GOV 51: PSET 01

Daniel Cabrera

GitHub Repository URL: https://github.com/daniel-e-cabrera/gov51-ps1.git

Question 2.1

```
#Loading in needed libraries and dataset
library(tidyverse)
```

```
Warning: package 'ggplot2' was built under R version 4.4.3
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   4.0.2     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

```
acs2024 <- read.csv("data/raw/acs2024.csv")
```

```
#Viewing the data
head(acs2024)
```

```
  YEAR SAMPLE SERIAL    CBSERIAL HHWT   CLUSTER STATEFIP GQ PERNUM PERWT SEX
1 2024 202401      1 2.02401e+12   41 2.024e+12        1  3      1    41   1
2 2024 202401      2 2.02401e+12   52 2.024e+12        1  3      1    52   1
3 2024 202401      3 2.02401e+12   31 2.024e+12        1  3      1    31   2
4 2024 202401      4 2.02401e+12    4 2.024e+12        1  4      1     4   2
5 2024 202401      5 2.02401e+12   19 2.024e+12        1  3      1    19   1
6 2024 202401      6 2.02401e+12   64 2.024e+12        1  4      1    64   1
  AGE EDUC EDUCD EMPSTAT EMPSTATD INCTOT TRANTIME
1  59    2    25       3       30  18500        0
2  43    6    64       3       30      0        0
3  75    6    63       3       30  27100        0
4  22    7    71       3       30   1000        0
5  51    6    63       3       30      0        0
6  20    7    71       3       30      0        0
```

Question 2.2

TRANTIME is a 3-digit numeric variable reporting the total amount of time, in minutes, that it usually took the respondent to get from home to work last week. Therefore, a value of 0 in TRANTIME means "N/A". Or, in other words, that the individual works from home, does not work, or for any other reason does not commute to work.

Additionally, PERWT is a 6-digit numeric variable indicating how many people in the U.S. population are represented by a given person in an IPUMS sample. More specifically, PERWT tells us how much each person weighs in the data. As mentioned in lecture, sometimes we weigh some individuals (such as minorities) more heavily than others because we lack more of them needed to be representative of our population. For example, if we lack Hispanics in our sample, and needed more to be representative of the U.S. population, then we weigh the Hispanics in the sample more heavily. Additionally, it should be noted that PERWT has two implied decimals. For example, a PERWT value of 010461 should be interpreted as 104.61. PERWT also will help us later when we want to calculate something that is weighted or something that we want to be nationally representative.

Question 2.3

```r
#Creating New Variables
acs2024 <- acs2024 |>
  mutate(female = case_when(SEX == 2 ~ 1,
                            SEX == 1 ~ 0))

acs2024 <- acs2024 |>
```

```
  mutate(less_hs = case_when(EDUC %in% 0:2 ~ 1, TRUE ~ 0),
         hs_only = case_when(EDUC %in% 3:5 ~ 1, TRUE ~ 0),
         some_college = case_when(EDUC %in% 6:7 ~ 1,  TRUE ~ 0),
         college_only = case_when(EDUC %in% 8:10 ~ 1, TRUE ~ 0),
         advanced_degree = case_when(EDUC == 11 ~ 1, TRUE ~ 0))

acs2024 <- acs2024 |>
  mutate(employed = case_when(EMPSTAT == 1 ~ 1, TRUE ~ 0),
         unemployed = case_when(EMPSTAT == 2 ~ 1,  TRUE ~ 0),
         not_in_labor_force = case_when(EMPSTAT == 3 ~ 1, TRUE ~ 0))

acs2024 <- acs2024 |> mutate (INCTOT_clean = case_when(INCTOT %in% c(9999998, 9999999) ~ NA_
```

