Car Industry during Covid

Jiayue He, Xuefei Wang

December 3, 2021

Contents

Project Description

For this Capstone project, we choose to focus on Economics. Our leader, Jiayue He majors in applied statistics and minored in economics. Therefore, in the beginning, we would like to study the affection of producing masks on our environmental economics. However, after doing some research, we found it hard for us to combine producing masks on environment and economy together. Then, we decided to move to some more common topics that also involved economics or environmental economics. Our new issue is based on car performance before and after COVID to do some further research. Due to this being an observational study, we couldn't draw a cause-effect conclusion in this report.

Our datasets are from two websites, one is provided by the U.S. Environment Protection Agency (EPA), and another is from the Automotive industry Portal, MarkLines.

Research Questions

Research Question 1: How much does COVID affect vehicle sales (automotive economy)? Research Question 2: How does the performance of the automobile industry affect our environment before and after COVID?

Variables

Variable	Types	Units	Definition
Manufacturer	Categorical	/	Different car brands
Year	Categorical	/	Year from 2018 to 2020
Month	Quantitative	/	Month in 2018-2020
X2.Cycle.MPG	Quantitative	miles per gallon	Compliance fuel economy measured by "2-cycle" tests
Real.World.MPG	Quantitative	miles per gallon	Estimated real-world fuel economy measured by "5-cycle" tests
Real.World.MPG_City	Quantitative	miles per gallon	Estimated real-world fuel economy measured by "5-cycle" tests for city
Real.World.MPG_Hwy	Quantitative	miles per gallon	Estimated real-world fuel economy measured by "5-cycle" tests for highway
Real.World.CO2	Quantitative	g/mi	Estimated real-world CO2 measured by "5-cycle" tests
Real.World.CO2_City	Quantitative	g/mi	Estimated real-world CO2 measured by "5-cycle" tests for city
Real.World.CO2_Hwy	Quantitative	g/mi	Estimated real-world CO2 measured by "5-cycle" tests for highway
Weight	Quantitative	lbs	Car weights
Footprint	Quantitative	square footage	Carbon footprint
Engine.Displacement	Quantitative	cubic inches	Total volume of air/fuel mixture an engine can draw in during one complete engine cycle
Horsepower	Quantitative	hp	The power a car engine produces
Fuel.Delivery.GDI	Quantitative	gdi rate	A fuel delivery system in gasoline internal combustion engines
Sale	Quantitative	million in units	Total sale per month
Covid	Quantitative	/	Whether the year is 2020 or not

To exactly demonstrate the COVID, we create one column named "Covid" to indicate whether the model year is during COVID. Sale is created by adding sale units for each branch per month.

Exploratory Data Analysis (EDA)

From the barplot below, we can find obviously how COVID affected vehicle sales. From the bars of 2018 and 2019, we can see that there is no much difference of sales in each month. If no COVID, case of 2020 perhaps has the similar sale trend. In fact, however, sales in 2020 illustrated å sharp change in some months, especially from March to June, which was the worst period caused by COVID. After June, the sales of 2020 started to recover and has been par on with data of 2018 and 2019 during the second half of the year.

