Car Prices for Dummys

Dustin Leatherman, Kyra Kemp, Danielle Deng, Zhichun Ke

11/16/2019

Introduction

@Kyra to fill this in

The Data

This dataset is a representative sample of over eight hundred 2005 GM cars from the Kelly Blue Book.

Due to the right-skewed nature of Price, a log transformation has been applied and displayed alongside for comparison.
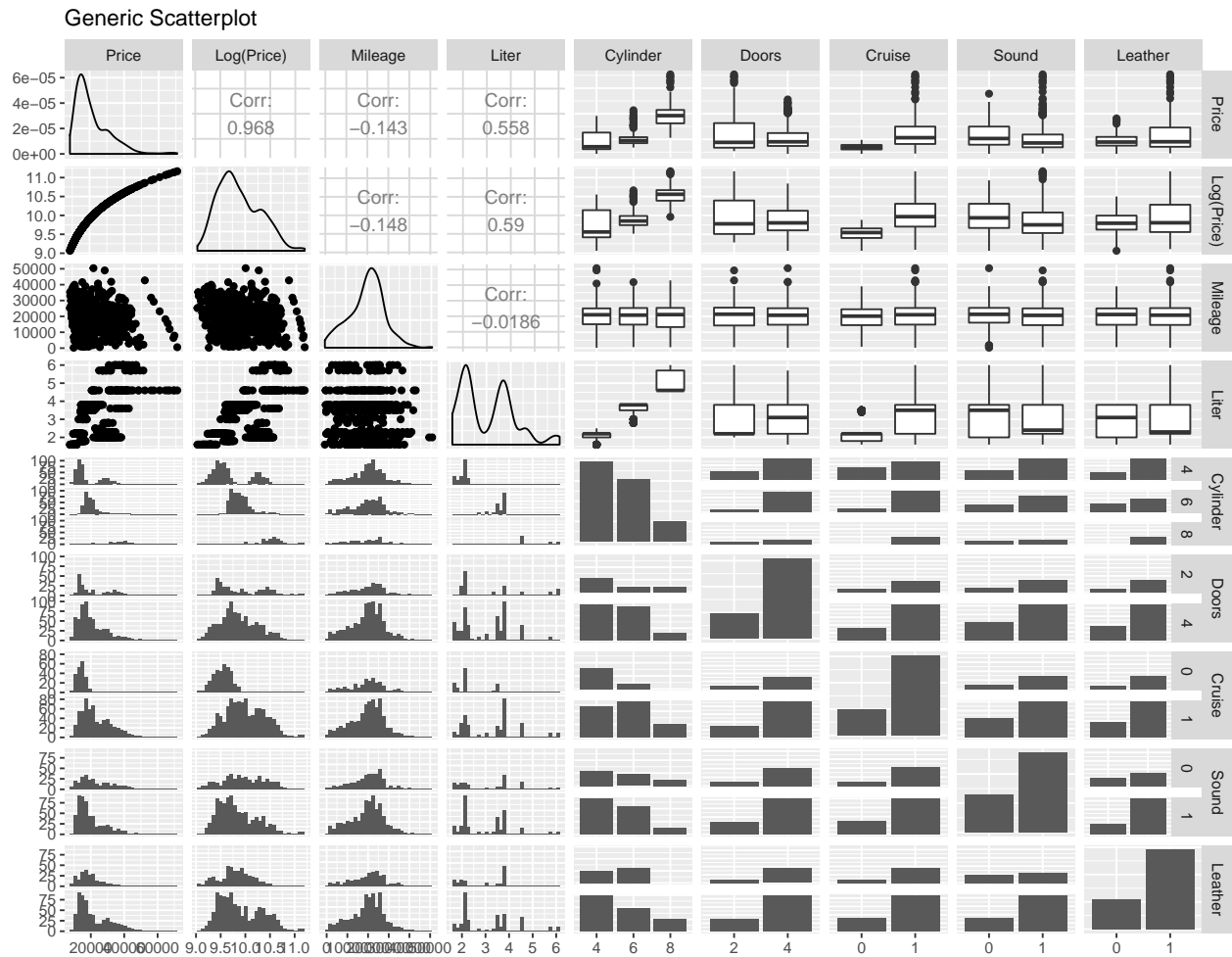
@Zhichun @Danielle to fill this in

Figure 1. A general sense of the relationship between variables is provided.

Summary (cont.)

Figure 2. Relationships between variables are explored in regards to Type.

Summary (cont.)

