viagogo take home test

nabeil kizilbash

Data Cleaning

```
library(readxl)
Data <- read_excel("C:/Users/Murtz.Kizilbash/Desktop/viagogotest/Spring 2018 - Product Case Data.xlsx")

Data$Land <- as.factor(Data$Land)
Data$Bounce <- as.factor(Data$Bounce)
Data$Purchase <- as.factor(Data$Purchase)
Data$Channel <- as.factor(Data$Channel)
Data$`User Type` <- as.factor(Data$`User Type`)
Data$datefactor <- as.factor(Data$Date)


head(Data)
```

```
## # A tibble: 6 x 9
##   Date                Channel `User Type` Land  Bounce Purchase
##   <dttm>              <fct>   <fct>       <fct> <fct>  <fct>
## 1 2014-10-10 00:00:00 Affili~ Returning ~ 0     0      0
## 2 2014-10-10 00:00:00 Affili~ Returning ~ 1     0      0
## 3 2014-10-10 00:00:00 Affili~ Returning ~ 1     1      0
## 4 2014-10-10 00:00:00 Affili~ Returning ~ 0     0      1
## 5 2014-10-10 00:00:00 Affili~ Returning ~ 1     0      1
## 6 2014-10-10 00:00:00 Affili~ New User    0     0      0
## # ... with 3 more variables: Visitors_Control <dbl>,
## #   Visitors_Variant <dbl>, datefactor <fct>
```

```
#check for missing values
table(is.na(Data))
```

```
##
## FALSE
## 11340
```

```
#no missing values

#summary of data
summary(Data)
```

```
##       Date                    Channel             User Type   Land
##  Min.   :2014-10-10   Affiliate   :210   New User       :630   0:504
##  1st Qu.:2014-10-15   Direct      :210   Returning User:630   1:756
##  Median :2014-10-20   Email       :210
##  Mean   :2014-10-20   Paid Search :210
##  3rd Qu.:2014-10-25   SEO         :210
```

```
##  Max.   :2014-10-30   Social Media:210
##
##  Bounce   Purchase Visitors_Control  Visitors_Variant       datefactor
##  0:1008   0:756    Min.   :   30.0   Min.   :   25.0   2014-10-10: 60
##  1: 252   1:504    1st Qu.:  265.5   1st Qu.:  260.2   2014-10-11: 60
##                    Median : 1091.5   Median : 1070.0   2014-10-12: 60
##                    Mean   : 2378.0   Mean   : 2417.3   2014-10-13: 60
##                    3rd Qu.: 2993.8   3rd Qu.: 2995.5   2014-10-14: 60
##                    Max.   :19938.0   Max.   :19045.0   2014-10-15: 60
##                                                        (Other)   :900
```

Data Analysis

The first step is to look at the difference between conversion and bounce rates of the different subgroups of the test population.

```
#Conversion and bounce rates test
test1 <- subset(Data, Data$Channel == 'Affiliate' & Data$`User Type` == 'Returning User')
head(test1)
```

