Big Data 2: Case Study - 1

1. Complete the following tasks using the infix command %>% to connect all tasks in sequence. Show the resulting dataframe after each task (40%):

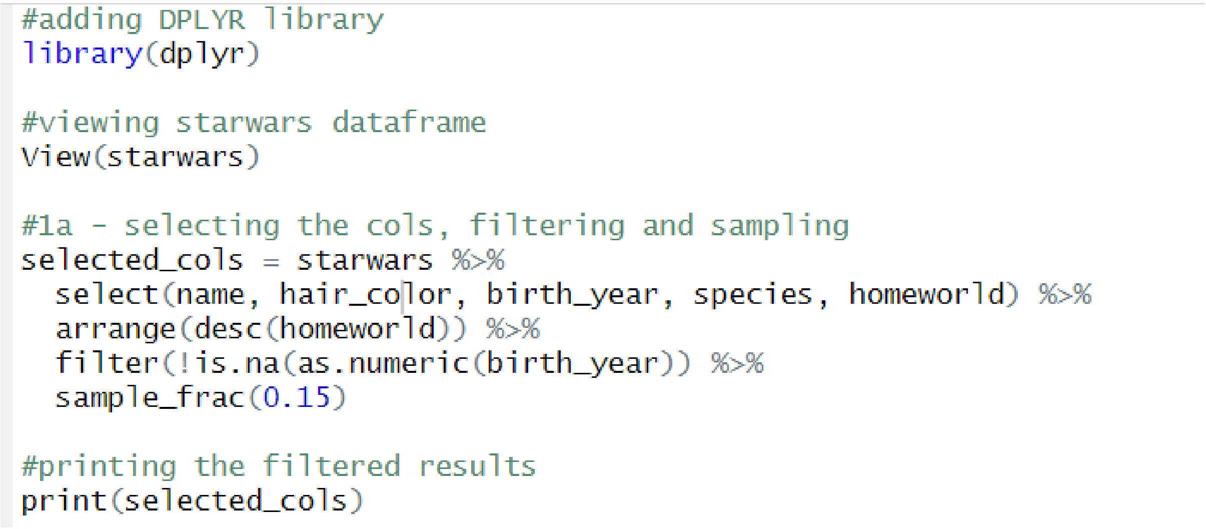
   a) Select the columns (name, hair_color, birth_year, species, and homeworld), then arrange by homeworld in descending order, and filter birth_year that is a number, then sample 15% of the result.

   b) Filter out all the species whose skin color may be grey, show the result.

   Solution:

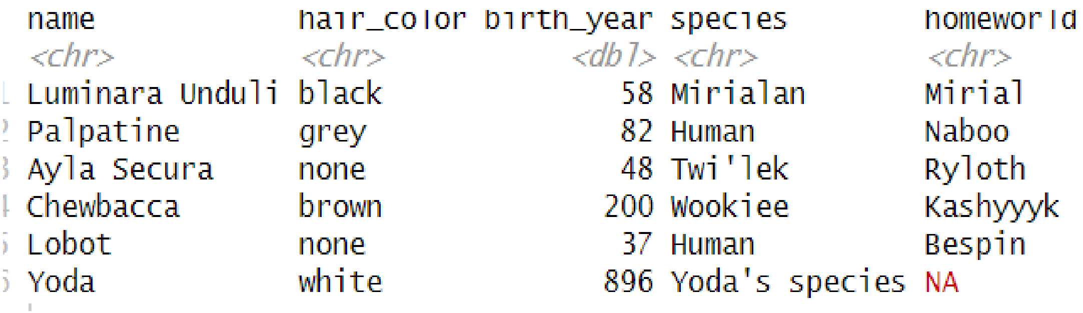
   <u>Source code (1a) -</u>

```
#adding DPLYR library
library(dplyr)

#viewing starwars dataframe
View(starwars)

#1a - selecting the cols, filtering and sampling
selected_cols = starwars %>%
select(name, hair_color, birth_year, species, homeworld) %>%
arrange(desc(homeworld)) %>%
filter(!is.na(as.numeric(birth_year)) %>%
sample_frac(0.15)

#printing the filtered results
print(selected_cols)
```



```
#adding DPLYR library
library(dplyr)

#viewing starwars dataframe
View(starwars)

#1a - selecting the cols, filtering and sampling
selected_cols = starwars %>%
  select(name, hair_color, birth_year, species, homeworld) %>%
  arrange(desc(homeworld)) %>%
  filter(!is.na(as.numeric(birth_year)) %>%
  sample_frac(0.15)

#printing the filtered results
print(selected_cols)
```

Output (1a):



