Assignment 2

Joyce Wu(5897736) & Yen-Chen Hsu(5897444)

2023-10-01

Maximizing ROI: A Comprehensive Analysis of Effective Advertising Strategy for Star Digital

This is an analysis to measure causal effect of display advertising on sales conversion. We are interesting in finding out both a) is the advertising effective and b) which site(s) should Star Digital advertise on if it is in fact effective.

In our analysis, we examined two key aspects: the impact of advertising on the likelihood of purchase and the cost-effective channel of our advertising sites. Our analysis demonstrates that increasing the frequency of advertising impression positively affect probability of purchase. Additionally, the data suggests that Site 6 is the preferred platform for advertising due to its potential of delivering a higher ROI. These insights should guide Star Digital in optimizing its display advertising strategy to improve effectiveness and return on investment.

```r
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
```

```r
library(readxl)
library(ggplot2)
```

Analysis Detail

```r
star<-read_excel("/Users/joycewu/Library/CloudStorage/GoogleDrive-wu001370@umn.edu/My Drive/Fall/Inferen
```

Loading data set for analysis

1. Descriptive Statistic:

   First, let's look at the structure and summary of the data:

```r
print(summary(star))
```

```
##        id             purchase           test            imp_1
##  Min.   :     27   Min.   :0.0000   Min.   :0.000   Min.   :  0.0000
##  1st Qu.: 353880   1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:  0.0000
##  Median : 708344   Median :1.0000   Median :1.000   Median :  0.0000
##  Mean   : 708953   Mean   :0.5029   Mean   :0.895   Mean   :  0.9309
##  3rd Qu.:1062738   3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:  0.0000
##  Max.   :1413367   Max.   :1.0000   Max.   :1.000   Max.   :296.0000
##      imp_2             imp_3               imp_4             imp_5
##  Min.   :  0.000   Min.   :  0.00000   Min.   :  0.000   Min.   : 0.00000
##  1st Qu.:  0.000   1st Qu.:  0.00000   1st Qu.:  0.000   1st Qu.: 0.00000
##  Median :  0.000   Median :  0.00000   Median :  0.000   Median : 0.00000
##  Mean   :  3.428   Mean   :  0.09477   Mean   :  1.589   Mean   : 0.04897
##  3rd Qu.:  2.000   3rd Qu.:  0.00000   3rd Qu.:  0.000   3rd Qu.: 0.00000
##  Max.   :373.000   Max.   :148.00000   Max.   :225.000   Max.   :51.00000
##      imp_6
##  Min.   :  0.000
##  1st Qu.:  0.000
##  Median :  1.000
##  Mean   :  1.784
##  3rd Qu.:  2.000
##  Max.   :404.000
```

