LINEAR REGRESSION AND TIME SERIES

MSDA3055-03-F23

DATASET-CO2 EMISSIONS

BY: AKANSHA PRUTHI

TRUPTI SUDHIR

&

PRASANTH REDDY KODI

ANALYZING CO2 EMISSIONS AND THEIR ENVIRONENTAL IMPACT

Introduction

Objective & Research Questions:

Our objective was to comprehensively analyze CO2 emissions data to understand their environmental impacts and explore strategies for emission reduction.

The research questions are:

1. What factors contribute to CO2 emissions and how can they be mitigated?

2. What are the current trends in CO2 emissions and their associated environmental impacts?

Motivation:

The project aimed to address concerns about climate change, global warming, unusual weather patterns, and ecosystem disruptions by identifying emission patterns and mitigation strategies.

Null Hypothesis (H0): There is no significant relationship between the predictor variables (such as engine size, cylinders, fuel consumption, etc.) and CO2 emissions in the population.

Alternative Hypothesis (H1): There is a significant relationship between the predictor variables and CO2 emissions in the population.

Data Description

The dataset sourced from Kaggle includes variables such as vehicle characteristics, fuel consumption, and CO2 emissions. It comprises both quantitative (e.g., Engine Size, Fuel Consumption) and qualitative (e.g., Make, Model) variables. This is the preview of the data.

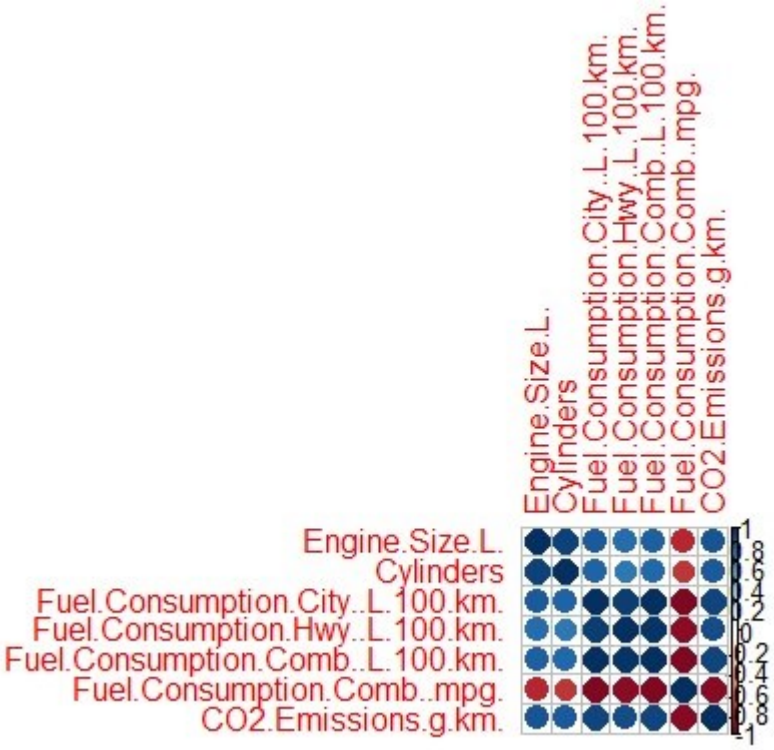
Make	Model	Vehicle Cl	Engine Siz	Cylinders	Transmiss	Fuel Type	Fuel Cons	Fuel Cons	Fuel Cons	Fuel Cons	CO2 Emissions(g/km)
ACURA	ILX	COMPACT	2	4	AS5	Z	9.9	6.7	8.5	33	196
ACURA	ILX	COMPACT	2.4	4	M6	Z	11.2	7.7	9.6	29	221
ACURA	ILX HYBRI	COMPACT	1.5	4	AV7	Z	6	5.8	5.9	48	136
ACURA	MDX 4WD	SUV - SMA	3.5	6	AS6	Z	12.7	9.1	11.1	25	255
ACURA	RDX AWD	SUV - SMA	3.5	6	AS6	Z	12.1	8.7	10.6	27	244
ACURA	RLX	MID-SIZE	3.5	6	AS6	Z	11.9	7.7	10	28	230
ACURA	TL	MID-SIZE	3.5	6	AS6	Z	11.8	8.1	10.1	28	232
ACURA	TL AWD	MID-SIZE	3.7	6	AS6	Z	12.8	9	11.1	25	255
ACURA	TL AWD	MID-SIZE	3.7	6	M6	Z	13.4	9.5	11.6	24	267
ACURA	TSX	COMPACT	2.4	4	AS5	Z	10.6	7.5	9.2	31	212
ACURA	TSX	COMPACT	2.4	4	M6	Z	11.2	8.1	9.8	29	225
ACURA	TSX	COMPACT	3.5	6	AS5	Z	12.1	8.3	10.4	27	239
ALFA ROM	4C	TWO-SEA	1.8	4	AM6	Z	9.7	6.9	8.4	34	193
ASTON M	DB9	MINICOM	5.9	12	A6	Z	18	12.6	15.6	18	359
ASTON M	RAPIDE	SUBCOMP	5.9	12	A6	Z	18	12.6	15.6	18	359
ASTON M	V8 VANTA	TWO-SEA	4.7	8	AM7	Z	17.4	11.3	14.7	19	338
ASTON M	V8 VANTA	TWO-SEA	4.7	8	M6	Z	18.1	12.2	15.4	18	354
ASTON M	V8 VANTA	TWO-SEA	4.7	8	AM7	Z	17.4	11.3	14.7	19	338
ASTON M	V8 VANTA	TWO-SEA	4.7	8	M6	Z	18.1	12.2	15.4	18	354
ASTON M	VANQUIS	MINICOM	5.9	12	A6	Z	18	12.6	15.6	18	359
AUDI	A4	COMPACT	2	4	AV8	Z	9.9	7.4	8.8	32	202
AUDI	A4 QUATT	COMPACT	2	4	AS8	Z	11.5	8.1	10	28	230
AUDI	A4 QUATT	COMPACT	2	4	M6	Z	10.8	7.5	9.3	30	214
AUDI	A5 CABRI	SUBCOMP	2	4	AS8	Z	11.5	8.1	10	28	230

