Final Project

Adem Hoskin

2022-05-09

Question of Interest

Question: Is there a relationship between average faculty salary (per month) and the tuition revenue made per full time student? (Modeling will be used) Columns: tuition__revenue__per__fte, faculty__salary Why it was chosen: I think it can interesting to see if and how staff salaries are determined based on the revenue of a school. I would guess that the more revenue made per student, the more the staff is paid; however, the relationship could very well be different.

Preprocessing

```
college_reduced <- college%>%
  select(TUITFTE, AVGFACSAL, INSTNM, MAIN)
```

The first two are the two variables we are looking at to answer the question. The other two are the names of the institution and whether they are the main campus or not. (Will be used to to create a faceted scatterplot to see the significance of the category.)

```
college_renamed<-college_reduced %>%
  rename(tuition_revenue_per_fte =TUITFTE, avg_faculty_salary=AVGFACSAL, institution_name=INSTI
```

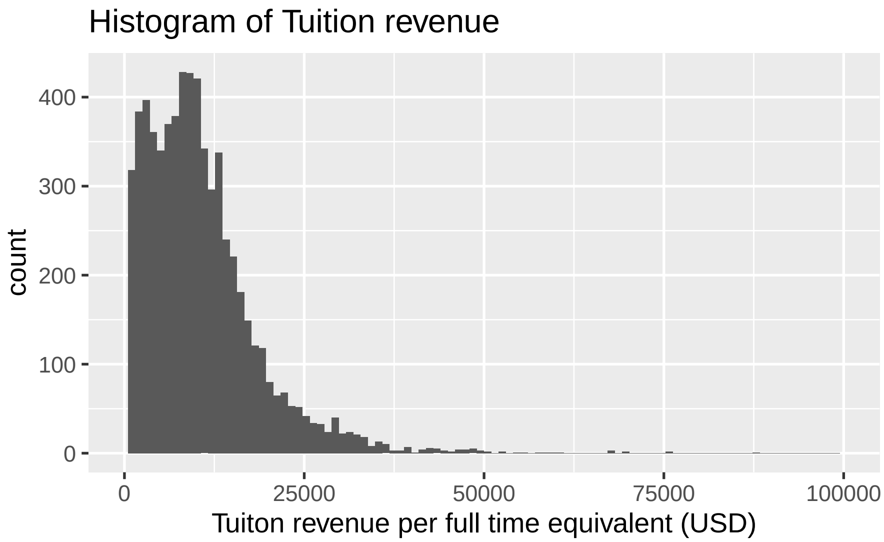
Renaming the columns to these names makes it easiter to understand what the numbers mean.

```
college_recoded <- college_renamed %>%
  mutate(
    main_campus_recoded = recode(
        main_campus,
        `0` = "not a main campus ",
        `1` = "main campus"
    )
  )
```

Makes the main__campus column easier to read. ## Visualization I am first making a histogram of the tuition revenue variable

