Multicollinearity Analysis

Dataset Background:

This dataset contains information on various models of cars taken from the 1974 Motor Trends US magazine. The goal is to predict the mileage of the cars given the different specifications.

Dataset Glimpse:

model	mpg	cyl	disp	hp	drat	wt
Mazda RX4	21	6	160	110	3.9	2.62
Mazda RX4 Wag	21	6	160	110	3.9	2.875
Datsun 710	22.8	4	108	93	3.85	2.32
Hornet 4 Drive	21.4	6	258	110	3.08	3.215
Hornet Sportabout	18.7	8	360	175	3.15	3.44
Valiant	18.1	6	225	105	2.76	3.46
Duster 360	14.3	8	360	245	3.21	3.57
Merc 240D	24.4	4	146.7	62	3.69	3.19
Merc 230	22.8	4	140.8	95	3.92	3.15
Merc 280	19.2	6	167.6	123	3.92	3.44

Total Number of Rows: 32

Total Number of Columns: 8

Column Details:

- model – Identifier to identify the car.
- mpg – Miles per Gallon, or mileage of the car.
- cyl – Number of Cylinders in the car.
- disp – Displacement (in cu. Inches).
- hp – Gross Horsepower.
- drat – Rear Axle Ratio.
- wt – Weight of the car (in thousand lbs).

Main Dependent Variable = mpg.

Using the SPSS Software EViews, we have analysed the data:

Descriptive Statistics:

	MPG	CYL	DISP	HP	DRAT	WT
Mean	20.09063	6.187500	230.7219	146.6875	3.596563	3.217250
Median	19.20000	6.000000	196.3000	123.0000	3.695000	3.325000
Maximum	33.90000	8.000000	472.0000	335.0000	4.930000	5.424000
Minimum	10.40000	4.000000	71.10000	52.00000	2.760000	1.513000
Std. Dev.	6.026948	1.785922	123.9387	68.56287	0.534679	0.978457
Skewness	0.640440	-0.183129	0.400272	0.761436	0.278873	0.443786
Kurtosis	2.799467	1.319032	1.910317	3.052233	2.435116	3.172471
Jarque-Bera	2.241155	3.946399	2.437707	3.095820	0.840234	1.090038
Probability	0.326091	0.139011	0.295569	0.212692	0.656970	0.579831
Sum	642.9000	198.0000	7383.100	4694.000	115.0900	102.9520
Sum Sq. Dev.	1126.047	98.87500	476184.8	145726.9	8.862322	29.67875
Observations	32	32	32	32	32	32

Inferences:

- The variable mpg is extremely right skewed, ranging between 10.4 to 33.9 miles per gallon.
- The variable cyl is slightly left skewed, ranging between 4 to 8 cylinders.
- The variable disp is slightly right skewed, ranging between 71 to 472 cubic inches.
- The variable hp is slightly right skewed, ranging between 52 to 335 hp.
- The variable drat is slightly right skewed, ranging between 2.76 to 4.53.
- The variable wt is slightly right skewed, ranging between 1.51 to 5.42 thousand lbs.
- There is no missing data.

Correlation Analysis:

	MPG	CYL	DISP	HP	DRAT	WT
MPG	1.000000	-0.852162	-0.847551	-0.776168	0.681172	-0.867659
CYL	-0.852162	1.000000	0.902033	0.832447	-0.699938	0.782496
DISP	-0.847551	0.902033	1.000000	0.790949	-0.710214	0.887980
HP	-0.776168	0.832447	0.790949	1.000000	-0.448759	0.658748
DRAT	0.681172	-0.699938	-0.710214	-0.448759	1.000000	-0.712441
WT	-0.867659	0.782496	0.887980	0.658748	-0.712441	1.000000

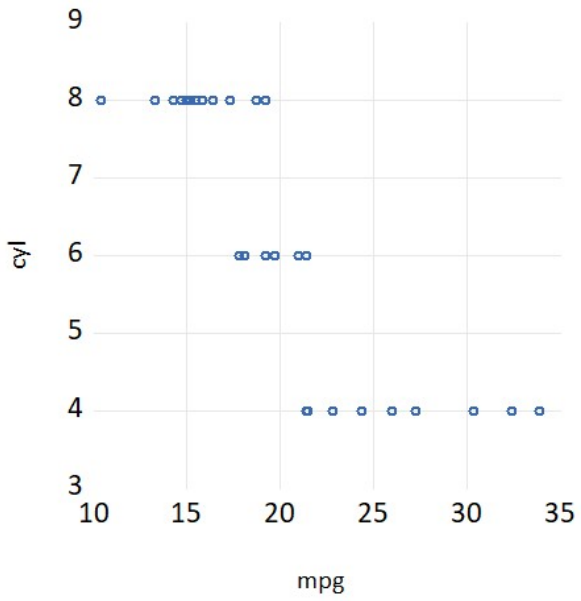
Inferences:

- Variables mpg and cyl have sufficient high degree of negative linear correlation, having correlation coefficient -0.85.
- Variables mpg and disp have sufficient high degree of negative linear correlation, having correlation coefficient -0.84.
- Variables mpg and hp have sufficient high degree of negative linear correlation, having correlation coefficient -0.77.
- Variables mpg and drat have moderate degree of positive linear correlation, having correlation coefficient 0.68.
- Variables mpg and wt have sufficient high degree of negative linear correlation, having correlation coefficient -0.87.
- Variables cyl and disp have very high degree of positive linear correlation, having correlation coefficient value 0.90.
- Variables cyl and hp have sufficient high degree of positive linear correlation, having correlation coefficient 0.83.
- Variables cyl and drat have moderate degree of negative linear correlation, having correlation coefficient -0.69.
- Variables cyl and wt have sufficient high degree of positive linear correlation, having correlation coefficient 0.78.
- Variables disp and hp have sufficient high degree of positive linear correlation, having correlation coefficient 0.79.
- Variables disp and drat have moderate degree of negative linear correlation, having correlation coefficient -0.71.
- Variables disp and wt have sufficient high degree of positive linear correlation, having correlation coefficient 0.88.
- Variables hp and drat have only the possibility of negative linear correlation, having correlation coefficient -0.45.
- Variables hp and wt have moderate degree of positive linear correlation, having correlation coefficient 0.66.
- Variables drat and wt have moderate degree of negative linear correlation, having correlation coefficient -0.71.

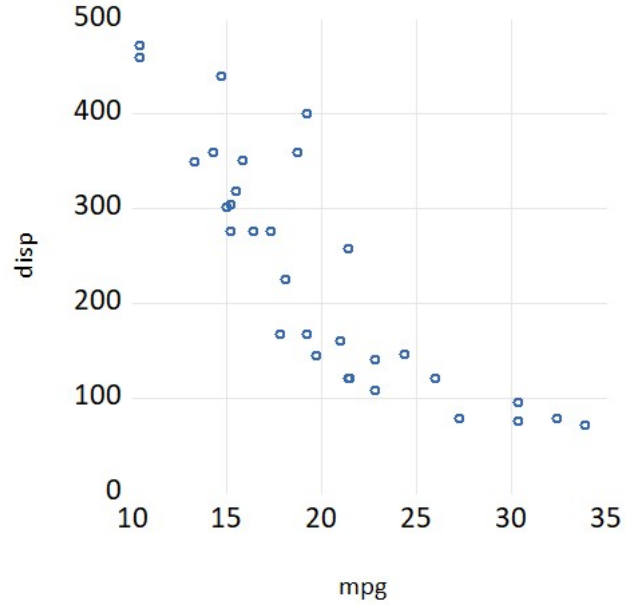
Since the variables are highly correlated with each other, we may infer the problem of multicollinearity. However, to confirm the same, we need to run a few more tests as follows:

- Plot scatter plots and check for correlation.
- Build a regression model and check for very high R^2 value (above 0.9).
- Build a regression model and check for statistically insignificant variables using t-values.

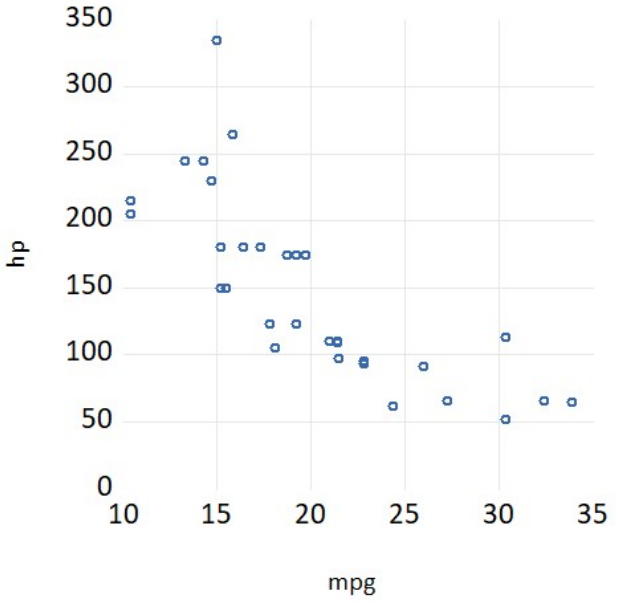
Scatter Plots:



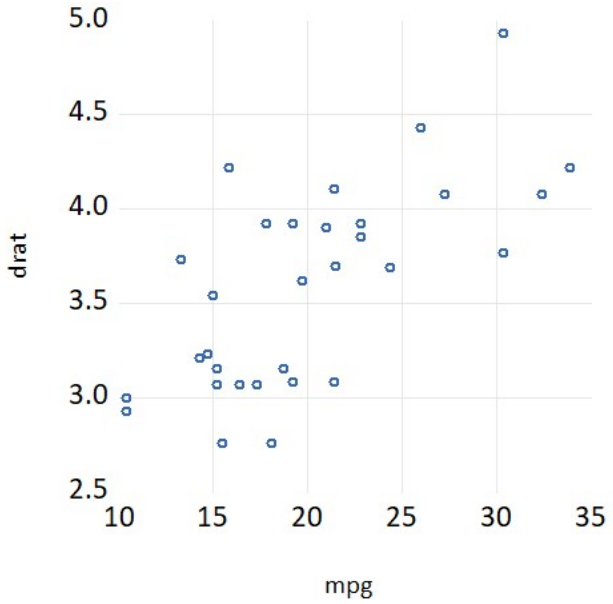
Inference: automobiles with 4 cylinders have the highest mileage.



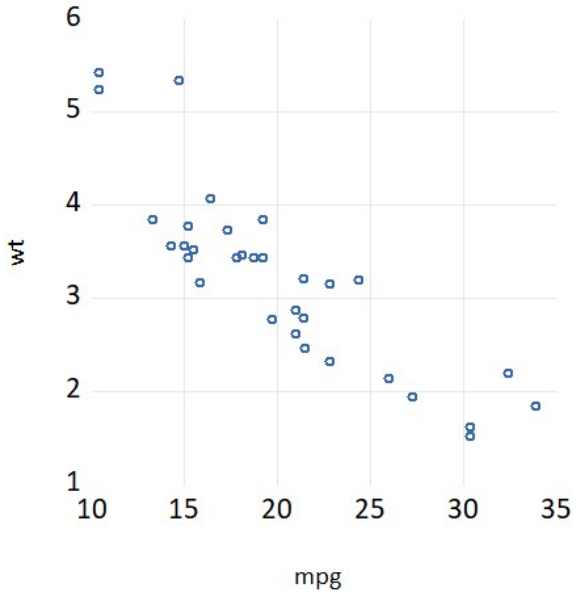
Inference: variables mpg and disp have sufficient high degree of negative linear correlation.



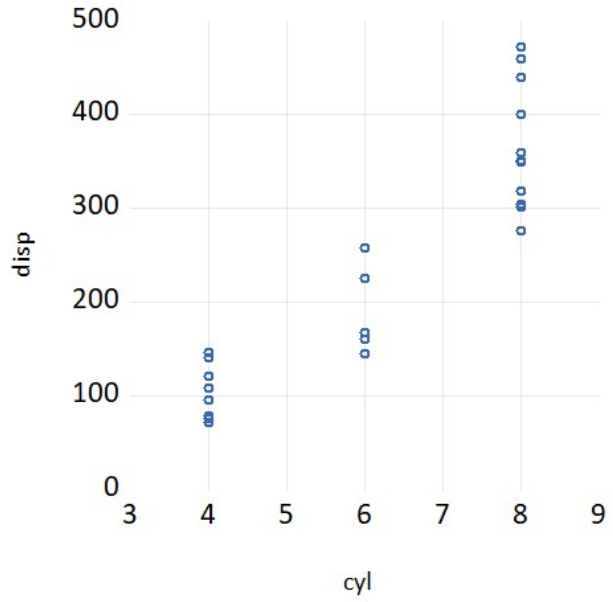
Inference: variables mpg and hp have sufficient high degree of negative linear correlation.



