Quiz7-2

```
if (!require("ggplot2")) install.packages("ggplot2")
```

```
Loading required package: ggplot2
```

```
if (!require("dplyr")) install.packages("dplyr")
```

```
Loading required package: 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
```

```
if (!require("car")) install.packages("car")
```

```
Loading required package: car
```

```
Loading required package: carData
```

```

Attaching package: 'car'
```

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

    recode
```

```r
library(ggplot2)
library(dplyr)
library(car)
```

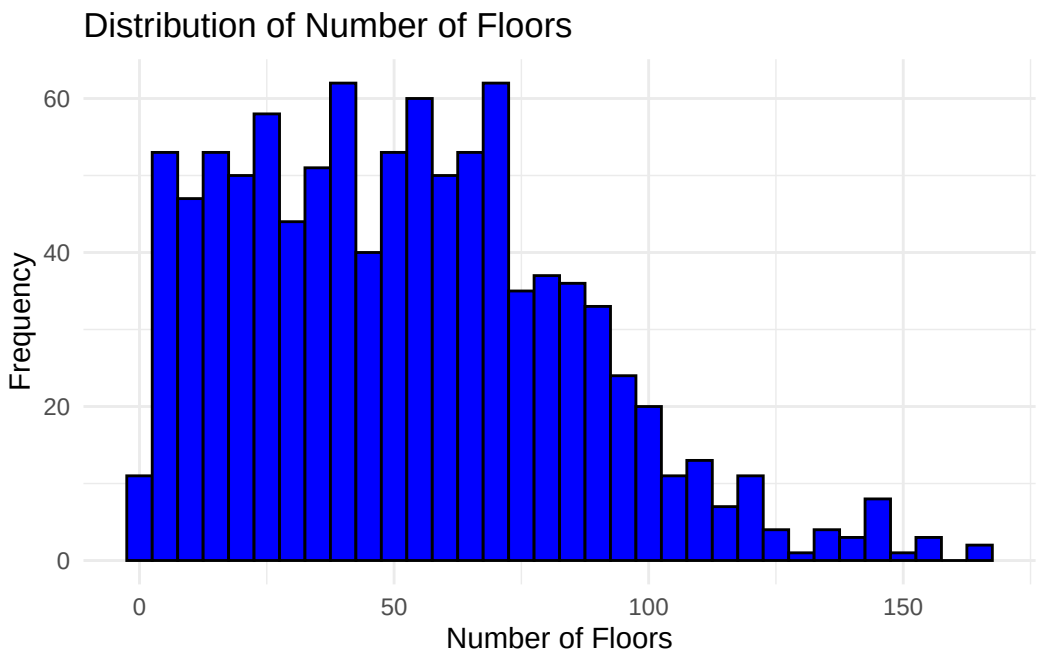
```r
set.seed(1)

n <- 1000

year_of_construction <- sample(1700:2020, n, replace = TRUE)
location_zone <- sample(c("Central", "Suburban", "Outskirts"), n, replace = TRUE, prob = c(0
building_type <- sample(c("Residential", "Commercial", "Mixed-use"), n, replace = TRUE)


number_of_floors <- round(runif(n, min = 1, max = 100) *
                            (year_of_construction / 2020)^2 *
                            (ifelse(location_zone == "Central", 1.5, 1)) *
                            (ifelse(building_type == "Commercial", 1.2, 1))
                          )


buildings_df <- data.frame(year_of_construction, location_zone, building_type, number_of_floo
```

```r
# 1. Summary Statistics of Number of Floors
summary(buildings_df$number_of_floors)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00   25.00   50.00   52.47   73.00  167.00
```

```r
# 2. Distribution of Number of Floors
ggplot(buildings_df, aes(x = number_of_floors)) +
  geom_histogram(binwidth = 5, fill = "blue", color = "black") +
```

