Assignment 1

YOUR NAME HERE

Invalid Date

Questions

1. Using the `vtable` package, create a table of summary statistics from the `econmath` data that includes the mean, standard deviation, minimum, and maximum for variables: *score*, *hsgpa*, *study*, *age*.

```
econmath %>%
  select(score, hsgpa, study, age) %>%
  sumtable(summ=c('mean(x)','sd(x)','min(x)','max(x)'))
```

2. Compute the summary statistics table as you did in (1), but group the data by whether or not the student took a high school economics course. Comment on the differences across groups in the mean of these variables.

INSERT COMMENTS HERE

```
econmath %>%
  select(score, hsgpa, study, age, econhs) %>%
  sumtable(summ=c('mean(x)','sd(x)','min(x)','max(x)'),
           group = 'econhs')
```

Table 1: Summary Statistics

Variable	Mean	Sd	Min	Max
score	73	13	20	98
hsgpa	3.3	0.34	2.4	4.3
study	14	7.8	0	50
age	19	0.94	18	29

Table 2: Summary Statistics

econhs	0				1			
Variable	Mean	Sd	Min	Max	Mean	Sd	Min	Max
score	73	13	20	98	72	13	23	96
hsgpa	3.3	0.34	2.4	4.1	3.4	0.35	2.4	4.3
study	14	7.7	0	48	14	8	0	50
age	19	0.94	18	29	19	0.94	18	28

3. Using the `ggplot2` package, produce a violin plot of *score* across the two values of *econhs* (note, you may need to look up violin plots to familiarize yourself with them). Create a title for the graph and relabel the x and y axes with more intuitive names. Describe the relationship between these two variables. [NOTE: when you define the aesthetics in your plot, you will need to declare *econhs* as a factor variable using `as.factor(econhs)`]

```
ggplot(econmath, aes(y = hsgpa, x = as.factor(econhs))) +
  geom_violin() +
  labs(title = "Density of Economics Scores",
       x = "High School GPA", y = "Density")
```

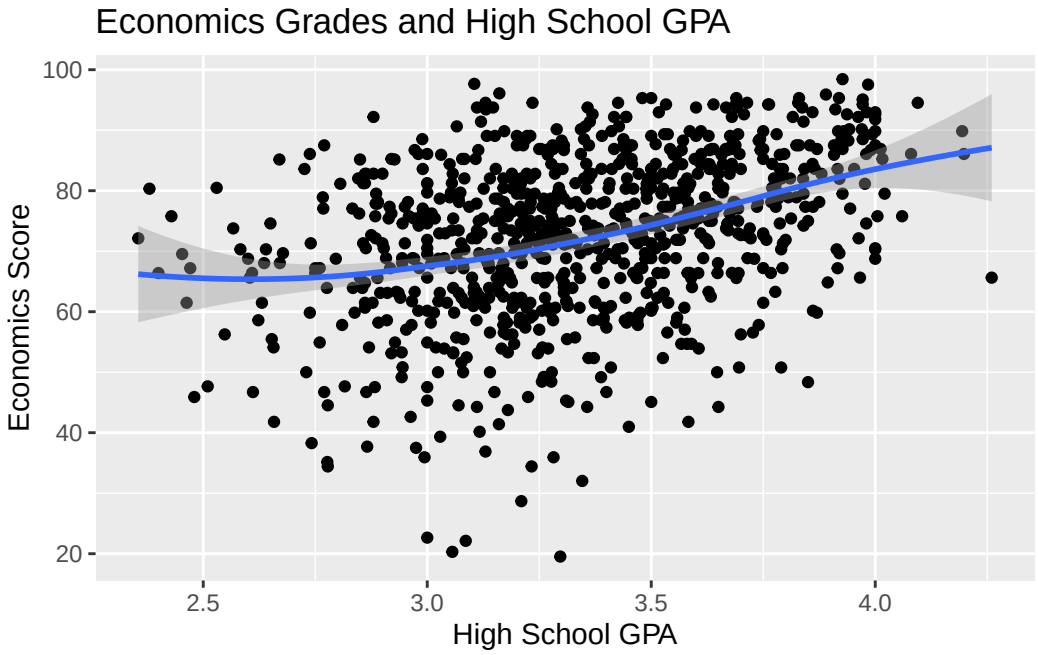
4. Using the `ggplot2` package, produce a scatterplot with *score* on the y-axis and *hsgpa*

on the x-axis. Layer on top of that a **loess** regression line (again, look up what a loess function is). Create a title for the graph and relabel the x and y axes with more intuitive names. Describe the relationship between these two variables.

