IE 48A - Final

Hacı Mehmet İnce

9/14/2020

Contents

PART I:

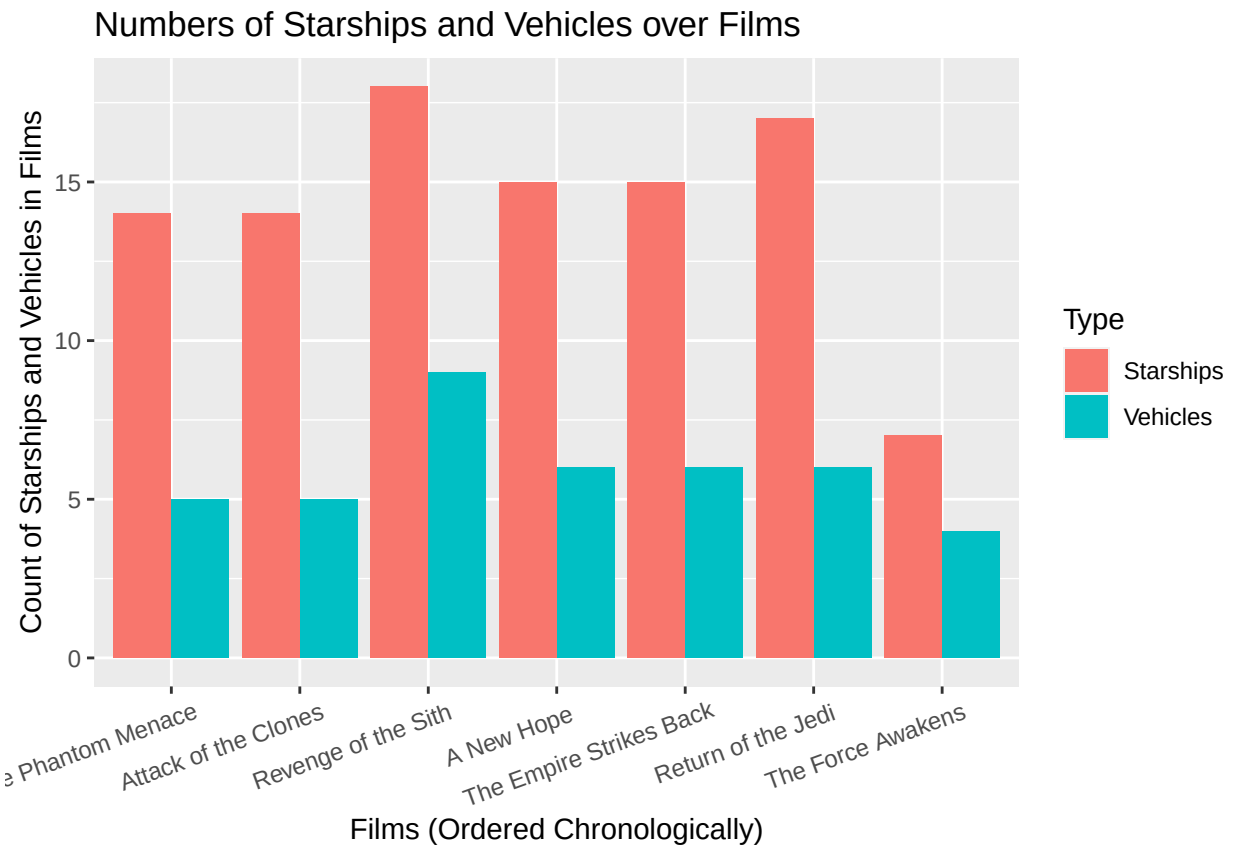
1. It may make sense to look at a country and see how the increase is. However, it is not correct to look at two different countries and say that "if it has increased this much in one, it will increase in the same amount or rate in the other". First, the way data is collected may differ from country to country. The methods of identifying patients or how many of them are determined are very important. There are also many factors that affect the spread of the disease. Measures in some countries may be more effective, or there may be less interaction between people. It is therefore wrong to attribute great significance to these comparisons.

