Quora Data Science Challenge

Mariam Walaa

11/10/2021

```
knitr::opts_chunk$set(fig.width=12, fig.height=8)
```

```
library(tidyverse)
library(lubridate)
```

```
user_activity_pre <- readr::read_csv("data/t3_user_active_min_pre.csv")
user_activity <- readr::read_csv("data/t1_user_active_min.csv")
user_attributes <- readr::read_csv("data/t4_user_attributes.csv")
user_variant <- readr::read_csv("data/t2_user_variant.csv")
```

Data

```
full_df <-
  user_activity_pre %>%
  bind_rows(user_activity) %>%
  filter(active_mins <= (24 * 60)) %>%
  left_join(user_attributes, by = "uid") %>%
  left_join(user_variant %>% select(-dt), by = "uid") %>%
  filter(lubridate::year(signup_date) >= 2009) %>%
  mutate(variant_number = recode(variant_number, `0` = "Control", `1` = "Treatment"))

glimpse(full_df)
```

```
## Rows: 2,256,021
## Columns: 7
## $ uid            <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ dt             <date> 2018-09-24, 2018-11-08, 2018-11-24, 2018-11-28, 2018-1~
## $ active_mins    <dbl> 3, 4, 3, 6, 6, 1, 8, 5, 8, 2, 1, 2, 3, 2, 3, 3, 1, 4, 1~
## $ gender         <chr> "male", "male", "male", "male", "male", "male", "male",~
## $ user_type      <chr> "non_reader", "non_reader", "non_reader", "non_reader",~
## $ variant_number <chr> "Control", "Control", "Control", "Control", "Control", ~
## $ signup_date    <date> 2018-09-24, 2018-09-24, 2018-09-24, 2018-09-24, 2018-0~
```

T-Test (Experiment Data)

```r
first_test <-
    user_activity %>%
    filter(active_mins <= (24 * 60)) %>%
    left_join(user_variant %>% select(-dt), by = "uid") %>%
    mutate(variant_number = recode(variant_number, `0` = "Control", `1` = "Treatment"))  %>%
    group_by(dt, variant_number) %>%
    summarise(total_mins = sum(active_mins), .groups = 'drop')

t.test(first_test %>% filter(variant_number == "Control") %>% ungroup()  %>% select(total_mins),
       first_test %>% filter(variant_number == "Treatment") %>% ungroup()  %>% select(total_mins))
```

```
##
##  Welch Two Sample t-test
##
## data:  first_test %>% filter(variant_number == "Control") %>% ungroup() %>% select(total_mins) and fi
## t = 89.307, df = 217.27, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  84284.32 88088.44
## sample estimates:
## mean of x mean of y
##  114326.2   28139.8
```

T-Test (Pre-Experiment Included)

```r
second_test <-
  user_activity_pre %>%
  bind_rows(user_activity) %>%
  filter(active_mins <= (24 * 60)) %>%
  left_join(user_attributes, by = "uid") %>%
  left_join(user_variant %>% select(-dt), by = "uid") %>%
  filter(user_type != "new_user") %>%
  mutate(variant_number = recode(variant_number, `0` = "Control", `1` = "Treatment"))  %>%
  group_by(dt, variant_number) %>%
  summarise(total_mins = sum(active_mins), .groups = 'drop')

t.test(second_test %>% filter(variant_number == "Control") %>% ungroup()  %>% select(total_mins),
       second_test %>% filter(variant_number == "Treatment") %>% ungroup()  %>% select(total_mins))
```

```
##
##  Welch Two Sample t-test
##
## data:  second_test %>% filter(variant_number == "Control") %>% ungroup() %>% select(total_mins) and s
## t = 100.83, df = 499.12, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  86652.70 90096.67
## sample estimates:
## mean of x mean of y
## 109272.83  20898.15
```

Analysis

```
full_df %>%
  group_by(uid, variant_number) %>%
  count() %>% ungroup(uid) %>% count()
```