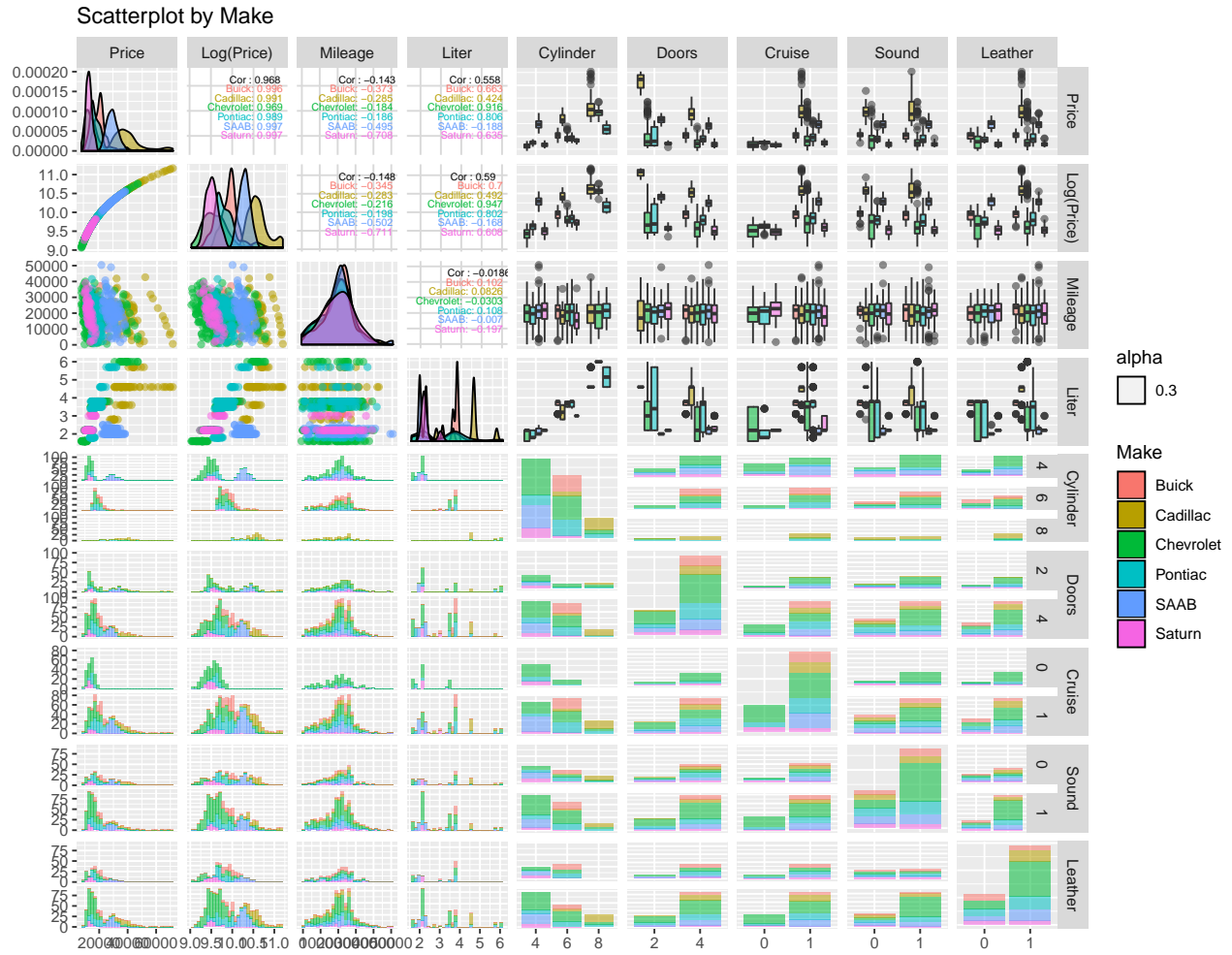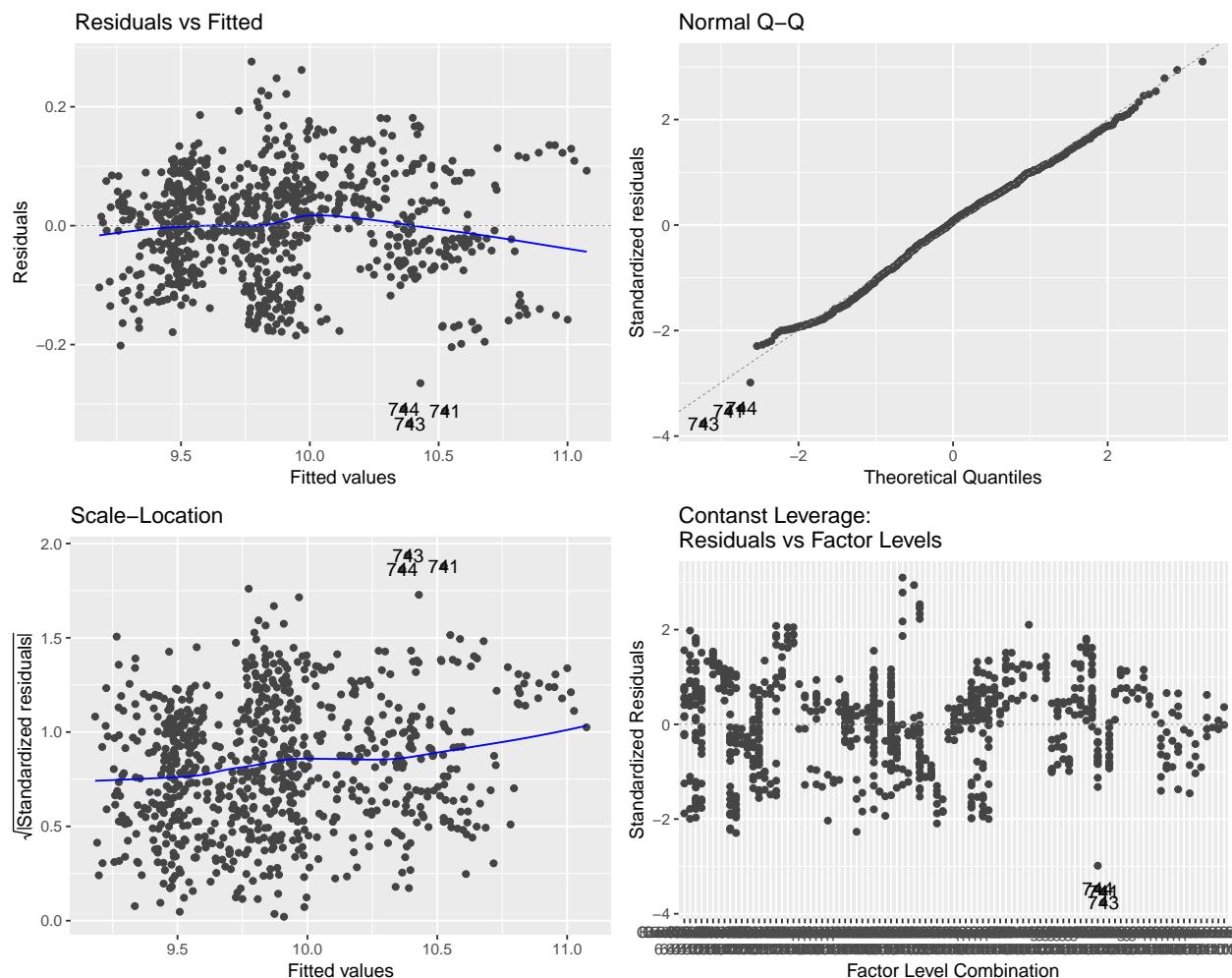Figure 3. Relationships between variables are explored in regards to Make.

