Tipologia i Cicle de Vida de les Dades

PRÀCTICA 2

Narcís Coll Follia

https://github.com/ncoll7/Neteja-analisi-estudiants

1. Descripció del dataset. Perquè és important i quina pregunta/problema pretén respondre?

El dataset amb el que treballarem en aquesta pràctica és un dataset que aporta informació de diferents estudiants d'institut dels Estats Units (*high school*) i les seves notes en diferents matèries.

L'enllaç al dataset és: https://www.kaggle.com/spscientist/students-performance-in-exams

Crec que és interessant perquè ens ajudarà veure com afecten diverses variables pròpies de l'estudiant a les seves notes.

El dataset consta de 8 variables i 1000 files (estudiants). En concret les variables són:

- **gender:** Sexe de l'estudiant
- **race.ethnicity:** La raça del estudiant. Les diferents races estan codificades en A, B, C, D i E i n o se'ns explica a la documentació quina raça és, per tant parlarem en els mateixos termes codificats.
- **parental.level.of.education:** Nivell d'educació dels pares.
- **lunch:** Si l'estudiant percep ajudes en el pagament del dinar. Entenc que aquesta variable ens intenta mostrar la situació econòmica de la família de l'estudiant.
- **test.preparation.course:** Si l'estudiant ha realitzat el curs de preparació dels exàmens.
- **math.score:** La nota de l'estudiant en l'examen de mates.
- **reading.score:** La nota de l'estudiant en l'examen de lectura.
- **writing.score:** La nota de l'estudiant en l'examen de redacció.

Així doncs amb aquest dataset podem veure com influencien diverses variables a la nota d'un estudiant i com es relacionen les diverses notes de les diferents matèries entre elles.

En concret ens interessa saber si fer el curs de preparació realment serveix per treure més bones notes per cada assignatura, com es relacionen les notes de les diferents assignatures entre elles i finalment com afecten globalment les diferents variables del dataset a les notes dels alumnes.

2. Integració i selecció de les dades d'interès a analitzar.

En primer lloc llegim les dades amb l'R i comprovem que la lectura s'ha efectuat de forma correcta:

```
'data.frame':   1000 obs. of  8 variables:
 $ gender                     : chr  "female" "female" "female" "male" ...
 $ race.ethnicity             : chr  "group B" "group C" "group B" "group A" ...
 $ parental.level.of.education: chr  "bachelor's degree" "some college" "master's degree" "associate's degree" ...
 $ lunch                      : chr  "standard" "standard" "standard" "free/reduced" ...
 $ test.preparation.course    : chr  "none" "completed" "none" "none" ...
 $ 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 ...
```

Veiem que sí que s'ha efectuat de forma correcta però ens interessa tenir les variables que l'R reconeix com a caràcters com a factors per fer més fàcil el seu posterior anàlisi, així que transformem aquestes variables en factors:

```
'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 ...
```

La variable parental.level.of.education és un factor amb uns nivells que podem ordenar (de menys nivell de educació a més). Ho ordenem simplement per tenir una presentació més bona a l'hora de representar la variable gràficament:

```
[1] "some high school"   "high school"        "some college"       "associate's degree" "bachelor's degree"
[6] "master's degree"
```

Finalment, en planificació de futurs anàlisi que poden ser interessants, creem una nova variable mean.score que és la mitjana entre les notes de les 3 matèries estudiades.

3. Neteja de les dades.

3.1. Les dades contenen zeros o elements buits? Com gestionaries aquests casos?

Per identificar els elements buits busquem diferents elements al nostre dataset. Per començar busquem si hi ha instàncies codificades explícitament com a valors buits (NA en R) i veiem que no n'hi ha cap. Després busquem altres codificacions que s'utilitzen sovint com a valors buits, en concret "?", "" i " ". No en trobem cap.

Finalment busquem si hi ha missings codificats com a 0. Només en trobem un i està en una variable numèrica on el 0 està dins el domini de la variable.

Per comprovar que no ens deixem cap valor buit, revisem les categories de les variables categòriques que no hi hagi missings codificats de manera diferent, i fem un resum numèric de les variables numèrics que no hi hagin valors fora de domini que ens podrien indicar un missing (per exemple un 9999).

