Homework 3

Lilit Ivanyan

2025-03-16

```r
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
df <- read.csv("mobiles_dataset.csv")
```

Part 1

Hint: Convert all the currencies to dollars using the following logic: 1 PKR = 0.0036 USD 1 INR = 0.011 USD 1 CNY = 0.14 USD 1 AED = 0.27 USD

```r
df <- df %>%
  rename(
    Pakistan_Price = `Launched.Price.Pakistan.PKR`,
    India_Price = `Launched.Price.India.INR`,
    China_Price = `Launched.Price.China.CNY`,
    USA_Price = `Launched.Price.USA.USD`,
    Dubai_Price = `Launched.Price.Dubai.AED`
  )

currency_rates <- c(Pakistan = 0.0036, India = 0.011, China = 0.14, USA = 1.0, Dubai = 0.27)

df <- df %>%
  mutate(
    Pakistan_Price_USD = Pakistan_Price * currency_rates["Pakistan"],
    India_Price_USD = India_Price * currency_rates["India"],
    China_Price_USD = China_Price * currency_rates["China"],
    USA_Price_USD = USA_Price * currency_rates["USA"],
    Dubai_Price_USD = Dubai_Price * currency_rates["Dubai"]
  )
```

1. Does battery capacity influence the launched price of a smartphone? Check this variability across all currencies. Is there any type of difference between behaviors?

```
correlations <- df %>%
  summarize(
    Correlation_Pakistan = cor(Battery.Capacity.mAh, Pakistan_Price_USD),
    Correlation_India = cor(Battery.Capacity.mAh, India_Price_USD),
    Correlation_China = cor(Battery.Capacity.mAh, China_Price_USD),
    Correlation_USA = cor(Battery.Capacity.mAh, USA_Price_USD),
    Correlation_Dubai = cor(Battery.Capacity.mAh, Dubai_Price_USD)
  )

print(correlations)
```

```
##   Correlation_Pakistan Correlation_India Correlation_China Correlation_USA
## 1          -0.06091272       -0.01905429       -0.04104334      -0.0411368
##   Correlation_Dubai
## 1       -0.04890776
```

As the correlation is between the battery capacity and different currencies is small negative it means that the battery capacity does not influence the launched capacity significantly. For different currencies, there is minor differences between correlation, for example it is a bit smaller for Pakistan currency compared to other currencies and for the case of Indian currency it is the highest among all, but the difference is not that significant.

2. Does RAM size impact the price of smartphones? Check this variability across all currencies. Is there any type of difference between behaviors?

```
df$RAM_numeric <- as.numeric(gsub("GB", "", df$RAM))
```

```
## Warning: NAs introduced by coercion
```

```
correlations <- df %>%
  summarize(
    Correlation_Pakistan = cor(RAM_numeric, Pakistan_Price_USD, use = "complete.obs"),
    Correlation_India = cor(RAM_numeric, India_Price_USD, use = "complete.obs"),
    Correlation_China = cor(RAM_numeric, China_Price_USD, use = "complete.obs"),
    Correlation_USA = cor(RAM_numeric, USA_Price_USD, use = "complete.obs"),
    Correlation_Dubai = cor(RAM_numeric, Dubai_Price_USD, use = "complete.obs")
  )

print(correlations)
```

```
##   Correlation_Pakistan Correlation_India Correlation_China Correlation_USA
## 1             0.409115         0.4169188          0.422733       0.4628016
##   Correlation_Dubai
## 1         0.4739097
```

The correlation between the RAM and the price of the smartphones is high positive meaning that the size of the RAM impacts the price significantly. The higher the RAM, higher is the price of the smartphone. There is small difference between different currencies. For example, the highest correlation is with AED, meaning that the influence is higher with that currency when with others, but in general, the correlation is quite close between all of them.

