Lab 8

2023-07-05

1. Read in data file and set up data frame for analysis.

```
realestate <- read.csv('RealEstate.csv')
head(realestate)
```

```
##   No X1.transaction.date X2.house.age X3.distance.to.the.nearest.MRT.station
## 1  1            2012.917         32.0                                84.87882
## 2  2            2012.917         19.5                               306.59470
## 3  3            2013.583         13.3                               561.98450
## 4  4            2013.500         13.3                               561.98450
## 5  5            2012.833          5.0                               390.56840
## 6  6            2012.667          7.1                              2175.03000
##   X4.number.of.convenience.stores X5.latitude X6.longitude
## 1                              10    24.98298     121.5402
## 2                               9    24.98034     121.5395
## 3                               5    24.98746     121.5439
## 4                               5    24.98746     121.5439
## 5                               5    24.97937     121.5425
## 6                               3    24.96305     121.5125
##   Y.house.price.of.unit.area
## 1                       37.9
## 2                       42.2
## 3                       47.3
## 4                       54.8
## 5                       43.1
## 6                       32.1
```

```
#1a.
house <- data.frame(
  realestate$X2.house.age,       realestate$X3.distance.to.the.nearest.MRT.station,
realestate$X4.number.of.convenience.stores,
realestate$Y.house.price.of.unit.area)

#1b.
colnames(house) <- c('age','mrt.dist','num.conv.stores','price')

# You can also use dplyr to do this.
```

