Final Project

Kiana Johnson

2023-05-09

Introduction

The question I'm exploring is: "Is there a relationship between the percentage of first-generation students and the graduation rate of each college?" The two columns I will be using are "PAR_ED_PCT_1STGEN" and "C100_4." The "PAR_ED_PCT_1STGEN" column will be the explanatory variable, while the "C100_4" will be the response variable. Because both of these columns contain continuous variables and values, I will be utilizing a linear model to answer my question. This question is interesting because first-generation students are typically characterized as having more grit and a good work ethic. I would like to see how the data reflects this sentiment through the success of these students, which I would define with a degree. I think this will be very insightful and confirm whether this is true.

Preprocessing

In this first block of code, I am piping the original college dataset and selecting the specific columns I'll be working with for this project. It includes, not only my two chosen variables, but also other columns of interest that I'll be using for my data analysis. I'm assigning it to a placeholding variable, "college_reduced1" for the time being, so I can work on my next step.

```
college_reduced1 <- college %>%
  select(PAR_ED_PCT_1STGEN,
         C100_4,
         DISTANCEONLY,
         POVERTY_RATE,
         REGION)
```

In this second block of code I'm piping my previously coded dataset variable, "college_reduced1", and renaming the columns so that they are easier to keep track of and understand. I'm assigning this new "renamed" dataset to another placeholding variable, "college_reduced2", to prepare for my next step.

```
college_reduced2 <- college_reduced1 %>%
  rename(
    "share_firstgeneration" = PAR_ED_PCT_1STGEN,
    "completion_rate_4yr_100nt" = C100_4,
    "distance_only" = DISTANCEONLY,
    "poverty_rate" = POVERTY_RATE,
    "region" = REGION
    )
```

In this third code chunk, I am piping the the "recoded" data variable, "college_reduced2", into the "mutate" function, where I recode the values in one of the two categorical variables I've decided to explore, "region." I'm using the "recode" function instead of the "rename" function for these variables because the values under the column also need to be tweaked. I assign this "recoded" dataset to my final placeholding variable, "college_reduced3", for my final preprocessing step.

```
college_reduced3 <- college_reduced2 %>%
  mutate(
   region = recode(
       region,
       `1` = "New England",
       `2` = "Mid East",
       `3` = "Great Lakes",
       `4` = "Plains",
       `5` = "Southeast",
       `6` = "Southwest",
       `7` = "Rocky Mountains",
       `8` = "Far West",
       `9` = "Outlying Areas",
   )
   )
```

In this fourth code chunk, I am piping the "region recoded" into the "mutate" function, yet again, to recode the "distance_only" variable, just as I did before. I am finally assigning this to the complete variable, "college_reduced", I will be referring to for the remainder of my project.

```
college_reduced <- college_reduced3 %>%
  mutate(
    distance_only = recode(
      distance_only,
      `1` = "Distance Education Only",
      `0` = "Not Distance Education Only"
    )
  )
```

In this final code chunk, I am using the "head" function to see the first 6 rows of my "college_reduced", to verify that the necessary changes I made were successful.

```
head(college_reduced) %>%
  select(share_firstgeneration, completion_rate_4yr_100nt, poverty_rate, region)
```

