ESP106 Final Project

Naomi

2023-03-06

ESP 106 Final Project

BACKGROUND

The relationship between built environment and travel behavior is an ever-evolving topic to explore. Whether the the availability and quality of transportation infrastructure shape people's mode choice has still been questionable. This study aims to observe how frequent households in metropolitan area with rail infrastructure travel by train in daily basis. Furthermore, it will also look at the variance of train trip frequencies based on the household income.

Numerous studies indicated that socio-demographics are important variables in exploring trips generation (Mukherjee & Raguram (2022), Currans, et. al. (2020), Handy (2015)). Boarnet et.al. (2020) found that train trips are taken more by Californian households with higher income (more than $50,000) in a neighborhood that is dense in terms of jobs and population. It came to a suggestion to provide housing for higher income households nearby the train stations. Wang & Woo (2017) also suggested an investment in suburban Transit-Rich Neighborhood that has diverse groups of income.

METHOD

Therefore, this study explores the data set of National Household Travel Survey (NHTS) 2017 to see the relationship between household income and daily train trips made, in national and California-state level. It also utilizes routes and stops data from Sacramento Regional Transit (SacRT) Light Rail service to see what income level group live near train station, based on this regional case study.

National Household Travel Survey

National Household Travel Survey 2017 is the latest travel diary survey that describe person-level and household-level travel behavior and choices in a national scale. For this study, I utilize Data Summary of NHTS 2017 on GitHub that is made publicly by Westat Center for Transportation, Technology & Safety Research (2016/2021).

Please note that to access the summarize NHTS data from Westat GitHub repository, R package of devtools is needed to be installed first, then "summarizeNHTS" package could be installed from their GitHub.

```r
#install.packages('devtools')
#devtools::install_github('Westat-Transportation/summarizeNHTS')
library(summarizeNHTS)
```

```
## Loading required package: ggplot2
```

```
library(terra)
```

```
## terra 1.7.9
```

```
library(sf)
```

```
## Linking to GEOS 3.9.3, GDAL 3.5.2, PROJ 8.2.1; sf_use_s2() is TRUE
```

```
library(ggplot2)
library(tidyr)
```

```
##
## Attaching package: ’tidyr’
```

```
## The following object is masked from ’package:terra’:
##
##     extract
```

```
library(dplyr)
```

```
##
## Attaching package: ’dplyr’
```

```
## The following objects are masked from ’package:terra’:
##
##     intersect, union
```

```
## The following object is masked from ’package:summarizeNHTS’:
##
##     select_all
```

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

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

```
library(scales)
```

```
##
## Attaching package: ’scales’
```

```
## The following object is masked from ’package:terra’:
##
##     rescale
```

Then I download the NHTS 2017 data and keeping it into a master data set.

```r
setwd("C:/Users/Benny Panjaitan/Documents/GitHub/esp106-Naomi/final")
unzip("data.zip")

# sign below must be removed in order to download the NHTS 2017 data
#download_nhts_data("2017", "C:/Users/Benny Panjaitan/Documents/GitHub/esp106-Naomi/final")
nhts <- read_data("2017")
```

```
## =========================================================================

## Reading Person Weights

## =========================================================================

## Reading Household Weights

## =========================================================================

## Reading Trip Table.

## =========================================================================

## Reading Person Table.

## =========================================================================

## Reading Household Table.

## =========================================================================

## Reading Vehicle Table.

## =========================================================================

## Derived variables:
## MSA_AGG, PRMACT_AGG, WHYTRP90_AGG successfully added!
```

I tidied up the NHTS 2017 data base to contain only the following variables:

1. Weighted household trip rate ("W")
2. Household travel frequency by train ("TRAIN")
3. States where the household live ("HHSTATE")
4. Household income level ("HHFAMINC")
5. Availability of rail infrastrucutr ("RAIL")

```
nhts_sum <- summarize_data(
    data = nhts,
    agg = "household_trip_rate",
    label = F,
    by = c("HHSTATE", "TRAIN", "HHFAMINC", "RAIL")
)
```

