**Homework 2**

**Group 10 Submission**

Achint Khanijo, Kartikay Nigam, Shweta Siddha, Manisha Gupta.

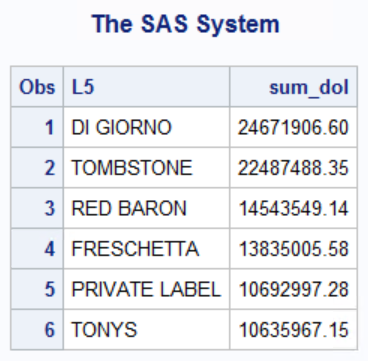
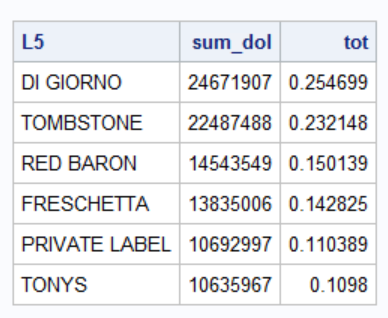
Using the store level scanner data provided to you in class (filename will be similar to spagsauc\_groc\_1114\_1165). You need to link UPC from this file with information in the prod\_sauce.xls file to get brand names etc.

Prepare a report that summarizes the data. At a minimum, it should answer the following questions.

Summarize data: No need for statistical testing for these questions.

1. What are the top 6 brands in the category in terms of dollar sales? What are the market shares of the 6 brands (assuming there are only 6 brands in the market).

Top six brands in terms of dollar sales are Di Giorno, Tombstone, Red\_Baron, Freschetta,Private\_Label and their market shares are 25%, 23%, 15%, 14%,11% and 10% respectively. In terms of market share captured by these top brands, Di Giorno and Tombstone emerges as big players. The combined market share of Di Giorno and Tombstone is approx. 50% of the market share captured by top 6 brands.

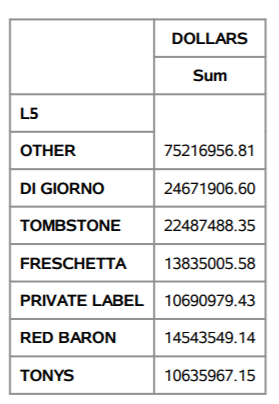
 

1. Which companies are the major players in the category? Which company owns which brands?

Kraft Foods, Schwan Food, General Mills, Private Label, Nestle USA INC and Pinnacle Foods are the top six companies in terms of dollar sales and number of units sold. Both Kraft food and Schwan Food are capturing 65% of the total market.

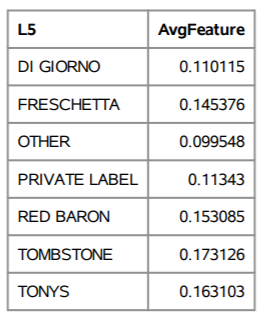
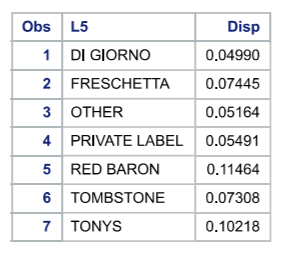
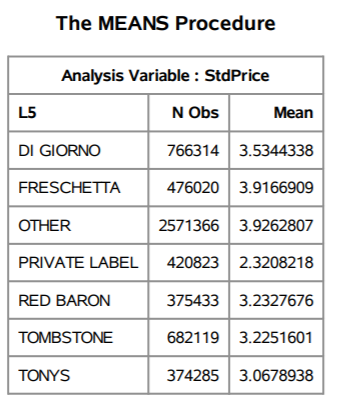
 

1. Create a 7th brand called “Other” that has all other brands that are not in the top 6.



The new category “other” contains 75216956.81 dollar revenue.

1. Find average prices, display, features of each of the 7 brands.



Most retailers these days use various promotional activities like placing items for display, advertisement of products and customer centric offers. From the above results we can observe that how these promotional activities and average price of a product affect the sales of commodities and their significance in improving the sales of brands.

