MATH 324: Multiple Linear Regression project

Caitlin Garcia

Collaboration rules:

You may consult with up to two classmates for help with this project, but use your own data (must have different make/model/zip codes). Please identify who you collaborate with here:

Read this document before you submit it to ensure there is not a ton of extra output that does not contribute to the analysis or communication. Also, I recommend using the spell-checker in RStudio (Edit -> Check Spelling). Note that you will need to closely follow the instructions on the Canvas assignment page to complete this project successfully.

Introduction

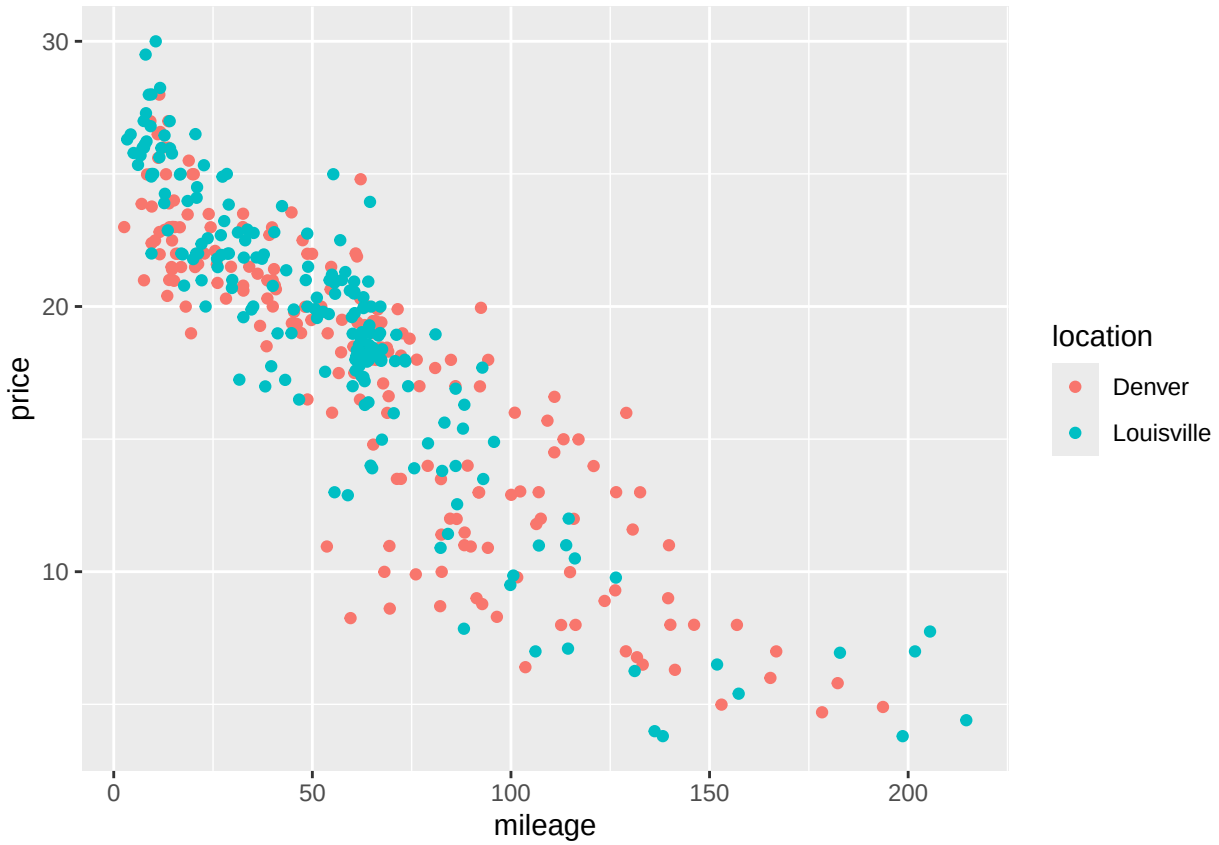
For this project, I chose to investigate used car prices for the Nissan Altima, comparing prices in two distinct geographical locations: Denver, Colorado, and Louisville, Kentucky. Frankly, choosing this car was a random choice, as I have nearly no knowledge about cars. I selected these locations because of their significant distance from each other, ensuring that there would be meaningful differences in the used car markets, such as regional economic factors, demand, and local vehicle availability. Before analyzing the data, I anticipate finding price differences between the two locations, with potentially higher prices in Denver due to higher cost of living and more urban demand compared to Louisville. However, this is just an assumption, and the data analysis will ultimately reveal whether this prediction holds true.

Research question 1

Assuming a linear relationship between price and mileage, is there a difference in price between the locations?

Exploratory data analysis

To explore whether a difference in price exists between the two locations while accounting for mileage, I first examined the basic statistics and distribution of the data. I plotted the relationship between price and mileage, using different colors to distinguish between the two locations. Denver, represented in (I believe) red, and Louisville, in blue, showed similar patterns, thought Louisville appeared to have slightly higher prices for vehicles with comparable mileage. The plots also seem to hint at Louisville having a faster depreciation rate.



Figure(s):

EDA TABLE 1 | |sample size | mean price | sd of price | mean mileage | sd of mileage | |:———:|:———-:|:———:|:———:|:———:|:———:| |Denver| 196 | 17.55 | 5.60 | 63.98 | 41.76 | |Louisville| 196 | 19.26 | 5.40 | 55.05 | 38.89 |

Comments: The summary statistics, particularly the mean price, provides an early indication that Louisville might have higher prices for cars, thought it is harder to compare between similar mileage, as their mean mileages have some difference. There also seems to be more variation in price and mileage in Denver than there is in Louisville.

