# Format d´entrega

La documentació generada en la realització de la prova d´avaluació continuada es troba allotjada en **GitHub** al següent repositori:

* <https://github.com/rsanchezs/data-cleaning>

En aquest repositori es poden trobar els següents fitxers:

* Aquest document en formats **pdf** i **docx** amb el nom rsanchezs\_PAC2.
* Tots els documents, gràfics, imatges i taules utilitzades en aquest PAC.



# Exercici 1

Després de llegir el recurs “Clean Data”, contesta les següents preguntes amb les teves pròpies paraules.

1. Quines són les principals etapes d’un projecte analític en ciència de dades?. Descriu amb les teves paraules cadascuna de les etapes. Aquest procés ha de realitzar-se una sola vegada o per contra és un procés iteratiu? (màxim 250 paraules).

Tot seguit es mostra una figura amb la qual ens propossem exposar les diferents etapes d´un projecte analític.

Com es pot apreciar un projecte analític no és un procés que es dugui a terme en una sola fase i generalment s´ha de passar per diverses fases.

Aquestes fases són: entendre i expressar un tipus d´objectiu, entendre les dades amb les que treballarem, preparació de les dades, modelatge, avaluació i desplegament. Convé fer ressaltar que, aquest procés no és lineal, sinó es realimenta i continua.

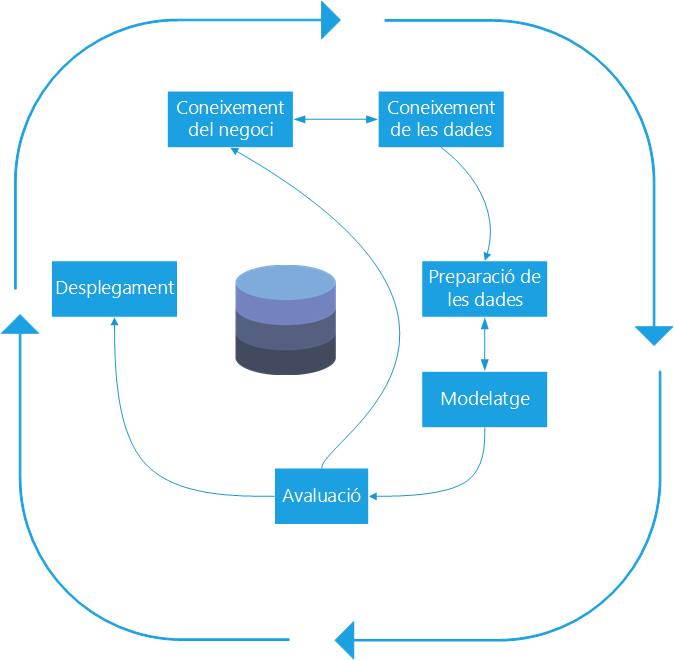


Fig. 1: Etapes d´un projecte analític

A continuació amb la següent taula es detallen les activitats que li corresponen a cada fase:

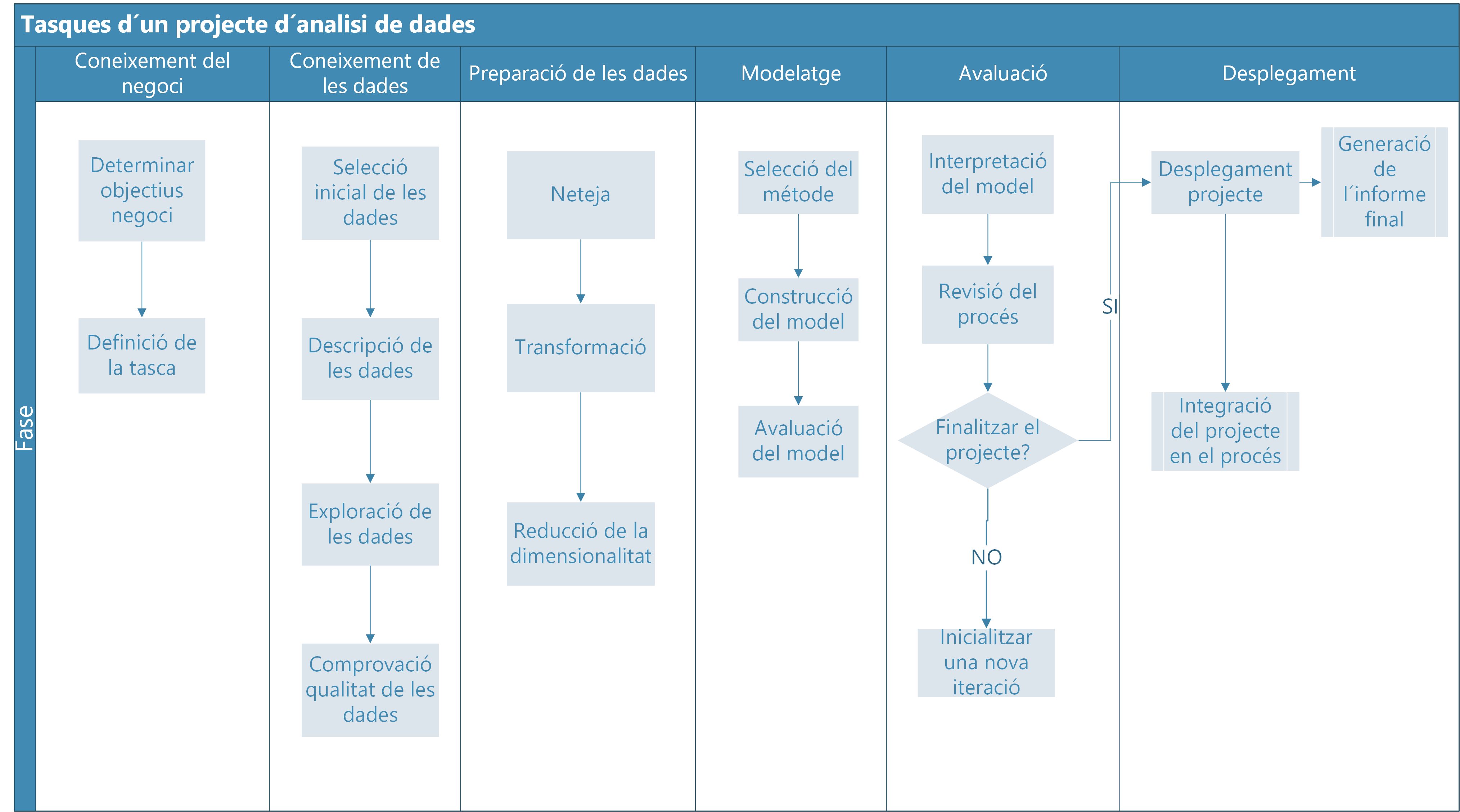


Fig. 2 Tasques d´un projecte analític

1. Suposa que es desitja transferir un grup arxius entre una empresa A i una altra empresa B. La primera treballa amb OSX i Linux com a sistema operatiu, mentre que la segona treballa amb Windows. Quin programa de compressió utilitzaria per a aquesta transmissió i per què? (màxim 100 paraules).

En primer lloc, convé fer ressaltar que quan es realitza una compressió d´un arxiu cal tindre en compte els següntes factors:

* La velocitat de compressió i descompressió.
* El ratio de compressió.
* La interoperabilitat de la solucio de compressió.
* Escollir el software de compressió:
* Zip
* gzip
* bzip2
* RAR

En la següent taula es mostren les diferents opcions:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Velocitat | Ratio | Interoperabilitat |
| gzip | Ràpida | Normal | Linux i OSX |
| bzip2 | Lenta | Petit | Linux i OSX |
| zip | Normal | Alt | Linux, OSX i Windows |
| RAR | Lenta | Normal | Windows |

Com es pot apreciar en la taula la millor opció sería realitzar la compressió amb **zip**.

1. Explica amb les teves pròpies paraules, quina és la diferència entre els zeros, les dades buides i les dades nul·les?. Amb l’ajuda d’un exemple real on existeix una pèrdua de dades, explica quan han de col·locar-se zeros i quan aquestes dades es deixaran com a dades buides durant el procés de data cleaning (màxim 200 paraules).

## Zeros

El valor zero es representa mitjançant el símbol **0**. Es tracta de un quantitat mesurable i que té un significat en un sistema numéric. Utilitzem zero quan ens referim a valors de tipus numéric.

