Lab 4 - Whats the Explanation?

If you haven't already, be sure to install `rmarkdown`

You'll also want to install `mosaicData`.

You can do both in this code chunk! Just make sure that after you run it, you delete it before trying to knit your file.

```
#after running, be sure to delete this code chunk before knitting!
#install.packages("rmarkdown")
#install.packages("mosaicData")
```

Part 1 - Explaining variation in number of births.

Question 1

Create a histogram of the `births` variable (using ggplot2). Your histogram should:

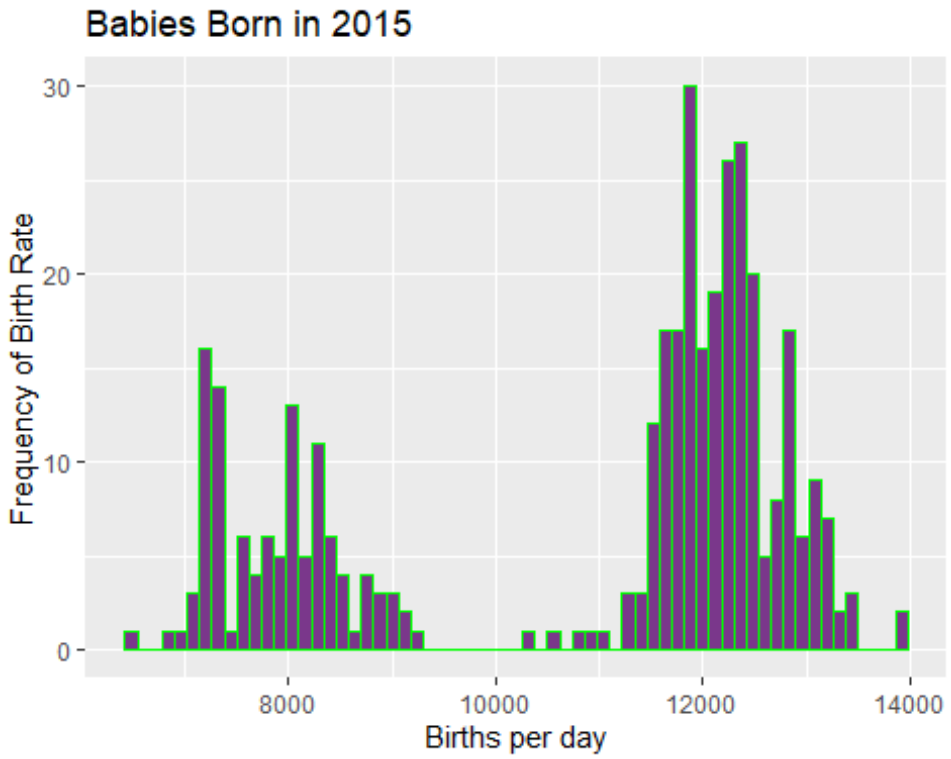
- Use a fill color of your choice
- Define a border color to better define the bins
- Add an appropriate title
- Adjusting number of bins is optional.

Also report the results from the summary function when summarizing that variable.

Include the image of your histogram.

Include your R code for this question.

```
#write your ggplot code here. Use the play button on the right to preview
OldBirths <- Births2015

ggplot(data=OldBirths, aes(x=births)) + geom_histogram(color='green',
fill='mediumorchid4', bins=80, binwidth = 120) + labs(stat="identity", title
= "Babies Born in 2015", x = "Births per day", y = "Frequency of Birth Rate")
```



Include the numeric summary output

```
#code summary here
cat("\014")
```

```
shell("cls")
OldBirths_summary <- summary(OldBirths$births)
sprintf("The summary statistic for the births in 2015  - minimum, 1st
Quartile, Median, Mean, 3rd Quartile, and Max - is: %f", OldBirths_summary)
```

```
## [1] "The summary statistic for the births in 2015  - minimum, 1st
Quartile, Median, Mean, 3rd Quartile, and Max - is: 6515.000000"
## [2] "The summary statistic for the births in 2015  - minimum, 1st
Quartile, Median, Mean, 3rd Quartile, and Max - is: 8431.000000"
## [3] "The summary statistic for the births in 2015  - minimum, 1st
Quartile, Median, Mean, 3rd Quartile, and Max - is: 11883.000000"
## [4] "The summary statistic for the births in 2015  - minimum, 1st
Quartile, Median, Mean, 3rd Quartile, and Max - is: 10899.991781"
## [5] "The summary statistic for the births in 2015  - minimum, 1st
Quartile, Median, Mean, 3rd Quartile, and Max - is: 12339.000000"
## [6] "The summary statistic for the births in 2015  - minimum, 1st
Quartile, Median, Mean, 3rd Quartile, and Max - is: 13949.000000"
```

Briefly describe what you see. How would you describe the shape of this distribution? Is it what you expected, or is it difficult for you to explain?

(write your response here!)

```
print("The shape of this distribution resembles two normally distributed
graphs split in half")
```

```
## [1] "The shape of this distribution resembles two normally distributed
graphs split in half"
```

(now is a good time to try knitting to a word doc to see if everything works! All of your code will need to be running though).

