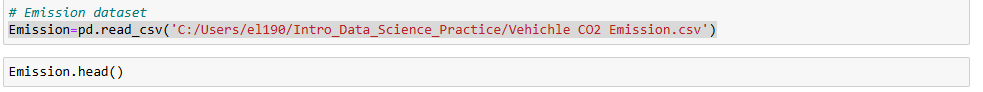
**Term Project Task 2**

The business problem I wanted to answer was to determine what vehicles or features produced the highest CO2 emissions or had the biggest impact on CO2 emissions. Answering this question will help car manufacturers design cars that are less harmful to the environment, and it will also provide compliance with government environmental regulations.

**Data Preprocessing**

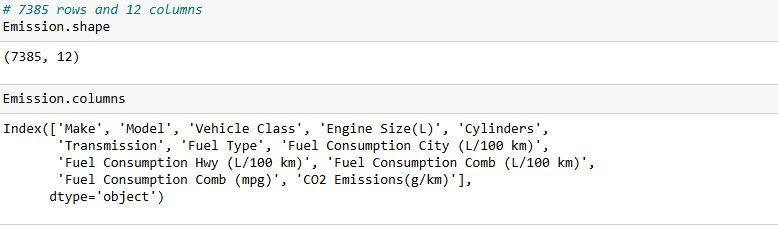
The dataset I will be using is data on vehicle CO2 emissions collected from Sahan, B. R. (n.d.). Vehicle CO2 Emissions Dataset. Kaggle: https://www.kaggle.com/datasets/brsahan/vehicle-co2-emissions-dataset/data. The data contains 12 features with 7385 observations. It contains both categorical and numerical values. The variables are brand, vehicle type, engine size, number of cylinders, transmission type, fuel type, city fuel consumption, highway fuel consumption, combined fuel consumption, combined fuel consumption per gallon, CO2 emissions in grams. To analyze the data, I will be using python in a Jupiter notebook. I collected the data and placed it in a panda’s data frame



A screenshot of a computer

AI-generated content may be incorrect.

I looked at the shape, columns, dtypes of the dataset. We have in the data set 7385 samples with 12 features that contained both categorical and numerical values.



A screenshot of a computer program

AI-generated content may be incorrect.

One of the first things I did was change fuel type codes into what the codes represented

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

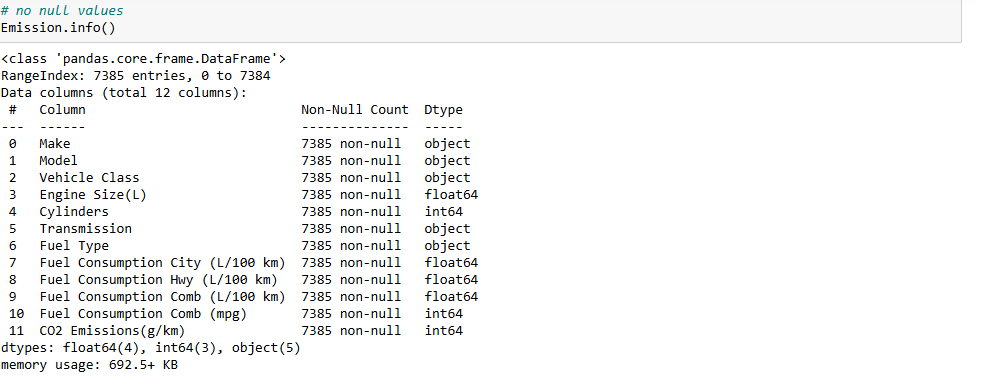
AI-generated content may be incorrect.

I also looked at the statistics of the data

A screenshot of a graph

AI-generated content may be incorrect.

I wanted to identify whether there were any null values in the data but found none.