```
#Viewing the data with new variables
head(acs2024)
```

```
  YEAR SAMPLE SERIAL    CBSERIAL HHWT   CLUSTER STATEFIP GQ PERNUM PERWT SEX
1 2024 202401      1 2.02401e+12   41 2.024e+12        1  3      1    41   1
2 2024 202401      2 2.02401e+12   52 2.024e+12        1  3      1    52   1
3 2024 202401      3 2.02401e+12   31 2.024e+12        1  3      1    31   2
4 2024 202401      4 2.02401e+12    4 2.024e+12        1  4      1     4   2
5 2024 202401      5 2.02401e+12   19 2.024e+12        1  3      1    19   1
6 2024 202401      6 2.02401e+12   64 2.024e+12        1  4      1    64   1
  AGE EDUC EDUCD EMPSTAT EMPSTATD INCTOT TRANTIME female less_hs hs_only
1  59    2    25       3       30  18500        0      0       1       0
2  43    6    64       3       30      0        0      0       0       0
3  75    6    63       3       30  27100        0      1       0       0
4  22    7    71       3       30   1000        0      1       0       0
5  51    6    63       3       30      0        0      0       0       0
6  20    7    71       3       30      0        0      0       0       0
  some_college college_only advanced_degree employed unemployed
1            0            0               0        0          0
2            1            0               0        0          0
3            1            0               0        0          0
4            1            0               0        0          0
5            1            0               0        0          0
6            1            0               0        0          0
  not_in_labor_force INCTOT_clean
1                  1        18500
2                  1            0
3                  1        27100
4                  1         1000
```

```
5                   1             0
6                   1             0
```

Question 2.4

If we treated 0 or 99 which in this data set, represents "N/A" rather than a true value, then any manipulation of the data is going to be incorrect. More specifically, a very low value of 0 or a high value of 99 will impact our summary statistics such as our averages, ranges, medians, etc. However, we don't want them to since they do not represent a true value.

For example, with our SEX variable, we have the numeric values of either 1 for male, 2 for female, or 9 for N/A. Therefore, we have to remove the variable 9 because it does not represent a true value. If we keep it in our data set, our summary statistics will be incorrectly skewed up. In other words, all of our values will be incorrectly higher than they should be.

Question 2.5

```
#Creating summary statistics table
stats <- data.frame(
  Variable = c("Age",
               "Female",
               "Less Than Highschool",
               "High School Only",
               "Some College",
               "College Only",
               "Advanced Degree",
               "Employed",
               "Unemployed",
               "Not in Labor Force",
               "Commute Time (mins)",
               "Total Income ($)"),

  N = c(length(acs2024$AGE),
        length(acs2024$female),
        length(acs2024$less_hs),
        length(acs2024$hs_only),
        length(acs2024$some_college),
        length(acs2024$college_only),
        length(acs2024$advanced_degree),
        length(acs2024$employed),
        length(acs2024$unemployed),
```

```
      length(acs2024$not_in_labor_force),
      length(acs2024$TRANTIME),
      sum(!is.na(acs2024$INCTOT_clean))),

Mean = c(mean(acs2024$AGE, na.rm = TRUE),
         mean(acs2024$female, na.rm = TRUE),
         mean(acs2024$less_hs, na.rm = TRUE),
         mean(acs2024$hs_only, na.rm = TRUE),
         mean(acs2024$some_college, na.rm = TRUE),
         mean(acs2024$college_only, na.rm = TRUE),
         mean(acs2024$advanced_degree, na.rm = TRUE),
         mean(acs2024$employed, na.rm = TRUE),
         mean(acs2024$unemployed, na.rm = TRUE),
         mean(acs2024$not_in_labor_force, na.rm = TRUE),
         mean(acs2024$TRANTIME, na.rm = TRUE),
         mean(acs2024$INCTOT_clean, na.rm = TRUE)),

SD = c(sd(acs2024$AGE, na.rm = TRUE),
       sd(acs2024$female, na.rm = TRUE),
       sd(acs2024$less_hs, na.rm = TRUE),
       sd(acs2024$hs_only, na.rm = TRUE),
       sd(acs2024$some_college, na.rm = TRUE),
       sd(acs2024$college_only, na.rm = TRUE),
       sd(acs2024$advanced_degree, na.rm = TRUE),
       sd(acs2024$employed, na.rm = TRUE),
       sd(acs2024$unemployed, na.rm = TRUE),
       sd(acs2024$not_in_labor_force, na.rm = TRUE),
       sd(acs2024$TRANTIME, na.rm = TRUE),
       sd(acs2024$INCTOT_clean, na.rm = TRUE)),

