Refining the Regression Model: Exploring the Automobile Dataset, Variable Selection, Transformation, Validation, and Car Price Prediction

Quang Duy Tran[1], Maiqi Zhang[1], Baixue Zhang[2]

[1]Department of Computer Science, Data Science, San José State University

[2]Department of Mathematics and Statistics, Statistics, San José State University

SJSU SAN JOSÉ STATE UNIVERSITY

Overview

- Research Objectives/Questions
- Data Exploration
- Variable Selection
- Baseline Model
- Outliers - Leverage points - Influential points
- Transformation Model
- Data Validation
- Inference

Research Objectives/Questions

1. Which predictors contribute significantly to the price of a brand-new car?

2. How well can we predict the price of a brand-new car on the smaller subset of predictors?

Data Exploration

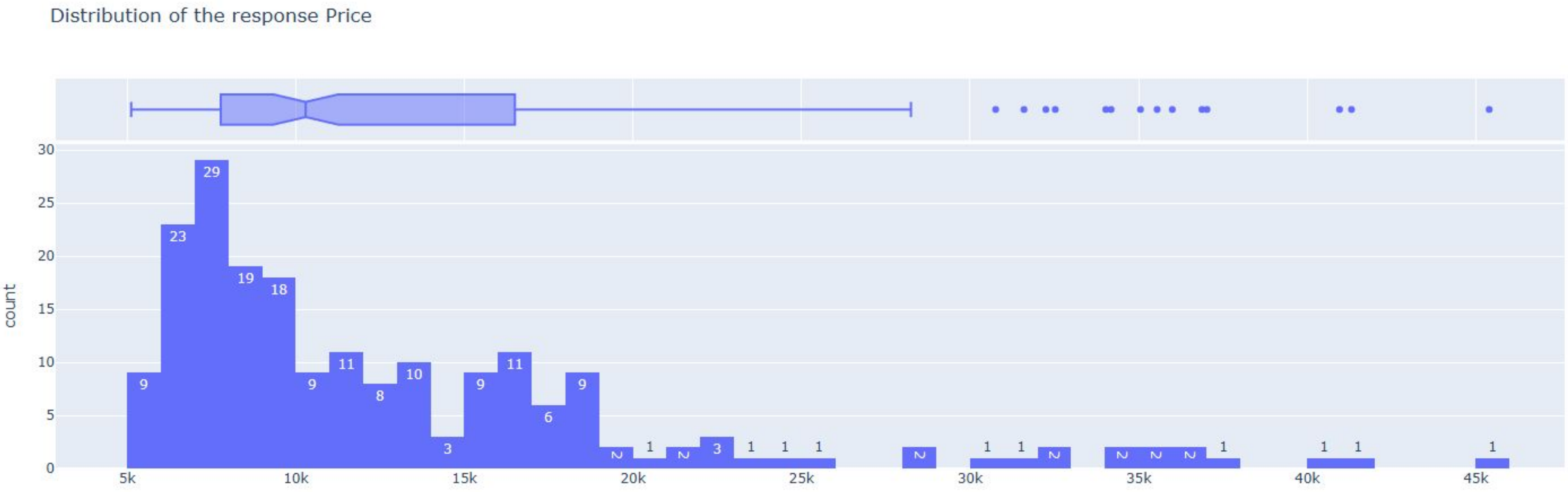
- 205 observations
- 26 variables:
    - 10 categorical
    - 16 continuous
- Response = Price
- No duplicated observations
- 46 NaN observations (22%)

```
-------------------- Dataset Info --------------------
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 205 entries, 0 to 204
Data columns (total 26 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   symboling          205 non-null    int64
 1   normalized_losses  164 non-null    float64
 2   make               205 non-null    object
 3   fuel_type          205 non-null    object
 4   aspiration         205 non-null    object
 5   num_doors          203 non-null    object
 6   body_style         205 non-null    object
 7   drive_wheels       205 non-null    object
 8   engine_location    205 non-null    object
 9   wheel_base         205 non-null    float64
 10  length             205 non-null    float64
 11  width              205 non-null    float64
 12  height             205 non-null    float64
 13  curb_weight        205 non-null    int64
 14  engine_type        205 non-null    object
 15  num_cylinders      205 non-null    object
 16  engine_size        205 non-null    int64
 17  fuel_system        205 non-null    object
 18  bore               201 non-null    float64
 19  stroke             201 non-null    float64
 20  compression_ratio  205 non-null    float64
 21  horsepower         203 non-null    float64
 22  peak_rpm           203 non-null    float64
 23  city_mpg           205 non-null    int64
 24  highway_mpg        205 non-null    int64
 25  price              201 non-null    float64
```

SJSU SAN JOSÉ STATE UNIVERSITY

Distribution of response Price

- Most cars are priced between $5,000 and $15,000.
- The distribution is right-skewed, with a long tail toward higher prices

