Analysis results

Taha Nasir

2022-07-29

# Loading libraries

In the first step, the libraries needed for the analysis are installed. If they’re pre-installed, then they’re loaded in RStudio using the library() function. require() function has also been used. Rmarkdown has been used to knit all the data into the word file.The following libraries have been loaded:

library(tidyr) library(ggplot2) library(dplyr) library(fBasics) library(kableExtra) library(funModeling) library(plotly) library(MASS) library(corrplot)

# Data import

The pre-processed data in excel has been imported into Rstudio using read.csv function.Column names have been defined. The code and dataset is as follows:

performance <- read.csv("StudentsPerformance.csv")  
  
names\_columns <- c("gender","race","parent\_education","lunch","test\_prep","math\_score","reading\_score","writing\_score")  
colnames(performance) <- names\_columns  
head(performance)

## gender race parent\_education lunch test\_prep math\_score  
## 1 female group B bachelor's degree standard none 72  
## 2 female group C some college standard completed 69  
## 3 female group B master's degree standard none 90  
## 4 male group A associate's degree free/reduced none 47  
## 5 male group C some college standard none 76  
## 6 female group B associate's degree standard none 71  
## reading\_score writing\_score  
## 1 72 74  
## 2 90 88  
## 3 95 93  
## 4 57 44  
## 5 78 75  
## 6 83 78

# Data manipulation

dplyr library has been used to manipulate the data. We have converted the variables with type=character to type=factor. Then the str() function has been used to analyse the structure of the performance data frame.

performance$gender<-factor(performance$gender)  
performance$race<-factor(performance$race)  
performance$parent\_education<-factor(performance$parent\_education)  
performance$lunch<-factor(performance$lunch)  
performance$test\_prep<-factor(performance$test\_prep)  
str(performance)

## '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 : Factor w/ 5 levels "group A","group B",..: 2 3 2 1 3 2 2 2 4 2 ...  
## $ parent\_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\_prep : 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 ...

# Descriptive statistics

The next step in the analysis is descriptive statistics. Summary statistics has been used to display the no.of various levels of the categorical variables and the mean, median, mode,min, max, 1st quartile and 3rd quartile of the numerical variables. IQR function will be used to determine the inter-quartile range of the numerical data. The use of summary() function and its output has been shown below:

summary(performance)

## gender race parent\_education lunch   
## female:518 group A: 89 associate's degree:222 free/reduced:355   
## male :482 group B:190 bachelor's degree :118 standard :645   
## group C:319 high school :196   
## group D:262 master's degree : 59   
## group E:140 some college :226   
## some high school :179   
## test\_prep math\_score reading\_score writing\_score   
## completed:358 Min. : 0.00 Min. : 17.00 Min. : 10.00   
## none :642 1st Qu.: 57.00 1st Qu.: 59.00 1st Qu.: 57.75   
## Median : 66.00 Median : 70.00 Median : 69.00   
## Mean : 66.09 Mean : 69.17 Mean : 68.05   
## 3rd Qu.: 77.00 3rd Qu.: 79.00 3rd Qu.: 79.00   
## Max. :100.00 Max. :100.00 Max. :100.00

Then, profiling\_num() function has been used to show the detailed descriptive statistics of the numerical variables.

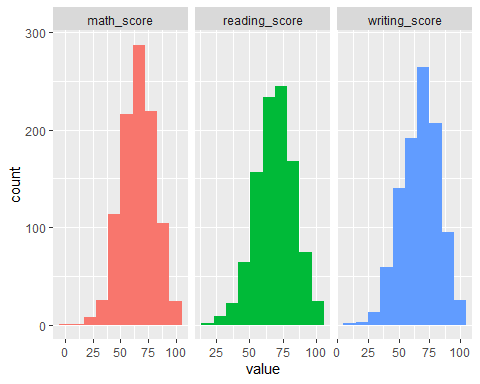
profiling\_num(performance)

## variable mean std\_dev variation\_coef p\_01 p\_05 p\_25 p\_50 p\_75  
## 1 math\_score 66.089 15.16308 0.2294342 27.99 40.95 57.00 66 77  
## 2 reading\_score 69.169 14.60019 0.2110800 31.99 44.00 59.00 70 79  
## 3 writing\_score 68.054 15.19566 0.2232882 31.98 42.95 57.75 69 79  
## p\_95 p\_99 skewness kurtosis iqr range\_98 range\_80  
## 1 90.05 98.01 -0.2785166 3.267597 20.00 [27.99, 98.01] [47, 86]  
## 2 92.00 100.00 -0.2587157 2.926081 20.00 [31.99, 100] [51, 87.1]  
## 3 92.00 100.00 -0.2890096 2.960808 21.25 [31.98, 100] [48, 87]

