Final Project

Muhammad Yahya Bin Haroon

2023-12-04

Introduction

How does the average income of the Faculty affect the graduation rate of a College.

i. The response variable is the graduation rate of a College (identified in the dataset as C150_4_POOLED) and the explanatory variable is the average income of the Faculty (identified in the dataset as AVGFACSAL). This is because the average income of the Faculty explains the changes in the graduation rate of a College. To do this, i will be using a Linear Model.

ii. This Question is interesting as many Institutions often cut costs when dealing with their Faculty. However, as the quote goes "A good teacher can inspire hope, ignite the imagination, and instill a love of learning". But a good faculty would also cost more. I intend to see what the statistics have to say on this topic.

Preprocessing

```
college_reduced <- college %>%
  select("UNITID", "INSTNM", "SCHTYPE", "AVGFACSAL", "C150_4_POOLED")
```

i. It is necessary to extract all relevant columns into a new dataset to make calculations easier and clearer. The columns "UNITID", "INSTNM" and "SCHTYPE" are needed to identify and categorize institutions. And the columns "AVGFACSAL" and "C150_4_POOLED" are necessary, as they are the columns that describe the salary_of_faculty and graduation_rate respectively.

```
names(college_reduced) <- c("university_id", "institution_name",
                            "type_of_school", "salary_of_faculty",
                            "graduation_rate")
```

ii. These columns need to be renamed to give clarity and make their usage easier.

```
college_reduced <- college_reduced %>%
  mutate(
    "type_of_school_renamed" = recode(
      type_of_school,
      '1' = "Public",
      '2' = "Private, Nonprofit",
      '3' = "Proprietary"
    )
  )
```

iii. The type_of_school column while dividing the schools into 3 categories, does not tell us what the categories are. Thus by renaming them, we will give more clarity and detail to our calculations and graphs.