No hem trobat cap missing en tot el dataset. En cas d'haver-ne trobat algun tenim diferents opcions:

Si la variable és categòrica i la quantitat de missings és important, podríem crear una nova categoria amb tots els missings d'aquesta variable, això ens podria evitar perdre la possible informació que ens podrien donar aquests missings i ens permetria estudiar si aquests missings es produeixen per alguna causa concreta.

Si la variable és categòrica però la quantitat de missings és petita podem realitzar una imputació simple d'aquests missings amb la moda de la variable o una imputació més completa utilitzant algun algoritme com el Knn.

Si la variable és contínua podem imputar els missings amb la mitjana de la variable.

3.2. Identificació i tractament de valors extrems.

Utilitzem resums numèrics en les variables numèriques i representacions en boxplot (veure apartat 5) per identificar els valors extrems.

Ja de per si soc bastant conservador a l'hora de tractar amb valors extrems. Normalment els identifico i al llarg dels anàlisis observo com es comporten les instàncies amb aquests valors extrems, però llevat que siguin casos molt obvis, o tingui suficient coneixement de la variable o utilitzi anàlisis que siguin molt sensibles a aquests valors, intento no modificar-los.

En el nostre cas al observar els boxplots veiem que les notes de les 3 matèries tenen algun valor que podríem considerar extrem (més clar en el cas de matemàtiques), però donat la naturalesa de les variables i que coneixem en detall el domini d'aquestes variables (0,100) decidim no tocar-los.

4. Anàlisi de les dades.

4.1. Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar).

Com que tenim un dataset no molt gran, abans de realitzar anàlisis més complexes, realitzo un anàlisi gràfic univariant per totes les variables del dataset. Per les variables categòriques utilitzo diagrames de barres i per les numèriques histogrames i boxplots (tots els gràfics a l'apartat 5).

Seguint amb els objectius de l'apartat 1 els anàlisis seran:

- Comparació de mitjanes de les tres matèries segons si els estudiants han realitzat el curs de preparació o no.
- Estudi de la relació entre les notes de les 3 matèries.
- Estudi global de com afecten totes les variables del dataset a la nota mitjana de l'estudiant.

4.2. Comprovació de la normalitat i homogeneïtat de la variància.

Per comprovar la normalitat de les variables numèriques hem realitzat els tests Kolmogorov-Smirnov i Shapiro-Wilk i en tots els casos han passat el primer però no el segon. He provat de realitzar la transformació BoxCox i tot i que les variables semblaven apropar-se encara més a la normalitat seguien passant el primer test però no el segon. Com que tenim un número elevat d'instàncies (1000) direm que assumim normalitat per el teorema central del límit.

Per les proves que volem fer ens interessa comprovar l'homoscedasticitat de les notes de les 3 matèries per la gent que ha fet el curs de preparació i per la que no. Per fer-ho hem realitzat els tests de Levene i de Flinger. En aquest cas els dos han coincidit per cada matèria de manera que:

Math.score compleix la hipòtesi d'homoscedasticitat:

```
> leveneTest(math.score ~ test.preparation.course, data = bd)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   1   0.533 0.4655
      998
> fligner.test(math.score ~ test.preparation.course, data = bd)

        Fligner-Killeen test of homogeneity of variances

data:  math.score by test.preparation.course
Fligner-Killeen:med chi-squared = 0.41276, df = 1, p-value = 0.5206
```

Reading.score també compleix la hipòtesi d'homoscedasticitat:

```
> leveneTest(reading.score ~ test.preparation.course, data = bd)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   1  1.0798  0.299
      998
> fligner.test(reading.score ~ test.preparation.course, data = bd)

        Fligner-Killeen test of homogeneity of variances

data:  reading.score by test.preparation.course
Fligner-Killeen:med chi-squared = 1.02, df = 1, p-value = 0.3125
```

Writing.score no compleix la hipòtesi d'homoscedasticitat:

```
> leveneTest(writing.score ~ test.preparation.course, data = bd)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value  Pr(>F)
group   1  5.9708 0.01472 *
      998
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> fligner.test(writing.score ~ test.preparation.course, data = bd)

        Fligner-Killeen test of homogeneity of variances

data:  writing.score by test.preparation.course
Fligner-Killeen:med chi-squared = 5.7598, df = 1, p-value = 0.0164
```

4.3. Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l'objectiu de l'estudi, aplicar proves de contrast d'hipòtesis, correlacions, regressions, etc. Aplicar almenys tres mètodes d'anàlisi diferents.

