Exercise 1:

- First, Reads the CSV file into an R dataframe.

```
> library(dplyr)
> setwd("C:/Users/PC/OneDrive - VietNam National University - HCM INTERNAT
IONAL UNIVERSITY/Desktop/DA/Lab/LAB5")
> data_houston <- read.csv("Zillow-Houston-TX.csv")
```

- Removes the prefix “results.” from all the variable names

```
names(data_houston) <- gsub("results.", "", names(data_houston))
```

- Then, writes the dataframe back to a CSV file, overwriting the original file.

```
> write.csv(data_houston, "Zillow-Houston-TX.csv", row.names = FALSE)
```

- Replaces all the missing values (NA) with 0. It also removes the rows where homeType is missing

```
> data_houston[is.na(data_houston)] <- 0
> data_houston <- data_houston[data_houston$homeType != "", ]
```

Exercise 2:

```
> data_houston %>%group_by(homeType) %>%summarise(
+   mean_price = mean(price[price != 0], na.rm = TRUE),
+   median_price = median(price[price != 0], na.rm = TRUE),
+   mode_price = names(which.max(table(price[price != 0])))
+     )
# A tibble: 4 × 4
  homeType      mean_price median_price mode_price
  <chr>              <dbl>        <dbl> <chr>
1 CONDO            474483.       211950 118000
2 LOT              256500        217500 75000
3 SINGLE_FAMILY    547277.       322500 80000
4 TOWNHOUSE        345333.       365000 151000
```

- The result provides the mean, median, and mode of the price for each type of home (homeType), excluding the prices that are 0 (which represent missing values). The mean is the average price, the median is the middle price when the prices are sorted, and the mode is the most frequent price.
  - ❖ CONDO: The mean price is approximately 474,483, the median price is 211,950, and the mode price is 118,000.
  - ❖ LOT: The mean price is 256,500, the median price is 217,500, and the mode price is 75,000.
  - ❖ SINGLE_FAMILY: The mean price is approximately 547,277, the median price is 322,500, and the mode price is 80,000.
  - ❖ TOWNHOUSE: The mean price is approximately 345,333, the median price is 365,000, and the mode price is 151,000.

Exercise 3:

```
> data_houston %>%group_by(homeType) %>%summarise(
+   variance_price = var(price[price != 0], na.rm = TRUE),
+   sd_price = sd(price[price != 0], na.rm = TRUE),
+   IQR_price = IQR(price[price != 0], na.rm = TRUE)
+   )
# A tibble: 4 × 4
  homeType      variance_price sd_price IQR_price
  <chr>                  <dbl>    <dbl>     <dbl>
1 CONDO          376952681667.  613965.    301475
2 LOT             47389500000   217691.    128250
3 SINGLE_FAMILY  211446963934.  459834.    370758.
4 TOWNHOUSE       34330333333.  185284.    184500
```

- The result provides the variance, standard deviation, and interquartile range (IQR) of the price for each type of home (homeType), excluding the prices that are 0 (which represent missing values). The variance and standard deviation is a measure of how spread out the prices are from their mean, and the IQR is the range within which the central 50% of the prices fall.
    - ❖ CONDO: The variance in price is approximately 376,952,681,667, the standard deviation is 613,965, and the IQR is 301,475.
    - ❖ LOT: The variance in price is 47,389,500,000, the standard deviation is 217,691, and the IQR is 128,250.
    - ❖ SINGLE_FAMILY: The variance in price is approximately 211,446,963,934, the standard deviation is 459,834, and the IQR is 370,758.
    - ❖ TOWNHOUSE: The variance in price is 34,330,333,333, the standard deviation is 185,284, and the IQR is 184,500.

Exercise 4:

```
> data_houston %>%group_by(homeType) %>%summarise(
+   ratio_90_10 = quantile(price[price != 0], 0.9, na.rm = TRUE) / quantil
e(price[price != 0], 0.1, na.rm = TRUE)
+   )
# A tibble: 4 × 2
  homeType      ratio_90_10
  <chr>               <dbl>
1 CONDO                8.97
2 LOT                  5.34
3 SINGLE_FAMILY        5.72
4 TOWNHOUSE            2.52
```

- The result provides the 90/10 ratio of the price for each type of home (homeType), excluding the prices that are 0 (which represent missing values). These ratios is a measure of the spread of the price distribution for each type of home. A higher ratio indicates a wider spread of prices.
    - ❖ CONDO: The 90/10 ratio is approximately 8.97. This means that the price at the 90th percentile is about 9 times the price at the 10th percentile.
    - ❖ LOT: The 90/10 ratio is approximately 5.34. This means that the price at the 90th percentile is about 5 times the price at the 10th percentile.
    - ❖ SINGLE_FAMILY: The 90/10 ratio is approximately 5.72. This means that the price at the 90th percentile is about 6 times the price at the 10th percentile.
    - ❖ TOWNHOUSE: The 90/10 ratio is approximately 2.52. This means that the price at the 90th percentile is about 2.5 times the price at the 10th percentile.

Exercise 5:

```
> # Convert acres to square feet
> data_houston$lotAreaValue[data_houston$lotAreaUnit == "acres"] <- data_h
ouston$lotAreaValue[data_houston$lotAreaUnit == "acres"] * 43560
> # Now all lotAreaValue is in square feet
> data_houston$lotAreaUnit <- "sqft"
> # Calculate covariance
> cov_price_lotAreaValue <- cov(data_houston$price[data_houston$price != 0
], data_houston$lotAreaValue[data_houston$price != 0], use = "na.or.comple
te")
> print(paste("The covariance between price and lotAreaValue is", cov_pric
e_lotAreaValue))
[1] "The covariance between price and lotAreaValue is -11572755727.6942"
```

- The covariance between price and lotAreaValue is approximately -11572755727.6942. This negative value indicates that there is a tendency for price and lotAreaValue to move in opposite directions. In other words, as the price increases, the lotAreaValue tends to decrease, and vice versa.

Exercise 6:

```
> # Calculate correlation
> cor_price_lotAreaValue <- cor(data_houston$price[data_houston$price != 0
], data_houston$lotAreaValue[data_houston$price != 0], use = "na.or.comple
te")
> print(paste("The correlation between price and lotAreaValue is", cor_pri
ce_lotAreaValue))
[1] "The correlation between price and lotAreaValue is -0.206124783643226"
```

- The correlation value is approximately -0.206, which indicates a weak negative relationship between these two variables. This means that as the lotAreaValue increases, the price tends to decrease slightly, and vice versa. However, the correlation is weak, so this trend is not very strong.