Model fitting

MODEL SUMMARY TABLE 1: (same slope different intercepts)

	estimate	test-statistic	p-value
intercept	25.283	92.314	<2e-16
mileage	-0.121	-37.795	<2e-16
locationLouisville	0.628	2.425	0.016

MODEL SUMMARY TABLE 2: (different slopes and different intercepts)

	estimate	test-statistic	p-value
intercept	25.048	75.122	<2e-16
mileage	-0.117	-26.836	<2e-16
locationLouisville	1.096	2.387	0.018

	estimate	test-statistic	p-value
mileage:locationLouisville	-0.008	-1.235	0.218

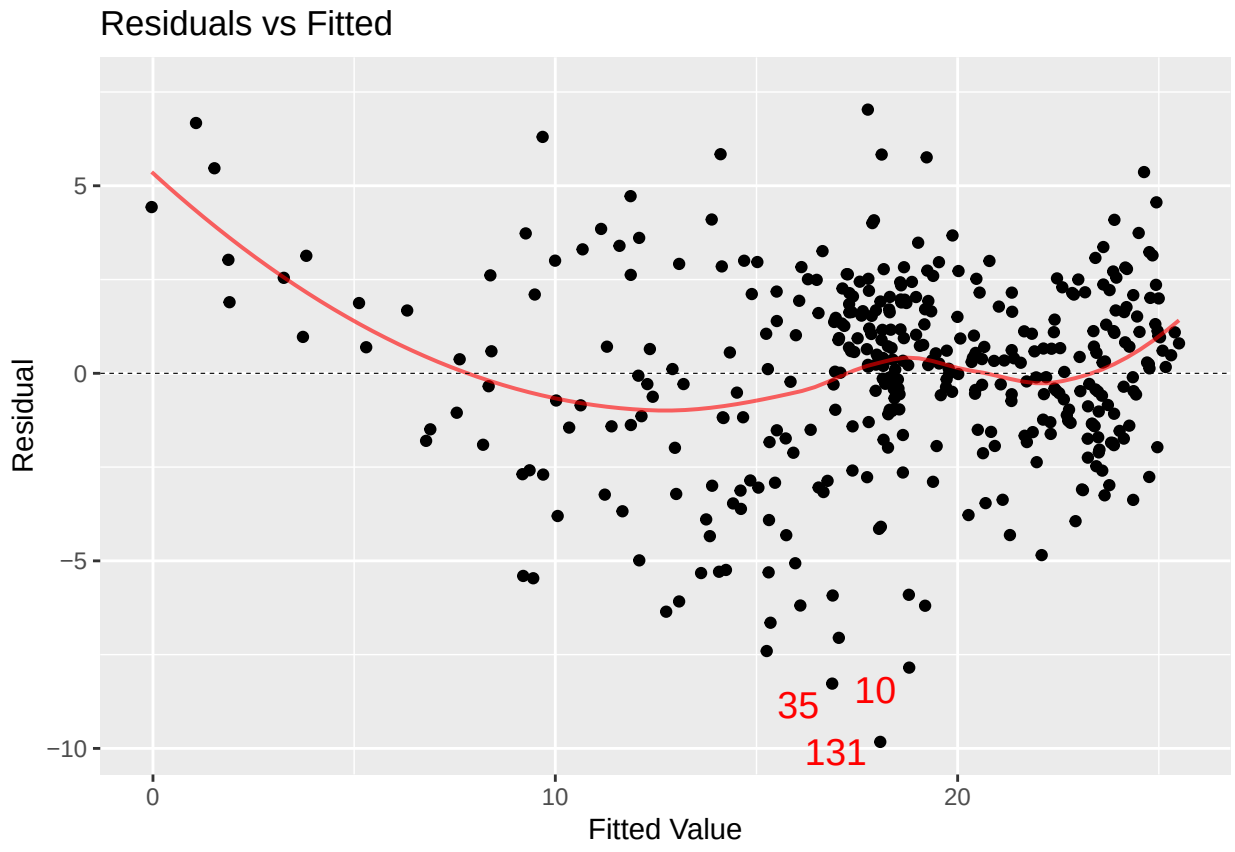
Fitted model for (location 1): $\widehat{price} = 25.283 - 0.121(mileage)$ Fitted model for (location 2): $\widehat{price} = 25.911 - 0.121(mileage)$

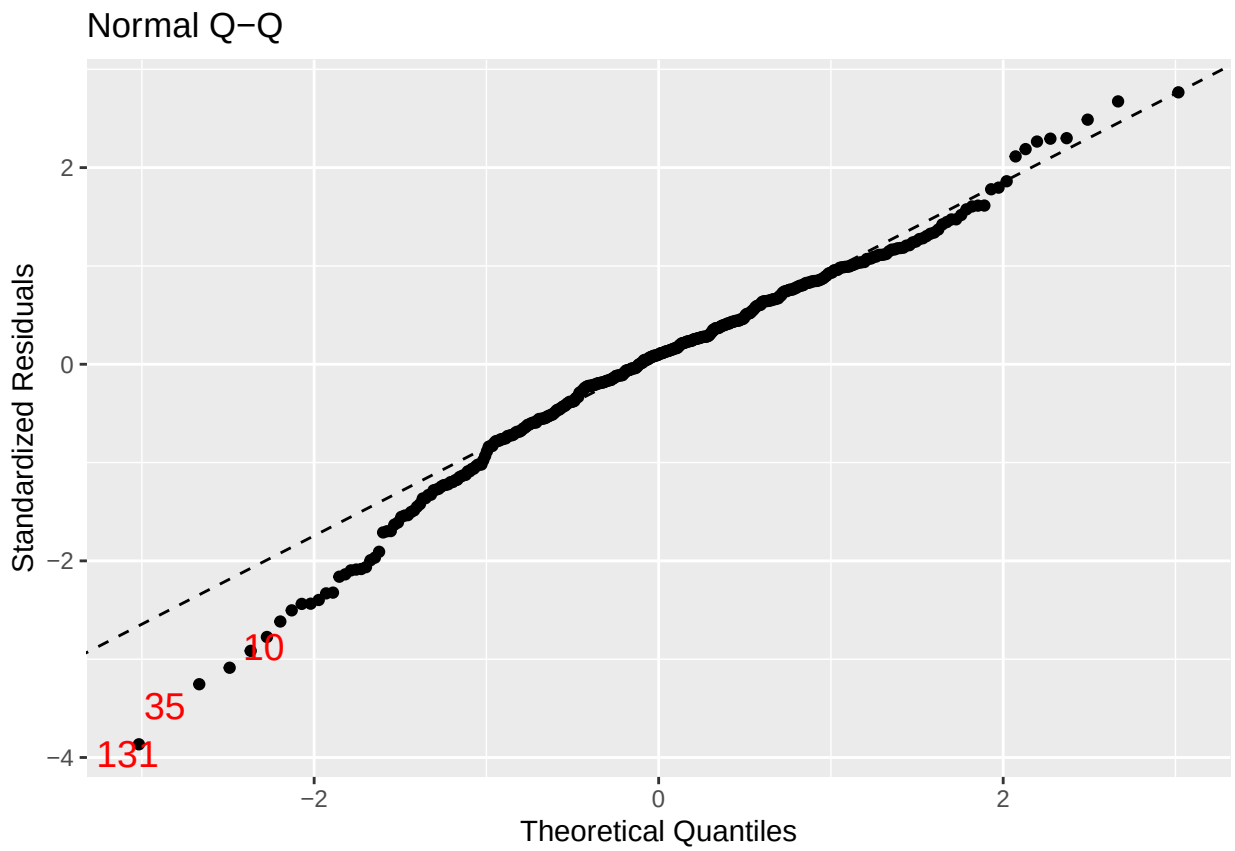
Comments When comparing the two models, we find that the second model, which allows for different slopes and intercepts between locations, does not significantly improve the fit over the first model. The interaction term between mileage and location is not statistically significant (p=0.218), meaning that there is no string evidence that the effect of mileage on price differs between Denver and Louisville. As a result, the first model, which assumes the same slope for both locations but different intercepts, is the more appropriate choice.

