COMPANY X A/B TESTING

JOEL DOMINGO

JOEL DOMINGO | JOELDOMINGO117@GMAIL.COM

INTRODUCTION

COMPANY X

Company X is a worldwide company that enables clients to make sophisticated multi-million-dollar marketing, pricing, staffing and operational decisions through offline A/B testing. Their customers are leaders in the retail, grocery, c-store and restaurant space.

As a data scientist on a team at **Company X**, this report will give an insight of our analysis (including methodology, results, and insights) to an audience of data scientists and business stakeholders.

PREMISE / THE EXPERIMENT

Suppose Company has a client, **Client X**, that sells a variety of snacks and beverages at all its stores. **Client X** suspects that by putting up new displays for **Brand Z's** candy, that consumers will purchase more of **Brand Z's** candy leading to higher revenue for candy. If **Client X's** intuition is correct, these new displays would lead to a multi-million dollar increase in total revenue across all stores; if they are incorrect, they will have wasted time and money putting up the new signs at best, and at worst they could see revenue from **Brand Z's** candy (or other candy brands) decrease.

To mitigate this risk, **Client X** decides to conduct an experiment in a subset of stores first before making the decision to roll out the new signs to all stores. **Client X** enlists the help of **Company X** to help them devise an experiment to detect whether the new signs will result in a ***statistically significant*** increase in revenue from **Brand Z's** candy.

NOTE: This report focuses on 3 proposed questions/problems regarding this scenario and will be aimed at presenting solutions to these problems.

PROBLEM STATEMENTS

QUESTION 1 – EXPERIMENT OVERVIEW

Outline the steps of an experiment, from start to finish, to detect whether putting up new displays for Brand Z's candy will result in an increase in revenue from Brand Z's candy. Suppose your audience for this outline is the client, who has some general statistical understanding (i.e., you don't need to get into the weeds of any techniques/algorithms) but is looking to understand more thoroughly how this experiment will be set up and analyzed from enlisted help—you, the expert data scientist.

QUESTION 2 – TEST/CONTROL SELECTION

One challenge in offline (brick and mortar) A/B testing compared to online A/B testing is that randomization is not possible at the customer level. Consequently, alternative methods must be used in choosing treatment and control groups. One way to overcome this challenge is to test at the store level and strategically select sets of stores to use as treatment and control groups.

Utilizing data and your statistical expertise, how could we intelligently improve upon random selection to obtain (1) a set of treatment stores that better represents the complete set of stores to which we want to roll out the changes, and (2) a set of control stores that are a better baseline for our experiment? What exact stores (by id) would you use in the control group and which in the treatment group for this experiment? Please explain how you came to this decision (e.g., methodology, algorithms, assumptions, etc.). Assume the data we've provided contains all stores.

QUESTION 3 – SIGNIFICANCE TESTING

NOTE: For this question, we will be referring to files provided in the assessment package.
Suppose we decided to conduct using the following stores:

- Treatment stores (q3_treatment_stores.csv)
- Control stores (qe_control_stores.csv)

Suppose also that we've completed collecting data for the experiment and are using the following time periods.

Pre-test period (X weeks prior to the start of the implementation	**Implementation period** (Period of time needed to get the experiment ready; in this case, the time to put the signs up across all stores)	**Test period** (X weeks after the end of the implementation period)
17/07/2016 – 15/10/2016	16/10/2016 – 12/11/2016	13/11/2016 – 12/02/2017

How would you quantify the impact of this test on revenue from **Brand Z's** candy sales? Should we recommend **Client X** roll out the new signs to all their stores? Explain your analysis and reasoning, and include any code used.

THE DATA

FILES PROVIDED

Brang Z's candies and products	Products_of_interest.csv
Treatment/test stores list	Q3_treatment_stores.csv
Control stores list	Q3_control_stores.csv
Store attributes	Store_attributes.csv
Transaction's data	Transactions.csv

DATA DICTIONARY (PROVIDED FILES)

STORE_ATTRIBUTES.CSV

STORE_ID	Store's unique identifier
ATTRIBUTE_ID	Name of the store attribute
ATTRIBUTE_TYPE	data type of the attributes value
ATTRIBUTE_INT_VAL	Value of the store attribute if attribute_type == integer
ATTRIBUTE_STR_VAL	Value of the store attribute if the attribute type == string
ATTRIBUTE_FLOAT_VAL	Value of the store attribute if the attribute_type == float

TRANSACTIONS.CSV

DATE_WEEK	Start date of the 7-day week the row's cumulative revenue value represents
STORE_ID	Store's unique identifier
PRODUCT_ID	Product's unique identifier
CURRENCY_CODE	Currency that the revenue is measured in
REVENUE	Revenue from sale of a specified product for a specific store in each week

FILES USED IN EACH PROBLEM

QUESTION → Files

QUESTION 1	- No files required
QUESTION 2	- products_of_interest.csv - transactions.csv
QUESTION 3	- Q3_treatment_stores.csv - Q3_control_stores.csv

QUESTION 1 – EXPERIMENT OVERVIEW

1. Outline the timeframes/time-periods for the following:

Time period	Description
Pre-trial period	Data from this period will allow us to select our test and control stores
Implementation period	Time needed to prepare the experiment in the stores
Trial / Test period	Data from this period will be used to evaluate the performance of the test stores against the selected control stores

2. Identifying potential test and control stores:

Trial stores should be established in this stage. The range of selection should ensure that most, if not all, segments/demographics or store profiles have representation within the trial. At the same time, control stores should be established based off their closeness (based off an agreed upon metric) in behavior to the trial stores.

For each trial store, 1 control store must be selected which is similar in characteristics. This 1-to-1 rule must be adhered to as to not risk skewing of the results. The method to which each control store is selected can be done mathematically to precisely select the closest performing store to each trial store.

