Car Price Prediction Multiple Linear Regression, STAT 510-01

Presenters: Gerry Cruz, Yushan Zhao, Kierra Manuel

The presentation should include the research goal of your study, a summary of the data set, how you build models, summary of main results, challenges, and possible future work.

Today's Agenda

- Introduction and Goal
- Summary of the Data Set
- EDA
- Variable and Model selection
- Regression Analysis
- Testing Model with a Prediction
- Challenges and Future Work

YZ

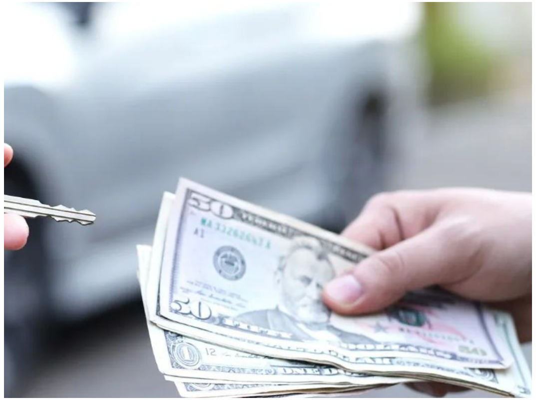
Introduction

A Chinese automobile company Geely Auto aspires to enter the US market by setting up their manufacturing unit there and producing cars locally to give competition to their US and European counterparts. They have contracted an automobile consulting company to understand the factors on which the pricing of cars depends. Specifically, they want to understand the factors affecting the pricing of cars in the American market, since those may be very different from the Chinese market. Based on various market surveys, the consulting firm has gathered a large data set of different types of cars across the America market.

Goal

We are required to model the price of cars with the available independent variables. It will be used by the management to understand how exactly the prices vary with the independent variables. They can accordingly manipulate the design of the cars, the business strategy etc. to meet certain price levels. Further, the model will be a good way for management to understand the pricing dynamics of a new market.

Introduction/Background



吉利汽车
GEELY AUTO

Car Price

Geely Auto

Car Market

YZ

Goal of our project

- Model the price of cars with the available independent variables.
- Use the model to understand how exactly the prices vary with the independent variables.
- Manipulate the design of the cars, the business strategy etc. to meet certain price levels.
- Use the model to understand the pricing dynamics of a new market.

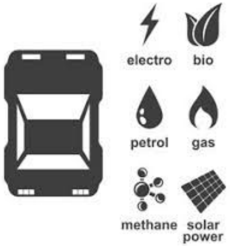
YZ

Summary of the Data Set

- 31 rows and 24 columns
- Dependent variable: car price
- Independent variables:
  - Categorical variables: carname, fueltype, drivewheel, enginelocation, enginetype, cylindernumber, enginesize, stroke, doornumber, carbody, fuelsystem, aspiration,
  - Continuous variables: wheelbase, carlength, carwidth, carheight, curbweight, boreratio, compressionratio, horsepower, peakrpm, citympg, highwaympg,
- Link for the data set: https://www.kaggle.com/datasets/hellbuoy/car-price-prediction

