Experiment Analysis

Gary Nguyen

Summary:

- A social media company has an ads product on their platform where companies and brands can use to market their products and services.
- **Overspending** is a situation where the ads generate more clicks than the company budgets for, and therefore incur a cost to the social medica company.
- The social media company therefore hypothesizes that a new advertising product, where companies pay whenever the ad show, will reduce overspending. The data scientists at the company ran an A/B experiment to decide whether this new product is indeed effective.

Summary of results:

a. There are 6,257 out of 7,733 (80.91%) campaigns that overspent in the control group, and 5,180 out of 5,721 (73.91%) campaigns that overspent the treatment group.
b. Based on the proportion z-test, there is convincing evidence that the new ad product reduces the proportion of overspending campaign. Based on logistic regression, medium and large company size contribute to a campaign's lower probability of overspending.
c. However, there is not enough evidence to conclude that the mean overspending amount in the treatment group is lower than the mean overspending amount in the control group, but there is strong evidence that the observations of overspend in the treatment group tend to be smaller than the observations of overspend in the control group, based on the Mann-Whitney-Wilcoxon rank sum test. Because of this inconclusiveness, in order to make inference about mean overspending, it is advisable to run a follow-up test.
d. There is evidence for the social media company to be concerned about the lower budget entered for campaigns with the new product, based on a t-test on the means of log transformations of campaign budgets.

Exploratory Data Analysis

Because the ranges of values of campaign budget, campaign spend and overspend are very wide, I performed log transformation on these variables to reduce skewness and make the data distribution approximately normal. From initial visualization (**Figure 1** and **Figure 2**), we can observe that the data on campaign budgets and campaign spend make sense. Larger companies tend to have larger campaign budgets and campaign spending. It is also noticeable that the dataset has a lot of outliers in both campaign spend and campaign budgets. I, however, decided against removing these outliers because there is no indication of errors.

Figure 1. Box plots of the log transformation of campaign budget in the control and treatment group, by company sizes