3. Do Apple devices have a higher price variation across different regions compared to other brands? In which country do Apple devices have the highest markup? Are there brands with more stable pricing across regions?

```
df <- df %>%
  mutate(Average_Price_USD = rowMeans(select(., Pakistan_Price_USD, India_Price_USD, China_Price_USD, US

brand_price_variation <- df %>%
  group_by(Company.Name) %>%
  summarize(
    Price_Range = max(Average_Price_USD, na.rm = TRUE) - min(Average_Price_USD, na.rm = TRUE),
    Price_SD = sd(Average_Price_USD, na.rm = TRUE)
  ) %>%
  arrange(desc(Price_SD))

print(brand_price_variation)
```

```
## # A tibble: 19 x 3
##    Company.Name Price_Range Price_SD
##    <chr>              <dbl>    <dbl>
##  1 Huawei            2091.     590.
##  2 Samsung           1887.     505.
##  3 Sony              1028      391.
##  4 Honor             1578.     387.
##  5 Tecno             1680.     362.
##  6 Google            1449.     347.
##  7 Xiaomi             837.     267.
##  8 Oppo              1286.     265.
##  9 Motorola          1201.     261.
## 10 Apple             1434.     256.
## 11 Vivo              1143      255.
## 12 OnePlus            874.     225.
## 13 Lenovo             505.     171.
## 14 POCO               362.     104.
## 15 Realme             425       97.9
## 16 Infinix            405.      94.3
## 17 iQOO               116       58.1
## 18 Nokia              141.      44.8
## 19 Poco                21.4     15.1
```

To check the price variation between different brands, we need to calculate the standard deviation of different brands. Higher the SD, higher price variations across different regions. From the table we can see that the highest price variation has Huawei smartphones and compared to all brands, Apple smartphones are on the 10th place with lower price variation.

```
apple_prices <- df %>%
  filter(Company.Name == "Apple") %>%
  summarize(
    Avg_Global_Price = mean(Average_Price_USD, na.rm = TRUE),
    Pakistan_Markup = mean(Pakistan_Price_USD, na.rm = TRUE) / Avg_Global_Price,
    India_Markup = mean(India_Price_USD, na.rm = TRUE) / Avg_Global_Price,
    China_Markup = mean(China_Price_USD, na.rm = TRUE) / Avg_Global_Price,
    USA_Markup = mean(USA_Price_USD, na.rm = TRUE) / Avg_Global_Price,
```

```r
    Dubai_Markup = mean(Dubai_Price_USD, na.rm = TRUE) / Avg_Global_Price
  )

max_markup_country <- names(which.max(apple_prices[2:6]))

print(paste("Apple devices have the highest markup in:", max_markup_country))
```

```
## [1] "Apple devices have the highest markup in: India_Markup"
```

The stabilility of the brand depends on the price variation. Lower the price variation, more stable the brand is. The brands which are the most stable are Poco, Nokia and iQOO.

4. Do all smartphone brands have flagship and budget-friendly models, or do some brands only focus on premium devices? Hint: Categorize brands into budget, mid-range, and premium segments (Budget: < \$300, Mid-range: \$300 - \$700, Premium: > \$700). Check how many models each brand has in each segment. Determine whether a brand covers all three segments or focuses only on premium/mid-range.

```r
df <- df %>%
  mutate(
    Price_Category = case_when(
      USA_Price_USD< 300 ~ "Budget",
      USA_Price_USD >= 300 & USA_Price_USD <= 700 ~ "Mid-range",
      USA_Price_USD > 700 ~ "Premium"
    )
  )

brand_segments <- df %>%
  group_by(Company.Name, Price_Category) %>%
  summarize(Count = n())
```

```
## `summarise()` has grouped output by 'Company.Name'. You can override using the
## `.groups` argument.
```

```r
print(head(brand_segments))
```

```
## # A tibble: 6 x 3
## # Groups:   Company.Name [3]
##   Company.Name Price_Category Count
##   <chr>        <chr>          <int>
## 1 Apple        Mid-range          8
## 2 Apple        Premium           89
## 3 Google       Mid-range         12
## 4 Google       Premium            9
## 5 Honor        Budget            29
## 6 Honor        Mid-range         37
```

From all the models, Apple is the only one that has almost only premium mdoels. Other brands have mid-range and budget-friendly brands as well.

5. Which region offers the most affordable smartphone prices on average? Are there any brands that price their phones significantly lower in one region compared to others?

```r
avg_prices_region <- df %>%
  summarize(
    Avg_Price_Pakistan = mean(Pakistan_Price_USD, na.rm = TRUE),
    Avg_Price_India = mean(India_Price_USD, na.rm = TRUE),
    Avg_Price_China = mean(China_Price_USD, na.rm = TRUE),
    Avg_Price_USA = mean(USA_Price_USD, na.rm = TRUE),
    Avg_Price_Dubai = mean(Dubai_Price_USD, na.rm = TRUE)
  )

cheapest_region <- names(which.min(avg_prices_region))

print(avg_prices_region)
```

```
##   Avg_Price_Pakistan Avg_Price_India Avg_Price_China Avg_Price_USA
## 1           449.9342        552.8237        530.7414      579.6238
##   Avg_Price_Dubai
## 1         586.029
```