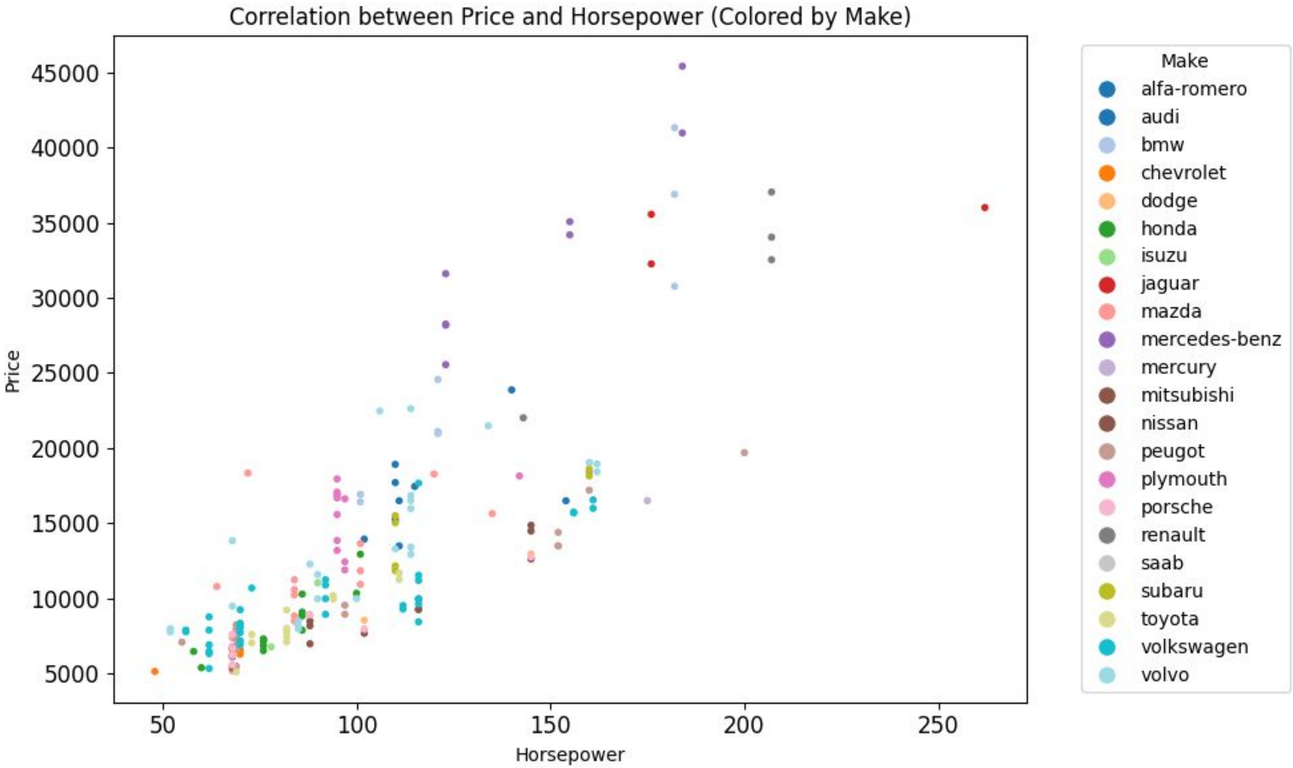


SJSU SAN JOSÉ STATE UNIVERSITY

Correlation between Price and Horsepower

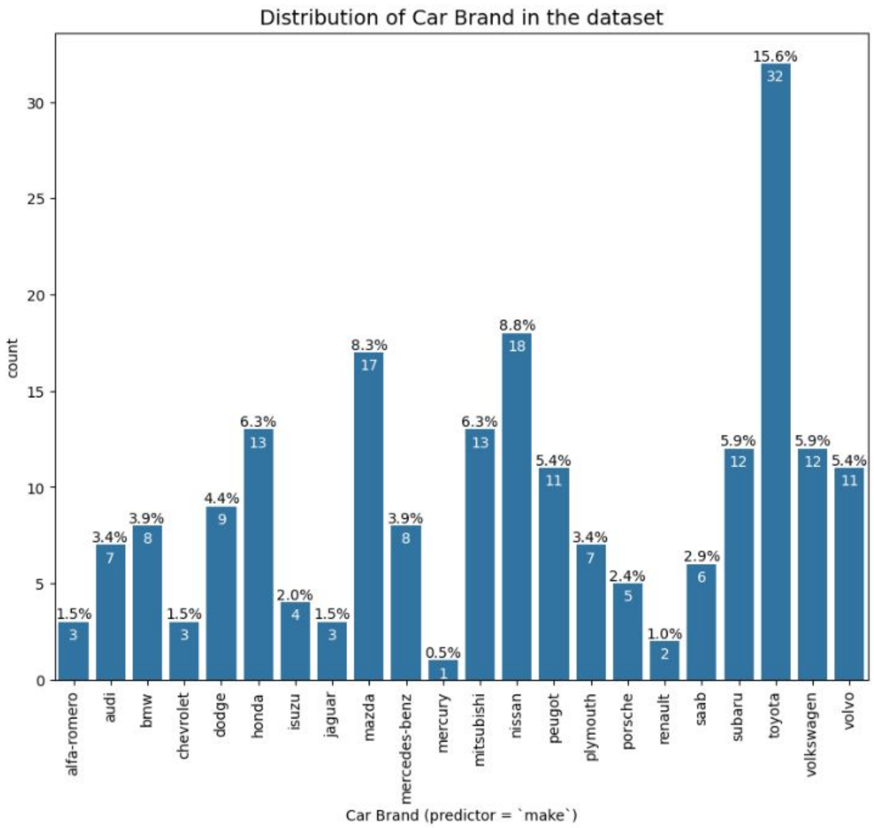
Toyota & Honda have low horsepower and are generally cheap

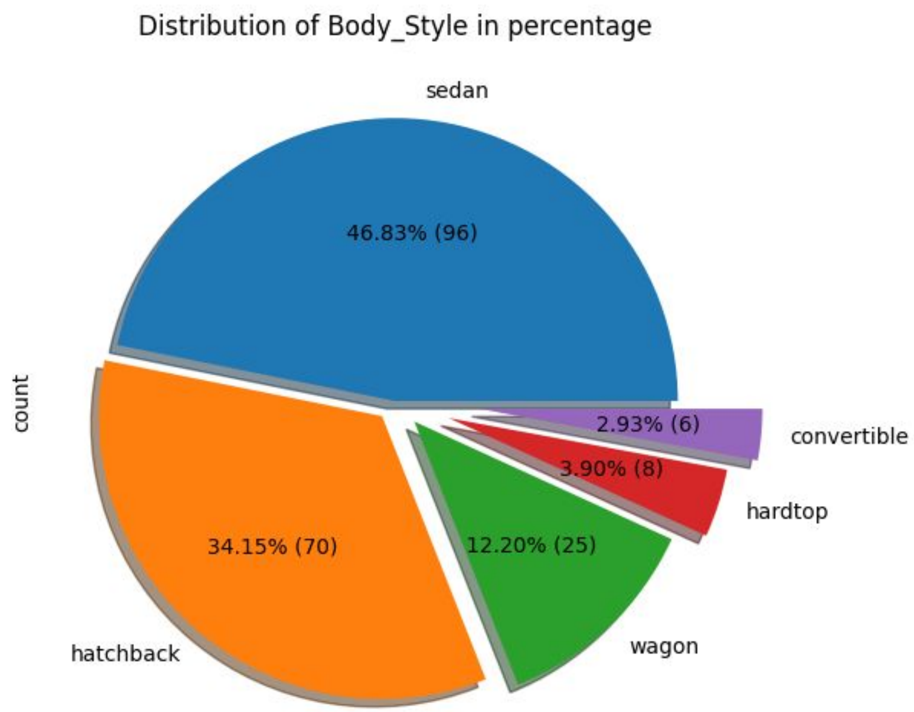
Mercedes-benz has moderate horsepower but expensive



SJSU SAN JOSÉ STATE UNIVERSITY

Data Exploration



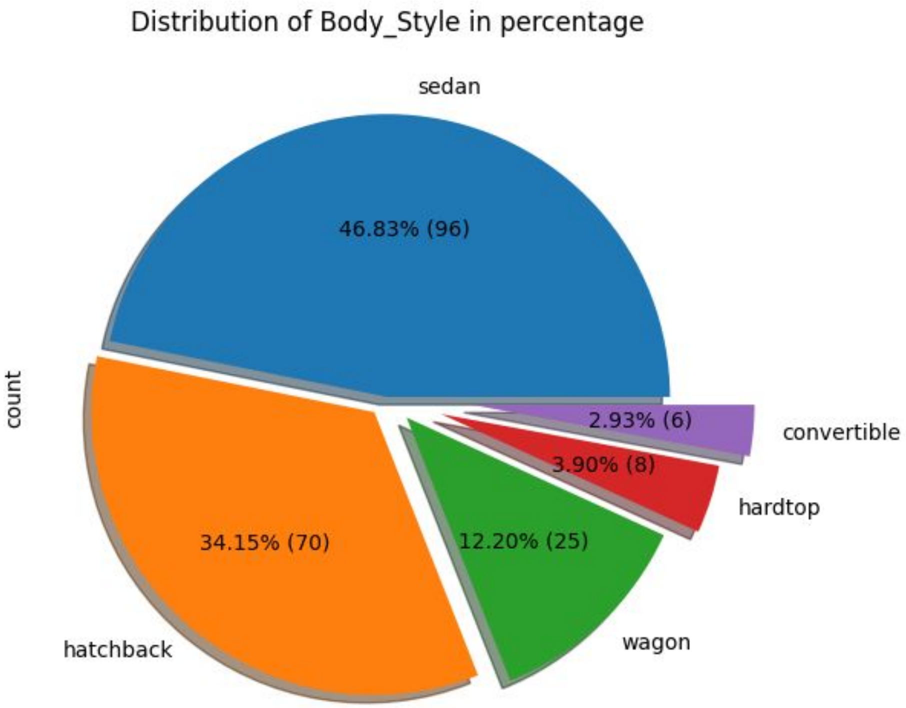


Data Exploration

Sedan

Hatchback

Wagon



SJSU SAN JOSÉ STATE UNIVERSITY

Research Objectives/Questions

1. Which predictors contribute significantly to the price of a brand-new car?

2. How well can we predict the price of a brand-new car on the smaller subset of predictors?

Variable selection

Outline

- Visual inspection
  - missing y
  - one level categorical variables
  - variables have a lot of missing values
- Address multicollinearity.
- R variables selection.
  - Forward, backward, stepwise.
  - Exhaustive selection(Impractical)
- Final variables

