PRÁCTICA 2 - TIPOLOGIA DE DADES

Memòria de la pràctica 2 de l'assignatura de Tipologia de Dades realitzada per Maria Sopena i Joana Llauradó. El codi usat per la pràctica es pot robar en el següent repositori GitHub:

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

   En aquesta pràctica utilitzarem pràctica utilitzarem el dataset lliure *Indicators of Heart Disease* de Keggel (https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease) . Aquest conjunt de dades es va obtenir a partor de l'enquesta anual del 2020 dels Centres per al Control i Prevenció de Malalties o, en anglès, Centers for Disease Control and Prevention (CDC). Aquestes centres i les dades que recullen son molt importants ja que permeten detectar factors de riscos i, per tant, prevenir malalties així com estudiar si hi ha algun brot d'una malaltria en concret. Això permet millorar la prevenció i detecció i, per tant, d'aturar, aquestes malalties.

   En aquesta pràctica realitzarem els següents estudis:
   - Estudi de Prevalencia: Estudiar la prevalencia de malalties (Cardiovascular, Infart, Diabetes, Athma, Malalria Renal i Cancer de Pell) estratificada per sexe,raça i edat.
   - Estudi d'Associacions: Segons les prevalencies anteriors escollirem esudiar de les assocaions més remarcables (asthma-edat, malalties cardiovasculars-edat, cancer pell-raça)
   - Estudi de Correlacions: Correlacio de les hores de son (factor de risc de malalties cardiovasculars) i edat. Correlació de l'activtiat física amb les hores de son.

   El dataset amb el que treballem conté dades de 319,795 pacients, un nombre molt elevat que ens ens permetrà fer anàlisis estadísticament molt robustos. Per altra banda, tenim 18 columnes que contenen informació sobre aquests pacients:

Nom variable	*Descripció*	*Tipus*	*Valors*
HeartDisease	*Presència de malaltia cardiovascular*	*categorica, chr*	*Yes /No*
BMI	*Index de massa corporal*	*continua, numerica*	*ex:16.6*

Smoking	*Persona fumadora*	*categorica, chr*	*Yes /No*
AlcoholDrinking	*Persona que ingereix alcohol*	*categorica, chr*	*Yes /No*
Stroke	*Pacient que ha patit un atac de cor*	*categorica, chr*	*Yes /No*
PhysicalHealth	*Índex que indica l'estat de salut física*	*numerica*	*0-30*
MentalHealth	*Índex que indica l'estat de salut física*	*numerica*	*0-30*
DiffWalking	*Dificultat en caminar*	*categorica, chr*	*Yes /No*
Sex	*Sexe del pacient*	*categorica, chr*	*Female/Male*
AgeCategory	*Categoria d'edat*	*Factor 13 levels*	*ex: 18-24, 25-29*
Race	*Ascencència*	*chr*	*ex: White*
Diabetic	*Pacient diabètic*	*categorica, chr*	*Yes /No*
PhysicalActivity	*Activitat físca*	*categorica, chr*	*Yes /No*
GenHealth	*estat de salut general*	*categorica, chr*	*ex: Very good", "good"*
SleepTime	*Hores dormides*	*numerica*	*ex: 5, 6*
Asthma	*Pacient amb asma*	*categorica, chr*	*Yes /No*
KidneyDisease	*Pacient amb malaltia renal*	*categorica, chr*	*Yes /No*
SkinCancer	*Cancer de pell*	*categorica, chr*	*Yes /No*

2. **Neteja de les dades.** *Pot ser el resultat d'addicionar diferents datasets o una subselecció útil de les dades originals, en base a l'objectiu que es vulgui aconseguir.*

   En aquest cas el conjunt de dades està complet i les seves dimensions sona dequades per poder realitzar anàlisis estadístics pel que no s'han integrats més datasets.

