Final Project

Aidan Pizzo

December 16, 2021

Dataset

For my final project, I analyzed a dataset from kaggle.com called "Students Performance in Exams" (https://www.kaggle.com/spscientist/students-performance-in-exams). This dataset contains the math, reading, and writing scores of 1000 students. Additional columns record the students' gender, race/ethnicity, parental level of education, lunch (whether or not they have free or reduced lunch), as well as whether or not they completed a test preparation course. I hope to understand how these factors influence the overall marks the students receive. One specific question I hope to answer is: which collection of these different factors result in the highest average academic score (average of math, reading, and writing)? Additionally, what collection of these different factors results in the lowest average academic score? What is the correlation between reading score and writing score? How does a parent's education level affect their child's academic scores?

data_munged.R

```
library(plyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
#load csv file
get_data <- function()
{
  d <- read.csv("../data/StudentsPerformance.csv", header=T, stringsAsFactors=F)
  return(d)
}

#eliminate spaces from column names, edit some column names
clean_column_names <- function(d)
{
  names(d) <- c("gender", "race.ethnicity.group", "parental.education.level",
```

```
                "lunch.type", "test.preparation.course", "math.score", "reading.score",
                "writing.score")
  return(d)
}
```

analysis.R

```
library(plyr)
library(dplyr)

#find the combination of factors with the best and the worst average test scores
find_best_worst_factors <- function(d)
{
  df <- plyr::ddply(d, .(gender, race.ethnicity.group, parental.education.level,
                         lunch.type, test.preparation.course), function(split_frame){
                           avg_math <- split_frame$math.score %>% sum /
                             split_frame$math.score %>%
                             length
                           avg_reading <- split_frame$reading.score %>% sum /
                             split_frame$reading.score %>%
                             length
                           avg_writing <- split_frame$writing.score %>% sum /
                             split_frame$writing.score %>%
                             length
                           avg_score <- (avg_math + avg_reading + avg_writing) / 3
    return(avg_score)
  })
  names(df) <- c("gender", "race.ethnicity.group", "parental.education.level",
    "lunch.type", "test.preparation.course", "avg.score")
  df <- dplyr::arrange(df, desc(avg.score))
  len <- df %>% nrow
  print("Factors with the highest average score: ")
  print(df[1,])
  print("Factors with the lowest average score: ")
  print(df[len,])
  return(df)
}

#sort dataframe by parent's education level
sort_by_parent_eduction <- function(d)
{
  df <- plyr::ddply(d, .(parental.education.level), function(split_frame){
                           avg_math <- split_frame$math.score %>% sum /
                             split_frame$math.score %>%
                             length
                           avg_reading <- split_frame$reading.score %>% sum /
                             split_frame$reading.score %>%
                             length
                           avg_writing <- split_frame$writing.score %>% sum /
                             split_frame$writing.score %>%
                             length
                           avg_score <- (avg_math + avg_reading + avg_writing) / 3
                           out_df <- data.frame(avg.math = avg_math, avg.reading = avg_reading,
```

```
                                    avg.writing = avg_writing, overall.avg = avg_score,
                                    stringsAsFactors=F)
                          return(out_df)
                        })
  df <- dplyr::arrange(df, desc(overall.avg))
  print(df)
  return(df)
}

#find the students who got 100s on all exams
get_100s <- function(d)
{
  df <- dplyr::filter(d, math.score == 100 & reading.score == 100 & writing.score == 100)
  print(df)
  print("Number of people with all perfect scores: ")
  len <- df %>% nrow
  print(len)
}

#find the subject with the most 100s scored
most_aced_section <- function(d)
{
  aced_math <- dplyr::filter(d, math.score == 100)
  num_aced_math <- aced_math %>% nrow
  aced_reading <- dplyr::filter(d, reading.score == 100)
  num_aced_reading <- aced_reading %>% nrow
  aced_writing <- dplyr::filter(d, writing.score == 100)
  num_aced_writing <- aced_writing %>% nrow
  subject <- c("Math", "Reading", "Writing")
  num_100s <- c(num_aced_math, num_aced_reading, num_aced_writing)
  num_aces <- data.frame(subject, num_100s, stringsAsFactors = F)
  max_aces <- which.max(num_100s)
  print("Subject with the most 100s is: ")
  print(subject[max_aces])
  return(num_aces)
}
```

