PROJECT

Group 7

Rishabh Bajaj
Atharva Talathi
Saurabh Chopda
Vivek Gogia
Aayushi Ghoghari

Data Cleaning

In the Dataset we observed that there were many variables which failed to prove meaningful insight and therefore were removed from the dataset.
The columns that were removed were as follows.

L1 was removed from the dataset as it had only one value Category- Coffee

Then the next columns L2, L3, L4 were removed as they had values which were covered in the rest of the dataset.

In the next step SY GE VENT ITEM were removed in data cleaning as these values were first used in finding the value of UPC and then they were removed.

After that Fat Content, Form columns were removed as they had more than 40% missing values, As the data was categorical, we were not able to use imputation method. Imputation method involves imputing median value of the column which would have created biasness and which in turn would directly affect the efficiency of our analysis

Data Preprocessing

The columns such as L5, Flavor/ scent and package are the categorical variables, so we have regrouped the columns taking in consideration the frequency of each value in the column of the dataset. Hence making the columns as:

- L5_Category
- Flavour_Category
- Package_Category

Following that we performed one hot Encoding on the above variables.
One hot Encoding is a method to convert the data to prepare it for an algorithm
And better prediction.
With one-hot, we convert each categorical value into a new categorical column and assign a binary value of 1 or 0 to those columns. Each integer value is represented as a binary vector. Thereafter, we converted all the newly formed dummy variables into numeric data type, which made it better to analyze the data and to run the regression analysis efficiently.

After making the dummy variables we merged the dataset using UPC and the IRI_key.

Then we have performed random sampling on the entire dataset.
As the dataset is very huge random sampling is a technique in which the entire dataset is sampled down to ensure that the results obtained should approximate what would have been obtained if the entire population set was measured.

OUR APPROACH ON DATA MODELLING

Our aim is to test and find the effects of 'price' per ounce (calculated according to the formula: Dollars/ (units*Size (in OZ))) and 'dollars' on the explanatory variables. We are trying to find what factors play a significant role on the price of coffee. There are various factors that affect the price, but we run the standardized regression model to find which factors are significant and insignificant.

After finding the significant factors and removing the insignificant factors, we look at the parameter estimates of the remaining explanatory variables in the model affecting the price of coffee and understand the effect they have on price.

Linear regression model on the overall dataset to assess the impact on price

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: price

Number of Observations Read	5000
Number of Observations Used	5000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	252.94865	22.99533	956.66	<.0001
Error	4988	119.89662	0.02404		
Corrected Total	4999	372.84527			

Root MSE	0.15504	R-Square	0.6784
Dependent Mean	0.42116	Adj R-Sq	0.6777
Coeff Var	36.81251		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.50305	0.02295	21.92	<.0001
D_0	1	0.07597	0.01353	5.61	<.0001
BRAND_FOLGERS	1	-0.03891	0.00650	-5.99	<.0001
BRAND_STARBUCKS	1	0.29831	0.01003	29.73	<.0001
BRAND_MILLSTONE	1	0.34579	0.00954	36.23	<.0001
BRAND_OTHER	1	0.04885	0.00567	8.61	<.0001
flavour_category_REGULAR	1	-0.01506	0.00528	-2.85	0.0044
flavour_category_VANILLA	1	0.07054	0.01090	6.47	<.0001
flavour_category_CHOCOLATE	1	0.08026	0.01041	7.71	<.0001
package_category_BAG	1	-0.19788	0.01835	-10.79	<.0001
package_category_BOX	1	0.63564	0.02160	29.43	<.0001
package_category_CONTAINER	1	-0.31496	0.01854	-16.99	<.0001

Effect of top 5 brands, flavors, and package categories on price

After looking at the data, we decided to run standardized regression to model the effect of various factors on the price per ounce of coffee.

After observing the P values for all variables and assuming the cut-off value as 0.05, we choose to remove D_1, A, B, C, None, and Brand_private. We omit these variables as they are insignificant and do not have any effect on the dependent variable.

From the above model, we can infer:

- The price of millstone is approximately 34% higher than the reference variable i.e., brand_maxwell.
- Parameter estimate shows that the price of chocolate flavor is approximately 8% higher than other flavor categories.
- The price of Box package category is approximately 63% higher than other package categories.
- D_0 variable shows that the deal on a particular brand is not on display. While running the model we found that if the deal is not displayed the price is approximately 7% higher.
- The flavor category regular, package category bag and brand Folgers have a negative effect on the price.