3. **Anàlisi de les dades.**

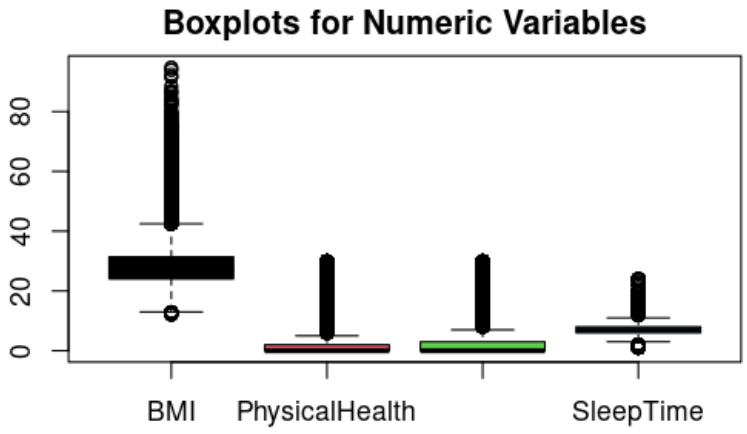
   3.1 Les dades contenen zeros o elements buits? Gestiona cadascun d'aquests casos.

   Ens trobem davant d'un dataset sense missing values, pel que està complet. A nivell de "0" només mirem les variables numèriques ja que en el cas de les categòriques el "0" fan referència a una categoria. Podem veure que de les 4 variables numèriques: BMI, SleepTime, PhsyclaHealth i MentalHEalth aquestes dues ultimes si que tenen 0, ara en aquest cap estem parlant d'index que van del 0 al 30 pel que en aquest cas no cal aplicar cap transformació.

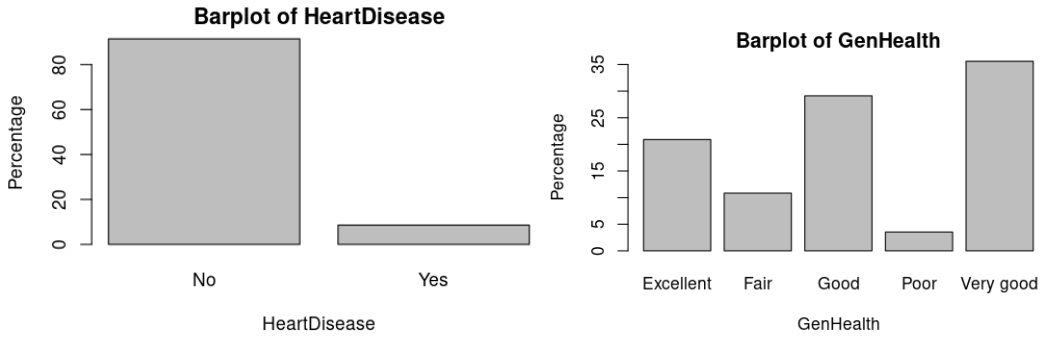
   3.2 Identifica i gestiona els valors extrems.

   Per tal de veure si tenim o no valors extrems mirarem la dsitribució de l es variables.

   En el cas de les numèriques (BMI, PhysicalHealth, MentalHealth, SleepTime) utilitzem els boxplots per veure si hi ha o no valors extrems. Obsrvem que la varibale amb més outliers és el BMI, podem identificar-lis mitjançant el mètode de Tukey (veure .RMD) tot i així no els hi apliquem cap transformació. En el cas de les altres variables tot i que begem outliers més avait ens indiqiuen que una gran part dels pacients tenen una valors elevats en quant a Mental i Phys



En el cas de les variables categòriques fem servir gràifcs de barres per veure'n la distriució. Aquí posem un parell d'exemples, la resta de gràfics es poden obtenir mitjançant el codi.



Finalment apliquem una transformaicó en la variable Diabetic per tal de que les categories siguin simplene "Yes" i "No" en comptes de tota una frase i la variable de GoodHealth la passem a numèrica.

4. Resolució del problema

4.1 Selecció dels grups de dades que es volen analitzar/comparar (p. e., si es volen comparar grups de dades, quins són aquests grups i quins tipus d'anàlisi s'aplicaran?).