Assess

Figures:

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





Comments While the QQ plot seems to be fine, showing that normality is met in this model, the residuals v. fitted plot seems to take a cornucopia shape, telling us that there is non-linearity and non-equal variance in this plot. To fix this, I would likely introduce a quadratic term.

Use

After accounting for the influence of mileage, there is a significant difference in price between the two locations. This means that cars in Louisville have a higher price even after accounting for mileage. In the model, the coefficient for locationLouisville is 0.628, with a p-value of 0.016, showing that this variable is statistically significant, which suggests that cars in Louisville are priced about $628 higher than those in Denver on average.

Research Question 2

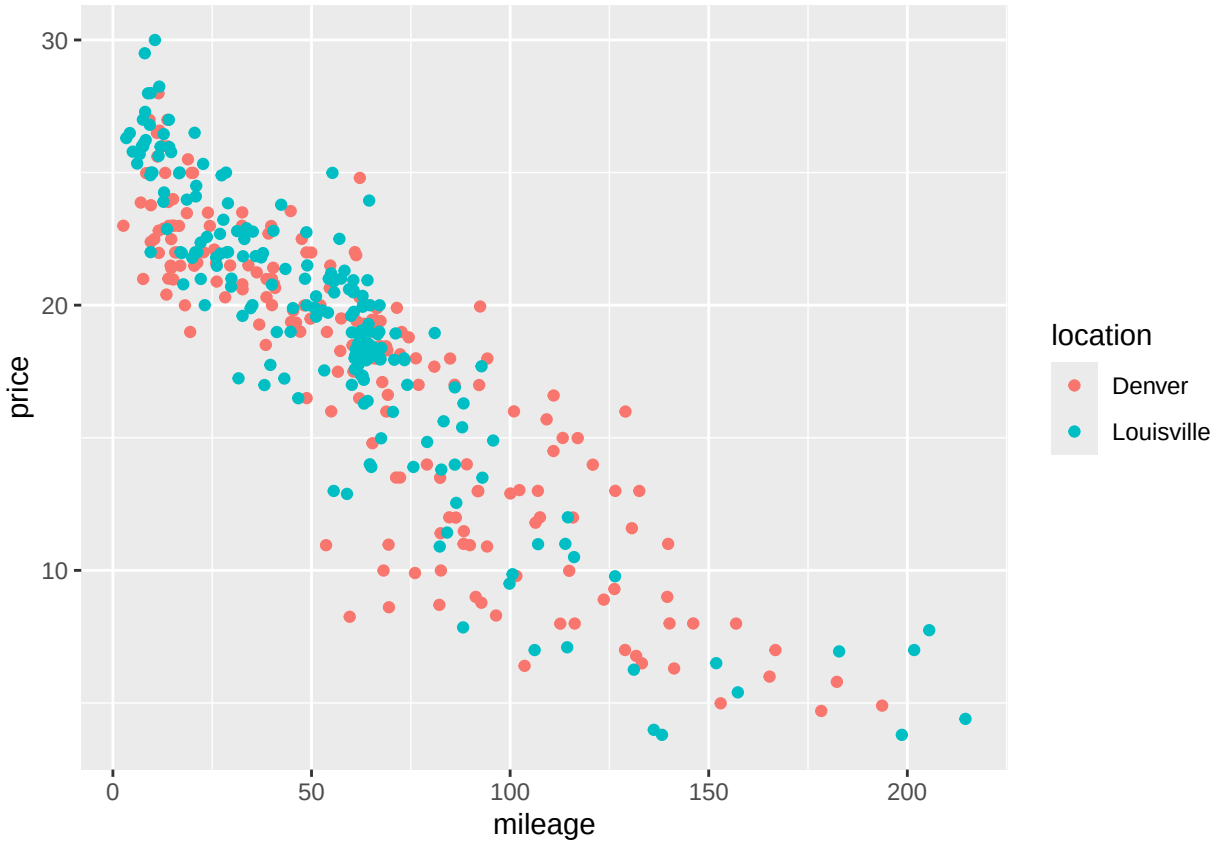
After accounting for a linear relationship between age and price and between mileage and price, is there a difference in price between the locations?

