**Exercise 1:**

Find missing value in variable **Price**

> summarize(new\_data\_metro\_zip, n = n(), na.rm = FALSE)

n na.rm

1 4584899 FALSE

> sum(is.na(new\_data\_metro\_zip$Price))

[1] 861121

The result shows that there are 861121 missing values in the **Price** variable

**Exercise 2:**

> mean(new\_data\_metro\_zip$Price, na.rm = TRUE)

[1] 322914.9

The argument na.rm = TRUE, which tells R to ignore any NA (missing) values in the **Price** variable before calculating the mean.

The result 322914.9 is the mean of the **Price** variable

**Exercise 3:**

> library(dplyr)

> region\_mean\_price <- new\_data\_metro\_zip %>% group\_by(RegionType) %>% summarize(mean\_price = mean(Price, na.rm = TRUE))

> print(region\_mean\_price)

# A tibble: 3 × 2

RegionType mean\_price

*<chr>* *<dbl>*

1 country 281886.

2 msa 237686.

3 zip 327720.

The average house price across all entries classified as ‘country’ is approximately 281886, as ‘msa’ is approximately 237686 and as ‘zip’ is approximately 327720

These results suggest that there are differences in the average house prices when grouped by **RegionType**. The average house price seems to be highest for entries classified as ‘zip’, and lowest for those classified as ‘msa’

**Exercise 4:**

> #median

> median\_price <- median(new\_data\_metro\_zip$Price, na.rm = TRUE)

> print(median\_price)

[1] 244411.6

> #standard deviation

> sd\_price <- sd(new\_data\_metro\_zip$Price, na.rm = TRUE)

> print(sd\_price)

[1] 302391.9

The median of **Price** is 244411.6

The standard deviation of **Price** is 302391.9

**Exercise 5:**

> region\_percent <- new\_data\_metro\_zip %>% group\_by(RegionType) %>% summarize(n = n()) %>% mutate(freq = n / sum(n))

> print(region\_percent)

# A tibble: 3 × 3

RegionType n freq

*<chr>* *<int>* *<dbl>*

1 country 285 0.0000622

2 msa 243390 0.0531

3 zip 4341224 0.947

The frequency (or percentage) of houses in each **RegionType**:

* with 285 houses, the percentage of houses are at the country level is 0.0000622 (or 0.00622%)
* with 243390 houses, the percentage of houses are at the MSA level is 0.0531 (or 5.31%)
* with 4341224 houses, the percentage of houses are at the zip code level is 0.947 (or 94.7%)

**Exercise 6:**

> region\_state\_summary <- new\_data\_metro\_zip %>% group\_by(RegionType, StateName) %>% summarize(n = n()) %>% mutate(freq = n / sum(n))

`summarise()` has grouped output by 'RegionType'. You can override using the `.groups` argument.

> print(region\_state\_summary)

# A tibble: 102 × 4

# Groups: RegionType [3]

RegionType StateName n freq

*<chr>* *<chr>* *<int>* *<dbl>*

1 country "" 285 1

2 msa "AK" 1140 0.00468

3 msa "AL" 5415 0.0222

4 msa "AR" 4845 0.0199

5 msa "AZ" 2850 0.0117

6 msa "CA" 9405 0.0386

7 msa "CO" 4845 0.0199

8 msa "CT" 1425 0.00585

9 msa "DE" 570 0.00234

10 msa "FL" 8265 0.0340

# ℹ 92 more rows

# ℹ Use `print(n = ...)` to see more rows

The frequency (or proportion) of houses in each combination of **RegionType** and **StateName:** all houses categorized at the country level make up 100% (1) of all country-level houses**,** while houses at the MSA level in AK make up approximately 0.00468 (or 0.468%) of all MSA-level houses**,** houses at the MSA level in AL make up approximately 0.0222 (or 2.22%) of all MSA-level houses,……

**Exercise 7:**

> state\_mean\_price <- new\_data\_metro\_zip %>% group\_by(StateName) %>% summarize(mean\_price = mean(Price, na.rm = TRUE))

> print(state\_mean\_price)

# A tibble: 52 × 2

StateName mean\_price

*<chr>* *<dbl>*

1 "" 281886.

2 "AK" 342410.

3 "AL" 210132.

4 "AR" 204382.

5 "AZ" 305145.

6 "CA" 652366.

7 "CO" 419162.

8 "CT" 384266.

9 "DC" 686295.

10 "DE" 319211.

# ℹ 42 more rows

# ℹ Use `print(n = ...)` to see more rows

The first row with **StateName** as ("") has a mean price of 281,886. The second row "AK", with a mean house price of 342,410. The third row "AL", with a mean house price of 210,132 …….  
**Exercise 8:**

First we create my\_table :

> my\_table <- new\_data\_metro\_zip %>% group\_by(RegionType) %>% summarize(mean = mean(Price, na.rm = TRUE), median = median(Price, na.rm = TRUE), sd = sd(Price, na.rm = TRUE))

> print(my\_table)

# A tibble: 3 × 4

RegionType mean median sd

*<chr>* *<dbl>* *<dbl>* *<dbl>*

1 country 281886. 267560. 75123.

2 msa 237686. 199186. 145763.

3 zip 327720. 248201. 308172.

Then we save it to a .png file:

> #save as .png file

> library(gridExtra)

> library(grid)

> grid\_table <- tableGrob(my\_table)

> png(filename = "my\_table.png")

> grid.draw(grid\_table)

> dev.off()

null device

1