Report on Analysis of Car Inventory

BASED ON 2014 INVENTORY

BY JOSH LIU, FOR ILLINI AUTO

This report is prepared by Josh, contracted by client Illini Auto.

Introduction

This data analysis report is based on the car inventory in 2004, on all models of cars in the inventory with an engine of 4, 6, 8 cylinders.

This report is compiled based on the dataset provided by the client. The dataset has 415 observations, each observation being a model of vehicles in the inventory of the client. Variables include Make, Model, MSRP, Invoice Price, Engize size, number of cylinders, horsepower, weight, length, and wheelbase length.

The purpose of this analysis is to identify underlying patterns between profitability and technical parameters of automobiles in inventory, in order to facilitate more informed decision-making in the operation of the dealership.

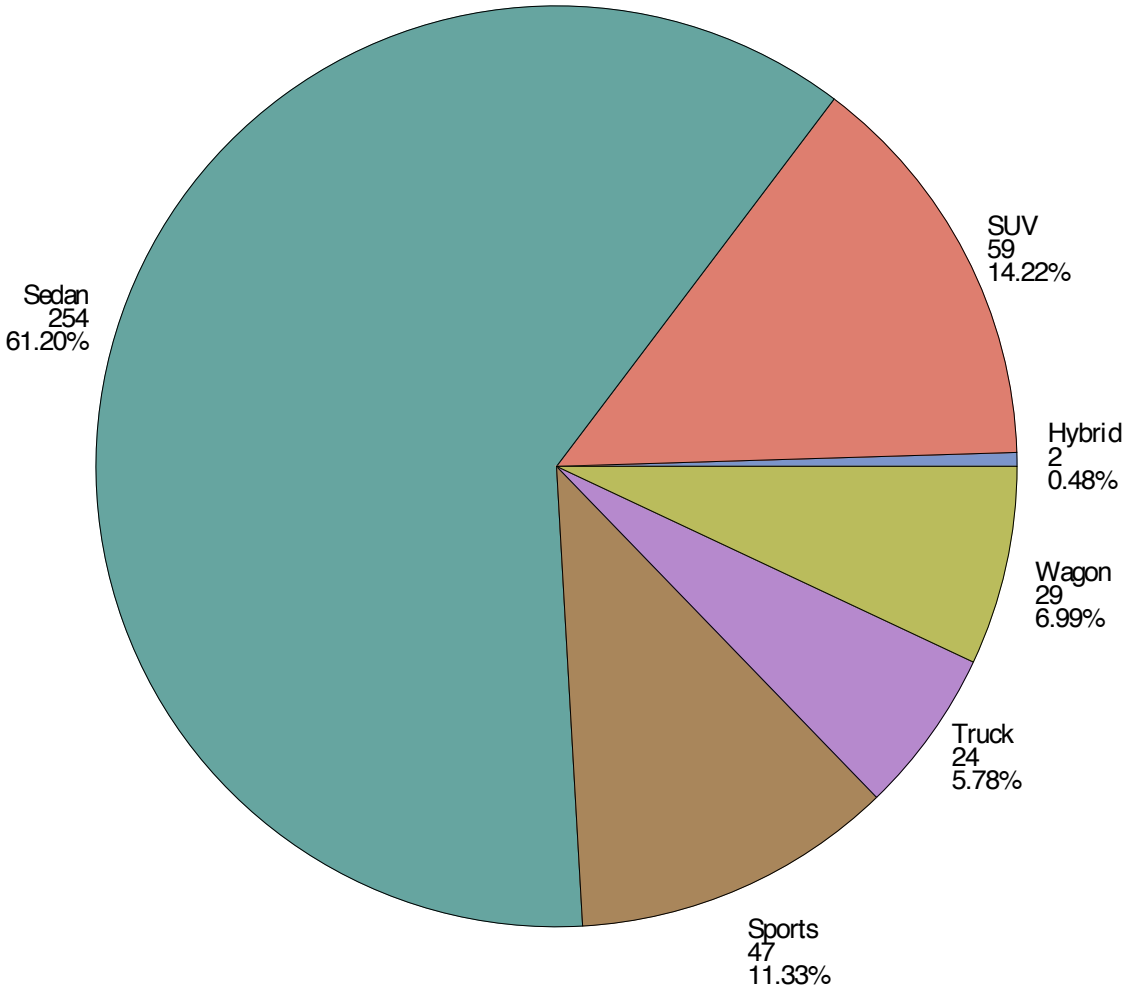
As requested by the client, this data analysis report will cover the following topics: 1. descriptive statistics of characteristics of cars from the previous year; 2. Analysis of effect of techinical parameters on profitability ; 3. Statistical modeling of profitability and interpretation; 4. Analysis of origins of cars: domestic or imports; 5. Classification of cars based on the number of cylinders; 6. Appendix: diagnostics of statistic models.

Part 1: Descriptive Statistics

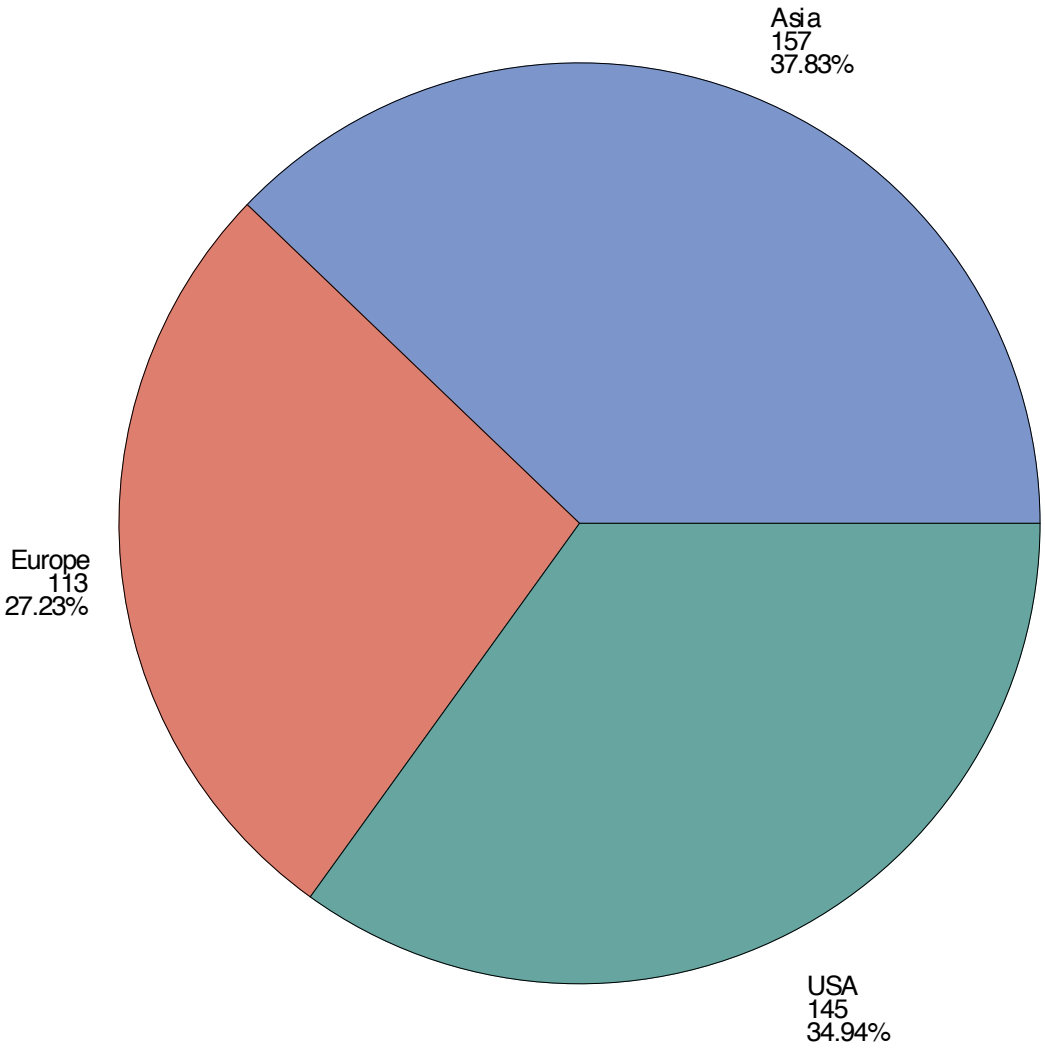
This table shows the top 20 makes that has the most models on sale in the dealership:

Obs	Make	COUNT	PERCENT
1	Toyota	28	6.74699
2	Chevrolet	27	6.50602
3	Mercedes-Benz	24	5.78313
4	Ford	22	5.30120
5	BMW	20	4.81928
6	Audi	19	4.57831
7	Nissan	17	4.09639
8	Honda	16	3.85542
9	Chrysler	15	3.61446
10	Volkswagen	14	3.37349
11	Mitsubishi	13	3.13253
12	Dodge	12	2.89157
13	Hyundai	12	2.89157
14	Jaguar	12	2.89157
15	Kia	11	2.65060
16	Lexus	11	2.65060
17	Mazda	11	2.65060
18	Pontiac	11	2.65060
19	Subaru	11	2.65060
20	Buick	9	2.16867

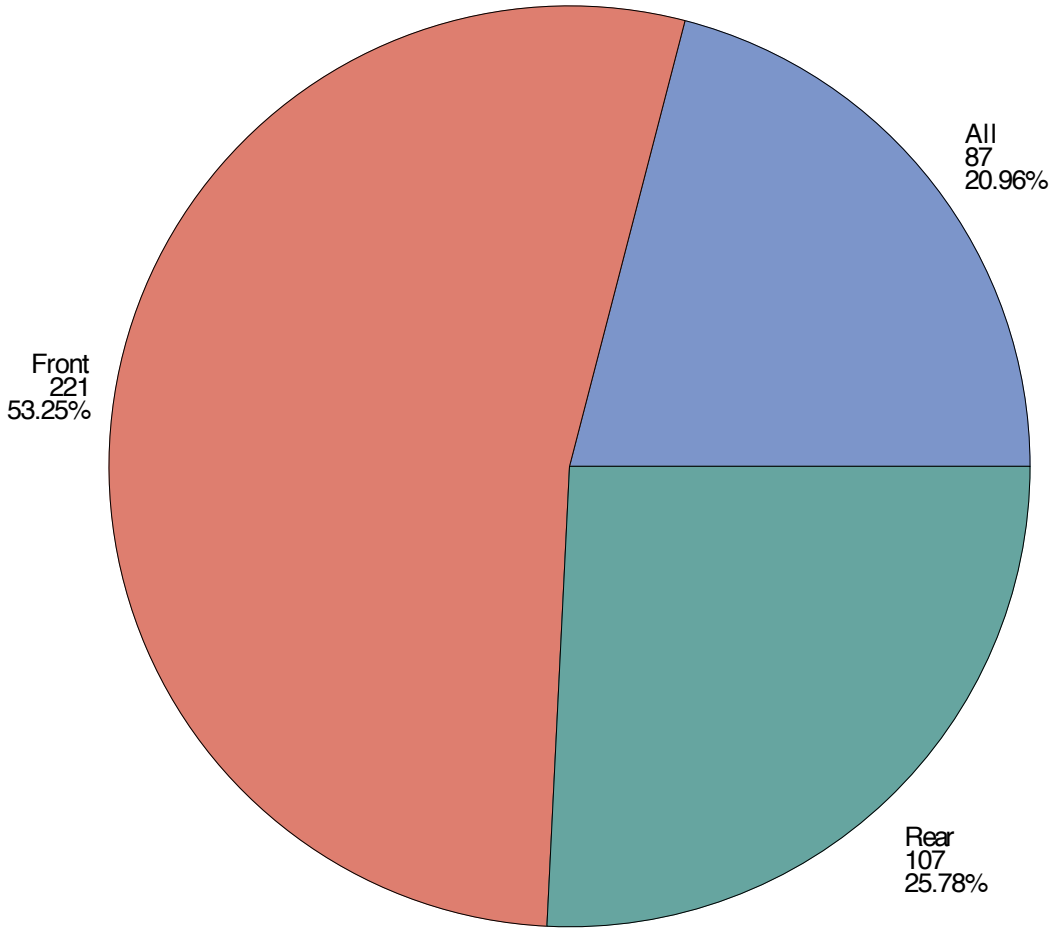
FREQUENCY of Type



FREQUENCY of Origin



FREQUENCY of DriveTrain



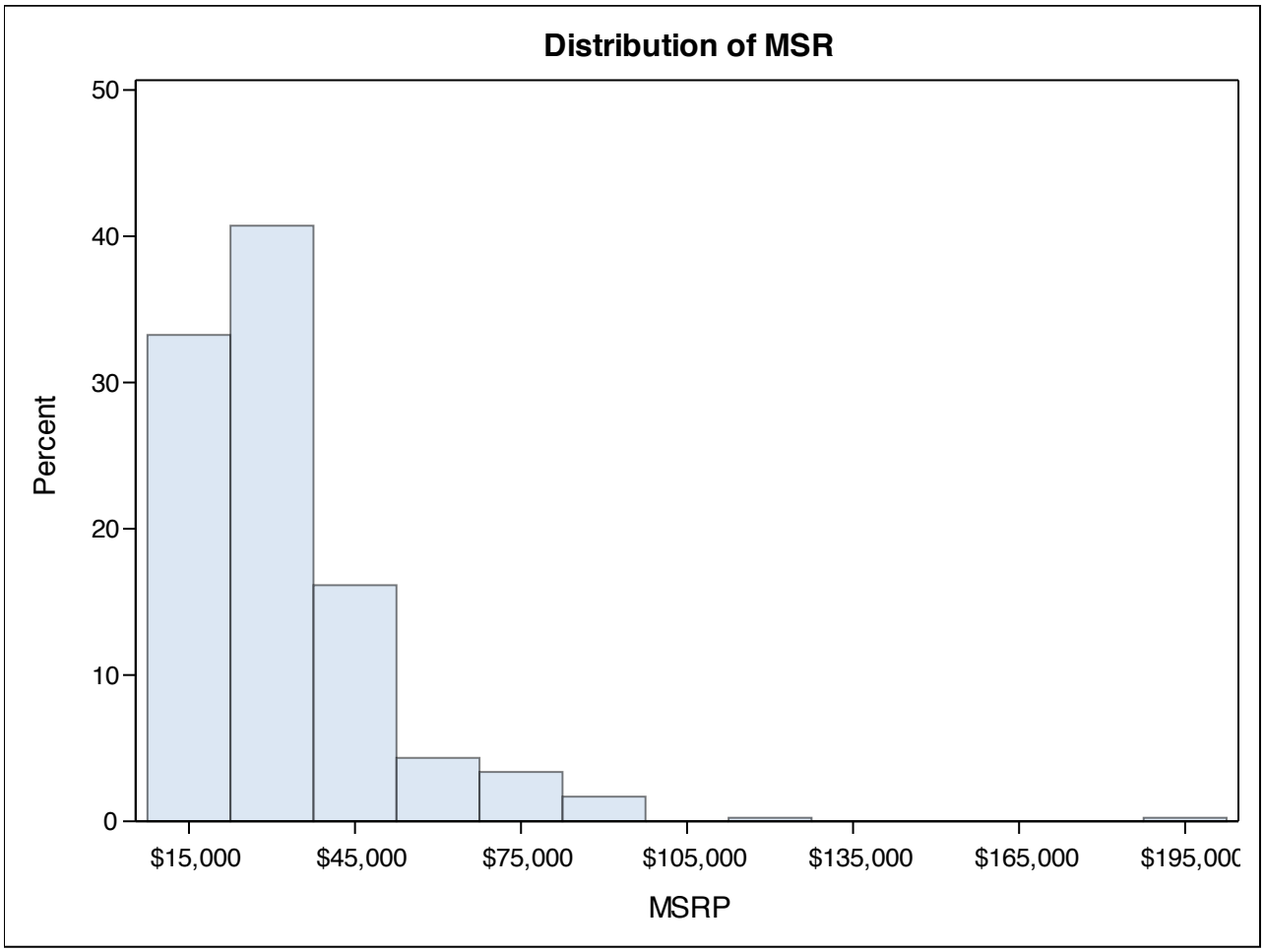
Distribution of MSRP

Variable:
MSRP

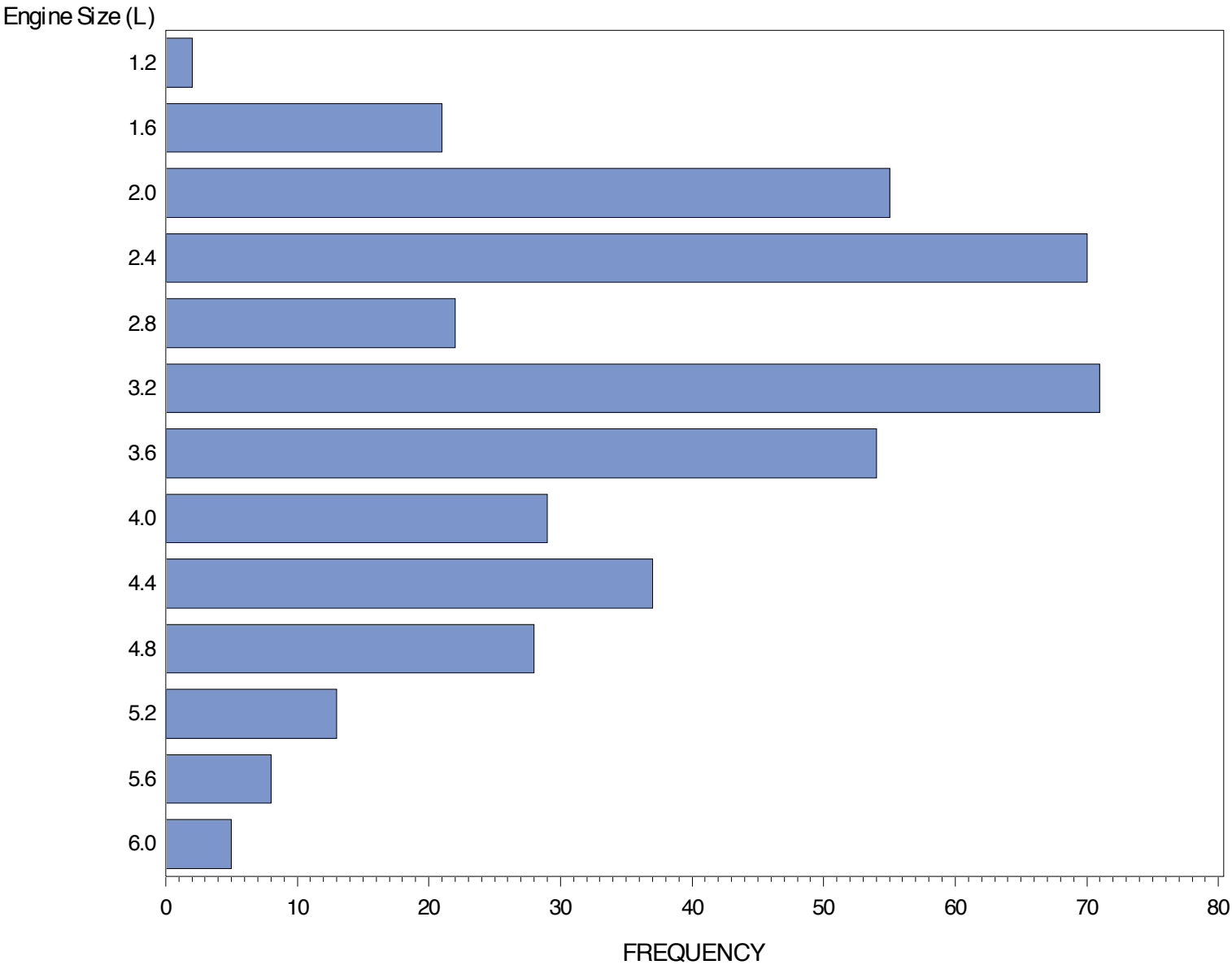
Basic Statistical Measures			
Location		**Variability**	
Mean	32035.80	**Std Deviation**	18282
Median	27339.00	**Variance**	334241033
Mode	13270.00	**Range**	182185
		Interquartile Range	17695

Note: The mode displayed is the smallest of 18 modes with a count of 2.

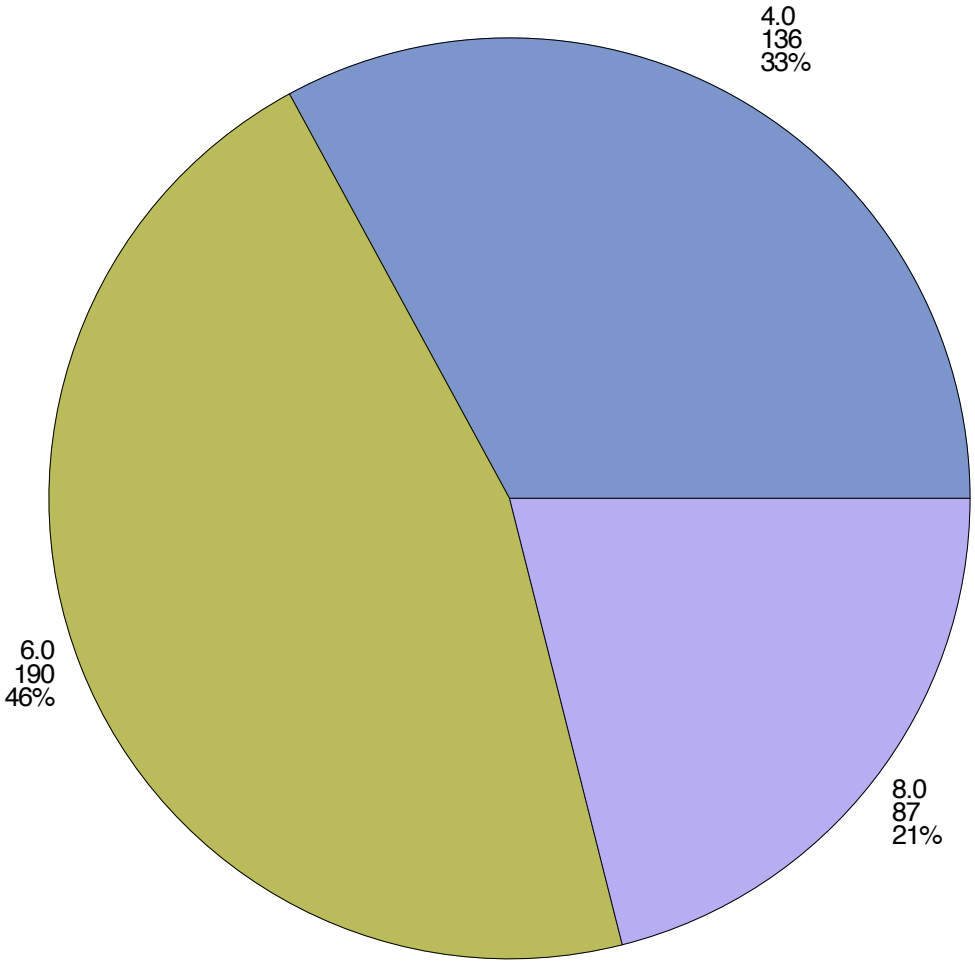
Distribution of MSRP



Most models on sale have MSRP between $15,000 and $45,000, with a few models having MSRP far outlying, as high as $195,000.

