An Analysis on the Impact of Electric Vehicles on Pollution and Health

STAD94 Statistics Project

Pao Zhu Vivian Hsu

August 13, 2020

Acknowledgement

I would like to express my special thanks to Dr. Ken Butler, a Lecturer at the University of Toronto, for providing me with this excellent opportunity to carry out a statistical project with real world data in an academic setting. Dr. Ken Butler has generously supervised me and provided me with meaningful feedback so that I could learn to improve myself over the course of this project.

Abstract

The objective of this study was to assess whether increasing the quantity of electric vehicles within a country results in a decline of ambient air pollution and improves population health. Through the use of data from various online platforms, a quantitative study was conducted to investigate this relationship. Vehicle, pollution, and health data was collected for 34 countries across America, Asia, and Europe for the years 2008 to 2019. The data was plotted using scatter plots and various regression models were built. In terms of pollution, regression models have indicated that pollutant levels decrease as the percent share of electric vehicles increase. However, the rate, form, and strength of this decrease varies by country. In terms of health, scatter plots have shown that countries tend to have lower DALYs and death rates per 100,000 population when there is a greater percent share of electric vehicles. However, the true effect of increased electric vehicles on health is unknown due to limitations on the number of years of data available for each country. Given these conclusions, it is suggested that greater efforts should be made to push electric vehicles in the market to reduce ambient air pollution. But further studies are required to better understand the relationship between electric vehicles and health.

Introduction

As the Earth's average temperature continues to rise over time, the search for environmentally sustainable transportation has become increasingly important to reduce the effects of global warming. Since the early 19th century, electric vehicles have been considered as a solution to reduce ambient air pollution and has been recently deemed as a method to improve public health (U.S. Department of Energy, n.d.-a, n.d.-b). However, the impact of these vehicles on reducing pollution and improving health has not been thoroughly evaluated. The objective of this study is to investigate the relationship between electric vehicles, ambient air pollution, and health. In particular, the study aims to determine whether increasing the quantity of electric vehicles within a country results in a decline of ambient air pollution and improves population health.

Methods

Using data from various online platforms, a quantitative study was conducted to investigate the relationship between electric vehicles, pollution, and health. To do so, data was collected from a total of 34 countries and categorized into three datasets – one on vehicles, pollution, and health. The first dataset focuses on the breakdown of vehicles by fuel type. The fuel types used were petrol, diesel, electric, hybrid, liquefied petroleum gas (LPG)/ natural gas, and other fuels. The second dataset involves ambient air pollution measurements by pollutant. The pollutants observed include carbon monoxide (CO), carbon dioxide (CO_2), nitrous oxides (NO_x), sulfur oxides (SO_x), greenhouse gases (GHG), and volatile organic compounds (VOC). Finally, the last dataset focuses on health measures by disease. The diseases investigated include various respiratory diseases, cardiovascular diseases, and cataracts. After data was collected for each of the datasets, they were then combined into a single dataset for analysis.

Data Collection for the Vehicle Dataset

The first dataset contains the breakdown of vehicles by fuel type. In particular, vehicles were categorized into six fuel types including petrol, diesel, electric, hybrid, LPG/ natural gas, and other fuels. For the purpose of this study, electric vehicles are defined as vehicles that only use electrical energy to operate and hybrid vehicles are defined as those that use both electrical energy and liquid fuels to operate. Establishing these definitions helps to ensure that the data for each category is consistent despite differences in organizational definitions. In addition, this also improves the ability to answer the research questions. It was originally planned that vehicle registrations for a small number of countries would be investigated for this dataset. However, complications arose due to variations in how every country manages their data. For example, some countries only release data on vehicle sales rather than registrations, others choose to keep their data entirely private, and others do not collect data at all. For this reason, it was later decided that both vehicle sales and registration data would be collected, and a greater number of countries would be included in order to increase the amount of data for the study and improve the prospects of observing an association. The final dataset includes sales and/ or registration data for 34 different countries including Canada, the United States, China, Japan, South Korea, Brazil, Australia, and 27 countries from Europe. Please see Appendix A for the full list of countries. The following describes the key components of the data collection process for each of the countries included.

Data for Canada and the United States was mainly obtained from government websites. For Canada, data was obtained from Statistics Canada. It provided data on the number of new motor vehicle registrations for the years 2011 to 2019 by six fuel categories, precisely petrol, diesel, battery electric, hybrid electric, plug-in hybrid electric, and other fuel types (Statistics Canada, 2020). Based on the definition of hybrid vehicles for this study, hybrid electric vehicles and plug-in hybrid electric vehicles were combined into a single measure for hybrid vehicles. For the United States, data was collected from the U.S. Energy Information Administration (EIA) along with the Alliance of Automobile Manufacturers. On a yearly basis, the EIA produces reports to project sales for light vehicles that run on petrol and diesel. While these reports emphasize sales projections, real data is available for a limited number of years. This can be found with careful inspection of the projection charts in each report. For example, the 2019 report had projections for up to 2050 but contained real data only for the year 2017 in the sales projection chart (U.S. Energy Information Administration, 2019b). By combining real data from the 2014 to 2020 reports, eight years of data from 2011 to 2018 were collected for petrol and diesel vehicles sales. For electric and hybrid vehicle sales, data was obtained from the Alliance of Automobile Manufacturers rather than the EIA. The EIA only publishes alternative fuel data for government agencies, transit agencies, and fuel provider companies. As a result, they no longer publish national figures on electric and hybrid vehicle usage (U.S. Energy Information Administration, 2019a). Thus, data from a sales dashboard on the Alliance of Automobile Manufacturers website was used instead (2019). Manual methods were used to collect this data because data files were not available for download. Nonetheless, the total sales for the years 2011 to 2018 were collected for fuel cell electric vehicles (FCEVs), battery electric vehicles (BEVs), plug-in hybrid electric vehicles (PHEVs), and hybrid electric vehicles (HEVs). By the definition of electric and hybrid vehicles for this study, FCEVs and BEVs were later grouped into the electric vehicles category while PHEVs and HEVs were grouped into the hybrid vehicles category.

For China, Japan, and South Korea, data was collected from online databases, a vehicle association website, and a government website respectively. For these countries, difficulties arose due to language barriers since most websites are mainly written in their respective languages. For example, data for China may have been present on the government website, however there was no effective way to translate it and so data was alternative sources were required. Consequently, data was obtained from CEIC Data and Statista instead. CEIC Data provided sales data on vehicles that run on petrol and diesel while Statista provided sales data on electric and hybrid vehicles (CEIC Data, n.d.; Statista, 2020a). For Japan, data was obtained from the Japan Automobile Manufacturers Association (JAMA). Two reports published in 2019 provided 2008 to 2018 data for new registrations of diesel passenger cars and *next-generation* passenger cars, which are cars that use hybrid, plug-in hybrid, electric, fuel cell, and/ or clean diesel power to operate (Japan Automobile Manufacturers Association, 2019a, 2019b). Based on the study definitions, data for hybrid and plug-in hybrid vehicles, electric and fuel cell vehicles, and clean diesel and diesel vehicles were each combined into a single category for hybrid, electric, and diesel vehicles respectively. Data for petrol cars were not available through JAMA and could not be found through other open source means. However, by using the percent share of *next-generation* cars (Japan Automobile Manufacturers Association, 2019b), an estimate of the total number of cars that use fuels other than the ones previously listed was derived for each year. Please see Appendix B for details on the calculation. Since these estimates were approximately two times greater than the amount of diesel cars, it is likely that majority of these cars use petrol. However, to avoid making assumptions, this value was placed in the other fuels category. Finally, for South Korea, vehicle registrations were obtained from the Ministry of Land, Infrastructure and Transport (MOLIT) website. While South Korea's government websites were also mainly written in their own language, a greater portion of the website could be translated by Google Translate. As a result, vehicle inspection data by vehicle size and fuel type was found. It was then extracted using a free online image to Excel workbook converter, which was required since the website's download feature was not in service at the time. The converted data was manually reviewed with the website to ensure accuracy. The result included data for petrol, diesel, LPG, electric, hybrid, compressed natural gas, and other fuels for 2016 and 2017. Lastly, LPG and compressed natural gas was then combined into one category as per the study's fuel categories.

Data for Brazil and Australia was obtained from Statista and the National Transport Commission respectively. For Brazil, Statista provides 2014 to 2019 data on new registrations for light vehicles that use petrol, diesel, electric, ethanol, and flexible fuel. Despite the resemblance between the terms *hybrid* and *flexible fuel*, flexible fuel vehicles are not hybrid vehicles since they can use either petrol or ethanol to run, but not electricity (Statista, 2020b). Thus, flexible fuel vehicles were not classified as hybrid vehicles but were combined with petrol vehicles instead. Note that they were not classified as other fuel vehicles because the number of vehicles using ethanol is very low and so it is likely that most flexible fuel cars are using petrol regardless of its flexible fuel capabilities. Ethanol vehicles were then classified as vehicles that use other fuels. While Brazil differs greatly from the other countries in terms of income-level, it was included in the study because it offers additional insight on how the relationship of interest could differ for lower-income countries. For Australia, 2016 to 2019 data on new light vehicle registrations were manually extracted from three reports by the National Transport Commission. This includes data on vehicles that use electric, petrol, diesel, and LPG fuels (National Transport Commission, 2018, 2019, 2020).

For the European countries, data was obtained from the European Automobile Manufacturers' Association (ACEA). The ACEA produces annual reports to describe the vehicles in use for countries in the European Union. Although the ACEA likely collected data before 2016, reports were only found for the years 2016 to 2019. Each report contains at least five years of data prior to the report year, however only one year is available for the breakdown of vehicles by fuel type (European Automobile Manufacturers' Association, 2017, 2018, 2019). This means only three years of data can be used. Extracting the data required more effort than initially expected. Firstly, the data which classifies vehicles by fuel type was presented as percentages of the total vehicle population rather than whole numbers. This is inconsistent with the data for all other countries. Thus, the percentages were converted to whole numbers by multiplying it with the total vehicle population for the given year. The total vehicle population was also provided within the same reports (European Automobile Manufacturers' Association, 2017, 2018, 2019). Another challenge with extracting the data was due to the PDF format of the reports. While data for other countries were also found in PDF format, this data was more challenging to deal with since copying and pasting the data into an Excel workbook resulted in messy strings of data where multiple data points were concatenated into a single string. In addition, since the data was presented in a slightly different way in each report, the data was also inconsistent with one another and required different approaches to organize it. Nonetheless, the data was cleaned and organized using R, resulting in three years of data for 27 different countries categorized by petrol, diesel, hybrid, electric, LPG/ natural gas, and other fuels.

After collecting vehicle data for each of the countries, the data was merged into a single dataset using R. This dataset served as a foundation for the next two datasets since pollution and health data would only need to be collected for these countries to determine an association between electric vehicles, pollution, and health.

Data Collection for the Pollution Dataset

The second dataset stores measurements of ambient air pollution by pollutant. As stated earlier, the six pollutants included in this dataset are CO, CO_2, NO_x, SO_x, GHG, and VOC. Data was obtained for these pollutants from the Organisation for Economic Co-operation and Development (OECD), which provides data from as far back as the 1960s to 2018. While the OECD offers multiple measurements for each pollutant (OECD, 2020), the selected measurements were tonnes per capita for CO_2 and GHG and kilograms per capita for CO, NO_x, SO_x, and VOC since these measurements contain the most data for each country. Ideally, more recent data was to be collected. However, most websites which do offer current data provide it by city and only for a limited number of months for free. Hence, older data was used instead. Initially, the data was taken from UNdata, a United Nations data service, which included data on CO_2, GHG, nitrous oxide (N_2O), methane (CH_4), nitrogen trifluoride (NF_3), hydrofluorocarbon (HFC), polyfluorinated chemicals (PFC), and sulfur hexafluoride (SF_6) (UNdata, 2020). However, OECD offered more common pollutants and provided at least some data for China compared to UNdata. Therefore, the OECD data was finally chosen as the source of pollution data and was organized in R to only include the 34 countries of interest.

Data Collection for the Health Dataset

The last dataset contains data on death rates, disability-adjusted life year (DALY) rates, and years of life lost (YLL) rates per 100,000 population for various diseases attributed to ambient air pollution. These diseases include ischaemic heart disease, stroke, cataracts, lower respiratory infections, various cancers, and chronic obstructive pulmonary disease. This data was collected from the World Health Organization (WHO), a multilateral organization specialized in leading global health responses (World Health Organization, n.d.-b). Once the data file was downloaded from the website, it was organized in R so that it only includes the 34 countries of interest.

Putting It All Together

After data was collected for vehicles, pollution, and health, the three datasets were merged into a single dataset using R. The percent share of electric vehicles for each country and year was computed and added to the data. This percentage allows direct comparisons to be made between countries despite differences in population size. The final dataset was then used for analysis to determine the impact of electric vehicles on pollution and health.

Analysis

Electric Vehicles and Pollution

To examine the relationship between electric vehicles and pollution, the percent share of electric vehicles were plotted against each of the pollutants in a faceted scatter plot as shown in Figure 1. On each plot, the points appear to follow multiple trends rather than a trend. This suggests that creating a single model to describe the relationship would not be effective and multiple models are required instead.

