Q1. Using the 1993 car data, do the following analysis in SAS. Please do not modify the original data using Excel or any other software. Your SAS code should run on the original data provided.

a. Find out the correlation between horsepower and midrange price. Comment on your findings (Is the correlation high or low? Is it significantly different from zero?).

Horsepower and midrange price are highly correlated. When the correlation was run, a correlation coefficient of 0.7882 was observed between horsepower and midrange price which falls in the large correlation area (.5 < | r |< ....). And we clearly see that it is not significantly different from zero.

b. Run a regression model to find out what factors influence the midrange price of the car. Some explanatory variables that you can use are *city MPG, Air bags standard, horsepower, Manual transmission, domestic or not.*

We run an OLS regression with

Indepedent variables : *city MPG, Air bags standard, horsepower, Manual transmission, domestic*

Dependent variable: *Midrange Price*

c. Based on the regression output of the above model, please answer the following questions:

1. **Comment on the model fit (Is the model a good model? Why do you think so?)**

Model Fit:

OLS Regression model with midrange price of the car as the dependent variable and city MPG, Air bags standard, horsepower, Manual transmission, domestic as independent variables, we find that R^2 =0.7256 and Adj. R^2= 0.7098.

- R^2 = 0.7256 implies that all of the five explanatory variables together explain 72.56% of the variance in the dependent variable - midrange price of the car.
- We think that the model has a reasonable fit, first and foremost because the p-value is 0.0001 which is much smaller than 0.05. And secondly as the value of R^2 is above 30%, **the model has a reasonable fit.**

2. **What is the interpretation of R^2 and adjusted R^2? Why do we need an adjusted R^2?**
   - The obtained value of R^2 and adjusted R^2 suggests that the 5 independent variables chosen above explain nearly 71% of variation in the dependent variable.
   - We need an Adjusted R^2, because as we add more explanatory variables to the model, adjusted R^2 increases and could in the limit become equal to 1. Adjusted R^2 imposes a penalty on any variable added to the model that has a very small explanatory power. Thus, as we add more variables to the model, Adjusted R^2 could actually go down and is therefore a more accurate measure of model fit. If

the difference between R^2 and adjusted R^2 is large (i.e., more than 5% points), adjusted R^2 is the more accurate indicator of model fit.

- Hence, in our case, after adjusted R^2 adds penalty with the 5 variables. The Adjusted R^2 is close to R^2, hence fulfilling the requirements.

3. Which of the variables have significant coefficients?

The variables CityMpg, Airbags, HorsePower and Domestic seem to have significant coefficients as their p values are less than 0.05. Only the variable Manual Transmission has the p value more than 0.05, i.e. 0.07. Thus we consider it to be an insignificant coefficient.

4. What is the interpretation of the coefficient for 'horsepower' and for 'domestic or not'?

For every unit increase in horsepower the midrange price increases by $5170

If the consumer buys a domestic car, then the midrange prices of the domestic variants i.e., the U.S. Manufacturers are $2440 cheaper than the international variants i.e., the Non-U.S. Manufacturers.

5. Of the above variables, what is the most important variable that affects the price? How did you determine this?

To check which variables affect the Midrange Price we use Standardized Regression Coefficients.

The SAS System

The REG Procedure
Model: Orig
Dependent Variable: MidrangePrice

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	19.50968	0.53960	36.16	<.0001	0
CityMPG	1	-1.57717	0.78124	-2.02	0.0466	-0.16328
AirBags	1	2.41761	0.61742	3.92	0.0002	0.25029
Horsepower	1	5.15673	0.80190	6.43	<.0001	0.53385
Manualtransmission	1	-1.20425	0.65695	-1.83	0.0702	-0.12467
Domestic	1	-2.44612	0.61381	-3.99	0.0001	-0.25324

From the above results we can predict the Horsepower to be the most important variable as its Standardized Estimate is 0.53385 which is higher than the Standardized Estimates of other variables. After the standardization we understand that the normalized coefficients help us determine the variable that best affects Midrange Price.

6. **What is the elasticity of midrange price with respect to HP (horsepower) (Elasticity is defined as the percentage change in midrange price due to a percent change in HP)?**

   We interpret from running a log-log model between Mid Range Price and Elasticity that one percent increase in Mid Range Price would yield 1.08% increase in HorsePower (HP)

7. **Check whether Horsepower has a non-linear effect on midprice? What do you find/conclude?**

   Running GLM on the variables, we use the squared term of Horsepower to find non-linearity. The p-value for squared term of Horsepower is > 0.05. Hence we fail to reject the null hypothesis. We conclude after running GLM that the relationship is not non-linear.

8. **Check if there is an interaction between HP and weight of the car. What do you conclude?**

We add an interaction variable Hp_weight to the regression model and run Horsepower and weight against the interaction variable.

We find that the p-value for the interaction variable is >0.05, hence we cannot reject the null hypothesis. The parameter estimate for interaction varaible Hp_Weight is not significant and is low, which is 0.00000361. Hence, we conclude that there is not a significant interaction effect between Horsepower and Weight.

d. What other variables do you think should be included in the regression model? Run alternate regressions including additional variables to improve the model fit. Which additional variables are significant and what is the interpretation of these new coefficients?

