User Referral Program

Mitul Shah

12/30/2016

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)
```

Getting a sense of the data

Now, let's load the data.

```
referral <- read.csv("referral.csv")
```

Now, let's have a look at the structure and the summary of the data.

```
## Looking at the structure and the summary
str(referral)
```

```
## 'data.frame':    97341 obs. of  6 variables:
##  $ user_id    : int  2 3 6 7 7 10 17 19 19 19 ...
##  $ date       : Factor w/ 56 levels "2015-10-03","2015-10-04",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ country    : Factor w/ 9 levels "CA","CH","DE",..: 5 1 5 8 7 3 8 8 9 5 ...
##  $ money_spent: int  65 54 35 73 35 36 25 69 17 29 ...
##  $ is_referral: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ device_id  : Factor w/ 17887 levels "AAASIUHCEETRZ",..: 3342 15670 1443 10783 10783 2019 11776 55
```

```
summary(referral)
```

```
##     user_id              date          country       money_spent
##  Min.   :    1   2015-11-14: 3303   UK     :15493   Min.   : 10.00
##  1st Qu.: 2020   2015-11-15: 3283   FR     :15396   1st Qu.: 27.00
##  Median : 4053   2015-10-31: 3233   US     :15280   Median : 42.00
##  Mean   : 6355   2015-11-22: 3220   IT     :11446   Mean   : 44.69
##  3rd Qu.:10286   2015-11-21: 3209   DE     :11093   3rd Qu.: 59.00
##  Max.   :20000   2015-11-08: 3164   ES     : 9831   Max.   :220.00
##                  (Other)   :77929   (Other):18802
##   is_referral            device_id
##  Min.   :0.0000   JOVUEUUQPQVXO:   35
##  1st Qu.:0.0000   XLJODRPXYKPRO:   34
##  Median :0.0000   KRGUOOGZKNQRQ:   33
##  Mean   :0.2878   KQMNMACBAEKPP:   32
##  3rd Qu.:1.0000   NWJQZEWLIUYHW:   30
##  Max.   :1.0000   OMCIHDOOQWZIG:   30
##                   (Other)      :97147
```

Let's now look at the first 10 rows of the data.

```
## Looking at the first 10 rows of the data
head(referral, n = 10)
```

```
##    user_id       date country money_spent is_referral     device_id
## 1        2 2015-10-03      FR          65           0 EVDCJTZMVMJDG
## 2        3 2015-10-03      CA          54           0 WUBZFTVKXGQQX
## 3        6 2015-10-03      FR          35           0 CBAPCJRTFNUJG
## 4        7 2015-10-03      UK          73           0 PRGXJZAJKMXRH
## 5        7 2015-10-03      MX          35           0 PRGXJZAJKMXRH
## 6       10 2015-10-03      DE          36           0 CVZCQLPXZCFUV
## 7       17 2015-10-03      UK          25           0 RCHOYRWHPOEVE
## 8       19 2015-10-03      UK          69           0 ICGUPKJIJFZUK
## 9       19 2015-10-03      US          17           0 ICGUPKJIJFZUK
## 10      19 2015-10-03      FR          29           0 ICGUPKJIJFZUK
```

Data Quality Issues

There is one strange thing which we observe in the first few rows of the data. There are some users who have multiple transactions on the same day in different countries. For example, we see user 7 having transactions in UK and MX on the same day while user 19 having transactions in UK, US and FR on the same day. This is highly unlikely to happen. Moreover, there are many such users in our data. But, for the sake of convenience, I have not removed all these users as there is still a possibility that a user can have these transactions while travelling. 2 consequent transactions by the same user in 2 different countries on the same date is possible but 3 or more consequent transactions by the same user in 3 or more different countries might seem unlikely.

There are a couple of more things which we need to check in our data. First, the number of unique users in our data should be equal to the number of unique devices as each user have their own device for the transaction. So let's check this.