Counts By Stratification

```
## # A tibble: 2 x 2
## # Groups:   variant_number [2]
##   variant_number     n
##   <chr>          <int>
## 1 Control        39886
## 2 Treatment       9964
```

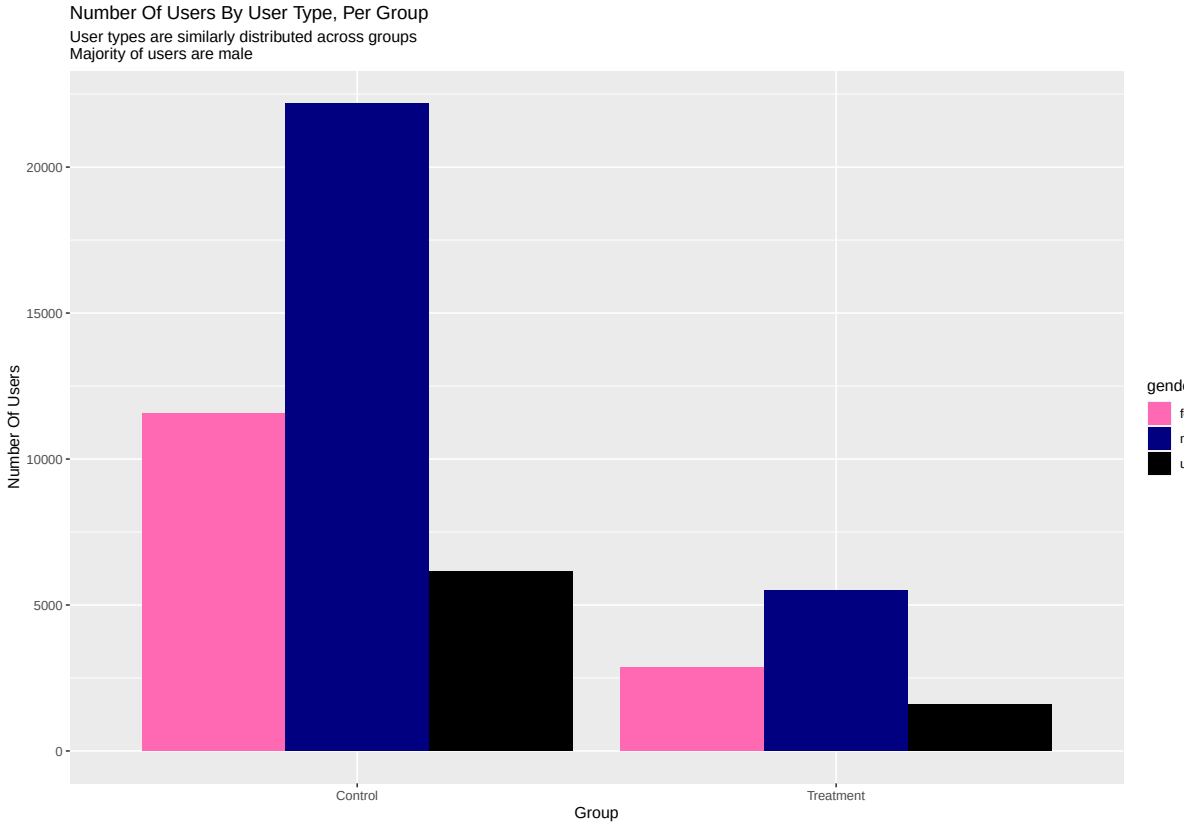
```
full_df %>%
    group_by(uid, variant_number, gender) %>%
    count() %>% ungroup(uid) %>% count()
```

```
## # A tibble: 6 x 3
## # Groups:   variant_number, gender [6]
##   variant_number gender      n
##   <chr>          <chr>   <int>
## 1 Control        female  11558
## 2 Control        male    22187
## 3 Control        unknown  6141
## 4 Treatment      female   2856
## 5 Treatment      male     5503
## 6 Treatment      unknown  1605
```

```
full_df %>%
  group_by(uid, variant_number, user_type) %>%
    count() %>% ungroup(uid) %>% count()
```

