# 3131 Assignment

2025-10-19

## Setup

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(stringr)
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

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

```
data = read.csv('hdb-resale-Jan-Jun2021.csv')
head(data)
```

```
##     month       town flat_type block      street_name storey_range
## 1 2021-01 ANG MO KIO    2 ROOM   170 ANG MO KIO AVE 4     07 TO 09
## 2 2021-01 ANG MO KIO    2 ROOM   170 ANG MO KIO AVE 4     01 TO 03
## 3 2021-01 ANG MO KIO    3 ROOM   216 ANG MO KIO AVE 1     04 TO 06
## 4 2021-01 ANG MO KIO    3 ROOM   223 ANG MO KIO AVE 1     07 TO 09
## 5 2021-01 ANG MO KIO    3 ROOM   223 ANG MO KIO AVE 1     10 TO 12
## 6 2021-01 ANG MO KIO    3 ROOM   331 ANG MO KIO AVE 1     04 TO 06
##   floor_area_sqm     flat_model lease_commence_date    remaining_lease
## 1             45       Improved               1986  64 years 01 month
## 2             45       Improved               1986  64 years 01 month
## 3             73 New Generation               1976 54 years 04 months
## 4             67 New Generation               1978  56 years 01 month
## 5             67 New Generation               1978           56 years
## 6             68 New Generation               1981           59 years
##   resale_price
## 1       225000
## 2       211000
## 3       275888
```

```
## 4       316800
## 5       305000
## 6       260000
```

```
dim(data)
```

```
## [1] 11527    11
```

```
any(is.na(data))
```

```
## [1] FALSE
```

```
colnames(data)
```

```
##  [1] "month"              "town"               "flat_type"
##  [4] "block"              "street_name"        "storey_range"
##  [7] "floor_area_sqm"     "flat_model"         "lease_commence_date"
## [10] "remaining_lease"    "resale_price"
```

```
data$town <- as.factor(data$town) # group into regions
levels(data$town)
```

```
##  [1] "ANG MO KIO"      "BEDOK"           "BISHAN"          "BUKIT BATOK"
##  [5] "BUKIT MERAH"     "BUKIT PANJANG"   "BUKIT TIMAH"     "CENTRAL AREA"
##  [9] "CHOA CHU KANG"   "CLEMENTI"        "GEYLANG"         "HOUGANG"
## [13] "JURONG EAST"     "JURONG WEST"     "KALLANG/WHAMPOA" "MARINE PARADE"
## [17] "PASIR RIS"       "PUNGGOL"         "QUEENSTOWN"      "SEMBAWANG"
## [21] "SENGKANG"        "SERANGOON"       "TAMPINES"        "TOA PAYOH"
## [25] "WOODLANDS"       "YISHUN"
```

```
data$flat_type <- as.factor(data$flat_type)
levels(data$flat_type)
```

```
## [1] "1 ROOM"           "2 ROOM"           "3 ROOM"           "4 ROOM"
## [5] "5 ROOM"           "EXECUTIVE"        "MULTI-GENERATION"
```

```
data$street_name <- as.factor(data$street_name)
data$storey_range <- as.factor(data$storey_range)
levels(data$storey_range)
```

```
##  [1] "01 TO 03" "04 TO 06" "07 TO 09" "10 TO 12" "13 TO 15" "16 TO 18"
##  [7] "19 TO 21" "22 TO 24" "25 TO 27" "28 TO 30" "31 TO 33" "34 TO 36"
## [13] "37 TO 39" "40 TO 42" "43 TO 45" "46 TO 48" "49 TO 51"
```

```
data$flat_model <- as.factor(data$flat_model)
levels(data$flat_model)
```

```
##  [1] "2-room"                 "Adjoined flat"          "Apartment"
##  [4] "DBSS"                   "Improved"               "Improved-Maisonette"
##  [7] "Maisonette"             "Model A"                "Model A-Maisonette"
## [10] "Model A2"               "Multi Generation"       "New Generation"
## [13] "Premium Apartment"      "Premium Apartment Loft" "Premium Maisonette"
## [16] "Simplified"             "Standard"               "Terrace"
## [19] "Type S1"                "Type S2"
```

## EDA: Exploring the variables and association

```
# --- Numeric summary ---
cat("Summary of resale_price:\n")
```

```
## Summary of resale_price:
```

```
print(summary(data$resale_price))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  180000  380000  468000  496545  585000 1250000
```

```
# Calculate descriptive statistics
mean_val   <- mean(data$resale_price, na.rm = TRUE)
sd_val     <- sd(data$resale_price, na.rm = TRUE)
min_val    <- min(data$resale_price, na.rm = TRUE)
q1_val     <- quantile(data$resale_price, 0.25, na.rm = TRUE)
median_val <- median(data$resale_price, na.rm = TRUE)
q3_val     <- quantile(data$resale_price, 0.75, na.rm = TRUE)
max_val    <- max(data$resale_price, na.rm = TRUE)
n_out      <- length(boxplot(data$resale_price, plot = FALSE)$out)

# Print numeric results
cat("\n--- Descriptive Statistics ---\n")
```

```
##
## --- Descriptive Statistics ---
```

```
cat("Mean: ", round(mean_val, 2), "\n")
```

```
## Mean:  496544.6
```

```
cat("Standard Deviation: ", round(sd_val, 2), "\n")
```

```
## Standard Deviation:  161965.1
```

```
cat("Min: ", round(min_val, 0), "\n")
```

```
## Min:  180000
```

```
cat("Q1: ", round(q1_val, 0), "\n")
```

```
## Q1:  380000
```

```
cat("Median: ", round(median_val, 0), "\n")
```

```
## Median:  468000
```

```
cat("Q3: ", round(q3_val, 0), "\n")
```

```
## Q3:  585000
```

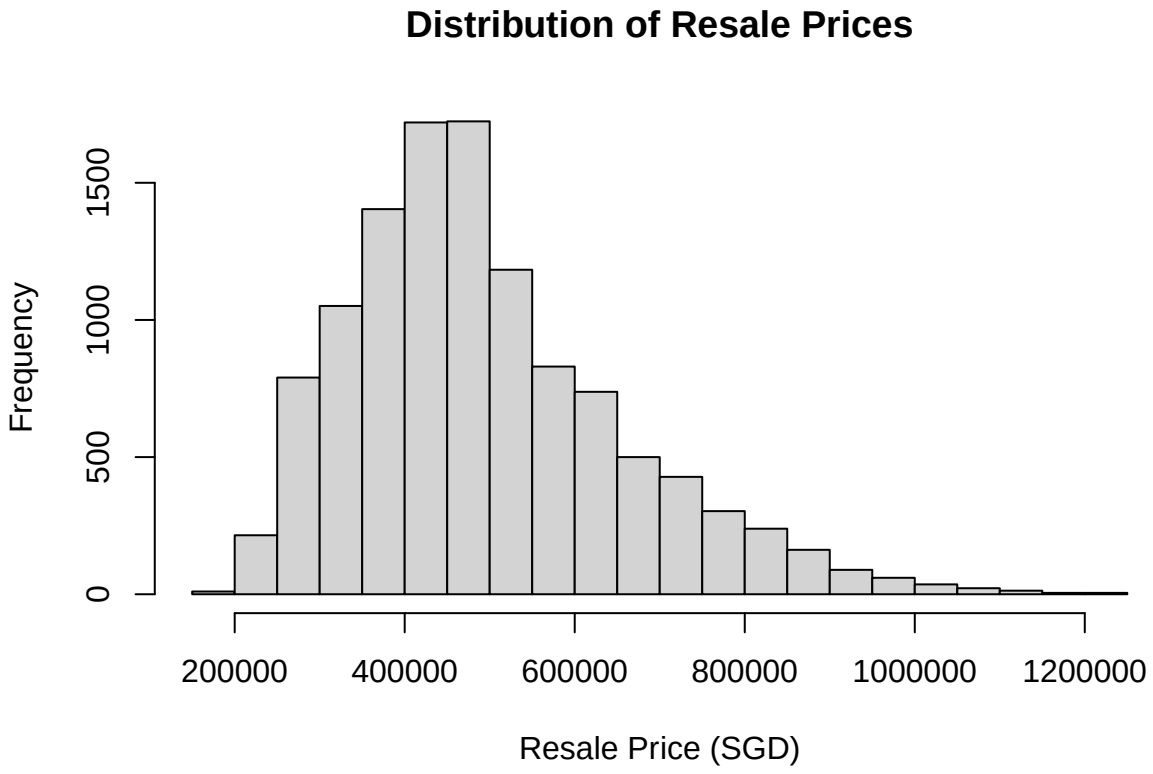
```
cat("Max: ", round(max_val, 0), "\n")
```

```
## Max:  1250000
```

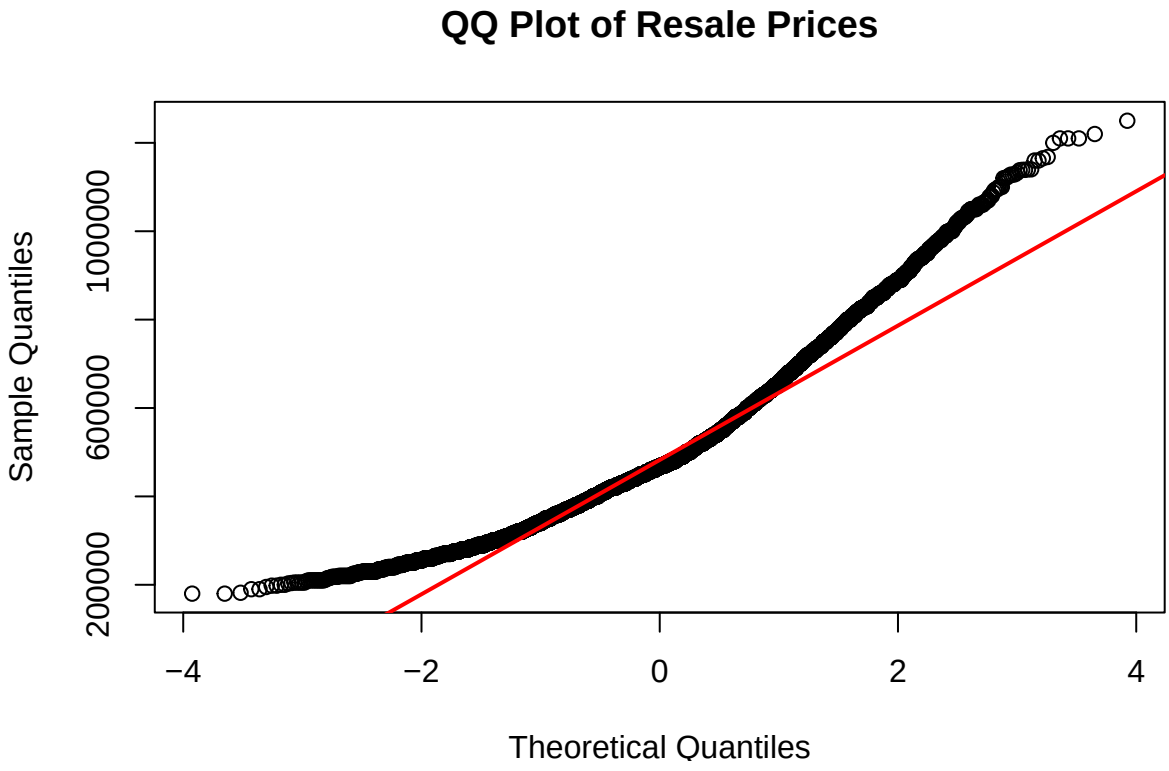
```
cat("Number of Outliers: ", n_out, "\n")
```

```
## Number of Outliers:  245
```

```
# --- Base R Plots (no colors) ---

# Histogram of resale prices
hist(data$resale_price,
     main = "Distribution of Resale Prices",
     xlab = "Resale Price (SGD)",
     breaks = 30)
```

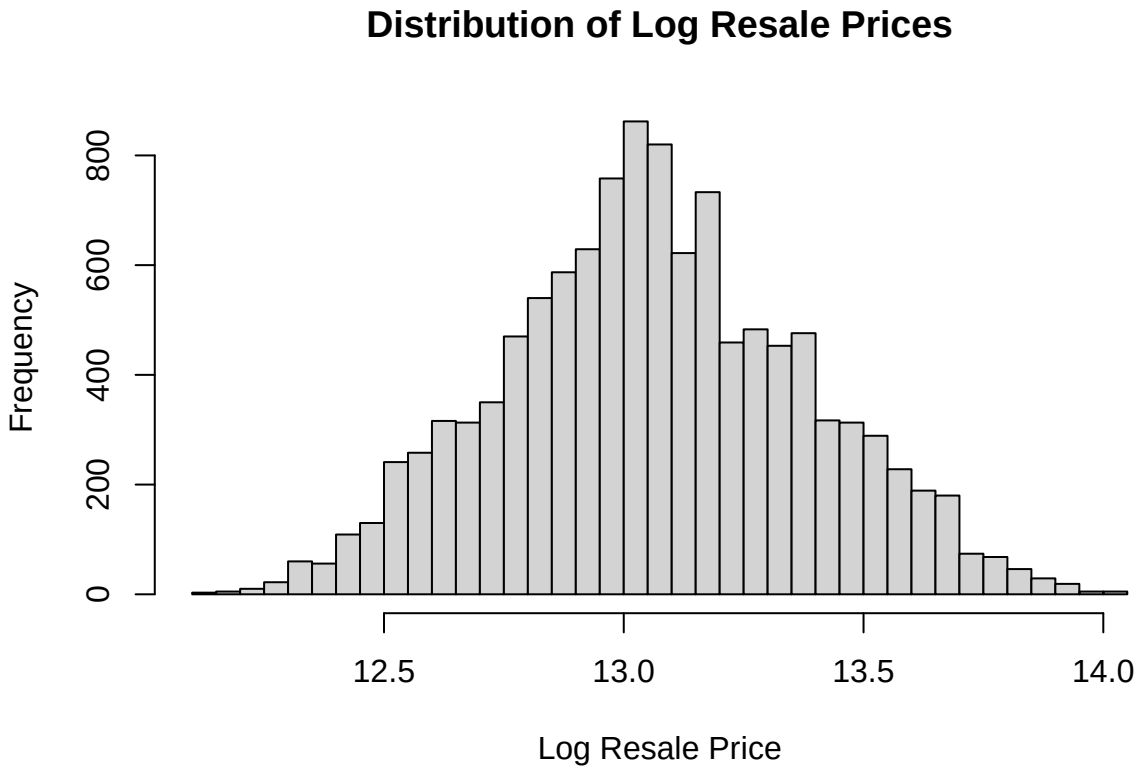



```r
# QQ plot for resale prices
qqnorm(data$resale_price, main = "QQ Plot of Resale Prices")
qqline(data$resale_price, col='red', lwd = 2)
```

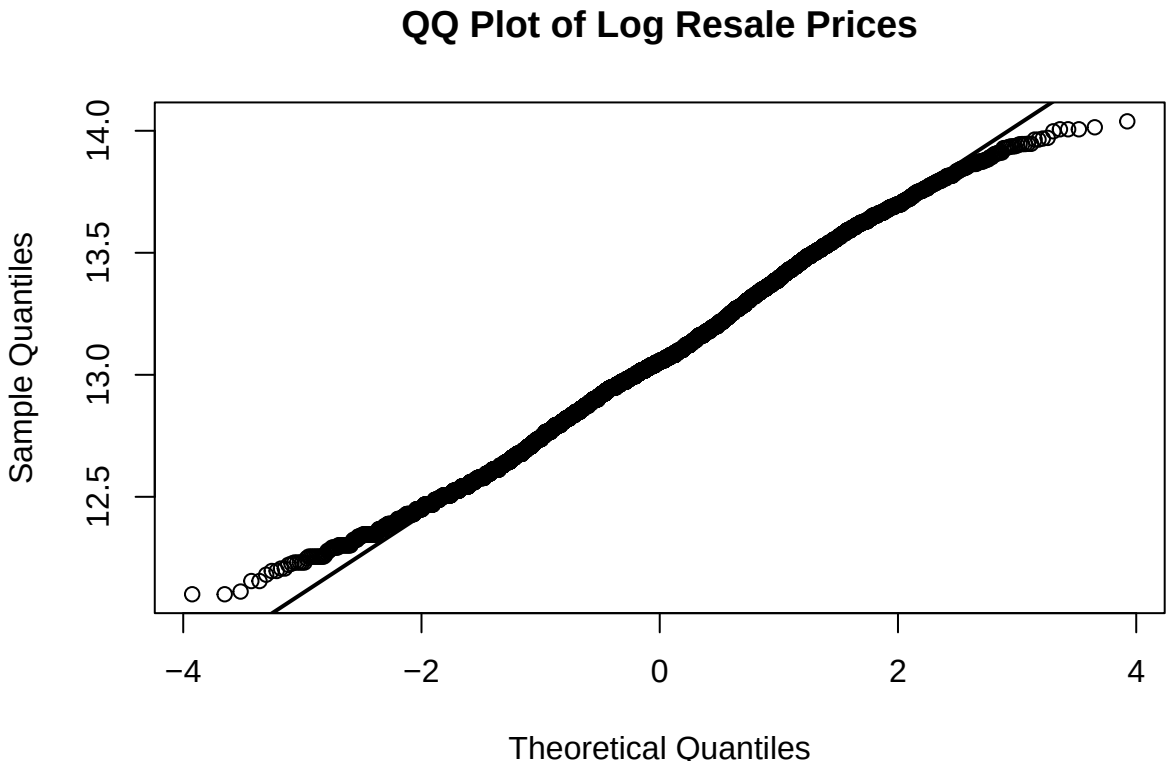



```r
# Histogram of log-transformed prices
hist(log(data$resale_price),
```

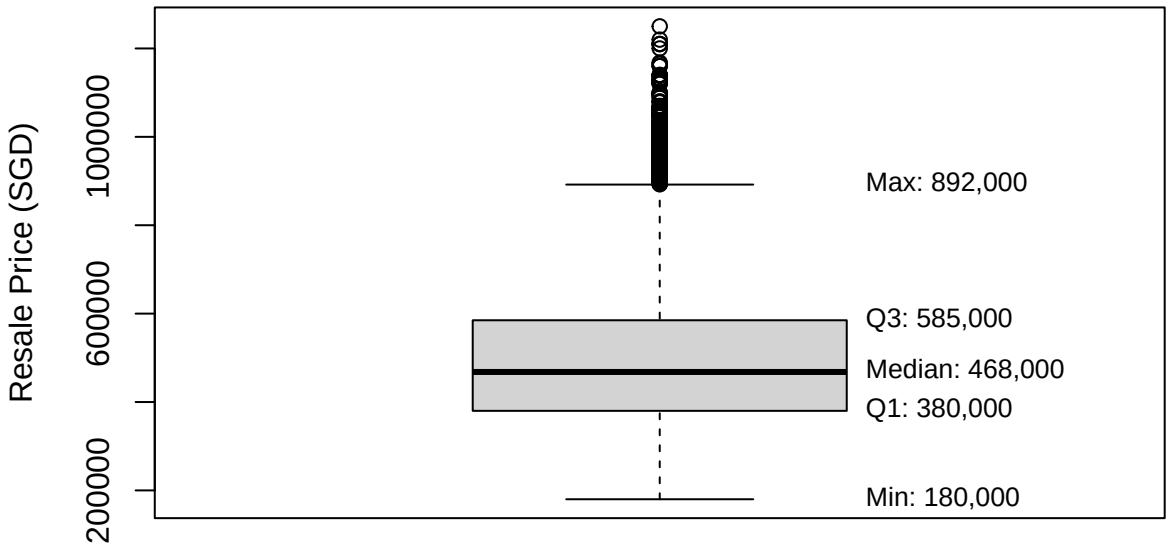
```
     main = "Distribution of Log Resale Prices",
     xlab = "Log Resale Price",
     breaks = 30)
```




```
# QQ plot for log-transformed prices
qqnorm(log(data$resale_price), main = "QQ Plot of Log Resale Prices")
qqline(log(data$resale_price), lwd = 2)
```



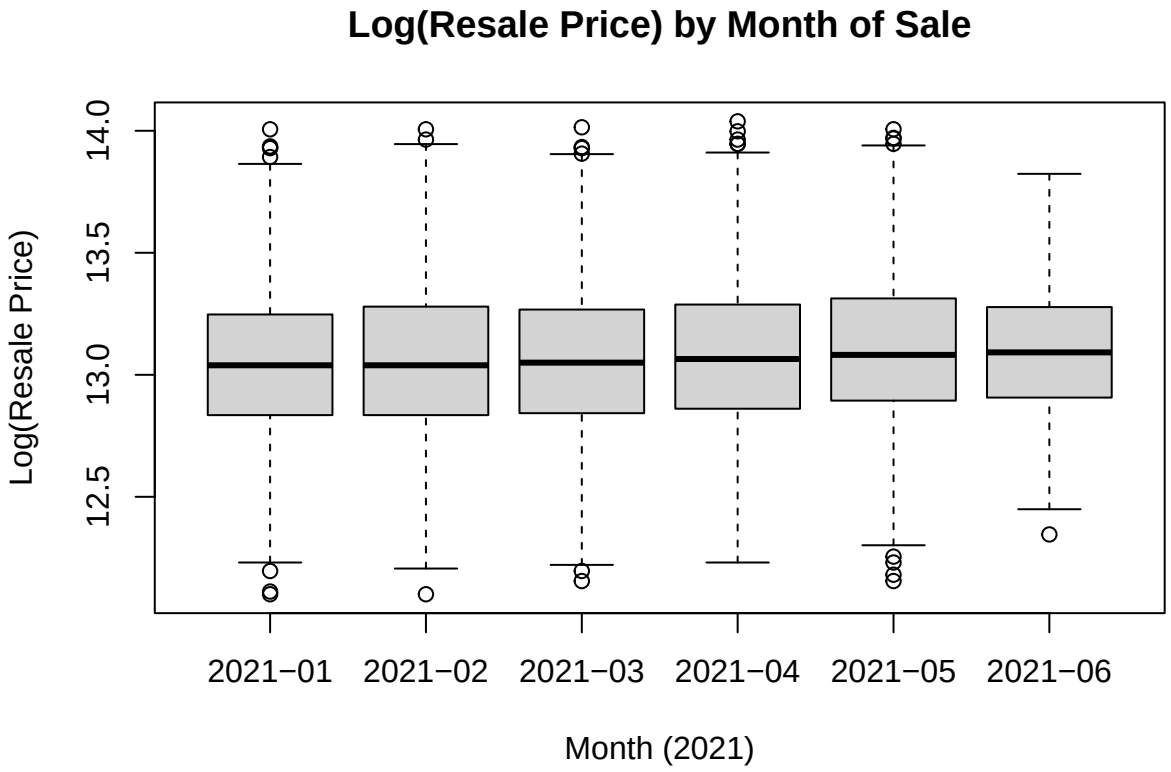

```r
# Boxplot of resale prices
bp <- boxplot(data$resale_price,
              main = "Boxplot of Resale Prices",
              ylab = "Resale Price (SGD)")

# Add five-number summary labels, spaced for clarity
stats <- as.numeric(bp$stats)
names <- c("Min", "Q1", "Median", "Q3", "Max")

# Adjust x position and text spacing for readability
text(x = 1.2,
     y = stats,
     labels = paste0(names, ": ", format(round(stats, 0), big.mark = ",")),
     pos = 4,
     cex = 0.8)
```



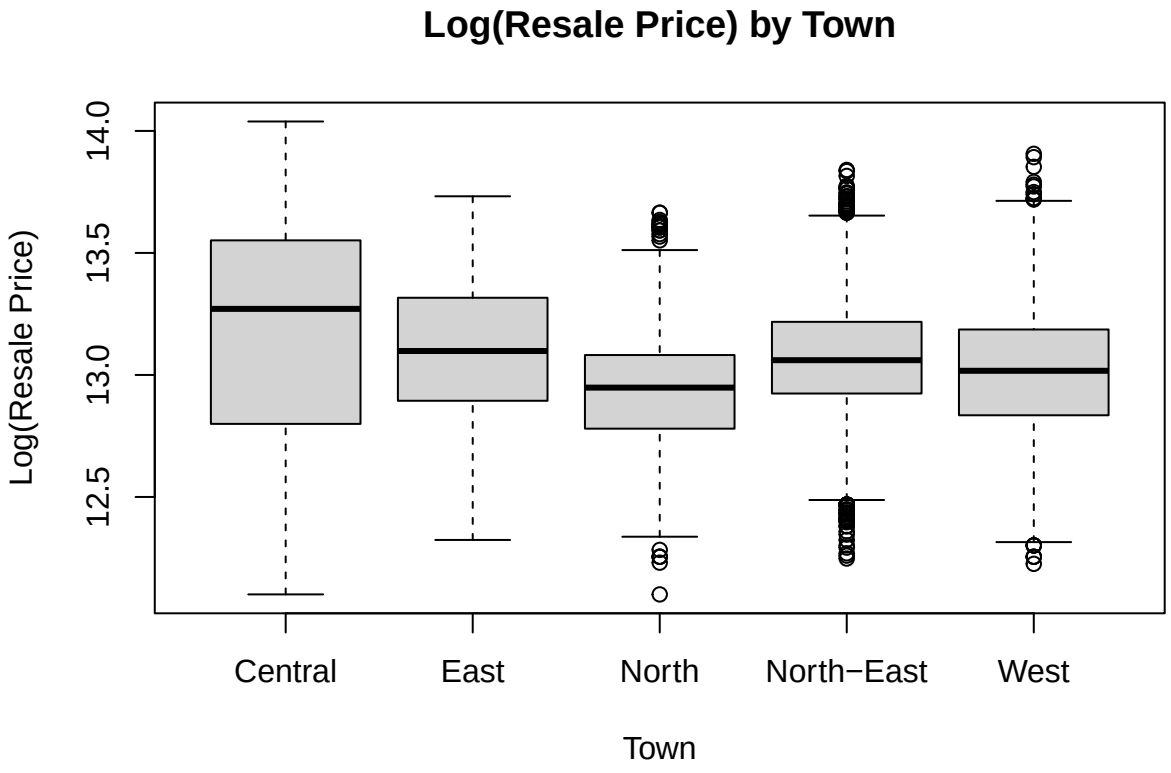

```
data <- data %>%
  mutate(town = case_when(
    town %in% c("BISHAN", "TOA PAYOH", "KALLANG/WHAMPOA", "CENTRAL AREA",
                "QUEENSTOWN", "BUKIT TIMAH", "MARINE PARADE", "BUKIT MERAH", "GEYLANG") ~ "Central",
    town %in% c("BEDOK", "TAMPINES", "PASIR RIS") ~ "East",
    town %in% c("WOODLANDS", "YISHUN", "SEMBAWANG") ~ "North",
    town %in% c("ANG MO KIO", "HOUGANG", "PUNGGOL", "SENGKANG", "SERANGOON") ~ "North-East",
    town %in% c("BUKIT BATOK", "BUKIT PANJANG", "CLEMENTI", "JURONG EAST",
                "JURONG WEST", "CHOA CHU KANG") ~ "West",
  ))
data$town <- factor(data$town, levels = c("Central", "East", "North", "North-East", "West"))
unique(data$town)
```

```
## [1] North-East East       Central    West       North
## Levels: Central East North North-East West
```

