Solution: Engagement Test - A/B Test

Ismat Ara

Description

A social network company decided to add a new feature called ‘Recommended Friends’, i.e., they suggest people you may know.

A model has been built to show 5 people to each user. These potential friends will be shown on the user newsfeed. Company ran A/B test to check if the model is actually improving the engagement. At first, the model is tested just on a subset of users to see how it performs compared to the newsfeed without the new feature.

The test has been running for sometime and your boss asks you to check the results. You are asked to check, for each user, the number of pages visited during their first session since the test started. If the number increased, the test is a success.

The goal of this project is to look at A/B test result and draw conclusions.

Check A/B Test Results

Let’s import the dataset first:

```
user <- read.csv('user_table.csv')
test <- read.csv('test_table.csv')
```

Check the data:

```
head(user)
```

```
##   user_id signup_date
## 1      34  2015-01-01
## 2      59  2015-01-01
## 3     178  2015-01-01
## 4     285  2015-01-01
## 5     383  2015-01-01
## 6     397  2015-01-01
```

```
head(test)
```

```
##   user_id       date browser test pages_visited
## 1  600597 2015-08-13      IE    0             2
## 2 4410028 2015-08-26  Chrome    1             5
## 3 6004777 2015-08-17  Chrome    0             8
## 4 5990330 2015-08-27  Safari    0             8
## 5 3622310 2015-08-07 Firefox    0             1
## 6 1806423 2015-08-28      IE    0             5
```

Check if user_id is duplicate or not

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

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

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

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

```
length(user$user_id)-length(test$user_id)
```

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

Looks like all the user_id are unique and the number of users in both user and test datasets are equal.

Let's combine the data:

```
combined <- merge(user,test, by='user_id',all=TRUE)
```

Check the combined data:

```
head(combined)
```

```
##   user_id signup_date       date browser test pages_visited
## 1      34  2015-01-01 2015-08-15  Chrome    0             6
## 2      59  2015-01-01 2015-08-12  Chrome    1             6
## 3     178  2015-01-01 2015-08-10  Safari    1             3
## 4     285  2015-01-01 2015-08-03   Opera    0             5
## 5     383  2015-01-01 2015-08-05 Firefox    1             9
## 6     397  2015-01-01 2015-08-27      IE    0             1
```

```
tail(combined)
```

```
##        user_id signup_date       date browser test pages_visited
## 99995  8999264  2015-08-31 2015-08-31  Chrome    1             4
## 99996  8999327  2015-08-31 2015-08-31  Safari    0             3
## 99997  8999539  2015-08-31 2015-08-31      IE    1             2
## 99998  8999550  2015-08-31 2015-08-31      IE    0             7
## 99999  8999709  2015-08-31 2015-08-31  Chrome    1             4
## 100000 8999849  2015-08-31 2015-08-31  Chrome    1             1
```

```
str(combined)
```

```
## 'data.frame':    100000 obs. of  6 variables:
##  $ user_id      : int  34 59 178 285 383 397 488 608 656 771 ...
##  $ signup_date  : Factor w/ 243 levels "2015-01-01","2015-01-02",..: 1 1 1 1 1 1 1
1 1 1 ...
##  $ date         : Factor w/ 31 levels "2015-08-01","2015-08-02",..: 15 12 10 3 5 2
7 10 30 31 7 ...
##  $ browser      : Factor w/ 5 levels "Chrome","Firefox",..: 1 1 5 4 2 3 1 4 1 3 ..
.
##  $ test         : int  0 1 1 0 1 0 0 0 1 0 ...
##  $ pages_visited: int  6 6 3 5 9 1 1 7 7 6 ...
```

Converting signup_date and date as Date.