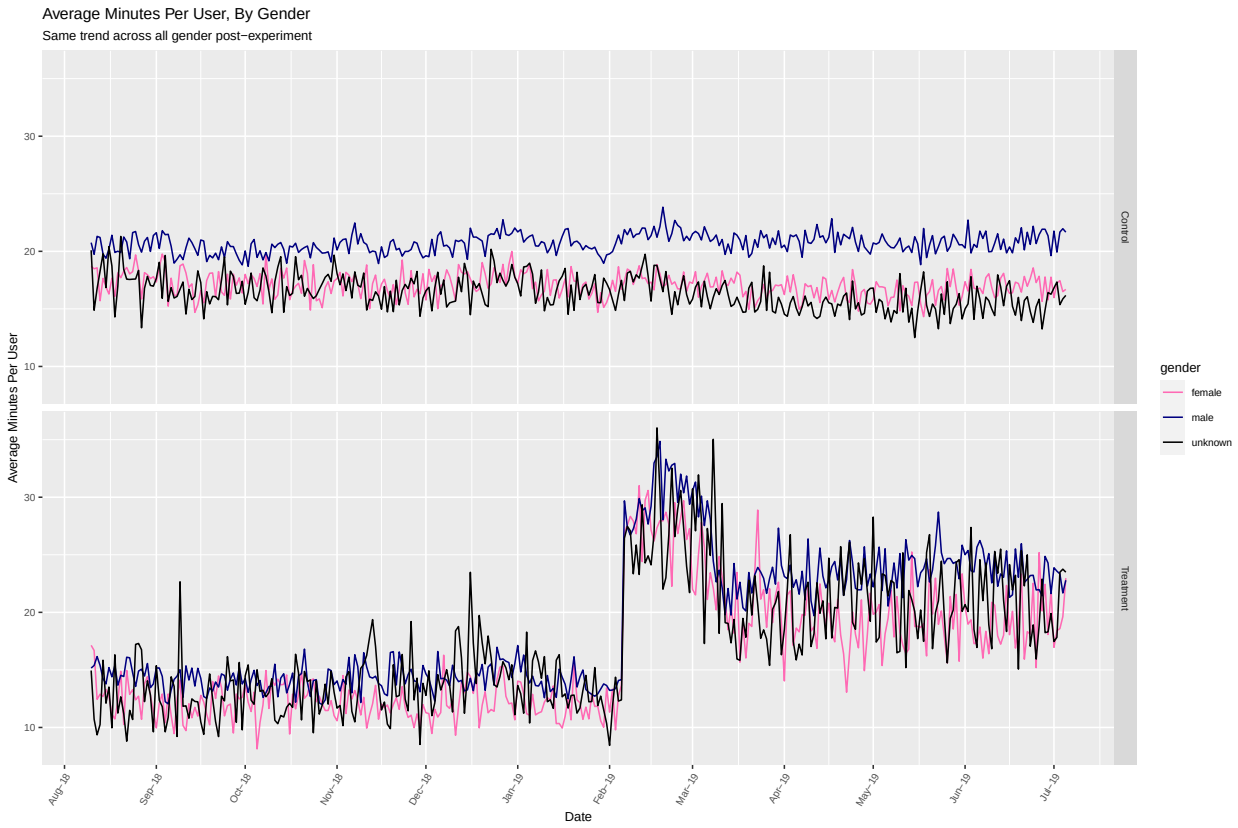
```
## # A tibble: 8 x 3
## # Groups:   variant_number, user_type [8]
##   variant_number user_type       n
##   <chr>          <chr>       <int>
## 1 Control        contributor   915
## 2 Control        new_user     3588
## 3 Control        non_reader  28650
## 4 Control        reader       6733
## 5 Treatment      contributor   129
## 6 Treatment      new_user     1210
## 7 Treatment      non_reader   7356
## 8 Treatment      reader       1269
```

```
full_df %>%
  group_by(uid, variant_number, gender) %>%
    count() %>% ungroup(uid) %>% count() %>%
    ggplot(aes(x = variant_number, y = n, fill = gender)) +
    geom_bar(stat = "identity", position='dodge') +
    ylab("Number Of Users") + xlab("Group") +
    ggtitle("Number Of Users By User Type, Per Group",
            subtitle = "User types are similarly distributed across groups\nMajority of users are male") +
    scale_fill_manual(values=c("hotpink", "navy", "black"))
```

