Stat 520 Project

Brian Bruxvoort

Load libraries and data

```
library(tidyverse)
```

```
Warning: package 'ggplot2' was built under R version 4.3.3
```

```
Warning: package 'tidyr' was built under R version 4.3.3
```

```
Warning: package 'purrr' was built under R version 4.3.3
```

```
Warning: package 'dplyr' was built under R version 4.3.3
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.5.0     v tibble    3.2.1
v lubridate 1.9.2     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 beco
```

```
library(tidymodels)
```

```
Warning: package 'tidymodels' was built under R version 4.3.3
```

```
-- Attaching packages ------------------------------------- tidymodels 1.2.0 --
v broom        1.0.5      v rsample      1.2.1
v dials        1.2.1      v tune         1.2.0
v infer        1.0.7      v workflows    1.1.4
v modeldata    1.3.0      v workflowsets 1.1.0
v parsnip      1.2.1      v yardstick    1.3.1
v recipes      1.0.10

Warning: package 'dials' was built under R version 4.3.3

Warning: package 'scales' was built under R version 4.3.3

Warning: package 'infer' was built under R version 4.3.3

Warning: package 'modeldata' was built under R version 4.3.3

Warning: package 'parsnip' was built under R version 4.3.3

Warning: package 'recipes' was built under R version 4.3.3

Warning: package 'rsample' was built under R version 4.3.3

Warning: package 'tune' was built under R version 4.3.3

Warning: package 'workflows' was built under R version 4.3.3

Warning: package 'workflowsets' was built under R version 4.3.3

Warning: package 'yardstick' was built under R version 4.3.3

-- Conflicts ---------------------------------------- tidymodels_conflicts() --
x scales::discard() masks purrr::discard()
x dplyr::filter()   masks stats::filter()
x recipes::fixed()  masks stringr::fixed()
x dplyr::lag()      masks stats::lag()
x yardstick::spec() masks readr::spec()
x recipes::step()   masks stats::step()
* Use suppressPackageStartupMessages() to eliminate package startup messages
```

```
library(corrplot)
```

```
Warning: package 'corrplot' was built under R version 4.3.2
```

```
corrplot 0.92 loaded
```

```
library(rpart)
```

```
Attaching package: 'rpart'
```

```
The following object is masked from 'package:dials':

    prune
```

```
library(rpart.plot)
library(randomForest)
```

```
randomForest 4.7-1.1
Type rfNews() to see new features/changes/bug fixes.
```

```
Attaching package: 'randomForest'
```

```
The following object is masked from 'package:dplyr':

    combine
```

```
The following object is masked from 'package:ggplot2':

    margin
```

```
library(caret)
```

```
Loading required package: lattice
```

```
Attaching package: 'caret'
```

```
The following objects are masked from 'package:yardstick':

    precision, recall, sensitivity, specificity

The following object is masked from 'package:purrr':

    lift
```

```
library(car)
```

```
Warning: package 'car' was built under R version 4.3.2

Loading required package: carData

Attaching package: 'car'

The following object is masked from 'package:dplyr':

    recode

The following object is masked from 'package:purrr':

    some
```

```
library(stats)
library(ranger)
```

```
Warning: package 'ranger' was built under R version 4.3.3


Attaching package: 'ranger'

The following object is masked from 'package:randomForest':

    importance
```

```
library(factoextra)
```

```
Warning: package 'factoextra' was built under R version 4.3.2
```

```
Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```r
realtor <- read.csv("~/Truman_Stat_520/Project/realtor-data.zip.csv")
pollution <- read.csv("~/Truman_Stat_520/Project/pollution_us_2000_2016.csv")
```

Clean Data

```r
realtor2 <- realtor %>%
  select(city, state, price) %>%
  rename(City = city, State = state, Price = price) %>%
  group_by(City, State) %>%
  summarise(
    MeanPrice = mean(Price, na.rm = TRUE),
    MedianPrice = median(Price, na.rm = TRUE),
    MaxPrice = max(Price, na.rm = FALSE),
    .groups = 'drop'
  )
```

```r
pollution2 <- pollution %>%
  select(State, City, NO2.AQI, O3.AQI, SO2.AQI, CO.AQI, Date.Local)
```

```r
joined_data <- inner_join(realtor2, pollution2, by = c("City", "State"))
```