From the results, we can see that Pakistan is the region that offers the most affordable smartphone prices on average.

```r
brand_price_comparison <- df %>%
  group_by(Company.Name) %>%
  summarize(
    Avg_Pakistan = mean(Pakistan_Price_USD, na.rm = TRUE),
    Avg_India = mean(India_Price_USD, na.rm = TRUE),
    Avg_China = mean(China_Price_USD, na.rm = TRUE),
    Avg_USA = mean(USA_Price_USD, na.rm = TRUE),
    Avg_Dubai = mean(Dubai_Price_USD, na.rm = TRUE)
  )

brand_discount_region <- brand_price_comparison %>%
  mutate(Min_Region = apply(select(., -Company.Name), 1, function(x) names(x)[which.min(x)]))

print(brand_discount_region)
```

```
## # A tibble: 19 x 7
##    Company.Name Avg_Pakistan Avg_India Avg_China Avg_USA Avg_Dubai Min_Region
##    <chr>               <dbl>     <dbl>     <dbl>   <dbl>     <dbl> <chr>
##  1 Apple                891.     1133.     1005.   1028.      995. Avg_Pakistan
##  2 Google               621.      774.      849.    755.      816. Avg_Pakistan
##  3 Honor                433.      537.      472.    608.      606. Avg_Pakistan
##  4 Huawei               661.     1131.      961.   1117.     1126. Avg_Pakistan
##  5 Infinix              158.      191.      219.    245.      246. Avg_Pakistan
##  6 Lenovo               227.      279.      295.    312.      317. Avg_Pakistan
##  7 Motorola             330.      371.      378.    433.      433. Avg_Pakistan
##  8 Nokia                186.      147.      161.    174       173. Avg_India
##  9 OnePlus              485.      503.      553.    609.      663. Avg_Pakistan
## 10 Oppo                 342.      481.      477.    505.      536. Avg_Pakistan
## 11 POCO                 210.      249.      286.    310.      323. Avg_Pakistan
## 12 Poco                 247.      275.      308.    290       287. Avg_Pakistan
## 13 Realme               249.      288.      275.    273.      269. Avg_Pakistan
```

```
## 14 Samsung                    746.      663.      683.    713.      741. Avg_India
## 15 Sony                      1180.     1008.      835.   1132.     1089. Avg_China
## 16 Tecno                      299.      405.      428.    472.      489. Avg_Pakistan
## 17 Vivo                       261.      393.      413.    469.      431. Avg_Pakistan
## 18 Xiaomi                     485.      630.      484.    560.      611. Avg_China
## 19 iQOO                       288.      484.      471.    399       405. Avg_Pakistan
```

As we can see from the results, for almost all the brands the region that has lower price compared to others is the Pakistan region. For almost all the brands the difference is quite significant.

