Funnel Analysis

Mitul Shah

12/28/2016

Let's load the required libraries first.

```
## Loading the required libraries
library(magrittr)
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
```

```
library(ggplot2)
```

Loading all the datasets

```
## Loading all datasets
user_table <- read.csv("user_table.csv")
home_page_table <- read.csv("home_page_table.csv")
search_page_table <- read.csv("search_page_table.csv")
payment_page_table <- read.csv("payment_page_table.csv")
payment_confirmation_table <- read.csv("payment_confirmation_table.csv")
```

Now, let's try to merge all the datasets without losing any users.

```
## Merging all datasets without losing any user
data <- merge(user_table, home_page_table, by = "user_id")
data <- merge(data, search_page_table, by = "user_id", all.x = TRUE)
data <- merge(data, payment_page_table, by = "user_id", all.x = TRUE)
data <- merge(data, payment_confirmation_table, by = "user_id", all.x = TRUE)
```

```
## Warning in merge.data.frame(data, payment_confirmation_table, by =
## "user_id", : column names 'page.x', 'page.y' are duplicated in the result
```

Now, let's give relevant names to all the columns.

```
## Renaming the columns
colnames(data) <- c("user_id", "date", "device", "sex", "home_page", "search_page", "payment_page", "pay
```

Now, let's set all the values in the home page column to 1 as everyone visited this page.

```
## Setting home_page variable to 1 throughout the column
data$home_page <- 1
```

Now, let's set search page, payment page and payment confirmation page values to 0 for those users who did not visit these pages.

```
## Setting search_page variable to 0 for people who did not visit this page
data$search_page <- as.character(data$search_page)
data$search_page[is.na(data$search_page)] = "0"

## Setting payment_page variable to 0 for people who did not visit this page
data$payment_page <- as.character(data$payment_page)
data$payment_page[is.na(data$payment_page)] = "0"

## Setting payment_confirmation_page variable to 0 for people who did not visit this page
data$payment_confirmation_page <- as.character(data$payment_confirmation_page)
data$payment_confirmation_page[is.na(data$payment_confirmation_page)] = "0"
```

We can also set search page, payment page and payment confirmation page values to 1 for those users who visited these pages.

```
## Setting search_page, payment_page and payment_confirmation_page to 1 for people who visited these pag
data$search_page <- ifelse(data$search_page == "0", 0, 1)
data$payment_page <- ifelse(data$payment_page == "0", 0, 1)
data$payment_confirmation_page <- ifelse(data$payment_confirmation_page == "0", 0, 1)
```

Now, let's check the structure and the summary of the data.

```
## Viewing the structure of the data
str(data)
```

```
## 'data.frame':    90400 obs. of  8 variables:
##  $ user_id                : int  17 28 37 38 55 72 112 136 139 158 ...
##  $ date                   : Factor w/ 120 levels "2015-01-01","2015-01-02",..: 111 119 52 82 32 11
##  $ device                 : Factor w/ 2 levels "Desktop","Mobile": 1 1 2 2 1 1 2 1 1 1 ...
##  $ sex                    : Factor w/ 2 levels "Female","Male": 2 2 2 1 2 2 2 2 1 1 ...
##  $ home_page              : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ search_page            : num  1 0 1 1 0 0 0 0 0 0 ...
##  $ payment_page           : num  0 0 0 1 0 0 0 0 0 0 ...
##  $ payment_confirmation_page: num  0 0 0 0 0 0 0 0 0 0 ...
```

```
## Checking the summary
summary(data)
```

```
##     user_id              date           device          sex
##  Min.   :    17   2015-02-08:  877   Desktop:60200   Female:45075
```

```
##  1st Qu.:247979   2015-02-07:  846   Mobile :30200   Male  :45325
##  Median :498267   2015-02-02:  845
##  Mean   :498710   2015-02-15:  835
##  3rd Qu.:749789   2015-02-25:  830
##  Max.   :999979   2015-02-21:  829
##                   (Other)   :85338
##    home_page  search_page   payment_page    payment_confirmation_page
##  Min.   :1   Min.   :0.0   Min.   :0.0000   Min.   :0.000
##  1st Qu.:1   1st Qu.:0.0   1st Qu.:0.0000   1st Qu.:0.000
##  Median :1   Median :0.5   Median :0.0000   Median :0.000
##  Mean   :1   Mean   :0.5   Mean   :0.0667   Mean   :0.005
##  3rd Qu.:1   3rd Qu.:1.0   3rd Qu.:0.0000   3rd Qu.:0.000
##  Max.   :1   Max.   :1.0   Max.   :1.0000   Max.   :1.000
##
```

We need to change the mode of the Date variable to date.

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

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

Assessing Data Quality

```
## Are there duplicates?
length(unique(data$user_id)) == length(data$user_id)
```

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

```
## Are there any users for which data says that they visited the payment_confirmation_page but not the p
data[which(data$payment_page == 0
                        & data$payment_confirmation_page == 1), ]
```

```
## [1] user_id                  date
## [3] device                   sex
## [5] home_page                search_page
## [7] payment_page             payment_confirmation_page
## <0 rows> (or 0-length row.names)
```

```
## Are there any users for which data says that they visited the payment_page but not the search_page?
data[which(data$search_page == 0
          & data$payment_page == 1), ]
```

```
## [1] user_id                  date
## [3] device                   sex
## [5] home_page                search_page
## [7] payment_page             payment_confirmation_page
## <0 rows> (or 0-length row.names)
```

```
## Are there any users for which data says that they visited the search_page but not the home_page?
data[which(data$home_page == 0
           & data$search_page == 1), ]
```

```
## [1] user_id                   date
## [3] device                    sex
## [5] home_page                 search_page
## [7] payment_page              payment_confirmation_page
## <0 rows> (or 0-length row.names)
```

The data quality looks good.