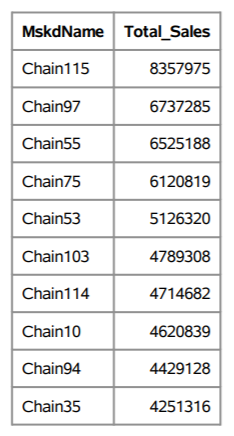
From the analysis in question 1 we observed that DI GIORNO brand has highest dollar sales, but it does not have highest average price of their frozen pizza, neither he has kept their items for display most number of times nor they have featured their products the most. But the fourth highest dollar sales brand FRESCHETTA has higher average price and a greater number of times their products for display and feature as compared to top sales brand.

1. What are the top 5 regions in terms of dollar sales?



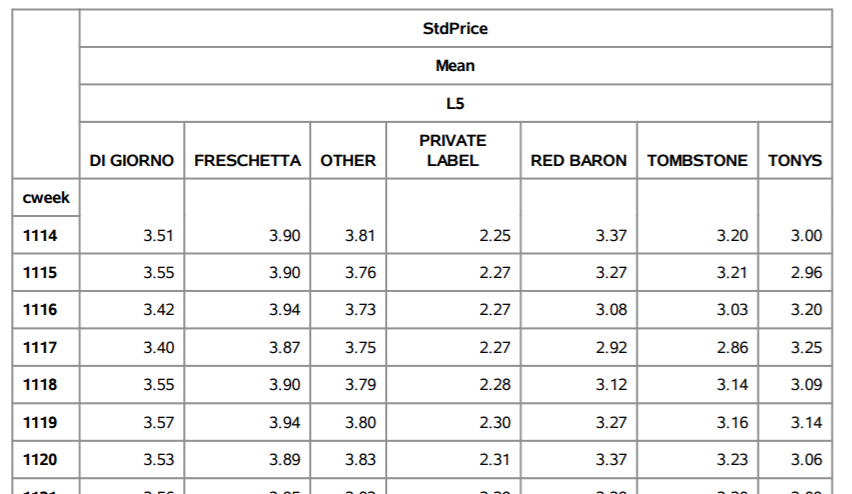
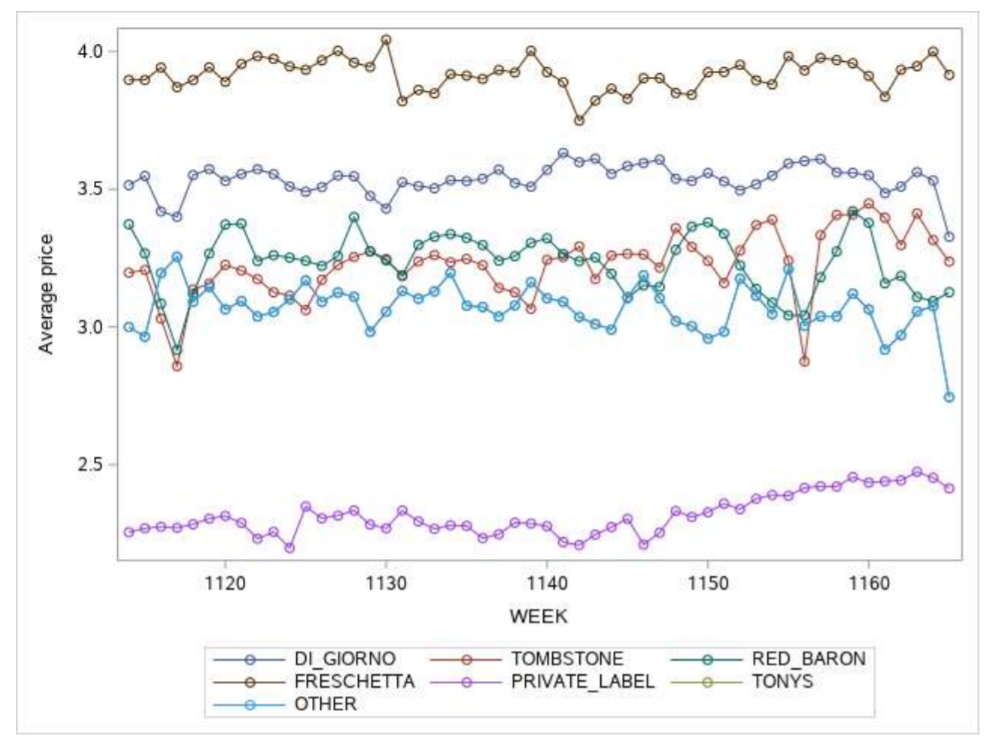
The top 5 regions or market for frozen pizza sales are Chicago, St Louis, Los Angeles, Milwaukee and New York, out of which Chicago has the highest dollar sales.