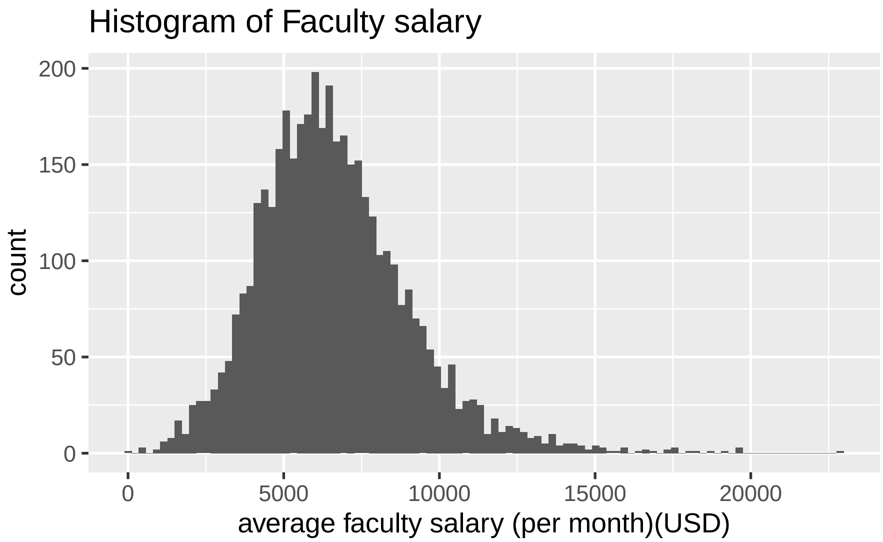
```
college_recoded%>%
  ggplot()+
  geom_histogram(aes(x=tuition_revenue_per_fte), bins=100)+xlim(0,100000)+
   labs(title="Histogram of Tuition revenue", x="Tuiton revenue per full time equivalent (USD)"
```

```
## Warning: Removed 469 rows containing non-finite values (stat_bin).
## Warning: Removed 2 rows containing missing values (geom_bar).
```



```
college_recoded%>%
  ggplot()+
  geom_histogram(aes(x=avg_faculty_salary), bins=100)+
  labs(title="Histogram of Faculty salary", x="average faculty salary (per month)(USD)")
```

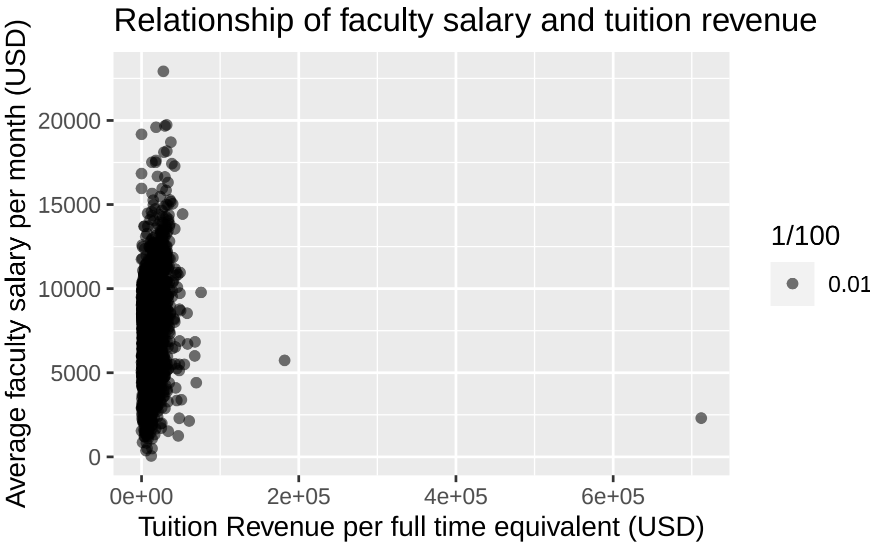
```
## Warning: Removed 2849 rows containing non-finite values (stat_bin).
```



In the first histogram, the distibution of the tuition revenue is very right skewed, whereas the second is slightly rightskewed. (The first one has an xlim in order to see the histogram better)

I am creating a scatter plot with the variables of my question to see any direct relationship.

```
college_recoded%>%
  ggplot() +
  geom_point(mapping = aes(x = tuition_revenue_per_fte, y = avg_faculty_salary, alpha = 1 / 10
  labs(title="Relationship of faculty salary and tuition revenue", x="Tuition Revenue per full
## Warning: Removed 2857 rows containing missing values (geom_point).
```



I have created a scatter plot with the variables of my question. We can see that almost of the entries fall in to a clump with only a couple outliers (Ultimate Medical Academy_Clearwater and SIT Graduate Institute). This falls in line with the histograms.