```
## Checking whether the number of unique users is equal to the number of unique devices
length(unique(referral$user_id))
```

```
## [1] 18809
```

```
length(unique(referral$device_id))
```

```
## [1] 17887
```

We see that the number of unique devices is about 1000 less than the number of unique users. This means that some of the users are using other users' devices for the purchase. There is a chance that some of these transactions might be fraud. But we do not have the data to comment anything on why the number of unique users is not equal to the number of unique devices. Again, I have not removed any rows from the data for the sake of convenience.

Second, as we are given that the new referral program began on 31st October, our data should not have value 1 in is_referral column before 31st October. We need to check this.

But before checking this, we need to change the mode of the date variable to date. I have also added another column to our data indicating whether the user made a transaction after the referral program began or before the referral program.

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

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

```
## Function to define users after the referral program
after_referral_user <- function(date){
  if(date >= "2015-10-31")
    return("Yes")
  if(date < "2015-10-31")
    return("No")
  else
    return(NA)
}

## Creating a new column for indicating after referral users
referral$is_after_referral_program <- sapply(referral$date, after_referral_user)
```

Now, let's check whether our data has 1 in is_referral column before the referral program.

```
## Subsetting the data before 31st Oct
data_before_referral <- referral[referral$date < "2015-10-31", ]

## Finding the rows which contain 1 in is_referral variable before 31st Oct
data_before_referral[data_before_referral$is_referral == 1, ]
```

```
## [1] user_id                  date
## [3] country                  money_spent
## [5] is_referral              device_id
## [7] is_after_referral_program
## <0 rows> (or 0-length row.names)
```

Our data doesn't have any observations which have value of is_referral variable to be 1 before the Referral Program. So this looks good.

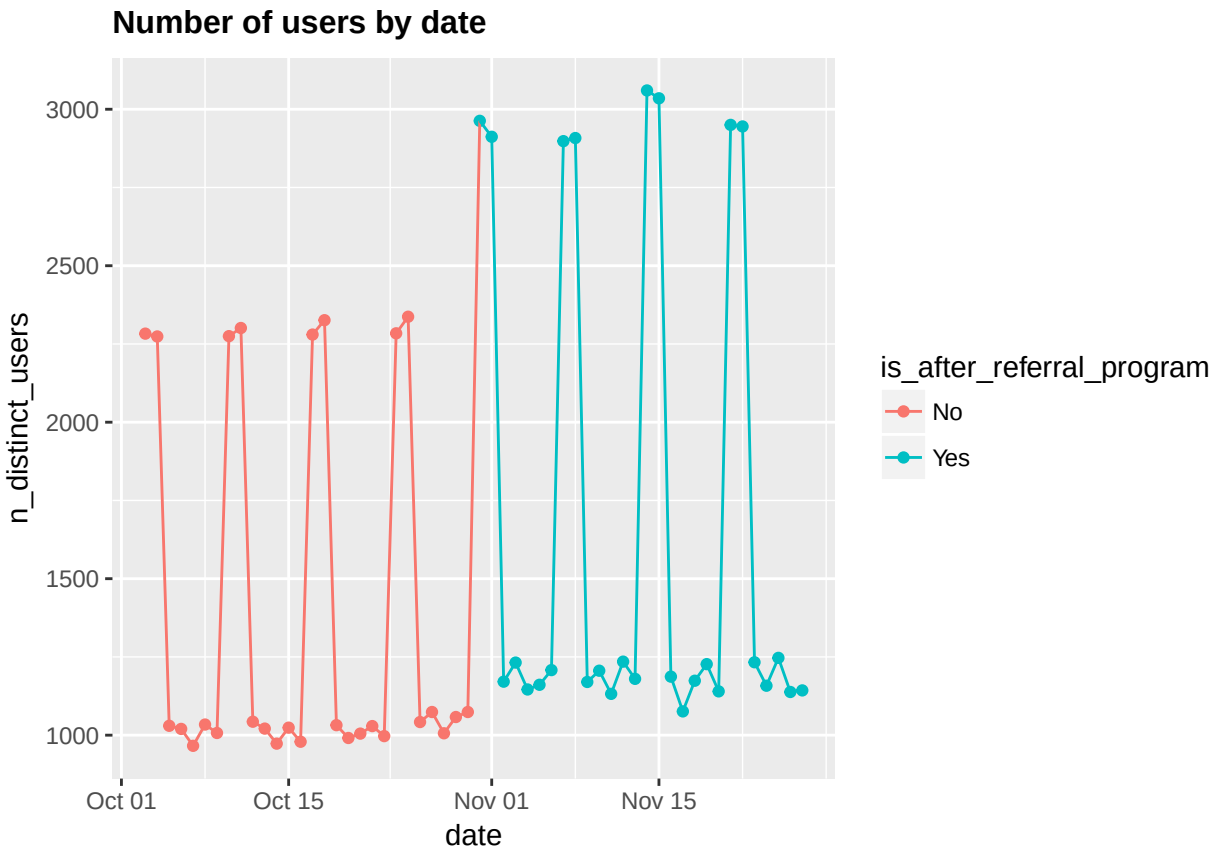
Impact of the Referral Program in terms of number of users

Let's check whether the referral program increased the number of users on the site.