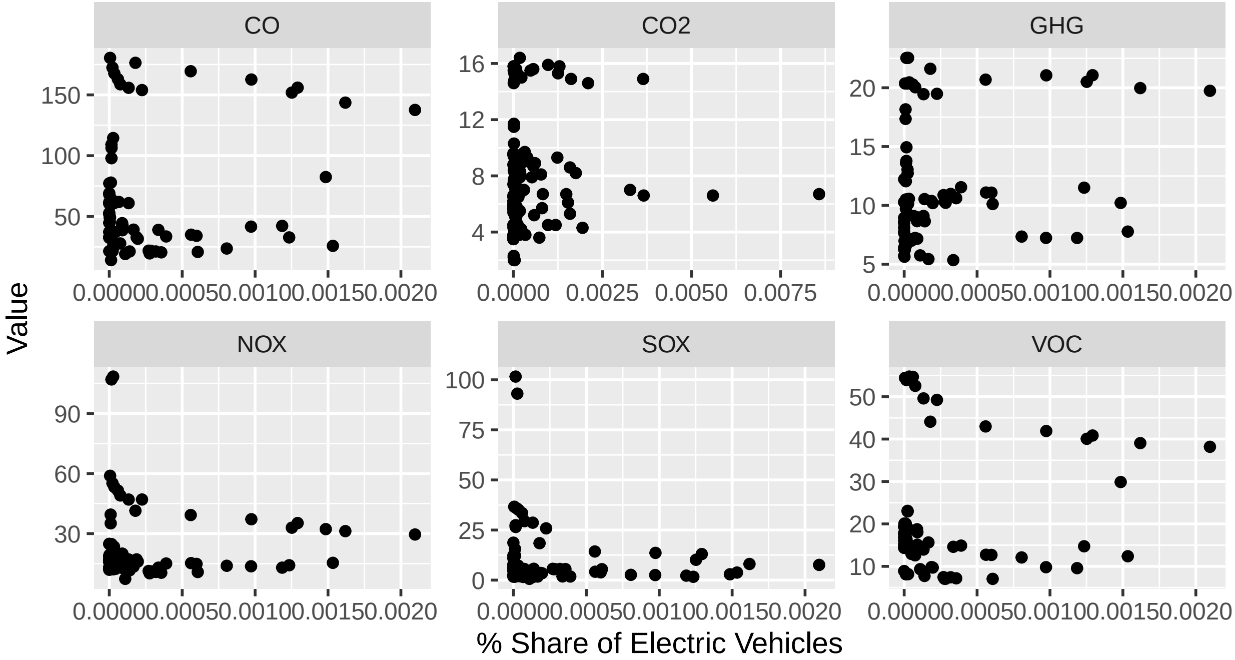


Figure 1: Percent share of electric vehicles against pollutant levels

Since the data is comprised of 34 different countries, differences between the countries could have produced multiple trends. For instance, people in Canada and the United States heavily rely on personal vehicles for transportation. This could differ from another country, such as Japan, which relies on public transit instead. These types of factors make air pollution greater in some places compared to others resulting in multiple trends being observed. Since countries within certain regions tend to have similar characteristics, WHO regions were used to group the countries together. The 34 countries were categorized into three WHO regions as outlined in Table 1 (World Health Organization, n.d.-a).

Table 1: Countries classified by WHO region

WHO Region	Countries
Americas	Brazil, Canada, United States
Western Pacific	Australia, China, Japan, South Korea
Europe	27 European countries

Using these regions, the points were plotted again in Figure 2 and the trends do appear to differ by region. For Europe, many of the points were very close to zero in terms of the percent of electric vehicles. These points could be coming from countries which generally do not use electric vehicles and so they would not have much value to assess the impact of electric vehicles. Removing them from the data could help to simplify analysis since they will not produce a trend anyways. For the Americas and the Western Pacific, clusters of outliers were present in some of the plots.

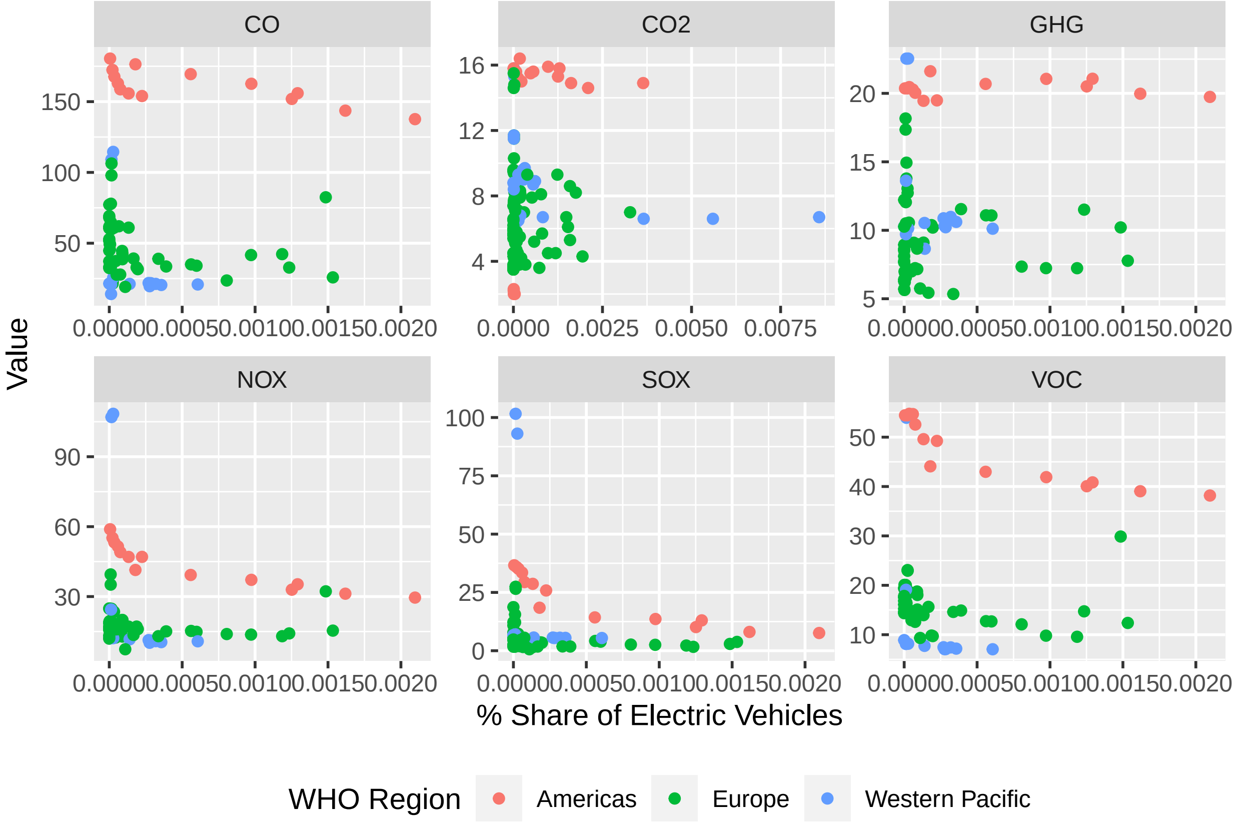


Figure 2: Percent share of electric vehicles against pollutant levels by WHO region

For instance, in the CO_2 plot for the Americas, the points for Brazil were outliers as shown in Figure 3. This could be due to the fact that Brazil is culturally and economically different from Canada and the United States.

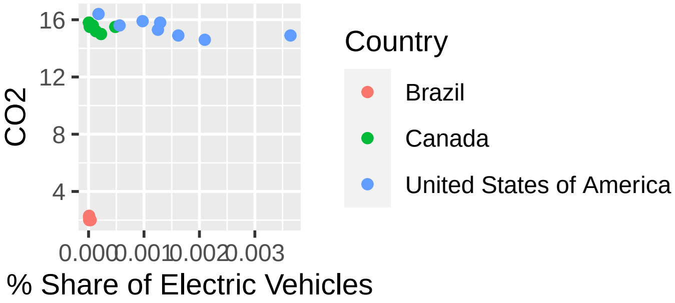


Figure 3: CO2 against percent share of electric vehicles for Americas by country

For the Western Pacific, there also appears to be a cluster of outliers in each of the plots. Since the Western Pacific consists of three East Asian countries (China, Japan, South Korea) and Australia, the outliers could be coming come from Australia since Australia differs drastically from the other three countries in terms of culture and demography. Figure 4, a plot of the Western Pacific coloured by country, confirms this. Note that there aren't many data points for South Korea and that the points do differ from China and Japan for most of the pollutants. As a result, South Korea may also be acting as an outlier in relation to China and Japan.

For the Western Pacific, the points for Australia were outliers as shown in Figure 4. Since the Western Pacific consists of three East Asian countries and Australia, it is expected that Australia would differ because their culture and demography are distinct. The points for South Korea were also be considered as outliers because they differ greatly from China and Japan for most of the pollutants and they only have one to two years of data regardless.

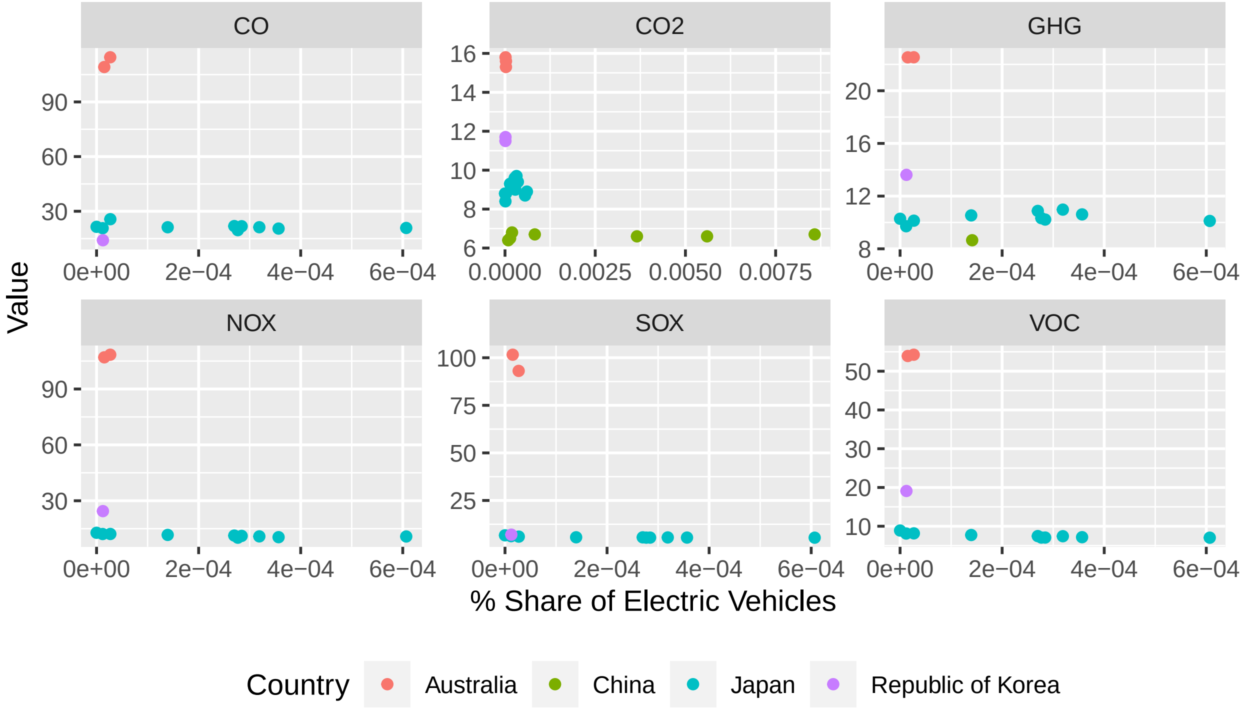


Figure 4: Pollutants against percent share of electric vehicles in Western Pacific

Overall, this meant using the WHO regions was not effective because it did not classify the countries well in terms of electric vehicle usage. A reclassification was required to deal with the outliers and identify European countries with near zero electric vehicle percentages. Table 2 outlines the classification used for the rest of this analysis.

Table 2: Reclassification of the countries

Region	Countries
North America	Canada, United States
East Asia	China, Japan
Northern Europe	Denmark, Estonia, Finland, Ireland, Latvia, Lithuania, Norway, Sweden, United Kingdom
Southern Europe	Croatia, Greece, Italy, Portugal, Slovenia, Spain
Eastern Europe	Czechia, Hungary, Poland, Romania, Slovakia
Western Europe	Austria, Belgium, Germany, Luxembourg, Netherlands, Switzerland

Note that Australia, Brazil, and South Korea have been removed in the new categories for two reasons. Firstly, they act as outliers when placed with other countries. Secondly, these countries only have 1-3 years of data with very low spread. Thus, grouping them with countries that have many years of data and larger spread would result in inaccurate analyses.

Canada and the United States

As border neighbours, Canada and the United States share many similar characteristics with one another. So initially, it was planned that a regression model would be created to describe the trends for both these countries together. However, based on Figure 5, Canada and the United States had considerably different trends despite them both having downward linear trends for each of the pollutants. In particular, Canada had a considerably steeper slope compared to the United States, which indicates that Canada's pollution levels decrease at a faster rate compared to the United States as the percent of electric vehicles increase. Due to these differences, separate models were built for each country.