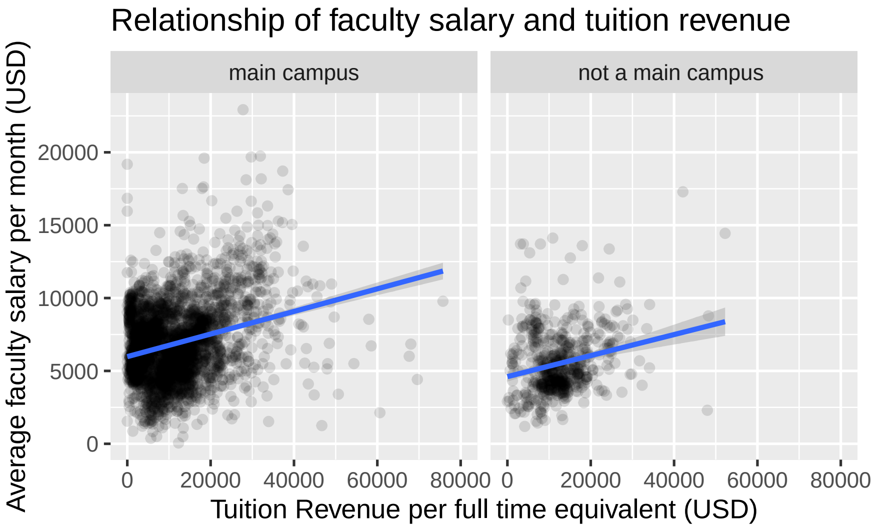
Here, I am creaing a faceted version of the same scatterplot (based on the main campus variable) with a linear model in order to better see if there is a strong linearity with the two variables (also removed the two outliters and added an alpha better see the density of the graph). I also made the graphs more zoomed with alpha to better see the density in points.

```
college_recoded%>%
  filter(institution_name != "Ultimate Medical Academy-Clearwater", institution_name !="SIT Gra
  ggplot(aes(x = tuition_revenue_per_fte, y = avg_faculty_salary))+
  geom_point(alpha=1/8)+
  xlim(0,80000)+
  geom_smooth(method=lm)+
  facet_wrap(vars(main_campus_recoded))+
  labs(title="Relationship of faculty salary and tuition revenue", x="Tuition Revenue per full
```

```
## `geom_smooth()` using formula 'y ~ x'

## Warning: Removed 2857 rows containing non-finite values (stat_smooth).

## Warning: Removed 2857 rows containing missing values (geom_point).
```



The graphs look quite similar however the graph with non-main campus has much less datapoints due to its nature as a secondary campus. This suggests that whether a campus is a main campus or not has little effect on salary. There is a little linearity, however it is very scattered and seems to not suggest a very strong relationship between the two variables.

Overall, these graphs suggest that tuition revenue may not have a great impact on

Summary Statistics

Summary Stats for Tuition revenue:

```
college_recoded%>%
  summarize(count=n(), mean= mean(tuition_revenue_per_fte, na.rm = TRUE), median= median(tuiti
```

count	mean	median	range	standdev	IQR
7058	10704	9148	0	12450.44	8910
7058	10704	9148	712078	12450.44	8910

Summary Statistics for Average Faculty Salary:

```
college_recoded%>%
  summarize(count=n(), mean= mean(avg_faculty_salary, na.rm = TRUE), median= median(avg_faculty
```

count	mean	median	range	standdev	IQR
7058	6631.144	6377	56	2459.03	2979
7058	6631.144	6377	22924	2459.03	2979

The count matches the number of rows (entries) in the dataframe. The mean represenst the average of the variable, median represent the midpoint value of the variable, the range is the difference between the maximum value and minimum value of the variable, the standard deviation represents the deviation of the values of a variable and the IQR represents the middle half of values (25-75 percentiles) in the variable. The standard deviation is noteworthy, whereas tuition's standard deviation is at 12450, salary's is at only 2459.