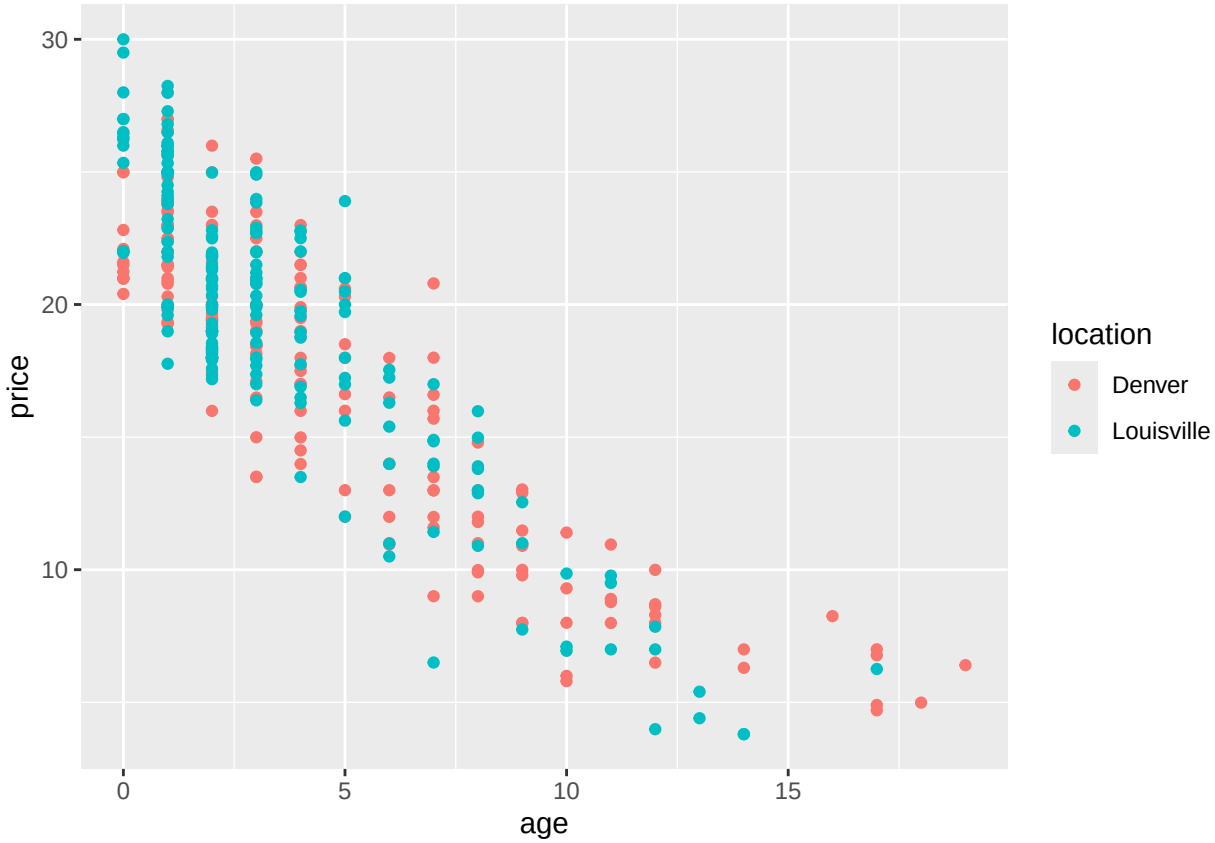
Exploratory data analysis

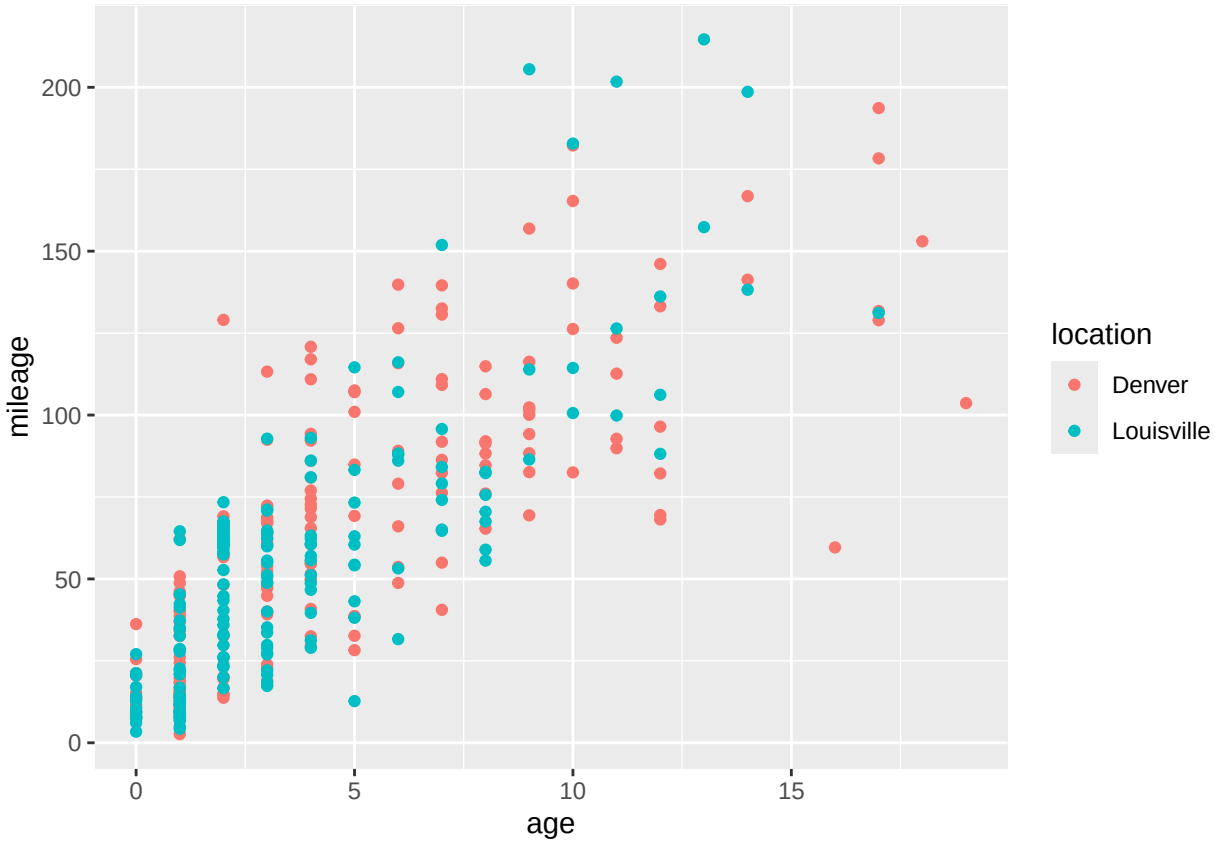
The exploratory data analysis showed significant relationships among the variables of interest. The correlation matrix especially indicated a strong negative correlation between price and age (r=-0.874), almost as significant as that between price and mileage (r = -0.888). This is also made clear in the Price v. Age scatter plot. Luckily, everything seems to have a linear trend. I see that there is a positive relationship between mileage and age, which makes me believe that age would be a good interaction term.

