Multiple Linear Regression Development

James Davin, Rui Gong

2024-03-13

Necessary library setup

Reading and summarizing the data we'll be working with

```
# read updated dataset
bmw_data <- read.csv("./BMW-pricing.csv", header=TRUE, as.is=TRUE)

# create summary
summary(bmw_data)
```

```
##   maker_key          model_key            mileage         engine_power
##  Length:4843        Length:4843        Min.   :    -64   Min.   :  0
##  Class :character   Class :character   1st Qu.: 102914   1st Qu.:100
##  Mode  :character   Mode  :character   Median : 141080   Median :120
##                                        Mean   : 140963   Mean   :129
##                                        3rd Qu.: 175196   3rd Qu.:135
##                                        Max.   :1000376   Max.   :423
##  registration_date      fuel           paint_color          car_type
##  Length:4843        Length:4843        Length:4843        Length:4843
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  feature_1       feature_2       feature_3       feature_4
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:2181      FALSE:1004      FALSE:3865      FALSE:3881
##  TRUE :2662      TRUE :3839      TRUE :978       TRUE :962
##
##
##
##  feature_5       feature_6       feature_7       feature_8
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:2613      FALSE:3674      FALSE:329       FALSE:2223
##  TRUE :2230      TRUE :1169      TRUE :4514      TRUE :2620
##
##
##
##      price          sold_at            obs_type
##  Min.   :   100   Length:4843        Length:4843
##  1st Qu.: 10800   Class :character   Class :character
##  Median : 14200   Mode  :character   Mode  :character
```

```
##  Mean   : 15828
##  3rd Qu.: 18600
##  Max.   :178500
```

Data cleaning

```
# checking missing values (we observe 0)
sum(is.na(bmw_data))
```

```
## [1] 0
```

Because we are not splitting our data into training/testing/validation nor engaging in machine learning, we remove the irrelevant "obs_type" (*observation type*) column.

```
bmw_data$obs_type <- NULL
```

We find inappropriate values for mileage and engine power.

```
# we check for inappropriate values
# min mileage should be 0; we get the index of those which are below 0
print(which(bmw_data$mileage < 0))
```

```
## [1] 2939
```

```
# min engine power is 0, but it should be > 0; we get the index of those which are below 0
print(which(bmw_data$engine_power <= 0))
```

```
## [1] 3766
```

```
# print the number of observations
print(length(bmw_data$mileage))
```

```
## [1] 4843
```

Owing to their relatively small total size of the data set and the extensive amount of time required to correctly impute the values for mileage and engine power, we opt to remove the entries with inappropriate observations. The number of entries reduces by 2.

```
# row 2939 has negative mileage values -- we opt to delete it
bmw_data <- bmw_data[-which(bmw_data$mileage < 0),]
# row 3765 has 0 engine power values -- we opt to delete it
bmw_data <- bmw_data[-which(bmw_data$engine_power == 0),]
# we should now have 4843 - 2 = 4831 observations
print(length(bmw_data$mileage))
```

```
## [1] 4841
```

The registration date by itself is not easily used in computations, so we use to create new features "age", "month_sold", "month_registered", and "year_registered". Note that all vehicles were sold during the same year (2018).

```
# create a new variable age and attach it to the same dataframe
# split the registration date and sold date vectors first, in order to calculate age
sold_at_split <- strsplit(bmw_data$sold_at, "/")
registration_split <- strsplit(bmw_data$registration_date, "/")

# create field specifying month sold
bmw_data$month_sold <- sapply(sold_at_split, function(x) as.integer(x[1]))
# create field specifying month registered
```

```r
bmw_data$month_registered <- sapply(registration_split, function(x) as.integer(x[1]))
# create field specifying year registered
bmw_data$year_registered <- sapply(registration_split, function(x) as.integer(x[3]))

# create a field specifying age of each car
bmw_data$age <- 2018 - bmw_data$year_registered + (1/12)*(bmw_data$month_sold - bmw_data$month_register
```

```r
summary(bmw_data)
```

```
##   maker_key          model_key            mileage         engine_power
##  Length:4841        Length:4841        Min.   :    476   Min.   : 25
##  Class :character   Class :character   1st Qu.: 103034   1st Qu.:100
##  Mode  :character   Mode  :character   Median : 141089   Median :120
##                                        Mean   : 141004   Mean   :129
##                                        3rd Qu.: 175217   3rd Qu.:135
##                                        Max.   :1000376   Max.   :423
##  registration_date      fuel           paint_color          car_type
##  Length:4841        Length:4841        Length:4841        Length:4841
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  feature_1       feature_2       feature_3       feature_4
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:2180      FALSE:1003      FALSE:3863      FALSE:3880
##  TRUE :2661      TRUE :3838      TRUE :978       TRUE :961
##
##
##
##  feature_5       feature_6       feature_7       feature_8
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:2611      FALSE:3672      FALSE:328       FALSE:2222
##  TRUE :2230      TRUE :1169      TRUE :4513      TRUE :2619
##
##
##
##      price          sold_at            month_sold    month_registered
##  Min.   :   100   Length:4841        Min.   :1.000   Min.   : 1.000
##  1st Qu.: 10800   Class :character   1st Qu.:3.000   1st Qu.: 3.000
##  Median : 14200   Mode  :character   Median :5.000   Median : 6.000
##  Mean   : 15817                      Mean   :4.927   Mean   : 6.131
##  3rd Qu.: 18600                      3rd Qu.:7.000   3rd Qu.: 9.000
##  Max.   :178500                      Max.   :9.000   Max.   :12.000
##  year_registered      age
##  Min.   :1990    Min.   : 0.5833
##  1st Qu.:2012    1st Qu.: 4.0833
##  Median :2013    Median : 4.8333
##  Mean   :2012    Mean   : 5.4335
##  3rd Qu.:2014    3rd Qu.: 5.8333
##  Max.   :2017    Max.   :28.0833
```

Data exploration (and continued cleaning)

Exploration of fuel partitioning

First we examine how fuel type influences key variables.

```
# partition the data by fuel type
bmw_fuel <- split(bmw_data, bmw_data$fuel)
bmw.diesel <- bmw_fuel$diesel
bmw.petrol <- bmw_fuel$petrol
bmw.hybrid <- bmw_fuel$hybrid_petrol
bmw.electric <- bmw_fuel$electro
```

We see that diesel is by far the most common type of fuel used by the sample.

```
print(nrow(bmw.diesel))
```

```
## [1] 4639
```

```
print(nrow(bmw.petrol))
```

```
## [1] 191
```

```
print(nrow(bmw.electric))
```

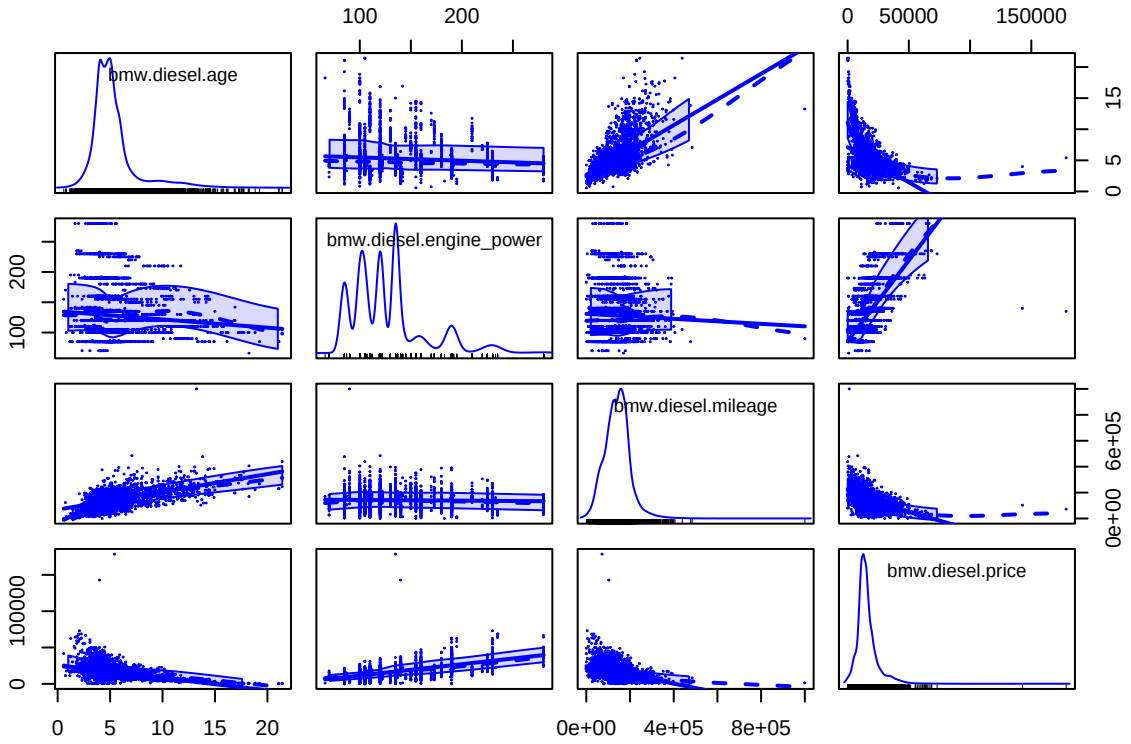
```
## [1] 3
```

```
print(nrow(bmw.hybrid))
```

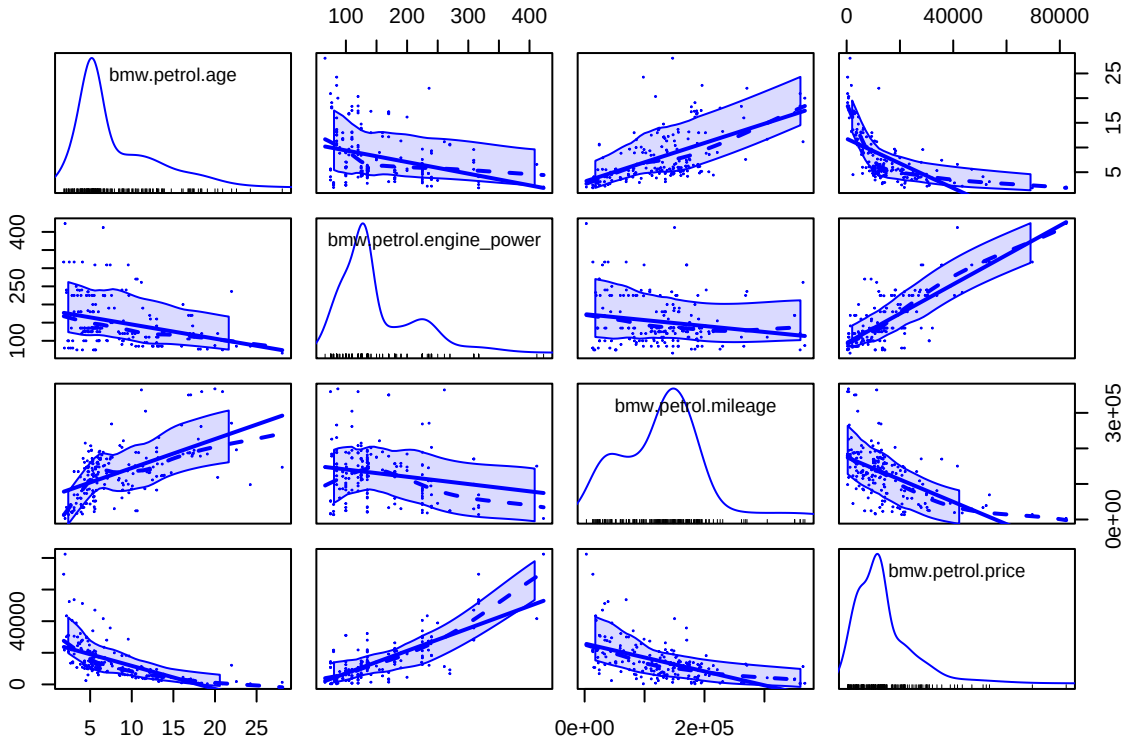
```
## [1] 8
```

Electric and hybrid BMWs make up $11/4,639 \approxeq 0.23\%$ of the dataset. Because of that, they lack sufficient information to develop scatterplots for, and we preclude them from our scatterplot analysis (R threw errors when attempted).

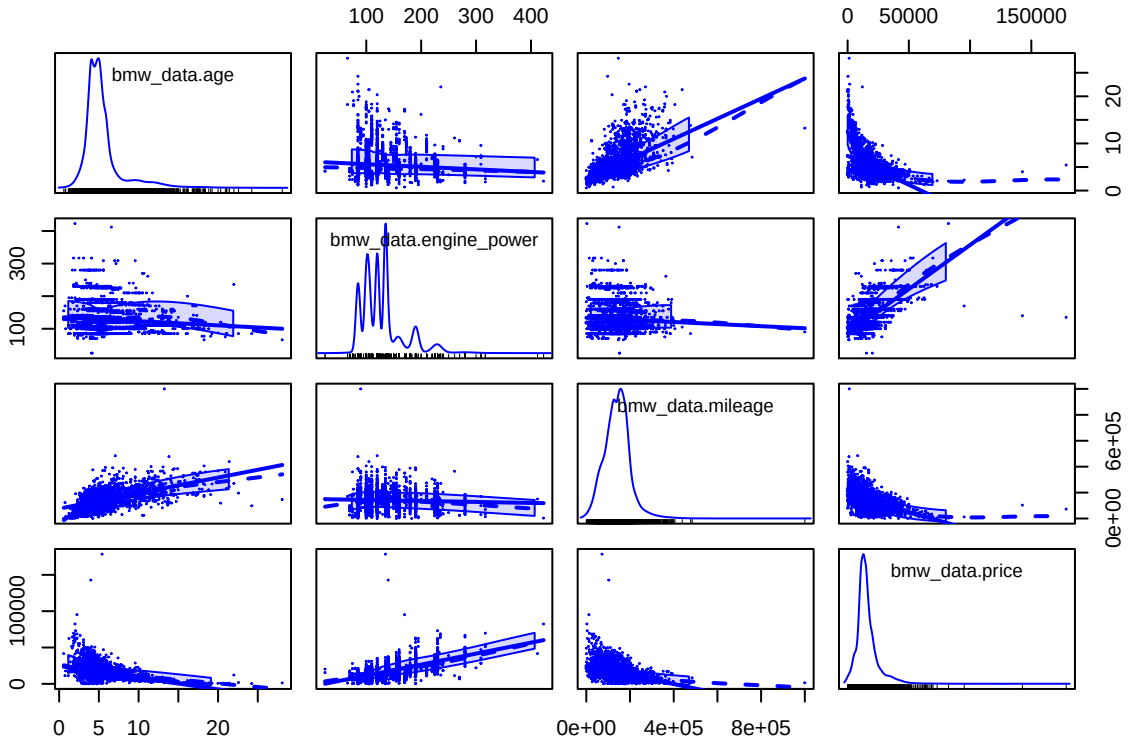
```
# create scatterplot matrices to study the distribution
scatterplotMatrix(~ bmw.diesel$age + bmw.diesel$engine_power + bmw.diesel$mileage + bmw.diesel$price,
                  pch=19, cex=0.1)
```




```
scatterplotMatrix(~ bmw.petrol$age + bmw.petrol$engine_power + bmw.petrol$mileage + bmw.petrol$price,
                  pch=19, cex=0.1)
```




```
# comparison group
scatterplotMatrix(~ bmw_data$age + bmw_data$engine_power + bmw_data$mileage + bmw_data$price,
                  pch=19, cex=0.1)
```



Exploration of age partitioning

```
bmw_data$age_class <- cut(bmw_data$age,
                      breaks = c(0, 2, 4, 8, 16, Inf),
                      labels = c("youngest", "young", "average", "old", "oldest"))

bmw_age<- split(bmw_data, bmw_data$age_class)
bmw.youngest <- bmw_age$youngest
bmw.young <- bmw_age$young
bmw.average <- bmw_age$average
bmw.old <- bmw_age$old
bmw.oldest <- bmw_age$oldest
bmw.ancient <- bmw_age$ancient
```

We see that we have quite a low number of "youngest" (0-2 years) and "oldest" ($>$ 16 years) cars ($\sim$ 2% of the sample) and cars of "average" age (4-8 years) are by far the common age ($\sim$ 66% of the sample).

```
print(nrow(bmw.youngest))
```

```
## [1] 53
```

```
print(nrow(bmw.young))
```

```
## [1] 1148
```

```
print(nrow(bmw.average))
```

```
## [1] 3144
```

```
print(nrow(bmw.old))
```

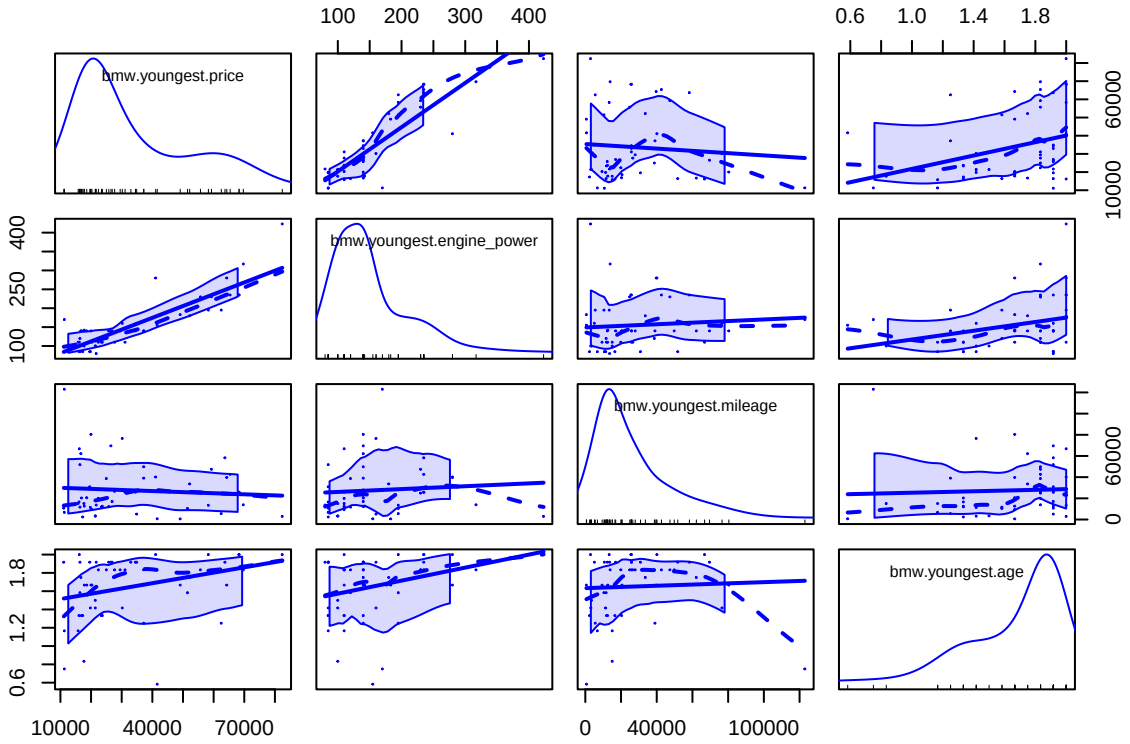
```
## [1] 449
```

```
print(nrow(bmw.oldest))
```

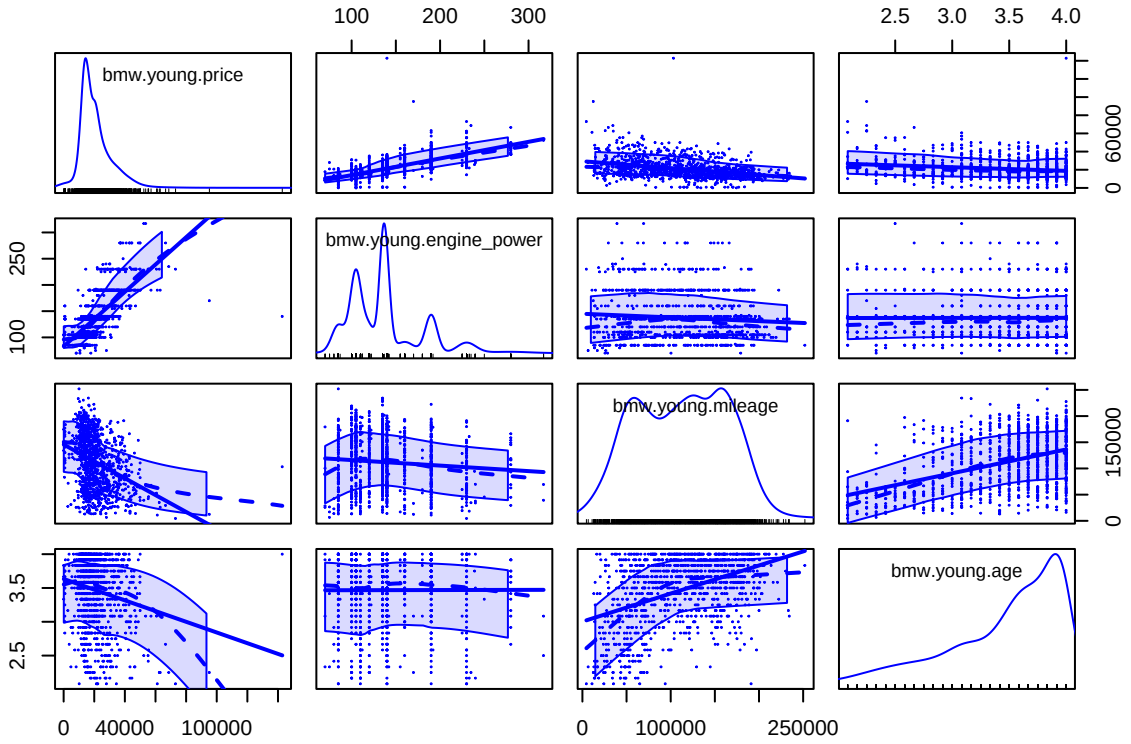
```
## [1] 47
```

Scatterplots by age partition

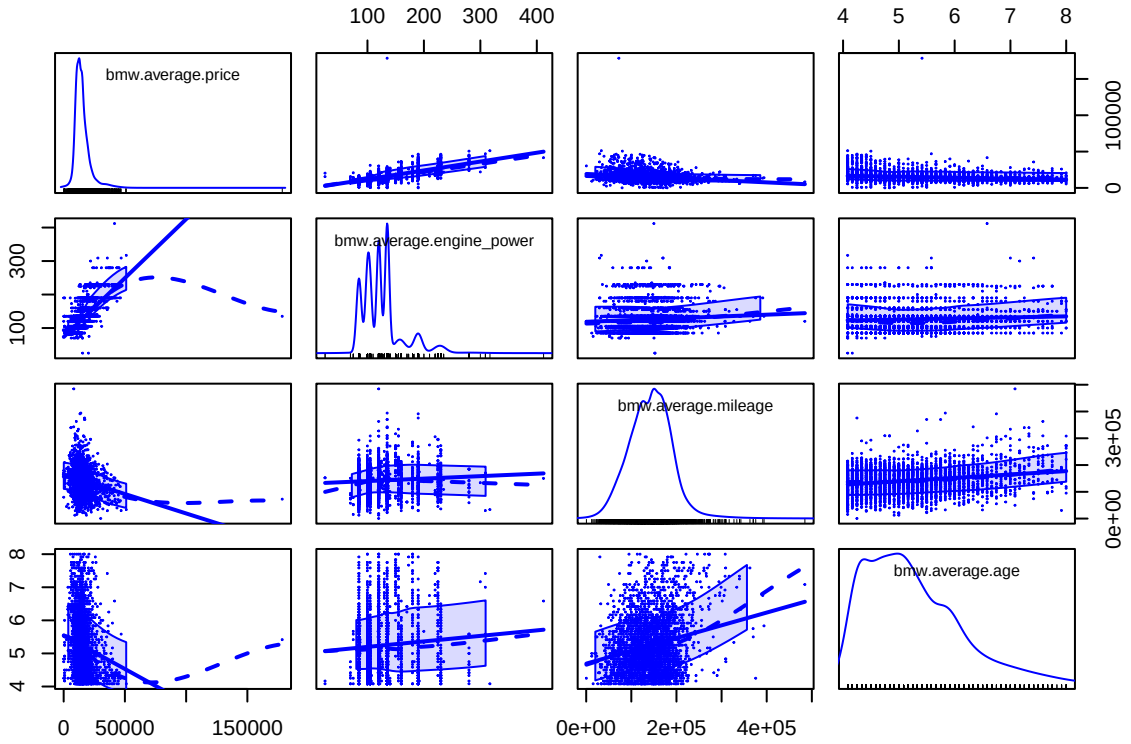
```
scatterplotMatrix(~ bmw.youngest$price + bmw.youngest$engine_power + bmw.youngest$mileage + bmw.youngest
                  pch=19, cex=0.1)
```



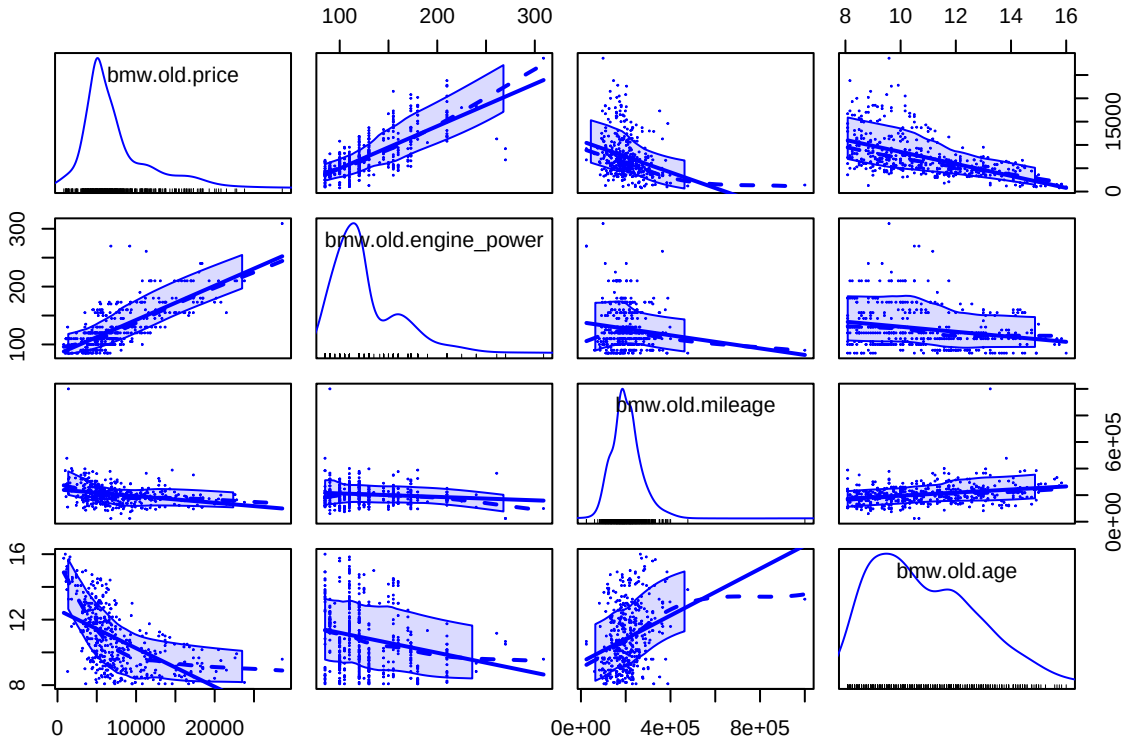

```
scatterplotMatrix(~ bmw.young$price + bmw.young$engine_power + bmw.young$mileage + bmw.young$age ,
                  pch=19, cex=0.1)
```



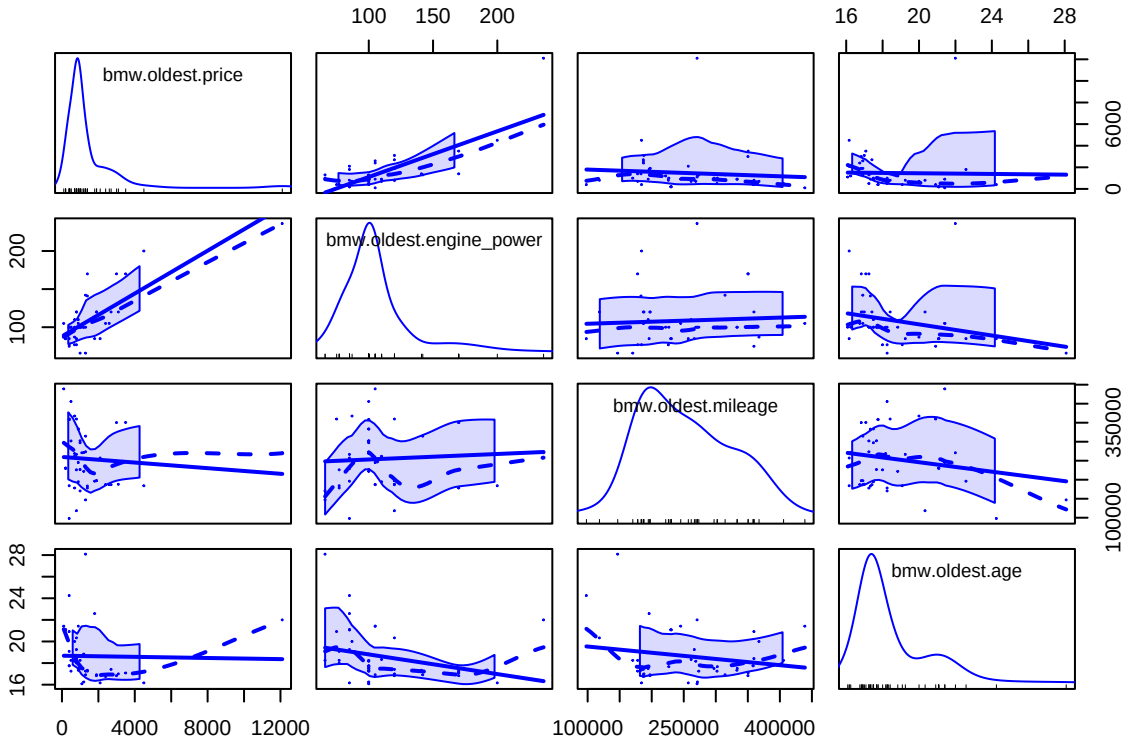

```
scatterplotMatrix(~ bmw.average$price + bmw.average$engine_power + bmw.average$mileage + bmw.average$ag
                  pch=19, cex=0.1)
```




```
scatterplotMatrix(~ bmw.old$price + bmw.old$engine_power + bmw.old$mileage + bmw.old$age,
                  pch=19, cex=0.1)
```



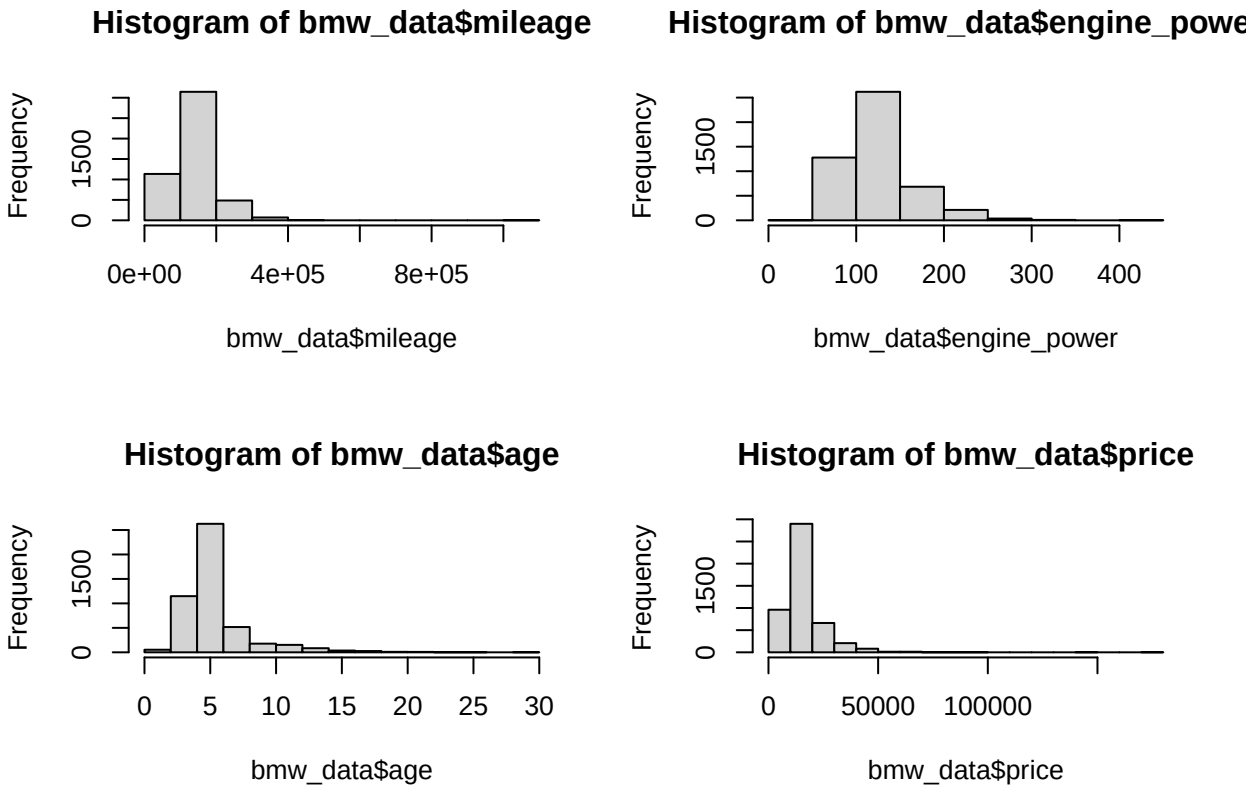

```
scatterplotMatrix(~ bmw.oldest$price + bmw.oldest$engine_power + bmw.oldest$mileage + bmw.oldest$age,
                  pch=19, cex=0.1)
```



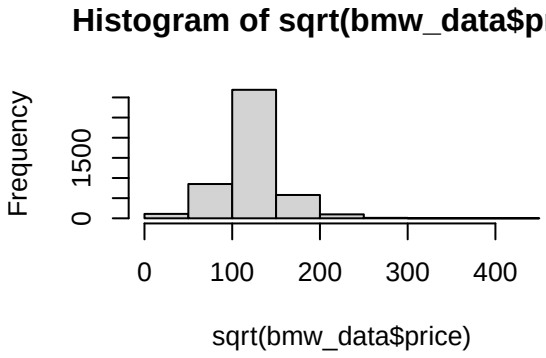
Interestingly, we see that engine power seems the most consistently linear predictor of price across the various age brackets. The youngest and oldest brackets are hardest to see correlations (linear or not) in between price and the other variables. We also see that mileage has a more pronounced negative correlation with price across all age brackets.

