Solution: Funnel_Analysis

Ismat Ara

Description

You are looking at data from an e-commerce website. The site is very simple and has just 4 pages:

- The first page is the home page. When you come to the site for the first time, you can only land on the home page as a first page.
- From the home page, the user can perform a search and land on the search page.
- From the search page, if the user clicks on a product, she will get to the payment page,where she is asked to provide payment information in order to buy that product.
- If she does decide to buy, she ends up on the confirmation page.

The company CEO isn't very happy with the volume of sales and, especially, of sales coming from new users. Therefore, she asked you to investigate whether there is something wrong in the conversion funnel or, in general, if you could suggest how conversion rate can be improved.

Specifically, she is interested in :

- A full picture of funnel conversion rate for both desktop and mobile.
- Some insights on what the product team should focus on in order to improve conversion rate as well as anything you might discover that could help improve conversion rate.

Data Analysis

Let's import the datasets first:

```
library(data.table)
users <- fread('user_table.csv')
home <-fread('home_page_table.csv')
search <- fread('search_page_table.csv')
payment <- fread('payment_page_table.csv')
confirmation <- fread('payment_confirmation_table.csv')
```

Checking the users dataset

```
head(users)
```

```
##    user_id       date  device    sex
## 1:  450007 2015-02-28 Desktop Female
## 2:  756838 2015-01-13 Desktop   Male
## 3:  568983 2015-04-09 Desktop   Male
## 4:  190794 2015-02-18 Desktop Female
## 5:  537909 2015-01-15 Desktop   Male
## 6:  993454 2015-03-03 Desktop   Male
```

The users table has 90,400 unique observations and four variables.

Let's check other tables as well:

```
head(home,3)
```

```
##    user_id      page
## 1:  313593 home_page
## 2:  468315 home_page
## 3:  264005 home_page
```

```
head(search,3)
```

```
##    user_id        page
## 1:   15866 search_page
## 2:  347058 search_page
## 3:  577020 search_page
```

```
head(payment,3)
```

```
##    user_id         page
## 1:  253019 payment_page
## 2:  310478 payment_page
## 3:  304081 payment_page
```

```
head(confirmation,3)
```

```
##    user_id                      page
## 1:  123100 payment_confirmation_page
## 2:  704999 payment_confirmation_page
## 3:  407188 payment_confirmation_page
```

Changing the column name of each pages which will help to indetify them in combined table

```
colnames(home) <- c('user_id','home_pg')
colnames(search) <- c('user_id','search_pg')
colnames(payment) <- c('user_id','payment_pg')
colnames(confirmation) <- c('user_id','converted')
```

Let's check 'confirmation' table again to see the column's name change

```
head(confirmation,3)
```

```
##    user_id                  converted
## 1:  123100 payment_confirmation_page
## 2:  704999 payment_confirmation_page
## 3:  407188 payment_confirmation_page
```

Now, let's check if there is any duplicate users in any table

```
length(unique(users$user_id))==length(users$user_id)
```

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

```
length(unique(home$user_id))==length(home$user_id)
```

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

```
length(unique(search$user_id))==length(search$user_id)
```

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

```
length(unique(payment$user_id))==length(payment$user_id)
```

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

```
length(unique(confirmation$user_id))==length(confirmation$user_id)
```

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

All tables gave TRUE value, that means there is no duplicate users in any table.

Now check everyone in one table is also in other tables

```
length(users$user_id)-length(home$user_id)
```

```
## [1] 0
```

Looks like all users in the users table also in the home page table.

```
length(home$user_id)-length(search$user_id)
```

```
## [1] 45200
```

The search page has 45,200 less users than the home page.

```
length(search$user_id)-length(payment$user_id)
```

```
## [1] 39170
```

The payment page has 39,170 less users than search page.

```
length(payment$user_id)-length(confirmation$user_id)
```

```
## [1] 5578
```

The payment confirmation page has 5,578 less users than payment page.

Now combining all the tables to create one dataset.

```
combined <- merge(users,home, by='user_id',all=TRUE)
combined <- merge(combined, search, by='user_id', all=TRUE)
combined <- merge(combined, payment, by='user_id', all=TRUE)
combined <- merge(combined, confirmation, by='user_id', all=TRUE)
```

Let’s check the structure of the combined data

```
str(combined)
```

```
## Classes 'data.table' and 'data.frame':   90400 obs. of  8 variables:
##  $ user_id   : int  17 28 37 38 55 72 112 136 139 158 ...
##  $ date      : chr  "2015-04-21" "2015-04-29" "2015-02-21" "2015-03-23" ...
##  $ device    : chr  "Desktop" "Desktop" "Mobile" "Mobile" ...
##  $ sex       : chr  "Male" "Male" "Male" "Female" ...
##  $ home_pg   : chr  "home_page" "home_page" "home_page" "home_page" ...
##  $ search_pg : chr  "search_page" NA "search_page" "search_page" ...
##  $ payment_pg: chr  NA NA NA "payment_page" ...
##  $ converted : chr  NA NA NA NA ...
##  - attr(*, ".internal.selfref")=<externalptr>
##  - attr(*, "sorted")= chr "user_id"
```

Now let’s make all the pages users visited as 1 (Yes) and if they didn’t visit then 0 (No).

```
combined$home_pg <- ifelse(combined$home_pg=='home_page',1,0)
combined$search_pg <- ifelse(combined$search_pg=='search_page',1,0)
combined$payment_pg <- ifelse(combined$payment_pg=='payment_page',1,0)
combined$converted <- ifelse(combined$converted=='payment_confirmation_page',1,0)
combined[is.na(combined)] <- 0
```