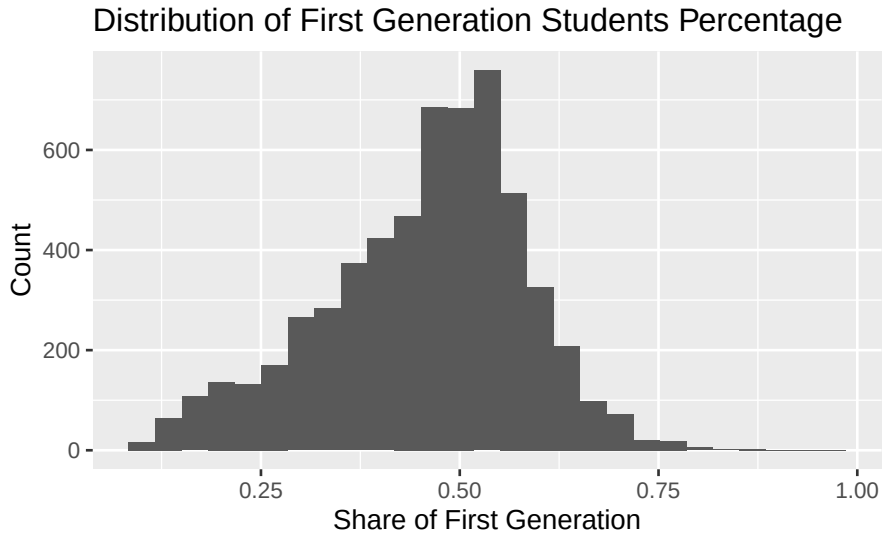
share_firstgeneration	completion_rate_4yr_100nt	poverty_rate	region
0.3658281	0.0381	14.88	Southeast
0.3412237	0.3179	10.91	Southeast
0.5125000	0.0000	10.65	Southeast
0.3101322	0.2270	9.37	Southeast
0.3434343	0.1129	16.96	Southeast
0.2257127	0.4404	10.05	Southeast

Visualization

I am now going to explore the answer to "Question 1: What type of variation occurs within my variables?"

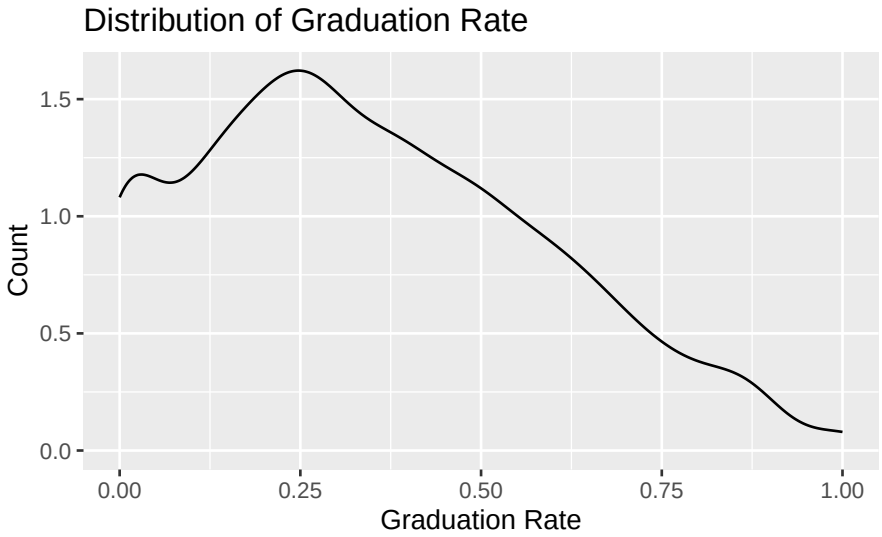
In this first chunk of code, I am creating a histogram representing the distribution of the percentage of first generation students, "share_firstgeneration". I'm graphing this in order to understand this distribution individually to see how it exists outside of other variables. The distribution of this graph is unimodal and normal, but slightly left-skewed with a center value of about 0.52.

```
college_reduced %>%
  ggplot() +
  geom_histogram(mapping = aes(x = share_firstgeneration), bins = 27) +
  labs(title = "Distribution of First Generation Students Percentage",
       x = "Share of First Generation",
       y = "Count")
```



In this second chunk of code, I am creating a density plot representing the distribution of the graduation rate, "completion_rate_4yr_100nt". I'm using this graph to learn more about the variable's variation and how it could possibly impact other variables. This density plot's distribution is slightly bimodal, but right-skewed. It has a center value at 0.25.

