Assignment 1 & 2

Shanay Patel

14/02/2021

Packages Installed

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.5     v dplyr   1.0.3
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

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

```
library(haven)
```

Hypothesis

This assignment attempts to understand the relationship between mode choice and the existing public infrastructure in a city (San Francisco-Oakland-Hayward, CA). The inferences drawn will be compared with those from two other cities with differing characteristics (Houston-The Woodlands-Sugar Land, TX & Washington-Arlington-Alexandria, DC-VA-MD-WV). The underlying assumption is that the quality of public transport infrastructure is best in D.C. followed by S.F., and then Houston.This assignment explores how this impacts mode choice of the cities' denizens.

Project Data Directory

The necessary data sets and directories are set up at the beginning of the assignment.

```
trips <- read_sav("C:/Users/Shana/Desktop/Urbana/UP 431 BL - Urban Transportation Modeling/Assignment 1
hh <- read_sav("C:/Users/Shana/Desktop/Urbana/UP 431 BL - Urban Transportation Modeling/Assignment 1 & 2
per <- read_sav("C:/Users/Shana/Desktop/Urbana/UP 431 BL - Urban Transportation Modeling/Assignment 1 &
veh <- read_sav("C:/Users/Shana/Desktop/Urbana/UP 431 BL - Urban Transportation Modeling/Assignment 1 &
```

Convert to Factors

```
trips <- as_factor(trips)
hh <- as_factor(hh)
per <- as_factor(per)
veh <- as_factor(veh)
```

San Francisco, Houston & Washington, D.C.

The data pertaining to San Francisco, Houston, and Washington D.C. are extracted from the data set.

```
sf_trips <- trips %>%
  filter(HH_CBSA == "San Francisco-Oakland-Hayward, CA")
ht_trips <- trips %>%
  filter(HH_CBSA == "Houston-The Woodlands-Sugar Land, TX")
dc_trips <- trips %>%
  filter(HH_CBSA == "Washington-Arlington-Alexandria, DC-VA-MD-WV")
```

The NHTS trip data will be used in Task 1 of Part 1: Univariates and Bivariates, which presents an aggregated mode choice using simplified mode categories.

```
sf_hh <- hh %>%
  filter(hh_cbsa == "San Francisco-Oakland-Hayward, CA")
ht_hh <- hh %>%
  filter(hh_cbsa == "Houston-The Woodlands-Sugar Land, TX")
dc_hh <- hh %>%
  filter(hh_cbsa == "Washington-Arlington-Alexandria, DC-VA-MD-WV")
```

The NHTS household data will be used in Tasks 2 and 7 of Part 1: Univariates and Bivariates, which summarize mode choice and car ownership by household income groups respectively.

```
sf_per <- per %>%
  filter(HH_CBSA == "San Francisco-Oakland-Hayward, CA")
ht_per <- per %>%
  filter(HH_CBSA == "Houston-The Woodlands-Sugar Land, TX")
dc_per <- per %>%
  filter(HH_CBSA == "Washington-Arlington-Alexandria, DC-VA-MD-WV")
```

The NHTS person data will be used in Task 6 of Part 1: Univariates and Bivariates, which summarizes public transit use by census tract level population density.

```
sf_veh <- veh %>%
  filter(HH_CBSA == "San Francisco-Oakland-Hayward, CA")
ht_veh <- veh %>%
  filter(HH_CBSA == "Houston-The Woodlands-Sugar Land, TX")
dc_veh <- veh %>%
  filter(HH_CBSA == "Washington-Arlington-Alexandria, DC-VA-MD-WV")
```

The NHTS vehicle data will be used in Task 11 of Part 1: Univariates and Bivariates, which shows annual household vehicle miles traveled by income group.

1. Part 1: Univariate and Birvariate Summaries

Modes of Transportation

The different modes in the NHTS data are examined, and cross-verified with Appendix C of the NHTS Travel Trends.The output below shows the categories in the variable “TRPTRANS,” for the San Francisco data set.

```
levels(sf_trips$TRPTRANS)
```

```
##  [1] "I prefer not to answer"
##  [2] "I don’t know"
##  [3] "Not ascertained"
##  [4] "Walk"
##  [5] "Bicycle"
##  [6] "Car"
##  [7] "SUV"
##  [8] "Van"
##  [9] "Pickup truck"
## [10] "Golf cart / Segway"
## [11] "Motorcycle / Moped"
## [12] "RV (motor home, ATV, snowmobile)"
## [13] "School bus"
## [14] "Public or commuter bus"
## [15] "Paratransit / Dial-a-ride"
## [16] "Private / Charter / Tour / Shuttle bus"
## [17] "City-to-city bus (Greyhound, Megabus)"
## [18] "Amtrak / Commuter rail"
## [19] "Subway / elevated / light rail / street car"
## [20] "Taxi / limo (including Uber / Lyft)"
## [21] "Rental car (Including Zipcar / Car2Go)"
## [22] "Airplane"
## [23] "Boat / ferry / water taxi"
## [24] "Something Else"
```

There are certain categories which do not fit with the categorization in the NHTS Travel Trends, and must be reorganized.

San Francisco

```
sf_trips <- sf_trips %>%
  mutate(sf_mode_short = fct_collapse(TRPTRANS,
                                   Walk = "Walk",
                                   Bicycle = "Bicycle",
                                   PV = c("Car", "SUV", "Van",
                                            "Pickup truck",
                                            "Golf cart / Segway",
                                            "Motorcycle / Moped",
                                            "RV (motor home, ATV, snowmobile)"),
                                   PT = c("Public or commuter bus",
                                               "Paratransit / Dial-a-ride",
                                               "City-to-city bus (Greyhound, Megabus)",
                                               "Amtrak / Commuter rail",
                                               "Subway / elevated / light rail / street car","Private /
                                   Missing = c("I prefer not to answer",
                                               "I don't know",
                                               "Not ascertained"),
                                   Other = c("School bus", "Taxi / limo (including Uber / Lyft)",
                                             "Rental car (Including Zipcar / Car2Go)",
                                             "Airplane", "Boat / ferry / water taxi",
                                             "Something Else")))
sf_trips <- sf_trips %>%
  filter(sf_mode_short != "Missing")
```

- Private Vehicles (PV) and Public Transportation (PT) are categorized based on Appendix C of the NHTS Travel Trends.
- Walk and Bicycle have not been categorized as "Non-motorized" to provide a clearer picture of mode choice.

The same process is repeated to standardize the data for the Houston and Washington D.C. metropolitan areas.

Houston

```
ht_trips <- ht_trips %>%
  mutate(ht_mode_short = fct_collapse(TRPTRANS,
                                   Walk = "Walk",
                                   Bicycle = "Bicycle",
                                   PV = c("Car", "SUV", "Van",
                                            "Pickup truck",
                                            "Golf cart / Segway",
                                            "Motorcycle / Moped",
                                            "RV (motor home, ATV, snowmobile)"),
                                   PT = c("Public or commuter bus",
                                               "Paratransit / Dial-a-ride",
                                               "City-to-city bus (Greyhound, Megabus)",
                                               "Amtrak / Commuter rail",
                                               "Subway / elevated / light rail / street car","Private /
                                   Missing = c("I prefer not to answer",
                                               "I don't know",
                                               "Not ascertained"),
                                   Other = c("School bus", "Taxi / limo (including Uber / Lyft)",
                                             "Rental car (Including Zipcar / Car2Go)",
                                             "Airplane", "Boat / ferry / water taxi",
                                             "Something Else")))
ht_trips <- ht_trips %>%
  filter(ht_mode_short != "Missing")
```

Washington D.C.

```
dc_trips <- dc_trips %>%
  mutate(dc_mode_short = fct_collapse(TRPTRANS,
                                   Walk = "Walk",
                                   Bicycle = "Bicycle",
                                   PV = c("Car", "SUV", "Van",
                                            "Pickup truck",
                                            "Golf cart / Segway",
                                            "Motorcycle / Moped",
                                            "RV (motor home, ATV, snowmobile)"),
                                   PT = c("Public or commuter bus",
                                               "Paratransit / Dial-a-ride",
                                               "City-to-city bus (Greyhound, Megabus)",
                                               "Amtrak / Commuter rail",
                                               "Subway / elevated / light rail / street car","Private /
                                   Missing = c("I prefer not to answer",
                                               "I don't know",
                                               "Not ascertained"),
                                   Other = c("School bus", "Taxi / limo (including Uber / Lyft)",
                                             "Rental car (Including Zipcar / Car2Go)",
                                             "Airplane", "Boat / ferry / water taxi",
                                             "Something Else")))
dc_trips <- dc_trips %>%
  filter(dc_mode_short != "Missing")
```

1.1 Mode Choice (and 2. Summary Across Metropolitan Areas)

1.1.1 Aggregated Mode Choice

This section also covers Task 2: Summary across metropolitan areas. Firstly, the Aggregated Mode Choice and Mode Share by Trip Purpose is shown for San Francisco. Thereafter, the aforementioned result is compared with the summaries of Washington D.C. and Houston. The second summary of comparison is Mode Choice by Income Group, which is carried out in a format similar to the first summary comparison.

San Francisco

```
sf_mode_share <- sf_trips %>%
  group_by(sf_mode_short) %>%
  tally(wt = WTTRDFIN) %>%
  mutate(sf_pct = n / sum(n))
```

Houston

```
ht_mode_share <- ht_trips %>%
  group_by(ht_mode_short) %>%
  tally(wt = WTTRDFIN) %>%
  mutate(ht_pct = n / sum(n))
```

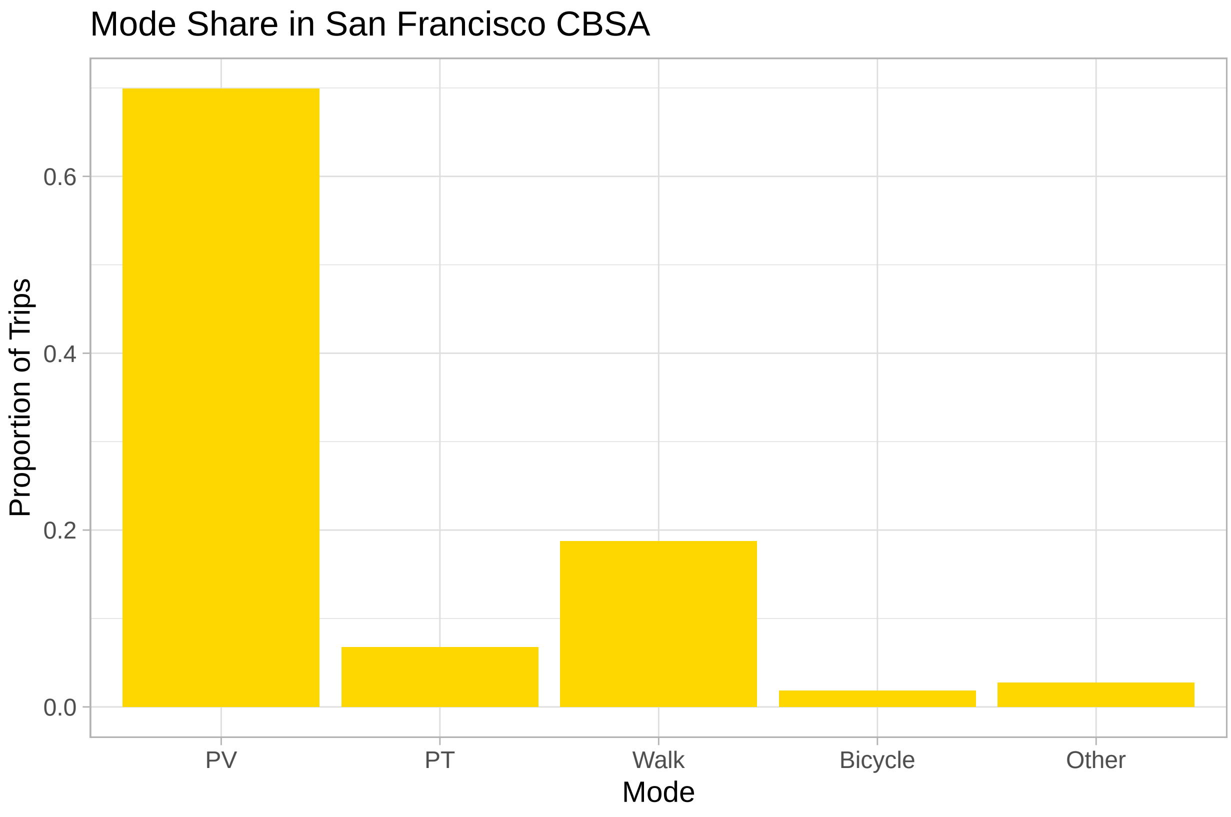
Washington D.C.

```
dc_mode_share <- dc_trips %>%
  group_by(dc_mode_short) %>%
  tally(wt = WTTRDFIN) %>%
  mutate(dc_pct = n / sum(n))
```

The mode share data generated above is summarized in the graphs below.