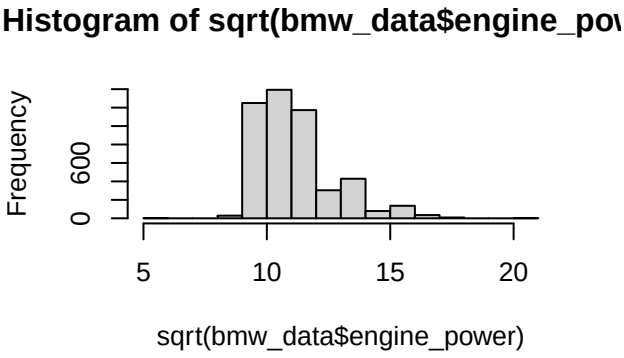
Observing the effects transformations on distributions

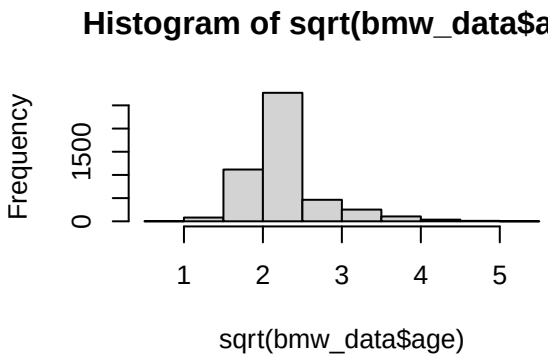
```
par(mfrow = c(2, 2))
hist(bmw_data$mileage) # right skewed
hist(bmw_data$engine_power) # right skewed
hist(bmw_data$age) # right skewed
hist(bmw_data$price) # right skewed
```

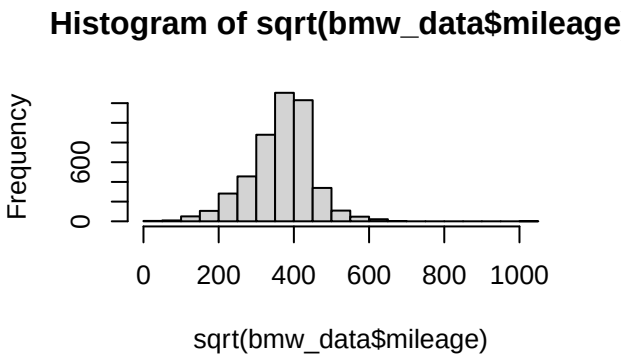



```r
par(mfrow = c(2, 2))
hist(sqrt(bmw_data$price)) # still right skewed
hist(sqrt(bmw_data$engine_power)) # slightly right skewed
hist(sqrt(bmw_data$age)) # slightly right skewed
hist(sqrt(bmw_data$mileage))# approximately normal (BEST FIT)
```

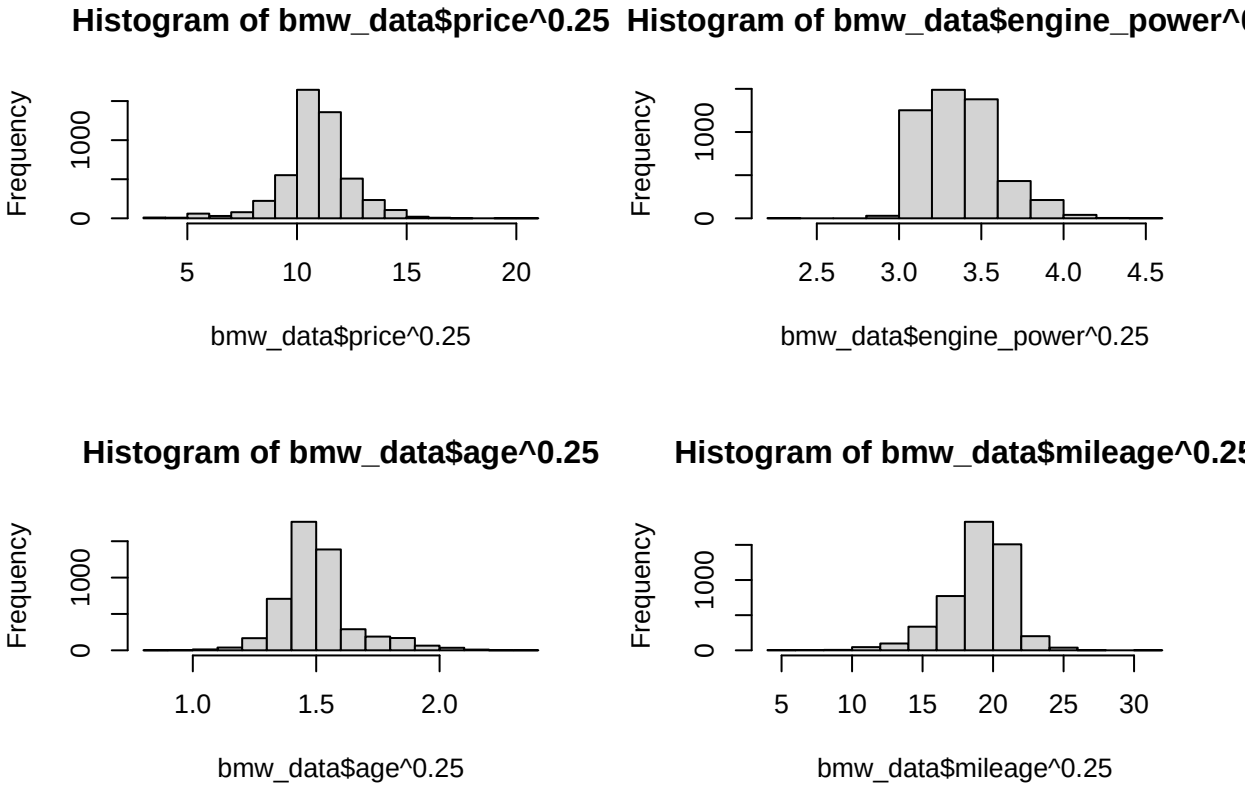




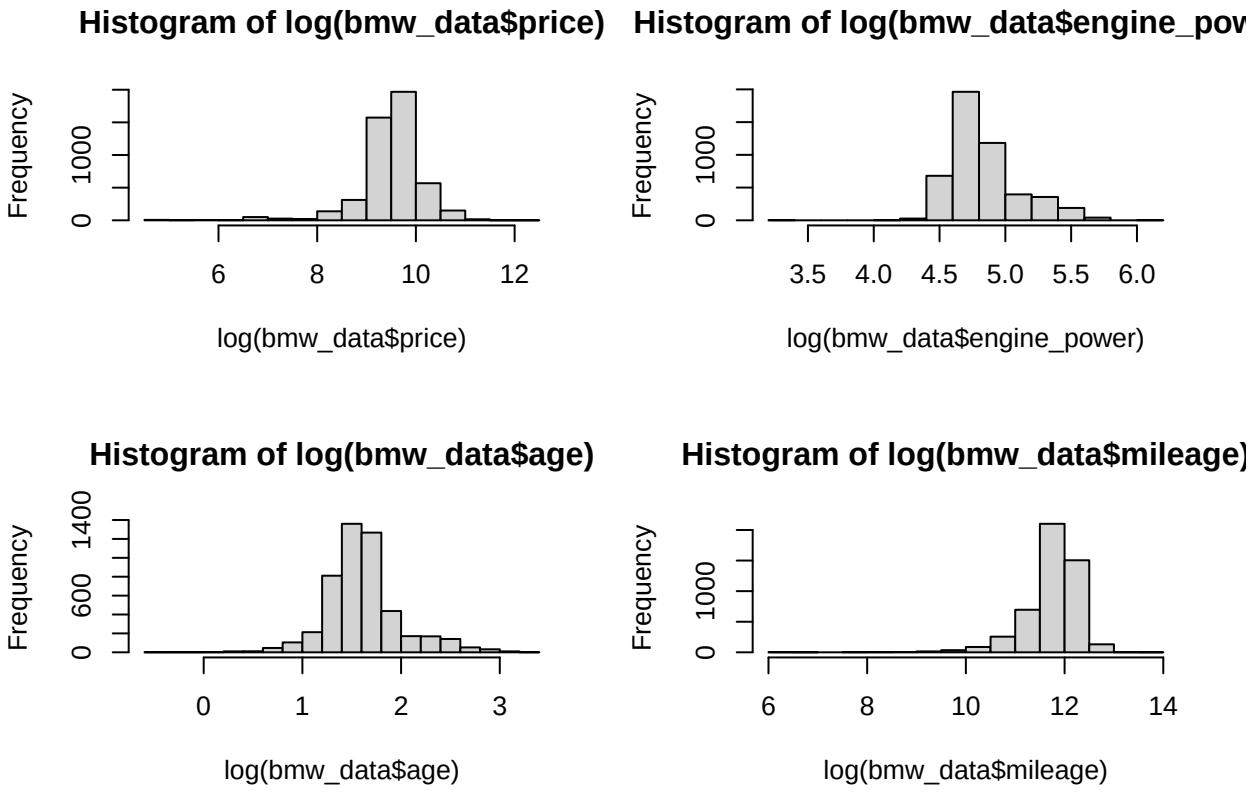





```r
par(mfrow = c(2, 2))
hist(bmw_data$price^0.25) # approximately normal (BEST FIT)
hist(bmw_data$engine_power^0.25) # slightly right skewed
hist(bmw_data$age^0.25) # approximately normal
hist(bmw_data$mileage^0.25) # slightly left skewed
```



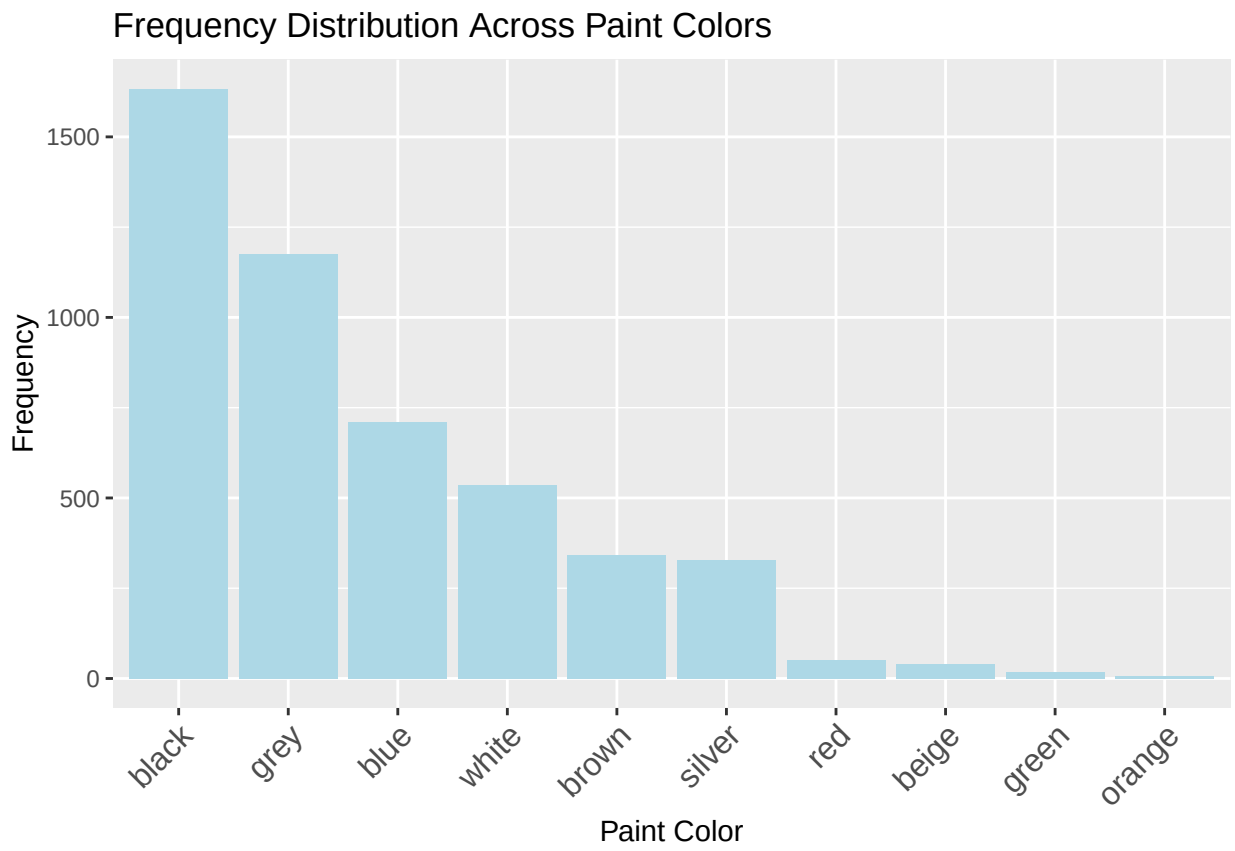

```r
par(mfrow = c(2, 2))
hist(log(bmw_data$price)) # slightly left skewed
hist(log(bmw_data$engine_power)) # slightly right skewed (BEST FIT)
hist(log(bmw_data$age)) # approximately normal (BEST FIT)
hist(log(bmw_data$mileage)) # left skewed
```



Frequency graphs

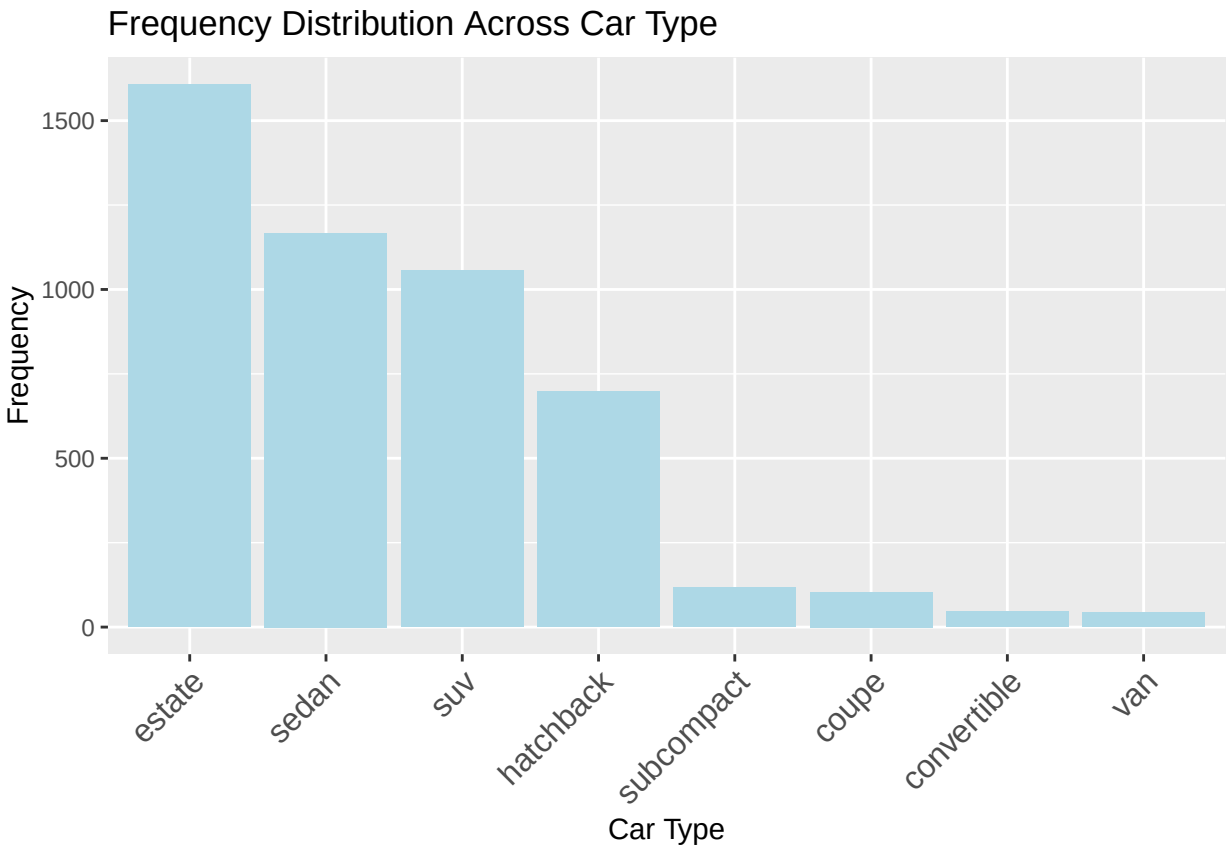
Paint color

```
# count the frequency of each paint color and arrange in descending order
paints_df <- bmw_data %>%
  count(paint_color)

# create a bar plot showing the frequency of each paint color
# use reorder for x to match the order by n
ggplot(paints_df, aes(x=reorder(paint_color, n, decreasing = TRUE), y=n)) +
  geom_bar(stat = "identity", fill="lightblue") + # adding bars
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size=12)) + # adjusting x labels
  xlab("Paint Color") +
  ylab("Frequency") +
  ggtitle("Frequency Distribution Across Paint Colors")
```



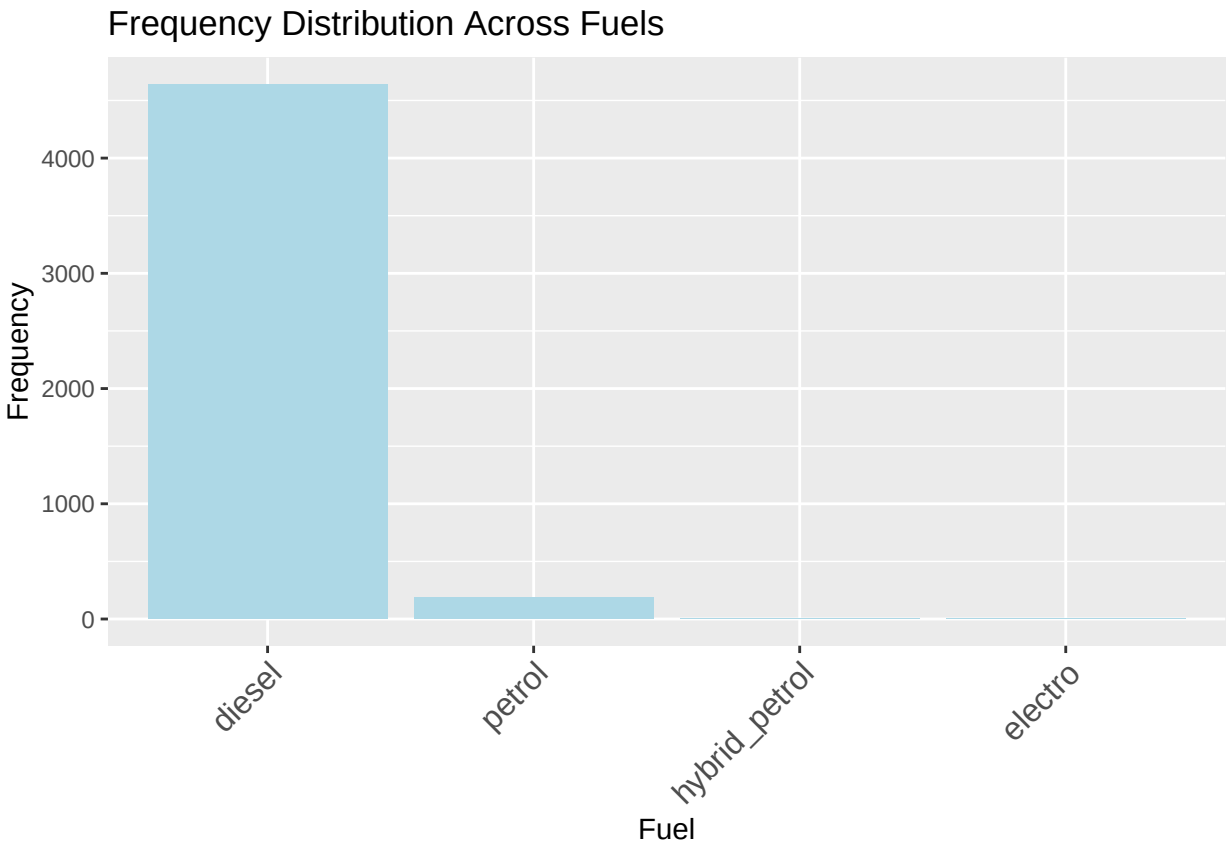
Car type

```
car_types_df <- bmw_data %>%
  count(car_type)

ggplot(car_types_df, aes(x=reorder(car_type, n, decreasing = TRUE), y=n)) +
  geom_bar(stat = "identity", fill="lightblue") + # adding bars
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size=12)) + # adjusting x labels
  xlab("Car Type") +
  ylab("Frequency") +
  ggtitle("Frequency Distribution Across Car Type")
```



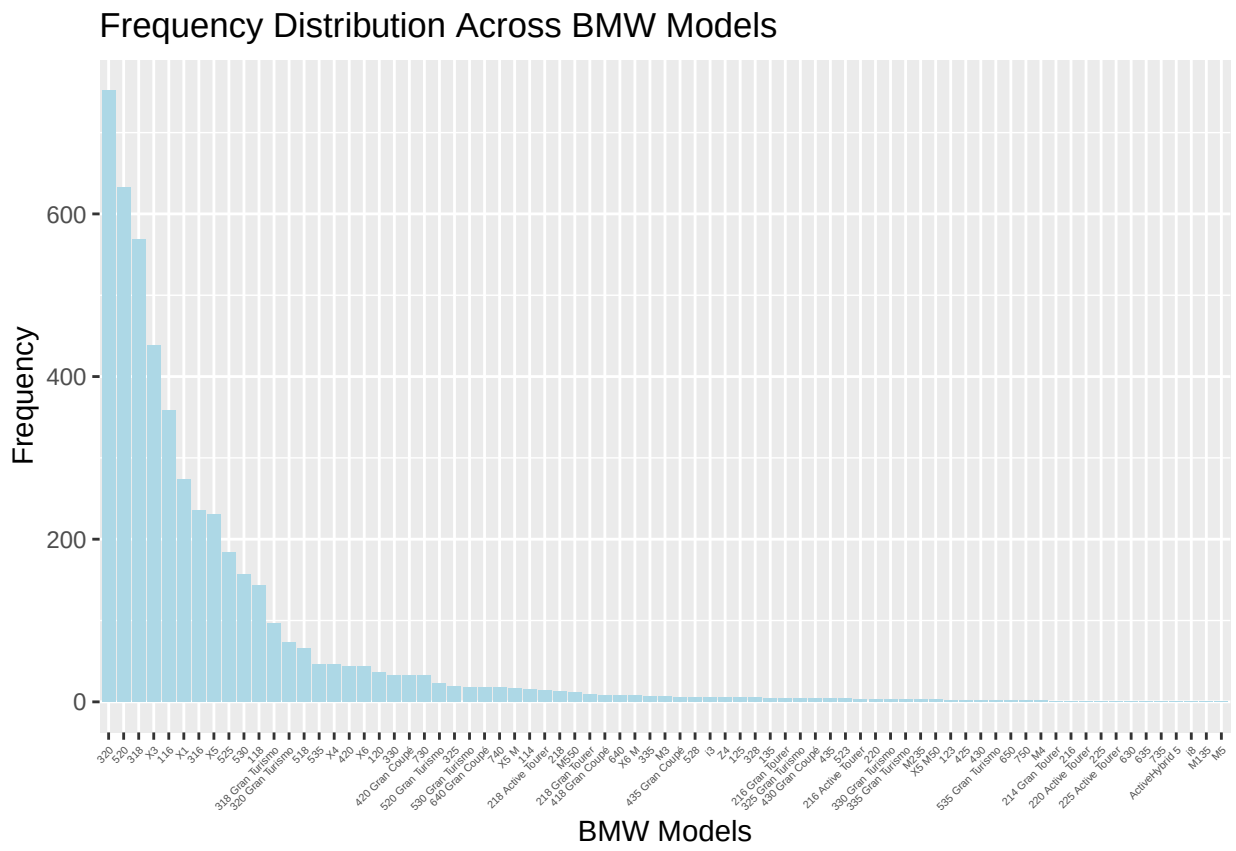
Fuel

```
fuels_df <- bmw_data %>%
  count(fuel)

ggplot(fuels_df, aes(x=reorder(fuel, n, decreasing = TRUE), y=n)) +
  geom_bar(stat = "identity", fill="lightblue") + # adding bars
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size=12)) + # adjusting x labels
  xlab("Fuel") +
  ylab("Frequency") +
  ggtitle("Frequency Distribution Across Fuels")
```



Model

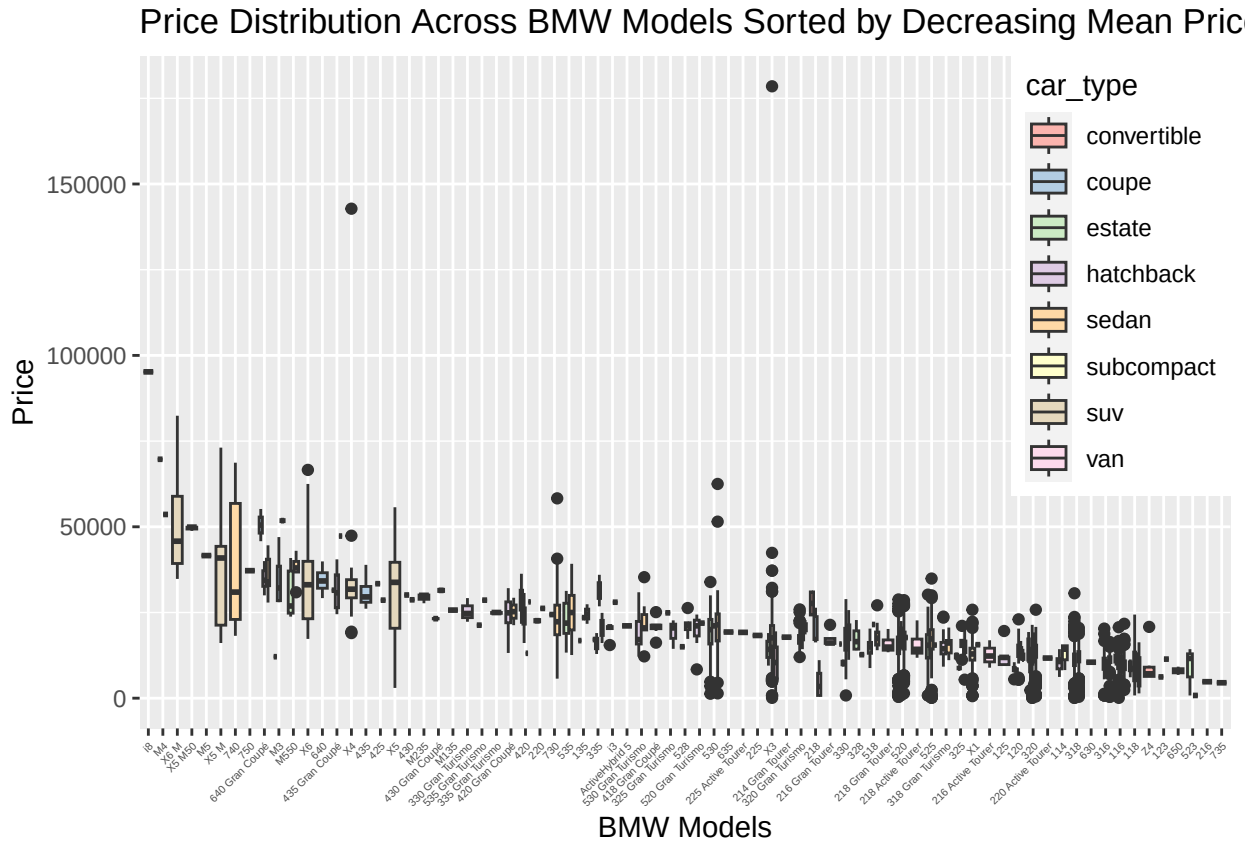
```
# Creating a new data frame with models sorted by their frequency
models_df <- bmw_data %>%
  count(model_key)
# This dataframe 'models_df' now contains models sorted by frequency

ggplot(models_df, aes(x=reorder(model_key, n, decreasing=TRUE), y=n)) +
  geom_bar(stat = "identity", fill="lightblue") + # adding bars
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size=4)) + # adjusting x labels
  xlab("BMW Models") +
  ylab("Frequency") +
  ggtitle("Frequency Distribution Across BMW Models")
```



Price distributions by categorical types

Model

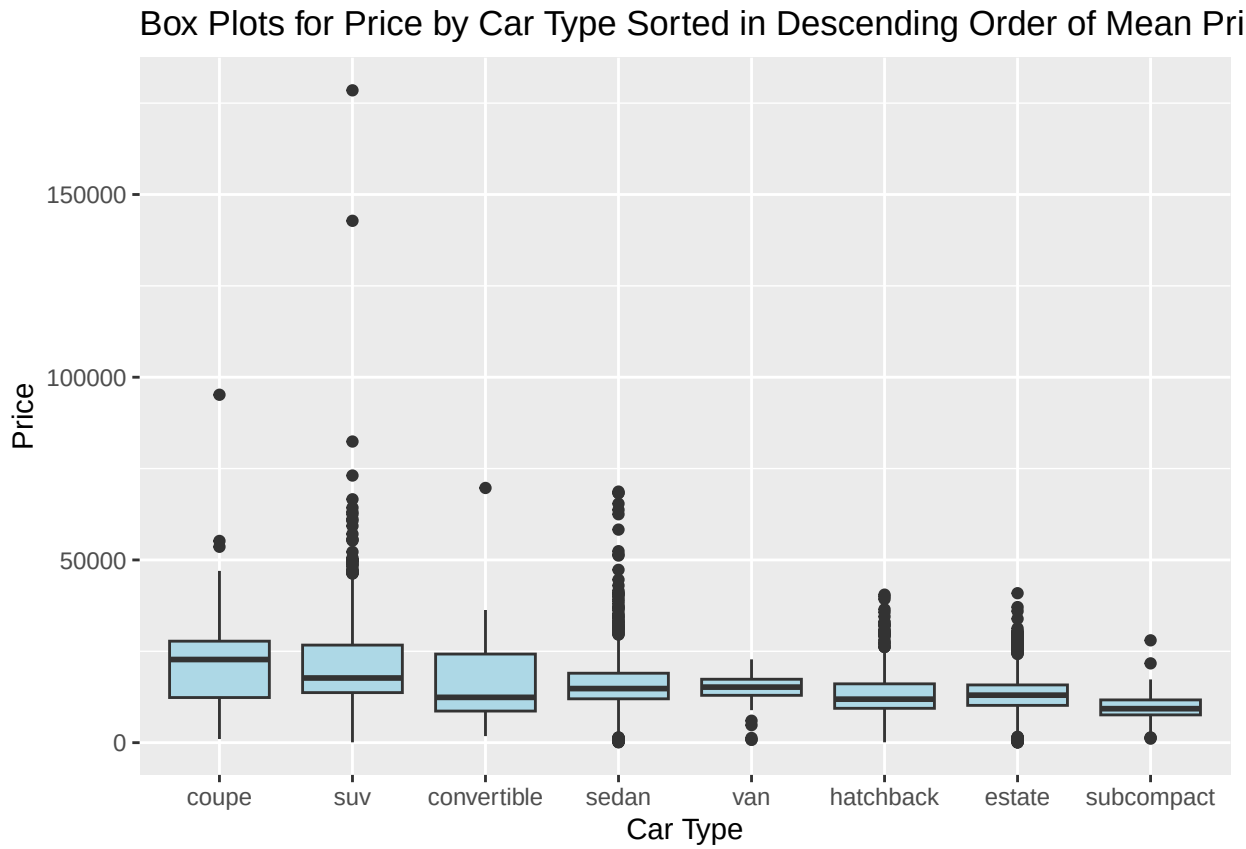
```
# Calculate mean price for each model
mean_by_model <- bmw_data %>%
  group_by(model_key) %>%
  summarize(mean_price = mean(price)) %>%
  arrange(desc(mean_price))

# Use the ordered model_key for plotting
bmw_data$model_key <- factor(bmw_data$model_key, levels = mean_by_model$model_key)

# Now plot with the models ordered by decreasing mean price
ggplot(bmw_data, aes(x=model_key, y=price, fill=car_type)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size=4),
        legend.position = c(1, 1), # Move the legend to the top-right corner
        legend.justification = c(1, 1)) + # Rotating x labels for legibility
  xlab("BMW Models") +
  ylab("Price") +
  ggtitle("Price Distribution Across BMW Models Sorted by Decreasing Mean Price") +
  scale_fill_brewer(palette="Pastel1")
```



Car type (e.g., sedan, hatchback)

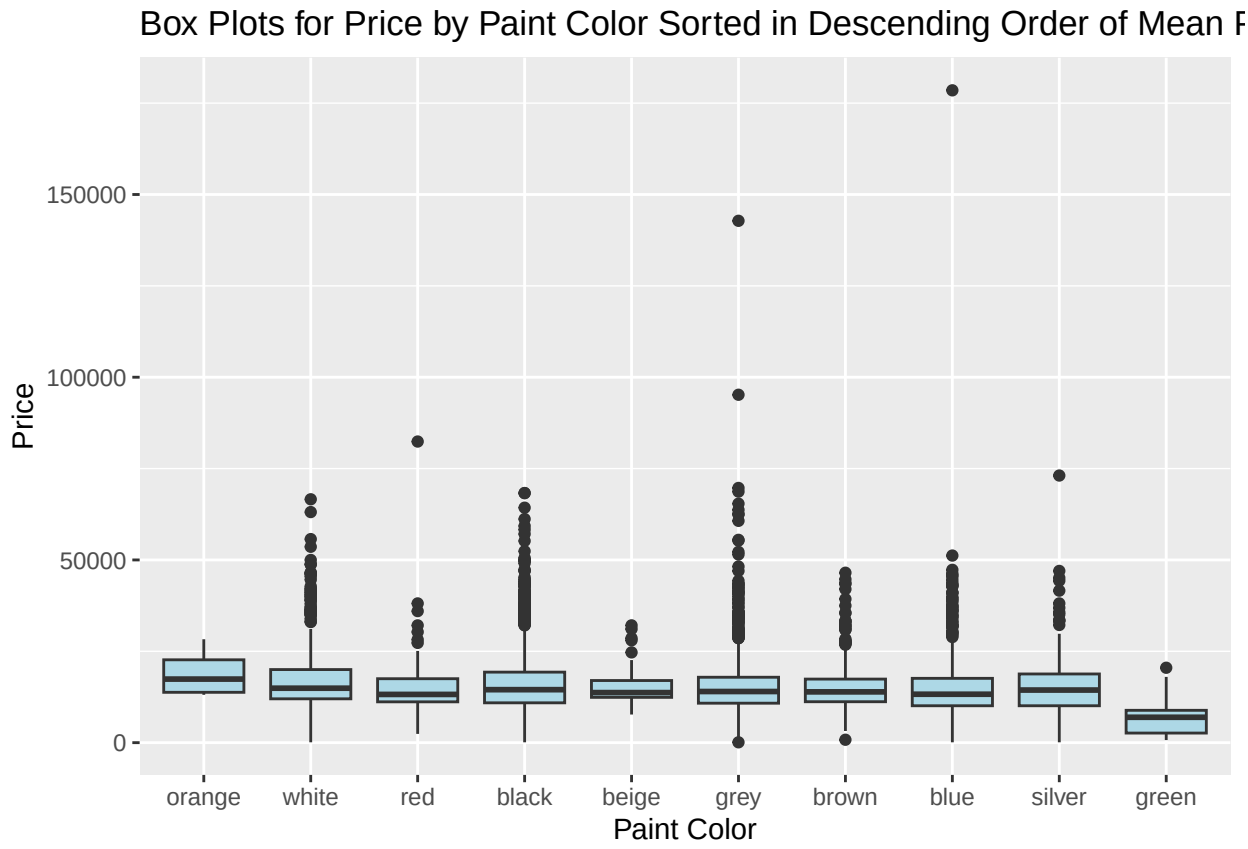
We note here that the mean price of the coupe is the most expensive, the SUV is second most expensive, and the subcompact is the cheapest.

```
mean_by_type <- bmw_data %>%
  group_by(car_type) %>%
  summarize(mean_price = mean(price)) %>%
  arrange(desc(mean_price))

bmw_data$car_type <- factor(bmw_data$car_type, levels=mean_by_type$car_type)
ggplot(bmw_data, aes(x=car_type, y=price)) +
    geom_boxplot(fill="lightblue") +
    xlab("Car Type") +
    ylab("Price") +
    ggtitle("Box Plots for Price by Car Type Sorted in Descending Order of Mean Price")
```