```
  theme_minimal() +
  labs(title = "Distribution of Number of Floors", x = "Number of Floors", y = "Frequency")
```

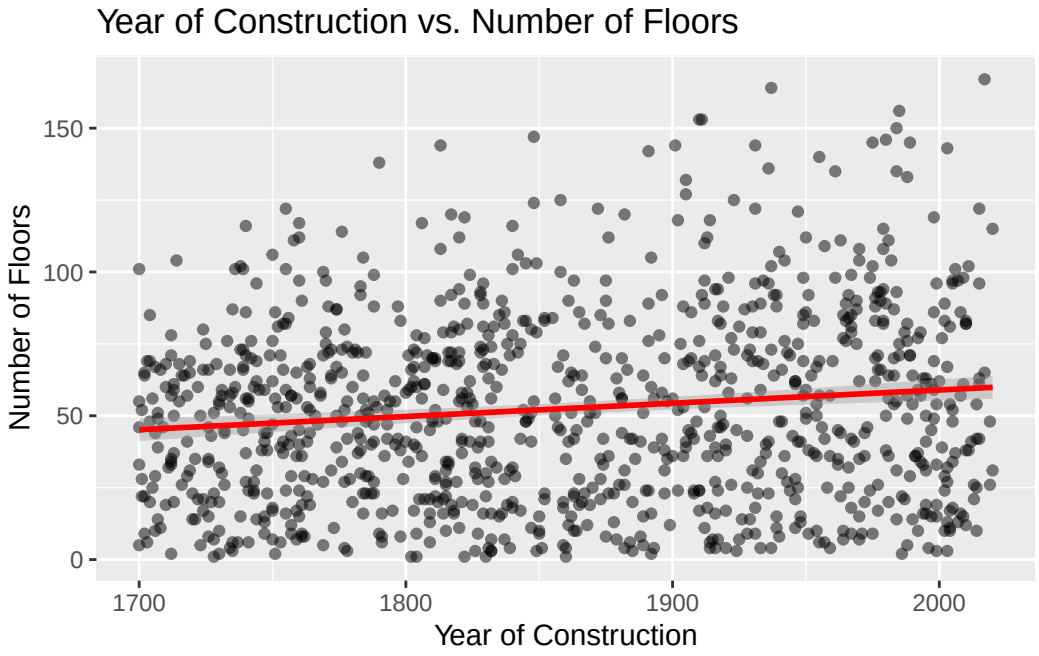



```
# 3. Year of Construction vs. Number of Floors
cor(buildings_df$year_of_construction, buildings_df$number_of_floors, use = "complete.obs")
```

```
[1] 0.1314269
```

```
ggplot(buildings_df, aes(x = year_of_construction, y = number_of_floors)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", color = "red") +
  labs(title = "Year of Construction vs. Number of Floors", x = "Year of Construction", y = "
```

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




```r
# 4. Effect of Location Zone on Number of Floors
anova_zone <- aov(number_of_floors ~ location_zone, data = buildings_df)
summary(anova_zone)
```

```
               Df Sum Sq Mean Sq F value Pr(>F)
location_zone   2  89642   44821   45.58 <2e-16 ***
Residuals     997 980438     983
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# 5. Effect of Building Type on Number of Floors
anova_type <- aov(number_of_floors ~ building_type, data = buildings_df)
summary(anova_type)
```