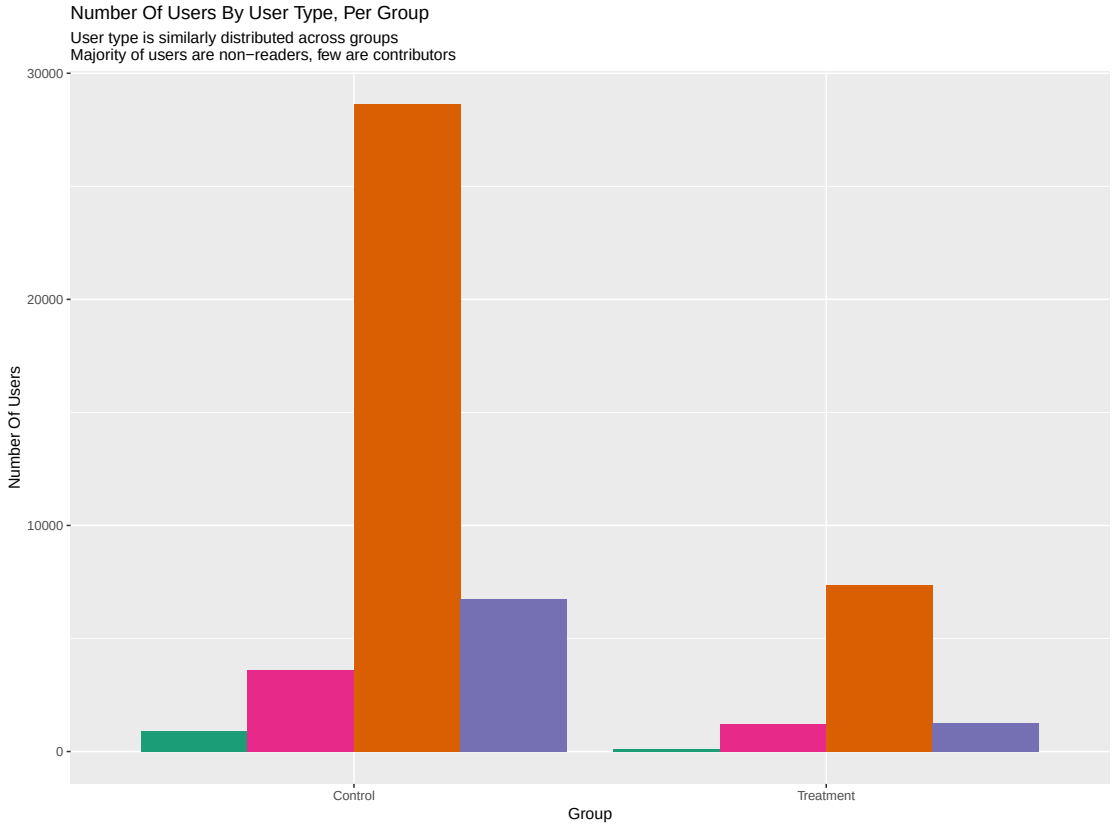


Trend By Gender

```
full_df %>%
  group_by(dt, gender, variant_number) %>%
    summarise(avg_min_per_usr = mean(active_mins), .groups = 'drop') %>%
    ggplot(aes(x = dt, y = avg_min_per_usr, color = gender)) +
    geom_line() + facet_grid(rows = vars(variant_number)) +
    scale_colour_manual(values=c("hotpink", "navy", "black")) +
    ggtitle("Average Minutes Per User, By Gender",
            subtitle = "Same trend across all gender post-experiment") +
    ylab("Average Minutes Per User") + xlab("Date") +
    scale_x_date(date_breaks = "months" , date_labels = "%b-%y") +
    theme(text = element_text(size=9), axis.text.x = element_text(angle=60, hjust=1))
```

