Jeu de données - Personnages Starwars

JH

2022-05-07

Introduction

Dans ce rapport nous pourrons observer les différentes caractéristiques des personnages de Starwars.

Présentation des données

```
##
## Attachement du package : 'dplyr'

## Les objets suivants sont masqués depuis 'package:stats':
##
##     filter, lag

## Les objets suivants sont masqués depuis 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
head(starwars)
```

```
## # A tibble: 6 x 14
##   name      height  mass hair_color skin_color eye_color birth_year sex   gender
##   <chr>      <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
## 1 Luke Sky~    172    77 blond      fair       blue            19   male  mascu~
## 2 C-3PO        167    75 <NA>       gold       yellow         112   none  mascu~
## 3 R2-D2         96    32 <NA>       white, bl~ red             33   none  mascu~
## 4 Darth Va~    202   136 none       white      yellow          41.9 male  mascu~
## 5 Leia Org~    150    49 brown      light      brown           19   fema~ femin~
## 6 Owen Lars    178   120 brown, gr~ light      blue            52   male  mascu~
## # ... with 5 more variables: homeworld <chr>, species <chr>, films <list>,
## #   vehicles <list>, starships <list>
```

```
names(starwars)
```

```
##  [1] "name"       "height"     "mass"       "hair_color" "skin_color"
##  [6] "eye_color"  "birth_year" "sex"        "gender"     "homeworld"
## [11] "species"    "films"      "vehicles"   "starships"
```

```
dim(starwars)
```

```
## [1] 87 14
```

```
ncol(starwars)
```

```
## [1] 14
```

```
nrow(starwars)
```

```
## [1] 87
```

Le jeu de données comprend les caractéristiques de 87 personnages. Ces caractéristiques sont décrites par 14 variables.

Sélection de données

Tout d'abord, nous allons sélectionner les individus 5, 25, 45, 65 et 85 dans le tableau de données.

```
starwars %>%
  slice(5, 25, 45, 65, 85)
```

```
## # A tibble: 5 x 14
##   name      height  mass hair_color skin_color eye_color birth_year sex   gender
##   <chr>      <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
## 1 Leia Org~    150    49 brown      light      brown             19 fema~ femin~
## 2 Lobot        175    79 none       light      blue              37 male  mascu~
## 3 Dud Bolt      94    45 none       blue, grey yellow            NA male  mascu~
## 4 Bail Pre~    191    NA black      tan        brown             67 male  mascu~
## 5 BB8           NA    NA none       none       black             NA none  mascu~
## # ... with 5 more variables: homeworld <chr>, species <chr>, films <list>,
## #   vehicles <list>, starships <list>
```

Le but est ensuite d'identifier l'ensemble des personnages féminins vivants sur la planète Tatooine.

```
starwars %>%
  filter(sex=="female") %>%
  filter(homeworld=="Tatooine")
```

```
## # A tibble: 2 x 14
##   name      height  mass hair_color skin_color eye_color birth_year sex   gender
##   <chr>      <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
## 1 Beru Whi~    165    75 brown      light      blue              47 fema~ femin~
## 2 Shmi Sky~    163    NA black      fair       brown             72 fema~ femin~
## # ... with 5 more variables: homeworld <chr>, species <chr>, films <list>,
## #   vehicles <list>, starships <list>
```

On sélectionne ensuite les femmes de plus de 2 mètres dans le but des les identifier. Comment s'appellent-elles ?

```
starwars %>%
  filter(sex=="female") %>%
  filter(height>200)
```

```
## # A tibble: 1 x 14
##   name    height  mass hair_color skin_color eye_color birth_year sex    gender
##   <chr>    <int> <dbl> <chr>      <chr>      <chr>           <dbl> <chr>  <chr>
## 1 Taun We    213    NA none       grey       black              NA female femini~
## # ... with 5 more variables: homeworld <chr>, species <chr>, films <list>,
## #   vehicles <list>, starships <list>
```

Il y a seulement un personnage féminin qui mesure plus de deux mètres, elle s'appelle Taun We.

Grâce à la commande suivante, on peut sélectionner les variables textuelles dans notre tableau de données.

```
starwars %>%
  select(where(is.character))
```

```
## # A tibble: 87 x 8
##    name           hair_color skin_color eye_color sex   gender homeworld species
##    <chr>          <chr>      <chr>      <chr>     <chr> <chr>  <chr>     <chr>
##  1 Luke Skywalker blond      fair       blue      male  mascu~ Tatooine  Human
##  2 C-3PO          <NA>       gold       yellow    none  mascu~ Tatooine  Droid
##  3 R2-D2          <NA>       white, bl~ red       none  mascu~ Naboo     Droid
##  4 Darth Vader    none       white      yellow    male  mascu~ Tatooine  Human
##  5 Leia Organa    brown      light      brown     fema~ femin~ Alderaan  Human
##  6 Owen Lars      brown, gr~ light      blue      male  mascu~ Tatooine  Human
##  7 Beru Whitesun~ brown      light      blue      fema~ femin~ Tatooine  Human
##  8 R5-D4          <NA>       white, red red       none  mascu~ Tatooine  Droid
##  9 Biggs Darklig~ black      light      brown     male  mascu~ Tatooine  Human
## 10 Obi-Wan Kenobi auburn, w~ fair       blue-gray male  mascu~ Stewjon   Human
## # ... with 77 more rows
```

Ainsi, on peut identifier 8 variables textuelles.