Min = c(min(acs2024$AGE, na.rm = TRUE),
        min(acs2024$female, na.rm = TRUE),
        min(acs2024$less_hs, na.rm = TRUE),
        min(acs2024$hs_only, na.rm = TRUE),
        min(acs2024$some_college, na.rm = TRUE),
        min(acs2024$college_only, na.rm = TRUE),
        min(acs2024$advanced_degree, na.rm = TRUE),
        min(acs2024$employed, na.rm = TRUE),
        min(acs2024$unemployed, na.rm = TRUE),
        min(acs2024$not_in_labor_force, na.rm = TRUE),
        min(acs2024$TRANTIME, na.rm = TRUE),
        min(acs2024$INCTOT_clean, na.rm = TRUE)),
```

```
  Max = c(max(acs2024$AGE, na.rm = TRUE),
          max(acs2024$female, na.rm = TRUE),
          max(acs2024$less_hs, na.rm = TRUE),
          max(acs2024$hs_only, na.rm = TRUE),
          max(acs2024$some_college, na.rm = TRUE),
          max(acs2024$college_only, na.rm = TRUE),
          max(acs2024$advanced_degree, na.rm = TRUE),
          max(acs2024$employed, na.rm = TRUE),
          max(acs2024$unemployed, na.rm = TRUE),
          max(acs2024$not_in_labor_force, na.rm = TRUE),
          max(acs2024$TRANTIME, na.rm = TRUE),
          max(acs2024$INCTOT_clean, na.rm = TRUE)))

knitr::kable(stats, digits = 2,
             caption = "Table 1: Summary Statistics for 2024 ACS Sample")
```

Table 1: Table 1: Summary Statistics for 2024 ACS Sample

Variable	N	Mean	SD	Min	Max
Age	3422888	43.39	24.03	0	96
Female	3422888	0.51	0.50	0	1
Less Than Highschool	3422888	0.18	0.39	0	1
High School Only	3422888	0.06	0.24	0	1
Some College	3422888	0.40	0.49	0	1
College Only	3422888	0.24	0.43	0	1
Advanced Degree	3422888	0.11	0.32	0	1
Employed	3422888	0.47	0.50	0	1
Unemployed	3422888	0.02	0.14	0	1
Not in Labor Force	3422888	0.35	0.48	0	1
Commute Time (mins)	3422888	10.81	19.83	0	195
Total Income ($)	2912790	54654.71	80080.40	-11500	1945000

Question 2.6

The maximum commute time of 195 minutes, or 3 hours and 15 minutes make sense for a daily commute time. At least for me, I am from rural Oregon and in the town Bandon, Oregon where I am from, many, many people commute almost two hours to get to work everyday from neighboring towns like Port Orford and Myrtle Point. Additionally, we expect our maximum commute time to be relatively high because it is quite literally the highest commute time value

we have of a very, very large sample (3422888). With this many samples, it is very likely that we are going to have at least one sample that is an extreme.

Question 3.1

```
#Calculating teh number of commuters with a commute time of 0
zero_commute = sum(acs2024$TRANTIME == 0)
print(zero_commute)
```

```
[1] 2062945
```

```
#Calculating the percent of commuters with a commute time of 0
percent_zero_commute <- (zero_commute / nrow(acs2024)) * 100
print(percent_zero_commute)
```

```
[1] 60.26914
```

These individuals (2062945 or about 60.3% of the total of respondents) with commune times of 0 are either people who work from home, individuals who do not work, their commute times are under a minute total, or are N/A.

Question 3.2