San Francisco

```
library(ggplot2)
sf_mode_share$sf_mode_short <- factor(sf_mode_share$sf_mode_short, levels = c("PV","PT","Walk","Bicycle
ggplot(sf_mode_share, aes(sf_mode_short, sf_pct)) +
  geom_col(fill = "gold") +
  labs(x = "Mode", y = "Proportion of Trips", title = "Mode Share in San Francisco CBSA",
       caption = "Source: NHTS (2017)") + theme_light()
```


Mode Share in San Francisco CBSA

Source: NHTS (2017)

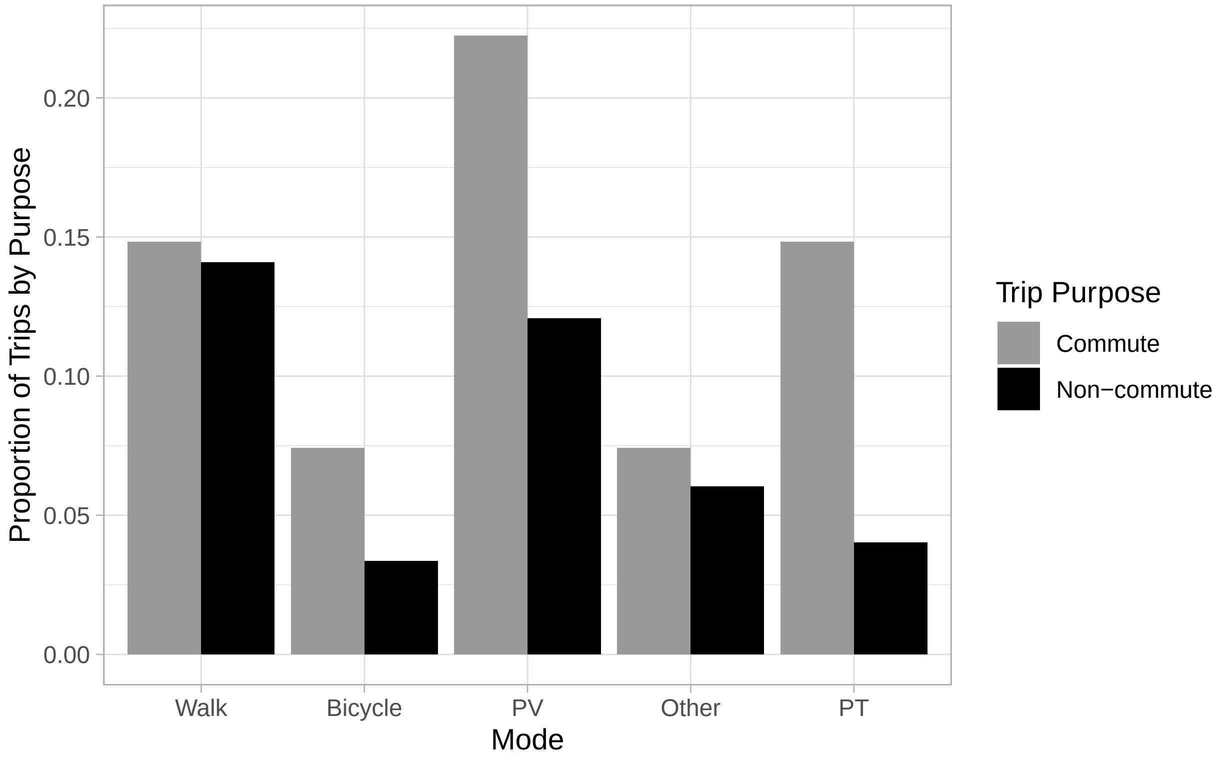
The graph above shows that private vehicles comprise almost 70% of daily commute in San Francisco. Approximately 7% of the population uses public transportation while a significant proportion of the population prefers to walk.

The data is sorted based on trip purpose; work-based trips are termed “commute trips” and other trips are termed “non-commute trips.”

```
sf_mode_share_commute <- sf_trips %>%
  mutate(sf_commuteTrp = fct_collapse(WHYTRP90,
                                      sf_commute_trip = "To/From Work",
                                      sf_non_commute_trip = c("Work-Related Business","Shopping","Other Fa
                                      Missing = "Refused / Don't Know")) %>%
  count(sf_commuteTrp, sf_mode_short, sf_wt = WTTRDFIN)%>%
  group_by(sf_commuteTrp) %>%
  mutate(sf_per = prop.table(n)*100)

sf_mode_share_commute <- sf_mode_share_commute %>%
  filter(sf_commuteTrp != "Missing")

#Error Log: Needed to filter out "Missing" values for correct visual representation.
```

```
ggplot(sf_mode_share_commute, aes(x = sf_mode_short, y = sf_per)) +
  geom_bar(aes(fill = sf_commuteTrp), position = "dodge", stat = 'identity') +
  labs(x = "Mode", y = "Proportion of Trips by Purpose", title = "Mode Share by Trip Purpose in San Fra
       caption = "Source: NHTS (2017)", fill= "Trip Purpose") +
  scale_fill_manual(values=c("#999999", "#000000", "#999999"), name="Trip Purpose",
                    breaks=c("sf_commute_trip","sf_non_commute_trip"),
                    labels=c("Commute", "Non-commute")) + theme_light()
```

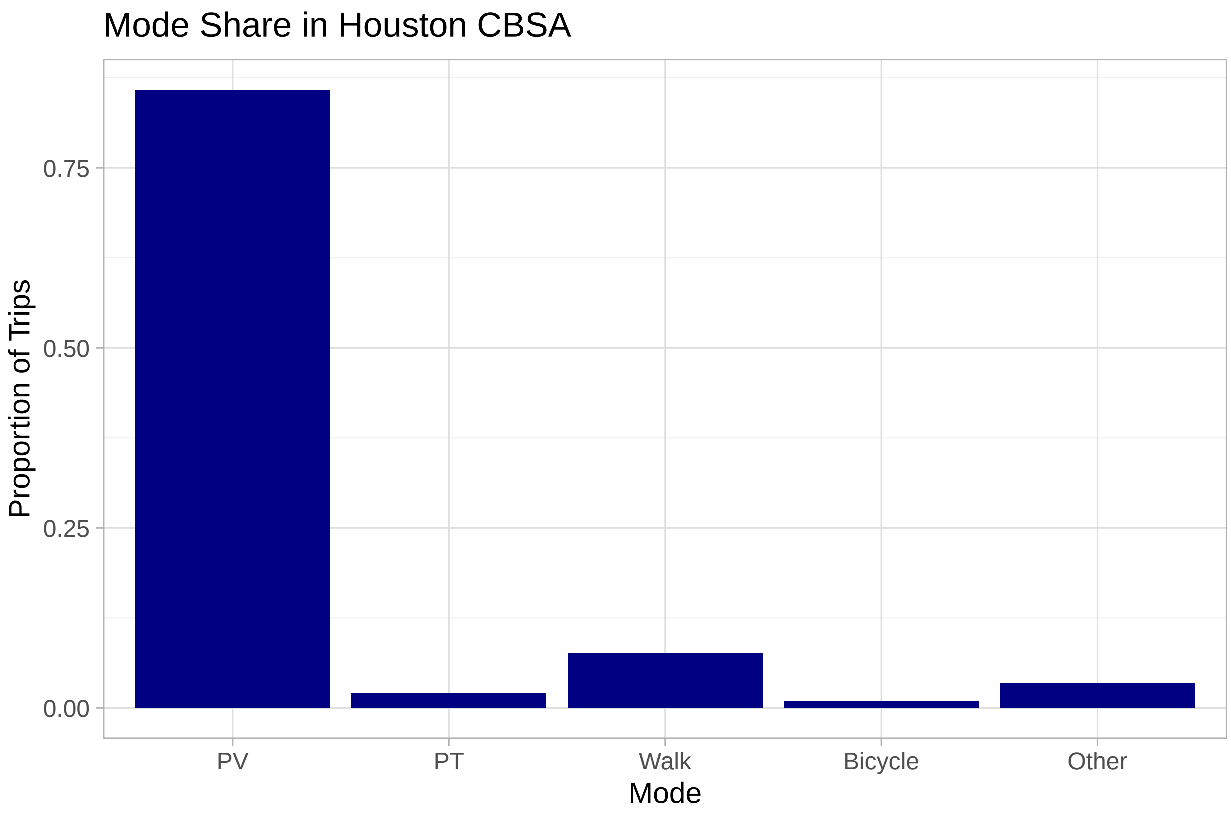
Mode Share by Trip Purpose in San Francisco CBSA



Source: NHTS (2017)

### Houston

```
ht_mode_share$ht_mode_short <- factor(ht_mode_share$ht_mode_short, levels = c("PV","PT","Walk","Bicycle
ggplot(ht_mode_share, aes(ht_mode_short, ht_pct)) +
  geom_col(fill = "navy") +
  labs(x = "Mode", y = "Proportion of Trips", title = "Mode Share in Houston CBSA",
       caption = "Source: NHTS (2017)") + theme_light()
```

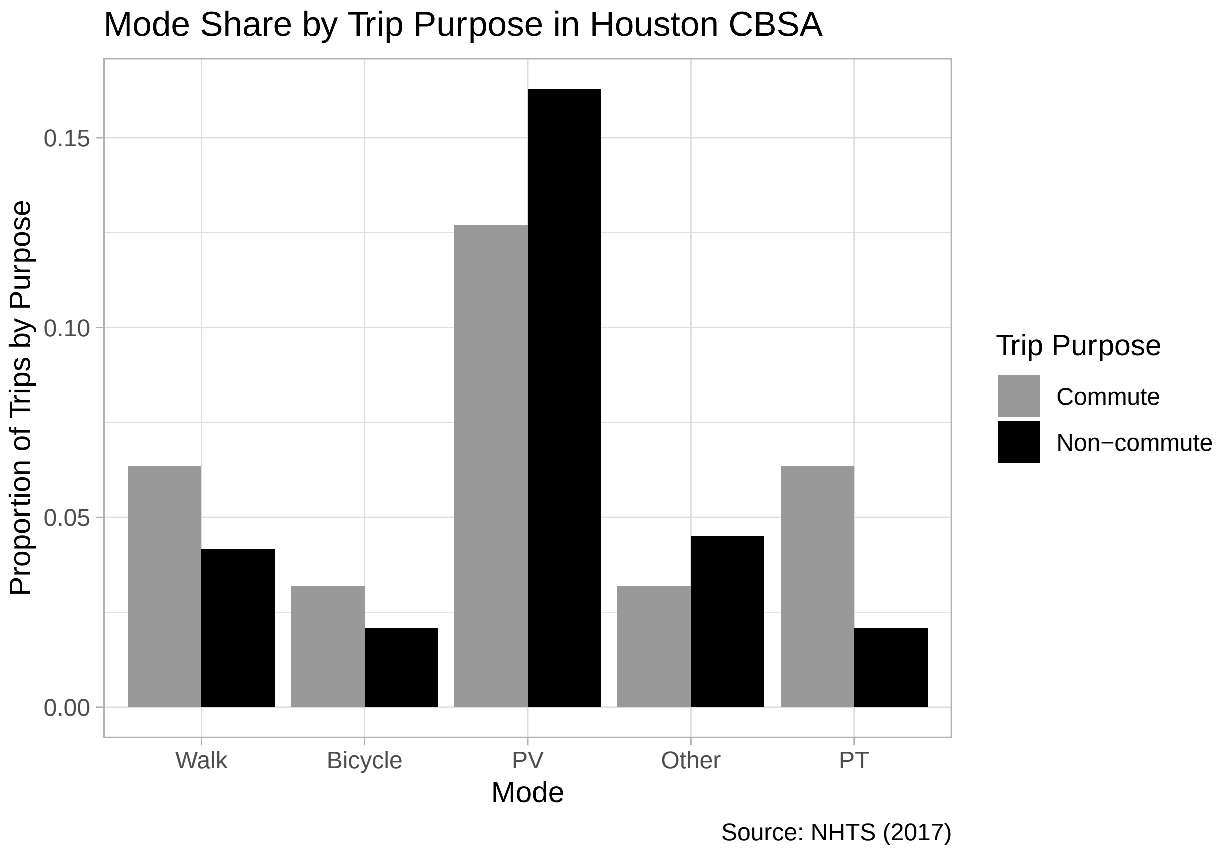


Source: NHTS (2017)