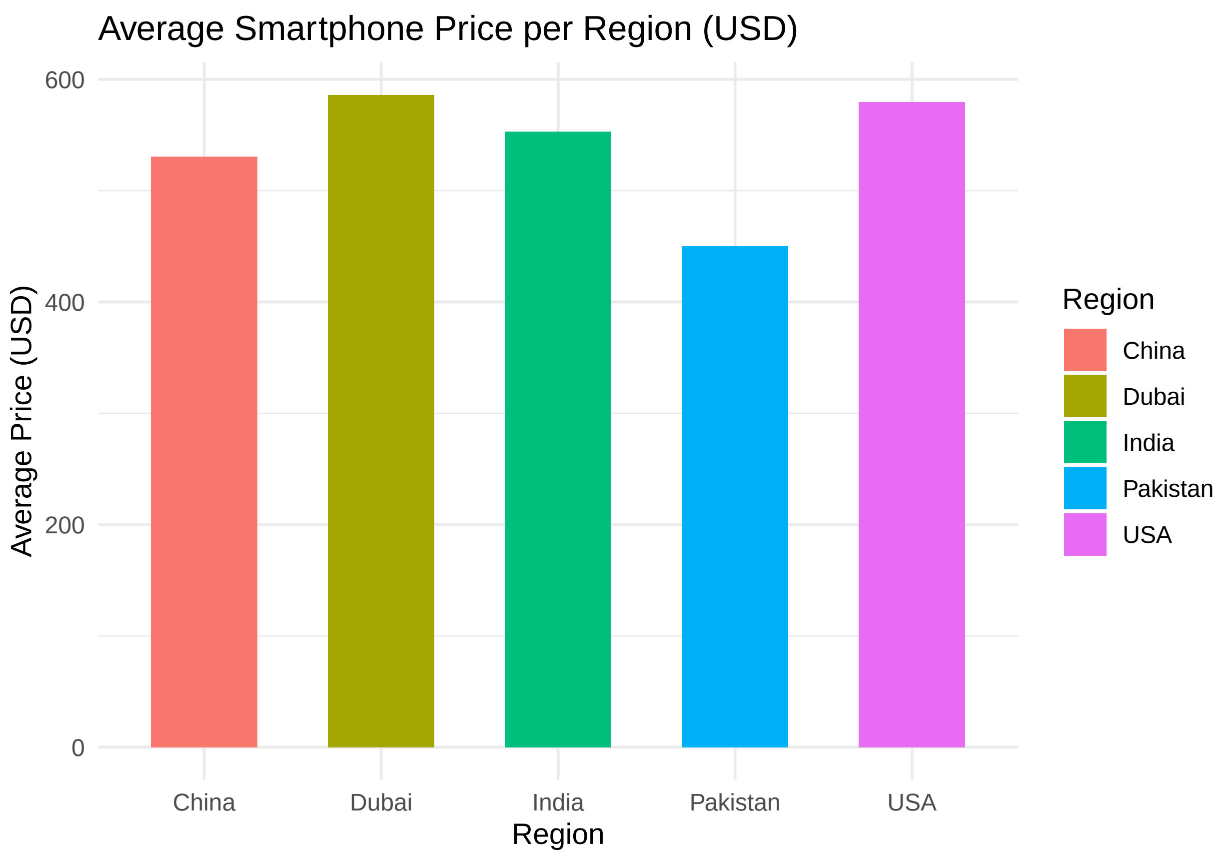
Part 2

1. Plot a bar chart for average price per region in USD.

```r
avg_prices_region <- df %>%
  summarize(
    Region = c("Pakistan", "India", "China", "USA", "Dubai"),
    Avg_Price = c(
      mean(Pakistan_Price_USD, na.rm = TRUE),
      mean(India_Price_USD, na.rm = TRUE),
      mean(China_Price_USD, na.rm = TRUE),
      mean(USA_Price_USD, na.rm = TRUE),
      mean(Dubai_Price_USD, na.rm = TRUE)
    )
  )
```

```
## Warning: Returning more (or less) than 1 row per `summarise()` group was deprecated in
## dplyr 1.1.0.
## i Please use `reframe()` instead.
## i When switching from `summarise()` to `reframe()`, remember that `reframe()`
##   always returns an ungrouped data frame and adjust accordingly.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

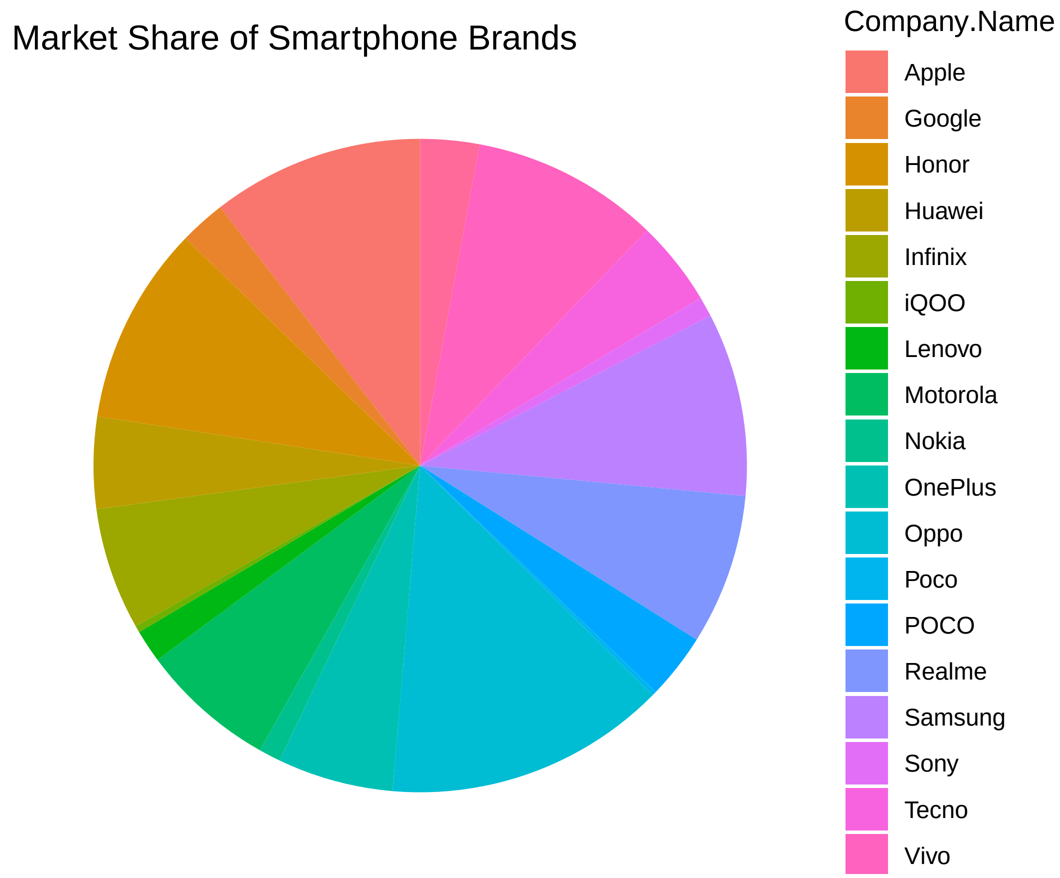
```r
ggplot(avg_prices_region, aes(x = Region, y = Avg_Price, fill = Region)) +
  geom_bar(stat = "identity", width = 0.6) +
  theme_minimal() +
  labs(title = "Average Smartphone Price per Region (USD)",
       x = "Region",
       y = "Average Price (USD)")
```



2. Create a pie chart of the market share of smartphone brands.