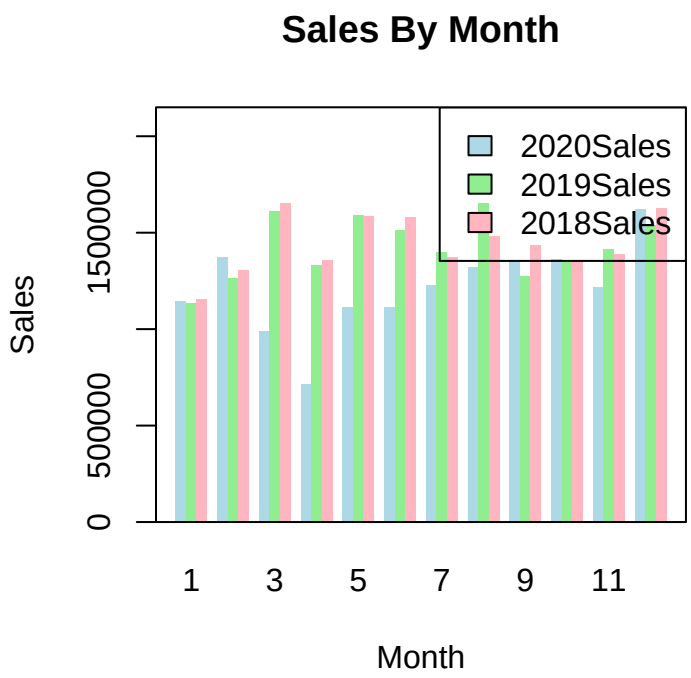


Figure 1: Car Sale Barplot

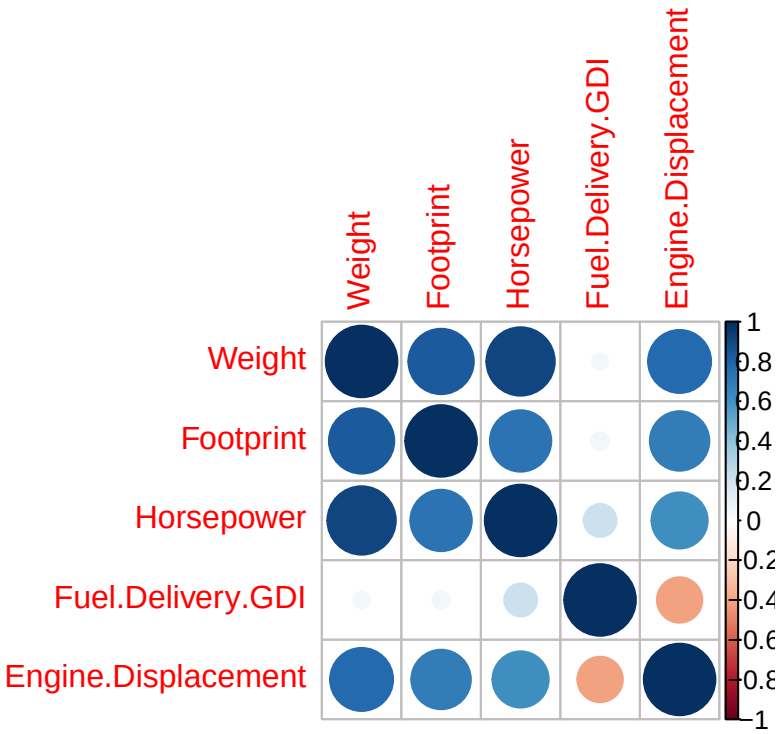


Figure 2: Correlation Plot

Collinearity happens when two or more explanatory variables are correlated with each other. An extreme situation called multicollinearity, where collinearity exists between three or more variables even if no pair of variables has a particularly high correlation. This means that there is redundancy between explanatory variables. We can see from this figure that four variables *Weight*, *Footprint*, *Horsepower*, *Engine.Displacement* show high relativity. Therefore, we decide to remove three of them and perserve *Weight* as one independent variable. Then we can retrain the linear regression model using *Country*, *Covid*, *Weight* and *Fuel.Delivery.GDI* as final explanatory variables.

Statistical Analysis

Research Question 1

Based on Figure 1, there is a big gap that happened in March 2020 and April 2020. With the effect of Covid, sales of May and June are also lower than the year before. The percentage of decreasing in March 2020 is 38.43% compared with March 2019. The decrease percentage for April 2020 is 46.62%. The percentage of May and June are around 30%. After that, the gap becomes small, then back to normal.

Research Question 2

To check how the performance of the automobile industry affect our environment before and after COVID, we filter out the related data including features like CO2, year, Footprint and others. We can omit the rows with missing values in some features, and thus obtain a clean dataset.

After data preprocessing, we fit a linear regression model using mentioned variables and set the *Real.World.CO2* emission as the response variable. At first, we select *Country*, *Covid*, *Weight*, *Footprint*, *Horsepower*, *Fuel.Delivery.GDI*, *Engine.Displacement* these seven variables as explanatory variables. Then we check the regression assumptions like linearity of the data, normality of residuals, homogeneity of residuals variance and independence of residuals error terms. They are appeared in Appendix.

