CAR PRICE PREDICTION

Maya Purandare and Troy Wu

Project Purpose and Goals

- This data is from Ward's Automotive Yearbook from **1985**
- We aim to use this dataset to find out the most influential factors in determining the car prices in the US market
- Useful for companies to accordingly manipulate the design of the cars, the business strategy etc. to meet certain price levels
- Research question: Which variables are significant in predicting the price of a car and which of those variables describe the car the best?
  - We used regression analysis to finalize a model that helps to best understand the pricing system of car market in the US

Dataset

- 205 Observations
- 25 Explanatory Variables (Increased to 37 for categorical variables)
- Dataset does not include car age or mileage so it is assumed this is the price of the car immediately after manufacturing

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	car_ID	symbolin	CarName	BrandLux	BrandPre	BrandSta	fueltypeG	aspiratio	doornum	carbodyC	carbodyH	carbodyS	carbodyV	drivewhe	drivewhe	enginelo	wheelbas	carlength	carwidth	carheight
2	1	3	alfa-rome	1	0	0	1	1	0	1	0	0	0	1	0	1	88.6	168.8	64.1	48.8
3	2	3	alfa-rome	1	0	0	1	1	0	1	0	0	0	1	0	1	88.6	168.8	64.1	48.8
4	3	1	alfa-rome	1	0	0	1	1	0	0	1	0	0	1	0	1	94.5	171.2	65.5	52.4
5	4	2	audi 100 l	1	0	0	1	1	1	0	0	1	0	0	1	1	99.8	176.6	66.2	54.3
6	5	2	audi 100ls	1	0	0	1	1	1	0	0	1	0	0	0	1	99.4	176.6	66.4	54.3
7	6	2	audi fox	1	0	0	1	1	0	0	0	1	0	0	1	1	99.8	177.3	66.3	53.1
8	7	1	audi 100ls	1	0	0	1	1	1	0	0	1	0	0	1	1	105.8	192.7	71.4	55.7
9	8	1	audi 5000	1	0	0	1	1	1	0	0	0	1	0	1	1	105.8	192.7	71.4	55.7
10	9	1	audi 4000	1	0	0	1	0	1	0	0	1	0	0	1	1	105.8	192.7	71.4	55.9
11	10	0	audi 5000	1	0	0	1	0	0	0	1	0	0	0	0	1	99.5	178.2	67.9	52

U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN
urbweigl	enginetyp	enginetyp	enginetyp	enginetyp	enginetyp	enginetyp	cylinderni	enginesiz	fuelsyste	fuelsyste	fuelsyste	boreratio	stroke	compres	horsepov	peakrpm	citympg	highwaym	price
2548	1	0	0	0	0	0	4	130	0	1	0	3.47	2.68	9	111	5000	21	27	13495
2548	1	0	0	0	0	0	4	130	0	1	0	3.47	2.68	9	111	5000	21	27	16500
2823	0	0	0	0	0	1	6	152	0	1	0	2.68	3.47	9	154	5000	19	26	16500
2337	0	1	0	0	0	0	4	109	0	1	0	3.19	3.4	10	102	5500	24	30	13950
2824	0	1	0	0	0	0	5	136	0	1	0	3.19	3.4	8	115	5500	18	22	17450
2507	0	1	0	0	0	0	5	136	0	1	0	3.19	3.4	8.5	110	5500	19	25	15250
2844	0	1	0	0	0	0	5	136	0	1	0	3.19	3.4	8.5	110	5500	19	25	17710
2954	0	1	0	0	0	0	5	136	0	1	0	3.19	3.4	8.5	110	5500	19	25	18920
3086	0	1	0	0	0	0	5	131	0	1	0	3.13	3.4	8.3	140	5500	17	20	23875
3053	0	1	0	0	0	0	5	131	0	1	0	3.13	3.4	7	160	5500	16	22	17859.2

Dataset

Feature Dictionary

SYM	Symboling	Int [-3, 3]
LUX	Brand - Luxury	0, 1
PREM	Brand - Premium	0, 1
STND	Brand - Standard	0,1
GAS	Fuel Type Gas (or Diesel)	0, 1
ASP	Aspiration Standard (or Turbo)	0, 1
4DOOR	Door Number (Four or Two)	0, 1
CONV	Car Body - Convertible	0, 1
HATCH	Car Body - Hatchback	0, 1
SEDAN	Car Body - Sedan	0, 1
WAGON	Car Body - Wagon	0, 1
RWD	Drive Wheel - Rwd	0, 1
FWD	Drive Wheel - Fwd	0, 1

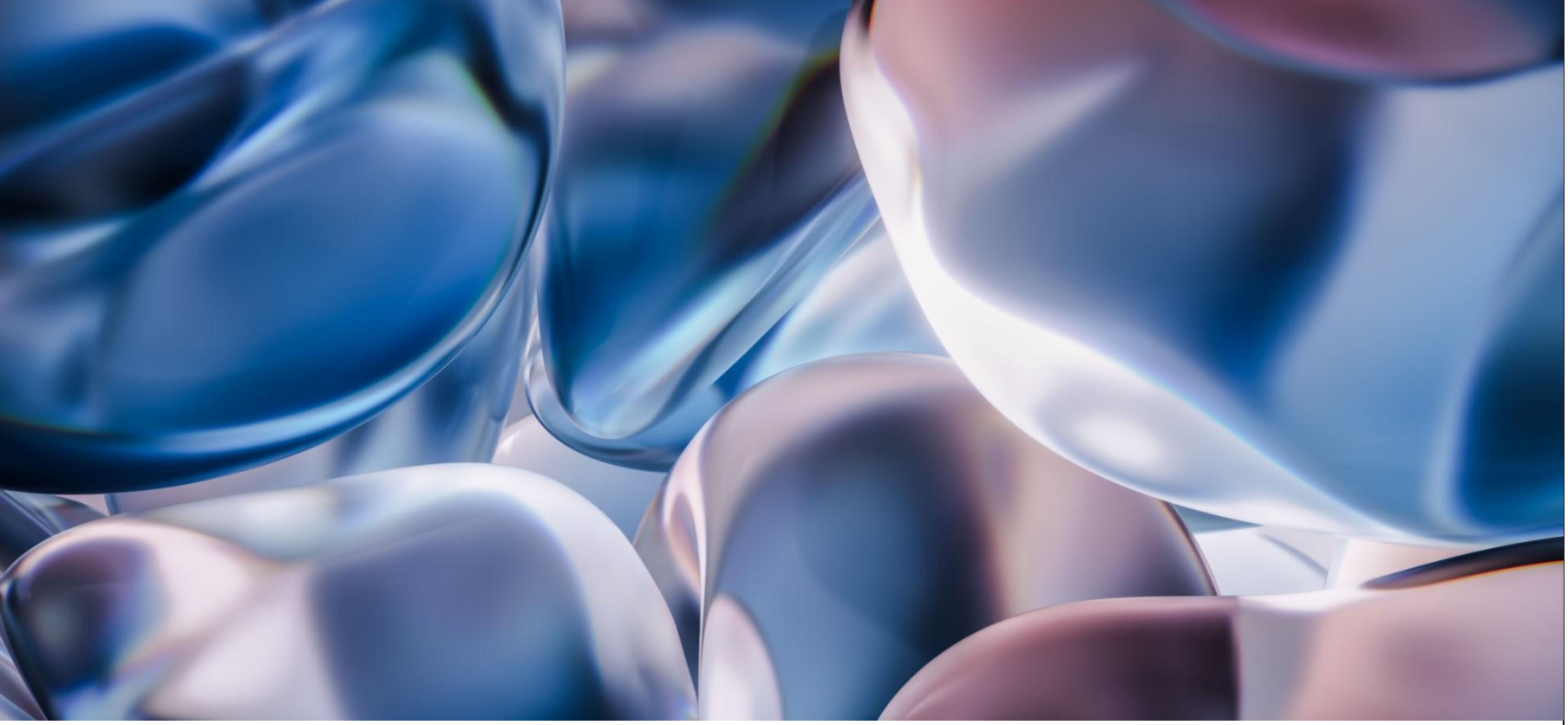
FRONT	Engine Location Front (or Back)	0, 1
WHEELBASE	Wheel Base	Numeric
LENGTH	Car Length	Numeric
WIDTH	Car Width	Numeric
HEIGHT	Car Height	Numeric
WEIGHT	Curb Weight	Numeric
DOHC	Engine Type - Dohc	0, 1
OHC	Engine Type - Ohc	0, 1
L	Engine Type - L	0, 1
ROTOR	Engine Type - Rotor	0, 1
OHCF	Engine Type - Ohcf	0, 1
OHCV	Engine Type - Ohcv	0, 1