```
#Filtering the data set to only include those with a commute time above 0
actual_commuters_subset <- acs2024 |>
  filter(TRANTIME > 0)
```

I chose to filter out all of the individuals who have a commute time that is more than 0. Simply put, if we want to study the commute time of individuals, then we need people with commute times to study. In other words, we need to only keep in our data set people who have commute times above 0 minutes. A similar way to think about this is if we wanted to study Harvard students. If we did, then we would need to filter from a data set the students who go to Harvard.

Question 3.3

```r
#Calculating the total number of commuters with a commute time above 0
commuters_above_zero_total <- actual_commuters_subset |>
  nrow()
print(commuters_above_zero_total)
```

```
[1] 1359943
```

```r
#Finding the min of my new filtered data set
commuters_above_zero_min <- min(actual_commuters_subset$TRANTIME)
print(commuters_above_zero_min)
```

```
[1] 1
```

After we have filtered out people who have a commute time above 0, we go from 3422888 respondents to only 1359943, losing about 60.3% of of our original group. The new minimum is now a commute time of 1 minute. This makes sense because we removed all the 0s from our original data set, which means that that next lowest possible commute time someone could have is 1. And surely enough, it appears that someone does have a commute time of 1 in our new subset.

Question 3.4

```r
#Creating summary statistics table with our new actual_commuters_subset
stats <- data.frame(
  Variable = c("Age",
               "Female",
               "Less Than Highschool",
               "High School Only",
               "Some College",
               "College Only",
               "Advanced Degree",
               "Employed",
               "Unemployed",
               "Not in Labor Force",
               "Commute Time (mins)",
               "Total Income ($)"),

  N = c(length(actual_commuters_subset$AGE),
```

```
      length(actual_commuters_subset$female),
      length(actual_commuters_subset$less_hs),
      length(actual_commuters_subset$hs_only),
      length(actual_commuters_subset$some_college),
      length(actual_commuters_subset$college_only),
      length(actual_commuters_subset$advanced_degree),
      length(actual_commuters_subset$employed),
      length(actual_commuters_subset$unemployed),
      length(actual_commuters_subset$not_in_labor_force),
      length(actual_commuters_subset$TRANTIME),
      length(actual_commuters_subset$INCTOT_clean)),

Mean = c(mean(actual_commuters_subset$AGE, na.rm = TRUE),
         mean(actual_commuters_subset$female, na.rm = TRUE),
         mean(actual_commuters_subset$less_hs, na.rm = TRUE),
         mean(actual_commuters_subset$hs_only, na.rm = TRUE),
         mean(actual_commuters_subset$some_college, na.rm = TRUE),
         mean(actual_commuters_subset$college_only, na.rm = TRUE),
         mean(actual_commuters_subset$advanced_degree, na.rm = TRUE),
         mean(actual_commuters_subset$employed, na.rm = TRUE),
         mean(actual_commuters_subset$unemployed, na.rm = TRUE),
         mean(actual_commuters_subset$not_in_labor_force, na.rm = TRUE),
         mean(actual_commuters_subset$TRANTIME, na.rm = TRUE),
         mean(actual_commuters_subset$INCTOT_clean, na.rm = TRUE)),

SD = c(sd(actual_commuters_subset$AGE, na.rm = TRUE),
       sd(actual_commuters_subset$female, na.rm = TRUE),
       sd(actual_commuters_subset$less_hs, na.rm = TRUE),
       sd(actual_commuters_subset$hs_only, na.rm = TRUE),
       sd(actual_commuters_subset$some_college, na.rm = TRUE),
       sd(actual_commuters_subset$college_only, na.rm = TRUE),
       sd(actual_commuters_subset$advanced_degree, na.rm = TRUE),
       sd(actual_commuters_subset$employed, na.rm = TRUE),
       sd(actual_commuters_subset$unemployed, na.rm = TRUE),
       sd(actual_commuters_subset$not_in_labor_force, na.rm = TRUE),
       sd(actual_commuters_subset$TRANTIME, na.rm = TRUE),
       sd(actual_commuters_subset$INCTOT_clean, na.rm = TRUE)),

