**MKT 6337.002**

**Predictive Analytics using SAS**

**Homework 1 Solutions**

**Group 2**

Eeshpaul Saggu – ess170530

Sumeet Kumar – sxk170831

Priyanshi Khatri – pxk173330

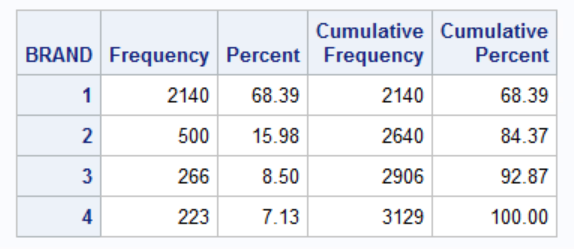
Pooja Arora – pxa170001

Rajdeep Arora– rxa174630

**Homework 1 Solutions:**

1. **How many brands are there in the data and what are their frequencies?**

Ans: There are total 4 brands in the data and their frequencies are listed below:



1. **Create a new variable called ‘name’ that explains the brand name as follows:**

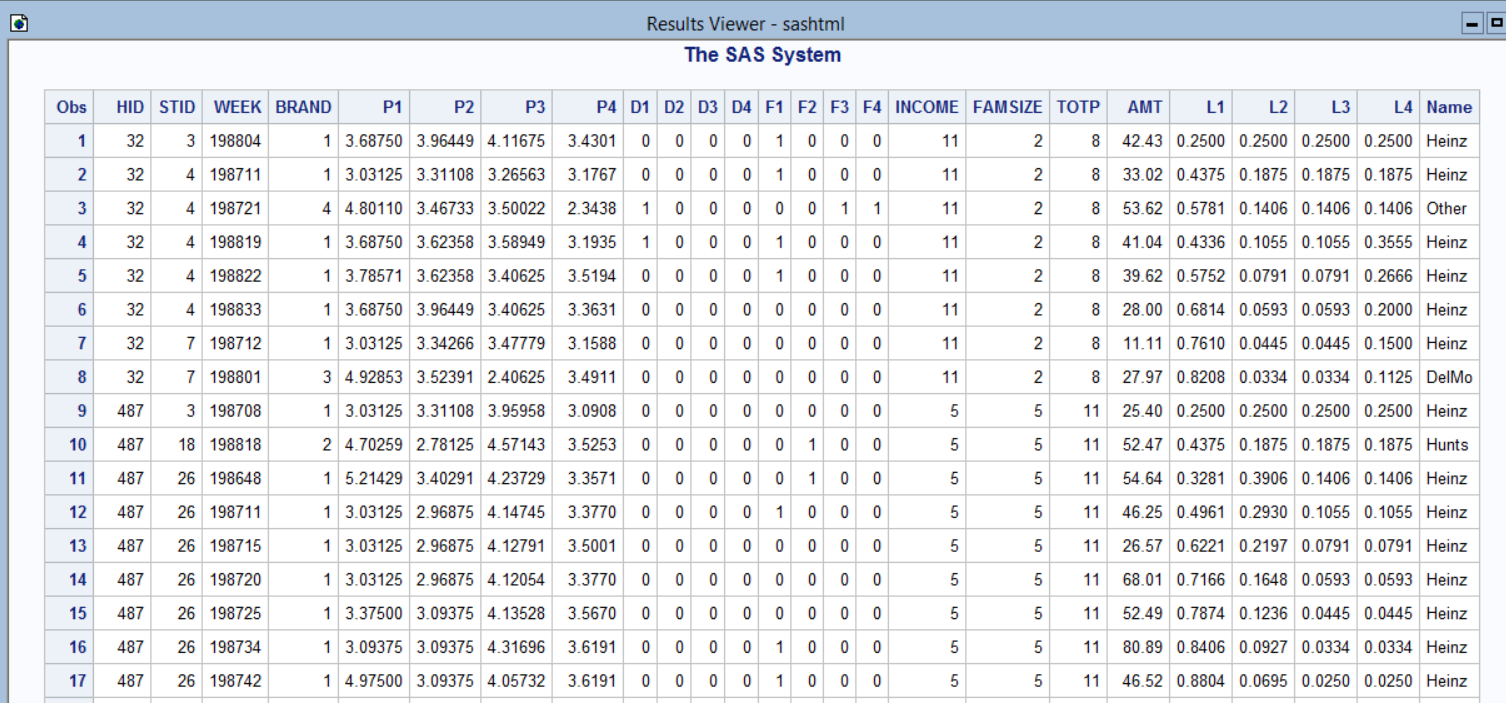
1=Heinz

2=Hunts

3=DelMonte

4=Other

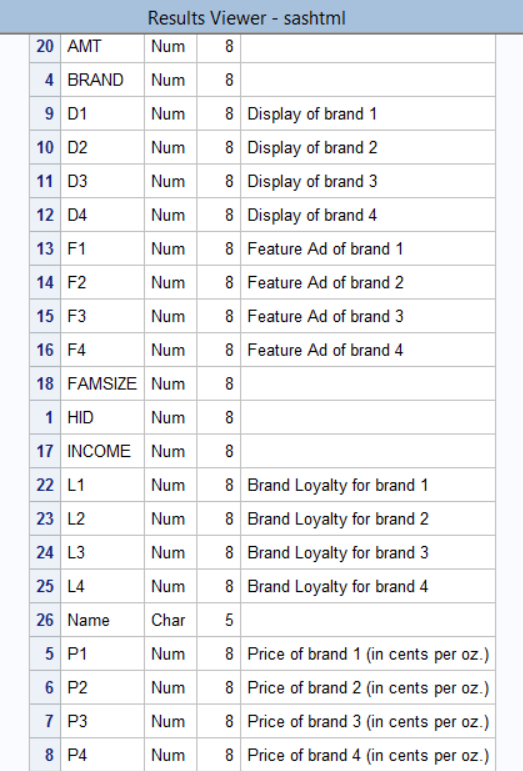
Ans: Below is the output showing a new variable ‘Name’ added to the dataset



