Sprint Project A/B Testing

GOAL OF THE TEST

The main goal of this A/B test was to test which of the 3 campaigns resulted in the highest sales.

> *Null-Hypothesis*: There is NO statistically significant difference between the campaigns in terms of generated sales. In other words, they are equally successful.
>
> *Alternate Hypothesis*: There is a statistically significant difference between the campaigns in terms of generated sales. In other words, one campaign performs better than the other.

For the analysis of A/B test results we will use a confidence level of 99% because, as we are comparing multiple campaigns in pairs, we suffer from the multiple testing problem.

TARGET METRIC

The target metric will be aggregated sales per campaign.

DATA VALIDATION AND EXPLORATION

Check how many distinct locations were in the trial:

count_distinct_locations
137

```
SELECT
    COUNT (DISTINCT location_id) AS count_distinct_locations

FROM `tc-da-1.turing_data_analytics.wa_marketing_campaign`
```

Check how many locations we had per campaign:

count_campaign1	count_campaign2	count_campaign3
43	47	47

```
SELECT
  SUM(CASE WHEN promotion = 1 THEN 1 ELSE 0 END) AS count_campaign1,
  SUM(CASE WHEN promotion = 2 THEN 1 ELSE 0 END) AS count_campaign2,
  SUM(CASE WHEN promotion = 3 THEN 1 ELSE 0 END) AS count_campaign3

FROM(SELECT
      location_id,
      promotion,
      Round(sum(sales_in_thousands),2) AS sales
    FROM `tc-da-1.turing_data_analytics.wa_marketing_campaign`
    GROUP BY location_id, promotion
    )
```

A perfectly equal distribution is not possible (137 locations / 3 = 45.6 stores thus, at best, we could do 45+46+46 stores for each campaign). But what we have is a good enough distribution.

Compare total sales per campaign

sales_campaign1	sales_campaign2	sales_campaign3
9993.03	8897.93	10408.52
= 34% of Total	= 30% of Total	= 36% of Total

```
SELECT
    SUM(CASE WHEN promotion = 1 THEN sales ELSE 0 END) AS sales_campaign1,
    SUM(CASE WHEN promotion = 2 THEN sales ELSE 0 END) AS sales_campaign2,
    SUM(CASE WHEN promotion = 3 THEN sales ELSE 0 END) AS sales_campaign3

FROM(SELECT
        location_id,
        promotion,
        ROUND(SUM(sales_in_thousands),2) AS sales
    FROM `tc-da-1.turing_data_analytics.wa_marketing_campaign`
    GROUP BY location_id, promotion
    )
```

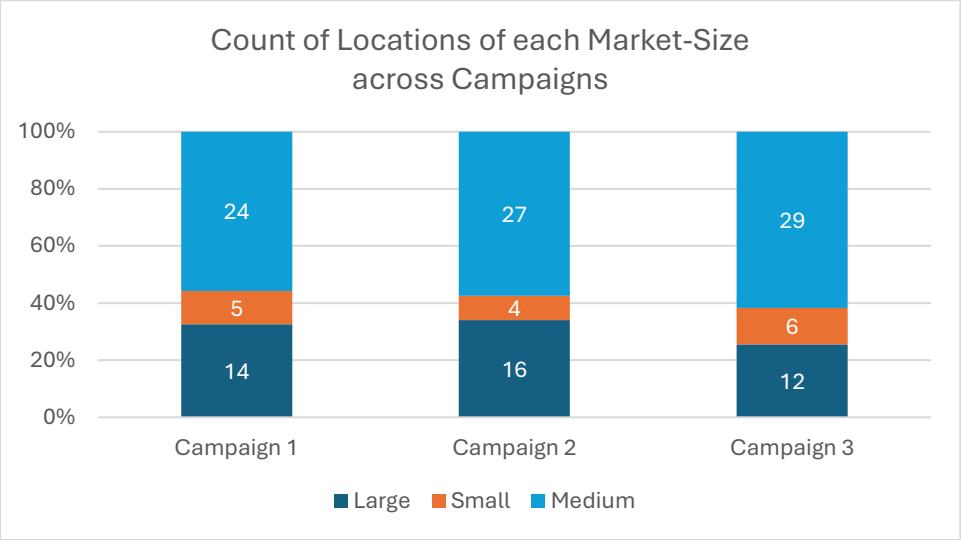
Sales are roughly the same among all 3 campaigns whereby campaign 3 generated the highest sales.