YZ

Car name

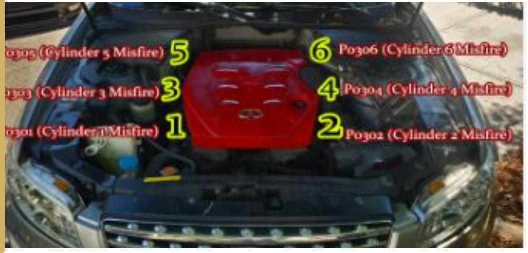


fuel type

drive wheel

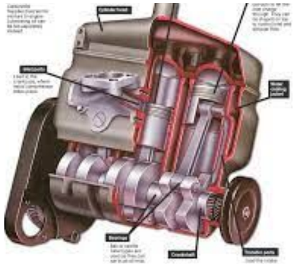
engine location

engine type



cylinder number

Engine size

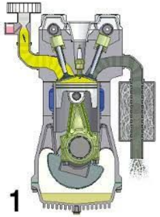


stroke

door number

car body

fuel system



Wheelbase

car length, car width, car height

curb weight

horsepower

●●●

YZ

Dataset

Un-Cleaned Data

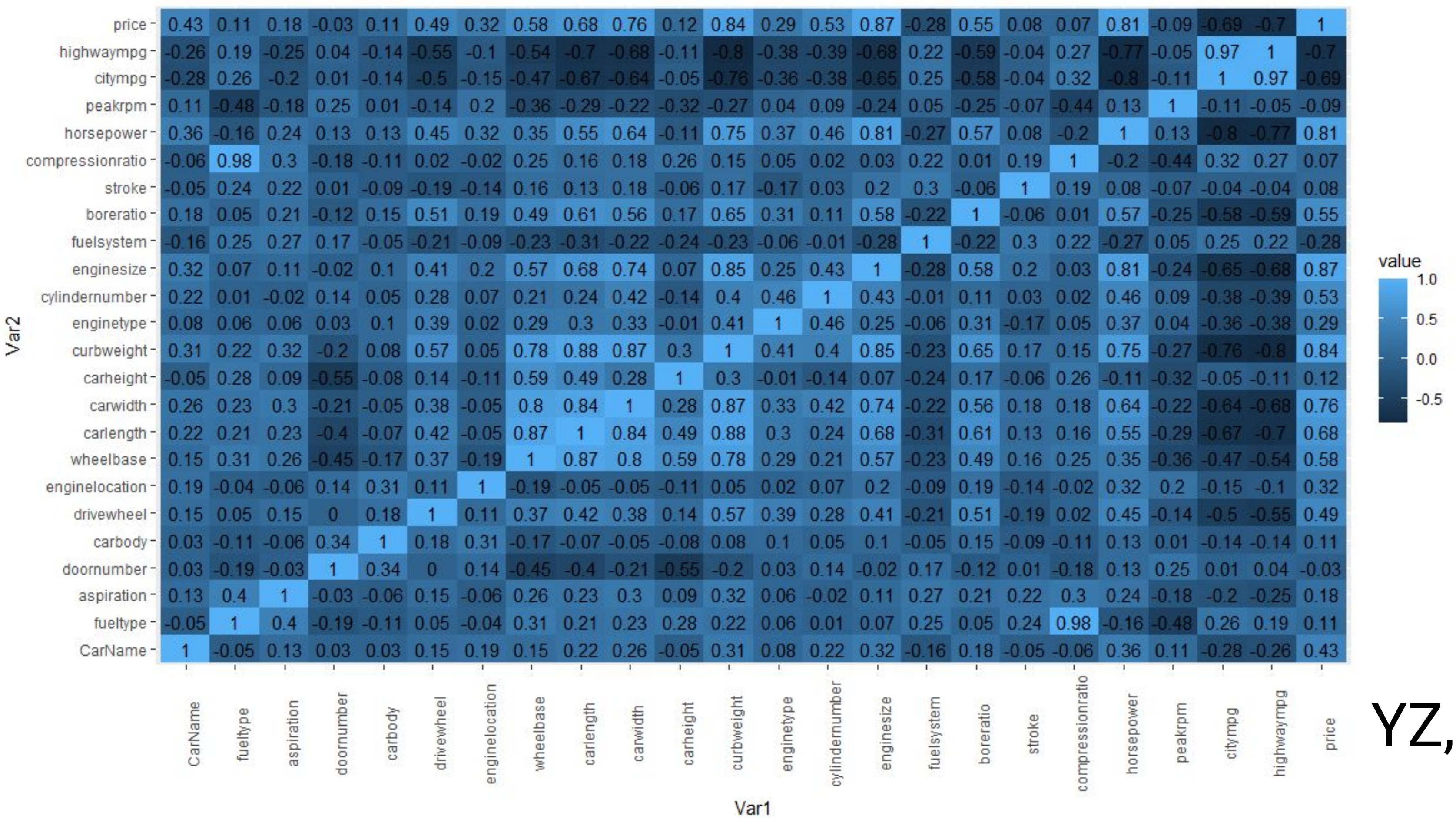
car_ID	symboling	CarName	fueltype	aspiration	doornumber	carbody
1	3	alfa-romero giulia	gas	std	two	convertible
2	3	alfa-romero stelvio	gas	std	two	convertible
3	1	alfa-romero Quadrifoglio	gas	std	two	hatchback
4	2	audi 100 ls	gas	std	four	sedan
5	2	audi 100ls	gas	std	four	sedan
6	2	audi fox	gas	std	two	sedan
7	1	audi 100ls	gas	std	four	sedan
8	1	audi 5000	gas	std	four	wagon
9	1	audi 4000	gas	turbo	four	sedan
10	0	audi 5000s (diesel)	gas	turbo	two	hatchback
11	2	bmw 320i	gas	std	two	sedan
12	0	bmw 320i	gas	std	four	sedan
13	0	bmw x1	gas	std	two	sedan
14	0	bmw x3	gas	std	four	sedan
15	1	bmw z4	gas	std	four	sedan
16	0	bmw x4	gas	std	four	sedan
17	0	bmw x5	gas	std	two	sedan
18	0	bmw x3	gas	std	four	sedan
19	2	chevrolet impala	gas	std	two	hatchback
20	1	chevrolet monte carlo	gas	std	two	hatchback
21	0	chevrolet vega 2300	gas	std	four	sedan

Cleaned Data

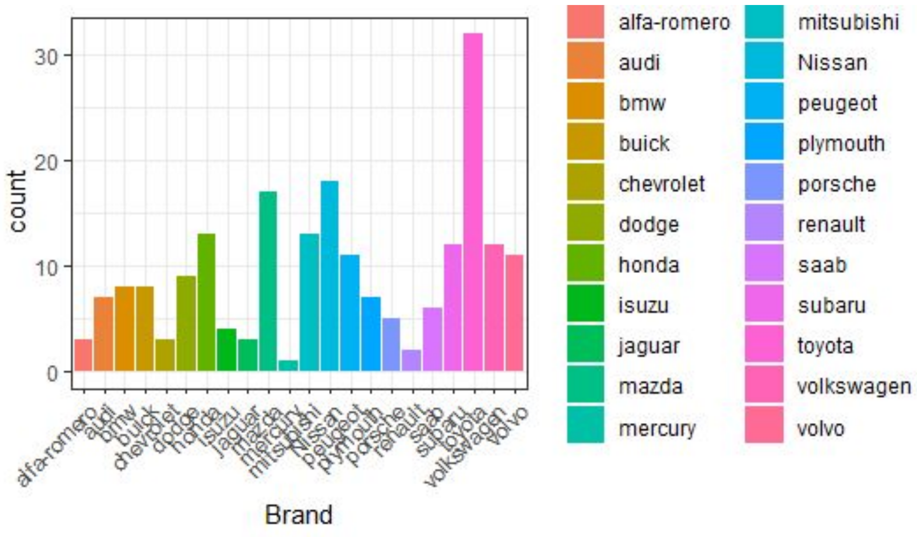
CarName	fueltype	aspiration	doornumber	carbody	drivewheel	enginelocation
18	0	0	1	4	1	0
18	0	0	1	4	1	0
18	0	0	1	1	1	0
12	0	0	0	0	0	0
12	0	0	0	0	2	0
12	0	0	1	0	0	0
12	0	0	0	0	0	0
12	0	0	0	2	0	0
12	0	1	0	0	0	0
12	0	1	1	1	2	0
11	0	0	1	0	1	0
11	0	0	0	0	1	0
11	0	0	1	0	1	0
11	0	0	0	0	1	0
11	0	0	0	0	1	0
11	0	0	0	0	1	0
11	0	0	1	0	1	0