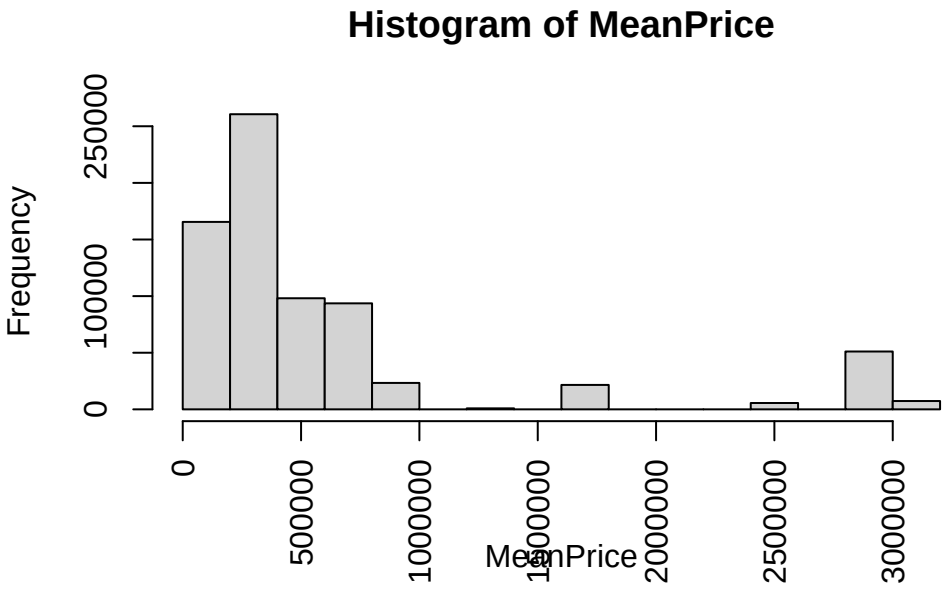
Data Exploration

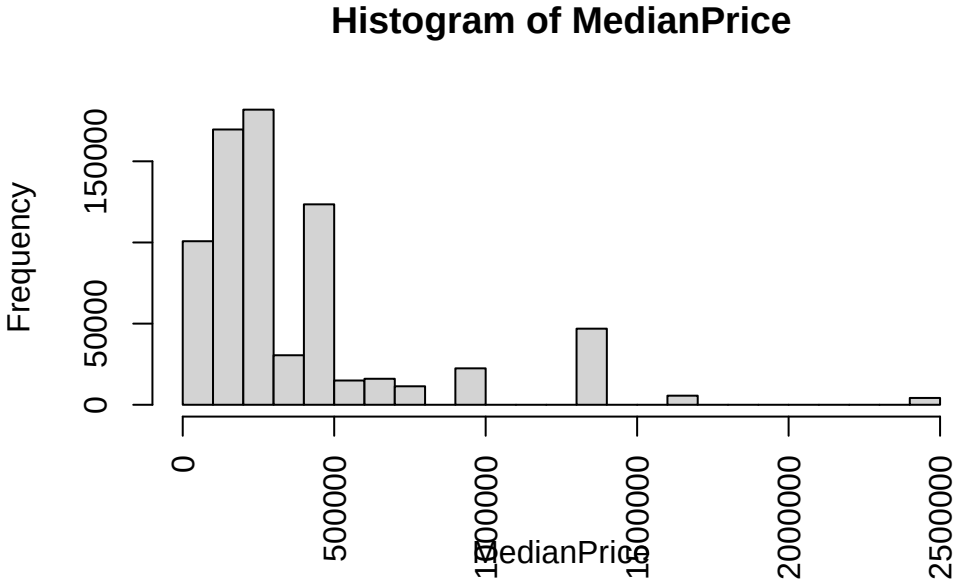
```r
summary(joined_data)
```

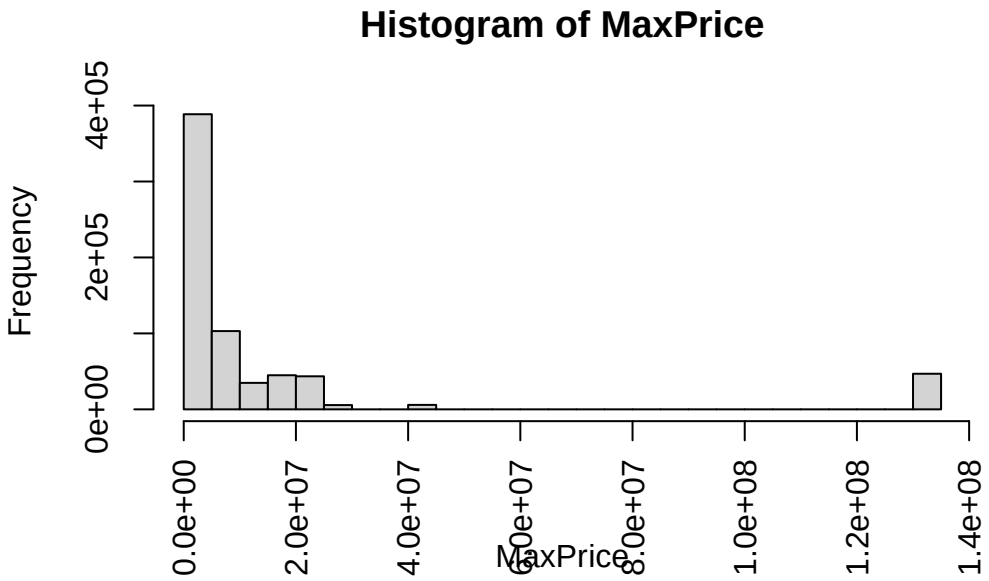
```
     City              State             MeanPrice        MedianPrice
 Length:727649      Length:727649      Min.   :  42765   Min.   :  45000
 Class :character   Class :character   1st Qu.: 204050   1st Qu.: 149000
 Mode  :character   Mode  :character   Median : 335658   Median : 277500
                                       Mean   : 621899   Mean   : 383565
                                       3rd Qu.: 614855   3rd Qu.: 430000
```

