Harsha Kankanamge

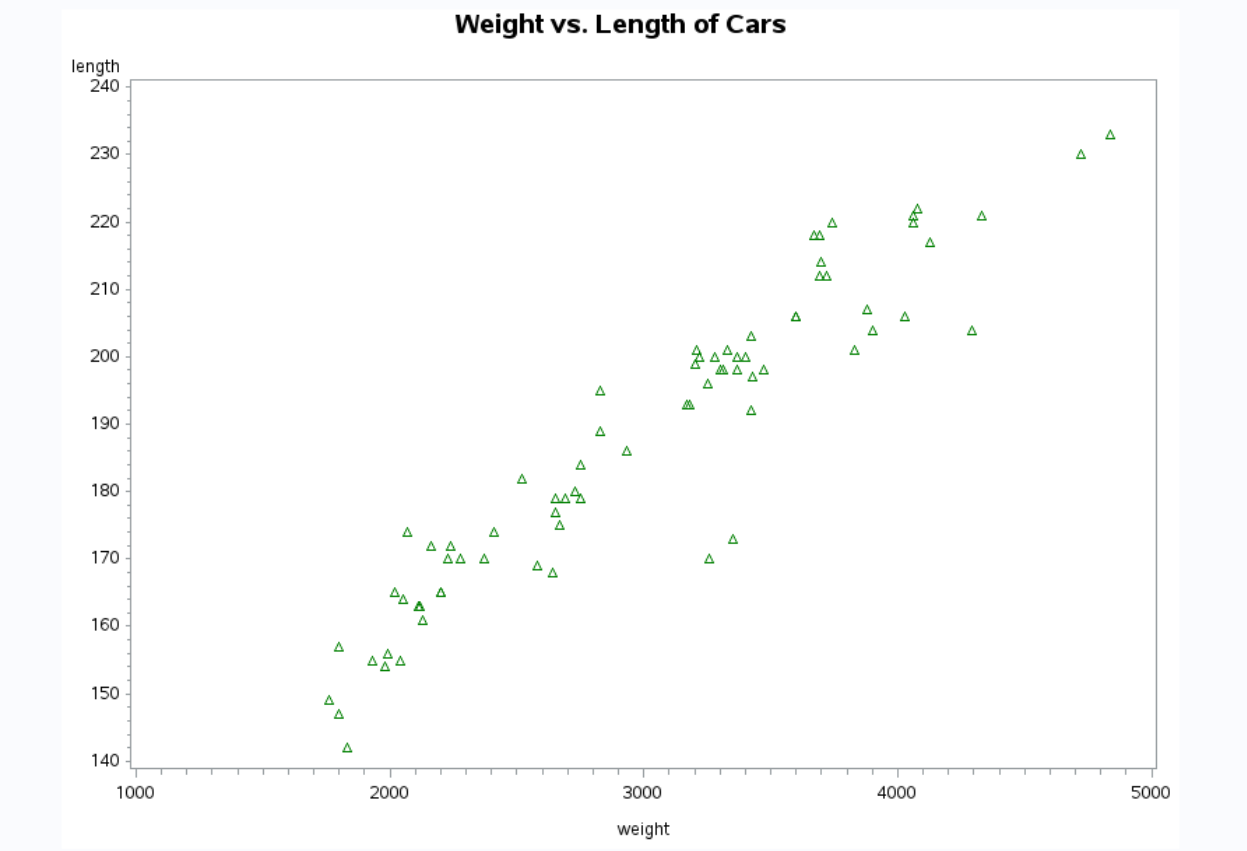
ACM 614 – Fall 2018 – Problem Set I

Part B.

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

1)Create a scatter plot that compares the weight of the cars to their length. Make the plot with green triangles. Make certain that weight is on the x axis and length is on the y axis. Title the plot 'Weight vs. Length of Cars' and make certain that no regression line is present.

```
proc gplot data=sa_hm.auto2;
    symbol v=triangle c=green;
    plot length*weight;
    Title 'Weight vs. Length of Cars';
run;
```



We can see very strong positive linear relationship between weight and length of the cars. When weight of the cars increases length of the cars also increases.

