A Statistical Examination of Factors Impacting Market Share in the Packaged Foods Industry

Helena Blumenau, Breck Emert & Maryann Stassen

Department of Biostatistics
University of Kansas, USA

December 10, 2023

List of Tables

List of Figures

Title

A Statistical Examination of Factors Impacting Market Share in the Packaged Foods Industry

Abstract

This study aims to determine the factors that influence market share for an individual product offered by a large packaged goods manufacturer, analyzing data from September 1999 to August 2022. Through regression modeling, the research focuses on understanding the influence of price, advertising exposure, discounts, promotions, month, and year on market share. R was selected to conduct the analysis, and after thorough data preprocessing, automatic model selection methods were used to aid the selection of the final predictive model.

Results revealed a significant positive correlation between market share and discounts and a strong negative correlation with pricing. Two time-based trends were incorporated into the model: an indicator variable for Spring and a quadratic term for the month. The final model, selected after an ANOVA review of the candidate models calculated by the Akaike Information Criterion, explained 79.5% of the variance in market share. The study concludes that market share is significantly influenced by pricing strategies, discounts, promotions, and the time of year. Future research is recommended to explore the specific effects of pricing strategies, and that expanding the dataset may give a more nuanced understanding of promotional impacts and seasonal effects. The study establishes a foundational approach for ongoing market share optimization and predictive modeling.

Introduction

In this investigation, the aim is to analyze and understand the factors influencing the market share of a specific product from a large packaged foods manufacturer. The dataset spans 36 consecutive months, from September 1999 to August 2002, and was sourced from Nielsen, a national database. The primary focus of the study is on market share, which serves as the response variable.

Various statistical methods were employed in our research, with a focus on regression modeling to quantify the relationships between the predictor variables and market share. The identification of strong correlations will guide the selection of the most influential factors for an optimized market share prediction model. The overarching objective is to identify potential weaknesses affecting revenue and explore strategies to enhance market share.

To achieve this goal, we considered the several key predictor variables:

1. **Price**: The cost of the product.
2. **Gross-Nielsen Rating Points**: A measure of the product's advertising exposure.
3. **Discount**: A binary variable indicating whether a discount was offered for the month.
4. **Promotion**: A binary variable indicating whether a promotion was offered for the month.
5. **Month**: The month in which the data was recorded
6. **Year**: The year in which the data was recorded

Through a thorough analysis of the relationships between these predictor variables and the response variable (market share), our objective is to pinpoint which factors exhibit the strongest correlations. This process is integral to developing a robust and predictive model that can be leveraged for decision-making. This approach will enable insights into potential weaknesses in revenue and illuminate strategies to bolster market share.

Primary Analysis Objectives

To investigate the association between market share and key predictor variables, including price, discount, gnrpoints, month, year, and promotion. Among these predictors, the objective of this study is to identify which variables are related to market share.

Materials and Methods

Data Collection

The dataset utilized in this research was sourced from Nielsen, a comprehensive national database. The dataset spans a duration of 36 consecutive months, from September 1999 to August 2002. Each entry in the dataset includes a unique identification number and information on the six variables for each month.

Data Preprocessing

An examination of the dataset was undertaken to identify essential preprocessing steps. This process involved a thorough assessment of data quality, including the identification of missing or duplicate values, as well as verification of data types to ensure R loaded the data types correctly. We found the dataset to be well-structured and did not require any alterations to the values.

Final Dataset

The names of the variables will hereby be referred to by their name in the dataset, for consistency, as visible in the data table below.

Table 1: Final Dataset

Variable Name	Data Type	Data Format	Description	Example
marketshare	Percentage	1.23	Average monthly market share for product (percent)	3.15
price	Numeric	1.23	Average monthly price of product (dollars)	2.19
gnrpoints	Numeric	123	Gross Nielsen Rating Points, an index of product exposure	498
discount	Factor	{1,0}	Presence of absence of a discount price offered during the month	1
promotion	Factor	{1,0}	Presence of absence of package promotion during the month	1
month	Character	tttt	Month (January - December)	January
date	Numeric	1.23	The date, normalized to 0 as the first date and increased 1 per year (months increase by 0.083)	1.50

Statistical Analysis

The data is available in .xlsx (Excel) format. The data analysis is done using the statistical software R, and the project focuses on multiple linear regression. Each of the predictor variables' distributions was explored individually (Appendix A), and the absence of missing values was confirmed. Automatic model selection methods were used to arrive at the final model. The model assumptions were assessed and confirmed, ending with a suggestion on the final predictive model for market share.

Primary Objective Analysis

Exploring individual predictors and the response variable is crucial for ensuring model assumptions later on. This exploration can assist in identifying any potential outliers or skewness in the data. If, at a later stage, we discover that the model does not fit the data well, it can serve as a reference point for determining where transformations may be necessary to create a more suitable model. After exploring the dataset variables, the subsequent step involves checking the linearity between predictors and the response variable, as well as examining the relationships between predictor variables to identify potential issues with multicollinearity. This step is important for developing a model that can provide accurate predictions. The methodology for the analysis involves modeling the relationship between the four predictor variables and marketshare. All statistical tests will be conducted with a significance level of 0.05, considering p-values less than 0.05 as statistically significant.

Results

Exploratory Analysis

Summary Statistics

To initiate the analysis, a preliminary exploration of the dataset was conducted using both graphical and descriptive methods. Descriptive statistics such as mean, median, and standard deviation were examined for each variable to gain insights into central tendencies and variations in the data. The average market share across the dataset was 2.66% and ranges from 2.23% to 3.16%. The product's price averages $2.32 and ranges from $2.12 to $2.78. Gross-Nielsen Rating Points (gnrpoints) has a mean of 388.06 with a range from 72 to 858. Discount and promotion have an average of 0.58 and 0.56 and a slight negative skewness, meaning that discounts and promotions are offered slightly more than they are not.

Skewness for the price was notably high at 0.95, indicating a right-skewed distribution. In contrast, market share and gnrpoints show less skewness with values of 0.27 and 0.26, respectively. The binary variables discount and promotion show negative skewness at -0.34 and -0.22, respectively, indicating that discounts and promotions are offered slightly more often than they are not. This serves as a descriptive statistic, but does not necessarily tell us the shape of the distribution (Table 2).

Table 2: Summary Statistics

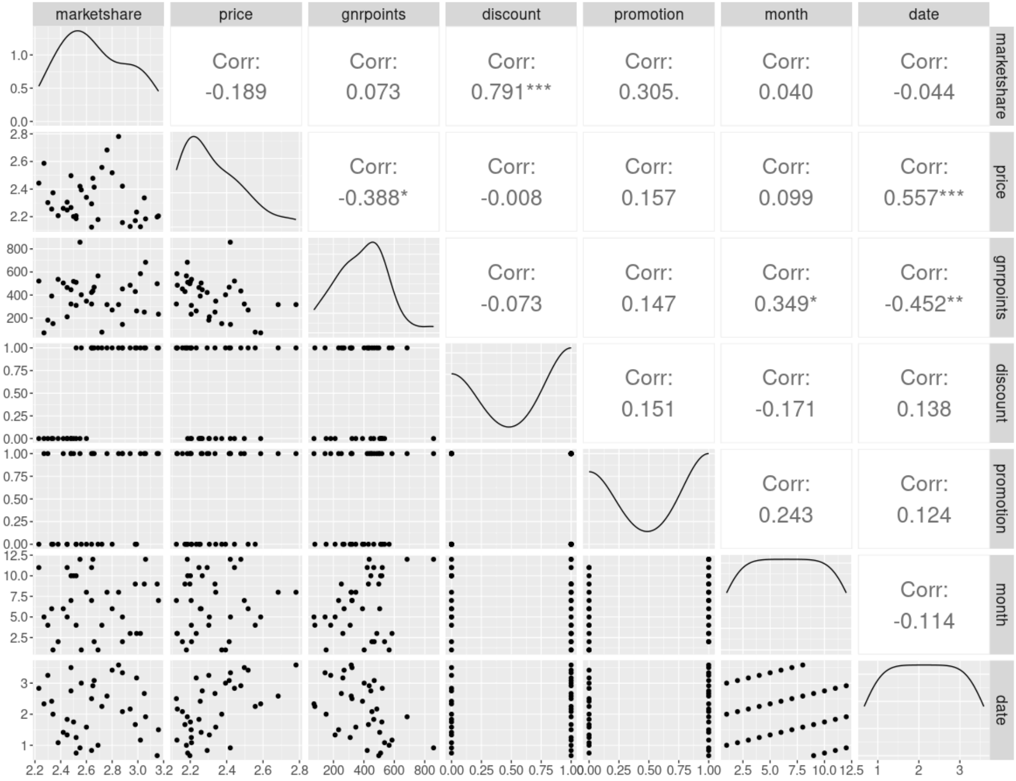
	Mean	Median	SD	Min	Max	Skewness
marketshare	2.66	2.65	0.26	2.23	3.16	0.27
price	2.32	2.28	0.16	2.12	2.78	0.95
gnrpoints	388.06	412	168.49	72	858	0.26
discount	0.58	1	0.5	0	1	-0.34
promotion	0.56	1	0.5	0	1	-0.22

Data Visualization

Graphical methods, including histograms and scatter plots, were used to assist in visualizing the distributions of the variables for unexpected patterns or trends in the data. Because each variable is manually chosen by the company, we do not expect or require specific distributions but instead aim to find values in ranges consistent for the context of each variable. Our findings revealed a relatively normal distribution for gnrpoints and the binary variables discount and promotion showed a balanced representation of both categories. The distribution of market share and of the predictor variables, price, gnrpoints, discount, and promotion can be seen visually in Appendix A: Figure A1 and Figure A2.

The bivariate distributions of each variable were examined in Figure 1. The scatterplot matrix indicates the absence of non-linear relationships concerning 'marketshare', the response variable (Figure 1). These scatterplots serve as a valuable reference to understand the relationships between the predictor variables with one another, as well as with the response variable. These scatterplots should be referenced to help understand the relationships mentioned further in the analysis, with detailed and full-scale scatterplots of marketshare's relationships in Appendix B.