Multicollinearity Test on price model

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: price

Number of Observations Read	5000
Number of Observations Used	5000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	252.94865	22.99533	956.66	<.0001
Error	4988	119.89662	0.02404		
Corrected Total	4999	372.84527			

Root MSE	0.15504	R-Square	0.6784
Dependent Mean	0.42116	Adj R-Sq	0.6777
Coeff Var	36.81251		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.50305	0.02295	21.92	<.0001	0
D_0	1	0.07597	0.01353	5.61	<.0001	1.00771
BRAND_FOLGERS	1	-0.03891	0.00650	-5.99	<.0001	1.37864
BRAND_STARBUCKS	1	0.29831	0.01003	29.73	<.0001	1.35976
BRAND_MILLSTONE	1	0.34579	0.00954	36.23	<.0001	1.42946
BRAND_OTHER	1	0.04885	0.00567	8.61	<.0001	1.55627
flavour_category_REGULAR	1	-0.01506	0.00528	-2.85	0.0044	1.38316
flavour_category_VANILLA	1	0.07054	0.01090	6.47	<.0001	1.16047
flavour_category_CHOCOLATE	1	0.08026	0.01041	7.71	<.0001	1.15883
package_category_BAG	1	-0.19788	0.01835	-10.79	<.0001	17.50028
package_category_BOX	1	0.63564	0.02160	29.43	<.0001	3.74388
package_category_CONTAINER	1	-0.31496	0.01854	-16.99	<.0001	17.70090

The above table shows us the diagnostics for multicollinearity using VIF (Variance Inflation Factor). These variables can be checked under the last column of Parameter Estimates.
We notice that there is a presence of multicollinearity because there are two VIF values which are above 10. We now omit ‘package_category_container’.

Linear regression model to assess the impact on dollars

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: DOLLARS

Number of Observations Read	5000
Number of Observations Used	5000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	3612859	401429	222.67	<.0001
Error	4990	8996041	1802.81380		
Corrected Total	4999	12608900			

Root MSE	42.45955	R-Square	0.2865
Dependent Mean	22.84740	Adj R-Sq	0.2852
Coeff Var	185.83978		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	200.16383	5.09204	39.31	<.0001
D_0		1	-153.39852	4.54182	-33.77	<.0001
D_1		1	-121.36411	7.85056	-15.46	<.0001
C		1	-48.70658	9.45264	-5.15	<.0001
None		1	-43.64814	2.80066	-15.58	<.0001
VOL_EQ	VOL_EQ	1	7.63726	1.07975	7.07	<.0001
BRAND_FOLGERS		1	5.54225	1.57873	3.51	0.0005
BRAND_STARBUCKS		1	10.86509	2.50542	4.34	<.0001
flavour_category_REGULAR		1	4.44032	1.31450	3.38	0.0007
package_category_CONTAINER		1	6.45665	1.43024	4.51	<.0001

Effect of top 5 brands, flavors, and package categories on dollars

After looking at the results of the model that we ran on ‘price’, we decided to test it on another variable: ‘dollars.

Assuming the cut-off value as 0.05, we choose to remove A, B, and Brand_millstone, brand_private, brand_other, Flavour_category_vanilla, flavour_category_chocolate, flavour_category_bag and flavour_category_box. We omit these variables as they are insignificant and have minimal effect on the dependent variable i.e., Dollars.

Looking at the results of the model, we conclude:

- Brand_Starbucks has approximately 10% higher dollar price than the reference variable i.e., Brand_maxwell.
- The no-display variable (D_0) has a negative effect on dollars so it means that having no deals displayed reduced the dollar sales of coffee products.
- C (small ad) feature had a large negative effect on the dollar sales.
- From a business standpoint, flavor category regular and package type container is effectively more profitable than other flavor and package type categories.

