CID: 01770410

MODULE: RETAIL AND MARKETING ANALYTICS

Introduction

The main aim of this report is to explore sales forecasting and attitudinal metrics analysis for effective marketing. With data-driven decision making becoming increasingly important for businesses to succeed in a competitive market, forecasting sales with models such as multiple linear regression and ARIMA can provide valuable information to help companies plan for growth. Analysing attitudinal metrics can also help businesses understand the effectiveness of their advertising and promotional campaigns in winning over customers. By using these metrics, businesses can make informed decisions on how to allocate resources efficiently and effectively to maximize their brand's success.

The dataset used for this report is the shampoo.xls dataset provided by the teacher, which includes information on sales, awareness, consideration, price, promotion, and liking of two different shampoo brands over a period of 96 months (8 years). The first part of the report will focus on sales forecasting for the two different shampoo brands using the Multiple Linear Regression model and the ARIMA model. The second part of the report will analyse the attitudinal metrics to determine the effectiveness of advertising and promotions for the two different shampoo brands. This analysis will provide insight into whether promotions have been effective in winning the minds and hearts of customers and how much additional investment in these campaigns could benefit the brands.

Sales Forecasting for Two Shampoo Brands

The dataset was examined for missing data, and none were found. It was then split into two datasets for the two different shampoo brands. To begin ARIMA forecasting, the datasets were converted into time series objects. Next, exploratory data analysis was performed by decomposing the time series into trend and seasonality components using the 'stl' function in R. The results of the decomposition are shown below:

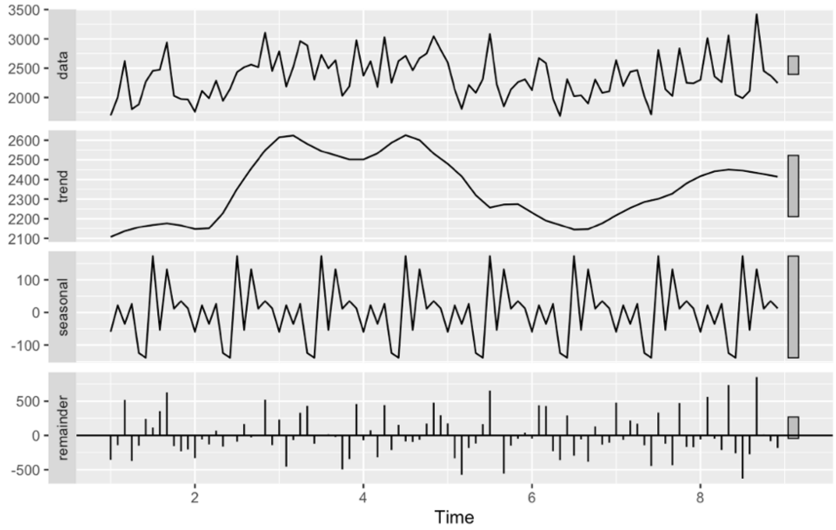


Figure 1 Result of Decomposition on Time Series Data for First Shampoo Brand

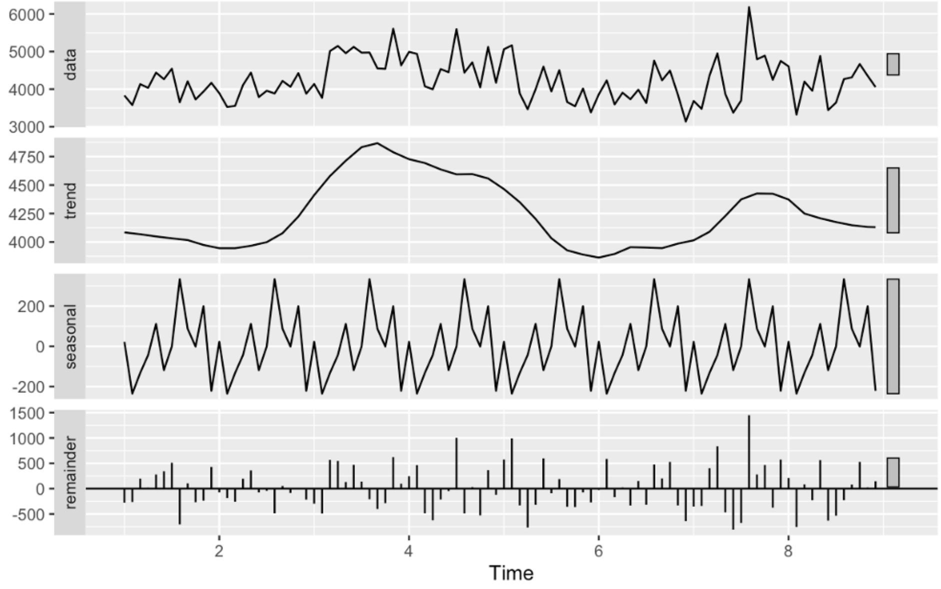


Figure 2 Result of Decomposition on Time Series Data for Second Shampoo Brand

From the plot above, the seasonal component is not as prominent as the trend component. Next step was to split the time series into test and train data for the purpose of the forecast. 80% of the dataset was assigned to the train dataset while the other 20% was assigned to the test dataset.

Next, we will conduct sales forecasting for both brands.

Brand 1- ARIMA Model

Stationarity tests were conducted on the time series data, including the Augmented Dickey Fuller, Phillips-Perron, and KPSS tests. The ADF test resulted in a p-value of 0.097, failing to reject the null hypothesis of non-stationarity. However, the PP test yielded a p-value of 0.01, indicating stationarity at a significance level of 0.05. The KPSS test produced a p-value of 0.1, further confirming the stationarity of the time series.

The subsequent step is to create ACF and PACF plots.