Visual Inspection — three problems spotted

- Missing y(price)→remove→201 observations left.
- Engine_location
  - 190 + front
  - 3 rear, but expensive 
- Normalized_losses
  - 41 missing values, ~20% of the observations

Address multicollinearity problems

VIF :

curb_weight: 16.047395

City_mpg: 26.424588

HWY_mpg: 24.428984

Drop curb_weight

One variable for car size

One variable for engine attributes

	length	width	wheel_ base	curb_w eight	highwa y_mpg	horse power	engine size	number of cylinder
length	1	0.84	0.87	0.87				
width	0.84	1	0.816	0.87				
wheel_ base	0.87	0.816	1	0.81				
curb_ weight	0.87	0.87	0.81	1	-0.813	<0.8	0.89	<0.8
Highway_ mpg				-0.813	1	-0.83	<0.8	<0.8
horse power				<0.8	-0.83	1	0.81	<0.8
engine size				0.89	<0.8	0.81	1	0.848
number of cylinder				<0.8	<0.8	<0.8	0.848	1

Variable selection by R

Issue: engine_location and Normalized_losses can not be fit together!

Solution: Consider two cases.

- Case1: Include engine_location and exclude normalized_losses. Later, include normalized_losses if engine_location is excluded. However, since engine_location was never dropped, normalized_losses could not be included.
- Case2: Include normalized_losses and exclude engine_location. Later, include engine_location if normalized_losses is excluded. Eventually, R excluded normalized_losses, allowing us to include engine_location.

Case 1: With engine_location.

Forward Variable selection

alpha	0.05/0.1	0.05/0.1	0.05/0.1	0.05/0.1	0.05/0.1	0.05/0.1	0.05/0.1	0.05/0.1	0.05/0.1	0.05/0.1	0.1	0.1
add	engine_size	make	curb_weight	engine_location	width	peak_rpm	aspiration	num_of_cylinders	engine_type	stroke	hwy_mpg	body_style
P	2.2 *10^-16	2.2*10^-6	1.66* 10 ^-6	0.00013	0.00535	0.00846	0.024	0.0014	0.01	0.025	0.053	0.097

Backward Variable selection

alpha	0.1/0.05	0.1/0.05	0.1/0.05	0.1/0.05	0.1/0.05	0.1/0.05	0.05	0.05	0.05	0.05	0.05	0.05
drop	number_of_doors	city_mpg	horse_power	drive_wheels	stoke	symboling	engine_type	hwy_mpg	compression_ratio	fuel_system	fuel_type	no
p	0.97	0.84	0.777	0.56	0.17	0.13	0.094	0.056	0.066	0.103	o.75	<0.05

__ :picked and __tossed :0.05 and 0.1.   __TBD :0.1 and __TBD 0.05

- Stepwise: aligns with forward and backward variable selection

Alpha (Drop or add)	0.05, add	0.1, drop	0.05, add	0.1, drop	0.05, add	0.1, drop	0.05, add
variables	engine_size	0	make	0	curb_weight	0	engine_location
p	2.2 *10^-16		2.2*10^-6		1.65* 10^-6		0.00013
0.1, drop	0.05, add	0.1, drop	0.05	0.1, drop	0.05, add	0.1, drop	0.05, add
0	width	0	peak_rmp	0	aspiration	0	# of cylinders
	0.00535		0.00846		0.00242	0	0.0014
0.1, drop	0.05, add	0.1, drop	0.05, add	0.1, drop	0.1, add	0.1, drop	0.1, add
0	engine_type	0	stoke	0	hwy_mpg	0	body_style
	0.00993		0.0025		0.0534		0.096

- R variable selection + multicollinearity elimination

Alpha = 0.05

Alpha = 0.1

```
Model 1: price ~ engine_size + make + width + engine_location + peak_rpm +
    aspiration + engine_type + stroke + body_style + height +
    bore
Model 2: price ~ engine_size + make + width + engine_location + peak_rpm +
    aspiration + engine_type + stroke + body_style + height +
    bore + compression_ratio + fuel_system + fuel_type
  Res.Df       RSS Df Sum of Sq      F Pr(>F)
1    159 589523454
2    152 546850551  7  42672902 1.6945 0.1143
```

Extra variables
Not significant

Conclusion of Case 1: with engine_locaiton

Second time Stepwise variable selection.

alpha	0.05/0.1 drop	0.05/0.1 add	0.05/0.1 drop	0.05/0.1 add	0.05 drop	end
Add or drop	Height	0	stroke	0	engine_type	
P	0.79		0.56		0.058	

```
Analysis of Variance Table

Model 1: price ~ engine_size + make + width + engine_location + peak_rpm +
    aspiration + engine_type + body_style + bore
Model 2: price ~ engine_size + make + width + engine_location + peak_rpm +
    aspiration + body_style + bore
  Res.Df       RSS Df Sum of Sq      F  Pr(>F)
1    161 591024164
2    164 619087193 -3 -28063028 2.5482 0.05777 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Case2: Without engine_location. __ and __ :0.05 and 0.1, __ :0.1 and __ 0.05

alpha	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.0 5	0.1	0.1	0.1
add	curb_weight(with F 748)	make	height	aspiration	body_style	wheel_base	num_of_cylinders	drive_wheels	engine_type	compression_ratio	length	number_of_doors	fuel_system	engine_size
P	2.2* 10 ^-16	2.2*10^-6	3.2*10 ^-10	0.00015	1.7 *10-5	0.000169	5.6 * 10^-5	0.0014	0.00495	0.0134	0.047	0.052	0.058	0.037

alpha	0.1/0.05	0.1/0.05	0.1/0.05	0.1/0.05	0.1/0.05	0.1/0.05	0.1/0.05	/0.1/0.05	0.1/0.05	0.1/0.05	0.1/0.05	0.1/0.05	0.1 /0.05	0.1/0.05	0.1/0.05
drop	symboling	engine_size	stoke	peak_rpm	normalized_losses	# of doors	city_mpg	aspiration	drive_wheels	hwy_mpg	compression_ratio	fuel_system	bore	body_style	fuel_type
p	0.98	0.59	0.508	0.47	0.426	0.97	0.95	0.83	0.81	0.24	o.16	0.15	0.19	0.145	0.1004

__ :picked and __ tossed :0.05 and 0.1. __ TBD :0.1 and __ TBD 0.05

- Stepwise: aligns with forward and backward variable selection

alpha	0.05/0.1 add	0.05/0.1 drop	0.05/0.1 add	0.05/0.1 drop	0.05/0.1 add	0.05/0.1 drop	0.05/0.1 add	0.05/0.1 drop	0.05/0.1 add	0.05/0.1 drop	end
add	curb_weight (with F 748)	0	make	0	height	0	aspiration	aspiration	aspiration	aspiration	
P	2.2* 10 ^-16		2.2*10^-6		3.2 *10^-10		0.00015	0.96	0.00015	0.96	

- R variable selection + multicollinearity elimination

Compare model with alpha 0.05 and alpha 0.1

