Finances in Education

Denise Dodd

2023-08-04

Table of Contents

Milestone 1

Introduction

In 2022, my state of Missouri was ranked the lowest in the country in terms of teacher pay [@Grumke2022b]. At the same time, many school districts in the state were moving to a 4-day school week due to a teacher shortage [@Grumke2022a]. Additionally, the state budget in 2022 had a $6 billion dollar surplus[@Keller2022]. This led me to wonder if there is any connection between finances and student achievement. This project specifically will focus on what effect money has on on a person's obtainment of a bachelor's degree based on state education expenditure per student, average household income, average teacher pay, and percent of people who have obtained a bachelor degree.

Research Questions

Questions that will be raised during this project are:

1) How does average teacher pay effect bachelor obtainment?
2) How does state education expenditure effect bachelor obtainment?
3) How does the average household income effect bachelor obtainment?
4) How does state education expenditure effect teacher pay?
5) At a certain point does bachelor obtainment flatten out regardless of other variables?

Approach

To answer these questions, I will first need to clean my data sets and transform them into a format that is compatible with one another and exclude any outliers. Next, I will review the correlation between the variables to determine if the variables increase/decrease together or in opposing directions and the strength of their relationships. I will then calculate the Coefficient of Determination to determine what percent of variability in bachelor obtainment is caused by each variable. The covariance will also be calculated to determine the strength of the relationships between the variables.

I will also be creating scatter plots and regression lines to visually represent the relationship between the variables.

How my approach addresses the problem.

By assessing these relationships and calculations, I hope to determine which (if any) financial variables contributes the largest variability in bachelor obtainment.

Data

I will be using the below data sets:

-Bachelor or Higher by State - This data set from the U.S. Federal Reserve has data for each state with variables including: 2020 Percent 25 or Older with a Bachelors Degree, Percent with Bachelors Degree From Preceding Year, and and identical variable with Percent with Bachelors Degree Year Ago From Period. The Federal Reserve shares the following regarding it's data collection methods: "Each Federal Reserve Bank gathers information on current economic conditions in its District through reports from Bank and Branch directors, plus interviews and online questionnaires completed by businesses, community organizations, economists, market experts, and other sources. Contacts are not selected at random; rather, Banks strive to curate a diverse set of sources that can provide accurate and objective information about a broad range of economic activities." [@FedRes2020b].

-Spending - This data set from the U.S. Census has extensive date across multiple tabs and variables. I will be utilizing the Total Elementary-Secondary Expenditure data for each state from this data set. The U.S. Census states that it collects "data about the economy and the people living in the United States

from many different sources. Some data are collected from respondents directly (including businesses), through the censuses and surveys we conduct. We also collect additional data from other sources. Primary sources for additional data are federal, state, and local governments, as well as some commercial entities." [@Census2020].

-Avg Income by State - This data set from the U.S. Federal Reserve has data for each state with variables including: 2020 Avg Income, Avg Income From Preceding Year, and and identical variable with Avg Income Year Ago From Period. Please see the details in the Bachelor or Higher by State table for details on the Federal Reserve's data collection methods. [@FedRes2020a].

-Teacher Pay - This data set from the National Center for Education Statistics details the average teacher salary for each state for the last school year of each decade beginning with the 1969/1970 school year. "NCES collects information through many surveys, using complex assessments, administrative sources, and samples of schools, institutions, and households" [@NCES2020a].

Required Packages

The below packages will be required for this project:

-readxl

-tidyverse

-dplyr

-ggplot2

-ltm

Plots and Table Needs

I will be utilize ggplot2 to create scatter plots and regression lines between the variables. Additionally, I will include histograms with normal curves to determine the distribution of finances and bachelor obtainment across the states.

Working with multiple different data frames containing a variety of variables can be overwhelming. I will be joining the data frames together into one data frame and filtering for only the necessary variables.

Questions for future steps

Due to cost of living differences, it will be difficult to compare Teacher Pay and Avg Income across the states. Is there a way to adjust for this in the data?

We don't know how many students are supported in each state by the the total state expenditure. Is there a way to find this information and convert this column to avg expenditure per student so we have comparable data points?

Average income was included to determine if a student's household income would have an effect on degree obtainment, but without knowing how many sets of the average income are contributing to the household and how many people that income is supporting, I'm unsure if this is a reliable variable. How can this variable be most accurately utilized within the context of this study?

Milestone 2

Updates from Milestone 1

I will be utilizing an additional data set which was not included in my Week 1 report.

- Enrollment - This data set from the National Center for Educational Statistics has enrollment data for each state. I will be using this table along with the Expenditure table referenced in Week 1 to find each state's average expenditure per student. "NCES collects information through many surveys, using complex assessments, administrative sources, and samples of schools, institutions, and households" [@NCES2020b].

How to import and clean my data.

The following steps will be taken to import the data:

-All data is in excel format so I will be using readxl library to load the data and place it into a data frame.

The following steps will be taken to clean the data:

-Subset of the data sets pertaining to 2020 data so data within the same time frame is being used.

-All money is being reported in dollars so equal units are being compared.

-Rename all columns containing states to have the column name "State" so I can join on this column for my final data frame.

-Select only relevant columns in my final data frame.

-Rename my columns for clarity and ease of use.

-Remove the District of Columbia from the final data frame as this is an outlier.

-Create an Avg_Exp_Per_Stud variable which will calculate each state's average expenditure per student by dividing the Expenditure variable by the Enrollment variable.

