Brazilian E-Commerce Public Dataset

Prashant Wakchaure | Student No. 20200126

19/12/2020

Task 1: Analysis

For this task, I went ahead and chose the "Brazilian E-Commerce Public Dataset by Olist" dataset from Kaggle: https://www.kaggle.com/olistbr/brazilian-ecommerce. It also suffices the minimum requirement of 2 categorical and 3 numerical variables. I'll perform various types of analysis on the dataset to infer out significant results in the forms of summaries, dataframes, tables and numerous plots. In the end, I'll also demonstrate the correlation between the numeric variables from the dataset.

1. Read the data into R.

```
comm <- read.csv("df.csv")
```

---

2. See the dimensions of the dataset.

The dataset has 1,13,367 rows and 32 columns.

```
nrow(comm)
```

```
[1] 113367
```

```
ncol(comm)
```

```
[1] 32
```

```
dim(comm)
```

```
[1] 113367     32
```

---

3. Check for null values.

There are no null values in the dataset.

```
is.null(comm)
```

```
[1] FALSE
```

4. Evolution of E-Commerce in Brazil

We will firstly see how the E-commerce in Brazil has flourished over time (years) according to our dataset.

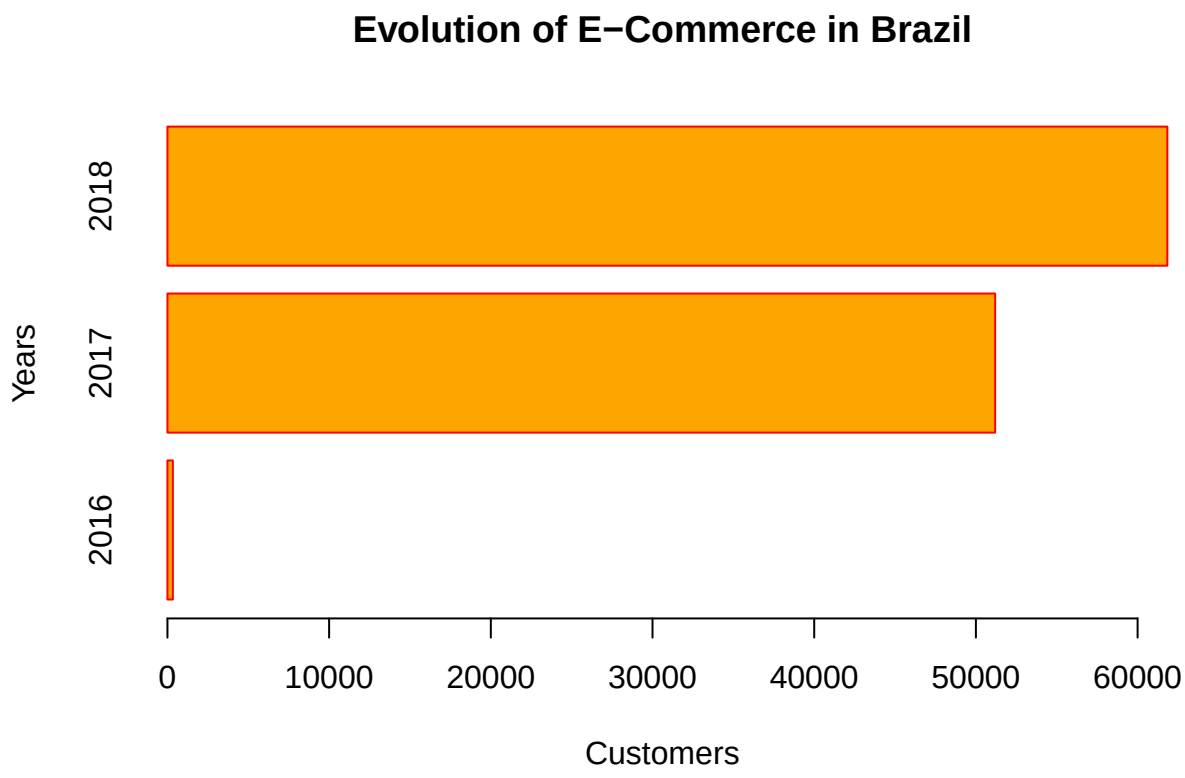
```
get_year <- as.POSIXct(comm$order_purchase_timestamp)
get_year <- format(get_year, "%Y")
head(get_year)
```

```
[1] "2017" "2017" "2017" "2017" "2017" "2017"
```

