Student Performance in Exams

Pinak Kapalay

December 31, 2018

Dataset: We have a data set named as students performance in a test. The Variables being: Gender, Race.Ethnicity, Parental Level fo Education, Lunch, Test Course Preparation, Math.score, Reading.score and Writing Score.

Inspiration: To find out, the significant variables affecting the scores of students and studying their relationship with the Math, Reading and Writing scores. Finally giving a suitable conclusion from the findings.

Title: Students Performance in the test.

library(ggplot2)  
library(corrplot)

## corrplot 0.84 loaded

library(MASS)  
library(DAAG)

## Loading required package: lattice

##   
## Attaching package: 'DAAG'

## The following object is masked from 'package:MASS':  
##   
## hills

library(HH)

## Loading required package: grid

## Loading required package: latticeExtra

## Loading required package: RColorBrewer

##   
## Attaching package: 'latticeExtra'

## The following object is masked from 'package:ggplot2':  
##   
## layer

## Loading required package: multcomp

## Loading required package: mvtnorm

## Loading required package: survival

##   
## Attaching package: 'survival'

## The following object is masked from 'package:DAAG':  
##   
## lung

## Loading required package: TH.data

##   
## Attaching package: 'TH.data'

## The following object is masked from 'package:MASS':  
##   
## geyser

## Loading required package: gridExtra

##   
## Attaching package: 'HH'

## The following objects are masked from 'package:DAAG':  
##   
## powerplot, vif

library(caret)

##   
## Attaching package: 'caret'

## The following object is masked from 'package:survival':  
##   
## cluster

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following object is masked from 'package:gridExtra':  
##   
## combine

## The following object is masked from 'package:MASS':  
##   
## select

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

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

library(RColorBrewer)  
library(ggthemes)  
library(ggrepel)  
library(gridExtra)  
library(descr)

sp <- read.csv("StudentsPerformance.csv", na.strings = c("","NA"))  
View(sp)  
str(sp)

## 'data.frame': 1000 obs. of 8 variables:  
## $ gender : Factor w/ 2 levels "female","male": 1 1 1 2 2 1 1 2 2 1 ...  
## $ race.ethnicity : Factor w/ 5 levels "group A","group B",..: 2 3 2 1 3 2 2 2 4 2 ...  
## $ parental.level.of.education: Factor w/ 6 levels "associate's degree",..: 2 5 4 1 5 1 5 5 3 3 ...  
## $ lunch : Factor w/ 2 levels "free/reduced",..: 2 2 2 1 2 2 2 1 1 1 ...  
## $ test.preparation.course : Factor w/ 2 levels "completed","none": 2 1 2 2 2 2 1 2 1 2 ...  
## $ math.score : int 72 69 90 47 76 71 88 40 64 38 ...  
## $ reading.score : int 72 90 95 57 78 83 95 43 64 60 ...  
## $ writing.score : int 74 88 93 44 75 78 92 39 67 50 ...

summary(sp)

## 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 math.score   
## free/reduced:355 completed:358 Min. : 0.00   
## standard :645 none :642 1st Qu.: 57.00   
## Median : 66.00   
## Mean : 66.09   
## 3rd Qu.: 77.00   
## Max. :100.00   
## reading.score writing.score   
## Min. : 17.00 Min. : 10.00   
## 1st Qu.: 59.00 1st Qu.: 57.75   
## Median : 70.00 Median : 69.00   
## Mean : 69.17 Mean : 68.05   
## 3rd Qu.: 79.00 3rd Qu.: 79.00   
## Max. :100.00 Max. :100.00

Comments: We view the Data Frame, Structure and Summary of the data frame sp.

apply(is.na(sp),2,sum)

## gender race.ethnicity   
## 0 0   
## parental.level.of.education lunch   
## 0 0   
## test.preparation.course math.score   
## 0 0   
## reading.score writing.score   
## 0 0

Comments: Checking for any NAs available in the data frame.

cat.math.score <- as.factor(sp$math.score)  
cat.reading.score <- as.factor(sp$reading.score)  
cat.writing.score <- as.factor(sp$writing.score)  
  
new\_sp <- cbind(sp,cat.math.score,cat.reading.score,cat.writing.score)  
  
View(new\_sp)  
str(new\_sp)

## 'data.frame': 1000 obs. of 11 variables:  
## $ gender : Factor w/ 2 levels "female","male": 1 1 1 2 2 1 1 2 2 1 ...  
## $ race.ethnicity : Factor w/ 5 levels "group A","group B",..: 2 3 2 1 3 2 2 2 4 2 ...  
## $ parental.level.of.education: Factor w/ 6 levels "associate's degree",..: 2 5 4 1 5 1 5 5 3 3 ...  
## $ lunch : Factor w/ 2 levels "free/reduced",..: 2 2 2 1 2 2 2 1 1 1 ...  
## $ test.preparation.course : Factor w/ 2 levels "completed","none": 2 1 2 2 2 2 1 2 1 2 ...  
## $ math.score : int 72 69 90 47 76 71 88 40 64 38 ...  
## $ reading.score : int 72 90 95 57 78 83 95 43 64 60 ...  
## $ writing.score : int 74 88 93 44 75 78 92 39 67 50 ...  
## $ cat.math.score : Factor w/ 81 levels "0","8","18","19",..: 53 50 71 28 57 52 69 21 45 19 ...  
## $ cat.reading.score : Factor w/ 72 levels "17","23","24",..: 45 63 68 30 51 56 68 16 37 33 ...  
## $ cat.writing.score : Factor w/ 77 levels "10","15","19",..: 51 65 70 21 52 55 69 16 44 27 ...

Comments: Changing the numeric variables into catagorical or factors for the chi-sq test. Appending the 3 newly created factor variables to the original dataset.

Checking independence of variables using a Chi-Sq test: Generally, a chi-sq test is employed to check the dependence of two variables (factors) with levels > 2. We will check the independence of Math, Reading and Writing scores with respect to Parental Level of Education and Race.Ethnicity.

