STAT448 Final Project

1.0 Project description

A distributor in Portugal is interested in sales differences across the regions and channels for different types of spending. Data was provided by the client on 2 Channels and 3 regions, of which the channels are Hotels/Restaurants/Cafes (Horeca), and the regions include Lisbon, Oporto and "Other" Portuguese cities. A brief summary of the channels and regions are presented below:

channelname		regionname		
Horeca	Retail	Lisbon	Oporto	Other
N	N	N	N	N
298	142	77	47	316

There are 298 Horeca and 142 Retail channels, with 77 of them come from Lisbon, 47 come from the city Oporto, and 316 from the other cities in Portugal. Each channel or region contains the sales of fresh food, milk, groceries, frozen food, detergent paper, and delicatessen food in units of 100s. Our goal is to understand the distribution of the sales of these products, whether one channel or region has larger sales in certain types of product than other channels or regions, and how these products are related to the total sales.

2.0 Statistical Analysis

Based on the information of data, our client proposed a list of 5 questions, and we answered them in the outline below:

Question 1

Provide a general descriptive overview of types of annual spending by distribution channel (Hotel/Restaurant/Café vs. Retail) and point out any differences in spending behavior you notice for the two channels

The SUMMARY Procedure

channelname=Horeca

Variable	Mean	Std Dev	Skewness	Minimum	Lower Quartile	Median
Fresh	134.7556040	138.3168750	2.5120836	0.0300000	40.4200000	95.8150000
Milk	34.5172483	43.5216557	4.6601864	0.5500000	11.6200000	21.5700000
Grocery	39.6213758	35.4551339	2.1183162	0.0300000	16.9400000	26.8400000
Frozen	37.4825168	56.4391250	5.2114478	0.2500000	8.3000000	20.5750000
Detergents_Paper	7.9056040	11.0409367	2.8571237	0.0300000	1.8300000	3.8550000
Delicassen	14.1595638	31.4742692	11.5218084	0.0300000	3.7900000	8.2100000

Variable	Upper Quartile	Maximum
Fresh	182.9100000	1121.51
Milk	40.5100000	439.5000000
Grocery	50.9100000	210.4200000
Frozen	45.7500000	608.6900000
Detergents_Paper	9.1200000	69.0700000
Delicassen	15.5000000	479.4300000

channelname=Retail

Variable	Mean	Std Dev	Skewness	Minimum	Lower Quartile	Median
Fresh	89.0432394	89.8771475	1.5939478	0.1800000	23.4300000	59.9350000
Milk	107.1650000	96.7963135	3.4131685	9.2800000	59.2100000	78.1200000
Grocery	163.2285211	122.6731809	2.9809453	27.4300000	92.1200000	123.9000000
Frozen	16.5261268	18.1280366	2.5268962	0.3300000	5.3200000	10.8100000
Detergents_Paper	72.6950704	62.9108970	2.6124251	3.3200000	36.7400000	56.1450000
Delicassen	17.5343662	19.5379705	3.7728408	0.0300000	5.5500000	13.5000000

Variable	Upper Quartile	Maximum
Fresh	122.3800000	444.6600000
Milk	122.2000000	734.9800000
Grocery	202.9200000	927.8000000
Frozen	21.9400000	115.5900000
Detergents_Paper	86.8200000	408.2700000
Delicassen	21.5700000	165.2300000