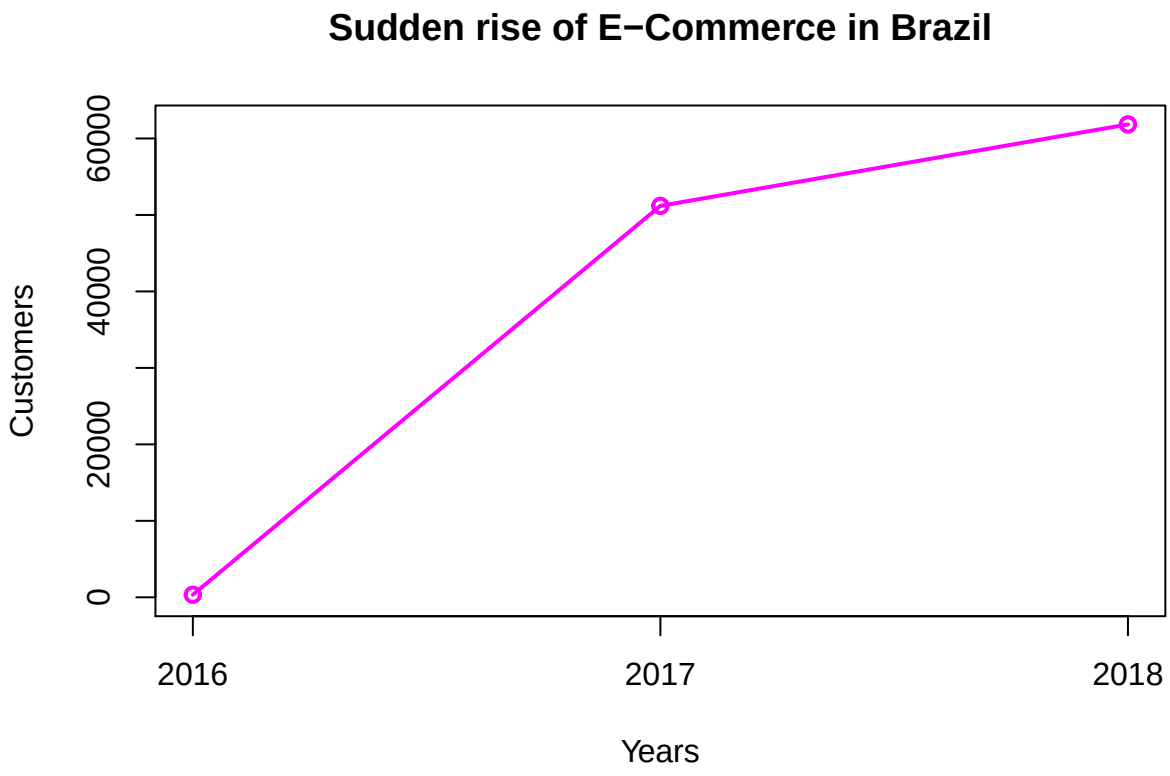
```
flourished <- table(get_year)
flourished
```

```
get_year
 2016  2017  2018
  335 51191 61841
```

```
barplot(flourished,xlab="Customers",ylab="Years",col="orange",horiz = TRUE,
main="Evolution of E-Commerce in Brazil",border="red")
```



```
plot(flourished,type = "o", col = "magenta", xlab = "Years", ylab = "Customers",
   main = "Sudden rise of E-Commerce in Brazil")
```



We see that there is a sudden rise after 2016 to 2017 and furthermore to 2018 in the E-Commerce in Brazil.

---

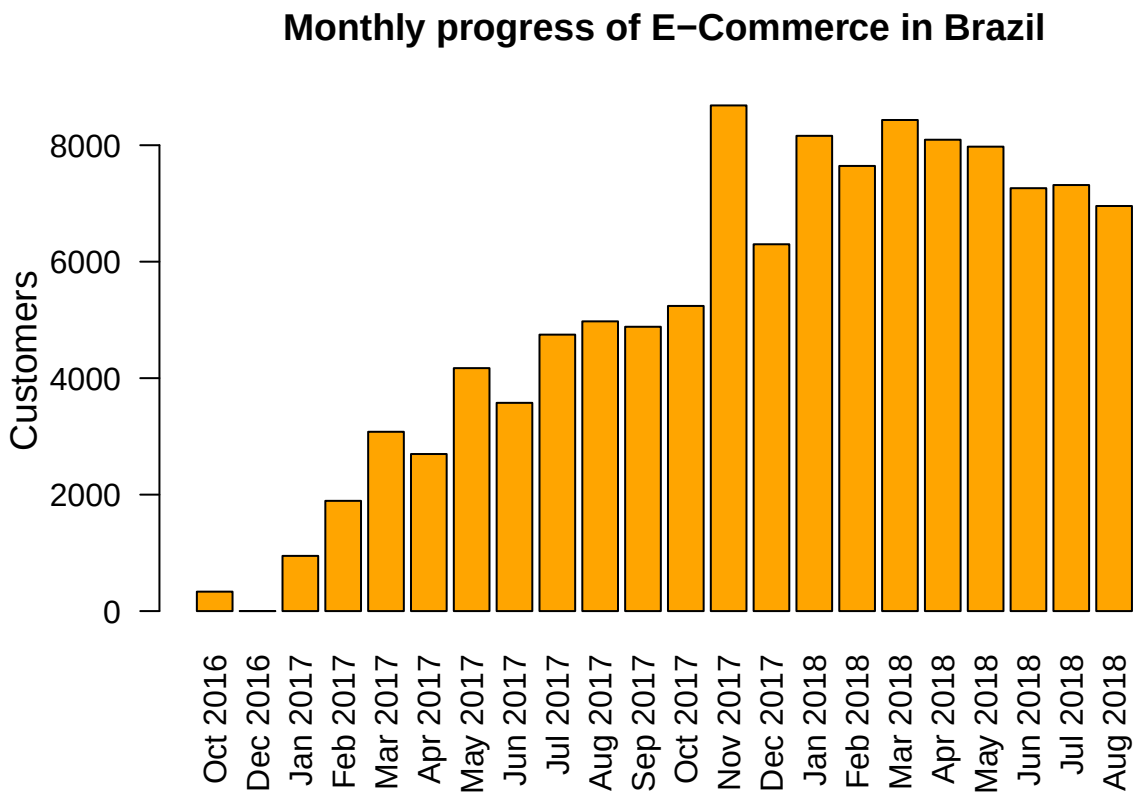
5. Monthly progress of E-Commerce in Brazil