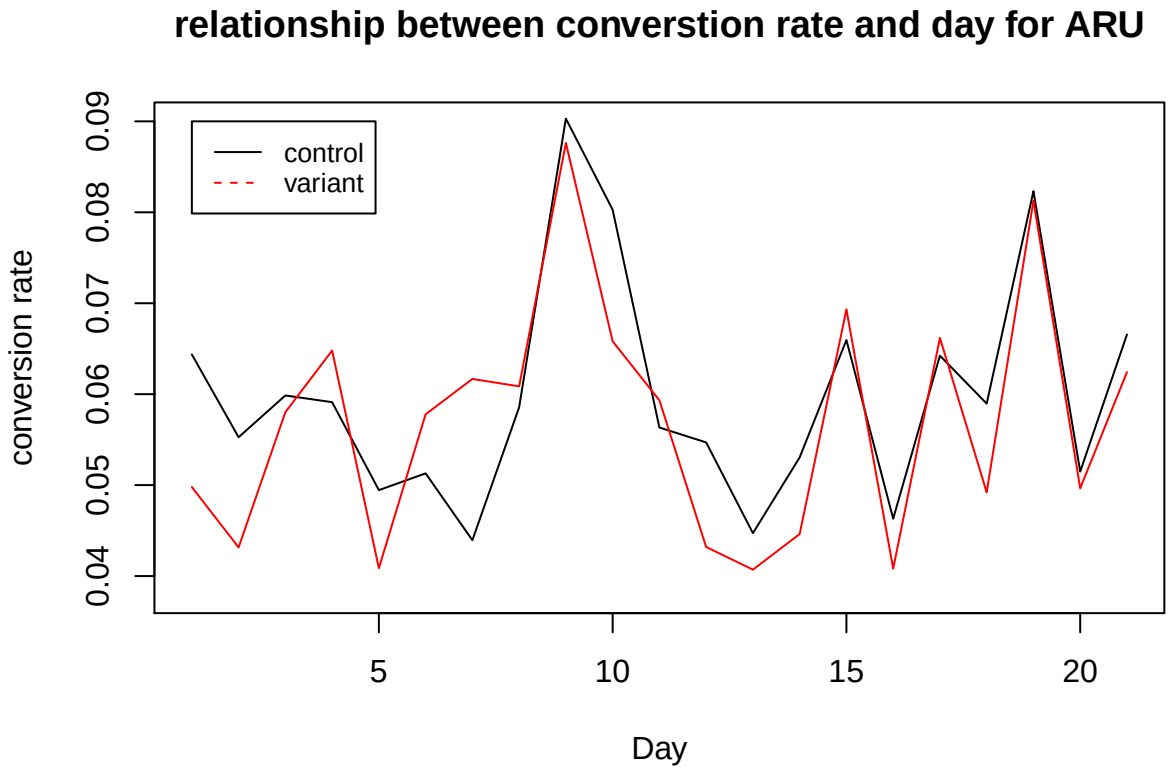
```
## # A tibble: 6 x 9
##   Date                Channel `User Type` Land  Bounce Purchase
##   <dttm>              <fct>   <fct>       <fct> <fct>  <fct>
## 1 2014-10-10 00:00:00 Affili~ Returning ~ 0     0      0
## 2 2014-10-10 00:00:00 Affili~ Returning ~ 1     0      0
## 3 2014-10-10 00:00:00 Affili~ Returning ~ 1     1      0
## 4 2014-10-10 00:00:00 Affili~ Returning ~ 0     0      1
## 5 2014-10-10 00:00:00 Affili~ Returning ~ 1     0      1
## 6 2014-10-11 00:00:00 Affili~ Returning ~ 0     0      0
## # ... with 3 more variables: Visitors_Control <dbl>,
## #   Visitors_Variant <dbl>, datefactor <fct>
```

```
#get conversion and bounce rates for each date
control_rates <- vector()
variant_rates <- vector()
bounce_control <- vector()
bounce_var <- vector()

dates <- unique(as.character(test1$datefactor))
for (x in dates){
  temp <- subset(test1, test1$datefactor == x)
  purchased <- sum(subset(temp, temp$Purchase ==1)$Visitors_Control)
  total <- sum(temp$Visitors_Control)
  crate <- purchased/total
  ctot_land <- sum(subset(temp,temp$Land ==1)$Visitors_Control)

  c_bounced <- sum(subset(temp, temp$Bounce ==1 & temp$Land ==1)$Visitors_Control)
  brate <- c_bounced/ctot_land


  bounce_control <- append(bounce_control, brate)
  control_rates <- append(control_rates, crate)
```