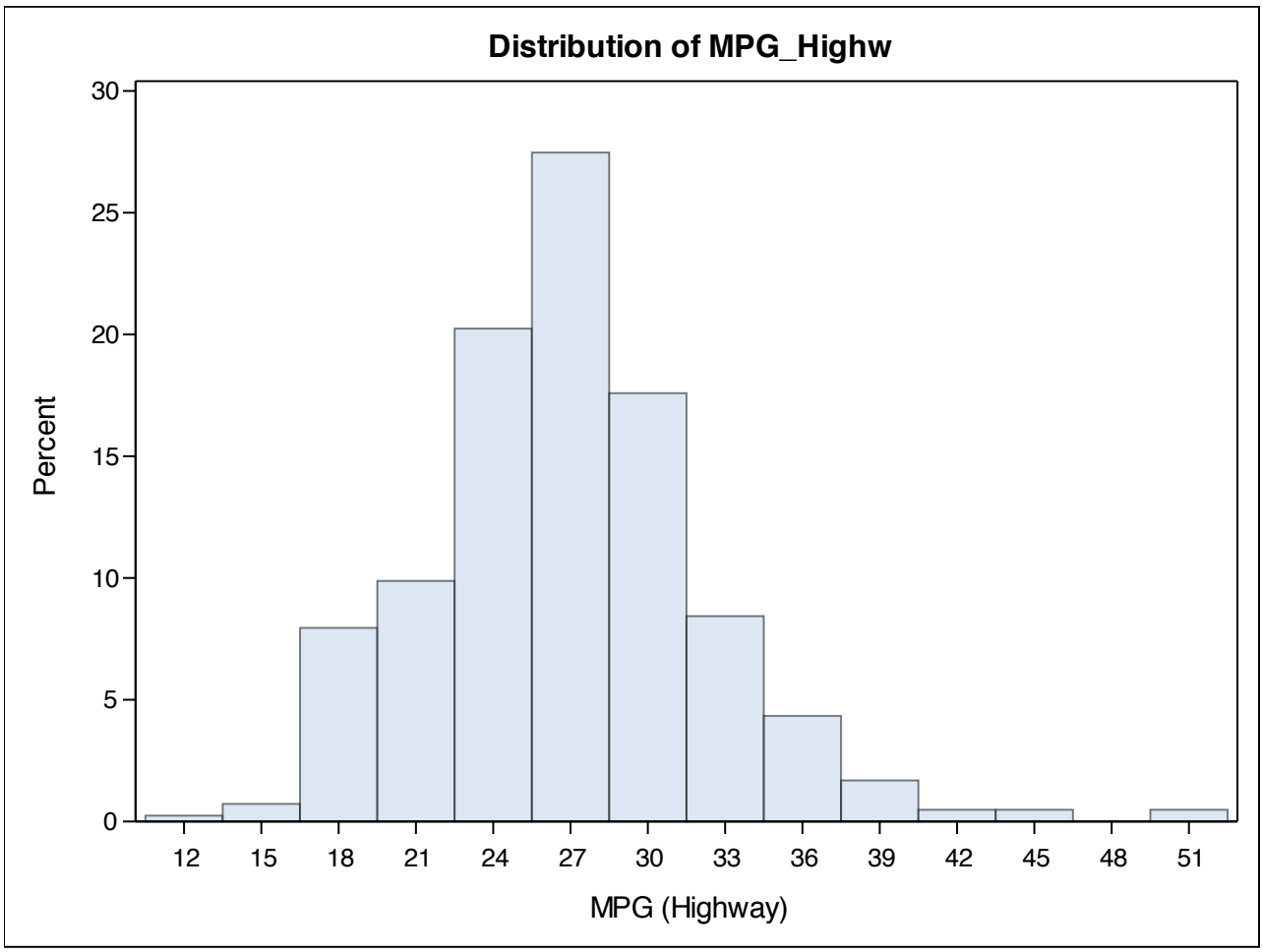

Engine Size (L)
1.2
1.6
2.0
2.4
2.8
3.2
3.6
4.0
4.4
4.8
5.2
5.6
6.0
0
10
20
30
40
50
60
70
80
FREQUENCY

FREQUENCY of Cylinders



Variable: MPG_Highway (MPG (Highway))

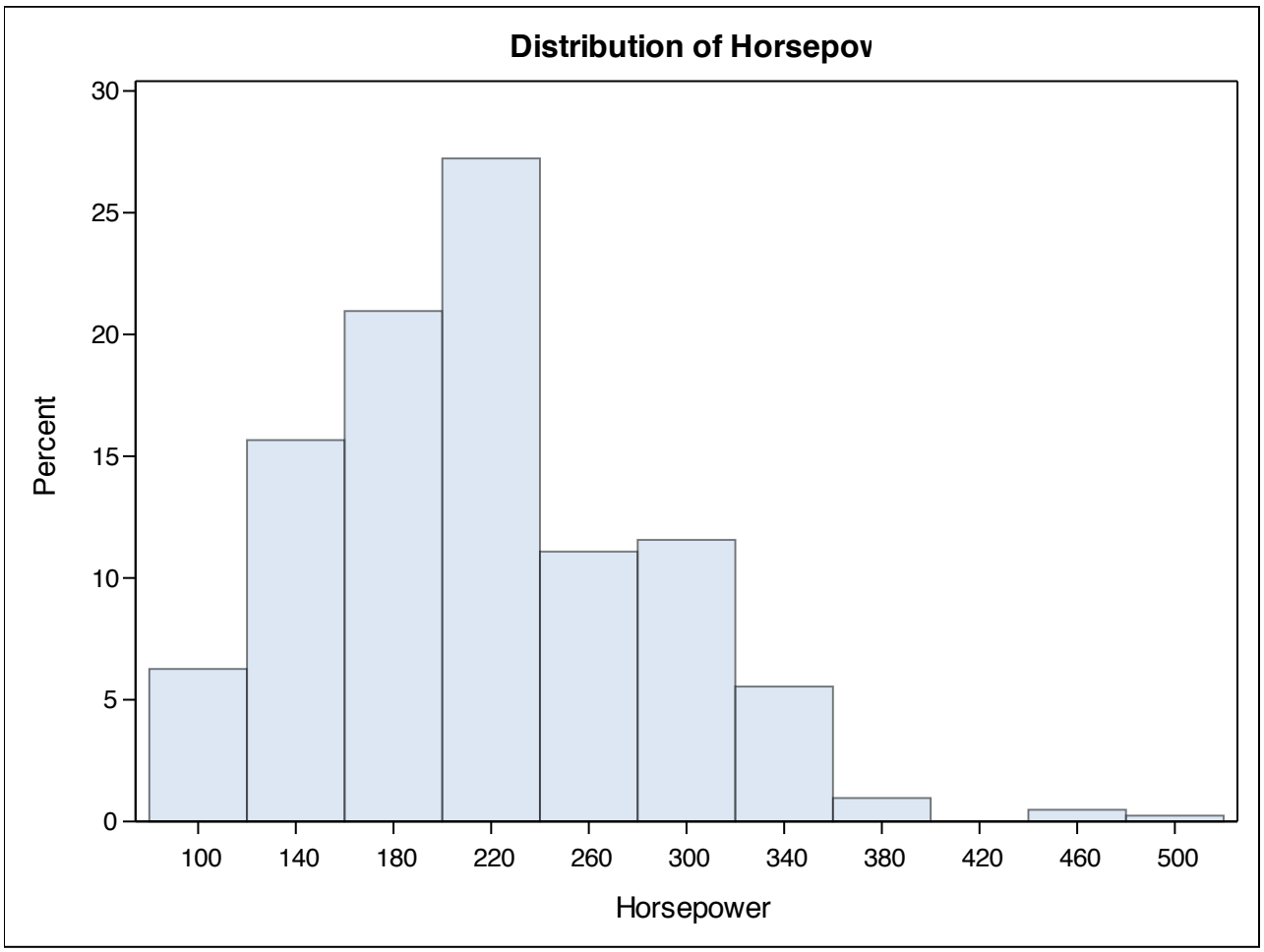
Basic Statistical Measures			
Location		**Variability**	
Mean	26.85542	**Std Deviation**	5.40873
Median	26.00000	**Variance**	29.25441
Mode	26.00000	**Range**	39.00000
		Interquartile Range	6.00000



Highway MPG is pretty nicely distributed, with most models having MPG between 21 and 33.

Variable:
Horsepower

Basic Statistical Measures			
Location		**Variability**	
Mean	213.2867	**Std Deviation**	67.49068
Median	210.0000	**Variance**	4555
Mode	200.0000	**Range**	400.00000
		Interquartile Range	85.00000

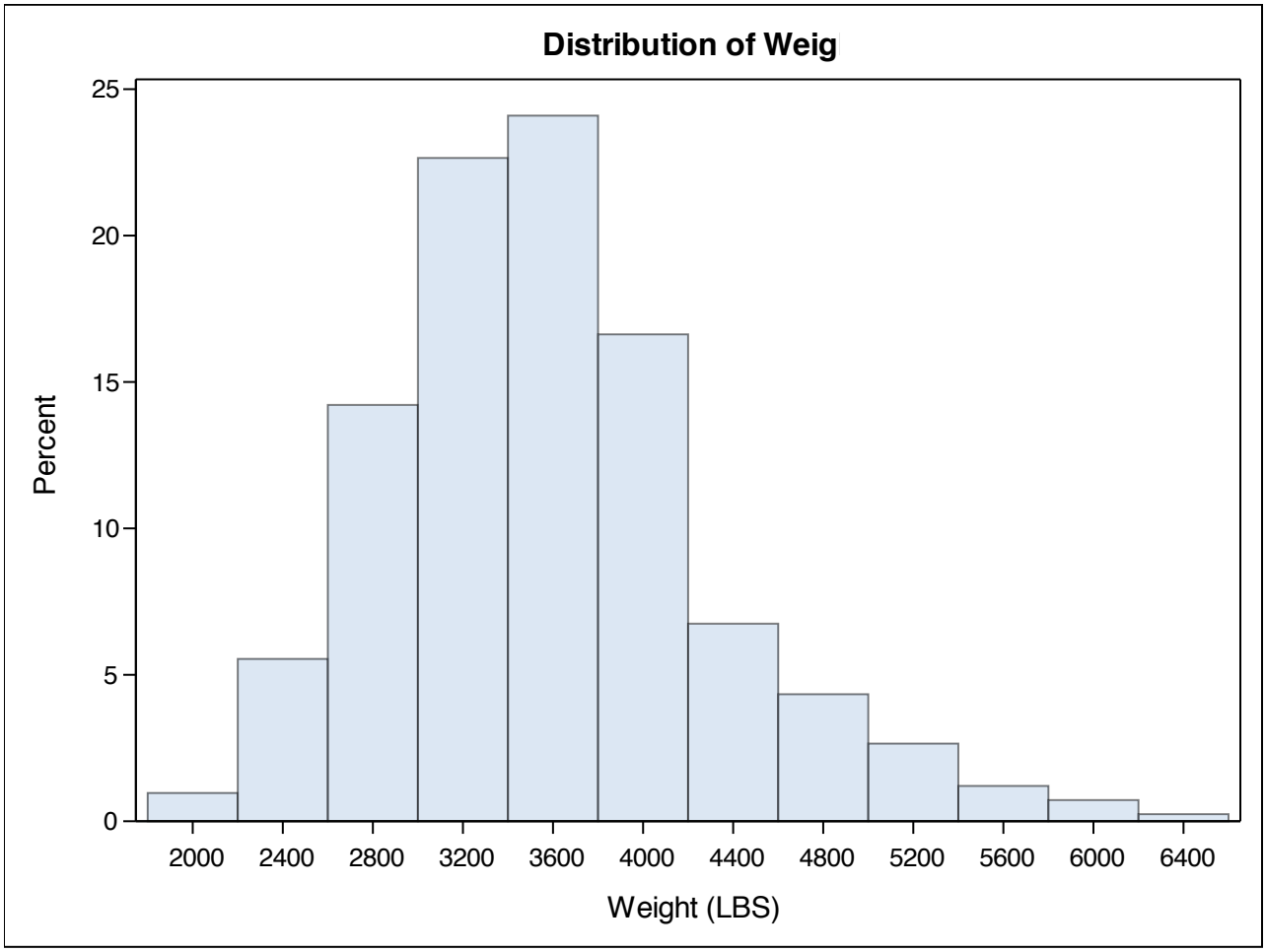


Horsepower of most models falls between 140 and 300. There are a few models with horse power less than 140 and higher than 380, which are considered to be outliers.

Variable: Weight (Weight (LBS))

Basic Statistical Measures			
Location		**Variability**	
Mean	3563.757	**Std Deviation**	736.78849
Median	3470.000	**Variance**	542857
Mode	3175.000	**Range**	4365
		Interquartile Range	887.00000

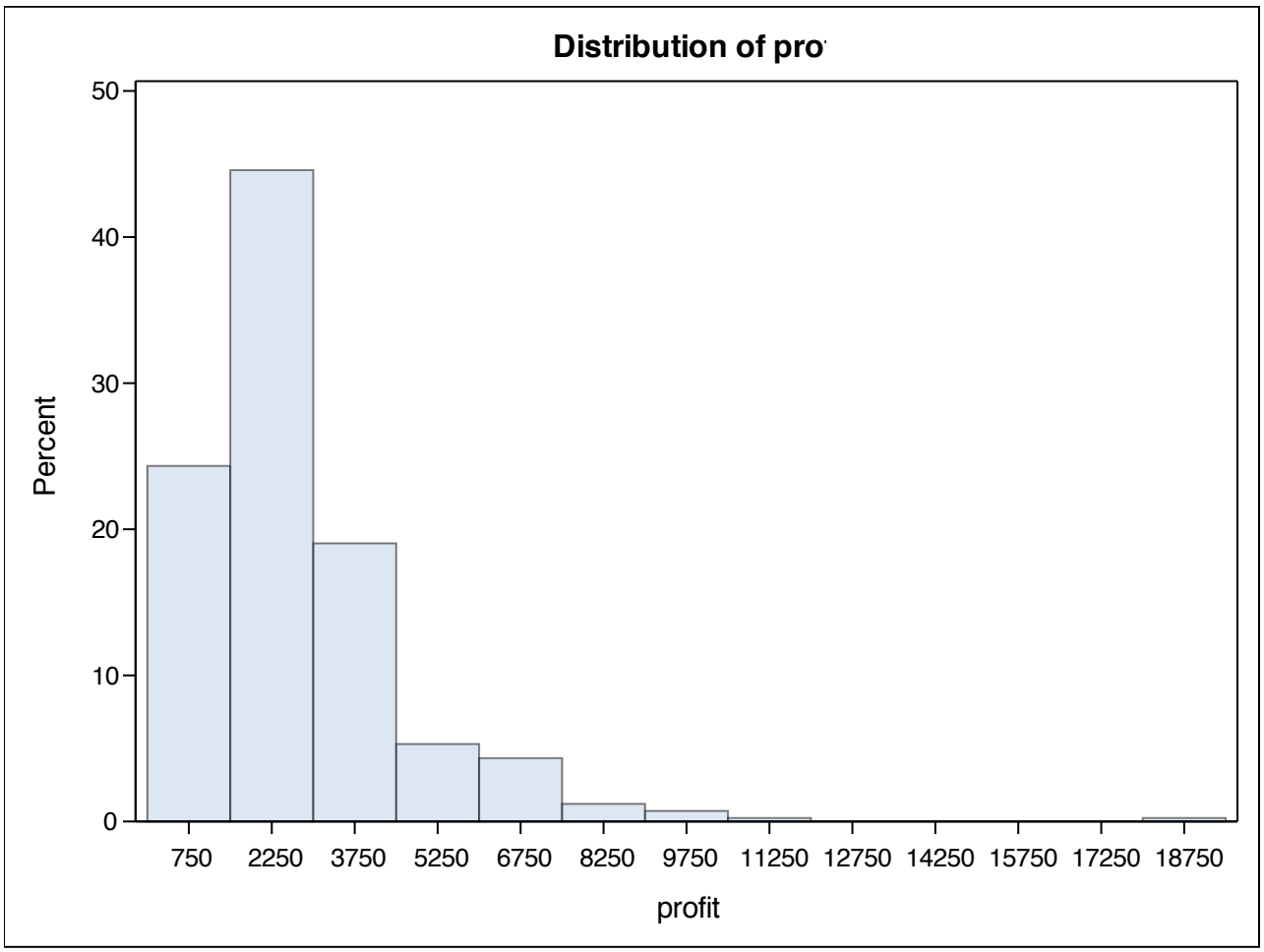
Note: The mode displayed is the smallest of 2 modes with a count of 4.


Distribution of Weig
Percent
25
20
15
10
5
0
2000
2400
2800
3200
3600
4000
4400
4800
5200
5600
6000
6400
Weight (LBS)

Variable:
profit

Basic Statistical Measures			
Location		**Variability**	
Mean	2721.186	**Std Deviation**	1908
Median	2298.000	**Variance**	3639460
Mode	730.000	**Range**	18752
		Interquartile Range	1864

Note: The mode displayed is the smallest of 16 modes with a count of 2.



Most models on sale have expected profit between $750 and $6,750, with one model having expected profit far outlying, as high as $18,750.

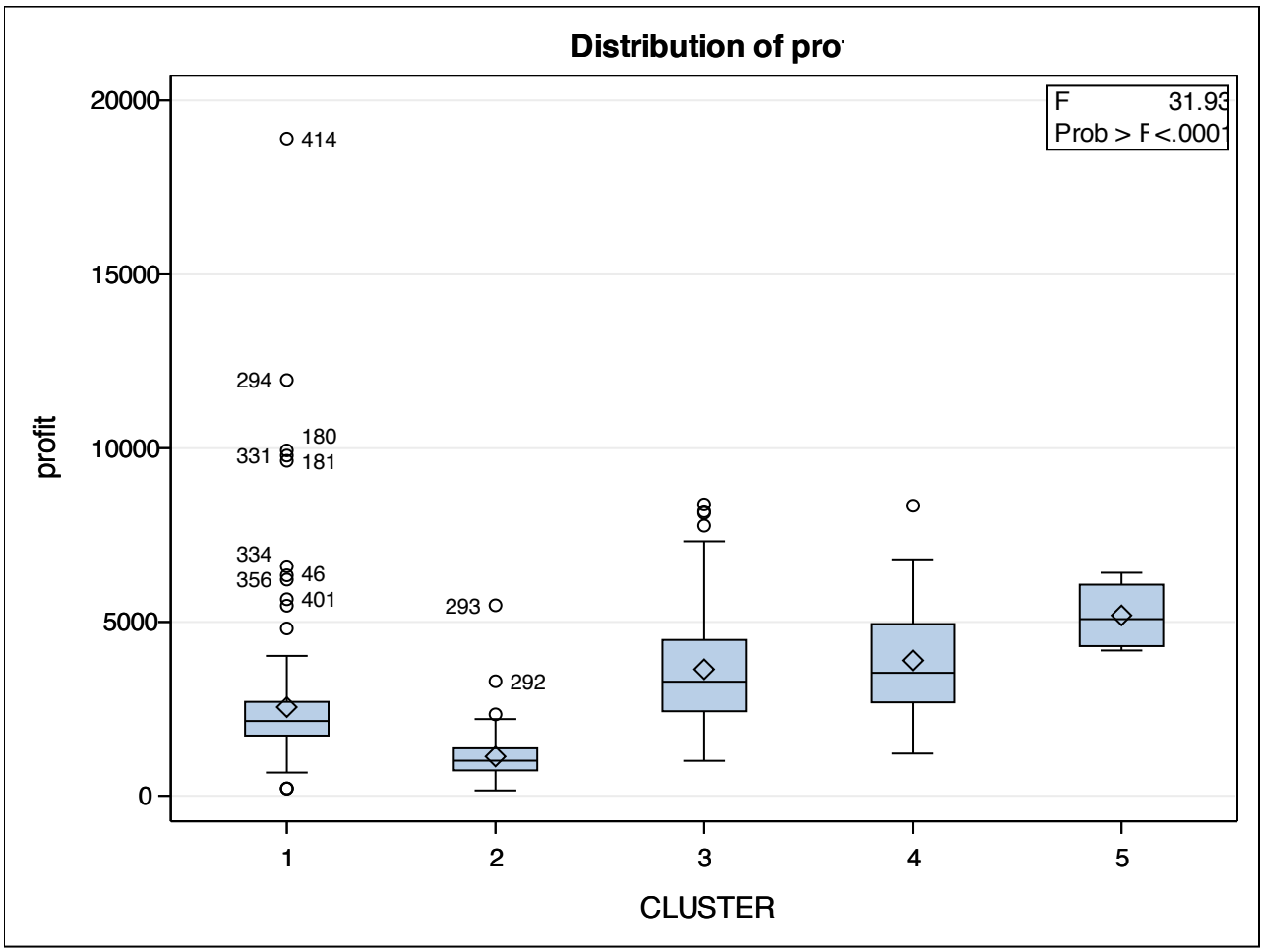
Part 2: Compare Expected Profit by Group defined by vehicle-related technical parameters

In this part, I will group all vehicles into tree groups based on Engine size, horse power, fuel efficienty, and vehicle size measures.

And then, I will perform a series of analyses to see if there is any difference of profit generated by the three groups of vehicles.

CLUSTER				
1	**2**	**3**	**4**	**5**
N	**N**	**N**	**N**	**N**
194	74	106	37	4

Dependent Variable: profit



This plot shows the distribution of profit by group. We can see clearly that Group 3's vehicles have larger profit than Group 1, and Group 1's vehicles have larger profit than Group 2.

Further analysis demonstrates a high correlation between profit and vehicle power and size parameters. See the following boxplots for more details.

