This document is a summary of the analysis performed to understand the drivers of sales volume. As part of our study we performed preliminary data investigation, built linear regression models, identified the best performing model and finally interpreted the results of implications to businesses and/or marketing managers.

Preliminary Data Investigation

In the given dataset we have various advertisement channels (store, billboard, and printouts), amount spent by competitors on advertising, price, and satisfaction as our independent variables and sales value as the dependent variable. Through preliminary data investigation we explored the data in a systematic manner and evaluated different ideas based on our business understanding.

We performed summary statistics analysis to help us with basic insights of the dataset. Below are our findings, refer to Figure 1 in Appendix for the full detailed summary.

- Minimum satisfaction was above 50, but the mean satisfaction is only 70 (not very high).
- Most of the expenditure goes towards the store advertising channel. That is, the minimum spend on in store advertisement is much higher than other two channels.
- The range of the price is small and steady.

Figure 2 in Appendix is the generalized pair plot that helped us identify if there are any concerns with the dataset, such as non-normal variables, and high correlations between any pairs. Within our given dataset, we do not have any variables that are skewed or highly correlated with other variables. Therefore, no further transformation was required. Furthermore, we noticed that the store and the billboard have a linear relationship with sales value.

As part of our final step under preliminary data investigation, we investigated correlation of independent variables with the dependent variable. Figure 3 in the Appendix shows that billboard followed by store are the most correlated variables with sales. All other variables have minimal to zero correlation with sales.

Model Comparison

We built seven linear regression models that predict sales volume based on various independent variables as mentioned earlier. To identify which model performed the best, we predicted on the hold out data and used R^2 as a measure of evaluation. The holdout prediction helps us avoid overfitting of the model. The results of R^2 are presented in Table 1 in the Appendix. Linear regression model trained as M7 performed the best with a test R^2 value of **~0.9283**. That is, the model with interactions on advertisement channels (store:billboard) best predicted the sales. Summary statistics and correlation plot for M7 are presented in Figures 4 & 5 in the Appendix.

Figure 6 in the Appendix shows the holdout prediction plot for M7. This outlines the actual hold out values (blue line) plotted against the predicted values (red line) and we can see the model predicted very well.

Next, we checked for violation assumptions, below are the observations from the analysis. Figure 7 in the Appendix shows the plots for this check.

- Normal Q-Q plot shows that residuals are normally distributed. No obvious patterns are visible.

- Contains a few outliers, nothing very significant.
- There is no violation of assumptions, therefore linear models can be used.

Discussion

Does advertising lead to higher sales volume?

There are 3 advertising channels: 1) store, 2) billboard and 3) printout. Based on our best Model (M7) all the 3 channels have positive coefficients. This implies that by increasing the advertising component leads to increase in sales volume. Refer to Summary statistics for details on values in Figure 4. To summarize:

- If a company increases the store advertising by $1, the average sales increases by $2.34.
- If a company increases the billboard advertising by $1, the average sales increases by $1.93
- If a company increases the printout advertising by $1, the average sales increases by $0.14

We see that Increase in printout does not lead in significant increase in sales, as discussed in the further question.

Which channels are more effective for advertising?

As p-value for store and store:billboard is much lower than 0.001, we can reject the null hypothesis that β_{store} and $\beta_{store:billboard}$ are equal to zero. This means that these two variables have an impact on the dependent variable. On the contrary, the p-value for billboard and printout is greater than 0.05, so we fail to reject null hypothesis.

Conclusion: Based on the coefficients and p-value, we can observe that store and synergy between store and billboard are the most effective channels for advertising. Though billboard alone isn't as effective a channel for advertising, pairing billboards with stores has a much greater impact on sales.

Note: The hierarchical principle states that if we include an interaction in a model, we should also include the main effects, even if the p-values associated with the principle coefficients are not significant. Here, in our exercise though billboard alone doesn't have a significant impact, but interaction between store and billboard has significant impact so we should keep billboard also for the sales prediction.

Are our customers price sensitive?

