ps1

Felicity Zhang

https://github.com/felicityxz17/gov51-ps1.git

2 Get to Know Your Data

2.1

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.2.0     v readr     2.1.6
v forcats   1.0.1     v stringr   1.5.1
v ggplot2   4.0.2     v tibble    3.3.1
v lubridate 1.9.5     v tidyr     1.3.2
v purrr     1.2.1
-- 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")
```

2.2

```
names(acs2024)
```

```
 [1] "YEAR"     "SAMPLE"   "SERIAL"   "CBSERIAL" "HHWT"     "CLUSTER"
 [7] "STATEFIP" "GQ"       "PERNUM"   "PERWT"    "SEX"      "AGE"
[13] "EDUC"     "EDUCD"    "EMPSTAT"  "EMPSTATD" "INCTOT"   "TRANTIME"
```

```
table(acs2024$EDUC)
```

```
     0      1      2      3      4      5      6      7      8     10
194921 231325 197656  62954  69309  79062 1023716 361037 239131 580925
    11
382852
```

`YEAR`: Year the household was included in the census; 4 digits

`SAMPLE`: IPUMS sample identifier; 6 digits (4 for year, 2 for sample)

`SERIAL`: Household serial number (unique identifying number)

`CBSERIAL`: Household serial number assigned by the Census Bureau

`HHWT`: Household weight (number of households in the U.S. population represented by the given sample household)

`CLUSTER`: Variable for variance estimation and correction; 11 digits

`STATEFIP`: State FIPS (Federal Information Processing Standards) code

`GQ`: Group quarters status (households, group quarters, or vacant units)

`PERNUM`: Person number within their household based on order of appearance on the census

`PERWT`: Person weight (number of people in the U.S. population represented by the given sample person); 2 implied decimals

`SEX`: Male (1) or female (2) or missing/blank (9)

`AGE`: Age in years

`EDUC`: Educational attainment measured by highest year of school or degree completed

`EMPSTAT`: Employment status

`INCTOT`: Total pre-tax personal income in the previous year; 7 digit code (9999999 = N/A)

`TRANTIME`: Total number of minutes to get from home to work in the previous week; 3 digits (000 = N/A)

2.3

```r
# SEX: 1 = Male; 2 = Female; 9 = Missing/blank
acs2024 <- mutate(acs2024, female = ifelse(SEX == 2, 1, 0))

# EDUC: 00-06 = N/A or no schooling through Grade 12; 07-10 = through college; 11 = 5+ years
acs2024 <- mutate(acs2024, less_than_hs = ifelse(EDUC < 06, 1, 0))
acs2024 <- mutate(acs2024, hs_only = ifelse(EDUC == 06, 1, 0))
acs2024 <- mutate(acs2024, some_college = ifelse((EDUC > 06) & (EDUC < 10), 1, 0))
acs2024 <- mutate(acs2024, college_only = ifelse(EDUC == 10, 1, 0))
acs2024 <- mutate(acs2024, advanced_degree = ifelse(EDUC > 10, 1, 0))

# EMPSTAT: 0 = N/A; 1 = Employed; 2 = Unemployed; 3 = Not in labor force
acs2024 <- mutate(acs2024, employed = ifelse(EMPSTAT == 1, 1, 0))
acs2024 <- mutate(acs2024, unemployed = ifelse(EMPSTAT == 2, 1, 0))
acs2024 <- mutate(acs2024, not_in_labor_force = ifelse(EMPSTAT == 3, 1, 0))
```

2.4

Accidentally treating "N/A" codes as real zeros would likely skew summary statistics by misinterpreting them as actual data points. For instance, the variable `TRANTIME` might become right-skewed due to the inclusion of `000` as indicating 0-minute transit time, incorrectly decreasing summary statistics like mean and median while also expanding the leftward range endpoint to include this value that does not otherwise make sense in the context of the data.

2.5