The boxplots above show larger vehicle, more powerful vehicle, vehicles that burn more gas bring in more profit. Without consideration of impact on the environment, the dealer is advised to sell more vehicles that come with a large-volume engine, with an engine of more cylinders, that burns a lot of gas per unit of distance traveled, vehicles of large length, and vehicles of heavy weight.

Rebalancing inventory towards more larger vehicles is a strategy that can increase expected profit for the dealer.

Part 3: Statistical Model Construction to Predict Expected Profit

In this part, I will develop a statistical model with Expected Profit as response variable. This part gives final result of the model. Intermediary process and diagnostics procedures are described in the appendix at the end of the report.

Vehicle Wheelbase By Group

Model: MODEL1
Dependent Variable: logprofit

Number of Observations Read	406
Number of Observations Used	406

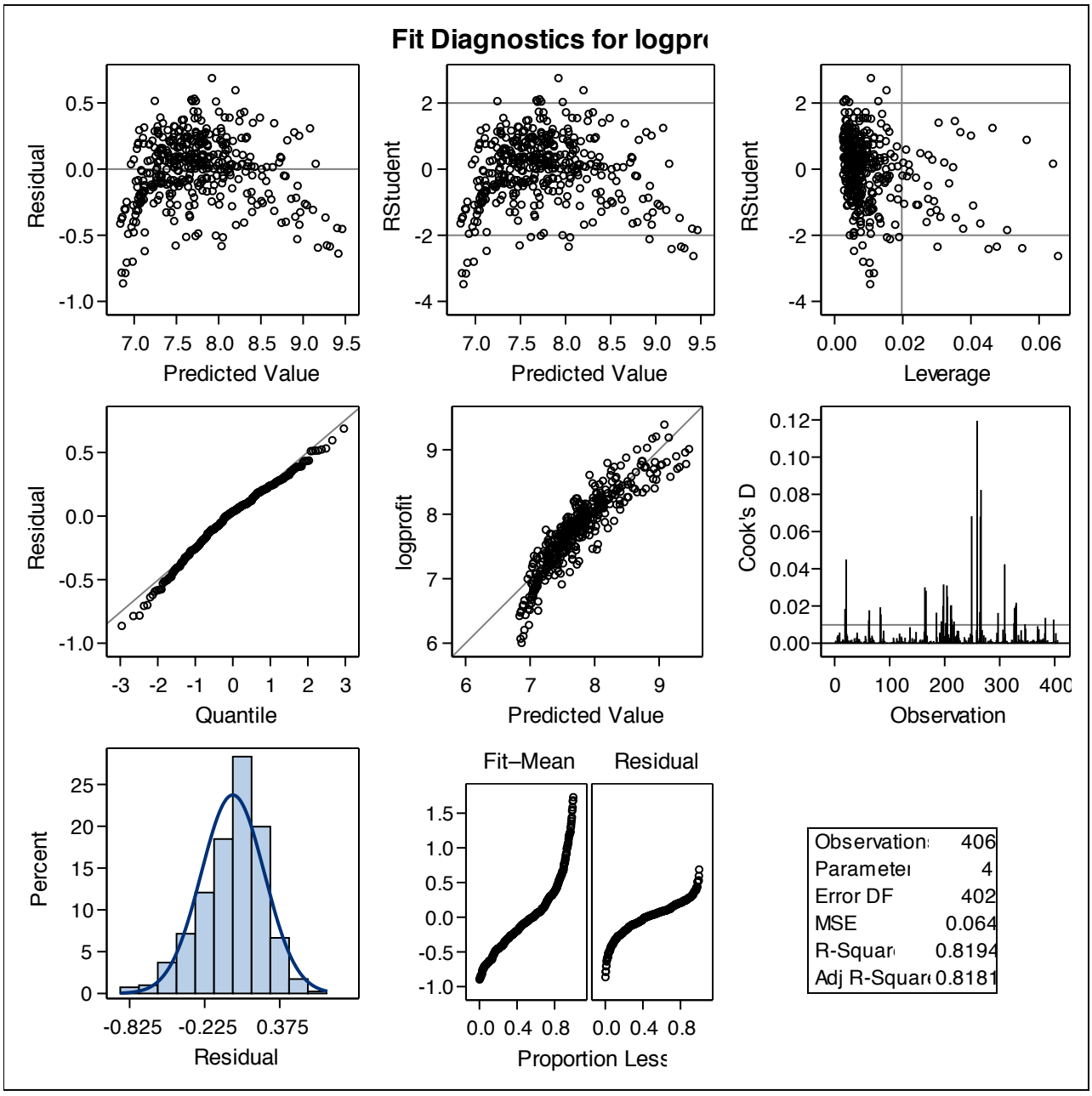
Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	116.81415	38.93805	608.00	<.0001
Error	402	25.74537	0.06404		
Corrected Total	405	142.55952			

Root MSE	0.25307	**R-Square**	0.8194
Dependent Mean	7.73261	**Adj R-Sq**	0.8181
Coeff Var	3.27273		

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	6.09590	0.06342	96.12	<.0001
MSRP		1	0.00002284	0.00000135	16.95	<.0001
Weight	Weight (LBS)	1	0.00013958	0.00002283	6.11	<.0001
Horsepower		1	0.00194	0.00037829	5.13	<.0001

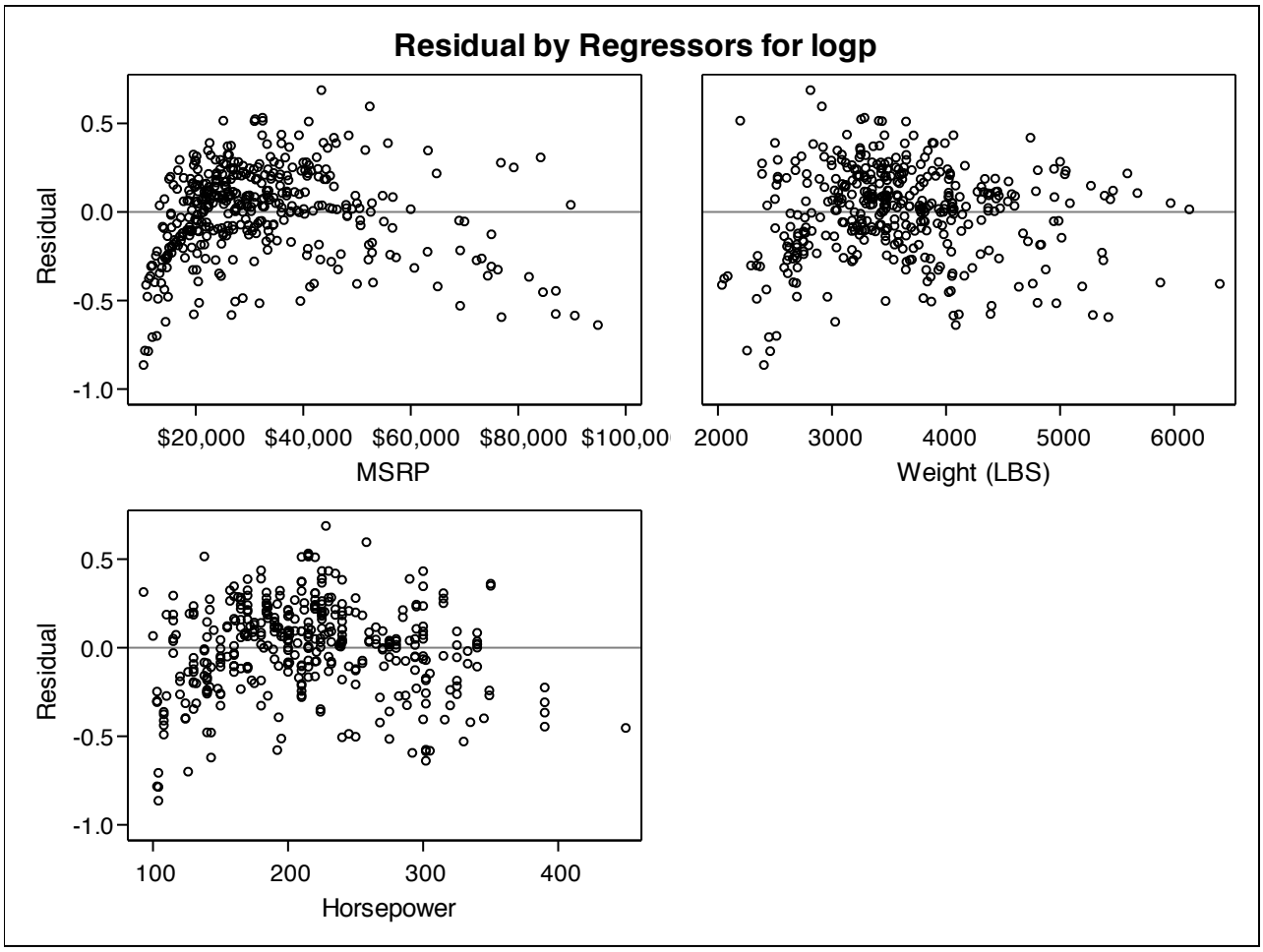
Vehicle Wheelbase By Group

Model: MODEL1
Dependent Variable: logprofit



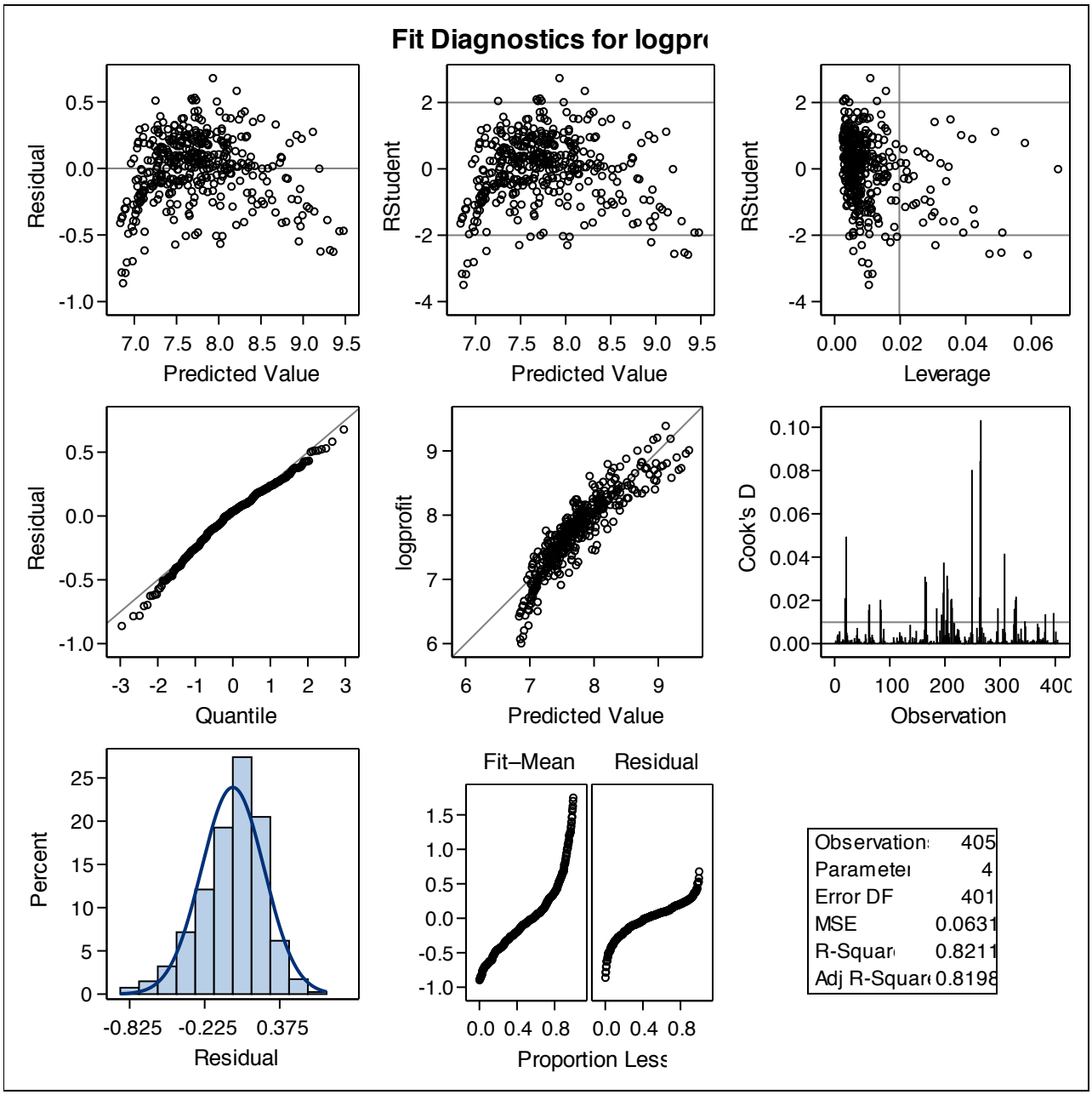
Vehicle Wheelbase By Group

Model: MODEL1
Dependent Variable: logprofit



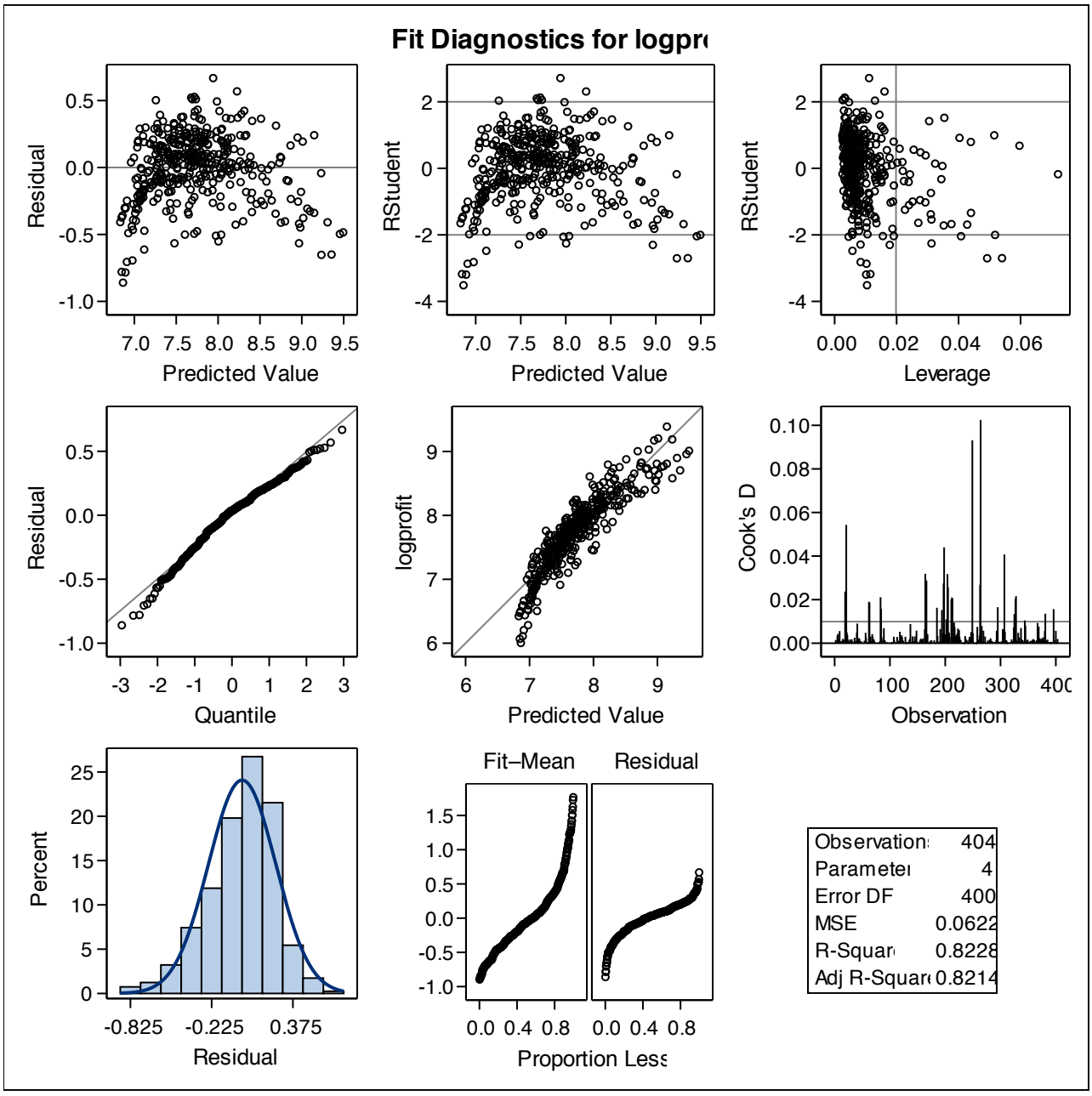
Vehicle Wheelbase By Group

Model: MODEL1
Dependent Variable: logprofit



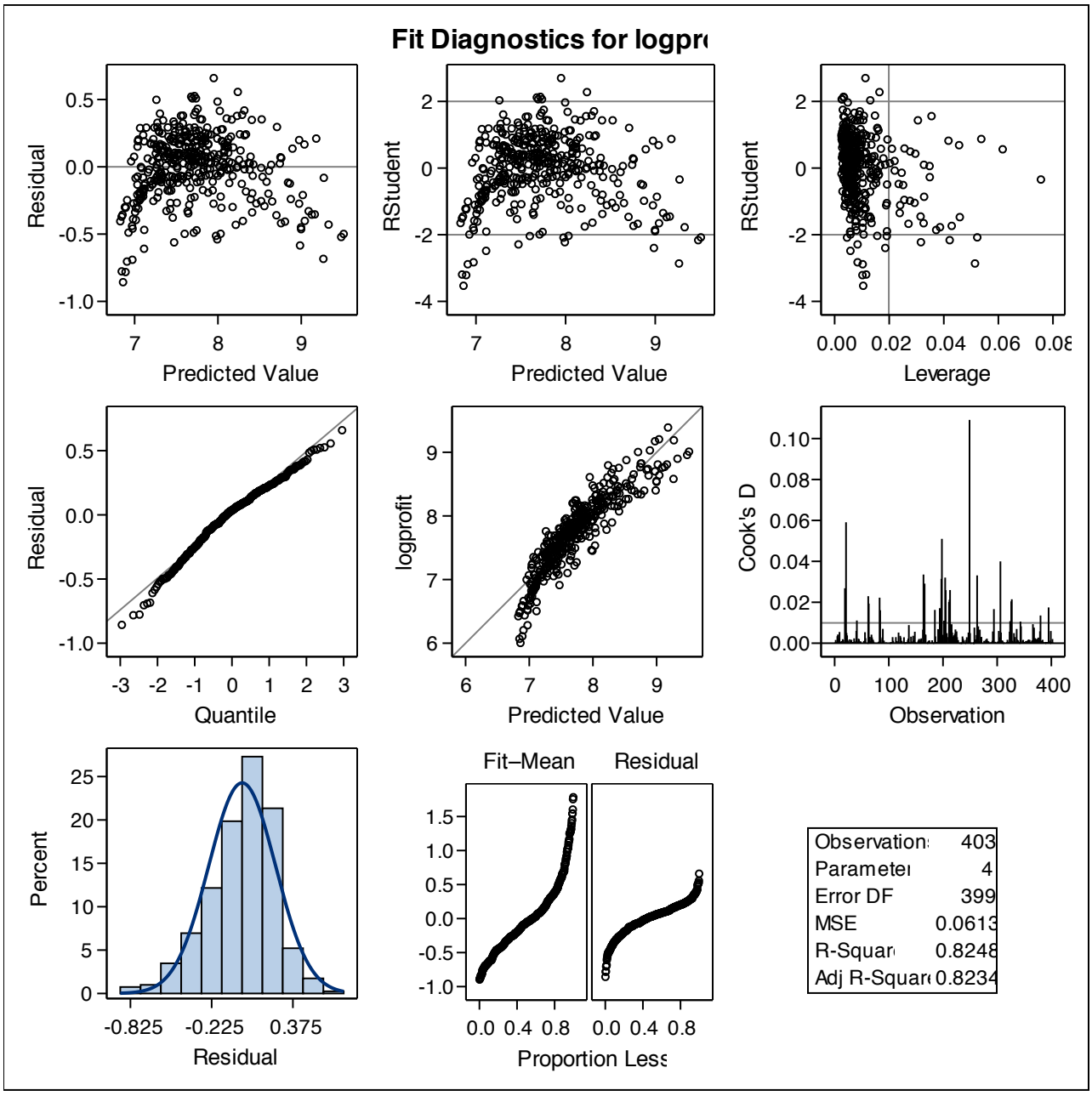
Vehicle Wheelbase By Group

Model: MODEL1
Dependent Variable: logprofit



Vehicle Wheelbase By Group

Model: MODEL1
Dependent Variable: logprofit



Vehicle Wheelbase By Group

Model: MODEL1
Dependent Variable: logprofit

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	114.81249	38.27083	635.92	<.0001
Error	398	23.95246	0.06018		
Corrected Total	401	138.76495			

