MATH 324: Multiple Linear Regression project

Shaurya Saxena

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 analysis, I chose to look at the Toyota Highlander, a midsize SUV known for its reliability and spaciousness. I'm comparing data from Aurora, Colorado (80016) and Los Angeles, California (90210).

I selected Aurora because there's a high number of Toyota Highlanders on the road here, likely due to the vehicle's reputation for handling Colorado's challenging weather. In contrast, Los Angeles offers a different perspective. The warm climate and urban environment may influence drivers to prioritize fuel efficiency and style, yet many families still value the Highlander's reliability and roominess.

I expect to find differences in pricing, mileage, and vehicle age between these two locations. In Aurora, the demand for reliable SUVs might mean higher prices and lower mileage due to better maintenance. Meanwhile, in Los Angeles, factors like urban driving and a faster turnover of vehicles could lead to different trends. This comparison will help illustrate how location affects the valuation of a popular vehicle like the Highlander.

Research question 1

Assuming a linear relationship between price and mileage, is there a difference in price between the locations?
To explore the relationship between price and mileage, I created summary statistics for both locations and visualized the data using scatterplots.

Exploratory data analysis

Figure(s):

EDA TABLE 1

	sample size	mean price	sd of price	mean mileage	sd of mileage
LA 90210	300	32.114	9.010	55.169	37.014
Aurora 80016	212	27.974	10.647	82.400	55.256

Comments: The scatterplot shows the relationship between price and mileage for Toyota Highlanders in both California and Colorado. The summary statistics reveal that California has a higher mean price than Colorado, despite Colorado having a higher mean mileage. This suggests that vehicles in California may be priced higher due to factors such as demand, market conditions, or differences in vehicle condition.

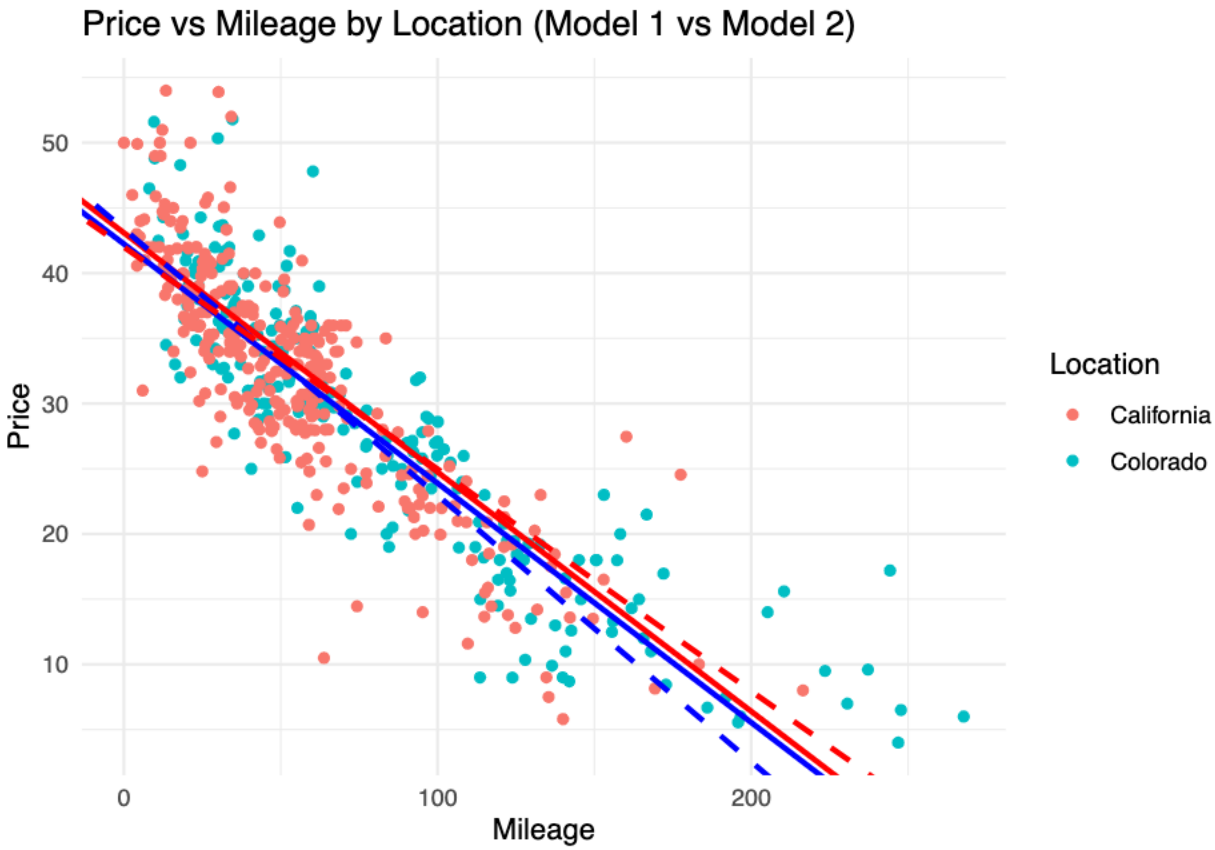
Model fitting

MODEL SUMMARY TABLE 1: (same slope different intercepts)

	estimate	test-statistic	p-value
intercept	42.236	107.130	<2e-16
mileage	-0.183	-37.683	<2e-16
location(CO)	0.856	1.829	0.0679

