Basics of programming

Jesus Sanchez

2022-10-17

#Exploring data

```
library(tidyverse)
```

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

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.1
## v readr   2.1.1     v forcats 0.5.1

## Warning: package 'stringr' was built under R version 4.1.3

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

```
starwars
```

```
## # A tibble: 87 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 S~    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 ~    202   136 none        white      yellow          41.9 male  mascu~
##  5 Leia O~    150    49 brown       light      brown           19   fema~ femin~
##  6 Owen L~    178   120 brown, grey light      blue            52   male  mascu~
##  7 Beru W~    165    75 brown       light      blue            47   fema~ femin~
##  8 R5-D4       97    32 <NA>        white, red red             NA   none  mascu~
##  9 Biggs ~    183    84 black       light      brown           24   male  mascu~
## 10 Obi-Wa~    182    77 auburn, wh~ fair       blue-gray       57   male  mascu~
## # ... with 77 more rows, and 5 more variables: homeworld <chr>, species <chr>,
## #   films <list>, vehicles <list>, starships <list>
```

```
dim(starwars) #87 obs and 14 variables
```

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

```
#str(starwars)
glimpse(starwars)
```

```
## Rows: 87
## Columns: 14
## $ name       <chr> "Luke Skywalker", "C-3PO", "R2-D2", "Darth Vader", "Leia Or~
## $ height     <int> 172, 167, 96, 202, 150, 178, 165, 97, 183, 182, 188, 180, 2~
## $ mass       <dbl> 77.0, 75.0, 32.0, 136.0, 49.0, 120.0, 75.0, 32.0, 84.0, 77.~
## $ hair_color <chr> "blond", NA, NA, "none", "brown", "brown, grey", "brown", N~
## $ skin_color <chr> "fair", "gold", "white, blue", "white", "light", "light", "~
## $ eye_color  <chr> "blue", "yellow", "red", "yellow", "brown", "blue", "blue",~
## $ birth_year <dbl> 19.0, 112.0, 33.0, 41.9, 19.0, 52.0, 47.0, NA, 24.0, 57.0, ~
## $ sex        <chr> "male", "none", "none", "male", "female", "male", "female",~
## $ gender     <chr> "masculine", "masculine", "masculine", "masculine", "femini~
## $ homeworld  <chr> "Tatooine", "Tatooine", "Naboo", "Tatooine", "Alderaan", "T~
## $ species    <chr> "Human", "Droid", "Droid", "Human", "Human", "Human", "Huma~
## $ films      <list> <"The Empire Strikes Back", "Revenge of the Sith", "Return~
## $ vehicles   <list> <"Snowspeeder", "Imperial Speeder Bike">, <>, <>, <>, "Imp~
## $ starships  <list> <"X-wing", "Imperial shuttle">, <>, <>, "TIE Advanced x1",~
```

```
head(starwars) #first 6 obs
```

```
## # 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 Sk~    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 V~    202   136 none        white      yellow          41.9 male  mascu~
## 5 Leia Or~    150    49 brown       light      brown           19   fema~ femin~
## 6 Owen La~    178   120 brown, grey light      blue            52   male  mascu~
## # ... with 5 more variables: homeworld <chr>, species <chr>, films <list>,
## #   vehicles <list>, starships <list>
```

```
tail(starwars) #last 6 obs
```

```
## # 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 Finn         NA    NA black      dark       dark              NA male   mascu~
## 2 Rey          NA    NA brown      light      hazel             NA female femin~
## 3 Poe Dam~     NA    NA brown      light      brown             NA male   mascu~
## 4 BB8          NA    NA none       none       black             NA none   mascu~
## 5 Captain~     NA    NA unknown    unknown    unknown           NA <NA>   <NA>
## 6 Padmé A~    165    45 brown      light      brown             46 female femin~
## # ... with 5 more variables: homeworld <chr>, species <chr>, films <list>,
## #   vehicles <list>, starships <list>
```

```
attach(starwars) #this is useful to stop writing starwars$..., now I can write any variable of starwars
hair_color
```

```
##  [1] "blond"         NA              NA              "none"
##  [5] "brown"         "brown, grey"   "brown"         NA
##  [9] "black"         "auburn, white" "blond"         "auburn, grey"
## [13] "brown"         "brown"         NA              NA
## [17] "brown"         "brown"         "white"         "grey"
## [21] "black"         "none"          "none"          "black"
## [25] "none"          "none"          "auburn"        "brown"
## [29] "brown"         "none"          "brown"         "none"
## [33] "blond"         "none"          "none"          "none"
## [37] "brown"         "black"         "none"          "black"
## [41] "black"         "none"          "none"          "none"
## [45] "none"          "none"          "none"          "none"
## [49] "white"         "none"          "black"         "none"
## [53] "none"          "none"          "none"          "none"
## [57] "black"         "brown"         "brown"         "none"
## [61] "black"         "black"         "brown"         "white"
## [65] "black"         "black"         "blonde"        "none"
## [69] "none"          "none"          "white"         "none"
## [73] "none"          "none"          "none"          "none"
## [77] "none"          "brown"         "brown"         "none"
## [81] "none"          "black"         "brown"         "brown"
## [85] "none"          "unknown"       "brown"
```

```
names(starwars) #names of my variables
```