```
               Df  Sum Sq Mean Sq F value   Pr(>F)
building_type   2   26574   13287    12.7 3.59e-06 ***
Residuals     997 1043505    1047
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# 6. Linear Regression Analysis
lmModel <- lm(number_of_floors ~ year_of_construction + location_zone + building_type, data =
summary(lmModel)
```

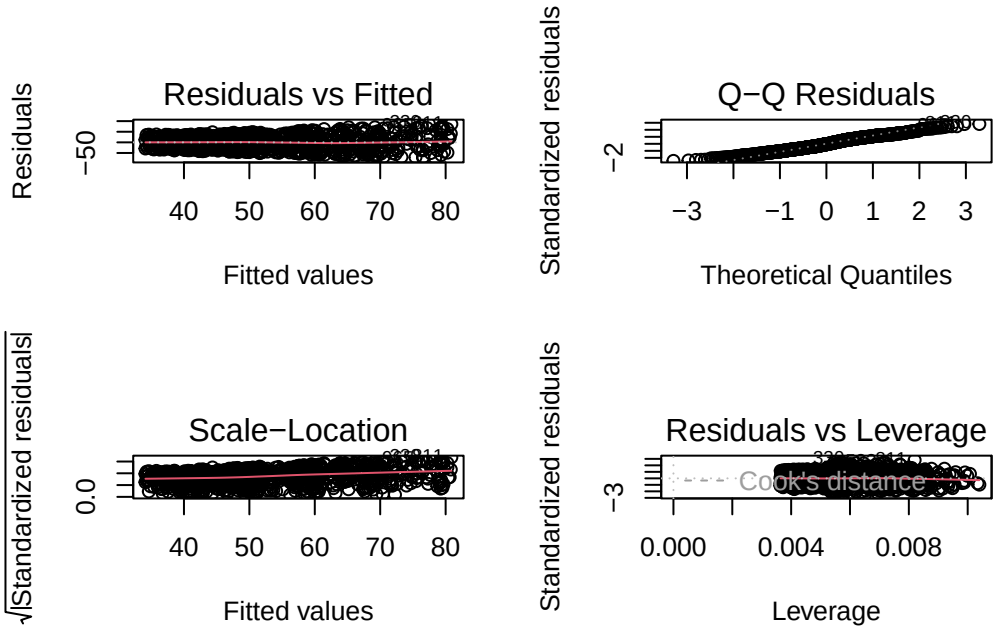
```
Call:
lm(formula = number_of_floors ~ year_of_construction + location_zone +
    building_type, data = buildings_df)

Residuals:
    Min      1Q  Median      3Q     Max
-74.392 -23.444  -1.622  24.611  86.551

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               -5.60399   19.55794  -0.287 0.774530
year_of_construction       0.04288    0.01043   4.110 4.28e-05 ***
location_zoneOutskirts   -19.62396    2.82507  -6.946 6.77e-12 ***
location_zoneSuburban    -20.95551    2.27095  -9.228  < 2e-16 ***
building_typeMixed-use   -12.21867    2.44835  -4.991 7.10e-07 ***
building_typeResidential  -9.02922    2.32131  -3.890 0.000107 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 30.73 on 994 degrees of freedom
Multiple R-squared:  0.123, Adjusted R-squared:  0.1186
F-statistic: 27.89 on 5 and 994 DF,  p-value: < 2.2e-16
```

```r
# 7. Interaction Effects
lmModel_interactions <- lm(number_of_floors ~ year_of_construction * location_zone + buildin
summary(lmModel_interactions)
```

```
Call:
lm(formula = number_of_floors ~ year_of_construction * location_zone +
    building_type, data = buildings_df)

Residuals:
    Min      1Q  Median      3Q     Max
-75.545 -23.380  -1.438  24.114  85.663
```

