**GROUP: 10**

**Shweta Siddha, Manisha Gupta, Achint Khanijio, Kartikay Nigam**

**Predictive Analytics on Frozen Pizza Dataset**

**OVERVIEW**

The data used in the report is grocery store scanner data of frozen pizza consumption. The analysis has been performed as Brand manager of FRESCHETTA to accelerate the sales of frozen pizza. We performed a competitive analysis amongst top performing pizza brands to bring business values and insights. And accordingly, the recommendation can be made.

**DATA PREPROCESSING:**

Before preparing models on the given dataset, we processed the data to bring them to same granularity. The data for product, grocery, delivery store, demographics and product company are available for across weeks and across panels. To bring the data to consistent level, all the files are merged and summarized at best possible levels.

**PROBLEM STATEMENT**:

* + - 1. As a brand owner of FRESCHETTA, we wanted to understand market share of other competitors and study how demographics affect brand choices.
      2. Analyzing customer using Segmentation based on RFM model and understanding customer clusters to strategically focus on target clusters.
      3. Time series analysis of weekly sales data and forecasting future sales to take necessary corrective and preventive actions.

**Case #1: Market share and Brand Preference analysis based on customer demographics and product promotions.**

To understand the factors influencing the brand preference of customers based on customer demographics. A multinomial logit model is used to predict the preference of brand choice, in this case we considered FRESCHETTA, DI GIORNIO, TOMBSTONE and rest all brands are categorized as OTHERS. Reasons for choosing these brands is DI GIORNIO is quite similar to FRESHETTA but has higher sales, as a brand manager of Freshetta we should analyzing the brand preference for one close competitor like DI GIORNIO and another with highest sales like TOMBSTONE.

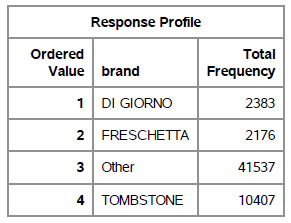
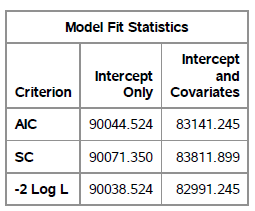
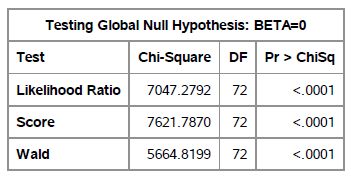
From the results, we can see that DI GIORNIO, TOMBSTONE and FRESHETTA are among the top brands in terms of dollar sales. To analyze the effect of demographic factors we ran a Multinomial Logistic regression model.

**MULTINOMIAL LOGIT MODEL**:

Baseline or reference categories for categorical variables:

Brands : Other HH Income Level: Level\_1 Gender : Female

Age : ‘65+’ Family Size: Size 1

**MODEL FIT STATISTICS:**

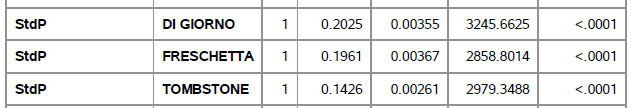
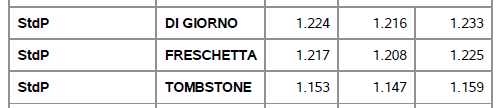
**AIC**: This is the Akaike Information Criterion. AIC, like Adjusted R-square in linear regression, penalize the log-likelihood for the number of predictors in the model. Ultimately, the model with the smallest AIC is considered good.

**SC**: This is the Schwarz Criterion. Like AIC, SC penalizes for the number of predictors in the model and the smallest SC is most desirable.

In the output above, the likelihood ratio chi-square of 7047.27 with a p-value < 0.0001 tells us that our model fits significantly better than an empty model (i.e., a model with no predictors).

