IE 48A 2020 Summer Term

Final Exam Answers

Fatma Nur Dumlupınar

09.14.2020

Contents

1 Part I: Short and Simple **2**

1.1 2

1.2 2

1.3 2

2 Part II: Extending Your Group Project **4**

3 Part III: Welcome to Real Life **6**

3.1 6

3.2 7

1 Part I: Short and Simple

1.1

In many analyzes, the total number of cases starting from the day of the first case was used as the comparison criterion. However, I think this is not a good measure of how much the virus has spread in a country. For example, for the same number of cases, the virus has spread more in less populated countries. These analyzes can be improved by using the ratio of the total number of cases to the population as a benchmark.

1.2

In order to concretize the research question, I think about which type of analysis can enable me to answer the question and which existing or extra variables will be needed for these analyzes. I examine the variables in the data set in terms of their usability, and their relationship between each other. With data cleaning, I increase the quality of the data and generate solutions for missing values. After completing the preliminary preparation and manipulation of the data, I search which statistical and visual techniques are suitable for the data and the question.

For the task to distribute funds, firstly I try to determine metrics to measure impact of the projects. These metrics can be how much budget they need, how much people can take advantage of the project, cruciality index of the topic of the project which measures if it is a basic need or not etc. Then, as performance measure, I try to minimize unaffordable prices of projects in total considering the following statement:

$$min \sum_{k}^{K}((need_{remaining,k} * (impactparameter_k))$$

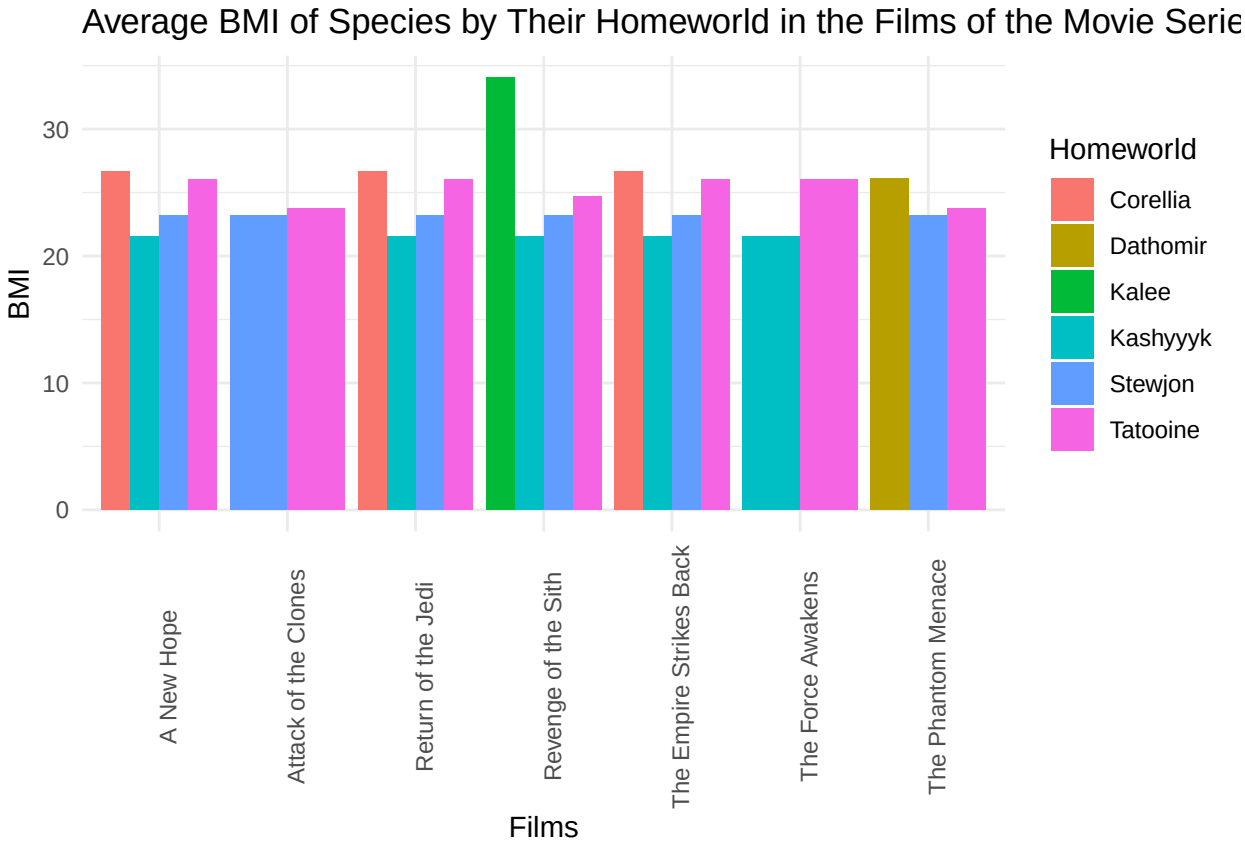
Summarizing the problem with such a mathematical statement can bring a interesting approach.

