housing-market-influence

Vasavi Kullanakoppal

2024-11-15

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggplot2)

# Load Data set
boston_housing <- read.csv("BostonHousing.csv")

# View the structure of the dataset
str(boston_housing)
```

```
## 'data.frame':    506 obs. of  14 variables:
##  $ crim   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
##  $ zn     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
##  $ indus  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
##  $ chas   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ nox    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
##  $ rm     : num  6.58 6.42 7.18 7 7.15 ...
```

```
##  $ age    : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
##  $ dis    : num  4.09 4.97 4.97 6.06 6.06 ...
##  $ rad    : int  1 2 2 3 3 3 5 5 5 5 ...
##  $ tax    : int  296 242 242 222 222 222 311 311 311 311 ...
##  $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
##  $ b      : num  397 397 393 395 397 ...
##  $ lstat  : num  4.98 9.14 4.03 2.94 5.33 ...
##  $ medv   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

```r
# Check for missing values
colSums(is.na(boston_housing))
```

```
##    crim      zn   indus    chas     nox      rm     age     dis     rad     tax
##       0       0       0       0       0       5       0       0       0       0
## ptratio       b   lstat    medv
##       0       0       0       0
```

```r
# Remove rows with missing values
boston_housing <- na.omit(boston_housing)
# Check for columns with all zeros
zero_columns <- sapply(boston_housing, function(col) all(col == 0))
zero_columns
```

```
##    crim      zn   indus    chas     nox      rm     age     dis     rad     tax
##   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
## ptratio       b   lstat    medv
##   FALSE   FALSE   FALSE   FALSE
```

```r
# Rename columns
colnames(boston_housing) <- c("CrimeRate", "ResidentialZone", "IndustrialProportion", "CharlesRiver", "NitrogenOxideConcentration",
                              "AvgRooms", "AgeOfHome", "DistanceToEmployment", "HighwayAccess", "TaxRate",
                              "PupilTeacherRatio", "BlackPopulation", "LowerStatus", "MedianValue")