presentation.R

```
library(ggplot2)

#boxplot showing difference in average score for each of three subjects
plot_gender <- function(d)
{
  p <- ggplot(d, aes(x = gender, y= math.score)) + geom_boxplot(outlier.colour="red") +
    labs(title="Math Score by Gender",x="gender", y = "math score")
  print(p)
  p <- ggplot(d, aes(x = gender, y= reading.score)) + geom_boxplot(outlier.colour="red")+
    labs(title="Reading Score by Gender",x="gender", y = " reading score")
  print(p)
  p <- ggplot(d, aes(x = gender, y= writing.score)) + geom_boxplot(outlier.colour="red")+
    labs(title="Writing Score by Gender",x="gender", y = "writing score")
  print(p)
```

```
}

#point graph showing correlation between reading score and writing score
plot_reading_writing <- function(d)
{
  p <- ggplot(d, aes(x = reading.score, y = writing.score)) + geom_point(size=1)+
    labs(title="Reading Score vs Writing Score", x = "reading score", y = "writing score")+
    geom_smooth()
  print(p)
}
```

configuration.R

```
source("data_munged.R")
source("analysis.R")
source("presentation.R")
```

FinalProject_AidanPizzo.R

```
source("configuration.R")

d <- get_data()
d <- clean_column_names(d)

df1 <- find_best_worst_factors(d)
```

```
## [1] "Factors with the highest average score: "
##   gender race.ethnicity.group parental.education.level   lunch.type
## 1 female              group D        bachelor's degree free/reduced
##   test.preparation.course avg.score
## 1               completed  97.66667
## [1] "Factors with the lowest average score: "
##     gender race.ethnicity.group parental.education.level   lunch.type
## 211   male              group B              high school free/reduced
##     test.preparation.course avg.score
## 211                    none  44.73333
```

```
df2 <- sort_by_parent_eduction(d)
```

```
##   parental.education.level avg.math avg.reading avg.writing overall.avg
## 1          master's degree 69.74576    75.37288    75.67797    73.59887
## 2        bachelor's degree 69.38983    73.00000    73.38136    71.92373
## 3       associate's degree 67.88288    70.92793    69.89640    69.56907
## 4             some college 67.12832    69.46018    68.84071    68.47640
## 5         some high school 63.49721    66.93855    64.88827    65.10801
## 6              high school 62.13776    64.70408    62.44898    63.09694
```

```
most_aced_section(d)
```