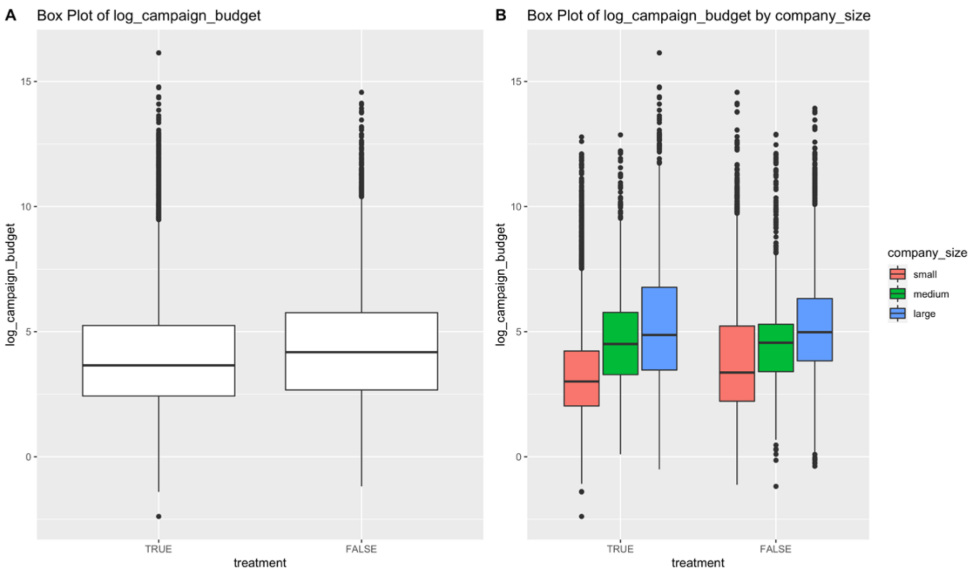
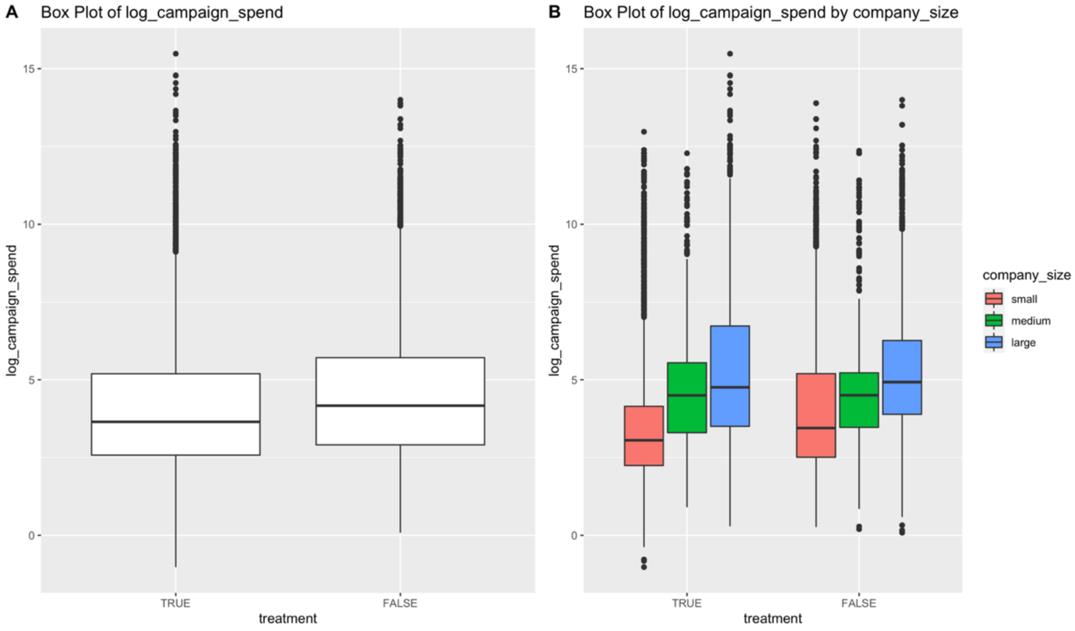
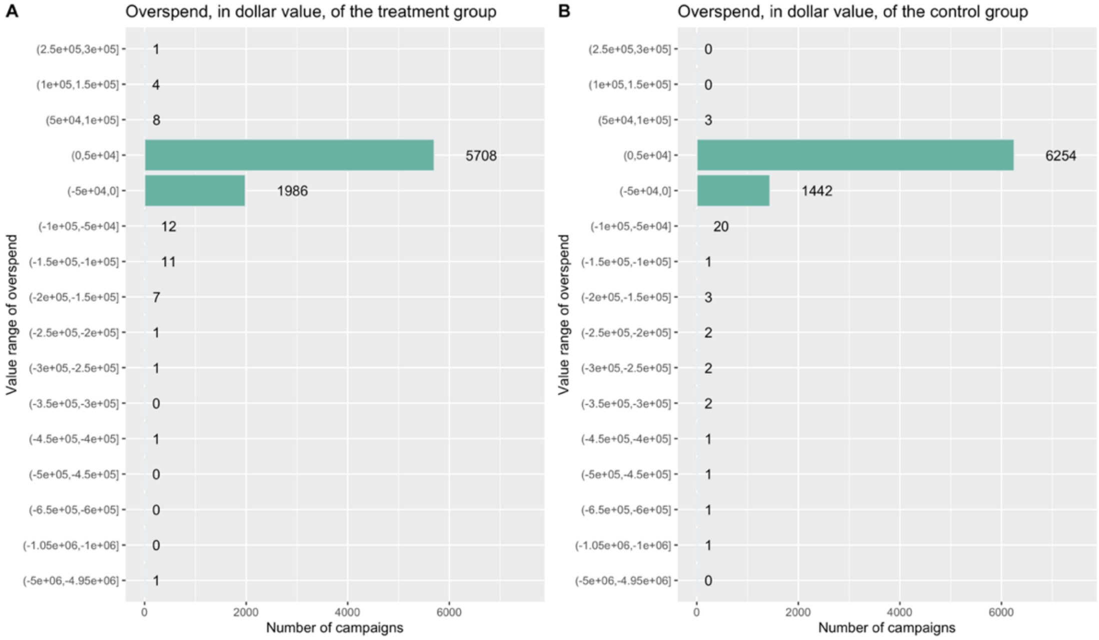


Figure 2. Box plots of the log transformation of campaign spend in the control and treatment group, by company sizes



From **Figure 3**, we can see that the distributions of dollar values of overspending are relatively similar between two groups, when grouped into bins of $50,000. The treatment group has more campaigns in the range of (-5e+04, 0] and the control group has more campaigns in the range of (0, 5e+04]. We can also observe a few more extreme values in the positive direction in the treatment group, and a few more extreme values in the negative directions in the control group. However, making conclusion based on these extreme values can be misleading because the absolute overspend dollar value is very dependent on the campaign size.

Figure 3. Overspend dollar value range in the control and treatment group



Part 1: Exploration - How many campaigns overspent in the control group vs. in the treatment group?

There are 6,257 out of 7,733 (80.91%) campaigns that overspent in the control group, and 5,180 out of 5,721 (73.91%) campaigns that overspent the treatment group. **Therefore, there are fewer campaigns that overspent more than 1% in the treatment group, which can be an indicator that the new product reduces overspending.** We would like to see whether this result is statistically significant or not, which will be examined in the next part.

Table 1: Number and percentages of overspending campaign in the control and treatment group

GROUP	Group Size	# of Campaigns Overspent	% of Campaigns Overspent
Control	7,733	6,257	80.91%
Treatment	7,741	5,721	73.91%

Part 2: Exploring the success of the new ads product by assessing whether the new ad product reduces overspending.

There are two ways to interpret this question:

- Was the new product effective at reducing the proportion of campaigns that overspent? And
- Was the new product effective at reducing the total overspend dollar?