Estudi 1:

Per aquest estudi volem saber si el fet de realitzar el curs de preparació ajuda a treure una nota més bona a l'examen per les diferents matèries.

Per fer-ho primer representarem un histograma per cada matèria dividint entre els estudiants que han fet el curs de preparació i els que no.

Sabem que la mitjana dels que han fet el curs és més gran que la dels que no. La pregunta és si aquesta diferència és estadísticament significativa o no. Per fer-ho farem un contrast d'hipòtesis de comparació de mitjanes.

Per les assignatures de matemàtiques i lectura realitzarem un test T ja que compleixen la hipòtesi d'homoscedasticitat mentre que per l'assignatura de redacció realitzarem un test de Mann-Whitney ja que no es compleix la hipòtesi d'homoscedasticitat.

En els tres tests les hipòtesis seran:

Hipòtesi nul·la: les notes de l'assignatura dels estudiants que han fet el curs de preparació són iguals o més petites dels que no l'han fet.

Hipòtesi alternativa: les notes de l'assignatura dels estudiants que han fet el curs de preparació són més grans dels que no l'han fet.

Els resultats i les conclusions de l'estudi es troben en els apartats 5 i 6 respectivament.

Estudi 2:

L'objectiu d'aquest estudi és veure la correlació que hi ha entre les notes de les diferents matèries, veure si els estudiants que treuen bones notes solen ser els mateixos per les diverses matèries.

Per fer-ho dibuixarem els gràfics de dispersió per parelles per les tres assignatures i després calcularem la matriu de correlacions de Pearson. Finalment realitzarem un test de correlació per veure si aquestes correlacions són significativament més grans que 0.

Els resultats i les conclusions de l'estudi es troben en els apartats 5 i 6 respectivament.

Estudi 3:

L'objectiu d'aquest estudi és veure com afecten les variables independents a la variable creada mean.score.

Per fer-ho crearem un model de regressió on mean.score serà la variable dependent i utilitzarem de variables explicatives totes les altres (menys les 3 notes de les assignatures ja que la variable dependent és una combinació lineal d'aquestes 3).

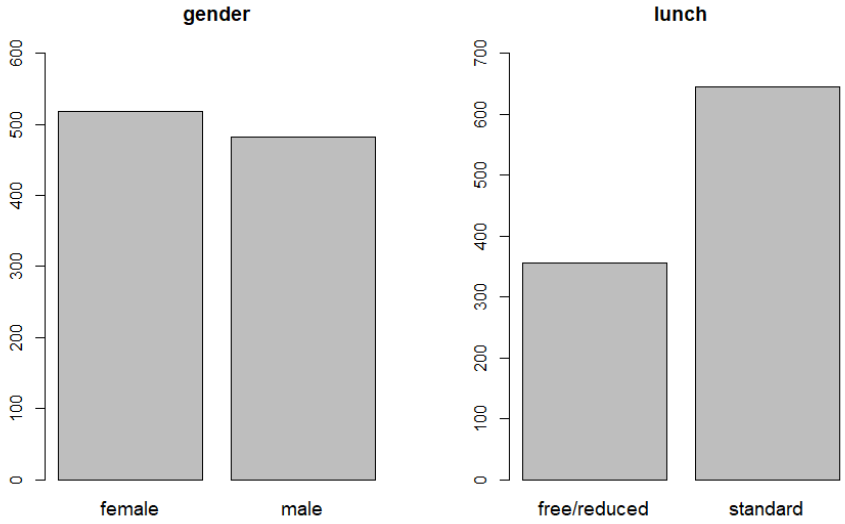
Un cop creat el model, avaluarem el compliment de les hipòtesis de normalitat i homoscedasticitat dels residus.

