Driving Sustainability: A Data-Driven Analysis of CO_2 Emissions Across Vehicle Specification

ANN MARY THOMAS, 22079609, ant0901@my.londonmet.ac.uk

ABSTRACT - **This study investigates the relationship between vehicle specifications—such as fuel type, engine capacity, and weight—and CO2 emissions among vehicles from different manufacturers. Using data from the European Environmental Agency, this project leverages Power BI to identify trends and correlations across attributes that may impact emissions levels. Linear Regression was applied to quantify the impact of these attributes on emissions. The analysis revealed that certain specifications significantly contribute to emissions, offering insights that can support eco-friendly decision-making by consumers and guide manufacturers toward lower-emission designs. This research aims to contribute to sustainable vehicle use and production by highlighting key factors affecting vehicle emissions.**

Keywords—CO2 Emissions, Vehicle Specifications, Power BI

INTRODUCTION

The increasing impact of vehicle emissions on climate change and air quality has made it essential to understand the factors contributing to CO2 emissions in the automotive sector. As vehicles remain a significant source of pollution, this study aims to analyze how specific vehicle characteristics—such as engine capacity, fuel type, and weight—influence CO2 emissions. By examining data from the European Environmental Agency, this project explores emission patterns across various European vehicle manufacturers, providing insights that can help consumers, manufacturers, and policymakers make more environmentally conscious decisions.

This research seeks to identify the vehicle specifications most associated with high and low emissions, helping to highlight models with a lower environmental impact. Focusing on major European vehicle manufacturers, the analysis provides a comparative view of emissions based on key attributes. The findings contribute to a broader understanding of sustainable practices in the automotive industry and support initiatives to reduce emissions through informed choices and innovations in vehicle design.

LITERATURE REVIEW

Numerous studies have examined the impact of vehicle characteristics on CO2 emissions, emphasizing the roles of engine capacity, fuel type, and vehicle weight. Studies found that diesel engines, while often more fuel-efficient, tend to produce higher CO2 emissions compared to petrol engines due to combustion differences [1] . Research by Conor Walsh and Alice Bows (2012) further highlights the influence of engine size, with larger engines typically emitting more CO2 due to their greater fuel requirements, making engine capacity a key determinant in evaluating vehicle emissions [2] . These studies underscore the complex relationship between fuel type and emissions, as alternative fuels such as electricity offer significant reductions in CO2 output, though infrastructure and production impact must be considered.

In addition to fuel type and engine capacity, vehicle weight has also been identified as a crucial factor in emissions levels. Heavier vehicles, such as SUVs and trucks, often exhibit higher CO2 emissions compared to lighter vehicles, as noted by Georgios and Biagio (2017) [3]. Their study observed that weight not only impacts emissions directly but also affects fuel consumption, further linking it to overall emissions output. However, while existing literature has explored individual factors in isolation, few studies have combined multiple vehicle characteristics to provide a holistic understanding of emissions patterns across different manufacturers and vehicle categories.

This study builds on previous research by offering a multi-dimensional analysis that includes fuel type, engine capacity, and weight to assess their combined impact on emissions. By examining data from various European manufacturers, this project contributes to the growing body of literature focused on reducing emissions in the automotive sector and provides valuable insights for both industry professionals and consumers seeking environmentally friendly transportation options.

METHODOLOGY

This section provides an overview of the data and outlines the six stages of the Data Analytical Life Cycle (DALC) as applied to this project.

A. Discover

The problem addressed in this analysis revolves around understanding the factors influencing CO_2 emissions in the automotive industry and identifying actionable insights to enhance sustainability efforts. The primary question guiding this analysis was: What are the key variables that drive CO_2 emissions, and how can manufacturers leverage this information to reduce their environmental impact? The relevance of this study lies in its potential to support sustainable decision-making by uncovering relationships between CO_2 emissions, engine capacity, vehicle mass, fuel type, and the use of emission reduction technologies. The results of this study were intended to inform both policymakers and manufacturers.

B. Data Preparation

The dataset used for this analysis underwent targeted preparation steps tailored specifically to the variables influencing CO_2 emissions. To begin with, columns unrelated to the focus of the study were removed, leaving only key variables such as mass, engine capacity, fuel type, and emission reduction technologies. For better clarity and ease of interpretation, column names were standardized, with abbreviations expanded—for instance, "CO_2_em" was renamed "CO_2_emission." During the cleaning process, rows with null or missing values in the mass column were eliminated to avoid biases or inaccuracies in the analysis caused by incomplete data entries. These null values likely arose from inconsistencies or gaps during the data collection phase.

