Student Performance Report

Jeremiah Perkins

12/10/2020

Part1

Introduction

This Student Performance Report is a quick Exploratory Data Analysis followed by some modeling that aims

```
library(fastDummies)
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------------

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts ------------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readxl)
library(knitr)
tinytex::install_tinytex()
```

```
## Warning: Detected an existing tlmgr at /Users/tasneemward/Library/TinyTeX/
## bin/x86_64-darwin/tlmgr. It seems TeX Live has been installed (check
## tinytex::tinytex_root()). You are recommended to uninstall it, although TinyTeX
## should work well alongside another LaTeX distribution if a LaTeX document is
## compiled through tinytex::latexmk().

## The directory /usr/local/bin is not writable. I recommend that you make it writable. See https://gitl

## Warning: Please run this command in your Terminal (password required):
##   sudo chown -R ‘whoami‘:admin /usr/local/bin

## TinyTeX installed to /Users/tasneemward/Library/TinyTeX
```

Now that we have the packages in, we can load in the data and while we are at it let’s change the name t

```
StudentsPerformance_1_ <- read_excel("~/Downloads/StudentsPerformance (1).xlsx")
data <- StudentsPerformance_1_
glimpse(data)
```

```
## Rows: 1,000
## Columns: 8
## $ gender                        <chr> "female", "female", "female", "male",...
## $ ‘race/ethnicity‘              <chr> "group B", "group C", "group B", "gro...
## $ ‘parental level of education‘ <chr> "bachelor’s degree", "some college", ...
## $ lunch                         <chr> "standard", "standard", "standard", "...
## $ ‘test preparation course‘     <chr> "none", "completed", "none", "none", ...
## $ ‘math score‘                  <dbl> 72, 69, 90, 47, 76, 71, 88, 40, 64, 3...
## $ ‘reading score‘               <dbl> 72, 90, 95, 57, 78, 83, 95, 43, 64, 6...
## $ ‘writing score‘               <dbl> 74, 88, 93, 44, 75, 78, 92, 39, 67, 5...
```

This dataset contains 8 columns and 1000 rows of data with variables including: gender, race/ethnicity,

Part2

Cleaning

There are some spaces between some of the variables so let’s clean that up a little bit. We also are goi

```
names(data)[names(data) == "math score"] <- "math.score"
names(data)[names(data) == "writing score"] <- "writing.score"
names(data)[names(data) == "reading score"] <- "reading.score"
names(data)[names(data) == "parental level of education"] <- "parent.edu"
names(data)[names(data) == "testing preparation course"] <- "test.prep"
data1 <- dummy_cols(data, select_columns = c("test preparation course", "gender"))
reading.score.mean <- mean(data$reading.score)
math.score.mean <- mean(data$math.score)
writing.score.mean <- mean(data$math.score)
data  <- data %>% mutate(math.above.below = ifelse(math.score < math.score.mean, "Below Mean", "Above Me
data  <- data %>% mutate(reading.above.below = ifelse(reading.score < reading.score.mean, "Below Mean",
data  <- data %>% mutate(writing.above.below = ifelse(writing.score < writing.score.mean, "Below Mean",
```

part3

Exploratory Data Analysis

We need to get some counts of our data, subset our data, correlation of our data, and graphs of our data