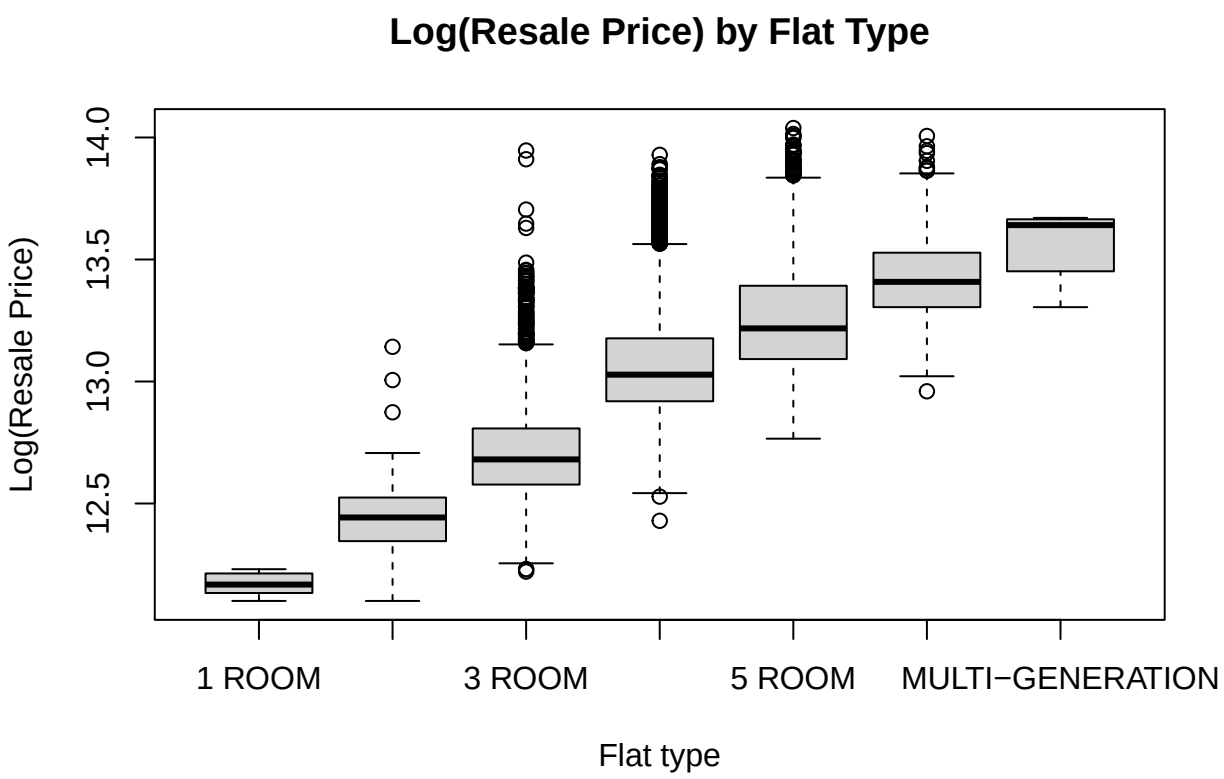
```
data <- data[, !(names(data) %in% c("street_name", "block", "flat_model"))]

# Boxplot: Log Resale Price by Month
boxplot(log(data$resale_price) ~ data$month,
        main = "Log(Resale Price) by Month of Sale",
        xlab = "Month (2021)",
        ylab = "Log(Resale Price)")
```




```r
# Boxplot: Log Resale Price by Town
boxplot(log(data$resale_price) ~ data$town,
        main = "Log(Resale Price) by Town",
        xlab = "Town",
        ylab = "Log(Resale Price)")
```

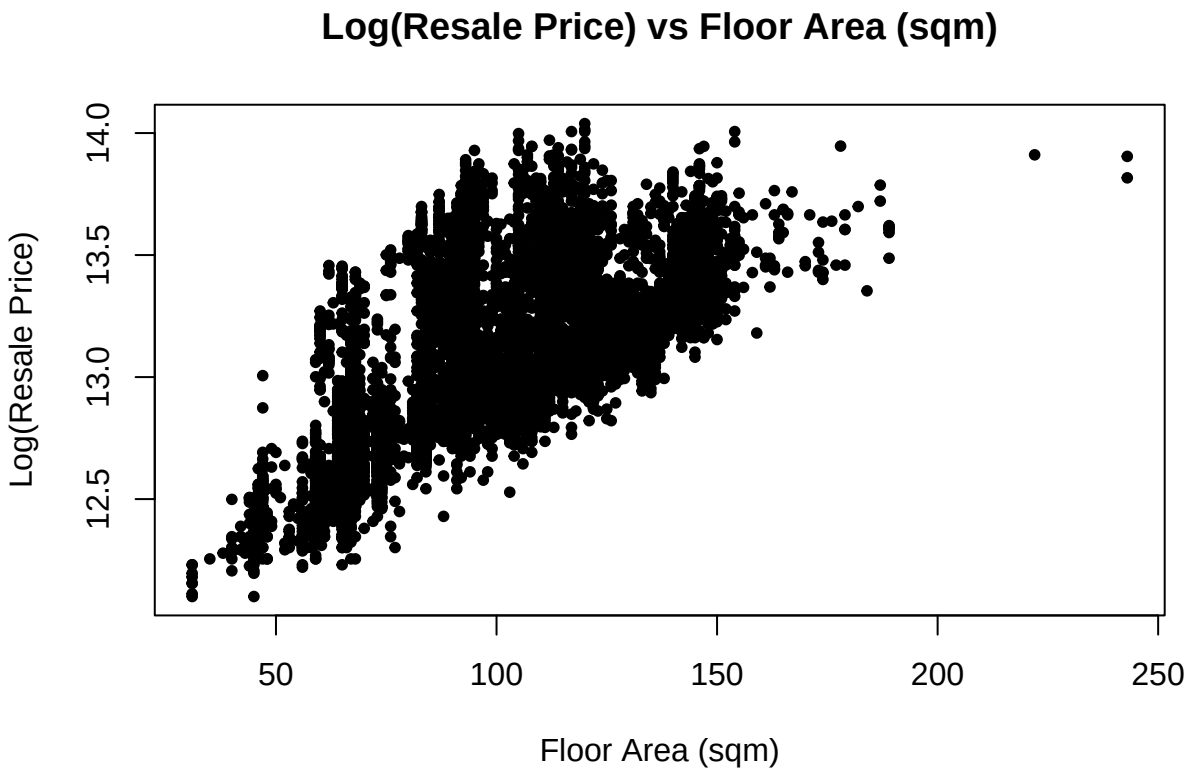


```r
# Boxplot: Log Resale Price by Flat Type
boxplot(log(data$resale_price) ~ data$flat_type,
        main = "Log(Resale Price) by Flat Type",
        xlab = "Flat type",
        ylab = "Log(Resale Price)",
        )
```




```r
# Convert Remaining Lease to numeric (months)

data <- data %>%
  mutate(
    years = as.numeric(str_extract(remaining_lease, "\\d+(?=\\s*years?)")),
    months = as.numeric(str_extract(remaining_lease, "\\d+(?=\\s*month)")),
    months = ifelse(is.na(months), 0, months),
    remaining_lease_months = years * 12 + months
  )

# Convert Storey Range to numerical midpoint (e.g., "01 TO 03" → 2)
data$storey_range <- sapply(strsplit(as.character(data$storey_range), " TO "), function(x) {
  low  <- as.numeric(x[1])
  high <- as.numeric(x[2])
  mean(c(low, high))
})

# --- Scatterplots with log resale price ---

# 1. Floor Area (sqm)
plot(data$floor_area_sqm, log(data$resale_price),
     pch = 20,
```

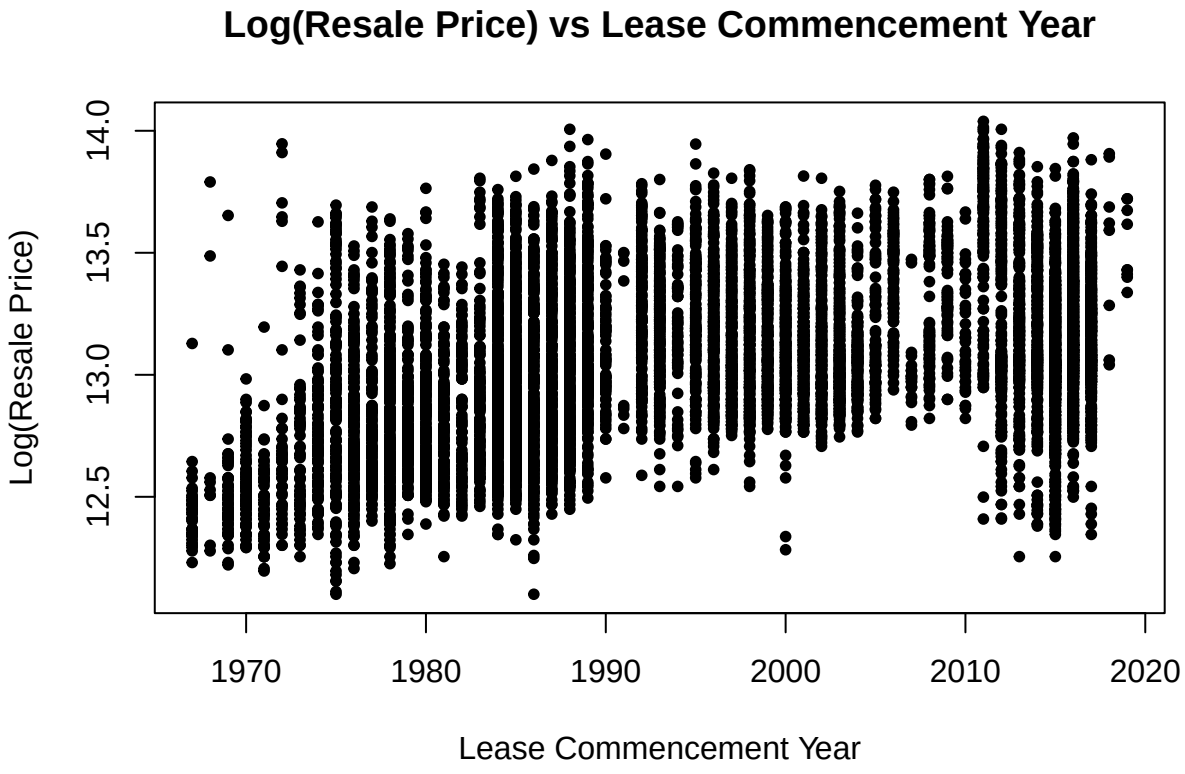
```
     main = "Log(Resale Price) vs Floor Area (sqm)",
     xlab = "Floor Area (sqm)",
     ylab = "Log(Resale Price)")
```




```
cat("Correlation (Log Resale Price ~ Floor Area): ",
    round(cor(log(data$resale_price), data$floor_area_sqm, use = "complete.obs"), 4), "\n")
```

```
## Correlation (Log Resale Price ~ Floor Area):  0.6848
```

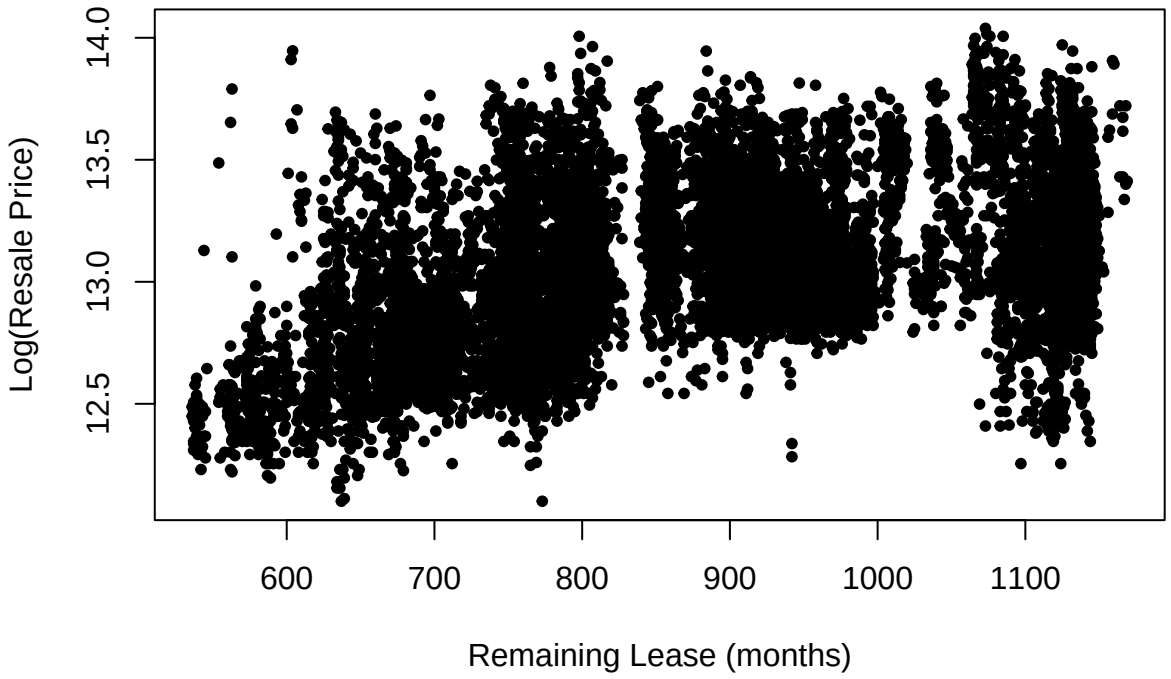
```
# 2. Lease Commencement Date
plot(data$lease_commence_date, log(data$resale_price),
     pch = 20,
     main = "Log(Resale Price) vs Lease Commencement Year",
     xlab = "Lease Commencement Year",
     ylab = "Log(Resale Price)")
```




```
cat("Correlation (Log Resale Price ~ Lease Commencement Year): ",
    round(cor(log(data$resale_price), data$lease_commence_date, use = "complete.obs"), 4), "\n")
```

```
## Correlation (Log Resale Price ~ Lease Commencement Year):  0.4098
```

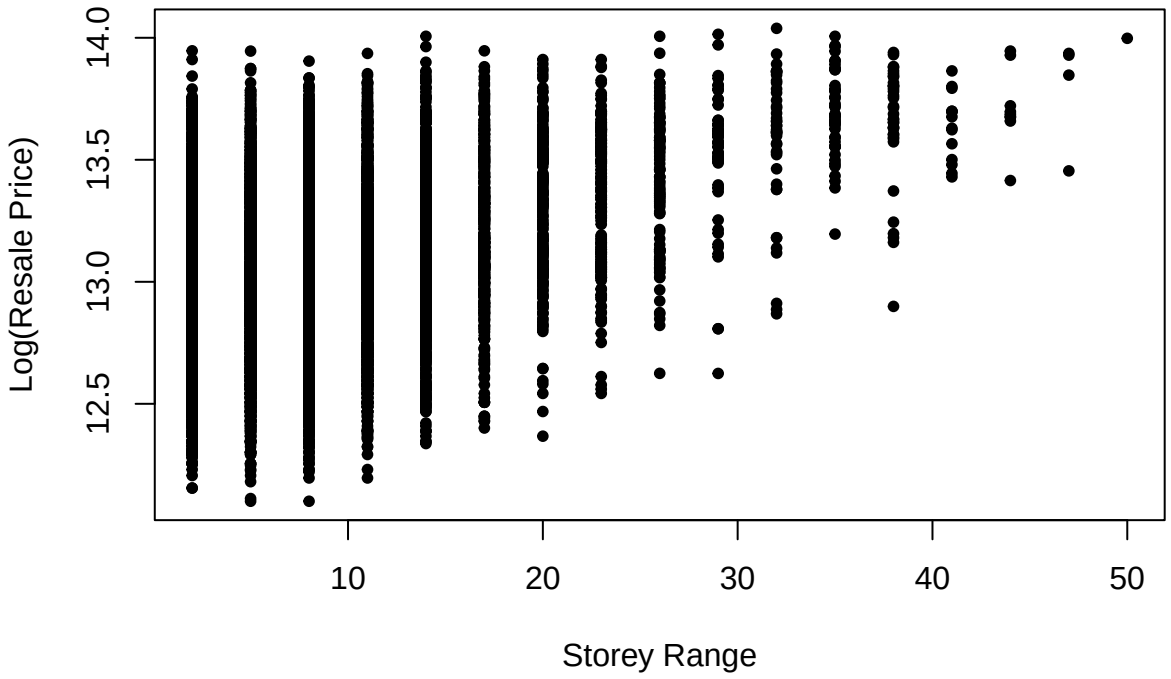
```
# 3. Remaining Lease (months)
plot(data$remaining_lease_months, log(data$resale_price),
     pch = 20,
     main = "Log(Resale Price) vs Remaining Lease (months)",
     xlab = "Remaining Lease (months)",
     ylab = "Log(Resale Price)")
```




```
cat("Correlation (Log Resale Price ~ Remaining Lease): ",
    round(cor(log(data$resale_price), data$remaining_lease_months, use = "complete.obs"), 4), "\n")
```

```
## Correlation (Log Resale Price ~ Remaining Lease):  0.4097
```

```
# 4. Storey Range
plot(data$storey_range, log(data$resale_price),
     pch = 20,
     main = "Log(Resale Price) vs Storey Range",
     xlab = "Storey Range",
     ylab = "Log(Resale Price)")
```




```
cat("Correlation (Log Resale Price ~ Storey Midpoint): ",
    round(cor(log(data$resale_price), data$storey_range, use = "complete.obs"), 4), "\n")
```

```
## Correlation (Log Resale Price ~ Storey Midpoint):  0.3488
```

```
data <- data[, !(names(data) %in% c("remaining_lease_months", "remaining_lease"))]
```

```
M0 = lm(log(resale_price) ~ month + town + flat_type + storey_range + floor_area_sqm + lease_commence_da
summary(M0)
```

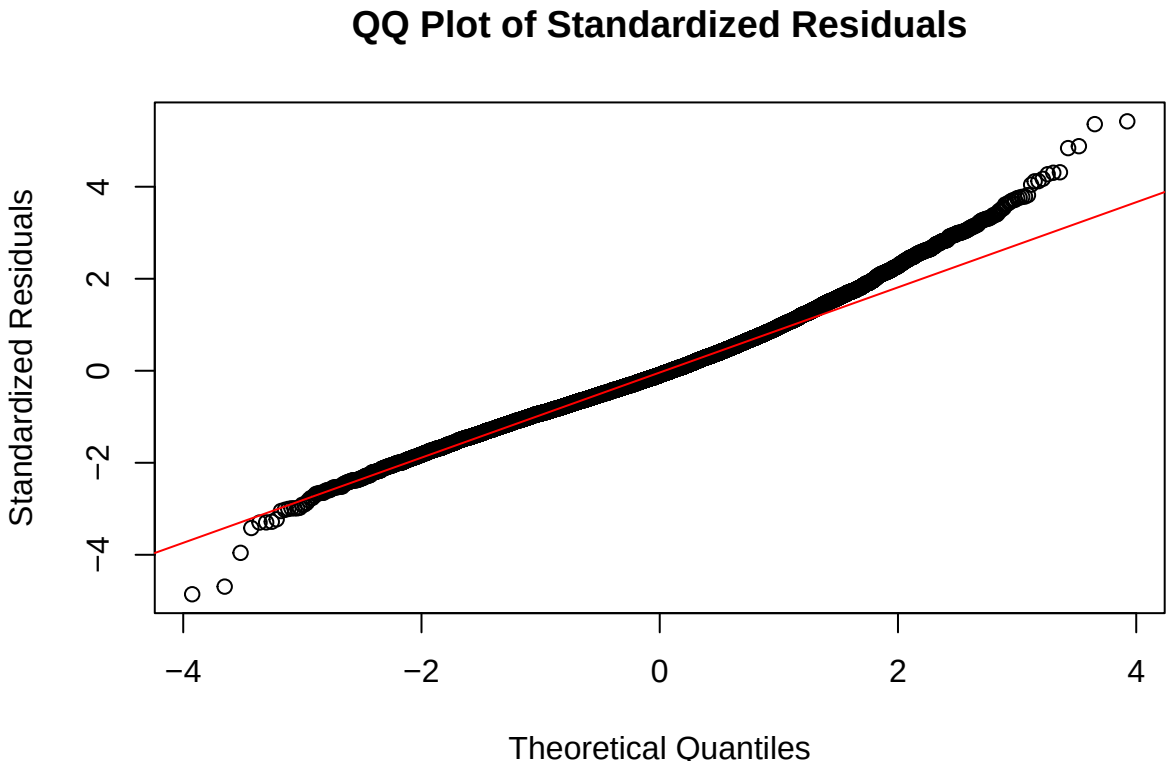
```
##
## Call:
## lm(formula = log(resale_price) ~ month + town + flat_type + storey_range +
##     floor_area_sqm + lease_commence_date, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.63630 -0.08762 -0.01244  0.07746  0.71643
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -5.0685191  0.2160316 -23.462  < 2e-16 ***
## month2021-02                0.0109463  0.0038917   2.813  0.00492 **
## month2021-03                0.0185823  0.0037711   4.928 8.44e-07 ***
## month2021-04                0.0278929  0.0038111   7.319 2.67e-13 ***
## month2021-05                0.0399064  0.0039944   9.990  < 2e-16 ***
## month2021-06                0.0553566  0.0124180   4.458 8.36e-06 ***
## townEast                   -0.1969752  0.0042704 -46.125  < 2e-16 ***
## townNorth                  -0.3749700  0.0046397 -80.817  < 2e-16 ***
## townNorth-East             -0.2928621  0.0038683 -75.709  < 2e-16 ***
## townWest                   -0.3205321  0.0040002 -80.128  < 2e-16 ***
```

```
## flat_type2 ROOM            0.1356094  0.0481335   2.817  0.00485 **
## flat_type3 ROOM            0.3290025  0.0473856   6.943 4.04e-12 ***
## flat_type4 ROOM            0.4009178  0.0483694   8.289  < 2e-16 ***
## flat_type5 ROOM            0.4301175  0.0495009   8.689  < 2e-16 ***
## flat_typeEXECUTIVE         0.4951049  0.0512553   9.660  < 2e-16 ***
## flat_typeMULTI-GENERATION  0.5662775  0.0695418   8.143 4.25e-16 ***
## storey_range               0.0097574  0.0002144  45.513  < 2e-16 ***
## floor_area_sqm             0.0075915  0.0001764  43.046  < 2e-16 ***
## lease_commence_date        0.0085703  0.0001063  80.602  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1323 on 11508 degrees of freedom
## Multiple R-squared:  0.8261, Adjusted R-squared:  0.8259
## F-statistic:  3038 on 18 and 11508 DF,  p-value: < 2.2e-16
```

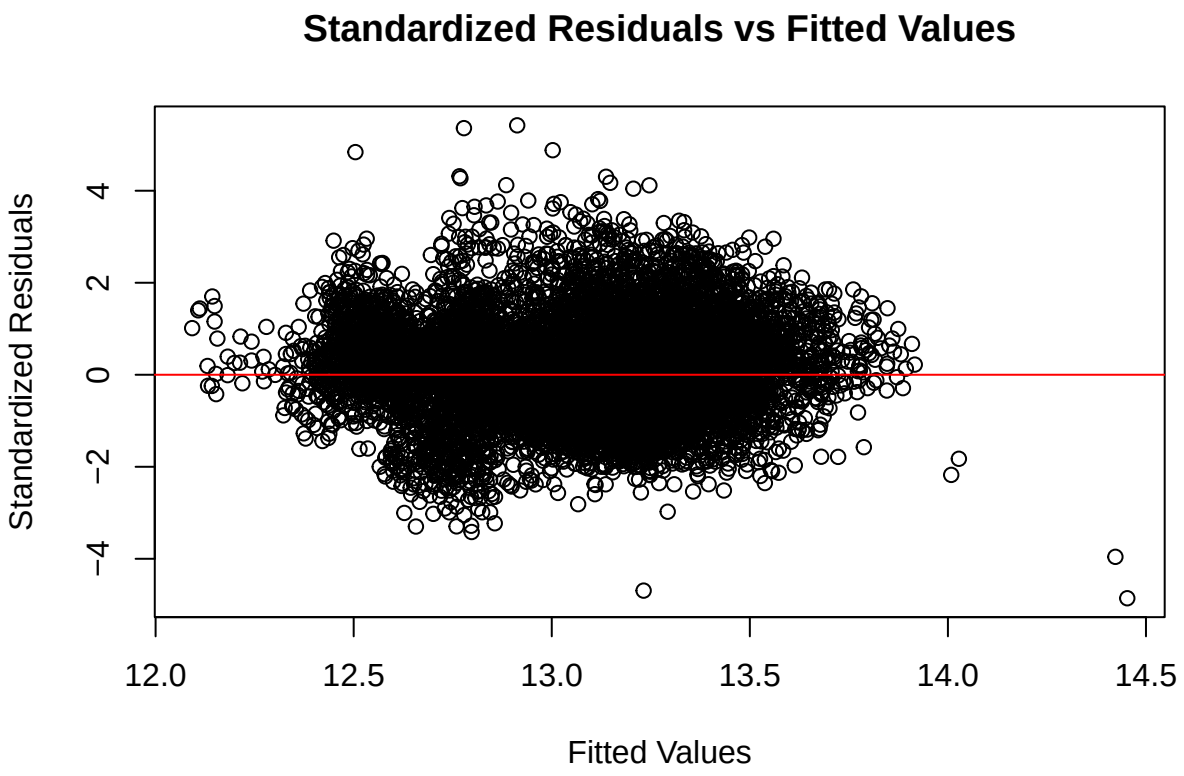
```r
anova(M0)
```

```
## Analysis of Variance Table
##
## Response: log(resale_price)
##                        Df Sum Sq Mean Sq  F value    Pr(>F)
## month                   5   3.44   0.687   39.274 < 2.2e-16 ***
## town                    4  61.71  15.427  881.287 < 2.2e-16 ***
## flat_type               6 684.85 114.142 6520.724 < 2.2e-16 ***
## storey_range            1  82.11  82.109 4690.712 < 2.2e-16 ***
## floor_area_sqm          1  11.39  11.393  650.856 < 2.2e-16 ***
## lease_commence_date     1 113.72 113.722 6496.732 < 2.2e-16 ***
## Residuals           11508 201.44   0.018
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

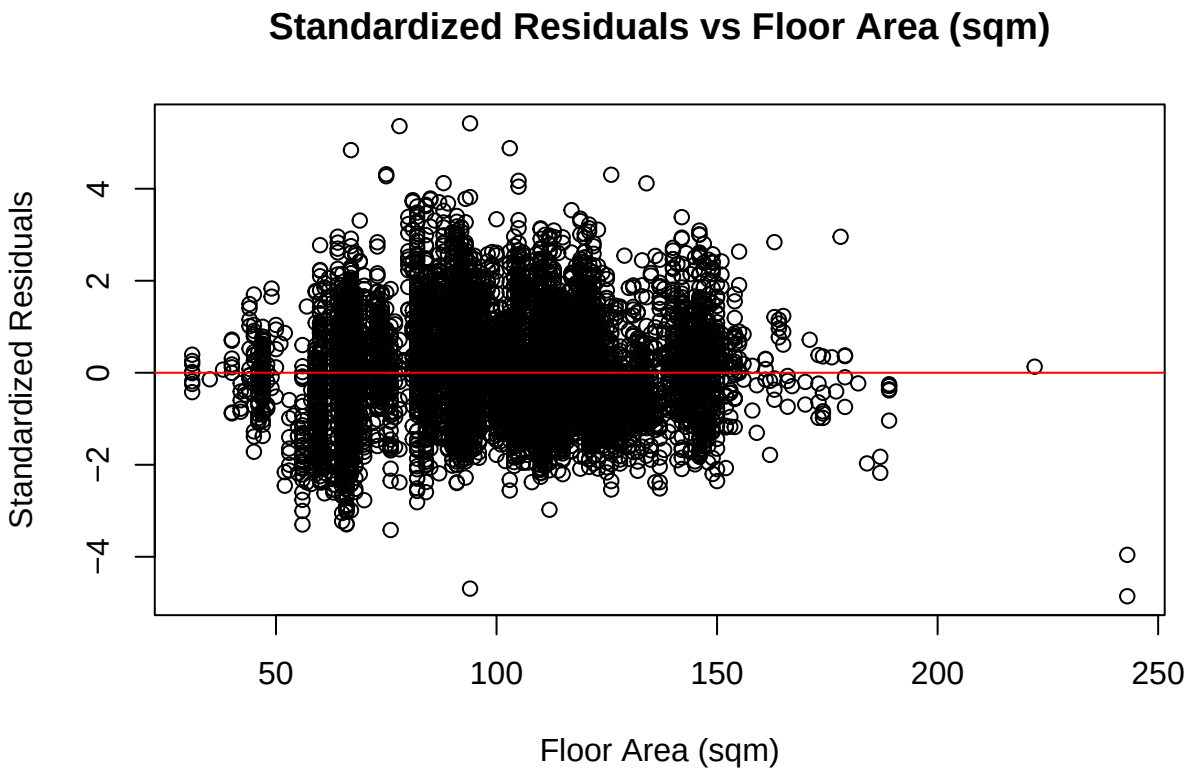
```r
# Standardized residuals
sr <- rstandard(M0)
fitted_vals <- fitted(M0)