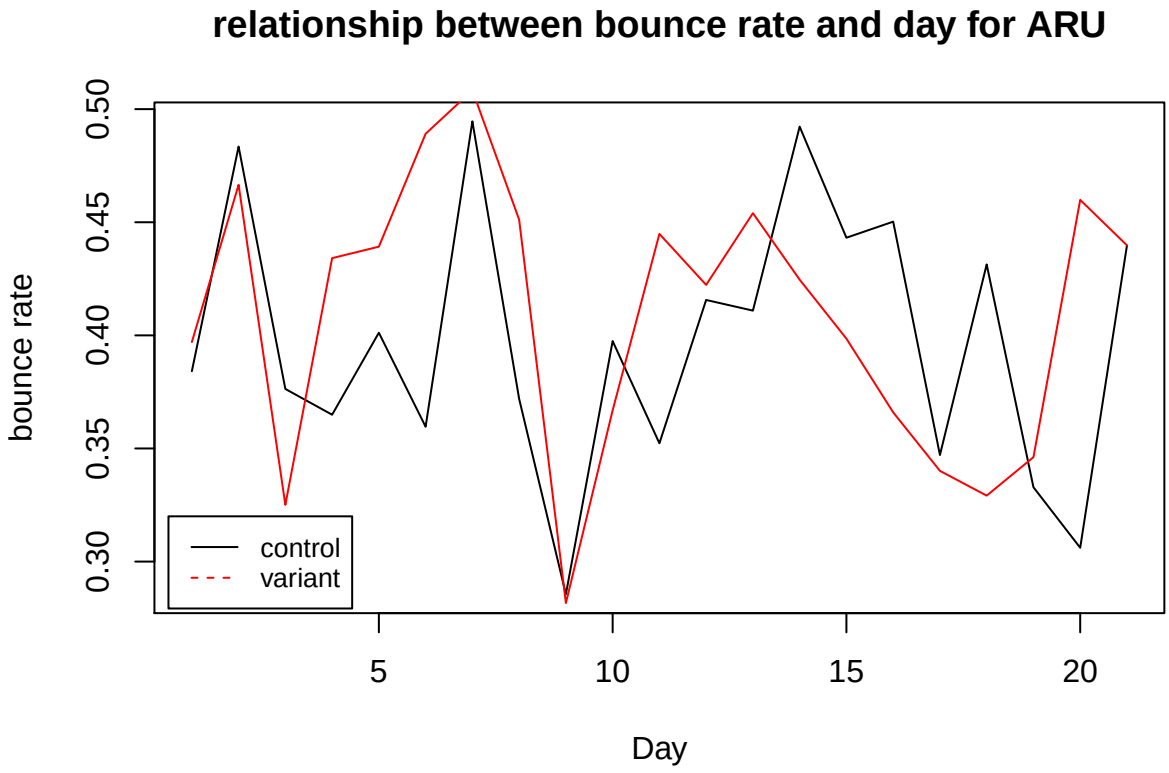
```
  vpurch <- sum(subset(temp, temp$Purchase ==1)$Visitors_Variant)
  vtotal <- sum(temp$Visitors_Variant)
  vtot_land <- sum(subset(temp,temp$Land ==1)$Visitors_Variant)
  vrate <- vpurch/vtotal
  v_bounced <- sum(subset(temp, temp$Bounce ==1  & temp$Land ==1)$Visitors_Variant)
  v_brate <- v_bounced/vtot_land

  variant_rates <- append(variant_rates, vrate)
  bounce_var <- append(bounce_var, v_brate)



}
conv_rate_date <- data.frame(dates, control_rates, variant_rates, bounce_control, bounce_var)
head(conv_rate_date)
```

```
##        dates control_rates variant_rates bounce_control bounce_var
## 1 2014-10-10    0.06436621    0.04978706      0.3841667  0.3970450
## 2 2014-10-11    0.05526316    0.04314905      0.4834544  0.4665536
## 3 2014-10-12    0.05985552    0.05801400      0.3763095  0.3250678
## 4 2014-10-13    0.05912197    0.06479535      0.3649114  0.4341392
## 5 2014-10-14    0.04944941    0.04086739      0.4011586  0.4392036
## 6 2014-10-15    0.05129241    0.05778689      0.3596111  0.4890956
```

```
#plotting two lines for the conversion rates
plot(as.numeric(conv_rate_date$dates), conv_rate_date$control_rates, type = 'l', xlab = 'Day', ylab = '
lines(conv_rate_date$dates, variant_rates, col = 'red')
legend(1, .09, legend=c("control", "variant"),
       col=c("black", "red"), lty=1:2, cex=0.8)
```




```
#plotting two lines for the bounce rates
plot(as.numeric(conv_rate_date$dates), conv_rate_date$bounce_control, type = 'l', xlab = 'Day', ylab =
lines(conv_rate_date$dates, bounce_var, col = 'red')
legend(.5, .32, legend=c("control", "variant"),
       col=c("black", "red"), lty=1:2, cex=0.8)
```




```
#conducting t-tests
t.test(conv_rate_date$control_rates, conv_rate_date$variant_rates, alternative = 'less')
```

```
##
##  Welch Two Sample t-test
##
## data:  conv_rate_date$control_rates and conv_rate_date$variant_rates
## t = 0.72783, df = 39.792, p-value = 0.7645
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##         -Inf 0.009460455
## sample estimates:
##  mean of x  mean of y
## 0.05985879 0.05700394
```

```
t.test(conv_rate_date$bounce_control, conv_rate_date$bounce_var, alternative = 'greater')
```

```
##
##  Welch Two Sample t-test
##
## data:  conv_rate_date$bounce_control and conv_rate_date$bounce_var
## t = -0.63537, df = 39.952, p-value = 0.7356
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.04233294         Inf
```

```
## sample estimates:
## mean of x mean of y
## 0.3971858 0.4087831
```