Count of data by gender

```
data %>% group_by(gender) %>% count(‘test preparation course‘)
```

```
## # A tibble: 4 x 3
## # Groups:   gender [2]
##   gender ‘test preparation course‘     n
##   <chr>  <chr>                     <int>
## 1 female completed                   184
## 2 female none                        334
## 3 male   completed                   174
## 4 male   none                        308
```

```
data %>% group_by(gender) %>% count(parent.edu)
```

```
## # A tibble: 12 x 3
## # Groups:   gender [2]
##    gender parent.edu             n
##    <chr>  <chr>              <int>
##  1 female associate’s degree   116
##  2 female bachelor’s degree     63
##  3 female high school           94
##  4 female master’s degree       36
##  5 female some college         118
##  6 female some high school      91
##  7 male   associate’s degree   106
##  8 male   bachelor’s degree     55
##  9 male   high school          102
## 10 male   master’s degree       23
## 11 male   some college         108
## 12 male   some high school      88
```

```
data %>% group_by(gender) %>% count(lunch)
```

```
## # A tibble: 4 x 3
## # Groups:   gender [2]
##   gender lunch            n
##   <chr>  <chr>        <int>
## 1 female free/reduced   189
## 2 female standard       329
## 3 male   free/reduced   166
## 4 male   standard       316
```

```
data %>% group_by(gender) %>% count(‘race/ethnicity‘)
```

```
## # A tibble: 10 x 3
## # Groups:   gender [2]
##    gender ‘race/ethnicity‘     n
##    <chr>  <chr>            <int>
##  1 female group A             36
##  2 female group B            104
##  3 female group C            180
##  4 female group D            129
##  5 female group E             69
##  6 male   group A             53
##  7 male   group B             86
##  8 male   group C            139
##  9 male   group D            133
## 10 male   group E             71
```

```
data %>% group_by(gender) %>% count(gender)
```

```
## # A tibble: 2 x 2
## # Groups:   gender [2]
##   gender     n
##   <chr>  <int>
## 1 female   518
## 2 male     482
```

```
data %>% group_by(gender) %>% count(math.score)
```

```
## # A tibble: 147 x 3
## # Groups:   gender [2]
##    gender math.score     n
##    <chr>       <dbl> <int>
##  1 female          0     1
##  2 female          8     1
##  3 female         18     1
##  4 female         19     1
##  5 female         22     1
##  6 female         23     1
##  7 female         24     1
##  8 female         26     1
##  9 female         27     1
## 10 female         29     3
## # ... with 137 more rows
```

```
data %>% group_by(gender) %>% count(reading.score)
```

```
## # A tibble: 132 x 3
## # Groups:   gender [2]
##    gender reading.score     n
##    <chr>          <dbl> <int>
##  1 female            17     1
##  2 female            24     1
##  3 female            29     1
##  4 female            31     1
##  5 female            32     1
##  6 female            34     2
##  7 female            38     2
##  8 female            39     2
##  9 female            40     1
## 10 female            41     2
## # ... with 122 more rows
```

```
data %>% group_by(gender) %>% count(writing.score)
```

```
## # A tibble: 136 x 3
## # Groups:   gender [2]
##    gender writing.score     n
##    <chr>          <dbl> <int>
##  1 female            10     1
##  2 female            23     1
##  3 female            27     1
##  4 female            28     1
##  5 female            30     1
##  6 female            32     2
##  7 female            33     2
##  8 female            36     1
##  9 female            38     3
## 10 female            39     1
## # ... with 126 more rows
```

```
data %>% group_by(gender) %>% count(‘race/ethnicity‘)
```

```
## # A tibble: 10 x 3
## # Groups:   gender [2]
##    gender ‘race/ethnicity‘     n
##    <chr>  <chr>            <int>
##  1 female group A             36
##  2 female group B            104
##  3 female group C            180
##  4 female group D            129
##  5 female group E             69
##  6 male   group A             53
##  7 male   group B             86
##  8 male   group C            139
##  9 male   group D            133
## 10 male   group E             71
```

```
data %>% group_by(‘race/ethnicity‘) %>% count(‘test preparation course‘)
```

```
## # A tibble: 10 x 3
## # Groups:   race/ethnicity [5]
##    ‘race/ethnicity‘ ‘test preparation course‘     n
##    <chr>            <chr>                     <int>
##  1 group A          completed                    31
##  2 group A          none                         58
##  3 group B          completed                    68
##  4 group B          none                        122
##  5 group C          completed                   117
##  6 group C          none                        202
##  7 group D          completed                    82
##  8 group D          none                        180
##  9 group E          completed                    60
## 10 group E          none                         80
```

```
data3 <- data %>% count(‘race/ethnicity‘)
table(data$math.above.below)
```

```
##
## Above Mean Below Mean
##        493        507
```

```
table(data$writing.above.below)
```

```
##
## Above Mean Below Mean
##        568        432
```

```
table(data$reading.above.below)
```

```
##
## Above Mean Below Mean
##        513        487
```

Counts are a good way of explaining data numerically, but they are not always the funniest thing to scr

Correlation

```
data %>% summarize(N = n(), r = cor(math.score, reading.score))
```

```
## # A tibble: 1 x 2
##       N     r
##   <int> <dbl>
## 1  1000 0.818
```

```
data %>% summarize(N = n(), r = cor(math.score, writing.score))
```

```
## # A tibble: 1 x 2
##       N     r
##   <int> <dbl>
## 1  1000 0.803
```

```
data %>% summarize(N = n(), r = cor(reading.score, writing.score))
```

```
## # A tibble: 1 x 2
##       N     r
##   <int> <dbl>
## 1  1000 0.955
```

The correlation between all the variables are above .8 which is a very high correlation. The correlati