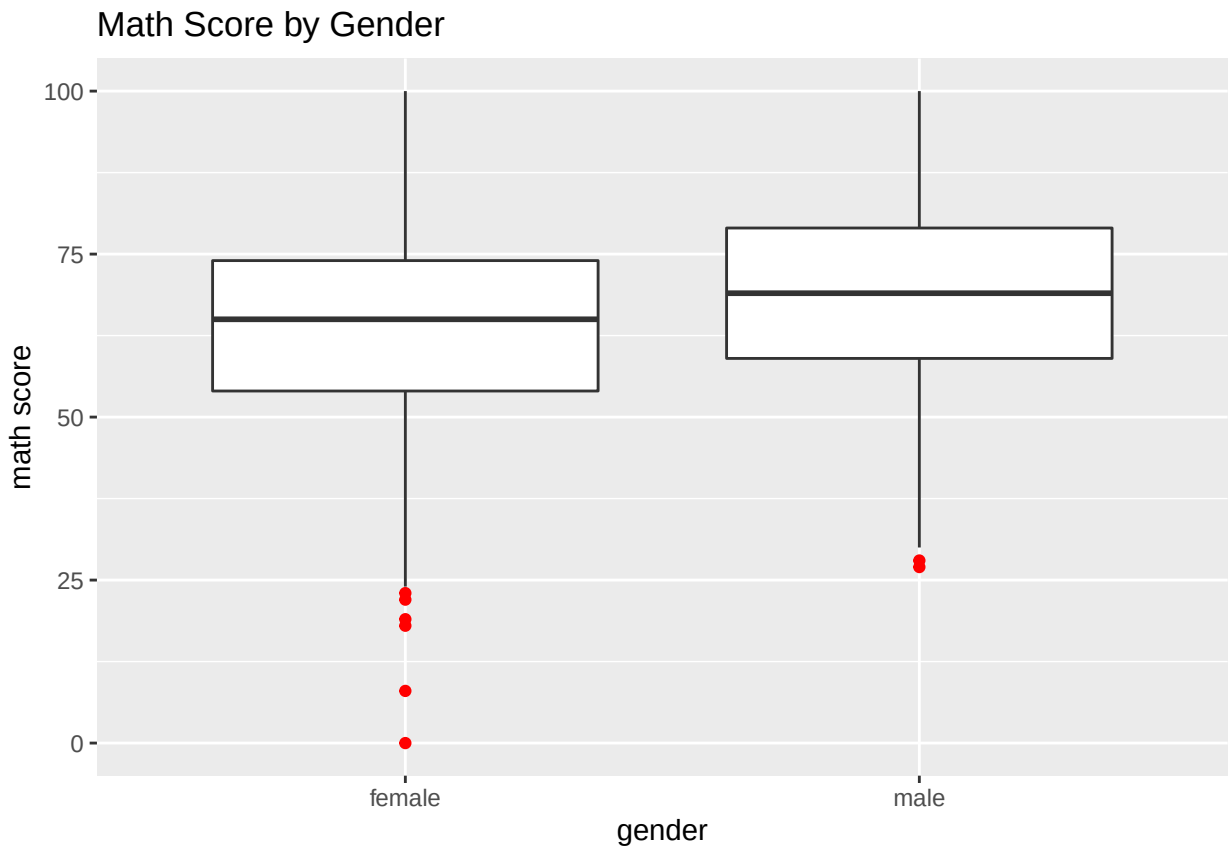
```
## [1] "Subject with the most 100s is: "
## [1] "Reading"
```

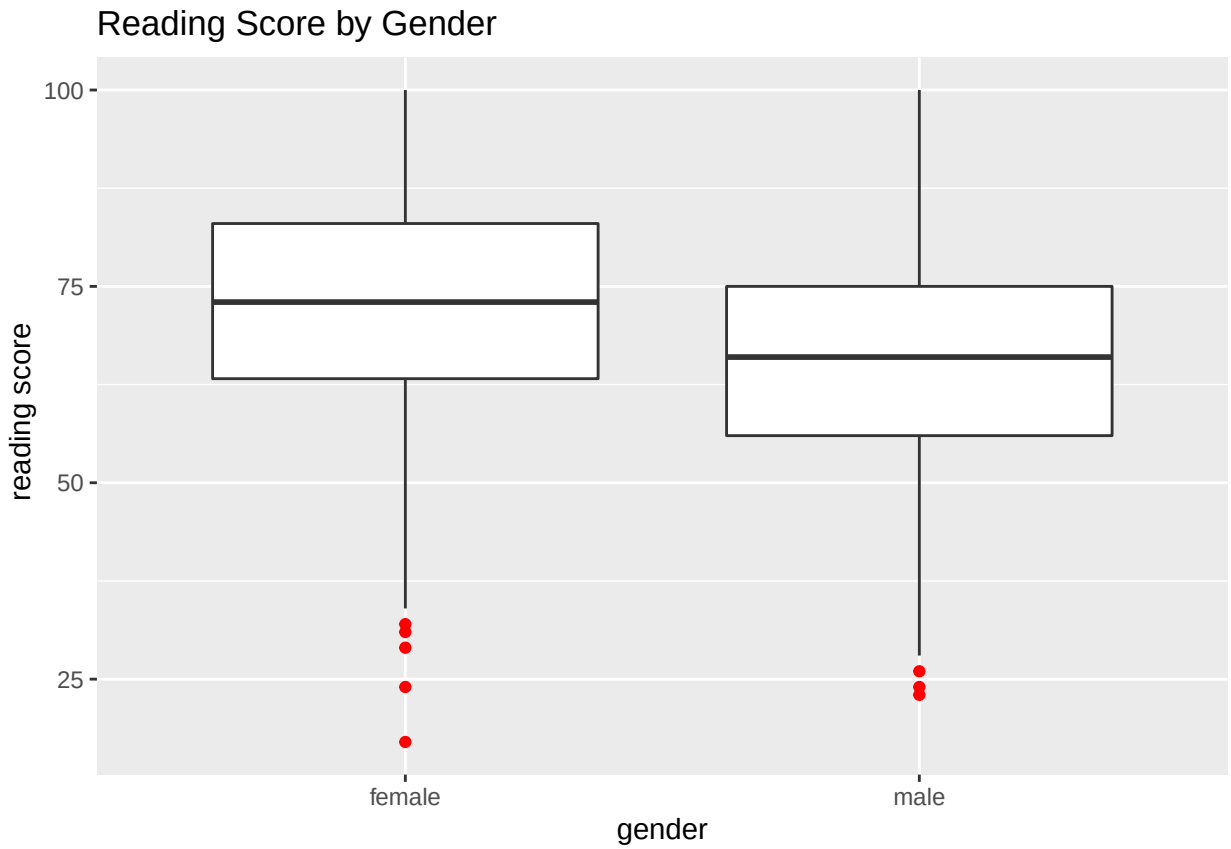
```
##   subject num_100s
## 1    Math        7
## 2 Reading       17
```

