Assignment 2

For Assignment 2, use the tidyverse package and avoid using base R functions.

Questions 1

- Download the BDHS data

```
bdhs2014 <- haven::read_dta("data/bdhs2014.dta")
bdhs2014 |> slice(1:6)
```

```
#> # A tibble: 6 x 76
#>   caseid    hidx v000   v001  v002  v003  v004  v008  v011  v012 v013    v015
#>   <chr>    <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl+l> <dbl+l>
#> 1 "      ~     1 BD6     306    29     2   306  1377  1066    25 3 [25-~ 1 [com~
#> 2 "      ~     1 BD6     568    87     7   568  1377  1161    18 1 [15-~ 1 [com~
#> 3 "      ~     1 BD6     298    37     4   298  1376  1076    25 3 [25-~ 1 [com~
#> 4 "      ~     1 BD6     289    95     2   289  1378  1100    23 2 [20-~ 1 [com~
#> 5 "      ~     1 BD6     337    25     2   337  1379   968    34 4 [30-~ 1 [com~
#> 6 "      ~     1 BD6     500    95     4   500  1376  1126    20 2 [20-~ 1 [com~
#> # i 64 more variables: v020 <dbl+lbl>, v024 <dbl+lbl>, v025 <dbl+lbl>,
#> #   v102 <dbl+lbl>, v106 <dbl+lbl>, v107 <dbl+lbl>, v113 <dbl+lbl>,
#> #   v116 <dbl+lbl>, v119 <dbl+lbl>, v120 <dbl+lbl>, v121 <dbl+lbl>,
#> #   v122 <dbl+lbl>, v123 <dbl+lbl>, v124 <dbl+lbl>, v125 <dbl+lbl>,
#> #   v127 <dbl+lbl>, v128 <dbl+lbl>, v129 <dbl+lbl>, v130 <dbl+lbl>,
#> #   v133 <dbl+lbl>, v135 <dbl+lbl>, v136 <dbl>, v137 <dbl>, v138 <dbl>,
#> #   v140 <dbl+lbl>, v150 <dbl+lbl>, v151 <dbl+lbl>, v152 <dbl+lbl>, ...
```

- The data contains information on different variables obtained for 7886 ever-married Bangladeshi women of reproductive age. Below are the definitions of different variables.

Variable	Description	Value Labels
caseid	case identification	
v002	household number	
v012	respondent's current age	
v101	region	1=Barisal, 2=Chittagong, 3=Dhaka, 4=Khulna, 5=Rajshahi, 6=Rangpur, 7=Sylhet
v102	type of place of residence	1=Urban, 2=Rural
v106	highest educational level	0=No education, 1=Primary, 2=Secondary, 3=Higher
v119	household has electricity	0=No, 1=Yes
v121	household has television	0=No, 1=Yes
v130	religion	1=Islam, 2=Hinduism, 3=Christianity, 4=Buddhism, 96=Others
v190	wealth index	1=Poorest, 2=Poorer, 3=Middle, 4=Richer, 5=Richest
v501	current marital status	0=Never married, 1=Married, 2=Living with partner, 3=Widowed, 4=Divorced, 5=Separated
v701	husband/partner's education level	0=No education, 1=Primary, 2=Secondary, 3=Higher, 8=Don't know

- To view the label of a specific variable, use the `var_label()` function from the `labelled` package:

```
library(labelled)
var_label(bdhs2014$v012)
```

```
#> [1] "respondent's current age"
```

- Alternatively, you can search for variable labels using keywords (e.g., `age` or `weight`):

```
labelled::look_for(bdhs2014, "weight")
```

```
#>  pos variable label                           col_type missing
#>  62  v437     respondent's weight in kilogra~ dbl+lbl  4
#>
#>
#>  76  bwt      Weight at Birth in grams        dbl      0
#>  values
```

```
#>  [9994] not present
#>  [9995] refused
#>  [9996] other
#>
```

Questions:

1. How many observations and variables do the `bdhs2014` data have?
2. Rename the variable `v130` to `religion`
3. Create a subset of the data for women with age (`v012`) greater than 25, and save it as `bdhs_age_20_plus`.
4. Create a new variable named `age_gap` which will be the difference of the age of husband (`v730`) and age of the women (`v012`).
5. Create a new variable as indicator of early child bearing using the mother's age at first birth (`v212`). The new variable named `ecb` is 1 if mother's age at first birth is less than 18 years, otherwise `ecb=0`.
6. Categorize the numeric variable age (`v012`) to some groups using the intervals [15-17), [18-19), [20-290), [30-39), [40-50). Set the new variable's name as `age_category`.
7. Sort the dataset in ascending order of the household number (`v002`)
8. Find the mean, median, mode, range, standard deviation, and IQR of the respondent's age (`v012`)
9. Create the frequency tables of the variables `ecb`, `age_cateogry`, and wealth index (`v190`), and `religion`.
10. Create the bivariate frequency table of the variables Education level (`v106`) and Type of residence (`v025`)

Question 2

- A `starwars` is a tibble in `dplyr` package containing 13 variables about the features of 13 characters in the movie.

```
library(dplyr)
data(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> <"A New Hope", "The Empire Strikes Back", "Return of the J~
#> $ vehicles   <list> <"Snowspeeder", "Imperial Speeder Bike">, <>, <>, <>, "Imp~
#> $ starships  <list> <"X-wing", "Imperial shuttle">, <>, <>, "TIE Advanced x1",~
```

Questions:

1. How many humans are contained in `starwars` overall? (Hint. use `count()`)
2. How many humans are contained in `starwars` by gender?
3. From which homeworld do the most individuals (rows) come from?
4. What is the mean height of all individuals with orange eyes from the most popular homeworld?
5. Compute the median, mean, and standard deviation of height for all droids.

Deadline

- Create a `qmd` file, render it as a PDF, and submit the assignment by 27 January 2025 on Google Classroom.