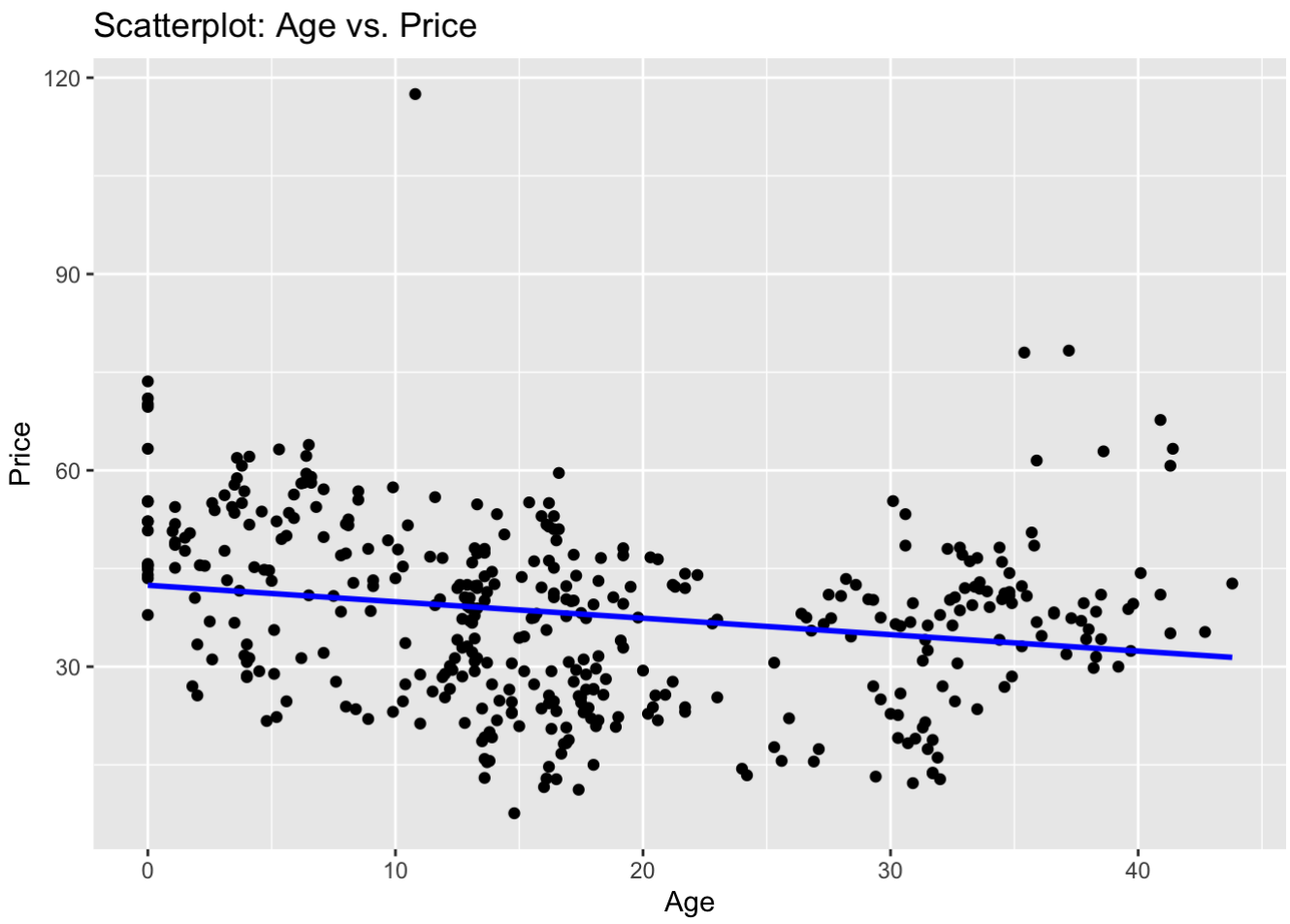
```
library(ggplot2)
#1c price vs age

plot_age <- ggplot(house, aes(x = age, y = price)) +
  geom_point() +
  geom_smooth(method = "lm", color = "blue", se = FALSE) +
  labs(x = "Age", y = "Price", title = "Scatterplot: Age vs. Price")



plot_age
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

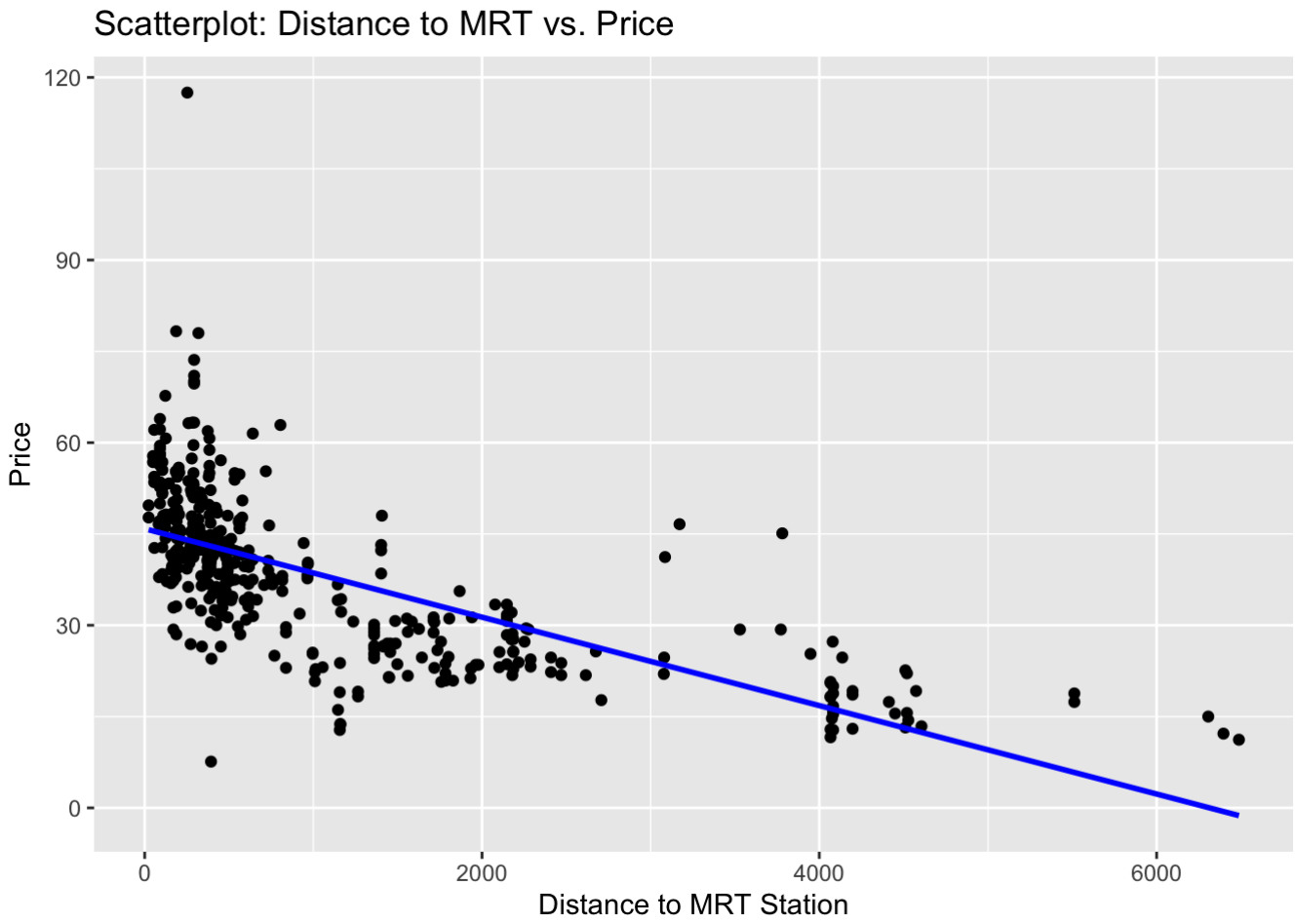



