Quora Data Challenge

```
library("tidyverse")
```

```
##   Attaching packages                tidyverse 1.2.1

##   ggplot2 3.2.1      purrr   0.3.3
##   tibble  2.1.3      dplyr   0.8.3
##   tidyr   1.0.0      stringr 1.4.0
##   readr   1.3.1      forcats 0.4.0

##   Conflicts                   tidyverse_conflicts()
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Conduct a t-test

Read data

```
t1 <- read_csv("t1_user_active_min.csv")
```

```
## Parsed with column specification:
## cols(
##   uid = col_double(),
##   dt = col_date(format = ""),
##   active_mins = col_double()
## )
```

```
t2 <- read_csv("t2_user_variant.csv")
```

```
## Parsed with column specification:
## cols(
##   uid = col_double(),
##   variant_number = col_double(),
##   dt = col_date(format = ""),
##   signup_date = col_date(format = "")
## )
```

Brief view

```
head(t1)
```

```
## # A tibble: 6 x 3
##     uid dt         active_mins
##   <dbl> <date>           <dbl>
```

```
## 1     0 2019-02-22           5
## 2     0 2019-03-11           5
## 3     0 2019-03-18           3
## 4     0 2019-03-22           4
## 5     0 2019-04-03           9
## 6     0 2019-04-06           1
```

```
head(t2)
```

```
## # A tibble: 6 x 4
##     uid variant_number dt         signup_date
##   <dbl>          <dbl> <date>     <date>
## 1     0              0 2019-02-06 2018-09-24
## 2     1              0 2019-02-06 2016-11-07
## 3     2              0 2019-02-06 2018-09-17
## 4     3              0 2019-02-06 2018-03-04
## 5     4              0 2019-02-06 2017-03-09
## 6     5              0 2019-02-06 2018-06-25
```

Find outliers and remove them

```
outlier_values <- boxplot.stats(t1$active_mins)$out

t1 <- t1[-which(t1$active_mins %in% outlier_values), ]
```

Join `t1` and `t2`

```
after <- t1 %>%
  group_by(uid) %>%
  summarise(
    mean_active_mins = mean(active_mins)
  ) %>%
  left_join(t2, by = "uid")
```

```
head(after)
```

```
## # A tibble: 6 x 5
##     uid mean_active_mins variant_number dt         signup_date
##   <dbl>            <dbl>          <dbl> <date>     <date>
## 1     0             3.31              0 2019-02-06 2018-09-24
## 2     1            19.8               0 2019-02-06 2016-11-07
## 3     2             2.43              0 2019-02-06 2018-09-17
## 4     3             3.21              0 2019-02-06 2018-03-04
## 5     4             1.95              0 2019-02-06 2017-03-09
## 6     5            12.4               0 2019-02-06 2018-06-25
```

t-test

```r
control <- after$mean_active_mins[which(after$variant_number == 0)]
treatment <- after$mean_active_mins[which(after$variant_number == 1)]
```

```r
t.test(control, treatment)
```

```
##
##  Welch Two Sample t-test
##
## data:  control and treatment
## t = -29.678, df = 14388, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.651482 -1.446849
## sample estimates:
## mean of x mean of y
##  5.367573  6.916739
```

Gathering additional data

Read data

```r
t3 <- read_csv("t3_user_active_min_pre.csv")
```

```
## Parsed with column specification:
## cols(
##   uid = col_double(),
##   dt = col_date(format = ""),
##   active_mins = col_double()
## )
```

```r
head(t3)
```

```
## # A tibble: 6 x 3
##     uid dt         active_mins
##   <dbl> <date>           <dbl>
## 1     0 2018-09-24           3
## 2     0 2018-11-08           4
## 3     0 2018-11-24           3
## 4     0 2018-11-28           6
## 5     0 2018-12-02           6
## 6     0 2018-12-04           1
```

Data manipulation

```r
outlier_values_t3 <- boxplot.stats(t3$active_mins)$out

t3 <- t3[-which(t3$active_mins %in% outlier_values_t3), ]
```

```r
before_after <- t3 %>%
  group_by(uid) %>%
  summarise(
    mean_active_mins_before = mean(active_mins)
  ) %>%
  right_join(after, by = "uid") %>%
  mutate(
    diff = mean_active_mins - mean_active_mins_before
  ) %>%
  select(uid, variant_number, diff)
```

```r
head(before_after)
```

```
## # A tibble: 6 x 3
##     uid variant_number    diff
##   <dbl>          <dbl>   <dbl>
## 1     0              0 -0.0256
## 2     1              0 -2.47
## 3     2              0 -1.27
## 4     3              0 -0.625
## 5     4              0 -0.407
## 6     5              0  8.23
```

t-test

```r
diff_control <- before_after$diff[which(before_after$variant_number == 0)]

diff_treatment <- before_after$diff[which(before_after$variant_number == 1)]
```

```r
t.test(diff_control, diff_treatment)
```

```
##
##  Welch Two Sample t-test
##
## data:  diff_control and diff_treatment
## t = -46.773, df = 12481, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.948675 -1.791915
## sample estimates:
## mean of x mean of y
## 0.1834917 2.0537865
```

Deeper dive

Read data

```
t4 <- read_csv("t4_user_attributes.csv")
```

```
## Parsed with column specification:
## cols(
##   uid = col_double(),
##   gender = col_character(),
##   user_type = col_character()
## )
```