2)Run a simple linear regression of length (dependent variable) on weight (independent variable). Based on this analysis, how long would you expect a 6,000 pound car to be?

```
proc reg data=sa_hm.auto2;
   model length= weight;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: length

Number of Observations Read	74
Number of Observations Used	74

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	32390	32390	613.27	<.0001
Error	72	3802.67784	52.81497		
Corrected Total	73	36193			

Root MSE	7.26739	R-Square	0.8949
Dependent Mean	187.93243	Adj R-Sq	0.8935
Coeff Var	3.86702		

Parameter Estimates

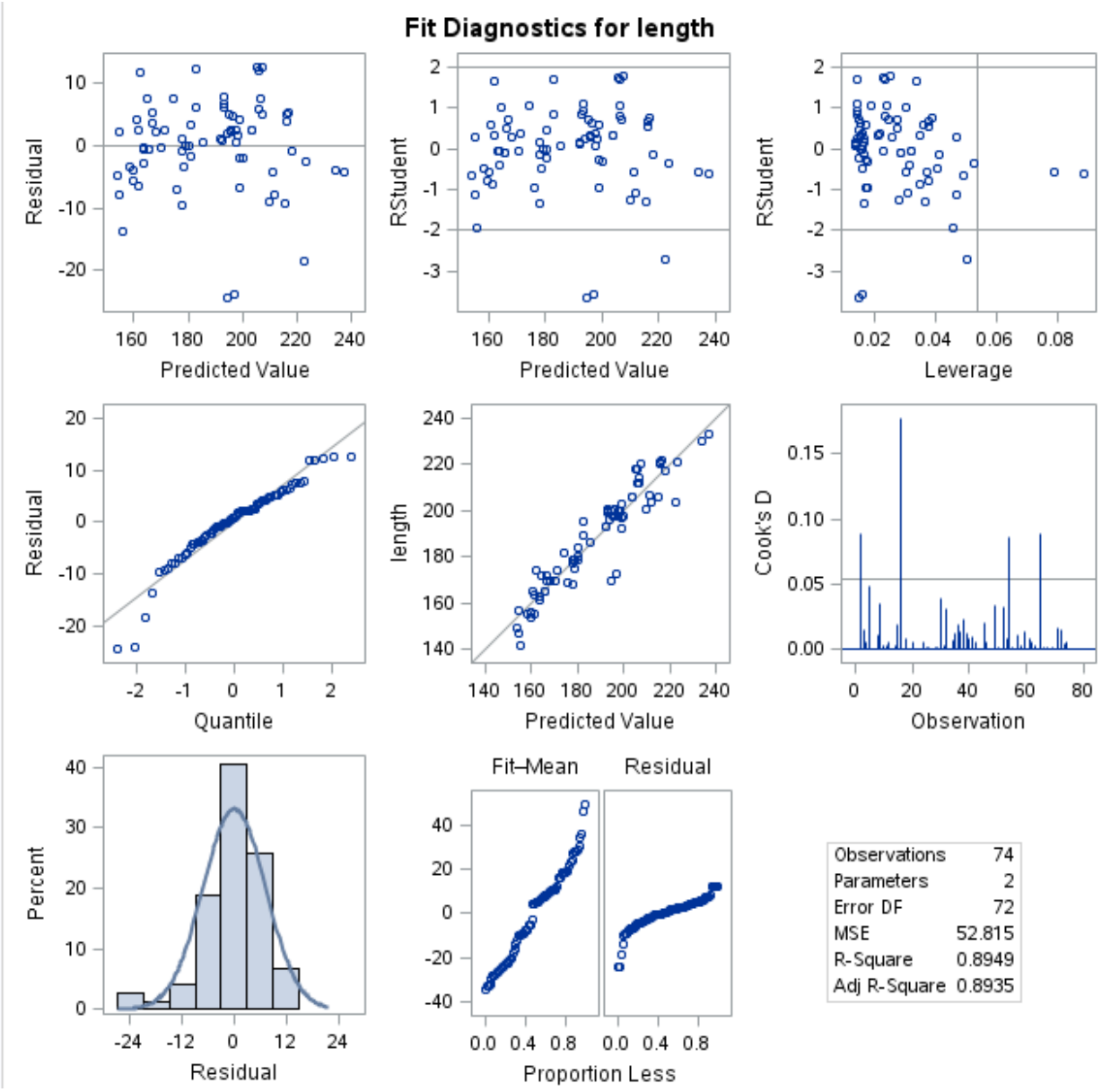
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	106.09652	3.41087	31.11	<.0001