Question 2

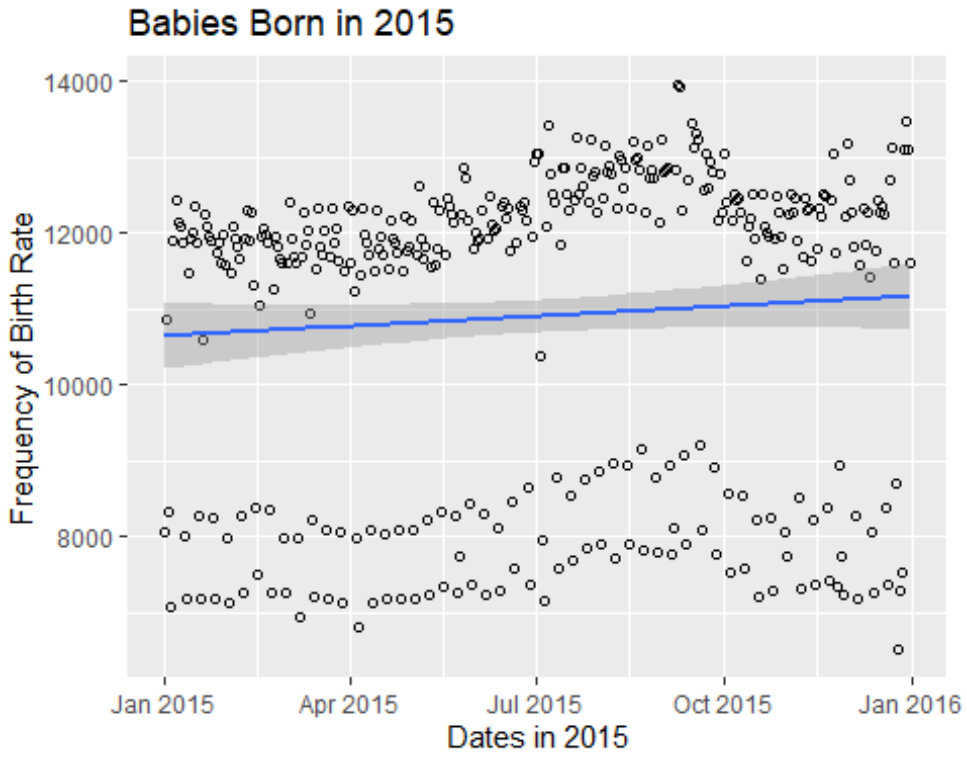
Next, let's try creating a scatterplot with `births` on the y axis and `date` on the x axis. This will help us see if the time of year might explain some of the variation we see in number of births.

Include the image of your histogram with an appropriate title.

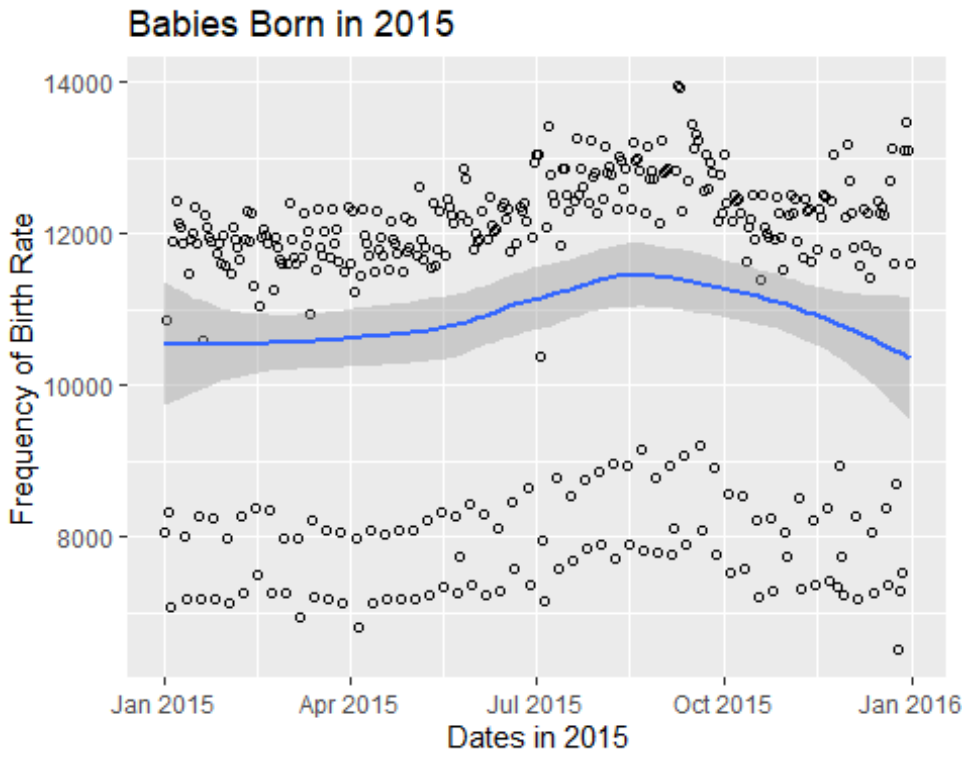
Include your R code for this question.

```
cat("\014")
```

```
shell("cls")
ggplot(OldBirths, aes(x=date, y=births)) + geom_point(shape=1) +
geom_smooth(method=lm) + labs(stat="identity", title = "Babies Born in 2015",
x = "Dates in 2015", y = "Frequency of Birth Rate")

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




```
# Add linear regression line
# By default includes 95% confidence region
ggplot(OldBirths, aes(x=date, y=births)) + geom_point(shape=1) +
geom_smooth() + labs(stat="identity", title = "Babies Born in 2015", x =
"Dates in 2015", y = "Frequency of Birth Rate")

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




```
# Use hollow circles
# Add a local regression smoothed fit curve with confidence region
```

from this point out, make your own R chunks using the + icon a little to the left of the "run" button up top.

