AppleStore

Alec Brooks

2024-10-27

In the ever-expanding mobile app marketplace, understanding user behavior and feedback is crucial for app developers and companies. One of the key metrics that influences the success of an app is its user rating. A high user rating can lead to more downloads, better visibility in app stores, and ultimately more revenue. However, many factors may influence how users rate an app, including whether the app is free or paid.

In this project, we will investigate whether the proportion of users giving high ratings (e.g., more than 3 stars) differs between costed apps (paid) and free apps in the mobile app market. Using a dataset from Kaggle titled "Mobile App Store", we will perform a **two-sample proportion hypothesis test** to determine if there is a statistically significant difference in user ratings between these two categories of apps.

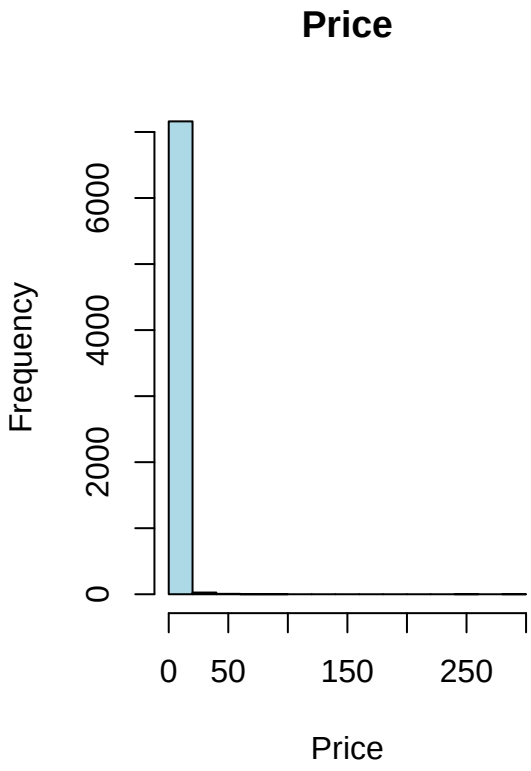
1) Please explore and learn about the data to understand this project better. Report what you found that should be enough to run the test.

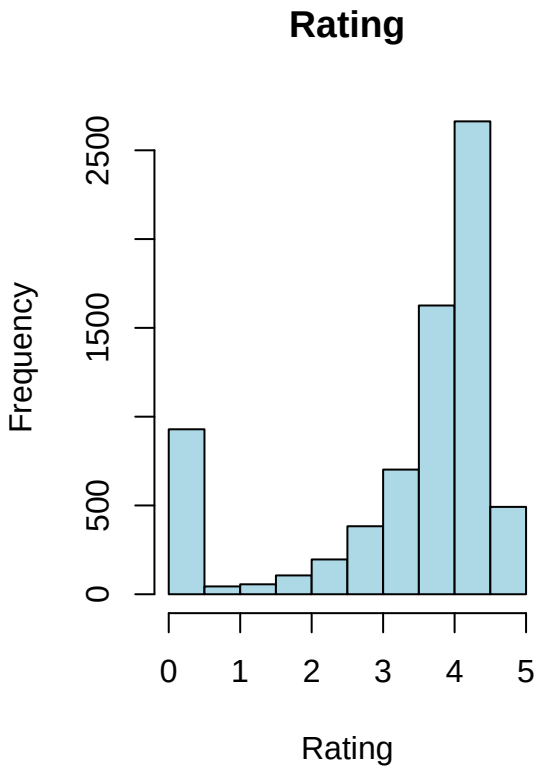
After reviewing the dataset dictionary, I identified three variables that will be useful for this analysis: id, price, and user_rating. The id variable will act as a unique identifier, with price and user_rating being the focus of the analysis. Although version ratings are included, the analysis will compare price to rating. Since there's no information on price changes across versions, using an average rating across versions seems most appropriate.

Some transformations will be needed as well. The price variable can be converted to a categorical variable, identifying items as either 'free' or 'paid,' because all we need to know is whether an app was free or not. Additionally, user_rating can also be made categorical, differentiating between ratings of more than 3 stars and less.