```r
#remove N/A values from inctot
INCTOT_modified = acs2024[acs2024$INCTOT != 9999999 & acs2024$INCTOT != 9999998, "INCTOT"]

acs2024_variable <- c("Age", "Female", "Less than High School", "High School Only", "Some Col

acs2024_n <- c(length(acs2024$AGE), sum(acs2024$female), sum(acs2024$less_than_hs), sum(acs20

acs2024_mean <- c(mean(acs2024$AGE), mean(acs2024$female), mean(acs2024$less_than_hs), mean(a

acs2024_sd <- c(sd(acs2024$AGE), sd(acs2024$female), sd(acs2024$less_than_hs), sd(acs2024$hs_

acs2024_min <- c(min(acs2024$AGE), min(acs2024$female), min(acs2024$less_than_hs), min(acs202

acs2024_max <- c(max(acs2024$AGE), max(acs2024$female), max(acs2024$less_than_hs), max(acs202
```

```
acs2024_stats <- data.frame(acs2024_variable, acs2024_n, acs2024_mean, acs2024_sd, acs2024_m

knitr::kable(acs2024_stats, col.names = c("Variable", "N", "Mean", "Std. Dev.", "Min", "Max"
```

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

Variable	N	Mean	Std. Dev.	Min	Max
Age	3422888	43.39	24.03	0	96
Female	1743242	0.51	0.50	0	1
Less than High School	835227	0.24	0.43	0	1
High School Only	1023716	0.30	0.46	0	1
Some College	600168	0.18	0.38	0	1
College Only	580925	0.17	0.38	0	1
Advanced Degree	382852	0.11	0.32	0	1
Employed	1608021	0.47	0.50	0	1
Unemployed	72315	0.02	0.14	0	1
Not in Labor Force	1192654	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

2.6

The maximum commute time, 195 minutes, does not seem to fall into the range of common values of commute times, but is still reasonable as a maximum duration of a daily commute. A one-way commute of a little over 3 hours is similar to a commute length from Salem, MA to Boston, MA, so the data point would make sense for a worker with a home in a rural area and a job in the city.

3 Who Should Be in Your Analysis?

3.1

```
sum(acs2024$TRANTIME == 0)
```

```
[1] 2062945
```

```r
sum(acs2024$TRANTIME == 0)/length(acs2024$TRANTIME == 0)
```

```
[1] 0.6026914
```

```r
head(acs2024[acs2024$TRANTIME == 0,c("AGE", "EMPSTAT", "INCTOT", "employed", "unemployed", "
```

```
   AGE EMPSTAT INCTOT employed unemployed not_in_labor_force
1   59       3  18500        0          0                   1
2   43       3      0        0          0                   1
3   75       3  27100        0          0                   1
4   22       3   1000        0          0                   1
5   51       3      0        0          0                   1
6   20       3      0        0          0                   1
7   66       3  16400        0          0                   1
8   28       3  30000        0          0                   1
9   18       3   5100        0          0                   1
10  64       3  12000        0          0                   1
11  21       3      0        0          0                   1
12  19       3      0        0          0                   1
13  55       3    600        0          0                   1
14  35       3      0        0          0                   1
16  44       3  30600        0          0                   1
17  19       1  15000        1          0                   0
18  32       3      0        0          0                   1
19  23       3      0        0          0                   1
20  19       3     40        0          0                   1
21  82       3  21400        0          0                   1
```

2,062,945 people out of the total 3,422,888 observations have `TRANTIME == 0`, representing a percentage of 60.26914%. A sample of the data indicates that these data points largely come from people who are not in the workforce, such as young adults (AGE between 18 and 22) and the retired (positive INCTOT from retirement pension and AGE values mostly over 50). However, it is also plausible that these data points might also include some people who are unemployed (as in the 17th sample entry above) or people who are employed but whose jobs do not require a commute, such as remote workers and care workers.

3.2

```r
commuters = acs2024[acs2024$TRANTIME != 0,]

sum(commuters$employed == 0)
```

```
[1] 0
```

Since the goal of this subset is to analyze commute times for people who actually commute to work, we should remove all data points with a transit time of 0, regardless of the reason that they have 0 transit time. All values greater than 0 are valid values for transit time, so the subset should be based on the condition `TRANTIME > 0`. As a sanity check, we can confirm that everyone in this subset is employed.

3.3