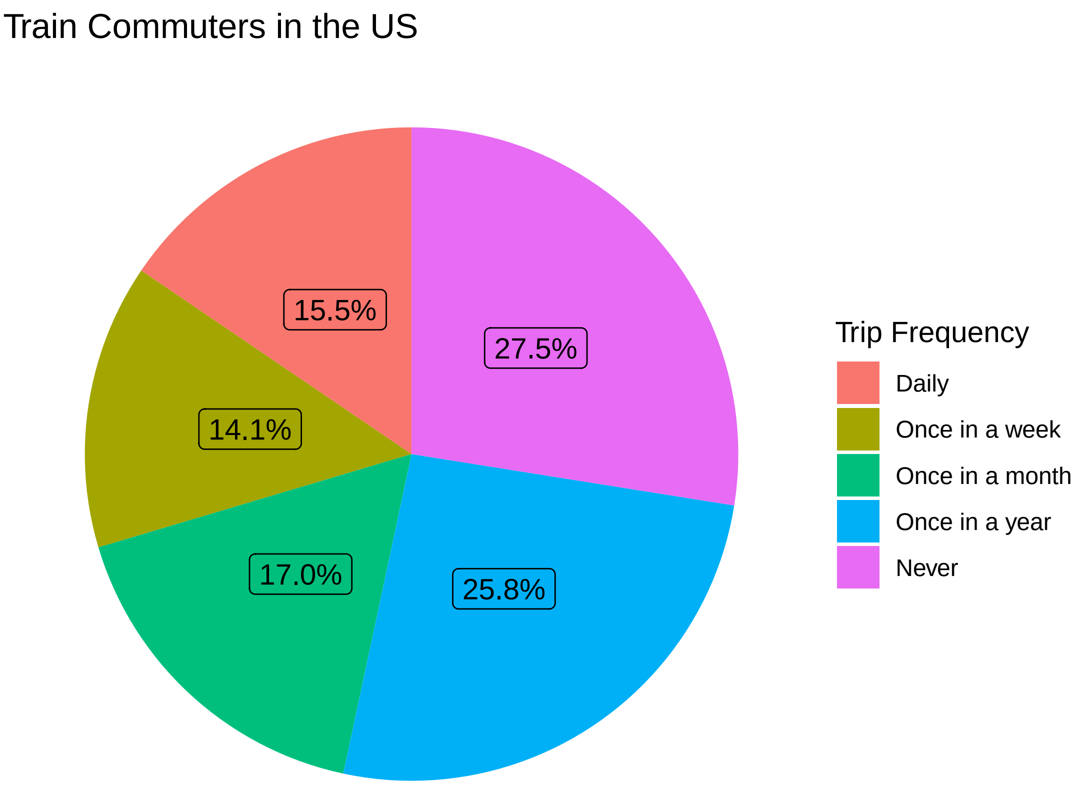
To control the study, “nhts_sum” data frame is subsetted to contain only train trips (“TRAIN”) made in metropolitan neighborhood with Rail infrastructure (“RAIL”). Besides, only household income (“HHFAMINC”) categories that have specified values are included in the data frame.

```
train_sum <- subset(nhts_sum, RAIL == "01" & TRAIN %in% c("01", "02", "03", "04", "05") & HHFAMINC %in%
```

Train Trips From the data, I look at the percentage of US households making train trips.

```
perc_freq <- train_sum %>%
  group_by(TRAIN) %>% # Variable to be transformed
  count() %>%
  ungroup() %>%
  mutate(perc = `n` / sum(`n`)) %>%
  arrange(perc) %>%
  mutate(labels = scales::percent(perc))
```

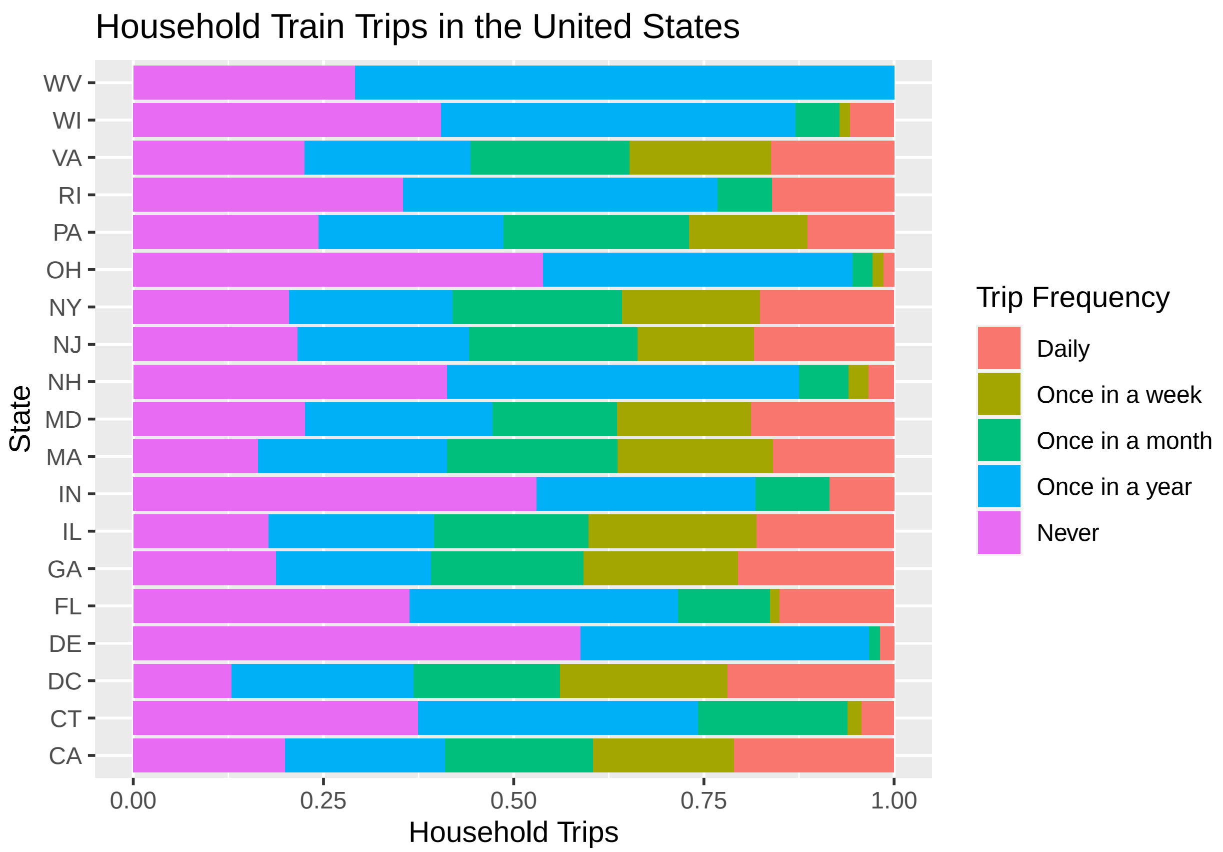
```
perc_freq %>%
  ggplot(aes(x = "", y = perc, fill = TRAIN)) +
  geom_col() +
  geom_label(aes(label = labels),
             position = position_stack(vjust = 0.5),
             show.legend = FALSE) +
  coord_polar(theta = "y") +
  theme_void() +
  scale_fill_discrete(name="Trip Frequency",labels = c("Daily", "Once in a week", "Once in a month", "O
  labs(title="Train Commuters in the US")
```



The result shows that only 15.5% of US households travel by train in daily basis, while 27.5% never travel by train.