Multicollinearity Test on Dollars

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: DOLLARS

Number of Observations Read	5000
Number of Observations Used	5000

Analysis of Variance

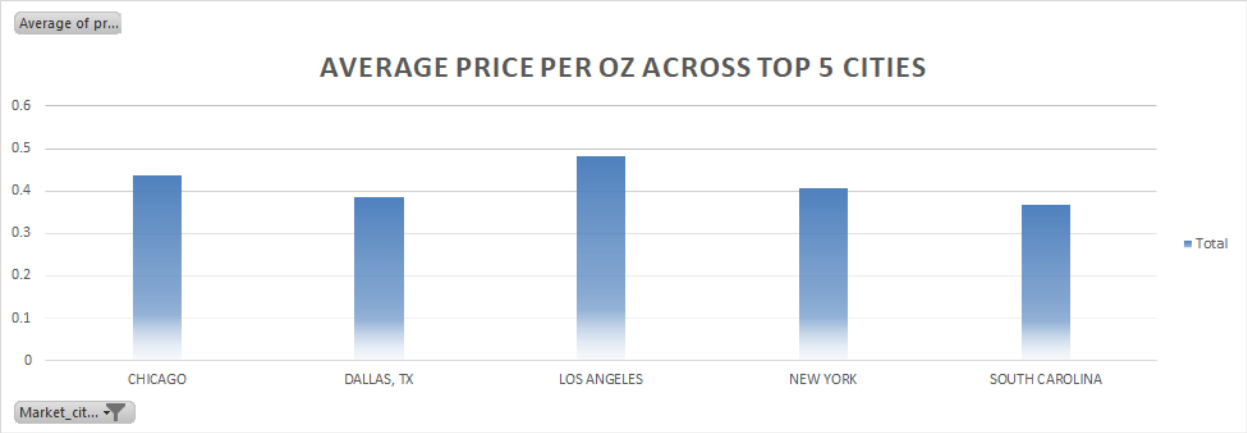
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	3612859	401429	222.67	<.0001
Error	4990	8996041	1802.81380		
Corrected Total	4999	12608900			

Root MSE	42.45955	R-Square	0.2865
Dependent Mean	22.84740	Adj R-Sq	0.2852
Coeff Var	185.83978		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	200.16383	5.09204	39.31	<.0001	0
D_0		1	-153.39852	4.54182	-33.77	<.0001	1.51381
D_1		1	-121.36411	7.85056	-15.46	<.0001	1.45736
C		1	-48.70658	9.45264	-5.15	<.0001	1.08558
None		1	-43.64814	2.80066	-15.58	<.0001	1.12679
VOL_EQ	VOL_EQ	1	7.63726	1.07975	7.07	<.0001	1.34344
BRAND_FOLGERS		1	5.54225	1.57873	3.51	0.0005	1.08425
BRAND_STARBUCKS		1	10.86509	2.50542	4.34	<.0001	1.13035
flavour_category_REGULAR		1	4.44032	1.31450	3.38	0.0007	1.14258
package_category_CONTAINER		1	6.45665	1.43024	4.51	<.0001	1.40449

The above table shows us the diagnostics for multicollinearity using VIF (Variance Inflation Factor). These variables can be checked under the last column of Parameter Estimates. We notice that there is a no presence of multicollinearity because there is no factor value above 10.



Given geographic diversity and coffee choices, we believe that it is utmost important to consider Cities and its relationship with price of coffee. Based on our analysis, we have noted that coffee has the most price among its customers in Los Angeles and Chicago.

Linear regression model with top 5 cities included to assess the impact on price

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: price

Number of Observations Read	10000
Number of Observations Used	10000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	19	581.77148	30.61955	1671.75	<.0001
Error	9980	182.79227	0.01832		
Corrected Total	9999	764.56375			

Root MSE	0.13534	R-Square	0.7609
Dependent Mean	0.42544	Adj R-Sq	0.7605
Coeff Var	31.81093		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.57754	0.01467	39.38	<.0001
A	1	-0.10256	0.00922	-11.12	<.0001
B	1	-0.09782	0.00861	-11.36	<.0001
C	1	-0.11018	0.02265	-4.86	<.0001
VOL_EQ	1	-0.10923	0.00254	-42.99	<.0001
BRAND_FOLGERS	1	-0.01824	0.00454	-4.02	<.0001
BRAND_STARBUCKS	1	0.29537	0.00654	45.14	<.0001
BRAND_MILLSTONE	1	0.31763	0.00646	49.20	<.0001
BRAND_PRIVATE	1	0.01361	0.00523	2.60	0.0093
BRAND_OTHER	1	0.04645	0.00434	10.69	<.0001
flavour_category_REGULAR	1	-0.00664	0.00325	-2.04	0.0411
flavour_category_VANILLA	1	0.04821	0.00666	7.24	<.0001
flavour_category_CHOCOLATE	1	0.05575	0.00673	8.29	<.0001
package_category_BAG	1	-0.15736	0.01284	-12.26	<.0001
package_category_BOX	1	0.64145	0.01468	43.69	<.0001
package_category_CONTAINER	1	-0.23015	0.01312	-17.55	<.0001
Market_cities_CHICAGO	1	0.04995	0.00893	5.59	<.0001
Market_cities_LOS_ANGELES	1	0.11111	0.00786	14.13	<.0001
Market_cities_NEW_YORK	1	0.03979	0.00804	4.95	<.0001
Market_cities_OTHER	1	0.03943	0.00602	6.55	<.0001