# Create a new variable for high crime neighborhoods
boston_housing <- boston_housing %>% mutate(HighCrime = ifelse(CrimeRate > median(CrimeRate), "Yes", "No"))
# Create interaction term for pollution and industrial zones
boston_housing <- boston_housing %>% mutate(PollutionIndustry = NitrogenOxideConcentration * IndustrialProportion)

head(boston_housing)
```

```
##   CrimeRate ResidentialZone IndustrialProportion CharlesRiver
```

```
## 1   0.00632              18                2.31           0
## 2   0.02731               0                7.07           0
## 3   0.02729               0                7.07           0
## 4   0.03237               0                2.18           0
## 5   0.06905               0                2.18           0
## 6   0.02985               0                2.18           0
##   NitrogenOxideConcentration AvgRooms AgeOfHome DistanceToEmployment
## 1                      0.538    6.575      65.2               4.0900
## 2                      0.469    6.421      78.9               4.9671
## 3                      0.469    7.185      61.1               4.9671
## 4                      0.458    6.998      45.8               6.0622
## 5                      0.458    7.147      54.2               6.0622
## 6                      0.458    6.430      58.7               6.0622
##   HighwayAccess TaxRate PupilTeacherRatio BlackPopulation LowerStatus
## 1             1     296              15.3          396.90        4.98
## 2             2     242              17.8          396.90        9.14
## 3             2     242              17.8          392.83        4.03
## 4             3     222              18.7          394.63        2.94
## 5             3     222              18.7          396.90        5.33
## 6             3     222              18.7          394.12        5.21
##   MedianValue HighCrime PollutionIndustry
## 1        24.0        No           1.24278
## 2        21.6        No           3.31583
## 3        34.7        No           3.31583
## 4        33.4        No           0.99844
## 5        36.2        No           0.99844
## 6        28.7        No           0.99844
```

```r
# Filter homes that border the Charles River
river_homes <- boston_housing %>% filter(CharlesRiver == 1)