Potential Models:

Statistical techniques like Regression or Time Series Analysis using R Studio were considered for the analysis.

Data Cleaning and Exploratory Data Analysis:

The data with us had no missing values, we also did a correlation matrix with numerical variables against Co2 and gained following insights:



- Fuel Consumption and CO2 Emissions: As expected, there is a strong positive correlation between these two variables (0.865). This means that as fuel consumption increases, CO2 emissions also increase.
- Engine Size and CO2 Emissions: There is also a positive correlation between Engine Size and CO2 Emissions (0.800). This makes sense, as larger engines typically burn more fuel and produce more emissions.
- Cylinders and CO2 Emissions: The correlation between Cylinders and CO2 Emissions is slightly weaker (0.704) than the correlation between Engine Size and CO2 Emissions. This suggests that the number of cylinders is not quite as important a factor in determining CO2 emissions as engine size.
- Fuel Consumption and Cylinders: There is a moderate negative correlation between Fuel Consumption and Cylinders (-0.534). This means that cars with fewer cylinders tend to have better fuel economy.
- City Fuel Consumption and Highway Fuel Consumption: There is a strong positive correlation between City Fuel Consumption and Highway Fuel Consumption (0.900). This is to be expected, as cars that are fuel-efficient in the city are likely to be fuel-efficient on the highway as well.

Regression Analysis

Simple Linear Regression:

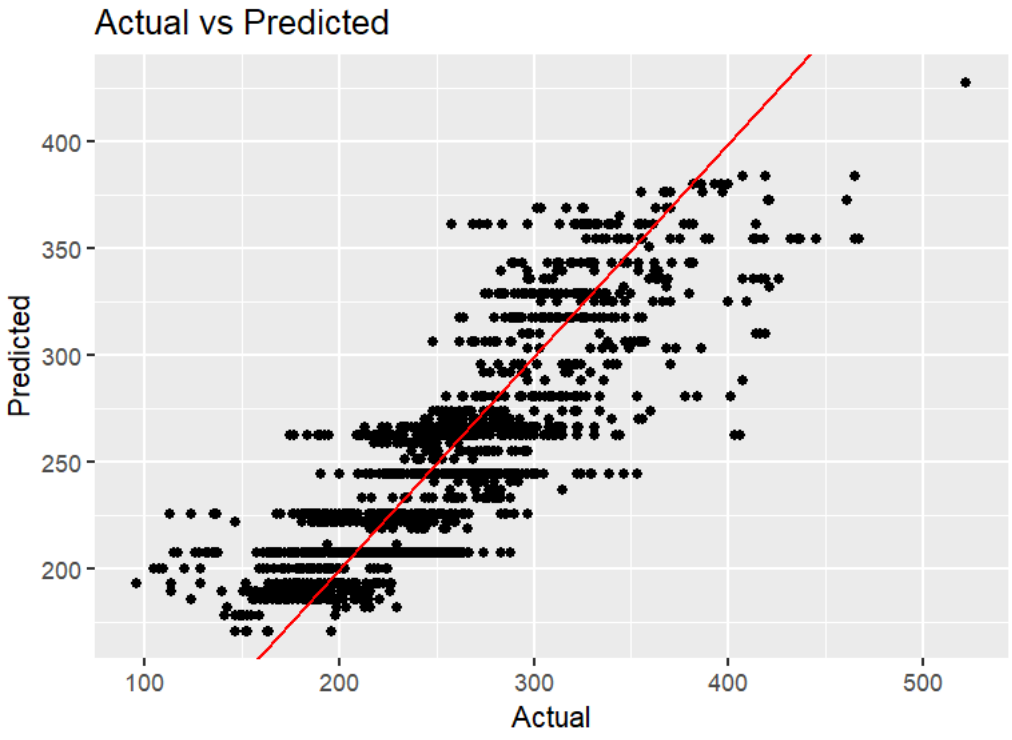
To begin with, we did a Simple Linear Regression model with only engine Size, as it exhibited the most correlation. An additional graph also shows Actual vs Predicted values.

```
Call:
lm(formula = CO2.Emissions.g.km. ~ Engine.Size.L., data = train_df)

Residuals:
     Min       1Q   Median       3Q      Max
-112.797  -18.397   -1.133   18.803  139.476

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     250.0393     0.4272   585.3   <2e-16 ***
Engine.Size.L.   49.7352     0.4274   116.4   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 30.72 on 5168 degrees of freedom
Multiple R-squared:  0.7238,    Adjusted R-squared:  0.7237
F-statistic: 1.354e+04 on 1 and 5168 DF,  p-value: < 2.2e-16
```