2. The first step is to understand the data. For this, general things such as the distribution of the data, its outliers, whether it has missing / incorrect information, and the average of its columns are checked. As a second step, correlations and relationships between columns are checked. Maybe unnecessary columns are found or interesting relationships are detected. Afterwards, various arguments are formed and analyzes are made to test them. In the given example, the impact is quite difficult to measure. If it were me, I would first group people according to given attributes. Later, I would identify the groups with the highest proportion of people in need. After checking whether these ratios were normal, I would think that groups with high rates were struggling for various reasons, and I would direct the funds to their benefit. As I mentioned earlier, if certain groups need help more than other groups, there are reasons. However, I think it is ethical to say what the data says if my data is not sufficient to detect these reasons. In other words, if there is no evidence, title should be "Pain Points in Our Society and Optimal Budget Allocation".

3. There are 2 trilogies in the films of the Starwars series that have been shot so far. In my opinion, the war and action scenes increase in the finals of the trilogies. In the Starwars series, this will especially result in an increase in the number of starships. So my argument is that the final movies of the trilogy will have more starships than previous movies. It can also be the case for vehicles. To examine this, I'll sort the movies and have the number of starships and vehicles in each movie plotted as a bar plot. Before, the libraries that will be required in this and the following sections are loaded.

```
library(dplyr)
library(tidyr)
library(zoo)
require(quantmod)
library(reshape)
library(ggplot2)
library(ggrepel)
```

```
data1 = starwars[,c("films","vehicles")] %>% unnest(films)  %>% unnest(vehicles)
data1$vehicles = "Vehicles"

data2 = starwars[,c("films","starships")] %>% unnest(films) %>% unnest(starships)
data2$starships = "Starships"

names(data1)[2] = "Type"
names(data2)[2] = "Type"

data3 = rbind(data1, data2)

data3$Type = as.factor(data3$Type)
filmnames = c("The Phantom Menace","Attack of the Clones","Revenge of the Sith",
              "A New Hope", "The Empire Strikes Back", "Return of the Jedi",
              "The Force Awakens")
data3$films <- factor(data3$films,levels = filmnames)

ggplot(data = data3, aes(x = films, fill = Type)) +
  geom_bar(position = "dodge") +
  ggtitle("Numbers of Starships and Vehicles over Films") +
  theme(
    axis.text.x = element_text(angle = 20, vjust = 0.9, hjust = 0.9)
  ) +
  xlab("Films (Ordered Chronologically)") + ylab("Count of Starships and Vehicles in Films")
```



The increase in the number of starships in the finals is seen, though not as much as I expected. While the

same is true for vehicles in one trilogy, there is no change in the other.

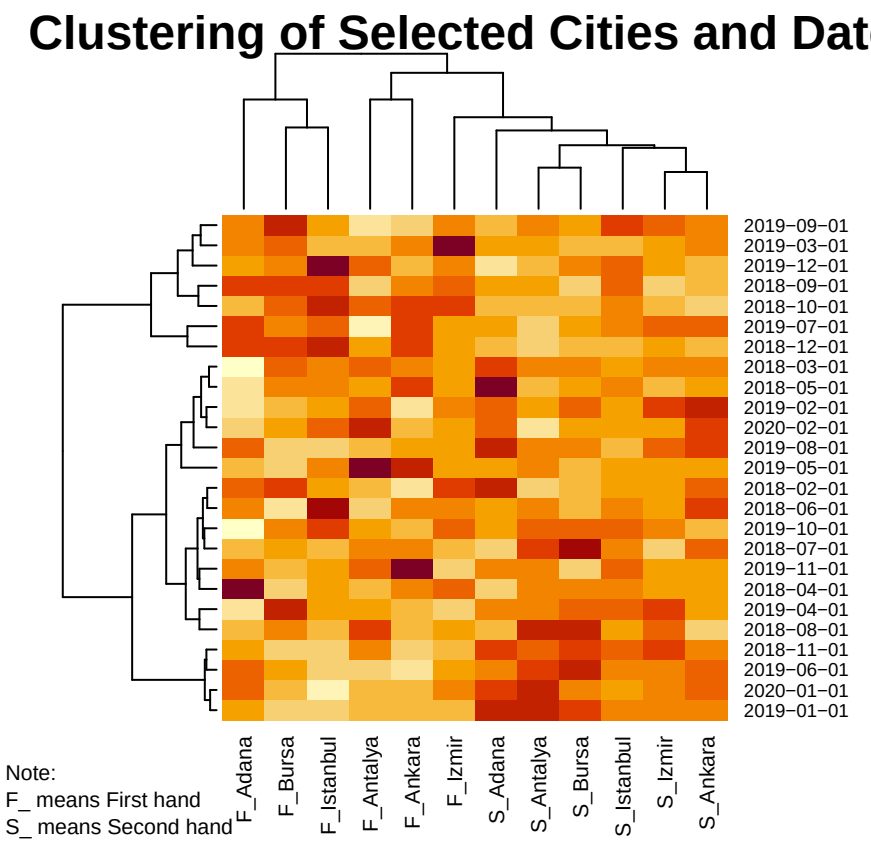
PART II:

In our group project, the cities were tried to be clustered according to the movement of the indexes, but it was not clear what the results were based on and how similar the cities were to each other. Because I thought clarity was very important, I decided to develop this part of our project. To avoid confusion, preprocessed data in the project report was saved to my computer and called directly for the final. NA values were filled in as 0 in the project, but having an index of 0 means that houses are free, which is quite illogical. Therefore, NA values were filled in by averaging the two months around. Subsequently, the monthly percentage changes were recorded in new "evds_diff" data. In "cols", the cities to be used are selected. "mindate" and "maxdate" were used to determine the date range. Selected columns are transferred to two new data, together with "Date" column. Two new data will be used, as two heat maps will be printed, one clustered and the other non-clustered.

```
url <- "https://raw.githubusercontent.com/pjournal/boun01-hmehmetince/gh-pages/data/preprocessed_evds.cs
evds <- read.csv(url(url, method="libcurl"))

evds = evds[,-1]
evds$Date = as.Date(evds$Date)


evds[1,][is.na(evds[1,])] = 1

for(i in c(2:165)){
  evds[,i] = na.approx(evds[,i])
}

evds_diff = evds

for(i in 2:165){
  evds_diff[,i] = unname(Delt(evds[,i]))
}

# NA values are removed
evds_diff = evds_diff[-1,]

cols = c("F_Adana","S_Adana","F_Istanbul","S_Istanbul","F_Ankara","S_Ankara","F_Antalya","S_Antalya","F_
mindate = "2018-02-01"
maxdate = "2020-02-01"

plot_data <- evds_diff[,c("Date",cols)]
plot_data <- plot_data[((plot_data$Date>=mindate) & (plot_data$Date<=maxdate)),]
plot_data2 <- plot_data # For ggplot heatmap
```

