Assignment2

Ssu Hsien Lee, Tsu Jung Liu

```
library(dplyr)
library(readxl)
library(ggplot2)
library(gridExtra)
library(pwr)
```

```
# import the dataset
star_data = read_excel("Star Digital Dataset.xls")
```

Setup

Executive Summary

Business Context

Star Digital identified a notable trend wherein a substantial portion of potential customers dedicated a significant amount of their time to online activities. Furthermore, there was a growing inclination among customers to consume media and make purchases through online channels. In response to this trend, Star Digital has been progressively boosting its allocation of advertising budget to online platforms, with a particular emphasis on banner ads.

Business Problem

To better quantify the impact of display advertising on sales conversion, Star Digital would like to initiate an experiment to assess user behavior in scenarios where individuals were exposed or not exposed to the online Star Digital's ads.

There are three key questions Star Digital want to address: 1. Is online advertising effective for Star Digital? 2. Is there a frequency effect of advertising on purchase? 3. Which sites should Star Digital advertise on?

Exploratory of the Dataset We first start from exploring the data set to check if there are any outliers or missing values in the data set.

From the exploration, there are 25303 rows and 10 columns of the star data set. Within the data set, there are no missing values. It also seems that purchase decisions are evenly distributed between the test and control groups, with approximately 50% of individuals in each group making a purchase.

```r
#display rows and columns
dim(star_data)
```

```
## [1] 25303     9
```

```r
# find location of missing values
which(is.na(star_data))
```

```
## integer(0)
```

```r
# count total missing values
print("Count of total missing values - ")
```

```
## [1] "Count of total missing values - "
```

```r
sum(is.na(star_data))
```

```
## [1] 0
```

```r
# purchase count
print("Count of total purchase of all consumers: ")
```

```
## [1] "Count of total purchase of all consumers: "
```

```r
table(star_data$purchase)
```

```
##
##     0     1
## 12579 12724
```

```r
# test control group count
print("Count of total purchase in test group: ")
```

```
## [1] "Count of total purchase in test group: "
```

```r
table(star_data[star_data$test == 1,]$purchase)
```

```
##
##     0     1
## 11213 11434
```

```r
print("Count of total purchase in control group: ")
```

```
## [1] "Count of total purchase in control group: "
```

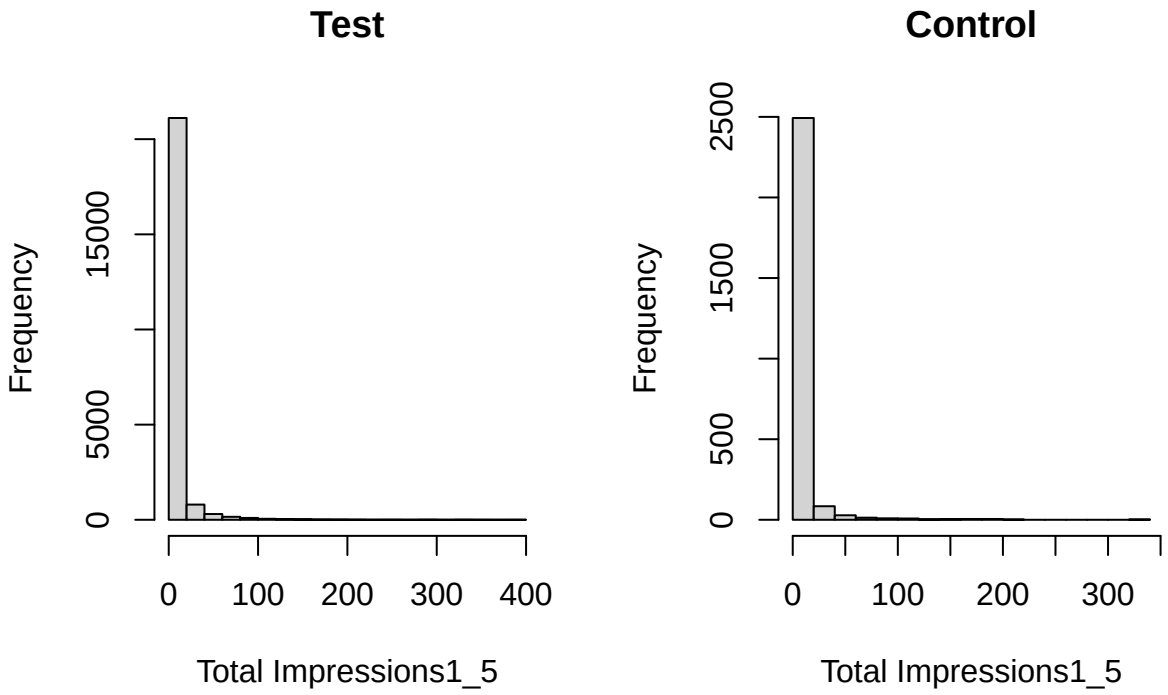
```
table(star_data[star_data$test == 0,]$purchase)
```

```
##
##    0    1
## 1366 1290
```