We will now see how the E-commerce in Brazil has flourished over time (months) according to our dataset.

```
get_month <- as.POSIXct(comm$order_purchase_timestamp)

library(zoo)
get_month <- as.yearmon(get_month, "%b%Y")
head(get_month)
```

```
[1] "Oct 2017" "Oct 2017" "Oct 2017" "Aug 2017" "Aug 2017" "Oct 2017"
```

```
flourished_months <- table(get_month)

barplot(flourished_months,names.arg=unique(sort(get_month)),
        las=2,cex.lab=1.2,
        ylab="Customers",col="orange",
        main="Monthly progress of E-Commerce in Brazil",
        border="black")
```



We see here that Nov 2017 has the highest number of customers, this might be due to Black Friday. And it is also evident that the customers in 2016 were very less. There is a significant distinguishment between the right and left side of Nov 2017.

---

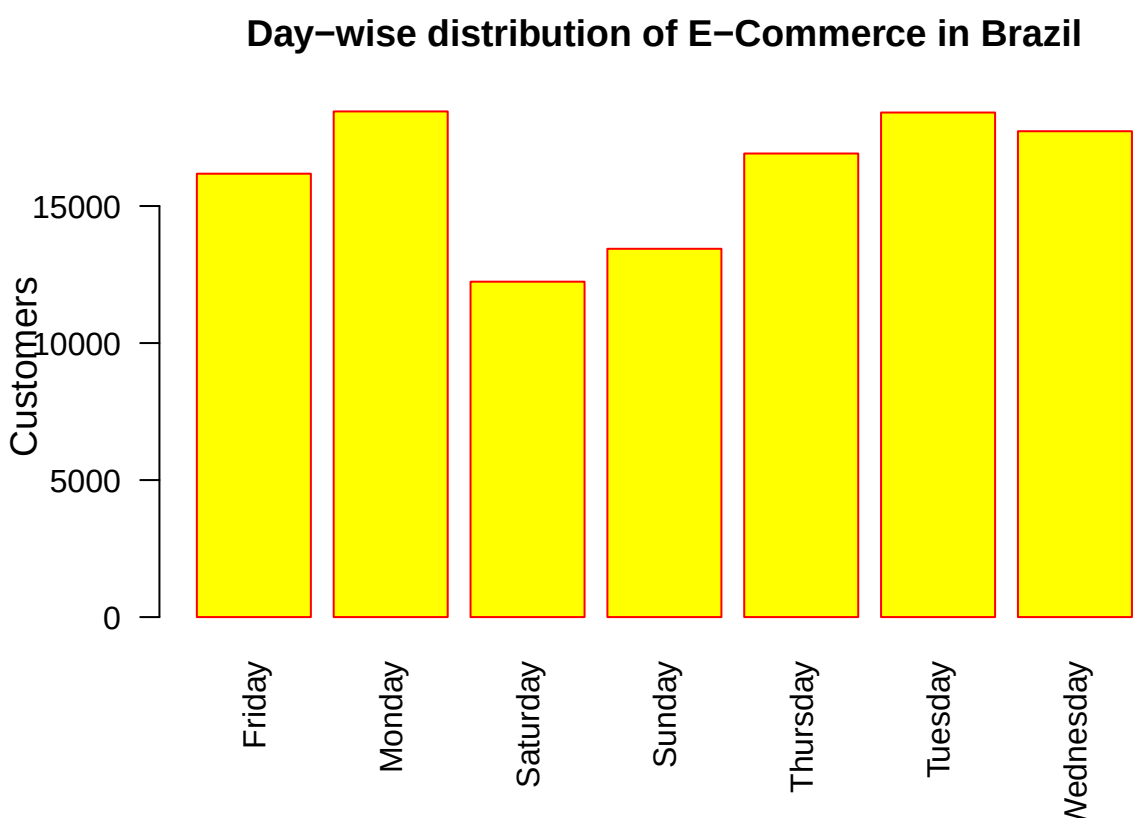
6. Day-wise distribution of E-Commerce in Brazil

We will now see how the E-commerce in Brazil has flourished over time (months) according to our dataset.

```
get_data <- as.POSIXct(comm$order_purchase_timestamp)
get_day <- format(get_data, "%A")
head(get_day)
```

```
[1] "Monday"    "Monday"    "Monday"    "Tuesday"   "Wednesday" "Monday"
```

```
flourished_day <- table(get_day)

barplot(flourished_day,names.arg=unique(sort(get_day)),
        las=2,cex.lab=1.2,
        ylab="Customers",col="yellow",
        main="Day-wise distribution of E-Commerce in Brazil",
        border="red")
```



It is evident that the online shopping is done mostly on weekdays. The weekends are not that occupied for online shopping as people prefer to go out. So, the E-Commerce industry was not totally capitalized during that time. But now, it is completely the different case.