head (river_homes)
```

```
##   CrimeRate ResidentialZone IndustrialProportion CharlesRiver
## 1   3.32105               0                19.58           1
## 2   1.12658               0                19.58           1
## 3   1.41385               0                19.58           1
## 4   3.53501               0                19.58           1
## 5   1.27346               0                19.58           1
## 6   1.83377               0                19.58           1
##   NitrogenOxideConcentration AvgRooms AgeOfHome DistanceToEmployment
```

```
## 1                     0.871    5.403     100.0              1.3216
## 2                     0.871    5.012      88.0              1.6102
## 3                     0.871    6.129      96.0              1.7494
## 4                     0.871    6.152      82.6              1.7455
## 5                     0.605    6.250      92.6              1.7984
## 6                     0.605    7.802      98.2              2.0407
##   HighwayAccess TaxRate PupilTeacherRatio BlackPopulation LowerStatus
## 1             5     403              14.7          396.90       26.82
## 2             5     403              14.7          343.28       12.12
## 3             5     403              14.7          321.02       15.12
## 4             5     403              14.7           88.01       15.02
## 5             5     403              14.7          338.92        5.50
## 6             5     403              14.7          389.61        1.92
##   MedianValue HighCrime PollutionIndustry
## 1        13.4       Yes          17.05418
## 2        15.3       Yes          17.05418
## 3        17.0       Yes          17.05418
## 4        15.6       Yes          17.05418
## 5        27.0       Yes          11.84590
## 6        50.0       Yes          11.84590
```

```r
# Group data by proximity to Charles River and calculate average prices
avg_price_by_river <- boston_housing %>% group_by(CharlesRiver) %>% summarize(AveragePrice = mean(MedianValue))
avg_price_by_river
```

```
## # A tibble: 2 x 2
##   CharlesRiver AveragePrice
##          <int>        <dbl>
## 1            0         22.1
## 2            1         28.4
```

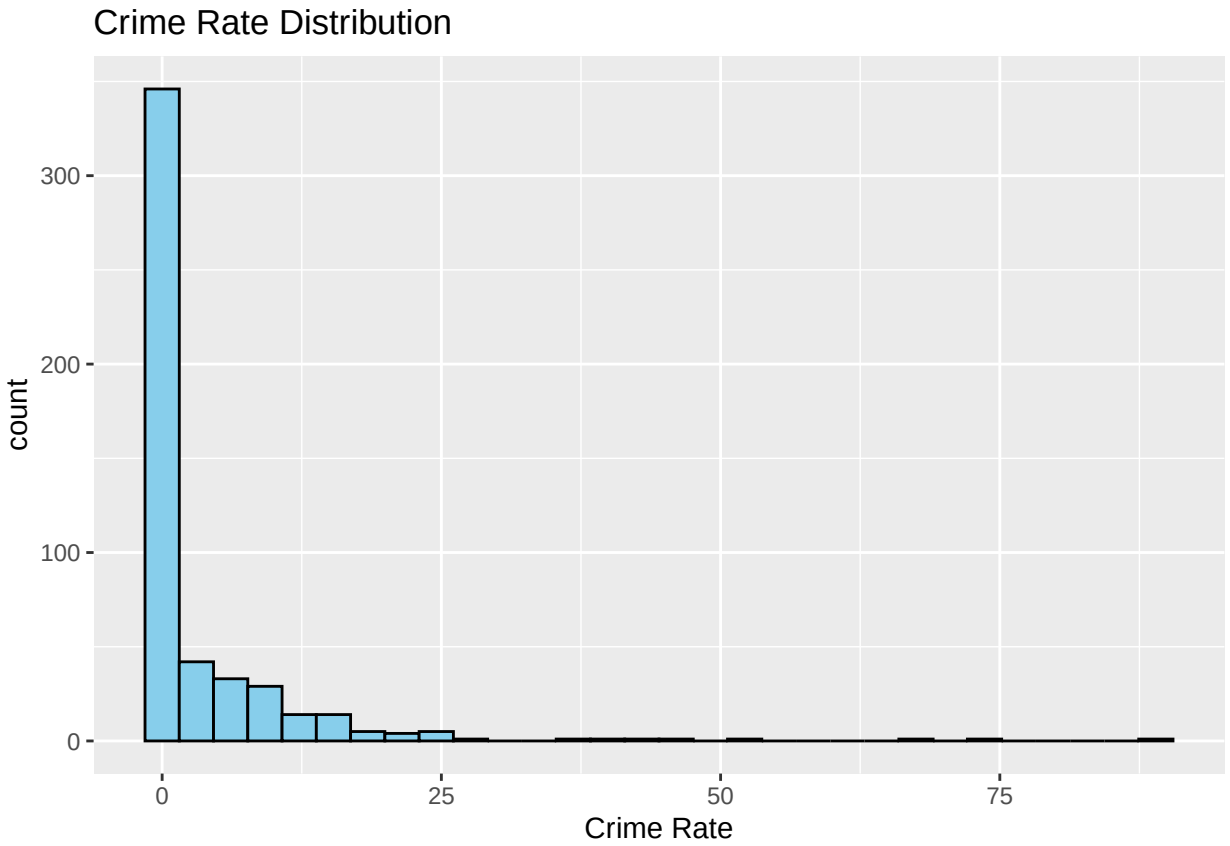
```r
