RESEARCH ARTICLE OPEN ACCESS

One-Way ANOVA (Analysis Of Variance) – Carbon Dioxide Emission of Vehicles and Gasoline Type Used

Parameswara Rao Kandregula*
*(Alumnus, Western Governors University, Houston, Texas
Email: param.index@gmail.com)

****

Abstract:

This paper will show the statistical analysis on carbon emission of vehicles based on gasoline type to check if there is any correlation between the gasoline type used in vehicle and carbon dioxide emission. It would show the data extraction, cleaning techniques and One-way ANOVA (Analysis of Variance) using publicly available vehicle carbon emission data and SAS, statistical software tool.

Keywords — **Statistics, ANOVA, analysis of variance, Carbon Dioxide Emission, Gasoline Type, Data Analytics, Data Cleaning, SAS**.

****

I. INTRODUCTION

Vehicle users apply type of gasoline mostly on basis of vehicle manufacturer recommendation, assuming it is better for vehicle in long run. But the recommended gas might not always be the best for environment. If any of the gasoline is statistically better than other gasoline in Carbon Dioxide emissions then we can encourage people to use the desired gasoline with lower carbon dioxide emission, encourage vehicle manufacturers to recommend the environment friendly gasoline options in addition to vehicle friendly and governments can also provide incentives for environment friendly gasolines. In the following sections of paper, we will apply statistical analysis to explore if there is any statistically significant difference in carbon dioxide emissions of a vehicle based on type of gasoline used – Regular, Premium and Midgrade.

II. DATA ANALYSIS KEY CONCEPTS

Data Analysis and Statistics is huge area, covering all of which is beyond the scope of this paper, but some of the key terms which will help understand this analysis better.

A. Statistical Significance and p-value

Finding impact of data on a result and relations between data are prime area of data analysis. But with so much of data, it is very important to take into account only that impacts results with mathematical confidence. The numerical threshold of measurement of a data element or group before it can be considered influencing the desired result is called Statistical Significance. One the formal definition of statistical significance from a Harvard Business article [1] is “Statistical significance helps quantify whether a result is likely due to chance or to some factor of interest”. It is often measured as p-value with value between 0 and 1.

B. Null Hypothesis

Null hypothesis is the core of all statistical results analysis. It is inverse of the hypothesis which is assumed before start of experiment. Based on experiment results data and the statistical significance of the result, the null hypothesis is either rejected or accepted. Normally null hypothesis is the negation of the relation (hypothesis) which is being tried to be proven by an experiment. Based on the results being above or

below the p-value, the hypothesis is either accepted or rejected.

III. DATA COLLECTION AND PREPARATION

Quality Data is the strength of any statistical analysis. Wrong sources of data or impurities in data can lead to incorrect conclusions. Thus, data should be collected from reliable sources and cleaned before starting any analysis.

A. Collection

To do this analysis, data is needed on Carbon dioxide emissions for vehicles using different types of gasoline.

Methodology Used – Data mining of existing documents. This data was collected by Environmental Protection Agency of USA and available on their website [2] for public view and use.

Advantages – i) The data about carbon emission is very accurate as it is collected by government agency in a lab under stringent guidelines of Environmental Protection Agency. Thus, it would not be biased ii) It had data for multiple years and various models of vehicles. Thus, helping the randomness of the data. Disadvantages – i) The data is having several extra unnecessary columns of data. ii) Since this is already collected data, it could have unequal distribution of desired factor of types of gasoline –Regular, Premium and Midgrade.

Challenges – i) This data cannot be collected by an individual or small organization as it would need heavy equipment to test the vehicle, measure and record it. ii) This data cannot be obtained from any common vehicle websites. This might be available with Car manufacturers, but they would not release it openly as it is not positive marketing factor. Thus, had to rely on a government agency to collect it and share.

B. Extraction and Preparation

Source of Data: The data was downloaded from United States Environment Protection Agency. (Updated: Wednesday June 13 2018). Datasets for All Model Years (1984–2019) [3]

It had: 39958 observations(rows) and 83 parameters (columns) in csv format.

Tool used to extract Data: Microsoft Excel

Reason / Advantages: i) Excel can handle large csv data file easily and transform to rows and columns easily. ii) Excel has many inbuilt easy to use tools to filter and clean data.

Disadvantages: i) Microsoft Excel needs license

Methods to Extract Data:

Online download and offline extraction – The source website do not have filters to extract data. Thus, the full file is downloaded, and the extraction is done offline. Disadvantage - This is more of manual process.

Data Cleansing – The data has 83 columns of irrelevant data, but we need mainly only the carbon dioxide emission quantity and type of gasoline and id – 3 columns

Data Cleansing – The data has about other types of fuels like ethanol, electricity but we are only interested in gasoline -regular, premium, midgrade. This gave us 37062 observations.