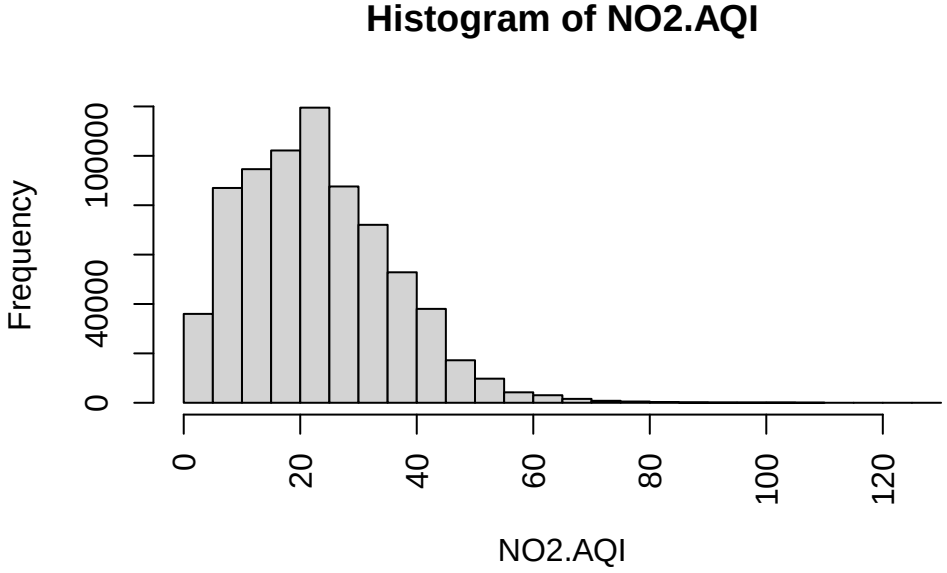
```
                                  Max.   :3070828   Max.   :2499000

   MaxPrice            NO2.AQI           O3.AQI          SO2.AQI
Min.   :    85000   Min.   :  0.00   Min.   :  0.00   Min.   :  0.0
1st Qu.:  1200000   1st Qu.: 14.00   1st Qu.: 25.00   1st Qu.:  1.0
Median :  3950000   Median : 23.00   Median : 32.00   Median :  4.0
Mean   : 16042976   Mean   : 23.59   Mean   : 36.06   Mean   : 10.3
3rd Qu.: 14000000   3rd Qu.: 32.00   3rd Qu.: 42.00   3rd Qu.: 13.0
Max.   :135000000   Max.   :128.00   Max.   :207.00   Max.   :200.0
NA's   :54382                                         NA's   :363665
    CO.AQI        Date.Local
Min.   : 0.0     Length:727649
1st Qu.: 2.0     Class :character
Median : 5.0     Mode  :character
Mean   : 4.9
3rd Qu.: 7.0
Max.   :64.0
NA's   :363830
```

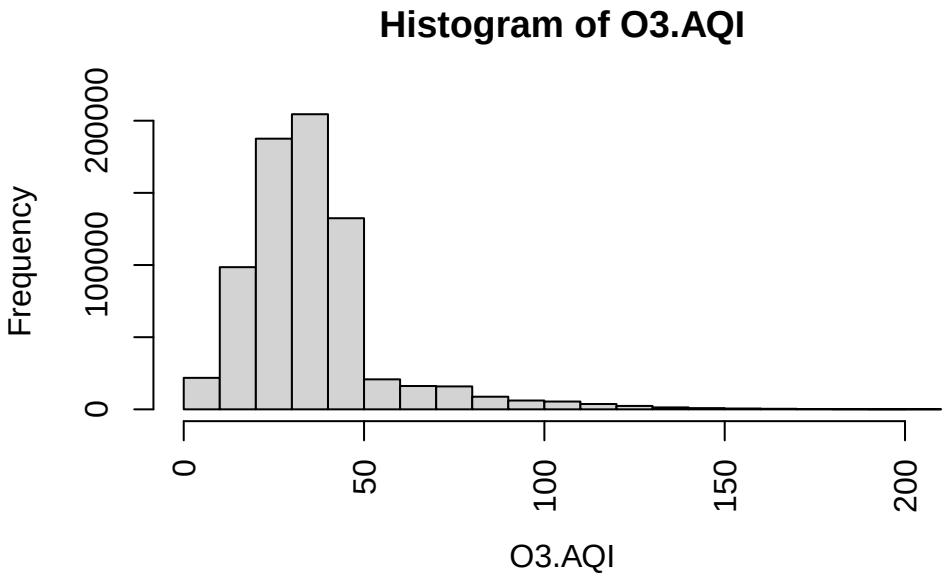
```r
numeric_columns <- sapply(joined_data, is.numeric)

numeric_variable_names <- names(joined_data)[numeric_columns]

output <- lapply(numeric_variable_names, function(var_name) {
  hist(joined_data[[var_name]], main = paste("Histogram of", var_name), xlab = var_name, x
  axis(1, las=2)
})
```

