Final Project

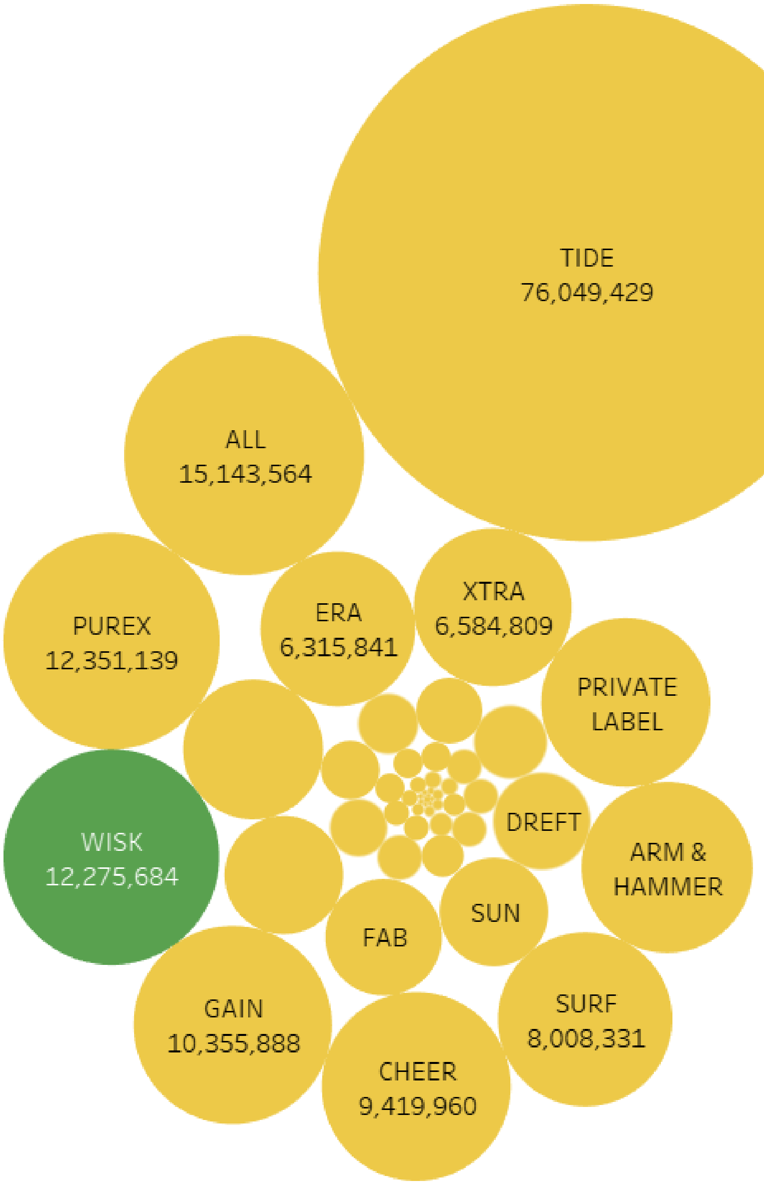
PREDICTIVE ANALYTICS USING SAS

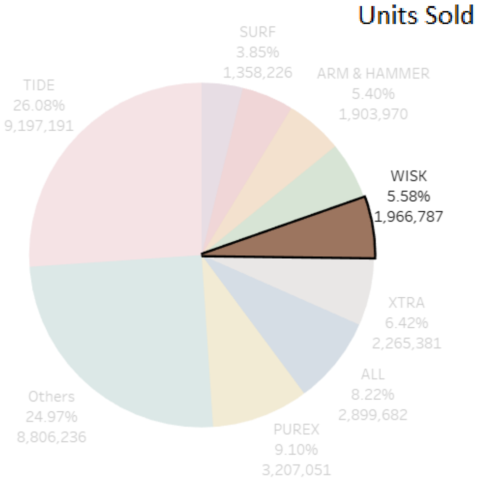
Rama Vidya Reddy
Sushanth Chintalapati
Sanjukta Shreekant Ajinkya
Sai Sumanth Chilukuri
Roshni Arul