Briefly describe what you see. What trends do you see between time of year and number of births? Does time of year explain most of the variability in births, or do you think there is still a lot of variability leftover?

```
print("Using both linear and local polynomial regession curves shows that the number of births increases in the later months of the year, with a visualized local maximum around July-October, though the variability is in question")
```

```
## [1] "Using both linear and local polynomial regession curves shows that
the number of births increases in the later months of the year, with a
visualized local maximum around July-October, though the variability is in
question"
```

Question 3

Using the data viewer, browse some of the other variables in this dataset and try to find a variable that explains why there are 2 distinct modes to the births variable.

Create a visualization with your 'best variable' on the x axis and the births variable on the y axis.

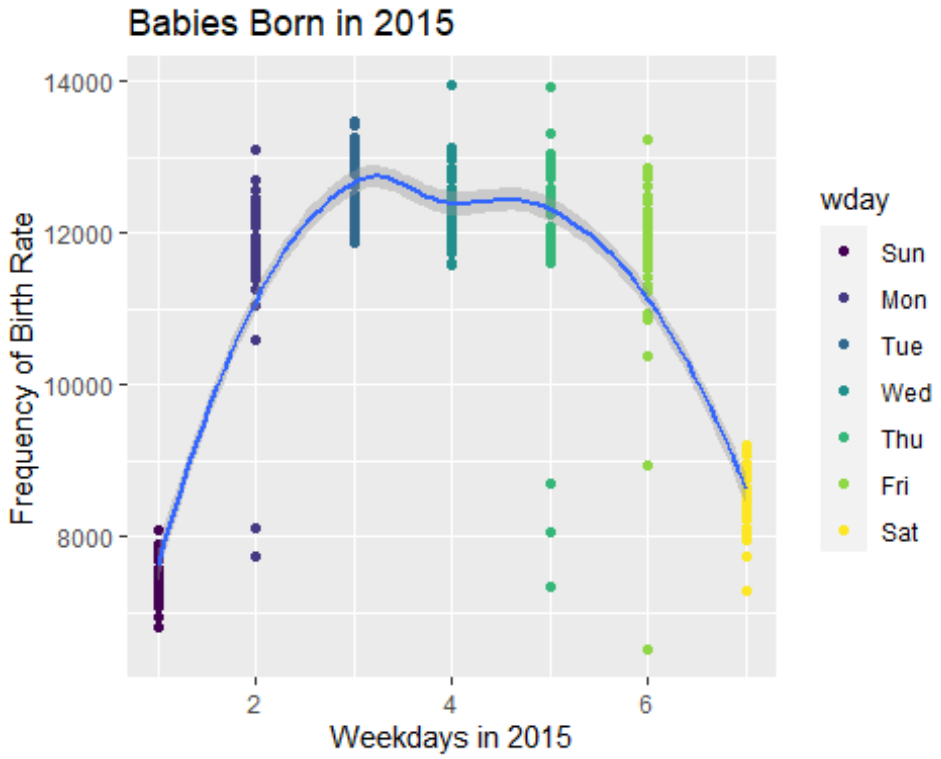
Include an image of your visualization with an appropriate title.

Include your R code for this question.

```
cat("\014")
```

```
shell("cls")
ggplot(OldBirths, aes(x=day_of_week, y=births)) + geom_point(aes(color=wday))
+ geom_smooth() + labs(stat="identity", title = "Babies Born in 2015", x =
"Weekdays in 2015", y = "Frequency of Birth Rate")
```

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



Briefly describe in context why this variable explains the bimodal variability in births. Feel free to do some internet searching for some insight if you're not sure!