Create a tp_vs_inc variable which returns TRUE if the Average Teacher Pay is above the Average Income and FALSE if the average Teacher Pay is below the Average Income.

```
library(readxl)
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```
## Warning: package 'forcats' was built under R version 4.3.2
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(dplyr)
library (ggplot2)

setwd("C:/Users/hadle/OneDrive/Documents/dsc520")

income_df <- read_excel("final/avg income by state.xls",
                        sheet = "2020", skip = 1)
names(income_df)[names(income_df) == "Name"] <- "State"


bachelor_df <- read_excel("final/Bachelor or Higher by State.xls",
                        sheet = "2020", skip = 0)


teacher_pay_df <- read_excel("final/Teacher Pay.xls",
                          sheet = "Sheet1", skip = 0)


spending_df <- read_excel("final/spending.xls",
    sheet = "1", skip = 7)
```

```
## New names:
## * `` -> `...1`
## * `` -> `...2`
## * `sources` -> `sources...4`
## * `sources` -> `sources...5`
## * `sources` -> `sources...6`
```

```
names(spending_df)[names(spending_df) == "...1"] <- "State"

enrollment_df <- read_excel("final/Enrollment.xls",
    sheet = "Digest 2013 Table 203.20", skip = 2)
```

```
## New names:
## * `` -> `...1`
## * `` -> `...15`
## * `` -> `...22`
```

```
names(enrollment_df)[names(enrollment_df) == "...1"] <- "State"

# Combine data frames.
combined_df <- dplyr::left_join(teacher_pay_df, spending_df, by ='State')
combined_df <- dplyr::left_join(combined_df, bachelor_df, by ='State')
combined_df <- dplyr::left_join(combined_df, income_df, by ='State')
combined_df <- dplyr::left_join(combined_df, enrollment_df, by ='State')

# Select needed columns for final data set.
final_df <- dplyr::select(combined_df, State, BD2020, Ex_Total, "TP2019-20", "2020", "Fall 2020")

# Rename columns for clarity.
names(final_df)[names(final_df) == "BD2020"] <- "Prct_Bachelor"
names(final_df)[names(final_df) == "Ex_Total"] <- "Expenditure"
```

```
names(final_df)[names(final_df) == "TP2019-20"] <- "Teacher_Pay"
names(final_df)[names(final_df) == "2020"] <- "Avg_Income"
names(final_df)[names(final_df) == "Fall 2020"] <- "Enrollment"

# Remove District of Columbia as this is an outlier.
final_df <- final_df[-c(9),]

# Create new variable calculating each states average expenditure per student.
final_df$Avg_Exp_Per_Stud <- as.numeric(final_df$Expenditure) / as.numeric(final_df$Enrollment)

#Create variable determining if teacher pay is above or below the average income.
final_df$tp_vs_inc <- (final_df$Avg_Income - final_df$Teacher_Pay) < 0
```

What does the final data set look like?

The final data set has 50 rows. The variables in each row are: State, Prct_Bachelor[@FedRes2020b], Expenditure[@Census2020], Teacher_Pay[@NCES2020a], Avg_Income[@FedRes2020a], Enrollment[@NCES2020b], Avg_Exp_Per_Stud, and tp_vs_inc.

A sample of the first five rows can be found below along with a summary of the data.

```
head(final_df)
```

```
## # A tibble: 6 x 8
##   State      Prct_Bachelor Expenditure Teacher_Pay Avg_Income Enrollment
##   <chr>              <dbl>       <dbl>       <dbl>      <dbl> <chr>
## 1 Alabama             27.8     8472016       54095      46119 715000
## 2 Alaska              31.9     2612622       70877      62715 147200
## 3 Arizona             33       9904670       50381      52313 1226000
## 4 Arkansas            24.9     5851201       49822      47123 485700
## 5 California          36.9    95419511       84659      70643 6543800
## 6 Colorado            44.2    13007701       57269      65352 931600
## # i 2 more variables: Avg_Exp_Per_Stud <dbl>, tp_vs_inc <lgl>
```

```
summary(final_df)
```

```
##     State           Prct_Bachelor    Expenditure        Teacher_Pay
##  Length:50          Min.   :23.10   Min.   : 1736898   Min.   :45192
##  Class :character   1st Qu.:30.23   1st Qu.: 3726160   1st Qu.:52385
##  Mode  :character   Median :33.40   Median : 9476182   Median :56882
##                     Mean   :33.88   Mean   :15500292   Mean   :59787
##                     3rd Qu.:37.42   3rd Qu.:17921825   3rd Qu.:65270
##                     Max.   :46.90   Max.   :95419511   Max.   :87543
##    Avg_Income     Enrollment        Avg_Exp_Per_Stud tp_vs_inc
##  Min.   :42698   Length:50          Min.   : 8.079   Mode :logical
##  1st Qu.:52132   Class :character   1st Qu.:12.405   FALSE:17
##  Median :55920   Mode  :character   Median :14.630   TRUE :33
##  Mean   :57447                      Mean   :15.174
##  3rd Qu.:62151                      3rd Qu.:17.721
##  Max.   :78685                      Max.   :29.307
```

What information is not self-evident?

It is not self evident how this information was gathered or how the sample population for each data set was determined.

It is also not self evident what the cost of living is in each location. This will play a factor in how heavily to weigh average household income and average teacher pay.

Another factor which is not self evident is how many people relocate after high school. Teacher pay, state education expenditure and to an extent average income are all variables pertaining to a student's high school learning. If they then relocate after high school their contribution to the bachelor percentage could be allocated to a different state.

Additionally, it is not self evident what the largest industries are in each state. Some industries don't require a bachelors and instead favor vocational training or alternate qualifications. If there are industry leaders in a state that doesn't require a bachelor's degree, we might see a lower percentage of degree obtainment in these states.

What are different ways you could look at this data?

First, I will review the correlation between the variables to determine if the variables increase/decrease together or in opposing directions and the strength of their relationships.

Next, I will be calculating the Coefficient of Determination to determine what percentage of the variability in one variable can be attributed to another variable.

Additionally, I will also be calculating the covariance between the variables to determine the strength of their relationships.

How do I plan to slice and dice the data?

As previously noted, the data will be sliced and diced in the following ways:

-The District of Columbia has been removed as this skewed high for the rest of the data set and appeared to be an outlier.

-Because cost of living can play a large factor in average income and teacher pay, an additional binary column has been added which determines if the average teacher pay is above or below the average income.

-In order to compare data as accurately as possible, an additional data set and variable has been added to calculate the average state expenditure per student.

How could I summarize my data to answer key questions?

As the above measurements will be calculated individually, I will present the data in a consolidated table. This will allow all information to be reviewed together and compared to one another. In reviewing this table, I will be able to determine which variables have the greatest effect on one another and which variables do not appear to have a relationship.

What types of plots and tables will help me to illustrate the findings to my questions?

I will be creating scatter plots with regression trend lines to visually determine how closely the variables are related and the strength of the trend line.

I will also be utilizing histograms to visually determine modes and trends. A normal curve will be added to the histograms to determine how much of the data lies under the curve and how much of the data lies outside of the curve.