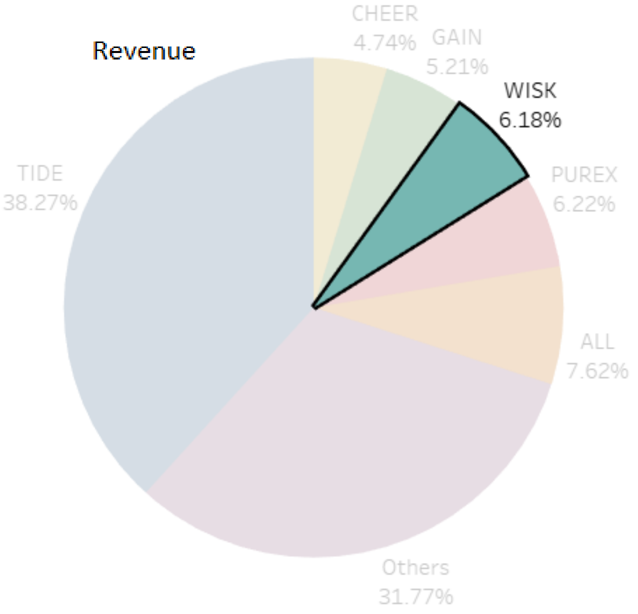
Group 10

UNIVERSITY OF TEXAS AT DALLAS | SPRING 2020

Descriptive Analysis
Market Share







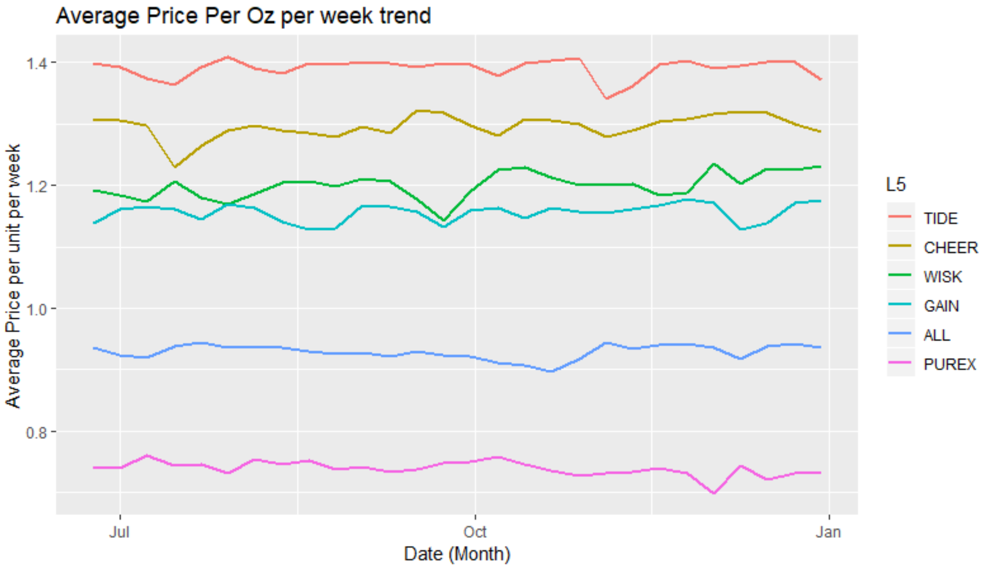
From the above plots it is clear the Tide is the market leader with 38.27% share in the market and selling over 9 Million units. On the other hand, our brand "WISK" has a market share of 6.18% having sold close to 2 Million units and making over 12 Million in dollar sales.

Store Location

Market Location	Dollar Sales
NEW YORK	1,397,801
LOS ANGELES	844,022
CHICAGO	704,554
BOSTON	632,258
WASHINGTON, DC	599,005
PHILADELPHIA	554,809
NEW ENGLAND	462,392
KANSAS CITY	40,350
GREEN BAY	34,039
EAU CLAIRE	32,823
TULSA,OK	23,584
DES MOINES	22,395
SPOKANE	22,234
OKLAHOMA CITY	15,867

The Highlighted regions are the top 7 locations and the others are the bottom 7 locations in terms of the revenue incurred. This makes perfect sense as all the densely populated areas account for the top sales while all the other poorly populated regions account for the lowest contribution to the revenue.

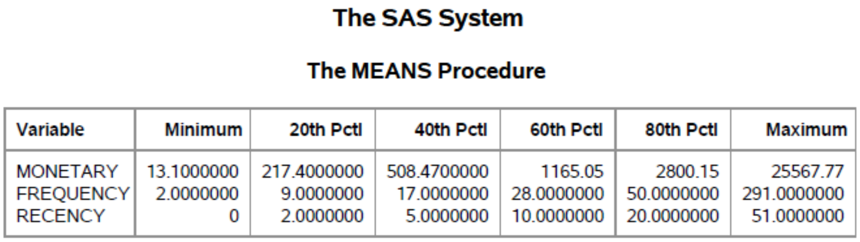
Price (per Oz) variation



From the above it is clear that the price variation per oz. is pretty much the same with time among all the brands, with "WISK" having the highest fluctuations. It is also clear that our brand falls under the average pricing bracket while "TIDE" falls in the high price (premium) and "PUREX" in the lowest priced ones among the top 6 most selling brands.

