320Final

Gabe Lavarte

5/21/2019

Question

I present you, the viewer of this article, a question: Which state has the highest rate of completing college in under 4 years for low income students (students whose family has made less than 30k for that year.), and do the schools with higher tuition costs actually graduate more low-income students on time?

Data Collection

We of course want to answer our question with some evidence, so we will need data for educational institutions in the U.S to eventually run some data analysis. I am providing the link here: https://collegescorecard.ed.gov/data/ for a quick download of the Folder containing all the CSV files called CollegeScorecard_Raw_Data. First you load the csv file called MERGED2015_16_PP, which we will be using for this tutorial, from the folder you just downloaded. You do this with a call to the load() function which will result in the data frame: MERGED2015_16_PP. Data frames can also be scraped from html tables of information on a website, but for simplicity, we can just use the provided College Scorecard Raw Data. If the join function doesn't work, add the dataframe to your environment by going to the files and selecting it there to import.

Below we start cleaning up the dataframe we just loaded by only selecting the columns, or attributes, specific to the question we are trying to answer here, which in this case will be the information of the insitutions and the data in relation to low income degrees percentages. After, we need to change the types of these attributes so that we can later run some Exploratory Data Analysis. Let's change the type of the low income rates to factors and change the NPT41_PUB, public school prices, type to an integer. I finish my tidy by removing empty values and replacing them with na, followed by a quick renaming of some of the column names for clarity in reading.

```
college_df <- MERGED2015_16_PP %>%
  select(LO_INC_COMP_ORIG_YR2_RT, LO_INC_COMP_ORIG_YR3_RT,LO_INC_COMP_ORIG_YR4_RT,
         INSTNM, CITY, STABBR, UNITID, NPT41_PUB, NPT41_PRIV)

college_df$LO_INC_COMP_ORIG_YR2_RT <- as.numeric(as.character(college_df$LO_INC_COMP_ORIG_YR2_RT))
```

```
## Warning: NAs introduced by coercion
```

```
college_df$LO_INC_COMP_ORIG_YR3_RT <- as.numeric(as.character(college_df$LO_INC_COMP_ORIG_YR3_RT))
```

```
## Warning: NAs introduced by coercion
```

```
college_df$LO_INC_COMP_ORIG_YR4_RT <- as.numeric(as.character(college_df$LO_INC_COMP_ORIG_YR4_RT))
```

```
## Warning: NAs introduced by coercion
```

```
college_df$NPT41_PUB <- as.numeric(college_df$NPT41_PUB)
college_df$NPT41_PRIV <- as.numeric(college_df$NPT41_PRIV)


#These following 6 lines of code change the null or abscent values from the data frame and converts the
college_df$LO_INC_COMP_ORIG_YR2_RT <- na_if(college_df$LO_INC_COMP_ORIG_YR2_RT, "PrivacySuppressed")
college_df$LO_INC_COMP_ORIG_YR3_RT <- na_if(college_df$LO_INC_COMP_ORIG_YR3_RT, "PrivacySuppressed")
```

```r
college_df$LO_INC_COMP_ORIG_YR4_RT <- na_if(college_df$LO_INC_COMP_ORIG_YR4_RT, "PrivacySuppressed")

college_df$LO_INC_COMP_ORIG_YR2_RT <- na_if(college_df$LO_INC_COMP_ORIG_YR2_RT, "NULL")
college_df$LO_INC_COMP_ORIG_YR3_RT <- na_if(college_df$LO_INC_COMP_ORIG_YR3_RT, "NULL")
college_df$LO_INC_COMP_ORIG_YR4_RT <- na_if(college_df$LO_INC_COMP_ORIG_YR4_RT, "NULL")

#Here I rename some of the columns for clarity
college_df <- college_df %>%
  rename(STATE = STABBR ,
         INSTITUTION = INSTNM,
         AVG_COST_PUB = NPT41_PUB,
         AVG_COST_PRIV = NPT41_PRIV)


# if all three columns are NA remove entity or boths costs are na remove

# Here we remove irrelvent states to the question I am  trying to answer which is
# in relation to the U.S 50 states.

college_df <- college_df %>%
  filter(STATE != "PW", STATE != "FM", STATE != "AS", STATE != "MP", STATE != "PR", STATE != "UM",
         STATE != "VI", STATE != "MH")

college_df <- as.data.frame(college_df)

head(college_df)
```

```
##   LO_INC_COMP_ORIG_YR2_RT LO_INC_COMP_ORIG_YR3_RT LO_INC_COMP_ORIG_YR4_RT
## 1              0.03225807              0.07246377              0.16874136
## 2              0.15739130              0.30489731              0.40150880
## 3                      NA              0.07386364              0.08965517
## 4              0.19075144              0.30879713              0.37478109
## 5              0.04138514              0.10015060              0.17095588
## 6              0.14306677              0.33912439              0.40700809
##                           INSTITUTION       CITY STATE UNITID AVG_COST_PUB
## 1            Alabama A & M University     Normal    AL 100654          315
## 2 University of Alabama at Birmingham Birmingham    AL 100663          345
## 3                  Amridge University Montgomery    AL 100690         1780
## 4 University of Alabama in Huntsville Huntsville    AL 100706          428
## 5            Alabama State University Montgomery    AL 100724          340
## 6          The University of Alabama Tuscaloosa    AL 100751          482
##   AVG_COST_PRIV
## 1          4072
## 2          4072
## 3          3889
## 4          4072
## 5          4072
## 6          4072
```

Exploratory Data Analysis

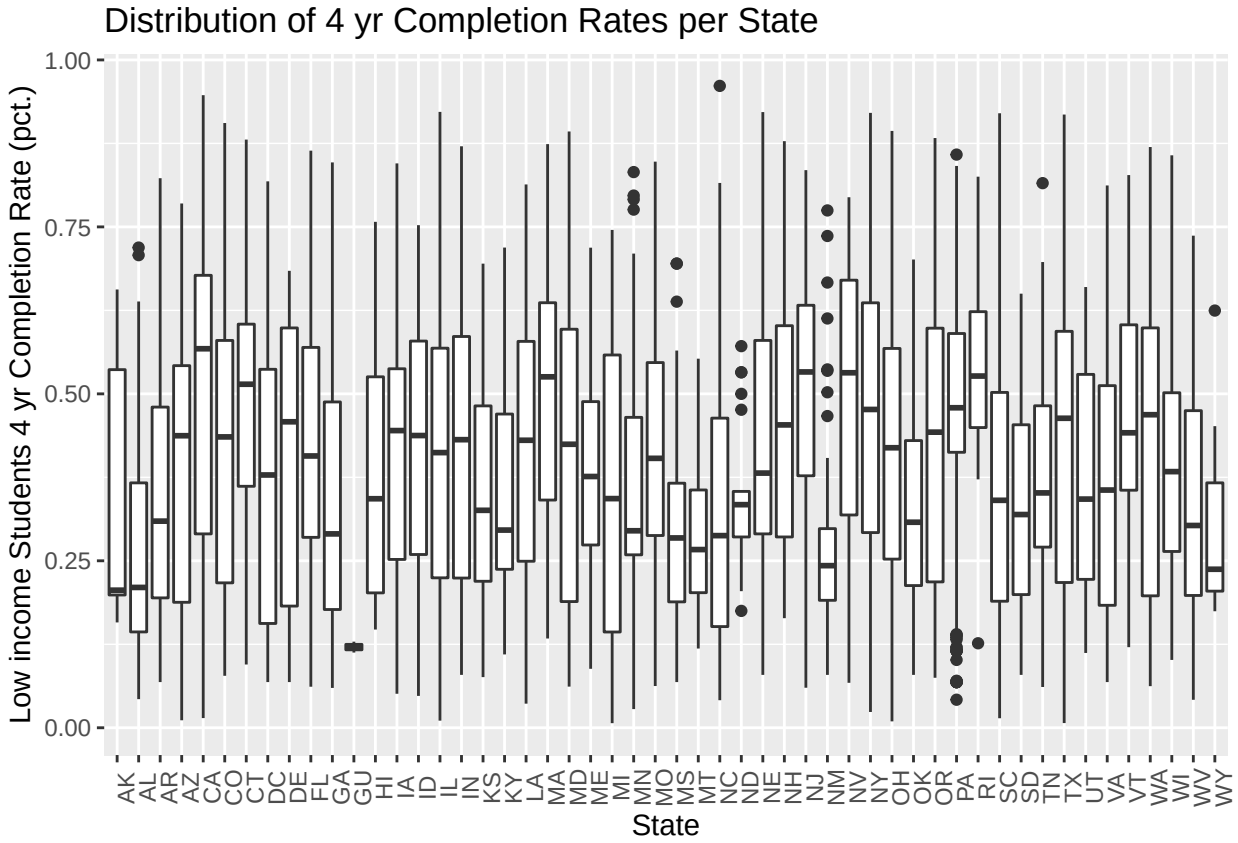
Now that we have our data set tidy, we have to do some Exploratory Data Analysis to see what we can learn from the table we have now before moving onto building the actual data model to test and answer our statistical question.