GC

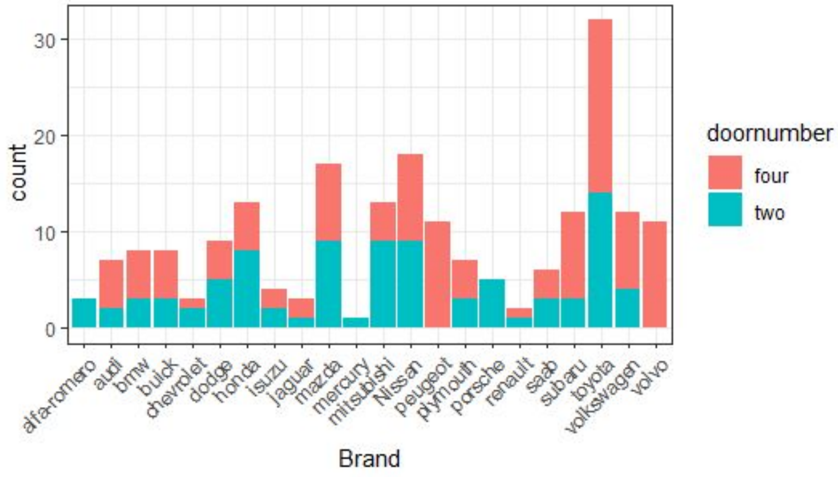
Correlation Heat Map



YZ, GC



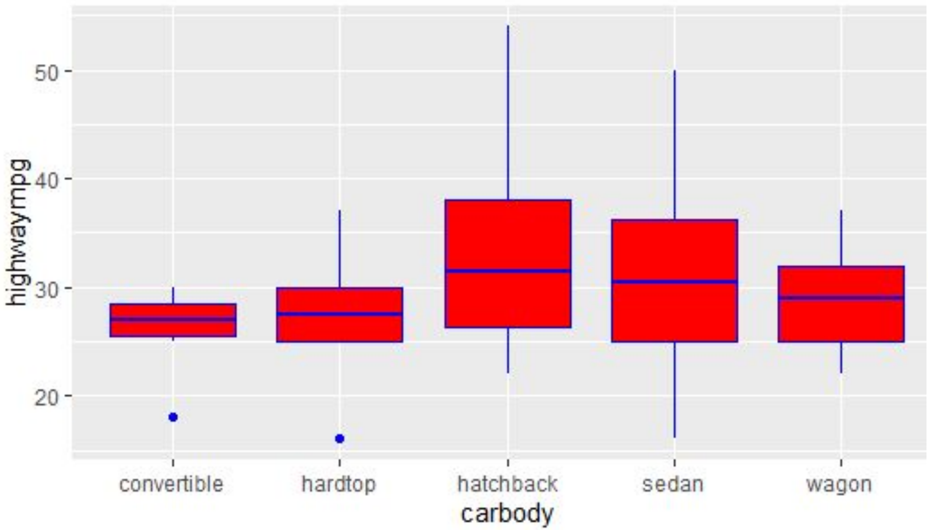
count
Brand
alfa-romero
audi
bmw
buick
chevrolet
dodge
honda
isuzu
jaguar
mazda
mercury
mitsubishi
Nissan
peugeot
plymouth
porsche
renault
saab
subaru
toyota
volkswagen
volvo