1. What are the top 10 store chains that sell a lot of your category in terms of dollar sales?



This analysis can help companies to understand which chains to focus more on, when planning for distribution of their products. This chains analysis can also help to making better decisions by concentrating more on regions and stores connected to these chains. The Chain115 has highest dollar sales followed by Chain97. This will help companies to tailor their product distribution accordingly.

1. What is the average price per unit of 7 brands by week? Plot the average price by week (I wish to see a line plot of price by week). Comment on your findings.

This is the analysis of 52 weeks of data for top 6 brands including others. The above trendline which is a plot between average price of all 7 brands plotted against 52 weeks shows that there is an upward trend in average prices over the weeks. The highest average price is in week 1159 to 1160, which November 12 to November 25, 2001. This is same week when Thanksgiving Day occurs. This can lead to conclusion that the prices are little hiked during the festive season of Thanksgiving as there is high foot-fall during this time of the month.

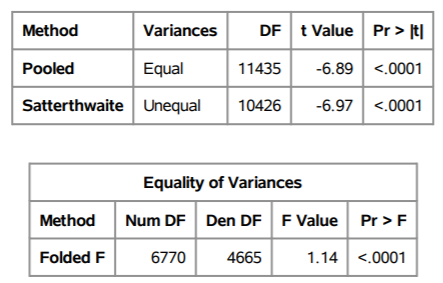
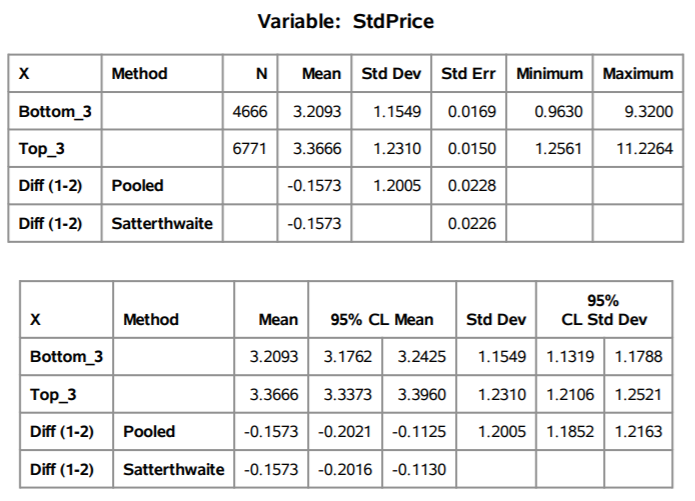
1. Assume you are manager of a brand (out of the top 6). Write a short paragraph stating what you learned from this descriptive analysis (steps 1-7).

As a brand manager of Freshchetta, we enjoy the 4th position among top 6 brands. We captured overall 8 % of the market whereas the top 2 big player - Di Giorno and Tomstone occupied more than 27% combined of the market alone, capturing 14% and 13% market share. The gap between dollar sales for top brand Di Giorno, and Freshchetta is huge. Our brand Freshchetta’s total sales is $10.8 million less than Di Giorno. Freshchetta deployed aggressive marketing and heavy discounting . As compared to Di Giorno, Freshchatta has 2% higher average display time and they offered 3% more features as opposed to Di Giorno.

Our brand Freshchetta ‘s avg price per week are higher than the top 6 brands but still its performance as compared to others are not outstanding. With these prices, it could have become the price leader. There might be a case that our company spending more on operational, marketing and using heavy discounting.

