How to buy a used car smarter
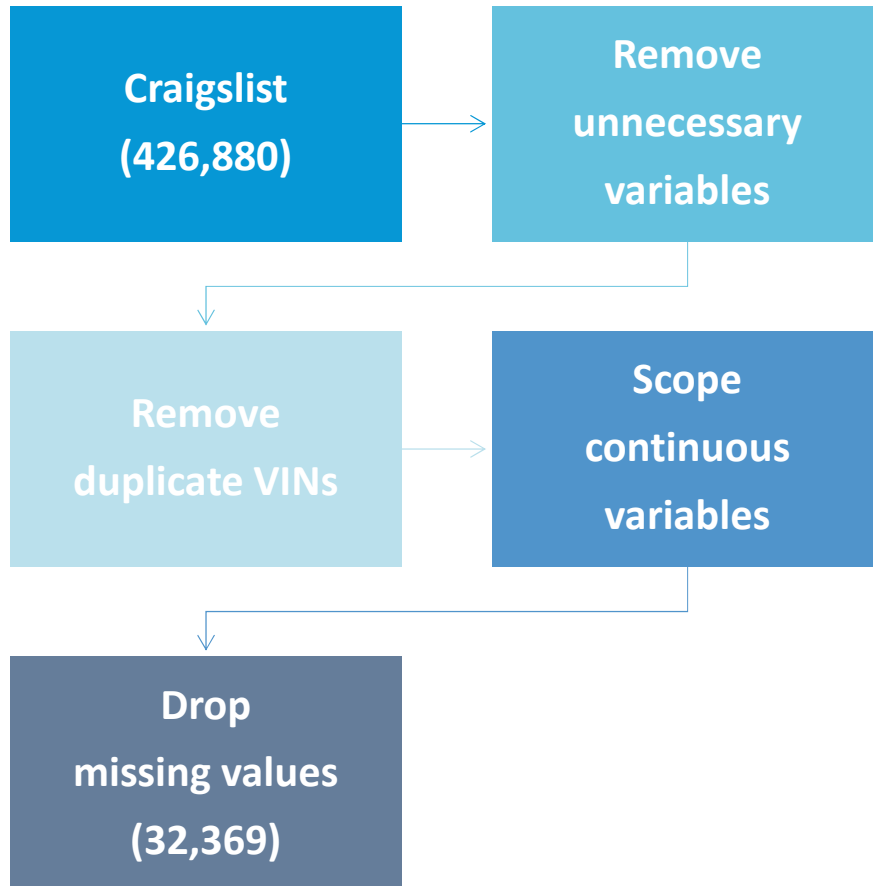
Minjung Lee

Introduction

This is an analysis to find out factors affecting the price of used cars and lay the foundation for wise consumption based on them.

Questions of Interest

- Which variables have major effects on the price of a used car?

- Does the price differ by entry year?

- Does the price differ by state?

Data

```
┌─────────────────┐     ┌─────────────────┐
│  Craigslist     │ ──> │  Remove         │
│  (426,880)      │     │  unnecessary    │
│                 │     │  variables      │
└─────────────────┘     └─────────────────┘
         │                       │
         v                       v
┌─────────────────┐     ┌─────────────────┐
│  Remove         │ ──> │  Scope          │
│  duplicate VINs │     │  continuous     │
│                 │     │  variables      │
└─────────────────┘     └─────────────────┘
         │
         v
┌─────────────────┐
│  Drop           │
│  missing values │
│  (32,369)       │
└─────────────────┘
```

Variables

- Response variable : *price*
- Predictors : *year, odometer, manufacturer, condition, type, color*, etc
- model of vehicle
  - representative of other variables (*cylinders, fuel, type*..)
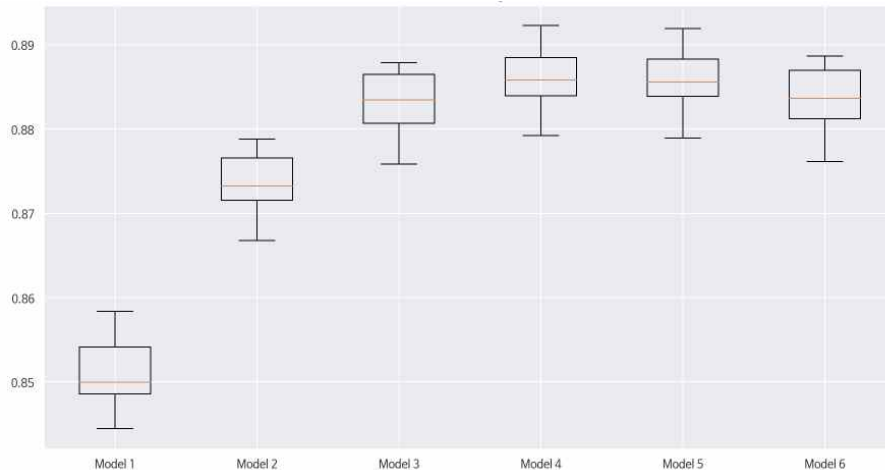 - removed to avoid multicollinearity