Figure 1: Scatterplot Matrix



Primary Objective Analysis

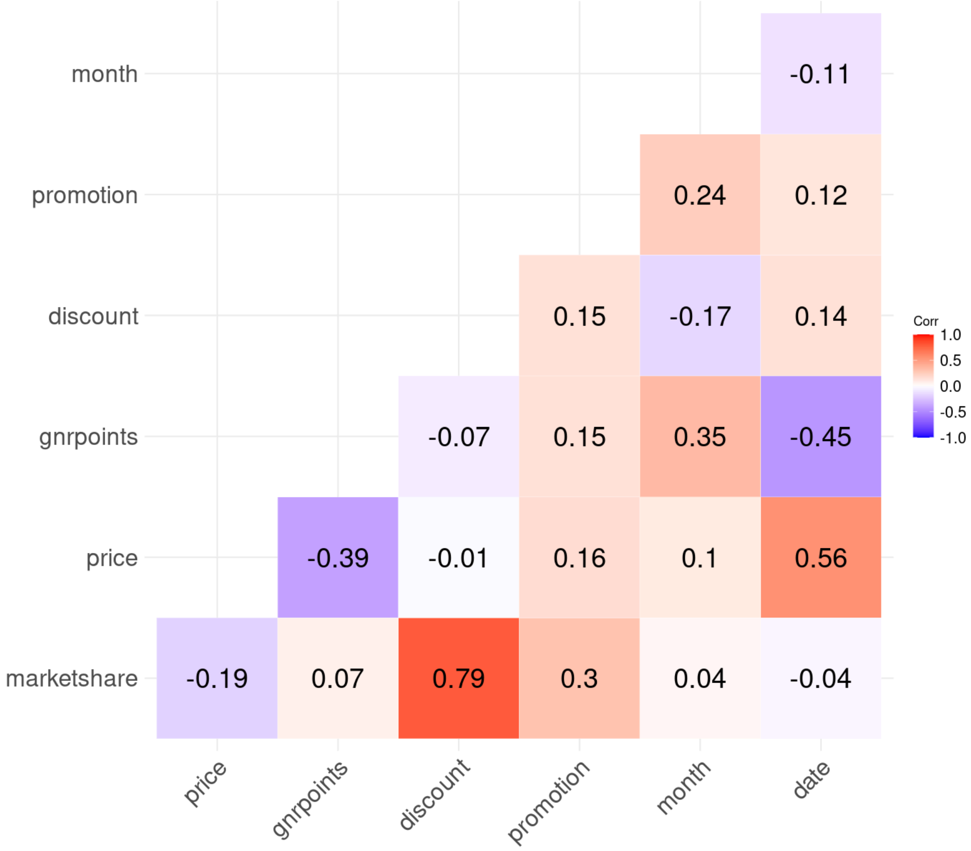
Correlation Analysis

The bivariate relationships, which were examined visually with a scatterplot matrix in Figure 1, are also examined through correlation coefficients. These were calculated to assess the relationships between the previously identified predictor variables and the response variable. The analysis of correlations between various predictor variables and the response variable are shown below in Figure 2.

A strong positive linear relationship was denoted by a correlation coefficient larger than 0.5. The largest association is seen between marketshare and the discount, observing a correlation coefficient of 0.79. A weak positive correlation of 0.3 is observed between marketshare and promotions. Conversely, a weak negative linear relationship is seen between marketshare and pricing, with a correlation of -0.19. A very weak positive linear relationship is seen between marketshare and gnrpoints, with a correlation of 0.07.

In this same figure, we also analyzed the relationship between the predictor variables. There is a moderate negative linear association seen between gnrpoints and price, with a correlation coefficient of -0.39. Beyond this, all associations have a correlation coefficient less than 0.2 (Figure 3). Issues relating to potential multicollinearity are addressed later on utilizing the variance inflation factor (Table 4).

Figure 2: Correlation Matrix for the Complete Dataset



Time-Based Trend Analysis

We investigated the potential impact of previous time points on marketshare through a lag-based analysis, predicated on the hypothesis that the predictor variables could have delayed effects on consumer behavior and market dynamics. However, our analysis did not uncover significant lagged or autocorrelated relationships, suggesting that marketshare for this product is influenced more by immediate factors, or that our dataset's scope is not wide enough to capture trends that cross multiple years. (Appendix D).

The coefficients of correlation of a basic linear model vary when split up seasonally, compared to the overall values. Table 3 presents correlation coefficients between various predictor variables across different seasons: Fall, Spring, Summer, and Winter. The correlation coefficients quantify the strength and direction of the linear relationships. Price exhibits negative correlations in the fall, summer, and winter, suggesting a tendency for lower prices to coincide with these seasons.

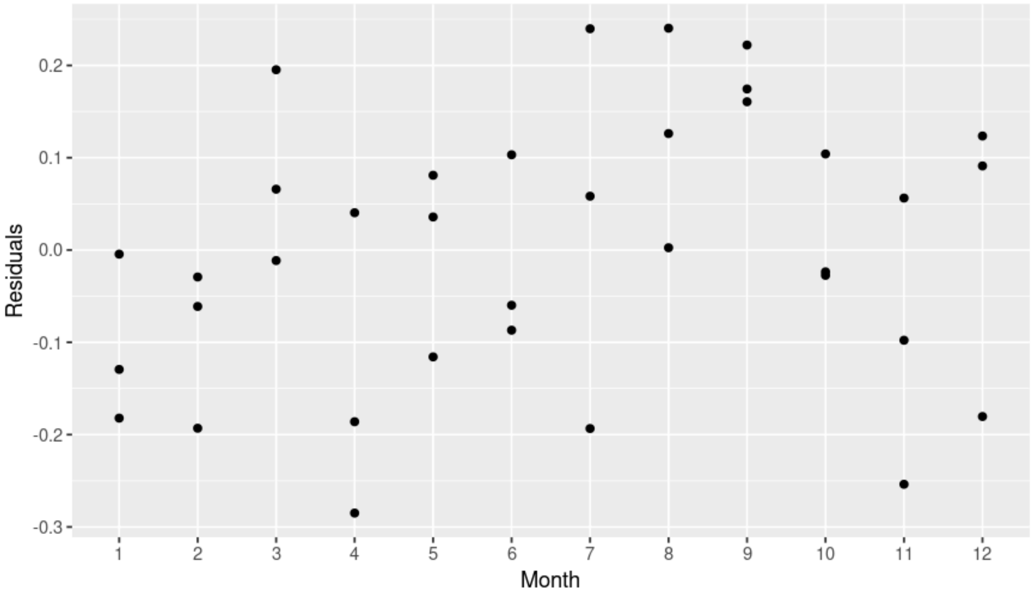
Gnrpoints shows varying correlations across seasons, with a positive correlation in the fall and a negative correlation in the spring. Discount displays strong, positive correlations in all four seasons, with the highest observed in the spring. Promotion indicates positive correlations across all seasons, with the strongest correlation in the spring (Table 3).

Table 3: Seasonal Correlation Coefficients for Predictor Variables

Season	price	gnrpoints	discount	promotion
Fall	-0.350	0.293	0.737	0.554
Spring	0.365	-0.217	0.875	-0.225
Summer	-0.324	0.119	0.712	0.339
Winter	-0.611	0.287	0.790	0.581

Additionally, when plotting the residuals of a simple linear model that includes the four non-temporal predictors, a quadratic trend emerges with respect to the month, characterized by a downward facing parabola. The parabola undergoes inversion in September, indicating a potential distinct trend (Figure 3). To validate this observation, separate variables can be tested to account for these two distinct trends. the parabola inverts in September, we can see this is possibly a separate trend, which can be verified by testing separate variables to account for these two observations.

Figure 3: Scatterplot of Model Residuals vs. Month



In response to the identified non-linearity, we proposed two potential variables. The initial variable, denoted as ‘isSpring’, serves as an indicator, assuming a value of True during the Spring season and False otherwise. The second variable introduces a quadratic predictor, integrated into the linear model as two variables: ‘month’ and ‘month_sq’. By incorporating the regression of both the month and its squared value, the model is better equipped to account for the inverted parabolic trend seen in Figure 3. Both variables underwent assessment in model selection, alongside other factors, before their final inclusion in the model.

Variable Selection

As our model incorporated multiple predictor variables, potential multicollinearity issues were addressed. Examination of the correlation matrix revealed the relationships among these predictors (Figure 2). Including variables that exhibit significant correlation can increase the variance of computed coefficients and limits their ability to accurately represent the true, underlying pattern (Alin, 2010). The basic correlation matrix indicates a low level of multicollinearity, with only price and date showing an absolute correlation above .5. The other strongest relationship is between gnrpoints and price, with a coefficient of -0.39 (Figure 2).

To examine the effects of potential multicollinearity, we employed the variance inflation factor (VIF) to quantify this phenomenon for each variable within the model (Kim, 2019). Table 4 shows that variance inflation factors for price, discount, and promotion are near 1, and isSpring has a moderate VIF of 6.27 due to being a transformation of month. The large VIF for month and month_sq was expected as month_sq is a linear transformation of month, and does not affect the interpretability of the model. A general rule of thumb is that VIF values greater than 10 indicate a problem with multicollinearity, and each variable shows a large potential for explanatory power of unique trends, so no variables were removed from model selection. These findings on multicollinearity were considered during model selection as it may affect the overall interpretability of the final regression coefficients.

Table 4: Variance Inflation Factors

price	discount	promotion	month	month_sq	isSpring
1.19	1.11	1.18	67.63	46.30	6.27

Model Selection

Model Choice

Multiple linear regression was selected as the primary method due to its capability to handle multiple predictor variables simultaneously, and was based on the assumption that marketshare is influenced linearly by pricing, advertising reach, promotions, and discounts. All relationships input to the model have been verified to be linear, or transformed to fit into the linear model. The regression analysis was conducted to quantify the impact of each predictor variable on marketshare and provide coefficients that indicate the strength and direction of these relationships.

The initial analysis establishes an optimal model based on the Akaike Information Criterion (AIC), a tool that assesses the predictive model's quality by striking a balance between its goodness of fit and simplicity (Vrieze, 2012). While alternative methods, such as Mallow's C_p, were employed for validation (refer to Appendix C), they are less favored for smaller sample sizes (Miyashiro & Takano, 2015). AIC results are interpreted by considering its calculation using the following formula:

$$AIC = -2k -2 \log(\hat{L})$$

Where:

- ***k*** is the number of estimated parameters in the model
- $\hat{L}$ is the maximized value of the likelihood function for the estimated model

Lower AIC values indicate better efficiency, suggesting that the model's fit compensates for the number of included variables.