Each **control store** will exist as it has previously (that is to operate as normal), with the change in strategy not applied during the implementation and testing period. Whereas each **trial store** will have the change applied, such change being the modification and implementation of **Brand Z displays.**

3. Prepare and conduct the experiment:

Allow the test to unfold during the trial period, making sure that during the pre-trial and trial period there are **no isolated/abnormal special offers** which can be misinterpreted as the effects of the experiment at either the control or trial stores. The data from both the trial and control stores will be collected during this period and will be used to evaluate the performance of the trial.

4. Post-trial evaluation:

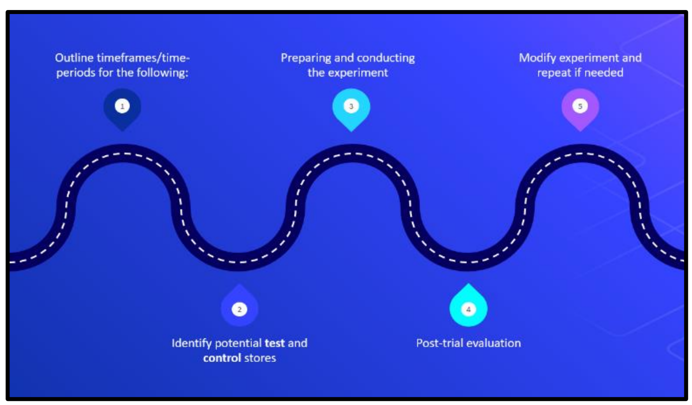
This stage focuses on the evaluation of data taken during the trial period. Performance data of the stores from the trial period will be collected from both the trial and control stores. The data collected will be filtered to relevant data pertaining to the **Brand Z products**. This detail can be modified should circumstances change or in future test runs.

The differences between each trial store and the corresponding control store will be evaluated. These differences will be determined to be either statistically significant (that is, we can confidently say these differences **were the effect of the trial**, and not by chance), or **insignificant**.

Determining a **significant** improvement will mean the experiment shows *successful uplift* and should be implemented in similar stores. **Significant decreases in performance** will mean the experiment has determined the strategy/change has inhibited performance and therefore should **not** be implemented in similar stores in the future under current circumstances.

5. Change parameters of the experiment and repeat if necessary

In the case that the differences are not statistically significant (or are inconclusive), we may need to change the parameters of the experiment (i.e., extending the time of the trial, including more test/control stores etc.)



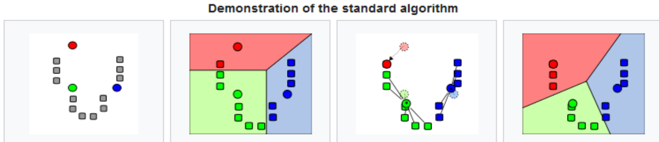
QUESTION 2 – TEST / CONTROL SELECTION

Part 1 **–** *'How could we intelligently improve upon random selection to obtain a set of* ***treatment stores*** *that better represents the complete set of stores that we want to roll out changes?'*

Under a generic approach, random selection of stores would be appropriate. However, due to the financial risk involved, it is essential to optimize our testing procedure to ensure that we attain reliable results and data to inform the business decision relating strategy. Therefore, to improve upon random selection, we can use the clustering algorithm known as K-Means.

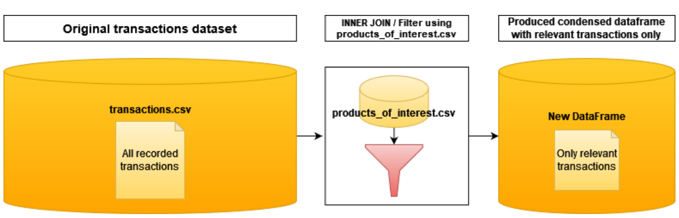
K-Means clustering is a machine learning algorithm which uses **vector quantization** with the purpose of partitioning *n* observations into *k* clusters, with *n* being the number of data points in the dataset, and *k* being the number of specified groupings.

Demonstration of the standard algorithm

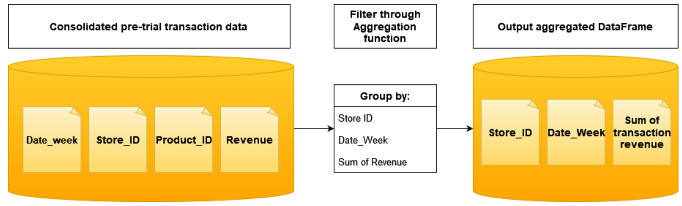


1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).
2. *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.
3. The centroid of each of the *k* clusters becomes the new mean.
4. Steps 2 and 3 are repeated until convergence has been reached.

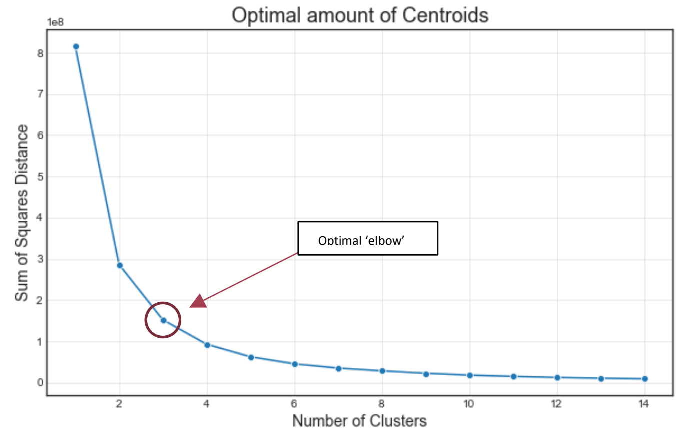
Prior to working with the K-Means clustering algorithm we were first required to isolate and clean the data that's relevant to this part of the problem. Using the transactions and relevant products dataset (*transactions.csv and products_of_interest.csv*), we can combine the two datasets and create one which represents the unique vectors pertaining to each store's performance. This can be done by only isolating transactions which involve only the products of interest, with the aim of de-noising our data.