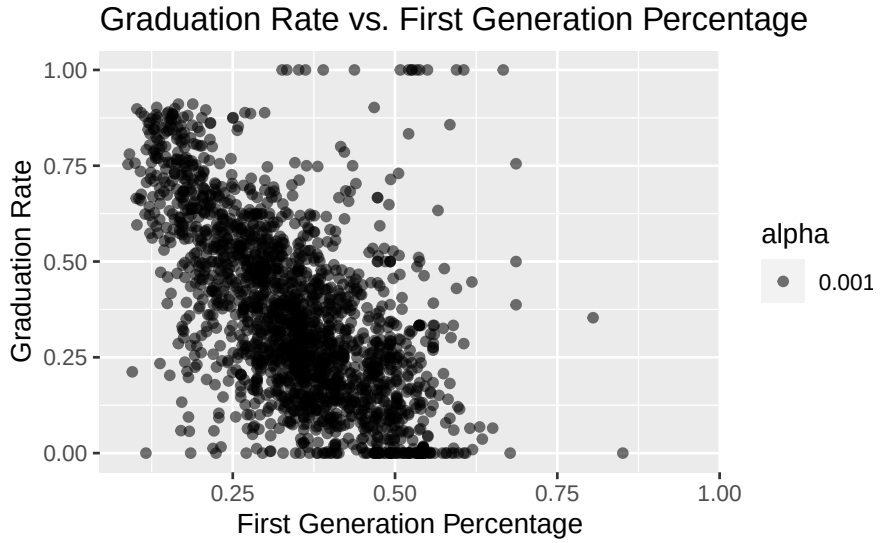
```
college_reduced %>%
  ggplot() +
  geom_density(mapping = aes(x = completion_rate_4yr_100nt)) +
  labs(title = "Distribution of Graduation Rate",
       x = "Graduation Rate",
       y = "Count")
```



Now, I am exploring "Question 2: What type of covariation occurs between my variables?"

In this first chunk of code, I'm graphing a scatterplot in order to model the covariation between both of my continuous variables, being first generation percentage and the graduation rate percentage. The graph shows that there seems to be a direct negative linear relationship between the two because of their high degree of correlation. While there are some poorly correlated areas, it is clear that as the first generation percentage increases, the graduation rate decreases.

```
#ab line
college_reduced %>%
  ggplot() +
  geom_point(mapping = aes(x = share_firstgeneration,
                           y = completion_rate_4yr_100nt, alpha = 0.001)) +
  labs(title = "Graduation Rate vs. First Generation Percentage",
       x = "First Generation Percentage",
       y = "Graduation Rate")
```

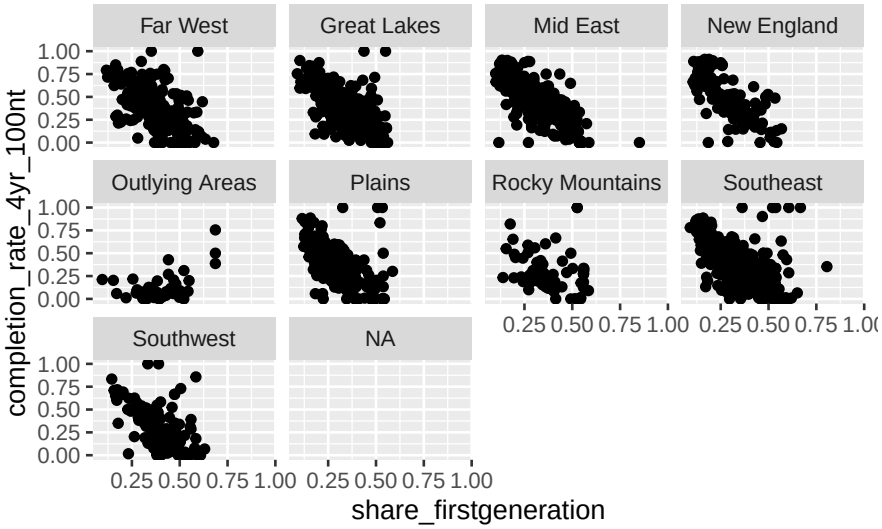


Now, while I've answered the first two questions, I now want to explore potential reasons as to why

this negative linear relationship exists between my two variables.