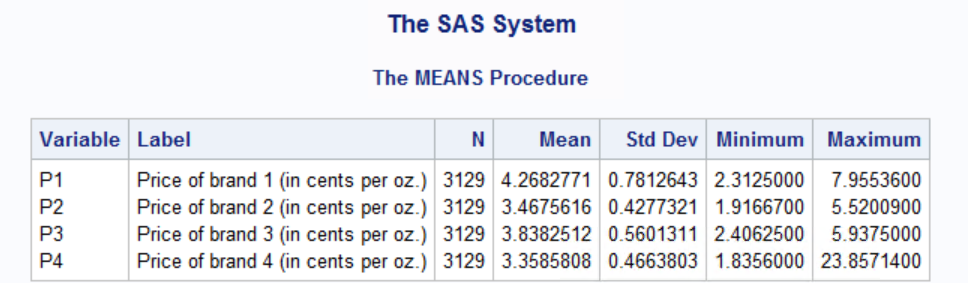
1. **Label p1-p4, d1-d4, f1-f4, l1-l4 to indicate prices, displays, features and loyalty for the four brands.**

Ans: Below is the output showing labels for prices, displays, features and loyalty.



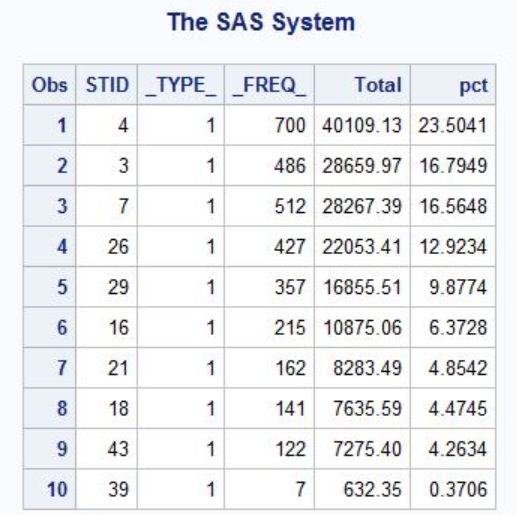


1. **What is the average price of different brands?**

Ans: Below tables shows average prices for all the four brands: -

1. **How many stores are available in the data and what is their share of purchases? Which are the largest stores?**

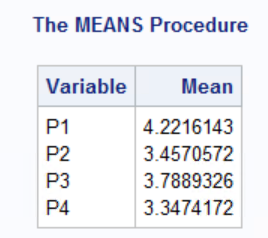
Ans: Based on the output below, there are 10 stores that are available in the data and the largest stores are the ones with STID 4,7 and 3. Their share of purchase is mentioned in the “pct” column.