Defining New Users

Let's define the new users as the users who came to the site after 2015-04-01.

```
## Function to define new users
define_new_user <- function(date){
  if(date >= "2015-04-01")
    return("New")
  if(date < "2015-04-01")
    return("Old")
  else
    return(NA)
}
```

Now, let's create a new column indicating whether the user is a new user or an old one.

```
## Creating a new column user_type
data$user_type <- sapply(data$date, define_new_user)
```

Full picture of Funnel for Desktop users

First, let's take all the desktop users by subsetting the data.

```
## Subsetting desktop users
desktop_users <- data %>% filter(device == "Desktop")
```

Now, let's try to analyze the overall conversion rate for the Desktop users by date and visualize them (both old users and the new users).

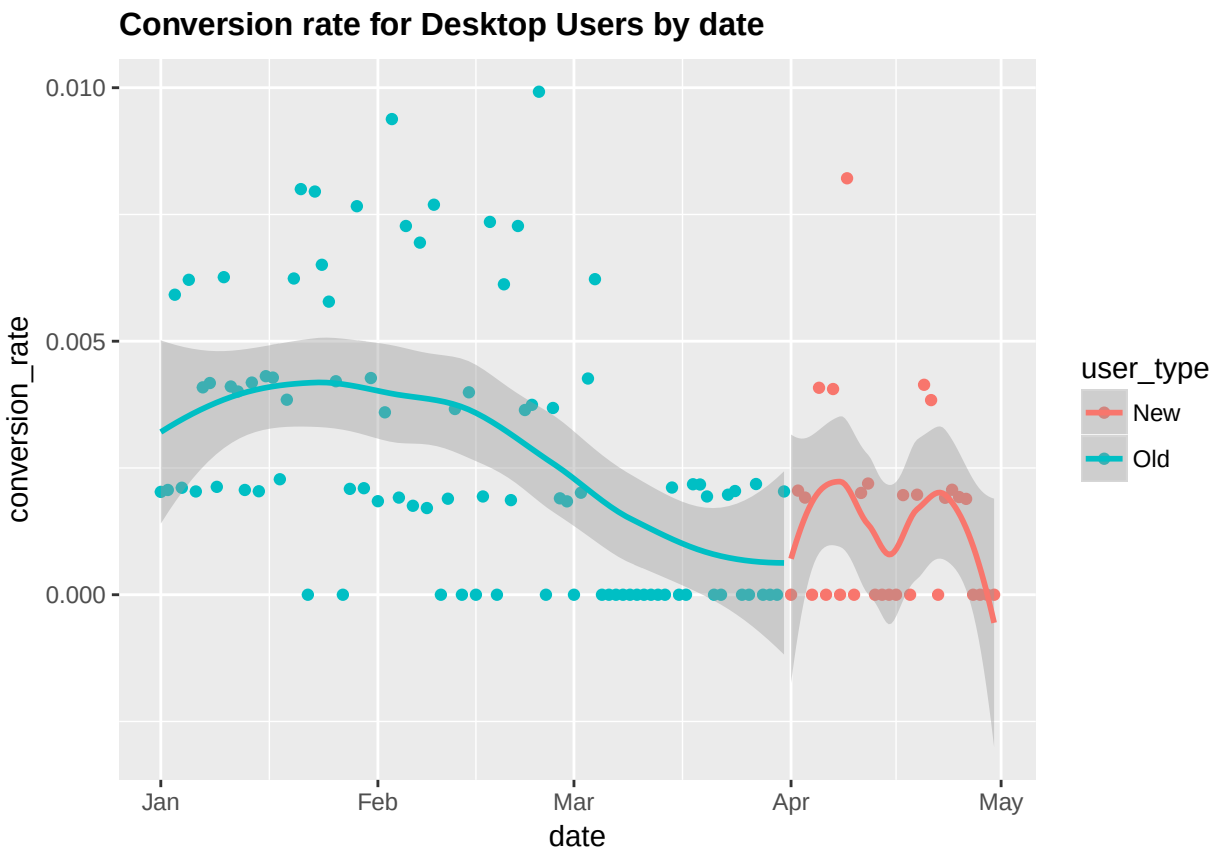
```
## Conversion rate for Desktop users by date
desktop_converted_by_date <- desktop_users %>%
                             group_by(date) %>%
                             summarise(conversion_rate =  mean(payment_confirmation_page))

## Creating a variable for user type
desktop_converted_by_date$user_type <- sapply(desktop_converted_by_date$date, define_new_user)

## Visualizing conversion rate by date
```

```
ggplot(desktop_converted_by_date,
       aes(date, conversion_rate, color = user_type)) +
       geom_point() +
       stat_smooth() +
       ggtitle("Conversion rate for Desktop Users by date") +
       theme(plot.title = element_text(size = 12, face = "bold"))
```

```
## `geom_smooth()` using method = 'loess'
```



There seems to be a strange unusual behaviour amongst the new users. But overall, the conversion rate has been decreasing heavily since 1st March.

Now, let's try to compare the conversion rate between the old users and the new users and try to find out whether the results are significant. We can use t-test in order to do this.

```
## Applying t-test to compare the conversion rate for old users and the new users
t.test(desktop_converted_by_date$conversion_rate[desktop_converted_by_date$user_type == "Old"], desktop
```

```
##
##  Welch Two Sample t-test
##
## data:  desktop_converted_by_date$conversion_rate[desktop_converted_by_date$user_type ==  and desktop
## t = 2.9883, df = 68.152, p-value = 0.003898
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.0004442751 0.0022298661
```

```
## sample estimates:
##   mean of x   mean of y
## 0.002811679 0.001474608
```

We see that the conversion rate for old users is 0.0028 while that for the new users is only 0.0014. Moreover, the p-value of less than 0.05 means that the results are not just by chance. Definitely, there is something wrong with the new Desktop users.

Let's try to investigate this further.

We can find out during which phase of the funnel the site is losing more users. To do this, we need to see how many users visiting the payment page also visited the payment confirmation page, how many users visiting the search page also visited the payment page and how many users who visited the home page also visited the search page.

Let's start from analyzing the number of users who visited payment confirmation page, given that they visited the payment page and visualize their results.