Color

```
mean_by_color <- bmw_data %>%
  group_by(paint_color) %>%
  summarize(mean_price = mean(price)) %>%
  arrange(desc(mean_price))

bmw_data$paint_color <- factor(bmw_data$paint_color, levels=mean_by_color$paint_color)
ggplot(bmw_data, aes(x=paint_color, y=price)) +
    geom_boxplot(fill="lightblue") +
    xlab("Paint Color") +
    ylab("Price") +
    ggtitle("Box Plots for Price by Paint Color Sorted in Descending Order of Mean Price")
```




```
print(mean_by_color)
```

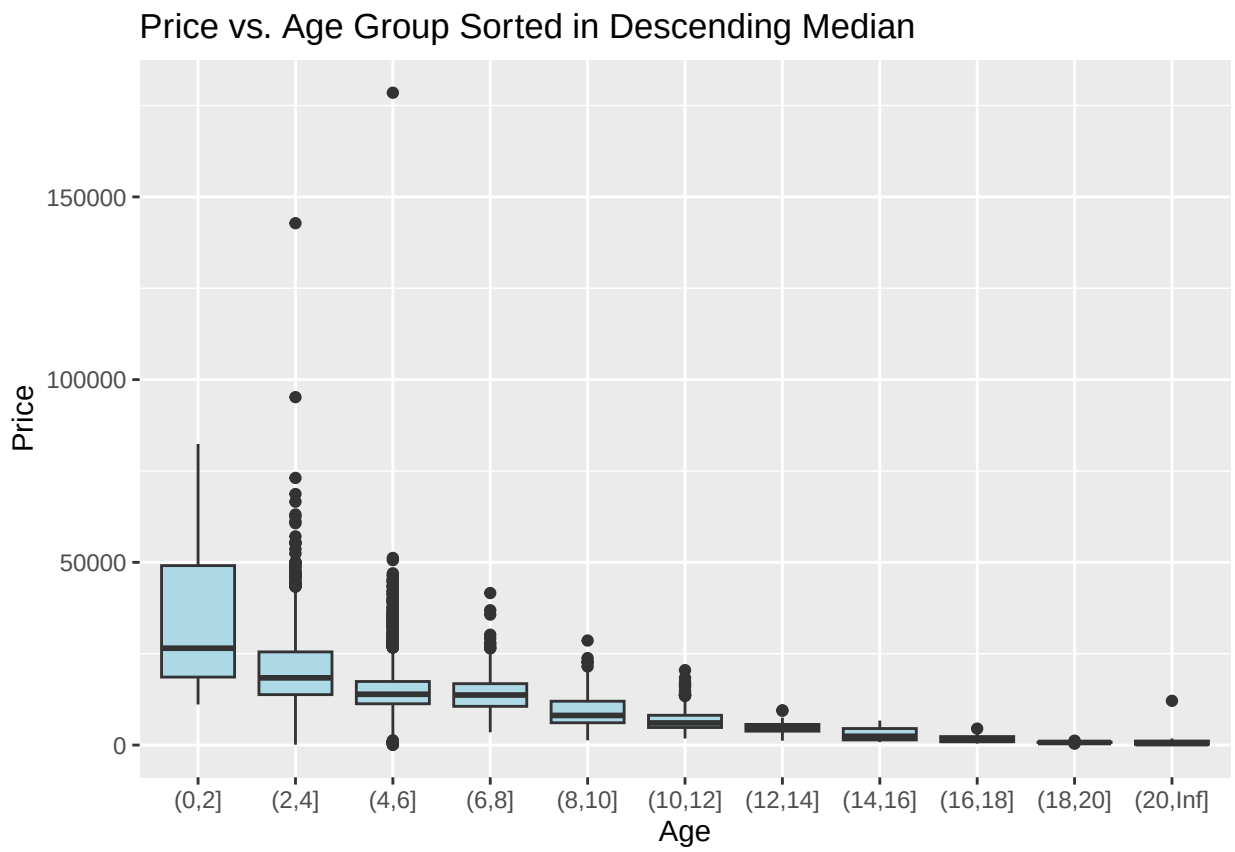
```
## # A tibble: 10 x 2
##    paint_color mean_price
##    <chr>            <dbl>
##  1 orange          18867.
##  2 white           17339.
##  3 red             16500
##  4 black           16138.
##  5 beige           15817.
##  6 grey            15596.
##  7 brown           15368.
##  8 blue            15115.
##  9 silver          14816.
## 10 green            7200
```

Interestingly, we see that of the common colors, white has the highest average price, while silver has the lowest. Coming in both with small samples, we see that orange has the highest mean and green has the lowest mean.

Age

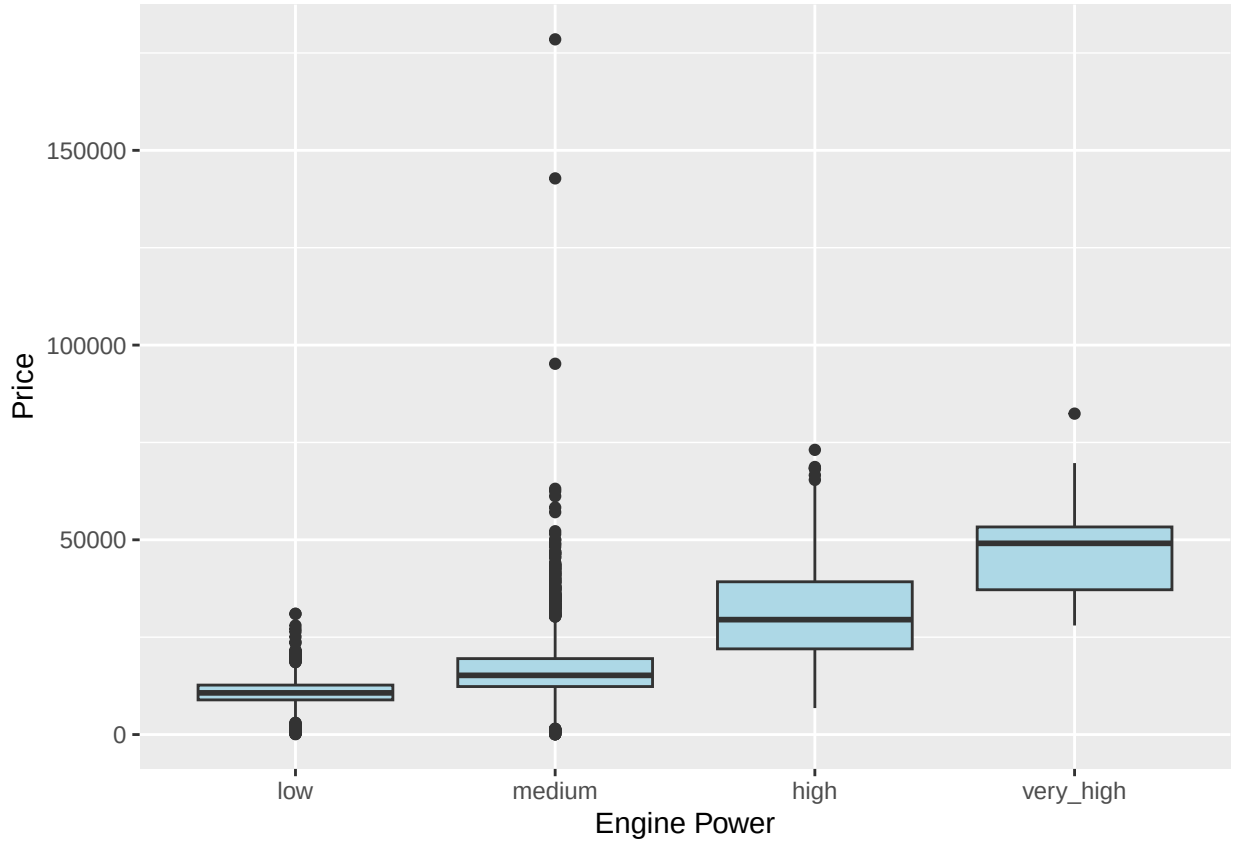
```
bmw_data$age_cat <- cut(bmw_data$age, breaks = c(0, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, Inf))
ggplot(bmw_data, aes(x=age_cat, y=price)) +
    geom_boxplot(fill="lightblue") +
    xlab("Age") +
```

```
    ylab("Price") +
    ggtitle("Price vs. Age Group Sorted in Descending Median")
```

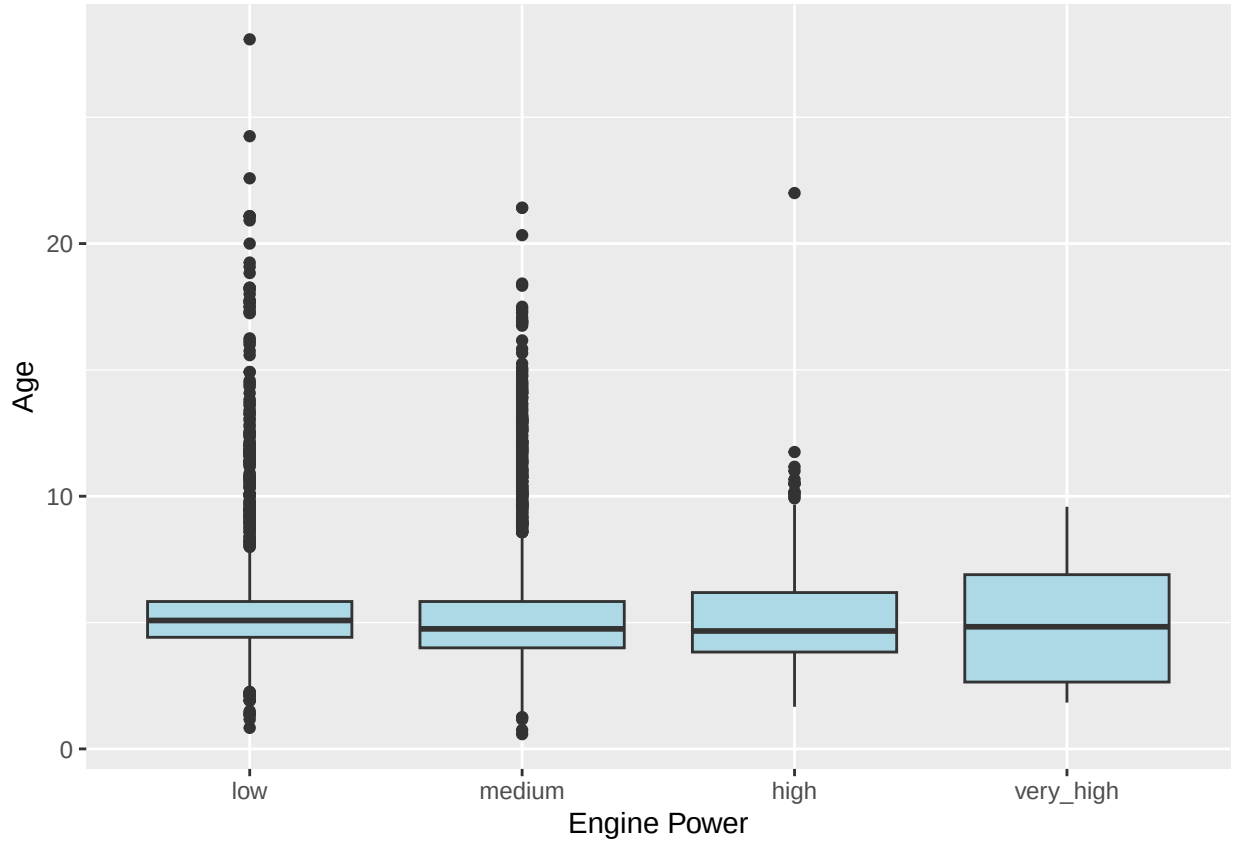


Engine power predicting price, age, and mileage

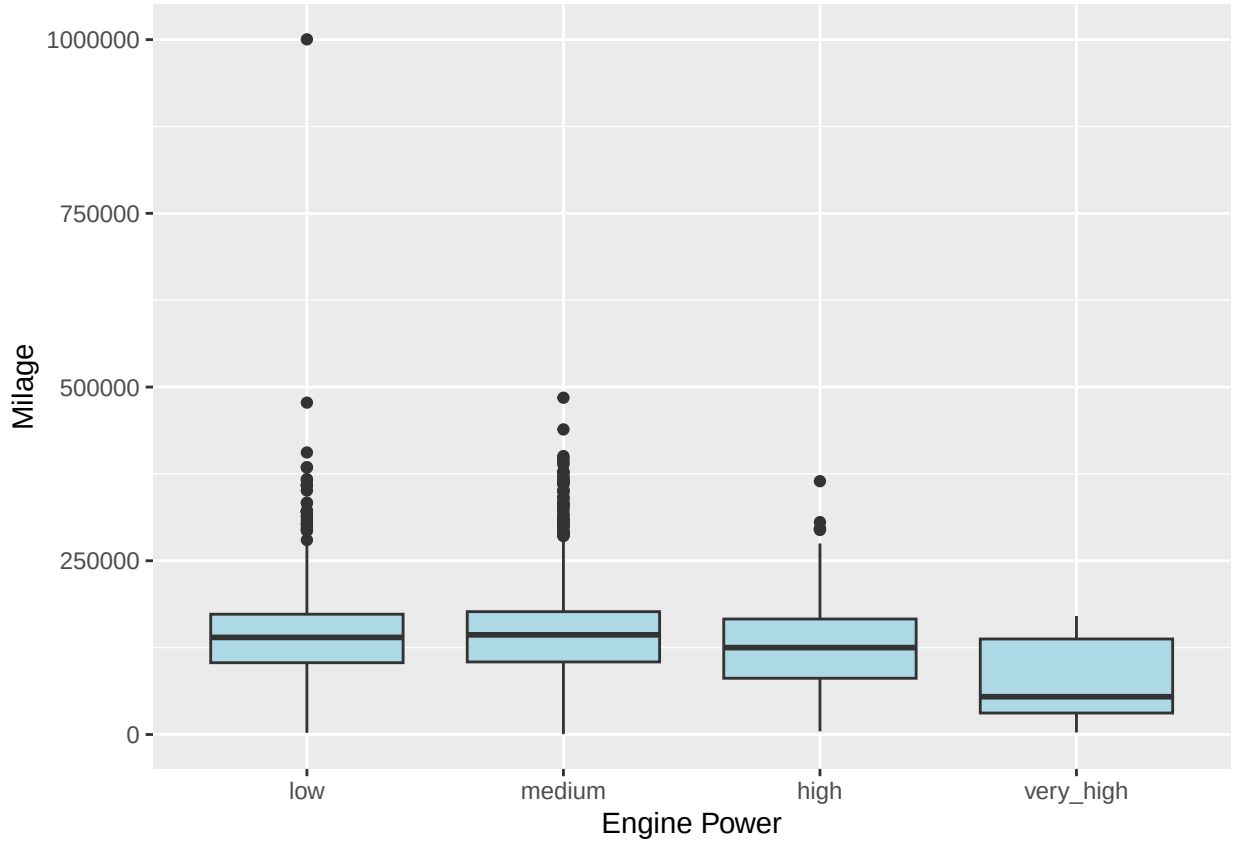
```
bmw_data$engine_cat <- cut(bmw_data$engine_power, breaks = c(0, 100, 200, 300, 600), labels = c('low',
ggplot(bmw_data, aes(x= engine_cat, y=price)) +
    geom_boxplot(fill="lightblue") +
    xlab("Engine Power") +
    ylab("Price")
```




```
ggplot(bmw_data, aes(x= engine_cat, y=age)) +
    geom_boxplot(fill="lightblue") +
    xlab("Engine Power") +
    ylab("Age")
```

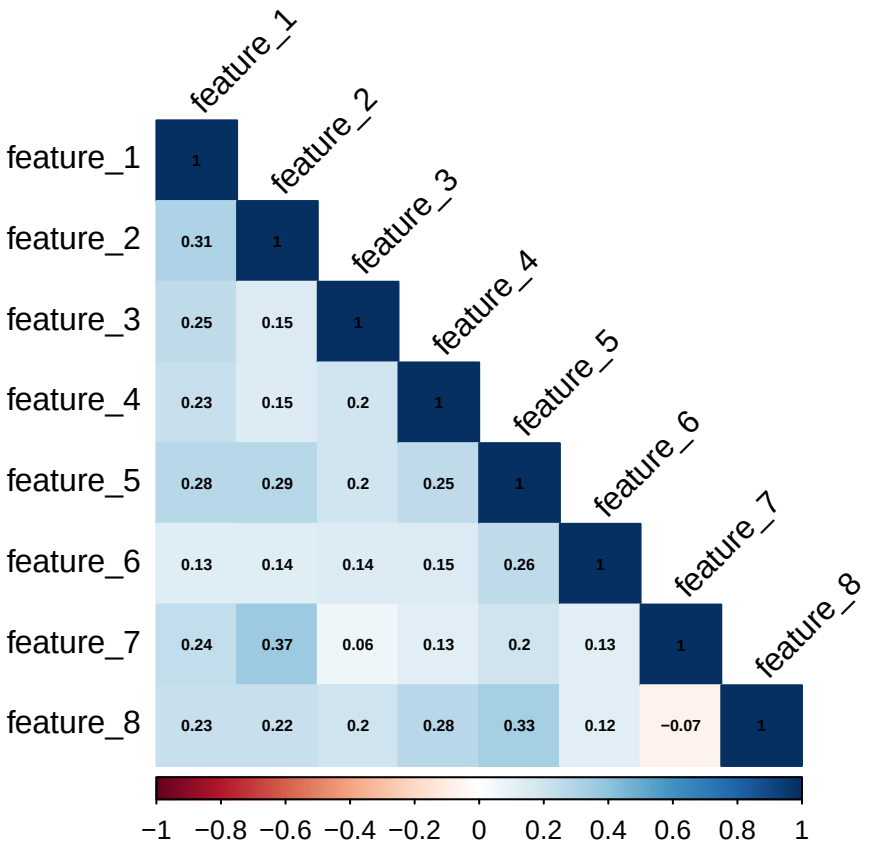



```
ggplot(bmw_data, aes(x= engine_cat, y= mileage)) +
    geom_boxplot(fill="lightblue") +
    xlab("Engine Power") +
    ylab("Milage")
```

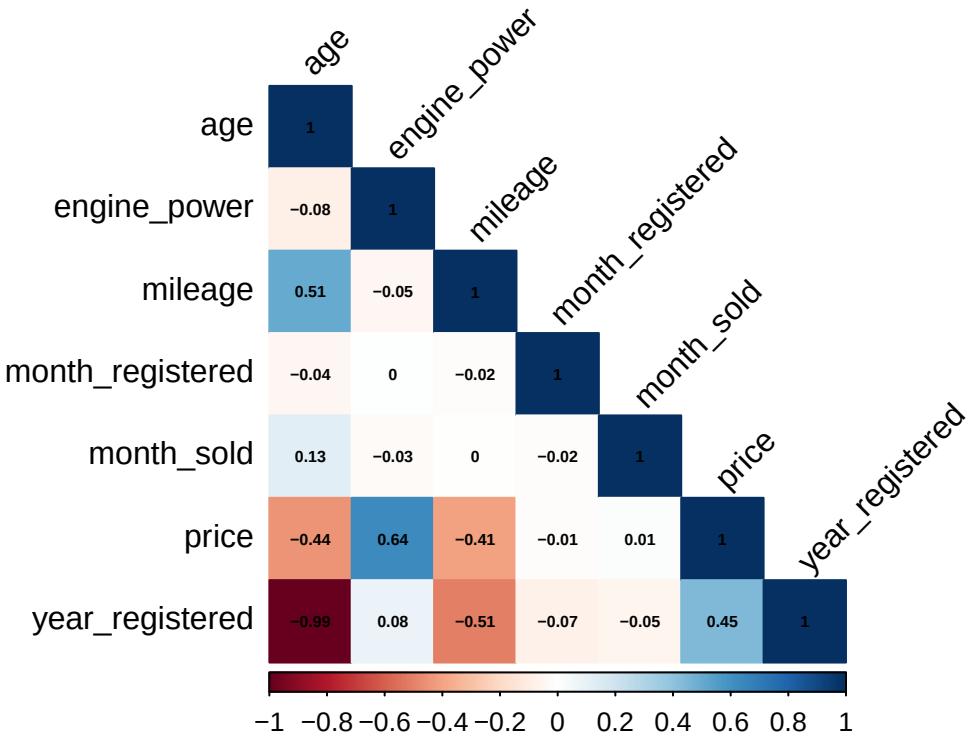


Correlation matrices between all features

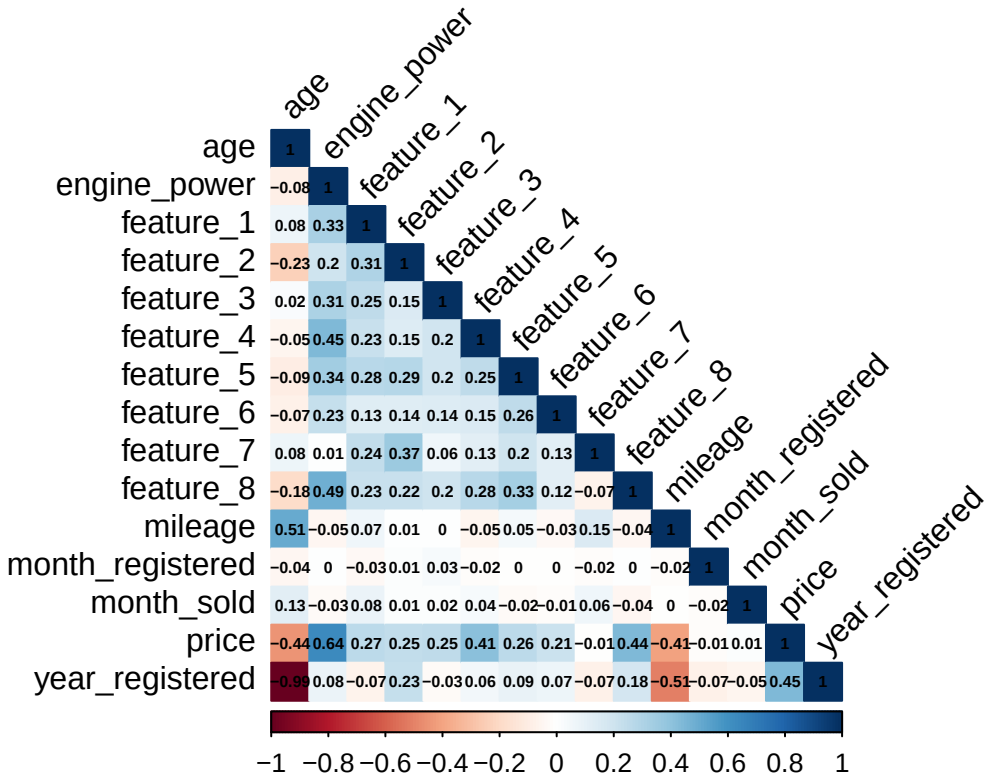
```
bmw_data_features_only <- select(bmw_data, starts_with("feature"))

remove_cols <- colnames(bmw_data_features_only)
bmw_data_minus_features <- bmw_data[, !(names(bmw_data) %in% remove_cols)]

helper_cor_plot <- function(dataframe, fontsize=1) {
  # convert any Boolean values to numeric values
  dataframe <- data.frame(lapply(dataframe, function(col) {
    if(is.logical(col)) as.numeric(col) else col
  }))
  cor_matrix <- cor(dataframe[sapply(dataframe, is.numeric)])
  cor_matrix <- cor_matrix[order(rownames(cor_matrix)), order(colnames(cor_matrix))]
  corrplot(cor_matrix, method = "color", type = "lower",
           tl.col = "black", tl.srt = 45, # Text label color and rotation
           addCoef.col = "black", number.cex = 0.5) # Add correlation coefficients to the plot
}

helper_cor_plot(bmw_data_features_only)
```




```
helper_cor_plot(bmw_data_minus_features)
```




```
helper_cor_plot(bmw_data, 0.5)
```



We observe medium-sized (> 0.3) correlations for the following...

- Feature 8 and Feature 5
- Feature 7 and Feature 2
- Feature 2 and Feature 1

Apart from those medium-sized correlations, we note that all features apart from Feature 7 and Feature 8 have *positive* correlations.

We see that price has the strongest correlations (> 0.4) with the age, engine power, Feature 4, Feature 8, and milage. Year Registered is also strongly correlated with price, but it essentially mirrors Age (-0.99 correlation) and therefore is discarded from consideration.

Interestingly, engine power and Feature 8 are strongly correlated (0.49), as are less surprisingly age and mileage (0.51). The strongest absolute correlation to price lies with engine power. This is slightly surprising as it may have been assumed that effective wear (which mileage is a proxy for) would have been the most correlated.

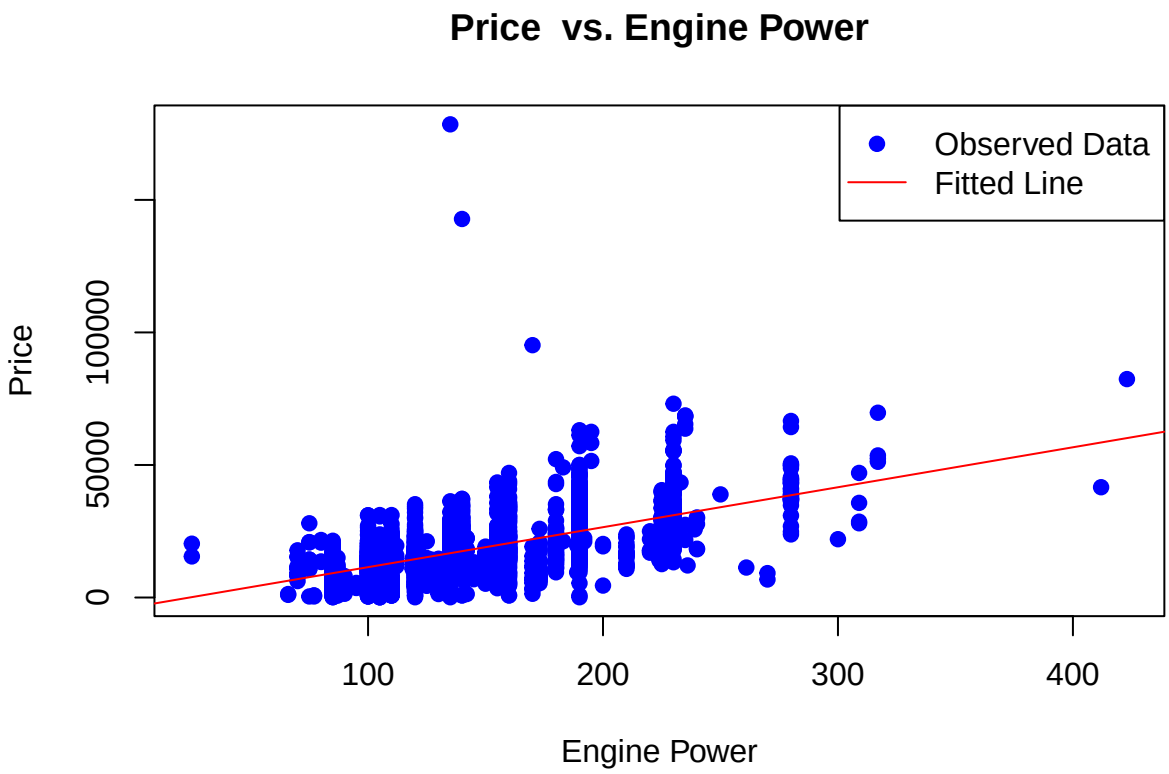
However, this finding is not unprecedented as in our graphs of cars partitioned by age, we saw that engine power was the most consistent predictor.

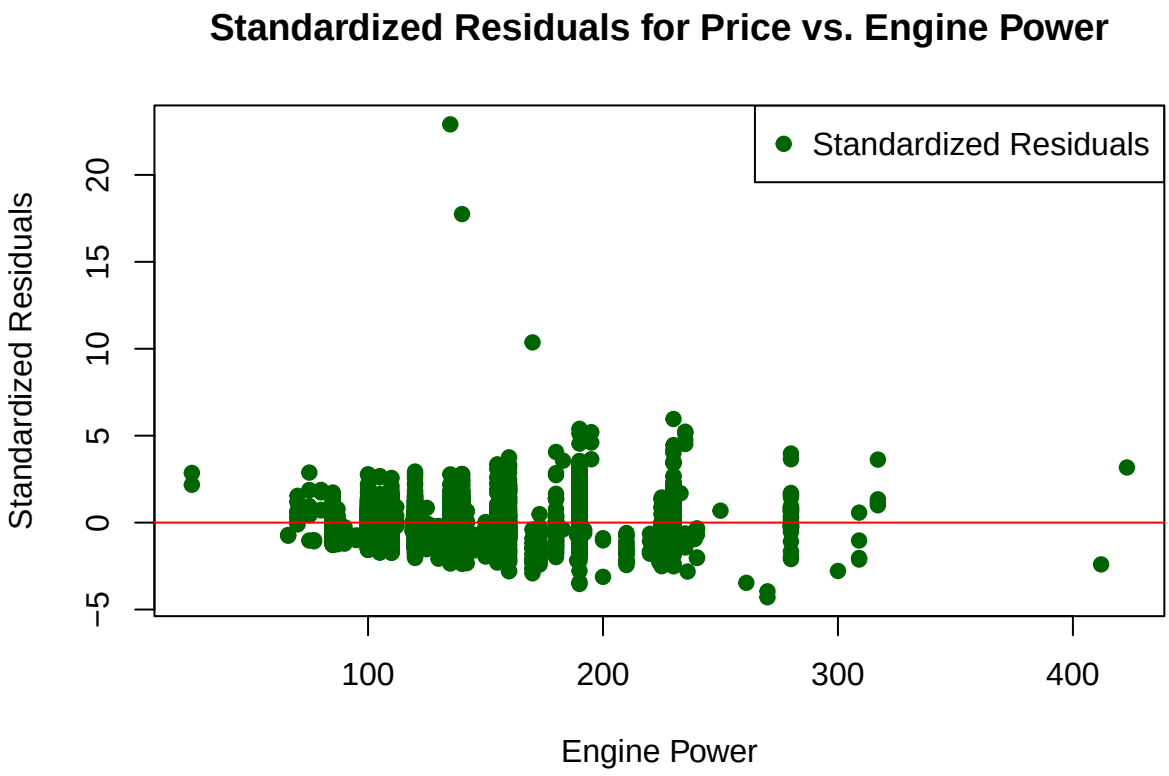
Defining models

Simple linear regression

```
# defining a helper function to plot the OLS and standardized residuals
plotModelAndResiduals <- function(x, y, x_name, y_name) {
  # Fitting the linear model
```

```r
  model <- lm(y ~ x)

  # Base plot for the model
  plot(x, y,
       main = paste(y_name, " vs.", x_name),
       xlab = x_name,
       ylab = y_name,
       pch = 19,
       col = "blue")
  abline(model, col = "red")

  # Adding a legend for the model plot
  legend("topright", legend = c("Observed Data", "Fitted Line"),
         col = c("blue", "red"), pch = c(19, NA), lty = c(NA, 1))

  # Calculating standardized residuals
  std_res <- rstandard(model)

  # Plot for standardized residuals
  plot(x, std_res,
       main = paste("Standardized Residuals for", y_name, "vs.", x_name),
       xlab = x_name,
       ylab = "Standardized Residuals",
       pch = 19,
       col = "darkgreen")

  # Adding a horizontal line at 0 in the residuals plot
  abline(h = 0, col = "red")

  # Adding a legend for the residuals plot
  legend("topright", legend = "Standardized Residuals", col = "darkgreen", pch = 19)

  return(model)
}


# fitting the linear model
y_1 <- bmw_data$price
x_1 <- bmw_data$engine_power
model_1 <- plotModelAndResiduals(x_1, y_1, "Engine Power", "Price")
```


Price vs. Engine Power
Observed Data
Fitted Line
Price
0
50000
100000
100
200
300
400
Engine Power




```
summary(model_1)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -30275  -3315    -22   2594 161778
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3630.343    351.377  -10.33   <2e-16 ***
## x             150.759      2.608   57.81   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7063 on 4839 degrees of freedom
## Multiple R-squared:  0.4085, Adjusted R-squared:  0.4084
## F-statistic:  3342 on 1 and 4839 DF,  p-value: < 2.2e-16
```

The model looks fairly good as-is apart from some major outliers that are lying far outside 5 standard deviations.

```
# calculate standardized residuals from the model
std_res <- rstandard(model_1)
```

```
# set threshold for standardized residuals
threshold <- 5.5

# find indices of data points with standardized residuals above the threshold
indices_above <- which(std_res > threshold)

# collect these data points from the original data
bmw_data_above <- bmw_data[indices_above, ]

# display dataframe
bmw_data_above
```

```
##      maker_key model_key mileage engine_power registration_date          fuel
## 91         BMW        i8   12402          170          1/1/2016 hybrid_petrol
## 4685       BMW        X4  103222          140          8/1/2014        diesel
## 4732       BMW      X5 M    4530          230          7/1/2016        diesel
## 4754       BMW        X3   72515          135          3/1/2013        diesel
##      paint_color car_type feature_1 feature_2 feature_3 feature_4 feature_5
## 91          grey    coupe      TRUE      TRUE     FALSE     FALSE     FALSE
## 4685        grey      suv      TRUE      TRUE     FALSE     FALSE      TRUE
## 4732      silver      suv      TRUE      TRUE     FALSE      TRUE     FALSE
## 4754        blue      suv     FALSE     FALSE      TRUE     FALSE     FALSE
##      feature_6 feature_7 feature_8  price  sold_at month_sold month_registered
## 91       FALSE     FALSE      TRUE  95200 4/1/2018          4                1
## 4685     FALSE      TRUE      TRUE 142800 8/1/2018          8                8
## 4732     FALSE     FALSE      TRUE  73100 8/1/2018          8                7
## 4754     FALSE     FALSE      TRUE 178500 8/1/2018          8                3
##      year_registered      age age_class age_cat engine_cat
## 91              2016 2.250000     young   (2,4]     medium
## 4685            2014 4.000000     young   (2,4]     medium
## 4732            2016 2.083333     young   (2,4]       high
## 4754            2013 5.416667   average   (4,6]     medium
```