Next, I compare the number of train ("TRAIN") trips made in different State ("HHSTATE").

```
train_sum %>%
ggplot(aes(x=HHSTATE, y=W, fill=TRAIN)) +
  geom_bar(position="fill", stat="identity") +
  labs(title="Household Train Trips in the United States",
       y = "Household Trips", x = "State") +
  scale_fill_discrete(name ="Trip Frequency", labels = c("Daily", "Once in a week", "Once in a month",
  scale_y_continuous(label = scales::comma) +
  coord_flip ()
```



Fraction from the bar plot above shows that states that have the highest share of daily train commuters are District of Columbia (DC), California (CA), and Georgia (GA).

Household Income To see the variability of household income in train trips made, the household income categories are modified from 11 NHTS categorization into 5 groups as below.

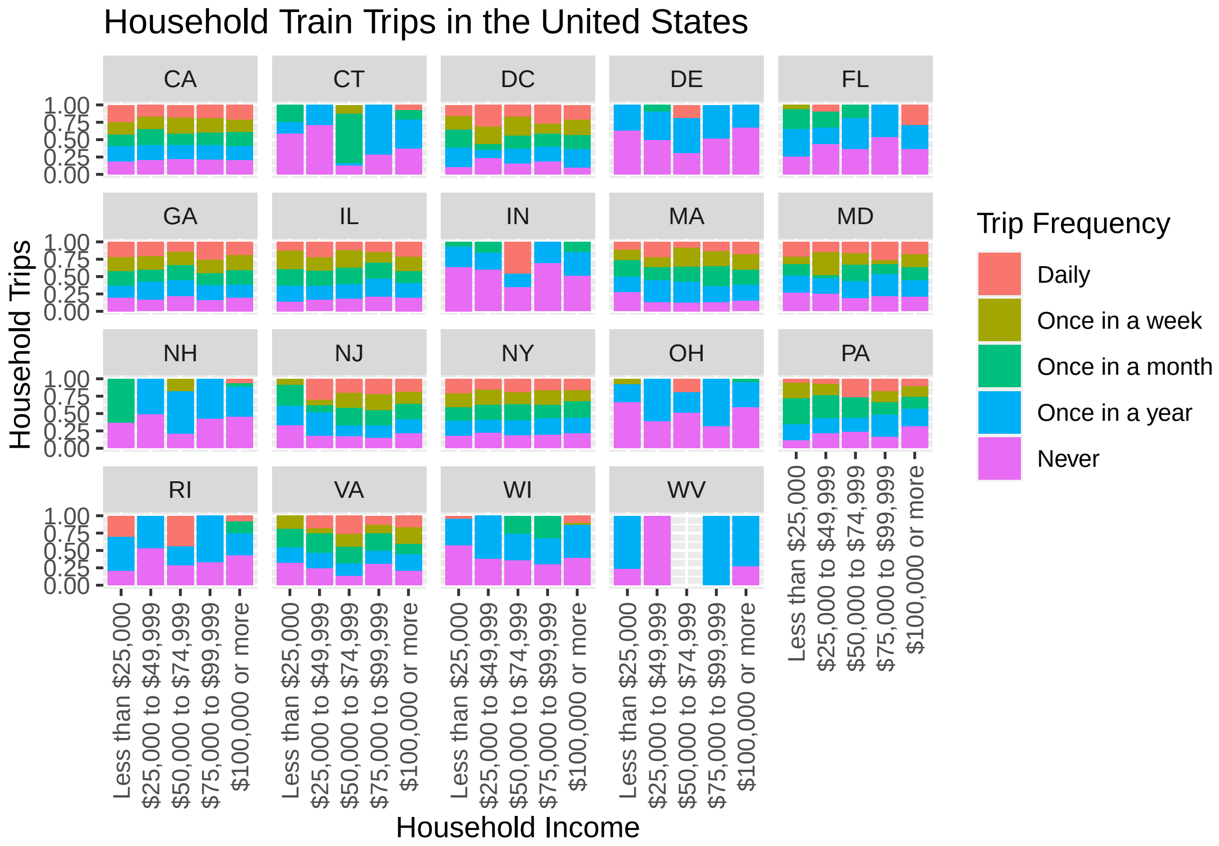
1. Less than $25,000
2. $25,000 to $49,999
3. $50,000 to $74,999
4. $75,000 to $99,999
5. $100,000 or more

```
train <- train_sum

train$HHFAMINC[train$HHFAMINC=="01"]<-"01"
train$HHFAMINC[train$HHFAMINC=="02"]<-"01"
train$HHFAMINC[train$HHFAMINC=="03"]<-"01"
train$HHFAMINC[train$HHFAMINC=="04"]<-"02"
train$HHFAMINC[train$HHFAMINC=="05"]<-"02"
train$HHFAMINC[train$HHFAMINC=="06"]<-"03"
train$HHFAMINC[train$HHFAMINC=="07"]<-"04"
train$HHFAMINC[train$HHFAMINC=="08"]<-"05"
```

```
train$HHFAMINC[train$HHFAMINC=="09"]<-"05"
train$HHFAMINC[train$HHFAMINC=="10"]<-"05"
train$HHFAMINC[train$HHFAMINC=="11"]<-"05"
```