Min = c(min(actual_commuters_subset$AGE, na.rm = TRUE),
        min(actual_commuters_subset$female, na.rm = TRUE),
        min(actual_commuters_subset$less_hs, na.rm = TRUE),
        min(actual_commuters_subset$hs_only, na.rm = TRUE),
```

```
          min(actual_commuters_subset$some_college, na.rm = TRUE),
          min(actual_commuters_subset$college_only, na.rm = TRUE),
          min(actual_commuters_subset$advanced_degree, na.rm = TRUE),
          min(actual_commuters_subset$employed, na.rm = TRUE),
          min(actual_commuters_subset$unemployed, na.rm = TRUE),
          min(actual_commuters_subset$not_in_labor_force, na.rm = TRUE),
          min(actual_commuters_subset$TRANTIME, na.rm = TRUE),
          min(actual_commuters_subset$INCTOT_clean, na.rm = TRUE)),

  Max = c(max(actual_commuters_subset$AGE, na.rm = TRUE),
          max(actual_commuters_subset$female, na.rm = TRUE),
          max(actual_commuters_subset$less_hs, na.rm = TRUE),
          max(actual_commuters_subset$hs_only, na.rm = TRUE),
          max(actual_commuters_subset$some_college, na.rm = TRUE),
          max(actual_commuters_subset$college_only, na.rm = TRUE),
          max(actual_commuters_subset$advanced_degree, na.rm = TRUE),
          max(actual_commuters_subset$employed, na.rm = TRUE),
          max(actual_commuters_subset$unemployed, na.rm = TRUE),
          max(actual_commuters_subset$not_in_labor_force, na.rm = TRUE),
          max(actual_commuters_subset$TRANTIME, na.rm = TRUE),
          max(actual_commuters_subset$INCTOT_clean, na.rm = TRUE))
)

knitr::kable(stats, digits = 1,
             caption = "Table 2: Summary Statistics for 2024 ACS Sample (Only Keeping Actual
```

Table 2: Table 2: Summary Statistics for 2024 ACS Sample (Only Keeping Actual Commuters)

Variable	N	Mean	SD	Min	Max
Age	1359943	43.4	15.3	16	96
Female	1359943	0.5	0.5	0	1
Less Than Highschool	1359943	0.0	0.2	0	1
High School Only	1359943	0.0	0.2	0	1
Some College	1359943	0.5	0.5	0	1
College Only	1359943	0.3	0.5	0	1
Advanced Degree	1359943	0.2	0.4	0	1
Employed	1359943	1.0	0.0	1	1
Unemployed	1359943	0.0	0.0	0	0
Not in Labor Force	1359943	0.0	0.0	0	0
Commute Time (mins)	1359943	27.2	23.3	1	195
Total Income ($)	1359943	72992.7	88726.0	-11500	1945000

Question 4.1

```
#creating a histogram for actual_commuters_subset
library(ggplot2)
library(scales)
```

```
Attaching package: 'scales'
```

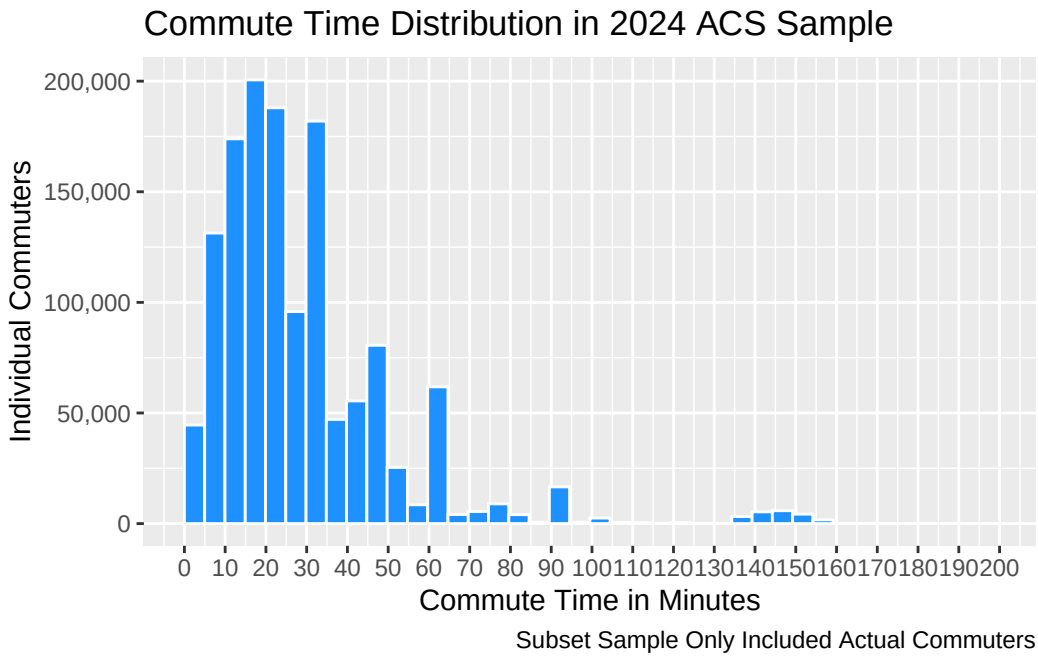
```
The following object is masked from 'package:purrr':