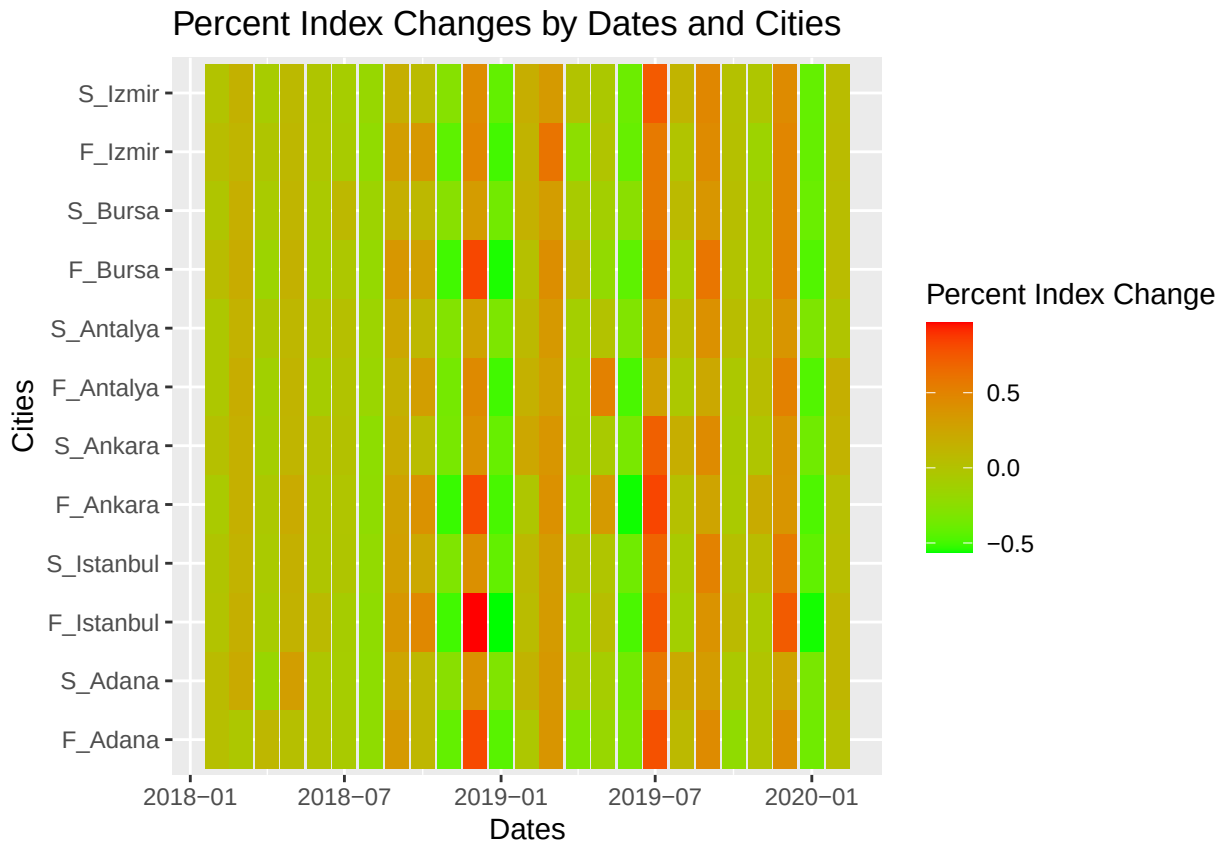
```
rownames(plot_data) = plot_data[,1]
plot_data = plot_data[,-1]

data <- as.matrix(plot_data)
heatmap(data, main = "Clustering of Selected Cities and Dates" , cex.main = 0.8, cexRow=0.7, cexCol=0.8
text<- "Note: \nF_ means First hand \nS_ means Second hand"
```

```
par(mar = c(8, 4, 3, 3),cex.main=0.5)
mtext(text, side = 1, line = 6, cex = 0.7, adj = 0)
```



The first heat map shows both which dates and which cities are close to each other and allows us to separate from where we want. However, it may be desirable to examine changes over time. Therefore, a new heat map, unclustered, is created.

```
long_plot_data2 <- melt(plot_data2, id.vars = "Date")

ggplot(long_plot_data2, aes(x = Date, y = variable, fill= value)) +
  geom_tile() +
  scale_fill_gradient(low="green", high="red") +
  labs(title = "Percent Index Changes by Dates and Cities", fill="Percent Index Change", x = "Dates", y
```



PART III:

The data covers the dates from January 2018 to August 2020. It is read from the created rds file. Used from 2019 to make the graphics clearer. I wonder, and as a result of this, the subject I examine is how sales change according to brands and time.