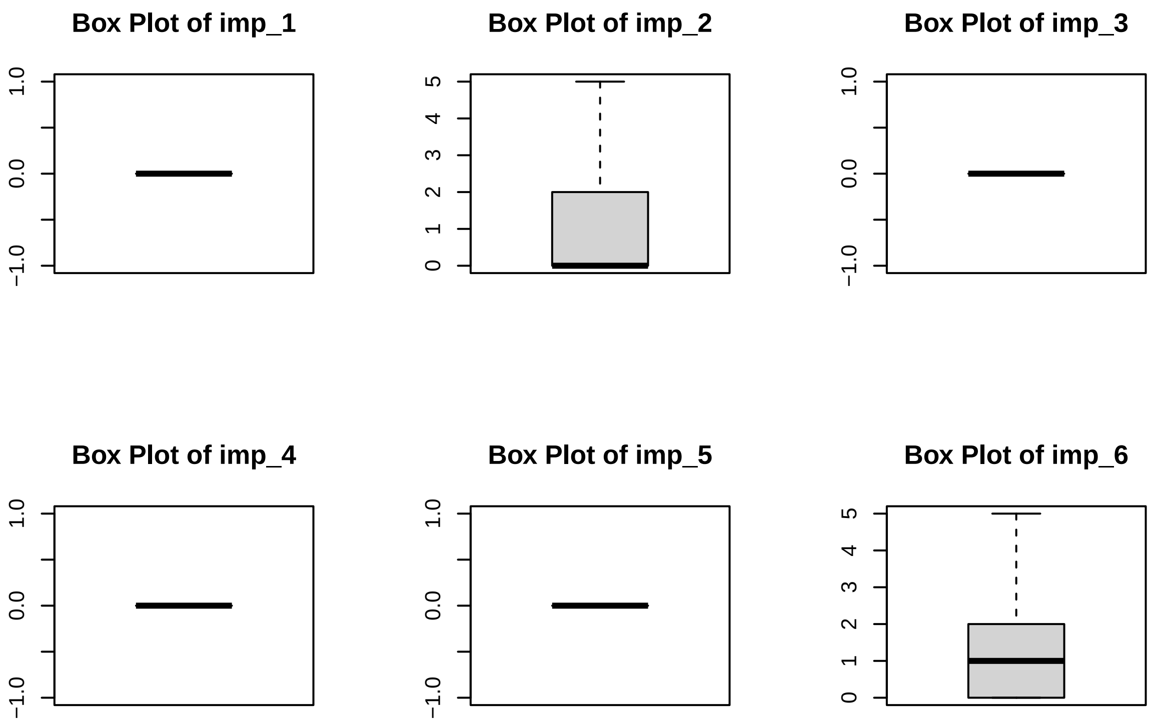
```r
print(str(star))
```

```
## tibble [25,303 x 9] (S3: tbl_df/tbl/data.frame)
##  $ id      : num [1:25303] 545716 893524 1372718 971359 59999 ...
##  $ purchase: num [1:25303] 1 1 1 1 1 1 0 0 1 0 ...
##  $ test    : num [1:25303] 1 1 1 1 1 0 1 0 1 1 ...
##  $ imp_1   : num [1:25303] 0 1 0 14 0 0 2 0 97 0 ...
##  $ imp_2   : num [1:25303] 1 0 0 37 0 1 272 0 214 0 ...
##  $ imp_3   : num [1:25303] 0 0 0 1 0 0 0 0 0 0 ...
##  $ imp_4   : num [1:25303] 0 17 10 7 13 0 18 0 11 0 ...
##  $ imp_5   : num [1:25303] 0 0 0 0 0 0 0 0 3 0 ...
##  $ imp_6   : num [1:25303] 0 1 0 7 0 0 2 1 13 2 ...
## NULL
```

