Udacity Business Analyst Nanodegree Project: Analyzing a Market Test

Jason Grenig

Step 1: Plan Your Analysis

To perform the correct analysis, you will need to prepare a data set. (500 word limit)
Answer the following questions to help you plan out your analysis:

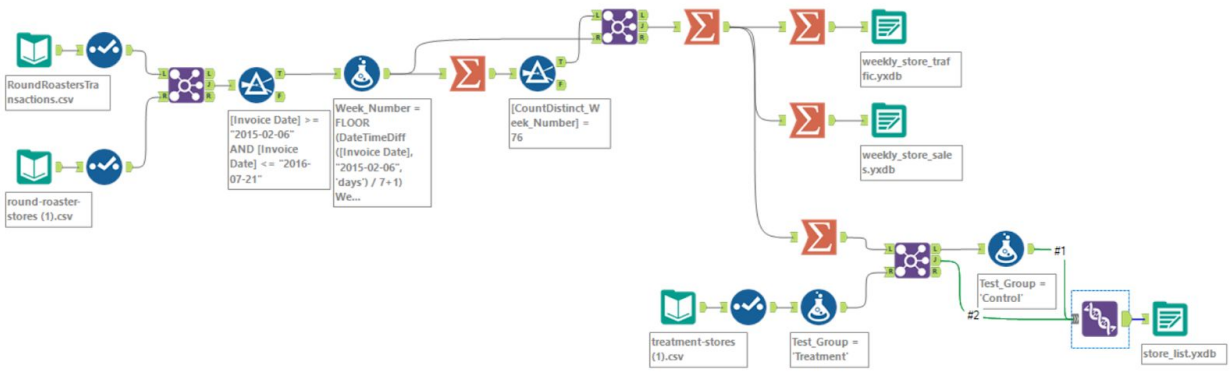
1. What is the performance metric you'll use to evaluate the results of your test?
   The performance metric used will be gross margin, as we are measuring the profitability of the television advertising in the 2 markets (Denver and Chicago).
2. What is the test period?
   The test period had a 12 week duration (2016-April-29 to 2016-July-21). Five stores in each of the test markets offered the updated menu along with television advertising.
3. At what level (day, week, month, etc.) should the data be aggregated?
   Data should be aggregated at the week level. We can expect customers of this coffee chain to make purchases on a weekly basis, so 1 cycle = 1 week.

Step 2: Clean Up Your Data

In this step, you should prepare the data for steps 3 and 4. You should aggregate the transaction data to the appropriate level and filter on the appropriate data ranges. You can assume that there is no missing, incomplete, duplicate, or dirty data. You're ready to move on to the next step when you have weekly transaction data for all stores.

In order to match treatment and control units, we first need to prep the data:

1. **Pull the historical data for the test period -** from the Transactions data, we'll filter by Invoice Date. The 12 week test period is from 2016-April-29 to 2016-July-21. Then we need 1 year + 12 weeks (76 weeks) of historical data prior to the test period, so we'll filter transactions from 2015-February-06 to 2016-July-21.
2. **Consolidate the data on a weekly basis -**
   a. Calc the Week_Number, Week_Start, Week_End dates.
   b. Filter w/ Count_Distinct on the Week_Number to exclude any stores that don't have 76 weeks of data.
   c. Group the data by store and by week to get the weekly store sales and weekly store traffic output .yxdb files. This will help us measure trend and seasonality.
   d. Bring in the Treatment Store data and combine with Control Store data, adding a Test_Group column to segregate stores for match pairing, and save as a store_list in an output .yxdb file.



Alteryx workflow of Data Cleaning

Step 3: Match Treatment and Control Units

In this step, you should create the trend and seasonality variables, and use them along with you other control variable(s) to match two control units to each treatment unit. Note: Calculate the number of transactions per store per week to calculate trend and seasonality.

Apart from trend and seasonality...

1. What control variables should be considered? Note: Only consider variables in the RoundRoastersStore file.
   Other control variables we can consider are:
   a. The City
   b. The State
   c. The Region
   d. The store's average monthly sales.
   e. The store's square footage.