The mode share in Houston is remarkably different from that of San Francisco, as is evident from the graph above. 85% of Houston's populace prefers to commute by private vehicles which is significant compared to the 2% who commute using public transportation.

```
ht_mode_share_commute <- ht_trips %>%
  mutate(ht_commuteTrp = fct_collapse(WHYTRP90,
                                      ht_commute_trip = "To/From Work",
                                      ht_non_commute_trip = c("Work-Related Business","Shopping","Other Fa
                                      Missing = "Refused / Don't Know")) %>%
  count(ht_commuteTrp, ht_mode_short, ht_wt = WTTRDFIN)%>%
  group_by(ht_commuteTrp) %>%
  mutate(ht_per = prop.table(n)*100)

ht_mode_share_commute <- ht_mode_share_commute %>%
  filter(ht_commuteTrp != "Missing")
```

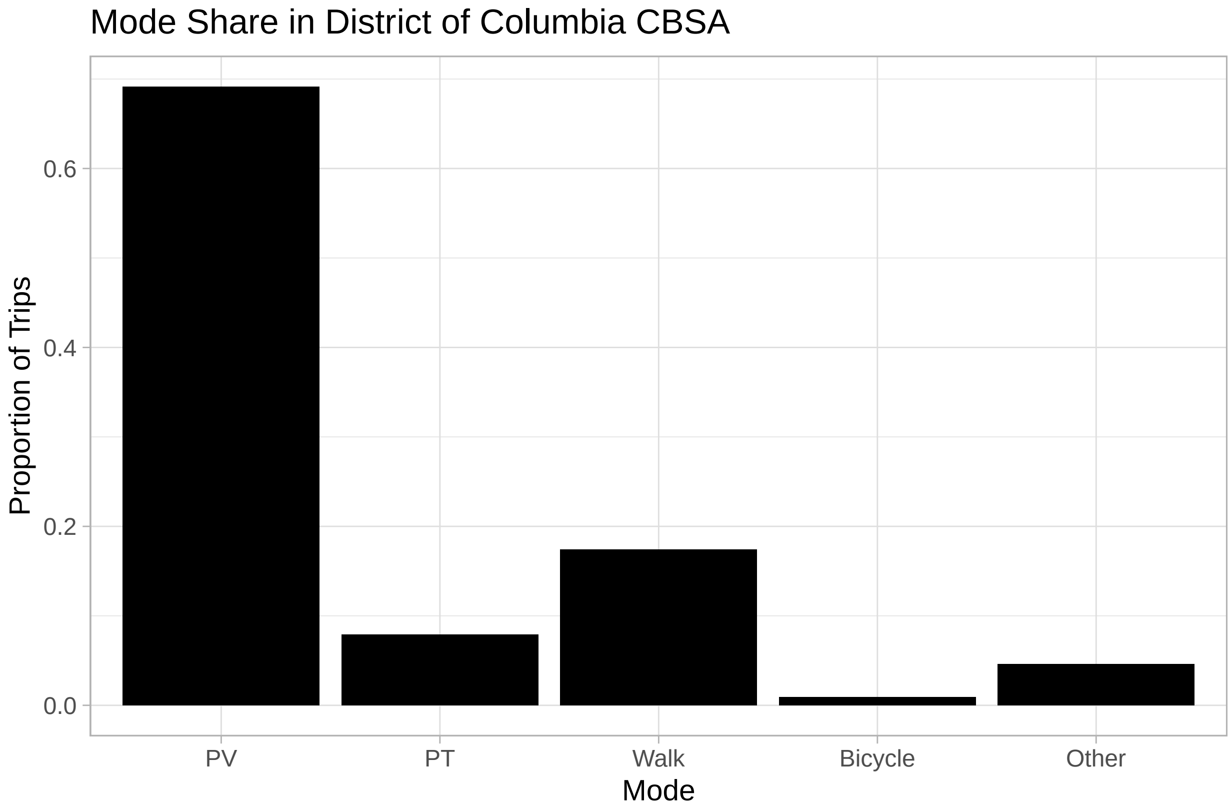
```
ggplot(ht_mode_share_commute, aes(x = ht_mode_short, y = ht_per)) +
  geom_bar(aes(fill = ht_commuteTrp), position = "dodge", stat = 'identity') +
  labs(x = "Mode", y = "Proportion of Trips by Purpose", title = "Mode Share by Trip Purpose in Houston
       caption = "Source: NHTS (2017)", fill= "Trip Purpose") +
  scale_fill_manual(values=c("#999999", "#000000", "#999999"), name="Trip Purpose",
                       breaks=c("ht_commute_trip","ht_non_commute_trip"),
                       labels=c("Commute", "Non-commute")) + theme_light()
```


Mode Share by Trip Purpose in Houston CBSA

Source: NHTS (2017)

### Washington D.C.

```
dc_mode_share$dc_mode_short <- factor(dc_mode_share$dc_mode_short, levels = c("PV","PT","Walk","Bicycle
ggplot(dc_mode_share, aes(dc_mode_short, dc_pct)) +
  geom_col(fill = "black") +
  labs(x = "Mode", y = "Proportion of Trips", title = "Mode Share in District of Columbia CBSA",
       caption = "Source: NHTS (2017)") + theme_light()
```

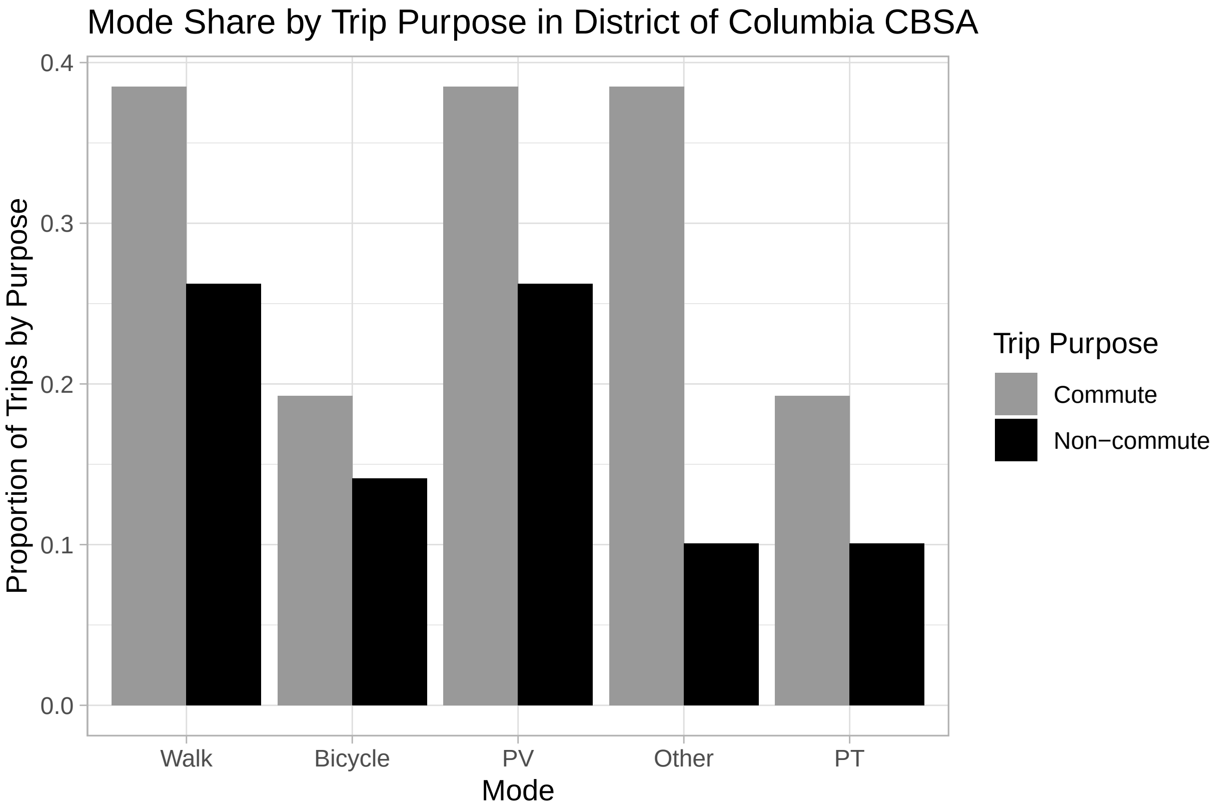


Source: NHTS (2017)

The mode choice scenario in Washington D.C. is similar to San Francisco, albeit the proportion of public transportation users is 8%, which is higher than that observed in San Francisco. Private vehicle is less dominant in Washington D.C. compared to Houston.

```
dc_mode_share_commute <- dc_trips %>%
  mutate(dc_commuteTrp = fct_collapse(WHYTRP90,
                                   dc_commute_trip = "To/From Work",
                                   dc_non_commute_trip = c("Work-Related Business","Shopping","Other Fa
                                   Missing = "Refused / Don't Know")) %>%
  count(dc_commuteTrp, dc_mode_short, dc_wt = WTTRDFIN)%>%
  group_by(dc_commuteTrp) %>%
  mutate(dc_per = prop.table(n)*100)

dc_mode_share_commute <- dc_mode_share_commute %>%
  filter(dc_commuteTrp != "Missing")
```

```
ggplot(dc_mode_share_commute, aes(x = dc_mode_short, y = dc_per)) +
  geom_bar(aes(fill = dc_commuteTrp), position = "dodge", stat = 'identity') +
  labs(x = "Mode", y = "Proportion of Trips by Purpose", title = "Mode Share by Trip Purpose in Distric
       caption = "Source: NHTS (2017)", fill= "Trip Purpose") +
  scale_fill_manual(values=c("#999999", "#000000", "#999999"), name="Trip Purpose",
                        breaks=c("dc_commute_trip","dc_non_commute_trip"),
                        labels=c("Commute", "Non-commute")) + theme_light()
```



Source: NHTS (2017)

# 1.1.2 Mode Choice by Houshold Income

The income levels are grouped in a simplified manner to produce results succinctly.

```
levels(sf_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"
```

```
levels(dc_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"
```

```
levels(ht_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"
```

