Student Performance Data based on Background

Akshay Thaper

My dataset

For my presentation, I downloaded Student Performance data from Kaggle (https://www.kaggle.com/spscientist/students-performance-in-exams#StudentsPerformance.csv). The inspiration for this dataset was to understand the influence of a student's background on his/her academic performance.

install packages outside of base R

```
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.3.0     v forcats 0.3.0
```

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

```
library(ggplot2)
```

Importing Data

I downloaded the file "StudentsPerformance.csv" and imported it into RStudio.

```
stuPer <- read.csv("StudentsPerformance.csv")
```

Its dimensions are 1000x8, with continuous variables such as test scores in reading, writing, and math, and categorical variables like race, gender, and socioeconomic characteristics. Here is a preview of the dataset:

```
head(stuPer)
```

```
##   gender race.ethnicity parental.level.of.education        lunch
## 1 female        group B           bachelor's degree     standard
## 2 female        group C                some college     standard
## 3 female        group B             master's degree     standard
## 4   male        group A          associate's degree free/reduced
## 5   male        group C                some college     standard
## 6 female        group B          associate's degree     standard
##   test.preparation.course math.score reading.score writing.score
## 1                    none         72            72            74
## 2               completed         69            90            88
## 3                    none         90            95            93
## 4                    none         47            57            44
## 5                    none         76            78            75
## 6                    none         71            83            78
```

Data Manipulation

Luckily, the dataset was already pretty cleaned up and the variables had reasonable names.

I manipulated the data by combining the three scores for reading, writing, and math into one composite score as a measure of student performance.

I also removed the columns for individual math, reading, and writing scores.

Once I had the composite score, I created a new column for the percentile rank of each student.

```r
stuPer <- stuPer %>% mutate(composite.score = math.score + reading.score + writing.score) %>%
  select(-math.score, -reading.score, -writing.score) %>%
  mutate(percentile.rank = percent_rank(composite.score)*100)
```

Next, I looked at the summary statistics:

```r
summary(stuPer)
```

```
##     gender    race.ethnicity     parental.level.of.education
##  female:518   group A: 89    associate's degree:222
##  male  :482   group B:190    bachelor's degree :118
##               group C:319    high school       :196
##               group D:262    master's degree   : 59
##               group E:140    some college      :226
##                              some high school  :179
##          lunch     test.preparation.course composite.score
##  free/reduced:355   completed:358           Min.   : 27.0
##  standard    :645   none     :642           1st Qu.:175.0
##                                             Median :205.0
##                                             Mean   :203.3
##                                             3rd Qu.:233.0
##                                             Max.   :300.0
##  percentile.rank
##  Min.   : 0.00
##  1st Qu.:24.72
##  Median :48.95
##  Mean   :49.68
##  3rd Qu.:74.67
##  Max.   :99.80
```

```r
summary(stuPer$composite.score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    27.0   175.0   205.0   203.3   233.0   300.0
```

In looking at the data, I suspected that there may be a few outliers. I used a formula to find the lower and upper bounds of the composite scores and excluded any outliers.

```r
#IQR is the inter-quartile range, 233 represents the 3rd Quartile and 175 represents the 1st Quartile
IQR <- 233 - 175
lowBound <- 175 - 1.5*IQR
highBound <- 233 + 1.5*IQR

cat("Composite scores below", lowBound, "and above", highBound, "will be excluded.
    There are no outliers on the high end because the max is 300.")
```

```
## Composite scores below 88 and above 320 will be excluded.
##     There are no outliers on the high end because the max is 300.
```

I created a new dataset without the outliers.