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

Let's check the summary of the combined data:

```
summary(combined)
```

```
##     user_id         signup_date             date
##  Min.   :     34   Min.   :2015-01-01   Min.   :2015-08-01
##  1st Qu.:2271007   1st Qu.:2015-03-08   1st Qu.:2015-08-08
##  Median :4519576   Median :2015-05-14   Median :2015-08-16
##  Mean   :4511960   Mean   :2015-05-11   Mean   :2015-08-16
##  3rd Qu.:6764484   3rd Qu.:2015-07-18   3rd Qu.:2015-08-24
##  Max.   :8999849   Max.   :2015-08-31   Max.   :2015-08-31
##     browser           test        pages_visited
##  Chrome :43427   Min.   :0.0000   Min.   : 0.000
##  Firefox:21758   1st Qu.:0.0000   1st Qu.: 3.000
##  IE     :21880   Median :1.0000   Median : 5.000
##  Opera  : 2127   Mean   :0.5015   Mean   : 4.604
##  Safari :10808   3rd Qu.:1.0000   3rd Qu.: 6.000
##                  Max.   :1.0000   Max.   :17.000
```

Check if there is any missing data:

```
any(is.na(combined))
```

```
## [1] FALSE
```

Good, there is no missing data.

t-test

First check if the data were equally distributed or not to both control and test set.

```
length(combined$user_id[combined$test==1])==length(combined$user_id[combined$test==0]
)
```

```
## [1] FALSE
```

It gives False result, that means data in both group are not equal. Let's see the difference:

```
length(combined$user_id[combined$test==1])-length(combined$user_id[combined$test==0])
```

```
## [1] 308
```

Let's check the general t-test:

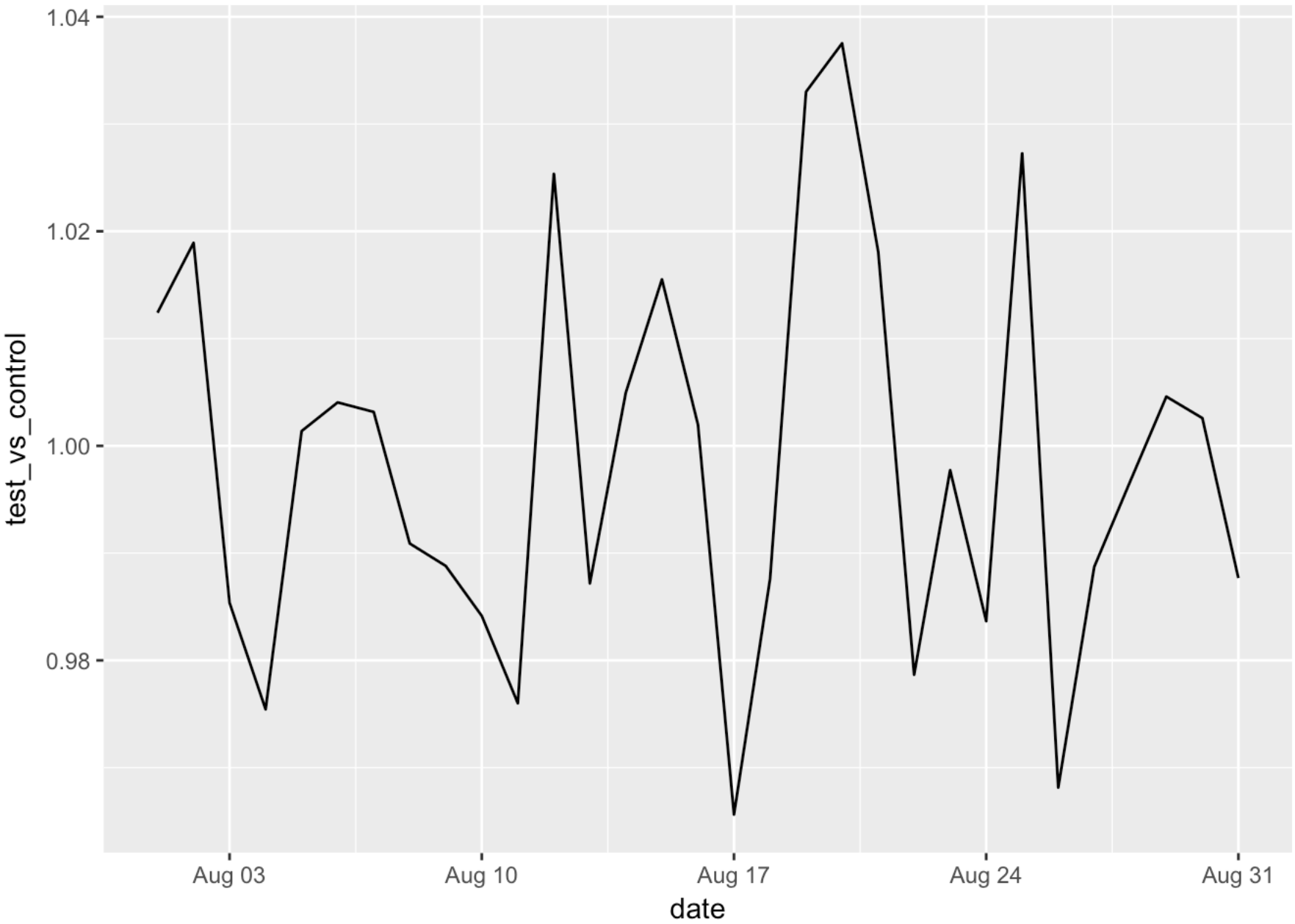
```
t.test(combined$pages_visited[combined$test==1], combined$pages_visited[combined$test
==0])
```

```
##
##  Welch Two Sample t-test
##
## data:  combined$pages_visited[combined$test == 1] and combined$pages_visited[combi
ned$test == 0]
## t = -0.55711, df = 95835, p-value = 0.5775
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.03931178  0.02190997
## sample estimates:
## mean of x mean of y
##  4.599693  4.608394
```