Statistical Analysis

1. Do large stores (top 3 stores) have higher average price per unit than small stores (stores ranked 8-10) for brand 1 (the top brand in Q1). Test and report your results and comments.



We performed t test to check whether top 3 brands has higher average price per unit is higher than bottom 3 brands. Our hypothesis testing are-

Ho: The avg price per unit of top 3 and bottom 3 brands are same.

Ha: The avg price per unit of top 3 and bottom 3 brands are not same.

To check the above on, we performed variance test where our null hypothesis is variance of avg price are same for both groups and alternative is variance in avg price per unit for both groups are not same. Since p value is less than 5%, we reject the null hypothesis and conclude that variance of avg price are not same for both the group.

Now, we checked the result of unequal group where p value is less than 5% we reject our null hypothesis i.e. The avg price per unit of top 3 and bottom 3 brands are same. With 95% confidence, we can say that avg price per unit of top 3 and bottom 3 are not same.

As the p value is less than alpha, we reject the null hypothesis, hence there is enough evidence to conclude that the variances are not equal.

Ha: Average of top3 <= Average of Bottom3

Ha: Average of top3 > Average of Bottom3

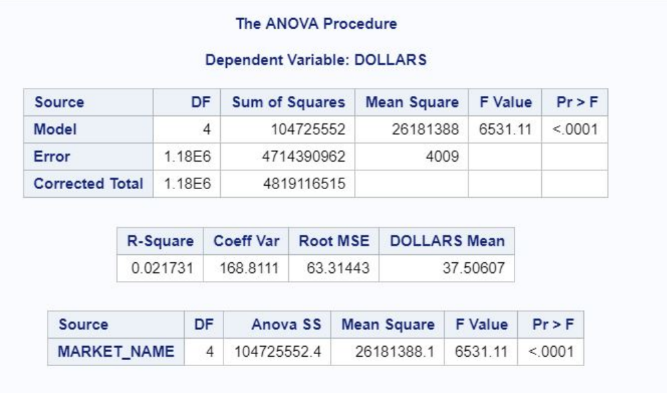
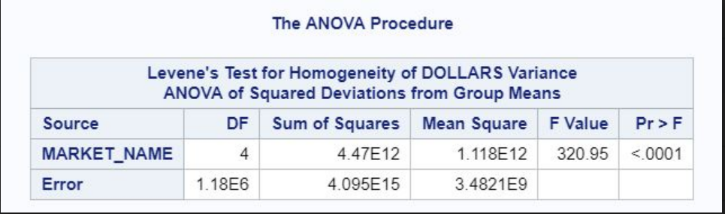
As the p-value is less than alpha (0.05), hence we reject the null, hence there is enough evidence to conclude that the average price of top 3 stores is higher than the average price of bottom 3 stores.

1. Develop three additional hypotheses linking useful variables to dollar sales, test them and report your findings.

Analysis 1 – Is there any difference between weekly sales in top 5 regions by dollar sales.

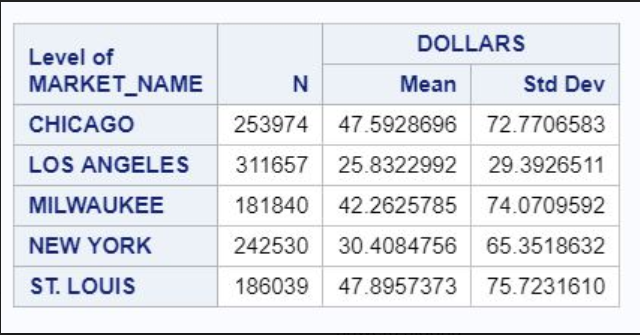
We conducted an ANOVA test to find the answer.   
Ho: There is heterogeneity Ha: There is no heterogeneity

Firstly, we can conclude that there is homogeneity in the variance of the 5 regions data based on Levene’s Test for Homogeneity.

