Multiple Linear Regression Analysis

mtcars

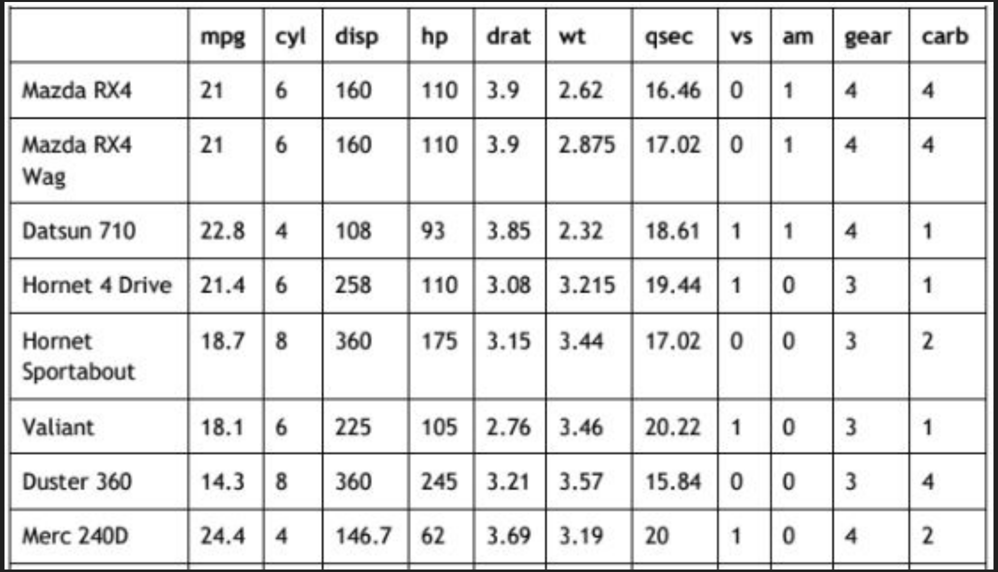
© Achmad Nafila Rozie

Outline

- Background and Objective
- Dataset
- Analysis Framework
- Findings
- Conclusion

Dataset

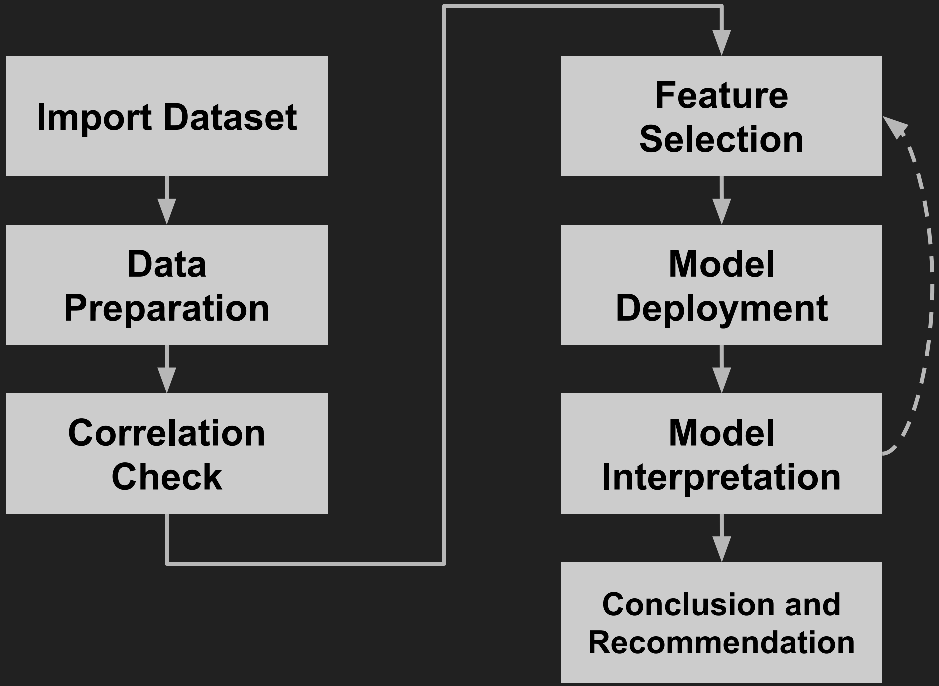
Given 10 variables to be checked further into their impact into the mpg variables



