Pricing Test Challenge

Mitul Shah

8/7/2017

```
## Loading the required libraries
library(magrittr)
library(dplyr)
library(ggplot2)
library(ggmap)
library(maps)
library(rworldmap)
```

Loading the data

```
## Loading the data
user_table <- read.csv("user_table.csv")
test_results <- read.csv("test_results.csv")
```

Checking Data Quality

```
## Are there any duplicate entries ?
length(unique(user_table$user_id)) == length(user_table$user_id) ## Looks good!
```

```
## [1] TRUE
```

```
length(unique(test_results$user_id)) == length(test_results$user_id) ## Looks good!
```

```
## [1] TRUE
```

```
## Merge two datasets
data = merge(test_results,user_table, by = "user_id", all.x = TRUE)

## Looking at the structure
str(data)
```

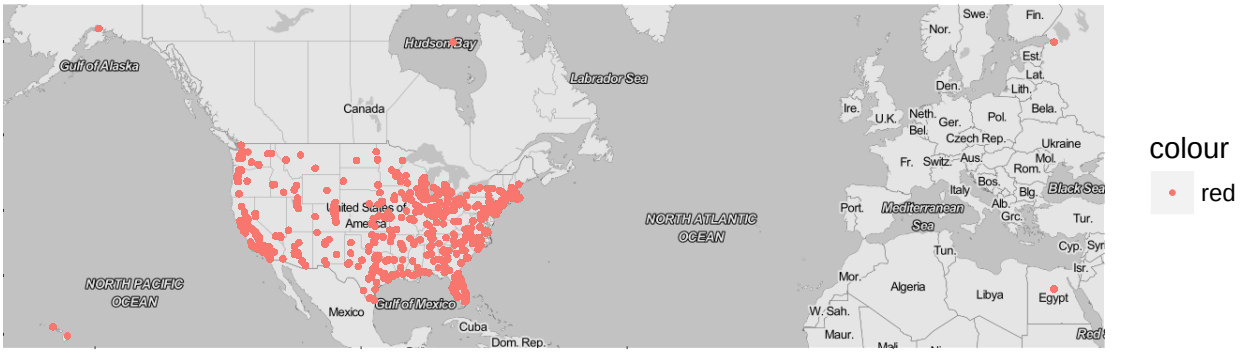
```
## 'data.frame':    316800 obs. of  12 variables:
##  $ user_id         : int  3 9 14 16 19 22 23 24 27 30 ...
##  $ timestamp       : Factor w/ 140931 levels "2015-03-02 00:04:12",..: 71600 90540 41454 125216 4489
##  $ source          : Factor w/ 12 levels "ads_facebook",..: 8 10 7 4 4 3 4 6 4 4 ...
##  $ device          : Factor w/ 2 levels "mobile","web": 2 1 1 1 1 2 2 1 1 2 ...
##  $ operative_system: Factor w/ 6 levels "android","iOS",..: 4 1 2 1 1 6 6 1 2 6 ...
##  $ test            : int  1 0 0 0 0 0 1 0 0 1 ...
##  $ price           : int  59 39 39 39 39 39 59 39 39 59 ...
```

```
##  $ converted       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ city            : Factor w/ 923 levels "Abilene","Akron",..: 587 100 909 23 2 857 112 655 NA 393
##  $ country         : Factor w/ 1 level "USA": 1 1 1 1 1 1 1 1 NA 1 ...
##  $ lat             : num  38.9 41.7 39.7 38 41.1 ...
##  $ long            : num  -94.8 -72.9 -75.5 -121.8 -81.5 ...
```

```
## Looking at the summary
summary(data)
```

```
##     user_id                      timestamp                source     
##  Min.   :      3   2015-04-12 11:51:16:    12   direct_traffic:60357  
##  1st Qu.: 249526   2015-04-04 17:38:26:    11   ads-google    :59379  
##  Median : 499022   2015-04-10 08:29:07:    11   ads_facebook  :53396  
##  Mean   : 499281   2015-05-25 07:27:08:    11   ads_other     :29876  
##  3rd Qu.: 749026   2015-03-06 12:23:20:    10   seo-google    :23175  
##  Max.   :1000000   2015-03-06 17:26:54:    10   ads-bing      :22873  
##                    (Other)            :316735   (Other)       :67744  
##     device       operative_system      test            price       
##  mobile:186471   android: 74935   Min.   :0.0000   Min.   :39.00  
##  web   :130329   iOS    : 95465   1st Qu.:0.0000   1st Qu.:39.00  
##                  linux  :  4135   Median :0.0000   Median :39.00  
##                  mac    : 25085   Mean   :0.3601   Mean   :46.21  
##                  other  : 16204   3rd Qu.:1.0000   3rd Qu.:59.00  
##                  windows:100976   Max.   :1.0000   Max.   :59.00  
##                                                                   
##    converted               city        country            lat        
##  Min.   :0.00000   New York   : 25748   USA :275616   Min.   :19.70  
##  1st Qu.:0.00000   Chicago    :  7153   NA's: 41184   1st Qu.:33.66  
##  Median :0.00000   Houston    :  6706                 Median :37.74  
##  Mean   :0.01833   San Antonio:  4633                 Mean   :37.11  
##  3rd Qu.:0.00000   Los Angeles:  4141                 3rd Qu.:40.70  
##  Max.   :1.00000   (Other)    :227235                 Max.   :61.18  
##                    NA's       : 41184                 NA's   :41184  
##       long        
##  Min.   :-157.80  
##  1st Qu.:-112.20  
##  Median : -88.93  
##  Mean   : -93.98  
##  3rd Qu.: -78.91  
##  Max.   :  30.31  
##  NA's   :41184
```

```
## Plotting the data on the world map
qmplot(long, lat, data = data, size = I(.5), darken = .1, color = "red") ## Something wrong!!
```



