Qualitative and Quantitative Analysis of Environmental Awareness in Transport Sector

Samyak Shah
Department of Electrical Engineering
Indian Institute of Technology Bombay
Mumbai, India
samyakshah@iitb.ac.in

Neelkamal Bhuyan
Department of Electrical Engineering
Indian Institute of Technology Bombay
Mumbai, India
180020058@iitb.ac.in

***Abstract*—Sustainable environmental practices are the need of the hour. Vehicular pollution along with industry emissions is the major contributor to air pollution. Electric vehicles are gradually gaining popularity and are being widely supported as these are not dependent on fossil fuels. Through this project we would use Car Fuel and CO_2 data to dig deeper into whether car manufacturers have adapted to the environment regulations set up by national government, how does fuel and car type affect its emissions and much more. At the end of this project we seek to get a clearer picture about how the emissions of conventional fuel run cars have changed over a decade.**

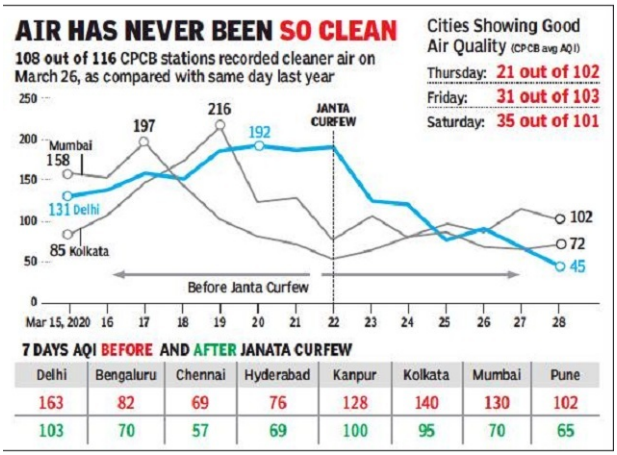
***Index Terms*—Data Analysis, Machine Learning, Air Pollution, Car Fuel Emissions, CO_2 Emission**

I. Introduction

Air pollution is a problem that scientists have been trying to solve for decades. Some of the constituents of car fuel emissions like carbon monoxide (CO) can lead to complicated health conditions. Dust in air causes allergies and troubles people with sensitive respiratory tracts. Vehicular emissions are considered among the major contributors to the declining air quality. Through this analysis we wish to answer four basic questions-

1) How has the vehicular emission levels changed over the span of a decade? We also wish to identify the rising or falling trends for various constituents of vehicular exhaust.
2) How have the European Standards for pollution control shaped over years?
3) How have different car manufacturers performed in adhering to the set regulations? Cars of which manufacturer fare badly compared to the rest in the amount of harmful emissions they produce? We also wish to discuss how the amount of pollution varies by engine type and the fuel type used by the cars.

Via this project we seek to present an extensive exploratory data analysis to the reader to enhance his or her understanding of the topic. Due to the importance this analysis holds to human life in general we use highly trustworthy data sources for our analysis. These are made available by the courtesy of the Vehicle Certification Agency (VCA), an Executive Agency for Transport.



Delhi	Bengaluru	Chennai	Hyderabad	Kanpur	Kolkata	Mumbai	Pune
163	82	69	76	128	140	130	102
103	70	57	69	100	95	70	65

Fig. 1. News Snippet from Times of India on March 29, 2020 [1]

Fig.1 above shows that the air quality improved significantly in the major cities of India after lockdown was imposed during the coronavirus pandemic. During the lockdown, non essential vehicle use was minimized by the regulations imposed. This led to a dramatic increase in the air quality in India. This ascertains that vehicle use and their emissions are the primary source of air pollution.

This Analysis Report takes a data driven approach to answer some common questions with data backed evidences. In this report we take the position that although, traditional car manufacturers have more or less changed their ways to manufacture more environment-friendly vehicles, the emission produced by them is still a significant non-zero value. This automatically will pave the way for better alternative options like cars, scooters and trucks that run on electrical power.