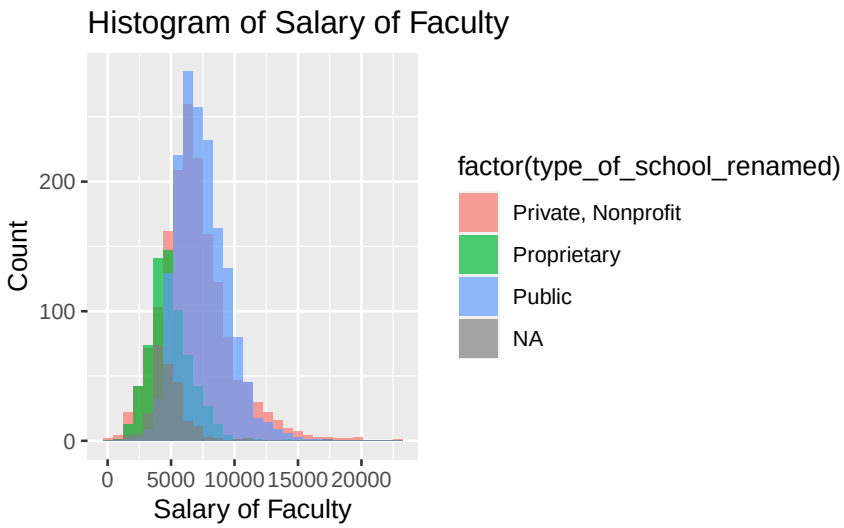
Visualization

i. The histogram of faculty salaries serves the purpose of visualizing the distribution of salaries across different types of schools, such as public, private non-profit, and proprietary institutions. This is particularly important for understanding the salary structure within these institutions and identifying any potential disparities among them.

```
college_reduced %>%
  ggplot() +
  geom_histogram(mapping = aes(x = salary_of_faculty,
                 fill = factor(type_of_school_renamed)), alpha = 0.7,
                 position = "identity") +
  labs(title = "Histogram of Salary of Faculty",
       x = "Salary of Faculty", y = "Count")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

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



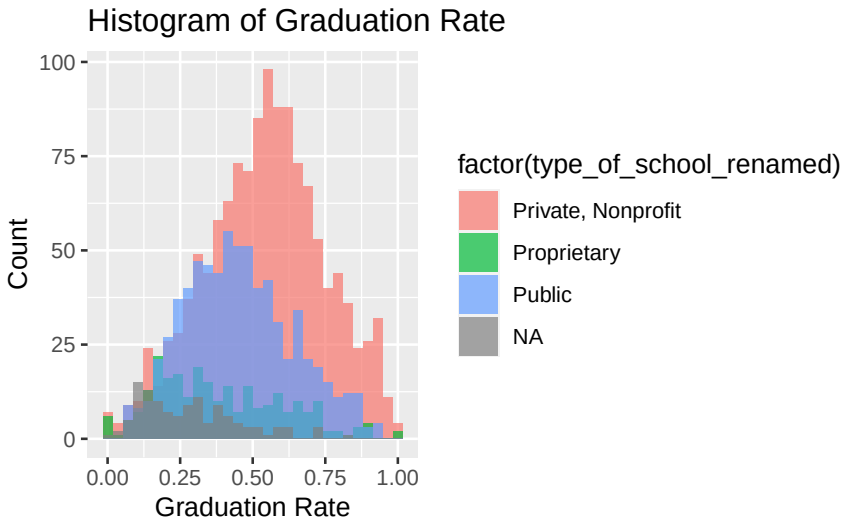
i. The Graphs are all Uni-modal and Right-skewed. There is high variance and high density at the start. From the graph, we can see that overall, Private,Nonprofit schools had the highest salaries, with Public schools coming in second. Proprietary schools come in third with significantly less salaries and NA come in last with the lowest salaries for their faculty.

ii. The histogram of graduation rates is used to illustrate the distribution of these rates across various school types. This visualization is key to understanding how graduation rates vary among different types of educational institutions, providing insights into the outcomes of education and the overall effectiveness of these schools.

```
college_reduced %>%
  ggplot() +
  geom_histogram(mapping = aes(x = graduation_rate,
                 fill = factor(type_of_school_renamed)), alpha = 0.7,
                 position = "identity") +
  labs(title = "Histogram of Graduation Rate",
       x = "Graduation Rate", y = "Count")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 4688 rows containing non-finite values (`stat_bin()`).
```

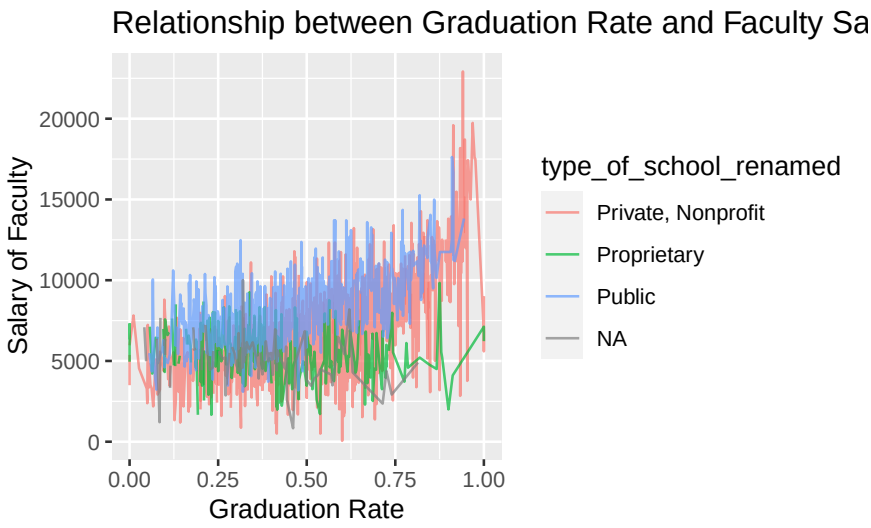


ii. The Graphs for Private,Nonprofit and Public schools are Uni-modal and Symmetric. There is high variance and high density in their middles. the Graphs for Proprietary Schools and NA are multi-modal and right-skewed. There is high variance and high density in their starts. From the graph, we can see that overall, Private,Nonprofit schools had the highest Graduation Rates, with Proprietary schools coming in as a close second. Public schools come in third with slightly a less Graduation Rates and NA come in last with lower Graduation Rates.

iii. The line chart depicting the relationship between graduation rates and faculty salaries, its purpose is to highlight trends or patterns, segregated by school type.

