Name : Chaitra Naik
Date : 10/29/2021
Title : Module 6 Project

Exploratory Data Analysis on

E-commerce Mobiles Dataset

Chosen from a reputed Indian e-commerce brand “Flipkart”

I have chosen this dataset because I am interested in the mobile phone segment and the launches that happen every year. This dataset shows specifications of several mobile brands in India, found on **Kaggle**.

This dataset has 2647 observations with 8 attributes

Attributes-

- Brand- Name of the Mobile Manufacturer
- Model- Model number of the Mobile Phone
- Color- Color of the model. Missing or Null values indicate no specified color of the model offered on the ecommerce website.
- Memory - RAM of the model (4GB,6GB,8GB, etc.)
- Storage- ROM of the model (32GB,64GB,128GB,256GB, etc.)
- Rating- Rating of the model based on reviews (out of 5). Missing or Null values indicate there are no ratings present for the model.
- Selling Price- Selling Price/Discounted Price of the model in INR when this data was scraped. Ideally price indicates the discounted price of the model
- Original Price- Actual price of the model in INR.

Key Findings:

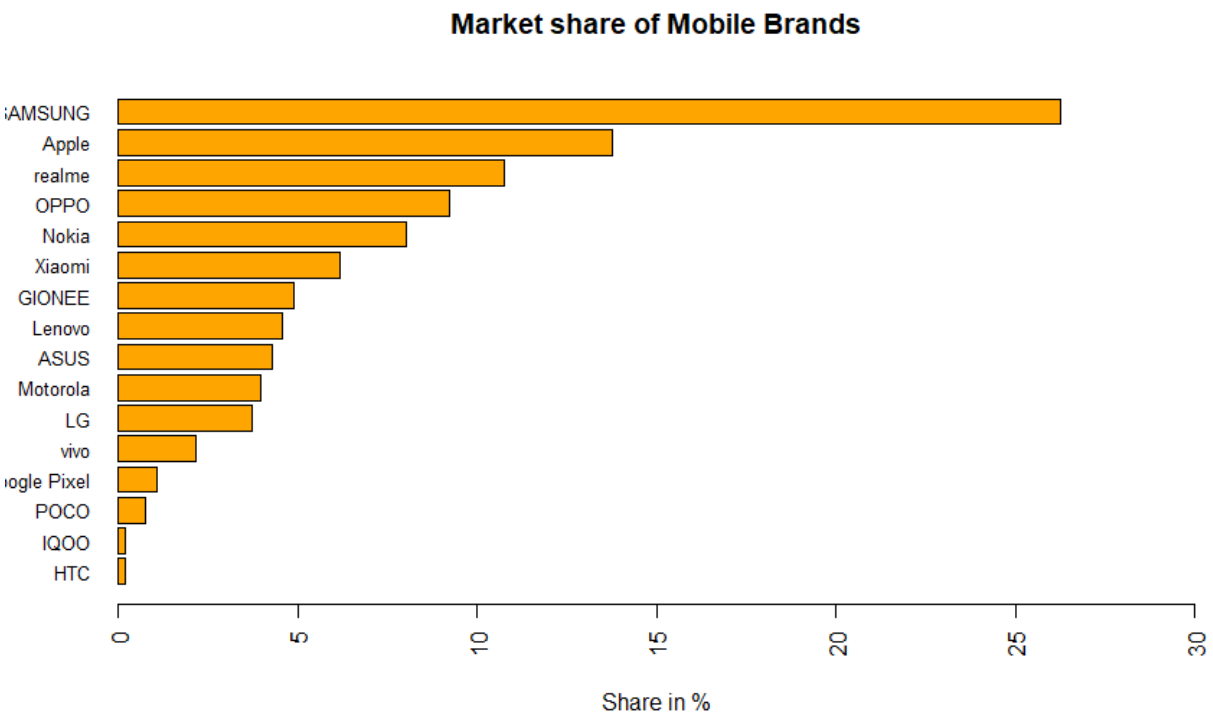
Background:

The study of the mobile_dataset.csv file had lot of missing values in many of the columns such as Model, Color, Memory, Storage, Selling Price and Original Price.

Selling Price and Original Price being quantitative variables, we need them in our further analysis for descriptive statistics. Hence the First step was to **clean the dataset** inorder to get rid of the missing values.

