Assignment 1

Meenakshi Vaidhiyanathan

2024-02-05

Dataset:

The dataset used for this project can be found here:
https://www.kaggle.com/datasets/yasserh/housing-prices-dataset.

Setup:

Libraries used :

```
##
## 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
```

Loading the dataset :

```
housingData <- read.csv("./housing.csv")
head(housingData)
```

```
##      price area bedrooms bathrooms stories mainroad guestroom basement
## 1 13300000 7420        4         2       3      yes        no       no
## 2 12250000 8960        4         4       4      yes        no       no
## 3 12250000 9960        3         2       2      yes        no      yes
## 4 12215000 7500        4         2       2      yes        no      yes
## 5 11410000 7420        4         1       2      yes       yes      yes
## 6 10850000 7500        3         3       1      yes        no      yes
##   hotwaterheating airconditioning parking prefarea furnishingstatus
## 1              no             yes       2      yes        furnished
## 2              no             yes       3       no        furnished
## 3              no              no       2      yes   semi-furnished
## 4              no             yes       3      yes        furnished
## 5              no             yes       2       no        furnished
## 6              no             yes       2      yes   semi-furnished
```

Descriptive statistics for quantitative variables:

The `beds` variable includes results on the number of houses with their respective bedrooms. For eg: 136 houses have 2 bedrooms.

```
beds<-table(housingData$bedrooms)
head(beds)
```

```
##
##   1   2   3   4   5   6
##   2 136 300  95  10   2
```

The `bath` variable includes results on the number of houses with their respective bathrooms. For eg: 401 houses have 1 bathroom.

```
bath<-table(housingData$bathrooms)
head(bath)
```

```
##
##   1   2   3   4
## 401 133  10   1
```

The `stories` variable includes results on the number of houses that are single, double, triple or quadruple storied building. For eg: 227 housese have are single storied.

```
stories<-table(housingData$stories)
head(stories)
```

```
##
##   1   2   3   4
## 227 238  39  41
```

The code below calculates the average price and the median value of the price of the houses.

```
mean(housingData$price)
```

```
## [1] 4766729
```

```
median(housingData$price)
```

```
## [1] 4340000
```

The `order()` function is used on the dataset housingData to express values of area in descending order. The `head()` function is then used to print the valueof the house with highest area while the `tail()` function is used to print the value of the house with lowest area.

```
highestAreaTable <-housingData[order(housingData$area,decreasing = TRUE),]
highAreaInfo<-head(highestAreaTable,1)
lowAreaInfo<-tail(highestAreaTable,1)
highAreaInfo
```

```
##      price  area bedrooms bathrooms stories mainroad guestroom basement
## 8 10150000 16200        5         3       2      yes        no       no
##   hotwaterheating airconditioning parking prefarea furnishingstatus
## 8              no              no       0       no      unfurnished
```

```
lowAreaInfo
```

```
##       price area bedrooms bathrooms stories mainroad guestroom basement
## 450 3150000 1650        3         1       2       no        no      yes
##     hotwaterheating airconditioning parking prefarea furnishingstatus
## 450              no              no       0       no      unfurnished
```

The `park` variable is used to store the number of houses with their respective parking facilities. For eg: 299 houses have zero parking facility.

```
park<-table(housingData$parking)
park
```

```
##
##   0   1   2   3
## 299 126 108  12
```

Descriptive statistics for categorical variables

The `num_ac` variable stores the number of houses with or without air conditioning.

```
num_ac<-table(housingData$airconditioning)
num_ac
```

```
##
##  no yes
## 373 172
```

The `num_furnishing` variable stores the number of houses that are furnished, semi-furnished and unfurnished.

```
num_furnishing<-table(housingData$furnishingstatus)
num_furnishing
```

```
##
##      furnished semi-furnished    unfurnished
##            140            227            178
```

The `num_hotwaterheating` variable stores the number of houses that are with or without hot water heating facility.

```
num_hotwaterheating<-table(housingData$hotwaterheating)
num_hotwaterheating
```

```
##
##  no yes
## 520  25
```

The num_mainroad variable stores the number of houses that are on the mainroad.

```
num_mainroad<-table(housingData$mainroad)
num_mainroad
```

```
##
##  no yes
##  77 468
```

The following code lists houses which have both air conditioning and hot water heating facilities. The filter() function filters the row of such houses that abide by the two conditions. The function was used as it supports conditions across multiple columns.

```
print(filter(housingData,(housingData$airconditioning== 'yes' & housingData$hotwaterheating=='yes')))
```

```
##     price area bedrooms bathrooms stories mainroad guestroom basement
## 1 3640000 2275        3         1       3      yes        no       no
##   hotwaterheating airconditioning parking prefarea furnishingstatus
## 1             yes             yes       0      yes   semi-furnished
```

Transform Variable

The dataset contained the absolute pricing of houses. It was transformed to be expressed as a multiple of 100,000.