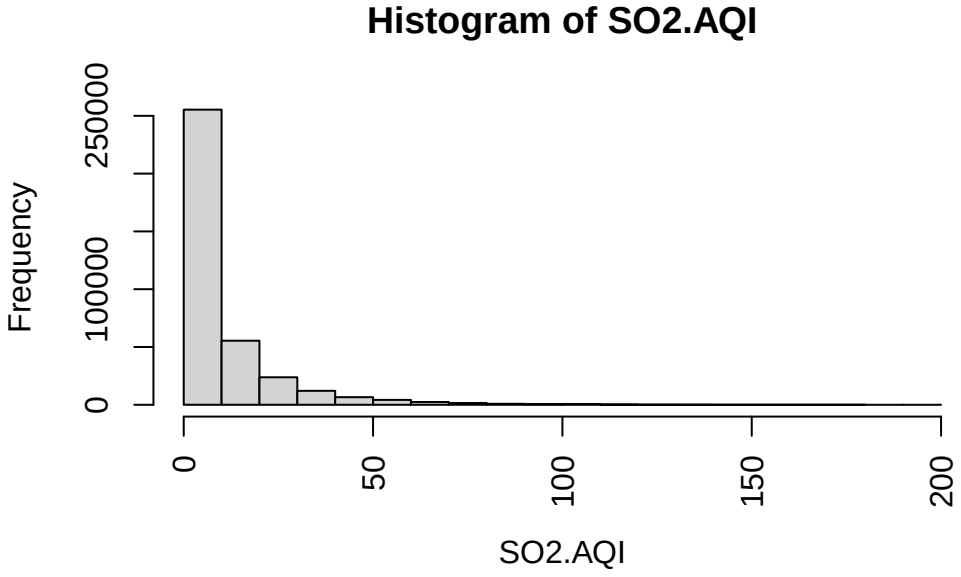

Histogram of MeanPrice
Frequency
0
100000
250000
0
500000
1000000
1500000
2000000
2500000
3000000
MeanPrice

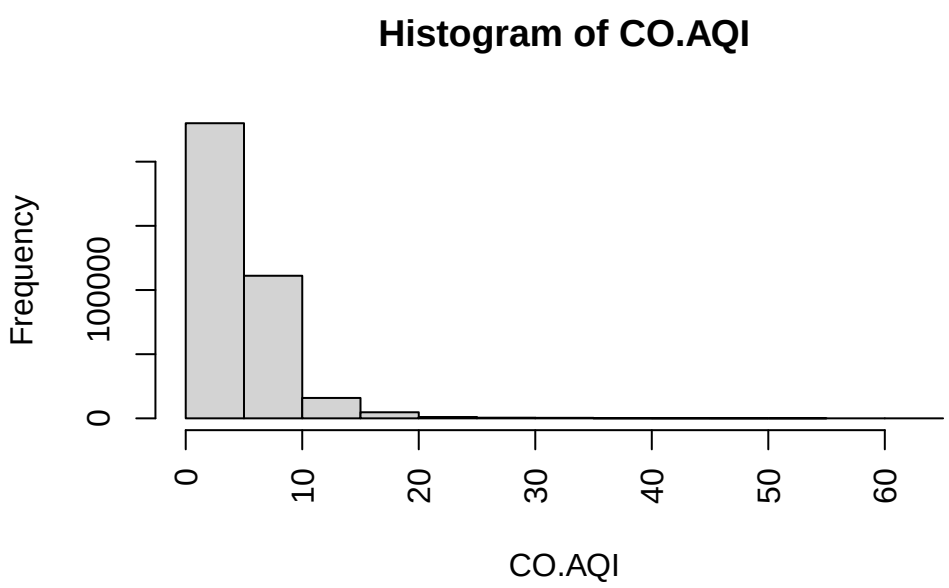

Histogram of MedianPrice
Frequency
0
50000
150000
0
500000
1000000
1500000
2000000
2500000
MedianPrice


Histogram of MaxPrice
Frequency
0e+00
2e+05
4e+05
0.0e+00
2.0e+07
4.0e+07
6.0e+07
8.0e+07
1.0e+08
1.2e+08
1.4e+08
MaxPrice


