Revving Up Your Predictions: A Multivariate Linear Regression Analysis of Auto Data

A Punchline report

The Issues:

The dataset contains information on four predictor variables: displacement, horsepower, weight, and acceleration, as well as one predicted (= response) variable: mpg. The aim of the project is to use multivariate linear regression to answer questions related to the usefulness of the predictors in predicting the response, the explanatory power of the predictors, and the accuracy of the predictions made by the model.

We address the questions:

- Determine if at least one predictor is significant in predicting the response variable.
- Identify whether all predictors contribute to explaining the response variable, or if only a subset of the predictors are useful.
- Evaluate the goodness of fit of the model to the data and calculate the predicted response value for a given set of predictor values, and assess the accuracy of the prediction.

Findings:

As part of EDA, calculated the correlation matrix of the dataset and plotted a heatmap, it showed some interesting correlation between the target and predictor variables weight, displacement. On further analysis confirmed the significance of these variables in predicting the dependent variable.

The OLS regression analysis reports an R-squared value of 0.739, indicating that approximately 74% of the variation in the response variable (mpg) is explained by the predictors (displacement, horsepower, weight, and acceleration). The coefficients for displacement, horsepower, and weight are all negative, indicating that

an increase in these predictors is associated with a decrease in mpg. The coefficient for acceleration is also negative, but not statistically significant at the 5% level.
The F-statistic is significant at the 5% level, indicating that at least one of the predictors is useful in predicting the response. Additionally, the t-tests for displacement and weight are significant at the 5% level, indicating that these predictors are useful in explaining the response. The t-test for horsepower is not significant at the 5% level, suggesting that it may not be a useful predictor. The standard errors for the coefficients indicate the precision of the estimates. The 95% confidence intervals for the coefficients are reported in the last column, indicating the range of values in which the true coefficient is likely to lie with 95% confidence. Overall, the model appears to fit the data relatively well, with a significant F-statistic and relatively high R-squared value.

Discussion:

The findings suggest that the full model, which included all four predictor variables (displacement, horsepower, weight, and acceleration), explained a significant portion of the variance in the response variable (mpg), with an R-squared value of 0.739. This means that about 74% of the variation in mpg can be explained by the combination of these four predictors.
However, upon examining the t-tests of individual predictor variables, we can see that only two of the four variables are statistically significant in predicting mpg - displacement and weight. Both of these variables have negative coefficients, indicating that as they increase, mpg tends to decrease. The adjusted R-squared value for this model is only slightly lower than the full model, at 0.735.
This suggests that, while the full model may have included some predictors that were not useful in predicting mpg, a simpler model with just displacement and weight can explain almost as much of the variation in mpg with greater statistical significance. The F-statistic for the reduced model is also much higher than the full model, indicating a better overall fit.
In practical terms, this could mean that car manufacturers and policymakers looking to improve fuel efficiency and reduce emissions may be able to focus more specifically on engine displacement and weight as factors to optimize, rather than also considering horsepower and acceleration, which

may not be as impactful in predicting fuel efficiency.

Appendix A: Methods

Data collection: The data set is a subset of the Auto data set, and the data was collected by the U.S. Environmental Protection Agency (EPA) in the 1970s.

Variable creation: The data set contains 4 predictor variables and one response variable. The predictor variables are displacement (engine displacement in cubic inches), horsepower (engine horsepower), weight (vehicle weight in pounds), and acceleration (time to accelerate from 0 to 60 miles per hour in seconds). The response variable is mpg (miles per gallon of the vehicle). The data set was collected to study the relationship between the predictor variables and the response variable, with the goal of building a predictive model for the fuel efficiency of a vehicle based on its specifications.

Analytic methods:

The dataset is loaded into the pandas dataframe and it has following variables: displacement, horsepower, weight, acceleration as predictors and mpg as response variable.

Performed exploratory data analysis to better understand the distribution and relationships between the variables, created scatter plots and histograms to visualize the relationships between the variables and used summary statistics to describe the central tendency, variability, and shape of the distributions. The correlation between predictors and the response variable is computed to analyse the linear relationship and heatmap representing the data is included in the report.

A multivariate linear regression model with all four predictor variables to predict the response variable, mpg and examined the model summary to evaluate the significance of the predictor variables and the overall fit of the model. Also calculated the R-squared value to determine the proportion of variance in the response variable that is explained by the predictor variables.