Data Transformation – Fuel Type and ID values are fine as expected. Co2 emission values had varying decimal places from 2 to 9. Changed the format of it to two decimal place numbers.

Imputation – There were no missing values found.

Steps to extract the data:

Initial data

It had: 39958 observations(rows) and 83 parameters (columns) as shown in Fig 1.a and Fig 1.b

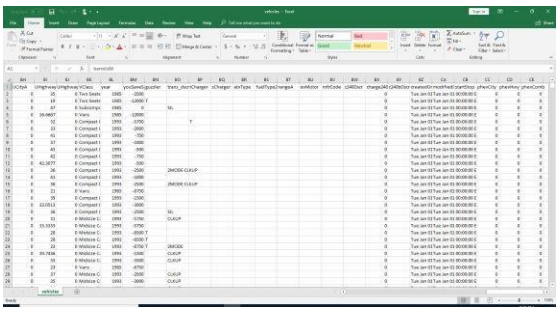

Fig. 1.a Source Data

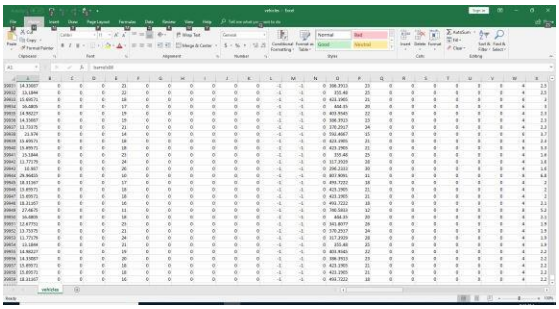

Fig. 1.b Source Data

Removed all unnecessary columns except Co2TailPipeGPM, FuelType and Id as shown in Fig. 2.

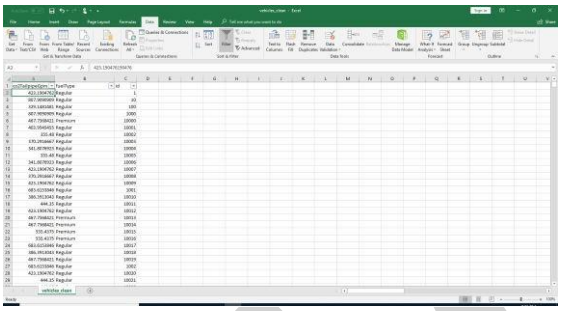

Fig. 2 Removed columns

Removed other fuel types: we get valid 37062 observations as shown in Fig. 3.a and 3.b

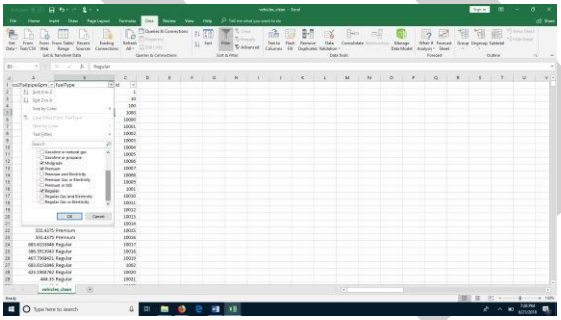

Fig. 3.a Removed Other Fuel Types

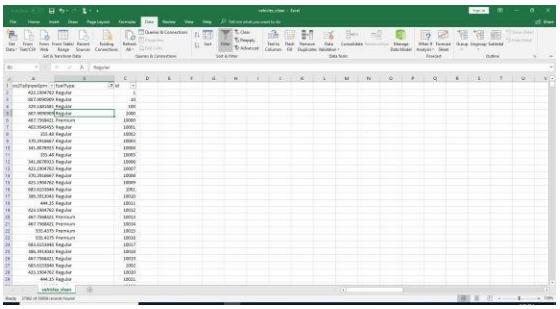

Fig. 3.b Removed Other Fuel Types

Changed data format of co2tailpipeGPm to two decimal formats as shown in Fig. 4.

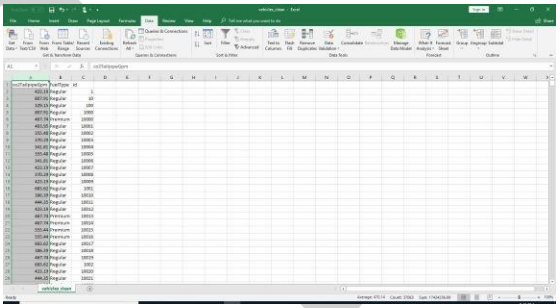

Fig. 4 Changed data format

Final File: 37062 observations and 3 columns as shown in Fig. 5.

