Vehicle CO2 Emissions

Louis Farmer
College of Computer Science and Engineering
California State University, Northridge
louis.farmer.89@my.csun.edu

Omar Marron
College of Computer Science and Engineering
California State University, Northridge
omar.marron.749@my.csun.edu

Jarod Wilson
College of Computer Science and Engineering
California State University, Northridge
jarod.wilson.501@my.csun.edu

I. Introduction

Our paper is going to be researching the area of car emissions and pollution. We hope to explore different fields in the area with the end goal of being able to visualize which factors affect a vehicle's emissions the most. Specifically, we will be researching how each of these factors affects the carbon dioxide (CO2) emission of a car. We believe that using the key features of a car, a machine learning model could be produced to accurately predict the emissions of a car.

Research like ours could be used to improve CO2 emissions within vehicles in the future, as companies will know which areas they should focus on optimizing the most due to which have the biggest impact. Some of the parameters we will be evaluating include the car's make, the engine's size in liters, how many cylinders the engine has, and the type of transmission, among many others.

To achieve our goal, we want to use linear regression and multiple linear regression. After evaluating our data with those models, we want to create competing models with techniques such as support vector regression, decision trees, and random forest to see how they compare with the original.

The goal is to create a predictive model that can accurately determine if a car will be CO2 efficient or not, and which areas of the car's design are affecting said emission the most.

II. Methodology

A. Dataset

The dataset that we are using is titled “CO2 Emission by Vehicles” and can be found here: https://www.kaggle.com/datasets/debajyotipodder/co2-emission-by-vehicles. The dataset has various pieces of data that help us distinguish vehicles, the main ones being the make and model of the vehicle. However, there are many more attributes that we will analyze. These include the class of the vehicle, the size of the engine, how many cylinders the engine has, the kind of transmission it has, the type of fuel it takes, and its fuel consumption values (both city, highway, and combined).

The dataset originates from data taken from the Canadian Government. This means we will have to look out for any models of cars that would have different regulations than standard. However, we can still say our data is valid for Canadian cars in particular.

B. Performance Metrics

Since we are primarily creating a prediction model that will predict whether a car will be CO2 efficient or not, accuracy will play a big part in performance and will be recorded and evaluated heavily. Along with that, precision is also very important because consistency matters a lot with a model like this. Low precision means the model is worthless. We also want to observe the Mean Absolute Error (MAE) which will indicate to us the absolute difference between the predicted

and actual values of CO2 emissions. A lower value here will mean we have a better fit and thus a more accurate model.

We will also measure performance through visualization by creating an AUC-ROC curve for each model and graphing them against each other. This will allow us a direct way of comparing the models in total, in addition to our numeric methods. Other performance measurements such as the precision, recall, and f1-score can be looked at for additional information and to help decide the best model assuming a few of them are close to each other on the AUC-ROC curve.

C. Setup

As far as the software is concerned, we will be conducting all of our experimentation inside a shared Google Colab Notebook. The experiment will be written in Python and will utilize many libraries that are also written in Python. For the models and data processing, the only library that will be used is sklearn from Scikit-Learn. This includes a few different modules such as LinearRegression, SVR, tree, and RandomForestRegressor; used for creating models of Linear Regression/Multiple Linear Regression, Support Vector, Decision Tree, and Random Forest respectively.

There are also some secondary modules that we will need to use in preprocessing. This includes standardizing and encoding our data. The libraries we will use to achieve this are make_classification and StandardScaler, both of which are from Scikit-Learn.

III. Related Works

In this section, we present a review of some of the works that relate to our field of research. We hope to provide a thorough and comprehensive overview of some of the existing literature, highlighting the key contributions and limitations of each paper we cite. By doing this, we hope to find existing gaps in the research literature which will then aid us in carrying out the best analysis of data possible. We hope to include a wide range of publications from different sources and ideologies. For each paper we cite, we hope to provide a critical evaluation, including how it is relevant to the research question and how it contributes to our understanding of our research topic.

A. Reference 1

This reference includes an introduction to and analysis of a projection model VERSIT+ LD, made to make projections for traffic CO2 emissions for a variety of traffic conditions. This source can be used to predict the impact of light-duty vehicles, in order to determine environmental impact and quality. While this model focuses mainly on the impact of vehicles in traffic ours focuses on the emissions of a given model, fuel type, etc. As such the focus of our models and features that we use are very different. The model they have focuses on the circumstances of a vehicle in motion while ours focuses on a vehicle's qualities.