Once the data is consolidated, we can use the ***date_week*** column to identify which week in the pre-trial data the transactions were completed, and aggregate them, thus having a complete time-series dataset aggregated by weekly revenue by store number.



When working with this algorithm under the Sci-kit Learn Python library, we require the input *k* to apply to *n* datapoints. To ascertain what our optimal number of *k* is, we can apply the ***'elbow method'***, which displayed the following results.



Using the elbow method, we can determine the optimal number of centroids or clusters to be between 2 and 4 clusters. With the information provided, we have decided that the data of stores can be split into 3 groups/clusters:

- Small stores / Low selling stores
- Medium stores / Average selling stores
- Large stores / High selling stores

The algorithm can therefore be run with the parameter ***n_clusters* = 3**. Once the model is initialized and fit to our data, converging on the 3 separate classes, we can label each week for any store with one of the 3 classes, then aggregate a final class/label for the store based off the modal value. An example is shown below (Clusters = **0** | **1** | **2**):

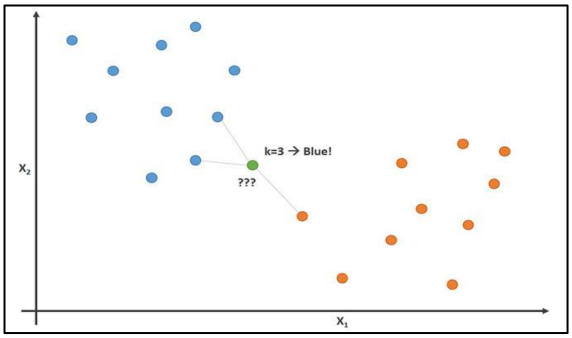
STORE ID	WEEK NUMBER	REVENUE	CLUSTER LABEL		MODAL CLUSTER
339	1	$ 1,216.29	1	What is the modal value	1
339	2	$ 971.42	1		
339	3	$ 988.36	1		
339	4	$ 1,056.12	1		
339	5	$ 1,219.31	1		
339	6	$ 936.02	1		
339	7	$ 1,106.60	1		
339	8	$ 1,175.39	1		
339	9	$ 1,205.55	1		
339	10	$ 1,351.91	0		
339	11	$ 1,045.50	1		
339	12	$ 1,072.45	1		
339	13	$ 926.21	1		

From here, we can select our trial stores, making sure to select ***at least*** one trial store from each existing cluster (clusters **zero, one or two**), to maximize the effectiveness of our strategy test.

Part 2 – ‘How could we intelligently improve upon random selection to obtain a set of corresponding control stores that are a better baseline for our experiment?

Once trial stores that represent the overall population have been established (i.e., trial stores representing each cluster/centroid group), we can establish a corresponding control store using the Nearest Neighbors algorithm.

This supervised learning algorithm is a classification technique that works by calculating the Euclidean distance between points on an *n* dimensional plane. The classification itself is also dependent on the parameter *k*, that being the number of neighbors or points it calculates the distance for at each iteration. The majority classification of **the nearest** *k* neighbors determines the classification of the current point on the plane, as shown below.

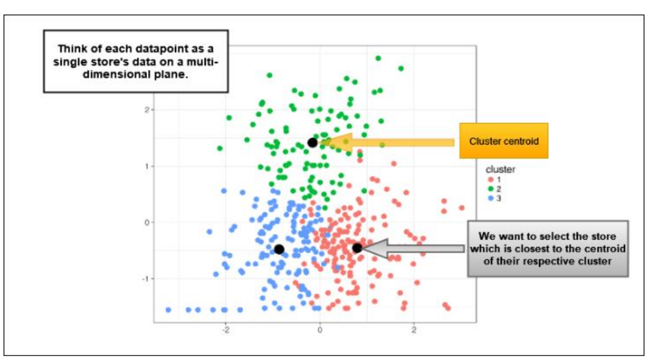


Although this is primarily a classification technique, we can repurpose this algorithm to fulfill our purposes of finding a control store. We must first understand that with the transformation of data we have performed so far, we have manipulated the data in such a way that each store has a revenue vector. Therefore, our solution to finding the most appropriate control store lays in the discovery of the store (within the same cluster) that has the closest Euclidean distance (or closest vector) to our chosen trial store, i.e., *what is the nearest neighbor to our trial store?*

Part 3 – ‘What are our chosen trial stores?

To ensure our business decision is made with the most information, we will use the K-Means clustering method outlined in part 1 and select 3 trial stores, one for each known cluster (small stores, medium stores and large stores). Once we have each cluster, we can separate their data and look at their weekly revenue.

For the purpose of covering as much information in the cluster as possible, we will identify our trial store based off it’s closeness to the center of the cluster. To do this, we will use the weekly revenue average of each store, and whichever is the closest to the population (cluster) average will be selected as the trial store.



The technical methods of how to transform the data can be found in the provided Jupyter Notebooks and Excel files.

The following show our results for the chosen trial stores:

Cluster Number	Population Average of Weekly Revenue ($)	Chosen Trial Store and AWR (Average Weekly Revenue)
Zero (0)	$1,526	**Store 382** with an AWR of **$1,525**
One (1)	$977	**Store 528** with an AWR of **$980**
Two (2)	$1,450	**Store 511** with an AWR of **$1,214**

	Store ID	1	2	3	4	5	6	7	8	9	10	11	12	13
0	382	1656.16	1455.67	1467.76	1424.14	1499.70	1486.68	1358.98	1648.40	1601.75	1713.10	1611.70	1483.53	1423.76
1	528	1110.99	937.21	1101.51	1067.19	955.40	917.08	1014.10	1050.49	928.44	974.68	894.57	911.17	880.77
2	511	1290.36	1248.46	1163.26	1287.34	1160.28	1159.02	1217.96	1321.20	1304.37	1243.14	1164.17	1201.23	1032.10

Part 4 – ‘What are the chosen control stores?’