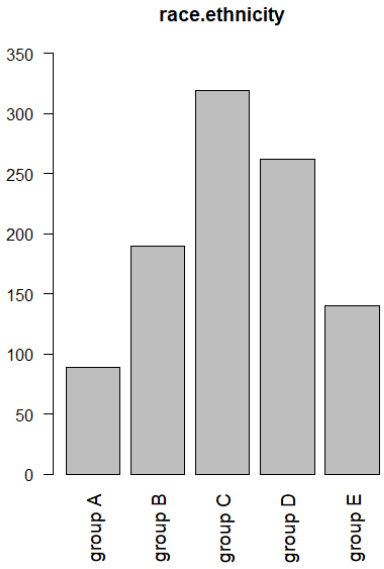
Després mirarem quins coeficients de les variables són significatius, què ens diuen aquests coeficients i avaluarem si és possible crear un model que expliqui bé la variable dependent.

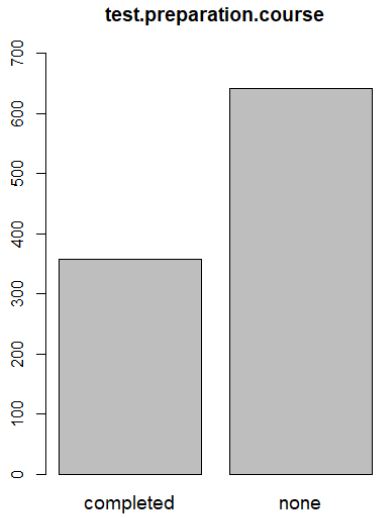
Els resultats i les conclusions de l'estudi es troben en els apartats 5 i 6 respectivament.

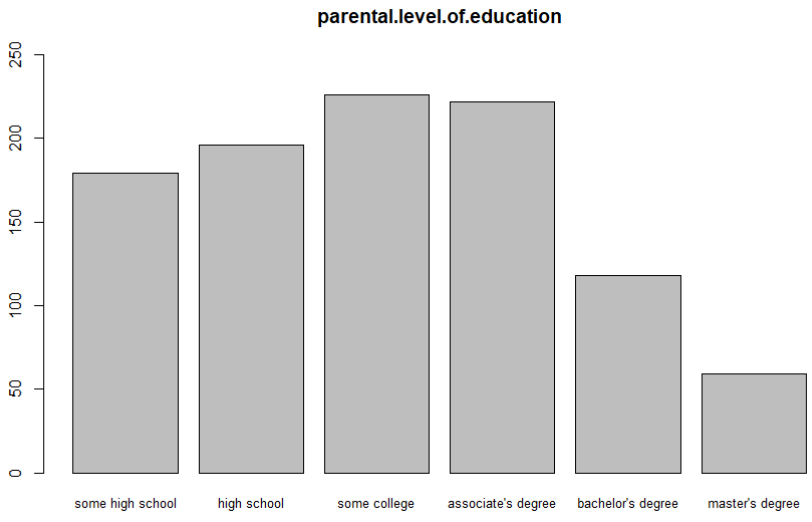
5. Representació dels resultats a partir de taules i gràfiques.

Anàlisi univariant variables categòriques:

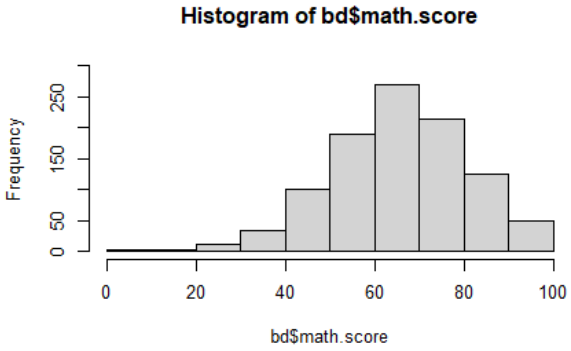


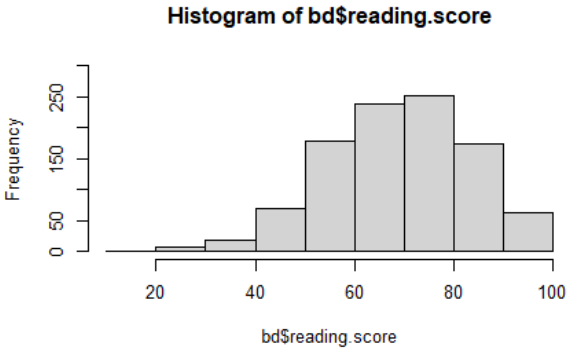


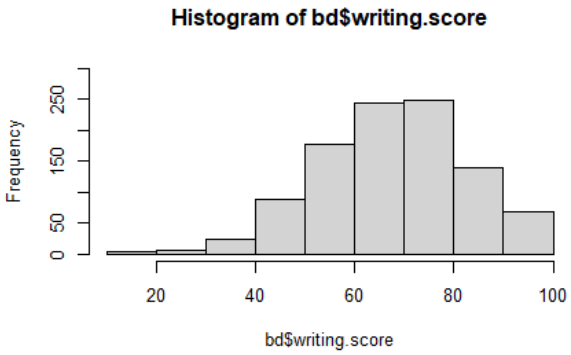


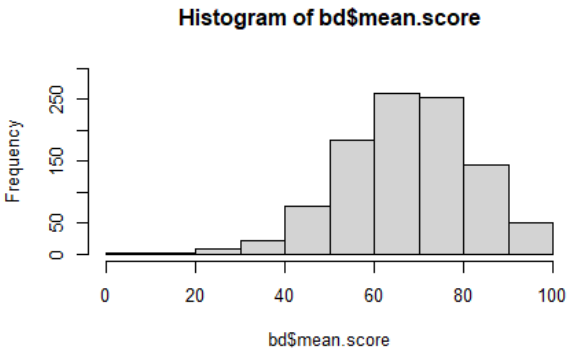


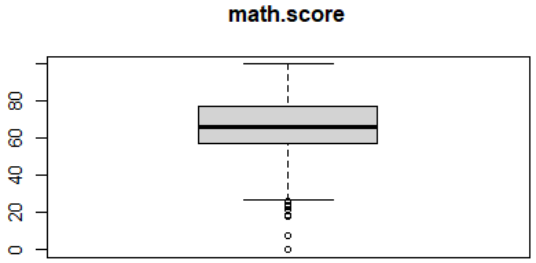
Anàlisi univariant variables numèriques:

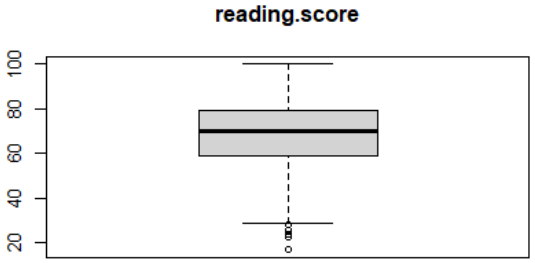


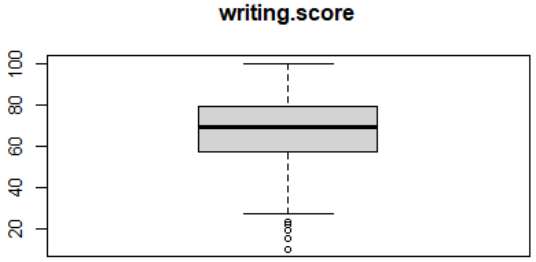


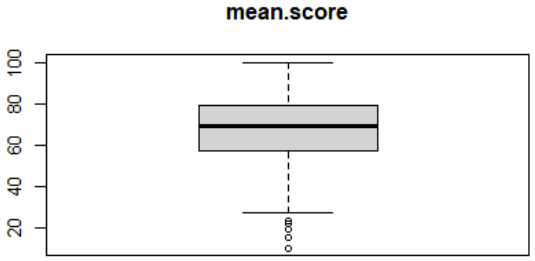






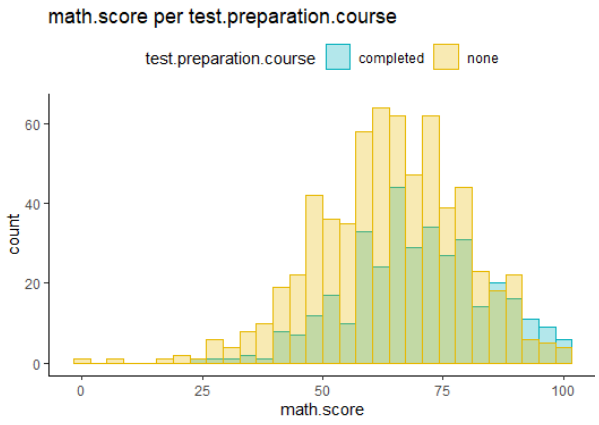


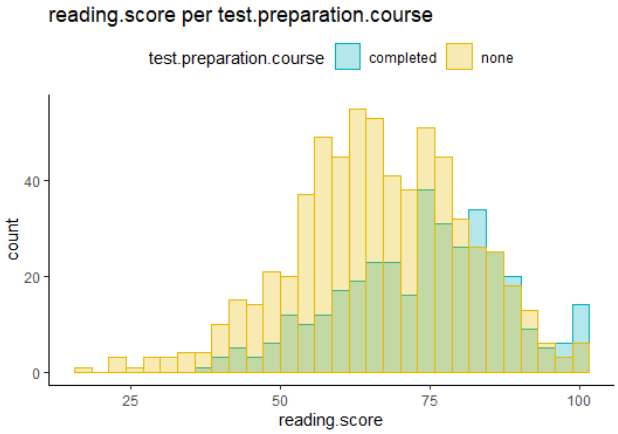


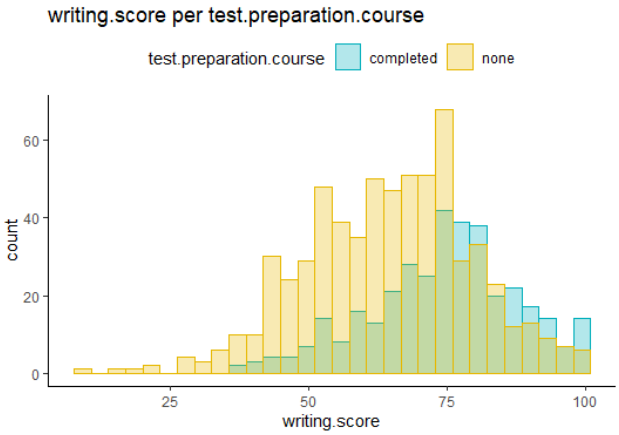


Estudi 1:

Histogrames notes assignatures per curs de preparació:







Test T comparació de mitjanes notes mates per curs de preparació:

```
        Two Sample t-test

data:  x and y
t = 5.7046, df = 998, p-value = 7.68e-09
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 3.996366      Inf
sample estimates:
mean of x mean of y
 69.69553  64.07788
```

Test T comparació de mitjanes notes lectura per curs de preparació:

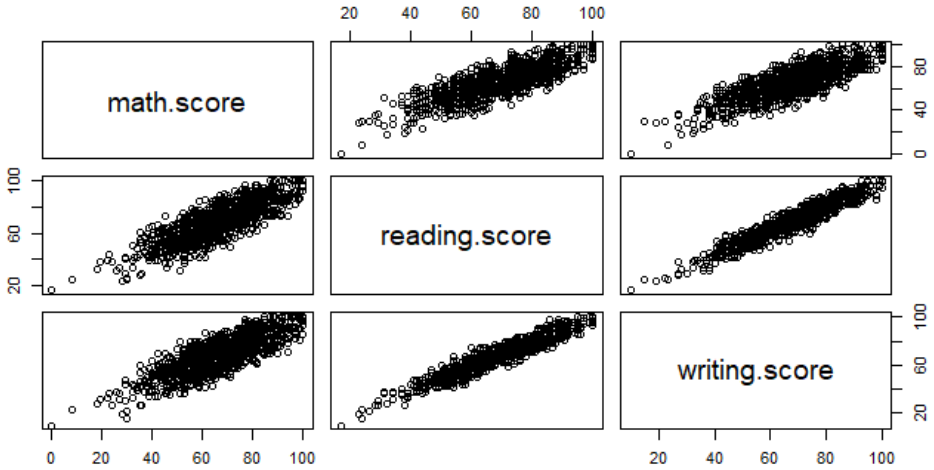
```
        Two Sample t-test

data:  x and y
t = 7.8717, df = 998, p-value = 4.541e-15
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 5.820307      Inf
sample estimates:
mean of x mean of y
 73.89385  66.53427
```