```
#1c price vs distance to mrt station

plot_mrt_dist <- ggplot(house, aes(x = mrt.dist, y = price)) +
  geom_point() +
  geom_smooth(method = "lm", color = "blue", se = FALSE) +
  labs(x = "Distance to MRT Station", y = "Price", title = "Scatterplot: Distance to
MRT vs. Price")

plot_mrt_dist
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

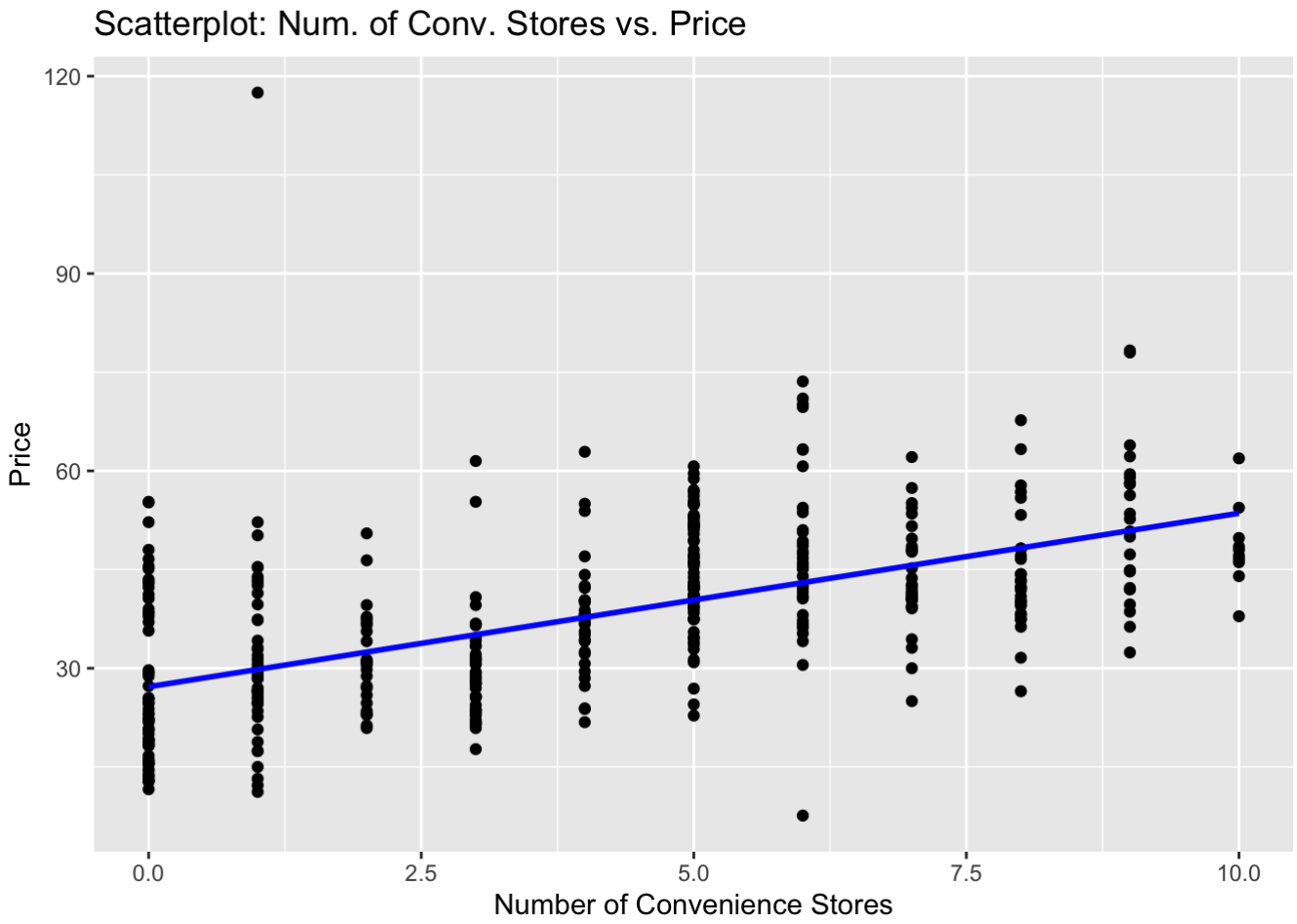



```
#1c price vs no. of convenience stores nearby

plot_num_conv_stores <- ggplot(house, aes(x = num.conv.stores, y = price)) +
  geom_point() +
  geom_smooth(method = "lm", color = "blue", se = FALSE) +
  labs(x = "Number of Convenience Stores", y = "Price", title = "Scatterplot: Num. of
Conv. Stores vs. Price")

plot_num_conv_stores
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
#1d Brief discussion about the plots
```