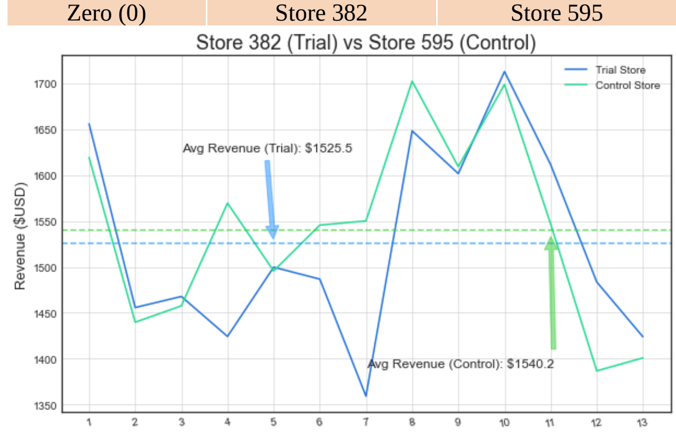
Using the method outlined in part 2, we can use the Nearest Neighbors algorithm to find the store with the closest Euclidean distance to each trial store (each representing one of the 3 clusters).

Prior to utilizing the algorithm, we were required to first clean the data. Within the provided transaction data, there exists stores which have weeks without any revenue from Brand Z candy. In this case, we were simply required to fill the np.NaN (null values) with zero, and therefore remove these rows (stores) from our data.

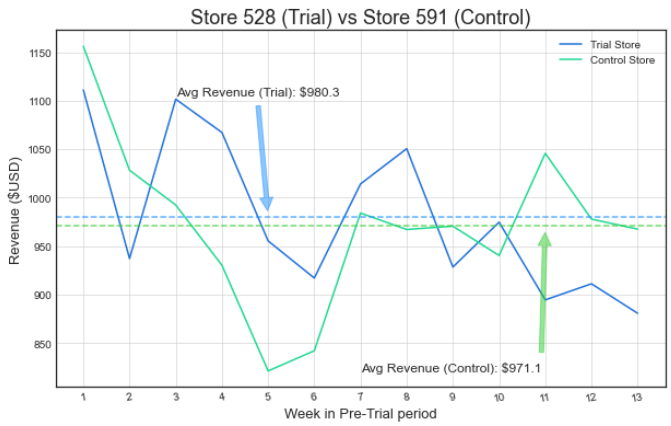
As mentioned, we set the *n_neighbors* parameter = 1, to find the single closest neighbor to the trial store. From here, the process is simple. Once the Neighbors model has been initialized to the correct parameter, we can assign our X value which the model will be told to identify. In this case, this is the row/vector of our chosen trial store.

Once the model has identified the trial store, we can fit the data to the rest of the training samples (cluster X data) and look at each stores Euclidean distance to store X. The training is hence complete, and we can output the nearest store index in the data. The following results show the chosen control stores as it pertains to each cluster trial store.

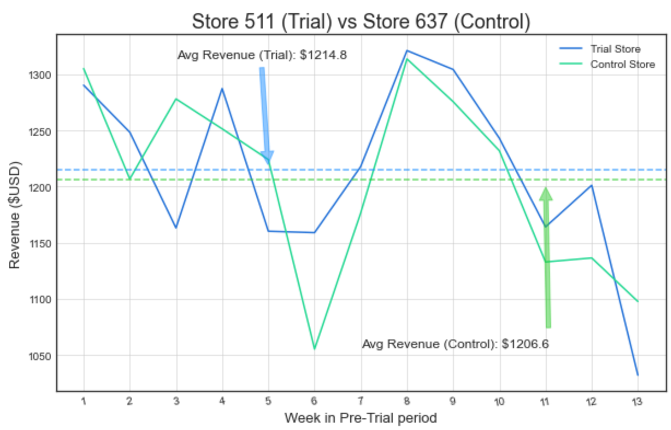
Cluster Number	Cluster Trial Store	Nearest Neighbor / Control Store
Zero (0)	Store 382	Store 595



Cluster Number	Cluster Trial Store	Nearest Neighbor / Control Store
One (1)	Store 528	Store 591



Cluster Number	Cluster Trial Store	Nearest Neighbor / Control Store
Two	Store 511	Store 637



From the trial stores of 382, 528 and 511 for clusters zero, one and two, we were able to find suitable control stores by calculating stores with the closest vectors via Euclidean distance.

Should the parameters of the experiment change, this process is reproducible and will result in similarly reliable results. However, this experiment cannot be reliable if it is decided that multiple control stores are chosen for one trial store. If there is a cluster which only has x amount of suitable control stores to the trial store, the algorithm will calculate x amount of control stores, one of which may be an outlier, thus skewing the aggregating results of the A/B test.

<u>Notes, assumptions and potential improvements:</u>

- When selecting and preparing the data for clustering, we filtered the data based off transaction data which applied only to **Brand Z's products**. This could be changed to include all transactions, to which the trial and control pairing would be different.

- Part 3 required that the data by **Store ID** was aggregated by their revenue by weeks during only the trial period.

- To determine the appropriate number of clusters, the elbow method was used. The 'elbow' we interpreted to be between 2 and 4 centroids, and in this case, we settled on 3 classes. Should this experiment be modified, the number of centroids can be changed. This would also change the number of control and trial stores required.

