ps1

Jenny Jia

2. Get To Know Your Data

1.

```
library(tidyverse)
```

```
-- 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   3.5.2     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.1.0
-- 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
```

```
dataset <- read_csv("data/raw/usa_00002.csv.gz") # reading my dataset
```

```
Rows: 3422888 Columns: 18
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (18): YEAR, SAMPLE, SERIAL, CBSERIAL, HHWT, CLUSTER, STATEFIP, GQ, PERNU...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

2.

STATEFIP = state FIPS code in which the household was located in AGE = an individual's age (as of their last birthday) SEX = an individual's gender (male or female) EDUC = an individual's education level (highest year completed) EMPSTAT = employment status TRANTIME = total time in minutes for commute to work INCTOT = total pre-tax personal income

TRANTIME = 0 means the individual does not/did not commmute to work, perhaps because they work from home. It does not mean that their commute was 0 minutes.

PERWT = Weight that IPUMS assigns to make their sample representative of the population of the US. Each person in the dataset represents multiple individuals in the population PERWT indicates how many people it represents

3.

```
dataset <- dataset |>
  mutate(
    female = if_else(SEX== 2, 1, 0) # creating binary 1 = female 0 = male
  )

dataset <- dataset |>
  mutate (
    less_than_hs = if_else(EDUC %in% 1:5, 1, 0),
    hs_only = if_else(EDUC == 6, 1, 0),
    some_college = if_else(EDUC %in% 7:8, 1, 0),
    college_only = if_else(EDUC == 9, 1, 0),
    advanced_degree = if_else(EDUC %in% 10:11, 1, 0)
  )

dataset <- dataset |>
mutate(
  employed = if_else(EMPSTAT == 1, 1, 0), # create employment dummy variables
  unemployed = if_else(EMPSTAT == 2, 1, 0),
  not_in_labor_force = if_else(EMPSTAT == 3, 1, 0)
)
```

```
summary(dataset$female) # checking to see if code worked
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  1.0000  0.5093  1.0000  1.0000
```

```r
table(dataset$employed)
```

```
      0       1
1814867 1608021
```

4.

Treating NA codes as actual zeros would distort the results because those values do not represent actual observations. For example, if TRANTIME = 0 means a person worked from home/had no real commute, counting it as a zero-minute commute would lower the average travel time to work. This would make it seem like people commute much less than they actually do and lead to misleading conclusions about transportation statistics.

5.

```r
dataset <- dataset |>
  mutate(
    INCTOT = if_else(INCTOT >= 9999999, NA, INCTOT) # had a problem with unrealistic total i
  )
```

```r
library(knitr)

dataset_table <- tibble(
  Variable = c(
    "Age (years)",
    "Female (1=female)",
    "Less than High School (1=yes)",
    "High School Only (1=yes)",
    "Some College (1=yes)",
    "College Only (1=yes)",
    "Advanced Degree (1=yes)",
    "Employed (1=yes)",
    "Unemployed (1=yes)",
    "Not in Labor Force (1=yes)",
    "Commute Time (mins)",
    "Total Income ($)"
  ),
  x = list(
```

```
    dataset$AGE,
    dataset$female,
    dataset$less_than_hs,
    dataset$hs_only,
    dataset$some_college,
    dataset$college_only,
    dataset$advanced_degree,
    dataset$employed,
    dataset$unemployed,
    dataset$not_in_labor_force,
    dataset$TRANTIME,
    dataset$INCTOT
  )
) |>
  mutate(
    N = map_int(x, ~ sum(!is.na(.x))),
    Mean = map_dbl(x, ~ mean(.x, na.rm = TRUE)),
    `Std. Dev.` = map_dbl(x, ~ sd(.x, na.rm = TRUE)),
    Min = map_dbl(x, ~ min(.x, na.rm = TRUE)),
    Max = map_dbl(x, ~ max(.x, na.rm = TRUE))
  ) |>
  select(-x)

kable(dataset_table, digits = 2,
      caption = "Summary Statistics for 2024 ACS Sample")
```

Table 1: Summary Statistics for 2024 ACS Sample

Variable	N	Mean	Std. Dev.	Min	Max
Age (years)	3422888	43.39	24.03	0	96
Female (1=female)	3422888	0.51	0.50	0	1
Less than High School (1=yes)	3422888	0.19	0.39	0	1
High School Only (1=yes)	3422888	0.30	0.46	0	1
Some College (1=yes)	3422888	0.18	0.38	0	1
College Only (1=yes)	3422888	0.00	0.00	0	0
Advanced Degree (1=yes)	3422888	0.28	0.45	0	1
Employed (1=yes)	3422888	0.47	0.50	0	1
Unemployed (1=yes)	3422888	0.02	0.14	0	1
Not in Labor Force (1=yes)	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

6.

The maximum commute time is 195 minutes, or over 3 hours, which personally doesn't make much sense to me. However, I understand that this is still a possibility for those who travel extremely far for work, or for those who take different methods of public transportation.

3 Who Should Be in My Analysis?

1.