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

```
length(starwars) #for a data set length will mean the number of variables
```

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

```
length(hair_color) #for a variable R will tell the number of obs
```

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

```
class(hair_color)
```

```
## [1] "character"
```

```
unique(hair_color) #name of unique obs
```

```
##  [1] "blond"         NA              "none"          "brown"
##  [5] "brown, grey"   "black"         "auburn, white" "auburn, grey"
##  [9] "white"         "grey"          "auburn"        "blonde"
## [13] "unknown"
```

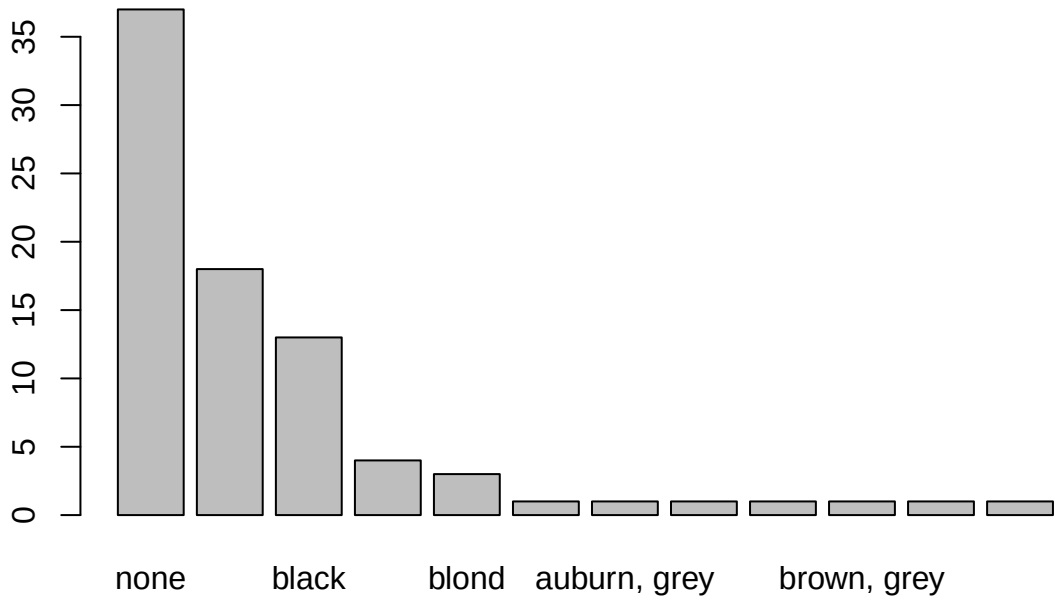
```r
#na: data is missing
#none: hair without a color or there´s no hair
#unknow: we don't know, maybe the character uses a hat, so we don't know the color

table(hair_color)
```

```
## hair_color
##        auburn  auburn, grey auburn, white         black         blond
##             1             1             1            13             3
##        blonde         brown   brown, grey          grey          none
##             1            18             1             1            37
##       unknown         white
##             1             4
```

```r
sort(table(hair_color), decreasing=T)
```

```
## hair_color
##          none         brown         black         white         blond
##            37            18            13             4             3
##        auburn  auburn, grey auburn, white        blonde   brown, grey
##             1             1             1             1             1
##          grey       unknown
##             1             1
```