Test Mann-Whitney comparació de mitjanes notes redacció per curs de preparació:

```
        Wilcoxon rank sum test with continuity correction

data:  x and y
W = 158809, p-value < 2.2e-16
alternative hypothesis: true location shift is greater than 0
```

Estudi 2:

Gràfics de dispersió per les notes de les tres assignatures:



Taula de correlacions:

	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

Tests correlacions:

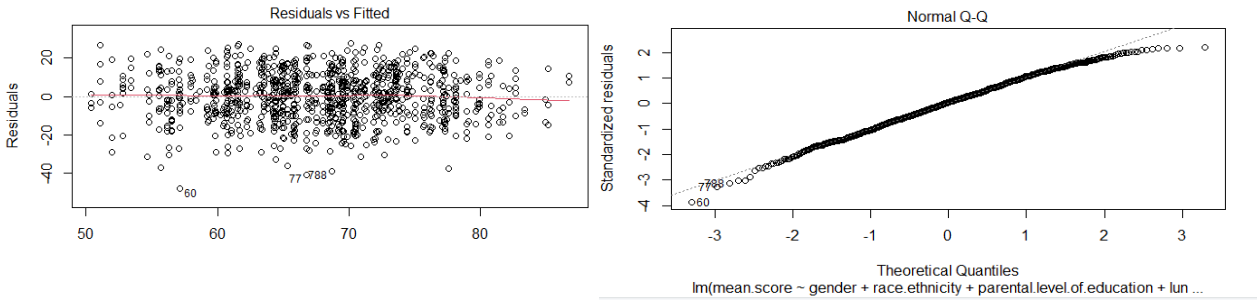
```
        Pearson's product-moment correlation

data:  bd$math.score and bd$reading.score
t = 44.855, df = 998, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7959276 0.8371428
sample estimates:
      cor
0.8175797
```

```
        Pearson's product-moment correlation

data:  bd$math.score and bd$writing.score
t = 42.511, df = 998, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7794321 0.8236517
sample estimates:
     cor
0.802642
```

```
        Pearson's product-moment correlation

data:  bd$writing.score and bd$reading.score
t = 101.23, df = 998, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9487506 0.9597921
sample estimates:
      cor
0.9545981
```

Estudi 3:

Propietats model:



Resum del model:

```
Call:
lm(formula = mean.score ~ gender + race.ethnicity + parental.level.of.education +
    lunch + test.preparation.course, data = bd)

Residuals:
    Min      1Q  Median      3Q     Max
-48.148  -8.298   0.646   8.736  27.522

Coefficients:
                                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                                 62.4008     1.7394  35.875  < 2e-16 ***
gendermale                                  -3.7242     0.7955  -4.682 3.24e-06 ***
race.ethnicitygroup B                        1.5290     1.6116   0.949 0.342983
race.ethnicitygroup C                        2.3855     1.5093   1.581 0.114296
race.ethnicitygroup D                        5.1258     1.5398   3.329 0.000904 ***
race.ethnicitygroup E                        6.9285     1.7081   4.056 5.38e-05 ***
parental.level.of.educationhigh school      -0.6325     1.3002  -0.486 0.626726
parental.level.of.educationsome college      3.6124     1.2563   2.875 0.004121 **
parental.level.of.educationassociate's degree  4.5400     1.2639   3.592 0.000344 ***
parental.level.of.educationbachelor's degree   7.0756     1.4848   4.765 2.17e-06 ***
parental.level.of.educationmaster's degree     8.6322     1.8871   4.574 5.38e-06 ***
lunchstandard                                8.7751     0.8275  10.605  < 2e-16 ***
test.preparation.coursenone                 -7.6386     0.8302  -9.201  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.49 on 987 degrees of freedom
Multiple R-squared:  0.2423,    Adjusted R-squared:  0.2331
F-statistic:  26.3 on 12 and 987 DF,  p-value: < 2.2e-16
```

6. Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?