Then repeated the analysis, but this time only used two predictor variables, displacement and weight, to fit a new multivariate linear

regression model and evaluated the significance of the predictor variables and the overall fit of the model using the same methods as before and compared the results to those of the previous model. Finally, performed a train-test split to evaluate the performance of the model on new, unseen data and trained the model on a subset of the data and tested it on the remaining data, calculating the root mean squared error to measure the accuracy of the model's predictions on the test data.

Appendix B: Results

The dataset contains 380 records and 5 columns. And the descriptive stats are as follows.

	displacement	horsepower	weight	acceleration	mpg
count	379.000000	379.000000	379.000000	379.000000	379.000000
mean	196.970976	104.897098	3008.567282	15.734037	22.770449
std	104.961018	39.499414	839.315990	2.770141	7.306728
min	68.000000	46.000000	1760.000000	9.000000	9.000000
25%	105.000000	75.000000	2260.000000	14.000000	16.500000
50%	156.000000	94.000000	2865.000000	15.600000	22.000000
75%	259.000000	120.000000	3632.500000	17.300000	28.000000
max	455.000000	225.000000	5140.000000	24.800000	46.600000

The kurtosis and skewness are calculated using standard dataframe methods and the values are as follows.

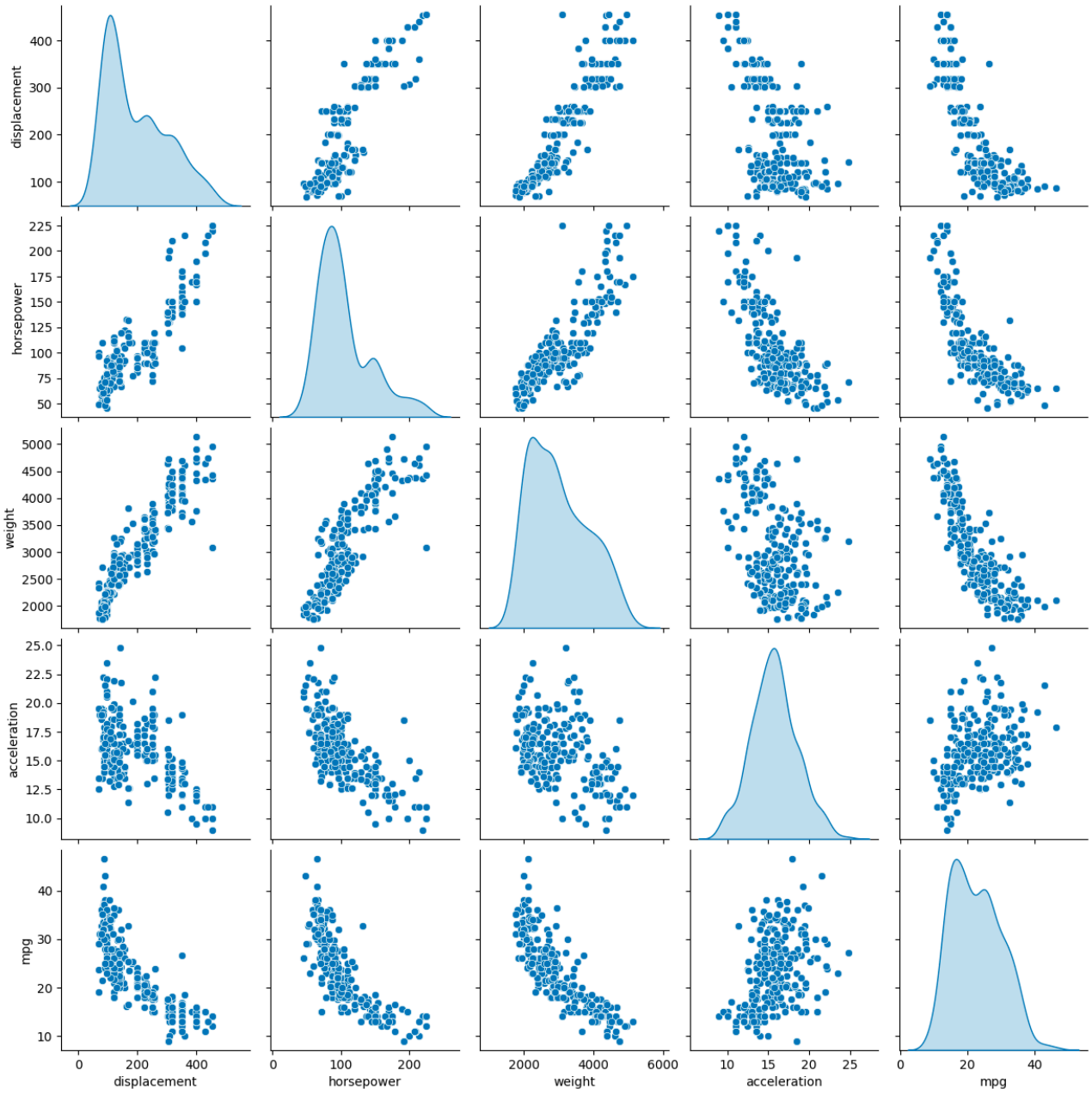
Skewness → displacement: 0.712113
Horsepower: 1.176262
Weight: 0.494648
Acceleration: 0.206091
Mpg: 0.395031