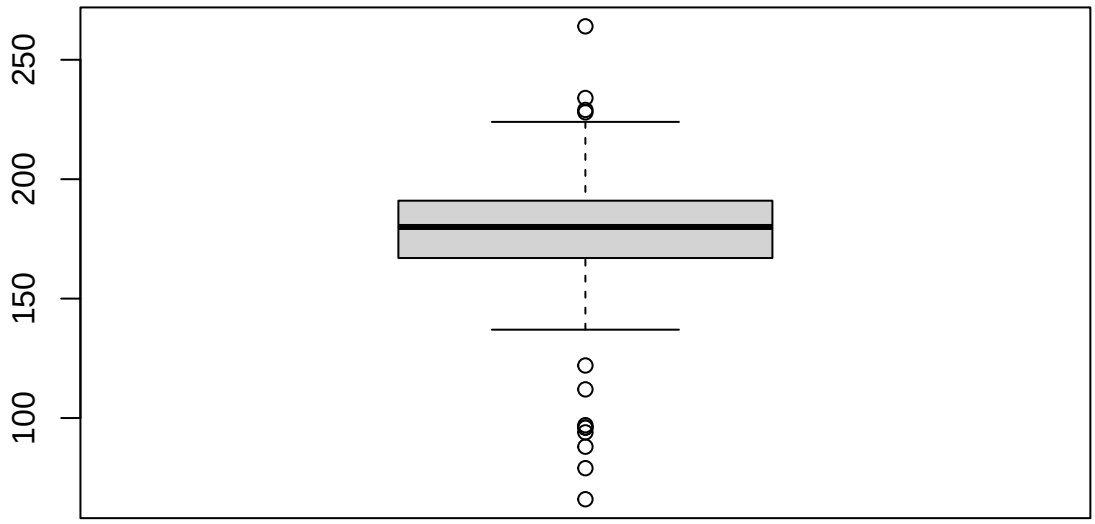
```r
View(sort(table(hair_color), decreasing=T))
barplot(sort(table(hair_color), decreasing=T))
```