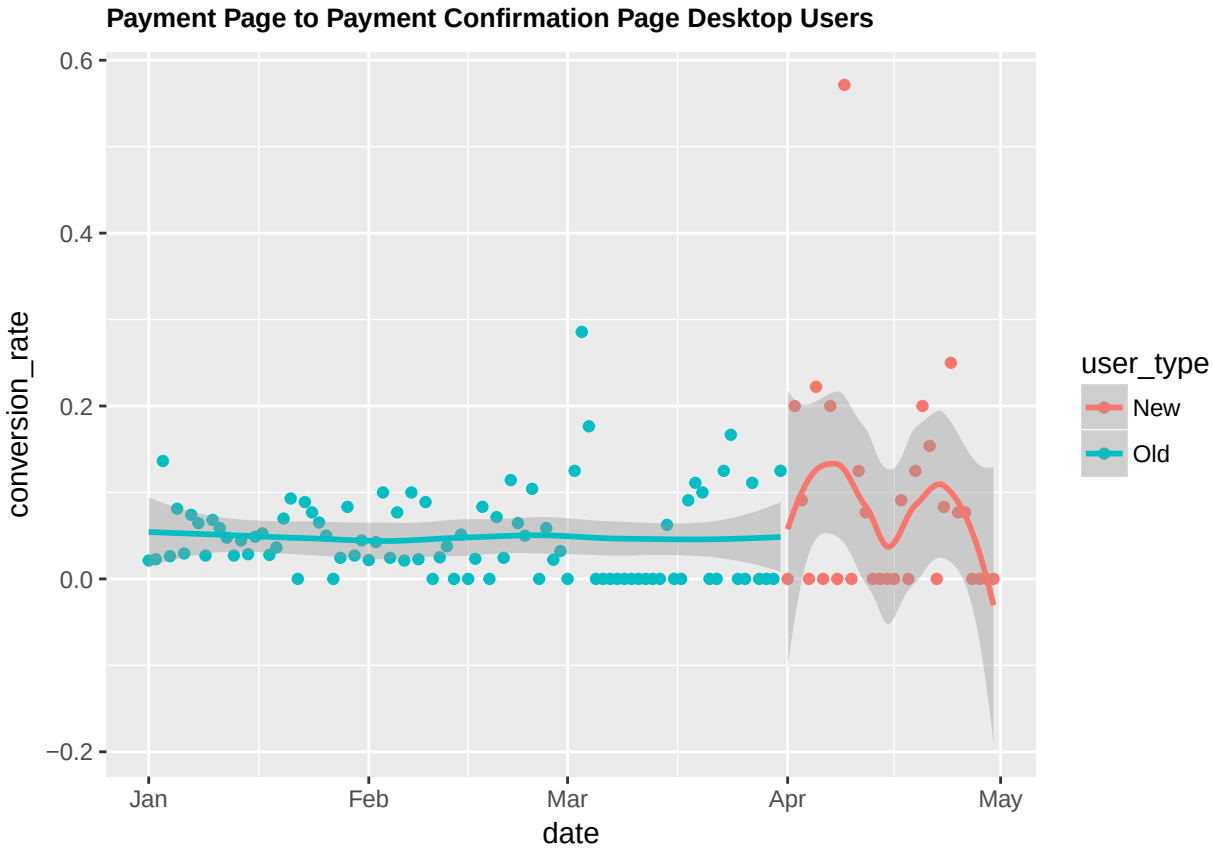
```
## Desktop users visiting payment page
desktop_payment_page_visitors <- desktop_users %>% filter(payment_page == 1)

## Payment page to payment_confirmation page
desktop_paymentpagevisitors_converted_by_date <- desktop_payment_page_visitors %>%
  group_by(date) %>%
  summarise(conversion_rate = mean(payment_confirmation_page))

## Creating a variable for new user
desktop_paymentpagevisitors_converted_by_date$user_type <- sapply(desktop_paymentpagevisitors_converted

## Visualizing conversion rate (Payment page to payment_confirmation page Desktop Users)
ggplot(desktop_paymentpagevisitors_converted_by_date,
       aes(date, conversion_rate, color = user_type)) +
       geom_point() +
       stat_smooth() +
       ggtitle("Payment Page to Payment Confirmation Page Desktop Users") +
       theme(plot.title = element_text(size = 10, face = "bold"))
```

```
## `geom_smooth()` using method = 'loess'
```



Again, visualizing these users gives strange results. The conversion rate seems to be almost constant for the old users while the new users seem to be behaving strangely. The reason for this also can be insufficient data for the new users.

Now, let's try to compare the conversion rates of these payment page visiting Desktop users using t-test.

```
## Comparing new users and old users (Payment page to payment_confirmation page)
t.test(desktop_paymentpagevisitors_converted_by_date$conversion_rate[desktop_paymentpagevisitors_conver
```

```
##
##  Welch Two Sample t-test
##
## data:  desktop_paymentpagevisitors_converted_by_date$conversion_rate[desktop_paymentpagevisitors_con
## t = -1.6098, df = 32.283, p-value = 0.1172
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.084141672  0.009840685
## sample estimates:
##  mean of x  mean of y
## 0.04763010 0.08478059
```

The results say that the conversion of new users is much more than the old users (almost 80% more). But the p-value of 0.1172 (>0.05) says that these results might just be by chance.

Now, let's analyze the number of users who visited payment page, given that they visited the search page and visualize their results.

