Car Industry during Covid

Jiayue He, Xuefei Wang

November 30, 2021

Contents

Project Description

For this Capstone project, we choose to focus on Economics. Our leader, Jiayue He is in majored of applied statistics and minor in economics. Therefore, in the beginning, we would like to study the affection of producing masks on our environmental economics. However, after doing some researches, we found it was hard for us to combine the affection of producing masks on environment and economy together. Then, we decided to move to some more common topics that also involved economics or environmental economics.

Research Questions

What are the overarching research questions that the client is targeting?

Question 1: What are some factors associated with environment in car industry?

Question 2:

Statistical Questions *(optional)*

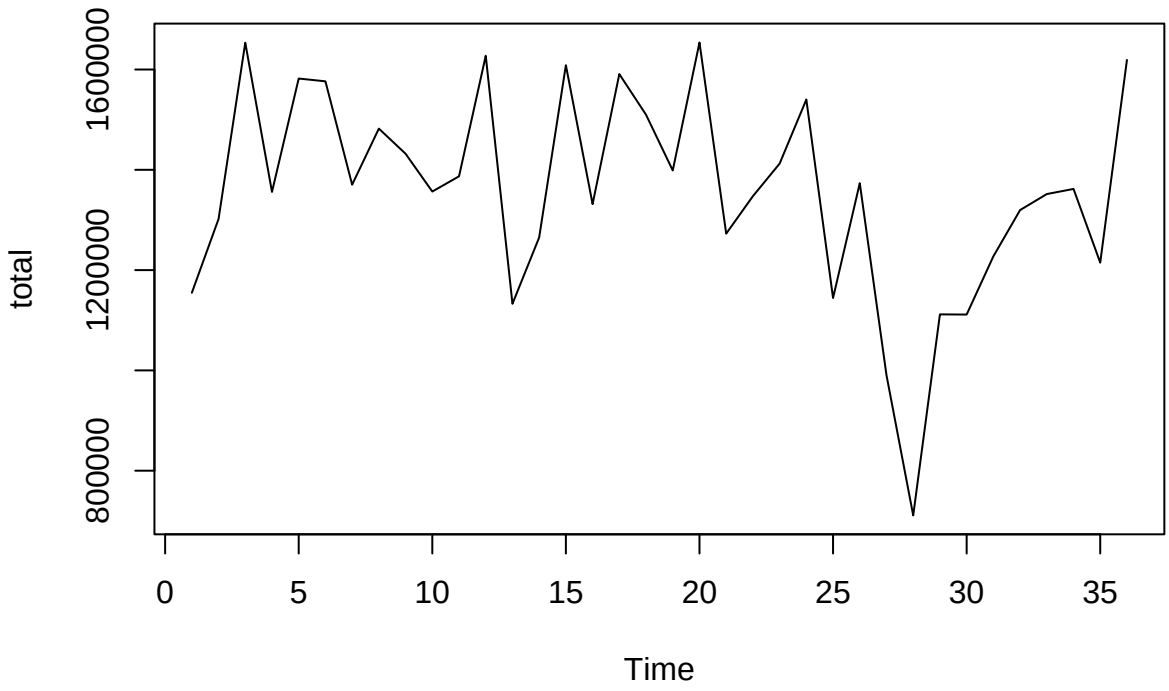
Question 1:

Question 2:

Variables

Variable	Types	Units	Definition
Manufacturer	Categorical	/	Different car brands
Year	Categorical	/	Year from 2018 to 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 Price	Quantitative	million in units	Total sale per year

Exploratory Data Analysis (EDA)



Statistical Analysis

Recommendations

Question 1:

Question 2:

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, there is a column that shows the sales in the same month but in 2019.

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.

In this data, a single observation unit represents one of light commercial vehicles, such as Toyota, Ford, and Honda etc. There are 12 tables that represent each month sales. Every brand of vehicles and its sales form a row. It contains 23 brands of vehicles, so it has 23 rows. Of course, we need to find a way to combine these 12 tables in our project. Therefore, the final information about our data is still in process.

Additional Considerations

Technical Appendix

Detailed information and a copy of code and or software results. Additional graphs and supporting figures may also be placed in the appendix.

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)
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")
variables <- data.frame(variables)
kable(variables,booktabs = T) %>%
  kable_styling("striped", latex_options = "HOLD_position", font_size = 7)

# The first resource cleaning
```

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

## 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))

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
brandfinal <- bind_rows(brandsale1,brandsale2,brandsale3)

## ascending the date
brandfinal <-
  brandfinal %>%
  arrange(Month)
brandfinal1 <-
  brandfinal %>%
  select(total)
plot.ts(brandfinal1)
```