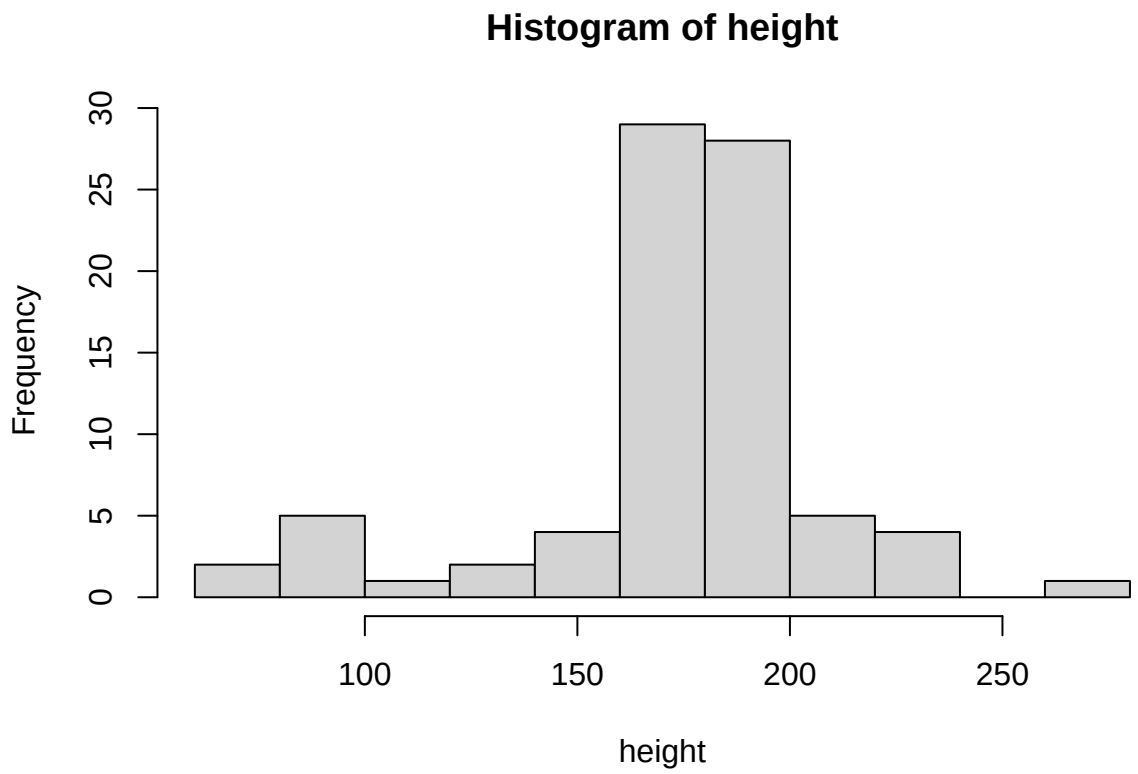

```
#pipes operators

starwars %>%
  select(hair_color) %>%
  count(hair_color) %>%
  arrange(desc(n)) %>%
  View()

View(starwars[is.na(hair_color),])  #selecting row where is.na is TRUE

summary(height)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    66.0   167.0   180.0   174.4   191.0   264.0       6
```

```
boxplot(height) #boxplot
```




```
hist(height) #histeogram
```



