Let’s together explore how much/ which people like Decaffeinated Folgers Coffee!

PREDICTIVE ANALYSIS USING SAS

**GROUP : 1**

Riti Kumari

Shreeya Bhosale

Paryag Mehta

Pooja Prasad

Shashank Shekhar

**Table of Contents**

1. Abstract ………………………………………………………………………….……………………………………………..3
2. Overview ……………………………………………………………………………………………………………………….3
3. Descriptive Analysis ……………………………………………………………………………………………………….3
4. Data Pre-processing ……………………………………………………………..……………………………………….4
5. Analysis ………………………………………………………………………………………………………………………….5
6. Brand Selection (MDC) …………………………………………………………………………………………….5
7. RFM Analysis ……………………………………………………………………………………………….…………..8
8. Trend Analysis …………………………………………………………………………………………………….……9
9. Inferences and Recommendations …………………………………………………………………………………12
10. References ……………………………………………………………………………………………………………………..12

**Predictive Analytics of Coffee data**

Shreeya Bhosale, Riti Kumari,

Paryag Mehta, Shashank Shekhar, Pooja Prasad

*The University of Texas at Dallas, Richardson, TX, USA*

**ABSTRACT**

We consider the task of analyzing household and store level data of the product coffee. There are 3 categories of coffee in the datasets and we have selected **GROUND DECAFFEINATED COFFEE**.

**OVERVIEW**

The data used in the report is grocery store scanner data of coffee consumption. The analysis is concentrated on the usage of “DECAF” coffee across the panels. The analysis has been performed as Brand manager of Folgers to accelerate the sales of decaf coffee. We performed a competitive analysis amongst top performing coffee brands to bring business values and insights. And accordingly, the recommendation can be made.

**DESCRIPTIVE ANALYSIS**

We performed various descriptive analysis and observed the following results.

* The DECAFFINATED Coffee captures around 12.1% of the total coffee share.

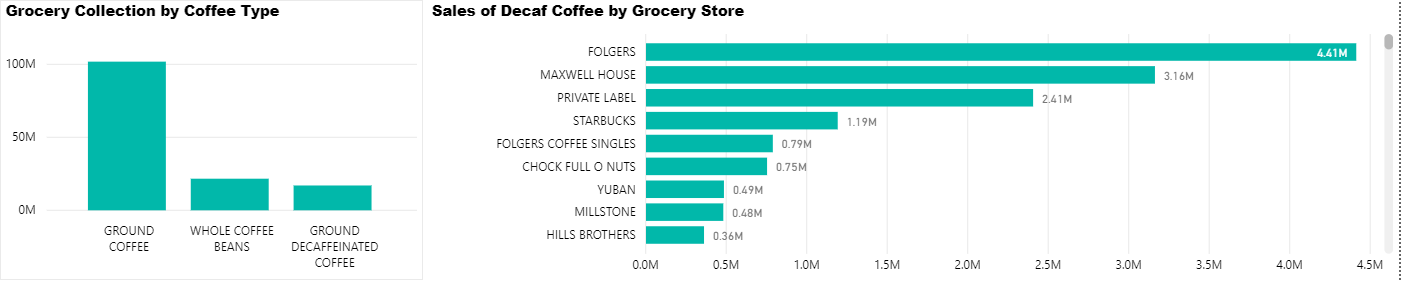


Figure 1: Grocery Collection by Coffee Type

* Folgers, Maxwell House, Private Label, Starbucks and Chock Full are major brands with concentration of 80% of sales at grocery stores.

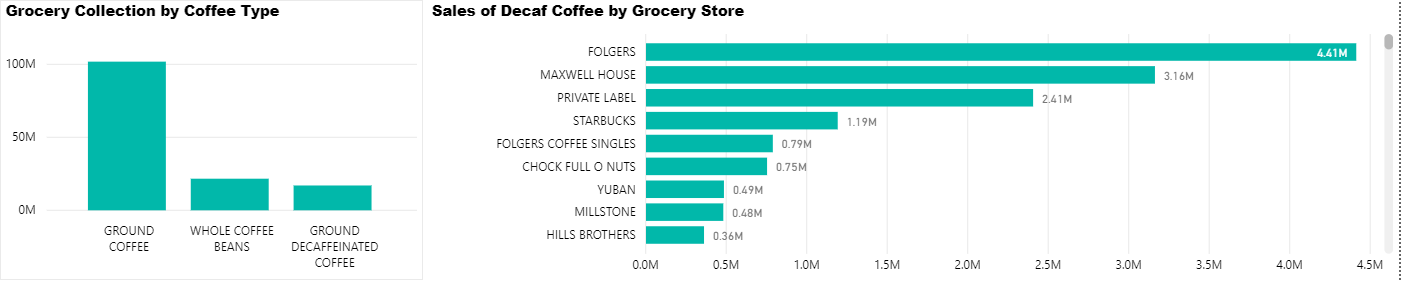


Figure 2: Sales of Decaf Coffee by Grocery Store

* The New York, Los Angeles, Chicago, Boston and New England has been the leading markets for sales.