## Dades buides

Podem trobar-nos amb dades buides quan no es coneix el valor per aquest atribut. Ens podem trobar amb els diferents tipus:

* **Cadenes de caracters**: s´utilitza la cadena “Desconegut” o no aplicable NA.
* **Dades numériques**: com hem vist en l´apartat anterior s´utilitza el valor 0.
* **Data**: les dades que representen dates i que no coneixem es poden representar mitjançant una data llunyana (1900-01-01) o primerenca (9999-12-31) en el temps.

## Dades nul.les

Les dades nul.les es representen mitjançant la paraula reservada NULL i ens indica que no existeix informació en un camp.

Aquesta falta d´informació pot venir o bé perquè no es aplicable o bé pel desconeixemnt del valor.

## Exemple

A continuació, amb l´ajuda del conjunt de dades diabetes veurem com tractar els valors desconeguts:

Realitzem un primer contacte amb el joc de dades, visualitzant la seva estructura i els 6 primers registres:

# Visualitzem les 6 primeres files  
head(diabetes)

## # A tibble: 6 x 10  
## Pregnancies Glucose `Blood pressure` Skinfold Insulin BMI  
## <int> <int> <int> <int> <int> <dbl>  
## 1 6 148 72 35 0 33.6  
## 2 1 85 66 29 0 26.6  
## 3 8 183 64 0 0 23.3  
## 4 1 89 66 23 94 28.1  
## 5 0 137 40 35 168 43.1  
## 6 5 116 74 0 0 25.6  
## # ... with 4 more variables: `Diabetes pedigree` <dbl>, Age <int>,  
## # Diabetes <chr>, weight <int>

Amb la següent instrucció podem comprobar el nombre de valors desconeguts per a cada atribut:

# Comprobació de valors desconeguts  
sapply(diabetes, function(x) sum(is.na(x)))

## Pregnancies Glucose Blood pressure Skinfold   
## 11 0 0 0   
## Insulin BMI Diabetes pedigree Age   
## 0 0 0 0   
## Diabetes weight   
## 0 0

Podem observar que, per a once observacions, el camp Pregnacies no té cap valor assignat. Això podría ocórrer per diversos motius:

* El mes senzill és que no s’hagi pogut obtenir aquesta informació, i en aquest cas podríem o bé emplenar aquests casos amb el valor més freqüent, o bé amb el valor mitjà.
* També podria ser que l’absència de valor per al camp embarassos ens indiques que aquestes files corresponen a pacients masculins, pels quals el camp no té sentit.