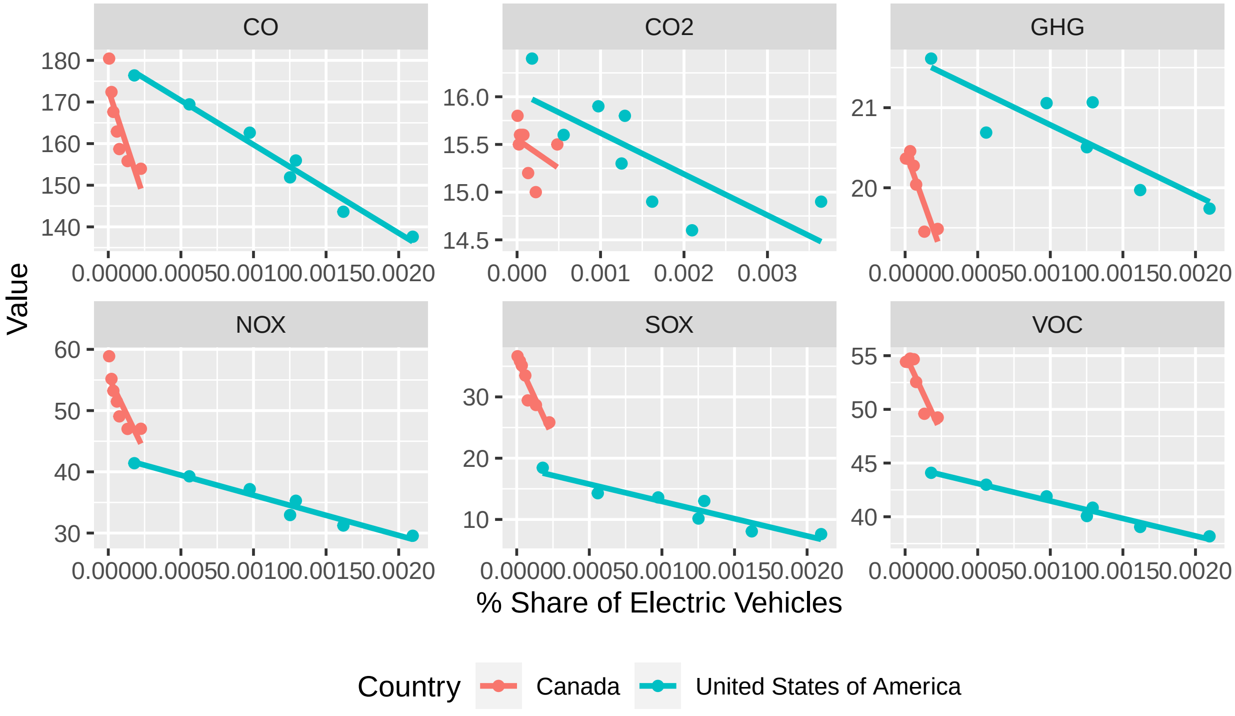


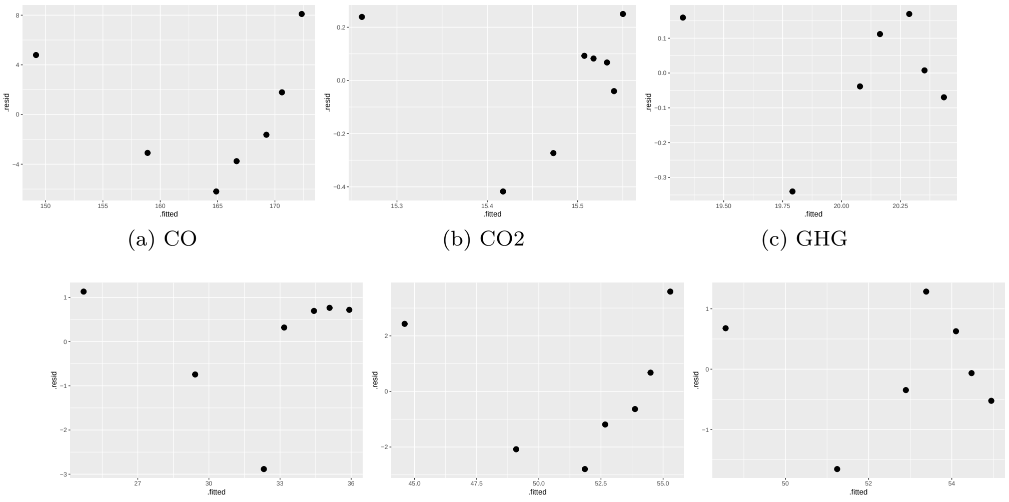
Figure 5: Pollutant levels against percent share of electric vehicles in North America

Since the points for each country were considerably linear, linear regression models were generated for each of the country-pollutant combinations as summarized in Table 3. For Canada, all of the models had a relatively high R^2 value except for CO_2, which had an R^2 value of 14%. Likewise for the United States, all of the models had a relatively high R^2 value except for CO_2, which had an R^2 value of 57%. This indicates that linear models explain the variability in the data well for all pollutants in Canada and the United States, except for CO_2.

Table 3: Linear models for Canada and the United States for each pollutant

Country	Pollutant	Intercept	X.Coef	R.squared
Canada	CO	172.94	-105908.40	0.72
Canada	CO2	15.55	-606.55	0.14
Canada	GHG	20.46	-5063.77	0.83
Canada	SOX	36.21	-51224.63	0.89
Canada	NOX	55.57	-48860.88	0.72
Canada	VOC	55.12	-29204.70	0.84
United States	CO	180.99	-21249.20	0.98
United States	CO2	16.05	-431.01	0.57
United States	GHG	21.66	-876.43	0.74
United States	SOX	18.59	-5638.17	0.89
United States	NOX	42.74	-6561.47	0.95
United States	VOC	44.72	-3254.10	0.97

To check if linear models were appropriate, the residuals were plotted against fitted values for each pollutant. In Figure 6, the points appeared to be following a somewhat quadratic trend for CO and NO_x. Therefore, a quadratic model was built for these pollutants. For CO_2, GHG, and VOC, there was an outlier on the top left corner of each plot. If this outlier was removed and the model was rebuilt for each pollutant, the R^2 values might improve. However, these outliers were not obvious after analyzing Figure 5 and additional methods were required to identify it. Due to the time constraints of this study, no model was selected for these pollutants. Finally, the points for SO_x did not appear to show any particular trend, yet they did not seem to be random either due to the cluster of points at the upper right-hand corner of the plot. Thus, no model could be built for SO_x.



(a) CO

(b) CO2

(c) GHG

(d) SOX

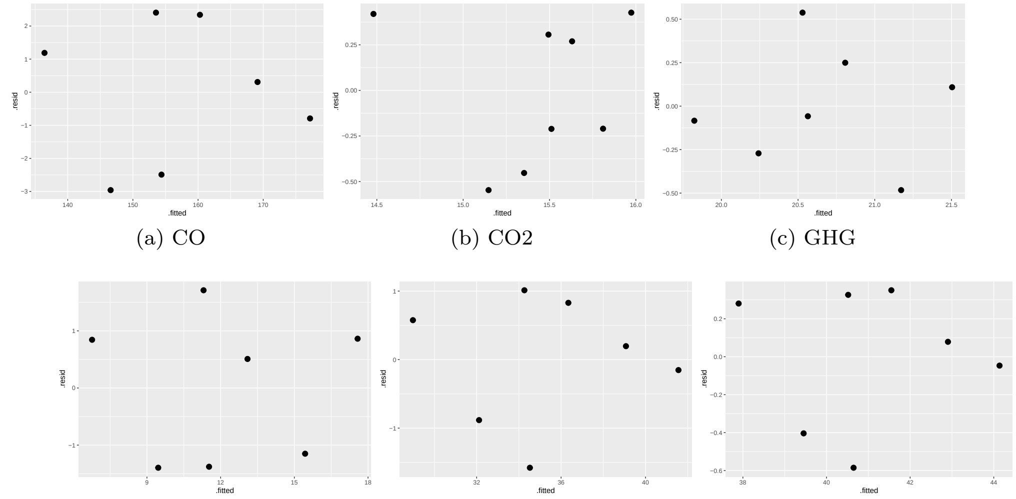
(e) NOX

(f) VOC

Figure 6: Residuals vs. fitted plots for linear models (a) to (f) for Canada

For the United States, the residuals against fitted values plots are shown in Figure 7. Unlike Canada, the points were quite random for all of the pollutants in the United States except CO_2. For SO_x and NO_x, the residuals range from -2 to 2.5 and -2 to 1 respectively. This suggests that there is possible skewness and a lack of normality. Although the points did not perfectly surround 0, this is still acceptable overall.

Like Canada, CO_2 also had an outlier for the United States and removing it could help to develop a better model. However, after analyzing the data, the outlier was not obvious and additional methods were required to identify it. Therefore, no model was selected for CO_2 due to the time constraints of this study. Linear models were appropriate to describe the trends for all other pollutants in the United States.



(a) CO

(b) CO2

(c) GHG

(d) SOX

(e) NOX

(f) VOC

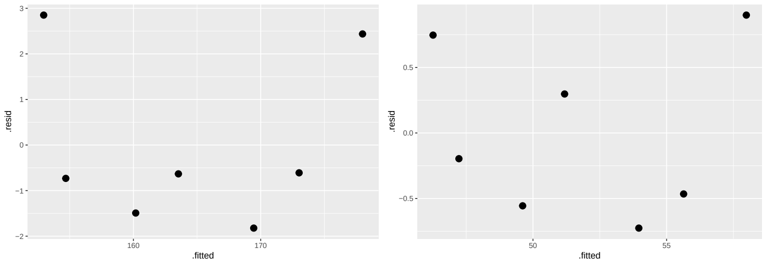
Figure 7: Residuals vs. fitted plots for linear models (a) to (f) for United States

As earlier stated, quadratic models were fitted for CO and NO_x in Canada because their residuals against fitted plots displayed a quadratic trend. Table 4 shows that the R^2 values have increased by over 20% with a quadratic model for both pollutants.

Table 4: Quadratic models for Canada for CO and NOX

Country	Pollutant	Intercept	X.Coef	X.squared.Coef	R.squared
Canada	CO	179.87	-334055.7	988333012	0.96
Canada	NOX	58.87	-157448.9	470402594	0.98

The residuals were then plotted against the fitted values to check if the quadratic models were suitable. This is shown in Figure 8. For NO_x, a quadratic model was suitable because the points on the plot were random and normal since the points centred around 0. However, a curve was still left over for CO which suggested that another type of curve may describe the data better.



(a) CO

(b) NOX

Figure 8: Residuals vs. fitted plots for quadratic models (a) and (b) for Canada

Thus, a rational ($x + \frac{1}{x}$) model was fitted as summarized in Table 5. With this model, the R^2 value was 93% which was lower than the R^2 value for the quadratic model. However, as shown in Figure 9, the residuals against fitted values were considerably more random for this model and normal compared to the quadratic model. Thus, the rational model was selected as the best model out of the three to explain CO trends in Canada as the percent of electric vehicles increase.

Table 5: Rational model for Canada for CO

Country	Pollutant	Intercept	X.Coef	one.over.X.Coef	R.squared
Canada	CO	165.94	-64654.74	8.98e-05	0.93

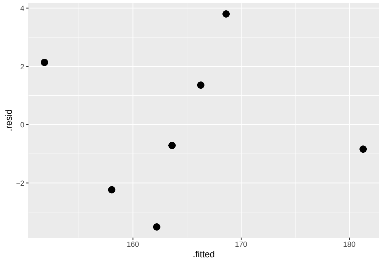
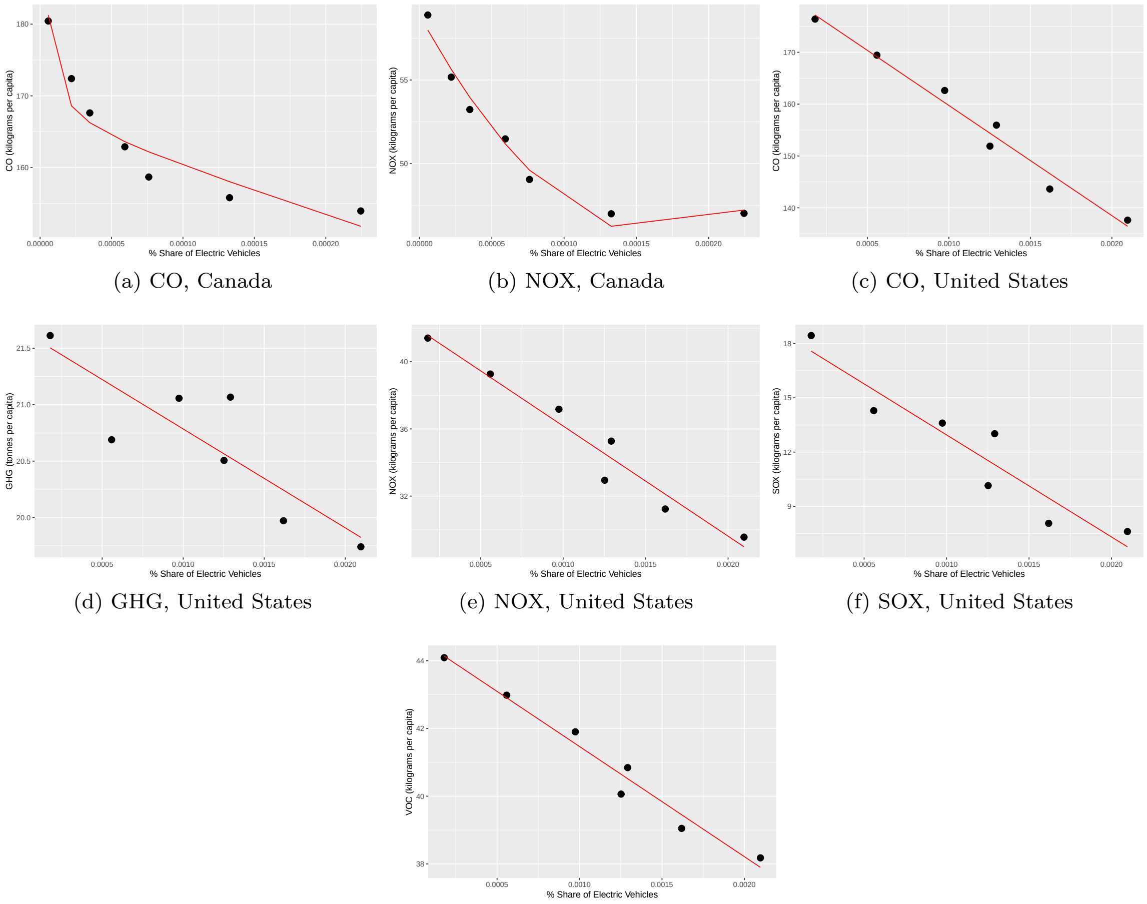


Figure 9: Residuals vs. fitted plots for CO rational model for Canada

Overall, a total of seven models were created for Canada and the United States as illustrated in Figure 10. These include a rational model for CO in Canada, a quadratic model for NO_x in Canada, and linear models for all pollutants except CO_2 in the United States.