For our analysis, we ran a regression model to check if cities play a vital role in affecting the price and hence from the above matrix, we can say that price varies in different cities and that location plays a major role in price of the product. Above, we can also see that the presence or absence of an in-store display highlighting doesn't have any effect on prices. Majorly, the type of packaging used, flavors and cities play a significant role in determining prices.

Multicollinearity analysis

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: price

Number of Observations Read	10000
Number of Observations Used	10000

Analysis of Variance

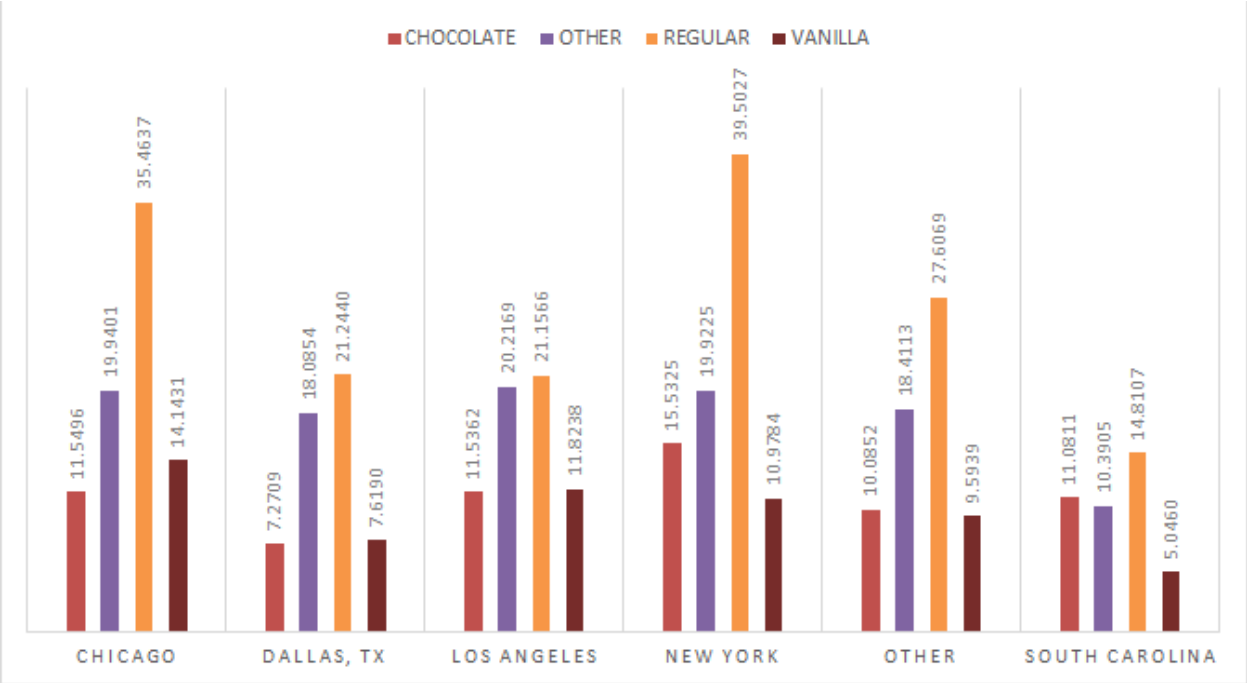
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	19	581.77148	30.61955	1671.75	<.0001
Error	9980	182.79227	0.01832		
Corrected Total	9999	764.56375			

Root MSE	0.13534	R-Square	0.7609
Dependent Mean	0.42544	Adj R-Sq	0.7605
Coeff Var	31.81093		