RFM (Recency, Frequency and Monetary)

RFM stands for Recency, Frequency, and Monetary value, each corresponding to some key customer trait. RFM is a method used for analyzing customer value. These RFM metrics are important indicators of a customer's behavior because frequency and monetary value affects a customer's lifetime value, and recency affects retention, a measure of engagement.
In our data we took a subset of all the laundry detergent brands from the scanner data to get the data about the brand that we were interested in the brand "Wisk".
By performing RFM analysis on the new data, we get the following data:



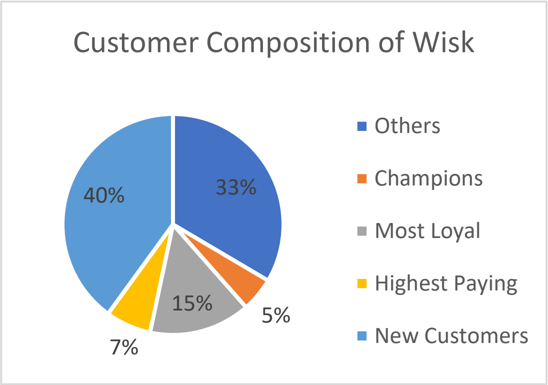
The SAS System

The MEANS Procedure

Variable	Minimum	20th Pctl	40th Pctl	60th Pctl	80th Pctl	Maximum
MONETARY	13.1000000	217.4000000	508.4700000	1165.05	2800.15	25567.77
FREQUENCY	2.0000000	9.0000000	17.0000000	28.0000000	50.0000000	291.0000000
RECENCY	0	2.0000000	5.0000000	10.0000000	20.0000000	51.0000000

Using this we further divided the data into different segments to acquire the different types of customers. By doing this we will be able to determine how to cater to the different customers.

CUSTOMER SEGMENTATION:



- **Champions/Best Customers** (Segment 1):
  These are people who are top priority customers who have a good rapport with the brand overall, which means they are heavy spenders, buy very frequently and also recently.
  Frequency and monetary values fall in the or greater than the 80th percentile region (>50 and >$2800) and recency in the less than 20th percentile region (<2).
  What can be done?
  These are the customers that will promote your brand. So, targeting them for new product launches is a good strategy for wider reach.

- **Loyal Customers** (Segment 2):
  These are the ones who have been buying your brand very frequently and naturally will also be spending more.
  So, if frequency is more than the 60th percentile (>28) then they fall in this category.
  What can be done?
  We can offer special discount coupons or rewards for such repeat customers.

- **Highest Paying** (Segment 3):
  These customers are the ones who spend the most on your brand compared to all the other customers.
  These customers have monetary value which is higher than the 80th percentile (>$2800).
  What can be done?
  These customers are willing to pay high products and are not very affected if prices are high. So, a good idea for such customers can be to offer them premium/luxury products if you have any in your brand.

- **New Customers** (Segment 4):
  These are the ones who have a good overall performance, are recent purchasers but are not very frequent buyers.
  Recency less than 20th percentile (<2), Frequency < 40th (<17) percentile and Monetary value > 60th percentile (>$1165).
  <u>What can be done?</u>
  Build a relationship with such customers by providing them personal assistance and give them offers. These people could become your potential loyal customers.

- **At Risk Customers** (Segment 0/Others):
  Customers who are spent a lot of money and bought often but hadn't made any purchase in the recent times.
  Frequency > 80th percentile (>50), Monetary > 80th percentile (>$2800).
  <u>What can be done?</u>
  Since they haven't made a purchase in a really long time, sending them reminder emails/promotions would be a great strategy to bring them back.

