Final Assignment

Alican Yılmaz

9/11/2020

Part 1: Short and Simple

1.

In the case of our country, Turkey, biggest problem to me is data transparency. I think metadata must be provided from each countries which shares the details of the data publicly. Although some parameters that affect the analyses are similar between each countries, such as R-nod, there are some significant distinctive characteristics countries have.(i.e The population of Portugal and Sweden is quite close to each other. However their accomodation density and culture differs very much. This affects both the simulation and statistical analyses) . Testing policies and demographic structures also differs from country to country which makes direct comparison less effective.

My solution to that would be, first, categorizing countries that are to be analyzed based on the government policies applied ,or demographic structures.To prevent the population difference, I would either normalise or take the percentage of cases to the specific threshold day for more reliable statistical result.

2.

My workflow for exploratory data analysis would be, first importing the raw data and preprocessing it to prepare for my analysis. Second, exploring and analyzing data to find some valuable information and gain insights about the subject that I am working on and finally, communicating my findings via reporting and/or visualization. In the example, the problem for impact is its ambiguity in terms of measurement. What the response variable will be highly affects the analysis. For example we can measure social impact by financial improvement or by happiness index. Or hypothetically speaking, based on our results some social groups might get unequally more positive impact than the other groups.(e.g white and black people or other minorities). Results may differ depending on how we approach the impact. Second point is, measuring also the combined effects of factors to the response variable is quite important.(Hypothetically speaking, funding both education and gender may have more combined impact than the mere sum of each factor. Or, reversely, funding education and job creating might have less combined impact than expected. For example, we can use two-way ANOVA to see those effects)

The problem with the last attitude(being more inclined to some action) is that, to me, one can almost always find some data which confirms their biased opinion. However, this does not mean that their claim is justified. “The questions that are asked by the data analyst”(thus their bias) and “how the data is collected” are also two important factors that affect the results.

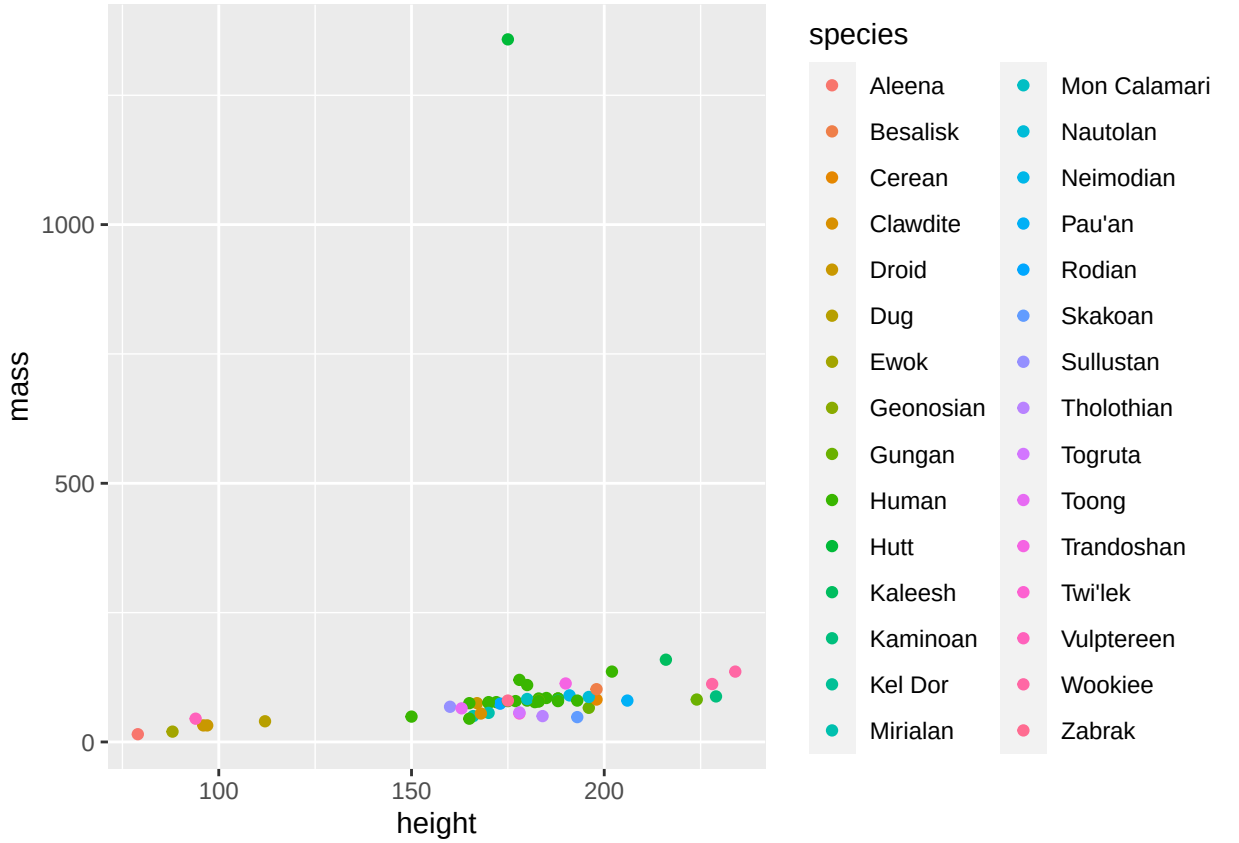
3.

