UP431 Lab2: Exploring NHTS Data (2) Home Activity

Using NHTS Data (2)

We will continue to use the Chicago metropolitan area data. You should get started by importing the data you have saved last week.Reading the RDS file back in is as simple as assigning the output of read_rds("filename") to a new object. Since read_rds function is coming from the tidyverse package, you should import the package before you start.

```
library(tidyverse)
chi_trips <- read_rds("C:/Lab0/2021_UP431/Lab1/Output/chi_trips.rds") # your path
```

If the data import was successful, you will see chi_trips on the Environment tab (typically on the right side of the source tab).

Check the column names using `names` function.

```
names(chi_trips)
```

```
##   [1] "HOUSEID"           "PERSONID"          "TDTRPNUM"
##   [4] "STRTTIME"          "ENDTIME"           "TRVLCMIN"
##   [7] "TRPMILES"          "TRPTRANS"          "TRPACCMP"
##  [10] "TRPHHACC"          "VEHID"             "TRWAITTM"
##  [13] "NUMTRANS"          "TRACCTM"           "DROP_PRK"
##  [16] "TREGRTM"           "WHODROVE"          "WHYFROM"
##  [19] "LOOP_TRIP"         "TRPHHVEH"          "HHMEMDRV"
##  [22] "HH_ONTD"           "NONHHCNT"          "NUMONTRP"
##  [25] "PSGR_FLG"          "PUBTRANS"          "TRIPPURP"
##  [28] "DWELTIME"          "TDWKND"            "VMT_MILE"
##  [31] "DRVR_FLG"          "WHYTRP1S"          "ONTD_P1"
##  [34] "ONTD_P2"           "ONTD_P3"           "ONTD_P4"
##  [37] "ONTD_P5"           "ONTD_P6"           "ONTD_P7"
##  [40] "ONTD_P8"           "ONTD_P9"           "ONTD_P10"
##  [43] "ONTD_P11"          "ONTD_P12"          "ONTD_P13"
##  [46] "TDCASEID"          "TRACC_WLK"         "TRACC_POV"
##  [49] "TRACC_BUS"         "TRACC_CRL"         "TRACC_SUB"
##  [52] "TRACC_OTH"         "TREGR_WLK"         "TREGR_POV"
##  [55] "TREGR_BUS"         "TREGR_CRL"         "TREGR_SUB"
##  [58] "TREGR_OTH"         "WHYTO"             "TRAVDAY"
##  [61] "HOMEOWN"           "HHSIZE"            "HHVEHCNT"
##  [64] "HHFAMINC"          "DRVRCNT"           "HHSTATE"
##  [67] "HHSTFIPS"          "NUMADLT"           "WRKCOUNT"
##  [70] "TDAYDATE"          "HHRESP"            "LIF_CYC"
##  [73] "MSACAT"            "MSASIZE"           "RAIL"
##  [76] "URBAN"             "URBANSIZE"         "URBRUR"
##  [79] "GASPRICE"          "CENSUS_D"          "CENSUS_R"
```

```
##  [82] "CDIVMSAR"           "HH_RACE"            "HH_HISP"
##  [85] "HH_CBSA"            "SMPLSRCE"           "R_AGE"
##  [88] "EDUC"               "R_SEX"              "PRMACT"
##  [91] "PROXY"              "WORKER"             "DRIVER"
##  [94] "WTTRDFIN"           "WHYTRP90"           "TRPMILAD"
##  [97] "R_AGE_IMP"          "R_SEX_IMP"          "VEHTYPE"
## [100] "OBHUR"              "DBHUR"              "OTHTNRNT"
## [103] "OTPPOPDN"           "OTRESDN"            "OTEEMPDN"
## [106] "OBHTNRNT"           "OBPPOPDN"           "OBRESDN"
## [109] "DTHTNRNT"           "DTPPOPDN"           "DTRESDN"
## [112] "DTEEMPDN"           "DBHTNRNT"           "DBPPOPDN"
## [115] "DBRESDN"            "mode_short"         "mode_short_carpool"
```

Task 1 Aggregate the mode choice by income group.

Which variable describes the income group?

```
# check levels
levels(chi_trips$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"
```

Simplify income group to lower (< 50% of median income), moderate (50% >= and < 80%), middle (80% - 12%), and upper (>= 120%) income groups. Use 2017 median income, which is about $61,500.

```
# collapse columns into 3 levels
# <50% (<30750), 50%<=<80% (30750-49200), 80%-120% (49200-73800), 120%< (73800)
# filter missing
mode_income <- chi_trips %>%
  mutate(
    hhincome_short = fct_collapse(
      HHFAMINC,
      "Lower" = c(
        "Less than $10,000",
        "$10,000 to $14,999",
        "$15,000 to $24,999",
        "$25,000 to $34,999"
      ),
      "Moderate" = c("$35,000 to $49,999"),
      "Middle" = c("$50,000 to $74,999"),
      "Upper" = c(
        "$75,000 to $99,999",
        "$100,000 to $124,999",
        "$125,000 to $149,999",
        "$150,000 to $199,999",
        "$200,000 or more"
      ),
      Missing = c("I prefer not to answer", "I don't know", "Not ascertained")
    )
  ) %>% filter(hhincome_short != "Missing")
```

```
levels(mode_income$hhincome_short)
```

```
## [1] "Missing"  "Lower"    "Moderate" "Middle"   "Upper"
```

Now you are ready to aggregate mode choice by income group.