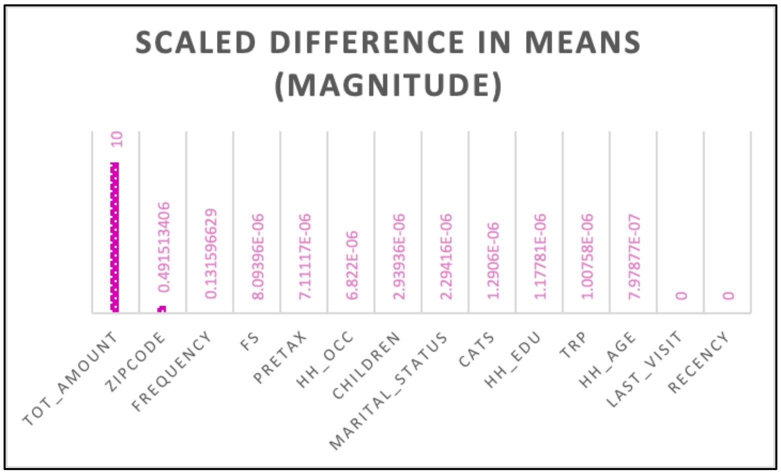
<u>Logit Model Analysis:</u>

While we analyzed the customers who bought Wisk, we wished to differentiate the customers who frequently bought our product (1) and the ones who did not (0).

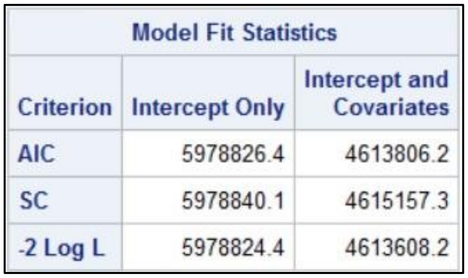
We used a logistic regression here. It is different from simple regression because the only assumption is that the error terms are not correlated. The dependant variable holds a binary levelled response [1/0]; so proc logistic is a good choice of procedure.

The total amount (sum of dollars) and the frequency (count of the week) was assessed. The customers who fell in the 75th percentile (third quartile) were used as a threshold for the creating new dependant variable 'Repeating_customers' and label them as the customers who frequently bought our product (1) and the ones who did not (0).

The independent variables 'pretax', 'FS', 'TRP', Zipcode, 'HH_AGE', 'HH_EDU', 'HH_OCC', 'Cats', Children, Marital_Status were chosen with the difference of means method.

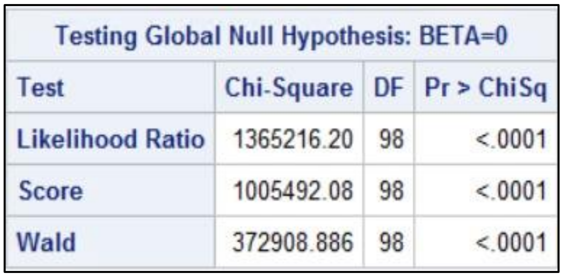


Partial Test



Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	5978826.4	4613806.2
SC	5978840.1	4615157.3
-2 Log L	5978824.4	4613608.2

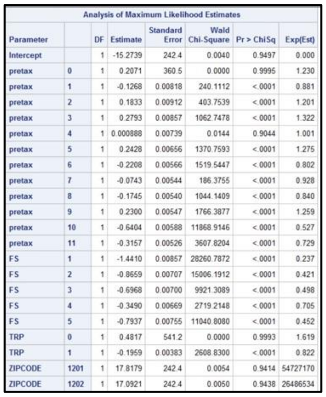
Global Test



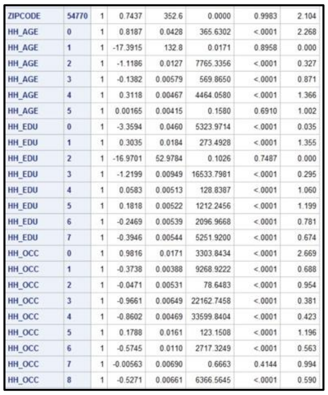
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1365216.20	98	<.0001
Score	1005492.08	98	<.0001
Wald	372908.886	98	<.0001

From the model fit statistics of the proc logistic regression, we see that the model with the explanatory variables performs better than the naïve model (the AIC, SC and -2LogL values are lower for the full model as compared to the null model).

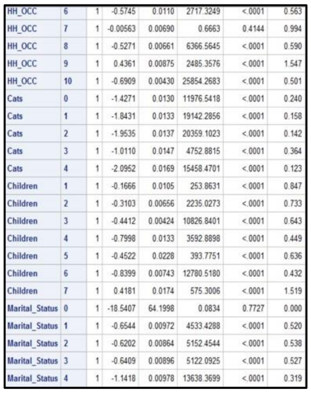
From Testing Global Null Hypothesis, our model is significant even at the 0.0001% level.



