ReadMe File

Research question:

Is there any significant association between total income and sex, age at immigration, and level of education of Canadian Immigrants from 2000 to 2015 who immigrated at the age of 20 years and older?

Dataset: Census 2016.csv

Variables:

TotInc => Total Income

Sex => Sex

AGEIMM => Age at immigration

HDGREE => Highest educational qualification

YRIMM => Year of immigration

R code:

##################################### Install these packages ####

install.packages("dplyr")

install.packages("knitr")

install.packages("stargazer")

install.packages("gt")

install.packages("kableExtra")

# Load necessary libraries

library(dplyr) # for data cleaning

library(knitr) # For output tables formation

library(gt) # For regression

library(kableExtra) # For exporting to .html

############################################## clean up data ####

# Load the CSV file and process the data

cd <- read.csv("Census 2016.csv")%>%

select(TotInc, Sex, AGEIMM, HDGREE, YRIMM) %>% # Select specified variables

filter(AGEIMM >= 5 & AGEIMM <= 13, # Filter AGEIMM range 5 to 13

YRIMM >= 2000 & YRIMM <= 2015, # Filter YRIMM range 2000 to 2015

!(TotInc %in% c(33, 88888888, 99999999)), # Remove TotInc = 33, 88888888, 99999999

TotInc >= 1000, # Remove cases with TOTINC less than 1000

!(HDGREE %in% c(88, 99))) # Remove cases with HDGREE = 88 or 99

glimpse(cd)

print(cd)

###################################### re-code sex and education ####

# re-code variables

cd[] <- lapply(cd, as.numeric) #convert all variables to numeric

cd <- cd %>%

mutate(Sex = case\_when( # Recoding Sex variable

Sex == 1 ~ "Female",

Sex == 2 ~ "Male"),

Sex = factor(Sex, levels = c("Female", "Male")))

cd <- cd %>%

mutate(HDGREE.R = case\_when( # Creating HDGREE.R variable

HDGREE >= 1 & HDGREE <= 8 ~ 1, # 1 to 8 = below bachelor

HDGREE == 9 ~ 2, # 9 = bachelor

HDGREE >= 10 & HDGREE <= 13 ~ 3), # 10 to 13 = above bachelor

HDGREE.R = factor(HDGREE.R, # Adding labels to HDGREE.R

levels = c(1, 2, 3), # Defining levels for the factor

labels = c("Below Bachelor", # Adding labels for each level

"Bachelor",

"Above Bachelor")))

print(cd)

################################## Descriptive statistics ####

# Summary statistics for Sex

sex\_summary <- cd %>%

group\_by(Sex) %>%

summarise(

Frequency = n()

) %>%

mutate(Percentage = round((Frequency / sum(Frequency)) \* 100, 2))

# Adding a row for total

sex\_summary <- sex\_summary %>%

bind\_rows(summarise(., Sex = "Total", Frequency = sum(Frequency), Percentage = sum(Percentage)))

# Summary statistics for HDGREE.R

hdgree\_summary <- cd %>%

group\_by(HDGREE.R) %>%

summarise(

Frequency = n()

) %>%

mutate(Percentage = round((Frequency / sum(Frequency)) \* 100, 2))

# Adding a row for total

hdgree\_summary <- hdgree\_summary %>%

bind\_rows(summarise(., HDGREE.R = "Total", Frequency = sum(Frequency), Percentage = sum(Percentage)))

# Print the summary tables

kable(sex\_summary, caption = "Summary Statistics for Sex (Categorical Variable)")

kable(hdgree\_summary, caption = "Summary Statistics for Highest Degree (Categorical Variable)")

########################## Multivariate regression analysis ####

# Fit the linear model

regression\_table\_gt <- regression\_table %>%

gt() %>%

tab\_header(

title = "Regression Results",

subtitle = "Total Income as Dependent Variable"

) %>%

fmt\_number(

columns = c(estimate, std.error, statistic, p.value), # Use c() instead of vars()

decimals = 3

) %>%

cols\_label(

term = "Term",

estimate = "Estimate",

std.error = "Std. Error",

statistic = "t-value",

p.value = "p-value"

) %>%

tab\_footnote(

footnote = "Regression coefficients with standard errors.",

locations = cells\_column\_labels(columns = c(estimate)) # Use c() instead of vars()

)

# Display the formatted table

regression\_table\_gt

################## Save all tables into one HTML file ####

# Create HTML content with kable and gt

html\_content <- "

<html>

<head>

<title>Combined Summary Tables</title>

</head>

<body>

<h1>Summary Statistics for Sex (Categorical Variable)</h1>

{sex\_table}

<h1>Summary Statistics for Highest Degree (Categorical Variable)</h1>

{degree\_table}

<h1>Regression Table</h1>

{regression\_table}

</body>

</html>

"

# Render the kable tables as HTML

sex\_table <- sex\_summary %>%

kable("html", caption = "Summary Statistics for Sex (Categorical Variable)") %>%

kable\_styling(bootstrap\_options = c("striped", "hover", "condensed", "responsive")) %>%

as.character()

hdgree\_table <- hdgree\_summary %>%

kable("html", caption = "Summary Statistics for Highest Degree (Categorical Variable)") %>%

kable\_styling(bootstrap\_options = c("striped", "hover", "condensed", "responsive")) %>%

as.character()

# Convert gt table to HTML

regression\_table\_html <- regression\_table\_gt %>%

as\_raw\_html()

# Replace placeholders in HTML content

html\_content <- gsub("\\{sex\_table\\}", sex\_table, html\_content)

html\_content <- gsub("\\{degree\_table\\}", hdgree\_table, html\_content)

html\_content <- gsub("\\{regression\_table\\}", regression\_table\_html, html\_content)

# Write the combined content to a file

writeLines(html\_content, "combined\_summary.html")

# Open the HTML file in the Viewer

rstudioapi::viewer("combined\_summary.html")

#################################### End ##################