# QQ plot
qqnorm(sr, main = "QQ Plot of Standardized Residuals", ylab = "Standardized Residuals")
qqline(sr, col = "red")
```



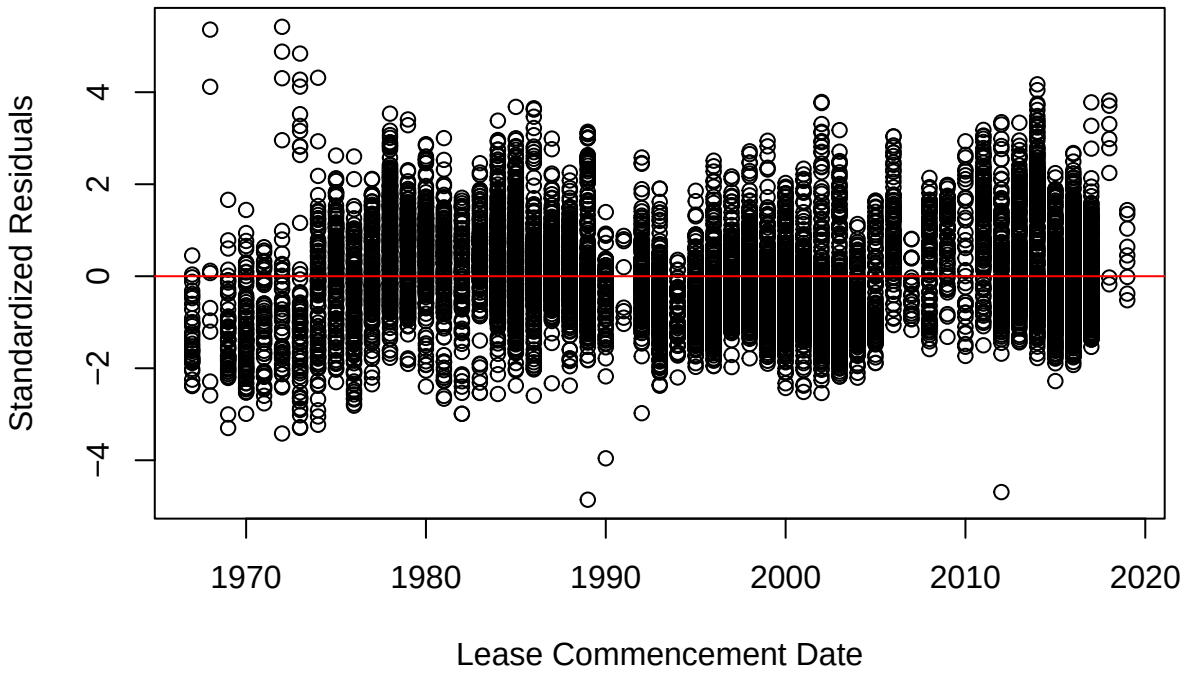

```r
# Residuals vs Fitted
plot(fitted_vals, sr,
     main = "Standardized Residuals vs Fitted Values",
     xlab = "Fitted Values",
     ylab = "Standardized Residuals")
abline(h = 0, col = "red")
```



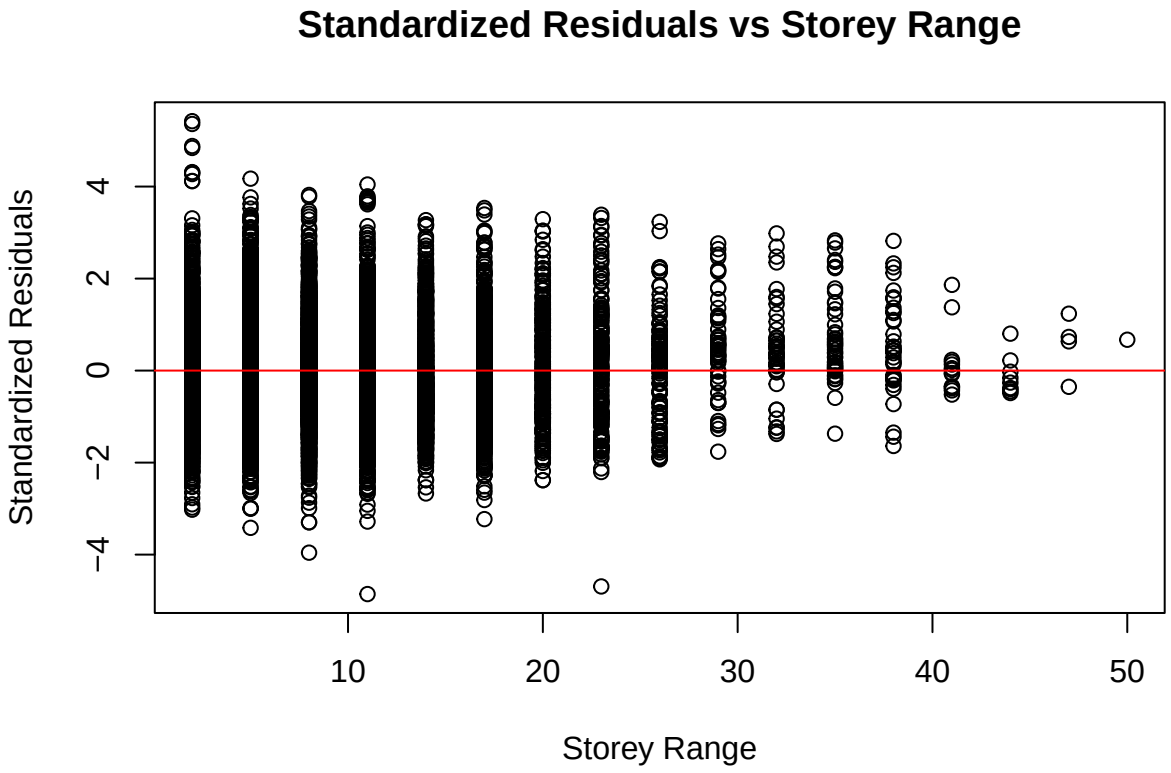

```r
# Residuals vs numeric predictors
plot(data$floor_area_sqm, sr,
     main = "Standardized Residuals vs Floor Area (sqm)",
     xlab = "Floor Area (sqm)",
     ylab = "Standardized Residuals")
abline(h = 0, col = "red")
```



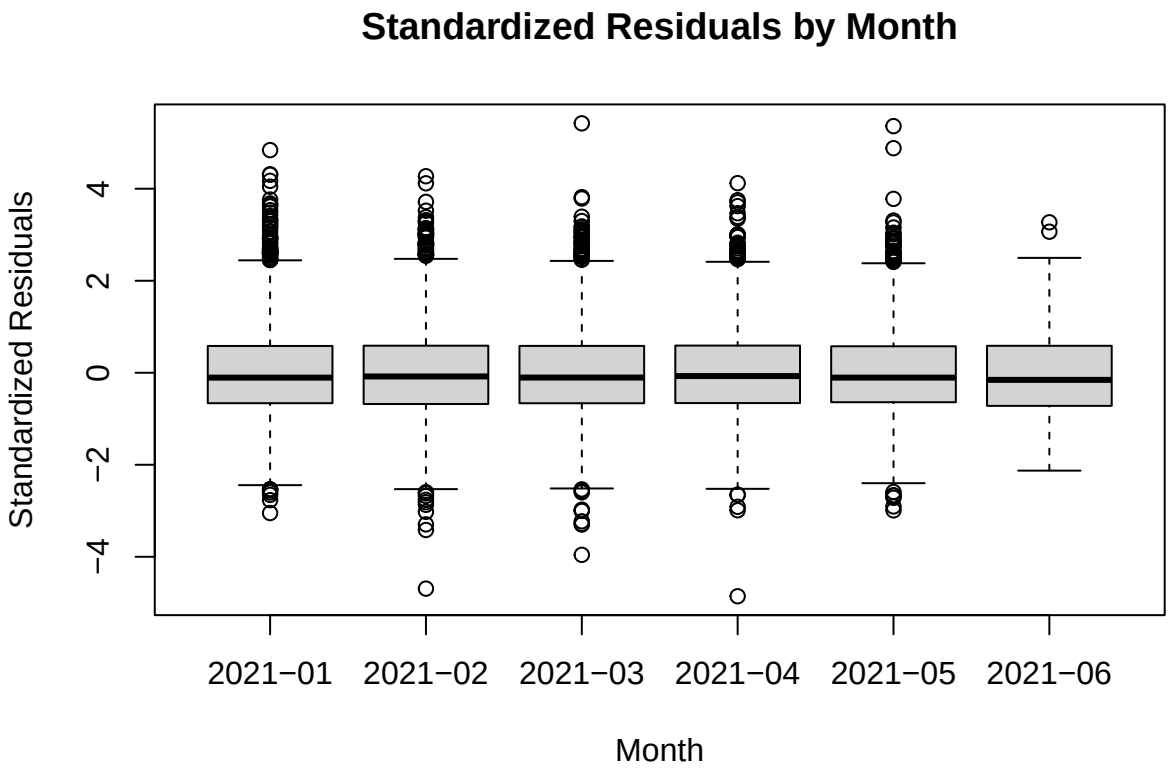

```
plot(data$lease_commence_date, sr,
     main = "Standardized Residuals vs Lease Commencement Date",
     xlab = "Lease Commencement Date",
     ylab = "Standardized Residuals")
abline(h = 0, col = "red")
```



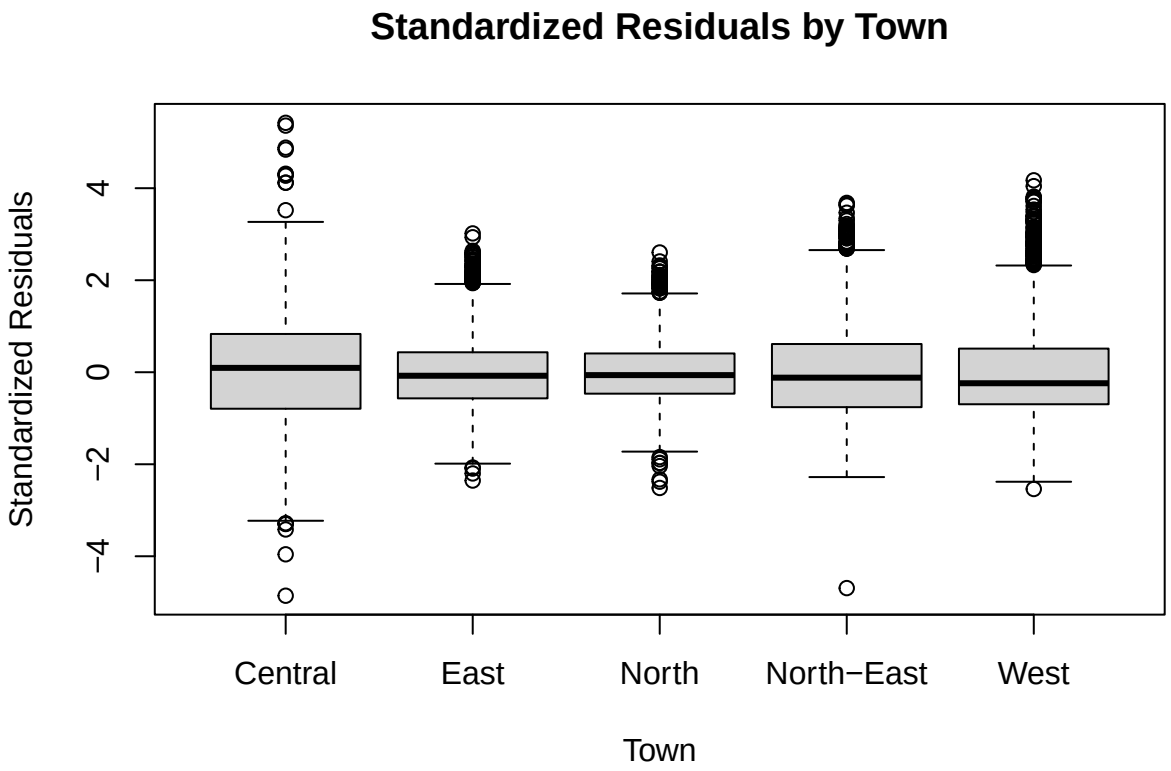
```r
plot(data$storey_range, sr,
     main = "Standardized Residuals vs Storey Range",
     xlab = "Storey Range",
     ylab = "Standardized Residuals")
abline(h = 0, col = "red")
```



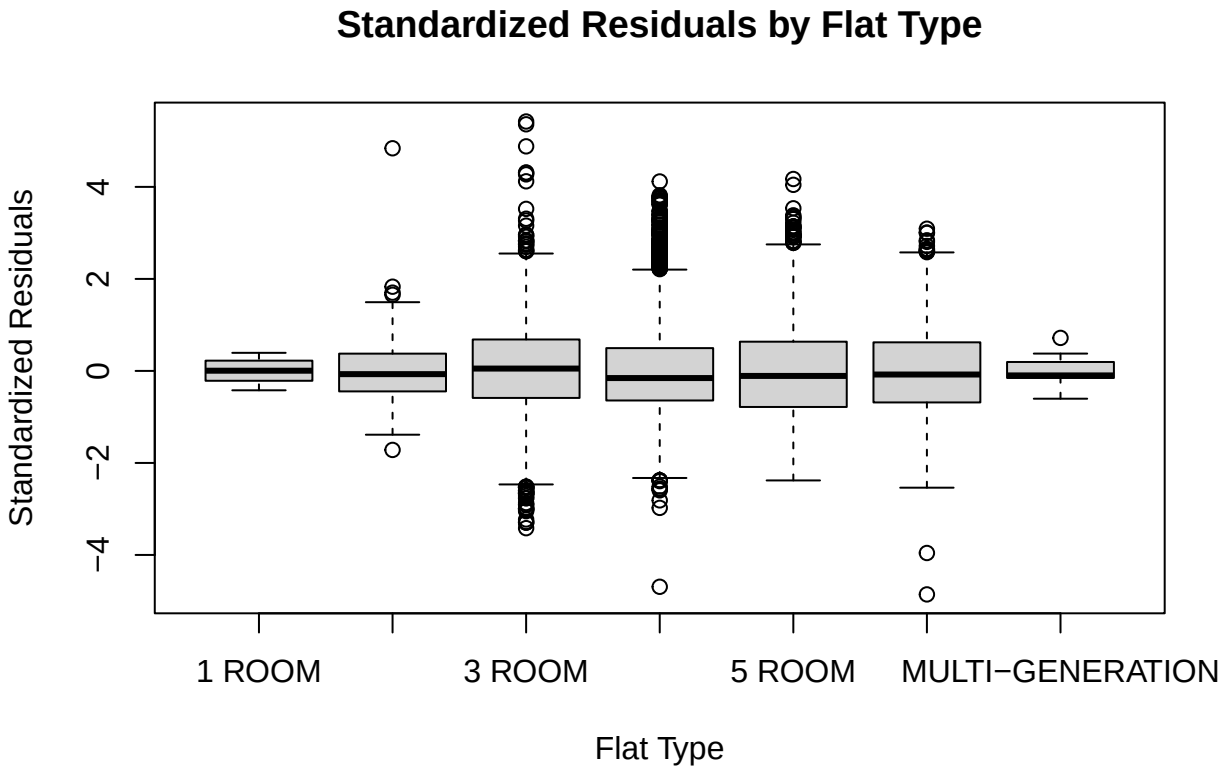

```r
# Residuals vs categorical variables using boxplots
boxplot(sr ~ data$month,
        main = "Standardized Residuals by Month",
        xlab = "Month",
        ylab = "Standardized Residuals")
```



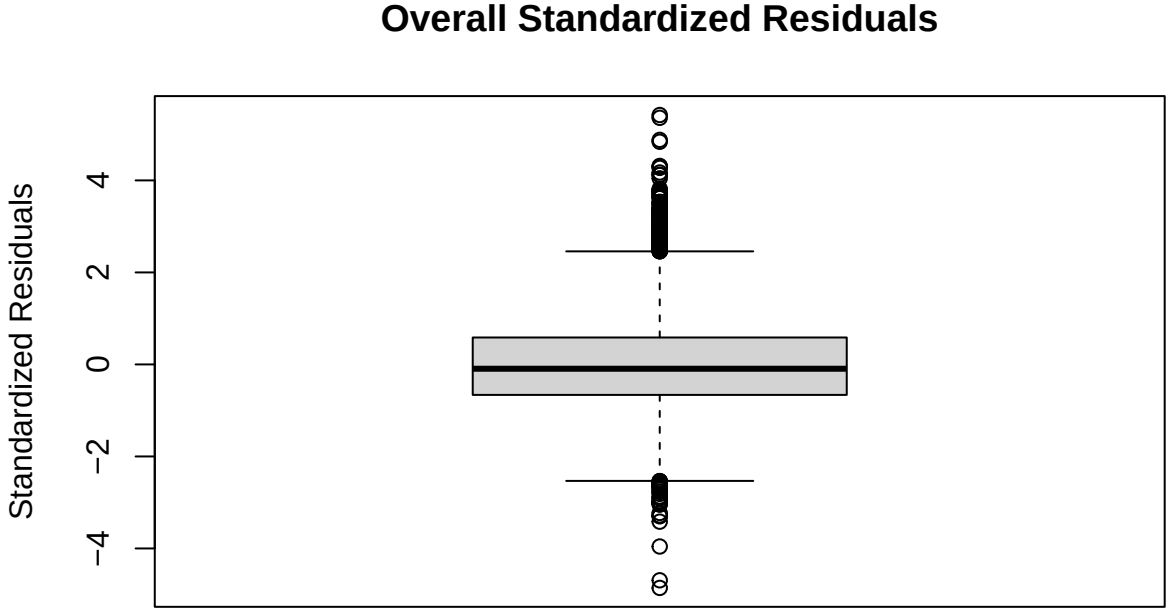

```
boxplot(sr ~ data$town,
        main = "Standardized Residuals by Town",
        xlab = "Town",
        ylab = "Standardized Residuals")
```



```r
boxplot(sr ~ data$flat_type,
        main = "Standardized Residuals by Flat Type",
        xlab = "Flat Type",
        ylab = "Standardized Residuals")
```




```r
# Overall boxplot to check outliers
boxplot(sr, main = "Overall Standardized Residuals", ylab = "Standardized Residuals")
```



```r
length(boxplot(sr, plot = FALSE)$out)
```

```
## [1] 240
```

```r
# Check for influential points
which(cooks.distance(M0) >= 1) # indices of influential points
```

```
## named integer(0)
```

```r
# Calculate VIF (Variance Inflation Factor) for each predictor
vif_values <- vif(M0)
print(vif_values)
```

```
##                           GVIF Df GVIF^(1/(2*Df))
## month                 1.006703  5        1.000668
## town                  1.314312  4        1.034754
## flat_type            13.366480  6        1.241180
## storey_range          1.198186  1        1.094617
## floor_area_sqm       11.468086  1        3.386456
## lease_commence_date   1.541270  1        1.241479
```

```r
data <- data %>%
  mutate(flat_type_grouped = case_when(
    flat_type %in% c("1 ROOM", "2 ROOM") ~ "Small",
    flat_type %in% c("3 ROOM", "4 ROOM") ~ "Medium",
    flat_type %in% c("5 ROOM", "EXECUTIVE", "MULTI-GENERATION") ~ "Large",
    TRUE ~ "Other"
  ))

M1 = lm(log(resale_price) ~ floor_area_sqm + lease_commence_date + town + storey_range + flat_type_grou
summary(M1)
```

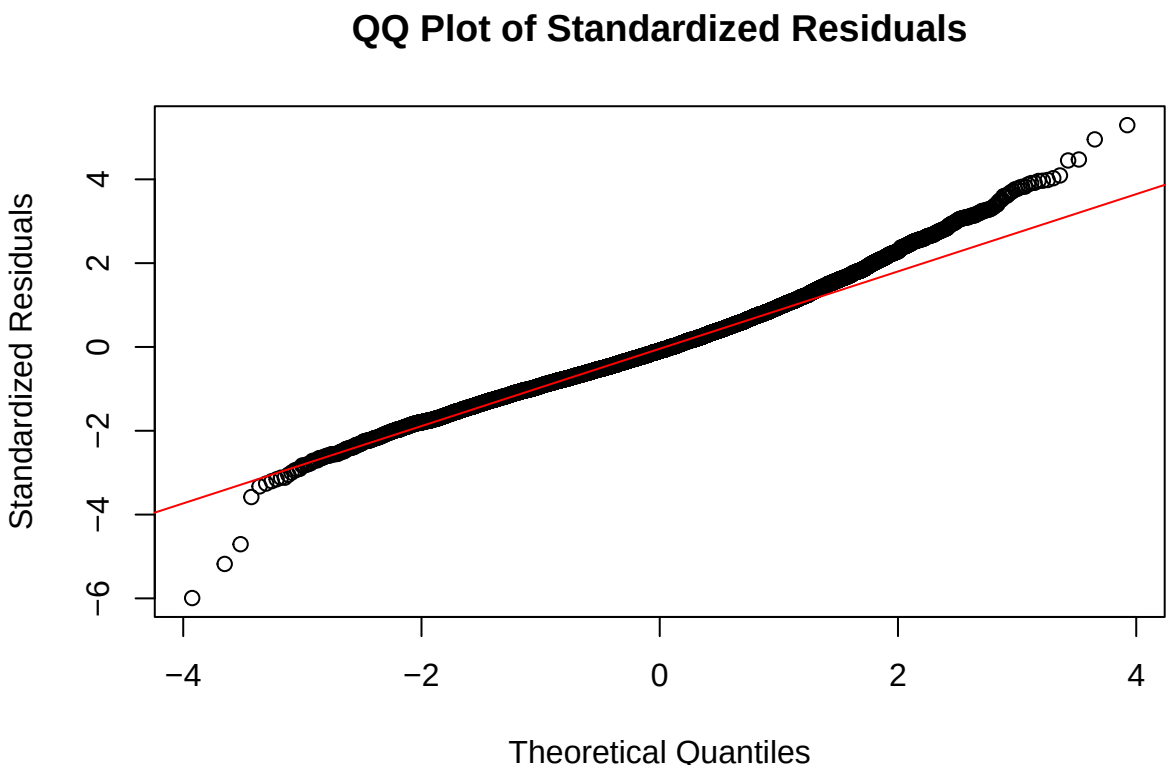
```
##
## Call:
## lm(formula = log(resale_price) ~ floor_area_sqm + lease_commence_date +
##     town + storey_range + flat_type_grouped, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.79934 -0.08901 -0.01213  0.07759  0.70855
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -5.718e+00  1.958e-01 -29.197   <2e-16 ***
## floor_area_sqm            9.473e-03  8.988e-05 105.399   <2e-16 ***
## lease_commence_date       9.011e-03  9.903e-05  90.992   <2e-16 ***
## townEast                 -2.009e-01  4.307e-03 -46.650   <2e-16 ***
## townNorth                -3.755e-01  4.691e-03 -80.051   <2e-16 ***
## townNorth-East           -2.929e-01  3.912e-03 -74.871   <2e-16 ***
## townWest                 -3.232e-01  4.040e-03 -79.998   <2e-16 ***
## storey_range              9.906e-03  2.164e-04  45.785   <2e-16 ***
## flat_type_groupedMedium   4.054e-03  4.264e-03   0.951    0.342
## flat_type_groupedSmall   -1.701e-01  1.264e-02 -13.464   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.1339 on 11517 degrees of freedom
## Multiple R-squared:  0.8218, Adjusted R-squared:  0.8216
## F-statistic:  5899 on 9 and 11517 DF,  p-value: < 2.2e-16
```

```
anova(M1)
```

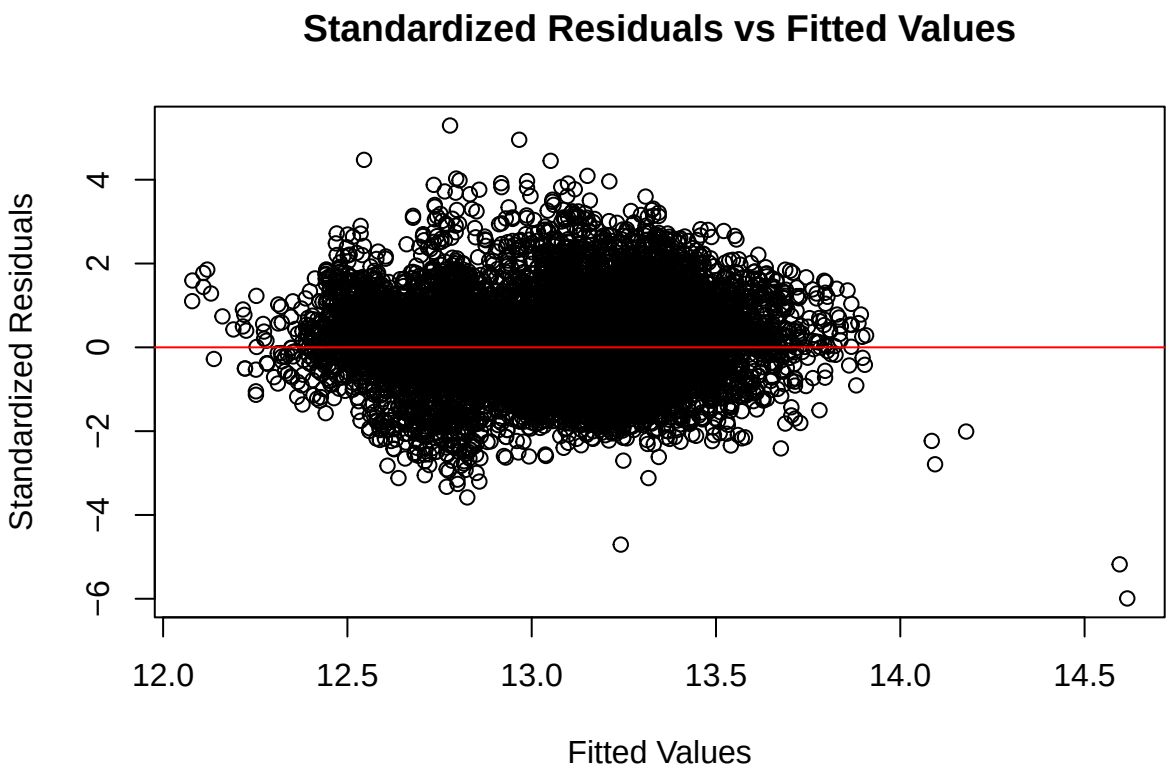
```
## Analysis of Variance Table
##
## Response: log(resale_price)
##                         Df Sum Sq Mean Sq F value    Pr(>F)
## floor_area_sqm           1 543.37  543.37 30300.9 < 2.2e-16 ***
## lease_commence_date      1 147.38  147.38  8218.3 < 2.2e-16 ***
## town                     4 217.98   54.49  3038.9 < 2.2e-16 ***
## storey_range             1  38.63   38.63  2154.3 < 2.2e-16 ***
## flat_type_grouped        2   4.77    2.39   133.1 < 2.2e-16 ***
## Residuals            11517 206.53    0.02
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Standardized residuals
sr <- rstandard(M1)
fitted_vals <- fitted(M1)