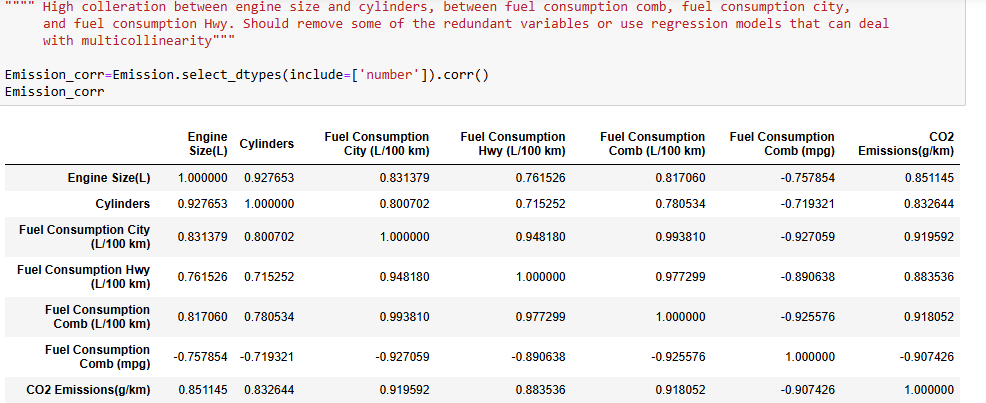


A screenshot of a computer

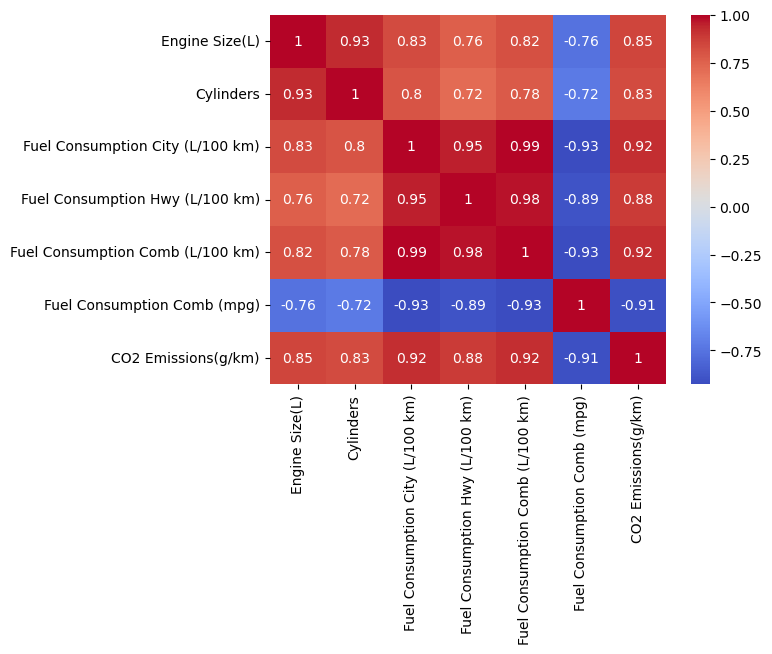
AI-generated content may be incorrect.

**Featuring Engineering for Linear Regression**

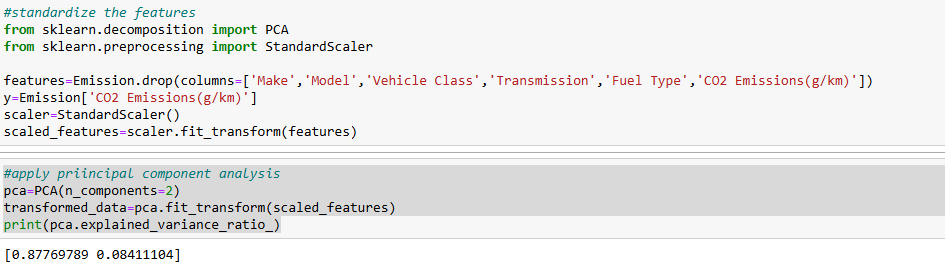
Next thing I wanted to look at was whether there was any correlation between the different features. The data showed a high correlation between the variables. This was data that contained multicollinearity. This will make it difficult to determine which variables have the largest impact on CO2 Emissions.