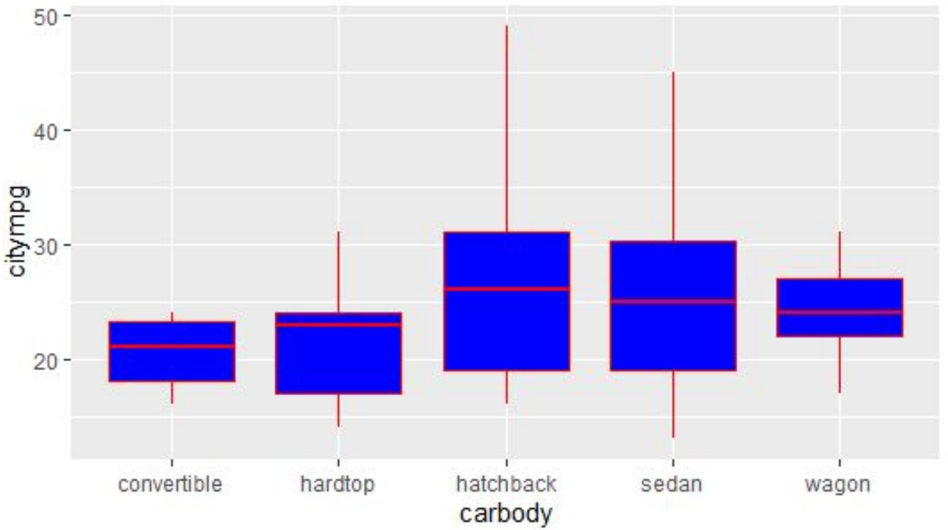


count
Brand
doornumber
four
two

GC

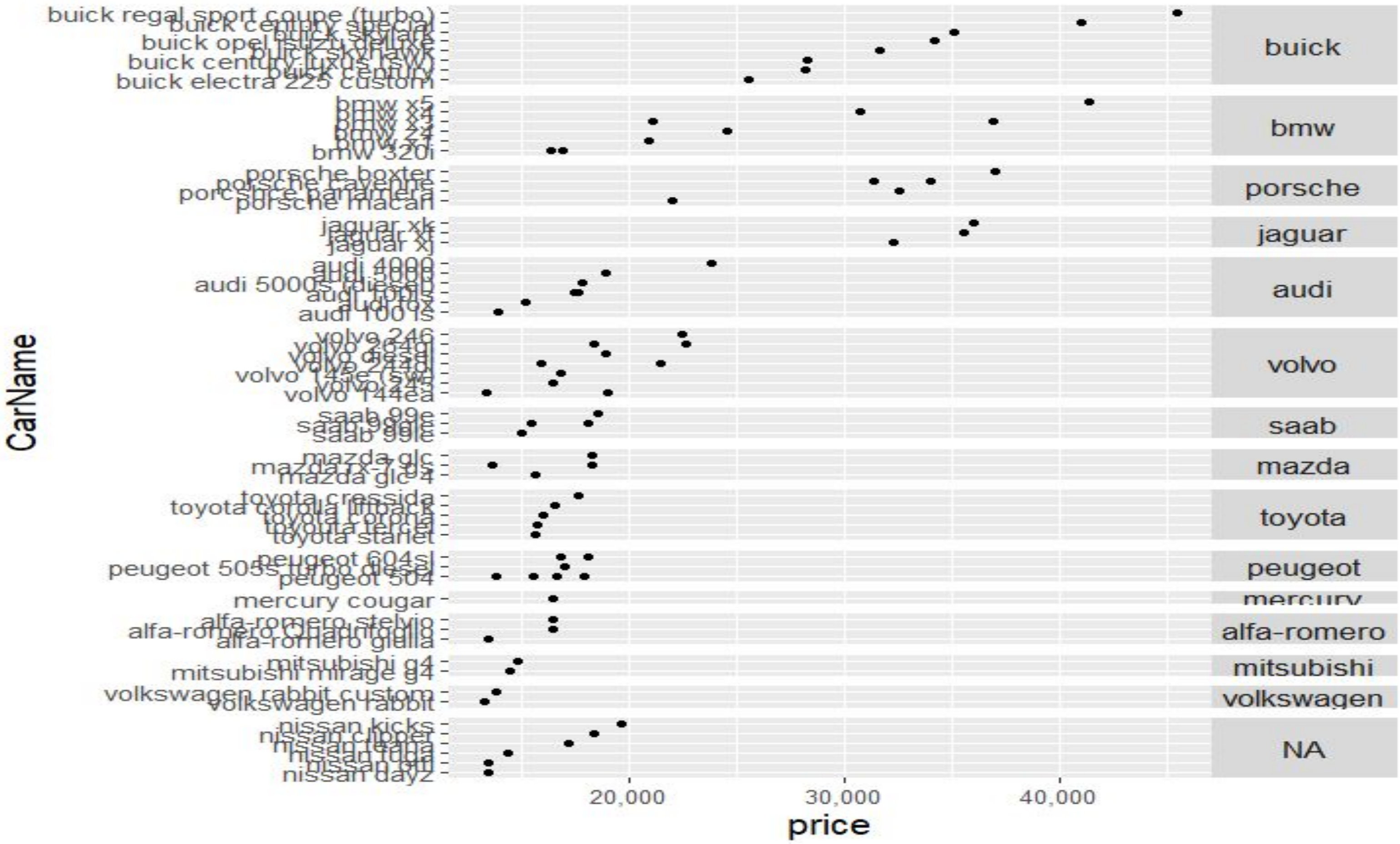
Which body type has the worst/best highway gallon per mile?





KM

Buick, BMW, Porsche & Jaguar are luxury cars



EDA

GC

How we build our model: model selection

- **How well those variables describe the price of a car?**
  - Use Multiple Linear Regression to create a model
    - Step 1: Used backwards stepwise regression
    - Step 2: show check for significant interaction terms
    - Step 3: show check of multiple regression model assumptions
    - Step 4: Box-cox method for a Power Transformation to satisfy line conditions
    - Step 5: Check multiple regression model assumptions again