1.3

In the dataset, there are many variables about personal information for each character. So, these variables can be used to comment about characters. Since I would like to see if there is a relationship between body size of characters and their homeworld, mass and height values are good options to find a quantitative result. So, average body mass index is calculated for characters grouped by their `homeworld`. In this way, we can make interpretations about body size of characters from different homeworld. For example, we can say Kaleesh species has bigger body than the others like **here**.

```
starwars1<-starwars %>% unnest(films) %>% unnest(vehicles) %>% unnest(starships)%>%
  group_by(homeworld,films )%>%
  summarise(bmi=mean(mass/((height/100)*(height/100))))

ggplot(starwars1,aes(x=films,y=bmi,fill=homeworld))+geom_col(position="dodge")+
  theme_minimal()+
  theme(axis.text.x = element_text(angle=90))+
  labs(x="Films",y="BMI" ,fill="Homeworld",
       title="Average BMI of Species by Their Homeworld in the Films of the Movie Series ")
```


Average BMI of Species by Their Homeworld in the Films of the Movie Serie
BMI
30
20
10
0
A New Hope
Attack of the Clones
Return of the Jedi
Revenge of the Sith
The Empire Strikes Back
The Force Awakens
The Phantom Menace
Films
Homeworld
Corellia
Dathomir
Kalee
Kashyyyk
Stewjon
Tatooine

2 Part II: Extending Your Group Project

When analyzing the effect of visa policy on number of tourists in the group project, we used Turkey's incoming tourist data with nationalities. Although this data can give some insight about whether there is an effect of visa policy on tourists or not, it is better to make further analysis to comment about preferences of tourists who are subject to different visa policies.

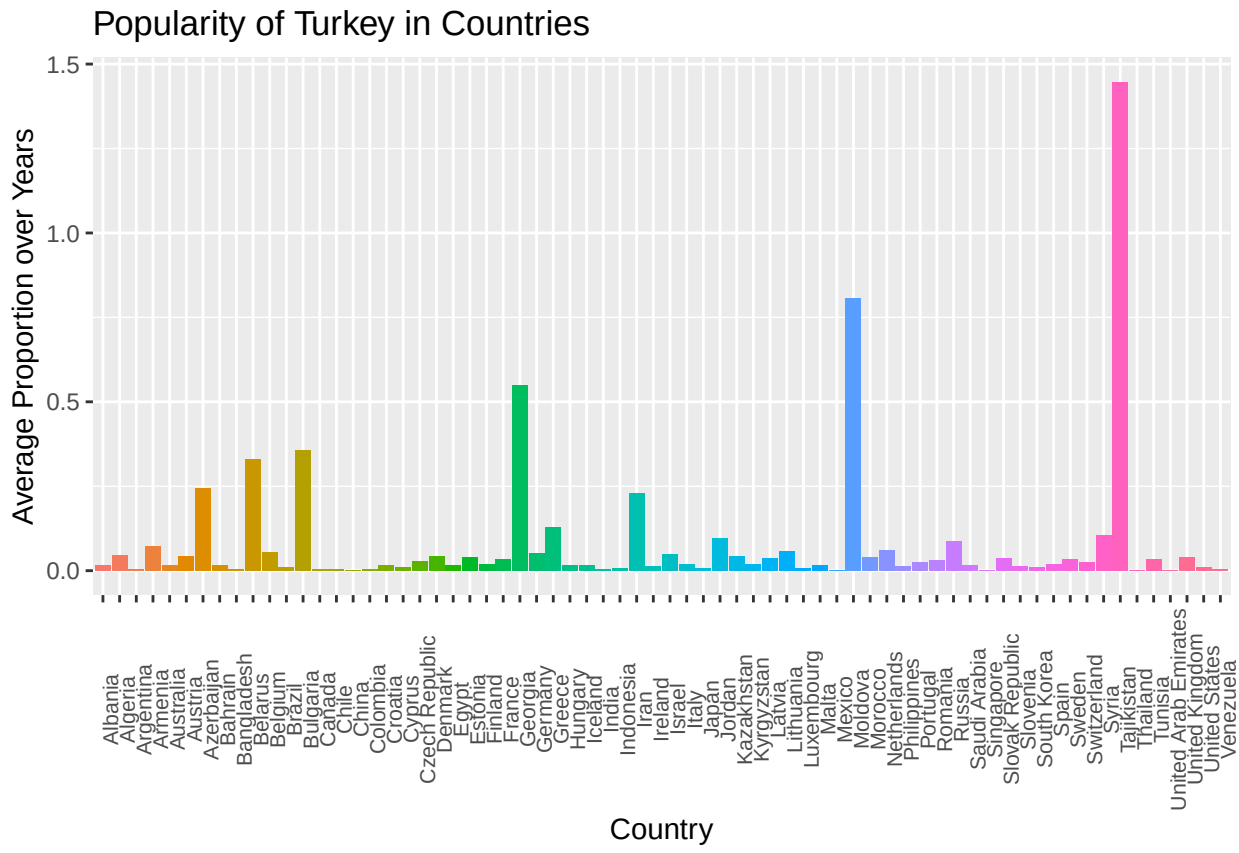
To determine the popularity of Turkey in a specified country, we need additional data which shows total number of citizens there who go abroad as tourists. In this way, we can compute the proportions that shows how many citizens preferred Turkey among all citizens going abroad. So, number of departures data is downloaded and preprocessed.

```
outgoing<-read_excel("Number_of_Departures.xls")
```