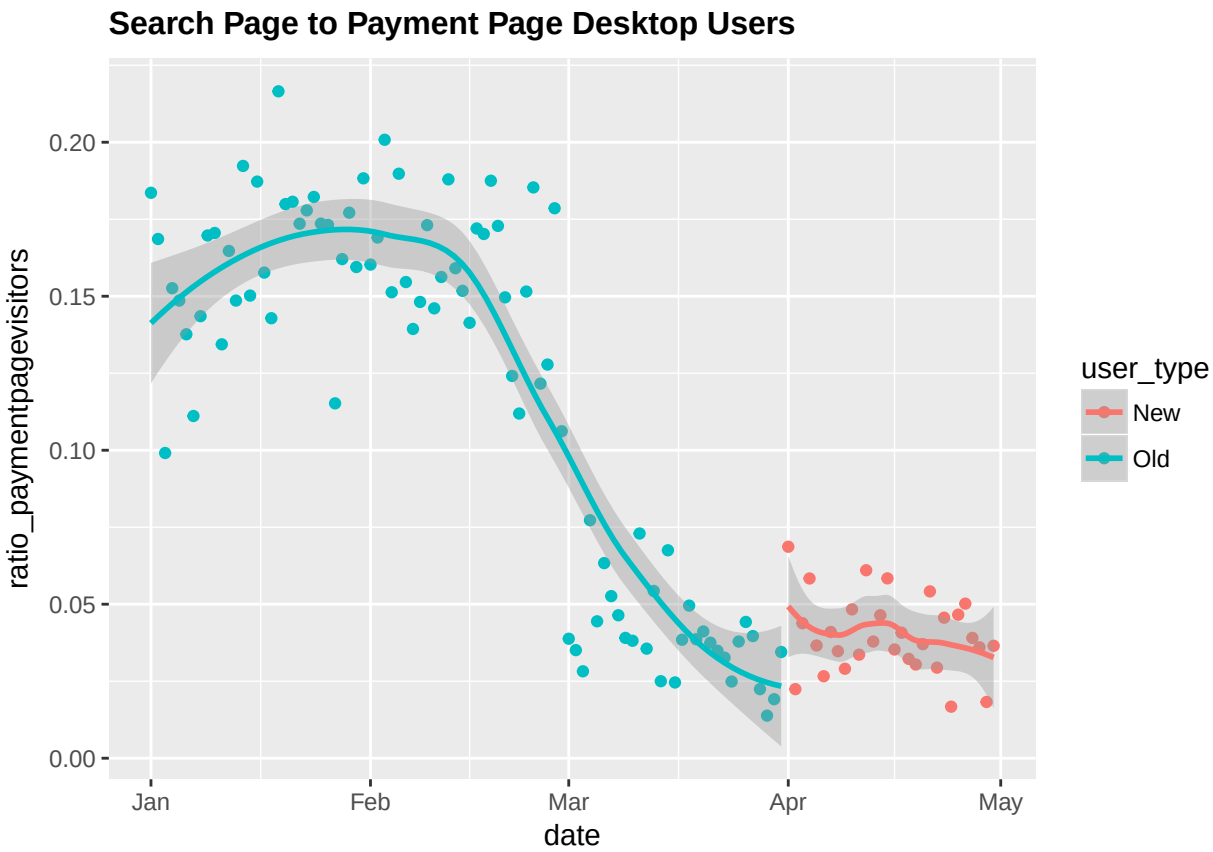
```
## Desktop users visiting search page
desktop_search_page_visitors <- desktop_users %>% filter(search_page == 1)

## Search page to Payment page Desktop users
desktop_searchpagevisitors_visitingpaymentpage_by_date <- desktop_search_page_visitors %>%
  group_by(date) %>%
  summarise(ratio_paymentpagevisitors = mean(payment_page))

## Creating a variable for new user
desktop_searchpagevisitors_visitingpaymentpage_by_date$user_type <- sapply(desktop_searchpagevisitors_v

## Visualizing Desktop users who visited payment page given they visited the search page
ggplot(desktop_searchpagevisitors_visitingpaymentpage_by_date,
       aes(date, ratio_paymentpagevisitors, color = user_type)) +
  geom_point() +
  stat_smooth() +
  ggtitle("Search Page to Payment Page Desktop Users") +
  theme(plot.title = element_text(size = 12, face = "bold"))
```

```
## `geom_smooth()` using method = 'loess'
```



From the graph, it seems that there is definitely something wrong with the search page as the number of users visiting the payment page after visiting the search page have decreased tremendously since mid of February. The machine learning software engineers definitely need to work on showing better search results to the users in order to prevent this in future.

Now, let's try to confirm these results by using t-test.

```
## Comparing new users and old users (Search Page to Payment Page)
t.test(desktop_searchpagevisitors_visitingpaymentpage_by_date$ratio_paymentpagevisitors[desktop_searchpa
       desktop_searchpagevisitors_visitingpaymentpage_by_date$ratio_paymentpagevisitors[desktop_searchpa
```

```
##
##  Welch Two Sample t-test
##
## data:  desktop_searchpagevisitors_visitingpaymentpage_by_date$ratio_paymentpagevisitors[desktop_searc
## t = 11.544, df = 107.45, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.06513870 0.09214776
## sample estimates:
##  mean of x  mean of y
## 0.11848400 0.03984077
```

The t-test shows the opposite results but this might just be due to insufficient data for new users (or the way we have chose to define the new users). The graph clearly shows that there is something wrong with the search page.