Kurtosis → displacement : -0.640071
Horsepower: 0.834916
Weight: -0.821477
Acceleration: 0.047970
Mpg: -0.551810

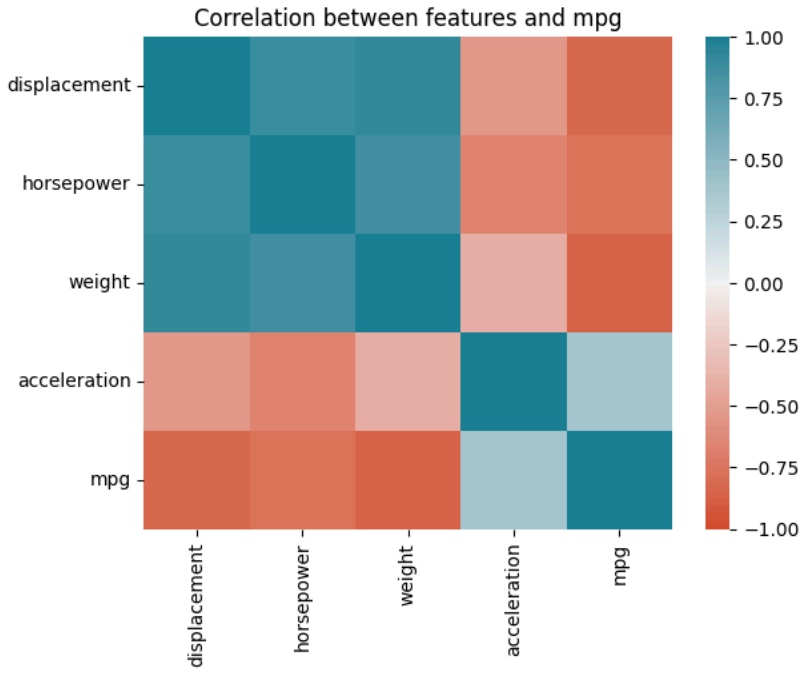
The skewness values for horsepower and weight indicate that their distributions are positively skewed, with longer tails on the right side of the distribution. Displacement and mpg have moderate positive skewness, while acceleration is only slightly positively skewed.

The kurtosis values for horsepower and weight indicate that their distributions are leptokurtic, meaning they have a sharp peak and heavy tails. Displacement and mpg have moderate platykurtic distributions, with flatter peaks and lighter tails. The kurtosis value for acceleration is close to zero, indicating a relatively normal distribution.

The scatterplots are created between the variables to analyse the distribution of variables and the scatter plots to visually analyse the relationship between variables.



The correlation between the variables in the dataset shows a linear relation for the variables displacement, weight with mpg. The resulting heatmap is shown below.



The OLS model is trained on whole dataset and the results are as follows.

OLS Regression Results

Dep. Variable:	mpg	R-squared:	0.739
Model:	OLS	Adj. R-squared:	0.736
Method:	Least Squares	F-statistic:	264.9
Date:	Mon, 20 Feb 2023	Prob (F-statistic):	1.06e-107
Time:	13:23:48	Log-Likelihood:	-1036.4
No. Observations:	379	AIC:	2083.
Df Residuals:	374	BIC:	2103.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	45.4041	2.090	21.720	0.000	41.294	49.515
displacement	-0.0155	0.006	-2.804	0.005	-0.026	-0.005
horsepower	-0.0251	0.014	-1.850	0.065	-0.052	0.002
weight	-0.0048	0.001	-6.997	0.000	-0.006	-0.003
acceleration	-0.1531	0.108	-1.423	0.156	-0.365	0.059