```
Coefficients:
                                            Estimate Std. Error t value
(Intercept)                                -28.02384   36.65975  -0.764
year_of_construction                         0.05491    0.01963   2.797
location_zoneOutskirts                      40.86959   57.10213   0.716
location_zoneSuburban                       -0.77282   45.56089  -0.017
building_typeMixed-use                     -12.29562    2.45086  -5.017
building_typeResidential                    -8.96191    2.32623  -3.853
year_of_construction:location_zoneOutskirts -0.03244    0.03059  -1.061
year_of_construction:location_zoneSuburban  -0.01083    0.02447  -0.443
                                           Pr(>|t|)
(Intercept)                                0.444792
year_of_construction                       0.005257 **
location_zoneOutskirts                     0.474328
location_zoneSuburban                      0.986470
building_typeMixed-use                     6.22e-07 ***
building_typeResidential                   0.000124 ***
year_of_construction:location_zoneOutskirts 0.289117
year_of_construction:location_zoneSuburban  0.658203
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 30.74 on 992 degrees of freedom
Multiple R-squared:  0.124, Adjusted R-squared:  0.1178
F-statistic: 20.07 on 7 and 992 DF,  p-value: < 2.2e-16
```

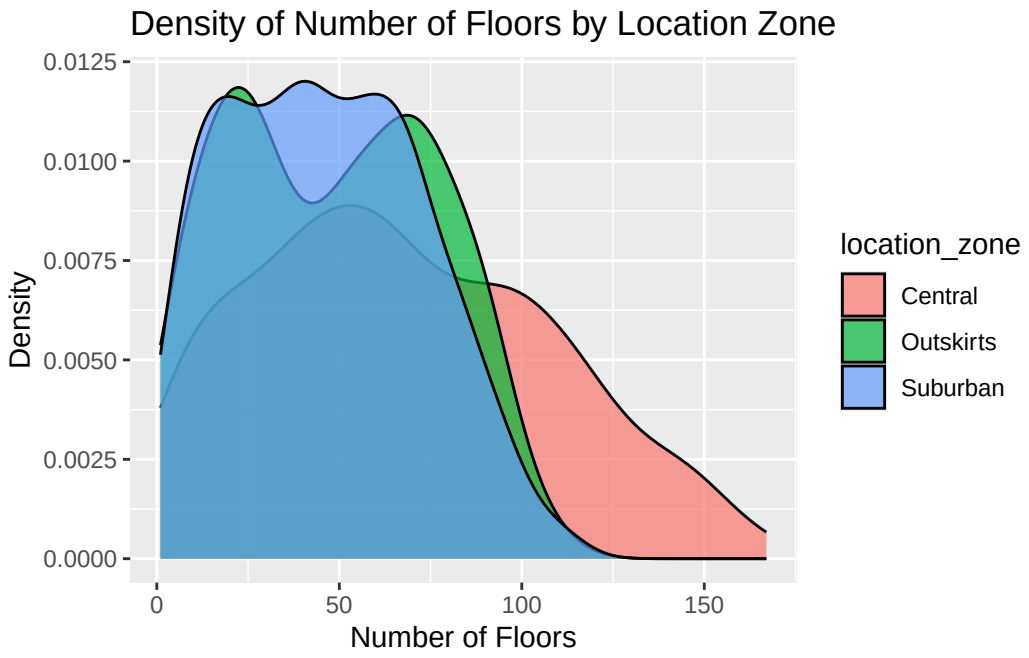
```r
# 8. Model Diagnostics
par(mfrow = c(2, 2))
plot(lmModel)
```




```
# 9. Predictive Accuracy
# Splitting the dataset into training (80%) and testing (20%) sets
set.seed(123)
training_indices <- sample(1:nrow(buildings_df), 0.8 * nrow(buildings_df))
training_data <- buildings_df[training_indices, ]
testing_data <- buildings_df[-training_indices, ]

lmModel_train <- lm(number_of_floors ~ year_of_construction + location_zone + building_type,
predicted_floors <- predict(lmModel_train, newdata = testing_data)

# Calculate Mean Squared Error (MSE)
mse <- mean((testing_data$number_of_floors - predicted_floors)^2)
print(paste("Mean Squared Error:", mse))
```

```
[1] "Mean Squared Error: 988.901876479925"
```

```
# 10. Density Plots by Category (Location Zone)
ggplot(buildings_df, aes(x = number_of_floors, fill = location_zone)) +
  geom_density(alpha = 0.7) +
  labs(title = "Density of Number of Floors by Location Zone", x = "Number of Floors", y = "I
```


Density of Number of Floors by Location Zone
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
Density
0
50
100
150
Number of Floors
location_zone
Central
Outskirts
Suburban