Ford Focus Price Analysis

Xinyi Gao, Yiyang Li, Damien MacFarland, Neha Sinha, Jinda Zhang

1 Introduction

The dataset contains contains information of price, transmission, mileage, fuel type, road tax, miles per gallon (mpg), and engine size. 500 out of 100,000 data are randomly selected to model the relationship between model price and other explanatory variables.

2 Exploratory Analysis

Firstly, we use glimpse() in R to investage how the dataset look like.

```
## Rows: 500
## Columns: 10
## $ price        <int> 11000, 18960, 13999, 16350, 15960, 4495, 14500, 18998, 11~
## $ year         <int> 2017, 2019, 2017, 2019, 2019, 2007, 2018, 2017, 2016, 201~
## $ transmission <fct> Manual, Manual, Manual, Manual, Manual, Automatic, Manual~
## $ mileage      <int> 50940, 824, 13631, 14410, 12388, 30000, 6575, 10222, 2484~
## $ fuelType     <fct> Diesel, Petrol, Petrol, Petrol, Diesel, Petrol, Petrol, P~
## $ engineSize   <dbl> 1.5, 1.0, 1.0, 1.0, 1.5, 1.6, 1.0, 2.0, 2.0, 1.0, 2.3, 1.~
## $ enginefactor <fct> 1.5, 1, 1, 1, 1.5, 1.6, 1, 2, 2, 1, 2.3, 1.5, 1, 1.5, 1.5~
## $ LogPrice     <dbl> 9.305651, 9.850087, 9.546741, 9.701983, 9.677841, 8.41072~
## $ LogMileage   <dbl> 10.838404, 6.714171, 9.520102, 9.575678, 9.424484, 10.308~
## $ age          <dbl> 4, 2, 4, 2, 2, 14, 3, 4, 5, 2, 2, 4, 2, 4, 3, 2, 1, 4, 2,~
```

Table 1: Summary statistics on price of 500 UK Used Car Data set.

n	Mean	St.Dev	Min	Q1	Median	Q3	Max
500	14343.6	4521.6	1850	11337.5	14496.5	17491.25	28930

Table 1 shows that the summaries of price of 500 UK Used Car Data set. For example the mean price 13480.5 pounds. We also note that the variability in the price as 4784.7 pounds. The messages can be easily seen the in the following boxplot which summarise the distribution of car price.

We can visualize the distribution of price by using summary in the following boxplot.