We can learn many things about the attribute properties like: - the central trends, including our means and medians. - spread, including our variance. - skew - and outliers.

let's quickly graph box plots of the distribution of completion rates across each state so that we can make distinctions on the range of the data, the min and max rates of completion. We also can see which states maintain outliers in their completion rates.

```
college_df %>%
  group_by(STATE) %>%
  ggplot(mapping = aes(x = STATE, y=LO_INC_COMP_ORIG_YR4_RT)) +
  geom_boxplot() +
  labs(title = "Distribution of 4 yr Completion Rates per State",
       x = "State",
       y = "Low income Students 4 yr Completion Rate (pct.)") +
  theme(axis.text.x = element_text(angle = 90))
```

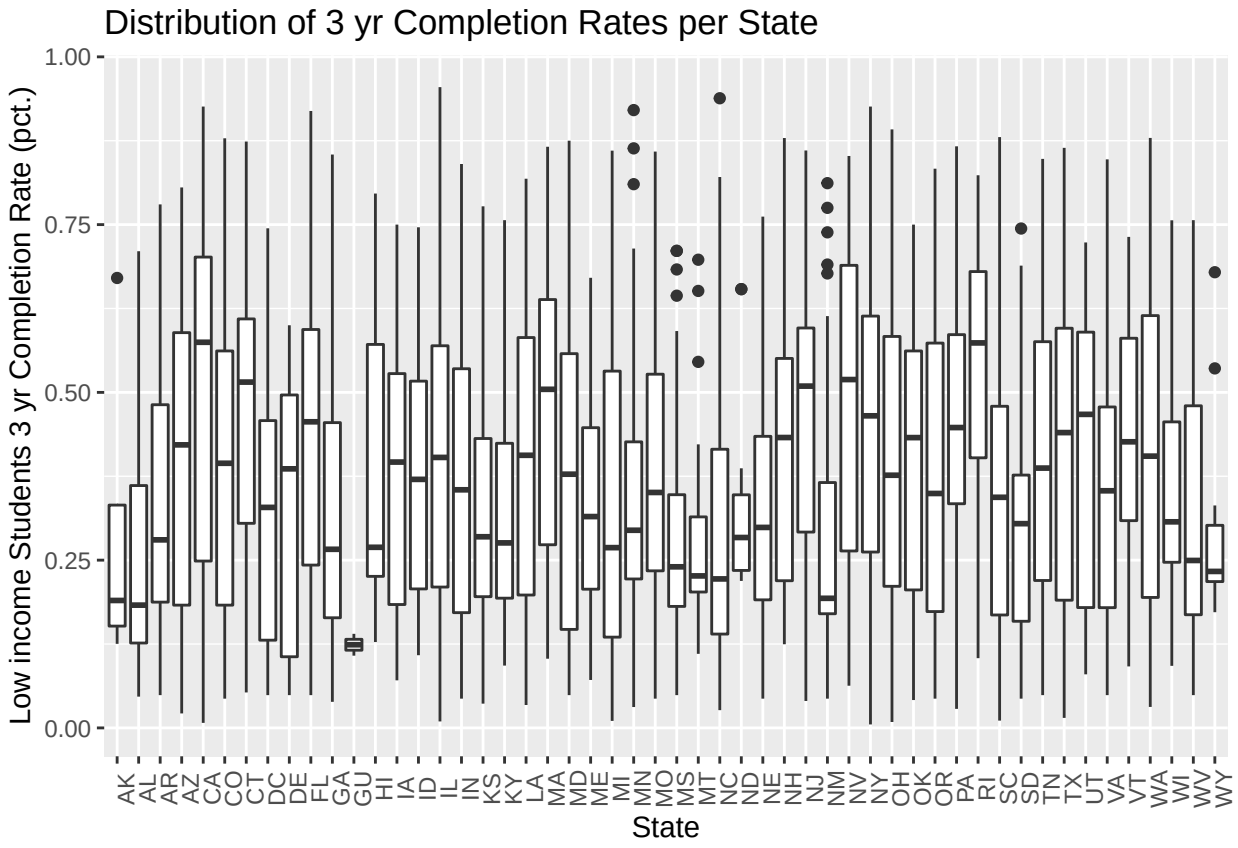
```
## Warning: Removed 1857 rows containing non-finite values (stat_boxplot).
```




```
college_df %>%
  group_by(STATE) %>%
  ggplot(mapping = aes(x = STATE, y=LO_INC_COMP_ORIG_YR3_RT)) +
  geom_boxplot() +
  labs(title = "Distribution of 3 yr Completion Rates per State",
```

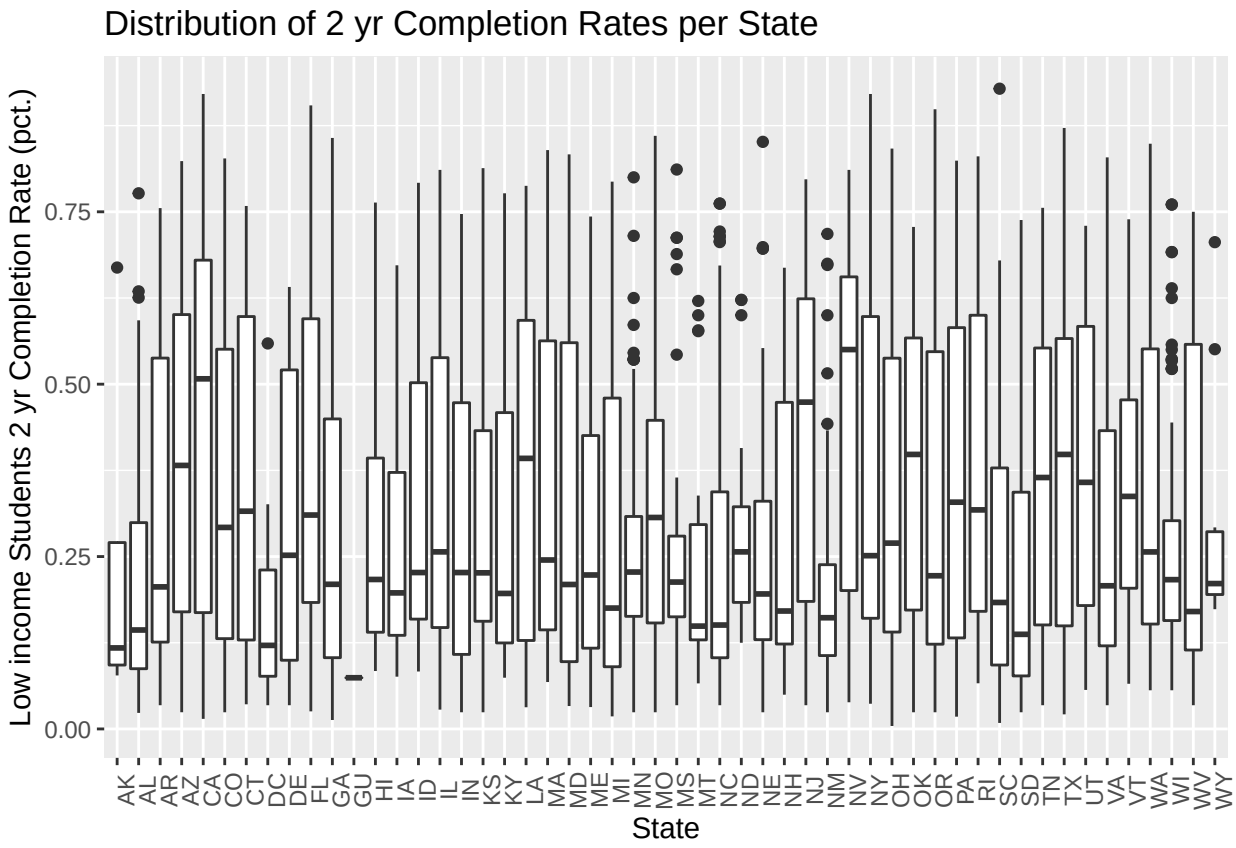
```
      x = "State",
      y = "Low income Students 3 yr Completion Rate (pct.)") +
  theme(axis.text.x = element_text(angle = 90))
```

```
## Warning: Removed 1851 rows containing non-finite values (stat_boxplot).
```

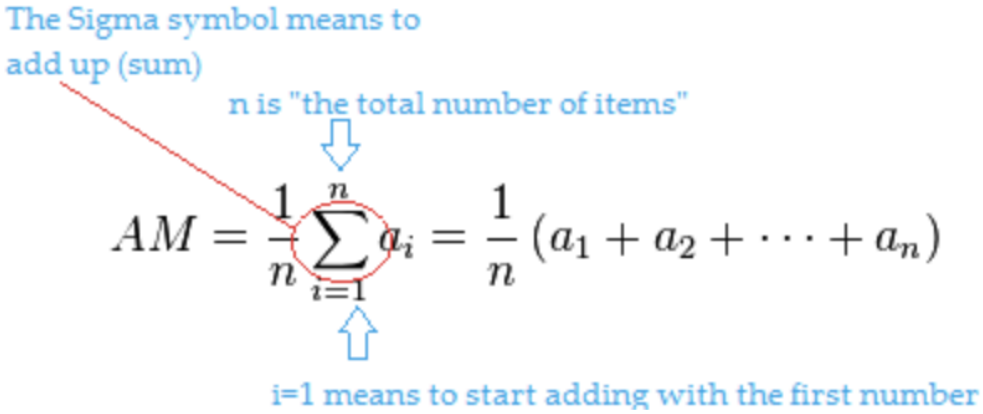



```
college_df %>%
  group_by(STATE) %>%
  ggplot(mapping = aes(x = STATE, y=LO_INC_COMP_ORIG_YR2_RT)) +
  geom_boxplot() +
  labs(title = "Distribution of 2 yr Completion Rates per State",
       x = "State",
       y = "Low income Students 2 yr Completion Rate (pct.)") +
  theme(axis.text.x = element_text(angle = 90))
```

```
## Warning: Removed 2165 rows containing non-finite values (stat_boxplot).
```



To help us answer our problem. We need to start grouping the educational institution entities by state to figure out which state of course has the best value in investing in an education. Below you can see how to calculate the means using r code to find the central tendancy of each state and their on time completion rates for students as well as their educational costs.



