Rocket Fuel and TaskaBella

Alex Osterbuhr

9/17/23

TaskaBella Advertising Campaign Effectiveness

This analysis examines the effectiveness of TaskaBella's advertising campaign.

Data Preparation and Exloration

Load and explore the TaskaBella data.

```
# Read data
#taskaBella_orig <- read_excel("B5896-XLS-ENG.xlsx",sheet=1)

taskaBella_orig <- read_excel("Rocket Fuel Spreadsheet.xlsx", sheet = 1)
```

user id: Unique identifier of the user

test: Whether the user was exposed to advertising or was in the control group. 1 if the user was exposed to the real ad, 0 if the user was in the control group and was shown a PSA.

converted: Whether the user converted. 1 if the user bought the handbag during the campaign, 0 if not.

tot_impr: The total number of ad impressions the user encountered. For users in the control group this counts the number of times they encountered the PSA.For exposed users it counts the number of times they were shown the ad.

mode_impr_day: Shows the day of the week on which the user encountered the most number of impressions. 1 means Monday, 7 means Sunday. For example if a given user encountered 2 impressions on Mondays, 3 on Tuesdays, 7 on Wednesdays, 0 on Thursdays and, Fridays, 9 on Saturdays and 2 on Sundays, this column takes the value of 6 (Saturday).

mode_impr_hour: Shows the hour of the day (0-23) in which the user encountered the most number of impressions.

```
# Explore the data
str(taskaBella_orig)
```

```
## tibble [588,101 × 6] (S3: tbl_df/tbl/data.frame)
##  $ user_id       : num [1:588101] 1069124 1119715 1144181 1435133 1015700 ...
##  $ test          : num [1:588101] 1 1 1 1 1 1 1 1 1 1 ...
##  $ converted     : num [1:588101] 0 0 0 0 0 0 0 0 0 0 ...
##  $ tot_impr      : num [1:588101] 130 93 21 355 276 734 264 17 21 142 ...
##  $ mode_impr_day : num [1:588101] 1 2 2 2 5 6 3 7 2 1 ...
##  $ mode_impr_hour: num [1:588101] 20 22 18 10 14 10 13 18 19 14 ...
```

```
summary(taskaBella_orig)
```

```
##     user_id             test        converted          tot_impr      
##  Min.   : 900000   Min.   :0.00   Min.   :0.00000   Min.   :   1.00  
##  1st Qu.:1143190   1st Qu.:1.00   1st Qu.:0.00000   1st Qu.:   4.00  
##  Median :1313725   Median :1.00   Median :0.00000   Median :  13.00  
##  Mean   :1310692   Mean   :0.96   Mean   :0.02524   Mean   :  24.82  
##  3rd Qu.:1484088   3rd Qu.:1.00   3rd Qu.:0.00000   3rd Qu.:  27.00  
##  Max.   :1654483   Max.   :1.00   Max.   :1.00000   Max.   :2065.00  
##  mode_impr_day   mode_impr_hour 
##  Min.   :1.000   Min.   : 0.00  
##  1st Qu.:2.000   1st Qu.:11.00  
##  Median :4.000   Median :14.00  
##  Mean   :4.026   Mean   :14.47  
##  3rd Qu.:6.000   3rd Qu.:18.00  
##  Max.   :7.000   Max.   :23.00
```

After exploring the data, all columns imported as numeric, but this is not true. We must identify the user_id column as an identifier, the test binary column must be converted to identify which user was in the test or control group, and the mode_impr_day column to identify days of the week easily. Also, I am noting there are potential outliers in the tot_impr column given the values in summary.

1. Advertising Effectiviness

Assess whether or not the campaign was effective, and provide justification using the analytical method of your choice. Consider whether or not additional customers converted as a result of the

campaign.

```
#Create new clear features that mirror the test and converted binary features into their true me
aning
taskaBella_orig$usertype <- ifelse(taskaBella_orig$test == 0, "Control", "Exposed")
taskaBella_orig$userbuy <- ifelse(taskaBella_orig$converted == 0, "No", "Yes")

#set days to new feature mirroring the mode_impr_day feature and it's numerical meaning
days <- c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday", "Sunday")
taskaBella_orig$dayofweek <- factor(days[taskaBella_orig$mode_impr_day], levels = days)

#set the above usertype and userbuy features to factors for leveling purposes for later use
taskaBella_orig$usertype <- factor(taskaBella_orig$usertype, levels = c("Control", "Exposed"))
taskaBella_orig$userbuy <- factor(taskaBella_orig$userbuy, levels = c("No", "Yes"))
```

I will run a t-test of usertype by converted to see the conversion rate. Null hypothesis is no change or increase of "Exposed" conversion compared to "Control". Hypothesis is there will be an increased conversion rate in "Exposed" compared to "Control".

```
t.test(taskaBella_orig$converted ~ taskaBella_orig$usertype, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  taskaBella_orig$converted by taskaBella_orig$usertype
## t = -7.3704, df = 588099, p-value = 0.0000000000001703
## alternative hypothesis: true difference in means between group Control and group Exposed is n
ot equal to 0
## 95 percent confidence interval:
##  -0.009738061 -0.005646845
## sample estimates:
## mean in group Control mean in group Exposed
##            0.01785411            0.02554656
```

