Understanding the Drivers of San Miguel's Sales through Marketing Mix Modeling

MKT 680 - Pricing Group Project

1. Introduction

Pernalonga is a large supermarket chain in Lunitunia, operating over 400 stores, selling 10,000 products in over 400 categories. Beer is one of the major categories that they offer to customers, as it makes up over 2% of their total sales. Mahou San Miguel, a leading beer brand in Spain, has been trying to increase its sales in that country, but they lag far behind the market leader Super Bock. So, the brand is interested to know whether or not it can increase the sales of San Miguel branded products by utilizing the correct marketing mix models. By exploring Pernalonga's transaction level data, we were able to solve this problem and come up with a marketing mix model for Mahou San Miguel using explanatory modeling techniques that will identify their brands best promotions and marketing vehicles to continue for next year, helping boost their revenue and drive significant results.

2. Business Understanding

Mahou San Miguel is a Spanish brewing company, founded in Madrid in 1890. It is the leading brand in the Spanish beer market. Its products include:

- 1890 Mahou Clásica
- 1957 San Miguel Especial
- 1969 Mahou Cinco Estrellas
- 1990 Laiker (Mahou's alcohol-free beer) Renamed as Mahou Sin
- 2001 San Miguel 0,0 (non-alcoholic beer)
- 2003 Mahou Negra, San Miguel 1516, and San Miguel ECO
- 2005 Mixta, Mahou shandy, and San Miguel 0,0% Manzana (apple-flavoured)
- 2007 San Miguel 0,0%, con Té sabor Limón (with tea lemon flavor)
- 2007 Alhambra Especial, Alhambra Reserva 1925, Mezquita, Alhambra Premium Lager, Alhambra Negra, and Alhambra Sin
- Selecta XV

However, San Miguel has not been able to replicate that success in the Lunitunia market. It has ~0.2% of the market and is significantly lagging behind the market leader Super Bock (which has 44% of the market share). Although it has been providing discounts (up to 30%), its has not been able to penetrate the market enough. Hence, there is a huge scope of using a more appropriate marketing mix model to optimize the campaigns and ensure higher sales.

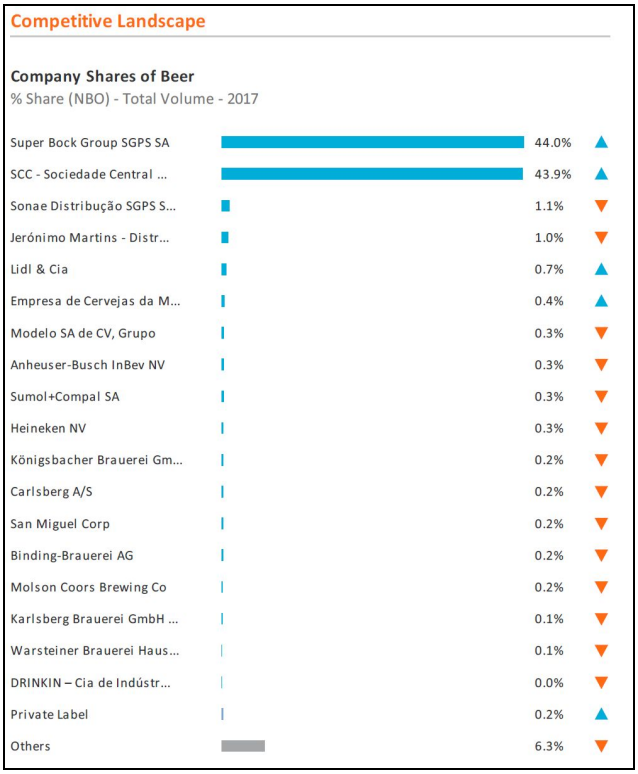


Figure 1: The ranking of San Miguel Corp in company shares of beer in Lunitunia market

One area where Mahou San Miguel can improve is their marketing promotions. Right now, they are unsure of which promotions are the key drivers of their revenue. Thus, there is ample opportunity to improve this part of their business and help drive key growth in a new market.

3. Data Understanding & Data Preparation

For this project, we have the transactional data of three different San Miguel beer products from 2016 up to 2018. The overview of each product's transactions is shown below, and these three products may be sold as a single can, a 6-pack, and a case.