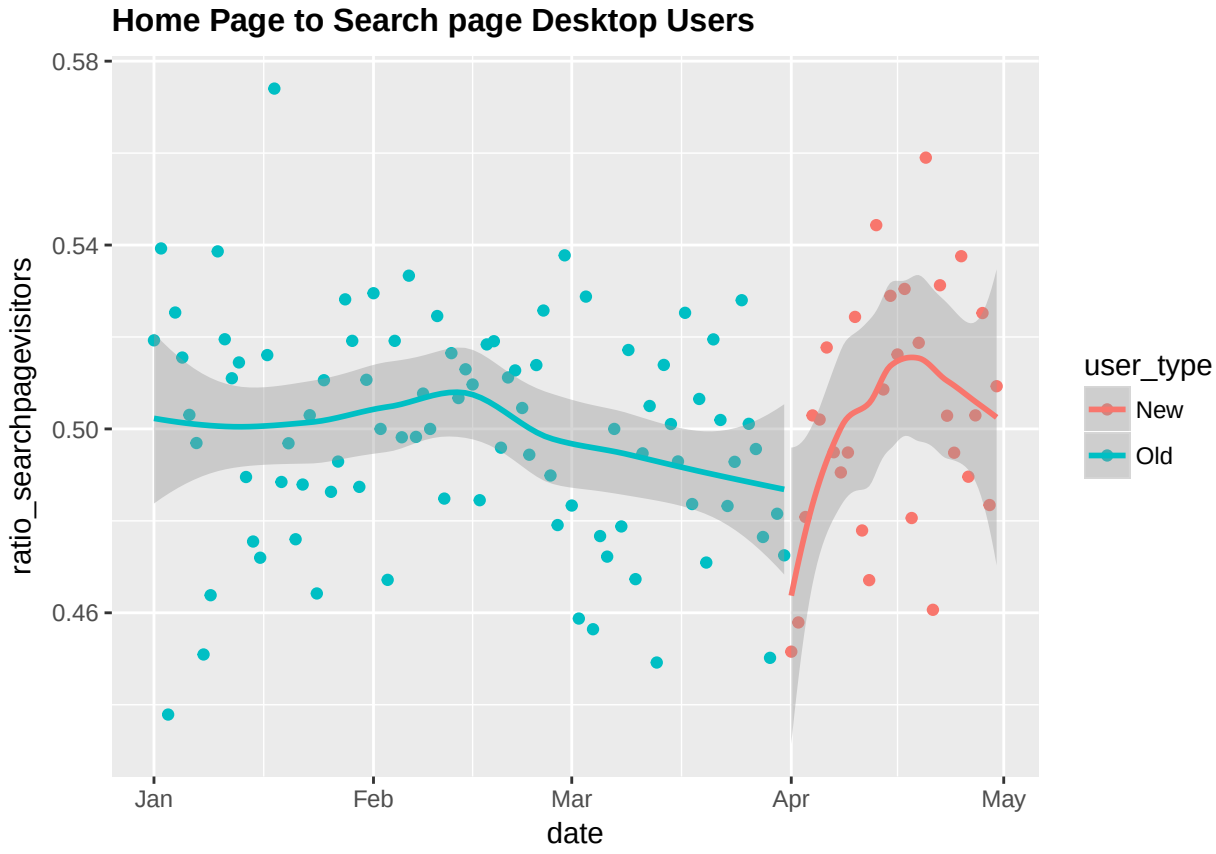
Now, let's try to analyze the users who visited the search page, given that they visited the home page and try to visualize them.

```
## Home page to Search page Desktop Users
home_to_searchpage_desktop_users <- desktop_users %>%
  group_by(date) %>%
  summarise(ratio_searchpagevisitors = mean(search_page))

## Creating a variable for new user
home_to_searchpage_desktop_users$user_type <- sapply(home_to_searchpage_desktop_users$date, define_new_u

## Home Page to Search Page visiting Desktop Users
ggplot(home_to_searchpage_desktop_users,
       aes(date, ratio_searchpagevisitors, color = user_type)) +
  geom_point() +
  stat_smooth() +
  ggtitle("Home Page to Search page Desktop Users") +
  theme(plot.title = element_text(size = 12, face = "bold"))
```

```
## `geom_smooth()` using method = 'loess'
```



We see that the number of users visiting the search page has increased for the new users. So the UI team is definitely doing well.

Let's confirm these results by using t-test.

```
## Comparing new users and old users (Home page to Search page)
t.test(home_to_searchpage_desktop_users$ratio_searchpagevisitors[home_to_searchpage_desktop_users$user_
       home_to_searchpage_desktop_users$ratio_searchpagevisitors[home_to_searchpage_desktop_users$user_
```

```
##
##  Welch Two Sample t-test
##
## data:  home_to_searchpage_desktop_users$ratio_searchpagevisitors[home_to_searchpage_desktop_users$use
## t = -0.74764, df = 45.877, p-value = 0.4585
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.015017446  0.006883455
## sample estimates:
## mean of x mean of y
## 0.4988412 0.5029082
```

Even the t-test confirm our result. But again the p-value of 0.45 (>0.05) says that our results might just be by chance.

Overall, I would say that the site is losing most of the Desktop users at the Search Page. The search results shown to the users need to be improved.

Mobile Users

Now, let's use the similar approach for the mobile users.

First, let's subset all the mobile users from the data.

```
## Subsetting mobile users
mobile_users <- data %>% filter(device == "Mobile")
```

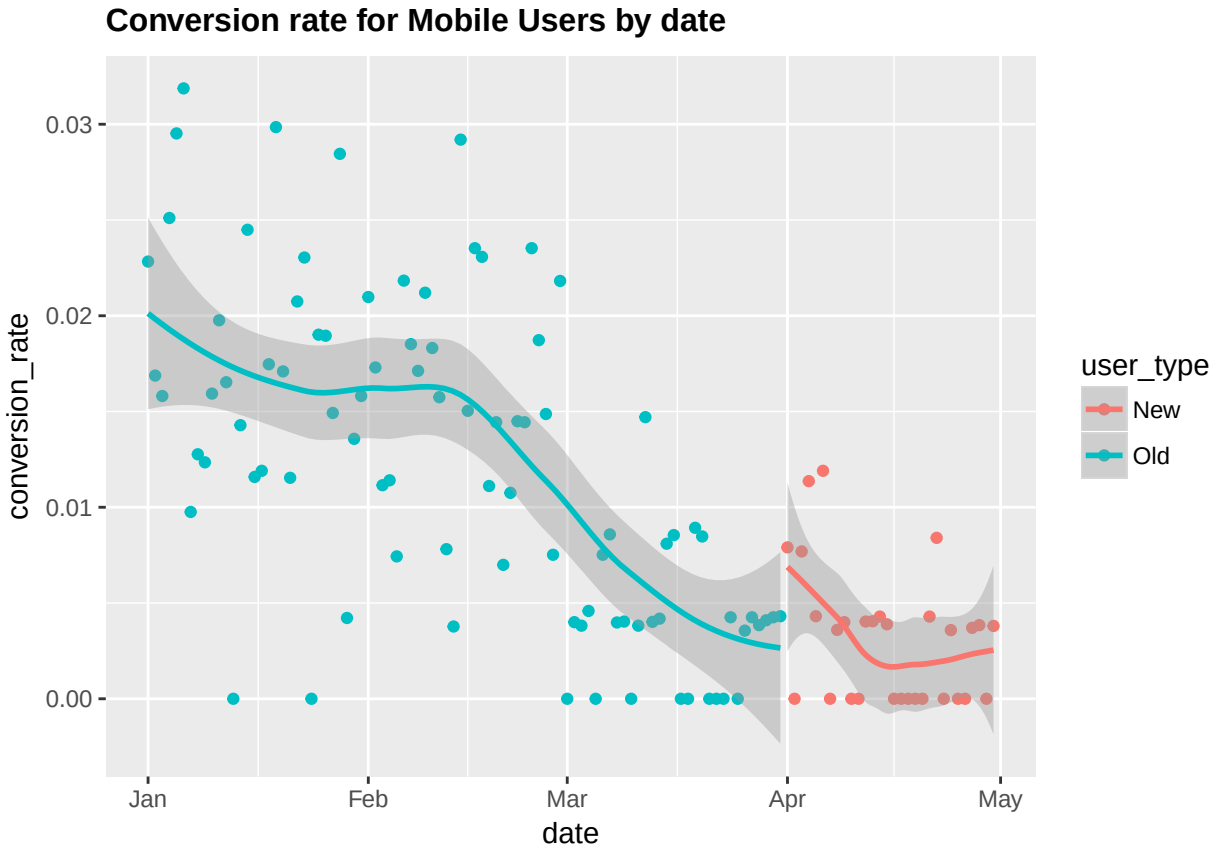
Now, let's try to visualize the overall conversion rate for the mobile users.

```
## Conversion rate for Mobile users by date
mobile_converted_by_date <- mobile_users %>%
  group_by(date) %>%
  summarise(conversion_rate = mean(payment_confirmation_page))

## Creating a variables for new user
mobile_converted_by_date$user_type <- sapply(mobile_converted_by_date$date, define_new_user)

## Visualizing conversion rate by date
ggplot(mobile_converted_by_date,
       aes(date, conversion_rate, color = user_type)) +
  geom_point() +
  stat_smooth() +
  ggtitle("Conversion rate for Mobile Users by date") +
  theme(plot.title = element_text(size = 12, face = "bold"))
```

```
## `geom_smooth()` using method = 'loess'
```