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

	estimate	test-statistic	p-value
intercept	43.380	84.537	<2e-16
mileage	-0.204	-26.425	<2e-16
location(CO)	-1.375	-1.724	0.085
mileage:Colorado	0.034	3.433	0.0006

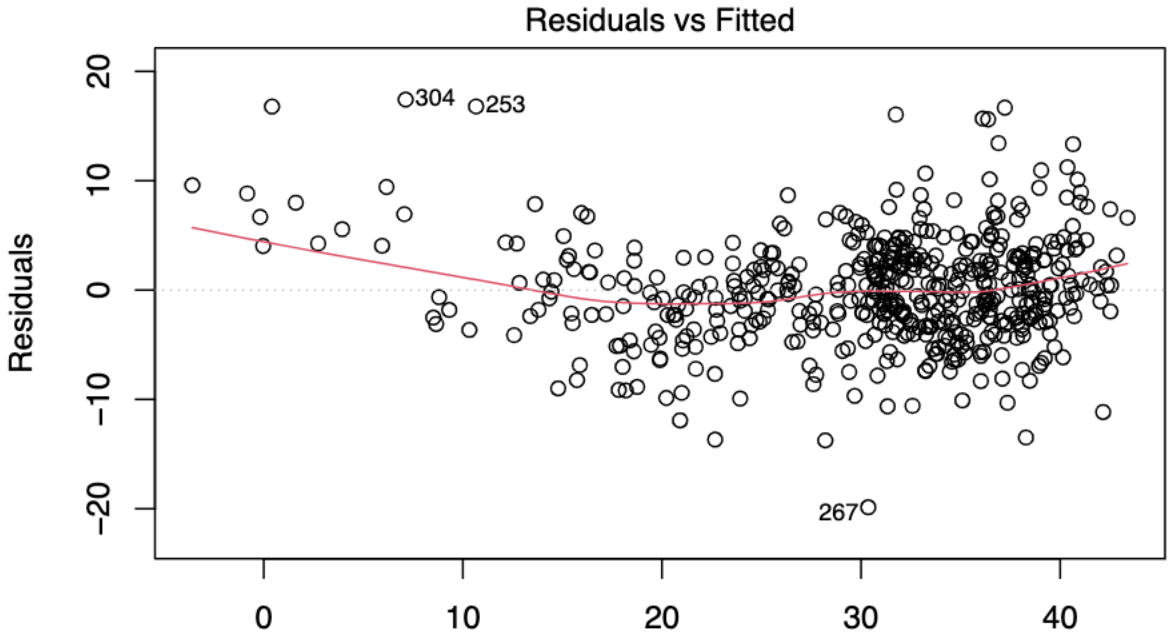


Fitted model for (California):
$Price = 43.380 - 0.204 \times mileage$

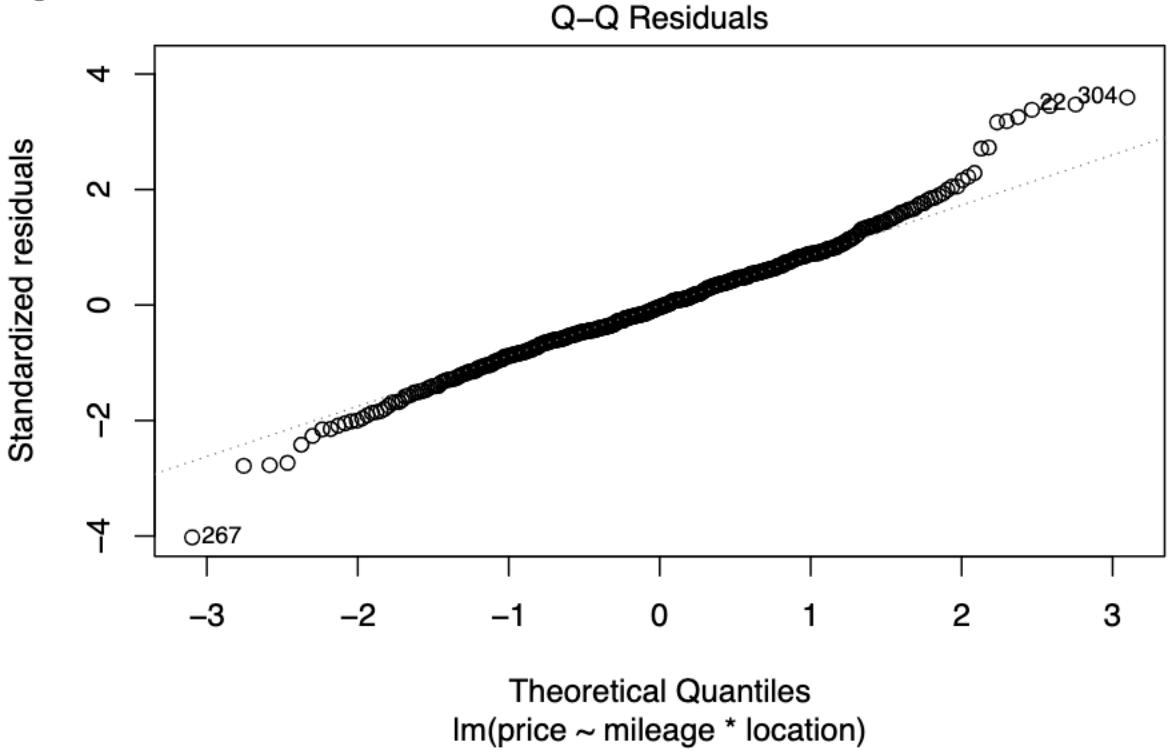
Fitted model for (Colorado):
$Price = 42.005 - 0.170 \times mileage$