A quadratic model will be used to determine if there is a flatten out point where no matter how much money is available, the percent of bachelor obtainment levels out.

Questions for future steps.

Are any other variables or data points needed?

Is there a more efficient way to smooth the data?

What is the best way to produce and display the quadratic model?

Milestone 3

A story / narrative that emerged from my data.

Introduction.

In 2022, my state of Missouri was ranked the lowest in the country in terms of teacher pay [@Grumke2022b]. At the same time, many school districts in the state were moving to a 4-day school week due to a teacher shortage [@Grumke2022a]. Additionally, the state budget in 2022 had a $6 billion dollar surplus[@Keller2022]. This led me to wonder if there is any connection between finances and student achievement.

In this project, I will be exploring the relationship between the percent of bachelor degree obtainment in each state and a variety of financial variables including each state's education expenditure, average income, and teacher pay.

The problem statement I addressed.

How much do financial variables (state education expenditure per student, average income, average teacher pay) factor in to a student's ability to obtain a bachelor's degree?

How I addressed this problem statement

To research how finances factor in to a student's ability to obtain a bachelor's degree, I will be using data from a range of sources including the U.S. Census and the U.S. Federal Reserve to create a data frame including each state's percent of bachelor obtainment[@FedRes2020b], education expenditure[@Census2020], K-12 enrollment[@NCES2020b], teacher pay[@NCES2020a], and average income[@FedRes2020a]. I will be using these variables to create two additional metrics of average expenditure per student, and a binary variable determining if average teacher pay is above or below the average income for the state.

I will first review the relationships between my variables by reviewing the correlation, coefficient of determination, and covariance between relevant variables.

I will then explore the data further with histograms and scatterplots with regression lines laid over them to understand the scope of the data.

Additionally, I will create a linear model and use this to create predictions pertaining to how financial variables might effect the percent of bachelor obtainment.

Finally, I will produce a quadratic model to determine if there is a point where the percent of bachelor obtainment flattens out regardless of how much financial support is available.

Analysis.

First, I will review the correlation between the variables to determine if the variables increase/decrease together or in opposing directions and the strength of their relationships.

Next, I will be calculating the Coefficient of Determination to determine what percentage of the variability in one variable can be attributed to another variable.

Additionally, I will also be calculating the covariance between the variables to determine the strength of their relationships.

Each of these measurements will be evaluated for the relevant variables below. I will summarize and detail my findings after the calculations.

~RELATIONSHIPS BETWEEN VARIABLES~

```
#Prct_Bachelor/Avg_Exp_Per_Stud
cor(final_df$Prct_Bachelor, final_df$Avg_Exp_Per_Stud)
```

```
## [1] 0.5838069
```

```r
cor(final_df$Prct_Bachelor, final_df$Avg_Exp_Per_Stud)^2
```

```
## [1] 0.3408305
```

```r
cov(final_df$Prct_Bachelor, final_df$Avg_Exp_Per_Stud)
```

```
## [1] 13.4312
```

```r
#Prct_Bachelor/Teacher_Pay
cor(final_df$Prct_Bachelor, final_df$Teacher_Pay)
```

```
## [1] 0.6456914
```

```r
cor(final_df$Prct_Bachelor, final_df$Teacher_Pay)^2
```

```
## [1] 0.4169174
```

```r
cov(final_df$Prct_Bachelor, final_df$Teacher_Pay)
```

```
## [1] 36110.81
```

```r
#Prct_Bachelor/Avg_Income
cor(final_df$Prct_Bachelor, final_df$Avg_Income)
```

```
## [1] 0.7836444
```

```r
cor(final_df$Prct_Bachelor, final_df$Avg_Income)^2
```

```
## [1] 0.6140985
```

```r
cov(final_df$Prct_Bachelor, final_df$Avg_Income)
```

```
## [1] 35813.67
```

```r
#Avg_Exp_Per_Stud/Teacher Pay
cor(final_df$Avg_Exp_Per_Stud, final_df$Teacher_Pay)
```

```
## [1] 0.7640131
```

```r
cor(final_df$Avg_Exp_Per_Stud, final_df$Teacher_Pay)^2
```

```
## [1] 0.583716
```

```
cov(final_df$Avg_Exp_Per_Stud, final_df$Teacher_Pay)
```

```
## [1] 32349.51
```

```
#Avg_Exp_Per_Stud/tp_vs_inc
biserial.cor(final_df$Avg_Exp_Per_Stud, final_df$tp_vs_inc)
```

```
## [1] -0.1166065
```

```
biserial.cor(final_df$Avg_Exp_Per_Stud, final_df$tp_vs_inc)^2
```

```
## [1] 0.01359708
```

```
cov(final_df$Avg_Exp_Per_Stud, final_df$tp_vs_inc)
```

```
## [1] 0.2328742
```

Relationship Between Variables Summary Table

Variable1	Variable2	Correlation	Coefficient of Determination	Covariance
Prct_Bachelor	Avg_Exp_Per_Stud	0.5838069	0.3408305	13.4312
Prct_Bachelor	Teacher_Pay	0.6456914	0.4169174	36110.81
Prct_Bachelor	Avg_Income	0.7836444	0.6140985	35813.67
Avg_Exp_Per_Stud	Teacher_Pay	0.7640131	0.583716	32349.51
Avg_Exp_Per_Stud	tp_vs_inc	-0.1166065	0.01359708	0.2328742

Based on the above correlation values, it appears that Prct_Bachelor/Avg Income and Avg_Exp_Per_stud/Teacher_Pay are the variables that have the strongest correlation. One can infer that there is some kind of relationship where either higher Percent of Bachelors contributes to a higher Average Income or vise versa and higher Average Expenditure per Student results in higher Teacher Pay or vice versa. I ran a biserial correlation on Avg_Exp_Per_Stud/tp_vs_inc because tp_vs_inc is a binary TRUE/FALSE variable. This is the only correlation that resulted in a negative correlation.