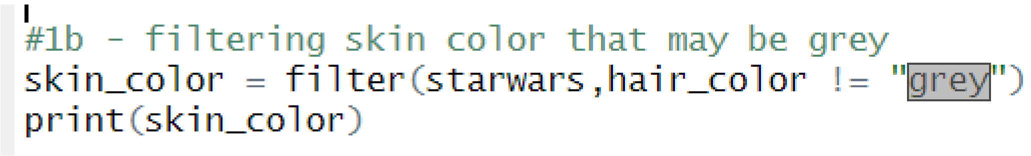
```
  name            hair_color birth_year species        homeworld
  <chr>           <chr>           <dbl> <chr>          <chr>
1 Luminara Unduli black              58 Mirialan       Mirial
2 Palpatine       grey               82 Human          Naboo
3 Ayla Secura     none               48 Twi'lek        Ryloth
4 Chewbacca       brown             200 Wookiee        Kashyyyk
5 Lobot           none               37 Human          Bespin
6 Yoda            white             896 Yoda's species NA
```

Source code (1b) -

#1b - filtering skin color that may be grey

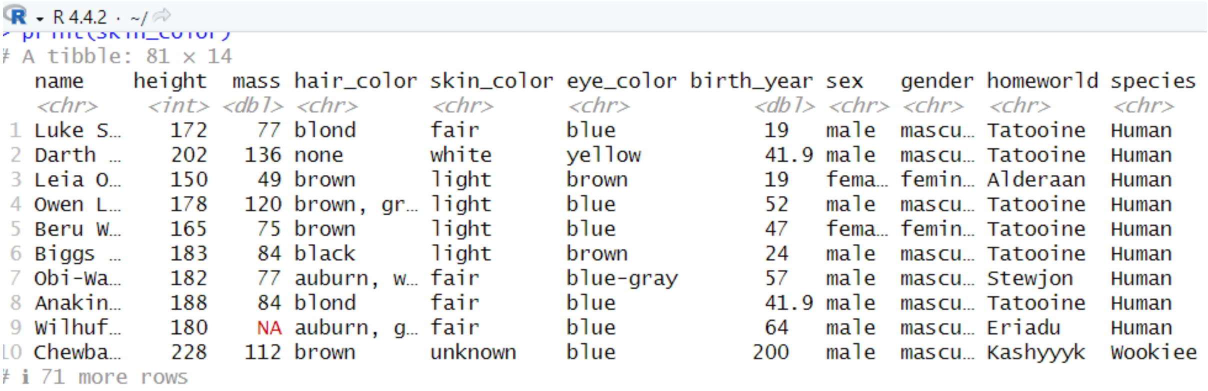
skin_color = filter(starwars,hair_color != "grey")

print(skin_color)



```
#1b - filtering skin color that may be grey
skin_color = filter(starwars,hair_color != "grey")
print(skin_color)
```

Output (1b):



```
R 4.4.2 · ~/
> print(skin_color)
# A tibble: 81 × 14
   name      height  mass hair_color skin_color eye_color birth_year sex    gender homeworld species
   <chr>      <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr>  <chr>  <chr>     <chr>  
 1 Luke S…      172    77 blond      fair       blue            19   male   mascu… Tatooine  Human  
 2 Darth …      202   136 none       white      yellow          41.9 male   mascu… Tatooine  Human  
 3 Leia O…      150    49 brown      light      brown           19   fema…  femin… Alderaan  Human  
 4 Owen L…      178   120 brown, gr… light      blue            52   male   mascu… Tatooine  Human  
 5 Beru W…      165    75 brown      light      blue            47   fema…  femin… Tatooine  Human  
 6 Biggs …      183    84 black      light      brown           24   male   mascu… Tatooine  Human  
 7 Obi-Wa…      182    77 auburn, w… fair       blue-gray       57   male   mascu… Stewjon   Human  
 8 Anakin…      188    84 blond      fair       blue            41.9 male   mascu… Tatooine  Human  
 9 Wilhuf…      180    NA auburn, g… fair       blue            64   male   mascu… Eriadu    Human  
10 Chewba…      228   112 brown      unknown    blue           200   male   mascu… Kashyyyk  Wookiee
# ℹ 71 more rows
```

2. Create and show a dataframe for each of the following (60%):

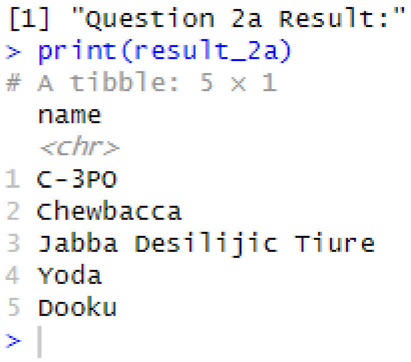
a) List the names (only) for characters whose birth_year > 100
b) List of (unique) films, along with the number of characters appearing in each film