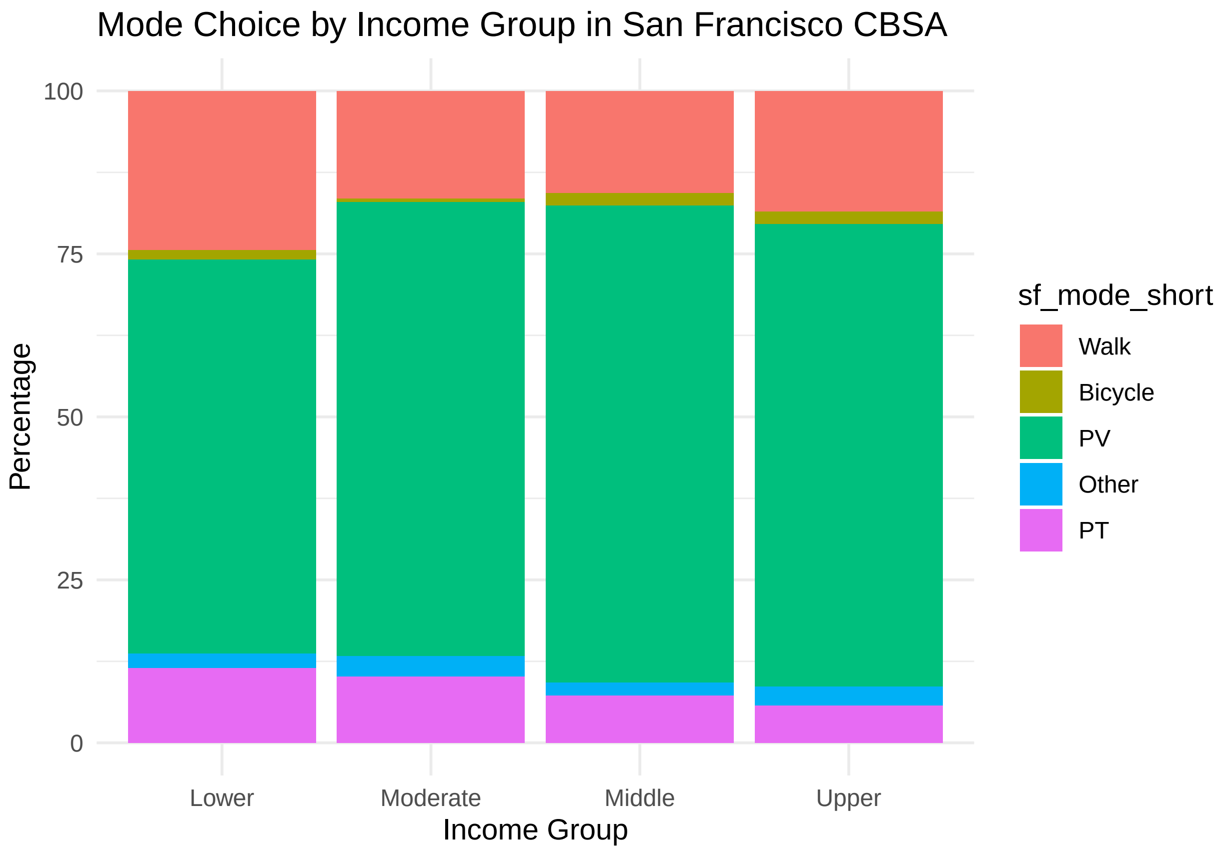
San Francisco

```
sf_hh <- sf_hh %>%
  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")
```

```
sf_mode_choice_income <- sf_trips %>%
  mutate(sf_incomegrp = 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(sf_incomegrp

  count(sf_incomegrp, sf_mode_short, wt = WTTRDFIN) %>%
  group_by(sf_incomegrp) %>%
  mutate(sf_per_incomegrp = prop.table(n)*100)

sf_mode_choice_income <- sf_mode_choice_income %>%
  filter(sf_incomegrp != "Missing")
```

```
ggplot(data=sf_mode_choice_income, aes(x=sf_incomegrp, y=sf_per_incomegrp, fill = sf_mode_short))+
  geom_histogram(binwidth = 2500, stat='identity')+
  labs(title = "Mode Choice by Income Group in San Francisco CBSA", x = "Income Group", y = "Percentage
  theme_minimal()
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



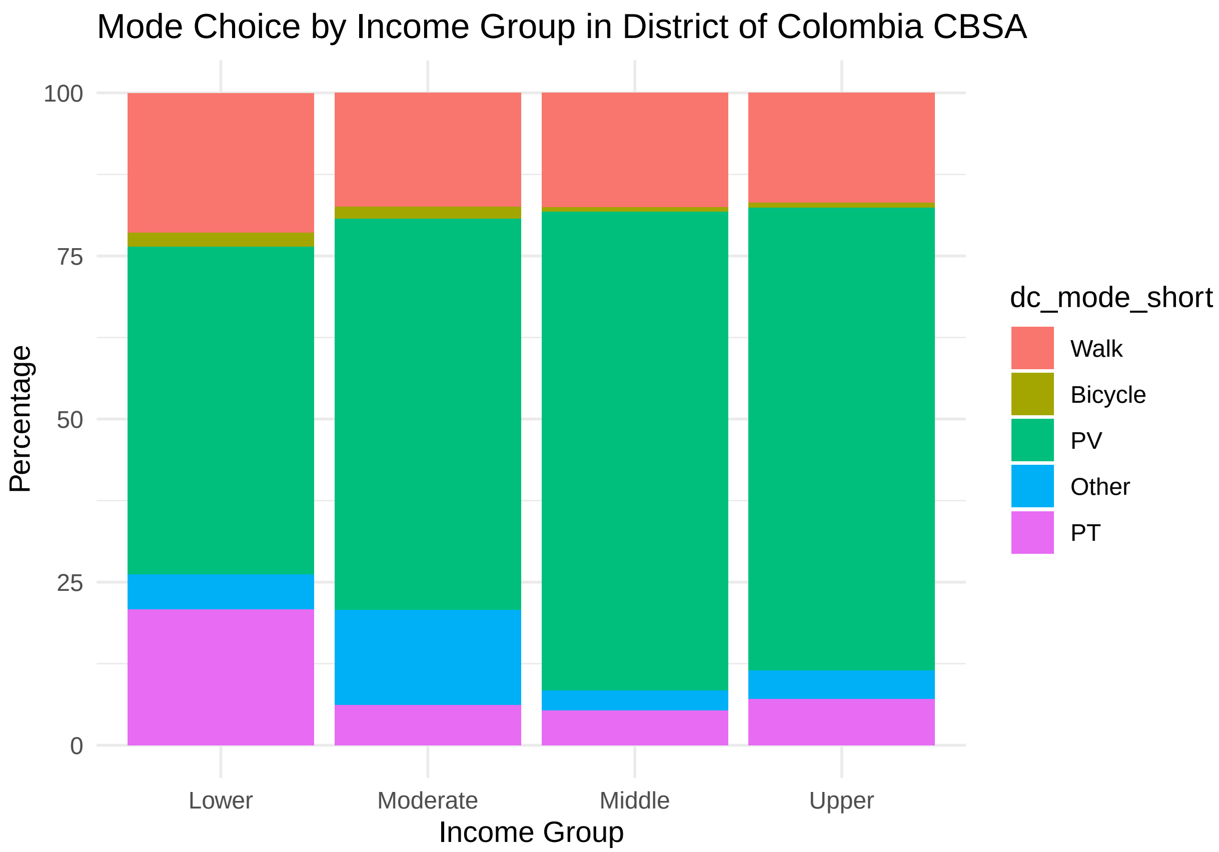
### Washington D.C.

```
dc_hh <- dc_hh %>%
  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")
```

```
dc_mode_choice_income <- dc_trips %>%
  mutate(dc_incomegrp = 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(dc_incomegrp

  count(dc_incomegrp, dc_mode_short, wt = WTTRDFIN) %>%
  group_by(dc_incomegrp) %>%
  mutate(dc_per_incomegrp = prop.table(n)*100)

dc_mode_choice_income <- dc_mode_choice_income %>%
  filter(dc_incomegrp != "Missing")
```

```
ggplot(data=dc_mode_choice_income, aes(x=dc_incomegrp, y=dc_per_incomegrp, fill = dc_mode_short))+
  geom_histogram(binwidth = 2500, stat='identity')+
  labs(title = "Mode Choice by Income Group in District of Colombia CBSA", x = "Income Group", y = "Perc
  theme_minimal()
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



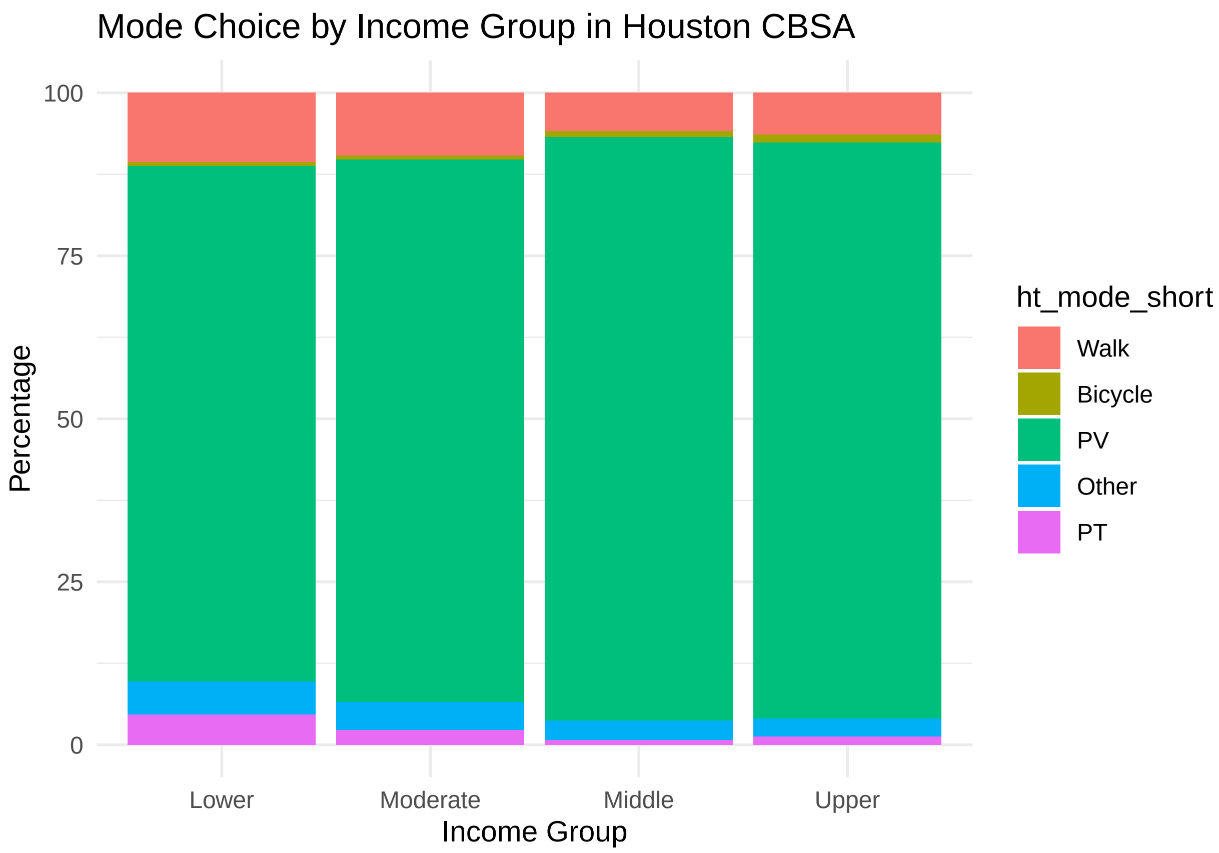
Houston

```
ht_hh <- ht_hh %>%
  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")
```

```
ht_mode_choice_income <- ht_trips %>%
  mutate(ht_incomegrp = 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(ht_incomegr

  count(ht_incomegrp, ht_mode_short, wt = WTTRDFIN) %>%
  group_by(ht_incomegrp) %>%
  mutate(ht_per_incomegrp = prop.table(n)*100)

ht_mode_choice_income <- ht_mode_choice_income %>%
  filter(ht_incomegrp != "Missing")
```

```
ggplot(data=ht_mode_choice_income, aes(x=ht_incomegrp, y=ht_per_incomegrp, fill = ht_mode_short))+
  geom_histogram(binwidth = 2500, stat='identity')+
  labs(title = "Mode Choice by Income Group in Houston CBSA", x = "Income Group", y = "Percentage")+
  theme_minimal()
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



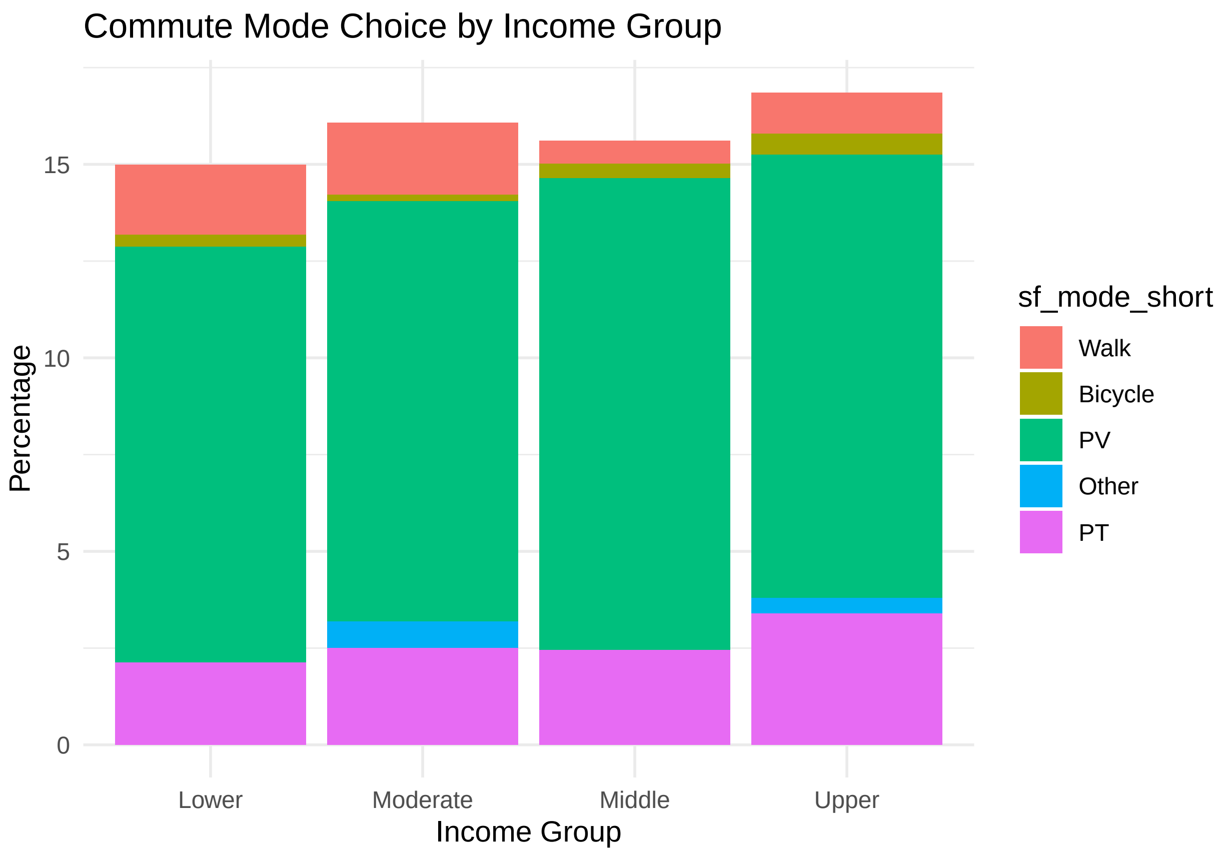
1.1.3 Commute Mode Choice by Houshold Income