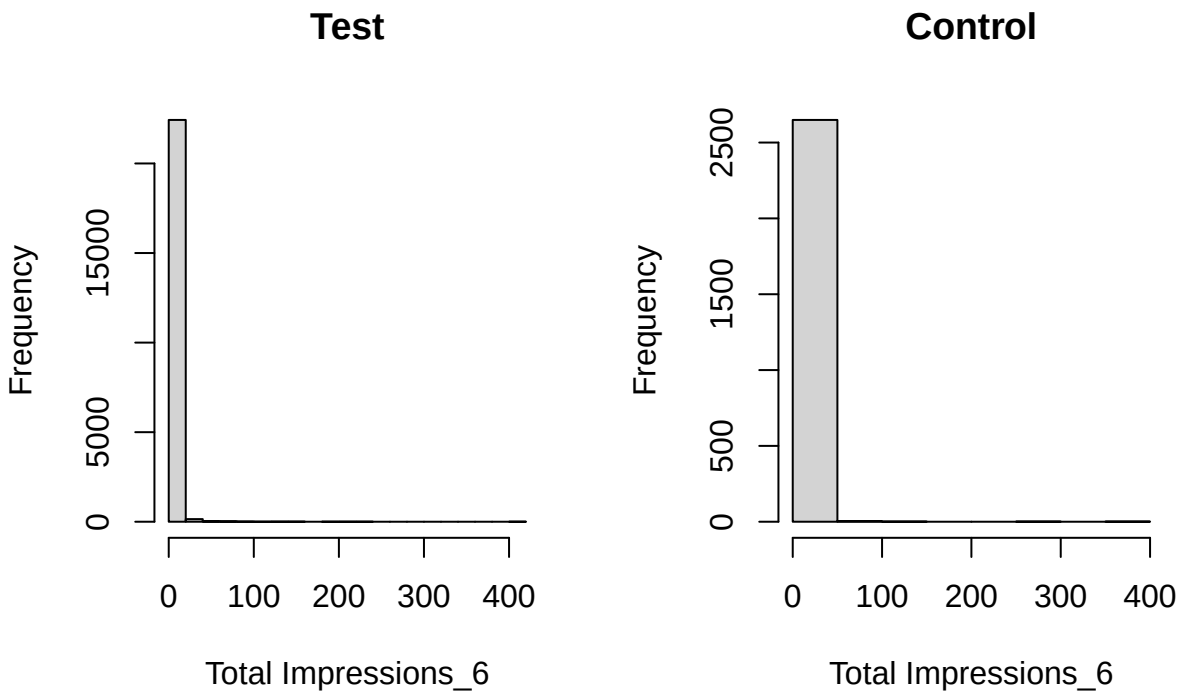
From the histogram chart below, it shows that the total impressions on website 1 to 5 and total impressions on website 6 are heavily right skewed in both test and control groups.

```
# Count the total imp_1~ imp_5
star_data = star_data %>% mutate(imp1_5 = star_data$imp_1 +
                                          star_data$imp_2 +
                                          star_data$imp_3 +
                                          star_data$imp_4 +
                                          star_data$imp_5 )
# histogram plot of impressions on test and control group
test = star_data %>% filter(test == 1)
control = star_data %>% filter(test == 0)

par(mfrow=c(1,2),oma = c(0, 0, 2, 0))
hist(test$imp1_5,xlab = 'Total Impressions1_5', main = "Test")
hist(control$imp1_5,xlab = 'Total Impressions1_5', main = "Control")
```



```
hist(test$imp_6,xlab = 'Total Impressions_6', main = "Test")
hist(control$imp_6,xlab = 'Total Impressions_6', main = "Control")
```



Experimental design

Having thoroughly examined the dataset, we are now commencing the experiment. This involves defining the test and control groups, addressing the sample size, and checking randomization to ensure the experiment's validity.

Test and Control Groups

Star Digital took into account the baseline conversion rate, campaign reach, minimum lift, and experiment power when deciding the 90%-10% allocation of consumers into test and control groups.

Test group: 90% consumers will be split into test groups, and they were exposed to a Star Control group: The rest of the 10% consumers will be split into control group, and they were shown a charity ad in replace.