Figures:

```
##               price    mileage        age
## price     1.0000000 -0.8876150 -0.8740035
## mileage -0.8876150  1.0000000  0.7742309
## age     -0.8740035  0.7742309  1.0000000
```




price
30
20
10
0
5
10
15
age
location
Denver
Louisville



Model fitting

```
##
## Call:
## lm(formula = price ~ mileage + age + as.factor(location), data = car_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3461 -1.4725 -0.0442  1.3524  5.4100
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   25.304878   0.211388 119.708   <2e-16 ***
## mileage                       -0.072217   0.003876 -18.630   <2e-16 ***
## age                           -0.688936   0.042320 -16.279   <2e-16 ***
## as.factor(location)Louisville  0.331369   0.200741   1.651   0.0996 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.967 on 388 degrees of freedom
## Multiple R-squared:  0.8758, Adjusted R-squared:  0.8749
## F-statistic: 912.3 on 3 and 388 DF,  p-value: < 2.2e-16
```

SUMMARY TABLE 3:

	estimate	test-statistic	p-value
intercept	25.305	119.708	<2e-16
mileage	-0.072	-18.630	<2e-16
age	-0.689	-16.279	<2e-16
location	0.331	1.651	0.100

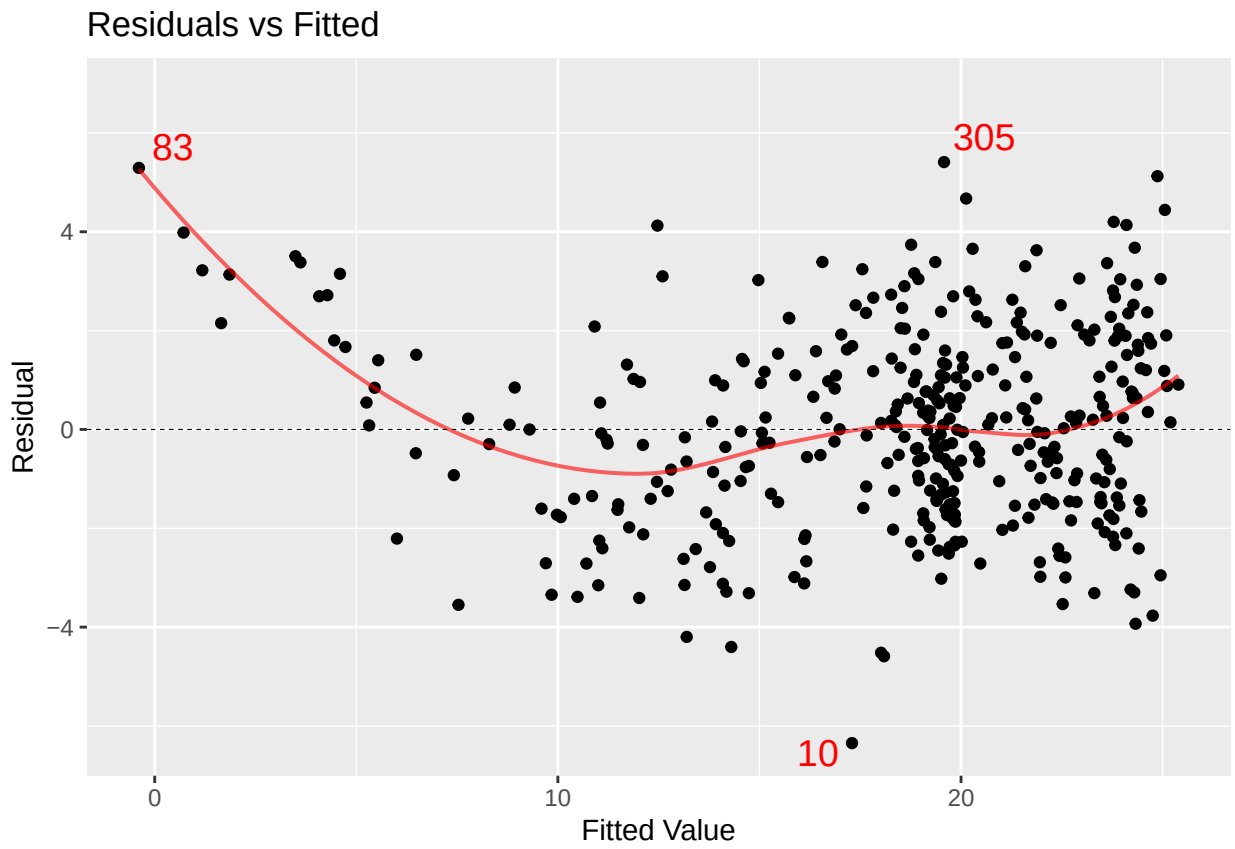
Interpretations in context

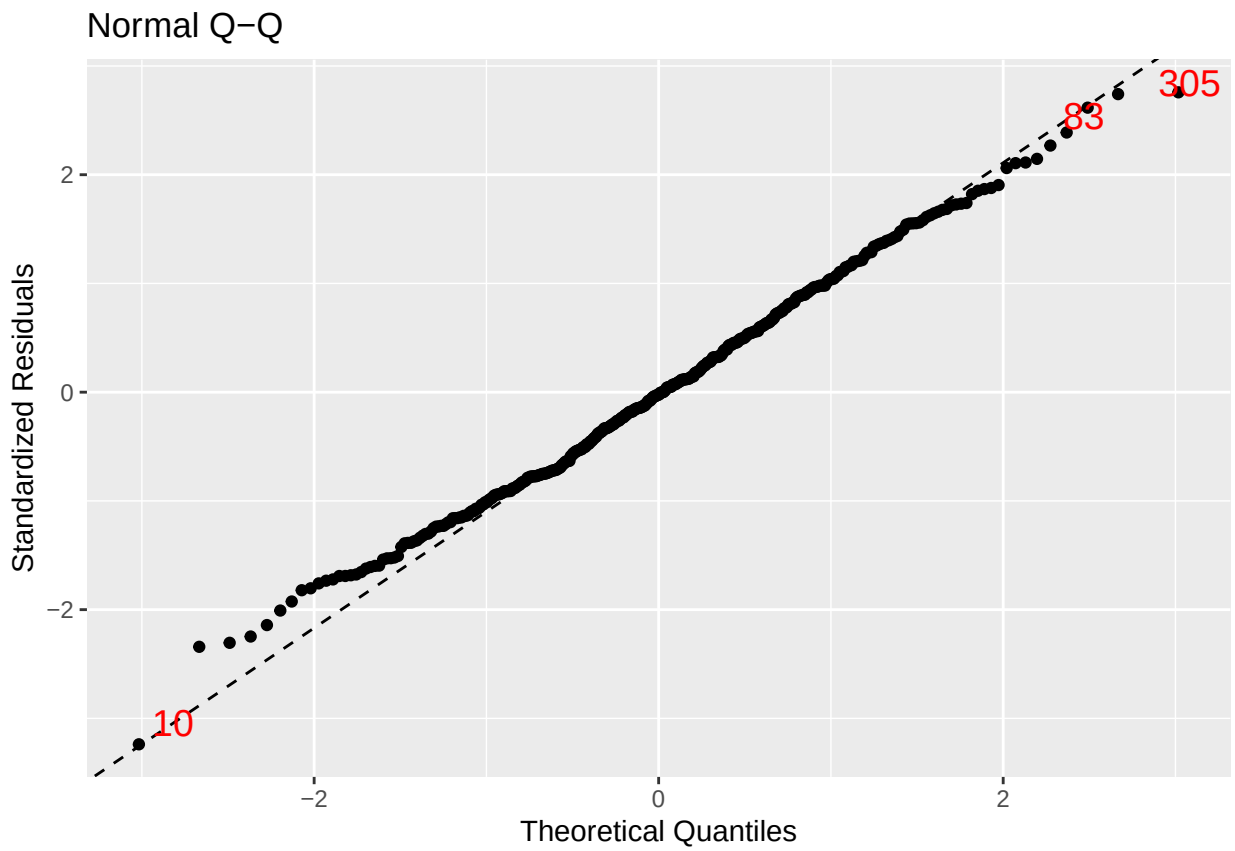
- intercept: $25,305, the expected price of a used car when both mileage and age are zero.
- mileage: When accounting for age and location, for each unit increase in mileage, the predicted price of a used car decreases by approximately $72.