Table 5: Model Selection based on AIC

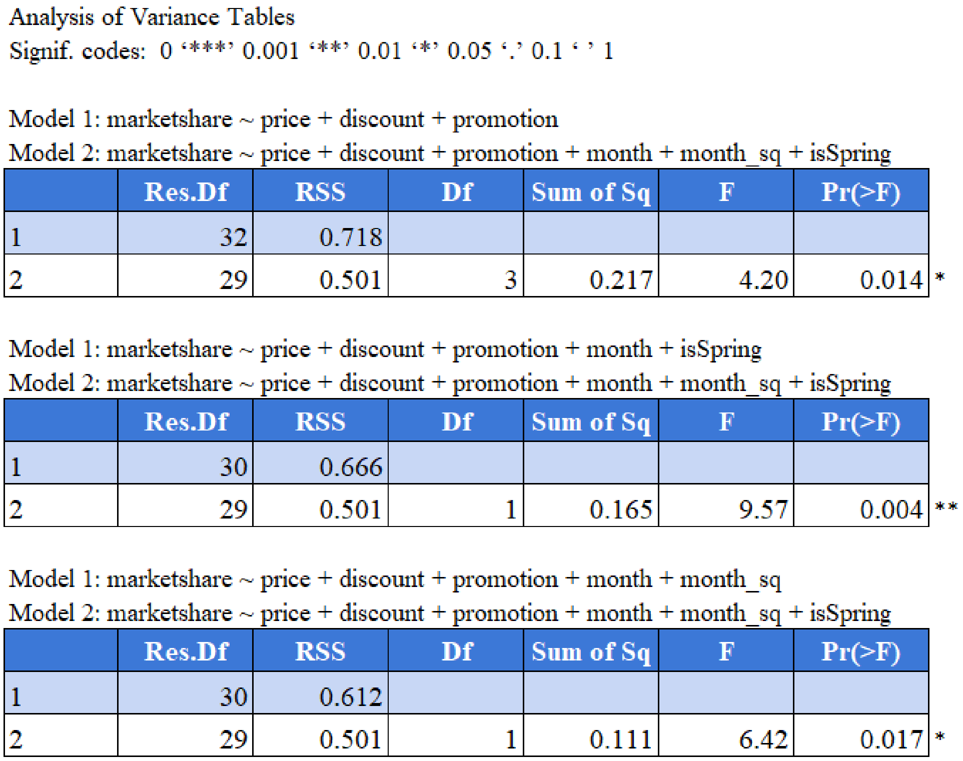
Included Variables						Size	AIC
price	discount	promotion	month	month_sq	isSpring	6	-35.8
price	discount	month	month_sq	isSpring		5	-35.3
discount	month	month_sq	isSpring			4	-34.4
price	discount	promotion	gnrpoints	month	month_sq	7	-33.9
discount	promotion	month	month_sq	isSpring		5	-33.7

Based on the AIC criteria, the model with the lowest value includes price, discount, promotion, month, month squared, and isSpring, with an AIC value of -35.8 (Table 5). The other automatic model selection techniques confirmed this choice (Appendix C).

We verified that our model was not only the best explanation of the trends in the sampled dataset, but also a good predictor, through a Leave-One-Out Cross-Validation (LOOCV) approach. This method was selected for its full utilization of the dataset, only leaving out one row per validation step, due the limited sample size of the dataset. Automatic feature selection was performed to find the model that minimizes the reduction in R-squared during LOOCV. The candidate model previously selected by AIC, with 6 predictor variables, also demonstrated the lowest Mean Absolute Error (MAE). The model's MAE of 0.120 indicates an average reduction of 12% from the observed marketshare values in out-of-sample predictions, lowering the R-squared to 0.690. In addition, it also held the largest overall R-squared value.

The ANOVA test, by assessing the increase in R-squared values relative to the loss of residual degrees of freedom, confirms that this model significantly enhances the fit over simpler models that were excluded based on previous criteria. The F-values, measuring the ratio of explained to unexplained variance, identify the variable's impact on marketshare.

Table 6: ANOVA Model Comparison Table



Analysis of Variance Tables
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Model 1: marketshare ~ price + discount + promotion
Model 2: marketshare ~ price + discount + promotion + month + month_sq + isSpring

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)	
1	32	0.718					
2	29	0.501	3	0.217	4.20	0.014	*

Model 1: marketshare ~ price + discount + promotion + month + isSpring
Model 2: marketshare ~ price + discount + promotion + month + month_sq + isSpring

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)	
1	30	0.666					
2	29	0.501	1	0.165	9.57	0.004	**

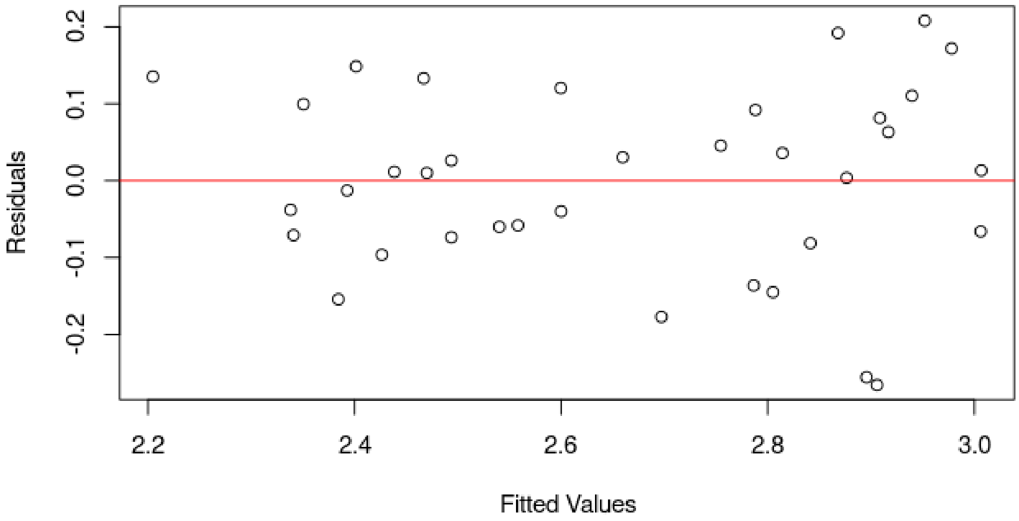
Model 1: marketshare ~ price + discount + promotion + month + month_sq
Model 2: marketshare ~ price + discount + promotion + month + month_sq + isSpring

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)	
1	30	0.612					
2	29	0.501	1	0.111	6.42	0.017	*

Residual Analysis

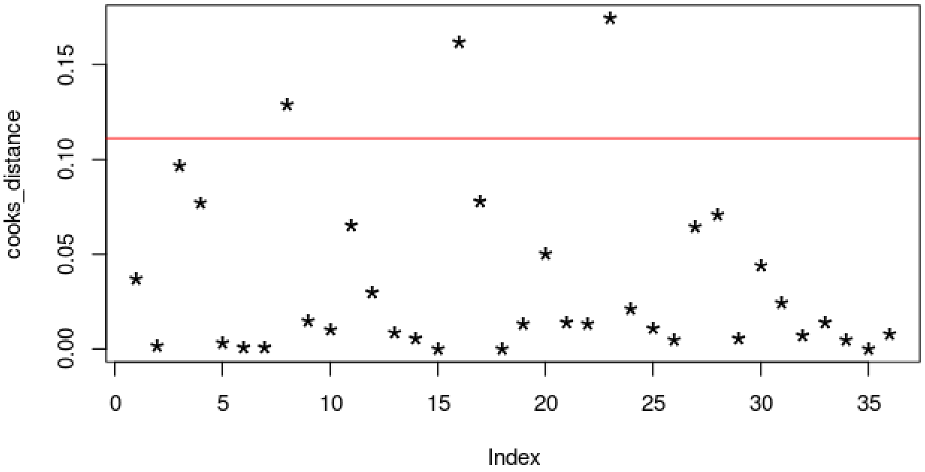
A residual versus fitted value plot was examined to assess homoscedasticity and identify potential systematic variance in the residuals. The plot did not exhibit patterns of heteroscedasticity, which supports our assumption that the residuals have constant variance across the range of fitted values. We also observed a scarcity of estimated points around a marketshare of 2.7; the scatterplot of the input data reveals noticeable gaps in the data around this marketshare value. Additionally, there are two distinct outliers around a residual of -0.3, which could potentially influence the model's estimates (Figure 4).

Figure 4: Residuals vs. Fitted Values



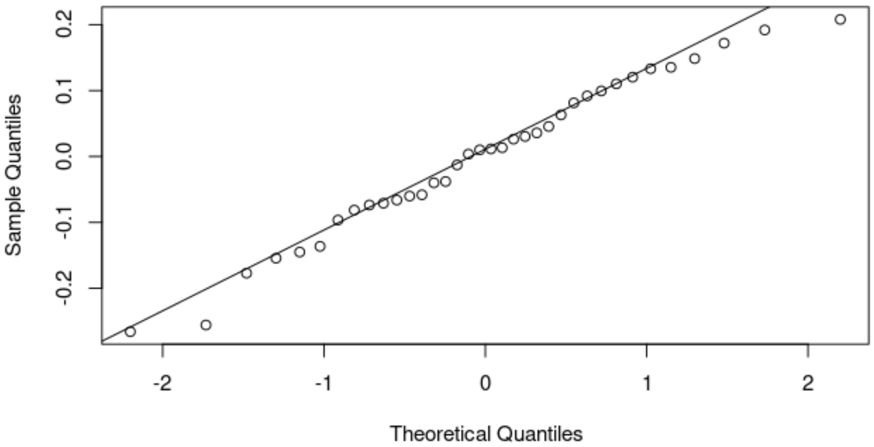
To examine these residual outliers for high leverage, we applied Cook's Distance for leverage calculation. Figure 5 shows the Cook's Distance leverage with a common threshold of 4/n to flag potential concerns. Notably, the points with high leverage do not align with the largest residuals in the model. An examination of the data that produced the high leverage points revealed no abnormalities.

Figure 5: Cook's Distance for Each Index



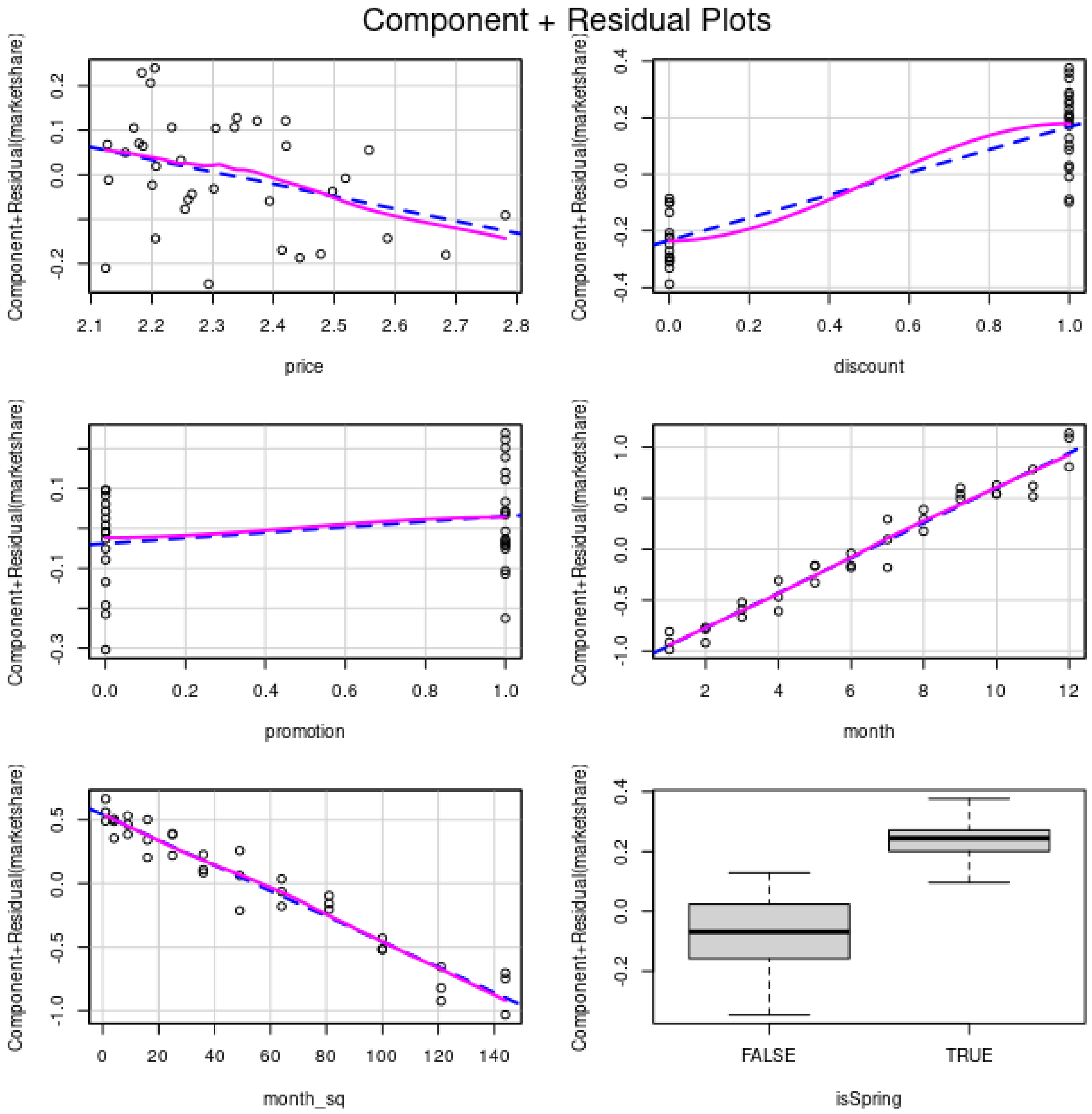
The linear model's assumptions are homoscedasticity, constant variance of the residuals, and that the data is best explained by the linear model. Each of these assumptions can be tested visually and statistically. First, the Quantile-Quantile plot shows the residuals vs. the expected values under normality. Deviations from the plotted straight line can indicate non-random error in the residuals. The plot of our model's residuals vs. the theoretical quantiles of residuals follows the line well, but has two outliers in the tails (Figure 6). The normality assumption for the residuals of our linear model was further evaluated using the Shapiro-Wilk test. The test yielded a statistic W = 0.978 with an associated p-value of 0.689. The statistically significant value near W = 1 does not provide sufficient evidence to conclude that the residuals are not normally distributed.

