Q2: Experiment Outcome [A/B Testing]

We are tasked with analyzing the outcome of an experiment recently conducted on a group of users to test the hypothesis on the effectiveness of a new feature. The feature will help in reducing campaign overspending

---

Exploratory Data Analysis

We will start with looking at the data..

	treatment	company_size	campaign_spend	campaign_budget
0	0	small	10.4477	3.9035
1	0	medium	3.7776	1.9872
2	0	medium	46.1880	55.4523
3	0	small	6.6271	6.5136
4	0	small	92.3405	83.1018
5	0	small	180.6259	328.1411
6	0	large	411.0601	404.3652
7	0	small	6.4808	5.3393
8	0	small	20.6899	12.2408
9	0	medium	28.5898	31.1672
10	0	small	40.9653	30.6545

From a cursory look, the data has two groups, treatment and control and there are three company sizes, small, meidum and elevation_range. Each is followed by the campaign budget and spend, respectively

	large	medium	small	All
False	2701	735	4297	7733
True	2445	692	4604	7741
All	5146	1427	8901	15474

A quick overview of the data summary

	campaign_spend	campaign_budget
count	1.547400e+04	1.547400e+04
mean	4.903037e+03	5.772614e+03
std	6.516692e+04	9.903381e+04
min	3.595000e-01	9.190000e-02
25%	1.517812e+01	1.279290e+01
50%	5.009005e+01	4.881625e+01
75%	2.365468e+02	2.523155e+02
max	5.289217e+06	1.024289e+07

75th percentile of both campaign spend and budget is in 236 and 252 dollars, however, their maxium values are way in the multi-millions!! For the purpose of visualization, we can remove some of the higher percentile values

95th percentile for campaign spend:

9488.173734999853

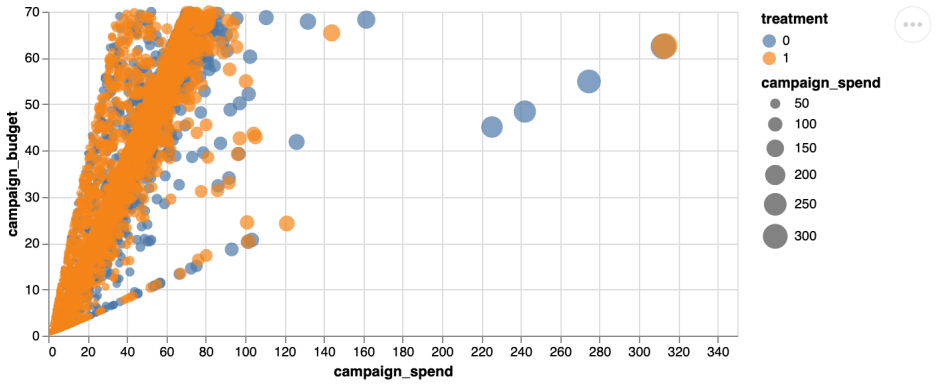
95th percentile for campaign budget:

10882.285809999985

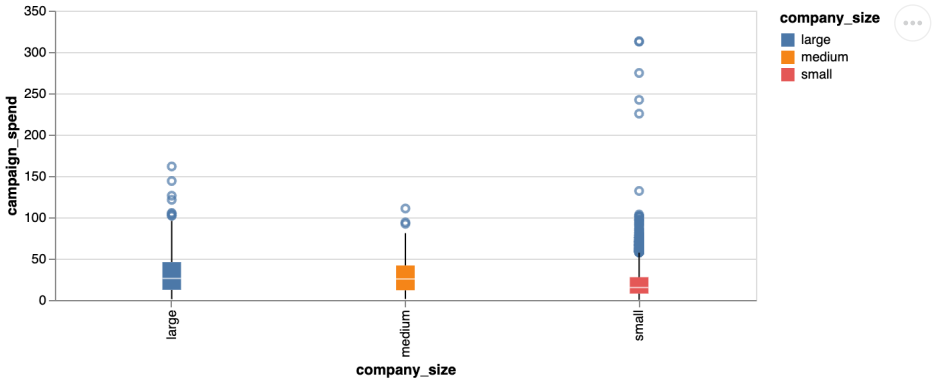
Checking for na

treatment	0
company_size	0

```
campaign_spend     0
campaign_budget    0
dtype: int64

A good visualization on the spread of data for campaign spend and budget
** data reduced to 75th percentile for better visualization **

Campaign Spend
0.00                                                                400.00

0.00                                                                400.00

Campaign Budget
0.00          70.00

0.00                                                                400.00
```

