Identify significant promotion & marketing vehicles

MKT_680_4100:

Marketing Models

Group 5

Yi Cai

Sameera Vattikuti

Nicholas Yang

Zeyu Zhu

Table of Contents

Chapter 1: Introduction

1.1 Business Background

Pernalonga, a leading supermarket chain with 421 stores which sells ~10K products in 430 categories with a consumer base of ~7900. Currently, 30% of their sales are through promotions (mostly in-store) executed in partnership with suppliers.

Beer is a frequently promoted category at Pernalonga. Mahou San Miguel, a Spanish brewer, sells three San Miguel beer products in Pernalonga's stores and regularly partners with Pernalonga to promote its products via weekly flyers and in-store displays. Mahou San Miguel also employs other marketing vehicles such as email, web (display banners and paid search), and traditional media (TV and Radio). They are interested in identifying promotion and marketing activities that drive significant incremental sales for continuation into 2018.

1.2 Problem Statement

The business problem could be translated to a data problem statement as 'Build models to identify contributions of marketing & promotional operations to sales & suggest best marketing/promotional activities for each brand.'

Chapter 2: Overview

2.1 Data Overview

For the purpose of our analysis, we look at the subset of transactional data related 3 products of Mahou. It consists of transactional data for ~8880 transactions at 410 stores related to 5084 unique customers. We have details like duration, vehicle of promotion, amount & unit of measurement of promotion and advertising activities. In addition to this, we also have data related to seasonality and a list of holidays for 2016 – 2017 that could have an impact on sale of beers.

2.1.1 Assumptions & Considerations

- The analysis is based on **weekly data**
- We create a variable **tran_wk & weeknum** in order to refer to transactions at a weekly level.
- The holiday data is binary coded.
- The types of promotions are dummy coded with metric assigned to each promotion & week.
- The data from all files are combined to result in a dataset of 30 columns & ~8880 rows
- Since we are interested in overall sales of a product at a weekly level, the lack of a proper unique transaction id does not matter in this context & therefore not addressed in the data
- Since all the products are of the same SKU – CT, model does not need to control for SKU

2.2 Business Overview

2.2.1 Product Overview - 138936951

For the given period, this product registers sale at 406 outlets. ~17000 units of this product were sold. It features in ~4500 transactions & sells ~3.7 units per transaction. At the beginning of the given period, the shelf price of the product was .69. It was increased to .75 and later to .79. Overall, this is the cheapest product from Mahou's portfolio. Generally, it was sold at a 4.5% discount.
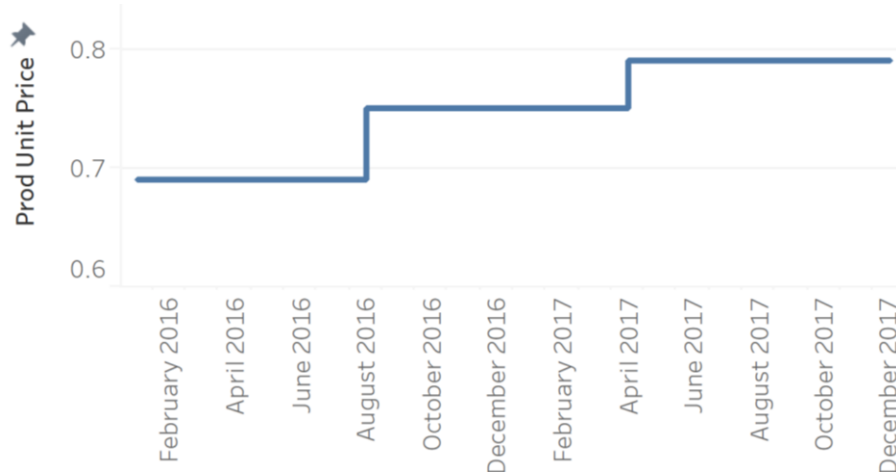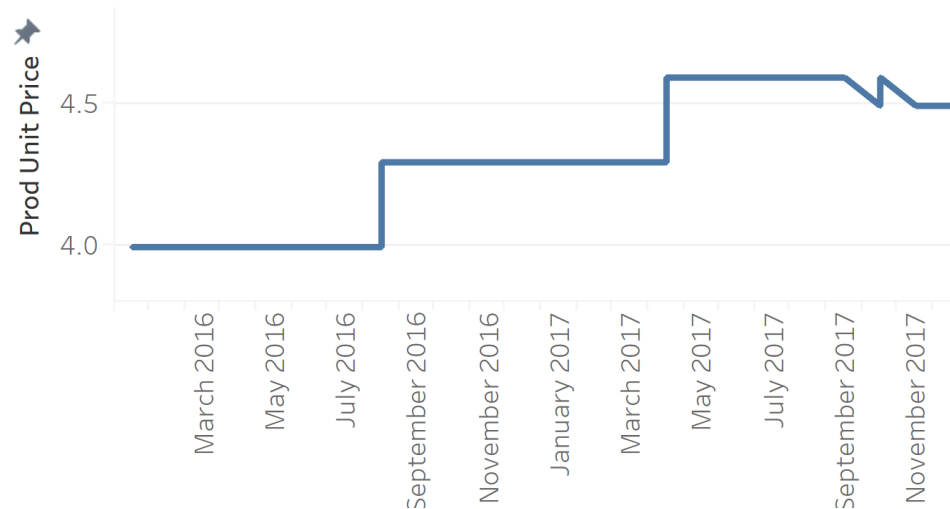
2.2.2 Product Overview - 138936952