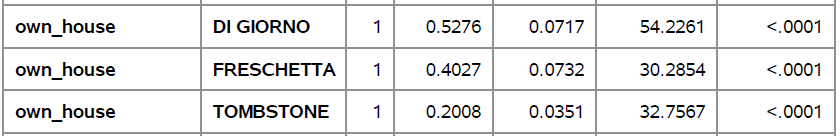
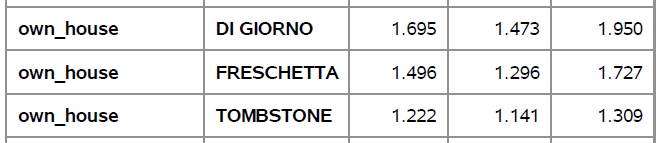
Based on all the above three criteria AIC, SC and -2Log L, we can see that the values of the model with intercept and covariates is less than the intercept only criterion. Therefore, the model which we used is a better fit than the intercept only model.

**Interpretation of coefficients**:

**StdPrice**: A one-unit increase in the variable StdPrice is associated with a 0.20 unit increase in the log odds of choosing brand FRESHETTA as compared to all other brands.

A one unit increase in StdPrice (weekly price of brand) increase the odds of choosing Freshetta over other brands by 21%.

**Type of Residential Possession**: It is quite interesting to see that whether the person is the owner of the house or living in rented house, is significant variable for all the pizza brands. The relative log odds of choosing brand Freshetta vs. Di Giorno, Tombstone and other brand will increase by 0.40 units if a person has his own home as compared to rented home. Hence, it’s odds ratio of choosing Freshetta increase by 49% if a person has his own home as compared to rented home.

**Age**: From the coefficients of age in categories: Age\_18\_24, Age\_25\_34, Age\_35\_44, Age\_45\_54, Age\_55\_64 and Age\_65+, we can see that a unit increase in the age in Age\_18\_24, odds of choosing brand Freshetta increases by 55% as compared to other brands.

**Gender**: If the gender is male, the odds of choosing the Freshetta as compared to other brands will decrease by 2.7%.

**HouseHold income level**: From the coefficients of House Hold income level in categories from level 2 to level 12, 1 unit increase in HH\_IncomeLevel\_12, odds of choosing brand Freshetta increases by 102.9% as compared to other brands and odds of choosing brand Di Giorno increases by 173.4%. It may be possible that Freshetta and Di Giorno are not significantly different, but on comparing their standard deviation and parameter estimates, as the difference in parameters is not 2 times more than standard deviation, we can say that choice of Di Giorno is not significantly preferable than Freshetta in high income group.

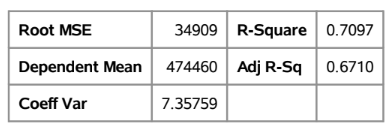
**FamilySize:** From the coefficients of Family Size level in categories from level 2 to level 6, in most cases the odds ratio is decreasing except for FamilySize\_2, where if the Family\_size\_2, the odds of choosing brand Freshetta increases by 8% as compared to Di Giorno, Tombstone and other brands.

**RECOMMENDATION**:

1. Demographics factors positively affect the choice of brands depending on the customer segmentation, but their McFadden R2 is only 0.07 (7%). The model with McFadden's values of 0.2 to 0.4 is a good fit. Hence, we can say that demographics does not significantly explain the choice of brand.
2. Increasing weekly price of pizza is comparatively more beneficial for FRESHETTA as compared to other brands. So, we can recommendation to business that we can increase the prices a little, such that the average prices are still less than that of Di Giorno’s. As we will still have the cheaper product of the two, the increase in price will not affect the demand of our pizzas and hence there will be a boost in the profits.
3. Customers in Age\_18\_24 prefer Freshetta the most in comparison to all other brands and all other age groups within Freshetta.
4. Freshetta is the brand most preferred in Household with lowest income level ('$10,000 to $11,999 per yr), highest in the income level 12 ('$100,000 and greater per year).
5. As most of the customer demographics are categorical it is quite difficult to check for their non linearity effect and interaction effect in this model.