Dummy Variables and log transformation:

We created dummy variables for categorical variables, except for “Make” and “Model”. Some variables exhibited positive skewness, suggesting potential transformation needs for symmetrical distributions.

```
                Vehicle.Class                  Engine.Size.L.                       Cylinders
                    3.6147386                      0.8088523                       1.1099644
                 Transmission                      Fuel.Type Fuel.Consumption.City..L.100.km.
                    4.9028665                      1.4999391                       0.8086761
Fuel.Consumption.Hwy..L.100.km. Fuel.Consumption.Comb..L.100.km.      Fuel.Consumption.Comb..mpg.
                    1.0787783                      0.8929529                       0.9766372
          CO2.Emissions.g.km.
                    0.5258801
```

Considering these values, applying a log transformation might help make their distributions more symmetrical, which can improve the linearity assumption in the regression model.

Train Test Split:

We now split our data into Train Data and Test Data with Co2 as the target variable.

Multiple Linear Regression:

Upon fitting the multiple linear regression, we get the following model and a plot for Actual vs Fitted values.

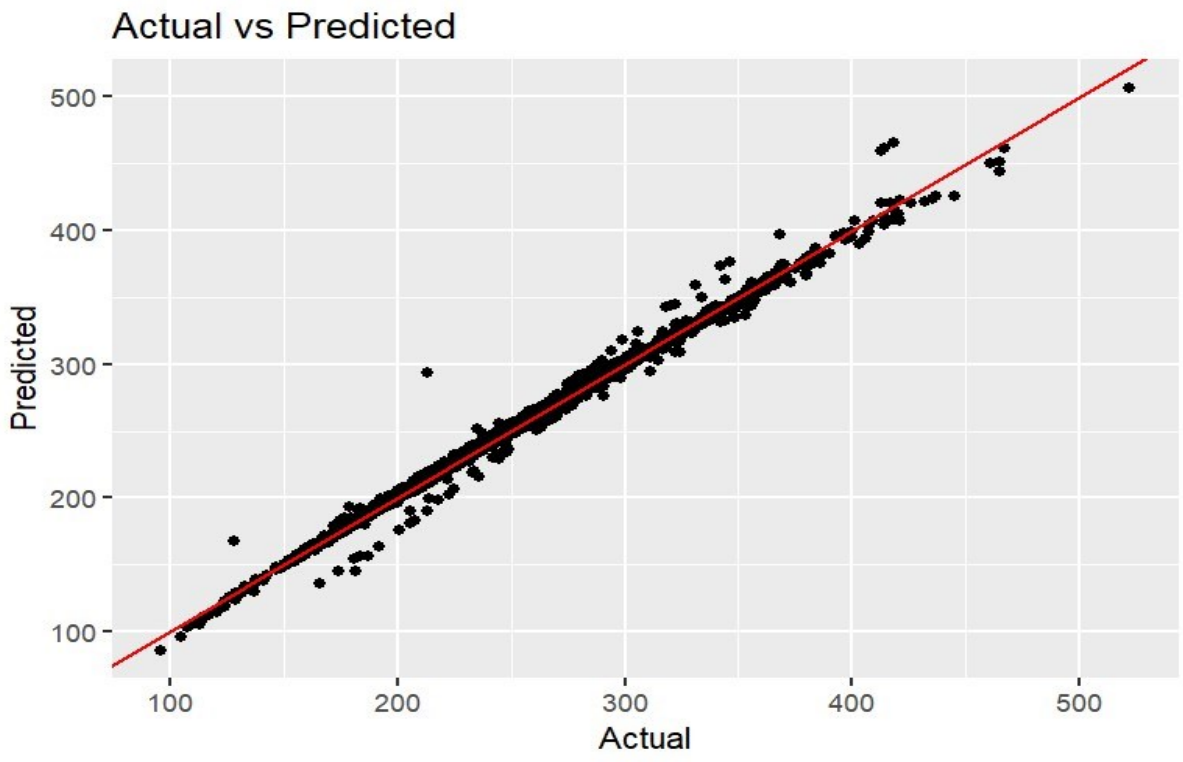
```
Call:
lm(formula = CO2.Emissions.g.km. ~ ., data = train_df)

Residuals:
    Min      1Q  Median      3Q     Max
-47.373  -2.271  -0.015   1.714  38.860

Coefficients: (4 not defined because of singularities)
                                      Estimate Std. Error  t value Pr(>|t|)
(Intercept)                           40.30646    2.27286   17.734  < 2e-16 ***
Vehicle.ClassxCOMPACT                  4.25380    0.90450    4.703 2.63e-06 ***
Vehicle.ClassxFULL-SIZE                4.92727    0.90775    5.428 5.96e-08 ***
Vehicle.ClassxMID-SIZE                 4.21859    0.89920    4.691 2.78e-06 ***
Vehicle.ClassxMINICOMPACT              3.27687    0.97302    3.368 0.000764 ***
Vehicle.ClassxMINIVAN                  4.26823    1.03731    4.115 3.94e-05 ***
Vehicle.ClassxPICKUP TRUCK - SMALL     7.08727    0.92421    7.668 2.07e-14 ***
Vehicle.ClassxPICKUP TRUCK - STANDARD  5.11253    0.84206    6.071 1.36e-09 ***
Vehicle.ClassxSPECIAL PURPOSE VEHICLE  6.65330    1.04150    6.388 1.83e-10 ***
Vehicle.ClassxSTATION WAGON - MID-SIZE 4.22981    1.16031    3.645 0.000270 ***
Vehicle.ClassxSTATION WAGON - SMALL    3.80892    0.94501    4.031 5.65e-05 ***
Vehicle.ClassxSUBCOMPACT               3.92938    0.91664    4.287 1.85e-05 ***
Vehicle.ClassxSUV - SMALL              4.39183    0.86183    5.096 3.60e-07 ***
Vehicle.ClassxSUV - STANDARD           4.94100    0.83319    5.930 3.22e-09 ***
Vehicle.ClassxTWO-SEATER               4.78610    0.91648    5.222 1.84e-07 ***
Vehicle.ClassxVAN - CARGO             -2.29441    1.53749   -1.492 0.135678
Vehicle.ClassxVAN - PASSENGER               NA         NA       NA       NA
Engine.Size.L.                         0.13315    0.15634    0.852 0.394446