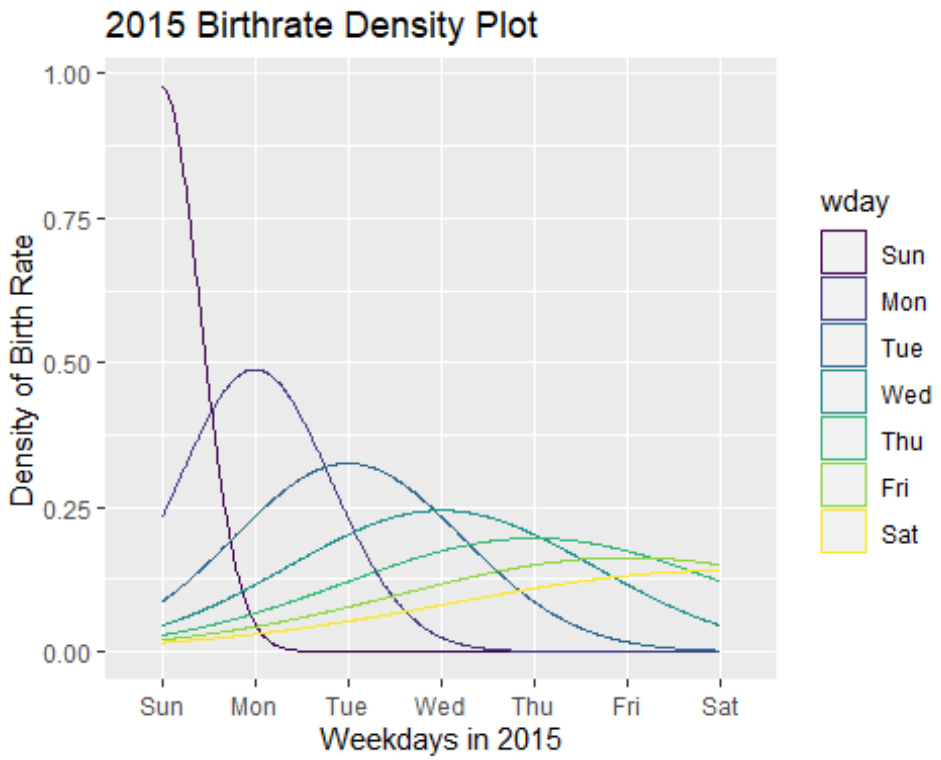
```
print("This graph of Birthrate vs Weekdays shows that we have a visual mode
of Births on Tuesdays as well as Wednesday/Thursdays, but the variability of
births per weekday is still in question since the following plot shows a
difference in birthrate densities per weekday. This second appearance of a
mode may explain the bimodal variability in births")
```

```
## [1] "This graph of Birthrate vs Weekdays shows that we have a visual mode
of Births on Tuesdays as well as Wednesday/Thursdays, but the variability of
births per weekday is still in question since the following plot shows a
difference in birthrate densities per weekday. This second appearance of a
mode may explain the bimodal variability in births"
```

```
ggplot(OldBirths, aes(x = wday, fill = births, color=wday)) +
geom_density(alpha = 0.3) + labs(stat="identity", title = "2015 Birthrate
Density Plot", x = "Weekdays in 2015", y = "Density of Birth Rate")
```

```
## Warning: The following aesthetics were dropped during statistical
transformation: fill
```

```
## ℹ This can happen when ggplot fails to infer the correct grouping
structure in
##   the data.
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```



Part 2 - Explaining variation in SAT Scores.

Question 4

Create a histogram of the `sat` variable. Your histogram should:

- Use a fill color of your choice
- Define a border color to better define the bins
- Add an appropriate title
- Adjusting number of bins is optional

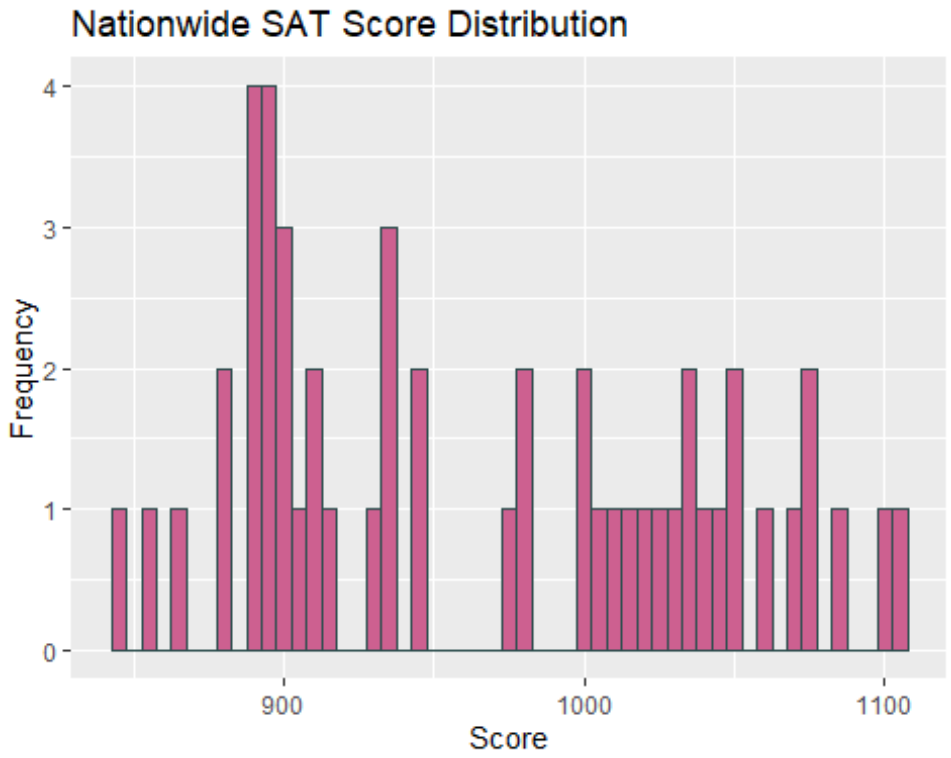
Also report the results from the summary function when summarizing that variable.

Include the image of your histogram.

Include your R code for this question.