The campaign was effective and we reject the null hypothesis given the p-value of the t test is substantially less than .05. To confirm this, I would like to identify total number of users in each group and exact count of converted users.

```
total_count_type <- taskaBella_orig %>%
  count(usertype) %>%
  rename(total = n)

total_group_convert <- taskaBella_orig %>%
  group_by(usertype)%>%
  count(userbuy) %>%
  rename(count = n)

total_count_type
```

```
## # A tibble: 2 × 2
##   usertype  total
##   <fct>     <int>
## 1 Control   23524
## 2 Exposed  564577
```

```
total_group_convert
```

```
## # A tibble: 4 × 3
## # Groups:   usertype [2]
##   usertype userbuy  count
##   <fct>    <fct>    <int>
## 1 Control  No       23104
## 2 Control  Yes        420
## 3 Exposed  No      550154
## 4 Exposed  Yes      14423
```

```
total_count_type %>%
  left_join(total_group_convert, by = 'usertype') %>%
  mutate(percentage_total = round((count / total) *100,2)) %>%
  select(!total)
```

```
## # A tibble: 4 × 4
##   usertype userbuy  count percentage_total
##   <fct>    <fct>    <int>            <dbl>
## 1 Control  No       23104            98.2
## 2 Control  Yes        420             1.79
## 3 Exposed  No      550154            97.4
## 4 Exposed  Yes      14423             2.55
```

In terms of percentage of each group converted, the exposed group had around .76% more conversions. Therefore, the ad campaign seems to indeed be more effective among the Exposed users, even if by a small margin.

```
#Estimate of customers who would have converted due to campaign using percent differences and to
tal exposed group
additional_converted <- round(((2.55 - 1.79)/ 100)* 564577, 0)
additional_converted
```

```
## [1] 4291
```

2. Profitability

Did TaskaBella make additional money, excluding advertising

cost? If so, how much more?

```
#We will take the percent difference stated above times total number of exposed users times $40 per conversion.
dollars_made <- additional_converted*40

paste('Additional money made: $',dollars_made, sep = '')
```

```
## [1] "Additional money made: $171640"
```

What was the cost of the campaign?

```
total_impressions <- sum(taskaBella_orig$tot_impr)
#$9 is average cost per 1000 impressions

campaign_cost <- round((total_impressions/1000) *9, 2)

paste('Ad Campaign Cost: $',campaign_cost, sep='')
```

```
## [1] "Ad Campaign Cost: $131374.64"
```

What was the ROI of the campaign?

```
#using total profit over cost times 100 (for a percentage)
profit <- dollars_made - campaign_cost

roi_percent <- (profit/campaign_cost)*100
paste('Total money made: $', profit, sep='')
```

```
## [1] "Total money made: $40265.36"
```

```
paste('ROI percentage: ', round(roi_percent,2), '%', sep='')
```

```
## [1] "ROI percentage: 30.65%"
```

This was essentially an A/B test. What was the opportunity cost of the control group?

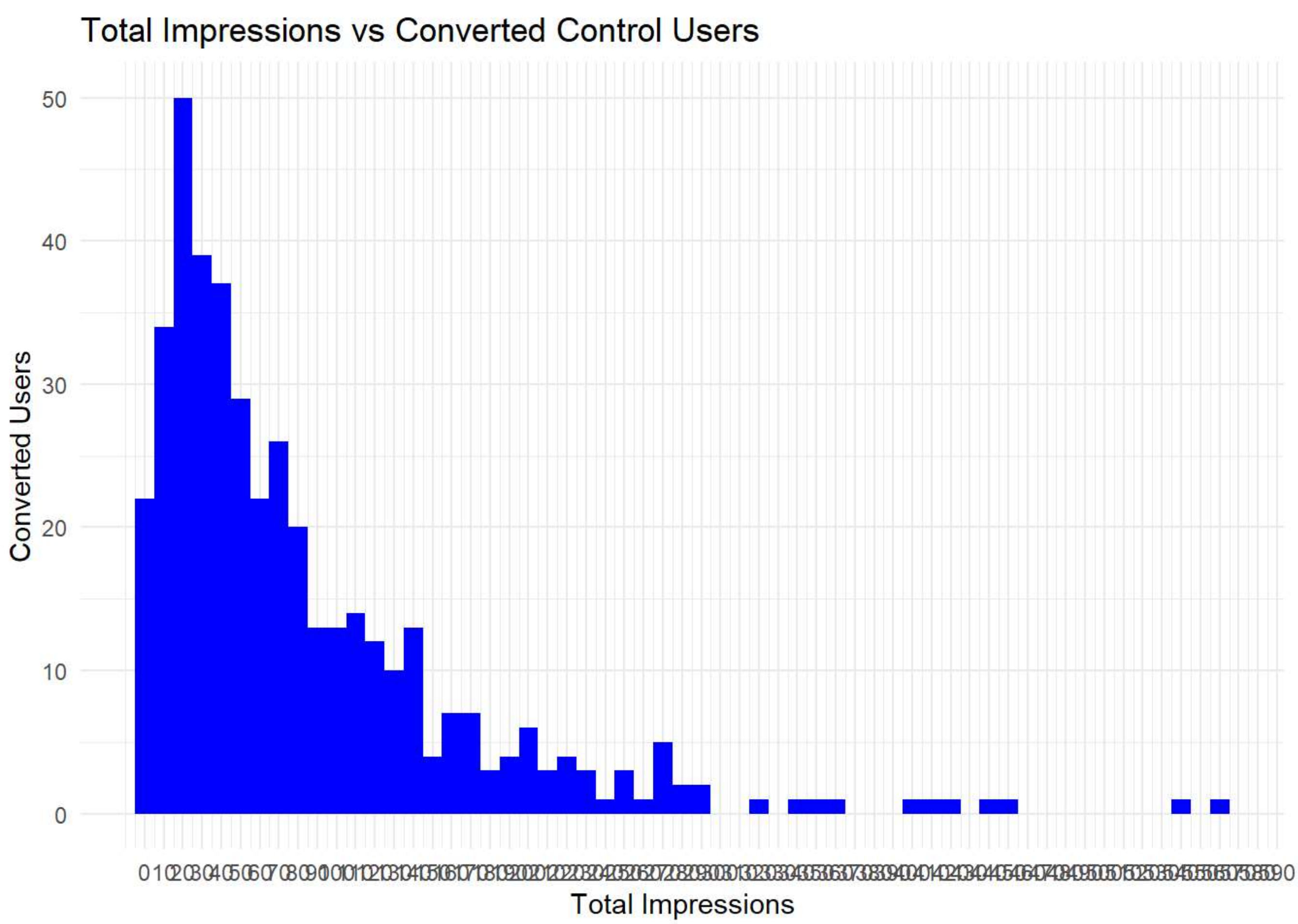
```
#Take our percent difference from above and multiply by total control group and the $40 their conversion would bring
opp_cost <- round(((2.55-1.79)/100)*23524*40, 2)
paste("Opportunity cost of the control group: $", opp_cost, sep='')
```

```
## [1] "Opportunity cost of the control group: $7151.3"
```