Checking the combined data:

```
head(combined)
```

```
##    user_id       date  device    sex home_pg search_pg payment_pg
## 1:      17 2015-04-21 Desktop   Male       1         1          0
## 2:      28 2015-04-29 Desktop   Male       1         0          0
## 3:      37 2015-02-21  Mobile   Male       1         1          0
## 4:      38 2015-03-23  Mobile Female       1         1          1
## 5:      55 2015-02-01 Desktop   Male       1         0          0
## 6:      72 2015-04-22 Desktop   Male       1         0          0
##    converted
## 1:         0
## 2:         0
## 3:         0
## 4:         0
## 5:         0
## 6:         0
```

Converting date as Date:

```
combined$date <- as.Date(combined$date)
```

Converting all variables except ‘user_id’ and ‘converted’ as factor:

```
combined$device <- as.factor(combined$device)
combined$sex <- as.factor(combined$sex)
combined$home_pg <- as.factor(combined$home_pg)
combined$search_pg <- as.factor(combined$search_pg)
combined$payment_pg <- as.factor(combined$payment_pg)
```

Checking the structure of the combined data again:

```
str(combined)
```

```
## Classes 'data.table' and 'data.frame':   90400 obs. of  8 variables:
##  $ user_id   : int  17 28 37 38 55 72 112 136 139 158 ...
##  $ date      : Date, format: "2015-04-21" "2015-04-29" ...
##  $ 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_pg   : Factor w/ 1 level "1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ search_pg : Factor w/ 2 levels "0","1": 2 1 2 2 1 1 1 1 1 1 ...
##  $ payment_pg: Factor w/ 2 levels "0","1": 1 1 1 2 1 1 1 1 1 1 ...
##  $ converted : num  0 0 0 0 0 0 0 0 0 0 ...
##  - attr(*, ".internal.selfref")=<externalptr>
##  - attr(*, "sorted")= chr "user_id"
```

Let’s check if there is any missing data:

```
colSums(is.na(combined))
```

```
##    user_id       date     device        sex    home_pg  search_pg 
##          0          0          0          0          0          0 
## payment_pg  converted 
##          0          0
```

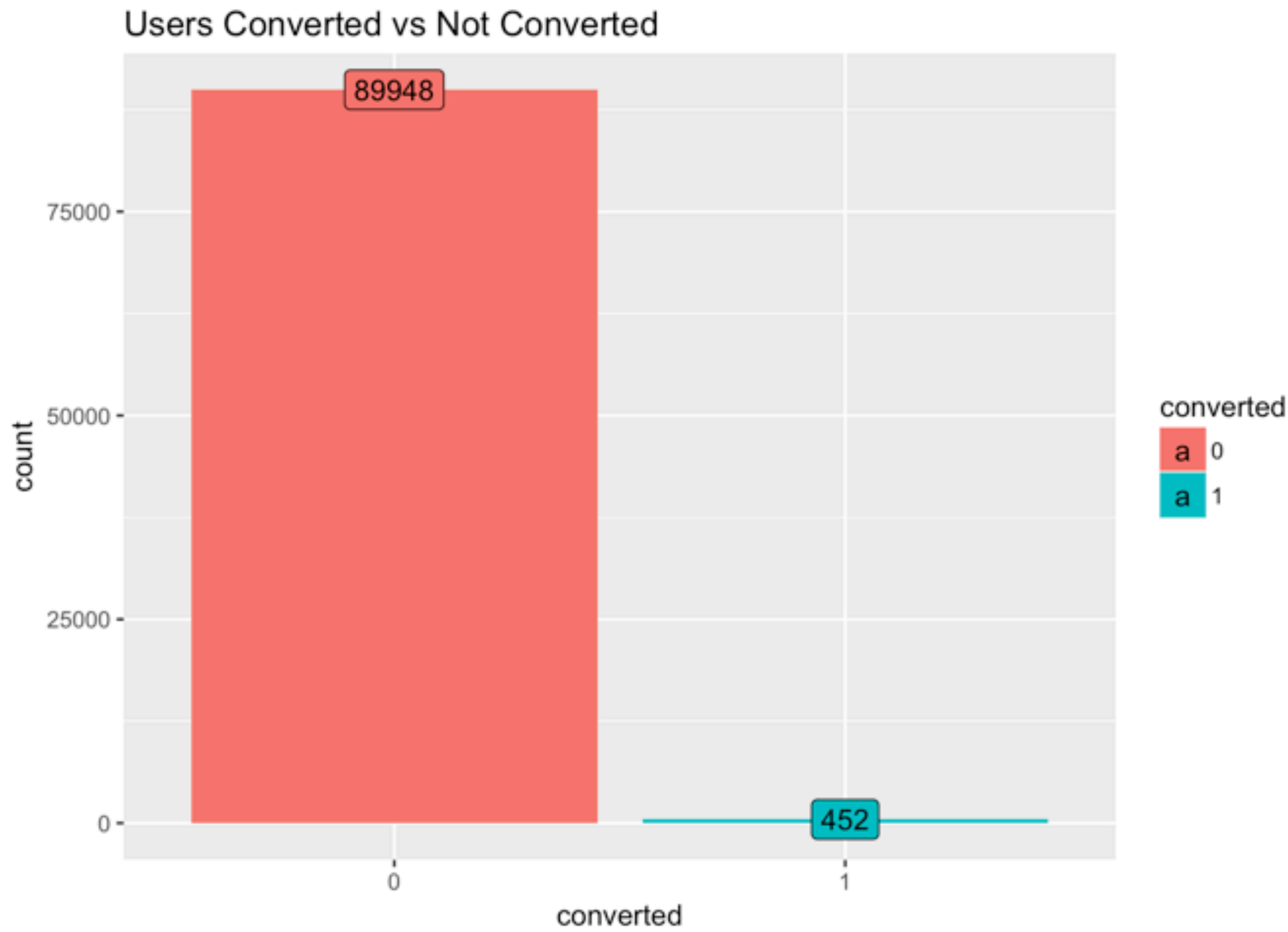
Looks good. There is no NAs in the data.