KM

How we build our model

- Step 1: Use Backwards stepwise regression to select variables

```
Step:  AIC=3241.3
price ~ CarName + fueltype + drivewheel + enginelocation + wheelbase + 
    carwidth + enginetype + cylindernumber + enginesize + stroke + 
    compressionratio + horsepower + peakrpm

                   Df Sum of Sq        RSS    AIC
<none>                          1315706397 3241.3
- compressionratio  1  17628383 1333334780 3242.0
- horsepower        1  25839112 1341545509 3243.3
- fueltype          1  30601932 1346308329 3244.0
- carwidth          1  46974022 1362680419 3246.5
- drivewheel        1  56610752 1372317149 3247.9
- stroke            1  58433177 1374139575 3248.2
- peakrpm           1  63788631 1379495028 3249.0
- wheelbase         1  70042664 1385749061 3249.9
- enginetype        1  87913403 1403619800 3252.6
- CarName           1  88897396 1404603793 3252.7
- cylindernumber    1 222851923 1538558320 3271.4
- enginelocation    1 250876662 1566583059 3275.1
- enginesize        1 453850116 1769556513 3300.1

Call:
lm(formula = price ~ CarName + fueltype + drivewheel + enginelocation + 
    wheelbase + carwidth + enginetype + cylindernumber + enginesize + 
    stroke + compressionratio + horsepower + peakrpm, data = data)

Coefficients:
     (Intercept)           CarName          fueltype        drivewheel    enginelocation  
      -53534.637           132.029          9242.231          1149.270         11194.717  
       wheelbase          carwidth        enginetype    cylindernumber        enginesize  
         192.613           520.832          -609.407          1377.244            91.229  
          stroke  compressionratio        horsepower           peakrpm  
       -2147.224          -505.609            23.616             1.749  

> #AIC=3241.3
```

GC

How we build our model

- Step 2: Show check for significant interaction terms

```
# model3 = update(model2, ~.+CarName:wheelbase+CarName:carwidth +CarName:enginetype+CarName:cylindernumber+CarName:enginesize+CarName:horsepower)
# summary(model3)
#We had higher adjR^2 but more p-values that are less signigficant
#Adjusted R-squared:  0.9095

# model3 = update(model2, ~.+CarName:wheelbase+CarName:carwidth +CarName:enginetype+CarName:cylindernumber+CarName:enginesize+CarName:horsepower)
# summary(model3)
#Adjusted R-squared:  0.9095, more significant individual values
#9 predictors are not significatn

# model3 = update(model2, ~.+CarName:wheelbase+CarName:carwidth +CarName:enginetype+CarName:cylindernumber+CarName:horsepower)
# summary(model3)
#Adjusted R-squared:  0.9095 , but more p-values that are less signigficant
##9 predictors are not significatn
```