I will try to answer both questions, because I think they are both useful to analyze and to understand the impact of the new product. I will present the final conclusion first, and then elaborate each point in analyses below.

Summary of Conclusions:

- There is convincing evidence that the new product reduces the proportion of overspending campaigns, based on the proportion Z-test.
- Based on logistic regression and ANOVA, company size plays a significant role in classifying a campaign as an overspending campaign. Specifically, medium and large companies are less likely to overspend in their campaigns.
- There is not enough evidence to conclude that mean overspend in the treatment group is lower than the mean overspend in the control group, based on permutation test and t-test. However, there is strong evidence that the observations of overspend in the treatment group tend to be smaller than the observations of overspend in the control group, based on the Mann-Whitney-Wilcoxon rank sum test. Because of these inconclusive findings and inability to quantify inference from the rank sum test, I recommend running a follow-up A/B test with a different design and higher statistical power, if budget allows.

First question: Was the new product effective at reducing the proportion of campaigns that overspent?

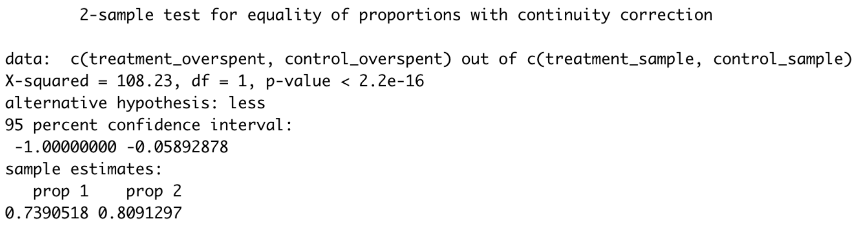
I defined "campaigns that overspent" as the campaigns with spend amount larger than budget amount. From initial data analysis, we know that the treatment group has fewer overspending campaigns than the control group. Specifically, there is 5,721 out of 7,741 (or 73.9%) overspending campaign in the treatment group, and 6,257 out of 7,733 (or 80.9%) overspending campaign in the control group.

I decided to use a one-sided, two-sample Z-test for proportion to test the difference. Proportions tend to (approximately) follow z-distribution, which is why this test is appropriate. Let p_t be the proportion of overspending campaign in the treatment group, and p_c be the proportion of overspending campaign in the control group, we have the following hypotheses:

- Null hypothesis (H_0): $p_t \geq p_c$
- Alternative hypothesis (H_a): $p_t < p_c$

After running the test (**Figure 4**), there is statistically significant evidence that there the new product reduces the proportion of overspending campaign by 7%.

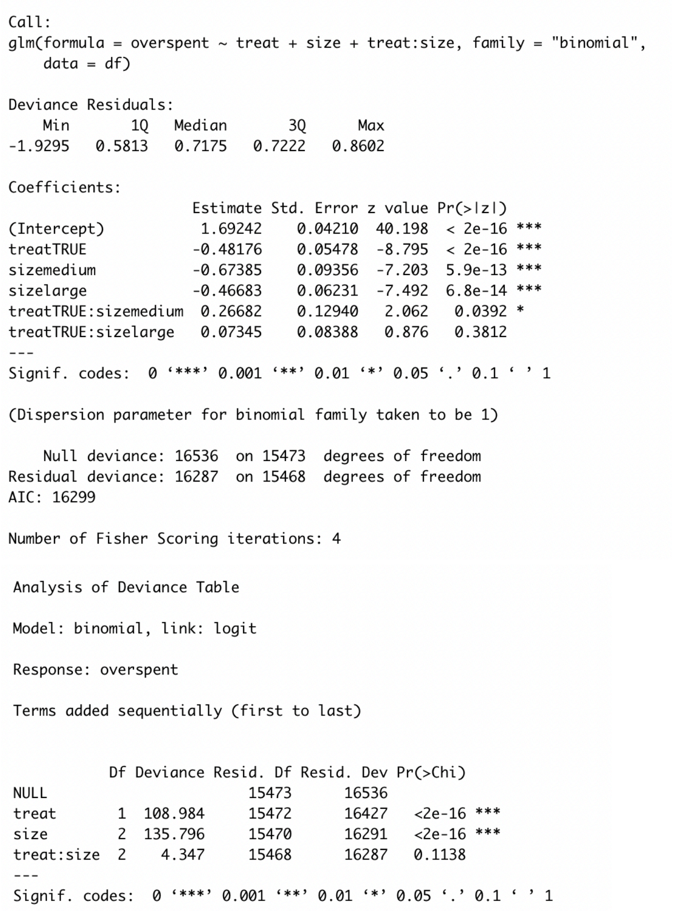
Figure 4. Proportion Z-test result for proportions of overspending campaigns in R



```
        2-sample test for equality of proportions with continuity correction

data:  c(treatment_overspent, control_overspent) out of c(treatment_sample, control_sample)
X-squared = 108.23, df = 1, p-value < 2.2e-16
alternative hypothesis: less
95 percent confidence interval:
 -1.00000000 -0.05892878
sample estimates:
   prop 1    prop 2 
0.7390518 0.8091297 
```

To assess the effect of company size on the treatment effect, I used a logistic regression model and two-way ANOVA test. From the output of the logistic regression in **Figure 5** (upper), we can observe statistical significance for the treatment dummy variable and company size categorical variables. For interaction effects, the interaction term between medium size and treatment effect is slightly statistically significant, but the interaction term between large size and treatment is not statistically significant. ANOVA test (**Figure 5** lower) revealed that there is not statistical significance for the interaction term.