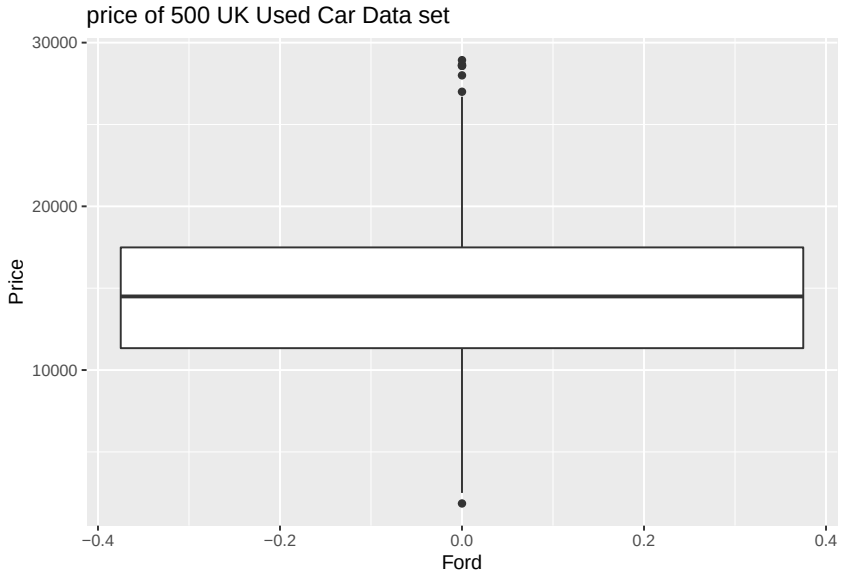
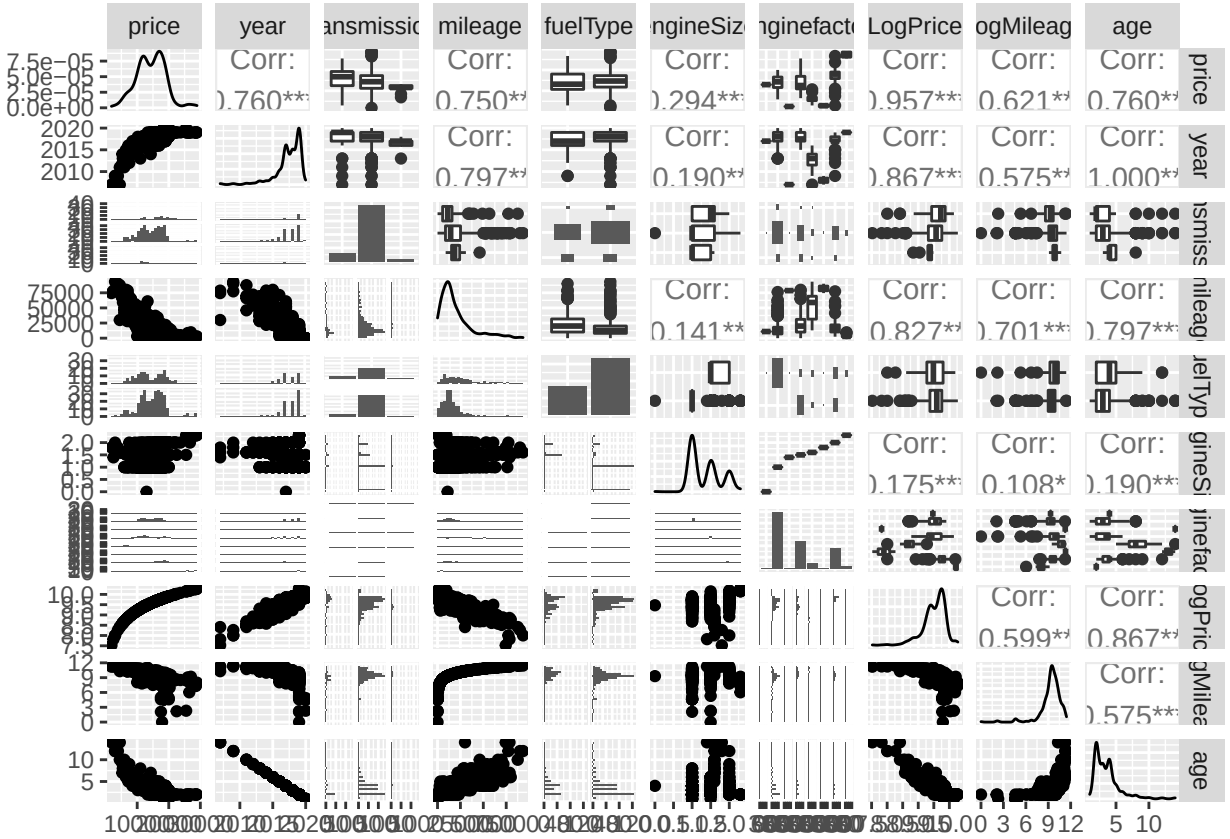


Figure 1: Price of 500 used Ford cars.

Then, we use the ggpairs function in the GGally package to generate an informative set of graphical and numerical summaries that illuminate the relationships bewteen pairs of variables.