---

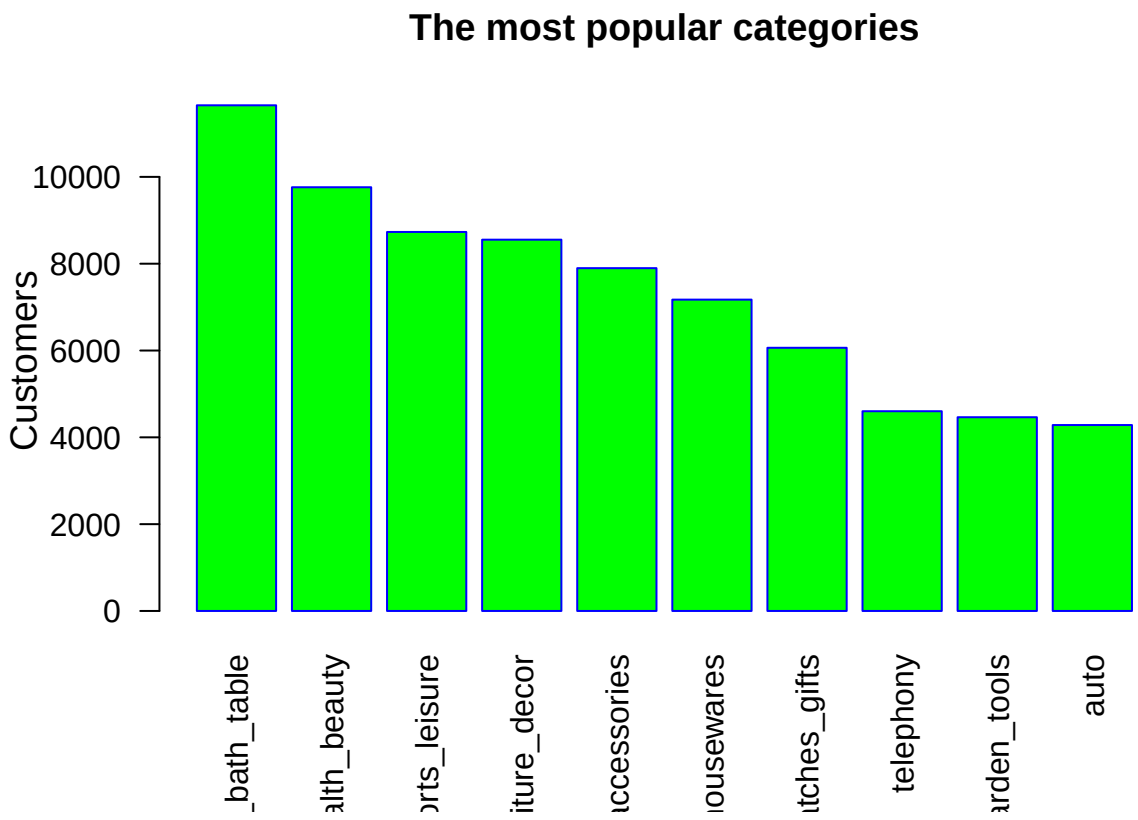
7. The most popular categories

We will now see the most popular categories which are purchased by the customers.

```
pop_cat <- as.data.frame(table(comm$product_category_name_english))
pop_cat_sort <- pop_cat[order(pop_cat$Freq,decreasing = TRUE),]
pop_cat_sort <- pop_cat_sort[1:10,]
pop_cat_sort
```

```
                   Var1  Freq
8         bed_bath_table 11649
44         health_beauty  9761
66        sports_leisure  8731
40       furniture_decor  8553
16 computers_accessories  7897
50            housewares  7172
71         watches_gifts  6063
69             telephony  4601
43          garden_tools  4463
6                  auto  4283
```

```
barplot(pop_cat_sort$Freq,names.arg=pop_cat_sort$Var1,
        las=2,cex.lab=1.2,
        ylab="Customers",col="green",
        main="The most popular categories",
        border="blue")
```



As we can see, the home accessories are purchased in abundance rather than any other products.

---

8. Average price of each Category

Herein I calculated the average price of each Category to see the most costly category.