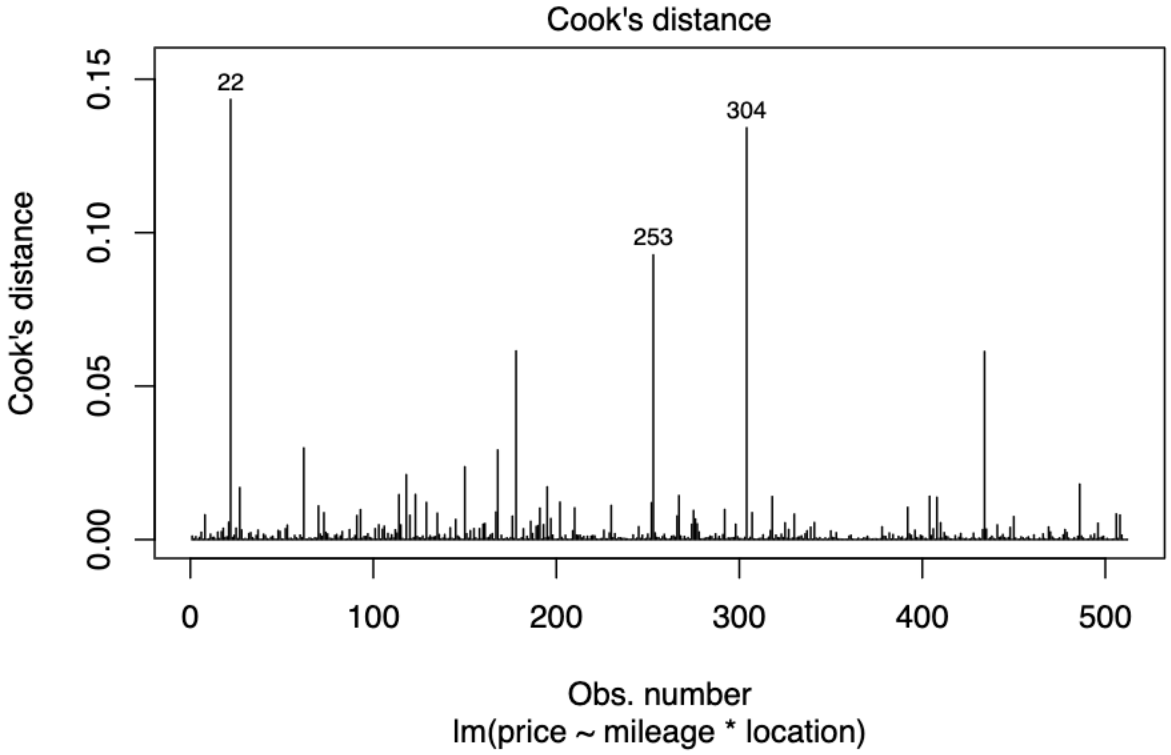
Comments
These results, as indicated on the graph above, show that while both locations have negative relationships between price and mileage, California has a slightly larger slope, suggesting that the mileage impacts price more significantly in California than in Colorado. Additionally, the intercept for California is higher - indicating that Toyota Highlanders in Los Angeles, California, on average - start at a higher price than their counterparts in Aurora, Colorado.

Assess



Figures:





Comments
The residuals vs. fitted plots shows a random scatter, which indicates that the linearity assumption is reasonable. The QQ plot indicates that the residuals follow a normal distribution, and the Cook's distance plot does not reveal any influential points that could negatively affect the regression results. Overall, the conditions for linear regression appear to be satisfied.

