Fuel Consumption Rating data from Kaggle

CARDIFF SCHOOL OF TECHNOLOGIES

4D Digidol | Data | Dylunio | Dyfodol

Cardiff Metropolitan University
Cardiff School of Technologies
Academic Year: 2022/2023
Term: 2
Module Name: Programming for Data Analysis_S2_22
Module Code: CIS7031
Module Leader: Sengar, Sandeep Singh **Lab Leader: Dhole, Priyesh**
MSc Programme: Data Science
Assignment Title: Fuel Consumption Rating **Link:** https://www.kaggle.com/datasets/rinichristy/2022-fuel-consumption-ratings

Student Name: Amritpal singh	**Student ID: ST20257747**
Feedback:	
Signature:	**Date:**

Introduction:

- https://www.kaggle.com/datasets/rinichristy/2022-fuel-consumption-ratings

The 2022 Fuel Efficiency Dataset provides a detailed and comprehensive overview of fuel economy, CO2 emissions and other vehicle characteristics for various vehicle models sold in Canada in 2022. This dataset is an essential resource for automotive scientists and stakeholders interested in understanding the factors that affect fuel economy. the fuel efficiency and environmental impact of vehicles. Natural Resources Canada, the government agency responsible for promoting sustainability and energy efficiency in Canada, has compiled a range of data.

The dataset contains information on fuel consumption indicators of vehicles, including the amount of fuel consumed per 100 km of city and highway driving and the fuel consumption indicator in the combined cycle. The dataset also includes information on vehicle CO2 emissions, e.g.Volume. the amount of greenhouse gases emitted per kilometer traveled.

Important point:-

Fuel consumption: City and highway fuel consumption ratings are shown in litres per 100 kilometres (L/100 km) - the combined rating (55% city, 45% hwy) is shown in L/100 km and in miles per imperial gallon (mpg)

CO2 emissions: the tailpipe emissions of carbon dioxide (in grams per kilometre) for combined city and highway driving

CO2 rating: the tailpipe emissions of carbon dioxide rated on a scale from 1 (worst) to 10 (best)

Smog rating: the tailpipe emissions of smog-forming pollutants rated on a scale from 1 (worst) to 10 (best)

Data Import: Importing the essential libraries and dataset using the pandas library is the first step in the analysis. The dataset is then investigated to obtain a deeper comprehension of its structure and content.

The following details are extracted:

Record format: There are X rows and Y columns. Column header list: Contains the names of all attributes and functions in the data set. Statistics Summary: Contains descriptive statistics such as mean, standard deviation, minimum, maximum, and quartiles. Displays the data types of each attribute. Also displays a different type of comparison. Data Visualisation: Using the matplotlib

and seaborn libraries, various graphs and plots are generated to obtain visual insights into the dataset.

The visualisations consist of:

Bar Charts: The number of vehicles is represented using bar charts based on attributes such as model year, make, model, engine size, number of cylinders, transmission type and fuel type. For example, a bar chart can be used to show the average CO2 emissions for each make of vehicle. This would allow us to compare the average emissions of different brands and find out which brands emit the most or the least. Likewise, a bar chart can be created to show the average CO2 emissions of a for each model year, showing emission trends over time

Histograms: Histograms show the distribution of city and highway fuel consumption, CO2 emissions and the CO2 emission index.Isotopes are a form of plots that illustrate the distribution of a continuous numeric variable. Histograms can be used to visualize the distribution of fuel consumption (city and highway), CO2 emissions and CO2 rating in the context of a range of fuel economy data. For example, a fuel economy histogram (City) would show the number of vehicles in different fuel economy zones. The x-axis shows the fuel consumption range (in l/100km or mpg) and the y-axis shows the number of vehicles in each range.

Boxplots: Bar charts are a form of data visualization used to show the distribution of a data set.They are useful for distinguishing the median, quartiles, and outliers in a data set. Boxplots can be used to examine the relationships between attributes such as number of cylinders, displacement, fuel type, fuel economy and CO2 emissions in the context of the fuel economy dataset. For example, a boxplot can be used to show the distribution of CO2 emissions for vehicles of different engine sizes. The boxplot would contain a vertical line representing the average CO2 emission value, a rectangle representing the interquartile range (IQR), and whiskers representing the range of CO2 emission values

Analysis and Insights: The following insights and observations are derived from the analysis: In the context of the Fuel consumption dataset, the "Top cars" analysis identifies the ten most prevalent automobile brands and models by count. This analysis can help companies in the automotive industry comprehend consumer preferences and demand by providing insights into the popularity of various vehicle brands and models on the market. To conduct this analysis on the Fuel consumption dataset, we can use the pandas library in Python to group the data by vehicle make and model and then enumerate the occurrences of each make and model. The results can then be sorted in descending order and the top 10 manufacturers and models can be selected. For instance,

1.Let's take a look at the top 10 vehicles from the smart make and model smart dataset.