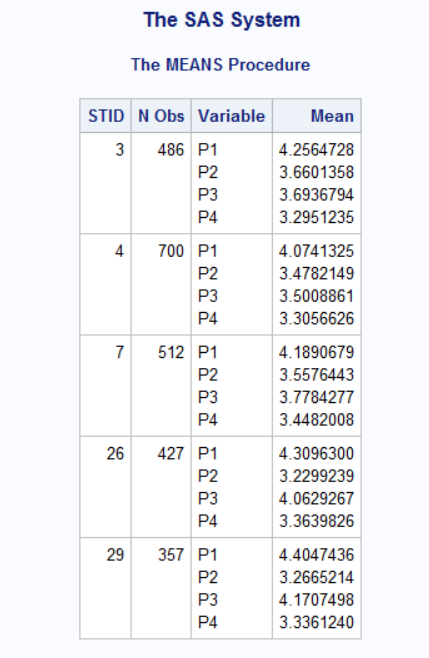
1. **What is the average price of different brands in the top 5 stores?**

Ans: The top 5 brands are the ones with STID 4,7,3,26 and 29. We created a separate dataset with top 5 stores and then calculated the average price.

Average price amongst all top 5 stores: -

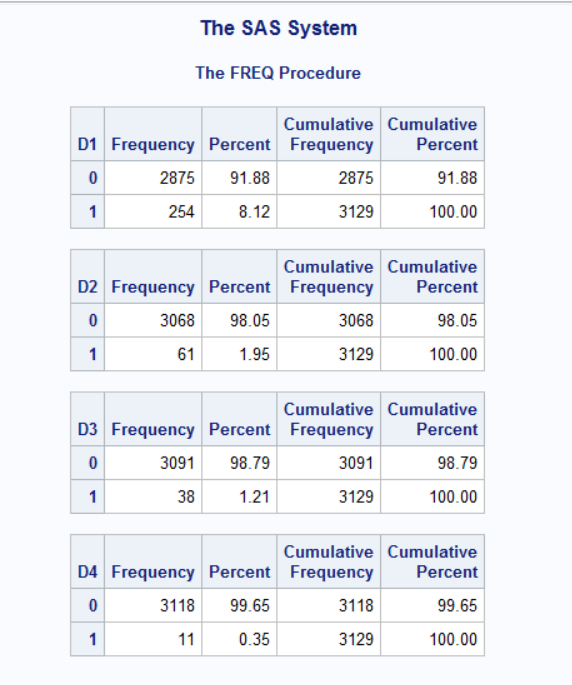


**Average price for each store**: -



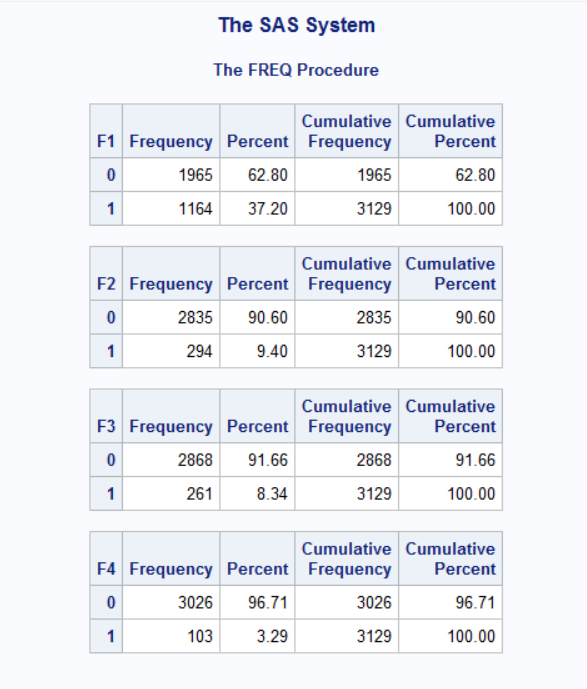
1. **What is the average display (percentage) of different brands?**