I wonder if we have students who performed well on one test, but did not so well on another. Let's look through the dataset and see if we can find such student. We also can also do a few other filter options that we will add to the code.

```
data %>% filter(math.score > 90 & reading.score < 75)
```

```
## # A tibble: 3 x 11
##   gender ‘race/ethnicity‘ parent.edu lunch ‘test preparati~ math.score
##   <chr>  <chr>            <chr>      <chr> <chr>                 <dbl>
## 1 male   group C          some coll~ stan~ none                     91
## 2 male   group E          associate~ free~ completed                91
## 3 male   group E          high scho~ stan~ none                     94
## # ... with 5 more variables: reading.score <dbl>, writing.score <dbl>,
## #   math.above.below <chr>, reading.above.below <chr>,
## #   writing.above.below <chr>
```

```
data %>% filter(reading.score > 90 & math.score < 75)
```

```
## # A tibble: 1 x 11
##   gender ‘race/ethnicity‘ parent.edu lunch ‘test preparati~ math.score
##   <chr>  <chr>            <chr>      <chr> <chr>                 <dbl>
## 1 female group D          high scho~ free~ none                     73
## # ... with 5 more variables: reading.score <dbl>, writing.score <dbl>,
## #   math.above.below <chr>, reading.above.below <chr>,
## #   writing.above.below <chr>
```

```
data %>% filter(parent.edu == "some high school")
```

```
## # A tibble: 179 x 11
##    gender ‘race/ethnicity‘ parent.edu lunch ‘test preparati~ math.score
##    <chr>  <chr>            <chr>      <chr> <chr>                 <dbl>
##  1 female group C          some high~ stan~ none                     69
##  2 female group B          some high~ free~ none                     18
##  3 female group C          some high~ stan~ none                     69
##  4 female group D          some high~ free~ none                     50
##  5 female group C          some high~ free~ completed                71
##  6 female group C          some high~ free~ none                      0
##  7 male   group A          some high~ free~ none                     39
##  8 female group D          some high~ stan~ none                     59
##  9 male   group B          some high~ stan~ none                     67
## 10 male   group D          some high~ free~ none                     45
## # ... with 169 more rows, and 5 more variables: reading.score <dbl>,
## #   writing.score <dbl>, math.above.below <chr>, reading.above.below <chr>,
## #   writing.above.below <chr>
```

```
data %>% filter(parent.edu == "some college")
```

```
## # A tibble: 226 x 11
##    gender ‘race/ethnicity‘ parent.edu lunch ‘test preparati~ math.score
##    <chr>  <chr>            <chr>      <chr> <chr>                 <dbl>
##  1 female group C          some coll~ stan~ completed                69
##  2 male   group C          some coll~ stan~ none                     76
##  3 female group B          some coll~ stan~ completed                88
##  4 male   group B          some coll~ free~ none                     40
##   5 male   group A          some coll~ stan~ completed                78
##   6 female group B          some coll~ free~ completed                65
##   7 male   group D          some coll~ stan~ none                     44
##   8 male   group B          some coll~ stan~ none                     69
##   9 female group D          some coll~ stan~ none                     69
##  10 female group B          some coll~ stan~ none                     63
## # ... with 216 more rows, and 5 more variables: reading.score <dbl>,
## #   writing.score <dbl>, math.above.below <chr>, reading.above.below <chr>,
## #   writing.above.below <chr>
```

```
data %>% filter(parent.edu == "some college" & `test preparation course` == "completed")
```

```
## # A tibble: 77 x 11
##    gender `race/ethnicity` parent.edu lunch `test preparati~ math.score
##    <chr>  <chr>            <chr>      <chr> <chr>                 <dbl>
##  1 female group C          some coll~ stan~ completed                69
##  2 female group B          some coll~ stan~ completed                88
##  3 male   group A          some coll~ stan~ completed                78
##  4 female group B          some coll~ free~ completed                65
##  5 male   group B          some coll~ free~ completed                59
##  6 male   group D          some coll~ stan~ completed                58
##  7 female group D          some coll~ free~ completed                58
##  8 male   group D          some coll~ stan~ completed                63
##  9 male   group A          some coll~ free~ completed                50
## 10 female group E          some coll~ stan~ completed                63
## # ... with 67 more rows, and 5 more variables: reading.score <dbl>,
## #   writing.score <dbl>, math.above.below <chr>, reading.above.below <chr>,
## #   writing.above.below <chr>
```

```
We can see there are only three students with a math score of over 90 and a reading score of less than
```