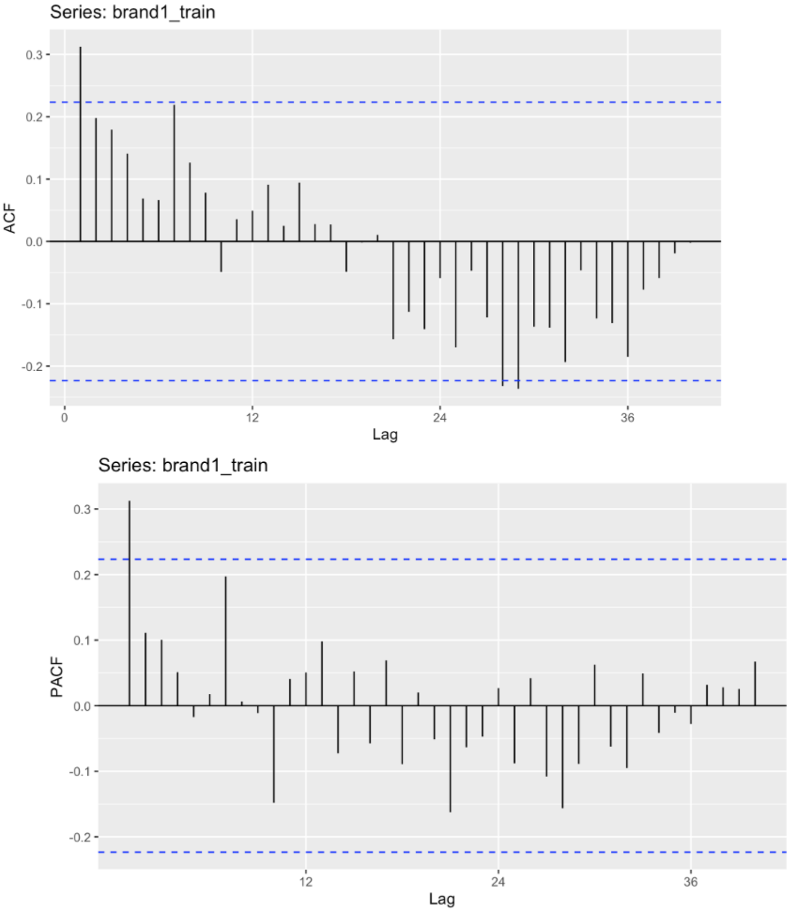
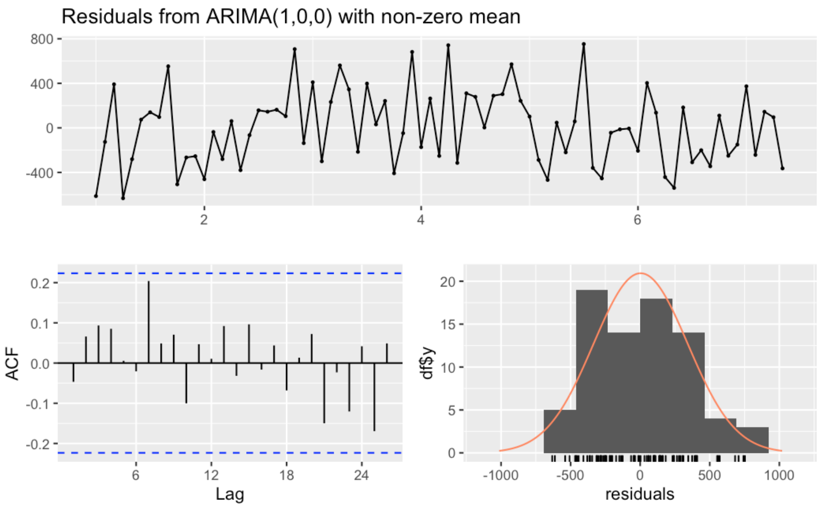
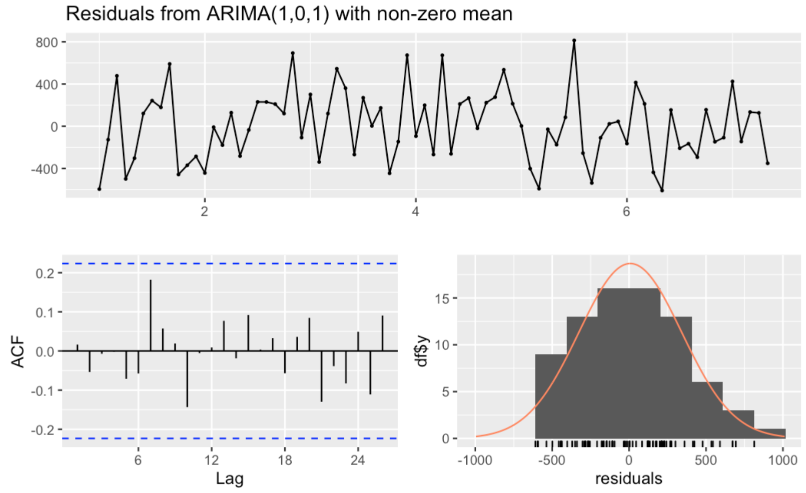


Fig 2. ACF and PACF plots

The ACF and PACF plots suggest low autocorrelation in the time series, which means the relationship between current and past values can be utilized for better forecasts. The auto.arima() function estimated the best ARIMA model, which is ARIMA(1,0,0) with non-zero mean and AIC of 1121.05. The next best models with the lowest AICs are ARIMA(1,0,1) and ARIMA(2,0,0) with non-zero mean.

The ARIMA() function was used to fit ARIMA models to the training data, followed by using the checkresiduals() function to evaluate the residuals' quality. The Ljung-Box test p-values (0.8243, 0.8759 and 0.8452) for the ARIMA models in the order they were mentioned were all above 5%, indicating independently distributed residuals. Residual plots showed a white noise pattern with no trends or significant autocorrelations, confirming the residuals' independence. The normal distribution of residuals indicates that all models are good.