Solution:

Source code (2a) -

```
result_2a <- starwars %>%

  # Filter on the birth_year column having a value > 100

  filter(birth_year > 100) %>%

  #Select specific columns(name)

  select(name)


# Display result

print("Question 2a Result:")

print(result_2a)
```

Output (2a):



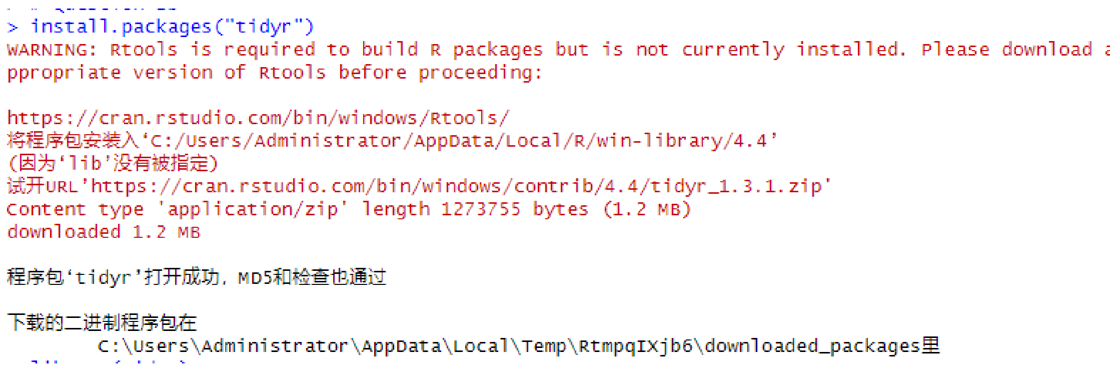
```
[1] "Question 2a Result:"
> print(result_2a)
# A tibble: 5 × 1
  name
  <chr>
1 C-3PO
2 Chewbacca
3 Jabba Desilijic Tiure
4 Yoda
5 Dooku
> 
```

Source code (2b) -

# Question 2b

install.packages("tidyr")

library(tidyr)



```
> install.packages("tidyr")
WARNING: Rtools is required to build R packages but is not currently installed. Please download a
ppropriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
将程序包安装入‘C:/Users/Administrator/AppData/Local/R/win-library/4.4’
(因为‘lib’没有被指定)
试开URL’https://cran.rstudio.com/bin/windows/contrib/4.4/tidyr_1.3.1.zip'
Content type 'application/zip' length 1273755 bytes (1.2 MB)
downloaded 1.2 MB

程序包‘tidyr’打开成功，MD5和检查也通过

下载的二进制程序包在
        C:\Users\Administrator\AppData\Local\Temp\RtmpqIXjb6\downloaded_packages里
```

result_2b <- starwars %>%

# Select the columns

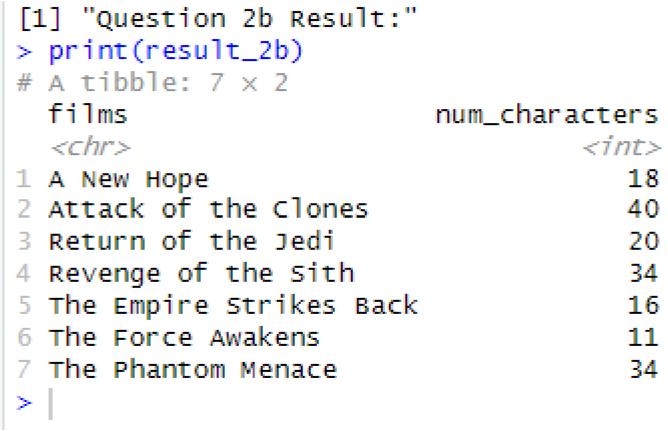
 select(name, films) %>%

# Expand the list of films

 tidyr::unnest(cols = films) %>%

```
# Group by films

  group_by(films) %>%



# Count the number of people in each group

  summarise(num_characters = n()) %>%


 # Sort by films name

  arrange(films)


# Display result

print("Question 2b Result:")

print(result_2b)
```

Output (2b):



```
[1] "Question 2b Result:"
> print(result_2b)
# A tibble: 7 × 2
  films                   num_characters
  <chr>                            <int>
1 A New Hope                          18
2 Attack of the Clones                40
3 Return of the Jedi                  20
4 Revenge of the Sith                 34
5 The Empire Strikes Back             16
6 The Force Awakens                   11
7 The Phantom Menace                  34
> 
```