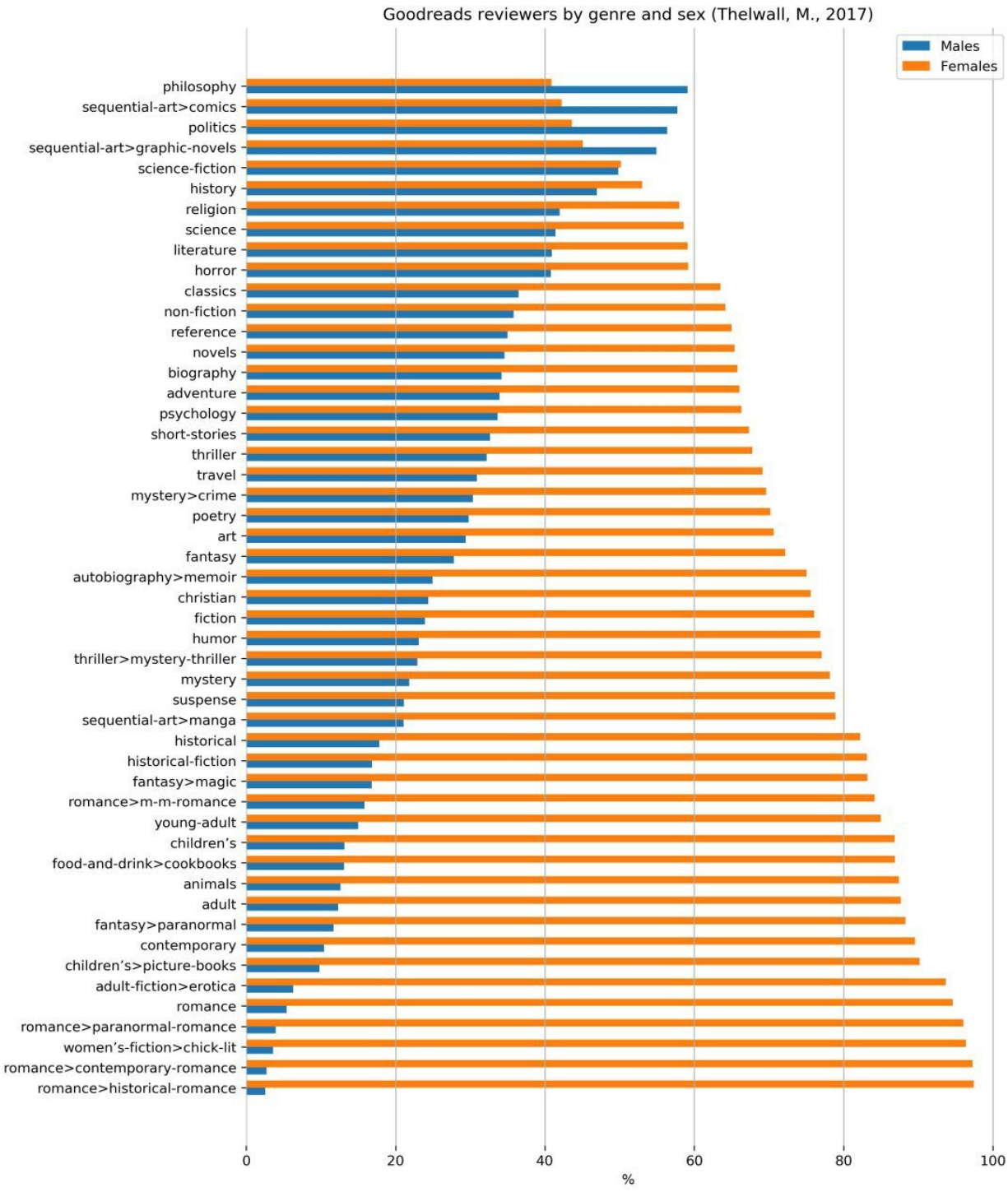
Midterm 1, Part 1

S&DS 361

2023-02-21

1. Visualization

The following visualization shows the percent of males and females who wrote reviews for various genres of books on Goodreads. Please give short answers (1-ish sentence) to the questions below.