- Given time constraints, we had chosen 1 test and control pairing per cluster. However, should the experiment be modified for further accuracy and reliability, it would be wise to select multiple trial stores per cluster, and therefore multiple control stores. Doing this ensures we have data more representative of the overall population of stores.

- When selecting a trial store from their respective clusters, we had chosen to use the average based off weekly revenue. There may be a more statistically accurate way to determine the closest datapoint to the centroid, which was not used, and leaves room for improvement.

- We calculated these stores under the assumption that there are no special offers or isolated changes in behavior during the trial period for all stores, which could skew and compromise the integrity of the data.

<u>Next Steps:</u>

Moving on to question 3, we would be able to use the selected trial and control stores to perform an A/B test. Despite being given the control and trial list in the next part of this assignment, we can assume this method outlined in this notebook is just as reliable at finding suitable control stores.

QUESTION 3 – QUANTIFYING TEST IMPACT

*‘How would you quantify the impact of this test on revenue from Brand Z’s candy sales*__*? Should we recommend Client X roll out the new signs to all their stores?*__*’*

During the months of the test period, we can analyze the performance difference between each trial store and their corresponding control store. Similarly, to the earlier sections of this report, we can segment the data in a way that looks at purely the products of interest and the revenue which comes from these products.

To both quantify and inform our recommendations, we can do a T-Test. This is an inferential statistic that is used to determine whether there is a **significant difference between the mean of two groups and how they are related.** It is a method of hypothesis testing that ultimately describes whether the difference between two groups can be sufficiently explained by a specific change applied to one of the groups, in this case, the trial group.

Having been given a list of trial and control pairs, and conscious of the time limitations of the experiment, we can save time by **assuming** there are 3 groups:

- Group 1 (smaller stores/low-selling stores)
- Group 2 (medium stores/medium-selling stores)
- Group 3 (large stores/high-selling stores)

Having these groups will allow us to more effectively understand which type of stores we should and should not implement the suggested changes into. For each group, we will have chosen 3 trial/control pairs in a similar manner to how we separated using clusters. The criterion for implementation is if we see any ‘significant uplift’ in 2/3 stores for that group. If trial **store X** has an uplift in sales higher than the 95^{th}% confidence interval of sales for the control store during the same period, we can confidently assume the uplift is due to the strategy changes in trial **store X**. Conversely, if a decrease in sales remains outside the 5^{th}% confidence interval of its control store, we can confidently assume the strategy had a negative impact on sales, assuming no other changes or outside forces had influenced this.

Furthermore, having 3 pairs in each cluster can help us produce a more informed result by using the modal result, and decreasing the result being of pure chance. For example, in smaller stores, if two trial stores show significant performance uplift (outside the 95^{th}% confidence interval) then we can say that the changes should be implemented in the smaller stores.

However, if less than 2/3 stores show significant uplift, or 2/3 show significant decreases (below the 5^{th}% confidence interval) in performance, then we can confidently recommend the strategy not be implemented in that cluster of stores. The stores will be chosen arbitrarily under the highlighted assumption.

The following stores per group have been selected:

Small Stores	
Trial Store	Control Store
Store 457	384
Store 431	539
Store 674	341

Medium Stores	
Trial Store	Control Store
Store 654	614
Store 392	363
Store 438	667

Large Stores	
Trial Store	Control Store
Store 408	584
Store 545	382
Store 609	521