Residuals from ARIMA(1,0,0) with non-zero mean


Residuals from ARIMA(1,0,1) with non-zero mean

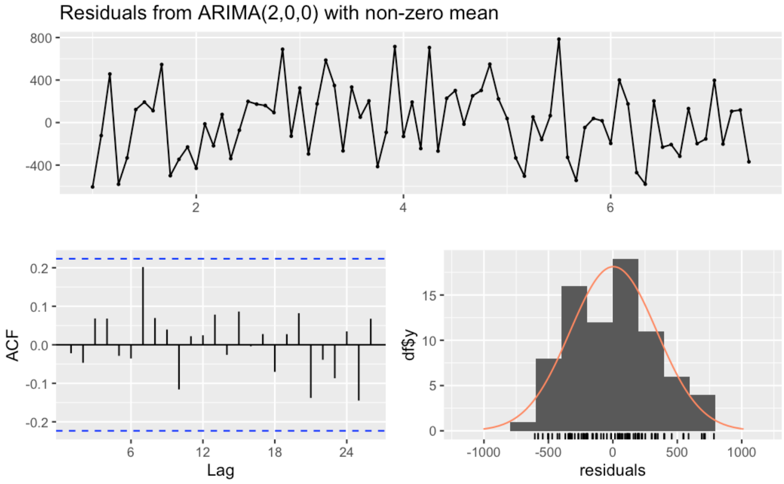
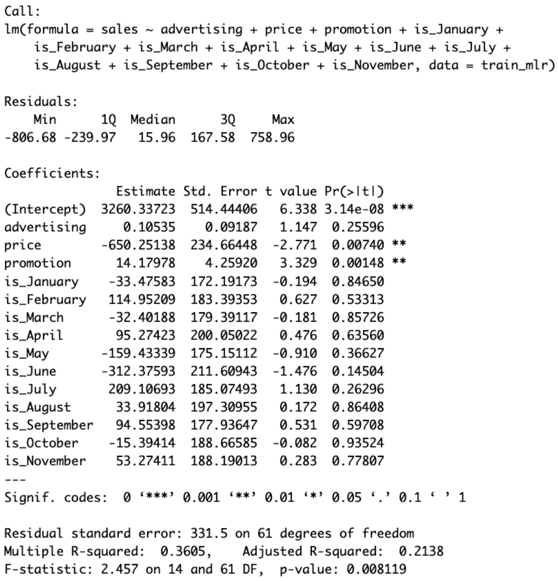


Fig 3. Result from checkresiduals() function for different ARIMA models

I used three ARIMA models to generate forecasts based on the test set and compared their accuracy with the actual values. The first model ARIMA(2,0,0) had the lowest root mean square error (RMSE) of 420.4542, while the other two models had RMSEs of 426.7279 and 424.1139. I will compare the accuracy of the ARIMA model with the lowest RMSE to that of the final multiple linear regression model that will be selected.

Brand 1- Multiple Linear Regression Model

The chosen independent variables for the regression model are advertising, price, and promotion, as they are expected to significantly impact shampoo sales. Additionally, the time variable was converted to dummy variables based on the 12 months of the year to account for seasonal effects. Results of the regression are shown below:



```
Call:
lm(formula = sales ~ advertising + price + promotion + is_January +
    is_February + is_March + is_April + is_May + is_June + is_July +
    is_August + is_September + is_October + is_November, data = train_mlr)

Residuals:
    Min      1Q  Median      3Q     Max
-806.68 -239.97   15.96  167.58  758.96

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  3260.33723  514.44406   6.338 3.14e-08 ***
advertising     0.10535    0.09187   1.147  0.25596
price        -650.25138  234.66448  -2.771  0.00740 **
promotion      14.17978    4.25920   3.329  0.00148 **
is_January    -33.47583  172.19173  -0.194  0.84650
is_February   114.95209  183.39353   0.627  0.53313
is_March      -32.40188  179.39117  -0.181  0.85726
is_April       95.27423  200.05022   0.476  0.63560
is_May       -159.43339  175.15112  -0.910  0.36627
is_June      -312.37593  211.60943  -1.476  0.14504
is_July       209.10693  185.07493   1.130  0.26296
is_August      33.91804  197.30955   0.172  0.86408
is_September   94.55398  177.93647   0.531  0.59708
is_October    -15.39414  188.66585  -0.082  0.93524
is_November    53.27411  188.19013   0.283  0.77807
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 331.5 on 61 degrees of freedom
Multiple R-squared:  0.3605,	Adjusted R-squared:  0.2138
F-statistic: 2.457 on 14 and 61 DF,  p-value: 0.008119
```

Fig 4. Result from Linear Regression Model

The linear regression model's R-squared is 0.3605, and the promotional and price variables have a significant effect on sales based in their p-values. Advertising is not significant at 5% level, and seasonal dummy variables do not significantly affect sales. The AIC algorithm helps in choosing the best variables. Using AIC, the model with price, promotion, and May, June and July dummy variables as independent variables had the lowest AIC- 880.55 with an R-squared of 33.4%. The log of the price variable was also taken, with no significant increase in the R-squared. However, the model with the log(price) addition gave the lowest RMSE- 417.0508.

Analysis and Discussion

Looking at the two models used for this brand, the MLR model gives the lowest RMSE- 417.0508. However, since the discrepancy among all the RMSE values for all the models is not much, the ARIMA(1,0,1) will be used for the twelve month forecast. The figure below shows the result of the forecast for the next one year:

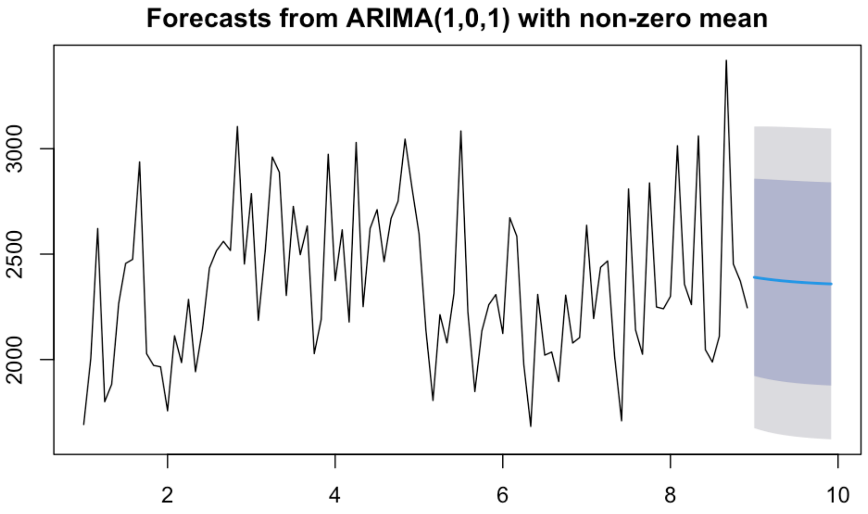


Fig 5. Result from ARIMA Forecast

A better forecast can be achieved by comparing with other models like the Holt-Winters and neural networks. Neural networks are particularly effective for forecasting because they can capture complex relationships and patterns in the data that may not be detected by an ARIMA model.