```
## Number of distinct users by date
n_users_by_date <- referral %>% group_by(date) %>%
  summarise(n_distinct_users = n_distinct(user_id))

## Creating a new column for indicating after referral users
n_users_by_date$is_after_referral_program <- sapply(n_users_by_date$date, after_referral_user)

## Visualizing number of users by date
ggplot(n_users_by_date, aes(date, n_distinct_users, color = is_after_referral_program, group = 1)) +
  geom_point() +
  geom_line() +
  ggtitle("Number of users by date") +
  theme(plot.title = element_text(size = 12, face = "bold"))
```



We can see that the referral program has clearly increased the number of users on the site.

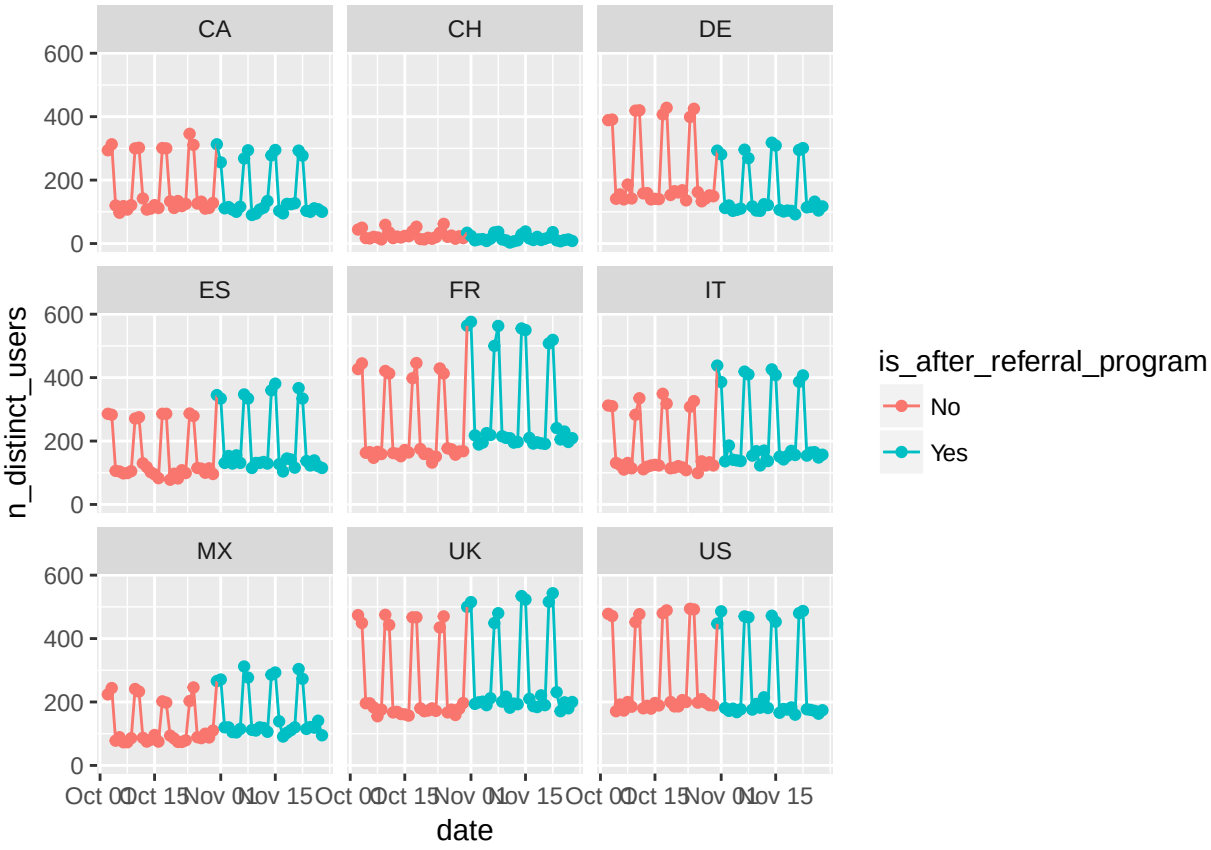
We also see that there are more number of users for 2 consecutive days in every week. These are Saturdays and Sundays. So more users are coming to the site in the weekend.

Let's also check the impact on the number of users before and after the referral program in each of the nine countries in our data.