Figure 6: Normal Q-Q Plot



The component residual plots visualize the linearity of the trend when each variable is left out of model-building. If the variable is a good choice for the linear model, we expect to find a linear trend in the residuals. Every variable has a clear linear trend, verifying the linear assumption further and identifying the need to include the variable (Figure 7).

Figure 7: Component Residual Plot



Final Model

The final regression model is summarized in table 7. The coefficients presented in the table quantify the impact of each predictor variable on the response variable. The statistical significance of each coefficient is important for determining the reliability of $\hat{Y}_i$. Additionally, measures such as the R-squared value, 0.795, tell us that 79.5 % of the sample variance in marketshare is accounted for by the model.

The estimated regression function from this data analysis is:

$$\hat{Y}_i = 2.380 - 0.278X_1 + 0.4014X_2 + 0.0686X_3 + 0.172X_4 - 0.009X_5 + 0.321X_6$$

where,

$\hat{Y}_i$ is the predicted marketshare

X_1 is the price

X_2 is the presence or absence of a discount

X_3 is the presence or absence of a promotion

X_4 is the month

X_5 is the month squared

X_6 is Spring

Table 7: Statistics Table for the Regression Model

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	2.380	0.433	5.497	0.0000
Price	-0.278	0.149	-1.867	0.0719
Discount	0.4014	0.047	8.590	0.0000
Promotion	0.0686	0.0479	1.434	0.1624
Month	0.172	0.052	3.297	0.0026
Month_sq	-0.009	0.0032	-3.093	0.0043
isSpring	0.321	0.127	2.535	0.017
Observations	36			
R^2	0.7953			
Adjusted R^2	0.753			
Residual Std. Error	0.1314 (df = 29)			
F Statistic	18.78 (df = 6, 29)			

The ANOVA table (Table 8) assesses the overall validity of our regression model, which ultimately includes six predictor variables: price, discount, promotion, month, month_sq (quadratic month), and isSpring (spring indicator). The predictor’s price, discount, and promotion show significant contributions to the model’s improvement, as indicated by their respective F-values (5.034, 88.2489 and 6.8299), and associated p-values (Table 8). These results suggest that managing pricing strategies, discounts, and promotions significantly influence marketshare. While the month predictor does not exhibit a significant impact, the quadratic term month_sq and the isSpring indicator also contribute meaningfully to the overall fit of the model.

Table 8: ANOVA Table for the Regression Model

	Df	Sum Sq	Mean Sq	F Value	Pr(>F)
Price	1	0.08693	0.08693	5.034	0.0326
Discount	1	1.523	1.523	88.2489	0.0000
Promotion	1	0.11791	0.11791	6.8299	0.01406
Month	1	0.05072	0.05072	2.9378	0.09720
Month_sq	1	0.05570	0.05570	3.2265	0.08288
isSpring	1	0.11090	0.11090	6.4242	0.01691
Residuals	29	0.500	0.01726		

Discussion & Conclusions

We examined four predictor variables in relation to market share: price, discount, promotion, and gnrpoints. All statistical analysis was conducted with a 95% confidence level and a significance threshold of α = 0.05. The predictor variables were explored individually, as well as with one another. Table 6 shows the estimated regression coefficient, the standard error, t-value, p-value associated with each of the predictors, R^2, R^2_a, residual standard error, and F statistics of the final model. Table 7 shows the ANOVA table for the final model F values and the corresponding p-values.

Exploring various model selection criteria, a final model was chosen based on the Akaike Information Criterion, but other methods, such as Mallows' Cp and Bayesian Information Criterion confirm that our final model was the most optimal for influencing market share. This model is one that includes price, discount, promotion, month, month_sq (quadratic month), and isSpring (spring indicator). After checking model assumptions, homoscedasticity, linearity, and normality, as well as confirming there were no issues with multicollinearity, the model explained 79.5% of the variation in market share. Because of this, it can be concluded that the model did a fair job at predicting market share.

However, there is still ~20% still left unexplained by the current model. One proposed future study would involve isolating the positive trend associated with discount. This would mean investigating whether the observed trend is influenced by the lowering price or the availability of discounts. A practical test could involve manipulating pricing strategies to observe the corresponding impact on discounts. Additionally, we could also acquire a more extensive dataset to verify the trend related to the promotion variable. This variable is currently included in the model without achieving a 95% confidence level. Future studies can focus on refining estimates for seasonal-specific regressions, offering a more specific understanding of how each season uniquely contributes to market share dynamics. A comprehensive analysis for each season could provide insights that might remain clouded by our current model. These steps together may ultimately improve the model's explanatory power, and provide a more nuanced perspective on market share overall.

Examining the correlation matrix (Figure 2) reminds us that individual variables may not be the most informative, and may need potential interactions or transformations to be more well understood. This is most evident in the representation of the 'date' variable, which currently fails to reveal any seasonal trends. To address this limitation, we propose a transformation of the data to be organized by year, as demonstrated in Appendix D, allowing us to unveil valuable insights limited by the current date variable.

Based on the results in Appendix D, we propose three distinct improvements aimed at maximizing market share with minimal cost:

1) Leveraging the lack of negative trend of market share with price during spring.
2) Reduce price during the summer months
3) Strategically time promotional activities

In summary, this project included a comprehensive exploration of factors influencing market share, employing a range of statistical techniques and analyses. Initial observations from the correlation matrix showed the need for further modeling strategies. Using the Akaike Information Criterion and the Analysis of Variance, we were able to better select a model, resulting in a final regression model with six predictor variables, including quadratic month variables and an indicator variable for spring. The ANOVA results affirmed the significance of the predictors included in the final model, accounting for most of the variance in market share. However, there is still approximately 20% of the variance that remains unexplained, suggesting avenues for future research, including isolating trends associated with discounts and refining the understanding of promotional impacts with a larger dataset. The proposal to transform data to highlight seasonal trends demonstrates a commitment to continuous improvement and understanding of the complicated dynamics related to market share. Overall, this project not only provides actionable insights for market share optimization, but also establishes a foundation for ongoing research and refinement in predictive modeling.

References

Alin, A. (2010). Multicollinearity. WIREs Computational Statistics, 2(3), 370–374. https://doi.org/10.1002/wics.84

Kim, J. H. (2019). Multicollinearity and misleading statistical results. *Korean Journal of Anesthesiology*, *72*(6), 558–569. https://doi.org/10.4097/kja.19087

Miyashiro, R., &amp; Takano, Y. (2015). Subset selection by Mallows' . Expert Systems with Applications, 42(1), 325–331. https://doi.org/10.1016/j.eswa.2014.07.056

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike Information Criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods*, *17*(2), 228–243. https://doi.org/10.1037/a0027127

Appendix

Appendix A: Descriptive Statistics

Graphical methods, including histograms and scatter plots, were used to assist in visualizing the distributions of the variables, and identify any potential patterns or trends in the data that would need to be addressed. Initial examination of market share indicates a slight positive skewness, as the whiskers to the right are slightly longer than the ones on the left. This can be seen in Figure A1(a,b). There are no identified outliers. Preliminary data analysis on the four predictor variables are shown below. The distribution of prices exhibits a slight skewness, with a right tail longer than the left, suggesting an asymmetry in figure A2 (a, b). Additionally, the presence of an outlier in the price data points requires further investigation to understand its impact and potential reasons for its extreme value. The gnrpoints variable displays an overall normal distribution, except for one potential outlier, seen in figure A2 (c, d). Regarding the binary variables, both discount and promotion show a slight imbalance, with slightly more occurrences than absences, as seen in Figure A2 (e, f).