I would plot the weight and height data by species. I have never watched any episodes of any starwars movies. And when I opened the data set, this is the first thing that I wondered about the data set.

```
library(dplyr)
library(tidyr)
library(ggplot2)
library(readr)
```

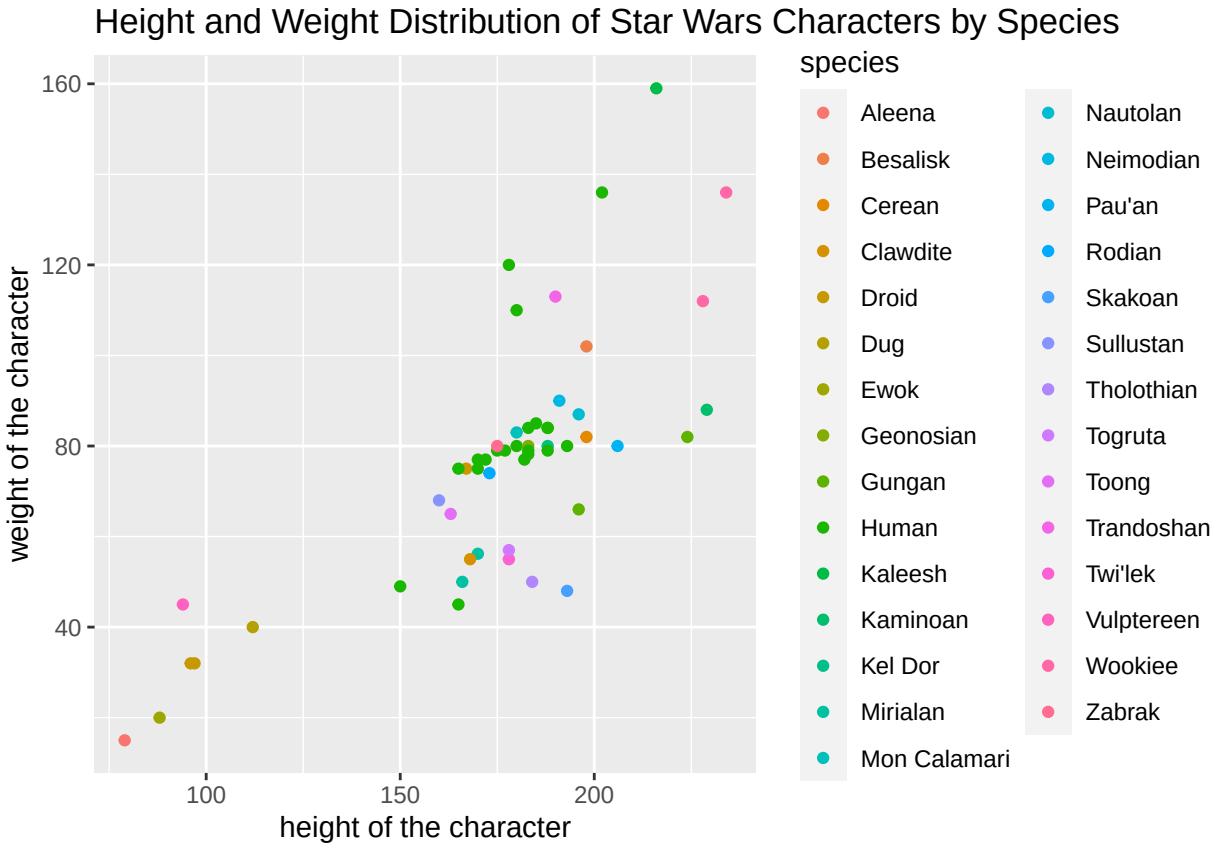
First, I am preparing my data for analysis.

```
f<-starwars %>% select(mass,height,homeworld,species)%>%
  filter(complete.cases(.))

f %>%
  ggplot(.,aes(x=height,y=mass,color=species))+geom_point()
```