1. As p-value for price is much less than 0.001, we can reject the null hypothesis that β_{price} is equal to zero.
2. The price coefficient of -199.3 represents the mean increase of sales in dollars for every additional one dollar decrease in price. If a company decreases the price by $1, the average sales increases by $199.3.

Based on the significant and the coefficient values of price, we can conclude that customers are price sensitive.

How do competitor's actions effect our sales?

As p-value for competitors' action is much lower than 0.001, we know that it has an impact on the dependent variable. Furthermore, competition coefficient is negative which means that as competitors increase their actions, our sales are negatively implicated. Additional $1 increase in competitors action decreases our sales by $1.53

Can analytics be used to predict the sales volume?

Yes, analytics can be used to predict sales volume as demonstrated in this exercise where we have used multiple linear regression as one of the methods to predict the sales. Multiple linear regression is used to explain the relationship between one continuous dependent variable and two or more independent variables. In our exercise,

sales is the dependent variable and store, billboard, printout, satisfaction level, competitor expenditure and price are the independent variables. For the equation of our best performed model refer to Equation 1 in the Appendix.

Multiple regression can help in predicting the sales volume by various means, such as:

First, it might be used to identify the strength of the effect that the independent variables have on a dependent variable. For example: We have used p-values to show how store advertising is significant in determining the sales but not printout advertising.

Second, it can be used to calculate the impacts of changes. We have seen how coefficient values calculate the change in sales with respect to change in one of the independent variables. For example, in the case of a store (β_{store} = 2.342) the average change in sales is $2.34 if the store advertising is increased by $1.

Third, it can also calculate how the synergy between the two independent variables impacts the dependent variable. For example, billboard advertising alone isn't effective for the sales, but store and billboard combined have a significant impact on the sales.

Relevancy of the exercise to the business

This exercise (Linear Regression) is extremely important for a business and especially marketing teams as it offers a way to measure the relationship of different variables to the output. Some of the outputs of the Linear Regression model include R^2, p-value, and estimate. R^2 explains how much of the variance in the dependent variable is explained by the independent variables collectively, p-value determines if there is statistical significance that a relationship exists between the independent and dependent variables at the population level, and the estimate sign shows if there is a negative or positive relationship while the value signifies the change in dependent variable given one unit change in the independent variable, holding other variables constant.

Understanding these outputs and the relationship between different factors can help marketers conduct a "Due to" analysis to understand and dissect the previous outcomes of business or marketing activities. This can be extended to a "What-if" analysis where the impact of different activities or factors can be simulated to gain actionable insights and prioritize the most profitable activities accordingly. This exercise also provides the ability to further predict future outcomes and optimize their business processes accordingly. As displayed through previous questions, this exercise enables businesses with the information they need to better plan and prepare.

Appendix

store	billboard	printout	sat	comp	price	sales
Min. :1150	Min. : 219	Min. : 26.0	Min. :54.00	Min. : 230.0	Min. : 85.00	Min. : 5819
1st Qu.:1831	1st Qu.: 845	1st Qu.: 626.0	1st Qu.:66.00	1st Qu.: 656.0	1st Qu.: 96.00	1st Qu.:15062
Median :1990	Median :1003	Median : 814.0	Median :70.00	Median : 788.0	Median :100.00	Median :17385
Mean :1993	Mean :1003	Mean : 806.7	Mean :69.45	Mean : 791.7	Mean : 99.75	Mean :17586
3rd Qu.:2153	3rd Qu.:1168	3rd Qu.: 979.2	3rd Qu.:73.00	3rd Qu.: 931.0	3rd Qu.:103.00	3rd Qu.:20038
Max. :2798	Max. :1791	Max. :1555.0	Max. :85.00	Max. :1339.0	Max. :117.00	Max. :33767