SUTVA

The SUTVA assumption is unlikely to be violated because both the test and control groups are fixed. This ensures that consumers in each group will not be exposed to advertisements from the other group.

Sample Size Analysis

Before the experiment, we will perform power test to ensure a sufficient sample size to yield practically significant outcomes. Data exploration revealed an imbalanced distribution in the number of observations between the control and test groups, with a 1:9 ratio. To address this, we will apply the 'pwr.2p2n.test' function to assess our sample size. We set a 5% level of significance and Type II error rate of 20%.

We evaluate the effect size (h) to determine our sample size requirements. The relatively small effect size of 0.057 indicates that the effect we are exploring is modest in size and may not have a significant real-world impact. As a result, we should consider increasing the sample size to better detect desired effects.

```
control_cnt = star_data %>% filter(test == 0) %>% nrow()
test_cnt = star_data %>% filter(test == 1)  %>% nrow()

pwr.2p2n.test(n1 = control_cnt,
              n2 = test_cnt,
              sig.level = .05,
              power = .8,
              alternative = c("two.sided"))
```

```
##
##      difference of proportion power calculation for binomial distribution (arcsine transformation)
##
##               h = 0.05748084
##              n1 = 2656
##              n2 = 22647
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: different sample sizes
```

Randomization Check

To assess the randomization effectiveness and ensure the comparability of the test and control groups, we employed a t-test. If the groups are statically significantly different, the conclusions of our analyses may not be reliable.

The results of the t-test revealed statistically significant differences between the test and control groups for individual websites (specifically, sites 1, 3, and 5), signified by p-values < 0.05. This suggests that the composition of individuals in these groups may not be similar.

However, when aggregating data across websites 1 to 5, considering Star Digital's inability to display ads on a specific site, the p-value is 0.9431, indicating that there is no enough evidence to say that the two groups are significantly different.

Similarly, in terms of their impressions on site 6, the p-value = 0.9431, meaning that there is also no enough evidence to assume that the test and control groups are not similar.

In conclusion, we can state that the test and control groups are similar to each other in this experiment.

```
t.test(imp_1 ~ test, data = star_data)
```

```
##
##  Welch Two Sample t-test
##
## data:  imp_1 by test
## t = -3.905, df = 3574.1, p-value = 9.596e-05
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.5961772 -0.1976257
## sample estimates:
## mean in group 0 mean in group 1
##       0.5756777       0.9725791
```

```r
# Output of tests for sites 2 to 5 are hidden to reduce redundancy as results are
# similar to that of site 1 (sites 3/5 significant, sites 2/4 insignificant)
t.test(imp_2 ~ test, data = star_data)
```

```
##
##  Welch Two Sample t-test
##
## data:  imp_2 by test
## t = -1.1451, df = 3392, p-value = 0.2522
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.8376354  0.2199595
## sample estimates:
## mean in group 0 mean in group 1
##        3.151355        3.460193
```

```r
t.test(imp_3 ~ test, data = star_data)
```

```
##
##  Welch Two Sample t-test
##
## data:  imp_3 by test
## t = 9.052, df = 2655.4, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  0.6291825 0.9771466
## sample estimates:
## mean in group 0 mean in group 1
##      0.81362952      0.01046496
```

```r
t.test(imp_4 ~ test, data = star_data)
```

```
##
##  Welch Two Sample t-test
##
## data:  imp_4 by test
## t = -0.74842, df = 3192.3, p-value = 0.4543
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.4000161  0.1790006
## sample estimates:
## mean in group 0 mean in group 1
##        1.490587        1.601095
```

```r
t.test(imp_5 ~ test, data = star_data)
```

```
##
##  Welch Two Sample t-test
##
## data:  imp_5 by test
## t = -1.975, df = 4339.8, p-value = 0.04834
```

```
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.0327378476 -0.0001201375
## sample estimates:
## mean in group 0 mean in group 1
##      0.03426205      0.05069104
```

```
t.test(imp_6 ~ test, data = star_data)
```

```
##
##  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
```

```
t.test(imp1_5 ~ test, data = star_data)
```

```
##
##  Welch Two Sample t-test
##
## data:  imp1_5 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
```

Address the questions:

Q1. Is online advertising effective for Star Digital?

In addressing the question of whether displaying online advertising effectively prompts consumers to purchase more subscriptions at Star Digital, a t-test was conducted on the purchase between the test and control groups.