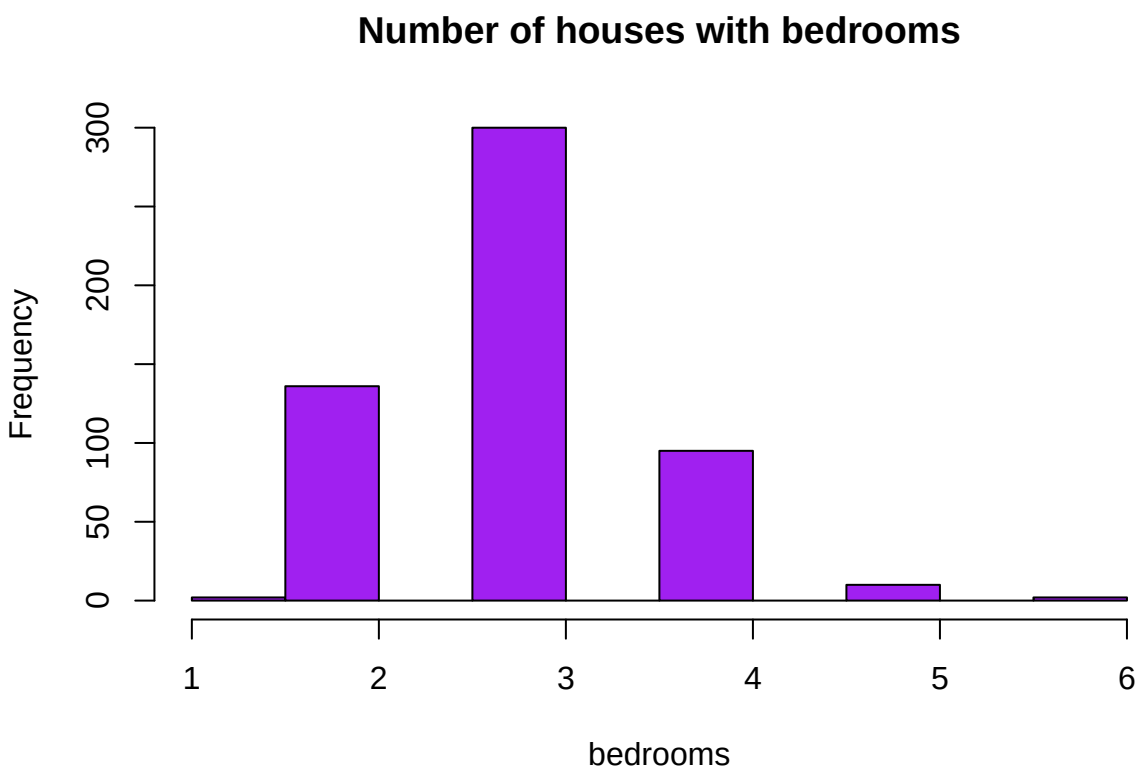
```
priceperhunk<-housingData$price/100000
head(priceperhunk, 30)
```

```
##  [1] 133.00 122.50 122.50 122.15 114.10 108.50 101.50 101.50  98.70  98.00
## [11]  98.00  96.81  93.10  92.40  92.40  91.00  91.00  89.60  88.90  88.55
## [21]  87.50  86.80  86.45  86.45  85.75  85.40  84.63  84.00  84.00  84.00
```

