PREDICTIVE ANALYTICS WITH SAS

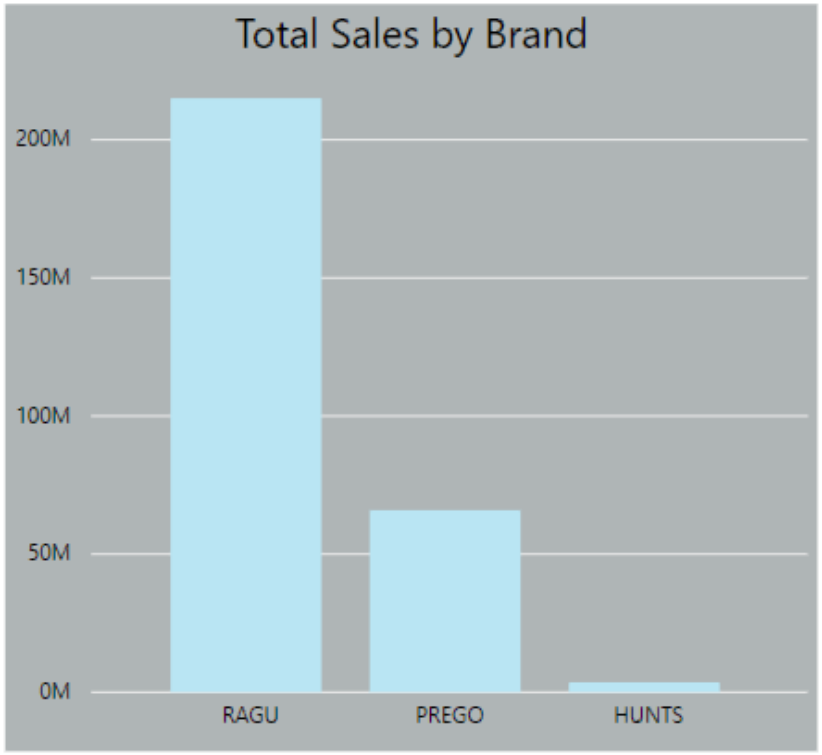
Marketing Insights for Prego Spaghetti Sauce

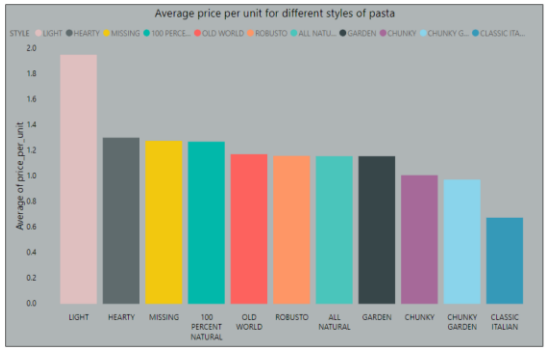
GROUP 5

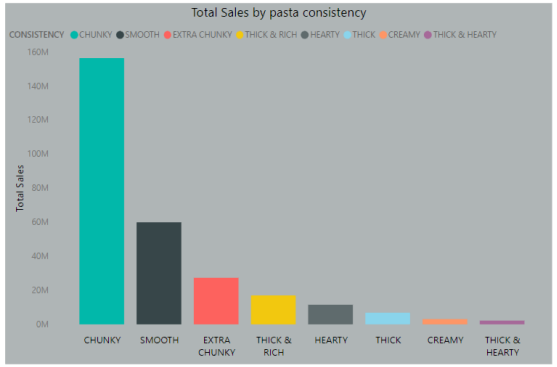
Executive Summary

Our project aims to provide marketing insights for the Prego brand of spaghetti sauce. We conducted store-wise and household analysis of 6.5 million rows to answer three essential marketing analytics questions -

* We used RFM to determine loyalty of customers to specific brands. Using that loyalty score, we conducted logistic regression and analyzed household characteristics to predict customer-behavior.
* We conducted survival analysis to determine the no. of months after which a customer will churn.
* We employed linear regression to determine what factors (display features, promotions, consistency, style) affect sales of spaghetti sauce.