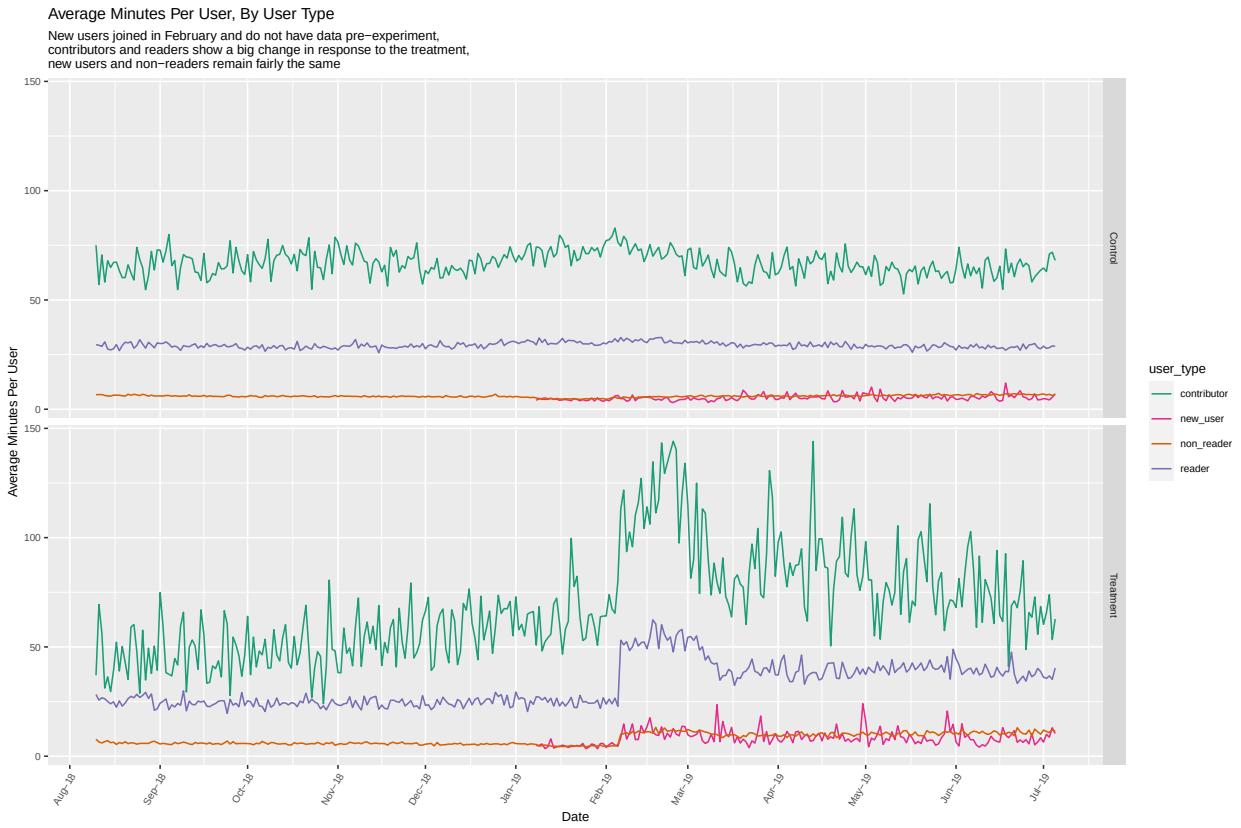



```
full_df %>%
  group_by(uid, variant_number, user_type) %>%
    count() %>% ungroup(uid) %>% count() %>%
    ggplot(aes(x = variant_number, y = n, fill = user_type)) +
    geom_bar(stat = "identity", position='dodge') +
    ylab("Number Of Users") + xlab("Group") +
    ggtitle("Number Of Users By User Type, Per Group",
            subtitle = "User type is similarly distributed across groups\nMajority of users are non-rea
    scale_fill_manual(values=c("#1b9e77", "#e7298a", "#d95f02", "#7570b3")) # https://colorbrewer2.org/
```