Overall the test is not winning, rather it's 0.2% drop.

let's plot day by day

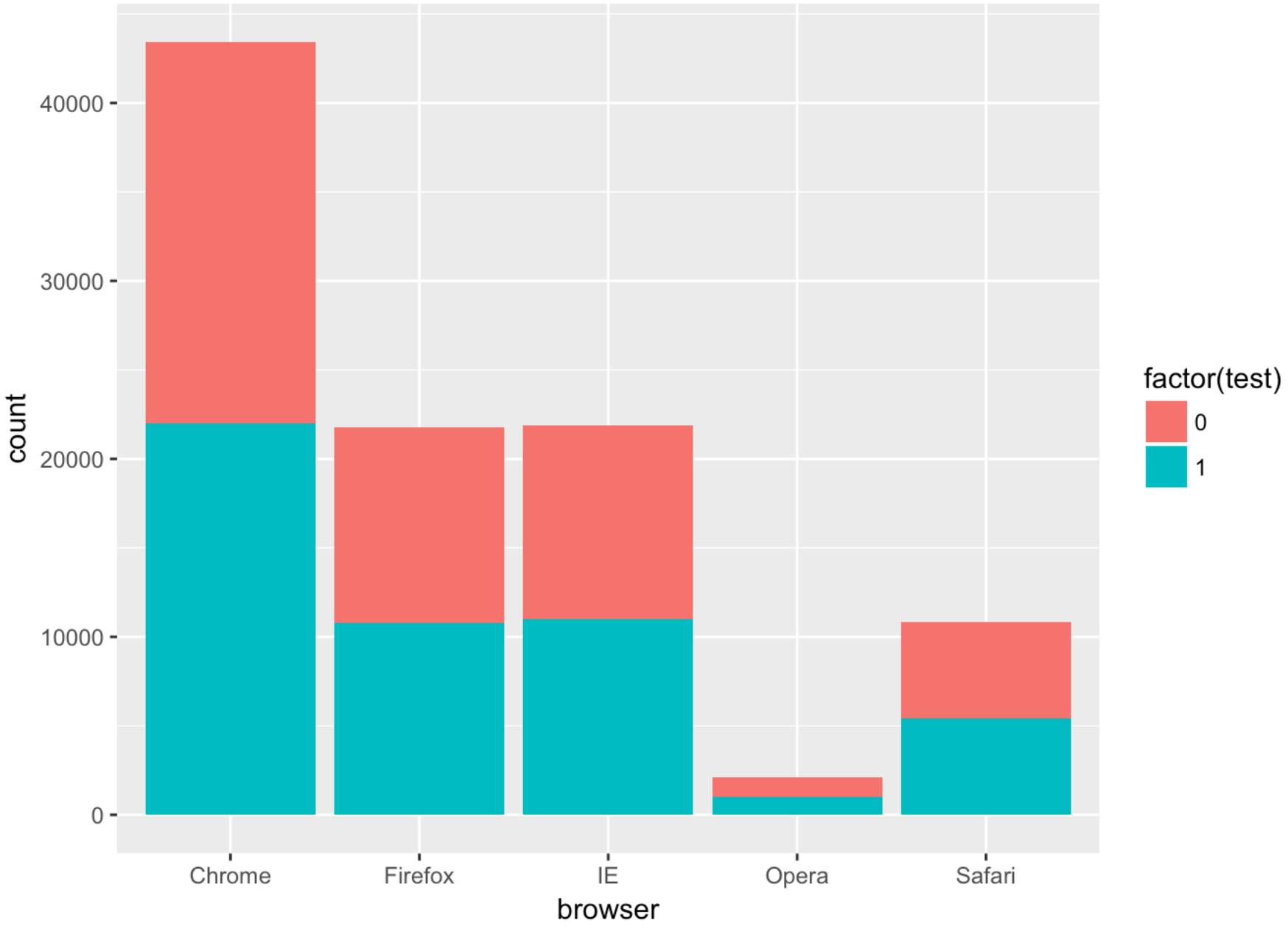
```
library(dplyr)
library(ggplot2)
data_by_day <- combined %>%
               group_by (date) %>%
               summarize(test_vs_control = mean(pages_visited[test==1])/
                                           mean(pages_visited[test==0]))

ggplot(data=data_by_day,aes(x=date,y=test_vs_control))+
  geom_line()
```