Goodreads reviewers by genre and sex (Thelwall, M., 2017)
Males
Females
philosophy
sequential-art>comics
politics
sequential-art>graphic-novels
science-fiction
history
religion
science
literature
horror
classics
non-fiction
reference
novels
biography
adventure
psychology
short-stories
thriller
travel
mystery>crime
poetry
art
fantasy
autobiography>memoir
christian
fiction
humor
thriller>mystery-thriller
mystery
suspense
sequential-art>manga
historical
historical-fiction
fantasy>magic
romance>m-m-romance
young-adult
children's
food-and-drink>cookbooks
animals
adult
fantasy>paranormal
contemporary
children's>picture-books
adult-fiction>erotica
romance
romance>paranormal-romance
women's-fiction>chick-lit
romance>contemporary-romance
romance>historical-romance
0
20
40
60
80
100
%

a. Do the title, axis labels, and other text clearly summarize the contents of the visualization? Why?

Some possible responses:

- It is a little odd to have just "%" as the x-axis label. Something like Percentage of Reviews.
- The y-axis labels are a little small.
- The y-axis labels are all lowercase. Capitalizing the first letter would look better.
- I'm not really sure what the > and – mean.
- etc

b. What would you change about the visualization? Include at least one additional comment different from your response given above.

- It is unnecessary to have both male% and female%, since it looks like male% = 100% - females%. I would simplify and show just male% or just female%
- The bars are really close together, especially since there are two bars for each genre. Makes it a little harder to read.
- I wish there were a vertical line at 50%.
- It might be nice to have a vertical line showing average male% or average female%.
- If there is only one bar per genre, there might be enough room to add text to the end of each bar showing the percentage for that genre.
- etc

2. Commenting code

Below are the first four and last four rows of `d`, the NBA games data that we worked with previously in class and on assignments.

```
      season      gid team score
1 Season2021 22000001  GSW    99
2 Season2021 22000001  BKN   125
3 Season2021 22000002  LAC   116
4 Season2021 22000002  LAL   109
```

```
         season      gid team score
4617 Season2022 22101229  SAC   116
4618 Season2022 22101229  PHX   109
4619 Season2022 22101230  UTA   111
4620 Season2022 22101230  POR    80
```

Below is some code that processes this data and creates a visualization. Please add comments to the code below everywhere there is a ## explaining what that chunk of code does.

```
## Find the average points scored for each team in each of the two seasons
## So there are two rows per team, one for the 2021 season, and one for 2022.

ds = d %>%
  group_by(team, season) %>%
  summarise(score= mean(score)) %>%

  ## Pivot wider so that instead of two rows per team (one for each season),
  ## there is one row per team, with a column for each season.

  pivot_wider(names_from  = season,
              values_from = score)

## Create a scatter plot showing
## average points in 2022 vs average points in 2021
## Include the team name as a label for each point.
## Left justify that text label.

ggplot(ds, aes(x = Season2021,
               y = Season2022,
               label = team))+
  geom_point()+
  geom_text(hjust=-.1)
```

3. dplyr

Suppose the data frame `d` contains the 4 columns `open.date`, `network`, `lev2` and `lev3` from the EV stations data that we worked with previously in class and on assignments. The first 6 rows of `d` are shown below.

```
   open.date       network lev2 lev3
1 2023-01-14           FLO    6   NA
2 2023-01-14           FLO    4   NA
3 2023-01-14    EV Connect   NA    2
4 2023-01-14    EV Connect    2   NA
5 2023-01-14 Blink Network    6   NA
6 2023-01-14 Blink Network    2   NA
```

Suppose we run the the following code.

```
dd = d %>%
  mutate(lev2 = ifelse(is.na(lev2), 0, lev2),
         lev3 = ifelse(is.na(lev3), 0, lev3))  %>%
  filter(lev2!=0)
```

```
head(dd,4)
```

Write the first four rows of `dd` below.

```
   open.date       network lev2 lev3
1 2023-01-14           FLO    6    0
2 2023-01-14           FLO    4    0
3 2023-01-14    EV Connect    2    0
4 2023-01-14 Blink Network    6    0
```

4. ggplot

```
head(mtcars,2)
```

```
              mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4      21   6  160 110  3.9 2.620 16.46  0  1    4    4
Mazda RX4 Wag  21   6  160 110  3.9 2.875 17.02  0  1    4    4
```