Figure A1: Distribution of Market Share

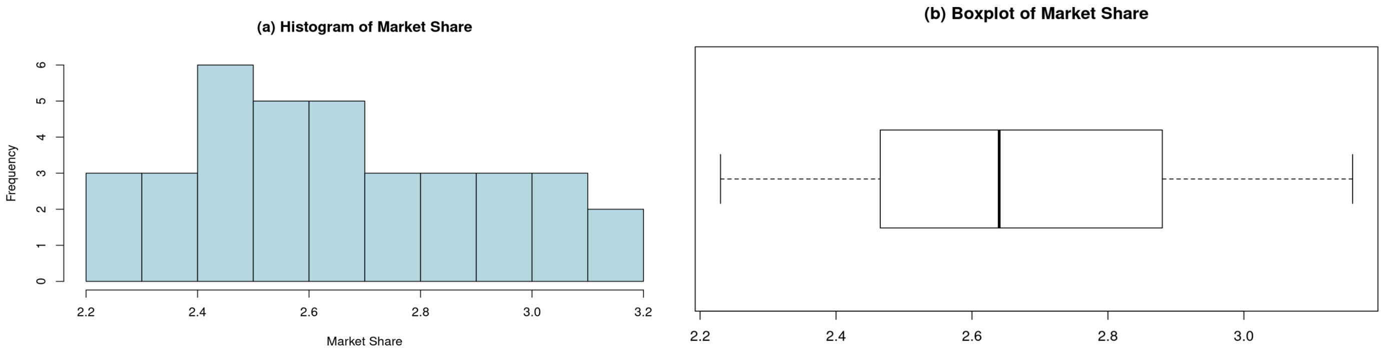
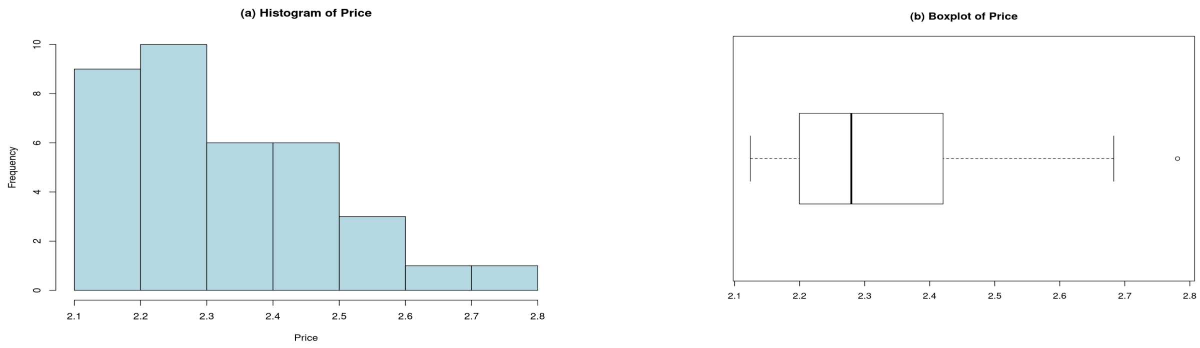
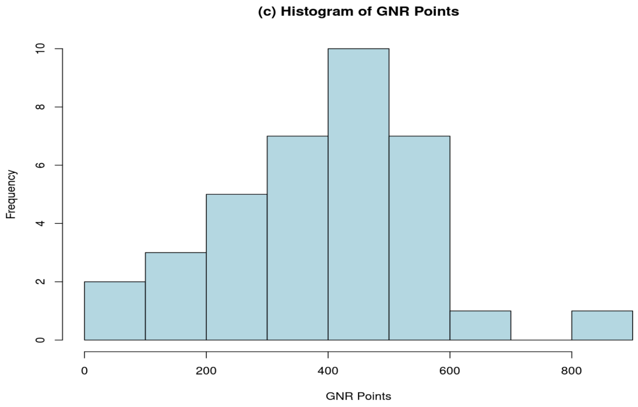
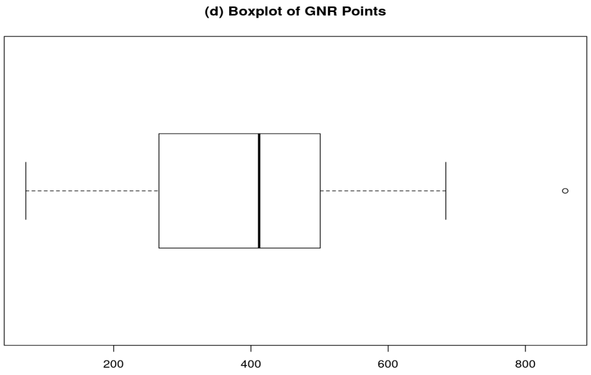
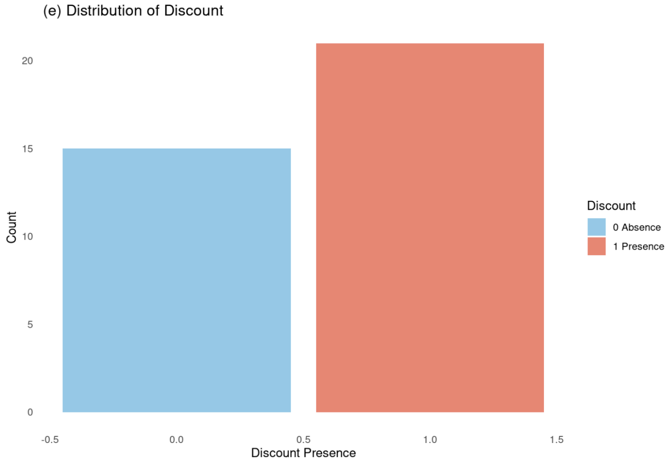


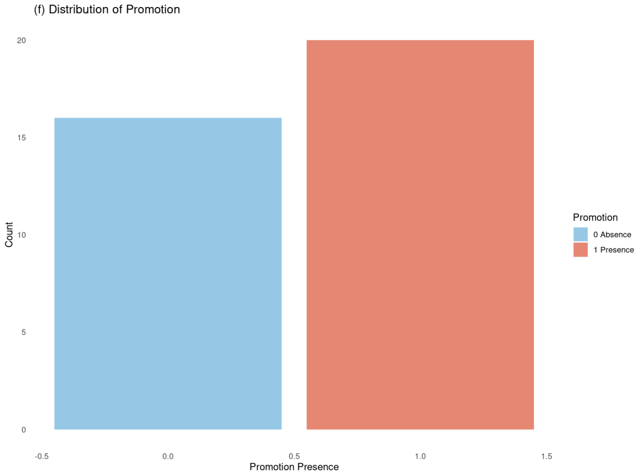
Figure A2: Distribution of predictor variables




(c) Histogram of GNR Points


(d) Boxplot of GNR Points


(e) Distribution of Discount


(f) Distribution of Promotion

Appendix B: Market Share vs. Predictor Variables

The analysis of the relationship between marketshare and the four predictor variables are shown below. Each data point represents an observation in the dataset. Market Share and price have a slight negative linear association, as seen in Figure B1. Marketshare and grnpoints, marketshare and promotion, and marketshare and discount, all have positive linear associations, seen in Figures B2, B3 and B4. Market Share and Date appear to have no linear association, seen in Figure B5.