(a) CO, Canada

(b) NOX, Canada

(c) CO, United States

(d) GHG, United States

(e) NOX, United States

(f) SOX, United States

(g) VOC, United States

Figure 10: Regression models (a) to (g) for Canada and United States

Overall, the regression models for Canada and the United States showed that an increase in the percent share of electric vehicles is associated with a decrease in pollution for various pollutants in Canada and the United States.

China and Japan

China and Japan also share many similar characteristics; they both have high population densities and similar cultures. However, the trends for China and Japan were also found to be quite different as shown in Figure 11. For China, only CO_2 had enough data to observe a trend. The trend appeared to be linear with a weak and unexpectedly positive slope of nearly zero. This suggested that an increase in the percent share of electric vehicles is not producing a significant impact on CO_2 levels. For Japan, a linear trend was seen for all pollutants except CO_2. For CO_2, the points were clustered towards the upper left-hand side of the plot. Since CO_2 was the only pollutant with more data for China, it was expected that Japan had a CO_2 trend but it was just on a different scale compared to China. The points in Japan were plotted again on its own to check this. In addition, CO has an outlier located on the upper left-hand corner of the plot. This was removed before building the model for CO since an outlier would mask the true trend.

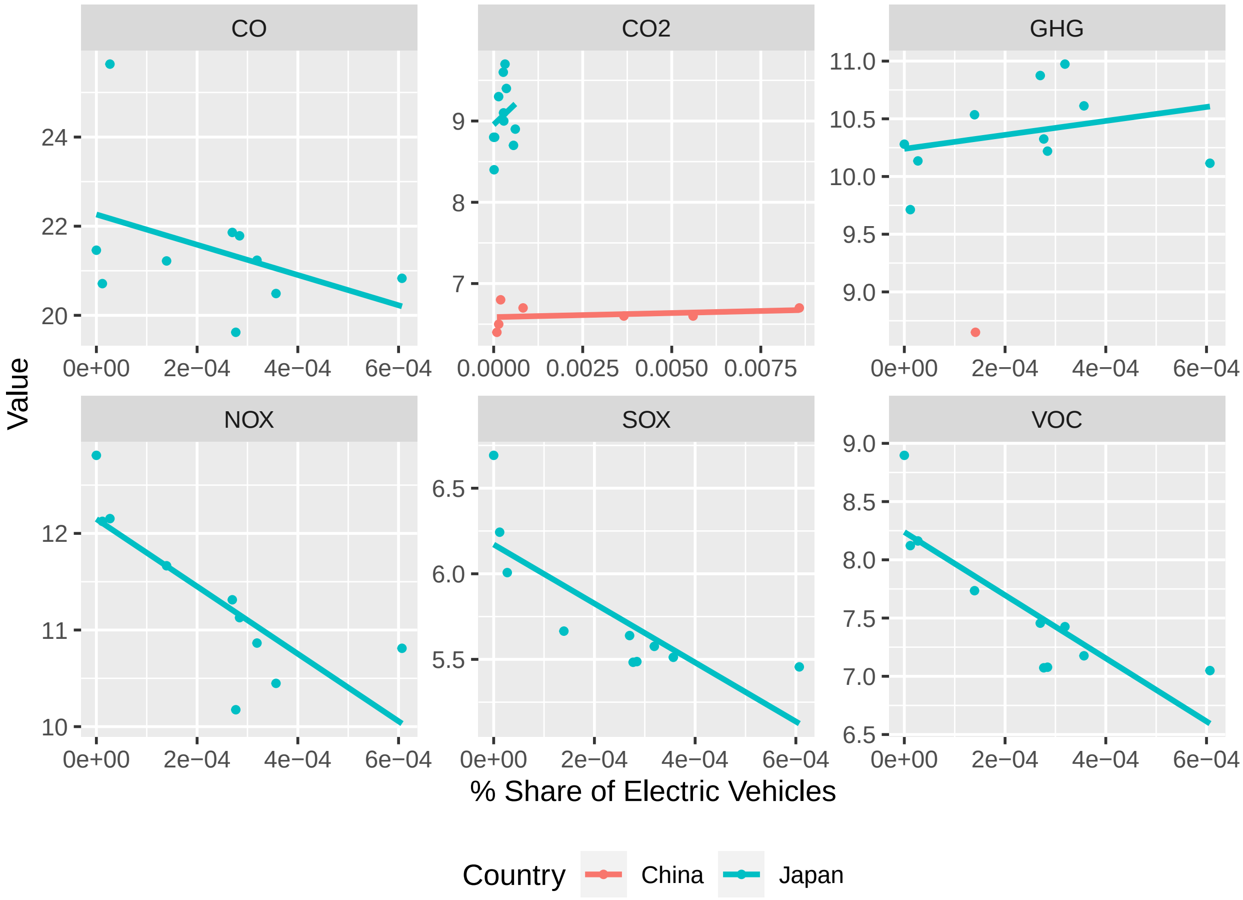


Figure 11: Pollutant levels against percent share of electric vehicles in East Asia

Figure 12 illustrates the CO_2 data against the percent share of electric vehicles for Japan. Unlike the other pollutants in Japan, the points for CO_2 followed a quadratic trend. Thus, a quadratic model would be more suitable for CO_2 while linear models were suitable for the rest.

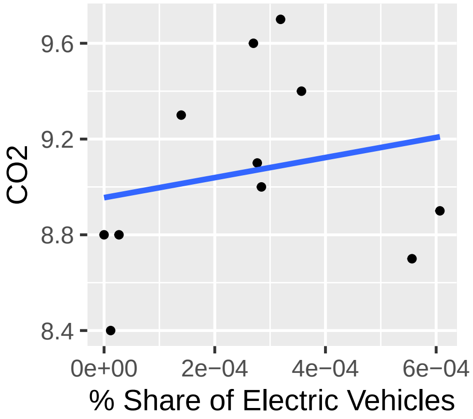


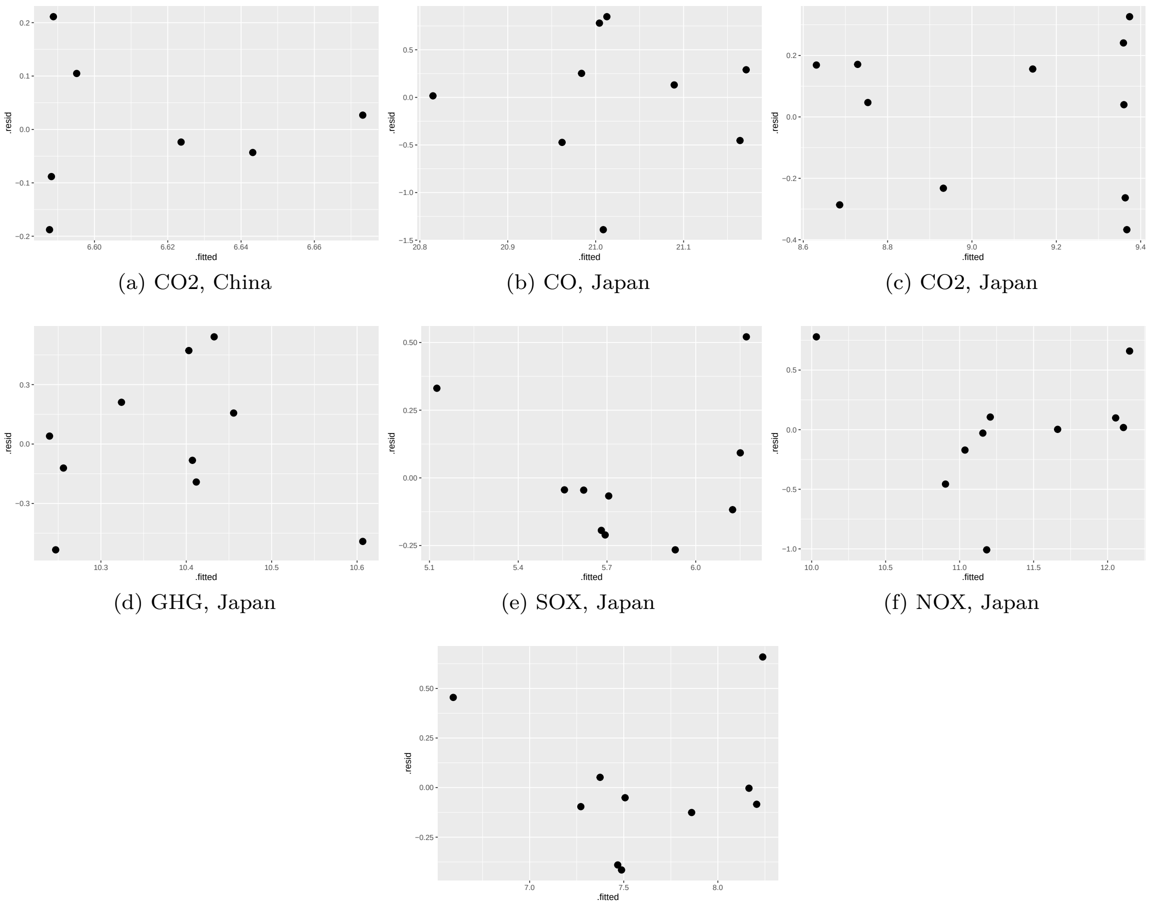
Figure 12: CO2 levels against percent share of electric vehicles in Japan

A quadratic model was fitted for CO_2 in Japan and linear models were fitted for all the other pollutants. Note that the outlier was removed prior to fitting the linear models for CO. For China, a linear model was fitted was CO_2. A summary of these models are shown in Table 6 below. The R^2 values are poor for all of the models, with 63% as the highest R^2 value for the CO_2, SO_x, and NO_x models in Japan. This suggests that the models do not explain for a lot of the variability in the data and may not accurately describe the true trends between electric vehicles and each of the pollutants.

Table 6: Rational model for Canada for CO

Country	**Pollutant**	**Intercept**	**X.Coef**	**X.squared.Coef**	**R.squared**
China	CO2	6.59	10.06	N/A	0.06
Japan	CO	21.17	-587.13	N/A	0.02
Japan	CO2	8.63	4729.13	-7527509.52	0.63
Japan	GHG	10.24	604.49	N/A	0.09
Japan	SOX	6.17	-1723.86	N/A	0.63
Japan	NOX	12.15	-3487.62	N/A	0.63
Japan	VOC	8.24	-2709.28	N/A	0.70

To check if better models could be fitted, the residuals were plotted against fitted values as shown in Figure 13.



(a) CO2, China

(b) CO, Japan

(c) CO2, Japan

(d) GHG, Japan

(e) SOX, Japan

(f) NOX, Japan

(g) VOC, Japan

Figure 13: Residuals vs. fitted plots for models (a) to (g) for China and Japan

For China, the points appeared to have a fanning effect, which indicated that there were differences in variability for each of the points. This meant that the homoscedasticity assumption in linear regression modelling was violated and so a linear model was not suitable for CO_2 in China. Thus, a model was not selected for China.

For Japan, the points on the CO, CO_2, and GHG plots were random which meant that the fitted models were suitable for the data. The points ranged from -1.5 to 1, -0.4 to 0.3, and -0.5 to 0.5 for CO, CO_2, and GHG respectively. This means there could be skewness for CO and CO_2. While the points for CO and CO_2 did not perfectly surround 0, it was still considered acceptable overall. However, the R^2 values were very low for these models, especially for CO and GHG. Hence, the the models may not be very reliable overall since they do not explain the variability in the data very well. The only model that could be considered decent is the quadratic model for CO_2 because the R^2 is 63%. For SO_x, NO_x, and VOC, the points did not seem to follow any particular trend. Although, they don't seem to be purely random either. This could be due to outliers, such as the point on the upper left-hand corner for NO_x, or it might just mean there is heterogeneity. Thus, linear models were not suitable for these pollutants. Due the time constraints of this study, no further actions were done to model these pollutants.