Data Visualization

Now let's visualize the data to see the general overview.

First let's see how many users visited the site and how may actually converted:

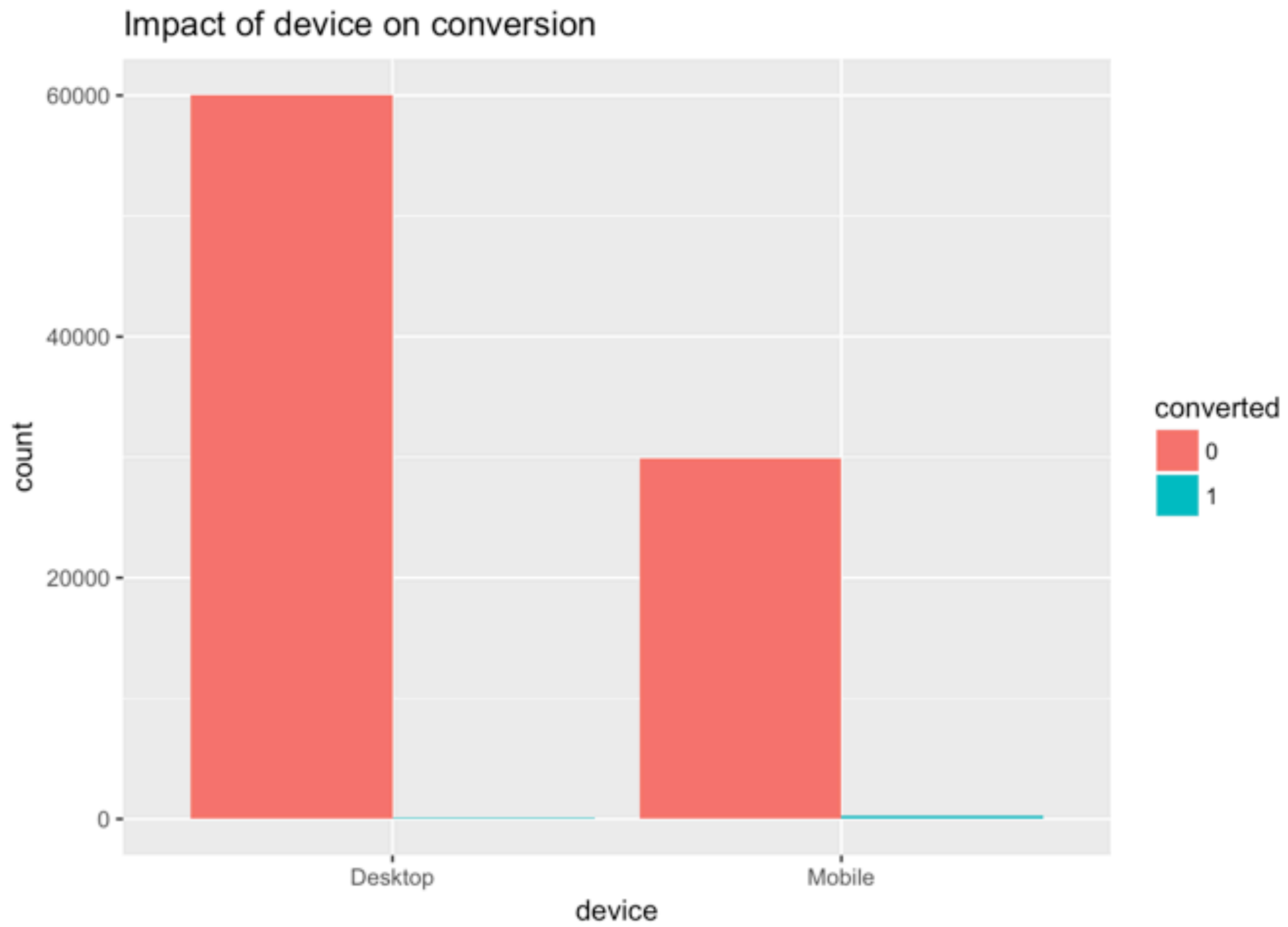
```
library(ggplot2)
library(gridExtra)
ggplot(combined, aes(x=factor(converted), fill=factor(converted)))+
  geom_bar()+
  labs(title='Users Converted vs Not Converted')+
  xlab('converted')+
  scale_fill_discrete(name='converted')+
  geom_label(stat='count',aes(label=..count..))
```



Even though 90,400 users visited the site but only 452 users converted which is only 0.5% of the total users.

Let’s see how device affect on conversion:

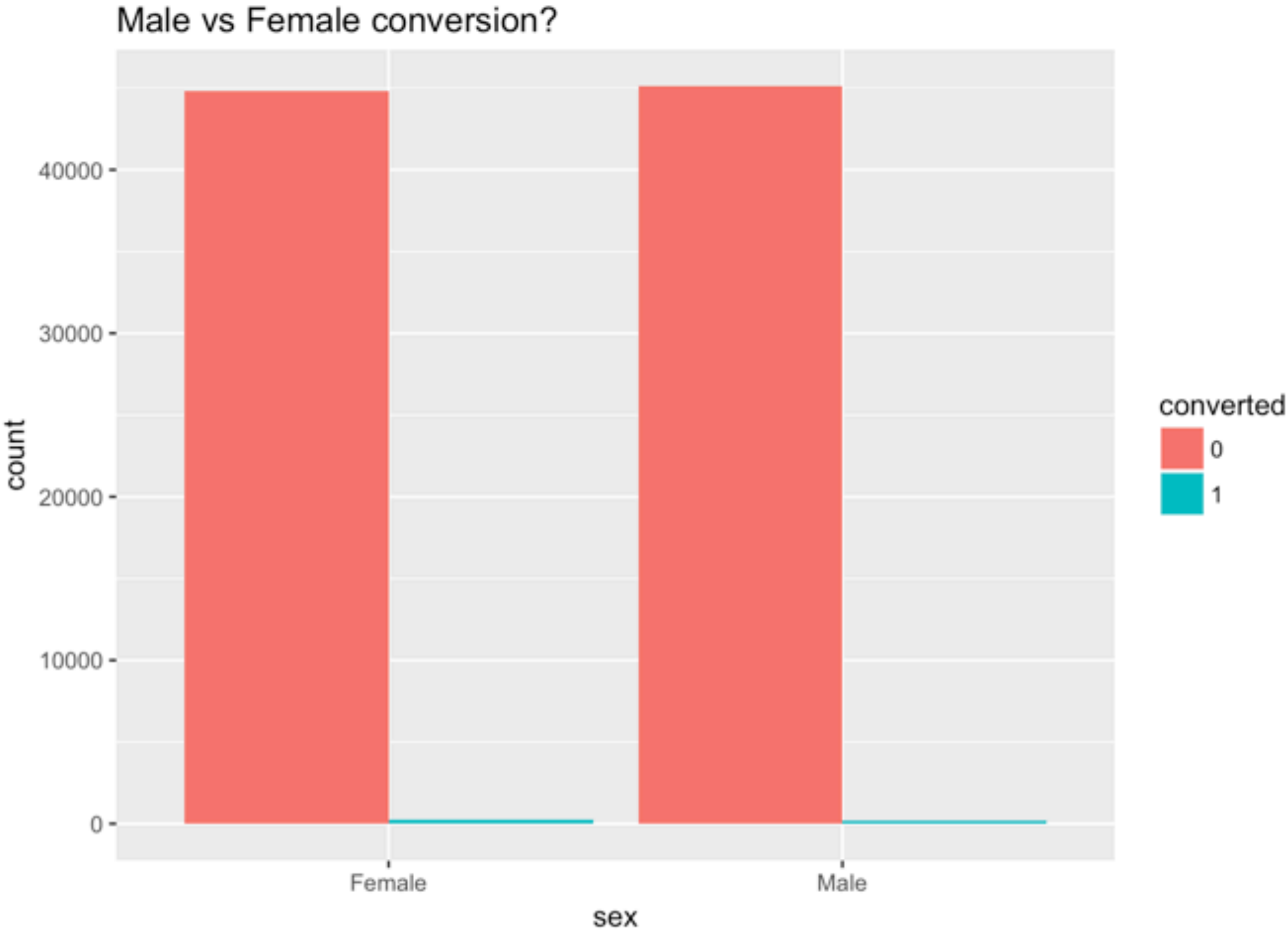
```
ggplot(combined, aes(x=device, fill=factor(converted)))+
  geom_bar(position='dodge')+
  labs(title='Impact of device on conversion')+
  scale_fill_discrete(name='converted')
```



This plot clearly shows that most people visited from desktop. The people who visited from mobile was half of the people of desktop but conversion from mobile was better than that of desktop. Later,we will take a deeper look at this.

Now check if gender is a factor for conversion:

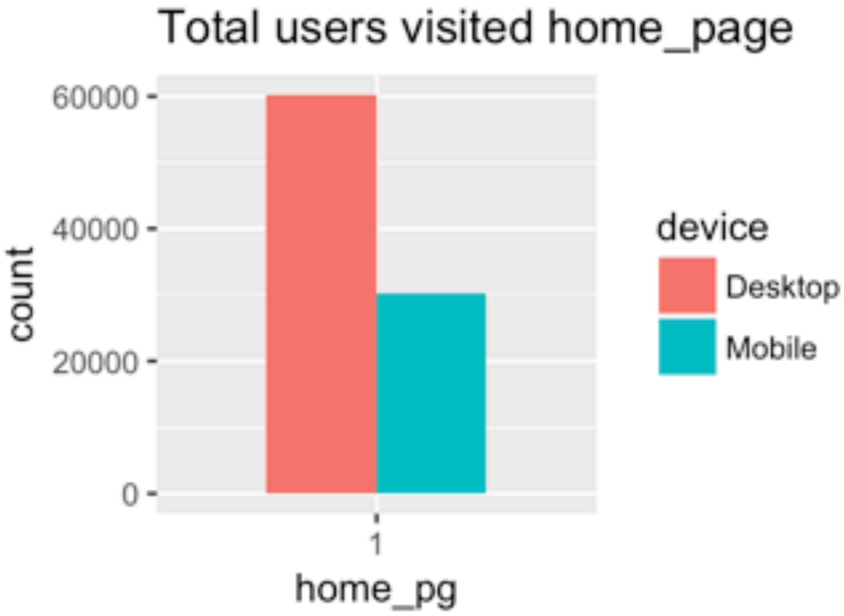
```
ggplot(combined, aes(x=sex, fill=factor(converted)))+
  geom_bar(position='dodge')+
  labs(title='Male vs Female conversion?')+
  scale_fill_discrete(name='converted')
```