I realized there is one character that distorts my data. It looks like a dwarf giant!?(Their height is 175 and weight is 1358.0) Removing that will make my data more readable.

```
f<-starwars %>% select(mass,height,homeworld,species)%>%
  filter(complete.cases(.))

f[!(f$mass==1358.0),] %>%
  ggplot(.,aes(x=height,y=mass,color=species))+geom_point()+
  labs(x="height of the character", y="weight of the character",
       title="Height and Weight Distribution of Star Wars Characters by Species")
```



From the graph, we can observe that height is concentrated between 150 and 200, whereas weight is concentrated between 70 and 90, which is mostly of "human" species. Also, we have some dwarf-like characters in the movie from Droid, Ewok and Aleena species. There are some giant-like creatures also from Kaleesh and Wookiee species.

Part 2: Extending Your Group Project

In our project we analyzed, from various perspectives(e.g. visa policy, exchange rate, seasonality etc.), the total tourists coming to Turkey by their nations. However, we forgot one important aspect: "Market Potential" and how much of this potential we have been benefited so far. It is important because it might give some insights as to how the future policies should be implemented. For my aim, first I will import the total population data from a reliable source, then I will adjust and join the `data.frames` for my analysis, and then finally I will make some analyses based on the "percentage" rather than "total number".

First, we get our visitor data:

```
mil<- read_csv("milliyetlere_gore_ziyaretci_sayisi.csv",guess_max = 100,
                      col_types = cols(
                        "Date" = col_date(format="%Y-%m")
                      ))

related_mil<-mil[1:150,]%>%
  filter(complete.cases(.))
```

And, we get our second data, total population by years and countries. Date is extracted from World Bank

```
total_population<-read_csv("total_population.csv")
total_population<-total_population %>%
```

```
  select(1,50:61)
tmp <- as.data.frame(t(related_mil[,-1]))
colnames(tmp) <- related_mil$Date
tmp <- cbind("Country Name" = rownames(tmp), tmp)
rownames(tmp) <- 1:nrow(tmp)
```

Now, lets analyze the market potential in 2019. In `part 4.1` we already determined the countries and number of tourists from that countries who has visited the Turkey most. These are, respectively:

Germany 376538.03 Russia 312899.25 United Kingdom 189129.39 Bulgaria 144545.25 Iran 134808.18 Georgia 132578.41 Netherlands 91578.99 France 71257.87 Ukraine 70682.76 Greece 55506.27

But, for a more reliable result, I will omit the the dates between `2020-04-01` and `2020-06-01`. Because the numbers during those months are quite lower than other months due to Covid-19.

```
#totalsums <-
df <- related_mil %>% select(-contains("Total"))
df <-df[1:147,]
```

Here, we obtained the average yearly total number of tourists by top 10 countries.

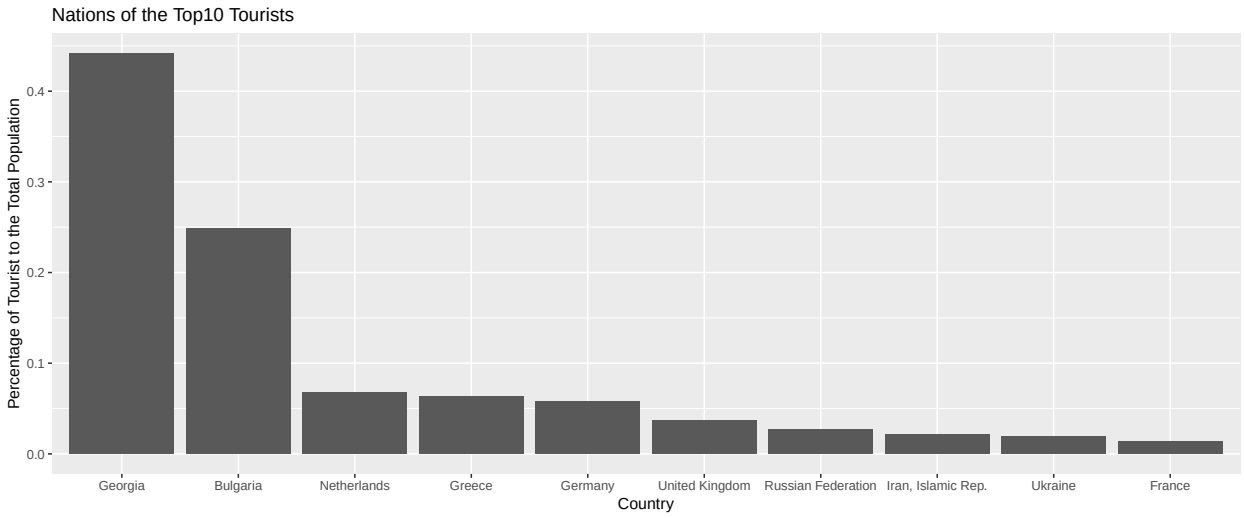
```
totalmeans <- df %>% select(-Date) %>% summarise(across(everything(),sum))%>% sort(decreasing = TRUE)
top10_m <- totalmeans %>% select(1:10)
x<-t(top10_m)
x <- cbind("Country.Name" = rownames(x), x)
rownames(x) <- 1:nrow(x)
x[,2]<-as.numeric(x[,2])/12
```

