Part B. **Harsha Kankanamge**

Using the dataset 'auto2' that was provided in class, conduct the following analysis and answer the following questions using SAS:

1)Using the auto2 dataset from the Class 2 walkthrough, create a summary report that details the average cost per car per car manufacturer (note: manufacturer is different than model).

```
data sa_hm.auto3;
 set sa_hm.auto2(keep=make price mpg rep78 hdroom trunk weight length
turn
        displ gratio foreign);
 length manufacturer $20;
 manufacturer = substr(make, 1, index(make, ' ') - 1);
run;

proc means mean data=sa_hm.auto3;
       class manufacturer;
              var price;
run;
```

The MEANS Procedure

Analysis Variable : price		
manufacturer	N Obs	Mean
AMC	3	4215.67
Audi	2	7992.50
BMW	1	9735.00
Buick	7	6075.29
Cad.	3	13930.33
Chev.	6	4372.33
Datsun	4	6006.50
Dodge	4	5055.50
Fiat	1	4296.00
Ford	2	4288.00
Honda	2	5149.00
Linc.	3	12852.33
Mazda	1	3995.00
Merc.	6	4913.83
Olds	7	6050.86
Peugeot	1	12990.00
Plym.	5	4820.00
Pont.	6	4878.83
Renault	1	3895.00
Subaru	1	3798.00
Toyota	3	5122.00
VW	4	6021.00
Volvo	1	11995.00

2) Merge the table from Question (1) above onto the auto dataset, and create a statistic for the price of a car relative to the average price for all cars in the dataset made by the car's manufacturer. Limit your analysis to manufacturers that have at least three cars on the list. Which car has the highest relative price? What is it, expressed as a percentage of the average manufacturer's price? Which car has the lowest relative price? What is it, expressed as a percentage of the average manufacturer's price?

Adding new column call 'mean_price'

```
proc sql;
  create table sa_hm.auto5 as
  select *,round(mean(price),1)  as mean_price
  from sa_hm.auto3
  group by manufacturer;
quit;
run;
proc print data=sa_hm.auto5;
run;
```

Obs	make	price	mpg	rep78	hdroom	trunk	weight	length	turn	displ	gratio	foreign	manufacturer	mean_price
1	AMC Pacer	4749	17	3	3.0	11	3350	173	40	258	2.53	0	AMC	4216
2	AMC Concord	4099	22	3	2.5	11	2930	186	40	121	3.58	0	AMC	4216
3	AMC Spirit	3799	22	.	3.0	12	2640	168	35	121	3.08	0	AMC	4216
4	Audi 5000	9690	17	5	3.0	15	2830	189	37	131	3.20	1	Audi	7993
5	Audi Fox	6295	23	3	2.5	11	2070	174	36	97	3.70	1	Audi	7993
6	BMW 320i	9735	25	4	2.5	12	2650	177	34	121	3.64	1	BMW	9735
7	Buick LeSabre	5788	18	3	4.0	21	3670	218	43	231	2.73	0	Buick	6075
8	Buick Regal	5189	20	3	2.0	16	3280	200	42	196	2.93	0	Buick	6075
9	Buick Skylark	4082	19	3	3.5	13	3400	200	42	231	3.08	0	Buick	6075
10	Buick Century	4816	20	3	4.5	16	3250	196	40	196	2.93	0	Buick	6075
11	Buick Riviera	10372	16	3	3.5	17	3880	207	43	231	2.93	0	Buick	6075
12	Buick Opel	4453	26	.	3.0	10	2230	170	34	304	2.87	0	Buick	6075
13	Buick Electra	7827	15	4	4.0	20	4080	222	43	350	2.41	0	Buick	6075
14	Cad. Deville	11385	14	3	4.0	20	4330	221	44	425	2.28	0	Cad.	13930
15	Cad. Eldorado	14500	14	2	3.5	16	3900	204	43	350	2.19	0	Cad.	13930
16	Cad. Seville	15906	21	3	3.0	13	4290	204	45	350	2.24	0	Cad.	13930
17	Chev. Nova	3955	19	3	3.5	13	3430	197	43	250	2.56	0	Chev.	4372
18	Chev. Monza	3667	24	2	2.0	7	2750	179	40	151	2.73	0	Chev.	4372
19	Chev. Monte Carlo	5104	22	2	2.0	16	3220	200	41	200	2.73	0	Chev.	4372
20	Chev. Malibu	4504	22	3	3.5	17	3180	193	31	200	2.73	0	Chev.	4372
21	Chev. Impala	5705	16	4	4.0	20	3690	212	43	250	2.56	0	Chev.	4372
22	Chev. Chevette	3299	29	3	2.5	9	2110	163	34	231	2.93	0	Chev.	4372
23	Datsun 510	5079	24	4	2.5	8	2280	170	34	119	3.54	1	Datsun	6007
24	Datsun 210	4589	35	5	2.0	8	2020	165	32	85	3.70	1	Datsun	6007
25	Datsun 200	6229	23	4	1.5	6	2370	170	35	119	3.89	1	Datsun	6007
26	Datsun 810	8129	21	4	2.5	8	2750	184	38	146	3.55	1	Datsun	6007
27	Dodge St. Regis	6342	17	2	4.5	21	3740	220	46	225	2.94	0	Dodge	5056
28	Dodge Diplomat	4010	18	2	4.0	17	3600	206	46	318	2.47	0	Dodge	5056
29	Dodge Colt	3984	30	5	2.0	8	2120	163	35	98	3.54	0	Dodge	5056
30	Dodge Magnum	5886	16	2	4.0	17	3600	206	46	318	2.47	0	Dodge	5056
31	Fiat Strada	4296	21	3	2.5	16	2130	161	36	105	3.37	1	Fiat	4296