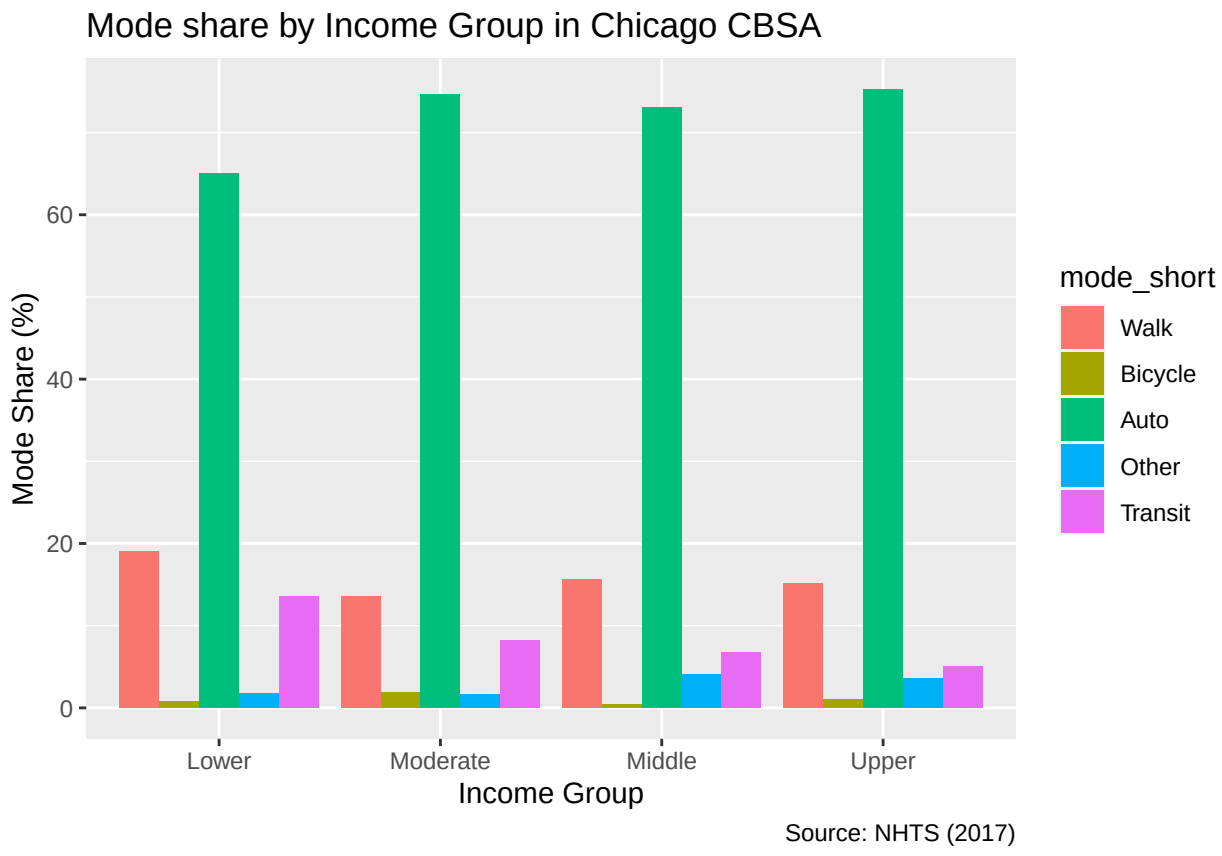
```
mode_by_income <- mode_income %>% count(mode_short, hhincome_short, wt=WTTRDFIN) %>%
  group_by(hhincome_short) %>%
  mutate(per = prop.table(n)*100) # make a new column
```

Visualize the data and elaborate your observation.

1) Use geom_bar position='dodge' option.

```
# Your code comes here
ggplot(mode_by_income, aes(hhincome_short, per)) +
  geom_col(aes(fill = mode_short), position = "dodge", stat = 'identity') +
  labs(x = "Income Group", y = "Mode Share (%)", title = "Mode share by Income Group in Chicago CBSA",
       caption = "Source: NHTS (2017)", fill= "mode_short")
```

```
## Warning: Ignoring unknown parameters: stat
```



2) Use `facet_wrap` to create separate graph for each income group.

```
ggplot(mode_by_income, aes(mode_short, per)) +
  geom_bar(aes(fill = mode_short), stat = 'identity') +
  facet_wrap( ~hhincome_short) +
  labs(x = "Trip Mode", y = "Mode Share (%)", title = "Mode share by Income Group in Chicago CBSA",
       caption = "Source: NHTS (2017)", fill= "mode_short") +
  guides(fill=guide_legend(title="Mode"))
```