- age: When accounting for mileage and location, for each additional year of age, the predicted price of a used car decreases by $689.
- location: Cars in Louisville are priced, on average, about $331 higher than those in Denver, though this is not a statistically significant variable as the p-value is about 0.05.

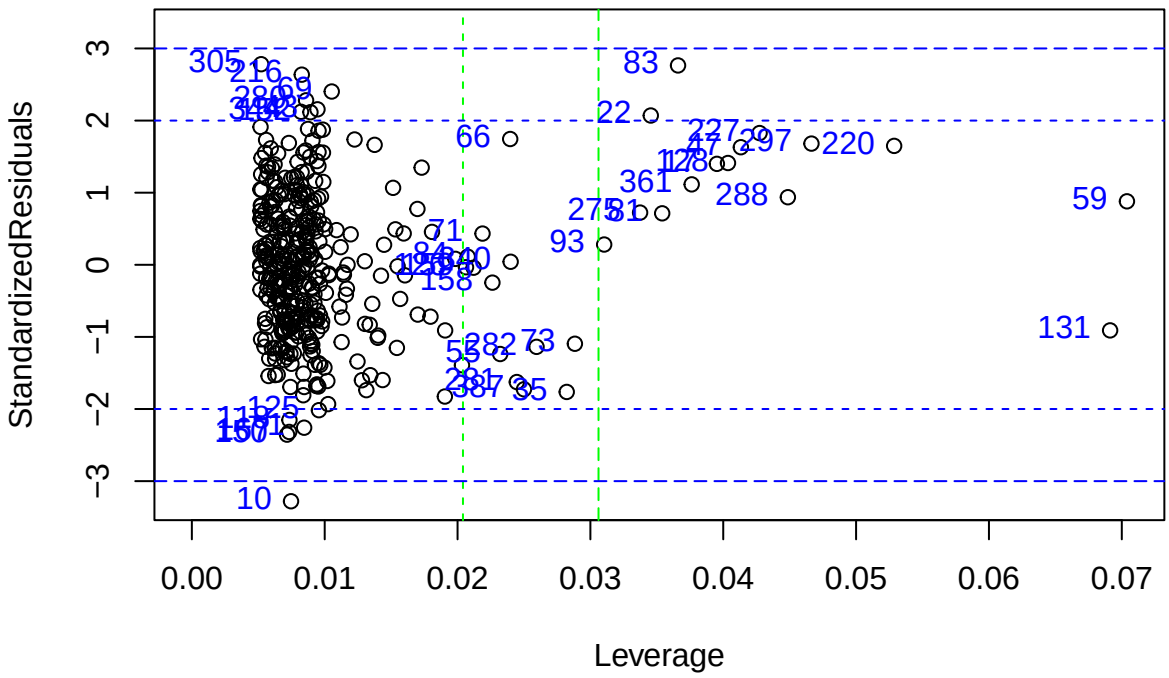
Assess

Figures:

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




Normal Q−Q
Standardized Residuals
Theoretical Quantiles
2
0
−2
−2
0
2
305
83
10



Comments Like with the last model we fitted, normality seems to be met, but there is a cornucopia effect in the residuals v. fitted. Therefore, there are issues with both linearity and equal variance. I would assume that, like the last model, having a polynomial term would help solve this issue. However, the Cook's plot shows that there is are no influential points at the very least.

Use

Because of the p-value for the location variable in this model (p = 0.1), we do not have significant evidence that there is a difference in price between the locations after accounting for both mileage and age. The lack of difference means that the observed difference in prices may not be reliable and could be due to random variation rather than a true effect of location.

Research question 3

What is the best model for predicting price using the variables available?