# Mean and median prices based on proximity to the Charles River
mean_median_price_by_river <- boston_housing %>% group_by(CharlesRiver) %>% summarize(MeanPrice = mean(MedianValue), MedianPrice = median(
mean_median_price_by_river
```

```
## # A tibble: 2 x 3
##   CharlesRiver MeanPrice MedianPrice
##          <int>     <dbl>       <dbl>
## 1            0      22.1        20.9
## 2            1      28.4        23.3
```

```r
# Calculate the median housing price by highway access
median_price_by_highway <- boston_housing %>% group_by(ResidentialZone) %>% summarize(MedianPrice = median(MedianValue))
median_price_by_highway
```

```
## # A tibble: 26 x 2
##    ResidentialZone MedianPrice
##              <dbl>       <dbl>
##  1             0          19.8
##  2            12.5        20.9
##  3            17.5        33
##  4            18          24
##  5            20          35.2
##  6            21          22.0
##  7            22          24.4
##  8            25          22.9
##  9            28          22.9
## 10            30          22.8
## # i 16 more rows
```

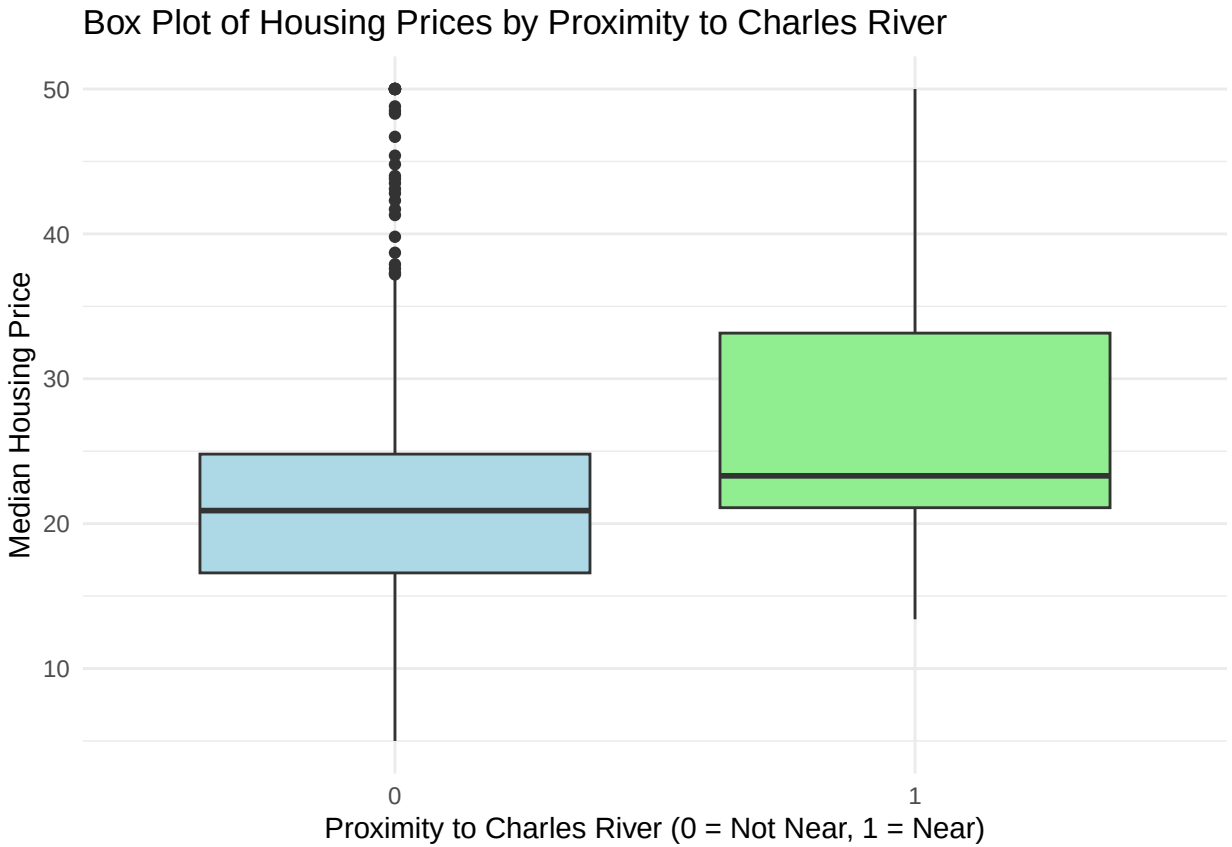
```r
# a. For Crime Rates

# Histogram for Crime Rate Distribution
ggplot(boston_housing, aes(x = CrimeRate)) +
  geom_histogram(bins = 30, color = "black", fill = "skyblue") +
  labs(title = "Crime Rate Distribution", x = "Crime Rate")
```




```r
# Descriptive statistics for CRIM (Crime Rate per Capita)
CrimeRate_stats <- boston_housing %>%
  reframe(
    mean_CrimeRate = mean(CrimeRate, na.rm = TRUE),
    median_CrimeRate = median(CrimeRate, na.rm = TRUE),
    sd_CrimeRate = sd(CrimeRate, na.rm = TRUE),
    min_CrimeRate = min(CrimeRate, na.rm = TRUE),
    max_CrimeRate = max(CrimeRate, na.rm = TRUE),
    CrimeRate_quantiles = quantile(CrimeRate, probs = c(0.25, 0.5, 0.75), na.rm = TRUE)
  )

print(CrimeRate_stats)
```

```
##   mean_CrimeRate median_CrimeRate sd_CrimeRate min_CrimeRate max_CrimeRate
## 1       3.647414          0.26169     8.637688       0.00632       88.9762
## 2       3.647414          0.26169     8.637688       0.00632       88.9762
## 3       3.647414          0.26169     8.637688       0.00632       88.9762
##   CrimeRate_quantiles
## 1             0.08199
## 2             0.26169
## 3             3.69311
```

