Car Value Analytics

Younseo Kim

Introduction

In our study, we are testing the hypothesis that certain factors - specifically, the age, mileage, model, and zip code of a used car - significantly influence its price. We hypothesize that older cars and those with higher mileage will generally be priced lower, while certain models and specific zip codes might have distinct impacts on the car's price. Our analysis aims to statistically validate whether these factors indeed play a significant role in determining the market value of used cars, using a data set of used car sales and employing linear regression with the natural logarithm of price as the dependent variable.

Load Data

```
data_ford16802 <- read.csv("C:/Users/leo/Desktop/ANOVAproject/data/ford16802.csv")
data_ford07640 <- read.csv("C:/Users/leo/Desktop/ANOVAproject/data/ford07640.csv")
data_honda16802 <- read.csv("C:/Users/leo/Desktop/ANOVAproject/data/honda16802.csv")
data_honda07640 <- read.csv("C:/Users/leo/Desktop/ANOVAproject/data/honda07640.csv")
data_toyota16802 <- read.csv("C:/Users/leo/Desktop/ANOVAproject/data/toyota16802.csv")
data_toyota07640 <- read.csv("C:/Users/leo/Desktop/ANOVAproject/data/toyota07640.csv")
```

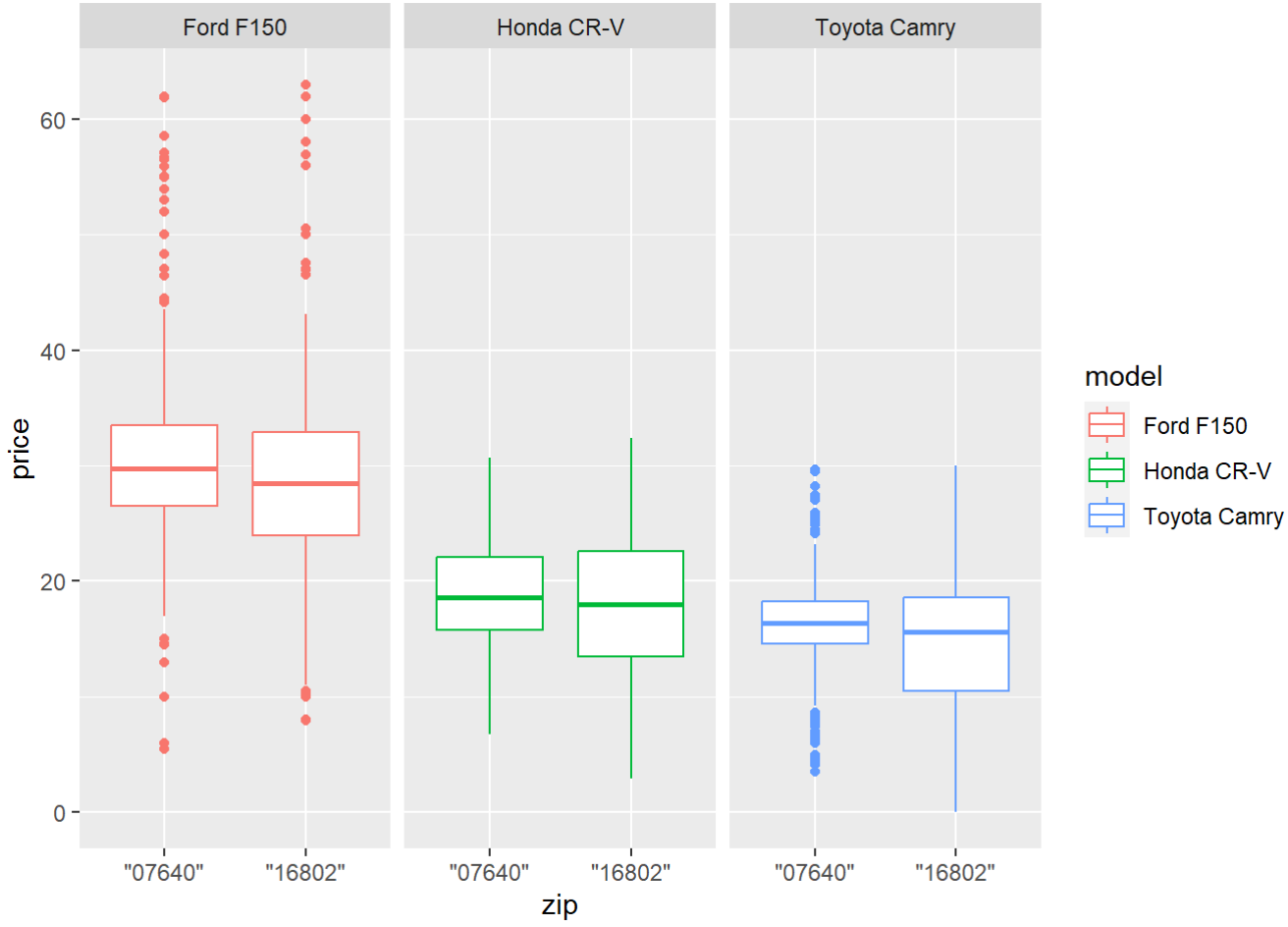
```
# combine used car data sets into one data set
used_cars <- rbind(data_ford16802, data_ford07640, data_honda16802, data_honda07640, data_toy
ota16802, data_toyota07640)

# calculate each car's age and create a new column
used_cars$age <- 2020 - used_cars$year

head(used_cars)
```

```
##   year  price mileage     zip     model age
## 1 2011 14.990 113.004 "16802" Ford F150   9
## 2 2013 22.999  64.362 "16802" Ford F150   7
## 3 2018 39.500  12.000 "16802" Ford F150   2
## 4 2016 22.000  86.456 "16802" Ford F150   4
## 5 2014 20.995  72.730 "16802" Ford F150   6
## 6 2016 38.998  24.996 "16802" Ford F150   4
```

I am going to use car data sets from zip codes State College, PA 16802 (rural) and Harrington Park, NJ 07640 (exurban). And, car models that I am going use are Ford F150 (truck), Toyota Camry (sedan) and Honda CR-V (suv). I am going to compare price of the cars by the cars' locations where they are being sold and models of the cars. And, I am going to see how the locations and the type of car affect the price of the cars.