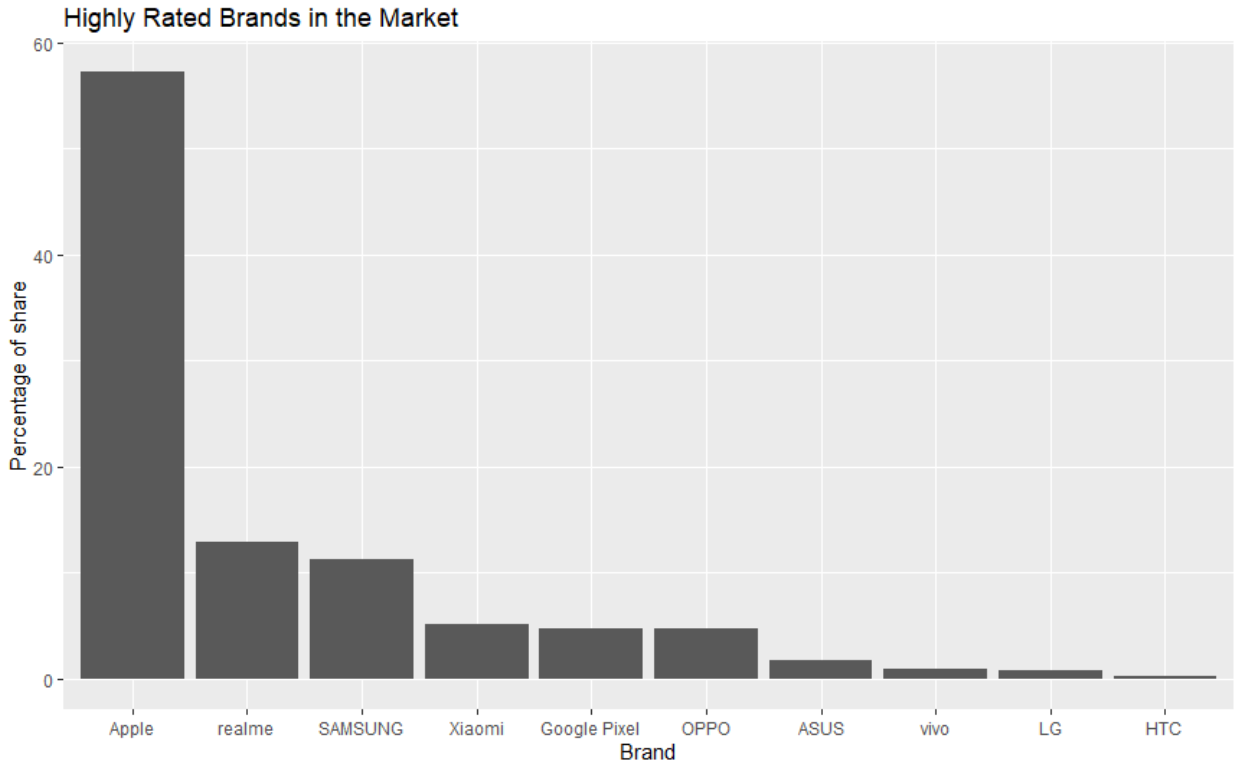
One way was to omit the values. However, Original Price had about **1678** rows of missing values out of the total **2647** records and omitting all of them would not be a feasible task. Hence the challenge was to fill the missing values of the models of each brand according to the specifications. This was done by grouping the dataset by Brand, Model, Memory and Storage columns. Next was to find the mean of the Selling Price of the existing values in the Original Price column. This would give the mean price of the models of the same brand matched with the same specifications. Thus, the missing values would be filled by the computed mean values. Hence the missing values were now removed and our final dataset is ready for processing.

Next step was to find the count of each brand and the percentage share of each brand compared to the total. The data was sorted in the increasing order and the resulting looked like this:



From this visualization, we can infer that the brand SAMSUNG has the highest market share of around 26% followed by Apple and realme. HTC has the least share in the Indian Market.

Next job is to find out the popular brands in the market. Hence filtering the dataset by the rating above or equal to 4.5 (Highly rated brands). The resulting graph looks like this.

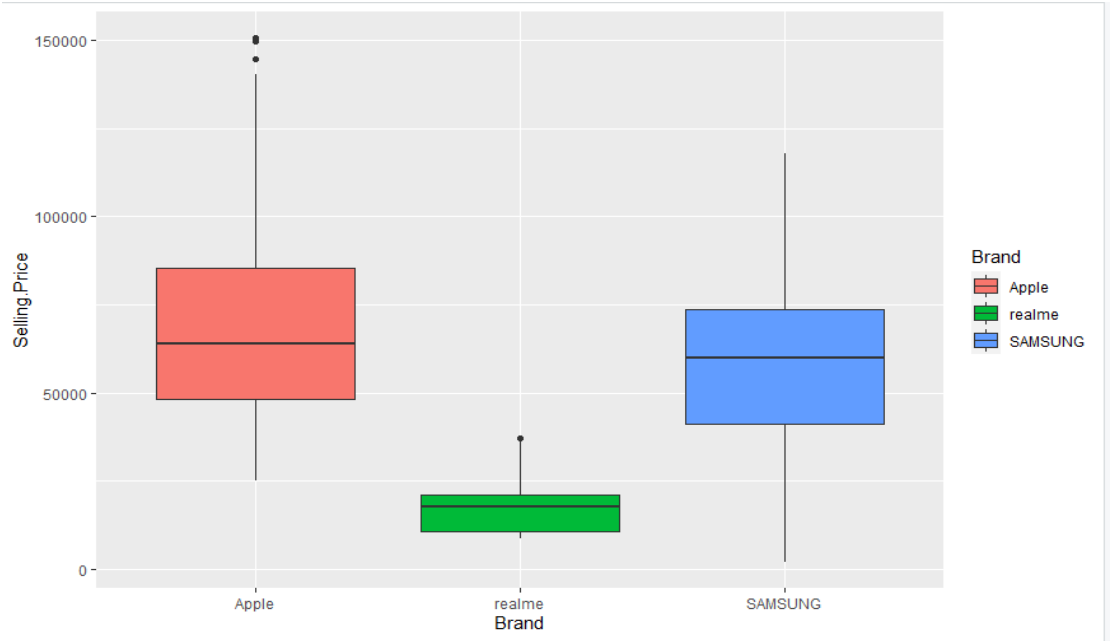


From this graph, we can infer that Apple is the most highly rated brand. Followed by this are realme and Samsung brands.