- high correlation between price and c(year(+ve),mileage(-ve),logmileage(-ve))

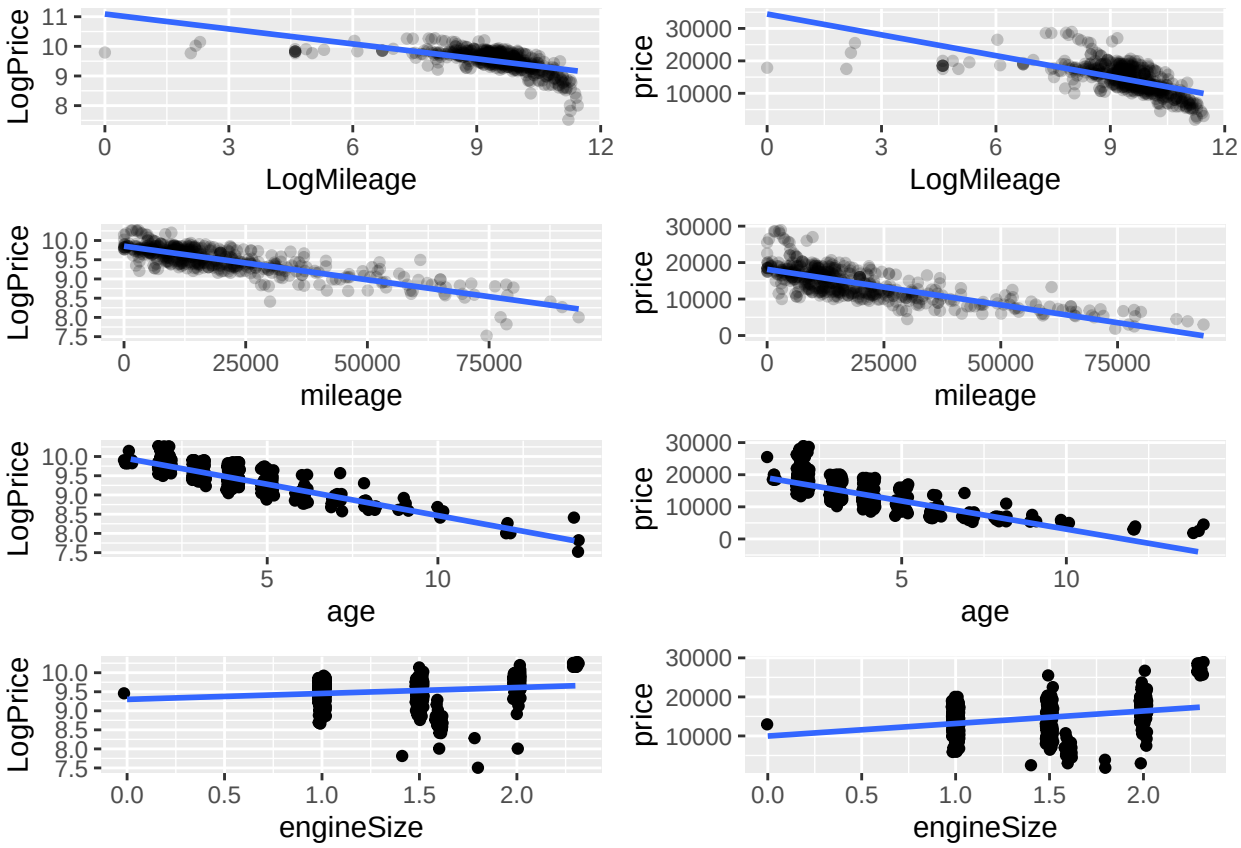
- year and mileage also have high correlation with oneanother
- log of price has very strong correlations with mileage and year and mildly strong correlation with log of mileage
- 2 levels of transmission and both fuel types seem to have the same distribution when compared to other variables so could probably be omitted. third transmission level had very few observations so hard to make judgment on any effects so can omit too.
- engine seize has weak to moderate correlation, keep for now and maybe discard later as we will only have a small number of variables anyway

We futher simplify the dataset by removing transmission, fuelType,year variables.