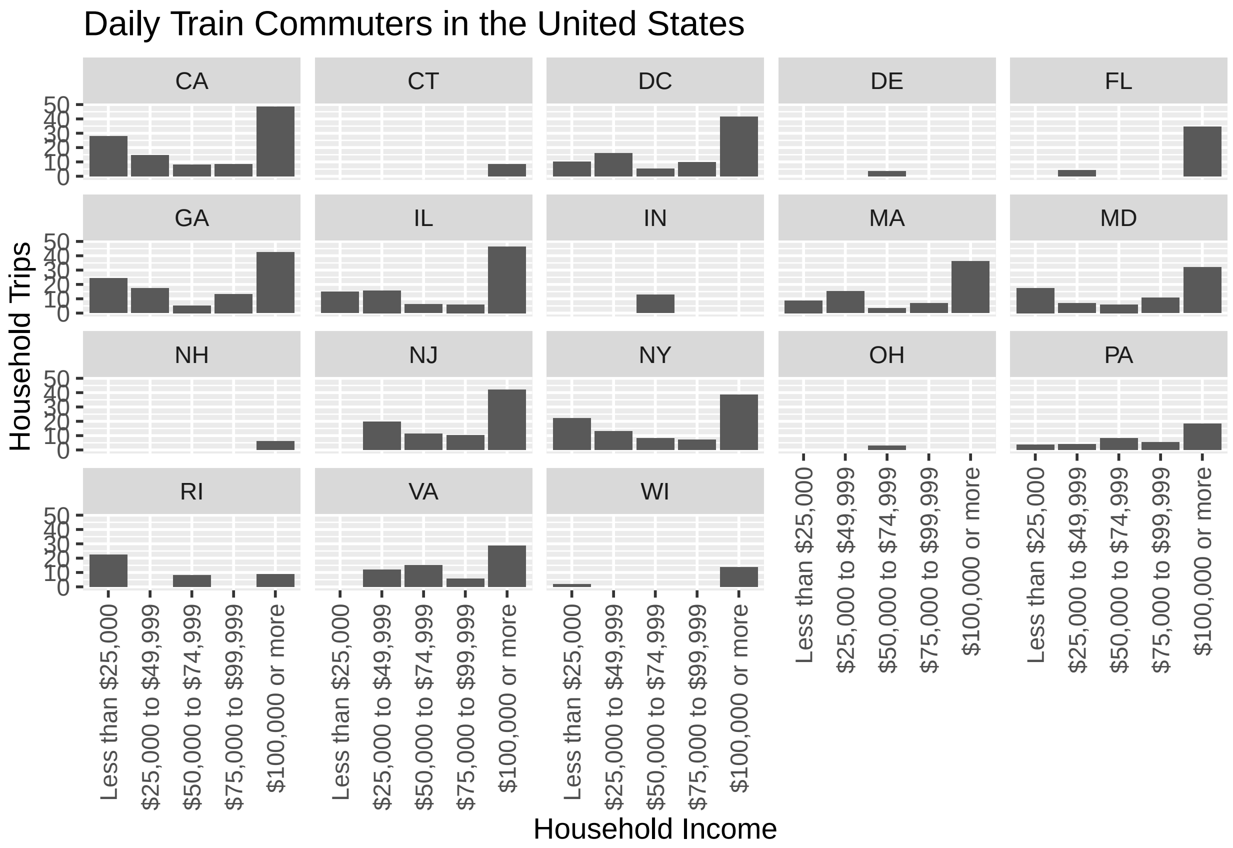
```
train %>%
ggplot(aes(x=HHFAMINC, y=W, fill=TRAIN)) +
  geom_bar(position="fill", stat="identity") +
  facet_wrap(~HHSTATE) +
  labs(title="Household Train Trips in the United States",
       y = "Household Trips", x = "Household Income") +
  scale_fill_discrete(name ="Trip Frequency", labels =c("Daily", "Once in a week", "Once in a month", "(
  scale_x_discrete(labels = c("Less than $25,000", "$25,000 to $49,999", "$50,000 to $74,999", "$75,000
  scale_y_continuous(label = scales::comma) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5))
```



From the income categorization, there are various pattern of train commuters' income in different state. Even, in some states, only some income groups reported that they made train trips in daily basis and no particular pattern of what specific income group commute by train daily in these states.

```
t1 <- subset(train, TRAIN =="01")

t1 %>%
ggplot(aes(x=HHFAMINC, y=W)) +
  geom_bar(position="stack", stat="identity") +
  facet_wrap (~HHSTATE) +
  labs(title="Daily Train Commuters in the United States",
       y = "Household Trips", x = "Household Income") +
  scale_x_discrete(labels = c("Less than $25,000", "$25,000 to $49,999", "$50,000 to $74,999", "$75,000
  scale_y_continuous(label = scales::comma) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5))
```



ANALYSIS

This section covers analyses to show the relationship between households daily train trips and their income in the chosen state, California. It will also depict how is the income level of households residing nearby the train stations of Sacramento Regional Transit (SacRT) Light Rail service.