weight	1	0.02710	0.00109	24.76	<.0001

Regression model for length:

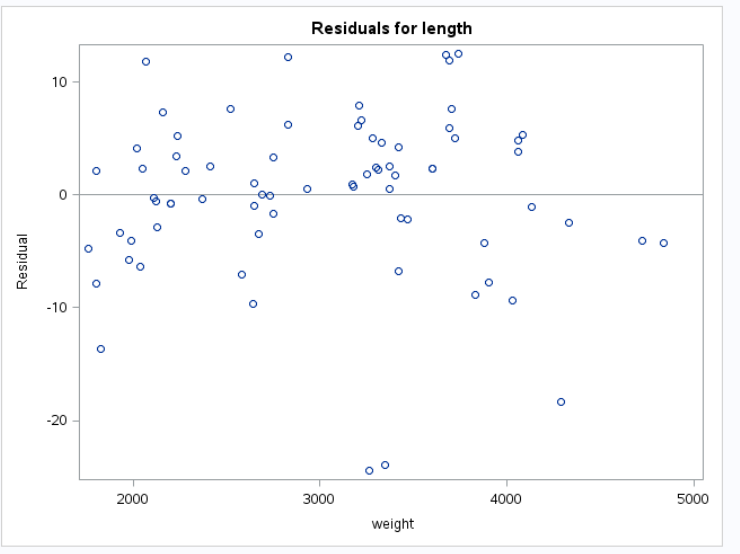
Length^ = 0.02710*weight+106.09652

If weight =6000 pound So estimated length would be:

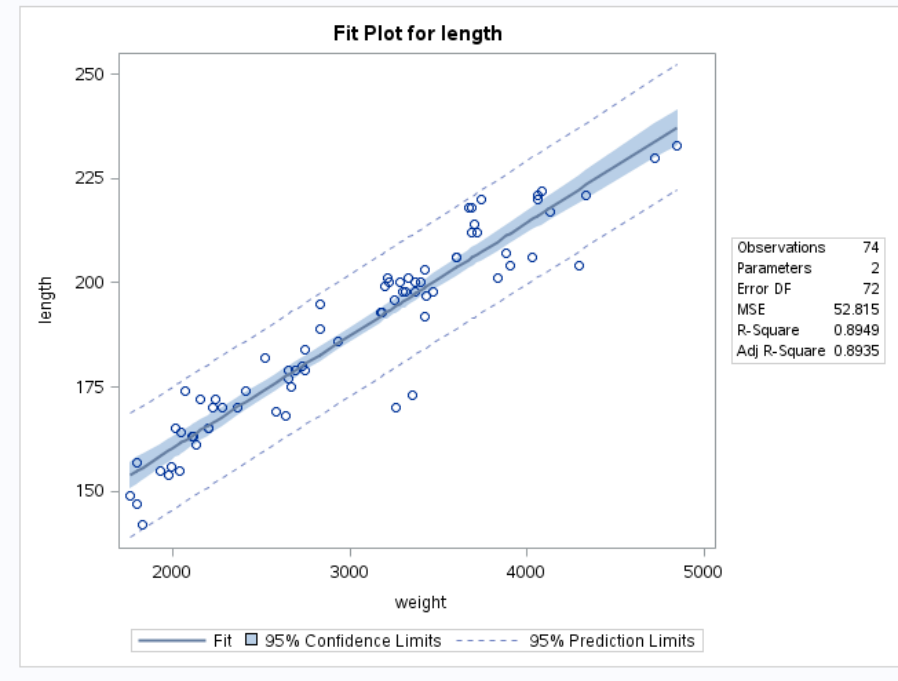
Length^=268.69652



Most of the data points are on or very closer to the line. We can assume residual are normal.



Residuals are randomly distributed around the zero line. Therefore, variances of residuals are homogeneous.



R^2=0.8949 So this is very good model for predict the length of the car using weight.

3)What is the average length difference of foreign cars to domestic cars? What is the average weight difference of foreign cars relative to domestic cars?