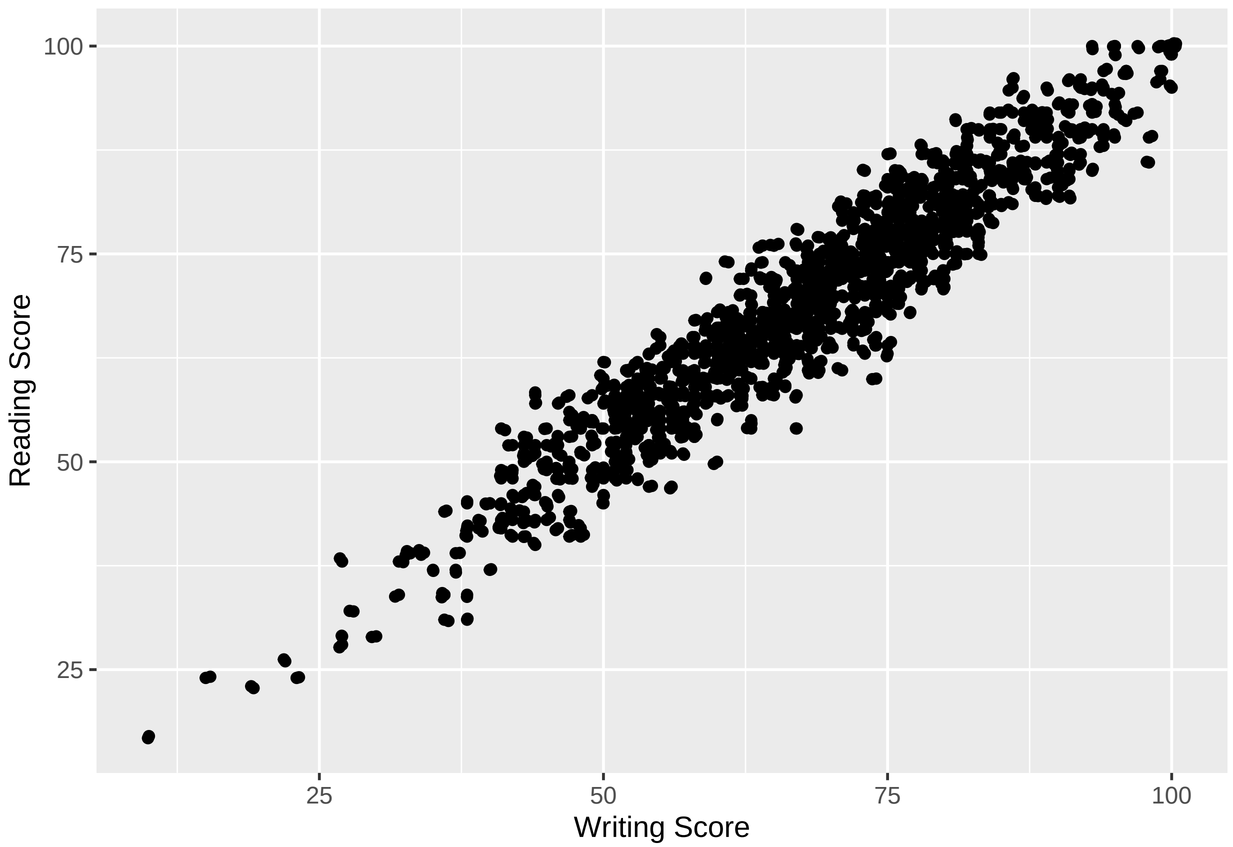
Part4

Graphs

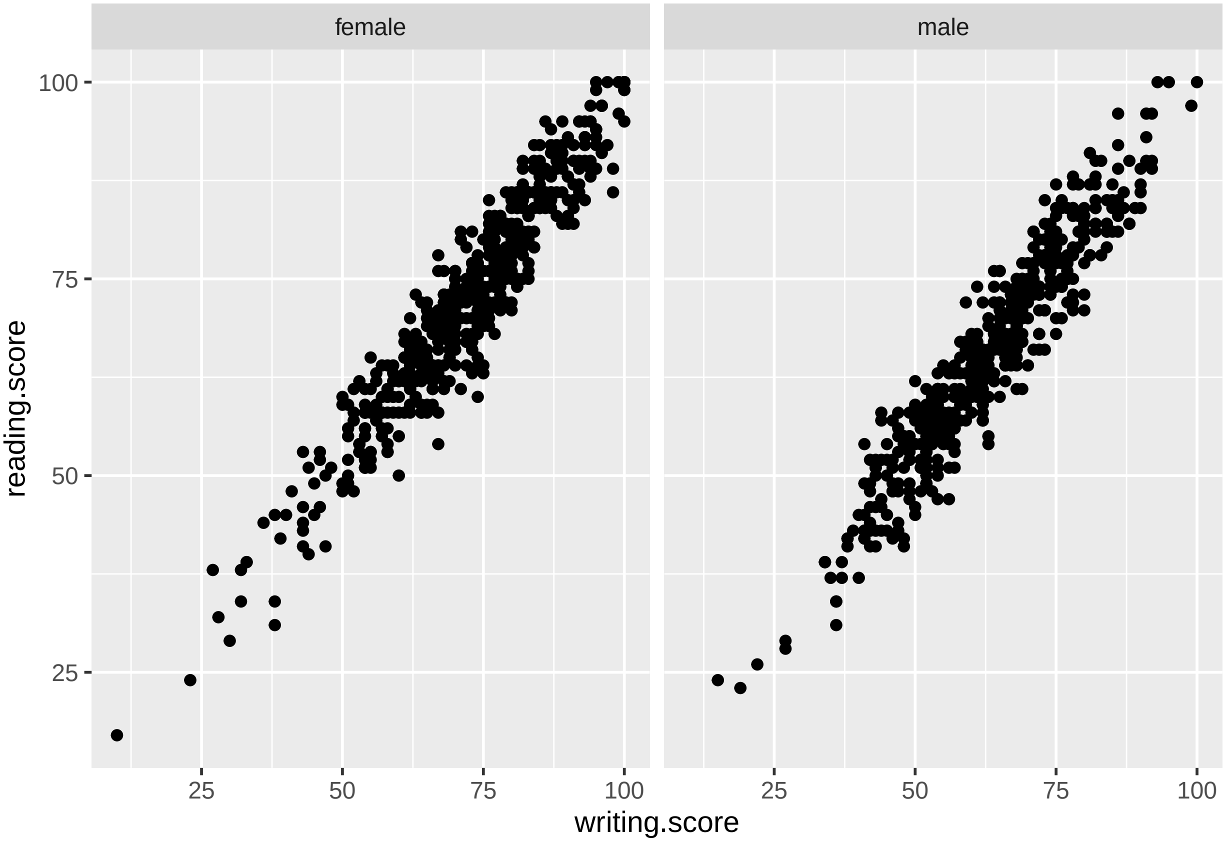
```
Graphs are a great way to visualize data and a way to visualize the numerical data you have been workin
```