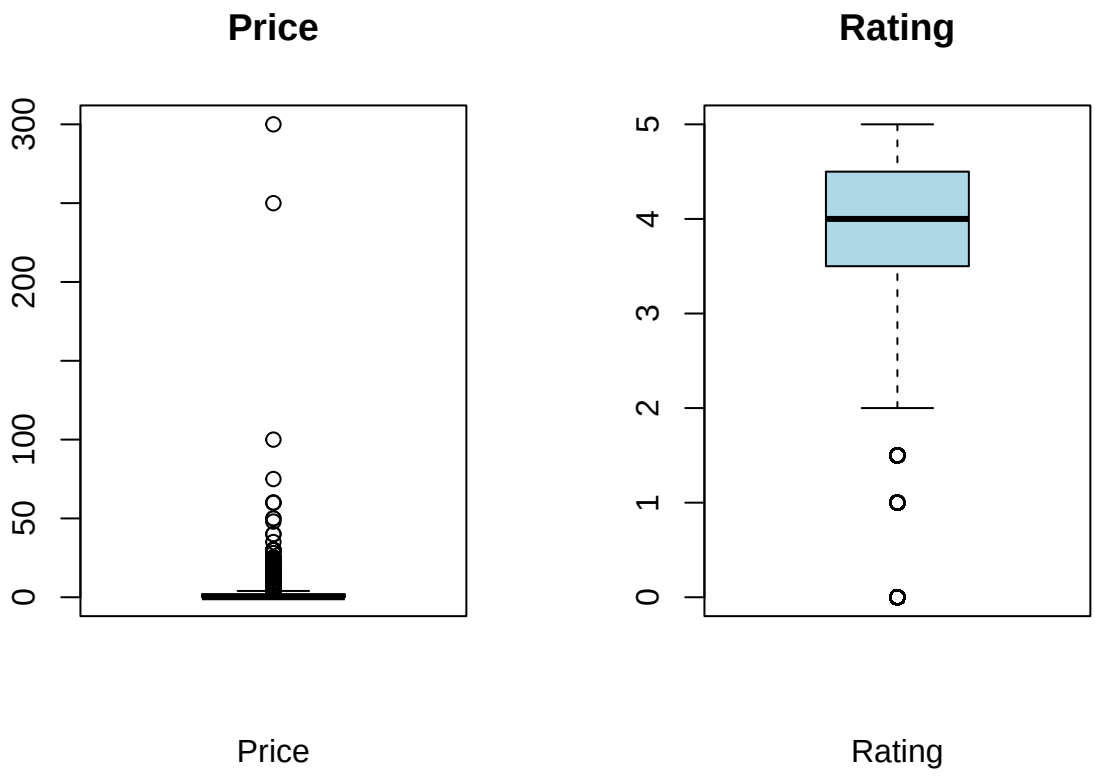
2) Let's visualize (graphs) a few variables that will be used for this project. What did you learn?

```
par(mfrow = c(1, 2))
hist(AppleStore$price, main = "Price", xlab = "Price", col = "lightblue")
hist(AppleStore$user_rating, main = "Rating", xlab = "Rating",col = "lightblue")
```






```r
par(mfrow = c(1, 2))
boxplot(AppleStore$price, main = "Price", xlab = "Price", col = "lightblue")
boxplot(AppleStore$user_rating, main = "Rating", xlab = "Rating",col = "lightblue")
```