The only model that was considered decent among the models for China and Japan was the quadratic model for CO_2. This is illustrated in Figure 14 below.

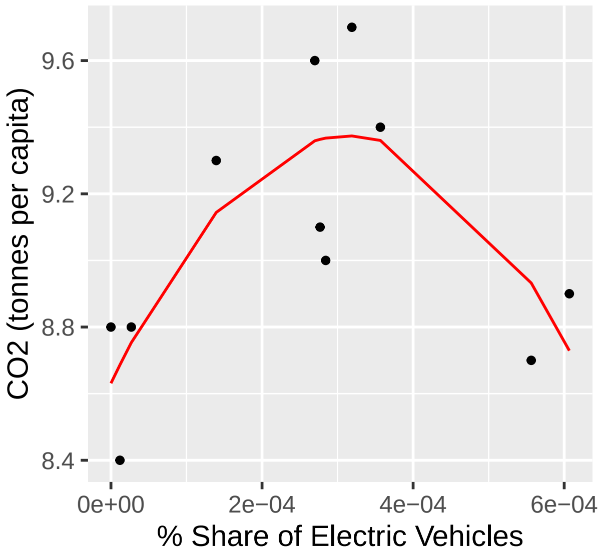


Figure 14: Regression models for CO2 in Japan

Overall, the regression model for Japan showed that an increase in the percent share of electric vehicles is associated with an initial increase and subsequent decrease in CO_2 in Japan.

Europe

As mentioned earlier, many of the data points for Europe were found to be near zero in terms of the percent share of the electric vehicles. It was suspected that this may be caused by countries that do not have any significant increase in their percent share of electric vehicles over time. To determine these countries, Europe was divided into four regions. Based on Figure 15, the points for Southern and Eastern Europe are mostly near zero in terms of the percent share of electric vehicles. Northern and Western Europe have greater percentages of electric vehicles but there is a fair bit of variation in the points. This suggests that it may be more reasonable to build models only for Northern and Western Europe, but only after potential outliers are removed.

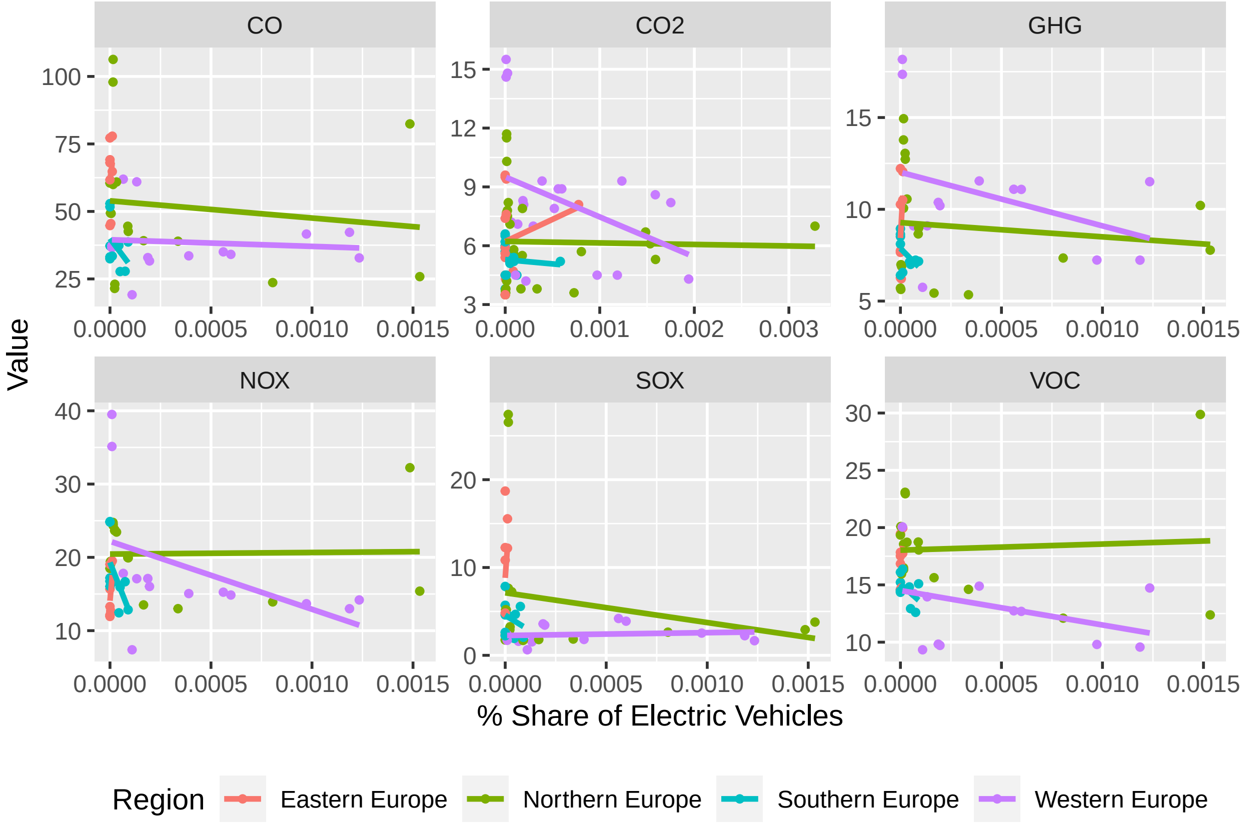


Figure 15: Pollutant levels against percent share of electric vehicles by region in Europe

In Figure 16, the data for Northern Europe was plotted again in attempt to identify outliers. Based on the plot, there did not appear to be any particular country that was producing outliers in the plots. It could be argued that Norway and Estonia are producing outliers, but they also seem to be fitting with the rest of the data as seen in the CO, CO_2, and GHG plots. Moreover, there was a lot of variability in the data and this variability did not seem equal for all points either. In other words, the homoscedasticity assumption in linear regression was violated for Northern Europe and so, a model could not be built for the region.

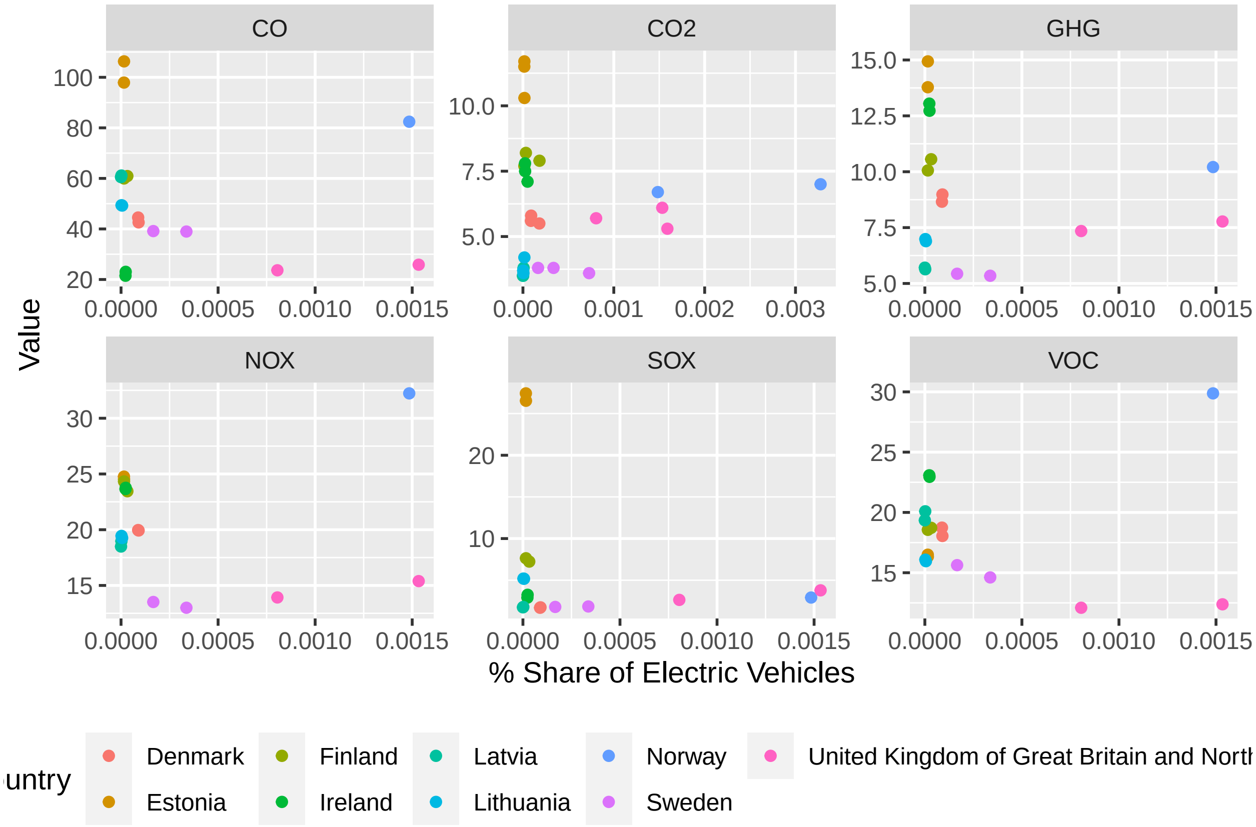


Figure 16: Pollutant levels against percent share of electric vehicles in Northern Europe

In a similar manner, the points for Western Europe were also plotted again as shown in Figure 17. For Western Europe, there was less variability in the data. However, the points for each country seemed to cluster with themselves. This suggested that a relationship between the percent share of electric vehicles and the pollutant levels do not exist for this data. Instead, the countries were just very different from one another and so building a model for Western Europe was not suitable either.

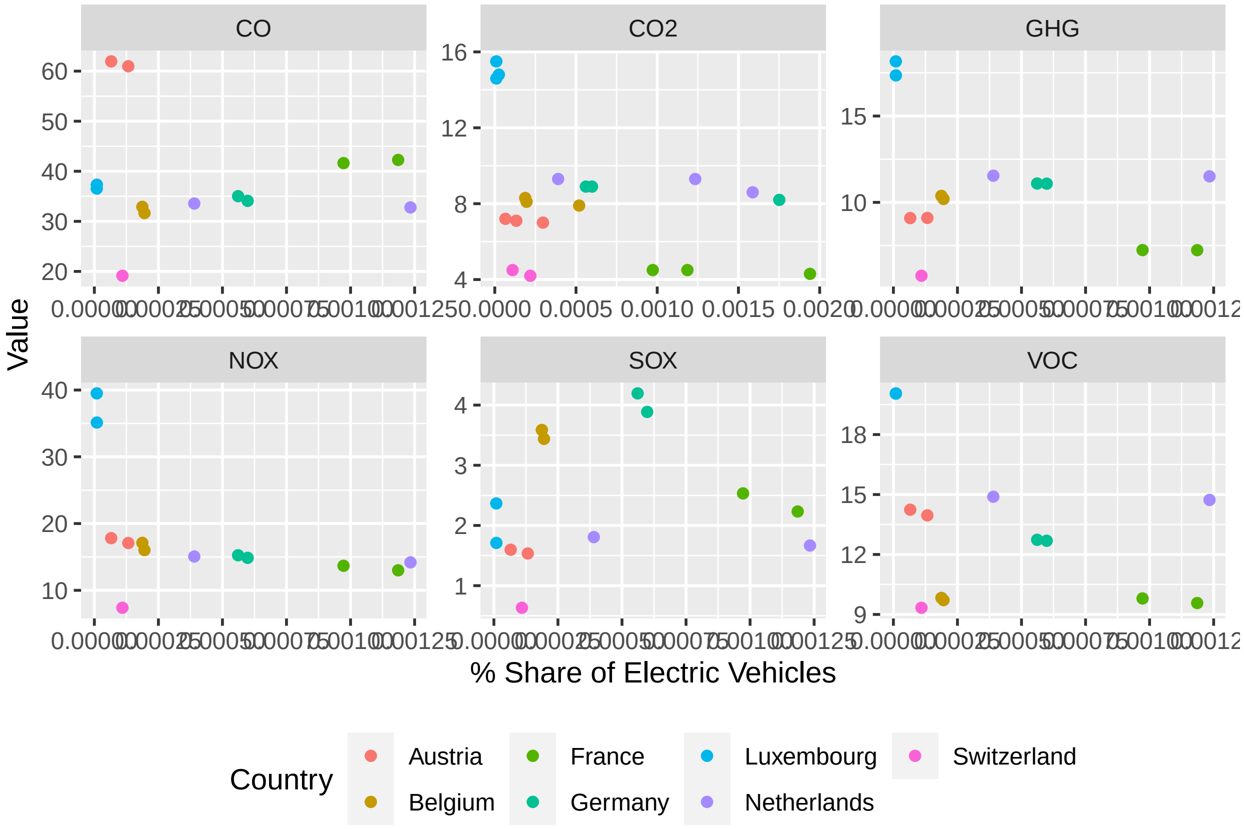


Figure 17: Pollutant levels against percent share of electric vehicles in Western Europe

Ultimately, no conclusions could be made for Europe because there was not enough data to observe a proper trend.

Electric Vehicles and Health

To examine the relationship between electric vehicles and health, the percent share of electric vehicles were plotted against the ambient air pollution attributable DALYs and death rate per 100,000 population as shown in Figure 18. Note that the WHO regions were used in this analysis since these regions are typically used by the WHO to report and analyze health data (World Health Organization, n.d.-a).