Mean Equation:

```
# the mean percent of low income (less than $30,000 in nominal family income) students who completed wi
college_df <- college_df %>%
  group_by(STATE) %>%
  mutate(MEAN_4YR = mean(LO_INC_COMP_ORIG_YR4_RT, na.rm = TRUE)) %>%

# the mean percent of low income (less than $30,000 in nominal family income) students who completed wi
  mutate(MEAN_3YR = mean(LO_INC_COMP_ORIG_YR3_RT, na.rm = TRUE)) %>%
```

```
# the mean percent of low income (less than $30,000 in nominal family income) students who completed wi
  mutate(MEAN_2YR = mean(LO_INC_COMP_ORIG_YR2_RT, na.rm = TRUE)) %>%

# the mean cost of college for low income (less than $30,000 in nominal family income) students at publ
  mutate(MEAN_PUB_COST = mean(AVG_COST_PUB, na.rm = TRUE)) %>%

# the mean cost of college for low income (less than $30,000 in nominal family income) students at priv
  mutate(MEAN_PRIV_COST = mean(AVG_COST_PRIV, na.rm = TRUE))

head(college_df)
```

```
## # A tibble: 6 x 14
## # Groups:   STATE [1]
##   LO_INC_COMP_ORI~ LO_INC_COMP_ORI~ LO_INC_COMP_ORI~ INSTITUTION CITY
##              <dbl>            <dbl>            <dbl> <fct>       <fct>
## 1           0.0323           0.0725           0.169  Alabama A ~ Norm~
## 2           0.157            0.305            0.402  University~ Birm~
## 3          NA                0.0739           0.0897 Amridge Un~ Mont~
## 4           0.191            0.309            0.375  University~ Hunt~
## 5           0.0414           0.100            0.171  Alabama St~ Mont~
## 6           0.143            0.339            0.407  The Univer~ Tusc~
## # ... with 9 more variables: STATE <fct>, UNITID <int>,
## #   AVG_COST_PUB <dbl>, AVG_COST_PRIV <dbl>, MEAN_4YR <dbl>,
## #   MEAN_3YR <dbl>, MEAN_2YR <dbl>, MEAN_PUB_COST <dbl>,
## #   MEAN_PRIV_COST <dbl>
```

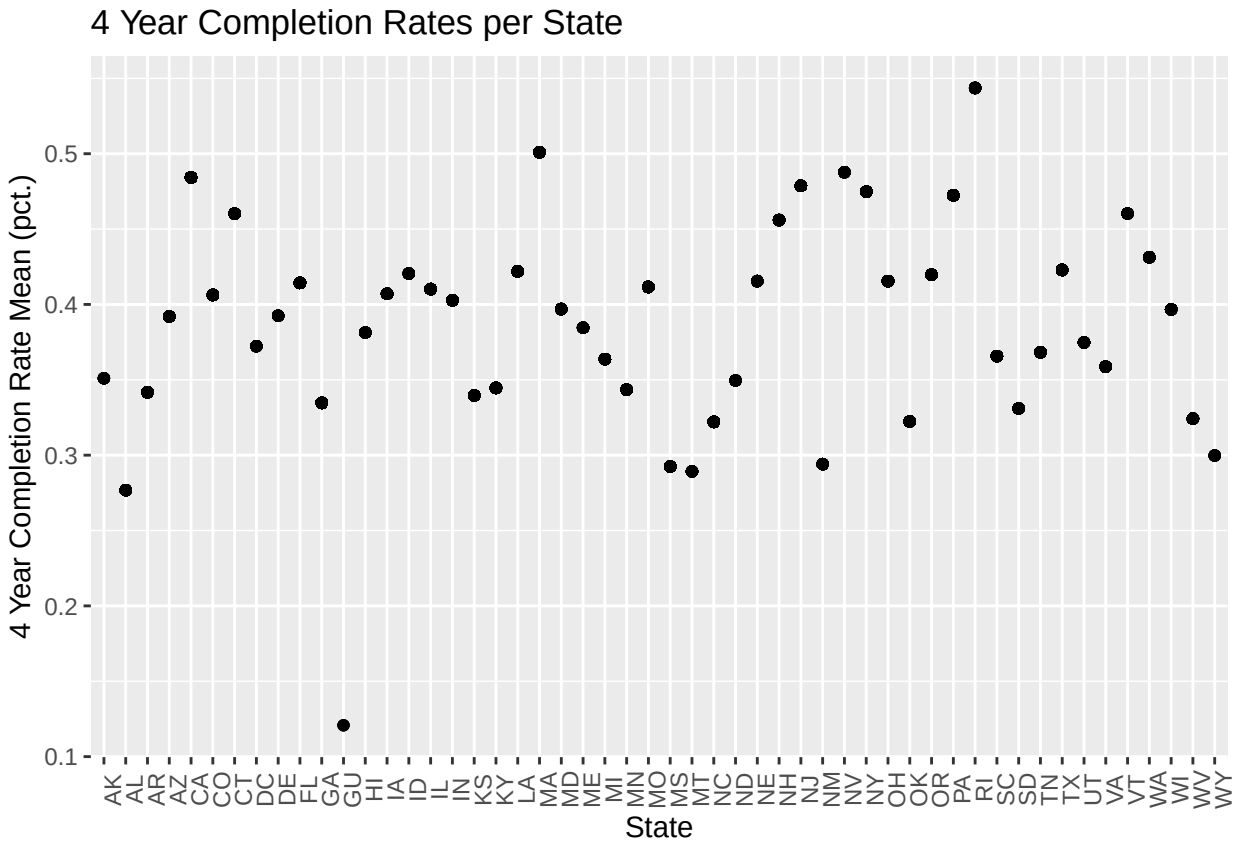
```
# the median percent of low income (less than $30,000 in nominal family income) students who completed
college_df <- college_df %>%
  group_by(STATE) %>%
  mutate(Median_4YR = median(LO_INC_COMP_ORIG_YR4_RT, na.rm = TRUE)) %>%

# the median percent of low income (less than $30,000 in nominal family income) students who completed
  mutate(Median_3YR = median(LO_INC_COMP_ORIG_YR3_RT, na.rm = TRUE)) %>%

# the median percent of low income (less than $30,000 in nominal family income) students who completed
  mutate(Median_2YR = median(LO_INC_COMP_ORIG_YR2_RT, na.rm = TRUE)) %>%

# the median cost of college for low income (less than $30,000 in nominal family income) students at pu
  mutate(Median_PUB_COST = median(AVG_COST_PUB, na.rm = TRUE)) %>%

# the median cost of college for low income (less than $30,000 in nominal family income) students at pr
  mutate(Median_PRIV_COST = median(AVG_COST_PRIV, na.rm = TRUE))

head(college_df)
```

```
## # A tibble: 6 x 19
## # Groups:   STATE [1]
##   LO_INC_COMP_ORI~ LO_INC_COMP_ORI~ LO_INC_COMP_ORI~ INSTITUTION CITY
##              <dbl>            <dbl>            <dbl> <fct>       <fct>
## 1           0.0323           0.0725           0.169  Alabama A ~ Norm~
## 2           0.157            0.305            0.402  University~ Birm~
## 3          NA                0.0739           0.0897 Amridge Un~ Mont~
## 4           0.191            0.309            0.375  University~ Hunt~
## 5           0.0414           0.100            0.171  Alabama St~ Mont~
```

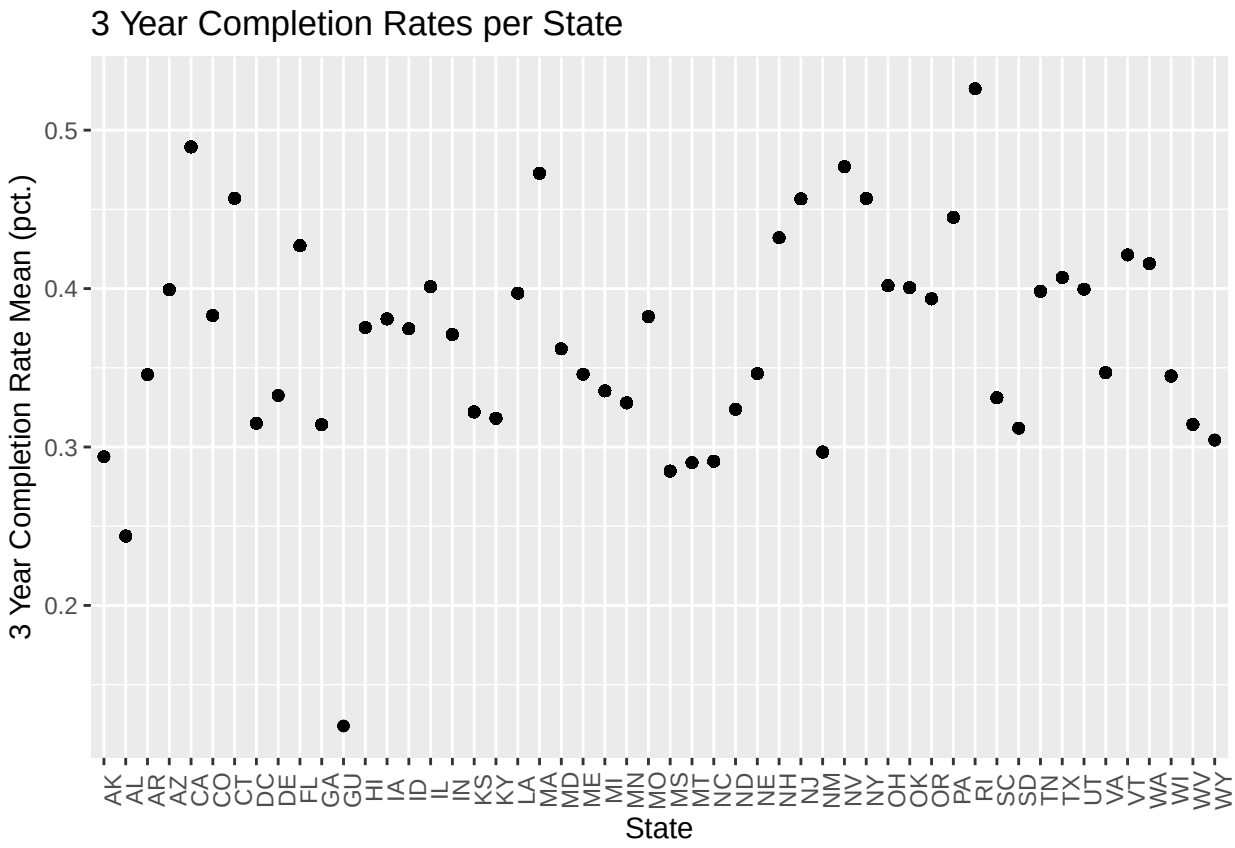
```
## 6           0.143            0.339            0.407  The Univer~ Tusc~
## # ... with 14 more variables: STATE <fct>, UNITID <int>,
## #   AVG_COST_PUB <dbl>, AVG_COST_PRIV <dbl>, MEAN_4YR <dbl>,
## #   MEAN_3YR <dbl>, MEAN_2YR <dbl>, MEAN_PUB_COST <dbl>,
## #   MEAN_PRIV_COST <dbl>, Median_4YR <dbl>, Median_3YR <dbl>,
## #   Median_2YR <dbl>, Median_PUB_COST <dbl>, Median_PRIV_COST <dbl>
```

We can now use plots to see how all the each state's mean costs, median costs, and graduation rates compare to one another before we eventually model what the best State to attend is. They key here is to make sure that the units are established and consistent among the plots so that people can make comparisons. You also want to ensure elligibility for presenting the EDA, so you want to make sure you prep the labels.