```
Analysis of Variance Table

Model 1: price ~ make + aspiration + body_style + drive_wheels + engine_location +
    wheel_base + height + engine_type + num_of_cylinders + compression_ratio +
    horsepower
Model 2: price ~ make + aspiration + num_of_doors + body_style + drive_wheels +
    engine_location + wheel_base + height + engine_type + num_of_cylinders +
    fuel_system + compression_ratio + horsepower
  Res.Df       RSS Df Sum of Sq      F Pr(>F)
1    156 605455638
2    148 576265522  8  29190116 0.9371 0.4879
```

Conclusion of Case 2: without engine_locaiton

alpha	0.05/0.1 drop	0.05/0.1 add	0.05 drop	0.05 add	0.05 drop	end
add	aspiration	0	drive_wheels	0	height	
P	0.79		0.073		0.19	

Compare model with alpha 0.05 and alpha 0.1

```
Analysis of Variance Table

Model 1: price ~ make + body_style + engine_location + wheel_base + engine_type +
    num_of_cylinders + compression_ratio + horsepower
Model 2: price ~ make + body_style + drive_wheels + engine_location +
    wheel_base + height + engine_type + num_of_cylinders + compression_ratio +
    horsepower
  Res.Df       RSS Df Sum of Sq      F Pr(>F)
1    162 633465830
2    159 606324718  3  27141112 2.3725 0.0724 .
---
```

Exhaustive search

When I start with an exhaustive search with R, it shows that R encountered a fatal error, R session restart.

Theoretically there are $\binom{22}{1}+\binom{22}{2}+\binom{22}{3}+\binom{22}{4}+\binom{22}{5}+\binom{22}{6}\cdots+\binom{22}{22}$ subsets.

Skip exhaustive search.

Variable selection final result

Start with including engine_location and exclude normalized_losses:

Model 1: price ~ engine_size + make + width + engine_location + peak_rpm + aspiration + body_style + bore (4+4)

Start with including normalized_losses and exclude engine_location:

Model2 :price ~ make + body_style + engine_location + wheel_base + engine_type + num_of_cylinders + compression_ratio + horsepower (3 +5)

____ means continuous ____ means categorical

Research Objectives/Questions