Below is the summary statisics of our reviesed dataset.

skim_type	skim_variable	n	factor.ordered	factor.n_unique	factor.top_counts	numeric.mean	numeric.sd	numeric.p25	numeric.p50	numeric.p75
factor	enginefactor	500	FALSE	8	1: 249, 1.5: 122, 2: 91, 1.6: 23	NA	NA	NA	NA	NA
numeric	price	500	NA	NA	NA	14343.65	4521.64	11337.50	14496.50	17491.25
numeric	mileage	500	NA	NA	NA	19477.50	17405.97	8228.75	13990.50	24434.50
numeric	engineSize	500	NA	NA	NA	1.36	0.41	1.00	1.20	1.60
numeric	LogPrice	500	NA	NA	NA	9.51	0.37	9.34	9.58	9.77
numeric	LogMileage	500	NA	NA	NA	9.40	1.31	9.02	9.55	10.10
numeric	age	500	NA	NA	NA	3.57	1.96	2.00	3.00	4.00

From summaries we see that engine size, logprice and logmileage are all on same scale now so probably advance with these variables, and year. We can verify that first by looking at plots.



- From the plot we may use the model price ~ mileage+enginesize+year or logprice ~ mileage+year+enginesize or drop enginesize in both.

- SHOULD WE HAVE YEAR? or should it be years ago/age? as year is relative to present day - it requires further work section.
- maybe try mileage and engine size since mileage and year are strongly correlated (multicolinarity avoided too)

3 Formal Analysis

Tables below show the estimate coefficients and their confidence intervals for their respective models

Table 2: Parameter estimates for model 1

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	12357.41	372.40	33.18	0	11625.73	13089.08
mileage	-0.21	0.01	-34.06	0	-0.22	-0.20
engineSize	4458.98	259.54	17.18	0	3949.05	4968.90

Table 3: Parameter estimates for model 2

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	18140.71	200.89	90.30	0	17746.02	18535.40
mileage	-0.20	0.01	-25.34	0	-0.21	-0.18

Table 4: Parameter estimates for model 3

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	14546.10	349.60	41.61	0	13859.22	15232.98
age	-1956.91	49.81	-39.28	0	-2054.79	-1859.04
engineSize	4977.11	235.88	21.10	0	4513.67	5440.55

Table 5: Parameter estimates for model 4

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	20611.80	273.66	75.32	0	20074.12	21149.48
age	-1756.77	67.27	-26.12	0	-1888.93	-1624.61

Table 6: Parameter estimates for model 5

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	14025.04	307.74	45.57	0	13420.40	14629.67
age	-1249.86	71.17	-17.56	0	-1389.70	-1110.02
engineSize	4930.56	205.77	23.96	0	4526.28	5334.84
mileage	-0.10	0.01	-12.54	0	-0.12	-0.08

Table 7: Parameter estimates for model 6

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	21484.63	1564.53	13.73	0.00	18410.46	24558.79
as.factor(age)2	-2943.21	532.67	-5.53	0.00	-3989.87	-1896.56
as.factor(age)3	-6267.74	541.96	-11.56	0.00	-7332.63	-5202.84
as.factor(age)4	-8509.63	539.82	-15.76	0.00	-9570.32	-7448.94
as.factor(age)5	-10147.24	575.86	-17.62	0.00	-11278.75	-9015.73
as.factor(age)6	-12192.62	614.22	-19.85	0.00	-13399.50	-10985.75
as.factor(age)7	-12828.70	720.94	-17.79	0.00	-14245.27	-11412.12
as.factor(age)8	-13928.31	755.69	-18.43	0.00	-15413.17	-12443.44
as.factor(age)9	-15389.03	979.49	-15.71	0.00	-17313.65	-13464.42
as.factor(age)10	-15360.29	1032.25	-14.88	0.00	-17388.57	-13332.01
as.factor(age)12	-18868.88	1135.39	-16.62	0.00	-21099.82	-16637.95
as.factor(age)14	-17667.32	1414.91	-12.49	0.00	-20447.49	-14887.15
as.factor(engineSize)1	-1823.94	1476.65	-1.24	0.22	-4725.43	1077.55
as.factor(engineSize)1.4	-1322.31	2461.50	-0.54	0.59	-6158.92	3514.31
as.factor(engineSize)1.5	-1114.46	1478.62	-0.75	0.45	-4019.80	1790.88
as.factor(engineSize)1.6	-945.59	1558.45	-0.61	0.54	-4007.79	2116.62
as.factor(engineSize)1.8	-344.03	2038.41	-0.17	0.87	-4349.32	3661.27
as.factor(engineSize)2	3327.38	1480.81	2.25	0.03	417.71	6237.04
as.factor(engineSize)2.3	8707.86	1544.24	5.64	0.00	5673.56	11742.16