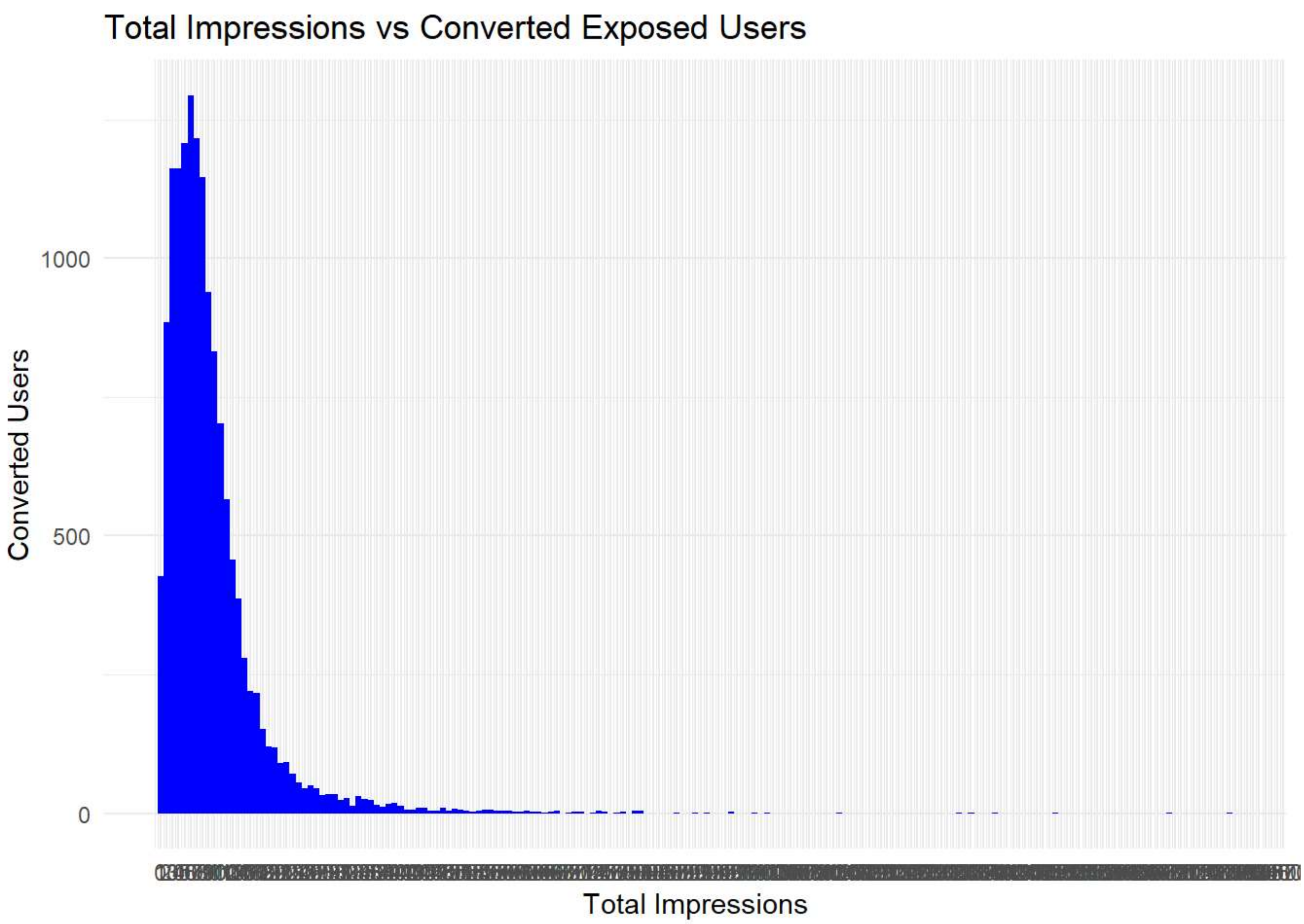
3. Impressions and Effectiveness

Create a chart of conversion rates as a function of the number of ads shown to users for both the control and experimental groups. Consider impressions in 10-unit chunks (1-10, 11-20, etc.). Keep in mind that conversion rate equates to the percentage of unique users who made a purchase.

```
#Histogram of Control conversions in bins by 10
taskaBella_orig %>%
 filter(userbuy %in% "Yes") %>%
  filter(usertype %in% 'Control') %>%
 ggplot( aes(x = tot_impr)) +
 geom_histogram(binwidth = 10, fill = "blue") +
 labs(x = "Total Impressions", y = "Converted Users", title = "Total Impressions vs Converted Co
ntrol Users")+
  scale_x_continuous(breaks = seq(0, max(taskaBella_orig$tot_impr) +10, by = 10))+
 theme_minimal()
```