We obtained yearly average total visitors of top 10 countries. Now, we will obtain the average total population of these countries to find how much of the total population prefers Turkey as their travel destination. And we will see if any change will occur in the top 10 list!

```
total_population_avg<-data.frame("Country_Name"=total_population[,1],
                                 Total_Population=rowMeans(total_population[,-1]))
x[2,1]="Russian Federation"
x[3,1]="United Kingdom"
x[5,1]="Iran, Islamic Rep."
x[7,1]="Netherlands"
s<-total_population_avg[c(54,201,80,20,81,111,175,76,247,88),]%>%
left_join(x,by=c("Country.Name"),"copy"=TRUE)%>%
mutate(perc=as.numeric(V2)/Total_Population)
```

After preparing the data for our analysis, we can plot the top10 countries by percentage chart now.

```
ggplot(s,aes(x=reorder(Country.Name, -perc), y=perc, fill("Country.Name" )))+
  geom_col()+ theme(legend.position = "None")+
  labs(x="Country",y="Percentage of Tourist to the Total Population",
    title="Nations of the Top10 Tourists")
```

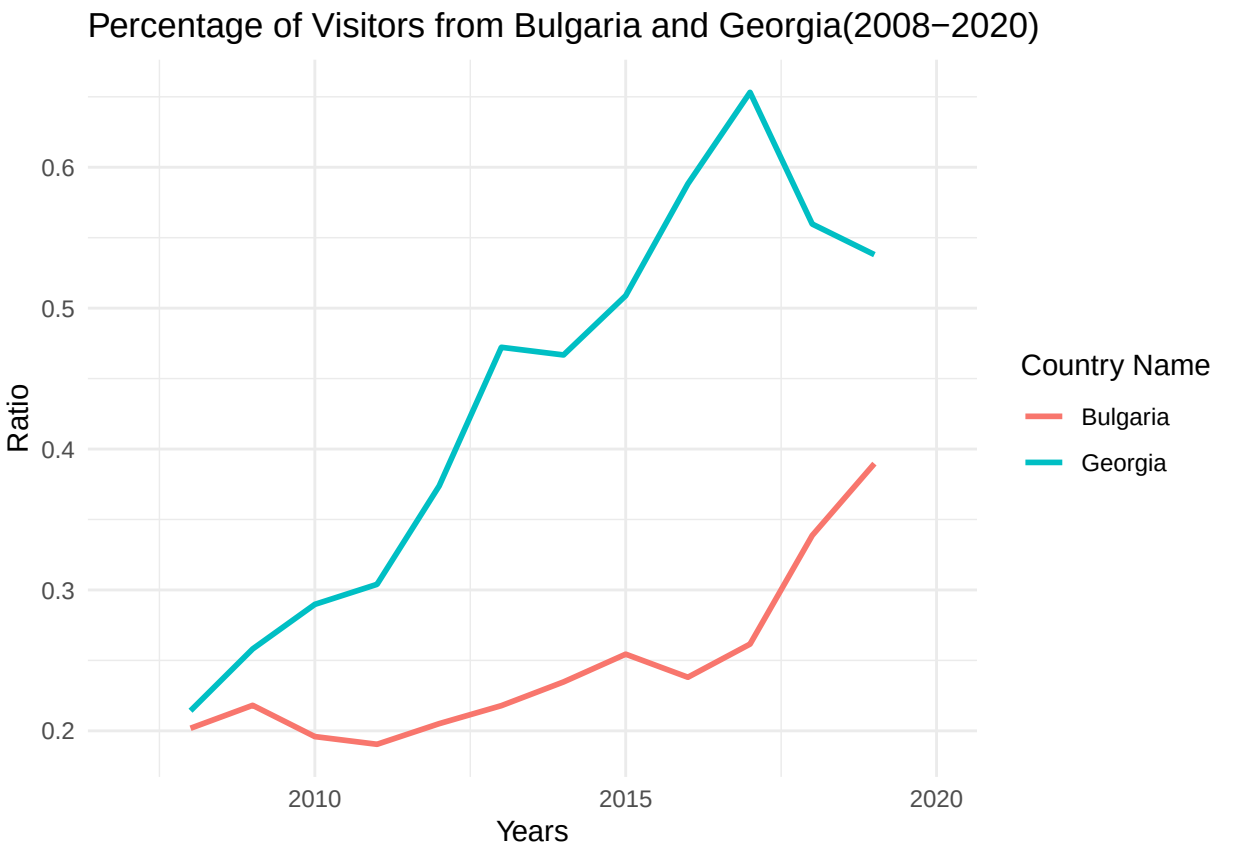


As you can see, the order completely changed now! The highest percentage of tourists compared to their population come from Georgia and Bulgaria. Top two countries - Germany,Russia- based only on the numbers are now 5th and 7th, respectively. This means a great number of tourists from these countries do not prefer Turkey, although huge numbers still visit here. Interestingly, Turkey hosts more than 0.4 of the population of Georgia which is a great share in the Georgia Market. However, here it is important to point out that same tourist can visit Turkey more than once, which can be another contributing factor that affects this percentage making it slightly higher than it really is.

```
b<-total_population[c(81,20),]%>%
  pivot_longer(-"Country Name")%>%
  rename(year=name)

a<-related_mil[1:144,] %>%
  select(Date,Georgia,Bulgaria)%>%
  mutate(year = format(Date, "%Y")) %>%
  pivot_longer(c(Georgia,Bulgaria)) %>%
  group_by(year,name)%>%
  summarise(yearly_tourist_number=sum(value))%>%
  rename("Country Name"="name")
```

```
bulgaria_georgia<-a%>%
  left_join(b,by=c("year","Country Name"))%>%
  mutate(perc=yearly_tourist_number/value)
```

```
bulgaria_georgia<-bulgaria_georgia %>%
     rename(country_name="Country Name")
bulgaria_georgia$year<-as.numeric(bulgaria_georgia$year)
bulgaria_georgia%>%
ggplot(.,aes(x=year,y=perc,color=country_name))+ geom_line(size=1)+
  labs(x="Years", y="Ratio", title="Percentage of Visitors from Bulgaria and Georgia(2008-2020)",
       color="Country Name")+
  xlim(2007,2020)+theme_minimal()
```