1. Which predictors contribute significantly to the price of a brand-new car?

Model 1: price ~ engine_size + make + width + engine_location + peak_rpm + aspiration + body_style + bore

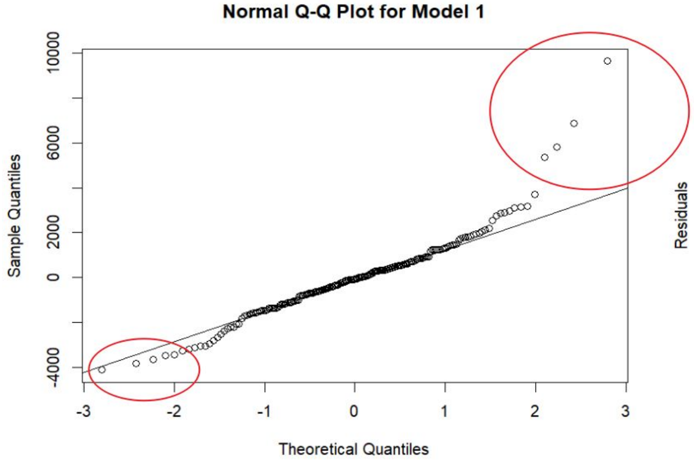
Model 2: price ~ make + body_style + engine_location + wheel_base + engine_type + num_of_cylinders + compression_ratio + horsepower

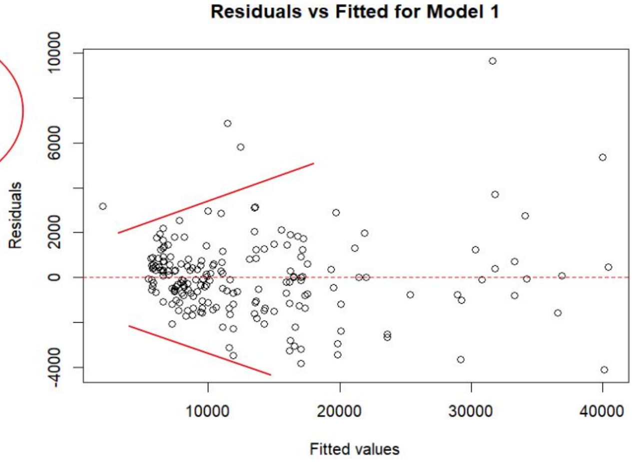
2. How well can we predict the price of a brand-new car on the smaller subset of predictors?

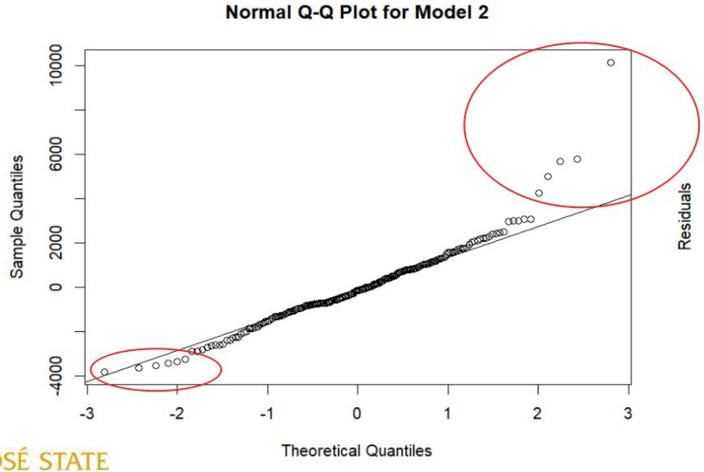
SJSU SAN JOSÉ STATE UNIVERSITY

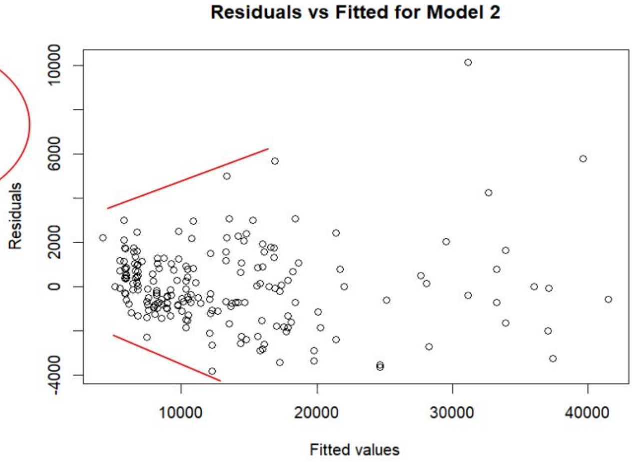
Model Adequacy and Reliability Check

1. QQ-Plot and Residual Plot
2. Outliers Detection
3. Leverage Point Detection
4. Influential Point Detection

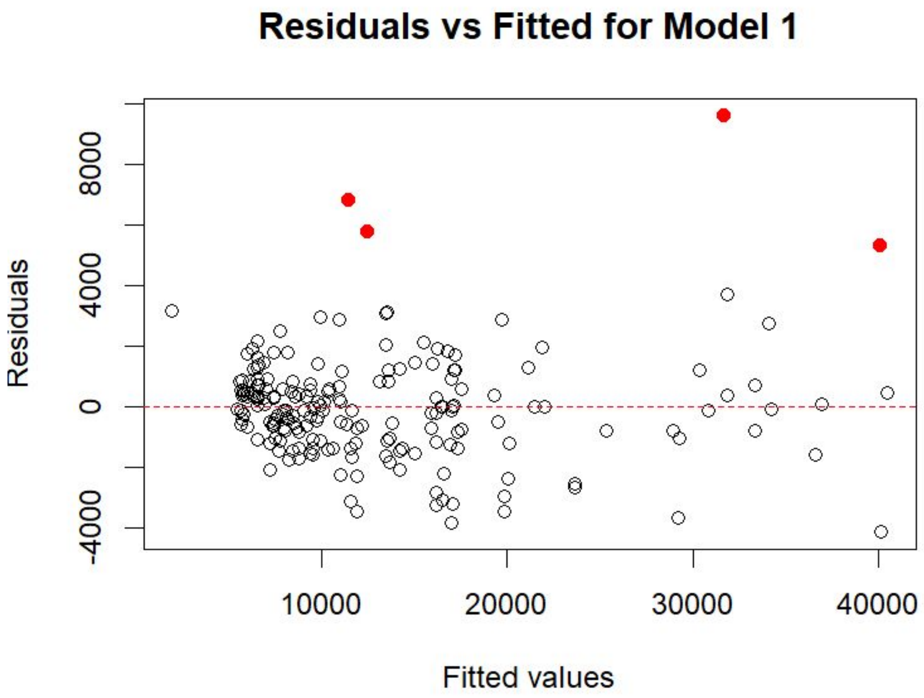

Normal Q-Q Plot for Model 1
Sample Quantiles
-4000
0
2000
6000
10000
-3
-2
-1
0
1
2
3
Theoretical Quantiles

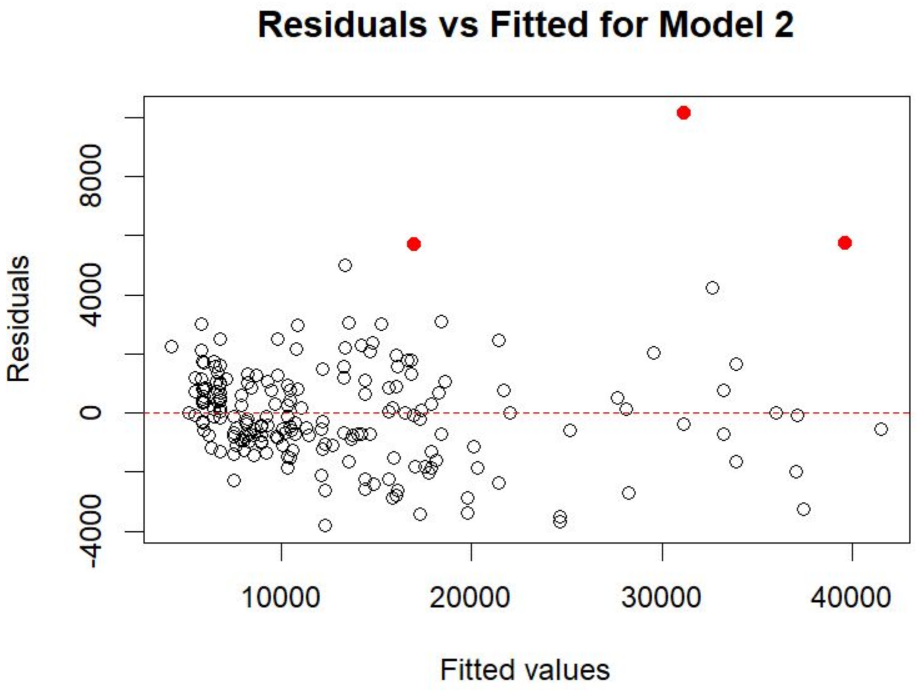

Residuals vs Fitted for Model 1
Residuals
-4000
0
2000
6000
10000
10000
20000
30000
40000
Fitted values


Normal Q-Q Plot for Model 2
Sample Quantiles
-4000
0
2000
6000
10000
-3
-2
-1
0
1
2
3
Theoretical Quantiles


Residuals vs Fitted for Model 2
Residuals
-4000
0
2000
6000
10000
10000
20000
30000
40000
Fitted values

Outliers Detection





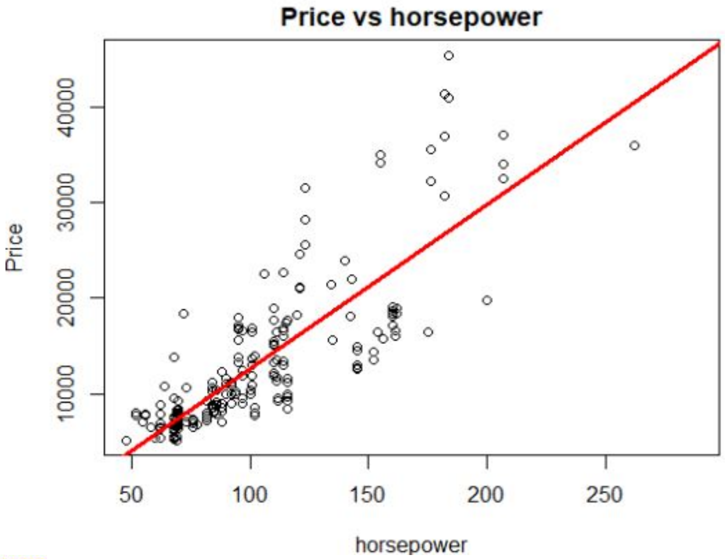
	AdjR2 in	AdjR2 out	MS_Res in	MS_Res out
Model 1 #60	0.9418388	0.9422036	3174806	3263245
Model 1 #68	0.9418388	0.9426322	3174806	3106651
Model 2 #72	0.9385753	0.9377706	3183245	3124377
Model 2 #199	0.9385753	0.9387636	3183245	3179110

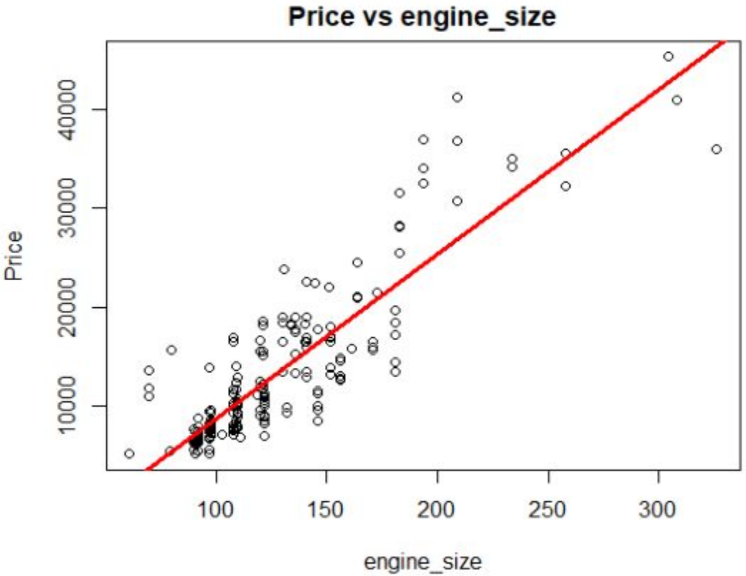
Leverage Points Detection

	AdjR2 in	AdjR2 out	Change Percentage	Estimator Name that Changed the most	Change Percentage
Model 1 #19	0.9418388	0.9429675	0.12	makechevrolet	245.46
Model 1 #66	0.9418388	0.9453205	0.369	makemercury	-174.98
Model 2 #2	0.9385753	0.9393009	7.73	makeaudi	-10237.97
Model 2 #3	0.9385753	0.9390276	4.82	makeaudi	17289.00

Influential Points Detection

No influential points detected.



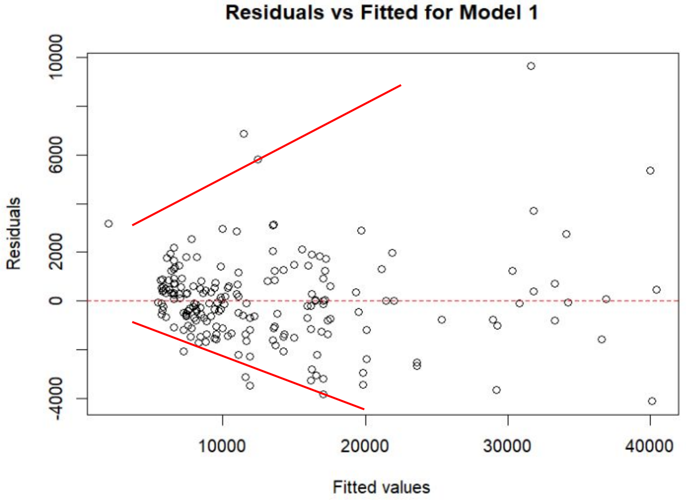