CYL	Cylinder Number	3, 4, 5, 6, 8, 12
SIZE	Engine Size	Numeric
CARB	Fuel System - Carburetor	0, 1
ELEC	Fuel System - Electric	0, 1
BORERATIO	Boreratio of Car	Numeric
STROKE	Stroke Inside Engine	Numeric
COMPRATIO	Compression Ratio of Car	Numeric
HP	Horsepower	Numeric
RPM	Car Peak RPM	Numeric
CITYMPG	Mileage in City	Numeric
HWMPG	Mileage on Highway	Numeric
PRICE	Price of Car	Numeric

Response Variable: PRICE

Price of Car

Mean	13276.71
St Deviation	7988.85
Max	45400.00
Min	5118.00



FEATURE SELECTION

Model 1

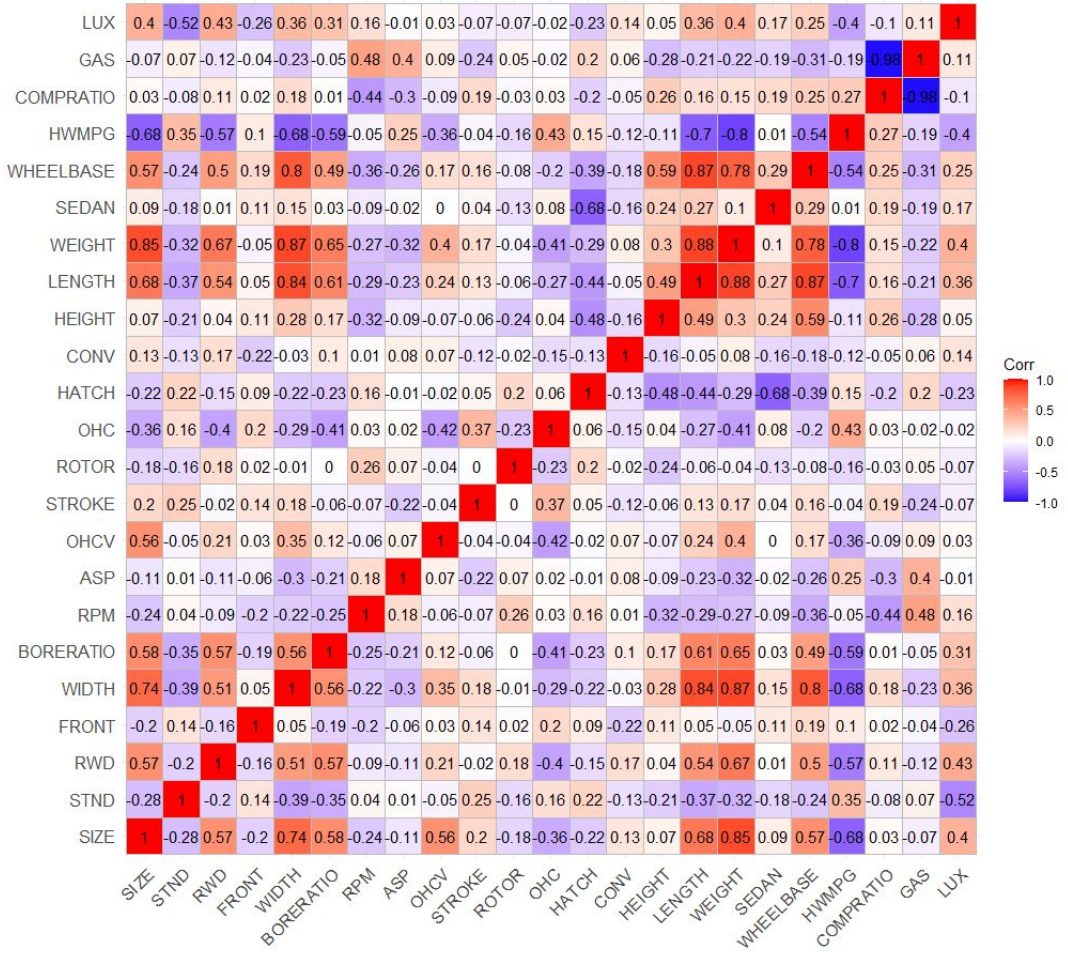
PRICE ~ SIZE + STND + RWD + FRONT + WIDTH + BORERATIO + RPM + ASP + OHCV + STROKE + ROTOR + OHC + HATCH + CONV + HEIGHT + LENGTH + WEIGHT + SEDAN + WHEELBASE + HWMPG + COMPRATIO + GAS + LUX