There might be some reason the users from some segment ended up in test or control and this affected the overall results.

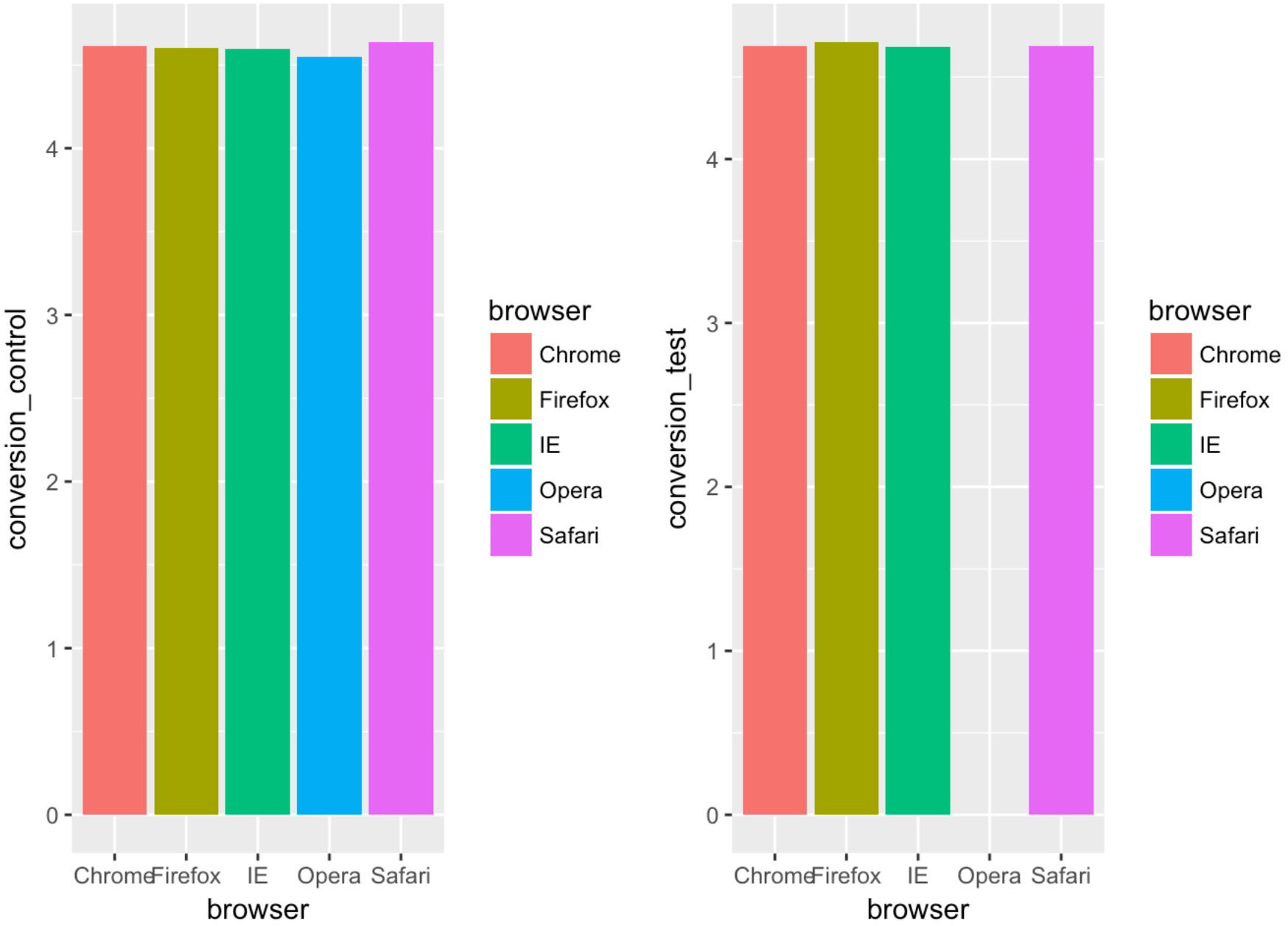
First let's check how the data are distributed in the browser segment.

```
ggplot(combined, aes(x=browser, fill=factor(test))) +geom_bar()
```