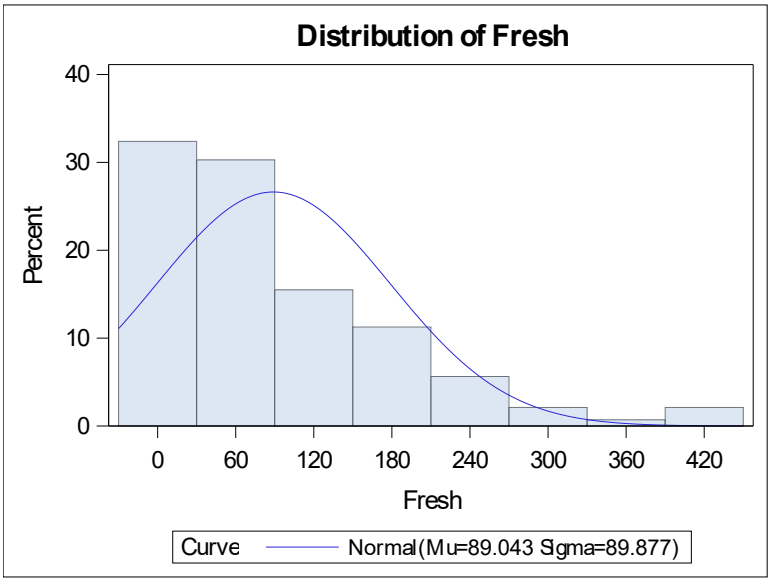
The summary table records the basic statistical measurements of all types of spending for each channel. For example, for the fresh food spending in Horeca (hotels/restaurants/cafes), the mean is around 135 with a standard deviation of around 138, which is a pretty high stand deviation. The median spending of fresh food is around 96, and the entire span from largest to smallest value is around 1121, which is computed from the difference between the maximum and minimum. The skewness is 2.51, which indicates the fresh food spending for Horeca is strongly right-skewed. Q1 and Q3 represents the 25th and 75th percentile value for all fresh food spending in Horeca respectively.

The other types of spending for both Horeca and Retails channels have similar conditions. The standard deviations for all spending are pretty large, and skewness is positively strong, meaning all spending are strongly right-skewed, so normality for data might not be satisfied. This also corresponds to the large value of range for each spending. To visually and quantitatively check if the normality assumption for spending is satisfied, we can look at the distribution of histograms and results of normality tests. The detailed explanation of histograms and normality tests are presented in Appendix A. Below is an example of histograms and tests we created:

The UNIVARIATE Procedure
Variable:
Fresh

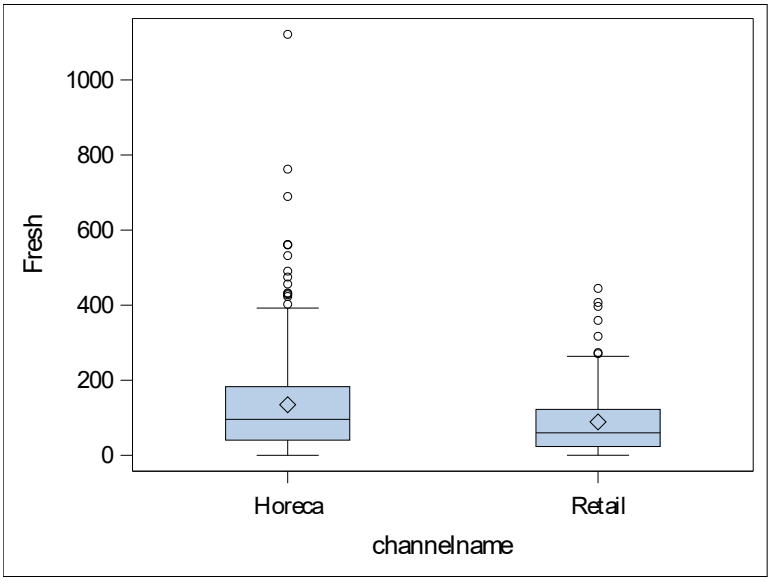
channelname=Horeca

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.78416	**Pr < W**	<0.0001
Kolmogorov-Smirnov	**D**	0.165019	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	2.458404	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	14.34703	**Pr > A-Sq**	<0.0050



The test statistic is significant, meaning fresh food spending does not follow normal distribution. This also corresponds to the histogram above since it is not symmetric.