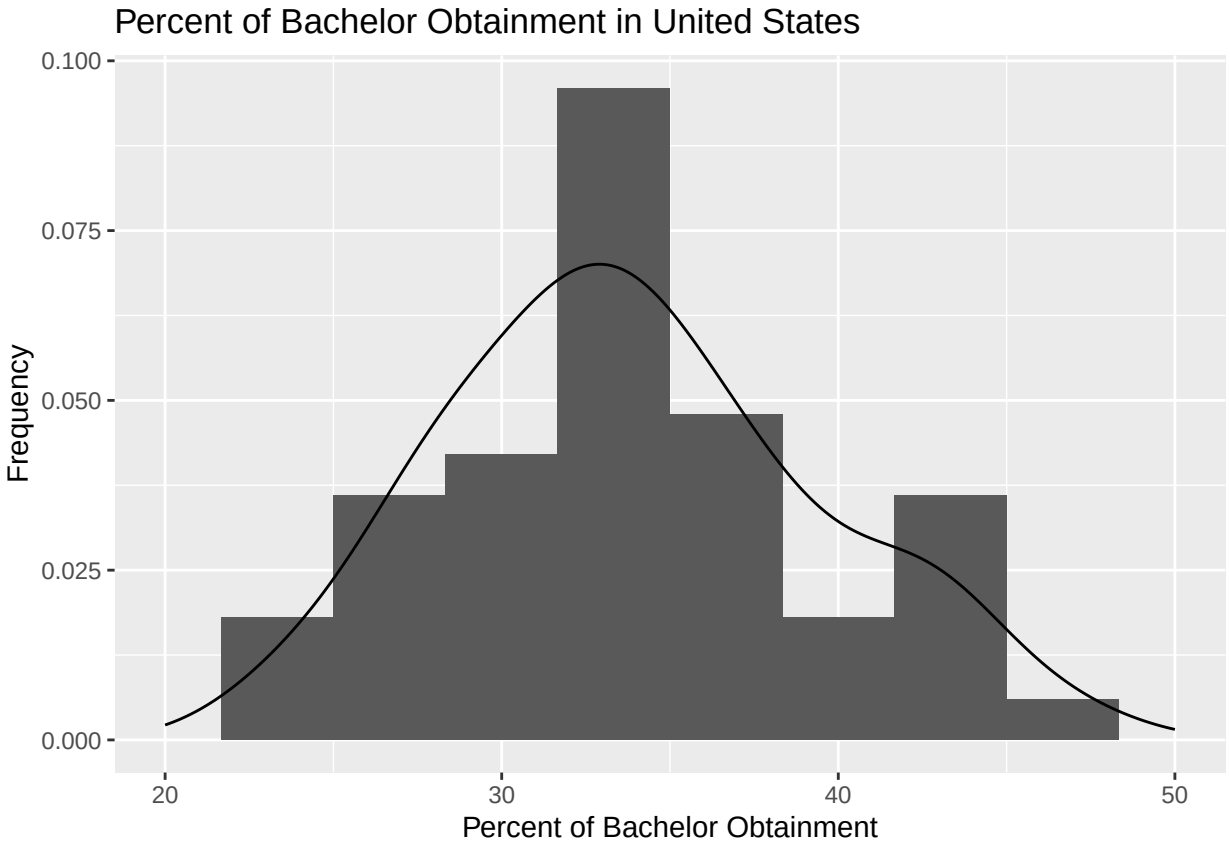
Based on the above coefficient of determination values, we can say that 61.4% of the variation in Prct_Bachelor can be explained by Avg_Income. The lowest coefficient of determination is Avg_Exp_Per_Stud/tp_vs_inc. This is incredibly low which indicates that one variables does not predict the other. If you know the percent of bachelor obtainment that a state has, it doesn't provide any insight into if their teachers are paid above or below the average income for the area and vice versa.

All of the above covariance values are positive indicating that as one variable increases, the other variables will increase as well. Prct_Bachelor/Teacher_Pay has the strongest covariance followed closely by Prct_Bachelor/Avg_Income and Avg_Exp_Per_Stud/Teacher Pay.