Figure 5. Outputs of logistic regression (upper) and ANOVA (lower)

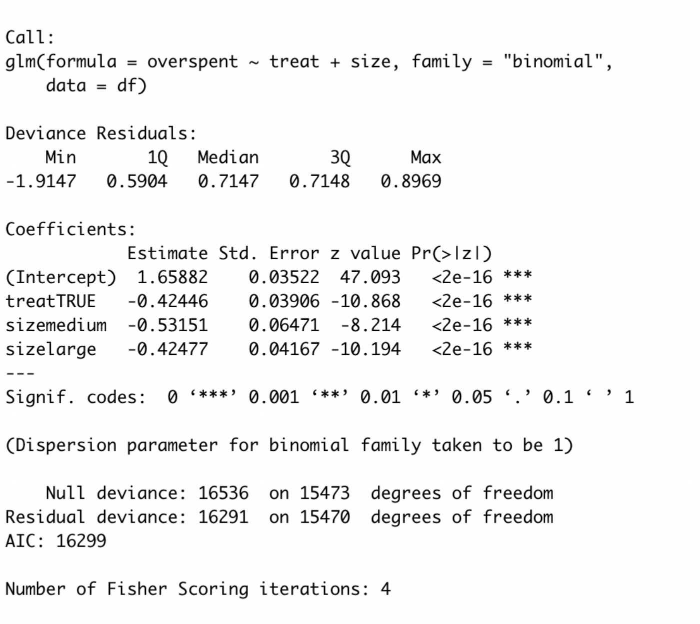


```
Call:
glm(formula = overspent ~ treat + size + treat:size, family = "binomial", 
    data = df)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.9295   0.5813   0.7175   0.7222   0.8602  

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)    
(Intercept)           1.69242    0.04210  40.198  < 2e-16 ***
treatTRUE            -0.48176    0.05478  -8.795  < 2e-16 ***
sizemedium           -0.67385    0.09356  -7.203  5.9e-13 ***
sizelarge            -0.46683    0.06231  -7.492  6.8e-14 ***
treatTRUE:sizemedium  0.26682    0.12940   2.062   0.0392 *  
treatTRUE:sizelarge   0.07345    0.08388   0.876   0.3812    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 16536  on 15473  degrees of freedom
Residual deviance: 16287  on 15468  degrees of freedom
AIC: 16299

Number of Fisher Scoring iterations: 4
```

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: overspent

Terms added sequentially (first to last)


           Df Deviance Resid. Df Resid. Dev Pr(>Chi)    
NULL                       15473      16536             
treat       1  108.984     15472      16427   <2e-16 ***
size        2  135.796     15470      16291   <2e-16 ***
treat:size  2    4.347     15468      16287   0.1138    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

To interpret the coefficients of the logistic regression, I removed the interaction terms, since they are not statistically significant and the coefficients cannot be easily interpreted their presence.

Figure 6. Outputs of logistic regression without interaction terms



```
Call:
glm(formula = overspent ~ treat + size, family = "binomial",
    data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9147   0.5904   0.7147   0.7148   0.8969

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.65882    0.03522  47.093   <2e-16 ***
treatTRUE   -0.42446    0.03906 -10.868   <2e-16 ***
sizemedium  -0.53151    0.06471  -8.214   <2e-16 ***
sizelarge   -0.42477    0.04167 -10.194   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 16536  on 15473  degrees of freedom
Residual deviance: 16291  on 15470  degrees of freedom
AIC: 16299

Number of Fisher Scoring iterations: 4
```

From output of the logistic regression model (Figure 6), the formulation of the final logistic regression function is:

$$\hat{p} = \sigma(1.659 + -0.424 \times treatment - 0.532 \times size\_medium - 0.425 \times size\_large)$$

$$\hat{p} = \frac{1}{1 + e^{-(1.659 + -0.424 \times treatment - 0.532 \times size\_medium - 0.425 \times size\_large)}}$$

- Keeping all other predictors constant , if the campaign is in the treatment group, the odds of being an overspending campaign will be $1 - e^{-0.424} = 34.5\%$ lower, with statistical significance.
- Keeping all other predictors constant, then the odds of being an overspending campaign in medium-sized company is $1 - e^{-0.532}$= 41.3% lower than in small-sized company (base level), with statistical significance.
- Keeping all other predictors constant, the odds ratio of being an overspending campaign in medium-sized company is $1 - e^{-0.425}$= 34.6% lower than in small-sized company (base level), with statistical significance.