We can see that Mathematics exam score average of students is 66.08. Reading exam score average of students is 69.16. Writing exam score average of students is 68.05. # Data Visualization For data visualization, three side-by-side histograms of the numerical variables will be plotted using the plot\_num() function as follows:

plot\_num(performance)

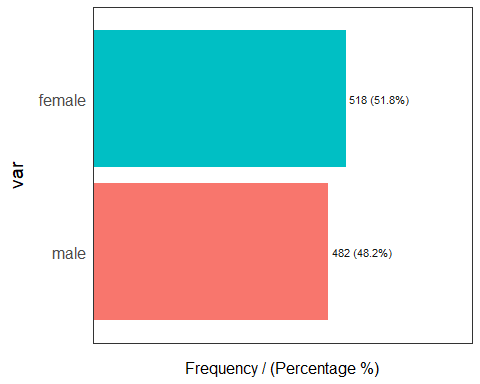
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =  
## "none")` instead.



The freq() function has been used to plot a table and a bar graph showing the frequency, percentage and cumulative frequency of the gender variable.

freq(performance$gender)

## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =  
## "none")` instead.

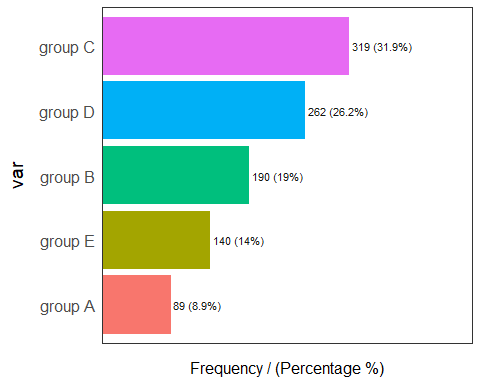


## var frequency percentage cumulative\_perc  
## 1 female 518 51.8 51.8  
## 2 male 482 48.2 100.0

The bar graph shows that the percentage of female students is higher that the male students in the dataset.Since the dataset contains the data of all the students of the public school in US. Hence, it can be concluded that the number of female students is higher than the number of male students in the school. Similarly freq() function has also been run on the race variable as follows:

freq(performance$race)

## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =  
## "none")` instead.

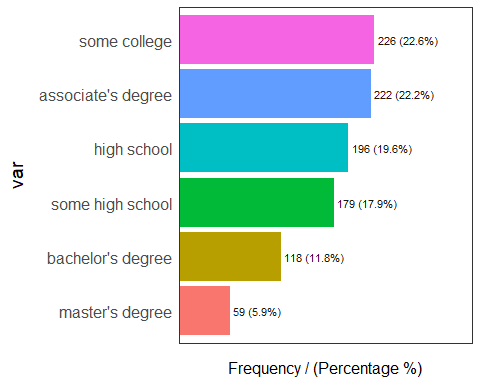


## var frequency percentage cumulative\_perc  
## 1 group C 319 31.9 31.9  
## 2 group D 262 26.2 58.1  
## 3 group B 190 19.0 77.1  
## 4 group E 140 14.0 91.1  
## 5 group A 89 8.9 100.0

From the bar graph of race vs frequency/percentage, we can conclude that the most common race ethnicity is Group C, while the least common is Group A. The parent education level variable of the students consists of 6 categories.

freq(performance$parent\_education)

## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =  
## "none")` instead.

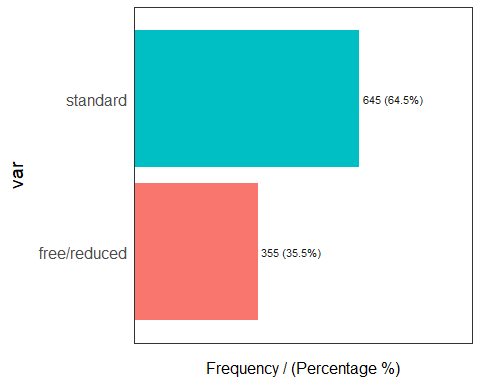


## var frequency percentage cumulative\_perc  
## 1 some college 226 22.6 22.6  
## 2 associate's degree 222 22.2 44.8  
## 3 high school 196 19.6 64.4  
## 4 some high school 179 17.9 82.3  
## 5 bachelor's degree 118 11.8 94.1  
## 6 master's degree 59 5.9 100.0

The most common education in the parent\_education variable is ‘some college’ while the least common is ‘master’s degree’. The detailed statistics and bar plot of the lunch variable is as follows:

freq(performance$lunch)

## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =  
## "none")` instead.