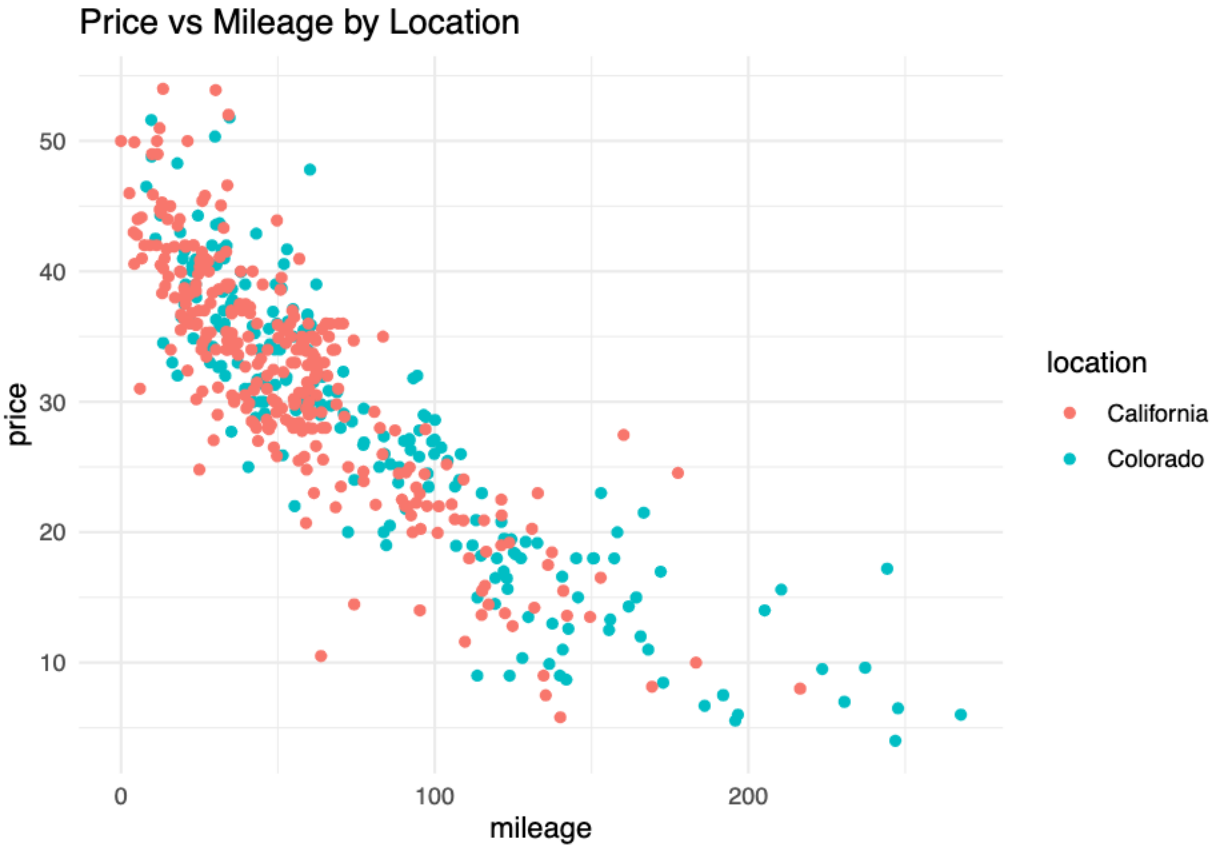
Use
After assessing the models, we can conclude that there is a significant difference in Toyota Highlander prices between California and Colorado after accounting for the mileage. Highlanders in California tend to be priced higher, which can be correlated to the market demand, conditions, and higher cost of living in more urban areas.