	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21	6	160	110	3.9	2.62	16.46	0	1	4	4
Mazda RX4 Wag	21	6	160	110	3.9	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.32	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.44	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.46	20.22	1	0	3	1
Duster 360	14.3	8	360	245	3.21	3.57	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.19	20	1	0	4	2

Sample of the dataset

Analysis Framework



Import Dataset

Data Preparation

Correlation Check

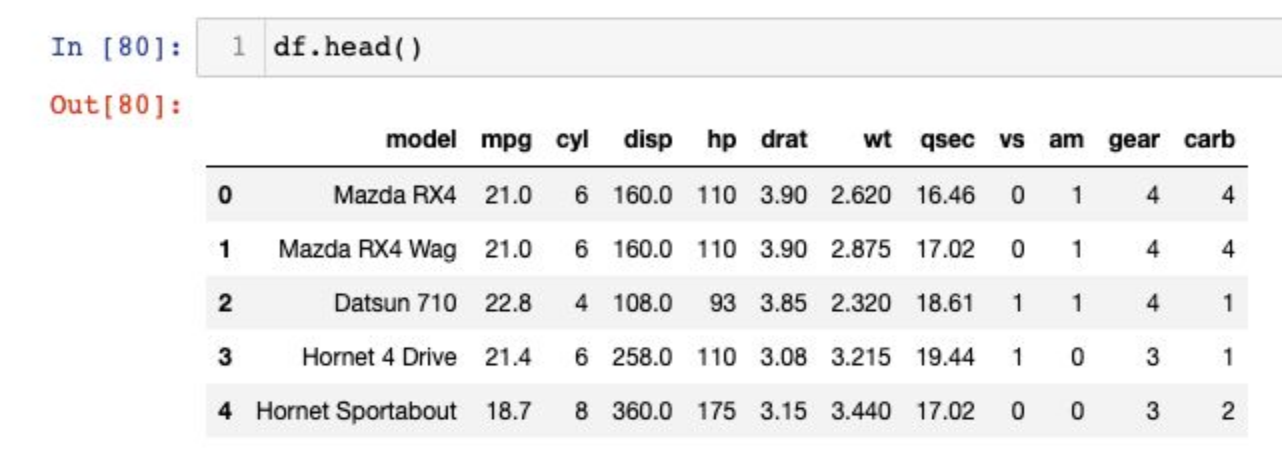
Feature Selection

Model Deployment

Model Interpretation

Conclusion and Recommendation

Findings - Import Dataset



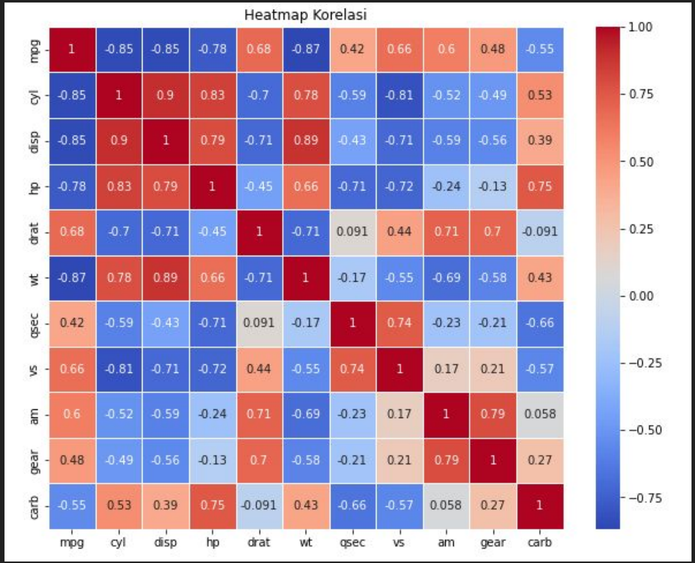
In [80]:

```
df.head()
```

Out[80]:

	model	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
0	Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
1	Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
2	Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
3	Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
4	Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2