```
ggplot(data = college_reduced, aes(x = graduation_rate,
       y = salary_of_faculty, color = type_of_school_renamed)) +
  geom_line(alpha = 0.7) +
  labs(title = "Relationship between Graduation Rate and Faculty Salary",
       x = "Graduation Rate", y = "Salary of Faculty")
```

```
## Warning: Removed 4689 rows containing missing values (‘geom_line()‘).
```



iii. It can be seen from this line chart, that Salaries of Faculty and Graduation Rates have different relationships according to the type of school. For Private, Nonprofit schools, there is a strong positive relationship, with it being observed that higher salaries for the faculty results in higher graduation rates. For Public schools, there is a slightly less strong positive relationship, with it being observed again that higher salaries for the faculty results in higher graduation rates. However, for proprietary schools and NA, it is observed that the opposite is true. Both graphs show that, there is a slightly negative relationship with it being observed that lower salaries for the faculty results in higher graduation rates.

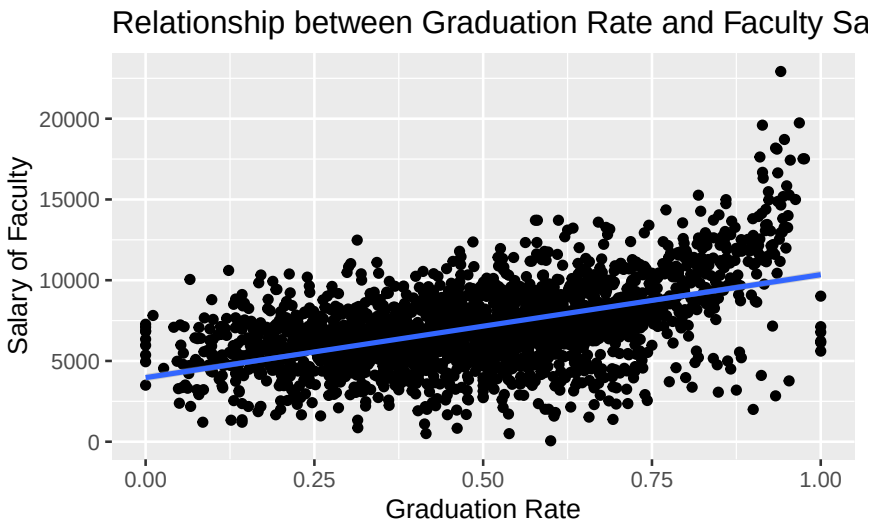
iv. The scatter plot with a linear regression line is designed to show the relationship between graduation rates and faculty salaries. This type of visualization is particularly appropriate for examining the correlation between these two continuous variables. The inclusion of a regression line is beneficial as it helps in understanding the general direction and strength of the relationship between graduation rates and faculty salaries. This approach provides a clear visual representation of any correlations that may exist.

```
college_reduced %>%
  ggplot(mapping = aes(x = graduation_rate, y = salary_of_faculty)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Relationship between Graduation Rate and Faculty Salary",
       x = "Graduation Rate", y = "Salary of Faculty")
```

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

```
## Warning: Removed 4762 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 4762 rows containing missing values (`geom_point()`).
```



iv. It can be seen from this scatter plot, that Salaries of Faculty and Graduation Rates have a positive linear relationship. While the gradient of the line of best-fit is not very high, we can see the general trend that institutions with higher Salaries for their Faculty have higher Graduation Rates.

Summary Statistics

```
college_reduced %>%
  group_by(type_of_school_renamed) %>%
  summarize(count = n())
```

type_of_school_renamed	count
Private, Nonprofit	1946
Proprietary	2540
Public	2066
NA	506