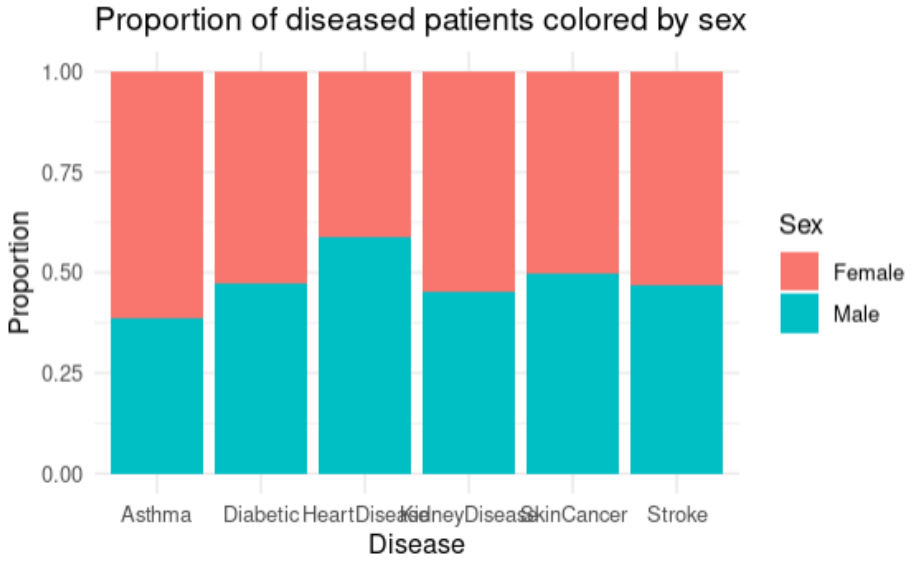
1 - Prevalència de malaltia segons el sexe, edat i raça: Es calcularà la freqüència de cada malaltia segons el sexe dels participants mitjançant 6 variables categòriques (la variable de sexe i cada variable corresponent a les malalties).
2 - Associació de malaltia amb sexe. Com que estem parlant de dos grups de variables categòriques apliquem un Chi Square Test per mirar l'associació.
3 - Prevalència i associació de càncer de pell segons raça. Per mirar la prevàlencia mirem altre cop la freqüència mentre que per estudiar l'assoaició utilitzem un model logisitic (GLM) ja que els dos grups de varibales (raça i sexe) son categòriques.
4 - Associaicó o relaicó d'hores de son segons l'edat. en aqeusts cas estem estudiant una variable numèrica amb una de categ`rocia pel que aplicarem un test d'ANOVA per veure si hi ha o no associació. També es podria aplciar el test de Krusal-Wallis test.
5 - Per estudiar la correlació entre es diversos factors de risc per les malalties cardiosvasculars i la diabetes fem una anàlszi de correlaicó aplican Pearson's ja que estem comparant variables categòriques.

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

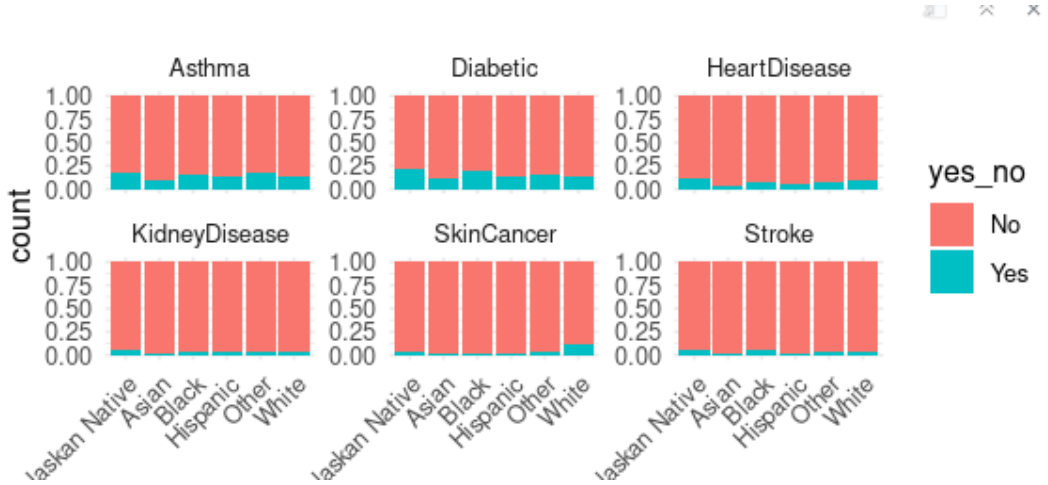
4.3. Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l'objectiu de l'estudi, aplicar proves de contrast d'hipòtesis, correlacions, regressions, etc. Aplicar almenys tres mètodes d'anàlisi diferents