Brand 2- ARIMA model.

Stationarity tests, including ADF, PP, and KPSS, were conducted on the time series data. The PP test indicated stationarity at a significance level of 0.05 with a p-value of 0.01, while the ADF test yielded a p-value of 0.513, failing to reject the null hypothesis of non-stationarity. The KPSS test further confirmed stationarity with a p-value of 0.1. ACF and PACF graphs will be plotted.

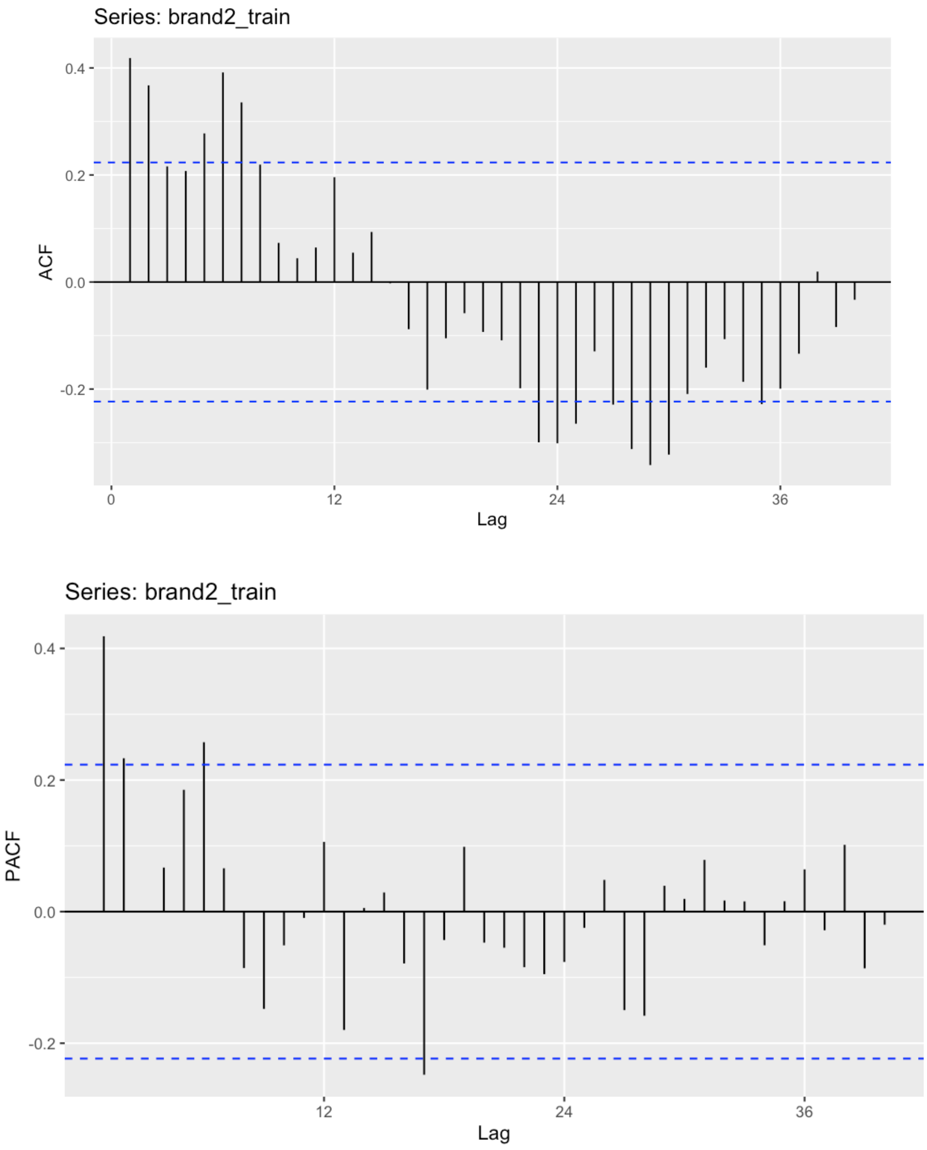


Fig 6. ACF and PACF plots

ACF plots showed significant upward and downward spikes that went beyond the confidence interval. This means that there is likely a lot of autocorrelations in this time series. The PACF plot shows no significant spikes beyond the confidence interval, indicating a stationary time series. The auto.arima() function selected ARIMA(2,0,0)(2,0,0)[12], with the next best models being ARIMA(2,0,0)(2,0,1)[12] and ARIMA(1,0,1)(2,0,0)[12]. All models had p-values greater than 0.05 in the Ljung-Box test, indicating independently distributed residuals. The residuals showed no trends, significant autocorrelations, and followed a normal distribution, indicating good model fits.