Variables	Coefficients	P-values	Is it significant
(Intercept)	209.764038	1.15e-08	Yes
CountryAmerica	25.171417	0.000243	Yes
CountryGerman	24.130255	0.000163	Yes
CountryJapan	-20.820965	0.001405	Yes
CountryKorea	-3.120265	0.702391	No
Covid	-11.051030	0.050606	Yes
Weight	0.045397	3.08e-07	Yes
Fuel.Delivery.GDI	-61.465645	$<$ 2e-16	Yes

From the linear regression result, we can find that the p-value of variable "Covid", around 0.05, which is a relatively significant value and means that "Covid" has much difference on the environment. Hence we can draw our conclusion: the performance of the automobile industry will affect our environment before and after COVID. The linear model is:

Real.World.CO2 = 209.764038 + CountryAmerica * 25.171417 + CountryGerman * 24.130255 + CountryJapan * -20.820965 + CountryKorea * -3.120265 + Covid * -11.051030 + Weight * 0.045397 + Fuel.Delivery.GDI * -61.465645

Recommendations

Our main research question is: How does the performance of the automobile industry affect our environment before and after COVID? To study this question, we plan to build a linear regression model to check how significant the variable of COVID is on environment. We first filter out the related data including features like CO2, year, Footprint and others. We can omit the rows with missing values in some features, and thus obtain a clean dataset. To exactly demonstrate the COVID, we create one column named "Covid" to indicate whether the model year is during COVID.

Resources

Source 1: https://www.epa.gov/automotive-trends/explore-automotive-trends-data#DetailedData

The first source we used is provided by the U.S. Environment Protection Agency's (EPA). EPA has collected data on every new light-duty vehicle model sold in the United States since 1975, either from testing performed by EPA at the National Vehicle Fuel and Emissions Laboratory in Ann Arbor, Michigan, or directly from manufacturers using official EPA test procedures. These data are collected to support several important national programs, including EPA criteria pollutant and GHG standards, the U.S. Department of Transportation's National Highway Traffic Safety Administration (NHTSA) Corporate Average Fuel Economy (CAFE) standards, and vehicle Fuel Economy and Environment labels. Thus, this expansive data set allows EPA to provide a uniquely comprehensive analysis of the automotive industry over the last 45 years.

Source 2: https://www.marklines.com/en/statistics/flash_sales/automotive-sales-in-usa-by-month-2020

The second data source is from Automotive industry Portal, MarkLines. This data source contains every month sale in 2020. In addition, we also used grepl the information for 2018 and 2019 from this website. MarkLine is intend to develop and grow the automotive industry by providing information services. This specific dataset was collected every month in 2020 and finished collecting in January 6th, 2021. All of this information is collected through purchases from third-party sources, as well as partnerships with other companies. We found this dataset on their company's official website.

Professor name: Sherry Wu

Contact information:

510 Kern Building

University Park, PA 16802

Email: sqw5740@psu.edu

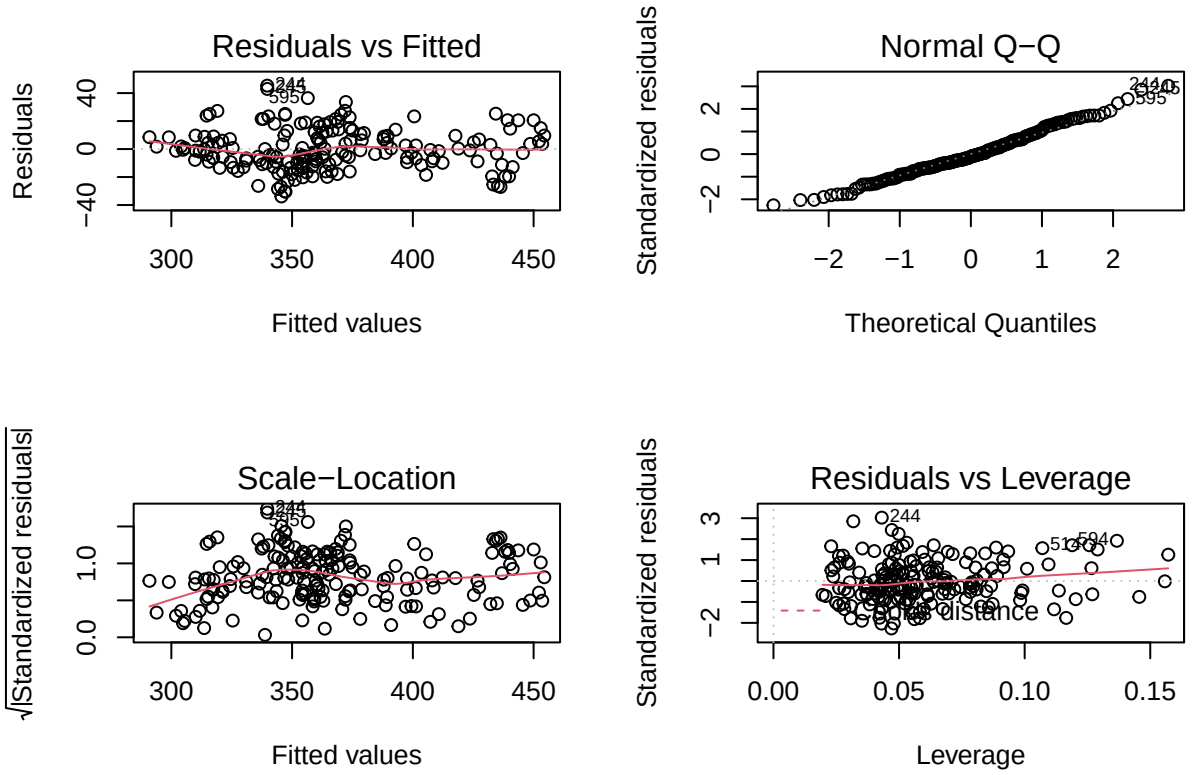
Phone: (814) 865-4921

Introduction:

Sherry Wu is a visiting assistant professor at Penn State University in the department of Economics, she is particularly interested in industrial organization. Dr.Wu is also the professor in the environmental economics class. Since our research goal is to analyze the impact of automotive industry on economy and environment, we think Dr.Wu will be very helpful in giving advice for our project.

Additional Considerations

Technical Appendix



We drew Residual vs. Fitted plots and Normal Q_Q plots to check assumptions. First of all, the residual vs. fitted plots show that the Linearity is met for all six models since those dots are randomly spaced around the line of residual(red line) that is 0. Secondly, the Independent assumption is met since every child is an individual observation unit. Then, from the Normal Q_Q plots, we observed that most of the dots are along the dashed line. We can state that the normality is met. Last but not least, according to residual vs. fitted plots, equal variance is also met. Distances between each dot seem to be the same.