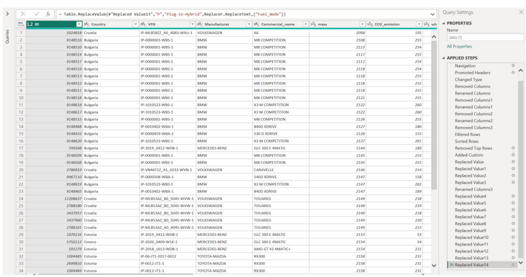

Figure 1 Data Cleaning using PowerBi

A detailed exploration of the dataset was conducted to uncover relationships and ensure data consistency. Scatter plots were specifically employed to explore potential linear relationships between CO_2 emissions and variables like mass and engine capacity, revealing critical insights for the modeling phase. Summary statistics provided further understanding of the data's distribution, helping to confirm the integrity of the remaining dataset. While preparing the data, categorical variables such as fuel type were standardized to maintain uniformity, ensuring consistent classification across the dataset. This tailored data preparation process ensured that the dataset was optimized for accurate analysis, setting a robust foundation for building the linear regression model and developing effective Power BI visualizations.

C. Plan Model

For the analysis, a linear regression model was chosen due to its simplicity, interpretability, and effectiveness in quantifying relationships between independent variables and CO_2 emissions as the dependent variable. This model was deemed appropriate for the dataset, as it allowed for examining the extent to which variables such as vehicle mass, engine capacity, and the adoption of emission reduction technologies influenced CO_2 emissions. The linear regression approach also enabled the identification of the strength and direction of these relationships through metrics like correlation coefficients and R-squared values.

D. Build Model

The linear regression model was implemented in Power BI using the Quick Measure feature, focusing on understanding the relationship between CO_2 emissions and two independent variables: mass and engine capacity. CO_2 emissions were designated as the dependent variable, while mass and engine capacity were selected as the primary predictors due to their significant influence on vehicle emissions. These variables were carefully chosen to reflect key physical and performance characteristics of vehicles that directly impact emission levels.

The regression model in Power BI was used to analyze the relative contribution of each independent variable to CO_2 emissions. Among the two variables, engine capacity emerged as the more accurate predictor, highlighting its stronger relationship with CO_2 emissions compared to mass. The model leveraged Power BI's capabilities to establish linear trends and assess the predictive power of these variables, setting the foundation for actionable insights into how vehicle attributes affect emissions. By focusing on these two predictors, the analysis provided a clear and concise understanding of their individual impact on emission levels.

E. Communicate

Communicating the results effectively was a critical part of this project, and the Power BI dashboard was designed with this purpose in mind. The dashboard was crafted to provide a clear, interactive, and visually appealing representation of the findings to facilitate data-driven decision-making. It included a variety of charts and metrics to cater to different types of stakeholders. For example, a gauge chart was used to highlight the average CO_2 emissions across the dataset, providing a quick, at-a-glance summary of the central metric. This was complemented by bar charts that showcased the contributions of different manufacturers to electric vehicle ranges, highlighting their strides toward sustainability.

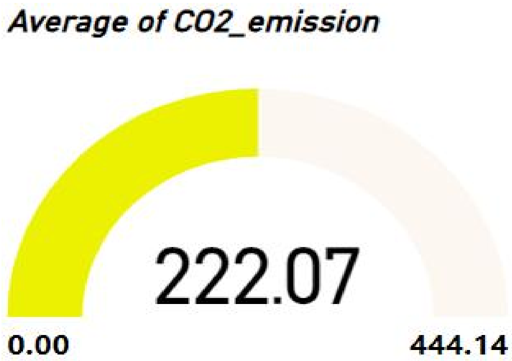

Figure 2 Gauge Chart of Avg_CO2 emission

Scatter plots were employed to illustrate the relationships between CO_2 emissions and key factors such as mass and engine capacity. These plots also incorporated trend lines to visually depict the direction and strength of the relationships. Cards displaying the correlation coefficients and R-squared values were included to summarize the strength of these relationships numerically. Additionally, line charts broke down CO_2 emissions by fuel type, providing a comparison of the environmental impacts of different fuel sources.

Each visual was accompanied by interactive slicers, such as filters for country or manufacturer, to enable stakeholders to drill down into the data based on specific contexts or areas of interest. The visual elements were laid out on a background with a sustainability theme to reinforce the goal of reducing CO_2 emissions. This combination of interactive features, varied visualizations, and thematic design ensured that the dashboard could effectively communicate complex findings in an intuitive manner, empowering users to make informed decisions.