EDA

```
gf_boxplot(price ~ zip | model, data = used_cars, color = ~model)
```




```
favstats(price ~ zip + model, data = used_cars)
```

```
##              zip.model   min       Q1  median       Q3    max     mean       sd
## 1    "07640".Ford F150 5.500 26.47125 29.7585 33.49250 61.900 30.64156 8.528525
## 2    "16802".Ford F150 7.995 23.91600 28.5000 32.85675 62.900 28.77388 8.846792
## 3   "07640".Honda CR-V 6.800 15.77500 18.6000 22.05150 30.663 18.76658 4.652094
## 4   "16802".Honda CR-V 2.900 13.47950 17.9965 22.56625 32.345 18.22998 6.120043
## 5 "07640".Toyota Camry 3.495 14.56600 16.4235 18.27375 29.687 16.09139 4.574641
## 6 "16802".Toyota Camry 0.000 10.50000 15.5975 18.60000 29.993 14.79024 5.681686
##     n missing
## 1 300       0
## 2 300       0
## 3 300       0
## 4 300       0
## 5 300       0
## 6 226       0
```

From the boxplot, it seems like Ford F150 (truck) has wider range of the price, while Honda CR-V (suv) and Toyota Camry (sedan) have narrower range of the price. Also, Honda CR-V and Toyota Camry have similar range of the price. For the zip codes, it seems like there is not that big of a difference in the range of price. Moreover, from the boxplot and the favstats chart, Ford F150's average price is higher than Honda CR-V and Toyota Camry's average price. Also, Honda CR-V seems to have slightly higher average price than Toyota Camry. In addition, Ford F150 seems to have higher standard deviation than Honda CR-V and Toyota Camry. For the zip codes, Harrington Park, NJ 07640 (exurban) seems to have slightly higher average price of used cars than State College, PA 16802 (rural). And, different zip code does not seem to make a huge difference between standard deviation. Overall, both model and zip have effect on the price, but model seems to have stronger effect on the price.

```
uc_anova <- aov(price ~ model + zip, data = used_cars)
summary(uc_anova)
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## model          2  64558   32279  726.85  < 2e-16 ***
## zip            1    651     651   14.66 0.000134 ***
## Residuals   1722  76474      44
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value of model is less than 2.2E-16 from the ANOVA table, and the p-value of zip is 0.0001336 from the ANOVA table. Since p-value of model is less than p-value of zip, we have sufficient evidence that the car model has strong effect on the price than zip. In order to use the p-values in the conclusion, conditions of ANOVA should be checked.

ANOVA Condition Check

Part I

```
# residual vs fitted plot
mplot(uc_anova, which = 1)
```

```
## Warning: The `augment()` method for objects of class `aov` is not maintained by the broom
team, and is only supported through the `lm` tidier method. Please be cautious in interpretin
g and reporting broom output.
##
## This warning is displayed once per session.
```

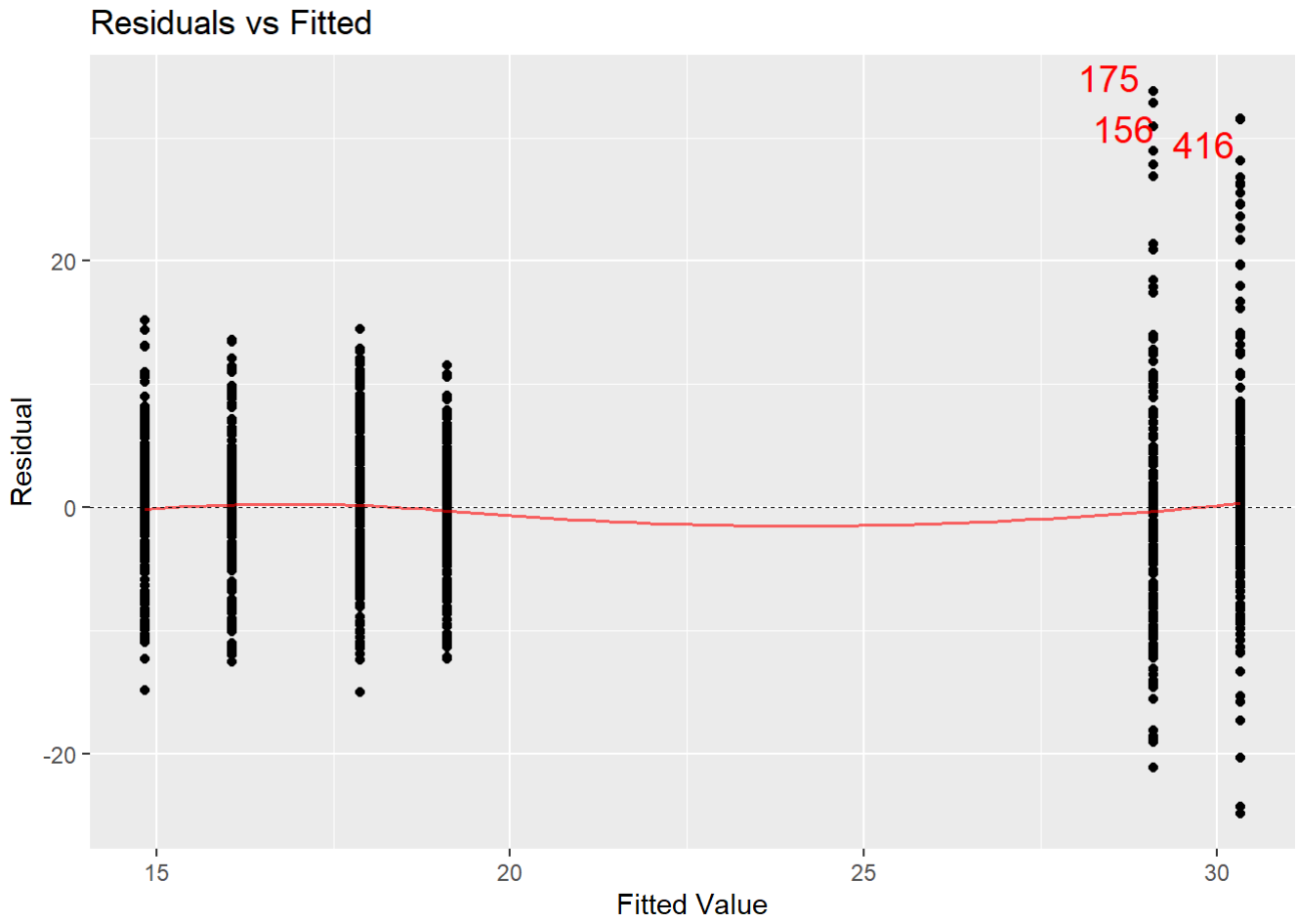
```
## mplot() doesn't know how to handle this kind of input.
```