VIFs

SIZE	12.07
WIDTH	7.13
LENGTH	10.94
WEIGHT	22.97
WHEELBASE	9.23
HWMPG	7.45
COMPRATIO	79.76
GAS	87.32

Adj R^2: 0.9197
Multiple R^2: 0.9288

Correlation Plot



Model 2

PRICE ~ SIZE + FRONT + BORERATIO + RPM + ASP + STROKE + SEDAN + HATCH + CONV + WAGON + WEIGHT

VIFs

SIZE	5.68
FRONT	1.49
BORERATIO	1.96
RPM	1.23
ASP	1.44
STROKE	1.23
SEDAN	8.9
HATCH	8.39
CONV	1.78
WAGON	4.85
WEIGHT	6.56

Adj R^2: 0.8584
Multiple R^2: 0.866

Model 3

PRICE ~ SIZE + FRONT + BORERATIO + RPM + ASP + STROKE

VIFs

SIZE	1.76
FRONT	1.17
BORERATIO	1.73
RPM	1.19
ASP	1.14
STROKE	1.20

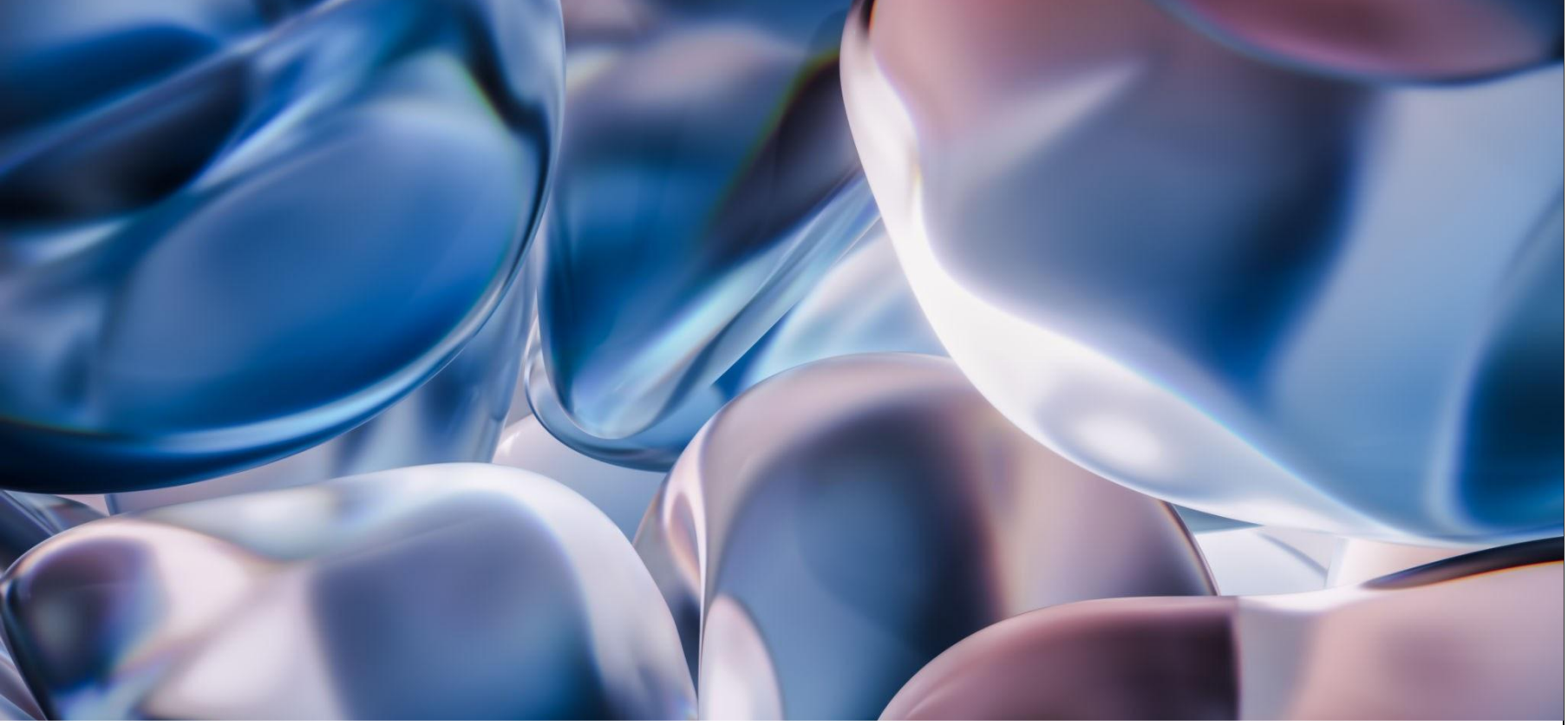
Adj R^2: 0.8149
Multiple R^2: 0.8203

Model 3 cont

PRICE ~ SIZE + FRONT + BORERATIO + RPM + ASP + STROKE

Summary Statistics

	Estimate	Std Error	t-value	p-value
INTERCEPT	-2442.87	6661.50	-0.37	0.71
SIZE	168.64	7.66	22.02	< 2e-16
FRONT	-7739.30	2159.36	-3.58	0.00043
BORERATIO	509.52	1169.19	0.436	**0.66**
RPM	2.15	0.551	3.90	0.00013
ASP	-2740.28	667.28	-4.11	5.87e-5
STROKE	-2618.18	841.45	-3.11	0.002136