```
sf_mode_choice_commute_income <- sf_trips %>%
 mutate(sf_commuteTrp = fct_collapse(WHYTRP90,
                                  sf_commute_trip = "To/From Work",
                                  sf_non_commute_trip = c("Work-Related Business","Shopping","Other Fa
                                  Missing = "Refused / Don't Know")) %>%
  mutate(sf_incomegrp = 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(sf_incomegr
```

```
count(sf_incomegrp, sf_mode_short, sf_commuteTrp, wt = WTTRDFIN) %>%
 group_by(sf_incomegrp) %>%
 mutate(sf_per_commute_incomegrp = prop.table(n)*100)

sf_mode_choice_commute_income <-  sf_mode_choice_commute_income %>%
 filter(sf_incomegrp != "Missing") %>%
  filter(sf_commuteTrp != "Missing") %>%
  filter(sf_commuteTrp == "sf_commute_trip")
```

```
ggplot(data=sf_mode_choice_commute_income, aes(x=sf_incomegrp, y=sf_per_commute_incomegrp, fill = sf_mo
  geom_histogram(binwidth = 2500, stat='identity')+
  labs(title = "Commute Mode Choice by Income Group", x = "Income Group", y = "Percentage")+
  theme_minimal()
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



The graph above shows only the mode choice representing commute choice by income group. This is achieved by filtering out the rows of non-commute trips; thus, the percentage representation is out of 100% of the data set and not the proportion of commute trips.

1.1.4 Mode choice by population density in the census tract of the household's home location (HTPPOPDN)

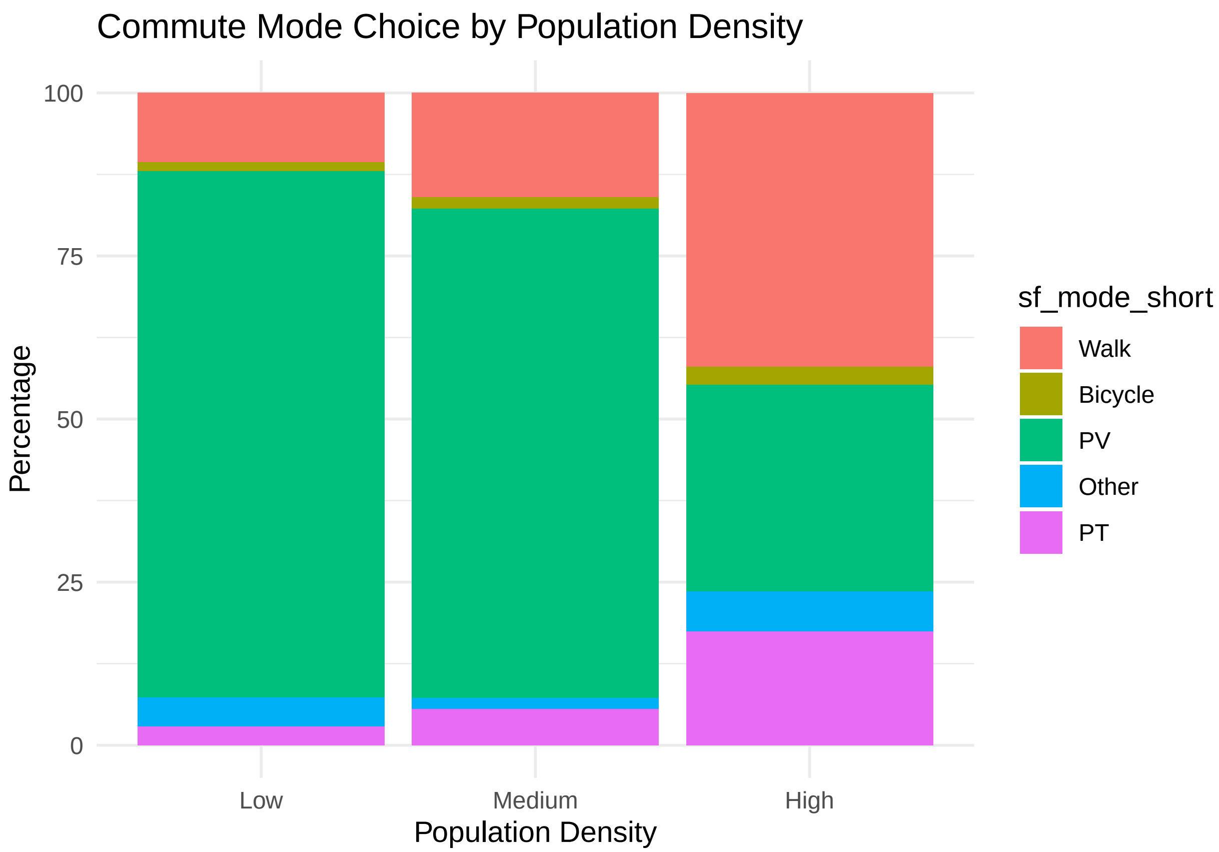
```
levels(sf_trips$OTPPOPDN)
```

```
## [1] "Not ascertained" "1,000-1,999"     "10,000-24,999"   "100-499"
## [5] "2,000-3,999"     "25,000-999,999"  "0-99"            "4,000-9,999"
## [9] "500-999"
```

```
sf_mode_choice_density <- sf_trips %>%
  mutate(sf_density = fct_collapse(OTPPOPDN,
                                   Low = c("0-99","100-499","1,000-1,999","500-999"),
                                   Medium = c("2,000-3,999","4,000-9,999","10,000-24,999"),
                                   High = "25,000-999,999",
                                   Missing = "Not ascertained"))

sf_mode_choice_density <-  sf_mode_choice_density %>%
  filter(sf_density != "Missing") %>%

 count(sf_density, sf_mode_short, wt = WTTRDFIN) %>%
  group_by(sf_density) %>%
  mutate(sf_per_density = prop.table(n)*100)
```

```
ggplot(data=sf_mode_choice_density, aes(x=sf_density, y=sf_per_density, fill = sf_mode_short))+
  geom_histogram(binwidth = 2500, stat='identity')+
  labs(title = "Commute Mode Choice by Population Density", x = "Population Density", y = "Percentage")+
  theme_minimal()
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



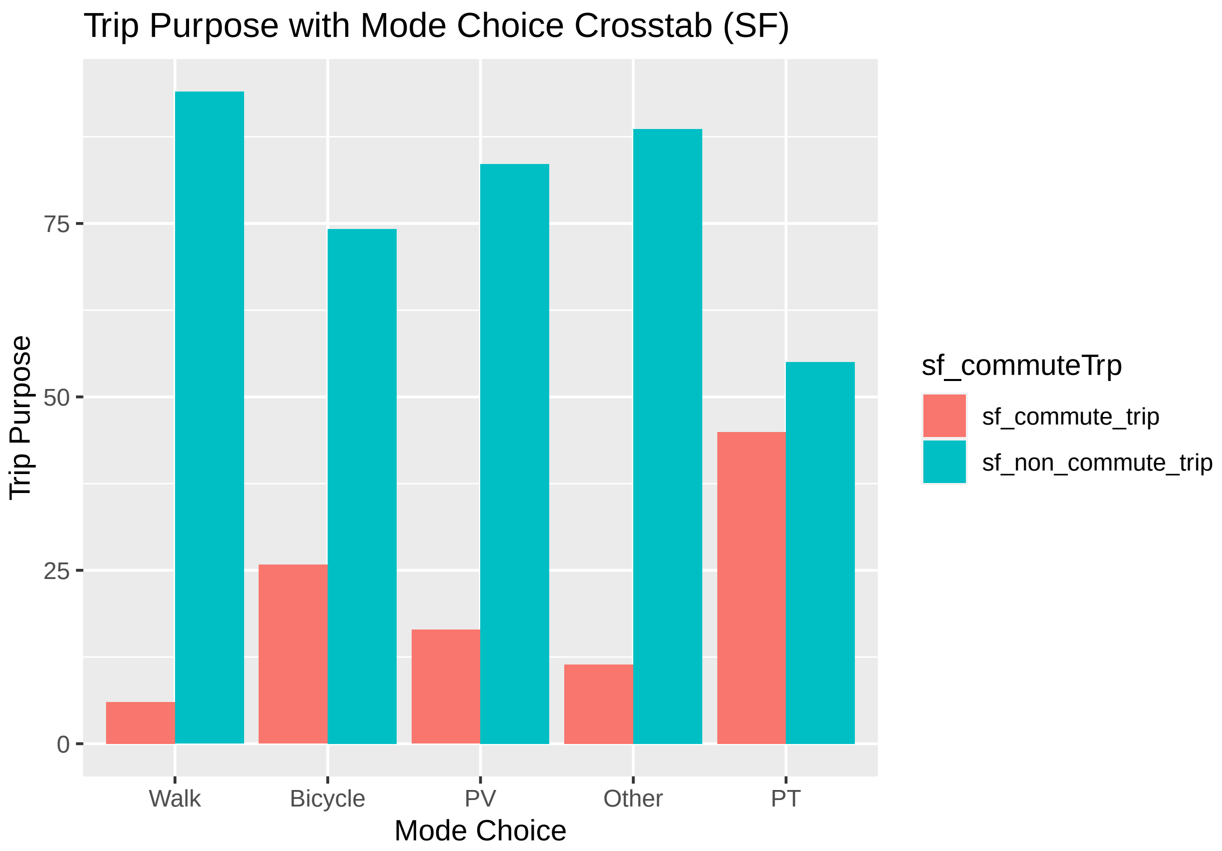
Since the data set being used is “trippub,” the variable for population density is “OTPPOPDN,” and not “HTPPOPDN” from “hhpub.”

1.1.5 A crosstab (bivariate table) of trip purpose with mode choice

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

 sf_trips <- sf_trips %>%
  mutate(sf_commuteTrp = fct_collapse(WHYTRP90,
                                   sf_commute_trip = "To/From Work",
                                   sf_non_commute_trip = c("Work-Related Business","Shopping","Other Fa
                                   Missing = "Refused / Don't Know"))


 sf_trips <-  sf_trips %>%
  filter(sf_commuteTrp != "Missing")


crosstab(df = sf_trips,
         x = sf_mode_short,
         y = sf_commuteTrp,
         weight = WTTRDFIN)
```

```
## # A tibble: 5 x 4
##   sf_mode_short sf_commute_trip sf_non_commute_trip          n
##   <fct>                   <dbl>               <dbl>      <dbl>
## 1 Walk                     6.01                94.0 1258811618.
## 2 Bicycle                 25.8                 74.2  125066279.
## 3 PV                      16.4                 83.6 4701741453.
## 4 Other                   11.4                 88.6  181934299.
## 5 PT                      45.0                 55.0  455824716.
```

```
crosstab(
  df = sf_trips,
  x = sf_mode_short,
  y = sf_commuteTrp,
  weight = WTTRDFIN,
  format = "long"
) %>%
  ggplot(aes(sf_mode_short, pct, fill = sf_commuteTrp)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "Mode Choice", y = "Trip Purpose", title = "Trip Purpose with Mode Choice Crosstab (SF)")
```



1.1.6 Public transit use (PTUSED) by census tract level population density (HTPPOPDN)—You need to use the NHTS Person File for this analysis, ‘perpub.sav

```
levels(sf_per$PTUSED)
```

```
##  [1] "Not ascertained"        "I don’t know"           "I prefer not to answer"
##  [4] "0"                      "1"                      "2"
##  [7] "3"                      "4"                      "5"
## [10] "6"                      "7"                      "8"
## [13] "9"                      "10"                     "11"
## [16] "12"                     "13"                     "14"
## [19] "15"                     "16"                     "17"
## [22] "18"                     "19"                     "20"
## [25] "21"                     "22"                     "23"
## [28] "24"                     "25"                     "26"
## [31] "27"                     "28"                     "29"
## [34] "30"
```