Scatter

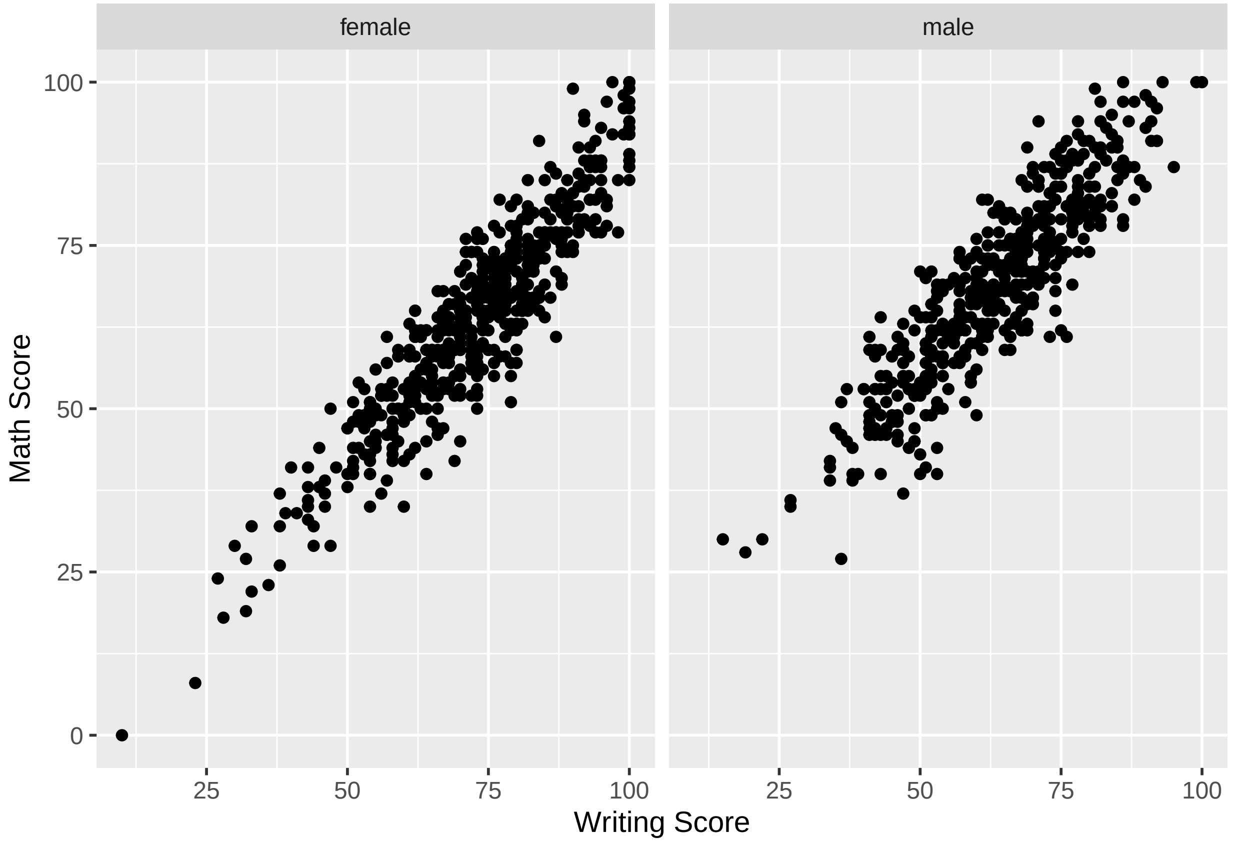
```
options(repr.plot.width = 5, repr.plot.height = 4)
ggplot(data, aes(x = writing.score, y = reading.score)) + geom_point() + geom_jitter() + labs(x = "Writ
```



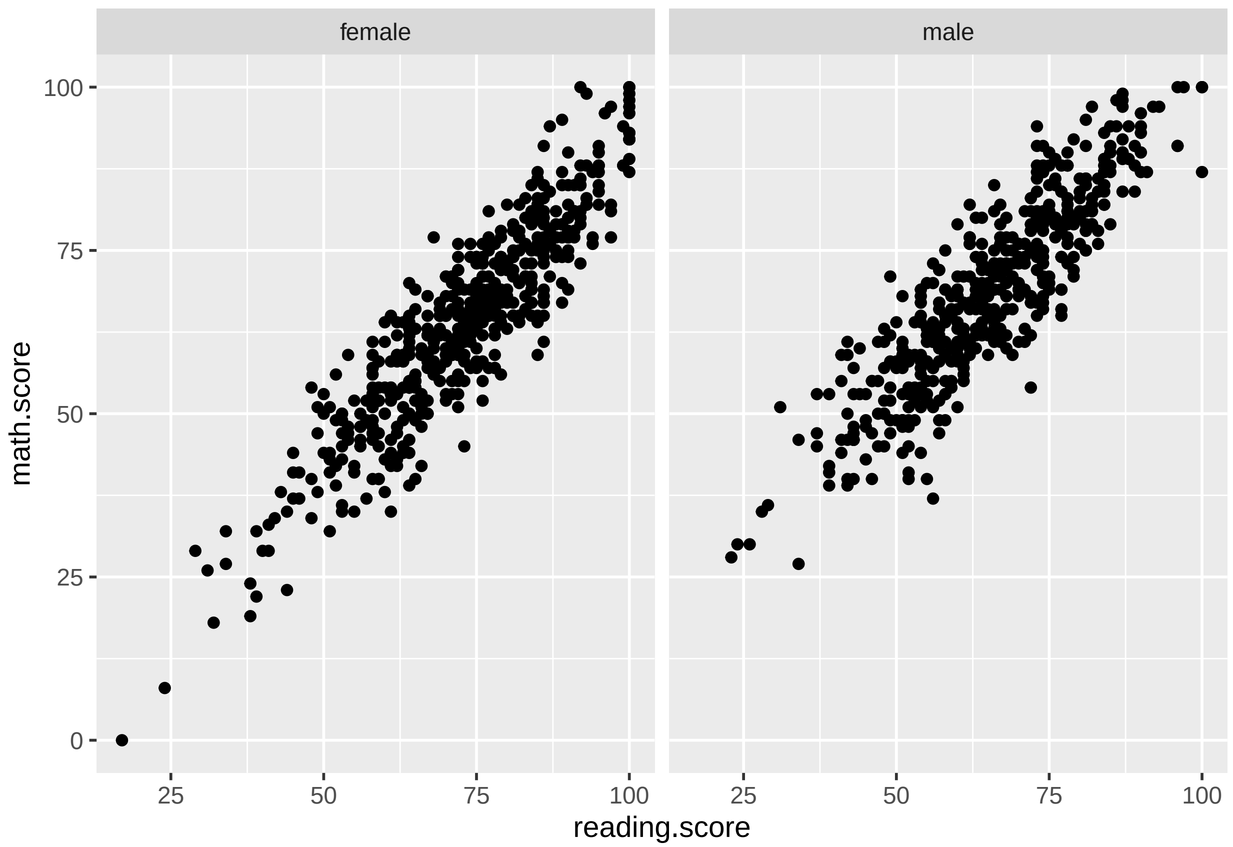
```
ggplot(data, aes(x = writing.score, y = reading.score)) + geom_point() + facet_wrap(~ gender)
```



```
ggplot(data, aes(x = writing.score, y = math.score)) + geom_point() + labs(x = "Writing Score", y = "Ma
```

