Opgaver lektion 3

Simon

19/11/2022

Inden opgaven læs følgende pakker!

```
library(tidyverse)
library(ggplot2)
```

Opgave 1:

Brug de følgende steps til at udregne BMI for Starwars karaktere. baseret på deres hår farve:

Step 1

Vælg de 4 variable vi ønsker at bruge (Vi beholder "name")

Fejrn "NA" værdier med funktionen "drop_na()"

```
starwars %>%
  select(name, hair_color, mass, height)%>%
  drop_na()
```

```
## # A tibble: 54 x 4
##    name               hair_color     mass height
##    <chr>              <chr>         <dbl>  <int>
##  1 Luke Skywalker     blond            77    172
##  2 Darth Vader        none            136    202
##  3 Leia Organa        brown            49    150
##  4 Owen Lars          brown, grey     120    178
##  5 Beru Whitesun lars brown            75    165
##  6 Biggs Darklighter  black            84    183
##  7 Obi-Wan Kenobi     auburn, white    77    182
##  8 Anakin Skywalker   blond            84    188
##  9 Chewbacca          brown           112    228
## 10 Han Solo           brown            80    180
## # ... with 44 more rows
## # i Use 'print(n = ...)' to see more rows
```

Step 2

Brug mutate til at lave en ny kolonne der viser BMI:

Formlen for BMI er følgende: $BMI = \frac{mass(kg)}{height(M)^2}$

Da vi skal have height i meter skal vi også bruge mutate til at øndre denne kolonne:

Husk vi stadig skal bruge ovenstående steps

```
starwars %>%
  select(name, hair_color, mass, height)%>%
  drop_na() %>%
  mutate(height= height/100)
```

```
## # A tibble: 54 x 4
##    name               hair_color     mass height
##    <chr>              <chr>         <dbl>  <dbl>
##  1 Luke Skywalker     blond            77   1.72
##  2 Darth Vader        none            136   2.02
##  3 Leia Organa        brown            49   1.5
##  4 Owen Lars          brown, grey     120   1.78
##  5 Beru Whitesun lars brown            75   1.65
##  6 Biggs Darklighter  black            84   1.83
##  7 Obi-Wan Kenobi     auburn, white    77   1.82
##  8 Anakin Skywalker   blond            84   1.88
##  9 Chewbacca          brown           112   2.28
## 10 Han Solo           brown            80   1.8
## # ... with 44 more rows
## # i Use 'print(n = ...)' to see more rows
```

Vi tilføjer nu BMI:

```
starwars %>%
  select(name, hair_color, mass, height)%>%
  drop_na() %>%
  mutate(height= height/100) %>%
  mutate(BMI= mass/(height^2))
```

```
## # A tibble: 54 x 5
##    name               hair_color     mass height   BMI
##    <chr>              <chr>         <dbl>  <dbl> <dbl>
##  1 Luke Skywalker     blond            77   1.72  26.0
##  2 Darth Vader        none            136   2.02  33.3
##  3 Leia Organa        brown            49   1.5   21.8
##  4 Owen Lars          brown, grey     120   1.78  37.9
##  5 Beru Whitesun lars brown            75   1.65  27.5
##  6 Biggs Darklighter  black            84   1.83  25.1
##  7 Obi-Wan Kenobi     auburn, white    77   1.82  23.2
##  8 Anakin Skywalker   blond            84   1.88  23.8
##  9 Chewbacca          brown           112   2.28  21.5
## 10 Han Solo           brown            80   1.8   24.7
## # ... with 44 more rows
## # i Use 'print(n = ...)' to see more rows
```

Hurtig Øvelse, Hvem har den højeste og mindste BMI? Brug Arrange() funktionen

Step 3

Vi kan nu udregne gennemsnit af BMI for hver hår farve:

Vi bruger en ny funktion kaldet "group_by()" Den opdeller kategoriske variable op i hver deres kategori og fungere derfor godt sammen med summarise funktionen:

```
starwars %>%
  select(name, hair_color, mass, height)%>%
  drop_na() %>%
  mutate(height= height/100) %>%
  mutate(BMI= mass/(height^2)) %>%
  group_by(hair_color) %>%
  summarise(mean(BMI))
```

```
## # A tibble: 9 x 2
##   hair_color    `mean(BMI)`
##   <chr>               <dbl>
## 1 auburn, white        23.2
## 2 black                22.8
## 3 blond                24.9
## 4 blonde               19.5
## 5 brown                24.5
## 6 brown, grey          37.9
## 7 grey                 26.0
## 8 none                 24.4
## 9 white                27.1
```

Så det ser ud til Starwars karaktere med brundt/gråt hår har den højeste BMI!