For the given period, this product registers sale at 400 outlets. ~11100 units of this product were sold. It features in ~3720 transactions & sells ~3 units per transaction. At the beginning of the given period, the shelf price of the product was 3.99. It was increased to 4.29 and later to 4.59. Finally, it was reduced slightly to 4.49 Generally, it was sold at a 4.9% discount.

2.2.3 Product Overview – 138936953

For the given period, this product registers sale at only 285 of the 410 outlets. ~1276 units of this product were sold. It features in ~635 transactions & sells ~2 units per transaction. At the beginning of the given period, the shelf price of the product was 14.99. It was increased to 15.99, then to 16.99 & finally reduced to 16.49. Overall, it is the most expensive product of Mahou. Generally, it was sold at a 7.8% discount.
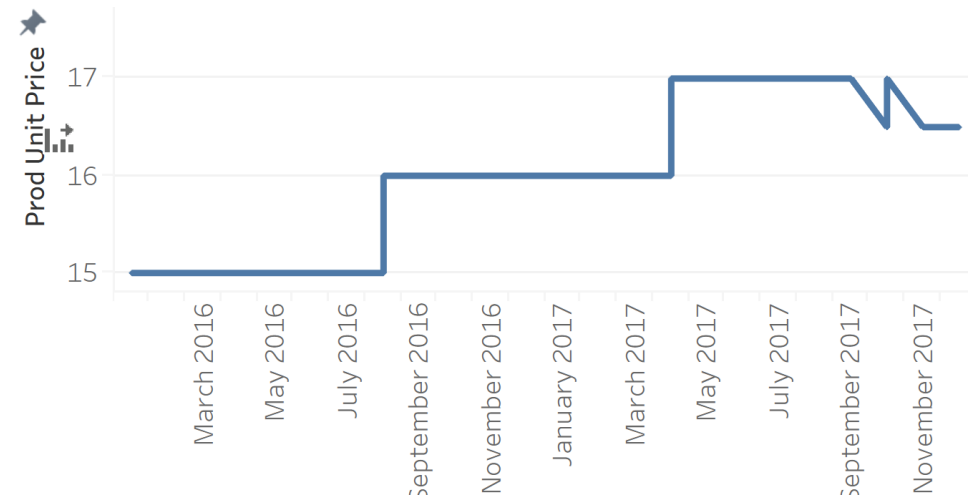
2.2.4 Assumptions & Considerations

Seasonality

The given data for indexed seasonality of sales highlights both seasonality & cyclicity of the beer industry sales. Can be seen that sales usually peak around July – possibly with the arrival of hot weather & hit lows at around December/January with extreme winters.

relationship between marketing efforts & sales. The effect of promotions is rarely linear. They usually need to reach a threshold for them to be effective. Also, the effect tapers off or flattens out after reaching a certain point. Multiplicative & logit forms are better at explaining the promotion-sales relationship.