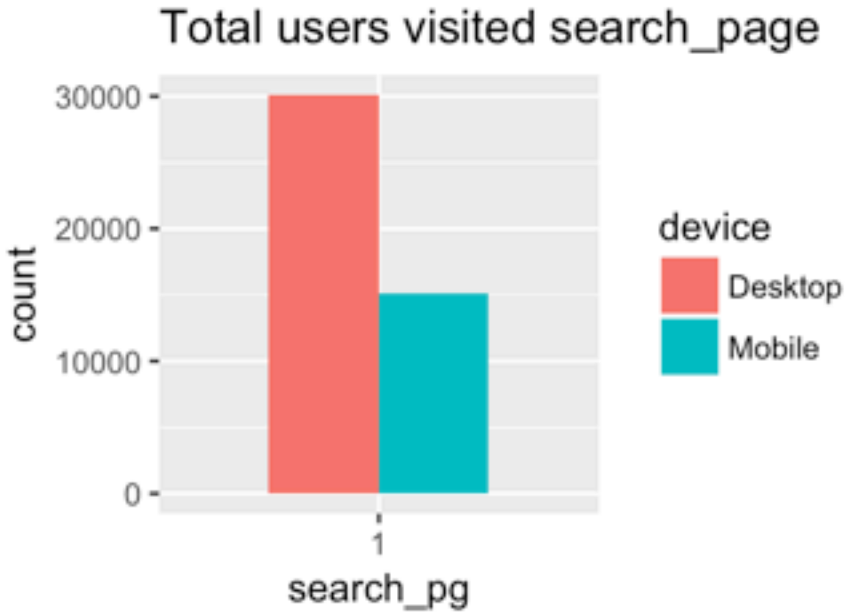


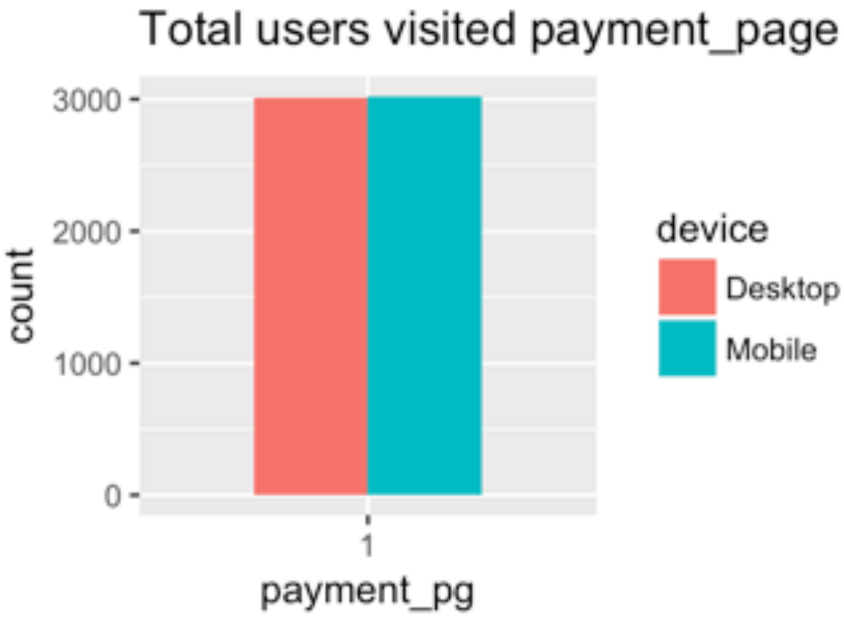
The people who visited the site, 50% was male and 50% was female and the conversion for both groups was pretty similar.

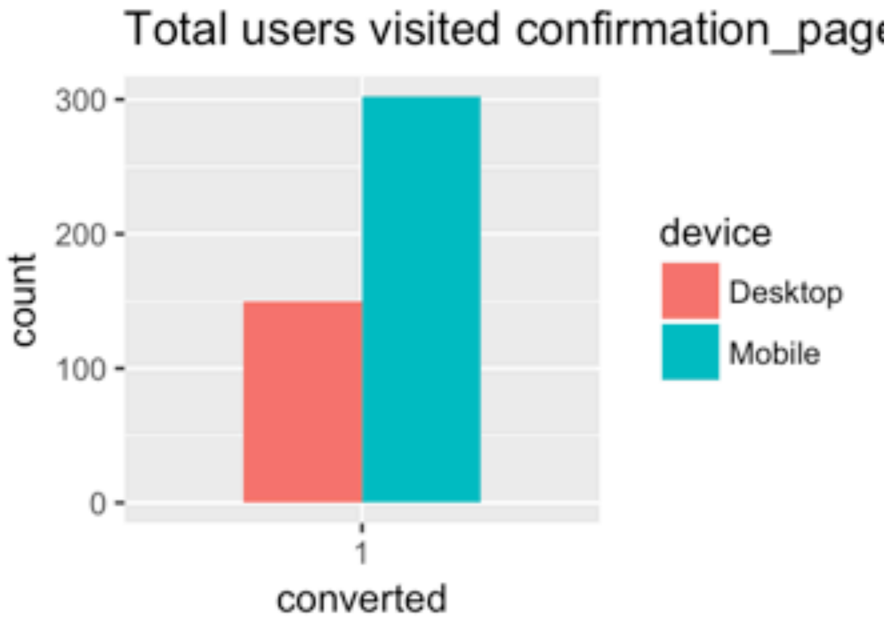
Let's take a deeper look at the data. The users who converted, visited all the pages like home_page first, then search_page and payment_page and ended up converting at confirmation_page. Now we will check how many people actually hit each pages.

```
p1 <- ggplot(combined, aes(x=home_pg, fill=device))+
  geom_bar(position='dodge',width=0.5)+ labs(title='Total users visited home_page')

p2 <- ggplot(combined[combined$search_pg==1], aes(x=search_pg, fill=device))+
  geom_bar(position='dodge',width=0.5)+ labs(title='Total users visited search_page')

p3 <- ggplot(combined[combined$payment_pg==1], aes(x=payment_pg, fill=device))+
  geom_bar(position='dodge',width=0.5)+ labs(title='Total users visited payment_page'
)

p4 <- ggplot(combined[combined$converted==1], aes(x=factor(converted), fill=device))+
  geom_bar(position='dodge',width=0.5)+ xlab('converted')+
  labs(title='Total users visited confirmation_page')

grid.arrange(p1,p2,p3,p4, ncol=2)
```