We found lot of inter X (independent) is highly correlated



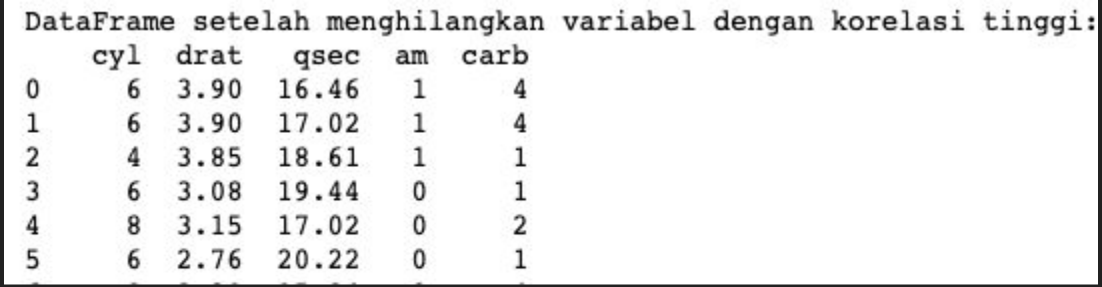
Heatmap Korelasi

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
mpg	1	-0.85	-0.85	-0.78	0.68	-0.87	0.42	0.66	0.6	0.48	-0.55
cyl	-0.85	1	0.9	0.83	-0.7	0.78	-0.59	-0.81	-0.52	-0.49	0.53
disp	-0.85	0.9	1	0.79	-0.71	0.89	-0.43	-0.71	-0.59	-0.56	0.39
hp	-0.78	0.83	0.79	1	-0.45	0.66	-0.71	-0.72	-0.24	-0.13	0.75
drat	0.68	-0.7	-0.71	-0.45	1	-0.71	0.091	0.44	0.71	0.7	-0.091
wt	-0.87	0.78	0.89	0.66	-0.71	1	-0.17	-0.55	-0.69	-0.58	0.43
qsec	0.42	-0.59	-0.43	-0.71	0.091	-0.17	1	0.74	-0.23	-0.21	-0.66
vs	0.66	-0.81	-0.71	-0.72	0.44	-0.55	0.74	1	0.17	0.21	-0.57
am	0.6	-0.52	-0.59	-0.24	0.71	-0.69	-0.23	0.17	1	0.79	0.058
gear	0.48	-0.49	-0.56	-0.13	0.7	-0.58	-0.21	0.21	0.79	1	0.27
carb	-0.55	0.53	0.39	0.75	-0.091	0.43	-0.66	-0.57	0.058	0.27	1

We will exclude certain independent variables that are highly correlated with one another to mitigate the risk of multicollinearity in the analysis