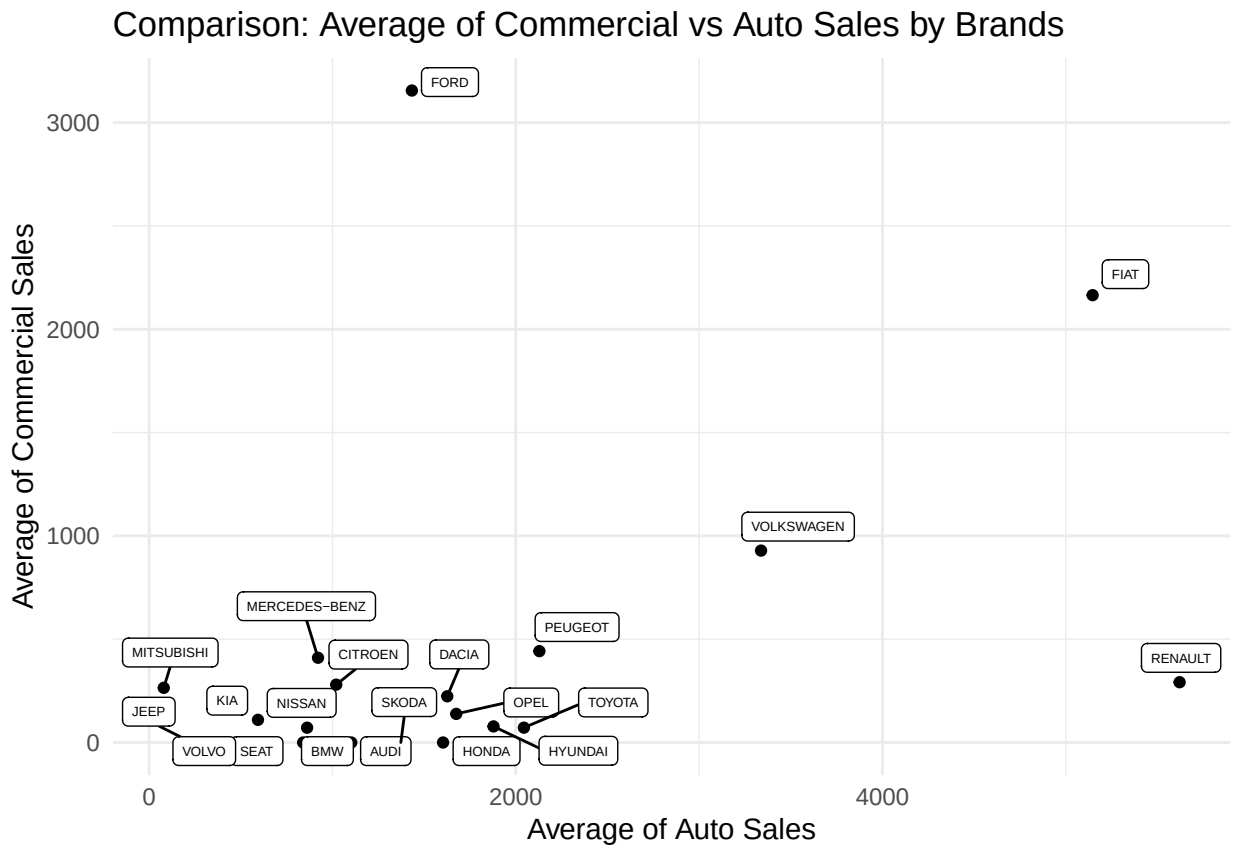
```
url <- "https://raw.githubusercontent.com/pjournal/boun01-hmehmetince/gh-pages/data/carsales"
cardata <- readRDS(url(url, method="libcurl"))

unique(cardata$brand)
```

```
##  [1] "ALFA ROMEO"    "ASTON MARTIN"  "AUDI"          "BENTLEY"
##  [5] "BMW"           "CHERY"         "CITROEN"       "DACIA"
##  [9] "DS"            "FERRARI"       "FIAT"          "FORD"
## [13] "HONDA"         "HYUNDAI"       "INFINITI"      "ISUZU"
## [17] "IVECO"         "JAGUAR"        "JEEP"          "KARSAN"
## [21] "KIA"           "LAMBORGHINI"   "LAND ROVER"    "LEXUS"
## [25] "MASERATI"      "MAZDA"         "MERCEDES-BENZ" "MINI"
## [29] "MITSUBISHI"    "NISSAN"        "OPEL"          "PEUGEOT"
## [33] "PORSCHE"       "RENAULT"       "SEAT"          "SKODA"
## [37] "SMART"         "SSANGYONG"     "SUBARU"        "SUZUKI"
## [41] "TOYOTA"        "VOLKSWAGEN"    "VOLVO"
```

```
cardata = cardata[cardata$year>2018,]
```