Data Analysis

I am now making a model based on my two variables to answer my question.

```
college_model <- lm(avg_faculty_salary~tuition_revenue_per_fte, data = college_recoded)
```

Now, I am using the tidy and glance functions to report back coefficents.

```
college_model %>%
  tidy()
```

term	estimate	std.error	statistic	p.value
(Intercept)	6367.7124935	48.2135405	132.07312	0
tuition_revenue_per_fte	0.0233042	0.0026714	8.72347	0

```
college_model %>%
  glance()
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nobs
0.0178005	0.0175666	2433.174	76.09893	0	1	-38714.96	77435.91	77454.94	24859491233	4199	4201

Looking at the estimate, we see the slope is about 233 (signalling the predicted progression of the modelled relationship) and if we look at our R^2, it is at 0.017. What this tells us is that the model doesn't capture the variablity well.

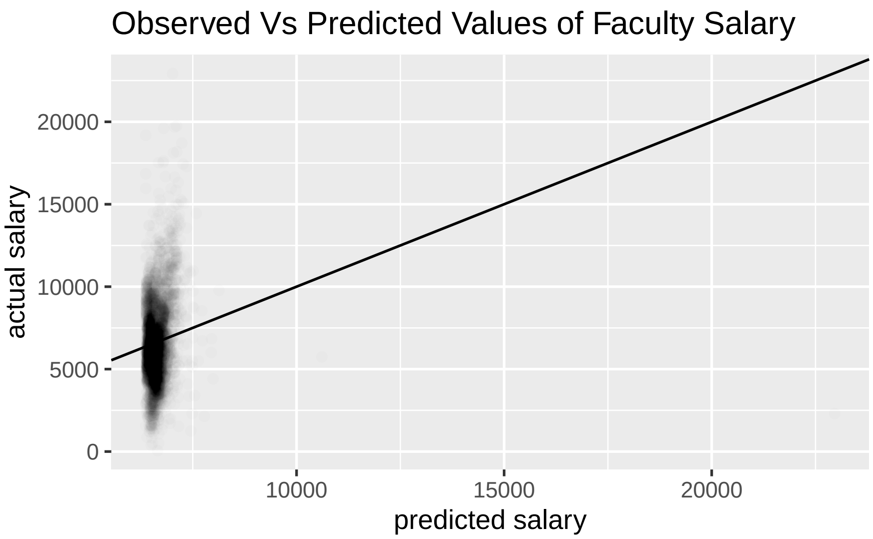
I will now conduct observed vs. predictions and residual vs. predicted plots but first I need to add residuals and predicted values to a data frame.

```
college_df <- college_recoded %>%
add_predictions(college_model) %>%
 add_residuals(college_model)
```

Now that is done, I will create a observed vs. prediction plot (using alpha to better see density).

```
college_df %>%
  ggplot() +
  geom_point(mapping = aes(x = pred, y = avg_faculty_salary), alpha= 1/100) +
  geom_abline(slope = 1, intercept = 0) +
  labs(title= "Observed Vs Predicted Values of Faculty Salary",x="predicted salary", y="actual
```

```
## Warning: Removed 2857 rows containing missing values (geom_point).
```

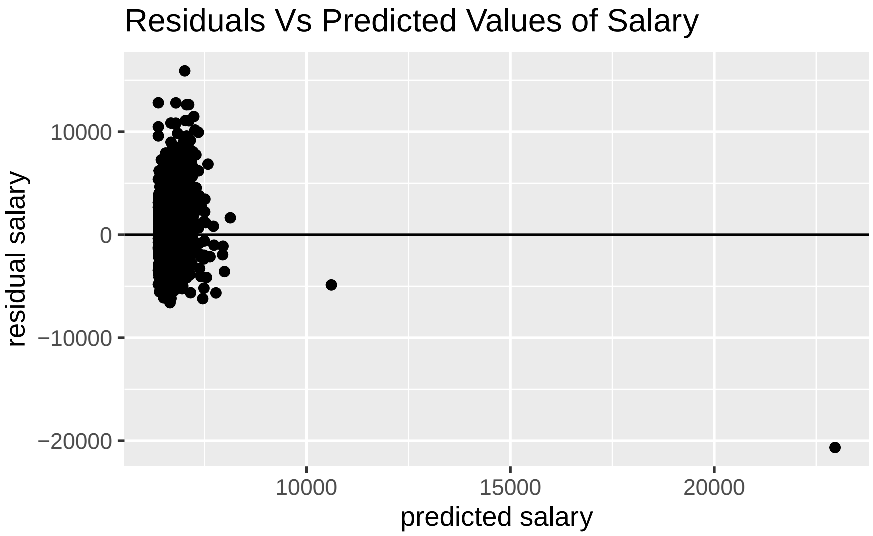


The predicted salary is greatly different from the actual salary being made, with many making less than expected and even more making more than predicted.

Next will be the residual vs predicted plot.