Checking Independence of math score:

table(new\_sp$parental.level.of.education,new\_sp$cat.math.score)

##   
## 0 8 18 19 22 23 24 26 27 28 29 30 32 33 34 35 36 37  
## associate's degree 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1  
## bachelor's degree 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 2  
## high school 0 1 0 0 0 1 0 0 1 0 1 1 0 1 2 2 2 0  
## master's degree 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## some college 0 0 0 1 1 0 0 0 0 1 0 0 1 0 0 2 0 0  
## some high school 1 0 1 0 0 0 1 0 1 0 1 1 2 0 0 1 0 1  
##   
## 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55  
## associate's degree 0 1 4 1 1 2 1 0 6 4 1 4 1 2 7 7 4 4  
## bachelor's degree 0 1 0 0 0 2 1 1 0 1 1 1 2 2 2 1 1 2  
## high school 1 1 1 3 3 0 2 3 2 1 2 2 5 1 4 6 4 3  
## master's degree 0 0 2 0 0 0 0 1 1 1 0 1 2 0 2 1 3 1  
## some college 0 0 2 1 2 0 1 2 1 1 1 5 3 1 2 7 5 4  
## some high school 2 1 1 1 0 1 4 2 1 3 6 4 2 5 1 2 1 4  
##   
## 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73  
## associate's degree 2 8 5 5 2 6 6 2 5 12 4 6 6 3 1 5 3 6  
## bachelor's degree 1 0 1 4 0 4 4 5 2 5 5 4 4 2 3 3 1 2  
## high school 2 8 7 6 6 5 7 4 4 4 8 4 5 7 6 5 7 5  
## master's degree 1 0 1 0 0 2 3 0 1 1 0 1 0 1 1 1 1 3  
## some college 1 0 11 8 5 6 9 11 3 5 4 6 6 14 5 8 3 4  
## some high school 2 2 0 9 3 4 6 4 5 9 3 5 5 5 2 4 3 7  
##   
## 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91  
## associate's degree 4 4 4 5 3 3 6 7 8 3 2 5 0 6 1 1 5 4  
## bachelor's degree 3 4 1 4 3 4 3 1 1 3 0 1 2 2 2 1 0 2  
## high school 5 6 4 2 2 1 1 4 2 0 2 0 2 2 3 1 1 0  
## master's degree 1 0 0 3 1 4 1 4 1 0 1 2 0 1 1 1 2 1  
## some college 5 5 8 5 2 7 2 3 4 2 4 3 2 4 6 0 0 2  
## some high school 7 2 4 5 3 3 4 3 2 0 2 3 2 1 2 2 0 0  
##   
## 92 93 94 95 96 97 98 99 100  
## associate's degree 1 1 2 1 1 3 1 0 2  
## bachelor's degree 2 1 1 0 2 1 0 1 2  
## high school 0 0 1 0 0 0 0 1 0  
## master's degree 1 0 1 1 0 0 0 0 0  
## some college 0 2 0 0 0 1 2 1 3  
## some high school 2 0 2 0 0 1 0 0 0

chisq.test(table(new\_sp$parental.level.of.education,new\_sp$cat.math.score))

## Warning in chisq.test(table(new\_sp$parental.level.of.education,  
## new\_sp$cat.math.score)): Chi-squared approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: table(new\_sp$parental.level.of.education, new\_sp$cat.math.score)  
## X-squared = 417.05, df = 400, p-value = 0.2683

The p-value for the test is 0.20, hence the two variables are not associated.

table(new\_sp$race.ethnicity,new\_sp$cat.math.score)

##   
## 0 8 18 19 22 23 24 26 27 28 29 30 32 33 34 35 36 37 38 39 40  
## group A 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 1 1 1 1  
## group B 0 1 1 1 0 1 1 0 0 0 0 1 1 0 0 0 1 1 1 0 2  
## group C 1 0 0 0 1 0 0 0 1 0 2 0 1 1 1 4 1 1 0 1 3  
## group D 0 0 0 0 0 0 0 1 1 0 1 0 0 0 0 1 0 0 0 2 4  
## group E 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 1 1 0 0  
##   
## 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61  
## group A 1 0 0 2 2 1 3 2 2 3 2 0 4 2 3 1 2 2 4 0 4  
## group B 2 1 0 1 1 2 3 5 5 3 1 7 4 5 2 0 4 8 5 7 6  
## group C 1 2 4 4 2 4 3 2 5 5 4 6 10 8 4 3 4 12 12 5 8  
## group D 1 2 1 2 3 2 2 2 4 2 3 5 3 3 8 4 4 3 8 4 7  
## group E 1 1 0 0 1 2 0 0 1 2 1 0 3 0 1 1 4 0 3 0 2  
##   
## 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82  
## group A 3 2 2 3 2 1 3 2 0 3 3 1 0 3 0 2 2 2 1 1 1  
## group B 7 5 2 8 6 6 3 4 3 3 4 6 5 3 3 4 1 4 2 3 6  
## group C 13 11 7 15 7 11 7 11 7 9 5 8 5 6 6 6 3 7 1 5 4  
## group D 8 7 6 8 4 7 7 14 4 7 3 8 10 7 7 9 6 6 10 7 4  
## group E 4 1 3 2 5 1 6 1 4 4 3 4 5 2 5 3 2 3 3 6 3  
##   
## 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100  
## group A 0 0 1 0 1 0 0 0 1 1 0 0 0 0 1 0 0 1  
## group B 0 1 1 0 4 4 0 3 2 0 0 2 0 0 1 0 0 0  
## group C 6 4 4 3 3 3 0 1 5 2 1 1 0 3 1 2 0 0  
## group D 0 2 5 2 2 7 4 2 0 1 1 0 1 0 1 1 0 1  
## group E 2 4 3 3 6 1 2 2 1 2 2 4 1 0 2 0 3 5

chisq.test(table(new\_sp$race.ethnicity,new\_sp$cat.math.score))

## Warning in chisq.test(table(new\_sp$race.ethnicity, new\_sp$cat.math.score)):  
## Chi-squared approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: table(new\_sp$race.ethnicity, new\_sp$cat.math.score)  
## X-squared = 358.6, df = 320, p-value = 0.06752

The p-value for the test is 0.065, which suggests that, the two variables are not associated much to each other. Hence, we can ignore race.ethnicity in this.