```
college_reduced %>%
  group_by(type_of_school_renamed) %>%
  summarize(
    count = n(),
    mean_gr = mean(graduation_rate, na.rm = TRUE),
    median_gr = median(graduation_rate, na.rm = TRUE),
    range_gr = max(graduation_rate, na.rm = TRUE)
    - min(graduation_rate, na.rm = TRUE),
    sd_gr = sd(graduation_rate, na.rm = TRUE),
    iqr_gr = IQR(graduation_rate, na.rm = TRUE)
  )
```

type_of_school_renamed	count	mean_gr	median_gr	range_gr	sd_gr	iqr_gr
Private, Nonprofit	1946	0.5411319	0.55225	1.0000	0.2060758	0.27550
Proprietary	2540	0.3897055	0.35165	1.0000	0.2192923	0.34345
Public	2066	0.4525291	0.43960	0.8879	0.1870922	0.26360
NA	506	0.2813555	0.26310	0.8139	0.1720242	0.24590

```
college_reduced %>%
  group_by(type_of_school_renamed) %>%
  summarize(
    count = n(),
    mean_salary = mean(salary_of_faculty, na.rm = TRUE),
    median_salary = median(salary_of_faculty, na.rm = TRUE),
    range_salary = max(salary_of_faculty, na.rm = TRUE)
    - min(salary_of_faculty, na.rm = TRUE),
    sd_salary = sd(salary_of_faculty, na.rm = TRUE),
    iqr_salary = IQR(salary_of_faculty, na.rm = TRUE)
  )
```

type_of_school_renamed	count	mean_salary	median_salary	range_salary	sd_salary	iqr_salary
Private, Nonprofit	1946	6877.952	6525.0	22868	2757.341	3043.0
Proprietary	2540	4871.344	4731.5	13506	1639.626	2031.0
Public	2066	7375.501	7136.5	16009	2002.998	2609.5
NA	506	4815.094	4513.0	16456	1597.583	1411.0

i. For Private, Nonprofit institutions, their mean salary is 6877.952 and the mean graduation rate is 0.54. The median Salary is 6525.0 and the median graduation rate is 0.55. The high mean and median salary is resulting in a high mean and median graduation rate.

ii. For Public institutions, their mean salary is 7375.501 and the mean graduation rate is 0.45. The median Salary is 7136.5 and the median graduation rate is 0.44. The high mean and median salary is resulting in a high mean and median graduation rate.

iii. For Proprietary institutions, their mean salary is 4871.3442 and the mean graduation rate is 0.39. The median Salary is 4731.5 and the median graduation rate is 0.35. The low mean and median salary is resulting in a low mean and median graduation rate.

iv. For NA, their mean salary is 4815.094 and the mean graduation rate is 0.28. The median Salary is 4731.5 and the median graduation rate is 0.26. The lowest mean and median salary is resulting in the lowest mean and median graduation rate.

Data Analysis

```
model_cr <- lm(data = college_reduced, graduation_rate ~ salary_of_faculty)
```

i. This line of code is used to create a linear regression model where graduation rate is being predicted based on salaries of the faculty.

```
model_cr%>%
  tidy()
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.1892260	0.0110032	17.19739	0
salary_of_faculty	0.0000427	0.0000015	29.27458	0

