CAnD3

Fellow name: Nicole Antunes Rezende

Assignment: Research Replicability and Workflow Management (RRWM)

Research question:

How does the region of origin influence the educational attainment of immigrants, controlling for sex, province of residence, and age at immigration?

COMPLETE CODE

```
# 1. Open the dataset using the package readr


install.packages('readr')

library (readr)


read_csv('pumf-98M0001-E-2016-individuals_F1.csv')


censusdata <- read_csv('pumf-98M0001-E-2016-individuals_F1.csv')


# 2. Select only the variables of our study (PPSORT, HDGREE, AGEIMM, POB, PR, Sex) using the package dplyr


install.packages("dplyr")

library(dplyr)


selected_censusdata <- censusdata |> select(PPSORT, HDGREE, AGEIMM, POB, PR, Sex)
```

```
# (Optional) have an idea of the variables using the function glimpse of the package tibble


install.packages('tibble')

library(tibble)

glimpse (selected_censusdata)


# 3. Recode Sex: Female from 1 to 0  and Male from 2 to 1.

        # 3.1 Transform the variable Sex into a categorical variable using the function factor


selected_censusdata$Sex <- ifelse(selected_censusdata$Sex == 1, 0, 1)


selected_censusdata$Sex <- factor (selected_censusdata$Sex, levels = c(0,1), labels = c('Female', 'Male'))


# 4. Recode POB, creating a new variable 'originregion', knowing that:

1 = "Born in Canada",

2 = "Born in USA",

3 to 6 = "Born in Latin America",

7 to 15 = "Born in Europe",

16 to 18 = "Born in Africa",

19 to 31 = "Born in Asia",

32= "Born in Oceania and others",

88= "Missing value"


table(selected_censusdata$POB)


library(dplyr)
```

```
selected_censusdata <- selected_censusdata |>
  mutate(originregion = case_when (
    POB == 1 ~ 1,
    POB == 2 ~ 2,
    POB >= 3 & POB <= 6 ~ 3,
    POB >= 7 & POB <= 15 ~ 4,
    POB >= 16 & POB <= 18 ~ 5,
    POB >= 19 & POB <= 31 ~ 6,
    POB == 32 ~ 7,
    POB == 88 ~ NA
    ))

selected_censusdata$originregion
table (selected_censusdata$originregion)

        # 4.1 Add the previous labels to the new variable 'originregion':

selected_censusdata$originregion <-  factor (selected_censusdata$originregion,
                                     levels = c(1,2,3,4,5,6,7, NA),
                                     labels = c("Born in Canada",
                                                "Born in USA",
                                                "Born in Latin America",
                                                "Born in Europe",
                                                "Born in Africa",
                                                "Born in Asia",
                                                "Born in Oceania and others"))
```

```
selected_censusdata$originregion

table (selected_censusdata$originregion)


print(selected_censusdata)


# 5. Recode HDGREE creating a new variable 'educlevel', knowing that:


1 = "No certificate, diploma or degree" ///

2 =  "Secondary (high) school diploma or equivalency certificate" ///

3 and 4 = "Trades or Apprenticeship certificate/diploma" ///

5 to 8 = "College, CEGEP or other non-university certificate or diploma;
University certificate or diploma below bachelor level" ///

9 = "Bachelor degree" ///

10 to 13 = "University certificate, diploma or degree above bachelor level"

88 or 99 = NA


table (selected_censusdata$HDGREE)


selected_censusdata <- selected_censusdata |>
  mutate(educlevel = case_when (
    HDGREE == 1 ~ 1,
    HDGREE == 2 ~ 2,
    HDGREE >= 3 & HDGREE <= 4 ~ 3,
    HDGREE >= 5 & HDGREE <= 8 ~ 4,
    HDGREE ==9  ~ 5,
    HDGREE >= 10 & HDGREE <= 13 ~ 6,
    HDGREE == 88 | HDGREE == 99 ~ NA
```

```
  ))


selected_censusdata$educlevel

table (selected_censusdata$educlevel)

# 6. Add the labels to the variable PR, following the information of the Census codebook (pdf) avaible with the Census dataset:


table (selected_censusdata$PR)


selected_censusdata$PR <-  factor (selected_censusdata$PR,
                                   levels = c(10, 11, 12, 13, 24, 35, 46, 47, 48, 59, 70),
                                   labels = c("Newfoundland and Labrador",
                                              "Prince Edward Island",
                                              "Nova Scotia",
                                              "New Brunswick",
                                              "Quebec",
                                              "Ontario",
                                              "Manitoba",
                                              "Saskatchewan",
                                              "Alberta",
                                              "British Columbia",
                                              "Northern Canada"))


selected_censusdata$PR

table (selected_censusdata$PR)


print(selected_censusdata)
```

7. Add the labels to the variable AGEIMM, following the information of the Census codebook (pdf) avaible with the Census dataset:

```
table (selected_censusdata$AGEIMM)

selected_censusdata$AGEIMM <-  factor (selected_censusdata$AGEIMM,
                        levels = c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, NA),
                        labels = c("0 to 4 years",
                                "5 to 9 years",
                                "10 to 14 years",
                                "15 to 19 years",
                                "20 to 24 years",
                                "25 to 29 years",
                                "30 to 34 years",
                                "35 to 39 years",
                                "40 to 44 years",
                                "45 to 49 years",
                                "50 to 54 years",
                                "55 to 59 yearss",
                                "60 years and over"))

selected_censusdata$AGEIMM
table (selected_censusdata$AGEIMM)

print(selected_censusdata)
```

8. Create one frequency table for the variables originregion and PR

```
crosstable1 <-  table(selected_censusdata$originregion, selected_censusdata$PR)

print(crosstable1)


        # 8.1 Export to a csv document using the function 'write.csv'


write.csv(crosstable1, 'output_crosstable1_code_NicoleAR.csv', row.names = T)


# 9. Perform a linear regression, using as dependent variable: educ level. The main independent variable is originregion. Control for Sex, PR and AGEIMM


model <-  lm(educlevel ~ originregion + Sex + PR + AGEIMM, data = selected_censusdata)


summary (model)


        # 9.1 Format the results using the function 'tidy' of the package named 'broom', before exporting it to a cvs document using the function 'write.csv'


install.packages('broom')

library (broom)


regression <-  tidy(model)


write.csv(regression, 'output_regression_code_NicoleaAR.csv', row.names = T)


# 10. Format the two output cvs documents (convert the text of the csv documents into separate columns, using the path: Data --> Text to Columns --> Delimited --> Delimiter: comma --> Finish)
```