This section is organized into 3 parts:

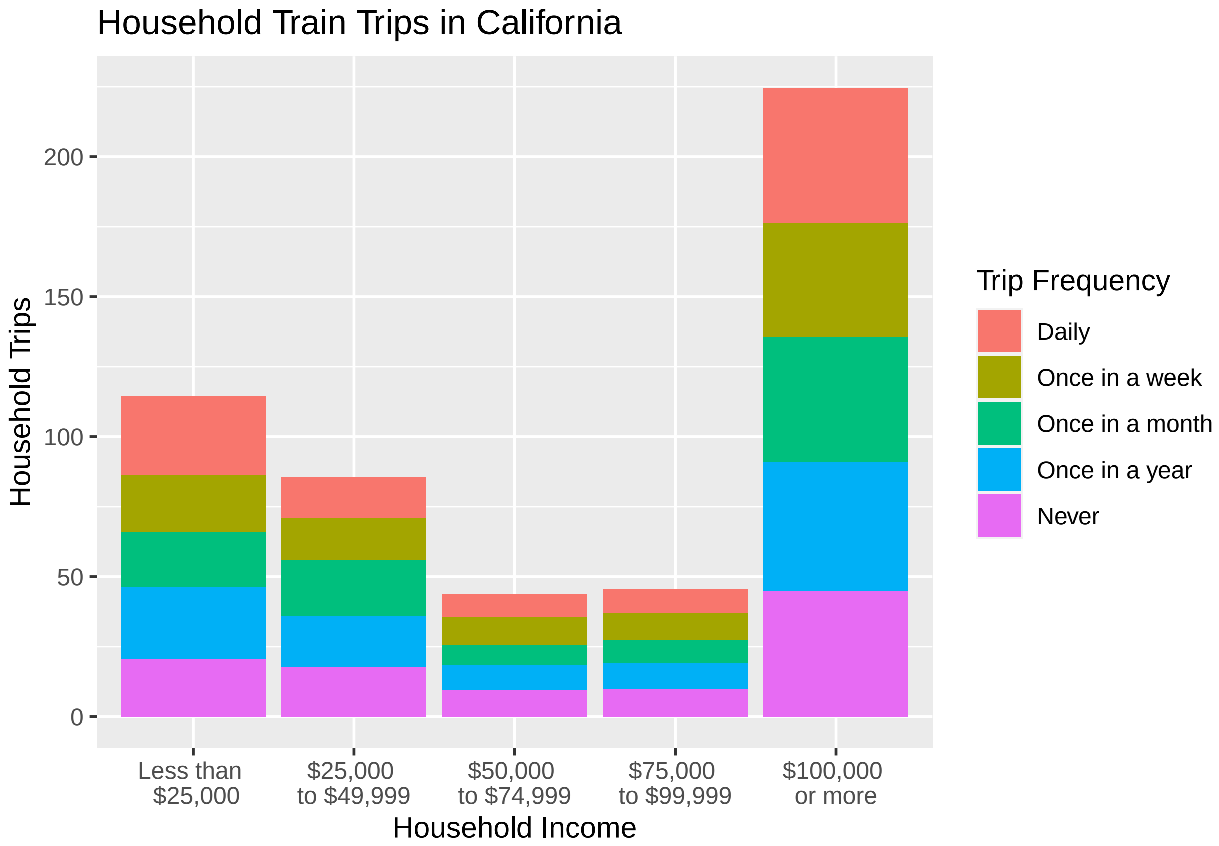
1. Descriptive statistics
2. Negative Binomial mixed model (NBMM)
3. Income level of Household along SacRT Light Rail service

Descriptive Statistics

Analysis takes California as the subject, for it shows high fraction and high count of train commuters in the United States, in general and in daily trip frequency, as shown in bar plot in the previous section. Below is the number of household train trips made in California.

```r
train_ca <- subset(train, HHSTATE == "CA")

train_ca %>%
ggplot(aes(x=HHFAMINC, y=W, fill=TRAIN)) +
  geom_bar(position="stack", stat="identity") +
  labs(title="Household Train Trips in California",
       y = "Household Trips", x = "Household Income") +
  scale_fill_discrete(name ="Trip Frequency", labels =c("Daily", "Once in a week", "Once in a month", "
  scale_x_discrete(labels = c("Less than \n $25,000", "$25,000 \nto $49,999", "$50,000 \nto $74,999", "
```