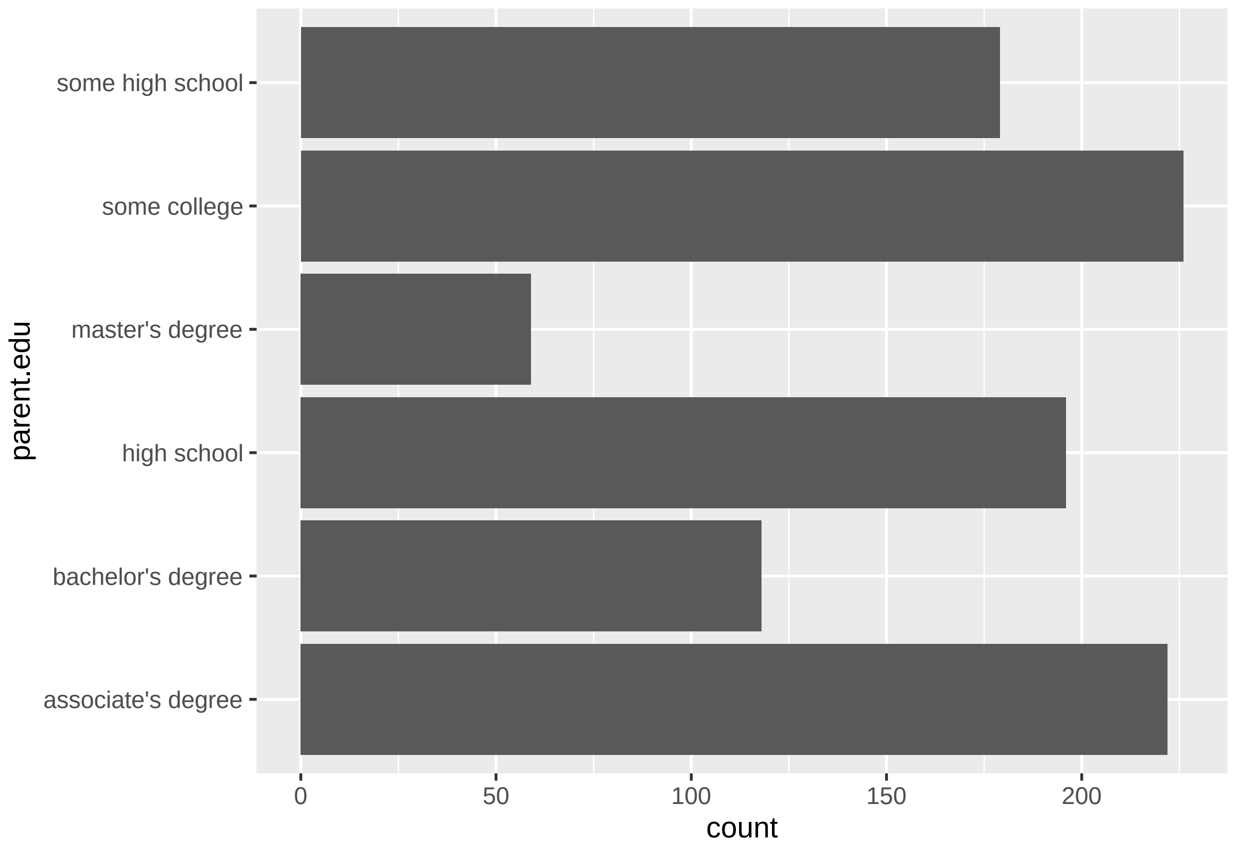


```r
ggplot(data, aes(x = reading.score, y = math.score)) + geom_point() + facet_wrap(~ gender)
```