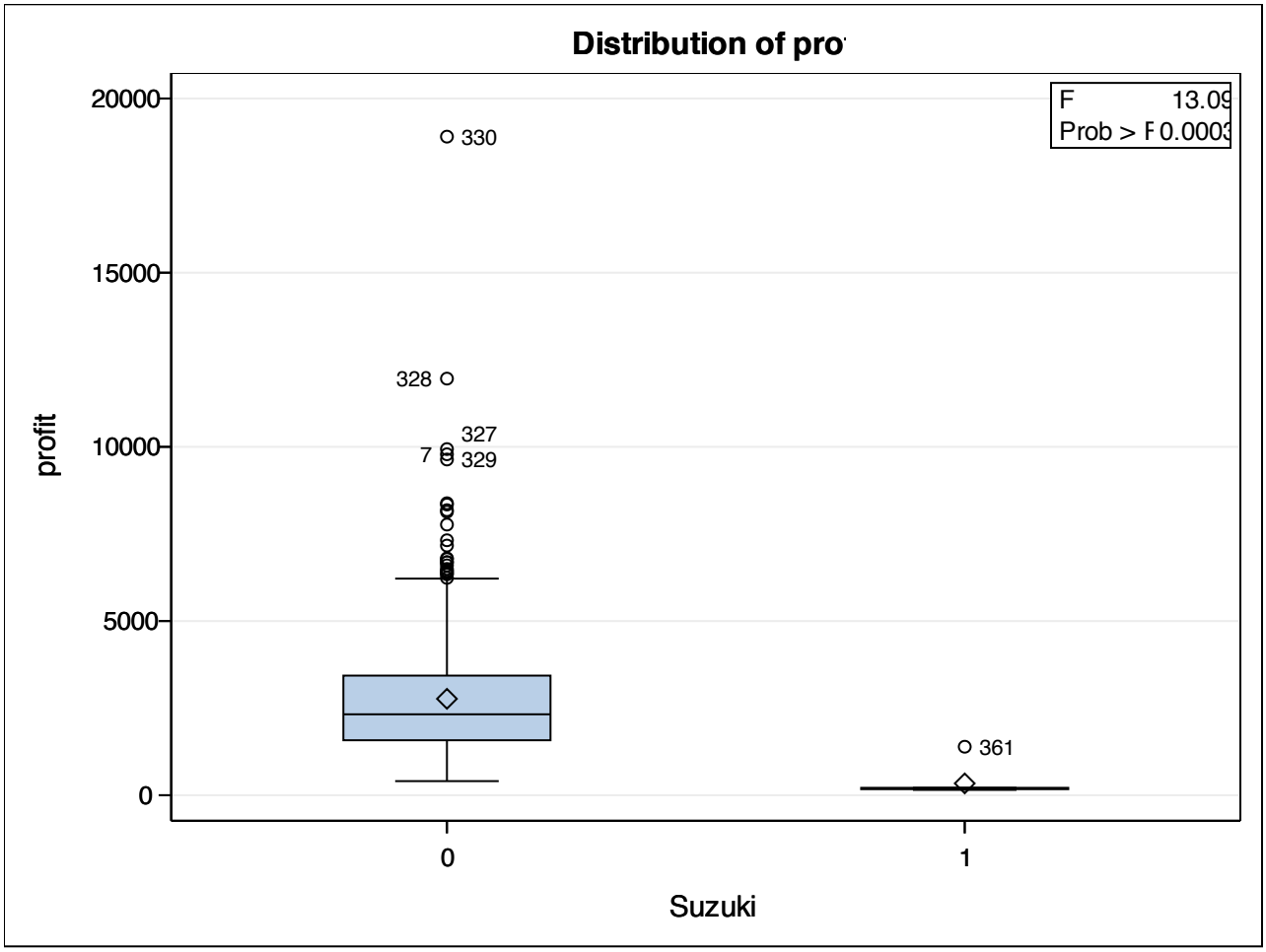
Root MSE	0.24532	**R-Square**	0.8274
Dependent Mean	7.72300	**Adj R-Sq**	0.8261
Coeff Var	3.17649		

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	6.08890	0.06170	98.69	<.0001
MSRP		1	0.00002613	0.00000144	18.14	<.0001
Weight	Weight (LBS)	1	0.00014819	0.00002231	6.64	<.0001
Horsepower		1	0.00137	0.00038118	3.60	0.0004

To analyze what variables affect expected profit the most, I developed a linear regression model and attempted to isolate three variables that have a strong bearing on expected profit.
During the analysis process, it happens that all Suzuki models, of which there are a total of 8, could not be accommodated in the final model. Therefore, analysis on expected profit is only conducted on non-Suzuki models. Suzuki vehicles will be analyzed in a separate part.
Three variables were identified as the most influential on the expected profit: MSRP, Weight, and Horsepower.
The model shows, holding all other variable constants, with MSRP increasing by $1,000, expected profit increases by 2.65%;
With Weight increasing by 100 lbs, expected profit increases by 1.50%;
With Horsepower increasing by 10, expected profit increases by 1.38%
Based on the statistical model, it is suggested that vehicles with a higher MSRP, a heavier Weight, and a stronger horsepower is expected to generate a larger profit.
Note that we excluded Suzuki from linear regression. Now I am going to compare profit of cars made by Suzuki, and all other cars.

Expected Profit: Suzuki or All Others

Dependent Variable: profit



Result shows that Suzuki cars have significanly less expected profit than other cars.

Part 4: Determination, based on other variables, whether a model of vehicle is domestic or an import

In this part, I will develop a logistic regression model, and identify which factors are associated with the odd of a certain vehicle being domestic. I will also try to predict whether a certain model of vehicle is domestic or an import. To conduct such an analysis, I re-coded the variable Origin into 'Domestic' if its value is 'USA', and 'Foreign' if otherwise.

Response Profile		
Ordered Value	DorF	Total Frequency
1	Domestic	145
2	Foreign	268

Probability modeled is DorF='Domestic'.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-8.3893	3.1484	7.1003	0.0077
EngineSize	1	3.7004	0.4774	60.0721	<.0001
Cylinders	1	-0.6333	0.2491	6.4640	0.0110
Horsepower	1	-0.0384	0.00576	44.4721	<.0001
MPG_Highway	1	0.2605	0.0847	9.4648	0.0021
MPG_City	1	-0.2956	0.1047	7.9669	0.0048
Weight	1	-0.00131	0.000479	7.5059	0.0061
Wheelbase	1	0.1022	0.0288	12.6049	0.0004

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
EngineSize	40.462	15.873	103.140
Cylinders	0.531	0.326	0.865
Horsepower	0.962	0.951	0.973
MPG_Highway	1.298	1.099	1.532
MPG_City	0.744	0.606	0.914
Weight	0.999	0.998	1.000
Wheelbase	1.108	1.047	1.172

This model shows that the odd of a car being domestic is associated with the following contributing factors: bigger engine size, better MPG on high way, and longer wheelbase.
With 0.1 additional unit increase in Engine Size, the odd of a vehicle being domestic is expected to increase by 44.78% (40.462 ^ 0.1 -1).
With one additional unit increase in highway MPG, the odd of a vehicle being domestic is expected to increase by 29.8 %.
With one additional unit increase in wheelbase, the odd of a vehicle being domestic is expected to increase by 10.8%.
This model also shows that the odd of a car being domestic is associated with the following undermining factors: less cylinders, less horsepower, lower MPG in the city, and lighter weight.

Odds Ratio Estimates		
Effect	Point Estimate	95% Wald Confidence Limits

With one more cylinder, the odd of a vehicle being domestic is expected to decrease by 46.9%.
With one additional unit increase in horsepower, the odd of a vehicle being domestic is expected to decrease by 3.8%.
With one additional unit increase in city MPG, the odd of a vehicle being domestic is expected to decrease by 25.6%.
With 100lbs increase in weight, the odd of a vehicle being domestic is expected to increase by 9.5%.
Part 5: Determination, based on other variables, on the number of cylinders a vehicle has

Stepwise Selection Summary									
Step	Number In	Entered	Removed	Label	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda
1	1	EngineSize		Engine Size (L)	0.8352	1038.68	<.0001	0.16483333	<.0001
2	2	MPG_City		MPG (City)	0.1225	28.55	<.0001	0.14463931	<.0001
3	3	Invoice			0.1136	26.13	<.0001	0.12821472	<.0001
4	4	MPG_Highway		MPG (Highway)	0.0168	3.48	0.0318	0.12606003	<.0001

Step	Number In	Entered	Removed	Average Squared Canonical Correlation	Pr > ASCC
1	1	EngineSize		0.41758334	<.0001
2	2	MPG_City		0.47868663	<.0001
3	3	Invoice		0.49373813	<.0001
4	4	MPG_Highway		0.50019651	<.0001

Discriminant analysis shows that the vehicles of different numbers of cylinders can be efficiently classified using variables MPG_City, MPG_Highway, Invoice, and Enginesize.
The result shows that we can classify vehicles with different numbers of cylinders, with a high probability of success, based on City MPG, Highway MPG, Invoice Price, and Enginesize.
Part 6: Conclusion
In this report, I conducted analysis on variables regarding pricing of vehicles, as well as physical characteristics of vehicles.
It was discovered that vehicles can be grouped into three classes: Lower expected profit, middle expected profit, and higher expected profit.
It was found there is a strong association betweeen a vehicle's size and horsepower, and a vehicle's expected profit.
A vehicle that is more powerful, that has a larger engine volume, that has larger physical dimensions, and that burns more gas, brings in a larger expected profit.
Without considering environment impact or 'Global Warming', Illini Auto is advised to expand their business on large vehicles, in order to generate more profit.
Statistical modeling also found linear correlation between expected profit and the following variables: MSRP, Weight, and Horsepower.
This confirms the suggestion derived from the previous model, that a bigger/heavier car is expected to generate more profit for Illini Auto.
In this report, I also conducted analysis on important factors that could help us classify a car's origin: domestic or import.
It was found that Engine Size, Highway MPG and Wheelbase are positive contributors to a car of being likely to be domestic.
In addition, Horsepower, City MPG and Weight are positive contributors to a car of being likely to be an import.
Illini Auto also asked for an analysis that could classify vehicles based on its cylinder number.
It was identified that City MPG, Highway MPG, Invoice, and Engine Size are effectively predictors of vehicles with different numbers of cylinders.
Part 2: Compare Expected Profit by Group defined by vehicle-related technical parameters
In this part, I aim to compare expected profit by groups of vehicles as defined by technical parameters.
First, I will perform cluster analysis based on engine size and power, fuel efficiency, and vehicle size measures. I will classify vehicles into different groups;
Second, I will compare expected profit by such groups.

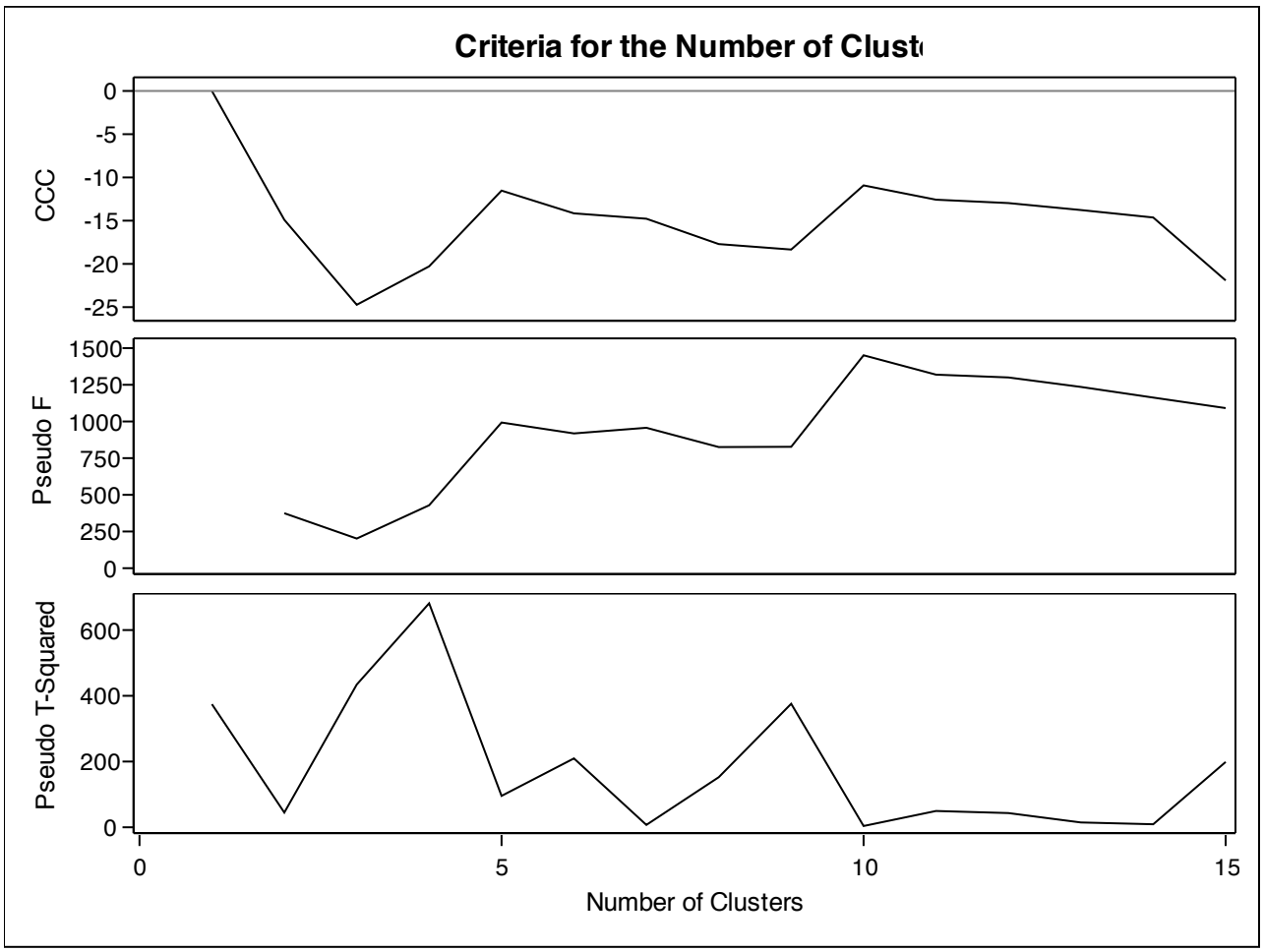
Eigenvalues of the Covariance Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	544928.865	542273.994	0.9949	0.9949
2	2654.871	2526.934	0.0048	0.9997
3	127.937	111.995	0.0002	1.0000
4	15.942	6.168	0.0000	1.0000
5	9.774	8.368	0.0000	1.0000
6	1.407	1.207	0.0000	1.0000
7	0.200		0.0000	1.0000

Root-Mean-Square Total-Sample Standard Deviation	279.7292

Root-Mean-Square Distance Between Observations	1046.651

Cluster History								
Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Approximate Expected R-Square	Cubic Clustering Criterion	Pseudo F Statistic
15	CL23	CL27	105	0.0064	.974	.991	-22	1091
14	CL24	OB267	26	0.0003	.974	.991	-15	1163
13	CL22	CL41	11	0.0006	.974	.990	-14	1235
12	CL26	CL142	17	0.0010	.973	.989	-13	1300
11	CL17	CL35	26	0.0023	.970	.987	-13	1319
10	CL15	OB330	106	0.0004	.970	.986	-11	1451
9	CL10	CL20	194	0.0277	.942	.983	-18	827
8	CL21	CL12	74	0.0080	.934	.980	-18	826
7	CL18	OB164	4	0.0005	.934	.976	-15	957
6	CL14	CL16	106	0.0154	.918	.968	-14	918
5	CL11	CL13	37	0.0118	.906	.956	-12	993
4	CL6	CL9	300	0.1486	.758	.934	-20	429
3	CL4	CL8	374	0.2630	.495	.886	-25	202
2	CL5	CL7	41	0.0193	.476	.747	-15	375
1	CL3	CL2	415	0.4756	.000	.000	0.00	.

Cluster History					
Number of Clusters	**Clusters Joined**		**Pseudo t-Squared**	**Norm RMS Distance**	**Tie**
15	CL23	CL27	199	0.2611	
14	CL24	OB267	9.2	0.2772	
13	CL22	CL41	14.7	0.2847	
12	CL26	CL142	43.2	0.3002	
11	CL17	CL35	49.6	0.3052	
10	CL15	OB330	3.9	0.3086	
9	CL10	CL20	376	0.3855	
8	CL21	CL12	152	0.3897	
7	CL18	OB164	7.2	0.4023	
6	CL14	CL16	210	0.4347	
5	CL11	CL13	95.5	0.6002	
4	CL6	CL9	681	0.7338	
3	CL4	CL8	434	1.0476	
2	CL5	CL7	44.8	1.1112	
1	CL3	CL2	375	1.7655	



CCC plot, Pseudo F and Pseudo T-squared plot all show 5 cluster is the optimal. Therefore, I will tentatively group all vehicles into 5 groups, subject to double-checking on group size.

CLUSTER				
1	2	3	4	5
N	N	N	N	N
194	74	106	37	4

There are too few vehicles in Cluster 5. Therefore, I'll remove those 2 vehicles form analysis.

CLUSTER			
1	**2**	**3**	**4**
N	**N**	**N**	**N**
194	74	106	37

Now that 4 groups are set up, the dataset is ready for some analysis.
Note that there are different numbers of vehicles in each group. Thus it is an unbalanced case of comparison of means.

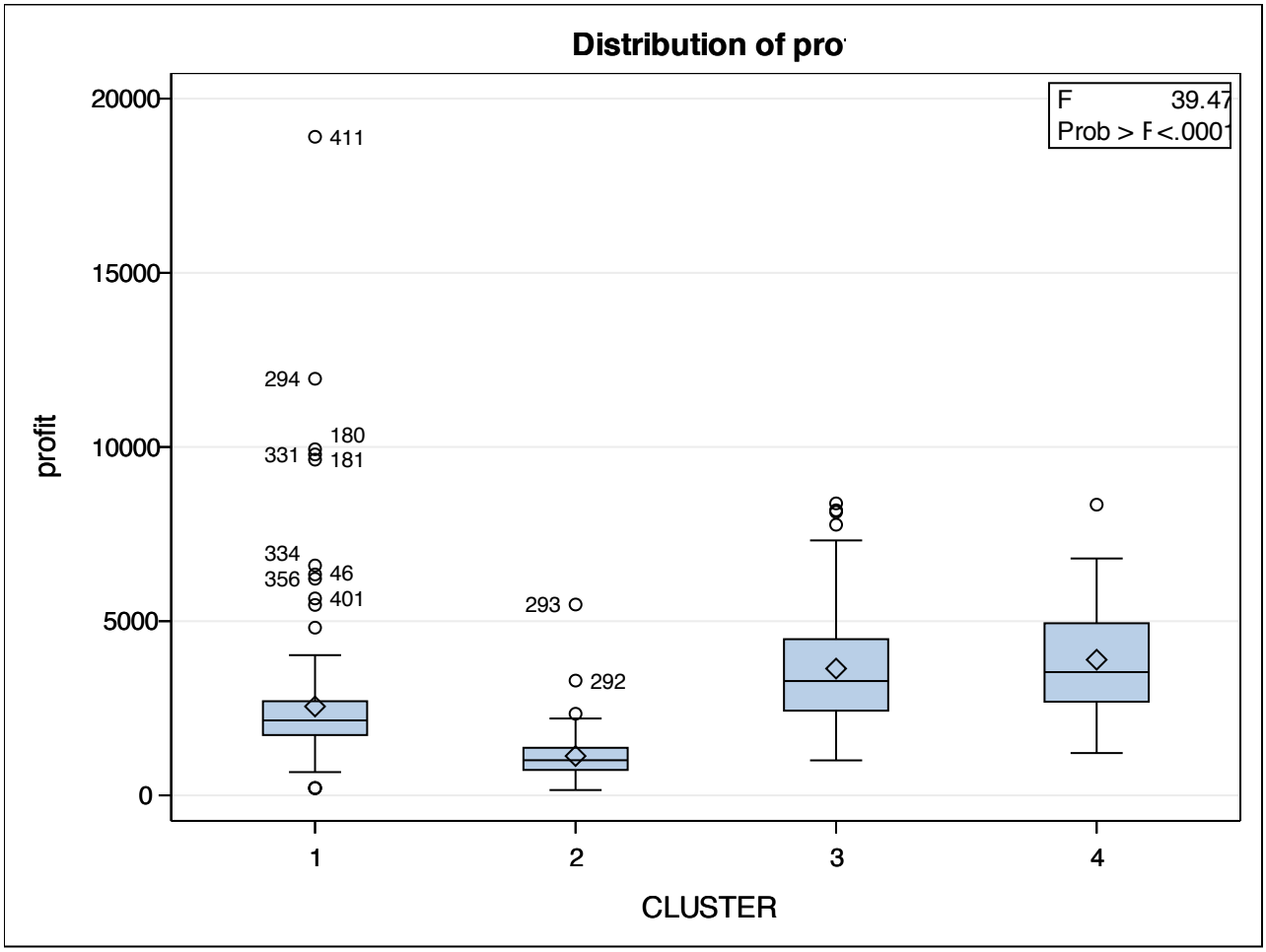
Dependent Variable: profit

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	333278778	111092926	39.47	<.0001
Error	407	1145455094	2814386		
Corrected Total	410	1478733873			

R-Square	Coeff Var	Root MSE	profit Mean
0.225381	62.19929	1677.613	2697.158

Source	DF	Type I SS	Mean Square	F Value	Pr > F
CLUSTER	3	333278778.4	111092926.1	39.47	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
CLUSTER	3	333278778.4	111092926.1	39.47	<.0001



Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

Least Squares Means for Effect CLUSTER				
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	1422.464893	831.165164	2013.764622
1	3	-1088.363256	-1611.076902	-565.649611
1	4	-1342.440513	-2118.797151	-566.083874
2	3	-2510.828149	-3166.406295	-1855.250003
2	4	-2764.905405	-3636.273743	-1893.537068
3	4	-254.077257	-1080.440957	572.286444

Analysis result shows our model explains 22.54% of the variation in Profit. The model tells us Group 3- and Group 4-vehicles have significantly higher expected profit than Group 1 vehicles, and Group 1 vehicles have significantly higher profit than Group 2 vehicles. There is no significantly difference in expected profit between Group 3 and Group 4.

Given the result, I will combine Group 3 and Group 4, and compare profit again.