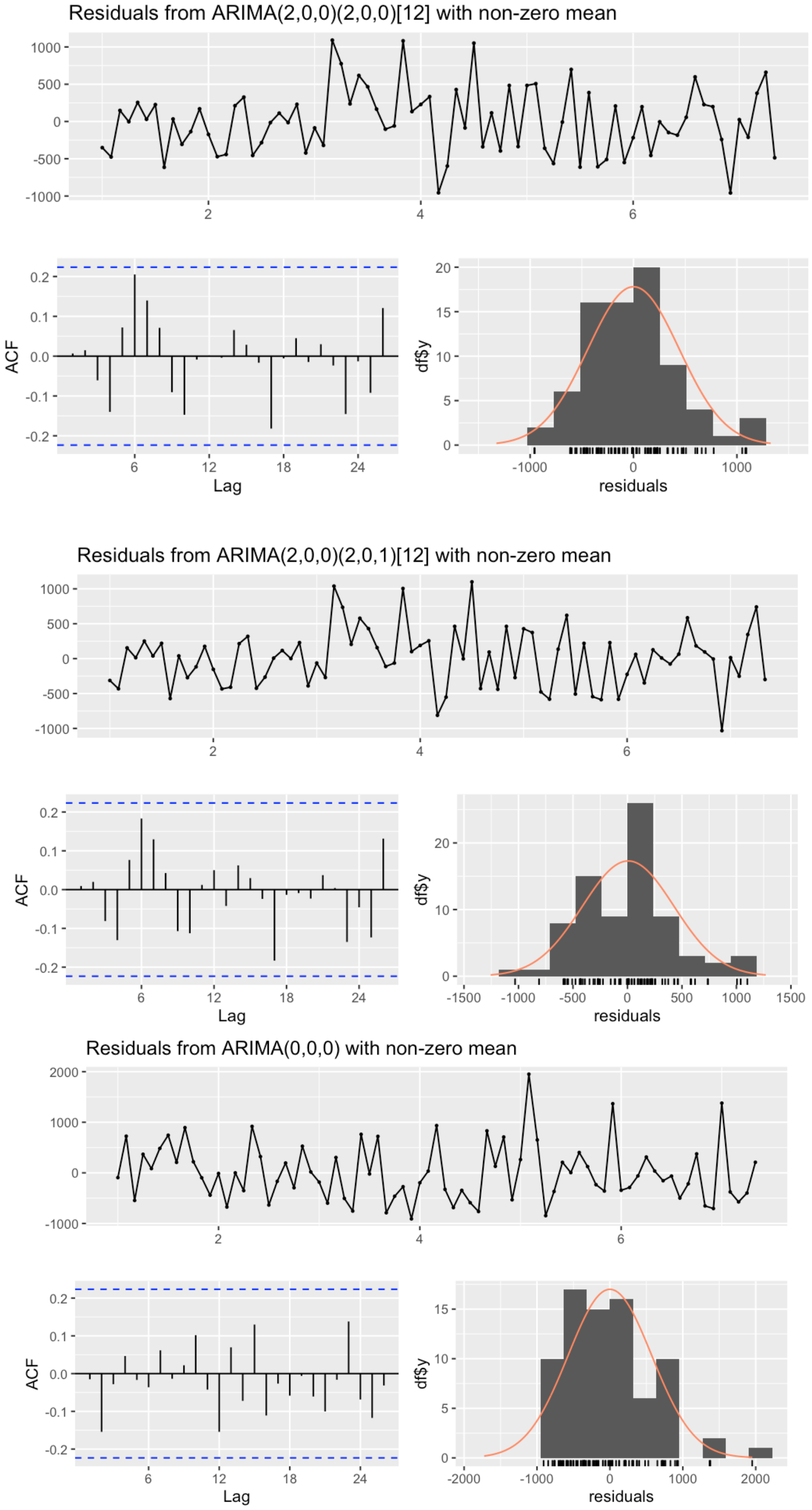
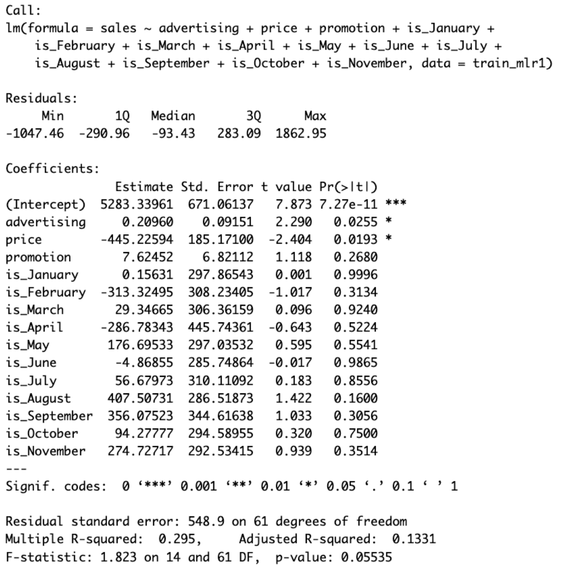


Fig 7. Result from checkresiduals() function for different ARIMA models

I used three ARIMA models to generate forecasts based on the test set and compared their accuracy with the actual values. The first model ARIMA(2,0,0)(2,0,0)[12] had the lowest root mean square error (RMSE) of 646.4412, when compared to the other two models.

Brand 2- Multiple Linear Regression Model.

Just like before, the independent variables in this multiple linear regression model will consist of advertising, price and promotion and the dummy variables for the dates. The results of the regression are shown below:



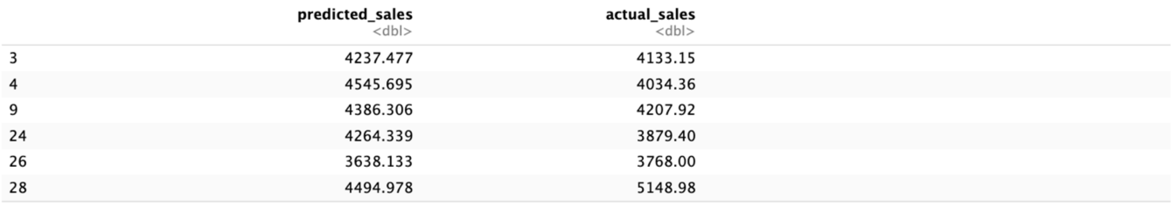
```
Call:
lm(formula = sales ~ advertising + price + promotion + is_January +
    is_February + is_March + is_April + is_May + is_June + is_July +
    is_August + is_September + is_October + is_November, data = train_mlr1)

Residuals:
     Min       1Q   Median       3Q      Max
-1047.46  -290.96   -93.43   283.09  1862.95

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  5283.33961  671.06137   7.873 7.27e-11 ***
advertising     0.20960    0.09151   2.290   0.0255 *
price        -445.22594  185.17100  -2.404   0.0193 *
promotion       7.62452    6.82112   1.118   0.2680
is_January      0.15631  297.86543   0.001   0.9996
is_February  -313.32495  308.23405  -1.017   0.3134
is_March       29.34665  306.36159   0.096   0.9240
is_April     -286.78343  445.74361  -0.643   0.5224
is_May        176.69533  297.03532   0.595   0.5541
is_June        -4.86855  285.74864  -0.017   0.9865
is_July        56.67973  310.11092   0.183   0.8556
is_August     407.50731  286.51873   1.422   0.1600
is_September  356.07523  344.61638   1.033   0.3056
is_October     94.27777  294.58955   0.320   0.7500
is_November   274.72717  292.53415   0.939   0.3514
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 548.9 on 61 degrees of freedom
Multiple R-squared:  0.295,	Adjusted R-squared:  0.1331
F-statistic: 1.823 on 14 and 61 DF,  p-value: 0.05535
```