```
## use methods("mplot") to see a list of available methods.
```

```
## mplot() doesn't know how to handle this kind of input.
```

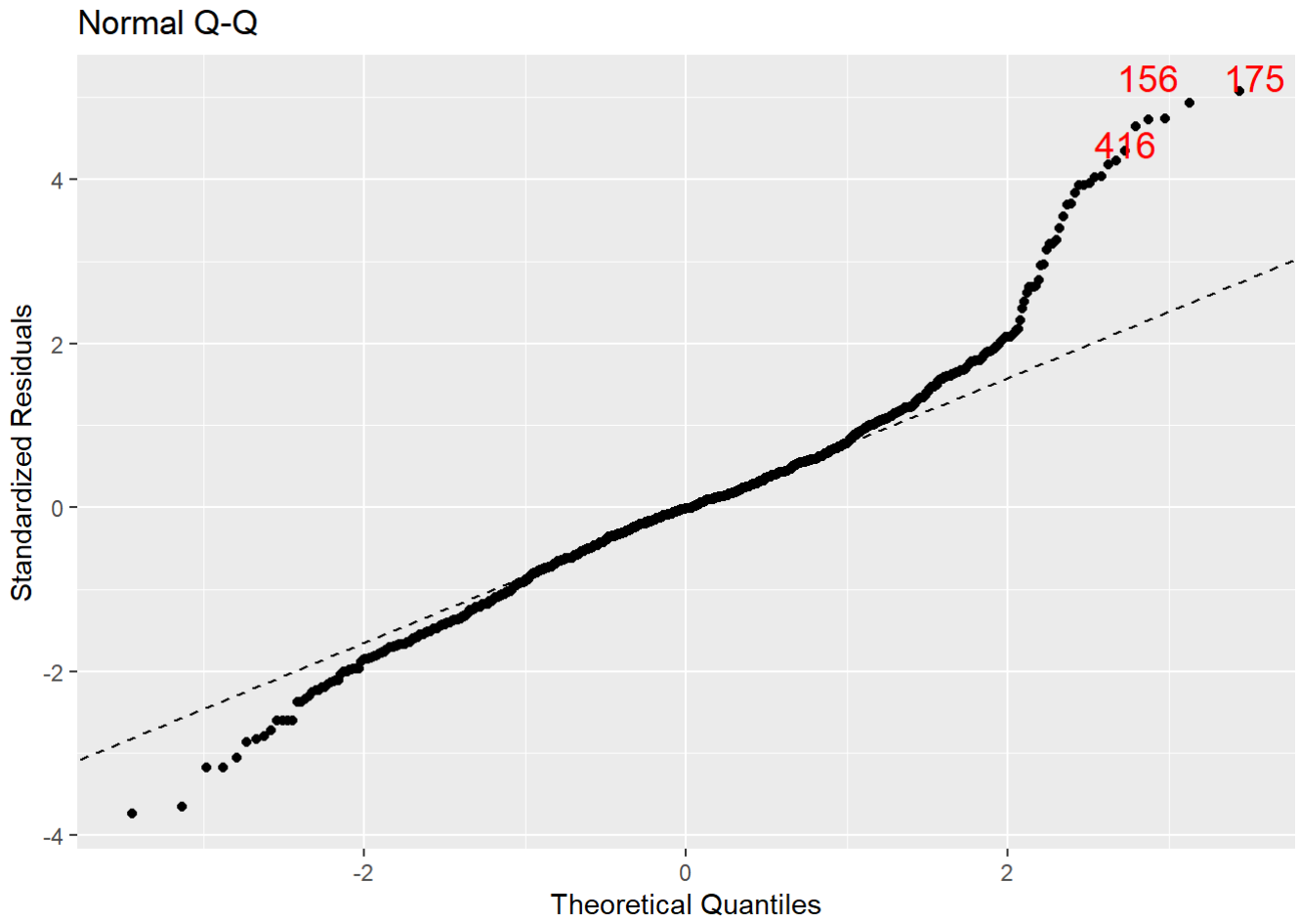
```
## use methods("mplot") to see a list of available methods.
```

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




```
# normal Q-Q plot
mplot(uc_anova, which = 2)
```

```
## mplot() doesn't know how to handle this kind of input.
## use methods("mplot") to see a list of available methods.
## mplot() doesn't know how to handle this kind of input.
## use methods("mplot") to see a list of available methods.
```



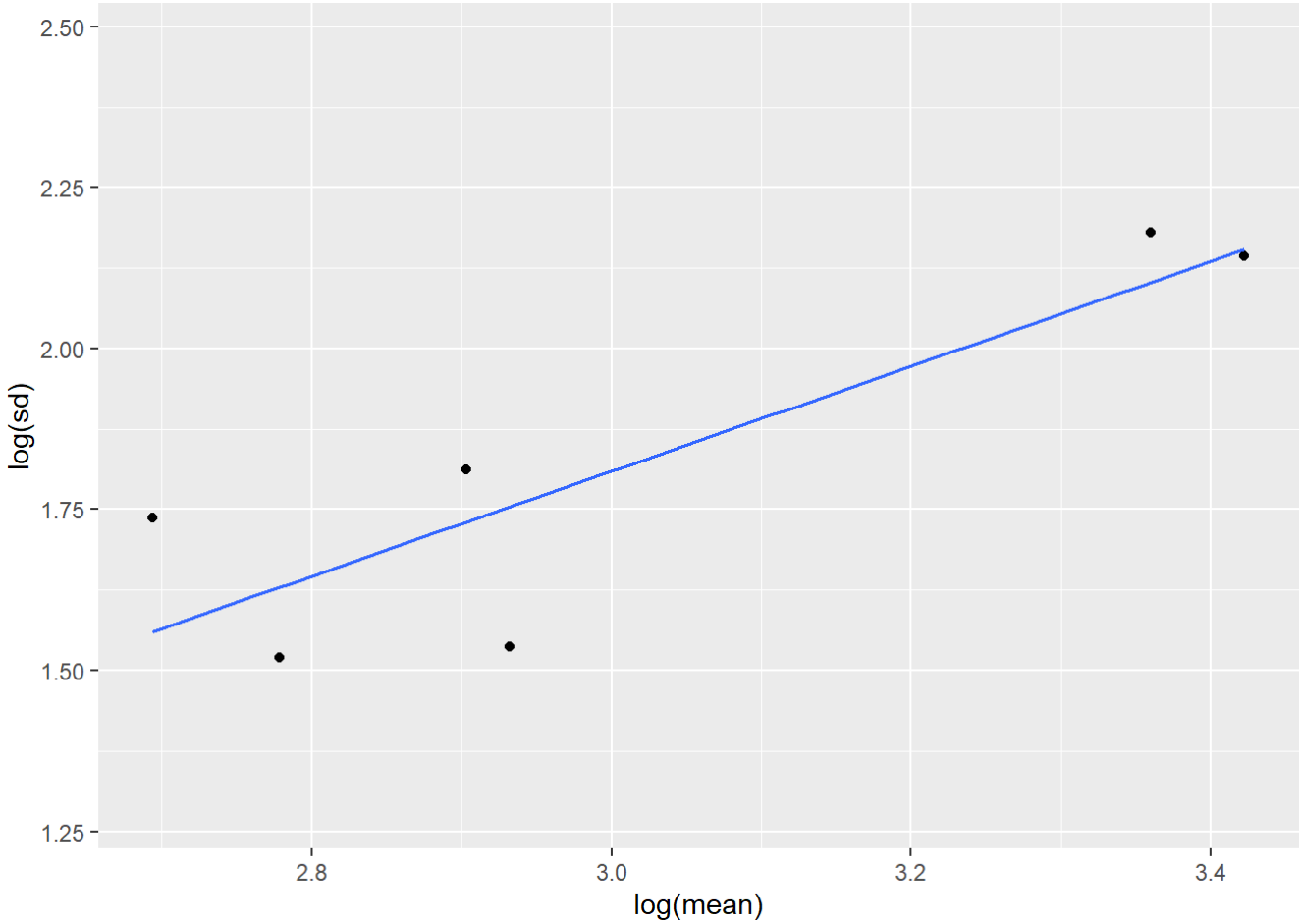
From the residuals vs fitted plot, it seems like equal variance condition for ANOVA is not met because the errors from the plot seem to form a fanning shape. From the normal Q-Q plot, it seems like normal distribution of error condition for ANOVA is not met because the residuals from the normal plot do not seem to lie along the line. This model does not meet the conditions of ANOVA, thus we would need to pre-process data in order to use ANOVA.