Analysis of Maximum Likelihood Estimates							
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept		1	-15.2739	242.4	0.0040	0.9497	0.000
pretax	0	1	0.2071	360.5	0.0000	0.9995	1.230
pretax	1	1	-0.1268	0.00818	240.1112	<.0001	0.881
pretax	2	1	0.1833	0.00912	403.7539	<.0001	1.201
pretax	3	1	0.2793	0.00857	1062.7478	<.0001	1.322
pretax	4	1	0.000888	0.00739	0.0144	0.9044	1.001
pretax	5	1	0.2428	0.00656	1370.7593	<.0001	1.275
pretax	6	1	-0.2208	0.00566	1519.5447	<.0001	0.802
pretax	7	1	-0.0743	0.00544	186.3755	<.0001	0.928
pretax	8	1	-0.1745	0.00540	1044.1409	<.0001	0.840
pretax	9	1	0.2300	0.00547	1766.3877	<.0001	1.259
pretax	10	1	-0.6404	0.00588	11868.9146	<.0001	0.527
pretax	11	1	-0.3157	0.00526	3607.8204	<.0001	0.729
FS	1	1	-1.4410	0.00857	28260.7872	<.0001	0.237
FS	2	1	-0.8659	0.00707	15006.1912	<.0001	0.421
FS	3	1	-0.6968	0.00700	9921.3089	<.0001	0.498
FS	4	1	-0.3490	0.00669	2719.2148	<.0001	0.705
FS	5	1	-0.7937	0.00755	11040.8080	<.0001	0.452
TRP	0	1	0.4817	541.2	0.0000	0.9993	1.619
TRP	1	1	-0.1959	0.00383	2608.8300	<.0001	0.822
ZIPCODE	1201	1	17.8179	242.4	0.0054	0.9414	54727170
ZIPCODE	1202	1	17.0921	242.4	0.0050	0.9438	26486534



ZIPCODE	54770	1	0.7437	352.6	0.0000	0.9983	2.104
HH_AGE	0	1	0.8187	0.0428	365.6302	<.0001	2.268
HH_AGE	1	1	-17.3915	132.8	0.0171	0.8958	0.000
HH_AGE	2	1	-1.1186	0.0127	7765.3356	<.0001	0.327
HH_AGE	3	1	-0.1382	0.00579	569.8650	<.0001	0.871
HH_AGE	4	1	0.3118	0.00467	4464.0580	<.0001	1.366
HH_AGE	5	1	0.00165	0.00415	0.1580	0.6910	1.002
HH_EDU	0	1	-3.3594	0.0460	5323.9714	<.0001	0.035
HH_EDU	1	1	0.3035	0.0184	273.4928	<.0001	1.355
HH_EDU	2	1	-16.9701	52.9784	0.1026	0.7487	0.000
HH_EDU	3	1	-1.2199	0.00949	16533.7981	<.0001	0.295
HH_EDU	4	1	0.0583	0.00513	128.8387	<.0001	1.060
HH_EDU	5	1	0.1818	0.00522	1212.2456	<.0001	1.199
HH_EDU	6	1	-0.2469	0.00539	2096.9668	<.0001	0.781
HH_EDU	7	1	-0.3946	0.00544	5251.9200	<.0001	0.674
HH_OCC	0	1	0.9816	0.0171	3303.8434	<.0001	2.669
HH_OCC	1	1	-0.3738	0.00388	9268.9222	<.0001	0.688
HH_OCC	2	1	-0.0471	0.00531	78.6483	<.0001	0.954
HH_OCC	3	1	-0.9661	0.00649	22162.7458	<.0001	0.381
HH_OCC	4	1	-0.8602	0.00469	33599.8404	<.0001	0.423
HH_OCC	5	1	0.1788	0.0161	123.1508	<.0001	1.196
HH_OCC	6	1	-0.5745	0.0110	2717.3249	<.0001	0.563
HH_OCC	7	1	-0.00563	0.00690	0.6663	0.4144	0.994
HH_OCC	8	1	-0.5271	0.00661	6366.5645	<.0001	0.590