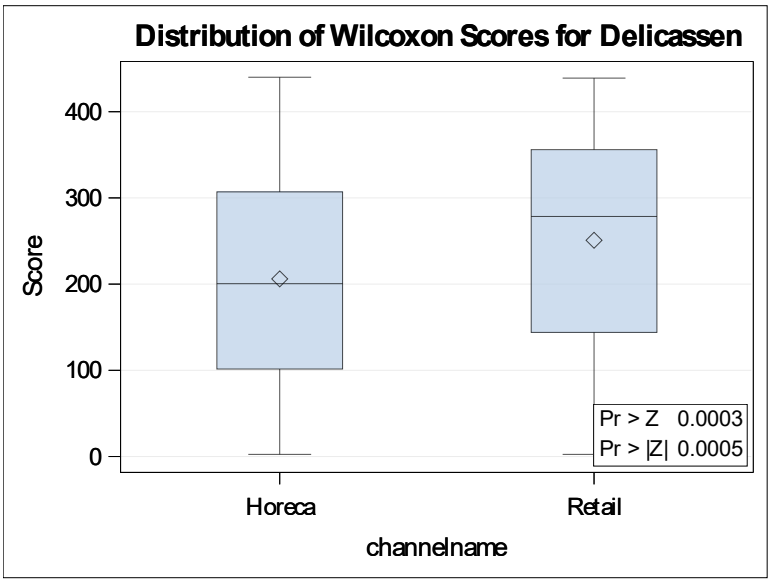
Looking at the summary statistics of different spending, we see that major differences are in terms of all variables except for Delicatessen. Horeca has greater value in the spending of fresh food and frozen food, and Retails has greater value in milk, grocery and detergents paper. The following boxplots show the relationship between the Horeca and Retails' spending in all of the food types and detergents paper. Here we show one example of the boxplots and the remaining ones are shown in Appendix A.



We could see there are many extreme observations that might influence the normality of data. Based on the results, we concluded that none of the spending follows normal distribution, which means their graphs are not symmetric.

The difference in Delicatessen, as can be seen in both the summary table and in the boxplot from Appendix A, is not obvious, so we performed a Wilcoxon two-sample test due to the non-normality.

Wilcoxon Two-Sample Test					
				t Approximation	
Statistic	Z	Pr > Z	Pr > |Z|	Pr > Z	Pr > |Z|
35620.50	3.4554	0.0003	0.0005	0.0003	0.0006
Z includes a continuity correction of 0.5.					



The p-value of Wilcoxon two-sample test for the Delicatessen is significant, which indicates Horeca and Retail does possess different spending in Delicatessen. The other types of spending have great difference between channels, and the Wilcoxon tests results in Appendix A indicate significant difference in spending between channels.

Question 2

The distributor is interested in relationships between types of spending as indicators of whether a business is a Retail business (as opposed to a Hotel/Restaurant/Café). Ignore the region and obtain your best model for being a Retail establishment as a function of spending variables. Interpret the relationships between spending and being a Retail channel for the distributor.

In this section, we used supervised learning models to predict the channel that each observation belongs to based on spending in our data.

We first used a logistic regression model using all spending as predictors, and the stepwise selection was performed to retain significant variables in the model. According to global test results in Appendix B, the p-value of global test is significant, which means at least one of the variables is significant. We find that the grocery spending and detergent paper spending are the significant predictors of the Retail channel. We also applied forward and backward selection, and achieved the same significant parameters as the stepwise selection. Details of forward and backward method are presented in Appendix B.

In this case, we treat the stepwise selection model as the final model. The variables included are the spending of grocery and detergents paper.

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-3.9880	0.3749	113.1395	<.0001
Grocery	1	0.0129	0.00454	8.0901	0.0045
Detergents_Paper	1	0.0928	0.0123	57.1691	<.0001

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
Grocery	1.013	1.004	1.022
Detergents_Paper	1.097	1.071	1.124

Looking at the maximum likelihood estimates, we see a statistically significant positive relationship between grocery spending and being Retail, and detergent paper spending and being Retail as well. With each additional unit increase in the spending of grocery food, there will be an expected increase of 0.0129 in the log-odds of becoming a Retail business, and the odds ratio shows this corresponds to a multiplicative increase of 1.3% of being a Retail business. Similarly, for each additional unit increase in the spending of detergent paper, there will be an expected increase of 0.0928 in the log-odds of becoming a Retail business. This corresponds to a multiplicative increase of 9.7% of being a Retail business.