#Cleaning data

```
library(tidyverse)
data()
View(starwars)
glimpse(starwars)
```

```
## Rows: 87
## Columns: 14
## $ name       <chr> "Luke Skywalker", "C-3PO", "R2-D2", "Darth Vader", "Leia Or~
## $ height     <int> 172, 167, 96, 202, 150, 178, 165, 97, 183, 182, 188, 180, 2~
## $ mass       <dbl> 77.0, 75.0, 32.0, 136.0, 49.0, 120.0, 75.0, 32.0, 84.0, 77.~
## $ hair_color <chr> "blond", NA, NA, "none", "brown", "brown, grey", "brown", N~
## $ skin_color <chr> "fair", "gold", "white, blue", "white", "light", "light", "~
## $ eye_color  <chr> "blue", "yellow", "red", "yellow", "brown", "blue", "blue",~
## $ birth_year <dbl> 19.0, 112.0, 33.0, 41.9, 19.0, 52.0, 47.0, NA, 24.0, 57.0, ~
## $ sex        <chr> "male", "none", "none", "male", "female", "male", "female",~
## $ gender     <chr> "masculine", "masculine", "masculine", "masculine", "femini~
## $ homeworld  <chr> "Tatooine", "Tatooine", "Naboo", "Tatooine", "Alderaan", "T~
## $ species    <chr> "Human", "Droid", "Droid", "Human", "Human", "Human", "Huma~
## $ films      <list> <"The Empire Strikes Back", "Revenge of the Sith", "Return~
## $ vehicles   <list> <"Snowspeeder", "Imperial Speeder Bike">, <>, <>, <>, "Imp~
## $ starships  <list> <"X-wing", "Imperial shuttle">, <>, <>, "TIE Advanced x1",~
```

```r
unique(starwars$gender) #displays data type in the obs of a specific column
```

```
## [1] "masculine" "feminine"  NA
```

```r
attach(starwars)
```

```
## The following objects are masked from starwars (pos = 3):
##
##     birth_year, eye_color, films, gender, hair_color, height,
##     homeworld, mass, name, sex, skin_color, species, starships,
##     vehicles
```

```r
starwars$gender <- as.factor(starwars$gender)
class(starwars$gender)#now gender is a factor
```

```
## [1] "factor"
```

```r
class(starwars$gender)
```

```
## [1] "factor"
```

```r
levels(starwars$gender)
```

```
## [1] "feminine"  "masculine"
```

```r
starwars$gender <- factor((starwars$gender), levels = c("f", "m"))
#changing levels

starwars %>% select(name, height, ends_with("color")) %>%
  names()
```

```
## [1] "name"       "height"     "hair_color" "skin_color" "eye_color"
```

```r
unique(starwars$hair_color)
```

```
##  [1] "blond"         NA              "none"          "brown"
##  [5] "brown, grey"   "black"         "auburn, white" "auburn, grey"
##  [9] "white"         "grey"          "auburn"        "blonde"
## [13] "unknown"
```

```r
starwars %>%
  select(name, height, ends_with("color")) %>%
  filter(hair_color %in% c("blond", "brown") & height < 180)
```

```
## # A tibble: 9 x 5
##   name                height hair_color skin_color eye_color
##   <chr>                <int> <chr>      <chr>      <chr>
## 1 Luke Skywalker         172 blond      fair       blue
```

```
## 2 Leia Organa              150 brown      light      brown
## 3 Beru Whitesun lars       165 brown      light      blue
## 4 Wedge Antilles           170 brown      fair       hazel
## 5 Wicket Systri Warrick     88 brown      brown      brown
## 6 Finis Valorum            170 blond      fair       blue
## 7 Cordé                    157 brown      light      brown
## 8 Dormé                    165 brown      light      brown
## 9 Padmé Amidala            165 brown      light      brown
```

```
#%in% works for group more than 1 variable

#missing data

mean(starwars$height) #we have a NA because there's missin values Na
```

```
## [1] NA
```

```
mean(starwars$height, na.rm = T)
```

```
## [1] 174.358
```

```
starwars %>%
  select(name, gender, hair_color, height) %>%
  na.omit()
```

```
## # A tibble: 0 x 4
## # ... with 4 variables: name <chr>, gender <fct>, hair_color <chr>,
## #   height <int>
```

```
starwars %>%
  select(name, gender, hair_color, height) %>%
  filter(!complete.cases(.)) #what obs we deleted
```

```