Figure B1: Scatter Plot of Market Share vs. Price

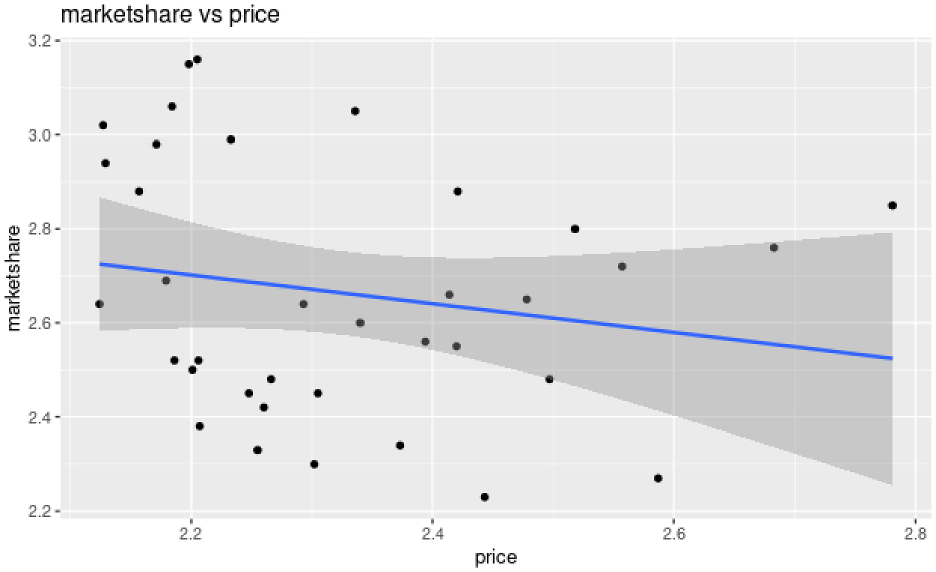


Figure B2: Scatter Plot of Market Share vs. GNRPoints

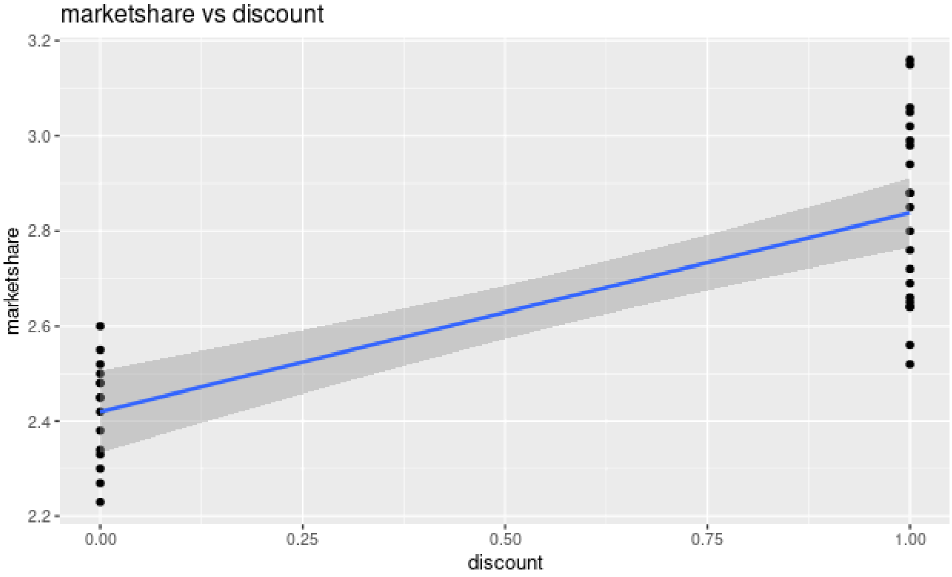


Figure B3: Scatter Plot of Market Share vs. Promotion

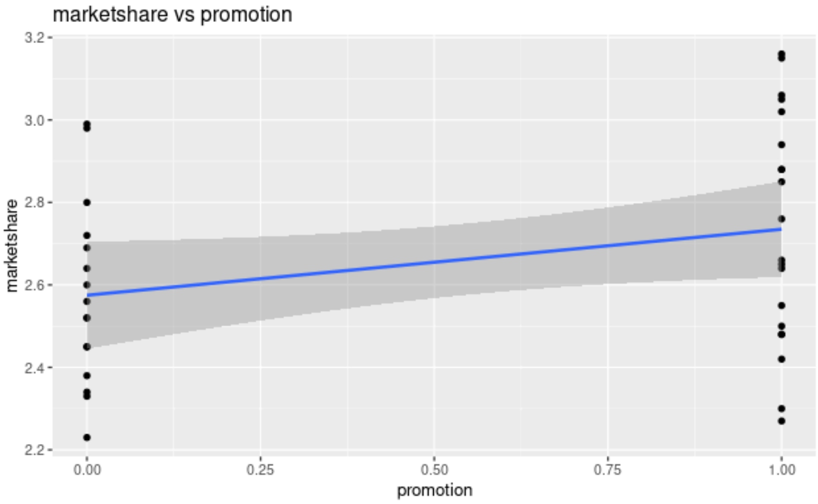


Figure B4: Scatterplot of Market Share vs. Discount

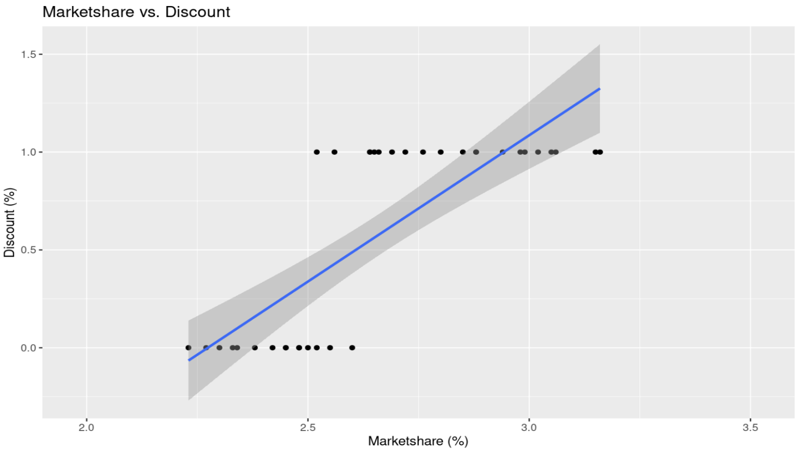
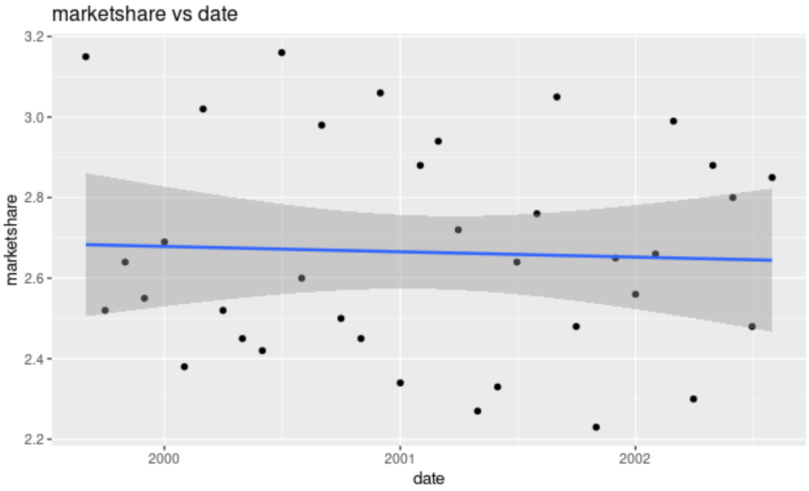


Figure B5: Scatter Plot of Market Share vs. Date



Appendix C: Model Selection Criteria

Additional model selection criteria were employed to validate and confirm the outcomes derived from the Akaike Information Criterion (AIC). Specifically, the results from tests including residual sum of squares (rss), adjusted R^2 and Mallows' Cp consistently favored model 6, confirming the selection based on AIC.

In the R^2 test, the values for models 1 - 6 were as follows: (0.6252543, 0.6604497, 0.7065091, 0.7622120 0.7808426, 0.7953459). We can see that this test favors model 5, as it has the largest R^2 value. Model 5 includes price, discount, month, month squared and isSpring. When looking at the Bayesian Information Criterion, the values for models 1 - 6 were as follows: (-28.16723, -28.13423, -29.79864, -33.79193, -33.14562, -32.02700). Again, we opt for the model with the smallest value, which in this case is model 5.

The other three tests used to verify model selection were consistent with what was identified in Table 5 using the AIC. The residual sum of squares for models 1 - 6 were as follows: (0.9167238, 0.8306269, 0.7179536, 0.5816903, 0.5361151, 0.5006361). For this test, we opt for the model with the lowest rss value, which we can see again is model 6, indicating that a model with price, discount, promotion, month, month squared and isSpring is considered optimal.

In the adjusted R^2 test, the values for models 1- 6 were as follows: (0.6142323, 0.6398709, 0.6789944, 0.7315296, 0.7443163, 0.7530037). We can see when opting for the model with the largest value, we would again favor model 6.

In the Mallows' Cp test, the values for model 1 - 6 were as follows: (21.102421, 18.115145, 13.588400, 7.695168, 7.055166, 7.000000). When using this criterion, we opt for the smallest Cp value, which in this case is model 6.

The AIC values and model selection can be seen in Table 5, which confirm that model 3 should be favored. This indicates that a model with price, discount, promotion, month, month squared and isSpring, is considered optimal.

Appendix D: Time-Based Analysis

Having modeled market share in response to various predictors, our focus shifts to understanding how the business can effectively manage these relationships. Based on the model found in this analysis, the marketshare was not distinctly maximized during any part of any year. Figure D1 highlights that the expected marketshare was maximized for all years during March and September, with a generally random trend observed. This prompts further exploration into the underlying dynamics.

The figures below also show the choices of values for the predictors made during each year. Figure D2 shows a line for the price set each year, with darker lines representing later years. Note that 1999 does not start until September. Price is maximized during spring to summer months, with two outliers in June and July of 2001. Because marketshare is negatively correlated with price at -0.19, we expect marketshare to be negatively impacted by higher pricing during these months (Figure D2). The time-trend for ad exposure follows a relatively inverse trend, peaking in Fall and Winter (Figure D3). To visualize the categorical variables ‘promotion’ and ‘discount’, the values were instead summed across each month, to get an understanding of the choices made relative to the time of year. The graph shows that discounts and promotions are offered in similar months, mostly in late Winter and lowest in late Fall (Figure D4).

Based on these results, we propose three distinct improvements aimed at maximizing market share with minimal cost:

1) Leveraging the lack of negative trend of market share with price during spring.
2) Reduce price during the summer months
3) Strategically time promotional activities