```
proc means data=sa_hm.auto2;
       class foreign;
              var length;
run;
```

The MEANS Procedure

Analysis Variable : length						
foreign	N Obs	N	Mean	Std Dev	Minimum	Maximum
0	52	52	196.1346154	20.0460537	147.0000000	233.0000000
1	22	22	168.5454545	13.6825481	142.0000000	193.0000000

Average length difference of foreign cars to domestic cars:

Length of domestic car -Length of foreign car =27.5891609

Average length of domestic car is 27.5891609 higher than foreign car

```
proc means data=sa_hm.auto2;
    class foreign;
    var weight;
run;
```

The MEANS Procedure

Analysis Variable : weight						
foreign	N Obs	N	Mean	Std Dev	Minimum	Maximum
0	52	52	3317.12	695.3637404	1800.00	4840.00
1	22	22	2315.91	433.0034542	1760.00	3420.00

Average weight difference of foreign cars to domestic cars:

weight of domestic car -weight of foreign car =1001.21 pound

Average weight of domestic car is 1001.21 pound higher than foreign car

4)After controlling for the weight of a car, how much shorter is a foreign car on average? Is this a statistically significant difference? (Hint: This will require a multivariate regression)

```
proc reg data=sa_hm.auto2;
   model length= weight foreign;
run;
```

Model: MODEL1
Dependent Variable: length

Number of Observations Read	74
Number of Observations Used	74

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	32395	16197	302.81	<.0001
Error	71	3797.77267	53.48976		
Corrected Total	73	36193			

Root MSE	7.31367	R-Square	0.8951
Dependent Mean	187.93243	Adj R-Sq	0.8921
Coeff Var	3.89165		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	107.04581	4.64861	23.03	<.0001
weight	1	0.02686	0.00137	19.64	<.0001
foreign	1	-0.69945	2.30975	-0.30	0.7629

Multivariate regression model:

Domestic car: 0
Foreign car: 1

Length= 0.02686*weight-0.69945*foreign+107.04581

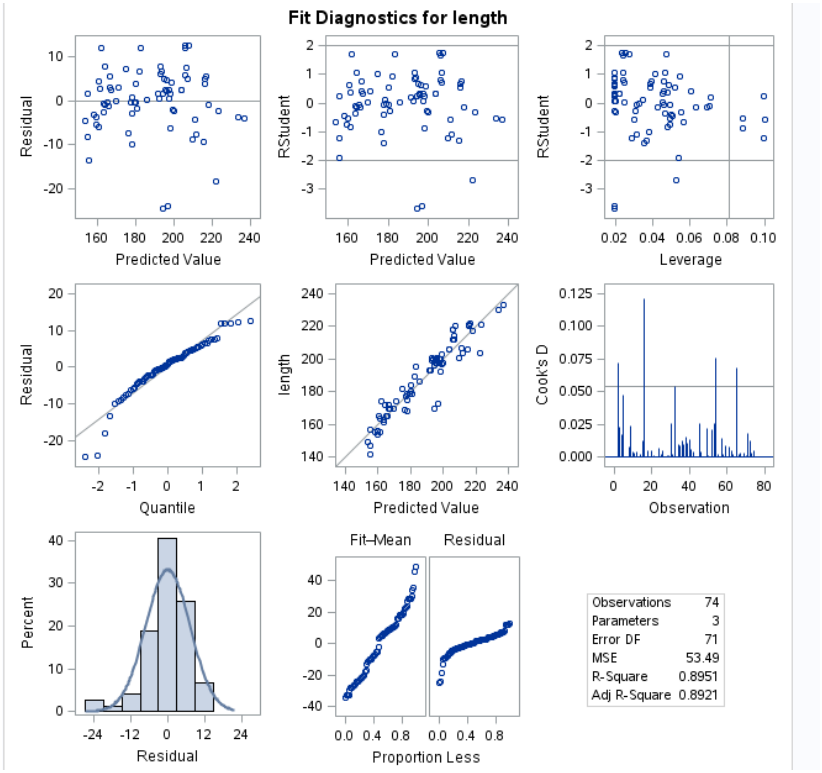
Average length of the foreign car is **0.69945** less than the domestic car after controlling the weight of the car.

Not significant because of high p-value and lower t value

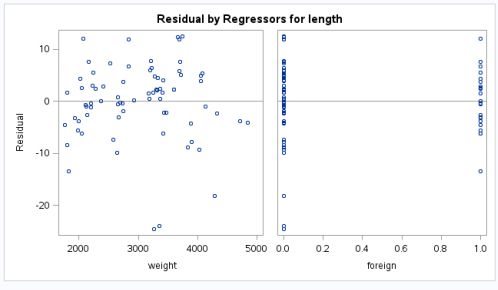
ß2=coefficient of foreign
H0: ß2=0
H1: ß2≠0

Test statistic t =-0.30
p-value=0.7629>0.05
Do no reject H0

Distinction of length of foreign car and domestic car is not significant for fixed weight.



Most of the data points are on or very closer to the line. We can assume residual are normal.



5)Do foreign cars provide better mileage per gallon, on average? Why or why not?