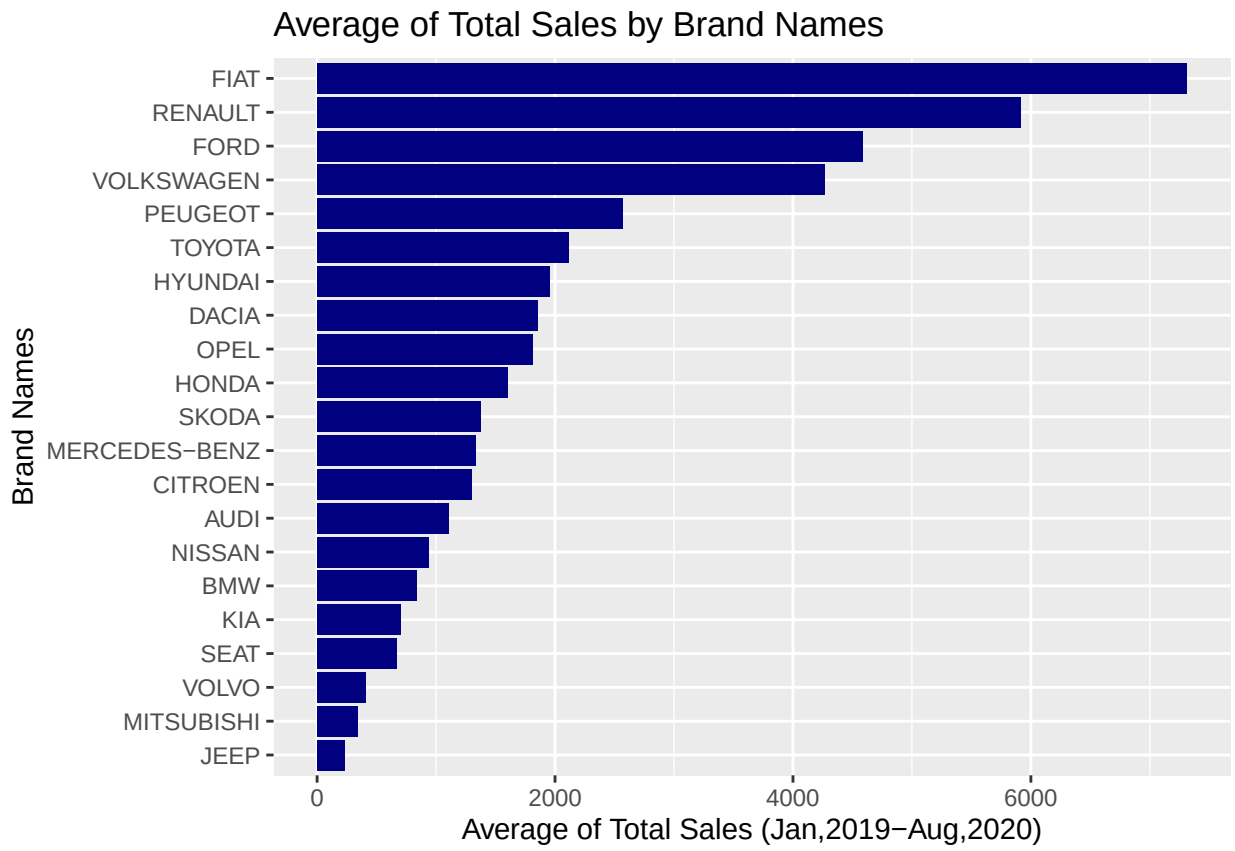
First, I grouped the data by brand and averaged the sales. So I can get general information about brands. I excluded the under-selling (less than 200 sales) brands from the data as they would not give meaningful results in plots. I analyzed the remaining commercial and auto sales on a plot. In general, some of brands that sell less only sell auto. Most of the sales of most brands come from auto. There appear to be 4 brands selling much more than others. Of these, Ford mostly sells commercials, while Renault almost always sells auto. Sales of Fiat and Volkswagen are more balanced.

```
grouped_means = cardata %>%
  group_by(brand) %>%
  summarise_at(vars(-c(year, month)),
               list(avg = mean)) %>%
  filter(total_total_avg > 200) %>%
  select(brand, auto_total_avg, comm_total_avg, total_total_avg)


ggplot(grouped_means) +
  geom_point(aes(auto_total_avg, comm_total_avg)) +
  theme_minimal() +
  geom_label_repel(aes(auto_total_avg, comm_total_avg,label=brand), size=1.6) +
  labs(x="Average of Auto Sales", y="Average of Commercial Sales", fill="Brand Names", title="Comparison
```