On a pu remarquer que les tailles des personnages étaient données en cm. Pour mettre en forme nos données, il peut être intéressant de créer une nouvelle colonne à partir d'une variable existente avec les tailles en mètres. C'est ce que nous allons faire avec les commandes suivantes :

```
starwars %>%
  mutate(height_meters = height/100) %>%
  select(name, height, height_meters) %>%
  arrange(height_meters)
```

```
## # A tibble: 87 x 3
##    name                  height height_meters
##    <chr>                  <int>         <dbl>
##  1 Yoda                      66          0.66
##  2 Ratts Tyerell             79          0.79
##  3 Wicket Systri Warrick     88          0.88
##  4 Dud Bolt                  94          0.94
##  5 R2-D2                     96          0.96
```

```
##  6 R4-P17                   96          0.96
##  7 R5-D4                    97          0.97
##  8 Sebulba                 112          1.12
##  9 Gasgano                 122          1.22
## 10 Watto                   137          1.37
## # ... with 77 more rows
```

Utilisation de ggplot2

Le package ggplot2 permet de concevoir des graphiques plus attractifs et plus complexes. C'est une extension de *tidyverse*.

Barplot

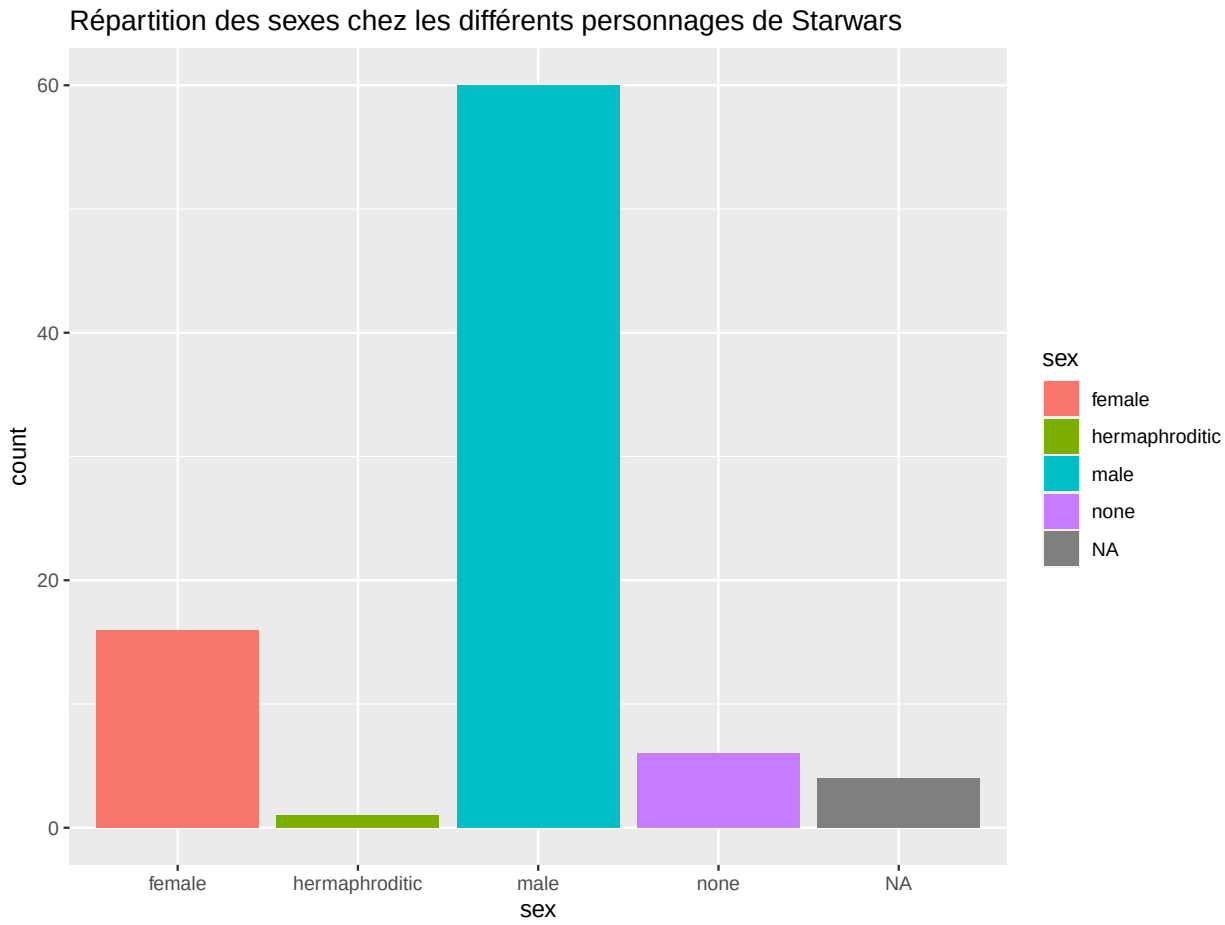
Il est possible de réaliser différents types de graphiques. On peut notamment réaliser des barplots comme ci-dessous :

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.6     v stringr 1.4.0
## v tidyr   1.2.0     v forcats 0.5.1
## v readr   2.1.2

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
ggplot(data=starwars) +
  geom_bar(mapping = aes (x = sex, fill = sex)) +
  ggtitle("Répartition des sexes chez les différents personnages de Starwars")
```



Cet exemple de graphique montre la répartition des sexes chez les personnages de Starwars.