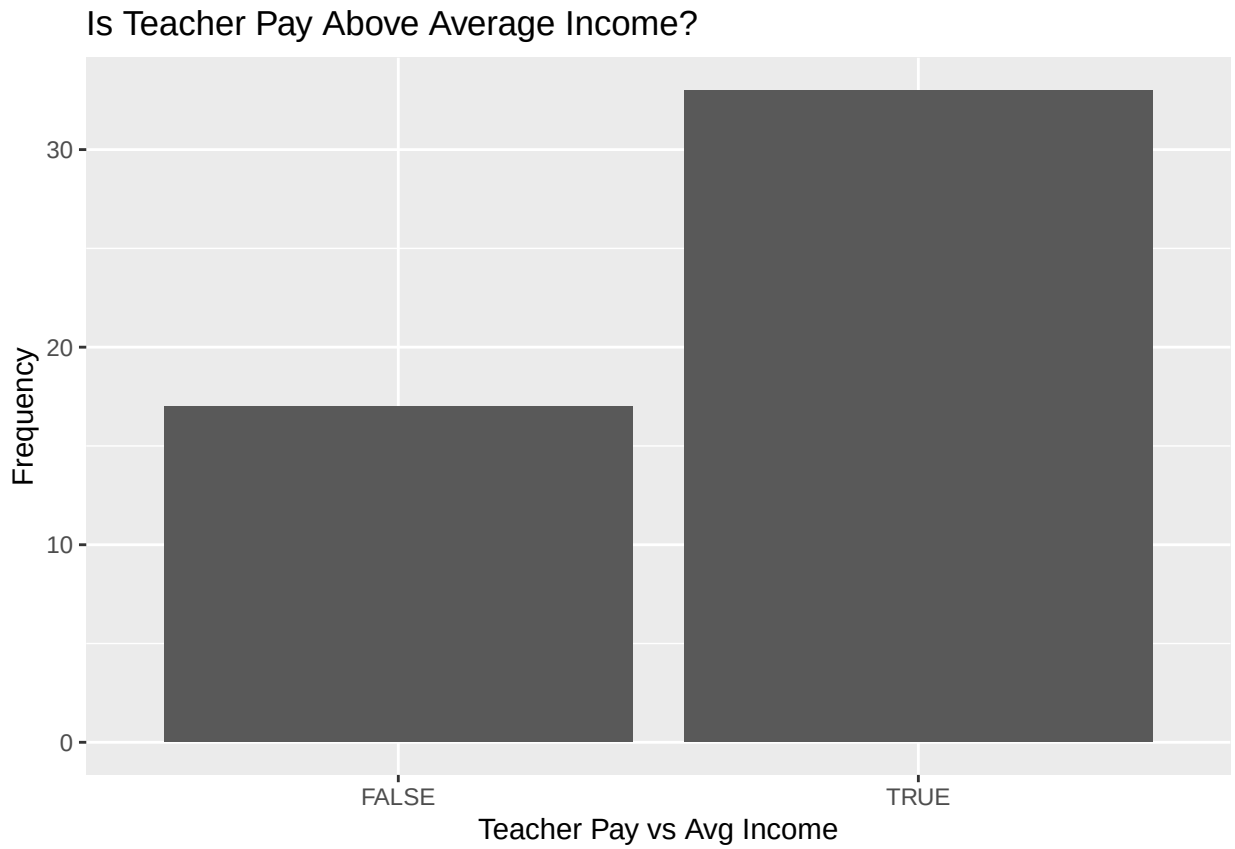
HISTOGRAMS

In an effort to get a better overall understanding of the data, I will be creating two histograms. The first histogram will display the frequency at which various percentages of bachelor obtainment can be found among the 50 states. This histogram will have a kernel density plot overlayed on top of it. The second histogram will provide context as to how many states pay their teachers above the average income or below the average income.

```
## Warning: Removed 2 rows containing missing values (`geom_bar()`).
```



The above histogram shows that the percent of bachelor obtainment has a tail on the right indicating a positive skew. It also appears that the mode percent of bachelor obtainment is around 33-34% and the spread is roughly 22%-48% bachelor obtainment.



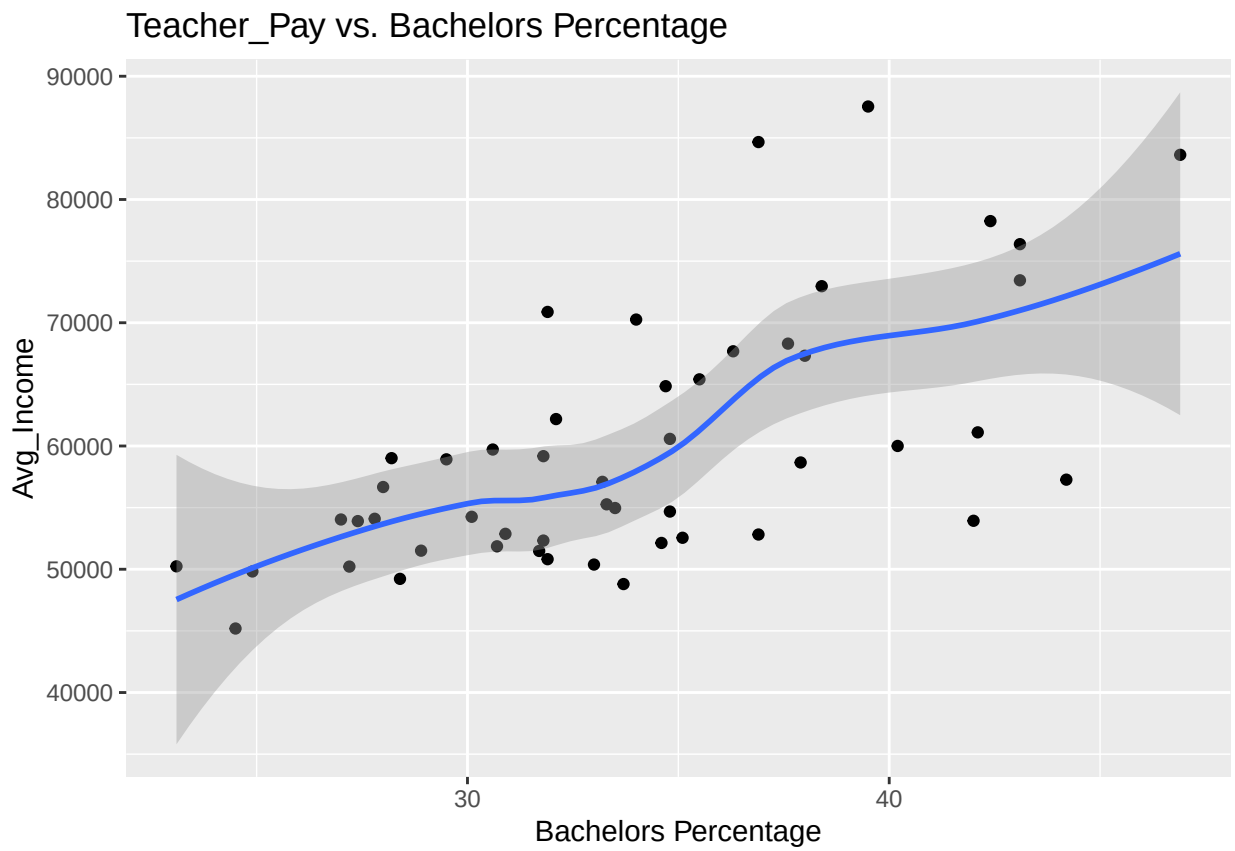
The above histogram shows that in a majority of states, teacher's make above the average income. I anticipate that teacher pay and average income will have a close relationship.

~SCATTERPLOT/REGRESSION LINES~

To further investigate the relationship between the Percent of bachelor obtainment and the various financial variables, I will be reviewing several graphs with trend lines. Analysis will follow the graphs.