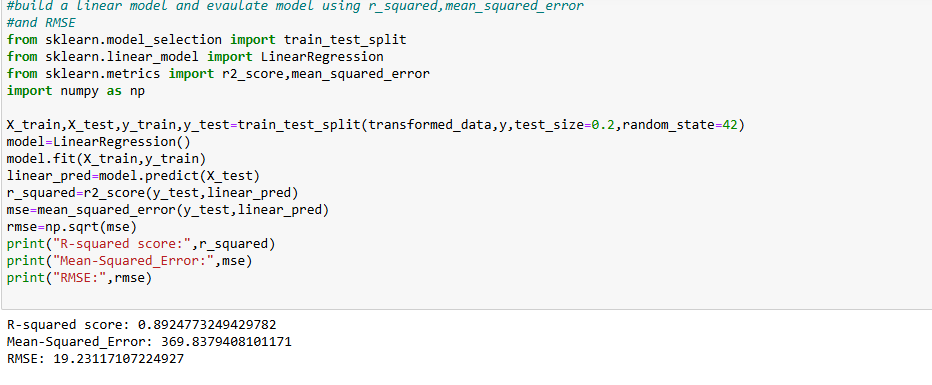


Before I applied a linear regression model I applied principal component analysis to the data. First, I standardized the numerical variables and then applied principal component analysis using two components to the standardized data and determined that the first component explained 87.7% of the variance in the data and the second component accounted for 8.41% of the variance.

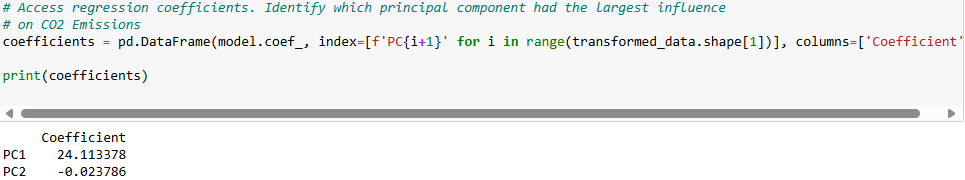


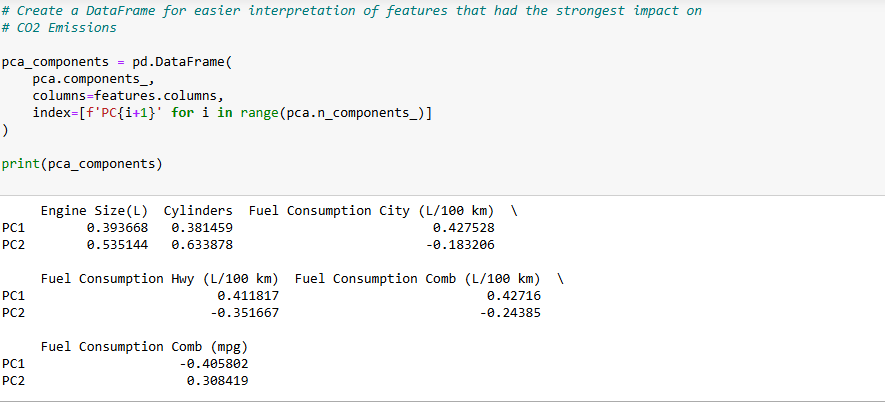
**Linear Regression Model**

Using the transformed data I used train\_test\_split on the data to build and fit a linear regression model. I evaluated the model using r squared, mean squared error and RMSE metrics and obtained the following results



The results show that the model is robust and a good fit for the data. Next, I wanted to identify the features that had the largest impact on CO2 emissions. First, I determined the regression coefficients for the principal components and determined that component 1 had the largest influence on CO2 emissions. After this I determined the features that had the greatest impact on CO2 emissions and identified that fuel consumption city, fuel consumption hwy and fuel consumption comb contributed the most to CO2 emissions.