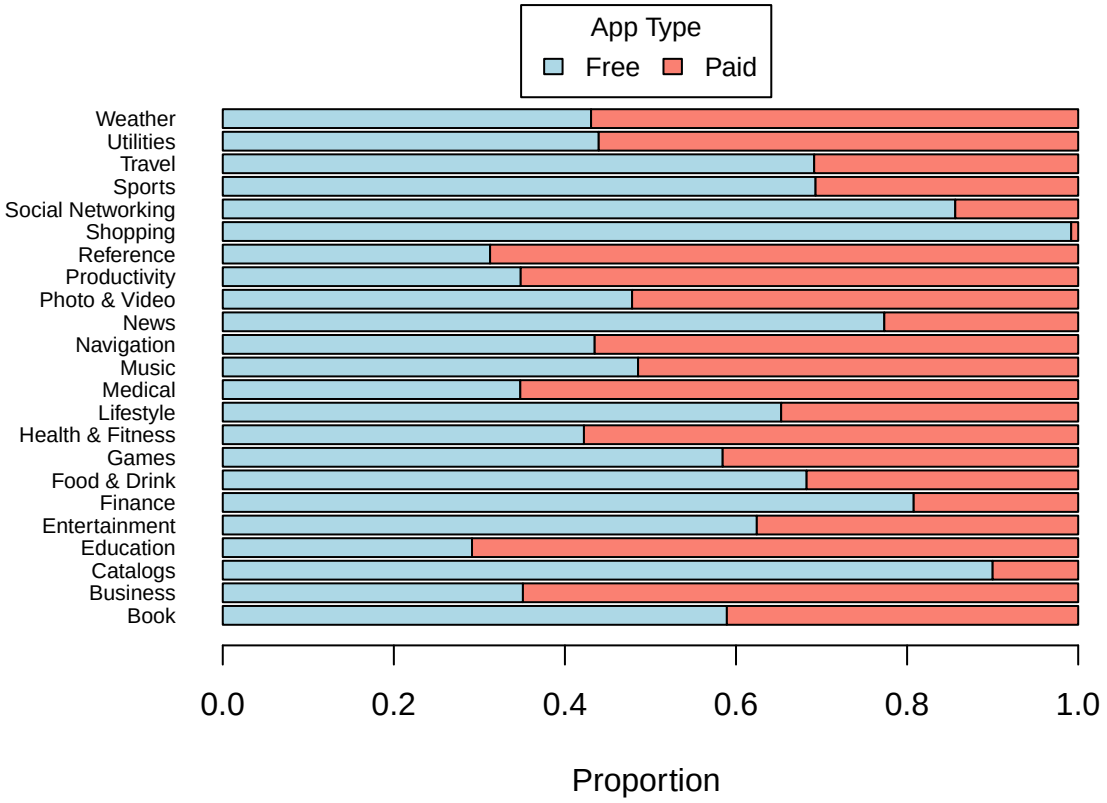


Examining the boxplots, we can see significant outliers in both variables. Most of the price data clusters below five dollars, with some large outliers reaching up to $300. Initially, I thought this was an issue in the data, but Googleing the app's name confirmed the price accuracy. This won't impact the test since the exact dollar amount is irrelevant, but it's interesting to note.

Looking at the ratings we see its skewed to the right, with a mean around four. There are some outliers on the lower end, which makes sense for user ratings, considering people often rate apps only when they have a very positive or very negative opinion of it. What's interesting, in the rating histogram, is that people tend to rate apps they like much more often than those they dislike, leading to an overall higher average."

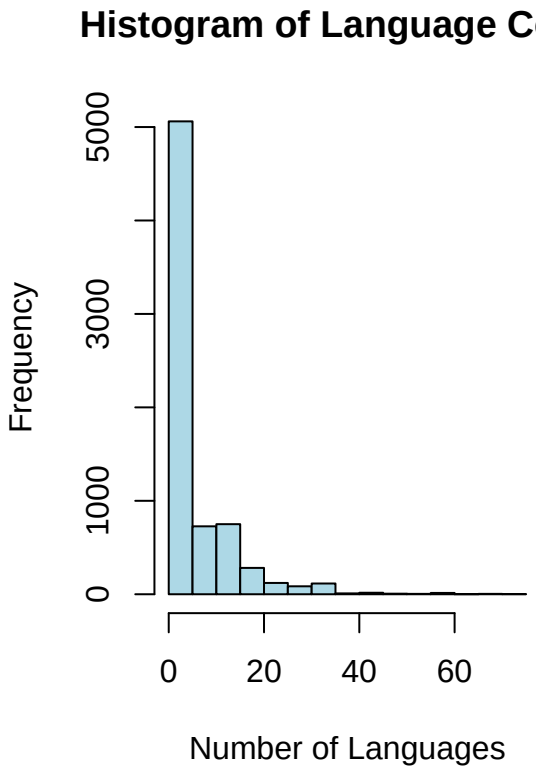
3) Out of curiosity, is there a difference in the genres offered between free and paid apps?