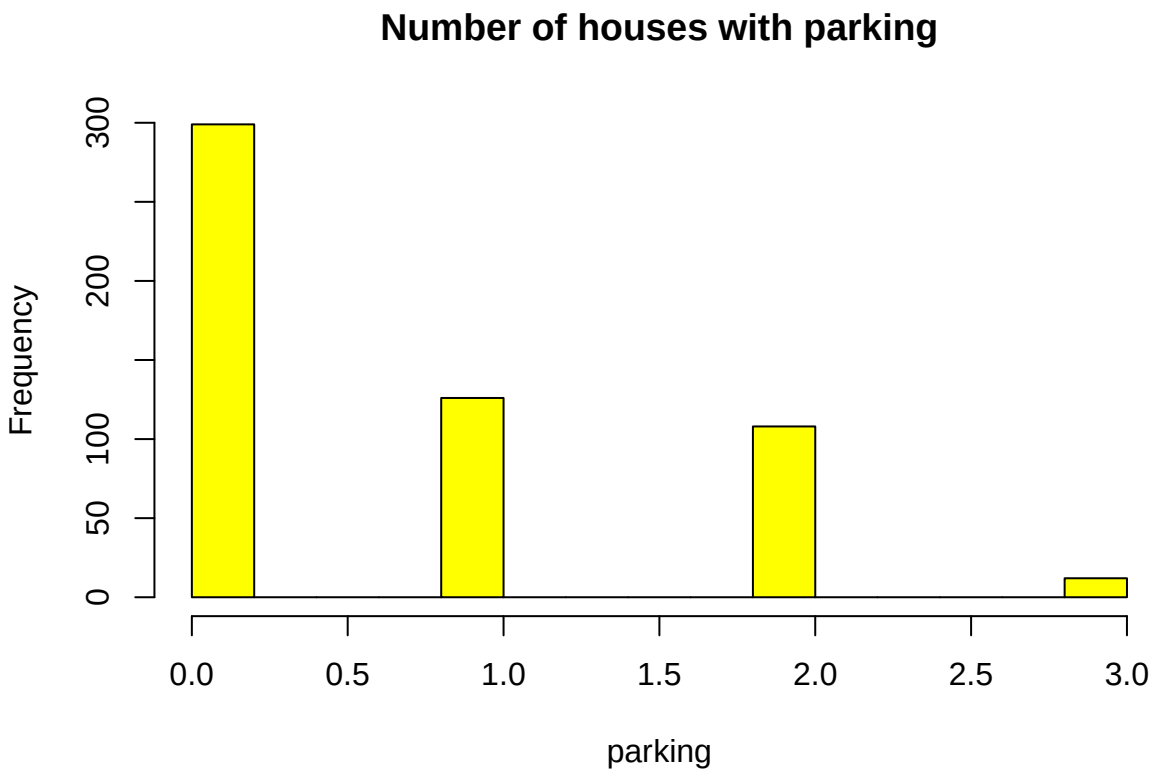
Plots

Plotting quantitative variable:

The following histogram `hist()` plots the number of houses in the dataset vs the number of bedrooms.