```
brand_market_share <- df %>%
  group_by(Company.Name) %>%
  summarize(Model_Count = n()) %>%
  arrange(desc(Model_Count))

ggplot(brand_market_share, aes(x = "", y = Model_Count, fill = Company.Name)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  theme_void() +
  labs(title = "Market Share of Smartphone Brands")
```

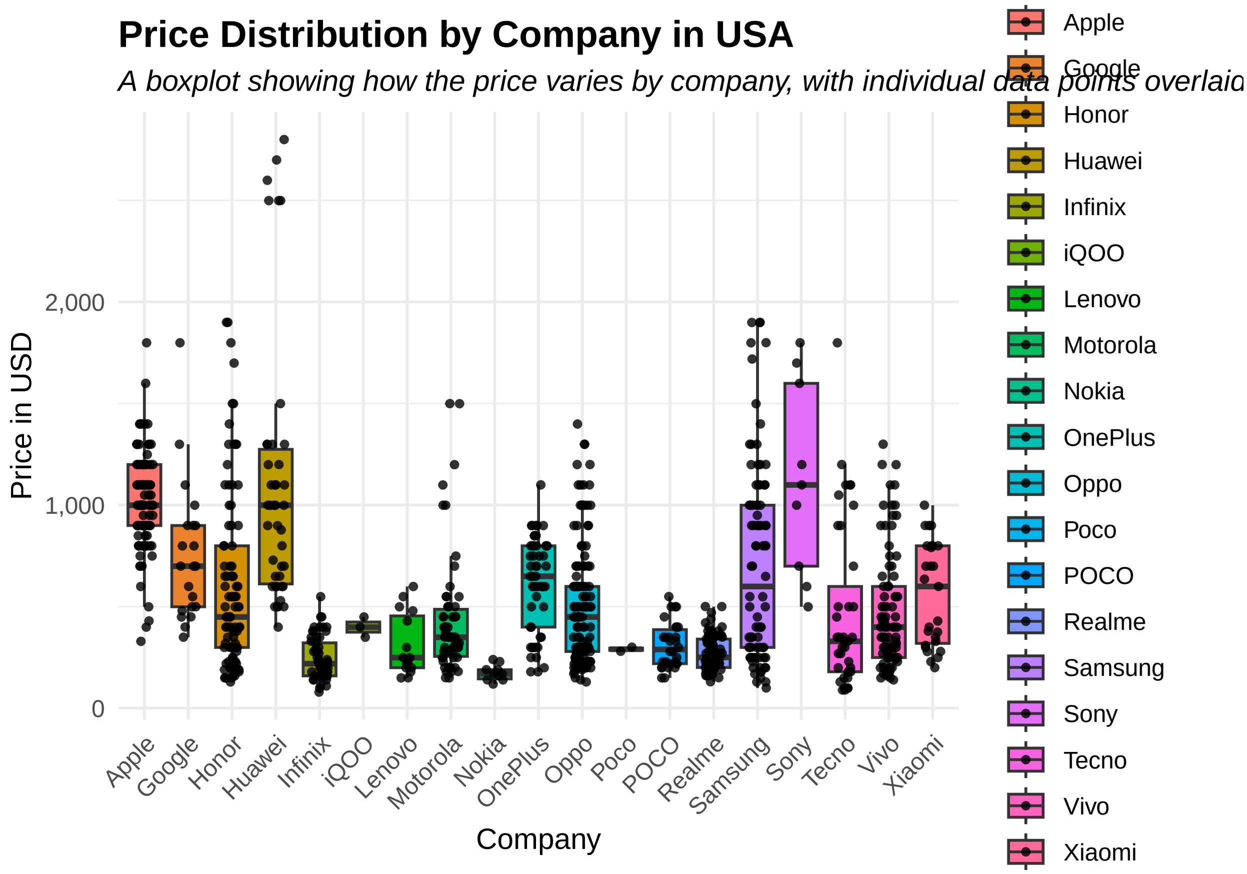


Part 3

1st Chart

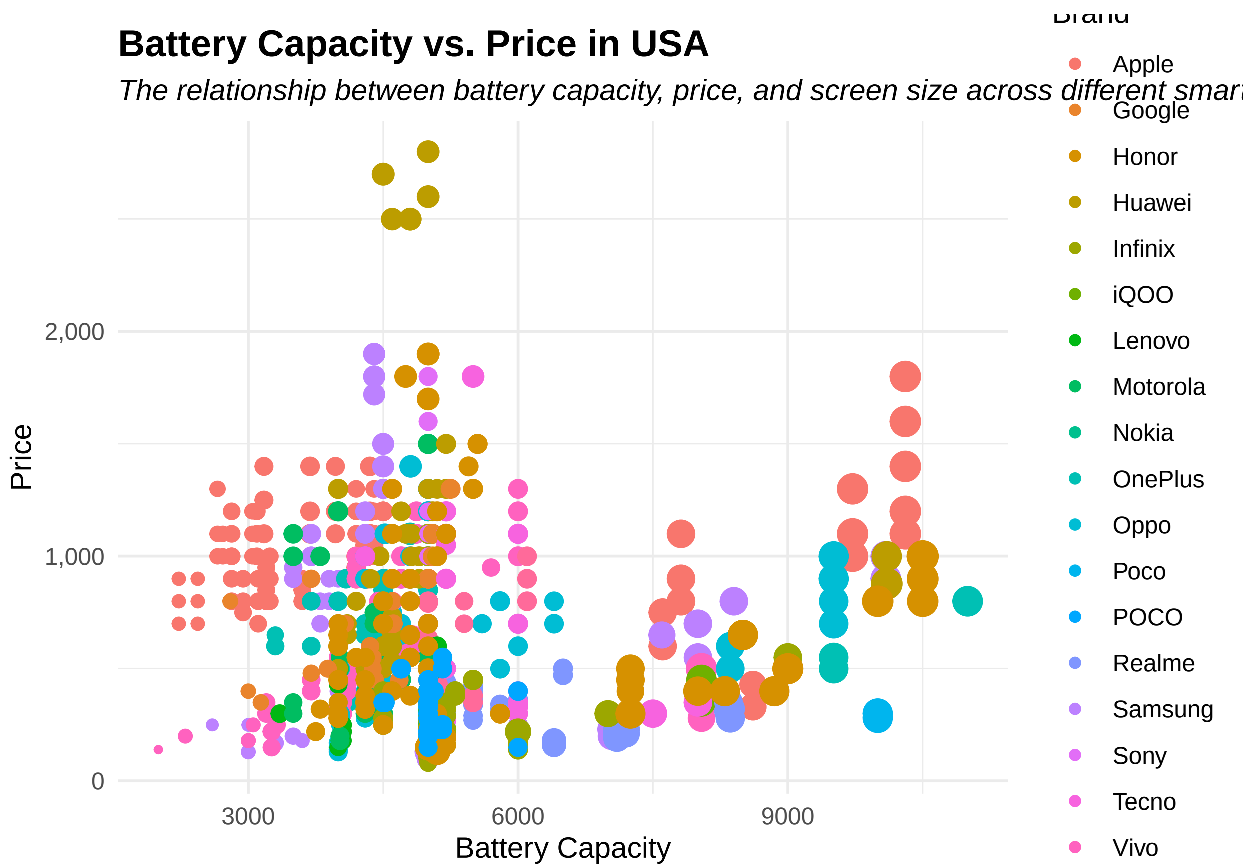
```
data <- read.csv("mobiles_dataset.csv")

ggplot(data, aes(x = Company.Name,
                 y = Launched.Price.USA.USD,
                 fill = Company.Name)) +
  geom_boxplot(outlier.shape = NA, alpha = 1) +  # Set alpha to 1 for solid colors
  geom_jitter(width = 0.2, size = 1, alpha = 0.8, color = "black") +
  scale_y_continuous(labels = scales::comma) +
  labs(title = "Price Distribution by Company in USA",
       subtitle = "A boxplot showing how the price varies by company, with individual data points overla
       x = "Company",
       y = "Price in USD") +
  theme_minimal() +
  theme(legend.position = "right",
        axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(face = "bold", size = 14),
        plot.subtitle = element_text(size = 11, face = "italic"))
```



2nd Chart