Adding frequency column to dataset to find out: frequency>2

```
proc sql;
  create table sa_hm.auto6 as
  select *,count(manufacturer)  as frequency
  from sa_hm.auto5
  group by manufacturer;
quit;
run;

data sa_hm.auto7;
       set sa_hm.auto6;
       if frequency>2;
run;

proc print data=sa_hm.auto7;
run;
```

Obs	make	price	mpg	rep78	hdroom	trunk	weight	length	turn	displ	gratio	foreign	manufacturer	mean_price	frequency
1	AMC Pacer	4749	17	3	3.0	11	3350	173	40	258	2.53	0	AMC	4216	3
2	AMC Concord	4099	22	3	2.5	11	2930	186	40	121	3.58	0	AMC	4216	3
3	AMC Spirit	3799	22	.	3.0	12	2640	168	35	121	3.08	0	AMC	4216	3
4	Buick LeSabre	5788	18	3	4.0	21	3670	218	43	231	2.73	0	Buick	6075	7
5	Buick Regal	5189	20	3	2.0	16	3280	200	42	196	2.93	0	Buick	6075	7
6	Buick Skylark	4082	19	3	3.5	13	3400	200	42	231	3.08	0	Buick	6075	7
7	Buick Century	4816	20	3	4.5	16	3250	196	40	196	2.93	0	Buick	6075	7
8	Buick Riviera	10372	16	3	3.5	17	3880	207	43	231	2.93	0	Buick	6075	7
9	Buick Opel	4453	26	.	3.0	10	2230	170	34	304	2.87	0	Buick	6075	7
10	Buick Electra	7827	15	4	4.0	20	4080	222	43	350	2.41	0	Buick	6075	7
11	Cad. Deville	11385	14	3	4.0	20	4330	221	44	425	2.28	0	Cad.	13930	3
12	Cad. Eldorado	14500	14	2	3.5	16	3900	204	43	350	2.19	0	Cad.	13930	3
13	Cad. Seville	15906	21	3	3.0	13	4290	204	45	350	2.24	0	Cad.	13930	3
14	Chev. Nova	3955	19	3	3.5	13	3430	197	43	250	2.56	0	Chev.	4372	6
15	Chev. Monza	3667	24	2	2.0	7	2750	179	40	151	2.73	0	Chev.	4372	6
16	Chev. Monte Carlo	5104	22	2	2.0	16	3220	200	41	200	2.73	0	Chev.	4372	6
17	Chev. Malibu	4504	22	3	3.5	17	3180	193	31	200	2.73	0	Chev.	4372	6
18	Chev. Impala	5705	16	4	4.0	20	3690	212	43	250	2.56	0	Chev.	4372	6
19	Chev. Chevette	3299	29	3	2.5	9	2110	163	34	231	2.93	0	Chev.	4372	6
20	Datsun 510	5079	24	4	2.5	8	2280	170	34	119	3.54	1	Datsun	6007	4
21	Datsun 210	4589	35	5	2.0	8	2020	165	32	85	3.70	1	Datsun	6007	4
22	Datsun 200	6229	23	4	1.5	6	2370	170	35	119	3.89	1	Datsun	6007	4
23	Datsun 810	8129	21	4	2.5	8	2750	184	38	146	3.55	1	Datsun	6007	4
24	Dodge St. Regis	6342	17	2	4.5	21	3740	220	46	225	2.94	0	Dodge	5056	4
25	Dodge Diplomat	4010	18	2	4.0	17	3600	206	46	318	2.47	0	Dodge	5056	4
26	Dodge Colt	3984	30	5	2.0	8	2120	163	35	98	3.54	0	Dodge	5056	4
27	Dodge Magnum	5886	16	2	4.0	17	3600	206	46	318	2.47	0	Dodge	5056	4
28	Linc. Versailles	13466	14	3	3.5	15	3830	201	41	302	2.47	0	Linc.	12852	3
29	Linc. Mark V	13594	12	3	2.5	18	4720	230	48	400	2.47	0	Linc.	12852	3
30	Linc. Continental	11497	12	3	3.5	22	4840	233	51	400	2.47	0	Linc.	12852	3

Filtering the data and adding difference column (frequency >2)

```
proc sql;
  create table sa_hm.auto8 as
  select manufacturer, make, price, mean_price,(price-mean_price)  as
Difference
  from sa_hm.auto7;
quit;
run;
proc print data=sa_hm.auto8;
run;
```