```r
length(commuters$TRANTIME)
```

```
[1] 1359943
```

```r
min(commuters$TRANTIME)
```

```
[1] 1
```

1,359,943 observations remain, which is consistent with the remaining number of data points from 3.1. The new minimum `TRANTIME` is 1 minute, which makes sense because it still qualifies as a commute and is reasonable for individuals who might live right next to their workplace.

3.4

```r
#remove N/A values from inctot
INCTOT_modified = commuters[commuters$INCTOT != 9999999 & commuters$INCTOT != 9999998, "INCT

commuters_variable <- c("Age", "Female", "Less than High School", "High School Only", "Some

commuters_n <- c(length(commuters$AGE), sum(commuters$female), sum(commuters$less_than_hs),
```

```
commuters_mean <- c(mean(commuters$AGE), mean(commuters$female), mean(commuters$less_than_hs)

commuters_sd <- c(sd(commuters$AGE), sd(commuters$female), sd(commuters$less_than_hs), sd(co

commuters_min <- c(min(commuters$AGE), min(commuters$female), min(commuters$less_than_hs), m

commuters_max <- c(max(commuters$AGE), max(commuters$female), max(commuters$less_than_hs), ma

commuters_stats <- data.frame(commuters_variable, commuters_n, commuters_mean, commuters_sd,

knitr::kable(commuters_stats, col.names = c("Variable", "N", "Mean", "Std. Dev.", "Min", "Ma
```

Table 2: Table 2. Summary Statistics for 2024 ACS Sample, Commuters Subset

Variable	N	Mean	Std. Dev.	Min	Max
Age	1359943	43.39	15.30	16	96
Female	642162	0.47	0.50	0	1
Less than High School	86605	0.06	0.24	0	1
High School Only	458344	0.34	0.47	0	1
Some College	302005	0.22	0.42	0	1
College Only	307542	0.23	0.42	0	1
Advanced Degree	205447	0.15	0.36	0	1
Employed	1359943	1.00	0.00	1	1
Unemployed	0	0.00	0.00	0	0
Not in Labor Force	0	0.00	0.00	0	0
Commute Time (mins)	1359943	27.22	23.31	1	195
Total Income ($)	1359943	72992.65	88726.05	-11500	1945000

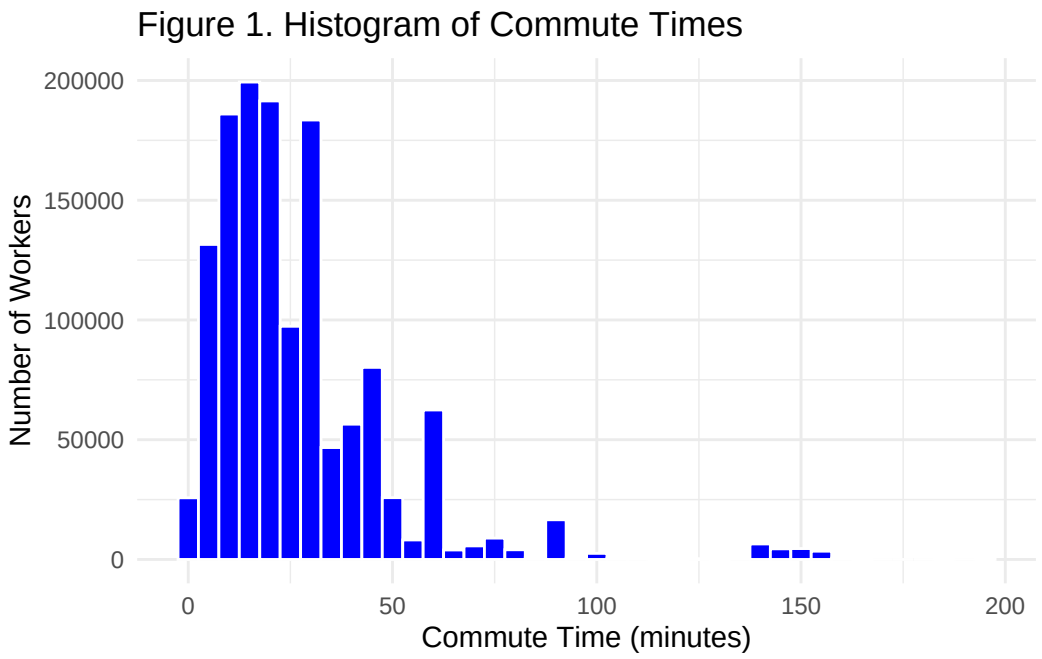
4 Visualize and Interpret

4.1