Checking Independence of reading score:

table(new\_sp$parental.level.of.education,new\_sp$cat.reading.score)

##   
## 17 23 24 26 28 29 31 32 34 37 38 39 40 41 42 43 44 45  
## associate's degree 0 0 0 0 0 0 1 0 0 1 0 0 0 1 1 4 0 2  
## bachelor's degree 0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 0 0 1  
## high school 0 0 2 0 0 2 0 0 2 0 0 1 0 2 1 1 2 3  
## master's degree 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0  
## some college 0 1 0 0 1 0 0 0 0 0 1 4 0 1 1 3 2 0  
## some high school 1 0 0 1 0 0 1 1 2 2 1 1 1 1 1 1 0 1  
##   
## 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63  
## associate's degree 0 0 2 1 0 5 2 3 5 5 5 6 4 3 2 8 4 3  
## bachelor's degree 1 0 0 2 0 2 0 0 1 1 2 0 5 2 3 2 4 3  
## high school 3 1 6 1 3 4 9 5 2 4 4 4 5 1 5 4 6 3  
## master's degree 0 0 0 0 0 0 0 2 0 0 1 1 2 2 1 1 0 1  
## some college 0 2 2 2 1 2 2 1 3 4 1 5 7 5 7 6 6 5  
## some high school 2 1 0 4 3 3 3 2 7 2 3 1 5 4 3 3 2 5  
##   
## 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81  
## associate's degree 3 7 5 5 5 3 6 5 4 8 6 8 11 10 6 2 3 3  
## bachelor's degree 4 1 3 2 3 1 3 2 7 5 5 5 1 0 4 5 0 3  
## high school 7 2 6 5 6 5 6 3 9 7 5 4 5 5 2 3 5 6  
## master's degree 2 1 1 3 2 0 3 2 2 0 1 0 1 1 2 1 1 4  
## some college 9 6 5 7 4 2 5 5 7 5 9 6 4 8 9 3 3 7  
## some high school 7 2 7 8 3 4 3 3 5 5 7 3 3 0 4 5 2 2  
##   
## 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 99  
## associate's degree 5 2 7 5 5 4 2 4 3 1 3 3 2 3 1 0 0  
## bachelor's degree 2 3 2 2 2 0 2 1 4 0 2 1 1 0 1 1 0  
## high school 3 1 2 1 1 2 2 1 0 2 1 1 0 1 0 0 1  
## master's degree 2 0 3 2 1 0 0 1 1 2 1 0 0 2 0 1 2  
## some college 3 5 4 5 6 5 1 4 5 0 1 1 0 2 2 2 0  
## some high school 5 3 8 3 4 2 2 1 4 1 2 0 0 0 0 1 0  
##   
## 100  
## associate's degree 4  
## bachelor's degree 8  
## high school 0  
## master's degree 2  
## some college 1  
## some high school 2

chisq.test(table(new\_sp$parental.level.of.education,new\_sp$cat.reading.score))

## Warning in chisq.test(table(new\_sp$parental.level.of.education,  
## new\_sp$cat.reading.score)): Chi-squared approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: table(new\_sp$parental.level.of.education, new\_sp$cat.reading.score)  
## X-squared = 376.85, df = 355, p-value = 0.2036

The p-value is 0.2036, hence both variables are independent.

table(new\_sp$race.ethnicity,new\_sp$cat.reading.score)

##   
## 17 23 24 26 28 29 31 32 34 37 38 39 40 41 42 43 44 45 46 47 48  
## group A 0 1 0 0 0 0 1 0 0 0 0 1 0 1 0 3 0 2 1 2 1  
## group B 0 0 2 0 0 1 0 1 0 0 2 1 0 1 1 3 2 1 2 0 4  
## group C 1 0 0 0 1 1 0 0 1 2 0 3 1 2 2 1 2 1 2 0 3  
## group D 0 0 0 0 0 0 1 0 2 1 0 1 0 2 4 0 0 0 1 2 2  
## group E 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 2 0 3 0 0 0  
##   
## 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69  
## group A 2 2 3 1 2 1 1 0 3 4 3 3 2 2 2 2 2 1 5 2 0  
## group B 2 2 2 4 4 7 3 4 1 4 1 4 4 6 4 6 8 4 6 5 3  
## group C 2 1 7 4 6 4 5 5 3 9 6 8 11 6 5 11 4 11 12 5 8  
## group D 3 1 1 6 1 6 4 1 8 8 7 4 6 4 7 8 4 10 4 7 3  
## group E 1 1 3 1 0 0 3 6 2 3 0 2 1 4 2 5 1 1 3 4 1  
##   
## 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90  
## group A 3 0 4 3 2 0 0 0 2 1 1 1 3 1 2 2 0 1 1 0 1  
## group B 7 3 5 2 1 4 5 4 4 2 2 4 5 4 5 7 5 2 0 2 4  
## group C 2 5 15 12 10 11 8 11 9 7 6 9 4 4 9 3 5 4 3 4 6  
## group D 11 10 7 7 11 7 3 8 10 6 3 7 7 2 7 0 6 3 2 5 4  
## group E 3 2 3 6 9 4 9 1 2 3 2 4 1 3 3 6 3 3 3 1 2  
##   
## 91 92 93 94 95 96 97 99 100  
## group A 0 1 1 0 0 2 0 0 1  
## group B 1 0 0 1 3 1 2 0 0  
## group C 2 3 4 1 2 1 0 0 3  
## group D 2 3 0 0 2 0 3 1 6  
## group E 1 3 1 1 1 0 0 2 7

chisq.test(table(new\_sp$race.ethnicity,new\_sp$cat.reading.score))

## Warning in chisq.test(table(new\_sp$race.ethnicity,  
## new\_sp$cat.reading.score)): Chi-squared approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: table(new\_sp$race.ethnicity, new\_sp$cat.reading.score)  
## X-squared = 334.17, df = 284, p-value = 0.02168

The p-value for the two variables is 0.02168, hence, the two variables are not independent of each other. They have relation with each other, i.e. race affects the reading score.