Multiplicative Model

There is a simple implicit interaction assumed between the causal variables which is true to reality where there are interactions among different types of promotions. The dependent variable here is unbound. However, sales have limits & is a bound quantity. This is not as easy to interpret & decompose as additive models.

Logit Model

Complex interactions are assumed among the variables which is possible in real promotions. The dependent variable is bound & in this case we take a limit of 10% of the maximum volume as max limit. This model is definitely not as interpretive & needs heuristics to calculate the sale value but generally perform better at modeling the relation among the variables here.

Standardizations

All the variables are normalized. The dependent variable is normalized by distribution. Price variables are normalized about the period mean. Percentage of discount is considered & previously stated metric are taken for appropriate promotion vehicles.

Chapter 3: Models & Evaluation

Multiplicative models were built for all products. We look at r-square, adjusted r-square, MAPE, F_Stat, dfR, dfE & Dubson Watson statistic. While we mainly use MAPE to make the decision. In general, logit

Unit of time (for analysis)

We take week as a unit of time for analysis as most of the data is available at a weekly level. This gives us more detailed idea of trends over the 2-year period. Weekly data also flattens the weekday vs weekend sales effect (which would be especially high for the beer category) while giving a better picture of yearly trends in sales & demand.

Promotion measurement

For flyers & store display, their presence was coded as 1 for relevant products & weeks. For paid search & web display, impressions were used as metric & circulation for email. For Radio & television, the given GRPs were converted to adstock which in turn were used to calculate reach.

2.3 Modeling Overview

For modeling, we have the option of using additive, multiplicative or logit models. Linear models rarely do a good job of explaining the

models perform better on all evaluation metrics than the multiplicative model. Especially the Dubson Watson statistic shows that there is lesser autocorrelation of errors for the logit models.

	Model	rsquare	adj_rsquare	MAPE	F_stat	dfR	dfE	DurbinWatson_stat
1	lm_prod1	0.1446	0.05365	0.2666	1.590	10	94	1.269
2	lm2_prod1	0.4959	0.44166	0.1166	9.147	10	93	2.258
3	lm_prod2	0.4200	0.35137	0.2112	6.122	11	93	1.452
4	lm2_prod2	0.6903	0.65331	0.1366	18.645	11	92	1.919
5	lm_prod3	0.7146	0.67938	0.5043	20.263	11	89	2.227
6	lm2_prod3	0.7958	0.77028	0.4710	31.178	11	88	2.358

The VIF & t-statistic look at the value of each independent variable in the model. Following are the values for products:

Product: 138936951

	VIF_lm	VIF_lm2	Pr(>|t|)_lm	Pr(>|t|)_lm2
discount	3.888	3.891	0.9897	0.015756
shelf_price	1.760	1.732	0.6921	0.002175
seasonality	2.307	2.288	0.1663	0.276822
holiday	2.437	2.393	0.7028	0.837140
flyer	2.392	2.433	0.2658	0.108042
paid_search	1.131	1.131	0.6640	0.241767
radio	6.510	6.437	0.7442	0.033230
tv	8.277	8.089	0.4651	0.003128
webdisplay	1.314	1.313	0.6870	0.835907
email	1.366	1.383	0.1641	0.150925

Product: 138936952

	VIF_lm	VIF_lm2	Pr(>|t|)_lm	Pr(>|t|)_lm2
discount	4.031	3.967	0.00479547	0.000006369641
shelf_price	1.746	1.726	0.02840996	0.000004839789
seasonality	2.400	2.379	0.00000254	0.000000001346
holiday	2.259	2.228	0.84697322	0.429110383610
store_display	1.361	1.365	0.98021825	0.875269738910
flyer	2.429	2.399	0.19362402	0.651648175119
paid_search	1.106	1.105	0.27589666	0.053945100510
radio	6.431	6.361	0.44698512	0.542443461580
tv	8.298	8.110	0.53769039	0.681103630087
webdisplay	1.345	1.344	0.49130129	0.197866773219
email	1.330	1.346	0.06919772	0.114606664654