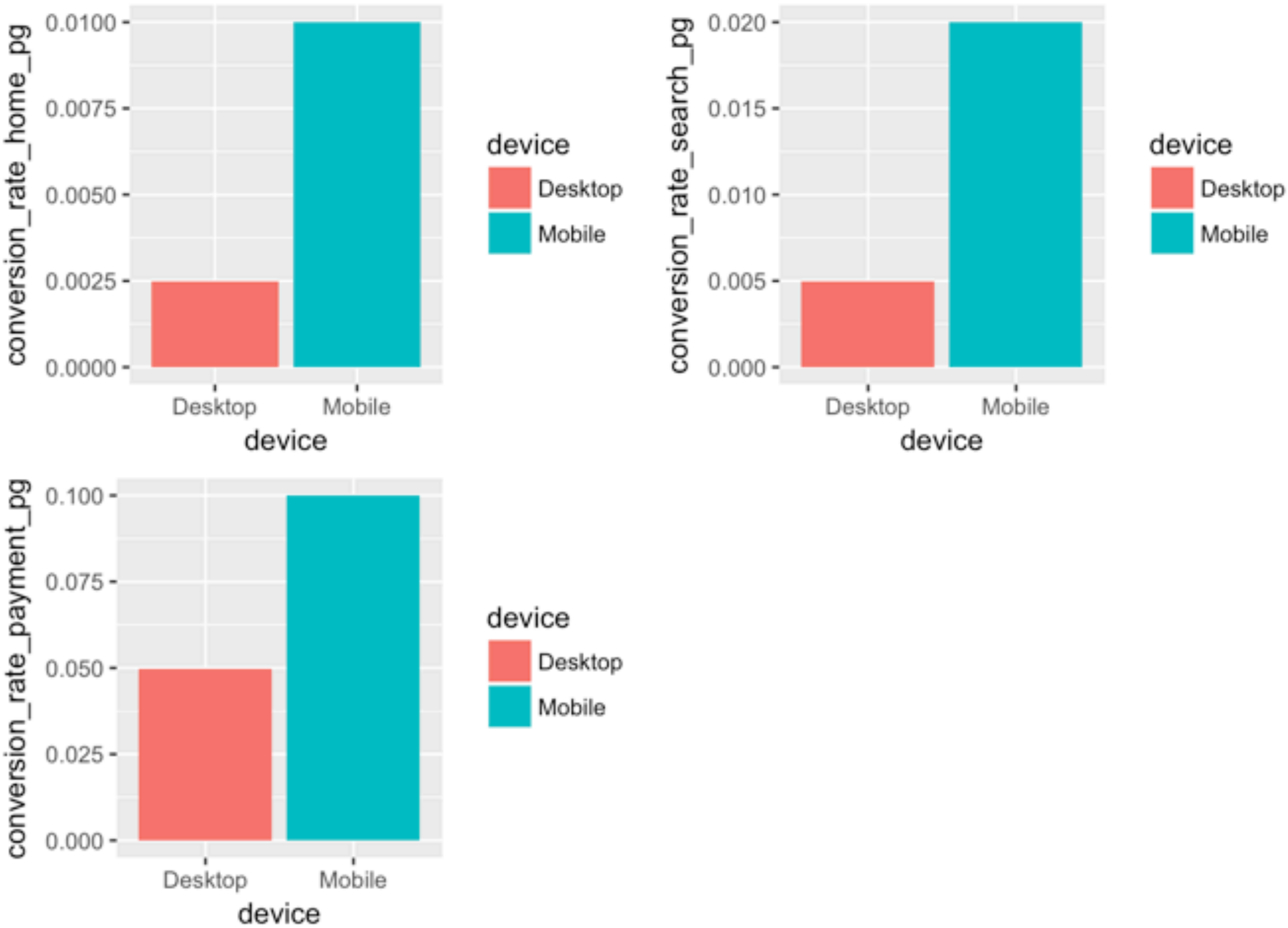


Total 90,400 users visited the home page and two-third of these users came from desktop users and one-third came from mobile users. 50% of the users who visited the home page also visited the search page and the other 50% left just after visiting the home page. the ratio of desktop and mobile users is similar to that of home page. There is a huge drop from search page to payment page. Approximately 6,000 users visited the payment page and the ratio of desktop and mobile users is same. And finally, only 452 users converted in the confirmation page. Interestingly, the conversion from mobile users were higher than that of desktop users even though more visitors were from desktop.

The large number of desktop user but much lower conversion rate, the reason might be the site is spending a lot of money on ads on desktop but the ads are attracting the wrong people. And maybe the desktop users had bad experience with the site.

Now, let’s check the conversion based on device:

```
library(dplyr)
data_device <- combined %>%
               group_by(device) %>%
               summarise(conversion_rate_home_pg=mean(converted[home_pg==1]),
                         conversion_rate_search_pg=mean(converted[search_pg==1]),
                         conversion_rate_payment_pg=mean(converted[payment_pg==1]))

p5 <- ggplot(data=data_device,aes(x=device,y=conversion_rate_home_pg))+
      geom_bar(stat='identity',aes(fill=device))

p6 <- ggplot(data=data_device,aes(x=device,y=conversion_rate_search_pg))+
      geom_bar(stat='identity',aes(fill=device))

p7 <- ggplot(data=data_device,aes(x=device,y=conversion_rate_payment_pg))+
      geom_bar(stat='identity',aes(fill=device))

grid.arrange(p5,p6,p7, ncol=2)
```



This plot shows the conversion of desktop users and mobile users. The conversion is increasing from one page to another because the total number of visitor is decreasing from one page to another.