EDA
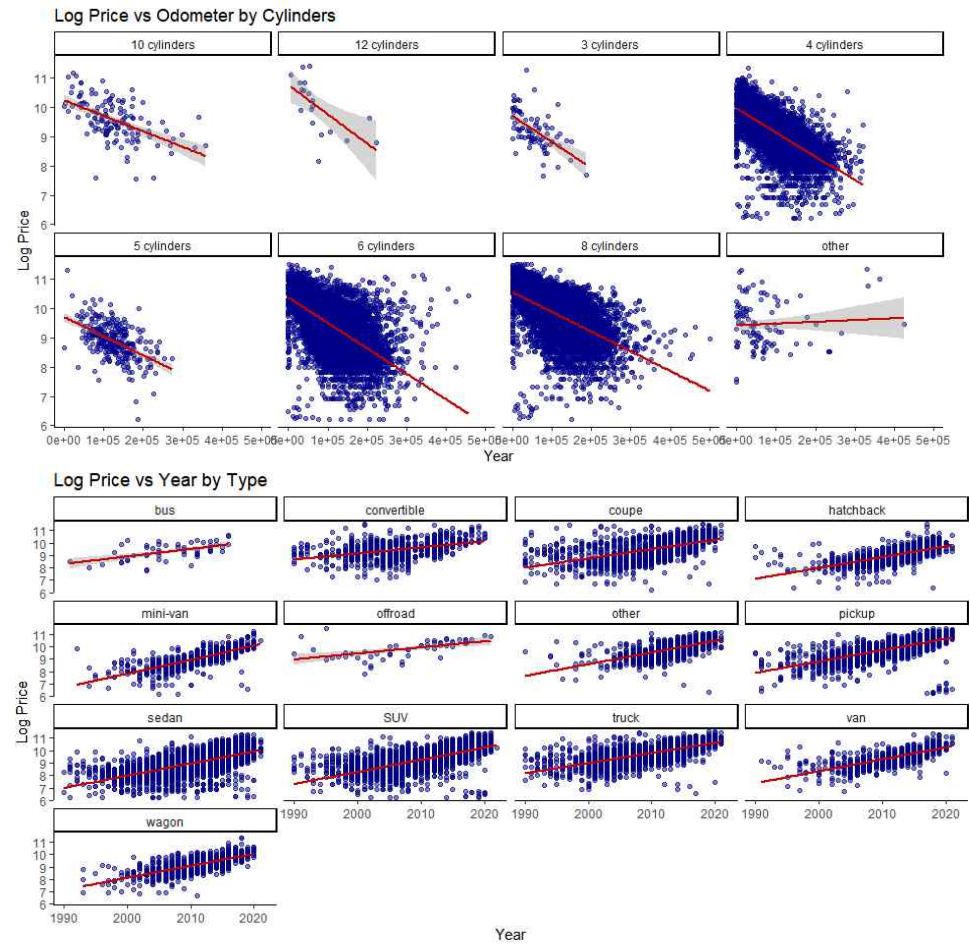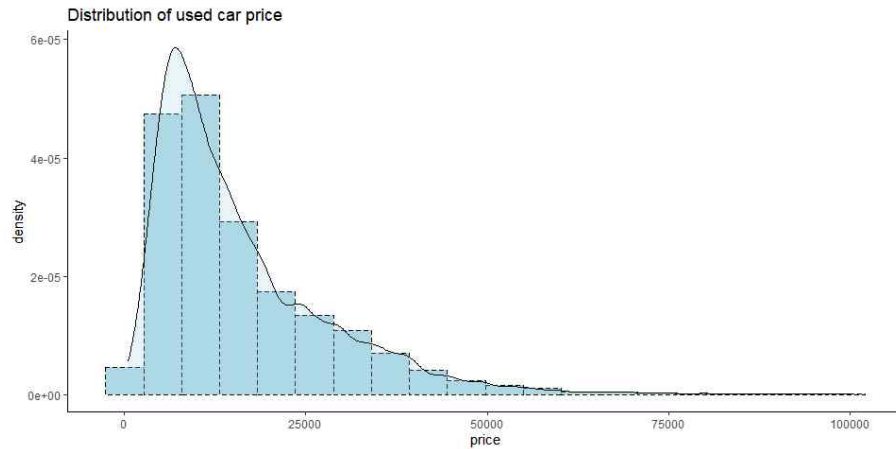
Model