The result can be further refined by categorizing population densities as per the standard incorporated in 1.4, for “OTPPOPDN”. For verification that both of the variables contain equivalent strings, a simple check is performed below.

```
levels(sf_per$HTPPOPDN)
```

```
## [1] "Not ascertained" "1,000-1,999"     "10,000-24,999"   "100-499"
## [5] "2,000-3,999"     "25,000-999,999"  "0-99"            "4,000-9,999"
## [9] "500-999"
```

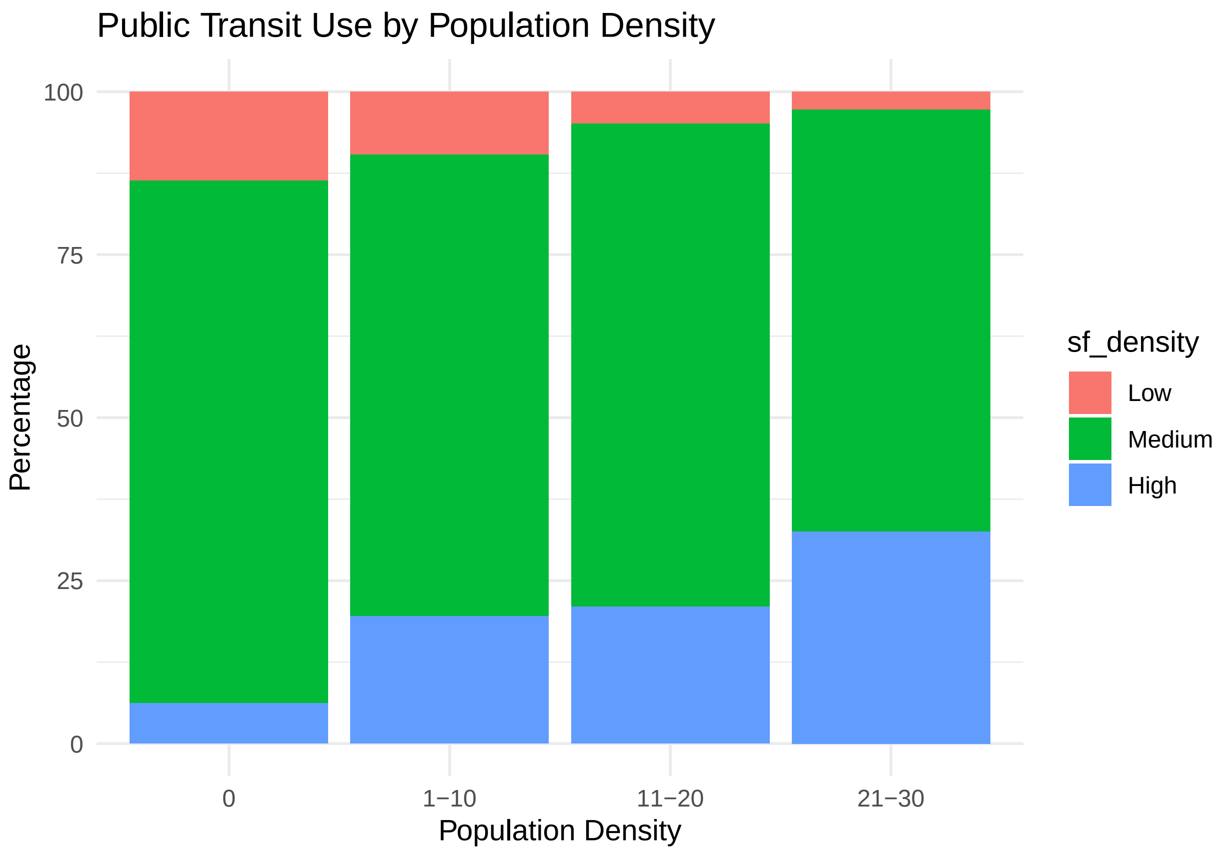
```
levels(sf_trips$OTPPOPDN)
```

```
## [1] "Not ascertained" "1,000-1,999"     "10,000-24,999"   "100-499"
## [5] "2,000-3,999"     "25,000-999,999"  "0-99"            "4,000-9,999"
## [9] "500-999"
```

```
sf_pt <- sf_per %>%
  mutate(sf_density = fct_collapse(HTPPOPDN,
                                   "Low" = c("0-99","100-499","1,000-1,999","500-999"),
                                   "Medium" = c("2,000-3,999","4,000-9,999","10,000-24,999"),
                                   "High" = "25,000-999,999",
                                   Missing = "Not ascertained")) %>%

  filter(sf_density != "Missing") %>%
  mutate(sf_ptused = fct_collapse(PTUSED,
                                  "0" = "0",
                                  "1-10" = c("1","2","3","4","5","6","7","8","9","10"),
                                  "11-20" = c("11","12","13","14","15","16","17","18","19","20"),
                                  "21-30" = c("21","22","23","24","25","26","27","28","29","30"),
                                  Missing = c("Not ascertained","I don't know","I prefer not to answer"
sf_pt <-  sf_pt %>%
  filter(sf_ptused != "Missing") %>%

count(sf_ptused, sf_density, wt = WTPERFIN) %>%
  group_by(sf_ptused) %>%
  mutate(sf_per_ptused = prop.table(n)*100)
```

```
ggplot(data=sf_pt, aes(x=sf_ptused, y=sf_per_ptused, fill = sf_density))+
  geom_histogram(binwidth = 2500, stat='identity')+
  labs(title = "Public Transit Use by Population Density", x = "Population Density", y = "Percentage")+
  theme_minimal()
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



# 1.2 Car Ownership

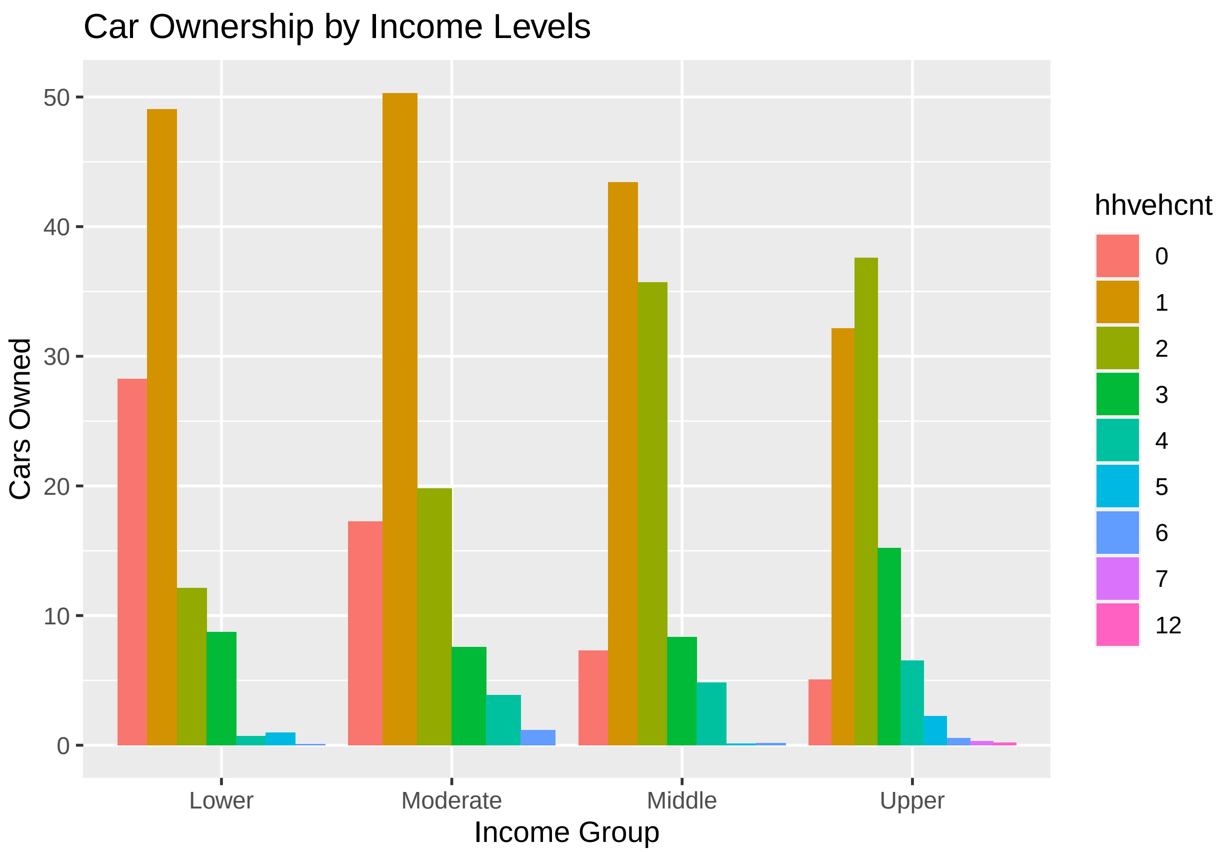
1.2.1 Car Ownership based on Household Income Levels

```
sf_hh <- sf_hh %>%
  mutate(sf_incomegrp = 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(sf_incomegr
```

```
crosstab(df = sf_hh,
         x = sf_incomegrp,
         y = hhvehcnt,
         weight = wthhfin)
```

```
## # A tibble: 4 x 11
##   sf_incomegrp   `0`   `1`   `2`   `3`   `4`   `5`    `6`   `7`  `12`       n
##   <fct>        <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
## 1 Lower        28.3   49.1  12.1  8.74 0.717 0.978 0.0870 0     0      330507.
## 2 Moderate     17.3   50.3  19.8  7.57 3.87  0     1.16   0     0      139140.
## 3 Middle        7.31  43.4  35.7  8.37 4.83  0.131 0.180  0     0      224262.
## 4 Upper         5.09  32.2  37.6 15.2  6.55  2.26  0.555  0.327 0.196 1175446.
```

```
crosstab(df = sf_hh,
         x = sf_incomegrp,
         y = hhvehcnt,
         weight = wthhfin,
         format = "long") %>%

  ggplot(aes(sf_incomegrp, pct, fill = hhvehcnt)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "Income Group", y = "Cars Owned", title = "Car Ownership by Income Levels")
```



An unexpected observation is the non-negligible amount of lower income group population owning 5+ vehicles.

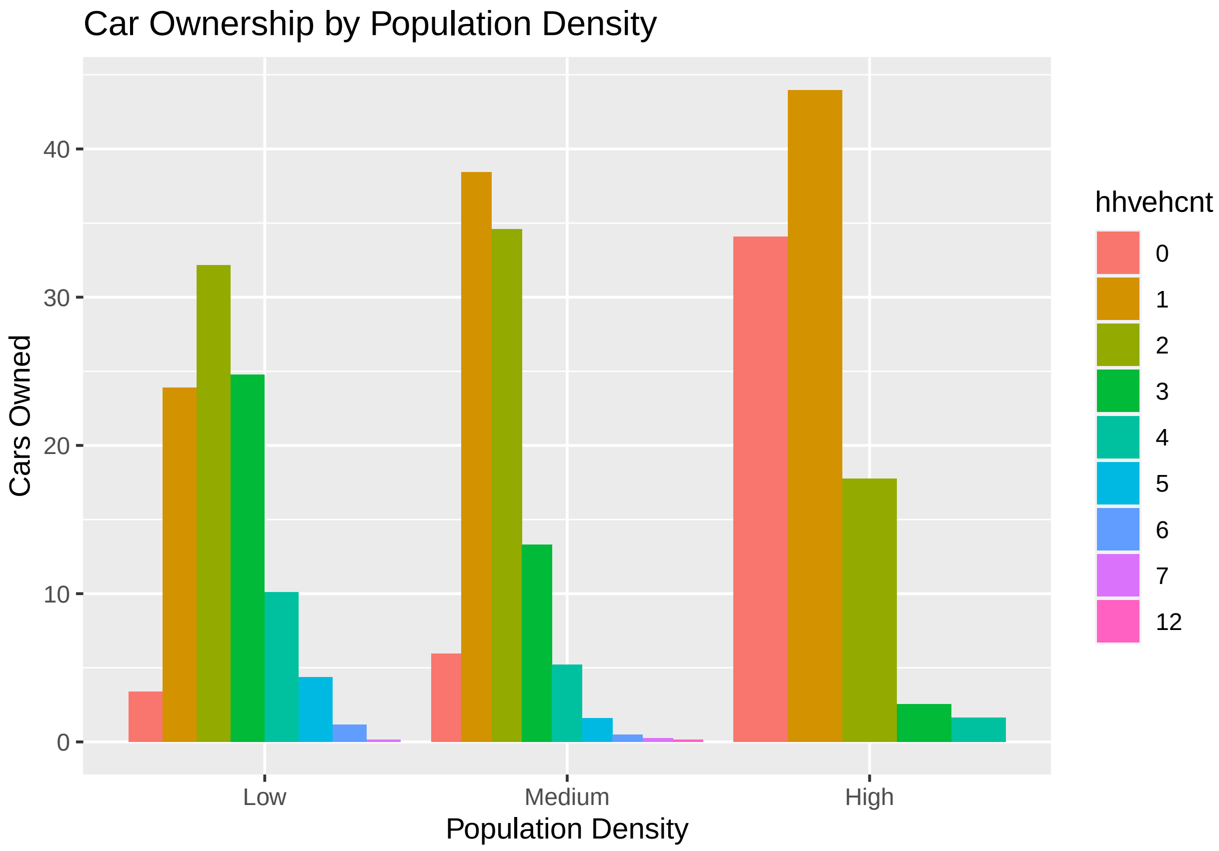
1.2.2 Car Ownership by Population Density