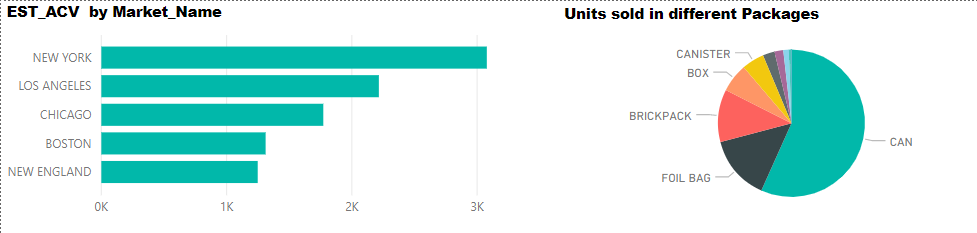


Figure 3: EST\_ACV by Market Name

* People prefer to buy coffee in cans followed by foil bags and brickpack. Sales in these three packages capture more than 80% of the sales.

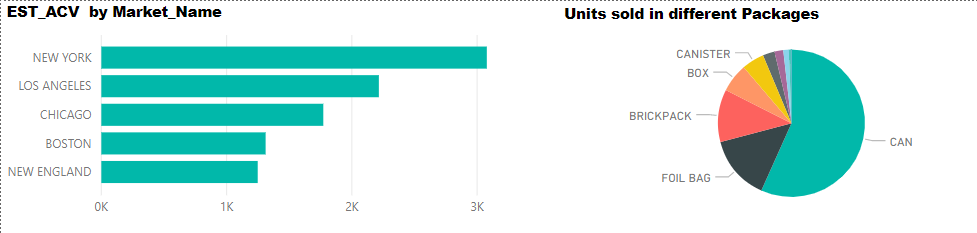


Figure 4: Units sold in different packages

**DATA PRE-PROCESSING**

We have the data of coffee sales at grocery stores by each week and by each panels. The sales were affected by the promotions and advertisements (Display & Features). The another dataset consists of the panel data. The sales were done from different store in different packages and weights in different time-frame. We know, the price is the strongest factor to drive the sales. Hence, there is need to calculate the weightage average price per ounce of each brand. Similarly, the display and features were standardised by calculating their weighted values.

Coffee grocery and prod\_data has been imported and joined using common key upc code. Weighted average price for each brand per week per store is calculated. Calculated the price per ounce for each brand per week per store per item. Calculated the market share for each brand per week per store per item. Calculated average of this market share across each brand per week per store. Flattened the weighted average price for each brand per week per store into respective columns as P1-P4 by using PROC TRANSPOSE. Similarly, the weighted average display and weighted average feature values are calculated for each brand per week per store. We filtered households who have chosen ground decaffeinated coffee from coffee panel grocery data. The measures, i.e. the units and the price paid by each panel per brand per week per store have been aggregated. Transposed data is then merged with coffee\_panel\_grocery data by per week per store. Demographic data has been appended to above dataset using common key using PANID.

**ANALYSIS**

In our analysis, we aim to address the following questions:

**Q3**

**Q2**

**Q1**

**BRAND SELECTION**

To answer **Q1**, weuseMDC (multinomial discrete choice) procedure because it analyzes models where the choice set consists of multiple alternatives. If a customer has chosen a brand 1 then it means that the utility of brand 1 must be greater than the utility of other brands for the customer. In our data, utility varies based on customer characteristics as well as characteristics of the choice alternatives.

* Table 1 shows the frequency distribution of the four choice alternatives. Folgers brand is most frequently chosen (35%)

|  |  |  |  |
| --- | --- | --- | --- |
| **Index** | **Choice** | **Frequency** | **Percent** |
| 1 | Folgers | 1767 | 34.82 |
| 2 | Maxwell House | 850 | 16.75 |
| 3 | Private Label | 1519 | 29.94 |
| 4 | Others | 938 | 18.49 |