Second question: Was the new product effective at reducing the raw overspend dollar?

This question is a relatively tricky to answer, because the overspend amount is very likely to be dependent on the size of each campaign's budget. For example, we can have a 10-million-campaign overspend by 1% (total overspending = $100,000), and 10 different $10,000 campaign underspend by 10% each (total underspending = $10,000), and still end up with a $90,000 overspend.

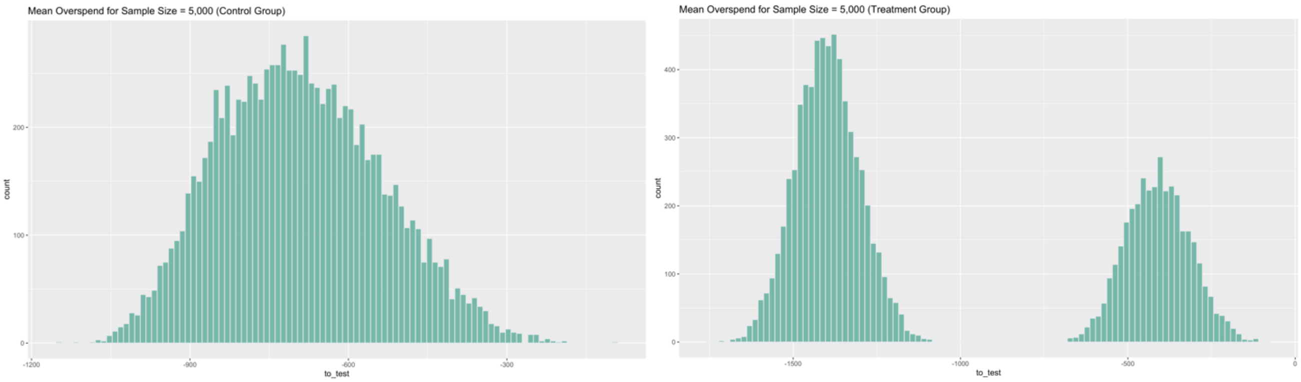
I will address this question with two variables: raw overspend and capped overspend. Raw overspend is defined as the difference between spend and budget. If a campaign underspends, the raw overspend will be negative. It is unknown whether the social media company cares about underspending and the possibility that underspending will offset overspending. Therefore, I created another variable that addresses overspending exclusively: capped overspending.

Capped overspend will be define as $\max(overspend, 0)$. Performing tests on this variable is more straightforward because the sampling statistics are more likely to be normally distributed as the means are not influenced by extreme negative values. The social media company is also more likely to care about this variable than raw overspend because the main objective is to reduce overspend in excess of budget, which is considered wasted opportunity.

Raw overspend

I won't be conducting a t-test because the sampling statistics (mean values of raw overspend) is not normally distributed even with a relatively large sample size (n = 5,000) in the treatment group (The plot on the right in Figure 7). The central limit theorem doesn't seem to apply appropriately in this case. Therefore, I will test whether the overspend amount is significantly different between the control and the treatment group using **a permutation test**.

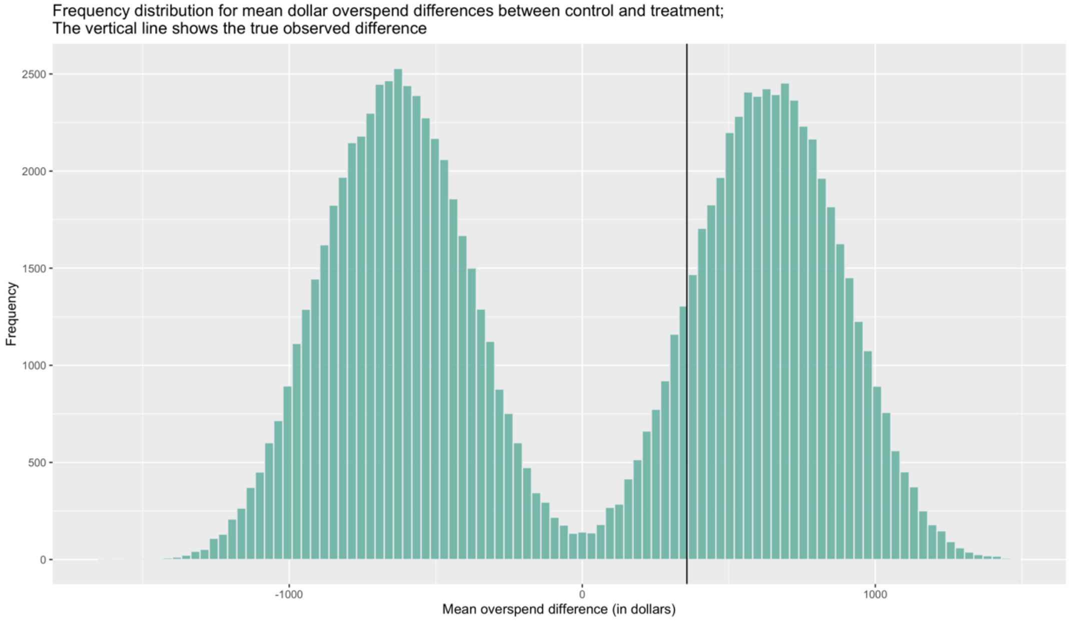
Figure 7. Sampling statistics (mean raw overspend) distribution between the control and treatment group