F. Measure Effectiveness / Apply Live

Ethical considerations were carefully addressed throughout the data analysis process. Data privacy was maintained by ensuring no sensitive or personally identifiable information was included in the dataset. The analysis process was transparent, with all data sources and methods clearly documented to avoid misrepresentation of results. Furthermore, biases were minimized by relying solely on statistically significant findings to draw conclusions. The methodology adopted in this project provided actionable insights, demonstrating how manufacturers can optimize vehicle designs to reduce CO_2 emissions. By identifying the relationships between key factors, this study contributes to the development of data-driven strategies for a more sustainable automotive industry.

RESULT AND DISCUSSION

The analysis revealed that engine capacity is the most influential factor affecting vehicle CO_2 emissions. This conclusion is supported by a strong correlation coefficient of 0.71 between engine capacity and CO_2 emissions. This relationship is clearly depicted in the scatter plot in Figure 1, where data points closely follow the trend line, confirming engine capacity as a key determinant of emissions. Larger engine vehicles are shown to emit significantly higher levels of CO_2, emphasizing the need for manufacturers to optimize engine technologies to reduce emissions.

In addition to engine capacity, mass showed a moderate positive correlation with CO_2 emissions, with a coefficient of 0.58. The variability in emissions for vehicles of similar mass, however, suggests that other factors, such as engine technology or fuel type, may play a more significant role in determining overall emission levels. This is illustrated in the scatter plot in Figure 2, where the distribution of data points indicates that mass alone does not fully explain the variation in emissions.

Figure 3 Scatter plot showing the relationship between engine capacity and CO_2 emissions.

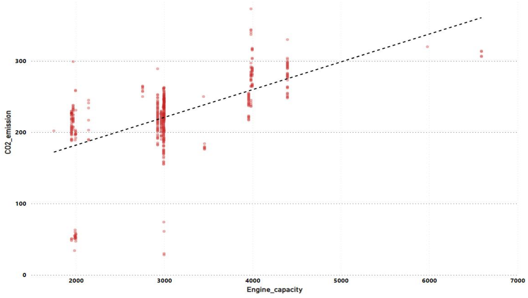

Figure 4 Scatter plot showing the relationship between mass and CO_2 emissions.

In terms of manufacturer performance, the data highlights that brands such as BMW and Mercedes-Benz have relatively lower average CO_2 emissions compared to other manufacturers. These results, as seen in the dashboard, likely reflect their focus on implementing cleaner engine technologies and improving fuel efficiency.

The dashboard further provides insights into the influence of fuel type on emissions. Vehicles powered by electric or hybrid technologies, such as diesel-electric or petrol-electric combinations, demonstrated the lowest average emissions, as visualized in the respective charts. Fully electric vehicles, as anticipated, showed negligible or zero CO_2 emissions, underscoring the importance of transitioning to electric mobility to achieve sustainability goals.

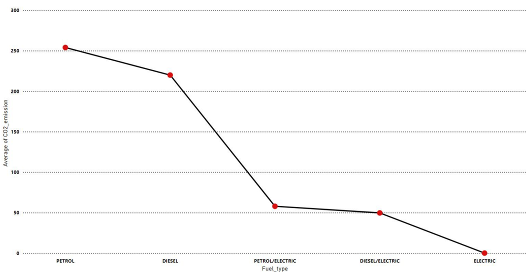

Figure 5 Avg CO2 emission by Fuel Type

From a geographic perspective, countries such as Estonia and Greece exhibited comparatively lower average CO_2 emissions. This observation may be attributed to their adoption of smaller engine vehicles or stricter emission regulations. Visualizations in the dashboard, including bar charts and scatter plots, support these findings by presenting trends across manufacturers, fuel types, and regions. These visuals allow for clear comparisons, effectively highlighting best practices in reducing vehicle emissions and guiding manufacturers and policymakers in implementing targeted emission reduction strategies.

PEER REVIEW

The feedback I received from my peer offered valuable insights that helped refine my analysis. One of the key suggestions was to extend the linear regression model to other relevant datasets to explore additional factors influencing CO_2 emissions. I realized that broadening the scope of the analysis could provide a more complete picture and enhance the model's predictive capabilities. Additionally, my peer emphasized the need for a more detailed explanation of how each feature in the dataset contributes to the broader issue of CO_2 emissions. This encouraged me to connect the individual variables more clearly to the overall sustainability context.