I have then filtered the dataset by the 3 popular brands – Apple , realme and SAMSUNG . I have calculated the min and max SP of each brand.,The mean and median of each brand.

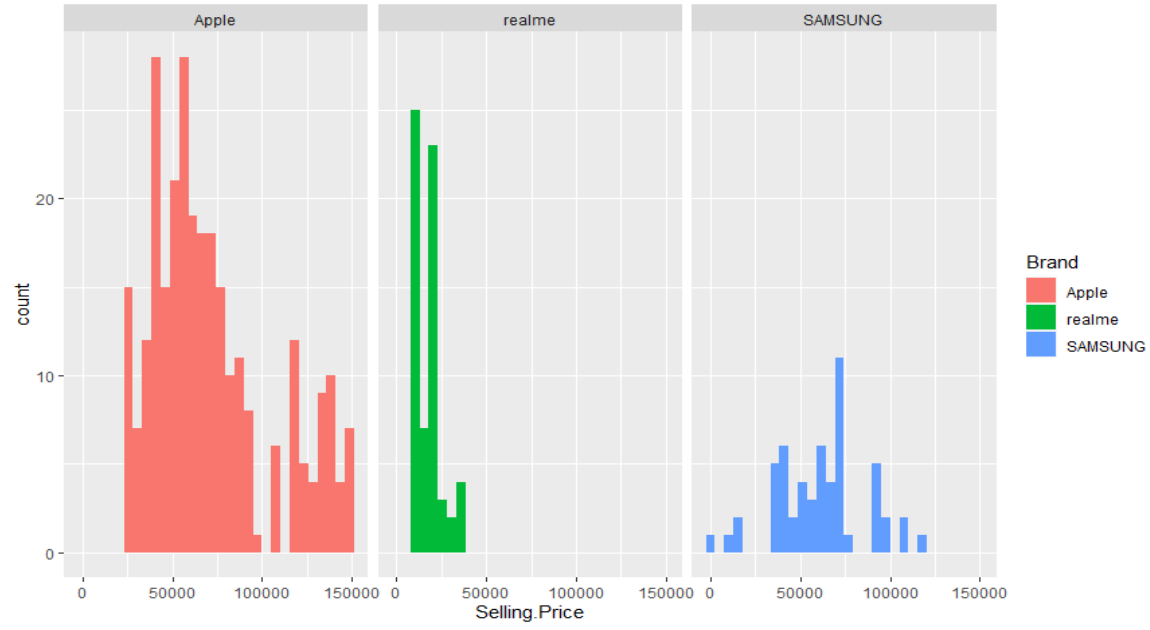
	Brand	min_Selling_Price	max_Selling_Price	Mean	Median
1	Apple	24999	150800	72093.54	63999
2	realme	8499	36999	17217.19	17499
3	SAMSUNG	1949	117990	61095.27	59999

Are the Popular brands expensive? Relation between Popular Brands Vs Selling Price is shown by a boxplot below:



Boxplot of Popular Brands Vs Selling Price shows that Brands like Apple and Samsung have premium phones at higher Selling Prices. As seen in the plot, the median Price of both brands is above 50000 INR. On the other hand, brands like realme offer much reasonable price phones catered to Indian market.

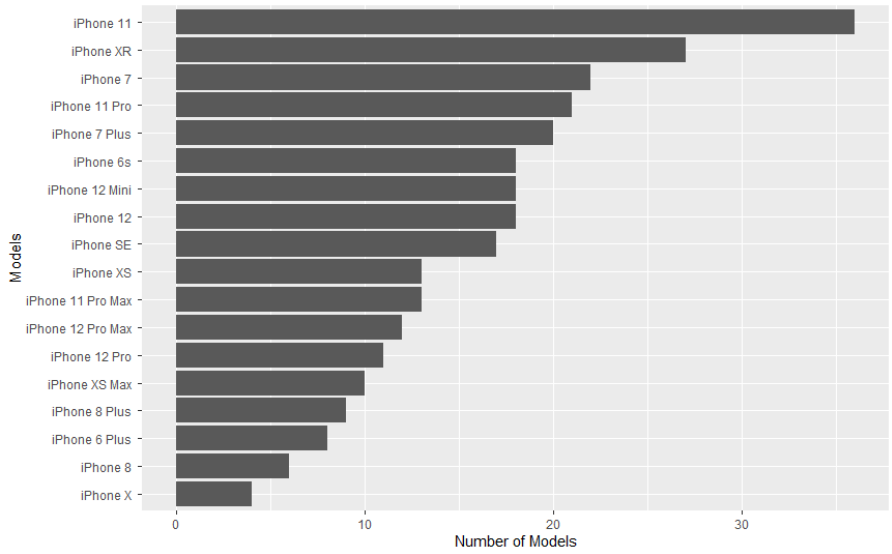
This can also be depicted by a histogram.