Checking the independence of writing scores:

table(new\_sp$parental.level.of.education,new\_sp$cat.writing.score)

##   
## 10 15 19 22 23 27 28 30 32 33 34 35 36 37 38 39 40 41  
## associate's degree 0 0 0 0 0 0 0 0 0 0 0 1 0 0 2 0 1 2  
## bachelor's degree 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0  
## high school 0 1 0 0 1 1 0 1 0 0 1 0 3 0 0 1 1 2  
## master's degree 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## some college 0 0 1 0 0 1 0 0 1 2 1 0 0 1 1 1 0 1  
## some high school 1 0 0 1 0 1 1 0 1 0 1 0 1 1 1 0 1 3  
##   
## 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59  
## associate's degree 1 2 2 1 1 1 2 1 4 2 3 7 4 2 5 6 7 1  
## bachelor's degree 0 0 1 0 1 3 2 0 0 2 0 0 2 1 1 3 1 1  
## high school 3 3 1 1 5 5 2 5 3 7 6 2 6 5 3 2 2 4  
## master's degree 0 0 0 0 1 0 0 0 1 0 1 0 3 0 0 0 1 1  
## some college 3 3 0 2 1 1 1 1 0 2 6 2 6 3 2 4 6 2  
## some high school 0 4 4 3 1 0 1 2 2 4 4 6 4 3 3 4 1 4  
##   
## 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77  
## associate's degree 2 3 4 6 5 5 2 7 7 5 7 2 6 11 9 1 6 7  
## bachelor's degree 1 3 6 0 2 1 4 2 3 2 5 1 5 2 6 4 3 0  
## high school 6 2 4 5 8 6 5 3 11 4 4 8 2 5 8 4 3 2  
## master's degree 0 1 1 1 0 1 2 3 2 1 1 1 3 1 3 4 0 1  
## some college 7 5 7 2 3 5 9 7 3 5 14 5 3 6 7 3 9 7  
## some high school 4 7 5 1 4 4 2 3 5 6 2 1 6 3 2 6 4 7  
##   
## 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95  
## associate's degree 5 4 8 3 6 3 4 3 3 2 3 2 4 2 4 3 2 5  
## bachelor's degree 1 4 2 7 2 5 2 0 1 4 1 3 1 1 2 1 0 2  
## high school 3 4 5 2 4 1 2 2 2 0 0 2 1 0 0 0 0 0  
## master's degree 2 0 0 1 2 0 4 2 0 2 0 1 1 2 0 1 1 1  
## some college 6 3 5 5 4 4 3 3 4 1 4 1 3 6 2 2 3 0  
## some high school 8 4 7 2 6 0 1 4 2 2 3 0 0 0 1 1 0 0  
##   
## 96 97 98 99 100  
## associate's degree 0 0 1 1 1  
## bachelor's degree 1 0 0 1 7  
## high school 0 0 0 0 1  
## master's degree 1 0 0 0 4  
## some college 1 1 1 2 0  
## some high school 1 1 0 0 1

chisq.test(table(new\_sp$parental.level.of.education,new\_sp$cat.writing.score))

## Warning in chisq.test(table(new\_sp$parental.level.of.education,  
## new\_sp$cat.writing.score)): Chi-squared approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: table(new\_sp$parental.level.of.education, new\_sp$cat.writing.score)  
## X-squared = 446.35, df = 380, p-value = 0.01063

The p-value for the test is 0.01063, hence both these variables are correlated and significant.

table(new\_sp$race.ethnicity,new\_sp$cat.writing.score)

##   
## 10 15 19 22 23 27 28 30 32 33 34 35 36 37 38 39 40 41 42 43 44  
## group A 0 0 1 0 0 0 0 0 0 0 1 0 1 0 0 0 0 2 0 5 2  
## group B 0 1 0 0 1 2 1 0 1 0 1 0 1 0 1 1 0 3 1 0 1  
## group C 1 0 0 0 0 1 0 1 0 2 0 1 1 1 0 1 1 1 5 5 2  
## group D 0 0 0 0 0 0 0 0 1 0 1 0 1 1 3 0 0 1 1 1 2  
## group E 0 0 0 1 0 0 0 0 0 0 0 0 0 0 2 0 2 1 0 1 1  
##   
## 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65  
## group A 1 1 2 2 2 2 1 1 3 2 2 1 2 3 0 2 2 4 1 5 3  
## group B 5 3 4 0 1 3 5 2 2 3 5 3 7 2 1 8 3 4 4 2 5  
## group C 0 2 1 1 5 3 7 8 9 9 6 5 4 6 5 4 9 8 2 7 9  
## group D 0 4 2 4 1 2 2 7 2 5 1 4 2 6 5 5 6 7 7 6 2  
## group E 1 0 1 1 0 0 2 2 1 6 0 1 4 1 2 1 1 4 1 2 3  
##   
## 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86  
## group A 1 2 1 3 2 0 1 3 3 0 2 1 0 1 1 2 3 0 0 1 2  
## group B 6 5 9 2 8 1 5 4 4 2 6 2 7 4 5 4 5 2 1 0 3  
## group C 9 7 11 7 10 8 8 6 12 7 7 9 4 6 8 6 6 3 7 6 3  
## group D 6 8 5 8 6 4 6 10 14 9 4 8 9 6 8 5 6 6 7 4 3  
## group E 2 3 5 3 7 5 5 5 2 4 6 4 5 2 5 3 4 2 1 3 1  
##   
## 87 88 89 90 91 92 93 94 95 96 97 98 99 100  
## group A 1 0 0 0 1 2 1 0 0 0 1 0 0 0  
## group B 5 2 2 0 2 4 1 1 0 1 0 0 0 0  
## group C 4 4 2 7 4 1 3 4 4 0 0 0 2 1  
## group D 0 3 5 2 2 0 1 0 1 3 0 2 2 7  
## group E 1 2 0 1 2 2 2 1 3 0 1 0 0 6

chisq.test(table(new\_sp$race.ethnicity,new\_sp$cat.writing.score))

## Warning in chisq.test(table(new\_sp$race.ethnicity,  
## new\_sp$cat.writing.score)): Chi-squared approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: table(new\_sp$race.ethnicity, new\_sp$cat.writing.score)  
## X-squared = 348.68, df = 304, p-value = 0.03946

The p-value for the test is 0.03946, hence both these variables are correlated and significant.

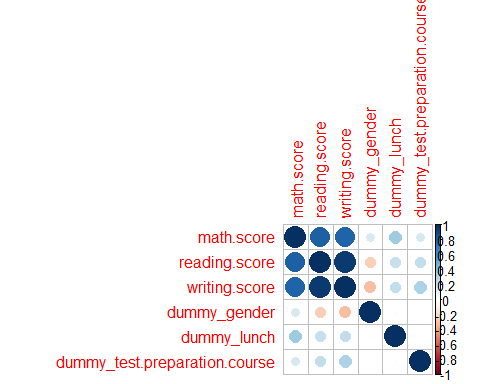
Conclusion from Chi Sq test suggests that, 1. The Parent’s level of education has an impact on the writing scores of the students. 2. The Race.ethnicity has an impact on reading as well as on the writing score. 3. Hence the plot of these variables are significant and rest combinations can very well be ignored.