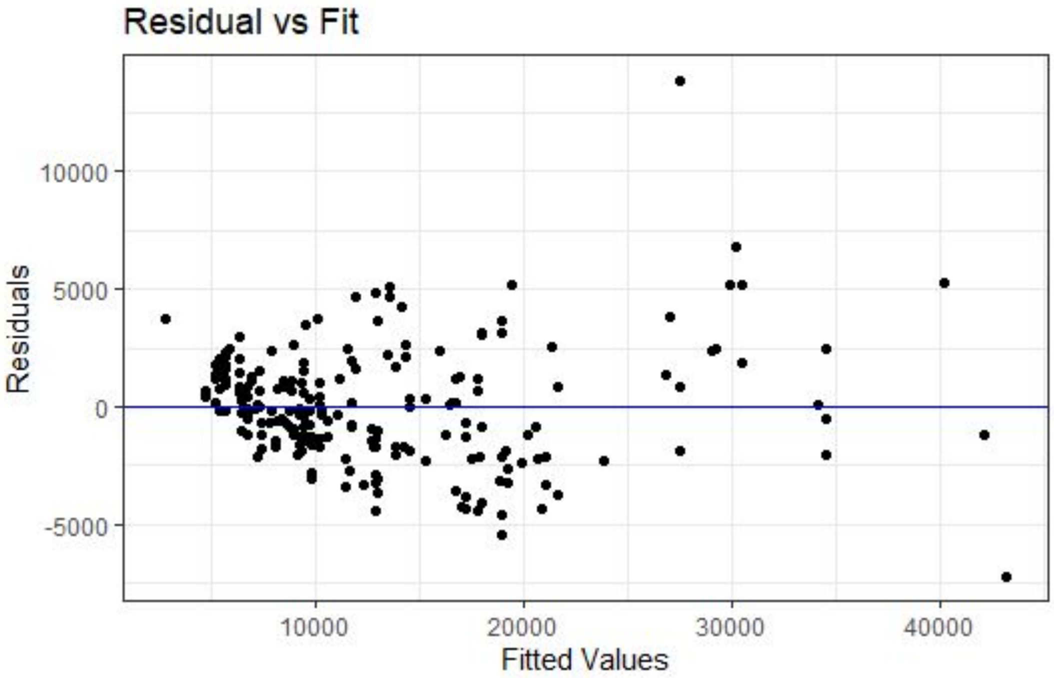
```
> anova(model2,model3)
Analysis of Variance Table

Model 1: price ~ CarName + fueltype + drivewheel + enginelocation + wheelbase +
    carwidth + enginetype + cylindernumber + enginesize + stroke +
    compressionratio + horsepower + peakrpm
Model 2: price ~ CarName + fueltype + drivewheel + enginelocation + wheelbase +
    carwidth + enginetype + cylindernumber + enginesize + stroke +
    compressionratio + horsepower + peakrpm + CarName:cylindernumber
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1    191 1315706397
2    190 1287335241  1  28371156 4.1873 0.0421 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

GC

How we build our model

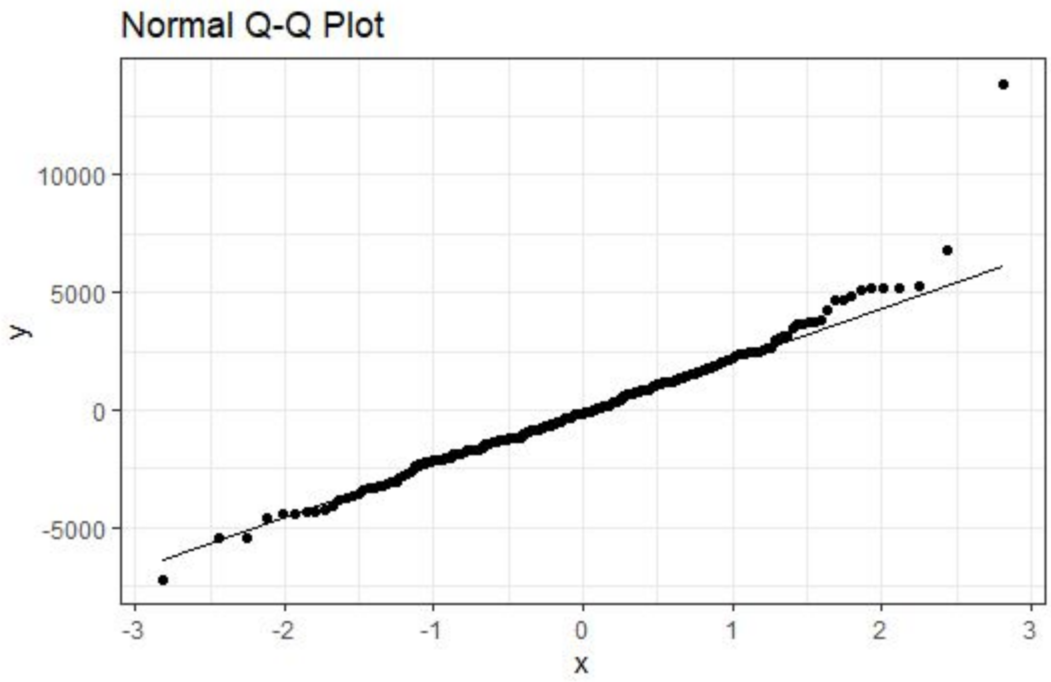
- Step 3: Show check of multiple regression model assumptions



KM

How we build our model

- Step 3: Show check of multiple regression model assumptions



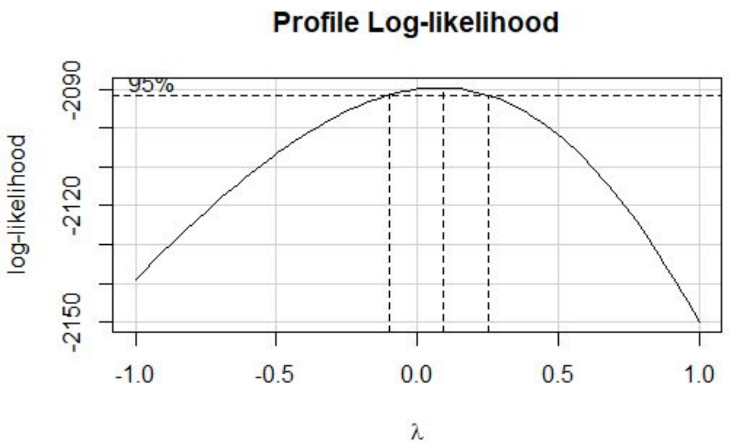
Shapiro-Wilk normality test

data: resid(model3)
W = 0.95893, p-value = 1.191e-05

KM

How we build our model

- Step 4: Box-cox method for a Power Transformation to satisfy line conditions (Lambda.opt = 0.09)

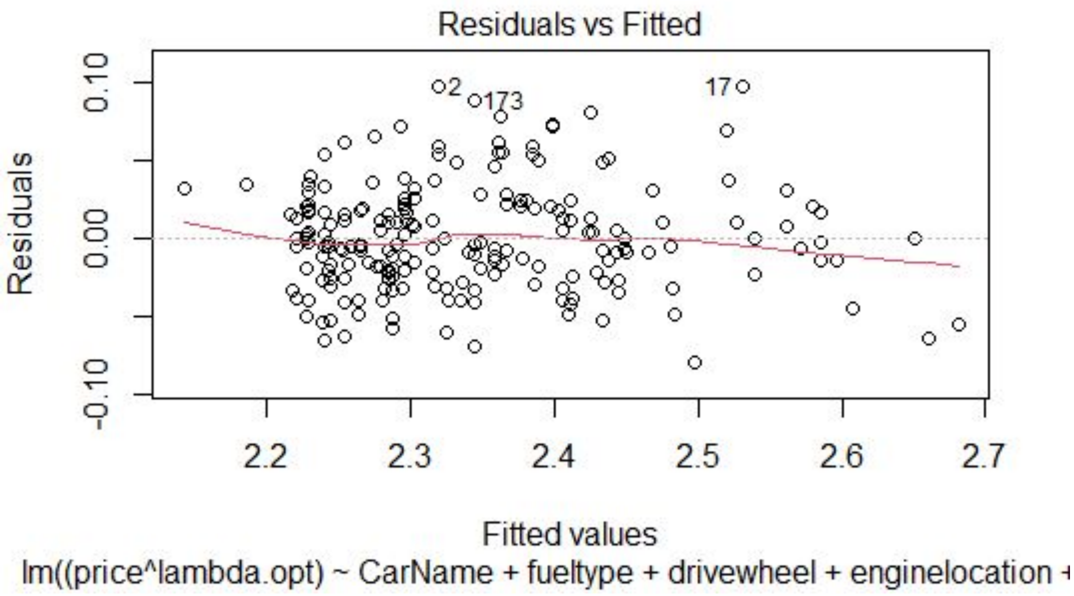