Ans: Below tables displays the average percentage display of different brands: -



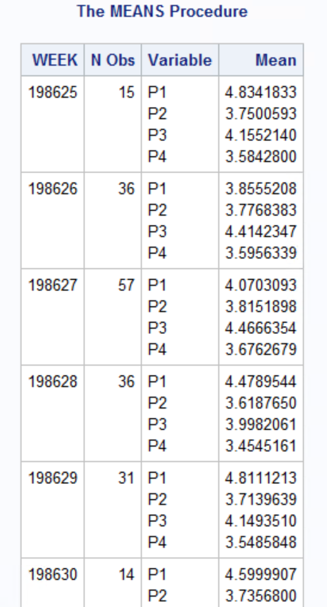
1. **What is the average feature (percentage) of different brands?**

**Ans:** Below tables displays the average percentage feature of different brands: -



1. **What is the average price of brands by week?**

**Ans: -** Weekly Average price of all brands

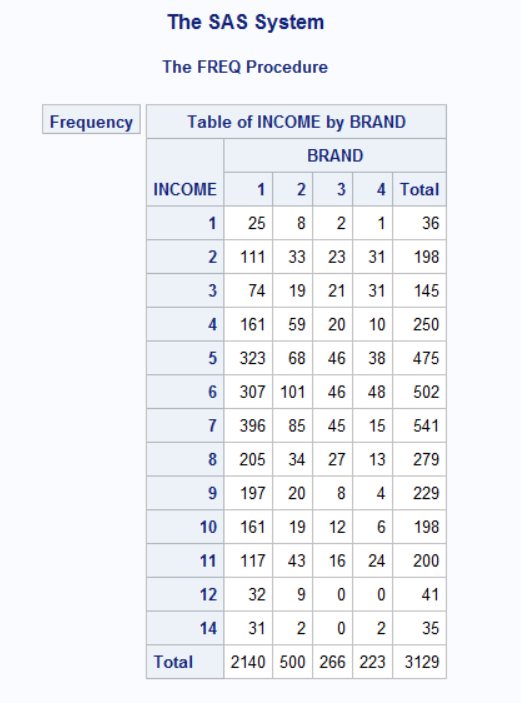


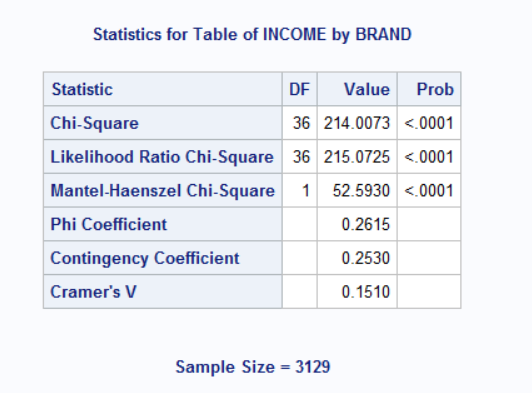
1. **Is there a relationship between income and brands?**

**Ans: -** Based on the dataset, the variables income and brand are categorical and to find the correlation between them, we will use chi-square test.

**H0: Income and brand are not correlated.**

**A0: Income and brand are correlated.**

****

****

As per the results of chi-square test, the probability of the hypothesis being true is less than .0001 so we reject the null and conclude that there is some relationship between the variables income and brand

1. **Is there a significant difference in price paid between large families and small families?**

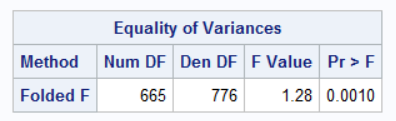
**Ans: -** Defined a new variable Price Which contains the price of the brand chosen by the family. Also, categorize the family size as Small (Famsize =1,2), Medium (Famsize = 3,4,5) & Large (Famsize = 6,7) and name it as FAMTYPE.

To check the significant difference, we use t-test.

To check for **equality of variance**, perform the F-test.