Table 1: Frequency distribution of coffee choice

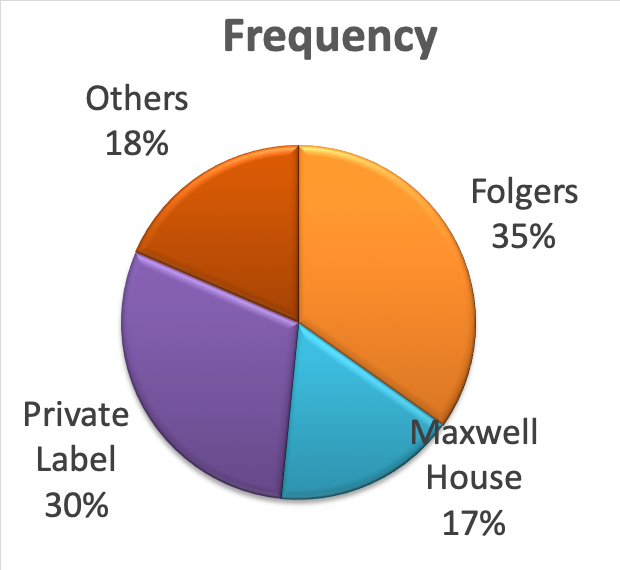


Figure 5: Frequency distribution of coffee choice

From MDC results, it is observed that consumer prefers Folgers followed by Private Label, Others and Maxwell.

The parameter estimates of MDC can be used to forecast the choice probability of individuals that are not in the input data set.

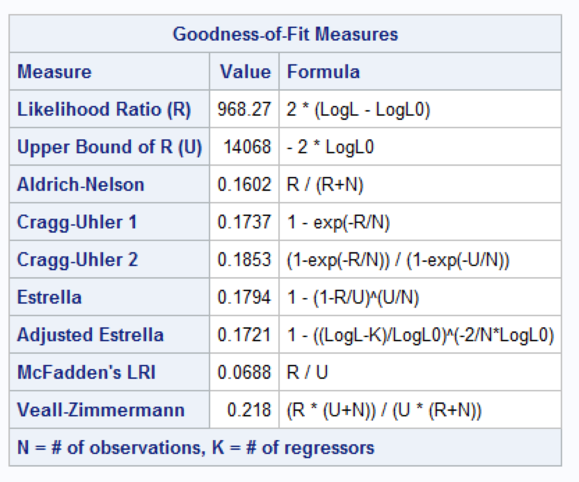
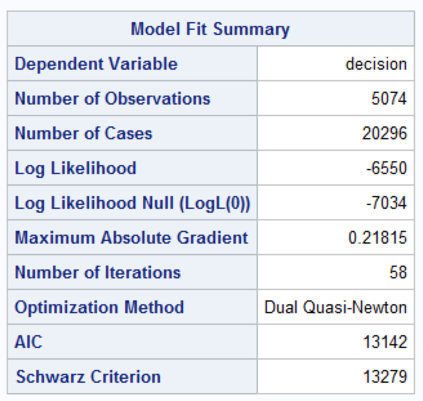


Table 2: Model fit summary and Goodness of Fit Measures

MDC procedure uses maximum likelihood method for model estimation. There is some improvement in log likelihood (-6550) over Null log likelihood (-7034) signifying the model is slightly better than intercept only model. McFadden likelihood ratio index (which is bounded by 0 and 1) indicates the model fit in multinomial logit model. Since the value is 0.0688 ~ 6.88%, the model is not a good fit. (1 corresponds to perfect fit).

Null hypothesis: All coefficients except for an intercept term are zero

Alternate hypothesis: At least one of the coefficients (except intercept term) is non-zero.

Conditional logit parameter estimates and interpretation: Reference brand is brand 1: FOLGERS