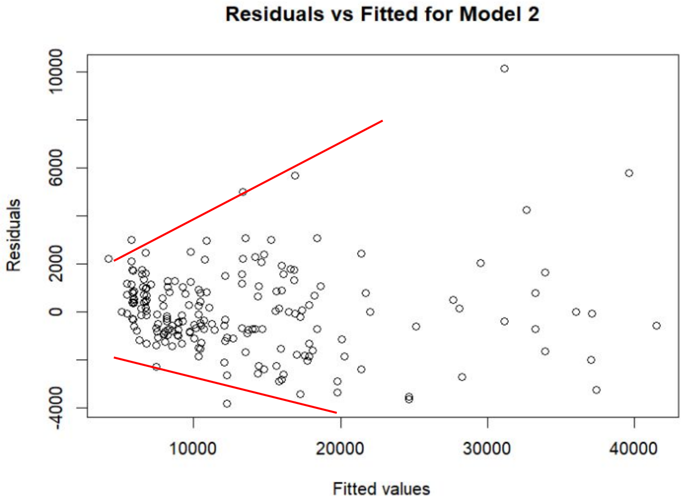
Model Transformation

1. Square Root Transformation

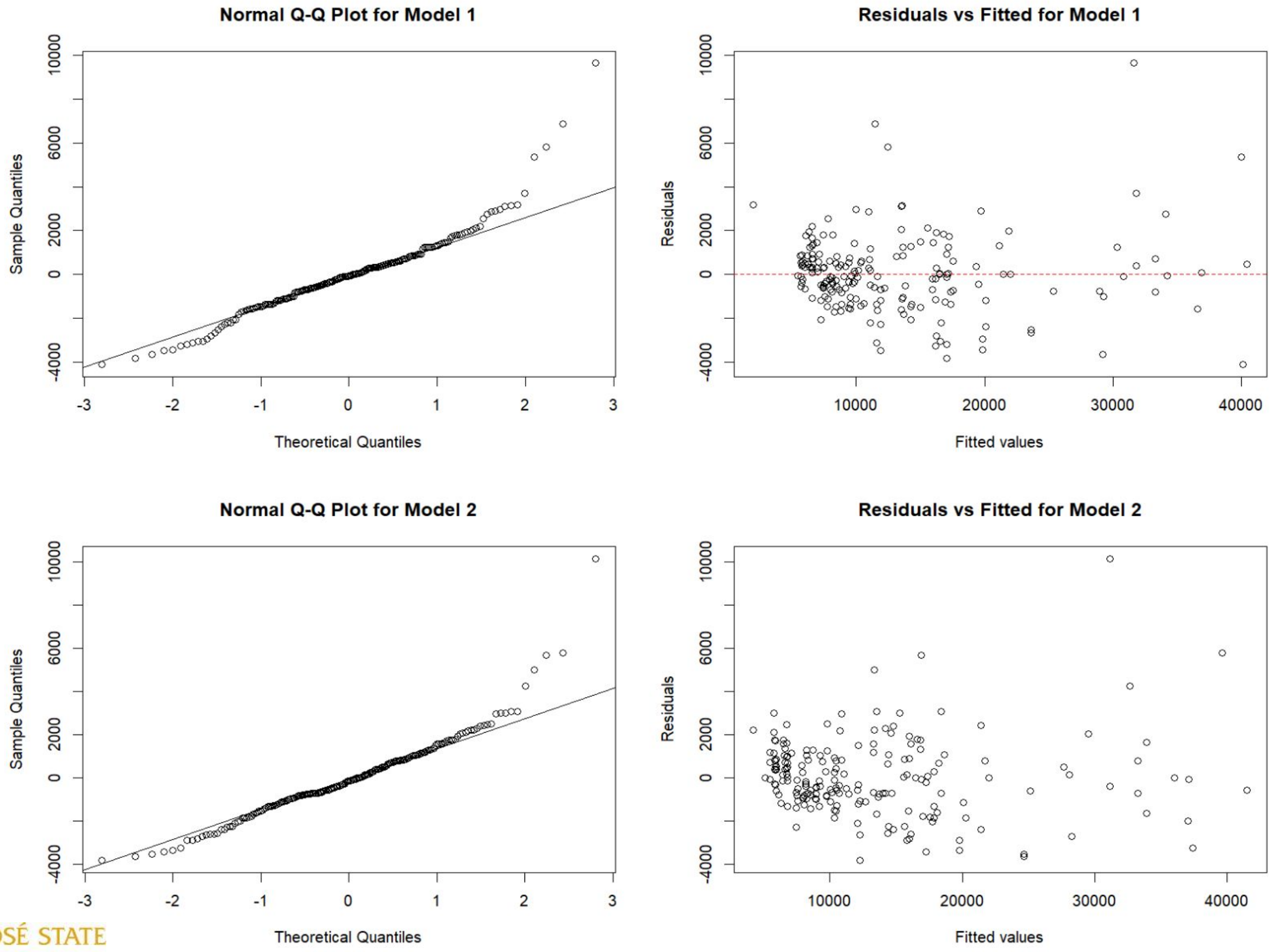
2. Log Transformation

- Optimal lambda for Model 1: 0.3
- Optimal lambda for Model 2: 0.2

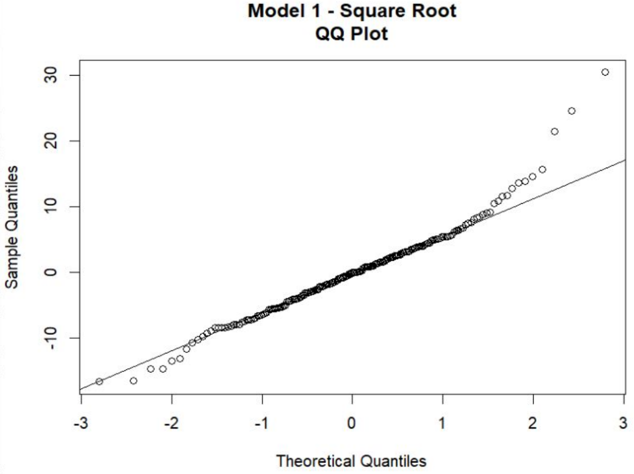


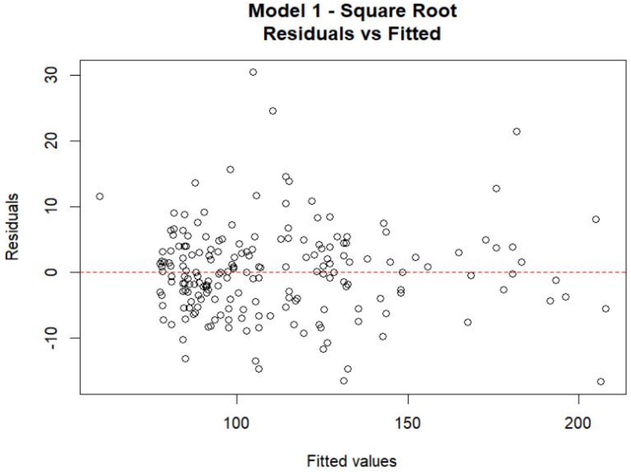


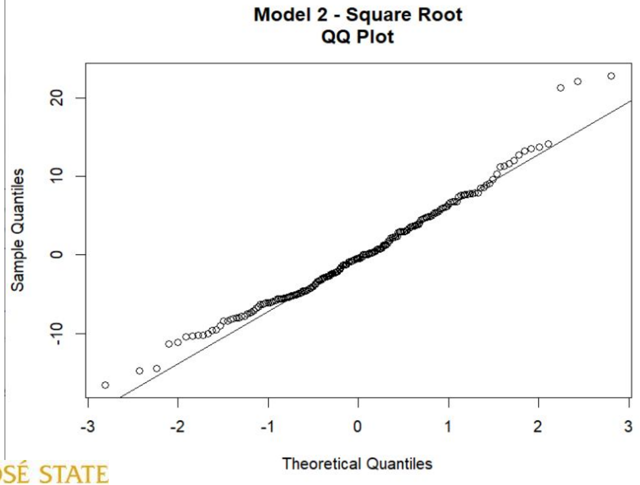
SJSU SAN JOSÉ STATE UNIVERSITY

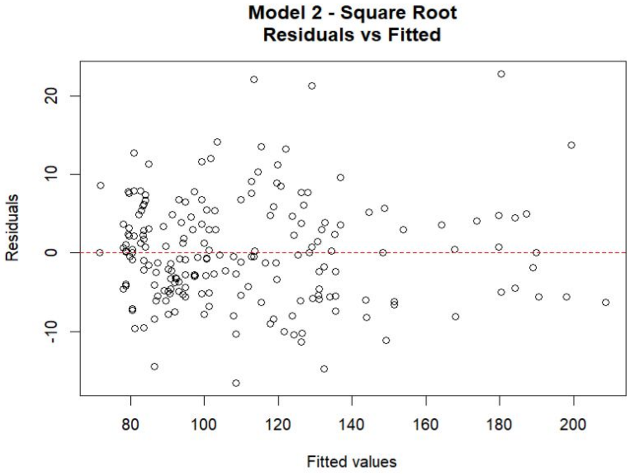


SJSU SAN JOSÉ STATE UNIVERSITY


Model 1 - Square Root
QQ Plot
Sample Quantiles
Theoretical Quantiles
-10
0
10
20
30
-3
-2
-1
0
1
2
3


Model 1 - Square Root
Residuals vs Fitted
Residuals
Fitted values
-10
0
10
20
30
100
150
200

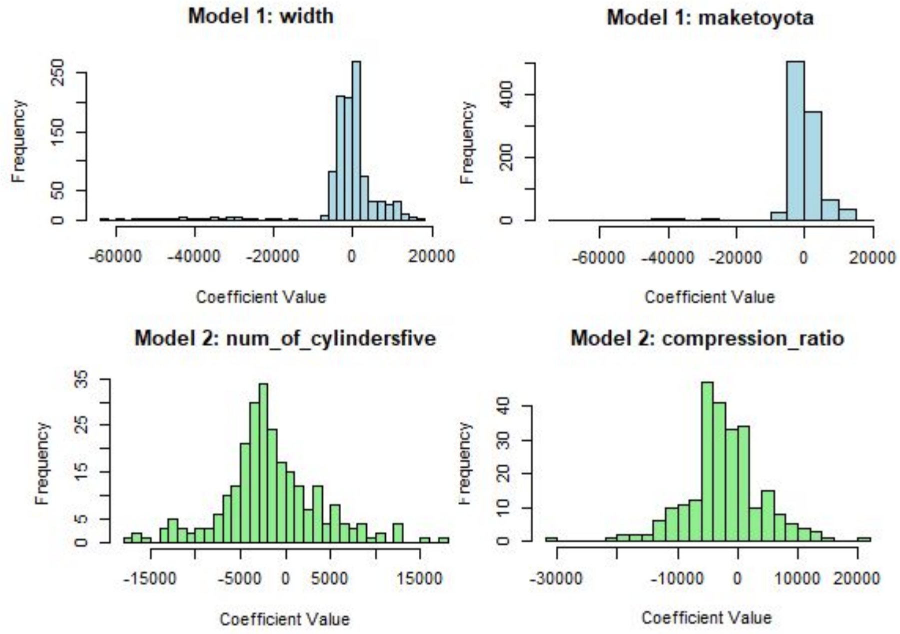

Model 2 - Square Root
QQ Plot
Sample Quantiles
Theoretical Quantiles
-10
0
10
20
-3
-2
-1
0
1
2
3

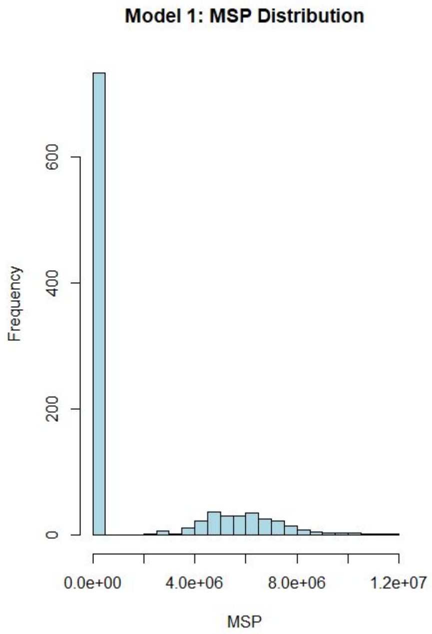

Model 2 - Square Root
Residuals vs Fitted
Residuals
Fitted values
-10
0
10
20
80
100
120
140
160
180
200

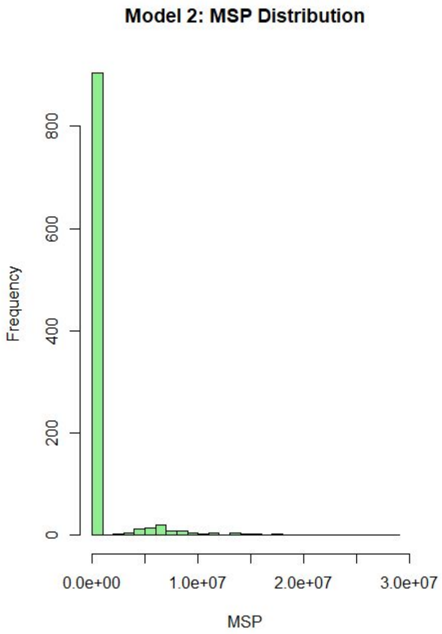
	Transformation	R^2	Adjusted R^2	Standard Error
Model 1	None	0.9414	0.9235	1443.658
Model 1	Square Root	0.9519	0.9431	1690.497
Model 2	None	0.9337	0.9122	1546,282
Model 2	Square Root	0.9532	0.9428	1677.584

Data Validation

- Model 2 has more MSP values close to 0 compared to model 1 MSP
- All distribution of the mean square error for prediction is normal with model 1 is slightly skewed







SJSU SAN JOSÉ STATE UNIVERSITY

Car Price Prediction Case in 1985

price ~ make + body_style + engine_location + wheel_base + engine_type + num_of_cylinders + compression_ratio + horsepower

Customer Requirement:

1. **Honda sedan** with **front** engine
2. **Wheelbase** of 100 inches
3. **OHC** (Overhead Cam) engine type
4. **4-cylinder** engine with **compression ratio** of 9.5, and 150 **horsepower**

Conclusion

1. Which predictors contribute significantly to the price of a brand-new car?

Best model: price ~ make + body_style + engine_location + wheel_base + engine_type + num_of_cylinders + compression_ratio + horsepower