The results show that the conversion rate for the mobile users has decreased almost continuously and has got worse.

Let's try to verify these results using t-test.

```
## Applying t-test to compare the conversion rate for old users and the new users
t.test(mobile_converted_by_date$conversion_rate[mobile_converted_by_date$user_type == "Old"], mobile_co
```

```
##
##  Welch Two Sample t-test
##
## data:  mobile_converted_by_date$conversion_rate[mobile_converted_by_date$user_type ==  and mobile_co
## t = 8.2183, df = 112.86, p-value = 3.866e-13
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.00683094 0.01117061
## sample estimates:
##   mean of x   mean of y
## 0.012156394 0.003155621
```

The conversion rate for the old users is 0.012 while that of new users is 0.003. Both these values of conversion rates are much higher than that of the corresponding group's Desktop users. Maybe the mobile app developers are doing a good job at the company!

Let's try to investigate this further.

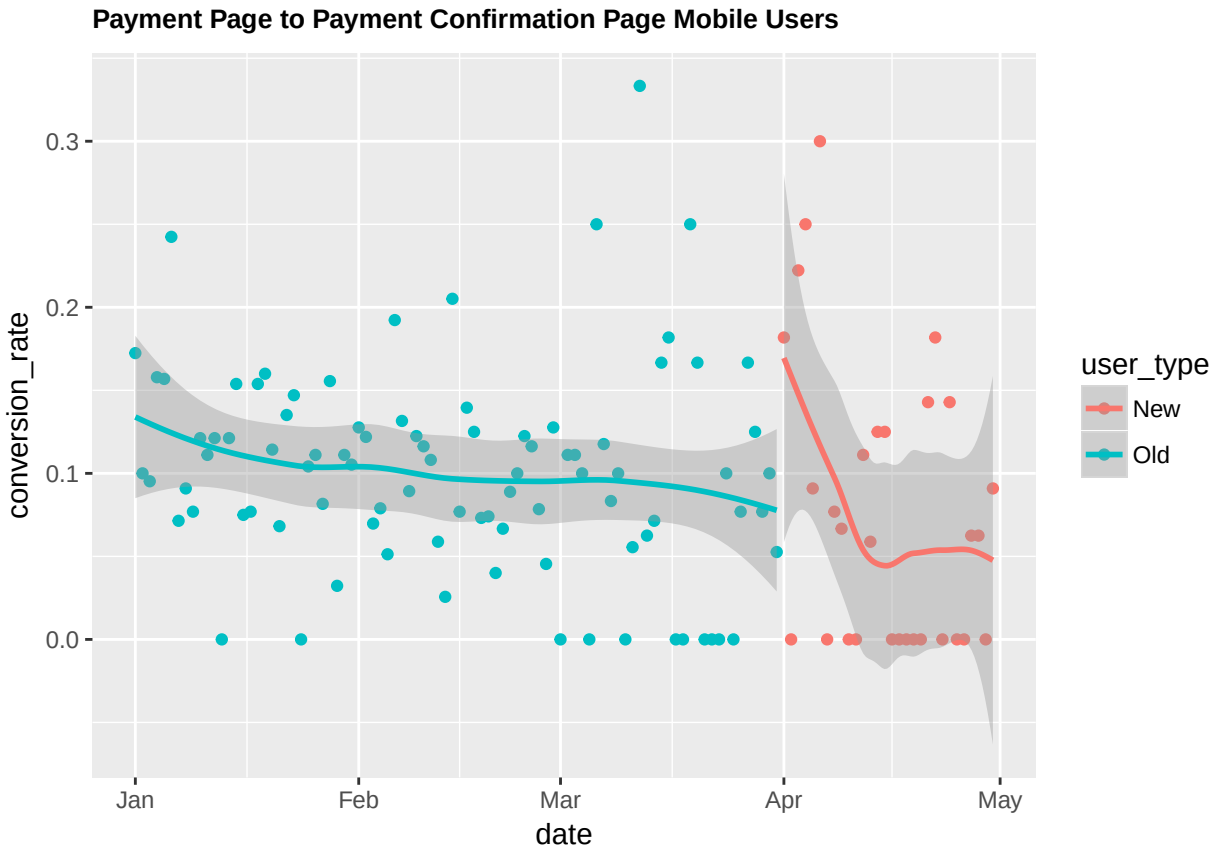
```
## Mobile users visiting payment page
mobile_payment_page_visitors <- mobile_users %>% filter(payment_page == 1)

## Payment page to payment_confirmation page
mobile_paymentpagevisitors_converted_by_date <- mobile_payment_page_visitors %>%
  group_by(date) %>%
  summarise(conversion_rate = mean(payment_confirmation_page))

## Creating a variable for new user
mobile_paymentpagevisitors_converted_by_date$user_type <- sapply(mobile_paymentpagevisitors_converted_by

## Visualizing conversion rate (Payment page to payment_confirmation page Mobile Users)
ggplot(mobile_paymentpagevisitors_converted_by_date,
       aes(date, conversion_rate, color = user_type)) +
  geom_point() +
  stat_smooth() +
  ggtitle("Payment Page to Payment Confirmation Page Mobile Users") +
  theme(plot.title = element_text(size = 10, face = "bold"))
```

```
## `geom_smooth()` using method = 'loess'
```




```
## Comparing new users and old users (Payment page to payment_confirmation page)
t.test(mobile_paymentpagevisitors_converted_by_date$conversion_rate[mobile_paymentpagevisitors_converte
```

```
##
```

```
##  Welch Two Sample t-test
##
## data:  mobile_paymentpagevisitors_converted_by_date$conversion_rate[mobile_paymentpagevisitors_conve
## t = 1.4217, df = 39.561, p-value = 0.163
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.01023182  0.05871021
## sample estimates:
##  mean of x  mean of y
## 0.10063638 0.07639718
```