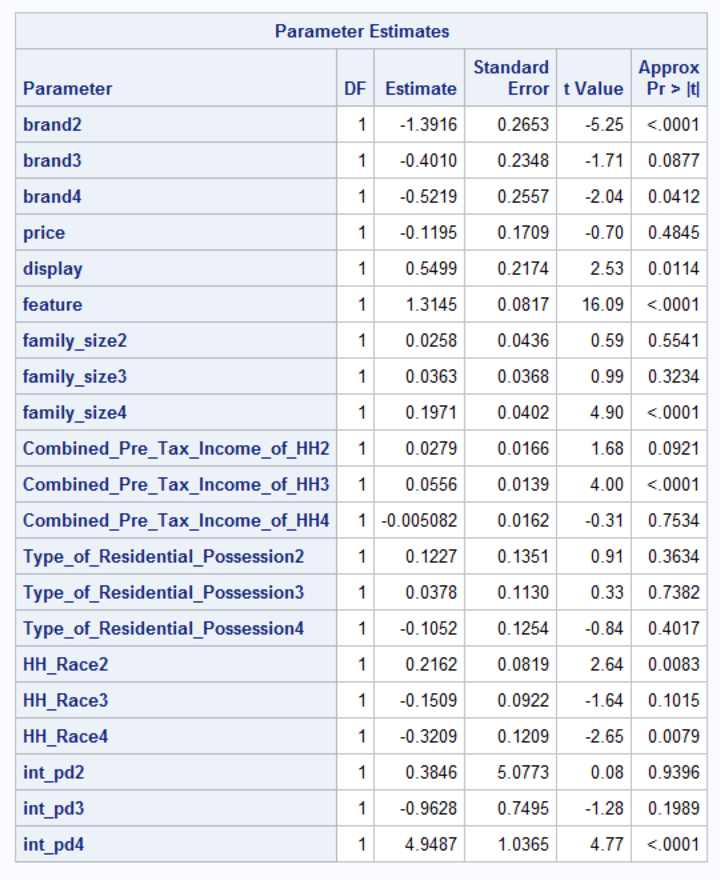


Table 3: Conditional logit parameter estimates

At a significance level of 5%

Price: As per the model estimation, price is not significant in determining likelihood of decaf coffee brand selection.

Display: Higher display leads to more likelihood of decaf coffee brand selection, i.e. high probability of selecting the brand on display.

Feature: Higher feature leads to more likelihood of decaf coffee brand selection. Feature has higher effect when compared to display.

There are 3 sets of coefficients estimated for each of combined pre-tax income for household, household race, residential possession and family size. Reference category is brand1 - FOLGERS.

Combined Pre-Tax Income for HH: With respect to the reference category (FOLGERS), those who have higher combined pre-tax income for household are more likely to choose OTHERS decaf coffee brand.

Parameter estimate of combined pre-tax income for household for selection of brand2 – MAXWELL and brand 4 – PRIVATE is not significant at 5% significance level which suggests household can select any of the brand1, brand2 and brand4.

Family Size: With respect to the reference category (FOLGERS), those who have more number of family members are more likely to choose PRIVATE decaf coffee brand.

Parameter estimate of Family Size for selection of brand2 – MAXWELL and brand 3 - OTHERS is not significant at 5% significance level which suggests household can select any of the brand1, brand2 and brand3.

Household Race: With respect to the reference category (FOLGERS), household who are of Hispanic race are more likely to choose MAXWELL and PRIVATE decaf coffee brand.

Parameter estimate of Household Race for selection of brand3 – OTHERS is not significant at 5% significance level which suggests household may select brand 1 or brand3.

Type of Residential Possession: With respect to the reference category (FOLGERS), None of the parameter estimates are significant at 5% significance level for decaf coffee brand selection which suggests household may select any of brand 1, 2, 3 or brand4.

Price\*Display: With respect to the reference category (FOLGERS), keeping price constant, if the coffee brand is displayed in the market, the probability of choosing brand4 -PRIVATE is more. Parameter estimate of Price\*Display for selection of brand2 – MAXWELL and brand 3-OTHERS are not significant at 5% significance level which suggests household may select brand 1 or brand 2 or brand3.

The predicted choice probabilities of individuals:

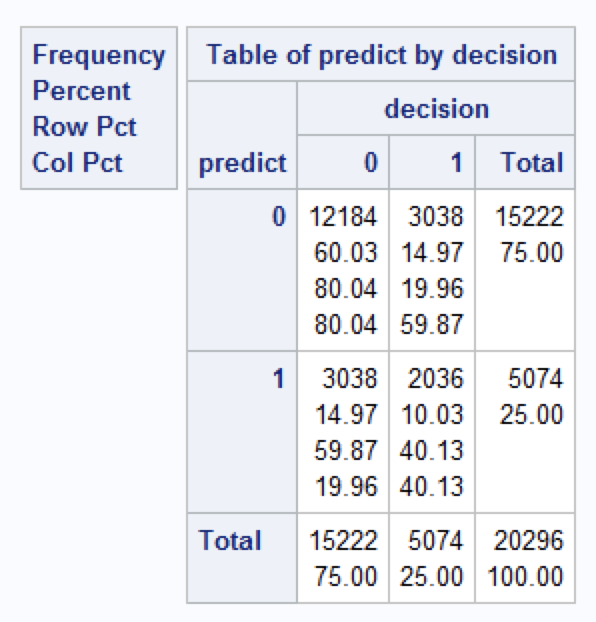


Table 4: Predicted choice probabilities

It is expected that the decision maker will choose mode /choice 1 based on predicted probabilities for all modes.

Frequency table of predict by decision shows that the expected and actual choices made by individual are almost accurate and prediction done by the model is 70.06% correct.

Precision = TP / (TP+FP) = 40.13%

Accuracy = (TP+TN) / (TP+FN+TN+FP) = 70.06%

Fall-out (% negative misclassified) = FP / (TN+FP) = 59.87%

Recall/ Hit Rate (% positive correctly classified) = TP / (TP+FN) = 40.13%

Own and cross price elasticity for FOLGERS.