ii. This tidy format makes it easier to understand and report the results of the linear model. The estimated effects, p value, standard error, and their statistical significance, can easily be seen all in a clean, tabular form. The model suggests that there is a positive relationship between faculty salary and Graduation Rate, as faculty salary increases, the response variable also increases, with each additional unit of salary corresponding to an increase in Graduation Rate. The small p-values for both terms suggest that these findings are statistically significant.

```
model_cr%>%
  glance()
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nobs
0.2719774	0.2716601	0.177452	857.0012	0	1	713.0277	-1420.055	-1402.839	72.23628	2294	2296

iii. This glance function is needed to get a quick overview of the model's goodness-of-fit and other essential diagnostic measures. These statistics suggest that the model explains about 27.2% (r.squared value) of the variance in Graduation Rate. The relatively low sigma suggests a good fit. However, the high deviance and BIC, and low logLik and AIC indicate that this model is not a very good fit to explain the graduation rate.

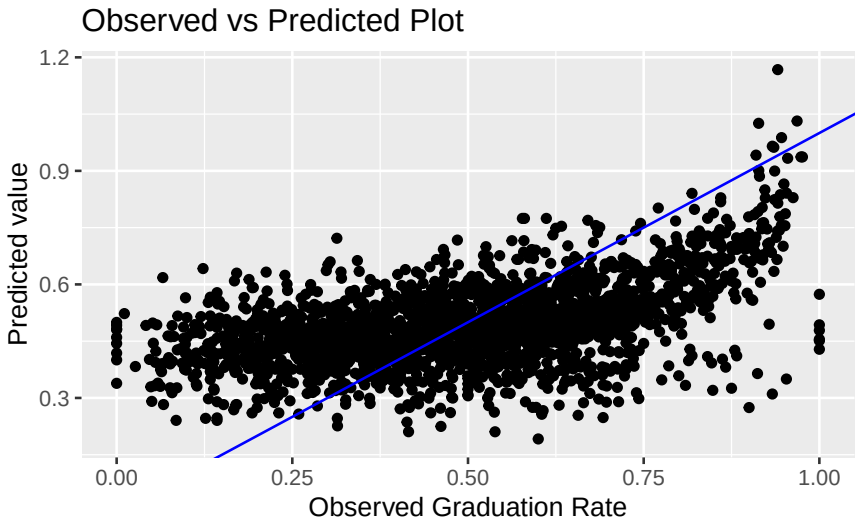
```
college_df <- college_reduced %>%
add_predictions(model_cr) %>%
add_residuals(model_cr)
```

iv. This code block is needed to create a new data frame from the college_reduced data frame which will in addition to the previous columns, hold the residual and prediction values that will be used in the following Data Analyses.

Observed vs. Predicted

```
college_df%>%
  ggplot() +
  geom_point(mapping = aes(x = graduation_rate , y = pred)) +
  geom_abline(slope = 1, intercept = 0, color = "blue") +
  labs(
    title = "Observed vs Predicted Plot",
    x = "Observed Graduation Rate",
    y = "Predicted value"
    )
```

```
## Warning: Removed 4762 rows containing missing values (‘geom_point()‘).
```

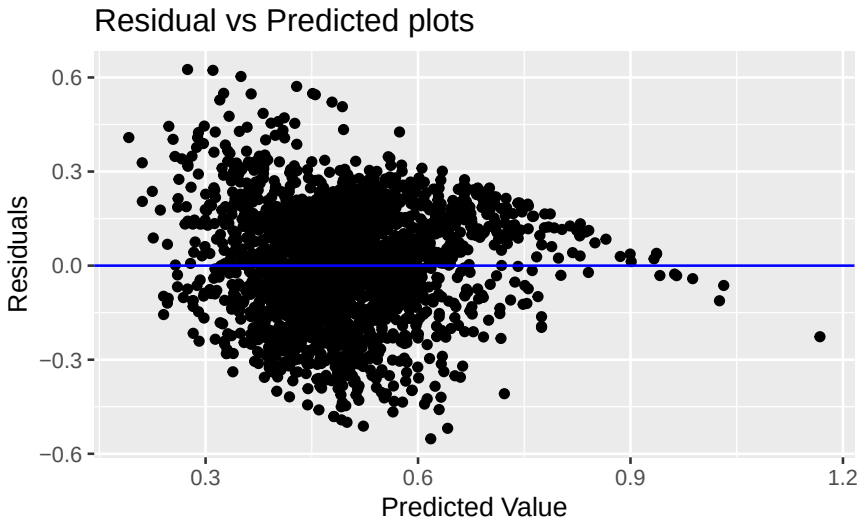


v. There are some points that are clustering around the 45 degree line, however, there is a significant number of clustering away from it. There are a significant number of overestimations and underestimations. These clusters deviating from the line suggest that suggest that this model is not a very good fit in predicting the graduation rate.

Residual vs. Predicted