In the following code chunk, I am faceting over the "region" variable to see if there is potentially any affect of location on this relationship. Specifically, I can observe that the highest density of schools are located in the "Great Lakes", "Southeast", "Mideast", and the "Far West". However, there does not seem to be a notable difference in how these graphs behave, besides the "Outlying Areas". I can hypothesize that as the area with the lowest density of schools, the "Outlying Areas" facet does not have enough data to compare to the other graphs.

```
college_reduced %>%
  ggplot() +
  geom_point(mapping = aes(x = share_firstgeneration,
                           y = completion_rate_4yr_100nt)) +
  facet_wrap(~ region)
```



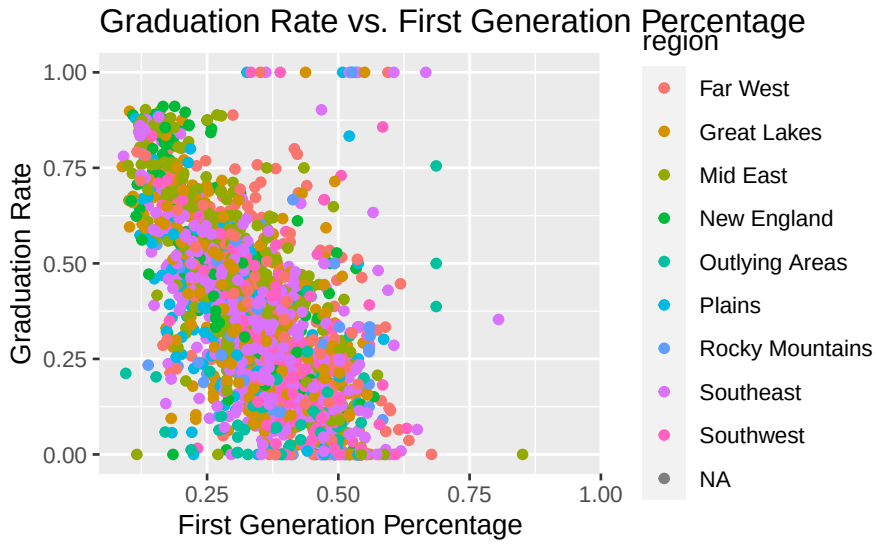

```
  labs(title = "Graduation Rate vs. First Generation Percentage",
       x = "First Generation Percentage",
       y = "Graduation Rate")
```

```
## $x
## [1] "First Generation Percentage"
##
## $y
## [1] "Graduation Rate"
##
## $title
## [1] "Graduation Rate vs. First Generation Percentage"
##
## attr(,"class")
## [1] "labels"
```

This next code chunk serves to verify my findings from the first graph. Here, I decided to color the complete "graduation rate vs. first generation percentage" graph by the "region" variable. There is no distinct color I can identify in a jumble of roughly evenly mixed ones, so this visual confirms

that there is no specific relationship with "region" and my negative linear relationship.

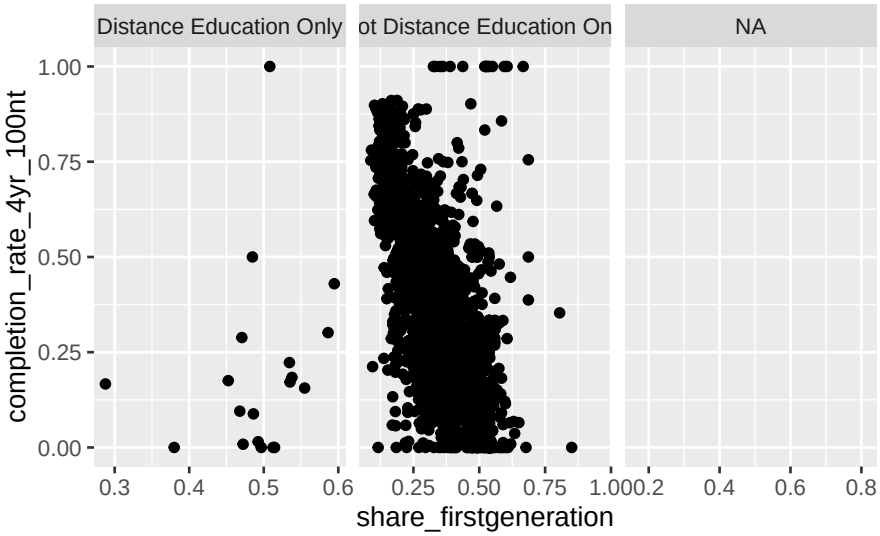
```
college_reduced %>%
  ggplot() +
  geom_point(mapping = aes(x = share_firstgeneration,
                           y = completion_rate_4yr_100nt,
                           color = region)) +
  labs(title = "Graduation Rate vs. First Generation Percentage",
       x = "First Generation Percentage",
       y = "Graduation Rate")
```