```
cat("\014")
```

```
shell("cls")
#?SAT
#head(SAT)
#str(SAT)

sat_data <- SAT
colnames(sat_data)[1] ="district"
#str(sat_data)
#head(sat_data)
#hist(sat_data$sat)
#ggplot(data=sat_data, aes(x=district)) +
geom_histogram(color='darkslategray', fill='hotpink3', bins=25, binwidth = 5)
+ labs(title = "Nationwide SAT Score Distribution", x = "State", y = "SAT
score")
ggplot(data=sat_data, aes(x=sat)) + geom_histogram(color='darkslategray',
fill='hotpink3', bins=25, binwidth = 5) + labs(title = "Nationwide SAT Score
Distribution", x = "Score", y = "Frequency")
```



Include the numeric summary output

```
#by(sat_data, sat_data$district, summary)
print("The summary statistics for the 'sat' variable is as follows:")

## [1] "The summary statistics for the 'sat' variable is as follows:"

summary(sat_data$sat)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   844.0   897.2   945.5   965.9  1032.0  1107.0

std_dev <- sd(sat_data$sat)
var_sat <- var(sat_data$sat)
```

Briefly describe what you see. Is this an expected amount of variability across state averages, or is this more or less variability than you expected?

```
sprintf("The standard deviation is %f and the variance is %f", std_dev,
var_sat)

## [1] "The standard deviation is 74.820558 and the variance is 5598.115918"

print("This is less than the variability I was expecting since I am only
familiar with the grading scheme of the SAT exam from 2016 onwards, and
despite this the nationwide averages surprise me since they do not approach a
normally-distributed form as the Central Limit Theorem suggests.")

## [1] "This is less than the variability I was expecting since I am only
familiar with the grading scheme of the SAT exam from 2016 onwards, and
despite this the nationwide averages surprise me since they do not approach a
normally-distributed form as the Central Limit Theorem suggests."
```

Question 5

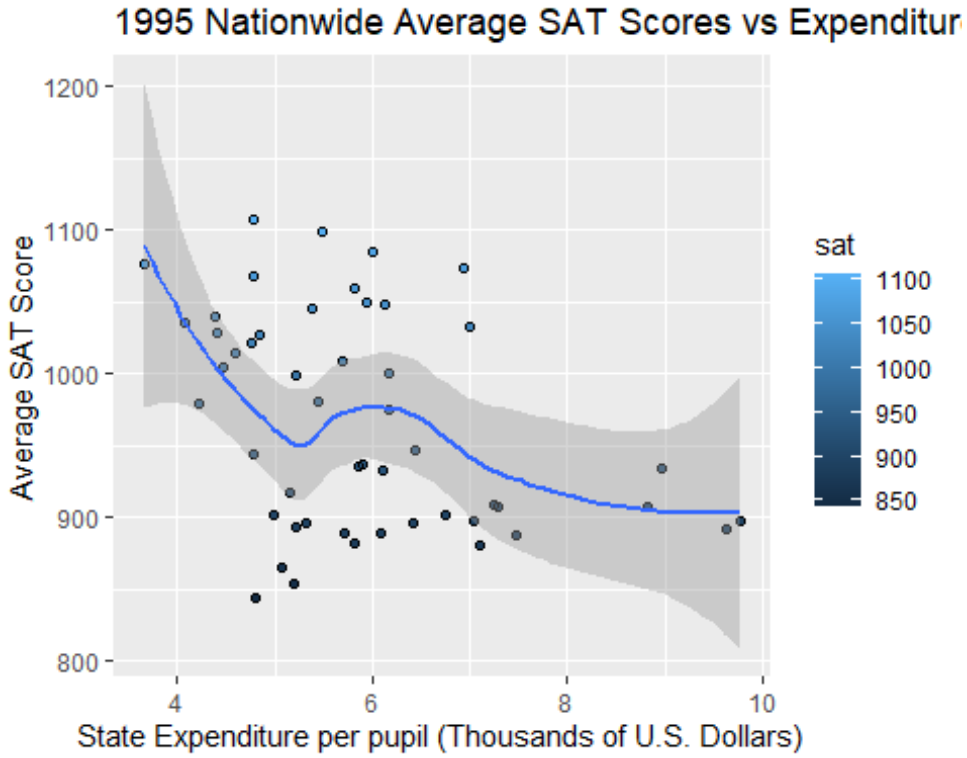
Create a scatterplot with each of the three variables: `expend`, `ratio`, `salary`, as the predictor variable (x axis) and `sat` listed as the response variable (y axis). **with R code for each Include the image of your scatterplot - expend**

Include the image of your scatterplot - ratio

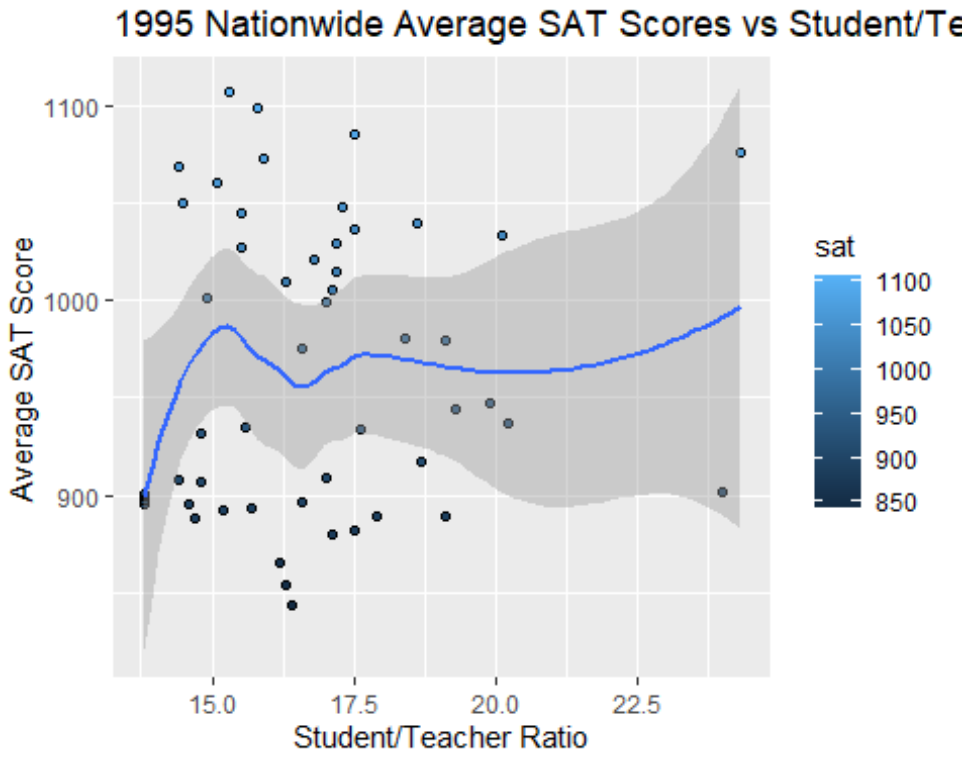
Include the image of your scatterplot - salary