According to Figure 18, there was a fanning effect shown for both the DALYs plot and the death rate plot. This indicates that linear regression modelling is not suitable since the homoscedasticity assumption was violated. Nonetheless, the data points did generally decrease as the percent of electric vehicles increased. This suggested that countries with a greater share of electric vehicles tend to have lower DALYs and death rates per 100,000 population. By region, Europe followed this general trend, likely because it consists of most of the points. However, the trend cannot be confirmed for the Americas and the Western Pacific since those regions only consist of 3-4 countries each. China is the only country with a considerably higher percent of electric vehicles based on the data. However, its DALYs and death rate per 100,000 population does not follow the apparent trend since it is very high, if not the highest, among all the other countries. In other words, China is an outlier.

According to Figure 18, a fanning effect was observed for both the DALYs plot and the death rate plot. This indicates that regression modelling was not suitable for this data since the homoscedasticity assumption for linear regression modelling was violated. Nonetheless, a general decreasing trend was observed for attributable DALYs and death rate per 100,000 population, based on the scatter plot. In other words, this suggests that countries with a greater share of electric vehicles tend to have lower DALYs and death rates per 100,000 population overall.

By region, Europe was seen to follow this general trend. However, this was likely due to the fact that most points on the graph were from Europe. For the Americas and the Western Pacific, the general trend could not be distinguished because there were only three to four countries of data for each region. China was observed as an outlier for this trend because it has both a very high percent share of electric vehicles and high DALYs and death rate per 100,000 population.

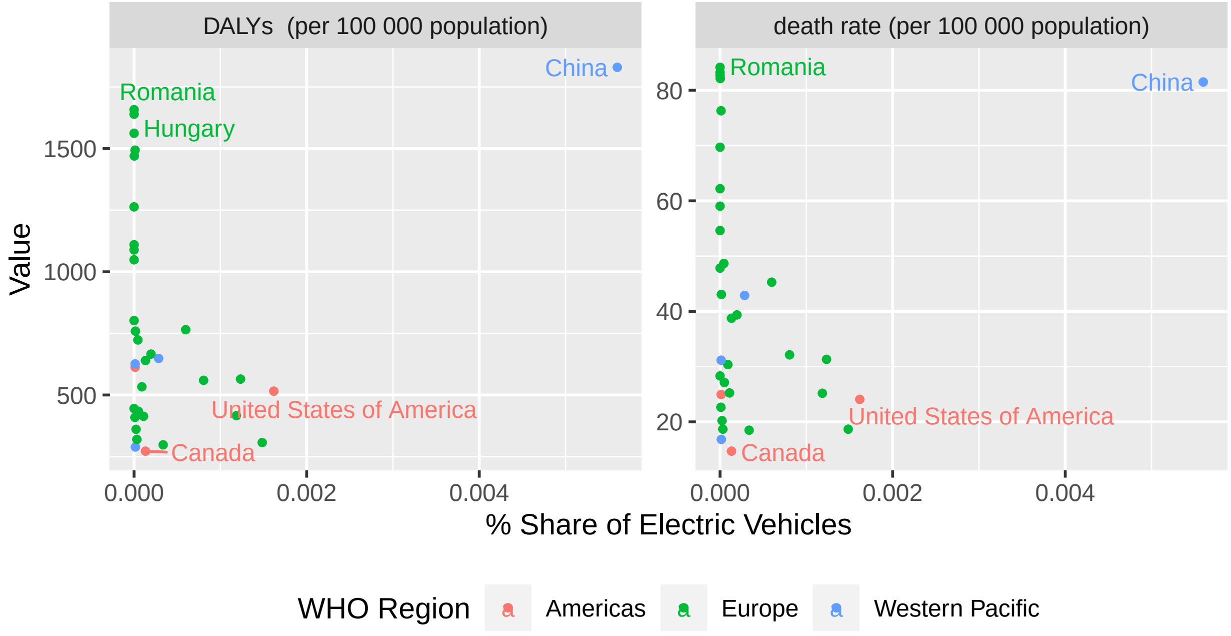


Figure 18: Pollutants against % share of electric vehicles

Discussion and Conclusions

Based on the pollution and health analyses, evidence shows that there is an association between electric vehicles, pollution, and health. In terms of pollution, regression models have indicated that pollutant levels decrease as the percent share of electric vehicles increase. The rate, form, and strength of this decrease varies from country to country and was even found to differ among countries that share similar characteristics, such as Canada and the United States. Since the fitted models explain the trends well, especially for Canada and the United States, it is suggested that greater efforts should be made to push more electric vehicles on the market since evidence shows that doing so could result in a decline of ambient air pollution. Note that this should be confirmed for a given country first because rate, form, and strength of this decrease varies.

In terms of health, the analysis has shown that countries tend to have lower DALYs and death rates per 100,000 population when there is a greater percent share of electric vehicles. This indicates that a population is generally healthier when a country has more electric vehicles, though large variations do exist between the countries observed. Because only one year of data was available for the health measures, it is difficult to conclude whether the observed trend applies for years other than 2016 or if an increase in electric vehicles truly results in a healthier population within a country. For example, if a country manages to increase its percent share of electric vehicles, it is unknown whether DALYs and death rates per 100,000 population will actually decrease for the given country. It is possible that doing so would have no significant effect, and that the trend observed in this study was due to factors that were unaccounted for, such as technological advancements, population demographics, or laws respective to a country. In other words, causal inference cannot be made with this analysis because it is unknown whether electric vehicles truly have an effect on health or if other factors are producing this trend. Further studies are required to better understand the relationship between electric vehicles and health in order to draw conclusions of causal inference.

Nevertheless, the results in this study should be carefully interpreted because a number of limitations have been encountered. Firstly, there was limited access to data for each of the variables of interest. Since the study used open source data, data was often missing, insufficient in amount, or inconsistent. One important example of this is the differences in the vehicle data; some countries had sales data while other countries had registration data. This complicated the analysis because sales data only contains a subset of the vehicles in a country while registration data contains all of the vehicles. Thus, the true trends could be different than the ones observed in the study. In general, data that is missing, insufficient, and inconsistent results in complications and inaccuracies with analyzing the data. If funding was available for this study, these issues can be avoided by buying full data sets from a select number of websites. Moreover, a translator could be hired to obtain and understand data in other languages if necessary.

Secondly, there was a time constraint of three months for the study. This meant data collection, analysis, and report writing had to be strategically completed so that the study could be done within the three months. Thus, only a limited amount of analysis and conclusions could be drawn, even if there was a lot of data collected. If more time was available for the study, it would have been interesting to include the p-values and confidence intervals for each of the linear model estimates and evaluate their significance in the report. For the quadratic and rational models, the p-values for each term could be investigated to test for significance as well. Additionally, if more time was available, extrapolations could be done for the models that were a good fit. Lastly, more of the data could have been used if there was more time. For example, the impact of electric vehicles on the burden of various diseases could have been studied.

Finally, the last limitation in this study involves the fact that the countries included in the study were not selected randomly. They were mainly selected for convenience or out of interest. In this way, the trends observed in the study may not be generalizable to other countries.

For future studies, it is recommended for this topic to be investigated for a longer period of time and that a source of funding is available to refine the data collected. While it is ideal for randomization to be done when selecting the countries, this will likely be unrealistic since many countries do not use electric vehicles or do not have the means to record vehicle data at a population level. Thus, countries would still need to be selected, at least until all countries start using electric vehicles. If the time constraints and lack of funding cannot be changed, then decreasing the scope of the study to a single region or country may help to simplify data collection and allow for a more thorough analysis to be made.

Appendices

Appendix A - List of Countries

The following is a list of all the countries used in this study in alphabetical order:
Australia
Austria
Belgium
Brazil
Canada
China
Croatia
Czechia
Denmark
Estonia
Finland
France
Germany
Greece
Hungary
Ireland
Italy
Japan
Latvia
Lithuania
Luxembourg
Netherlands
Norway
Poland
Portugal
Republic of Korea
Romania
Slovakia
Slovenia
Spain
Sweden
Switzerland
United Kingdom of Great Britain and Northern Ireland
United States of America

Appendix B - Estimating Other Fuel Cars for Japan

JAMA provides data on new passenger car registrations including the number of diesel cars, the number of *next-generation* cars, and the percent share of *next-generation* cars for the years 2008 to 2018. Using this data, the number of cars that use fuels other than diesel and *next-generation* fuels (hybrid, plug-in hybrid, electric, fuel cell, clean diesel) can be computed as shown below.

Let d be the number of diesel cars. Let g be the number of *next-generation* cars. Let p be the percent share of *next-generation* cars. Let x be the number of cars that use fuels other than diesel and *next-generation* fuels.

For 2008, $d = 1345133$, $g = 108518$, and $p = 0.026$. Then,

$$\begin{aligned} 108518 &= 0.026x \\ x &= 108518/0.026 \\ &\approx 4173769.23 \\ &\approx 4173770 \end{aligned}$$

there is a total of 4173770 new registrations for passenger cars. Given d and g,

$$108518 + 1345133 = 1453651$$

there are 1453651 cars using diesel and *next-generation* fuels. Then using this value,

$$4173770 - 1453651 = 2720119$$

there are 2720119 cars that use fuels other than diesel and *next-generation* fuels. Likewise was done for 2009 to 2018.