Own price elasticity (j wrt j) = (1-Prob(j))\*Xj\*β

Cross price elasticity (= (-Prob(j))\*Xj\*β

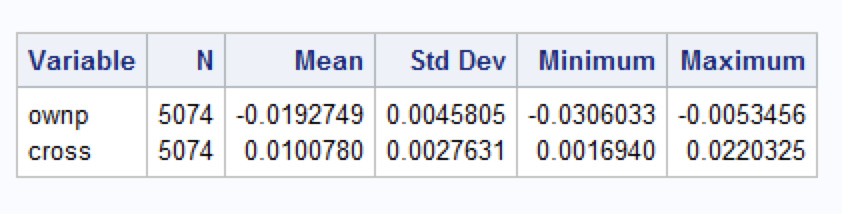


Table 5: Mean of Own and cross price elasticity

Mean value of own price elasticity here is negative indicating inverse relationship between quantity (product) demanded and price. As the price increases for Folgers, its demand will decrease.

Cross price determines the demand and price of other brand with respect to my brand. Here, the positive mean cross elasticity indicates that as price for other brand increases, the demand for Folgers will increase.

**RFM ANALYSIS**

With the help of the given data, we try to categorise the users in different segements and analyse who are our most valuable customers. For customer segmentation, their respective RFM score has been calculated. This includes the parameters like when has the customer made his last purchase, how many times, he has been purchasing and the most important, how much the company is getting revenue i.e. monetary worth of the customer.

The panel data has been ranked as per their recency, frequency and their monetary worth into five ranks. The same With RFM\_Score, the customers were divided into ‘Prime’ and ‘Non-Prime’ customers.

We have taken 443 score to be the baseline for the prime customer segmentation.

We ran RFM analysis by taking some features like Combined\_Pre\_Tax\_Income, Number of pets (Number\_of\_cats + Number\_of\_Dogs), Family\_Size, Marital\_Status into account.



Table 6: Customer distribution bewteen prime /Not prime

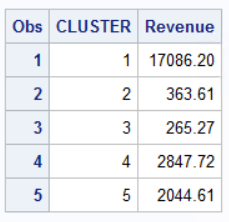


Table 7: Revenue by Cluster

We performed the clustering to categorize the customers into five cluster. As per the results, we can see that the customers lying in Cluster1 generates a lot of revenue for the coffee sales. This specifies that the panels in cluster 1 are loyal to the Folgers brand.

The customers lying in cluster 4 and 5 generates revenue for the company. However, they are not very loyal to the brand Folger. They are not the prime customers and lie in the switch zone. Hence, the efforts should be made to retain these customers.

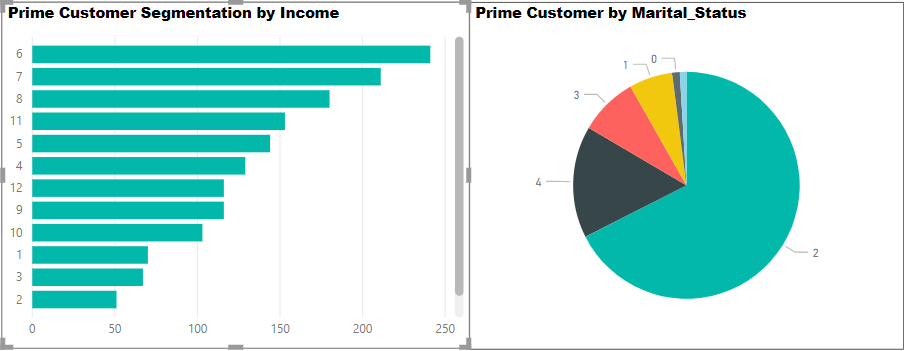


Figure 6: Customer segmentation by Income and Marital Status

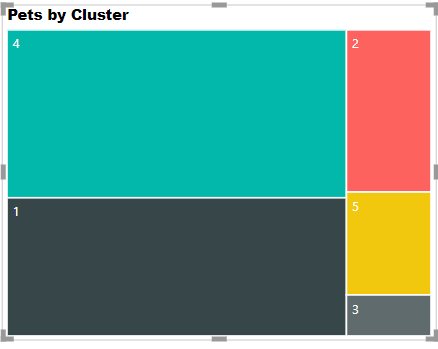


Figure 7: Customer segmentation by number of pets

As per the clustering analysis, we can see some common behavior/ characteristics of the prime customers:

* The prime customers income range across various income levels. However, the prime members with annual income between $25k to $60k makes the maximum worth as per the RFM analysis.
* The married group people constitute around 67% of the sales of the prime customers.
* The prime customers are generally very fond of pets.
* We also see that the non-prime customers are generally have no pets and belong to extreme income category. They need attention and should be given extra offers to earn their loyalty.