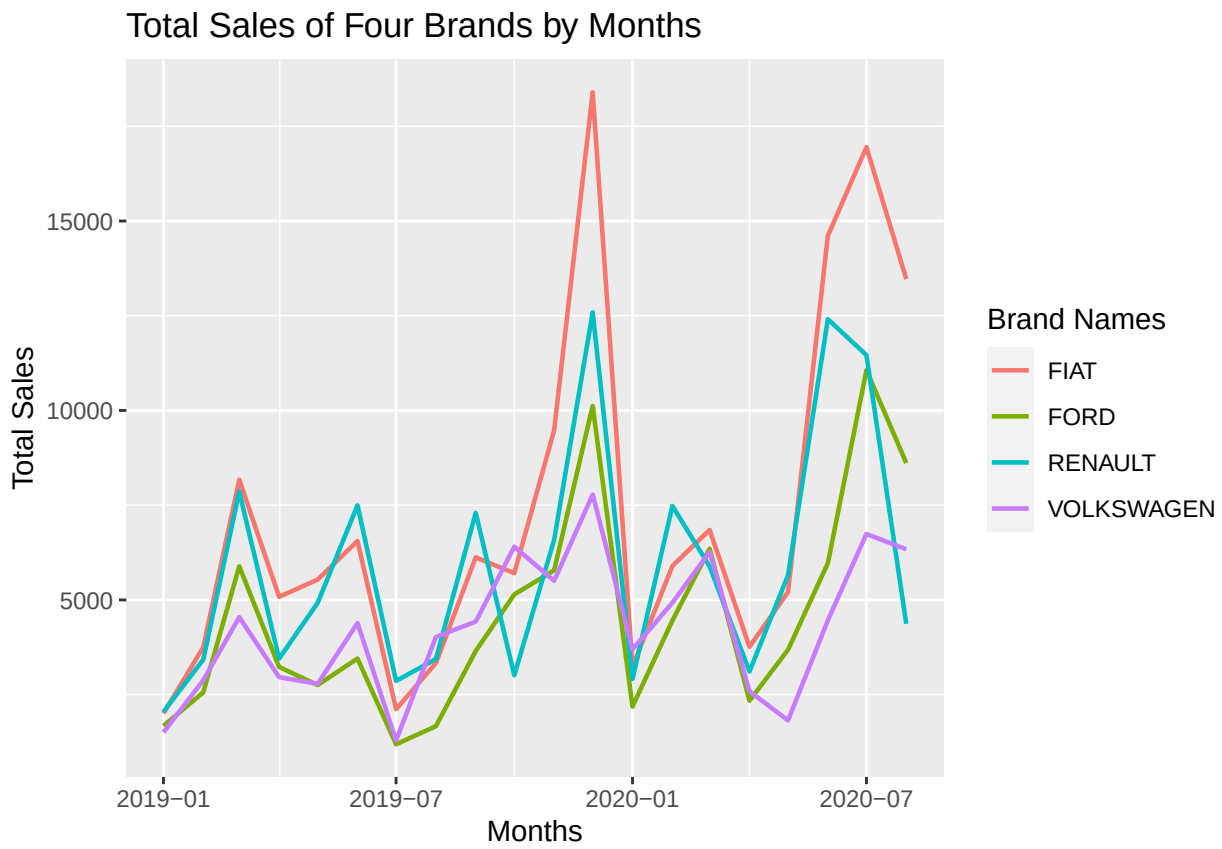
The vast majority of sales come from certain brands. I would like to examine these brands in detail. I created a bar plot to decide which brands to include and to compare the total sales of the brands, ranking them by their total sales.

```
ggplot(grouped_means, aes(reorder(brand,total_total_avg), total_total_avg)) +
  geom_bar(stat='identity', fill="#000080") +
  coord_flip() +
  labs(x="Brand Names", y="Average of Total Sales (Jan,2019-Aug,2020)", fill="Brand Names", title="Avera
```