```
proc means data=sa_hm.auto2;
       class foreign;
              var length weight mpg price;
run;
```

The MEANS Procedure

foreign	N Obs	Variable	N	Mean	Std Dev	Minimum	Maximum
0	52	length	52	196.1346154	20.0460537	147.0000000	233.0000000
		weight	52	3317.12	695.3637404	1800.00	4840.00
		mpg	52	19.8269231	4.7432972	12.0000000	34.0000000
		price	52	6072.42	3097.10	3291.00	15906.00
1	22	length	22	168.5454545	13.6825481	142.0000000	193.0000000
		weight	22	2315.91	433.0034542	1760.00	3420.00
		mpg	22	24.7727273	6.6111869	14.0000000	41.0000000
		price	22	6384.68	2621.92	3748.00	12990.00

Yes, foreign cars provide better mileage per gallon, on average.

Average MPG of foreign car is 24.7727273

Average MPG of domestic car is 19.8269231

On average foreign car has 4.9458042 higher MPG than domestic car.

```
PROC CORR DATA=sa_hm.auto2;
   VAR mpg weight length foreign;
RUN;
```

Pearson Correlation Coefficients, N = 74
Prob > |r| under H0: Rho=0

	mpg	weight	length	foreign
mpg	1.00000	-0.80717 <.0001	-0.79578 <.0001	0.39340 0.0005
weight	-0.80717 <.0001	1.00000	0.94601 <.0001	-0.59283 <.0001
length	-0.79578 <.0001	0.94601 <.0001	1.00000	-0.57020 <.0001
foreign	0.39340 0.0005	-0.59283 <.0001	-0.57020 <.0001	1.00000

If we can compare the correlation and mean value of length and weight for foreign and domestic car.

Length and weight have strong negative correlation with MPG. Since domestic car has higher length and weight it should has lower PMG compare with foreign car. Even though correlation between MPG and foreign is weak, it is positive. It means that foreign car has higher MPG compared with domestic cars.

If we build the regression model for mpg with weight.

```
proc reg data=sa_hm.auto2;
model mpg= weight;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: mpg

Number of Observations Read	74
Number of Observations Used	74

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			

Root MSE	3.43889	R-Square	0.6515
Dependent Mean	21.29730	Adj R-Sq	0.6467
Coeff Var	16.14707		