B. Reference 2

This paper is an analysis of fuel consumption and CO2 emissions from passenger vehicles, specifically in Europe. It takes into account a few different factors that are very relevant to our area of research, such as the weight of the vehicle and how that affects the amount of fuel it consumes. It also includes some areas of research that our data does not include, such as the effects of auxiliary systems such as air conditioning in a vehicle. The main things that we will be looking to use for our research, though, are the mass of the vehicle as well as some of the contextual information that this paper gives.

C. Reference 3

This reference is an official US government website that outlines many of the parameters that we are going to be using for our piece of research. It outlines things such as the year and type of car, as well as the miles per gallon (MPG) of the car and how that relates to the CO2 emission of the vehicle. This dataset is helpful because it includes some electric vehicles, which we can compare the gas-powered vehicles to for reference. We also get to see some of the trends that have been exhibited from around 1970 to the present year. One major difference between this set of data and ours is that they are focusing on the improvement of CO2 emission in vehicles over the years, whereas we're focused on seeing what contributes the most to the emission and if we can predict that emission.

D. Reference 4

This paper has a big emphasis on the future of energy usage in the car industry and how that will affect CO2 emissions. The studies take data from industries in China, France, Germany, India, Japan, and the United States, meaning it is very diverse. The main concept of the paper is looking into whether electric cars will have a positive effect on greenhouse gas emissions and whether the actual production of said cars in both a manufacturing and waste management sense will counteract the effects of those emissions. This differs from our research because we are not looking at electric vehicles at all, but it is interesting to see how the production of certain cars has an impact on the environment, despite the cars themselves not having a negative effect.

E. Reference 5

This reference focuses on hydrogen powered cars and how factors of the vehicle influence its emissions. This model predicts emissions within 4% of experimental values, revealing a lot about the inner workings of hydrogen emissions. Compared to our data set and future outputs, theirs not only predicts emissions, but also predicts various kinds of emissions, such as CO in addition to CO2. Their results help to show what kind of information we can generate from our model, and how that information might be used.

F. Reference 6

This reference is a comparison and exploration into whether it is better from a CO2 emission point of view to have an older car, or a better more modern car. It goes into detail regarding the scrappage schemes of old cars, and if that process, plus the production of a new car is good enough to offset the bad CO2 emissions of the old car.

G. Reference 7

This article explains how the European Parliament has approved new targets to reduce carbon dioxide (CO2) emissions from new cars by 37.5% by 2030, compared to 2021 levels. The current target is a 40% reduction by 2021, but there has been concern that car manufacturers may not meet this goal. The new legislation also sets a target for a 31% reduction in CO2 emissions from new vans by 2030. The European Parliament hopes that these targets will help to combat climate change and improve air quality in European cities.

H. Reference 8

In this reference the authors focus on car emissions in roadways, the main concern being pollution in civilian areas. Despite focusing mainly on highway emissions most of its data is sourced from cars focusing on features such as model year, emission standards, etc.. These features are somewhat similar to our own however our source includes features such as gas type, consumptions, etc.. While this source is focused on how emissions change with a given timeframe ours is focused on predicting the resulting emissions from a given vehicle. Thereby this source is more focused on environmental impact rather than creating a predictive model to substitute experimental data, or supplement it.

I. Reference 9

This article focuses on CO2 emissions from transportation, one of the leading sources of greenhouse gas emissions. The article highlights the significant increase in emissions from transportation over the past few decades, driven by the growth in the number of vehicles and travel demand. The article also discusses the different types of transportation and their contribution to emissions, with road transport being the largest emitter. The article further explores various strategies to reduce emissions from transportation, including promoting public transport, improving fuel efficiency, and transitioning to electric and alternative fuel vehicles. Finally, the article emphasizes the need for policymakers to implement a combination of policies and incentives to achieve significant emission reductions in transportation.