There are some locations in the data which are not in USA (for instance we can see a point in Egypt, etc.), but the data only has USA in the country variable. So, something is definitely wrong for these points. But as there are hardly such points, I haven't removed these locations from the data for the sake of convenience.

```
## Check for the test and the price column (these columns indicate the same thing)
nrow(subset(data, data$test == 1 & data$price == 39)) ## Something wrong!
```

```
## [1] 155
```

```
nrow(subset(data, data$test == 0 & data$price == 59)) ## Something wrong!
```

```
## [1] 210
```

Thus, there is something wrong with these 365 rows. Let's remove them.

```
## Removing the above rows
data <- filter(data, (data$test == 1 & data$price == 59) | data$test == 0 & data$price == 39)
```

Now, let's see the performance of the control and the test group.

Should the company sell its software for $39 or $59?

```
## Applying t-test to compare the performance of the control group and the test group
t.test(data$converted[data$test == 1], data$converted[data$test == 0])
```

```
##
##  Welch Two Sample t-test
##
## data:  data$converted[data$test == 1] and data$converted[data$test == 0]
## t = -9.0446, df = 260430, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.005285975 -0.003403056
## sample estimates:
##  mean of x  mean of y
## 0.01555505 0.01989956
```

The results show that the test group had a conversion rate of 0.0155 while the control group had the conversion rate of 0.0199.

In order to decide whether the company should sell its software for $39 or $59, we just need to compare 0.0155 * $59 and 0.0199 * $39. The first value (0.0155 * $59) is greater than the second value. Hence, the company should sell its software for $59 if the company just cares about generating higher revenue. On the other hand, if the company is a startup, it might be more interested in getting more users. In this case, it should continue to sell its software for $39.