Estudi 1:

Mirant només els histogrames es pot intuir una diferència entre les mitjanes dels grups dels estudiants que ha fet el curs de preparació i els que no, però no es veu de forma clara.

Quan realitzem els tres testos de comparació de mitjanes però, en els tres casos rebutgem la hipòtesi nul·la que la mitjana de les notes dels estudiants que han realitzat el curs de preparació sigui igual o més petita que la mitjana de les notes dels estudiants que no han realitzat el curs de preparació.

Per tant podem dir que de mitjana, els alumnes que han realitzat el curs de preparació han tret notes significativament millors que els alumnes no les han realitzat i, per tant, ens dona indicis per pensar que els cursos de preparació són útils per obtenir més bones notes.

Estudi 2:

Mirant els gràfics de dispersió ja podem veure com sembla que hi ha una correlació positiva entre les notes dels estudiants per les diferents assignatures i ja veiem que la més forta és entre les notes de lectura i redacció, cosa que és lògica per la naturalesa de les assignatures ja que és d'esperar que a una persona que se li doni bé la lectura també se li doni ve escriure.

Això ho confirmem amb la matriu de correlacions. Veiem que el coeficient de correlació de Pearson entre les notes de lectura i redacció és del 0.95, entre mates i lectura del 0.82 i entre mates i redacció del 0.80.

Finalment hem fet un test de correlacions per confirmar que les correlacions són significativament més grans que zero i efectivament així ens ho han indicat (cosa que podíem esperar pels elevats valors dels coeficients de correlació).

Així doncs en aquest estudi hem confirmat que hi ha una forta relació lineal entre les notes de les diferents assignatures tot i que evidentment no és una relació causa-efecte (no ho sabem però per lògica treure bones notes d'una assignatura no et farà treure bones notes en una altra). Ens indica que un alumne que ha tret bones notes en una assignatura és bastant probable que també hagi tret bones notes a les altres assignatures.

Estudi 3:

En aquest estudi hem vist com afecten les altres variables d'un estudiant a la seva nota mitjana.

Primer hem comprovat que es complien les hipòtesis de normalitat i homoscedasticitat dels residus i després hem observat el resum del model.

El primer que veiem és que totes les variables explicatives (totes categòriques) tenen almenys una categoria significativa.

El model ens diu que:

- S’esperen 3.7 punts menys pel fet de ser home respecte a les dones (amb els altres criteris constants).
- No s’esperen diferències significatives per les races de grup B i C respecte la del grup A però si que s’espera que els del grup D treguin 5.1 punts més que els del grup A i que els del grup E treguin 6.9 punts més que els del grup A.
- En quan a l’educació dels pares, s’espera que les notes de l’estudiant siguin cada cop més elevades segons el nivell d’educació dels pares respecte a l’estudiant que els seus pares no tenien la secundària, menys en el cas que els pares de l’estudiant tingui només la secundària acabada que no s’aprecien diferències significatives.
- Pels estudiants que no paguen la quota del dinar sencera s’espera que tingui 8.8 punts més que el que no.
- Els estudiants que no han fet el curs de preparació s’espera que tinguin 7.6 punts menys que els que sí que l’han fet.

Per tant el perfil d’estudiant que s’espera que tingui una nota mitjana més elevada és una noia, de raça ètnica E, que els seus pares tinguin estudis de màster, que pagui la quota estàndard de dinar i que hagi realitzat el curs de preparació.

De tota manera cal apuntar que el model té un R-quadrat de 0.24, per tant només el 24% de la variància total de la mitjana de les notes de l’estudiant està explicada per aquest model.

Taula de contribucions

Contribucions	Signa
Recerca prèvia	NCF
Redacció de les respostes	NCF
Desenvolupament codi	NCF