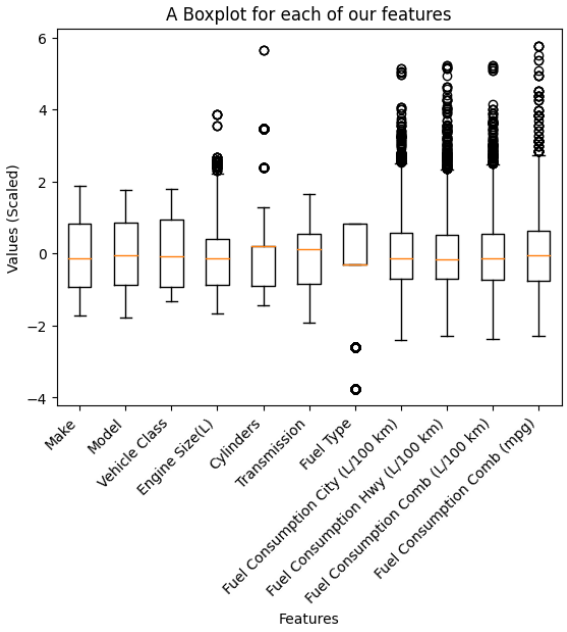
J. Reference 10

This paper takes a slightly different approach to analyzing CO2 emissions. Instead of looking at the car itself, it instead takes a look at what happens after the car is made and is looking for a buyer. In Europe, they introduced legislation which means that cars will have to display their CO2 output to the buyer. Each car is given a letter rating, similar to the power efficiency rating certain appliances get. This creates consumer awareness and may lead consumers to make more environmentally aware choices when buying a car. It also makes it easier for the consumer to see if a car is going to be fuel efficient or not which is a big factor to consider for most.

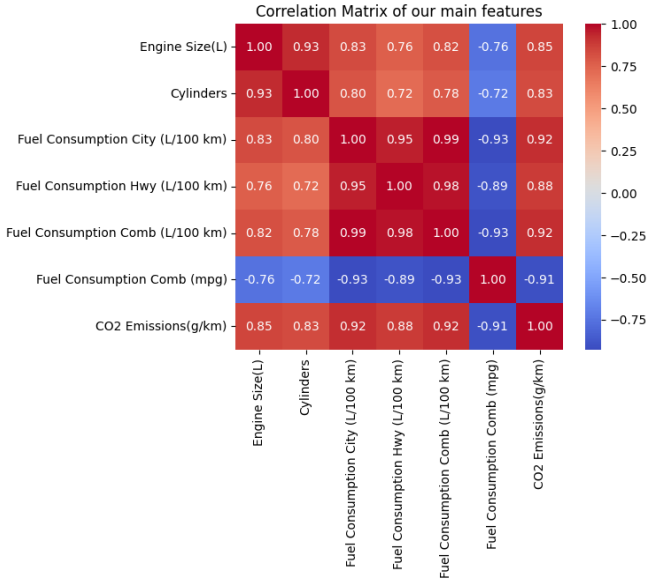
IV. Results

A. Visualization

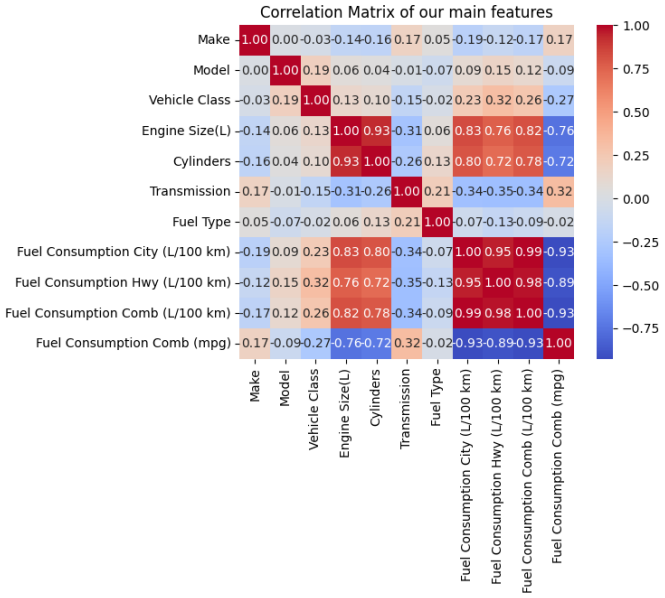
To evaluate the performance of our models, we used a combination of statistical metrics and visualizations. Before implementing any of the models we intended to use, we first explored the dataset and used a few different graphs to give us an idea of the data we were working with. The first one being the following box-plot diagram that allows us to see if our data has any outliers.



The boxplot shown above was made after the data had been scaled, due to there being many different kinds of data such as vehicle model year and cylinders inside the engine. As you can see, there are a few features of our data that have many outliers, and others that have very few outliers. The other visual representation we have of our data is a correlation matrix. This allows us to see which features inside our data are heavily related to each other and which ones aren't.