```
# create uc_log data for log(sd) vs log(mean) plot
uc_log <- used_cars %>%
  group_by(model, zip) %>%
  summarise (mean = mean(price), sd = sd(price))
```

```
## `summarise()` has grouped output by 'model'. You can override using the
## `.groups` argument.
```

```
# log(sd) vs log(mean) plot
gf_point(log(sd) ~ log(mean), data = uc_log) %>% gf_lm()
```

```
## Warning: Using the `size` aesthetic with geom_line was deprecated in ggplot2 3.4.0.
## ℹ Please use the `linewidth` aesthetic instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```




```
# finding slope
uc_log_slope <- lm(log(sd) ~ log(mean), data = uc_log)

summary(uc_log_slope)
```

```
##
## Call:
## lm(formula = log(sd) ~ log(mean), data = uc_log)
##
## Residuals:
##        1        2        3        4        5        6
## -0.01054  0.07738 -0.21686  0.08105 -0.10825  0.17723
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.6367     0.7133  -0.893   0.4225
## log(mean)     0.8154     0.2356   3.461   0.0258 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1603 on 4 degrees of freedom
## Multiple R-squared:  0.7497, Adjusted R-squared:  0.6871
## F-statistic: 11.98 on 1 and 4 DF,  p-value: 0.02579
```

In order for ANOVA model to meet ANOVA conditions, some kind of transformation is needed. By plotting a log(sd) vs log(mean) plot, we can decided what kind of transformation we need to perform. From Y^P, P is calculated by 1 - slope = 1 - 0.8154 = 0.1846. Since P is closer to 0 than 0.5, log transformation would improve the model to meet the ANOVA conditions.

Part II

```
# log transformation of price column
used_cars$log_price <- log(used_cars$price)

# remove outliers
used_cars <- used_cars[-c(1312, 1350),]

# ANOVA model after log transformation
uc_anova_log <- aov(log_price ~ model + zip, data = used_cars)

# residuals vs fitted plot
mplot(uc_anova_log, which = 1)
```

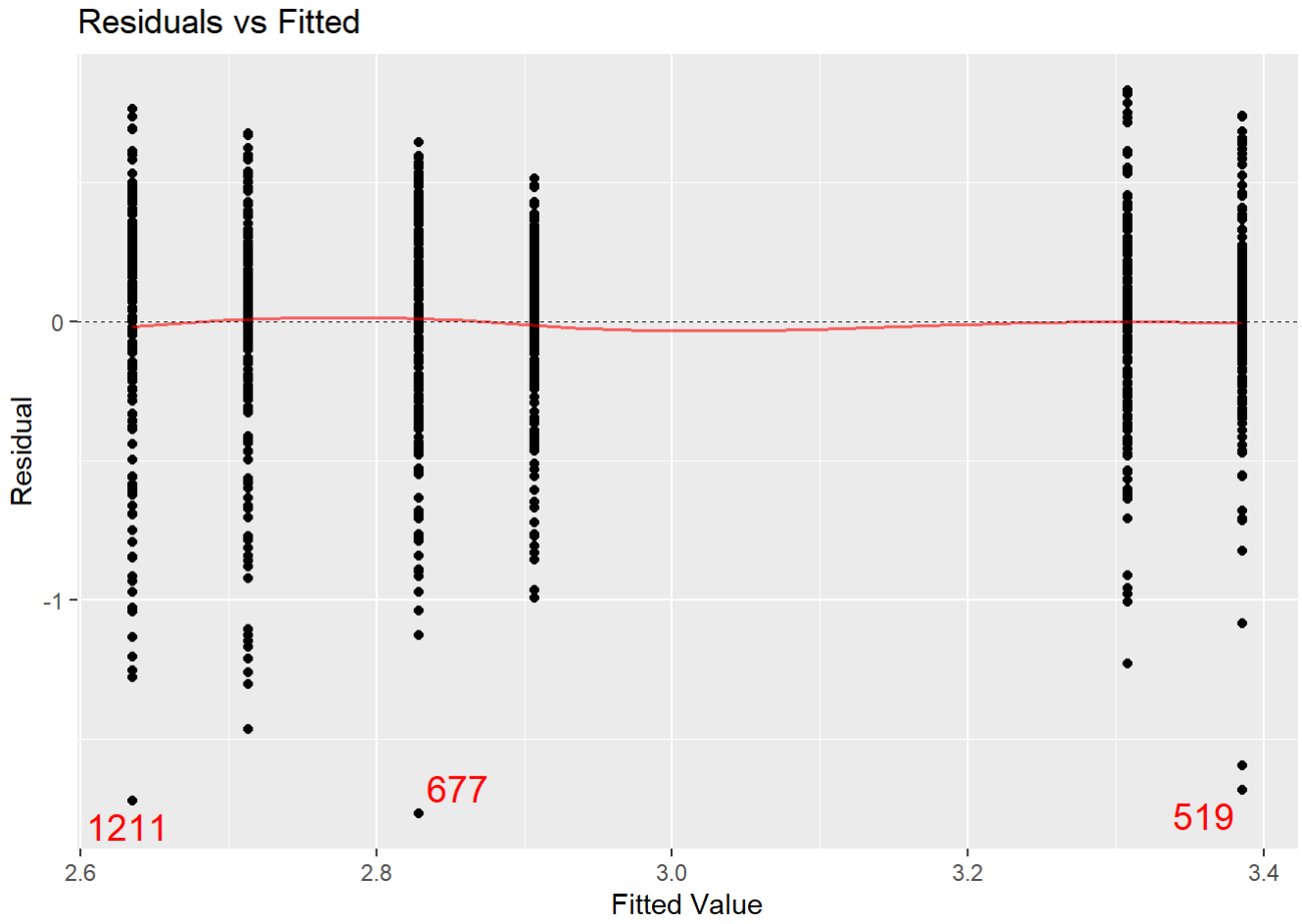
```
## mplot() doesn't know how to handle this kind of input.
```

```
## use methods("mplot") to see a list of available methods.
```

```
## mplot() doesn't know how to handle this kind of input.
```