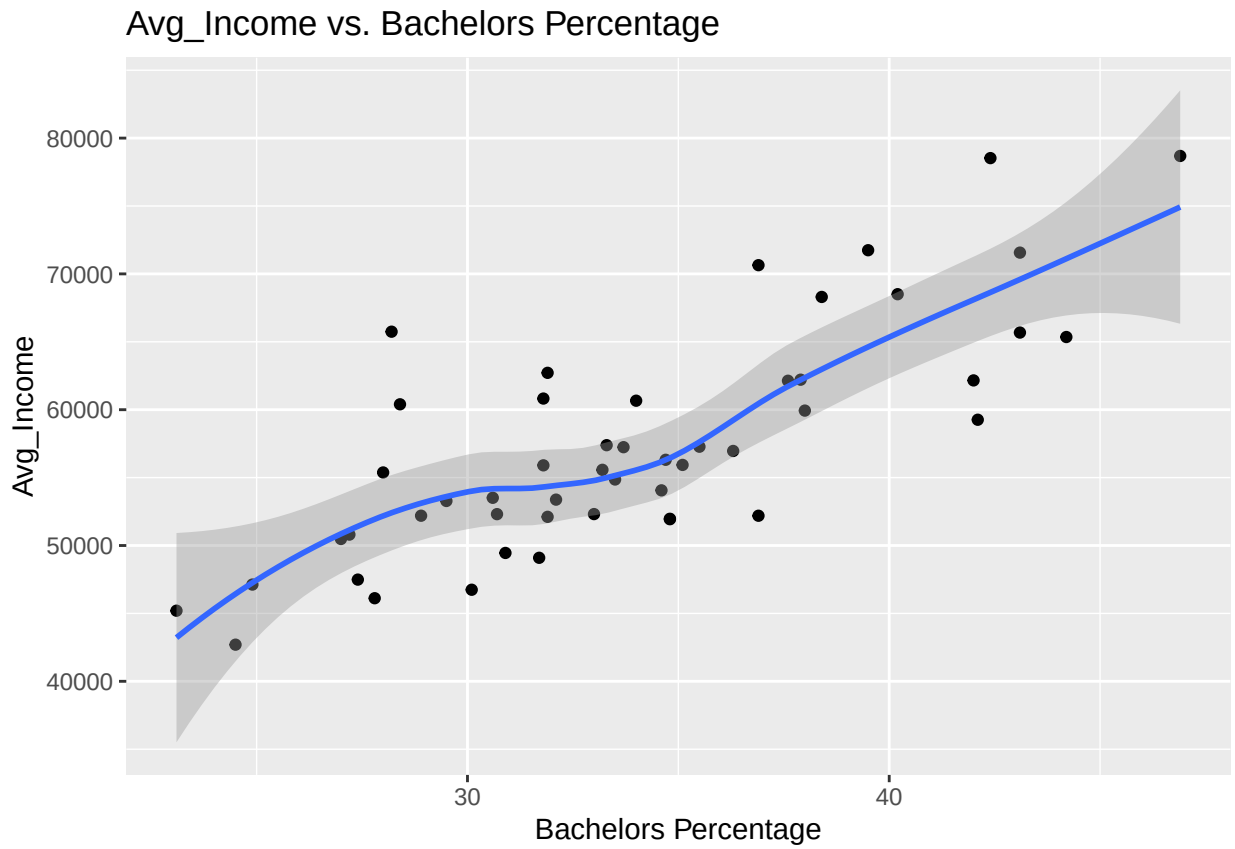
Teacher_Pay/Prct_Bachelor

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



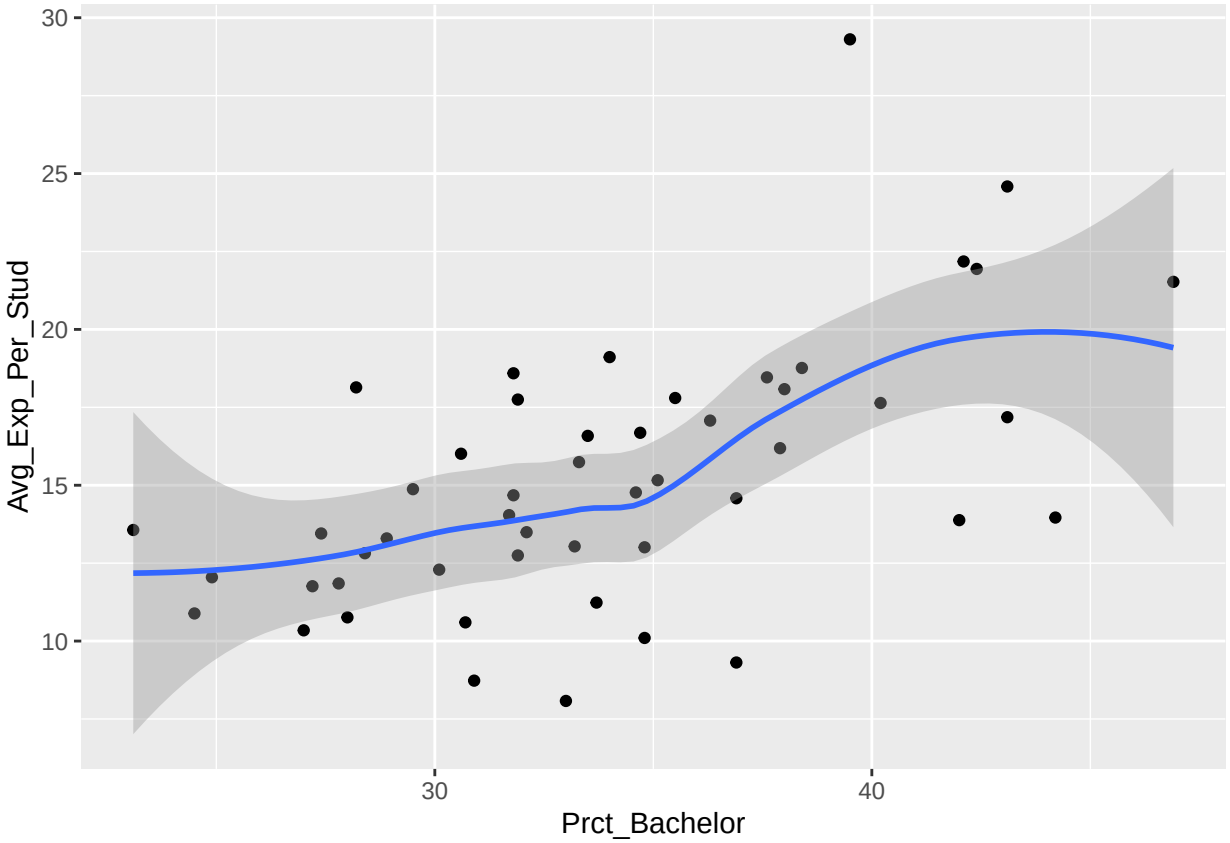
Avg_Income/Prct_Bachelor

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



Avg_Exp_Per_Stud/Prct_Bachelor

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



Based on the above graphs, it appears that Teacher_Pay/Prct_Bachelor and Avg_Income/Prct_Bachelor have a positive relationship where both variables increase at the same time, while Expenditure/Prct_Bachelor has a relatively flat regression line to begin with before a slight slope and curve.