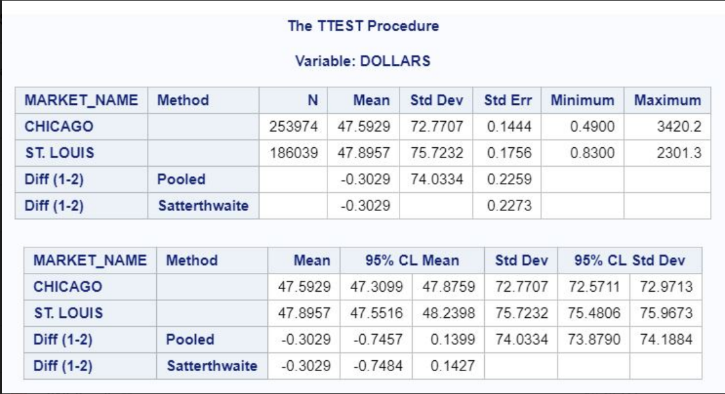
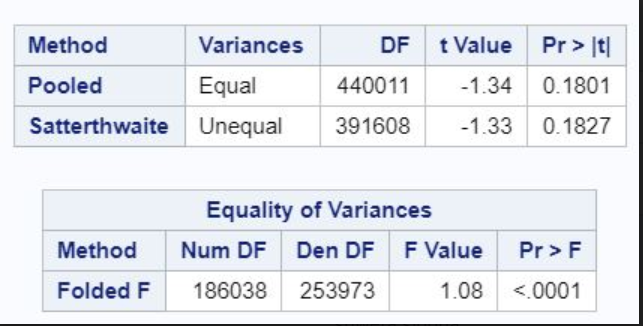


Then, based on the ANOVA output above, we can conclude that at least sales in two regions are different from each other

Now, looking at the means of Weekly Sales from the 5 regions, we see that sales of Chicago and St. Luis are relatively higher:



Now, comparing the Sales of Chicago and St. Louis, we do a t-test to find out where the sales are higher.  
Ho: Sales are not different from each other Ha: Sales are different from each other

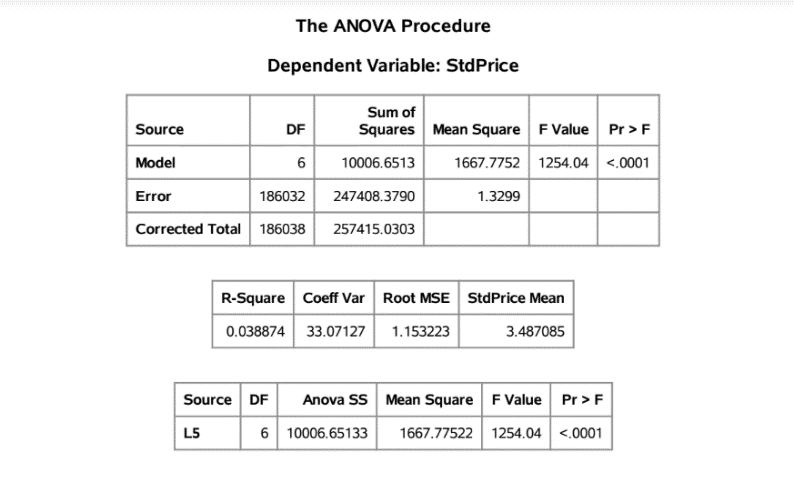
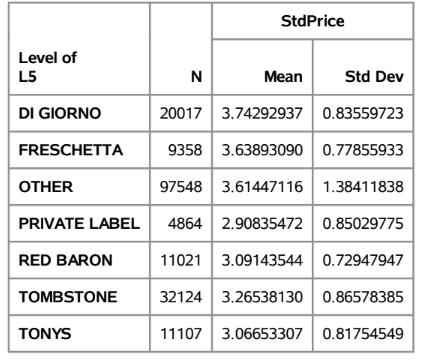
 

So, the sales in St Louis and Chicago are not statistically different from each other.

Analysis 2 – In one of the high sales markets, we try to find if the Average Weekly Price of Top 7 brands are different from each other.