Fig 8. Result from Linear Regression Model

The model shows that advertising spending and product price affect sales, while month of the year and promotional activity do not. The R-squared of 0.295 explains 29.5% of sales variation, making the model a useful starting point. AIC was used to select the most relevant variables for better prediction, resulting in a new recommended model with an AIC of 958.37 and variables including advertising, price, and February and August dummy variables. The R-squared of the new model was 23.4%, and adding a log function to the price variable did not significantly impact the R-squared. The model recommended using the AIC produced the lowest RMSE- 524.6293.

Analysis

The MLR model has a lower RMSE than the ARIMA model, making it the ideal choice for forecasting. The figure below shows that the MLR model performs well in predicting sales for the second shampoo brand, but it's important to compare its performance against other forecasting models and consider external factors that may affect accuracy. Evaluating the MLR model's performance against other models and performing sensitivity analysis can improve accuracy. Regularly updating the model with new data and monitoring its performance is also recommended.



	predicted_sales <dbl>	actual_sales <dbl>
3	4237.477	4133.15
4	4545.695	4034.36
9	4386.306	4207.92
24	4264.339	3879.40
26	3638.133	3768.00
28	4494.978	5148.98

Fig 9. Prediction Sales based on Actual Sales in the test set.

Mindset Metrics Analysis for Two Shampoo Brands

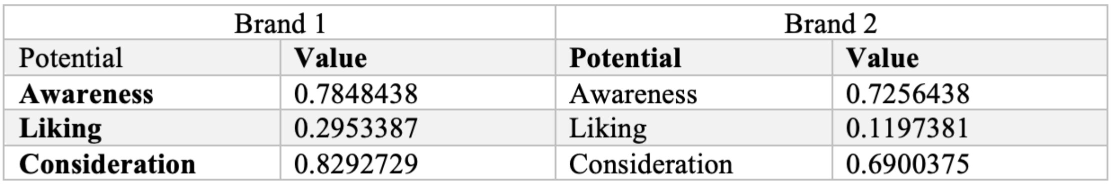
The following paragraphs will involve the computation of the potential, stickiness, and responsiveness of the mindset metrics. Subsequently, the conversion rate, which indicates the extent to which these metrics can influence sales, will also be calculated. Finally, the overall appeal of the attitudinal metrics can be estimated by combining all the results obtained from the analysis.

Potential

The potential for a product or service is its highest achievable sales in a specific market segment based on existing market factors. It can be calculated using the formula:

$$POT_t = \frac{[MAX - A_{t-1}]}{MAX}$$

Where $MAX = 100\%$ and $A_{t-1}]$ here represents the previous awareness/consideration. The potential for liking was calculated by converting liking scores to percentages and taking the average. Results for the potential for awareness, consideration, and liking of two shampoo brands are summarized in the table. (Hanssens et al., 2014).

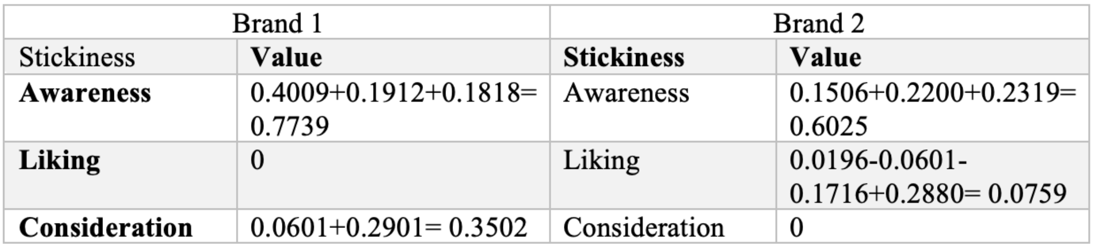


Brand 1		Brand 2	
Potential	**Value**	**Potential**	**Value**
Awareness	0.7848438	Awareness	0.7256438
Liking	0.2953387	Liking	0.1197381
Consideration	0.8292729	Consideration	0.6900375

The above results indicate that Brand 1 has a higher potential for awareness than Brand 2, with 78.5% of Brand 1's target audience having the potential to be aware of the shampoo brand compared to 72.6% for Brand 2. Additionally, Brand 1's target audience has a higher potential for liking the shampoo brand than Brand 2's target audience. Finally, a higher proportion of Brand 1's audience is considering purchasing its products than Brand 2's audience.

Stickiness

(Hanssens et al., 2014). defined stickiness as the degree to which a consumer's attitude toward a brand remains stable over time. To calculate the stickiness for each metric in the dataset, we used the "ar" function in R to calculate an autoregressive model for the awareness, liking, and consideration variables. The resulting coefficients from this function were then summed up to estimate the stickiness for each metric for the two shampoo brands. By using this method, we were able to determine how likely consumers were to maintain a positive attitude toward each brand over time.



Brand 1		Brand 2	
Stickiness	**Value**	**Stickiness**	**Value**
Awareness	0.4009+0.1912+0.1818= 0.7739	Awareness	0.1506+0.2200+0.2319= 0.6025
Liking	0	Liking	0.0196-0.0601-0.1716+0.2880= 0.0759
Consideration	0.0601+0.2901= 0.3502	Consideration	0

Brand 1 has higher stickiness value for awareness and consideration compared to Brand 2, indicating that their advertising and promotional efforts are more effective and likely to have a more sustaining impact. Brand 2 has a positive stickiness value for liking, indicating that their efforts are more effective in creating positive attitudes towards their product and that this effect is likely to have some sustaining power. Overall, Brand 1's efforts are more effective in creating and sustaining consideration among their target audience compared to Brand 2.