```
college_df %>%
  ggplot() +
  geom_point(mapping = aes(x = pred, y = resid)) +
  geom_abline(slope = 0, intercept = 0) +
  labs(title= "Residuals Vs Predicted Values of Salary",x="predicted salary", y="residual sala
```

```
## Warning: Removed 2857 rows containing missing values (geom_point).
```



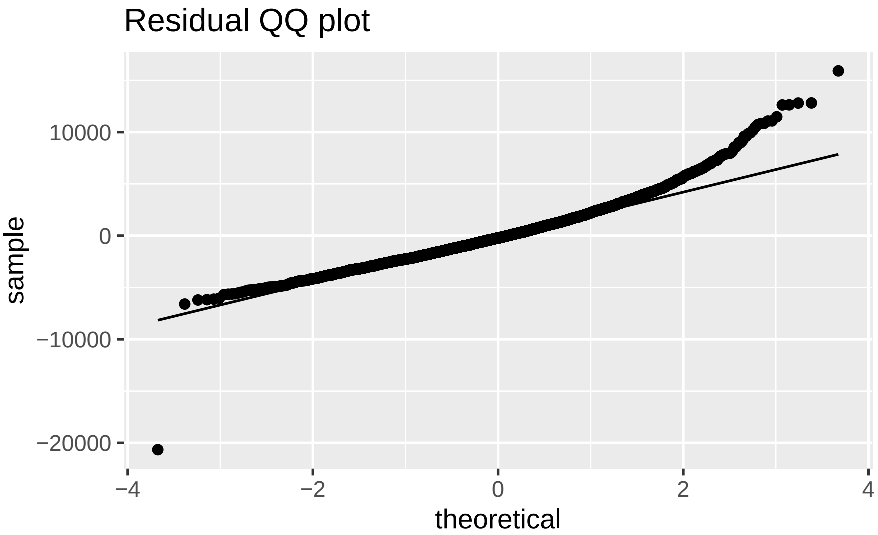
Here, plot does not fit close to the slope, with it being close together and clumped to a small range of salaries, this tells us how little the data fits to the model.

Now will we make a QQ plot

```
college_df %>%
  ggplot() +
  geom_qq(aes(sample = resid)) +
  geom_qq_line(aes(sample = resid))+
  labs(title="Residual QQ plot")
```

```
## Warning: Removed 2857 rows containing non-finite values (stat_qq).
```

```
## Warning: Removed 2857 rows containing non-finite values (stat_qq_line).
```



The QQ plot has points that considerable points that deviate from the line, the light-tailedness of the graph suggesting a gasp in the values.

Conclusion

Taking the information from all the sections, I will say that it does not seem that tuition revenue has a very substantial effect on salaries. It seems that salaries can vary wildly from each other at equal areas of revenue as was seen from the scatterplots. Furthermore, the standard deviations of tuition revenue and salaries contrasted greatly with each other. With the model, the R^2 shows an inability to account for the variability and the plots show how the theoretical values do not line up with the actual values.

With this conclusion, there seems to be other factors not accounted for in my research that likely influence salary more. (funding, location, overall profits that account for costs etc.) This suggests that faculty are not paid based on a percentage of overall revenue, they could rather perhaps be due to market forces of supply and demand.