~LINEAR MODEL~

Because we are primarily focusing on what financial variables impact a state's percent of bachelor obtainment, I will fit a linear model using Prct_Bachelor as outcome and the Expenditure, Teacher_Pay, and Avg_Income variables as predictors.

```
linear_model <- lm(Prct_Bachelor ~ Avg_Exp_Per_Stud + Teacher_Pay + Avg_Income, data =
final_df)
summary(linear_model)
```

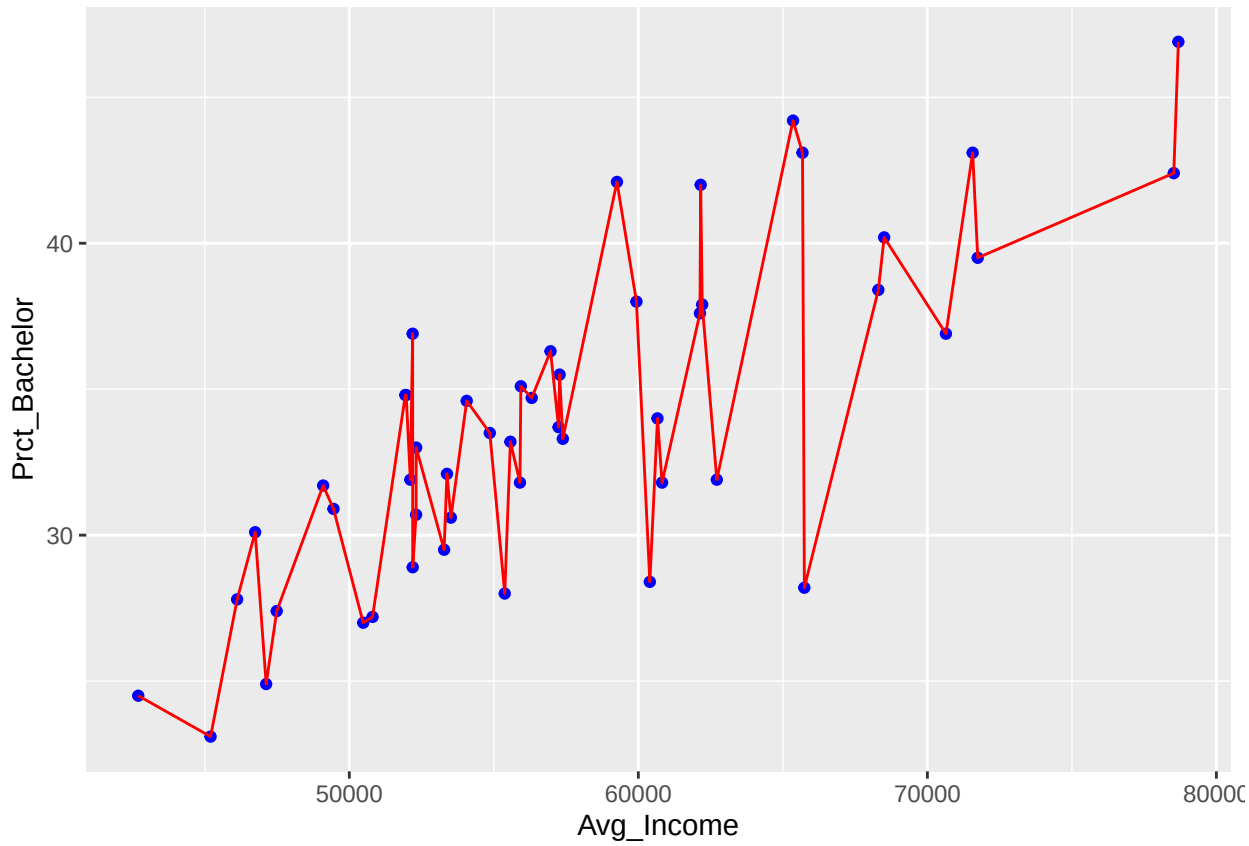
```
##
## Call:
## lm(formula = Prct_Bachelor ~ Avg_Exp_Per_Stud + Teacher_Pay +
##     Avg_Income, data = final_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6931 -2.1434  0.2644  1.8456  7.4142
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.693e+00  3.801e+00   0.972    0.336
## Avg_Exp_Per_Stud -1.978e-02  1.986e-01  -0.100    0.921
```

```
## Teacher_Pay       3.551e-05  9.109e-05   0.390    0.699
## Avg_Income        4.938e-04  1.068e-04   4.625 3.06e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.528 on 46 degrees of freedom
## Multiple R-squared:  0.6154, Adjusted R-squared:  0.5903
## F-statistic: 24.54 on 3 and 46 DF,  p-value: 1.245e-09
```

Based on the above p-values, it appears that Avg_Income is the only variable that has a statistically significant p-value less than .05, although the overall p-value of our model would also be considered statistically significant. All variables display very low standard error values and our overall model has a low residual value which tells us that there is little uncertainty in our estimations. The R-squared value tells us that 61.54% of the variability in Prct_Bachelor can be explained by these variables. This is very close to the corrective Adjusted R-squared value of 59.03%.

I've included an additional graph below plotting Prct_Bachelor vs Average Income. A version of this comparison has already been graphed above, but since Avg_Income is most closely linked to Prct_Bachelor it made sense to isolate another version of this graph for review. The below graph shows that while not a linear fit, overall as Avg_Income increases Prct_Bachelor increases as well.

```
#Plot the predictions against the original data
ggplot(data = final_df, aes(y = Prct_Bachelor, x = Avg_Income)) +
  geom_point(color='blue') +
  geom_line(color='red',data = final_df, aes(y= Prct_Bachelor, x= Avg_Income))
```