Figure 1: Summary statistics of the dataset

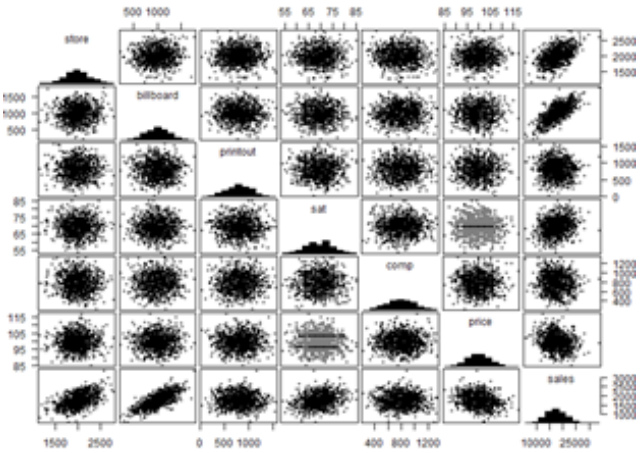

Figure 2: Generalized pair plots of the dataset

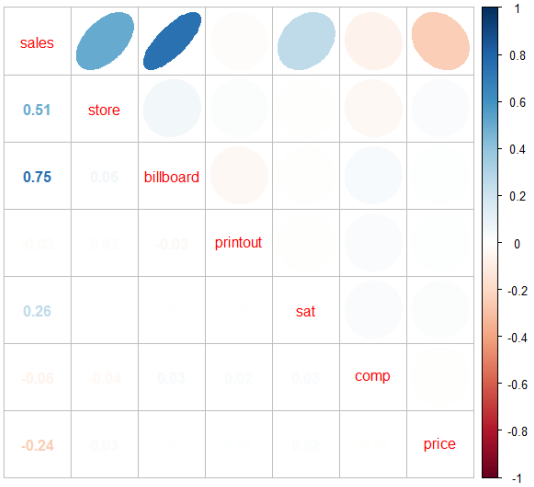

Figure 3: Correlation plot

```
Call:
lm(formula = sales ~ price + sat + comp + printout + (store +
    billboard)^2, data = train.df)

Residuals:
    Min      1Q  Median      3Q     Max
-2931.4  -707.9     2.0   691.6  3343.8

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      9.130e+03  1.552e+03   5.882 6.15e-09 ***
price           -1.993e+02  7.356e+00 -27.096  < 2e-16 ***
sat              1.993e+02  7.357e+00  27.088  < 2e-16 ***
comp            -1.534e+00  1.871e-01  -8.195 1.10e-15 ***
printout         1.396e-01  1.434e-01   0.974 0.330591
store            2.342e+00  6.235e-01   3.756 0.000186 ***
billboard        1.929e+00  1.207e+00   1.598 0.110518
store:billboard  4.486e-03  6.043e-04   7.423 3.15e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1000 on 742 degrees of freedom
Multiple R-squared:  0.9257,    Adjusted R-squared:  0.925
F-statistic:  1320 on 7 and 742 DF,  p-value: < 2.2e-16
```

Figure 4: Summary statistics of the best performed model – M7

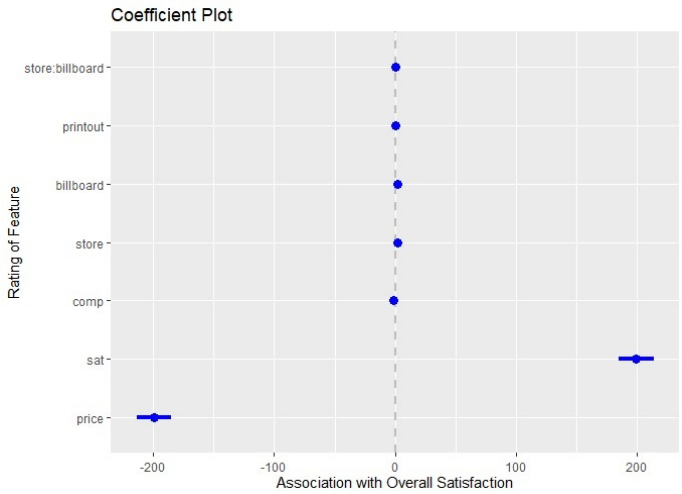

Figure 5: Coefficient plot of the best performed model - M7

Table 1: Train and Test R^2 results for seven linear regression models

Models	Variable Used	Train R^2	Test R^2
M1	Price	0.0596	0.0441
M2	Price, Store	0.3237	0.3121
M3	Price, Store, Billboard	0.8393	0.8541
M4	Price, Store, Billboard, Printout	0.8392	0.8570
M5	Price, Store, Billboard, Printout, Satisfaction	0.913	0.9158
M6	Price, Store, Billboard, Printout, Satisfaction, Comp.	0.9195	0.9202
M7	**Price, Printout, Satisfaction, Comp., Interactions - (Store + Billboard)^2**	**0.925**	**0.9283**