CLUSTER		
1	2	3
N	N	N
194	74	143

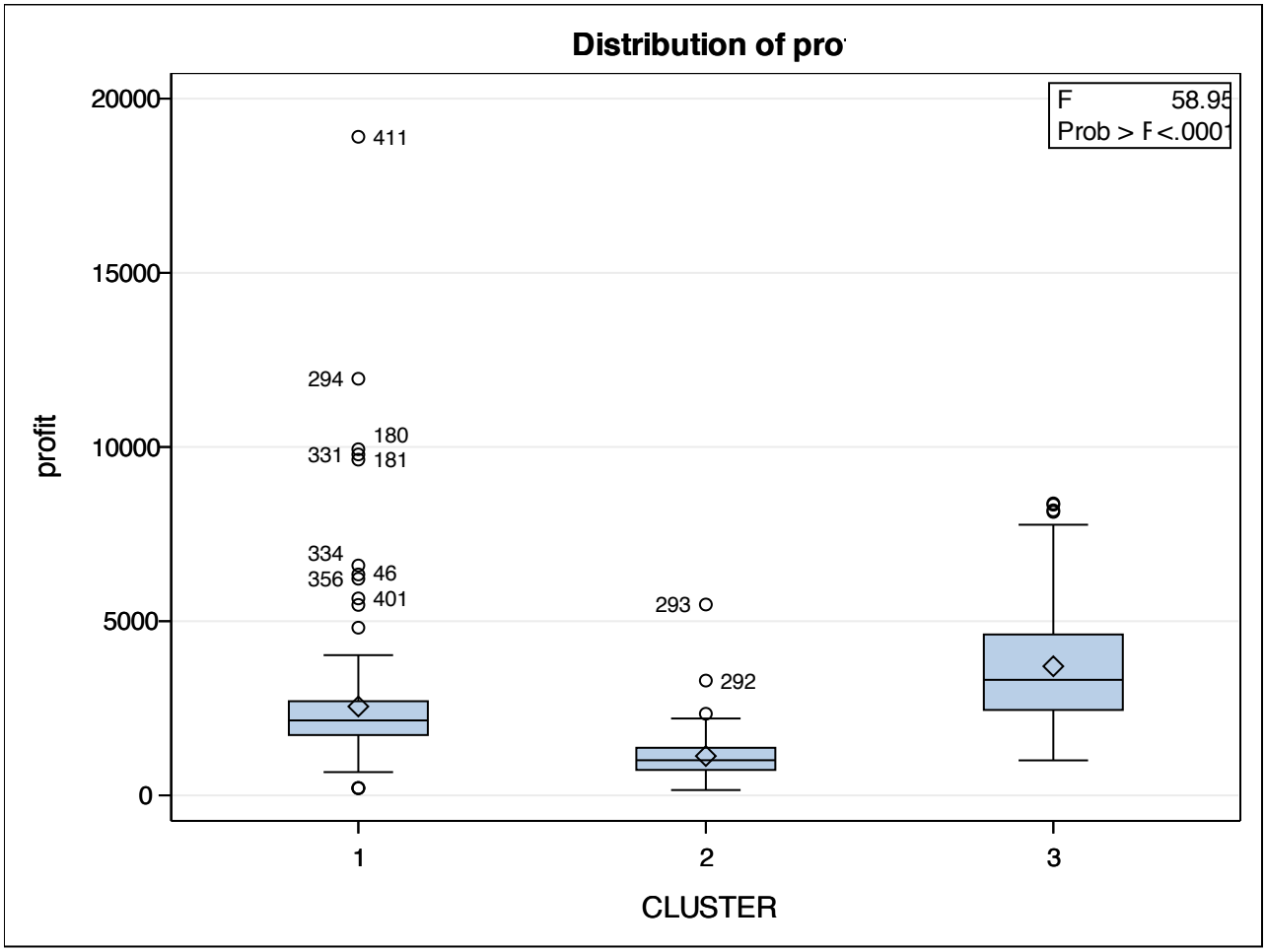
Dependent Variable: profit

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	331508249	165754125	58.95	<.0001
Error	408	1147225624	2811828		
Corrected Total	410	1478733873			

R-Square	Coeff Var	Root MSE	profit Mean
0.224184	62.17101	1676.850	2697.158

Source	DF	Type I SS	Mean Square	F Value	Pr > F
CLUSTER	2	331508249.0	165754124.5	58.95	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
CLUSTER	2	331508249.0	165754124.5	58.95	<.0001



Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

Least Squares Means for Effect CLUSTER				
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	1422.464893	883.528609	1961.401177
1	3	-1154.103525	-1588.846746	-719.360305
2	3	-2576.568418	-3141.417906	-2011.718930

Next, I will try to explore what characteristics result in the clustering, as well as the difference in expected profit across the 3 groups.
Principal Component analysis is conducted to check correlation.

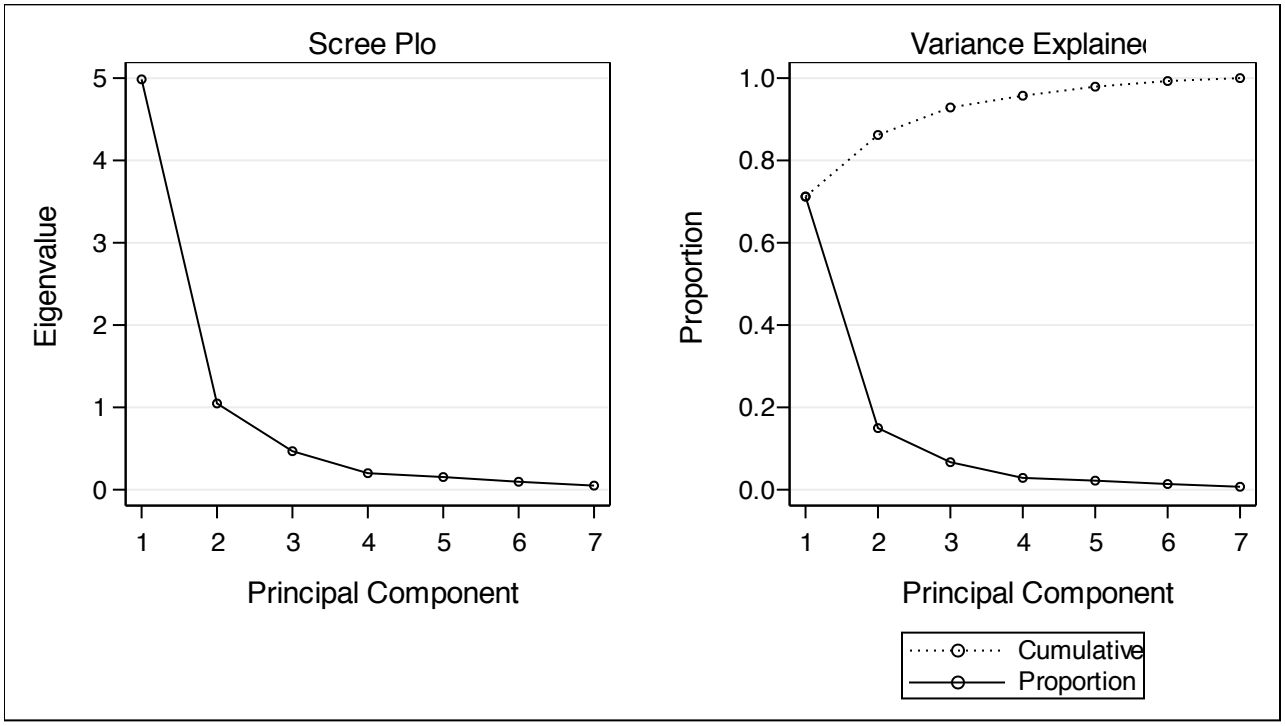
Observations	411
Variables	7

Simple Statistics							
	EngineSize	**Horsepower**	**MPG_City**	**MPG_Highway**	**Weight**	**Length**	**Wheelbase**
Mean	3.147688564	212.2335766	20.14111922	26.96107056	3539.119221	186.1338200	107.9659367
StD	1.026770101	66.9439805	4.82165215	5.32215277	696.161257	14.1436567	8.0990097

Correlation Matrix								
		EngineSize	**Horsepower**	**MPG_City**	**MPG_Highway**	**Weight**	**Length**	**Wheelbase**
EngineSize	Engine Size (L)	1.0000	0.7746	-.7365	-.7323	0.8181	0.6530	0.6479
Horsepower		0.7746	1.0000	-.6848	-.6453	0.6372	0.3831	0.3933
MPG_City	MPG (City)	-.7365	-.6848	1.0000	0.9307	-.7563	-.4908	-.5045
MPG_Highway	MPG (Highway)	-.7323	-.6453	0.9307	1.0000	-.7997	-.4417	-.5102
Weight	Weight (LBS)	0.8181	0.6372	-.7563	-.7997	1.0000	0.6818	0.7467
Length	Length (IN)	0.6530	0.3831	-.4908	-.4417	0.6818	1.0000	0.8850
Wheelbase	Wheelbase (IN)	0.6479	0.3933	-.5045	-.5102	0.7467	0.8850	1.0000

Eigenvalues of the Correlation Matrix				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	4.98507188	3.93759821	0.7122	0.7122
2	1.04747367	0.57973866	0.1496	0.8618
3	0.46773501	0.26731411	0.0668	0.9286
4	0.20042090	0.04684858	0.0286	0.9572
5	0.15357232	0.05735526	0.0219	0.9792
6	0.09621706	0.04670790	0.0137	0.9929
7	0.04950917		0.0071	1.0000

Eigenvectors								
		Prin1	**Prin2**	**Prin3**	**Prin4**	**Prin5**	**Prin6**	**Prin7**
EngineSize	Engine Size (L)	0.408449	-.060394	0.357087	0.262394	-.728473	-.310591	0.077592
Horsepower		0.343198	-.374815	0.693064	-.219592	0.444744	0.066201	-.104832
MPG_City	MPG (City)	-.390401	0.325802	0.379815	0.453289	0.076160	-.089342	-.614758
MPG_Highway	MPG (Highway)	-.387966	0.334192	0.474421	0.002860	-.025556	0.203966	0.685900
Weight	Weight (LBS)	0.414986	0.047511	-.127047	0.705300	0.232607	0.481210	0.162083
Length	Length (IN)	0.340340	0.578527	0.060675	-.424226	-.218387	0.488001	-.282822
Wheelbase	Wheelbase (IN)	0.352717	0.548711	-.054713	-.005799	0.404076	-.616327	0.168358

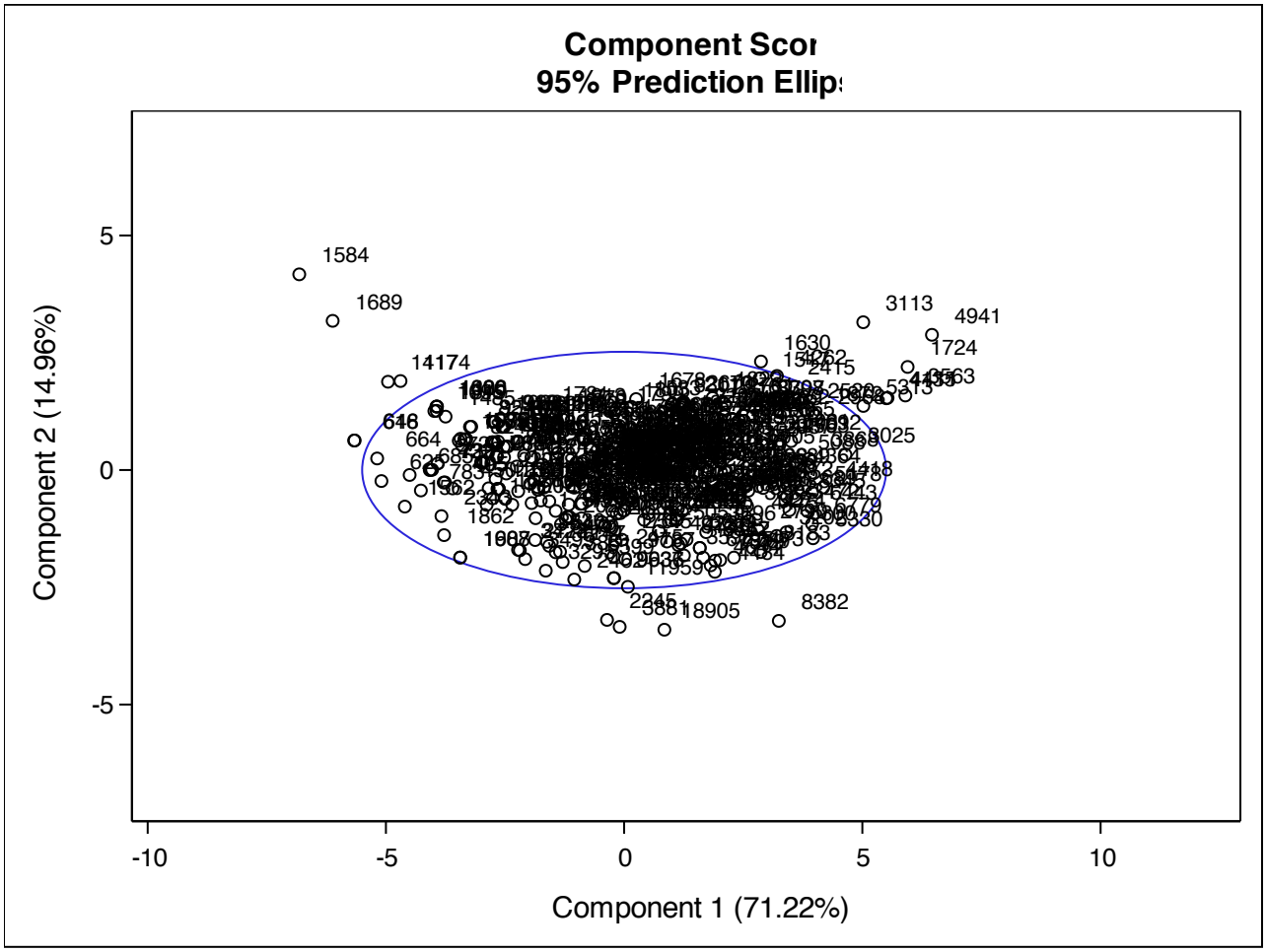


Correlation matrix shows clearly that Engine Size, horsepower, and vehicle size measure variables are positively correlated. City MPG and Highway MPG are positively correlation. Both City MPG and Highway MPG are negatively correlated with each of the variables related to engine size and vehicle size.

This result is consistent with our expectation based on anecdotal experience: larger vehicles and/or more powerful vehicles usually have lower fuel efficiency.

This result is also reflected in the Principal Component 1, which picks out positive values on all variables but City MPG and Highway MPG.

The second Principal Component picks out positive values on all but Engine Size and Horsepower, which are the only two indicators of vehicle power. All other variables make positive contribution to Principal Component 2.



The scoreplot, while difficult to read, seems to indicate vehicles with a larger profit tends to have a larger PC1 value. I begin to suspect larger vehicles / more powerful vehicles generate more profit than other vehicles.
I will have SAS conduct proportional-priors discriminant analysis to confirm my theory.

Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
743.219296	56	<.0001

Since the Chi-Square value is significant at the 0.1 level, the within covariance matrices will be used in the discriminant function.
Reference: Morrison, D.F. (1976) Multivariate Statistical Methods p252.

Multivariate Statistics and F Approximations					
S=2 M=2 N=200					
Statistic	**Value**	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Wilks' Lambda	0.18450402	76.27	14	804	<.0001
Pillai's Trace	0.91689969	48.74	14	806	<.0001
Hotelling-Lawley Trace	3.87033463	110.93	14	639.85	<.0001
Roy's Greatest Root	3.72269939	214.32	7	403	<.0001
NOTE: F Statistic for Roy's Greatest Root is an upper bound.					
NOTE: F Statistic for Wilks' Lambda is exact.					

Classification Results for Calibration Data: WORK.TREE_OUT
Cross-validation Results using Quadratic Discriminant Function

Posterior Probability of Membership in CLUSTER						
Obs	From CLUSTER	Classified into CLUSTER		1	2	3
92	**1**	**3**	*	0.4966	0.0000	0.5034
171	**1**	**2**	*	0.0901	0.9098	0.0001
180	**1**	**3**	*	0.1389	0.0000	0.8611
181	**1**	**3**	*	0.1423	0.0000	0.8577
204	**1**	**3**	*	0.3905	0.0000	0.6095
207	**1**	**3**	*	0.3163	0.0000	0.6837
208	**1**	**3**	*	0.4576	0.0000	0.5424
240	**1**	**3**	*	0.2723	0.0000	0.7277
262	**3**	**1**	*	0.5844	0.0000	0.4156
271	**3**	**1**	*	0.7148	0.0000	0.2852
279	**1**	**3**	*	0.0001	0.0000	0.9999
282	**1**	**3**	*	0.3206	0.0000	0.6794
292	**2**	**1**	*	0.9823	0.0002	0.0175
293	**2**	**1**	*	0.7561	0.0000	0.2438
294	**1**	**3**	*	0.1624	0.0000	0.8376
310	**3**	**1**	*	0.8422	0.0000	0.1578
314	**3**	**1**	*	0.6769	0.0000	0.3231
324	**1**	**3**	*	0.0739	0.0000	0.9261
356	**1**	**3**	*	0.4078	0.0000	0.5922
357	**1**	**3**	*	0.2539	0.0000	0.7461
368	**1**	**2**	*	0.3753	0.6247	0.0000
369	**1**	**2**	*	0.0000	1.0000	0.0000
379	**2**	**1**	*	0.9084	0.0819	0.0097
383	**3**	**1**	*	0.5703	0.0000	0.4297
389	**2**	**1**	*	0.5085	0.4913	0.0002
398	**2**	**1**	*	1.0000	0.0000	0.0000
411	**1**	**3**	*	0.0000	0.0000	1.0000

**** Misclassified observation***

Classification Summary for Calibration Data: WORK.TREE_OUT
Cross-validation Summary using Quadratic Discriminant Function

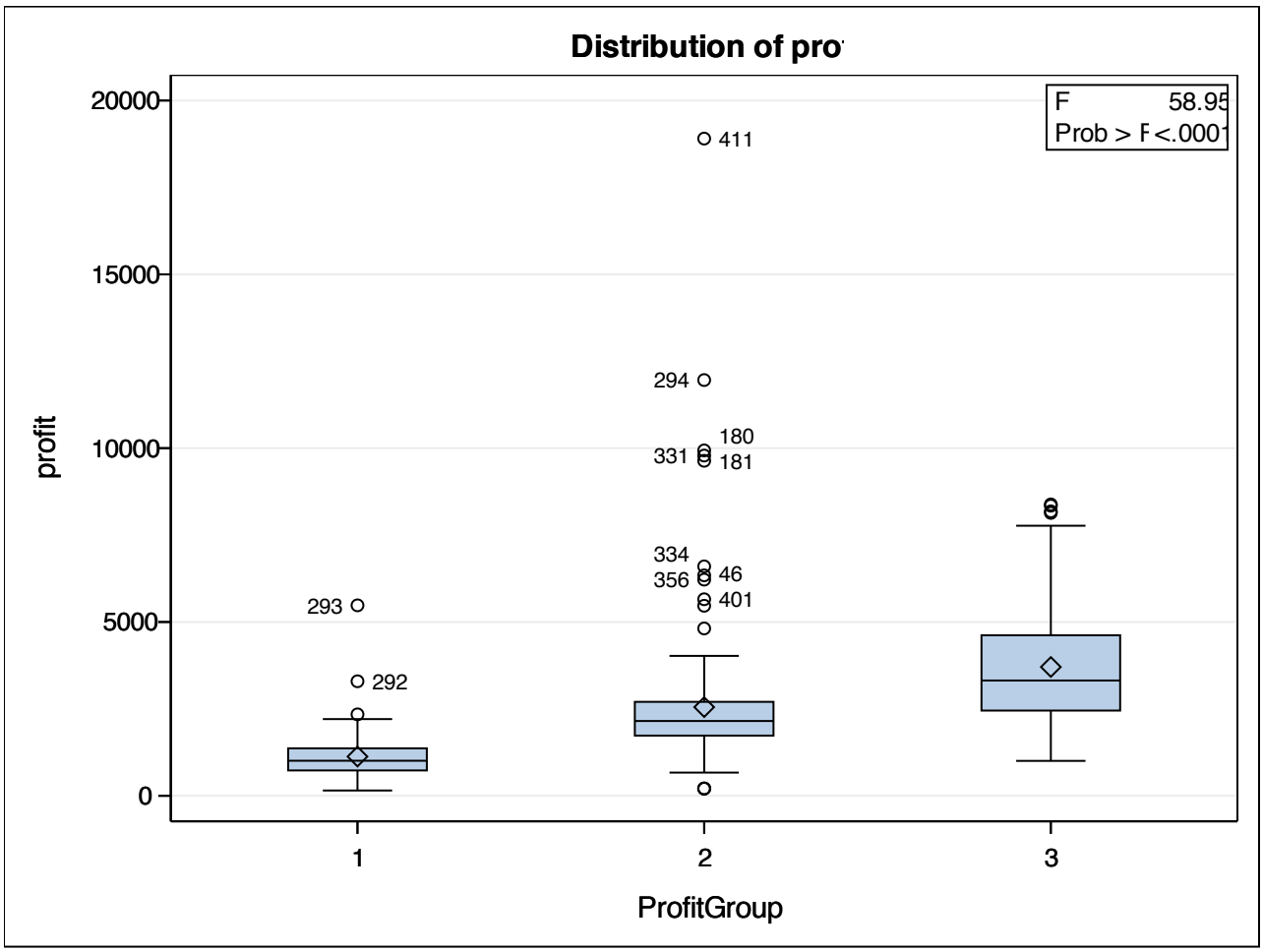
Number of Observations and Percent Classified into CLUSTER				
From CLUSTER	**1**	**2**	**3**	**Total**
1	177 91.24	3 1.55	14 7.22	194 100.00
2	5 6.76	69 93.24	0 0.00	74 100.00
3	5 3.50	0 0.00	138 96.50	143 100.00
Total	187 45.50	72 17.52	152 36.98	411 100.00
Priors	0.47202	0.18005	0.34793	

Error Count Estimates for CLUSTER				
	1	**2**	**3**	**Total**
Rate	0.0876	0.0676	0.0350	0.0657
Priors	0.4720	0.1800	0.3479	

Classification table shows an exceptionally high success rate when classifying vehicles into the correct category. This means my speculation, that vehicle profit is somewhat related to its size and gas-guzzling ability, might be true. However, I will have SAS perform yet another test to further confirm this speculation.
I will conduct Cumulative Logistic distribution.
To conduct cumulative logistic distribution, I will first re-code cluster based on profit. By ascending order of profit, clusters are ordered as Cluster 2, Cluster 1, Cluster 3. Therefore, Cluster 2 is coded into Profit group 1, Cluster 1 are recoded into Profit group 2, and Cluster 3 are recoded into Profit group 3.

Table of CLUSTER by ProfitGroup				
CLUSTER	**ProfitGroup**			
Frequency	**1**	**2**	**3**	**Total**
1	0	194	0	194
2	74	0	0	74
3	0	0	143	143
Total	74	194	143	411

Dependent Variable: profit



Model Information	
Data Set	WORK.TREE_OUT
Distribution	Multinomial
Link Function	Cumulative Logit
Dependent Variable	ProfitGroup

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept1	1	4254.679	3.2534E8	-6.377E8	6.3767E8	0.00	1.0000
Intercept2	1	5541.327	4.2032E8	-8.238E8	8.2382E8	0.00	1.0000
EngineSize	1	-6.9929	11895403	-2.331E7	23314554	0.00	1.0000
Horsepower	1	0.0474	151390.8	-296720	296720.6	0.00	1.0000
MPG_City	1	0.7670	2251493	-4412844	4412846	0.00	1.0000
MPG_Highway	1	-2.7791	3054598	-5986905	5986899	0.00	1.0000
Weight	1	-1.5359	116411.9	-228165	228161.5	0.00	1.0000
Length	1	2.1386	1473237	-2887489	2887493	0.00	1.0000
Wheelbase	1	-1.1600	1496121	-2932344	2932341	0.00	1.0000
Scale	0	1.0000	0.0000	1.0000	1.0000		

Note: The scale parameter was held fixed.