L'aplicació de les diferents proves estadístiques tal i com s'especifica en lapartat 4.1 s'ha realitzat en R. El codi es troba en el fixter .Rmd del gtihub. Aquí només adjuntarem alguns dels resultats i la seva interpretació.

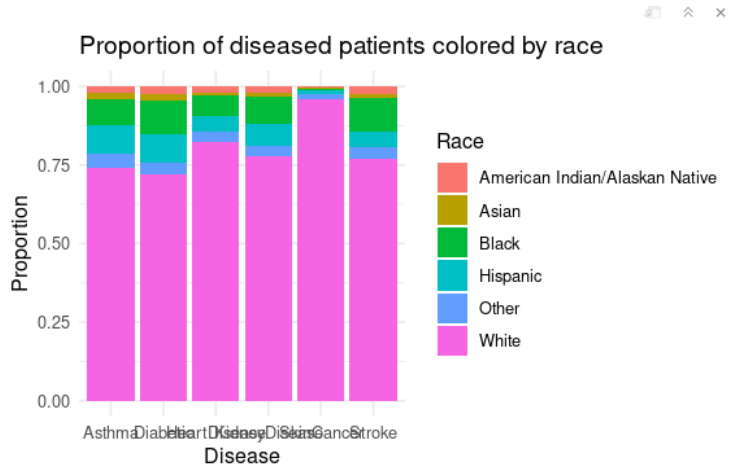
Prevalència de malaltia segons el sexe, edat i raça:



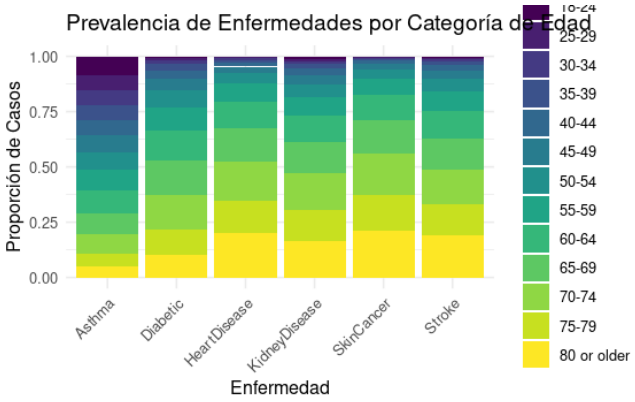
Aquests resultats ens mostren que tot i que la distribució per malaltia entre sexes és sobre el 50% en la majoria dels casos, veiem que sembla que en Asma positiu hi hagi un major percentage de dones i en HeartDisease d'homes.



Podem observar que la majoria de malalties les distribucions son semblant en les diferents ascendencies però que en el cas de càncer de pell si que es veu una diferència segons el tipus d'ascencencia.



Una altra manera de visualitzar-ho seria amb un stacked barplot on podem veure altre cop que en cas de cancer de pell l'ascendencia europea (white) és més prevalent en aquest tipus de cancer.