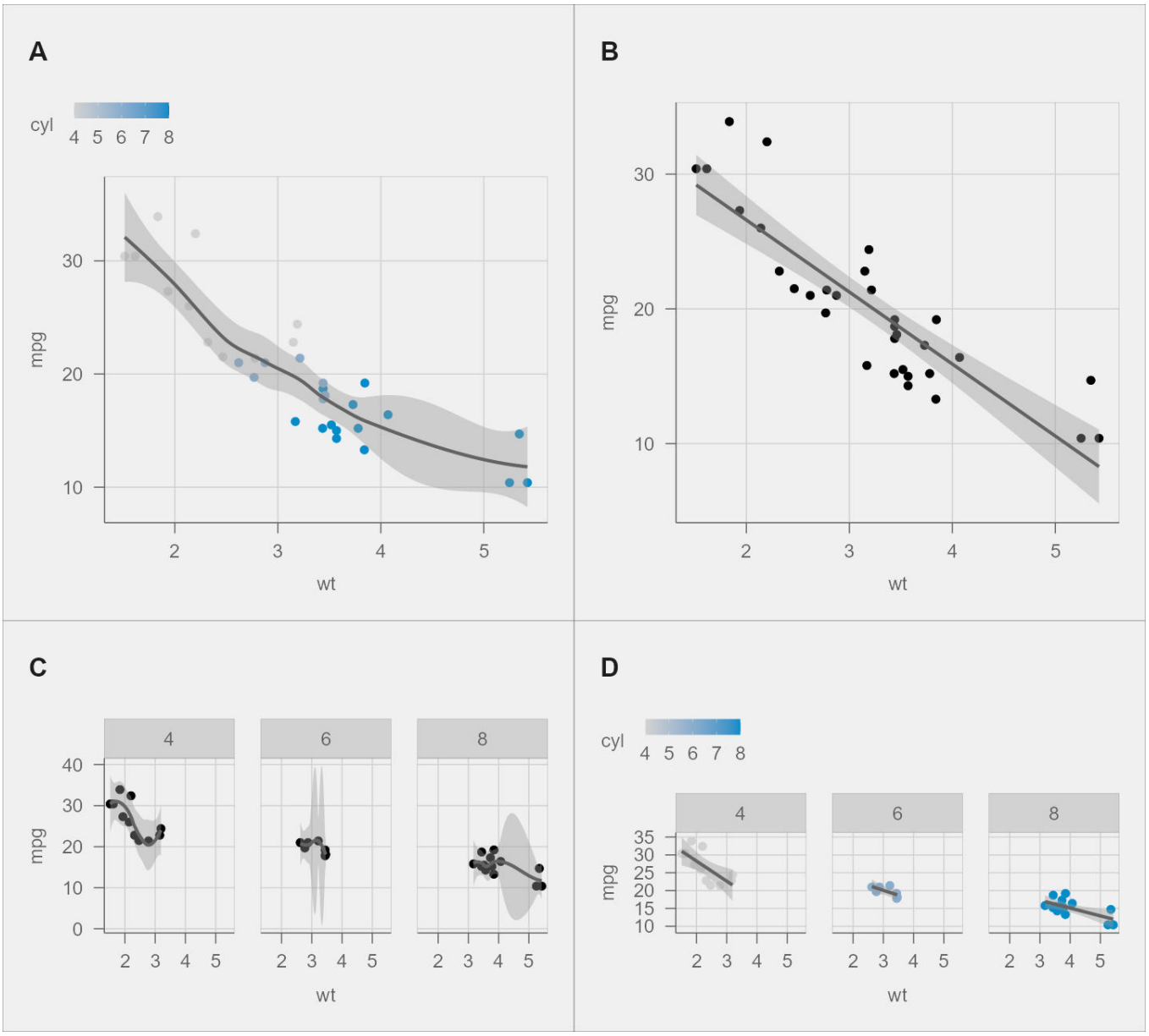
Below are 4 lines of code, each of which creates a visualization of the `mtcars` data. Below the code are 4 visualizations labeled A, B, C, and D, which were generated by one of the four lines of code. Match each line of code to the visualization it generates. Indicate your choice by writing A, B, C, or D in the blank to the left of each line of code.

```
_A_ ggplot(d=mtcars, aes(x=wt, y=mpg)) + geom_point(aes(color=cyl)) + geom_smooth(           )

_D_ ggplot(d=mtcars, aes(x=wt, y=mpg)) + geom_point(aes(color=cyl)) + geom_smooth(method='lm') + facet_wrap(~cyl)

_C_ ggplot(d=mtcars, aes(x=wt, y=mpg)) + geom_point(color='black' ) + geom_smooth(           ) + facet_wrap(~cyl)

_B_ ggplot(d=mtcars, aes(x=wt, y=mpg)) + geom_point(color='black' ) + geom_smooth(method='lm')
```