Obs	manufacturer	make	price	mean_price	Difference
1	AMC	AMC Pacer	4749	4216	533
2	AMC	AMC Concord	4099	4216	-117
3	AMC	AMC Spirit	3799	4216	-417
4	Buick	Buick LeSabre	5788	6075	-287
5	Buick	Buick Regal	5189	6075	-886
6	Buick	Buick Skylark	4082	6075	-1993
7	Buick	Buick Century	4816	6075	-1259
8	Buick	Buick Riviera	10372	6075	4297
9	Buick	Buick Opel	4453	6075	-1622
10	Buick	Buick Electra	7827	6075	1752
11	Cad.	Cad. Deville	11385	13930	-2545
12	Cad.	Cad. Eldorado	14500	13930	570
13	Cad.	Cad. Seville	15906	13930	1976
14	Chev.	Chev. Nova	3955	4372	-417
15	Chev.	Chev. Monza	3667	4372	-705
16	Chev.	Chev. Monte Carlo	5104	4372	732
17	Chev.	Chev. Malibu	4504	4372	132
18	Chev.	Chev. Impala	5705	4372	1333
19	Chev.	Chev. Chevette	3299	4372	-1073
20	Datsun	Datsun 510	5079	6007	-928

Statistic for Difference:

```
proc means data=sa_hm.auto8;
 var Difference;
 run;
```

The MEANS Procedure

Analysis Variable : Difference				
N	Mean	Std Dev	Minimum	Maximum
61	-0.0327869	1400.89	-2545.00	4320.00

Adding average manufacturing percentage column

```
proc sql;
  create table sa_hm.auto9 as
  select *,round(mean_price*100/price)  as A_M_percentage
  from sa_hm.auto8;
quit;
run;
proc print data=sa_hm.auto9;
run;
```

Obs	manufacturer	make	price	mean_price	Difference	A_M_percentage
1	AMC	AMC Pacer	4749	4216	533	89
2	AMC	AMC Concord	4099	4216	-117	103
3	AMC	AMC Spirit	3799	4216	-417	111
4	Buick	Buick LeSabre	5788	6075	-287	105
5	Buick	Buick Regal	5189	6075	-886	117
6	Buick	Buick Skylark	4082	6075	-1993	149
7	Buick	Buick Century	4816	6075	-1259	126
8	Buick	Buick Riviera	10372	6075	4297	59
9	Buick	Buick Opel	4453	6075	-1622	136
10	Buick	Buick Electra	7827	6075	1752	78
11	Cad.	Cad. Deville	11385	13930	-2545	122
12	Cad.	Cad. Eldorado	14500	13930	570	96
13	Cad.	Cad. Seville	15906	13930	1976	88
14	Chev.	Chev. Nova	3955	4372	-417	111
15	Chev.	Chev. Monza	3667	4372	-705	119
16	Chev.	Chev. Monte Carlo	5104	4372	732	86
17	Chev.	Chev. Malibu	4504	4372	132	97
18	Chev.	Chev. Impala	5705	4372	1333	77
19	Chev.	Chev. Chevette	3299	4372	-1073	133
20	Datsun	Datsun 510	5079	6007	-928	118
21	Datsun	Datsun 210	4589	6007	-1418	131
22	Datsun	Datsun 200	6229	6007	222	96
23	Datsun	Datsun 810	8129	6007	2122	74
24	Dodge	Dodge St. Regis	6342	5056	1286	80
25	Dodge	Dodge Diplomat	4010	5056	-1046	126
26	Dodge	Dodge Colt	3984	5056	-1072	127
27	Dodge	Dodge Magnum	5886	5056	830	86
28	Linc.	Linc. Versailles	13466	12852	614	95
29	Linc.	Linc. Mark V	13594	12852	742	95
30	Linc.	Linc. Continental	11497	12852	-1355	112

Top five: Highest relative price cars

```
PROC SORT DATA=sa_hm.auto9 OUT=sa_hm.auto10 ;
  BY descending Difference ;
RUN ;

proc print data=sa_hm.auto10;
run;
```

Obs	manufacturer	make	price	mean_price	Difference	A_M_percentage
1	Olds	Olds Toronado	10371	6051	4320	58
2	Buick	Buick Riviera	10372	6075	4297	59
3	Olds	Olds 98	8814	6051	2763	69
4	Datsun	Datsun 810	8129	6007	2122	74
5	Cad.	Cad. Seville	15906	13930	1976	88

Highest (Difference) relative price car is 'Olds Toronado'

Bottom five: Lowest relative price cars

56	Buick	Buick Opel	4453	6075	-1622	136
57	Merc.	Merc. Zephyr	3291	4914	-1623	149
58	Olds	Olds Starfire	4195	6051	-1856	144
59	Olds	Olds Omega	4181	6051	-1870	145
60	Buick	Buick Skylark	4082	6075	-1993	149
61	Cad.	Cad. Deville	11385	13930	-2545	122

Lowest (Difference) relative price car is Cad. Deville

What is it, expressed as a percentage of the average manufacturer's price:

When manufacture has more than one differences products, those difference products may have difference manufacturing price. So we can calculate average manufacturing price for all products.