    discard
```

```
The following object is masked from 'package:readr':

    col_factor
```

```
actual_commuters_subset_histogram <- ggplot(data = actual_commuters_subset,
                                            mapping = aes(x = TRANTIME)) +
  geom_histogram(bins = 40,
                 boundary = 0,
                 fill = "dodgerblue",
                 color = "white") +
  labs(x = "Commute Time in Minutes",
       y = "Individual Commuters",
       title = "Commute Time Distribution in 2024 ACS Sample",
       caption = "Subset Sample Only Included Actual Commuters") +
  scale_x_continuous(breaks = seq(0, 200, by = 10)) +
  scale_y_continuous(labels = comma)
print(actual_commuters_subset_histogram)
```



Question 4.2

In my histogram, we can see that that it is very much right skewed, or more specifically, that the majority of actual commuters have a commute time of around 5 minutes to 35 minutes. Notably, there is a dramatic decline in the number of respondents after 35 minutes, and then another large decline after about 65 minutes. Lastly, there are a few outliers visible on the graph around 135 minutes to 160 minutes.

Question 4.3

```
actual_commuters_subset_mean <- mean(actual_commuters_subset$TRANTIME)
print(actual_commuters_subset_mean)
```

```
[1] 27.21505
```

```
actual_commuters_subset_median <- median(actual_commuters_subset$TRANTIME)
print(actual_commuters_subset_median)
```

```
[1] 20
```

From my summary statistics table, I can see that my mean is about 27.2. Additionally, I can calculate my median which is 20. Comparing these two, we an see that the average for actual commuters is higher than the median number (middle value). This is most likely because the mean is taking into account those few outliers we had around the 135-160 minute mark. This tells me that my distribution is right skewed and yes, it does match with what I see in my histogram.

Question 4.4

```
ggsave("/Users/danielcabrera/Desktop/gov51-ps1/data/raw/code/output/commute_histogram.png", p
```

Question 5.1

```
weighted.mean(actual_commuters_subset$TRANTIME, actual_commuters_subset$PERWT)
```

```
[1] 27.19112
```

Question 5.2

The weighted mean is very similar to the unweighted mean I calculated earlier. The difference is only about 0.02.

Question 5.3

The unweighted mean of 27.215 shows us the average of our sample population which is a subset (only having actual commuters) of our original ACS 2024 sample data set. On the other hand, our **weighted commute time takes into account how many people in the actual U.S. population one of our sample individuals represents. By doing this, it can make our sample nationally representative.** More specifically then, the weighted mean of 27.191 tells us the best estimate for the average commute time in the U.S. population (not our sample population). And it does so by using the PERWT variable. As mentioned in lecture, sometimes we weigh some individuals (such as minorities) more heavily than others because we lack more of them needed to be representative of our population. For example, if we lack Hispanics in our sample, and needed more to be representative of the U.S. population, then we weigh the Hispanics in the sample more heavily.