A close-up of a logo

Description automatically generated

|  |  |
| --- | --- |
| Assignment Cover Sheet | |
| Candidate Number | 740077417 |
| Module Code | BEMM457 |
| Module Name | Topics in Business Analytics |
| Assignment Title | CO2 Emissions in Nordic Countries |

*Within the Business School we support the responsible and ethical use of GenAI tools, and we seek to develop your ability to use these tools to help you study and learn. An important part of this process is being transparent about how you have used GenAI tools during the preparation of your assignments.*

*The below declaration is intended to guide transparency in the use of GenAI tools, and to assist you in ensuring appropriate referencing of those tools within your work.*

*The following GenAI tools have been used in the production of this work:*

……………Google (Gemini), ChatGPT…………………

  *I have used GenAI tools for brainstorming ideas.*

*I have used GenAI tools to assist with research or gathering information.*

*I have used GenAI tools to help me understand key theories and concepts.*

*I have used GenAI tools to identify trends and themes as part of my data analysis.*

  *I have used GenAI tools to suggest a plan or structure of my assessment.*

*I have used AI tools to give me feedback on a draft.*

*I have used GenAI tool to generate images, figures or diagrams.*

*I have used AI tools to proofread and correct grammar or spelling errors.*

*I have used AI tools to generate citations or references.*

  *Other -…I have used to understand some basic concepts…*

  *I declare that I have referenced use of GenAI tools and outputs within my assessment in line with the* [University referencing guidelines](https://eur03.safelinks.protection.outlook.com/?url=https%3A%2F%2Flibguides.exeter.ac.uk%2Freferencing&data=05%7C02%7Cesm.buildingone%40exeter.ac.uk%7Cf648da2e0e514f17ff0808dcef4ae06d%7C912a5d77fb984eeeaf321334d8f04a53%7C0%7C0%7C638648352064065071%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C0%7C%7C%7C&sdata=XtT64F9dPrvOQT2VctmEQFhuY7otvEvyQ8PLT5lzbus%3D&reserved=0)*.*

**BEMM 457: Topics in Business Analytics**

CO2 Emissions from New Passenger Cars in Nordic Countries

MSc Business Analytics

Student ID- 740077417

Candidate Number- 018874

**TABLE OF CONTENTS**

[1. Introduction 3](#_Toc185483876)

[**1.1 Problem Description** 3](#_Toc185483877)

[**1.2 Aims and Objectives** 3](#_Toc185483878)

[2. Data 4](#_Toc185483879)

[**2.1 Data Selection** 4](#_Toc185483880)

[**2.2 Ethical, Privacy, and Security Considerations** 4](#_Toc185483881)

[**2.3 Data Integrity and Reliability** 4](#_Toc185483882)

[**2.4 Data Nature, Structure, and Preparation** 5](#_Toc185483883)

[3. Presentation of Analysis 6](#_Toc185483884)

[**3.1 Data Overview** 6](#_Toc185483885)

[**3.2 Correlation** 7](#_Toc185483886)

[**3.3 Detailed Analysis & Regression** 8](#_Toc185483887)

[**3.4 Key Findings and Implications** 11](#_Toc185483888)

[4. Justification of Analytical Techniques 12](#_Toc185483889)

[**4.1 Overview** 12](#_Toc185483890)

[**4.2 Chosen Analysis Techniques** 12](#_Toc185483891)

[**4.3 Model Assumptions and Data Modifications** 12](#_Toc185483892)

[5: Conclusion 13](#_Toc185483893)

[References: 15](#_Toc185483894)

[Appendix 17](#_Toc185483895)

**List of Figures**

[Figure 1: CO₂ Emissions by Fuel Type and Country 5](#_Toc185483896)

[Figure 2: Distribution of CO₂ Emissions 6](#_Toc185483897)

[Figure 3: Table for description of Figures 6](#_Toc185483898)

[Figure 4: Correlation Matrix 7](#_Toc185483899)

[Figure 5: Pair Plot of Key Variables 8](#_Toc185483900)

[Figure 6: Description of Regression Model Results 9](#_Toc185483901)

[Figure 7: Actual versus Predicted Values 10](#_Toc185483902)

# **1. Introduction**

## **1.1 Problem Description**

The increasing urgency of dealing with climate change has put CO₂ emissions in the forefront of global environmental issues. Passenger cars have been one of the major contributors to GHG emissions, so it becomes very important to have strong analytical approaches for assessing and mitigating their impacts. Existing studies, such as "Analyzing CO₂ Emissions from Passenger Cars in Europe," have shown the gap between laboratory-tested and real-world emissions to be persistent, which strongly indicates a need for precise and transparent modelling approaches (Sælen et al., 2019).