age vs price: the price seems to reach a 'baseline' between 'young' houses and 'older' houses. But younger houses are slightly more expensive than the baseline while the oldest houses seem to have a dramatic price spike. Meaning there is a link: the newer a house is the pricier it gets and this price diminishes until a certain extent. Once you take a look at really old houses the price spikes again.

MRT Station vs Price: This positively correlates the distance from the MRT Station with a house's price. Meaning that the closer a house is to the station the more it increases in value.

Number of convenience store to price: There seems to be a linear relation between the number of convenience stores and the price of a house. It seems that the more stores there are, the pricier a house becomes.

2. Individual correlations.

```
cor_age_price <- cor(house$age, house$price)

cor_mrt_dist_price <- cor(house$mrt.dist, house$price)

cor_num_conv_stores_price <- cor(house$num.conv.stores, house$price)

cor_age_price
```

```
## [1] -0.210567
```

```
cor_mrt_dist_price
```

```
## [1] -0.6736129
```

```
cor_num_conv_stores_price
```

```
## [1] 0.5710049
```

3. Correlation matrix for all variables.

```
cor(house)
```

```
##                        age    mrt.dist num.conv.stores      price
## age              1.00000000  0.02562205      0.04959251 -0.2105670
## mrt.dist         0.02562205  1.00000000     -0.60251914 -0.6736129
## num.conv.stores  0.04959251 -0.60251914      1.00000000  0.5710049
## price           -0.21056705 -0.67361286      0.57100491  1.0000000
```

```
# The y = house specification asks R to compute all correlations in the house
# data frame.

house |> cor(y = house, method = "pearson") |>  round(3)
```

```
##                   age mrt.dist num.conv.stores  price
## age             1.000    0.026           0.050 -0.211
## mrt.dist        0.026    1.000          -0.603 -0.674
## num.conv.stores 0.050   -0.603           1.000  0.571
## price          -0.211   -0.674           0.571  1.000
```

```
# The correlation coefficent between number of convenience stores and distance
# to mrt station is -0.603.
```

The correlation suggests that there is a negative correlation between the number of convenienve stores and the MRT Station distance. This means that as the MRT distance increases the number of convenience stores decreases.