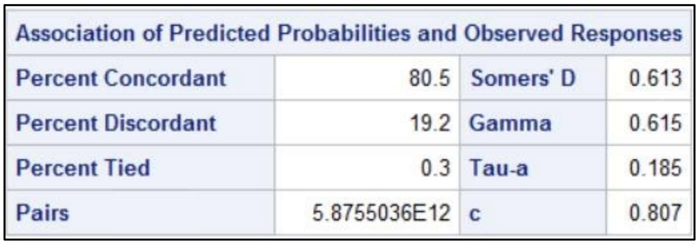


HH_OCC	6	1	-0.5745	0.0110	2717.3249	<.0001	0.563
HH_OCC	7	1	-0.00563	0.00690	0.6663	0.4144	0.994
HH_OCC	8	1	-0.5271	0.00661	6366.5645	<.0001	0.590
HH_OCC	9	1	0.4361	0.00875	2485.3576	<.0001	1.547
HH_OCC	10	1	-0.6909	0.00430	25854.2683	<.0001	0.501
Cats	0	1	-1.4271	0.0130	11976.5418	<.0001	0.240
Cats	1	1	-1.8431	0.0133	19142.2856	<.0001	0.158
Cats	2	1	-1.9535	0.0137	20359.1023	<.0001	0.142
Cats	3	1	-1.0110	0.0147	4752.8815	<.0001	0.364
Cats	4	1	-2.0952	0.0169	15458.4701	<.0001	0.123
Children	1	1	-0.1666	0.0105	253.8631	<.0001	0.847
Children	2	1	-0.3103	0.00656	2235.0273	<.0001	0.733
Children	3	1	-0.4412	0.00424	10826.8401	<.0001	0.643
Children	4	1	-0.7998	0.0133	3592.8898	<.0001	0.449
Children	5	1	-0.4522	0.0228	393.7751	<.0001	0.636
Children	6	1	-0.8399	0.00743	12780.5180	<.0001	0.432
Children	7	1	0.4181	0.0174	575.3006	<.0001	1.519
Marital_Status	0	1	-18.5407	64.1998	0.0834	0.7727	0.000
Marital_Status	1	1	-0.6544	0.00972	4533.4288	<.0001	0.520
Marital_Status	2	1	-0.6202	0.00864	5152.4544	<.0001	0.538
Marital_Status	3	1	-0.6409	0.00896	5122.0925	<.0001	0.527
Marital_Status	4	1	-1.1418	0.00978	13638.3699	<.0001	0.319

Interpreting Significant Coefficients:

- *The odds of a customer frequently purchasing our brand increases 1.322 times for a house with pre-tax income range $12,000 to $14,999 per year as compared to a house with a combined per-tax income of $ 100,000 and greater income per year.*
- *The odds of a customer frequently purchasing our brand increases 0.705 times for a family of four people as compared to a family of 6 or more people.*
- *The odds of a customer frequently purchasing our brand increases 1.619 times when the type of residential possession is owned as compared to rented.*
- *The odds of a customer frequently purchasing our brand increases 1.366 times when the age of the household is 45-54 as compared to the households with age 65+.*
- *The odds of a customer frequently purchasing our brand increases 1.355 times when the household education is some graduate school or less as compared to household education of post graduate work.*
- *The odds of a customer frequently purchasing our brand increases 2.669 times when the household as an occupation as 'other' as compared to a household which is not employed.*
- *The odds of a customer frequently purchasing our brand increases 0.364 times when the customer has three cats vs. a customer having 5+ cats.*

- *The odds of a customer frequently purchasing our brand increases 1.519 times when the household has children in age group (0-5), (6-11) and (12-17) as compared to family size>0 yet no children.*
- *The odds of a customer frequently purchasing our brand increases 0.538 times when the the customer is married vs. when the customer is separated.*



Association of Predicted Probabilities and Observed Responses			
Percent Concordant	80.5	Somers' D	0.613
Percent Discordant	19.2	Gamma	0.615
Percent Tied	0.3	Tau-a	0.185
Pairs	5.8755036E12	c	0.807

80.7% of the times our model will correctly sort a repeating customer from a non-repeating customer.

In the random pairs of observations from the group of customers who do not buy our brand frequently and the group of customers who buy our brand frequently, the predicted probability (phat) for the observation from the frequent customer group should be greater than the phat for the observation from the non-frequent customer group. The percent concordant for every observation pair in our model where p-hat (1) > p-hat (0) is 80.5%.

Managerial Recommendations:

In the above interpretations, the odds of being a frequent customer were observed. These were the customers who were above our threshold set in the 75th percentile. From the same table we can see which groups of customers are buying the detergent 'Wisk' less frequently. This group of customers obtained from our model should be targeted by the management. These customers are the ones in the 2nd quartile.
For example, some of our desired target groups could be; households with children in the age groups (6-11) & (2-17), households with a family size of 3, households with an occupation- craftsman or customers who are divorced.