Model Deployment and Interpretation

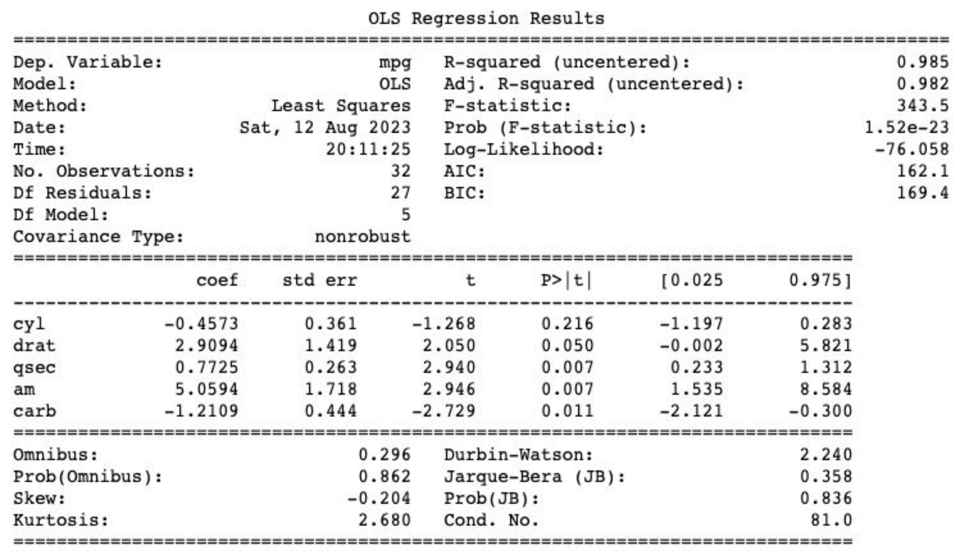
New Dataset after we eliminate 5 variables with correlation threshold = 0.7



```
DataFrame setelah menghilangkan variabel dengan korelasi tinggi:
   cyl  drat   qsec  am  carb
0    6  3.90  16.46   1     4
1    6  3.90  17.02   1     4
2    4  3.85  18.61   1     1
3    6  3.08  19.44   0     1
4    8  3.15  17.02   0     2
5    6  2.76  20.22   0     1
```

We found that there are 5 variables that has low correlation (correlation inter-X variables) <= 0.7

We generate quite “good” identifier with R-Squared 0.985, but 1 variable is not significant toward dependant variable

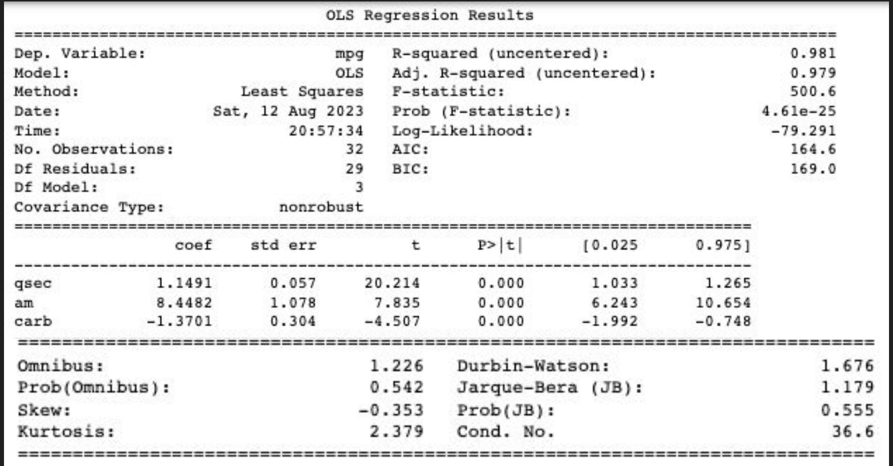


```
                            OLS Regression Results
=======================================================================================
Dep. Variable:                    mpg   R-squared (uncentered):                   0.985
Model:                            OLS   Adj. R-squared (uncentered):              0.982
Method:                 Least Squares   F-statistic:                              343.5
Date:                Sat, 12 Aug 2023   Prob (F-statistic):                    1.52e-23
Time:                        20:11:25   Log-Likelihood:                         -76.058
No. Observations:                  32   AIC:                                      162.1
Df Residuals:                      27   BIC:                                      169.4
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
cyl           -0.4573      0.361     -1.268      0.216      -1.197       0.283
drat           2.9094      1.419      2.050      0.050      -0.002       5.821
qsec           0.7725      0.263      2.940      0.007       0.233       1.312
am             5.0594      1.718      2.946      0.007       1.535       8.584
carb          -1.2109      0.444     -2.729      0.011      -2.121      -0.300
==============================================================================
Omnibus:                        0.296   Durbin-Watson:                   2.240
Prob(Omnibus):                  0.862   Jarque-Bera (JB):                0.358
Skew:                          -0.204   Prob(JB):                        0.836
Kurtosis:                       2.680   Cond. No.                         81.0
==============================================================================
```