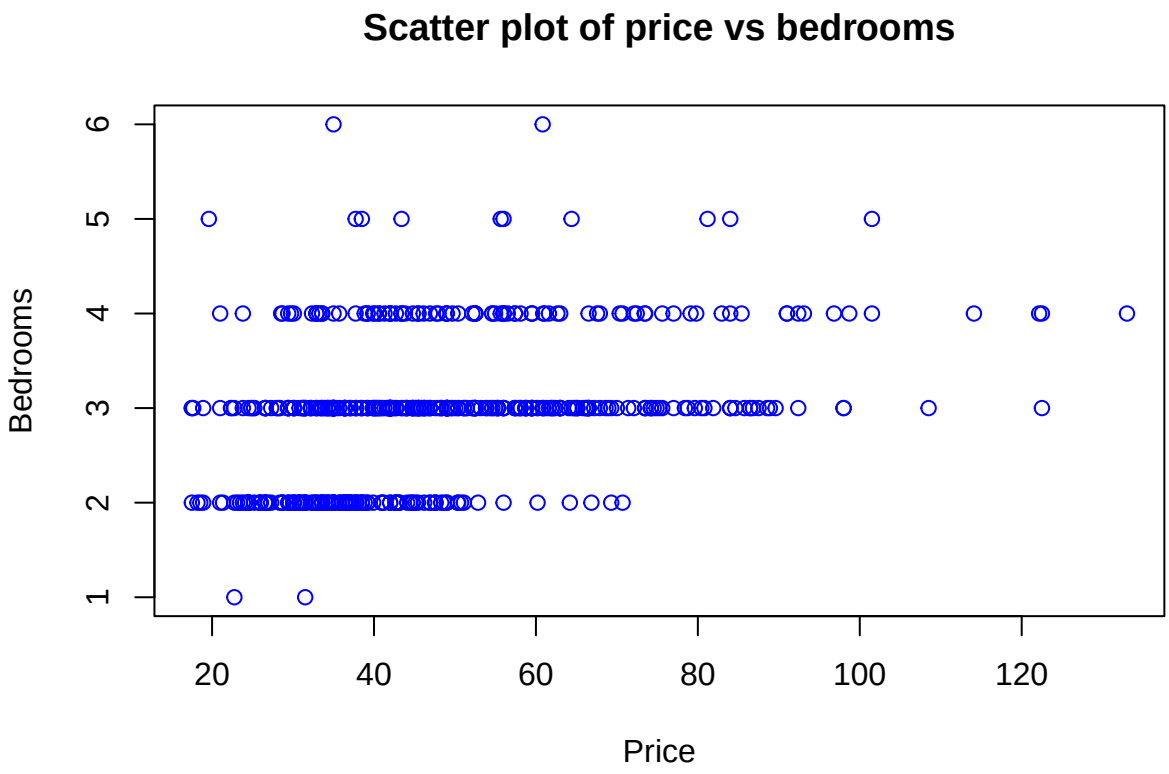


The following histogram `hist()` plots the number of houses in the dataset vs the number of parking spots.

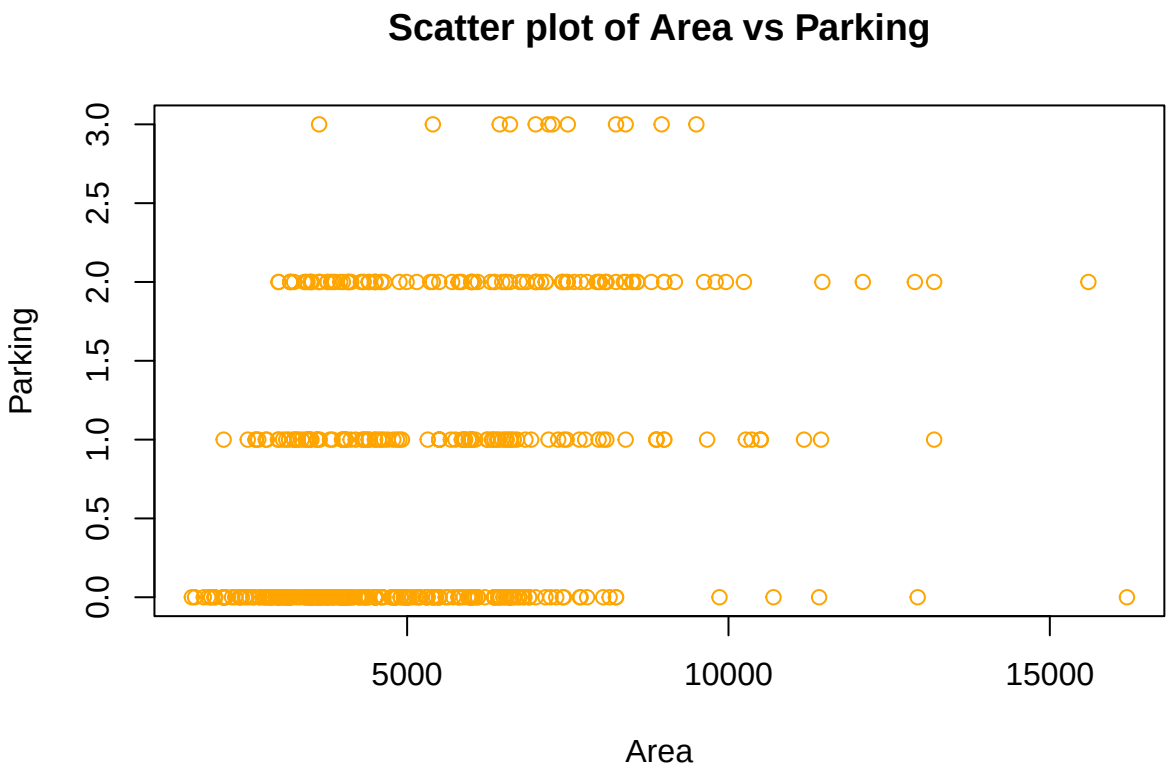


Plotting a scatter plot

The following plot depicts the price of the house in the dataset vs the number of bedrooms. The plot illustrates the variability in price based on the number of bedrooms.