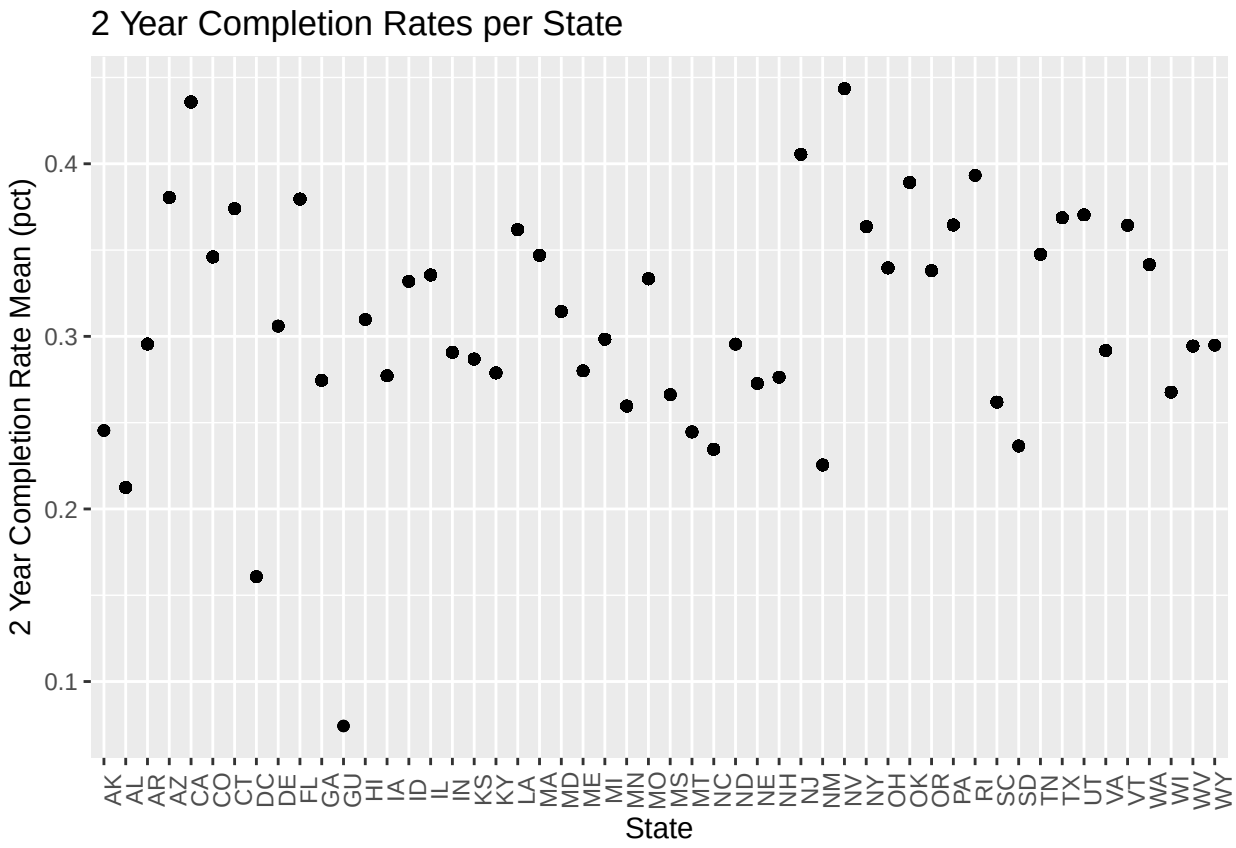
```
college_df %>%
  ggplot(mapping = aes(x = STATE, y = MEAN_4YR)) +
  geom_point() +
  labs(title = "4 Year Completion Rates per State",
       x = "State",
       y = "4 Year Completion Rate Mean (pct.)") +
  theme(axis.text.x = element_text(angle = 90))
```



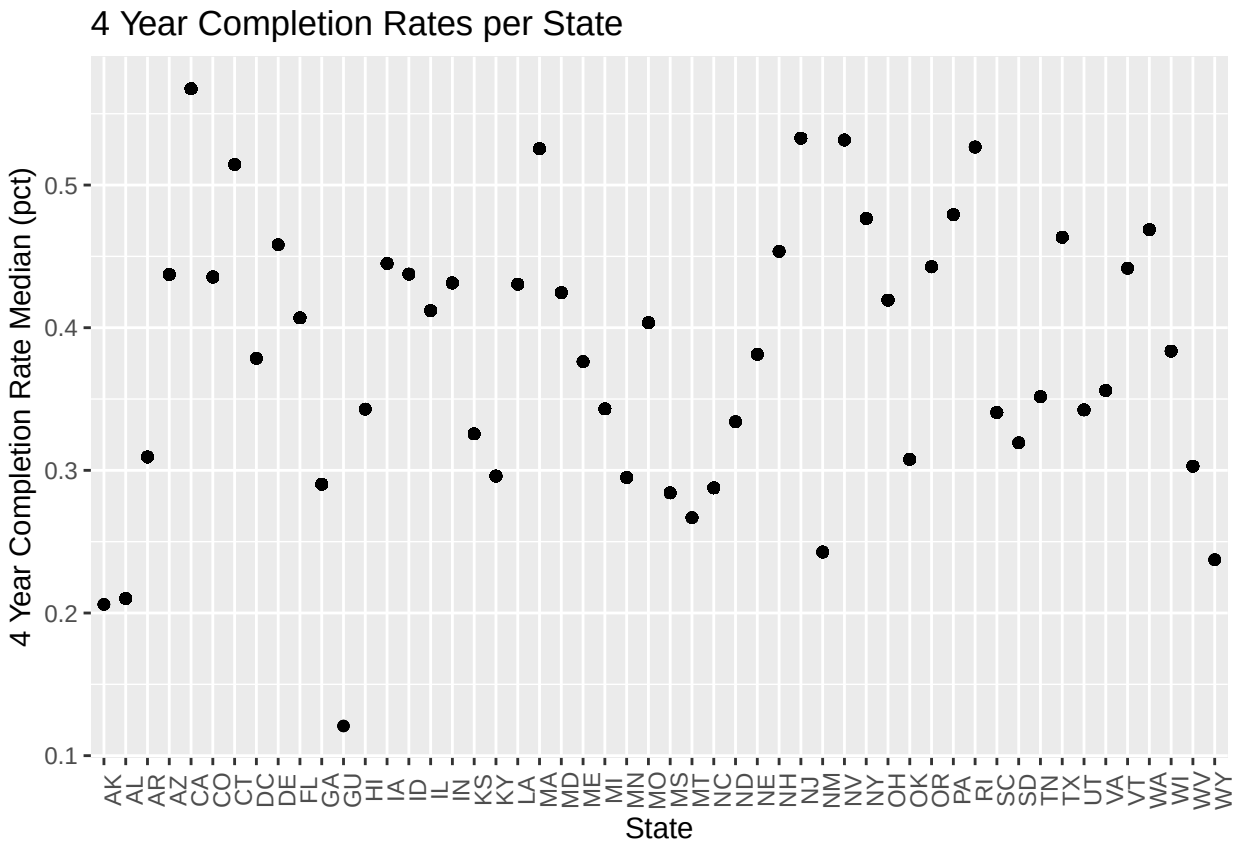

```
college_df %>%
  ggplot(mapping = aes(x = STATE, y = MEAN_3YR)) +
  geom_point() +
  labs(title = "3 Year Completion Rates per State",
       x = "State",
       y = "3 Year Completion Rate Mean (pct.)") +
  theme(axis.text.x = element_text(angle = 90))
```



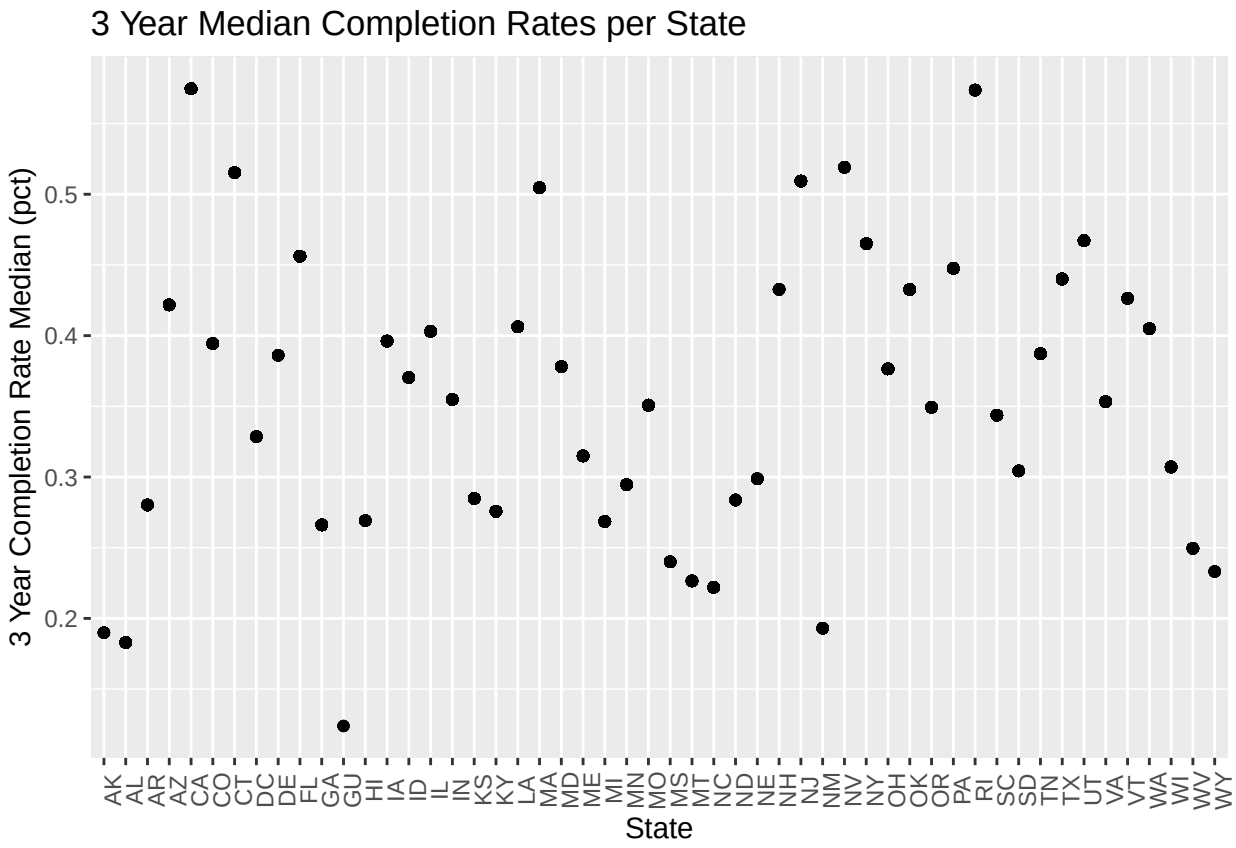

```
college_df %>%
  ggplot(mapping = aes(x = STATE, y = MEAN_2YR)) +
  geom_point() +
  labs(title = "2 Year Completion Rates per State",
       x = "State",
       y = "2 Year Completion Rate Mean (pct)") +
  theme(axis.text.x = element_text(angle = 90))
```