Interpretation

- **R-squared = 0.985** means 98.5% of variance in the model is able to be explained by these 5 variables.
- **Coefficient Overall**
  We found that there is 1 column (cyl) that considered not significant - p-value >= 0.05 so we will iterate the model later.
- **Coefficient Interpretation**
  Each coef show impact value between X (independent) variable into Y (dependent) variable .
  e.g. if the car is automatic (am) → means the car has higher mpg for 5.0594

New Regression Results after ‘cyl’ column is eliminated



```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:                    mpg   R-squared (uncentered):                   0.981
Model:                            OLS   Adj. R-squared (uncentered):              0.979
Method:                 Least Squares   F-statistic:                              500.6
Date:                Sat, 12 Aug 2023   Prob (F-statistic):                    4.61e-25
Time:                        20:57:34   Log-Likelihood:                         -79.291
No. Observations:                  32   AIC:                                      164.6
Df Residuals:                      29   BIC:                                      169.0
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
qsec           1.1491      0.057     20.214      0.000       1.033       1.265
am             8.4482      1.078      7.835      0.000       6.243      10.654
carb          -1.3701      0.304     -4.507      0.000      -1.992      -0.748
==============================================================================
Omnibus:                        1.226   Durbin-Watson:                   1.676
Prob(Omnibus):                  0.542   Jarque-Bera (JB):                1.179
Skew:                          -0.353   Prob(JB):                        0.555
Kurtosis:                       2.379   Cond. No.                         36.6
==============================================================================
```

Interpretation

- **R-squared = 0.981** means 98.1% of variance in the model is able to be explained by these 5 variables.
- **Coefficient Overall** All of the correlation was significantly impact toward the Y variable
- **Coefficient Interpretation**
  - 'qsec': A coefficient of 1.15 with a p-value of 0.000 indicates a positive effect on 'mpg' and is statistically significant.
  - 'am' (automatic/manual transmission): A coefficient of 8.45 with a p-value of 0.000 indicates a significant positive effect on 'mpg.'
  - 'carb' (carburetor): A coefficient of -1.37 with a p-value of 0.000 indicates a significant negative effect on 'mpg.'
- **Durbin-Watson**: A Durbin-Watson value of 1.67 indicates that there is no significant autocorrelation within the model.

Conclusion and Recommendation

Conclusion

- We could identify the mpg based on 3 independent variables: qsec, am, and carb.
- The equation of the last multicollinear regression is as follow:

$$mpg = 1.15 * (qsec) + 8.85 * (am) - 1.37 * (carb) + e$$

- am variable become the most significant variable into the increasing of mpg

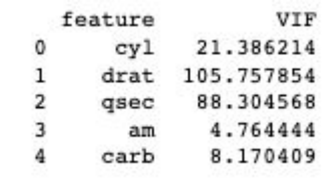
Alternative Recommendation(s)

- For users that likely to look **more 'economic'** which mean high mpg car - kindly to look for high qsec, automatic, and low number of carburetor.
- Oppositely, kindly to look for low qsec, manual, and high number of carburetor

End

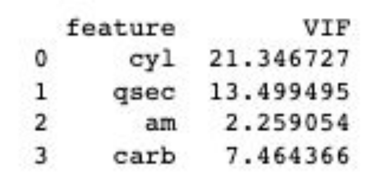
Feature Reduction to eliminate multicollinearity using VIF
(Variance Inflation Factor)

1st FIV



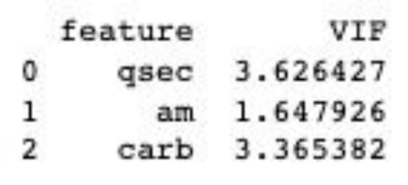
	feature	VIF
0	cyl	21.386214
1	drat	105.757854
2	qsec	88.304568
3	am	4.764444
4	carb	8.170409

2nd FIV



	feature	VIF
0	cyl	21.346727
1	qsec	13.499495
2	am	2.259054
3	carb	7.464366

3rd FIV



	feature	VIF
0	qsec	3.626427
1	am	1.647926
2	carb	3.365382