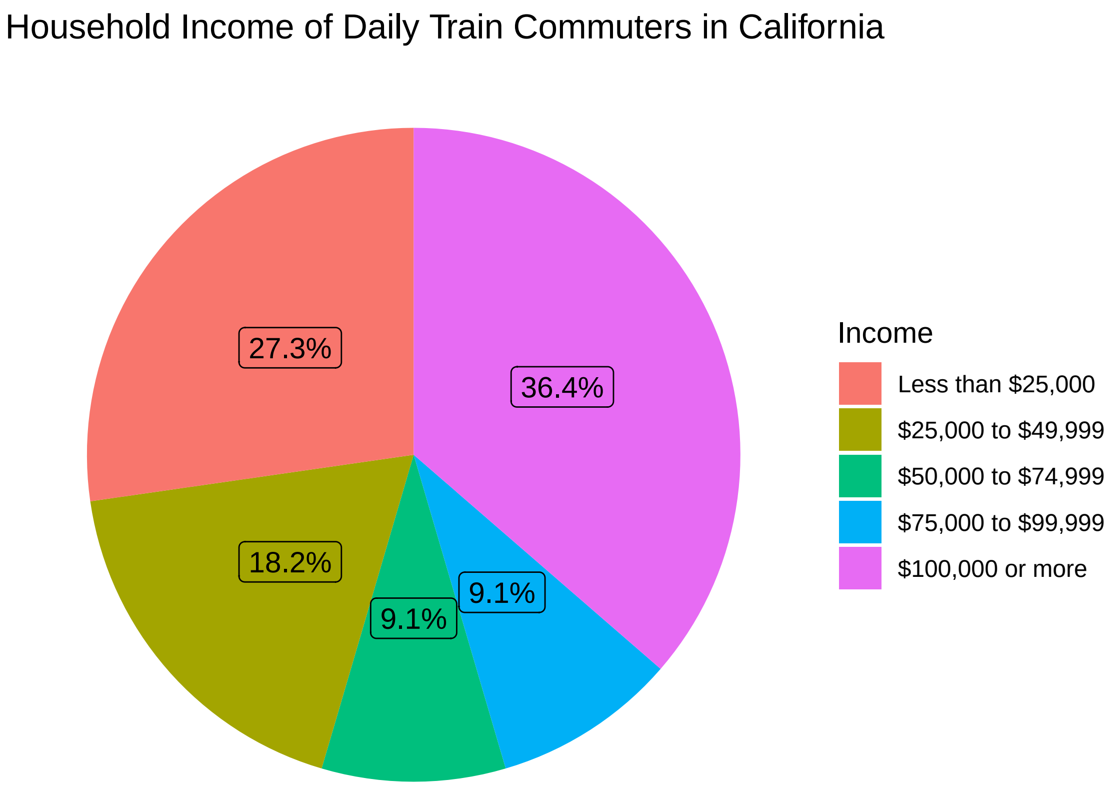
The highest income category shows a significant number of household trips made. It can be implied that households with income higher than $100,000 makes more trips than other income groups. However, to see how significant income level is to the train trips, I will analyse the train trips made daily.

```r
ca1 <- subset(train_ca, TRAIN =="01")

perc_inc <- ca1 %>%
  group_by(HHFAMINC) %>% # Variable to be transformed
  count() %>%
  ungroup() %>%
  mutate(perc = `n` / sum(`n`)) %>%
  arrange(perc) %>%
  mutate(labels = scales::percent(perc))
```

```
pie_ca <- perc_inc %>%
  ggplot(aes(x = "", y = perc, fill = HHFAMINC)) +
  geom_col() +
  geom_label(aes(label = labels),
             position = position_stack(vjust = 0.5),
             show.legend = FALSE) +
  coord_polar(theta = "y") +
  theme_void() +
  scale_fill_discrete(name="Income",labels = c("Less than $25,000", "$25,000 to $49,999", "$50,000 to $
  labs(title="Household Income of Daily Train Commuters in California")

pie_ca
```



From the pie chart, it is shown that the largest share of daily train commuters is household within the highest income group (more than $100,000) and the second one is the lowest income group (less than $25,000).

Negative Binomial Mixed Model (NBMM)

In order to run the Negative Binomial Mixed Model (NBMM), I install an R package of "NBZIMM" (Negative Binomial and Zero-Inflated Mixed Models) from https://github.com/nyiuab/NBZIMM

```
library(remotes)
install_github("nyiuab/NBZIMM", force=T, build_vignettes=F)
```

```
## Downloading GitHub repo nyiuab/NBZIMM@HEAD
## -- R CMD build -----------------------------------------------------------------
```

```
## WARNING: Rtools is required to build R packages, but no version of Rtools compatible with R 4.2.2 wa
##
## Please download and install Rtools 4.2 from https://cran.r-project.org/bin/windows/Rtools/ or https:/
```

```
##          checking for file 'C:\Users\Benny Panjaitan\AppData\Local\Temp\Rtmpuuocx3\remotesde44dba3d6
##       -  preparing 'NBZIMM':
##    checking DESCRIPTION meta-information ...  v  checking DESCRIPTION meta-information
##       -  checking for LF line-endings in source and make files and shell scripts
##       -  checking for empty or unneeded directories
##       -  building 'NBZIMM_1.0.tar.gz'
##
##
```

```
## Installing package into 'C:/Users/Benny Panjaitan/AppData/Local/R/win-library/4.2'
## (as 'lib' is unspecified)
```

```
library(NBZIMM)
```

For this analyses, negative binomial is chosen based on similar study that have been done (Boarnet et. al., 2020). The model works well with numerous variables, yet this study only focus on demographic variable of household income. Therefore, the modeling goes as below.

```
model <- glmm.nb(fixed = W ~ HHFAMINC,
              random = ~ 1 | TRAIN, data = train_ca, zi_fixed = ~1)
```

```
## Loading required namespace: nlme
```

```
## Loading required namespace: MASS
```

```
##
## Attaching package: 'nlme'
```

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

```
##
## Attaching package: 'MASS'
```