Compare Market-Size distribution among the campaigns

market_size	count_c1	count_c2	count_c3
Large	14	16	12
Small	5	4	6
Medium	24	27	29

```
SELECT
  market_size,
  COUNTIF(promotion = 1) AS count_c1,
  COUNTIF(promotion = 2 ) AS count_c2,
  COUNTIF(promotion = 3 ) AS count_c3,

FROM(SELECT
        DISTINCT location_id,
        promotion,
        market_size
    FROM `tc-da-1.turing_data_analytics.wa_marketing_campaign`
    )
GROUP BY market_size
```



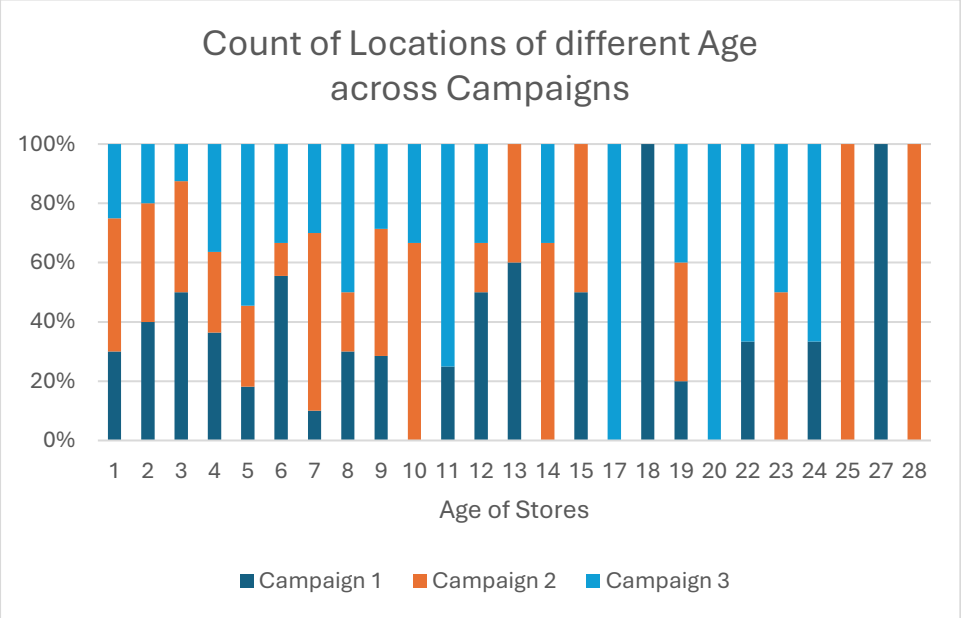
The medium market size occupied the biggest share among all three campaign groups and the small market size occupied the least. But all three market sizes are roughly equally distributed among each campaign.

	Campaign 1	Campaign 2	Campaign 3		Total Avg
Large	33%	34%	26%	➡	31%
Small	12%	9%	13%	➡	11%
Medium	56%	57%	62%	➡	58%
	100%	100%	100%		100%