Product: 138936953

	VIF_lm	VIF_lm2	Pr(>|t|)_lm	Pr(>|t|)_lm2
discount	9.450	9.340	0.000184475916126	0.0000054873625534
shelf_price	1.738	1.710	0.000000019940868	0.0000000000034931
seasonality	2.387	2.363	0.000000000004633	0.0000000000008445
holiday	2.363	2.315	0.561134669971038	0.5283414240383802
store_display	5.021	5.167	0.074833523383383	0.0080640509349726
flyer	3.232	3.211	0.952937540158413	0.3335691409770082
paid_search	1.142	1.142	0.352088163313476	0.2304987161999729
radio	6.754	6.671	0.395152437425925	0.9954151522634486
tv	8.545	8.351	0.495826003478104	0.6871829183778713
webdisplay	1.364	1.364	0.666802006628787	0.4289438726911361
email	1.417	1.447	0.762855464076750	0.6318899961602763

The VIF of radio & TV indicate that they have a high level of collinearity with the other variables. From a marketing POV, these are the most expensive & large-scale promotions when compared to the rest. Therefore, they are usually done in combination with lot of the other ways of promotions. So the extent & effects of these will also be similar to other promotions.

Product: 138936951

```
Call:
lm(formula = log(sale_amount/(330 - sale_amount)) ~ discount +
    self_price + seas_index + as.factor(holi_index) + as.factor(flyer) +
    paid_search + radio + tv + web_display + email, data = transaction_product1[-1,
    ])

Residuals:
    Min      1Q  Median      3Q     Max
-1.5085 -0.3326 -0.0447  0.2622  2.2662

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)               4.2719     1.2278    3.48  0.00077 ***
discount                  3.5636     1.4488    2.46  0.01576 *
self_price               -5.4117     1.7163   -3.15  0.00218 **
seas_index                0.0894     0.0818    1.09  0.27682
as.factor(holi_index)1   -0.0368     0.1785   -0.21  0.83714
as.factor(flyer)1         0.6354     0.3916    1.62  0.10804
paid_search               0.0674     0.0572    1.18  0.24177
radio                    -1.4788     0.6842   -2.16  0.03323 *
tv                        2.9690     0.9786    3.03  0.00313 **
web_display               0.0128     0.0616    0.21  0.83591
email                     0.0915     0.0632    1.45  0.15092
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.548 on 93 degrees of freedom
Multiple R-squared:  0.496,     Adjusted R-squared:  0.442
F-statistic: 9.15 on 10 and 93 DF,  p-value: 0.000000000233
```

The intercept is the most significant feature as expected, since base forms a large part of the sales volume. Shelf price is the next most significant variable which has a negative effect on the volume. Discount has a positive effect on sales while interestingly, radio seems to negative coefficient meaning increase in radio reach impacts the log odds of sales by -1.48.

Product: 138936952

```
Call:
lm(formula = log(sale_amount/(227 - sale_amount)) ~ discount +
    self_price + seas_index + as.factor(holi_index) + as.factor(store_display) +
    as.factor(flyer) + paid_search + radio + tv + web_display +
    email, data = transaction_product2[-1, ])

Residuals:
    Min      1Q  Median      3Q     Max
-0.8982 -0.1806 -0.0389  0.1885  1.0405

Coefficients:
                          Estimate Std. Error t value    Pr(>|t|)
(Intercept)                 3.9217     0.8402    4.67 0.0000103268 ***
discount                    5.2195     1.0897    4.79 0.0000063696 ***
self_price                 -0.9883     0.2034   -4.86 0.0000048398 ***
seas_index                  0.3822     0.0567    6.74 0.0000000013 ***
as.factor(holi_index)1      0.0930     0.1171    0.79        0.429
as.factor(store_display)1   0.0314     0.1994    0.16        0.875
as.factor(flyer)1          -0.0627     0.1385   -0.45        0.652
paid_search                 0.0750     0.0384    1.95        0.054 .
radio                       0.2827     0.4624    0.61        0.542
tv                         -0.2747     0.6662   -0.41        0.681
web_display                -0.0549     0.0424   -1.30        0.198
email                       0.0675     0.0424    1.59        0.115
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.372 on 92 degrees of freedom
Multiple R-squared:  0.69,      Adjusted R-squared:  0.653
F-statistic: 18.6 on 11 and 92 DF,  p-value: <0.0000000000000002
```