Trial Store	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10	Week 11	Week 12	Week 13	Week 14	Avg rev	Group	Selection for relevant group
492	678.43	538.03	613.59	576.21	597.89	616.23	699.74	671.59	699.65	986.97	1048.64	910.02	855.23	78.15	683.5979	1	✗
642	794.37	917.98	751.9	787.39	678.77	636.78	577.19	589.48	717	777.65	889.35	830.26	869.1	126.19	710.2436	1	✗
582	910.79	783.11	794.35	795.67	677.9	538.57	594.22	707.82	782.27	826.4	973.55	1070.31	775.22	120.22	739.3143	1	✗
457	858.55	774.14	841.1	790.88	899.3	712.35	653.72	811.1	799.3	726.29	920.37	919.34	826.88	93.04	759.0257	1	✓
431	1166.59	1128.58	997.12	895.83	890.32	830.03	779.24	927.67	1128.91	985.94	1021.15	999.86	1138.84	127.55	929.8307	1	✓
674	1104.4	1460.58	1036.16	823.63	811.75	803.1	1316.41	891.72	1140.64	923.04	927.78	989.14	856.41	120.47	943.2307	1	✓
565	1116.66	1116.84	1059.78	913.85	793.86	808.99	874.11	971.61	1001.64	1095.7	1147.93	1107.11	1178.22	147.27	952.3979	1	✗
399	1009.06	1289.08	972.55	1032.69	933.16	833.34	837.78	920.23	1082.65	1127.23	1143.28	1192.52	1252.88	172.59	985.6457	1	✗
436	1195.25	1360.28	954.69	989.98	1090.73	1100.3	1031.3	936.07	1070.01	1209.56	1308.68	1239.8	1130.28	160.48	1055.529	2	✗
454	1218.91	1247.81	1107.03	1096.4	1102.17	1089.16	878.04	969.17	1088.26	1139.61	1222.15	1424.65	1145.06	171.09	1062.822	2	✗
344	1420.55	1173.26	1021.92	1094.65	1039.22	940.83	806.61	1097.1	1260.11	1183.78	1359.03	1385.74	1209.94	113.7	1079.031	2	✗
366	1391.01	1181.96	1149.95	1133.88	1094.74	974.73	839.79	1083.19	1343.77	1173.44	1193.44	1308.48	1338.72	187.83	1099.638	2	✗
654	1323.85	1394.72	1153	1084.16	963.11	926	900.06	1132.18	1312.38	1250.63	1544.9	1485.32	1169.61	133.75	1126.691	2	✓
392	1101.31	1169.55	1118.15	1185.11	1230.62	1236.79	1078.27	1125.33	1233.05	1395.5	1222.29	1402.84	1380.1	216	1149.636	2	✓
438	1433.99	1304.57	1162.46	1194.98	1154.83	993.27	935.87	1078.57	1259.22	1469.81	1431.58	1560.86	1374.23	119.79	1176.716	2	✓
424	1178.59	1227.88	1239.56	1222.61	1303.27	1279.35	1134.36	1184.75	1480.76	1377.82	1406.53	1463.34	1424.05	205.1	1223.426	2	✗
367	1371.46	1344.02	1202.97	1156.99	1235.56	1128.97	1160.27	1333.89	1279.18	1458.66	1325.44	1551.08	1290.43	306.37	1224.664	2	✗
458	1388.16	1285.93	1248.77	1173.18	1158.55	1123.07	1121.14	1471.54	1397.78	1528.03	1359.76	1687.13	1663.18	218.81	1273.216	2	✗
505	1486.46	1336.83	1458.74	1374.71	1321.64	1170.53	856.98	1132.13	1517.43	1563.96	1805.56	1772.48	1466.1	148.63	1315.156	3	✗
651	1535.57	1512.87	1410.99	1348.98	1251.25	1202.19	1241.5	1102.85	1521.78	1456.97	1450.98	1768.99	1489.03	187.78	1320.124	3	✗
408	1674.11	1415.1	1491.69	1521.01	1518.35	1220.89	1030.56	1238.5	1593.58	1724.62	1574.65	1620.07	1605.39	190.03	1387.039	3	✓
545	1444.72	2062.01	1493.21	1383.94	1326.49	1377.45	1660.85	1564.41	1397.73	1301.38	1402.03	1531.29	1534.5	172.34	1403.739	3	✓
609	1870.55	1761.64	1762.73	1746.05	1621.11	1435.66	1258.26	1419.94	1699.77	1813.39	1759.14	1934.58	1789.58	254.47	1580.491	3	✓
502	1822.16	2002.61	1803.81	1834.93	1729.77	1571.98	1081.14	1602.94	1916.9	2104.14	2150.34	2115.33	1890.69	175.28	1700.144	3	✗
451	2393.61	1901.92	1715.61	2010.32	2116.27	1473.85	1343.07	1822.1	2029.79	2225.04	2048.27	1979.89	2278.96	273.88	1829.47	3	✗

The trial stores have been **SORTED** by **WEEKLY AVERAGE REVENUE** and subsequently split into 3 groups **(small, medium and large)**

for each group are selected based off their closeness to the center or population average for the

Performing the hypothesis tests – Group 1 (Small)

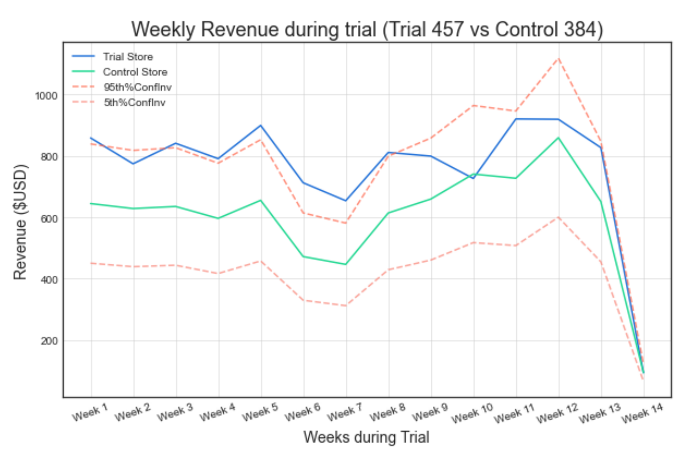
Small Stores (Group 1.1)	
Trial Store 457	Control Store 384

Hypothesis

We will test with a null hypothesis that there is no difference between the trial store and control store during the trial period.

Results

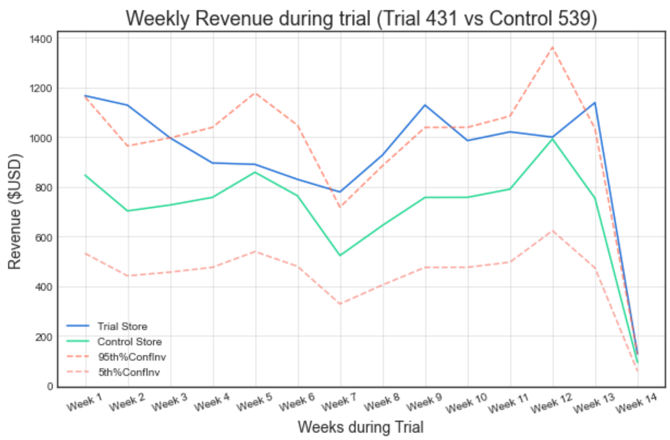
In this group, the trial only displayed ‘significant’ performance uplift in 46% of the weeks in the trial. We can conclude that it was not effective in this trial store and therefore fail to reject h0.



Small Stores (Group 1.2)	
Trial Store 431	Control Store 539

Hypothesis
We will test with a null hypothesis that there is no difference between the trial store and control store during the trial period.

Results
In this group, the trial only displayed 'significant' performance uplift in 46% of the weeks in the trial. We can conclude that it was not effective in this trial store and therefore fail to reject h0.



Small Stores (Group 1.3)	
Trial Store 674	Control Store 341

Hypothesis
We will test with a null hypothesis that there is no difference between the trial store and control store during the trial period.

Results
In this group, the trial showed 'significant performance' in only 46% of the weeks in the trial. Furthermore, the trial store performance was significant **below** the performance of the control store. We can conclude that it was not effective in this trial store and therefore fail to reject h0.