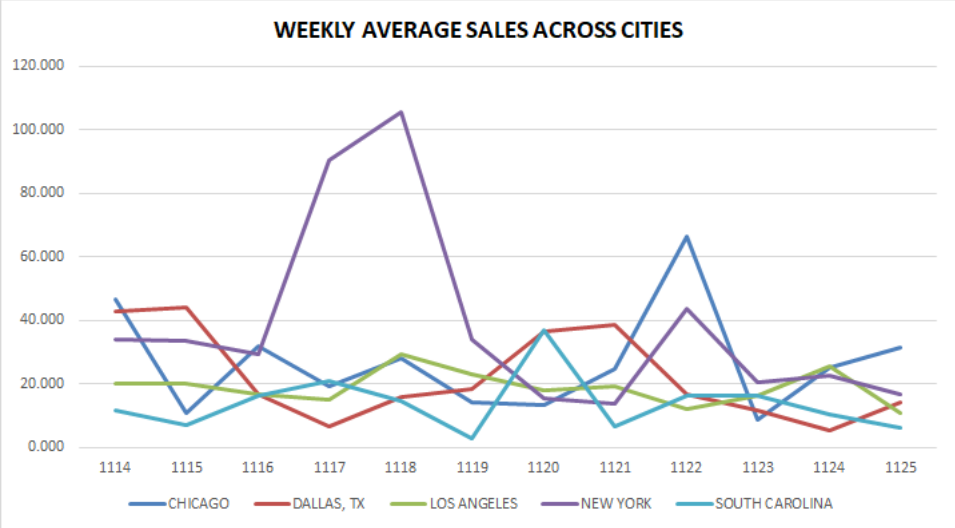
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.57754	0.01467	39.38	<.0001	0
A	1	-0.10256	0.00922	-11.12	<.0001	1.00759
B	1	-0.09782	0.00861	-11.36	<.0001	1.01055
C	1	-0.11018	0.02265	-4.86	<.0001	1.00468
VOL_EQ	1	-0.10923	0.00254	-42.99	<.0001	1.45745
BRAND_FOLGERS	1	-0.01824	0.00454	-4.02	<.0001	1.82592
BRAND_STARBUCKS	1	0.29537	0.00654	45.14	<.0001	1.62999
BRAND_MILLSTONE	1	0.31763	0.00646	49.20	<.0001	1.76590
BRAND_PRIVATE	1	0.01361	0.00523	2.60	0.0093	1.61880
BRAND_OTHER	1	0.04645	0.00434	10.69	<.0001	2.33916
flavour_category_REGULAR	1	-0.00664	0.00325	-2.04	0.0411	1.37379
flavour_category_VANILLA	1	0.04821	0.00666	7.24	<.0001	1.16233
flavour_category_CHOCOLATE	1	0.05575	0.00673	8.29	<.0001	1.15412
package_category_BAG	1	-0.15736	0.01284	-12.26	<.0001	22.49813
package_category_BOX	1	0.64145	0.01468	43.69	<.0001	4.56237
package_category_CONTAINER	1	-0.23015	0.01312	-17.55	<.0001	23.18540
Market_cities_CHICAGO	1	0.04995	0.00893	5.59	<.0001	1.68457
Market_cities_LOS_ANGELES	1	0.11111	0.00786	14.13	<.0001	2.15698
Market_cities_NEW_YORK	1	0.03979	0.00804	4.95	<.0001	2.06248
Market_cities_OTHER	1	0.03943	0.00602	6.55	<.0001	3.45411

The variance inflation factor (VIF) tells us about the multicollinearity between variables and from the above result we can see that packaging in bags and container is highly correlated. Hence, we remove one of the variables, to improve the efficiency of our model. Having multicollinearity in a model means the two variables which have underlying effect that they measure will be accounted twice across the variables.



Upon analyzing the coffee flavors across cities, we also observed that the demand for regular flavor is higher in New York and Chicago. This means more customers prefer regular flavor of coffee over chocolate, vanilla, and other flavors. Whether this rise was significant, will be explored through the regression analyses in the next section.