We create and append the dummy variables (only for factors with level 2) and convert the catagorical values into numeric.

new\_sp$dummy\_gender = ifelse(new\_sp$gender=="male",1,0)  
new\_sp$dummy\_lunch = ifelse(new\_sp$lunch=="standard",1,0)  
new\_sp$dummy\_test.preparation.course = ifelse(new\_sp$test.preparation.course=="completed",1,0)  
  
str(new\_sp)

## 'data.frame': 1000 obs. of 14 variables:  
## $ gender : Factor w/ 2 levels "female","male": 1 1 1 2 2 1 1 2 2 1 ...  
## $ race.ethnicity : Factor w/ 5 levels "group A","group B",..: 2 3 2 1 3 2 2 2 4 2 ...  
## $ parental.level.of.education : Factor w/ 6 levels "associate's degree",..: 2 5 4 1 5 1 5 5 3 3 ...  
## $ lunch : Factor w/ 2 levels "free/reduced",..: 2 2 2 1 2 2 2 1 1 1 ...  
## $ test.preparation.course : Factor w/ 2 levels "completed","none": 2 1 2 2 2 2 1 2 1 2 ...  
## $ math.score : int 72 69 90 47 76 71 88 40 64 38 ...  
## $ reading.score : int 72 90 95 57 78 83 95 43 64 60 ...  
## $ writing.score : int 74 88 93 44 75 78 92 39 67 50 ...  
## $ cat.math.score : Factor w/ 81 levels "0","8","18","19",..: 53 50 71 28 57 52 69 21 45 19 ...  
## $ cat.reading.score : Factor w/ 72 levels "17","23","24",..: 45 63 68 30 51 56 68 16 37 33 ...  
## $ cat.writing.score : Factor w/ 77 levels "10","15","19",..: 51 65 70 21 52 55 69 16 44 27 ...  
## $ dummy\_gender : num 0 0 0 1 1 0 0 1 1 0 ...  
## $ dummy\_lunch : num 1 1 1 0 1 1 1 0 0 0 ...  
## $ dummy\_test.preparation.course: num 0 1 0 0 0 0 1 0 1 0 ...

numric <- new\_sp[sapply(new\_sp,is.numeric)]  
descrCor <- cor(numric)  
View(descrCor)  
corrplot(descrCor)



Here, we find correlation of all the numeric variables and plot it using a corrplot.

highlycorrelated = findCorrelation(descrCor,cutoff = 0.8)  
highlycorrelated

## [1] 3 2

Math score Linear Model:

sp.math <- new\_sp[c(6,12,13,14)]  
  
set.seed(108)  
s <- sample(1:nrow(sp.math),0.75\*nrow(sp.math)) # Using 75% of the original data as out training data and remaining as the testing data.  
train <- sp.math[s,]  
test <- sp.math[-s,]  
View(s)  
View(train)  
View(test)  
ncol(test)

## [1] 4

fit <- lm(math.score~. ,data = train)  
summary(fit)

##   
## Call:  
## lm(formula = math.score ~ ., data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -54.998 -9.739 -0.120 9.934 34.759   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 54.9980 1.0146 54.204 < 2e-16 \*\*\*  
## dummy\_gender 4.2724 0.9952 4.293 1.99e-05 \*\*\*  
## dummy\_lunch 10.4762 1.0362 10.110 < 2e-16 \*\*\*  
## dummy\_test.preparation.course 5.9708 1.0361 5.763 1.21e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 13.58 on 746 degrees of freedom  
## Multiple R-squared: 0.1712, Adjusted R-squared: 0.1679   
## F-statistic: 51.38 on 3 and 746 DF, p-value: < 2.2e-16

The summary gives us the p-values and hence the significant variables which determine the values of math score. The p-values suggest a strong correlation between math score Vs gender,lunch and test.prep.course.

Reading score Linear Model:

sp.read <- new\_sp[c(7,12,13,14)]  
  
set.seed(108)  
s <- sample(1:nrow(sp.read),0.75\*nrow(sp.read))  
train <- sp.read[s,]  
test <- sp.read[-s,]  
View(s)  
View(train)  
View(test)  
ncol(test)

## [1] 4

fit <- lm(reading.score~. ,data = train)  
summary(fit)

##   
## Call:  
## lm(formula = reading.score ~ ., data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -48.707 -8.442 -0.318 9.558 34.416   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 65.7068 0.9750 67.395 < 2e-16 \*\*\*  
## dummy\_gender -8.0264 0.9562 -8.394 2.36e-16 \*\*\*  
## dummy\_lunch 6.7347 0.9957 6.764 2.72e-11 \*\*\*  
## dummy\_test.preparation.course 7.9033 0.9956 7.939 7.51e-15 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 13.05 on 746 degrees of freedom  
## Multiple R-squared: 0.1875, Adjusted R-squared: 0.1842   
## F-statistic: 57.39 on 3 and 746 DF, p-value: < 2.2e-16

The summary gives us the p-values and hence the significant variables which determine the values of reading score. The p-values suggest a strong correlation between reading score Vs gender,lunch and test.prep.course.