Appendix C - R Code Used in Analysis

```r
knitr::opts_chunk$set(echo=F, warning = F, message = F)

library(knitr)
library(kableExtra)
library(tidyverse)
library(readxl)
library(ggrepel)

# Load the (cleaned) data and compute the % share of electric vehicles.
ALLdata_raw <- read_excel("DataPrep/ALLdata.xlsx")
ALLdata_raw %>%
  mutate("%Electric"=100*Electric/sum(ifelse(is.na(Petrol),0,Petrol),
                                      ifelse(is.na(Diesel),0,Diesel),
                                      ifelse(is.na(Hybrid),0,Hybrid),
                                      ifelse(is.na(Electric),0,Electric),
                                      ifelse(is.na(`LPG+NaturalGas`),0,`LPG+NaturalGas`),
                                      ifelse(is.na(Other),0,Other)),
         "%Hybrid"=100*Hybrid/sum(ifelse(is.na(Petrol),0,Petrol),
                                  ifelse(is.na(Diesel),0,Diesel),
                                  ifelse(is.na(Hybrid),0,Hybrid),
                                  ifelse(is.na(Electric),0,Electric),
                                  ifelse(is.na(`LPG+NaturalGas`),0,`LPG+NaturalGas`),
                                  ifelse(is.na(Other),0,Other))) %>%
  select(Country:Other,`%Electric`,`%Hybrid`,VehicleType:Value) ->
  ALLdata

# Dataframe of electric vehicles and pollution
ALLdata %>%
  select(-c(Sex:`Total_death rate (per 100 000 population)`)) %>%
  unique() %>%
  filter(!is.na(Pollutant)) %>%
  filter(!is.na(Value)) ->
  electricPoll
```

```r
# Dataframe of electric vehicles and pollution for each region (both WHO and new)
electricPoll %>%
  select(-Measure) %>%
  pivot_wider(names_from = Pollutant,values_from = Value) -> electricPoll_wider

electricPoll_wider %>% filter(`Region`=="North America") -> electricPoll_NA
electricPoll_NA %>% filter(Country=="Canada") -> electricPoll_NA_CAN
electricPoll_NA %>% filter(Country=="United States of America") -> electricPoll_NA_US

electricPoll_wider %>% filter(`Region`=="East Asia") -> electricPoll_EA
electricPoll_EA %>% filter(Country=="China") -> electricPoll_EA_CHN
electricPoll_EA %>% filter(Country=="Japan") -> electricPoll_EA_JP

electricPoll_wider %>% filter(`Region`=="Northern Europe") -> electricPoll_NE
electricPoll_wider %>% filter(`Region`=="Western Europe") -> electricPoll_WE
remove(electricPoll_wider)

# Dataframe of electric vehicles and health
ALLdata %>%
  select(Country:`Total_death rate (per 100 000 population)`) %>%
  unique() %>%
  filter(!is.na(Sex)) %>%
  pivot_longer("Ischaemic heart disease_DALYs  (per 100 000 population)":"Total_death rate (per 100 000
               names_to="Disease_Measure", values_to="Value") %>%
  separate(Disease_Measure,into=c("Disease","Measure"),sep="_") ->
  electricHlth

# Scatter plots of % share of electric vehicles against value of pollutant
ggplot(electricPoll,aes(x=`%Electric`,y=Value)) +
  geom_point() +
  facet_wrap(~Pollutant, scales="free") +
  labs(x="% Share of Electric Vehicles")

# Table of the countries classified by WHO region
kable(rbind(
  c("WHO Region", "Countries"),
  c("Americas", "Brazil, Canada, United States"),
  c("Western Pacific", "Australia, China, Japan, South Korea"),
  c("Europe", "27 European countries")),
  format="latex",
  caption = "Countries classified by WHO region") %>%
  kable_styling(latex_options = c("HOLD_position")) %>%
  row_spec(0, bold = T, background = "#dceaf2")

# Scatter plots of % electric vehicles against value of pollutant with WHO regions labelled
ggplot(electricPoll,aes(x=`%Electric`,y=Value,colour=`WHO Region`)) +
  geom_point() +
  facet_wrap(~Pollutant, scales="free") +
  labs(x="% Share of Electric Vehicles") +
  scale_fill_continuous(guide = guide_legend()) +
  theme(legend.position="bottom")

# Dataframe for electric vehicles and pollution for the Americas WHO region
```

```r
electricPoll %>%
  select(-Measure) %>%
  pivot_wider(names_from = Pollutant,values_from = Value) %>%
  filter(`WHO Region`=="Americas") ->
  electricPoll_A

# Scatter plot showing Brazil as an outlier in the Americas WHO region
ggplot(electricPoll_A,aes(x=`%Electric`,y=CO2,colour=Country)) +
  geom_point() +
  labs(x="% Share of Electric Vehicles")

# Dataframe for electric vehicles and pollution for the Western Pacific WHO region
electricPoll %>% filter(`WHO Region`=="Western Pacific") -> electricPoll_WP

# Scatter plot of the Western Pacific to check what country the outliers are from
ggplot(electricPoll_WP,aes(x=`%Electric`,y=Value,colour=`Country`)) +
  geom_point() +
  facet_wrap(~Pollutant, scales="free") +
  labs(x="% Share of Electric Vehicles") +
  scale_fill_continuous(guide = guide_legend()) +
  theme(legend.position="bottom")

# Table of the countries using the new classification
kable(rbind(
  c("Region", "Countries"),
  c("North America", "Canada, United States"),
  c("East Asia", "China, Japan"),
  c("Northern Europe", "Denmark, Estonia, Finland, Ireland, Latvia, Lithuania, Norway, Sweden, United Ki
  c("Southern Europe", "Croatia, Greece, Italy, Portugal, Slovenia, Spain"),
  c("Eastern Europe", "Czechia, Hungary, Poland, Romania, Slovakia"),
  c("Western Europe", "Austria, Belgium, Germany, Luxembourg, Netherlands, Switzerland")),
  format="latex",
  caption = "Reclassification of the countries") %>%
  kable_styling(latex_options = c("HOLD_position", "scale_down")) %>%
  row_spec(0, bold = T, background = "#dceaf2")

# Dataframes for the new regions by category (North America, East Asia, Europe)
electricPoll %>%
  filter(Country=="Canada" | Country=="United States of America") -> electricPoll_America
electricPoll %>%
  filter(Country=="China" | Country=="Japan") -> electricPoll_Asia
electricPoll %>%
  filter(`WHO Region`=="Europe") -> electricPoll_Europe
electricPoll %>%
  filter(`Region`=="Northern Europe") -> electricPoll_EuropeN
electricPoll %>%
  filter(`Region`=="Western Europe") -> electricPoll_EuropeW

# Scatter plot for North America
ggplot(electricPoll_America,aes(x=`%Electric`,y=Value,colour=Country)) +
  geom_point() +
  geom_smooth(method="lm",se=F) +
  facet_wrap(~Pollutant, scales="free") +
  labs(x="% Share of Electric Vehicles") +
  scale_fill_continuous(guide = guide_legend()) +
  theme(legend.position="bottom")

# Linear regression models for Canada by pollutant
NA_CAN_CO_1 <- lm(CO~`%Electric`,electricPoll_NA_CAN)
NA_CAN_CO2_1 <- lm(CO2~`%Electric`,electricPoll_NA_CAN)
NA_CAN_GHG_1 <- lm(GHG~`%Electric`,electricPoll_NA_CAN)
NA_CAN_SOX_1 <- lm(SOX~`%Electric`,electricPoll_NA_CAN)
NA_CAN_NOX_1 <- lm(NOX~`%Electric`,electricPoll_NA_CAN)
NA_CAN_VOC_1 <- lm(VOC~`%Electric`,electricPoll_NA_CAN)

# Linear regression models for United States by pollutant
NA_US_CO_1 <- lm(CO~`%Electric`,electricPoll_NA_US)
NA_US_CO2_1 <- lm(CO2~`%Electric`,electricPoll_NA_US)
NA_US_GHG_1 <- lm(GHG~`%Electric`,electricPoll_NA_US)
NA_US_SOX_1 <- lm(SOX~`%Electric`,electricPoll_NA_US)
NA_US_NOX_1 <- lm(NOX~`%Electric`,electricPoll_NA_US)
NA_US_VOC_1 <- lm(VOC~`%Electric`,electricPoll_NA_US)

# All linear models for Canada and United States
NAlinearMods <- list(NA_CAN_CO_1,NA_CAN_CO2_1,NA_CAN_GHG_1,NA_CAN_SOX_1,NA_CAN_NOX_1,
                     NA_CAN_VOC_1,NA_US_CO_1,NA_US_CO2_1,NA_US_GHG_1,NA_US_SOX_1,
                     NA_US_NOX_1,NA_US_VOC_1)

# Function that creates summary columns (R^2 vals or coefficients)
# If `coef` is "R^2" then it will compute the R^2 values
# If `coef` is an int, it will compute the coefficients located at index `coef`
# `models` is a list of all the regression models to create summaries for
extract_coefs <- function(coef, models, digits=2){
  summary_col <- c()
  if(coef=="R^2"){
    for(model in models){
      summary_col <- c(summary_col, round(summary(model)$r.squared,digits))
    }
  } else {
    for(model in models){
      summary_col <- c(summary_col, round(summary(model)$coefficients[coef],digits))
    }
  }
  return(summary_col)
}

# Table that summarizes Canada and United States linear models
NAlinearTable <- data.frame(
  Country = rep(c("Canada", "United States"),each=6),
  Pollutant = rep(c("CO","CO2","GHG","SOX","NOX","VOC"),2),
  Intercept = extract_coefs(1,NAlinearMods),
  "X-Coef" = extract_coefs(2,NAlinearMods),
  R.squared = extract_coefs("R^2",NAlinearMods)
)

kable(NAlinearTable,
  format="latex",
  caption = "Linear models for Canada and the United States for each pollutant") %>%
  kable_styling(latex_options = c("HOLD_position")) %>%
  row_spec(0, bold = T, background = "#dceaf2")

# Residuals vs. fitted plots for Canada's linear models
ggplot(NA_CAN_CO_1,aes(y=.resid, x=.fitted))+geom_point(size=3.5)
ggplot(NA_CAN_CO2_1,aes(y=.resid, x=.fitted))+geom_point(size=3.5)
ggplot(NA_CAN_GHG_1,aes(y=.resid, x=.fitted))+geom_point(size=3.5)
ggplot(NA_CAN_SOX_1,aes(y=.resid, x=.fitted))+geom_point(size=3.5)
ggplot(NA_CAN_NOX_1,aes(y=.resid, x=.fitted))+geom_point(size=3.5)
ggplot(NA_CAN_VOC_1,aes(y=.resid, x=.fitted))+geom_point(size=3.5)

# Residuals vs. fitted plots for United States for US' linear models
ggplot(NA_US_CO_1,aes(y=.resid, x=.fitted))+geom_point(size=3.5)
ggplot(NA_US_CO2_1,aes(y=.resid, x=.fitted))+geom_point(size=3.5)
ggplot(NA_US_GHG_1,aes(y=.resid, x=.fitted))+geom_point(size=3.5)
ggplot(NA_US_SOX_1,aes(y=.resid, x=.fitted))+geom_point(size=3.5)
ggplot(NA_US_NOX_1,aes(y=.resid, x=.fitted))+geom_point(size=3.5)
ggplot(NA_US_VOC_1,aes(y=.resid, x=.fitted))+geom_point(size=3.5)

# Quadratic models for CO and NOX in Canada
NA_CAN_CO_2 <- lm(CO~`%Electric`+I(`%Electric`^2),electricPoll_NA_CAN)
NA_CAN_NOX_2 <- lm(NOX~`%Electric`+I(`%Electric`^2),electricPoll_NA_CAN)
NAquadraticMods <- list(NA_CAN_CO_2, NA_CAN_NOX_2)

# Table that summarizes Canada quadratic models
NAquadraticTable <- data.frame(
  Country = rep("Canada",each=2),
  Pollutant = c("CO","NOX"),
  Intercept = extract_coefs(1,NAquadraticMods),
  "X-Coef" = extract_coefs(2,NAquadraticMods),
  "X.squared-Coef" = extract_coefs(3,NAquadraticMods),
  R.squared = extract_coefs("R^2",NAquadraticMods)
)

kable(NAquadraticTable,
  format="latex",
  caption = "Quadratic models for Canada for CO and NOX") %>%
  kable_styling(latex_options = c("HOLD_position")) %>%
  row_spec(0, bold = T, background = "#dceaf2")

# Residuals vs. fitted plots for Canada's quadratic models
ggplot(NA_CAN_CO_2,aes(y=.resid, x=.fitted))+geom_point(size=3.5)
ggplot(NA_CAN_NOX_2,aes(y=.resid, x=.fitted))+geom_point(size=3.5)

# Rational model for CO in Canada
NA_CAN_CO_3 <- lm(CO~`%Electric`+I(`%Electric`^(-1)),electricPoll_NA_CAN)
NArationalMod <- list(NA_CAN_CO_3)

# Table that summarizes Canada's rational model
NArationalTable <- data.frame(
  Country = c("Canada"),
  Pollutant = c("CO"),
  Intercept = extract_coefs(1,NArationalMod),
  "X-Coef" = extract_coefs(2,NArationalMod),
  "one.over.X-Coef" = extract_coefs(3,NArationalMod,7),
  R.squared = extract_coefs("R^2",NArationalMod)
)

kable(NArationalTable,
  format="latex",
  caption = "Rational model for Canada for CO") %>%
  kable_styling(latex_options = c("HOLD_position")) %>%
  row_spec(0, bold = T, background = "#dceaf2")

# Residuals vs. fitted plots for Canada's rational model
ggplot(NA_CAN_CO_3,aes(y=.resid, x=.fitted))+geom_point(size=3.5)

# Dataframes for each country-pollutant combination that has a decent model for Canada and US
electricPoll_America %>%
  filter(Country=="Canada", Pollutant=="CO") -> electricPoll_CANCO
electricPoll_America %>%
  filter(Country=="Canada", Pollutant=="NOX") -> electricPoll_CANNOX
electricPoll_America %>%
  filter(Country=="United States of America", Pollutant=="CO") -> electricPoll_USCO
electricPoll_America %>%
  filter(Country=="United States of America", Pollutant=="GHG") -> electricPoll_USGHG
electricPoll_America %>%
  filter(Country=="United States of America", Pollutant=="SOX") -> electricPoll_USSOX
electricPoll_America %>%
  filter(Country=="United States of America", Pollutant=="NOX") -> electricPoll_USNOX
electricPoll_America %>%
  filter(Country=="United States of America", Pollutant=="VOC") -> electricPoll_USVOC

# Plot the final models for Canada and US on a scatter plot of the data
ggplot(electricPoll_CANCO,aes(x=`%Electric`,y=Value)) +
  geom_point(size=3.5) +
  labs(x="% Share of Electric Vehicles", y="CO (kilograms per capita)") +
  geom_line(data = NA_CAN_CO_3, aes(y = .fitted), colour="red")

ggplot(electricPoll_CANNOX,aes(x=`%Electric`,y=Value)) +
  geom_point(size=3.5) +
  labs(x="% Share of Electric Vehicles", y="NOX (kilograms per capita)") +
  geom_line(data = NA_CAN_NOX_2, aes(y = .fitted), colour="red")

ggplot(electricPoll_USCO,aes(x=`%Electric`,y=Value)) +
  geom_point(size=3.5) +
  labs(x="% Share of Electric Vehicles", y="CO (kilograms per capita)") +
  geom_line(data = NA_US_CO_1, aes(y = .fitted), colour="red")

ggplot(electricPoll_USGHG,aes(x=`%Electric`,y=Value)) +
  geom_point(size=3.5) +
  labs(x="% Share of Electric Vehicles", y="GHG (tonnes per capita)") +
  geom_line(data = NA_US_GHG_1, aes(y = .fitted), colour="red")
```

```
ggplot(electricPoll_USNOX,aes(x=`%Electric`,y=Value)) +
  geom_point(size=3.5) +
  labs(x="% Share of Electric Vehicles", y="NOX (kilograms per capita)") +
  geom_line(data = NA_US_NOX_1, aes(y = .fitted), colour="red")

ggplot(electricPoll_USSOX,aes(x=`%Electric`,y=Value)) +
  geom_point(size=3.5) +
  labs(x="% Share of Electric Vehicles", y="SOX (kilograms per capita)") +
  geom_line(data = NA_US_SOX_1, aes(y = .fitted), colour="red")

ggplot(electricPoll_USVOC,aes(x=`%Electric`,y=Value)) +
  geom_point(size=3.5) +
  labs(x="% Share of Electric Vehicles", y="VOC (kilograms per capita)") +
  geom_line(data = NA_US_VOC_1, aes(y = .fitted), colour="red")

# Scatter plot for East Asia
ggplot(electricPoll_Asia,aes(x=`%Electric`,y=Value, colour=Country)) +
  geom_point(size=1) +
  geom_smooth(method="lm",se=F) +
  facet_wrap(~Pollutant, scales="free") +
  labs(x="% Share of Electric Vehicles") +
  scale_fill_continuous(guide = guide_legend()) +
  theme(legend.position="bottom")

# Dataframe of electric vehicles and pollution for Japan
electricPoll_Asia %>%
  select(-Measure) %>%
  pivot_wider(names_from = Pollutant,values_from = Value) %>%
  filter(Country=="Japan") ->
  electricPoll_Japan

# Scatter plot of CO2 in Japan
ggplot(electricPoll_Japan,aes(x=`%Electric`, y=CO2)) +
  geom_point(size=1) +
  geom_smooth(method="lm",se=F) +
  labs(x="% Share of Electric Vehicles")

# Dataframe with removed outlier for CO
electricPoll_EA_JP %>% filter(CO <= 24) -> electricPoll_EA_JP_noOutlier

# Regression models for China and Japan
EA_CHN_CO2_1 <- lm(CO2~`%Electric`,electricPoll_EA_CHN)
EA_JP_CO_1 <- lm(CO~`%Electric`,electricPoll_EA_JP_noOutlier)
EA_JP_CO2_1 <- lm(CO2~`%Electric`+I(`%Electric`^2),electricPoll_EA_JP)
EA_JP_GHG_1 <- lm(GHG~`%Electric`,electricPoll_EA_JP)
EA_JP_SOX_1 <- lm(SOX~`%Electric`,electricPoll_EA_JP)
EA_JP_NOX_1 <- lm(NOX~`%Electric`,electricPoll_EA_JP)
EA_JP_VOC_1 <- lm(VOC~`%Electric`,electricPoll_EA_JP)

# All models for China and Japan
EAmodels <- list(EA_CHN_CO2_1,EA_JP_CO_1,EA_JP_CO2_1,EA_JP_GHG_1,EA_JP_SOX_1,
                 EA_JP_NOX_1,EA_JP_VOC_1)
```

```r
# Table that summarizes the models for China and Japan
EAmodelsTable <- data.frame(
  Country = c("China",rep(c("Japan"),6)),
  Pollutant = c("CO2", "CO", "CO2", "GHG", "SOX", "NOX", "VOC"),
  Intercept = extract_coefs(1,EAmodels),
  "X-Coef" = extract_coefs(2,EAmodels),
  "X.squared-Coef" = c(rep(c("N/A"),2),extract_coefs(3,list(EA_JP_CO2_1)),rep(c("N/A"),4)),
  R.squared = extract_coefs("R^2",EAmodels)
)

kable(EAmodelsTable,
  format="latex",
  caption = "Rational model for Canada for CO") %>%
  kable_styling(latex_options = c("HOLD_position")) %>%
  row_spec(0, bold = T, background = "#dceaf2")

# Residuals vs. fitted plots for China and Japan's models
ggplot(EA_CHN_CO2_1,aes(y=.resid, x=.fitted))+geom_point(size=3.5)
ggplot(EA_JP_CO_1,aes(y=.resid, x=.fitted))+geom_point(size=3.5)
ggplot(EA_JP_CO2_1,aes(y=.resid, x=.fitted))+geom_point(size=3.5)
ggplot(EA_JP_GHG_1,aes(y=.resid, x=.fitted))+geom_point(size=3.5)
ggplot(EA_JP_SOX_1,aes(y=.resid, x=.fitted))+geom_point(size=3.5)
ggplot(EA_JP_NOX_1,aes(y=.resid, x=.fitted))+geom_point(size=3.5)
ggplot(EA_JP_VOC_1,aes(y=.resid, x=.fitted))+geom_point(size=3.5)

# Dataframes for the only decent model in Japan
electricPoll_Asia %>% filter(Country=="Japan", Pollutant=="CO2") -> electricPoll_JPCO2

# Plot the final model for Japan on a scatter plot of the data
ggplot(electricPoll_JPCO2,aes(x=`%Electric`,y=Value)) +
  geom_point(size=1) +
  labs(x="% Share of Electric Vehicles", y="CO2 (tonnes per capita)") +
  geom_line(data = EA_JP_CO2_1, aes(y = .fitted), colour="red")

# Scatter plot of Europe
ggplot(electricPoll_Europe,aes(x=`%Electric`,y=Value,colour=Region)) +
  geom_point(size=1) +
  geom_smooth(method="lm",se=F) +
  facet_wrap(~Pollutant, scales="free") +
  labs(x="% Share of Electric Vehicles") +
  scale_fill_continuous(guide = guide_legend()) +
  theme(legend.position="bottom")

# Scatter plot of Northern Europe
ggplot(electricPoll_EuropeN,aes(x=`%Electric`,y=Value,colour=Country)) +
  geom_point() +
  facet_wrap(~Pollutant, scales="free") +
  labs(x="% Share of Electric Vehicles") +
  scale_fill_continuous(guide = guide_legend()) +
  theme(legend.position="bottom")

# Scatter plot of Western Europe
ggplot(electricPoll_EuropeW,aes(x=`%Electric`,y=Value,colour=Country)) +
  geom_point() +
  facet_wrap(~Pollutant, scales="free") +
  labs(x="% Share of Electric Vehicles") +
  scale_fill_continuous(guide = guide_legend()) +
  theme(legend.position="bottom")

# Dataframe of health data for all diseases and both sexes
# Labels are also created for the upcoming graph
electricHlth %>%
  filter(Disease=="Total", Sex=="Both sexes") %>%
  mutate(Label=ifelse(`%Electric`>0.0015 | Country=="Canada" |
                        (grepl("death", Measure, fixed = TRUE) & Value>84) |
                        (grepl("DALY", Measure, fixed = TRUE) & Value>1600),
                      Country, "")) ->
  electricHlth_Total

# Scatter plot of electric vehicles against DALYs and death rate
ggplot(electricHlth_Total,aes(x=`%Electric`,y=Value, colour=`WHO Region`, label=Label)) +
  geom_point(size=1) +
  geom_text_repel(size=3) +
  facet_wrap(~Measure, scales="free") +
  labs(x="% Share of Electric Vehicles") +
  scale_fill_continuous(guide = guide_legend()) +
  theme(legend.position = "bottom", legend.box = "vertical")
```