lots of levels of engine size as a factor have 0 in their CI therefore cannot use model 6

Table 8: Parameter estimates for model 7

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	14301.99	680.87	21.00	0	12964.18	15639.80
as.factor(age)2	-2890.98	661.90	-4.37	0	-4191.52	-1590.44
as.factor(age)3	-6396.84	673.37	-9.50	0	-7719.90	-5073.78
as.factor(age)4	-8785.92	669.90	-13.12	0	-10102.17	-7469.66
as.factor(age)5	-10657.16	714.24	-14.92	0	-12060.53	-9253.79
as.factor(age)6	-12856.62	761.54	-16.88	0	-14352.93	-11360.31
as.factor(age)7	-13593.74	850.72	-15.98	0	-15265.28	-11922.20
as.factor(age)8	-15092.95	852.61	-17.70	0	-16768.20	-13417.70
as.factor(age)9	-16799.56	1125.75	-14.92	0	-19011.48	-14587.63
as.factor(age)10	-17326.36	1123.59	-15.42	0	-19534.04	-15118.68
as.factor(age)12	-20235.50	1245.35	-16.25	0	-22682.42	-17788.58
as.factor(age)14	-19558.44	1241.10	-15.76	0	-21997.02	-17119.87
engineSize	5126.95	203.52	25.19	0	4727.07	5526.84

Table 9: Parameter estimates for model 8

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	14334.88	583.02	24.59	0	13189.32	15480.43
as.factor(age)2	-2124.00	569.68	-3.73	0	-3243.33	-1004.66
as.factor(age)3	-4969.50	586.42	-8.47	0	-6121.73	-3817.27
as.factor(age)4	-6557.09	597.43	-10.98	0	-7730.95	-5383.23
as.factor(age)5	-8077.57	641.39	-12.59	0	-9337.81	-6817.33
as.factor(age)6	-7874.28	751.35	-10.48	0	-9350.57	-6397.99
as.factor(age)7	-8747.26	813.91	-10.75	0	-10346.49	-7148.04
as.factor(age)8	-10817.45	797.24	-13.57	0	-12383.90	-9251.00
as.factor(age)9	-10687.48	1067.17	-10.02	0	-12784.31	-8590.65
as.factor(age)10	-10758.71	1080.60	-9.96	0	-12881.93	-8635.49
as.factor(age)12	-11763.59	1240.93	-9.48	0	-14201.84	-9325.35
as.factor(age)14	-13612.73	1152.29	-11.81	0	-15876.81	-11348.65
engineSize	5104.48	174.28	29.29	0	4762.05	5446.91
mileage	-0.10	0.01	-13.35	0	-0.11	-0.08

Table 10: Parameter estimates for model 9

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	22095.87	701.72	31.49	0	20717.16	23474.58
LogMileage	-940.71	78.55	-11.98	0	-1095.04	-786.38
engineSize	4970.82	207.96	23.90	0	4562.24	5379.41
age	-1593.31	53.39	-29.84	0	-1698.21	-1488.42