Step 4

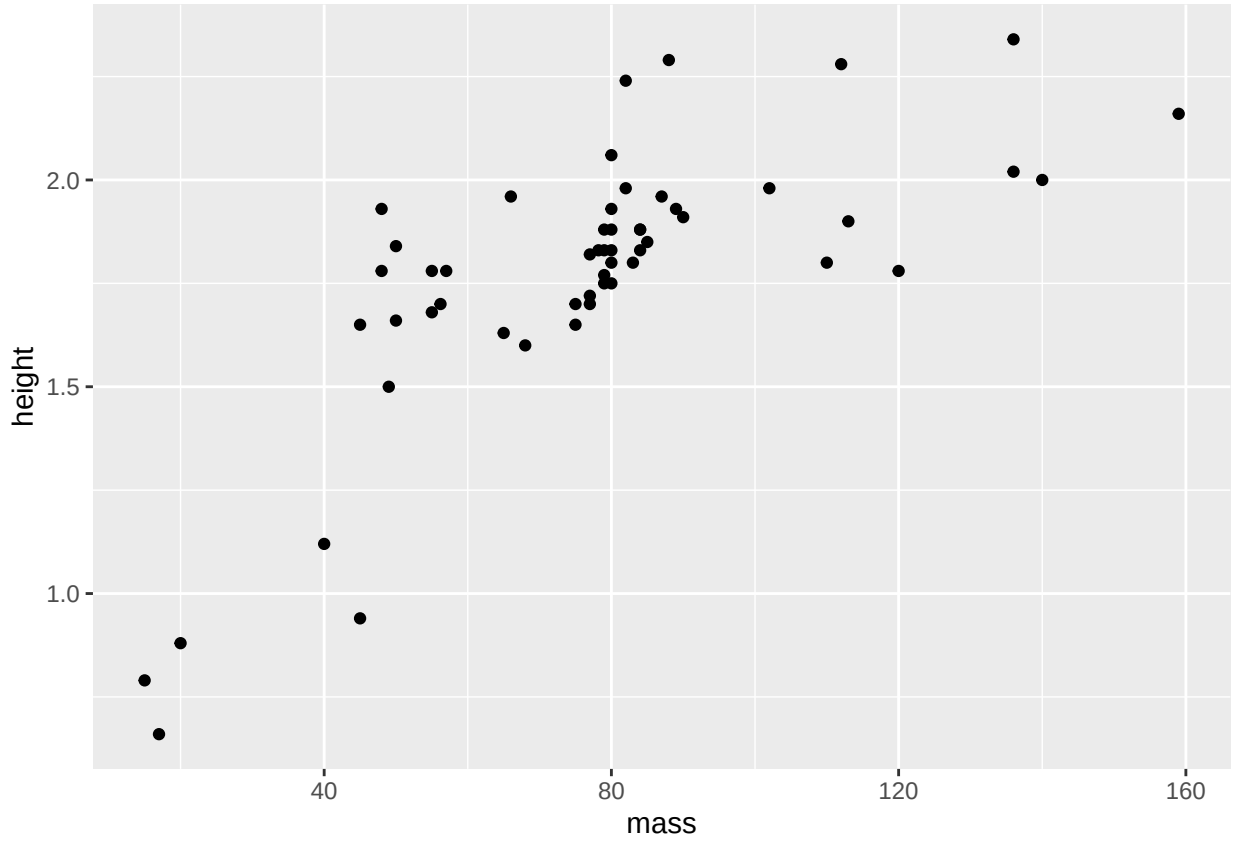
Visualisering lad os bruge datasættet fra **step 2** derfor gemmer vi dette som et nyt dataset kaldet "starwars_plots"

```
starwars_plots= starwars %>%
  select(name, hair_color, mass, height)%>%
  drop_na() %>%
  mutate(height= height/100) %>%
  mutate(BMI= mass/(height^2))
```

Step 5

Lav et plot der viser Ma

```
starwars_plots %>% ggplot(aes(x = mass, y= height)) +
  geom_point()
```



opgave 2

Jeg har hentet data på arbejdsløshedsraten direkte inde fra OECD. Her kræves der først en del arbejde med dataet føør vi kan bruge det til noget....

```
library(readxl)
OECD_data <- read_excel("OECD data.xlsx")
glimpse(OECD_data)
```

```
## Rows: 305
## Columns: 8
## $ LOCATION     <chr> "DNK", "DNK", "DNK", "DNK", "DNK", "DNK", "DNK", "DNK", "~
## $ INDICATOR    <chr> "UNEMP", "UNEMP", "UNEMP", "UNEMP", "UNEMP", "UNEMP", "UN~
## $ SUBJECT      <chr> "25_74", "25_74", "25_74", "25_74", "25_74", "25_74", "25~
## $ MEASURE      <chr> "PC_LF", "PC_LF", "PC_LF", "PC_LF", "PC_LF", "PC_LF", "PC~
## $ FREQUENCY    <chr> "Q", "Q", "Q", "Q", "Q", "Q", "Q", "Q", "Q", "Q", "Q", "Q~
## $ TIME         <chr> "2005-Q1", "2005-Q2", "2005-Q3", "2005-Q4", "2006-Q1", "2~
## $ Value        <dbl> 4.833333, 4.433333, 4.100000, 3.433333, 3.366667, 3.36666~
## $ `Flag Codes` <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
```