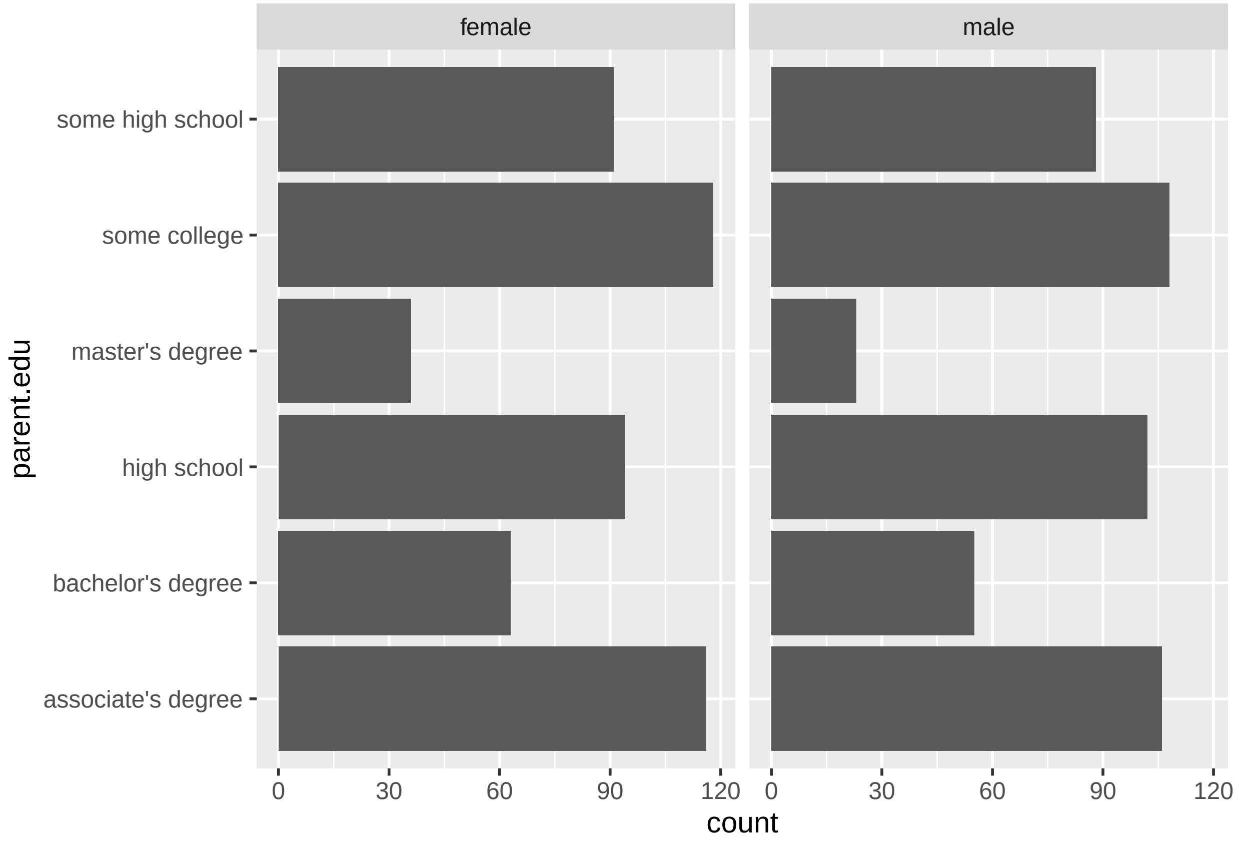


Bar

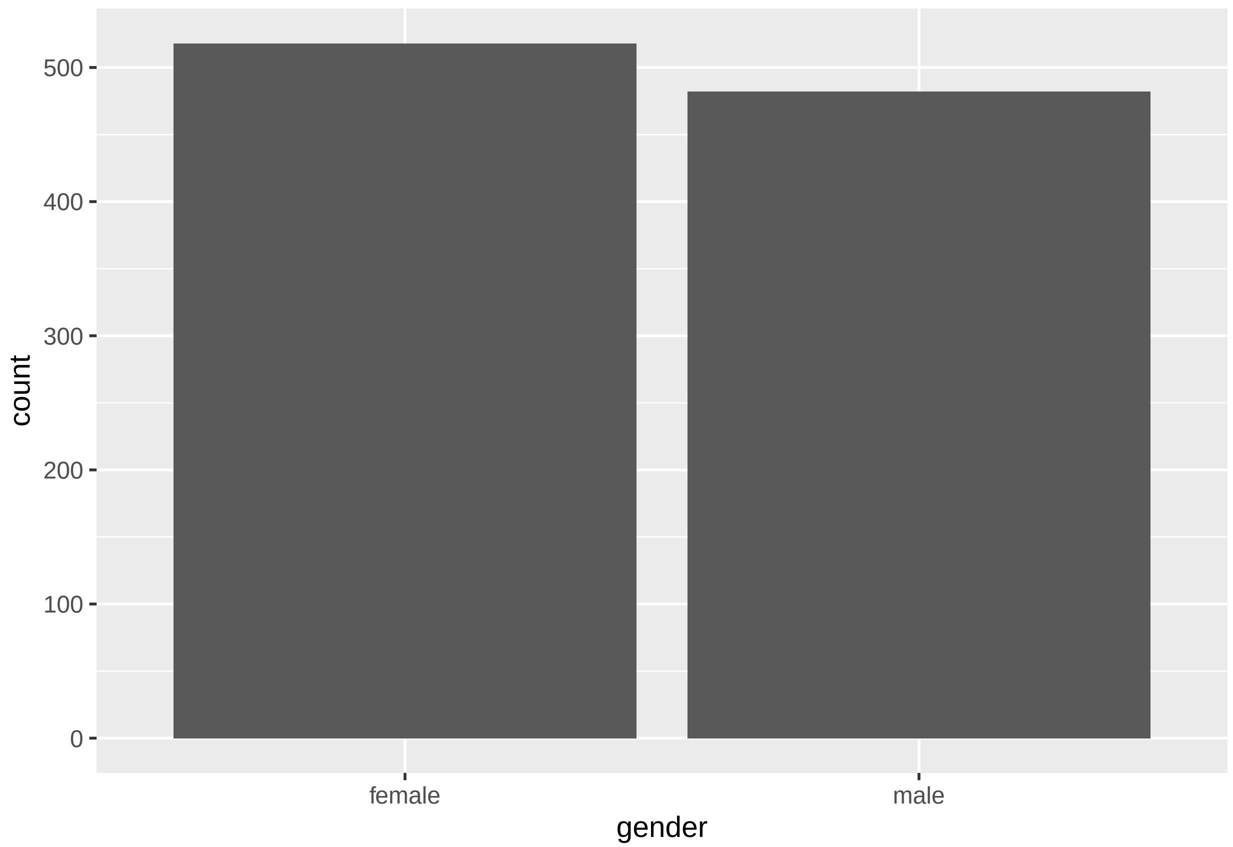
```
ggplot(data, aes(x = parent.edu)) + geom_bar() + coord_flip()
```



```r
ggplot(data, aes(x = parent.edu)) + geom_bar() +  facet_wrap(~ gender) + coord_flip()
```



```r
ggplot(data, aes(x = gender)) + geom_bar()
```