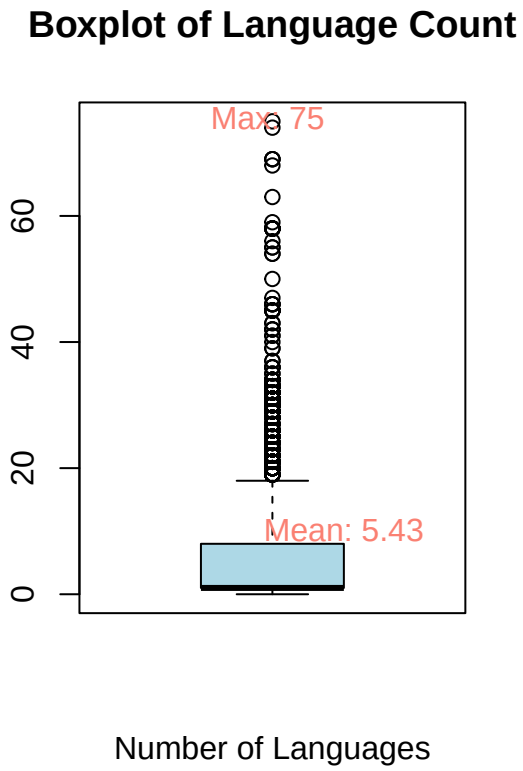
```
genre_proportions <- prop.table(rbind(table(AppleStore[AppleStore$price == 0, ]$prime_genre),
                    table(AppleStore[AppleStore$price > 0, ]$prime_genre)), margin = 2)
par(mar = c(4, 8, 4, 2))
barplot(genre_proportions, horiz = TRUE, col = c("lightblue", "salmon"),
        main = "",
        xlab = "Proportion", ylab = "", cex.names = 0.7, las = 1)
legend("top", inset = c(-0.1, -0.15), legend = c("Free", "Paid"),
       fill = c("lightblue", "salmon"),
       title = "App Type", cex = 0.8, xpd = TRUE, horiz = TRUE)
```



4) Out of another curiosity, how many languages are usually supported in each app? How many languages could be the most?

```
par(mfrow = c(1, 2))
hist(AppleStore$lang.num, col = "lightblue", main = "Histogram of Language Count",
     xlab = "Number of Languages", ylab = "Frequency")
boxplot(AppleStore$lang.num, main = "Boxplot of Language Count",
        xlab = "Number of Languages", col = "lightblue")
text(1.2, mean(AppleStore$lang.num), labels =
       paste("Mean:", round(mean(AppleStore$lang.num), 2)), col = "salmon", pos = 3)
text(1.2, max(AppleStore$lang.num), labels =
       paste("Max:", max(AppleStore$lang.num)), col = "salmon", pos = 2)
```





5) Clean up: Now, let's get ready to run the hypothesis test. What numbers do we need to run the test?

```
#remove cols
AppleStore <- AppleStore[,c(2,6,9)]
#fix price
AppleStore$price[AppleStore$price > 0] <- "paid"
AppleStore$price[AppleStore$price == 0] <- "free"
#fix ratings
AppleStore$user_rating[AppleStore$user_rating <= 3] <- "<=3"
AppleStore$user_rating[AppleStore$user_rating > 3] <- ">3"
```

6) Perform the hypothesis test and interpret the outcome. What does it mean?

Null Hypothesis: The proportion of paid apps with average ratings of greater than three is equal to the proportion of free apps with average ratings of greater than three.

Alternative Hypothesis: The proportion of paid apps with average ratings of greater than three is *not* equal to the proportion of free apps with average ratings of greater than three.