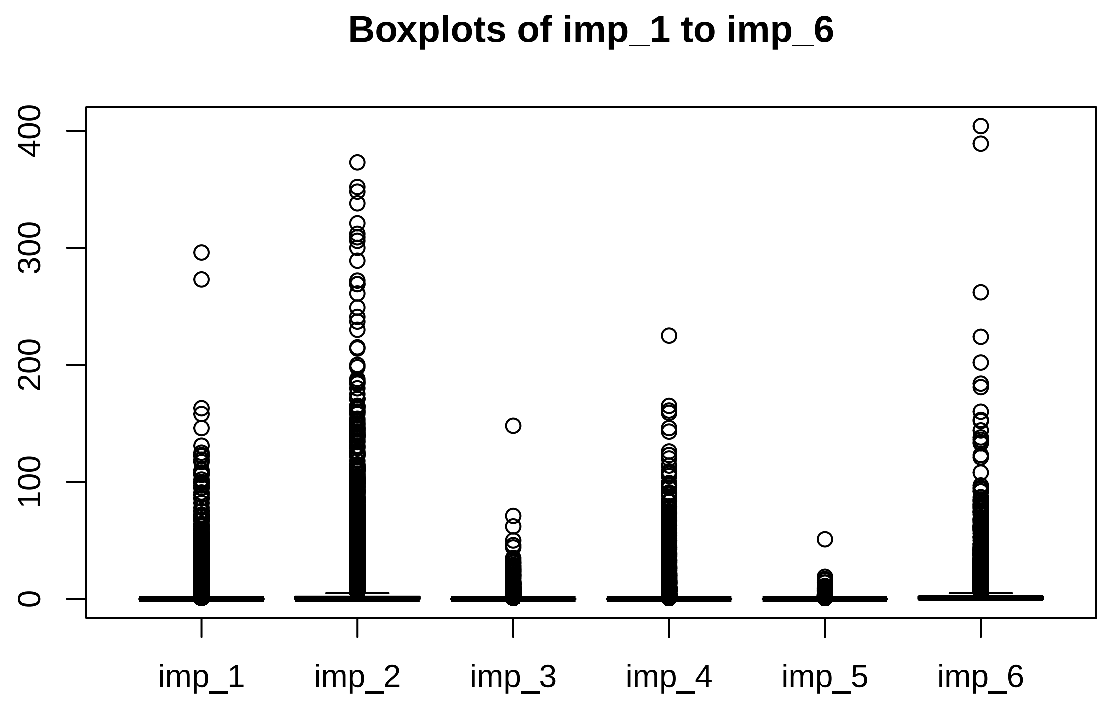
In order to understand more about the distribution of 6 websites from the experiment, we plot out the IQR plot for each website.

```r
#imp_1~imp_6 stands for the number of ad impression for either Star Digital or charity
Q1 <- apply(star[, 4:9], 2, quantile, probs=0.25)
Q3 <- apply(star[, 4:9], 2, quantile, probs=0.75)
IQR_values <- Q3 - Q1

# Identify potential outliers
lower_bound <- Q1 - 1.5 * IQR_values
upper_bound <- Q3 + 1.5 * IQR_values

# Create box plots for imp_1 to imp_6
par(mfrow=c(2, 3))  # Set up a 2x3 grid for multiple box plots
```

```
for (i in 4:9) {
  boxplot(star[, i], main=paste("Box Plot of", colnames(star)[i]), outline=FALSE)
}
```



```
# Create box plot for imp_1 to imp_6
boxplot(star[, 4:9], main="Boxplots of imp_1 to imp_6")
```



From above plots, we observed that the variables ‘imp_2’ and ‘imp_6’ displayed a notably higher number of outliers compared to the other websites. If possible, management team should also dive in the nature of these outliers in website 2 & 6, to find out what causes more impression to happen on these sites.

Then, we proceed to assess the result of the experiment. The measurable objective in the data set will be “purchase”, hence, we would calculate the ad impression to purchase rate and unique user purchase rate.

```
star = star %>% mutate(total_imp = imp_1+imp_2+imp_3+imp_4+imp_5+imp_6)

#unique ID to purchase rate
purchase_count <- sum(star$purchase == 1)
total_records <- nrow(star)
purchase_rate_uuid <- purchase_count/total_records
cat("purchase per unique ID: ", purchase_rate_uuid)
```

```
## purchase per unique ID:  0.5028653
```

```
#impression-to-purchase rate
impression_to_purchase_rate <- purchase_count / sum(star$total_imp)
cat("Impression-to-Purchase Rate:", impression_to_purchase_rate)
```

```
## Impression-to-Purchase Rate: 0.06385274
```

2. Randomization check: Here, we use t.test to understand if test and control groups are randomly given treatment by assessing whether total_imp varies across 2 groups. #small p-value = different H0 = not difference in total impression between test and control group; H1 = there’s difference of total impression between test and control group)

```r
t.test(total_imp ~ test, star)
```

```
##
##  Welch Two Sample t-test
##
## data:  total_imp by test
## t = 0.12734, df = 3204.4, p-value = 0.8987
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.8658621  0.9861407
## sample estimates:
## mean in group 0 mean in group 1
##        7.929217        7.869078
```

From t.test result, we report p-value 0.8987, which is quite high, indicating that there is not enough evidence to reject the null hypothesis. In other words, based on the data and the statistical test performed, it does not appear that there is a significant difference in the "total_imp" between control and test group