We think that Type, EngineSize, NumberOfCylinders should be included in the regression model

Regression Model 1:

```
proc GLM data = cars2;
class Type;
model MidrangePrice = Type EngineSize NumberOfCylinders cityMPG
Airbags horsepower Manualtransmission domestic;run;
```

Variables: *Type EngineSize NumberOfCylinders* cityMPG Airbags horsepower Manualtransmission domestic

R^2 = 0.752132

With the chosen variables R^2 explains 75% of the variance in the model.

Regression Model 2:

```
proc GLM data = cars2;
class Manufacturer;
model MidrangePrice = Manufacturer EngineSize NumberOfCylinders
cityMPG horsepower Manualtransmission domestic;run;
```

Variables: *Manufacturer EngineSize NumberOfCylinders* cityMPG horsepower Manualtransmission domestic

R^2 = 0.913563

With the chosen variables R^2 explains 91% of the variance in the model

Regression Model 3:

```
proc GLM data = cars2;
class Manufacturer Type;
model MidrangePrice = Manufacturer Type EngineSize NumberOfCylinders
horsepower Manualtransmission domestic;
```

Variables: *Manufacturer Type EngineSize NumberOfCylinders* horsepower Manualtransmission domestic

R^2 = 0.933024

With the chosen variables R^2 explains 93% of the variance in the model

Regression Model 2 seems to be the best model so far and we have been able to improve the model fit considerably by using 7 variables.

Summary:

We checked for the additional variables, *EngineSize, Type, Manufacturer, NumberOfCylinders* and find that these additional variables create some improvement in the model. Especially Manufacturer and Type, they explain the model the best and we attain an R^2 of 0.93.

EngineSize and NumberOfCylinders do have an impact over the R^2 but not so much as the other 2 variables that have been added in the model.

The added variables are significant as they are above the threshold p-value 0.05. These added variables increase the model fit and adjusted R^2 being closer assures us about the model fit.

Obviously Type and Manufacturer are clearly one of the most important variables that people look for when buying cars. This explains why the model works so much better when we introduce these variables into our model.

Q2. The file "diamond.dat" has data on the cut, color, clarity, carat and prices of diamonds in US dollars.

Cut is classified as: Fair, Good, Very good and Ideal (These levels are in order of quality of cut with 'Ideal' being the best and 'Fair' being the worst)
Color: D and E
Clarity is classified as: VVS1, VVS2, VS1 and VS2
(These levels are in order of clarity with VVS1 being the best and VS2 the worst)

We are interested in finding out using dummy variable regressions how to price different combinations of attributes of diamonds.

1. **Is there a relationship between cut and clarity? Run a test and discuss what you find/conclude.proc**
   We run chi square into our model and check for the relationship. As seen in the model, we can reject the null hypothesis and conclude that there is a relationship between cut & clarity

2. **Is there a difference in price between D and E color? Find out using a t-test. What do you find?**

   First, we need to perform an f test to check the for equality of variances:
   H_o= Variances are equal
   H_1= Variances are not equal

- When the p-value (shown under "Pr>F") is greater than 0.05, then the variances are **equal** then read the **"Pooled"** section of the result

- When the p-value (shown under "Pr>F") is no more than 0.05, then the variances are **unequal** then read the **"Satterthwaite"** section of the result

As the p-value is more than 0.05 we conclude that we cannot reject the null hypothesis and that the variances are equal.

Null hypothesis: there is no price difference between D&E
Alternate hypothesis: There is a significant price difference between color D&E

As the p value is less than 0.05, we reject the null hypothesis and conclude that there is significant price difference between the color D&E.

3. Using price as a dependent variable, run a regression model (USE PROC REG) to answer the following questions. Note that you have to create dummy variables for cut, color and clarity before you can do the regression.

a. Is there a significant difference in prices between color "D" and "E"?

Ho= There is no significant difference in prices between color D & E
H1= There is a significant difference in prices between color D & E

We observe from the t-test that f-svalue is less than 0.05, thus we reject the null hypothesis and conclude that there is a significant difference of prices between the color D & E.

b. How much more price would an Ideal cut command over a Good cut diamond?

Note: As there are 4 variants in Cut and Clarity, we take N-1 dummy variables, that is 3 dummy variables for each.

Cut Dummy variables:

Cut_Fair
Cut_Good
Cut_Ideal

Clarity Dummy variables:

Clarity_VVS1
Clarity_VVS2
Clarity_VS1

There are 2 variants of Color, hence we use just 1 dummy variable for Color = Color1

If the cut is Ideal, the price of diamond will be $266.51 higher than Good cut diamond.

c. How much more price would a VVS2 command over a VS1 diamond?

VVS2 diamond will command $2820.52 more than VS1 diamond

d. Which variables are significant?

All the variables are significant except for color of the diamond. The t-value of Color is -1.00 which is not significant and we can't reject the null hypothesis at 95% Confidence Interval.

e. Comment on the model fit.

Model Fit:

Regression model with *prices* of diamond as the dependent variable and *Cut_Fair Cut_Good Cut_Ideal Clarity_VVS1 Clarity_VVS2 Clarity_VS1 Color1 Carat* as independent variables, we find that R^2 =0.9359 and Adj. R^2= 0.9338

- R-Square = 0.9359 implies that all of the eight explanatory variables together explain 93.59% of the variance in the dependent variable prices of the diamond.
- We think that the model has a reasonable fit first and foremost because the p-value is 0.0001 which is much smaller than 0.5. And secondly as the value of R^2 is above 30%, **the model has a reasonable fit.**
- Adjusted R^2 also explains that with the addition of each explanatory variable its value is not going down or is not very different from R^2 and we conclude that the model has a reasonable fit.