The t-test says that the new users have a bit lower conversion rate after they visited the payment page. But the p-value of 0.163 (>0.05) indicates that these results might just be by chance.

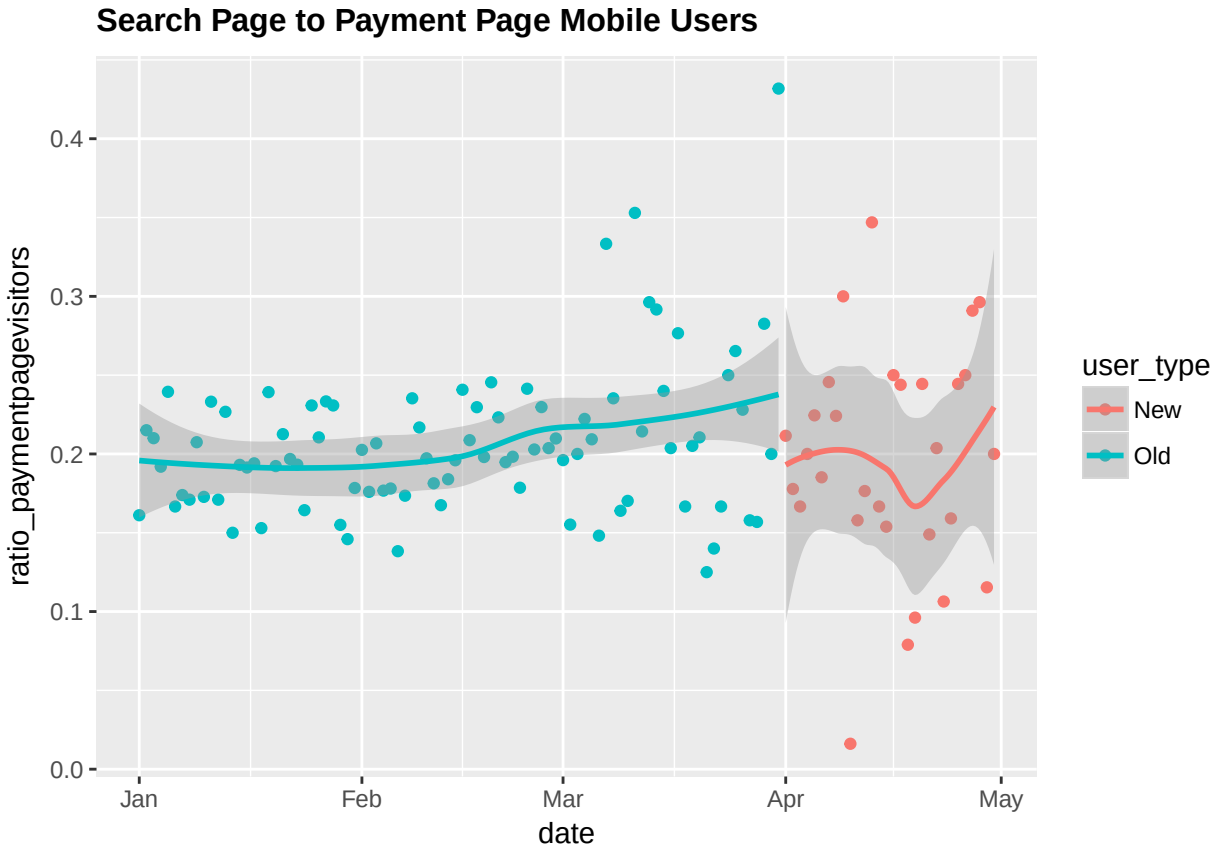
```
## Mobile users visiting search page
mobile_search_page_visitors <- mobile_users %>% filter(search_page == 1)

## Search page to Payment page Mobile users
mobile_searchpagevisitors_visitingpaymentpage_by_date <- mobile_search_page_visitors %>%
  group_by(date) %>%
  summarise(ratio_paymentpagevisitors = mean(payment_page))

## Creating a variable for new user
mobile_searchpagevisitors_visitingpaymentpage_by_date$user_type <- sapply(mobile_searchpagevisitors_visi

## Visualizing Mobile users who visited payment page given they visited the search page
ggplot(mobile_searchpagevisitors_visitingpaymentpage_by_date,
       aes(date, ratio_paymentpagevisitors, color = user_type)) +
  geom_point() +
  stat_smooth() +
  ggtitle("Search Page to Payment Page Mobile Users") +
  theme(plot.title = element_text(size = 12, face = "bold"))
```

```
## `geom_smooth()` using method = 'loess'
```




```
## Comparing new users and old users (Search Page to Payment Page)
t.test(mobile_searchpagevisitors_visitingpaymentpage_by_date$ratio_paymentpagevisitors[mobile_searchpag
       mobile_searchpagevisitors_visitingpaymentpage_by_date$ratio_paymentpagevisitors[mobile_searchpag
```

```
##
##  Welch Two Sample t-test
##
## data:  mobile_searchpagevisitors_visitingpaymentpage_by_date$ratio_paymentpagevisitors[mobile_searchp
## t = 0.70423, df = 37.728, p-value = 0.4856
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.01853865  0.03830984
## sample estimates:
## mean of x mean of y
## 0.2059507 0.1960651
```