Choose

Final fitted model: Fitted model for (location 1): $\widehat{price} = 18.105 - 96.945(mileage^1) + 10.167(mileage^2)$
Fitted model for (location 2): $\widehat{price} = 18.700 - 96.945(mileage^1) + 10.167(mileage^2)$

summary of final model:

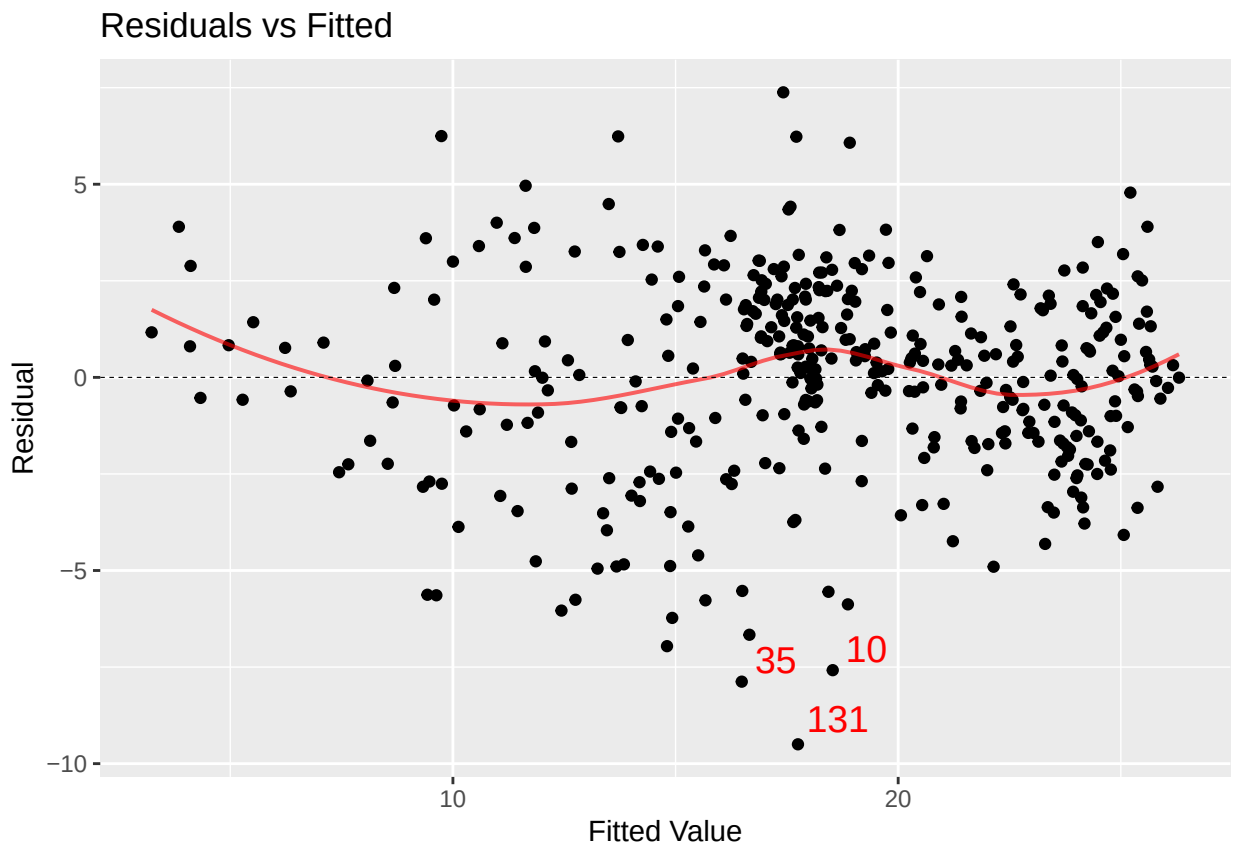
```
##
## Call:
## lm(formula = price ~ poly(mileage, 2) + location, data = car_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.4994 -1.4376  0.2656  1.7194  7.3793
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          18.1047     0.1791 101.088  < 2e-16 ***
## poly(mileage, 2)1  -96.9453     2.5144 -38.556  < 2e-16 ***
## poly(mileage, 2)2   10.1668     2.5003   4.066 5.79e-05 ***
## locationLouisville   0.5956     0.2541   2.344   0.0196 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.499 on 388 degrees of freedom
## Multiple R-squared:  0.7996, Adjusted R-squared:  0.798
## F-statistic: 515.9 on 3 and 388 DF,  p-value: < 2.2e-16
```

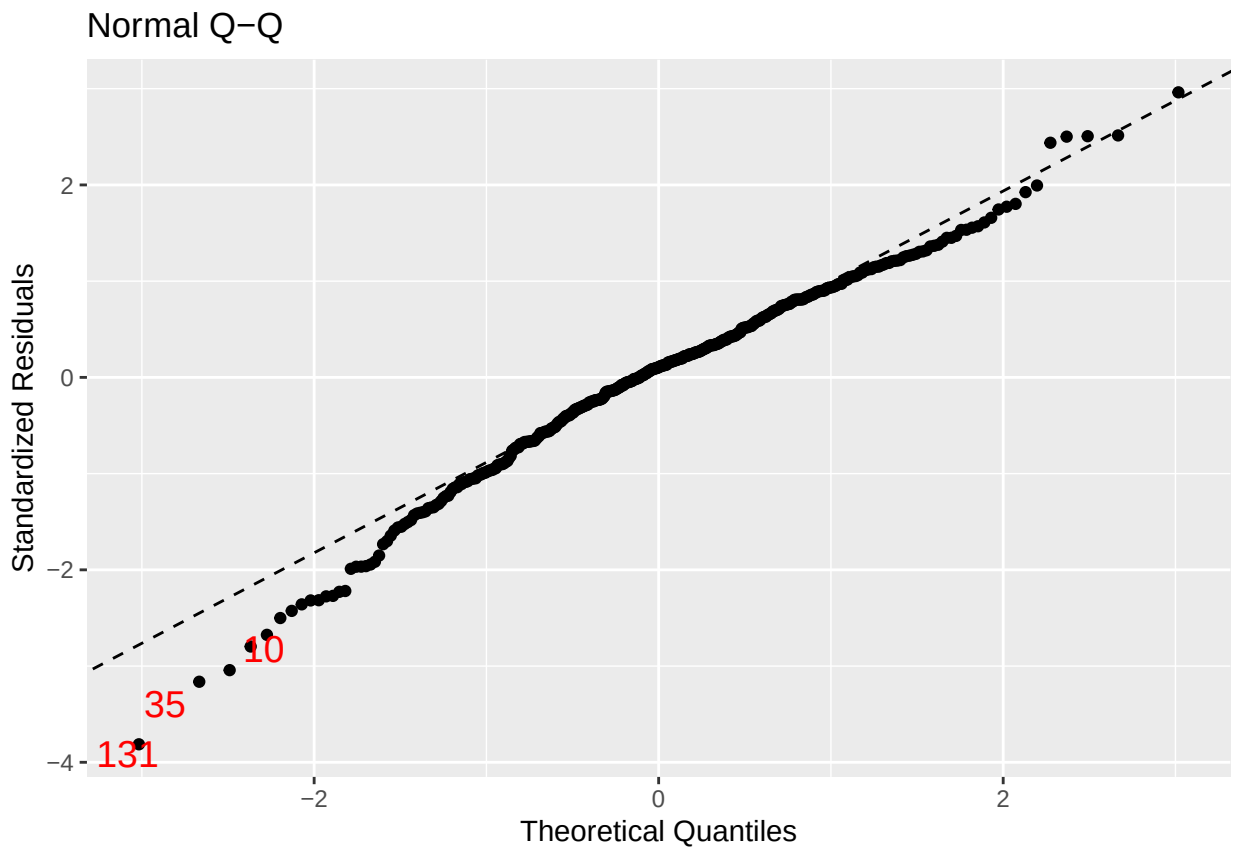
Comments I chose a model that did not include age because when age was included in my second model, location was not statistically significant anymore. Since the goal is to predict price between two locations, I need to somehow include location in my model. I tried multiple different models that included age and location (i.e. trying quadratics and interaction terms), but all of them followed the trend of age making location not significant, so I ultimately had to take it out. So, instead, I decided to make a better version of the first model, where the issues with linearity and equal variance are addressed.