Furthermore, my peer recommended expanding the literature review with more references to research on vehicle emissions, which broadened my understanding of the topic and helped identify gaps in my analysis. They also suggested incorporating ethical considerations regarding data handling, which made me reflect on the responsibility of data scientists in ensuring privacy and the ethical use of data. This feedback has been instrumental in shaping my approach to data analysis, making it more comprehensive, insightful, and ethically grounded, and will guide me in refining my final report.

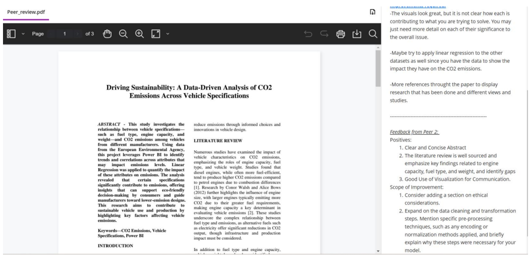

Figure 6 peer review

CONCLUSION AND RECOMMENDATION

The analysis provided valuable insights into the factors influencing CO_2 emissions from vehicles, with a particular focus on engine capacity and mass. The linear regression model demonstrated that engine capacity is the most significant factor, with a strong correlation coefficient of 0.71, indicating that vehicles with larger engines generally emit higher levels of CO_2. The scatter plot further reinforced this relationship, showing a clear upward trend in emissions as engine capacity increased. While mass also exhibited a moderate positive correlation (0.58) with CO_2 emissions, its contribution to the overall emissions was less pronounced, suggesting that other factors, such as engine technology or fuel type, may play a more critical role in influencing emissions.

The findings highlight the importance of considering multiple vehicle characteristics when evaluating emissions, with engine capacity emerging as a key predictor. For future work, expanding the analysis to include additional variables such as fuel type, vehicle type, or alternative fuel sources could provide a more comprehensive understanding of CO_2 emissions. Incorporating measures of fuel efficiency, such as miles per gallon or electric range, would further enhance the model's ability to identify vehicles that are both environmentally friendly and fuel-efficient. In terms of model improvement, refining the linear regression approach by exploring non-linear relationships or incorporating more sophisticated models could increase the precision of the predictions and offer better clarity in identifying key drivers of CO_2 emissions.

References

[1] L. S. John, E. Russell and H. Robert, "CO2 Emission Benefit of Diesel (versus Gasoline) Powered Vehicles," Environmental Science & Technology, vol. 38, no. 12, pp. 3217-3223, 2004.

[2] W. Conor and B. Alice, "Size matters: Exploring the importance of vessel characteristics to inform estimates of shipping emissions," Applied Energy, vol. 98, pp. 128-137, 2012.

[3] F. Georgios and C. Biagio, "Fuel consumption and CO2 emissions from passenger cars in Europe – Laboratory versus real-world emissions," Progress in Energy and Combustion Science, vol. 60, pp. 97-131, 2017.

[4] J. Gareth, W. Daniela, H. Trevor, T. Robert and T. Jonathan, Linear Regression, Springer Nature Link, 2023.

POWERBI LINK : https://londonmet-my.sharepoint.com/:u:/g/personal/ant0901_my_londonmet_ac_uk/EYNXJErvRCVFtbFmlVtcF1EBP0lxdyk8ahPNUPQfQ1rIgA?e=jsDZXy