In the later sections of the paper, we produce a systematic machine learning model which can predict the amount of CO_2 emissions given the features of the car like the manufacturer, fuel type etc. as the input. This machine learning model has been a result of tuned hyper parameters and extensive data cleaning procedures. Mean Squared Error of the predicted values with the real values was used as an evaluation metric for the model. We have included the code files in the form of Python Notebooks in the submission zip folder. The readers

are requested to use it to peruse the code and the resulting graphs and to gain familiarity with the data in general. The code is written to be reproducible.

II. Datasets

We have chosen the Car Fuel and Emissions 2000-2013 dataset [2] from data.world. This dataset was originally published by the Vehicle Certification Agency (VCA), an Executive Agency of the Department for Transport. It is assumed that the data is released under the UK Open Government License. The dataset in itself is a combination of 32 different smaller files whose names have been mentioned in the column named 'file' (It is the first (or the zeroth) column).
The dataset has 45111 rows and 30 columns. The attributes (columns) include details like Maufacturer name, Fuel Type, year of production, CO emissions etc. The dataset although

year	manufacturer	model	description	euro_stan	tax_band	transmissi	transmissi
2000	Alfa Romeo	145 Range	1.6 Twin Spar	2		M5	Manual
2000	Alfa Romeo	145 Range	1.8 Twin Spar	2		M5	Manual
2000	Alfa Romeo	145 Range	Cloverleaf	2		M5	Manual
2000	Alfa Romeo	146 Range	1.6 Twin Spar	2		M5	Manual
2000	Alfa Romeo	146 Range	1.8 Twin Spar	2		M5	Manual
2000	Alfa Romeo	146 Range	Ti	2		M5	Manual
2000	Alfa Romeo	156 Range	1.6 Twin Spar	2		M5	Manual
2000	Alfa Romeo	156 Range	1.8 Twin Spar	2		M5	Manual
2000	Alfa Romeo	156 Range	2.0 Selespeec	2		SAT5	Automatic

Fig. 2. A few sample entries in the dataset

appears complete and clean on the first glimpse, presents quite some challenges for analysis. A few of the columns have a lot of null values and they need to be dealt appropriately. There are a few outliers too, which may be a result of incorrect data entries while the dataset was typed or published. Some of the datatypes need to be corrected too. The fuel cost column was split in the dataset into two incomplete columns. We transform it into a single column with no null values.
For the Machine Learning part with the data, we did an extensive analysis to understand the correlation between the independent variables and removed a few data columns in case of low correlation observed. The remaining columns were chosen as the input to the machine learning model. We understand that one needs to be careful while choosing the columns and rows before modelling any Machine Learning model. This ensures that our model is genuine and gives the best possible result.

III. Related Work

Since the dataset used in the project isn't very easy to procure and is neither available on popular sites like Kaggle or UCI Machine Learning Repository, little previous work has been done (as much as we could find) with the dataset.
Although a popular subject, most of the analysis reports are made by media outlets, their work is less descriptive and little to no insight is provided by them about how they deal with the data.
On the contrary our project holds relevant as it can be the one-stop solution to most of the data driven questions about Car Fuel Emissions. The project encapsulates our best efforts; we are open to feedback from students and data science researchers. The authors of the report can be reached out via emails mentioned at the top of the report.
One of the works by Tong, Hung and Cheung named- 'On-Road Motor Vehicle Emissions and Fuel Consumption in Urban Driving Conditions'[3] follows somewhat similar path as we do, but the dataset we chose is different and the context of their work and the questions answered in their report are more related to mathematical modelling than analysis using the data provided.

IV. Analysis Pipeline

The core targets of the analysis are:

1) Study how much work has the conventional car industry done, in reducing the carbon footprint and other pollutant emissions of their models, over a decade.
2) The relationship between the mechanical attributes of the car and its emissions
3) Manufacturer-wise analysis of how important to them sustainability and eco-friendliness was.
4) Evolution of Euro Standards over a decade through the general emissions levels of vehicles over the years .