Average manufacturing price= $\frac{\textbf{Sum of manufacturer price for all products}}{\textit{Number of difference products}}$

Percentage of the average manufacturer's price= $\frac{\textit{Average manufacturer's prices} * \textbf{100}}{\textbf{Manufacturer price}}$

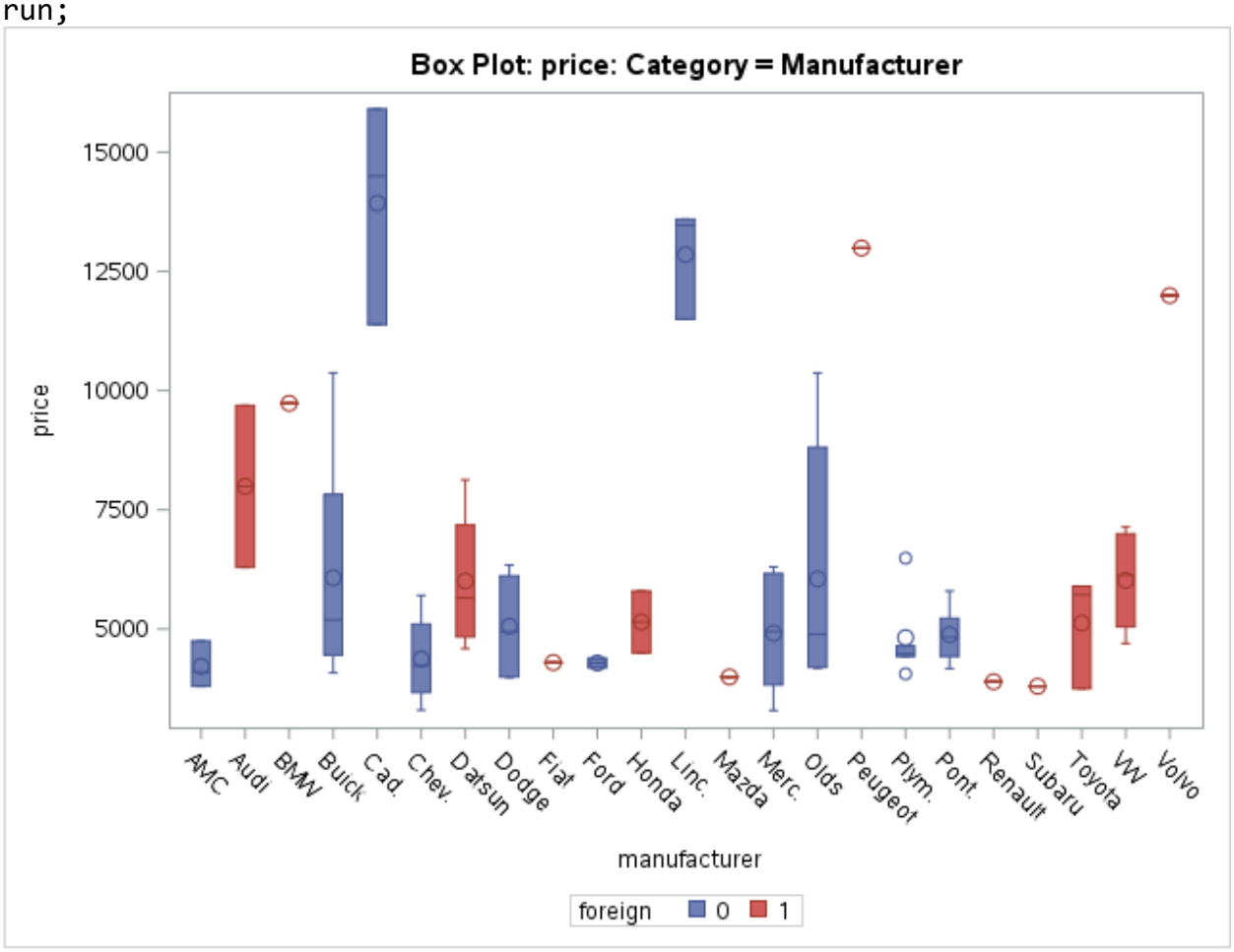
Which means if Buick Riviera manufacturing price is $100 Buick average manufacturing price will be $59.

Simply it means, 59% of its manufacturing price is its average manufacturing price.

3) Use the PROC BOXPLOT function to graph a boxplot for the price ranges and the miles per gallon reported for each car manufacturer. Which manufacturer has the largest amount of variability in reported price? Which manufacturer has the lowest variability in terms of miles per gallon? How do you define variability?

price ranges

```
proc sgplot data=sa_hm.auto3;
title "Box Plot: price: Category = Manufacturer";
vbox price / category=manufacturer Group=foreign;
run;
```



Buick manufacturer has the largest amount of variability in reported prices.

It is 6290.

