Team 5b
Swathi Mudhelli
Harshal Vaza
Meixi Sun
Alejandra Miramontes
Nayab (Naby) Syeda

STAR DIGITAL TEAM ASSIGNMENT

Randomization check

To understand if the test and control groups are similar (i.e., if we are comparing apples to apples), we found a t-test to be appropriate:

a) t.test(Timp ~ star$test)

```
        Welch Two Sample t-test

data:  Timp by star$test
t = 0.12734, df = 3204.4, p-value = 0.8987
alternative hypothesis: true difference in means 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
```

Based on results of the t-test, the control group saw 7.92 ad impressions, whereas the test group saw 7.86 ad impressions. Although there is a small difference in the mean values, the p-value is significantly large. This implies that the averages are not statistically different. Therefore, we can conclude that the randomization is valid.

Q1. Is online advertising effective for Star Digital? In other words, is there a difference in conversion rate between the treatment and control groups?

To answer this question, we performed a test as noted by: t.test(star$purchase ~ star$test).

```
        Welch Two Sample t-test

data:  star$purchase by star$test
t = -1.8713, df = 3309.2, p-value = 0.06139
alternative hypothesis: true difference in means 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
```

From the t-test, we found the mean purchase of the control group is 48% and the test group is at 50%. We concluded that there is a difference between the two. The p-value is a small number (although slightly greater >5%), but we conclude that it is marginally significant. This implies that there is a difference between the mean (conversion rate) of both groups. Therefore, the ads are effective.

We continued and calculated the lift ratio to further understand the impact of ad effectiveness.

Lift = (mean(test)-mean(control))/mean(control) =(0.504-0.485)/0.485 = 0.0395 or 3.95%

We calculated the lift using the control group purchase mean (conversion rate) as the baseline. The lift is 3.95%, indicating that showing ads increases the probability of purchase of each customer by 3.95%. Therefore, we can conclude that ads are effective.

Shown below is the logistic regression to check for effectiveness:

```
Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.186  -1.186   1.169   1.169   1.202

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.05724    0.03882  -1.474   0.1404
test         0.07676    0.04104   1.871   0.0614 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 35077  on 25302  degrees of freedom
Residual deviance: 35073  on 25301  degrees of freedom
AIC: 35077

Number of Fisher Scoring iterations: 3
```

The p-value of the test is 0.0614, which is marginally significant (<10%), but not highly significant. The coefficient of the test is 0.07676. We can say that showing ads is at the least marginally effective.

Q2. Is there a frequency effect of advertising on purchase? In particular, the question is whether increasing the frequency of advertising (number of impressions) increases the probability of purchase?

In order to understand the frequency effect, we calculate two variables:

1) Timp = total sum of impressions 1-6
2) Timp_t = Timp*test (interaction term)

The beta coefficient of the interaction term explains the effect of showing more ads to the test group vs control group. In other words, it captures the effect of showing more ads to the test group.

With these variables, we fit a logistic regression model noted below:
Y = c+ b1 (Timp)+ b2 (Timp*Test)
B1-> captures how much customers who are shown ads (real or charity) are likely to purchase
B2-> captures how likely customers who saw real ads are to purchase, as the ads shown increase, frequency goes up

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.9091  -1.1272   0.1306   1.2150   1.2485

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.181875   0.014584 -12.471  < 2e-16 ***
Timp         0.016228   0.002676   6.065 1.32e-09 ***
Timp_t       0.015055   0.002930   5.139 2.76e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 35077  on 25302  degrees of freedom
Residual deviance: 34190  on 25300  degrees of freedom
AIC: 34196

Number of Fisher Scoring iterations: 5
```

The p value of both the total_impressions and the interaction term is highly significant. Both the coefficients are positive and significant. The coefficient of Timp (b1) is positive, implying customers who saw charity ads are also likely to purchase more. However, the coefficient of interaction term (b2) being positive implies that, showing more real ads increases purchase probability (by 1.5%). Therefore, we can conclude that there is a frequency effect of advertising.

Q3. How does the effectiveness of Sites 1-5 compare with that of Site 6?

To compare the effectiveness of advertising in sites 1-5 vs site 6, we use a logistic regression model as follows:

Y = a + b1 (Imp-15) + b2 (Imp_15 * test)+ b3 (Imp_6) + b4 (imp_6*test)

In this model, we include the interaction terms and compare the coefficients b2 and b4 to understand which sites are more effective. If b2>b4, sites 1-5 are more effective and if b4>b2, site 6 is more effective. We also include the total impressions from these sites to control for the effect of different customers seeing different ads (charity ad, real ads) and the different purchase probabilities.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.171919   0.014669 -11.720  < 2e-16 ***
Imp_15       0.019603   0.003265   6.005 1.92e-09 ***
Imp_15_t     0.014437   0.003562   4.054 5.04e-05 ***
star$imp_6   0.004068   0.004263   0.954   0.3399
Imp_6_t      0.013344   0.005321   2.508   0.0121 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 35077  on 25302  degrees of freedom
Residual deviance: 34166  on 25298  degrees of freedom
AIC: 34176

Number of Fisher Scoring iterations: 5
```