Product: 138936953

```
Call:
lm(formula = log(sale_amount/(83 - sale_amount)) ~ discount +
    self_price + seas_index + as.factor(holi_index) + as.factor(store_display) +
    as.factor(flyer) + paid_search + radio + tv + web_display +
    email, data = transaction_product3[-1, ])

Residuals:
    Min      1Q  Median      3Q     Max
-1.5312 -0.4403  0.0643  0.4440  1.1324

Coefficients:
                          Estimate Std. Error t value     Pr(>|t|)
(Intercept)               10.56978    1.57648    6.70 0.00000000185988 ***
discount                  11.13382    2.30008    4.84 0.00000548736255 ***
self_price                -0.82402    0.10223   -8.06 0.00000000000349 ***
seas_index                 0.78438    0.09381    8.36 0.00000000000284 ***
as.factor(holi_index)1    -0.12513    0.19766   -0.63           0.5283
as.factor(store_display)1 -1.57269    0.58011   -2.71           0.0081 **
as.factor(flyer)1          0.31399    0.32293    0.97           0.3336
paid_search                0.07826    0.06481    1.21           0.2305
radio                      0.00448    0.77675    0.01           0.9954
tv                         0.44657    1.10533    0.40           0.6872
web_display                0.05626    0.07080    0.79           0.4289
email                     -0.03506    0.07292   -0.48           0.6319
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.606 on 88 degrees of freedom
Multiple R-squared:  0.796,     Adjusted R-squared:  0.77
F-statistic: 31.2 on 11 and 88 DF,  p-value: <0.0000000000000002
```
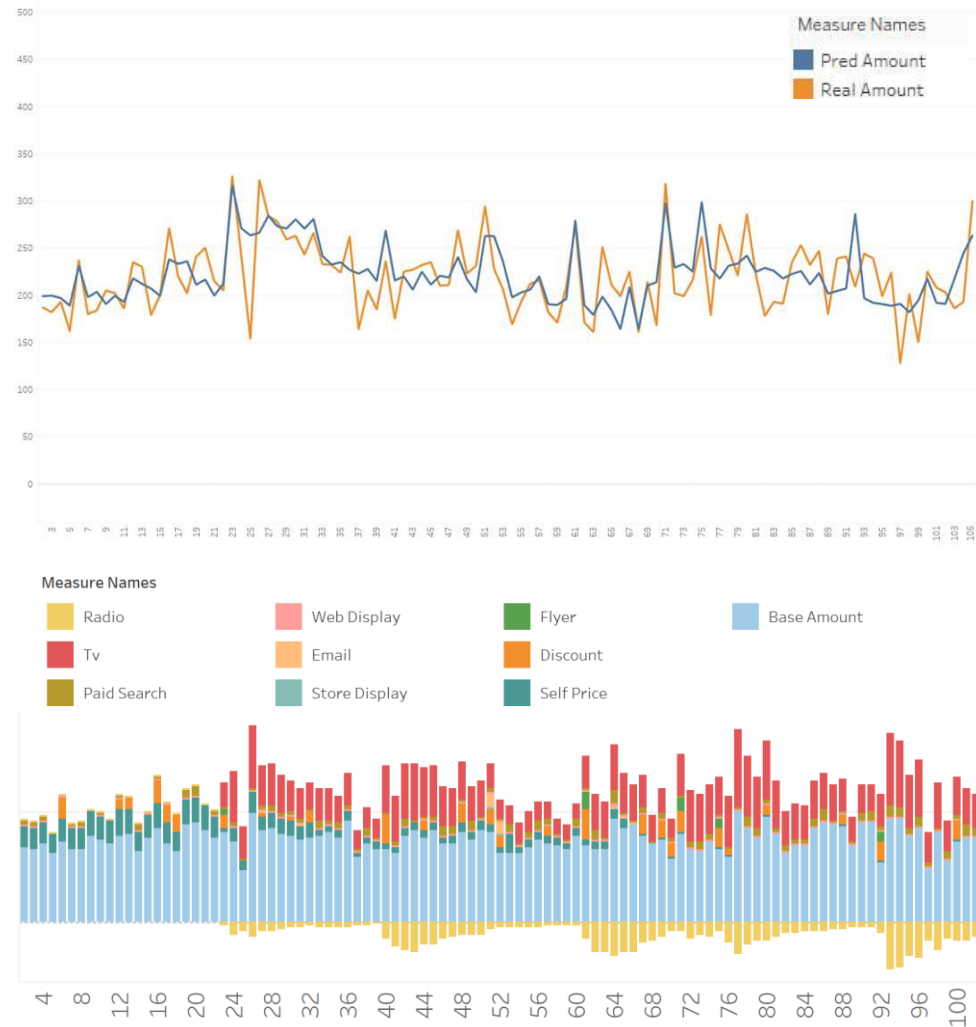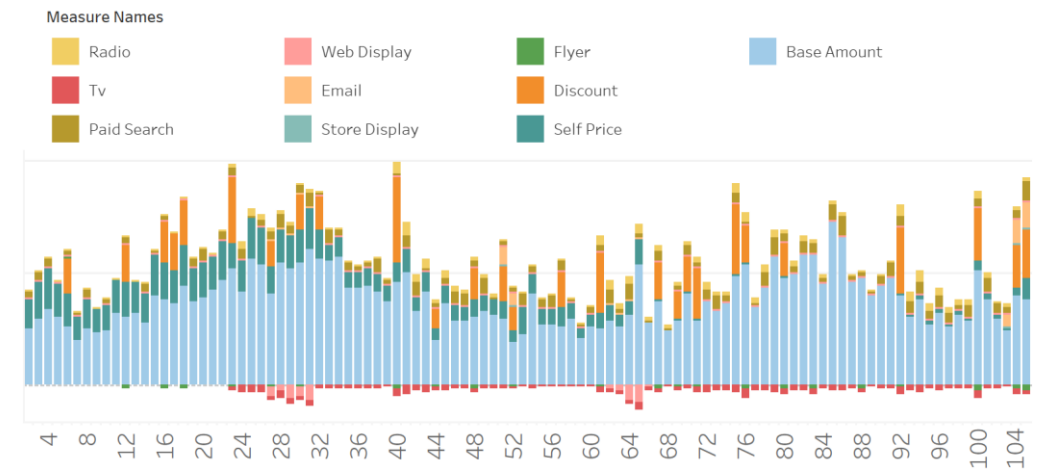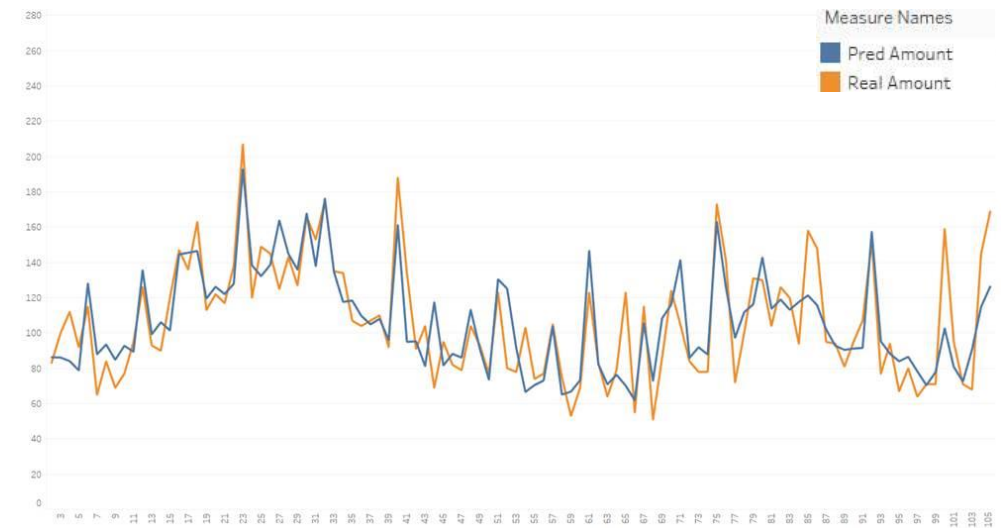
3.1 Decomposition