Analysis

Prior to fitting any model, a full model will be fit first in order to ensure that model assumptions are correct.

Normality

The residuals are close to the line barring a few outliers on the left tail. From a visual perspective, data appear normal but a Shapiro-Wilk test indicates there is convincing evidence that the data is non-normal (p-value = 0.002606). There are three outliers identified on the left tail that may be influencing this outcome. This must be explored.

statistic	p.value	method
0.9939636	0.0026057	Shapiro-Wilk normality test

Linearity

There do not appear to be any patterns in the residuals indicating that a linear regression model is appropriate for the data.

Constant Variance

The spread around y = 0 on the Residuals vs Fitted plot appears to be constant barring the same potential outliers present in the normal probability plot; however, the Brown-Forsyth test indicates that there is convincing evidence that the variance is not constant (p-value = 0.00016).

| statistic	p.value	method
14.33485	0.0001644	modified robust Brown-Forsythe Levene-type test based on the absolute deviations from the media

x
0.9578966

```
## Model :
## log.price ~ Mileage + Make + Type + Liter + cylinder.factor +
##     doors.factor + cruise.factor + sound.factor + leather.factor
##
## Complete :
##               (Intercept) Mileage MakeCadillac MakeChevrolet MakePontiac
## doors.factor4 0           0       0            0             0
##               MakeSAAB MakeSaturn TypeCoupe TypeHatchback TypeSedan
## doors.factor4 0        0          0         1             1
##               TypeWagon Liter cylinder.factor6 cylinder.factor8
## doors.factor4 1         0     0                0
##               cruise.factor1 sound.factor1 leather.factor1
## doors.factor4 0              0             0
```

There is multicollinearity present due to the high correlation between Liter and Cylinder. (see *Figure 1*). This further confirmed given a correlation coefficient of 0.958. Additionally, the **Doors** predictor is linearly dependent on the Indicator variables TypeWagon, TypeHatchback, and TypeSedan. This means that this coefficient cannot be estimated and thus should be removed from the model.