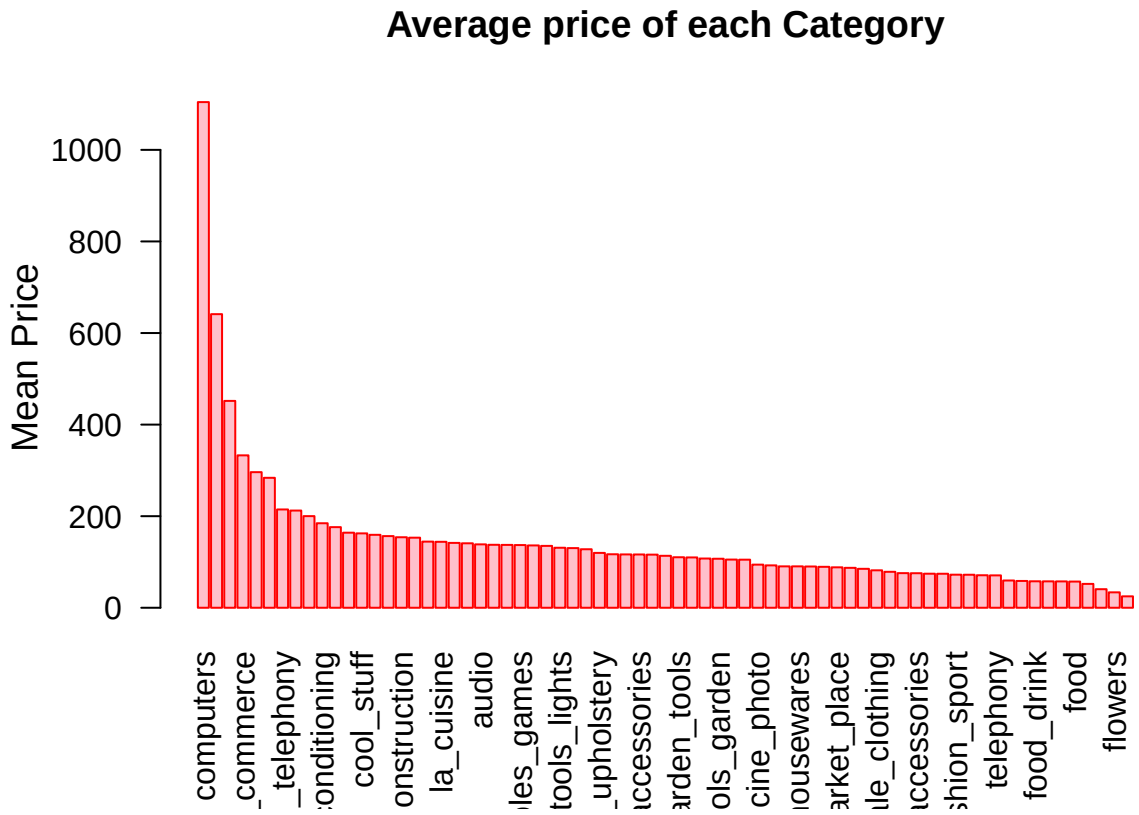
```
avg <- data.frame(comm$product_category_name_english, comm$price)

colnames(avg) = c("Categories", "Price")
avg <- avg[order(avg$Price,decreasing = TRUE),]

x <- aggregate(.~Categories, data=avg, mean)
x <- x[order(x$Price,decreasing = TRUE),]
x
```

```
                Categories      Price
```