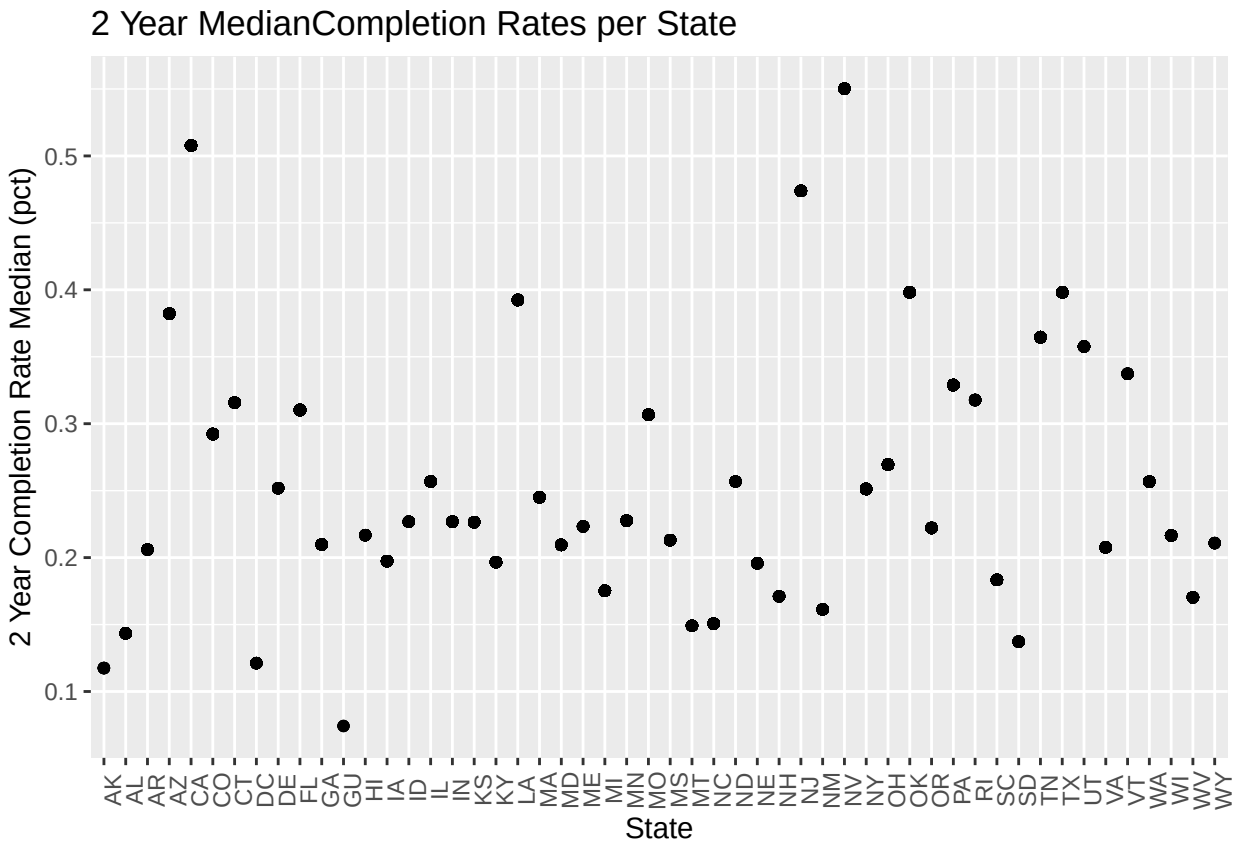

```
college_df %>%
  ggplot(mapping = aes(x = STATE, y = Median_4YR)) +
  geom_point() +
  labs(title = "4 Year Completion Rates per State",
       x = "State",
       y = "4 Year Completion Rate Median (pct)") +
  theme(axis.text.x = element_text(angle = 90))
```



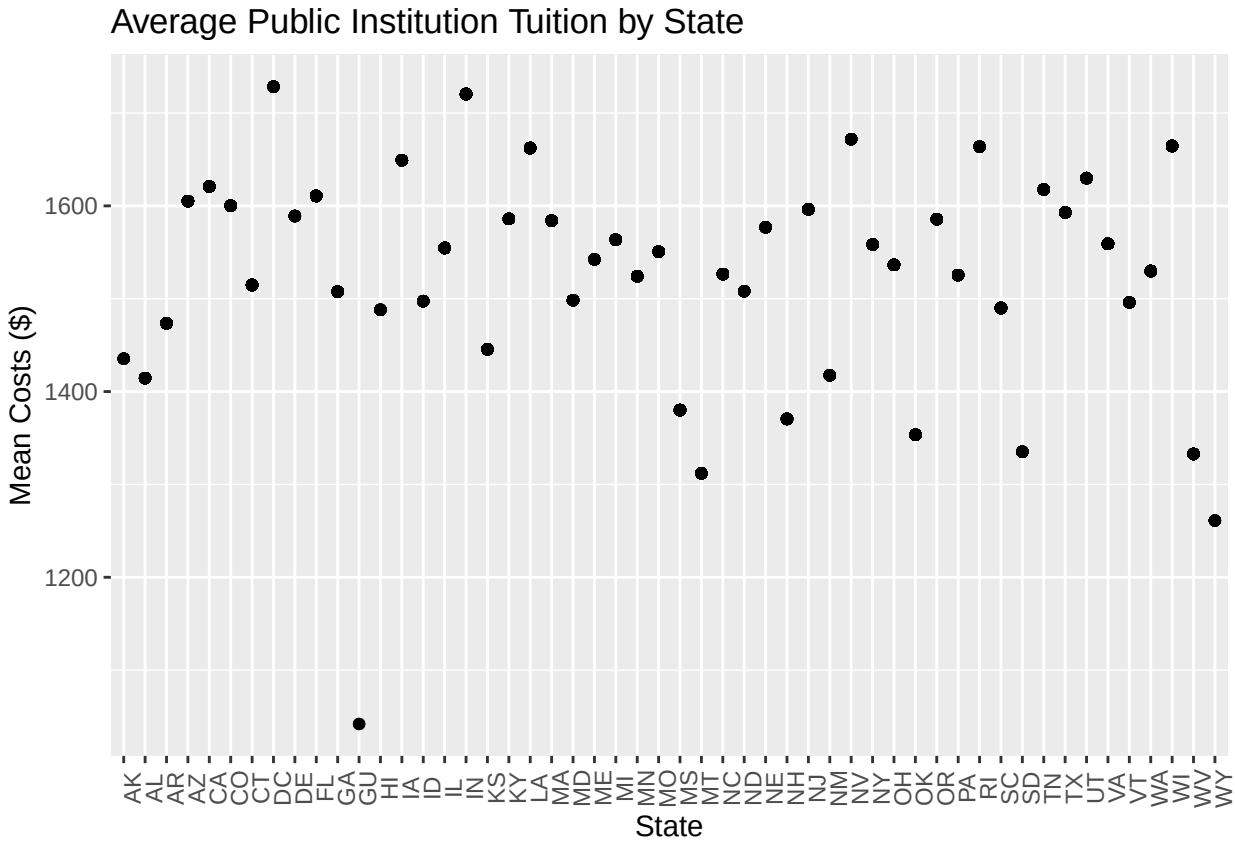

```
college_df %>%
  ggplot(mapping = aes(x = STATE, y = Median_3YR)) +
  geom_point() +
  labs(title = "3 Year Median Completion Rates per State",
       x = "State",
       y = "3 Year Completion Rate Median (pct)") +
  theme(axis.text.x = element_text(angle = 90))
```