2. We also have different distributions with different factors (fuel type, cylinder, displacement, gearbox, fuel consumption, make, model)

3. Find a comparison of two and more than two factors to get the best result. The range of vehicles is analysed by engine size and shows the most common engine sizes. To perform the analysis, we can use the fuel consumption dataset to examine the distribution of engine capacities among different cars. We can calculate the frequency or count of each engine size in a data set and then display that data in bar or pie charts. The dataset is analysed to determine the positioning of the vehicle based on the number of cylinders. The structure of automobiles based on transmission variants was examined.

Fuel economy: The fuel consumption distribution (city and highway) is displayed graphically and shows a spectrum of fuel consumption. Studying the distribution of CO2 emissions and CO2 ratios sheds light on the environmental impact of vehicles. The five vehicles with the highest and lowest average fuel consumption are determined. The analysis of the CO2 emissions and CO2 indices shows the environmental impact of the vehicles included in the data set. You can visualize the distribution of CO2 emissions and CO2 indicators using a histogram or boxplot. These displays can help identify the range of CO2 emissions and classifications, the most common emissions and classifications, and outliers. The five vehicles with the highest and lowest average consumption provide information about fuel efficiency. Average fuel economy can be calculated by averaging city and highway fuel economy rates. The five vehicles with the highest average fuel economy can help consumers identify fuel-efficient vehicles, while the five vehicles with the lowest average fuel economy can help consumers identify fuel-efficient vehicles with a large environmental impact.

Linear regression model: A linear regression model relates a dependent variable (target) to one or more independent variables (trait) in a data set. The fuel economy dataset uses a linear regression model to determine fuel economy based on several independent variables including displacement, transmission type, CO2 emissions and smog. The linear regression model estimates the line that best fits the data points, thereby minimizing the sum of squared residuals between predicted and actual values. The goal is to create a model that can accurately determine fuel consumption based on independent variables, which can then be used to increase fuel efficiency and reduce CO2 emissions. The algorithm learns the relationship between the independent variables and the target variable by using the data set to train the model.After training, the model is evaluated for its ability to precisely predict petroleum consumption. This is accomplished by dividing the dataset into training and testing sets and comparing the predicted and actual values. Mean squared error and coefficient of determination are two metrics used to evaluate the efficacy of a model. The mean squared error measures the average squared

difference between the predicted and actual values, whereas the coefficient of determination indicates the proportion of the target variable's variance that can be explained by the independent variables. By determining the optimal parameters and weights for each variable, the linear regression model is able to accurately predict fuel consumption and provide insight into how various factors influence fuel efficiency and CO2 emissions.

Conclusion:

Analysis of the fuel consumption dataset provides valuable insights into various aspects of vehicles such as make, model, engine size, fuel consumption and CO2 emissions. Data patterns and relationships in data are highlighted through data visualization and statistical analysis. The linear regression model provides a tool to estimate fuel consumption based on important characteristics. The results of this study can be used to understand the fuel efficiency and emissions of the and make informed decisions regarding vehicle selection and environmental impact.

Limitations and Future Work:

It is important to realize that this analysis has some limitations. The dataset used is a sample from a specific time period and may not represent the entire vehicle population. Further investigation and analysis could include a larger and more diverse data set covering a wider range of vehicle models and features. Additionally, exploring additional algorithms and machine learning techniques can provide further insights and increase accuracy.

Bibliography

1. Aggarwal, C. C. (2015). Outlier analysis. Springer

2. Tabachnick, B. G., & Fidell, L. S. (2013). Using multivariate statistics. Pearson.

3. Liu, S., & Wang, X. (2016). A novel method for elimination of outliers in the context of partial least squares regression. Chemometrics and Intelligent Laboratory Systems, 152, 44-50.

4. Zuo, Y., Peng, H., & Zhang, Y. (2019). Robust principal component analysis with complex noise: A least squares approach. IEEE Transactions on Signal Processing, 67(8), 2171-2184.
5. Intellipaat , learn code from there and taking classes for different concept
6. Simplilearn a e-platform with one of the finest faculty with amazing content .