Now, since the "region" variable came out inconclusive, I am going to explore whether or not distance education plays a role into this negative linear relationship between my two original variables.

In this next code chunk, I am facetting the original scatterplot over the "distance_only" variable. Here, I find that there aren't enough distance education only schools available in this database. Therefore, there is no pattern or trend to observe. Perhaps, if there were more distance education schools in the dataset, I would find something different. In the meantime, this is another dead end.

```
college_reduced %>%
  ggplot() +
  geom_point(mapping = aes(x = share_firstgeneration,
                           y = completion_rate_4yr_100nt)) +
  facet_wrap(~ distance_only, scales = "free_x")
```




```
  labs(title = "Graduation Rate vs. First Generation Percentage",
       x = "First Generation Percentage",
       y = "Graduation Rate")
```

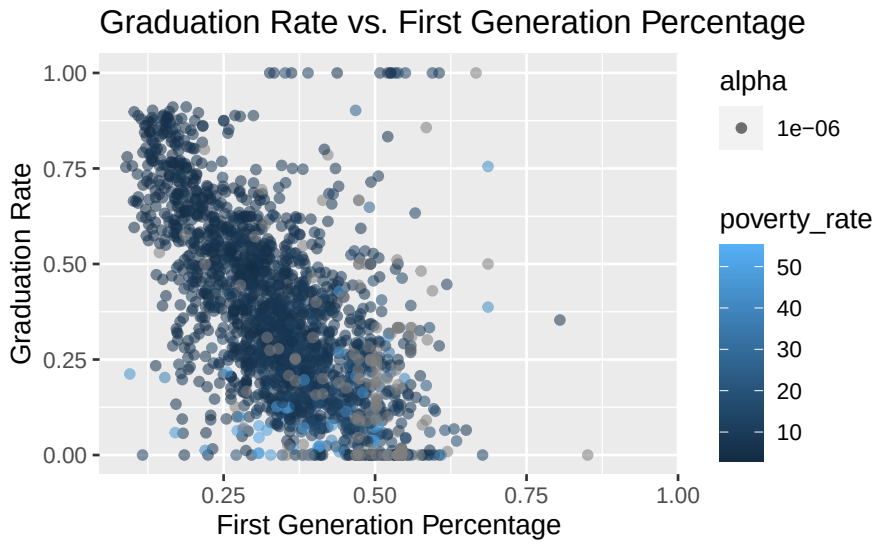
```
## $x
## [1] "First Generation Percentage"
##
## $y
## [1] "Graduation Rate"
##
## $title
## [1] "Graduation Rate vs. First Generation Percentage"
##
## attr(,"class")
## [1] "labels"
```

Now, I am going to explore the potential influence of poverty rate on this negative linear relationship between my two variables.

In this next code chunk, I am using the "color" function to add a third continuous variable to my original "graduation rate vs. first generation percentage" graph. With this, I can see that as the graduation rate decreases, the poverty rate seems to grow higher. While the vast majority of these schools are both low poverty or N/A values, the pattern is still clear. I also notice that as the first generation percentage increases, there is an increase in poverty rate. This leads me to believe that there is a relationship between these three variables, and that poverty rate plays a role in this negative linear relationship.

```
college_reduced %>%
  ggplot() +
  geom_point(mapping = aes(x = share_firstgeneration,
                           y = completion_rate_4yr_100nt,
                           alpha = 0.000001,
                           color = poverty_rate)) +
  labs(title = "Graduation Rate vs. First Generation Percentage",
```

```
       x = "First Generation Percentage",
       y = "Graduation Rate")
```



Summary Statistics

Now, I will be generating my summary statistics for both of my variables, "share_firstgeneration" and "completion_rate_4yr_100nt".