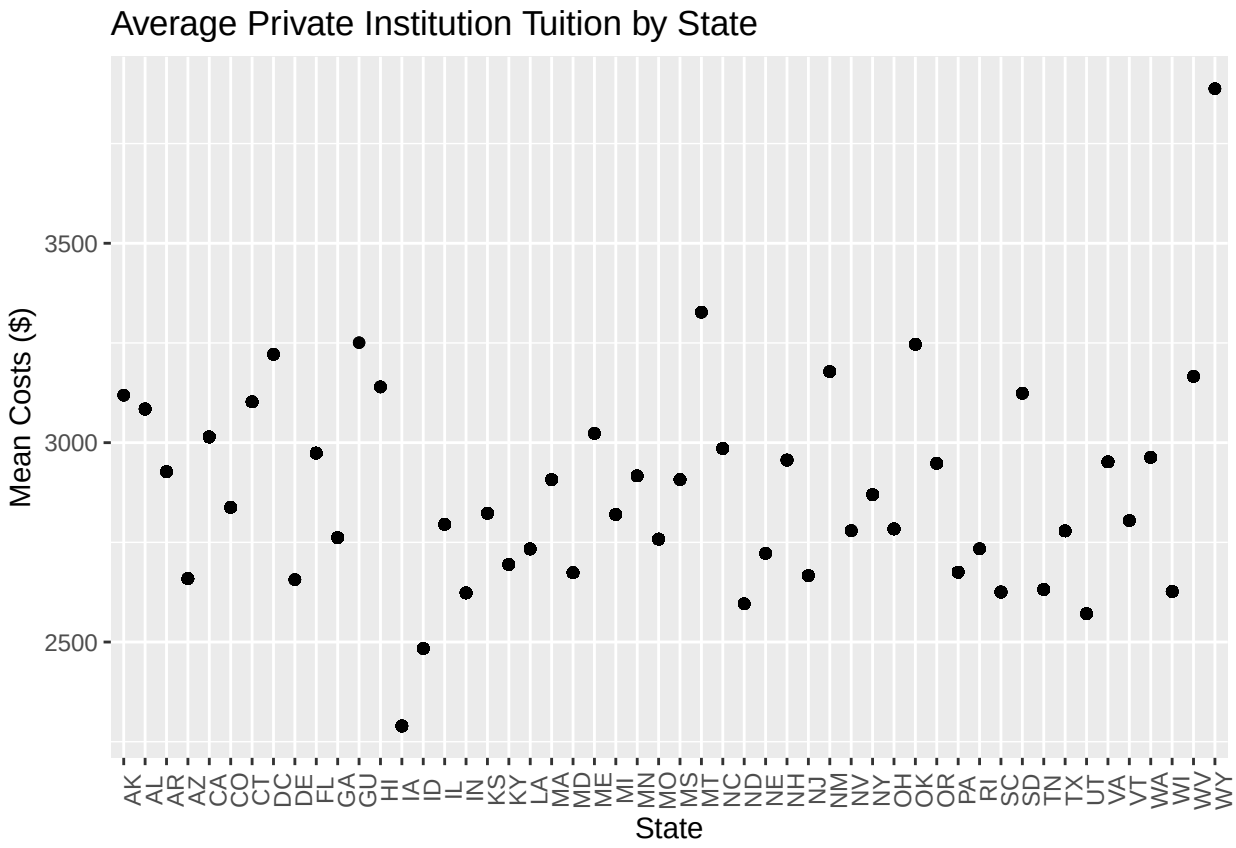

```
college_df %>%
  ggplot(mapping = aes(x = STATE, y = Median_2YR)) +
  geom_point() +
  labs(title = "2 Year MedianCompletion Rates per State",
       x = "State",
       y = "2 Year Completion Rate Median (pct)") +
  theme(axis.text.x = element_text(angle = 90))
```



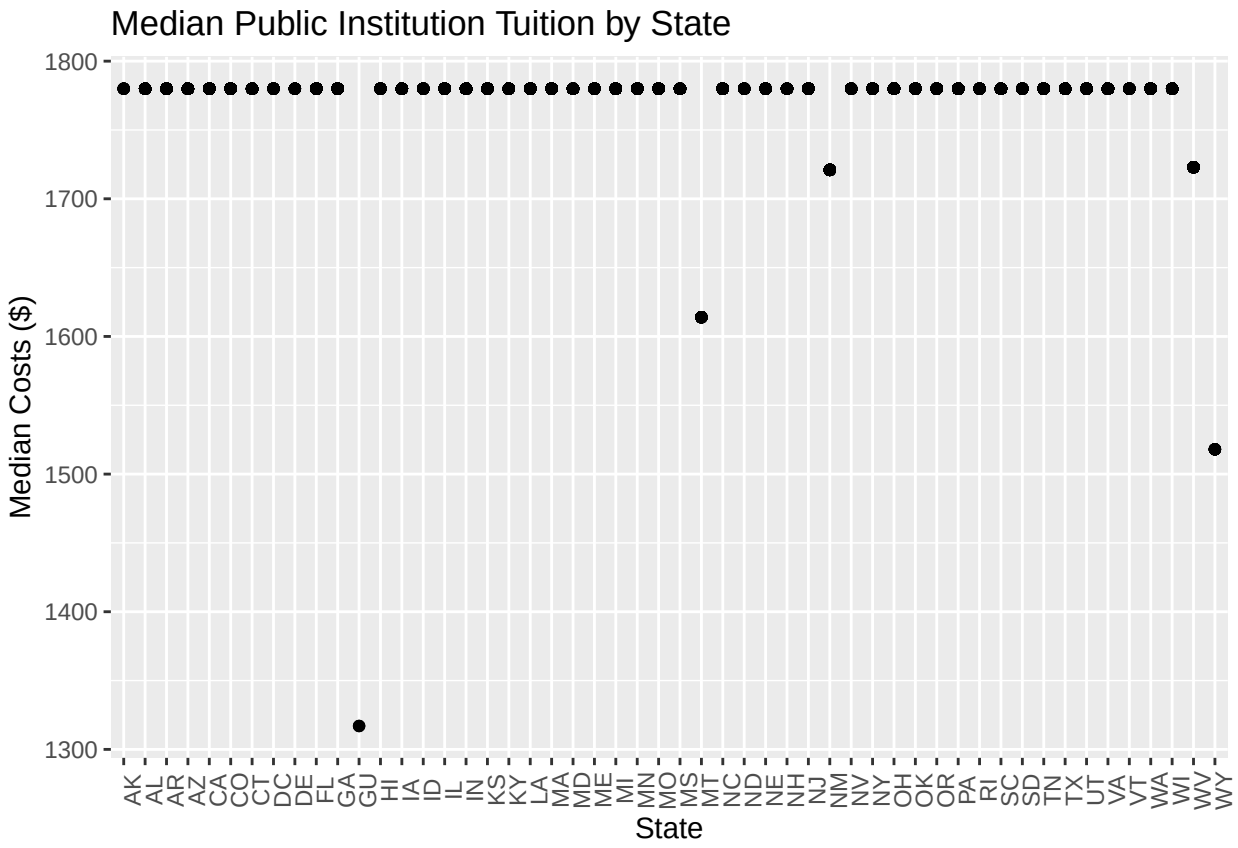

```
college_df %>%
  ggplot(mapping = aes(x=STATE, y = MEAN_PUB_COST)) +
  geom_point() +
  labs(title = "Average Public Institution Tuition by State",
       x = "State",
       y = "Mean Costs ($)") +
  theme(axis.text.x = element_text(angle = 90))
```




```
college_df %>%
  ggplot(mapping = aes(x=STATE, y = MEAN_PRIV_COST)) +
  geom_point() +
  labs(title = "Average Private Institution Tuition by State",
       x = "State",
       y = "Mean Costs ($)") +
  theme(axis.text.x = element_text(angle = 90))
```



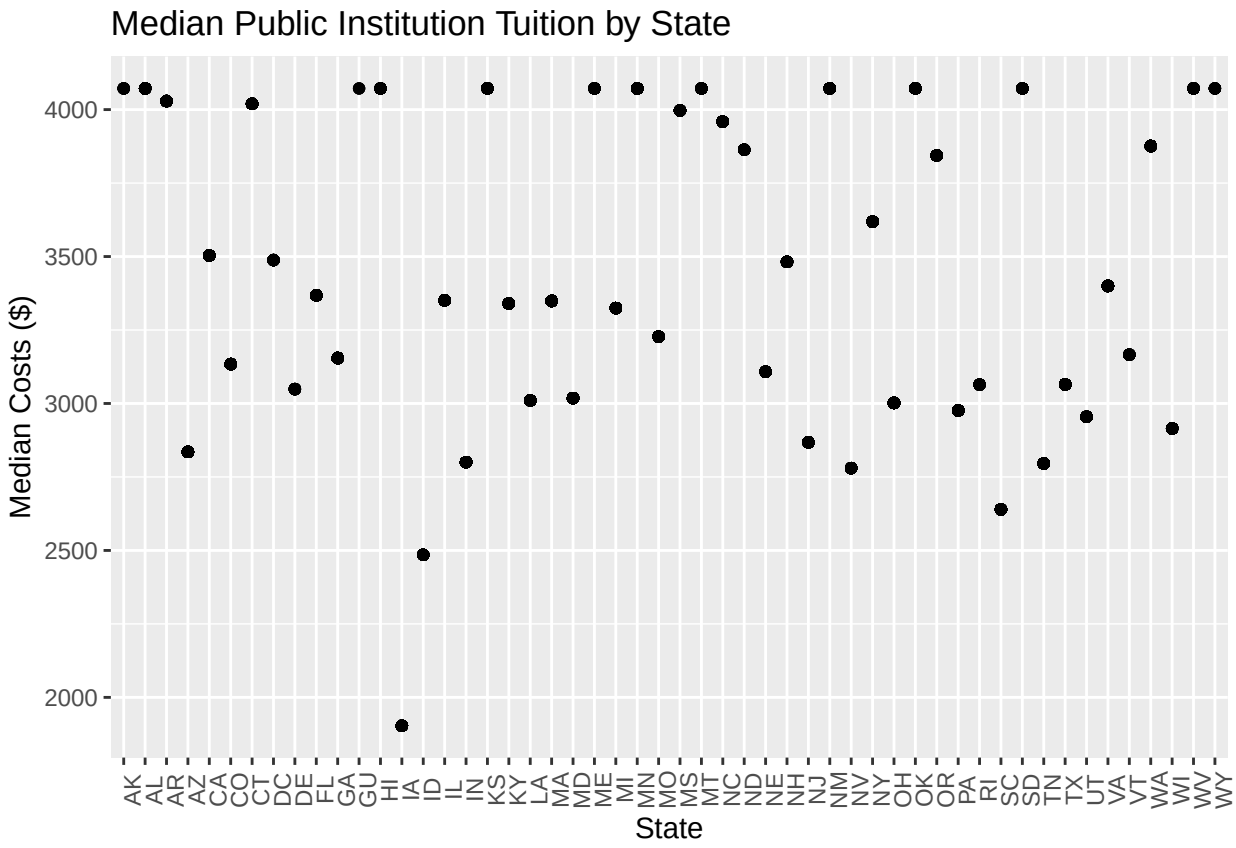

```
college_df %>%
  ggplot(mapping = aes(x=STATE, y = Median_PUB_COST)) +
  geom_point() +
  labs(title = "Median Public Institution Tuition by State",
       x = "State",
       y = "Median Costs ($)") +
  theme(axis.text.x = element_text(angle = 90))
```




```
college_df %>%
  ggplot(mapping = aes(x=STATE, y = Median_PRIV_COST)) +
  geom_point() +
  labs(title = "Median Public Institution Tuition by State",
       x = "State",
       y = "Median Costs ($)") +
  theme(axis.text.x = element_text(angle = 90))
```

$$\text{var}(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Figure 1: Variance