References

Alliance of Automobile Manufacturers. (2019). *Advanced technology vehicle sales dashboard.* https://autoalliance.org/energy-environment/advanced-technology-vehicle-sales-dashboard/

CEIC Data. (n.d.). *China automobile sales: Annual.* https://www.ceicdata.com/en/china/automobile-sales-annual

European Automobile Manufacturers' Association. (2017). *ACEA report vehicles in use europe 2017.* https://www.acea.be/uploads/statistic_documents/ACEA_Report_Vehicles_in_use-Europe_2017.pdf

European Automobile Manufacturers' Association. (2018). *ACEA report vehicles in use europe 2018.* https://www.acea.be/uploads/statistic_documents/ACEA_Report_Vehicles_in_use-Europe_2018.pdf

European Automobile Manufacturers' Association. (2019). *ACEA report vehicles in use europe 2019.* https://www.acea.be/uploads/statistic_documents/ACEA_Report_Vehicles_in_use-Europe_2019.pdf

Japan Automobile Manufacturers Association. (2019a). *Motor vehicle statistics of japan.* http://www.jama-english.jp/publications/MVS2019.pdf

Japan Automobile Manufacturers Association. (2019b). *The motor industry of japan 2019.* http://www.jama-english.jp/publications/The_Motor_Industry_of_Japan_2019.pdf

Ministry of Land, Infrastructure and Transport. (n.d.). *The results of vehicle inspection.* http://stat.molit.go.kr/portal/cate/statView.do?hRsId=509&hFormId=5896&hSelectId=5896&sStyleNum=2&sStart=2016&sEnd=2017&hPoint=00&hAppr=1

National Transport Commission. (2018). *Carbon dioxide emissions intensity for new australian light vehicles 2017.*

National Transport Commission. (2019). *Carbon dioxide emissions intensity for new australian light vehicles 2018.*

National Transport Commission. (2020). *Carbon dioxide emissions intensity for new australian light vehicles 2019.*

OECD. (2020). *Air and ghg emissions (indicator).* https://doi.org/10.1787/93d10cf7-en

Statista. (2020a, January). *Annual sales volume of new energy vehicles in china from 2011 to 2019, by type.*

Statista. (2020b, January). *Number of registered new light vehicles in brazil from 2014 to 2019, by fuel type.*

Statistics Canada. (2020). *Table 20-10-0021-01 new motor vehicle registrations.* https://doi.org/10.25318/2010002101-eng

UNdata. (2020). *UNdata explorer.* http://data.un.org/Explorer.aspx

U.S. Department of Energy. (n.d.-a). *Electric vehicle benefits.* https://www.energy.gov/timeline/timeline-history-electric-car#:~:text=Around%201832%2C%20Robert%20Anderson%20develops,an%20English%20inventor%20in%201884.

U.S. Department of Energy. (n.d.-b). *Electric vehicle benefits.* https://www.energy.gov/eere/electricvehicles/electric-vehicle-benefits

U.S. Energy Information Administration. (2014). *Annual energy outlook 2014 table: Light duty vehicle sales by technology type.* https://www.eia.gov/outlooks/aeo/data/browser/#/?id=48-AEO2014&region=1-0&cases=ref2014&start=2011&end=2040&f=A&linechart=ref2014-d102413a.4-48-AEO2014.1-0~ref2014-d102413a.5-48-AEO2014.1-0~ref2014-d102413a.30-48-AEO2014.1-0~ref2014-d102413a.31-48-AEO2014.1-0&map=ref2014-d102413a.5-48-AEO2014.1-0&ctype=linechart&sourcekey=0

U.S. Energy Information Administration. (2015). *Annual energy outlook 2015 table: Light duty vehicle sales by technology type.* https://www.eia.gov/outlooks/aeo/data/browser/#/?id=48-AEO2015&region=1-0&cases=ref2015&start=2012&end=2040&f=A&linechart=ref2015-d021915a.4-48-AEO2015.1-0~ref2015-

d021915a.5-48-AEO2015.1-0~ref2015-d021915a.30-48-AEO2015.1-0~ref2015-d021915a.31-48-AEO2015.1-0&map=ref2015-d021915a.5-48-AEO2015.1-0&ctype=linechart&sourcekey=0

U.S. Energy Information Administration. (2016). *Annual energy outlook 2016 table: Light duty vehicle sales by technology type.* https://www.eia.gov/outlooks/aeo/data/browser/#/?id=48-AEO2016&region=1-0&cases=ref2016&start=2013&end=2040&f=A&linechart=ref2016-d032416a.4-48-AEO2016.1-0~ref2016-d032416a.5-48-AEO2016.1-0~ref2016-d032416a.30-48-AEO2016.1-0~ref2016-d032416a.31-48-AEO2016.1-0&map=ref2016-d032416a.5-48-AEO2016.1-0&ctype=linechart&sourcekey=0

U.S. Energy Information Administration. (2017). *Annual energy outlook 2017 table: Light duty vehicle sales by technology type.* https://www.eia.gov/outlooks/aeo/data/browser/#/?id=48-AEO2017&region=1-0&cases=ref2017&start=2015&end=2050&f=A&linechart=ref2017-d120816a.4-48-AEO2017.1-0~ref2017-d120816a.5-48-AEO2017.1-0~ref2017-d120816a.30-48-AEO2017.1-0~ref2017-d120816a.31-48-AEO2017.1-0&map=ref2017-d120816a.5-48-AEO2017.1-0&ctype=linechart&sourcekey=0

U.S. Energy Information Administration. (2018). *Annual energy outlook 2018 table: Light duty vehicle sales by technology type.* https://www.eia.gov/outlooks/aeo/data/browser/#/?id=48-AEO2018&region=1-0&cases=ref2018&start=2016&end=2050&f=A&linechart=ref2018-d121317a.4-48-AEO2018.1-0~ref2018-d121317a.5-48-AEO2018.1-0~ref2018-d121317a.30-48-AEO2018.1-0~ref2018-d121317a.31-48-AEO2018.1-0&map=ref2018-d121317a.5-48-AEO2018.1-0&ctype=linechart&sourcekey=0

U.S. Energy Information Administration. (2019a). *Alternative fuel vehicle data.* https://www.eia.gov/renewable/afv/index.php

U.S. Energy Information Administration. (2019b). *Annual energy outlook 2019 table: Light duty vehicle sales by technology type.* https://www.eia.gov/outlooks/aeo/data/browser/#/?id=48-AEO2019&region=1-0&cases=ref2019&start=2017&end=2050&f=A&linechart=ref2019-d111618a.4-48-AEO2019.1-0~ref2019-d111618a.5-48-AEO2019.1-0~ref2019-d111618a.30-48-AEO2019.1-0~ref2019-d111618a.31-48-AEO2019.1-0&map=ref2019-d111618a.5-48-AEO2019.1-0&ctype=linechart&sourcekey=0

U.S. Energy Information Administration. (2020). *Annual energy outlook 2020 table: Light duty vehicle sales by technology type.* https://www.eia.gov/outlooks/aeo/data/browser/#/?id=48-AEO2020&region=1-0&cases=ref2020&start=2018&end=2050&f=A&linechart=ref2020-d112119a.4-48-AEO2020.1-0~ref2020-d112119a.5-48-AEO2020.1-0~ref2020-d112119a.30-48-AEO2020.1-0~ref2020-d112119a.31-48-AEO2020.1-0&map=ref2020-d112119a.5-48-AEO2020.1-0&ctype=linechart&sourcekey=0

World Health Organization. (n.d.-a). *Alphabetical list of who member states.* https://www.who.int/choice/demography/by_country/en/

World Health Organization. (n.d.-b). *Burden of disease.* https://apps.who.int/gho/data/node.main.BODAMBIENTAIR?lang=en