However, not every countries in visa policy analysis data `ranked_final` are also in new data because the data is not from the same source. This might cause not to be able to analyze all the countries, but still we can make analysis for common 58 countries in the datasets.

```
joined<-inner_join(outgoing,ranked_final)
proportions_yearly<-joined%>%filter(complete.cases(.)) %>%group_by(Country,Year,Visa)%>%
  summarise(proportion=total/`Yearly Outgoing Tourists`)%>%
  arrange(desc(proportion))
mean_proportion<-proportions_yearly%>%ungroup(Year)%>%summarise(avg_proportion=mean(proportion))
```

After the data frame in visa policy analysis and the new data frame are joined, proportion values as popularity metrics are calculated for all countries. In the first plot, we can see popularity of Turkey in countries.

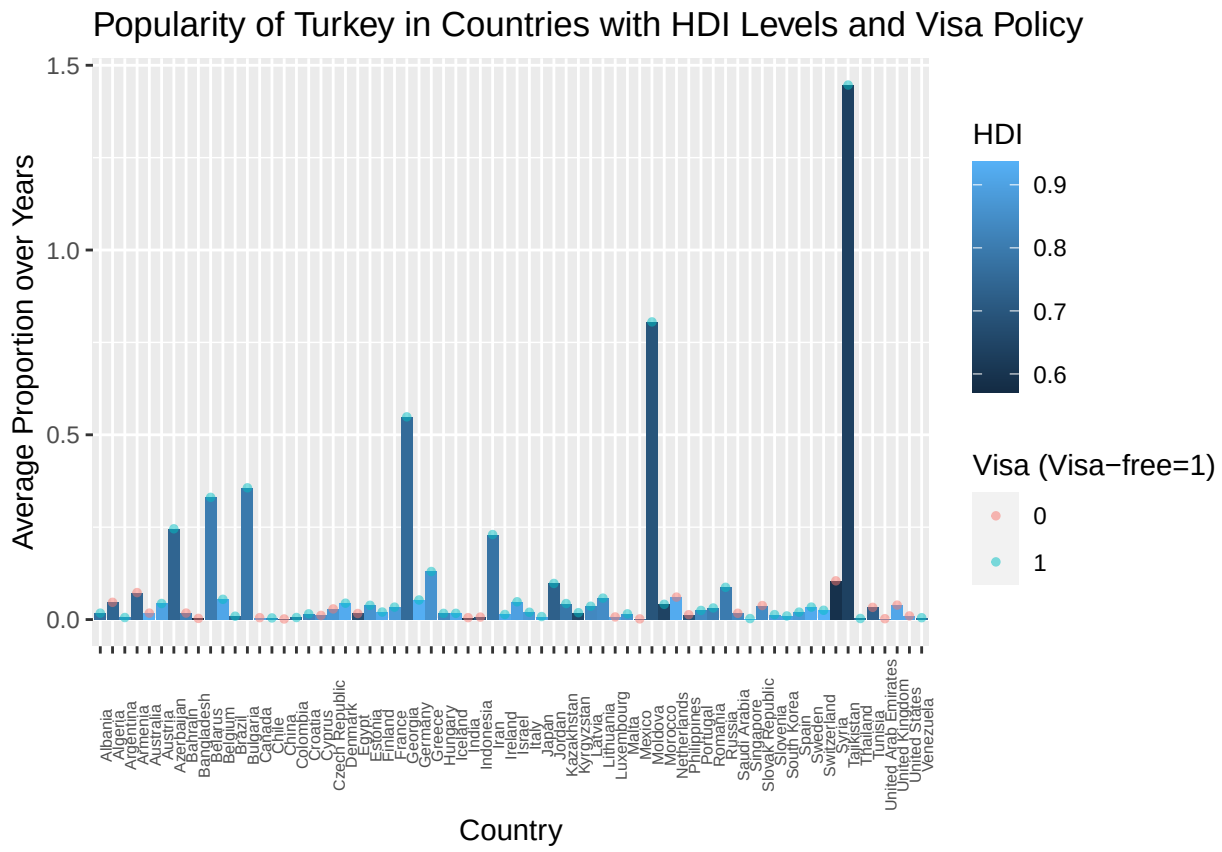


Proportion values are expected to be less than or equal to 1 because it represents those who prefer Turkey among all citizens going abroad as tourists. However, Tajikistan's proportion value is more than 1, which arises probably from mismatch of two datasets.

Now, popularity values should be compared with visa policies to decide if there is an effect of visa on the numbers. Visa requirement data is available in the joined data frame.

However, it is hard to say that high or low popularity values are only because of the visa policy. Therefore, HDI level , an index measuring development level of a country considering several indicators, in the `avg_HDI` data can be added to the plot to see whether citizens of a country have enough economic and social potential to visit Turkey.