part5

Models

The models I created aimss to predict math score and writing score, the dependent variables, from a vari
Y = a + bX + e_i where Y = dependent variable, a is the intercept, b is the coefficient, X is the indepe

```
model <- lm(math.score ~ gender_female + writing.score + reading.score + ‘test preparation course_comple
model1.2 <- lm(math.score ~ gender_female + writing.score + reading.score + ‘test preparation course_non
model2 <- lm(math.score ~ gender_male + writing.score + reading.score + ‘test preparation course_comple
model2.2 <- lm(math.score ~ gender_male + writing.score + reading.score + ‘test preparation course_none
model3 <- lm(math.score ~ gender_male + writing.score + reading.score + ‘test preparation course_none‘
model4 <-  lm(math.score ~ gender_female + writing.score + reading.score + ‘test preparation course_comp
model5 <- lm(writing.score ~ gender_female + math.score + parent.edu + ‘test preparation course_none‘,
```

```
model
```

```
##
## Call:
## lm(formula = math.score ~ gender_female + writing.score + reading.score +
##     ‘test preparation course_completed‘, data = data1)
##
## Coefficients:
##                          (Intercept)                         gender_female
##                             6.3112                             -13.6332
##                      writing.score                        reading.score
##                             0.6978                               0.2983
## ‘test preparation course_completed‘
##                            -3.5816
```

```
model1.2
```

```
##
## Call:
## lm(formula = math.score ~ gender_female + writing.score + reading.score +
##     ‘test preparation course_none‘, data = data1)
##
## Coefficients:
##                     (Intercept)                    gender_female
##                          2.7296                         -13.6332
##                   writing.score                    reading.score
##                          0.6978                           0.2983
## ‘test preparation course_none‘
##                          3.5816
```

```
model2
```

```
##
## Call:
## lm(formula = math.score ~ gender_male + writing.score + reading.score +
##     ‘test preparation course_completed‘, data = data1)
##
## Coefficients:
##                        (Intercept)                          gender_male
##                            -7.3220                              13.6332
##                      writing.score                        reading.score
##                             0.6978                               0.2983
## ‘test preparation course_completed‘
##                            -3.5816
```

```
model2.2
```

```
##
## Call:
## lm(formula = math.score ~ gender_male + writing.score + reading.score +
##     ‘test preparation course_none‘, data = data1)
##
## Coefficients:
##                     (Intercept)                      gender_male
##                        -10.9037                          13.6332
##                   writing.score                    reading.score
##                          0.6978                           0.2983
## ‘test preparation course_none‘
##                          3.5816
```

```
model3
```

```
##
## Call:
## lm(formula = math.score ~ gender_male + writing.score + reading.score +
##     ‘test preparation course_none‘ + parent.edu + ‘race/ethnicity‘,
##     data = data1)
##
## Coefficients:
##                    (Intercept)                     gender_male
##                      -12.46248                        13.68198
##                  writing.score                   reading.score
##                        0.76920                         0.23052
## ‘test preparation course_none‘     parent.edubachelor’s degree
##                        4.00276                        -1.27468
##          parent.eduhigh school       parent.edumaster’s degree
##                        0.77949                        -2.23118
##         parent.edusome college      parent.edusome high school
##                        0.41811                         0.82611
##        ‘race/ethnicity‘group B         ‘race/ethnicity‘group C
##                        0.92754                         0.24643
##        ‘race/ethnicity‘group D         ‘race/ethnicity‘group E
##                       -0.03263                         5.25524
```

```
model4
```

```
##
## Call:
## lm(formula = math.score ~ gender_female + writing.score + reading.score +
##     ‘test preparation course_completed‘ + parent.edu + ‘race/ethnicity‘,
##     data = data1)
##
## Coefficients:
##                         (Intercept)                         gender_female
##                             5.22226                             -13.68198
##                       writing.score                         reading.score
##                             0.76920                               0.23052
## ‘test preparation course_completed‘           parent.edubachelor’s degree
##                            -4.00276                              -1.27468
##               parent.eduhigh school             parent.edumaster’s degree
##                             0.77949                              -2.23118
##              parent.edusome college            parent.edusome high school
##                             0.41811                               0.82611
##             ‘race/ethnicity‘group B               ‘race/ethnicity‘group C
##                             0.92754                               0.24643
##             ‘race/ethnicity‘group D               ‘race/ethnicity‘group E
##                            -0.03263                               5.25524
```

```
model5
```

```
##
## Call:
```

```
## lm(formula = writing.score ~ gender_female + math.score + parent.edu +
##     ‘test preparation course_none‘, data = data1)
##
## Coefficients:
##                   (Intercept)                   gender_female
##                        9.6864                         13.2902
##                    math.score     parent.edubachelor’s degree
##                        0.8337                          1.9695
##          parent.eduhigh school      parent.edumaster’s degree
##                       -1.6460                          3.2240
##         parent.edusome college     parent.edusome high school
##                       -0.2701                         -1.4847
## ‘test preparation course_none‘
##                       -5.2755
```

part6

Conclusion

```
We started this R project with a question. The question was, can we predict student test scores from va
```