```
cat("\014")
```

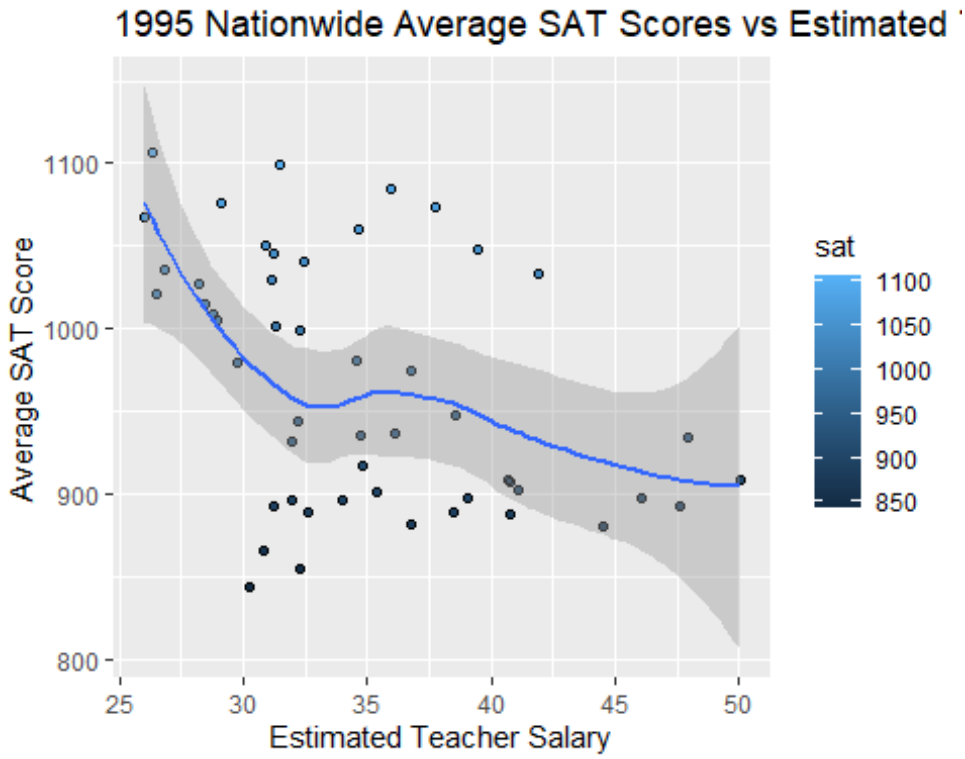
```
shell("cls")

#geom_smooth(method=lm, se=FALSE, fillrange=TRUE)

ggplot(sat_data, aes(x=expend, y=sat)) + geom_point(aes(color=sat)) +
geom_point(shape=1) + geom_smooth() + labs(stat="identity", title = "1995
Nationwide Average SAT Scores vs Expenditure per Pupil", x = "State
Expenditure per pupil (Thousands of U.S. Dollars)", y = "Average SAT Score")