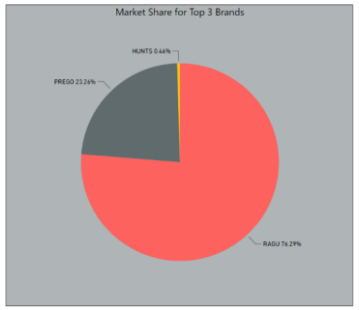


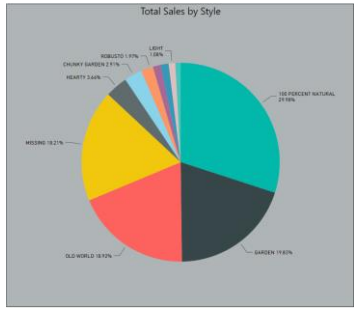


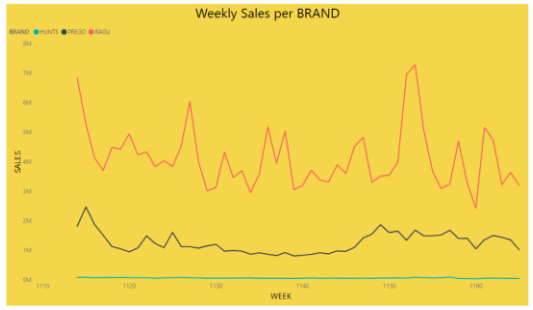


EXPLORATORY DATA ANALYSIS

- Top 3 brands (Overall based on Total Sales across all Stores): **Ragu>Prego>Hunts**
- Market share of Ragu is the highest at 76% generating >200M sales, followed by Prego and Hunts
- 29.98% of the total sales of spaghetti sauce came from the style “100 PERCENT NATURAL” followed by Garden at 19.83%
- While the total sales were good, we wanted to check which style had the highest price per unit. Our analysis revealed that the “LIGHT” style had the highest price per unit and “CLASSIC ITALIAN” had the lowest price per unit.
- An analysis of the sales based on “CONSISTENCY” of the sauce showed that “CHUNKY” had the highest sales while “THICK & HEARTY” had the lowest. In terms of price per unit, “CREAMY” was the most pricy while “THICK & HEARTY” was the least.







Performing RFM to identify Loyal Customers

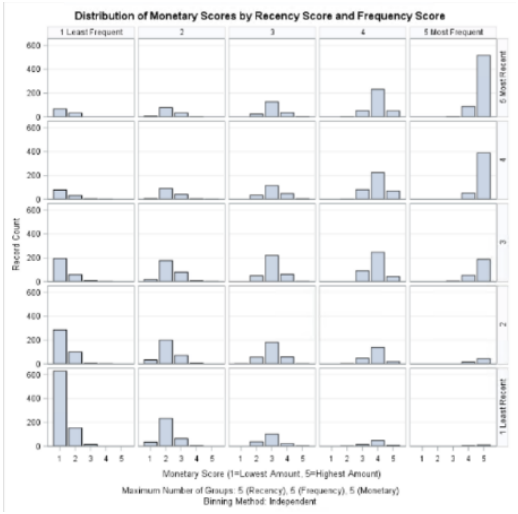
- The distribution of the Sum of Transactions amount by Recency & Frequency is accumulated towards the right
- From the output of the Correlation, we could see the Frequency and the Monetary and highly correlated. Hence, we are neglecting the Frequency over Monetary as money is important factor
- Assigning highest value of 5 to each, we ranked customers based on the sum of individual scores for the two factors(R & M). Those who scored 8 or above on with a minimum of 4 for each - recency and monetary - were considered loyal.

Break-down of Frequency of Customers buying a Brand

The FREQ Procedure

Table of BRAND by recency_score

Frequency / Row Pct / Col Pct	recency_score (Recency Score (1=Least Recent, 5=Most Recent))					
BRAND	1	2	3	4	5	Total
FRANCESCO	95 7.69 8.50	162 13.11 11.81	240 19.42 13.04	377 30.50 19.73	362 29.29 18.05	1236
PREGO	388 13.34 34.74	498 17.12 36.30	662 22.76 35.98	644 22.14 33.70	717 24.65 35.76	2909
RAGU	634 15.46 56.76	712 17.37 51.90	938 22.88 50.98	890 21.71 46.57	926 22.59 46.18	4100
Total	1117	1372	1840	1911	2005	8245



Break-down of Frequency of Customers buying a Brand

The CORR Procedure

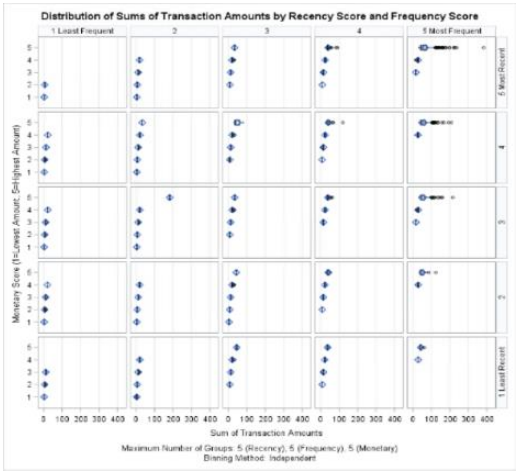
3 Variables: recency_score frequency_score monetary_score