- p_1 is the proportion of paid apps with ratings > 3
- p_2 is the proportion of free apps with ratings > 3

$$H_0 : p_1 = p_2$$
$$H_a : p_1 \neq p_2$$

```r
paid_pop <- sum(AppleStore$price == "paid")
free_pop <- sum(AppleStore$price == "free")
paid_proportion <- sum(AppleStore$price == "paid" & AppleStore$user_rating == ">3")
free_proportion <- sum(AppleStore$price == "free" & AppleStore$user_rating == ">3")

prop.test(x = c(paid_proportion, free_proportion), n = c(paid_pop, free_pop),
          alternative = "two.sided", conf.level = 0.95, correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  c(paid_proportion, free_proportion) out of c(paid_pop, free_pop)
## X-squared = 66.409, df = 1, p-value = 3.664e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.06306219 0.10194386
## sample estimates:
##    prop 1    prop 2
## 0.8083413 0.7258383
```

With an effective p-value of 0, we can reject our null hypothesis, showing that there is enough statistical evidence to conclude that ratings are significantly different between free and paid apps. Further evidence supporting the rejection of the null hypothesis is the 95% confidence interval, which shows $0.0631 < p < 0.1049$, indicating that $p_1 \neq p_2$. The estimated difference is $0.80 - 0.72 = 0.08$, suggesting an approximate 8% difference between the two proportions.

7) How did you feel after completing the project?

Good. I feel like i understand hypothesis testing alot better now, and could use it in future settings. (prop.test)

full Code

```r
library(readr)
AppleStore <- read_csv("AppleStore.csv")

par(mfrow = c(1, 2))
hist(AppleStore$price, main = "Price", xlab = "Price", col = "lightblue")
hist(AppleStore$user_rating, main = "Rating", xlab = "Rating",col = "lightblue")

par(mfrow = c(1, 2))
boxplot(AppleStore$price, main = "Price", xlab = "Price", col = "lightblue")
boxplot(AppleStore$user_rating, main = "Rating", xlab = "Rating",col = "lightblue")
```

```r
genre_proportions <- prop.table(rbind(table(AppleStore[AppleStore$price == 0, ]$prime_genre),
                    table(AppleStore[AppleStore$price > 0, ]$prime_genre)), margin = 2)
par(mar = c(4, 8, 4, 2))
barplot(genre_proportions, horiz = TRUE, col = c("lightblue", "salmon"),
        main = "",
        xlab = "Proportion", ylab = "", cex.names = 0.7, las = 1)
legend("top", inset = c(-0.1, -0.15), legend = c("Free", "Paid"),
       fill = c("lightblue", "salmon"),
       title = "App Type", cex = 0.8, xpd = TRUE, horiz = TRUE)

par(mfrow = c(1, 2))
hist(AppleStore$lang.num, col = "lightblue", main = "Histogram of Language Count",
     xlab = "Number of Languages", ylab = "Frequency")
boxplot(AppleStore$lang.num, main = "Boxplot of Language Count",
        xlab = "Number of Languages", col = "lightblue")
text(1.2, mean(AppleStore$lang.num), labels =
       paste("Mean:", round(mean(AppleStore$lang.num), 2)), col = "salmon", pos = 3)
text(1.2, max(AppleStore$lang.num), labels =
       paste("Max:", max(AppleStore$lang.num)), col = "salmon", pos = 2)

#remove cols
AppleStore <- AppleStore[,c(2,6,9)]
#fix price
AppleStore$price[AppleStore$price > 0] <- "paid"
AppleStore$price[AppleStore$price == 0] <- "free"
#fix ratings
AppleStore$user_rating[AppleStore$user_rating <= 3] <- "<=3"
AppleStore$user_rating[AppleStore$user_rating > 3] <- ">3"

paid_pop <- sum(AppleStore$price == "paid")
free_pop <- sum(AppleStore$price == "free")
paid_proportion <- sum(AppleStore$price == "paid" & AppleStore$user_rating == ">3")
free_proportion <- sum(AppleStore$price == "free" & AppleStore$user_rating == ">3")

prop.test(x = c(paid_proportion, free_proportion), n = c(paid_pop, free_pop),
          alternative = "two.sided", conf.level = 0.95, correct = FALSE)
```