The dataset rows will be grouped by the criterion of specific categorical variable to analyse the change of specific output variables as the value of the grouping variable changes. The categorical variables observed in the dataset are:

1) manufacturer
2) euro standard
3) transmission type
4) fuel type
5) Year

Following are the output variables in the dataset, they indicate the level of pollution (in any form):

1) noise level
2) CO_2 emission
3) CO emissions
4) NO_x emissions
5) particulate emissions

Before starting any analysis we will observe the dataset using the 'display' function in python to observe the type and values of the attributes.
Then, we print the number of null values in each column, in case of a significant number of null values, the specific rows are removed, or the entire column may be removed; the final decision is dependent on what column it is and how important it is for a particular part of the analysis pipeline.
We then change some data types for ease of operations in the later stages of the analysis. Fuel Cost data is given some under 6000 miles aggregate and rest under 12000 miles aggregate. This will be combined under one column of average fuel cost/mile.
Note: We only attach a few figures in this report; for viewing all the figures please refer the Python notebook containing all the code and the results.

A. Year Wise Analysis

The brief steps involved for this part are:

1) Take mean of all emission data, noise level and fuel price.
2) Plot time-series line graph for each type of emission and noise level.
3) Plot line graph (time-series) of both year avg fuel price and avg CO_2 emission and superimpose the graphs for better comparative analysis.
4) Repeat step 3 for the case of CO emissions and Nitrous Oxides (NO_x) emissions.

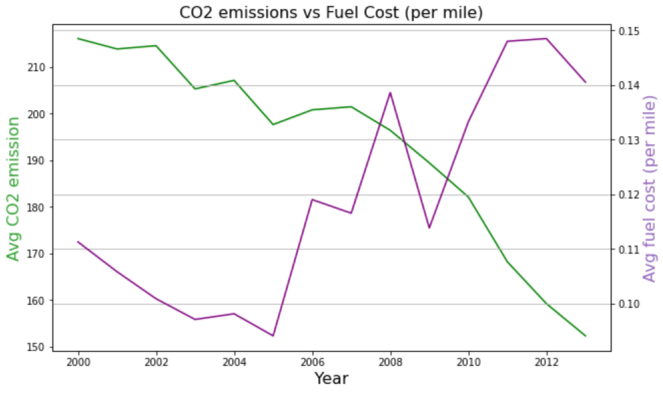


Fig. 3. Average CO_2 emissions and average fuel price year-wise

B. Euro Standard Wise Analysis

The brief steps involved for this part are:

1) For each Euro standard take the average of each type of emission and noise level.
2) Plot Bar Graph, separately for each type of emission and noise level
3) For each Euro standard take
   a) Average of engine capacity
   b) The more frequently occurring fuel type
   c) The average fuel cost, separately for each fuel type

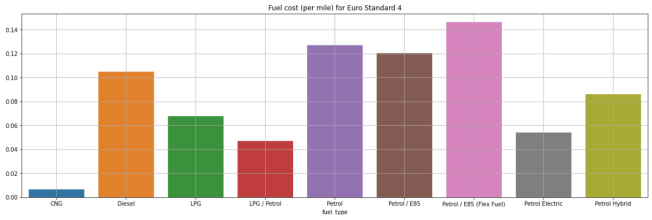


Fig. 4. Average Fuel Cost for Euro Standard 4

4) Use Bar Plot for Fuel cost per mile for each European standard, line plot for engine capacity and bar plots for fuel type count.

C. Manufacturer and Year Wise Analysis

The brief steps involved for this part are:

1) Look at percentage of CO_2, CO and NO_x emissions contributed by each manufacturer through pie chart
2) Use **Spider-Chart** for comparing two companies on various parameters related to efficiency. The starter code of the spider plots was inspired from python-gallery [4].