Once we have the model for all products, logit is selected as the best model for all products. Following is a representation of predicted vs actual sales & the decomposition for each product.
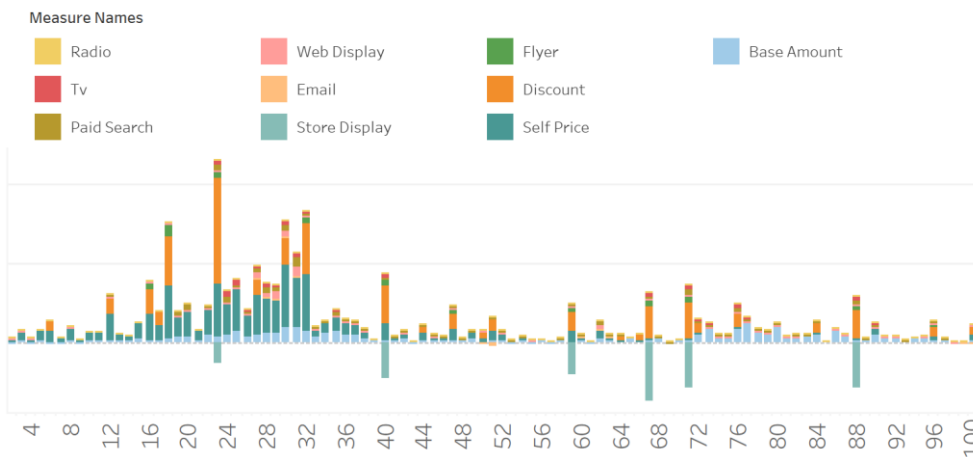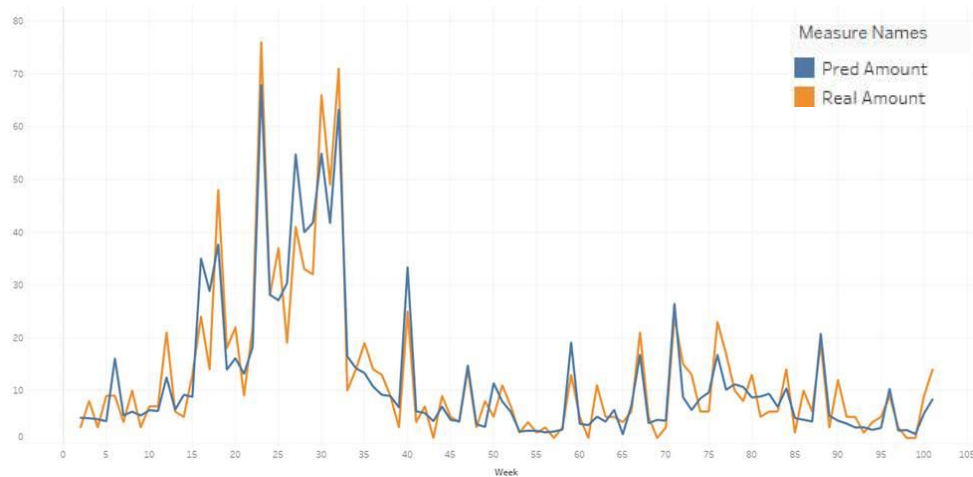
Product: 138936951

Product: 138936952

Product: 138936953

The base for all products includes seasonality & holiday index. The 'DuTo' of each variable is calculated by keeping the values of other variables to zero or minimum. Then the data is scaled back to reflect the actual volumes as closely as possible.

Chapter 4: Conclusion

Product: 138936951

Television is the most effective mode of promotion when it comes to impacting sales.

Base forms a large portion of the volume & TV shows the biggest lift in sales, followed by discount. Shelf price has an impact on the volumes as well. Paid search & discount give consistent but small improvements in volume whenever used.

Radio has a negative impact. Probably because this coincides heavily with their increments in pricing.

Product: 138936952

Specifically for promotions, paid search has the most positive impact on sales, followed by radio.

Once again, the product has a large base. Discount has the most impact on volumes. Shelf price has a consistent effect on the volumes as well.

TV, web display & flyers have slight negative impact, probably due to the fact that these are rolled out at weeks where price adjustments have been done.

Product: 138936953

The base forms a very small part of the volume. Since this is an expensive product, shelf price highly impacts the volume. Discount has big impact as well. In terms of marketing promotions, flyers, tv, web display have consistent positive impact. Shelf display has a negative impact & this could once again be because the displays coincide with price increases.