R Script

```
# clean up & set default chunk options
rm(list = ls())
knitr::opts_chunk$set(echo = FALSE)

# load packages
library(readxl)
library(ggplot2)
library(tidyverse)
library(dplyr)
library(mosaic)
```

```r
library(lubridate)
library(data.table)
library(plotly)
library(dygraphs)
library(corrplot)
library(kableExtra)

# inputs
environment <- read.csv("environment.csv")
sale <- read.csv("TOTALSA.csv")
brandsale <- read_excel("TotalSalebyBrand.xlsx")
variables <- read_excel("variables.xlsx")
kable(variables,booktabs = T) %>%
  kable_styling(latex_options = "HOLD_position", font_size = 7)

# The first resource cleaning
## selections
year <- c('2018','2019','2020')
environment1 <- filter(environment, Model.Year %in% year)

model <- c('All')

#environment1 <- filter(environment1, Manufacturer != model)
environment1 <- filter(environment1, ï..Manufacturer != model)

months <- c('2018-01-01','2018-02-01','2018-03-01','2018-04-01','2018-05-01',
            '2018-06-01','2018-07-01','2018-08-01','2018-09-01','2018-10-01',
            '2018-11-01','2018-12-01','2019-01-01','2019-02-01','2019-03-01',
            '2019-04-01','2019-05-01','2019-06-01','2019-07-01','2019-08-01',
            '2019-09-01','2019-10-01','2019-11-01','2019-12-01','2020-01-01',
            '2020-02-01','2020-03-01','2020-04-01','2020-05-01','2020-06-01',
            '2020-07-01','2020-08-01','2020-09-01','2020-10-01','2020-11-01',
            '2020-12-01')
sale1 <- filter(sale, DATE %in% months)
sale1$date <- ymd(months)

## rename variables to become more appropriate
names(environment1)[names(environment1) ==
                      'Model.Year'] <- 'Year'
names(environment1)[names(environment1) ==
                      'ï..Manufacturer'] <- 'Manufacturer'
names(environment1)[names(environment1) ==
                      'Real.World.CO2..g.mi.'] <- 'Real.World.CO2'
names(environment1)[names(environment1) ==
                      'Real.World.CO2_City..g.mi.'] <- 'Real.World.CO2_City'
names(environment1)[names(environment1) ==
                      'Real.World.CO2_Hwy..g.mi.'] <- 'Real.World.CO2_Hwy'
names(environment1)[names(environment1) ==
                      'Weight..lbs.'] <- 'Weight'
names(environment1)[names(environment1) ==
                      'Horsepower..HP.'] <- 'Horsepower'
names(environment1)[names(environment1) ==
                      'Footprint..sq..ft..'] <- 'Footprint'
```

```r
names(environment1)[names(environment1) ==
                      'Fuel.Delivery...Gasoline.Direct.Injection..GDI.'] <-
  'Fuel.Delivery.GDI'

## select variables that can be used
environment1 <-
  environment1 %>%
  select(Manufacturer, Year, X2.Cycle.MPG, Real.World.MPG, Real.World.MPG_City,
         Real.World.MPG_Hwy, Real.World.CO2, Real.World.CO2_City, Real.World.CO2_Hwy,
         Weight, Footprint, Horsepower, Fuel.Delivery.GDI, Engine.Displacement)

## change form of variables
environment1$Year <- as.factor(environment1$Year)
environment1$X2.Cycle.MPG <- as.numeric(environment1$X2.Cycle.MPG)
environment1$Real.World.MPG <- as.numeric(environment1$Real.World.MPG)
environment1$Real.World.MPG_City <- as.numeric(environment1$Real.World.MPG_City)
environment1$Real.World.MPG_Hwy <- as.numeric(environment1$Real.World.MPG_Hwy)
environment1$Real.World.CO2 <- as.numeric(environment1$Real.World.CO2)
environment1$Real.World.CO2_City <- as.numeric(environment1$Real.World.CO2_City)
environment1$Real.World.CO2_Hwy <- as.numeric(environment1$Real.World.CO2_Hwy)
environment1$Weight <- as.numeric(environment1$Weight)
environment1$Footprint <- as.numeric(environment1$Footprint)
environment1$Horsepower <- as.numeric(environment1$Horsepower)
environment1$Fuel.Delivery.GDI <- as.numeric(environment1$Fuel.Delivery.GDI)

sale1 <-
  sale1 %>%
  mutate(dates = seq(as.Date("2018-01-01",  format = "%Y-%m-%d"), length.out = 36, by = "month"))

environment1$Engine.Displacement[environment1$Manufacturer == "Tesla"] <- 0
environment1$Engine.Displacement <- as.numeric(environment1$Engine.Displacement)

## Tesla carbon emission

environment1$Real.World.CO2[environment1$Manufacturer == "Tesla" &
                              environment1$Year == "2018"] <- 400
environment1$Real.World.CO2[environment1$Manufacturer == "Tesla" &
                              environment1$Year == "2019"] <- 420
environment1$Real.World.CO2[environment1$Manufacturer == "Tesla" &
                              environment1$Year == "2020"] <- 400

# The second resource
## change name
names(brandsale)[2] <- "Sale2020"
names(brandsale)[3] <- "Sale2019"
names(brandsale)[4] <- "Sale2018"