Cylinders                              0.92983    0.11372    8.176 3.66e-16 ***
TransmissionxA10                       1.66499    1.35810    1.226 0.220266
TransmissionxA4                       -5.32458    1.08922   -4.888 1.05e-06 ***
TransmissionxA5                       -1.07144    0.89389   -1.199 0.230731
TransmissionxA6                       -1.06501    0.68991   -1.544 0.122726
TransmissionxA7                        1.74448    1.04149    1.675 0.093996 .
TransmissionxA8                        0.57779    0.70070    0.825 0.409639
TransmissionxA9                        0.97037    0.72672    1.335 0.181843
```

```
TransmissionxAM5                  0.01206    2.78367    0.004 0.996543
TransmissionxAM6                  2.23414    0.84233    2.652 0.008018 **
TransmissionxAM7                  0.68614    0.66718    1.028 0.303802
TransmissionxAM8                 -0.19267    0.94106   -0.205 0.837782
TransmissionxAM9                  4.03803    3.38586    1.193 0.233075
TransmissionxAS10                 1.06043    0.80815    1.312 0.189523
TransmissionxAS4                 -1.89047    4.71109   -0.401 0.688230
TransmissionxAS5                 -1.36157    1.29478   -1.052 0.293042
TransmissionxAS6                 -0.05545    0.67536   -0.082 0.934563
TransmissionxAS7                 -0.65896    0.72501   -0.909 0.363446
TransmissionxAS8                  0.19039    0.67383    0.283 0.777533
TransmissionxAS9                  0.36108    0.90774    0.398 0.690813
TransmissionxAV                   0.77439    0.76065    1.018 0.308696
TransmissionxAV10                -0.65094    1.88487   -0.345 0.729847
TransmissionxAV6                 -1.29536    0.84833   -1.527 0.126833
TransmissionxAV7                  0.08526    0.87210    0.098 0.922123
TransmissionxAV8                 -1.10186    1.14687   -0.961 0.336721
TransmissionxM5                  -0.12831    0.76597   -0.168 0.866976
TransmissionxM6                  -0.14898    0.67515   -0.221 0.825359
TransmissionxM7                        NA         NA       NA       NA
Fuel.TypexD                      30.23957    0.49495   61.096  < 2e-16 ***
Fuel.TypexE                    -107.89613    0.46597 -231.551  < 2e-16 ***
Fuel.TypexN                            NA         NA       NA       NA
Fuel.TypexX                       0.63312    0.20315    3.117 0.001840 **
Fuel.TypexZ                            NA         NA       NA       NA
Fuel.Consumption.City..L.100.km. 23.42841    2.97105    7.886 3.80e-15 ***
Fuel.Consumption.Hwy..L.100.km.  12.24839    1.56648    7.819 6.42e-15 ***
Fuel.Consumption.Comb..L.100.km. 23.90900    4.44805    5.375 7.99e-08 ***
Fuel.Consumption.Comb..mpg.      -5.24672    0.22685  -23.129  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.649 on 5119 degrees of freedom
Multiple R-squared:  0.9937,    Adjusted R-squared:  0.9937
F-statistic: 1.624e+04 on 50 and 5119 DF,  p-value: < 2.2e-16
```



In addition to the above model, we also tested on Stepwise Model, for whose we go the following scores.

