Task 1 Solution Proposal

Sander Eide Dahling

2023-06-18

Task 1 Solution Proposal

First we need to import all relevant packages to the memory of R. I will make use of the Tidyverse in this task, as it is relatively simple.

```r
# Loading relevant libraries into memory
library(tidyverse)
```

Before loading in the data, you need to make sure that the working directory of the program is the same as the location of the file. I make sure this is the case by running this code:

```r
# Getting the current path of the source file
current_path = dirname(rstudioapi::getActiveDocumentContext()$path)

# Then you are able to set the working directory as this file path
setwd(current_path)
```

In my example, the working directory will be set to **C:/dir/git/nhh-course-help/GUIDES/Data-Science-Academy/R-Training-Tasks/Task1-Top-Unis/[solution]**. This might not look the same to you, especially if you are on a Mac.

With the working directory set up correctly, we can start by loading in the data. The data used for this task is a data set gathered from Kaggle. If you are unable to download the data set from this link, you can find the data in the "data" folder in the Task1 folder.

```r
# Importing the data using the read_csv() function from readr

df <- read_csv("../data/Top-Largest-Universities.csv")

# Objective 1
df_1 <- df %>%
  filter(`Distance / In-Person` == "In-Person")

# Completed Objective 1
head(df_1)
```

```
## # A tibble: 6 x 8
##    Rank Institution            Locat~1 Conti~2 Founded Affil~3 Dista~4 Enrol~5
##   <dbl> <chr>                  <chr>   <chr>   <chr>   <chr>   <chr>     <dbl>
## 1     2 California Community C~ Califo~ North ~ 1967    Public  In-Per~ 2133846
## 2     3 National University, B~ Gazipu~ Asia    1992    Public  In-Per~ 2097182
## 3     6 Islamic Azad University Iran    Asia    1982    Private In-Per~ 1000000
## 4    10 National Technological~ Mexico  North ~ 1948 a~ Public  In-Per~  620000
## 5    11 State University of Ne~ New Yo~ North ~ 1948    Public  In-Per~  606232
```

```
## 6     12  Tribhuvan University    Kirtip~ Asia     1959    Public  In-Per~  604437
## # ... with abbreviated variable names 1: Location, 2: Continent,
## #   3: Affiliation, 4: 'Distance / In-Person', 5: Enrollment
```

```r
# Objective 2

# Here we make use of pipes and dplyr in order to summarise the data frame by Continent, and compute

df_2 <- df_1 %>%
  group_by(Continent) %>%
  summarise(average_enrollment = mean(Enrollment)) %>%
  arrange(desc(average_enrollment))

# Completed Objective 2
df_2
```

```
## # A tibble: 5 x 2
##   Continent      average_enrollment
##   <chr>                       <dbl>
## 1 Asia                      432233.
## 2 North America             324739.
## 3 Africa                    213600
## 4 South America             197179
## 5 Europe                    124482.
```

```r
# Objective 3

# Here we need to re-use some of the code from the previous task and make some changes to the summar

df_3 <- df_1 %>%
  group_by(Continent) %>%
  summarise(average_enrollment = mean(Enrollment),
            number_of_institutions = n()) %>%
  arrange(desc(average_enrollment))

# Completed Objective 3
df_3
```

```
## # A tibble: 5 x 3
##   Continent      average_enrollment number_of_institutions
##   <chr>                       <dbl>                  <int>
## 1 Asia                      432233.                     21
## 2 North America             324739.                     27
## 3 Africa                    213600                       2
## 4 South America             197179                       4
## 5 Europe                    124482.                      5
```

```r
# Objective 4

# Our starting position is once again the data frame df_1, where we need to create a new summarisati

# Creating a copy of the data frame, and add a new column
df_4 <- df_1 %>%
  mutate(PrivateDummy = ifelse(Affiliation == "Private", 1, 0)) %>%
  group_by(Continent) %>%
  summarise(private_percentage = mean(PrivateDummy)) %>%
  arrange(desc(private_percentage))
```

```r
# Completed Objective 4
df_4
```

```
## # A tibble: 5 x 2
##   Continent     private_percentage
##   <chr>                      <dbl>
## 1 South America              0.25
## 2 Asia                       0.143
## 3 North America              0.0370
## 4 Africa                     0
## 5 Europe                     0
```

```r
# Completed Objective 4
df_4
```