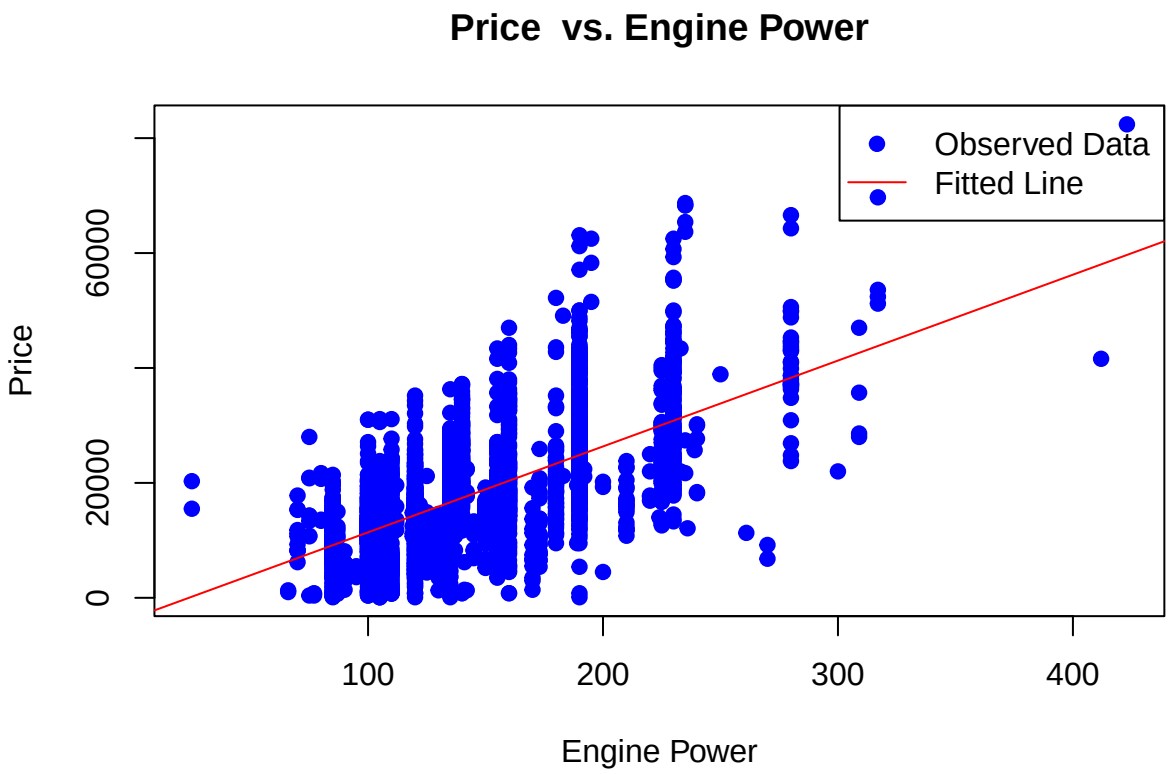
```
print(length(bmw_data_above$model_key))
```

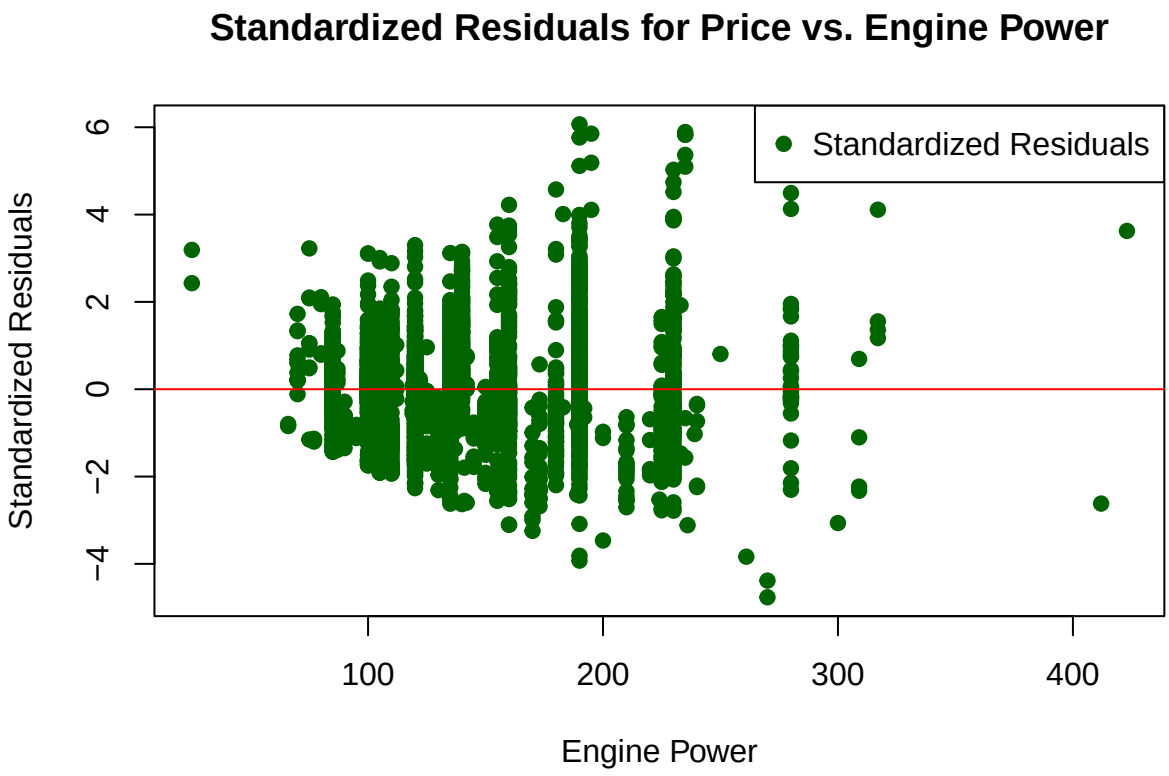
```
## [1] 4
```

When observing the 4 major outliers present in this model, we note that they are all relatively young, all have Feature 8, none have Feature 6, and most have Feature 1 and Feature 2 and yet don't have Feature 3 or Feature 4. All are different models with varying mileage (4,530 - 103,222). Three of the cars are SUVs. They make up $\sim 0.1\%$ of our data, yet with the maximal standardize residual of 22.9, they have a disproportionately high effect on our model. for that reason, we remove them all.

```
bmw_data_cleaned <- bmw_data[-indices_above, ]
```

```
y_2 <- bmw_data_cleaned$price
x_2 <- bmw_data_cleaned$engine_power
model_2 <- plotModelAndResiduals(x_2, y_2, "Engine Power", "Price")
```


Price vs. Engine Power
Observed Data
Fitted Line
Price
0
20000
60000
100
200
300
400
Engine Power



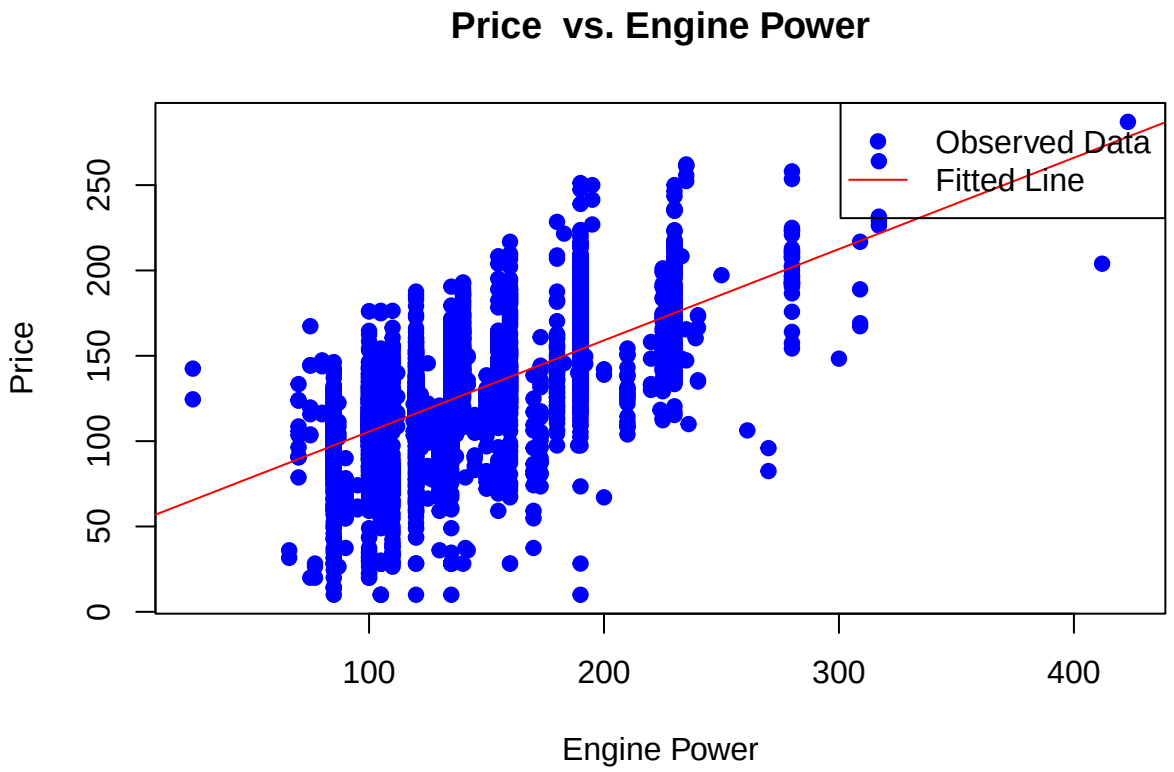

```
summary(model_2)
```

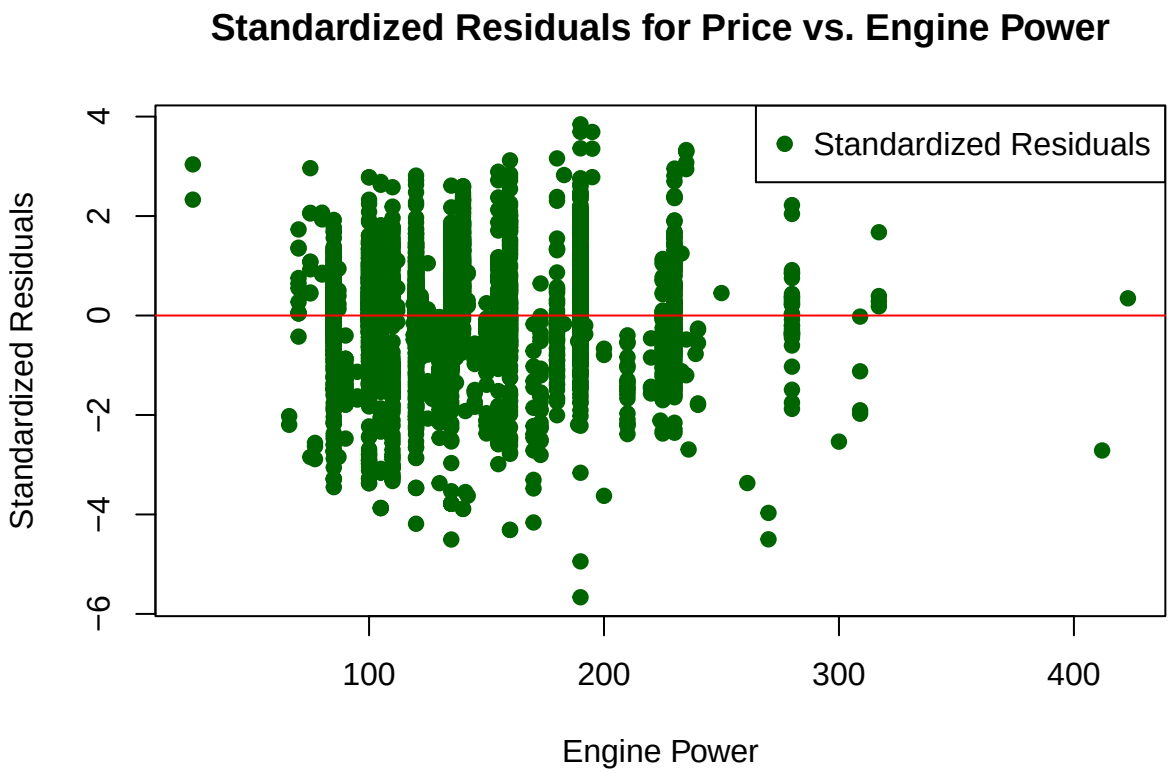
```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -30006  -3231     41   2669  38249
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3544.231    313.956  -11.29   <2e-16 ***
## x             149.447      2.331   64.12   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6307 on 4835 degrees of freedom
## Multiple R-squared:  0.4596, Adjusted R-squared:  0.4594
## F-statistic:  4111 on 1 and 4835 DF,  p-value: < 2.2e-16
```

By simply removing the four outliers, we were able to greatly improve our standardized residuals graph and brought our R^2 value up to 0.4596, a significant improvement. However, we see nonconstant variance in which the residuals grow larger as engine power goes long. Therefore, the model is not valid as-is.

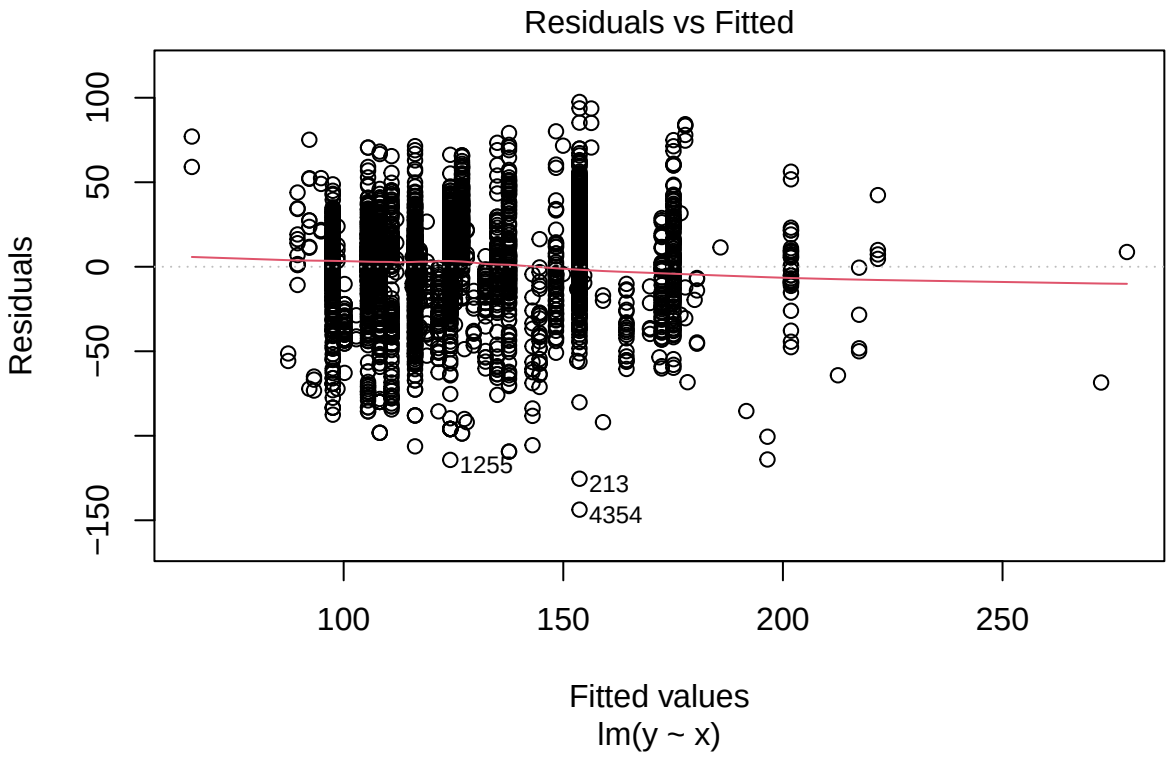
```
y_3 <- sqrt(bmw_data_cleaned$price)
x_3 <- bmw_data_cleaned$engine_power
```

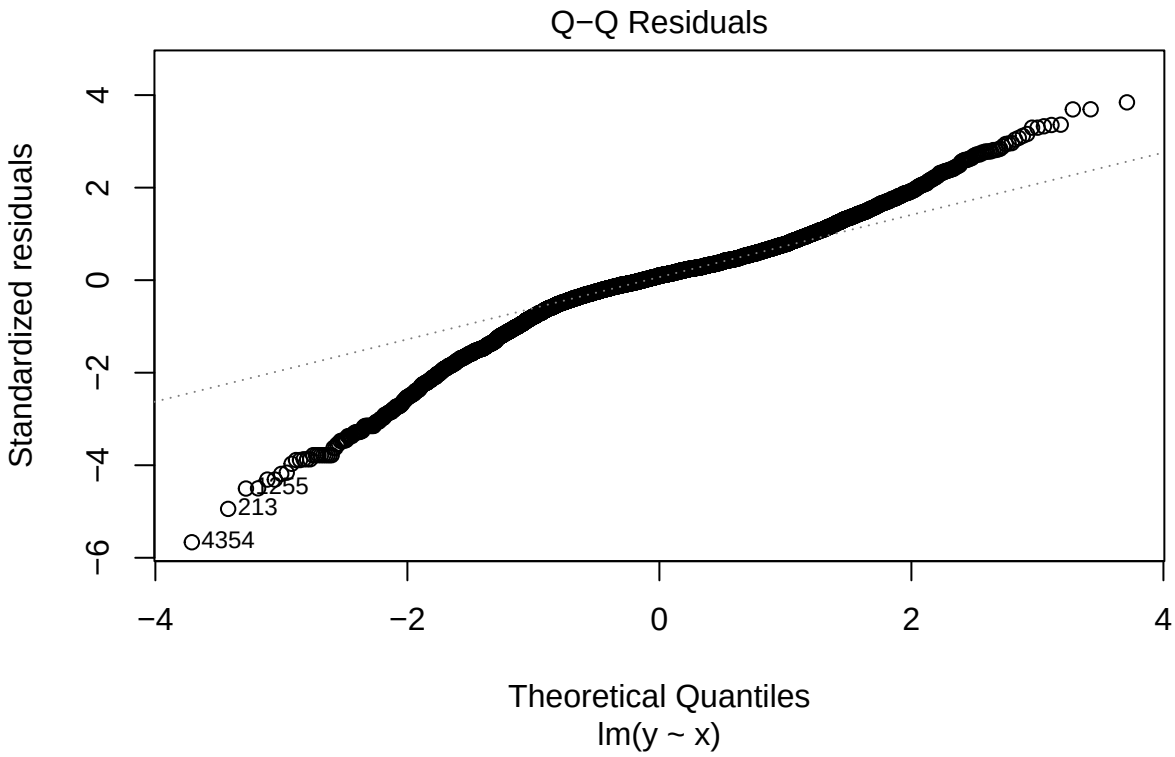
```
model_3 <- plotModelAndResiduals(x_3, y_3, "Engine Power", "Price")
```

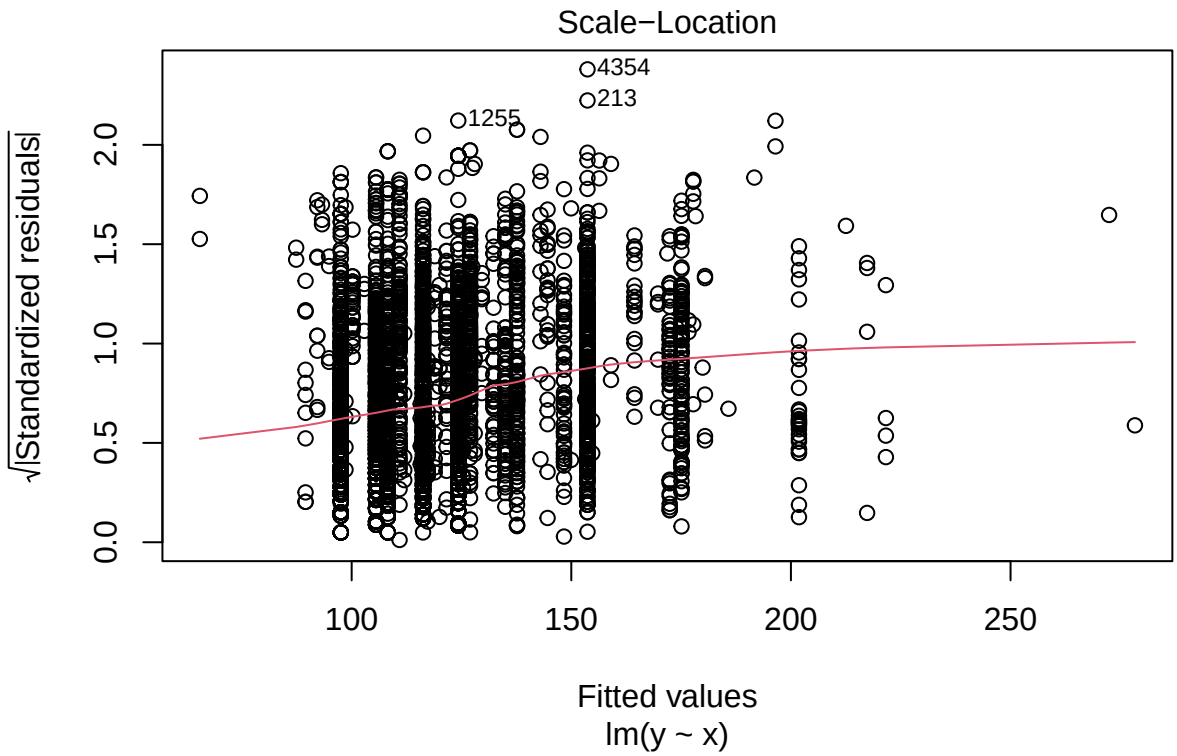


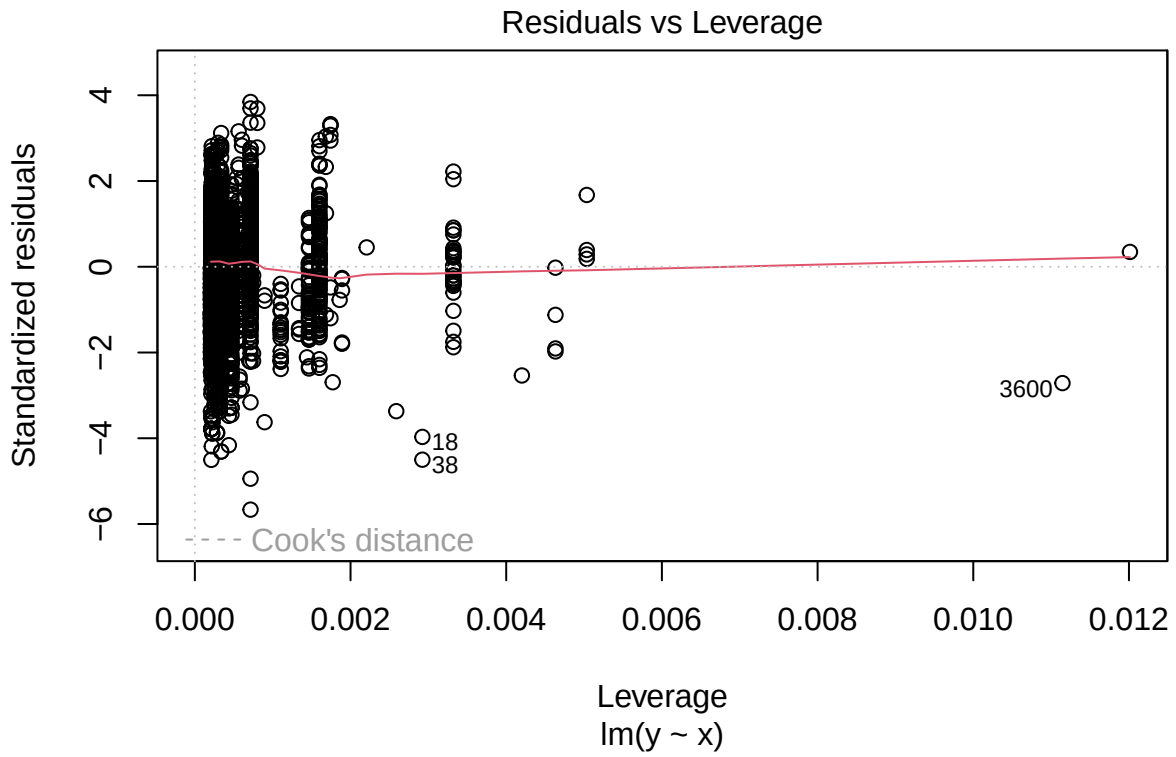

Standardized Residuals for Price vs. Engine Power
Standardized Residuals
Standardized Residuals
4
2
0
−2
−4
−6
100
200
300
400
Engine Power

```
plot(model_3)
```


Residuals vs Fitted
Residuals
100
50
0
−50
−150
1255
213
4354
100
150
200
250
Fitted values
lm(y ~ x)


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(y ~ x)
4
2
0
−2
−4
−6
−4
−2
0
2
4
1255
213
4354


Scale−Location
4354
213
1255
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
100
150
200
250
Fitted values
lm(y ~ x)




```
summary(model_3)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -143.694   -9.819    2.470   13.203   97.504
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 52.063570   1.263150   41.22   <2e-16 ***
## x            0.534895   0.009377   57.04   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.38 on 4835 degrees of freedom
## Multiple R-squared:  0.4023, Adjusted R-squared:  0.4021
## F-statistic:  3254 on 1 and 4835 DF,  p-value: < 2.2e-16
```

By taking the square root of price, we are able to greatly enhance the data's adherence to our assumptions of constant variance and a mean of errors equal to 0. We notice that the median residual is at a low of 2.47, compared to prior median residuals of -52595 and -22.

We note that we have potential outlier observations in 18, 38, 3600, 1255, 213, 4354, and 255. In the theoretical quantiles plot and scale-location plot, we see 213 and 4354 occur in both, and they are the only

two observations to both be marked as outliers more than once.

```
potential_outlier <- bmw_data[213,]
print(potential_outlier)
```

```
##     maker_key model_key mileage engine_power registration_date   fuel
## 213       BMW       316   94829           85          8/1/2013 diesel
##     paint_color car_type feature_1 feature_2 feature_3 feature_4 feature_5
## 213        grey   estate     FALSE     FALSE     FALSE     FALSE     FALSE
##     feature_6 feature_7 feature_8 price  sold_at month_sold month_registered
## 213     FALSE     FALSE     FALSE 12600 4/1/2018          4               8
##     year_registered      age age_class age_cat engine_cat
## 213            2013 4.666667   average   (4,6]        low
```

```
potential_outlier <- bmw_data[4354,]
print(potential_outlier)
```

```
##      maker_key model_key mileage engine_power registration_date   fuel
## 4356       BMW        X1  168996          100         11/1/2010 diesel
##      paint_color car_type feature_1 feature_2 feature_3 feature_4 feature_5
## 4356       brown      suv     FALSE     FALSE     FALSE     FALSE     FALSE
##      feature_6 feature_7 feature_8 price  sold_at month_sold month_registered
## 4356     FALSE      TRUE      TRUE  5800 5/1/2018          5              11
##      year_registered age age_class age_cat engine_cat
## 4356            2010 7.5   average   (6,8]        low
```

At first glance, nothing seems exceptionally wrong with these items. Both vehicles are perhaps cheap, particularly the SUV, but they also only share two features of a potential 16 amongst them and have medium mileage and low engine power. These outliers are due for further investigation later. In our research, we came across the RPubs modeling done here (https://rpubs.com/Adetya/650497) in which approximately 10 SUVs were grealty skewing the model. Therefore, we will cross-validate our work in removing (or not) assumptions with their work as it becomes necessary/pertinent.

Multiple linear regression

From our previous work with correlation matrices, we can see that the potentially most predictive elements for price included engine power, mileage, age, Feature 3, and Feature 8. However, other features had positive effects and we did not investigate the effect of paint and model.

We convert the model, color, paint, and month-sold and month-registered information all to categorical data for R to interpret as separate classes.

```
# setting the categorical variables to factors (so R treats them as categorical)
bmw_data_cleaned$model_key <- as.factor(bmw_data_cleaned$model_key)
bmw_data_cleaned$car_type <- as.factor(bmw_data_cleaned$car_type)
bmw_data_cleaned$paint_color <- as.factor(bmw_data_cleaned$paint_color)
bmw_data_cleaned$month_sold <- as.factor(bmw_data_cleaned$month_sold)
bmw_data_cleaned$fuel <- as.factor(bmw_data_cleaned$fuel)
```

We now create a multiple linear regression including all potentially informative features that do not exhibit immediately obvious multicollinearity (i.e., year registered) and shall use summary information from the model to see which features are most important. We also exclude month registered as that contains information already present in age feature and seems illogical to influence a selling price later on.

```
mlr_all <- lm(formula = price ~
                engine_power + age + mileage + model_key + car_type + paint_color + month_sold + fuel +
                feature_1 + feature_2 + feature_3 + feature_4 + feature_5 + feature_6 + feature_7 + fea
```

```
               data = bmw_data_cleaned)
summary(mlr_all)
```

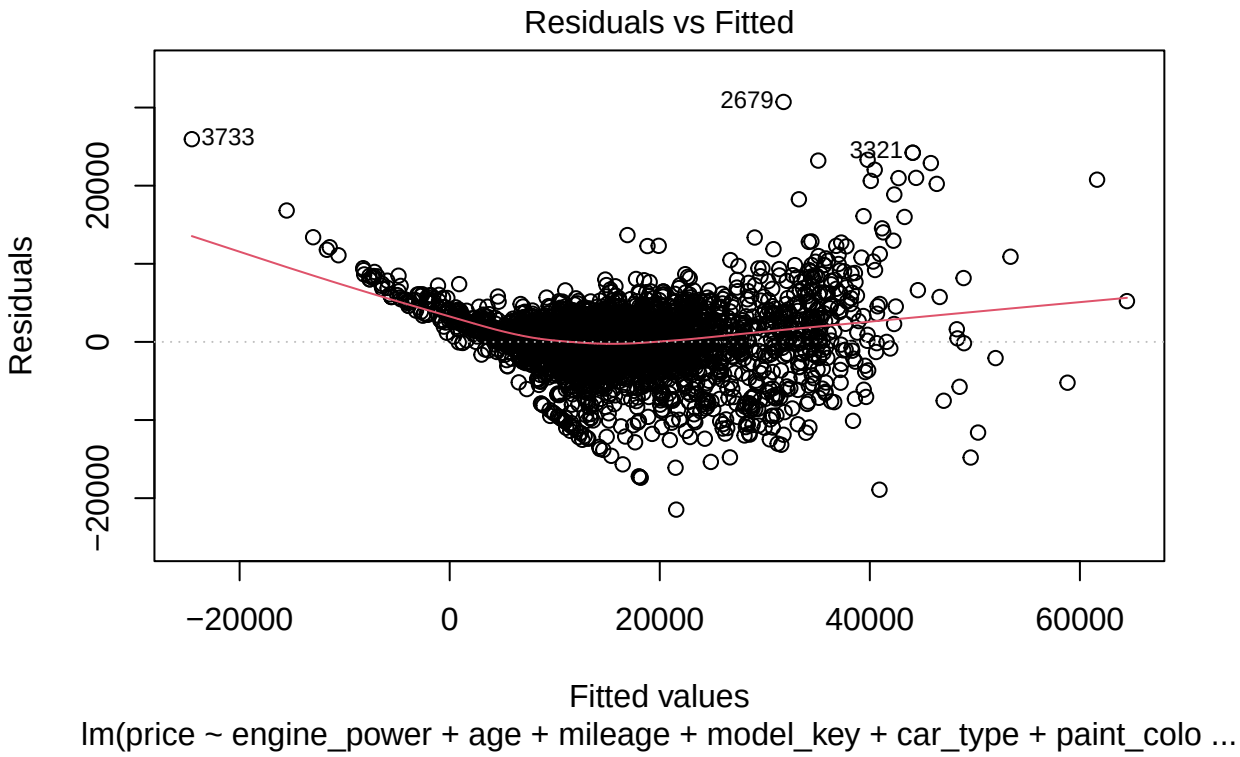
```
##
## Call:
## lm(formula = price ~ engine_power + age + mileage + model_key +
##     car_type + paint_color + month_sold + fuel + feature_1 +
##     feature_2 + feature_3 + feature_4 + feature_5 + feature_6 +
##     feature_7 + feature_8, data = bmw_data_cleaned)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21468.0  -1386.5     14.7   1480.4  30714.9
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  4.283e+04  3.285e+03  13.038  < 2e-16 ***
## engine_power                 5.198e+01  3.965e+00  13.108  < 2e-16 ***
## age                         -1.059e+03  2.923e+01 -36.232  < 2e-16 ***
## mileage                     -3.095e-02  1.079e-03 -28.699  < 2e-16 ***
## model_keyX6 M               -4.751e+03  3.286e+03  -1.446   0.1483
## model_keyX5 M50             -6.565e+03  3.664e+03  -1.792   0.0733 .
## model_keyM5                 -1.744e+04  4.424e+03  -3.942 8.21e-05 ***
## model_keyX5 M               -1.328e+04  3.187e+03  -4.165 3.17e-05 ***
## model_key740                -1.369e+04  2.744e+03  -4.991 6.22e-07 ***
## model_key750                -1.960e+04  3.629e+03  -5.400 6.99e-08 ***
## model_key640 Gran Coupé     -1.490e+04  2.737e+03  -5.443 5.50e-08 ***
## model_keyM3                 -1.399e+04  2.874e+03  -4.870 1.16e-06 ***
## model_keyM550               -1.952e+04  2.798e+03  -6.977 3.43e-12 ***
## model_keyX6                 -1.481e+04  3.101e+03  -4.776 1.84e-06 ***
## model_key640                -1.588e+04  2.882e+03  -5.509 3.81e-08 ***
## model_key435 Gran Coupé     -2.024e+04  2.973e+03  -6.810 1.10e-11 ***
## model_keyX4                 -1.769e+04  3.128e+03  -5.655 1.65e-08 ***
## model_key435                -2.371e+04  3.133e+03  -7.569 4.49e-14 ***
## model_key425                -1.906e+04  3.630e+03  -5.250 1.59e-07 ***
## model_keyX5                 -1.696e+04  3.071e+03  -5.522 3.54e-08 ***
## model_key430                -2.257e+04  3.613e+03  -6.246 4.57e-10 ***
## model_keyM235               -2.560e+04  3.294e+03  -7.770 9.54e-15 ***
## model_key430 Gran Coupé     -2.442e+04  3.164e+03  -7.717 1.45e-14 ***
## model_keyM135               -2.537e+04  4.417e+03  -5.744 9.80e-09 ***
## model_key330 Gran Turismo   -2.651e+04  3.355e+03  -7.900 3.43e-15 ***
## model_key535 Gran Turismo   -2.698e+04  3.620e+03  -7.451 1.10e-13 ***
## model_key335 Gran Turismo   -2.697e+04  3.341e+03  -8.071 8.75e-16 ***
## model_key420 Gran Coupé     -2.369e+04  2.760e+03  -8.581  < 2e-16 ***
## model_key420                -2.345e+04  2.700e+03  -8.686  < 2e-16 ***
## model_key220                -2.797e+04  3.381e+03  -8.273  < 2e-16 ***
## model_key730                -2.137e+04  2.720e+03  -7.856 4.87e-15 ***
## model_key535                -2.534e+04  2.666e+03  -9.505  < 2e-16 ***
## model_key135                -2.930e+04  3.123e+03  -9.382  < 2e-16 ***
## model_key335                -2.705e+04  2.909e+03  -9.300  < 2e-16 ***
## model_keyi3                 -2.487e+04  4.093e+03  -6.076 1.33e-09 ***
## model_keyActiveHybrid 5     -3.414e+04  4.900e+03  -6.968 3.66e-12 ***
## model_key530 Gran Turismo   -2.510e+04  2.773e+03  -9.049  < 2e-16 ***
## model_key418 Gran Coupé     -2.423e+04  3.004e+03  -8.066 9.12e-16 ***
```