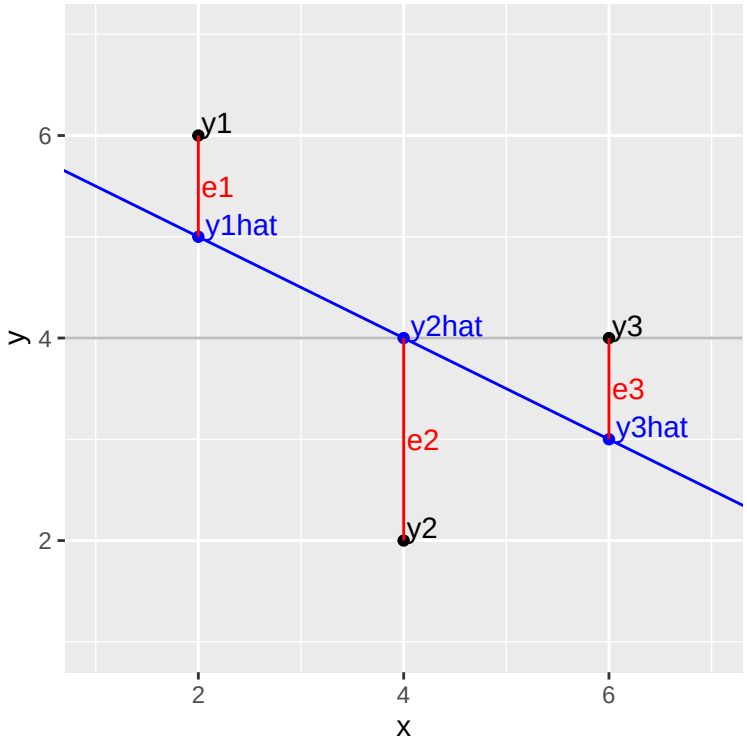
5. Regression

Consider the following three data points, linear model, and scatter plot with the regression line from the model.

$$(x_1, y_1) = (2, 6), \quad (x_2, y_2) = (4, 2), \quad (x_3, y_3) = (6, 4)$$

```
x1=2; y1=6;
x2=4; y2=2;
x3=6; y3=4;
d = data.frame(x=c(x1,x2,x3),
               y=c(y1,y2,y3))
m = lm(y~x, data=d)
m$coefficients
```

```
(Intercept)           x
        6.0        -0.5
```



Use this information to answer the following questions. Do the calculations by hand and show your work.

a. Label $y_1, y_2,$ and y_3 on the graph.

b. Label $\hat{y}_1, \hat{y}_2$, and $\hat{y}_3$, the predicted values of y corresponding to x_1, x_2, and x_3.

c. Label the parts of the graph that represent the error terms (residuals) e_1, e_2, and e_3?

d. What is $\bar{y}$, the sample mean of y?

$$\frac{1}{3}(6+2+4)$$

e. Compute SSE for this model.

$$SSE = \sum_{1}^{3}(y_j - \hat{y_j})^2 = (y_1 - \hat{y_1})^2 + (y_2 - \hat{y_2})^2 + (y_3 - \hat{y_3})^2 = (6-5)^2 + (2-4)^2 + (4-3)^2 = 6$$

f. Compute SST for this model.

$$SST = \sum_{1}^{3}(y_j - \hat{y_j})^2 = (y_1 - \hat{y_1})^2 + (y_2 - \hat{y_2})^2 + (y_3 - \hat{y_3})^2 = (6-4)^2 + (2-4)^2 + (4-4)^2 = 8$$

g. Compute R^2 for this model.