Compare Age distribution among the campaigns

	Count Locations		
age_of_store	Campaign 1	Campaign 2	Campaign 3
1	24	36	20
2	8	8	4
3	16	12	4
4	16	12	16
5	8	12	24
6	20	4	12
7	4	24	12
8	12	8	20
9	8	12	8
10	0	16	8
11	4	0	12
12	12	4	8
13	12	8	0
14	0	8	4
15	4	4	0
17	0	0	4
18	8	0	0
19	4	8	8
20	0	0	4
22	4	0	8
23	0	4	4
24	4	0	8
25	0	4	0
27	4	0	0
28	0	4	0

```sql
SELECT
    age_of_store,
    COUNTIF(promotion = 1) AS count_c1_locations,
    COUNTIF(promotion = 2) AS count_c2_locations,
    COUNTIF(promotion = 3) AS count_c3_locations,
FROM `tc-da-1.turing_data_analytics.wa_marketing_campaign`
GROUP BY age_of_store
ORDER BY age_of_store
```



We can see that not every age group is equally distributed among the three campaigns. For example, we have 4 stores with the age of 28 but all 4 of them are in campaign 2.

Let's double check that insight with Quartiles:

	Min_Age	1st_Quartile	Median	Mean	3rd_Quartile	Max_Age
Campaign 1	1	3	6	8.3	12	27
Campaign 2	1	3	7	8	10	28
Campaign 3	1	5	8	9.2	12	24

```sql
SELECT
    promotion AS Campaign,
    MIN(age_of_store) AS Min_Age,
    APPROX_QUANTILES(age_of_store, 100)[OFFSET(25)] AS first_Quartile,
    APPROX_QUANTILES(age_of_store, 100)[OFFSET(50)] AS Median,
    ROUND(AVG(age_of_store),1) AS Mean,
    APPROX_QUANTILES(age_of_store, 100)[OFFSET(75)] AS third_Quartile,
    MAX(age_of_store) AS Max_Age

FROM `tc-da-1.turing_data_analytics.wa_marketing_campaign`
GROUP BY promotion
```

Looking at the Quartiles distribution, you can see that the values are roughly the same for each campaign. The majority of the stores (within each campaign) are between 10 – 12 years old or younger. Overall, we are happy enough with the age distribution among the 3 campaigns.

CONCLUSION: the distribution of various variables is very roughly equal among all three campaigns. Hence, as a sample ratio mismatch is outside of the scope of this sprint, we consider the samples as similar, and the A/B testing results will be considered to be meaningful.

CALCULATIONS T-TEST

First, we are aggregating sales of all four weeks per location_ID. The resulting table has 137 rows, which is in alignment with the above data validation queries, telling us we have 137 distinct stores in our sample.

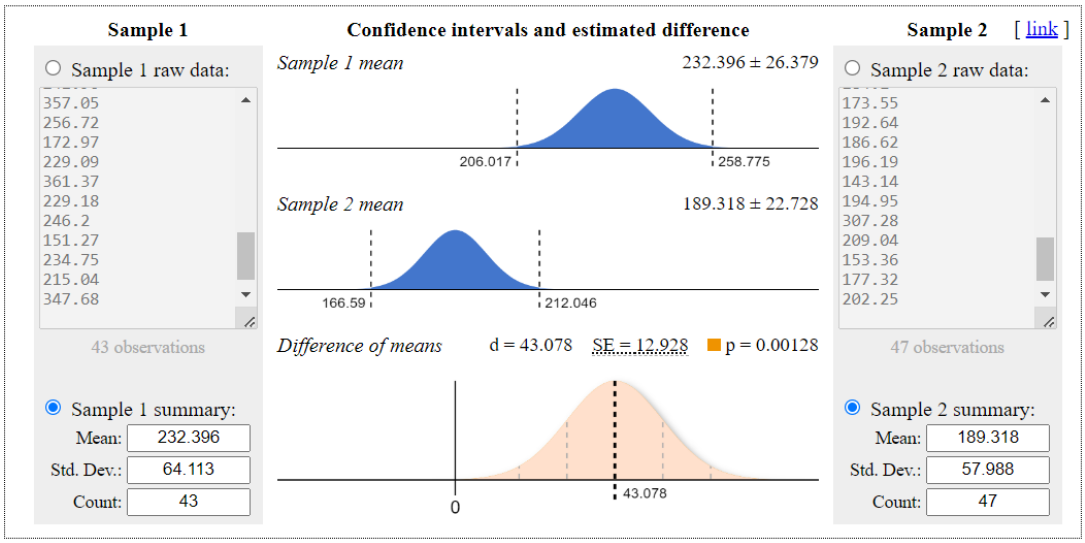
```
SELECT
  location_id,
  promotion,
  ROUND(SUM(sales_in_thousands),2) AS sales
FROM `tc-da-1.turing_data_analytics.wa_marketing_campaign`
GROUP BY location_id, promotion
```