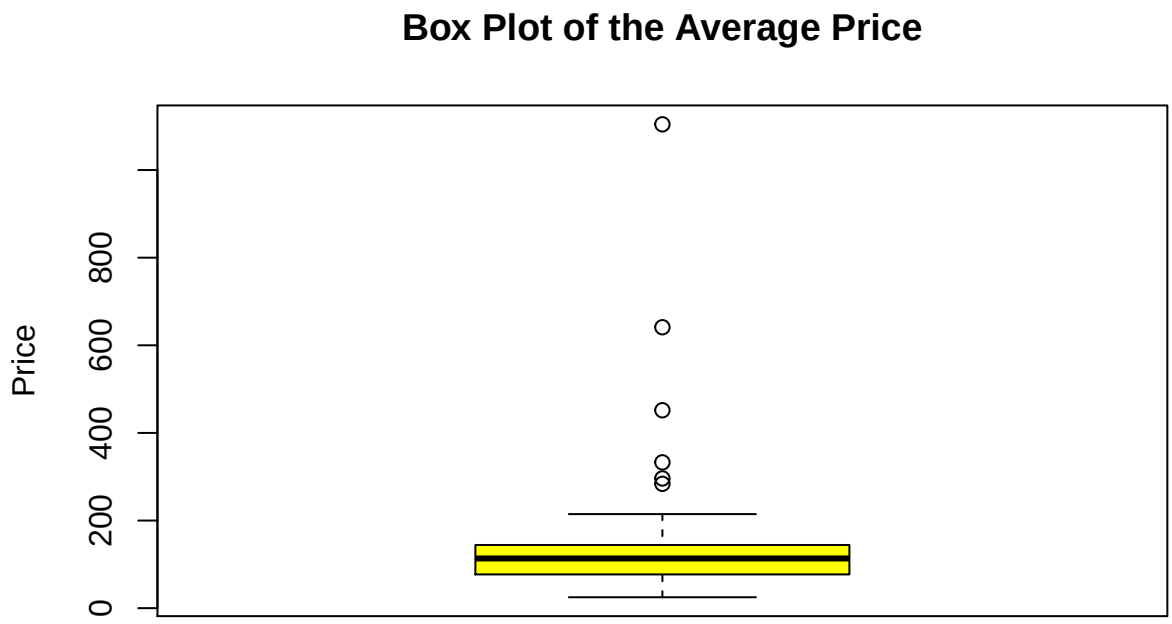
15	computers	1104.31769
65	small_appliances_home_oven_and_coffee	641.19413
46	home_appliances_2	451.72996
1	agro_industry_and_commerce	332.83569
57	musical_instruments	296.10856
64	small_appliances	283.81782
35	fixed_telephony	214.68766
20	construction_tools_safety	212.29737
71	watches_gifts	200.09283
2	air_conditioning	184.56116
39	furniture_bedroom	176.03950
52	kitchen_dining_laundry_garden_furniture	163.93361
21	cool_stuff	162.43949
58	office_furniture	159.30609
56	music	156.45500
18	construction_tools_construction	153.95071
23	costruction_tools_tools	153.07524
51	industry_commerce_and_business	144.40378
53	la_cuisine	143.99875
62	security_and_services	141.64500
6	auto	140.76156
5	audio	138.44569
41	furniture_living_room	137.48633
49	home_construction	137.28081
17	consoles_games	136.92281
48	home_confort	136.09074
7	baby	135.10208
19	construction_tools_lights	131.01903
44	health_beauty	130.25440
54	luggage_accessories	127.78099
42	furniture_mattress_and_upholstery	119.77950
3	art	116.79531
70	toys	116.44613
16	computers_accessories	116.35278
60	perfumery	116.01054
66	sports_leisure	113.43692
43	garden_tools	110.29690
61	pet_shop	109.88156
63	signaling_and_security	107.49060
22	costruction_tools_garden	107.03130
59	party_supplies	105.22644
45	home_appliances	104.99542
14	cine_photo	94.17423
8	bed_bath_table	92.52589
67	stationery	90.38793
50	housewares	90.37754
68	tablets_printing_image	90.11851
32	fashion_shoes	89.39799
55	market_place	88.26923
40	furniture_decor	87.19149
9	books_general_interest	85.03069
31	fashion_male_clothing	81.78348
10	books_imported	78.48915
4	arts_and_craftmanship	75.58375

```
29                fashion_bags_accessories   75.43242
26                             dvds_blu_ray   74.36478
30               fashion_childrens_clothes   74.27857
33                           fashion_sport   72.22483
34             fashion_underwear_beach   72.10464
11                    books_technical   70.96647
69                          telephony   70.72875
25                              drinks   59.52510
28            fashio_female_clothing   58.55422
38                         food_drink   57.77394
13              christmas_supplies   57.65697
27                     electronics   57.56195
37                            food   57.19841
12                cds_dvds_musicals   52.14286
24               diapers_and_hygiene   40.56189
36                           flowers   33.63758
47                home_comfort_2   24.94097
```

```
barplot(x$Price,names.arg=x$Categories,las=2,cex.lab=1.2,ylab="Mean Price",col="pink",
main="Average price of each Category",border="red")
```




```
boxplot(x$Price,ylab = "Price",
        main = "Box Plot of the Average Price", col = "yellow")
```



As evident from the graphs above, it is clear that the computers are the costliest & the home products are comparatively cheaper. That's the reason why the computers are not most popular categories, since the public tend to buy the cheaper products more.

---

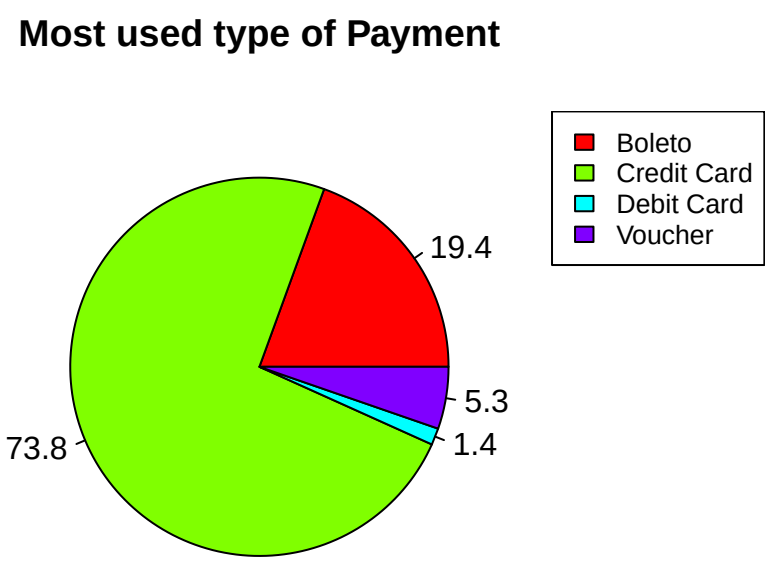
9. Most frequent type of Payment

Herein, I wanted to see the most frequent type of payment, so I plotted a pie chart for the same.

```
pop_cat <- as.data.frame(table(comm$payment_type))
pop_cat
```

```
         Var1  Freq
1      boleto 22040
2 credit_card 83695
3  debit_card  1621
4     voucher  6011
```

```
piepercent<- round(100*pop_cat$Freq/sum(pop_cat$Freq), 1)
pie(pop_cat$Freq,piepercent, main = "Most used type of Payment", col = rainbow(length(pop_cat$Freq)))
legend("topright", c("Boleto","Credit Card","Debit Card","Voucher"), cex = 0.8,
   fill = rainbow(length(pop_cat$Freq)))
```



It is clearly evident that most of the customers use Credit Card, which is around 73.8% customers.

