Assignment 2

Yiwei Huang

2023

Section 1: Descrption of the data

The dataset is measuing housing price in in the suburbs of chicago and features of the houses. The data was uploaded to Kaggle and made public.

The research questions this data could help to answer is what features contribute to the housing price in Chicago most?

It’s saved in csv format. It’s delimited by comma.

Section 2: Reading the data into R

```
#use read.csv to read in the data, it's a base R function
df <- read.csv('~/Downloads/realest.csv')
```

Section 3: Clean the data

```
#rename col
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
df1 = rename(df,Squarefeet=Space)
#get rid of lot col
df1 = select(df1,-Lot)
```

Section 4: Characteristics of the data

```
#This next chunk is inline code. Inline code puts the text with the output of the function in my docume
library(knitr)
#Create a dataframe with column names and brief descriptions
data <- data.frame(
  Column_Name = c("Price", "Bedroom", "Room", "Space",
                  "Lot","Tax","Bathroom","Garage","Condition"),
  Description = c("price of house",
                  "number of bedrooms",
                  "number of rooms",
                  "size of house (in square feet)",
                  "width of a lot",
                  "amount of annual tax",
                  "number of bathrooms",
                  "number of garage",
                  "condition of house (1 if good , 0 otherwise)")
)
```

This dataframe has 157 rows and 9 columns. The names of the columns and a brief description of each are in the table below:

Column Name	Description
Price	price of house
Bedroom	number of bedrooms
Room	number of rooms
Space	size of house (in square feet)
Lot	width of a lot
Tax	amount of annual tax
Bathroom	number of bathrooms
Garage	number of garage
Condition	condition of house (1 if good , 0 otherwise)

‘

Section 5: Summary statistics

```
# Pick three columns of the dataframe
df_3cols = select(df,Bedroom, Bathroom, Garage)
# Use a summary function to get the following summaries of these columns
get_summary <- function(x) {
  result <- c(
    min_value = min(x, na.rm = TRUE),
    max_value = max(x, na.rm = TRUE),
    mean_value = mean(x, na.rm = TRUE),
    num_missing = sum(is.na(x))
  )
  return(result)
```

```r
}
bedroom_summary <- get_summary(df_3cols$Bedroom)
bathroom_summary <- get_summary(df_3cols$Bathroom)
garage_summary <- get_summary(df_3cols$Garage)

final_summary <- rbind(bedroom_summary, bathroom_summary, garage_summary)
print(final_summary)
```

```
##                  min_value max_value mean_value num_missing
## bedroom_summary          1         8  3.1666667           1
## bathroom_summary         1         3  1.4807692           1
## garage_summary           0         2  0.8461538           1
```