We conducted a randomization check to impression on Site 1 through 5 as well as on Site 6. We can also conclude that there is no significant differences between the number of impressions for these 2 grouped Sites between the control and test group.

```r
# sum total impression on Site 1 through 6
star = star %>% mutate(site15_imp = imp_1+imp_2+imp_3+imp_4+imp_5)
t.test(site15_imp ~ test, star)
```

```
##
##  Welch Two Sample t-test
##
## data:  site15_imp by test
## t = -0.071371, df = 3268.6, p-value = 0.9431
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.8402427  0.7812196
## sample estimates:
## mean in group 0 mean in group 1
##        6.065512        6.095024
```

```r
t.test(imp_6 ~ test, star)
```

```
##
##  Welch Two Sample t-test
##
## data:  imp_6 by test
## t = 0.43156, df = 2898.4, p-value = 0.6661
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.3176712  0.4969729
## sample estimates:
## mean in group 0 mean in group 1
##        1.863705        1.774054
```

3. Sample Size check:

```
power.t.test(delta=.1, sd = 1, sig.level = .05, power = 0.8, type = "two.sample",alternative = "two.sid
```

```
##
##      Two-sample t test power calculation
##
##               n = 1570.737
##           delta = 0.1
##              sd = 1
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Given the management team's goal of boosting the purchase conversion rate by 1% with the intervention, we set the delta as 0.1, and the minimum sample size for control and test group should be around 1570. We have a big enough sample size for both groups.

Question 1 - Is online advertising effective for Star Digital?

```
summary(lm(purchase ~ test, star))
```

```
##
## Call:
## lm(formula = purchase ~ test, data = star)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5049 -0.5049  0.4951  0.4951  0.5143
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.485693   0.009701  50.064   <2e-16 ***
## test        0.019186   0.010255   1.871   0.0614 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5 on 25301 degrees of freedom
## Multiple R-squared:  0.0001383,  Adjusted R-squared:  9.882e-05
## F-statistic: 3.501 on 1 and 25301 DF,  p-value: 0.06135
```

From the above regression, we can obtain the p-value for 'test' as 0.0614, greater than the acceptable significance level of 0.05. This indicates that we do not have enough evidence to reject the null hypothesis: the experiment of displaying Star Digital ads to the test group has no effect on their purchase behavior. We should also note that the sample data set was deliberately chosen, we should not arbitrarily conclude that the ad display experiment is useless.

Although we did not have sufficient evidence to support the experiment's effectiveness, we did observe a positive correlations between the test group and the average number of purchases increase by 1.91%.

Question 2 - Whether increasing frequency of advertising increases the probability of purchase?

To answer this question, we would first inspect the impression frequency of test and control group.

```r
summary(lm(purchase ~ test*total_imp, data = star))
```

```
##
## Call:
## lm(formula = purchase ~ test * total_imp, data = star)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.89562 -0.47994 -0.05711  0.51280  0.53228
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.4651265  0.0101335  45.900  < 2e-16 ***
## test           0.0111885  0.0107209   1.044   0.2967
## total_imp      0.0025937  0.0004131   6.278 3.49e-10 ***
## test:total_imp 0.0010362  0.0004408   2.351   0.0188 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.4942 on 25299 degrees of freedom
## Multiple R-squared:  0.02317,    Adjusted R-squared:  0.02306
## F-statistic:   200 on 3 and 25299 DF,  p-value: < 2.2e-16
```

To discuss the effectiveness of impression frequency on the probability of purchase, we can refer to the p-value of the total impression and the interaction of the test and total impression. Both of them demonstrate very small p-values, much lower than the significant level 0.05: .3.49e -10 for total impression, and 0.0188 for the interaction of total impression and test.

This indicates that we have enough evidence to reject the null hypothesis, in other words, the total impression number does effects number of purchases.

The output also provides insight into how purchase event changes for a one unit increase of impression. A one unit increase in ‘total_imp’ is associated with an estimated increase of approximately 0.0025937 in the purchase. Furthermore, given the positive interaction effect of total impression and the test, we can know that for the test group as total impression increase, the purchase increases simultaneously.

Question 3 - Which site should Star Digital advertise on? (Site 6 or in Site 1 through 5)

Here we present a short summary table of the number of users, purchases, impression on Site 6 or on Site 1 through 5 in order to get a clearer sense of the total ROI for both the test and control groups.