```
model4 = lm((price^lambda.opt) ~ CarName + fueltype + drivewheel + enginelocation +
       wheelbase + carwidth + enginetype + cylindernumber + enginesize +
       stroke + compressionratio + horsepower + peakrpm + CarName:cylindernumber,
     data = data)
```

KM

How we build our model

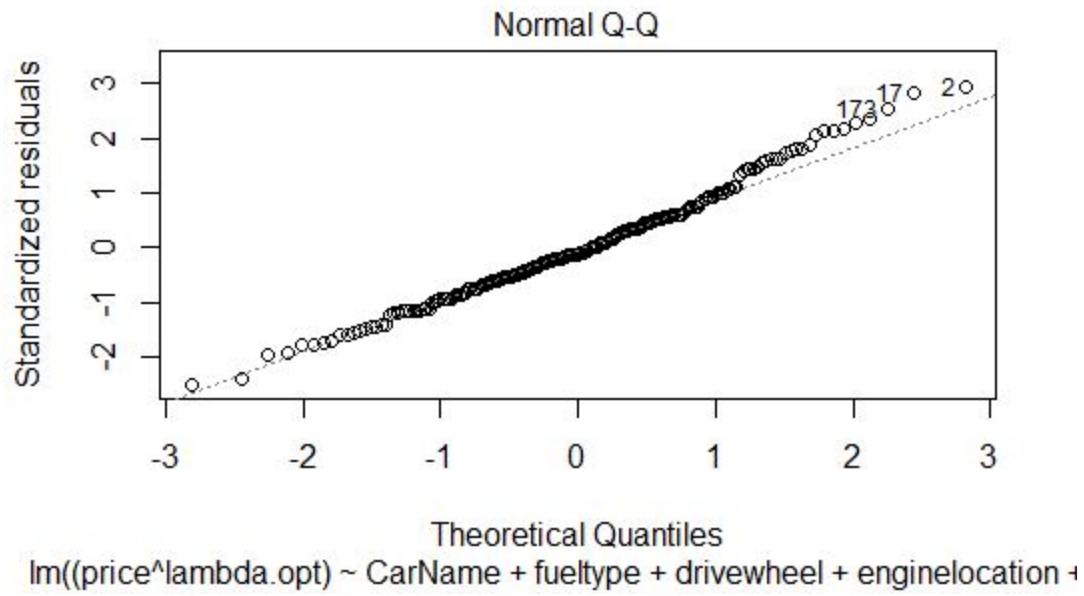
- Step 5: Check multiple regression model assumptions again



KM

How we build our model

- Step 5: Check multiple regression model assumptions again

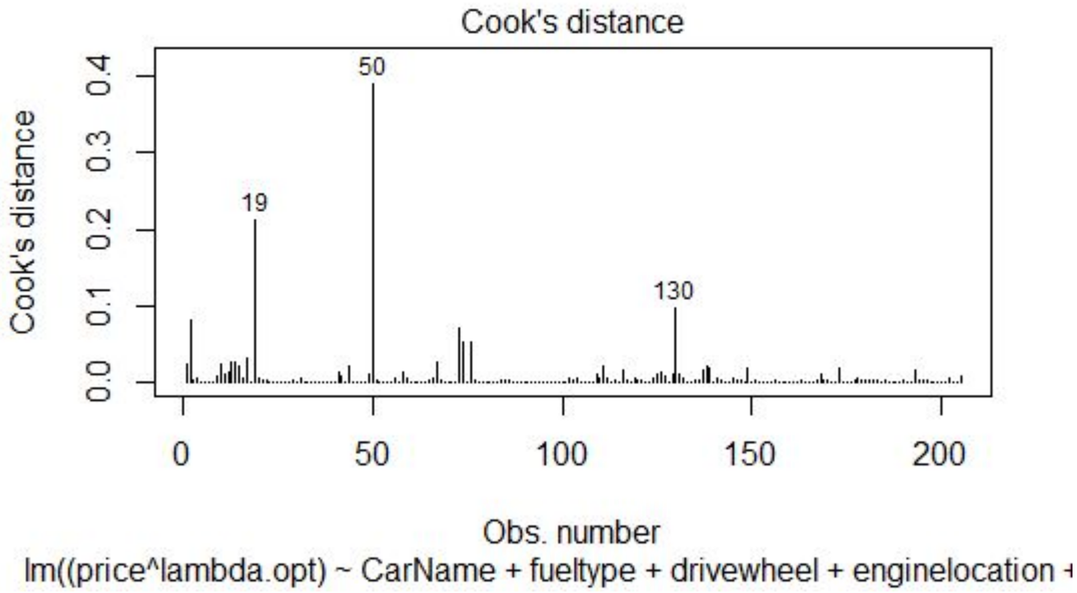


KM

How we build our model

- Step 5: Check multiple regression model assumptions again

Check high leverage points



KM

Our final model (Yay!!)

Car price^0.09 = car namex1+fuel typex2+drive wheelx3+engine locationx4+ wheel basex5+ car widthx6+ engine typex7+ cylinder numberx8+ engine sizex9+ strokex10+ compression ratiox11+ horse powerx12+ peak rpmx13+car name*cylinder number

YZ

Our final model (Yay!!)

- **How well those variables describe the price of a car?**

  The results from below indicate, our model represents 89% of the variation in the average price of a car.