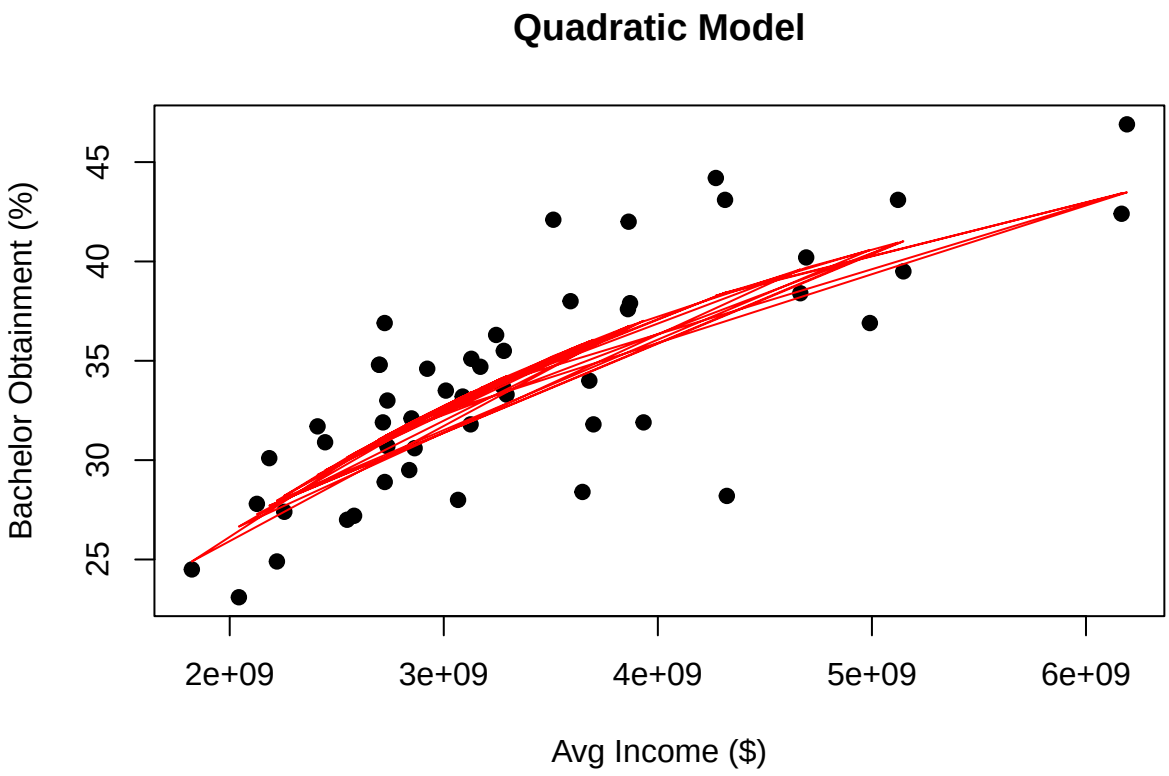
QUADRATIC MODEL

The final question posed, was if there was ever a point where regardless of how much financial support is provided, is there a point where the percent of bachelor obtainment flattens out. To research this, I have developed and graphed a quadratic model below. The quadratic model does not rise and fall like a parabola, but the slope does seem to begin the flattening out process on the right side of the graph indicating that there will be a point where regardless of how much the average income is in a state, the percent of bachelor obtainment will not go above a certain level.

```
final_df$Avg_Income2 <- final_df$Avg_Income^2
quad_model <- lm(Prct_Bachelor ~ Avg_Income + Avg_Income2, data =
final_df)
summary(quad_model)
```

```
##
## Call:
## lm(formula = Prct_Bachelor ~ Avg_Income + Avg_Income2, data = final_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.2571  -1.8800   0.3486   2.2455   6.8596
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.585e+01  2.069e+01  -0.766   0.4474
## Avg_Income   1.192e-03  6.938e-04   1.718   0.0923 .
## Avg_Income2 -5.568e-09  5.735e-09  -0.971   0.3366
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.462 on 47 degrees of freedom
## Multiple R-squared:  0.6217, Adjusted R-squared:  0.6056
## F-statistic: 38.62 on 2 and 47 DF,  p-value: 1.201e-10
```

```
quadratic = quad_model$coefficient[3]*final_df$Avg_Income2 + quad_model$coefficient[2]*final_df$Avg_Inc
plot(final_df$Avg_Income2, final_df$Prct_Bachelor, main="Quadratic Model",
xlab="Avg Income ($)", ylab="Bachelor Obtainment (%)", pch=19)
par(new = TRUE)
lines(final_df$Avg_Income2,quadratic, col="red")
```



Implications.

The above findings tell us that the percent of people earning a bachelor in each state and the average income in each state have the greatest impact on one another. It's difficult to tell if the average income is higher because jobs requiring a bachelors can trend towards paying at a higher rate, or if the percent receiving a bachelors in each state is higher because it is expensive to obtain a bachelors so it's more accessible to those with a higher average income.

Average teacher pay is also closely associated with percent of bachelor obtainment, which makes sense because the average teacher pay and average income are also very closely related.

Average state education expenditure per student is the least closely associated with the percent of bachelor obtainment which I find interesting and would be curious about doing a deeper dive into how the various states are spending their education funds.

To answer the specific questions raised in step 1 of this project:

1) How does average teacher pay effect bachelor obtainment? The higher the average teacher pay the higher percent of bachelor obtainment. However, this is more likely due to the cost of living/average income in each state more than the average teacher pay.

2) How does state education expenditure effect bachelor obtainment? The average state expenditure per student does not appear to have a large effect on bachelor obtainment.

3) How does the average household income effect bachelor obtainment? This appears to be the strongest relationship. The higher the average income is the more likely

4) How does state education expenditure effect teacher pay? Average state education expenditure per student has a high correlation with teacher pay.

5) At a certain point does bachelor obtainment flatten out regardless of other variables? The only variable that proved statistically significant to bachelor obtainment is Average Income. The above quadratic model shows that there will be a point where regardless of how large the average income is in a state, the percent of bachelor obtainment will level out.

Limitations.

This data is only providing a small subset of relevant data. There could be non-financial reasons as to why someone would not wish to pursue a bachelors degree such as a goal to work in a profession that requires alternate training or skills. It would be interesting to include what the largest industry is in each state to determine how this would effect the data.

There are also two variables, Avg_Exp_Per_Stud and tp_vs_inc, that were created based on data from different sources. These variables were created to smooth out the data, but it would have been more ideal if these variables were using data from the same source to perform their calculations to ensure better accuracy.

Because we are comparing k-12 data and how it effects bachelor obtainment, it would be useful to know what percent of the population obtained their bachelor out of state or no longer reside in the state where they received their k-12 education.

Concluding Remarks

Overall, it appears that the more money that is in a community (average income, teacher pay) the more likely a student is to obtain a bachelors degree. The amount of money the state spends on education would be a poor indicator of how many students go on to obtain a bachelors.