2. What is the correlation between your each potential control variable and your performance metric?
   Correlation testing for the control variables:
   a. City, State, and Region Testing -
      i. By checking the distinct count of stores by City and by State, we see that there aren't enough stores in either to do the control-treatment match pairing.
      ii. However, if we look at distinct count of stores by region, there are 42 stores in the Central region and 91 stores in the West region. This is enough for the pairing, so we can use region as a control variable for our test.

Record #	Region	City	State	CountDistinct_StoreID
1	Central	Altoona	IA	1
2	Central	Amarillo	TX	1
3	Central	Anoka	MN	1
4	Central	Arlington	TX	1
5	Central	Arlington Heights	IL	1
6	Central	Barrington	IL	1
7	Central	Blaine	MN	1
8	Central	Brookfield	WI	1
9	Central	Chesterfield	MO	1
10	Central	Chicago	IL	3
11	Central	Cottage Grove	MN	1
12	Central	Dallas	TX	2
13	Central	Des Moines	IA	1
14	Central	Elk Grove Village	IL	1
15	Central	Fort Worth	TX	1
16	Central	Greendale	WI	1
17	Central	Houston	TX	3
18	Central	Lawrence	KS	1
19	Central	Lockport	IL	1
20	Central	Lucas	TX	1
21	Central	Madison	WI	1
22	Central	Melrose Park	IL	1
23	Central	Mount Prospect	IL	1
24	Central	Nixa	MO	1
25	Central	Northbrook	IL	1

Control Variable Test for Matching by City

Record #	Region	State	CountDistinct_StoreID
1	Central	IA	2
2	Central	IL	15
3	Central	KS	2
4	Central	MN	4
5	Central	MO	4
6	Central	TX	12
7	Central	WI	3
8	West	AZ	10
9	West	CA	42
10	West	CO	15
11	West	ID	2
12	West	NM	2
13	West	NV	2
14	West	OR	8
15	West	WA	10

Control Variable Test for Matching by State

Record #	Region	CountDistinct_StoreID
1	Central	42
2	West	91

Control Variable Test for Matching by Region

b. Average Monthly Sales and Square Footage Testing -
   i. To test the Average Monthly Sales and Square Footage, we can join the Transaction and Store files, then run an Association Analysis.
   ii. The Average Monthly Sales is highly correlated to the Gross Margin, at 0.988 and a p-value of 0.0 - well within the 0.05 significance threshold.
   iii. The Square Footage (Sq_Ft) has no correlation at -0.02 and a p-value of 0.8.

iv. Therefore we should only keep the Average Monthly Sales as a control variable.

Pearson Correlation Analysis

Focused Analysis on Field Sum_Gross.Margin

	Association Measure	p-value	
AvgMonthSales	0.988219	0.00000	***
Sq_Ft	-0.020353	0.81612	

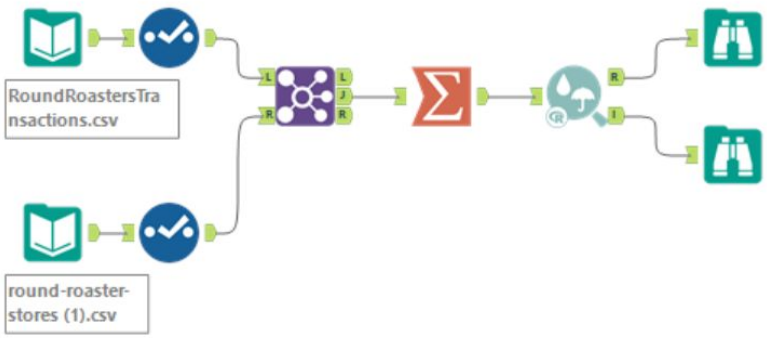
Full Correlation Matrix

	Sum_Gross.Margin	Sq_Ft	AvgMonthSales
Sum_Gross.Margin	1.000000	-0.020353	0.988219
Sq_Ft	-0.020353	1.000000	-0.046967
AvgMonthSales	0.988219	-0.046967	1.000000