Histogram of NO2.AQI
Frequency
0
40000
100000
0
20
40
60
80
100
120
NO2.AQI

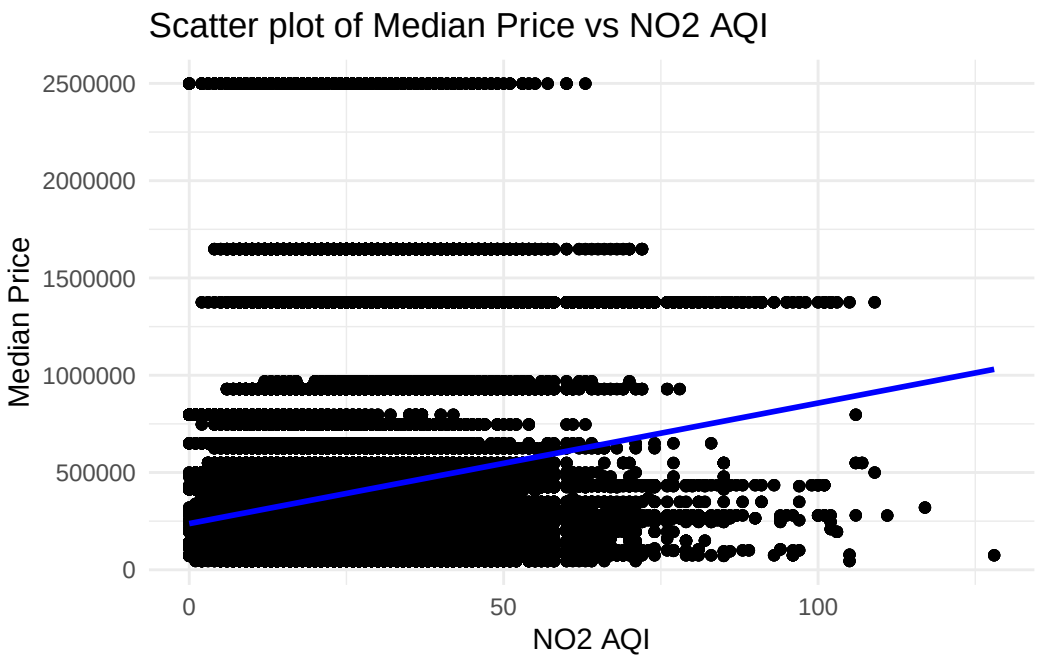

Histogram of O3.AQI
Frequency
0
100000
200000
0
50
100
150
200
O3.AQI


Histogram of SO2.AQI
Frequency
0
100000
250000
0
50
100
150
200
SO2.AQI

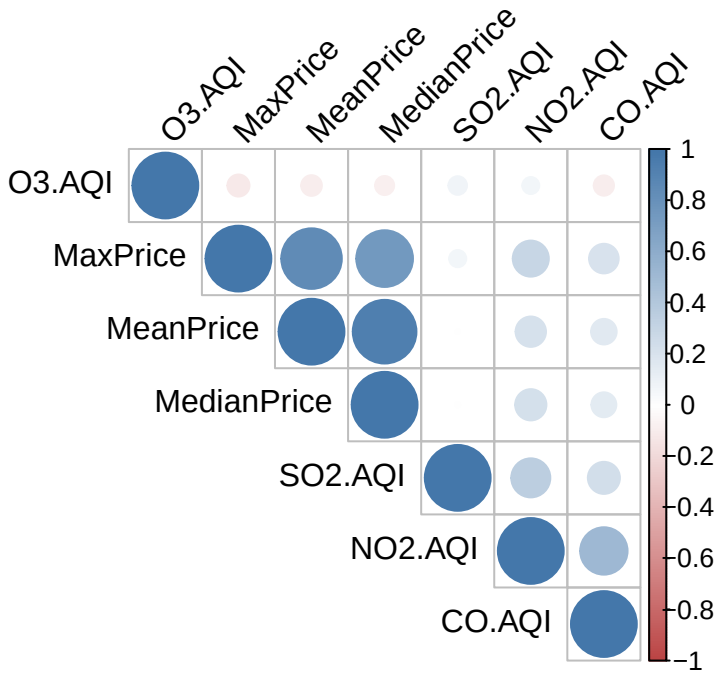



```
ggplot(joined_data, aes(x = NO2.AQI, y = MedianPrice)) +
  geom_point() +
  geom_smooth(method = "lm", col = "blue") +
  theme_minimal() +
  labs(title = "Scatter plot of Median Price vs NO2 AQI", x = "NO2 AQI", y = "Median Price
```

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




```r
cor_matrix <- cor(joined_data[, sapply(joined_data, is.numeric)], use = "complete.obs")
corrplot(cor_matrix, method = "circle", type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45,
         col = colorRampPalette(c("#BB4444", "white", "#4477AA"))(200))
```




```r
joined_data$MaxPrice[is.na(joined_data$MaxPrice)] <- mean(joined_data$MaxPrice, na.rm = TR

# Median imputation
joined_data$SO2.AQI[is.na(joined_data$SO2.AQI)] <- median(joined_data$SO2.AQI, na.rm = TRU
joined_data$CO.AQI[is.na(joined_data$CO.AQI)] <- median(joined_data$CO.AQI, na.rm = TRUE)
```

```r
summary(joined_data)
```

```
    City              State             MeanPrice        MedianPrice
Length:727649      Length:727649      Min.   :  42765   Min.   :  45000
Class :character   Class :character   1st Qu.: 204050   1st Qu.: 149000
Mode  :character   Mode  :character   Median : 335658   Median : 277500
                                      Mean   : 621899   Mean   : 383565
                                      3rd Qu.: 614855   3rd Qu.: 430000
                                      Max.   :3070828   Max.   :2499000
   MaxPrice            NO2.AQI          O3.AQI          SO2.AQI
Min.   :    85000   Min.   :  0.00   Min.   :  0.00   Min.   :  0.000
1st Qu.:  1223600   1st Qu.: 14.00   1st Qu.: 25.00   1st Qu.:  4.000
Median :  4900000   Median : 23.00   Median : 32.00   Median :  4.000
Mean   : 16042976   Mean   : 23.59   Mean   : 36.06   Mean   :  7.144
3rd Qu.: 15986000   3rd Qu.: 32.00   3rd Qu.: 42.00   3rd Qu.:  4.000
```

```
 Max.   :135000000   Max.   :128.00   Max.   :207.00   Max.   :200.000
    CO.AQI        Date.Local
 Min.   : 0.000   Length:727649
 1st Qu.: 5.000   Class :character
 Median : 5.000   Mode  :character
 Mean   : 4.953
 3rd Qu.: 5.000
 Max.   :64.000
```