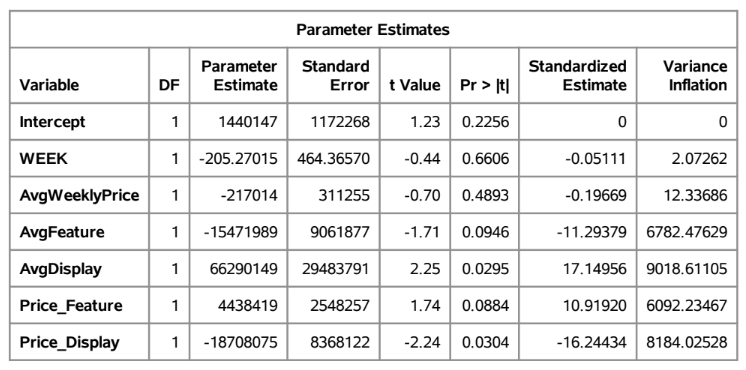
**Understanding market share from product data**:

**LINEAR REGRESSION Model**:



From the above results we can see that using product data we are able to explain 67% variation in the dollar sales which is reasonably good.

**Interpretation of coefficients**:



Null Hypothesis: Beta coefficient of all independent variable is zero.

Alternate Hypothesis: Beta coefficient of at least one independent variable is not zero.

From the above model we can observe that as the p-value for coefficient for Average Display and interaction variable between price and display is less than 0.05. Therefore, their coefficients are significant.

**Price Elasticity** = 217014 \* 3.5351 / 474459.7

= 1.6169

A $1 increase in price would decrease the sales by 217014 units.

**Display**: If there is a price promotion in the form of display, sales would increase by 66290149 units.

**Heteroskedasticity analysis:** VIF for some variables is greater than 10 and Condition Index greater than 100, so yes there is multicollinearity in the model.

**Case#2: Segmentation of customer based on RFM model**

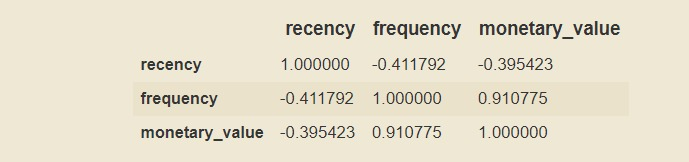
To quantitively segregate the customer based on their past behavior and build strategies based on target customer, it is very important to understand their purchase patterns. One of the methods to segment the customer is RFM (Recency Frequency and Monetary value) model.

**Determining the value of Recency, Frequency and Monetary value**

**RECENCY (R):** Days since last purchase  
**FREQUENCY (F):** Total number of purchases  
**MONETARY VALUE (M):** Total money this customer spent

1. For brand FRESHETTA, the data set containing the purchase information of each panel ID across week was filtered.
2. The following metrics were calculated
   1. Recency by Week– Since recency is calculated for a point in time, and the last WEEK is 1165, we will use 1166 to calculate recency
   2. Frequency– How often a customer purchased? Count of the number of weeks of purchase for each customer.
   3. Monetary – How much a customer spent for the product? – Sum of all the dollar amount paid by a customer made during purchase across weeks in the dataset.
3. Four segments are created based upon the frequency distribution and univariate results using quartile value of 0.25,0.5 and 0.75.
4. Each customer based upon their recency, frequency and total money spent on purchase were tagged with Recency, Frequency and Monetary Scores with ‘1’ being highest and ‘5’ being lowest.
5. Lowest recency, highest frequency and monetary amounts are our best customer.

**Correlation between RFM matrices**:



From the above matrix, we can see that there is high correlation between frequency and monetary value. It is quite reasonable as more and more customer visit the store, they might generate more revenue. So, while analyzing the customer segments based on clusters formed from the three matrices; this fact needs to be taken into consideration.