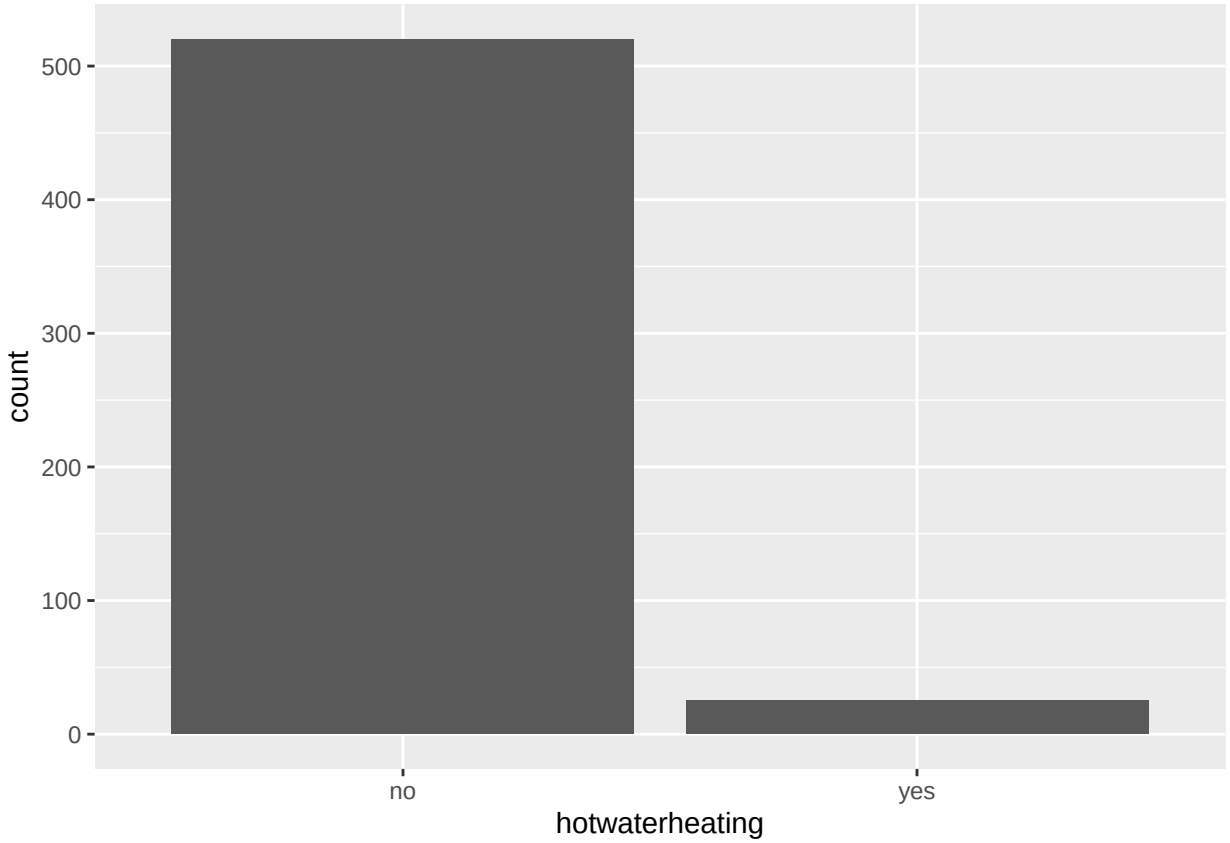


The following plot illustrates the area of the house vs the number of parking spots available.



Plotting of a categorical variable

- The following plot illustrates a categorical value i.e. `hotwaterheating` of values yes or no.
- The function `ggplot()` was used as it allowed for plotting of categorical data. The `geom_bar()` function was used to depict a bar graph. The `aes()` or aesthetic function was needed to specify the categorical variable against the axis of the graph. Documentation for the `aes()` function can be found at https://www.rdocumentation.org/packages/ggplot2/versions/3.4.4/topics/aes.



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.