Multiple Linear Regression

```r
linear.model <- lm(MedianPrice ~ NO2.AQI + SO2.AQI + O3.AQI + CO.AQI, data = joined_data)
summary(linear.model)
```

```
Call:
lm(formula = MedianPrice ~ NO2.AQI + SO2.AQI + O3.AQI + CO.AQI,
    data = joined_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1010593  -221238  -111895    82665  2371699

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 292527.12    1308.63  223.54   <2e-16 ***
NO2.AQI       6642.54      37.42  177.51   <2e-16 ***
SO2.AQI      -1894.00      40.45  -46.82   <2e-16 ***
O3.AQI       -1704.19      21.72  -78.46   <2e-16 ***
CO.AQI        1880.68     174.34   10.79   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 371200 on 727644 degrees of freedom
Multiple R-squared:  0.05531,   Adjusted R-squared:  0.05531
F-statistic: 1.065e+04 on 4 and 727644 DF,  p-value: < 2.2e-16
```

```r
# Make predictions
predictions <- predict(linear.model, newdata = joined_data)
```

```r
# Actual values
actuals <- joined_data$MedianPrice

# Calculating the residuals
residuals <- predictions - actuals

# Calculating RMSE
rmse <- sqrt(mean(residuals^2))

# Calculating MAE
mae <- mean(abs(residuals))

# Print the results
cat("RMSE: ", rmse, "\n")
```

```
RMSE:  371212.5
```

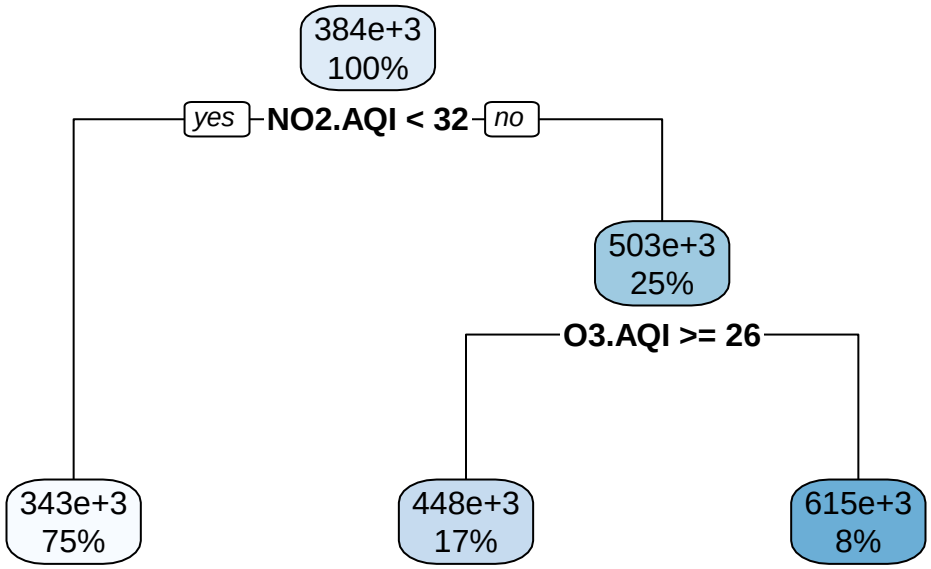
```r
cat("MAE: ", mae, "\n")
```

```
MAE:  253927.4
```

CART

```r
cart.model <- rpart(MedianPrice ~ NO2.AQI + O3.AQI + SO2.AQI + CO.AQI,
                    data=joined_data,
                    method="anova")
```

```r
rpart.plot(cart.model)
```




```
predictions <- predict(cart.model, newdata=joined_data)
```

```
actual_values <- joined_data$MedianPrice

# MAE
mae <- mean(abs(predictions - actual_values))
print(paste("Mean Absolute Error (MAE):", mae))
```

```
[1] "Mean Absolute Error (MAE): 253855.013891151"
```

```
# MSE
mse <- mean((predictions - actual_values)^2)
print(paste("Mean Squared Error (MSE):", mse))
```

```
[1] "Mean Squared Error (MSE): 139391762037.145"
```

```
# RMSE
rmse <- sqrt(mse)
print(paste("Root Mean Squared Error (RMSE):", rmse))
```

```
[1] "Root Mean Squared Error (RMSE): 373352.061782368"
```

```r
# R-squared
rss <- sum((predictions - actual_values)^2)
tss <- sum((actual_values - mean(actual_values))^2)
rsquared <- 1 - (rss/tss)
print(paste("R-squared (R²):", rsquared))
```

```
[1] "R-squared (R²): 0.0443901148961726"
```