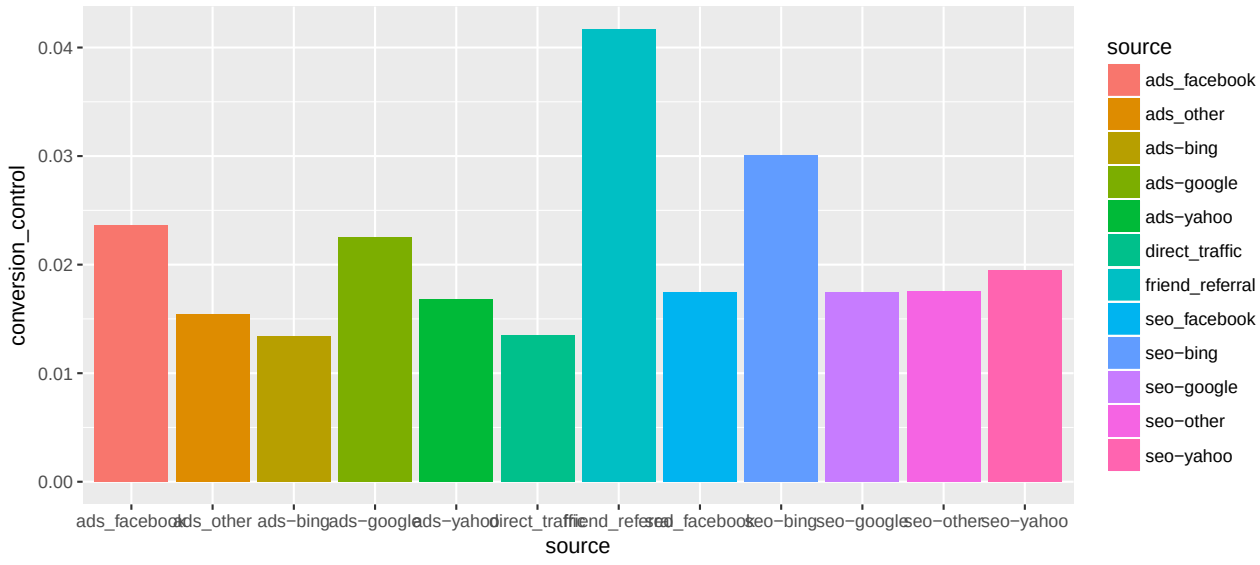
```
## Changing the mode of timestamp
data$timestamp <- as.Date(data$timestamp, format = "%Y-%m-%d")
```

```
## Warning in strptime(x, format, tz = "GMT"): unknown timezone 'zone/tz/
## 2018c.1.0/zoneinfo/America/New_York'
```

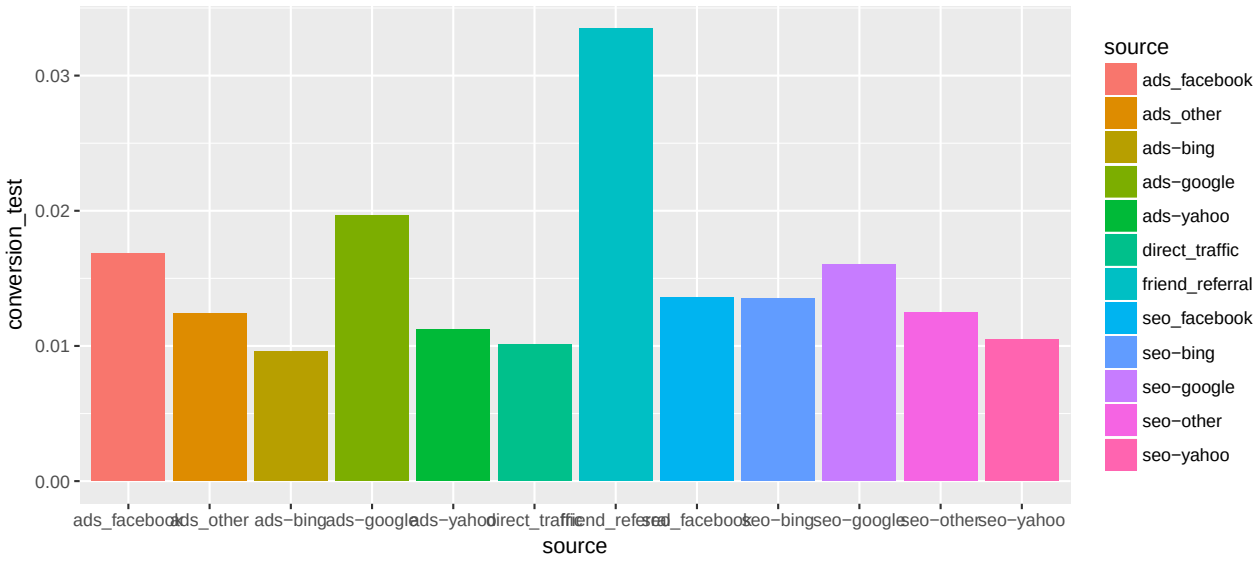
Main Findings from the data

```
## Conversion rate by source
data_source <- data %>%
  group_by(source) %>%
  summarise(conversion_control = mean(converted[test == 0]), conversion_test = mean(converted[test == 1]
```

```
## Plotting conversion rate by source for the control group
ggplot(data = data_source, aes(x = source, y = conversion_control)) +
  geom_bar(stat = "identity", aes(fill = source))
```