Now I have to do the same thing but with every possible combination of classes for the users

```
#initializing empty arrays
total_control <- vector()
total_variant <- vector()
total_cbounce <- vector()
total_vbounce <- vector()
category <- vector()
conv_p_values <- vector()
bounce_p_values <- vector()
channel_names <- unique(as.character(Data$Channel))
user_types <- unique(as.character(Data$`User Type`))

#running on each combination
for (x in channel_names){
  for (y in user_types){

  for (d in dates){

    temporary_table <- subset(Data, Data$Channel == x & Data$`User Type` == y & Data$datefactor == d)
    purch <- sum(subset(temporary_table, temporary_table$Purchase ==1)$Visitors_Control)
    tot <- sum(temporary_table$Visitors_Control)
    c_rate <- purch/tot
    c_t_land <- sum(subset(temporary_table,temporary_table$Land ==1)$Visitors_Control)

    cb <- sum(subset(temporary_table, temporary_table$Bounce ==1 & temporary_table$Land == 1)$Visitors_
    br <- cb/c_t_land

    total_cbounce <- append(total_cbounce, br)
    total_control <- append(total_control, c_rate)

    v_purch <- sum(subset(temporary_table, temporary_table$Purchase ==1)$Visitors_Variant)
    v_total <- sum(temporary_table$Visitors_Variant)
    v_t_land <- sum(subset(temporary_table,temporary_table$Land ==1)$Visitors_Variant)
    v_rate <- v_purch/v_total
    total_variant <- append(total_variant, v_rate)

    vb <- sum(subset(temporary_table, temporary_table$Bounce ==1 & temporary_table$Land == 1)$Visitors_
    vbr <- vb/v_t_land
    total_vbounce <- append(total_vbounce, vbr)

  }
    type_user <- paste(x,y, sep = ' ')
    category <- append(category, type_user)

    p <- t.test(total_control, total_variant, alternative = 'less')$p.value
    bp_val <- t.test(total_cbounce, total_vbounce, alternative = 'greater')$p.value
    conv_p_values <- append(conv_p_values, p)
    bounce_p_values <- append(bounce_p_values, bp_val)
```