```
sf_hh <- sf_hh %>%
  mutate(sf_density = fct_collapse(htppopdn,
                                   "Low"  = c("0-99","100-499","1,000-1,999","500-999"),
                                   "Medium" = c("2,000-3,999","4,000-9,999","10,000-24,999"),
                                   "High" = "25,000-999,999",
                                   Missing = "Not ascertained"))
```

```
crosstab(df = sf_hh,
         x = sf_density,
         y = hhvehcnt,
         weight = wthhfin)
```

```
## # A tibble: 3 x 11
##   sf_density   `0`   `1`   `2`   `3`   `4`   `5`   `6`   `7`  `12`        n
##   <fct>      <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>    <dbl>
## 1 Low         3.39  23.9  32.1 24.8  10.1   4.37 1.16  0.162 0      189631.
## 2 Medium      5.96  38.4  34.6 13.3   5.21  1.59 0.484 0.259 0.168 1369583.
## 3 High       34.1   44.0  17.8  2.53  1.64  0    0     0     0      310141.
```

```
crosstab(df = sf_hh,
         x = sf_density,
         y = hhvehcnt,
         weight = wthhfin,
         format = "long") %>%

  ggplot(aes(sf_density, pct, fill = hhvehcnt)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "Population Density", y = "Cars Owned", title = "Car Ownership by Population Density")
```



# 1.3 Travel Demand

1.3.1 Number of Trips on Travel Day by Car Ownership

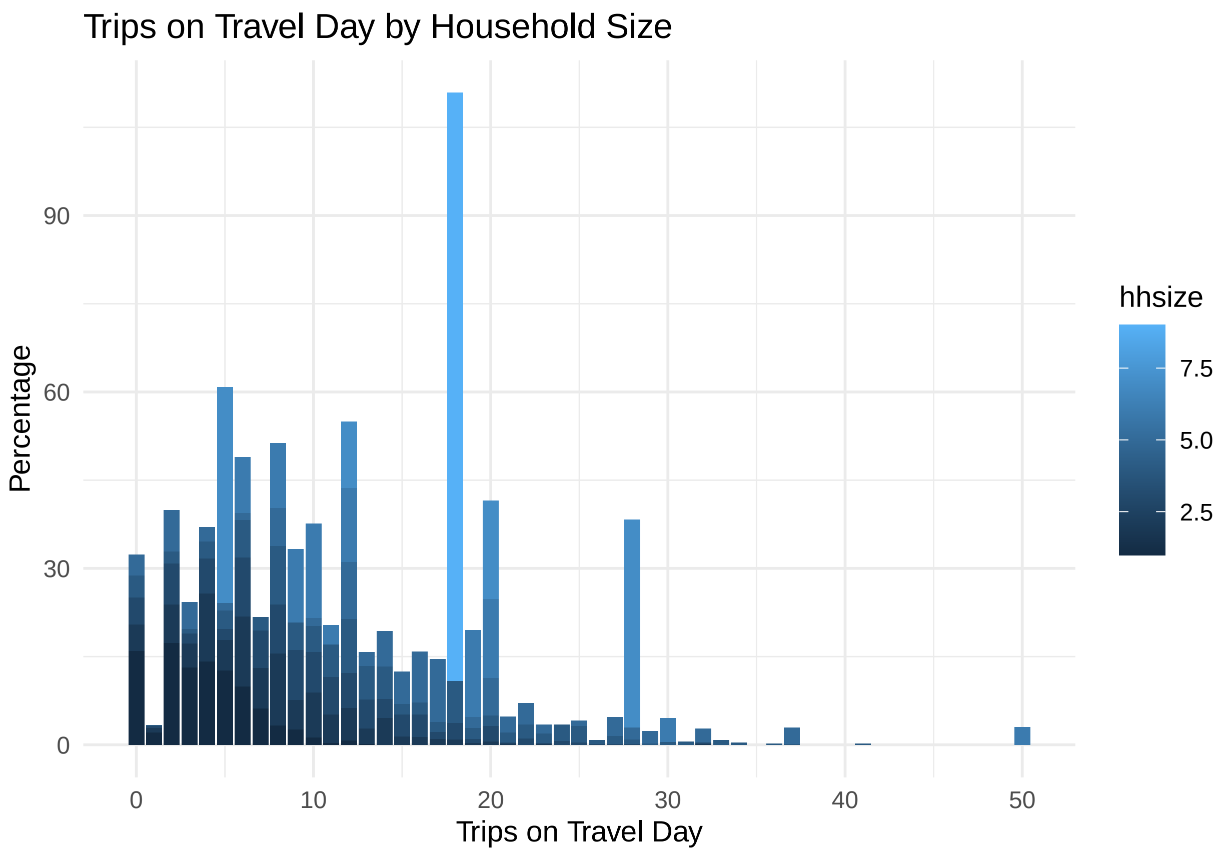
```
# sf_hh <- sf_hh %>%
  # mutate(sf_cars = fct_collapse(hhvehcnt,
    #                             Zero = 0,
     #                            One = 1,
      #                           Two = 2,
       #                          "Three+" = c(3,4,5,6,7,8,9,10,11,12),
        #                         Missing = "Not ascertained")) %>%
  # filter(sf_cars != "Missing")

# I cannot get "mutate" or even "transmute" to collapse the HHVEHCNT column into simplified categories.
```

```
#sf_hh <- sf_tripveh %>%
 # mutate(sf_cars = fct_collapse(hhvehcnt,
  #                                 "Zero" = 0,
   #                                "One" = 1,
    #                               "Two" = 2,
     #                          "Three or more" = c(3,4,5,6,7,8,9,10,11,12))) %>%

# I cannot get "mutate" or even "transmute" to collapse the HHVEHCNT column into simplified categories.

#count(cnttdhh, hhvehcnt, wt = wthhfin) %>%
 # group_by(hhvehcnt) %>%
  # mutate(sf_per_tripveh = prop.table(n)*100)

# ggplot(data=sf_hh, aes(x=cnttdhh, y=sf_per_tripveh, fill = hhvehcnt))+
  # geom_histogram(binwidth = 2500, stat='identity')+
  # labs(title = "Trips on Travel Day by Car Ownership", x = "Trips on Travel Day", y = "Percentage")+
  # theme_minimal()
```

1.3.2 Number of Trips on Travel Day by Household Size

```
sf_hh %>%
  count(cnttdhh, hhsize, wt = wthhfin) %>%
  group_by(hhsize) %>%
  mutate(sf_per_triphh = prop.table(n)*100) %>%

  ggplot(aes(x=cnttdhh, y=sf_per_triphh, fill = hhsize))+
  geom_histogram(binwidth = 2500, stat='identity')+
  labs(title = "Trips on Travel Day by Household Size", x = "Trips on Travel Day", y = "Percentage")+
  theme_minimal()
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



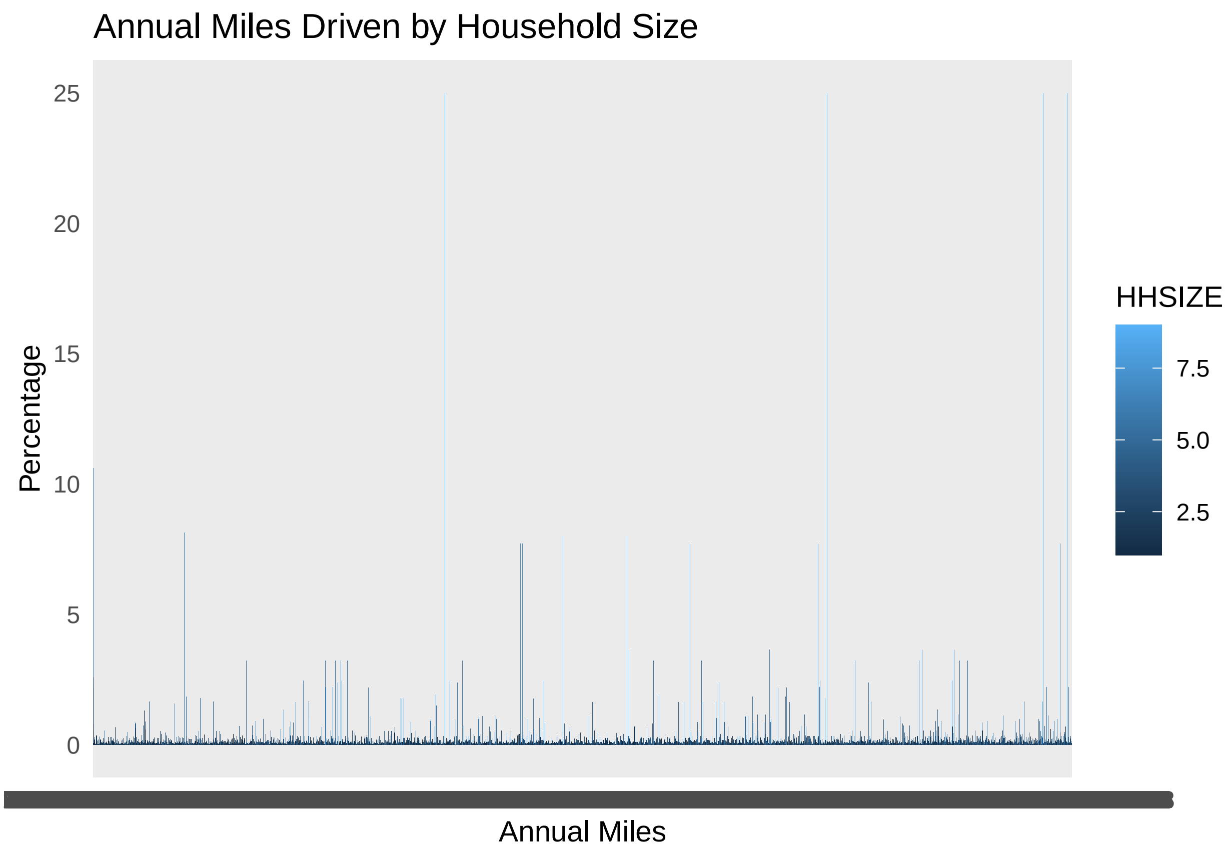
```
# Note to self: When running ggplot2 with %>% pipeline, do not specify data set (i.e. ggplot("data = x"

# Query: How to adjust scale so as to not return 2.5, 7.5, etc.?
```

1.3.3 Number of Trips on Travel Day by Household Size

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

```
sf_veh %>%
  count(sf_bestmile, HHSIZE, wt = WTHHFIN) %>%
  group_by(HHSIZE) %>%
  mutate(sf_per_bestmile = prop.table(n)*100) %>%

  ggplot(aes(x=sf_bestmile, y=sf_per_bestmile, fill = HHSIZE))+
  geom_histogram(binwidth = 2500, stat='identity')+
  labs(title = "Annual Miles Driven by Household Size", x = "Annual Miles", y = "Percentage")+
  theme_minimal()
```