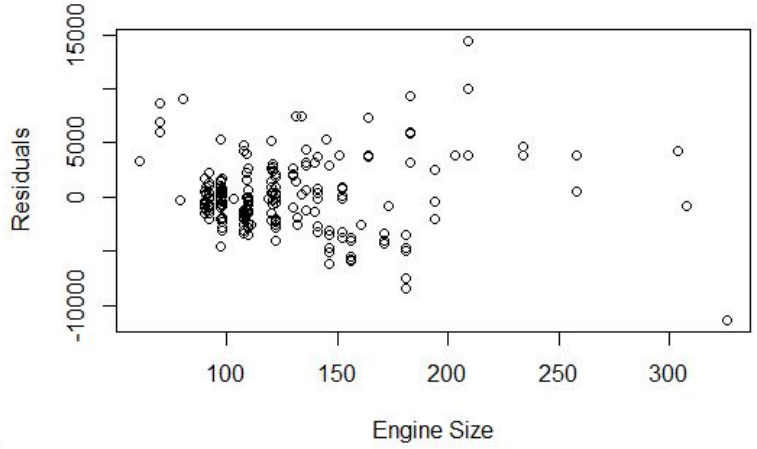


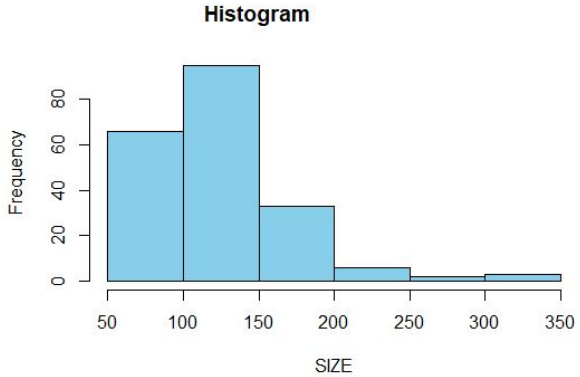
VARIABLES

Explanatory Variables: SIZE

Engine Size

Mean	126.91
St Deviation	41.64
Max	326.00
Min	61.00

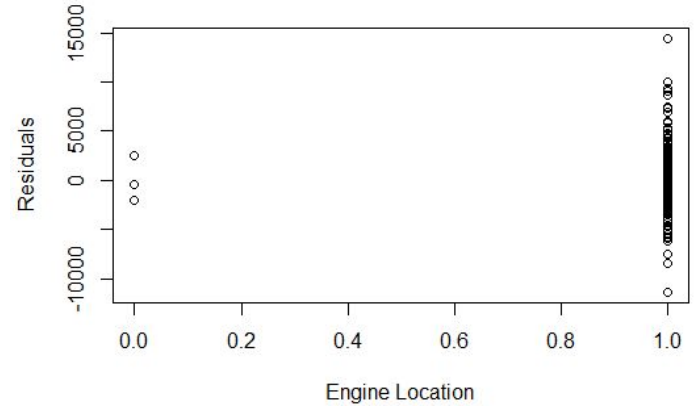


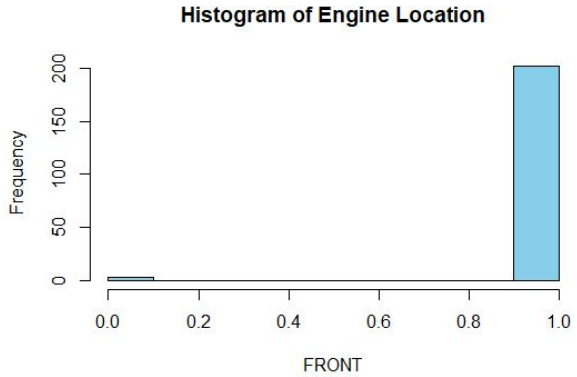


Explanatory Variables: FRONT

Engine Location - Front or Back

Mean	0.99
St Deviation	0.12
Max	1.00
Min	0.00

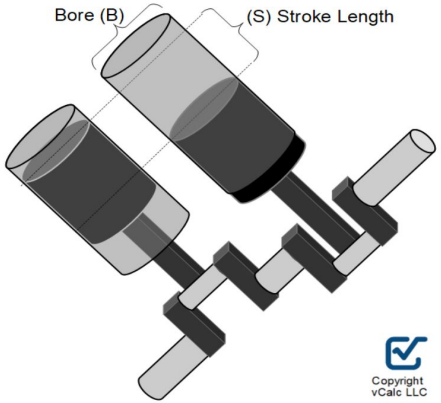


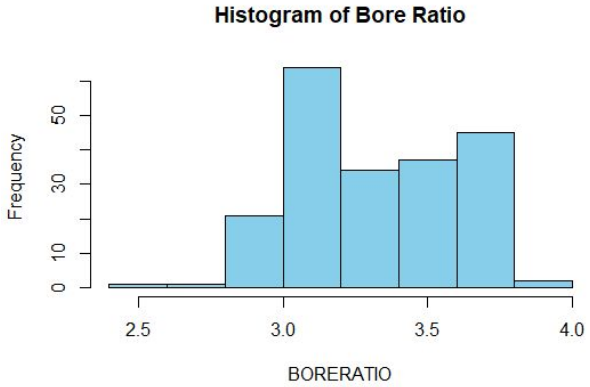


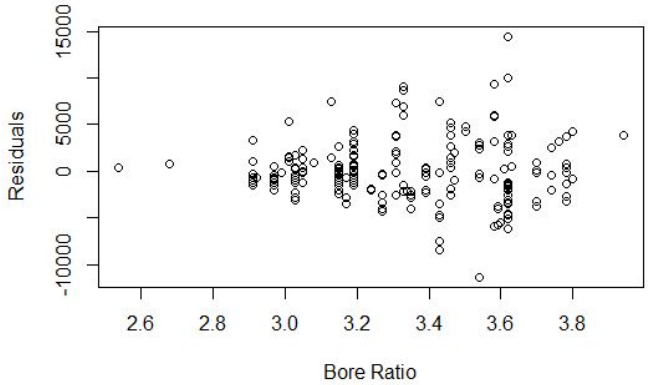
Explanatory Variables: BORERATIO

ratio of the cylinder bore (diameter) to the stroke (distance traveled by the piston)

