R Code & Output\_RRWM Activity\_Genan

Genan Hamad

2025-10-02

# Project title: RRWM Data Activity CAnD3

## Owner: Genan Hamad

## Date: Oct 2, 2025

Citation of Data: Statistics Canada. 2019. Census of Population, 2016 [Canada] Public Use Microdata File (PUMF): Individuals File. Statistics Canada [producer and distributor], accessed September 10, 2021. ID: pumf-98M0001-E-2016-individuals. The following tasks are included in this script:  
Create a subset of data from the Education and Income variables Recode the Education variable Generate a descriptive statistics table for the new subset of data Run a simple linear regression

# Import and Read CSV data file using base R  
housing\_data <- read.csv("housing\_data\_pumf-98M0001-E-2016-individuals\_F1.csv")

# Preview data  
head(housing\_data)

## PPSORT WEIGHT WT1 WT2 WT3 WT4 WT5 WT6 WT7 WT8 WT9 WT10 WT11 WT12  
## 1 453141 37.03728 0 0 0 0 0 0 0 0.0000 0 0 0 0  
## 2 923226 37.03728 0 0 0 0 0 0 0 0.0000 0 0 0 0  
## 3 385097 37.03728 0 0 0 0 0 0 0 0.0000 0 0 0 0  
## 4 732612 37.03728 0 0 0 0 0 0 0 592.5964 0 0 0 0  
## 5 143665 37.12091 0 0 0 0 0 0 0 0.0000 0 0 0 0  
## 6 459269 37.03728 0 0 0 0 0 0 0 592.5964 0 0 0 0  
## WT13 WT14 WT15 WT16 ABOID AGEGRP AGEIMM ATTSCH BedRm BFNMEMB CapGn  
## 1 0.0000 0 0 592.5964 6 11 99 1 5 0 99999999  
## 2 0.0000 0 0 592.5964 6 5 99 9 5 0 99999999  
## 3 0.0000 0 0 592.5964 6 2 99 9 5 0 99999999  
## 4 0.0000 0 0 0.0000 6 12 99 1 4 0 99999999  
## 5 593.9346 0 0 0.0000 6 15 99 1 2 0 99999999  
## 6 0.0000 0 0 0.0000 6 19 99 1 3 0 -3300  
## CFInc CFInc\_AT CfSize CFSTAT CHDBN ChldC CIP2011 CIP2011\_STEM\_SUM  
## 1 30 27 4 2 99999999 2500 8 10  
## 2 30 27 4 4 99999999 99999999 99 99  
## 3 30 27 4 4 99999999 99999999 99 99  
## 4 20 19 4 2 4400 99999999 6 1  
## 5 16 15 2 1 99999999 99999999 5 4  
## 6 28 26 2 1 99999999 99999999 13 11  
## Citizen CitOth CMA CONDO COW CQPPB DETH123 DIST DPGRSUM DTYPE EFDecile  
## 1 1 99 505 0 1 99999999 1 9 1 1 8  
## 2 1 99 505 0 9 99999999 2 9 1 1 8  
## 3 1 99 505 0 9 99999999 2 9 1 1 8  
## 4 1 99 999 0 1 99999999 2 4 1 1 4  
## 5 1 99 999 0 1 99999999 1 6 1 1 4  
## 6 1 99 999 0 9 8300 2 9 1 1 9  
## EfDIMBM EFInc EFInc\_AT EfSize EICBN EmpIn ETHDER FOL FPTWK GENSTAT  
## 1 26 30 27 4 99999999 95000 5 1 1 4  
## 2 26 30 27 4 99999999 99999999 36 1 9 4  
## 3 26 30 27 4 99999999 99999999 36 1 9 4  
## 4 16 20 19 4 99999999 40000 35 1 1 4  
## 5 13 16 15 2 99999999 29000 8 2 1 4  
## 6 26 28 26 2 99999999 99999999 38 1 9 4  
## GovtI GTRfs HCORENEED\_IND HDGREE HHInc HHInc\_AT HHMRKINC HHSIZE HHTYPE  
## 1 99999999 99999999 0 4 30 