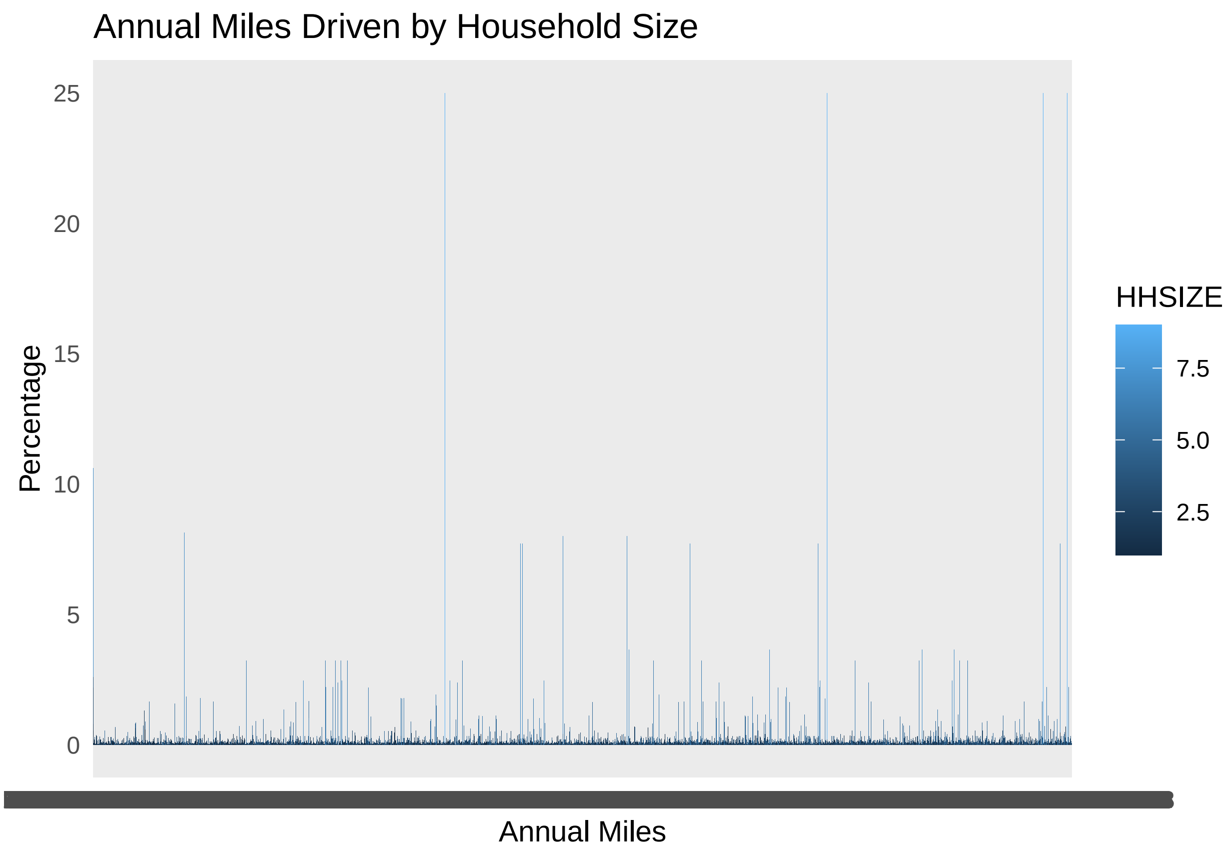
```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



```
agg_veh <- aggregate(as.numeric(as.character(sf_bestmile))~HOUSEID + HHFAMINC + WTHHFIN + HBHUR + HHSIZ

agg_veh <- agg_veh %>% rename("HHVMT" = "as.numeric(as.character(sf_bestmile))")
```

```
sf_veh %>%
  count(sf_bestmile, HHSIZE, wt = WTHHFIN) %>%
  group_by(HHSIZE) %>%
  mutate(sf_per_bestmile = prop.table(n)*100) %>%

  ggplot(aes(x=sf_bestmile, y=sf_per_bestmile, fill = HHSIZE))+
  geom_histogram(binwidth = 2500, stat='identity')+
  labs(title = "Annual Miles Driven by Household Size", x = "Annual Miles", y = "Percentage")+
  theme_minimal()
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

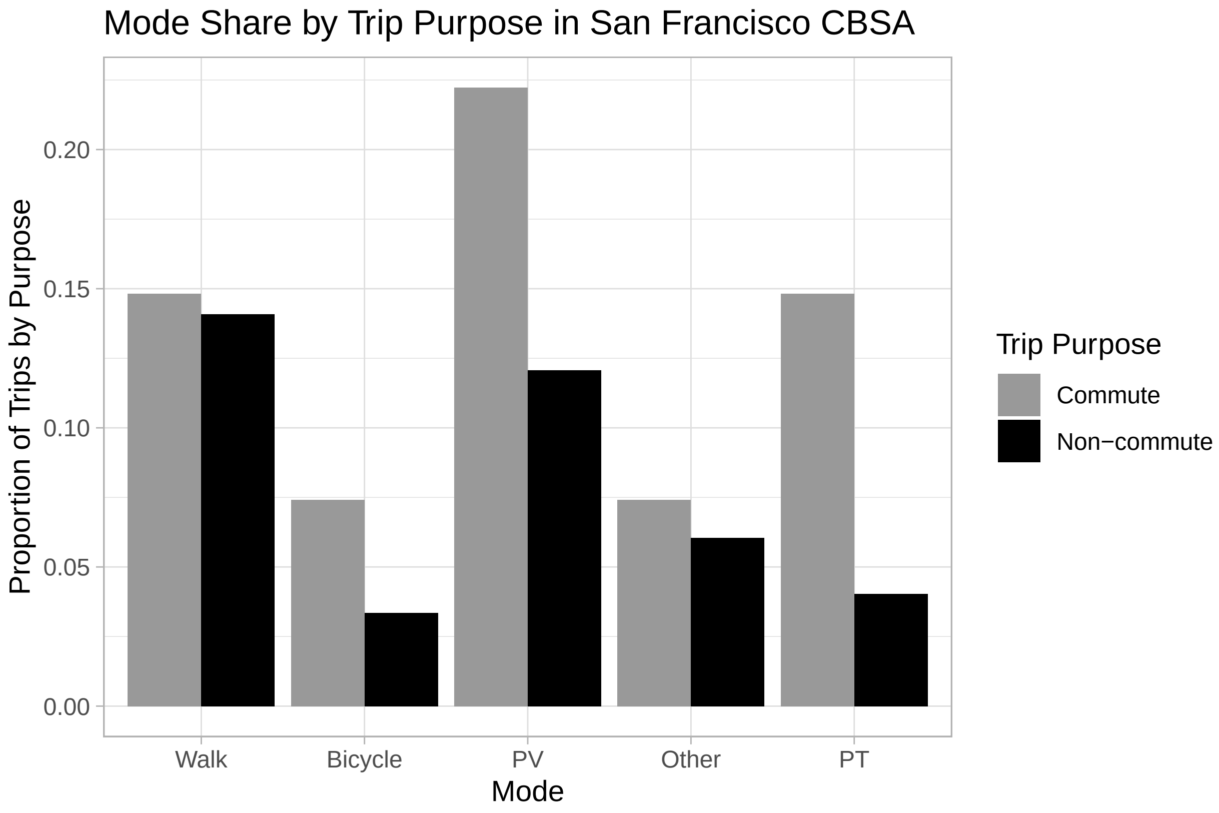


```
# Query: Despite aggregating the sf_bestmile variable, I am getting the same overly-complicated result.
```

1.3.3 Three Additional Crosstabs (WIP)

```
sf_mode_share_commute <- sf_trips %>%
  mutate(sf_commuteTrp = fct_collapse(WHYTRP90,
                                      sf_commute_trip = "To/From Work",
                                      sf_non_commute_trip = c("Work-Related Business","Shopping","Other Fa
                                      Missing = "Refused / Don't Know")) %>%
  count(sf_commuteTrp, sf_mode_short, sf_wt = WTTRDFIN)%>%
  group_by(sf_commuteTrp) %>%
  mutate(sf_per = prop.table(n)*100)

sf_mode_share_commute <- sf_mode_share_commute %>%
  filter(sf_commuteTrp != "Missing")

#Error Log: Needed to filter out "Missing" values for correct visual representation.
```

```
ggplot(sf_mode_share_commute, aes(x = sf_mode_short, y = sf_per)) +
  geom_bar(aes(fill = sf_commuteTrp), position = "dodge", stat = 'identity') +
  labs(x = "Mode", y = "Proportion of Trips by Purpose", title = "Mode Share by Trip Purpose in San Fra
       caption = "Source: NHTS (2017)", fill= "Trip Purpose") +
  scale_fill_manual(values=c("#999999", "#000000", "#999999"), name="Trip Purpose",
                       breaks=c("sf_commute_trip","sf_non_commute_trip"),
                       labels=c("Commute", "Non-commute")) + theme_light()
```

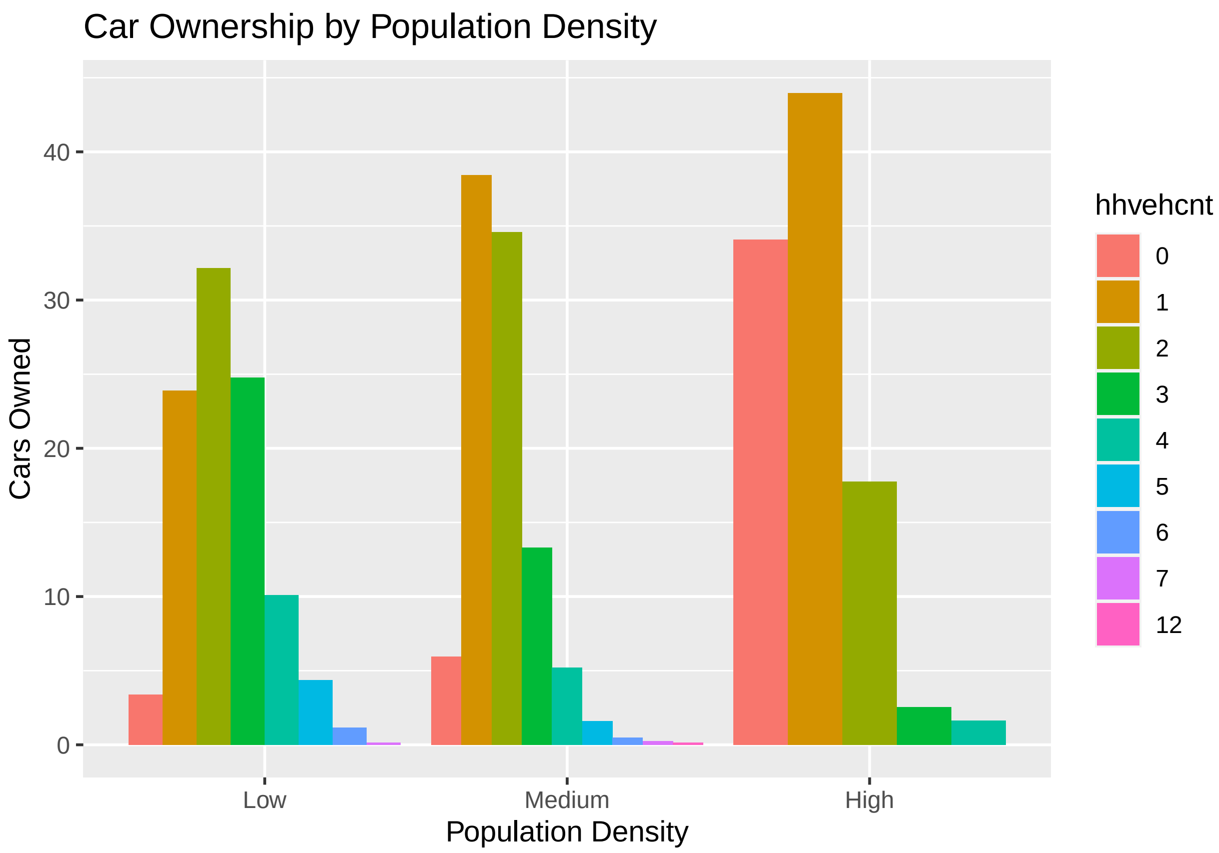


Source: NHTS (2017)

```
crosstab(df = sf_hh,
         x = sf_density,
         y = hhvehcnt,
         weight = wthhfin)
```

```
## # A tibble: 3 x 11
##   sf_density   `0`   `1`   `2`   `3`   `4`   `5`   `6`   `7`  `12`        n
##   <fct>      <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>    <dbl>
## 1 Low         3.39  23.9  32.1 24.8  10.1   4.37 1.16  0.162 0      189631.
## 2 Medium      5.96  38.4  34.6 13.3   5.21  1.59 0.484 0.259 0.168 1369583.
## 3 High       34.1   44.0  17.8  2.53  1.64  0    0     0     0      310141.
```

```
crosstab(df = sf_hh,
         x = sf_density,
         y = hhvehcnt,
         weight = wthhfin,
         format = "long") %>%

  ggplot(aes(sf_density, pct, fill = hhvehcnt)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "Population Density", y = "Cars Owned", title = "Car Ownership by Population Density")
```


Car Ownership by Population Density
Cars Owned
0
10
20
30
40
Low
Medium
High
Population Density
hhvehcnt
0
1
2
3
4
5
6
7
12