The diagram shows how each of our features are heavily correlated with one another. Fuel Consumption shows an inverse correlation, but it still shows a strong trend. This shows the power of our dataset, and helps illustrate how useful each of our features are. However this lacks information regarding our categorical features.



Taking the results again after label encoding, we still see some areas where features are highly related. A few that are more obvious than others are the different kinds of fuel consumption types like city, highway, and combined. Another example is the engine size being heavily related to the number of cylinders the engine has. The only other area that has strong positively correlating relationships is the engine size and cylinders with the different fuel consumption types. These are areas that would be expected to be linked heavily, but there are some areas that are strongly negatively correlated such as the engine size with the combined fuel consumption. There are also various features that you would not expect to have any kind of correlation such as the make and model, or transmission and model.

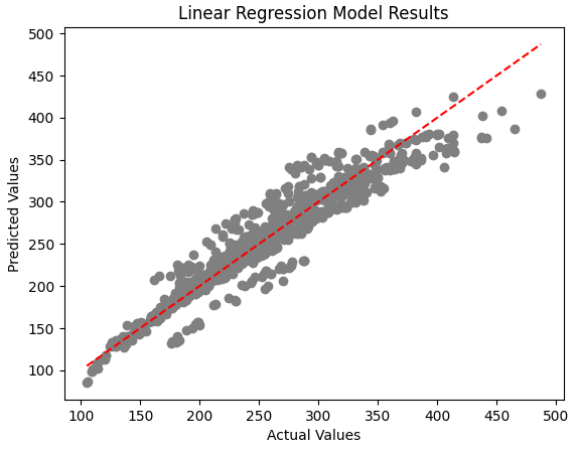
B. Data Preparation

Before working on the model we also performed NaN detection and label encoding as previously mentioned. Fortunately there were no NaN results, checking both for the recognized NaN and for potentially other changes such as odd data types. Checking each value count for each categorical row also shows no odd values. Further data-preparation such as noise reduction does not seem necessary at first, and we would like to try with our raw dataset before adjusting anything for skew or outliers. We will however perform normalization, doing so after splitting our dataset. Using a pipeline would simplify the process, as well as maybe using ensemble with hard voting to test our model, however this does not seem necessary for now.

The goal of our research was to compare various models, including linear regression, support vector regression, decision tree, and random forest regressors. In our comparisons we'll look to determine which model performs the best without modification, after which we can then attempt hyperparameter tuning and feature engineering to improve our model further.

Our training data consisted of 80% of our total dataset, leaving 20% for us to do testing on. Another validation set would be good, however given our dataset is so small we cannot make one out of the original set, and do not have a proper source for an external set.

C. Linear Regression



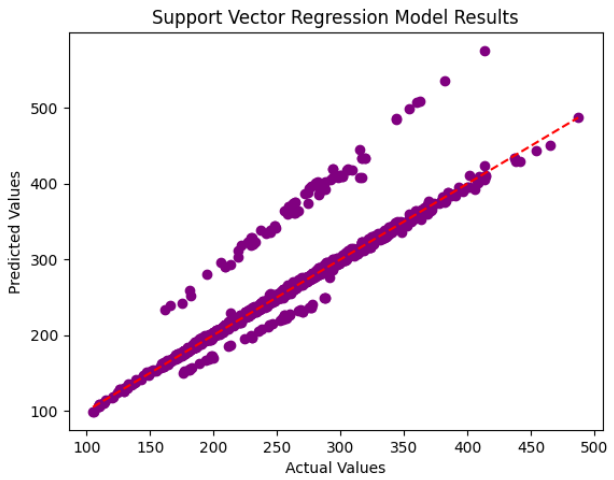
We first fitted and ran a linear regression model as a baseline to compare the other models to. The linear regression model gave us a R2 score of 0.9141, meaning that it explains around 91% of the variance in the target variable. We also found an RMSE of 17.2, which further shows how performant this model is. For the purposes of each model we will mainly look at R2 scores.

To determine if the model is overfitting we have the initial graph which shows no signs of overfitting. In addition, we performed K-Folds cross validation on the model, and got

results that were within +/- 0.03 of 0.9, which is fairly consistent. For future models we will simply note if the cross validation scores were consistent or not.