This first code chunk demonstrates the count of observations, mean, median, minimum, maximum, standard deviation, range and interquartile range of the "share_firstgeneration" variable. Since these are both continuous variables, there is no need to group_by anything.

```
college_reduced %>%
  summarize(
    n = n(),
    mean = mean(share_firstgeneration, na.rm = TRUE),
    median = median(share_firstgeneration, na.rm = TRUE),
    min = min(share_firstgeneration, na.rm = TRUE),
    max = max(share_firstgeneration, na.rm = TRUE),
    std = sd(share_firstgeneration, na.rm = TRUE),
    range = range(share_firstgeneration, na.rm = TRUE),
    iqr = IQR(share_firstgeneration, na.rm = TRUE)
    )
```

n	mean	median	min	max	std	range	iqr
7058	0.4557885	0.47609	0.08867	0.957265	0.1267972	0.088670	0.1670384
7058	0.4557885	0.47609	0.08867	0.957265	0.1267972	0.957265	0.1670384

This second code chunk demonstrates the count of observations, mean, median, minimum, maximum, standard deviation, range and interquartile range of the "completion_rate_4yr_100nt" variable. Since these are both continuous variables, there is no need to group_by anything.

```r
college_reduced %>%
  summarize(
    n = n(),
    mean = mean(completion_rate_4yr_100nt, na.rm = TRUE),
    median = median(completion_rate_4yr_100nt, na.rm = TRUE),
    min = min(completion_rate_4yr_100nt, na.rm = TRUE),
    max = max(completion_rate_4yr_100nt, na.rm = TRUE),
    std = sd(completion_rate_4yr_100nt, na.rm = TRUE),
    range = range(completion_rate_4yr_100nt, na.rm = TRUE),
    iqr = IQR(completion_rate_4yr_100nt, na.rm = TRUE)
    )
```

n	mean	median	min	max	std	range	iqr
7058	0.3499199	0.32	0	1	0.2398309	0	0.3519
7058	0.3499199	0.32	0	1	0.2398309	1	0.3519

Data Analysis

In this first code chunk, I am creating my linear model by the name of "fg_cr_model" with the response variable (completion_rate_4yr_100nt) and the explanatory variable (share_firstgeneration) to start off my data analysis.

```r
fg_cr_model <- lm(completion_rate_4yr_100nt ~ share_firstgeneration, data = college_reduced)
```

In these next two code chunks, I am going to assess my model's coefficients and performance using both the "tidy" and "glance" functions. These give me the slope, the intercept, and the R-Squared value. The "share_firstgeneration" variable's slope value is -1.2041320. The intercept value is 0.7761905. The R squared value the "fg_cr_model" yields is 0.3682816, meaning that roughly 37% of the variation between these two variables is represented. This means that this model is just "okay" in variance.

```r
fg_cr_model %>%
  tidy()
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.7761905	0.0133143	58.29757	0
share_firstgeneration	-1.2041320	0.0360381	-33.41278	0

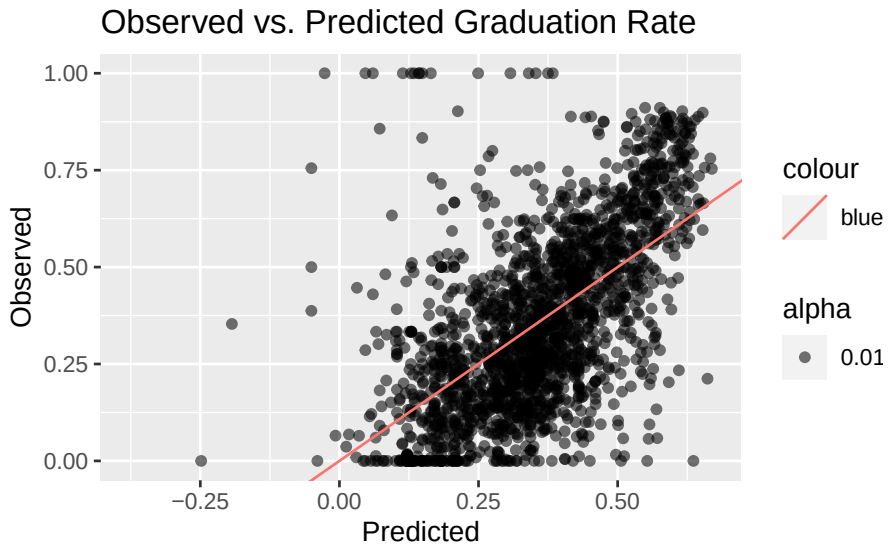
```r
fg_cr_model %>%
  glance() %>%
  select(r.squared)
```

r.squared
0.3682816

In these next two code chunks, I'm testing how well the assumptions of the linear model, "fg_cr_model", are met using an "observed vs. predicted plot." To achieve this, I'm using both the "add_predictions" and "add_residuals" functions to analyze them. Next, I'm using a scatterplot to graph the actual observed "completion_rate_4yr_100nt" values and the predictions we just generated. In doing so, I'm plotting an abline with a slope of 1 and an intercept of zero. This is because a prediction's aim is to be as close to the original as possible. While there is some poor correlation and stray points in some places, the points seem to follow the line generally.