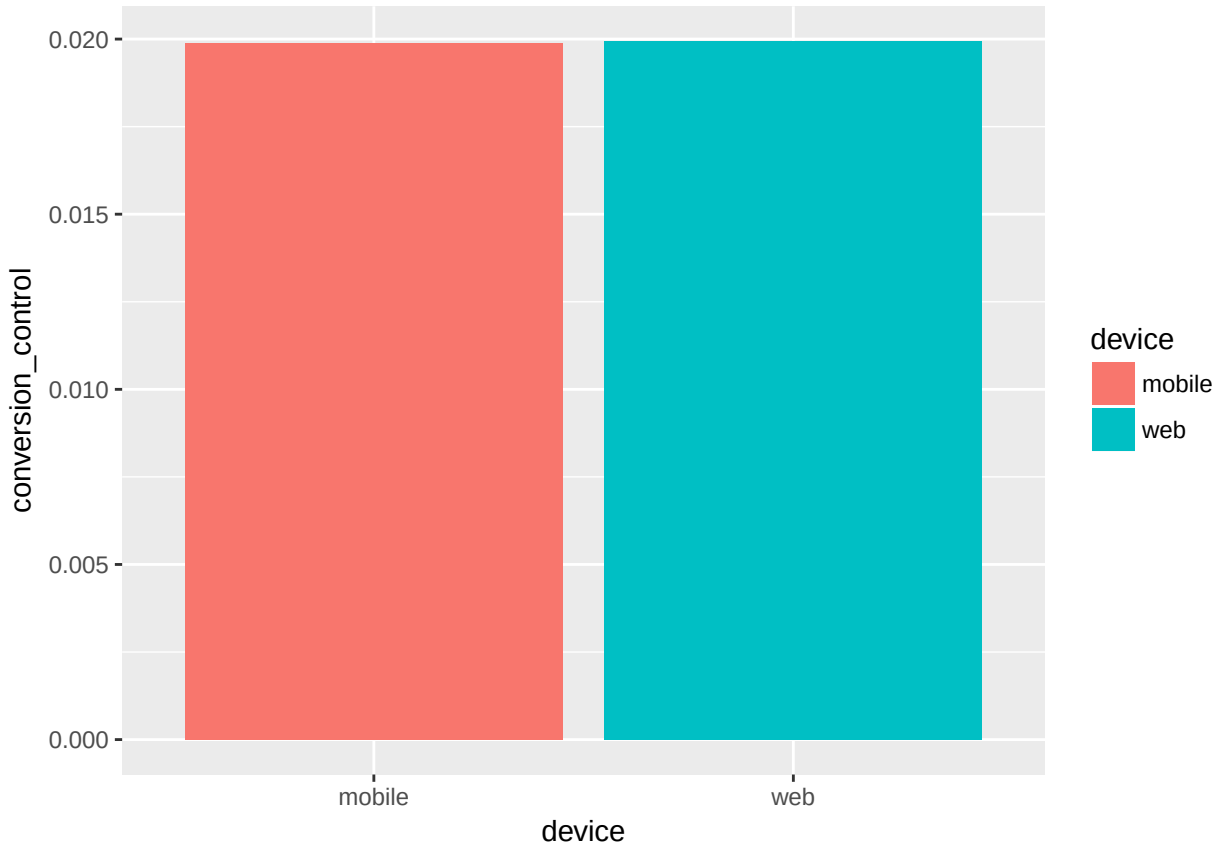

```
## Plotting conversion rate by source for the test group
ggplot(data = data_source, aes(x = source, y = conversion_test)) +
  geom_bar(stat = "identity", aes(fill = source))
```