```
## Number of users by date and country
n_users_by_date_and_country <- referral %>% group_by(date, country)   %>% summarise(n_distinct_users = n

## Creating a new column for indicating after referral users
n_users_by_date_and_country$is_after_referral_program <- sapply(n_users_by_date_and_country$date, after

## Visualize it!
ggplot(n_users_by_date_and_country, aes(date, n_distinct_users, color = is_after_referral_program, group
  geom_point() +
  geom_line() +
  facet_wrap(~country, ncol = 3)
```



From the above plots, we see that the number of users have decreased in Switzerland and Germany after the referral program. So the referral program might not be a good idea in these countries if we are just trying to increase the number of users. But we still need to compare the revenue generated in these countries before and after the referral program to check the impact.

The referral program seems to be working really well in terms of number of users in Spain, France, Italy and Mexico as we clearly see the increase in the number of users in these countries. But Canada, US and UK doesn't seem to have a huge impact on the number of users after the referral program.

Impact in terms of Revenue Generated

```r
## Data before the referral program
data_before_referral <- referral[referral$date < "2015-10-31", ]

## Data after the referral program
data_after_referral <- referral[referral$date >= "2015-10-31", ]

## Subsetting the referred users
referred_users <- data_after_referral[data_after_referral$is_referral == 1, ]

## Comparing average money spent per transaction before referral and after referral
t.test(data_before_referral$money_spent, data_after_referral$money_spent)
```

```
##
##  Welch Two Sample t-test
##
## data:  data_before_referral$money_spent and data_after_referral$money_spent
## t = -31.067, df = 96157, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.782069 -4.214485
## sample estimates:
## mean of x mean of y
##  42.38178  46.88006
```

```r
## Number of unique referred users
length(unique(referred_users$user_id))
```

```
## [1] 12715
```

To estimate the impact, we just need to compare the total money spent before referral program which is equal to 42.38178 * 47341 (47341 is the number of transactions before the referral program) and the difference between total money spent after referral program and 10 * number of unique referred users, i.e. 46.88006 * 50000 - 10 * 12715. So we are just comparing $2006396 and $2216853. We can clearly see that the revenue generated after the referral program was more than that of the revenue generated before the referral program.

But there is a risk associated in measuring the impact of the referral program with this approach. By using this approach, we are not measuring the actual impact of the referral program as the revenue which we calculated were in different time periods. The users might behave differently in different time periods. It is possible that the revenue generated in the 2nd period (after the referral program began) was just due to the seasonal behavioural changes of the users.

So, the better approach to measure the actual impact of the referral program is to compare the revenue generated in the same time period, i.e. by comparing the revenue generated without the referred users (after the referral program began) and the revenue generated with the referred users.

```r
## Subsetting the users who were not referred after the referral program began
not_referred_users <- data_after_referral[data_after_referral$is_referral == 0, ]

## Revenue generated from the users who were not referred and came to the site by themselves after the
sum(not_referred_users$money_spent)
```

```
## [1] 1028216
```

```
## Revenue generated from referred users after the referral program
sum(referred_users$money_spent)
```

```
## [1] 1315787
```

So, now we are comparing $1028216 (this would have been the revenue if only the non-referred users had come to the site) and 1028216 + 1315787 - 10 * 12715 = $2216853.

Now, we see the actual impact of the referral program on the site. The revenue generated has more than doubled! So the impact of the referral program is huge on the site.

Conclusions

1. The referral program is definitely working well in all the countries as we see that the revenue generated has more than doubled because of the referred users coming to the site.
2. We cannot estimate the impact of the referral program by comparing the data before the referral program and the data after the referral program because we might miss the seasonal changes in the users' behaviour of making the transactions on the site.
3. Based on the data, we see that the number of users making the transactions is really high on the weekends. So it might be a good idea to offer some sort of discount on the weekends. This might increase the number of users on the weekends heavily generating higher revenue.
4. The company has only offered the discount to the user who is referring other users and the user gets 10$ credit when a new user makes a transaction on the site. It might be a good idea to offer some credit to the new user who makes the transaction, the first time on the site, as well.