A permutation test is a procedure in which we assume the null hypothesis is correct. Then, we combine the control and treatment group into one group, draw a sample of n_c size from the combined group (n_c is equal to group size of the original control group), then draw a sample of n_t size from the combined group (n_t is equal to group size of the original treatment group). We then calculate the test statistics (in our case, the difference in mean overspend between the n_c-sized sample and the n_t-sized sample). We repeat the procedure N times. In this case, I repeated the procedure 100,000 times to ensure randomization.

If the observed difference (mean overspend in the original control group – mean overspend in the original treatment group) lies outside the majority of the permutation distribution, then we can conclude the chance variation is not responsible, meaning the difference is statistically significant. In contrast, if the observed difference lies within the permuted differences, then it is likely that the observed difference arises from chance variations.

Figure 8. Permuted and observed differences in mean overspend

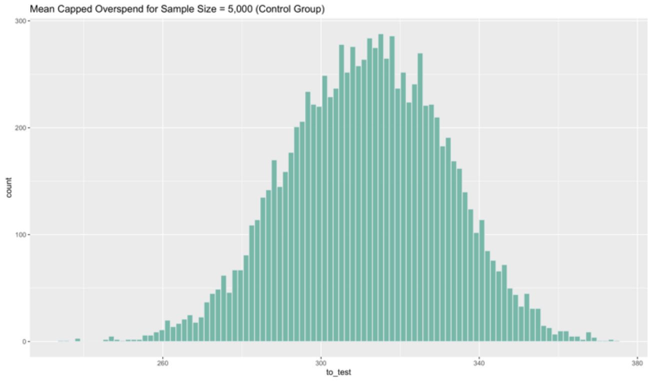


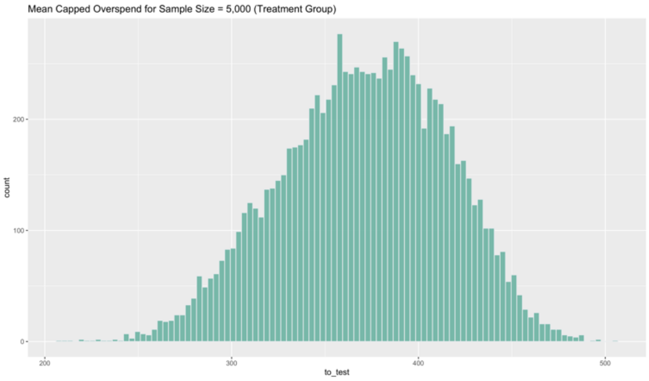
From Figure 8, we can see that the observed difference lies well within the permutation distribution. In fact, the mean difference of random permutations exceeds the observed difference between control and treatment 43.5% of the time, which is far greater than the 5% threshold of significance. Therefore, we don't have enough evidence to conclude that the mean values of raw overspend differ significantly between the control and treatment group.

Capped overspend

With capped overspend, I was able to verify that with sufficiently large sample size, the sampling statistics distribution look relatively normal for both the control and treatment group (Figure 8). With enough sample size, the Central Limit Theorem will hold, and we can conduct a t-test. Without knowing the population variance, I will use a one-sided, two-sample t-test to test whether the mean differences in capped overspend is statistically significant. Specifically, I will perform a Welch's t-test of unequal variances directly. When the variances are equal, this t-test will be identical to the normal t-test, so it is appropriate to just use Welch's t-test directly.

Figure 8. Sampling statistics (mean capped overspend) distribution between the control and treatment group

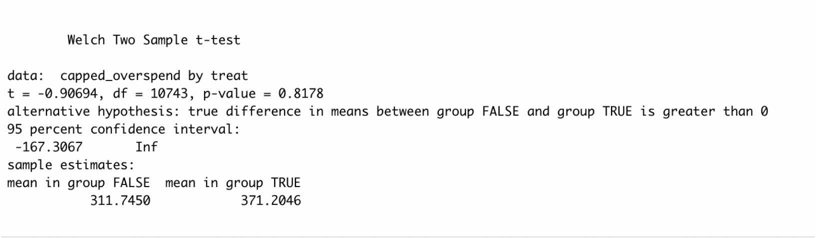




Let μ_c be the mean value of capped overspend in the control group, and μ_t the mean value of capped overspend in the treatment group.