Mean	3.33
St Deviation	0.27
Max	3.94
Min	2.54



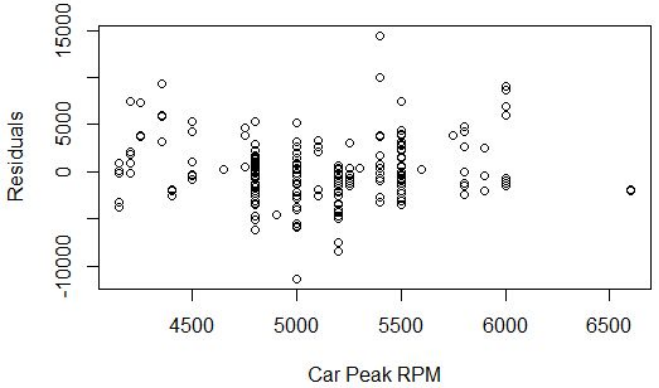


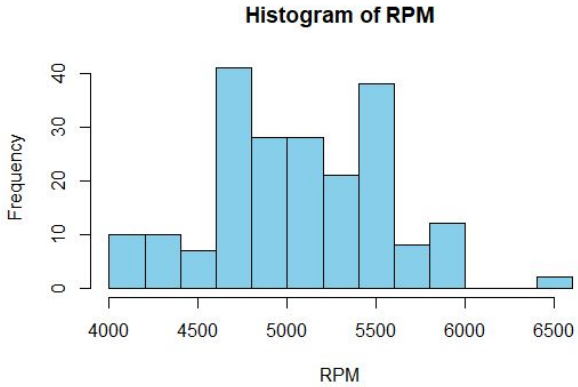


Explanatory Variables: RPM

Peak Revolutions Per Minute

Mean	5125.12
St Deviation	476.99
Max	6600.00
Min	4150.00

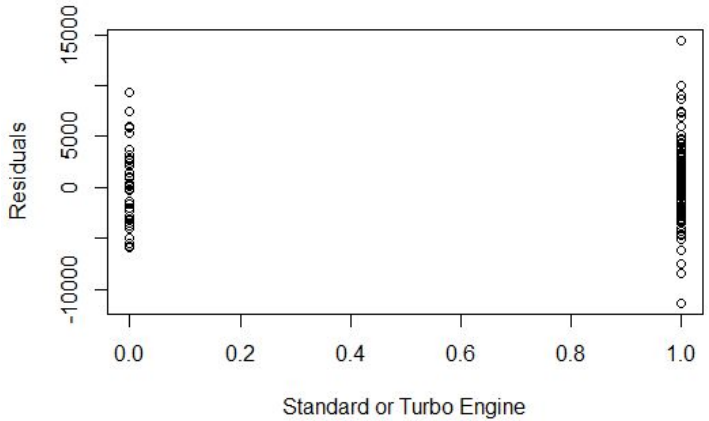


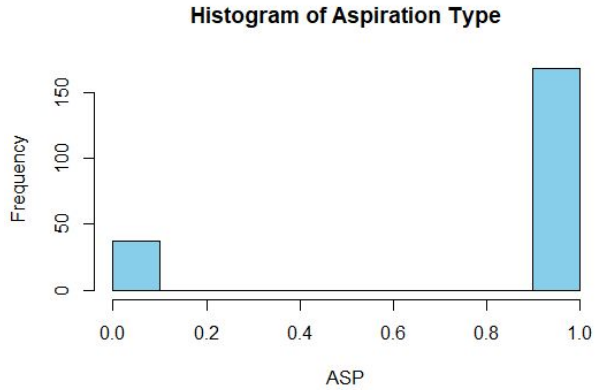


Explanatory Variables: ASP

Aspiration - Standard or Turbo

Mean	0.82
St Deviation	0.39
Max	1.00
Min	0.00

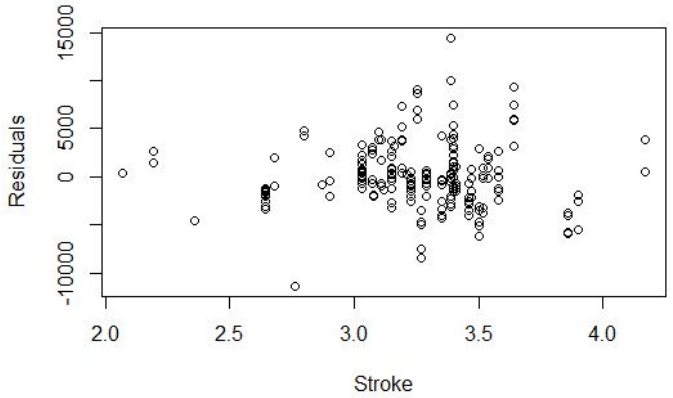


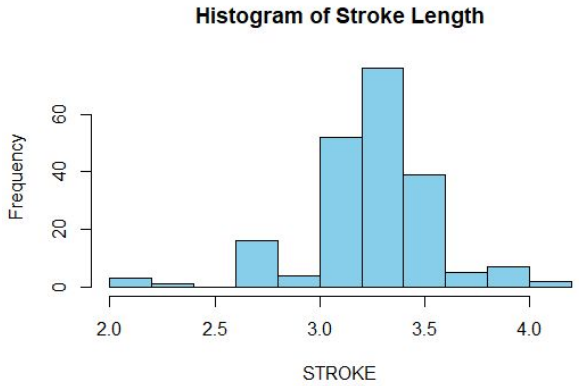


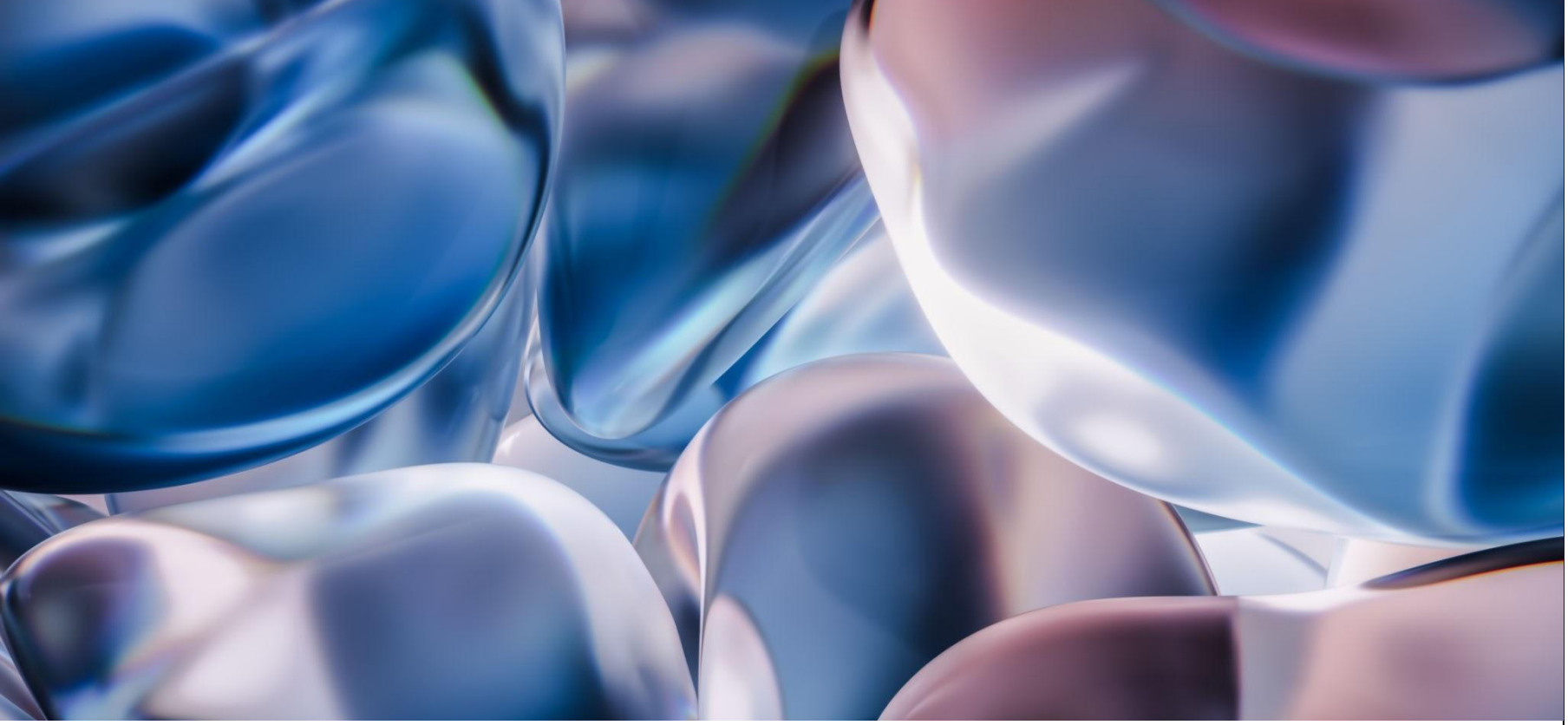
Explanatory Variables: STROKE

Stroke Length

Mean	3.26
St Deviation	0.31
Max	4.17
Min	2.07







GLOBAL TEST

Underlying Model

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2)$$

Regression Model

$$\hat{y} = -2442.87 + 168.64 x_1 + -7739.30 x_2 + 509.52 x_3 + 2.15 x_4 + -2740.28 x_5 + -2618.18 x_6$$

Interpretation Ex

For every one unit increase in size (x1), the mean price is estimated to increase $168.64, given all other variables held constant

Hypothesis Test

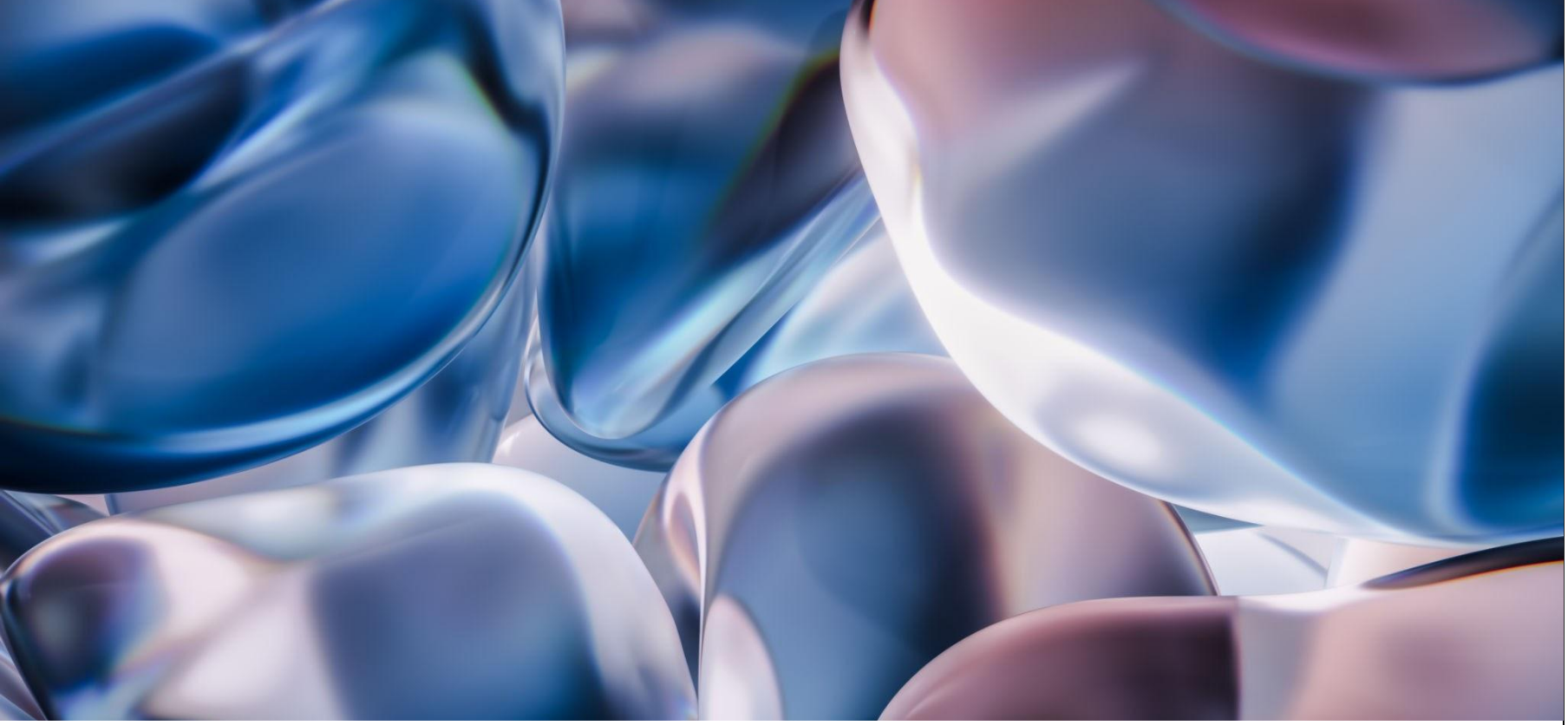
H0: $\beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$