Second, we have to conduct a two-tailed t-test, with (as mentioned in the “Goal of the test” paragraph) a Confidence Interval of 99%. A two-tailed test, in contrast to a one-tailed t-test, is appropriate if you want to determine if there is any difference between groups you are comparing, which is our case.

As a repetition, our Null Hypothesis claims that there is no statistically significant difference in the performance of the 3 campaigns.

An online calculator (Evan Miller A/B test calculator) is used to determine the p-values.

t-test Campaign 1 (Sample 1 in picture) **vs. Campaign 2** (Sample 2 in picture)



Verdict: Sample 1 mean is greater

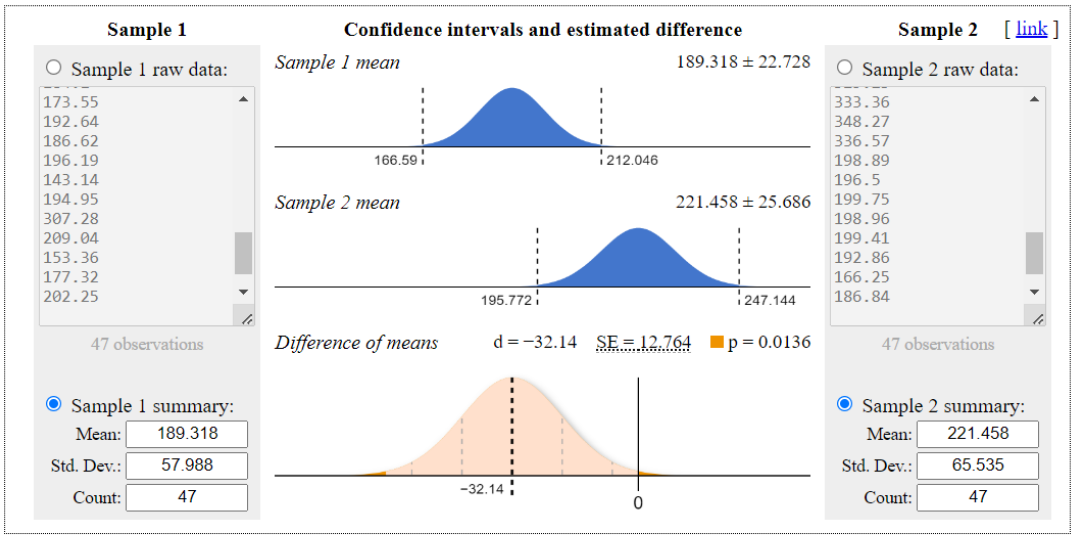
Hypothesis: d = 0 d ≤ 0 d ≥ 0

Confidence: 99%

We have a significant difference!

The p-value is 0.00128, a significantly small p-value, indicating that we have strong evidence against the null hypothesis and that the difference between Campaign 1 and 2 is statistically significant. In other words, Campaign 1 is better than Campaign 2.

t-test Campaign 2 (Sample 1 in picture) **vs. Campaign 3** (Sample 2 in picture)