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{6}{8} = \frac{1}{4} = 0.25$$

h. Name 3 assumptions of a simple linear regression model. Here are four:

i. Linear relationship between x and y.
ii. Errors ϵ are normally distributed.
iii. Errors have constant variance. The variance does not depend on x.
iv. Errors are independent.

6. Multiple Regression

In this question we'll analyze the data `FirstYearGPA.csv`, a new data set. The handout that accompanies this exam contains

- the first 2 rows of the data
- a `ggpairs` plot
- 4 models, along with the `summary` output of those models,

which you will need to use to answer the questions below. Some column definitions:

- GPA is grade point average in first year of college,
- HSGPA is grade point average in high school,
- SATV is SAT Verbal score,

- SATM is SAT math score,
- HU is the number of credit hours of humanities courses in high school.

a. What percentage of the variation in `GPA` is explained by the model `m1`?

$R^2 = 0.1997$, so 19.97%.

b. Is `m1` useful for predicting `GPA`? Give at least two parts of the `summary(m1)` output that support your answer.

Yes, R^2 is much greater than 0. `HSGPA` is significant. The sign and effect size of `HSGPA` are practically meaningful.

c. When $x = 3.5$, we get $\hat{y} = 3.12$. Which of the intervals below is the 95% confidence interval for $\hat{y}$, and which is the 95% prediction interval for y, when $x = 3.5$? How can you tell which is which?

```
   fit lwr  upr
1 3.12 2.3 3.95
```

```
   fit  lwr  upr
1 3.12 3.07 3.18
```

The first is the prediction interval for y, it is much wider.

(continued)

d. Is there any evidence of collinearity that we should be worried about when building a multiple regression model? Explain.

`SATV` and `SATM` are somewhat correlated. So maybe, but it's not huge.

e. Which of the models `m1` thru `m4` would you consider to be the best? Why?

`m4`, highest adjusted R^2, significant predictors, practical meaningful effect sizes.

f. Given what you know about `m1` thru `m4`, what is the next model you would try for `m5`? Why?

I would try `GPA ~ HSGPA + SATV + HU`. Since `m2` and `m4` were better than `m3`, and `SATV` and `HU` are more correlated with `GPA` than `SATM`, it's worth trying both `SATV` and `HU` together with `HSGPA`.

g. If a student got a 4.0 in HSGPA, and 800 on SATV, what can you say about her expected GPA in the first year of college, according to `m2`? (rounded to the nearest 0.0001)

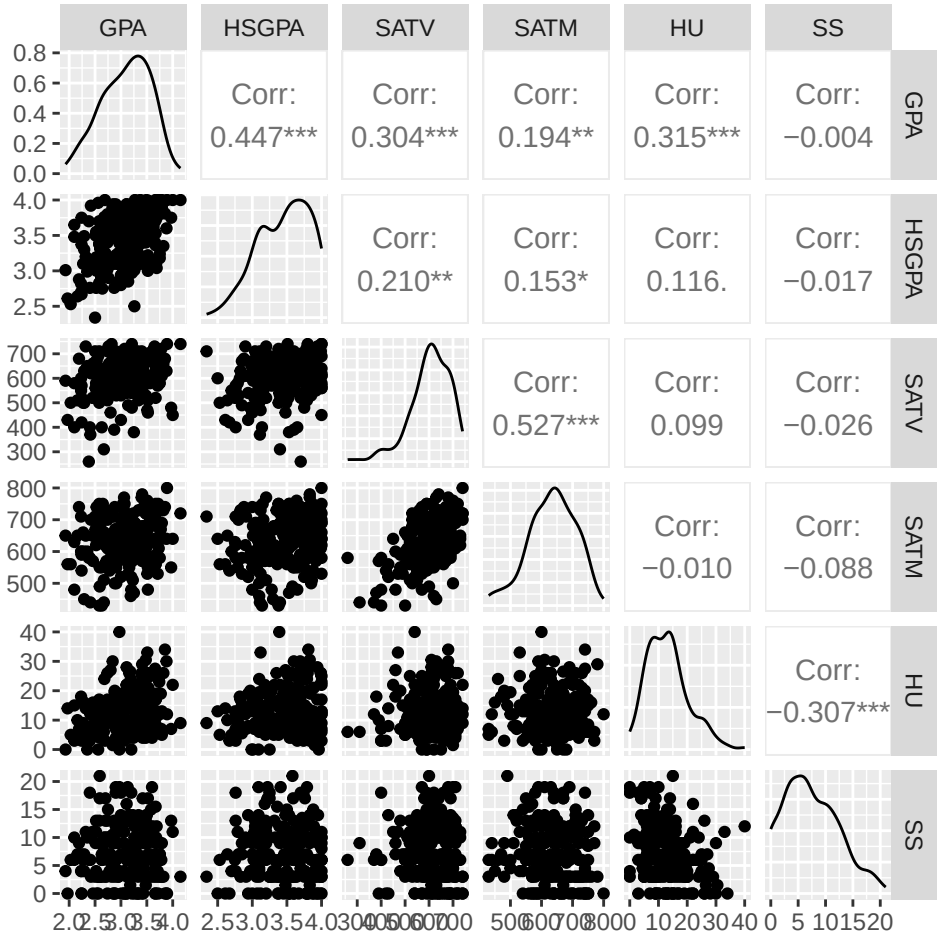
$$E(GPA) = 0.6351 + 0.4975(4.0) + 0.0012(800) = 3.5851$$