Now that we see the central tendancies for each region lets also check for the spread of states mean and median costs and graduation rates. This requires us to calculate the variance to arrive at the spread. It's as simple as using the var() r function. What this does is calculates the average distance from the mean for each state and the institutions in its state. Here is an attached photo of the Variance equation:

```
var_yr2_rt <- college_df$LO_INC_COMP_ORIG_YR2_RT %>%
  var(na.rm = TRUE)
var_yr3_rt <- college_df$LO_INC_COMP_ORIG_YR3_RT %>%
  var(na.rm = TRUE)
var_yr4_rt <- college_df$LO_INC_COMP_ORIG_YR4_RT %>%
  var(na.rm = TRUE)

var_avg_cost_pub <- college_df$AVG_COST_PUB %>%
  var(na.rm = TRUE)

var_avg_cost_pub <- college_df$AVG_COST_PRIV %>%
  var(na.rm = TRUE)
```

We can then square root the variance to calculate the standard deviation of that vector of data. This means we can standardize the costs between private and public institution costs. This wouldn't make sense of course for our question we set out to answer since we want to directly compare the costs between private and public institutions to see which ones reap a better value for the student.

Now, lets look to see if there is any skew in the data. It would be good to know if some states are skewed to lower academic completion rates, meaning that the range is larger and more dense to the left of the distribution of academic completion rates for that particular state.

```
college_df %>%
  group_by(STATE) %>%

# Ranges of 3/4 and 1/4 quartiles for completion in 4 years at original institution. If yr41_depth
# and yr42_dpeth have a large differnece, then this can tell you that there is a skew for the
# rates of completion in 4,3, or 2 years.
  mutate(q1_yr4_depth = quantile(LO_INC_COMP_ORIG_YR4_RT, 1/4, na.rm = TRUE),
         q3_yr4_depth = quantile(LO_INC_COMP_ORIG_YR4_RT, 3/4, na.rm = TRUE),
         yr4_1_depth = Median_4YR - q1_yr4_depth,
         yr4_2_depth = q3_yr4_depth - Median_4YR) %>%

  mutate(q1_yr3_depth = quantile(LO_INC_COMP_ORIG_YR3_RT, 1/4, na.rm = TRUE),
         q3_yr3_depth = quantile(LO_INC_COMP_ORIG_YR3_RT, 3/4, na.rm = TRUE),
         yr3_1_depth = Median_3YR - q1_yr3_depth,
         yr3_2_depth = q3_yr3_depth - Median_3YR) %>%

  mutate(q1_yr2_depth = quantile(LO_INC_COMP_ORIG_YR2_RT, 1/4, na.rm = TRUE),
         q3_yr2_depth = quantile(LO_INC_COMP_ORIG_YR2_RT, 3/4, na.rm = TRUE),
         yr2_1_depth = Median_2YR - q1_yr2_depth,
         yr2_2_depth = q3_yr2_depth - Median_2YR) %>%

  # We can do the same calculations to look for a skew in the costs for public and private institutions
  mutate(q1_pub_depth = quantile(AVG_COST_PUB, 1/4, na.rm = TRUE),
         q3_pub_depth = quantile(AVG_COST_PUB, 3/4, na.rm = TRUE),
         pub_1_depth = Median_PUB_COST - q1_pub_depth,
         pub_2_depth = q3_pub_depth - Median_PUB_COST) %>%

  mutate(q1_priv_depth = quantile(AVG_COST_PRIV, 1/4, na.rm = TRUE),
         q3_priv_depth = quantile(AVG_COST_PRIV, 3/4, na.rm = TRUE),
         priv_1_depth = Median_PRIV_COST - q1_priv_depth,
         priv_2_depth = q3_priv_depth - Median_PRIV_COST) %>%

  select(yr4_2_depth, yr4_1_depth, yr3_2_depth,yr3_1_depth, yr2_2_depth, yr2_1_depth,
         pub_2_depth, pub_1_depth, priv_2_depth, priv_1_depth)
```

```
## Adding missing grouping variables: `STATE`
```

```
## # A tibble: 7,437 x 11
## # Groups:   STATE [52]
##    STATE yr4_2_depth yr4_1_depth yr3_2_depth yr3_1_depth yr2_2_depth
##    <fct>       <dbl>       <dbl>       <dbl>       <dbl>       <dbl>
##  1 AL          0.157      0.0666       0.178      0.0565       0.156
##  2 AL          0.157      0.0666       0.178      0.0565       0.156
##  3 AL          0.157      0.0666       0.178      0.0565       0.156
##  4 AL          0.157      0.0666       0.178      0.0565       0.156
##  5 AL          0.157      0.0666       0.178      0.0565       0.156
##  6 AL          0.157      0.0666       0.178      0.0565       0.156
```

```
##  7 AL          0.157      0.0666       0.178      0.0565       0.156
##  8 AL          0.157      0.0666       0.178      0.0565       0.156
##  9 AL          0.157      0.0666       0.178      0.0565       0.156
## 10 AL          0.157      0.0666       0.178      0.0565       0.156
## # ... with 7,427 more rows, and 5 more variables: yr2_1_depth <dbl>,
## #   pub_2_depth <dbl>, pub_1_depth <dbl>, priv_2_depth <dbl>,
## #   priv_1_depth <dbl>
```

You can now determine from comparing depth 1 to depth 2 whether or not you believe the difference in range between them, can tell you if ir might be skewed. This means if depth 1 is much bigger than depth 2 it might be skewed to the depth 1 side, meaning skewed towards lower rates and tuition costs.

Now after looking at Central Tendancy, Spread, and Skew on the data we have decided to look at above, we are able to move onto using hypothesis testing.

Linear Regression Model

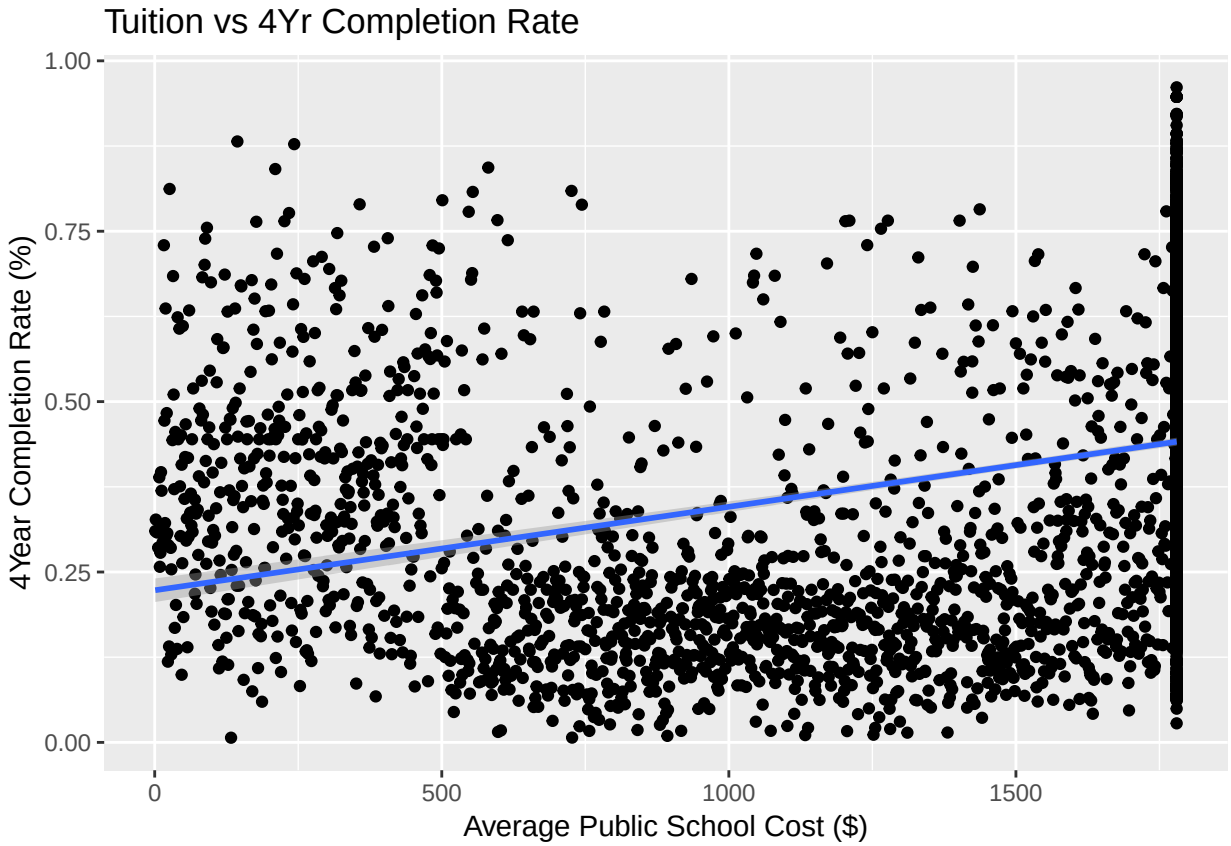
Linear Model in depth expained: https://www.mathworks.com/discovery/linear-model.html