**TIME-SERIES ANALYSIS**

We have weekly sales data from January 2001 to December 2001. The average sales of the Folgers data over the time shows a time-series data with no trend or seasonality.

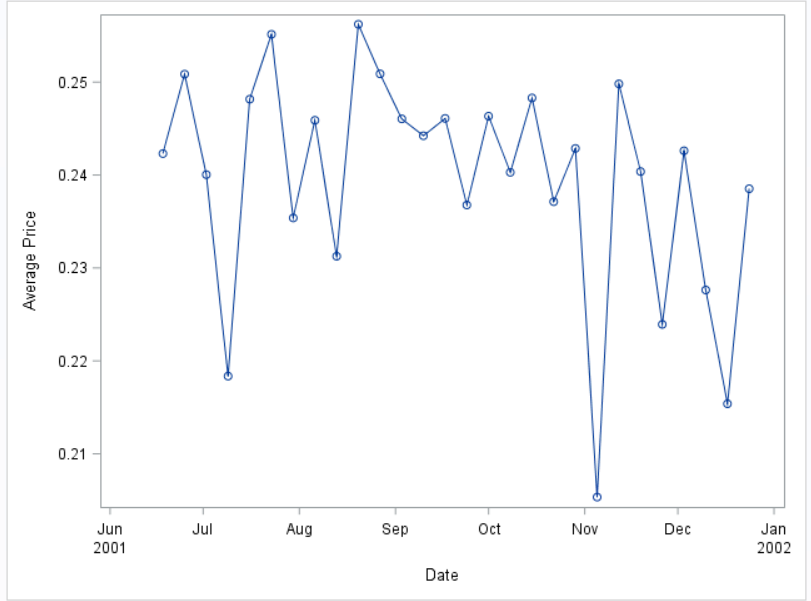


Figure 8: Time series plot of Average Price of Folgers

We can see that the data-sets are not randomly selected but from the same source. Hence, there are high chances of the target variable to be the dependent on previous lag values. Hence, we checked for the auto-correlation.

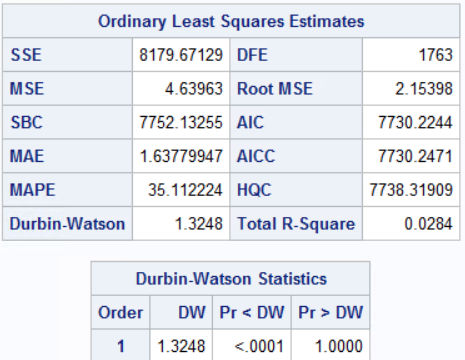


Figure 9: Durbin-Watson Statistics

As per the **Durbin-Watson Statistics**, we can see that the value is 1.32 (range of 0 to 2). This signifies the positive auto-correlation between the target variable at time t with the lag value at kth interval.

We also need to check if the time-series stationarity. We performed the **Dicker-Fullers test**.

H0: There is presence of a unit root

H1: Time-series is either stationarity, trend stationarity or explosive root depending on the test used.

From DF- test, we get the p-value as 0.0024. At 5% of significance level, we **reject** the null hypothesis.

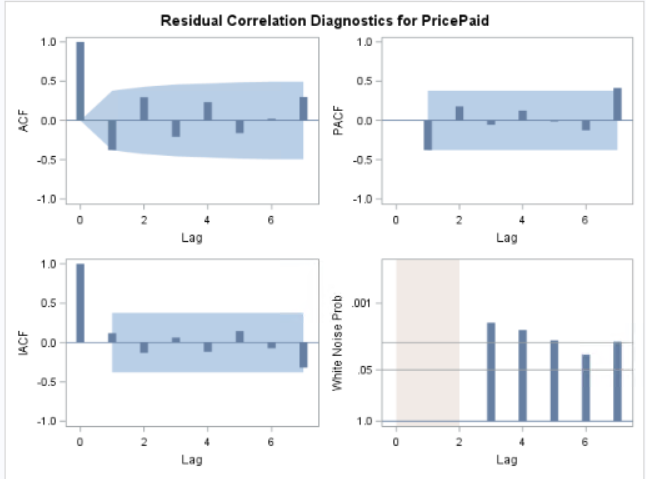


Figure 10: Residual Correlation Diagnostics

Plotting the correlogram plot, from ACF plot, we can see that lag 1 and lag 12 are correlated with the time t.

With PACF plot, we can see that lag 7 is correlated with time t.

Hence, we need to run the ARMA model with MA at lag 1 and lag 12 and AR model with lag 7.