Product ID	Median Unit Price ($)	Total Quantity	Total Sales ($)
138936951	0.75	22834	17029.66
138936952	4.29	11120	47637.90
138936953	14.99	1276	19957.74

Table 1: Transaction data of three targeted beer products

The important national holidays, seasonality indices, and marketing vehicles used on San Miguel products are provided at a weekly level. Among all 7 marketing vehicles, TV, radio, paid

search, web display, and emails are implemented to advertise all three kinds of products universally. Flyers are implemented for each product separately in different weeks, and store displays are only used on products that are sold in 6-pack and in a case. A paid search and web displays are measured in impressions, and TV and radio advertisements are measured in unit of GRP, or Gross Ratings Points.

Three marketing mix models are going to be built - one for each product. It is important to note that the last week in 2015 and the last week in 2017, which are the New Year weeks, are removed from the transaction table as those weeks are only partially recorded (most of the week took place in either 2015 or 2018, of which we do not have data). Keeping these data points would skew our models, since our marketing mix models are constructed at a weekly level with the dependent variable as the total quantity sold. For example, the first weekend had only 11 sales, whereas next year's New Year's week had over 10x that amount. The independent variables used in our model are discussed in details below.

Price:

We took the weighted average shelf price in that week.

Discount:

We used the average discount rate, which is calculated based on the the total discount amount over the the total sales amount

Seasonality Index:

It is simply merged from the seasonality index table given.

Holiday:

Holidays are transformed into dummy variables to encode them as 0/1's because various holidays may have a different impact on San Miguel's sales. We used our domain knowledge to remove some of the holidays, such as religious holidays and weeks before the actual holidays, that are given in the list because we believe that people normally do not organize big celebrations or gatherings that could potentially have an impact on beer sales. These discarded holidays are:

POPEVISIT, *ALLSAINTS*, *IMMACULATE*, *EASTER*, *CORPUS*, *RESTORATION*, *PrLIBERTY*, *PreEASTER*, *PrASSUMPTION*, and *PrXMAS*.

TV & Radio Reach:

TV reach and Radio reach are aggregate time series causal values in the model. Since we are given the half-life values of TV advertisements and Radio advertisements, as well as the Reach formula based on GRP, we will transform the given GRP values into TV and radio reach for each product over every week.

Paid search, Web display, Email, Store display, Flyer:

These marketing vehicle values in each week can be directly appended as they would only have impact on the sale quantities of products within a week.

Substitute Index:

Because San Miguel is such a small company in this country, it is heavily affected by its substitutes, or other larger beer companies. Thus, to account for this, we implemented a variable that represents the average price of its substitutes. The substitute index we chose is the weighted average shelf price of all substitutes in that week. We defined substitute as as items within the same category of the product of interest that are rarely bought together and have negative effect on each other. Additionally, we only looked at those products that were comparable in price to the San Miguel products. We used association rules to find the substitutes for each San Miguel product using the package "arules" in R, which utilized the APRIORI algorithm. To get the substitutes of a product, we applied APRIORI function with support and confidence at 0, so it can generate all of the co-purchased items with in the same category with the lift. Lift gives the correlation between item A and item B in the rule A=>B (how purchasing A affects purchasing B).If the lift is smaller than 1, the presence of A will have negative effect on B. In this way, we listed the substitutes for each product of interest by looking at the pairs of products generated with APRIORI function that have lift value smaller than 1. Once we have the substitutes, we can calculate the weighted average shelf price of the substitutes as our substitute index.

We also plotted the correlation matrix between the numeric variables in the models of three products individually, which are displayed in the appendix. Although some of the variables show correlations (e.g. tv_reach and radio_reach), since they are marketing vehicles that we want to investigate on, we decide to include them in our model. Meanwhile, after removing some of the holidays as mentioned above, the multicollinearity issues among holiday factors were minimized.

4. Building the models

As our goal in this project was to identify the key promotion and marketing activities, we first needed to come up with a model that allowed us to see the effects of each campaign with respect to sales units. To accurately model this relationship for each of three San Miguel products, we implemented a multiplicative model for each product to best fit the response function of all sales vehicles and baseline variables because the independent variables may interact with each other. The base form for this model is found below:

$$\log(y_t) = \sum_i \beta_i g_i(x_{it}) + \epsilon_t$$

Our dependent variable was the natural log transformation of sales units, and our independent variables included all marketing campaigns, including TV and Radio Reach and paid search impressions, as well as factors important to our baseline, such as a seasonality index and markers for important holidays. Additionally, we included weekly statistics about the products, consisting of the weighted average unit price and the average discount percentage, and the average unit

price of their substitute goods to account for that aspect as well. As previously mentioned, we used our domain knowledge to remove certain holidays from the model, as they had no bearing on San Miguel beer sales.