	Model	RMSE
1	Simple Linear	30.732020
2	Multiple Linear	5.423867
3	Stepwise Forward	5.423867
4	Stepwise Backward	5.424548
5	Stepwise Both	5.424548

Based on above results, the “Multiple Linear Regression model” emerged as the final choice due to its superior performance (lower RMSE) compared to Simple Linear Regression. The MLR's ability to include multiple predictors provided a more comprehensive understanding of CO2 emissions' influencing factors.

- The coefficients revealed interesting findings:
- Engine size had a significant positive impact on CO2 emissions.
- Fuel consumption variables showed strong positive correlations with CO2 emissions.

NOTE:

We did not exclude “Fuel Consumption Comb (L/100 km)” and “Fuel Consumption Comb (mpg)” because after trying that the RMSE of the new model performed worse. Including all these variables allows the model to capture potential nuances and variations in CO2 emissions based on different types of fuel consumption metrics. While they share the theme of fuel consumption, their individual contributions might vary due to their specific contexts or units of measurement.

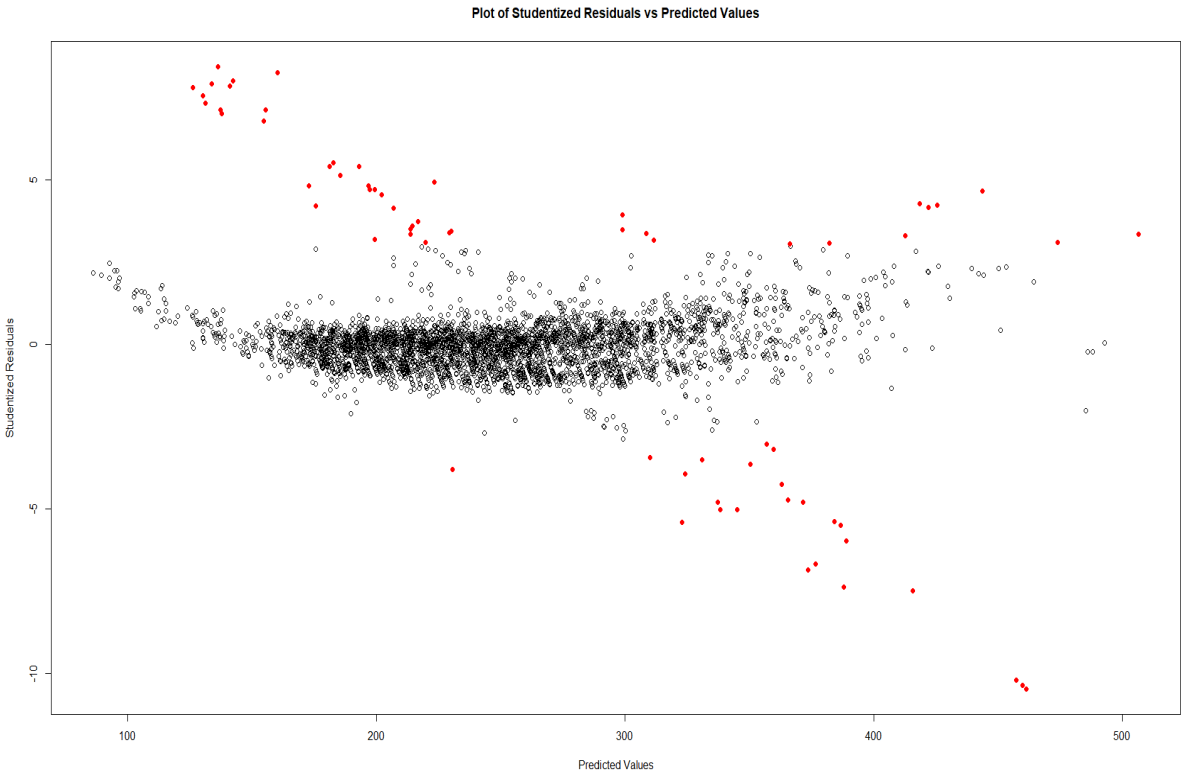
	Model	RMSE
1	Simple Linear	30.732020
2	Multiple Linear	5.761963
3	Stepwise Forward	5.761963
4	Stepwise Backward	5.762563
5	Stepwise Both	5.762563