# QQ plot
qqnorm(sr, main = "QQ Plot of Standardized Residuals", ylab = "Standardized Residuals")
qqline(sr, col = "red")
```

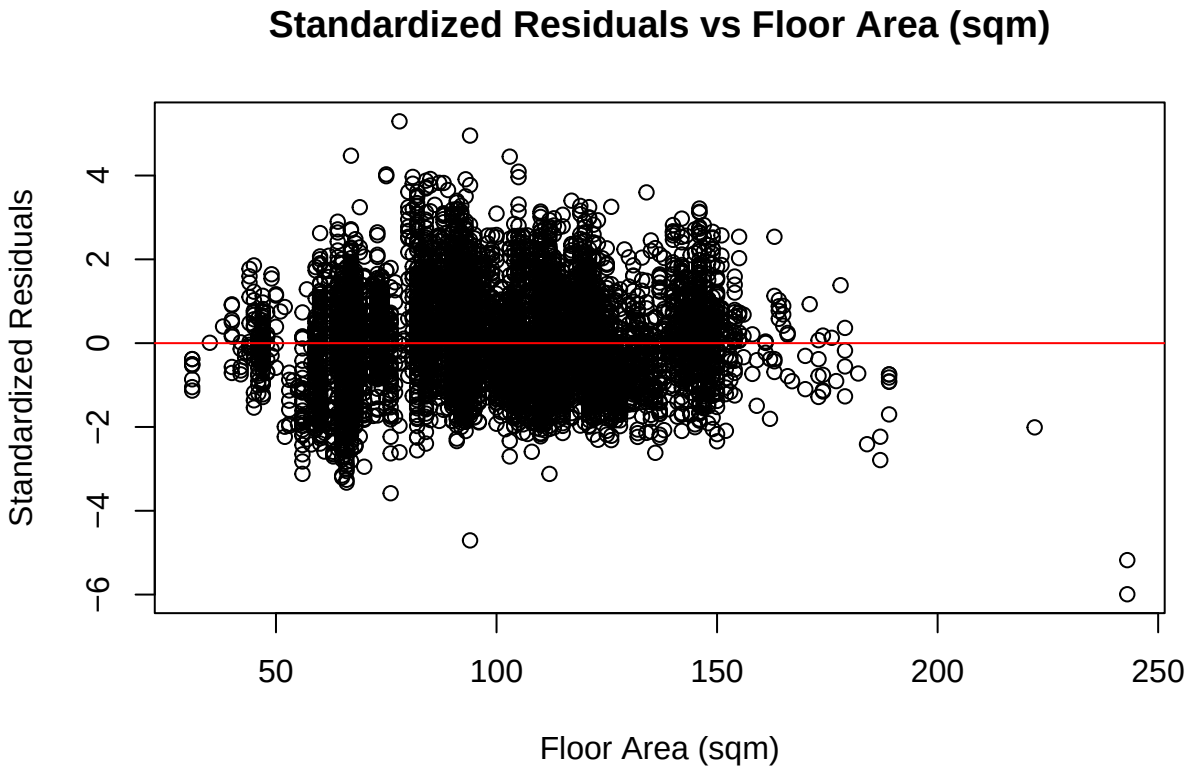



```
# Residuals vs Fitted
plot(fitted_vals, sr,
     main = "Standardized Residuals vs Fitted Values",
     xlab = "Fitted Values",
```

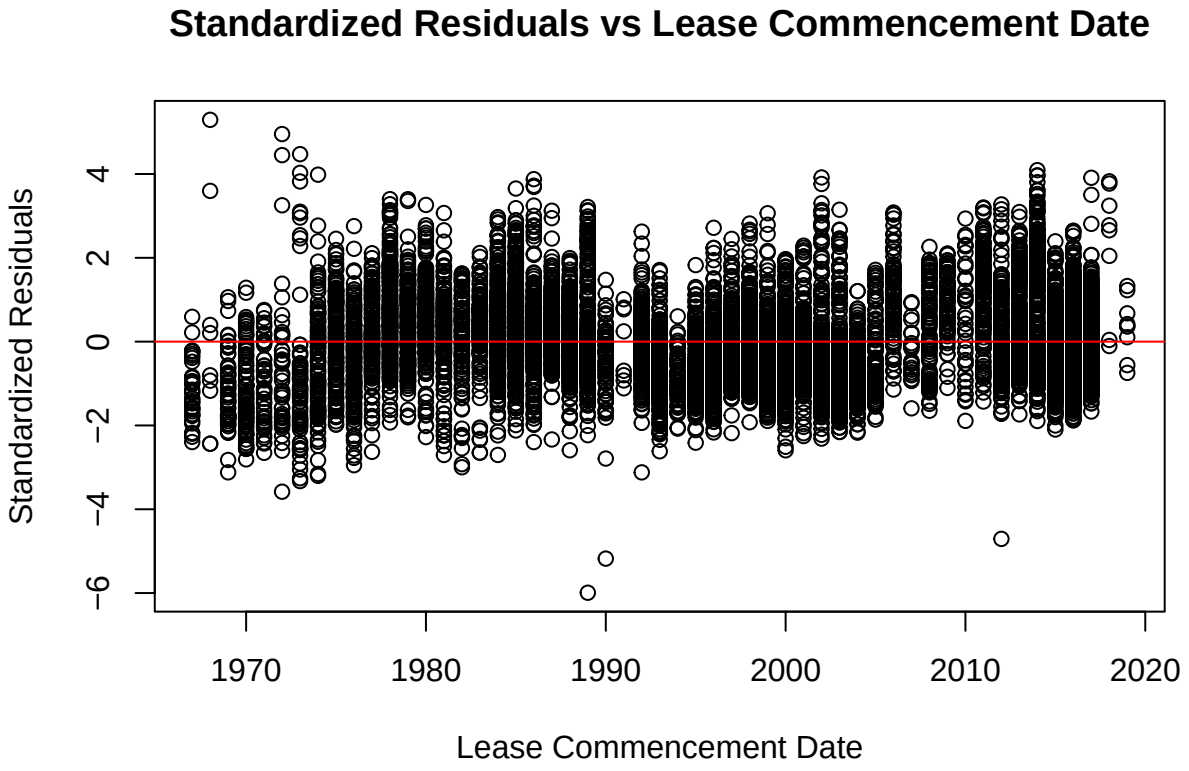
```
     ylab = "Standardized Residuals")
abline(h = 0, col = "red")
```



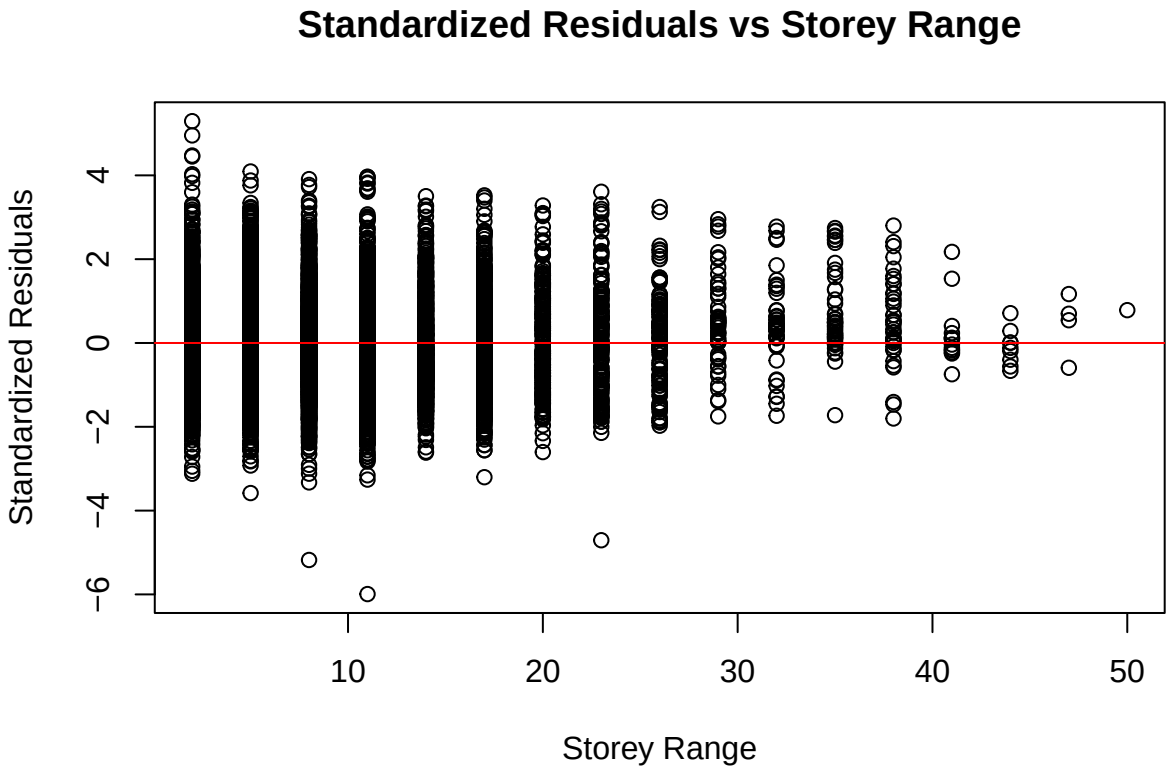

```
# Residuals vs numeric predictors
plot(data$floor_area_sqm, sr,
     main = "Standardized Residuals vs Floor Area (sqm)",
     xlab = "Floor Area (sqm)",
     ylab = "Standardized Residuals")
abline(h = 0, col = "red")
```



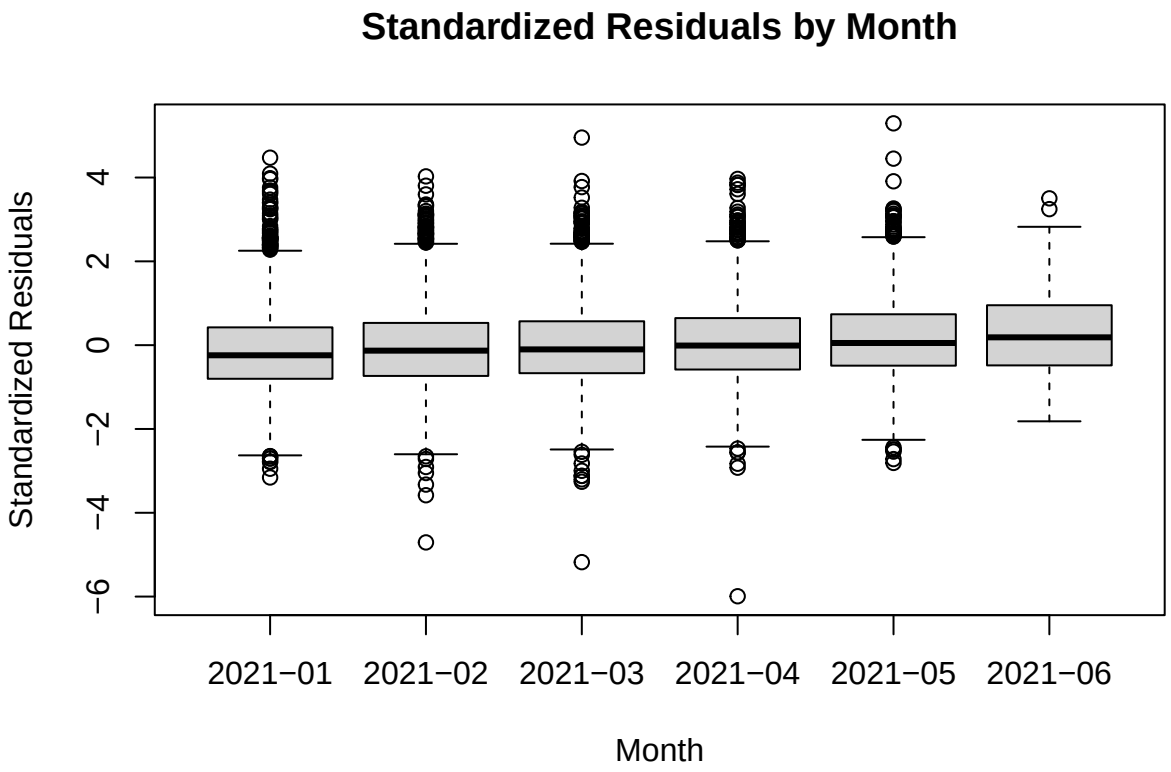

```
plot(data$lease_commence_date, sr,
     main = "Standardized Residuals vs Lease Commencement Date",
     xlab = "Lease Commencement Date",
     ylab = "Standardized Residuals")
abline(h = 0, col = "red")
```



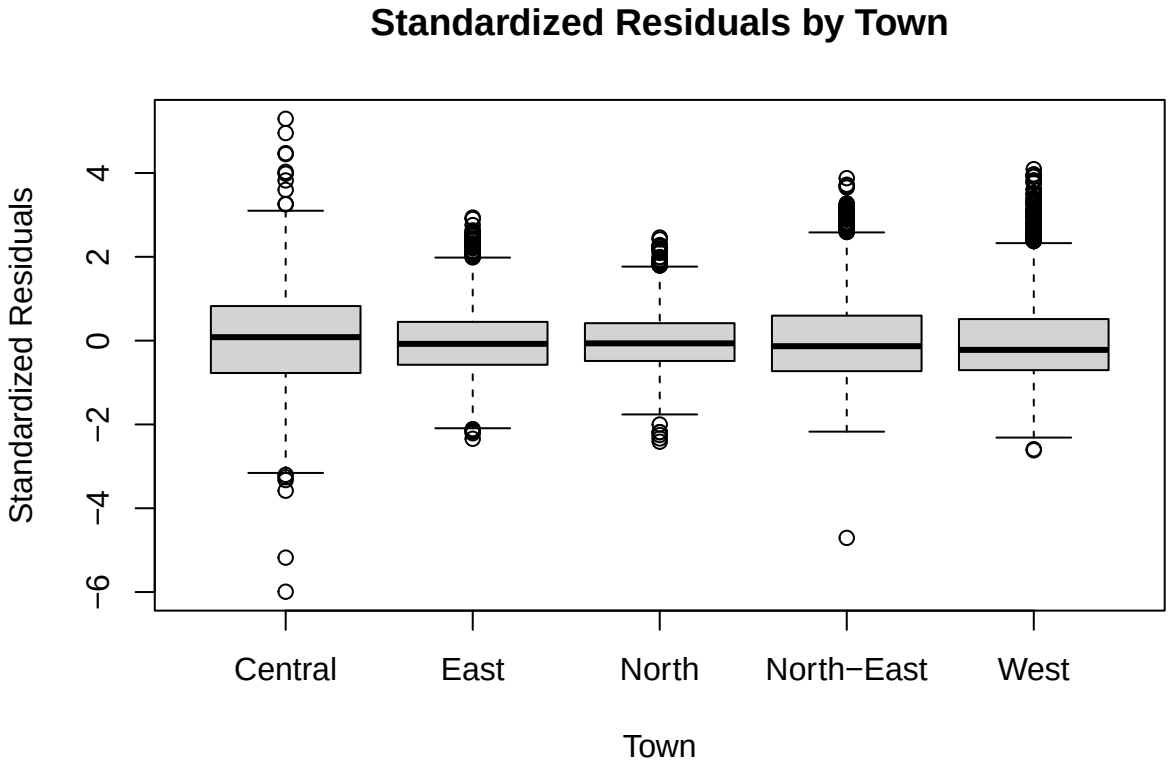
```r
plot(data$storey_range, sr,
     main = "Standardized Residuals vs Storey Range",
     xlab = "Storey Range",
     ylab = "Standardized Residuals")
abline(h = 0, col = "red")
```



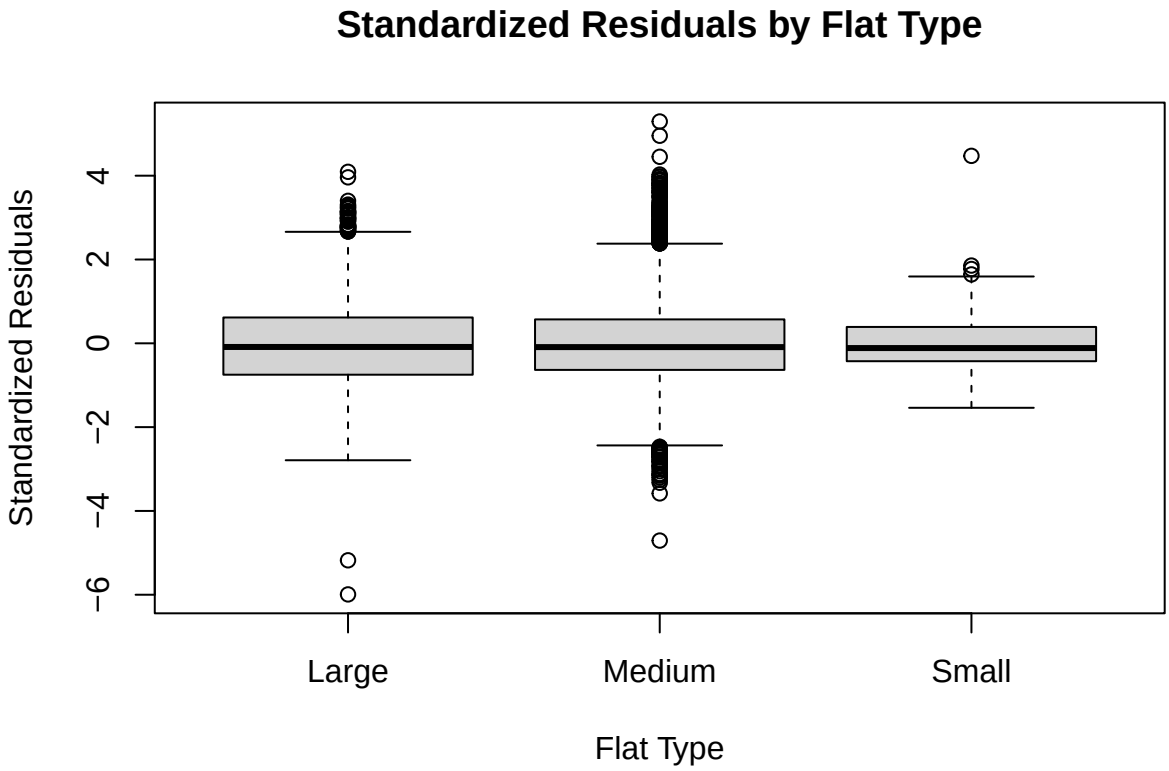

```r
# Residuals vs categorical variables using boxplots
boxplot(sr ~ data$month,
        main = "Standardized Residuals by Month",
        xlab = "Month",
        ylab = "Standardized Residuals")
```



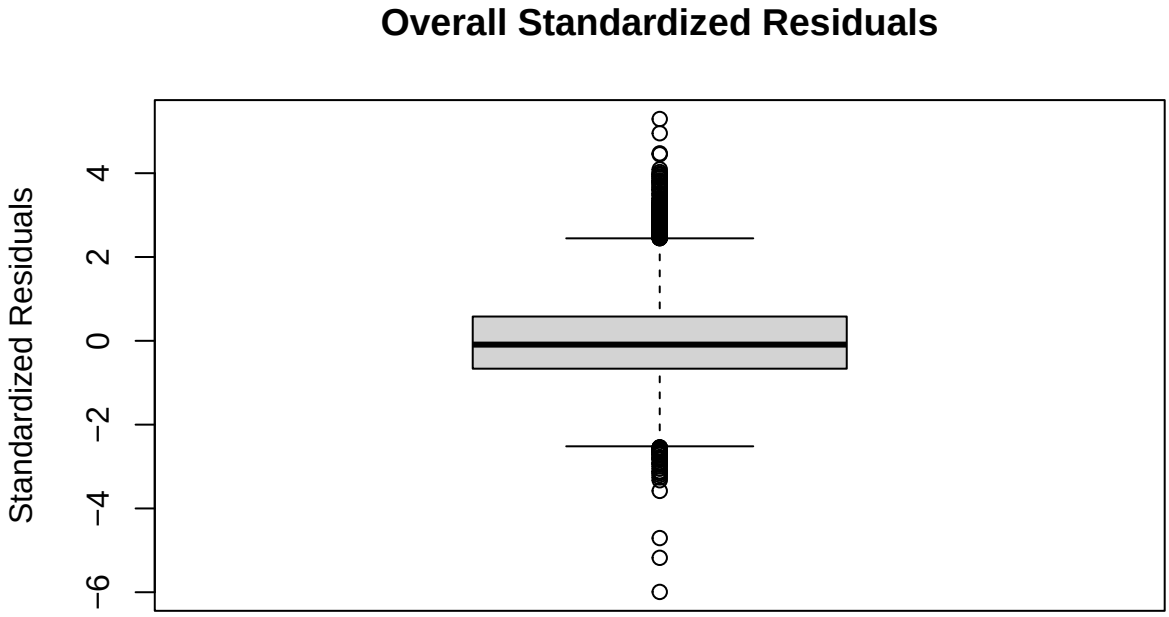

```
boxplot(sr ~ data$town,
        main = "Standardized Residuals by Town",
        xlab = "Town",
        ylab = "Standardized Residuals")
```



```r
boxplot(sr ~ data$flat_type_grouped,
        main = "Standardized Residuals by Flat Type",
        xlab = "Flat Type",
        ylab = "Standardized Residuals")
```




```r
# Overall boxplot to check outliers
boxplot(sr, main = "Overall Standardized Residuals", ylab = "Standardized Residuals")
```



```r
length(boxplot(sr, plot = FALSE)$out)
```

```
## [1] 254
```

```r
# Check for influential points
which(cooks.distance(M1) >= 1) # indices of influential points
```

```
## named integer(0)
```

```r
# Calculate VIF (Variance Inflation Factor) for each predictor
vif_values <- vif(M1)
print(vif_values)
```

```
##                         GVIF Df GVIF^(1/(2*Df))
## floor_area_sqm      2.907419  1        1.705116
## lease_commence_date 1.305075  1        1.142399
## town                1.293871  4        1.032729
## storey_range        1.191307  1        1.091470
## flat_type_grouped   2.831934  2        1.297241
```

```r
# Regroup into Large and Small-Medium
data <- data %>%
  mutate(flat_type_grouped = case_when(
    flat_type %in% c("1 ROOM", "2 ROOM", "3 ROOM", "4 ROOM") ~ "Small-Medium",
    flat_type %in% c("5 ROOM", "EXECUTIVE", "MULTI-GENERATION") ~"Large",
    TRUE ~ "Other"
  ))

# Fit the model
M2 <- lm(log(resale_price) ~ floor_area_sqm + lease_commence_date + town + storey_range + flat_type_grou

# Check summary and ANOVA
summary(M2)
```

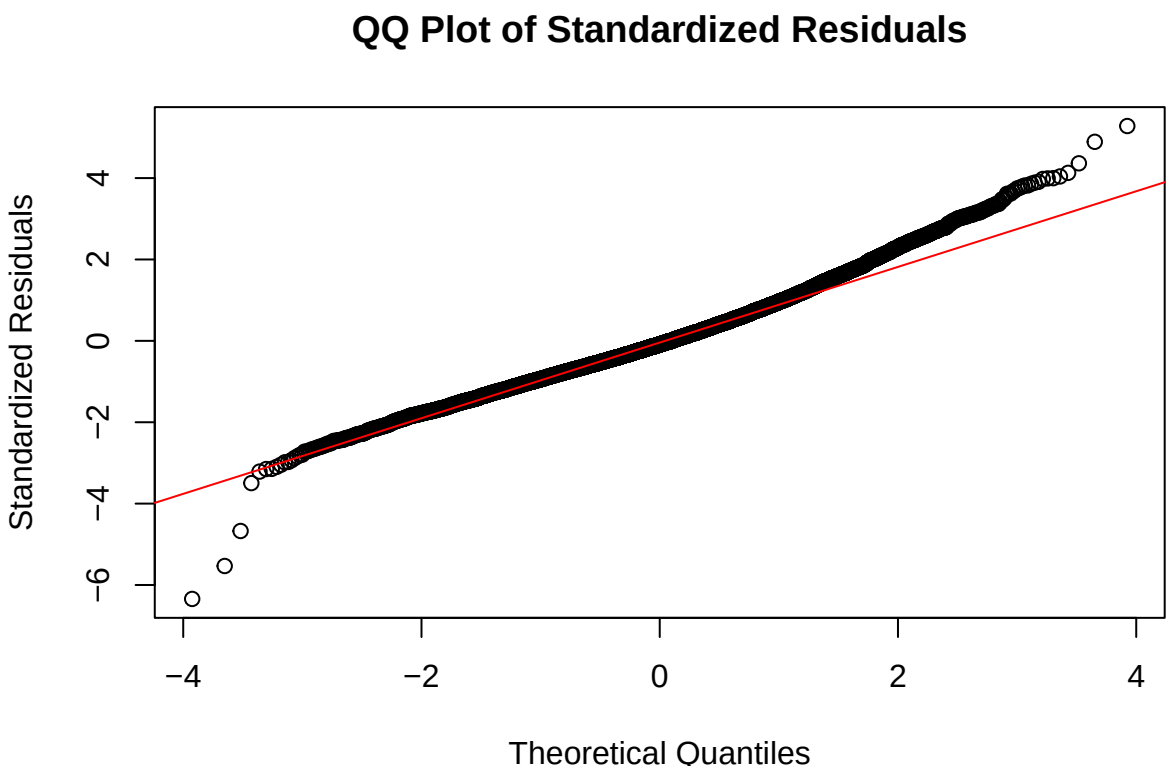
```
##
## Call:
## lm(formula = log(resale_price) ~ floor_area_sqm + lease_commence_date +
##     town + storey_range + flat_type_grouped, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.85543 -0.09048 -0.01244  0.07916  0.71420
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -5.606e+00  1.978e-01 -28.341  < 2e-16 ***
## floor_area_sqm                 9.933e-03  8.592e-05 115.603  < 2e-16 ***
## lease_commence_date            8.926e-03  9.994e-05  89.313  < 2e-16 ***
## townEast                      -2.018e-01  4.352e-03 -46.376  < 2e-16 ***
## townNorth                     -3.761e-01  4.741e-03 -79.334  < 2e-16 ***
## townNorth-East                -2.921e-01  3.954e-03 -73.877  < 2e-16 ***
## townWest                      -3.235e-01  4.083e-03 -79.238  < 2e-16 ***
## storey_range                   1.006e-02  2.184e-04  46.059  < 2e-16 ***
## flat_type_groupedSmall-Medium  1.775e-02  4.219e-03   4.206 2.62e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.1353 on 11518 degrees of freedom
## Multiple R-squared:  0.8179, Adjusted R-squared:  0.8178
## F-statistic:  6467 on 8 and 11518 DF,  p-value: < 2.2e-16
```