## # A tibble: 87 x 4
##    name               gender hair_color    height
##    <chr>              <fct>  <chr>          <int>
##  1 Luke Skywalker     <NA>   blond            172
##  2 C-3PO              <NA>   <NA>             167
##  3 R2-D2              <NA>   <NA>              96
##  4 Darth Vader        <NA>   none             202
##  5 Leia Organa        <NA>   brown            150
##  6 Owen Lars          <NA>   brown, grey      178
##  7 Beru Whitesun lars <NA>   brown            165
##  8 R5-D4              <NA>   <NA>              97
##  9 Biggs Darklighter  <NA>   black            183
## 10 Obi-Wan Kenobi     <NA>   auburn, white    182
## # ... with 77 more rows
```

```
starwars %>%
  select(name, gender, hair_color, height) %>%
  filter(!complete.cases(.)) %>%
  drop_na(height)
```

```
## # A tibble: 81 x 4
##    name               gender hair_color    height
##    <chr>              <fct>  <chr>          <int>
##  1 Luke Skywalker     <NA>   blond            172
##  2 C-3PO              <NA>   <NA>             167
##  3 R2-D2              <NA>   <NA>              96
##  4 Darth Vader        <NA>   none             202
##  5 Leia Organa        <NA>   brown            150
##  6 Owen Lars          <NA>   brown, grey       178
##  7 Beru Whitesun lars <NA>   brown            165
##  8 R5-D4              <NA>   <NA>              97
##  9 Biggs Darklighter  <NA>   black            183
## 10 Obi-Wan Kenobi     <NA>   auburn, white    182
## # ... with 71 more rows
```

```
starwars %>%
  select(name, gender, hair_color, height) %>%
  filter(!complete.cases(.)) %>%
  mutate(hair_color = replace_na(hair_color, "none"))
```

```
## # A tibble: 87 x 4
##    name               gender hair_color    height
##    <chr>              <fct>  <chr>          <int>
##  1 Luke Skywalker     <NA>   blond            172
##  2 C-3PO              <NA>   none             167
##  3 R2-D2              <NA>   none              96
##  4 Darth Vader        <NA>   none             202
##  5 Leia Organa        <NA>   brown            150
##  6 Owen Lars          <NA>   brown, grey       178
##  7 Beru Whitesun lars <NA>   brown            165
##  8 R5-D4              <NA>   none              97
##  9 Biggs Darklighter  <NA>   black            183
## 10 Obi-Wan Kenobi     <NA>   auburn, white    182
## # ... with 77 more rows
```

```
#replacing all NA values from hair_color

#Duplicates-------------------------------

Names <- c("Peter", "John", "Andrew", "Peter")
Age <- c(22,33,44,22)

friends <- data.frame(Names, Age)
duplicated(friends) #reporting duplicates
```

```
## [1] FALSE FALSE FALSE  TRUE
```

```
friends[!duplicated(friends), ] #the archaic method
```

```
##    Names Age
## 1  Peter  22
## 2   John  33
## 3 Andrew  44
```

```
friends %>% distinct() #using tydiverse
```

```
##    Names Age
## 1  Peter  22
## 2   John  33
## 3 Andrew  44
```

```
#recording variables----------------------------------

starwars %>% select(name, gender)
```

```
## # A tibble: 87 x 2
##    name               gender
##    <chr>              <fct>
##  1 Luke Skywalker     <NA>
##  2 C-3PO              <NA>
##  3 R2-D2              <NA>
##  4 Darth Vader        <NA>
##  5 Leia Organa        <NA>
##  6 Owen Lars          <NA>
##  7 Beru Whitesun lars <NA>
##  8 R5-D4              <NA>
##  9 Biggs Darklighter  <NA>
## 10 Obi-Wan Kenobi     <NA>
## # ... with 77 more rows
```

```
class(starwars$gender)
```

```
## [1] "factor"
```

```
starwars$gender <- as.factor(starwars$gender)
class(starwars$gender) #now we can recode the variable
```

```
## [1] "factor"
```

```
levels(starwars$gender)
```

```
## [1] "f" "m"
```

```
starwars %>%
  select(name, gender) %>%
  mutate(gender_coded = recode(gender,
                         "masculine"= 1,
                         "feminine" = 2))
```

```
## # A tibble: 87 x 3
##    name               gender gender_coded
##    <chr>              <fct>         <dbl>
##  1 Luke Skywalker     <NA>             NA
```

```
##  2 C-3PO              <NA>             NA
##  3 R2-D2              <NA>             NA
##  4 Darth Vader        <NA>             NA
##  5 Leia Organa        <NA>             NA
##  6 Owen Lars          <NA>             NA
##  7 Beru Whitesun lars <NA>             NA
##  8 R5-D4              <NA>             NA
##  9 Biggs Darklighter  <NA>             NA
## 10 Obi-Wan Kenobi     <NA>             NA
## # ... with 77 more rows
```

#Manipulating data

#Visualise

#Analyse