We can use a simple linear regression model to see how one continious variable Y relates to a numeric or continuous variable X. As an example, we can use the pipeline we have already built and analyze the relationship between cost of the institution and the completion rate in 2, 3, and 4 years at the institution. We will use the average cost denoted under the columns AVG_COST_PUB and AVG_COST_PRIV as our X variable to see how the Y variable, the completion rate fluxuates based on cost.

```
college_df %>%
  ungroup() %>%
  ggplot(mapping = aes(x = AVG_COST_PUB, y = LO_INC_COMP_ORIG_YR4_RT)) +
  geom_point() +
  geom_smooth(method = lm) +
  labs(title = "Tuition vs 4Yr Completion Rate",
       y = "4Year Completion Rate (%)",
       x = "Average Public School Cost ($)")
```

```
## Warning: Removed 1857 rows containing non-finite values (stat_smooth).
```

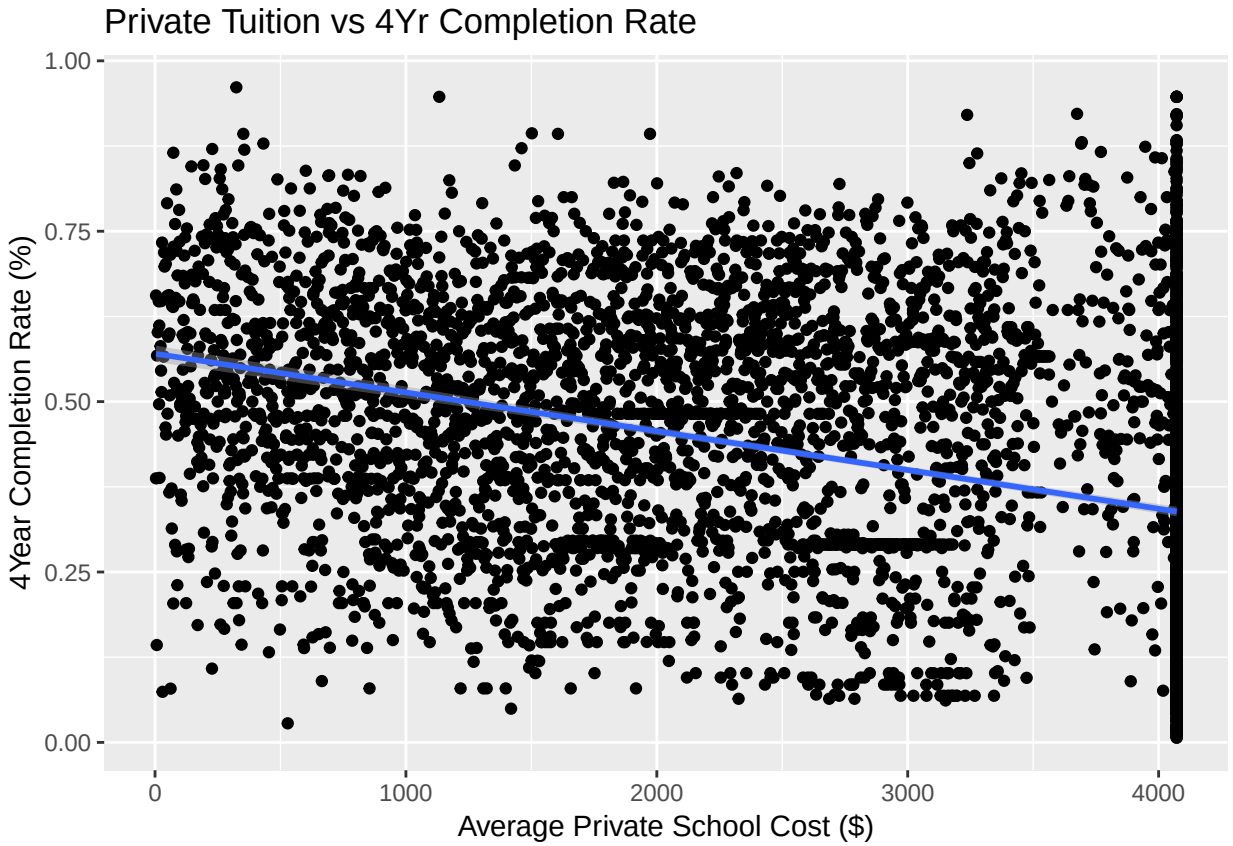
```
## Warning: Removed 1857 rows containing missing values (geom_point).
```




```
college_df %>%
  ungroup() %>%
  ggplot(mapping = aes(x = AVG_COST_PRIV, y = LO_INC_COMP_ORIG_YR4_RT)) +
  geom_point() +
  geom_smooth(method = lm) +
  labs(title = "Private Tuition vs 4Yr Completion Rate",
       y = "4Year Completion Rate (%)",
       x = "Average Private School Cost ($)")
```

```
## Warning: Removed 1857 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 1857 rows containing missing values (geom_point).
```

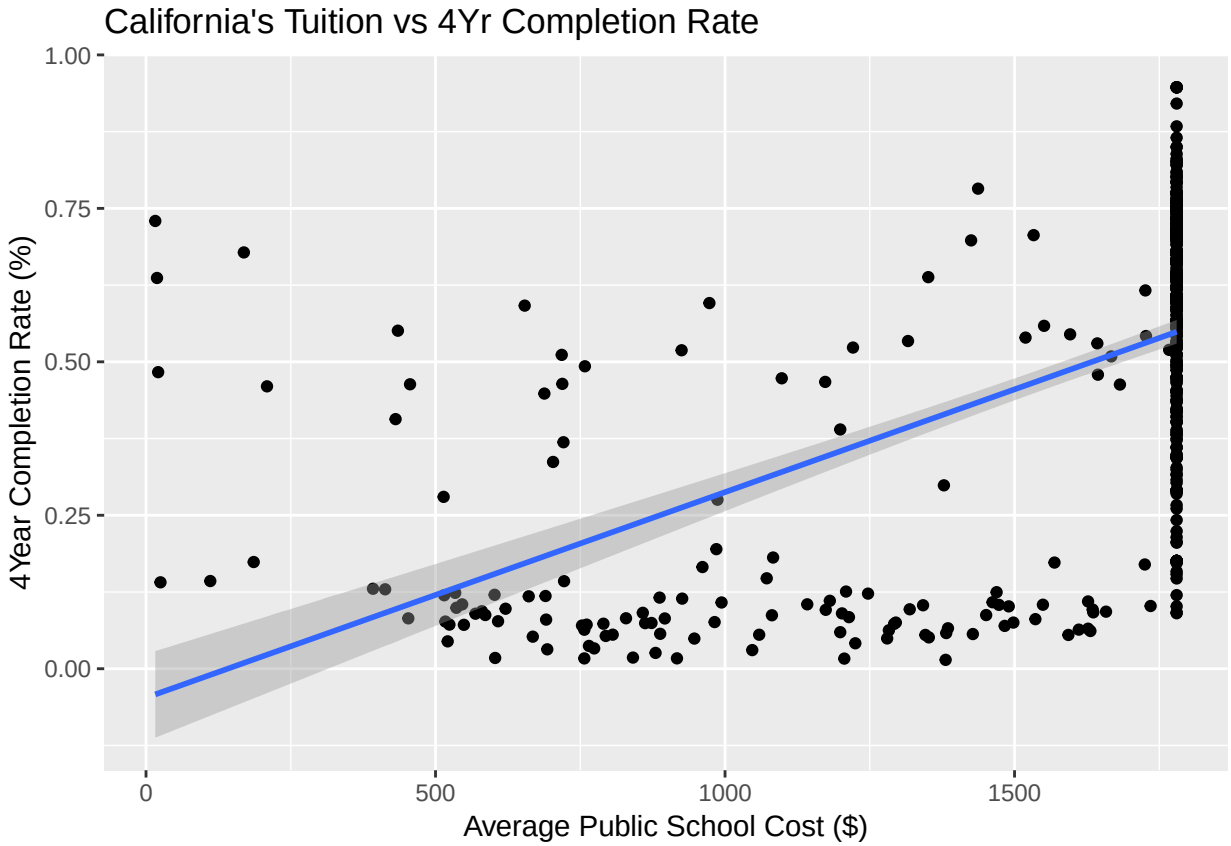


In this specific example, I have just graphed the average cost of public schools and private schools on the x-axis and the completion rate in 4 years for low income students at those schools on the y-axis. Note that this includes all schools regardless of the state. In the next code block we will use the filter() function to select a specific state we want to test to see if that positive correlation between averge public school cost and the completion rate in 4 years.

```
college_df %>%
  ungroup() %>%
  filter(STATE == "CA") %>%
  ggplot(mapping = aes(x = AVG_COST_PUB, y = LO_INC_COMP_ORIG_YR4_RT)) +
  geom_point() +
  geom_smooth(method = lm) +
  labs(title = "California's Tuition vs 4Yr Completion Rate",
       y = "4Year Completion Rate (%)",
       x = "Average Public School Cost ($)")
```

```
## Warning: Removed 222 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 222 rows containing missing values (geom_point).
```



We learn from this linear model, that in California there is a slight positive corelation between average public school cost and low income student's 4 year graduation rate. the positive corelation is not as strong as the one found in the example above when we included all states.

Final Insights

Tying it back towards our initial questions when we embarked on this pipeline tutorial, we were trying to answer whether public institutions with a higher average cost of attendance yielded higher completion rates at institutions. We found from our linear models for public and private school costs that the completion rates for lower income students increased in public schools as the school was more pricey, but for students at private institutions the oppositite might be true due to the negative correlation.

Now we must rememeber to answer questions ethically based on the results we have found, and not overstep any conclusions.

Our data set gave us information on public and private higher education institutions for the year 2015 merged with 2016. If we wanted to maintain a more accurate assessment for the general relationship between cost of attendance and that school's chance for higher completion for lower income students. This would hopefully tell us if that institution is worth it's cost since even the poorer students complete their tenure there more often than in other institutions.

We have also learned from our mean scatterplot and the box plots of completion rates among each state, which states on average have higher completion rates. We can pull from out 4 year completion rate boxplot and see that California has the highest average and the box is higher on for California which tells us that there are more schools with the higher completion rate of around .55 for the low income students.

Now, you can search for a data set and begin to draw conclusions using this data science pipeline to load a

data set, clean it up, perform exploratory data analysis on it, and then build our linear regression model to answer one of our questions.