LR Statistics For Type 3 Analysis			
Source	**DF**	**Chi-Square**	**Pr > ChiSq**
EngineSize	1	0.00	1.0000
Horsepower	1	0.00	1.0000
MPG_City	1	0.00	1.0000
MPG_Highway	1	0.00	1.0000
Weight	1	321.91	<.0001
Length	1	0.00	1.0000
Wheelbase	1	0.00	1.0000

SAS returned ridiculously large/small standard errors. Therefore, I conclude (quasi-)complete separation occurred, indicating (almost) perfect classification occurred.

Part 3: Statistical Model Construction to Predict Expected Profit

LR Statistics For Type 3 Analysis			
Source	**DF**	**Chi-Square**	**Pr > ChiSq**

I identified 13 explanantory variables (Type, Origin, DriveTrain, MSRP, Invoice, EngineSize, Cylinders, Horsepower, MPG_City, MPG_Highway, Weight, Wheelbase, and Length).Before conducting any analysis, we know that MSRP and Invoice cannot be in the model at the same time. The first step to construct a linear regression model is to construct 13 main-effect-only models, each with one explanatory variable, and eliminate those not significant.

Dependent Variable: profit

R-Square	Coeff Var	Root MSE	profit Mean
0.171311	64.20881	1747.241	2721.186

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Type	5	258120236.1	51624047.2	16.91	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Type	5	258120236.1	51624047.2	16.91	<.0001

Variable 'Type' is significant at the .0001 level. However, its R-square is only .1713. Therefore, it will be eliminated.

Dependent Variable: profit

R-Square	Coeff Var	Root MSE	profit Mean
0.161790	64.34107	1750.840	2721.186

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Origin	2	243774958.6	121887479.3	39.76	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Origin	2	243774958.6	121887479.3	39.76	<.0001

Variable 'Origin' is significant at the .0001 level. However, its R-square is only .1618. Therefore, it will be eliminated.

Dependent Variable: profit

R-Square	Coeff Var	Root MSE	profit Mean
0.206218	62.61269	1703.807	2721.186

Source	DF	Type I SS	Mean Square	F Value	Pr > F
DriveTrain	2	310716769.9	155358384.9	53.52	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
DriveTrain	2	310716769.9	155358384.9	53.52	<.0001

Variable 'DriveTrain' is significant at the .0001 level. However, its R-square is only .2062. Therefore, it will be eliminated.

Model: MODEL1
Dependent Variable: profit

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	1304053349	1304053349	2657.22	<.0001
Error	413	202682922	490758		
Corrected Total	414	1506736271			

Root MSE	700.54099	**R-Square**	0.8655
Dependent Mean	2721.18554	**Adj R-Sq**	0.8652
Coeff Var	25.74396		

Variable 'MSRP' will not be eliminated in next step.

Model: MODEL1
Dependent Variable: profit

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	1258573609	1258573609	2094.56	<.0001
Error	413	248162661	600878		
Corrected Total	414	1506736271			

Root MSE	775.16328	**R-Square**	0.8353
Dependent Mean	2721.18554	**Adj R-Sq**	0.8349
Coeff Var	28.48623		

Variable 'Invoice' is significant. However, because it cannot be in the model along with MSRP, and it has a smaller F statistic, it will be eliminated.

Model: MODEL1
Dependent Variable: profit

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	481638582	481638582	194.05	<.0001
Error	413	1025097689	2482077		
Corrected Total	414	1506736271			

Root MSE	1575.46080	**R-Square**	0.3197
Dependent Mean	2721.18554	**Adj R-Sq**	0.3180
Coeff Var	57.89612		

Variable 'EngineSize' will not be eliminated in next step.

Dependent Variable: profit

R-Square	Coeff Var	Root MSE	profit Mean
0.367374	55.86765	1522.312	2724.855

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Cylinders	1	553110090.1	553110090.1	238.67	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Cylinders	1	553110090.1	553110090.1	238.67	<.0001

Variable 'Cylinders' will not be eliminated in next step.

Model: MODEL1
Dependent Variable: profit

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	909130299	909130299	628.29	<.0001
Error	413	597605972	1446988		
Corrected Total	414	1506736271			

Root MSE	1202.90807	**R-Square**	0.6034
Dependent Mean	2721.18554	**Adj R-Sq**	0.6024
Coeff Var	44.20529		

Variable 'HorsePower' will not be eliminated in next step.

Model: MODEL1
Dependent Variable: profit

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	350048791	350048791	124.99	<.0001
Error	413	1156687480	2800696		
Corrected Total	414	1506736271			

Root MSE	1673.52803	**R-Square**	0.2323
Dependent Mean	2721.18554	**Adj R-Sq**	0.2305
Coeff Var	61.49996		

Variable 'MPG_City' will not be eliminated in next step.

Model: MODEL1
Dependent Variable: profit

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	299079860	299079860	102.28	<.0001
Error	413	1207656411	2924108		
Corrected Total	414	1506736271			

Root MSE	1710.00220	**R-Square**	0.1985
Dependent Mean	2721.18554	**Adj R-Sq**	0.1966
Coeff Var	62.84034		

Variable 'MPG_Highway' will not be eliminated in next step.

Model: MODEL1
Dependent Variable: profit

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	322239519	322239519	112.36	<.0001
Error	413	1184496751	2868031		
Corrected Total	414	1506736271			

Root MSE	1693.52617	**R-Square**	0.2139
Dependent Mean	2721.18554	**Adj R-Sq**	0.2120
Coeff Var	62.23487		

Variable 'Weight' will not be eliminated in next step.

Model: MODEL1
Dependent Variable: profit

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	45003895	45003895	12.72	0.0004
Error	413	1461732376	3539304		
Corrected Total	414	1506736271			

Root MSE	1881.30369	**R-Square**	0.0299
Dependent Mean	2721.18554	**Adj R-Sq**	0.0275
Coeff Var	69.13544		

Variable 'WheelBase' is significant at the .0005 level. However, it's R-square is merely 0.03. Therefore, it will be eliminated in next step.

Model: MODEL1
Dependent Variable: profit

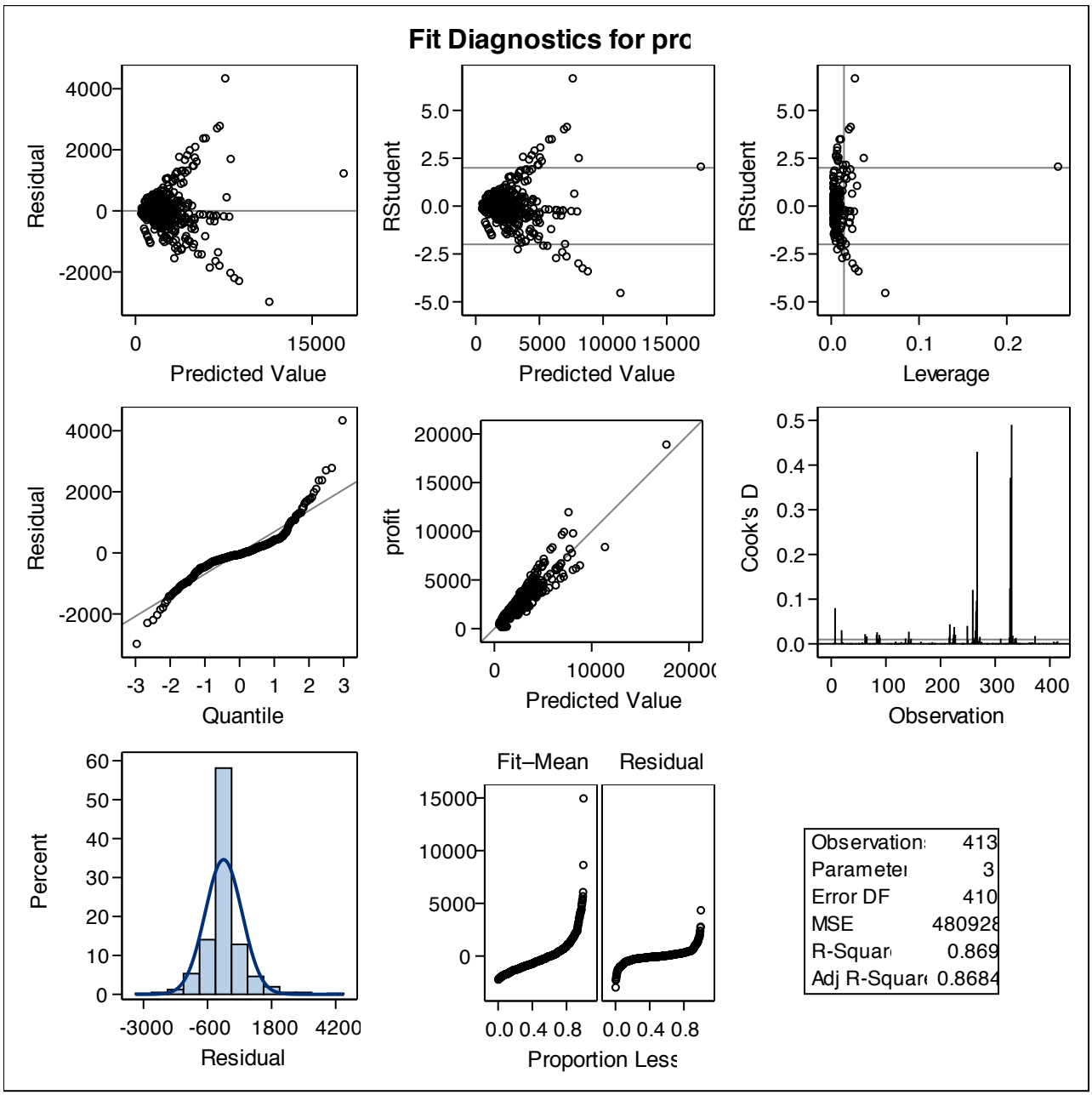
Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	62663244	62663244	17.92	<.0001
Error	413	1444073026	3496545		
Corrected Total	414	1506736271			

Root MSE	1869.90504	**R-Square**	0.0416
Dependent Mean	2721.18554	**Adj R-Sq**	0.0393
Coeff Var	68.71656		

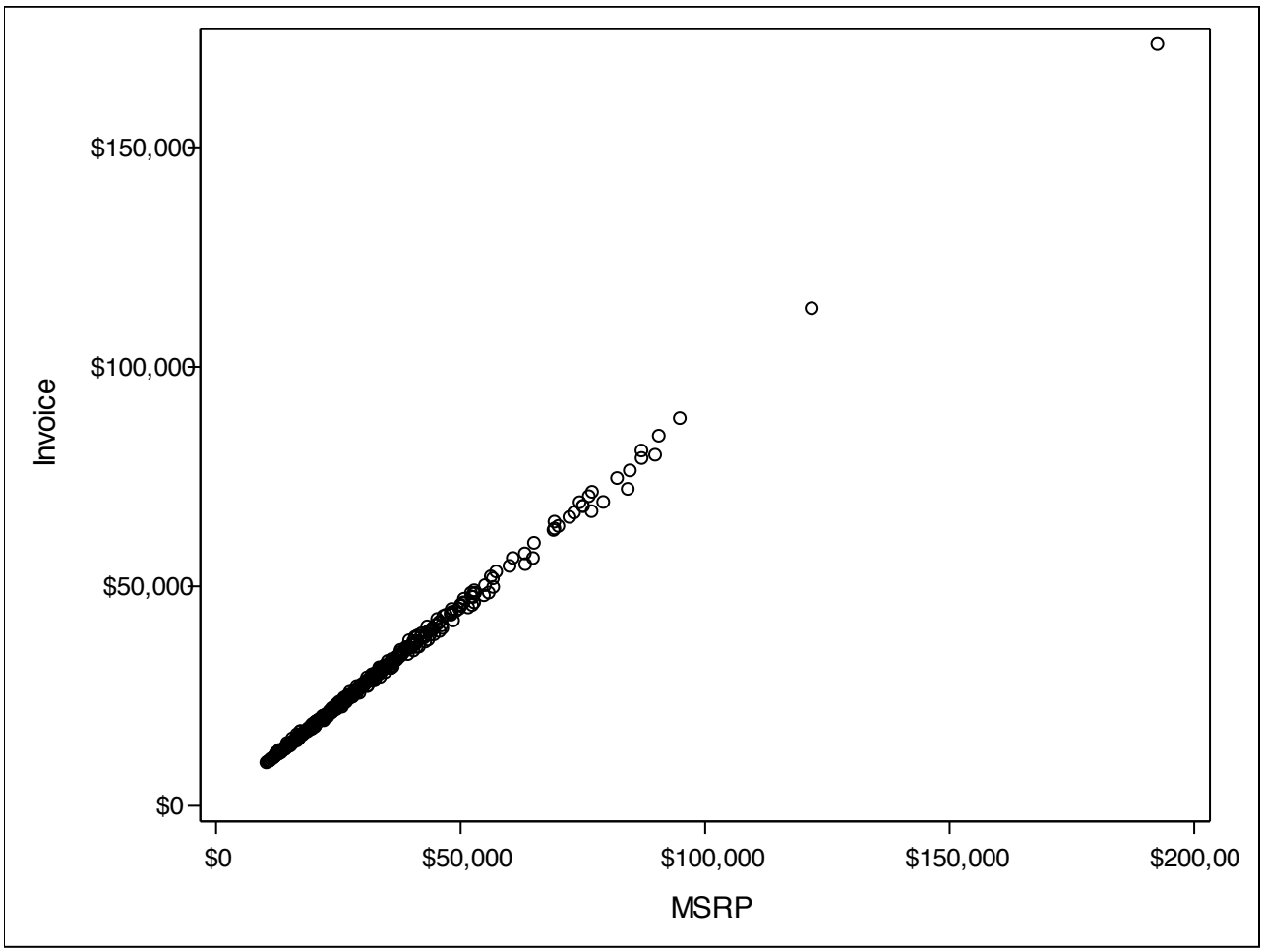
Variable 'Length' is significant at the .0001 level. However, it's R-square is merely 0.04. Therefore, it will be eliminated in next step.

In the next step, I will put all explanatory variables that were not previously eliminated into a linear model, and have SAS perform forward selection.

Model: MODEL1
Dependent Variable: profit



Diagnostics panel indicates violation of Normality assumption and Constant Variance assumption. Therefore, I will transform Profit by taking its natural logarithm, and refit the model.



Note that after transforming Profit, there is no longer perfect linear relationship among MSRP, Invoice, and sqrt(profit). However, it is found that MSRP and Invoice are highly correlated. Therefore, I will only keep MSRP in the model

Model: MODEL1
Dependent Variable: logprofit

Number of Observations Read	415
Number of Observations Used	413
Number of Observations with Missing Values	2

Forward Selection: Step 1

Variable MSRP Entered: R-Square = 0.6305 and C(p) = 143.2397

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	121.09397	121.09397	701.32	<.0001
Error	411	70.96583	0.17267		
Corrected Total	412	192.05980			

Variable	**Parameter Estimate**	**Standard Error**	**Type II SS**	**F Value**	**Pr > F**
Intercept	6.75053	0.04125	4624.50047	26782.9	<.0001
MSRP	0.00002959	0.00000112	121.09397	701.32	<.0001

Bounds on condition number: 1, 1

Forward Selection: Step 2

Variable Weight Entered: R-Square = 0.7011 and C(p) = 39.7095

Model: MODEL1
Dependent Variable: logprofit

Forward Selection: Step 2

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	134.65517	67.32759	480.87	<.0001
Error	410	57.40463	0.14001		
Corrected Total	412	192.05980			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	5.92795	0.09146	588.12551	4200.56	<.0001
MSRP	0.00002455	0.00000113	66.21210	472.91	<.0001
Weight	0.00027596	0.00002804	13.56120	96.86	<.0001

Bounds on condition number: 1.259, 5.036

Forward Selection: Step 3

Variable Horsepower Entered: R-Square = 0.7150 and C(p) = 20.9919

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	137.31750	45.77250	341.98	<.0001
Error	409	54.74230	0.13384		
Corrected Total	412	192.05980			

Model: MODEL1
Dependent Variable: logprofit

Forward Selection: Step 3

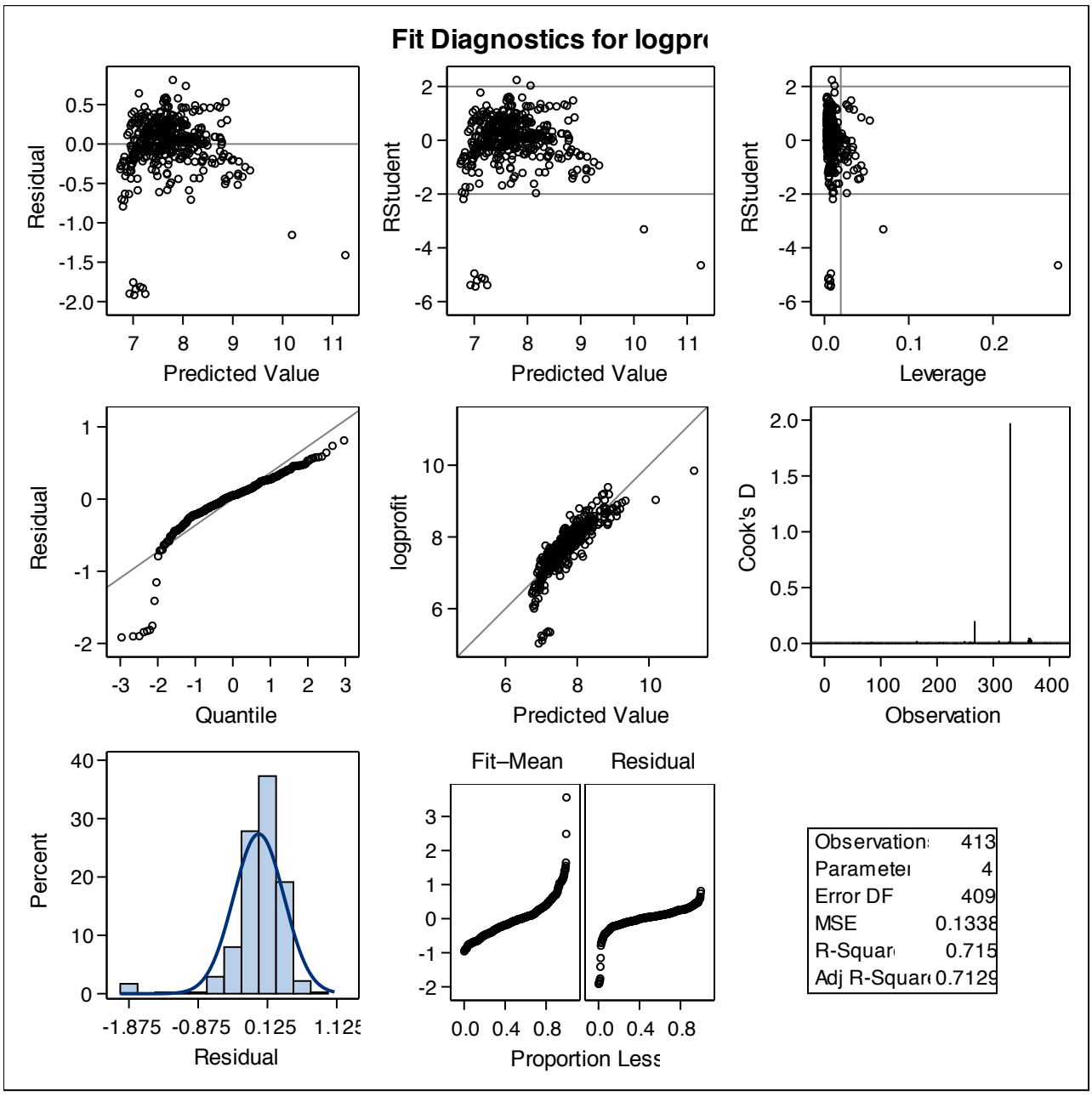
Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	5.87642	0.09017	568.45764	4247.16	<.0001
MSRP	0.00001878	0.00000170	16.32726	121.99	<.0001
Horsepower	0.00240	0.00053790	2.66233	19.89	<.0001
Weight	0.00019881	0.00003242	5.03455	37.61	<.0001

Bounds on condition number: 4.0746, 26.468

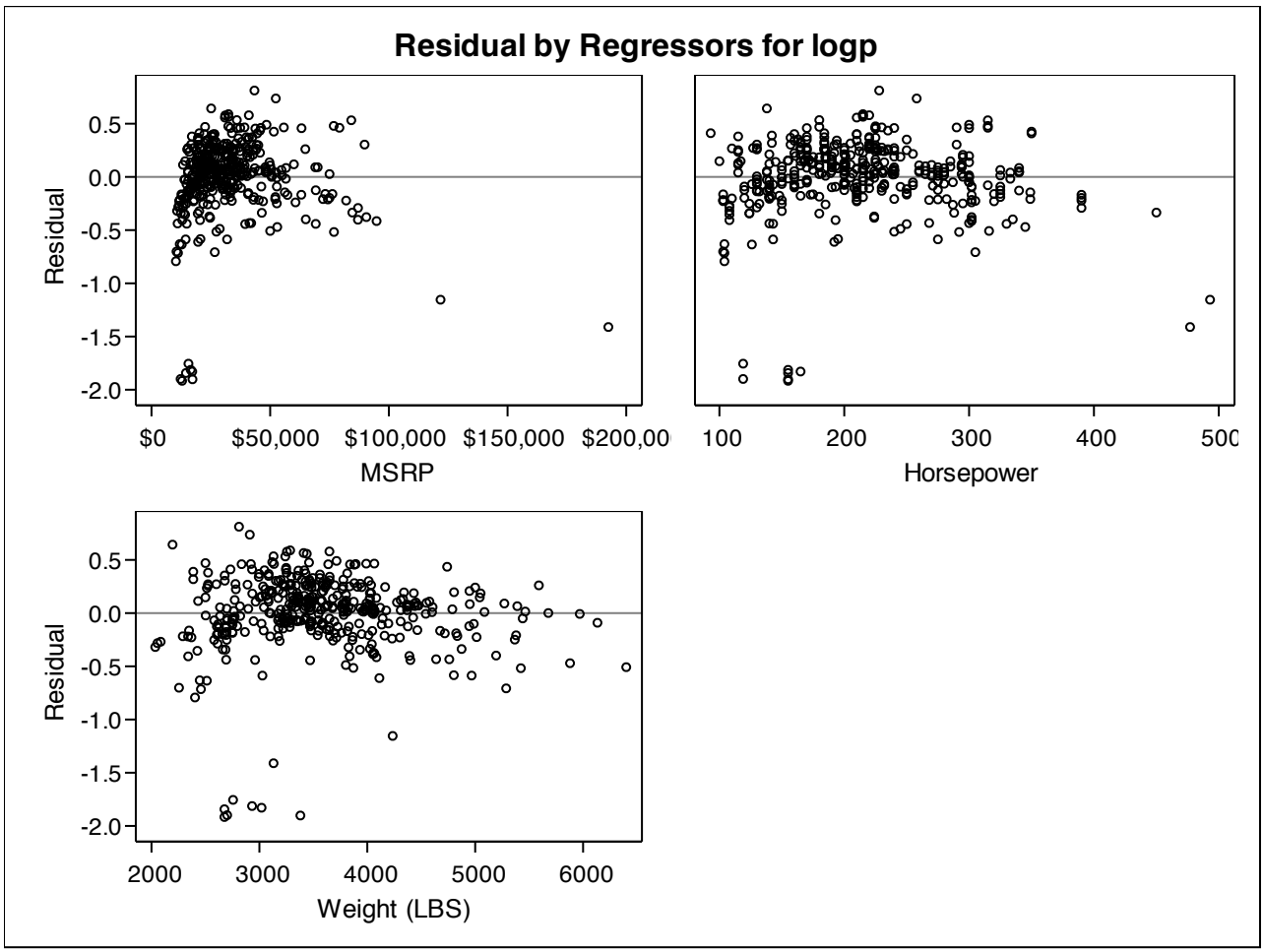
No other variable met the 0.0500 significance level for entry into the model.