Group 1 Conclusion (Smaller Stores)

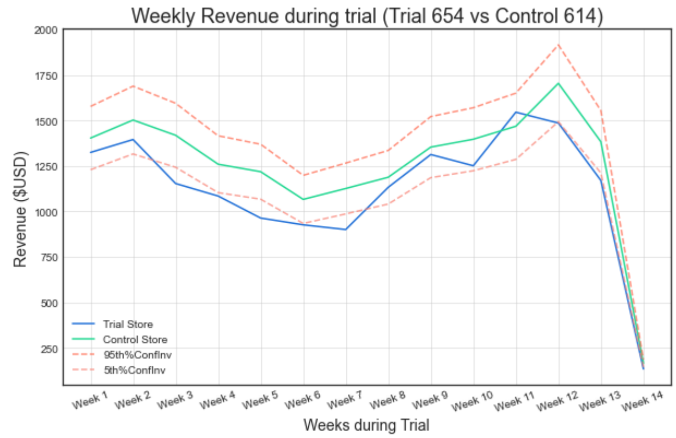
Smaller stores did not show any sustained significant uplift in any of the 3 trial stores compared to their respective control stores. As a result, we cannot confidently say that the strategy will be effective in other smaller stores. Therefore, we **do not recommend** spending **$X** on implementing the new strategy in additional smaller stores.

Performing the hypothesis tests – Group 2 (Medium)

Medium Stores (Group 2.1)	
Trial Store 654	Control Store 614

Hypothesis
We will test with a null hypothesis that there is no difference between the trial store and control store during the trial period.

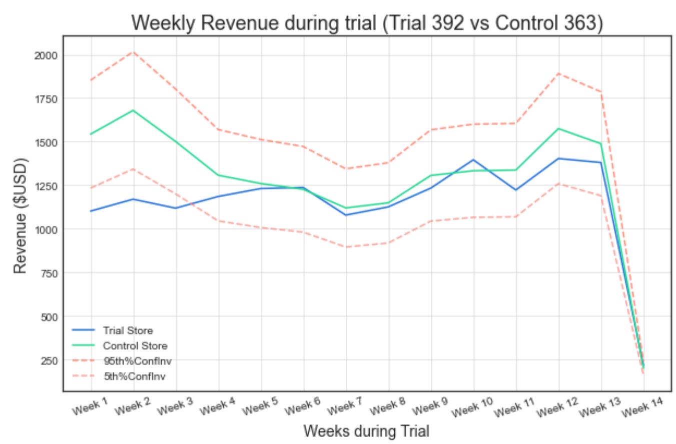
Results
In this group, the trial displayed 'significant' performance **decreases** in 54% of the weeks in the trial. We can conclude that it was **detrimental** to performance in this trial store and reject h0.



Medium Stores (Group 2.2)	
Trial Store 392	Control Store 363

Hypothesis
We will test with a null hypothesis that there is no difference between the trial store and control store during the trial period.

Results
In this group, the trial only displayed 'significant' performance **decreases** in 23% of the weeks in the trial. We can conclude that it was not effective in this trial store and therefore fail to reject h0.



Medium Stores (Group 2.3)	
Trial Store 438	Control Store 614

Hypothesis
We will test with a null hypothesis that there is no difference between the trial store and control store during the trial period.

Results
In this group, the trial only displayed 'significant' performance **uplift** in 31% of the weeks in the trial. We can conclude that it was not effective in this trial store and therefore fail to reject h0.

Group 2 Conclusion (Medium Stores)

Medium-sized stores did not show any sustained significant uplift in any of the 3 trial stores compared to their respective control stores. In fact, one group produced significant performance decreases as a result of the strategy. As a result, we cannot confidently say that the strategy will be effective in other medium stores. Therefore, we **do not recommend** spending **$X** on implementing the new strategy in additional medium-sized stores.

Performing the hypothesis tests – Group 3 (Large)

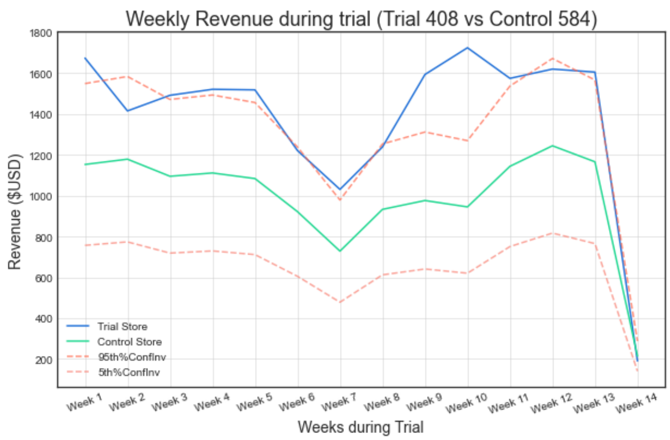
Large Stores (Group 3.1)	
Trial Store 408	Control Store 584

Hypothesis

We will test with a null hypothesis that there is no difference between the trial store and control store during the trial period.

Results

In this group, the trial displayed 'significant' performance **uplift** in 70% of the weeks in the trial. We can conclude that it was positively effective in this trial store and therefore reject h0.