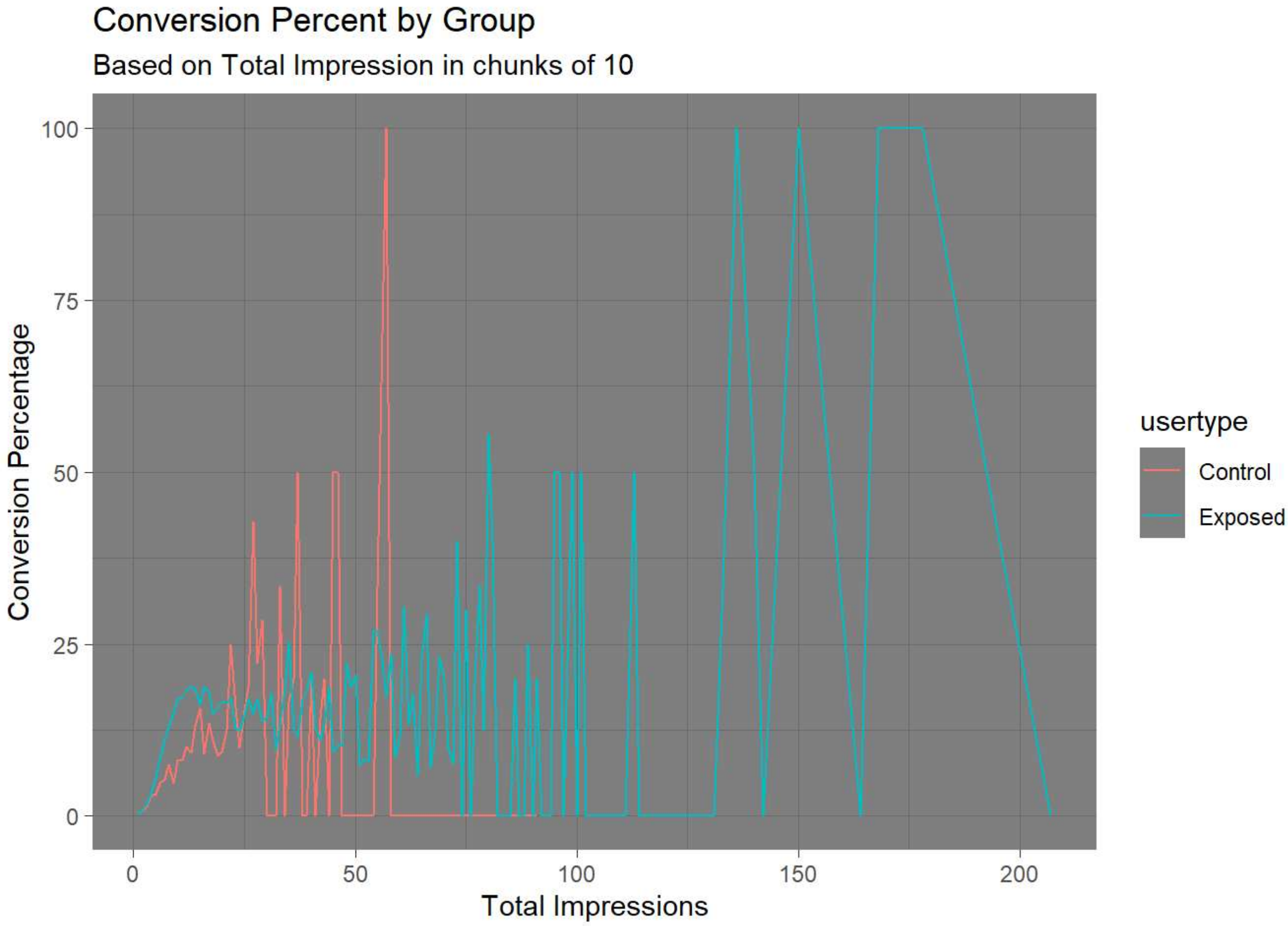


```
#Histogram of Control conversions in bins by 10
taskaBella_orig %>%
 filter(userbuy %in% "Yes") %>%
 filter(usertype %in% 'Exposed') %>%
 ggplot( aes(x = tot_impr)) +
 geom_histogram(binwidth = 10, fill = "blue") +
 labs(x = "Total Impressions", y = "Converted Users", title = "Total Impressions vs Converted Ex
posed Users")+
  scale_x_continuous(breaks = seq(0, max(taskaBella_orig$tot_impr) +10, by = 10))+
 theme_minimal()
```




```
#Using cut to convert tot_impr to a factor and create bins by 10 then plot to a line

taskaBella_orig %>%
  group_by(usertype, tot_impr = cut(tot_impr, breaks = seq(0, max(taskaBella_orig$tot_impr) +10,
by = 10))) %>%
  summarise(conv_rate = sum(userbuy == "Yes") / n_distinct(user_id)*100) %>%
  ggplot(aes(x = as.numeric(tot_impr), y = conv_rate))+
  geom_line(aes(color = usertype))+
  labs(x= 'Total Impressions', y= 'Conversion Percentage', title = 'Conversion Percent by Grou
p', subtitle = 'Based on Total Impression in chunks of 10')+
  theme_dark()
```



Consider the standard errors for each 10-unit impression grouping.