The multiplicative increases refer to the odds ratio estimates in the table above, and the odds ratio is significant because neither 95% confidence interval includes 1.

Table of channelname by _INTO_			
channelname	**_INTO_(Formatted Value of the Predicted Response)**		
Frequency	**Horeca**	**Retail**	**Total**
Horeca	282	16	298
Retail	25	117	142
Total	307	133	440

The frequency table compares the number of true region names to the predicted groups, and the _INTO_ value represents the channel the number of observations would have been placed into based on the model. There are 399 of all 440 channels predicted correctly, which consists of 288 from Horeca and 117 from Retail, and only 41 channels were misclassified, which leads to a misclassification rate of 9.3182%

We also fitted a logistic regression model using all the spending and the two-way interactions of these spending.

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-4.2007	0.3981	111.3341	<.0001
Grocery	1	0.0153	0.00461	10.9965	0.0009
Detergents_Paper	1	0.1060	0.0132	64.0187	<.0001
Grocery*Detergents_P	1	-0.00013	0.000023	32.1963	<.0001

Table of channelname by _INTO_			
channelname	**_INTO_(Formatted Value of the Predicted Response)**		
Frequency	**Horeca**	**Retail**	**Total**
Horeca	281	17	298
Retail	25	117	142
Total	306	134	440

By adding the interaction term, the stepwise selection fits the model with the same variables as the previous model with the extra interaction term between the two variables. All the terms in this model are significant, but looking at the frequency table, one more region was misclassified compared to the first model that did not include the interaction term. This means the first model using only the spending of fresh food and detergent papers has better prediction accuracy, and intuitively, the spending of grocery and detergents paper should not be correlated, which means the interaction is meaningless.

We also checked the diagnostic plots for the first model with only main effects and all assumptions are satisfied. Related plots are presented in Appendix B. Based on the logistic regression results, prediction accuracy and model assumption checks, we select the first model as the best model.

Question 3:

The distributor believes that the regions are not very different in terms of spending. Ignoring channel, use the spending to variables to check the distributor's claim that the regions cannot be distinguished well based on spending alone. Provide and interpret your best model and state how it supports and refutes the distributor's claim.

Our final conclusion is that the regions cannot be distinguished well based on the spending alone.

We performed discriminant analysis to cluster the data and predict the region where each observation comes from.

We first use a stepwise selection with a significance threshold of 0.05, but in that case none of the variables was selected due to insignificance. As, we increased the value of threshold and finally set the significance threshold to be 0.33. The results of the model under this threshold are easy to interpret and have the best prediction accuracy compared to the other models. This high p-value threshold is a hint that the regions are hard to distinguish based on spending alone.

The summary table for the stepwise selection is listed below:

Stepwise Selection Summary								
Step	Number In	Entered	Removed	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda
1	1	Fresh		0.0052	1.13	0.3235	0.99484805	0.3235
2	2	Frozen		0.0095	2.10	0.1239	0.98536222	0.1688
3	3	Delicassen		0.0059	1.28	0.2785	0.97958797	0.1740
4	4	Detergents_Paper		0.0056	1.22	0.2967	0.97411898	0.1790
5	5	Milk		0.0080	1.75	0.1744	0.96629284	0.1352
6	4		Delicassen	0.0021	0.45	0.6408	0.96828141	0.0808

Looking at the summary table above, the stepwise method selects the variables with p-value less than 0.33. In this case, fresh food, frozen food, detergents paper and milk spending are the significant variables being selected. Delicatessen was removed because of its insignificance. However, the Partial R-square, an indicator to show how much of the variation in the response variable could be explained by the model, is almost equal to 0 for all the selected variables. This means the predictions of which region each observation comes from is hard to be explained by the model. This is another hint that the regions are hard to distinguish based on spending.