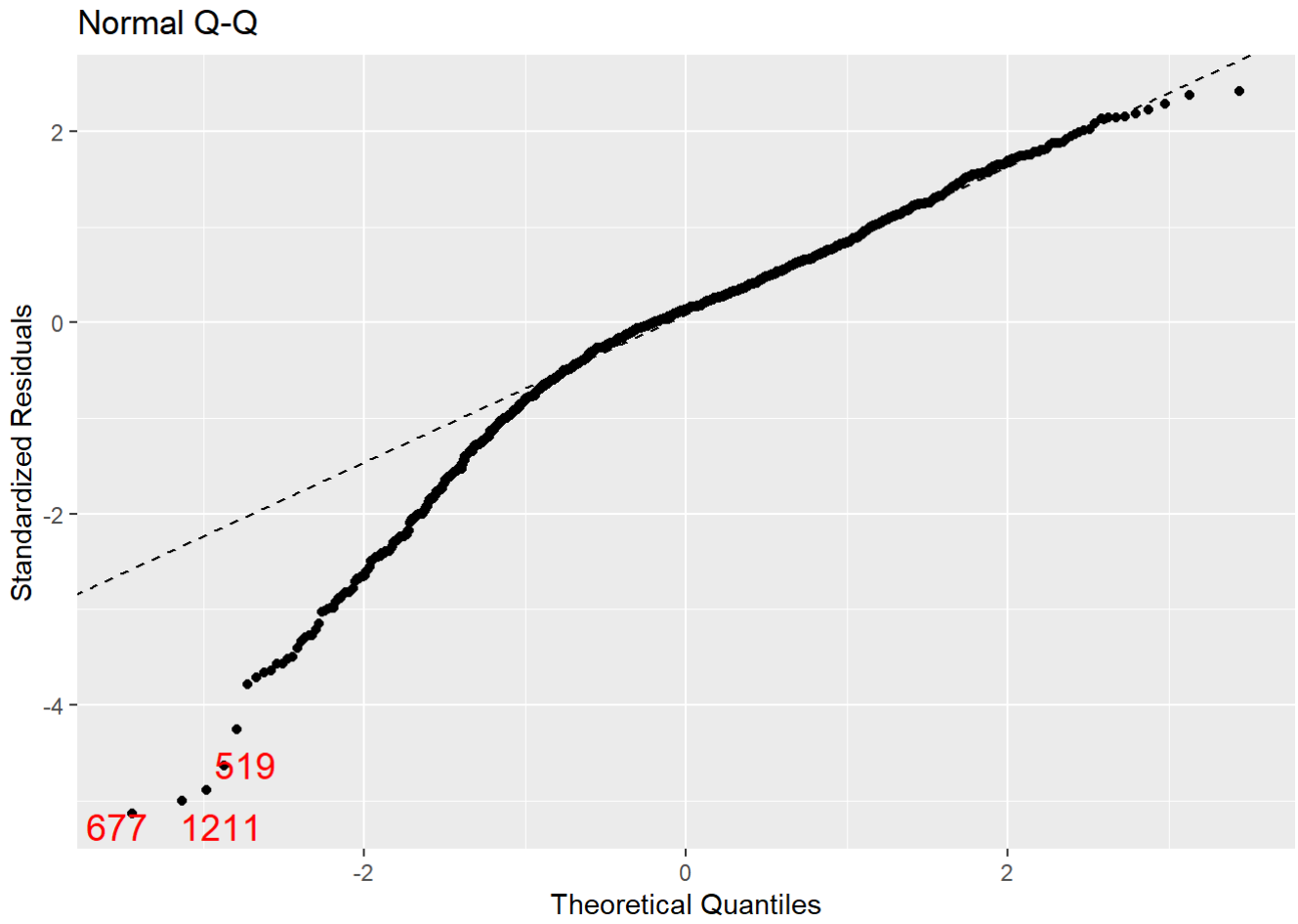
```
## use methods("mplot") to see a list of available methods.
```

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



```
# normal Q-Q plot
mplot(uc_anova_log, which = 2)
```

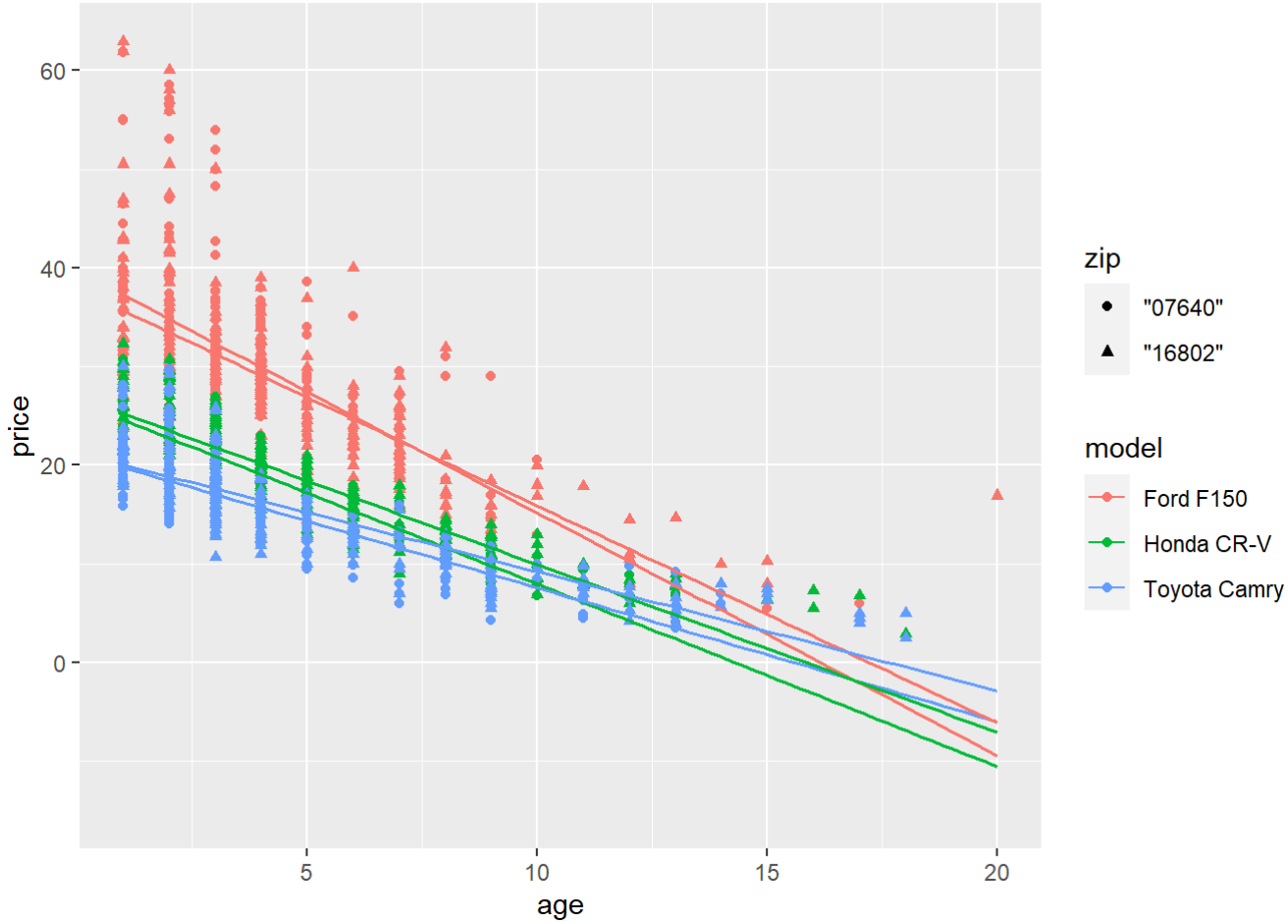
```
## mplot() doesn't know how to handle this kind of input.
## use methods("mplot") to see a list of available methods.
## mplot() doesn't know how to handle this kind of input.
## use methods("mplot") to see a list of available methods.
```