Plot shows the comparison of prices of the three top rated brands. Apple and Samsung both offer premium expensive models whereas realme offers affordable phones within 50000 INR. We can also conclude that Apple offers phones in a varied price range, realme mostly in the lower price range and Samsung has phones in only certain buckets of price ranges (higher and lower).

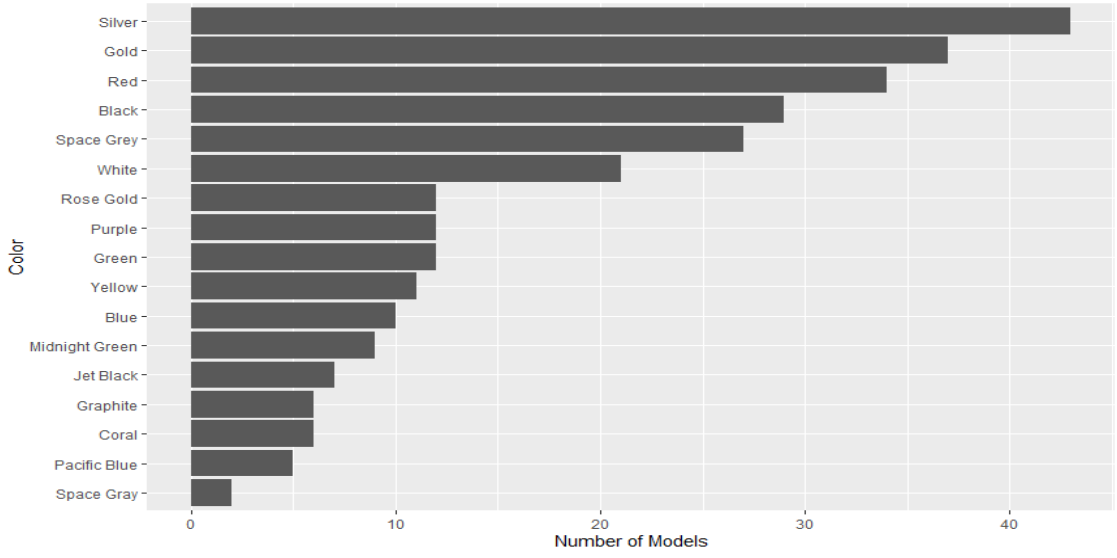
Most Preferred Model of Each of the Top Rated Brands

- **Iphone**



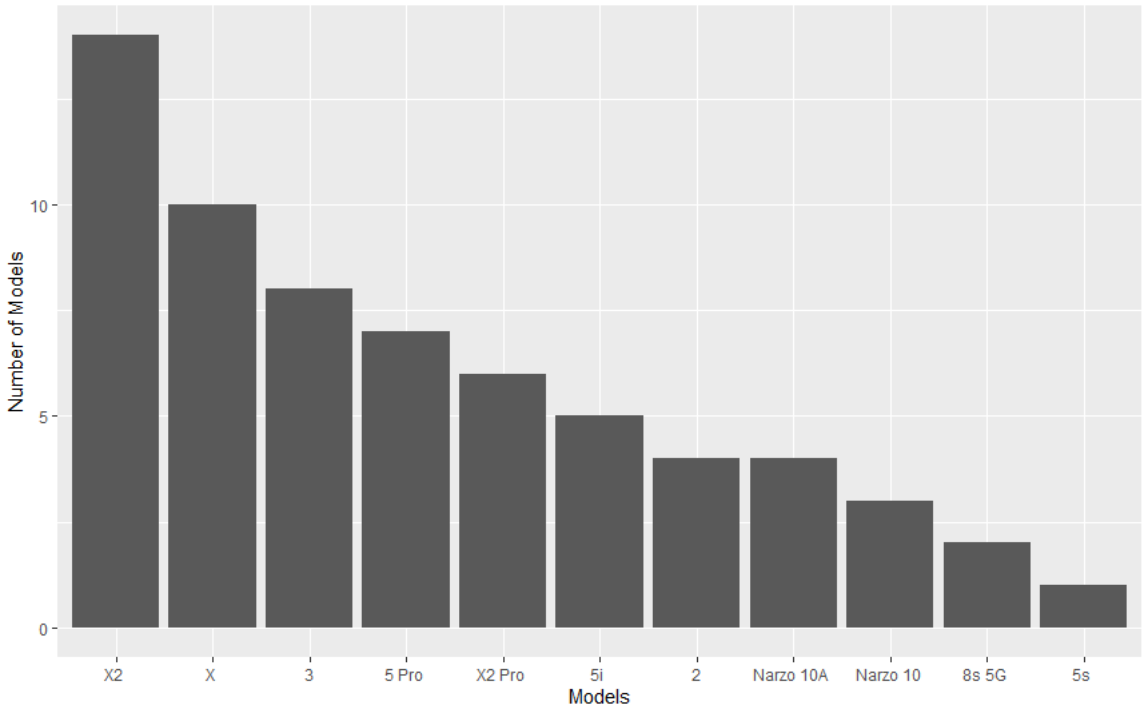
The plot shows that Iphone11 is the most preferred model among the Iphone brand, followed by Iphone XR and iPhone 7.Though Iphone 12 and its versions were more recently launched, Iphone 11 still continues to be preferred model.