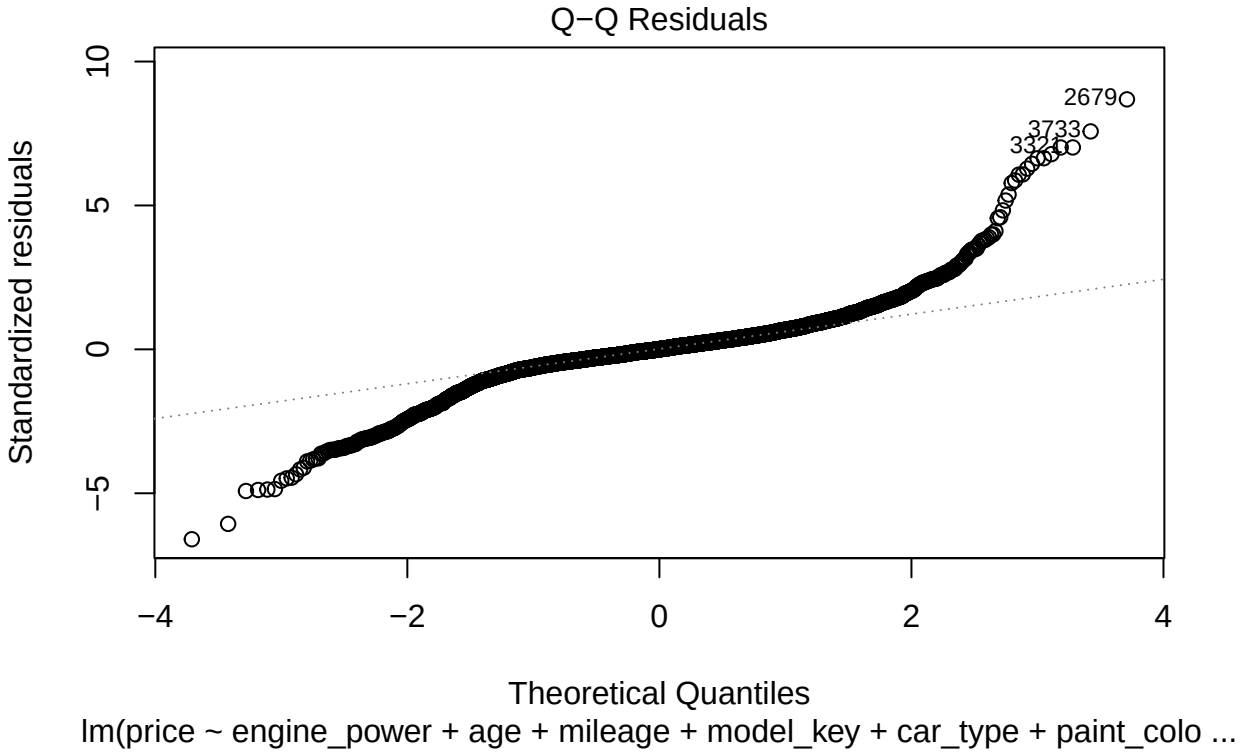
```
## model_key325 Gran Turismo  -2.433e+04  3.203e+03  -7.597 3.64e-14 ***
## model_key528               -2.595e+04  3.007e+03  -8.629  < 2e-16 ***
## model_key520 Gran Turismo  -2.385e+04  2.799e+03  -8.522  < 2e-16 ***
## model_key530               -2.563e+04  2.654e+03  -9.659  < 2e-16 ***
## model_key635               -2.522e+04  4.415e+03  -5.711 1.19e-08 ***
## model_key225 Active Tourer -2.484e+04  4.652e+03  -5.340 9.75e-08 ***
## model_key225               -3.151e+04  4.437e+03  -7.101 1.43e-12 ***
## model_keyX3                -2.587e+04  3.092e+03  -8.367  < 2e-16 ***
## model_key214 Gran Tourer   -2.211e+04  4.719e+03  -4.685 2.88e-06 ***
## model_key320 Gran Turismo  -2.742e+04  2.715e+03 -10.102  < 2e-16 ***
## model_key218               -2.783e+04  2.873e+03  -9.688  < 2e-16 ***
## model_key216 Gran Tourer   -2.316e+04  3.504e+03  -6.609 4.29e-11 ***
## model_key330               -2.925e+04  2.700e+03 -10.832  < 2e-16 ***
## model_key328               -2.873e+04  3.079e+03  -9.332  < 2e-16 ***
## model_key518               -2.554e+04  2.751e+03  -9.287  < 2e-16 ***
## model_key218 Gran Tourer   -2.566e+04  3.225e+03  -7.957 2.20e-15 ***
## model_key520               -2.580e+04  2.688e+03  -9.599  < 2e-16 ***
## model_key218 Active Tourer -2.403e+04  3.140e+03  -7.654 2.35e-14 ***
## model_key525               -2.616e+04  2.679e+03  -9.768  < 2e-16 ***
## model_key318 Gran Turismo  -2.719e+04  2.741e+03  -9.919  < 2e-16 ***
## model_key325               -2.992e+04  2.755e+03 -10.861  < 2e-16 ***
## model_keyX1                -2.893e+04  3.116e+03  -9.282  < 2e-16 ***
## model_key216 Active Tourer -2.462e+04  3.653e+03  -6.739 1.78e-11 ***
## model_key125               -3.313e+04  3.113e+03 -10.641  < 2e-16 ***
## model_key120               -3.038e+04  2.733e+03 -11.114  < 2e-16 ***
## model_key320               -2.870e+04  2.685e+03 -10.691  < 2e-16 ***
## model_key220 Active Tourer -2.816e+04  4.633e+03  -6.080 1.30e-09 ***
## model_key114               -3.075e+04  2.904e+03 -10.588  < 2e-16 ***
## model_key318               -2.824e+04  2.713e+03 -10.408  < 2e-16 ***
## model_key630               -3.105e+04  4.414e+03  -7.033 2.31e-12 ***
## model_key316               -2.872e+04  2.738e+03 -10.491  < 2e-16 ***
## model_key116               -3.062e+04  2.744e+03 -11.161  < 2e-16 ***
## model_key118               -3.034e+04  2.727e+03 -11.125  < 2e-16 ***
## model_keyZ4                -3.041e+04  3.032e+03 -10.028  < 2e-16 ***
## model_key123               -3.346e+04  3.680e+03  -9.092  < 2e-16 ***
## model_key650               -4.232e+04  3.597e+03 -11.767  < 2e-16 ***
## model_key523               -2.365e+04  3.217e+03  -7.351 2.30e-13 ***
## model_key216               -3.398e+04  4.673e+03  -7.271 4.14e-13 ***
## model_key735               -2.967e+04  4.426e+03  -6.703 2.29e-11 ***
## car_typesuv                -7.535e+02  1.678e+03  -0.449   0.6534
## car_typeconvertible         2.968e+03  7.277e+02   4.078 4.62e-05 ***
## car_typesedan               2.591e+02  5.599e+02   0.463   0.6436
## car_typevan                -5.772e+03  1.395e+03  -4.139 3.55e-05 ***
## car_typehatchback          -3.929e+02  6.059e+02  -0.648   0.5168
## car_typeestate             -1.754e+03  5.618e+02  -3.123   0.0018 **
## car_typesubcompact          6.376e+02  7.000e+02   0.911   0.3624
## paint_colorwhite            1.284e+03  1.536e+03   0.836   0.4034
## paint_colorred              1.244e+03  1.611e+03   0.772   0.4400
## paint_colorblack            8.944e+02  1.533e+03   0.583   0.5596
## paint_colorbeige            7.311e+02  1.636e+03   0.447   0.6550
## paint_colorgrey             7.788e+02  1.535e+03   0.507   0.6119
## paint_colorbrown            9.000e+02  1.544e+03   0.583   0.5601
## paint_colorblue             4.559e+02  1.537e+03   0.297   0.7668
## paint_colorsilver           6.480e+02  1.545e+03   0.420   0.6748
```

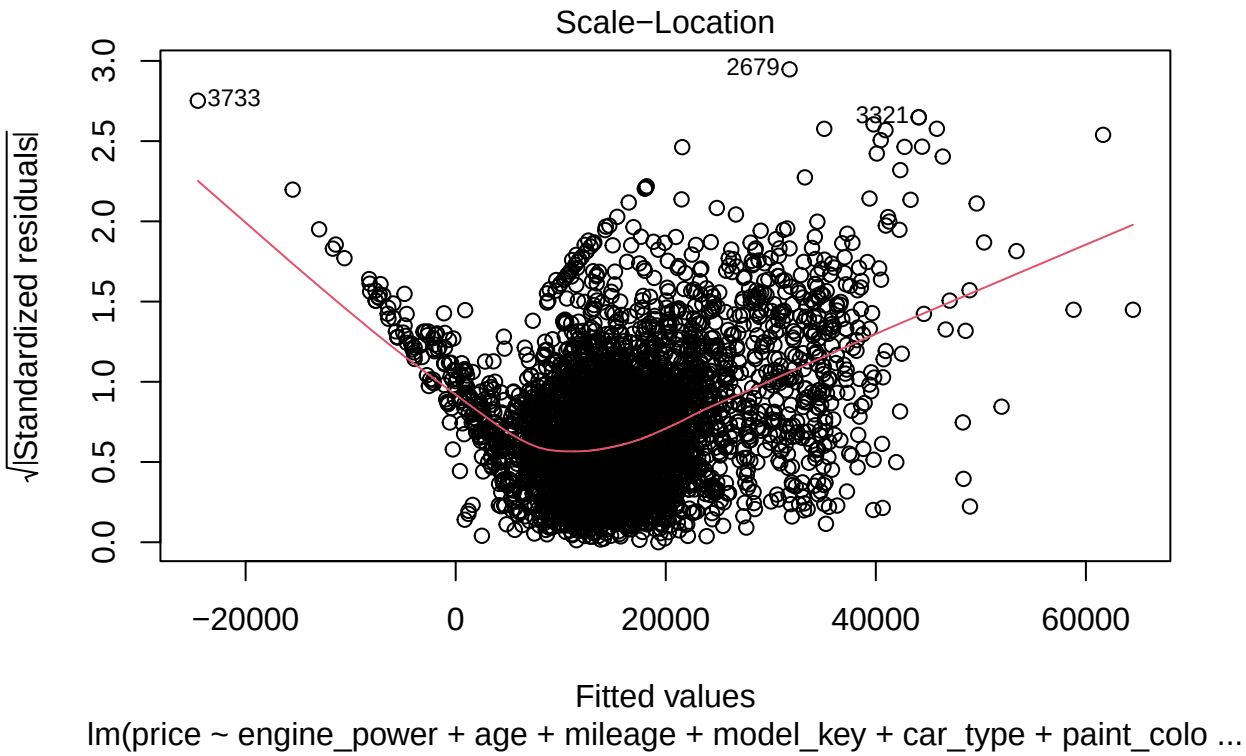
```
## paint_colorgreen            6.330e+02  1.762e+03   0.359   0.7195
## month_sold2                -3.073e-01  2.965e+02  -0.001   0.9992
## month_sold3                -1.546e+02  2.824e+02  -0.547   0.5842
## month_sold4                 1.162e+02  2.852e+02   0.408   0.6837
## month_sold5                 2.490e+02  2.801e+02   0.889   0.3740
## month_sold6                 1.597e+02  2.903e+02   0.550   0.5822
## month_sold7                 4.882e+02  2.955e+02   1.652   0.0985 .
## month_sold8                 6.519e+02  2.974e+02   2.192   0.0284 *
## month_sold9                 5.290e+02  3.498e+02   1.512   0.1305
## fuelelectro                 1.099e+03  3.598e+03   0.305   0.7601
## fuelhybrid_petrol           4.559e+03  2.080e+03   2.192   0.0285 *
## fuelpetrol                 -4.210e+02  3.240e+02  -1.300   0.1938
## feature_1TRUE               5.780e+02  1.252e+02   4.618 3.98e-06 ***
## feature_2TRUE              -2.877e-01  1.572e+02  -0.002   0.9985
## feature_3TRUE               6.930e+02  1.409e+02   4.918 9.05e-07 ***
## feature_4TRUE               1.435e+03  1.752e+02   8.190 3.34e-16 ***
## feature_5TRUE              -1.498e+02  1.252e+02  -1.197   0.2315
## feature_6TRUE               1.320e+03  1.321e+02   9.991  < 2e-16 ***
## feature_7TRUE               1.015e+03  2.446e+02   4.149 3.39e-05 ***
## feature_8TRUE               1.311e+03  1.294e+02  10.133  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3560 on 4725 degrees of freedom
## Multiple R-squared:  0.8317, Adjusted R-squared:  0.8278
## F-statistic: 210.4 on 111 and 4725 DF,  p-value: < 2.2e-16
```

```
plot(mlr_all)
```

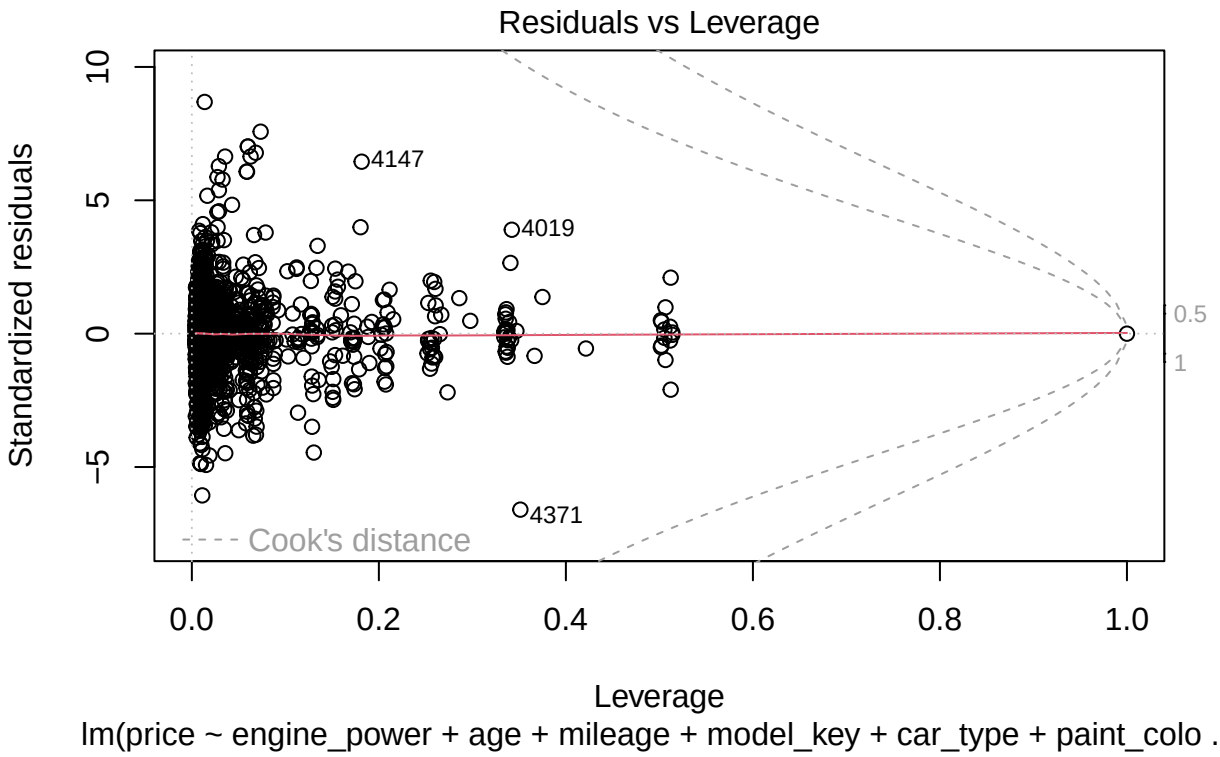
```
## Warning: not plotting observations with leverage one:
##   56, 107, 1899, 2925, 3153, 3600, 4798, 4799, 4820, 4821
```


Residuals vs Fitted
Residuals
Fitted values
lm(price ~ engine_power + age + mileage + model_key + car_type + paint_colo ...
3733
2679
3321
20000
0
−20000
−20000
0
20000
40000
60000


Q−Q Residuals
Standardized residuals
10
5
0
−5
2679
3733
−4
−2
0
2
4
Theoretical Quantiles
lm(price ~ engine_power + age + mileage + model_key + car_type + paint_colo ...




```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



We observe that when plotting a naive MLR with all features we have a quadratic pattern in the plain and square-rooted standardized residuals and several points are marked as outliers, chief amongst them being entries 3733, 2679, and 3321. We also see that the points 56, 107, 1899, 2925, 3153, 4798, 4799, and 4820 have reported leverages of **1** and thereforee were not plotted by R. Before we spend time inquiring further, we must whittle down our model to the most important features and consider transformations.

According to the summary output, all features but Feature 2 ($p \approx 0.99$) and Feature 5 ($p \approx 0.2$) were significant. No colors had any significance. Month sold does interestingly have some significance in the event of it being September ($p \approx 0.03$) or August ($p \approx 0.095$).

It may be tempting to say that the car market is hotter during the later Summer season; however, it is hard to generalize this finding because the dataset was sampled from a specific auction house.

Whether a car is a convertible, hatchback, or estate seems informative, and a vast majority of the models contain informative information as well regarding price. We see that all but two models possess signficance codes indiciating $p \approx 0$.

With respect to fuel type, we see that if a vehicle has the fuel class hybrid-petrol, that provides significant information, but the other two classes of fuel apart from diesel are not informative.

A new model could remove paint color, Feature 2, and Feature 5.

```
mlr_2 <- lm(formula = price ~
              engine_power + age + mileage + model_key + car_type + month_sold + fuel +
              feature_1 + feature_3 + feature_4 + feature_6 + feature_7 + feature_8,
              data = bmw_data_cleaned)
summary(mlr_2)
```

```
##
## Call:
```

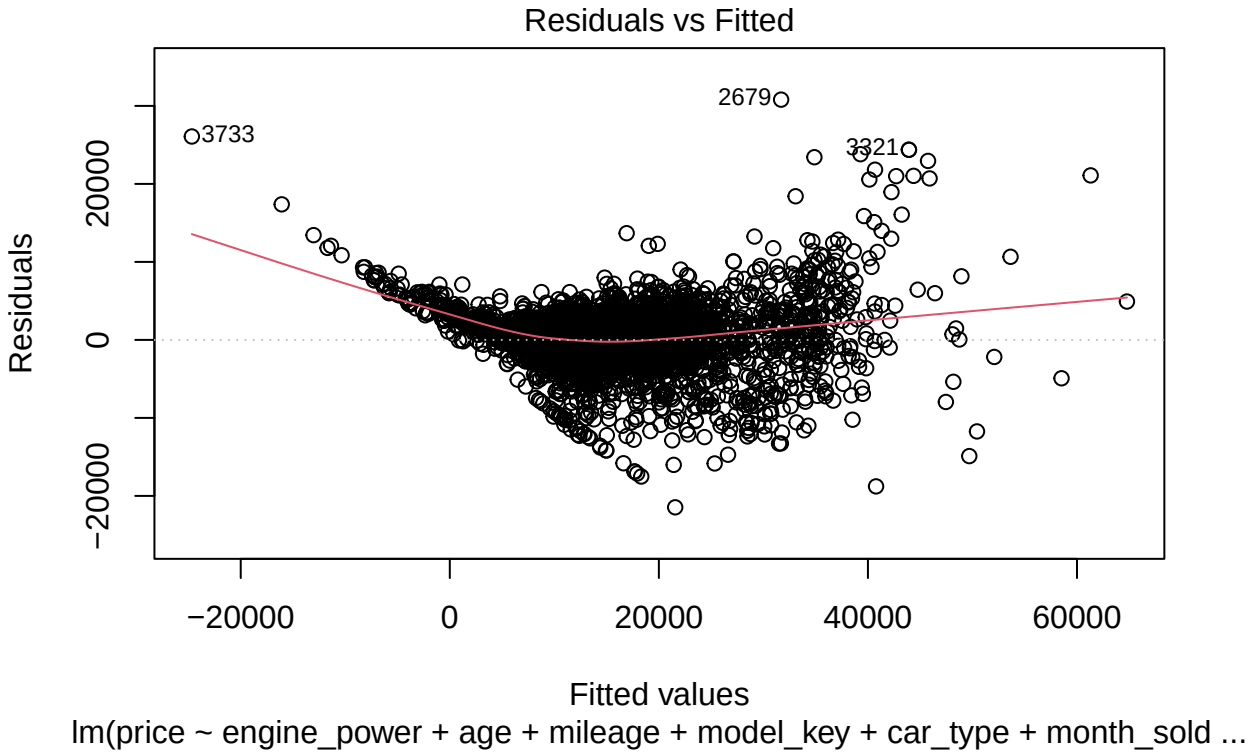
```
## lm(formula = price ~ engine_power + age + mileage + model_key +
##     car_type + month_sold + fuel + feature_1 + feature_3 + feature_4 +
##     feature_6 + feature_7 + feature_8, data = bmw_data_cleaned)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -21462  -1385      0   1496  30803
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  4.405e+04  2.863e+03  15.385  < 2e-16 ***
## engine_power                 5.175e+01  3.948e+00  13.110  < 2e-16 ***
## age                         -1.069e+03  2.771e+01 -38.571  < 2e-16 ***
## mileage                     -3.081e-02  1.073e-03 -28.714  < 2e-16 ***
## model_keyX6 M               -4.781e+03  3.285e+03  -1.455 0.145605
## model_keyX5 M50             -6.473e+03  3.666e+03  -1.766 0.077482 .
## model_keyM5                 -1.721e+04  4.424e+03  -3.890 0.000102 ***
## model_keyX5 M               -1.348e+04  3.185e+03  -4.231 2.37e-05 ***
## model_key740                -1.389e+04  2.741e+03  -5.068 4.16e-07 ***
## model_key750                -1.985e+04  3.626e+03  -5.474 4.63e-08 ***
## model_key640 Gran Coupé     -1.514e+04  2.733e+03  -5.540 3.19e-08 ***
## model_keyM3                 -1.431e+04  2.873e+03  -4.982 6.52e-07 ***
## model_keyM550               -1.998e+04  2.794e+03  -7.152 9.86e-13 ***
## model_keyX6                 -1.493e+04  3.098e+03  -4.819 1.49e-06 ***
## model_key640                -1.631e+04  2.877e+03  -5.669 1.52e-08 ***
## model_key435 Gran Coupé     -2.058e+04  2.971e+03  -6.927 4.88e-12 ***
## model_keyX4                 -1.791e+04  3.123e+03  -5.735 1.03e-08 ***
## model_key435                -2.385e+04  3.132e+03  -7.614 3.20e-14 ***
## model_key425                -1.961e+04  3.628e+03  -5.405 6.81e-08 ***
## model_keyX5                 -1.716e+04  3.068e+03  -5.594 2.35e-08 ***
## model_key430                -2.301e+04  3.612e+03  -6.370 2.06e-10 ***
## model_keyM235               -2.595e+04  3.292e+03  -7.882 3.97e-15 ***
## model_key430 Gran Coupé     -2.462e+04  3.161e+03  -7.790 8.20e-15 ***
## model_keyM135               -2.579e+04  4.411e+03  -5.846 5.38e-09 ***
## model_key330 Gran Turismo   -2.682e+04  3.347e+03  -8.012 1.41e-15 ***
## model_key535 Gran Turismo   -2.710e+04  3.618e+03  -7.490 8.19e-14 ***
## model_key335 Gran Turismo   -2.736e+04  3.338e+03  -8.197 3.15e-16 ***
## model_key420 Gran Coupé     -2.395e+04  2.753e+03  -8.699  < 2e-16 ***
## model_key420                -2.379e+04  2.694e+03  -8.830  < 2e-16 ***
## model_key220                -2.825e+04  3.338e+03  -8.463  < 2e-16 ***
## model_key730                -2.163e+04  2.714e+03  -7.970 1.98e-15 ***
## model_key535                -2.560e+04  2.662e+03  -9.615  < 2e-16 ***
## model_key135                -2.934e+04  3.122e+03  -9.398  < 2e-16 ***
## model_key335                -2.716e+04  2.903e+03  -9.356  < 2e-16 ***
## model_keyi3                 -2.489e+04  4.090e+03  -6.086 1.25e-09 ***
## model_keyActiveHybrid 5     -3.419e+04  4.895e+03  -6.985 3.25e-12 ***
## model_key530 Gran Turismo   -2.538e+04  2.769e+03  -9.166  < 2e-16 ***
## model_key418 Gran Coupé     -2.467e+04  2.995e+03  -8.237 2.25e-16 ***
## model_key325 Gran Turismo   -2.453e+04  3.199e+03  -7.669 2.09e-14 ***
## model_key528                -2.639e+04  3.001e+03  -8.792  < 2e-16 ***
## model_key520 Gran Turismo   -2.422e+04  2.792e+03  -8.676  < 2e-16 ***
## model_key530                -2.595e+04  2.648e+03  -9.798  < 2e-16 ***
## model_key635                -2.554e+04  4.415e+03  -5.786 7.69e-09 ***
## model_key225 Active Tourer -2.523e+04  4.617e+03  -5.464 4.89e-08 ***
```

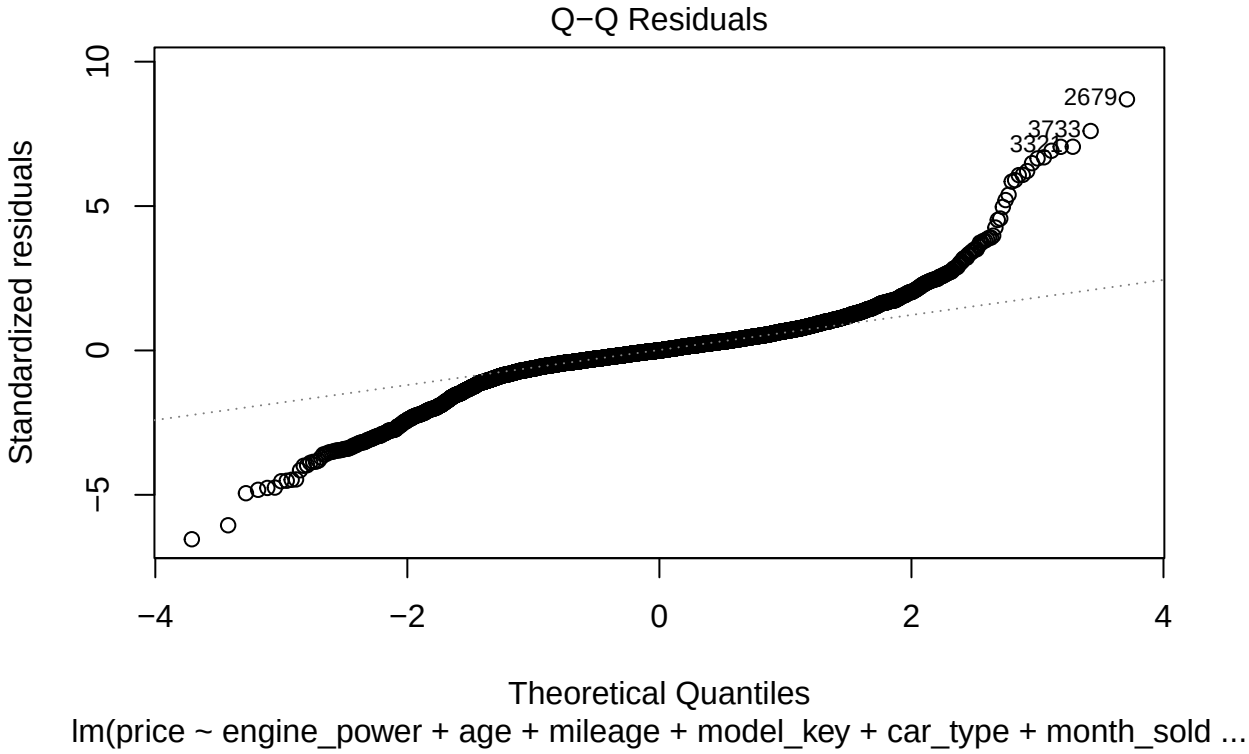
```
## model_key225               -3.168e+04  4.438e+03  -7.139 1.08e-12 ***
## model_keyX3                -2.607e+04  3.088e+03  -8.442  < 2e-16 ***
## model_key214 Gran Tourer   -2.189e+04  4.682e+03  -4.675 3.02e-06 ***
## model_key320 Gran Turismo  -2.771e+04  2.709e+03 -10.227  < 2e-16 ***
## model_key218               -2.816e+04  2.869e+03  -9.818  < 2e-16 ***
## model_key216 Gran Tourer   -2.348e+04  3.502e+03  -6.706 2.23e-11 ***
## model_key330               -2.948e+04  2.696e+03 -10.932  < 2e-16 ***
## model_key328               -2.913e+04  3.074e+03  -9.477  < 2e-16 ***
## model_key518               -2.587e+04  2.741e+03  -9.438  < 2e-16 ***
## model_key218 Gran Tourer   -2.589e+04  3.221e+03  -8.036 1.16e-15 ***
## model_key520               -2.608e+04  2.682e+03  -9.725  < 2e-16 ***
## model_key218 Active Tourer -2.424e+04  3.139e+03  -7.724 1.37e-14 ***
## model_key525               -2.645e+04  2.673e+03  -9.895  < 2e-16 ***
## model_key318 Gran Turismo  -2.749e+04  2.736e+03 -10.047  < 2e-16 ***
## model_key325               -3.025e+04  2.750e+03 -11.000  < 2e-16 ***
## model_keyX1                -2.913e+04  3.113e+03  -9.357  < 2e-16 ***
## model_key216 Active Tourer -2.492e+04  3.650e+03  -6.826 9.82e-12 ***
## model_key125               -3.332e+04  3.103e+03 -10.736  < 2e-16 ***
## model_key120               -3.070e+04  2.728e+03 -11.254  < 2e-16 ***
## model_key320               -2.896e+04  2.679e+03 -10.811  < 2e-16 ***
## model_key220 Active Tourer -2.848e+04  4.634e+03  -6.147 8.54e-10 ***
## model_key114               -3.096e+04  2.898e+03 -10.686  < 2e-16 ***
## model_key318               -2.847e+04  2.709e+03 -10.512  < 2e-16 ***
## model_key630               -3.124e+04  4.413e+03  -7.079 1.66e-12 ***
## model_key316               -2.894e+04  2.734e+03 -10.587  < 2e-16 ***
## model_key116               -3.090e+04  2.738e+03 -11.286  < 2e-16 ***
## model_key118               -3.063e+04  2.723e+03 -11.250  < 2e-16 ***
## model_keyZ4                -3.074e+04  3.015e+03 -10.195  < 2e-16 ***
## model_key123               -3.350e+04  3.674e+03  -9.116  < 2e-16 ***
## model_key650               -4.253e+04  3.597e+03 -11.824  < 2e-16 ***
## model_key523               -2.398e+04  3.208e+03  -7.473 9.26e-14 ***
## model_key216               -3.426e+04  4.670e+03  -7.336 2.57e-13 ***
## model_key735               -2.977e+04  4.427e+03  -6.725 1.96e-11 ***
## car_typesuv                -7.613e+02  1.678e+03  -0.454 0.650143
## car_typeconvertible         3.043e+03  7.249e+02   4.198 2.74e-05 ***
## car_typesedan               2.510e+02  5.572e+02   0.451 0.652354
## car_typevan                -5.747e+03  1.395e+03  -4.121 3.84e-05 ***
## car_typehatchback          -3.597e+02  6.039e+02  -0.596 0.551502
## car_typeestate             -1.764e+03  5.593e+02  -3.153 0.001624 **
## car_typesubcompact          6.398e+02  6.984e+02   0.916 0.359663
## month_sold2                -1.353e+01  2.965e+02  -0.046 0.963602
## month_sold3                -1.840e+02  2.825e+02  -0.651 0.514905
## month_sold4                 8.905e+01  2.853e+02   0.312 0.754945
## month_sold5                 2.170e+02  2.800e+02   0.775 0.438464
## month_sold6                 1.192e+02  2.902e+02   0.411 0.681137
## month_sold7                 4.491e+02  2.953e+02   1.521 0.128386
## month_sold8                 6.319e+02  2.974e+02   2.125 0.033619 *
## month_sold9                 4.769e+02  3.495e+02   1.365 0.172458
## fuelelectro                 1.017e+03  3.598e+03   0.283 0.777535
## fuelhybrid_petrol           4.405e+03  2.081e+03   2.117 0.034299 *
## fuelpetrol                 -3.998e+02  3.235e+02  -1.236 0.216481
## feature_1TRUE               5.364e+02  1.232e+02   4.354 1.37e-05 ***
## feature_3TRUE               6.895e+02  1.405e+02   4.907 9.58e-07 ***
## feature_4TRUE               1.401e+03  1.742e+02   8.042 1.11e-15 ***
```

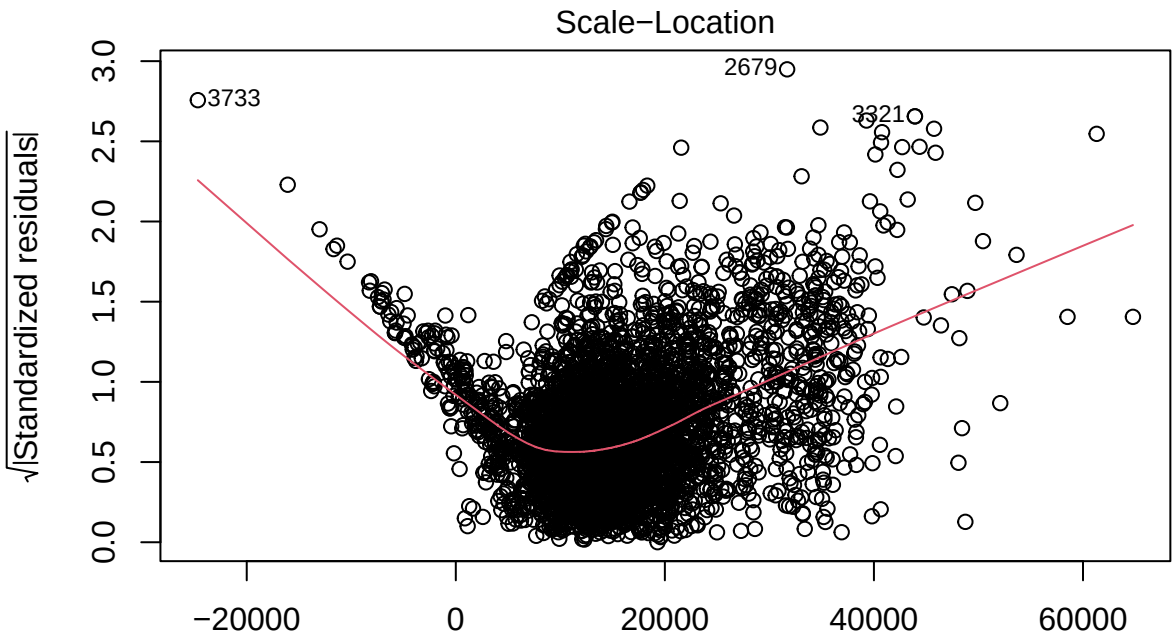
```
## feature_6TRUE               1.323e+03  1.298e+02  10.199  < 2e-16 ***
## feature_7TRUE               9.296e+02  2.267e+02   4.100 4.21e-05 ***
## feature_8TRUE               1.296e+03  1.280e+02  10.124  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3563 on 4736 degrees of freedom
## Multiple R-squared:  0.831,  Adjusted R-squared:  0.8275
## F-statistic: 232.9 on 100 and 4736 DF,  p-value: < 2.2e-16
```