Ha: At least one $\beta_i \neq 0$

TS ~ $F_{6, 198, 0.05} = 150.6$

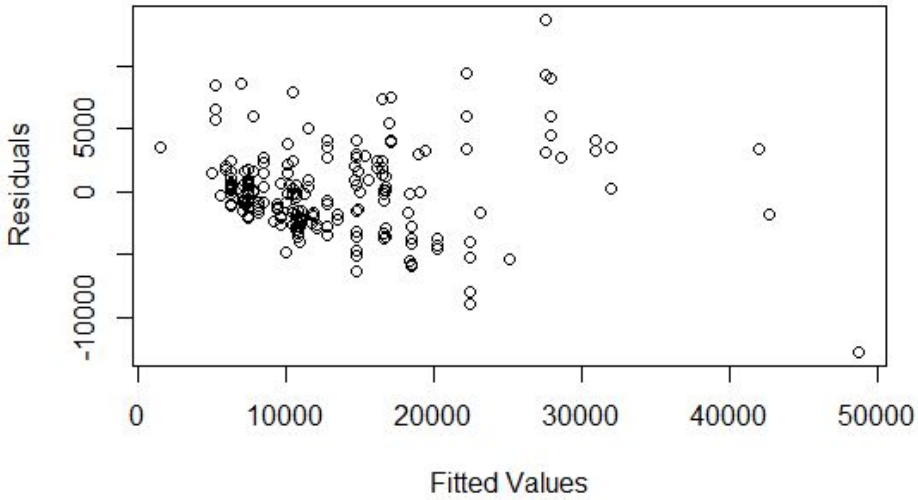
CV $F_{6, 198, 0.05} = 2.177$

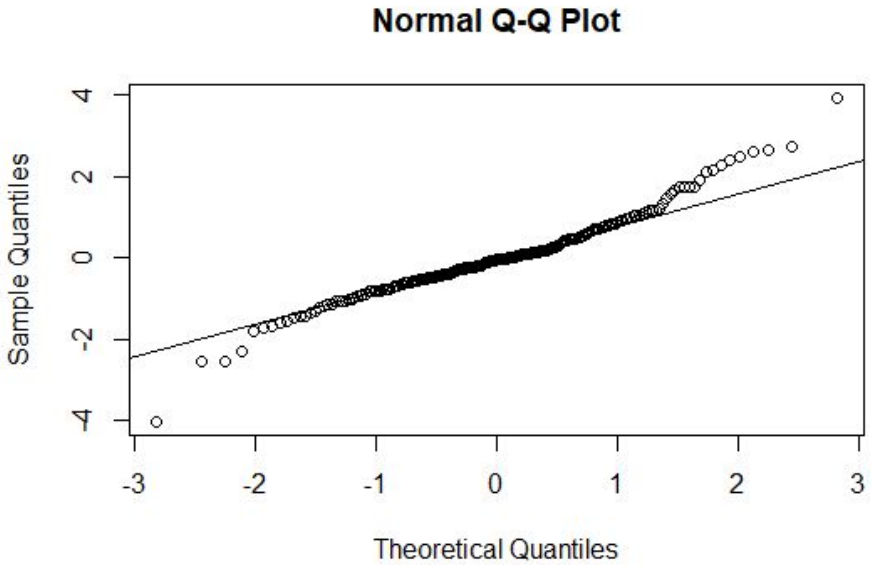
Since 150.6 > 2.177 , We reject the null hypothesis and can claim a linear relationship between variables and car price with a 5% significance level



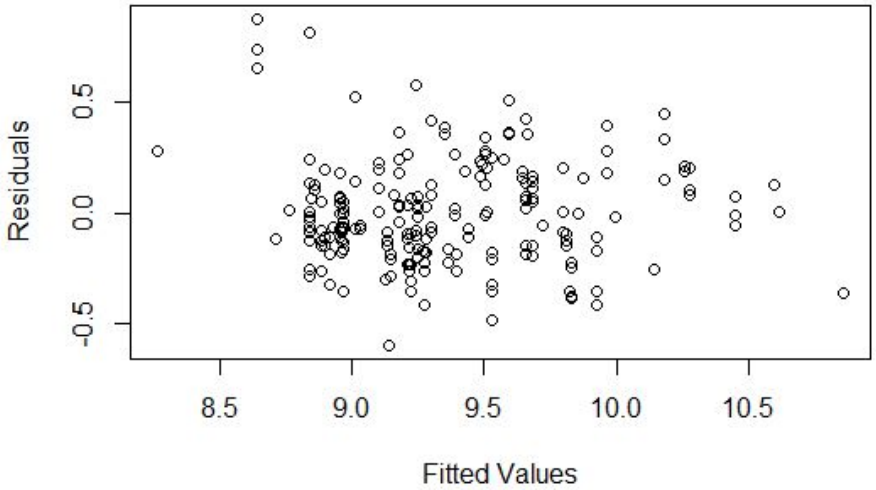
MODEL ADEQUACY CHECKING

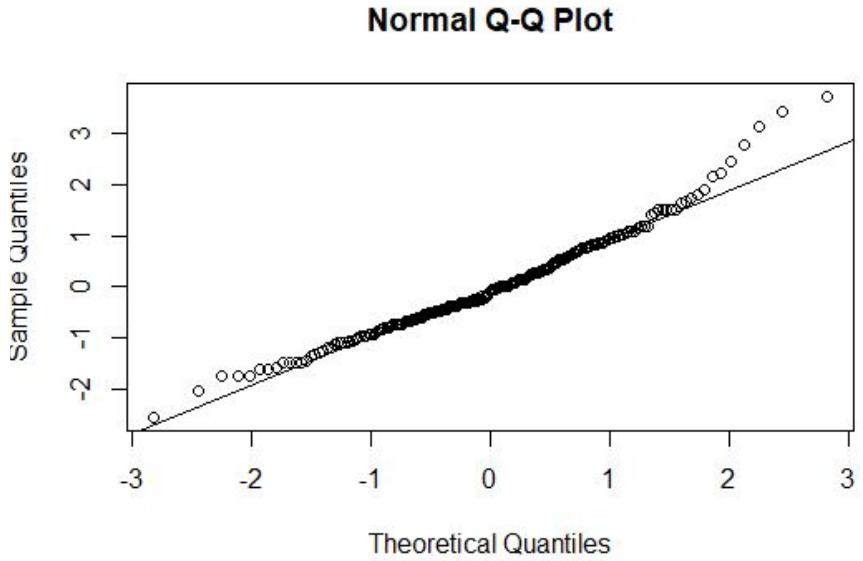
Residual v Fitted Graph has a funnel shape, not constant variance