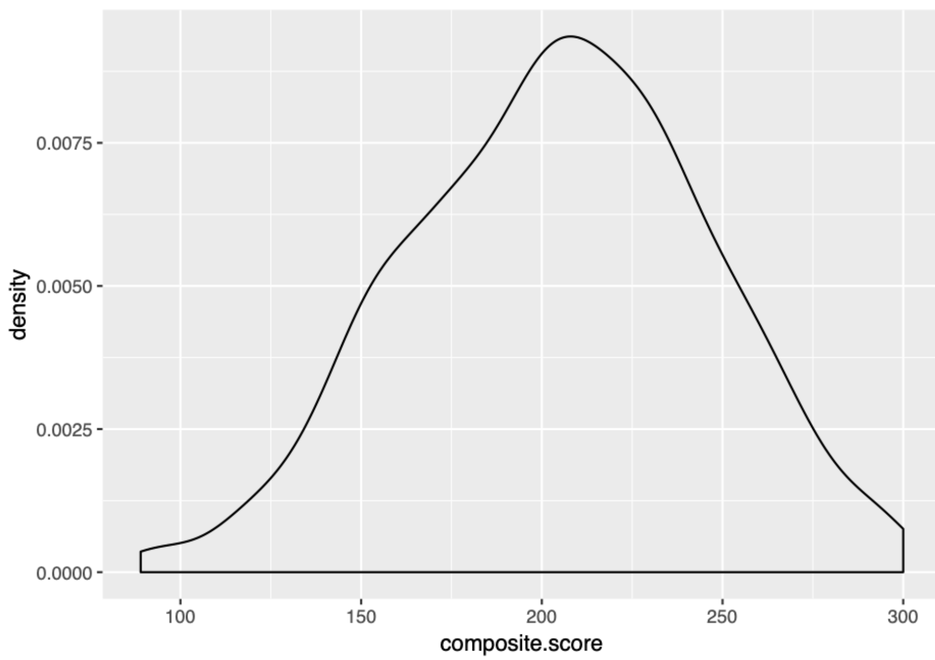
```
stuPerOut <-stuPer
stuPerOut <- stuPerOut %>% filter(stuPer$composite.score > lowBound)
summary(stuPerOut$composite.score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    89.0   175.0   206.0   204.3   234.0   300.0
```

Graphing

I wanted to see if the scores were normally distributed, so I graphed a density plot of the composite scores after taking care of the outliers.

```
ggplot(stuPerOut, aes(composite.score)) + geom_density()
```




```
shapiro.test(stuPerOut$composite.score)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  stuPerOut$composite.score
## W = 0.99547, p-value = 0.004912
```

The p-value of the Shapiro-Wilk test indicates that we should reject the null hypothesis and that this is not a normally distributed dataset. I will use non-parametric statistics to analyze this data.

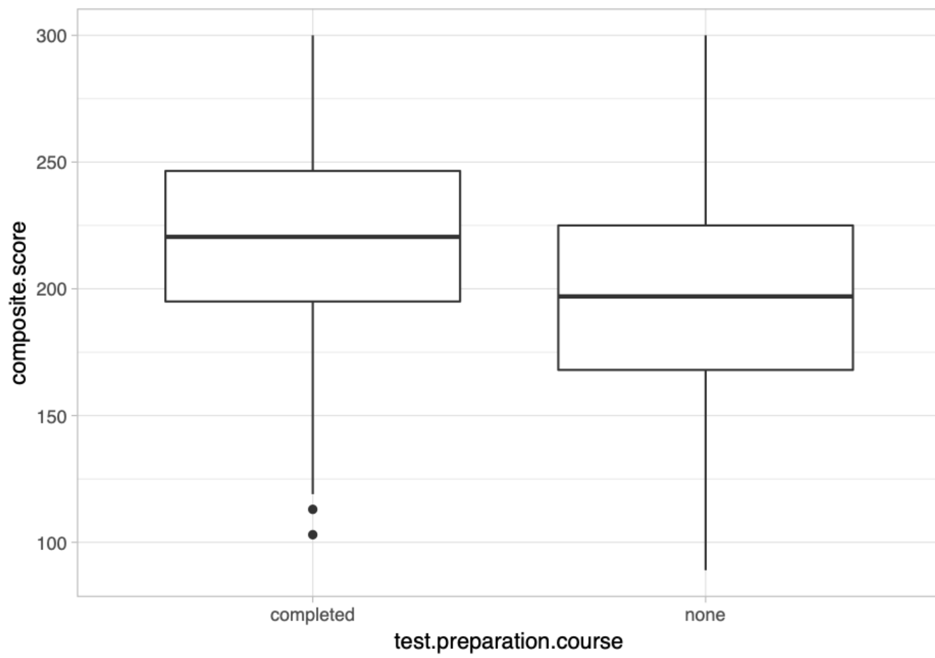
Statistical Analysis

Statistical analysis of one continuous variable (composite test score) and one categorical variable (completion of prep course):

Question of interest - Does the completion of the prep course correlate with higher composite scores?

I first graphed the scores of students who completed the prep course and those who did not take a prep course:

```r
ggplot(stuPerOut, aes(x=test.preparation.course, y = composite.score)) + geom_boxplot() + theme_light()
```



I used a non-parametric test (Wilcox) to test the null hypothesis that completion of the prep course and composite scores are independent of each other.

```r
wilcox.test(composite.score ~ test.preparation.course, data = stuPerOut)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  composite.score by test.preparation.course
## W = 147810, p-value = 3.56e-15
## alternative hypothesis: true location shift is not equal to 0
```

Since the p-value was below .05, this indicates that we should reject the null hypothesis and that there is a correlation between completing the prep course and the composite scores.