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




```
ggplot(sat_data, aes(x=ratio, y=sat)) + geom_point(aes(color=sat)) +
geom_point(shape=1) + geom_smooth() + labs(stat="identity", title = "1995
Nationwide Average SAT Scores vs Student/Teacher Ratio", x = "Student/Teacher
Ratio", y = "Average SAT Score")

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




```
ggplot(sat_data, aes(x=salary, y=sat)) + geom_point(aes(color=sat)) +
geom_point(shape=1) + geom_smooth() + labs(stat="identity", title = "1995
Nationwide Average SAT Scores vs Estimated Teacher Salary", x = "Estimated
Teacher Salary", y = "Average SAT Score")

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



Briefly describe what you see. What is the relationship between these variables? Do these variable relationships seem expected to you or surprising?

```
print("Using 'geom_smooth(method=lm, se=FALSE, fillrange=TRUE)' to generate
linear regression features shows that there is a negative relationship in the
first and thrid graphs, yet a positive slope in the second graph.")
```

```
## [1] "Using 'geom_smooth(method=lm, se=FALSE, fillrange=TRUE)' to generate
linear regression features shows that there is a negative relationship in the
first and thrid graphs, yet a positive slope in the second graph."
```

```
print("In addition, local polynomial/non-parametric regression agrees in that
that there is no graphically demonstrated benefit in Increased State
Expenditure per Pupil, nor in increased Teacher Salary. However, there
appears to be a positive relationship between Student/Teacher Ratio and
Average SAT Score.")
```

```
## [1] "In addition, local polynomial/non-parametric regression agrees in
that that there is no graphically demonstrated benefit in Increased State
Expenditure per Pupil, nor in increased Teacher Salary. However, there
appears to be a positive relationship between Student/Teacher Ratio and
Average SAT Score."
```

```
print("This is surprising as the modern conclusion is generally that as State
Expenditure per Pupil and Teacher Salary increases so does Student
Performance, and that a lower Student/Faculty ratio leads to better test
scores.")
```

```
## [1] "This is surprising as the modern conclusion is generally that as
State Expenditure per Pupil and Teacher Salary increases so does Student
Performance, and that a lower Student/Faculty ratio leads to better test
scores."
```

Question 6

First: look at the data viewer and sort the data by frac (click on the column header). We're looking for a noticeable gap where we can separate the High percentage states and Low percentage states.

At what percentage might you sensibly choose to use as a cut-off (i.e., where there is a large gap)?

Second: Create a new variable in the SAT data frame called frac_bin which will now label each state as "High" if above the cut-off value and "Low" if below the cut-off value. Use an ifelse function to create this new variable and be sure to assign it to SAT$frac_bin.

Include the ifelse code you used to complete this.

```
cat("\014")
```

```
shell("cls")
avg_sat_score = mean(c(max(sat_data$frac), min(sat_data$frac)))
sat_frac_divider_ceil = subset(x = sat_data, subset = frac >= avg_sat_score)
sat_frac_divider_floor = subset(x = sat_data, subset = frac <= avg_sat_score)
sprintf("Dividing the SAT dataset into two subsets based on the average
percentage of all eligible students taking the SAT. This is obtained by
taking the average of the maximum and minimum 'frac' values in the SAT
dataset, %f and %f, to get %f", max(sat_data$frac), min(sat_data$frac),
avg_sat_score)
```

```
## [1] "Dividing the SAT dataset into two subsets based on the average
percentage of all eligible students taking the SAT. This is obtained by
taking the average of the maximum and minimum 'frac' values in the SAT
dataset, 81.000000 and 4.000000, to get 42.500000"
```

```
#ggplot(sat_frac_divider_ceil, aes(x=frac, y=sat)) +
geom_point(aes(color=frac)) + geom_point(shape=1) + geom_smooth()

#ggplot(sat_frac_divider_floor, aes(x=frac, y=sat)) +
geom_point(aes(color=frac)) + geom_point(shape=1) + geom_smooth()


sat_data$frac_bin = ifelse(sat_data$frac >= avg_sat_score, "1", "0")
```

Question 7

Now recreate the scatterplot with the expend variable on the x axis and `sat` on the y axis, but add a color aesthetic with `frac_bin`. **Include image of your plot here**