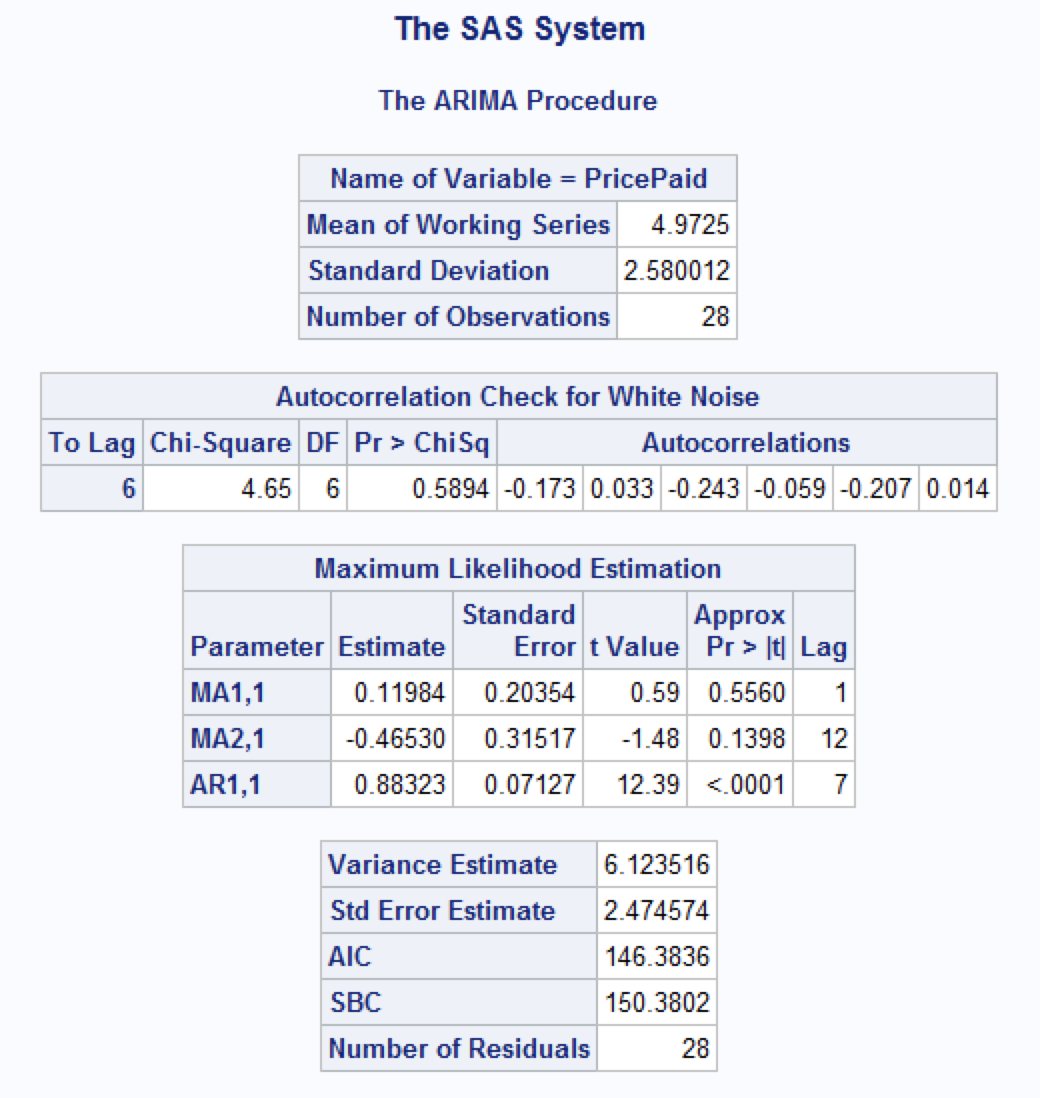


Figure 11: ARIMA statistics

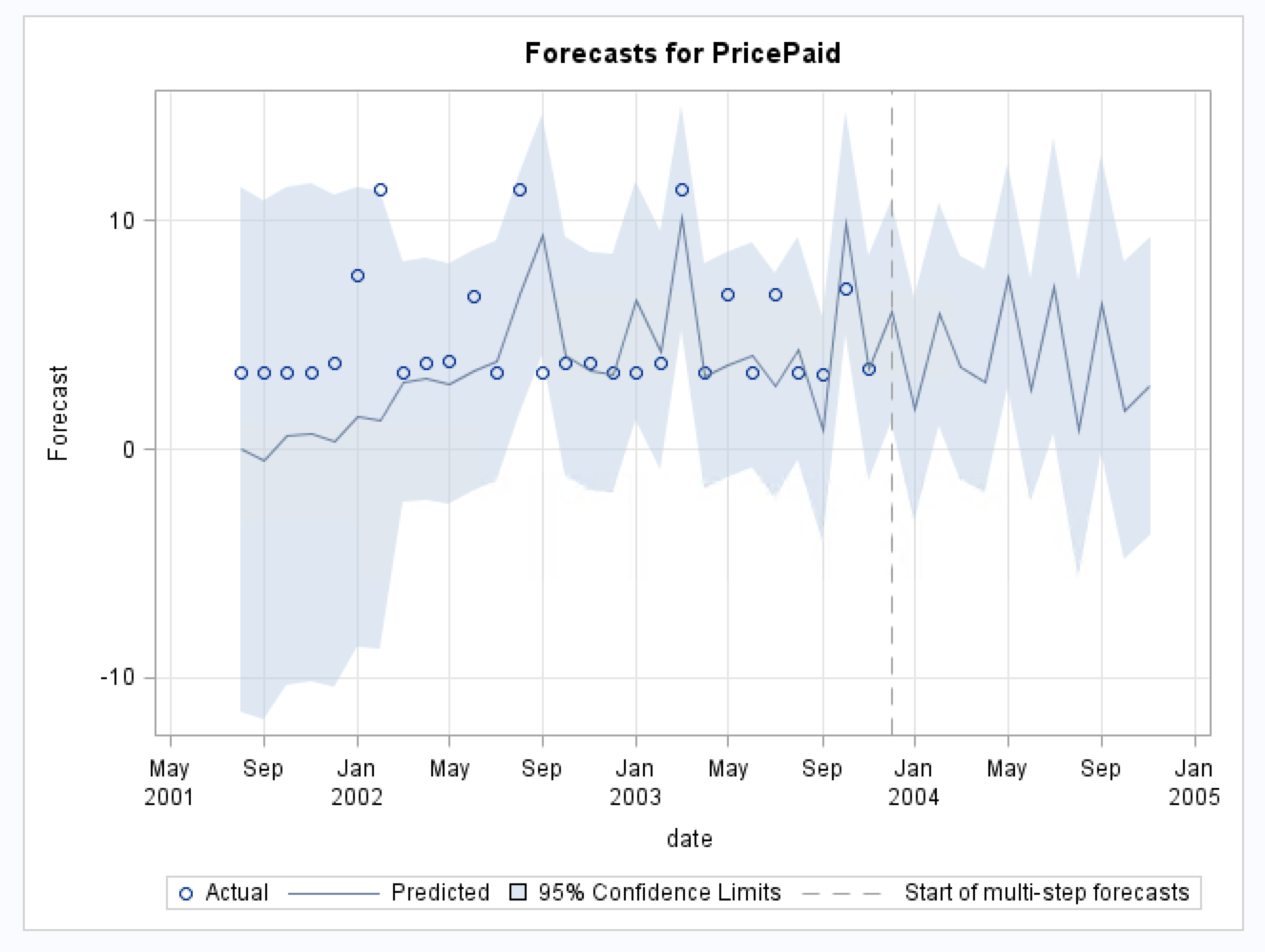


Figure 12: Forecast for PricePaid

From the forecasted price, we can see that the price for Decaffeinated Folgers coffee is expected to be stationary in the coming months. The price will rise followed by the decline in price.

**INFERENCES & RECOMMENDATIONS**

After panel regression, RFM and demographic analysis, we recommend the following marketing strategies:

* As the brand manager of Folgers, we see that it has been leading the market. The next big player is Private Label (Altria Group). But the average price for both the brands differ a lot which signifies that both have different user-group. Hence, Folgers should not compete with the Private Label group but with Maxwell and chock-full brand as they share the same user-group.
* The prime customers are generally the middle-range income married people. We can introduce the vouchers/offers to attract the new customers and retain the old customers.
* The prime customers also have pets. The cats/dogs related food/biscuits can clubbed with the coffee and introduce in the market.
* There are a few clusters, who are not prime customers, but they have good monetary worth. This infers that they are less frequent but when they are coming they are buying in bulk quantity. The discounts on the same will help get their loyalty and hence in long-run will help the company with the customer-lifetime value.
* As a brand manager of Folgers, we are not having data for other companies, but introducing clubbed offers/ price-cut and special discounts in future may help in user acquisition. Also, new strategies need to be formed/modified as per the other companies’ strategies as the sales of my company is highly dependent on the competitive market.

REFERENCES

1. <https://onlinecourses.science.psu.edu/stat501/node/358/>
2. <http://support.sas.com/documentation/cdl/en/etsug/60372/HTML/default/viewer.htm#etsug_macros_sect007.htm>
3. <https://support.sas.com/documentation/onlinedoc/ets/132/mdc.pdf>
4. <https://searchdatamanagement.techtarget.com/definition/RFM-analysis>
5. <https://www.putler.com/rfm-analysis/>
6. <https://clevertap.com/blog/rfm-analysis/>