Performing log transformation successfully fixed unequal variance of errors, but even after the transformation, some errors from the normal plot do not lie along the line. Thus, we will proceed with caution.

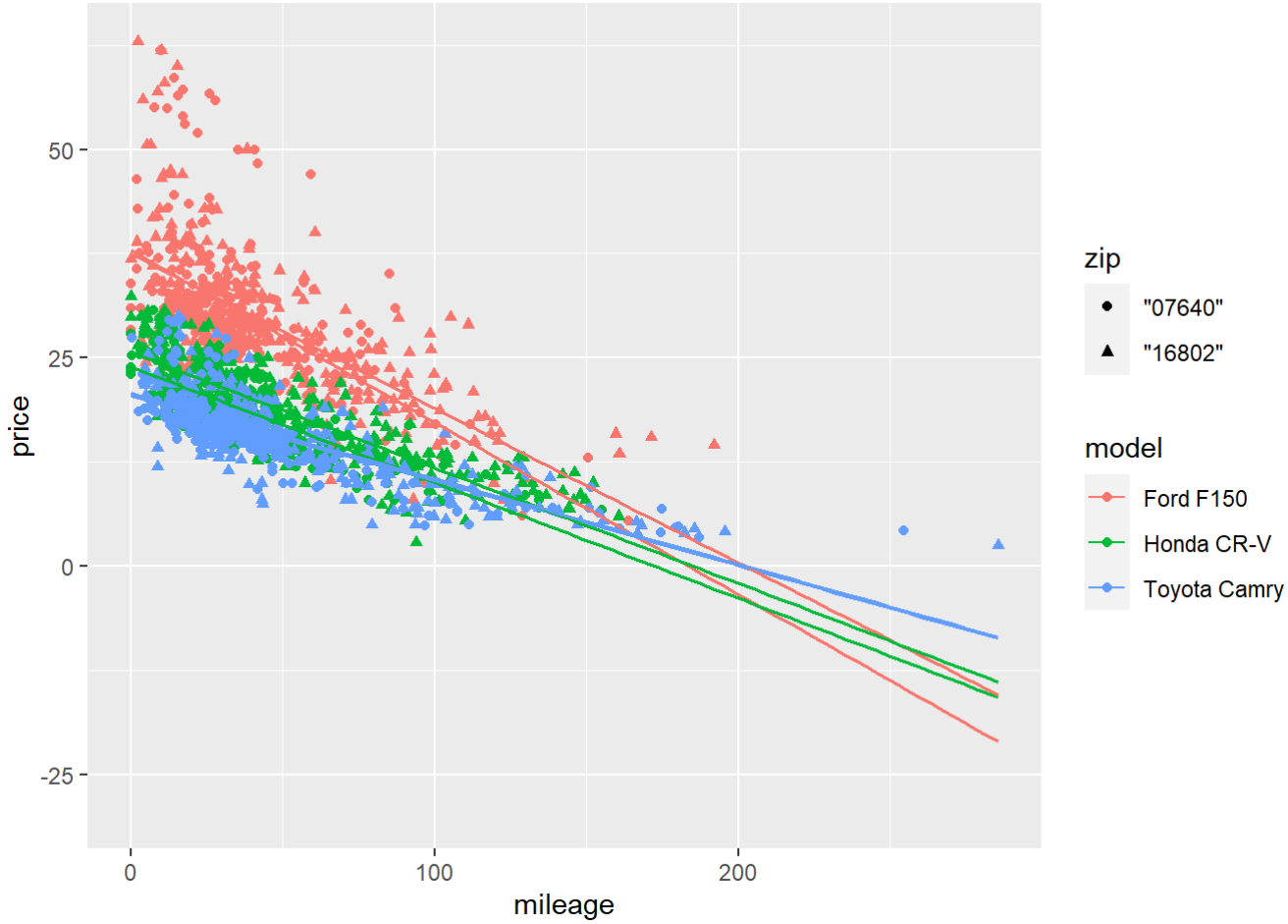
Part III

```
gf_point(price ~ age, col = ~ model, shape = ~ zip, data = used_cars) %>% gf_lm()
```



This data also has age and mileage variable. From price vs age graph, there are some interactions after age of 7.5 year, but since these cars' original price do not differ dramatically, after 7.5 years, which is pretty old, the cars' price would be similar. Thus, we should account age until around 7.5 year. Considering that, there is no interaction between car model and age.