Where does the impression count make the biggest difference?

```
taskaBella_orig$impressionrange <- cut(taskaBella_orig$tot_impr,
                                    breaks = seq(0, 2070, by = 10),
                                    labels = paste0(seq(0, 2060, by = 10), "-", seq(9, 2070, by
= 10)))

impressionbuckets <- group_by(taskaBella_orig, impressionrange)

taskaBella_orig_impressions <- summarise(impressionbuckets, Users = n(), Conversions = sum(conve
rted), TotalImpressions = sum(tot_impr))

taskaBella_orig_impressions <-taskaBella_orig_impressions %>%
  group_by(impressionrange)

taskaBella_orig_impressions$conversion_rate <- (taskaBella_orig_impressions$Conversions / taskaB
ella_orig_impressions$Users) * 100

taskaBella_orig_impressions <- taskaBella_orig_impressions %>%
  arrange(desc(conversion_rate))

taskaBella_orig_impressions
```

```
## # A tibble: 124 × 5
## # Groups:   impressionrange [124]
##    impressionrange Users Conversions TotalImpressions conversion_rate
##    <fct>           <int>       <dbl>            <dbl>           <dbl>
##  1 1350-1359           1           1             1354           100
##  2 1490-1499           1           1             1491           100
##  3 1670-1679           1           1             1680           100
##  4 1770-1779           1           1             1778           100
##  5 790-799             9           5             7161            55.6
##  6 940-949             2           1             1897            50
##  7 950-959             2           1             1907            50
##  8 980-989             2           1             1969            50
##  9 1000-1009           2           1             2013            50
## 10 1120-1129           2           1             2257            50
## # … with 114 more rows
```

Seen here, some impression ranges had a 100% conversion rate because there was only one user in those ranges. Below, I create this same table, but where the amount of users in a range is above 100, and then over 200. Depending on your goal, this filter needs to be set appropriately.

```
taskaBella_orig_impressions %>%
  filter(Users > 100) %>%
  arrange(desc(conversion_rate))
```

```
## # A tibble: 37 × 5
## # Groups:   impressionrange [37]
##    impressionrange Users Conversions TotalImpressions conversion_rate
##    <fct>           <int>       <dbl>            <dbl>           <dbl>
##  1 340-349           137          34            47273            24.8
##  2 120-129          2398         442           300592            18.4
##  3 150-159          1274         234           198055            18.4
##  4 130-139          1882         344           254556            18.3
##  5 110-119          2913         526           335785            18.1
##  6 160-169          1094         196           180843            17.9
##  7 330-339           140          25            46932            17.9
##  8 210-219           541          95           116498            17.6
##  9 270-279           248          43            68294            17.3
## 10 250-259           296          51            75541            17.2
## # … with 27 more rows
```

```
taskaBella_orig_impressions %>%
  filter(Users > 200) %>%
  arrange(desc(conversion_rate))
```

```
## # A tibble: 30 × 5
## # Groups:   impressionrange [30]
##    impressionrange Users Conversions TotalImpressions conversion_rate
##    <fct>           <int>       <dbl>            <dbl>           <dbl>
##  1 120-129          2398         442           300592            18.4
##  2 150-159          1274         234           198055            18.4
##  3 130-139          1882         344           254556            18.3
##  4 110-119          2913         526           335785            18.1
##  5 160-169          1094         196           180843            17.9
##  6 210-219           541          95           116498            17.6
##  7 270-279           248          43            68294            17.3
##  8 250-259           296          51            75541            17.2
##  9 100-109          3649         617           384466            16.9
## 10 90-99            4731         783           450841            16.6
## # … with 20 more rows
```

Based on this information, I think the 340-349 range is the optimal range but needs to be explored a bit more. 120-129 and 150-159 are safe ranges to say the campaign makes the biggest difference.