27 30 4 2  
## 2 99999999 99999999 0 99 30 27 30 4 2  
## 3 99999999 99999999 0 99 30 27 30 4 2  
## 4 300 4700 0 9 20 19 19 4 2  
## 5 99999999 99999999 0 6 16 15 14 2 1  
## 6 99999999 15100 0 2 28 26 25 2 1  
## HLAEN HLAFR HLANO HLBEN HLBFR HLBNO IMMCAT5 IMMSTAT IncTax Invst KOL  
## 1 1 0 1 0 0 0 1 1 24000 99999999 1  
## 2 1 0 1 0 0 0 1 1 99999999 99999999 3  
## 3 1 0 1 0 0 0 1 1 99999999 99999999 3  
## 4 1 0 1 0 0 0 1 1 4000 99999999 3  
## 5 1 0 1 0 1 0 1 1 4000 99999999 3  
## 6 1 0 1 0 0 0 1 1 14000 3800 1  
## LFACT LICO LICO\_AT LOC\_ST\_RES LOCSTUD LoLIMA LoLIMB LoMBM LSTWRK LWAEN LWAFR  
## 1 1 1 1 1 6 1 1 1 3 1 0  
## 2 99 1 1 9 99 1 1 1 9 9 9  
## 3 99 1 1 9 99 1 1 1 9 9 9  
## 4 1 1 1 1 6 1 1 1 3 1 0  
## 5 1 1 1 1 2 1 1 1 3 1 0  
## 6 13 1 1 4 99 1 1 1 1 9 9  
## LWANO LWBEN LWBFR LWBNO MarStH MOB1 Mob5 MODE MrkInc MTNEn MTNFr MTNNO  
## 1 0 0 0 0 2 1 2 2 97000 1 0 1  
## 2 9 9 9 9 1 1 2 9 99999999 1 0 1  
## 3 9 9 9 9 1 1 2 9 99999999 1 0 1  
## 4 0 0 0 0 3 8 1 2 41000 1 0 1  
## 5 0 0 0 0 3 1 1 2 30000 0 1 1  
## 6 9 9 9 9 2 1 1 9 67000 1 0 1  
## NAICS NOC16 NOCS NOL NOS OASGI OtInc PKID0\_1 PKID15\_24 PKID2\_5 PKID25  
## 1 12 24 8 1 1 99999999 1000 0 0 0 0  
## 2 99 99 99 1 1 99999999 99999999 0 0 0 0  
## 3 99 99 99 1 1 99999999 99999999 0 0 0 0  
## 4 9 8 2 6 1 99999999 99999999 0 1 0 0  
## 5 14 6 2 1 1 99999999 99999999 0 0 0 0  
## 6 99 99 99 1 1 6700 99999999 0 0 0 0  
## PKID6\_14 PKIDS POB POBF POBM POWST PR PR1 PR5 PresMortG PRIHM PWDUR PWLEAVE  
## 1 1 1 1 1 1 2 35 35 35 1 1 2 6  
## 2 1 1 1 1 1 9 35 35 35 1 0 9 9  
## 3 1 1 1 1 1 9 35 35 35 1 0 9 9  
## 4 0 1 1 1 1 4 35 88 35 1 0 2 5  
## 5 0 0 1 1 1 5 11 11 11 1 1 2 1  
## 6 0 0 1 1 1 9 35 35 35 0 1 9 9  
## PWOCC PWPR REGIND REPAIR Retir ROOMS SempI Sex SHELCO SSGRAD Subsidy  
## 1 1 99 0 1 99999999 11 99999999 2 2500 5 9  
## 2 9 99 0 1 99999999 11 99999999 2 2400 99 9  
## 3 9 99 0 1 99999999 11 99999999 1 2500 99 9  
## 4 1 35 0 2 99999999 8 21000 1 1300 8 9  
## 5 1 11 0 3 99999999 5 99999999 1 700 6 9  
## 6 9 99 0 2 63700 7 99999999 2 1100 4 9  
## Tenur TotInc TotInc\_AT VALUE VisMin Wages WKSWRK WRKACT YRIMM  
## 1 1 97000 73000 450000 13 95000 6 11 9999  
## 2 1 99999999 99999999 440000 13 99999999 9 99 9999  
## 3 1 99999999 99999999 440000 13 99999999 9 99 9999  
## 4 1 46000 41000 839779 13 19000 6 11 9999  
## 5 1 30000 26000 60000 13 29000 5 9 9999  
## 6 1 82000 69000 839779 13 99999999 9 1 9999

## Variable selection

# Create a new dataset with only Wages (income variable) and HDGREE (education variable) columns  
data2 <- housing\_data[c("Wages", "HDGREE")]  
  