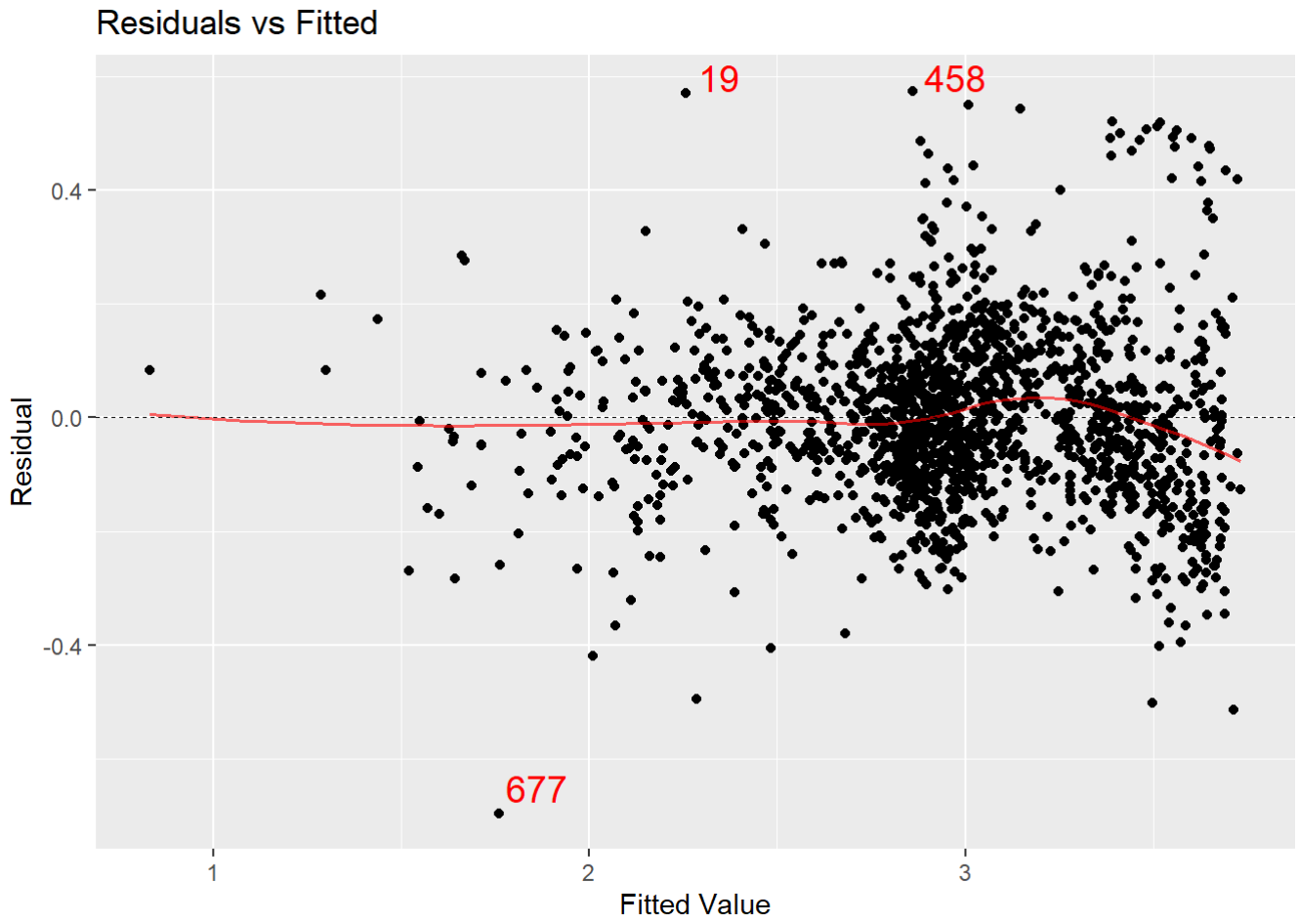
```
gf_point(price ~ mileage, col = ~ model, shape = ~ zip, data = used_cars) %>% gf_lm()
```



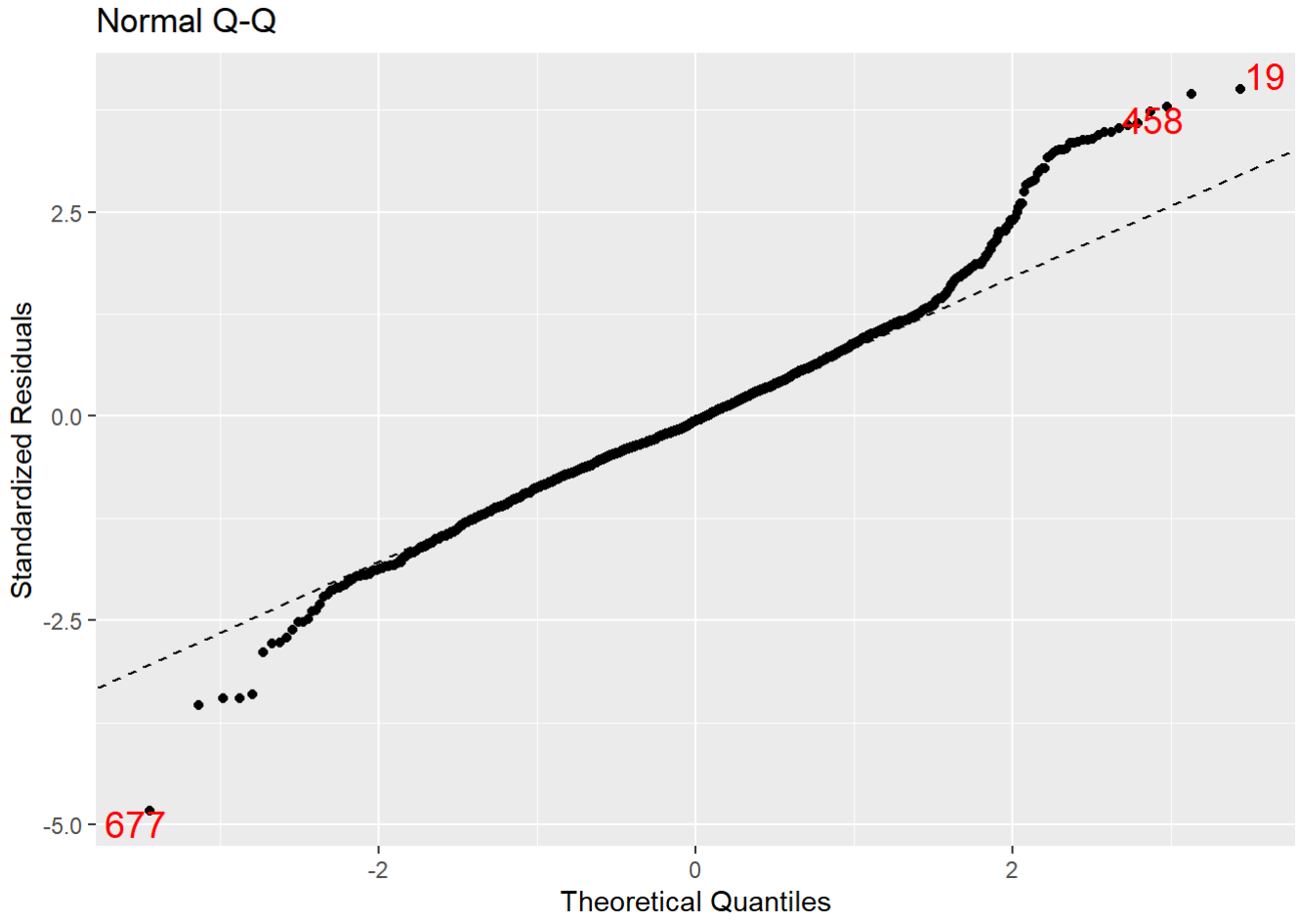
For price vs mileage graph, there is one interaction a little bit before 100,000 mileage and many interaction after 100,000 mileage. According to 100,000 mileage rule, used cars with around 100,000 mileage would be considered unreliable. Since all the interactions are located beyond 100,000 mileage, we should only account mileage that is below 100,000 mileage. Since Honda CR-V and Toyota Camry's original price is not dramatically different, it is possible that the cars' used price become similar after riding it around 100,000 mileage. Considering that, there is no interaction between car model and mileage. Considering some possible conditions, we can say there is no interaction for neither age nor mileage.

```
uc_age_mileage <- lm(log(price) ~ age + mileage + model + zip, data = used_cars)

# residual vs fitted plot
mplot(uc_age_mileage, which = 1)
```