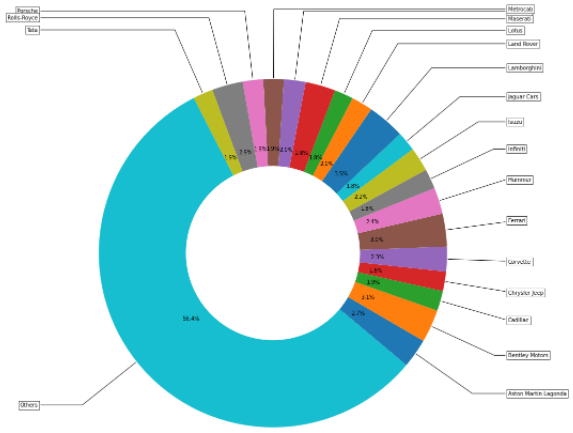


Fig. 5. Manufacturer-wise CO_2 emissions, pie chart

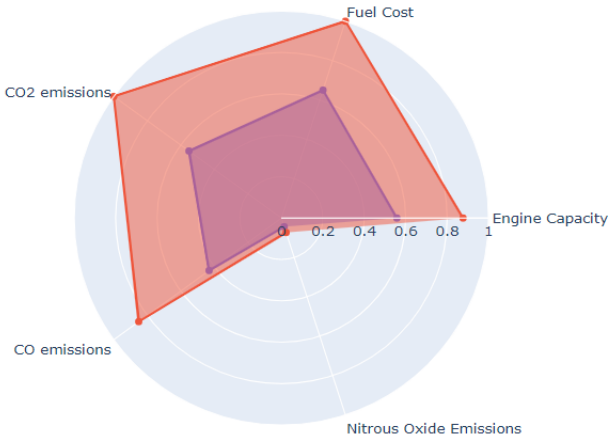


Fig. 6. Spider Plot, red is Lamborghini, blue is Porsche

D. Using Machine Learning

The brief steps involved for this part are:

1) Print a heatmap of the variables to know their correlation.
2) Only retain the columns on which we wish to do analysis or regression further. Drop all the other columns.
3) Remove all the rows which have null value in any of the columns. This is our way of removing the entries (rows) where the data is incomplete.
4) See if we can characterise the dependent variables as a function of the independent variables by looking at their one-one scatter plots.

5) Remove outlier points. This is our way of removing points where there may have been a data entry mistake.
6) We observe from the heatmap and the scatter plots that CO emissions and Nitrous Oxide emissions have almost no dependence on the independent variables.
7) Apply linear regression on CO_2 emissions with the engine type and the fuel cost as the input variables. Split the dataset into training, dev and test data. We will use RMSE as the performance metric for the model.
8) Due to limited computational resources, we try using a neural network with two layers and four hidden units in each. We will employ keras for the same.
9) To concretize the results obtained by the linear regression model in step 6 and the neural network in step 7, use Kernel SVM to compare and validate the results.

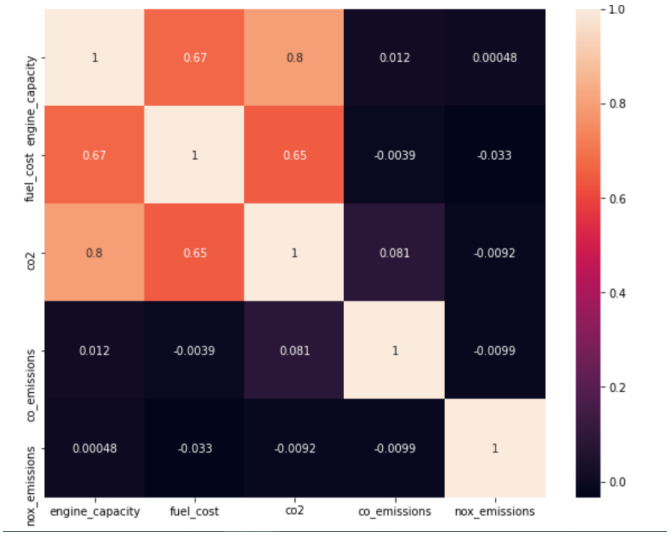


Fig. 7. Correlation heatmap between the variables

V. Results

In this section we will do a subsection-wise listing of the conclusions, implications and results:

A. Year Wise Analysis

The major conclusions are:

1) CO_2 emissions have been steadily reducing over the years.
2) Except a peak in 2005, CO emissions are also reducing over the years.
3) Nitrous Oxide emissions have not reduced significantly over the years. In addition to that we have high NO_x pollution in 2003,2004 and 2009.
4) Significant decrease in Noise Pollution has been observed between 2008 and 2011.
5) From the second set of graphs in this section it is clear that as we are trying to **reduce emissions** it is getting **costlier to run a vehicle**.