```r
anova(M2)
```

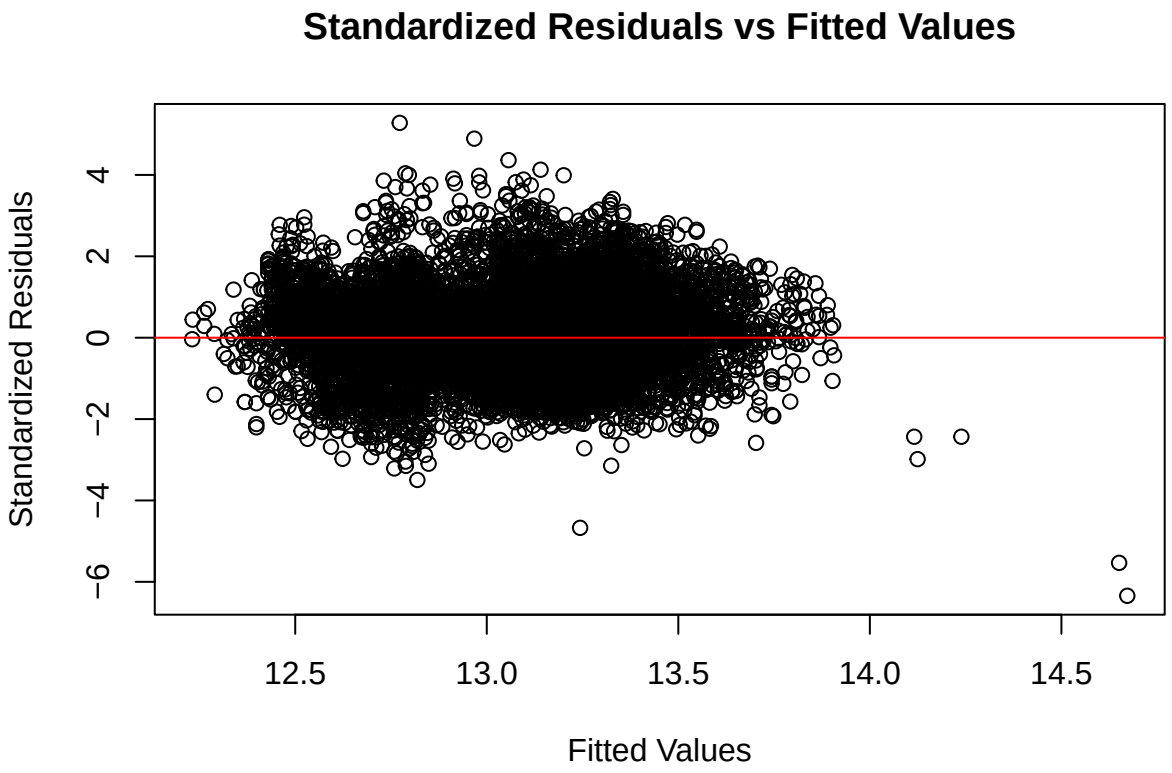
```
## Analysis of Variance Table
##
## Response: log(resale_price)
##                         Df Sum Sq Mean Sq   F value    Pr(>F)
## floor_area_sqm           1 543.37  543.37 29664.410 < 2.2e-16 ***
## lease_commence_date      1 147.38  147.38  8045.700 < 2.2e-16 ***
## town                     4 217.98   54.49  2975.027 < 2.2e-16 ***
## storey_range             1  38.63   38.63  2109.077 < 2.2e-16 ***
## flat_type_grouped        1   0.32    0.32    17.692 2.617e-05 ***
## Residuals            11518 210.98    0.02
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Standardized residuals
sr <- rstandard(M2)
fitted_vals <- fitted(M2)

# QQ plot
qqnorm(sr, main = "QQ Plot of Standardized Residuals", ylab = "Standardized Residuals")
qqline(sr, col = "red")
```

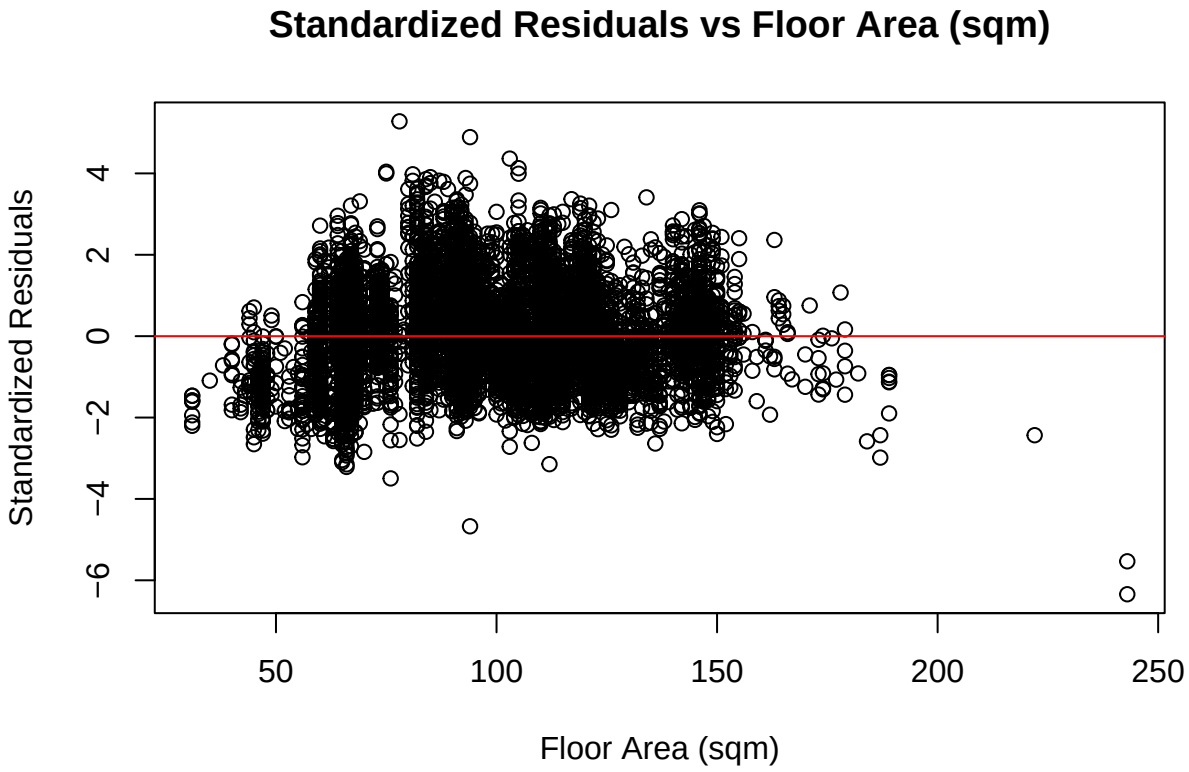



```r
# Residuals vs Fitted
plot(fitted_vals, sr,
     main = "Standardized Residuals vs Fitted Values",
```

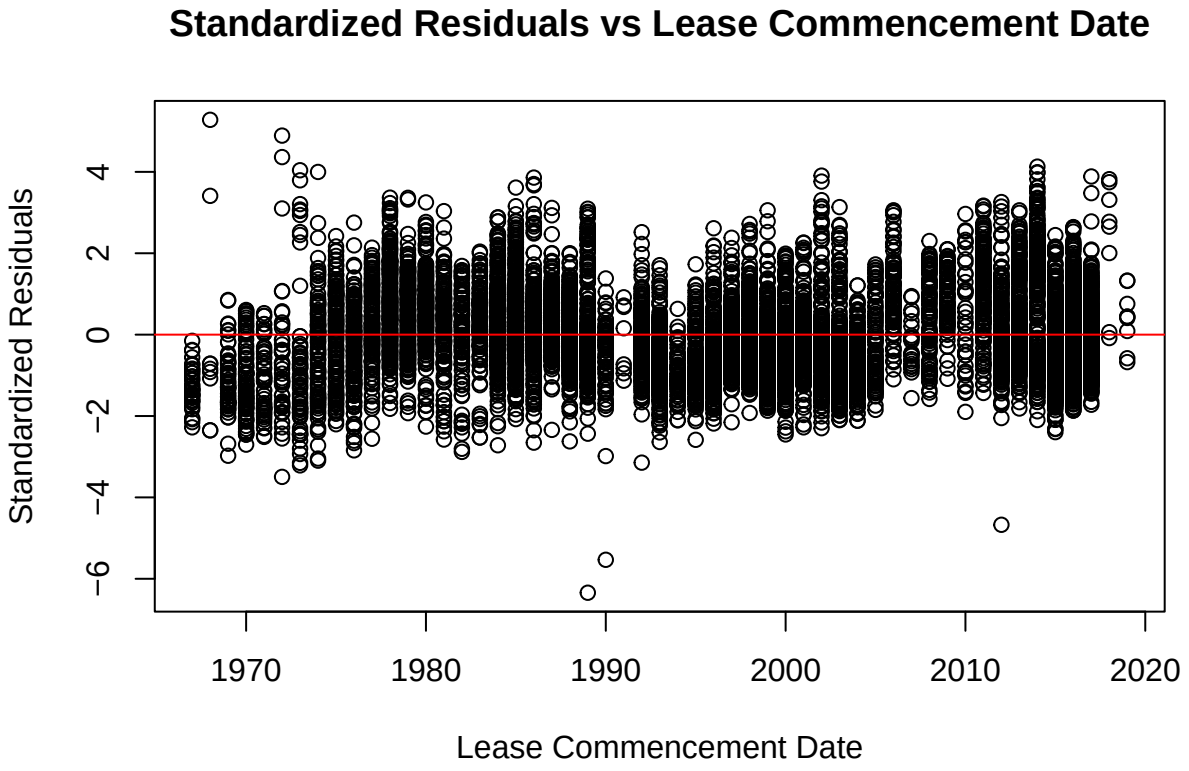
```
     xlab = "Fitted Values",
     ylab = "Standardized Residuals")
abline(h = 0, col = "red")
```



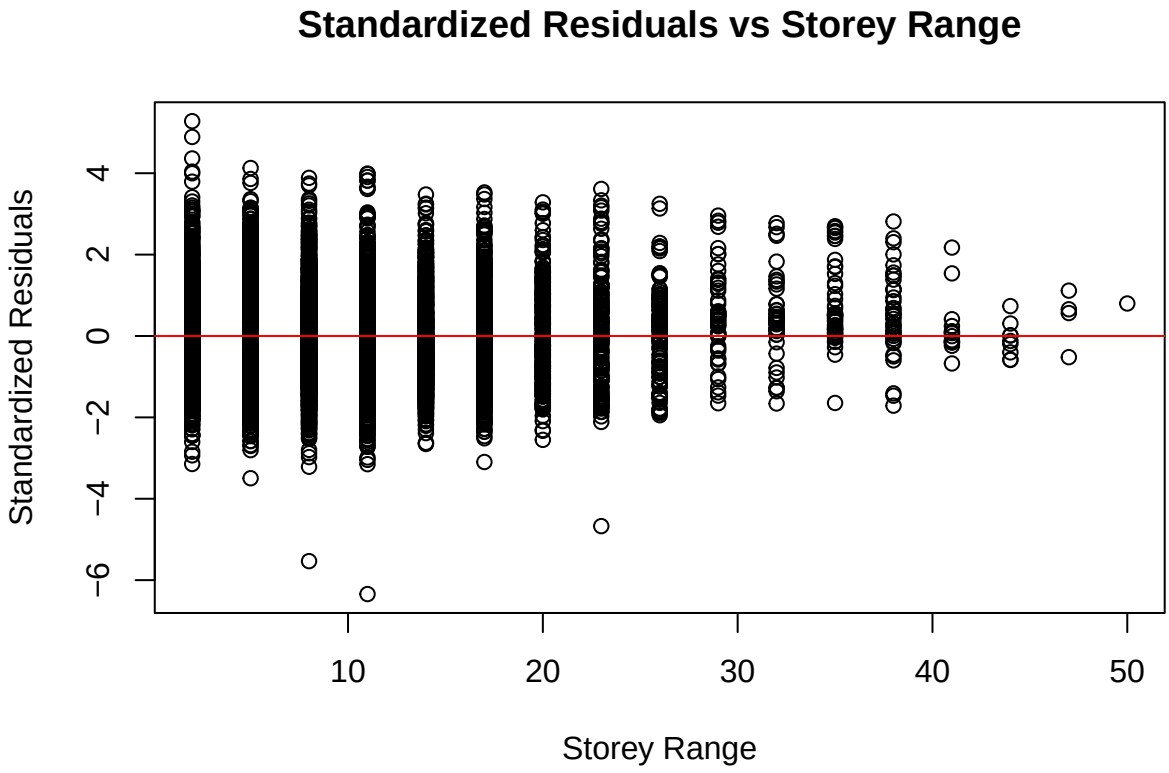

```
# Residuals vs numeric predictors
plot(data$floor_area_sqm, sr,
     main = "Standardized Residuals vs Floor Area (sqm)",
     xlab = "Floor Area (sqm)",
     ylab = "Standardized Residuals")
abline(h = 0, col = "red")
```



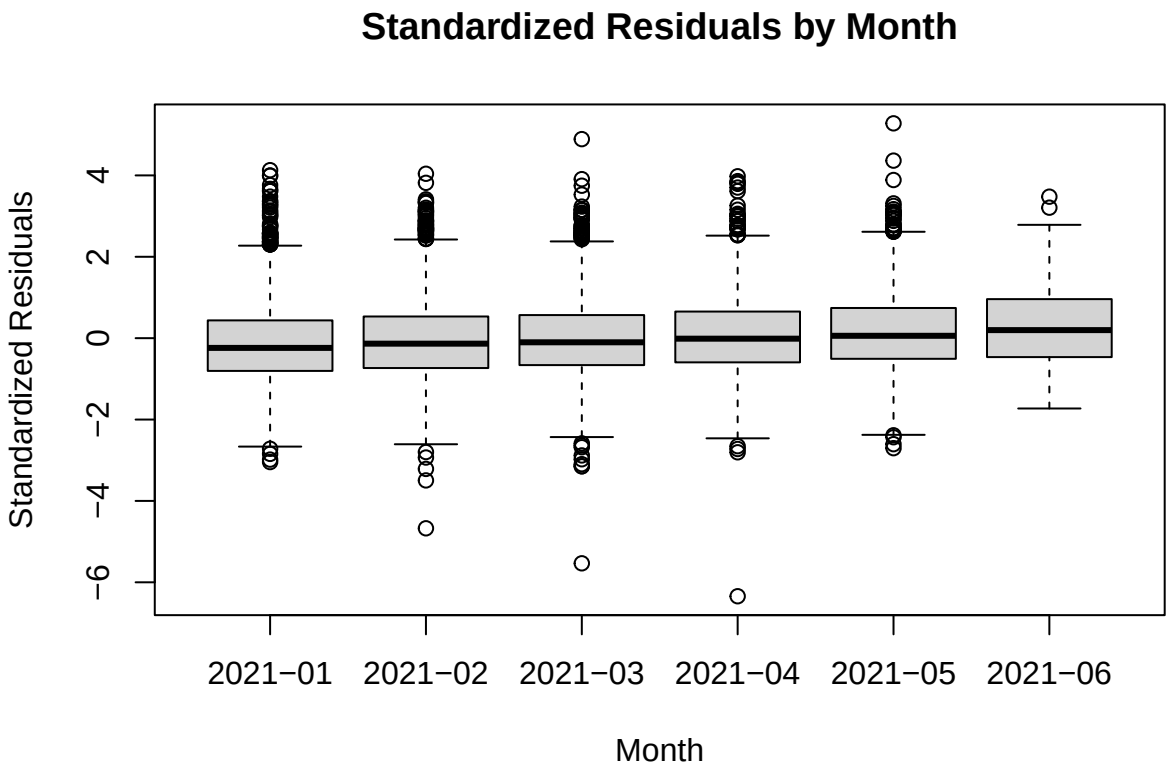

```
plot(data$lease_commence_date, sr,
     main = "Standardized Residuals vs Lease Commencement Date",
     xlab = "Lease Commencement Date",
     ylab = "Standardized Residuals")
abline(h = 0, col = "red")
```



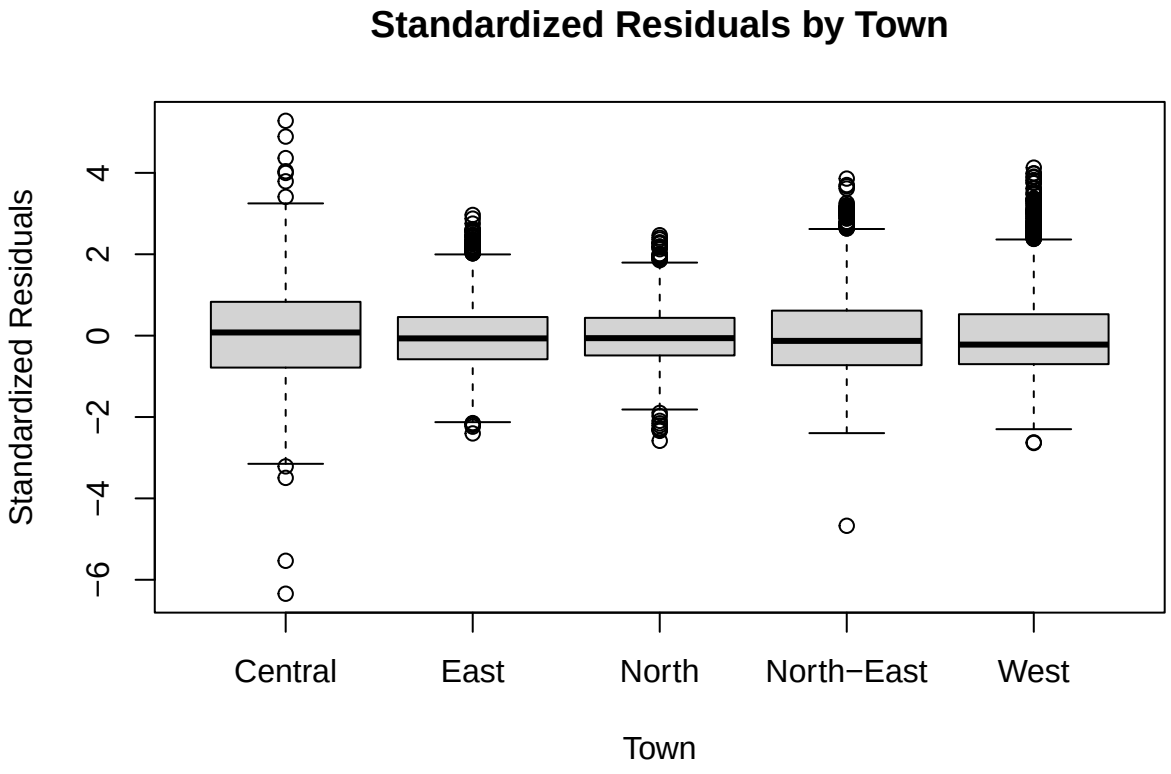
```r
plot(data$storey_range, sr,
     main = "Standardized Residuals vs Storey Range",
     xlab = "Storey Range",
     ylab = "Standardized Residuals")
abline(h = 0, col = "red")
```



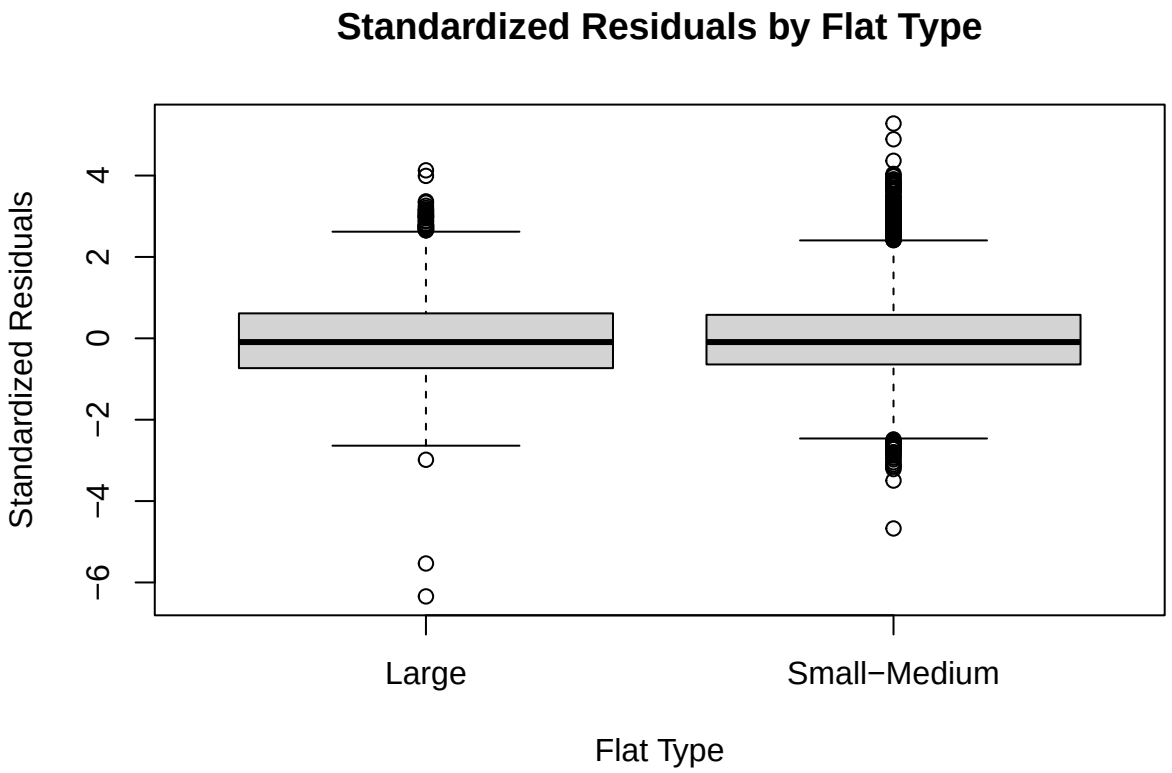

```r
# Residuals vs categorical variables using boxplots
boxplot(sr ~ data$month,
        main = "Standardized Residuals by Month",
        xlab = "Month",
        ylab = "Standardized Residuals")
```



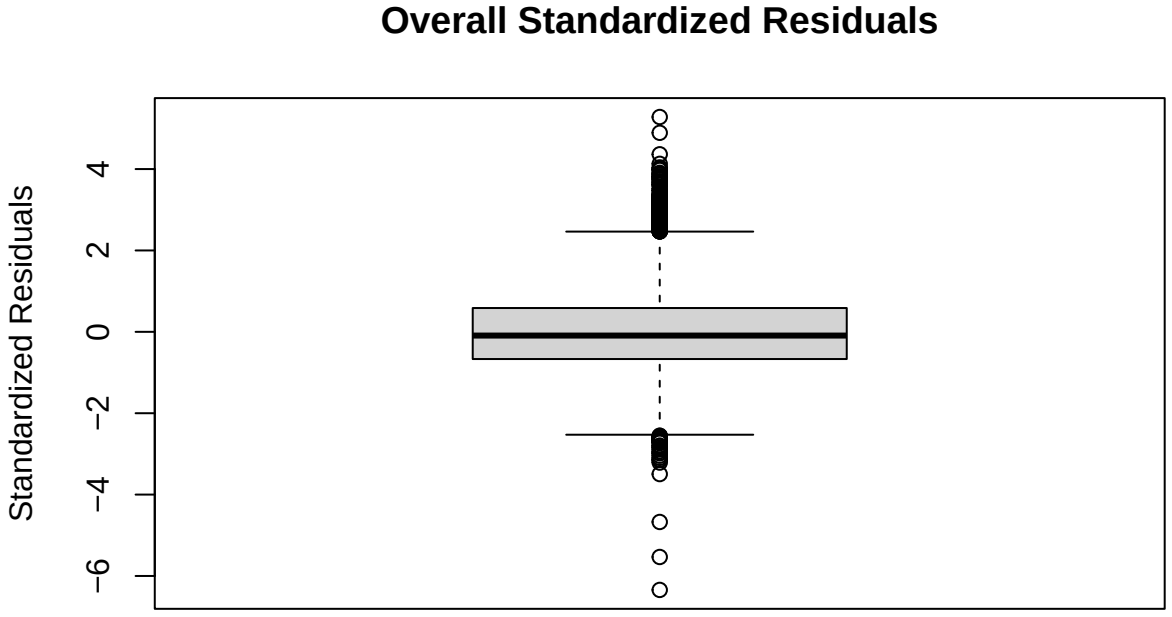

```
boxplot(sr ~ data$town,
        main = "Standardized Residuals by Town",
        xlab = "Town",
        ylab = "Standardized Residuals")
```



```r
boxplot(sr ~ data$flat_type_grouped,
        main = "Standardized Residuals by Flat Type",
        xlab = "Flat Type",
        ylab = "Standardized Residuals")
```




```r
# Overall boxplot to check outliers
boxplot(sr, main = "Overall Standardized Residuals", ylab = "Standardized Residuals")
```



```r
length(boxplot(sr, plot = FALSE)$out)
```

```
## [1] 223
```

```r
# Check for influential points
which(cooks.distance(M2) >= 1) # indices of influential points
```

```
## named integer(0)
```

```r
# Calculate VIF (Variance Inflation Factor) for each predictor
vif_values <- vif(M2)
print(vif_values)
```

```
##                         GVIF Df GVIF^(1/(2*Df))
## floor_area_sqm      2.601164  1        1.612812
## lease_commence_date 1.301180  1        1.140693
## town                1.292592  4        1.032601
## storey_range        1.188840  1        1.090339
## flat_type_grouped   2.528199  1        1.590031
```

```r
# Fit the model
M3 <- lm(log(resale_price) ~ floor_area_sqm + lease_commence_date + town + I(storey_range^1.34) + flat_

# Check summary and ANOVA
summary(M3)
```

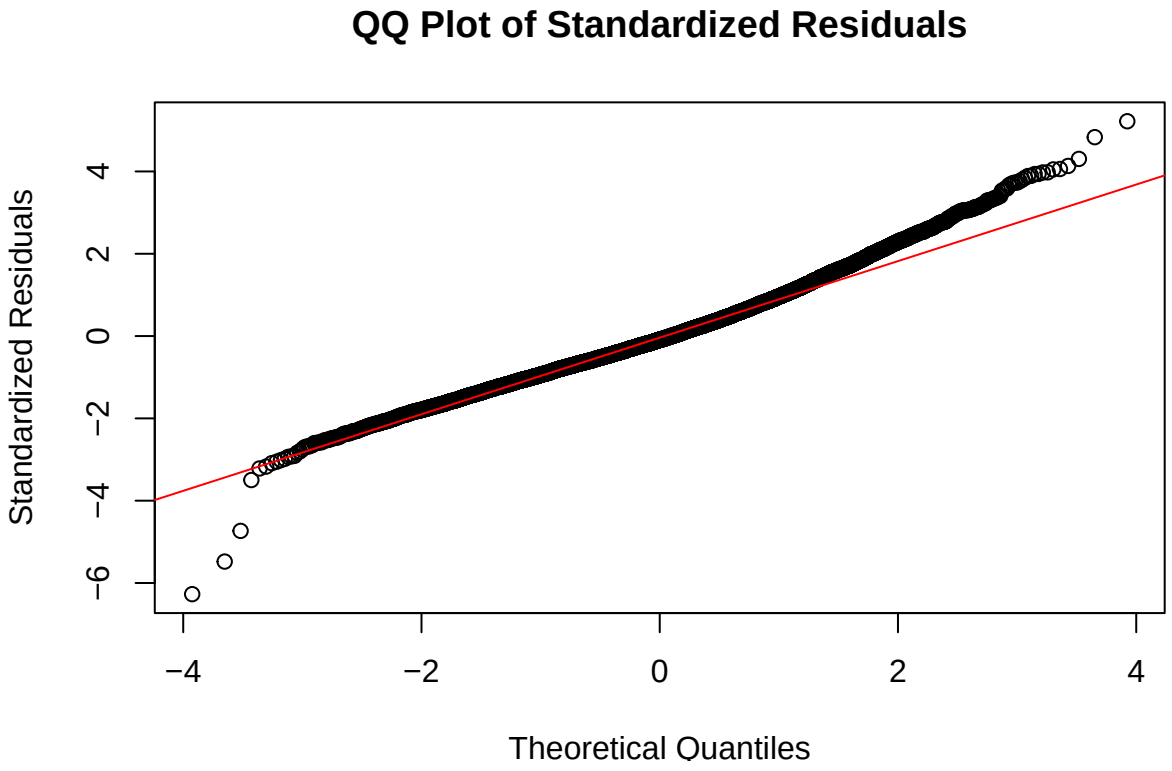
```
##
## Call:
## lm(formula = log(resale_price) ~ floor_area_sqm + lease_commence_date +
##     town + I(storey_range^1.34) + flat_type_grouped, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.84362 -0.08996 -0.01200  0.07936  0.70429
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -5.502e+00  1.977e-01 -27.828  < 2e-16 ***
## floor_area_sqm                 9.919e-03  8.568e-05 115.776  < 2e-16 ***
## lease_commence_date            8.885e-03  9.981e-05  89.022  < 2e-16 ***
## townEast                      -1.984e-01  4.354e-03 -45.573  < 2e-16 ***
## townNorth                     -3.724e-01  4.744e-03 -78.490  < 2e-16 ***
## townNorth-East                -2.870e-01  3.965e-03 -72.386  < 2e-16 ***
## townWest                      -3.204e-01  4.082e-03 -78.487  < 2e-16 ***
## I(storey_range^1.34)           3.190e-03  6.809e-05  46.844  < 2e-16 ***
## flat_type_groupedSmall-Medium  1.706e-02  4.207e-03   4.055 5.04e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.135 on 11518 degrees of freedom
## Multiple R-squared:  0.8189, Adjusted R-squared:  0.8188
## F-statistic:  6509 on 8 and 11518 DF,  p-value: < 2.2e-16
```