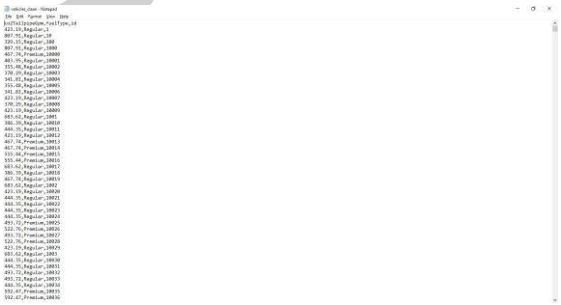

Fig. 5 Final file

IV. ANALYSIS AND OUTPUT

Quality Data is the strength of any statistical analysis. Wrong sources of data or impurities in data can lead to incorrect conclusions. Thus, data should be collected from reliable sources and cleaned before starting any analysis.

A. Analysis Technique

Analysis Technique: One Way ANOVA

Reason: ANOVA is appropriate as I want to compare the continuous dependent variable - carbon dioxide emission means of vehicles based on different types of gasoline (Regular, Medium and Premium gasoline) i.e categorical independent variable.

Advantages of ANOVA: i) It is very good in analyzing the difference in groups means when the independent variables are continuous and dependent is categorical. ii) We need not do individual t-test on each pair.

Disadvantages of ANOVA: One of the assumptions of ANOVA is homoscedasticity (equality of variances), which is sometimes tough to satisfy especially when data set is small or categorical values are too many.

B. Tools Used

To do this analysis, software used is: SAS

Reason and Advantages: 1) SAS has good inbuilt ANOVA methods (GLM) 2) SAS has good statistical graphical output. 2) I am more acquainted with SAS and have university edition of SAS.

Disadvantages of SAS: 1) SAS is primarily licensed [4] 2) Since I am using university edition, it is little slow to handle large data.

C. Calculation and Code

Calculations: We try to check if the mean of carbon dioxide emission for the three gasolines types is statistically equal or not.

Our Null Hypothesis: Mu(C02 Emissions of Premium Gas) = Mu(C02 Emissions of Regular Gas) = Mu(C02 emissions of Midgrade Gas).

SAS code used for calculation of one-way ANOVA is shown in Fig. 6.

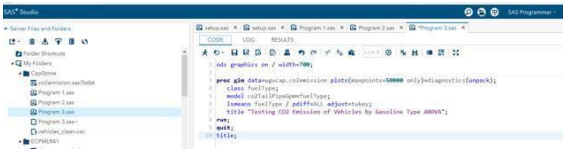

Fig. 6 SAS Code Snippet

D. Outputs and Derivations

Assumption of Equal Variances:

There is no pattern as shown in Fig. 7, thus we can assume there is equal variance.

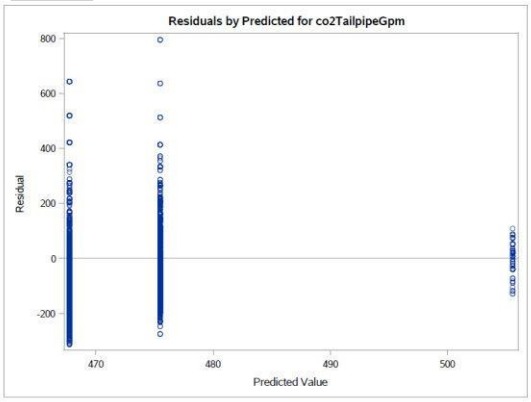

Fig. 7 Output: Residuals

Assumption of Residuals are Normally Distributed:

As shown in Fig. 8, The plot looks approximately bell-shape, thus residuals are normally distributed.

Main ANOVA result for null hypothesis:

The p value is less than 0.05 as shown in Fig. 9, thus we reject the null hypothesis.