5. Model Diagnostics and Results

All the models were significant at p-values < 0.05 level. Upon running our regression model, the results revealed that the following variables as statistically significant by product.

Product ID	Positive Impact	Negative Impact
138936951	Discount percent, seasonality index, paid search	Average shelf price, Holiday (Labor), Holiday (Republic)
138936952	Discount percent, seasonality index	Average shelf price
138936953	Discount percent, seasonality index	Average shelf price

Table 2: Explanatory model stats of three products

Based on our model's evaluation, the error of the models in predicting the sales were as follows:

Product ID	RMSE	MAE	MAPE
138936951	25.403	19.475	0.093
138936952	16.978	12.47	0.1229
138936953	5.178	3.339	0.4179

Table 3: Evaluation of y prediction of three products

As we can see, MAPE is the highest for product 3 even though RMSE is low. This is because the sales quantities are low for product 3 in general comparing to the other two products, in fact some weeks had no sales at all. Conversely, for product 1 which has the most data, MAPE is quite low. Thus, MAPE does a better job in evaluating the model than measures like RMSE and MAE as it takes into account the overall magnitude of the quantity of each product.

6. Calculating DueTos

In a marketing mix model, the base is computed as the sum of sales components that are not considered in decision making or cannot be affected or changed by the decision maker. For multiplicative models, we can calculate DueTo_{it} as the gap between the predicted value of the dependent variable when a sales driver is added and the predicted sales of which that sales driver is kept at its baseline. To find the DueTos, we first need to establish what x_{itBASE} will be. To set the x_{itBASE}, we need to implement a bit of domain knowledge. For media and promotions, x_{itBASE} is generally 0. So for factors such as the holidays or email promotions, the x_{itBASE} is set to 0. For values regarding price and discount, we chose to set their x_{itBASE} values to the average over two

years. Thus, our x_{itBASE} contained either 0's for the marketing terms, or averages for price and discount related terms.

Our process to finding the DueTos was the following:

- Select a product x_i and its corresponding x_{itBASE}
- Calculate y_{hat}, the estimated sales quantity of the product, and y_{hat} where $x_{it} = x_{itBASE}$ for all independent variables in our model on a weekly level
- Find the difference of the two values y_{hat} and y_{hat} ($x_{it} = x_{itBASE}$), which are the DueTos without scaling
- Sum all the raw DueTos for each week to see if there is a need to unbias these values. Yet since we have many insignificant independent variables in our model, their coefficients are very likely to be off which can skew our y_{hat} ($x_{it} = x_{itBASE}$) significantly. In this case, when summing up all the raw DueTos, we decided to scale down the values of DueTos of the insignificant independent variables while keeping the significant ones unchanged.
- Perform any debiasing necessary to make sure that the sum of DueTos is equal to the actual value of the weekly sales quantity.

Thus, we are left with our finalized DueTos after scaling properly for all variables for each of the three products.

7. Results

After finding the results, we found that some of our DueTos came out to be very largely negative, despite taking place on non-significant variables. For example, the add of a holiday factor would decrease by 50 on sales units compared to the baseline of sales which shows lesser than negative 50 on DueTo of this variable, despite not being significant in the model for that product and only appearing a few times in the dataset. We are unable to pinpoint the exact point of the error, as the formulation of the y_{hat}'s in this section is the same as above, especially for the single can product where it showed accurate y_{hat} with a low MAPE. Because of the very largely negative DueTos, as mentioned above, to calculate the sum of the DueTos for each week before debiasing them, we scaled down the large DueTos of some insignificant independent variables. After scaling them to the actual value of the dependent variables, the results are shown below:

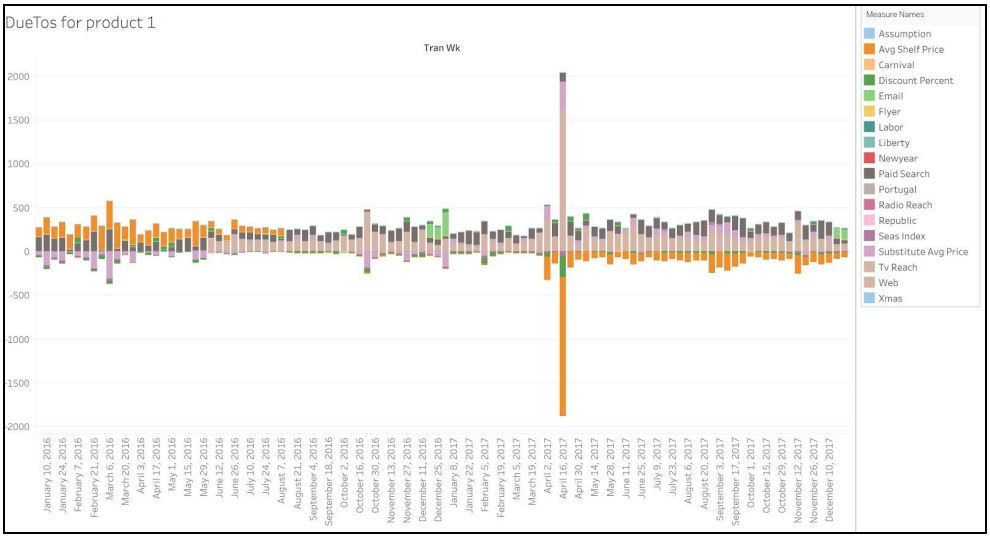


Figure 2: DueTos for product 1 by week

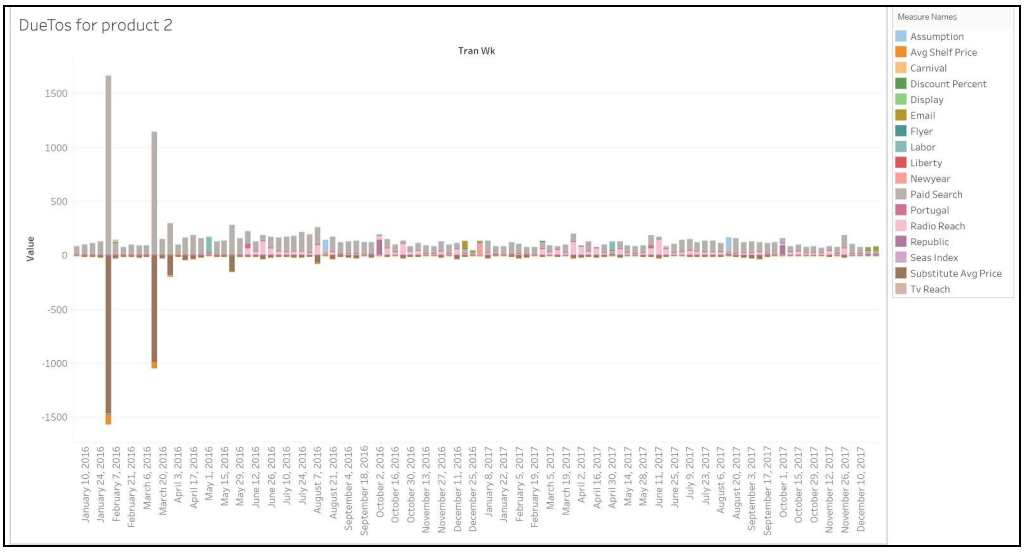


Figure 3: DueTos for product 2 by week

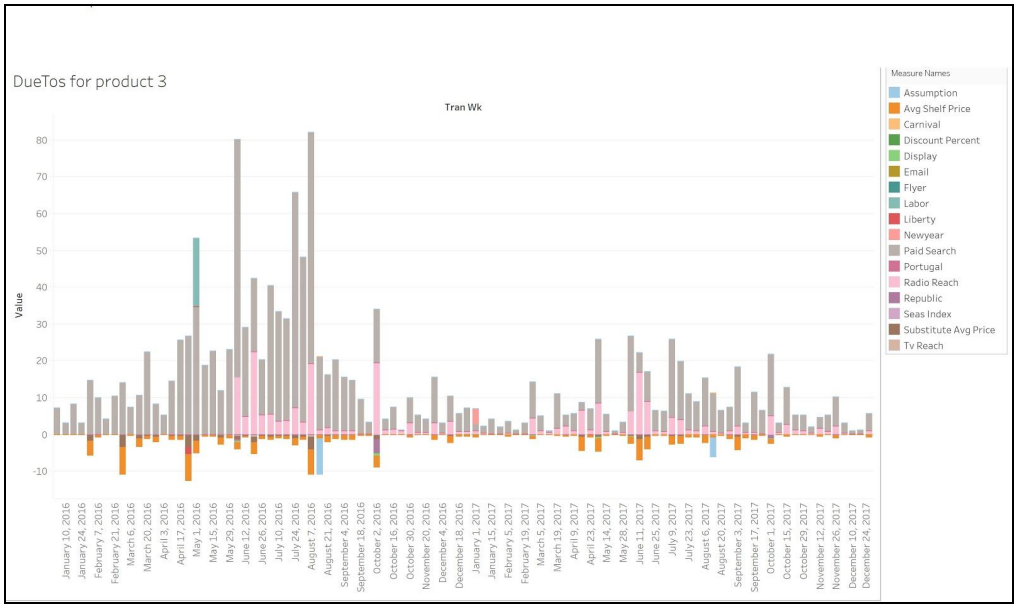


Figure 4: DueTos for product 3 by week

8. Conclusion and Next Steps

After completing our analysis, we recommend that San Miguel continues to use the Paid Search and Email advertising campaigns, as their low cost, high impressions nature bodes well for their smaller company. These vehicles are effective ways to boost their profile, and this their revenue for the upcoming year. On the opposite side, we recommend that they ease off the TV and Radio advertising, as they are more expensive and they do not bring near enough revenue/publicity to justify the costs.

As for next steps, we would definitely want to double back an improve our models and DueTos so that we can be more confident in the results that we find. We chose multiplicative model instead of the additive model because we understand that there are interactions among independent variables. Another model we could use is logit model which shows complex implicit interactions among the causal values, and it sets the dependent variable as a bounded value which may help improve the model performance. We recognize that something somewhere probably went wrong during the calculations of DueTos, but we feel that our processes and thinking were sound throughout this section as well as the whole the project. However, we would strive to improve in these areas in the future.

Appendix:

Statistical Summary of the Model Based on Product 1 (single can) and Variable Significance

```
> summary(model_prod1)

Call:
lm(formula = log(weekly_qty) ~ avg_shelf_price + discount_percent +