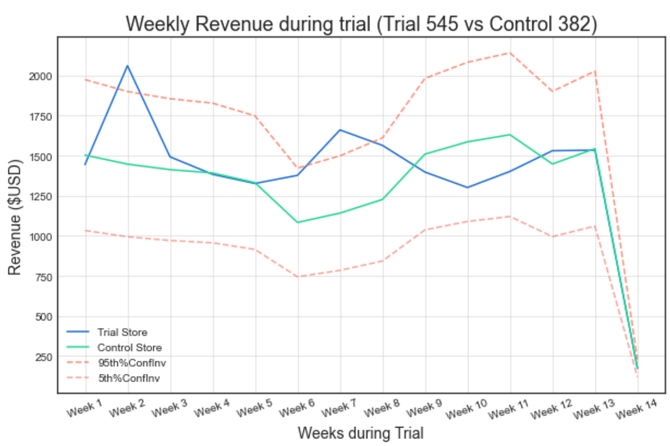
Large Stores (Group 3.2)	
Trial Store 545	Control Store 382

Hypothesis

We will test with a null hypothesis that there is no difference between the trial store and control store during the trial period.

Results

In this group, the trial displayed 'significant' performance **uplift** in only 15% of the weeks in the trial. We can conclude that it was not effective in this trial store and therefore fail to reject h0.



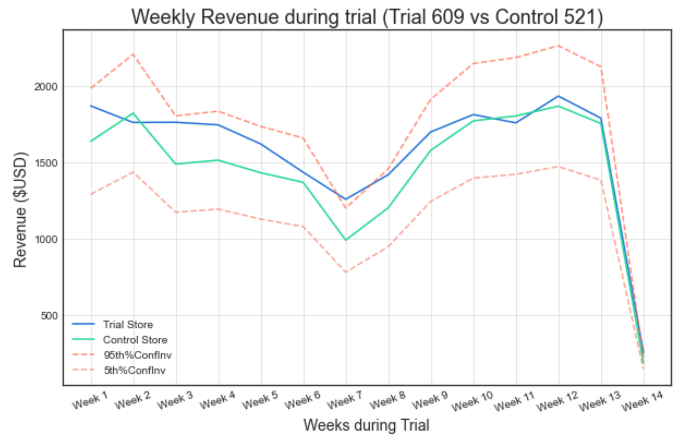
Large Stores (Group 3.3)	
Trial Store 609	Control Store 521

Hypothesis

We will test with a null hypothesis that there is no difference between the trial store and control store during the trial period.

Results

In this group, the trial displayed 'significant' performance **uplift** in only 7% of the weeks in the trial. We can conclude that it was not effective in this trial store and therefore fail to reject h0.



Group 3 Conclusion (Large Stores)

Large-sized stores only produced sustained **significant uplift** in only 1 of the 3 trial stores as a result of the implementation of the Brand Z candy strategy. Therefore, we do **not recommend** spending **$X** on implementing the new strategy in all larger stores.

However, the next steps should be to plan and identify whether value exists in testing the strategy in stores which are significantly similar to trial store 408 (group 3.1) and control store 585, since the effect of the strategy showed promising **sustained significant uplift** as a result of the strategy implementation.

Should a store architype produce sustainable **significant uplift** outside the 95th% confidence interval, we can quantify this monetarily by looking at the average percentage increase in revenue within those stores as a result of the strategy change. For example, if large stores display significant uplift in all 3 pairs of stores with an average revenue increase of 25%, we can confidently say that the implementation of strategy within further stores of similar characteristics will be forecasted to produce an estimated 20-25% in additional sales revenue. Alternatively, the same can be said for significant performance decreases and therefore should not be implemented.

Final Recommendation

Given the results of our A/B testing on the pre-determined trial and control store, neither group 1,2 or 3 produced enough evidence of sustained 'significant' uplift. We found that none of the groups showed differences in performance enough so that they were influenced by the proposed strategy to implement higher frequencies of **Brand Z candy** with the purpose of increasing sales. Therefore, our final recommendation is to **not** move forward with the strategy implementation in further stores, as it presents a high volume of financial risk. Alternatively, given more information, the processes and testing strategies performed in this experiment may be repeated with alternative parameters to further inform future decisions in this topic.

Source code and references:

Question 2 Notebook

Part 1:
https://github.com/jdomingo117/ProjectsJoeldomingo/blob/main/AB%20Testing%20Project%20-%20Company%20X/Code/Question%202%20Part%201.ipynb

Part 2:
https://github.com/jdomingo117/ProjectsJoeldomingo/blob/main/AB%20Testing%20Project%20-%20Company%20X/Code/Question%202%20Part%202.ipynb

Question 3 Notebook

https://github.com/jdomingo117/ProjectsJoeldomingo/blob/main/AB%20Testing%20Project%20-%20Company%20X/Code/Question%203.ipynb

Generated files

Includes:

- Exported DataFrames
- Excel files (csv)
- Screenshot of processes and formulas

https://github.com/jdomingo117/ProjectsJoeldomingo/tree/main/AB%20Testing%20Project%20-%20Company%20X/My%20Files

Provided files

https://github.com/jdomingo117/ProjectsJoeldomingo/tree/main/AB%20Testing%20Project%20-%20Company%20X/Provided%20files

Python Libraries Used

SciKit Learn (includes NearestNeighbors & KMeans Algorithms)
Buitinck, L. et al., 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*. pp. 108–122.

NumPy
Harris, C.R., Millman, K.J., van der Walt, S.J. et al. *Array programming with NumPy*. Nature 585, 357–362 (2020). DOI: 10.1038/s41586-020-2649-2. (Publisher link).

Pandas
McKinney, W. & others, 2010. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*. pp. 51–56.

MatPlotLib.pyplot
Hunter, J.D., 2007. Matplotlib: A 2D graphics environment. *Computing in science & engineering*, 9(3), pp.90–95.

Seaborn
Waskom, M. et al., 2017. *mwaskom/seaborn: v0.8.1 (September 2017)*, Zenodo. Available at: https://doi.org/10.5281/zenodo.883859.

Contact info

Email: joeldomingo117@gmail.com
LinkedIn: https://www.linkedin.com/in/joeldomingo117/
GitHub: github.com/jdomingo117/ProjectsJoeldomingo