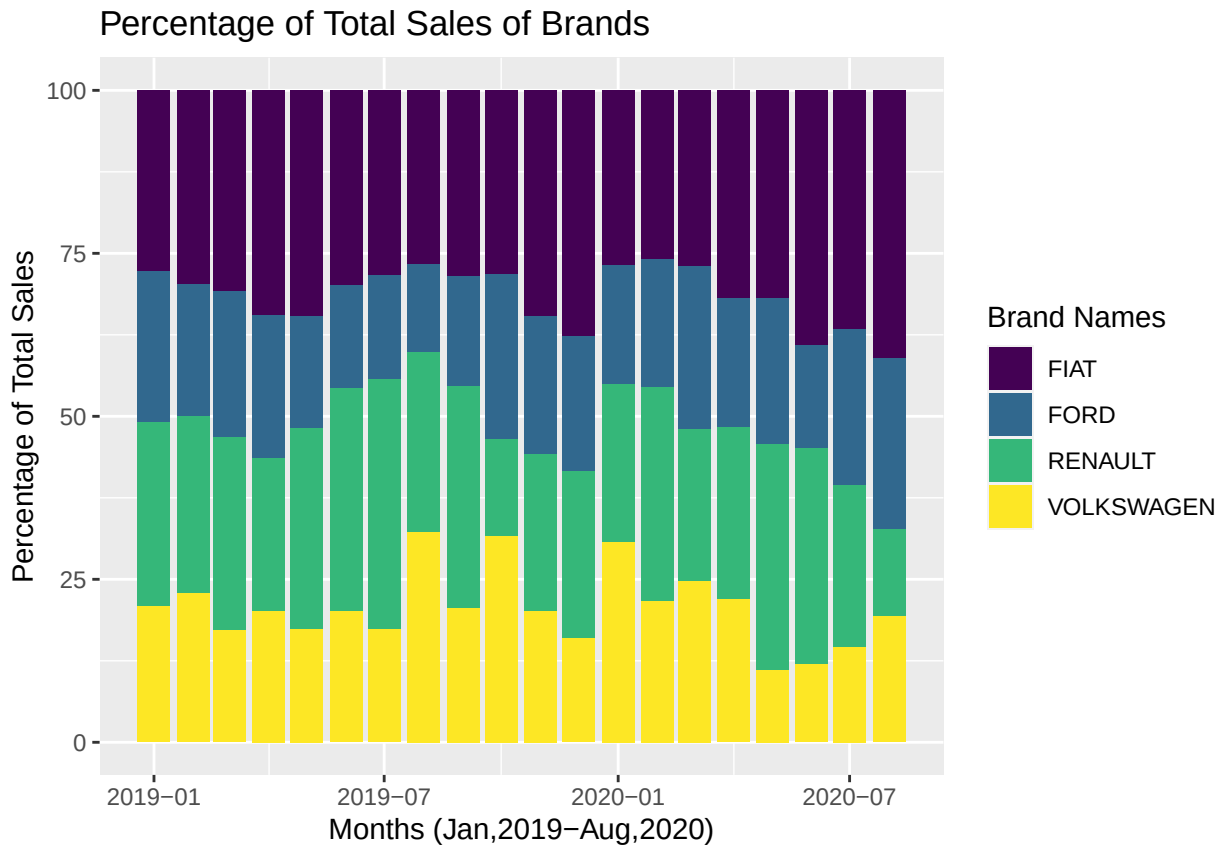
After the four brands I mentioned earlier, Fiat, Renault, Ford, and Volkswagen, there is a significant decrease in total sales. So I will continue my analysis using only these brands. I filtered the data to include these brands. Then I'd like to see how they changed over time using a line chart.

```
names_best_four = grouped_means[order(grouped_means$total_total_avg,decreasing = TRUE),][1:4,]$brand


cardata4brand = cardata[(cardata$brand %in% names_best_four),c("brand","auto_total","comm_total","total_

cardata4brand$date = as.Date(paste(cardata4brand$year,cardata4brand$month,"1",sep = "-"))

ggplot(cardata4brand, aes(x=date,fill=brand,color=brand)) +
  geom_line(aes(y = total_total),size=0.8) +
  labs(x = "Months", y = "Total Sales",color="Brand Names",
       title="Total Sales of Four Brands by Months")
```



Their sales often go up and down together. In some months, some may rise and fall more than others. Unless there is a change, we can say that their market shares should remain mostly the same. So when does it change and what is its cause? I created a new plot to see how these shares change. To create it, I calculated what percentage of the monthly sales of the brands accounted for the total sales of the four brands. So I can see when and how the market shares have changed.

```
perc_car4brand  = cardata4brand %>% group_by(date) %>%
  mutate(percentage = total_total/sum(total_total)*100)

ggplot(data=perc_car4brand,aes(x=date,y=percentage,
                               fill=ordered(brand))) +
  geom_bar(stat="identity") +
  labs(fill="Brand Names", title = "Percentage of Total Sales of Brands",
       x="Months (Jan,2019-Aug,2020)",y="Percentage of Total Sales")
```



The most interesting of the changes is that Renault's share has dropped drastically while Fiat's share has recently increased. In the first chart I created, we saw that Fiat sales were mostly commercial and Renault sales were mostly auto. Maybe because of the epidemic, people mostly stayed at their home and didn't need the new auto. In addition, commercial sales may have increased, as cargo traffic and the number of home orders increase rapidly.