From the graph above, we can see the yearly percentage changes of top 2 countries, namely Georgia and Bulgaria. Peak year for Georgia is found to be 2017, with %65 percentage value. For Bulgaria it is %39 in 2019. However,again, it is important to point out that same tourist can visit Turkey more than once, which can be another contributing factor that affects this percentage making it slightly higher than it really is. Still, this graph gives a nice information about how the market potential changes yearly within a country. For example, starting from 2016, Turkey attracts more and more visitors from Bulgaria. While, this percentage tends to decrease in Georgia after 2017. This percentage method also helps us to analyze further about what went well in Bulgaria and what went wrong in Georgia. And new advertisement, touristic policies specific to a country can be implemented based on these results.

Part 3: Welcome to Real Life

3.1.a)Data gathering from multiple sources:

The data for my analysis has been gathered from Otomotiv Distribütörleri Dernegi (ODD). First I will try to make it easier to be analyzed, and then, save it in `.RData` format.

First we convert all the excel files into one data frame.

```
library(readxl)
col_names_vector<-c("brand_name","auto_dom","auto_imp","auto_total","comm_dom",
  "comm_imp","comm_total","total_dom","total_imp","total_total")
setwd("C:/Users/Alican/Desktop/car_sales")

file.list <- list.files(pattern='*.xlsx')
data_frames_not_bound<- lapply(file.list, read_excel,skip=7,col_names=col_names_vector,n_max=42)
bound_data_frames <- bind_rows(data_frames_not_bound)
```

We add month and year column and replace NA values with 0. Now our data is ready for analyze.

```
bound_data_frames<-bound_data_frames%>%
  mutate(year=2020)
bound_data_frames$year[1:504] <- 2019
bound_data_frames<-bound_data_frames%>%
  mutate(month=0)%>%
  mutate_all(~replace(., is.na(.), 0))

for (i in 1:12){
 bound_data_frames$month[(1+42*(i-1)):(42*i)]<- i
}
for (i in 1:8){
  bound_data_frames$month[(504+1+42*(i-1)):(504+42*i)]<- i
}
head(bound_data_frames)
```

```
## # A tibble: 6 x 12
##   brand_name auto_dom auto_imp auto_total comm_dom comm_imp comm_total total_dom
##   <chr>         <dbl>    <dbl>      <dbl>    <dbl>    <dbl>      <dbl>     <dbl>
## 1 ALFA ROMEO        0        8          8        0        0          0         0
## 2 ASTON MAR~        0        2          2        0        0          0         0
## 3 AUDI              0      495        495        0        0          0         0
## 4 BENTLEY           0        2          2        0        0          0         0
## 5 BMW               0      352        352        0        0          0         0
## 6 CITROEN           0       80         80        0      126        126         0
## # ... with 4 more variables: total_imp <dbl>, total_total <dbl>, year <dbl>,
## #   month <dbl>
```