Linear regression model with top 5 cities included to assess the impact on dollars

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: DOLLARS

Number of Observations Read	10000
Number of Observations Used	10000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	5209434	347296	196.30	<.0001
Error	9984	17663656	1769.19627		
Corrected Total	9999	22873090			

Root MSE	42.06181	R-Square	0.2278
Dependent Mean	23.32580	Adj R-Sq	0.2266
Coeff Var	180.32315		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	136.37713	4.45233	30.63	<.0001
D_0	1	-104.95772	3.48493	-30.12	<.0001
D_1	1	-44.94892	5.55756	-8.09	<.0001
A	1	25.78642	3.69884	6.97	<.0001
None	1	-29.80343	2.43741	-12.23	<.0001
VOL_EQ	1	9.59718	0.77081	12.45	<.0001
BRAND_FOLGERS	1	3.07499	1.40653	2.19	0.0288
BRAND_STARBUCKS	1	12.14805	1.98040	6.13	<.0001
BRAND_MILLSTONE	1	-4.29064	1.96255	-2.19	0.0288
BRAND_PRIVATE	1	-4.96097	1.61798	-3.07	0.0022
BRAND_OTHER	1	-3.17183	1.32257	-2.40	0.0165
flavour_category_REGULAR	1	3.81470	0.93742	4.07	<.0001
package_category_CONTAINER	1	7.39943	1.04734	7.06	<.0001
Market_cities_CHICAGO	1	6.04513	2.42854	2.49	0.0128
Market_cities_NEW_YORK	1	11.67167	2.07972	5.61	<.0001
Market_cities_OTHER	1	3.41924	1.29970	2.63	0.0085

Effect on dollar sales

- After running regression, we see that the flavor category regular, market cities Chicago and New York are highly significant as observed in the graph in the section above.
- D_0(no display) has a huge negative effect on the dollar sales as expected.
- Brand Starbucks has the highest dollar sales as compared to other brands. It is approximately 12% higher than the other brands.
- New York city has the highest dollar sales as compared to all the other cities.

Multicollinearity Analysis

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: DOLLARS

Number of Observations Read	10000
Number of Observations Used	10000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	20	5218255	260913	147.48	<.0001
Error	9979	17654835	1769.19878		
Corrected Total	9999	22873090			

Root MSE	42.06184	R-Square	0.2281
Dependent Mean	23.32580	Adj R-Sq	0.2266
Coeff Var	180.32328		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	134.93432	4.75177	28.40	<.0001	0
D_0	1	-104.85559	3.48756	-30.07	<.0001	1.66270
D_1	1	-45.00355	5.55818	-8.10	<.0001	1.57456
A	1	25.31599	3.79548	6.67	<.0001	1.76750
C	1	-5.87080	7.46572	-0.79	0.4317	1.13006
None	1	-30.41118	2.58799	-11.75	<.0001	1.92714
VOL_EQ	1	9.95963	0.79152	12.58	<.0001	1.46414
BRAND_FOLGERS	1	2.91104	1.41034	2.06	0.0390	1.82697
BRAND_STARBUCKS	1	12.72814	2.03219	6.26	<.0001	1.62734
BRAND_MILLSTONE	1	-3.65424	2.00182	-1.83	0.0680	1.75786
BRAND_PRIVATE	1	-4.76091	1.62591	-2.93	0.0034	1.61972
BRAND_OTHER	1	-2.87952	1.34902	-2.13	0.0328	2.33486
flavour_category_REGULAR	1	3.60924	1.00986	3.57	0.0004	1.37383
flavour_category_VANILLA	1	-0.22761	2.06548	-0.11	0.9123	1.15842
flavour_category_CHOCOLATE	1	0.80879	2.08277	0.39	0.6978	1.14479
package_category_BOX	1	4.33296	2.37884	1.82	0.0686	1.24002
package_category_CONTAINER	1	7.72549	1.08549	7.12	<.0001	1.64419
Market_cities_CHICAGO	1	7.15130	2.77664	2.58	0.0100	1.68540
Market_cities_LOS_ANGELES	1	2.24641	2.44520	0.92	0.3583	2.15930
Market_cities_NEW_YORK	1	12.78478	2.49865	5.12	<.0001	2.06152
Market_cities_OTHER	1	4.63509	1.86985	2.48	0.0132	3.45258

We run the multicollinearity VIF test and see that there is no multicollinearity present amongst the variables.