# View first few rows of data2  
head(data2)

## Wages HDGREE  
## 1 95000 4  
## 2 99999999 99  
## 3 99999999 99  
## 4 19000 9  
## 5 29000 6  
## 6 99999999 2

## recoding the HDGREE (education variable)

# recode the HDGREE (education variable), where as the values (1, 2, 3, 4, 5, 6, 7)   
# are coded as "low\_edu" and the values (8, 9, 10, 11, 12, 13) are coded as "high\_edu"  
  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

data2 <- data2 %>%  
 mutate(HDGREE = case\_when(  
 HDGREE %in% 1:7 ~ "low\_edu",  
 HDGREE %in% 8:13 ~ "high\_edu",  
 TRUE ~ NA\_character\_ # for unexpected/missing values  
 ))  
  
# show summary of data2  
summary(data2)

## Wages HDGREE   
## Min. : 1 Length:930421   
## 1st Qu.: 32000 Class :character   
## Median : 110000 Mode :character   
## Mean :45959314   
## 3rd Qu.:99999999   
## Max. :99999999

# Preview data2  
head(data2)

## Wages HDGREE  
## 1 95000 low\_edu  
## 2 99999999 <NA>  
## 3 99999999 <NA>  
## 4 19000 high\_edu  
## 5 29000 low\_edu  
## 6 99999999 low\_edu

## Descriptive Statistics Table

# generate a descriptive statistics table for data2   
# data2 contains:  
# "Wages": numeric variable and "HDGREE": categorical variable ("low\_edu" and "high\_edu")  
  
# Install required packages  
install.packages("summarytools", repos = "https://cloud.r-project.org")

## Installing package into 'C:/Users/saleh/AppData/Local/R/win-library/4.5'  
## (as 'lib' is unspecified)

## package 'summarytools' successfully unpacked and MD5 sums checked  
##   
## The downloaded binary packages are in  
## C:\Users\saleh\AppData\Local\Temp\RtmpCiDAG8\downloaded\_packages

install.packages("officer", repos = "https://cloud.r-project.org")

## Installing package into 'C:/Users/saleh/AppData/Local/R/win-library/4.5'  
## (as 'lib' is unspecified)

## package 'officer' successfully unpacked and MD5 sums checked  
##   
## The downloaded binary packages are in  
## C:\Users\saleh\AppData\Local\Temp\RtmpCiDAG8\downloaded\_packages

# Load them  
library(summarytools)  
library(officer)  
  
# Create a discriptive statistics table for data2, using dfSummary()   
desc\_table <- dfSummary(data2)  
dfSummary(data2)

## Data Frame Summary   
## data2   
## Dimensions: 930421 x 2   
## Duplicates: 929827   
##   
## ---------------------------------------------------------------------------------------------------------------------  
## No Variable Stats / Values Freqs (% of Valid) Graph Valid Missing   
## ---- ------------- --------------------------------- --------------------- --------------------- ---------- ---------  
## 1 Wages Mean (sd) : 45959314 (49751667) 216 distinct values : 930421 0   
## [integer] min < med < max: : : (100.0%) (0.0%)   
## 1 < 110000 < 1e+08 : :   
## IQR (CV) : 99967999 (1.1) : :   
## : :   
##   
## 2 HDGREE 1. high\_edu 197119 (25.8%) IIIII 764686 165735   
## [character] 2. low\_edu 567567 (74.2%) IIIIIIIIIIIIII (82.2%) (17.8%)   
## ---------------------------------------------------------------------------------------------------------------------

## Regression Analysis

# perform a simple linear regression using data2, where:  
# Outcome variable: Wages (numeric)  
# Predictor: HDGREE (categorical: "low\_edu", "high\_edu")  
  
# Use HDGREE as a factor  
data2$HDGREE <- factor(data2$HDGREE, levels = c("low\_edu", "high\_edu"))  
  
# Run a simple linear regression analysis   
model <- lm(Wages ~ HDGREE, data = data2)  
  
# View the regression model summary  
summary(model)

##   
## Call:  
## lm(formula = Wages ~ HDGREE, data = data2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -38087492 -38053493 -26036393 61912506 73945606   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 38087493 62791 606.6 <2e-16 \*\*\*  
## HDGREEhigh\_edu -12033100 123673 -97.3 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 47300000 on 764684 degrees of freedom  
## (165735 observations deleted due to missingness)  
## Multiple R-squared: 0.01223, Adjusted R-squared: 0.01223   
## F-statistic: 9467 on 1 and 764684 DF, p-value: < 2.2e-16