```
head(t4)
```

```
## # A tibble: 6 x 3
##     uid gender user_type
##   <dbl> <chr>  <chr>
## 1     0 male   non_reader
## 2     1 male   reader
## 3     2 male   non_reader
## 4     3 male   non_reader
## 5     4 male   non_reader
## 6     5 female non_reader
```

Join t2 and t4

```
info <- t2 %>%
  left_join(t4, by = "uid")
```

```
head(info)
```

```
## # A tibble: 6 x 6
##     uid variant_number dt         signup_date gender user_type
##   <dbl>          <dbl> <date>     <date>      <chr>  <chr>
## 1     0              0 2019-02-06 2018-09-24  male   non_reader
## 2     1              0 2019-02-06 2016-11-07  male   reader
## 3     2              0 2019-02-06 2018-09-17  male   non_reader
## 4     3              0 2019-02-06 2018-03-04  male   non_reader
## 5     4              0 2019-02-06 2017-03-09  male   non_reader
## 6     5              0 2019-02-06 2018-06-25  female non_reader
```

Views

```
info %>%
  group_by(variant_number) %>%
  summarise(
    cnt = n()
  )
```

```
## # A tibble: 2 x 2
##   variant_number   cnt
##            <dbl> <int>
## 1              0 40000
## 2              1 10000
```

```
info %>%
  group_by(variant_number, gender) %>%
  summarise(
    cnt = n()
  )
```

```
## # A tibble: 6 x 3
## # Groups:   variant_number [2]
##   variant_number gender    cnt
##            <dbl> <chr>   <int>
## 1              0 female  11607
## 2              0 male    22237
## 3              0 unknown  6156
## 4              1 female   2870
## 5              1 male     5520
## 6              1 unknown  1610
```

```
info %>%
  group_by(variant_number, user_type) %>%
  summarise(
    cnt = n()
  )
```

```
## # A tibble: 8 x 3
## # Groups:   variant_number [2]
##   variant_number user_type     cnt
##            <dbl> <chr>       <int>
## 1              0 contributor   915
## 2              0 new_user     3653
## 3              0 non_reader  28699
## 4              0 reader       6733
## 5              1 contributor   129
## 6              1 new_user     1235
## 7              1 non_reader   7367
## 8              1 reader       1269
```

```
info %>%
  group_by(variant_number, gender, user_type) %>%
  summarise(
    cnt = n()
  )
```

```
## # A tibble: 24 x 4
## # Groups:   variant_number, gender [6]
##    variant_number gender  user_type     cnt
##             <dbl> <chr>   <chr>       <int>
```

```
##  1              0 female  contributor   223
##  2              0 female  new_user     1176
##  3              0 female  non_reader   8387
##  4              0 female  reader       1821
##  5              0 male    contributor   596
##  6              0 male    new_user     1747
##  7              0 male    non_reader  15768
##  8              0 male    reader       4126
##  9              0 unknown contributor    96
## 10              0 unknown new_user      730
## # … with 14 more rows
```

The comprehensive dataset

```
data <- t1 %>%
  group_by(uid) %>%
  summarise(
    after_mean_active_mins = mean(active_mins)
  ) %>%
  left_join(info, by = "uid")

data <- t3 %>%
  group_by(uid) %>%
  summarise(
    before_mean_active_mins = mean(active_mins)
  ) %>%
  right_join(data, by = "uid") %>%
  select(
    uid,
    variant_number,
    gender, user_type,
    after_mean_active_mins,
    before_mean_active_mins
    ) %>%
  mutate(
    diff = after_mean_active_mins - before_mean_active_mins
  )
```

```
head(data)
```

```
## # A tibble: 6 x 7
##     uid variant_number gender user_type after_mean_acti… before_mean_act…
##   <dbl>          <dbl> <chr>  <chr>                <dbl>             <dbl>
## 1     0              0 male   non_read…             3.31              3.33
## 2     1              0 male   reader               19.8              22.3
## 3     2              0 male   non_read…             2.43              3.7
## 4     3              0 male   non_read…             3.21              3.83
## 5     4              0 male   non_read…             1.95              2.36
## 6     5              0 female non_read…            12.4               4.2
## # … with 1 more variable: diff <dbl>
```

```r
result <- tibble(gender = as.character(), user_type = as.character(), diff = as.numeric())
```

```r
for (i in unique(data$gender)){
  for (j in unique(data$user_type)){
    slice = data %>%
      filter(
        gender == i,
        user_type == j
      )

    control = slice$diff[which(slice$variant_number == 0)]
    treatment = slice$diff[which(slice$variant_number == 1)]

    test = t.test(control, treatment)

    diff = as.numeric(test$estimate[2] - test$estimate[1])

    tmp = tibble(gender = i, user_type = j, diff = diff)

    result = rbind(result, tmp)
  }
}
```

```r
result %>% arrange(desc(result$diff))
```

```
## # A tibble: 12 x 3
##    gender  user_type      diff
##    <chr>   <chr>         <dbl>
##  1 unknown reader       2.56
##  2 male    new_user     2.54
##  3 male    reader       2.16
##  4 unknown new_user     2.10
##  5 male    non_reader   1.99
##  6 female  non_reader   1.96
##  7 unknown non_reader   1.83
##  8 female  reader       1.70
##  9 female  new_user     1.55
## 10 female  contributor  0.997
## 11 unknown contributor -0.0322
## 12 male    contributor -0.0430
```