$$log(price)_i = \beta_0 + \beta_1 yearc_i + \beta_2 agec_i^3 + \beta_3 odometerc_i + \beta_4 manufacturer_i + \beta_5 cylinders_i + \beta_6 fuel_i$$

$$+\beta_7 type_i + \beta_8 condition_i + \beta_9 transmission_i + \beta_{10} title\_status_i + \beta_{11} state_i + \beta_{12} paint\_color_i$$

$$+\beta_{13} yearc_i : manufacturer_i + \beta_{14} odometerc_i : type_i + \beta_{15} yearc_i : type_i + \beta_{16} odometerc_i : fuel_i + \beta_{17} cylinders_i : odometerc_i$$

Model Selection

- K-fold Cross Validation

- Forward/Backward/Stepwise selection

  with AIC/BIC

  - AIC= BIC+ *color*

Result

- Adjusted R-Squared : 0.7915
- p-value : < 2.2e-16

- price-year

- Advanced Model Validation
- Multicollinearity
- Model Assessment
  - QQ plot

```
Call:
lm(formula = log(price) ~ yearc + year3 + condition + cylinders +
    fuel + odometerc + title_status + transmission + paint_color +
    state + manufacturer + type + yearc:manufacturer + yearc:type +
    fuel:odometerc + cylinders:odometerc + odometerc:type, data = car_removed)

Residuals:
    Min      1Q  Median      3Q     Max
-4.4551 -0.1499  0.0166  0.1787  2.0072

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            1.057e+01  6.966e-02 151.727  < 2e-16 ***
yearc                  8.772e-02  7.809e-03  11.233  < 2e-16 ***
year3                 -1.245e-04  3.823e-06 -32.570  < 2e-16 ***
conditionfair         -6.428e-01  1.624e-02 -39.580  < 2e-16 ***
conditiongood         -7.189e-02  4.623e-03 -15.550  < 2e-16 ***
conditionlike new      1.711e-03  7.116e-03   0.240 0.810022
conditionnew           5.722e-02  3.203e-02   1.786 0.074035 .
conditionsalvage      -8.380e-01  5.746e-02 -14.584  < 2e-16 ***
cylinders12 cylinders  2.973e-01  9.842e-02   3.021 0.002525 **
cylinders3 cylinders  -8.468e-01  7.241e-02 -11.695  < 2e-16 ***
cylinders4 cylinders  -6.467e-01  3.312e-02 -19.523  < 2e-16 ***
cylinders5 cylinders  -4.993e-01  4.080e-02 -12.239  < 2e-16 ***
cylinders6 cylinders  -3.647e-01  3.282e-02 -11.112  < 2e-16 ***
cylinders8 cylinders  -8.257e-02  3.279e-02  -2.518 0.011806 *
cylindersother        -6.385e-01  5.763e-02 -11.080  < 2e-16 ***
fuelelectric          -4.764e-01  6.174e-02  -7.717 1.23e-14 ***
fuelgas               -4.875e-01  1.086e-02 -44.902  < 2e-16 ***
fuelhybrid            -4.698e-01  2.108e-02 -22.288  < 2e-16 ***
fuelother             -3.796e-01  2.172e-02 -17.482  < 2e-16 ***