.rownames	GVIF	Df	GVIF..1..2.Df..
Mileage	1.008172	1	1.004078
Make	8.933621	5	1.244809
Type	2.937632	4	1.144194
Liter	19.558325	1	4.422479
cylinder.factor	39.557783	2	2.507887
cruise.factor	1.587586	1	1.259994
sound.factor	1.151701	1	1.073173
leather.factor	1.247231	1	1.116795

.rownames	GVIF	Df	GVIF..1..2.Df..
Mileage	1.007870	1	1.003927
Make	4.091844	5	1.151309
Type	2.246857	4	1.106488
Liter	1.876033	1	1.369684
cruise.factor	1.552723	1	1.246083
sound.factor	1.136262	1	1.065956
leather.factor	1.196471	1	1.093833

Checking for Variance Inflation Factors for the model full model sans **Doors** yields extreme multicollinearity for cylinder and Liter. Given that Liter is more precise than Cylinder in terms of measurement and Cylinder has a higher correlation with log.price, Liter will be dropped from the model. This produces a drastic decrease in mulitcollinearity to a more acceptable level
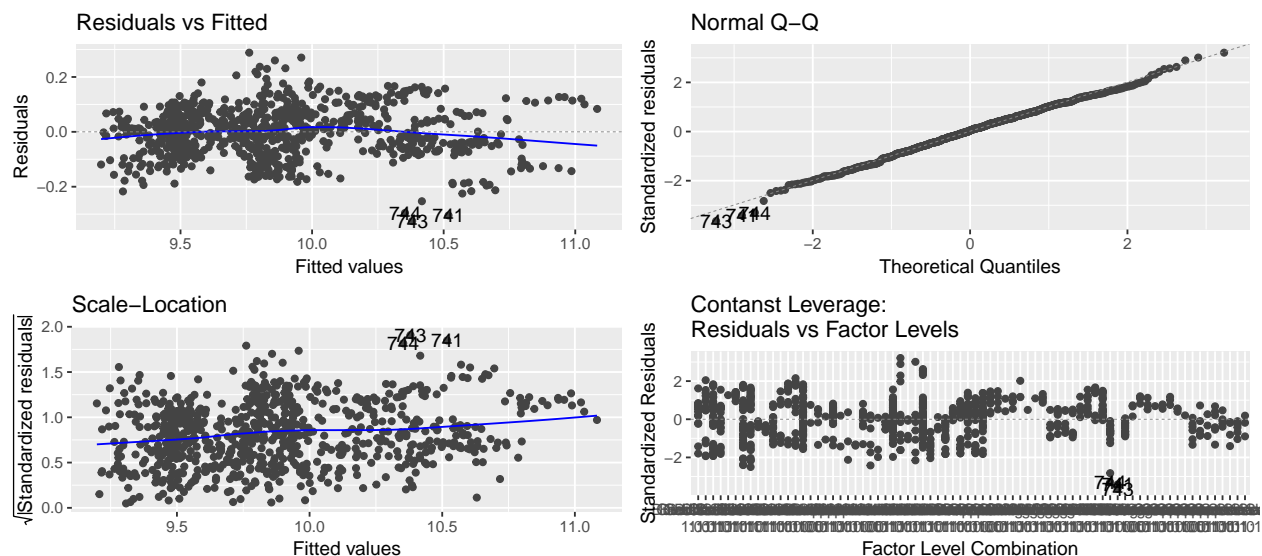
Question 1

How does Type, Cylinder, and Liter affect average Price? How does this change when Mileage is taken into account?

Appendices

Variable	Description
Price	suggested retail price of the used 2005 GM car in excellent condition. The condition of a car can greatly affect price. All cars in this data set were less than one year old when priced and considered to be in excellent condition.
Mileage	number of miles that the car has been driven.
Make	manufacturer of the car such as Saturn, Pontiac, and Chevrolet.
Model	Specific models for each car manufacturer such as Ion, Vibe, Cavalier.
Trim	Specific type of car model suc has SE Sedan 4D, Quad Coupe 2D.
Type	Body type such as sedan, coupe, etc.
Cylinder	Number of cylinders in the engine.
Liter	A more specific meaasure of engine size.
Doors	Number of doors
Cruise	Indicator representing whether the car has cruise control (1 = cruise).
Sound	Indicator representing whether the car has upgraded speakers (1 = sound).
Leather	Indicator representing whether the car has leather seats (1 = leather).