```
plot(mlr_2)
```

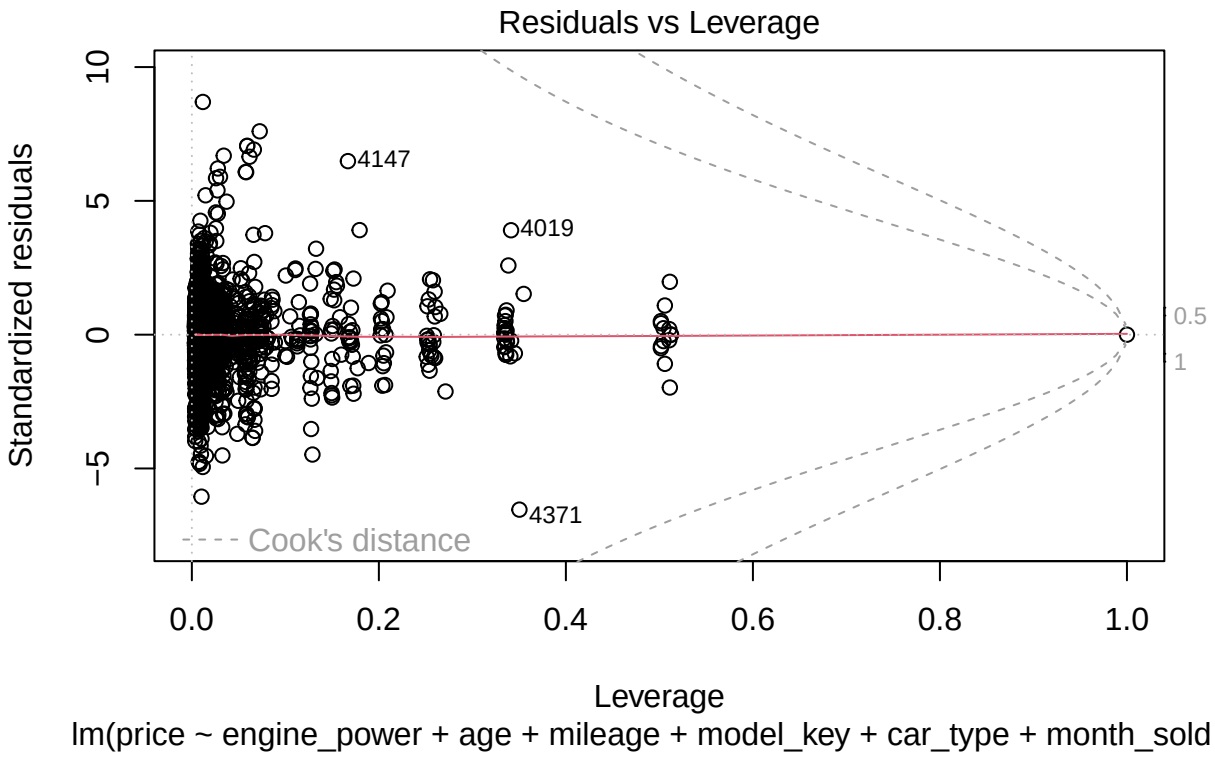
```
## Warning: not plotting observations with leverage one:
##   56, 107, 1899, 2925, 3153, 3600, 4798, 4799, 4820, 4821
```




Q−Q Residuals
Standardized residuals
10
5
0
−5
2679
3733
−4
−2
0
2
4
Theoretical Quantiles
lm(price ~ engine_power + age + mileage + model_key + car_type + month_sold ...




```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



In this new model, month sold has now fluctuated such that August appears more significant. The model is similarly weak compared to the prior model in terms of its alignment with our assumptions. Therefore we shall take a more aggressive approach, reducing our model further. This time we will take away month sold and fuel.

```
mlr_3 <- lm(formula = price ~
              engine_power + age + mileage + model_key + car_type +
              feature_1 + feature_3 + feature_4 + feature_6 + feature_7 + feature_8,
              data = bmw_data_cleaned)
summary(mlr_3)
```

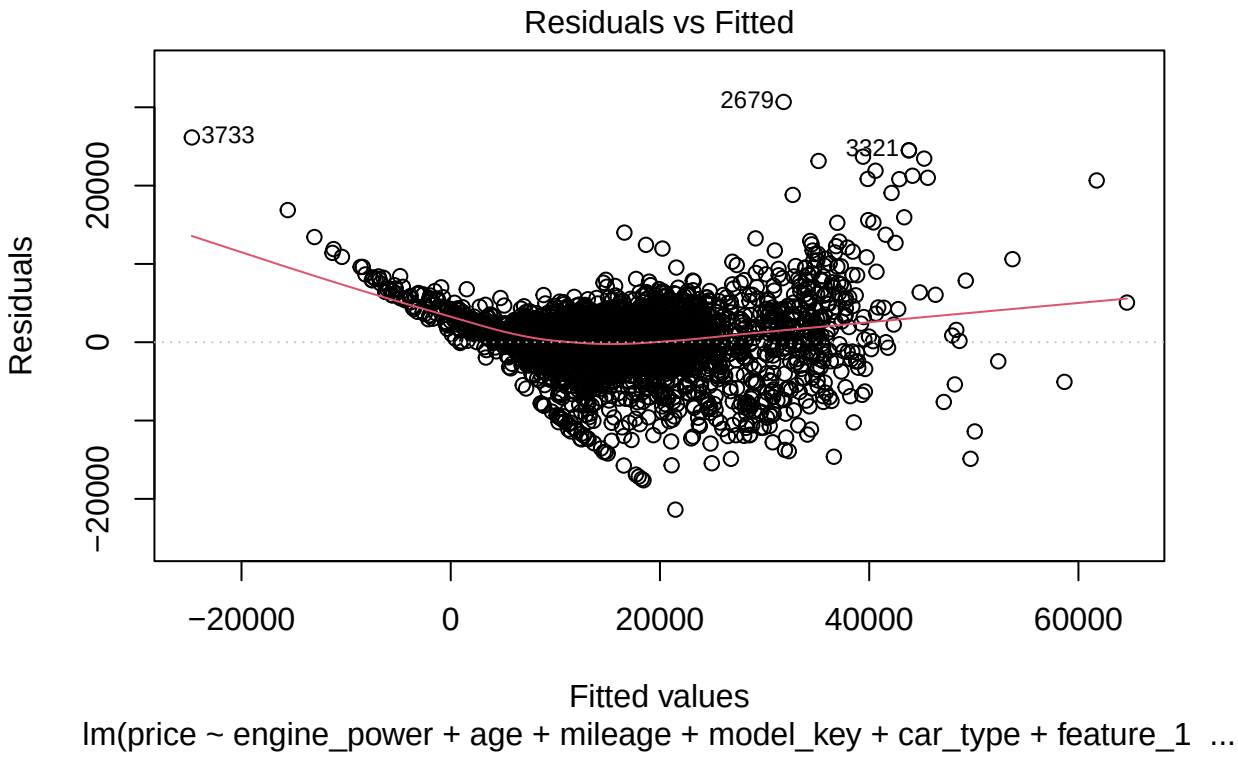
```
##
## Call:
## lm(formula = price ~ engine_power + age + mileage + model_key +
##     car_type + feature_1 + feature_3 + feature_4 + feature_6 +
##     feature_7 + feature_8, data = bmw_data_cleaned)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21373.6  -1398.9      0.4   1492.9  30678.9
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 4.349e+04  2.850e+03  15.261  < 2e-16 ***
## engine_power                5.171e+01  3.900e+00  13.259  < 2e-16 ***
## age                        -1.064e+03  2.645e+01 -40.211  < 2e-16 ***
## mileage                    -3.083e-02  1.065e-03 -28.962  < 2e-16 ***
```

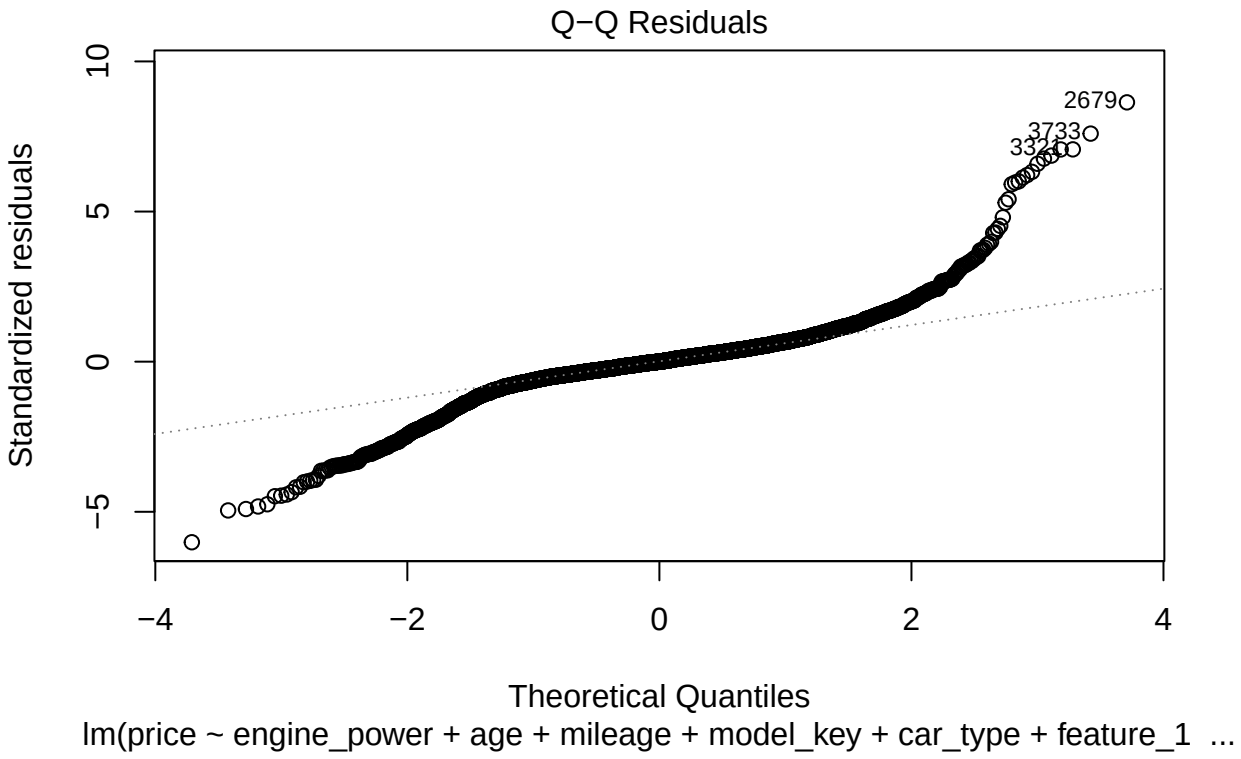
```
## model_keyX6 M              -4.279e+03  3.281e+03  -1.304 0.192229
## model_keyX5 M50            -6.003e+03  3.659e+03  -1.640 0.100990
## model_keyM5                -1.662e+04  4.424e+03  -3.756 0.000175 ***
## model_keyX5 M              -1.291e+04  3.180e+03  -4.059 5.02e-05 ***
## model_key740               -1.335e+04  2.736e+03  -4.881 1.09e-06 ***
## model_key750               -1.920e+04  3.619e+03  -5.305 1.18e-07 ***
## model_key640 Gran Coupé    -1.439e+04  2.724e+03  -5.284 1.32e-07 ***
## model_keyM3                -1.403e+04  2.876e+03  -4.879 1.10e-06 ***
## model_keyM550              -1.923e+04  2.784e+03  -6.906 5.66e-12 ***
## model_keyX6                -1.419e+04  3.091e+03  -4.591 4.53e-06 ***
## model_key640               -1.579e+04  2.868e+03  -5.507 3.84e-08 ***
## model_key435 Gran Coupé    -2.017e+04  2.972e+03  -6.787 1.29e-11 ***
## model_keyX4                -1.725e+04  3.118e+03  -5.530 3.37e-08 ***
## model_key435               -2.354e+04  3.135e+03  -7.509 7.09e-14 ***
## model_key425               -1.903e+04  3.624e+03  -5.250 1.58e-07 ***
## model_keyX5                -1.652e+04  3.063e+03  -5.393 7.26e-08 ***
## model_key430               -2.228e+04  3.604e+03  -6.183 6.83e-10 ***
## model_keyM235              -2.575e+04  3.294e+03  -7.818 6.58e-15 ***
## model_key430 Gran Coupé    -2.382e+04  3.155e+03  -7.548 5.25e-14 ***
## model_keyM135              -2.573e+04  4.418e+03  -5.824 6.11e-09 ***
## model_key330 Gran Turismo  -2.621e+04  3.340e+03  -7.845 5.29e-15 ***
## model_key535 Gran Turismo  -2.650e+04  3.622e+03  -7.316 2.98e-13 ***
## model_key335 Gran Turismo  -2.667e+04  3.331e+03  -8.006 1.48e-15 ***
## model_key420 Gran Coupé    -2.318e+04  2.747e+03  -8.438  < 2e-16 ***
## model_key420               -2.323e+04  2.688e+03  -8.640  < 2e-16 ***
## model_key220               -2.771e+04  3.339e+03  -8.301  < 2e-16 ***
## model_key730               -2.099e+04  2.706e+03  -7.756 1.06e-14 ***
## model_key535               -2.495e+04  2.655e+03  -9.397  < 2e-16 ***
## model_key135               -2.902e+04  3.125e+03  -9.287  < 2e-16 ***
## model_key335               -2.673e+04  2.905e+03  -9.202  < 2e-16 ***
## model_keyi3                -2.154e+04  3.116e+03  -6.912 5.41e-12 ***
## model_keyActiveHybrid 5    -2.919e+04  4.416e+03  -6.609 4.29e-11 ***
## model_key530 Gran Turismo  -2.469e+04  2.762e+03  -8.941  < 2e-16 ***
## model_key418 Gran Coupé    -2.411e+04  2.990e+03  -8.062 9.44e-16 ***
## model_key325 Gran Turismo  -2.376e+04  3.193e+03  -7.440 1.19e-13 ***
## model_key528               -2.604e+04  3.003e+03  -8.673  < 2e-16 ***
## model_key520 Gran Turismo  -2.350e+04  2.785e+03  -8.437  < 2e-16 ***
## model_key530               -2.530e+04  2.642e+03  -9.579  < 2e-16 ***
## model_key635               -2.451e+04  4.410e+03  -5.557 2.89e-08 ***
## model_key225 Active Tourer -2.488e+04  4.621e+03  -5.385 7.58e-08 ***
## model_key225               -3.121e+04  4.432e+03  -7.042 2.17e-12 ***
## model_keyX3                -2.547e+04  3.085e+03  -8.256  < 2e-16 ***
## model_key214 Gran Tourer   -2.098e+04  4.681e+03  -4.481 7.62e-06 ***
## model_key320 Gran Turismo  -2.709e+04  2.704e+03 -10.021  < 2e-16 ***
## model_key218               -2.744e+04  2.866e+03  -9.573  < 2e-16 ***
## model_key216 Gran Tourer   -2.295e+04  3.501e+03  -6.555 6.17e-11 ***
## model_key330               -2.878e+04  2.688e+03 -10.704  < 2e-16 ***
## model_key328               -2.881e+04  3.076e+03  -9.369  < 2e-16 ***
## model_key518               -2.518e+04  2.737e+03  -9.199  < 2e-16 ***
## model_key218 Gran Tourer   -2.502e+04  3.219e+03  -7.774 9.30e-15 ***
## model_key520               -2.542e+04  2.677e+03  -9.497  < 2e-16 ***
## model_key218 Active Tourer -2.352e+04  3.138e+03  -7.497 7.77e-14 ***
## model_key525               -2.578e+04  2.666e+03  -9.670  < 2e-16 ***
## model_key318 Gran Turismo  -2.680e+04  2.731e+03  -9.816  < 2e-16 ***
```

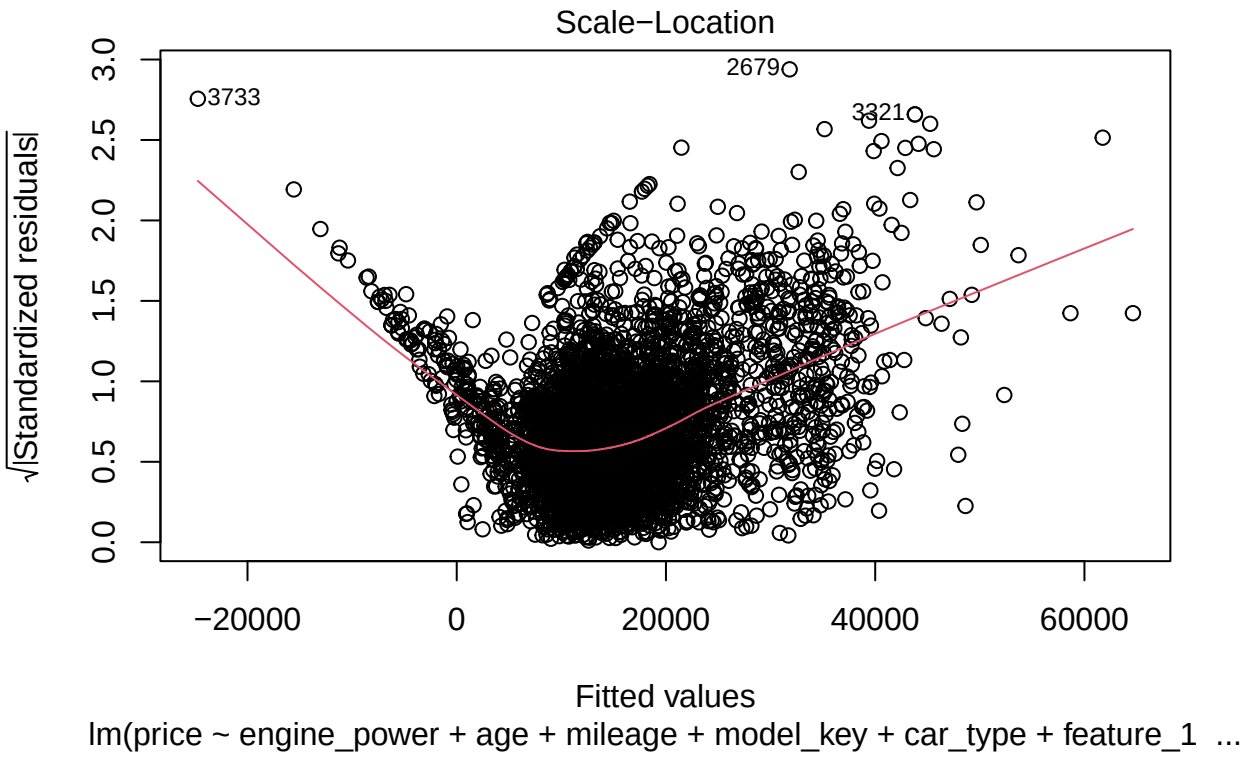
```
## model_key325               -2.963e+04  2.745e+03 -10.793  < 2e-16 ***
## model_keyX1                -2.850e+04  3.110e+03  -9.167  < 2e-16 ***
## model_key216 Active Tourer -2.411e+04  3.650e+03  -6.606 4.38e-11 ***
## model_key125               -3.263e+04  3.097e+03 -10.538  < 2e-16 ***
## model_key120               -3.006e+04  2.721e+03 -11.049  < 2e-16 ***
## model_key320               -2.832e+04  2.675e+03 -10.587  < 2e-16 ***
## model_key220 Active Tourer -2.801e+04  4.637e+03  -6.041 1.65e-09 ***
## model_key114               -3.040e+04  2.898e+03 -10.489  < 2e-16 ***
## model_key318               -2.776e+04  2.704e+03 -10.266  < 2e-16 ***
## model_key630               -3.103e+04  4.418e+03  -7.023 2.48e-12 ***
## model_key316               -2.824e+04  2.730e+03 -10.347  < 2e-16 ***
## model_key116               -3.020e+04  2.734e+03 -11.048  < 2e-16 ***
## model_key118               -2.999e+04  2.717e+03 -11.036  < 2e-16 ***
## model_keyZ4                -3.051e+04  3.017e+03 -10.112  < 2e-16 ***
## model_key123               -3.292e+04  3.670e+03  -8.968  < 2e-16 ***
## model_key650               -4.197e+04  3.599e+03 -11.661  < 2e-16 ***
## model_key523               -2.369e+04  3.212e+03  -7.376 1.92e-13 ***
## model_key216               -3.388e+04  4.674e+03  -7.249 4.87e-13 ***
## model_key735               -2.951e+04  4.433e+03  -6.658 3.09e-11 ***
## car_typesuv                -6.983e+02  1.681e+03  -0.415 0.677892
## car_typeconvertible         2.934e+03  7.251e+02   4.047 5.28e-05 ***
## car_typesedan               1.641e+02  5.571e+02   0.295 0.768308
## car_typevan                -5.767e+03  1.396e+03  -4.132 3.66e-05 ***
## car_typehatchback          -4.068e+02  6.046e+02  -0.673 0.501104
## car_typeestate             -1.859e+03  5.589e+02  -3.326 0.000888 ***
## car_typesubcompact          4.394e+02  6.966e+02   0.631 0.528217
## feature_1TRUE               5.758e+02  1.230e+02   4.681 2.93e-06 ***
## feature_3TRUE               6.887e+02  1.405e+02   4.901 9.85e-07 ***
## feature_4TRUE               1.375e+03  1.740e+02   7.906 3.29e-15 ***
## feature_6TRUE               1.310e+03  1.298e+02  10.099  < 2e-16 ***
## feature_7TRUE               1.015e+03  2.263e+02   4.483 7.52e-06 ***
## feature_8TRUE               1.298e+03  1.281e+02  10.133  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3570 on 4747 degrees of freedom
## Multiple R-squared:   0.83,  Adjusted R-squared:  0.8268
## F-statistic: 260.4 on 89 and 4747 DF,  p-value: < 2.2e-16
```

```
plot(mlr_3)
```

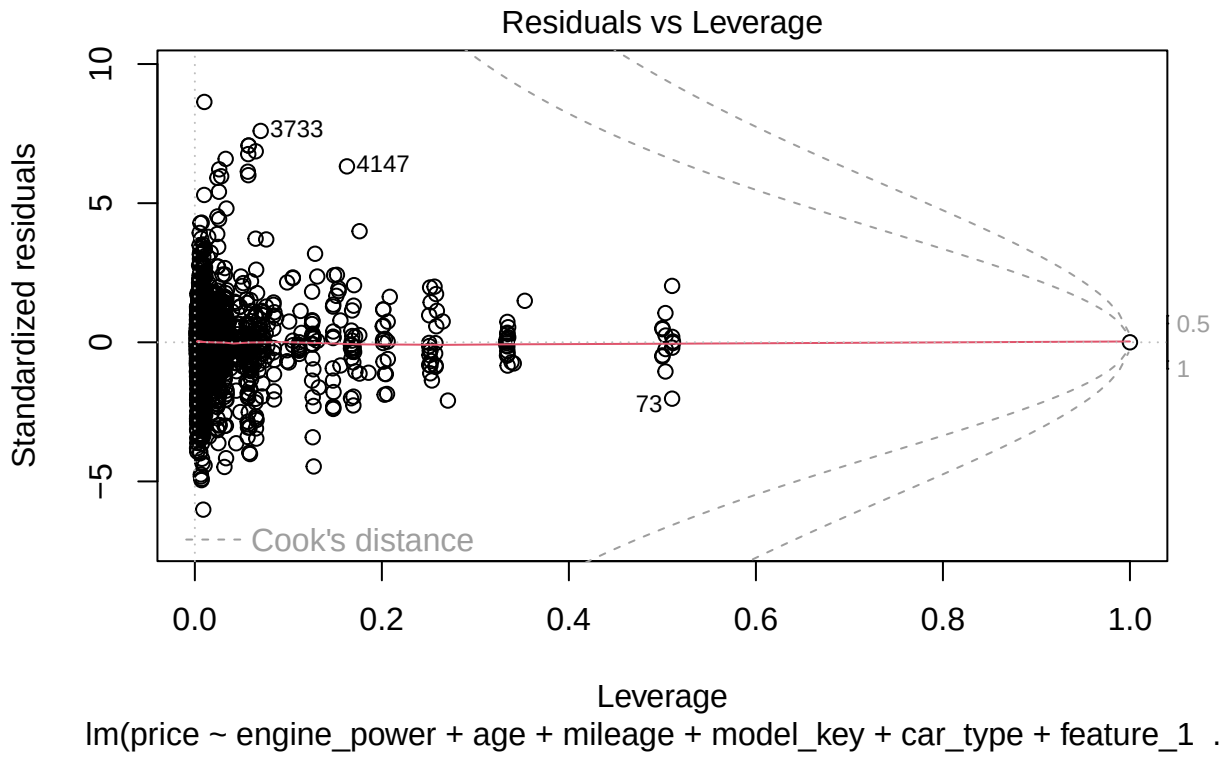
```
## Warning: not plotting observations with leverage one:
##   56, 107, 1899, 2925, 3153, 3600, 4798, 4799, 4820, 4821
```


Residuals vs Fitted
3733
2679
3321
Residuals
20000
0
−20000
−20000
0
20000
40000
60000
Fitted values
lm(price ~ engine_power + age + mileage + model_key + car_type + feature_1 ...


Q−Q Residuals
Standardized residuals
10
5
0
−5
2679
3733
−4
−2
0
2
4
Theoretical Quantiles
lm(price ~ engine_power + age + mileage + model_key + car_type + feature_1 ...




```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



Our adjusted R^2 has remained roughly the same despite the removals, and our F-Statistic has continued to increase, from 210 to 260. We now keep only the three most statistically signficant features (4, 6, & 8).