Responsiveness

The responsiveness of an attitude metric is the immediate change of the metric in response to a marketing stimulus. To get the responsiveness model, the log of both sides of the equation below will be taken. The resulting model will be a log-linear model.

$$Y_t = cY_{t-1}^{\gamma} X_{1t}^{\beta_1} X_{2t}^{\beta_2} e_t^{u}$$

Y represents the attitudinal metric while X_i where i= 1,2 is a marketing instrument.

Taking the log of both sides,

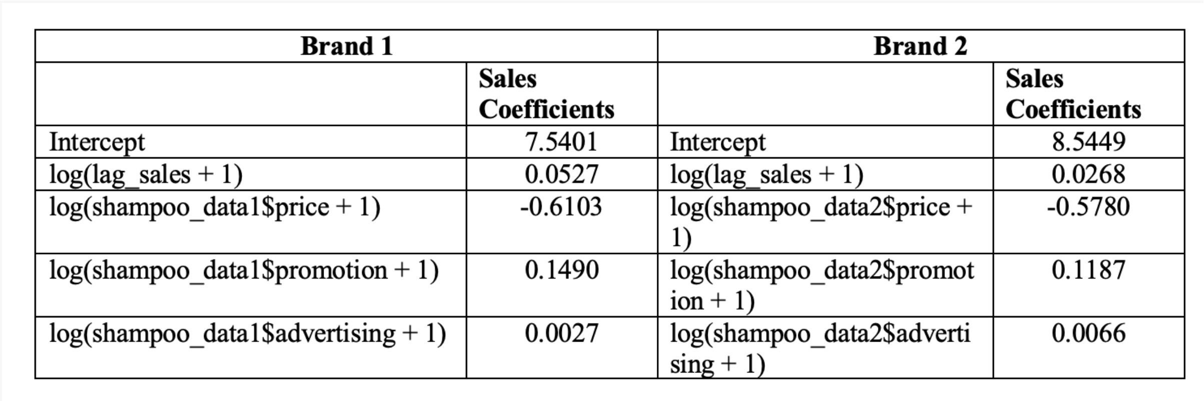
$$log(Y_t) = c' + \gamma log(Y_{t-1}) + \beta_1 log(X_{1t}) + \beta_2 log(X_{2t}) + e'_t$$

In R, this log-linear model was fitted for the mindset metrics and sales to obtain the responsiveness. See the tables below for results.

Brand 1		**Brand 2**	
	Awareness Coefficients		**Awareness Coefficients**
Intercept	3.1842	Intercept	3.6729
log(lag_aware + 1)	0.1265	log(lag_aware + 1)	0.0319
log(shampoo_data1$price + 1)	-0.5381	log(shampoo_data2$price + 1)	-0.3309
log(shampoo_data1$promotion + 1)	0.0341	log(shampoo_data2$promotion + 1)	0.0076
log(shampoo_data1$advertising + 1)	0.0056	log(shampoo_data2$advertising + 1)	0.0103

Brand 1		Brand 2	
	Consideration Coefficients		**Consideration Coefficients**
Intercept	2.7520	Intercept	3.5087
log(lag_consideration + 1)	0.0223	log(lag_consideration + 1)	0.0382
log(shampoo_data1$price + 1)	0.0152	log(shampoo_data2$price + 1)	-0.0657
log(shampoo_data1$promotion + 1)	0.0142	log(shampoo_data2$promotion + 1)	-0.0211
log(shampoo_data1$advertising + 1)	0.0030	log(shampoo_data2$advertising + 1)	0.0015

Brand 1		Brand 2	
	Liking Coefficients		**Liking Coefficients**
Intercept	1.7200	Intercept	2.0194
log(lag_liking + 1)	-0.0115	log(lag_liking + 1)	0.0014
log(shampoo_data1$price + 1)	0.0665	log(shampoo_data2$price + 1)	-0.0271
log(shampoo_data1$promotion + 1)	0.0010	log(shampoo_data2$promotion + 1)	-0.0034
log(shampoo_data1$advertising + 1)	0.0010	log(shampoo_data2$advertising + 1)	0.0004



Brand 1		Brand 2	
	Sales Coefficients		**Sales Coefficients**
Intercept	7.5401	Intercept	8.5449
log(lag_sales + 1)	0.0527	log(lag_sales + 1)	0.0268
log(shampoo_data1$price + 1)	-0.6103	log(shampoo_data2$price + 1)	-0.5780
log(shampoo_data1$promotion + 1)	0.1490	log(shampoo_data2$promotion + 1)	0.1187
log(shampoo_data1$advertising + 1)	0.0027	log(shampoo_data2$advertising + 1)	0.0066

Due to the extensive nature of the calculations and the fact that not all coefficients are essential for the analysis, the responsiveness calculation results will be summarised in the final analysis. Subsequently, the next step will be to estimate the extent to which the attitudinal metrics can translate into sales - conversion.

Conversion

To estimate the conversion rate, a multiplicative funnel model will be used to calculate the sales revenue. Then, to get the conversion model, the log of both sides of the equation will be taken. The multiplicative funnel model for sales revenue is shown below:

$$S_t = cS_{t-1}^{\gamma} A_t^{\beta_1} C_t^{\beta_2} L_t^{\beta_3} e_t^{u}$$
$$S_t = Sales\ revenue$$
$$A_t = Awareness\ Metric$$
$$L_t = Liking\ Metric$$
$$C_t = Consideration\ Metric$$

Using R, the log of both sides of the equation was taken and the results below were gotten for both brands of shampoo.