The mean purchase rate for the control group was observed to be 48%, while the test group exhibited a mean purchase rate of 50%. At first glance, this might suggest a difference in purchasing behavior between the two groups. However, the p-value reveals a value of 0.06139, exceeding the acceptable significance level of 0.05. Consequently, there is insufficient evidence to confidently assert that the consumers in the test or control group significantly influence purchasing at Star Digital.

Although the p-value is above the significance level, it's only slightly larger than 0.05, indicating a marginal significance. So the cautious interpretation is that there is a difference in the mean conversion rates of both groups. Therefore, online advertising appears to have some effectiveness for Star Digital. Considering this conclusion, we recommend that Star Digital persists in displaying online ads to enhance the purchasing behavior of consumers.

It's important to note that this marginal significance could be influenced by the sample size. Increasing the sample size is recommended, as it has the potential to yield a more precise and conclusive result.

```
t.test(purchase ~ test, data = star_data)
```

```
##
##  Welch Two Sample t-test
##
## data:  purchase by test
## t = -1.8713, df = 3309.2, p-value = 0.06139
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.039289257  0.000916332
## sample estimates:
## mean in group 0 mean in group 1
##       0.4856928       0.5048792
```

Q2. Is there a frequency effect of advertising on purchase?

In this context, our analysis involves a linear regression using total impressions as the independent variable and an interaction term to analyze the test effect.

First, we evaluate the effect of total impressions on purchase. The outcome p-value is relatively small (3.49e-10), indicating it is statistically significant. The coefficient is 0.0025937, which is positive, suggesting that more ad impressions will increase consumers' purchase.

To analyze the impact of increased ad impressions in the test group, we examined the interaction term (test:total_imp). The resulting p-value (0.0188) indicates statistical significance ($p < 0.05$). The positive coefficient (0.0010362) reveals that higher ad impressions have a stronger effect on purchase behavior within the test group compared to the control group.

We can conclude that a higher frequency of advertising indeed increases the likelihood of making a purchase. Additionally, our analysis reveals that consumers exposed to advertisements for Star Digital exhibit a higher propensity for purchasing compared to those exposed to advertisements for a charity organization. In summary, we recommend that Star Digital increase ad exposure on websites to enhance brand visibility, ultimately leading to higher conversion rates and boost sales.

```
star_data = star_data %>% mutate(total_imp = imp_1+imp_2+imp_3+imp_4+imp_5+imp_6)
summary(lm(purchase ~ test*total_imp, data = star_data))
```

```
##
## Call:
## lm(formula = purchase ~ test * total_imp, data = star_data)
##
## 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
```

Q3. Which sites should Star Digital advertise on?

Now that we understand the positive impact of increased ad impressions on consumer purchase behavior, the next question is where the company should invest. We will compare the effectiveness of advertising on websites 1 to 5 versus website 6 using a linear regression model.

The p-value (9.11e-11) and the positive coefficient (0.003) associated with ad impressions on websites 1 to 5 strongly support their significant impact on increasing purchases. In contrast, we lack sufficient evidence to draw the same conclusion for ad impressions on website 6. Additionally, when assessing the test effect on different websites, the p-value exceeds 0.05; therefore, we can't confidently concluding that ad impressions on websites 1 to 5 and website 6 have different effects on the control and test groups. In conclusion, we recommend the company focus its investment on websites 1 to 5.

```
summary(lm(purchase ~ test*imp1_5 + test*imp_6, data = star_data))
```

```
##
## Call:
## lm(formula = purchase ~ test * imp1_5 + test * imp_6, data = star_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.96643 -0.48127 -0.06395  0.51493  0.53375
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.4653462  0.0101323  45.927  < 2e-16 ***
## test        0.0121117  0.0107286   1.129    0.259
## imp1_5      0.0030780  0.0004747   6.484 9.11e-11 ***
## imp_6       0.0008997  0.0009167   0.982    0.326
## test:imp1_5 0.0007301  0.0005037   1.449    0.147
## test:imp_6  0.0014738  0.0010489   1.405    0.160
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4941 on 25297 degrees of freedom
## Multiple R-squared:  0.02359,    Adjusted R-squared:  0.0234
## F-statistic: 122.3 on 5 and 25297 DF,  p-value: < 2.2e-16
```

Conclusion

Based on the findings from the experiment, we have determined that online advertising significantly influences purchases for Star Digital. Furthermore, the analysis indicates a positive coefficient between advertising frequency and the likelihood of making a purchase. Notably, the effectiveness of advertising on websites 1 to 5 surpasses that of website 6.

Given these insights, we highly recommend Star Digital to maintain a consistent display of online advertisements on websites 1 to 5. This strategic approach is poised to not only boost purchase rates but also contribute to a substantial increase in overall revenue for the company.