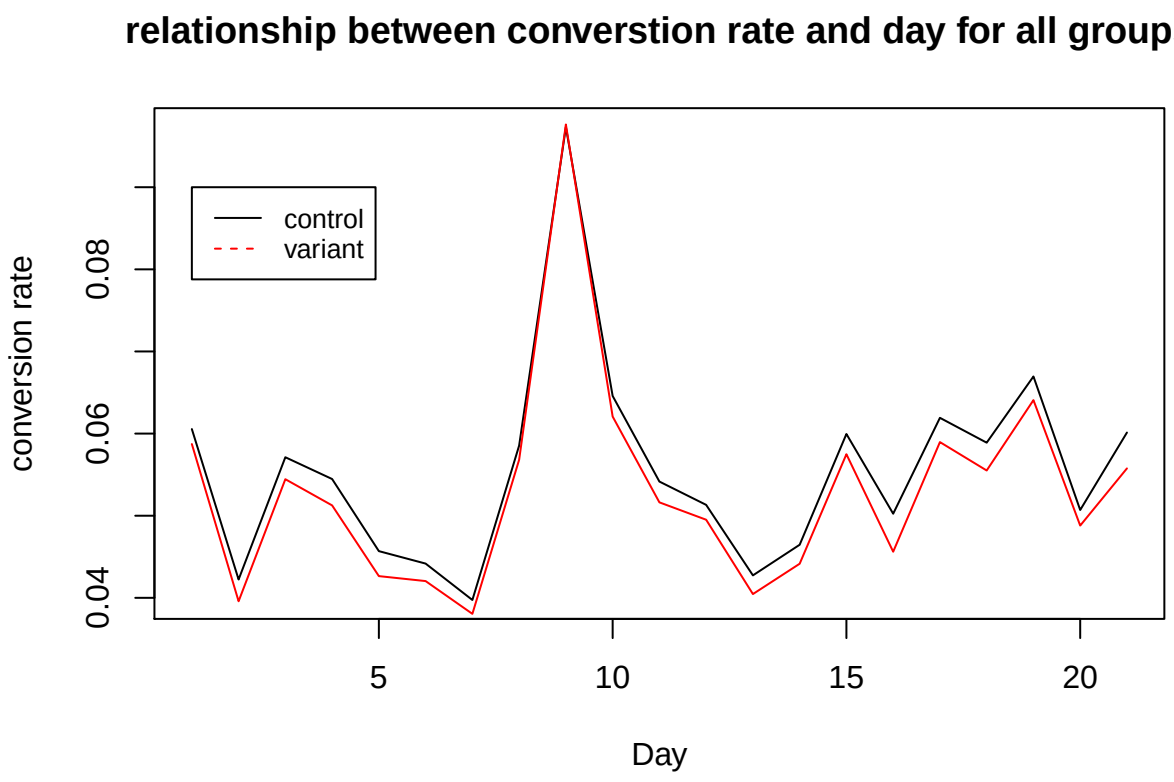
```
    total_control <- vector()
    total_variant <- vector()
    total_cbounce <- vector()
    total_vbounce <- vector()
}
}

data.frame(category, conv_p_values, bounce_p_values)
```

```
##                       category conv_p_values bounce_p_values
## 1     Affiliate Returning User     0.7645104       0.7355960
## 2           Affiliate New User     0.3428622       0.6828296
## 3        Direct Returning User     0.7263521       0.9055620
## 4             Direct New User     0.8121440       0.8352863
## 5         Email Returning User     0.6862309       0.8549620
## 6               Email New User     0.7729310       0.8022885
## 7   Paid Search Returning User     0.9246129       0.7290684
## 8         Paid Search New User     0.4653839       0.8715967
## 9           SEO Returning User     0.7907829       0.5978817
## 10                SEO New User     0.6948196       0.8235721
## 11 Social Media Returning User     0.5486956       0.9529502
## 12       Social Media New User     0.4828771       0.2967265
```