```
joined_proportion_HDI<-inner_join(avg_HDI, mean_proportion)
visa_stat<-ranked_final%>%select(Visa,Country)%>%group_by(Country)%>%summarise(Visa=mean(Visa))
joined_proportion_HDI_visa<-inner_join(joined_proportion_HDI,visa_stat)
```



Turkey hosts tourists from countries with "high" and "very high" HDI level most as it can be seen in the development level analysis in our project. So, if popularity of Turkey in a country with high HDI level is low, the reason might be visa restrictions. So, citizens in Australia, Canada, United States, Luxembourg and United Arab Emirates can be affected by visa policy negatively, but low level of tourists can also result from distance for the countries Australia, Canada and United States.

On the other hand, if popularity of Turkey in a country with low HDI level is high, the reason might be that Turkey allows visa-free travel for these countries. Tajikistan, Moldova can be given as examples for these countries. Also, since HDI levels in countries Azerbaijan, Belarus, Bulgaria, Georgia, Iran are not so high, they can also be considered the same.

3 Part III: Welcome to Real Life

3.1

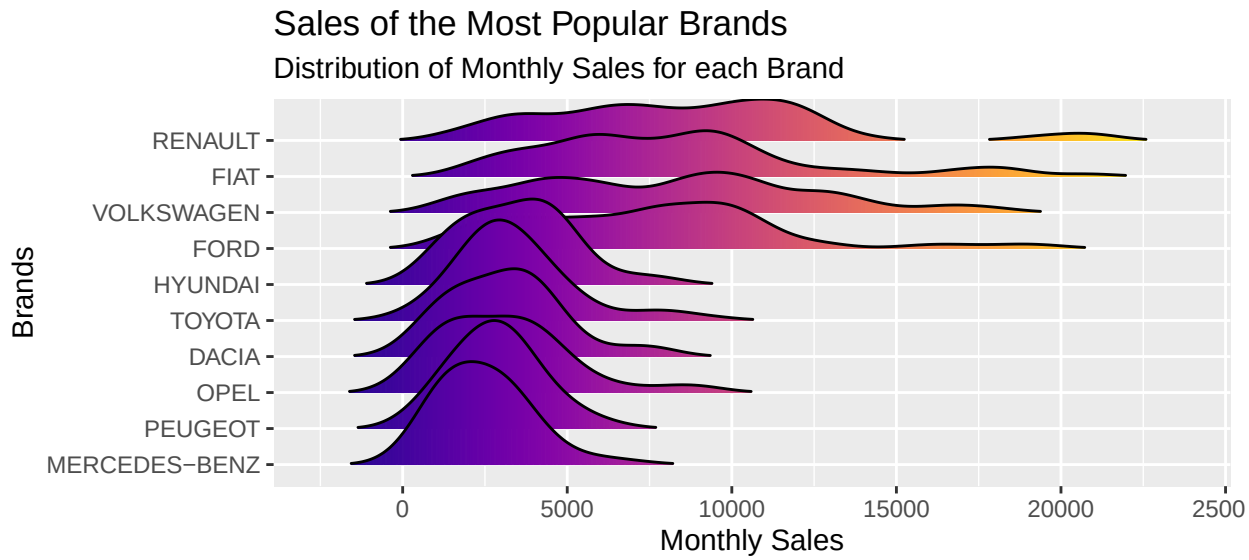
```
df<-readRDS(gzcon(url(
"https://github.com/pjournal/boun01g-data-mine-r-s/blob/gh-pages/Final%20TakeHome/all_data?raw=true")))
```