Most Preferred Color in the Iphone



More than 40% users prefer **Silver** color in the iPhone. Followed by Gold and then Red.

- **Realme**



Plot shows X2 is the most preferred model in the realme brand, followed by **X** and **3** models.

- **Most Preferred Color in Realme**

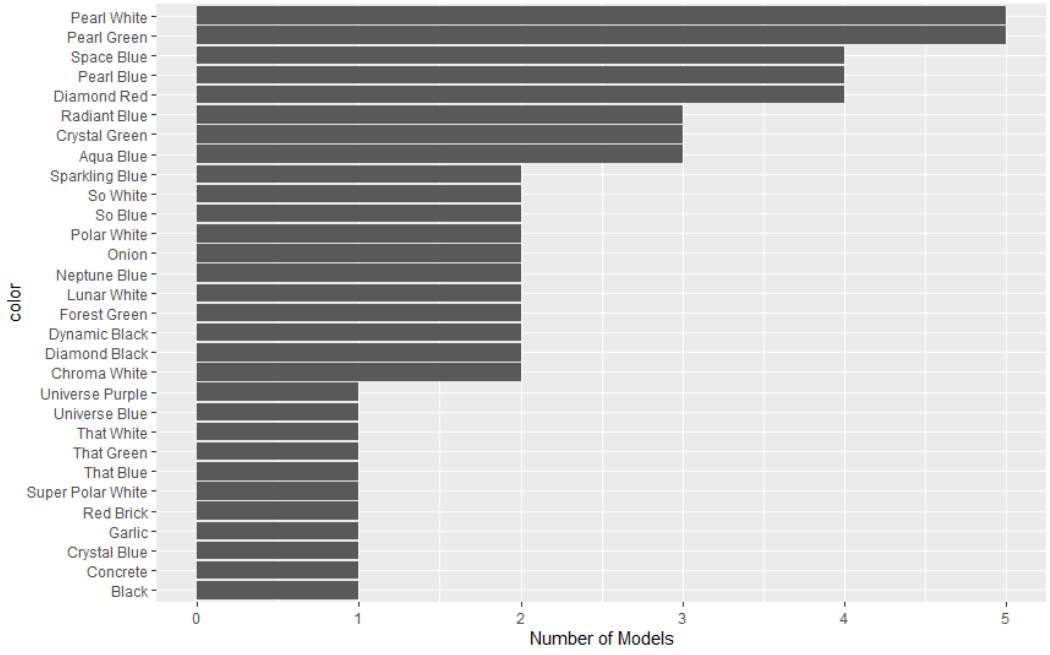
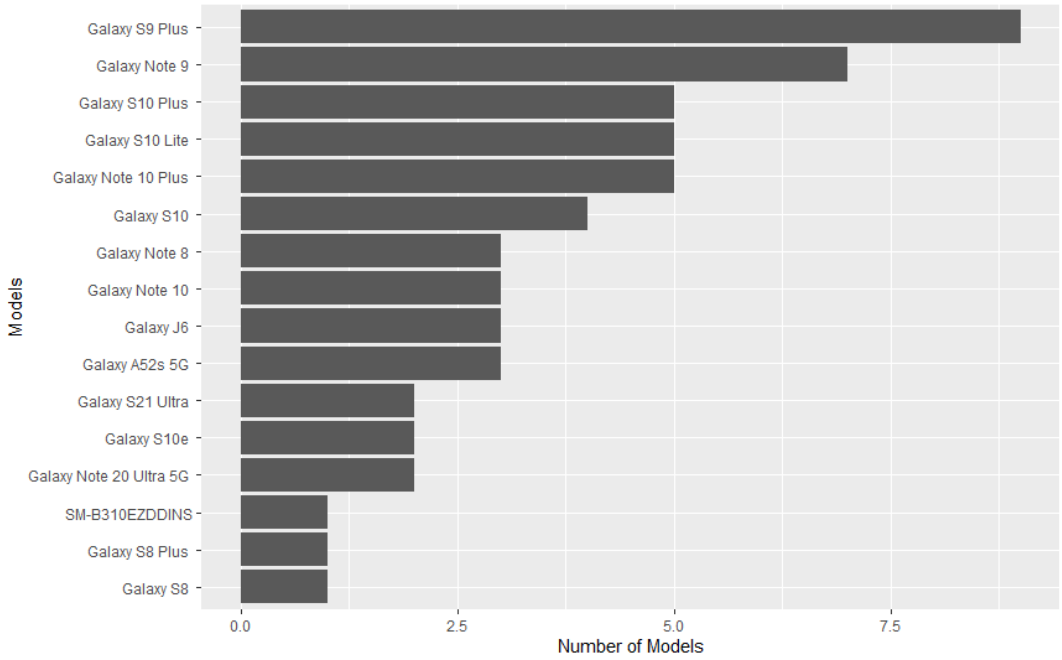