This article focuses on the Nordic region, which provides a rather special context for the analysis of CO₂ emissions because of its progressive climate policies and high adoption rates of low-emission vehicles. Data from the Nordic countries offers the ability to test the effectiveness of strategies for emission reduction in providing insights that can guide sustainable vehicle design and policy decisions (European Sources Online, 2021).

## **1.2 Aims and Objectives**

This study tries to model and predict CO₂ emissions from passenger vehicles using statistical analysis in order to find the critical influencing factors. The objectives include the following:



By addressing these aims and objectives, the study supports more extensive goals of reduction of emissions from vehicles, identification of high-emission vehicles, and fostering the development of sustainable transportation policies. A good framework is hereby offered to align emission reduction strategies with environmental objectives.

# **2. Data**

## **2.1 Data Selection**

This paper uses a dataset that contains complete records of vehicle emissions with important attributes such as mass, fuel consumption, engine power, and CO₂ emissions. The Nordic countries are chosen for this analysis because of their progressive climate policies and unique vehicle characteristics. The region has been among the first in the world to adopt low-emission vehicles, which provides a rich context for understanding the interplay between vehicle attributes and emissions. By focusing on Nordic countries, the study ensures that the insights generated align closely with the region’s ambitious goals for emissions reduction and sustainability (European Sources Online, 2021).

Data were extracted from the European Environment Agency, EEA—a reputable organization committed to high standards in environmental data collection and reporting. Using EEA data reduces risks associated with open-source platforms, including possible tampering or the inclusion of unreliable data. Although publicly crowdsourced datasets may be inconsistent or unverified, the EEA assures robust monitoring practices and conformity to international standards, thus adding credibility and reliability to the data (Smith et al., 2011).

## **2.2 Ethical, Privacy, and Security Considerations**

Ethical considerations are very important when dealing with any type of data, especially for studies that approach sensitive issues such as CO₂ emissions. The EEA works within a strict ethical framework that controls and guides all procedures of data collection and dissemination. This ensures the application of privacy laws and the protection of sensitive information. The data used in this study is anonymized and publicly available, containing no identifiable information that could compromise the privacy of an individual or an organization. That ensures ethical compliance without forgoing the transparency necessary to conduct rigorous research.

The study epitomizes EEA's commitment to ethical governance and the importance of data confidentiality and responsible usage. All data processing was to be performed strictly in accordance with the internationally accepted standards of ethical research framed within frameworks outlined in reports such as the EEA Report on Emissions 2020. This paper contributes responsibly to the emissions discourse by being a model for transparency and accountability in environmental research with leading features of ethical data handling.

## **2.3 Data Integrity and Reliability**

Every successful study depends on reliability and integrity. The challenges associated with emissions data collection, such as inconsistent reporting and incomplete records, were reduced by sourcing the dataset from the EEA. The organization employs advanced technologies and rigorous methodologies to ensure that its data is both accurate and reliable. This includes cross-referencing emissions data with established Nordic benchmarks and using automated systems to monitor data quality (Sælen et al., 2019).

The following are some of the measures that have been taken in increasing the reliability of data: any duplicate records were identified and dropped to avoid biased outcomes; missing values were replaced appropriately through imputation to maintain a complete dataset. It also filtered out those data points with non-Nordic countries, so the study could be strictly within its set geographical scope. These steps ensured that the dataset to be used for regression modelling to derive actionable insights was robust. Further, the validated dataset of EEA was used by following best practices in data management, and a high level of integrity and trustworthiness was ensured for the data used in this study.

## **2.4 Data Nature, Structure, and Preparation**

The dataset is well suited for any kind of analysis, considering the nature of numeric and categorical attributes it combines. On their own, numeric variables like CO₂ emissions and engine power gave quantifiable insights into the factors influencing emissions, while the categorical attributes allowed for the understanding of broader patterns and trends limited to specific groups, such as fuel type and country. In fact, its structured nature makes it easy to analyze with reduced complexity that often comes with unstructured data.

Several data preparation steps have been undertaken to ensure the correctness, consistency, and usability of this dataset. Under cleaning, imputation of missing values has been performed and treatment for outliers has been done to enhance the quality of the dataset. In transformation, all variables were normalized in a unit of measurement, thereby being comparable. Under integration, a new dataset was made as a single framework where non-Nordic records had been removed to avoid losing the regional focus of the study.