- Null Hypothesis (H0): $\mu_t \geq \mu_c$
- Alternative Hypothesis (HA): $\mu_t < \mu_c$

Figure 9. T-test for difference in mean capped overspend



```
        Welch Two Sample t-test

data:  capped_overspend by treat
t = -0.90694, df = 10743, p-value = 0.8178
alternative hypothesis: true difference in means between group FALSE and group TRUE is greater than 0
95 percent confidence interval:
 -167.3067       Inf
sample estimates:
mean in group FALSE  mean in group TRUE 
           311.7450            371.2046 
```

From the t-test result in Figure 9, we can conclude that there is not enough evidence to reject the null hypothesis (p = 0.8178 > 0.05). In fact, the mean capped overspend of the treatment group is higher than the mean capped overspend of the control group. This means that the mean overspending amount in the treatment group is actually higher than that of the control group, if we don't care about underspending.

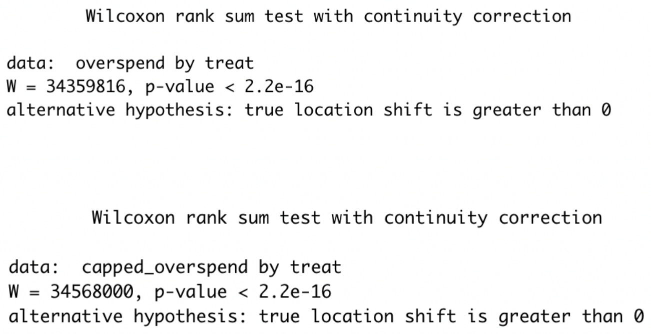
Mann-Whitney-Wilcoxon Test

The Mann-Whitney-Wilcoxon test checks whether the distributions of two variables are the same. However, we cannot make any inference about test statistics (e.g., mean or median), because this test is a non-parametric test. Let X be the variable that represent either raw overspend or capped overspend.

- Null hypothesis (H_0): The population distributions of overspend are the same between the control and treatment group, $P(X_{control} > X_{treatment}) \leq 0.5$
- Alternative hypothesis (H_a): Observations from control tend to be larger than observations from treatment, $P(X_{control} > X_{treatment}) > 0.5$

From Figure 10, there is evidence that overspend (both raw and capped) in control group tends to be larger than overspend in the treatment group. However, it is difficult to quantify the result because this is a non-parametric test. Because of inconclusive findings, I recommend running a follow-up test with clearer design and higher power, to investigate further.

Figure 10. Mann-Whitney-Wilcoxon Test for raw overspend (upper) and capped overspend (lower)



```
	Wilcoxon rank sum test with continuity correction

data:  overspend by treat
W = 34359816, p-value < 2.2e-16
alternative hypothesis: true location shift is greater than 0



	Wilcoxon rank sum test with continuity correction

data:  capped_overspend by treat
W = 34568000, p-value < 2.2e-16
alternative hypothesis: true location shift is greater than 0
```

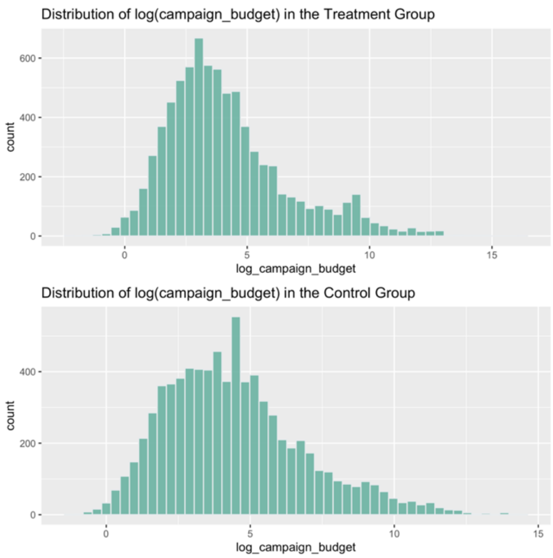
Question 3: Assessing guardrail metrics - whether the new ads product reduces the budgets that campaign entered?

Conclusions:

- **There is evidence for the product managers to be concerned about the lower budget for the campaigns for the new product.** I arrived at this conclusion after performing a t-test on the means of log transformations of campaign budgets.

The log transformation reduces the skewness of campaign budget. From Figure 11, we can see that the distributions of log campaign budget are relatively normal in both the control and treatment group. Without knowing the variances of either population distributions, a t-test is appropriate in this case.

Figure 11. Distribution of the log transformation of campaign budget in the control and treatment group

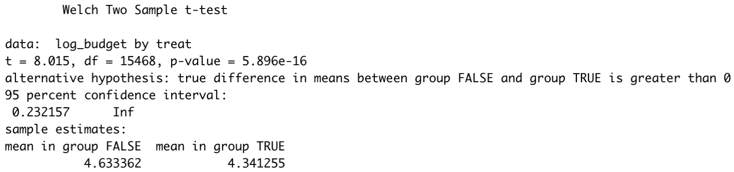


Specifically, I performed a one-sided, two-sample t-test. Let μ_c be the mean value of the log transformation of campaign budget in the control group, and μ_t the mean value of the log transformation of campaign budget in the treatment group.