We choose either linear or quadratic discriminant analysis to see the predictions, of which depends on the chi-square test.

The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
199.897611	20	<.0001

Since the Chi-Square value is significant at the 0.1 level, the within covariance matrices will be used in the discriminant function.
Reference: Morrison, D.F. (1976) Multivariate Statistical Methods p252.

The chi-square test has a significant p-value, and accordingly, we used quadratic discriminant analysis instead of the linear one. This means different regions possess different covariance structures, so within covariance matrices were computed to measure the distance between each observation. The corresponding results show predictions of the regions, and under the predictions the within group distance, the distance between observations for each region, is minimized

Multivariate Statistics and F Approximations					
S=2 M=0.5 N=216					
Statistic	**Value**	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Wilks' Lambda	0.96828141	1.76	8	868	0.0808
Pillai's Trace	0.03177014	1.76	8	870	0.0823
Hotelling-Lawley Trace	0.03270438	1.77	8	617.68	0.0797
Roy's Greatest Root	0.03098627	3.37	4	435	0.0099
NOTE: F Statistic for Roy's Greatest Root is an upper bound.					
NOTE: F Statistic for Wilks' Lambda is exact.					

Since the sample size for each region is not the same, we placed a proportional prior such that the prior probability for each region is the same as the true proportion in the original data. After fitting the model, we used the proportion of the predictions to compare to the true proportion, and see how much prediction accuracy we achieved.

For the quadratic discriminant analysis, we first checked the MANOVA tests result. Three tests have an insignificant p-value under 5% significant level, which implies the model might not be able to provide some amount of discriminating power between the three regions.

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.WHOLESALE
Cross-validation Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into regionname				
From regionname	**Lisbon**	**Oporto**	**Other**	**Total**
Lisbon	0 0.00	1 1.30	76 98.70	77 100.00
Oporto	0 0.00	3 6.38	44 93.62	47 100.00
Other	0 0.00	9 2.85	307 97.15	316 100.00
Total	0 0.00	13 2.95	427 97.05	440 100.00
Priors	0.175	0.10682	0.71818	

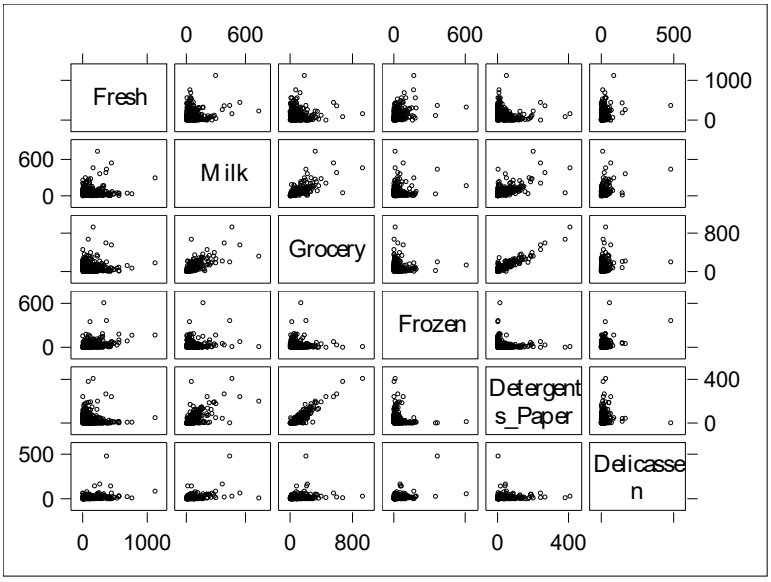
Error Count Estimates for regionname				
	Lisbon	**Oporto**	**Other**	**Total**
Rate	1.0000	0.9362	0.0285	0.2955
Priors	0.1750	0.1068	0.7182	

For prediction accuracy, we used cross-validation to estimate prediction error. The cross-validation results show that the observations for all 3 regions are mixed together. Almost all predictions were classified into the regions other than Lisbon and Oporto, and only 307 cities from other regions and 3 cities from Oporto were classified correctly. All of the predictions for Lisbon are misclassified; the error rate for original Lisbon and Oporto regions are 100% and 93.62% respectively. Though the overall error rate is just 29.55% because of correct predictions from other regions, the noticeably misclassified items indicate the three regions are not pretty distinguishable based on spending alone. This means the three regions should have very similar characteristics with respect to spending.