```r
# The average crime rate is around 3.65. The median crime rate is lower at 0.26.
# This indicates that the data is skewed which means most areas have a low crime rate,
# but there are a few with very high crime rates.
# The high standard deviation of 8.64 suggests the idea of outliers with exceptionally high crime rates.
# The lowest crime rate is 0.00632 and the maximum is 88.9762, which may be the outliers affecting the mean.

# b. For Charles River

# Box plot for MEDV by Charles River
ggplot(boston_housing, aes(x = factor(CharlesRiver), y = MedianValue)) +
  geom_boxplot(fill = c("lightblue", "lightgreen")) +
  labs(title = "Box Plot of Housing Prices by Proximity to Charles River",
       x = "Proximity to Charles River (0 = Not Near, 1 = Near)",
       y = "Median Housing Price") +
  theme_minimal()
```




```r
# Summary table by proximity to Charles River
avg_price_by_river <- boston_housing %>%
  group_by(CharlesRiver) %>%
  summarize(MeanPrice = mean(MedianValue), MedianPrice = median(MedianValue), SDPrice = sd(MedianValue))

# Descriptive statistics for Proximity to Charles River
CharlesRiver_stats <- boston_housing %>%
  group_by(CharlesRiver) %>%
  summarize(count = n()) %>%
  mutate(proportion = count / sum(count))
print(CharlesRiver_stats)
```

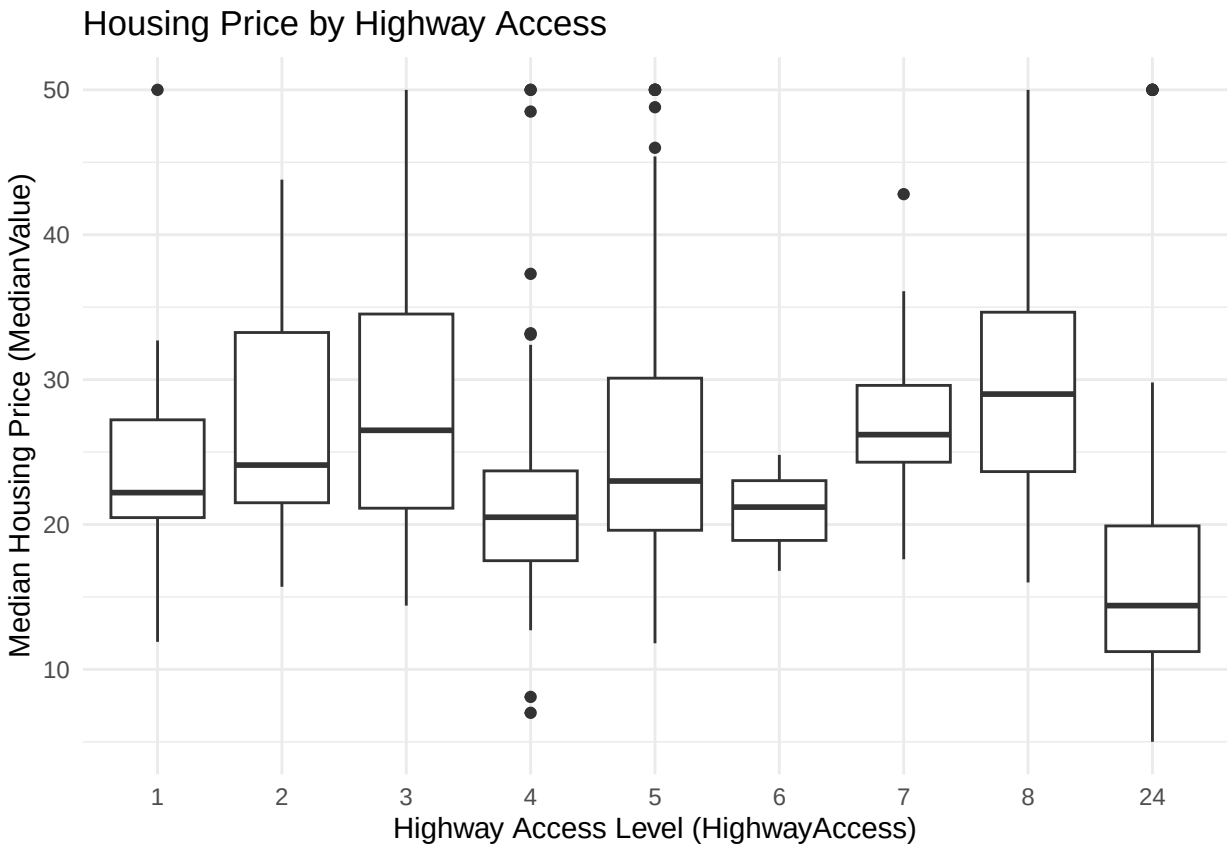
```
## # A tibble: 2 x 3
##   CharlesRiver count proportion
##          <int> <int>      <dbl>
## 1            0   466     0.930
## 2            1    35     0.0699
```