H0: Variances are equal

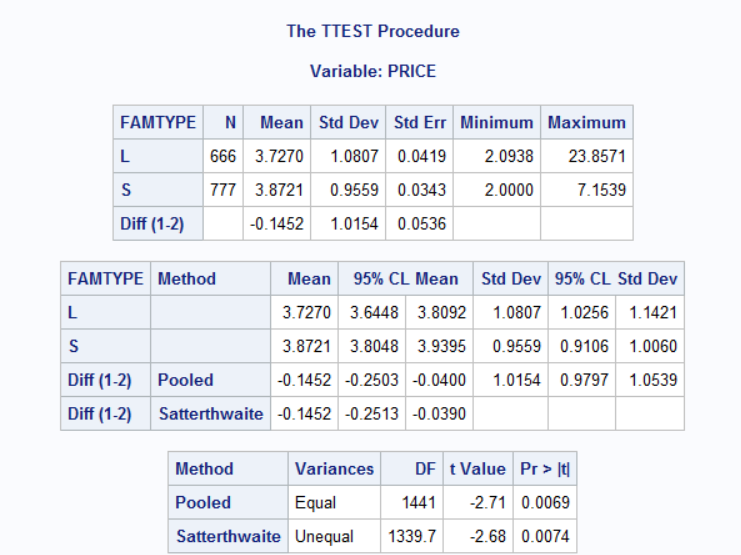
H1: Variances are unequal



Based on this output, the p-value is less than 0.05 so we reject the Null hypothesis. So we can conclude that Variances are unequal.

H0: Average price paid by small families is significantly same to large families

H1: Average price paid by small families is not significantly same to large families



From the output of t-test, we see that p-value is less than 0.05 so we reject the null hypothesis and we can conclude that there is a significant difference between the price paid by Small & large Families.

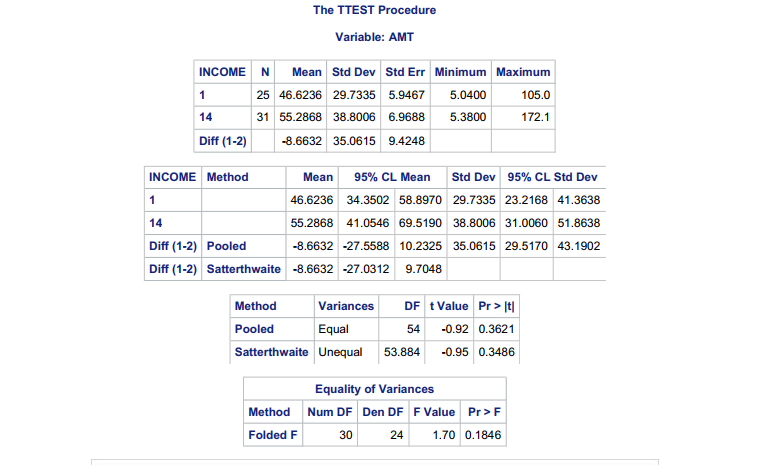
1. **Develop three additional hypotheses using the data, test them and report your findings.**
2. **Hypothesis 1**

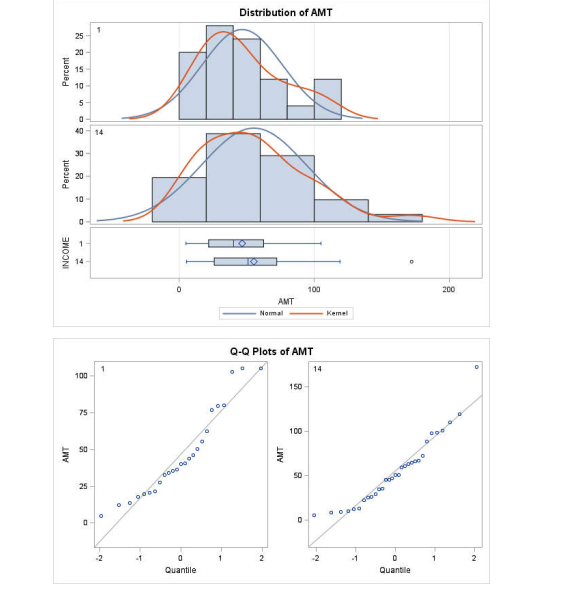
Test whether with higher the income, there is a significant difference expenditure made by households?

**Null Hypothesis** – There is no significant difference between expenditure of higher and lower income group at 95% of Confidence Interval.

**Alternate Hypothesis** – There is a significant difference between expenditure of higher and lower income group at 95% of Confidence Interval.

**Test Results** – As the p-value of F-test is greater than 5% and p-value of T-tests are greater than 5%, we are unable to reject the Null hypothesis hence stating there is no significant difference between expenditure of higher and lower income group at 95% of Confidence Interval.