Summary of Forward Selection								
Step	**Variable Entered**	**Label**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
1	MSRP		1	0.6305	0.6305	143.240	701.32	<.0001
2	Weight	Weight (LBS)	2	0.0706	0.7011	39.7095	96.86	<.0001
3	Horsepower		3	0.0139	0.7150	20.9919	19.89	<.0001

Model: MODEL1
Dependent Variable: logprofit



Model: MODEL1
Dependent Variable: logprofit



Selection process picks out MSRP, Weight, and Horsepower
Residual plot shows several observations have exceptionally large residual. I will find out which they are.

Top 10 Vehicles with Largest Residuals

Obs	Make	Model	MSRP	Invoice	residual	residualabs
1	Suzuki	Aeno S 4dr	$12,884	$12,719	-1.91721	1.91721
2	Suzuki	Verona LX 4dr	$17,262	$17,053	-1.90242	1.90242
3	Suzuki	Forenza S 4dr	$12,269	$12,116	-1.89942	1.89942
4	Suzuki	Aerio LX 4dr	$14,500	$14,317	-1.84398	1.84398
5	Suzuki	Vitara LX	$17,163	$16,949	-1.82969	1.82969
6	Suzuki	Aerio SX	$16,497	$16,291	-1.81377	1.81377
7	Suzuki	Forenza EX 4dr	$15,568	$15,378	-1.75562	1.75562
8	Porsche	911 GT2 2dr	$192,465	$173,560	-1.41023	1.41023
9	Mercedes-Benz	SL55 AMG 2dr	$121,770	$113,388	-1.15476	1.15476
10	Porsche	Boxster convertible 2dr	$43,365	$37,886	0.81119	0.81119

It turns out the 9 outlying residuals are from Suzuki vehicles and a Porsche. I will remove these 9 vehicles from analysis.

Model: MODEL1
Dependent Variable: logprofit

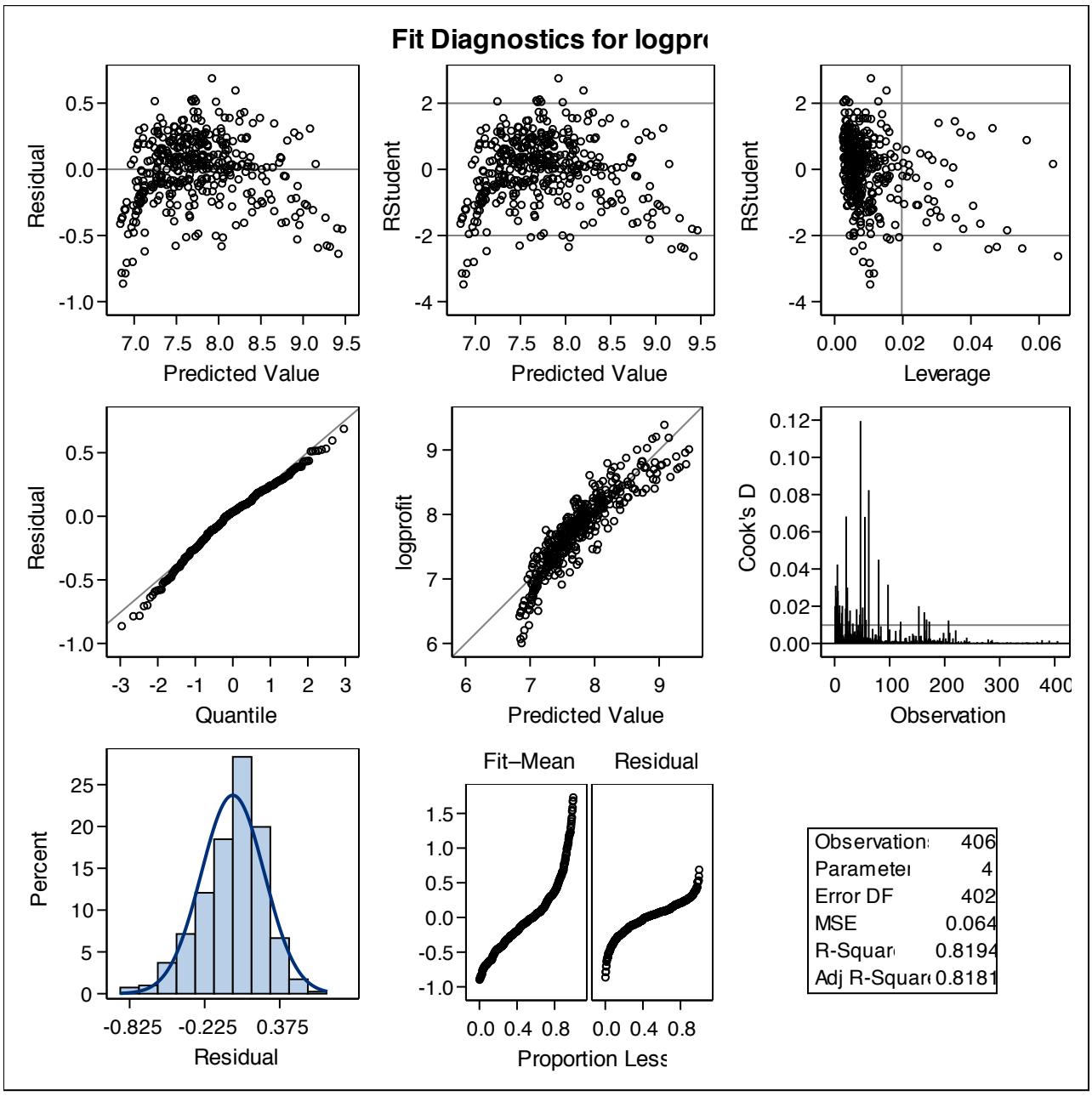
Number of Observations Read	406
Number of Observations Used	406

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	116.81415	38.93805	608.00	<.0001
Error	402	25.74537	0.06404		
Corrected Total	405	142.55952			

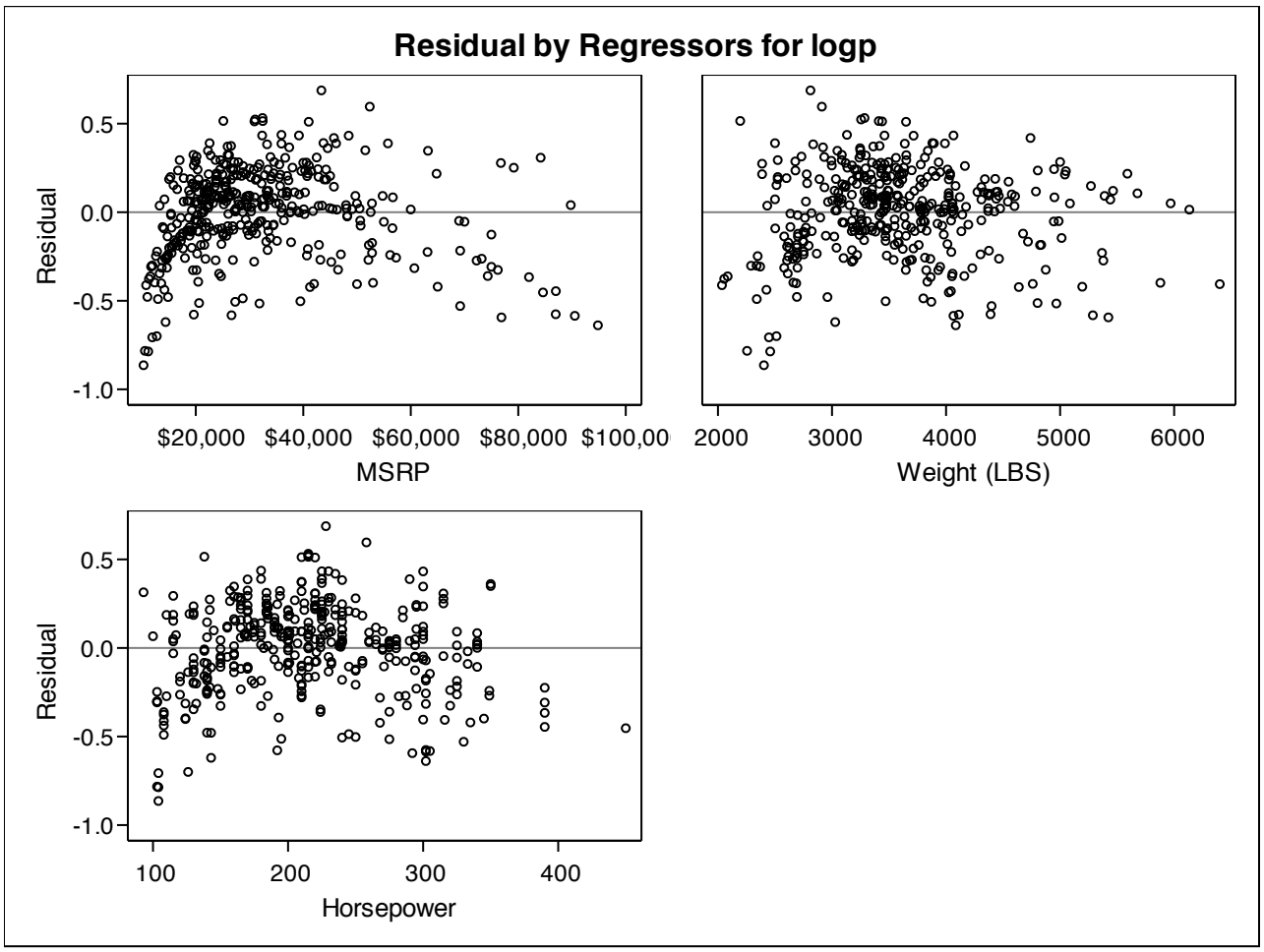
Root MSE	0.25307	**R-Square**	0.8194
Dependent Mean	7.73261	**Adj R-Sq**	0.8181
Coeff Var	3.27273		

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	6.09590	0.06342	96.12	<.0001
MSRP		1	0.00002284	0.00000135	16.95	<.0001
Weight	Weight (LBS)	1	0.00013958	0.00002283	6.11	<.0001
Horsepower		1	0.00194	0.00037829	5.13	<.0001

Model: MODEL1
Dependent Variable: logprofit



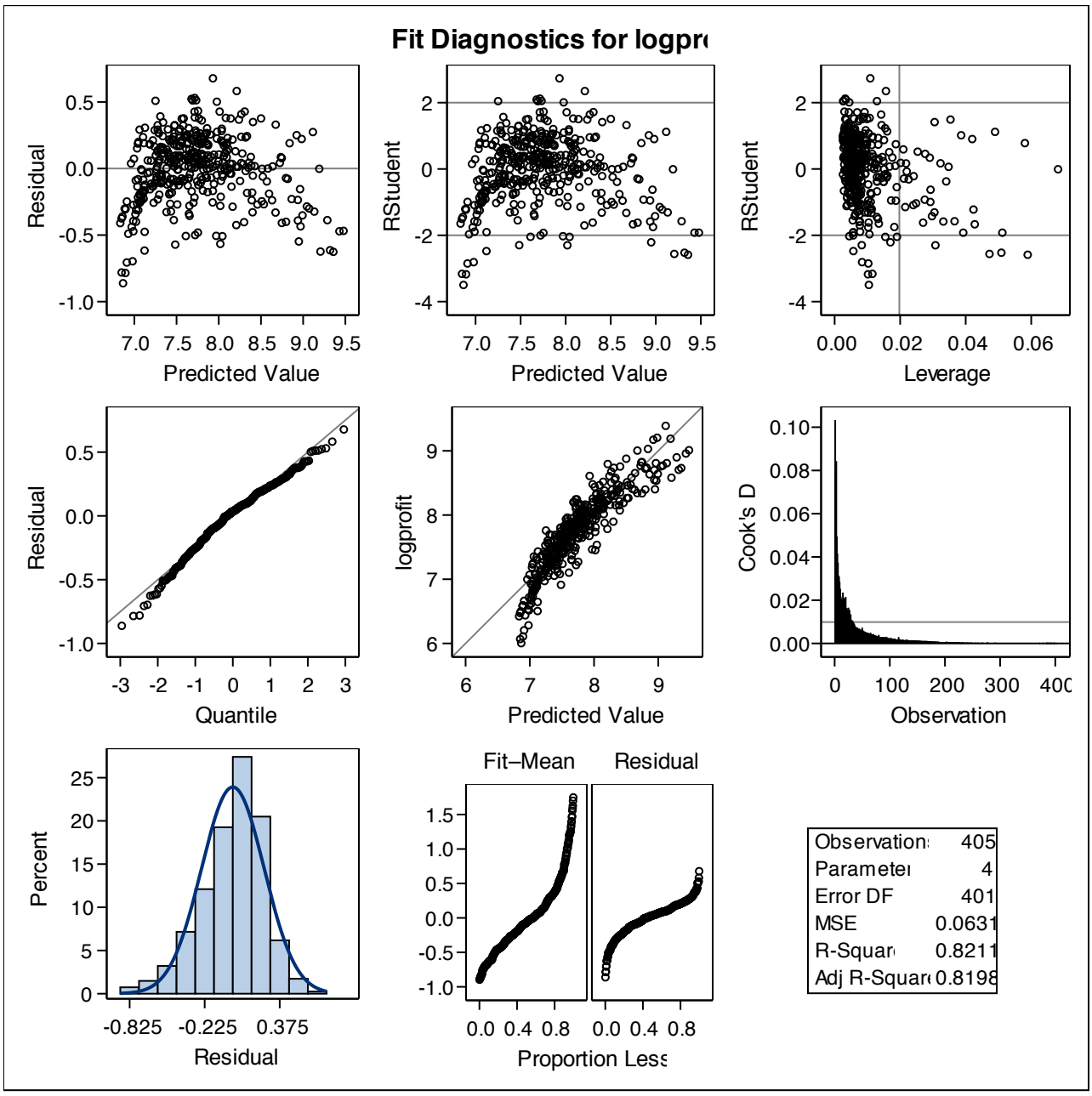
Model: MODEL1
Dependent Variable: logprofit



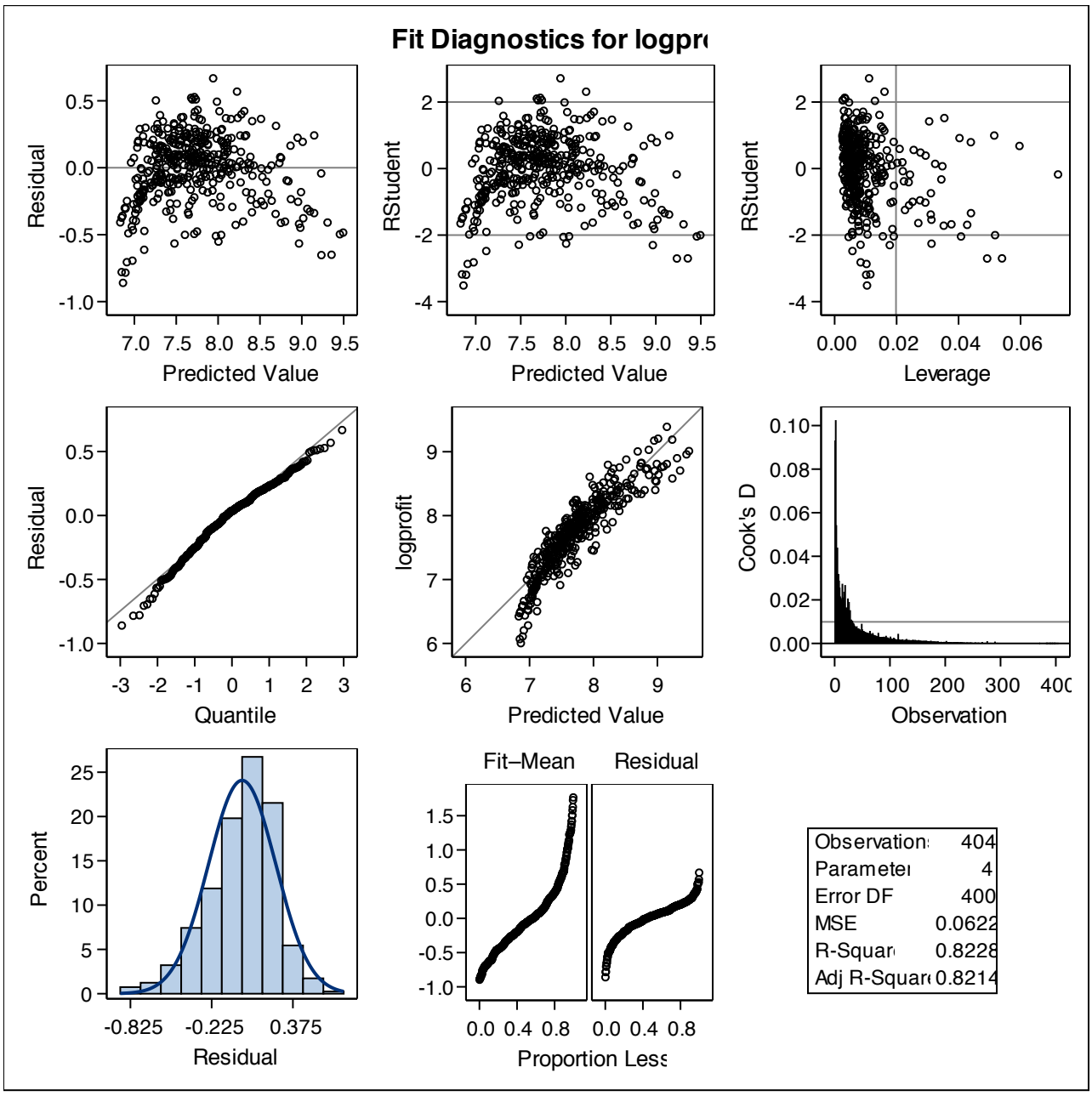
Residual plots show no strong indication of violation of equal variance assumption. q-q plot also shows no strong indication of violation of normality assumption. There is one observation, however, with exceptionally large Cook's DIstance. I will try to remove this point from our model.