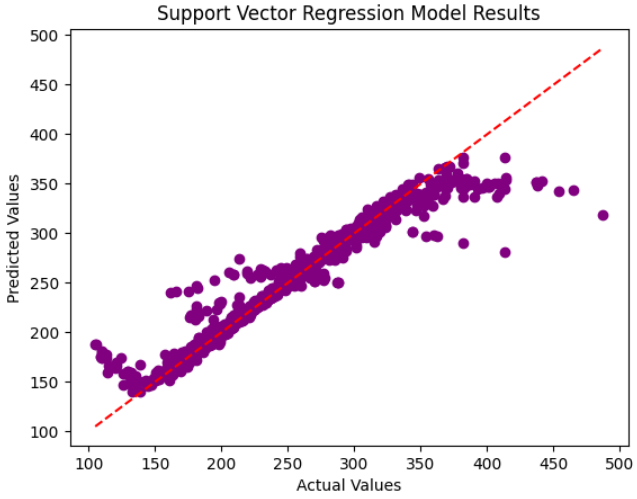
If we compare this with the other models we fitted and trained, it performed worse than average. The only model that performed worse than the linear regression model was the support vector regression model, which scored 0.8076 as its R2 score. This may be due to our dataset being potentially non-linear, which may indicate that we would want to switch our SVM kernel from Linear to RBF.

D. Support Vector Regression (Linear SVM)



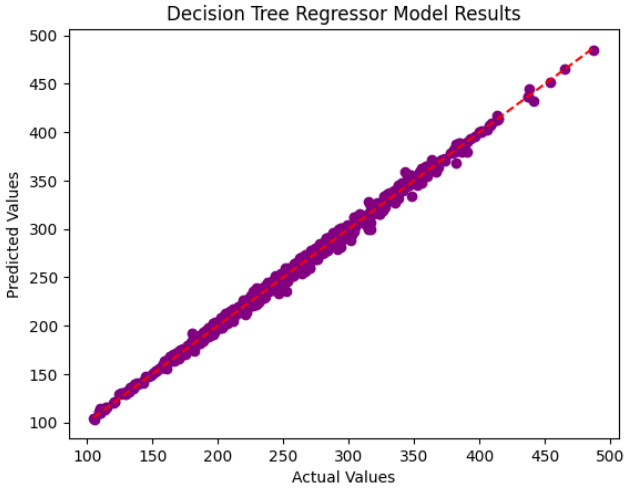
Our Support Vector Regression model performed far worse than other models, with only a 0.81 R2 score. This could be for a few reasons, first being that we used Linear SVM and that it is struggling to make a linear hyperplane for our nonlinear dataset. By switching our kernel from Linear to RBF we may see better results, which we will also attempt.

E. Kernel RBF SVM



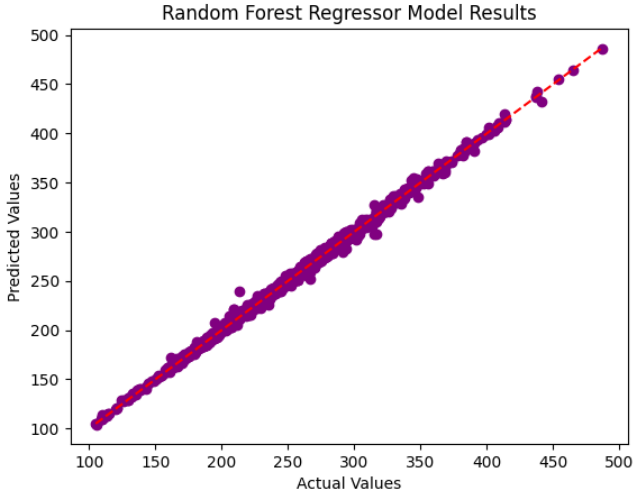
After switching our kernel we get far better results. An r2 score of 0.925 with a RMSE of 16 is a massive improvement from our previous results, however for both kernels the cross validation scores indicate some issues. For Linear SVM the cross validation scores would not run at all, either indicating a problem in our methods or an issue with the model. For RBF, the cross validation scores were wildly inaccurate. For both of these, and given the graph, it could just indicate an error in our methods, however it would be best to stay cautious and assume that these models might be overfitting.

F. Decision Tree Regressor



The decision tree regressor and random forest regressor were a little ahead in performance. The decision tree regressor scored an R2 score of 0.997, meaning it was very accurate with a few inaccuracies here and there. In addition, its cross validation scores are consistent across the board.

G. Random Forest Regressor



The Random Forest Regressor scored a high score of 0.998, meaning that it was the most accurate predictor of our data, and predicted the results almost perfectly, with a few visible outliers that it did not get correct. Given Random Forest is an ensemble of Decision Tree, it does stand to reason that it would be better than Decision Tree, so the results here

are unsurprising. The R2 score though is indeed quite high, and its cross validation remains consistent, so this isn't due to overfitting.