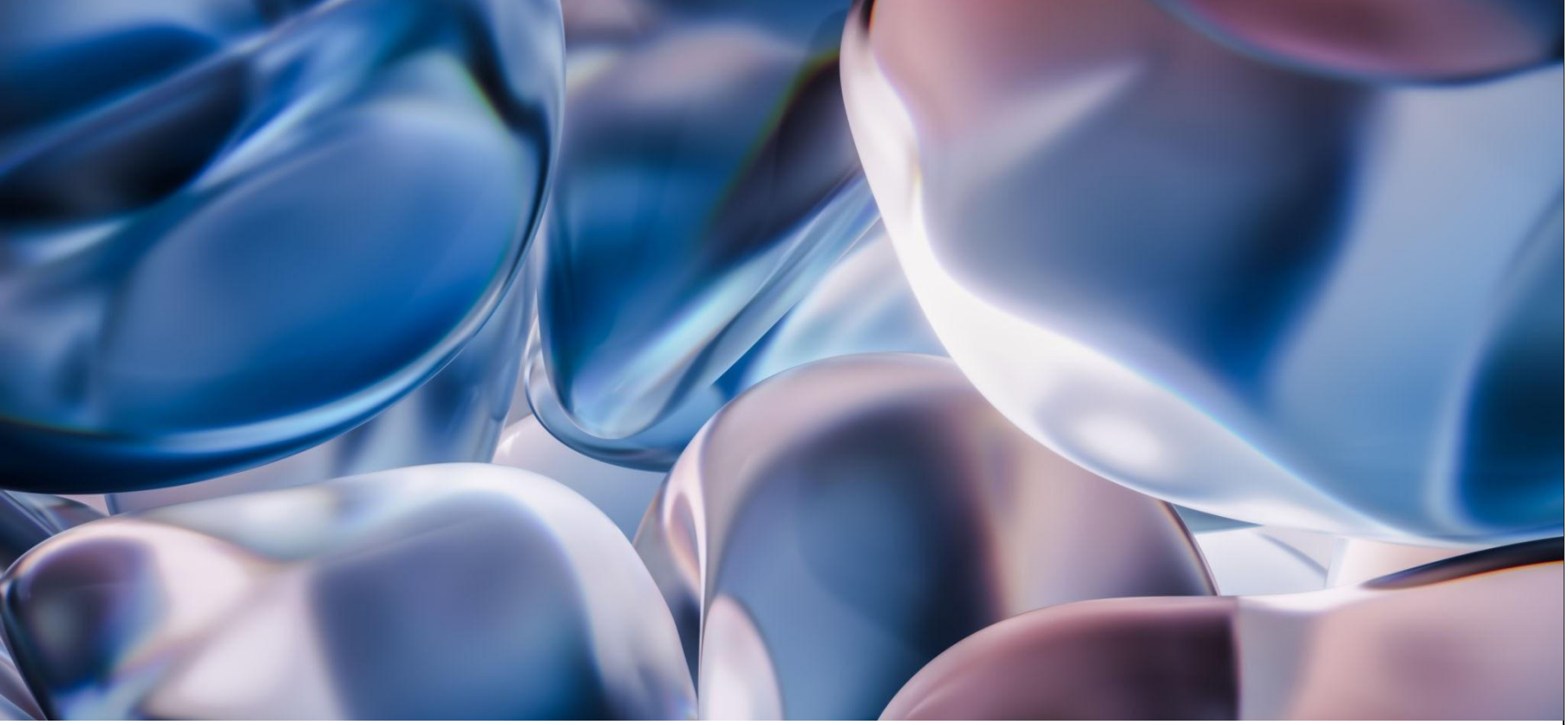




After performing Log transformation on Price and Engine Size, Bore Ratio, Peak RPM, and Stroke





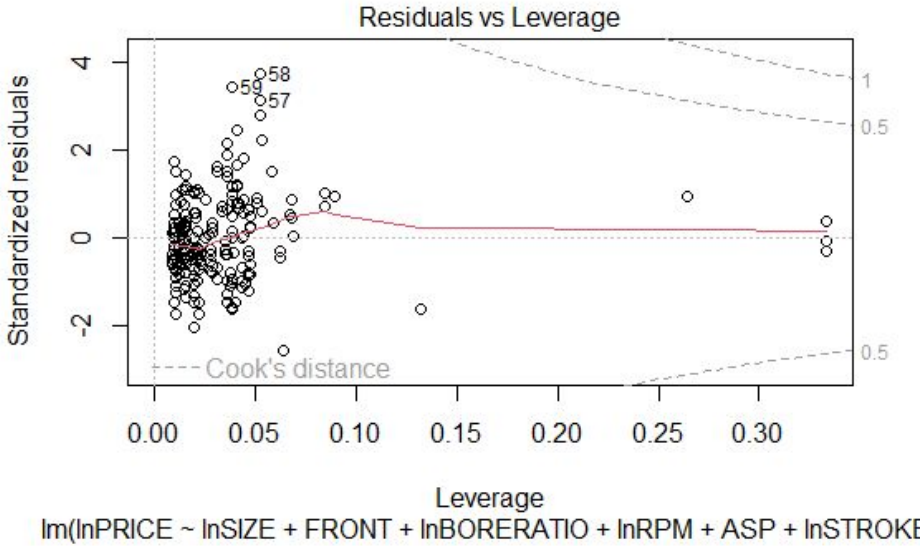


INFLUENCE AND LEVERAGE

Leverage

Studentized Residuals; Outliers: 57, 58, 59

Cook's D; Leverage: none
Df fits; Leverage: 9 15 50 56 57 58 59 67 135 139 193 204



Leverage cont

Why are these leverage points?

Obs 57, 58, and 59 all have abnormally small sized engines

Other observations likely showed up as leverage points because their engine stats were similar to those of other cars but they had much higher/ lower prices due to other factors not included in our model such as car size, body type, and brand.