Alle navne er med stort, hvilket bliver lidt nederen i længden... Så jeg ændre til små bogstaver:

```
names(OECD_data) <- tolower(names(OECD_data))
```

```
OECD_data %>%
  count(location)
```

```
## # A tibble: 5 x 2
##   location     n
##   <chr>    <int>
## 1 DEU         61
## 2 DNK         61
## 3 ITA         61
## 4 OECD        61
## 5 USA         61
```

Step 1

Brug nu select funktionen til at vælge de tre kolonner: "location","value" og "time".

```
OECD_data %>%
  select(location, value, time)
```

```
## # A tibble: 305 x 3
##    location value time
##    <chr>    <dbl> <chr>
##  1 DNK       4.83 2005-Q1
##  2 DNK       4.43 2005-Q2
##  3 DNK       4.1  2005-Q3
##  4 DNK       3.43 2005-Q4
##  5 DNK       3.37 2006-Q1
##  6 DNK       3.37 2006-Q2
##  7 DNK       3.17 2006-Q3
##  8 DNK       3.03 2006-Q4
##  9 DNK       3.3  2007-Q1
## 10 DNK       3    2007-Q2
## # ... with 295 more rows
## # i Use 'print(n = ...)' to see more rows
```

Vi vil nu gerne lave en kolonne med henholdsvis Danmarks, Tysklands, Italians, USAs og OECDs arbejdsløshedsrate.

Dette kan gøres på flere måder:

1. Brug filter() funktionen til at vælge observationer kun for Danmark, gem dette i et dataset → Gør nu det samme for de andre lande → brug left_join() funktionen til at sammensætte de 5 individuelle dataset.

```
DK_dat=OECD_data %>%
  select(location, value, time) %>%
  filter(location == "DNK") %>%
  rename(dk_value = value)
```

```
DEU_dat=OECD_data %>%
  select(location, value, time) %>%
  filter(location == "DEU") %>%
  rename(deu_value = value)

ITA_dat=OECD_data %>%
  select(location, value, time) %>%
  filter(location == "ITA") %>%
  rename(ita_value = value)

OECD_dat=OECD_data %>%
  select(location, value, time) %>%
  filter(location == "OECD")%>%
  rename(oecd_value = value)

USA_dat=OECD_data %>%
  select(location, value, time) %>%
  filter(location == "USA")%>%
  rename(usa_value = value)

option_1_data= DK_dat %>%
  left_join(DEU_dat, by = c("time")) %>%
  left_join(ITA_dat, by = c("time")) %>%
  left_join(OECD_dat, by = c("time")) %>%
  left_join(USA_dat, by = c("time")) %>%
  select(time,dk_value, deu_value, ita_value, oecd_value, usa_value)
```

Denne metode tager en del tid og kode, nogle gange er det nødvendigt, men til lige præcis den her opgave kan vi bruge pivot_wider funktionen som gør det meget nemmere:

2. Brug pivot_wider() funktionen

- Søg pivot_wider funktionen op under "Help"
- brug "namess_from" og "values_from" til at få dit dataset.

```
OECD_data %>%
  select(location, value, time) %>%
  pivot_wider( names_from = location, values_from= value)
```

```
## # A tibble: 61 x 6
##    time      DNK   DEU   ITA   USA  OECD
##    <chr>   <dbl> <dbl> <dbl> <dbl> <dbl>
##  1 2005-Q1  4.83 10.3    6.4  4.11  5.72
##  2 2005-Q2  4.43 10.6    6.4  3.98  5.70
##  3 2005-Q3  4.1  10.7    6.1  3.93  5.61
##  4 2005-Q4  3.43 10.4    6.2  3.93  5.58
##  5 2006-Q1  3.37 10.1    5.9  3.70  5.42
##  6 2006-Q2  3.37  9.73   5.6  3.69  5.27
##  7 2006-Q3  3.17  9.43   5.4  3.59  5.18
##  8 2006-Q4  3.03  9.17   5.3  3.42  5.01
##  9 2007-Q1  3.3   8.63   5.1  3.57  4.94
```

```
## 10 2007-Q2  3     8.2    4.9  3.51  4.76
## # ... with 51 more rows
## # i Use ‘print(n = ...)‘ to see more rows
```