    seas_index + tv_reach + radio_reach + flyer + email + paid_search +
    web + holiday_NEWYEAR + holiday_CARNIVAL + holiday_LIBERTY +
    holiday_LABOR + holiday_PORTUGAL + holiday_ASSUMPTION + holiday_REPUBLIC +
    holiday_XMAS + substitute_avg_price, data = prod_1_weekly)

Residuals:
     Min       1Q   Median       3Q      Max
-0.41429 -0.07220  0.01075  0.07640  0.32054

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          5.537e+00  3.231e-01  17.135  < 2e-16 ***
avg_shelf_price     -1.146e+00  5.533e-01  -2.071  0.04139 *
discount_percent     1.902e+00  3.685e-01   5.162 1.58e-06 ***
seas_index           2.064e-05  1.024e-05   2.016  0.04691 *
tv_reach             2.624e-01  2.567e-01   1.022  0.30966
radio_reach         -1.055e-01  1.782e-01  -0.592  0.55552
flyer               -6.676e-02  1.121e-01  -0.596  0.55307
email                4.593e-07  5.038e-07   0.912  0.36457
paid_search          2.321e-06  1.033e-06   2.247  0.02725 *
web                  7.976e-08  1.726e-07   0.462  0.64518
holiday_NEWYEAR      1.460e-01  1.696e-01   0.861  0.39167
holiday_CARNIVAL    -6.907e-02  1.184e-01  -0.583  0.56132
holiday_LIBERTY     -1.456e-01  1.078e-01  -1.350  0.18045
holiday_LABOR       -3.291e-01  1.083e-01  -3.038  0.00316 **
holiday_PORTUGAL    -1.929e-01  1.431e-01  -1.348  0.18136
holiday_ASSUMPTION   2.659e-02  1.005e-01   0.265  0.79203
holiday_REPUBLIC    -3.787e-01  1.292e-01  -2.931  0.00434 **
holiday_XMAS        -2.190e-01  1.361e-01  -1.609  0.11139
substitute_avg_price 4.628e-01  3.978e-01   1.163  0.24788
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1341 on 85 degrees of freedom
Multiple R-squared:  0.5151,    Adjusted R-squared:  0.4124
F-statistic: 5.016 on 18 and 85 DF,  p-value: 1.522e-07
```