Matrix of Corresponding p-values

	Sum_Gross.Margin	Sq_Ft	AvgMonthSales
Sum_Gross.Margin		0.81612	0.00000
Sq_Ft	0.81612		0.59138
AvgMonthSales	0.00000	0.59138	

Association Analysis Results



Alteryx Association Analysis Workflow

3. What control variables will you use to match treatment and control stores?
Based on the above control variable testing, we can use average monthly sales and region as the discrete control variables. Then for the continuous control variables, we can use the weekly number of invoices per store and weekly total gross margin per store (our measures of trend and seasonality).

4. Please fill out the table below with your treatment and control stores pairs:

Treatment Store	Control Store 1	Control Store 2
1664	7162	8112
1675	1580	1807
1696	1964	1863
1700	2014	1630

1712	8162	7434
2288	9081	2568
2293	12219	9524
2301	3102	9238
2322	2409	3235
2341	12536	2383

Record #	Controls	Treatments	Distance	Region	AvgMonthSales	Test_Group
1	7162	1664	0.478595	Central	11000	Treatment
2	8112	1664	1.034443	Central	11000	Treatment
3	1580	1675	0.45634	Central	15000	Treatment
4	1807	1675	0.560454	Central	15000	Treatment
5	1964	1696	0.312367	Central	10000	Treatment
6	1863	1696	0.489137	Central	10000	Treatment
7	2014	1700	0.810402	Central	15000	Treatment
8	1630	1700	0.91618	Central	15000	Treatment
9	8162	1712	0.671441	Central	19000	Treatment
10	7434	1712	0.793269	Central	19000	Treatment
11	9081	2288	0.277932	West	14000	Treatment
12	2568	2288	0.714134	West	14000	Treatment
13	12219	2293	0.348583	West	11000	Treatment
14	9524	2293	0.656038	West	11000	Treatment
15	3102	2301	0.381248	West	11000	Treatment
16	9238	2301	0.434646	West	11000	Treatment
17	2409	2322	0.171431	West	14000	Treatment
18	3235	2322	0.45125	West	14000	Treatment
19	12536	2341	0.39796	West	11000	Treatment
20	2383	2341	0.423792	West	11000	Treatment

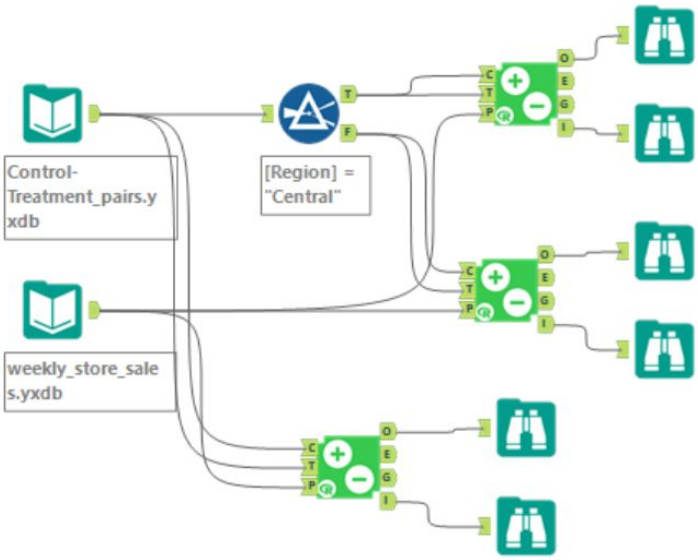
Complete Control-Treatment Matched Pairing

Step 4: Analysis and Writeup

Conduct your A/B analysis and create a short report outlining your results and recommendations. (250 words limit)

Answer these questions. Be sure to include visualizations from your analysis:

1. What is your recommendation - Should the company roll out the updated menu to all stores?
   After reviewing the AB Analysis results, the company should roll out the updated menu with gourmet sandwiches and wine offerings to the rest of the stores. The overall lift is 40.7%, which is higher than the 18% lift threshold needed to roll out the changes to all stores. Looking at the Time Comparison Plots, we can also see that gross margin was similar historically, but the treatment groups in both regions significantly outperformed the control groups during the test.