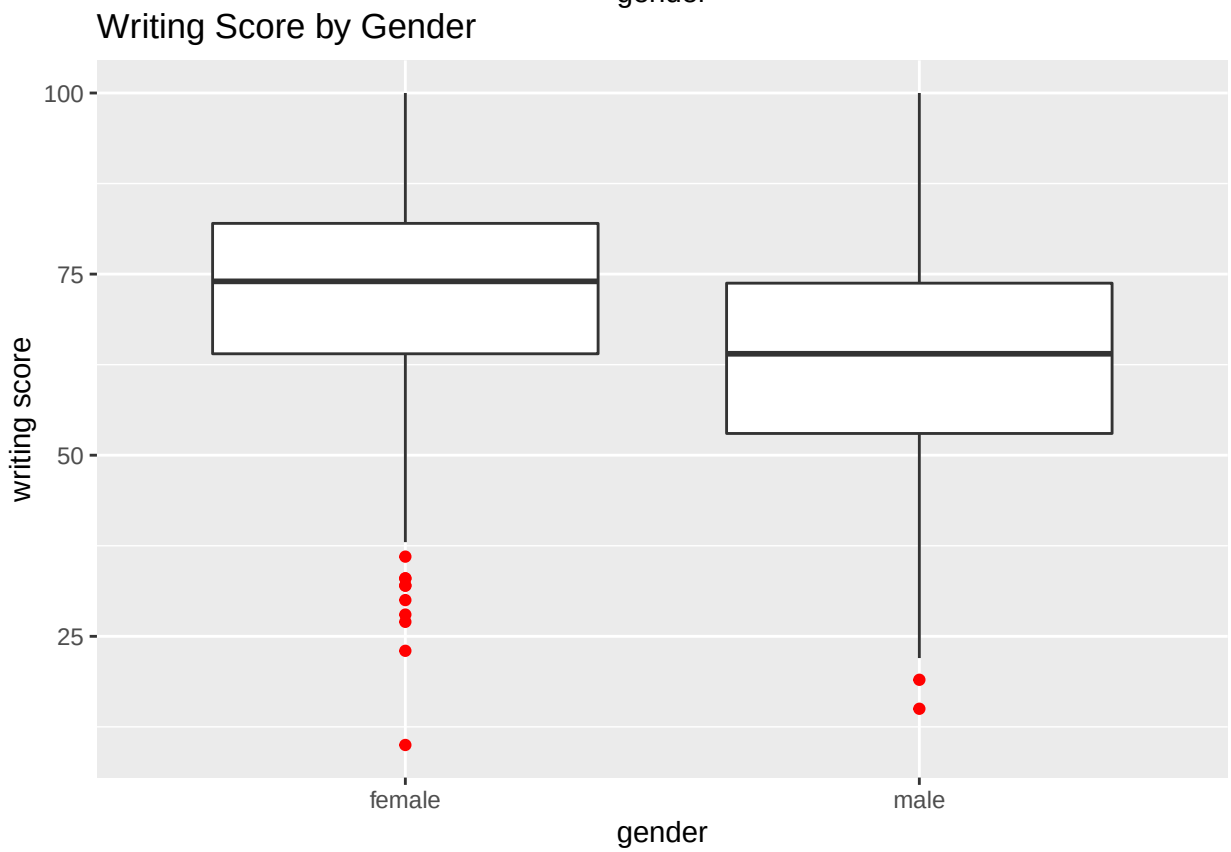
```
## 3 Writing       14
```

The following plots contain box plots separated by gender, containing scores for the math, reading, and writing exams.

```
plot_gender(d)
```





