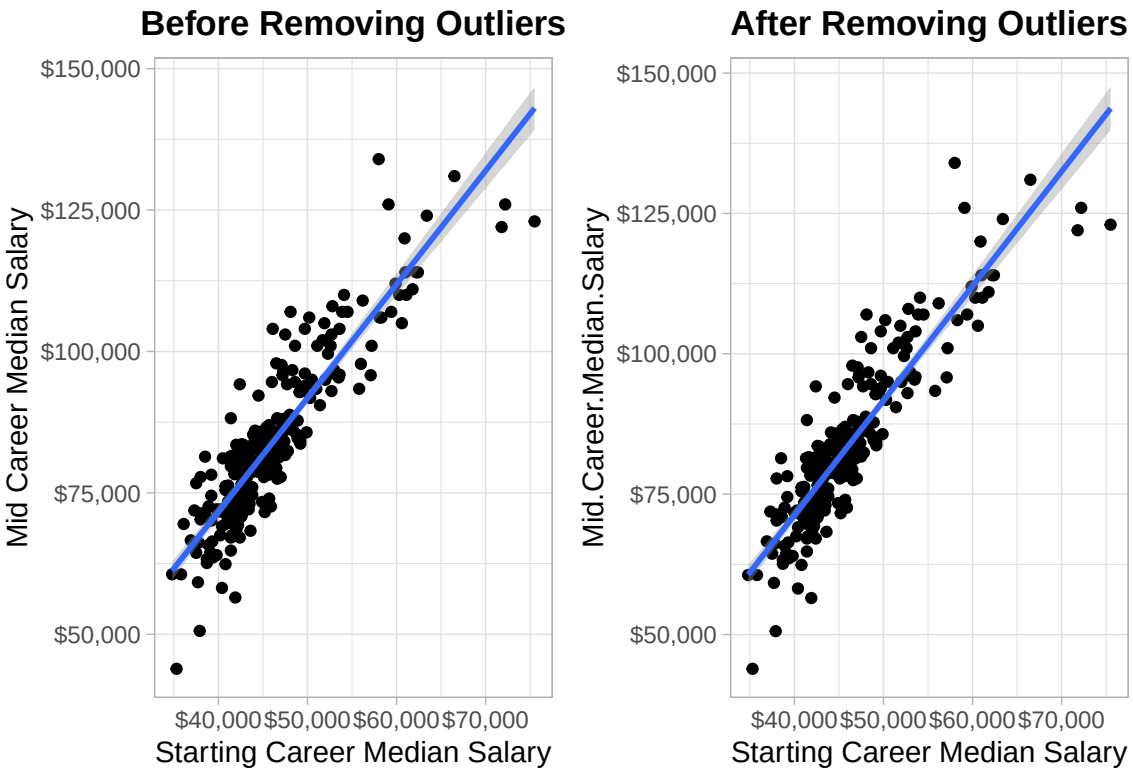

Figure 8: Data and Linear Regression w and w/out Outliers

Let us briefly visualize the difference that removing these outliers has on the linear model. For visualization purposes we can compare the starting and mid-career median salary data with the addition of the most simple linear regression line for y ~ x (i.e mid-career median salary ~ starting median salary). Note that according to the AIC test above this is not model which fits the data best, but it is not too much worse. What we see in Figure 8 is that the linear model changes very minimally after we remove the 15 observations which fall above the Cook's distance cut-off value, 0.016. Although there are some observations which, according to Cook's distance, might be influential outliers, what we see here is that these data seem to minimally affect the well-fitting linear model.

Conclusion

Choosing which college to attend is an important decision for the millions of incoming students every year. One of the many deciding factors is the future salary one might recieve after graduating and finding a job. We found through this analysis that there seems to be a strong linear relationship between the mid-career median salary one will end up recieving, and ones starting median salary as well as the college type and region. Overall, Engineering and Ivy-League schools return the highest mid-career salaries while on the other hand, State and Party schools return the lowest mid-career salaries. In addition, schools in California and the Northeast U.S return the best mid-career salaries when averaging over all the school types. It is also to note that for all the Liberal Arts students out there, colleges in the Southern U.S return the highest starting and mid-career salaries. All in all, if a college student wants the best chance at recieving a high salary later in their career, three important factors to consider are the colleges region, type, and starting salaries after graduation.