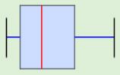


Alteryx AB Analysis Workflow

2. What is the lift from the new menu for West and Central regions (include statistical significance)?
    a. The West region got a lift of 37.9%, with a 99.5% confidence level.
    b. The Central region got a lift of 43.5%, with a 99.6% confidence level.

AB Test Analysis for Sum_Gross Margin

Time: 2018-12-25 19:41:34



37.9

Average Lift, %

526.5

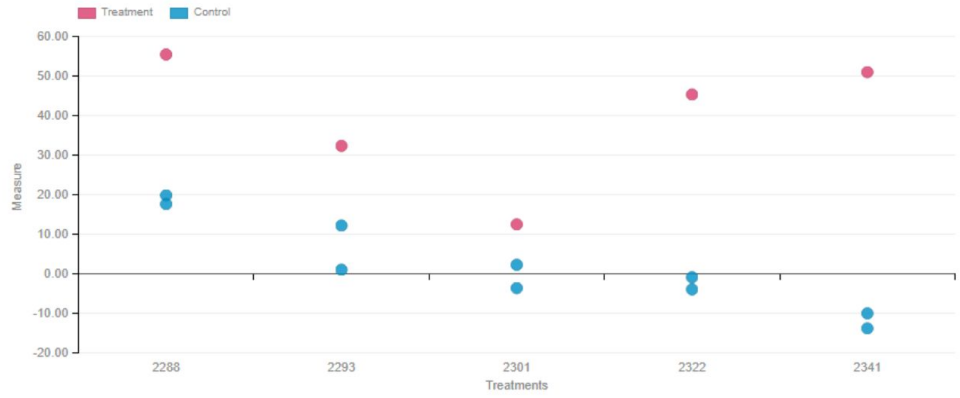
Expected Impact, Sum_Gross Margin

99.5

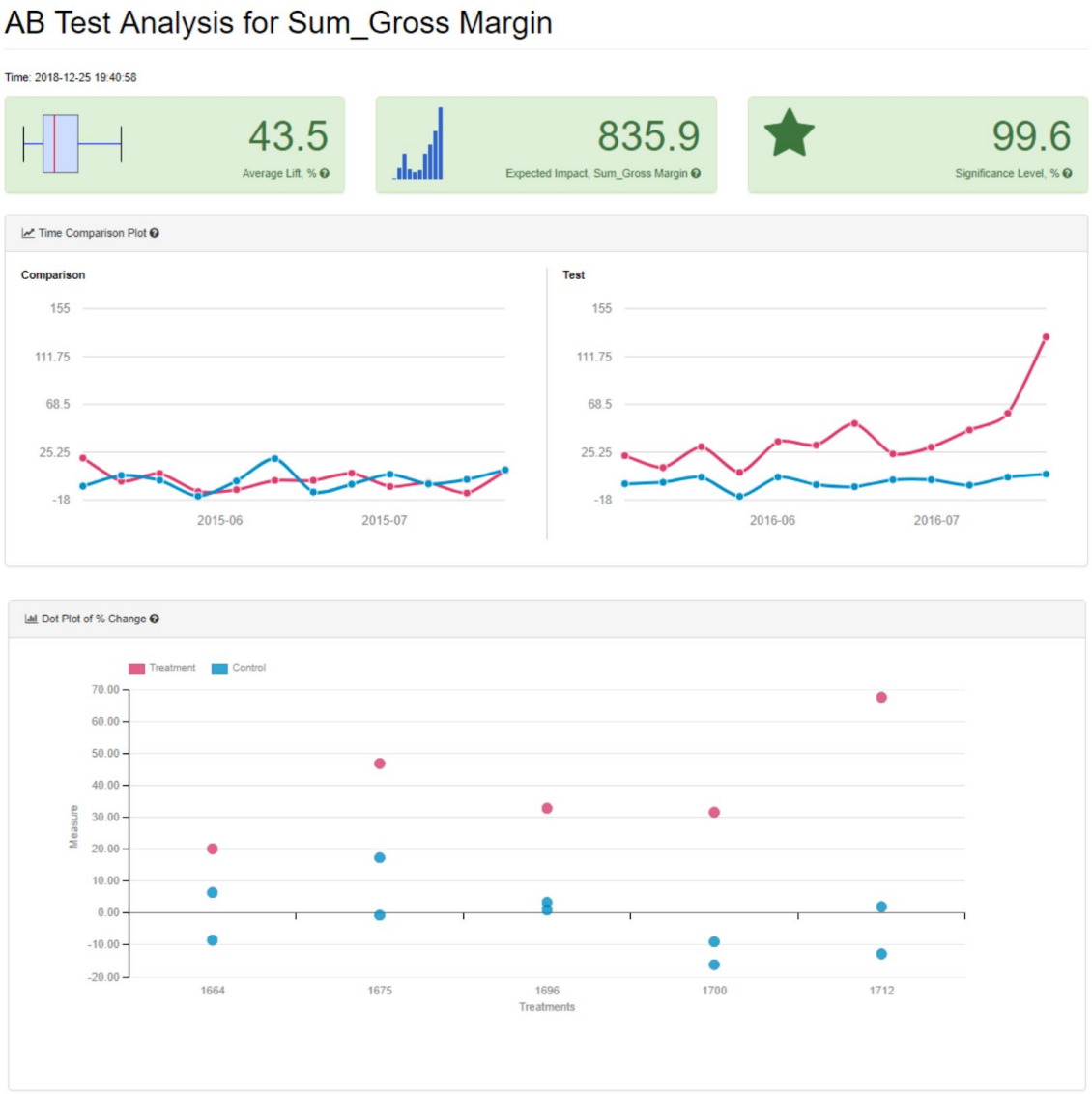
Significance Level, %