Now we can save our data into R.Data file using `save(bound_data_frames, file = "*.RData")`

```
save(bound_data_frames, file = "car_sales_2019_2020_August.RData")
```

3.1.b) Exploratory Data Analysis

First, let's see total sales of brands between these years:

```
top_10_sales<-bound_data_frames%>%
  select(brand_name,total_total)%>%
  group_by(brand_name)%>%
  summarise(total_sales=sum(total_total))%>%
  arrange(desc(total_sales))%>%
  slice(1:10)
top_10_sales
```

```
## # A tibble: 10 x 2
##    brand_name total_sales
##    <chr>            <dbl>
##  1 FIAT            146227
##  2 RENAULT         118250
##  3 FORD             91769
##  4 VOLKSWAGEN       85329
##  5 PEUGEOT          51415
##  6 TOYOTA           42323
##  7 HYUNDAI          39124
##  8 DACIA            36996
##  9 OPEL             36270
```

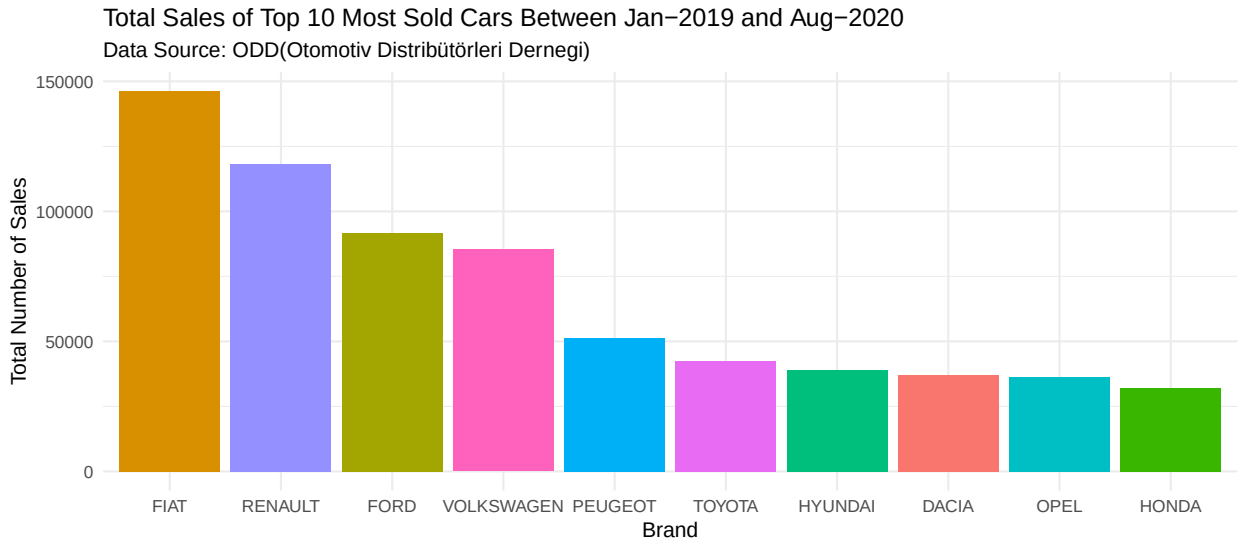
```
## 10 HONDA            32063
```

```
least_10_sales<-bound_data_frames%>%
  select(brand_name,total_total)%>%
  group_by(brand_name)%>%
  summarise(total_sales=sum(total_total))%>%
  arrange(desc(total_sales))%>%
  slice(33:42)
least_10_sales
```

```
## # A tibble: 10 x 2
##    brand_name   total_sales
##    <chr>              <dbl>
##  1 JAGUAR               365
##  2 ALFA ROMEO           320
##  3 LEXUS                178
##  4 SMART                 91
##  5 MASERATI              67
##  6 FERRARI               35
##  7 ASTON MARTIN          31
##  8 LAMBORGHINI           22
##  9 BENTLEY               18
## 10 INFINITI               0
```