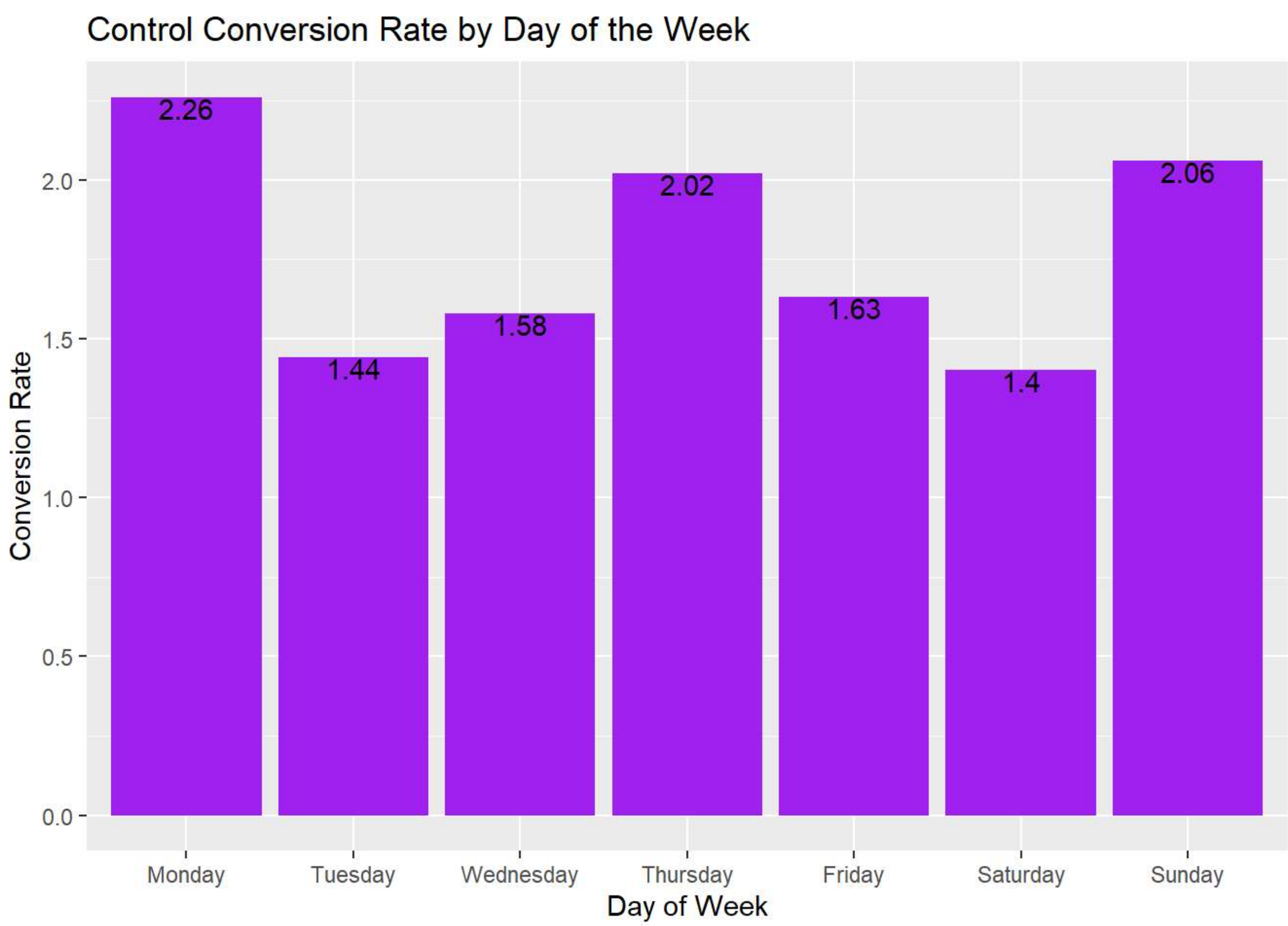
4. Time and Effectiveness

Create a chart showing conversion rates for the control and

exposed groups by day of week.

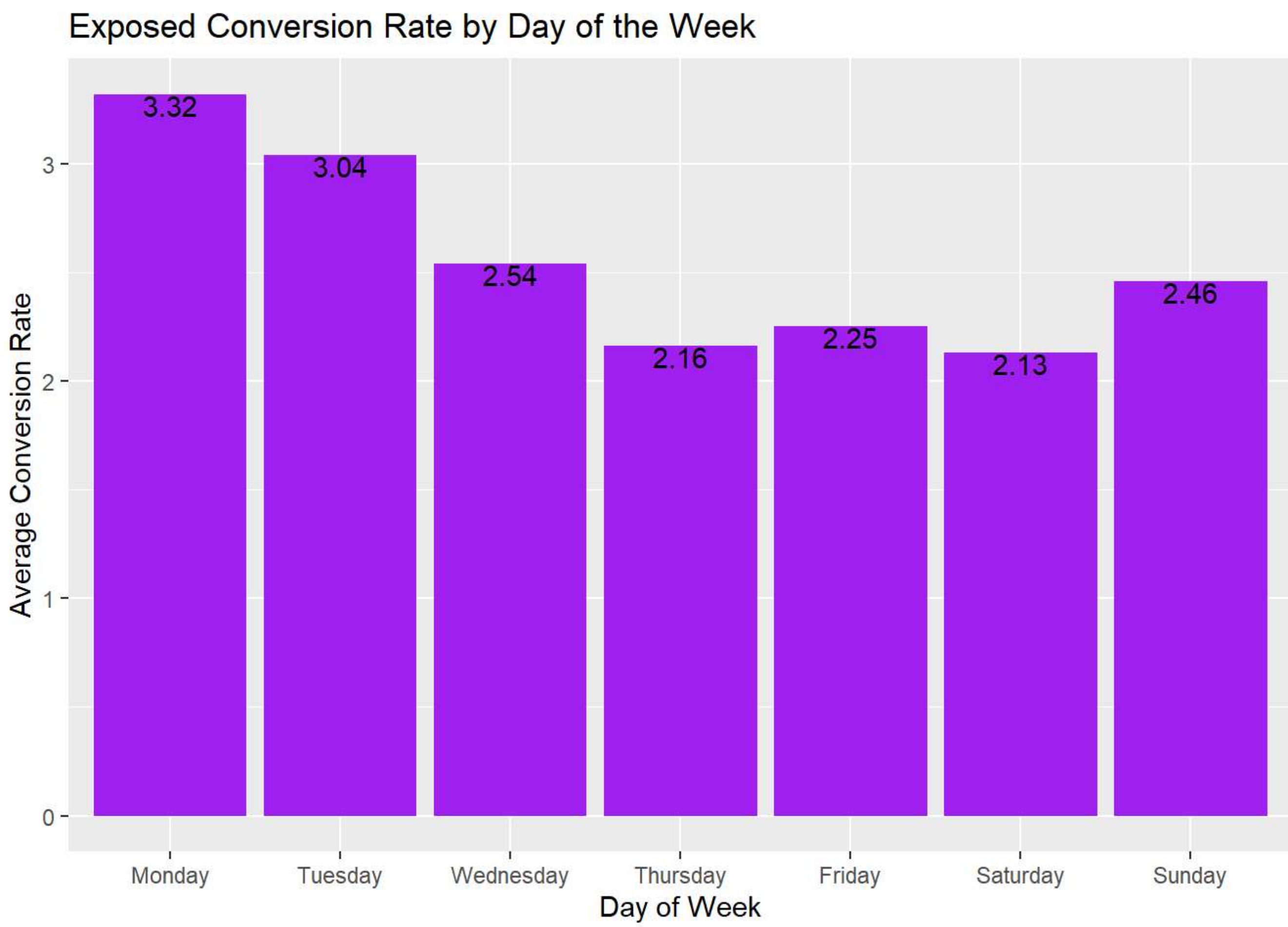
```
Control_by_day <- taskaBella_orig %>%
  filter(usertype %in% 'Control') %>%
  group_by(dayofweek) %>%
  mutate(user_count_day = length(user_id)) %>%
  mutate(conversions_day = sum(converted)) %>%
  select(usertype, dayofweek, user_count_day, conversions_day) %>%
  distinct(dayofweek, .keep_all = TRUE) %>%
  mutate(rate_by_day = round((conversions_day/user_count_day)*100,2))


Control_by_day %>%
  ggplot(aes(x = dayofweek, y = rate_by_day))+
  geom_col(fill = 'purple')+
  geom_text(aes(label = rate_by_day), vjust = 'inward')+
  labs(x = 'Day of Week', y = 'Conversion Rate', title = 'Control Conversion Rate by Day of the
Week')
```



```
Exposed_by_day <- taskaBella_orig %>%
  filter(usertype %in% 'Exposed') %>%
  group_by(dayofweek) %>%
  mutate(user_count_day = length(user_id)) %>%
  mutate(conversions_day = sum(converted)) %>%
  select(usertype, dayofweek, user_count_day, conversions_day) %>%
  distinct(dayofweek, .keep_all = TRUE) %>%
  mutate(rate_by_day = round((conversions_day/user_count_day)*100,2))


Exposed_by_day %>%
  ggplot(aes(x = dayofweek, y = rate_by_day))+
  geom_col(fill = 'purple')+
  geom_text(aes(label = rate_by_day), vjust = 'inward')+
  labs(x = 'Day of Week', y = 'Average Conversion Rate', title = 'Exposed Conversion Rate by Day
of the Week')
```