Time Comparison Plot

Dot Plot of % Change



West Region AB Test Analysis Results

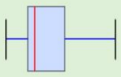


Central Region AB Test Analysis Results

3. What is the lift from the new menu overall?
   Overall, the new menu results in a 40.7% lift, with a 100% confidence level.

AB Test Analysis for Sum_Gross Margin

Time: 2018-12-25 19:42:08



40.7

Average Lift, %

681.2

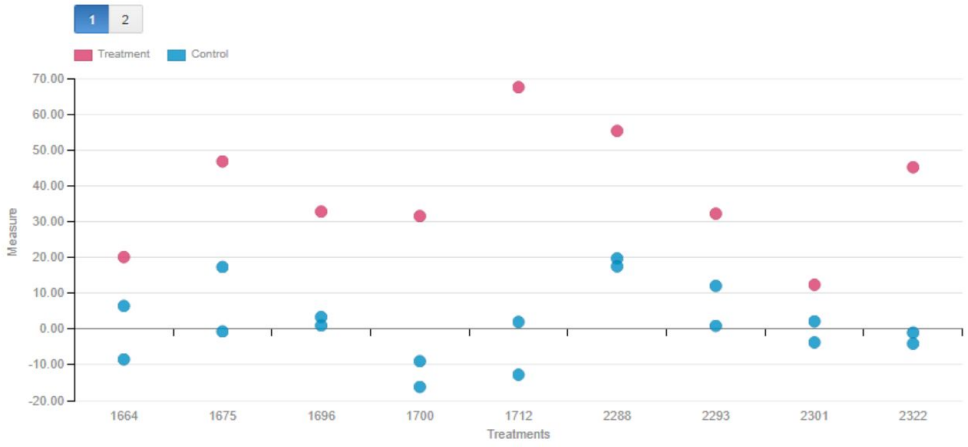
Expected Impact, Sum_Gross Margin

100

Significance Level, %

Time Comparison Plot

Dot Plot of % Change



Overall AB Test Analysis Results