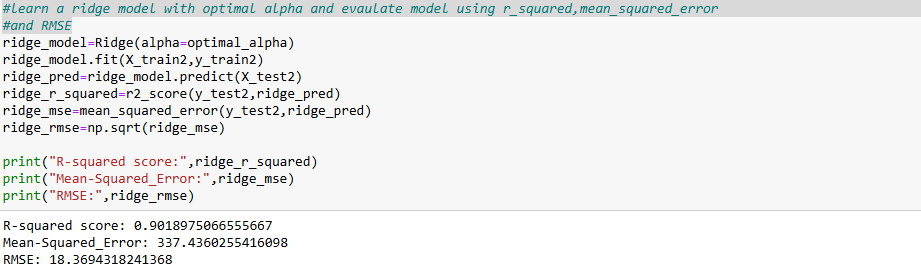
**Ridge Regression Model**

Next, I used Ridge Regression Model on the data set to identify the most significant features to CO2 emissions. I used train\_test\_split to split the standardized data and a range of alpha values to determine the optimal alpha to use

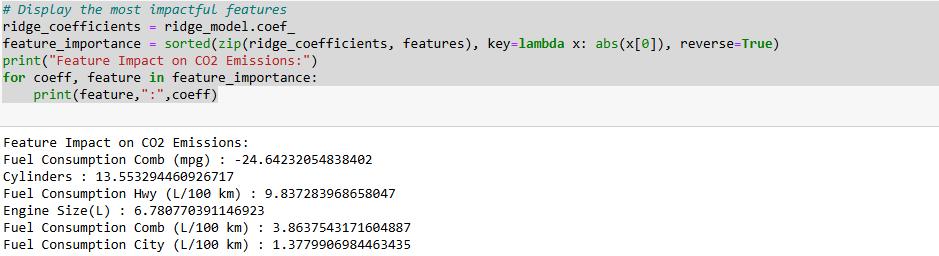
A screen shot of a computer

AI-generated content may be incorrect.

Using the optimal alpha, I built and fitted a ridge regression model to the data and evaluated using the same metrics as before and received the following results



This shows that the ridge regression model performed slightly better than the linear regression model with PCA. The model effectively handled the multicollinearity in the data. Using ridge regression, I identified the most significant features in the data. It clearly indicated that fuel consumption comb(mpg) was the most significant feature. It shows that as the fuel consumption comb(mpg) increased CO2 emissions decreased.



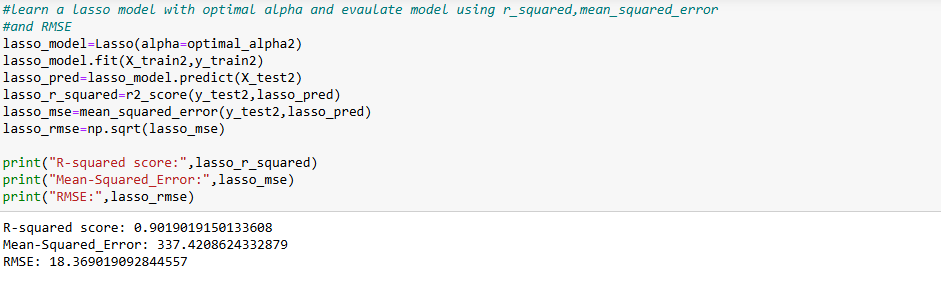
**Lasso Regression Model**

I used Lasso regression to identify the most significant features that impact CO2 emissions. I used train\_test\_split to split the standardized data and determined the optimal alpha to use