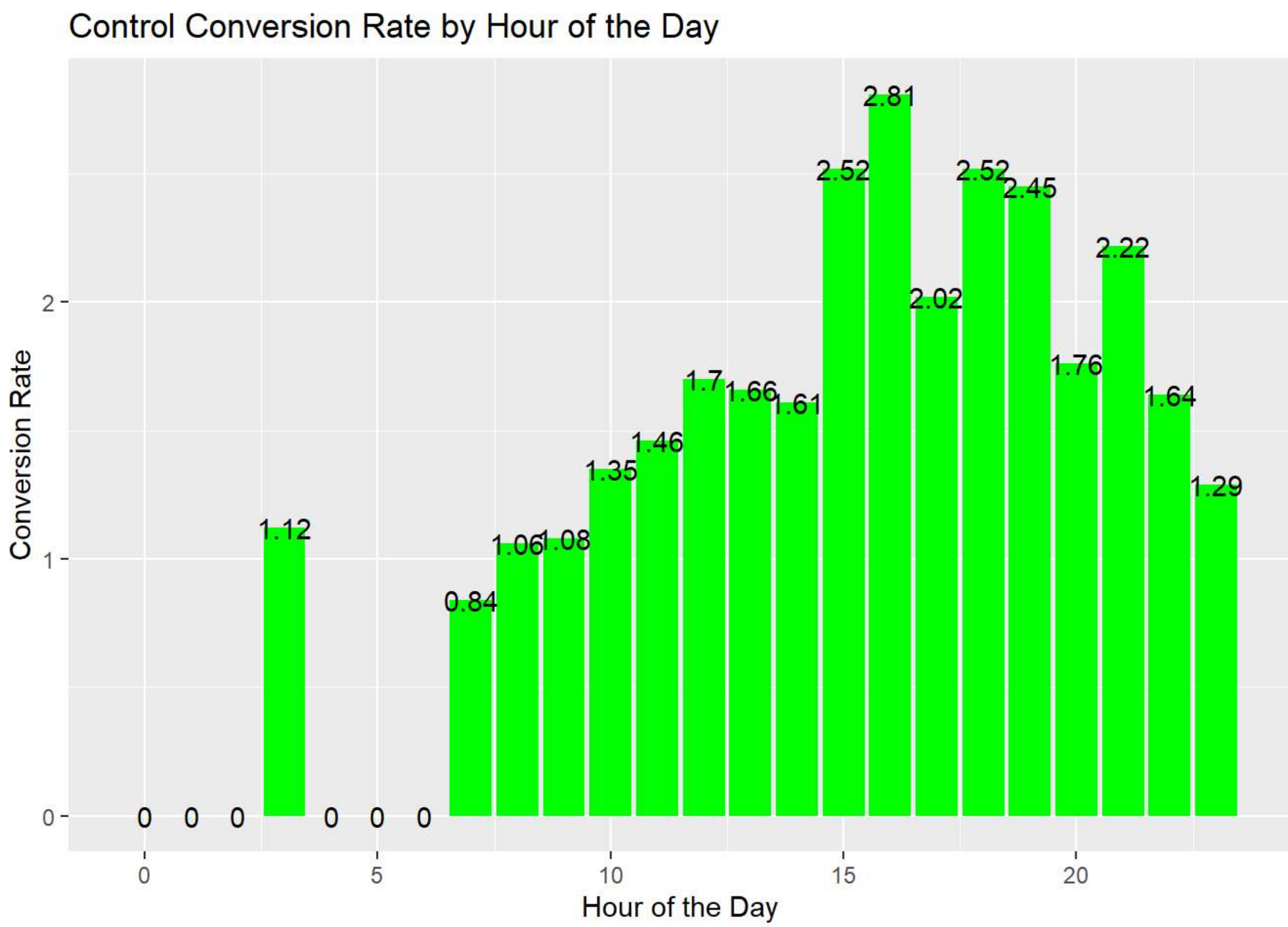


Create a chart showing conversion rates for the control and

exposed groups by time of day.