```r
anova(M3)
```

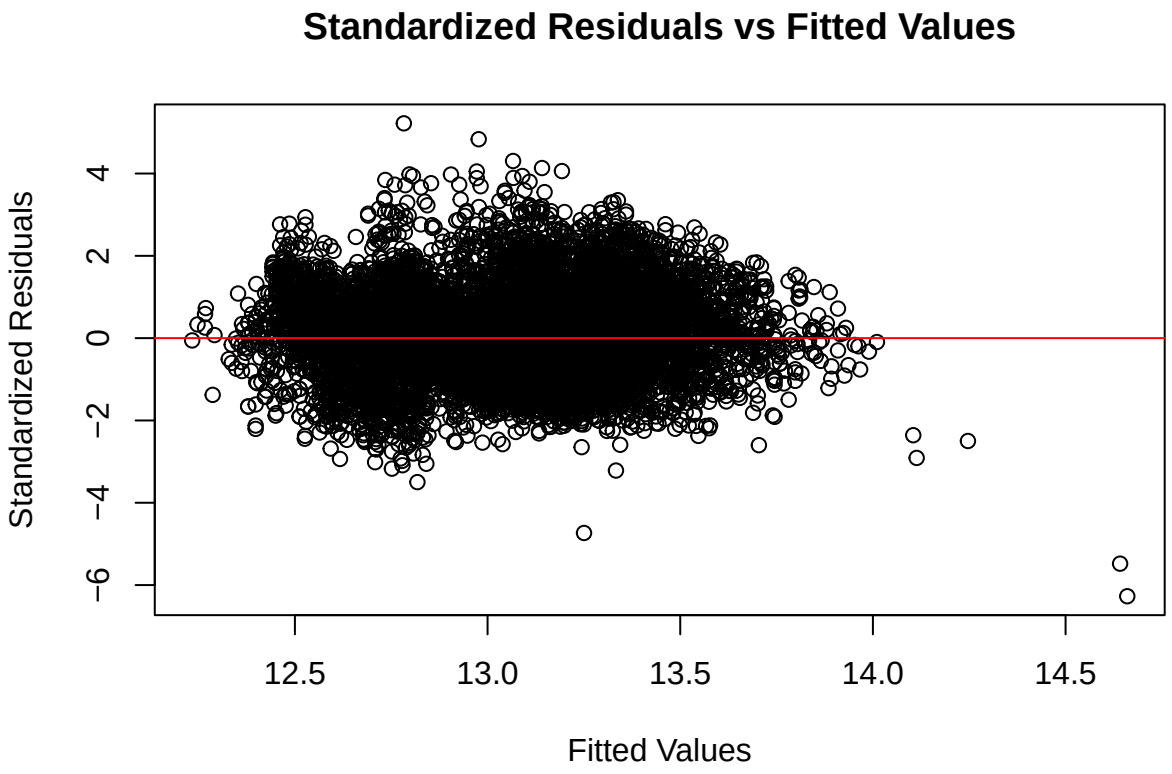
```
## Analysis of Variance Table
##
```

```
## Response: log(resale_price)
##                          Df Sum Sq Mean Sq   F value    Pr(>F)
## floor_area_sqm            1 543.37  543.37 29823.103 < 2.2e-16 ***
## lease_commence_date       1 147.38  147.38  8088.741 < 2.2e-16 ***
## town                      4 217.98   54.49  2990.943 < 2.2e-16 ***
## I(storey_range^1.34)      1  39.78   39.78  2183.317 < 2.2e-16 ***
## flat_type_grouped         1   0.30    0.30    16.447 5.037e-05 ***
## Residuals             11518 209.86    0.02
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

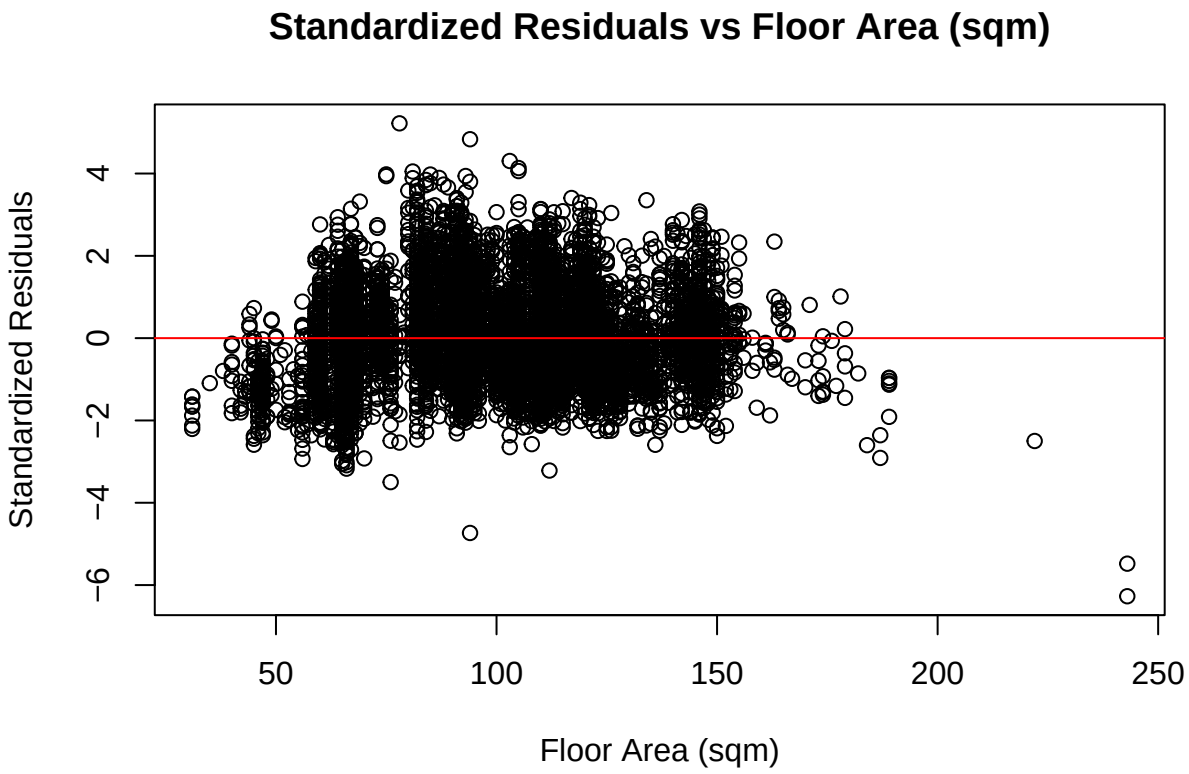
```r
# Standardized residuals
sr <- rstandard(M3)
fitted_vals <- fitted(M3)

# QQ plot
qqnorm(sr, main = "QQ Plot of Standardized Residuals", ylab = "Standardized Residuals")
qqline(sr, col = "red")
```



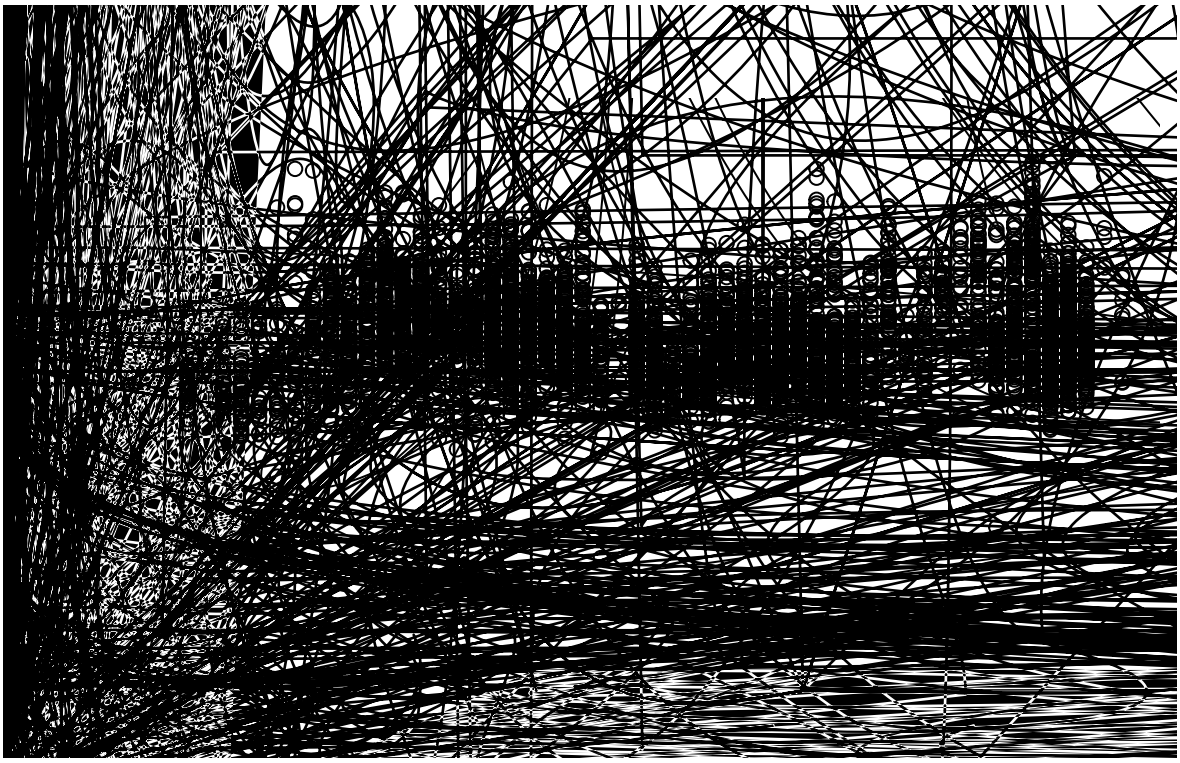

```r
# Residuals vs Fitted
plot(fitted_vals, sr,
     main = "Standardized Residuals vs Fitted Values",
     xlab = "Fitted Values",
     ylab = "Standardized Residuals")
abline(h = 0, col = "red")
```



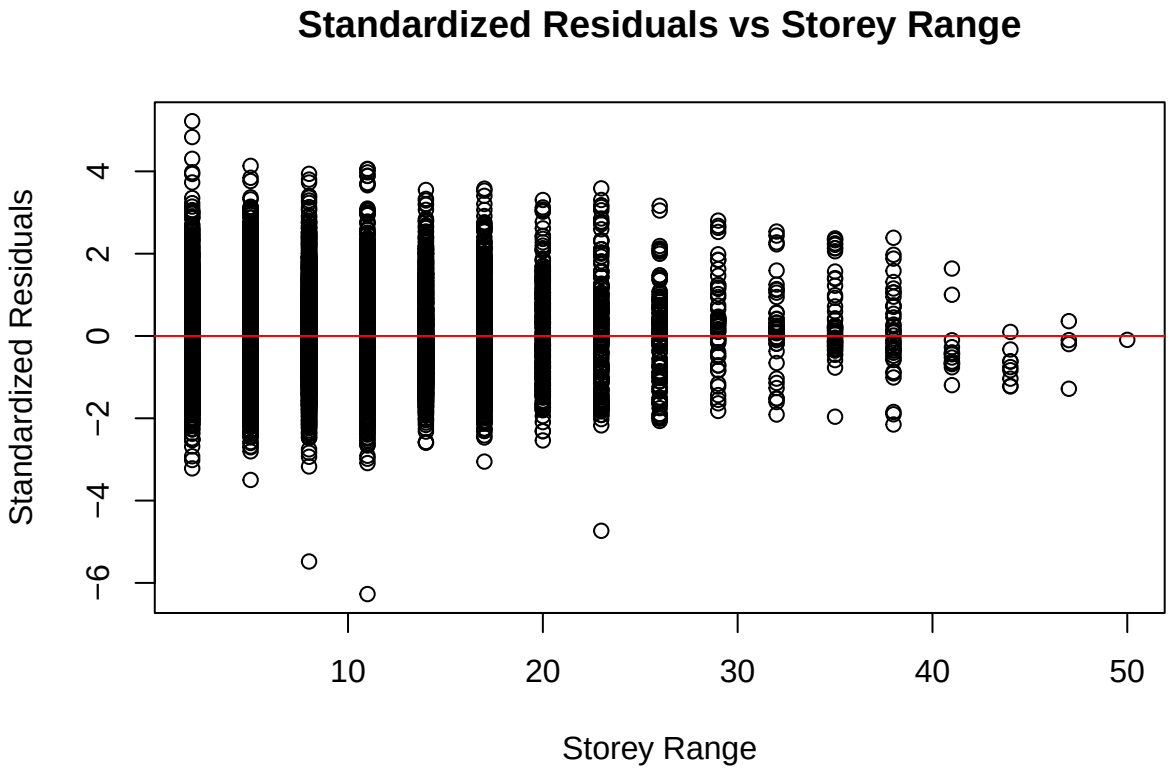

```r
# Residuals vs numeric predictors
plot(data$floor_area_sqm, sr,
     main = "Standardized Residuals vs Floor Area (sqm)",
     xlab = "Floor Area (sqm)",
     ylab = "Standardized Residuals")
abline(h = 0, col = "red")
```



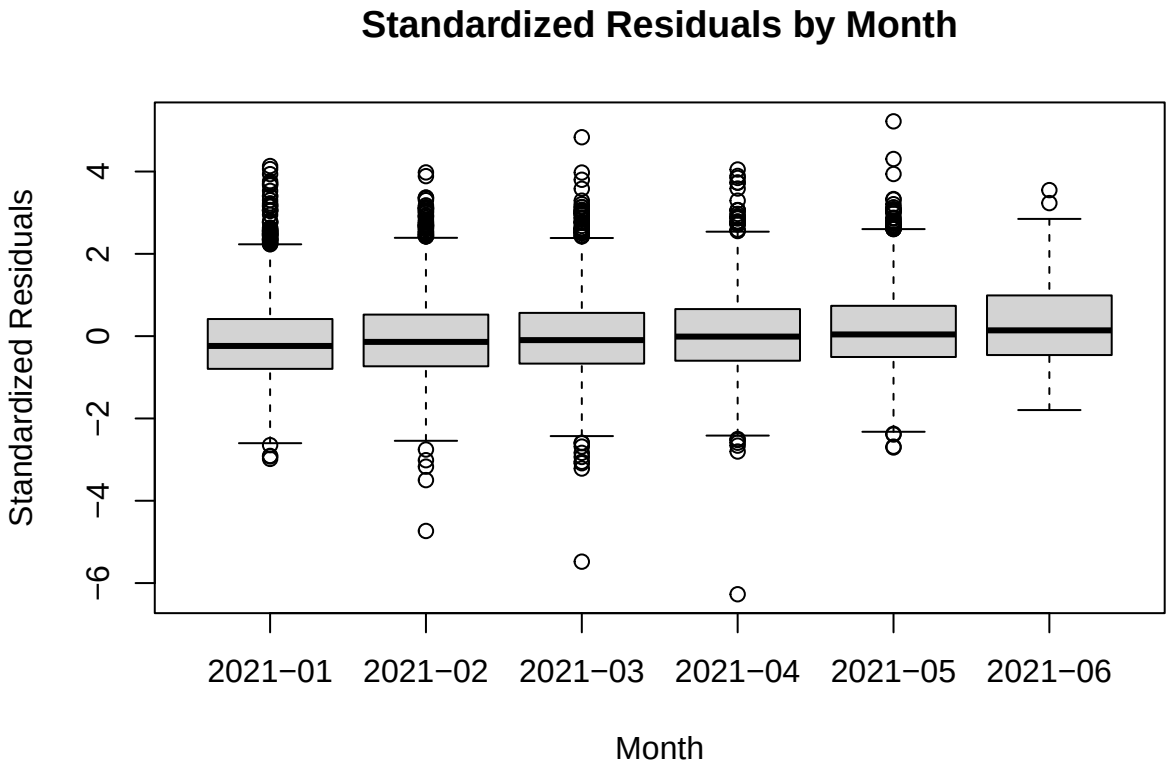

```
plot(data$lease_commence_date, sr,
     main = "Standardized Residuals vs Lease Commencement Date",
     xlab = "Lease Commencement Date",
     ylab = "Standardized Residuals")
abline(h = 0, col = "red")
```



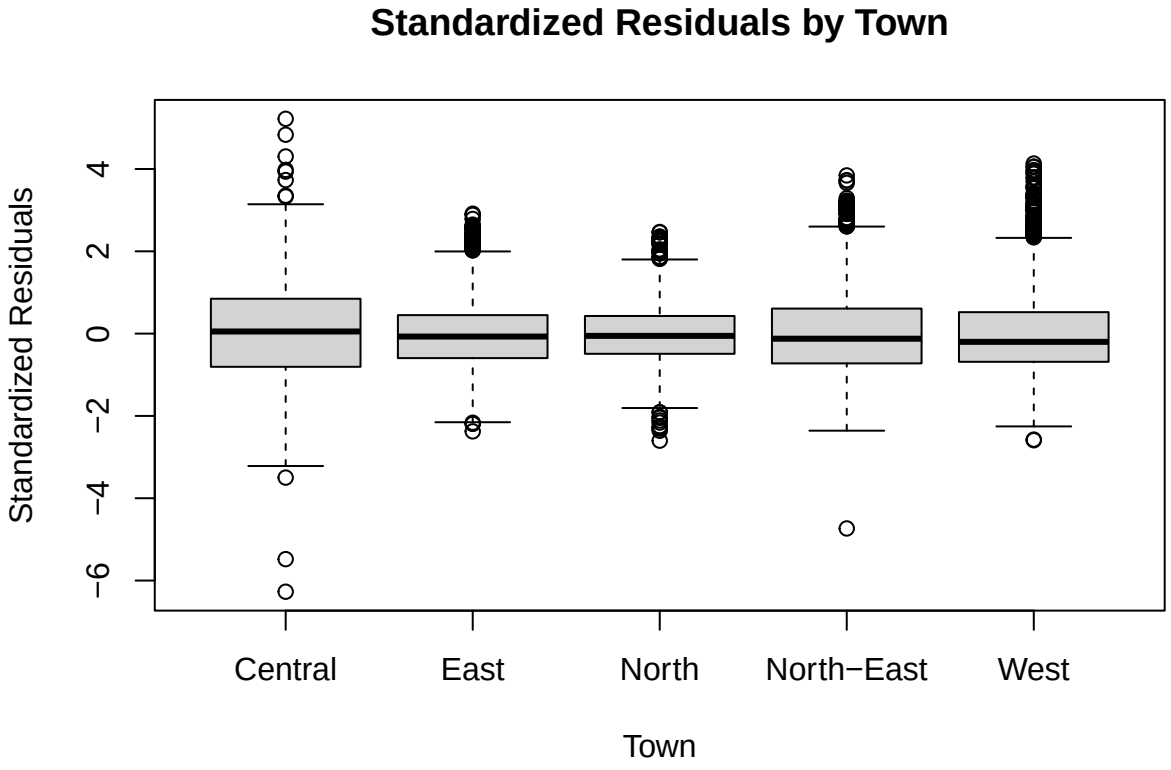
```r
plot(data$storey_range, sr,
     main = "Standardized Residuals vs Storey Range",
     xlab = "Storey Range",
     ylab = "Standardized Residuals")
abline(h = 0, col = "red")
```



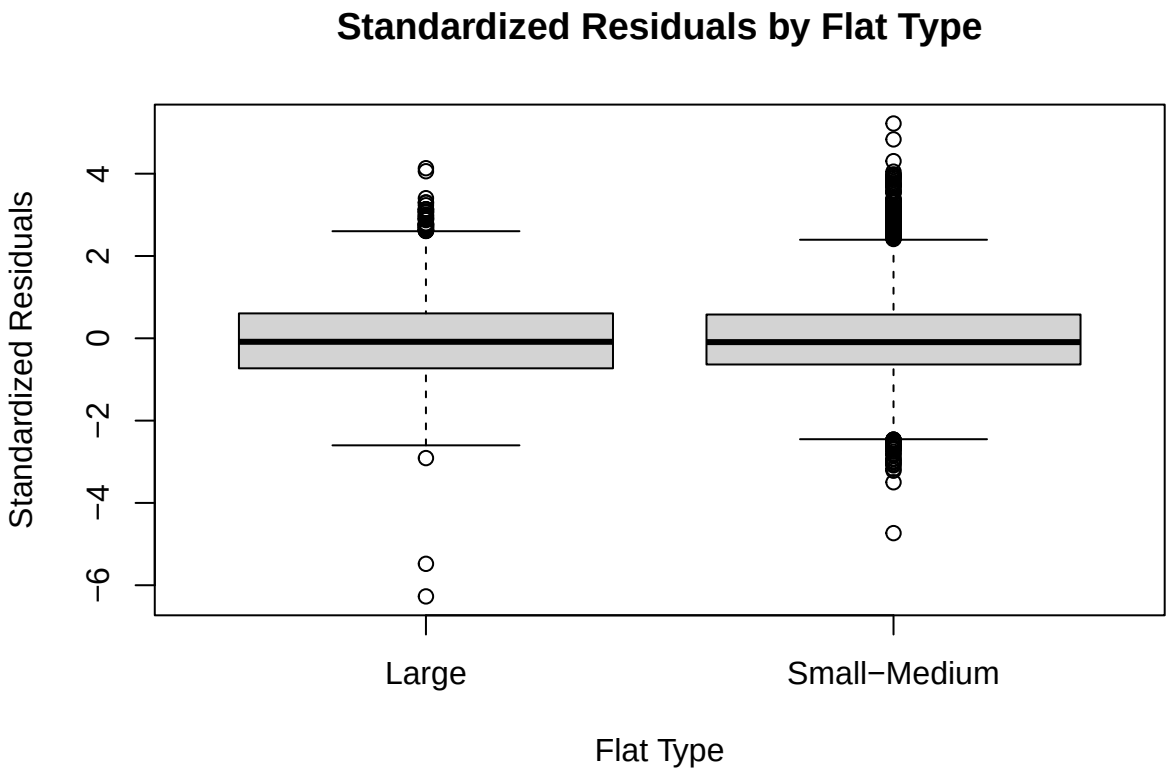

```r
# Residuals vs categorical variables using boxplots
boxplot(sr ~ data$month,
        main = "Standardized Residuals by Month",
        xlab = "Month",
        ylab = "Standardized Residuals")
```



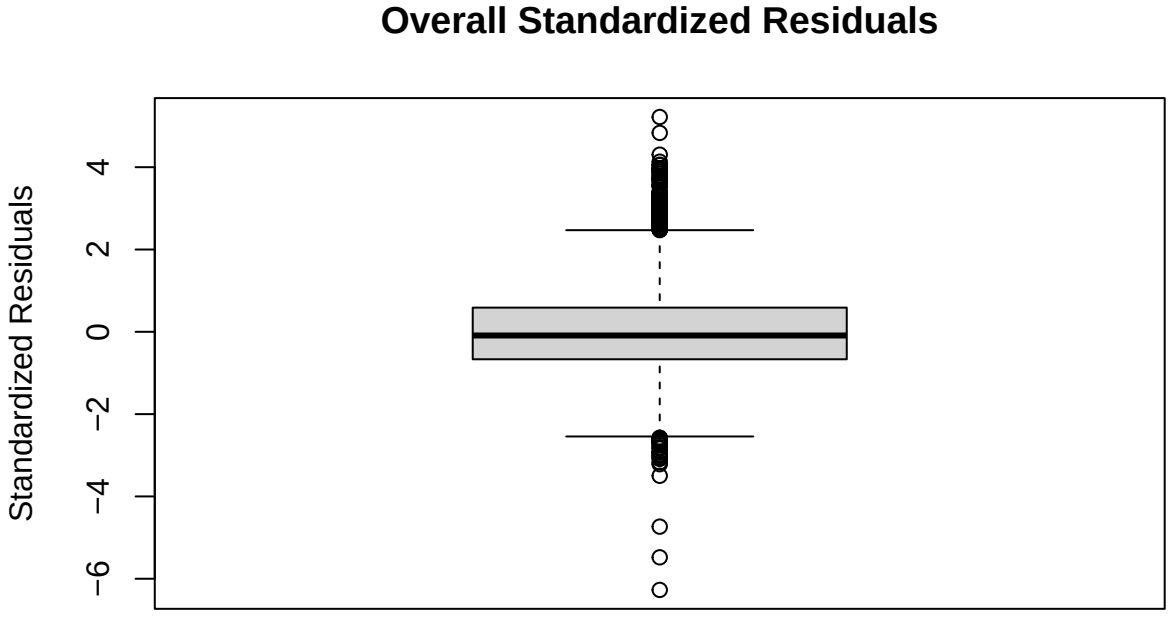

```
boxplot(sr ~ data$town,
        main = "Standardized Residuals by Town",
        xlab = "Town",
        ylab = "Standardized Residuals")
```



```r
boxplot(sr ~ data$flat_type_grouped,
        main = "Standardized Residuals by Flat Type",
        xlab = "Flat Type",
        ylab = "Standardized Residuals")
```




```r
# Overall boxplot to check outliers
boxplot(sr, main = "Overall Standardized Residuals", ylab = "Standardized Residuals")
```



```r
length(boxplot(sr, plot = FALSE)$out)
```

```
## [1] 209
```

```r
# Check for influential points
which(cooks.distance(M3) >= 1) # indices of influential points
```

```
## named integer(0)
```

```r
# Calculate VIF (Variance Inflation Factor) for each predictor
vif_values <- vif(M3)
print(vif_values)
```

```
##                           GVIF Df GVIF^(1/(2*Df))
## floor_area_sqm        2.600231  1        1.612523
## lease_commence_date   1.304689  1        1.142230
## town                  1.307856  4        1.034118
## I(storey_range^1.34) 1.198913  1        1.094949
## flat_type_grouped     2.527400  1        1.589780
```

```r
data <- data %>%
  mutate(storey_grouped = case_when(
    storey_range <= 5 ~ "Low",
    storey_range <= 11 ~ "Mid",
    storey_range <=20 ~ "High",
    storey_range >20 ~ "Very High",
  ))

data$storey_grouped <- factor(data$storey_grouped, levels = c("Low", "Mid", "High", "Very High"))

# Re-run regression
M4 <- lm(log(resale_price) ~ floor_area_sqm + lease_commence_date + town + flat_type_grouped + storey_g
summary(M4)
```