Calculating price Difference

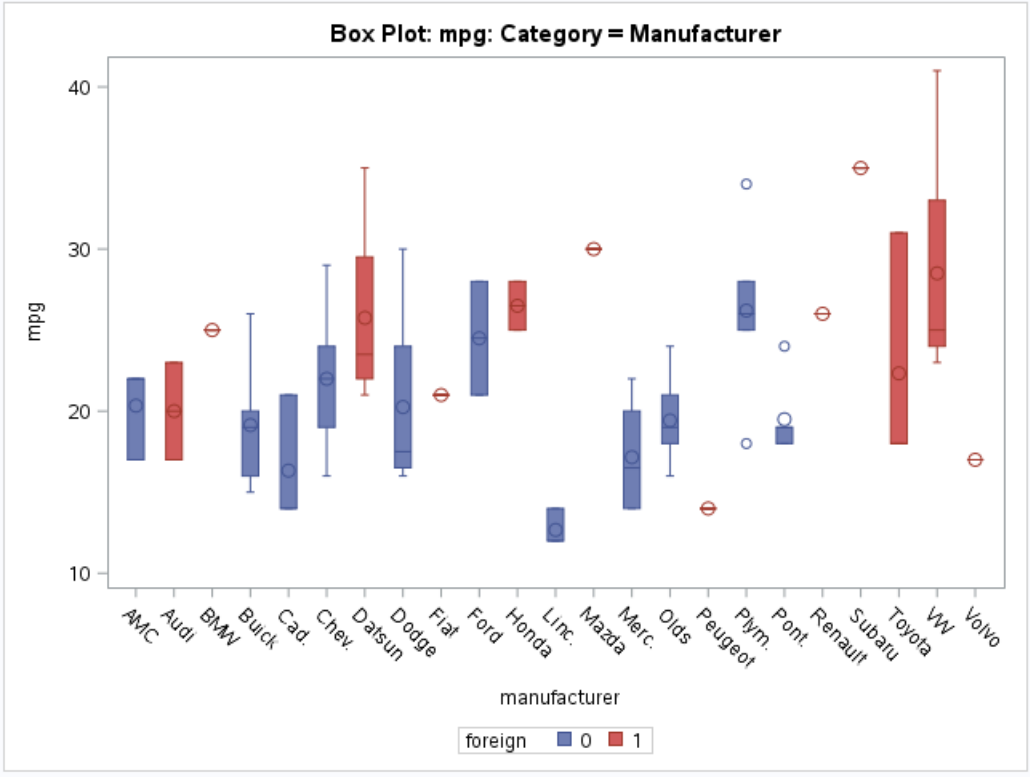
```
proc sql ;
  create table sa_hm.auto13 as
  select distinct manufacturer,range(price) as difference
  from sa_hm.auto3
  group by manufacturer;
quit;
run;

proc print data=sa_hm.auto13;
run;
```

Obs	manufacturer	difference
1	AMC	950
2	Audi	3395
3	BMW	0
4	Buick	6290
5	Cad.	4521
6	Chev.	2406
7	Datsun	3540
8	Dodge	2358
9	Fiat	0
10	Ford	202
11	Honda	1300
12	Linc.	2097
13	Mazda	0
14	Merc.	3012
15	Olds	6190
16	Peugeot	0
17	Plym.	2426
18	Pont.	1626
19	Renault	0
20	Subaru	0
21	Toyota	2151
22	VW	2443
23	Volvo	0

Miles per gallon:

```
proc sgplot data=sa_hm.auto3;
title "Box Plot: mpg: Category = Manufacturer";
vbox mpg / category=manufacturer Group=foreign;
run;
```



Renault, Subaru, Volvo, Peugeot, Mazda, Fiat, BMW manufacturer have the lowest variability in terms of miles per gallon (0)

```
proc sql ;
  create table sa_hm.auto14 as
  select distinct manufacturer,range(mpg) as difference_mpg
  from sa_hm.auto3
  group by manufacturer;
quit;

run;
proc print data=sa_hm.auto14;
run;
```

Variability is a range which means data points in a statistical distribution or data set diverge from the average, or mean, value as well as the extent to which these data points differ from each other. Commonly used measures of variability: range, variance and standard deviation.

Obs	manufacturer	difference_mpg
1	AMC	5
2	Audi	6
3	BMW	0
4	Buick	11
5	Cad.	7
6	Chev.	13
7	Datsun	14
8	Dodge	14
9	Fiat	0
10	Ford	7
11	Honda	3
12	Linc.	2
13	Mazda	0
14	Merc.	8
15	Olds	8
16	Peugeot	0
17	Plym.	16
18	Pont.	6
19	Renault	0
20	Subaru	0
21	Toyota	13
22	VW	18
23	Volvo	0