In the binded data, there are 12 columns and 3219 rows, which include car sales for the years from 2014 to 2020. The sales are divided into 4 categories. These are domestic and imported autos, domestic and imported light commercial vehicles.

Also, total values are provided with different combinations like total auto sales, total light commercial vehicle sales, total domestic car sales, imported car sales, and grand total. So, there are 9 columns related to sales, 2 columns for years and months and 1 column for brand name.

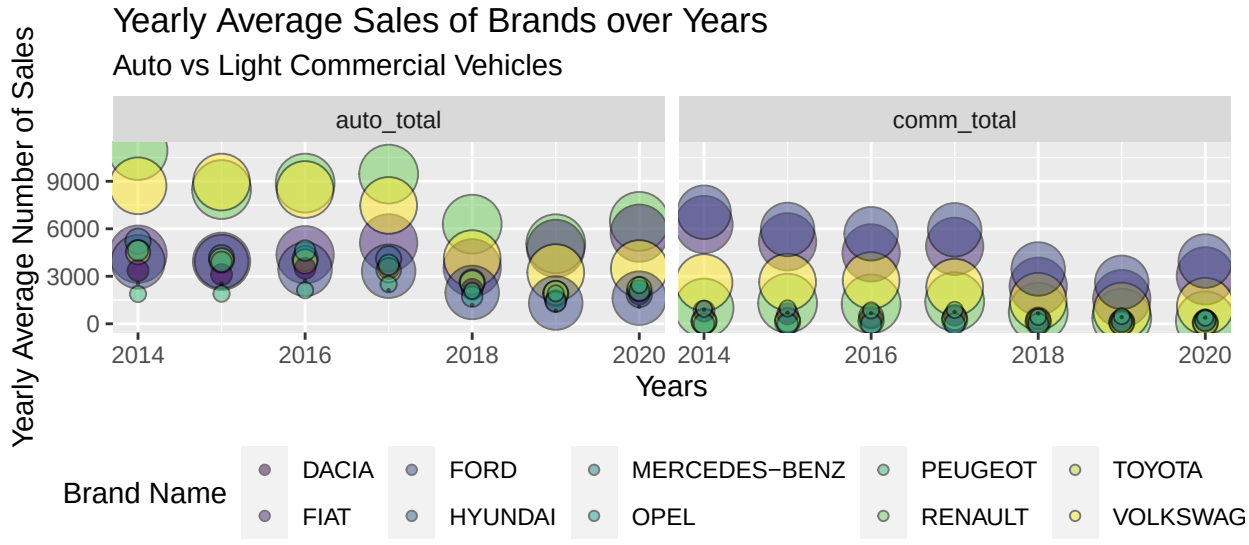
3.2

First, I decided to determine the most popular 10 brands on which I make the analysis. For this, main data is grouped brand name, and popularity values of brands are calculated considering how much percentage they have in the grand total of 7 years. Then, their monthly total sales, in other words 12 sales data for each year, are plotted on the first plot. Brands are ordered on the y axis from the most popular one to the least.