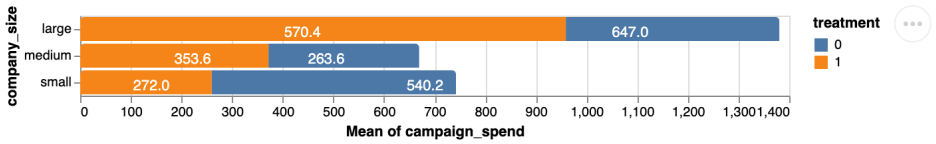


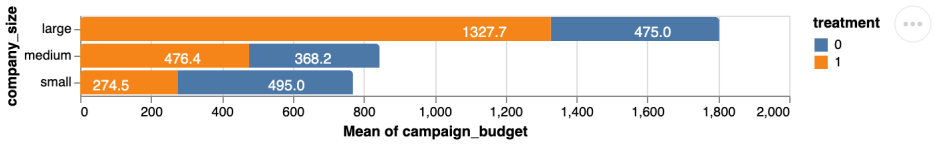

```
Adding a histogram - it updates with the filter above
```



The data is consistent with the inutuition that the company size follows the budget and spend- although there are quite a few instances when the spend is quite high for smaller companies

```
Lets take a look at some of the mean campaign spend and budget
** Greater than 95 Percentile data removed, to reflect a better picture**
```





Adding new variables for over spend

Assumption: Since we are considering if overspend happened and not by how much, we are going to include all the data at this time

We can introduce two new variables,
percentage overspend and assign 1 to it if crossed 1% of the budget

	treatment	company_size	campaign_spend	campaign_budget	percentage	over_spend
0	False	small	10.4477	3.9035	167.649545	1.0
1	False	medium	3.7776	1.9872	90.096618	1.0
2	False	medium	46.1880	55.4523	-16.706791	0.0
3	False	small	6.6271	6.5136	1.742508	1.0
4	False	small	92.3405	83.1018	11.117328	1.0
5	False	small	180.6259	328.1411	-44.954808	0.0
6	False	large	411.0601	404.3652	1.655657	1.0
7	False	small	6.4808	5.3393	21.379207	1.0
8	False	small	20.6899	12.2408	69.024083	1.0
9	False	medium	28.5898	31.1672	-8.269591	0.0

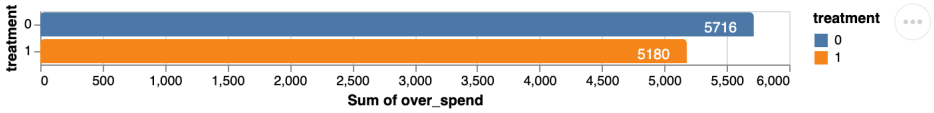
Note: It is important to remember that over spend is 1 when the spend is 1% greater than the budget

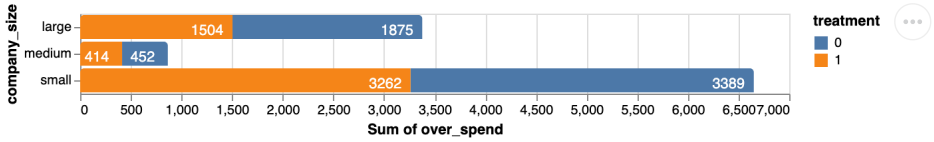
The table below shows the proportion of over spend across groups and overall

	0	1
False	0.4406	0.5246
True	0.5594	0.4754

		False	True
large	0	0.1068	0.1216
large	1	0.2425	0.1943
medium	0	0.0366	0.0359
medium	1	0.0585	0.0535
small	0	0.1174	0.1734
small	1	0.4383	0.4214

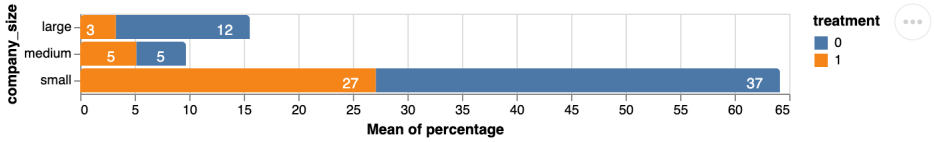
The barplots will give us a better breakdown of total number overall and for each group





From this data it is clear that there are definitely more over spend occurences in the control group overall and across all the companies At an overall level, the number of over spend campaigns are 5716 vs 5180 in control vs treatment,respectively

We can also look at the mean percentage for each group in overspending



...definitely a reduction in percentage of overspend (except for middle size)
but it is too early to tell

Hypothesis Testing for the new feature

We will begin by looking at an overall effectiveness and will then delve in to specific groups, a.k.a company size

For this purpose, we will first consider the overall propotion in the control and test for the campaigns which had an over spend in the budget

We will be using the z-test for this since we meet the conditions of it:

1. Your sample size is greater than 30.
2. Data points should be independent from each other. In other words, one data point isn't related or doesn't affect another data point.
3. Your data should be normally distributed. However, for large sample sizes (over 30) this doesn't always matter.
4. Your data should be randomly selected from a population, where each item has an equal chance of being selected.
5. Sample sizes should be equal or close if at all possible.