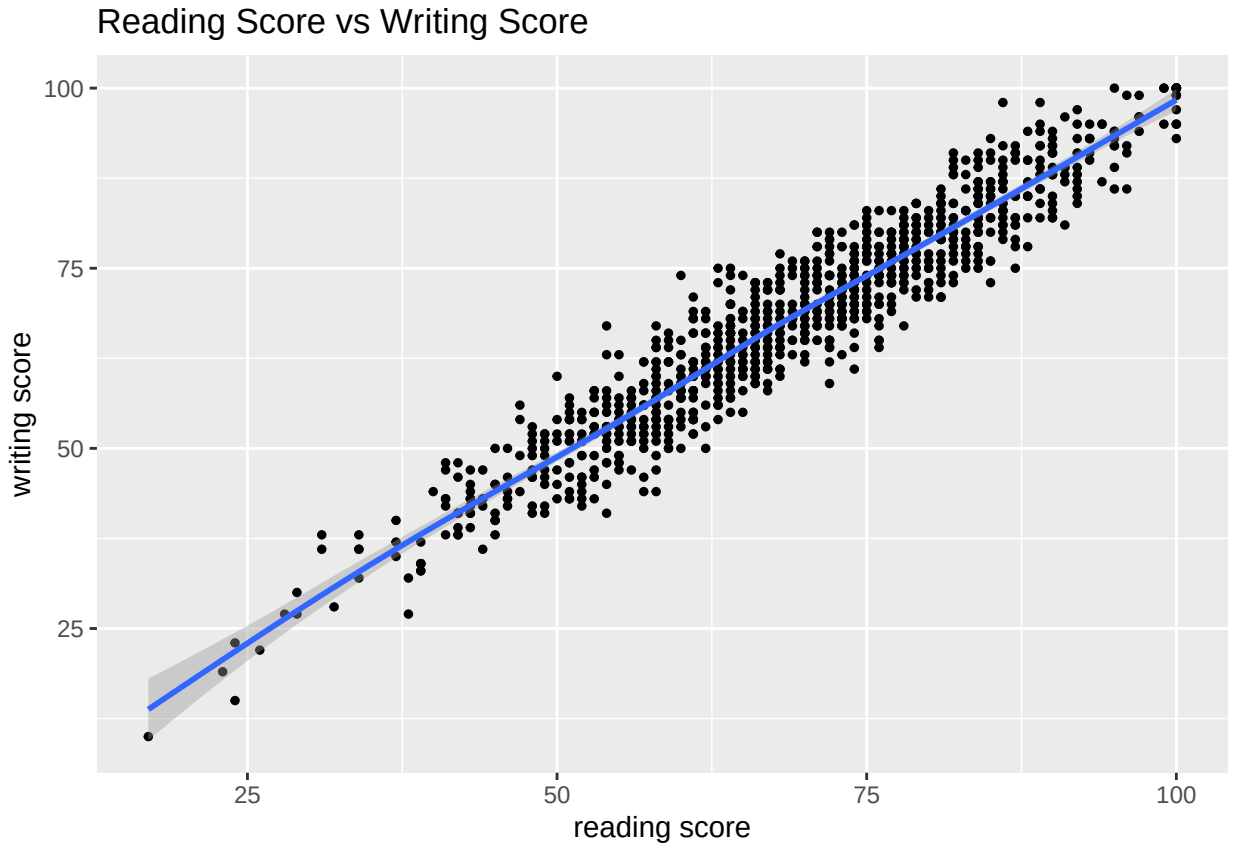


The following plot is a point graph that shows the correlation between the reading score and the writing

score. The best fit line is shown in blue as a geom_smooth graph.

```
plot_reading_writing(d)
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



AWS

I verify that I have terminated all EC2 instances, deleted all AMI, volumes, and snapshots, and deleted any S3 buckets that I created.