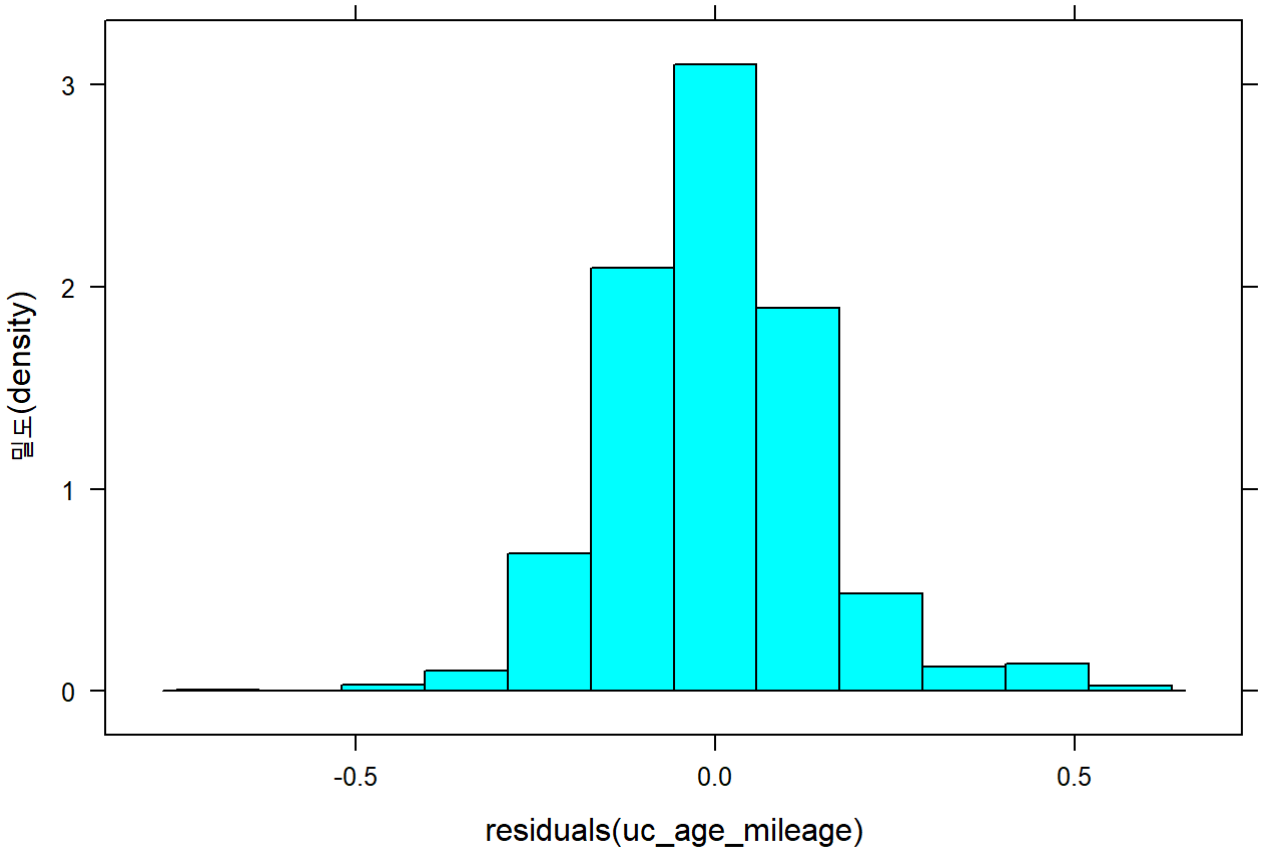
```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# normal Q-Q plot
mplot(uc_age_mileage, which = 2)
```



```
histogram(residuals(uc_age_mileage))
```



From the residual vs fitted plot, it seems like there is no fanning shape, thus equal variance condition of errors condition is met. From the Normal plot, it seems like many errors do not lie along the line, but from the histogram, it seems like there is no important skewness. Therefore, normal distribution of errors condition is also met. Both equal variance of errors condition and normal distribution of errors condition are met, so we can make ANOVA conclusion.

Conclusion

```
summary(uc_age_mileage)
```

```
##
## Call:
## lm(formula = log(price) ~ age + mileage + model + zip, data = used_cars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.69563 -0.08954 -0.00852  0.08153  0.57429
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.7591969  0.0081116 463.434  < 2e-16 ***
## age               -0.0705166  0.0018900 -37.309  < 2e-16 ***
## mileage           -0.0038489  0.0001619 -23.780  < 2e-16 ***
## modelHonda CR-V   -0.4108105  0.0084626 -48.544  < 2e-16 ***
## modelToyota Camry -0.6011225  0.0087760 -68.496  < 2e-16 ***
## zip"16802"         0.0430536  0.0071920   5.986 2.61e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1457 on 1718 degrees of freedom
## Multiple R-squared:  0.8934, Adjusted R-squared:  0.893
## F-statistic:  2878 on 5 and 1718 DF,  p-value: < 2.2e-16
```

Based on the analysis of the provided linear regression model, we conclude that age, mileage, specific car models (Honda CR-V and Toyota Camry), and the zip code have a significant impact on the price of used cars. The model indicates that as a car ages or as its mileage increases, its price tends to decrease, with estimated reductions of approximately 6.84% ((e^-0.0705166 - 1) * 100) per year of age and 0.38% ((e^-0.0038489 - 1) * 100) per unit increase in mileage. Moreover, Honda CR-Vs and Toyota Camry are priced lower than other models in the data set, by about 33.52% ((e^-0.4108105 - 1) * 100) and 45.38% ((e^-0.6011225 - 1) * 100) respectively. Additionally, cars sold in zip code “16802” are priced about 4.40% ((e^0.0430536 - 1) * 100) higher than those in other areas. The model’s high R-squared value of 0.8934 suggests that these factors collectively explain a significant portion of the variability in used car prices.

```
anova(uc_age_mileage)
```

```
## Analysis of Variance Table
##
## Response: log(price)
##             Df  Sum Sq Mean Sq  F value    Pr(>F)
## age          1 181.405 181.405 8548.408 < 2.2e-16 ***
## mileage      1  15.188  15.188  715.686 < 2.2e-16 ***
## model        2 108.045  54.023 2545.728 < 2.2e-16 ***
## zip          1   0.760   0.760   35.836 2.607e-09 ***
## Residuals 1718  36.458   0.021
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA analysis conclusively demonstrates that age, mileage, car model, and zip code are all statistically significant factors affecting the price of used cars. Age and mileage have the most pronounced effects, strongly suggesting that the market value of a car decreases as it gets older and as its mileage increases. The model of the car also plays a critical role in determining its price, while the impact of the selling location (zip code) is comparatively smaller but still significant. This analysis provides valuable insights for understanding the pricing dynamics in the used car market.