Parameter Estimates

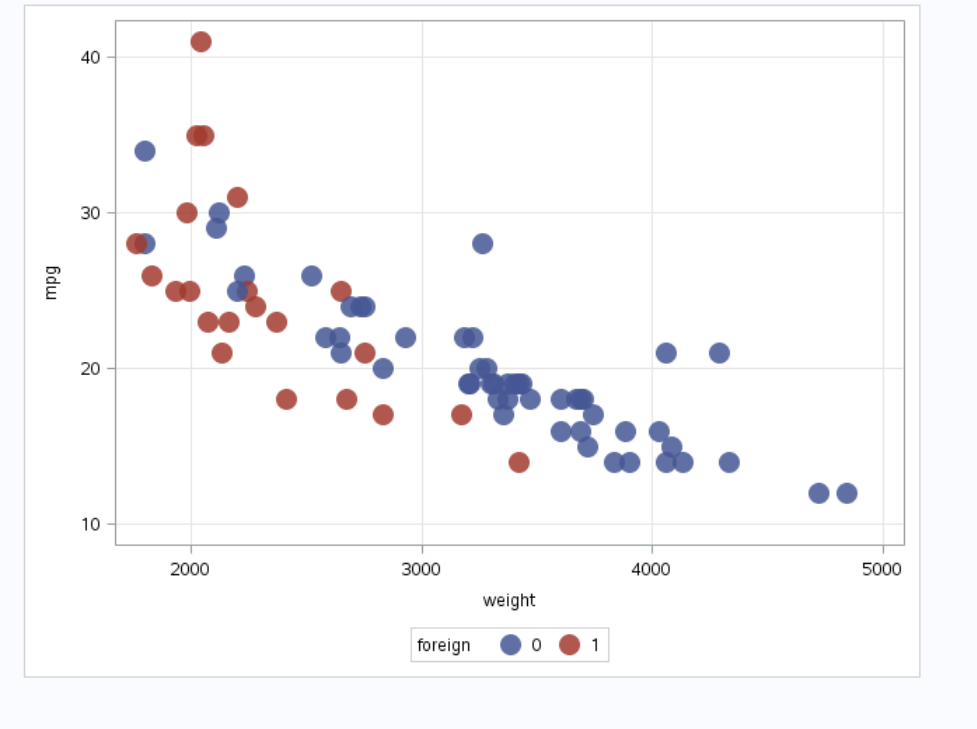
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	39.44028	1.61400	24.44	<.0001
weight	1	-0.00601	0.00051788	-11.60	<.0001

Mpg^= -0.00601*weight+39.44028

Since the domestic car has higher weight, domestic car has lower mpg compared with foreign car.

```
proc sgplot data=sa_hm.auto2;
   scatter x=weight y=mpg / group=foreign
      markerattrs=(symbol=circlefilled size =15) transparency=0.15;
   xaxis grid;
   yaxis grid;
run;
```

Domestic car: 0 (blue)
Foreign car: 1 (red)



If cars have lower weight so that cars have higher mpg according to the graph.

In here we can see that foreign car has lower weight and higher mpg compared with domestic cars.

The reasons are that domestic cars have lower MPG compared with foreign cars.

1 Higher weight of domestic cars compared with foreign cars.

2 Higher length of domestic cars compared with foreign cars.

6)After controlling for the weight and length of a car, do foreign cars provide better mileage per gallon? Is this a statistically significant finding? With 95 percent confidence, what is the estimated range of the average effect that being foreign-made has on a car's mileage per gallon?

```
proc glm data=sa_hm.auto2;
model mpg= weight length foreign/solution clparm;
run;
```

The GLM Procedure

Dependent Variable: mpg

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1645.288898	548.429632	48.10	<.0001
Error	70	798.170563	11.402437		
Corrected Total	73	2443.459459			

R-Square	Coeff Var	Root MSE	mpg Mean
0.673344	15.85530	3.376749	21.29730

Source	DF	Type I SS	Mean Square	F Value	Pr > F
weight	1	1591.990203	1591.990203	139.62	<.0001
length	1	24.090421	24.090421	2.11	0.1505
foreign	1	29.208272	29.208272	2.56	0.1140

Source	DF	Type III SS	Mean Square	F Value	Pr > F
weight	1	84.74324180	84.74324180	7.43	0.0081
length	1	26.00119804	26.00119804	2.28	0.1355
foreign	1	29.20827198	29.20827198	2.56	0.1140

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	50.53701292	6.24583515	8.09	<.0001	38.08008750	62.99393835
weight	-0.00436563	0.00160137	-2.73	0.0081	-0.00755947	-0.00117179
length	-0.08274318	0.05479417	-1.51	0.1355	-0.19202670	0.02654035
foreign	-1.70790387	1.06711033	-1.60	0.1140	-3.83618832	0.42038057

Mpg^=-0.00436563*weight-0.08274318*length-1.70790387*foreign+50.53701292

After controlling for the weight and length of a car, foreign cars do not provide better mileage per gallon according to the above model. Foreign cars have 1.70790387 less mpg than domestic cars on average according to the model after controlling the weight and length.

Not significant because of high p-value and lower t value

ß3=coefficient of foreign
H0: ß3=0
H1: ß3≠0

Test statistic t =-1.6
p-value=0.1140>0.05
Do no reject H0
Distinction of mpg of foreign car and domestic car is not significant for fixed weight and length.