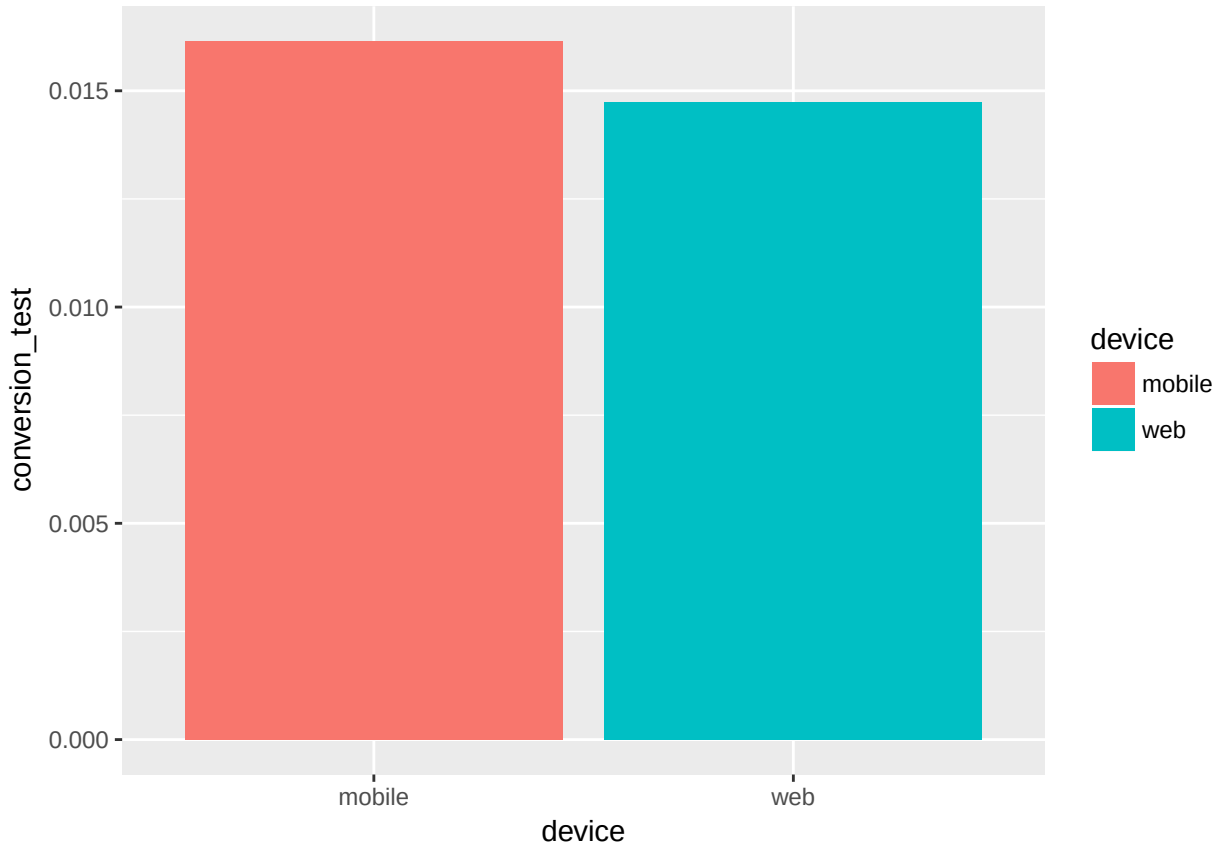
The friend referral has led to the maximum conversion rate. Hence, it might be helpful for the marketing team to send out emails with some sort of deals (like if you refer the product to someone and someone buys it, you get 10% discount and you can refer to as many friends as you can). This might help the company to maximize its revenue.

```
## Conversion rate by device
data_device <- data %>%
  group_by(device) %>%
  summarise(conversion_control = mean(converted[test == 0]), conversion_test = mean(converted[test == 1]
```

```
## Plotting conversion rate by device for the control group
ggplot(data = data_device, aes(x = device, y = conversion_control)) +
  geom_bar(stat = "identity", aes(fill = device))
```