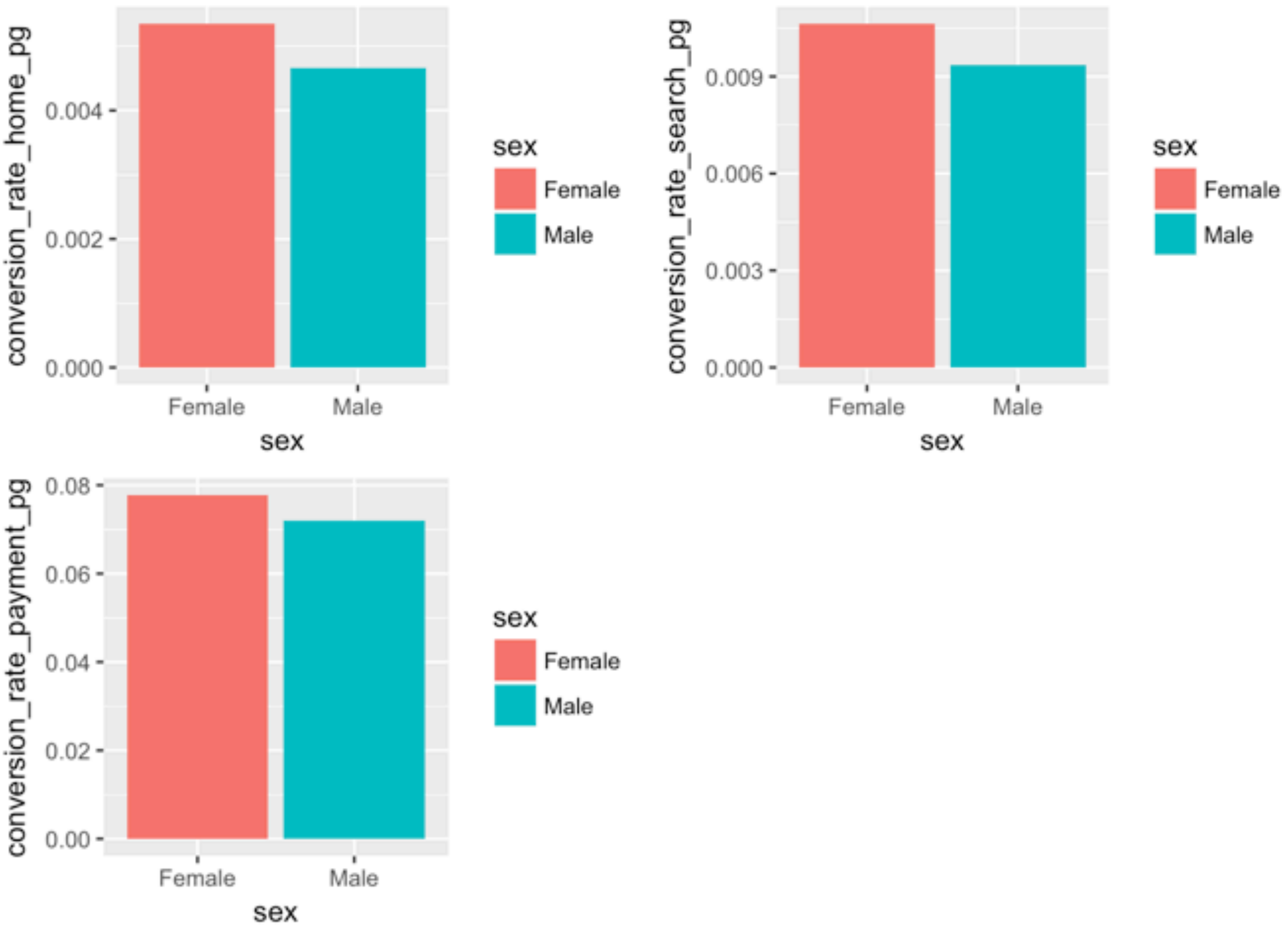
Now check how users are distributed between male and female users and how they converted.

```
p8 <- ggplot(combined, aes(x=home_pg, fill=sex))+
      geom_bar(position='dodge',width=0.5)+
      labs(title='Male vs Female visited home_pg')

p9 <- ggplot(combined[combined$search_pg==1], aes(x=search_pg, fill=sex))+
      geom_bar(position='dodge',width=0.5)+
      labs(title='Male vs Female visited search_pg')

p10 <- ggplot(combined[combined$payment_pg==1], aes(x=payment_pg, fill=sex))+
       geom_bar(position='dodge',width=0.5)+
       labs(title='Male vs Female visited payment_pg')

p11 <- ggplot(combined[combined$converted==1], aes(x=factor(converted), fill=sex))+
       geom_bar(position='dodge',width=0.5)+ xlab('converted')+
       labs(title='Male vs Female converted')

grid.arrange(p8,p9,p10,p11, ncol=2)
```

This plot shows male vs female visitors in different pages. The male and female visitors in home,search and payment pages are aproximately 50/50 but more female converted than male.