Effect of doubling the vol_eq on dollars sales

Root MSE	42.06364	R-Square	0.2278
Dependent Mean	23.32580	Adj R-Sq	0.2265
Coeff Var	180.33097		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	136.65706	4.51717	30.25	<.0001
D_0	D_0	1	-104.85084	3.49719	-29.98	<.0001
D_1	D_1	1	-44.96844	5.55806	-8.09	<.0001
A	A	1	25.79783	3.69913	6.97	<.0001
None	None	1	-29.83619	2.43915	-12.23	<.0001
squarevol		1	0.31288	0.85112	0.37	0.7132
VOL_EQ	VOL_EQ	1	8.73247	2.47533	3.53	0.0004
BRAND_FOLGERS	BRAND_FOLGERS	1	3.07525	1.40660	2.19	0.0288
BRAND_STARBUCKS	BRAND_STARBUCKS	1	12.24013	1.99627	6.13	<.0001
BRAND_MILL STONE	BRAND_MILLSTONE	1	-4.39269	1.98217	-2.22	0.0267
BRAND_PRIVATE	BRAND_PRIVATE	1	-4.97244	1.61835	-3.07	0.0021
BRAND_OTHER	BRAND_OTHER	1	-3.16351	1.32282	-2.39	0.0168
flavour_category_REGULAR	flavour_category_REGULAR	1	3.82099	0.93762	4.08	<.0001
package_category_CONTAINER	package_category_CONTAINER	1	7.46415	1.06207	7.03	<.0001
Market_cities_CHICAGO	Market_cities_CHICAGO	1	6.04462	2.42865	2.49	0.0128
Market_cities_NEW_YORK	Market_cities_NEW_YORK	1	11.66957	2.07982	5.61	<.0001
Market_cities_OTHER	Market_cities_OTHER	1	3.43281	1.30028	2.64	0.0083

We have made a new column “squarevol” which is (vol_eq*vol_eq) to check if it has any different effect on the dollar sales. We can infer from this that doubling the vol_eq will have no effect on the dollars or sales which is our dependent variable.

Interaction effect model to check the combined effect of BRAND and CITY on Price.

Root MSE	0.15109	R-Square	0.7020
Dependent Mean	0.42544	Adj R-Sq	0.7015
Coeff Var	35.51393		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.22747	0.00787	28.89	<.0001
interaction_effect		1	-0.01583	0.02769	-0.57	0.5675
A	A	1	-0.12579	0.01028	-12.24	<.0001
B	B	1	-0.11673	0.00961	-12.15	<.0001
C	C	1	-0.15850	0.02526	-6.27	<.0001
BRAND_FOLGERS	BRAND_FOLGERS	1	-0.03362	0.00505	-6.66	<.0001
BRAND_STARBUCKS	BRAND_STARBUCKS	1	0.31413	0.00738	42.54	<.0001
BRAND_MILLSTONE	BRAND_MILLSTONE	1	0.35702	0.00715	49.95	<.0001
BRAND_PRIVATE	BRAND_PRIVATE	1	0.02491	0.00583	4.27	<.0001
BRAND_OTHER	BRAND_OTHER	1	0.06411	0.00482	13.29	<.0001
flavour_category_REGULAR	flavour_category_REGULAR	1	-0.02107	0.00361	-5.83	<.0001
flavour_category_VANILLA	flavour_category_VANILLA	1	0.07721	0.00739	10.45	<.0001
flavour_category_CHOCOLATE	flavour_category_CHOCOLATE	1	0.09143	0.00744	12.29	<.0001
package_category_BAG	package_category_BAG	1	0.10922	0.00363	30.12	<.0001
package_category_BOX	package_category_BOX	1	0.98926	0.00800	123.59	<.0001
Market_cities_CHICAGO	Market_cities_CHICAGO	1	0.02684	0.00995	2.70	0.0070
Market_cities_LOS_ANGELES	Market_cities_LOS_ANGELES	1	0.12385	0.00877	14.12	<.0001
Market_cities_NEW_YORK	Market_cities_NEW_YORK	1	0.05049	0.00910	5.55	<.0001
Market_cities_OTHER	Market_cities_OTHER	1	0.03690	0.00672	5.49	<.0001

We have added a new column of 'Interaction effect' which is the combination of Brand_Starbucks and Brand_New_York. We ran a regression model with the dependent variable as price. Here we get a p-value of 0.5675 which is greater than 0.05 so it is not significant. We can infer that the combined effect of Brand(Starbucks) and City(New York) do not have any significant effect to price.