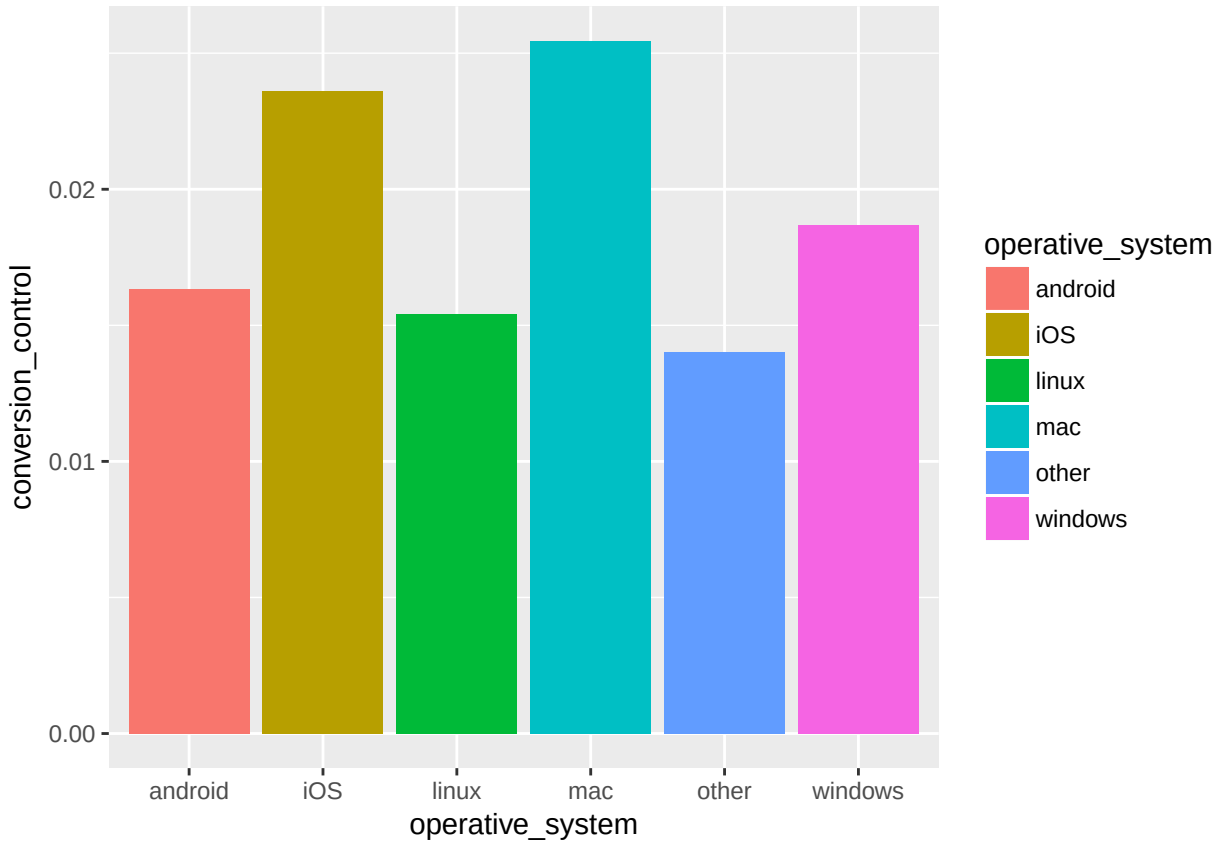

```
## Plotting conversion rate by device for the test group
ggplot(data = data_device, aes(x = device, y = conversion_test)) +
  geom_bar(stat = "identity", aes(fill = device))
```