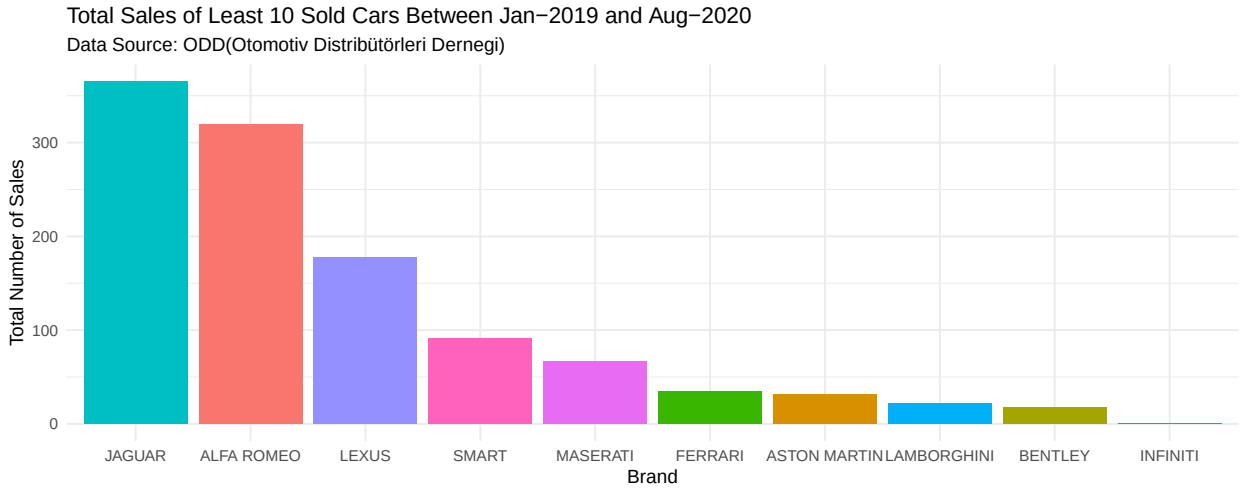
From the tables above, top3 brands are found to be Fiat, Renault and Ford, respectively and the least 3 are Lamborghini, Bentley and Infiniti, respectively. Least 10 sales are constituted of all luxury cars. You can see both most and least sold cars from the graphs below:

```
top_10_sales%>%
  ggplot(.,aes(x=reorder(brand_name, -total_sales), y=total_sales, fill=brand_name))+
  geom_col() +theme_minimal() + theme(legend.position = "None") +
  labs(x="Brand",y="Total Number of Sales",
    title="Total Sales of Top 10 Most Sold Cars Between Jan-2019 and Aug-2020",
    subtitle = "Data Source: ODD(Otomotiv Distribütörleri Derneği)")
```



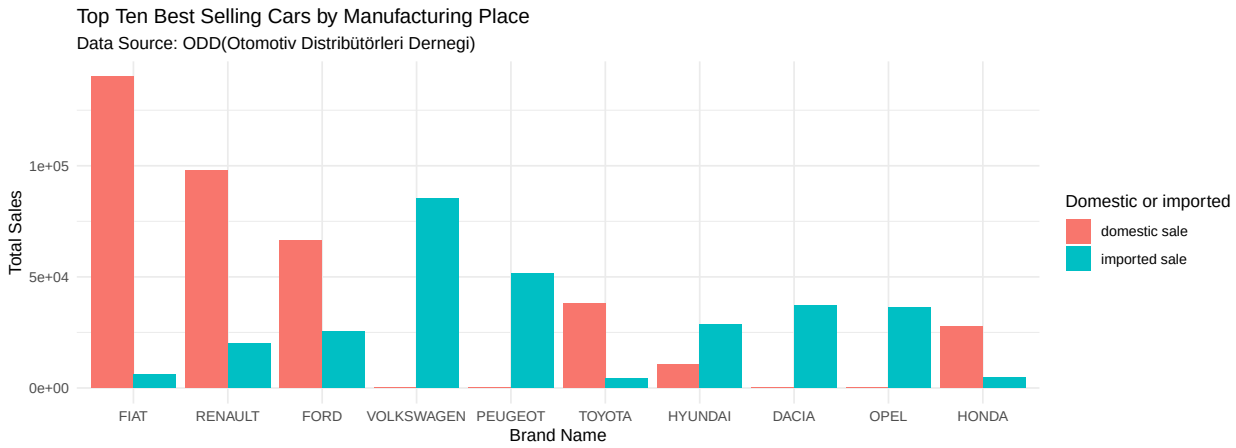

```
least_10_sales%>%
  ggplot(.,aes(x=reorder(brand_name, -total_sales), y=total_sales, fill=brand_name))+
```

```
  geom_col() +theme_minimal() + theme(legend.position = "None")+
  labs(x="Brand",y="Total Number of Sales",
    title="Total Sales of Least 10 Sold Cars Between Jan-2019 and Aug-2020",
    subtitle = "Data Source: ODD(Otomotiv Distribütörleri Derneği)")
```



Now, let's see the distribution of domestic/imported cars of top 10 cars.