```r
# There are 35 homes and only a small proportion of 6.9% of homes are located near the Charles River
# and 466 homes and majority of homes (93%) are situated away from the river.
# The box plot shows that the proximity to the Charles River has a positive influence on housing prices.
# Also, it shows that there is more variability in prices for homes near the river,
# possibly due to different property types or amenities.

# c. For Highway Access

# Descriptive statistics for Accessibility to Highways
HighwayAccess_stats <- boston_housing %>%
  group_by(HighwayAccess) %>%
  summarize(count = n()) %>%
  mutate(proportion = count / sum(count))
print(HighwayAccess_stats)
```

```
## # A tibble: 9 x 3
##   HighwayAccess count proportion
##           <int> <int>      <dbl>
## 1             1    20     0.0399
## 2             2    23     0.0459
## 3             3    38     0.0758
## 4             4   109     0.218
## 5             5   113     0.226
## 6             6    26     0.0519
## 7             7    17     0.0339
## 8             8    23     0.0459
## 9            24   132     0.263
```

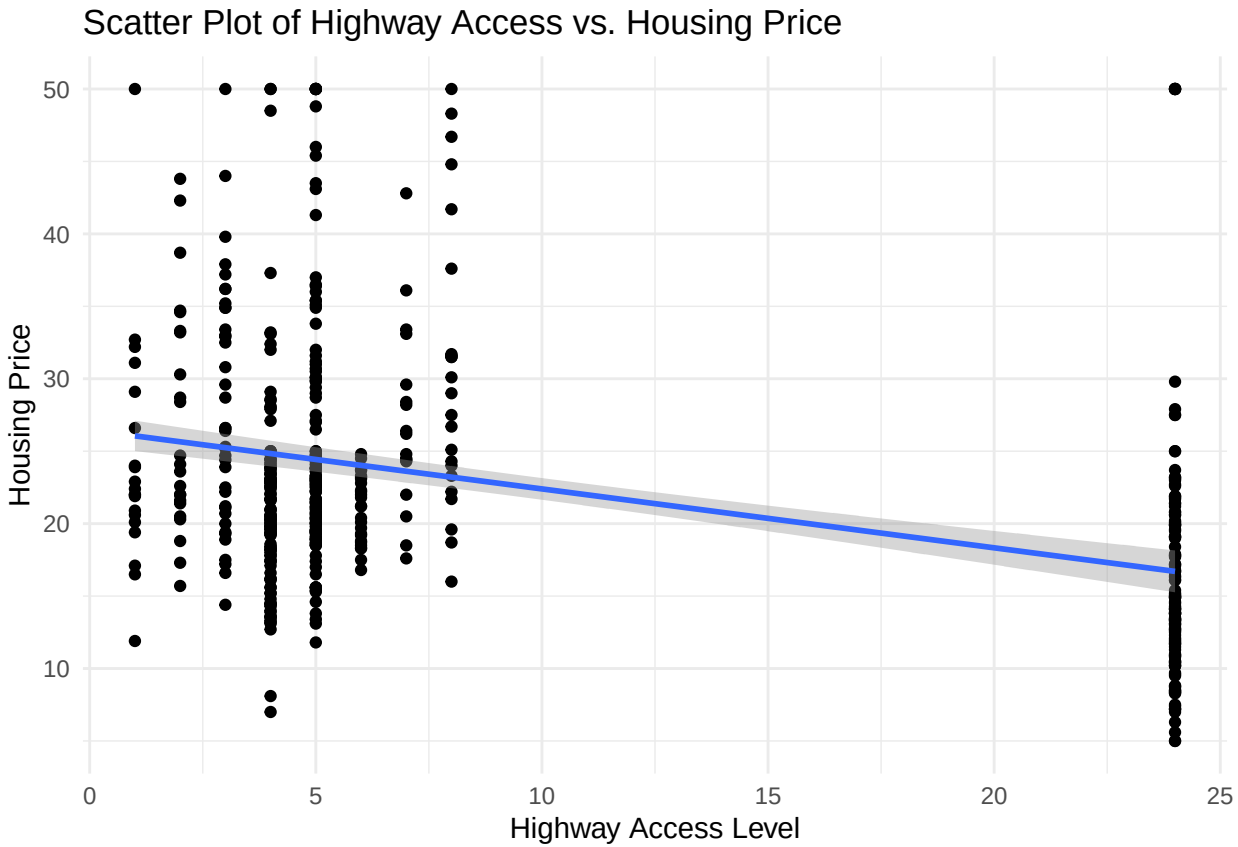
```r
# Box plot for MEDV by Highway Access
ggplot(boston_housing, aes(x = factor(HighwayAccess), y = MedianValue)) +
  geom_boxplot() +
  labs(title = "Housing Price by Highway Access",
       x = "Highway Access Level (HighwayAccess)",
       y = "Median Housing Price (MedianValue)") +
```

```
  theme_minimal()
```




```
#Scatter plot with trend line
ggplot(boston_housing, aes(x = HighwayAccess, y = MedianValue)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Scatter Plot of Highway Access vs. Housing Price",
       x = "Highway Access Level",
       y = "Housing Price") +
  theme_minimal()
```

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




```
# The scatter plot with a trend line shows a weak negative relationship
# between highway access level and median housing price,
# suggesting that homes closer to major highways tend to have slightly lower prices.
# However, the wide scatter of points indicates that this alone is not a strong predictor of housing prices,
# and other factors likely influence the variability observed.


# Multivariable Regression Analysis

# Multiple regression model including Highway Access, Proximity of Charles River and Crime Rate to predict Median Value

model <- lm(MedianValue ~ HighwayAccess + CharlesRiver + CrimeRate, data = boston_housing)
```

```r