Verdict: No significant difference

Hypothesis: d = 0 d ≤ 0 d ≥ 0

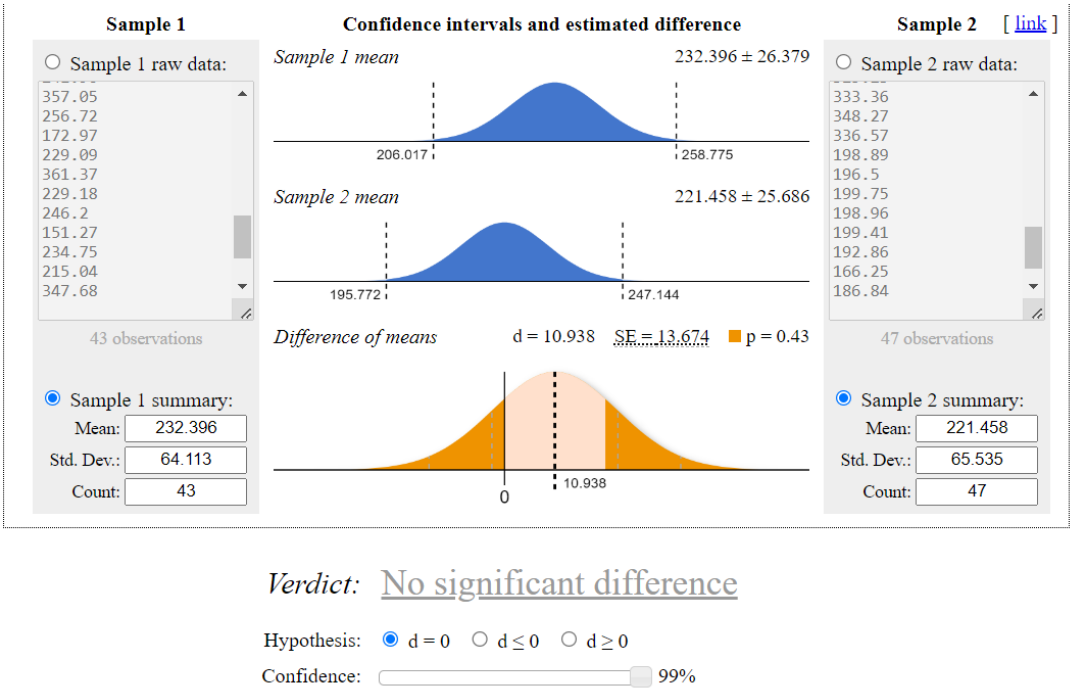
Confidence: 99%

We have no statistically significant difference.

The p-value is 0.0136, which is slightly more than 1% and hence over our 99% Confidence Interval threshold. Consequently, Campaign 2 and 3 are performing equally well.

(If we would be satisfied with a 95% Confidence Interval, then there would be a statistically significant difference, indicating that Campaign 3 performs better than Campaign 2)

t-test Campaign 1 (Sample 1 in picture) **vs. Campaign 3** (Sample 2 in picture)



Verdict: No significant difference

Hypothesis: d = 0 d ≤ 0 d ≥ 0

Confidence: 99%

Surprisingly, there is no statistically significant difference, with a p-value of 0.43 (well over the 1% threshold). Hence, Campaign 1 and 3 are equally good.

Based on the previous two t-tests, we would have expected that Campaign 1 is better than Campaign 3, simply because we already know that Campaign 3 and 2 are equally good and that Campaign 2 is worse than Campaign 1.

EVALUATION AND RECOMMENDATION

We have strong evidence that Campaign 1 is statistically more effective than Campaign 2.

However, Campaign 3 is statistically equally well performing as Campaign 1 and Campaign 2, which creates some discrepancies. Looking closer at the p-value between Campaign 3 and 2, we can see that we almost reached a statistically significant difference between them, meaning we almost were able to conclude that Campaign 3 was better than Campaign 2. This result would have been more in alignment with our expectations.

In conclusion, while Campaign 1 is clearly better than Campaign 2, it remains unclear whether Campaign 3 is truly equally well performing as Campaign 1 or as Campaign 2. Consequently, it remains unclear whether Campaign 1 is truly the overall best performing one or whether Campaign 1 and 3 are equally good. It is recommended to revisit the test design, especially as we know that we do have a small sample ratio mismatch, and to re-evaluate whether Campaign 1 is significantly better than Campaign 3.