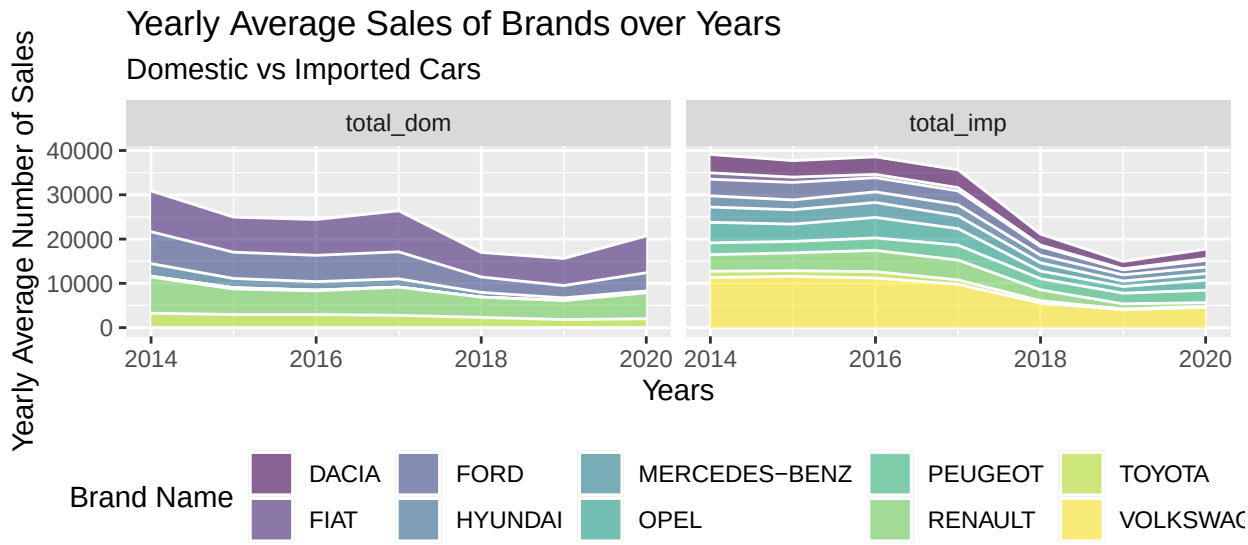
It is possible to see the monthly sales values of the most popular 4 brands are distributed in wider range than the others. However, minimum values of all the brands are close to each other.

For the second analysis, popularity values are merged to the main data. Also, pivot_longer() function is used to make the data suitable for the rest of the plots. Similarly, the most popular 10 brands are used. Yearly average sales are calculated for 7 years for auto and light commercial vehicles. The size of cycles are scaled with the popularity values of brands.



Although Renault has the highest yearly average auto sales, the best seller of light commercial vehicles is Ford. Sales are fluctuating but there is a slump in 2018.

In addition, yearly average sales for domestic and imported cars are also calculated for the third plot.



The same decrease has also occured in domestic and imported car sales as expected. This starts in 2018 because of increasing exchange and interest rates and continue to decrease with smaller difference in 2019. However, the sales start to increase again in 2020 because of decreasing interest rates.

So far, popularity values which are calculated by considering their grand total values are used in the analysis. However, the popularity change by each category. The popular brands in each category can be seen below:

Brand Name	Popularity for Autos
RENAULT	11.603067
VOLKSWAGEN	9.430720
FIAT	6.792983
HYUNDAI	5.013116
TOYOTA	4.737802
OPEL	4.608813
FORD	4.117826
DACIA	4.053384
PEUGEOT	3.213710
NISSAN	2.930032

Brand Name	Popularity for LCV
FORD	7.2066148
FIAT	5.6888336
VOLKSWAGEN	2.8994461
RENAULT	1.3862769
PEUGEOT	1.0496234
MERCEDES-BENZ	0.9494806
CITROEN	0.8837389
DACIA	0.6453206
MITSUBISHI	0.4934188
TOYOTA	0.3821653

Brand Name	Popularity for Domestic Cars
FIAT	11.4708711
RENAULT	8.2886268
FORD	7.3888089
TOYOTA	3.8500377
HYUNDAI	2.2100391
HONDA	2.0527707
ISUZU	0.2593399
KARSAN	0.1641864
CITROEN	0.1612725
PEUGEOT	0.1209806

Brand Name	Popularity for Imported Cars
VOLKSWAGEN	12.330167
RENAULT	4.700717
DACIA	4.698705
OPEL	4.666882
PEUGEOT	4.142353
FORD	3.935632
MERCEDES-BENZ	3.705704
NISSAN	3.091053
HYUNDAI	3.023739
SKODA	2.747104