Random Forest

```r
rf.model <- ranger(MedianPrice ~ NO2.AQI + O3.AQI + SO2.AQI + CO.AQI,
                   data=joined_data,
                   method="anova")
```

```
Warning in ranger(MedianPrice ~ NO2.AQI + O3.AQI + SO2.AQI + CO.AQI, data =
joined_data, : Unused arguments: method
```

```
Growing trees.. Progress: 12%. Estimated remaining time: 3 minutes, 47 seconds.
Growing trees.. Progress: 25%. Estimated remaining time: 3 minutes, 8 seconds.
Growing trees.. Progress: 37%. Estimated remaining time: 2 minutes, 40 seconds.
Growing trees.. Progress: 49%. Estimated remaining time: 2 minutes, 10 seconds.
Growing trees.. Progress: 62%. Estimated remaining time: 1 minute, 37 seconds.
Growing trees.. Progress: 75%. Estimated remaining time: 1 minute, 4 seconds.
Growing trees.. Progress: 87%. Estimated remaining time: 32 seconds.
Growing trees.. Progress: 100%. Estimated remaining time: 0 seconds.
```

```r
rf.predictions <- predict(rf.model, data=joined_data, type = "response")
```

```
Predicting.. Progress: 81%. Estimated remaining time: 7 seconds.
```

```r
predicted_values <- rf.predictions$predictions

# Actual values
actual_values <- joined_data$MedianPrice
```

```r
# MAE
mae <- mean(abs(predicted_values - actual_values))
print(paste("Mean Absolute Error (MAE):", mae))
```

```
[1] "Mean Absolute Error (MAE): 214292.857314077"
```

```r
# MSE
mse <- mean((predicted_values - actual_values)^2)
print(paste("Mean Squared Error (MSE):", mse))
```

```
[1] "Mean Squared Error (MSE): 104828990677.271"
```

```r
# RMSE
rmse <- sqrt(mse)
print(paste("Root Mean Squared Error (RMSE):", rmse))
```

```
[1] "Root Mean Squared Error (RMSE): 323773.054279183"
```

```r
# R-squared
rss <- sum((predicted_values - actual_values)^2)
tss <- sum((actual_values - mean(actual_values))^2)
rsquared <- 1 - (rss/tss)
print(paste("R-squared (R²):", rsquared))
```

```
[1] "R-squared (R²): 0.281337589304867"
```

PCA

```r
# Extract predictor variables
predictors <- joined_data[, c("NO2.AQI", "SO2.AQI", "O3.AQI", "CO.AQI")]

# Perform PCA on the predictors
pca_results <- prcomp(predictors, center = TRUE, scale. = TRUE) # Standardizing variables
summary(pca_results)
```

```
Importance of components:
                          PC1    PC2    PC3    PC4
Standard deviation     1.2260 1.0208 0.9308 0.7671
Proportion of Variance 0.3758 0.2605 0.2166 0.1471
Cumulative Proportion  0.3758 0.6363 0.8529 1.0000
```

```r
# Extracting the scores of the first two PCs
pc_scores <- pca_results$x[, 1:2]

# Prepare the data for regression
# Combine the PC scores with the MedianPrice
regression_data <- data.frame(cbind(pc_scores, MedianPrice = joined_data$MedianPrice))

# Linear Regression using the principal components to predict MedianPrice
model <- lm(MedianPrice ~ ., data = regression_data)
summary(model)
```

```
Call:
lm(formula = MedianPrice ~ ., data = regression_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1243530  -222229  -115389    75025  2321007

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 383565.0      440.3  871.20   <2e-16 ***
PC1          47140.2      359.1  131.27   <2e-16 ***
PC2         -37676.2      431.3  -87.36   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 375600 on 727646 degrees of freedom
Multiple R-squared:  0.03304,   Adjusted R-squared:  0.03304
F-statistic: 1.243e+04 on 2 and 727646 DF,  p-value: < 2.2e-16
```

```r
predictions <- predict(model, regression_data)

# Actual values
```

```r
actuals <- regression_data$MedianPrice

# Calculate MAE
mae <- mean(abs(predictions - actuals))

# Calculate MSE
mse <- mean((predictions - actuals)^2)

# Calculate RMSE
rmse <- sqrt(mse)

# Print the metrics
cat("MAE:", mae, "\n")
```

```
MAE: 255133.1
```

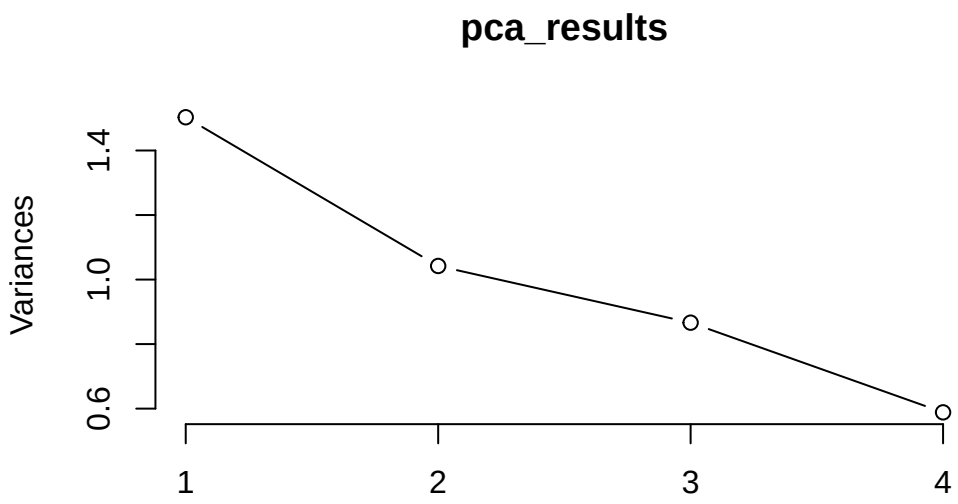
```r
cat("MSE:", mse, "\n")
```

```
MSE: 141047372856
```

```r
cat("RMSE:", rmse, "\n")
```

```
RMSE: 375562.7
```

