

**PRINCIPLES OF STATISTICAL MODELLING**

**EXPLORATORY DATA ANALYSIS ON FUEL ECONOMY DATA**

By Rohit Anand

Advisor: Dr. Stefan Kettemann

Table of Contents

[Motivation 3](#_Toc42959697)

[Dataset Description 3](#_Toc42959698)

[Mathematical Formulations 4](#_Toc42959699)

[Missing Values 4](#_Toc42959700)

[Outliers 6](#_Toc42959701)

[Principal Component Analysis 7](#_Toc42959702)

[Clustering 7](#_Toc42959703)

[Correlation 8](#_Toc42959704)

[Conclusion 10](#_Toc42959705)

[Resources 10](#_Toc42959706)

# Motivation

In most of the countries, transportation has become the major cause of lung and other health problems. It is also the primary source of global warming pollution. New technologies are replacing reliability on natural fuel resources by electric cars while others who continue using it can be a source of increasing air pollution – unless preventive measures are being taken.

The dataset for this project has been extracted from [US Environmental Protection Agency](https://fueleconomy.gov/feg/download.shtml).

This data is composed ( ***Action of measuring data*** ) of the testing of vehicles done at the EPA laboratory in Michigan and also from the vehicle manufacturers in the year 2008. The test data is sent to the Department of Energy and Transportation and the Internal Revenue Services of the US so that they can carry out better administration and monitoring of the state economy related propagandas ( ***Events,*** ω ). If the MPG(Miles per gallon) values are too high on the fuel economy labels, it is required by the EPA that the manufacturers change or update these values on the vehicles. ***Universe,*** Ω can be related to the vehicle models manufactured in the US.

## Dataset Description

|  |  |  |
| --- | --- | --- |
| **all\_alpha\_08.xlsx (Year: 2008)** | | |
| **Column** | **Description** | **Important for analysis(1-5)?**  **1: lowest and 5: highest** |
| Model | Vechile model and make year | 1 |
| Displ | Engine displacement(in ltr): sweep volume of pistons | 4 |
| Cyl | Number of cylinders in an engine | 4 |
| Trans | Transmission type + number of gears | 2 |
| Drive | 2/4/all wheel drive | 3 |
| Fuel | Fuel used in the engine | 5 |
| Sales Area | Area | 3 |
| Stnd | Vehicle emission standard code | 2 |
| Stnd Description | Vehicle emission standard code description | 1 |
| Underhood ID | Engine Family | 2 |
| Veh class | EPA Vehicle class | 3 |
| Air pullution score | Smog rating | 5 |
| City MPG | Fuel economy in city (miles per gallon) | 4 |
| Hwy MPG | Fuel economy in highway (miles per gallon) | 4 |
| Cmb MPG | Fuel economy in city+highway (miles per gallon) | 4 |
| Greenhouse gas rating | Vehicle tailpipe emissions on CO2 (Range: 1 - 10) | 5 |

# Mathematical Formulations

A brief explanation of the ***Universe,*** the ***Events*** and the ***Action of measuring data*** has been reflected in the above section. In this part, we will concentrate more on the specific ***Random Variables*** which are of importance in this report and their respective ***Data Value Space*** (DVS). Other less important features have been removed.

|  |  |
| --- | --- |
| **Random Variables ( *X* )** | **Data Value Space and Data Type** |
| Displ | {2,3,4,5,6,7} ***originally continuous but converted to discrete*** |
| Cyl | {2, 4, 5, 6, 8, 10, 12, 16} |
| Air pullution score | {1.0, 3.0, 4.0, 6.0, 7.0, 8.0, 9.0, 9.5} |
| City MPG | {8.0, 9.0, 10.0, 11.0, 12.0, 13.0, 14.0, 15.0, 16.0, 17.0, 18.0, 19. 0, 20.0, 21.0, 22.0, 23.0, 24.0, 25.0, 26.0, 27.0, 28.0, 29.0, 33.0, 34.0, 35.0, 40.0, 48.0} |
| Hwy MPG | {12.0, 13.0, 14.0, 15.0, 16.0, 17.0, 18.0, 19.0, 20.0, 21.0, 22.0, 23.0, 24.0, 25.0, 26.0, 27.0, 28.0, 29.0, 30.0, 31.0, 32.0, 33.0, 34.0, 35.0, 36.0, 37.0, 45.0} |
| Cmb MPG | {10.0, 11.0, 12.0, 13.0, 14.0, 15.0, 16.0, 17.0, 18.0, 19. 0, 20.0, 21.0, 22.0, 23.0, 24.0, 25.0, 26.0, 27.0, 28.0, 29.0, 30.0, 31.0, 32.0, 34.0, 42.0, 46.0} |
| Greenhouse gas rating | {0.0, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0} |

## 

## Missing Values

|  |  |
| --- | --- |
| **Columns** | **Row count with null values** |
| model | 0 |
| displ | 0 |
| cyl | 199 |
| trans | 199 |
| drive | 93 |
| fuel | 0 |
| cert\_region | 0 |
| veh\_class | 0 |
| air\_pollution\_score | 0 |
| city\_mpg | 199 |
| hwy\_mpg | 199 |
| cmb\_mpg | 199 |
| greenhouse\_gas\_score | 199 |
| smartway | 0 |

Before fixing/removing the missing values, data cleaning and preprocessing needs to be done.