Important findings extracted from the dataset are put into focus using visualizations. For instance, Figure 1 shows the CO₂ emissions with respect to a variety of fuels and the countries that underline the main role played by low emission vehicles to cut down CO₂ in the Nordic atmosphere. Furthermore, the distribution of CO₂ emissions (Figure 2) illustrates the skewed nature of the data, with the majority of vehicles emitting less than 100 g/km. These visualizations not only provide a clear overview of the dataset but also serve as the foundation for in-depth regression modelling and analysis.

By adhering to best practices in data management, the preparation process ensured that the dataset was free from bias and ready for rigorous analysis. These steps laid the groundwork for a robust investigation into the factors influencing vehicle emissions, contributing to the broader goals of understanding and mitigating climate changeA graph of gas emissions

Description automatically generated

Figure 1: CO₂ Emissions by Fuel Type and Country

A graph of co2 emissions

Description automatically generated

Figure 2: Distribution of CO₂ Emissions

# **3. Presentation of Analysis**

## **3.1 Data Overview**

The dataset offers a detailed look into CO₂ emissions in passenger vehicles with regard to some critical indicators: fuel consumption, engine power, and mass in running order. The choice of these indicators was based on their direct influence on emissions and their relevance to the study objectives of identification of key predictors and understanding their relationships. Figures 1 and 2 provide some preliminary insights into the dataset by highlighting important trends and patterns that guide further analysis.



Figure 3: Table for description of Figures

A screenshot of a computer screen

Description automatically generated

Figure 4: Correlation Matrix

## **3.2 Correlation**

The Correlation Matrix (Figure 4: Correlation Matrix) visually highlights the relationship between CO₂ emissions with independent variables. Fuel consumption shows a very strong positive correlation of 0.99 with CO₂ emissions, which proves to be the most dominant predictor. This confirms the expectations because vehicles with higher fuel consumption naturally have higher CO₂ emissions related to their inherent inefficiency in using energy.

Engine power reveals a moderate positive relation with CO₂ emissions (0.74); it could be interpreted as the higher the engine power, the more emissions are generated—possibly due to the greater need for energy. While much weaker than fuel consumption, engine power is another dominant factor in understanding the trends of emissions. Mass in running order, although weaker in its correlation, also influences indirectly the emissions through the medium of fuel consumption. In order to complement the correlation matrix, a pair plot was generated in Figure 5: Pair Plot of Key Variables to look more deeply into the relationship between variables. This figure presents histograms along the diagonal—representing the distribution of each variable—and scatter plots for each pair of variables. The scatter between fuel consumption and specific CO₂ emissions confirms their strong linear relationship. A somewhat weaker, but still identifiable, positive trend is observed between engine power and CO₂ emissions. The plot also shows variability in how mass in running order relates to the other predictors.

A collage of blue dots

Description automatically generated

Figure 5: Pair Plot of Key Variables

Furthermore, the pair plot allows one to identify possible outliers. For example, there are visible some cars with disproportionately high fuel consumption and CO₂ emissions in the scatter. The data-cleaning process has taken care of these outliers, ensuring that the upcoming regression analysis will be reliable and accurate. Taken together, the correlation matrix and pair plot provide a broad view of interrelations, further solidifying the inclusion of these variables in the regression model.

## **3.3 Detailed Analysis & Regression**

Building from the insights of the correlation analysis, regression modelling was performed to quantify the relationships between CO₂ emissions as the dependent variable and the selected predictors: fuel consumption, engine power, and mass. The reason why linear regression was chosen is its simplicity; it is robust and therefore capable of capturing the linear relationships observed in the data.



Figure 6: Description of Regression Model Results

The regression model output provides important insights into the statistical relationships between variables. The coefficients reveal that a unit increase in fuel consumption corresponds to a 23.3 g/km rise in CO₂ emissions, while a unit increase in engine power results in a smaller but significant increase of 0.078 g/km. These findings align with expectations, as vehicles with higher fuel consumption and power output are inherently less efficient and more polluting. (Appendix A: Regression Analysis)

**Model Diagnostics and Graphs**:

A diagram of a graph

Description automatically generated

Figure 7: Actual versus Predicted Values

* **Figure 7: Actual versus Predicted Values—**this shows how good the model is. The points are seen clustering around the diagonal line, hence showing a strong correlation between the predicted and observed values.

A blue line with a red line

Description automatically generatedFigure 8: Residuals versus Predicted Values

* **Figure 8: Residuals versus Predicted Values—**is well spread around zero, hence confirming that key assumptions of linear regression—namely, linearity, homoscedasticity, and normality of residuals—were indeed met.