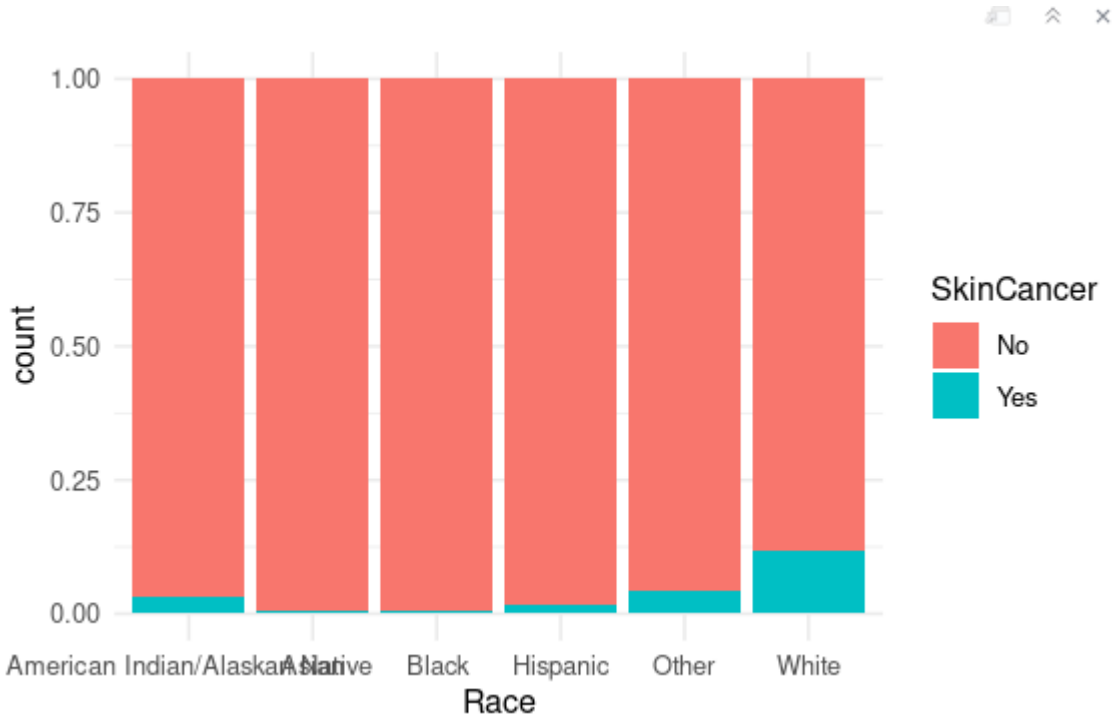


En el següent plot veurem d'entre les diferents categores d'edats quines malalties tenen major prevlaència. Observem doncs que eart disesase i stroke son majoritariament m alalties de població envellida.

Associació de malaltia amb sexe

Després de realitza un Pearsons Chi squarted Test per veure si el sexe està correlacionat amb Asma, Heart Disease i SkinCancer, obtenim un P valor de 2.2E-16, 2.2E-16 i 2.18E-14, respectivament. Per tant podem rebutjant en els tres casos les hipòtesis nul·les i afirmar que sí que hi ha correlació entre el sexe ia questes malalties.

Prevalència i associació de càncer de pell segons raça



Tal com esperàvem, hi ha un percentage més alt de persones amb pell blanc que tenen cancer de pell. En el cas del model obtenim els següents resultats:

```
Call:
glm(formula = SkinCancer ~ Race + Sex + AgeCategory + Smoking,
    family = "binomial", data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8626  -0.5258  -0.2899  -0.1097   4.1085

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -6.61528    0.14588 -45.346  < 2e-16 ***
RaceAsian            -1.34356    0.15648  -8.586  < 2e-16 ***
RaceBlack            -1.82427    0.11671 -15.631  < 2e-16 ***
RaceHispanic         -0.46992    0.09383  -5.008 5.50e-07 ***
RaceOther             0.38537    0.09266   4.159 3.20e-05 ***
RaceWhite             1.08620    0.07957  13.651  < 2e-16 ***
```