```
fg_cr_df <- college_reduced %>%
  add_predictions(fg_cr_model) %>%
  add_residuals(fg_cr_model)
```

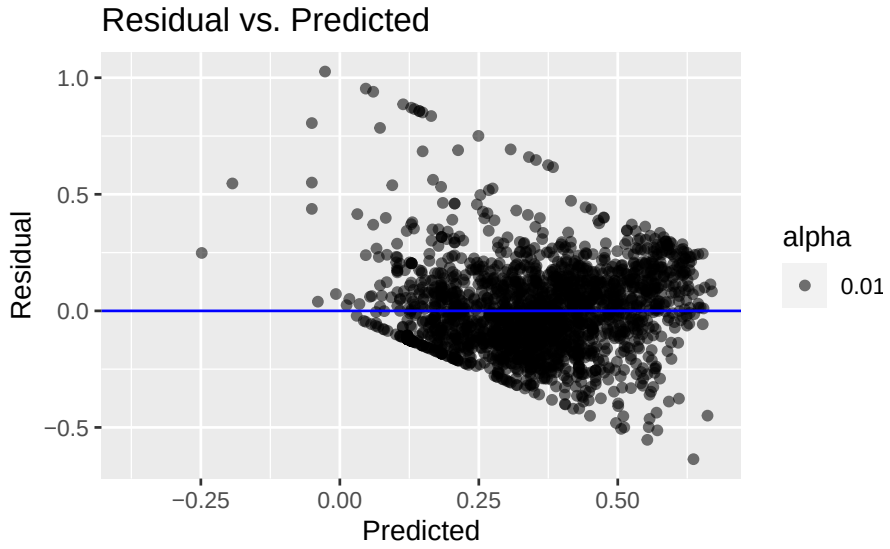
```
fg_cr_df %>%
  ggplot() +
  geom_point(mapping = aes(x = pred,
                           y = completion_rate_4yr_100nt,
                           alpha = 0.01)) +
  geom_abline(mapping = aes(slope = 1,
                            intercept = 0,
                            col = "blue")) +
  labs(title = "Observed vs. Predicted Graduation Rate",
       x = "Predicted",
       y = "Observed")
```



Using the residual and predicted values we generated in the previous part, this code chunk will be creating a "predicted vs. residual" plot. This plot tells me that the linear model has somewhat of a constant variability in the residuals. While there is poor correlation in some areas, the points seem to follow the hline I generated as well.

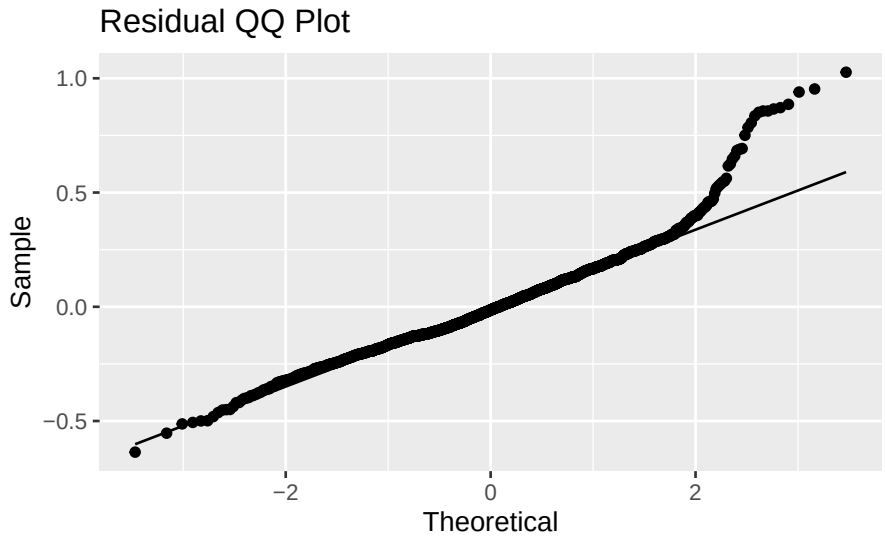
```
fg_cr_df %>%
  ggplot() +
  geom_point(mapping = aes(x = pred, y = resid, alpha = 0.01)) +
  geom_hline(yintercept = 0, col = "blue") +
  labs(title = "Residual vs. Predicted",
```