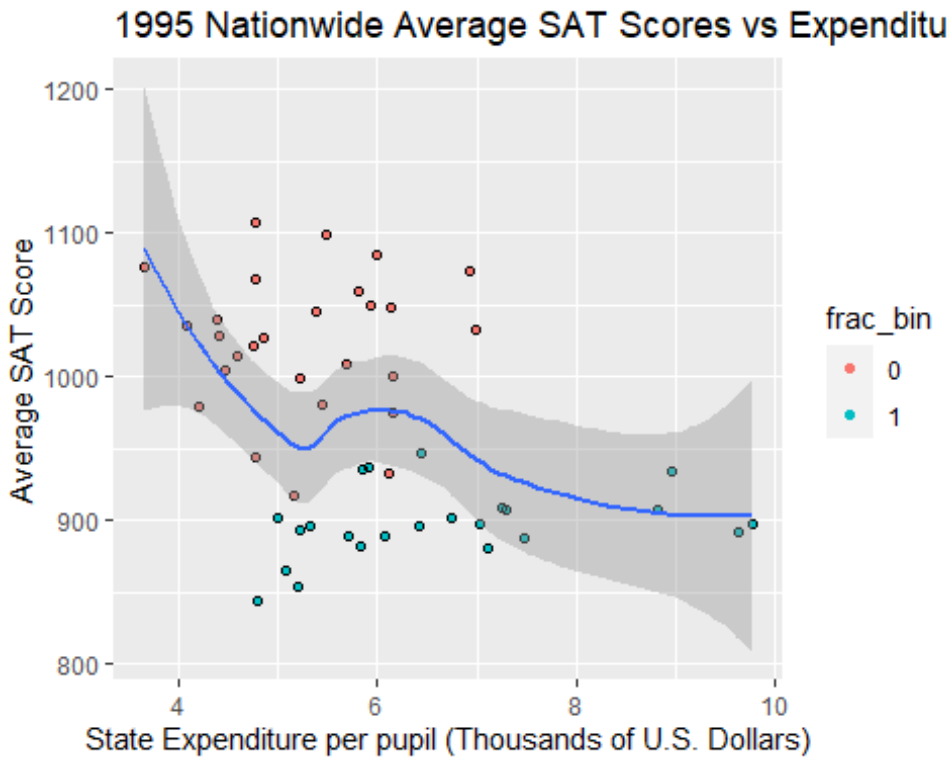
Briefly describe what you see.

- Is there any association between the percent of eligible students in a state taking the SAT and the state's avg score?
- If only focusing on states with a high fraction of students taking the SAT or only focusing on states with a low fraction, is there an association between expenditure and SAT scores?
- Do you have any ideas as to why the fraction of students taking the SAT in a state might explain score differences?

```
cat("\014")
```

```
shell("cls")
ggplot(sat_data, aes(x=expend, y=sat)) + geom_point(aes(color=frac_bin)) +
geom_point(shape=1) + geom_smooth() + labs(stat="identity", title = "1995
Nationwide Average SAT Scores vs Expenditure per Pupil, Relative to Average
Percentage of All Eligible Students Taking the SAT", x = "State Expenditure
per pupil (Thousands of U.S. Dollars)", y = "Average SAT Score")
```

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




```
print("There appears to be an association between the percent of eligible
students in a state taking the SAT and the state’s avg score wherein the
lower percentage of eligible students taking the SAT corresponds to a higher
average ")
```

```
## [1] "There appears to be an association between the percent of eligible
students in a state taking the SAT and the state’s avg score wherein the
lower percentage of eligible students taking the SAT corresponds to a higher
average "
```

```
print("- If only focusing on states with a high fraction of students taking
the SAT or only focusing on states with a low fraction, there does appear to
be a positive trend in that increased State Expenditure corresponds to higher
Average SAT scores.")
```

```
## [1] "- If only focusing on states with a high fraction of students taking
the SAT or only focusing on states with a low fraction, there does appear to
be a positive trend in that increased State Expenditure corresponds to higher
Average SAT scores."
```

```
print("We can see an example of Simpson’s Paradox - wherein a statistical relationship between two variables can be reversed by including additional factors in the analysis. By including the variable 'percentage of all eligible students taking the SAT'  , we saw a reversal of the relationship shown in the original scatterplot between SAT score and Expenditure per student.")
```

```
## [1] "We can see an example of Simpson’s Paradox - wherein a statistical relationship between two variables can be reversed by including additional factors in the analysis. By including the variable 'percentage of all eligible students taking the SAT'  , we saw a reversal of the relationship shown in the original scatterplot between SAT score and Expenditure per student."
```

```
print("The fraction of students taking the SAT in a state might explain score differences since a further analysis reveals that not all students take in SAT. Within states with the same percent of students taking the SAT, we see a positive relationship between Expenditure on Students and Average SAT Scores.")
```

```
## [1] "The fraction of students taking the SAT in a state might explain score differences since a further analysis reveals that not all students take in SAT. Within states with the same percent of students taking the SAT, we see a positive relationship between Expenditure on Students and Average SAT Scores."
```

```
print("When only a few students take the SAT in a state, these might only be the best students i.e. those who will score highest. If you have every student - meaning the whole student body upon which a State spends money on, you include both the high and low scoring student populations, bringing the average SAT score down as compared to states where there is only a small percentage of their 'best' students taking SAT. This is not a fair state-to-state comparison.")
```

```
## [1] "When only a few students take the SAT in a state, these might only be the best students i.e. those who will score highest. If you have every student - meaning the whole student body upon which a State spends money on, you include both the high and low scoring student populations, bringing the average SAT score down as compared to states where there is only a small percentage of their 'best' students taking SAT. This is not a fair state-to-state comparison."
```

```
print("Within similar states, i.e. similar percentage of students taking the SAT, we see that spending more money is associated with higher SAT scores. Therefore we must interpret the data's correlations with caution and context.")
```

```
## [1] "Within similar states, i.e. similar percentage of students taking the SAT, we see that spending more money is associated with higher SAT scores. Therefore we must interpret the data's correlations with caution and context."
```