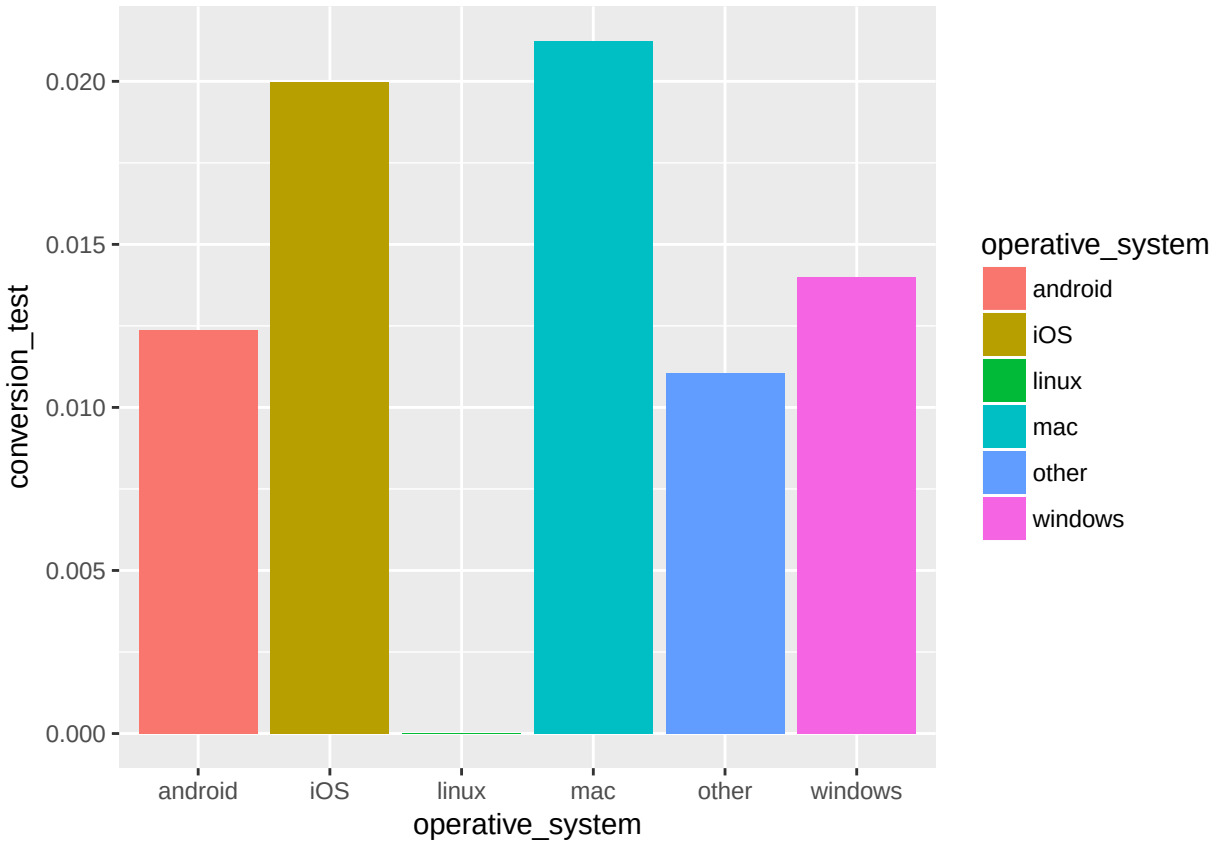
The type of device hasn't affected the conversion rate too much. It seems to be an insignificant variable. Mobile users as well as web users have almost the same conversion rate.

```
## Conversion rate by operative system
data_operative_system <- data %>%
  group_by(operative_system) %>%
  summarise(conversion_control = mean(converted[test == 0]), conversion_test = mean(converted[test == 1]
```

```
## Plotting conversion rate by operative system for the control group
ggplot(data = data_operative_system, aes(x = operative_system, y = conversion_control)) +
  geom_bar(stat = "identity", aes(fill = operative_system))
```




```
## Plotting conversion rate by operative system for the test group
ggplot(data = data_operative_system, aes(x = operative_system, y = conversion_test)) + geom_bar(stat =
```



The company product seems to be performing well among the Apple users (iOS and Mac has a higher conversion rate) while the linux users seem to be buying the software very less often than other users. The product team definitely need to find out why the linux users have a lower conversion rate and try to improve the product for them.