```
zero_commute <- sum(dataset$TRANTIME == 0, na.rm = TRUE) # where TRANTIME = 0

total <- nrow(dataset) # total observations

zero_commute_percent <- zero_commute / total * 100 # percentage

zero_commute
```

```
[1] 2062945
```

```
zero_commute_percent
```

```
[1] 60.26914
```

According to the results, 2,062,945 people do not commute to work, which is about 60% of the total population. While this number seems rather high, this makes sense, if you consider the fact that students, children, unemployed, and retired people all don't have a commute to work.

2.

```
commuters <- dataset |>
  filter( # now I will filter for only people that actually have a commute to work
    EMPSTAT == 1, # employed people
    TRANTIME > 0, # has a real commute
    AGE >= 16
  )
```

I restricted the sample to individuals who are employed, at least 16 years old, and have a commute time greater than zero. This removes children, retirees, unemployed people, and people who work from home. Also, although the legal minimum working age is 14, 16 is the age used by the Bureau of Labor Statistics, so I shall stick with 16 instead of 14.

3.

```
nrow(commuters) # new number of observations
```

```
[1] 1359943
```

```
min(commuters$TRANTIME) # new minimum commute time
```

```
[1] 1
```

After restricting the dataset to employed individuals aged 16 or older with a commute time greater than zero, 1,359,943 observations remain. The minimum commute time is now 1 minute, which makes sense because individuals with no commute were excluded from the analysis.

4.

```
library(knitr)

commuters_table <- tibble(
  Variable = c(
    "Age (years)",
    "Female (1=female)",
    "Less than High School (1=yes)",
    "High School Only (1=yes)",
    "Some College (1=yes)",
    "College Only (1=yes)",
    "Advanced Degree (1=yes)",
    "Employed (1=yes)",
    "Unemployed (1=yes)",
    "Not in Labor Force (1=yes)",
    "Commute Time (mins)",
```

```
    "Total Income ($)"
  ),
  x = list(
    commuters$AGE,
    commuters$female,
    commuters$less_than_hs,
    commuters$hs_only,
    commuters$some_college,
    commuters$college_only,
    commuters$advanced_degree,
    commuters$employed,
    commuters$unemployed,
    commuters$not_in_labor_force,
    commuters$TRANTIME,
    commuters$INCTOT
  )
) |>
  mutate(
    N = map_int(x, ~ sum(!is.na(.x))),
    Mean = map_dbl(x, ~ mean(.x, na.rm = TRUE)),
    `Std. Dev.` = map_dbl(x, ~ sd(.x, na.rm = TRUE)),
    Min = map_dbl(x, ~ min(.x, na.rm = TRUE)),
    Max = map_dbl(x, ~ max(.x, na.rm = TRUE))
  ) |>
  select(-x)