```
ggplot(data, aes(x = Battery.Capacity.mAh,
                 y = Launched.Price.USA.USD,
                 color = Company.Name,
                 size = Screen.Size.inches)) +
  geom_point(alpha = 1) +  # Set alpha to 1 for solid colors
  scale_size_continuous(range = c(1, 5), guide = "none") +  # Keep sizes but remove from legend
  scale_y_continuous(labels = scales::comma) +
  labs(title = "Battery Capacity vs. Price in USA",
       subtitle = "The relationship between battery capacity, price, and screen size across different s
       x = "Battery Capacity",
       y = "Price",
       color = "Brand") +
  theme_minimal() +
  theme(legend.position = "right",
        plot.title = element_text(face = "bold", size = 14),
        plot.subtitle = element_text(size = 11, face = "italic"))
```



3rd Chart

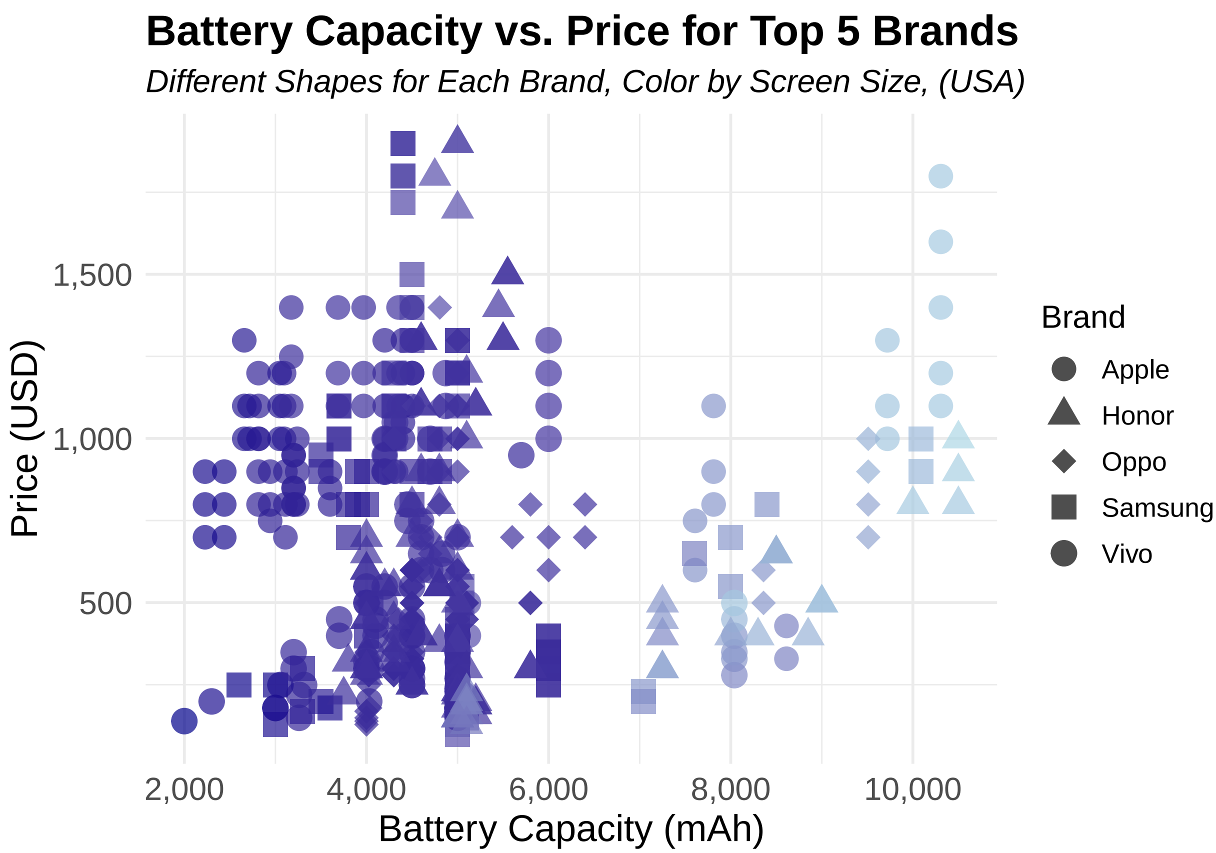
```r
data <- read.csv("mobiles_dataset.csv")