APPENDIX

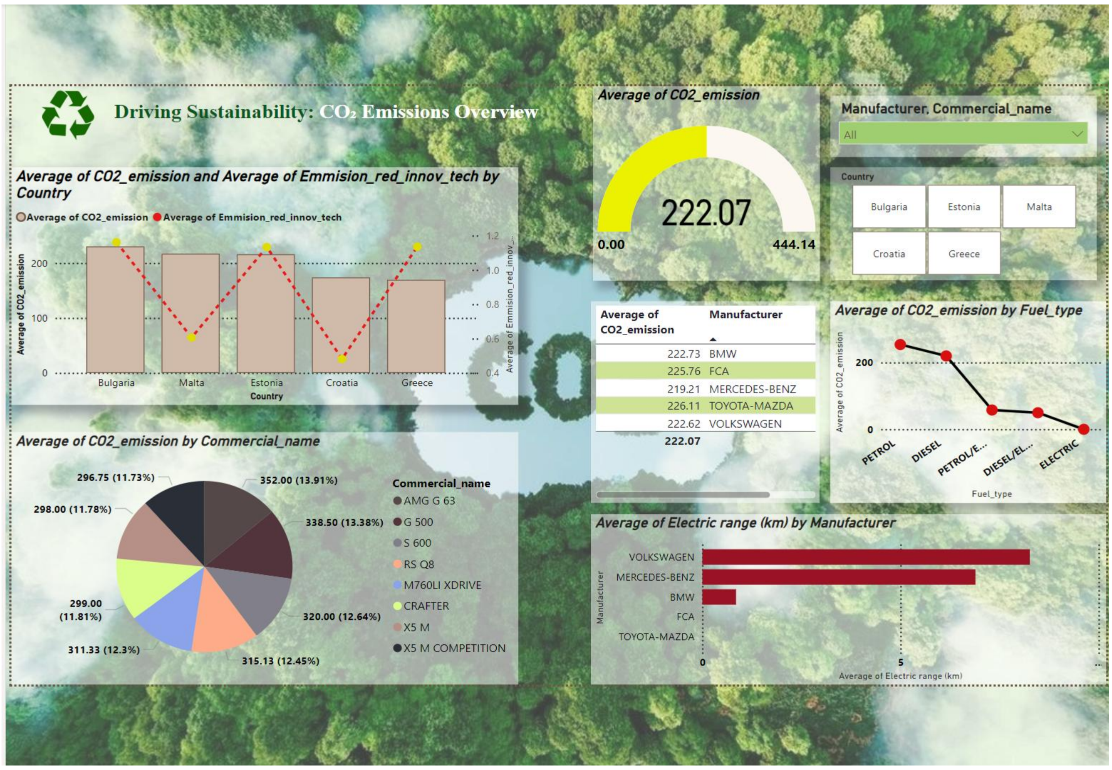

Figure 7 PowerBi dashboard 1

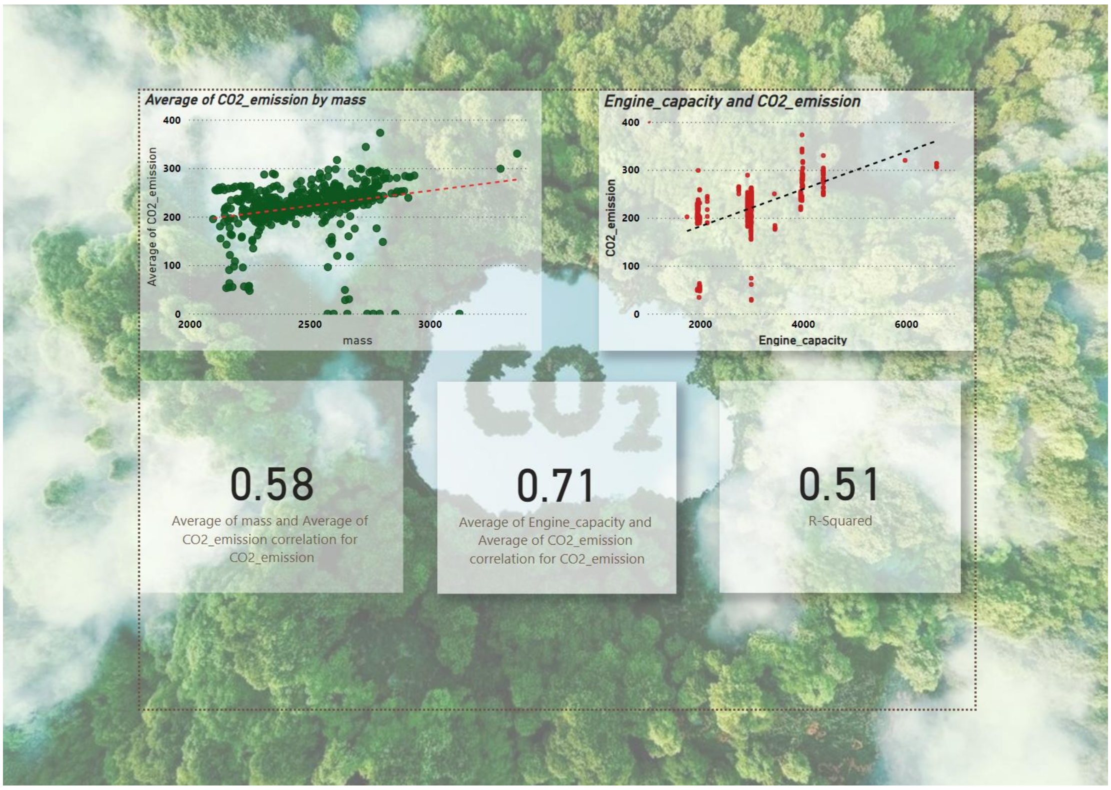

Figure 8 PowerBi dashboard 2