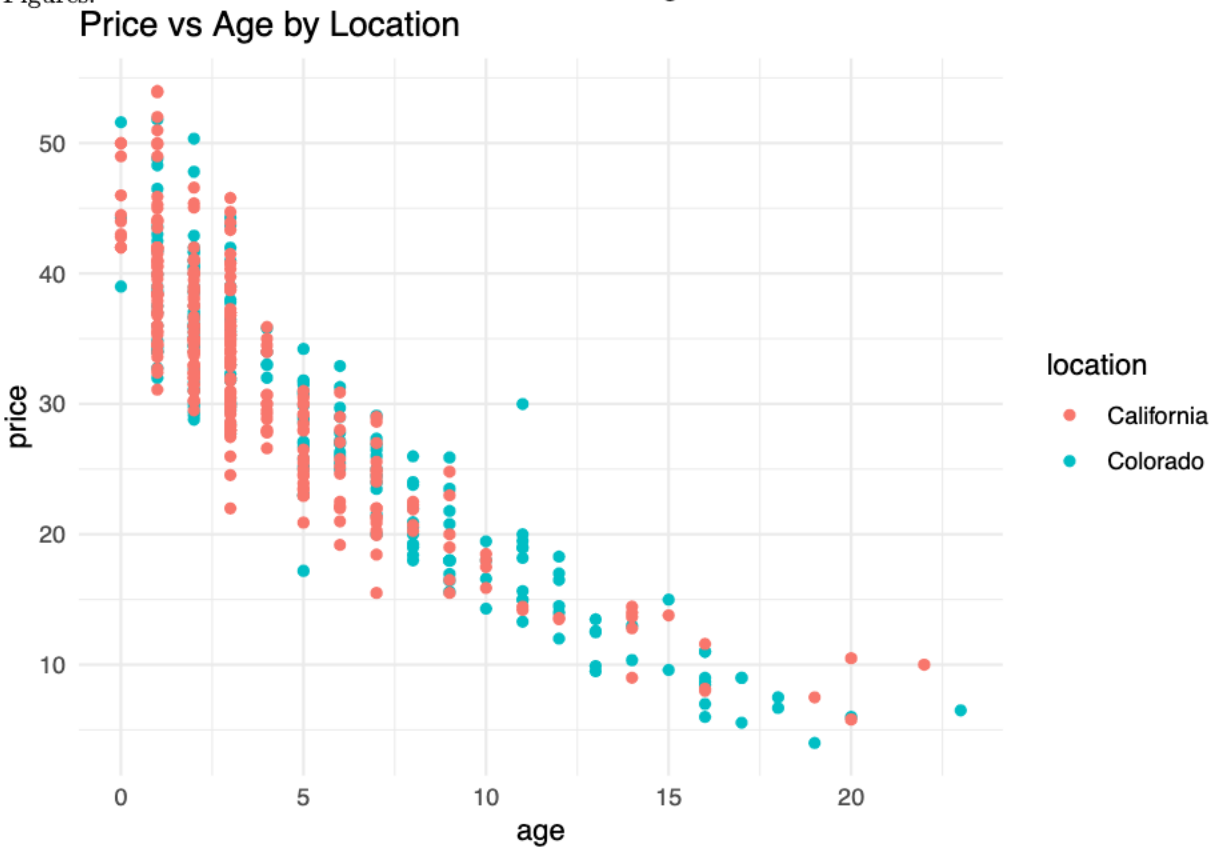
Research Question 2

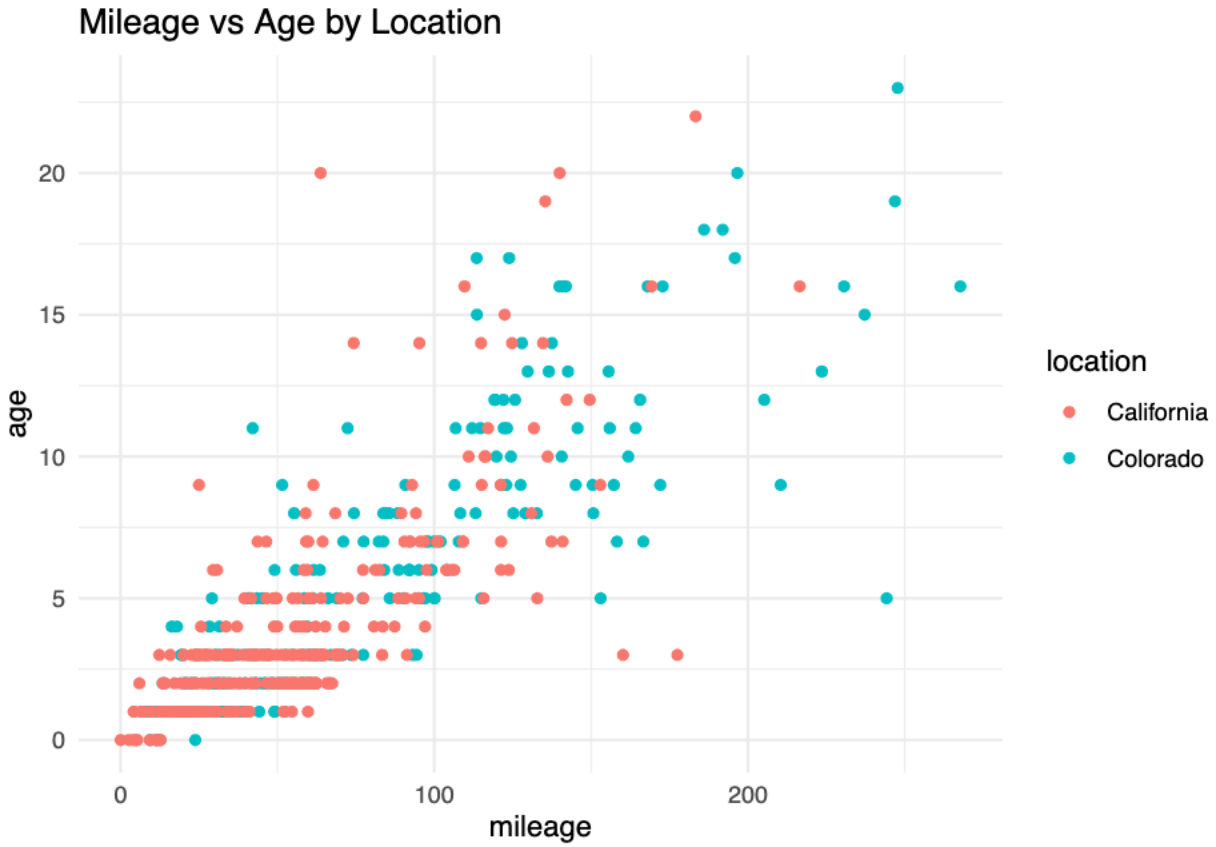
After accounting for a linear relationship between age and price and between mileage and price, is there a difference in price between the locations?

Exploratory data analysis

Figures:







Comments

The scatterplots show varying relationships between the price, age, and mileage for both locations. Notably, older vehicles tend to be priced lower across both locations, indicating depreciation by age. The relationship between mileage and age also shows a positive correlation, suggesting that cars with higher mileage are generally older.

Model fitting

To investigate if location has an effect on car prices after controling for mileage and age, a linear model is fitted with mileage, age, and location as predictors.

SUMMARY TABLE 3:

	estimate	test-statistic	p-value
intercept	42.085	0.307	<2e-16
mileage	-0.091	0.006	<2e-16
age	-1.263	0.069	<2e-16
location	1.022	0.364	0.0052