---

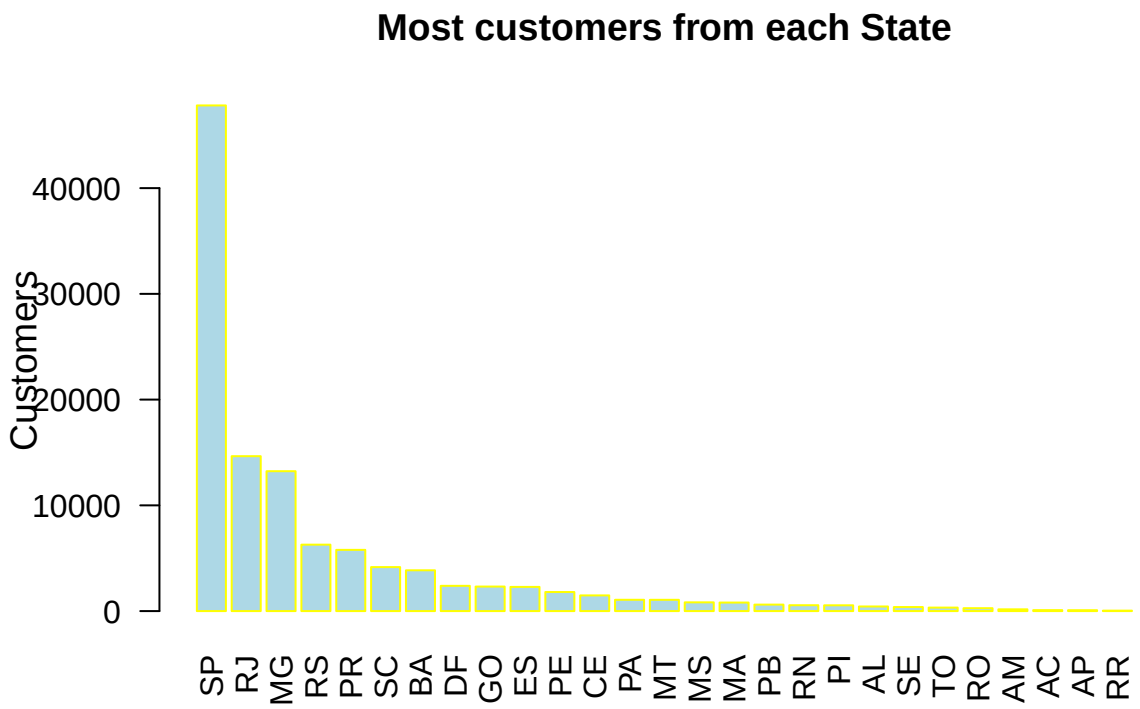
10. Distribution of customers over the states

Herein, we see the no. of customers from each states in the given dataset.

```
most_cus <- as.data.frame(table(comm$customer_state))

most_cus_sort <- most_cus[order(most_cus$Freq,decreasing = TRUE),]
head(most_cus_sort)
```

```
   Var1  Freq
26   SP 47819
19   RJ 14648
11   MG 13230
23   RS  6282
18   PR  5790
24   SC  4161
```

```
barplot(most_cus_sort$Freq,names.arg=most_cus_sort$Var1,
        las=2,cex.lab=1.2,
        ylab="Customers",col="lightblue",
        main="Most customers from each State",
        border="yellow")
```



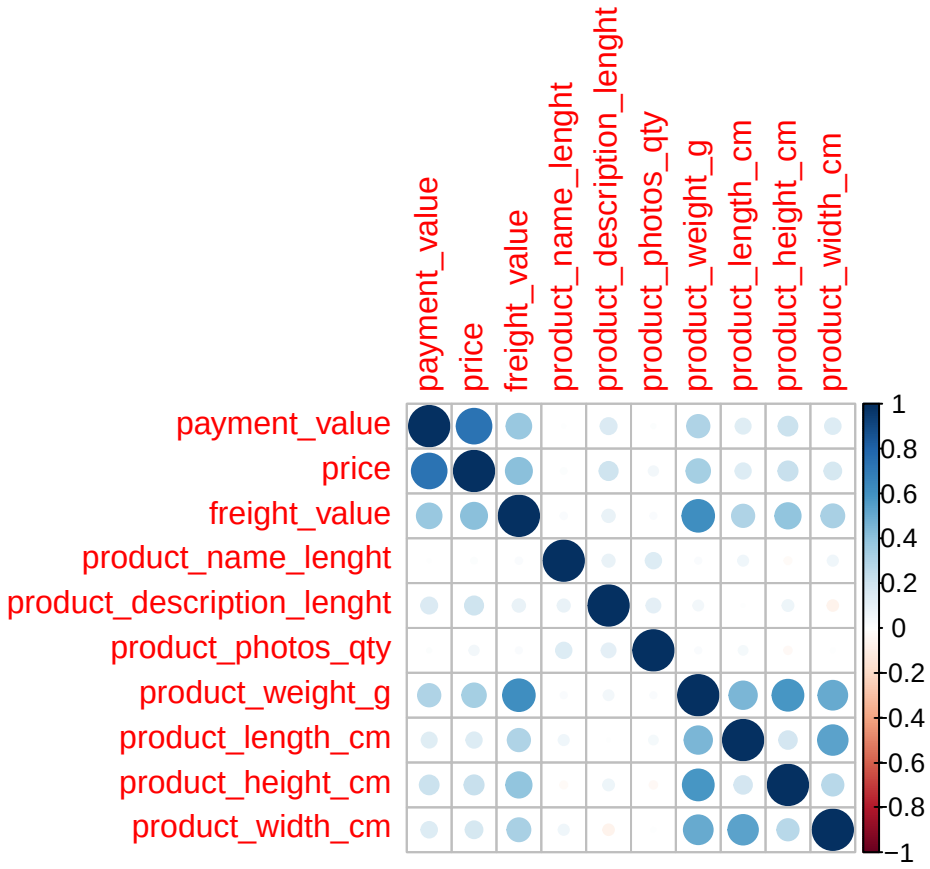
We see that Sao Paolo has the most customers in Brazil as it is the most famous city worldwide and is a cultural center of Brazil.