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

```
## The following object is masked from 'package:terra':
##
##     area
```

```
## Computational iterations: 2
## Computational time: 0.002 minutes
```

```
summary(model)
```

```
## Linear mixed-effects model fit by maximum likelihood
##   Data: train_ca
##   AIC BIC logLik
##    NA  NA     NA
##
## Random effects:
##  Formula: ~1 | TRAIN
##         (Intercept)  Residual
## StdDev: 4.065849e-06 0.4554007
##
## Variance function:
##  Structure: fixed weights
##  Formula: ~invwt
## Fixed effects:  W ~ HHFAMINC
##                 Value  Std.Error DF   t-value p-value
## (Intercept) 2.0322821 0.07095486 46 28.641903  0.0000
## HHFAMINC02  0.1145341 0.11073834 46  1.034277  0.3064
## HHFAMINC03  0.1335644 0.13925109 46  0.959162  0.3425
## HHFAMINC04  0.1804753 0.13838763 46  1.304129  0.1987
## HHFAMINC05  0.3864584 0.09127671 46  4.233921  0.0001
##  Correlation:
##            (Intr) HHFAMINC02 HHFAMINC03 HHFAMINC04
## HHFAMINC02 -0.641
## HHFAMINC03 -0.510  0.326
## HHFAMINC04 -0.513  0.329      0.261
## HHFAMINC05 -0.777  0.498      0.396      0.399
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -1.70304801 -0.35436181 -0.09382013  0.37282149  5.25883190
##
## Number of Observations: 55
## Number of Groups: 5
```

The p-value shows that the highest income group (more than $100,000) and the lowest income group (less than $25,000) have the most significant effect toward train trips taken daily.

Income level of Household along SacRT Light Rail service

The last part of this section aims to see whether the findings of relationship between daily train trips and household income will be proved spatially. The spatial analysis will be taken from a case study of Sacramento Regional Transit (SacRT) Light Rail in Sacramento, one of Metropolitan area with more than 1 million population that has rail infrastructure in California. Data sets are obtained from SacRT data repository and census data of household income.

The light rail data, containing route and stops data, are loaded in simple feature (sf) format.

```
route <- read_sf("data/lr_routes/Light_Rail_Routes.shp") %>%
  st_transform(6417)
stops <- read_sf("data/lr_stations.geojson") %>%
  st_transform(6417)
```

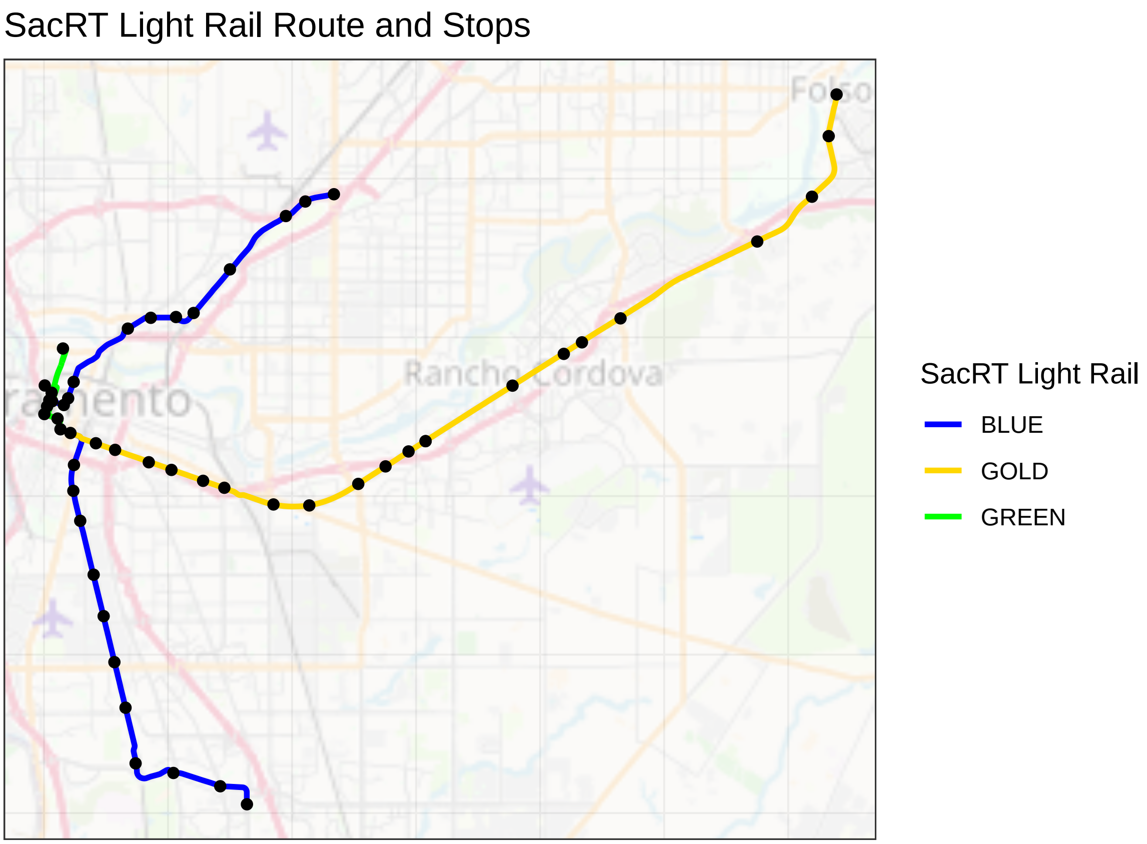
```
ggplot() +
  ggspatial::annotation_map_tile(alpha = 0.3) +
  geom_sf(data = route, aes(color = Line), lwd = 1) +
  geom_sf(data = stops) +
  scale_color_identity(guide = "legend", name = "SacRT Light Rail") +
  theme_bw() +
  theme(axis.text = element_blank(),
        axis.ticks = element_blank()) +
  labs(title="SacRT Light Rail Route and Stops")
```

```
## Zoom: 10

## Fetching 4 missing tiles

##   |                                                                              |

## ...complete!
```



Above is the SacRT Light Rail network coverage in Sacramento county, from the city of Sacramento to the city of Rancho Cordova, up to the city of Folsom.