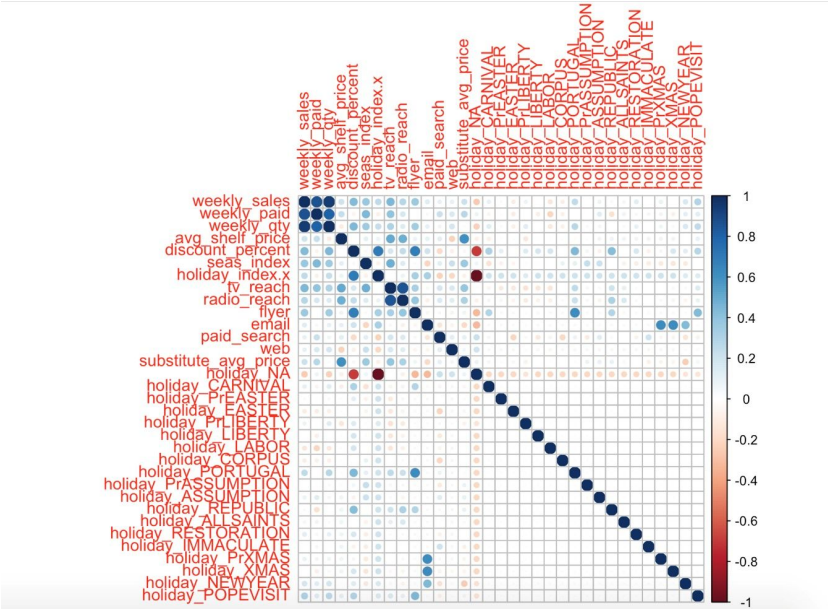
Statistical Summary of the Model Based on Product 2 (6-pack) and Variable Significance

```
> summary(model_prod2)

Call:
lm(formula = log(weekly_qty) ~ avg_shelf_price + discount_percent +
    seas_index + tv_reach + radio_reach + flyer + display + email +
    paid_search + web + holiday_NEWYEAR + holiday_CARNIVAL +
    holiday_LIBERTY + holiday_LABOR + holiday_PORTUGAL + holiday_ASSUMPTION +
    holiday_REPUBLIC + holiday_XMAS + substitute_avg_price, data = prod_2_weekly)

Residuals:
     Min       1Q   Median       3Q      Max
-0.40032 -0.08534  0.00224  0.07484  0.51712

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          6.447e+00  8.009e-01   8.050 4.82e-12 ***
avg_shelf_price     -4.664e-01  1.042e-01  -4.477 2.37e-05 ***
discount_percent     2.461e+00  6.788e-01   3.626 0.000493 ***
seas_index           8.565e-05  1.483e-05   5.776 1.26e-07 ***
tv_reach            -1.847e-01  3.590e-01  -0.514 0.608347
radio_reach          6.831e-02  2.514e-01   0.272 0.786474
flyer               -3.971e-02  8.112e-02  -0.490 0.625755
display             -7.286e-02  1.049e-01  -0.695 0.489165
email                2.269e-07  6.992e-07   0.325 0.746344
paid_search          1.990e-06  1.402e-06   1.420 0.159417
web                 -3.076e-07  2.366e-07  -1.300 0.197178
holiday_NEWYEAR      3.391e-01  2.419e-01   1.402 0.164578
holiday_CARNIVAL     6.483e-03  1.672e-01   0.039 0.969156
holiday_LIBERTY     -7.962e-02  1.463e-01  -0.544 0.587830
holiday_LABOR        1.339e-01  1.475e-01   0.908 0.366549
holiday_PORTUGAL     2.342e-02  1.894e-01   0.124 0.901899
holiday_ASSUMPTION   2.289e-01  1.411e-01   1.623 0.108436
holiday_REPUBLIC     1.611e-01  1.764e-01   0.913 0.363703
holiday_XMAS         7.317e-02  1.947e-01   0.376 0.708020
substitute_avg_price -1.684e-01  1.521e-01  -1.107 0.271290
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.185 on 84 degrees of freedom
Multiple R-squared:  0.6876,    Adjusted R-squared:  0.617
F-statistic: 9.732 on 19 and 84 DF,  p-value: 3.51e-14
```