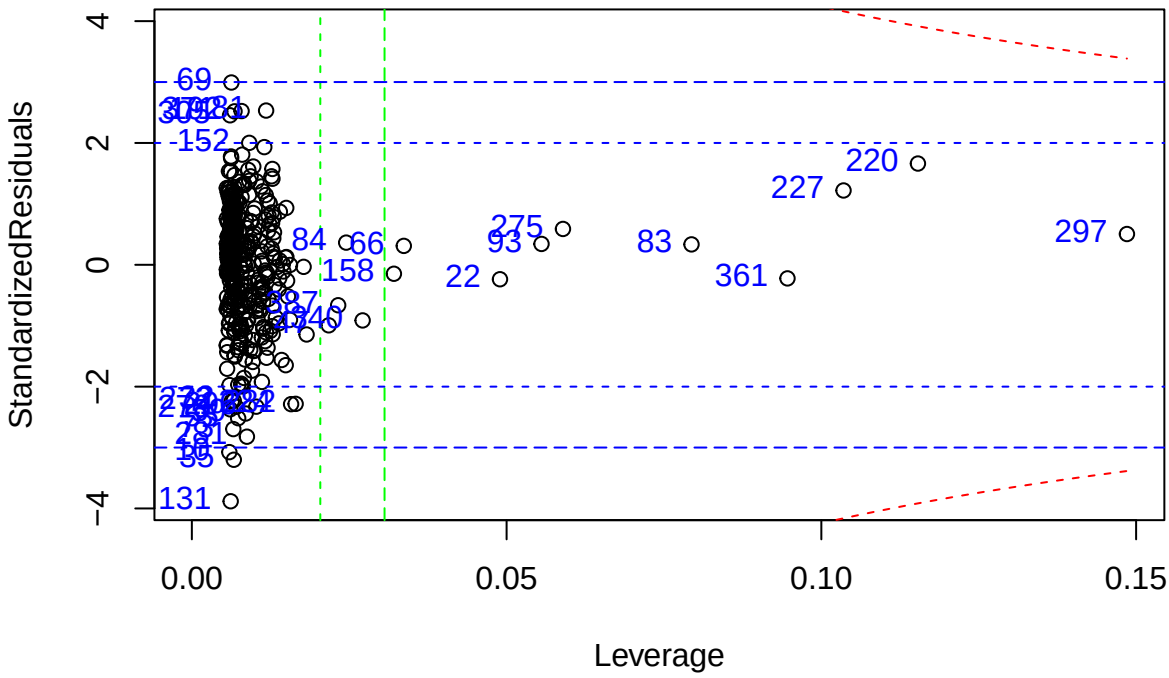
Assess

Figures:

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


Residuals vs Fitted
Residual
5
0
−5
−10
10
20
Fitted Value
35
10
131


Normal Q−Q
Standardized Residuals
Theoretical Quantiles
2
0
−2
−4
−2
0
2
10
35
131



Comments To address the non-linearity and non-equal variance given in the first model, I included a second-degree polynomial term for mileage to better capture the relationship between mileage and price. While there are more points to the end of the graph than at the front, this is a large improvement to the conditions I was checking. Normality is met, as evidenced by the QQ plot,thought there is departure from the line at the very front. After assessing the Cook's plot for this model, there does not seem to be any influential points that need to be removed. All points with high residuals have very low leverage, and points with high leverage have residuals near 0.

Use

Doing Denver, I used the "predict" function with mileage = 40, scaled down to fit the numbers in the model. The predicted price of the car is $20,377.12. A confidence interval for this is (15,448.82, 25,205.41).

```
##        fit      lwr      upr
## 1 20.37712 15.44882 25.30541
```

Conclusion

In this analysis of used car prices across different locations, I found that both mileage and age significantly influence a car's price, with mileage being the most substantial contributor. The models indicated that as mileage increases, the price of a vehicle tends to decrease. The results also suggested some location-specific effects, where in Louisville, prices were slightly higher compared to Denver, though the significance varied depending on the inclusion of certain variables. Overall, the models demonstrated a strong fit, particularly with the inclusion of polynomial terms for mileage, which helped capture the non-linear relationship between

mileage and price. The adjusted r-squared for this model was 0.798, which has a good amount of explanation, but could be improved upon if age could be used. This does make sense in the real world, and could be used for predicting the costs for used Altimas, though improvements could be made to the model by including different variables or using different locations. While the models are effective in predicting prices based on available data, there are other factors that were not recorded in this data that definitely have an effect, such as vehicle condition or market trends.