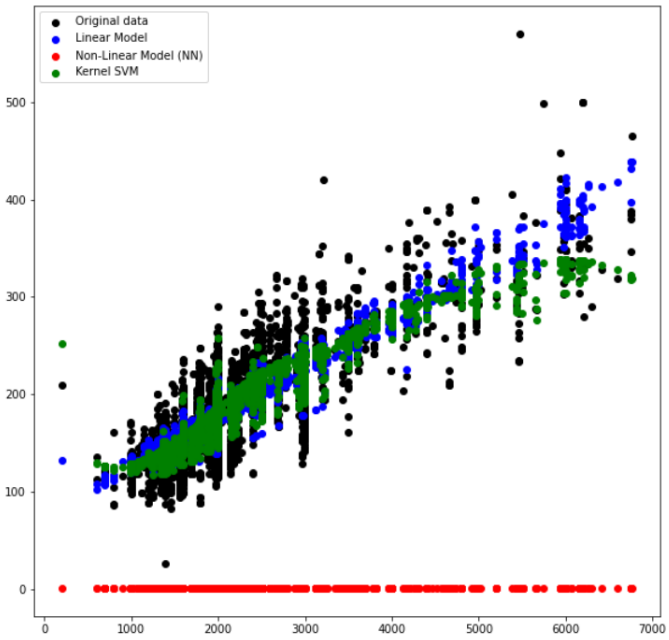


Fig. 8. CO_2 emissions versus the engine type. The graph shows the original data alongside the outputs using Kernel SVM, Neural Networks and Linear Model.

B. Euro Standard Wise Analysis

The major conclusions are:

1) Across all Euro Standards CO_2, CO and Nitrous Oxide emissions are decreasing except from the peak in NO_x emissions from standard 2 to 3
2) This goes to show that the governments have been constantly trying to reduce impact of transport on environment
3) In addition to that Noise pollution also seems to decrease in general, except for the case of standard 5 to 6.
4) Apart from the increase from Euro Standard 2 to 3, the engine capacity of vehicles, in general, seem to decrease.
5) In all Euro standards, Petrol cars seem to be the costliest to run, based on money spent on fuel/mile
6) Diesel-run vehicles follow petrol-run ones very closely in all Euro Standards
7) It is observed across the Euro Standards that coupling traditional petrol/diesel vehicles with hybrid/electric technology reduces the cost of running significantly
8) All electric cars, introduced in Euro Standard 6, have a significantly lower cost of running compared to petrol and diesel counterparts. **Note:** Fuel cost refers to cost of running based on money spent on fuel, am indication of mileage.
9) Across all Euro standards Petrol and Diesel cars are the most sought after with Petrol being more popular among the two (except in Euro Standard 5).

C. Manufacturer and Year Wise Analysis

The major conclusions are:

1) Most companies have done a good job at reducing CO2 emissions with a peak in one year
2) Metrocab, Micro Compact Car and Perodua have increasing trend and hence need to focus on reducing CO2 emissions in their cars
3) Corvette had an increasing trend initially but in recent years seem to have attended to this issue as seen by the decreasing trend thereafter
4) Daihatsu and MG Rover group had decreasing emissions but now have increasing emissions
5) Most companies have done a good job at reducing CO emissions with a peak in one year
6) Some companies like Suzuki, Roll-Royce and Ford have a decrease and then a significant increase in CO emissions
7) Companies like Isuzu, MG Rover, Bentley, Abarth and Infiniti have increasing trend showing that they need to focus on reducing CO emissions.
8) Dodge performs the worst with increasing NO_x emissions each year
9) Ferrari and Lamborghini do a good job with decreasing NO_x emissions each year
10) Hummer does not seem to have made any progress at all in this regard.
11) Even though all companies more or less contribute equally to CO_2 emissions, it is important to note that Lamborgini contributes the most, individually.
12) Even though all companies more or less contribute equally to CO emissions, it is important to note that Corvette and Daewoo Cars contribute the most, individually.
13) Metrocab, individually, contributes the most to Nitrous Oxide emissions. This can be expected as its a taxicab manufacturer and therefore doesn't have environmental protection as one of its primary goals