Figure D1: Predicted Market share by Year

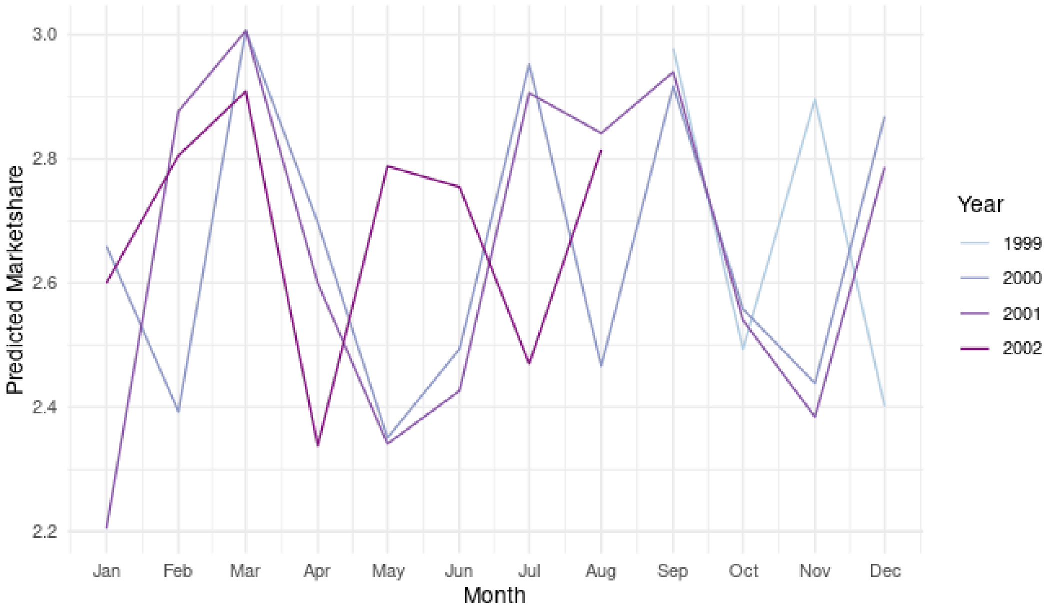


Figure D2: Price by Month

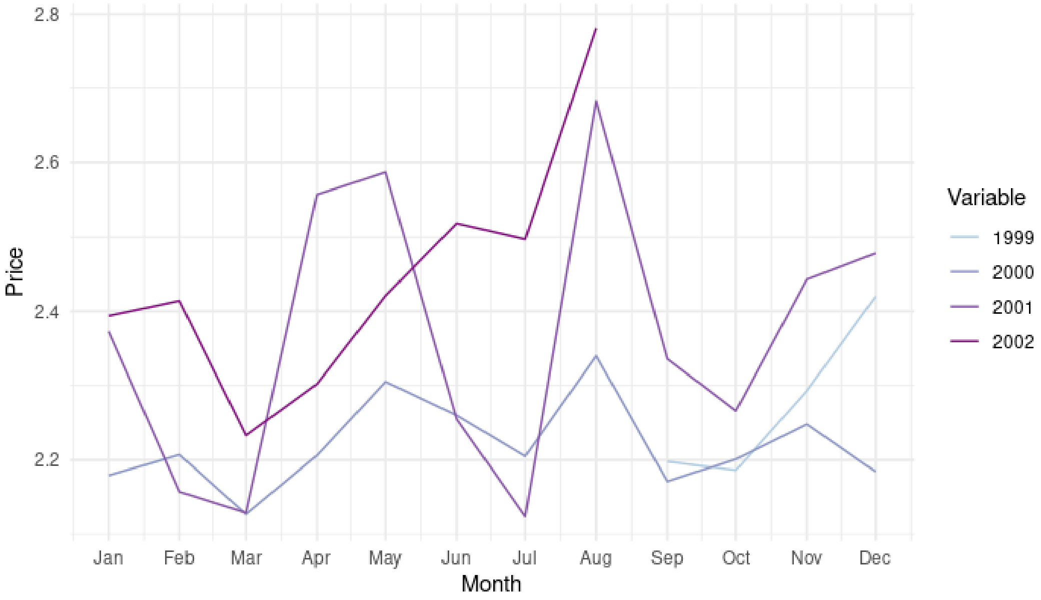


Figure D3: Ad Exposure by Year

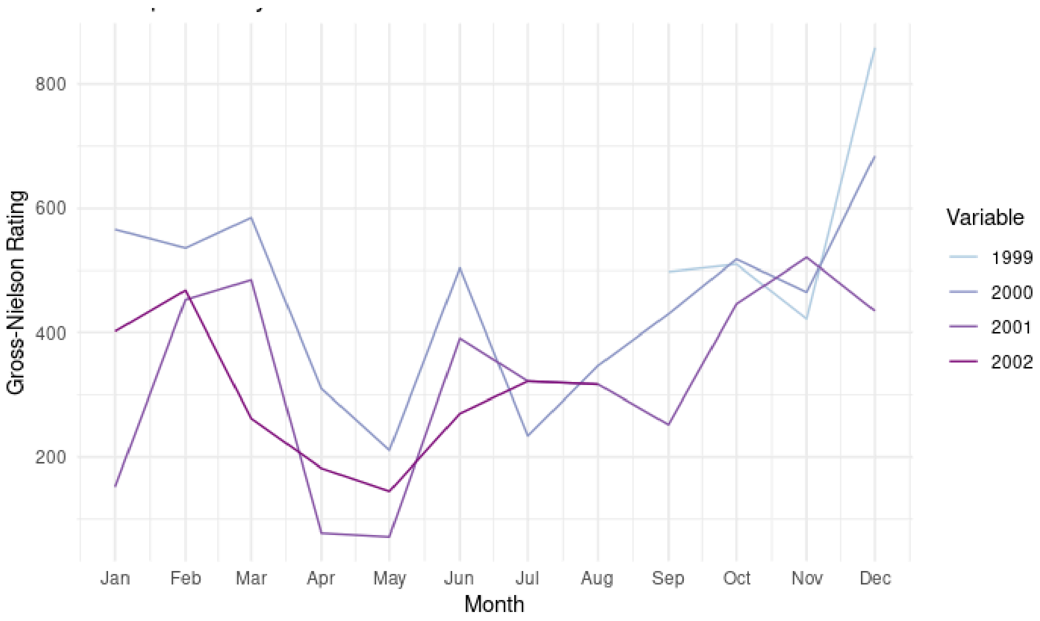


Figure D4: Count of Promotions and Discounts Offered in Each Month

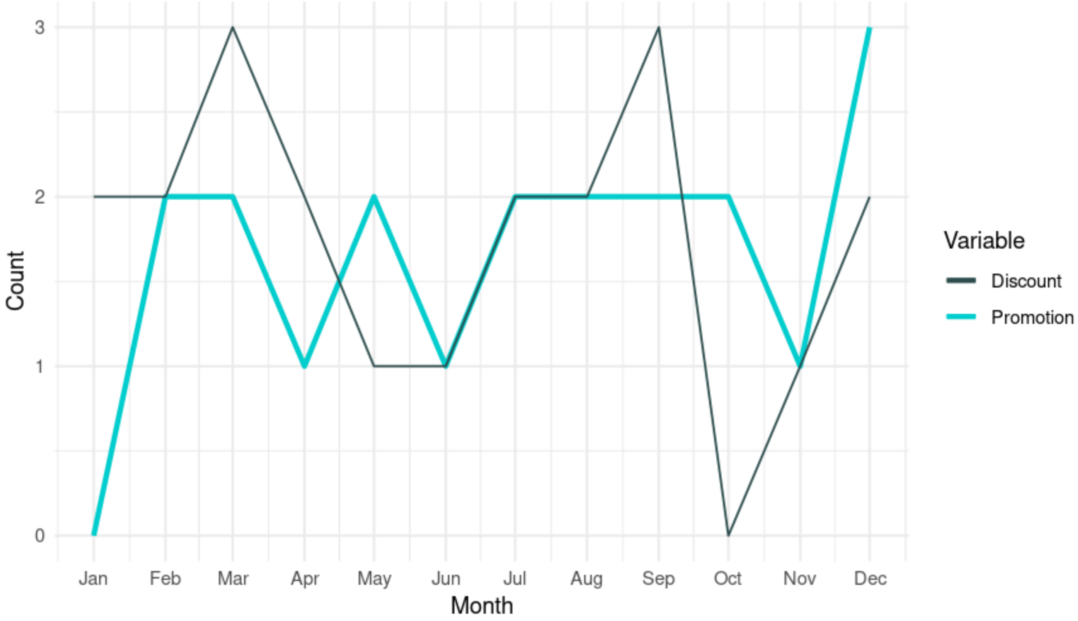


Figure D5: Market Share – Autocorrelation Function

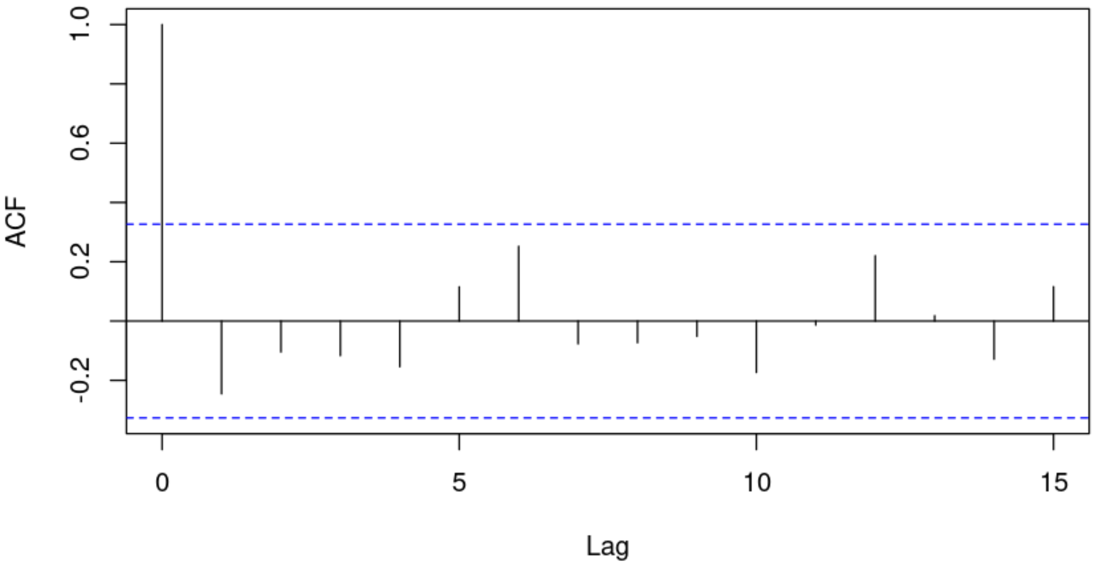
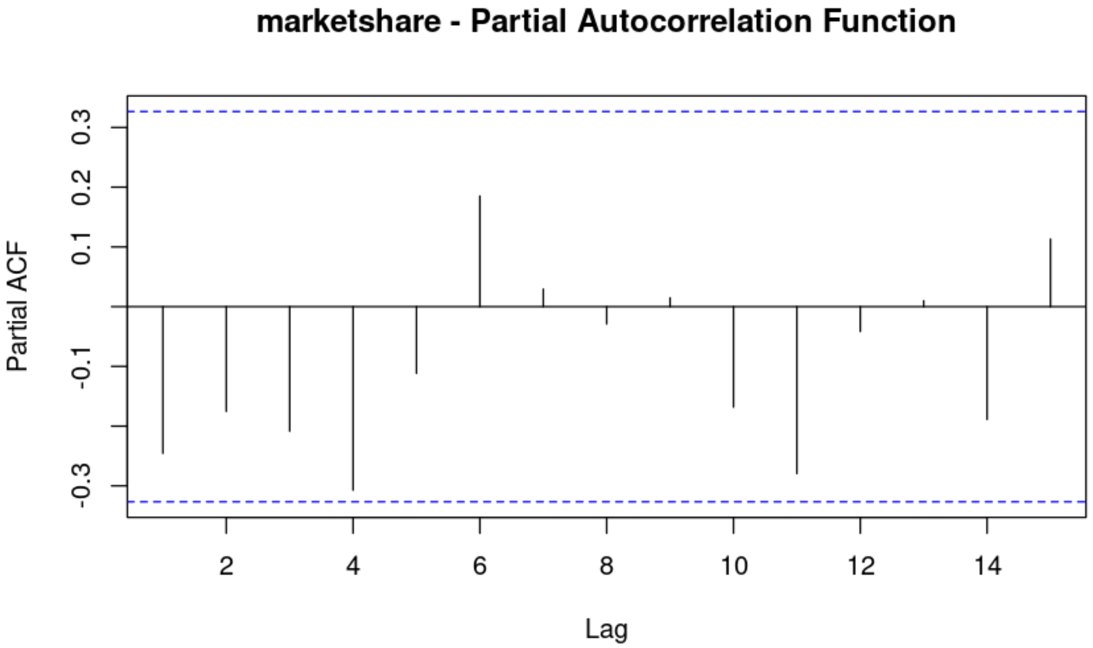


Figure D6: Market Share – Partial Autocorrelation Function



Appendix E: R Code

~~~
---
title: "Marketshare Analysis"
output: pdf_document
date: "2023-12-08"
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)

library(dplyr)
library(readxl)
library(ggplot2)
library(ggpubr)
library(lubridate)
library(car)
library(gridExtra)
library(moments)
library(broom)
library(forecast)
library(leaps)
library(MASS)
library(ggcorrplot)
library(grid)
library(purrr)
library(SimDesign)
library(RColorBrewer)
library(knitr)
library(caret)
library(mlbench)
```
~~~

~~~~
# Data Exploration

```{r Basic Model Setup}
df <- read_excel("market_share.xlsx")

# Modify the Dataframe into df_num and df_mod
# Replace month with numeric, and add a date column
df_mod <- df

df_mod <- df %>%
  mutate(
    # Create a numeric representation of the month
    month_num = match(month, month.abb),
    # Create a numeric representation of the date
    date = (year - 1999) + (month_num - 1) / 12
  ) %>%
  dplyr::select(-month, -year, -idnum) %>%
  rename(month = month_num)

df_num <- df_mod %>%
  select_if(is.numeric)
```

```{r Stats Table, Cor/Scatter Matrices, fig.width=13, fig.height=10}
# Calculate descriptive statistics

# Descriptive Statistics Function
descriptive_stats <- function(x) {
  c(mean = mean(x, na.rm = TRUE),
    median = median(x, na.rm = TRUE),
~~~~

```
    sd = sd(x, na.rm = TRUE),
    min = min(x, na.rm = TRUE),
    max = max(x, na.rm = TRUE),
    skewness = skewness(x, na.rm = TRUE))
}

# Format and print the descriptive statistics
stats_table <- as.data.frame(lapply(df_num, descriptive_stats))
stats_table <- t(stats_table)
print(stats_table)



# Correlation Matrix
corr_matrix <- cor(df_num)
ggcorrplot(corr_matrix, type = "lower", hc.order = FALSE, outline.color = "white",
        lab = TRUE,
        lab_size = 8,
        tl.cex = 20) +
  theme(legend.text = element_text(size = 13))

# VIF
lm_full <- lm(marketshare ~ ., data=df_num)
summary(lm_full)
vif(lm_full)

# Scatterplot Matrix
ggpairs(df_num, progress = FALSE, lab_size = 8, upper = list(continuous = wrap("cor", size =
7))) +
  theme(
    axis.text.x = element_text(size = 12),
    axis.text.y = element_text(size = 12),
    strip.text.x = element_text(size = 14),
    strip.text.y = element_text(size = 14))
```

~~~
```{r Dotplots, fig.height=8, fig.width=8}
# Create dotplots to visualize the distributions

# Dotplot function
create_dotplot <- function(var_name) {
  ggplot(df, aes_string(x = var_name)) +
    geom_dotplot(method="histodot", stackdir = 'up', stackratio = 1.5, dotsize = 0.5) +
    theme(axis.title.y = element_blank(),
          axis.title.x = element_blank(),
          axis.text.y = element_blank(),
          axis.ticks.y = element_blank()) +
    labs(title = toTitleCase(var_name))
}

# Format and print the dotplots
plot_list <- lapply(names(df_num)[-6], function(var_name) create_dotplot(var_name))

grid.arrange(
  arrangeGrob(plot_list[[1]], nrow = 1),
  arrangeGrob(plot_list[[2]], plot_list[[3]], plot_list[[4]], plot_list[[5]], ncol = 2),
  heights = c(2, 4),
  top = textGrob("Dotplot Variable Distributions", gp = gpar(fontface = "bold", fontsize = 20))
)

~~~