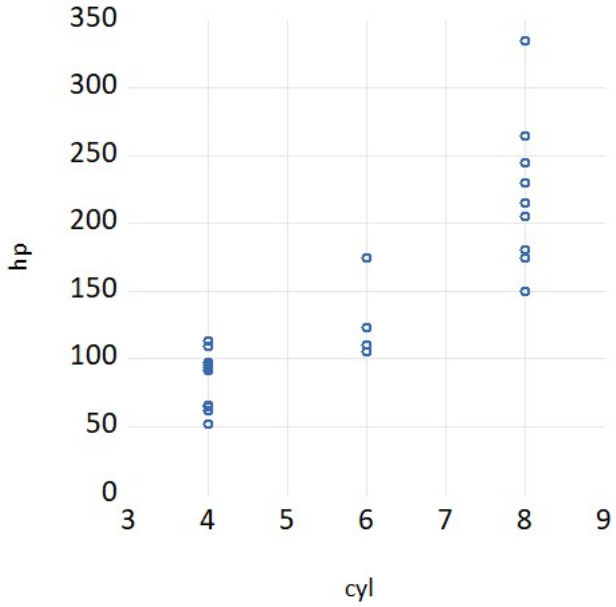
Inference: variables mpg and drat have moderate degree of positive linear correlation.



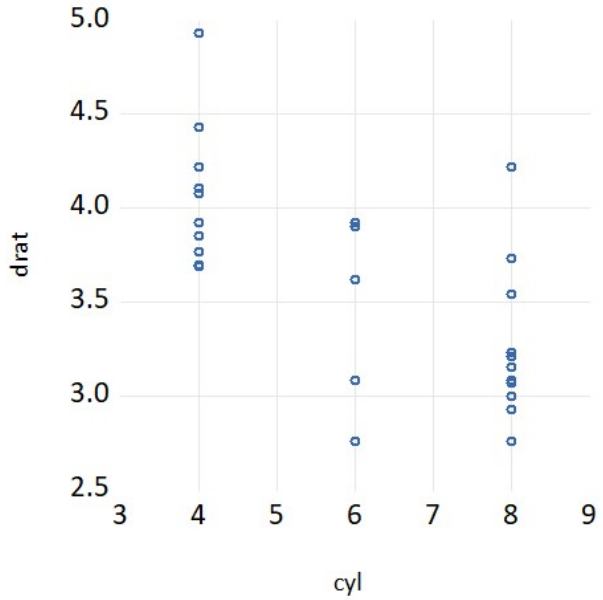
Inference: variables mpg and wt have sufficient high degree of negative linear correlation.



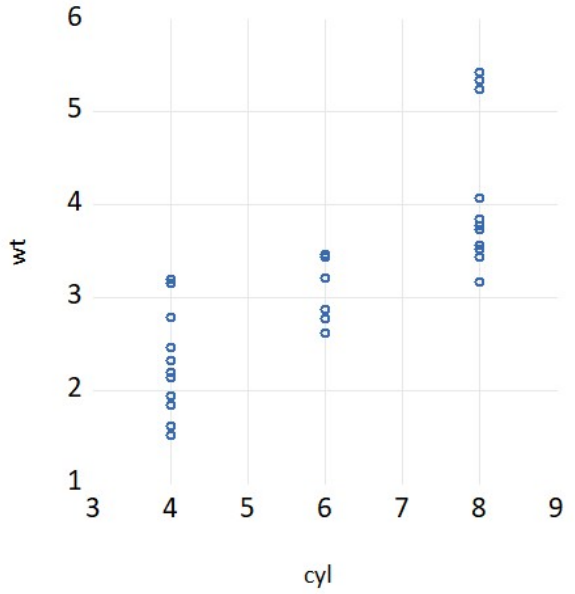
Inference: automobiles having 8 cylinders have the highest displacement.



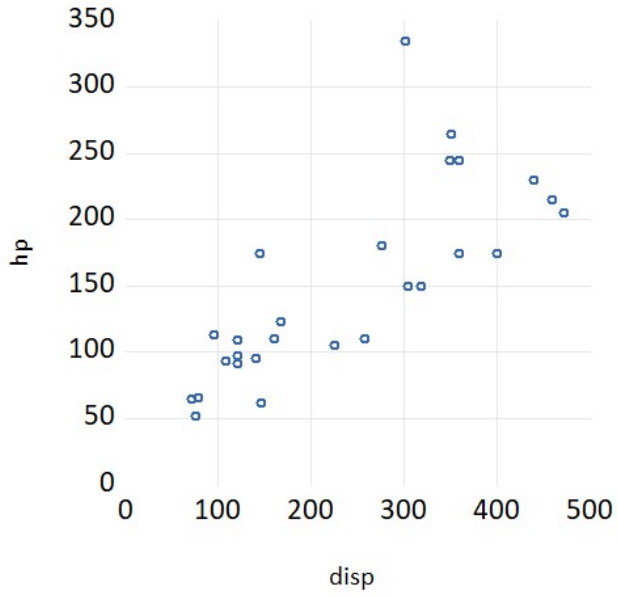
Inference: automobiles with 8 cylinders yield the highest power.