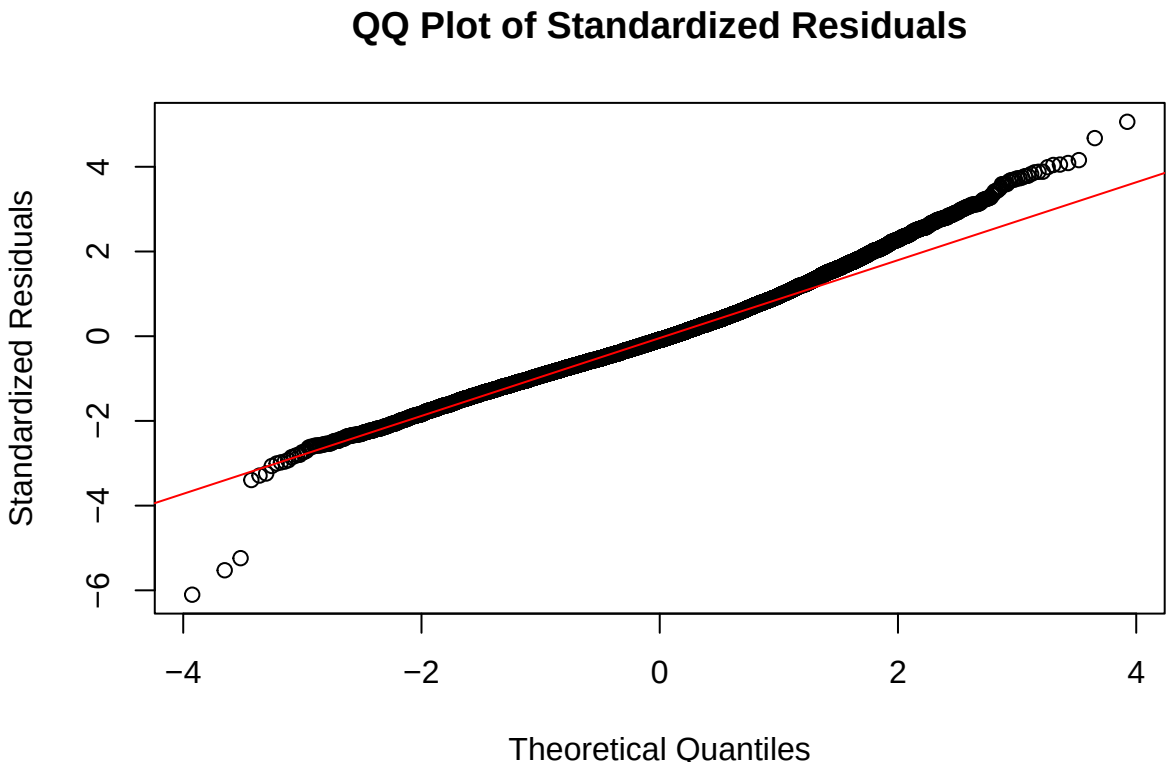
```
##
## Call:
## lm(formula = log(resale_price) ~ floor_area_sqm + lease_commence_date +
##     town + flat_type_grouped + storey_grouped, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.83282 -0.09085 -0.01157  0.07893  0.69299
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -5.951e+00  2.001e-01 -29.735  < 2e-16 ***
## floor_area_sqm                  9.899e-03  8.697e-05 113.830  < 2e-16 ***
## lease_commence_date             9.124e-03  1.009e-04  90.394  < 2e-16 ***
## townEast                       -2.048e-01  4.416e-03 -46.370  < 2e-16 ***
## townNorth                      -3.801e-01  4.817e-03 -78.919  < 2e-16 ***
## townNorth-East                 -2.940e-01  4.033e-03 -72.895  < 2e-16 ***
## townWest                       -3.252e-01  4.142e-03 -78.519  < 2e-16 ***
## flat_type_groupedSmall-Medium  1.655e-02  4.273e-03   3.872 0.000109 ***
## storey_groupedMid               4.686e-02  2.897e-03  16.173  < 2e-16 ***
## storey_groupedHigh              9.307e-02  3.832e-03  24.291  < 2e-16 ***
## storey_groupedVery High         2.684e-01  6.785e-03  39.565  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1369 on 11516 degrees of freedom
## Multiple R-squared:  0.8136, Adjusted R-squared:  0.8134
## F-statistic:  5026 on 10 and 11516 DF,  p-value: < 2.2e-16
```

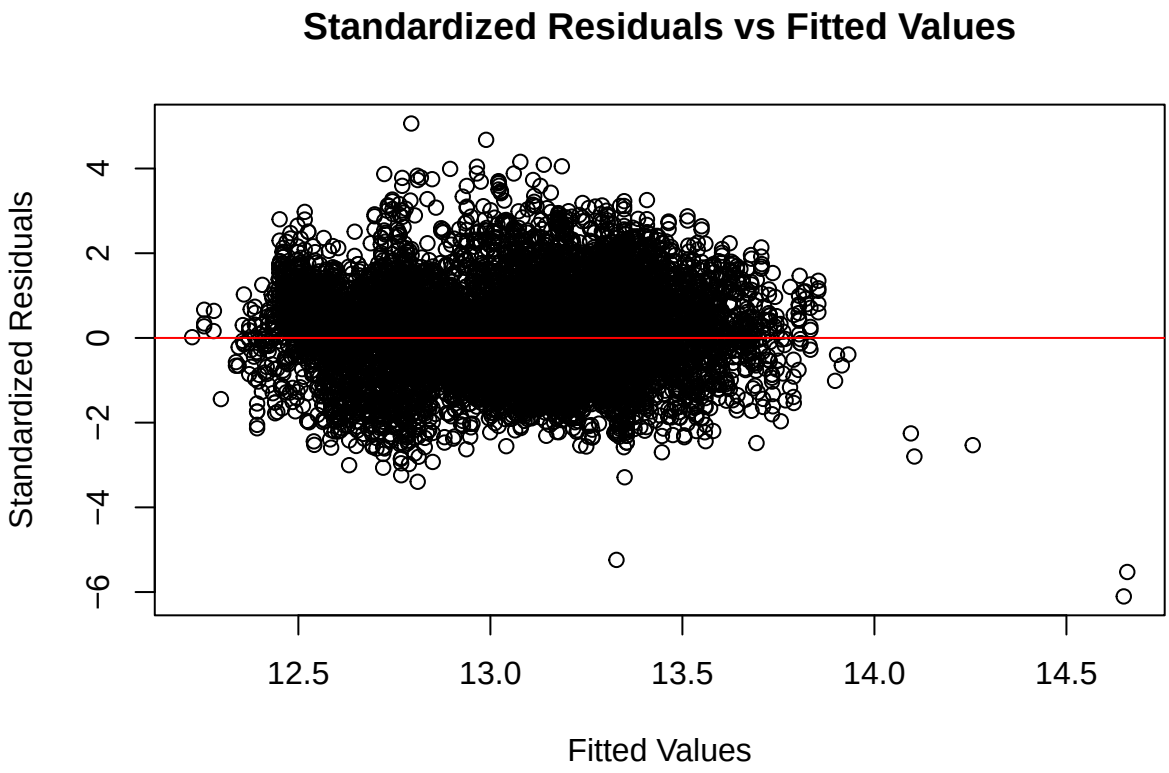
```
anova(M4)
```

```
## Analysis of Variance Table
##
## Response: log(resale_price)
##                         Df Sum Sq Mean Sq    F value  Pr(>F)
## floor_area_sqm           1 543.37  543.37 28972.3938 < 2e-16 ***
## lease_commence_date      1 147.38  147.38  7858.0083 < 2e-16 ***
## town                     4 217.98   54.49  2905.6255 < 2e-16 ***
## flat_type_grouped        1   0.10    0.10     5.2504 0.02196 *
## storey_grouped           3  33.86   11.29   601.7372 < 2e-16 ***
## Residuals            11516 215.98    0.02
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

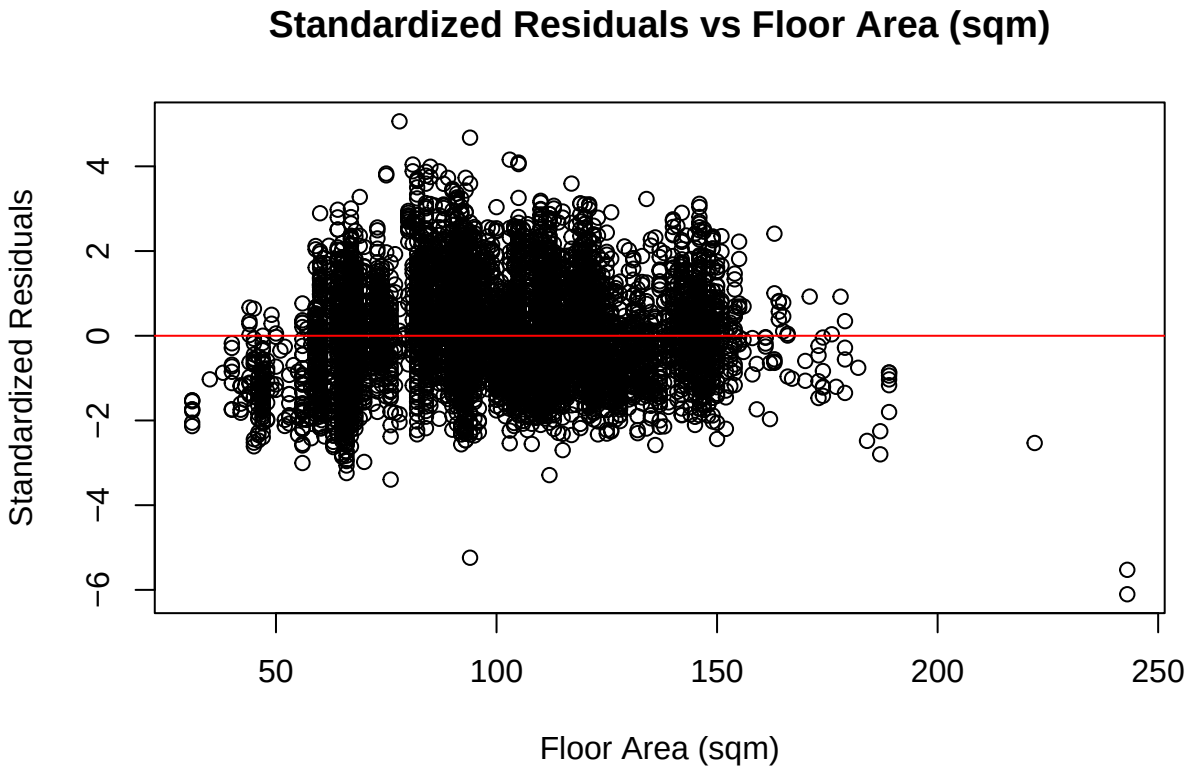
```
# Standardized residuals
sr <- rstandard(M4)
fitted_vals <- fitted(M4)

# QQ plot
qqnorm(sr, main = "QQ Plot of Standardized Residuals", ylab = "Standardized Residuals")
qqline(sr, col = "red")
```



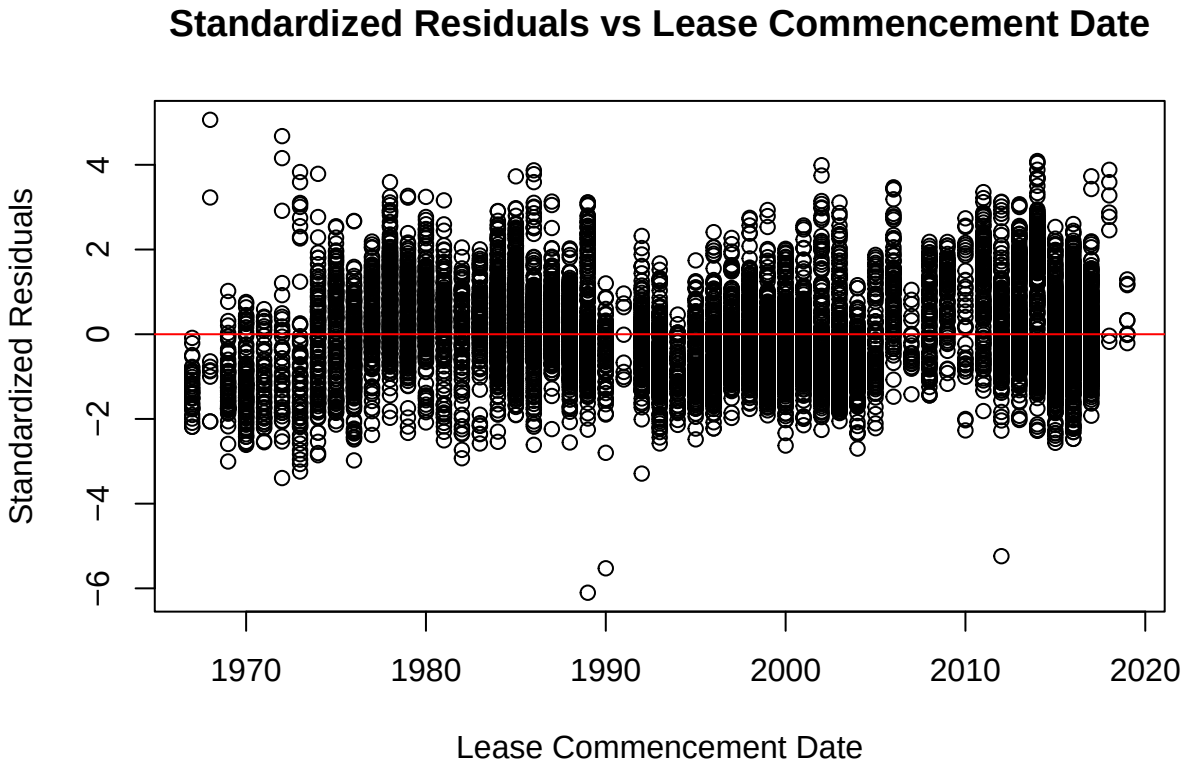
```r
# Residuals vs Fitted
plot(fitted_vals, sr,
     main = "Standardized Residuals vs Fitted Values",
     xlab = "Fitted Values",
     ylab = "Standardized Residuals")
abline(h = 0, col = "red")
```



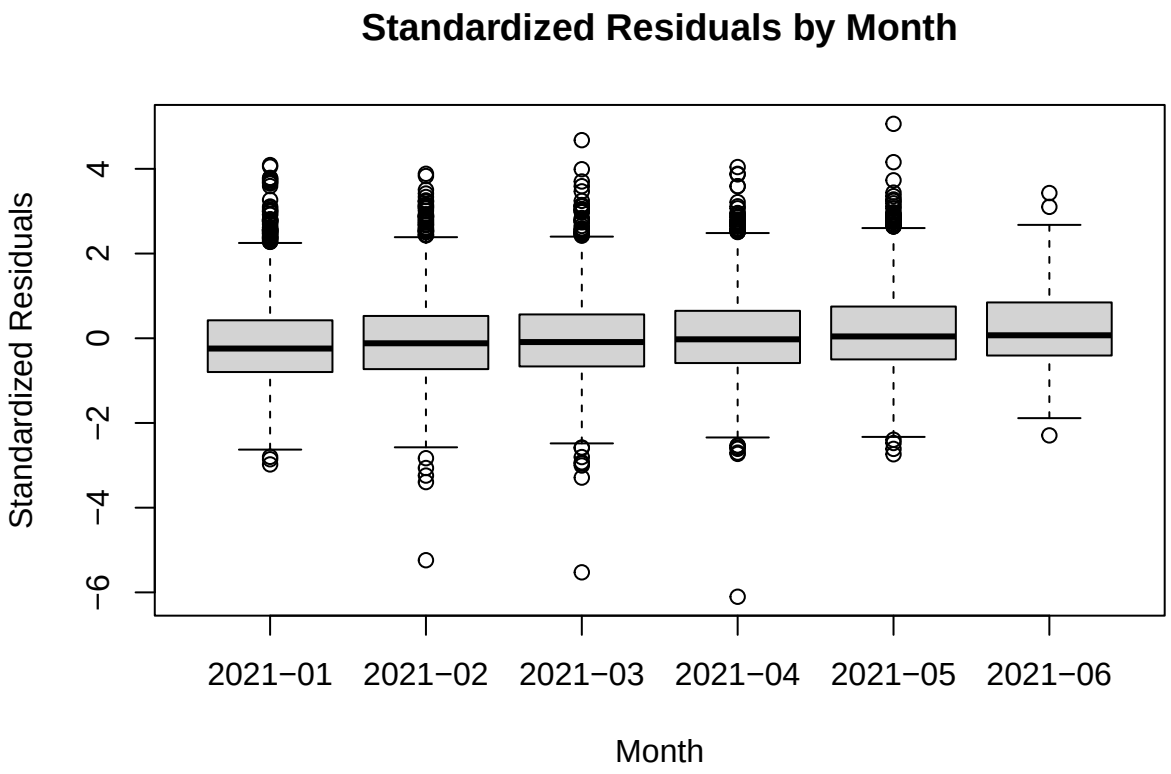

```r
# Residuals vs numeric predictors
plot(data$floor_area_sqm, sr,
     main = "Standardized Residuals vs Floor Area (sqm)",
     xlab = "Floor Area (sqm)",
     ylab = "Standardized Residuals")
abline(h = 0, col = "red")
```



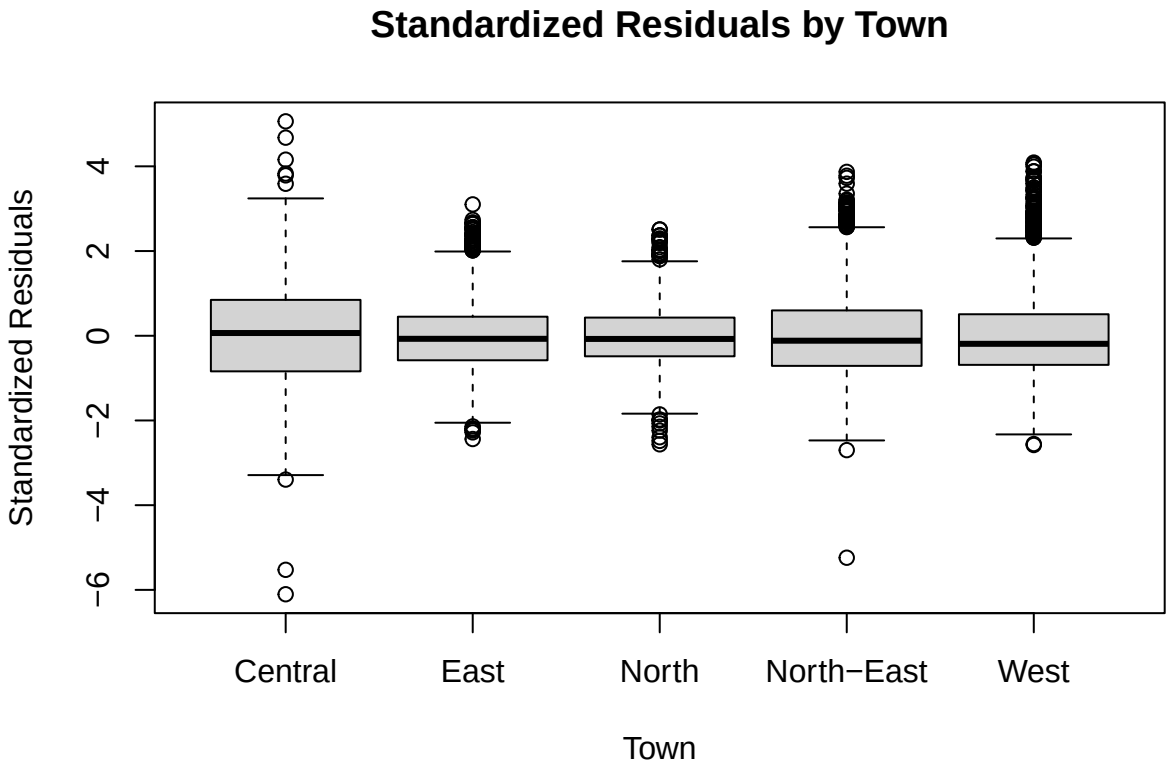

```
plot(data$lease_commence_date, sr,
     main = "Standardized Residuals vs Lease Commencement Date",
     xlab = "Lease Commencement Date",
     ylab = "Standardized Residuals")
abline(h = 0, col = "red")
```



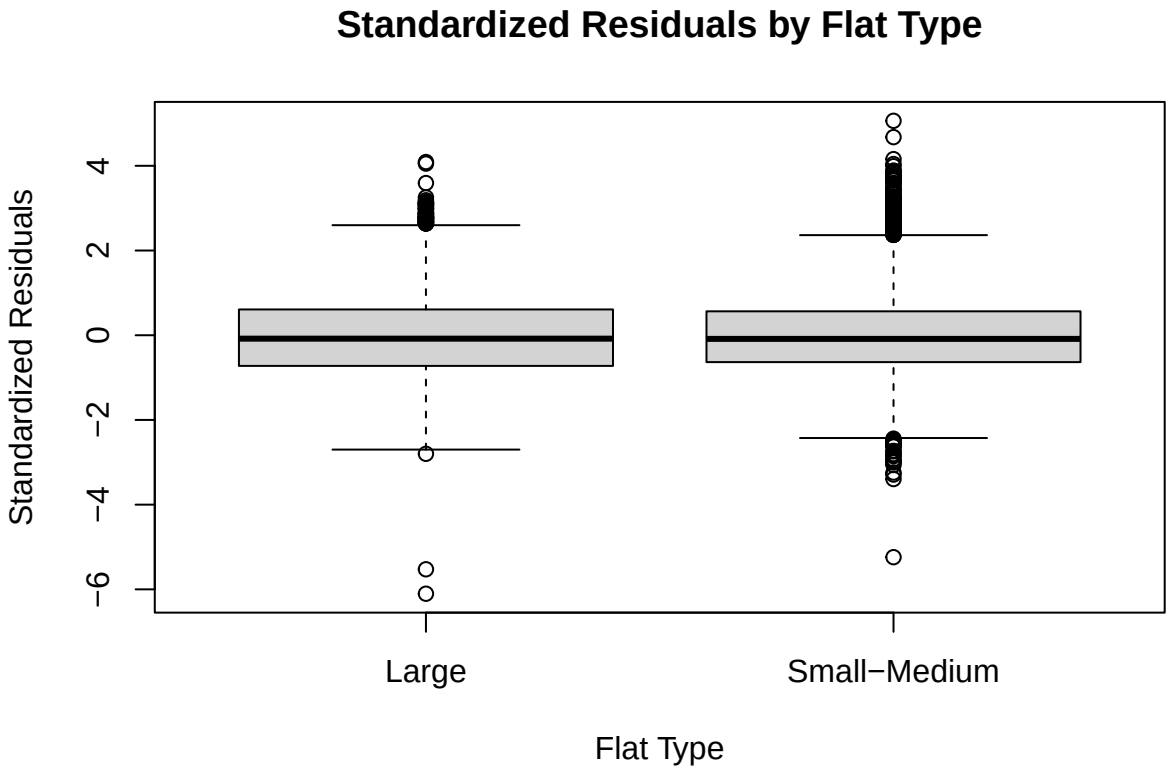
```r
# Residuals vs categorical variables using boxplots
boxplot(sr ~ data$month,
        main = "Standardized Residuals by Month",
        xlab = "Month",
        ylab = "Standardized Residuals")
```



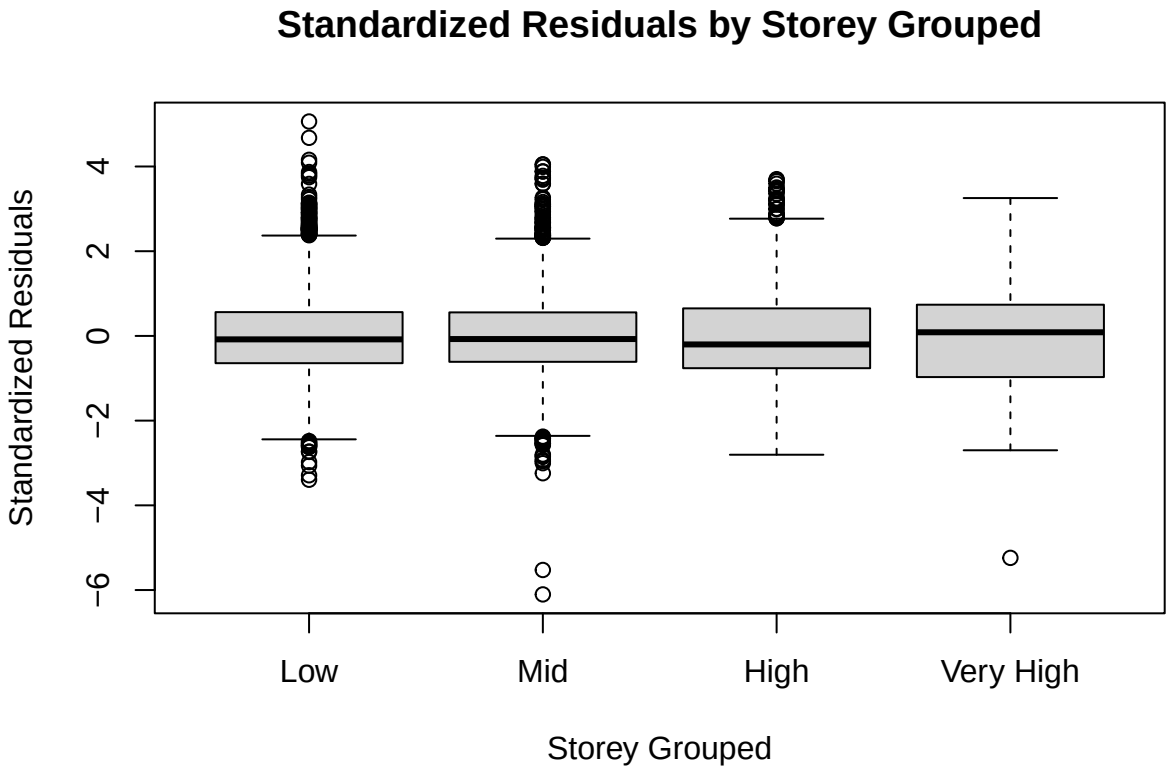

```r
boxplot(sr ~ data$town,
        main = "Standardized Residuals by Town",
        xlab = "Town",
        ylab = "Standardized Residuals")
```



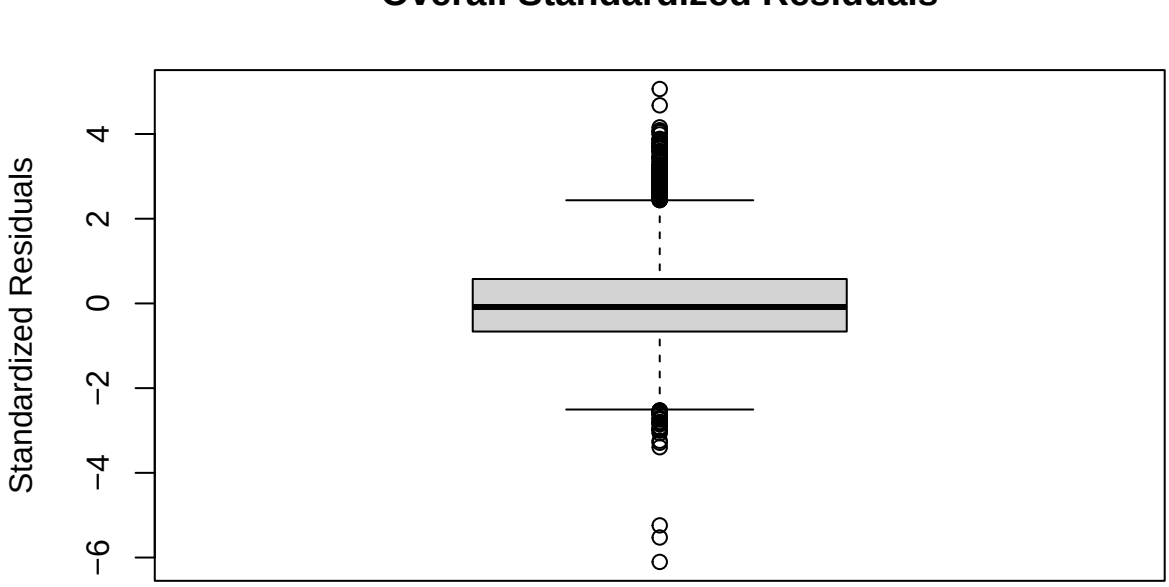

```
boxplot(sr ~ data$flat_type_grouped,
        main = "Standardized Residuals by Flat Type",
        xlab = "Flat Type",
        ylab = "Standardized Residuals")
```