Simple Statistics

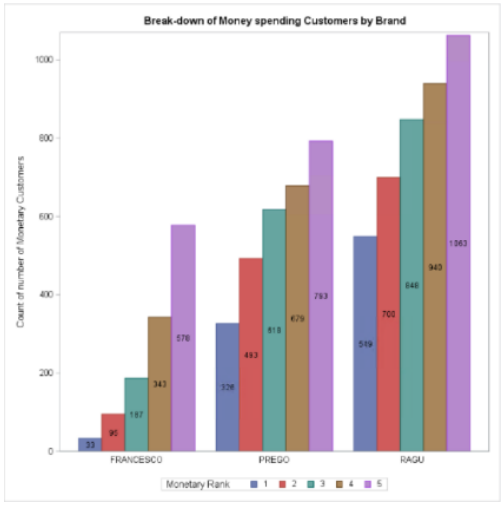
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
recency_score	6704	2.99045	1.41186	20048	1.00000	5.00000	Recency Score (1=Least Recent, 5=Most Recent)
frequency_score	6704	2.93929	1.47344	19705	1.00000	5.00000	Frequency Score (1=Least Frequent, 5=Most Frequent)
monetary_score	6704	2.99970	1.41442	20110	1.00000	5.00000	Monetary Score (1=Lowest Amount, 5=Highest Amount)

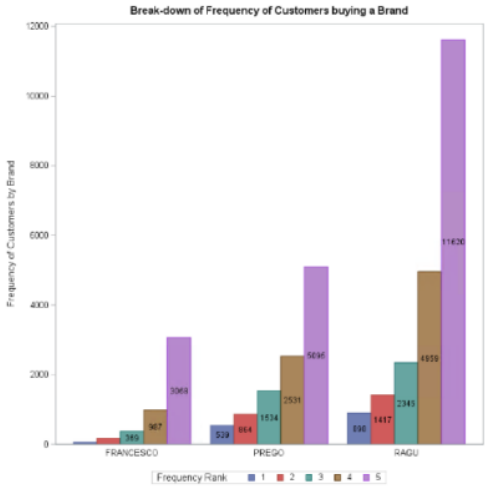
Pearson Correlation Coefficients, N = 6704
Prob > |r| under H0: Rho=0

	recency_score	frequency_score	monetary_score
recency_score Recency Score (1=Least Recent, 5=Most Recent)	1.00000	0.55845 <.0001	0.54424 <.0001
frequency_score Frequency Score (1=Least Frequent, 5=Most Frequent)	0.55845 <.0001	1.00000	0.92315 <.0001
monetary_score Monetary Score (1=Lowest Amount, 5=Highest Amount)	0.54424 <.0001	0.92315 <.0001	1.00000



- 9% of the customers have a RM score of 10
- We are considering RM > 8 and if R>4, M>4, then customer is Loyal else Not loyal
- 35.82% of the customers are the most valuable customers as they have RM score>=8
- Most frequently bought brand is RAGU followed by Prego and Francesco
- Count for number of Monetory customers is highest for Ragu followed by Prego and Francesco





Break-down of Money spending Customers by Brand

The FREQ Procedure

rm_score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
2	421	7.59	421	7.59
3	494	8.91	915	16.50
4	581	10.47	1496	26.97
5	608	10.96	2104	37.93
6	738	13.30	2842	51.23
7	718	12.94	3560	64.18
8	704	12.69	4264	76.87
9	756	13.63	5020	90.50
10	527	9.50	5547	100.00

Customer Loyalty Traits Using Regression

We applied economic theory, subject knowledge and utilized the assistance of statistical selection models like stepwise that could predict whether a particular customer would be loyal or not. Model with the smallest AIC, SC and -2 Log L is considered the best model.

Logistic Regression with all possible demographics which we obtained after removing few redundant and irrelevant demographics.

Intercept value of 2 Log L= 31179.530 (without X variables)
Intercept and covariates value of 2 Log L = 28850.529 (with X variables)
So here the improvement of model = McFadden's R-square
= diff. in (-2LogL)/Null model's (-2logL)
= (31179.530 - 28850.529)/ 31179.530 = 7.46%
= Therefore, **R-square value is 7.46%** which signifies improvement in the model

Compared to Intercept only (Null model i.e no X variable) the AIC and BIC Decreased for intercept and Covariates. The Best Model is with lower AIC and BIC.
Compared to the Null model, the model with explanatory variables have reduced the AIC and SC(BIC) which shows there is an improvement in the model.

% of Concordance: Concordance tells how the model is when paired with events than with that a no events and also how good it is predicting within the model. The value of predicting the random pair of loyal customer than not loyal customer is 68.3%. This means when running the model with new data set we can at least expect accuracy of 68.3% in prediction.

Naive Ratio: 7148/(7148 + 19803)= 26.52% and Number of Correct Prediction= 20128
Hit Ratio is No. of correct prediction/(Total predictions) so **Hit ratio** =20128/26951= 74.68%
There is an improvement of 48.16% from the naive ratio.
The **Area under Curve** also shows that the model correct prediction 68.3%.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	31181.530	28976.529
SC	31189.731	29493.241
-2 Log L	31179.530	28850.529

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2329.0007	62	<.0001
Score	2201.5976	62	<.0001
Wald	1939.2782	62	<.0001

Association of Predicted Probabilities and Observed Responses

Percent Concordant	68.3	Somers' D	0.366
Percent Discordant	31.7	Gamma	0.366
Percent Tied	0.0	Tau-a	0.143
Pairs	141551844	c	0.683

Response Profile

Ordered Value	Loyal	Total Frequency
1	1	7148
2	0	19803

Probability modeled is Loyal=1.