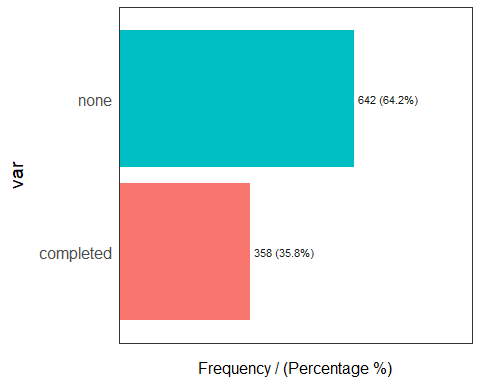


## var frequency percentage cumulative\_perc  
## 1 standard 645 64.5 64.5  
## 2 free/reduced 355 35.5 100.0

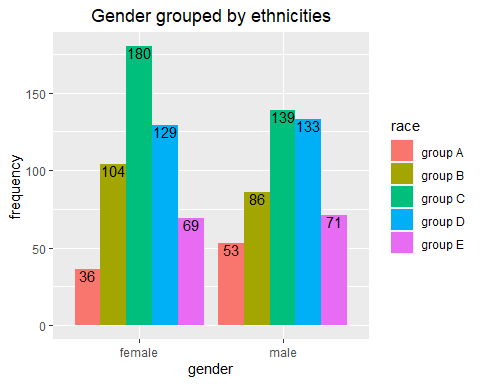
64.5% of the students eat lunch at standard price and 35.5% eat free/reduced lunch.

freq(performance$test\_prep)

## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =  
## "none")` instead.



## var frequency percentage cumulative\_perc  
## 1 none 642 64.2 64.2  
## 2 completed 358 35.8 100.0

From the table and bar plot, we can conclude that while 642 of the students did not complete the exam preparation course, 358 student did. In the next step, the distribution of gender grouped by race will be visualized using a bar plot.The code has been hidden by using echo=FALSE argument.Only the plot will be shown.Both the geom\_bar() and geom\_text()functions have been used to generate the plot.

# t-test

## 1st Research Question The 1st research question for the t-test is as follows: ### Are there any differences between the exam scores of female and male students? To answer this question, three null and main hypothesis have been formulated. The first null hypothesis for this research question is: **H0** : There is no difference between the mathematics scores of male and female students. The first main hypothesis for this research question is: **H1** : There is a difference between the mathematics scores of male and female students. Now, the t-test has been conducted after the formulation of the hypothesis.

attach(performance)  
t.test(math\_score~gender, var.equal=TRUE)

##   
## Two Sample t-test  
##   
## data: math\_score by gender  
## t = -5.3832, df = 998, p-value = 9.12e-08  
## alternative hypothesis: true difference in means between group female and group male is not equal to 0  
## 95 percent confidence interval:  
## -6.952285 -3.237737  
## sample estimates:  
## mean in group female mean in group male   
## 63.63320 68.72822

***t=-5.398*** and ***p value < 0.05***

Hence, H0 is rejected. There is a difference between gender and math test score. The second null hypothesis for this research question is: **H0**: There is no difference between the reading scores of male and female students. The second main hypothesis for this research question is: **H1**: There is a difference between the reading scores of male and female students. The t-test has been conducted after the formulation of the second hypothesis.

t.test(reading\_score~gender, var.equal=TRUE, data=performance)

##   
## Two Sample t-test  
##   
## data: reading\_score by gender  
## t = 7.9593, df = 998, p-value = 4.681e-15  
## alternative hypothesis: true difference in means between group female and group male is not equal to 0  
## 95 percent confidence interval:  
## 5.375946 8.894212  
## sample estimates:  
## mean in group female mean in group male   
## 72.60811 65.47303

***t=7.9593*** and ***p value <0.05*** H0 is rejected. There is a difference between gender and reading test score. The third null and main hypothesis are as follows: **H0**: There is no difference between the writing scores of male and female students. **H1**: There is a difference between the writing scores of male and female students. The t-test has been conducted after the formulation of the third hypothesis as well.

t.test(writing\_score~gender,var.equal=TRUE )

##   
## Two Sample t-test  
##   
## data: writing\_score by gender  
## t = 9.9796, df = 998, p-value < 2.2e-16  
## alternative hypothesis: true difference in means between group female and group male is not equal to 0  
## 95 percent confidence interval:  
## 7.35558 10.95638  
## sample estimates:  
## mean in group female mean in group male   
## 72.46718 63.31120

***t=9.9796*** and ***p value <0.05*** H0 is rejected. There is a difference between gender and writing test score.

# Chi-squared test:

The H0 and H1 are as follows:

**H0**: Completing the test preparation course is not related to gender.

**H1**: Completing the test preparation course is related to gender.

chi\_tab1<- table(gender, test\_prep)  
  
chisq.test(chi\_tab1)

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: chi\_tab1  
## X-squared = 0.015529, df = 1, p-value = 0.9008