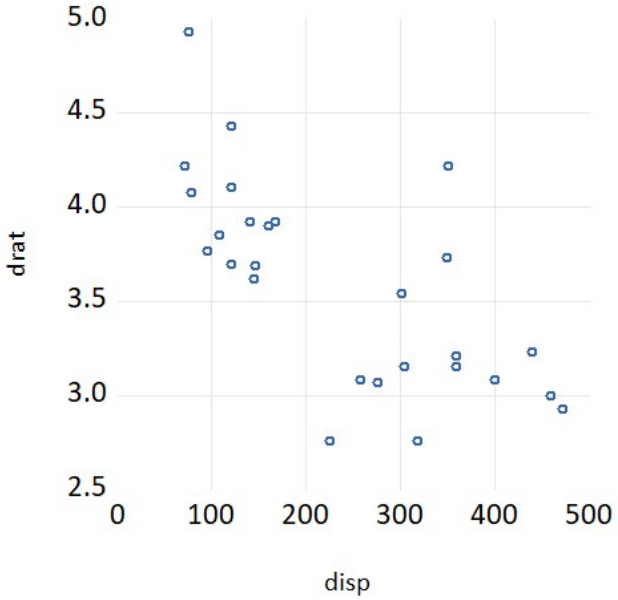
Inference: automobiles with 4 cylinders have the highest rear axle ratio.



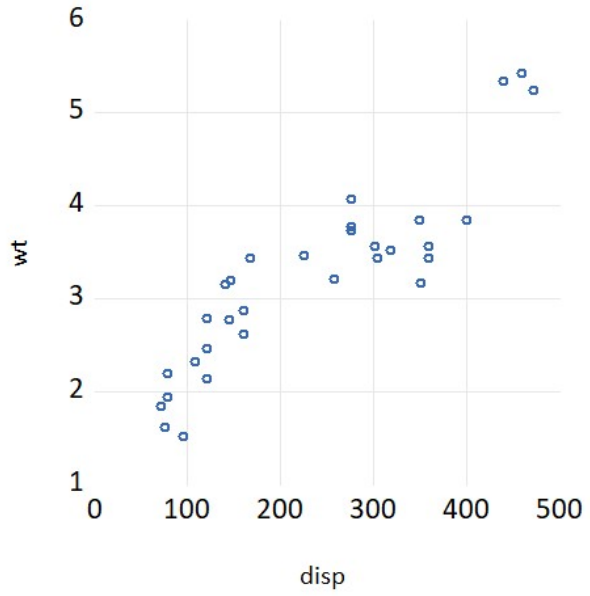
Inference: automobiles with 8 cylinders are the heaviest.



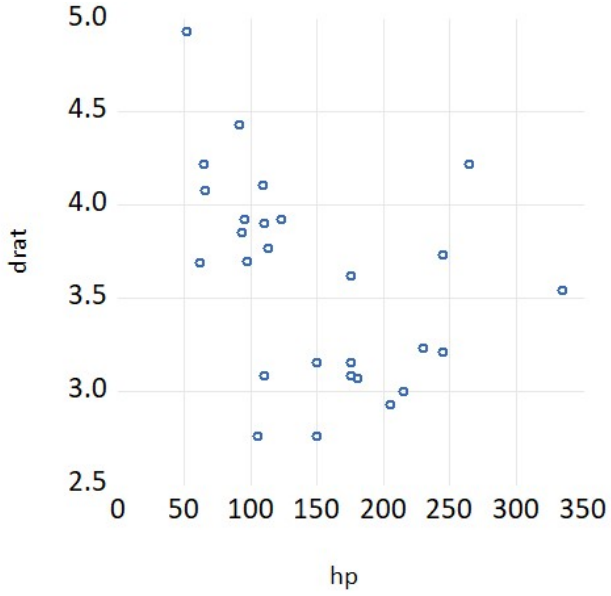
Inference: variables disp and hp have sufficient high degree of positive linear correlation.