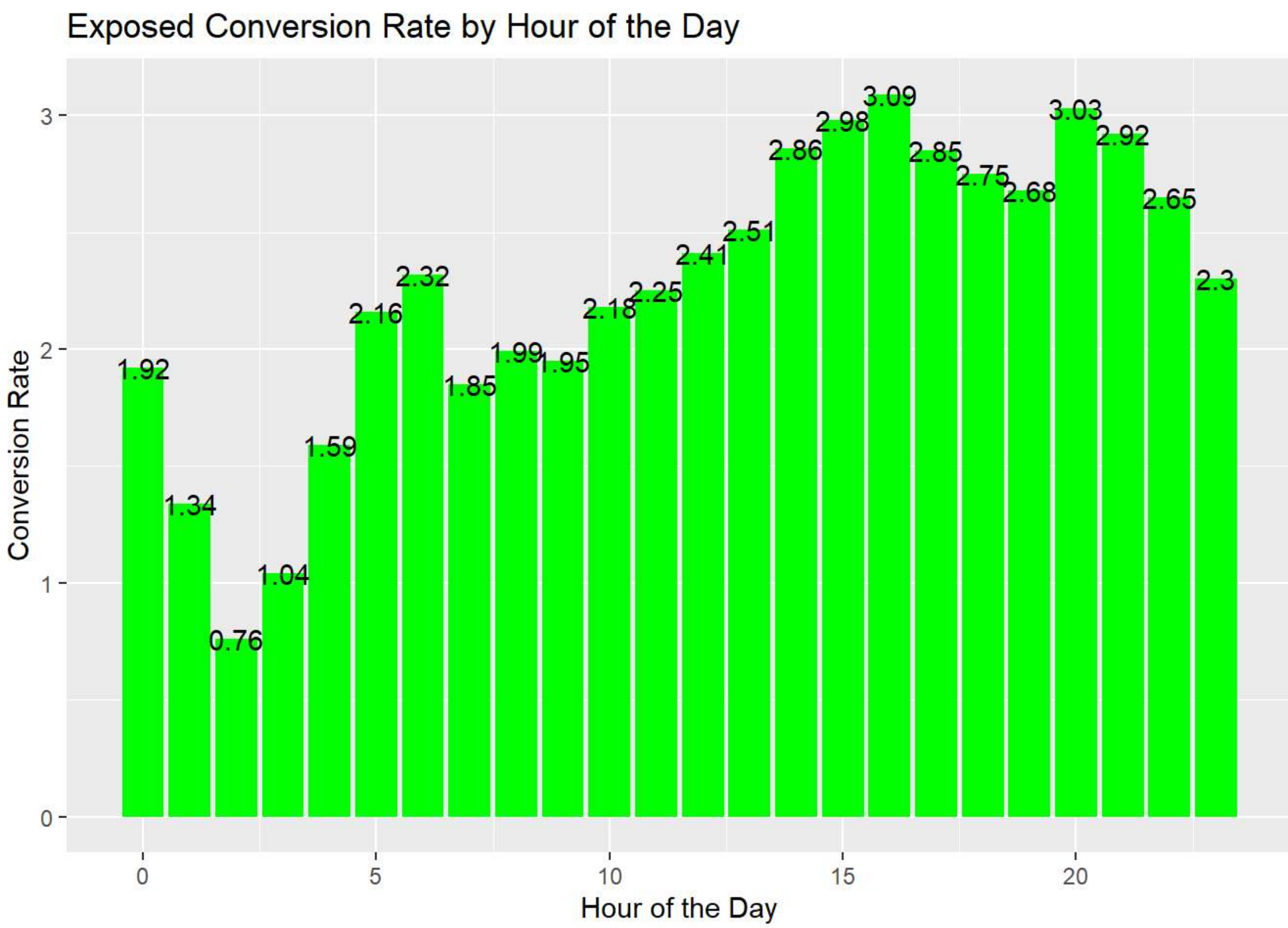
```
Control_by_hour <- taskaBella_orig %>%
  filter(usertype %in% 'Control') %>%
  group_by(mode_impr_hour) %>%
  mutate(user_count_hour = length(user_id)) %>%
  mutate(conversions_hour = sum(converted)) %>%
  select(usertype, mode_impr_hour, user_count_hour, conversions_hour) %>%
  distinct(mode_impr_hour, .keep_all = TRUE) %>%
  mutate(rate_by_hour = round((conversions_hour/user_count_hour)*100,2))


Control_by_hour %>%
  ggplot(aes(x = mode_impr_hour, y = rate_by_hour))+
  geom_col(fill = 'green')+
  geom_text(aes(label = rate_by_hour), position = 'dodge')+
  labs(x = 'Hour of the Day', y = 'Conversion Rate', title = 'Control Conversion Rate by Hour of
the Day')
```



```
Exposed_by_hour <- taskaBella_orig %>%
  filter(usertype %in% 'Exposed') %>%
  group_by(mode_impr_hour) %>%
  mutate(user_count_hour = length(user_id)) %>%
  mutate(conversions_hour = sum(converted)) %>%
  select(usertype, mode_impr_hour, user_count_hour, conversions_hour) %>%
  distinct(mode_impr_hour, .keep_all = TRUE) %>%
  mutate(rate_by_hour = round((conversions_hour/user_count_hour)*100,2))


Exposed_by_hour %>%
  ggplot(aes(x = mode_impr_hour, y = rate_by_hour))+
  geom_col(fill = 'green')+
  geom_text(aes(label = rate_by_hour), position = 'dodge')+
  labs(x = 'Hour of the Day', y = 'Conversion Rate', title = 'Exposed Conversion Rate by Hour of
the Day')
```



Based on your analysis, when would you choose to run ads?

Based on these graphs, the best choice in my opinion is Mondays at around 4pm (or 1600 hrs). As a range, Monday, Tuesday, and Wednesday from 1pm to 5pm and 7pm to 10pm is optimal.