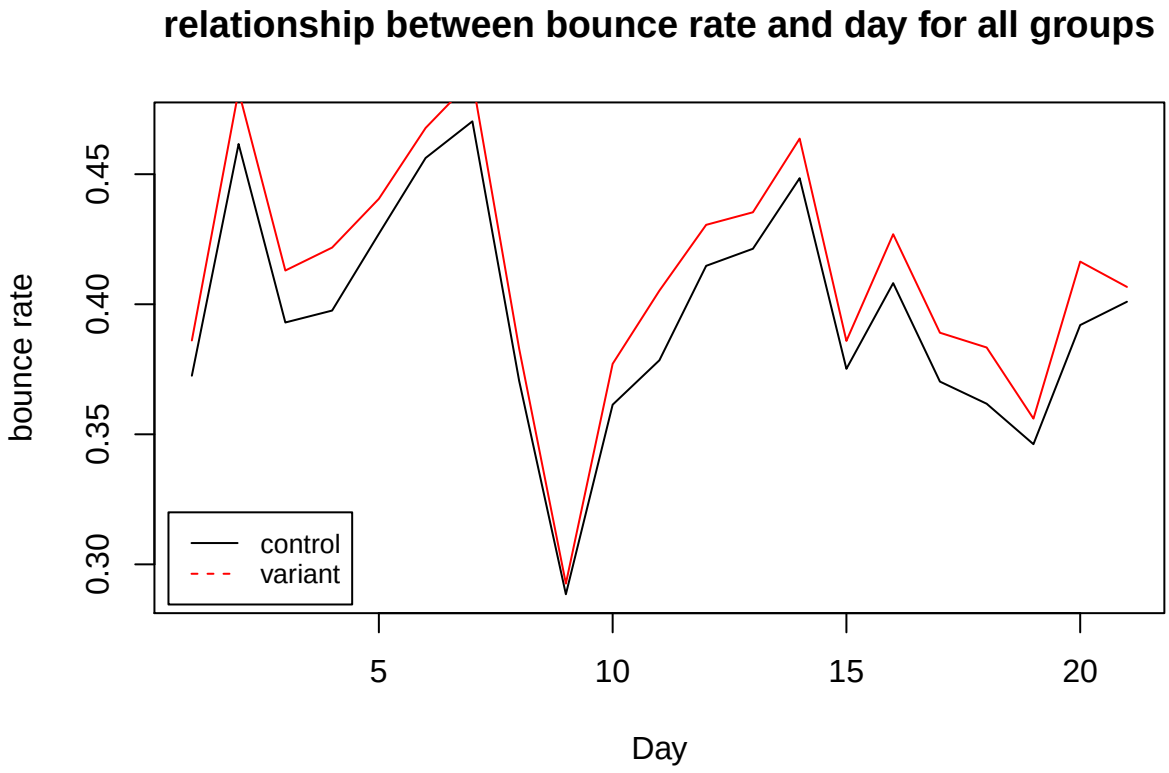
Extra Agg Analysis

```
test1 = Data


dates <- unique(as.character(test1$datefactor))

control_rates <- vector()
variant_rates <- vector()
bounce_control <- vector()
bounce_var <- vector()
for (x in dates){
  temp <- subset(test1, test1$datefactor == x)
  purchased <- sum(subset(temp, temp$Purchase ==1)$Visitors_Control)
  total <- sum(temp$Visitors_Control)
  crate <- purchased/total
  ctot_land <- sum(subset(temp,temp$Land ==1)$Visitors_Control)

  c_bounced <- sum(subset(temp, temp$Bounce ==1 & temp$Land ==1)$Visitors_Control)
  brate <- c_bounced/ctot_land


  bounce_control <- append(bounce_control, brate)
  control_rates <- append(control_rates, crate)

  vpurch <- sum(subset(temp, temp$Purchase ==1)$Visitors_Variant)
  vtotal <- sum(temp$Visitors_Variant)
```

```
  vtot_land <- sum(subset(temp,temp$Land ==1)$Visitors_Variant)
  vrate <- vpurch/vtotal
  v_bounced <- sum(subset(temp, temp$Bounce ==1  & temp$Land ==1)$Visitors_Variant)
  v_brate <- v_bounced/vtot_land

  variant_rates <- append(variant_rates, vrate)
  bounce_var <- append(bounce_var, v_brate)



}
conv_rate_date <- data.frame(dates, control_rates, variant_rates, bounce_control, bounce_var)

#plotting two lines for the conversion rates
plot(as.numeric(conv_rate_date$dates), conv_rate_date$control_rates, type = 'l', xlab = 'Day', ylab = '
lines(conv_rate_date$dates, variant_rates, col = 'red')
legend(1, .09, legend=c("control", "variant"),
       col=c("black", "red"), lty=1:2, cex=0.8)
```




```
#plotting two lines for the bounce rates
plot(as.numeric(conv_rate_date$dates), conv_rate_date$bounce_control, type = 'l', xlab = 'Day', ylab =
lines(conv_rate_date$dates, bounce_var, col = 'red')
legend(.5, .32, legend=c("control", "variant"),
       col=c("black", "red"), lty=1:2, cex=0.8)
```




```
#conducting t-tests
t.test(conv_rate_date$control_rates, conv_rate_date$variant_rates)
```