kable(
  commuters_table,
  digits = 2,
  caption = "Summary Statistics for Commuters Subset"
)
```

Table 2: Summary Statistics for Commuters Subset

Variable	N	Mean	Std. Dev.	Min	Max
Age (years)	1359943	43.39	15.30	16	96
Female (1=female)	1359943	0.47	0.50	0	1
Less than High School (1=yes)	1359943	0.05	0.22	0	1
High School Only (1=yes)	1359943	0.34	0.47	0	1
Some College (1=yes)	1359943	0.22	0.42	0	1
College Only (1=yes)	1359943	0.00	0.00	0	0

Variable	N	Mean	Std. Dev.	Min	Max
Advanced Degree (1=yes)	1359943	0.38	0.48	0	1
Employed (1=yes)	1359943	1.00	0.00	1	1
Unemployed (1=yes)	1359943	0.00	0.00	0	0
Not in Labor Force (1=yes)	1359943	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

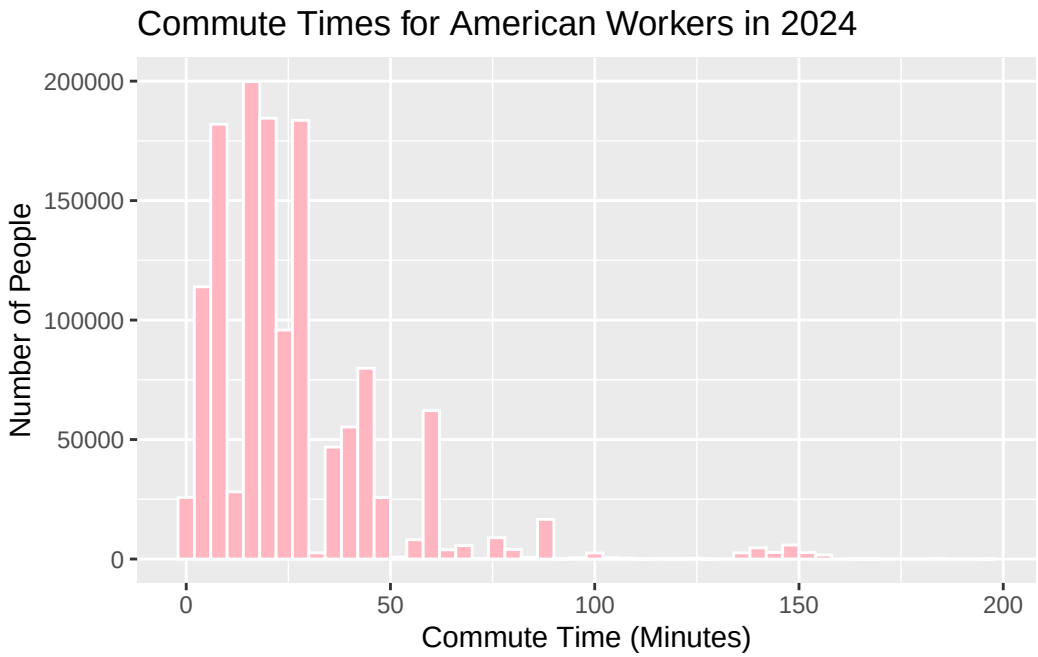
5.

This step was completed using the commit function.

4. Visualize and Interpret

1.

```
ggplot(commuters, aes(x = TRANTIME)) +
  geom_histogram(binwidth = 4, fill = "lightpink", color = "white") +
  labs(
    title = "Commute Times for American Workers in 2024",
    x = "Commute Time (Minutes)",
    y = "Number of People"
  )
```



2.

The distribution of commute times is skewed right, which means most people have relatively short commutes to work, although a select few people have longer ones that exceed 100 minutes. The center, or median, appears to be around 20-30 minutes. There are spikes every 10 minutes, for example, at 20 minutes. This could indicate that people are rounding their commute, as it is hard to report a very precise commute time.

3.

```
mean <- mean(commuters$TRANTIME, na.rm = TRUE) # doing some calculations for precise values
median <- median(commuters$TRANTIME, na.rm = TRUE)
mean
```

```
[1] 27.21505
```

```
median
```

```
[1] 20
```

In the commuters subset, the mean commute time is 27.21 minutes and the median is 20 minutes, so the mean is larger. This indicates a right-skewed distribution where a small number of very long commutes pull the average upward. This matches the histogram, which shows most commute times clustered below about 40 minutes with a long right tail. # 4.

```
ggsave(
  filename = "output/commute_histogram.png",
  width = 7,
  height = 5
)
```

5.

Committing results should show on GitHub.

5. The Weight of Evidence

1.

```
weighted_mean <- weighted.mean(
  commuters$TRANTIME,
  commuters$PERWT,
  na.rm = TRUE
)

weighted_mean
```

```
[1] 27.19112
```

2.

The weighted and unweighted means are slightly different. The weighted mean is 27.19 minutes, whereas the unweighted is 27.21 minutes. The difference is 0.02 minutes.

3.

The unweighted mean represents the average commute time among the individuals in the survey sample, while the weighted mean represents the estimated average commute time for the entire US population using survey weights. They might differ because some demographic groups can be overrepresented in the sample, and weights correct for that imbalance. In this case, the numbers are essentially identical, which suggests that the sample is already fairly representative of the population. "‘