---

11. Correlation between the numeric variables

Herein, I wanted to see the highest correlation between 2 variables, so firstly I plotted the correlation heatmap and then used some plots and other calculations to show which variables have the highest correlation.

```
x <- comm[, sapply(comm, class) == "numeric"]
corr_mat <- round(cor(x),2)

corrplot::corrplot(cor(x))
```




```
fp_cor <- round(cor(comm$freight_value, comm$price),2)
cat("The correlation between the freight value and item's price is",
    fp_cor," ")
```

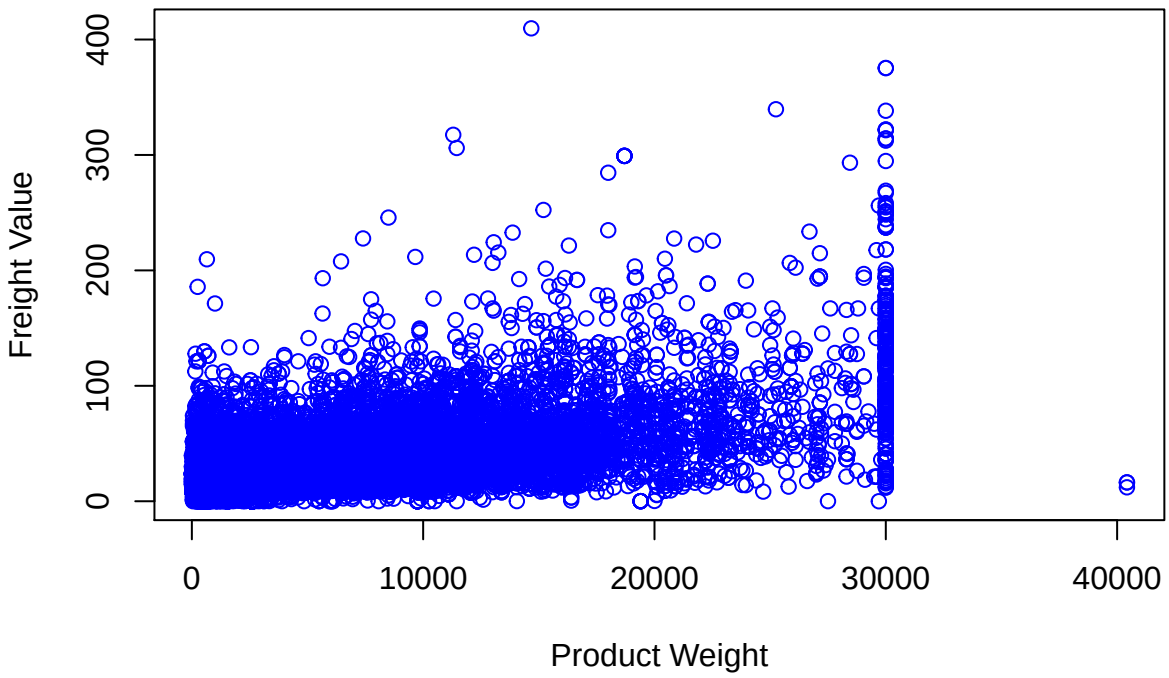
```
The correlation between the freight value and item's price is 0.41
```

```
plot(comm$price,comm$freight_value, xlab = "Item price",
     ylab = "Freight Value",
     main = 'Relationship between the items price and the freight values',
     col = 'red')
```


```
fw_cor <- round(cor(comm$freight_value, comm$product_weight_g),2)
cat("The correlation between the freight value and the products weight is",
    fw_cor)
```

```
The correlation between the freight value and the products weight is 0.61
```

```
plot(comm$product_weight_g,comm$freight_value, xlab = "Product Weight",
     ylab = "Freight Value",
     main = 'Relationship between the freight value and the product weights',
     col = 'blue')
```



From the correlation heatmap and the plots and calculations, we can see that there is some probable amount of correlation between the freight values, prices and item's weights.