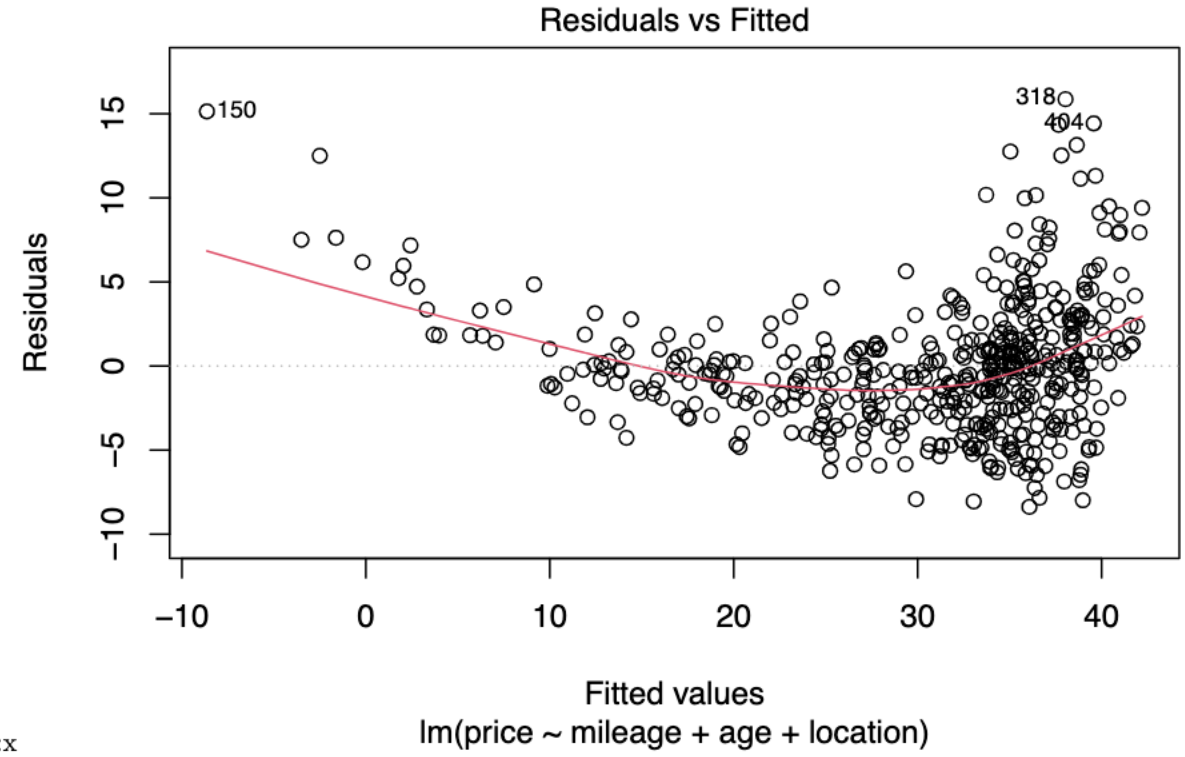
Interpretations in context

- intercept (42.06): The estimated base price of a car in California with 0 mileage and 0 years of age is $\$42,058$.
- mileage (-0.091): For each additional mile, the car price decreases by approximately \$0.091, holding age and location constnat. This negative effect is highly statistically significant because $p < 2e - 16$.
- age (-1.263): For each additional year of age, the car price deceases by approximately $\$1,263$, which is also highly statistically significant because $p < 2e - 16$. This shows that the older the car, the more the price tends to drop.

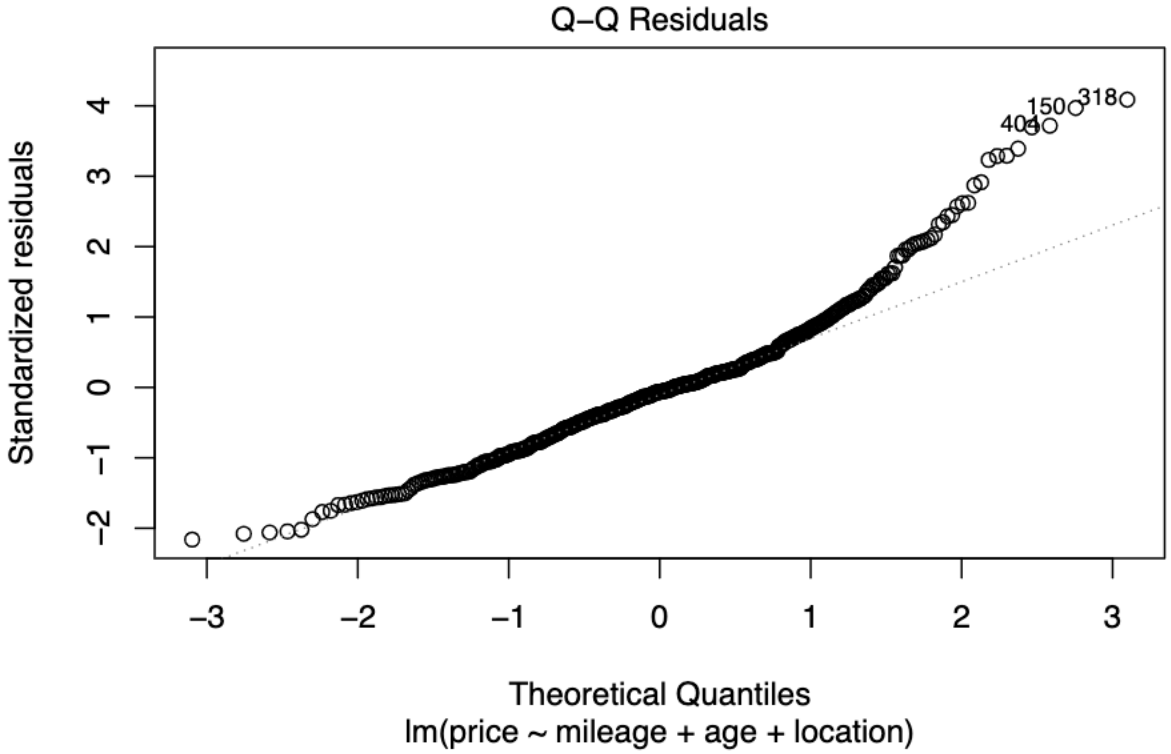
- location (CO = 1.022): Cars in Colorado are priced, on average, $\$1,022$ higher than cars in California., controlling for mileage and age. This result is statistically significant because $p = 0.0052$, indicating a difference in pricing between the two locations.