* Replacing blank space by an underscore in the column names.
* *Cyl* has values in format: (6 cyl), (4, cyl) etc
  + The actual number of cylinders, i.e, 6, 4 was extracted from the column
* 199 rows had missing values for the columns *cyl, trans, city\_mpg, highway\_mpg, cmb\_mpg, greenhouse\_gas\_score*.
* 63 duplicate rows found and dropped
* For the features with properties of an hybrid vehicle, the data was in a multivalued format, *fuel\_type\_1/fuel\_type\_2,* which was then split and again appended to the main dataframe with separate rows. List of columns exhibiting such properties: [*'fuel', 'air\_pollution\_score', 'city\_mpg', 'hwy\_mpg', 'cmb\_mpg', 'greenhouse\_gas\_score'*]
* In the end, the datatype of relevant columns were changed according to the requirement and analysis to be carried out in the rest of the report.
* Figures below highlight the distribution among the important features which has been presented in the report.

A screenshot of a cell phone

Description automatically generated

The number of cylinders ( *Fig 1* ) among different car model has

* Mean = 6
* Skewness = 0.90
* Kurtosis = 1.46
* Standard Deviation = 1.82

*Fig 1*

A close up of a map

Description automatically generatedThe Piston Displacement inside each cylinder in an engine (*Fig 2*) among different car model has

* Mean = 4.02 cm
* Skewness = 0.44
* Kurtosis = -0.70
* Standard Deviation = 1.31

*Fig 2*

A screenshot of a cell phone

Description automatically generatedThe Mileage in Miles per gallon in highway among different car model ( *Fig 3* ) has

* Mean = 23.44 miles/gallon
* Skewness = 0.17
* Kurtosis = -0.06
* Standard Deviation = 4.95

*Fig 3*

# Outliers

Outlier detection method

* The Z-score has been calculated cosidering the observations 3 standard deviations away from the mean.
* It has been found that 10.01% oft he observations are falling outside the observations which are 3 standard devations from the mean. Therefore, these are not included in out further analysis
* Reason for removing outliers: Principal component analysis is highly affected by outliers

A picture containing clock

Description automatically generated

A picture containing clock, door

Description automatically generated

*Fig 4 Fig 5*

A screenshot of a cell phone

Description automatically generated A picture containing clock

Description automatically generated

*Fig 6 Fig 7*

* Comparing the above figures, we can notice the significant amount of reduction in outliers. ( *Fig 4 vs Fig 6*) and ( *Fig 5 vs Fig 7*)

# Principal Component Analysis

* Finding the principal components which explain most of the variance in the dataset.
* Scaling the observations prove benefitial for PCA and clustering.
* Optimal number of components was selected to be 3 which explains 94.78% of the variance in the features ( *Fig 8* ).

A close up of a map

Description automatically generated

*Fig 8*

# Clustering

* This report uses K-Means clustering as a parameter to recognize clusters in the dataset which can be benefitial for finding relationships.
* Eucledean distance has been used to find the distance between cluster centroids and the obervation points.
* Finding the optimal number of clusters has always been a challenge. This report highlights one of the method, WCSS curve or the elbow curve *( Fig 9 )* to figure out the best value of k.
* WCSS: Weighted Cluster Sum of Squares

A screenshot of a cell phone

Description automatically generated

*Fig 9*

* Since it appears that choosing the value of k can be difficult. So, k=2 ( *Fig 10* ) or k=3 ( *Fig 11* ) can be choosen and plotted.

A close up of a map

Description automatically generatedA close up of a map

Description automatically generated

*Fig 10 Fig 11*

# Correlation

Correlation between the features expressed as a heat map:

A screenshot of a cell phone

Description automatically generated

*Fig 12*

* The values on the square boxes in this matrix *( Fig 12 )* represents the correlation coefficient between the important features.
* Correlation coefficient tells about the relationship between a pair of features which ranges between -1 to 1.
* The more this values is closer to 1 means a strong positive relationship.
* The more this values is closer to -1 means a strong negative relationship.

*A screenshot of a cell phone

Description automatically generatedFig 13* represents Vehicle Mileage in Highway ( miles per gallon ) vs Vehicle Mileage in Highway and city ( miles per gallon )

This figure shows a high positive correlation between these two features with = 0.98

*Fig 13*

A screenshot of a cell phone

Description automatically generated

*Fig 14* represents Piston displacement in a engine cylinder(in cm) vs Vehicle Mileage in Highway and city ( miles per gallon )

This figure shows a high negative relationship between these two features with = - 0.81

*Fig 14*

*Fig 15*A close up of a piece of paper

Description automatically generatedrepresents Greenhouse Gas Score vs Air Pollution Score of the vehicle models.

This figure shows almost no relationship between these two features with = 0.28

*Fig 15*

# Formulae

* Mean, =
* Standard Deviation, =
* Skewness =
* Covariance, ***cov ( x , y )*** *=*
* Correlation, =

Where *x* and y are random variables

# Conclusion

Since the report took a step wise process oft he analyzing the Fuel Economy Data, it can be therefore concluded that the Greenhouse Gas Score is highly dependent on the features:

* Cyl, i.e, Number of Cylinders in a vehicle engine
* Displ, i.e, Piston Displacement in a cylinder of an engine
* City\_mpg, i.e, Mileage of a vehicle measured inside a city ( in miles per gallon )
* Hwh\_mpg, i.e, Mileage of a vehicle measured on a highway ( in miles per gallon )
* Cmb\_mpg, i.e, Mileage of a vehicle measured in a city and on highway combined

( in miles per gallon )

The code repository has been published in Github which can be accessed through this link

# Resources

<https://fueleconomy.gov/feg/EPAGreenGuide/GreenVehicleGuideDocumentation.pdf>

<https://fueleconomy.gov/feg/download.shtml>

<https://www.epa.gov/compliance-and-fuel-economy-data/data-cars-used-testing-fuel-economy>

<https://www.ucsusa.org/transportation/technologies>