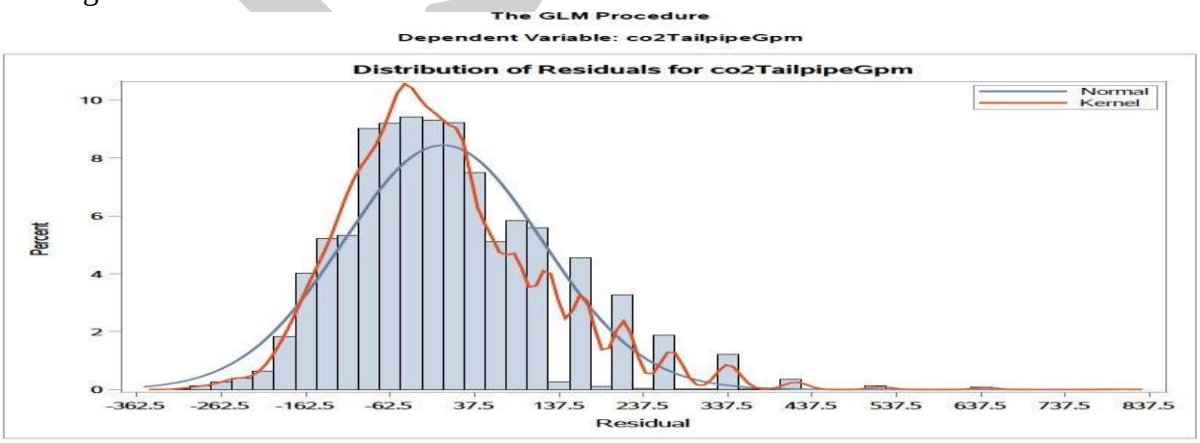

Fig. 8 Distribution of Residuals

Testing CO2 Emission of Vehicles by Gasoline Type ANOVA

The GLM Procedure

Dependent Variable: co2TailpipeGpm

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	585050.9	292525.4	20.97	<.0001
Error	37059	516920425.6	13948.6		
Corrected Total	37061	517505476.4			

R-Square	Coeff Var	Root MSE	co2TailpipeGpm Mean
0.001131	25.12118	118.1041	470.1376

Fig. 9 Output: Null Hypothesis p value

Least Square Means Comparison in between gasoline types:

As shown by comparison graph in 10.b and data in Fig 10.a, the p values for each is less than 0.05 thus we reject null hypothesis of equal variance between each type of gasoline.

Testing CO2 Emission of Vehicles by Gasoline Type ANOVA

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

fuelType	co2TailpipeGpm LSMEAN	LSMEAN Number
Midgrad	505.547083	1
Premium	475.477729	2
Regular	467.733045	3

Least Squares Means for effect fuelType
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: co2TailpipeGpm

i/j	1	2	3
1		0.0347	0.0050
2	0.0347		<.0001
3	0.0050	<.0001	

Fig. 10.a Output: Null Hypothesis (Equal Variance)

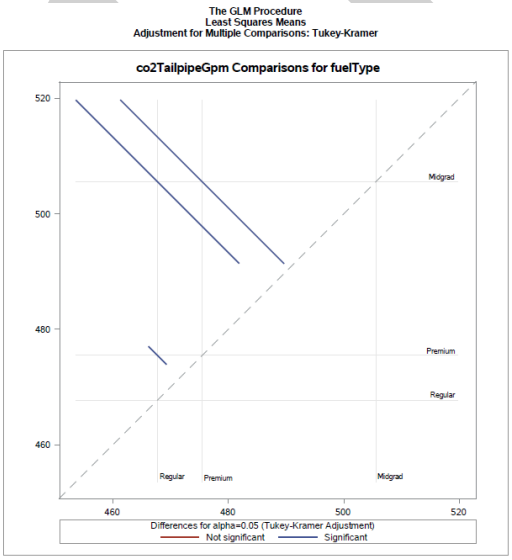

Fig. 10.b Output: Null Hypothesis (Equal Variance)

V. Data Summary and Conclusions

Summary: i) At alpha level of 0.05, the p value of the result is small enough to reject our null hypothesis and accept the alternate hypothesis that there is significant difference in carbon-dioxide emissions of vehicles based on gasoline type used - Regular, Premium and Midgrade. ii) The LS Means p value confirms the same and it shows the biggest difference is between regular and premium as its p value is less than 0.0001. iii) This also shows the powerful engine of SAS and its libraries. With just couple of lines of code and proper dataset, lot of useful data, derivations and graphs are emitted for analysis.

Limitation of Analysis: i) This study does not tell which gasoline type has lower co2 emissions than other ii) The dataset used from all types of vehicles with different makes, models, engine type, age etc thus the unknown error contribution is also high.

Recommended Course of Action: Since this analysis shows, the type of gasoline has significant different on C02 emission, customers should be encouraged to use the fuel with lower C02 emission and help the earth been clean. Also, government can do price study and provide subsidy for the price of gas with lower emission.

Approaches for future study of the data: i) Use pair wise one side t-test on all the gasoline types so we can find out which gasoline is doing better than other. ii) Since electricity cars are also picking up but their percentage is very less, we can analyze to compare the electricity types also like hybrid vs pure electric car. iii) We can study the impact of age of car on C02 emission, does it degrade or remain constant.

ACKNOWLEDGMENT

I did this research for graduate Capstone of my Master of Science in Data Analytics and would like to thank Western Governors University and my course instructor Dr. William Sewell.

REFERENCES

[1] Refresher on Statistical Significance [Online]. Available: https://hbr.org/2016/02/a-refresher-on-statistical-significance
[2] Fuel Economy Data [Online]. Available: https://www.fueleconomy.gov/feg/download.shtml
[3] Fuel Economy Estimates Data 1984 to 2019 Car Models [Online]. Available: https://www.fueleconomy.gov/feg/epadata/vehicles.csv.zip
[4] https://www.sas.com