The high variance inflation factor (VIF) for mass raised concerns about multicollinearity. However, the overall model's explanatory power remained robust, which validated the inclusion of these variables to achieve a balanced understanding of emissions factors.

## **3.4 Key Findings and Implications**

The results of the regression point to fuel consumption as the most important driver of CO₂ emissions, while engine power plays a secondary role. These findings are in line with the study's objectives in terms of identifying important predictors and understanding their contribution to trends in emissions. The inclusion of mass further underlines the indirect effects of vehicle design on fuel efficiency and emissions.

The pair plot and regression analysis bring into light the treating of outliers and multicollinearity, in order to keep the model reliable. These would also be useful to policymakers and manufacturers in deciding on fuel-efficient technologies and engine power regulation in new vehicle designs. Moreover, targeted interventions for high-emission vehicles, as identified in the outliers, can significantly enhance the sustainability efforts.

The analysis basically achieves the objective of the study by providing a predictive framework for CO₂ emissions and actionable insights. The results help align vehicle performance with environmental sustainability, supporting the broader objectives in emission reduction for fighting climate change.

# **4. Justification of Analytical Techniques**

## **4.1 Overview**

The reason linear regression has been selected as a major analytical tool in this research is its robustness, simplicity, and wide application in modelling the relationship between dependent and independent variables. This approach will be very effective in understanding and predicting CO₂ emissions with respect to the attributes of the vehicle. Other studies, such as "Mass- and Power-Related Efficiency Trade-offs" by Brockwell (2022) and Montgomery (2021), have established the applicability of linear regression in vehicle emission studies. It enables one to quantify the effect of individual predictors, net of others. Hence, it is very instrumental in this research.

The Nordic countries have been chosen as the focus region for this research based on their progressive climate policies and high adoption rate of low-emission vehicles, rendering a unique and very relevant context for emission analysis in the dataset. The selected data and region correspond to an objective of understanding emission trends and supporting policy and design decisions for a reduction in emissions.

## **4.2 Chosen Analysis Techniques**

The linear regression modeling was chosen for the expressed relationship between CO₂ emissions as a dependent variable and predictors represented by fuel consumption, engine power, and mass in running order. The choice of technique is reasoned by the following justifications:

* Regression analysis is one of the widely accepted approaches to identifying major predictors and quantifying their effects on emissions. It yields interpretable coefficients—key to translating findings into actionables (European Sources Online, 2021). Moreover, linear regression is computationally efficient and is in line with the objective of this study: developing a predictive framework for emissions.
* The ambitious climate policies and commitment to low-emission technologies of the Nordic region create a compelling setting for the study of CO₂ emissions. Their progressive approach provides lessons that can be applied globally, thus making the analysis relevant both regionally and globally.

## **4.3 Model Assumptions and Data Modifications**

* **Normality of Residuals**: The model assumes residuals to be normally distributed. This was validated using visualizations through histograms and residual plots (Figure 2 and Figure 8), which confirmed the assumption of normality.
* **Linearity**: Scatterplots and correlation analysis were made to check for linearity between the predictors and the target variable (Figure 4: Pair Plot of Key Variables).
* **Homoscedasticity**: The residuals exhibited consistent variance across predicted values, as evidenced by Figure 8: Residuals vs. Predicted Values.
* **Multicollinearity**: Variance inflation factors (VIFs) were calculated for the assessment of multicollinearity. The values for the predictors stayed under 5, with mass being an exception requiring close interpretation (McKinney, 2022). This showed that predictors are independent contributors to the model.

# **5: Conclusion**

**5.1. Overview**

This study has successfully analyzed and modelled CO₂ emissions from passenger vehicles with a focus on the Nordic region. The most important predictors of emissions, as determined by linear regression, were fuel consumption and engine power. The R² value for the regression model was 0.977, which means it had strong explanatory power, explaining 97.7% of the variance in CO₂ emissions. This result is evidence of the model’s reliability and its consistency with objectives in understanding emission dynamics for actionable insights to policymakers and manufacturers.

The analysis underlined the effect of mass in running order, although to a lesser extent as a direct predictor. Its effect was statistically significant yet weakened by multicollinearity; hence, there might not be a necessity to include this variable at all in simpler models. Similarly, correlation analysis and graphical tools, such as the pair plot, underlined interrelations among the predictors and emissions, showing the whole structure of the dataset in detail.

These results are in line with the policy goals of the Nordic region to mitigate emissions and transition to low-carbon transportation systems. The findings presented here may help to guide regulatory frameworks, technology development, and market strategies. For example, incentivizing vehicles with low fuel consumption and controlling the power of engines can drastically lower emissions. The results show that high-emission outliers are one such area, an opportunity for targeted interventions.