Writing score Linear Model:

sp.write <- new\_sp[c(8,12,13,14)]  
  
set.seed(108)  
s <- sample(1:nrow(sp.write),0.75\*nrow(sp.write))   
train <- sp.write[s,]  
test <- sp.write[-s,]  
View(s)  
View(train)  
View(test)  
ncol(test)

## [1] 4

fit <- lm(writing.score~. ,data = train)  
summary(fit)

##   
## Call:  
## lm(formula = writing.score ~ ., data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -54.347 -8.707 0.073 9.490 30.293   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 64.3472 0.9636 66.780 < 2e-16 \*\*\*  
## dummy\_gender -10.2200 0.9451 -10.814 < 2e-16 \*\*\*  
## dummy\_lunch 7.5800 0.9841 7.703 4.25e-14 \*\*\*  
## dummy\_test.preparation.course 10.3280 0.9839 10.497 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 12.9 on 746 degrees of freedom  
## Multiple R-squared: 0.2697, Adjusted R-squared: 0.2668   
## F-statistic: 91.84 on 3 and 746 DF, p-value: < 2.2e-16

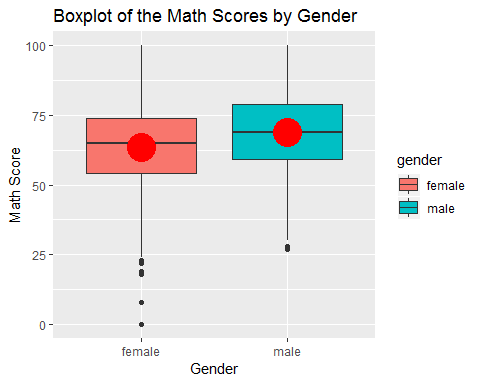
The summary gives us the p-values and hence the significant variables which determine the values of writing score. The p-values suggest a strong correlation between writing score Vs gender,lunch and test.prep.course.

Now we plot significant variables vs the scores and study the plots:

1. Boxplot for Math Score Vs Gender

group\_by(sp,gender) %>%  
 ggplot(aes(x = factor(gender), y = math.score, fill = gender))+geom\_boxplot(colours = rev(brewer.pal(10,'Spectral'))) +  
 stat\_summary(fun.y=mean, geom="point", shape=20, size=15, color="red", fill="red") +  
 xlab("Gender") +ylab("Math Score")+ggtitle('Boxplot of the Math Scores by Gender')

## Warning: Ignoring unknown parameters: colours

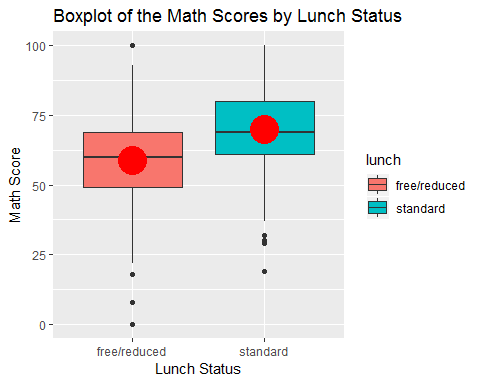


It can be clearly seen that the male students have a higher mark range compared to the female students. The mean for the male catagory is higher compared to the female.

1. Boxplot for Math Score Vs Lunch

group\_by(sp,lunch) %>%  
 ggplot(aes(x = factor(lunch), y = math.score, fill = lunch)) +   
 geom\_boxplot(colours = rev(brewer.pal(10,'Spectral'))) +  
 stat\_summary(fun.y=mean, geom="point", shape=20, size=15, color="red", fill="red") +  
 xlab("Lunch Status") +  
 ylab("Math Score") +ggtitle('Boxplot of the Math Scores by Lunch Status')

## Warning: Ignoring unknown parameters: colours

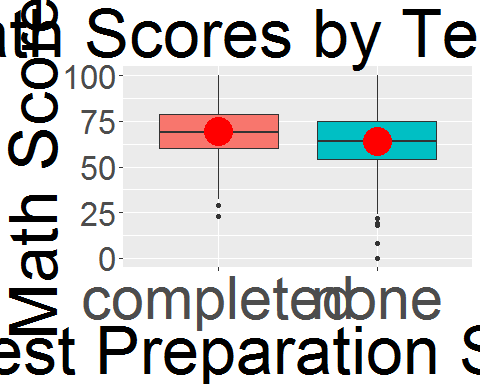


It can very well be concluded that, the students who had standard lunch, have fared well in Math as compared to those who had free/reduced lunch.

1. Boxplot for Math Score Vs Test Preparation Course

sp %>%  
 group\_by(test.preparation.course) %>%  
 ggplot(aes(x = factor(test.preparation.course), y = math.score, fill = test.preparation.course)) +   
 geom\_boxplot(colours = rev(brewer.pal(10,'Spectral'))) +  
 stat\_summary(fun.y=mean, geom="point", shape=20, size=15, color="red", fill="red") +  
 xlab("Test Preparation Status") +  
 ylab("Math Score") +  
 theme(legend.position ='none',axis.title.y = element\_text(size=50), axis.text.y = element\_text(size = 25),axis.title.x = element\_text(size=50),axis.text.x = element\_text(size = 45), plot.title = element\_text(size=50, hjust = 0.5)) +  
 ggtitle('Boxplot of the Math Scores by Test Preparation Status')

## Warning: Ignoring unknown parameters: colours

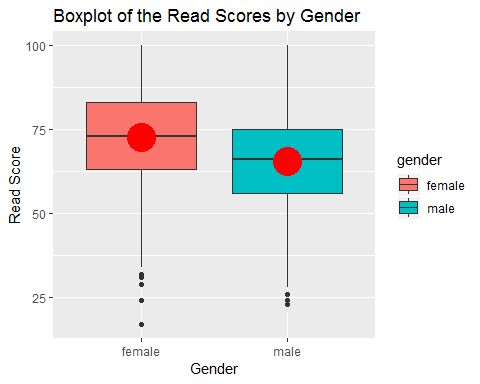


The students who had completed the test preparation course have done well in math compared to those who have not.

1. Boxplot for Reading Score Vs Gender

group\_by(sp,gender) %>%  
 ggplot(aes(x = factor(gender), y = reading.score, fill = gender))+geom\_boxplot(colours = rev(brewer.pal(10,'Spectral'))) +  
 stat\_summary(fun.y=mean, geom="point", shape=20, size=15, color="red", fill="red") +  
 xlab("Gender") +ylab("Read Score")+ggtitle('Boxplot of the Read Scores by Gender')

## Warning: Ignoring unknown parameters: colours

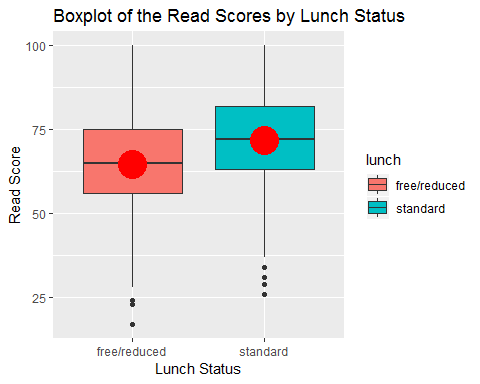


From this plot, it can be concluded, that the female students have performed well at reading and overall have a higer score compared to the male students. Even the mean for the female catagory is higher compared to the male catagory.