4) Run a regression testing whether or not there is a curvilinear relationship between a car's MPG and its weight and length. Use a backward selection process to retain all variables that are statistically significant with a P value of 0.1 or less. What is the improvement in model fit relative to a simple linear model? Use the SGPLOT procedure to create a chart that plots both actual and predicted MPG relative to the significant variables.

```
proc reg data=sa_hm.auto2;
model mpg= length weight /selection = backward;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: mpg

Number of Observations Read	74
Number of Observations Used	74

Backward Elimination: Step 0

All Variables Entered: R-Square = 0.6614 and C(p) = 3.0000

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1616.08062	808.04031	69.34	<.0001
Error	71	827.37884	11.65322		
Corrected Total	73	2443.45946			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	47.88487	6.08787	720.96132	61.87	<.0001
length	-0.07959	0.05536	24.09042	2.07	0.1549
weight	-0.00385	0.00159	68.72348	5.90	0.0177

Bounds on condition number: 9.5177, 38.071

Backward Elimination: Step 1

Variable length Removed: R-Square = 0.6515 and C(p) = 3.0673

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1591.99020	1591.99020	134.62	<.0001
Error	72	851.46926	11.82596		
Corrected Total	73	2443.45946			

Mpg=-0.07959*length – 0.00385*weight + 47.88487 in this model length is not significant

p-value=0.1549>0.1

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	39.44028	1.61400	7061.67623	597.13	<.0001
weight	-0.00601	0.00051788	1591.99020	134.62	<.0001

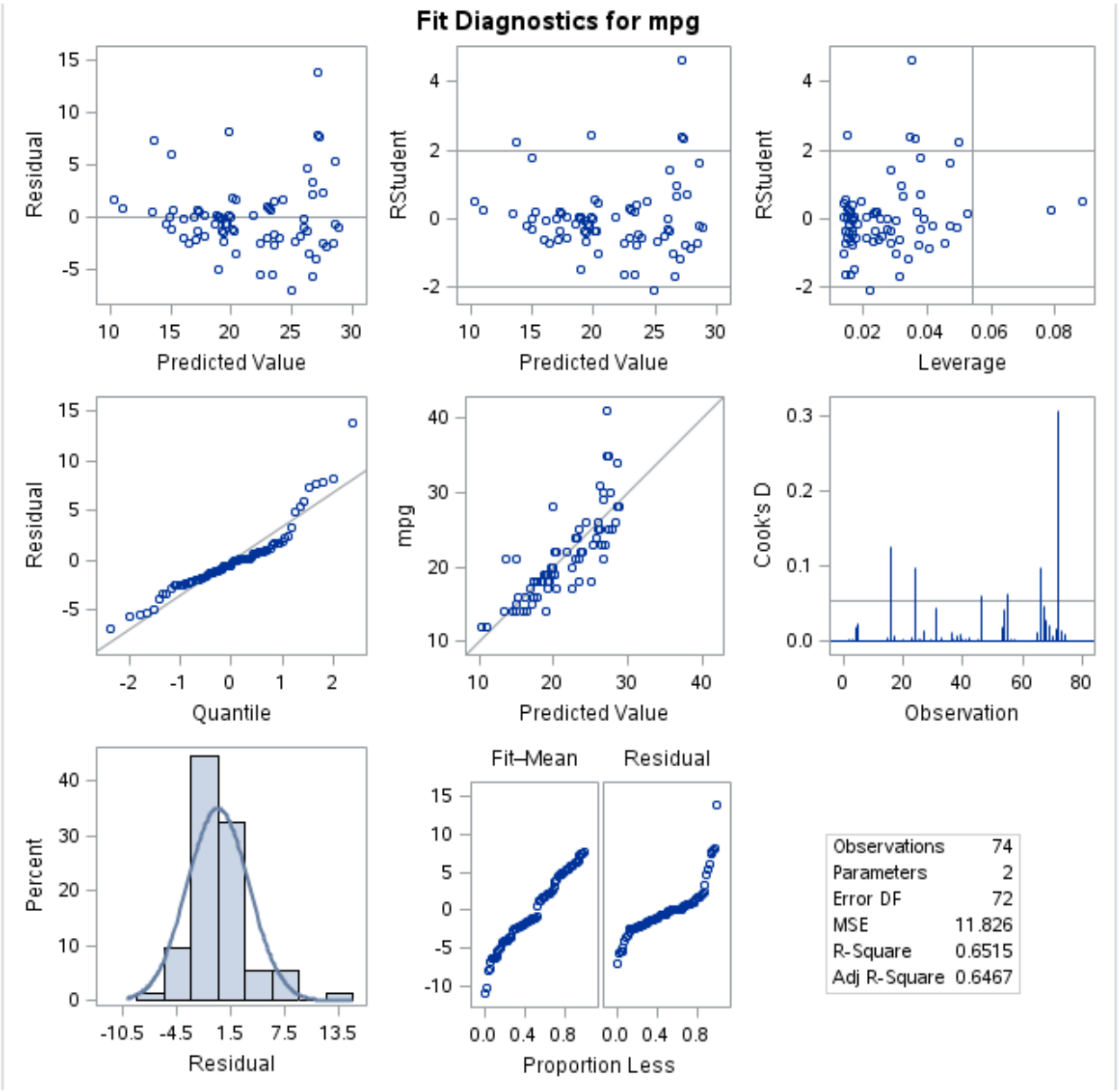
Bounds on condition number: 1, 1

All variables left in the model are significant at the 0.1000 level.

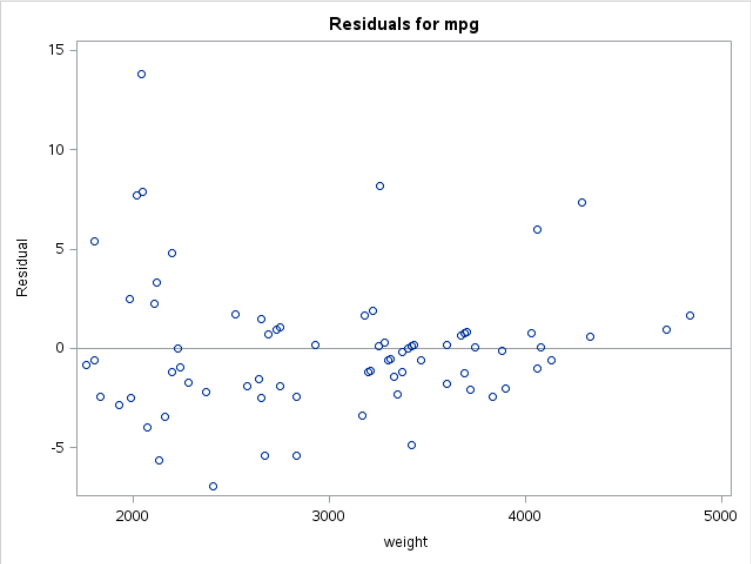
Summary of Backward Elimination							
Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	length	1	0.0099	0.6515	3.0673	2.07	0.1549

Mpg = - 0.00601*weight+39.44028

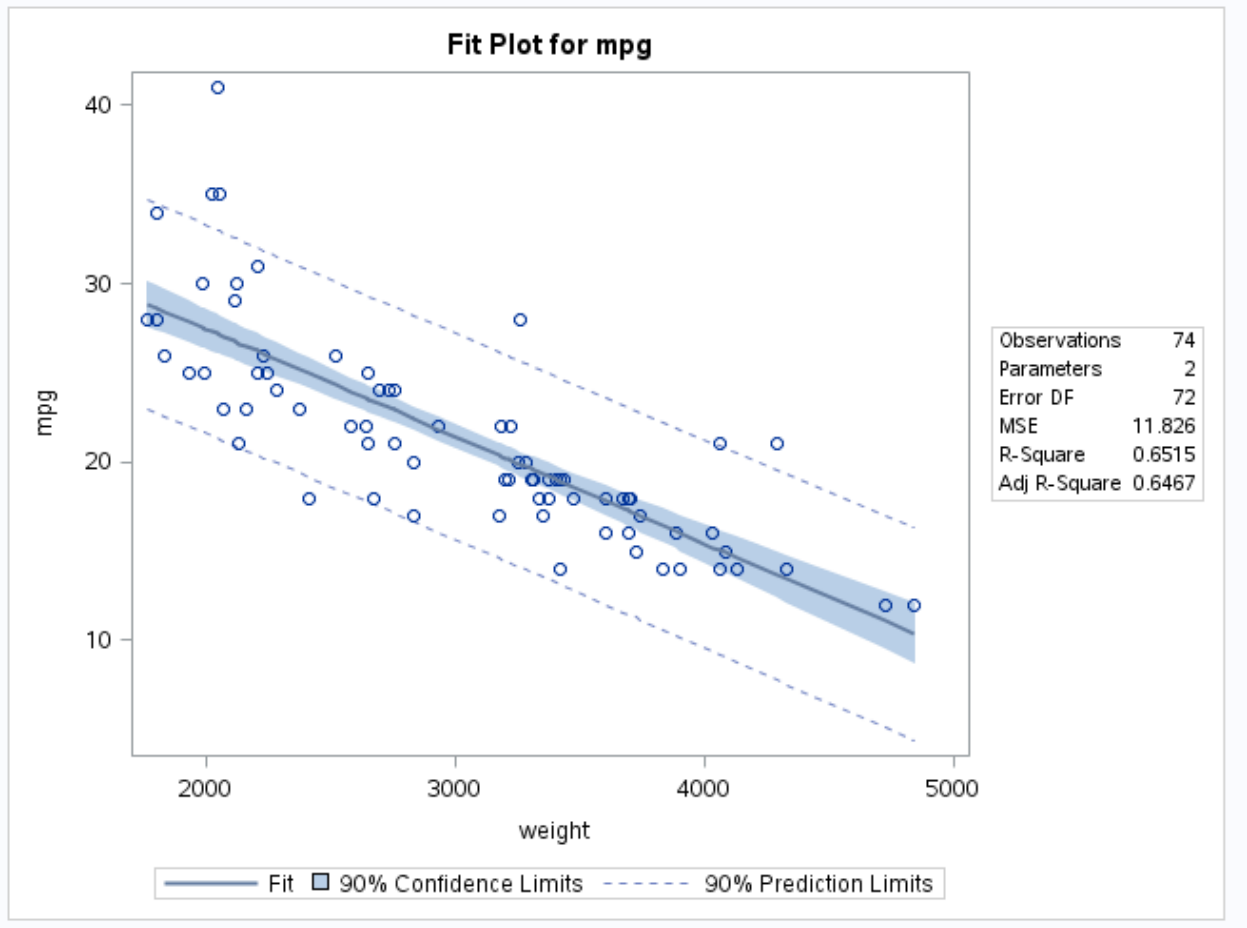
A coefficient is significant. R^2 =0.6515 R^2 is very close even model with length so this model is better. Less variable and very high F=134.62 p-value <0.0001



All the residual points are fitting onto the line. Which mean residuals are normally distributed.



We do not see the any pattern here. So residuals are homogeneous.

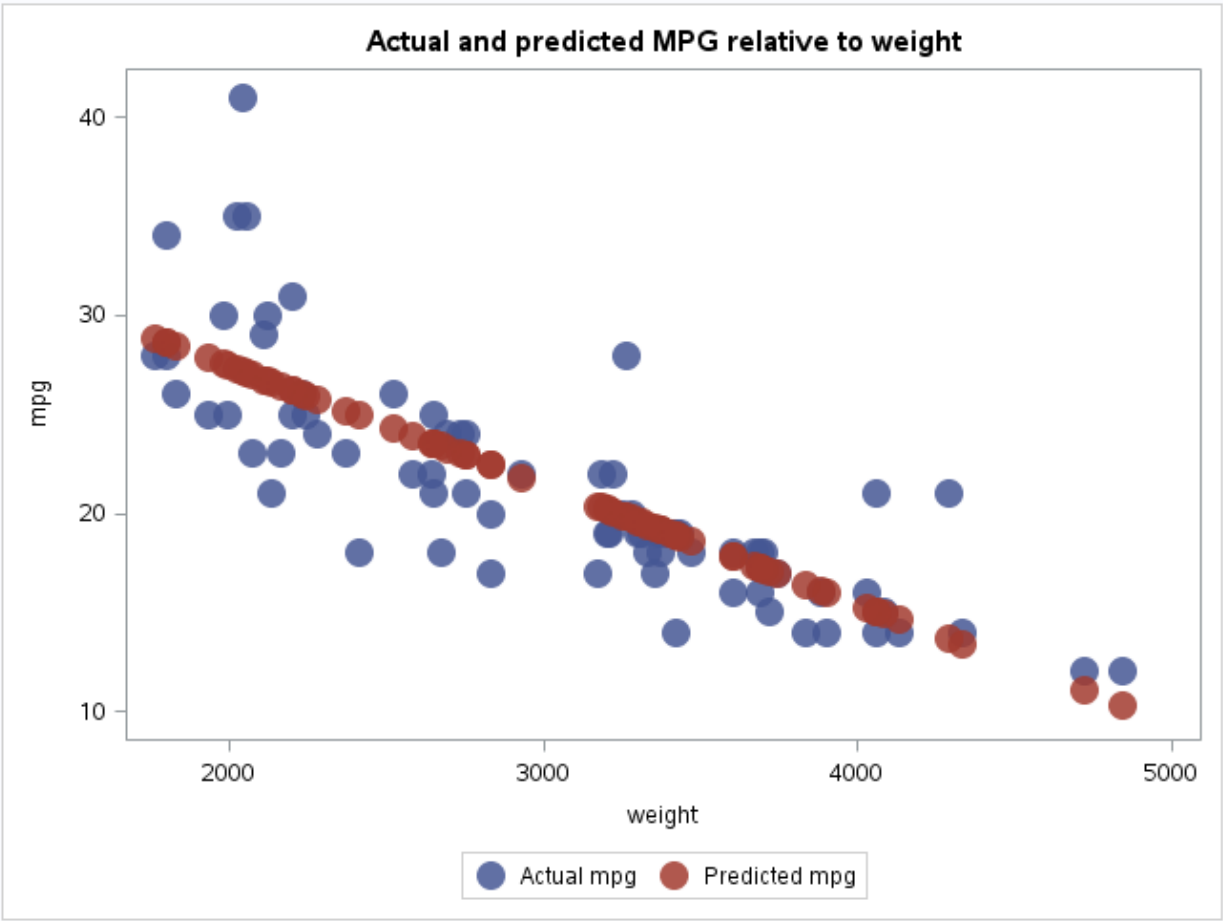