Omnibus:	47.140	Durbin-Watson:	2.141
Prob(Omnibus):	0.000	Jarque-Bera (JB):	71.149
Skew:	0.802	Prob(JB):	3.55e-16
Kurtosis:	4.389	Cond. No.	3.40e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.4e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

The statistical summary for the model trained only using displacement and weight is as follows:

OLS Regression Results

Dep. Variable:	mpg	R-squared:	0.737
Model:	OLS	Adj. R-squared:	0.735
Method:	Least Squares	F-statistic:	525.8
Date:	Mon, 20 Feb 2023	Prob (F-statistic):	1.18e-109
Time:	13:23:48	Log-Likelihood:	-1038.2
No. Observations:	379	AIC:	2082.
Df Residuals:	376	BIC:	2094.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	42.5726	1.015	41.954	0.000	40.577	44.568
displacement	-0.0169	0.005	-3.552	0.000	-0.026	-0.008
weight	-0.0055	0.001	-9.178	0.000	-0.007	-0.004

Omnibus:	42.976	Durbin-Watson:	2.147
Prob(Omnibus):	0.000	Jarque-Bera (JB):	62.383
Skew:	0.760	Prob(JB):	2.84e-14
Kurtosis:	4.281	Cond. No.	1.64e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.64e+04. This might indicate that there are strong multicollinearity or other numerical problems.

Finally, trained the dataset is divided into train and test data then calculated the same metrics and RMSE and R-squared are reported as

RMSE: 3.80824

r-square: 0.67922

Appendix C: Code

The statistical analysis is performed using following code and a linear model is trained using sklearn package of python.

a. Importing required libraries and loading the dataset into pandas dataframe

```
df = pd.read_excel("./Datasets/auto_data_namala_durga.xls")
df.head()
```

✓ 0.0s

b. Dataframe describe method is used to generate descriptive stats of all the numerical columns.

```
df.describe()
```
✓ 0.0s

	displacement	horsepower	weight	acceleration	mpg
count	379.000000	379.000000	379.000000	379.000000	379.000000
mean	196.970976	104.897098	3008.567282	15.734037	22.770449
std	104.961018	39.499414	839.315990	2.770141	7.306728
min	68.000000	46.000000	1760.000000	9.000000	9.000000
25%	105.000000	75.000000	2260.000000	14.000000	16.500000
50%	156.000000	94.000000	2865.000000	15.600000	22.000000
75%	259.000000	120.000000	3632.500000	17.300000	28.000000
max	455.000000	225.000000	5140.000000	24.800000	46.600000

c. To computed skewness & kurtosis pandas offers skew, kurtosis methods.

```
df.skew()
```
✓ 0.0s

```
displacement    0.712113
horsepower      1.176262
weight          0.494648
acceleration    0.206091
mpg             0.395031
dtype: float64
```

```
df.kurtosis()
```
✓ 0.0s

```
displacement   -0.640071
horsepower      0.834916
weight         -0.821477
acceleration    0.047970
mpg            -0.551810
dtype: float64
```

d. Heat map is plotted using the following code.

```
ax = sns.heatmap(df.corr(), vmin=-1, vmax=1, cmap=sns.diverging_palette(20, 220, as_cmap=True))
ax.set_title("Correlation between features and mpg")
```

✓ 0.1s

e. Kde plots and scatter plots for each variables are computed using below code.

```
sns.pairplot(df, diag_kind='kde')
```

✓ 2.3s

f. Finally, linear model is trained using following code and summary method gives the OLS regression results.

```
import statsmodels.api as sm

X, y = df.iloc[:, :-1], df.iloc[:, -1]
X = sm.add_constant(X)
lm = sm.OLS(y, X).fit()
```

✓ 2.8s

```
lm.summary()
```

✓ 0.0s

g. Additional code for other analysis is shown below.

- For train, test dataset splitting

```
from sklearn.model_selection import train_test_split

train_x, test_x, train_y, test_y = train_test_split(df.iloc[:,:-1], df.iloc[:, -1],test_size=0.2,random_state=23)
```

✓ 0.0s

```
train_x_const = sm.add_constant(train_x)
lm = sm.OLS(train_y, train_x_const).fit()
lm.summary()
```

- For computing RMSE & R-square

```
pred_y = lm.predict(sm.add_constant(test_x))
print(f'RMSE: {np.sqrt(((pred_y - test_y) ** 2).mean()):.5f}')
```

```
from sklearn.metrics import r2_score

print(f"r-square: {r2_score(test_y, pred_y): .5f}")
```