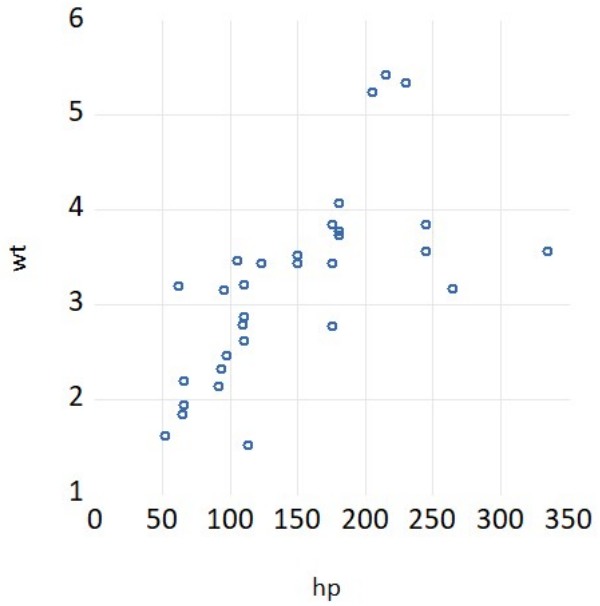
Inference: variables disp and drat have moderate degree of negative linear correlation.



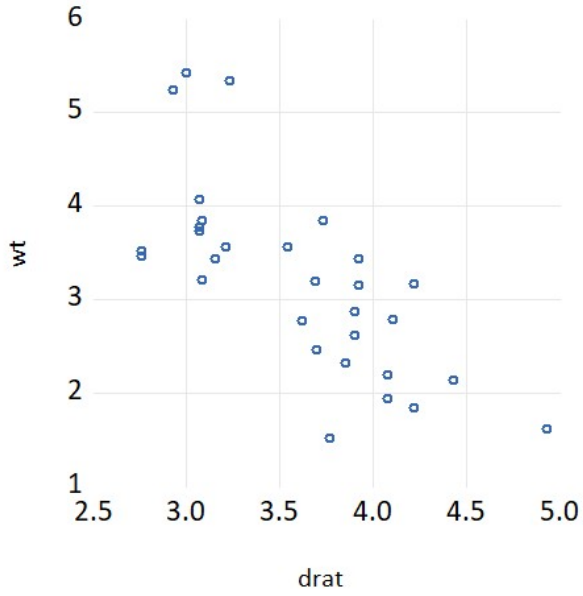
Inference: variables disp and wt have sufficient high degree of positive linear correlation.



Inference: variables hp and drat have only the possibility of negative linear correlation.



Inference: variables hp and wt have moderate degree of positive linear correlation.



Inference: variables drat and wt have moderate degree of negative linear correlation.

From the above scatter plots, we may infer the presence of multicollinearity. However, to confirm the same, we need to perform the following tests:

- Build a regression model and check for very high R^2 value (above 0.9).
- Build a regression model and check for statistically insignificant variables using t-values.

Regression Model:

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	36.00836	7.571436	4.755816	0.0001
CYL	-1.107486	0.715882	-1.547023	0.1339
DISP	0.012357	0.011896	1.038818	0.3085
HP	-0.024017	0.013279	-1.808691	0.0821
DRAT	0.952207	1.390847	0.684624	0.4996
WT	-3.673287	1.059002	-3.468631	0.0018

R-squared	0.981348	Mean dependent var	20.09063
Adjusted R-squared	0.923710	S.D. dependent var	6.026948
S.E. of regression	2.537610	Akaike info criterion	4.867683
Sum squared resid	167.4261	Schwarz criterion	5.142509
Log likelihood	-71.88293	Hannan-Quinn criter.	4.958780
F-statistic	29.77331	Durbin-Watson stat	1.850141
Prob(F-statistic)	0.000000		

Estimate Equation:

mpg = 36.01 + (-1.11)(cyl) + (0.01)(disp) + (-0.02)(hp) + (0.95)(drat) + (-3.67)(wt)

Inferences:

- The model has very good explanatory power, having very high R^2 value of 0.98.
- The variables cyl, disp, hp, and drat are statistically insignificant, having p-values 0.13, 0.31, 0.08, and 0.50.
- The variable wt is statistically significant, having p-value 0.02.
- The intercept has a p-value of 0.0001, and is statistically significant.

From the inferences made using the high value of R^2, correlation matrix, and numerous statistically insignificant variables, we can confirm the presence of multicollinearity.