```
Residual standard error: 2625 on 191 degrees of freedom
Multiple R-squared:  0.8989,    Adjusted R-squared:  0.8921
F-statistic: 130.7 on 13 and 191 DF,  p-value: < 2.2e-16
```

YZ

Regression Analysis

```
Call:
lm(formula = (price^lambda.opt) ~ CarName + fueltype + drivewheel +
    enginelocation + wheelbase + carwidth + enginetype + cylindernumber +
    enginesize + stroke + compressionratio + horsepower + peakrpm +
    CarName:cylindernumber, data = data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.080081 -0.023134 -0.004182  0.019577  0.098205

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             1.248e+00  1.494e-01   8.354 1.36e-14 ***
CarName                 2.405e-03  5.379e-04   4.471 1.34e-05 ***
fueltype                9.392e-02  5.975e-02   1.572 0.117648
drivewheel              2.201e-02  5.507e-03   3.997 9.17e-05 ***
enginelocation          1.111e-01  2.526e-02   4.401 1.80e-05 ***
wheelbase               3.479e-03  8.218e-04   4.233 3.59e-05 ***
carwidth                8.973e-03  2.734e-03   3.282 0.001229 **
enginetype             -8.216e-03  2.376e-03  -3.458 0.000672 ***
cylindernumber          2.674e-02  5.604e-03   4.773 3.62e-06 ***
enginesize              5.944e-04  1.566e-04   3.796 0.000198 ***
stroke                 -1.887e-02  1.014e-02  -1.860 0.064373 .
compressionratio       -3.884e-03  4.309e-03  -0.901 0.368572
horsepower              9.514e-04  1.659e-04   5.735 3.79e-08 ***
peakrpm                 1.007e-05  7.880e-06   1.279 0.202611
CarName:cylindernumber -1.660e-03  4.595e-04  -3.612 0.000389 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03571 on 190 degrees of freedom
Multiple R-squared:     0.9,	Adjusted R-squared:  0.8926
F-statistic: 122.2 on 14 and 190 DF,  p-value: < 2.2e-16
```

- Carname (categorical predictor) is significant predictor of car price^0.09 after controlling the remaining variables.
- Horsepower (continuous predictor) is significant predictor of car price^0.09 after controlling the remaining variables.

KM

Prediction using our Model

A car fanatic would like to know how much their dream car would cost? We are able to answer this question using our model.

First, we gather all the information of the customers dream car: A larger size, electric Porsche with all wheel drive and higher wheelbase, for increased tire traction, and engine specifics for a larger, powerful, efficient engine.

What is the car price^0.09 for a Porsche with the above specifications?

```
> predictionans
       fit      lwr     upr
1 1.882564 1.541358 2.22377
```

According to our results, this customers dream car is

approximately (122.403, 7187.106), unit is Dollar.

KM

Confidence interval (cont'd)

We also test for the average price range of our customers dream car using a confidence interval.

What is the average car price^0.09 for a Porsche with the previous specifications?

```
confidenceans
     fit      lwr      upr
1.882564 1.548707 2.216421
```

According to our results, the average prices for this customers dream car range from, $129.05 to $6926.84 (unit is k). These finding match theoretical principal of our prediction interval being smaller than our confidence interval.

KM

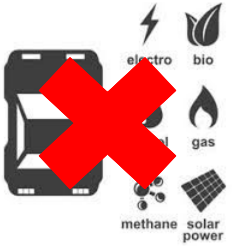
Results

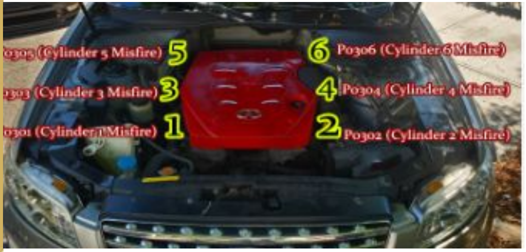
A multiple linear regression was conducted to investigate whether car name, fuel type, drive wheel, engine location, wheel base, car width, engine type, cylinder number, engine size, stroke, compression ratio, horse power, peak rpm, and the interaction between car name and cylinder number would predict car price.

YZ

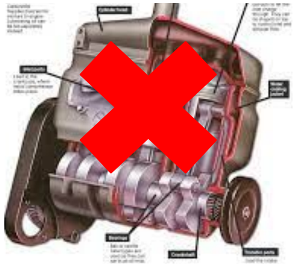
Results (continued)

- Car name, drive wheel, engine location, wheel base, car width, engine type, cylinder number, engine size, horsepower, and the interaction between car name and cylinder number were significant predictors of car price^0.09, controlling for remaining variables.
- However, fuel type, engine size, stroke, compression ratio, and peak rpm were not strong predictors of car price^0.09.
- The results highlight that car name, drive wheel, engine location, wheel base, car width, engine type, cylinder number, engine size, horsepower, and the interaction between car name and cylinder number are important for car price.

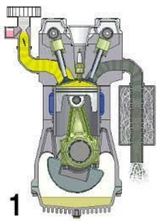




Car name fuel type drive wheel engine location engine type cylinder number



Engine size stroke door number car body fuel system



•••

Wheelbase car length, car width, car height curb weight horsepower

YZ

Challenges

- Cleaning Dataset
  - Checking for faulty data
  - Creating Dummy variables
  - Relabeling the carname
- Computing confidence interval and prediction interval while our model is a little bit complicated
  - Figuring out what the values are

GC

Further Work

- Split data into Train/Test
- Apply PCA
- More EDA

GC

Thank You