In addition to the SacRT Light Rail network, demographic data of household income data is loaded in simple feature (sf) format. The household income is categorized based on census tract. Then, the income data is joined with the route map to see how the income level of households living around SacRT Light Rail stations is.

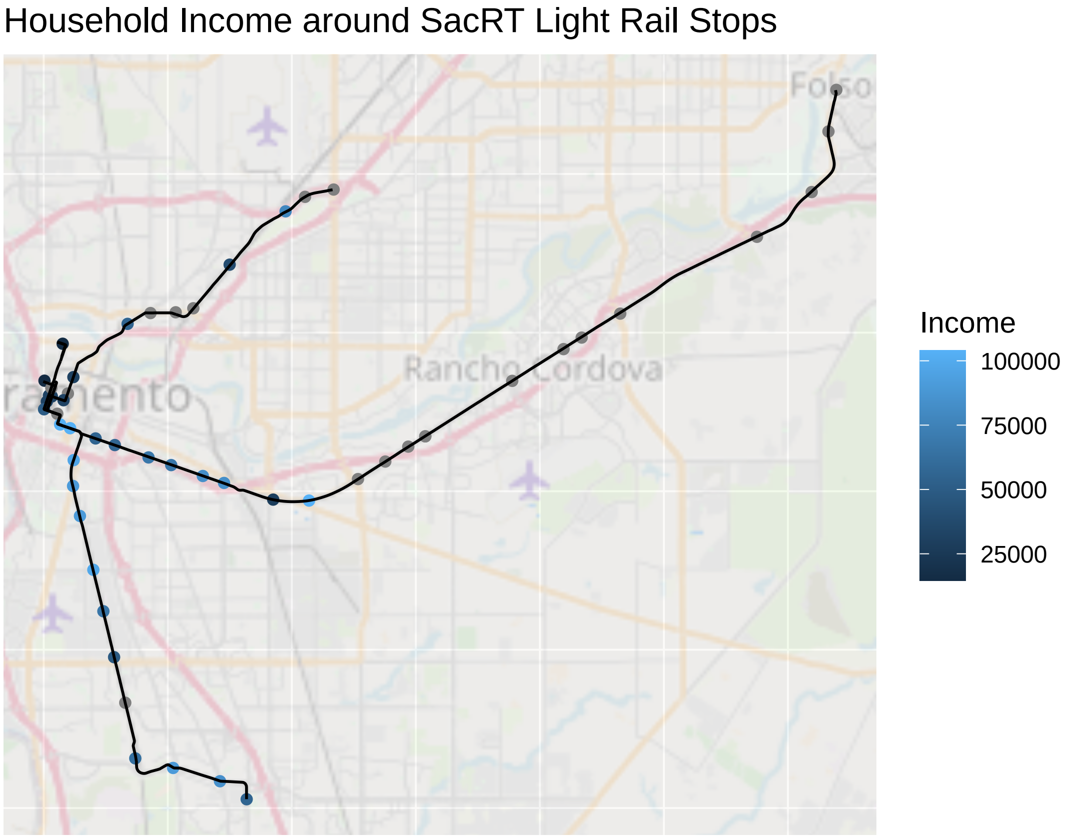
```
income <- read_sf("data/sacramento_income.gpkg") %>%
  st_transform(6417)

stops_income <- sf::st_join(stops, income, join = st_intersects, left = T)

names(stops_income)[names(stops_income) == 'estimate'] <- 'Income'
```

```
ggplot() +
  ggspatial::annotation_map_tile(alpha = 0.3) +
  geom_sf(data = stops_income, aes(color = Income)) +
  geom_sf(data = route) +
  theme(axis.text = element_blank(),
        axis.ticks = element_blank()) +
  labs(title="Household Income around SacRT Light Rail Stops")
```

```
## Zoom: 10
```



From the map above, it is seen that Sacramento RT light rail commuters income variations are consistent with findings on previous section specifically in downtown area. Households nearby stations on the northern part of downtown are categorized in the lowest income group (less than $25,000), while households nearby stations on the southern part of downtown are those who have income more than $100,000.

Although this pattern is interesting to see, it does not have sufficient base to support the causal relationship between the household income of daily train commuters and the households' resident proximity to train station. It is because people living nearby the stations are not directly implied as the train commuters.

People may live near to train stops without being the train riders, while the train commuters may be people living outside the census tract of the train stops.

CONCLUSION

This study reveals that 27.5% of U.S. households residing in metropolitan area with rail infrastructure never travel by train, while 15.5% travel by train in daily basis. In national level, daily train commuters with income more than $100,000 have the highest share. In California, 36.4% of its daily train commuters has income more than $100,000, while 27.3% earn less than $25,000. These groups show significance in daily train trips made, but not in a linear model.