As per the analysis, below are the top 10 customers:



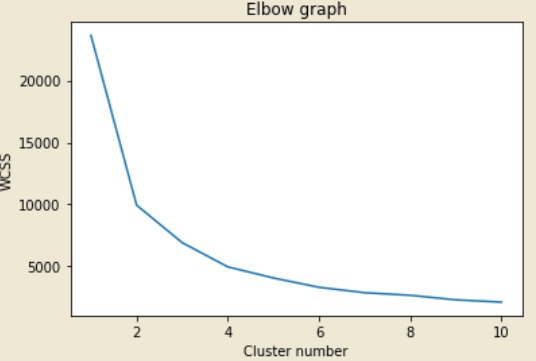
**Interpretation**

* Panel Id 3338764 has frequency: 71, monetary value: $556.83 and recency: 1 week.
* Panel Id 3316943 has frequency: 92, monetary value: $549.05 and recency: 1 week and so on.

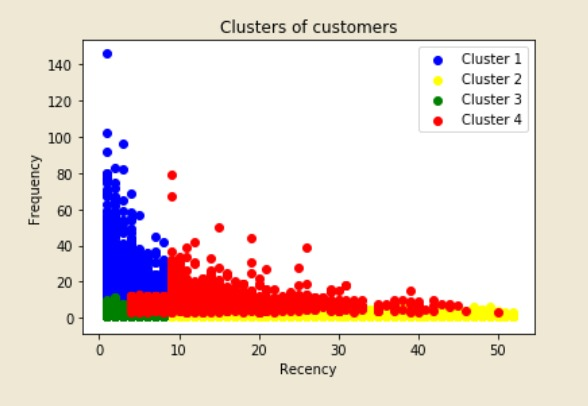
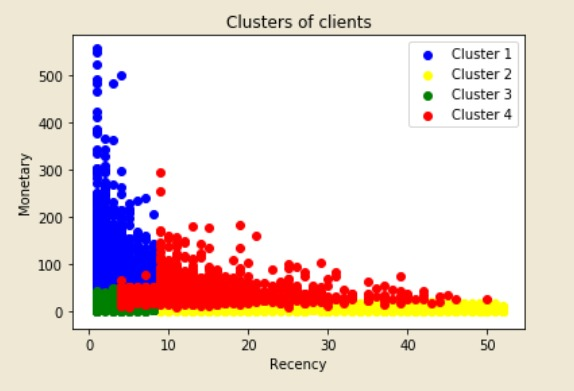
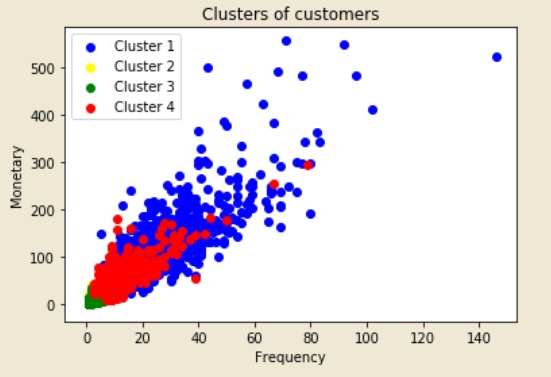
**CLUSTER ANALYSIS**:

Our main goal will be to group these panel ids in customer segments by using the clustering technique: K-Means.

Using elbow curve to determine number of clusters:



Our Elbow graph shows that an optimal number of clusters would be 4. So, we have created 4 cluster based on the score value generated for RFM matrices.