- No coefficient for age =13 as a factor, therefore we could not predict a price for that scenario
- From the below table we would choose 5 (age as factors do not work for ages not defined in the model) & (engineSize as factor always contained 0's in its confidence intervals) (model 5: 83% Radj, min AIC & BIC) nothing gained from logging mileage

```
## # A tibble: 9 x 6
##   model r.squared adj.r.squared   p.value   AIC   BIC
##   <int>     <dbl>         <dbl>     <dbl> <dbl> <dbl>
## 1     1     0.726         0.725 2.03e-140 9195. 9212.
## 2     2     0.563         0.562 1.31e- 91 9426. 9439.
## 3     3     0.777         0.777 7.21e-163 9091. 9108.
## 4     4     0.578         0.577 2.37e- 95 9409. 9422.
## 5     5     0.831         0.830 5.29e-191 8956. 8977.
## 6     6     0.898         0.895 2.30e-225 8732. 8816.
## 7     7     0.841         0.837 1.46e-185 8944. 9003.
## 8     8     0.884         0.880 3.09e-217 8790. 8853.
## 9     9     0.827         0.826 1.09e-188 8966. 8987.
```

3.1 Log models

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	9.51	0.03	343.61	0	9.45	9.56
mileage	0.00	0.00	-40.15	0	0.00	0.00
engineSize	0.26	0.02	13.76	0	0.23	0.30

CI for mileage is 0, so drop in this model.

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	9.85	0.01	709.10	0	9.83	9.88
mileage	0.00	0.00	-32.87	0	0.00	0.00

CI for mileage is 0, so we may drop this model.

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	9.71	0.02	449.04	0	9.67	9.75
age	-0.18	0.00	-57.01	0	-0.18	-0.17
engineSize	0.32	0.01	21.56	0	0.29	0.34

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	10.09	0.02	590.14	0	10.06	10.13
age	-0.16	0.00	-38.77	0	-0.17	-0.16

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	9.67	0.02	560.98	0	9.64	9.70
age	-0.12	0.00	-30.33	0	-0.13	-0.11
engineSize	0.31	0.01	26.98	0	0.29	0.33
mileage	0.00	0.00	-17.34	0	0.00	0.00

as above, mileage is 0, try log of mileage

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	10.05	0.05	216.83	0	9.96	10.14
age	-0.16	0.00	-45.24	0	-0.17	-0.15
engineSize	0.31	0.01	22.89	0	0.29	0.34
LogMileage	-0.04	0.00	-8.07	0	-0.05	-0.03

```
## # A tibble: 6 x 6
##   model r.squared adj.r.squared   p.value   AIC   BIC
##   <int>     <dbl>         <dbl>     <dbl> <dbl> <dbl>
## 1     1     0.772         0.771 4.55e-160 -312. -295.
## 2     2     0.685         0.684 7.59e-127 -152. -140.
## 3     3     0.871         0.871 4.14e-222 -599. -582.
## 4     4     0.751         0.751 1.58e-152 -271. -259.
## 5     5     0.920         0.919 1.79e-271 -834. -813.
## 6     6     0.886         0.886 9.82e-234 -659. -638.
```

- we would select model logm3 from the logs, but only after checking assumptions of model lm5 87% R, p is significant, negative AIC and BIC???? No zero's in the confidence intervals, need to explain what confidence intervals infer

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	9.71	0.02	449.04	0	9.67	9.75
age	-0.18	0.00	-57.01	0	-0.18	-0.17
engineSize	0.32	0.01	21.56	0	0.29	0.34

3.2 check model assumptions

autoplot will not knit, need ggplot or plot, or to add a wrapper

not great

ALL ASSUMPTIONS WORK WELL - I WOULD PROCEDE WITH: lm(data=fordfocus, LogPrice~age + engineSize)

$$\widehat{\text{LogPrice}} = \widehat{\beta}_0 + \widehat{\beta}_1 \text{age} + \widehat{\beta}_2 \text{engineSize}$$

add example of how this works if we have a made up scenario and how changes in explanatory variables affect the outcome variable

4 Conclusions

TBD