Handout (3 pages)

```
d = read.csv('data/FirstYearGPA.csv')
d = d %>% select(-X)
head(d,2)
```

```
   GPA HSGPA SATV SATM Male HU SS FirstGen White CollegeBound
1 3.06  3.83  680  770    1  3  9        1     1            1
2 4.15  4.00  740  720    0  9  3        0     1            1
```

```
ggpairs(d[,c(1:4,6:7)])
```



```r
m1 = lm(GPA ~ HSGPA       , data=d)
m2 = lm(GPA ~ HSGPA + SATV, data=d)
m3 = lm(GPA ~ HSGPA + SATM, data=d)
m4 = lm(GPA ~ HSGPA + HU  , data=d)

summary(m1)
```

```
Call:
lm(formula = GPA ~ HSGPA, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-1.10565 -0.31329  0.05871  0.29485  0.82291

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.17985    0.26194   4.504 1.09e-05 ***
HSGPA        0.55501    0.07542   7.359 3.78e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4174 on 217 degrees of freedom
Multiple R-squared:  0.1997,    Adjusted R-squared:  0.196
F-statistic: 54.15 on 1 and 217 DF,  p-value: 3.783e-12
```

```r
summary(m2)
```

```
Call:
lm(formula = GPA ~ HSGPA + SATV, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-0.97894 -0.27639  0.02867  0.30133  0.87956

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.6351217  0.2955033   2.149  0.03272 *
HSGPA       0.4975320  0.0750569   6.629 2.66e-10 ***
SATV        0.0012283  0.0003373   3.641  0.00034 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4061 on 216 degrees of freedom
Multiple R-squared:  0.246, Adjusted R-squared:  0.239
F-statistic: 35.23 on 2 and 216 DF,  p-value: 5.711e-14
```

```r
summary(m3)
```

```
Call:
lm(formula = GPA ~ HSGPA + SATM, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-1.00720 -0.31027  0.04086  0.31148  0.83620

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.7579762  0.3274774   2.315   0.0216 *
HSGPA       0.5305151  0.0757139   7.007 3.06e-11 ***
SATM        0.0007985  0.0003772   2.117   0.0354 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4141 on 216 degrees of freedom
Multiple R-squared:  0.216, Adjusted R-squared:  0.2087
F-statistic: 29.75 on 2 and 216 DF,  p-value: 3.869e-12
```

```r
summary(m4)
```

```
Call:
lm(formula = GPA ~ HSGPA + HU, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-1.04272 -0.28375  0.05263  0.26621  0.91674

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.087416   0.251617   4.322 2.36e-05 ***
HSGPA       0.516624   0.072705   7.106 1.72e-11 ***
HU          0.017163   0.003772   4.550 8.93e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3996 on 216 degrees of freedom
Multiple R-squared:  0.2697,    Adjusted R-squared:  0.263
F-statistic: 39.89 on 2 and 216 DF,  p-value: 1.808e-15
```