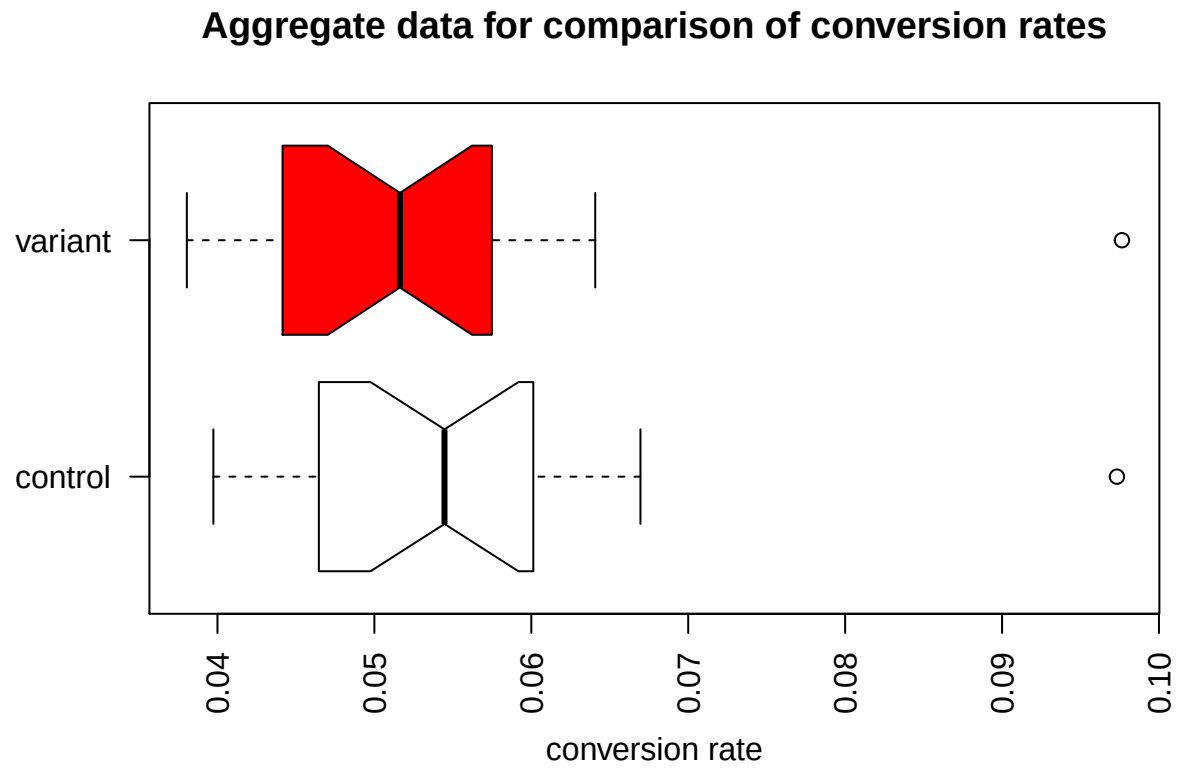
```
##
##  Welch Two Sample t-test
##
## data:  conv_rate_date$control_rates and conv_rate_date$variant_rates
## t = 0.64584, df = 39.954, p-value = 0.5221
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.005339713  0.010354774
## sample estimates:
##  mean of x  mean of y
## 0.05561333 0.05310580
```

```
t.test(conv_rate_date$bounce_control, conv_rate_date$bounce_var)
```

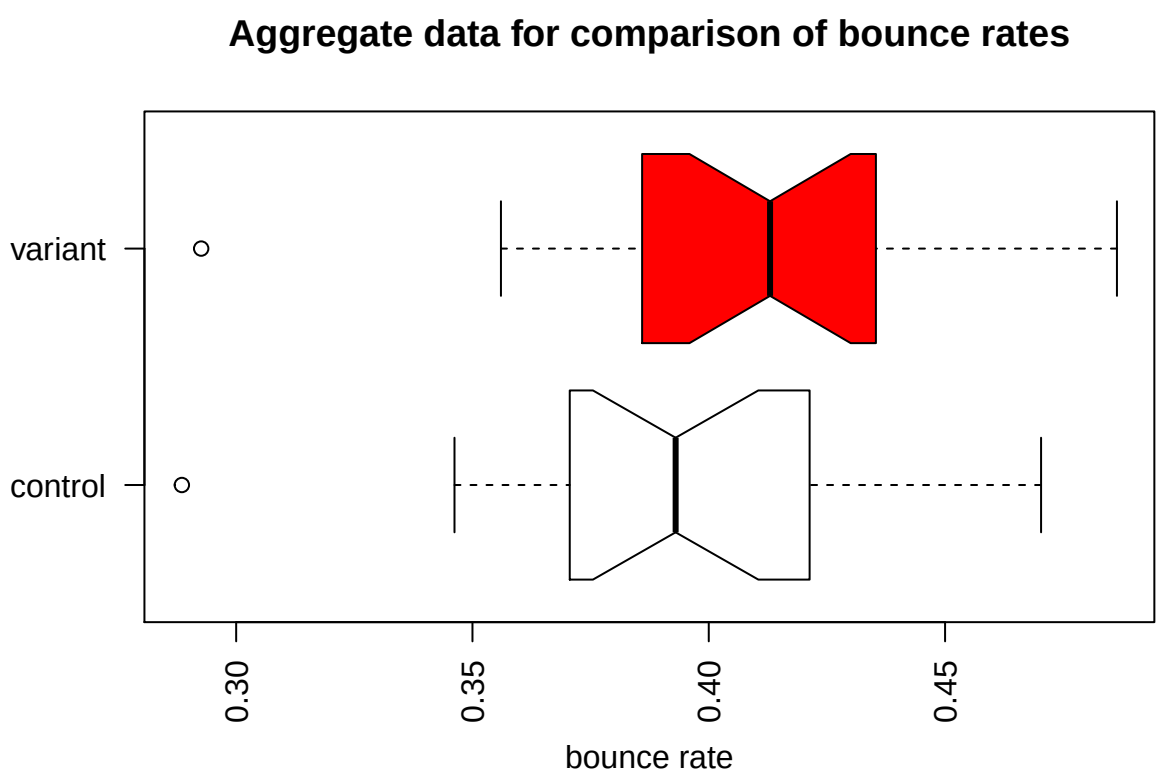
```
##
##  Welch Two Sample t-test
##
## data:  conv_rate_date$bounce_control and conv_rate_date$bounce_var
## t = -1.1648, df = 39.936, p-value = 0.251
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.04337822  0.01166030
```

```
## sample estimates:
## mean of x mean of y
## 0.3960114 0.4118704
```

```
#boxplots of data
boxplot(control_rates, variant_rates,
main = "Aggregate data for comparison of conversion rates",
at = c(1,2),
names = c("control", "variant"),
las = 2,
col = c("white","red"),
border = "black",
horizontal = TRUE,
notch = TRUE, xlab = 'conversion rate'
)
```




```
boxplot(bounce_control, bounce_var,
main = "Aggregate data for comparison of bounce rates",
at = c(1,2),
names = c("control", "variant"),
las = 2,
col = c("white","red"),
border = "black",
horizontal = TRUE,
notch = TRUE, xlab = 'bounce rate'
)
```