- Null Hypothesis (H_0): $\mu_t \geq \mu_c$
- Alternative Hypothesis (H_a): $\mu_t < \mu_c$

From the result in Figure 12, with p-value < 2.2e-16 < 0.05, we can reject the null hypothesis. The mean of campaign budget in the control group is $e^{4.63-4.34} = 1.34$ times the mean campaign budget in the treatment group. This means the treatment effect decreases the campaign budget by approximately 26%.

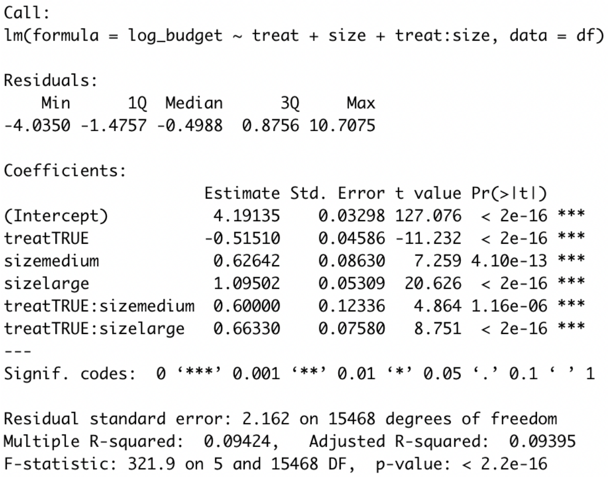
Figure 12. Result of the two-sample t-test on the log transformation of campaign budget.



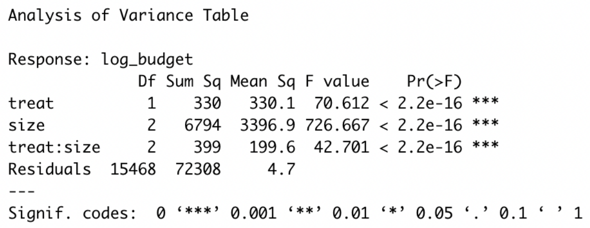
```
        Welch Two Sample t-test

data:  log_budget by treat
t = 8.015, df = 15468, p-value = 5.896e-16
alternative hypothesis: true difference in means between group FALSE and group TRUE is greater than 0
95 percent confidence interval:
 0.232157      Inf
sample estimates:
mean in group FALSE  mean in group TRUE 
           4.633362            4.341255 
```

After performing linear regression and two-way ANOVA (Figure 13) for the treatment effect and company size, we can observe an interaction effect between company size and treatment on log(campaign_budget), with high statistical significance (p < 2.2e-16). Therefore, the log transformations of campaign budgets are affected by both company sizes and treatment effect.

Figure 13. Outputs of linear regression (upper) and ANOVA (lower) between log campaign budgets and treatment & company size



```
Call:
lm(formula = log_budget ~ treat + size + treat:size, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.0350 -1.4757 -0.4988  0.8756 10.7075 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)           4.19135    0.03298 127.076  < 2e-16 ***
treatTRUE            -0.51510    0.04586 -11.232  < 2e-16 ***
sizemedium            0.62642    0.08630   7.259 4.10e-13 ***
sizelarge             1.09502    0.05309  20.626  < 2e-16 ***
treatTRUE:sizemedium  0.60000    0.12336   4.864 1.16e-06 ***
treatTRUE:sizelarge   0.66330    0.07580   8.751  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.162 on 15468 degrees of freedom
Multiple R-squared:  0.09424,	Adjusted R-squared:  0.09395 
F-statistic: 321.9 on 5 and 15468 DF,  p-value: < 2.2e-16
```



```
Analysis of Variance Table

Response: log_budget
              Df Sum Sq Mean Sq F value    Pr(>F)    
treat          1    330   330.1  70.612 < 2.2e-16 ***
size           2   6794  3396.9 726.667 < 2.2e-16 ***
treat:size     2    399   199.6  42.701 < 2.2e-16 ***
Residuals  15468  72308     4.7                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Therefore, we should incorporate the interaction terms when interpreting the treatment effect sizes. It’s not appropriate to interpret to individually interpret coefficients when interaction terms are presence, so we can use the formulation below to predict campaign budgets:

$$\begin{aligned}\log(campaign\ budget) &\\ = \ 4.19\ +\ -0.52 \times treatment\ +\ \ 0.63 \times size\_medium + \ \ 1.10 \times size\_large\\ + 0.60 \times treatment\ \times size\_medium + 0.66 \times treatment\ \times size\_large\end{aligned}$$

Where:

- $treatment = 1$ when the campaign is in the treatment group, otherwise 0
- $size\_medium = 1$ when the campaign is from a medium-sized company, otherwise 0
- $size\_large = 1$ when the campaign is from a large-sized company, otherwise 0
- If $size\_medium = 0$ and $size\_large = 0$, then the campaign is from a small-sized company