top_brands <- c("Apple", "Honor", "Oppo", "Samsung", "Vivo")
data_filtered <- data %>% filter(Company.Name %in% top_brands)

shapes <- c("Apple" = 16, "Honor" = 17, "Oppo" = 18, "Samsung" = 15, "Vivo" = 19)

ggplot(data_filtered, aes(x = Battery.Capacity.mAh,
                          y = Launched.Price.USA.USD,
                          shape = Company.Name)) +
  geom_point(aes(color = Screen.Size.inches), size = 4, alpha = 0.7) +
  scale_shape_manual(values = shapes) +
  scale_color_gradient(low = "darkblue", high = "lightblue", guide = "none") +
  scale_y_continuous(labels = scales::comma) +
  scale_x_continuous(labels = scales::comma) +
  labs(title = "Battery Capacity vs. Price for Top 5 Brands",
       subtitle = "Different Shapes for Each Brand, Color by Screen Size, (USA)",
       x = "Battery Capacity (mAh)",
       y = "Price (USD)",
       shape = "Brand") +
  theme_minimal() +
  theme(legend.position = "right",
        plot.title = element_text(face = "bold", size = 16),
        plot.subtitle = element_text(size = 12, face = "italic"),
```

```
      axis.title = element_text(size = 14),
      axis.text = element_text(size = 12),
      legend.title = element_text(size = 12),
      legend.text = element_text(size = 10))
```



Part 4

```
data <- read.csv("mobiles_dataset.csv")

data$RAM <- as.numeric(gsub("GB", "", data$RAM))
data$Front.Camera <- as.numeric(gsub("MP", "", data$Front.Camera))
data$Back.Camera <- as.numeric(gsub("MP", "", data$Back.Camera))
data$Mobile.Weight <- as.numeric(gsub("g", "", data$Mobile.Weight))

data <- na.omit(data)

numeric_data <- data %>% select(RAM, Front.Camera, Back.Camera, Mobile.Weight,
                                Battery.Capacity.mAh, Screen.Size.inches,
                                Launched.Price.USA.USD)

cor_matrix <- cor(numeric_data)
print(cor_matrix)
```

```
##                                RAM Front.Camera Back.Camera Mobile.Weight
## RAM                     1.00000000    0.5042749   0.4830147   0.037621622
## Front.Camera            0.50427494    1.0000000   0.6187113  -0.316214918
## Back.Camera             0.48301474    0.6187113   1.0000000  -0.365288891
## Mobile.Weight           0.03762162   -0.3162149  -0.3652889   1.000000000
## Battery.Capacity.mAh    0.17797906   -0.2307056  -0.1916805   0.856422376
## Screen.Size.inches      0.08699954   -0.2620625  -0.3154501   0.976485453
## Launched.Price.USA.USD  0.64461997    0.2654214   0.2501874   0.009296035
##                        Battery.Capacity.mAh Screen.Size.inches
## RAM                              0.17797906         0.08699954
## Front.Camera                    -0.23070557        -0.26206254
## Back.Camera                     -0.19168054        -0.31545014
## Mobile.Weight                    0.85642238         0.97648545
## Battery.Capacity.mAh             1.00000000         0.88613071
## Screen.Size.inches               0.88613071         1.00000000
## Launched.Price.USA.USD          -0.00873267         0.01654036
##                        Launched.Price.USA.USD
## RAM                               0.644619967
## Front.Camera                      0.265421413
## Back.Camera                       0.250187430
## Mobile.Weight                     0.009296035
## Battery.Capacity.mAh             -0.008732670
## Screen.Size.inches                0.016540365
## Launched.Price.USA.USD            1.000000000
```

The analysis shows that RAM is the strongest predictor of mobile phone prices, with a high correlation (0.64), indicating that devices with more RAM tend to be more expensive. Camera specifications also have a moderate impact, with correlations of 0.27 (front camera) and 0.25 (back camera), showing that while better cameras contribute to higher prices, they are not the primary determinant. Interestingly, battery capacity and screen size have little to no influence on price, indicating that brands do not significantly charge more for bigger batteries or larger displays. Additionally, mobile weight is strongly correlated with battery size (0.86) and screen size (0.97) but does not directly impact pricing.

To conclude, RAM and brand perception are the most significant factors influencing price, while battery capacity, screen size, and weight have minimal impact.