As p value =0.9>0.05, hence H0 is accepted. This means that gender is not all related to The H0 and H1 for the next chi-squared test are as follows: **H0**: Lunch eating status is not related to gender. **H1**: Lunch eating status is related to gender.

chi\_tab2<- table(gender, lunch)  
  
chisq.test(chi\_tab2)

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: chi\_tab2  
## X-squared = 0.37174, df = 1, p-value = 0.5421

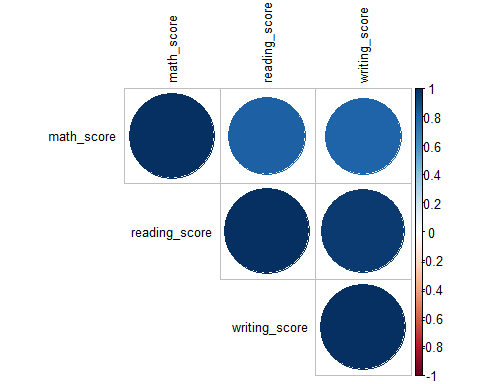
As p value=0.5421>0.05, H0 is accepted.Thus means that lunch eating status is not related to gender.

# Correlation

There is a strong positive correlation between the variables as shown by the correlation matrix.

num\_data <- performance[, c( "math\_score", "reading\_score", "writing\_score")]  
m<-cor(num\_data)  
m

## math\_score reading\_score writing\_score  
## math\_score 1.0000000 0.8175797 0.8026420  
## reading\_score 0.8175797 1.0000000 0.9545981  
## writing\_score 0.8026420 0.9545981 1.0000000

The correlation plot is shown below: 

# Multiple Linear Regression

First of all, linear regression is used to determine the extent to which the mathematics score can be explained by the gender and reading score.

reg1<- lm(math\_score~ gender+reading\_score, data=performance)  
summary(reg1)

##   
## Call:  
## lm(formula = math\_score ~ gender + reading\_score, data = performance)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -22.7447 -4.3918 -0.0747 4.1590 18.1338   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -5.2230 1.1056 -4.724 2.64e-06 \*\*\*  
## gendermale 11.8614 0.4292 27.634 < 2e-16 \*\*\*  
## reading\_score 0.9483 0.0147 64.523 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.577 on 997 degrees of freedom  
## Multiple R-squared: 0.8122, Adjusted R-squared: 0.8119   
## F-statistic: 2157 on 2 and 997 DF, p-value: < 2.2e-16

***math.score=−5.2230+11.8614gender+0.9483reading***

R-squared = 0.8122, this implies that 81.2% of the mathematics score can be explained by the variables of gender and reading test score. The second regression analysis is used to determine the extent to which the mathematics score can be explained by the variables gender, test\_prep and reading\_score.

reg2<- lm(math\_score~ gender+test\_prep+reading\_score, data=performance)  
summary(reg2)

##   
## Call:  
## lm(formula = math\_score ~ gender + test\_prep + reading\_score,   
## data = performance)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -23.1746 -4.2620 -0.1302 4.1333 17.5549   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -7.1570 1.2351 -5.795 9.19e-09 \*\*\*  
## gendermale 11.9632 0.4279 27.955 < 2e-16 \*\*\*  
## test\_prepnone 1.5327 0.4457 3.438 0.000609 \*\*\*  
## reading\_score 0.9614 0.0151 63.661 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.541 on 996 degrees of freedom  
## Multiple R-squared: 0.8144, Adjusted R-squared: 0.8139   
## F-statistic: 1457 on 3 and 996 DF, p-value: < 2.2e-16

***math.score=−7.1570+11.9632gender+1.5327course+0.9614reading***

R-squared = 0.8144,this implies that 81.4% of the mathematics score can be explained by the variables of gender,reading test scores and test preparation course. The third regression analysis is run to determine the extent to which gender and test\_prep variables are used to explain the writing score.

reg3<-lm(writing\_score~ gender+test\_prep, data=performance)  
summary(reg3)

##   
## Call:  
## lm(formula = writing\_score ~ gender + test\_prep, data = performance)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -58.925 -8.711 0.317 9.296 32.289   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 78.8971 0.8372 94.23 <2e-16 \*\*\*  
## gendermale -9.2137 0.8665 -10.63 <2e-16 \*\*\*  
## test\_prepnone -9.9722 0.9031 -11.04 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 13.69 on 997 degrees of freedom  
## Multiple R-squared: 0.1898, Adjusted R-squared: 0.1882   
## F-statistic: 116.8 on 2 and 997 DF, p-value: < 2.2e-16

***writing.score=78.8971−9.2137gender−9.9722course*** R-squared = 0.1898, this implies that 19% of the writing score can be explained by the variables of gender and test preparation course.