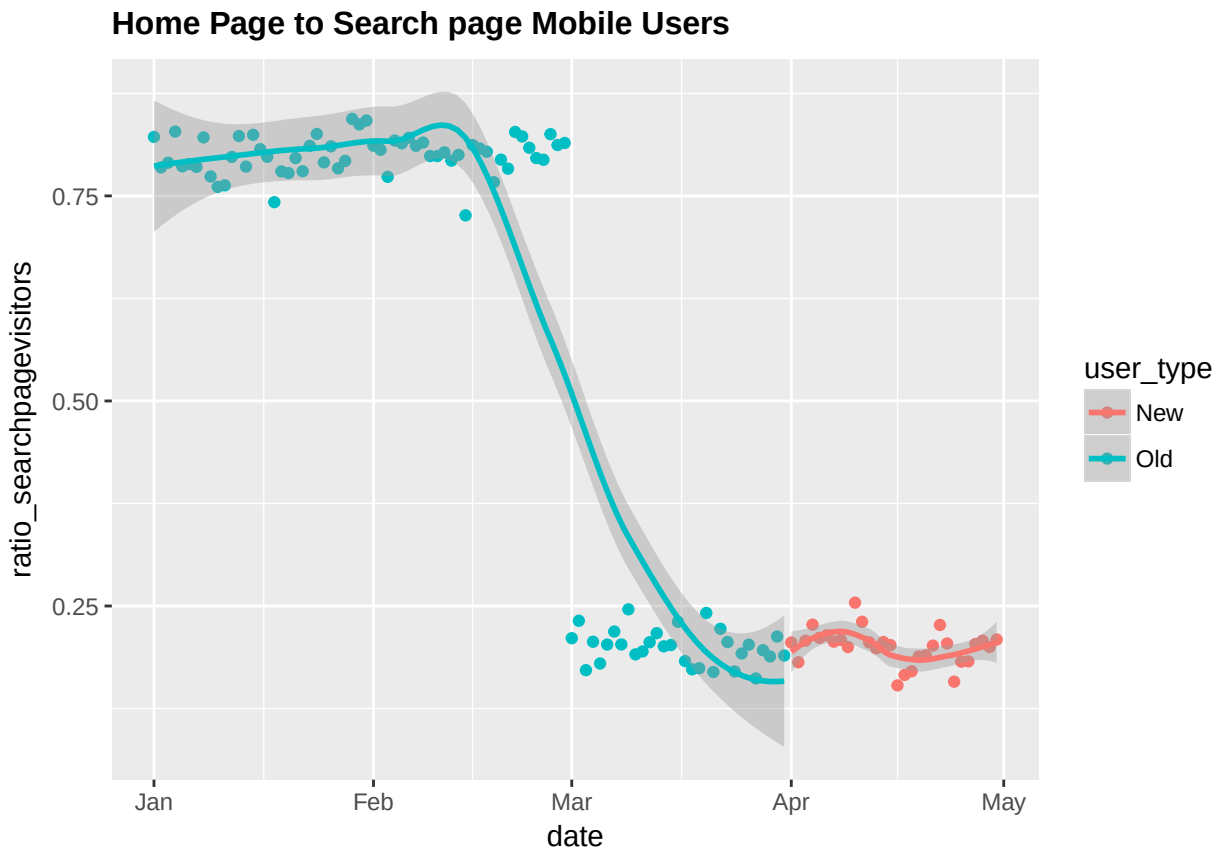
The plot as well as the t-test results indicate less percentage of users visiting payment page after they have visited the search page.

```
## Home page to Search page Mobile Users
home_to_searchpage_mobile_users <- mobile_users %>%
  group_by(date) %>%
  summarise(ratio_searchpagevisitors = mean(search_page))

## Creating a variable for new user
home_to_searchpage_mobile_users$user_type <- sapply(home_to_searchpage_mobile_users$date, define_new_us
```

```
## Home Page to Search Page visiting Mobile Users
ggplot(home_to_searchpage_mobile_users,
       aes(date, ratio_searchpagevisitors, color = user_type)) +
  geom_point() +
  stat_smooth() +
  ggtitle("Home Page to Search page Mobile Users") +
  theme(plot.title = element_text(size = 12, face = "bold"))
```

```
## `geom_smooth()` using method = 'loess'
```




```
t.test(home_to_searchpage_mobile_users$ratio_searchpagevisitors[home_to_searchpage_mobile_users$user_ty
       home_to_searchpage_mobile_users$ratio_searchpagevisitors[home_to_searchpage_mobile_users$user_ty
```

```
##
##  Welch Two Sample t-test
##
## data:  home_to_searchpage_mobile_users$ratio_searchpagevisitors[home_to_searchpage_mobile_users$user
## t = 12.862, df = 91.915, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.3325771 0.4540409
## sample estimates:
## mean of x mean of y
## 0.5934283 0.2001193
```

The plot and the t-test clearly indicates that most of the users are not going further after visiting the home page. There is a steep drop seen in the graph. Moreover, the t-test also gives a p-value of less than 0.05 which indicates that these results are not just by chance. So the home page needs a lot of work for the mobile users.

Thus, for the Mobile Users, I would say that the funnel is losing most of its users on the home page. While for the Desktop Users, it was the search page.