Ho: Avg Price is same for all brands in St. Louis

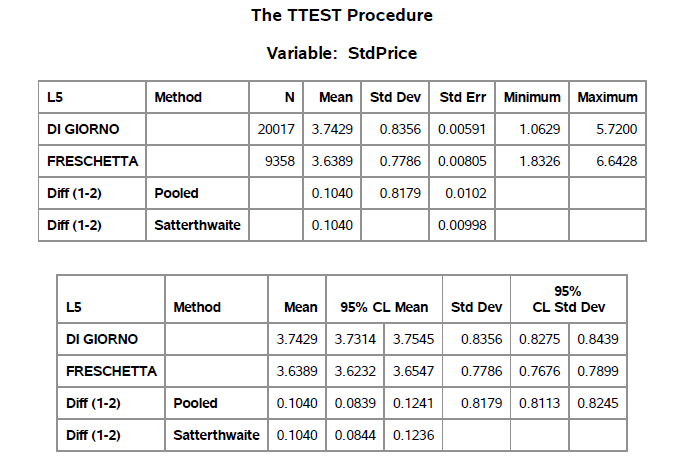
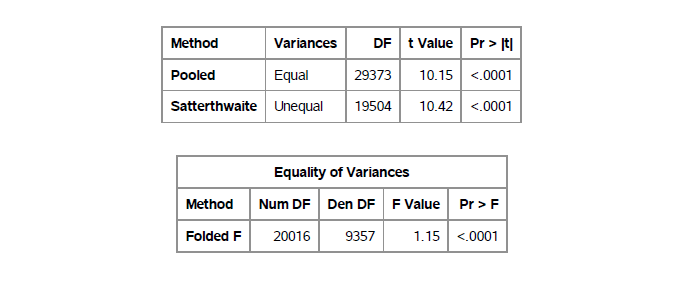
Ha: Avg Price is different for at least two brands in St. Louis

From the above two outputs, we see that the Avg Prices are clearly different for different brands.

Now let us compare the Avg Price of our brand, Freschetta with the Di Giorno which has the highest average price.

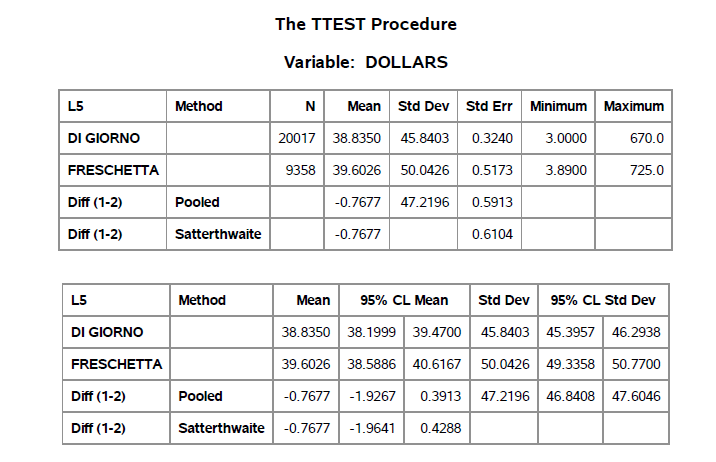
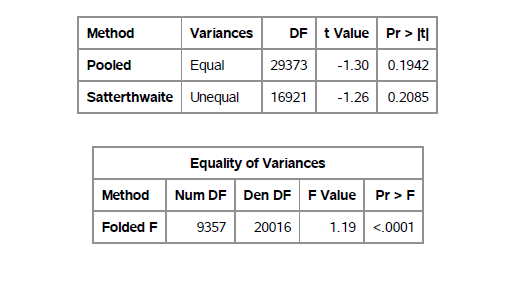
Ho: Avg Prices are same between the two brands in St Louis   
Ha: Avg Prices are different from each other in St. Louis

We can conclude that the average prices are different. Looking at the means, Freschetta which is our brand has cheaper prices than Di Giorno.

Analysis 3 – Are the sales of Freschetta and Di Giorno different from each other in St. Louis.