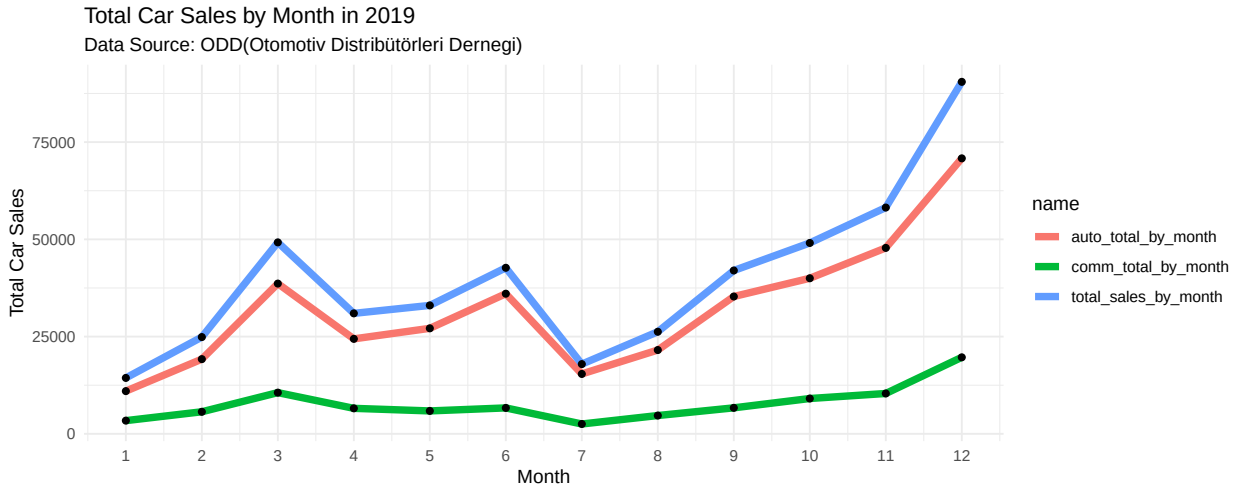
```
imp_or_dom<-bound_data_frames%>%
  select(brand_name,total_dom,total_imp,total_total)%>%
  group_by(brand_name)%>%
  summarise(allmonthstotal_dom=sum(total_dom),allmonthstotal_imp=sum(total_imp),
    total_sum=sum(total_total))%>%
  arrange(desc(total_sum))%>%
  slice(1:10)
colnames(imp_or_dom) <- c("brand_name","domestic sale","imported sale","total_sale")
imp_or_dom[,1:3]%>%
  pivot_longer(-brand_name,names_to="dom_or_imp")%>%
  ggplot(.,aes(x=reorder(brand_name,-(value)),y=value,fill=dom_or_imp))+
  geom_bar(stat="identity", position=position_dodge())+
  labs(x="Brand Name", y = "Total Sales", fill="Domestic or imported",
    title="Top Ten Best Selling Cars by Manufacturing Place",
    subtitle = "Data Source: ODD(Otomotiv Distribütörleri Derneği)") +
  theme_minimal()
```



From the graph, we can observe that in Fiat, Renault and Ford domestic types are sold significantly higher than imported types. Volkswagen, Peugeot, Dacia and Opel do not have any domestic types that has been sold.

Now, I wondered what is the peak month in the Car Sales in 2019:

```
sales_by_month<-bound_data_frames%>%
  filter(year==2019)%>%
  select(total_total,comm_total,auto_total,month)%>%
  group_by(month)%>%
  summarise(total_sales_by_month=sum(total_total),
    comm_total_by_month=sum(comm_total),auto_total_by_month=sum(auto_total))


sales_by_month %>% pivot_longer(.,-month) %>% ggplot(.,aes(x=month,y=value,color=name)) +
  geom_line(size=2)+
  labs(x="Month",y="Total Car Sales",title="Total Car Sales by Month in 2019",
    subtitle = "Data Source: ODD(Otomotiv Distribütörleri Derneği)")+
  geom_point(color="black")+
  scale_x_continuous(breaks = round(seq(min(sales_by_month$month),
    max(sales_by_month$month), by = 1),1))+
  theme_minimal()
```



From the graph, we can observe that most sales is observed in the last month of 2019, and least is during in January. Commercial Automobiles have been sold way less than normal cars unsurprisingly, and their sale trend is mostly stable. Both of the car types have been sold more in March and December. Also, both of the car types' sales has increased continuously after July until December.

Reference:

- Stackoverflow/how-can-i-read-multiple-excel-files-into-r
- mef-bda503/Instructor-Example
- Group Bıktık-R-tık Progress Journal