Obs	Make	Model	cookd
1	Mercedes-Benz	CL500 2dr	0.11952
2	Mercedes-Benz	SL500 convertible 2dr	0.08238
3	Mercedes-Benz	G500	0.06821
4	Mercedes-Benz	S500 4dr	0.06800
5	Audi	RS 6 4dr	0.04500

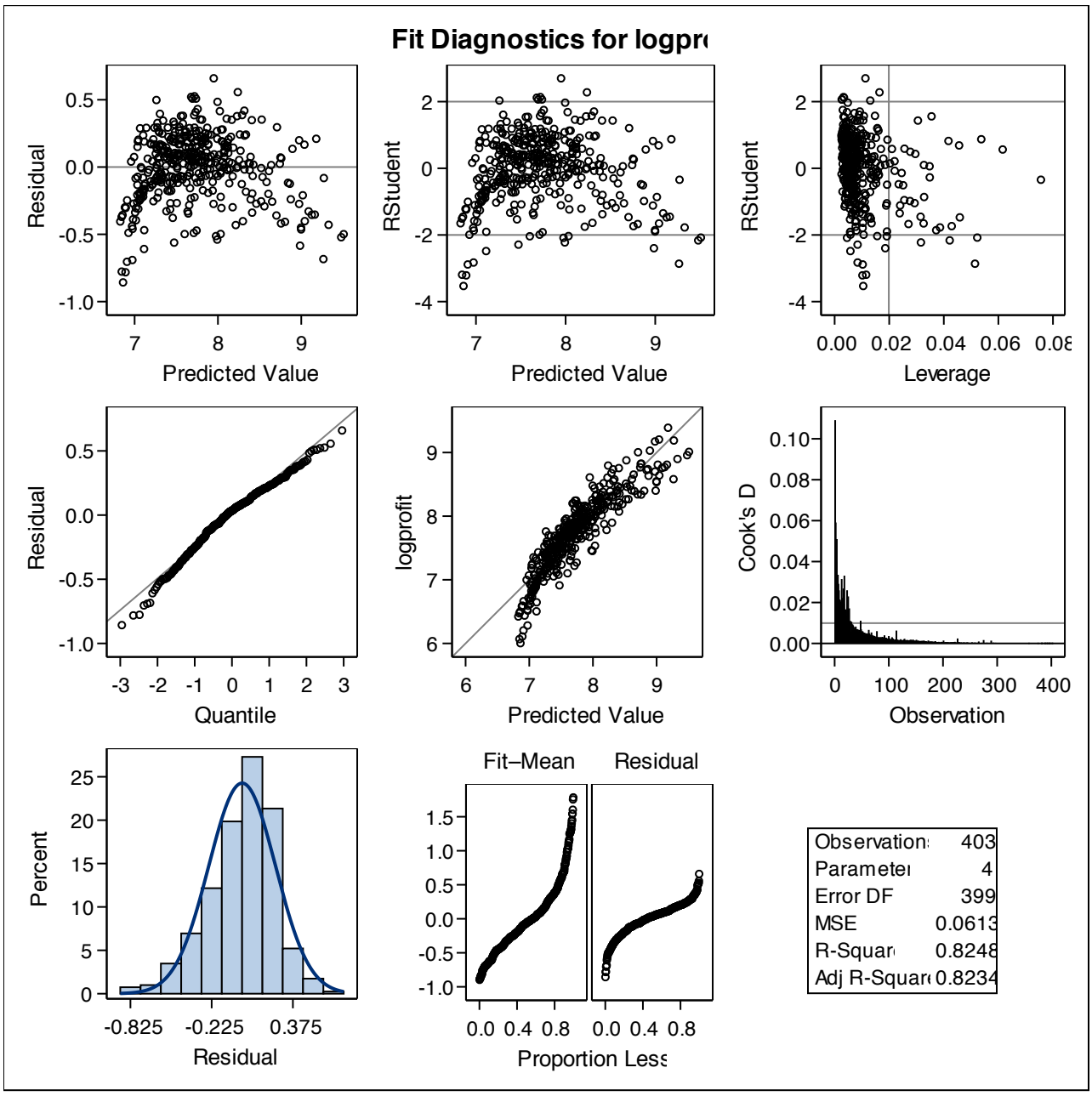
Model: MODEL1
Dependent Variable: logprofit



Model: MODEL1
Dependent Variable: logprofit



Model: MODEL1
Dependent Variable: logprofit



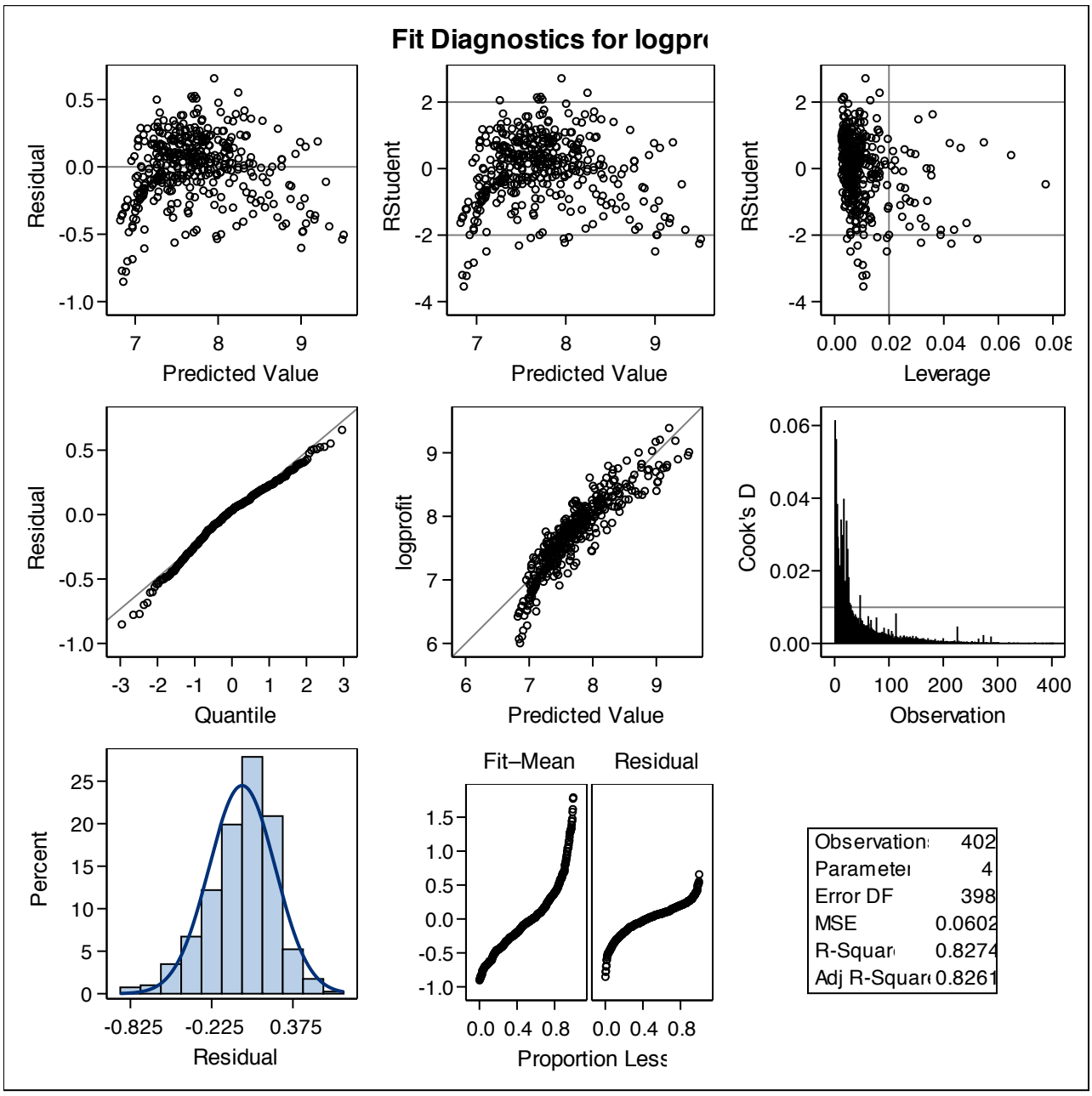
Model: MODEL1
Dependent Variable: logprofit

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	114.81249	38.27083	635.92	<.0001
Error	398	23.95246	0.06018		
Corrected Total	401	138.76495			

Root MSE	0.24532	**R-Square**	0.8274
Dependent Mean	7.72300	**Adj R-Sq**	0.8261
Coeff Var	3.17649		

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	6.08890	0.06170	98.69	<.0001
MSRP		1	0.00002613	0.00000144	18.14	<.0001
Weight	Weight (LBS)	1	0.00014819	0.00002231	6.64	<.0001
Horsepower		1	0.00137	0.00038118	3.60	0.0004

Model: MODEL1
Dependent Variable: logprofit



This is the final model. Interpretation in the body of the report is based on this model.
Note that we excluded Suzuki from linear regression. Now I am going to compare profit of cars made by Suzuki, and all other cars.

Expected Profit: Suzuki or All Others

Class Level Information		
Class	**Levels**	**Values**
Suzuki	2	0 1

Number of Observations Read	415
Number of Observations Used	415

Expected Profit: Suzuki or All Others

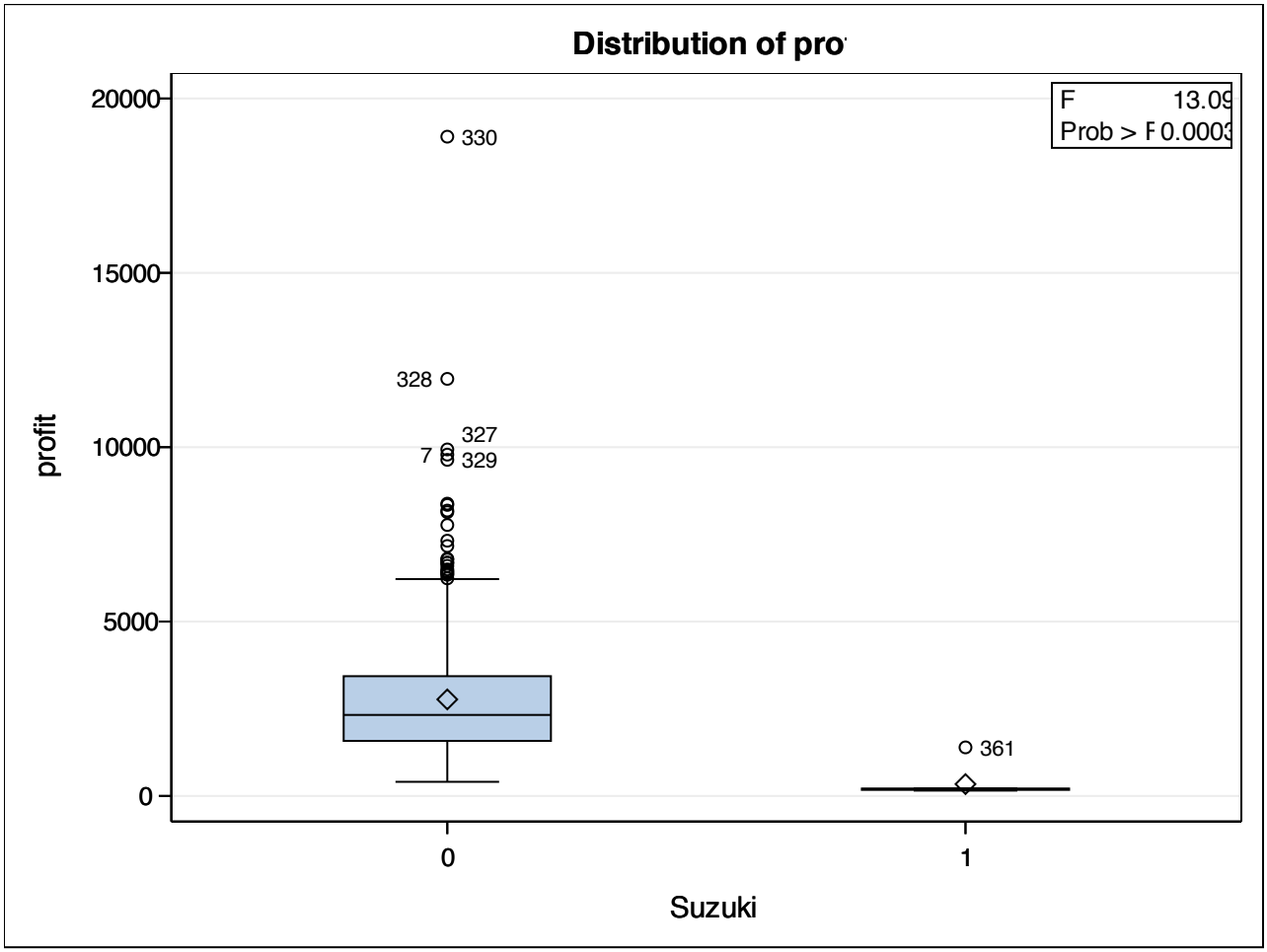
Dependent Variable: profit

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	46290817	46290817	13.09	0.0003
Error	413	1460445454	3536188		
Corrected Total	414	1506736271			

R-Square	Coeff Var	Root MSE	profit Mean
0.030723	69.10500	1880.475	2721.186

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Suzuki	1	46290816.75	46290816.75	13.09	0.0003

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Suzuki	1	46290816.75	46290816.75	13.09	0.0003



Expected Profit: Suzuki or All Others

Dependent Variable: profit

Result shows that Suzuki cars have significanly less expected profit than other cars.
Part 4: Determination, based on other variables, whether a model of vehicle is domestic or an import (Logistic Regression)
In this part, I will develop a logistic regression model, and predict whether a certain model of vehicle is domestic or an import.
To conduct such an analysis, I re-coded the variable Origin into Domestic if its value is 'USA', and Foreign if otherwise.

Response Profile		
Ordered Value	DorF	Total Frequency
1	Domestic	145
2	Foreign	268

Probability modeled is DorF='Domestic'.

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-12.0147	4.0142	8.9583	0.0028
Type	Hybrid	1	-12.0525	1247.1	0.0001	0.9923
Type	SUV	1	1.0934	0.7507	2.1212	0.1453
Type	Sedan	1	-0.4545	0.5798	0.6144	0.4331
Type	Sports	1	0.6672	0.8698	0.5884	0.4430
Type	Truck	1	-0.4625	0.9075	0.2597	0.6103
DriveTrain	All	1	-1.2955	0.5936	4.7621	0.0291
DriveTrain	Front	1	0.1985	0.4523	0.1926	0.6608
EngineSize		1	3.6997	0.5205	50.5299	<.0001
Cylinders		1	-0.7114	0.2759	6.6470	0.0099
Horsepower		1	-0.0394	0.00637	38.1850	<.0001
MPG_Highway		1	0.2806	0.1079	6.7651	0.0093
MPG_City		1	-0.3120	0.1275	5.9846	0.0144
Weight		1	-0.00135	0.000584	5.3142	0.0212
Wheelbase		1	0.1128	0.0493	5.2251	0.0223
Length		1	0.0173	0.0252	0.4740	0.4912

Based on the initial result, I will remove several variables, one at a time, and refit the model until all terms are significant.

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-6.7897	3.3157	4.1932	0.0406
DriveTrain	**All**	1	-1.1367	0.5676	4.0103	0.0452
DriveTrain	**Front**	1	0.3395	0.4061	0.6990	0.4031
EngineSize		1	3.7157	0.4959	56.1464	<.0001
Cylinders		1	-0.7353	0.2635	7.7866	0.0053
Horsepower		1	-0.0380	0.00600	40.1142	<.0001
MPG_Highway		1	0.2099	0.0907	5.3507	0.0207
MPG_City		1	-0.2651	0.1133	5.4787	0.0192
Weight		1	-0.00085	0.000527	2.6258	0.1051
Wheelbase		1	0.0779	0.0454	2.9499	0.0859
Length		1	0.00320	0.0236	0.0183	0.8923

Variable 'Type' is removed in this step.

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-6.7022	3.2532	4.2445	0.0394
DriveTrain	**All**	1	-1.1342	0.5675	3.9947	0.0456
DriveTrain	**Front**	1	0.3524	0.3943	0.7987	0.3715
EngineSize		1	3.7286	0.4872	58.5570	<.0001
Cylinders		1	-0.7348	0.2636	7.7738	0.0053
Horsepower		1	-0.0382	0.00592	41.4782	<.0001
MPG_Highway		1	0.2134	0.0871	6.0078	0.0142
MPG_City		1	-0.2697	0.1082	6.2177	0.0126
Weight		1	-0.00086	0.000526	2.6774	0.1018
Wheelbase		1	0.0826	0.0295	7.8650	0.0050

Variable 'Length' is removed in this step.

Model Information	
Data Set	WORK.CARS
Response Variable	DorF
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Response Profile		
Ordered Value	**DorF**	**Total Frequency**
1	Domestic	145
2	Foreign	268

Probability modeled is DorF='Domestic'.

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	537.346	372.264
SC	541.369	404.451
-2 Log L	535.346	356.264

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	179.0820	7	<.0001
Score	140.5301	7	<.0001
Wald	92.2143	7	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-8.3893	3.1484	7.1003	0.0077
EngineSize	1	3.7004	0.4774	60.0721	<.0001
Cylinders	1	-0.6333	0.2491	6.4640	0.0110
Horsepower	1	-0.0384	0.00576	44.4721	<.0001
MPG_Highway	1	0.2605	0.0847	9.4648	0.0021
MPG_City	1	-0.2956	0.1047	7.9669	0.0048
Weight	1	-0.00131	0.000479	7.5059	0.0061
Wheelbase	1	0.1022	0.0288	12.6049	0.0004

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
EngineSize	40.462	15.873	103.140
Cylinders	0.531	0.326	0.865
Horsepower	0.962	0.951	0.973
MPG_Highway	1.298	1.099	1.532
MPG_City	0.744	0.606	0.914
Weight	0.999	0.998	1.000
Wheelbase	1.108	1.047	1.172

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	**DF**	**Pr > ChiSq**
12.6204	8	0.1256

Variable 'DriveTrain' is removed in this step. This is the final model, based on which the result is reported to the client. Global test statistics all show our model is significantly better than an intercept-only model. In addition, Hosmer-Lemeshow test does not provide strong evidence for lack of fit. In addition, each term in the model is significant at least or almost at .01 level. Therefore, I conclude this model is a good fit.

This model shows that the odd of a car being domestic is associated with the following contributing factors: bigger engine size, better MPG on high way, and longer wheelbase.

This model also shows that the odd of a car being domestic is associated with the following undermining factors: less cylinders, less horsepower, lower MPG in the city, and lighter weight.

Now I will produce a classification table to show how good our model predicts the origin of a car.

Table of DorF by DorFpredicted			
DorF	**DorFpredicted**		
Frequency Row Pct	**Domestic**	**Foreign**	**Total**
Domestic	98 67.59	47 32.41	145
Foreign	26 9.63	244 90.37	270
Total	124	291	415

Statistics for Table of DorF by DorFpredicted

Statistic	DF	Value	Prob
Chi-Square	1	151.2414	<.0001
Likelihood Ratio Chi-Square	1	152.3732	<.0001
Continuity Adj. Chi-Square	1	148.4878	<.0001
Mantel-Haenszel Chi-Square	1	150.8769	<.0001
Phi Coefficient		0.6037	
Contingency Coefficient		0.5168	
Cramer's V		0.6037	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	98
Left-sided Pr <= F	1.0000
Right-sided Pr >= F	<.0001
Table Probability (P)	<.0001
Two-sided Pr <= P	<.0001

Sample Size = 415

The classification table shows that the model classifies Foreign vehicles correctly at a high success rate of 90%, and classifies Domestic vehicles correctly at a success rate of 67%. Chi-square test statistics also show that the predicted Domestic-Or-Foreign is strongly associated with Actually Domestic-or-Foreign

Part 5: Determination, based on other variables, on the number of cylinders a vehicle has

In the dataset provided by the client, all vehicles have 4, 6, or 8 cylinders. The number of cylinders a vehicle has is an important matrix that an auto dealer uses to structure inventory. In this part, I will design a model that could potentially help us to classify a vehicle based on information from other variables.

I will have SAS conduct discriminant analysis.

Stepwise Selection Summary									
Step	Number In	Entered	Removed	Label	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda
1	1	EngineSize		Engine Size (L)	0.8352	1038.68	<.0001	0.16483333	<.0001
2	2	MPG_City		MPG (City)	0.1225	28.55	<.0001	0.14463931	<.0001
3	3	Invoice			0.1136	26.13	<.0001	0.12821472	<.0001
4	4	MPG_Highway		MPG (Highway)	0.0168	3.48	0.0318	0.12606003	<.0001

Step	Number In	Entered	Removed	Average Squared Canonical Correlation	Pr > ASCC
1	1	EngineSize		0.41758334	<.0001
2	2	MPG_City		0.47868663	<.0001
3	3	Invoice		0.49373813	<.0001
4	4	MPG_Highway		0.50019651	<.0001

First, I had SAS select variables for further discriminant analysis. I will use only these variables for further exploration. I will use proportional-priors in analysis.

Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
593.398575	20	<.0001

Since the Chi-Square value is significant at the 0.1 level, the within covariance matrices will be used in the discriminant function.
Reference: Morrison, D.F. (1976) Multivariate Statistical Methods p252.

Multivariate Statistics and F Approximations					
S=2 M=0.5 N=202.5					
Statistic	**Value**	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Wilks' Lambda	0.12606003	184.83	8	814	<.0001
Pillai's Trace	1.00039303	102.08	8	816	<.0001
Hotelling-Lawley Trace	5.92961088	301.23	8	579.11	<.0001
Roy's Greatest Root	5.75531677	587.04	4	408	<.0001
NOTE: F Statistic for Roy's Greatest Root is an upper bound.					
NOTE: F Statistic for Wilks' Lambda is exact.					

Classification Results for Calibration Data: WORK.CARS
Cross-validation Results using Quadratic Discriminant Function

Posterior Probability of Membership in Cylinders						
Obs	From Cylinders	Classified into Cylinders		4	6	8
29	**6**	**4**	*	0.5045	0.4954	0.0001
32	**6**	**4**	*	0.5431	0.4568	0.0001
47	**6**	**8**	*	0.0000	0.4536	0.5464
67	**6**	**4**	*	0.7552	0.2448	0.0000
81	**6**	**8**	*	0.0000	0.3423	0.6577
86	**4**	**6**	*	0.0616	0.9384	0.0000
87	**6**	**8**	*	0.0000	0.4899	0.5101
134	**8**	**6**	*	0.0000	0.8509	0.1491
143	**6**	**8**	*	0.0000	0.3660	0.6340
144	**4**	**6**	*	0.1063	0.8935	0.0001
147	**6**	**8**	*	0.0000	0.3759	0.6241
187	**6**	**4**	*	0.8218	0.1782	0.0000
215	**6**	**4**	*	0.5883	0.4117	0.0000
231	**8**	**6**	*	0.0000	0.8250	0.1750
232	**8**	**6**	*	0.0000	0.7842	0.2158
266	**4**	**6**	*	0.0619	0.9380	0.0000
325	**6**	**8**	*	0.0000	0.3334	0.6666
326	**6**	**8**	*	0.0000	0.3380	0.6620
327	**6**	**8**	*	0.0000	0.4065	0.5935
328	**6**	**8**	*	0.0000	0.0000	1.0000
333	**4**	**6**	*	0.3439	0.6561	0.0000
334	**4**	**6**	*	0.0567	0.9432	0.0000
335	**4**	**6**	*	0.1703	0.8297	0.0000
336	**4**	**6**	*	0.0383	0.9616	0.0001
337	**4**	**6**	*	0.2348	0.7648	0.0005
355	**4**	**6**	*	0.4332	0.5668	0.0000
360	**6**	**4**	*	0.7178	0.2822	0.0000
365	**6**	**4**	*	0.6087	0.3912	0.0000
404	**8**	**6**	*	0.0000	0.7026	0.2974
408	**8**	**6**	*	0.0000	0.6862	0.3138

**** Misclassified observation***

Classification Summary for Calibration Data: WORK.CARS
Cross-validation Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into Cylinders				
From Cylinders	**4**	**6**	**8**	**Total**
4	127 93.38	9 6.62	0 0.00	136 100.00
6	7 3.68	174 91.58	9 4.74	190 100.00
8	0 0.00	5 5.75	82 94.25	87 100.00
Total	134 32.45	188 45.52	91 22.03	413 100.00
Priors	0.3293	0.46005	0.21065	

Error Count Estimates for Cylinders				
	4	**6**	**8**	**Total**
Rate	0.0662	0.0842	0.0575	0.0726
Priors	0.3293	0.4600	0.2107	

Multivariate statistics returned results significant at the .0001 level, indicating strong evidence to support differences among multivariate means.
Chi-square statistic returned significant result, indicating strong evidence of different covariances between cars of different Cylinder. We conclude there are differences in covariances. Therefore, within covariance matrices are used in analysis
Result shows our model achieves very high success rate when classifying vehicles based on the number of cylinders. Result is reported in the body of the project.