D. Head to Head Manufacturer Comparison

1) Renault and Nissan perform more or less same in all the areas, which is expected in the low price segment as their is a lot of competition
2) Honda and Toyota perform more or less same in all the areas, which is expected in the lower mid-range segment as their is a also a fair amount of competition
3) Volvo and Volkswagen perform similar in all areas except for two. In CO emissions, Volvo seems to perform worse than Volkswagen while the opposite is true in the case of Nitrous Oxides emissions
4) Audi and BMW perform similar in all areas except CO emissions where Audi performs a lot worse than BMW. Hence, it is advisable to buy a BMW
5) Lamborghini performs very bad in comparison to Porsche in all the areas showing that Lamborghini doesn't have environmental protection as one of its goals

E. Using Machine Learning

The major conclusions are:

1) CO_2 emissions have a strong correlation with the dependent variables
2) CO and Nitrous Oxide emissions do not seem to have a strong correlation with the dependent variables
3) It goes to show that we may find a definite function for CO2 emissions but not necessarily for CO and Nitrous Oxides (in terms of fuel cost and engine type)
4) Looking at scatter plots with respect to both Engine Capacity and Fuel Cost, we can go ahead with Linear Model
5) Comparing results of Linear Model, NN model and Kernel SVM, we see that Linear Model performs far better than a Non-Linear model
6) We also see that Kernel SVM follows the Linear Model, further validating that a Linear Model is being followed by CO_2 emissions
7) We can therefore conclude that a Linear model represents the CO_2 emissions' dependence on Engine Capacity and Fuel Cost better than a Non-Linear Model

VI. Discussion

At the end, we can say that we have got a fair answers to the basic questions on which we built this project. The pollution has surely reduced due to the stricter rules from the controlling agencies, it's still non-zero and very significant. Since the parent dataset was supplied by an agency in the UK and contains information about the European Standards, an interesting extension of the project may be to analyse, using similar methods, the data for the Indian, Chinese, African or the Middle Eastern market. Although good datasets may be difficult to procure for every case.

The results which we obtained for the Exploratory Data Analysis are fairly strong as we could observe the distinguishing trends quite well. The results of the machine learning model may be a bit weak since we were unable to use larger neural network model due to the constraint of computational power we had.

We strongly believe that this dataset can serve pretty well to the students of Energy Sciences and Transport Engineering for their academic use or reference. The analysis report may be very useful for the car manufacturers in understanding the competition and will help them introspect their own performance in terms of adherence to regulation and environment-friendliness. Our analysis is of course useful for the agencies which regulate the Eurpoean Standards and monitor the emissions caused by cars.

Our project opens up some important possibilities and opportunities for extensions- for example we can expand our analysis to more recent data and see if the trends change or not. We can include more data rows from the newly popular electric vehicles and understand how much they perform better than vehicles run by traditional fuels.

We would be glad to receive more ideas about alternate approaches and project extensions from the reader.

ACKNOWLEDGMENT

The COVID-19 pandemic has shifted our college in the online mode; studying, work and communication has been more difficult than before. We would like to thanks the instructors for the course DS 203- Prof. Amit Sethi, Prof. Manjesh Hanawal, Prof. Sunita Sarawagi and Prof. S. Sudarshan along with the excellent team of Teaching Assistants for the course who guided us and ensured that our doubts are solved on time. All their suggestions have been important guiding steps for our project.

REFERENCES

[1] Article. Bhattacharya A. 'Delhi has best AQI ever, 'good air' day in 34 other cities'. Times of India: online news website. March 29, 2020.

[2] Dataset. Vehicle Certification Agency. Title of Dataset: 'Car Fuel Emissions 2000-2013'. London, United Kingdom. 2016

[3] Tong H., Hung W., Cheung C., 'On-Road Motor Vehicle Emissions and Fuel Consumption in Urban Driving Conditions'. Journal of the Air Waste Management Association, Volume 50, 2000 - Issue 4. December 27, 2011.

[4] Computer Code. Holtz Y. ' 390 The Basic Radar Chart'. The Python Graph Gallery. 2019.