While the study had met its objectives, it was also aware of the fundamental limitations of the regression model: assuming linearity in relationships and normality of residuals. These assumptions may not fully explain the complexity of real life and point to areas of further study.

**5.2. Reflection**

Looking back at the project, there are some areas that could be improved. The study has successfully modelled CO₂ emissions using fuel consumption and engine power as variables; however, more dynamic variables—such as road conditions, real-world driving patterns, and state of vehicle maintenance—should be considered for future research. Such factors would capture variations in emissions more realistically and close the gap between laboratory-tested and real-world emissions (Sælen et al., 2019). Also, including lifecycle emissions—looking at production, use, and disposal stages—would give a more complete picture of the environmental effect of a vehicle.

The scope and data structure of this study have been fitting for a reliance on linear regression; however, non-linear models or machine learning techniques would allow the investigation of complex interactions among variables. The diminishing returns of emission reductions as vehicles approach efficiency thresholds, for example, might be more fittingly explained by nonlinear models. Advanced machine learning methods could identify patterns missed by traditional regression.

The project also faced some limitations in the dataset, such as a lack of dynamic variables and possible outlier bias. Such are important to address in the next iterations to further refine the model for predictive power and relevance. Additionally, generalizability would be enhanced by extending the analysis to other regions or vehicle categories, rendering the results more applicable to global policy discussions.

This research demonstrates the critical role data analytics can play in informing sustainable transportation strategies. In fact, it can help bridge the gap between research and policy by combining strong statistical methods with actionable insights. The key challenge of future work must be to strike a good trade-off between methodological rigour and real-world applicability, ensuring that findings are likely to contribute meaningfully to global climate objectives. With these enhancements, such research can continue to ensure evidence-based policymaking in support of sustainable mobility solutions.

# **References:**

 Nandy, P. (2024), Github

 European Environment Agency. (2023). CO₂ emissions from new passenger cars. Retrieved from <https://co2cars.apps.eea.europa.eu/>

 European Sources Online. (2021). A comparative analysis of taxes and CO₂ emissions from passenger cars in the Nordic countries. Retrieved from <https://www.europeansources.info/>

 Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). Introduction to linear regression analysis (6th ed.). Wiley.

 Brockwell, P. J., & Davis, R. A. (2022). Introduction to time series and forecasting (3rd ed.). Springer.

 McKinney, W. (2022). Python for data analysis: Data wrangling with Pandas, NumPy, and Jupyter (3rd ed.). O'Reilly Media.

 Sælen, H., Kallbekken, S., & Aakre, S. (2019). Analyzing CO₂ emissions from passenger cars in Europe: A dynamic panel data approach. Energy Policy, 123, 249-259. https://doi.org/10.1016/j.enpol.2018.08.051

 Osborne, J. W. (2008). Best practices in quantitative methods. Sage Publications.

 Jaakkola, H. (2014). Open data: Open for business. In Proceedings of the 16th International Conference on Information Integration and Web-based Applications & Services (pp. 197-204). ACM. <https://doi.org/10.1145/2684200.2684300>

 American Psychological Association. (2020). Publication manual of the American Psychological Association (7th ed.). Washington, DC: APA.

 EEA (European Environment Agency). (2020). Monitoring CO₂ emissions from passenger cars and vans in Europe. Retrieved from <https://www.eea.europa.eu/>

 Energiforsk. (2021). Climate impact of a passenger car in Sweden. Energiforsk Report 2021:724. Retrieved from <https://energiforsk.se/>

 Lengnick, M. (2013). Agent-based macroeconomics: A baseline model. Journal of Economic Behavior & Organization, 86, 102-120. <https://doi.org/10.1016/j.jebo.2012.12.021>

 Madsen, C. (2012). Technology adoption and adaptation in Canada’s west coast shipyards, 1918-1950. Business and Economic History On-Line, 10, 1-56. Retrieved from <http://www.thebhc.org/>

# **Appendix**

**Appendix A:** Regression Analysis



Table 1: Regression Analysis

**Appendix B:** Knowledge Graph

A diagram of a graph

Description automatically generated

**Appendix C :** CO2 Emissions by Country

A graph of co2 emissions

Description automatically generated

**Appendix D:** CO2 Emissions by Fuel Types

A graph of a graph showing the cost of carbon dioxide

Description automatically generated

**Appendix E:** CO2 Emissions by Manufacturer

A graph of co2 emissions

Description automatically generated