```
mlr_4 <-  lm(formula = price ~
              engine_power + age + mileage + model_key + car_type + feature_4 + feature_6 + feature_8,
              data = bmw_data_cleaned)
summary(mlr_4)
```

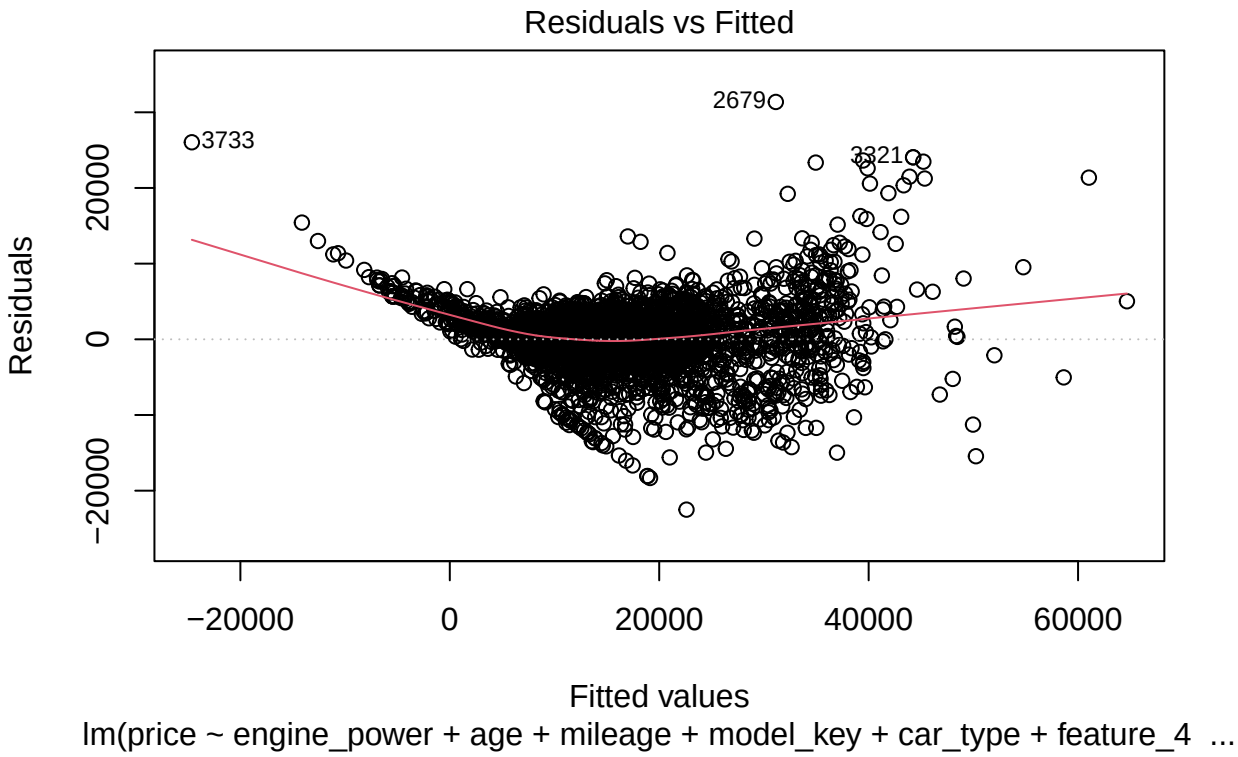
```
##
## Call:
## lm(formula = price ~ engine_power + age + mileage + model_key +
##     car_type + feature_4 + feature_6 + feature_8, data = bmw_data_cleaned)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22503.8  -1371.8     52.1   1530.2  31364.4
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 4.485e+04  2.858e+03  15.691  < 2e-16 ***
## engine_power                5.154e+01  3.916e+00  13.164  < 2e-16 ***
## age                        -1.050e+03  2.658e+01 -39.491  < 2e-16 ***
## mileage                    -3.033e-02  1.065e-03 -28.472  < 2e-16 ***
## model_keyX6 M              -3.789e+03  3.309e+03  -1.145 0.252220
## model_keyX5 M50            -5.658e+03  3.691e+03  -1.533 0.125389
## model_keyM5                -1.584e+04  4.462e+03  -3.550 0.000388 ***
## model_keyX5 M              -1.239e+04  3.207e+03  -3.863 0.000114 ***
```

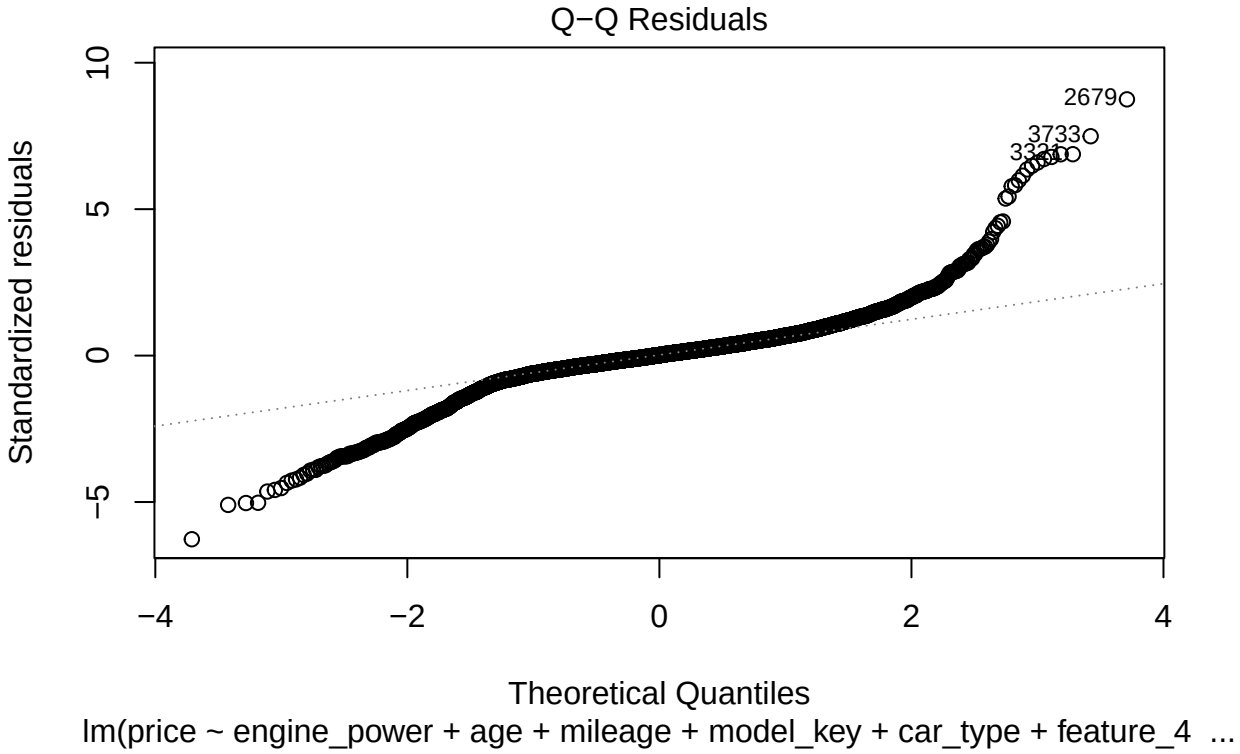
```
## model_key740               -1.278e+04  2.757e+03  -4.636 3.64e-06 ***
## model_key750               -1.844e+04  3.648e+03  -5.056 4.45e-07 ***
## model_key640 Gran Coupé    -1.387e+04  2.746e+03  -5.051 4.57e-07 ***
## model_keyM3                -1.408e+04  2.902e+03  -4.853 1.26e-06 ***
## model_keyM550              -1.905e+04  2.808e+03  -6.785 1.30e-11 ***
## model_keyX6                -1.374e+04  3.117e+03  -4.408 1.07e-05 ***
## model_key640               -1.527e+04  2.891e+03  -5.284 1.32e-07 ***
## model_key435 Gran Coupé    -2.015e+04  2.997e+03  -6.722 2.00e-11 ***
## model_keyX4                -1.716e+04  3.146e+03  -5.454 5.17e-08 ***
## model_key435               -2.354e+04  3.159e+03  -7.451 1.09e-13 ***
## model_key425               -1.978e+04  3.652e+03  -5.415 6.42e-08 ***
## model_keyX5                -1.623e+04  3.089e+03  -5.253 1.56e-07 ***
## model_key430               -2.264e+04  3.635e+03  -6.228 5.14e-10 ***
## model_keyM235              -2.547e+04  3.322e+03  -7.666 2.13e-14 ***
## model_key430 Gran Coupé    -2.396e+04  3.183e+03  -7.528 6.15e-14 ***
## model_keyM135              -2.550e+04  4.457e+03  -5.722 1.12e-08 ***
## model_key330 Gran Turismo  -2.587e+04  3.370e+03  -7.676 1.97e-14 ***
## model_key535 Gran Turismo  -2.599e+04  3.653e+03  -7.114 1.29e-12 ***
## model_key335 Gran Turismo  -2.750e+04  3.358e+03  -8.190 3.34e-16 ***
## model_key420 Gran Coupé    -2.308e+04  2.769e+03  -8.335  < 2e-16 ***
## model_key420               -2.349e+04  2.710e+03  -8.666  < 2e-16 ***
## model_key220               -2.838e+04  3.362e+03  -8.443  < 2e-16 ***
## model_key730               -2.040e+04  2.727e+03  -7.479 8.87e-14 ***
## model_key535               -2.466e+04  2.677e+03  -9.211  < 2e-16 ***
## model_key135               -2.895e+04  3.151e+03  -9.187  < 2e-16 ***
## model_key335               -2.651e+04  2.929e+03  -9.052  < 2e-16 ***
## model_keyi3                -2.152e+04  3.142e+03  -6.849 8.37e-12 ***
## model_keyActiveHybrid 5    -2.910e+04  4.455e+03  -6.531 7.23e-11 ***
## model_key530 Gran Turismo  -2.443e+04  2.784e+03  -8.777  < 2e-16 ***
## model_key418 Gran Coupé    -2.394e+04  3.016e+03  -7.936 2.58e-15 ***
## model_key325 Gran Turismo  -2.398e+04  3.220e+03  -7.449 1.11e-13 ***
## model_key528               -2.607e+04  3.028e+03  -8.609  < 2e-16 ***
## model_key520 Gran Turismo  -2.305e+04  2.808e+03  -8.209 2.84e-16 ***
## model_key530               -2.511e+04  2.664e+03  -9.426  < 2e-16 ***
## model_key635               -2.373e+04  4.447e+03  -5.337 9.88e-08 ***
## model_key225 Active Tourer -2.510e+04  4.659e+03  -5.388 7.46e-08 ***
## model_key225               -3.178e+04  4.469e+03  -7.112 1.32e-12 ***
## model_keyX3                -2.552e+04  3.111e+03  -8.204 2.96e-16 ***
## model_key214 Gran Tourer   -2.022e+04  4.722e+03  -4.282 1.89e-05 ***
## model_key320 Gran Turismo  -2.723e+04  2.726e+03  -9.991  < 2e-16 ***
## model_key218               -2.788e+04  2.887e+03  -9.654  < 2e-16 ***
## model_key216 Gran Tourer   -2.257e+04  3.530e+03  -6.394 1.77e-10 ***
## model_key330               -2.878e+04  2.711e+03 -10.619  < 2e-16 ***
## model_key328               -2.890e+04  3.101e+03  -9.321  < 2e-16 ***
## model_key518               -2.493e+04  2.759e+03  -9.035  < 2e-16 ***
## model_key218 Gran Tourer   -2.463e+04  3.243e+03  -7.595 3.67e-14 ***
## model_key520               -2.526e+04  2.699e+03  -9.359  < 2e-16 ***
## model_key218 Active Tourer -2.323e+04  3.163e+03  -7.344 2.42e-13 ***
## model_key525               -2.563e+04  2.688e+03  -9.534  < 2e-16 ***
## model_key318 Gran Turismo  -2.698e+04  2.753e+03  -9.801  < 2e-16 ***
## model_key325               -2.968e+04  2.769e+03 -10.720  < 2e-16 ***
## model_keyX1                -2.870e+04  3.135e+03  -9.153  < 2e-16 ***
## model_key216 Active Tourer -2.355e+04  3.679e+03  -6.401 1.69e-10 ***
## model_key125               -3.296e+04  3.121e+03 -10.562  < 2e-16 ***
```

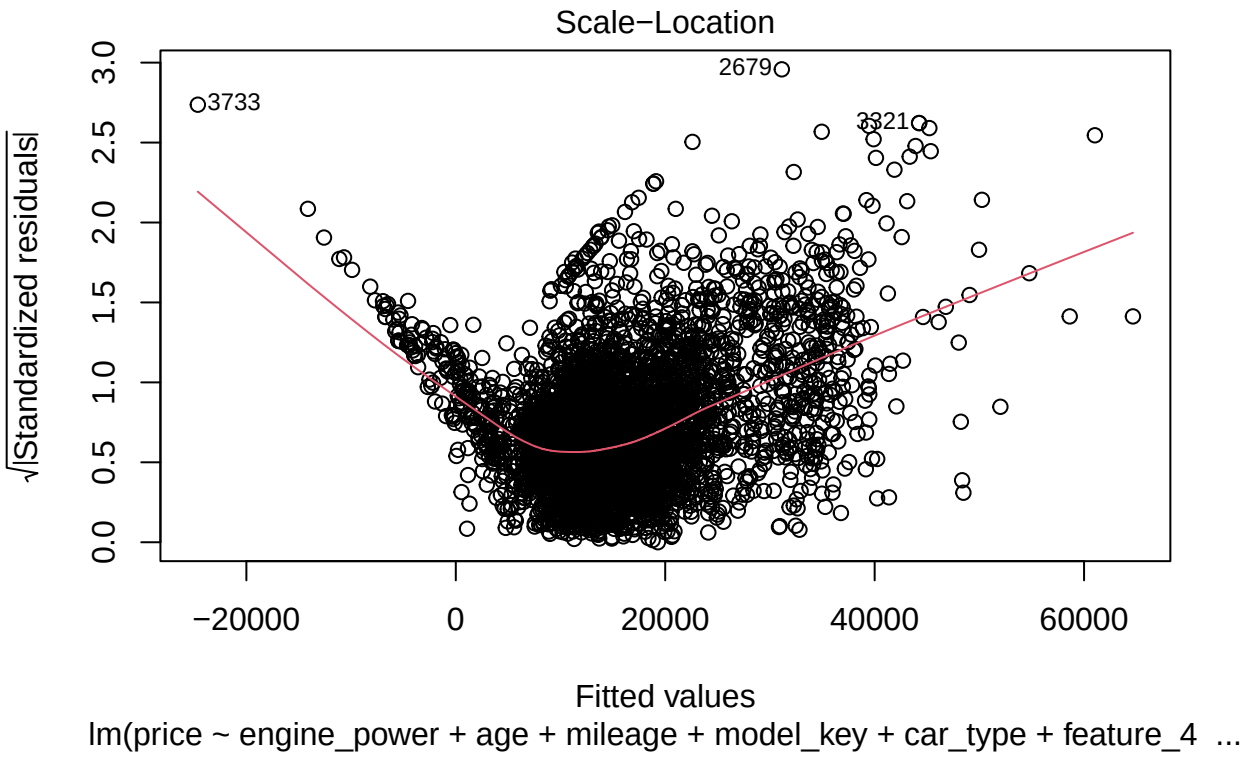
```
## model_key120               -3.046e+04  2.742e+03 -11.108  < 2e-16 ***
## model_key320               -2.849e+04  2.696e+03 -10.564  < 2e-16 ***
## model_key220 Active Tourer -2.785e+04  4.676e+03  -5.957 2.75e-09 ***
## model_key114               -3.082e+04  2.920e+03 -10.554  < 2e-16 ***
## model_key318               -2.791e+04  2.726e+03 -10.237  < 2e-16 ***
## model_key630               -3.048e+04  4.454e+03  -6.843 8.72e-12 ***
## model_key316               -2.848e+04  2.751e+03 -10.352  < 2e-16 ***
## model_key116               -3.056e+04  2.755e+03 -11.090  < 2e-16 ***
## model_key118               -3.035e+04  2.738e+03 -11.083  < 2e-16 ***
## model_keyZ4                -3.071e+04  3.043e+03 -10.094  < 2e-16 ***
## model_key123               -3.269e+04  3.702e+03  -8.831  < 2e-16 ***
## model_key650               -4.196e+04  3.631e+03 -11.557  < 2e-16 ***
## model_key523               -2.399e+04  3.239e+03  -7.407 1.52e-13 ***
## model_key216               -3.470e+04  4.711e+03  -7.366 2.06e-13 ***
## model_key735               -2.942e+04  4.472e+03  -6.578 5.29e-11 ***
## car_typesuv                -9.126e+02  1.695e+03  -0.539 0.590237
## car_typeconvertible         3.046e+03  7.302e+02   4.171 3.09e-05 ***
## car_typesedan               3.986e+01  5.618e+02   0.071 0.943436
## car_typevan                -6.335e+03  1.404e+03  -4.512 6.59e-06 ***
## car_typehatchback          -4.386e+02  6.097e+02  -0.719 0.471994
## car_typeestate             -1.969e+03  5.635e+02  -3.494 0.000480 ***
## car_typesubcompact          4.764e+02  7.025e+02   0.678 0.497688
## feature_4TRUE               1.638e+03  1.718e+02   9.534  < 2e-16 ***
## feature_6TRUE               1.509e+03  1.291e+02  11.693  < 2e-16 ***
## feature_8TRUE               1.275e+03  1.285e+02   9.928  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3602 on 4750 degrees of freedom
## Multiple R-squared:  0.8269, Adjusted R-squared:  0.8237
## F-statistic: 263.8 on 86 and 4750 DF,  p-value: < 2.2e-16
```

```
plot(mlr_4)
```

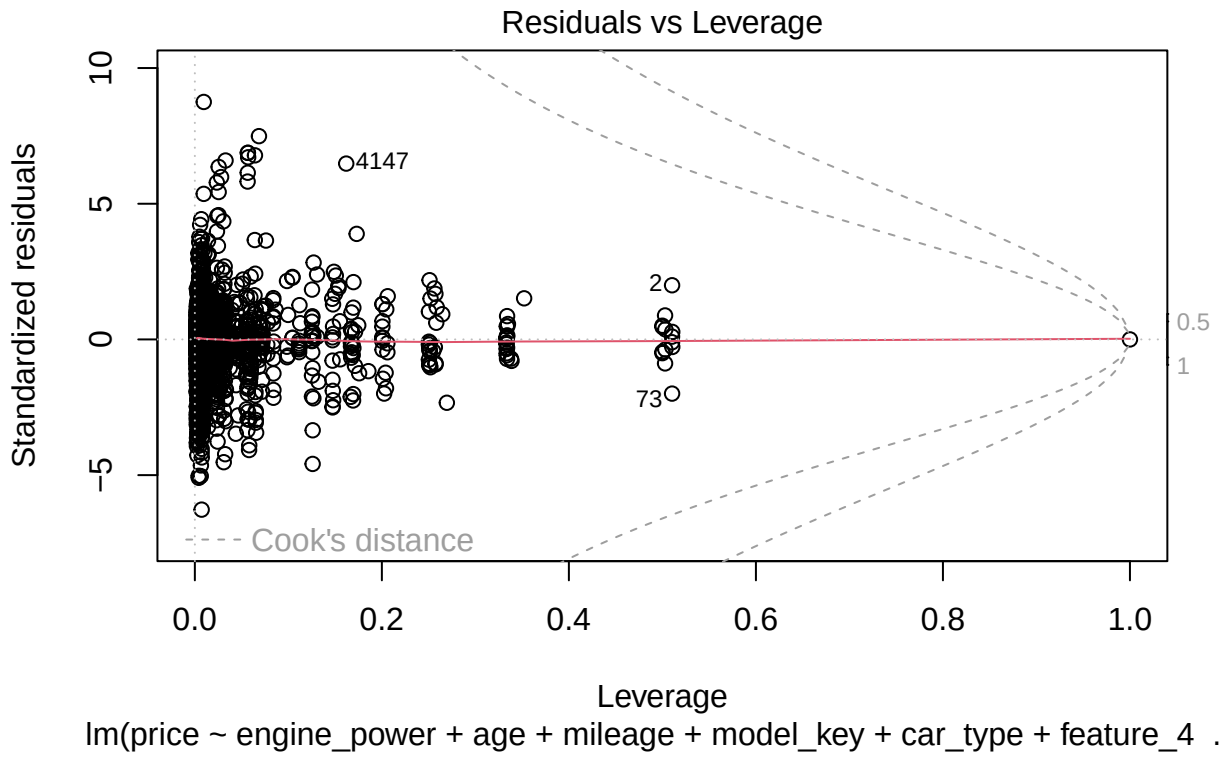
```
## Warning: not plotting observations with leverage one:
##   56, 107, 1899, 2925, 3153, 3600, 4798, 4799, 4820, 4821
```


Residuals vs Fitted
2679
3733
3321
Residuals
20000
0
−20000
−20000
0
20000
40000
60000
Fitted values
lm(price ~ engine_power + age + mileage + model_key + car_type + feature_4 ...


Q−Q Residuals
Standardized residuals
10
5
0
−5
2679
3733
−4
−2
0
2
4
Theoretical Quantiles
lm(price ~ engine_power + age + mileage + model_key + car_type + feature_4 ...




```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



Our model has become simpler without much change at all in its key statistics or adherence to our assumptions. Let's try to further simplify the model.

```
mlr_5 <-  lm(formula = price ~
              engine_power + age + mileage + model_key + car_type,
              data = bmw_data_cleaned)
summary(mlr_5)
```

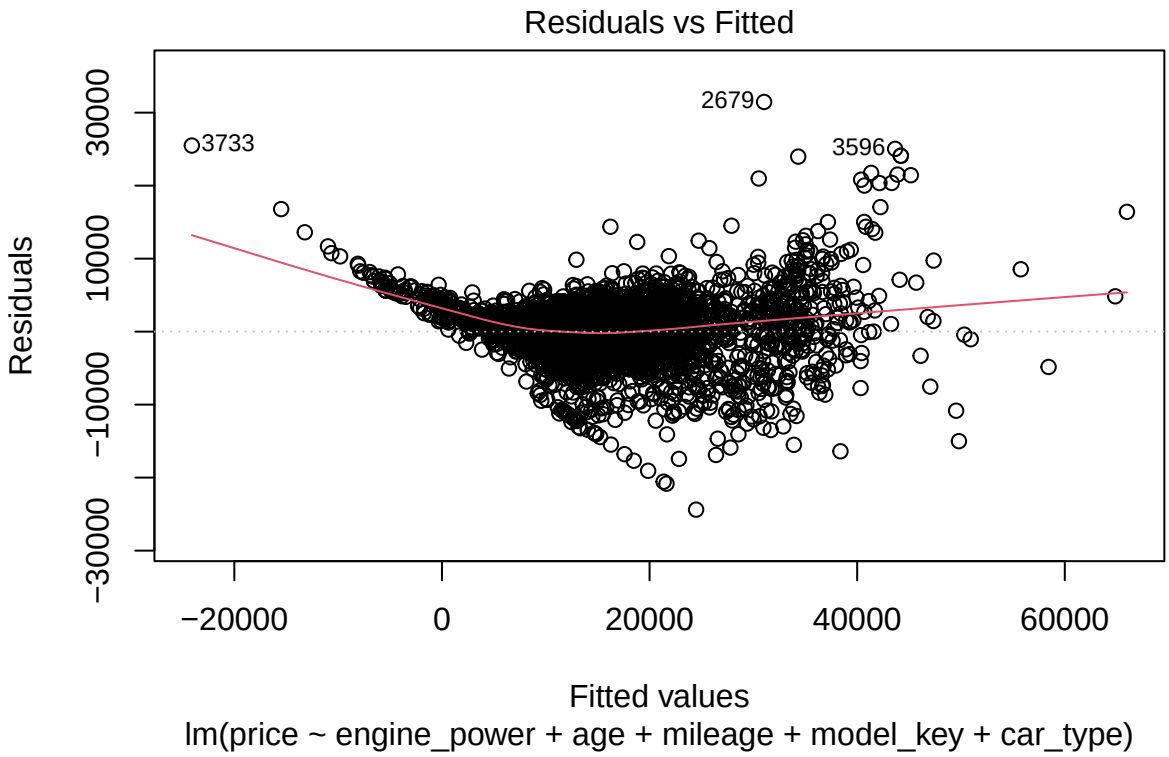
```
##
## Call:
## lm(formula = price ~ engine_power + age + mileage + model_key +
##     car_type, data = bmw_data_cleaned)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24385.0  -1510.2     32.5   1765.8  31463.6
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 4.284e+04  2.961e+03  14.469  < 2e-16 ***
## engine_power                6.651e+01  3.977e+00  16.722  < 2e-16 ***
## age                        -1.113e+03  2.713e+01 -41.033  < 2e-16 ***
## mileage                    -2.962e-02  1.104e-03 -26.831  < 2e-16 ***
## model_keyX6 M              -2.908e+03  3.429e+03  -0.848 0.396396
## model_keyX5 M50            -5.444e+03  3.827e+03  -1.422 0.154959
## model_keyM5                -1.676e+04  4.628e+03  -3.622 0.000295 ***
## model_keyX5 M              -1.109e+04  3.323e+03  -3.338 0.000850 ***
```

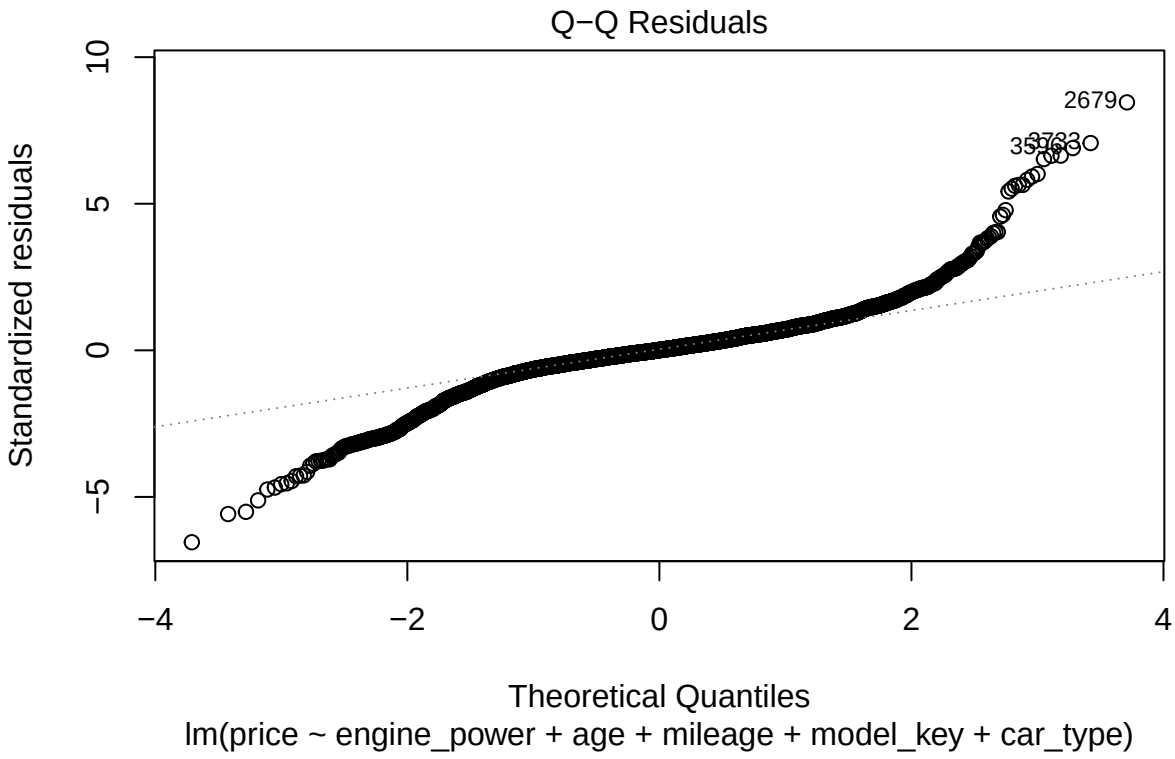
```
## model_key740               -1.118e+04  2.855e+03  -3.915 9.15e-05 ***
## model_key750               -1.745e+04  3.776e+03  -4.621 3.91e-06 ***
## model_key640 Gran Coupé    -1.252e+04  2.845e+03  -4.402 1.10e-05 ***
## model_keyM3                -1.419e+04  3.009e+03  -4.714 2.50e-06 ***
## model_keyM550              -1.679e+04  2.907e+03  -5.778 8.06e-09 ***
## model_keyX6                -1.269e+04  3.228e+03  -3.931 8.57e-05 ***
## model_key640               -1.399e+04  2.997e+03  -4.669 3.11e-06 ***
## model_key435 Gran Coupé    -1.828e+04  3.106e+03  -5.885 4.26e-09 ***
## model_keyX4                -1.597e+04  3.258e+03  -4.902 9.78e-07 ***
## model_key435               -2.125e+04  3.274e+03  -6.492 9.36e-11 ***
## model_key425               -1.815e+04  3.787e+03  -4.793 1.69e-06 ***
## model_keyX5                -1.493e+04  3.198e+03  -4.667 3.14e-06 ***
## model_key430               -2.068e+04  3.769e+03  -5.487 4.29e-08 ***
## model_keyM235              -2.452e+04  3.445e+03  -7.118 1.26e-12 ***
## model_key430 Gran Coupé    -2.090e+04  3.295e+03  -6.342 2.47e-10 ***
## model_keyM135              -2.537e+04  4.621e+03  -5.490 4.23e-08 ***
## model_key330 Gran Turismo  -2.401e+04  3.493e+03  -6.873 7.10e-12 ***
## model_key535 Gran Turismo  -2.494e+04  3.788e+03  -6.585 5.03e-11 ***
## model_key335 Gran Turismo  -2.411e+04  3.475e+03  -6.937 4.55e-12 ***
## model_key420 Gran Coupé    -2.064e+04  2.868e+03  -7.197 7.10e-13 ***
## model_key420               -2.150e+04  2.809e+03  -7.657 2.30e-14 ***
## model_key220               -2.668e+04  3.485e+03  -7.655 2.33e-14 ***
## model_key730               -1.876e+04  2.823e+03  -6.646 3.36e-11 ***
## model_key535               -2.265e+04  2.772e+03  -8.172 3.84e-16 ***
## model_key135               -2.687e+04  3.265e+03  -8.227 2.45e-16 ***
## model_key335               -2.450e+04  3.036e+03  -8.072 8.69e-16 ***
## model_keyi3                -1.898e+04  3.254e+03  -5.834 5.78e-09 ***
## model_keyActiveHybrid 5    -2.722e+04  4.620e+03  -5.892 4.07e-09 ***
## model_key530 Gran Turismo  -2.218e+04  2.881e+03  -7.698 1.66e-14 ***
## model_key418 Gran Coupé    -2.143e+04  3.124e+03  -6.860 7.74e-12 ***
## model_key325 Gran Turismo  -2.220e+04  3.338e+03  -6.652 3.22e-11 ***
## model_key528               -2.464e+04  3.138e+03  -7.852 5.02e-15 ***
## model_key520 Gran Turismo  -2.108e+04  2.908e+03  -7.249 4.86e-13 ***
## model_key530               -2.312e+04  2.758e+03  -8.381  < 2e-16 ***
## model_key635               -2.319e+04  4.610e+03  -5.031 5.06e-07 ***
## model_key225 Active Tourer -2.441e+04  4.831e+03  -5.053 4.51e-07 ***
## model_key225               -2.924e+04  4.633e+03  -6.310 3.04e-10 ***
## model_keyX3                -2.414e+04  3.222e+03  -7.491 8.10e-14 ***
## model_key214 Gran Tourer   -1.930e+04  4.896e+03  -3.942 8.21e-05 ***
## model_key320 Gran Turismo  -2.486e+04  2.823e+03  -8.809  < 2e-16 ***
## model_key218               -2.658e+04  2.993e+03  -8.880  < 2e-16 ***
## model_key216 Gran Tourer   -2.118e+04  3.660e+03  -5.785 7.70e-09 ***
## model_key330               -2.733e+04  2.810e+03  -9.726  < 2e-16 ***
## model_key328               -2.765e+04  3.215e+03  -8.603  < 2e-16 ***
## model_key518               -2.322e+04  2.859e+03  -8.122 5.78e-16 ***
## model_key218 Gran Tourer   -2.356e+04  3.362e+03  -7.008 2.76e-12 ***
## model_key520               -2.360e+04  2.796e+03  -8.441  < 2e-16 ***
## model_key218 Active Tourer -2.214e+04  3.279e+03  -6.754 1.61e-11 ***
## model_key525               -2.353e+04  2.784e+03  -8.452  < 2e-16 ***
## model_key318 Gran Turismo  -2.518e+04  2.853e+03  -8.827  < 2e-16 ***
## model_key325               -2.850e+04  2.870e+03  -9.927  < 2e-16 ***
## model_keyX1                -2.823e+04  3.250e+03  -8.686  < 2e-16 ***
## model_key216 Active Tourer -2.291e+04  3.814e+03  -6.005 2.06e-09 ***
## model_key125               -3.142e+04  3.236e+03  -9.711  < 2e-16 ***
```

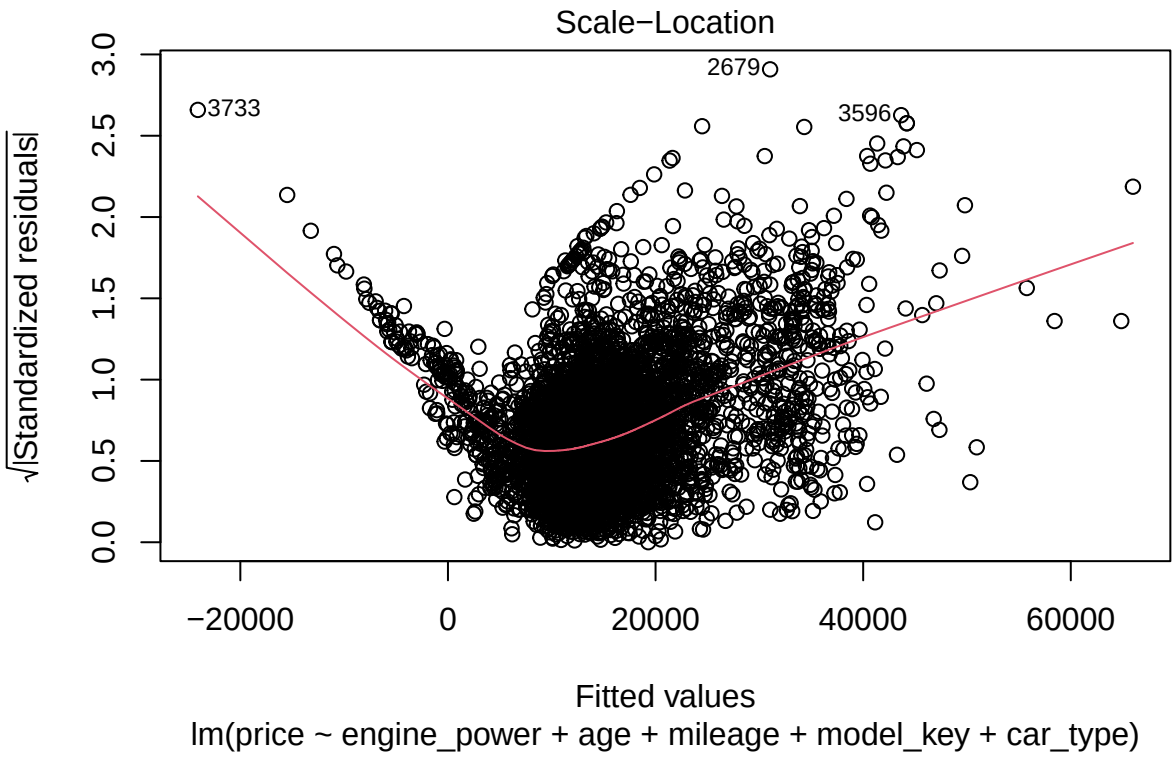
```
## model_key120               -2.886e+04  2.841e+03 -10.159  < 2e-16 ***
## model_key320               -2.686e+04  2.794e+03  -9.614  < 2e-16 ***
## model_key220 Active Tourer -2.682e+04  4.848e+03  -5.533 3.32e-08 ***
## model_key114               -2.897e+04  3.026e+03  -9.572  < 2e-16 ***
## model_key318               -2.635e+04  2.825e+03  -9.327  < 2e-16 ***
## model_key630               -2.792e+04  4.618e+03  -6.047 1.59e-09 ***
## model_key316               -2.700e+04  2.851e+03  -9.470  < 2e-16 ***
## model_key116               -2.898e+04  2.855e+03 -10.149  < 2e-16 ***
## model_key118               -2.864e+04  2.838e+03 -10.091  < 2e-16 ***
## model_keyZ4                -2.931e+04  3.154e+03  -9.294  < 2e-16 ***
## model_key123               -2.997e+04  3.837e+03  -7.811 6.93e-15 ***
## model_key650               -4.368e+04  3.763e+03 -11.609  < 2e-16 ***
## model_key523               -2.224e+04  3.355e+03  -6.627 3.80e-11 ***
## model_key216               -3.406e+04  4.885e+03  -6.972 3.56e-12 ***
## model_key735               -2.809e+04  4.636e+03  -6.059 1.47e-09 ***
## car_typesuv                 2.352e+02  1.756e+03   0.134 0.893431
## car_typeconvertible         3.389e+03  7.563e+02   4.481 7.59e-06 ***
## car_typesedan              -1.037e+02  5.825e+02  -0.178 0.858757
## car_typevan                -6.120e+03  1.456e+03  -4.202 2.69e-05 ***
## car_typehatchback          -6.942e+02  6.321e+02  -1.098 0.272184
## car_typeestate             -2.054e+03  5.845e+02  -3.514 0.000445 ***
## car_typesubcompact          9.804e+01  7.283e+02   0.135 0.892915
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3736 on 4753 degrees of freedom
## Multiple R-squared:  0.8136, Adjusted R-squared:  0.8104
## F-statistic:   250 on 83 and 4753 DF,  p-value: < 2.2e-16
```

```
plot(mlr_5)
```

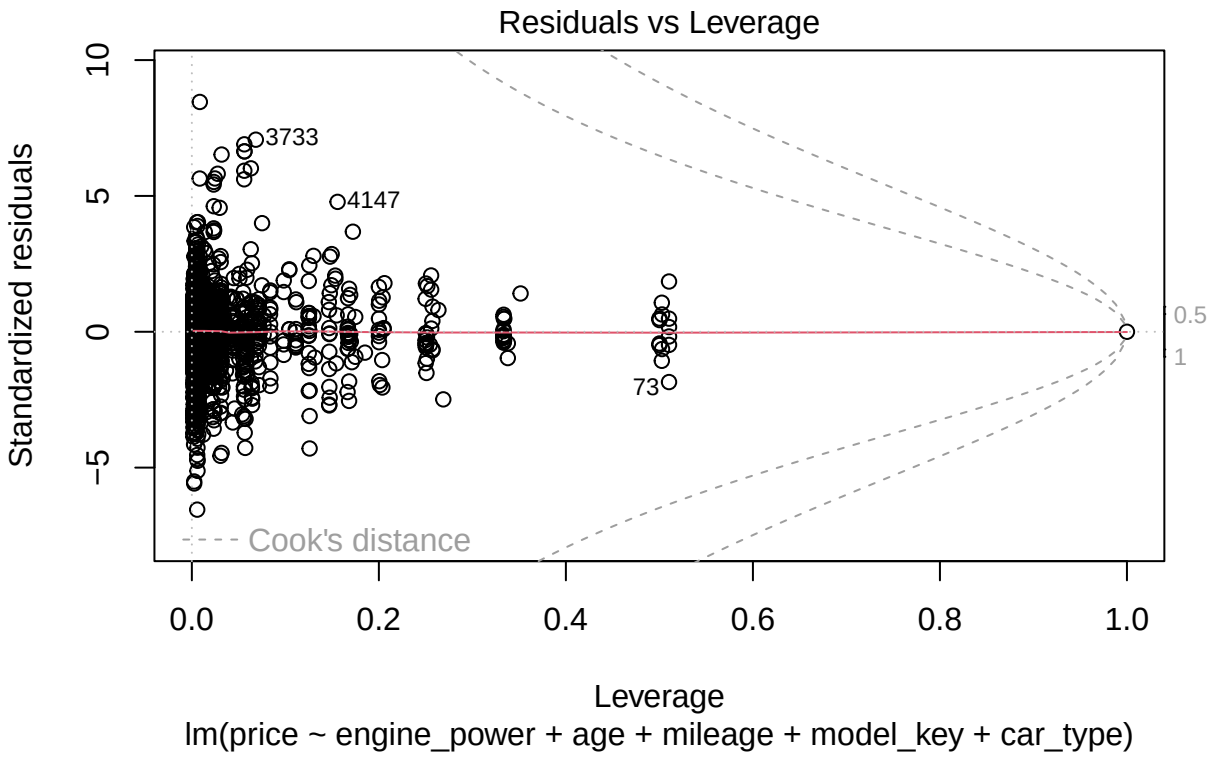
```
## Warning: not plotting observations with leverage one:
##   56, 107, 1899, 2925, 3153, 3600, 4798, 4799, 4820, 4821
```