```r
boxplot(sr ~ data$storey_grouped,
        main = "Standardized Residuals by Storey Grouped",
        xlab = "Storey Grouped",
        ylab = "Standardized Residuals")
```




```r
# Overall boxplot to check outliers
boxplot(sr, main = "Overall Standardized Residuals", ylab = "Standardized Residuals")
```



```r
length(boxplot(sr, plot = FALSE)$out)
```

```
## [1] 234
```

```r
# Check for influential points
which(cooks.distance(M4) >= 1) # indices of influential points
```

```
## named integer(0)
```

```r
# Calculate VIF (Variance Inflation Factor) for each predictor
vif_values <- vif(M4)
print(vif_values)
```

```
##                         GVIF Df GVIF^(1/(2*Df))
## floor_area_sqm      2.602745  1        1.613302
## lease_commence_date 1.296305  1        1.138554
## town                1.322763  4        1.035584
## flat_type_grouped   2.532907  1        1.591511
## storey_grouped      1.223789  3        1.034231
```

```r
M5 <- lm(log(resale_price) ~ log(floor_area_sqm) + lease_commence_date + town + flat_type_grouped + sto
summary(M5)
```

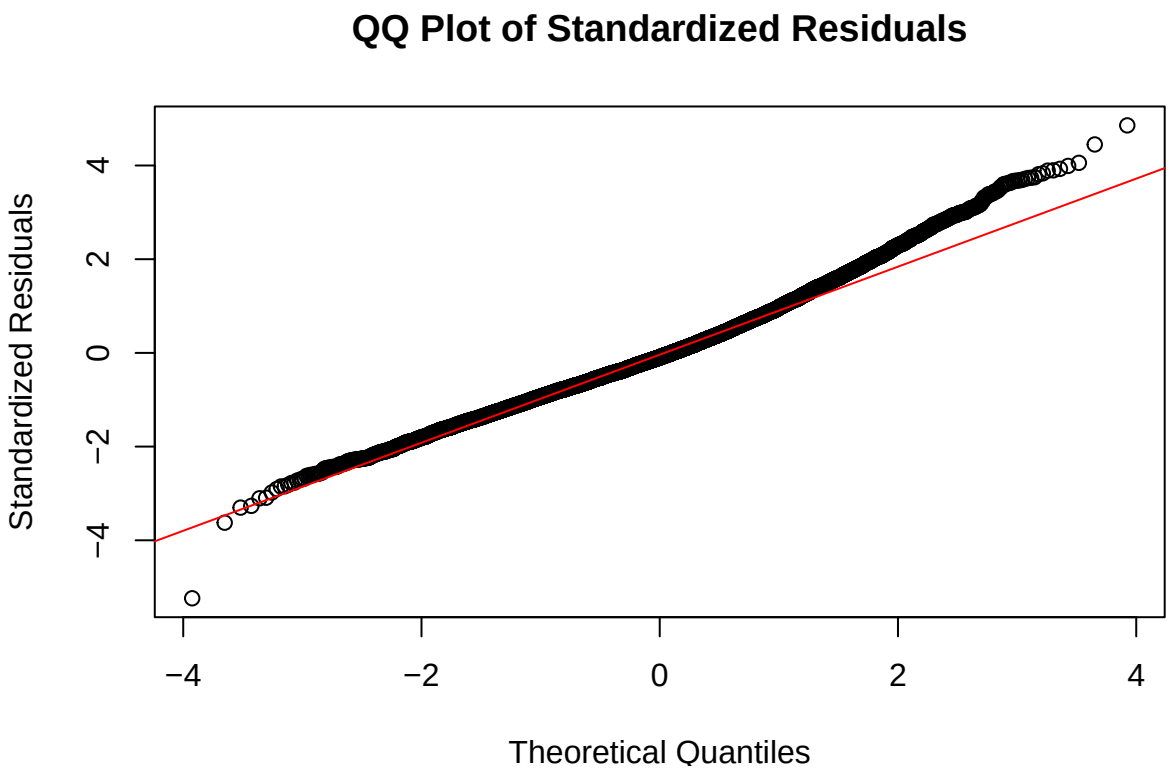
```
##
## Call:
## lm(formula = log(resale_price) ~ log(floor_area_sqm) + lease_commence_date +
##     town + flat_type_grouped + storey_grouped, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.71115 -0.09138 -0.01284  0.08084  0.65985
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -7.6604026  0.1990357  -38.49  < 2e-16 ***
## log(floor_area_sqm)             0.8766372  0.0075960  115.41  < 2e-16 ***
## lease_commence_date             0.0084839  0.0001007   84.25  < 2e-16 ***
## townEast                       -0.2049308  0.0043834  -46.75  < 2e-16 ***
## townNorth                      -0.3806238  0.0047814  -79.60  < 2e-16 ***
## townNorth-East                 -0.3009224  0.0040067  -75.10  < 2e-16 ***
## townWest                       -0.3276645  0.0041134  -79.66  < 2e-16 ***
## flat_type_groupedSmall-Medium -0.0279434  0.0039358   -7.10 1.32e-12 ***
## storey_groupedMid               0.0443847  0.0028757   15.44  < 2e-16 ***
## storey_groupedHigh              0.0870690  0.0038013   22.91  < 2e-16 ***
## storey_groupedVery High         0.2595887  0.0067333   38.55  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1359 on 11516 degrees of freedom
## Multiple R-squared:  0.8163, Adjusted R-squared:  0.8162
## F-statistic:  5118 on 10 and 11516 DF,  p-value: < 2.2e-16
```

```r
anova(M5)
```

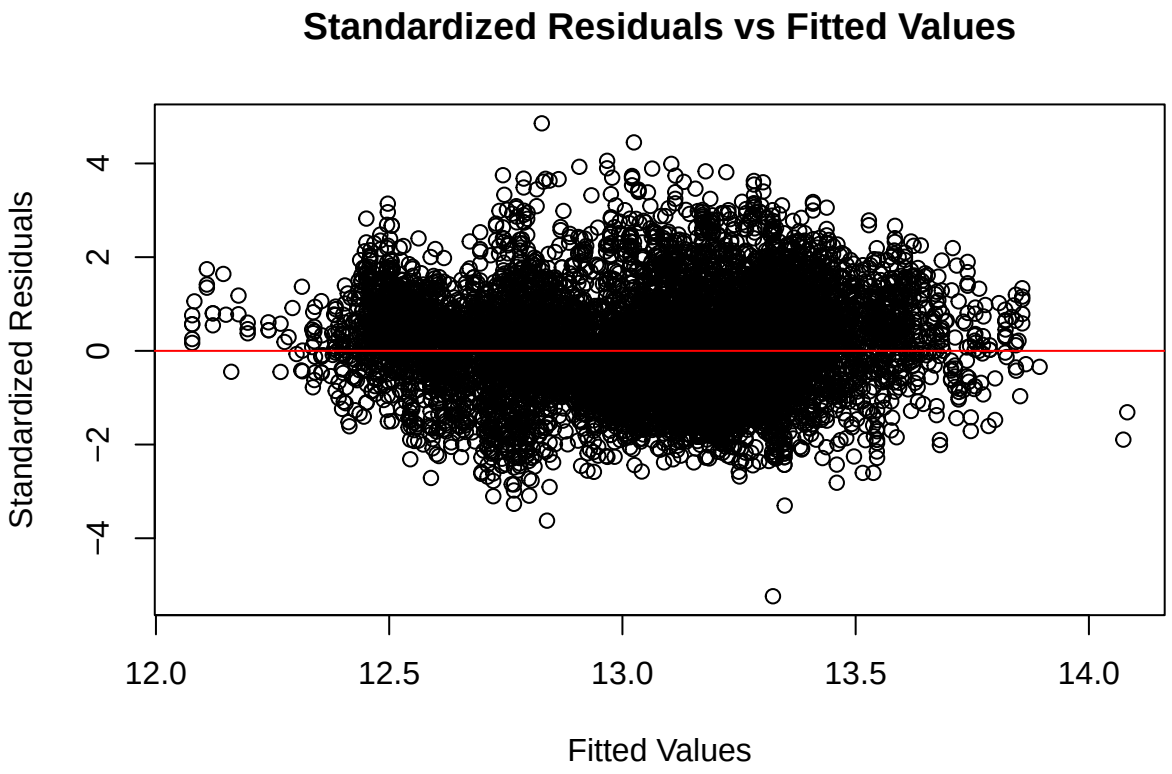
```
## Analysis of Variance Table
##
## Response: log(resale_price)
```

```
##                         Df Sum Sq Mean Sq   F value    Pr(>F)
## log(floor_area_sqm)      1 571.77  571.77 30937.409 < 2.2e-16 ***
## lease_commence_date      1 120.24  120.24  6505.963 < 2.2e-16 ***
## town                     4 221.34   55.33  2993.999 < 2.2e-16 ***
## flat_type_grouped        1   1.22    1.22    65.801 5.494e-16 ***
## storey_grouped           3  31.26   10.42   563.838 < 2.2e-16 ***
## Residuals            11516 212.83    0.02
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

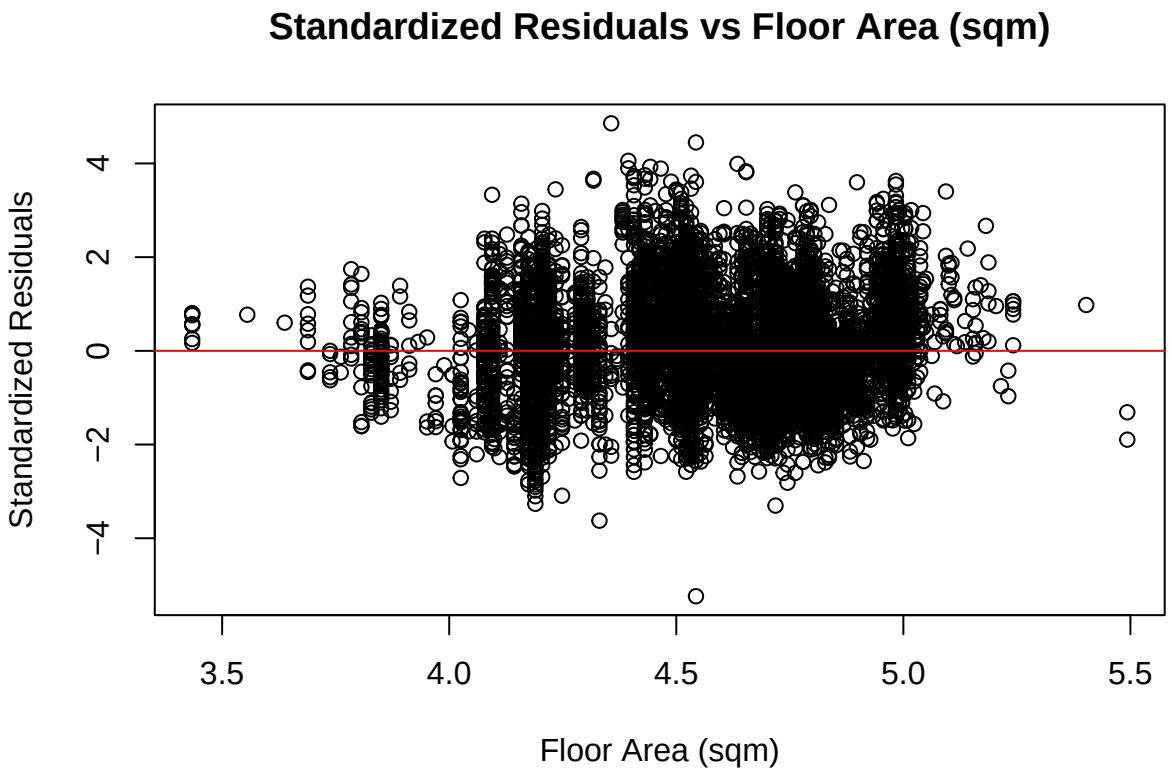
```r
# Standardized residuals
sr <- rstandard(M5)
fitted_vals <- fitted(M5)

# QQ plot
qqnorm(sr, main = "QQ Plot of Standardized Residuals", ylab = "Standardized Residuals")
qqline(sr, col = "red")
```



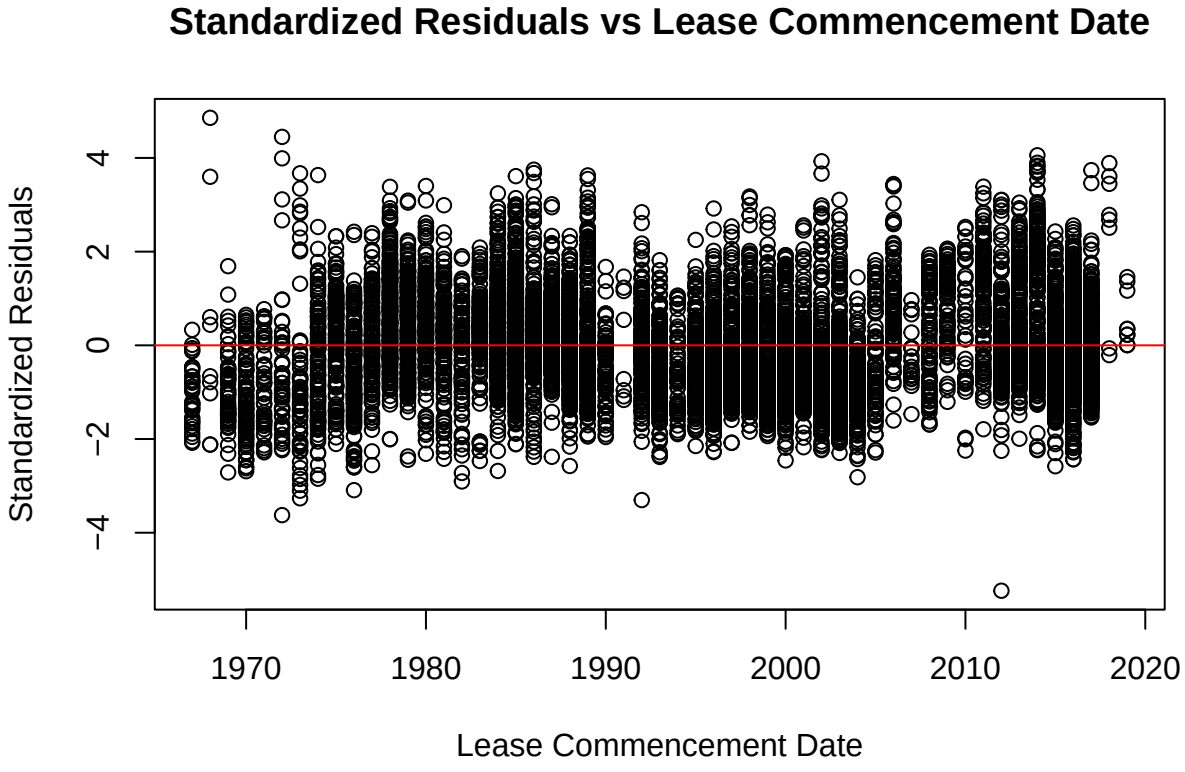

```r
# Residuals vs Fitted
plot(fitted_vals, sr,
     main = "Standardized Residuals vs Fitted Values",
     xlab = "Fitted Values",
     ylab = "Standardized Residuals")
abline(h = 0, col = "red")
```



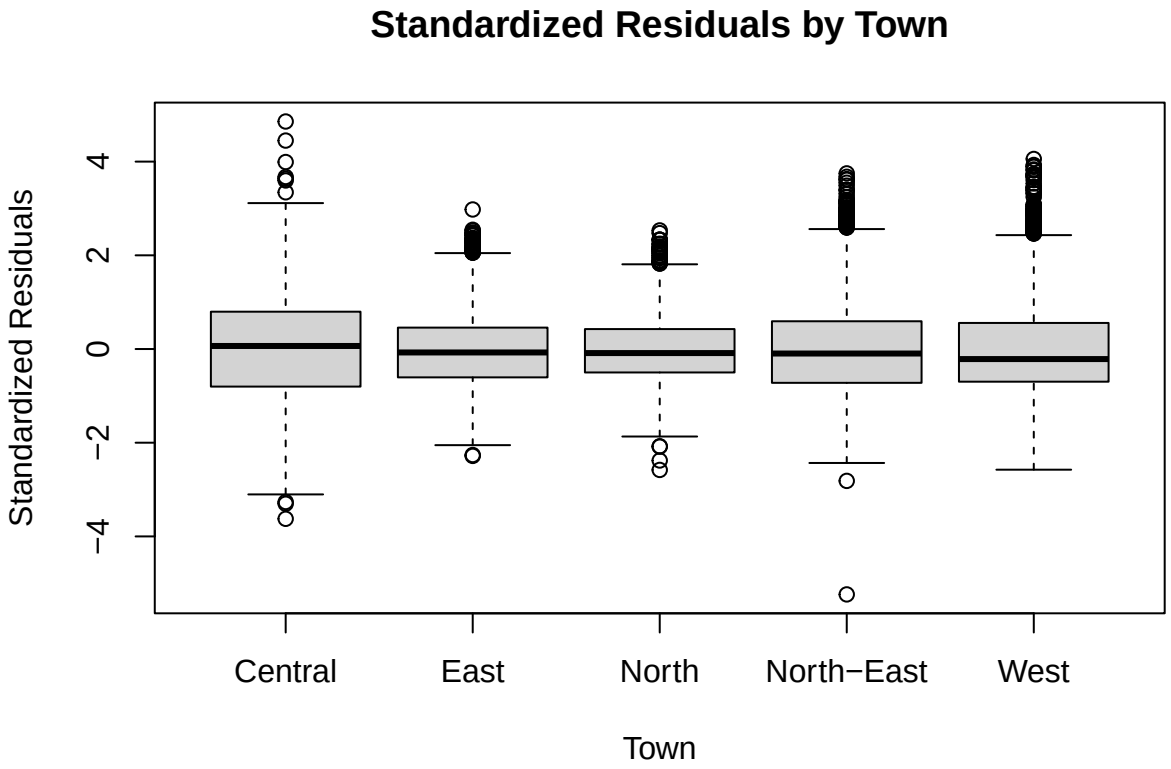

```
# Residuals vs numeric predictors
plot(log(data$floor_area_sqm), sr,
     main = "Standardized Residuals vs Floor Area (sqm)",
     xlab = "Floor Area (sqm)",
     ylab = "Standardized Residuals")
abline(h = 0, col = "red")
```



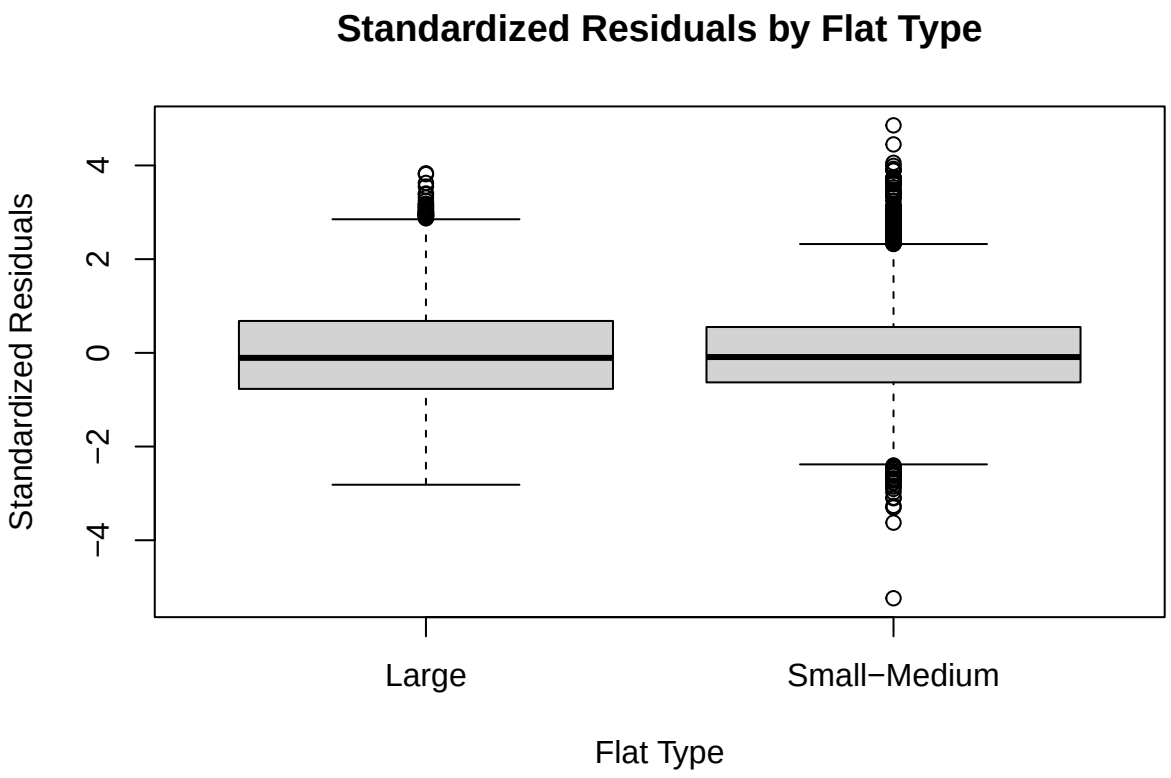

```
plot(data$lease_commence_date, sr,
     main = "Standardized Residuals vs Lease Commencement Date",
     xlab = "Lease Commencement Date",
     ylab = "Standardized Residuals")
abline(h = 0, col = "red")
```



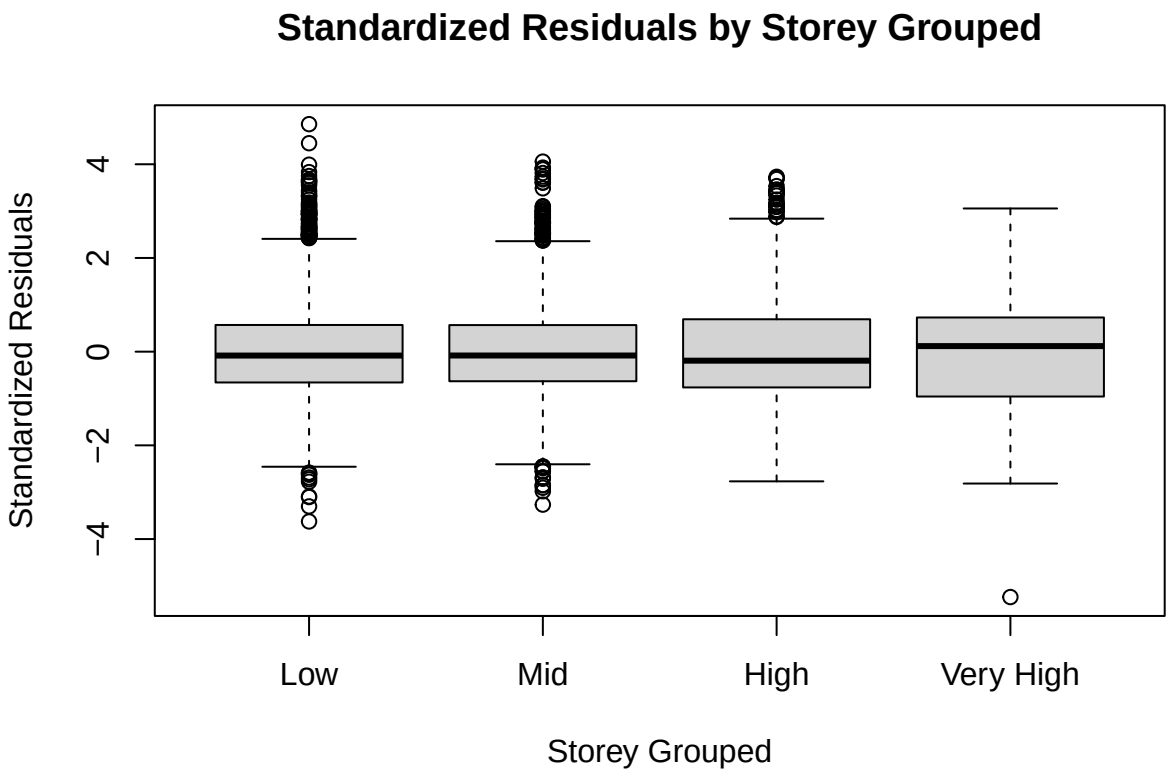
```r
boxplot(sr ~ data$town,
        main = "Standardized Residuals by Town",
        xlab = "Town",
        ylab = "Standardized Residuals")
```



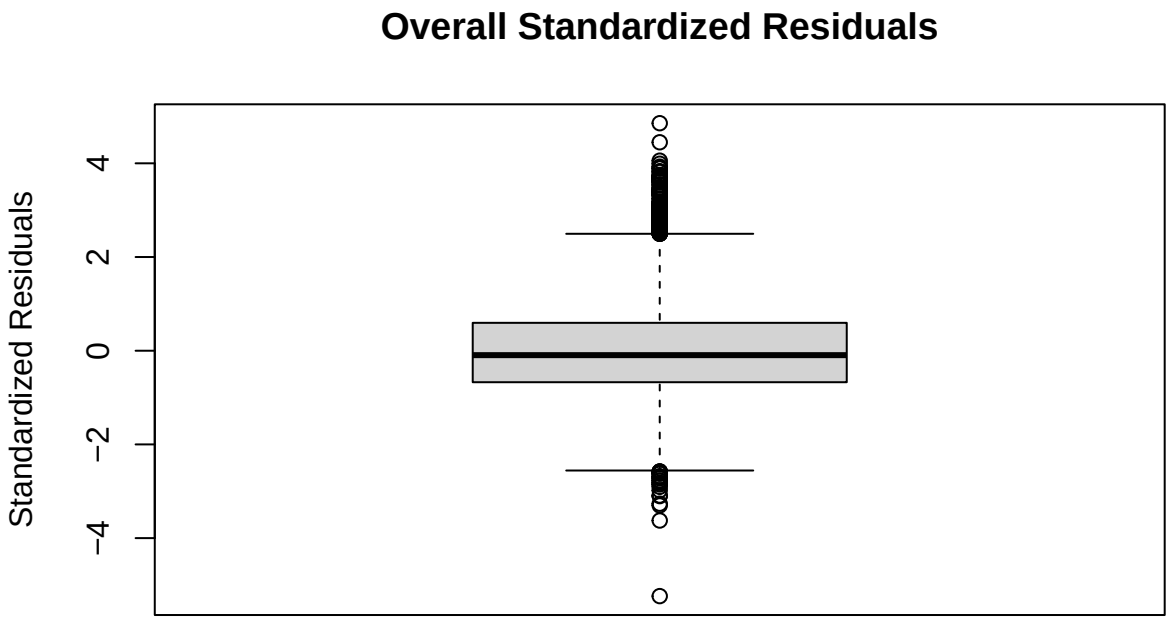

```r
boxplot(sr ~ data$flat_type_grouped,
        main = "Standardized Residuals by Flat Type",
        xlab = "Flat Type",
        ylab = "Standardized Residuals")
```




```
boxplot(sr ~ data$storey_grouped,
        main = "Standardized Residuals by Storey Grouped",
        xlab = "Storey Grouped",
        ylab = "Standardized Residuals")
```



```r
# Overall boxplot to check outliers
boxplot(sr, main = "Overall Standardized Residuals", ylab = "Standardized Residuals")
```




```r
length(boxplot(sr, plot = FALSE)$out)
```

```
## [1] 209
```

```r
# Check for influential points
which(cooks.distance(M5) >= 1) # indices of influential points
```

```
## named integer(0)
```

```r
# Calculate VIF (Variance Inflation Factor) for each predictor
vif_values <- vif(M5)
print(vif_values)
```

```
##                         GVIF Df GVIF^(1/(2*Df))
## log(floor_area_sqm) 2.263650  1        1.504543
## lease_commence_date 1.309283  1        1.144239
## town                1.320606  4        1.035373
## flat_type_grouped   2.180453  1        1.476636
## storey_grouped      1.222067  3        1.033989
```