2. How well can we predict the price of a brand-new car on the smaller subset of predictors?

Our model demonstrates robustness despite the presence of outliers and leverage points, as their removal does not significantly alter predicted car prices.

The `make` variable is most impacted by these points, highlighting the strong influence of car brand on the price.

Applying a square root transformation effectively mitigates the impact of outliers, further enhancing the model's performance, reliability and accurately predict new car prices.

Appendix

Pearson correlation coefficient table

	symboling	normalized_losses	wheel_base	length	width	height	curb_weight	engine_size	bore	stroke	num_of_cylinders
symboling	1.000000000	0.51838797	-0.52046477	-0.33621705	-0.2198496	-0.47399437	-0.2523723	-0.1102384	-0.25701277	-0.020538841	0.02354329
normalized_losses	0.518387968	1.00000000	-0.06400101	0.02911438	0.1048565	-0.41708077	0.1228602	0.2038412	-0.03616694	0.065626988	0.26588542
wheel_base	-0.520464770	-0.06400101	1.00000000	0.87196801	0.8159350	0.55876376	0.8105069	0.6504878	0.58048403	0.164011960	0.31381957
length	-0.336217051	0.02911438	0.87196801	1.00000000	0.8391841	0.50515596	0.8703550	0.7266664	0.64905924	0.116049120	0.39015769
width	-0.219849642	0.10485650	0.81593501	0.83918412	1.0000000	0.29840309	0.8706493	0.7800176	0.57504802	0.192891028	0.50786485
height	-0.473994373	-0.41708077	0.55876376	0.50515596	0.2984031	1.00000000	0.3693631	0.1165051	0.26150092	-0.095364375	-0.05496260
curb_weight	-0.252372341	0.12286025	0.81050693	0.87035496	0.8706493	0.36936307	1.0000000	0.8888474	0.64664028	0.171691317	0.59630323
engine_size	-0.110238431	0.20384120	0.65048780	0.72666638	0.7800176	0.11650514	0.8888474	1.0000000	0.59733622	0.296693139	0.77088755
bore	-0.257012766	-0.03616694	0.58048403	0.64905924	0.5750480	0.26150092	0.6466403	0.5973362	1.00000000	-0.105464066	0.13659466
stroke	-0.020538841	0.06562699	0.16401196	0.11604912	0.1928910	-0.09536437	0.1716913	0.2966931	-0.10546407	1.000000000	0.13093041