Residuals vs Fitted
Residuals
Fitted values
lm(price ~ engine_power + age + mileage + model_key + car_type)
3733
2679
3596
30000
10000
−10000
−30000
−20000
0
20000
40000
60000


Q−Q Residuals
Standardized residuals
10
5
0
−5
2679
−4
−2
0
2
4
Theoretical Quantiles
lm(price ~ engine_power + age + mileage + model_key + car_type)




```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



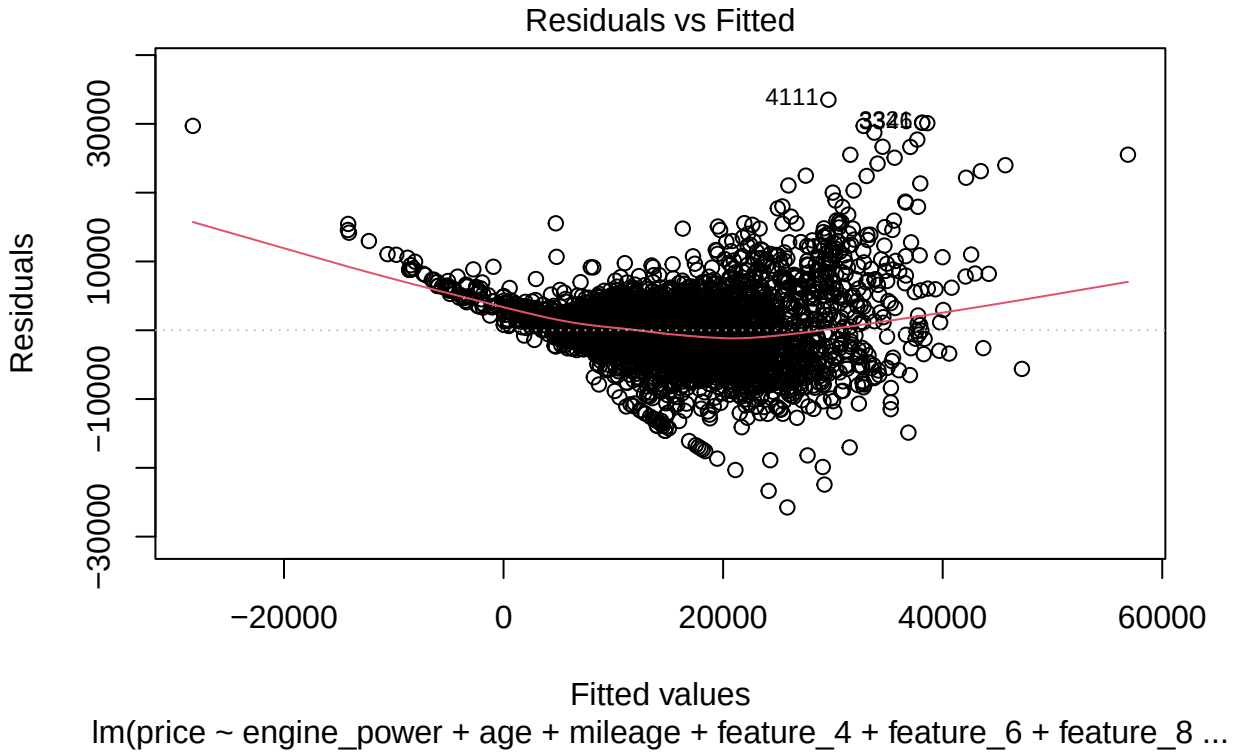
Upon removal of the features, we see a worsening of the model across the board, and yet the understandability of the model is not markedly increased. We shall now try removing the model key.

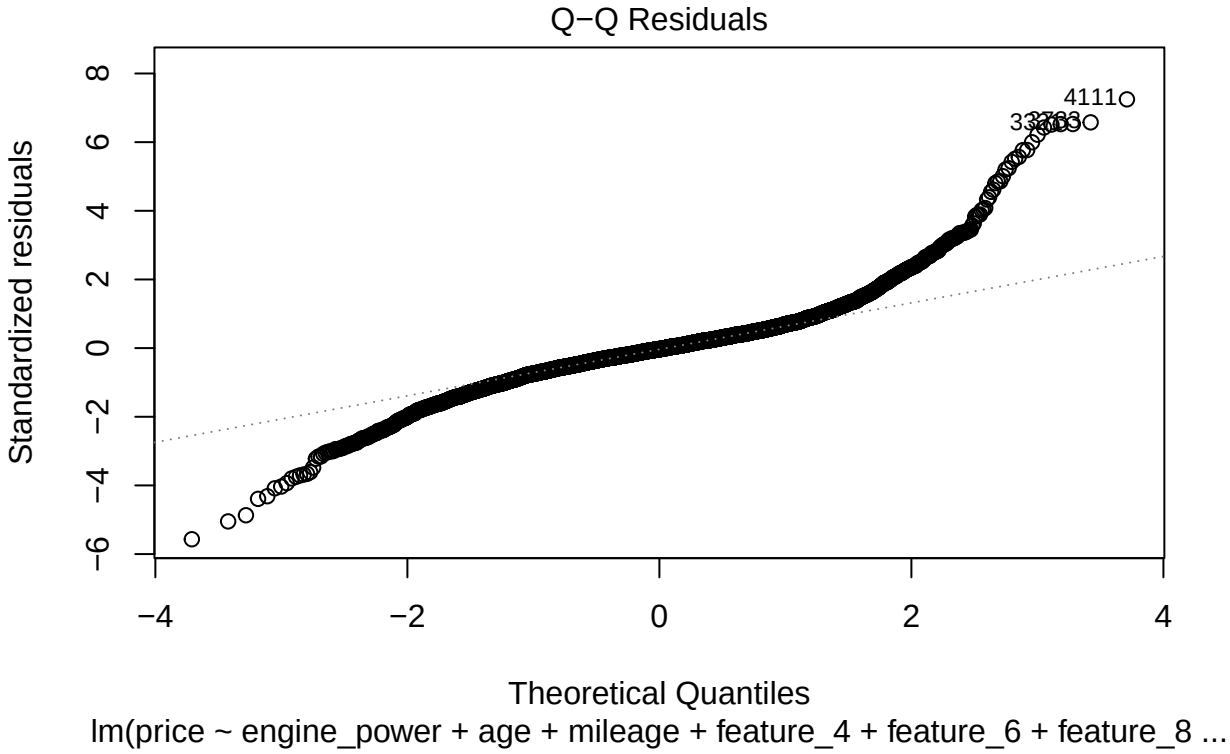
```
mlr_6 <-  lm(formula = price ~
              engine_power + age + mileage + feature_4 + feature_6 + feature_8,
              data = bmw_data_cleaned)
summary(mlr_6)
```

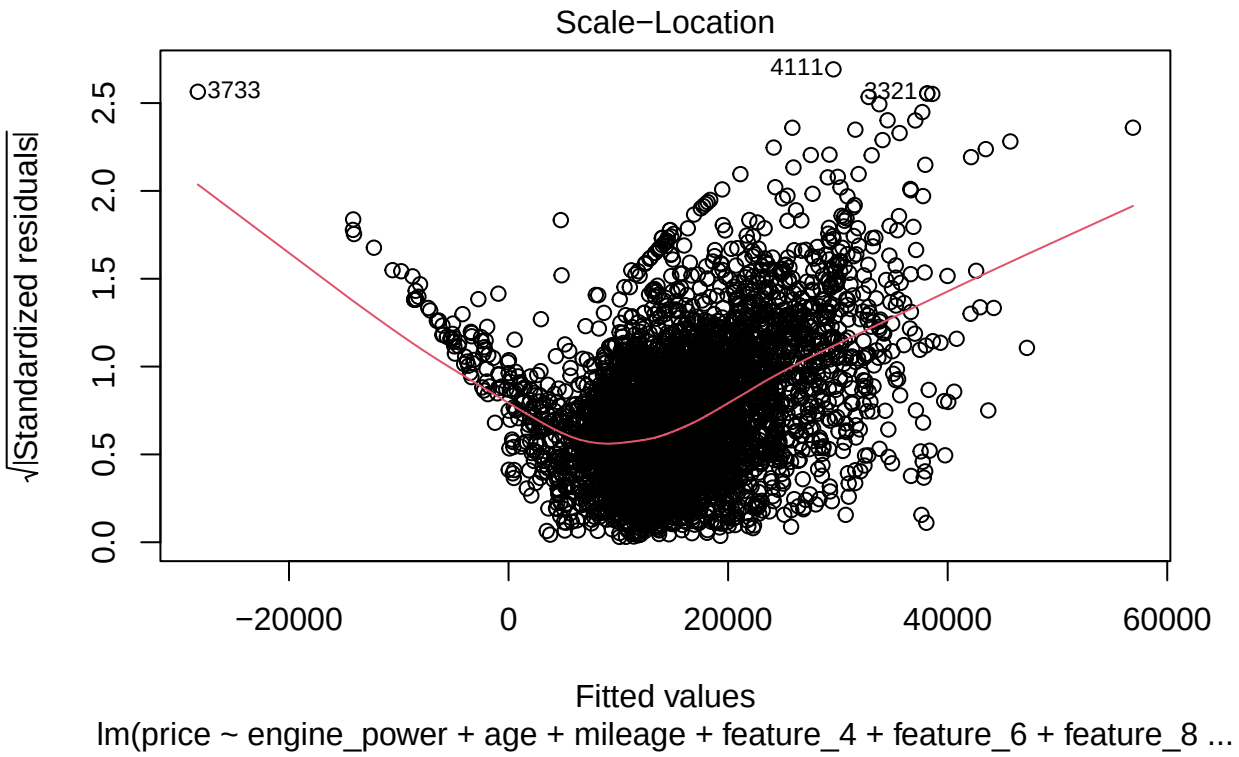
```
##
## Call:
## lm(formula = price ~ engine_power + age + mileage + feature_4 +
##     feature_6 + feature_8, data = bmw_data_cleaned)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -25752  -2277   -138   1945  33508
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.283e+03  3.062e+02  30.315  < 2e-16 ***
## engine_power   1.126e+02  2.142e+00  52.563  < 2e-16 ***
## age           -9.128e+02  3.100e+01 -29.445  < 2e-16 ***
## mileage       -3.561e-02  1.288e-03 -27.657  < 2e-16 ***
## feature_4TRUE  3.241e+03  1.875e+02  17.291  < 2e-16 ***
## feature_6TRUE  1.008e+03  1.604e+02   6.287 3.53e-10 ***
## feature_8TRUE  1.889e+03  1.559e+02  12.113  < 2e-16 ***
## ---
```

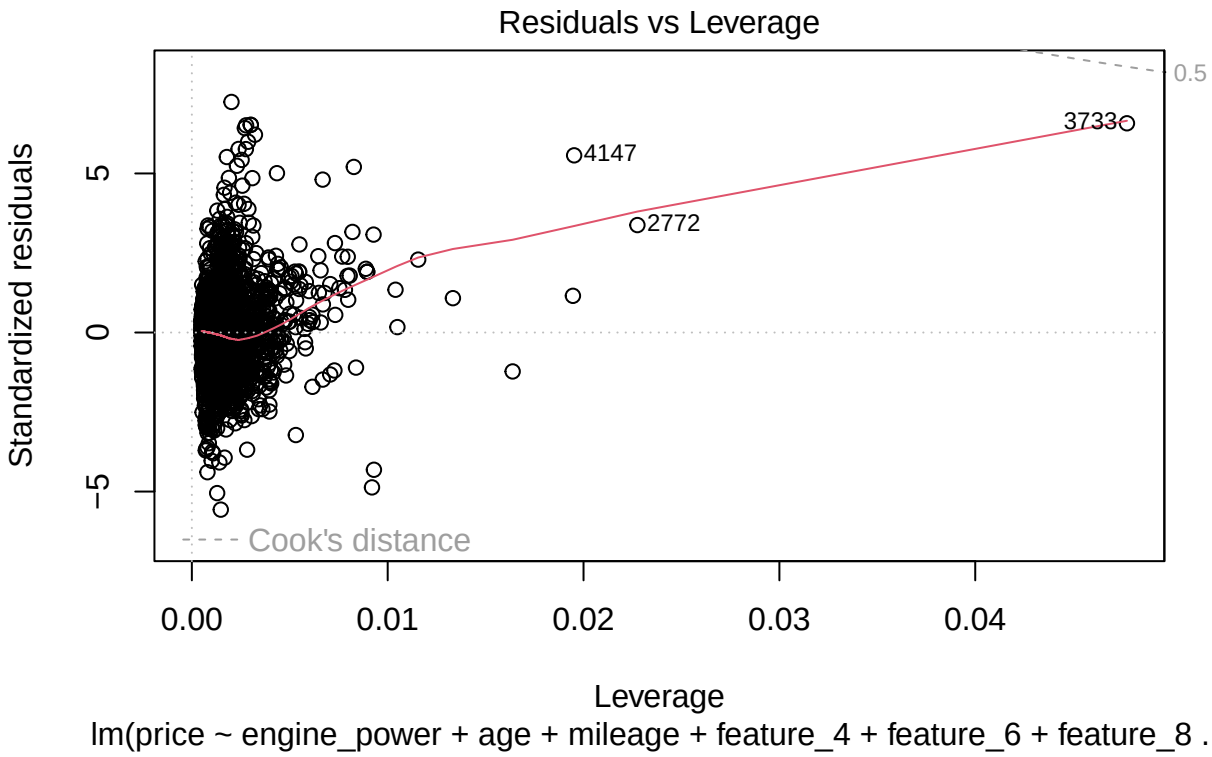
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4628 on 4830 degrees of freedom
## Multiple R-squared:  0.7093, Adjusted R-squared:  0.7089
## F-statistic:  1964 on 6 and 4830 DF,  p-value: < 2.2e-16
```

```
plot(mlr_6)
```




Q−Q Residuals
Standardized residuals
8
6
4
2
0
−2
−4
−6
4111
−4
−2
0
2
4
Theoretical Quantiles
lm(price ~ engine_power + age + mileage + feature_4 + feature_6 + feature_8 ...


Scale−Location
3733
4111
3321
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
−20000
0
20000
40000
60000
Fitted values
lm(price ~ engine_power + age + mileage + feature_4 + feature_6 + feature_8 ...



We see that the model's adjusted R^2 has decreasing by 0.1, yet the F-statistic has blown up to $> 1,900$. We should investigate further with an ANOVA test.

```
anova_result <- anova(mlr_6, mlr_4)
print(anova_result)
```

```
## Analysis of Variance Table
##
## Model 1: price ~ engine_power + age + mileage + feature_4 + feature_6 +
##     feature_8
## Model 2: price ~ engine_power + age + mileage + model_key + car_type +
##     feature_4 + feature_6 + feature_8
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1   4830 1.0347e+11
## 2   4750 6.1623e+10 80 4.1845e+10 40.318 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see that the full model provides a significantly better fit for the data, and therefore we should likely consider the model of the car in our analysis. This makes intuitive sense given our graph in which we saw that models could hugely differ in their distribution of price, and many models had means twice that of other models.

As we have done transformations and analysis of various models, we should now consider further removal of outliers.

```
# accumulated outliers for mlr_4 based on leverage and diagnostic graphs
outlier_indices <- c(56, 107, 1899, 2925, 3153, 3600, 4798, 4799, 4820, 4821, 2679,
```

```
                      3596, 3733, 73, 4147)
```

To remove outliers, we set a threshold of $\frac{10 \cdot p}{n}$ where p is the number of parameters and n is the number of observations (our sample size). We then unionize the resulting high-leverage data points with our prior outliers.

```
leverages <- hatvalues(mlr_4)

p <- length(coef(mlr_4))
n <- nrow(bmw_data_cleaned)
print(p)
```

```
## [1] 87
```

```
print(n)
```

```
## [1] 4837
```

```
threshold <- (10 * p) / n
print(threshold)
```

```
## [1] 0.1798636
```

```
high_leverage_indices <- which(leverages >= threshold)
print(length(high_leverage_indices))
```

```
## [1] 84
```

```
outlier_indices_expanded <- union(outlier_indices, high_leverage_indices)
print(length(outlier_indices_expanded))
```

```
## [1] 88
```

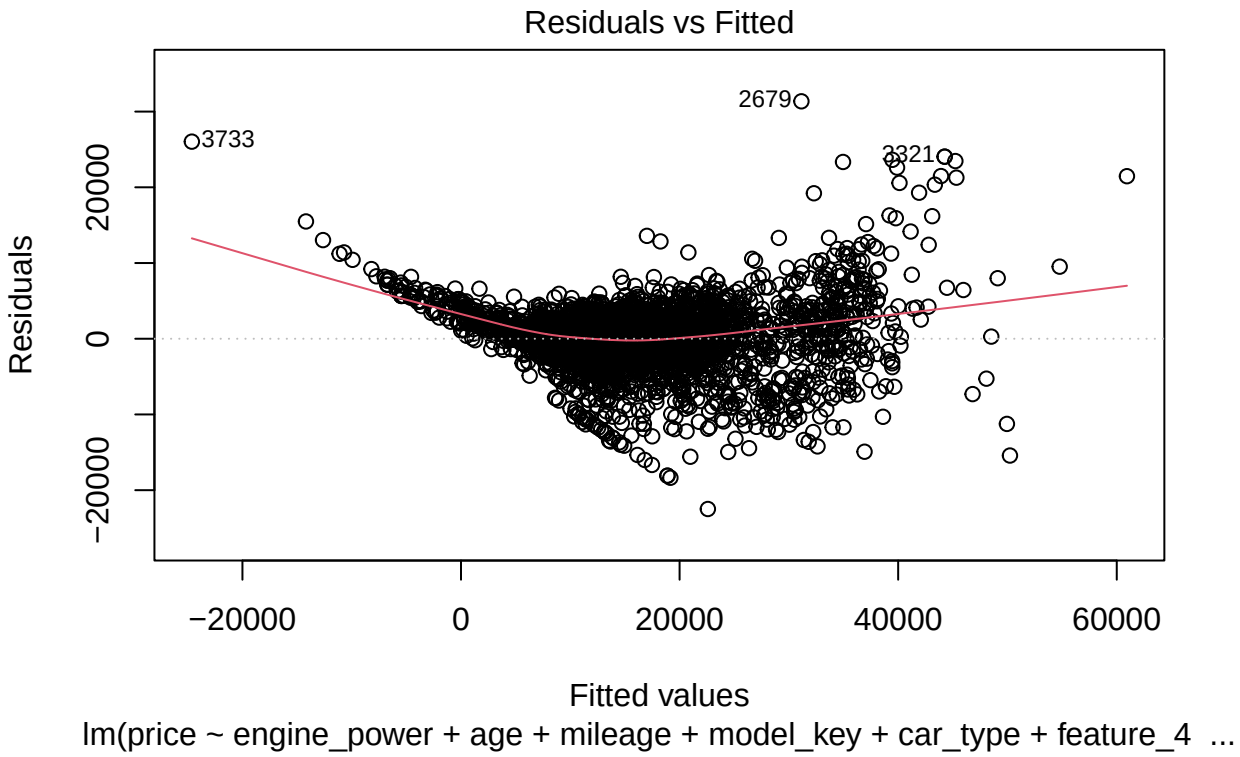
We refit our model having now removed our extensive number of outliers.

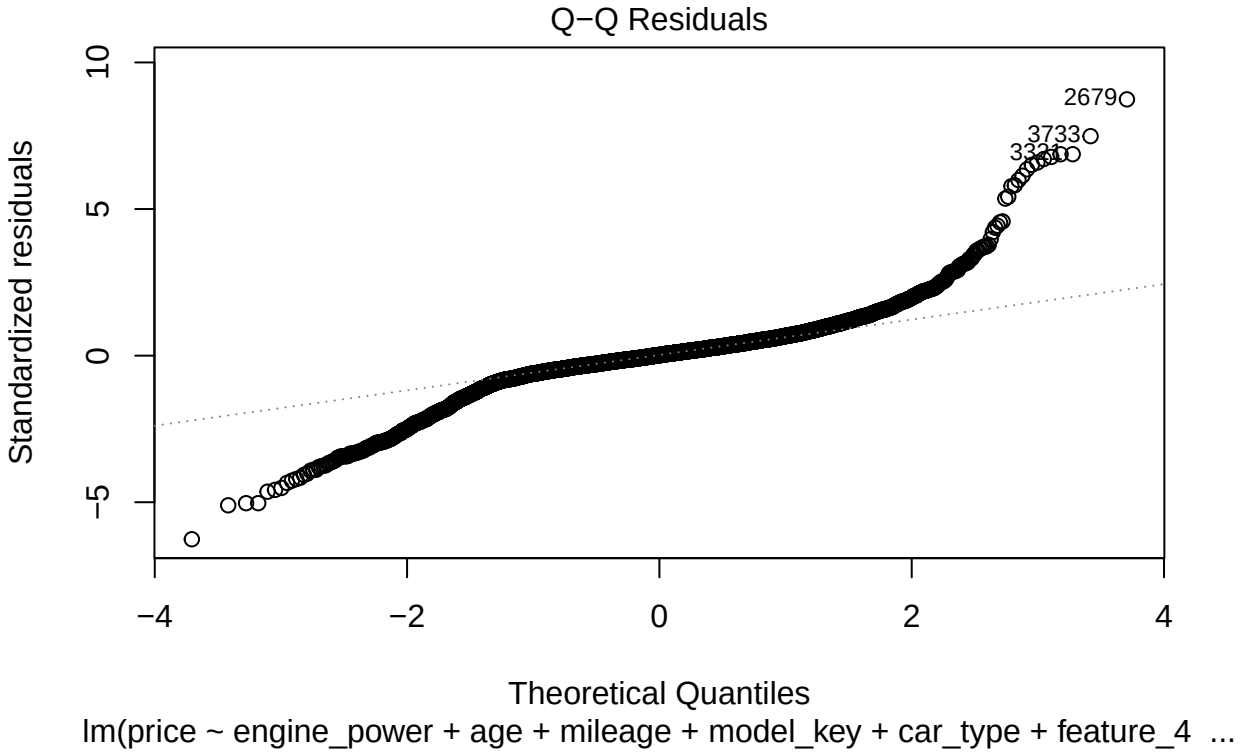
```
bmw_data_reduced <- bmw_data_cleaned[-outlier_indices_expanded, ]
mlr_4_v2 <-  lm(formula = price ~
              engine_power + age + mileage + model_key + car_type + feature_4 + feature_6 + feature_8,
              data = bmw_data_reduced)
summary(mlr_4_v2)
```

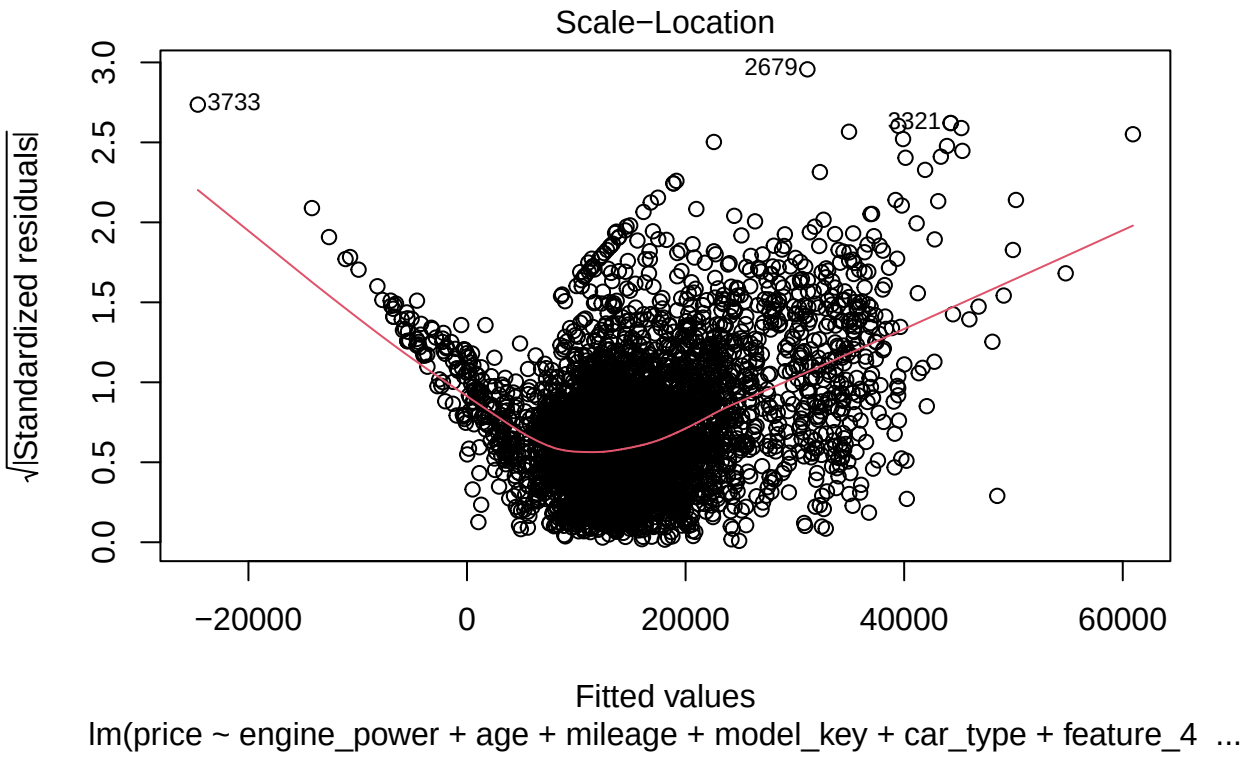
```
##
## Call:
## lm(formula = price ~ engine_power + age + mileage + model_key +
##     car_type + feature_4 + feature_6 + feature_8, data = bmw_data_reduced)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -22487  -1379     70   1543  31354
##
## Coefficients: (1 not defined because of singularities)
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 4.032e+04  1.629e+03  24.749  < 2e-16 ***
## engine_power                5.096e+01  3.929e+00  12.970  < 2e-16 ***
## age                        -1.053e+03  2.672e+01 -39.416  < 2e-16 ***
## mileage                    -3.030e-02  1.073e-03 -28.243  < 2e-16 ***
## model_keyX5 M              -8.630e+03  1.582e+03  -5.456 5.13e-08 ***
## model_key740               -7.879e+03  1.644e+03  -4.793 1.69e-06 ***
## model_key640 Gran Coupé    -8.978e+03  1.622e+03  -5.535 3.28e-08 ***
```

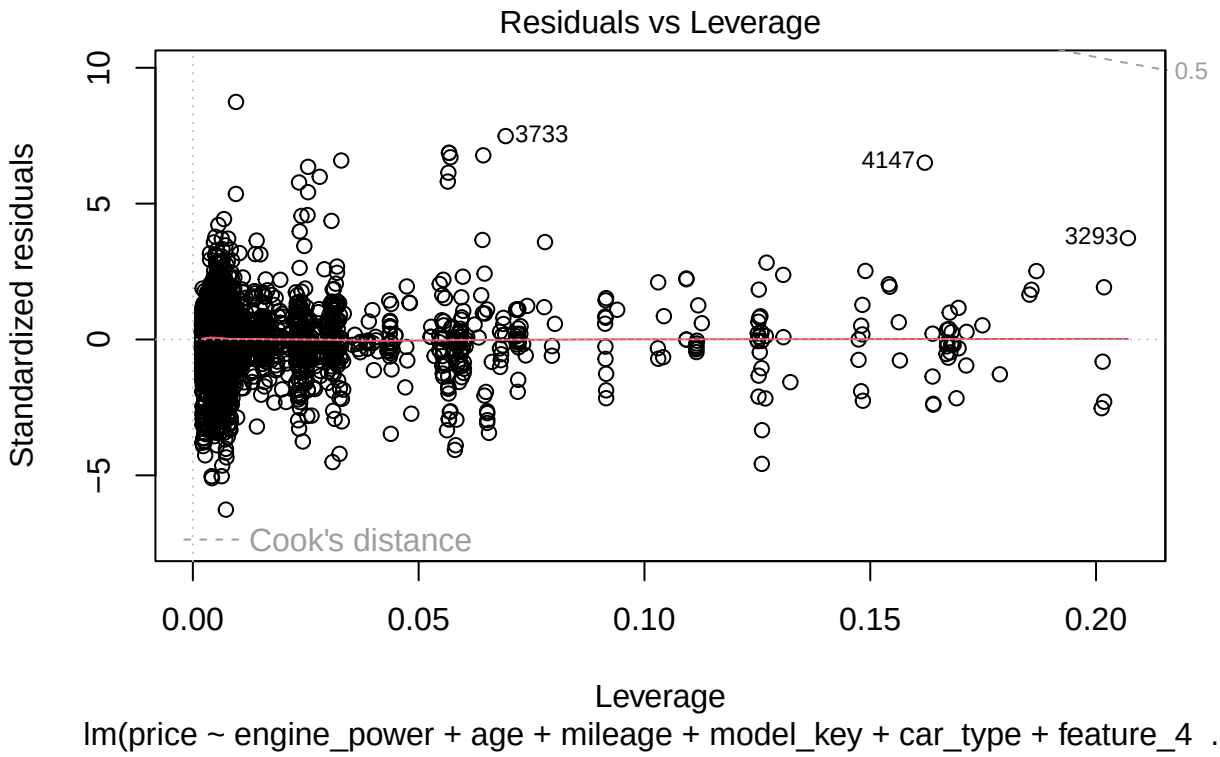
```
## model_keyM3                -9.280e+03  1.903e+03  -4.876 1.12e-06 ***
## model_keyM550              -1.409e+04  1.778e+03  -7.927 2.78e-15 ***
## model_keyX6                -9.967e+03  1.396e+03  -7.141 1.07e-12 ***
## model_key640               -1.057e+04  1.809e+03  -5.846 5.39e-09 ***
## model_key435 Gran Coupé    -1.448e+04  2.147e+03  -6.742 1.75e-11 ***
## model_keyX4                -1.343e+04  1.439e+03  -9.330  < 2e-16 ***
## model_keyX5                -1.247e+04  1.321e+03  -9.435  < 2e-16 ***
## model_key420 Gran Coupé    -1.822e+04  1.618e+03 -11.259  < 2e-16 ***
## model_key420               -1.886e+04  1.474e+03 -12.795  < 2e-16 ***
## model_key730               -1.551e+04  1.566e+03  -9.905  < 2e-16 ***
## model_key535               -1.975e+04  1.504e+03 -13.134  < 2e-16 ***
## model_key335               -2.171e+04  1.915e+03 -11.334  < 2e-16 ***
## model_keyi3                -1.669e+04  2.171e+03  -7.686 1.84e-14 ***
## model_key530 Gran Turismo  -1.954e+04  1.664e+03 -11.742  < 2e-16 ***
## model_key418 Gran Coupé    -1.908e+04  2.001e+03  -9.538  < 2e-16 ***
## model_key528               -2.118e+04  2.053e+03 -10.318  < 2e-16 ***
## model_key520 Gran Turismo  -1.818e+04  1.683e+03 -10.798  < 2e-16 ***
## model_key530               -2.022e+04  1.455e+03 -13.894  < 2e-16 ***
## model_keyX3                -2.178e+04  1.366e+03 -15.944  < 2e-16 ***
## model_key320 Gran Turismo  -2.236e+04  1.541e+03 -14.514  < 2e-16 ***
## model_key218               -2.309e+04  1.818e+03 -12.701  < 2e-16 ***
## model_key330               -2.392e+04  1.546e+03 -15.471  < 2e-16 ***
## model_key518               -2.010e+04  1.581e+03 -12.712  < 2e-16 ***
## model_key218 Gran Tourer   -1.941e+04  2.423e+03  -8.011 1.42e-15 ***
## model_key520               -2.041e+04  1.489e+03 -13.704  < 2e-16 ***
## model_key218 Active Tourer -1.801e+04  2.314e+03  -7.783 8.62e-15 ***
## model_key525               -2.075e+04  1.481e+03 -14.014  < 2e-16 ***
## model_key318 Gran Turismo  -2.214e+04  1.576e+03 -14.048  < 2e-16 ***
## model_key325               -2.488e+04  1.642e+03 -15.149  < 2e-16 ***
## model_keyX1                -2.502e+04  1.407e+03 -17.782  < 2e-16 ***
## model_key120               -2.564e+04  1.585e+03 -16.183  < 2e-16 ***
## model_key320               -2.365e+04  1.482e+03 -15.951  < 2e-16 ***
## model_key114               -2.600e+04  1.840e+03 -14.129  < 2e-16 ***
## model_key318               -2.308e+04  1.524e+03 -15.144  < 2e-16 ***
## model_key316               -2.367e+04  1.558e+03 -15.192  < 2e-16 ***
## model_key116               -2.572e+04  1.573e+03 -16.354  < 2e-16 ***
## model_key118               -2.551e+04  1.555e+03 -16.409  < 2e-16 ***
## model_keyZ4                -2.612e+04  2.147e+03 -12.162  < 2e-16 ***
## car_typesuv                        NA         NA      NA       NA
## car_typeconvertible         3.093e+03  7.791e+02   3.969 7.31e-05 ***
## car_typesedan              -1.711e+02  5.840e+02  -0.293 0.769603
## car_typevan                -6.944e+03  1.564e+03  -4.441 9.18e-06 ***
## car_typehatchback          -6.763e+02  6.399e+02  -1.057 0.290662
## car_typeestate             -2.192e+03  5.855e+02  -3.743 0.000184 ***
## car_typesubcompact          2.444e+02  7.291e+02   0.335 0.737416
## feature_4TRUE               1.642e+03  1.729e+02   9.498  < 2e-16 ***
## feature_6TRUE               1.511e+03  1.300e+02  11.626  < 2e-16 ***
## feature_8TRUE               1.257e+03  1.291e+02   9.735  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3604 on 4695 degrees of freedom
## Multiple R-squared:  0.8203, Adjusted R-squared:  0.8183
## F-statistic: 404.5 on 53 and 4695 DF,  p-value: < 2.2e-16
```

```
plot(mlr_4_v2)
```


Residuals vs Fitted
2679
3733
3321
Residuals
20000
0
−20000
−20000
0
20000
40000
60000
Fitted values
lm(price ~ engine_power + age + mileage + model_key + car_type + feature_4 ...


Q−Q Residuals
Standardized residuals
10
5
0
−5
2679
3733
−4
−2
0
2
4
Theoretical Quantiles
lm(price ~ engine_power + age + mileage + model_key + car_type + feature_4 ...


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3733
2679
3321
−20000
0
20000
40000
60000
Fitted values
lm(price ~ engine_power + age + mileage + model_key + car_type + feature_4 ...



Despite our lump-sum removal of 311 outliers, our model hasn't improve significantly. Not only does this inform us that simply removing terms based on their leverage is not meaningful, but it also shows the danger of removing many data points simply in hopes of improving a model's reported statistics. We need to be more precise in the way we improve our model. Currently, our model seems roughly "capped" at an adjusted R^2 of ~ 0.8.