Interaction effect model to check the combined effect of FLAVOUR and BRAND on Price.

Root MSE	0.15098	R-Square	0.7024
Dependent Mean	0.42544	Adj R-Sq	0.7019
Coeff Var	35.48743		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.22706	0.00787	28.85	<.0001
interaction_effect3		1	-0.07642	0.01958	-3.90	<.0001
A	A	1	-0.12551	0.01027	-12.22	<.0001
B	B	1	-0.11713	0.00960	-12.20	<.0001
C	C	1	-0.15825	0.02524	-6.27	<.0001
BRAND_FOLGERS	BRAND_FOLGERS	1	-0.03387	0.00504	-6.72	<.0001
BRAND_STARBUCKS	BRAND_STARBUCKS	1	0.31401	0.00728	43.12	<.0001
BRAND_MILLSTONE	BRAND_MILLSTONE	1	0.36397	0.00736	49.44	<.0001
BRAND_PRIVATE	BRAND_PRIVATE	1	0.02442	0.00583	4.19	<.0001
BRAND_OTHER	BRAND_OTHER	1	0.06381	0.00482	13.24	<.0001
flavour_category_REGULAR	flavour_category_REGULAR	1	-0.02050	0.00361	-5.67	<.0001
flavour_category_VANILLA	flavour_category_VANILLA	1	0.07630	0.00739	10.33	<.0001
flavour_category_CHOCOLATE	flavour_category_CHOCOLATE	1	0.10367	0.00807	12.85	<.0001
package_category_BAG	package_category_BAG	1	0.10854	0.00363	29.93	<.0001
package_category_BOX	package_category_BOX	1	0.98931	0.00800	123.69	<.0001
Market_cities_CHICAGO	Market_cities_CHICAGO	1	0.02655	0.00995	2.67	0.0076
Market_cities_LOS_ANGELES	Market_cities_LOS_ANGELES	1	0.12403	0.00877	14.15	<.0001
Market_cities_NEW_YORK	Market_cities_NEW_YORK	1	0.04970	0.00897	5.54	<.0001
Market_cities_OTHER	Market_cities_OTHER	1	0.03701	0.00671	5.51	<.0001

We have added a new column of 'Interaction effect3' which is the combination of flavour_category_CHOCOLATE and BRAND_MILLSTONE. We ran a regression model with the dependent variable as price. We can observe that the p-value is less than 0.05, so it is significant to our dependent variable. We can infer that the combined effect of Flavor (Chocolate) and Brand (Millstone) has a positive impact on Price. 0.35 0.08 The effect of chocolate flavor before the interaction effect was 0.08 on price and after running this model it increased to 0.10. So, we can conclude that increase in 1 unit of price increases 0.10 units of chocolate flavor sold if there is a combined effect.
The effect of Millstone brand before the interaction effect was 0.34 and it increases to 0.36 which is 0.02 units increased. We can conclude that increase in 1 unit of price increases 0.36 units of Millstone brand sold.

Insights:

There are many factors affecting the price per oz of the coffee products. After running the model on price, we can see that:

- Brand Millstone has the highest price per oz compared to other brands while brand Folgers has the lowest.
- Flavor category chocolate has the highest price per oz compared to other flavor categories while flavor regular has the lowest.
- City Los Angeles has the highest price per oz compared to other cities while New York has the lowest.
- Package category box has the highest price per oz while container has the lowest.

After looking over some of the insights, we can conclude that if the customer base is price sensitive, then we would recommend keeping Brand Folgers, flavor category regular in stock and having stores in New York.

Market Strategy:

- If the Customer is price sensitive. Analyzing the price effect on brands will help to increase sales.
  To get better sales we suggest keeping less products of brand Starbucks with flavor chocolate as they have higher price when compared to other brands.

- Packaging also plays an important part in sales.
  It was observed that package category box had the highest dollar value sales compared to another package category. We suggest that if there are products with packaging in box there can be an increase in sales.

- Flavor category chocolate had a higher dollar value sale than any another flavor category.
  We suggest if a store wants to increase their sales, they can keep chocolate flavor products in stock.

- The city of New York observed higher dollar value sales compared to other cities, so to get more sales more stores can be opened in New York.

- Surprisingly having a deal on display does not have a positive impact on the dollar sales. That means that if a store has an advertising display it does not guarantee there will be an increase in sales.