Any higher than our current results may lead to overfitting, as no model will ever have 100% accuracy unless it's overfitting. Initially we planned to perform tuning and feature engineering, however given the performance of our Random Forest model, this does not seem necessary, and we can confidently say that our initial hypothesis and thesis was correct, that a model can be produced for our specific use case.

Discussion

Following the execution of our analysis, we looked over what we found and drew some observations and conclusions. Overall, the dataset proved to be very predictable, with only a few categories seeming to be harder to do in this area. We noticed strong accuracy with all of the models we tested, which indicates that most of the data correlates with much of the other data, meaning this would be a good dataset to carry out further analysis on and could be used effectively for real applications. As was discussed in the results section, the random forest classification model scored the highest, meaning this would most likely be the most effective model to use for further training and use. However our data is limited to specific models and makes, so usage of a validation from outside our original data set was not an available option.

Limitations

There were very few limitations in our project, at least none that prevented us from producing a model. Our data set was clean, and reasonably balanced. Since we were working with regression models, issues with an imbalance in target class were less of an issue. However, while our model performs well some of the features we include may cause issues in applying this model to future vehicles. Features such as car model and make may not be applicable for newer car models or makes. Furthermore, our data set is using information from a Canadian source, meaning that our data and model is likely only applicable in Canada where regulations may be different than other countries. In the future, it might be a good idea to retrain the model using information from the Bureau of Transportation Statistics or another agency in order to get information that would be applicable in the United States.

Conclusion

The results show that there are trends in our given feature set, and that it is possible to predict emissions based on a car's features. We found that Random Forest was the best for this as it produced a model with very little variance. We believe this model can be used by car manufacturers as a way to get early information on car emissions in order to meet guidelines. It can also be used as a way to double check their own measurements, though the results from the model should not be treated as 100% accurate. This model should however only be applied to very specific models and makes, or similar, that the model was trained on, due to its training features. In the future this model could be applied to these cases should the features be broadened and more defining features be removed.

References

[1] R. Smit, R. Smokers, and E. Rabé, "A new modeling approach for road traffic emissions: VERSIT+" *Transportation Research Part D: Transport and Environment. Volume 12, Issue 6. 2007.*

[2] G. Fontaras, N.-G. Zacharof, and B. Ciuffo, *Fuel consumption and CO2 emissions from passenger cars in Europe – Laboratory versus real-world emissions*. 2017.

[3] "Highlights of the Automotive Trends Report," EPA. [Online]. Available: https://www.epa.gov/automotive-trends/highlights-automotive-trends-report. 2022.

[4] Gómez Vilchez, J. J. & Jochem, P. (2020) *Powertrain technologies and their impact on greenhouse gas emissions in key car markets*. Transportation research. Part D, Transport and environment. [Online]. 2020, Vol. 80.

[5] T. Ho, V. Karri, D. Lim, D. Barret, "An investigation of engine performance parameters and artificial intelligent emission prediction of hydrogen powered car" *International Journal of Hydrogen Energy*. Volume 33, Issue 14. 2008.

[6] Kagawa, S. et al. (2013) *Better cars or older cars?: Assessing CO2 emission reduction potential of passenger vehicle replacement programs*. Global environmental change. [Online]. 2013, Vol. 23.

[7] European Parliament. 2023. CO2 emissions from cars: Facts and figures (infographics): News: European parliament. (February 2023). Retrieved March 10, 2023 from https://www.europarl.europa.eu/news/en/headlines/society/20190313STO31218/co2-emissions-from-cars-facts-and-figures-infographics

[8] L. E. Yu, L. M. Hildemann, and W. R. Ott. "A Mathematical Model for Predicting Trends in Carbon Monoxide Emissions and Exposures on Urban Arterial Highways." *Journal of the Air & Waste Management Association*. Volume 46. Issue 5. 2012

[9] Hannah Ritchie. 2020. Cars, planes, trains: *Where do CO2 emissions from transport come from?* (October 2020). Retrieved March 10, 2023 from https://ourworldindata.org/co2-emissions-from-transport

[10] Haq, G. & Weiss, M. (2016) CO2 *Labeling of passenger cars in Europe: Status, challenges, and future prospects*. Energy policy. [Online]. 2016, Vol. 95