Brand 1		**Brand 2**	
	Conversion Coefficients		**Conversion Coefficients**
Intercept	6.4998	Intercept	7.6321
log(lag_sales + 1)	0.0518	log(lag_sales + 1)	0.0176
log(shampoo_data1$awareness)	0.1661	log(shampoo_data2$awareness)	0.2055
log(shampoo_data1$consideration)	-0.4244	log(shampoo_data2$consideration)	0.0673
log(shampoo_data1$liking)	0.9723	log(shampoo_data2$liking)	-0.1887

Results Summary

The table below shows a summary of the results for the first brand:

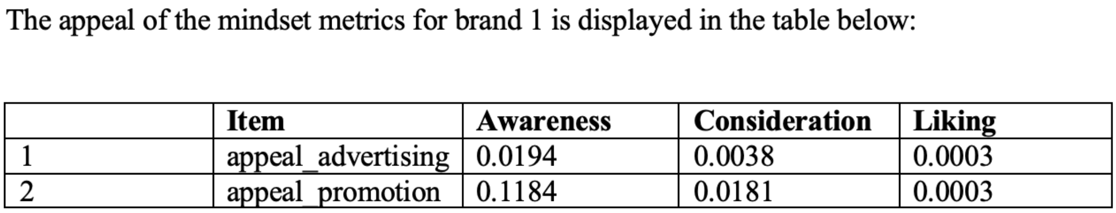
	Item	**Awareness**	**Consideration**	**Liking**
1	Potential	0.7848	0.8293	0.2953
2	Stickiness	0.7739	0.3502	0
3	Responsiveness to Advertising	0.0056	0.0030	0.0010
4	Responsiveness to Promotion	0.0341	0.0142	0.0010
5	Conversion	0.1661	-0.4244	0.9723

The table below shows a summary of the results for the second brand:

	Item	Awareness	Consideration	Liking
1	Potential	0.7256	0.6900	0.1197
2	Stickiness	0.6025	0	0.0759
3	Responsiveness to Advertising	0.0103	0.0015	0.0004
4	Responsiveness to Promotion	0.0076	-0.0211	-0.0034
5	Conversion	0.2055	0.0673	-0.1887

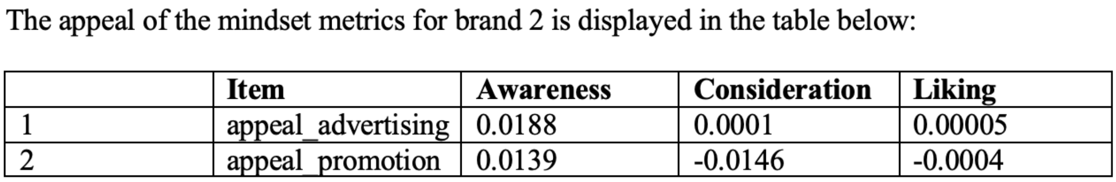
The next step will be to estimate the appeal of each mindset metric. The formula is shown below:

$$Appeal \ = Potential_k \ X \ Responsiveness_k^{(i)} \ X \ \frac{1}{(1 - Stickiness_k)}$$



The appeal of the mindset metrics for brand 1 is displayed in the table below:

	Item	Awareness	Consideration	Liking
1	appeal_advertising	0.0194	0.0038	0.0003
2	appeal_promotion	0.1184	0.0181	0.0003



The appeal of the mindset metrics for brand 2 is displayed in the table below:

	Item	Awareness	Consideration	Liking
1	appeal_advertising	0.0188	0.0001	0.00005
2	appeal_promotion	0.0139	-0.0146	-0.0004

Analysis

Brand 1 has higher "Awareness" and "Consideration" values for both "appeal_advertising" and "appeal_promotion", indicating their effectiveness in increasing brand awareness and consideration. However, both metrics have low "Liking" values, implying a weak emotional connection with the brand. On the other hand, Brand 2 has low "Awareness" and "Consideration" values for both metrics, and negative "Liking" values, suggesting a negative emotional connection. To improve Brand 1's emotional connection with the audience, the marketing strategy should focus on improving brand liking. For Brand 2, the strategy should improve brand awareness and consideration.

Now, we would like to investigate what will happen if the investment in promotion is tripled for the two different brands.

The first step is to estimate the new values of the attitudinal metrics. The formula below will be used:

$$New_{M_i} = Start_{M_i} * (Advertising_{new}/Advertising_{old})^{Responsiveness_{Promotion_{M_i}}}$$

The next step is to calculate the short run gain of each mindset metric. This is estimated by dividing the new value of the metric by the old value of the metric and subtracting the result by 1. After this, the long run gain for the metrics will be calculated using the formula below:

$$LRGain_{M_i} = Gain_{M_i}/(1 - Carryover_{M_i})$$

$Carryover_{M_i}$ represents the lagged values of the metrics that were estimated from the responsiveness model. The final step is to estimate the new conversion rate given the long run gain. The formula is as follows:

$$Conversion_{M_i} = LRGain_{M_i} * Conversion_{M_i}$$

The conversion rates for different metrics were summed up for Brand 1 (0.007 + 0 - 0.006 = 0.001) and for Brand 2 (0.002 + 0.001 - 0.002 = 0.001). These results indicate that despite tripling the investment in promotion, the sales only increased by 0.1% for both brands. In the long run, this translates to an 18.8% increase in net sales for Brand 1, which is calculated as the 18.7% increase in sales due to transactions (i.e., 18.8% - 0.1%). Similarly, Brand 2 experiences a net effect on long run sales of a 14.3% increase due to a 14.2% increase in sales due to transactions.

Marketing managers for both brands should prioritise investment in promotion for long-term net sales growth resulting from increased sales transactions. Brand 1 should focus on improving sales transactions, while Brand 2 should invest in promotion for better brand awareness and consideration. A comprehensive marketing strategy that accounts for short-term and long-term goals, and considers competition and market changes, is crucial for success.

Conclusion

In summary, this report has accomplished its primary objective of delivering sales forecasting and attitudinal metrics analysis for two shampoo brands. The sales forecasting was carried out using multiple linear regression and ARIMA models, while the analysis of attitudinal metrics was performed to evaluate the efficacy of advertising and promotional campaigns. The results of the analysis provide valuable insights that can assist these shampoo brands in planning for growth and allocating resources efficiently. To enhance the accuracy and dependability of the findings, it is recommended that further research be conducted on other forecasting and analysis techniques. In conclusion, this report provides actionable information that can inform decision-making processes for the future success of these shampoo brands.

References

1. Hanssens, D.M., Pauwels, K.H., Srinivasan, S., Vanhuele, M. and Yildirim, G. (2014). Consumer Attitude Metrics for Guiding Marketing Mix Decisions. *Marketing Science*, [online] 33(4), pp.534–550. doi:https://doi.org/10.1287/mksc.2013.0841.