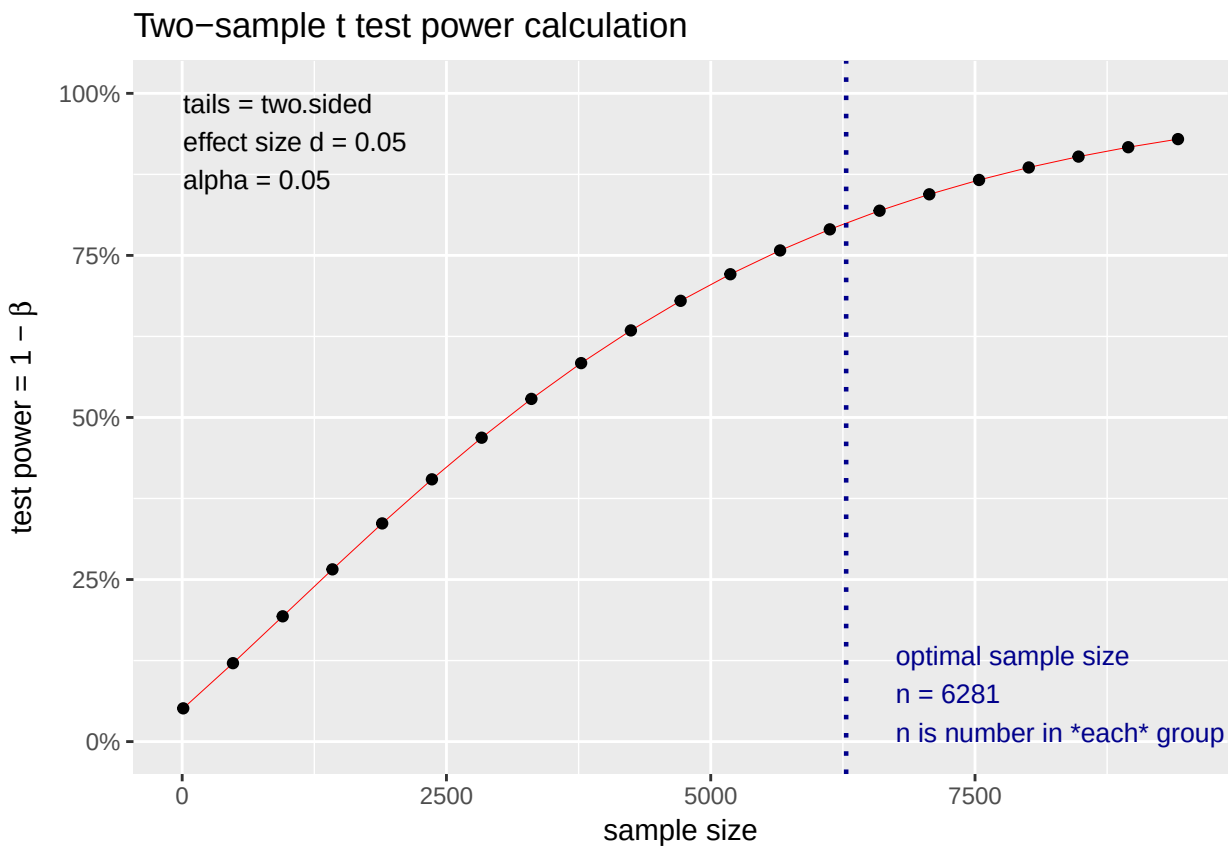
Detecting Sample Size

```
library(pwr)

test <- power.t.test(d = 0.05, sig.level = 0.05, power = 0.8, alternative = "two.sided")
test
```

```
##
##      Two-sample t test power calculation
##
##               n = 6280.064
##           delta = 0.05
##              sd = 1
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
plot(test)
```


Two−sample t test power calculation
tails = two.sided
effect size d = 0.05
alpha = 0.05
optimal sample size
n = 6281
n is number in *each* group
test power = 1 − β
100%
75%
50%
25%
0%
0
2500
5000
7500
sample size