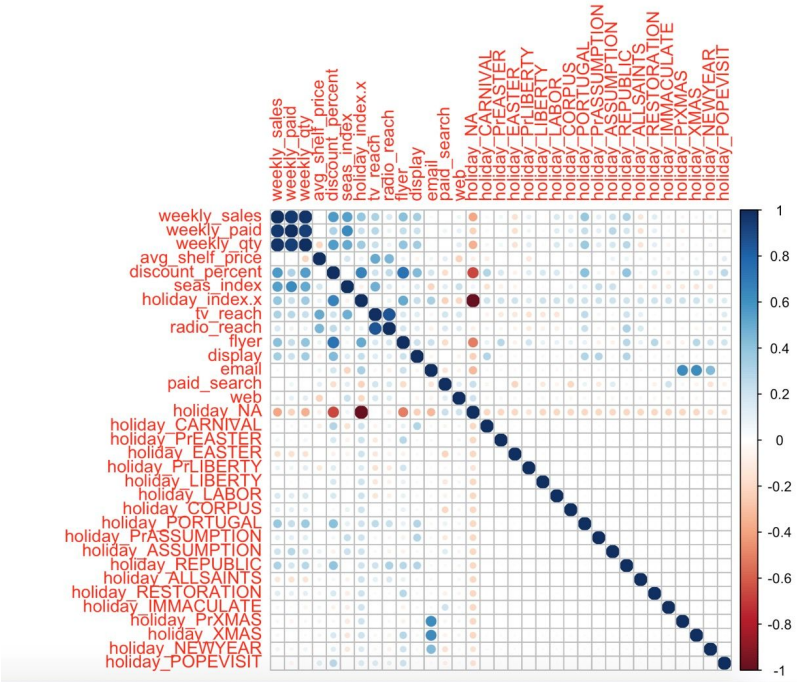
Statistical Summary of the Model Based on Product 3 (case) and Variable Significance

```
> summary(model_prod3)

Call:
lm(formula = log(weekly_qty) ~ avg_shelf_price + discount_percent +
    seas_index + tv_reach + radio_reach + flyer + display + email +
    paid_search + web + holiday_NEWYEAR + holiday_CARNIVAL +
    holiday_LIBERTY + holiday_LABOR + holiday_PORTUGAL + holiday_ASSUMPTION +
    holiday_REPUBLIC + holiday_XMAS + substitute_avg_price, data = prod_3_weekly)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4089 -0.2524  0.0758  0.2775  0.9635

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          8.557e+00  1.464e+00   5.844 1.06e-07 ***
avg_shelf_price     -5.823e-01  9.515e-02  -6.120 3.28e-08 ***
discount_percent     9.560e+00  2.602e+00   3.675 0.000429 ***
seas_index           3.366e-04  4.023e-05   8.367 1.54e-12 ***
tv_reach            -1.240e+00  1.013e+00  -1.224 0.224393
radio_reach          8.624e-01  7.068e-01   1.220 0.226033
flyer               -8.196e-02  4.041e-01  -0.203 0.839808
display             -8.174e-01  6.075e-01  -1.346 0.182251
email               -3.823e-06  1.994e-06  -1.917 0.058747 .
paid_search          4.572e-06  4.150e-06   1.102 0.273951
web                  2.430e-07  6.754e-07   0.360 0.719978
holiday_NEWYEAR      5.878e-01  6.670e-01   0.881 0.380858
holiday_CARNIVAL    -7.144e-01  5.130e-01  -1.393 0.167596
holiday_LIBERTY     -7.101e-01  4.901e-01  -1.449 0.151254
holiday_LABOR        2.400e-01  5.941e-01   0.404 0.687281
holiday_PORTUGAL    -6.435e-01  5.986e-01  -1.075 0.285604
holiday_ASSUMPTION  -5.943e-01  3.896e-01  -1.525 0.131115
holiday_REPUBLIC    -6.131e-01  5.902e-01  -1.039 0.302075
holiday_XMAS         9.698e-01  6.206e-01   1.563 0.122056
substitute_avg_price -1.299e-02  7.009e-02  -0.185 0.853478
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5197 on 80 degrees of freedom
Multiple R-squared:  0.7763,    Adjusted R-squared:  0.7231
F-statistic: 14.61 on 19 and 80 DF,  p-value: < 2.2e-16
```

Product 1 (single can) Correlation of Independent Values



Product 2 (6-pack) Correlation of Independent Values



Product 3 (case) Correlation of Independent Values