```
college_df %>%
  ggplot() +
  geom_point(mapping = aes(x = pred, y = resid)) +
  geom_hline(yintercept = 0, color = "blue") +
  labs(
    title = "Residual vs Predicted plots",
    x = "Predicted Value",
    y ="Residuals"
    )
```

```
## Warning: Removed 4762 rows containing missing values (‘geom_point()‘).
```



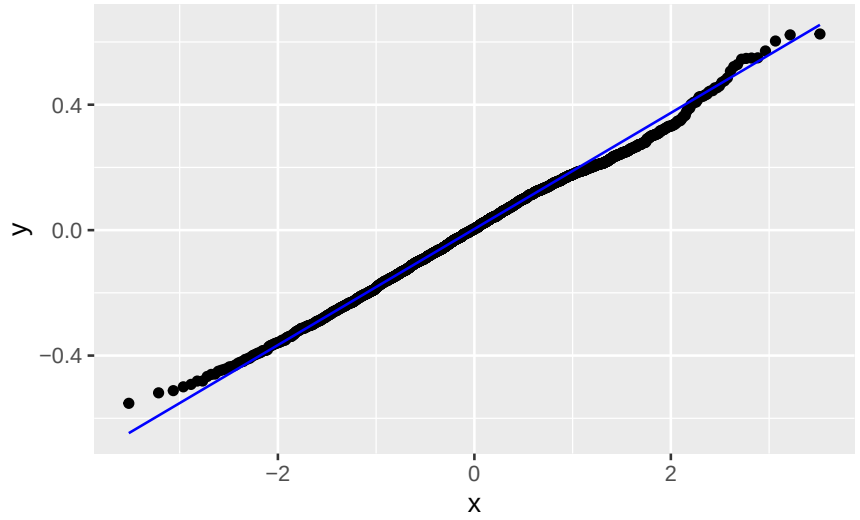
vi. In my opinion, the model does not meet the assumption of constant variability. The plots on the graph are not evenly distributed throughout the chart. The plots are not following an equal trend along the intercept even when the slope is not rising. A funnel shape is being observed in this plot, meaning there is heteroscedasticity.

Q-Q Plot

```
college_df %>%
ggplot() +
geom_qq(aes(sample = resid)) +
geom_qq_line(aes(sample = resid), color = "blue")
```

```
## Warning: Removed 4762 rows containing non-finite values (‘stat_qq()‘).
```

```
## Warning: Removed 4762 rows containing non-finite values (‘stat_qq_line()‘).
```



vii. Almost all of the plots a falling onto the straight line. However, there is a slight s-shaped curve, meaning the tails of my data are just slightly heavier than a normal distribution. Overall, for the most part, as almost all of the plots a falling onto the straight line, the slight deviations can be ignored. Thus it can be concluded that the residuals of a regression model are normally distributed.

Conclusion

i. Based on these analyses, it can be concluded that the graduation rate of an institution is somewhat significantly influenced by the salary of the faculty. With an R-squared value of 0.27, this indicates that approximately 27% of the variation in the graduation rates of institutions can be explained by variations in faculty salaries.

ii. In terms of my original question: How does the average income of the Faculty affect the graduation rate of a College, the overall trend shows that yes, the average income of the Faculty does affect the graduation rate of a College. It is seen that in Private, Nonprofit and public institutions, a higher income faculty results in a higher graduation rate. However, in Proprietary institutions and NA, the opposite is seen, that a lower income faculty results in a higher graduation rate.

iii. The analyses from the different sections support each other, as the r-square value, p-value, summary statistics and all the different graphical representations all pointed towards the same conclusion.

iv. From these Analyses, it is observed that faculty salaries have a notable impact on college graduation rates. The findings imply that investing in faculty through competitive salaries can potentially improve student outcomes. However, the distinct trends in different types of institutions highlight the need for further research to understand the unique factors at play. These insights are crucial for policymakers and educational administrators, as they underscore the significance of faculty compensation in educational strategies.