Ho: Dollar Sales are same between the two brands in St Louis   
Ha: Dollar Sales are different from each other in St. Louis

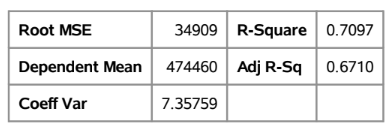
Based on the t-test, the dollar sales of our brand are not significantly different from Di Giorno’s dollar sales.

As we earlier concluded that the avg price of Freschetta is less than that of Di Giorno’s and now we know that dollar sales are still not different. This means that we are selling more units than Di Giorno.

So, our recommendation to business is 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.

*Note*: *The exact amount by which the prices should be increased would need more data and analyses*

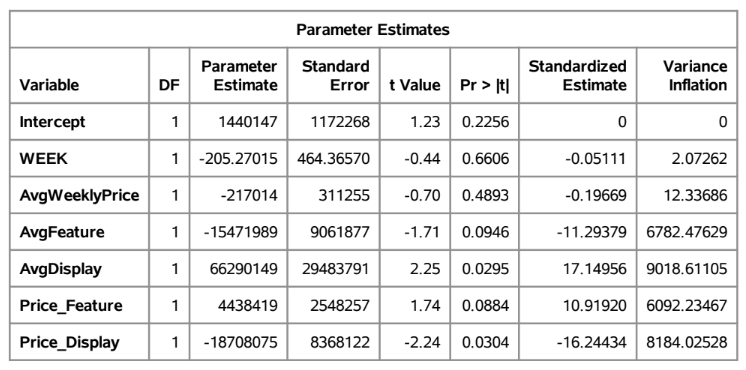
1. For the top brand: run a regression model with weekly dollar sales as dependent variable. Use average weekly price per unit, average display, average feature, and other useful variables in your regression and answer the following questions:
2. What is the R-sq and adj R-sq of the model?



R2 =0.71 implies that all the 6 explanatory variables including the interaction variables explain 71% of the variance in the dependent variable Total Weekly Sales. R2 lies in a range between 0 and 1 with values closer to 1.00 indicating a very good model fit. However, R2 =0.71 can be considered reasonably good model.

R2 has a property that it increases with the increase in the number of explanatory variables in the model. This is a misleading assumption to explain the variance in dependent variable with more independent variables. Therefore, adjusted R2 imposes a penalty on any variable added to the model that has a very small explanatory power. In the above regression model Adjusted R2 is 0.67. It implies that all the 6 explanatory variables including the interaction variables explain 67% of the variance in the dependent variable.

1. Which coefficients are significant?



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

Alternate Hypothesis: Beta coefficient of atleast 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.

1. Which variables are most important in explaining sales?

As the coefficient of Average display and interaction term for price and display are significant. Therefore, price and display are most important in explaining sales.

1. Interpret the meaning of the price coefficient? What is the price elasticity?

A $1 increase in price would decrease the sales by 217014 units. This is inferred from the negative sign on the coefficient, which suggests that as the price increases, sales would decrease.

Price elasticity of demand is defined as the percentage change in quantity corresponding to a percentage change in price.

PE = 217014 \* 3.5351 / 474459.7 = 1.6169

1. Interpret the meaning of the display coefficient?

Display is a variable that takes the value 1 or 0 depending on whether there is a promotion in the form of display in the store or not, during that week. The interpretation of the coefficient of a Display is that when there is a price promotion in the form of display, sales would increase by 66290149 units.

1. Test whether there is an interaction between display, feature and price. Comment on your findings.

As we ran the regression model with the interaction terms between feature and price; and display and price. In that interaction term for price and display is significant at 95% significant level but the interaction term between price and feature is not so significant at 95% confident level.

1. Test whether the effect of price is non-linear. Comment on your findings.

On running the regression model between Price, Price squared term against the weekly average sales, both the terms are not found significant. Hence there is no direct relationship between price and sales in this scenario.

1. Test using VIF and COLLIN whether there is multicollinearity in the model? Comment on your findings.

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

1. Test for presence of heteroscedasticity using White test. Do A WLS if needed. Comment on your findings.

As there is not much change in SSE value with (5.543E10) and without (5.483E10) WHITE test. Hence there is no heteroscedasticity in the model. Also, the error is quite evenly distributed in residual plot.