Looks like the distribution of users in different browsers is equal. Now check the conversion in the browser segment:

```
library(gridExtra)
data_browser <- combined %>%
                group_by (browser) %>%
                summarize(conversion_test = mean(pages_visited[test==1]),
                          conversion_control = mean(pages_visited[test==0]))

p1 <- ggplot(data=data_browser,aes(x=browser,y=conversion_control))+
  geom_bar(stat='identity',aes(fill=browser))

p2 <- ggplot(data=data_browser,aes(x=browser,y=conversion_test))+
  geom_bar(stat='identity',aes(fill=browser))

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



Interesting! There is no conversion in the test group for Opera browser. There might be some bug issue. Let's remove the data of Opera from both test and control sets.

```
data_updated <- subset(combined, !combined$browser=='Opera')
```

Let's check the general t-test again:

```
t.test(data_updated$pages_visited[data_updated$test==1], data_updated$pages_visited[d
ata_updated$test==0])
```

```
##
##  Welch Two Sample t-test
##
## data:  data_updated$pages_visited[data_updated$test == 1] and data_updated$pages_v
isited[data_updated$test == 0]
## t = 5.4743, df = 92316, p-value = 4.404e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.05468628 0.11568527
## sample estimates:
## mean of x mean of y
##  4.694989  4.609804
```

Looks like the test is winning at 1.85%. Now let's check the novelty effect.

Separate the data of new user.

```
data_new_user <- subset(data_updated, data_updated$signup_date==data_updated$date)
```

Now check t-test for new user:

```
t.test(data_new_user$pages_visited[data_new_user$test==1], data_new_user$pages_visite
d[data_new_user$test==0])
```

```
##
##  Welch Two Sample t-test
##
## data:  data_new_user$pages_visited[data_new_user$test == 1] and data_new_user$page
s_visited[data_new_user$test == 0]
## t = -1.0809, df = 19563, p-value = 0.2797
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.11657577  0.03370165
## sample estimates:
## mean of x mean of y
##  4.593712  4.635149
```

Even though the test is winning overall but it's not winning for the new users. It's novelty effect.

In conclusion we can say that the test is not winning.