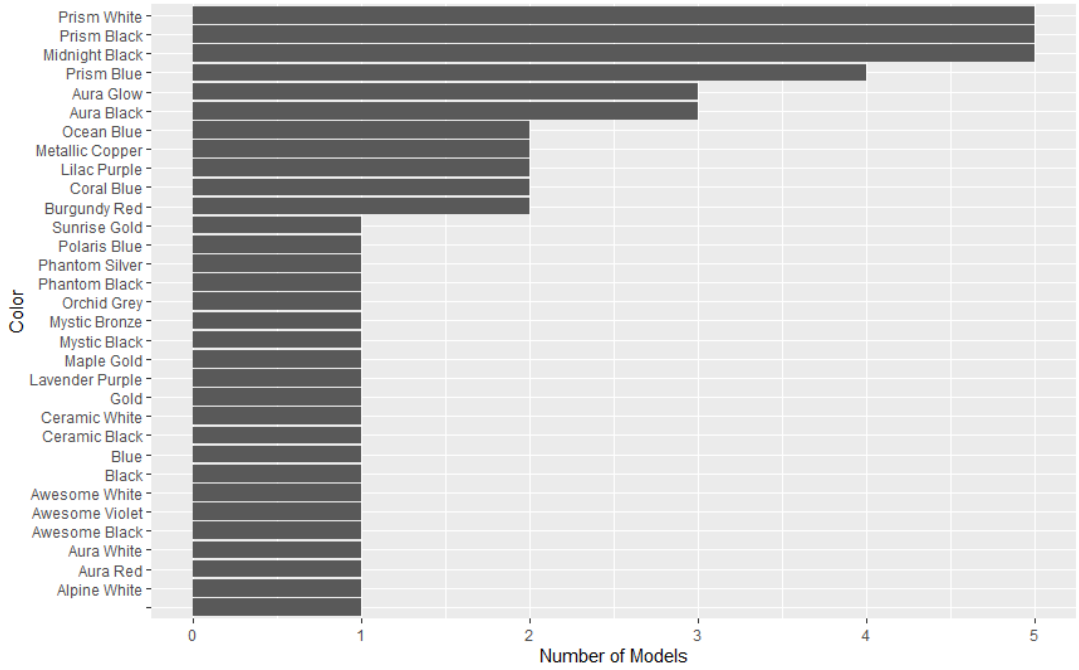
Figure shows that Pearl White and Pearl Green are the most preferred colors for realme

- **Samsung**



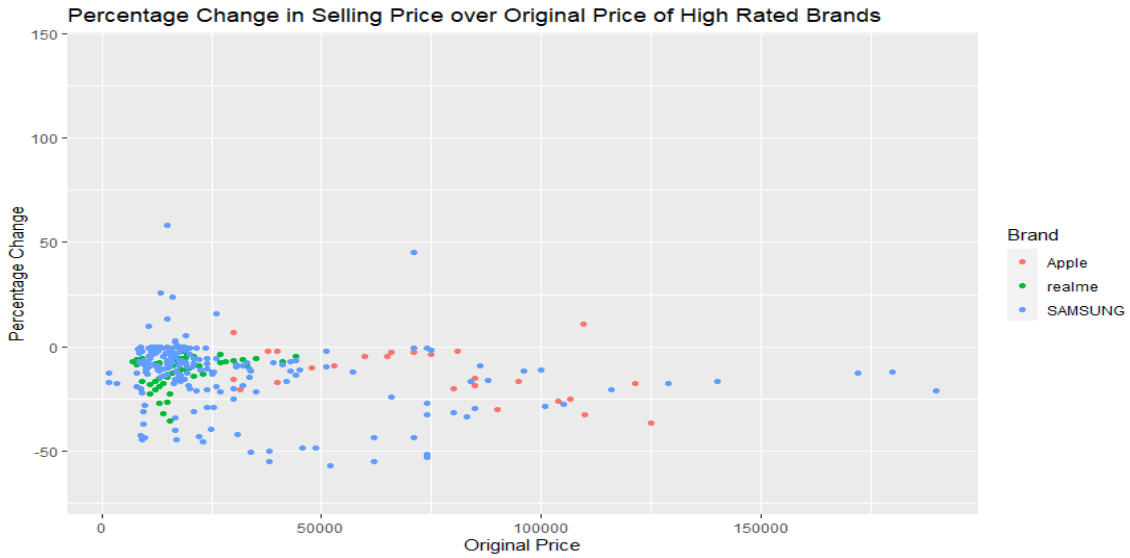
Plot shows Galaxy S9 Plus is the most preferred model in the Samsung series, followed by is Galaxy Note 9 .

- **Most Preferred Color in Samsung**



We can infer from this plot that Prism White, Prism Black and Midnight Black are the most preferred colors for Samsung brand.