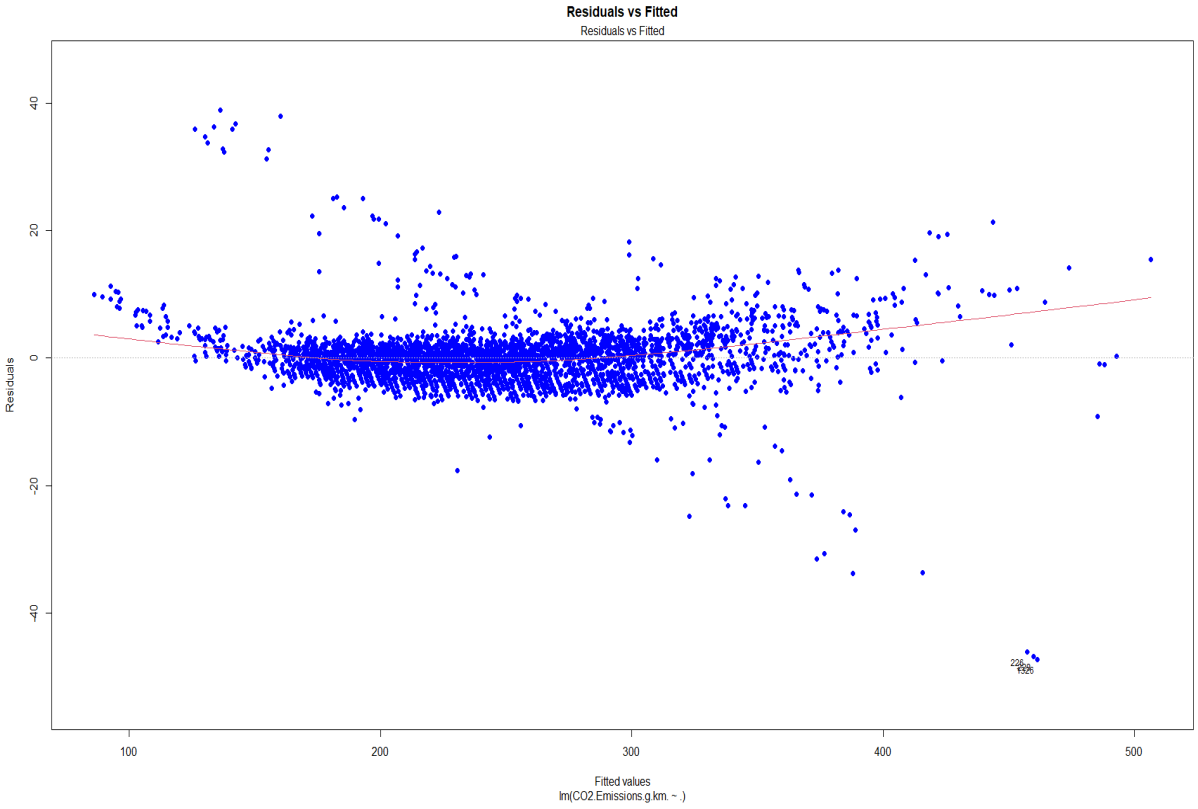
In addition to this, we also tried other models like Ridge-Lasso, Elastic Net Models, but were unable to get the desired results.

```
> # Summary of Elastic Net model
> summary(elastic_net_model)
           Length Class      Mode
a0           74   -none-     numeric
beta       3996   dgCMatrix  S4
df           74   -none-     numeric
dim           2   -none-     numeric
lambda       74   -none-     numeric
dev.ratio    74   -none-     numeric
nulldev       1   -none-     numeric
npasses       1   -none-     numeric
jerr          1   -none-     numeric
offset        1   -none-     logical
call          4   -none-     call
nobs          1   -none-     numeric
```

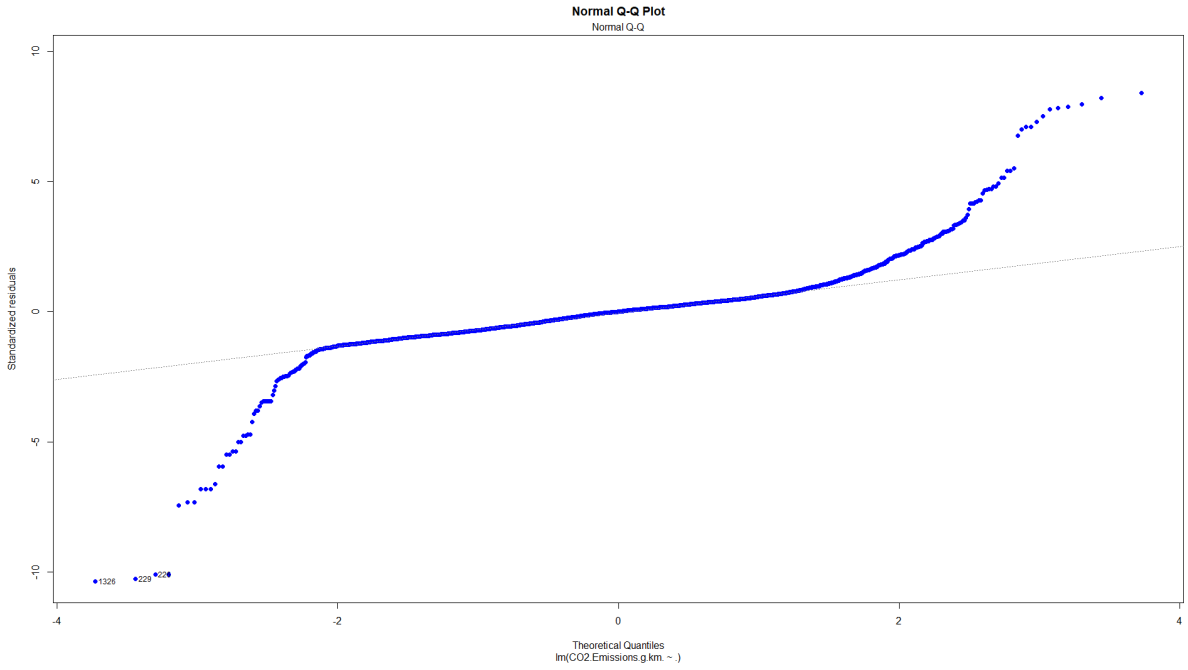
Detecting Outliers:

We plotted studentized residuals vs Predicted values and set a threshold of 3. This shows us the potential outliers. Additionally, Residuals vs Fitted values shows the same outliers.

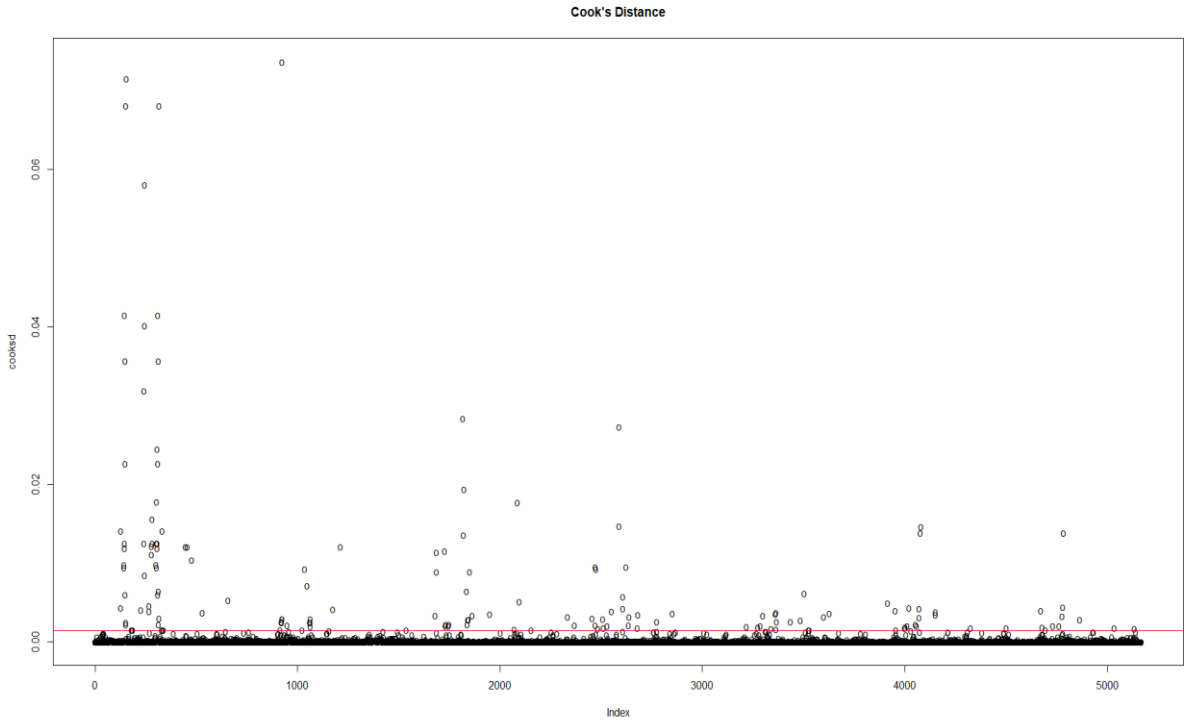




- A Normal Q-Q plot shows the presence of outliers.



- Cooks distance plot shows the presence of outliers along with a visible pattern.



We will further discuss the implication of presence of data point outliers in our dataset. This may affect our model in several ways like

- Inaccuracy of model predictions
- Violations of model assumptions
- Decrease in model representation of data points

ANOVA TABLE:

Analysis of Variance Table

Response: CO2.Emissions.g.km.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Vehicle.Class	15	6044200	402947	18645.805	< 2.2e-16	***
Engine.Size.L.	1	7695053	7695053	356078.043	< 2.2e-16	***
Cylinders	1	459787	459787	21276.003	< 2.2e-16	***
Transmission	26	624307	24012	1111.113	< 2.2e-16	***
Fuel.Type	3	226774	75591	3497.892	< 2.2e-16	***
Fuel.Consumption.City..L.100.km.	1	2375815	2375815	109937.582	< 2.2e-16	***
Fuel.Consumption.Hwy..L.100.km.	1	104828	104828	4850.764	< 2.2e-16	***
Fuel.Consumption.Comb..L.100.km.	1	1086	1086	50.237	1.547e-12	***
Fuel.Consumption.Comb..mpg.	1	11560	11560	534.946	< 2.2e-16	***
Residuals	5119	110625	22			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