summary(model)
```

```
##
## Call:
## lm(formula = MedianValue ~ HighwayAccess + CharlesRiver + CrimeRate,
##     data = boston_housing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.427  -4.989  -1.856   2.989  32.853
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    25.4795     0.5688  44.797  < 2e-16 ***
## HighwayAccess  -0.2516     0.0540  -4.660 4.07e-06 ***
## CharlesRiver    5.7643     1.4456   3.987 7.68e-05 ***
## CrimeRate      -0.2484     0.0547  -4.540 7.05e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.23 on 497 degrees of freedom
## Multiple R-squared:  0.2101, Adjusted R-squared:  0.2053
## F-statistic: 44.07 on 3 and 497 DF,  p-value: < 2.2e-16
```

```r
# Residuals of -17.427 and 32.853 shows some variability in the model's accuracy.
# The residuals of about -5 to +3 suggest reasonably consistent prediction accuracy within a typical range.
# With highway access and crime rate at zero and no proximity to the Charles River,
# the average predicted housing price is 25.48 units.
# Each unit increase in highway access is associated with a decrease in housing price by 0.2516 units.
# The p-value 4.07e-06 indicates this relationship is statistically significant.
# Being located near the Charles River is associated with an increase in housing prices by 5.7643.
# The p-value 7.68e-05 indicates this positive relationship is statistically significant as well.
# Each unit increase in the crime rate is associated with a decrease in housing prices by 0.2484.
# The p-value 7.05e-06 makes this a statistically significant predictor.

# Model Fit:

# Residual Standard Error shows the model's predictions are off by about 8.23 units.
# R-squared shows that about 21.01% of the variation in housing prices
```

```
# is explained by Highway Access, Charles River proximity, and Crime Rate combined.
# Adjusted R-squared of 0.2053 shows the model has some predictive power,
# though a large portion of housing price variability remains unexplained.
# The very low p-value < 2.2e-16 means that, collectively,
# the predictors significantly explain the variability in housing prices.

# Despite these significant relationships, the R-squared value of ~21% indicates
# that the model only captures a small portion of the variability in housing prices,
# suggesting that other factors are also important in determining housing values.
# Further modeling, potentially incorporating more variables, could improve the predictive power.
```