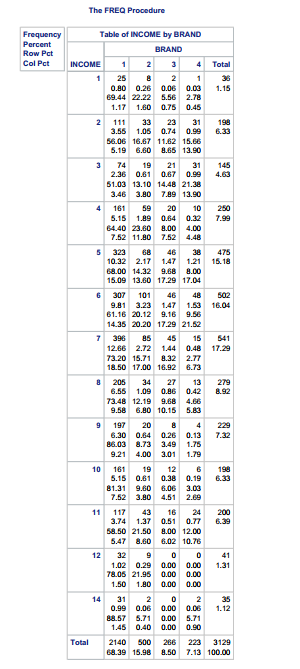
**b) Hypothesis 2:**

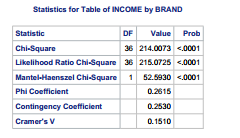
Test whether there is a significant relationship between number of members in a households and income level?

**Null Hypothesis** – There is no significant relationship between number of members in a households and income level at 95% of confidence interval.

**Alternate Hypothesis** - There is a significant relationship between number of members in a households and income level at 95% of confidence interval.

**Test Results** – Based upon the chi-square test, as the p-value is less <0.001 and hence 5%, we can safely reject the null hypothesis. This concludes that there is a significant relationship between number of members in a households and income level at 95% of confidence interval.







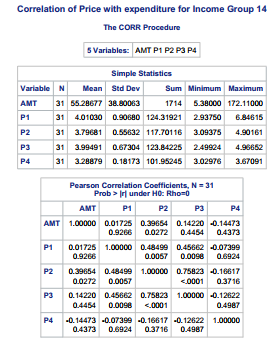
1. **Hypothesis 3:**

Test whether for Income group 14, there is any significant relationship between price of a brand and expenditure made.

**Null Hypothesis** – There is no significant relationship between price of a brand and expenditure made by any household at 95% of confidence interval.

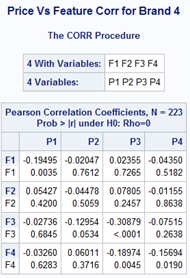
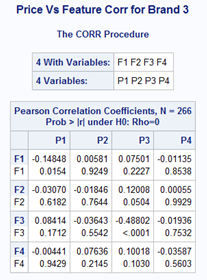
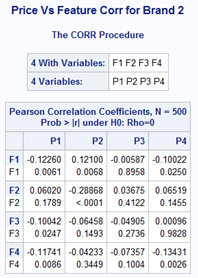
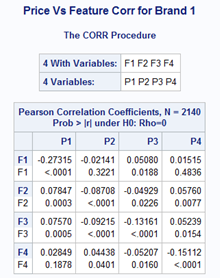
**Alternate Hypothesis** - There is a significant relationship between price of a brand and expenditure made by any household at 95% of confidence interval.

**Test Results** – Based upon the correlation, p-value is significant for only Brand 2 at 95% of confidence interval. Also, as the correlation value is positive, it shows positive relationship between price increase in brand 3 and expenditure done by Income group 14.

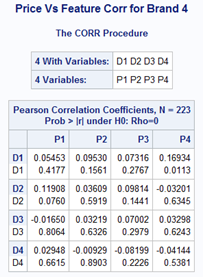
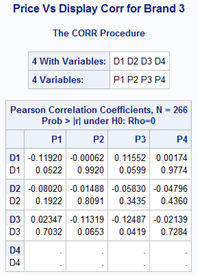
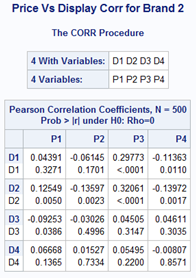
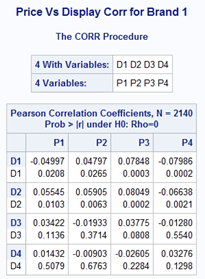


**13. What is the correlation between prices and display and feature of brand 1, 2, 3, 4? Which are the three largest correlations. What is the interpretation of these correlation coefficients?**

Price Vs Feature Correlation:



Price Vs Display Correlation:



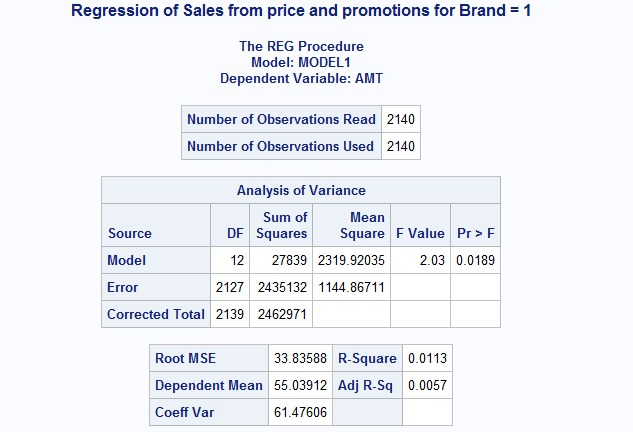
Above are the correlation matrices for all the 4 brands. By looking at the matrices above we can say that the top three correlated parameters are:

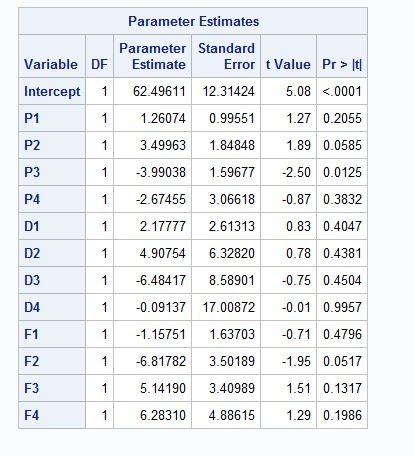
1. Correlations between P3 and F3 for Brand4: - 0.31
2. Correlations between P1 and F1 for Brand1: - 0.27
3. Correlations between P1 and F1 for Brand4: - 0.19

All the three highest correlations are not very high. Being featured has a negative correlation with price. It could be that products that are featured already have a marked down price. Whether a product is a display product or not does not have a significant correlation with price. The correlation between price and display is lesser than even price and features.

**14. How much are sales affected by prices and promotions? Write a paragraph to report on the results of your findings based on a suitable regression model. Comment on model fit (r-square, f-test, etc), t-values, coefficients and their meaning.**

**BRAND 1: -**





**F-test**As p-value is less than 5%, model is overall significant for all the independent variables taken together using F-test.   
 **P-value & t-value:**

Only P3 is significant at 5% of significance value for the current model of Brand 1.

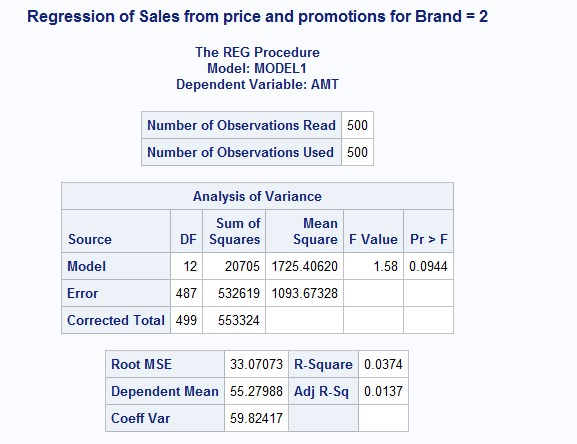
**Coefficients:**

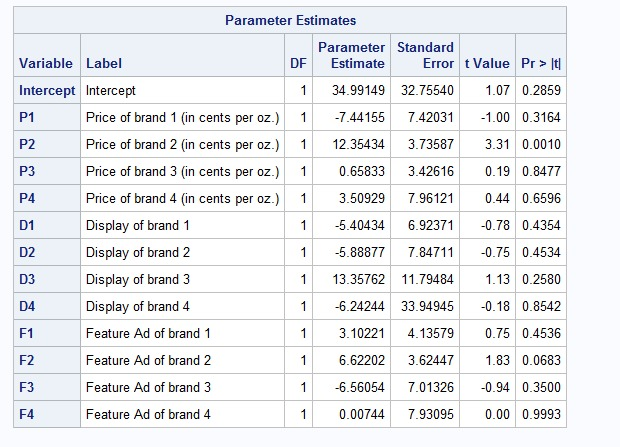
The coefficient for P3 is -3.99. It indicates that an increase of one cent per oz. in the price of Hunts ketchup is associated with an decrease of $3.99 in the sales amount when controlling for the rest of the variables in the model.

**Goodness of fit:**

Adjusted R-Square is 1.13% which means only 1.13% of the change in sales is being explained by the model. P-value of the model (Pr > F) is significant at 5%. Therefore, we do reject the null-hypothesis that all Beta coefficients in the model equal to zero. Overall, the model does a very poor job of predicting sales.