>

In this ANOVA table:

- All predictor variables (Vehicle.Class, Engine.Size.L., Cylinders, Transmission, Fuel.Type, Fuel.Consumption.City..L.100.km., Fuel.Consumption.Hwy..L.100.km., Fuel.Consumption.Comb..L.100.km., Fuel.Consumption.Comb..mpg.) have extremely small p-values (< 0.001). This suggests that all these variables are statistically significant in explaining the variance in CO2 emissions.
- The residuals represent unexplained variance that the model could not capture. The residual mean square is used to estimate the error variance of the model.
- The overall F-statistic for the model is extremely high, indicating that the set of variables together significantly contribute to explaining the variance in CO2 emissions.
- All the predictor variables in this ANOVA table have extremely small p-values (e.g., < 0.001), it suggests strong evidence against the null hypothesis for each of these variables. Hence, we have statistical grounds to reject the null hypothesis and infer that these variables do have a significant effect on predicting CO2 emissions in your model.
- The small p-values indicate that these variables are associated with changes in CO2 emissions, providing statistical support for their inclusion in the model.

Discussions and limitations

Verdict:

Based on all the evidence from the above statistical tests, **WE REJECT THE NULL HYPOTHESIS** and can say that there is a significant relationship between the predictor variables and CO2 emissions in the population.

Method overview and findings:

- The final model that we selected is a Multiple Linear Regression model.
- We selected the model based on the RMSE score and the R Squared value.
- The coefficients with higher absolute values tend to contribute more to the predicted outcome (CO2 emissions in this case). Looking at the coefficients with larger absolute values.
- Fuel.TypexE: With an estimate of -107.89613, this coefficient seems to have a substantial impact on reducing CO2 emissions. Vehicles with fuel type 'E' contribute significantly to lower CO2 emissions compared to others.
- Fuel.TypexD: Conversely, 'Fuel.TypexD' with an estimate of 30.23957 has a substantial positive impact on CO2 emissions. This fuel type ('D') appears to contribute more to emissions.

- Fuel.Consumption.City..L.100.km. and Fuel. Consumption. Hwy..L.100.km. These coefficients have relatively high values (23.42841 and 12.24839 respectively), indicating that higher fuel consumption in both city and highway conditions contributes notably to increased CO2 emissions.
- Cylinders: The 'Cylinders' coefficient with an estimate of 1.70001 suggests that as the number of cylinders increases, CO2 emissions tend to rise.
- These coefficients highlight the variables that contribute more significantly to predicting higher or lower CO2 emissions.

Limitations

- The dataset in our hands has not been treated with outliers and possibly explains the reason for higher R squared value.
- More details and insights can be gained with advanced Machine Learning Models like Random Forests or Neural Networks.

Conclusion:

- Our aim was to get an answer for the research question about Co2 emissions. The dataset in our hands had very few variables and we took that as a challenge and tried to fit the best model with the resources in our hand.
- We started by fitting a Simple Linear Regression model for Co2 against Engine Cylinders, as it was correlated the most.
- We then went on to create Dummy Variables and perform Log Transformation for the variables to standardize them.
- We created a Multiple Linear Regression model removing Car Make and Model categories.
- A model was fit and its RMSE score was compared with other models, and we found that MLR had the lowest score.
- We also plotted Actual vs Predicted values and found that most of the points followed the fitted line.
- From multiple graphs we found that there are outliers and may possibly cause increased R score.
- ANOVA table confirms that we can Reject Null Hypothesis statistically.

Recommendations:

The coefficients and their significance from our model indicate the impact of various vehicle characteristics on CO2 emissions. Based on these coefficients, here are some recommendations for reducing CO2 emissions:

Vehicle Class Impact: Prioritize smaller vehicle classes such as Compact, Subcompact, and Small SUVs. They tend to have lower CO2 emissions compared to larger classes like SUV Standard or Pickup Trucks.

Transmission Preferences: Consider vehicles with AM6 transmissions, as they seem to contribute less to CO2 emissions than some other transmission types.

Fuel Type Impact: Ethanol (E85) seems to contribute significantly less to CO2 emissions compared to other fuel types. Vehicles running on E85 could be recommended for reduced emissions.

Fuel Consumption Efficiency: Emphasize vehicles with better fuel efficiency in the city, highway, and combined driving. Lower fuel consumption per 100 km translates to reduced CO2 emissions.

Cylinder and Engine Size: Lower cylinder count and engine size tend to contribute less to emissions. This indicates that vehicles with smaller engines and fewer cylinders are likely to emit less CO2.

Wheelbase Consideration: Wheelbase (SWB, LWB, EWB) doesn't show a significant impact based on the coefficients. Hence, while considering emissions reduction, focus more on other factors like vehicle class and transmission.

4WD/4X4 and AWD Impact: The impact of 4WD/4X4 or AWD on CO2 emissions is not explicitly evident from the coefficients. The model suggests other factors might have a more substantial influence on emissions.