UP431 Lab2: Exploring NHTS Data (2)

Using NHTS Data (2)

This week, we will explore household data and vehicle data. Get ready with `tidyverse`, `haven` and NHTS SPSS dataset.

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------------
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## Warning: package 'readr' was built under R version 4.0.3
```

```
## -- Conflicts ------------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(haven)
```

```
## Warning: package 'haven' was built under R version 4.0.3
```

Using Household Data

```
hh <- read_sav("C:/Lab0/2021_UP431/Lab1/Data/spss (2)/hhpub.sav")
```

Begin with exploring what variables household data contains.

```
hh <- as_factor(hh)
names(hh)
```

```
##  [1] "houseid"    "travday"    "sampstrat"  "homeown"    "hhsize"
##  [6] "hhvehcnt"   "hhfaminc"   "pc"         "sphone"     "tab"
## [11] "walk"       "bike"       "car"        "taxi"       "bus"
## [16] "train"      "para"       "price"      "place"      "walk2save"
## [21] "bike2save"  "ptrans"     "hhrelatd"   "drvrcnt"    "cnttdhh"
## [26] "hhstate"    "hhstfips"   "numadlt"    "youngchild" "wrkcount"
```

```
## [31] "tdaydate"   "hhresp"     "lif_cyc"    "msacat"     "msasize"
## [36] "rail"       "urban"      "urbansize"  "urbrur"     "scresp"
## [41] "census_d"   "census_r"   "cdivmsar"   "hh_race"    "hh_hisp"
## [46] "hh_cbsa"    "resp_cnt"   "webuse17"   "smplsrce"   "wthhfin"
## [51] "hbhur"      "hthtnrnt"   "htppopdn"   "htresdn"    "hteempdn"
## [56] "hbhtnrnt"   "hbppopdn"   "hbresdn"
```

Filter rows that are from `Chicago-Naperville-Elgin, IL-IN-WI` CBSA.

```
chi_hh <- hh %>% filter(hh_cbsa == "Chicago-Naperville-Elgin, IL-IN-WI")
```

Task 1

What do you think would be the relationship between car ownership and income group? What variable would you use to analyze the relationship?

Like lab 1, use `levels` and `fct_collapse` to see how income group is recorded in the dataset and customize it.

```
levels(chi_hh$hhfaminc)
```

```
##  [1] "I prefer not to answer" "I don't know"           "Not ascertained"
##  [4] "Less than $10,000"      "$10,000 to $14,999"     "$15,000 to $24,999"
##  [7] "$25,000 to $34,999"     "$35,000 to $49,999"     "$50,000 to $74,999"
## [10] "$75,000 to $99,999"     "$100,000 to $124,999"   "$125,000 to $149,999"
## [13] "$150,000 to $199,999"   "$200,000 or more"
```

```
# Your code comes here
```

a) Calculate the percentage of household with no vehicle for each income group. Don't forget that every analysis should consider weight!

```
# Your code comes here
```

Is there any trend?

b) Calculate the mean of vehicle number for each income group. You can use `weighted.mean` to easily calculate the weighted mean.

```
# Your code comes here
```

c) Calculate the mean number of vehicle to driver ratio (HHVEHCNT/WRKCOUNT), for each income group.

```
# Your code comes here
```

Why can't we calculate the ratio with the previous code?

```
# Your code comes here
```

d) Visualze Task 1 (a) into a bar graph using `geom_bar`.

```
# Your code comes here
```

d-1) Crosstab: bivariate frequency tables with percent numbers. Just another way of doing a similar task.

```
#install.packages("pollster")  #a package for survey analysis
library(pollster)


crosstab(df = chi_hh,
         x = hhincome_short,
         y = hhvehcnt,
         weight = wthhfin)

crosstab(
  df = chi_hh,
  x = hhincome_short,
  y = hhvehcnt,
  weight = wthhfin,
  format = "long"
) %>%
  ggplot(aes(hhincome_short, pct, fill = hhvehcnt)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "Income Group", y = "People with no vehicle(%)", title = "Population with No Vehicle by Incor
```

#Using Vehicle Data

Import vehicle data. Explore the variables and filter rows from Chicago-Naperville-Elgin, IL-IN-WI.

```
veh <- read_sav("C:/Lab0/2021_UP431/Lab1/Data/spss (2)/vehpub.sav") # your path
veh <- as_factor(veh)
names(veh)
chi_veh <- veh %>% filter(HH_CBSA == "Chicago-Naperville-Elgin, IL-IN-WI")
```

Vehicle data consists one row for *each* vehicle. It means that a household with three vehicles will have three rows in the vehicle data.View the dataset and check `HOUSEID` to see what it means!

Task 2

Assume that you need a VMT value in a household level. You would need to aggregate the BESTMILE variable in the vehicle file to a household level using HOUSEID.Before that, make sure that there are no non-numeric values in BESTMILE.

```
chi_veh <- chi_veh %>%
  mutate(BESTMILE_new = fct_collapse(BESTMILE, Missing = c("Not ascertained"))) %>%
  filter(BESTMILE != "Missing")
```

a) Use `aggregate` function. Let's leave HHFAMINC,WTHHFIN for the next task, and also HHSIZE, HBHUR, HHVEHCNT, DRVRCNT for linear regression in the last task.Name the aggregated VMT as `HHVMT`.

```
# Your code comes here
```

b) Caculate the mean annual household VMT by four income group made in task 1.

```
# Your code comes here
```

```
# Your code comes here
```

c) Make an ANOVA test to check the relationship between annual household VMT and income group.

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.0.3
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.0.3
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
# Your code comes here
```

Task 3

Lastly, conduct a linear regression analysis at the household level.

HHVMT = f(HHSIZE, HHVEHCNT/WRKCOUNT, INCGROUP –> dummy, HBHUR –> dummy)

a) Create dummy variables (https://www.marsja.se/create-dummy-variables-in-r/). You can either use a libary to make it fast, or you can make dummy variables by yourself using `ifelse`.

```
#install.packages("fastDummies")
library(fastDummies)
```

```
## Warning: package 'fastDummies' was built under R version 4.0.3
```

If you want to use `fastDummies`, rename any variable with an empty space(' ') in the name, since you will not be able to access the variable by name if there is a space.

```
agg_veh <- agg_veh %>%
  mutate(HBHUR_new = fct_collapse(
    HBHUR,
    "SmallTown" = c("Small Town"),
    "SecondCity" = c("Second City"),
    "Missing" = c("Not ascertained")
  )) %>%
  filter(HBHUR_new != "Missing")
```

```
# Your code comes here
agg_veh <- dummy_cols(agg_veh, select_columns = c("HHINCOME_SHORT","HBHUR_new"))
#agg_veh
```

a-2) IF you want to use `ifelse`, here is an example.

```
#exampleData$Var_A <- ifelse(exampleData$Var == 'A', 1, 0)
#exampleData$Var_B <- ifelse(exampleData$Var == 'B', 1, 0)
```

b) Run a regression and print the result using `summary` function.

```
# Your code comes here
```

Do you see any significant relationship?

c) What additional variables do you want to include in the regression?

```
# Your code comes here
```