```
       x = "Predicted", y = "Residual")
```



In this last code chunk, I will be creating a QQ Plot. This graph essentially tells me where the residuals of my data would be normally distributed. The linear model is normally distributed for the most part. While there are some values that stray away from the qq line, the majority are directly on it or within close range.

```
fg_cr_df %>%
  ggplot() +
  geom_qq(aes(sample = resid)) +
  geom_qq_line(aes(sample = resid)) +
  labs(title = "Residual QQ Plot",
       x = "Theoretical",
       y = "Sample")
```



Conclusion

An overall conclusion I can draw from my analyses is that there are many factors that go into this dataset. Exploring the complicated relationship between American colleges and the reasons for their statistics can lead us to many places. With over 2,000 columns to choose from, the nitty gritty details are so important to pay attention to and acknowledge. Each of these variables have an affect on each other, in some way or another, but some are more apparent than others. While determining that is a huge task within itself, using common sense is a great way to start off and figure out which values could potentially have a relation to each other. In my case specifically, I was able to conclude that there does seem to be a strong covariation and relationship between the percentage of first generation students ("share_firstgeneration") and the graduation rate ("completion_rate_4yr_100nt") in U.S. colleges. However, it was not in the way I expected. I originally wanted to explore the idea of first generation students being more motivated, but within the first few blocks of code, my own bias had been proven wrong. While there is indeed a relationship between both my response and explanatory variables, it was a negative linear relationship, while my original thoughts assumed a positive one. As the percentage of first generation students increased, the graduation rate decreased. The graphs and different sections of analysis demonstrated this relationship clearly, validating the existence of the relationship. While the linear model I generated yielded an explanation for a rounded 37% of the data, it is still considered "okay" and viable. I think that this could be because of the many N/A values in the dataset, so a lot is unaccounted for.

However, I did find some potential confounding factors that could be important to consider when delving into the negative linear relationship I found. I looked through a couple different variables including region, distance education, and poverty rate. I found that there is no correlation with region and that there is not enough data for distance education. However, I did find that the poverty rate of the schools played a role in the decreased graduation rate, along with first generation student percentage. This opens the door to a conversation of wealth inequality in the United States. The data that I was able to sift through and gather my findings from showed me that this is a poverty-related issue. While grit and hardwork goes a long way, someone's financial situation may set them many paces behind. This can perhaps shed some light on things to focus on and how to combat against systems that unfairly set others behind others financially and in education. While a solution for working towards equal opportunity and resources would certainly need more in-depth analysis, the issue of poverty is definitely a good place to start.