Trend By User Type

```
full_df %>%
  group_by(dt, user_type, variant_number) %>%
    summarise(avg_min_per_usr = mean(active_mins), .groups = 'drop') %>%
    ggplot(aes(x = dt, y = avg_min_per_usr, color = user_type)) +
    geom_line() +
    facet_grid(rows = vars(variant_number)) +
    scale_colour_manual(values=c("#1b9e77", "#e7298a", "#d95f02", "#7570b3")) + # https://colorbrewer2.
    ggtitle("Average Minutes Per User, By User Type",
            subtitle = "New users joined in February and do not have data pre-experiment,\ncontributors
    ylab("Average Minutes Per User") + xlab("Date") +
    scale_x_date(date_breaks = "months" , date_labels = "%b-%y") +
    theme(text = element_text(size=9), axis.text.x = element_text(angle=60, hjust=1))
```



Summary

Three units of analysis were considered in conducting this t-test:

1. Total minutes per user in each group (number of data points in each group = number of users in that group)
2. Total minutes per user per day in each group (number of data points in each group = number of users in that group x number of days of the experiment)
3. Total minutes per group per day (number of data points in each group = number of days of the experiment)

The chosen unit of analysis was **Total Minutes Per Group Per Day**. The reason for this selection is we aim to compare the difference in minutes spent on the site between the Control Group and Treatment Group. This comparison is time-dependent as the user activity data is aggregated up to each user's daily activity (i.e., each user has a single record for their activity on a given day). Therefore, to determine whether there is a difference in time spent on the app by the group that was given the new UI design, we compare the day-by-day total minutes spent on the site by all users in each group on a given day.

The number of data points in each group is then 150 days as the experiment runs from February to July.

The other two choices would not be correct since the data points within a group would not be dependent. Each data point associated with a certain user would be dependent and this violates the assumption of independence for data points within a group in a t-test.

To conduct the first t-test, I use the standard Welch Two Sample t-test implemented through the R base function t.test(). I remove records with active minutes greater than 24*60 minutes per day as well as users who signed up earlier than 2009 because those are clear logging errors. I use a left join to combine the

user_activity data with the user_variant data, and I do a simple recoding of the variant number. I then group by the date and variant number, and summarize by summing up the active_mins for each date and group in the data.

Based on these findings, I recommend pushing the new UI design to production, since we find a [84284.32, 88088.44] confidence interval in the difference between the average total time spent for both groups, with a large t-value of 89.307 indicating a large difference in the two groups and a small p-value indicating that there is stronger evidence in favor of the alternative hypothesis (that the true difference in means is not equal to 0).

To compute the updated treatment effect by applying the pre-experiment data, I apply the same cleaning steps as in the first t-test, in addition to binding the rows of the user_activity_pre data with the user_activity data. The pre-experiment data does not change my conclusion about the treatment effect – I still recommend that the UI design is pushed to production.

The disaggregation by gender shows that trends are similar across genders within each group, but trends are vastly different between groups. In general, the treatment group shows an increase in total time spent per day once experiment begins.

The disaggregation by user shows that trends vary by user type. In the treatment group, new users and non-readers have the same overall trend while readers and contributors show an increase once experiment starts. In the control group, the trend remains same across all four user types, but the contributors and readers generally spend more time than non-readers and new users.

The plot disaggregating by user type shows that new users enter post-experiment. Given this new information, I would perform a new t-test excluding the new users from the combined data.

Looking at distributions of user types and genders within the control and treatment group, a large majority of users in both groups are male and non-readers. In the treatment group, the contributor group is the smallest group of users with large variance in time spent over time. I recommend the product team attempts the experiment with a more balanced stratification of user types within each group, or at least more contributors.