This is not a statistically significant finding

Estimated range of the average effect that being foreign-made has on a car's mileage per gallon: -3.83618832 to 0.42038057

95% confidence interval for mean mpg for domestic cars and foreign cars.

```
proc sql;
create view work.cars as
  select * from sa_hm.auto2
    order by foreign;
quit;
proc means data=work.cars lclm uclm alpha=0.05;
  var mpg;
by foreign;
run;
```

The MEANS Procedure

foreign=0

Analysis Variable : mpg	
Lower 95% CL for Mean	Upper 95% CL for Mean
18.5063807	21.1474655

foreign=1

Analysis Variable : mpg	
Lower 95% CL for Mean	Upper 95% CL for Mean
21.8414912	27.7039633

7)Create a summary table that reports average price of a car by its repair status ranking and whether or not it is foreign or domestic.

```
proc means data=sa_hm.auto2;
class foreign rep78;
var price;
run;
```

The MEANS Procedure

Analysis Variable : price							
foreign	rep78	N Obs	N	Mean	Std Dev	Minimum	Maximum
0	1	2	2	4564.50	522.5519113	4195.00	4934.00
	2	8	8	5967.63	3579.36	3667.00	14500.00
	3	27	27	6607.07	3661.27	3291.00	15906.00
	4	9	9	5881.56	1592.02	3829.00	8814.00
	5	2	2	4204.50	311.8340905	3984.00	4425.00
1	3	3	3	4828.67	1285.61	3895.00	6295.00
	4	9	9	6261.44	1896.09	3995.00	9735.00
	5	9	9	6292.67	2765.63	3748.00	11995.00

```
proc means mean data=sa_hm.auto2;

class foreign rep78;
var price;
run;
```

The MEANS Procedure

Analysis Variable : price			
foreign	rep78	N Obs	Mean
0	1	2	4564.50
	2	8	5967.63
	3	27	6607.07
	4	9	5881.56
	5	2	4204.50
1	3	3	4828.67
	4	9	6261.44
	5	9	6292.67

8)What are some of the most difficult or confusing aspects of SAS to you so far?

How to decode the categorical variable back to its normal name?

Calculate the mean difference by categorical variable directly without doing manually.

Can we analyze the big data set with SAS as Python?

9) **[ADVANCED QUESTION]** Create a summary tables that details the average price of a car by manufacturer of a car (note, this is not the make of the car). What is the highest priced manufacturer, on average? What is the lowest, on average?

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

The MEANS Procedure

Analysis Variable : price						
manufacturer	N Obs	N	Mean	Std Dev	Minimum	Maximum
AMC	3	3	4215.67	485.6267428	3799.00	4749.00
Audi	2	2	7992.50	2400.63	6295.00	9690.00
BMW	1	1	9735.00	.	9735.00	9735.00
Buick	7	7	6075.29	2257.92	4082.00	10372.00
Cad.	3	3	13930.33	2313.71	11385.00	15906.00
Chev.	6	6	4372.33	911.3044863	3299.00	5705.00
Datsun	4	4	6006.50	1573.12	4589.00	8129.00
Dodge	4	4	5055.50	1236.39	3984.00	6342.00
Fiat	1	1	4296.00	.	4296.00	4296.00
Ford	2	2	4288.00	142.8355698	4187.00	4389.00
Honda	2	2	5149.00	919.2388155	4499.00	5799.00
Linc.	3	3	12852.33	1175.50	11497.00	13594.00
Mazda	1	1	3995.00	.	3995.00	3995.00
Merc.	6	6	4913.83	1239.38	3291.00	6303.00
Olds	7	7	6050.86	2486.49	4181.00	10371.00
Peugeot	1	1	12990.00	.	12990.00	12990.00
Plym.	5	5	4820.00	955.6874489	4060.00	6486.00
Pont.	6	6	4878.83	582.4851643	4172.00	5798.00
Renault	1	1	3895.00	.	3895.00	3895.00
Subaru	1	1	3798.00	.	3798.00	3798.00
Toyota	3	3	5122.00	1193.32	3748.00	5899.00
VW	4	4	6021.00	1166.44	4697.00	7140.00
Volvo	1	1	11995.00	.	11995.00	11995.00

The highest priced manufacturer is “Cad” ($13930.33) on average

The lowest priced manufacturer is “Subaru” ($3798) on average