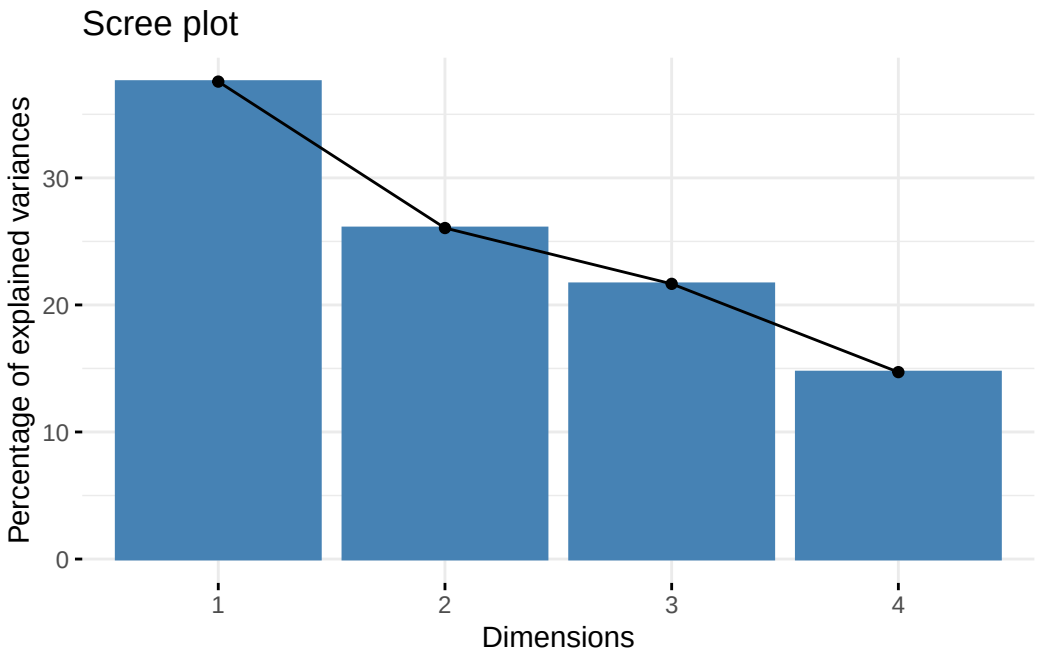
```r
plot(pca_results, type = "lines")
```



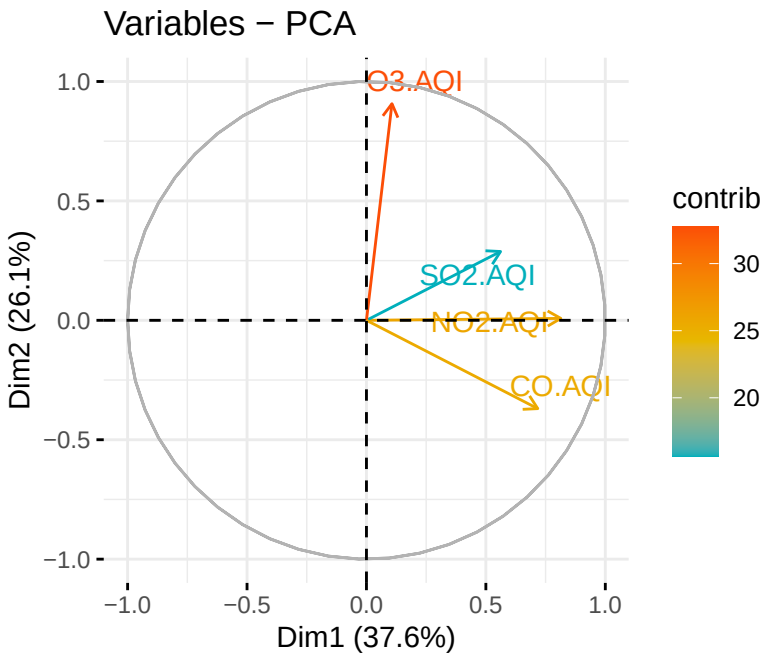

```
summary(pca_results)
```

```
Importance of components:
                          PC1    PC2    PC3    PC4
Standard deviation     1.2260 1.0208 0.9308 0.7671
Proportion of Variance 0.3758 0.2605 0.2166 0.1471
Cumulative Proportion  0.3758 0.6363 0.8529 1.0000
```

```
fviz_eig(pca_results)
```




```
fviz_pca_var(pca_results,
             col.var = "contrib", # Color by contributions to the PC
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE     # Avoid text overlapping
)
```


Variables − PCA
O3.AQI
SO2.AQI
NO2.AQI
CO.AQI
contrib
30
25
20
1.0
0.5
0.0
−0.5
−1.0
Dim2 (26.1%)
Dim1 (37.6%)