Check the conversion based on gender:

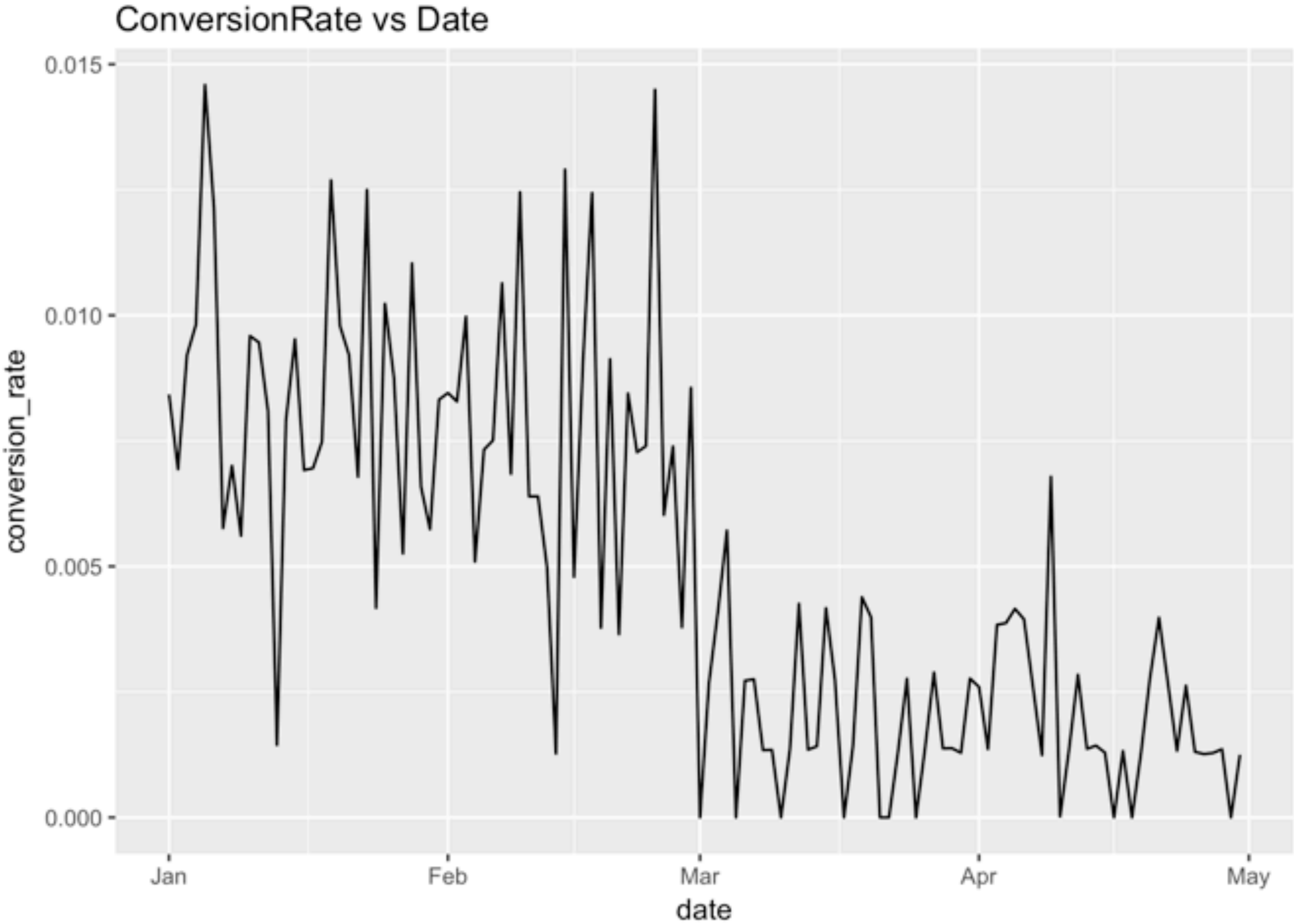
```
data_sex <- combined %>%
            group_by(sex) %>%
            summarise(conversion_rate_home_pg=mean(converted[home_pg==1]),
                      conversion_rate_search_pg=mean(converted[search_pg==1]),
                      conversion_rate_payment_pg=mean(converted[payment_pg==1]))

p12 <- ggplot(data=data_sex,aes(x=sex,y=conversion_rate_home_pg))+
       geom_bar(stat='identity',aes(fill=sex))

p13 <- ggplot(data=data_sex,aes(x=sex,y=conversion_rate_search_pg))+
       geom_bar(stat='identity',aes(fill=sex))

p14 <- ggplot(data=data_sex,aes(x=sex,y=conversion_rate_payment_pg))+
       geom_bar(stat='identity',aes(fill=sex))

grid.arrange(p12,p13,p14, ncol=2)
```



This plot clearly shows that female users converted more than male users at different pages.

Now check how the conversion varies over time:

```
data_date <- combined %>%
             group_by(date) %>%
             summarize(conversion_rate = mean(converted))

ggplot(data=data_date, aes(x=date, y=conversion_rate))+
  geom_line()+
  labs(title='ConversionRate vs Date')
```



Looks like the conversion in January and February is higher than in March and April. Seasonality might be the reason. There might also be some bug or a bad product change or competitor might be the reason.

Conclusion

- 90,400 users hit the home page but it dropped down to 50% to the search page. The reason might be the users didn't find the site as interesting. The product team can make the home page more attractive and informative.
- Another drop down from search page to payment page. Even though 50% of the initial users visited the search page, only approximately 7% of the initial users hit the payment page. The users might find that the site is not the right place what they are looking for or they might face some problem to search anything.
- Even though approximately 7% of the initial users hit the payment page but all of them didn't confirm

the payment. Only 0.5% of the initial users converted. The reason might be the user decided at the last moment not to buy the product or they might face any problem during payment. The product team should check the payment page and make it easier and secure for payment.

- Among 90,400 users, two-third of them use desktop and one-third use mobile but the conversion rate from mobile users was better. Marketing team should take a look if they are spending more money on desktop ads to attract wrong people as the conversion rate from desktop users is very low. On the other hand, marketing team should take action to increase the mobile users as conversion from mobile users is higher and it's a priority.
- The male and female users were pretty similar but female conversion was a bit higher than male conversion. Marketing team can try to attract more male users besides female users.