1. Boxplot for Reading Score Vs Lunch

group\_by(sp,lunch) %>%  
 ggplot(aes(x = factor(lunch), y = reading.score, fill = lunch)) +   
 geom\_boxplot(colours = rev(brewer.pal(10,'Spectral'))) +  
 stat\_summary(fun.y=mean, geom="point", shape=20, size=15, color="red", fill="red") +  
 xlab("Lunch Status") +  
 ylab("Read Score") +ggtitle('Boxplot of the Read Scores by Lunch Status')

## Warning: Ignoring unknown parameters: colours

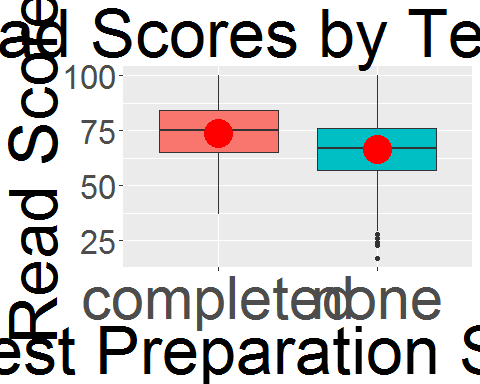


The students who had a standard lunch performed well at reading compared to those who had free/reduced lunch.

1. Boxplot for Reading Score Vs Test Preparation Course

sp %>%  
 group\_by(test.preparation.course) %>%  
 ggplot(aes(x = factor(test.preparation.course), y = reading.score, fill = test.preparation.course)) +   
 geom\_boxplot(colours = rev(brewer.pal(10,'Spectral'))) +  
 stat\_summary(fun.y=mean, geom="point", shape=20, size=15, color="red", fill="red") +  
 xlab("Test Preparation Status") +  
 ylab("Read Score") +  
 theme(legend.position ='none',axis.title.y = element\_text(size=50), axis.text.y = element\_text(size = 25),axis.title.x = element\_text(size=50),axis.text.x = element\_text(size = 45), plot.title = element\_text(size=50, hjust = 0.5)) +  
 ggtitle('Boxplot of the Read Scores by Test Preparation Status')

## Warning: Ignoring unknown parameters: colours

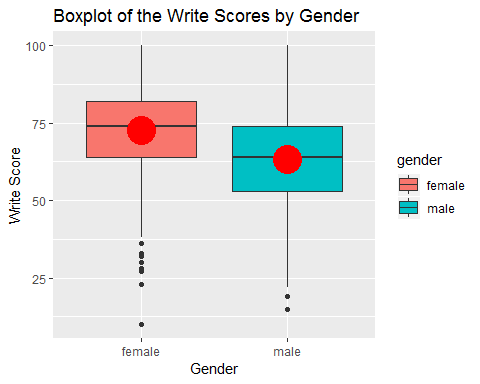


It is clearly visible that the students who completed the test preparation course have performed better at reading compared to those who have not.

1. Boxplot for Writing Score Vs Gender

group\_by(sp,gender) %>%  
 ggplot(aes(x = factor(gender), y = writing.score, fill = gender))+geom\_boxplot(colours = rev(brewer.pal(10,'Spectral'))) +  
 stat\_summary(fun.y=mean, geom="point", shape=20, size=15, color="red", fill="red") +  
 xlab("Gender") +ylab("Write Score")+ggtitle('Boxplot of the Write Scores by Gender')

## Warning: Ignoring unknown parameters: colours

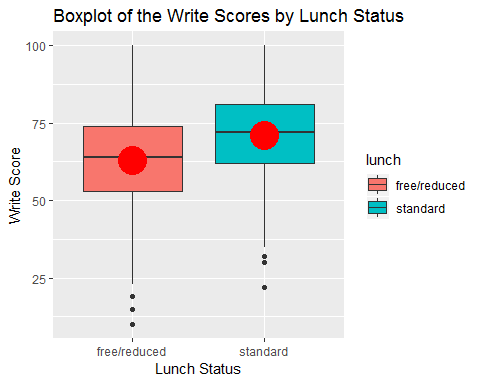


From the above plot, it can be inferred that the female students have performed better at writing compared to th male students.

1. Boxplot for Writing Score Vs Lunch

group\_by(sp,lunch) %>%  
 ggplot(aes(x = factor(lunch), y = writing.score, fill = lunch)) +   
 geom\_boxplot(colours = rev(brewer.pal(10,'Spectral'))) +  
 stat\_summary(fun.y=mean, geom="point", shape=20, size=15, color="red", fill="red") +  
 xlab("Lunch Status") +  
 ylab("Write Score") +ggtitle('Boxplot of the Write Scores by Lunch Status')

## Warning: Ignoring unknown parameters: colours

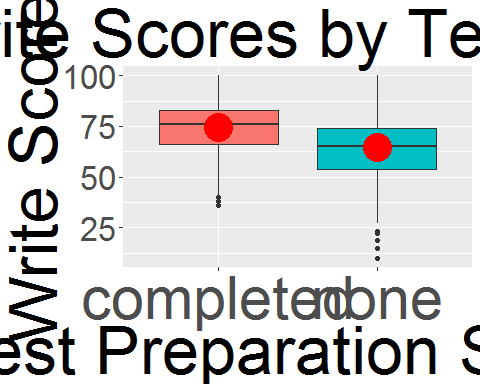


The students who had a standard lunch fared well at writing compared to those who had free/reduced lunch.