odometerc             -1.304e-06  7.652e-07  -1.704 0.088304 .
title_statuslien      -6.553e-04  2.768e-02  -0.024 0.981112
title_statusmissing   -9.960e-01  1.444e-01  -6.896 5.44e-12 ***
title_statusparts only -1.526e+00 1.774e-01  -8.601  < 2e-16 ***
title_statusrebuilt   -1.971e-01  1.278e-02 -15.428  < 2e-16 ***
title_statussalvage   -2.532e-01  2.255e-02 -11.228  < 2e-16 ***
transmissionmanual     9.970e-02  9.411e-03  10.595  < 2e-16 ***
transmissionother      1.092e-01  1.003e-02  10.884  < 2e-16 ***
paint_colorblue       -3.129e-02  7.528e-03  -4.156 3.25e-05 ***
paint_colorbrown      -3.823e-02  1.341e-02  -2.851 0.004358 **
paint_colorcustom      5.717e-03  1.320e-02   0.433 0.664943
paint_colorgreen      -1.288e-02  1.336e-02  -0.964 0.334952
paint_colorgrey       -2.434e-02  7.219e-03  -3.371 0.000750 ***
paint_colororange      6.461e-02  2.721e-02   2.374 0.017589 *
paint_colorpurple     -5.146e-02  3.732e-02  -1.379 0.167927
paint_colorred        -1.565e-02  7.913e-03  -1.978 0.047943 *
paint_colorsilver     -4.118e-02  6.686e-03  -6.159 7.42e-10 ***
paint_colorwhite      -1.212e-02  6.112e-03  -1.984 0.047283 *
paint_coloryellow      3.724e-02  2.540e-02   1.466 0.142601
stateal               -1.241e-01  2.440e-02  -5.088 3.65e-07 ***
statear               -1.389e-01  3.041e-02  -4.567 4.97e-06 ***
stateaz               -2.305e-01  2.233e-02 -10.321  < 2e-16 ***
stateca               -1.524e-01  2.013e-02  -7.572 3.76e-14 ***
stateco               -1.782e-01  2.273e-02  -7.840 4.64e-15 ***
statect               -2.772e-01  2.538e-02 -10.923  < 2e-16 ***
statedc               -2.080e-01  2.697e-02  -7.713 1.26e-14 ***
```
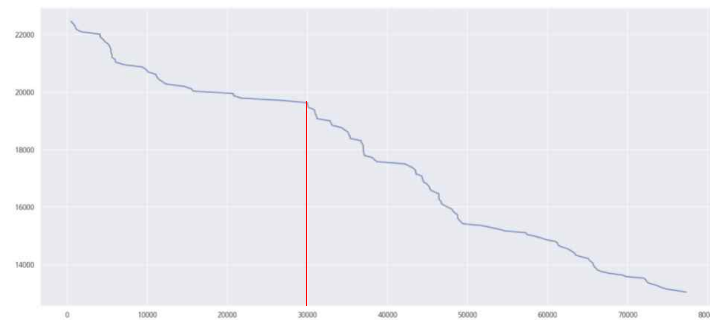
- price-state
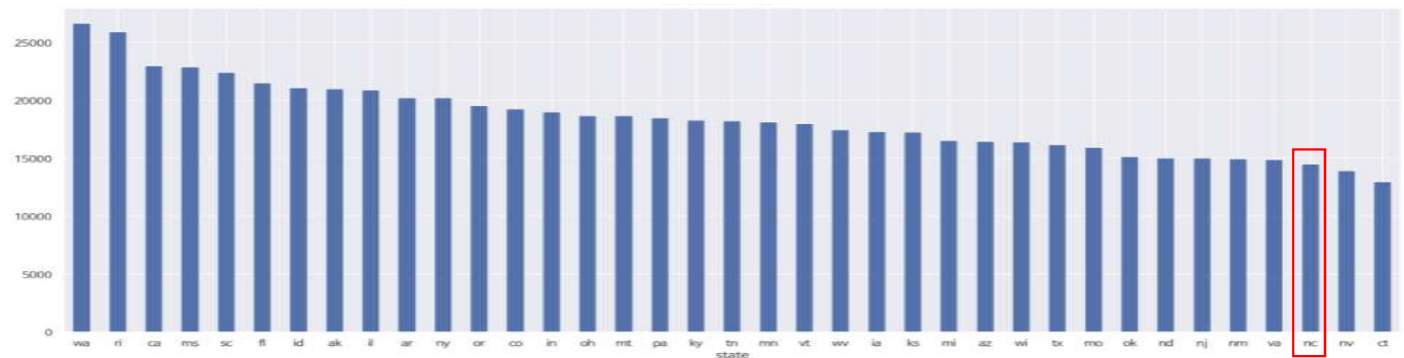
Limitations

- Characteristics of Craigslist
  - This is not sales data, but data registered for sales.
  - Since the seller, the individual, directly inputs the information, there is a case of inputting nonsensical data.

- The type and price of preferred cars may vary depending on the situation.
  - For example, a small car is preferred when the economy is bad.
  - The observation period is short. (The dataset is for three months in 2021.)

- This is data about only transactions between individuals.
  - Transaction data through used car companies such as CarMax or Carvana were not reflected.