num_of_cylinders	0.023543289	0.26588542	0.31381957	0.39015769	0.5078648	-0.05496260	0.5963032	0.7708876	0.13659466	0.130930406	1.00000000
compression_ratio	-0.139021791	-0.12997093	0.29396760	0.18896778	0.2615303	0.23743151	0.2265128	0.1435677	0.01921597	0.240894808	0.06300331
horsepower	-0.003668657	0.29090559	0.51450686	0.66672597	0.6787789	0.03226392	0.7885094	0.8098548	0.55710740	0.149314989	0.61773846
peak_rpm	0.199797806	0.24067647	-0.29249053	-0.23910434	-0.2359063	-0.25123623	-0.2620855	-0.2872601	-0.31584138	-0.008568987	-0.11971892
city_mpg	0.088912095	-0.23693364	-0.57663540	-0.71687663	-0.6621225	-0.19455902	-0.7595379	-0.6958896	-0.58561823	-0.021380833	-0.48333020
highway_mpg	0.149309477	-0.18969131	-0.60826982	-0.71783122	-0.6893674	-0.22164557	-0.7871670	-0.7113644	-0.58672907	-0.013974079	-0.51826633

	compression_ratio	horsepower	peak_rpm	city_mpg	highway_mpg
symboling	-0.13902179	-0.003668657	0.199797806	0.08891209	0.14930948
normalized_losses	-0.12997093	0.290905591	0.240676469	-0.23693364	-0.18969131
wheel_base	0.29396760	0.514506864	-0.292490530	-0.57663540	-0.60826982
length	0.18896778	0.666725972	-0.239104336	-0.71687663	-0.71783122
width	0.26153025	0.678778916	-0.235906329	-0.66212250	-0.68936743
height	0.23743151	0.032263922	-0.251236231	-0.19455902	-0.22164557
curb_weight	0.22651275	0.788509418	-0.262085506	-0.75953792	-0.78716702
engine_size	0.14356771	0.809854784	-0.287260069	-0.69588958	-0.71136436
bore	0.01921597	0.557107399	-0.315841384	-0.58561823	-0.58672907
stroke	0.24089481	0.149314989	-0.008568987	-0.02138083	-0.01397408
num_of_cylinders	0.06300331	0.617738464	-0.119718918	-0.48333020	-0.51826633
compression_ratio	1.00000000	-0.162893609	-0.418726319	0.27951325	0.22244152
horsepower	-0.16289361	1.000000000	0.074931817	-0.83717978	-0.82797250
peak_rpm	-0.41872632	0.074931817	1.000000000	-0.05493781	-0.03437238
city_mpg	0.27951325	-0.837179780	-0.054937813	1.00000000	0.97199680
highway_mpg	0.22244152	-0.827972503	-0.034372382	0.97199680	1.00000000