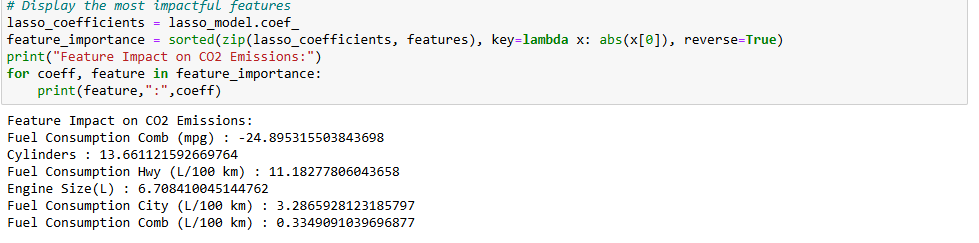
A white screen with blue text

AI-generated content may be incorrect.

Using the optimal alpha, I built and fitted a lasso regression model to the data and evaluated using the same metrics as before and received the following results



These are similar results to the ridge regression models and using lasso regression, I identified the most significant features in the data which was again fuel consumption comb(mpg)



**Results and Conclusion**

We make a comparison of the three models by observing the metrics

|  |  |  |  |
| --- | --- | --- | --- |
| Metric | Linear Regression with PCA | Ridge Regression with alpha =10 | Lasso Regression with alpha =0.0663 |
| R- squared | 0.895 | 0.9019 | 0.9019 |
| Mean squared error | 369.8379 | 337.4360 | 337.4209 |
| RMSE | 19.2312 | 18.3694 | 18.3690 |

We note that lasso regression performed slightly better than linear regression with PCA and ridge regression by a small amount. All models have a high R squared value suggesting a strong fit to the data. Lasso performed better in mean squared error and RMSE compared to the other models.

We also look at the most important features for each model by looking at the feature coefficients

|  |  |  |
| --- | --- | --- |
| Feature | Ridge Regression with alpha = 10 | Lasso Regression with alpha = 0.0663 |
| Fuel Consumption Comb (mpg) | -24.6423 | -24.8953 |
| Cylinders | 13.5533 | 13.6611 |
| Fuel Consumption Hwy (L/100 km) | 9.83738 | 11.1828 |
| Engine Size (L) | 6.7808 | 6.7084 |
| Fuel Consumption Comb (L/100 km) | 3.8637 | 0.3349 |
| Fuel Consumption City (L/100 km) | 1.3780 | 3.2866 |

|  |  |  |
| --- | --- | --- |
| Feature | PC1 | PC2 |
| Fuel Consumption Comb (mpg) | -0.405802 | 0.308419 |
| Cylinders | 0.381459 | 0.633878 |
| Fuel Consumption Hwy (L/100 km) | 0.411817 | -0.351667 |
| Engine Size (L) | 0.393668 | 0.535144 |
| Fuel Consumption Comb (L/100 km) | 0.427160 | -0.243850 |
| Fuel Consumption City (L/100 km) | 0.427528 | -0.183206 |

We see that fuel consumption comb (mpg) is the most significant feature for ridge and lasso regression and is also important for linear regression with PCA since we recall that PC1 had a larger influence on CO2 emissions compared to PC2.

If we display a graph, we can observe the inverse relationship between fuel consumption comb and CO2 emissions. So, the best way to decrease CO2 emissions is by increasing fuel consumption comb(mpg)

A graph of a graph showing the amount of gas consumption

AI-generated content may be incorrect.

Sorting the data by cars with high fuel consumption comb(mpg) we can identify that the top 20 vehicles that have 4 cylinders and use regular gasoline are the cars that produce the least amount of CO2 emissions. The 20 cars with the least amount of fuel consumption comb (mpg) are vehicles that have 8 cylinders and have ethanol blend as their fuel type. These are the cars that produce the highest CO2 emissions.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

My recommendation for any company who wish to build cars that produce low CO2 emissions is to make cars that have only 4 cylinders, use regular gasoline and have high fuel consumption comb (mpg).