## calculate the total for each year
brandsale1 <-
  brandsale %>%
  select(Sale2020, Month) %>%
  group_by(Month)%>%
  summarise(total = sum(Sale2020))
```

```r
brandsale2 <-
  brandsale %>%
  select(Sale2019, Month) %>%
  group_by(Month)%>%
  summarise(total = sum(Sale2019))

brandsale3 <-
  brandsale %>%
  select(Sale2018, Month) %>%
  group_by(Month)%>%
  summarise(total = sum(Sale2018))

## rename the date
brandsale1$Month[brandsale1$Month == '1'] <- '2020-01'
brandsale1$Month[brandsale1$Month == '2'] <- '2020-02'
brandsale1$Month[brandsale1$Month == '3'] <- '2020-03'
brandsale1$Month[brandsale1$Month == '4'] <- '2020-04'
brandsale1$Month[brandsale1$Month == '5'] <- '2020-05'
brandsale1$Month[brandsale1$Month == '6'] <- '2020-06'
brandsale1$Month[brandsale1$Month == '7'] <- '2020-07'
brandsale1$Month[brandsale1$Month == '8'] <- '2020-08'
brandsale1$Month[brandsale1$Month == '9'] <- '2020-09'
brandsale1$Month[brandsale1$Month == '10'] <- '2020-10'
brandsale1$Month[brandsale1$Month == '11'] <- '2020-11'
brandsale1$Month[brandsale1$Month == '12'] <- '2020-12'

brandsale2$Month[brandsale2$Month == '1'] <- '2019-01'
brandsale2$Month[brandsale2$Month == '2'] <- '2019-02'
brandsale2$Month[brandsale2$Month == '3'] <- '2019-03'
brandsale2$Month[brandsale2$Month == '4'] <- '2019-04'
brandsale2$Month[brandsale2$Month == '5'] <- '2019-05'
brandsale2$Month[brandsale2$Month == '6'] <- '2019-06'
brandsale2$Month[brandsale2$Month == '7'] <- '2019-07'
brandsale2$Month[brandsale2$Month == '8'] <- '2019-08'
brandsale2$Month[brandsale2$Month == '9'] <- '2019-09'
brandsale2$Month[brandsale2$Month == '10'] <- '2019-10'
brandsale2$Month[brandsale2$Month == '11'] <- '2019-11'
brandsale2$Month[brandsale2$Month == '12'] <- '2019-12'

brandsale3$Month[brandsale3$Month == '1'] <- '2018-01'
brandsale3$Month[brandsale3$Month == '2'] <- '2018-02'
brandsale3$Month[brandsale3$Month == '3'] <- '2018-03'
brandsale3$Month[brandsale3$Month == '4'] <- '2018-04'
brandsale3$Month[brandsale3$Month == '5'] <- '2018-05'
brandsale3$Month[brandsale3$Month == '6'] <- '2018-06'
brandsale3$Month[brandsale3$Month == '7'] <- '2018-07'
brandsale3$Month[brandsale3$Month == '8'] <- '2018-08'
brandsale3$Month[brandsale3$Month == '9'] <- '2018-09'
brandsale3$Month[brandsale3$Month == '10'] <- '2018-10'
brandsale3$Month[brandsale3$Month == '11'] <- '2018-11'
brandsale3$Month[brandsale3$Month == '12'] <- '2018-12'

## combine three data
```

```r
brandfinal <- bind_rows(brandsale1,brandsale2,brandsale3)

## ascending the date
brandfinal <-
  brandfinal %>%
  arrange(Month)

# clean data for the first model
names(environment)[names(environment) == 'ï..Manufacturer'] <- 'Manufacturer'
research.data <- environment
research.data$Country <- "all"
research.data$Country[research.data$Manufacturer %in% c("Toyota","Mazda","Honda","Subaru","Nissan")] <-
research.data$Country[research.data$Manufacturer %in% c("BMW","Mercedes","VW")] <- "German"
research.data$Country[research.data$Manufacturer %in% c("GM","Ford","Tesla","FCA")] <- "America"
research.data$Country[research.data$Manufacturer %in% c("Hyundai","Kia")] <- "Korea"

colnames(research.data)[4] <- "Year"
colnames(research.data)[9] <- "Real.World.CO2"
colnames(research.data)[12] <- "Weight"
colnames(research.data)[13] <- "Footprint"
colnames(research.data)[15] <- "Horsepower"
colnames(research.data)[29] <- "Fuel.Delivery.GDI"

research.data <- research.data[c("Real.World.CO2","Country" ,"Year" ,"Weight" ,"Footprint" ,"Horsepower

research.data <- research.data %>% drop_na()

clean_data <- research.data[!(research.data$Real.World.CO2=="-" | research.data$Country =="-" | researcl

clean_data <- transform(clean_data, Real.World.CO2 = as.numeric(Real.World.CO2),Year = as.numeric(Year)

clean_data$Covid <- as.integer(clean_data$Year == 2020)
gbs <- aggregate(brandsale[,2:4], list(brandsale$Month), FUN=sum)
mx <- t(as.matrix(gbs[-1]))
colors <- c("lightblue", "lightgreen", "lightpink")

barplot(mx,main='Sales By Month', names = c(1:12), ylab='Sales', xlab='Month',border=F, beside = TRUE, c
# add a legend
box()
legend('topright',fill=colors,legend=c('2020Sales','2019Sales', '2018Sales'))
# check Collinearity
model_corr_matrix <- cor(clean_data %>%
                           select(Weight, Footprint, Horsepower, Fuel.Delivery.GDI, Engine.Displacemen
                         use = "pairwise.complete.obs")

corrplot::corrplot(model_corr_matrix)
# linear regression with Covid
fit.lm1 <- lm(Real.World.CO2 ~ Country + Covid + Weight + Footprint + Horsepower + Fuel.Delivery.GDI + I

par(mfrow = c(2, 2))
plot(fit.lm1)
```