Per tant, tal com esperàvem, que Race contribueix significativament al cancer en tots els casos. Si ens fixem en el primer valor (Estimate) podem saber si l'associació és positiva o negativa. Aquest valor fa referència al log odds ratio de l'associació, el simbol ens indica si aquesta associació és positiva o negativa i el número la intensitat d'aquesta. Per tant veiem que *RaceAsian*, *RaceBlack*, *RaceHispanic* tenen una associació negativa, és a dir, en menys risc de càncer sent la raça negra la associació més forta. En canvi, *RaceOthers* i *RaceWhite* tenen una associació positivia suggerint un increment del risc en aquests dos grups racials.

Associació d'hores de son

```
                Df Sum Sq Mean Sq F value Pr(>F)
AgeCategory     12  14310    1192   591.1 <2e-16 ***
Residuals   319782 645142       2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El baix valor de p (<0,001) indica que les diferències observades no són probablement a causa dels atzar. Pel que podem concloure que hi Hi ha diferències significatives en les mitjanes de 'SleepTime' entre diferents nivells de 'AgeCategory'. Com que el resultat és significatiu també mirarem si hi ha diferències segons els grups d'edat mitjantcant una prova de TukeyHSD i veiem les següents dfierències signficiatives varis grups (veure codi) per exemple les hores de son son significativament diferents entre el grup d'edat 31-34 i el grup d'edadt 18-24. En canvi sembla ser que no hi ha diferenències significatives a nivell d'hores de son entre els més joves (18-24) i el grup de 65-69 anys d'edat.

Correlació entre es diversos factors de risc

El coeficient de correlació entre l'Activitat Física i el Temps de Son és de 0,0038, indicant una relació lineal molt feble o pràcticament nul·la. De manera similar, el coeficient de correlació entre Consum d'Alcohol i Temps de Son és de -0,0051, assenyalant una altra associació molt feble. El coeficient de correlació per a Fumar i Temps de Son és de -0.0303. Sembla que hi ha una lleugera tendència per a aquells que consumeixen més alcohol a tenir temps de son lleugerament més curts. Tot i això la correlacio entre aquestes variibales i les hores de son és feble i, per tant no podem concñiure res definitiu.

5. **Representació dels resultats** *a partir de taules i gràfiques. Aquest apartat es pot respondre al llarg de la pràctica, sense la necessitat de concentrar totes les representacions en aquest punt de la pràctica*

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

   En resum hem observat que:
   - Hi ha algunes malalties més prevalents segons el sexe. En el cas de l'asma és més prevalent en dones mentre que en el cas de Heart Disease és més prevaent en homes.
   - La raça blanca és més freqüent tenit càncer de pell en comparació amb altres races. També veiem que *RaceAsian*, *RaceBlack*, *RaceHispanic*

tenen menys risc de càncer sent la raça negra la associació més forta. En canvi, *RaceOthers* i *RaceWhite* tenen un major risc de patir aquest càncer.
- Observem que els patrons de descnas o hores de son son diversos en els diferents grups d'edat i que en més d'un grup les diferencies son signficiatvies. A més a més, si mirem l'impacte que té realitzar activtiat física, beure alcohol i fumar en dormir més o menys hores veiem que les correlacions entre aqeustes variables son febles pel que no podem concliure que siguin asuestes facotrs de risc els responsables de d'hormir més o menys hores.

Per tant, podem concliure que efectivament aquest conjunt de dades permet respondre a les nostres preguntes i que pot servir per detectar brots de malalties o bé factors de risc per aquestes.

7. Contribucions

Contribucions	Signatura	
Inevstigació prèvia	Maria Sopena	Joana Llauradó
Redacció de les respostes	Maria Sopena	Joana Llauradó
Desenvolupmanet del codi	Maria Sopena	Joana Llauradó
Participació del vídeo	Maria Sopena	Joana Llauradó

8. Vídeo

enllaç