This will be a one-tail z-test since we are only concerned with control if it is larger or not than test

Setting the null and alternative Hypothesis--

```
H[null] = There is no difference between the propotions of the two groups, i.e.
any change in proportions is due to chance whereas the alternative hypothesis is that
control group has a higher propotion than test
```

$$H_0 : proportion_{control} = proportion_{test}$$

$$H_a : proportion_{control} > proportion_{test}$$

```
Setting a significance level to 0.05

overspend_control, sample_size_control = (5716, 7733)
overspend_test, sample_size_test = (5180, 7741)

Running the z-test...

z_stat: 9.540, p_value: 0.000

Reject the null hypothesis - suggest the alternative hypothesis is true
```

The p-value suggests that the we can reject the null hypothesis, we can infer that the change in proportions of over spending in test is not due to chance and the new feature is definitely working

Similar to our above methodology, we can use z-test for each company size to highlight the effectiveness of the new feature

---

Setting up z-test for small companies

Setting the null and alternative Hypothesis--

```
H[null] = There is no difference between the propotions of the two groups
of small companies, i.e.
any change in proportions is due to chance whereas the alternative hypothesis is that
control group has a higher propotion than test
```

$$H_0 : proportion_{control} = proportion_{test}$$

$$H_a : proportion_{control} > proportion_{test}$$

```
Setting a significance level to 0.05
```

```
Running the z-test...

z_stat: 8.697, p_value: 0.000

Reject the null hypothesis - suggest the alternative hypothesis is true
```

The p-value suggests that the we can reject the null hypothesis, we can infer that the change in proportions of over spending in test is not due to chance and the new feature is definitely working

---

Setting up z-test for medium size companies

Setting the null and alternative Hypothesis--

```
H[null] = There is no difference between the propotions of the two groups
of medium size companies, i.e.
any change in proportions is due to chance whereas the alternative hypothesis is that
control group has a higher propotion than test
```

$$H_0 : proportion_{control} = proportion_{test}$$

$$H_a : proportion_{control} > proportion_{test}$$

```
Setting a significance level to 0.05

Running the z-test...

z_stat: 0.645, p_value: 0.259
```

Fail to reject the null hypothesis - we have nothing else to say

The p-value suggests that the we cannot reject the null hypothesis, we can infer that any change in proportions of over spending in the test is due to chance and the new feature is not working for medium size companies

---

Setting up z-test for large size companies

Setting the null and alternative Hypothesis--

```
H[null] = There is no difference between the propotions of the two groups
of large companies, i.e.
any change in proportions is due to chance whereas the alternative hypothesis is that
control group has a higher propotion than test
```

$$H_0 : proportion_{control} = proportion_{test}$$

$$H_a : proportion_{control} > proportion_{test}$$

```
Setting a significance level to 0.05

Running the z-test...

z_stat: 5.964, p_value: 0.000

Reject the null hypothesis - suggest the alternative hypothesis is true
```

The p-value suggests that the we can reject the null hypothesis, we can infer that the change in proportions of over spending in test is not due to chance and the new feature is definitely working

---

Checking for any changes in budget in the Experiment

```
Here we will try to determine if there is intentional lowering of Budgets
```

For this purpose, we will analyze the campaign budgets set by teams in overspend_control and test group

We will again be using the z-test for this since we meet the conditions.

This will be a one-tail z-test since we are only concerned with if budget in contol is larger than test

Setting the null and alternative Hypothesis--

```
H[null] = There is no difference between the mean of budgets of the two groups, i.e.
any change in budget is due to chance, whereas the alternative hypothesis is that
control group has a higher budget than test [intentional lowering for the latter]
```

$$H_0 : meanofbudget_{control} = meanofbudget_{test}$$

$$H_a : meanofbudget_{control} > meanofbudget_{test}$$

```
Setting a significance level to 0.05

Running the z-test...

ztest and p-value:

 -1.4196728095175308  0.9221485212106051
```

accept null hypothesis

The p-value suggests that the we cannot reject the null hypothesis, we can refer that any change in setting up campaign budge in test is due to chance and intentional lowering of budget has been done

---