```r
star %>% group_by(test) %>% summarize(count_users = n(),
                                      total_purchase = sum(purchase),
                                      total_imp = sum(total_imp),
                                      total_6imp = sum(imp_6),
                                      total_15imp = sum(site15_imp),
                                      cost = (25*(total_15imp/1000)+
                                                20*(total_6imp/1000)),
                                      revenue = (1200*total_purchase),
                                      ROI = ((revenue-cost) / cost))
```

```
## # A tibble: 2 x 9
##    test count_users total_purchase total_imp total_6imp total_15imp  cost
##   <dbl>       <int>          <dbl>     <dbl>      <dbl>       <dbl> <dbl>
## 1     0        2656           1290     21060       4950       16110  502.
## 2     1       22647          11434    178211      40177      138034 4254.
## # i 2 more variables: revenue <dbl>, ROI <dbl>
```

For this question, we will only like to see the correlation of the impressions with purchases for the testing group. We will filter this subset of data, and focus on observing the difference in effectiveness between impressions on Site 1 through 5 versus impression on Site 6.

```
star_test = star %>% filter(test == 1) %>% mutate(site15_imp = imp_1+imp_2+imp_3+imp_4+imp_5)
```

Next, we run an model to test the correlation between the total impression on Site 1 through 5 and purchases. The result of this model shows a statistical significance of a positive correlation where purchases increase around 0.0039 with one unit increase of impression on Site 1 through 5.

```
summary(lm(purchase ~ site15_imp, data = star_test))
```

```
##
## Call:
## lm(formula = purchase ~ site15_imp, data = star_test)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0053 -0.4850  0.0589  0.5111  0.5189
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.4811201  0.0034383  139.93   <2e-16 ***
## site15_imp  0.0038981  0.0001674   23.29   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4941 on 22645 degrees of freedom
## Multiple R-squared:  0.02339,    Adjusted R-squared:  0.02335
## F-statistic: 542.3 on 1 and 22645 DF,  p-value: < 2.2e-16
```

We did the same analysis to observe the correlation between the total impression on Site 6 and purchases. The result shows a statistical significance of a positive correlation where purchases increase around 0.0037 with one unit increase of impression on Site 6.

```
summary(lm(purchase ~ imp_6, data = star_test))
```

```
##
## Call:
## lm(formula = purchase ~ imp_6, data = star_test)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0637 -0.5020  0.4019  0.5017  0.5017
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.4983241  0.0034406  144.84  < 2e-16 ***
## imp_6       0.0036950  0.0005118    7.22 5.38e-13 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.4994 on 22645 degrees of freedom
## Multiple R-squared:  0.002296,   Adjusted R-squared:  0.002252
## F-statistic: 52.12 on 1 and 22645 DF,  p-value: 5.377e-13
```

Here is a tabular summary of how impression of both Site 1 through 5 versus Site 6 has on ROI taking into account the unit price per impression as well as the expected revenue increase in dollar amount. In summary, we would recommend Star Digital advertise more on Site 6 than on Site 1 through 5, due to a higher expected ROI yield.

```
table = data.frame(
  Site = c("Site_1-5", "Site_6"),
  Revenue_per_impression_increase = c("0.0038981*$1200 = $4.67", "0.0036950*$1200 = $4.43"),
  Cost_per_impression = c("$0.025", "$0.020"),
  ROI_per_impression_increase = c("185.8","220.5")
)
print(table)
```

```
##       Site Revenue_per_impression_increase Cost_per_impression
## 1 Site_1-5         0.0038981*$1200 = $4.67             $0.025
## 2   Site_6         0.0036950*$1200 = $4.43             $0.020
##   ROI_per_impression_increase
## 1                       185.8
## 2                       220.5
```