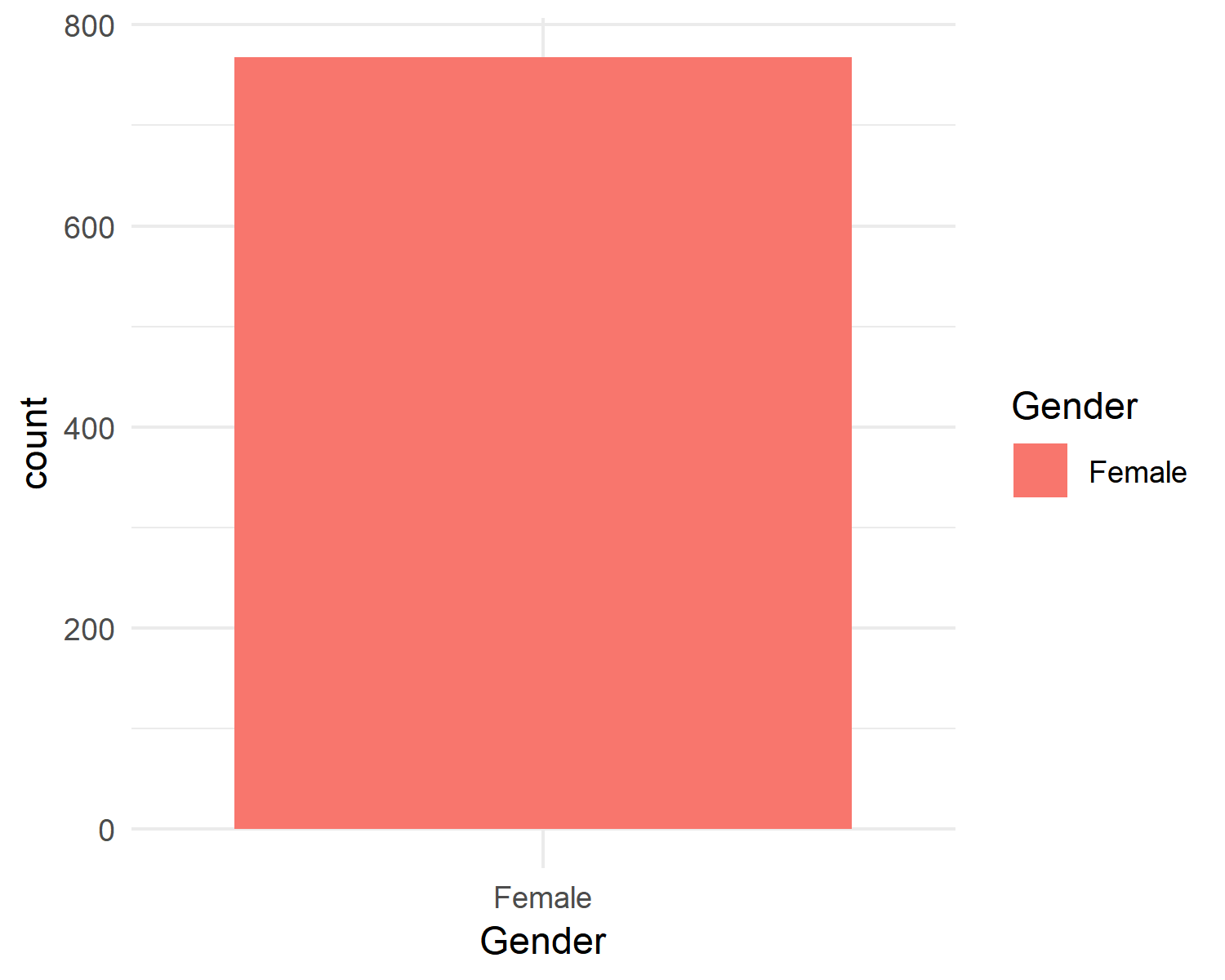
En qualsevol d’aquests dos casos, hauríem de transformar el conjunt de dades. En el primer supòsit, afegint un valor per defecte que fos raonable:

# Mitjançant una condició lógica obtenim els valors desconeguts  
missing\_values\_pregnacies <- is.na(diabetes$Pregnancies)  
# Reemplacem els valors desconeguts amb 0  
diabetes$Pregnancies[missing\_values\_pregnacies] <- 0

En el segon, creant un nou camp que contingui informació sobre si el pacient és un home o no.

Amb el següent diagrama podem comprobar la distribució de la nova variable Gender:

# Visualitzem la distribució de la variable  
diabetes %>% ggplot(aes(Gender)) +  
 geom\_bar(aes(fill = Gender)) +  
 theme\_minimal()



# Exercici 2

Després de llegir el recurs “Data Mining: concepts and techniques” contesta les següents preguntes.

1. Explica amb les teves pròpies paraules quines són les principals etapes del pre-processament de dades (màxim 400 paraules). Posa un exemple d’un projecte analític i indica a grans trets que s’hauria de fer en cada etapa.
2. Explica breument amb les teves pròpies paraules, tres dels factors que poden influir en l’estimació de la qualitat de les dades (màxim 100 paraules).

# Exercici 3

Després de llegir el recurs “Data Cleaning Basics”, contesta les següents preguntes:

1. Què es considera un extrem score ? Quins són els seus possibles efectes en els resultats finals? (màxim 150 paraules).
2. Explica dues possibles causes que poden donar lloc a l’aparició d’extrem scores i proposa una possible solució juntament amb un breu exemple pràctic per a cada cas (màxim 200 paraules).

# Bibliografia

[1]:Megan Squire (2015). Clean Data . Packt Publishing Ltd. Capítols 1 i 2.

[2]:Jiawei Han, Micheine Kamber, Jian Pei (2012). Data mining: concepts and techniques . Morgan Kaufmann. Capítol 3.

[3]:Jason W. Osborne (2010). Data Cleaning Basics: Best Practices in Dealing with Extreme Scores. Newborn and Infant Nursing Reviews; 10 (1): pp. 1527-3369 .