1. Boxplot for Writing Score Vs Test Preparation Course

sp %>%  
 group\_by(test.preparation.course) %>%  
 ggplot(aes(x = factor(test.preparation.course), y = writing.score, fill = test.preparation.course)) +   
 geom\_boxplot(colours = rev(brewer.pal(10,'Spectral'))) +  
 stat\_summary(fun.y=mean, geom="point", shape=20, size=15, color="red", fill="red") +  
 xlab("Test Preparation Status") +  
 ylab("Write Score") +  
 theme(legend.position ='none',axis.title.y = element\_text(size=50), axis.text.y = element\_text(size = 25),axis.title.x = element\_text(size=50),axis.text.x = element\_text(size = 45), plot.title = element\_text(size=50, hjust = 0.5)) +  
 ggtitle('Boxplot of the Write Scores by Test Preparation Status')

## Warning: Ignoring unknown parameters: colours

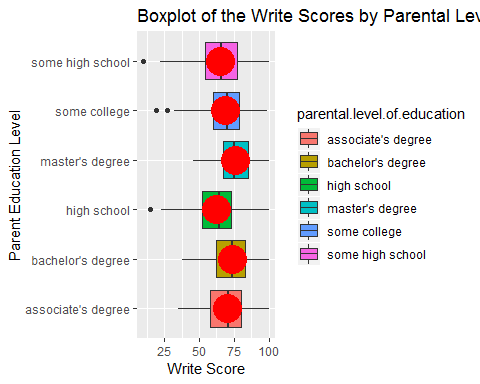


The students who completed the test preparation score have performed far better compared to the students who did not.

1. Boxplot for Writing Score Vs Parental Level of Education

group\_by(sp,parental.level.of.education) %>%  
 ggplot(aes(x = factor(parental.level.of.education), y = writing.score, fill = parental.level.of.education)) +   
 geom\_boxplot(colours = rev(brewer.pal(10,'Spectral'))) +  
 stat\_summary(fun.y=mean, geom="point", shape=20, size=15, color="red", fill="red") +  
 xlab("Parent Education Level") +ylab("Write Score") +coord\_flip() +  
 ggtitle('Boxplot of the Write Scores by Parental Level of Education')

## Warning: Ignoring unknown parameters: colours

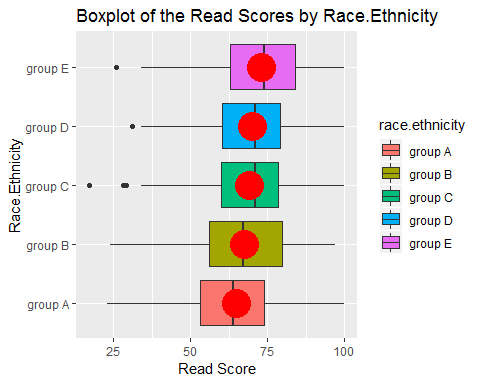


Students whose parents had a Master’s Degree performed the best and the students, whose parents were high school graduates performed the worst. Hence, the parental level of education was found to have an impact on the student’s performance in the writing test.

1. Boxplot for Reading Score Vs Race.Ethnicity

group\_by(sp,race.ethnicity) %>%  
 ggplot(aes(x = factor(race.ethnicity), y = reading.score, fill = race.ethnicity)) +   
 geom\_boxplot(colours = rev(brewer.pal(10,'Spectral'))) +  
 stat\_summary(fun.y=mean, geom="point", shape=20, size=15, color="red", fill="red") +  
 xlab("Race.Ethnicity") +ylab("Read Score") +coord\_flip() +  
 ggtitle('Boxplot of the Read Scores by Race.Ethnicity')

## Warning: Ignoring unknown parameters: colours

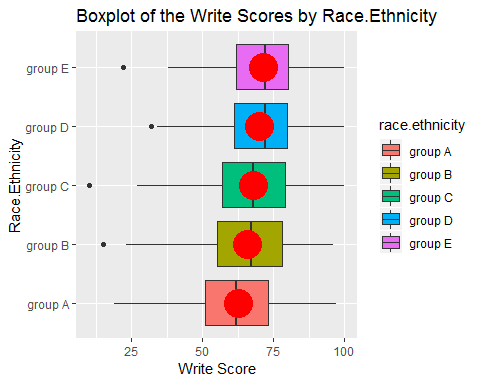


The students belonging to the Group E race have secured the highest range of marks and those belonging to Group A have secured the least range of marks.

1. Boxplot for Writing Score Vs Race.Ethnicity

group\_by(sp,race.ethnicity) %>%  
 ggplot(aes(x = factor(race.ethnicity), y = writing.score, fill = race.ethnicity)) +   
 geom\_boxplot(colours = rev(brewer.pal(10,'Spectral'))) +  
 stat\_summary(fun.y=mean, geom="point", shape=20, size=15, color="red", fill="red") +  
 xlab("Race.Ethnicity") +ylab("Write Score") +coord\_flip() +  
 ggtitle('Boxplot of the Write Scores by Race.Ethnicity')

## Warning: Ignoring unknown parameters: colours



The students belonging to the Group E race have secured the highest range of marks and those belonging to Group A have secured the least range of marks.

Conclusions:

1. The male students have a higher mark range for a Math test, compared to the female students. The mean for the male catagory is higher compared to the female.
2. It can very well be concluded that, the students who had standard lunch, have fared well in Math as compared to those who had free/reduced lunch.
3. The students who had completed the test preparation course have done well in Math compared to those who have not.
4. It can be concluded, that the female students have performed well at reading and overall have a higer score compared to the male students. Even the mean for the female catagory is higher compared to the male catagory.
5. The students who had a standard lunch performed well at reading compared to those who had free/reduced lunch.
6. It is clearly visible that the students who completed the test preparation course have performed better at reading compared to those who have not.
7. It can be inferred that the female students have performed better at writing compared to th male students.
8. The students who had a standard lunch fared well at writing compared to those who had free/reduced lunch.
9. The students who completed the test preparation score have performed far better in the writing test, compared to the students who did not.
10. Students whose parents had a Master’s Degree performed the best and the students, whose parents were high school graduates performed the worst in writing test. Hence, the parental level of education was found to have an impact on the student’s performance in the writing test.
11. The students belonging to the Group E race have secured the highest range of marks and those belonging to Group A have secured the least range of marks in the reading test.
12. The students belonging to the Group E race have secured the highest range of marks and those belonging to Group A have secured the least range of marks in a writing test.

Deductions:

1. Math Test : According to the observations, a Male student who has a standard lunch and completes the test preparation course has the highest chances of getting a good Math score.
2. Reading Score: A female student, who has standard lunch, completes the test course preparation and belongs to Group E ethnicity has the best possibility to get a good reading score.
3. Writing Score: A female student, who has a standard lunch, completes the test course preparation, belonging to Group E Ethnicity and having Parents with Master’s degree, has the highest possibility of getting a good writing score.