INSERT COMMENTS HERE

```
p<-ggplot(econmath, aes(y = score, x = hsgpa)) +
  geom_point() +
  geom_smooth(method = "loess") +
  labs(title = "Economics Grades and High School GPA",
       x = "High School GPA", y = "Economics Score")
p
```

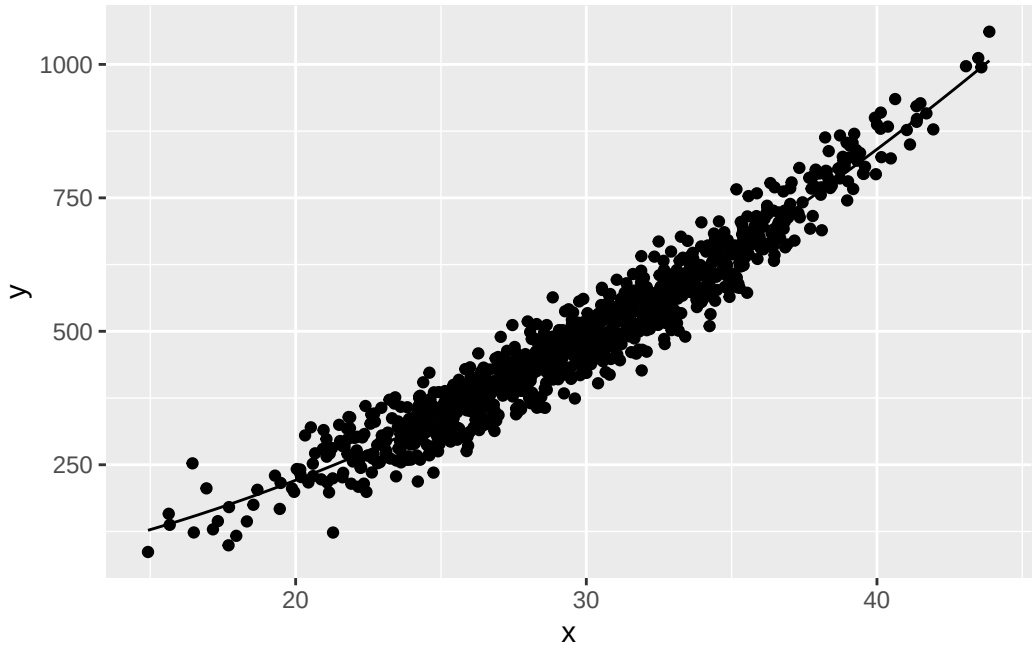
```
`geom_smooth()` using formula = 'y ~ x'
```



5. Suppose that the process that generates the data is $y = 1 + x + 0.5 * x^2 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 40)$. This means that the Conditional Expecation Function (CEF) is $E[y|x] = 1 + x + 0.5 * x^2$. The code below creates the data for x and y. Plot the conditional expectation function on top of a scatterplot of the data.

```r
data <- tibble(x = rnorm(1000,30,5),
               y = 1 + x + 0.5*x^2 + rnorm(1000,0,40))

ggplot(data, aes(x = x, y = y)) +
  geom_point() +
  geom_function(fun = function(x) 1 + x + 0.5*x^2)
```



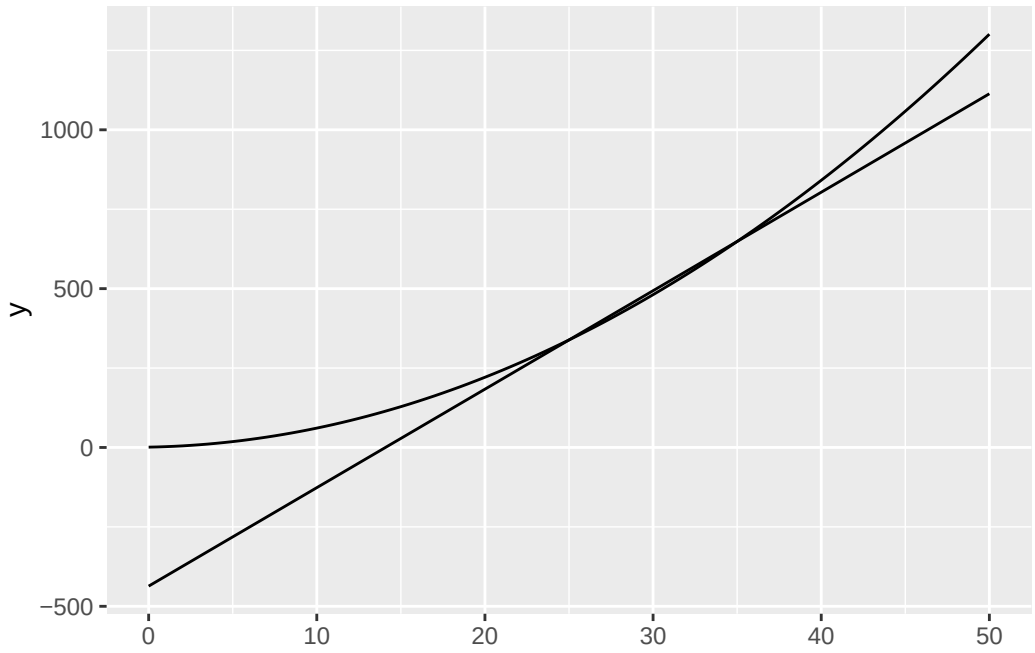
5. Suppose you are interested in the Population Regression of y on x. Compute the population regression slope and intercept. A useful piece of information for this question is that for a Normal random variable x, the covariance between x and x^2 is $(E[x])^3 + 3E[x]Var[x] - E[x]((E[x])^2 + Var[x])$.

INSERT COMMENTS HERE

6. Plot the Population Regression Function (PRF) with the CEF and comment on the quality of the approximation.

INSERT COMMENTS HERE

```r
ggplot() +
  geom_function(fun = function(x) 1 + x + 0.5*x^2) +
  geom_function(fun = function(x) -436.5 + 31*x ) +
  xlim(0,50)
```


y
1000
500
0
−500
0
10
20
30
40
50