```
library(ggplot2)

ggplot(commuters, aes(x = TRANTIME)) + geom_histogram(
  binwidth = 5,
  fill = "#0000FF",
  color = "white"
) +
```

```
  labs(
    title = "Figure 1. Histogram of Commute Times",
    x = "Commute Time (minutes)",
    y = "Number of Workers"
  ) + theme_minimal()
```



4.2

The distribution is clearly right-skewed rather than symmetric, with a long right tail extending to values around 155-160 minutes and beyond (up to 195, from the findings of 2.6). The center of the distribution appears to be somewhere in the range of 20-35 minutes, aligning with the median value of 20 minutes from 3.4, Table 2. The most common commute times, represented by "spikes" in the graph, seem to be between 10-25 minutes and 25-30 minutes, with smaller spikes occurring near 40-45 minutes and 60-65 minutes. There is also a very small relative spike around 90-95 minutes. (These spikes are possibly partly attributable to individual rounding to "nice" times like 15, 30, 60, and 90 minutes.)

4.3

```
cat("The mean commute time from 3.4, Table 2 is", round(mean(commuters$TRANTIME), 2), "minute
```

```
The mean commute time from 3.4, Table 2 is 27.22 minutes.
```

```
cat("The median commute time is", median(commuters$TRANTIME), "minutes.")
```

```
The median commute time is 20 minutes.
```

Since the mean is greater than the median, we can confirm that the distribution is right-skewed, consistent with the histogram in 4.1, Figure 1.

4.4

```
ggsave("output/commute_histogram.png", width = 8, height = 5)
```

5 The Weight of Evidence

5.1

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

```
[1] 27.19112
```

5.2

Yes, the weighted mean is slightly different from the unweighted mean calculated earlier.

```
cat("The weighted mean from 5.1 is approximately", round(weighted.mean(commuters$TRANTIME, c
```

```
The weighted mean from 5.1 is approximately 27.19 minutes.
```

```
cat("The unweighted mean from 4.3 is approximately", round(mean(commuters$TRANTIME), 2), "mi
```

```
The unweighted mean from 4.3 is approximately 27.22 minutes.
```

```
cat("The unweighted mean is slightly greater than the weighted mean by approximately", round
```

```
The unweighted mean is slightly greater than the weighted mean by approximately 0.02 minutes
```

5.3

The unweighted mean indicates that within the given sample, the average commute of respondents who commuted to work was approximately:

```
cat(round(mean(commuters$TRANTIME), 2), "minutes.")
```

```
27.22 minutes.
```

The weighted mean indicates that if respondents' commute times are reweighted to match the proportion of the population they represent, then the average commute among individuals in the U.S. who commute to work is estimated by the sample to be approximately:

```
cat(round(weighted.mean(commuters$TRANTIME, commuters$PERWT),2), "minutes.")
```

```
27.19 minutes.
```

These numbers might differ if the sample is not representative of the general population; for instance, if the sample includes a large number of individuals living in rural areas who need to make long daily commutes to jobs in urban centers, and the proportion of these individuals in the overall sample is greater than their true population proportion across the U.S., then the unweighted mean would be higher than the weighted mean. However, the fact that the two means are very close in the given sample indicates that the distribution of individual respondents within the sample is representative of the distribution across the entire population.