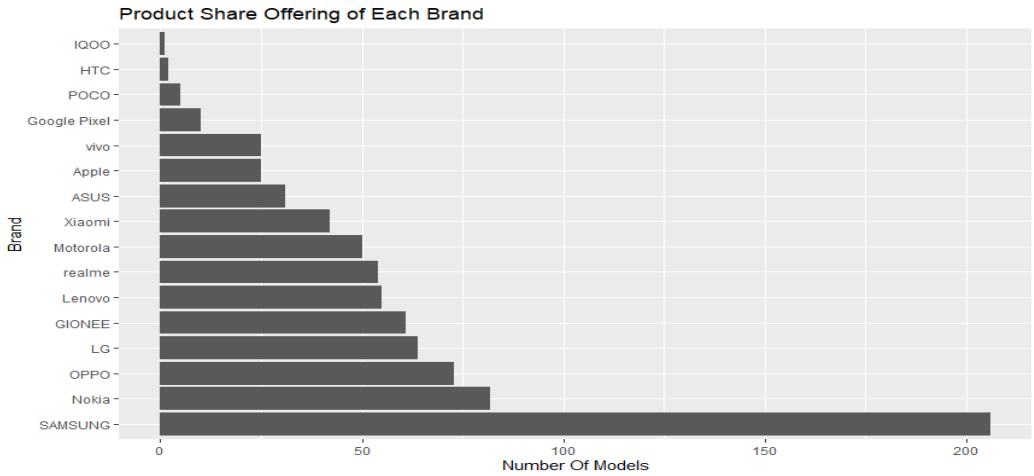
Next, we want to calculate the percentage change in the Selling Price Over Original Price of each of the top three brands.



Scatter Plot shows the percentage change of Selling Price over Original Price for each of the brands: Apple, realme and Samsung. Negative percentage denotes that discount are being offered on these brands. Samsung is being given more discount, going upto 50% , for the other two brands, max discount is around 25%. Discounts for Samsung and Apple after certain price range (75k) is almost same (under 25%).

Since there is heavy discount on these phones, we can make assumptions that this dataset is being collated around some festive season or Flipkart days, as discounts are usually higher during this time. Another thing to note here is that how are the companies making profits with such heavy discounts, more analysis on their different product line is required to answer this question.

Brand with the most product offerings for the Market



Barplot shows the brands and their respective model offerings to the market.We can infer that Samsung offers most number of models (206 models). Followed by Nokia(82) and OPPO(73) brands.

Now, that we have the Brands with the most product offerings,let us find the min and the max Selling Price .

	Brand	min_Selling_Price	max_Selling_Price
1	Nokia	1000	56299
2	OPPO	4999	60990
3	SAMSUNG	1099	169999

BIBLIOGRAPHY :

- *count number of rows in a data frame in R based on group*. (n.d.). Stack Overflow. https://stackoverflow.com/questions/25293045/count-number-of-rows-in-a-data-frame-in-r-based-on-group
- *RPubs - Reordering a ggplot bar chart axis*. (n.d.). RPubs - Reordering a Ggplot Bar Chart Axis. https://www.rpubs.com/dvdunne/reorder_ggplot_barchart_axis#:%7E:text=Sometimes%20when%20creating%20a%20bar,the%20bars%20in%20alphabetical%20order.&text=A%20simple%20way%2C%20to%20reorder,to%20use%20the%20reorder%20command.
- *How to count how many values per level in a given factor?* (2014, September 30). Stack Overflow https://stackoverflow.com/questions/26114525/how-to-count-how-many-values-per-level-in-a-given-factor
- Z. (2021, May 26). *How to Calculate the Mean by Group in R (With Examples)*. Statology. https://www.statology.org/r-mean-by-group/
- S. (n.d.). *R Basics*. Summarise. http://statseducation.com/Introduction-to-R/modules/tidy%20data/summarise/

APPENDIX:

R- Script:

```
# Import libraries
library(FSA)        #Import FSA    library
library(FSAdata)
library(magrittr)   #Import magrittr library
library(plyr)       #Import plyr   library
library(dplyr)      #Import dplyr  library
library(tidyr)      #Import tidyr  library
```

```
library(tidyverse)  #Import tidyverse  library
library(plotrix)    #Import plotrix  library



# Set the current working directory
getwd()
setwd("C:/Users/chait/Desktop/CPS-NEU/Assignments/Module6")

# read the dataset
ph <- read.csv("mobile_dataset.csv",header = TRUE ,sep = ",")
class(ph)   #dataframe

#CLEANING THE DATASET
ph1 <- ph %>% na.omit()  # omit the entries with missing values
view(ph1)

ph2<-ph1 %>%
 group_by(Brand,Model,Memory,Storage) %>%
  summarise_at(vars(Original.Price), list(name = mean))  # find the avg of the original price column
colnames(ph2)[colnames(ph2) == "name"] <- "Original_Price_Mean"

ph3 <- merge(x = ph, y = ph2, by = c("Brand","Model","Memory","Storage"), all.x = TRUE)

ph4 <- ph3 %>%
  mutate(Original.Price = coalesce(Original.Price,Original_Price_Mean))#fill the missing values of the org
price with the mean of the org price

df = subset(ph4, select = -c(Original_Price_Mean) )   #Final cleaned dataset
```

```
summary(df)                                  #Summary of the cleaned dataset

# Find the count of each brand
ph_count <- table(df$Brand)
ph_count

ph_count <- 100*prop.table(ph_count)

# Sort the data by the increasing order of the percent count
ph_count <- ph_count[order(ph_count)]

# Bar plot of different brands
barplot(ph_count,names.arg = paste(names(ph_count)),horiz = TRUE,cex.names = 0.8,las=2,col =
"orange",xlim = c(0,30),
        main = "Market share of Mobile Brands",xlab="Share in %")



#Filtering out the popular brands based on ratings
pop_brands <- filter(df, Rating >= 4.5)



pop_table <- group_by(pop_brands,Brand)
pop_table <- summarise(pop_table, count=n())
per <-  round(100*prop.table(pop_table$count),1)
pop_br_tab <- pop_table %>% add_column(per)

#bar chart of the highest rated brands in market

ggplot(data=pop_br_tab,aes(x=reorder(Brand,-per),y=per)) + geom_bar(stat="identity")+
```

```
  labs(title = "Highly Rated Brands in the Market",
     x = "Brand", y = "Percentage of share")

#filter the data by the popular 3 brands
pop_apple <-pop_brands %>% filter(Brand == "Apple" | Brand == "realme" | Brand =="SAMSUNG")

pop_min_SP <- pop_apple %>%
  group_by(Brand) %>%
  summarise(min_Selling_Price = min(Selling.Price,na.rm = TRUE))



pop_max_SP <- pop_apple %>%
  group_by(Brand) %>%
  summarise(max_Selling_Price = max(Selling.Price,na.rm = TRUE))



# finding mean and median of popular 3 brands

#Mean
mean_pop <- pop_apple %>%
            group_by(Brand) %>%
            summarise_at(vars(Selling.Price), list(name = mean))
colnames(mean_pop)[colnames(mean_pop) == "name"] <- "Mean"



#Median
median_pop <- pop_apple %>%
  group_by(Brand) %>%
  summarise_at(vars(Selling.Price), list(name = median))
```

```
colnames(median_pop)[colnames(median_pop) == "name"] <- "Median"

#Summarise the min,max,mean and median values into single dataframe df_value
df_values <- data.frame(pop_min_SP,pop_max_SP,mean_pop,median_pop)

df_values <- merge(pop_min_SP, pop_max_SP,mean_pop,median_pop, by = 'Brand')
df_values <- subset (df_values, select = -c(Brand.1,Brand.2,Brand.3))
view(df_values)



# Are the popular phones expensive? what is the relation between popular brands Vs prices
ggplot(pop_apple,aes(x = Brand , y= Selling.Price,fill = Brand )) + geom_boxplot()

#histogram
pop_brands %>%
  filter(Brand == "Apple" | Brand == "realme" | Brand =="SAMSUNG") %>%
  ggplot(aes(x = Selling.Price,fill = Brand)) + geom_histogram(na.rm = TRUE)    + facet_wrap(~Brand)



#check the most preferred model of each brand

# Apple
pop_apple1 <- pop_brands %>% filter(Brand == "Apple")
ggplot(data=pop_apple1,aes(x=Model)) + geom_bar()  + coord_flip()

#color most preferred
ggplot(data=pop_apple1,aes(x=Color)) + geom_bar() + coord_flip()
```

```
# Realme
pop_real <- pop_brands %>% filter(Brand == "realme")
ggplot(data=pop_real,aes(x = Model)) + geom_bar()

#color most preferred
ggplot(data=pop_real,aes(x=Color)) + geom_bar() +coord_flip()

#Samsung
pop_sam <- pop_brands %>% filter(Brand == "SAMSUNG")
ggplot(data=pop_sam,aes(x=Model)) + geom_bar() + coord_flip()

# most preferred color
ggplot(data=pop_sam,aes(x=Color)) + geom_bar() + coord_flip()

#Brand with the most product offerings to the market
No_of_models <- df %>%
  group_by(Brand) %>%
  summarise(sum_of_models = length(unique(Model)))
No_of_models

# Bar plot of the brand with the most product offerings
ggplot(data=No_of_models,aes(x=reorder(Brand,-sum_of_models),y=sum_of_models)) +
geom_bar(stat="identity") +
  labs(title = "Product Share Offering of Each Brand",
       x = "Brand", y = "Number Of Models") + coord_flip()

#Now that we have found the brands with most product offerings,find out the min cost of each brand

min_SP <- df %>% filter(Brand== "SAMSUNG" | Brand == "Nokia" | Brand =="OPPO") %>%
```

```
  group_by(Brand) %>%
  summarise(min_Selling_Price = min(Selling.Price,na.rm = TRUE))
min_SP

#Find the max cost of each brand
max_SP <- df %>% filter(Brand== "SAMSUNG" | Brand == "Nokia" | Brand =="OPPO") %>%
  group_by(Brand) %>%
  summarise(max_Selling_Price = max(Selling.Price,na.rm = TRUE))
max_SP

#merge the values into a dataframe
df_model <- data.frame(min_SP,max_SP)

df_model <- merge( min_SP,max_SP,by = 'Brand')
view(df_model)

##########################################
# calculate the percent change in the price
df1 <- df%>% group_by(Brand,Model,Memory,Storage) %>%
       mutate(pct_change = ((Selling.Price - Original.Price)/Original.Price ) * 100)
view(df1)
# plot the graph of Profit Vs Original Price
df1 %>%
  filter(Brand == "Apple" | Brand == "realme" | Brand =="SAMSUNG") %>%
  ggplot(aes(x = Original.Price,y = pct_change,col=Brand)) + geom_point(na.rm = TRUE) + ylim(-70,140)+
  labs(title = "Percentage Change in Selling Price over Original Price of High Rated Brands",
       x = "Original Price", y = "Percentage Change")
```