**BRAND 2: -**





**F-test**As p-value is greater than 5%, model is overall not significant for all the independent variables taken together using F-test.   
 **P-value & t-value:**

Only P2 is significant at 5% of significance value for the current model of Brand 2.

**Coefficients:**

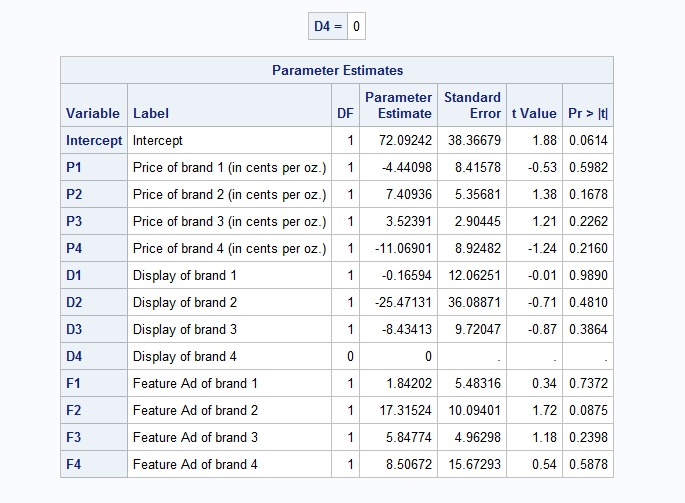
The coefficient for P2 is 12.35. It indicates that an increase of one cent per oz. in the price of Hunts ketchup is associated with an increase of $12.35 in the sales amount when controlling for the rest of the variables in the model.

**Goodness of fit:**

Adjusted R-Square is 1.37% which means only 1.13% of the change in sales is being explained by the model. P-value of the model (Pr > F) is not significant at 5%. Therefore, we do not reject the null-hypothesis that all Beta coefficients in the model equal to zero. Overall, the model does a very poor job of predicting sales.

**BRAND 3: -**





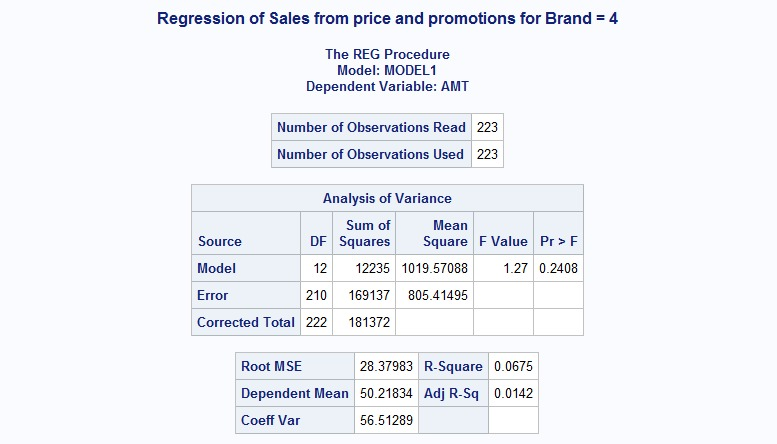
**F-test**As p-value is greater than 5%, model is overall not significant for all the independent variables taken together using F-test.   
 **P-value & t-value:**

No explanatory variable is significant at 95% of Confidence Interval.

**Goodness of fit:**

Adjusted R-Square is 4.72% which means only 4.72% of the change in sales is being explained by the model. P-value of the model (Pr > F) is not significant at 5%. Therefore, we do not reject the null-hypothesis that all Beta coefficients in the model equal to zero. Overall, the model does a very poor job of predicting sales.

**BRAND 4: -**





**F-test**As p-value is greater than 5%, model is overall not significant for all the independent variables taken together using F-test.   
 **P-value & t-value:**

Only P3 & P4 is significant at 5% of significance value for the current model of Brand 2.

**Coefficients:**

The coefficient for P3 is -10.74 and P4 is 2.88. It indicates that an increase of one cent per oz. in the price of Hunts ketchup is associated with an increase of $10.74 & $2.88 respectively in the sales amount when controlling for the rest of the variables in the model individually.

**Goodness of fit:**

Adjusted R-Square is 6.75% which means only 6.75% of the change in sales is being explained by the model. P-value of the model (Pr > F) is not significant at 5%. Therefore, we do not reject the null-hypothesis that all Beta coefficients in the model equal to zero. Overall, the model does a very poor job of predicting sales.