Assess

- L: To assess the linearity using the residuals vs. fitted plot, the linearity condition looks to be met as the points look randomly dispersed around the horizontal axis, additionally there is no noticeable pattern.
- N: To assess the normality using the Q-Q Residuals plot, the normality condition is met as most of the points are along the line of the Q-Q plot.
- E: To assess the equal variance using the residuals vs. fitted plot, the equal variance condition does not look to be met because residuals don't look equally distributed, and there is some flaring.



Figures:x



Comments
The exploratory data analysis indicated that older vehicles tend to be priced lower due to depreciation, while a positive correlation between mileage and age suggests that higher mileage typically means older cars. Model 3 backed up these findings, showing that Toyota Highlanders in Colorado are, on average, priced $1,022 higher than those in California, even after accounting for mileage and age, highlighting the significant role of location in car pricing. However, the residual analysis showed some issues with equal variance, indicating areas for model improvement. Overall, these results underscore the importance of vehicle characteristics and geographical factors for buyers and sellers in the market.

Use
Based on the results from our fitted model, there is a statistically significant difference in car prices between the two locations after accounting for both mileage and age. Specifically, cars in Colorado are priced, on average, $1,022 higher than cars in California, holding mileage and age constant. This result, which is supported by a p-value of 0.0052, indicates that the observed difference in prices is unlikely to be due to random chance. In context, this suggests that geographical location plays an important role in determining price beyond just the car's attributes. The higher average prices in Colorado could be driven by various factors, such as market demand, economic conditions, or differences in the types of cars sold between locations. Thus, buyers and sellers should consider location when setting or negotiating car prices, as regional factors significantly influence pricing.

Research question 3

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

Choose

In finding the best model, I considered serveral variations of the linear regression model. And after testing several models, the final model that best predicts price includes both mileage and age as predictors, with

interaction terms for location and mileage as well as location and age. I chose this model as my final model because it had the highest R^2 value of 0.853 out of all the models I tried, indicating it explains a good proportion of the variance in car prices.

Final fitted model:
$Price = 43.20 - 0.11 \times mileage - 1.20 \times Location_{CO} - 1.25 \times Age + 0.037 \times Mileage \times Location_{CO} - 0.038 \times Age \times Location_{CO}$

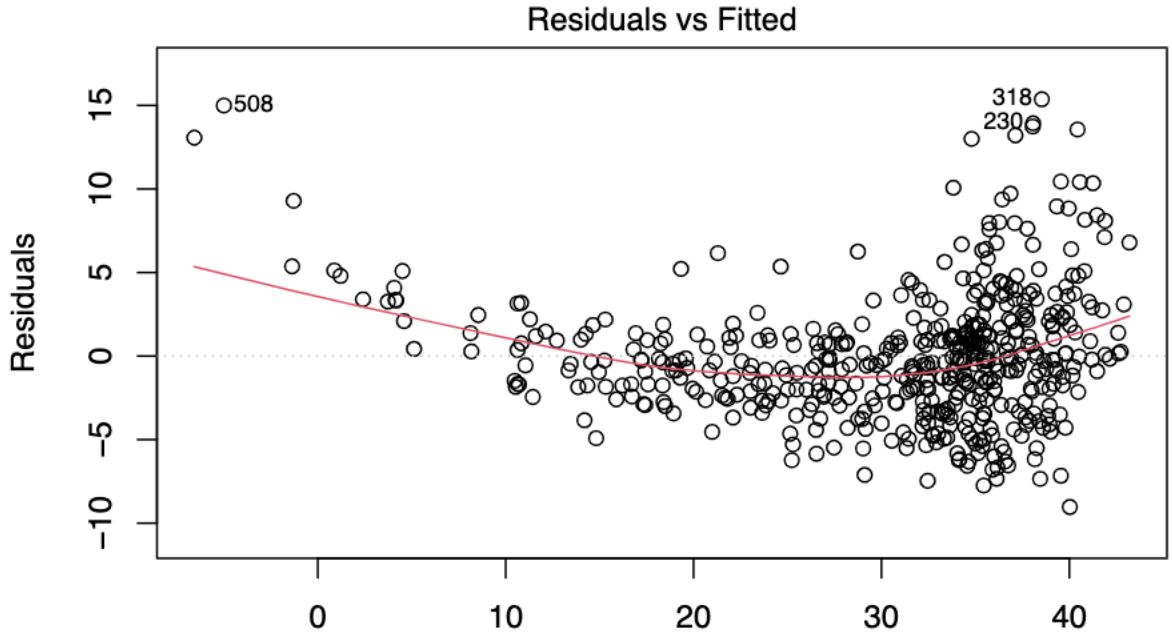
Summary of Final Model:

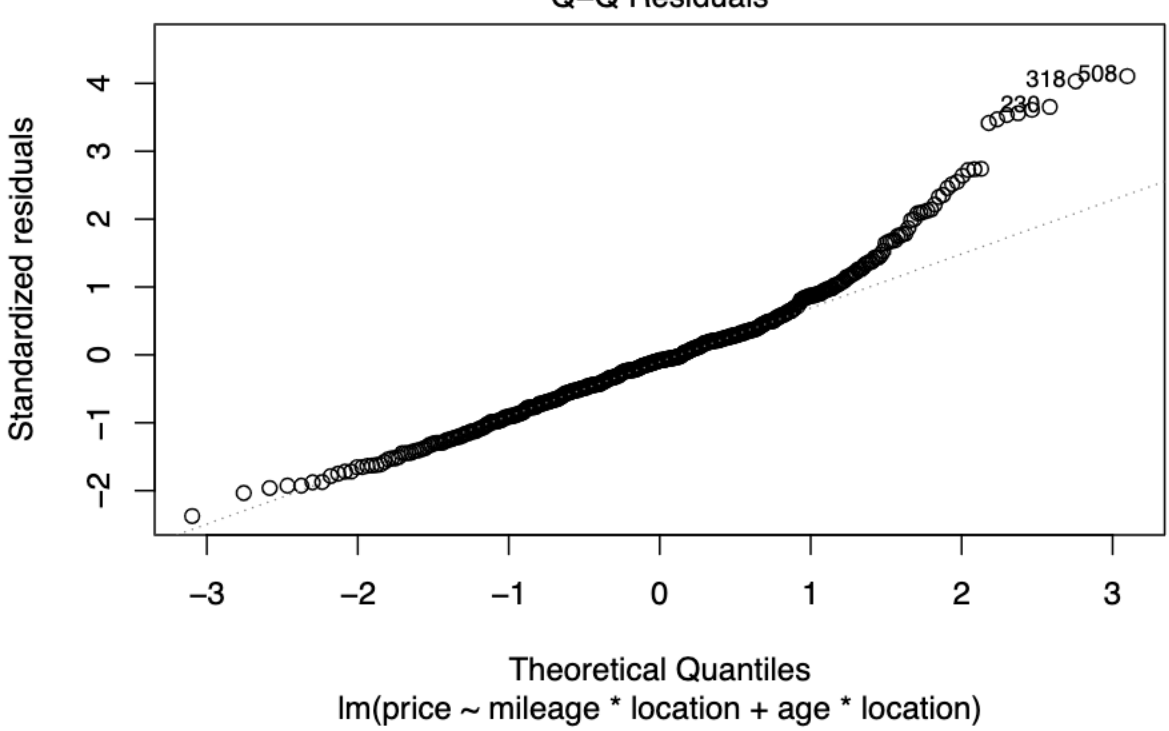
```
##
## Call:
## lm(formula = price ~ mileage * location + age * location, data = car_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0304 -2.4350 -0.3068  1.6451 15.3659
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               43.199962   0.396907 108.842  < 2e-16 ***
## mileage                   -0.113402   0.008983 -12.624  < 2e-16 ***
## locationColorado          -1.201804   0.616902  -1.948  0.05195 .
## age                       -1.245872   0.092038 -13.537  < 2e-16 ***
## mileage:locationColorado   0.036540   0.012556   2.910  0.00377 **
## locationColorado:age      -0.037985   0.136814  -0.278  0.78141
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.824 on 506 degrees of freedom
## Multiple R-squared:  0.853,  Adjusted R-squared:  0.8515
## F-statistic: 587.2 on 5 and 506 DF,  p-value: < 2.2e-16
```

Comments
This models suggests that mileage and age both have both significant effects on price, with a small but noticeable difference in how these variables affect prices in California versus Colorado.

Assess

Figures:





The residuals versus fitted plot shows that the model meets the linearity assumption, as the points are randomly dispersed around the horizontal line. The Q-Q plot indicates that the residuals are mostly normally distributed, though some outliers are present. Overall, with a high R-squared value of 0.853, the model explains a significant portion of the variance in car prices, indicating good predictive performance.

Use Using this model, to predict the price for a 3-year-old Toyota Highlander with 40,000 miles in California would be:

$$Price = 43.20(1,000) - 0.11 \times 40,000 - 1.25(1,000) \times 3 = \$35,050$$

95% prediction interval
The 95% prediction interval is [$27,399, $42,453].

We are 95% confident that the price of an individual Toyota Highlander in California that is 3 years old and has 40,000 miles will fall between $27,399 and $42,453.

Conclusion

Through my analysis of car prices for Toyota Highlanders in California and Colorado, I learned that both mileage and age are significant factors influencing vehicle pricing, with mileage being the most substantial contributor to price depreciation. The results indicated that as mileage increases, the price of the vehicle decreases, proving the common notion that higher usage typically correlates with more wear and tear. Additionally, the model revealed that cars in Colorado had an average price that was slightly higher than in California when controlling for mileage and age. This suggests that regional demand and market conditions can impact pricing significantly. Overall, I believe my model is effective, as it explains a substantial proportion of the variance in car prices, which can be pointed to by the high R-squared value. However, it could be improved by incorporating additional variables such as vehicle condition, accident history, service records, and market trends. Doing so can provide a more comprehensive view of the factors influencing price. Furthermore, considering geographical differences in the availability of certain models, economic factors, and the local cost of living could make the model's predictive power more accurate. An interesting observation was the potential influence of climate on vehicle maintenance; for instance, cars in Colorado may benefit from less harsh conditions compared to those in coastal California, potentially leading to different depreciation rates.