```{r Marketshare Scatters}
```

~~~
# Plot marketshare vs. every predictor

y_var <- "marketshare"
independent_vars <- setdiff(names(df_mod), y_var)

# Scatterplot Function
plot_function <- function(indep_var) {
  ggplot(df_mod, aes_string(x = indep_var, y = y_var)) +
  geom_point() +
  geom_smooth(method = "lm") +
    labs(title = paste(y_var, "vs", indep_var),
         x = indep_var,
         y = y_var)
}

# Print the scatterplots
plots <- map(independent_vars, plot_function)
walk(plots, print)

```

# Time Series Analysis

```{r Time Series dataframes}
# Generate separate dataframes needed

# Create a df stacked by year
df_stacked <- df %>%
  mutate(
    month_num = match(month, month.abb),  # Months to 1-12
    norm_date = make_date(2000, month_num, 1)  # Normalize to year 2000 (arbitrary)
~~~

```
  )

# Create a df arranged by date
df_lag <- df_mod %>% arrange(date)

# Create a df with labeled seasons
df_seasons <- df_mod %>%
  mutate(
    season = case_when(
      month %in% c(1, 2, 3) ~ "Winter",
      month %in% c(4, 5, 6) ~ "Spring",
      month %in% c(7, 8, 9) ~ "Summer",
      month %in% c(10, 11, 12) ~ "Fall",
      TRUE ~ NA_character_
    )
  )
```

```{r Yearly Trends}
# Visualize the yearly trends for the variables

# Extravagant color scheme
bu_pu_colors <- brewer.pal(5, "BuPu")[c(2:5)]

# Plot stacked price by year
ggplot(df_stacked, aes(x = norm_date, y = price, group = factor(year), color = factor(year))) +
  geom_line() +
  scale_x_date(date_breaks = "1 month", date_labels = "%b") +  # Abbreviated month name
x-axis labels
  labs(title = "Price by Year",
       x = "Month",
       y = "Price",
       color = "Variable") +
  theme_minimal() +
  scale_color_manual(values = bu_pu_colors)

# Plot stacked gnrpoints by year
ggplot(df_stacked, aes(x = norm_date, y = gnrpoints, group = factor(year), color = factor(year)))
+
  geom_line() +
  scale_x_date(date_breaks = "1 month", date_labels = "%b") +  # Abbreviated month name
x-axis labels
  labs(title = "Ad Exposure by Year",
       x = "Month",
       y = "Gross-Nielson Rating",
       color = "Variable") +
  theme_minimal() +
  scale_color_manual(values = bu_pu_colors)



# Group by month and sum promotion and discount variables
monthly_sums <- df_stacked %>%
  group_by(norm_date) %>%
  summarise(promotion_sum = sum(promotion, na.rm = TRUE),
            discount_sum = sum(discount, na.rm = TRUE)) %>%
  ungroup()

# Plot summed promotion and discount variables vs. grouped month
ggplot(monthly_sums, aes(x = norm_date)) +
  geom_line(aes(y = promotion_sum, color = "Promotion"), size = 1.2) +
  geom_line(aes(y = discount_sum, color = "Discount")) +
```

~~~
  scale_x_date(date_breaks = "1 month", date_labels = "%b", limits = c(make_date(2000, 1, 1),
make_date(2000, 11, 31))) +
  labs(title = "Count of Promotions and Discounts Offered in Each Month",
       x = "Month",
       y = "Count",
       color = "Variable") +
  theme_minimal() +
  scale_color_manual(values = c("Promotion" = "cyan3", "Discount" = "darkslategray"))

```


```{r AutoCorrelation}
# Test for autocorrelation trends

# Graph autocorrelation for marketshare
acf(df_lag$marketshare, main="marketshare - Autocorrelation Function")
pacf(df_lag$marketshare, main="marketshare - Partial Autocorrelation Function")

# Graph autocorrelation for gnrpoints
acf(df_lag$gnrpoints, main="gnrpoints - Autocorrelation Function")
pacf(df_lag$gnrpoints, main="gnrpoints - Partial Autocorrelation Function")

```


```{r Correlations by Season}
# Analyze variables by season

# Fit models for marketshare vs. all predictors by season
models <- df_seasons %>%
  group_by(season) %>%
  do(model = lm(marketshare ~ price + gnrpoints + discount + promotion, data = .))
~~~

```
# Extract coefficients into a data frame
coefficients_df <- models %>%
  summarize(season = first(season),
            price_coef = coef(model)["price"],
            gnrpoints_coef = coef(model)["gnrpoints"],
            discount_coef = coef(model)["discount"],
            promotion_coef = coef(model)["promotion"]) %>%
  arrange(season)

# Print the coefficients table
print(coefficients_df)


# Compute the correlation for each predictor with marketshare by season
correlation_df <- df_seasons %>%
  group_by(season) %>%
  summarize(
    price_cor = cor(marketshare, price, use = "complete.obs"),
    gnrpoints_cor = cor(marketshare, gnrpoints, use = "complete.obs"),
    discount_cor = cor(marketshare, discount, use = "complete.obs"),
    promotion_cor = cor(marketshare, promotion, use = "complete.obs")
  ) %>%
  arrange(season)

# Print the correlations table
print(correlation_df)


# Extract detailed statistics for each coefficient
detailed_stats <- models %>%
```

~~~~
  rowwise() %>%
  do(tidy(.$model)) %>%
  filter(term != "(Intercept)") %>%
  ungroup()

# Print the detailed statistics
print(detailed_stats)
```


# Model Selection

```{r Dataframe with isSpring indicator and quadratic month}
# Add isSpring indicator variable
df_mod <- df_mod %>% mutate(isSpring = month %in% c(1, 2, 3))

# Add quadratic month term
df_mod$month_sq <- df_mod$month^2

```


```{r Basic Model Residuals, Stacked by Month}
# Create a stacked residual plot to visualize seasonal trends in the residuals

# Create the basic model
basic_model <- lm(marketshare ~ price + promotion + discount + isSpring, data = df_mod)
summary(basic_model)

basic_model_res <- residuals(basic_model)
summary(basic_model_res)

# Add the residuals to the dataframe
~~~~

~~~
df_stacked$residuals <- basic_model_res

# Plot residuals vs. months using ggplot2
ggplot(df_stacked, aes(x = as.factor(month_num), y = residuals)) +
  geom_point() +
  xlab("Month") +
  ylab("Residuals") +
  ggtitle("Residuals vs. Months")

qqnorm(basic_model_res)
qqline(basic_model_res)

```


```{r AIC}
# Perform AIC

# Dynamic predictor selection
chosen_predictors <- c("price", "discount", "promotion", "gnrpoints", "month", "month_sq",
"isSpring")

# Perform regsubsets for the chosen predictors
formula <- as.formula(paste("marketshare ~", paste(chosen_predictors, collapse=" + ")))
models <- regsubsets(formula, data=df_mod, nbest=7)
models_summary <- summary(models)

# Empty df for storage
model_info <- data.frame(size = integer(), AIC = double(), row.names = character())

# Calculate AIC for the models
for (i in 1:length(models_summary$cp)) {
~~~

~~~
  included_predictors <- chosen_predictors[models_summary$which[i,][-1]]
  formula <- as.formula(paste("marketshare ~", paste(included_predictors, collapse=" + ")))
  model <- lm(formula, data=df_mod)
  aic_value <- AIC(model)

  # Append the model information
  model_info <- rbind(model_info, data.frame(size = sum(models_summary$which[i,][-1]), AIC
= aic_value, row.names = paste(included_predictors, collapse=" + ")))
}

# Order models by AIC and format
ordered_models <- model_info[order(model_info$AIC),]
ordered_models$AIC <- round(ordered_models$AIC, digits = 2)
kable(ordered_models, format = "html", table.attr = "style='width:70%;'", caption = "Model
Selection based on AIC")

```


```{r Cp}
# Perform Cp

# Cp on all possible predictors
models <- regsubsets(marketshare ~ price + discount + promotion + gnrpoints + month +
month_sq + isSpring, data=df_mod, nbest=7)
models_summary <- summary(models)

# Print model variables ordered by R^2
model_info <- data.frame(models_summary$which, Cp=models_summary$cp)
ordered_models <- model_info[order(model_info$Cp), -1]
print(ordered_models)
~~~

~~~
# Summarize the lowest cp model
cp_lm <- lm(marketshare ~ price + discount + promotion + month + month_sq + isSpring,
data=df_mod)
summary(cp_lm)

qqnorm(residuals(cp_lm))
qqline(residuals(cp_lm))



# Chosen final model, for now
final_model <- lm(marketshare ~ price + discount + promotion + month + month_sq + isSpring,
data=df_mod)

```


```{r Centered Model}
# Test if centering the model shows benefit to VIF

df_mod_centered <- df_mod %>%
  mutate(across(c(price, gnrpoints, month, month_sq), ~ . - mean(.)))

testingModel <- lm(marketshare ~ price + discount + promotion + month + month_sq + isSpring,
data = df_mod_centered)
vif(testingModel)
summary(testingModel)

testResiduals <- residuals(testingModel)
qqnorm(testResiduals)
qqline(testResiduals)
~~~

~~~~
```


```{r Cross Validation}
# Perform LOOCV

# Run with LOOCV control
control <- trainControl(method = "LOOCV")
model_loocv <- train(marketshare ~ price + discount + promotion + month + month_sq +
isSpring,
                     data = df_mod,
                     method = "lm",
                     trControl = control)

# Results
print(model_loocv)



# Perform RFE

# isSpring to numeric
df_mod$isSpring <- as.numeric(df_mod$isSpring)

# Run with RFE control
control <- rfeControl(functions = lmFuncs,
                      method = "LOOCV")
results <- rfe(df_mod[,-which(names(df_mod) == "marketshare")],
               df_mod$marketshare,
               sizes = c(1:ncol(df_mod)-1),
               rfeControl = control)

# Results
~~~~

~~~
print(results)

result_variables <- results$optVariables
print(result_variables)
```

```{r Model Assumptions, fig.height=7, fig.width=7}
# Verify model assumptions for the potential final model

quiet(crPlots(final_model))
shapiro.test(residuals(final_model))

```

```{r Leverage}
# Verify leverage for observed outliers

plot(cp_lm$fitted.values, residuals(final_model), xlab = "Fitted Values", ylab = "Residuals",
main = "Residuals vs. Fitted Values")
abline(h = 0, col = "red")

# Calculate Cook's distance for the model
cooks_distance <- cooks.distance(final_model)

# Plot Cook's distance
plot(cooks_distance, pch = "*", cex = 2, main = "Cook's distance")
abline(h = 4 / length(cooks_distance), col = "red")

# Show the corresponding rows in the dataframe
high_cooks_indices <- order(cooks_distance, decreasing = TRUE)
df_mod[high_cooks_indices[1:3], ]
~~~

```
```