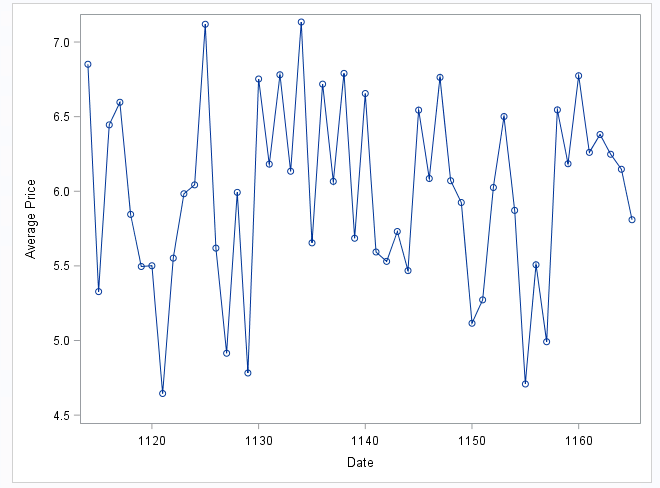
  

**RECOMMENDATIONS**:

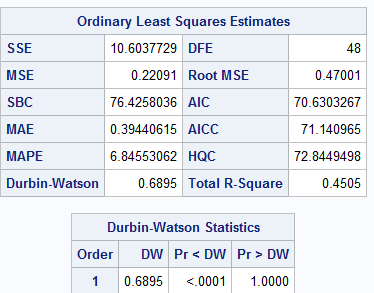
1. If we know our most valued customer, we can plan to design promotional or any special offers for targeted customers.
2. It will help in identifying the most potential customers.
3. It can help in improving the quality of service, loyalty and retention.
4. Improve customer relationship based on better understanding needs of segments.
5. It can help in designing new products and enhancing existing products as per customer needs.
6. Can provide opportunities for upselling and cross selling.

**Case #3: Time Series Analysis of Dollar sales over weeks**

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



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.



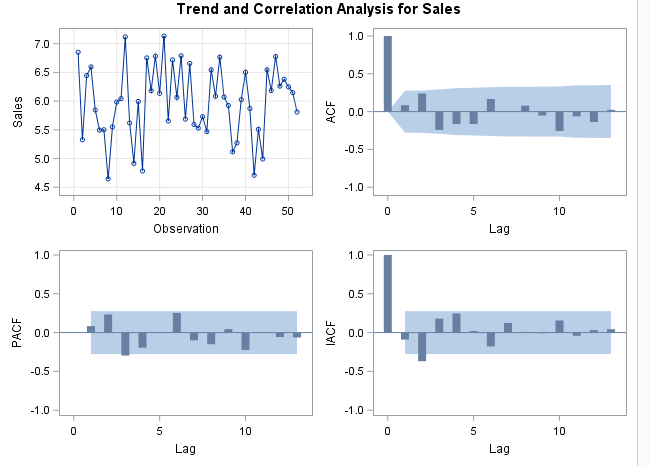
As per the Durbin-Watson Statistics, we can see that the value is 0.7 (range of 0 to 2). This signifies the positive auto-correlation between the target variable at time t with the lag value at kth interval.

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

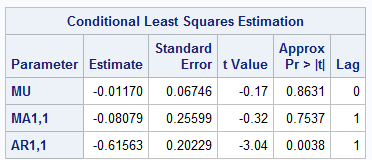
H0: There is presence of a unit root

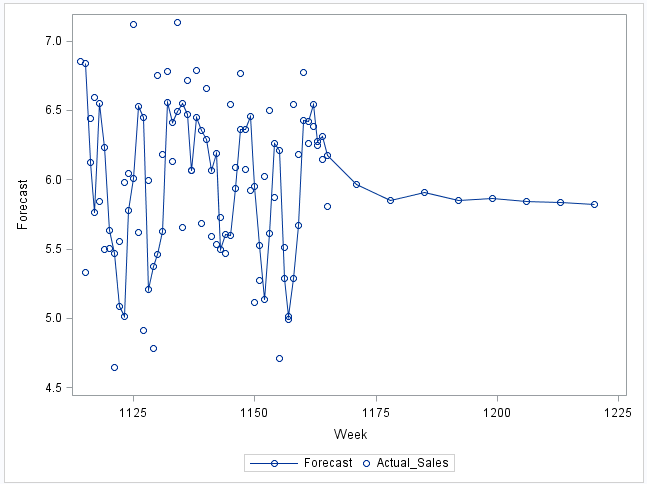
H1: Time-series is either stationarity, trend stationarity or explosive root depending on the test used.

From DF- test, we get the p-value as 0.001. At 5% of significance level, we reject the null hypothesis.



As we can see, AR1,1 fits better on our data





From the forecasted sales, we can see that the sales for Freschetta Frozen pizza is expected to be stationary in the coming months. The sales will rise followed by the decline in price.