Part 1 Conclusion

I chose a t-test to look at the difference in means for both the control and the variant group to see if there was a difference in the average bounce and conversion rates for each cohort of users.

Based on the above t-tests that we ran we can see that for every group combination there is not a statistically significant (alpha = .05) difference that exposing users to the variant either increased our conversion rate, or decreased the bounce rate. Using this information we should not immediately switch to the variant in favor of the control.

Since this is the case I would recommend looking at the amount of time and work necessary to integrate this variant into our existing business and see whether or not it is worth doing. It does make more sense to show customers the events that are close to them and occuring relatively soon, but the data does not support that this makes an actual difference, and in fact on an aggregate level, customers who were shown the variant had lower conversion rates and higher bounce rates.

I would like to run this experiment for a little longer to see if we can notice a substantial change as it only ran for less than a month, so with some additional time maybe more information will become available and we can be more sure of the results. On day number 9 of the trial there was an extremely high conversion rate and an extremely low bounce rate so this could be something that is worth looking into.

It would also be interesting to see how much the pages really differed between the variant and the control, if there was little to no difference in the events that were being shown to the users then it makes sense that there would not be a noticeable difference as the pages being shown to both groups would be essentially the same. I also would like to see if there was a difference in the amount of time spent on the platform and to see if the variant had an effect on that.

So in conclusion, the variant neither increased the conversion rate or decreased the bounce rate for any group by a statistically significant amount. With that being said, it does make sense to show users events that are closer to them and are occuring relatively soon, as that may increase their likelihood to buy a ticket due to convenience. I would consider running this test for an addional length of time to see if the results change, and I would look to see how much the pages truly differ in content.

Part 2

1. List five potential improvements you would make to the current page. Explain your reasoning.

Include the price of the tickets such as 'as low as _____', this could draw attention to the event as people might have previously dismissed it as being too expensive.

Track user behavior and show them events similar to ones that they purchased previously. For example, if they recently purchased country music tickets, I would show them and highlight another country music song.

Enable people to 'favorite' an artist or a team so that those events show up at the top of the queue. Some people are willing to travel to events that are far from them just because they are die hard fans, highlighting events that they would be more interested in would give us a chance to increase sales.

Include the number of tickets available for the event, this implies scarcity and could incentivise people to purchase tickets if they were on the fence.

Maybe instead of all of these individual events we could display categories (concerts, sporting events, etc), this would allow users to navigate to their desired attraction faster without scrolling through the website.

2. What additional data would you like to know to help you assess which idea to prioritize?

The cost and manpower required to implement each of these technologies and the predicted amount that it could affect business.

3. How would you measure the success or failure of each of these changes?

Conversion rate, looking at the increase in the sales/revenue, bounce rate, average length of time spent on the platform, clickthrough rate, referral rate.