Equation 1: Best performed model (M7) prediction equation

Sales = $\beta_{intercept}$ + β_{price}* (price) + β_{store}* (store) +$\beta_{billboard}$* (billboard) +$\beta_{printout}$* (printout) +$\beta_{sat.}$* (satisfaction) + $\beta_{comp.}$* (comp.) + $\beta_{store:billboard}$* (store * billboard)

OR

Sales = 9,130 - 199.30* price + 2.342 * store + 1.929 * billboard + 0.1396 * printout + 199.3 * satisfaction -1.534 * comp. + 0.004486 * (store * billboard)

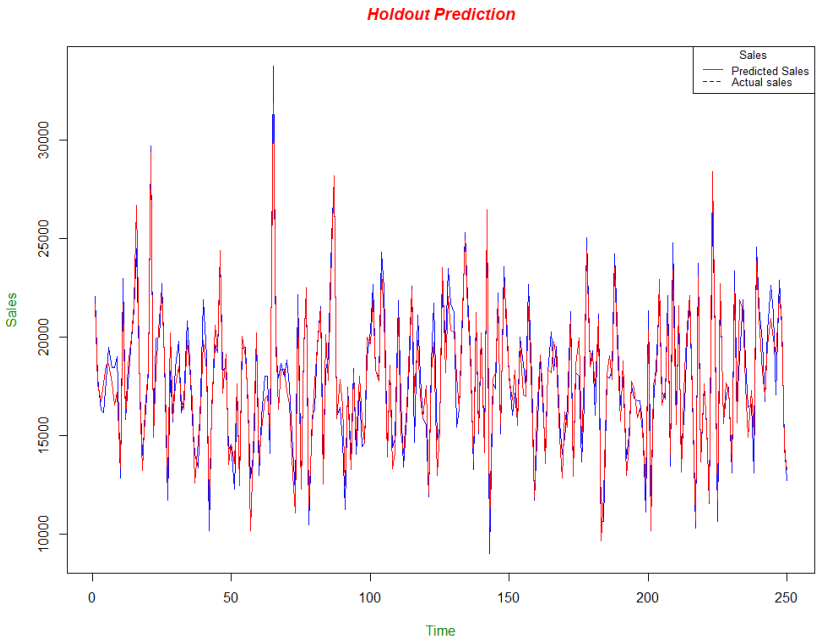

Figure 6: Holdout Prediction of the best performed model - M7

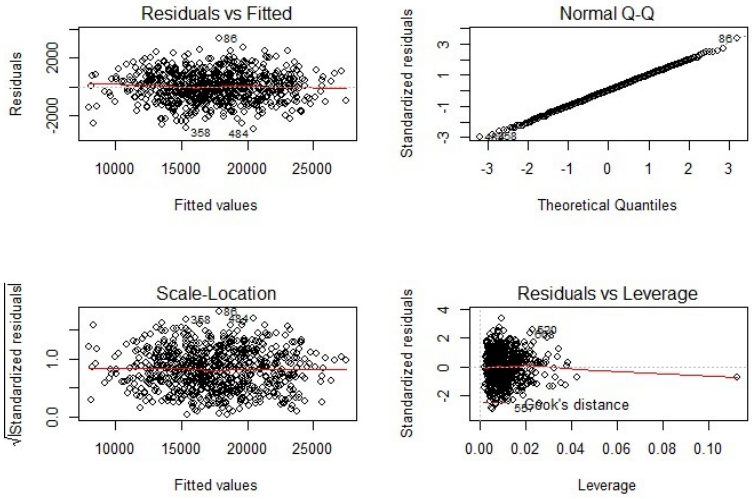

Figure 7: Checking for violation of assumptions on best performed model - M7