While analyzing the coefficients of interaction terms of Imp_15 and Imp_6, it appears as though sites 1-5 have a higher and statistically significant coefficient when compared to site 6. However, to conclude if sites 1-5 are better than site 6, we need to understand how significant the difference between the coefficients of the interaction terms is. To do this, we used a 5% confidence interval.

```
                    5 %         95 %
(Intercept) -0.196075706 -0.14781682
Imp_15       0.014504218  0.02524695
Imp_15_t     0.008361069  0.02008938
star$imp_6  -0.002341177  0.01250855
Imp_6_t      0.003657380  0.02172931
```

Based on a 5% confidence interval, both the coefficients lie in the CI range of each other. Hence, we can not conclude if site 1-5 is better than site 6 for advertisement.

Optional Challenge Question: Which sites should Star Digital advertise on? In particular, should it put its advertising dollars in Site 6 or in Sites 1 through 5?

The given data is oversampled, which biases the intercept but not the coefficients. We are able to use the regression model above to understand if advertising had an impact. But we cannot calculate the extent of impact due to the biased intercept. In order to correct the bias, we use the offset and run the model again.

- Offset is defined as = [(1-PCR)/PCR]
  [(1-SCR)/SCR]
- Where PCR is population conversion rate (0.00153) and SCR is sample conversion rate (0.5)

- offset1 = [(1-0.00153)/0.00153] / [(1-0.5)/0.5] = 6.480956

- We run a new logistic regression with the above offset value.

model4<- glm(purchase ~ Imp_15+Imp_15_t+imp_6+Imp_6_t, offset = offset1,data=star, family="binomial")

```
Coefficients:
             Estimate Std. Error  z value Pr(>|z|)
(Intercept) -6.652876   0.014669 -453.518  < 2e-16 ***
Imp_15       0.019603   0.003265    6.005 1.92e-09 ***
Imp_15_t     0.014437   0.003562    4.054 5.04e-05 ***
imp_6        0.004068   0.004263    0.954   0.3399
Imp_6_t      0.013344   0.005321    2.508   0.0121 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 35077  on 25302  degrees of freedom
Residual deviance: 34166  on 25298  degrees of freedom
AIC: 34176

Number of Fisher Scoring iterations: 5
```

From the p-values, clearly the coefficients of imp_15 and the interaction term of imp_15_t are highly significant, implying that these terms can better predict the purchase probability.
To understand which sites provide better financial incentive, we can compare purchase probabilities.

Assuming two scenarios,
1) Showing ads on sites 1-5 and not site 6 (real ads are shown to test group on sites 1-5, and the charity ads are shown to control group on site 6)

a) Let X: log(purchase)= intercept + coeff(imp_15) + coeff(imp_15_t) + coeff(imp_6)
= -6.652876+0.019603+0.014437+0.004068
= > P(purchase) = 1/(1+exp(x)) = 0.00134

2) Showing ads on site 6 and not site 1-5 ((real ads are shown to test group on sites 6, and the charity ads are shown to control group on site 1-5):

b) Let Y: log(purchase)= intercept + coeff(imp_6) + coeff(imp_6_t) + coeff(imp_5)
= -6.652876+0.004068+0.013344+0.019603
= > P(purchase) =1/(1+exp(y)) = 0.00133

From the formulae above, the probability of purchase is slightly higher for sites 1-5. However, the cost of advertising is higher for sites 1-5. So, the decision to advertise on sites 1-5 or site 6 should be based on ROI (return on investment).

Metric	Sites 1-5	Site 6
Cost of advertising (per impression)	$0.025	$0.020
Purchase probability	0.00134	0.00133
Contribution per purchase	$1200	$1200
Net benefit = P(purchase)*contribution	$1.61	$1.60
ROI = (Net benefit - cost)/cost * 100	63.33%	79.23%

Based on ROI above, it is beneficial to advertise on Site 6 due to higher return on investment.