Final Regression Model

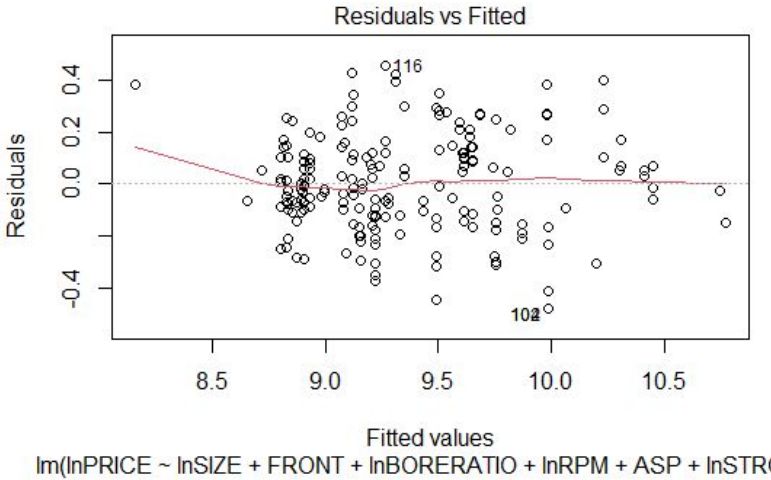
$$\ln\hat{y} = -4.04 + 1.73\log(x_1) + -0.14x_2 + 0.05\log(x_3) + 0.76\log(x_4) + -0.21x_5 + -0.90\log(x_6)$$

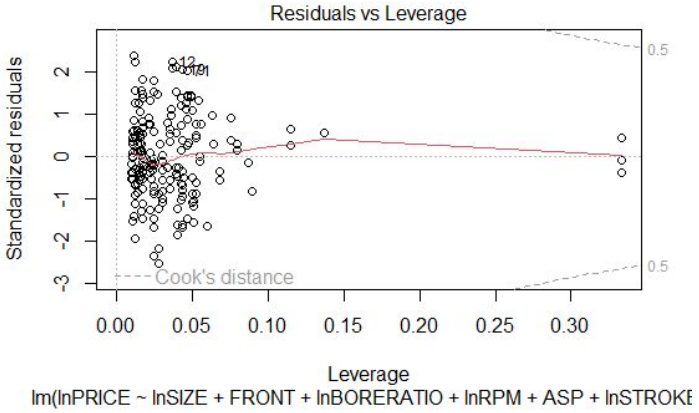
Interpretation Ex.

When a car has an engine in the front (x2), the predicted price decreases 13.06% given all other variables held constant

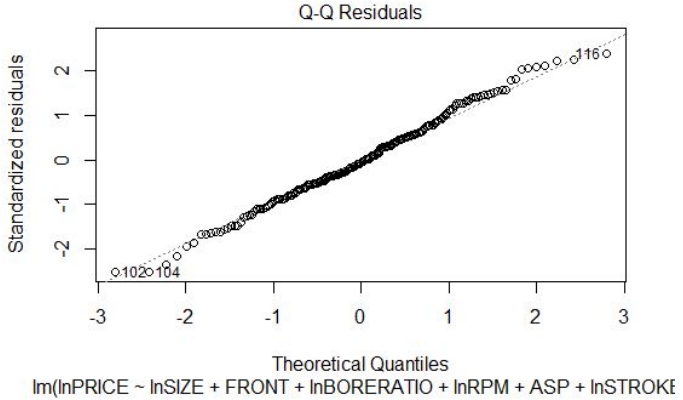
For every 1% increase in Bore Ratio (x3), the predicted price increases 5.13% given all other variables held constant

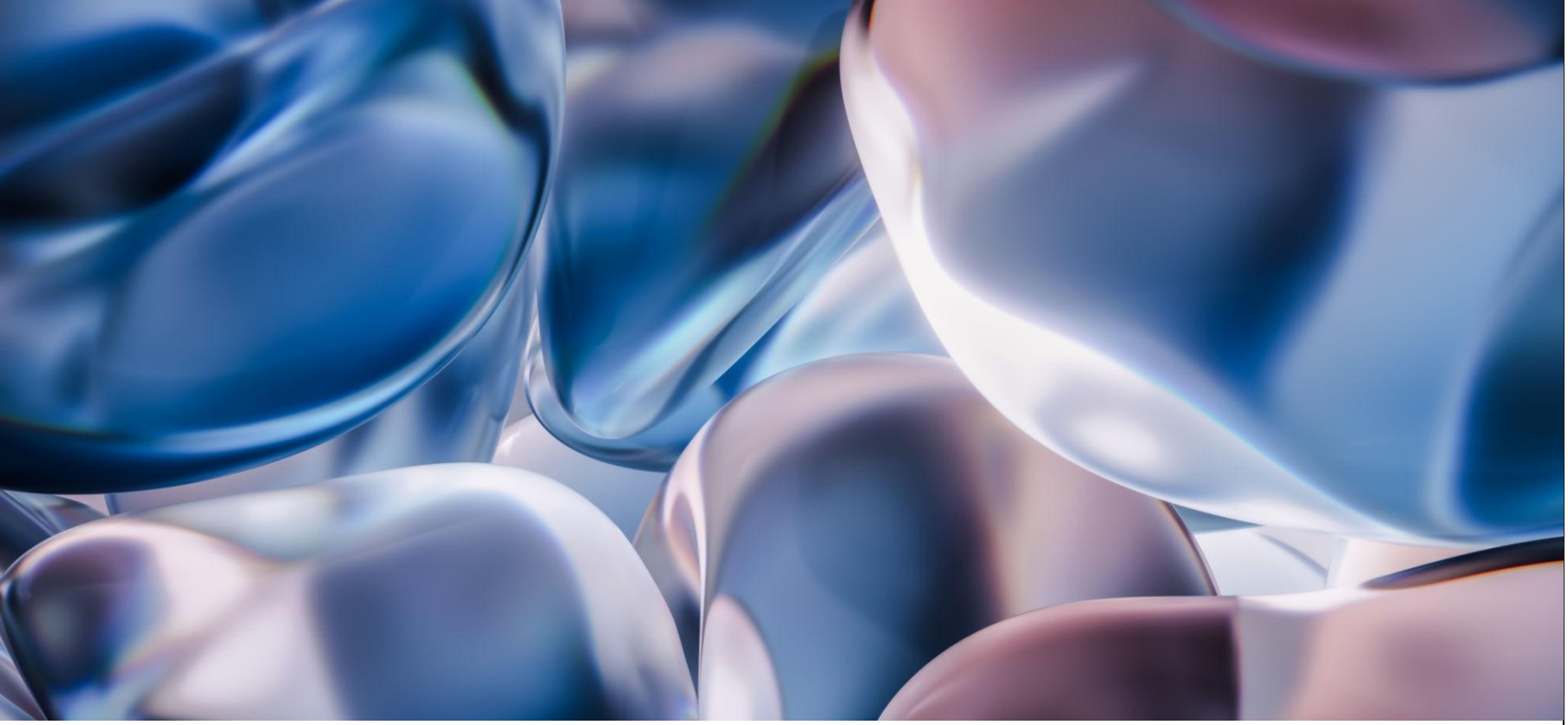
Final Regression Model cont





Adj R^2: 0.8519
Multiple R^2: 0.8566
85.66 % Variability in price can be explained by the Regression model. 14.34% left in error





CONCLUSION

Final Thoughts

1. Through our regression analysis, we find out that the size of the engine is influential to the final car prices due to its correlation between size and peak performance of engine
2. While size of engine is one of the deciding factor of the car prices, the type of the car or engine doesn't have as high of a influence on the car prices and thus wasn't included in the final regression model
3. Expectedly, performance of the vehicle is correlational to the car prices. This further demonstrates with bigger cars, bigger engine is installed, and therefore have higher prices
4. Surprisingly, the location of the engine also plays a role in determining the prices of the cars in the US market

Thank You

Any Questions?