Plotting actual values vs predicted values

```
proc sql;
  create table sa_hm.auto15 as
  select *,-0.00601*weight+39.44028  as predict_mpg
  from sa_hm.auto2;
quit;
run;

proc sgplot data=sa_hm.auto15;
   title "MPG with weight";
  Scatter X = weight Y = mpg /
 LEGENDLABEL = 'Actual mpg' markerattrs=(symbol=circlefilled size =15)
transparency=0.15;
 Scatter X = weight Y = predict_mpg/
 LEGENDLABEL = 'Predicted mpg' markerattrs=(symbol=circlefilled size
=15) transparency=0.15;

run;
```



5) Your friend Joaquin likes to go gambling. She's planning on going to Niagara Falls to play the Roulette Wheel (American Style). She plans on bringing $1,000 to make 100 $10 bets on landing on 00 on the wheel. She will not re-invest any winnings or play more than 100 times. Given your new knowledge of SAS programming you offer to write a program in SAS for her to estimate her expected losses. Write a program to estimate how much Lorena will lose by playing roulette 100 times.

```
data loss;

expected_losses=100*(350*(1/38)-10*(37/38));

run;

proc print data=loss;
run;
```

Obs	expected_losses
1	-52.6316

Estimate lose by playing roulette 100 times: 52.6316