Question 4:

The distributor is interested in understanding how frozen food sales are related to sales of other products. Obtain your best model for frozen food spending as a function of other types of spending and interpret the model for the distributor.

As a demonstration of pairwise correlation exploration of different types of spending, a pairwise scatter plot follows below.



The frozen food spending will be the response variable, and the goal is to find the best model for frozen food spending as a function of other types of spending. The plots of other types of spending vs Frozen will give some indications that which type of spending might have strong or weak linear relationship with the frozen food spending. In this plot, it follows that the points in all plots related to Frozen are collected in the low left corner, so it is hard to directly see the relationships. We assume fresh food and delicatessen have positive linear relationship to frozen food, and the milk, grocery and detergents paper have negative linear relationship to frozen food. Later we will perform model selection methods to select the best model.

The plots of predictors vs predictors might show some multicollinearity between predictors. If two predictors are strongly correlated (either positively or negatively), at least one of them should be removed. We can observe that among all pairs of predictors, milk, grocery and detergents paper have strong positive relationship with each other. This suggests we might need to remove at least one of them from the model.

We started with the full model, and applied the stepwise selection first for model selection.

Stepwise selection:

Dependent Variable: Frozen								
Summary of Stepwise Selection								
Step	**Variable Entered**	**Variable Removed**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
1	Delicassen		1	0.1528	0.1528	48.6660	79.02	<.0001
2	Fresh		2	0.0666	0.2194	12.5650	37.28	<.0001
3	Detergents_Paper		3	0.0164	0.2359	5.1558	9.38	0.0023

The stepwise method retains delicatessen, fresh food and detergents paper in the model. The global test results show our model is significant, and all of the variables included are significant under 5% significant level. The partial R-square represents the amount of variation in the response variable that could be explained by each predictor, and the model R-square is a cumulative of partial R-squares. According to the stepwise method, only 23.59% of the variation of frozen food spending can be explained by the three spending we selected, which is a relatively low proportion.

We also used forward selection and backward elimination to compare the results.

Forward selection:

Summary of Forward Selection							
Step	**Variable Entered**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
1	Delicassen	1	0.1528	0.1528	48.6660	79.02	<.0001
2	Fresh	2	0.0666	0.2194	12.5650	37.28	<.0001
3	Detergents_Paper	3	0.0164	0.2359	5.1558	9.38	0.0023

Backward elimination:

Variable	**Parameter Estimate**	**Standard Error**	**Type II SS**	**F Value**	**Pr > F**
Intercept	14.12094	3.18598	35620	19.64	<.0001
Fresh	0.09588	0.01670	59765	32.96	<.0001
Detergents_Paper	-0.13189	0.04306	17016	9.38	0.0023
Delicassen	0.58324	0.07468	110588	60.99	<.0001

Summary of Backward Elimination							
Step	**Variable Removed**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
1	Grocery	4	0.0003	0.2411	4.1902	0.19	0.6629
2	Milk	3	0.0052	0.2359	5.1558	2.97	0.0855

Forward selection achieves the same parameters as the stepwise selection, and all the variables included are significant. The backward elimination removes grocery and milk from the model, so it achieves the same result as the forward and stepwise selection. This means under 5% significant level, stepwise, forward and backward selection each ended up with the same final model. The final model is regarded as the best model, and we presented again along with some parameter interpretation and model analysis.

Final model:

We first performed some diagnostic checks. In the Cooks distance plot Figure 4.1, two of the observations have a Cooks distance greater than 0.5 as an evidence of potential outliers. We removed them and checked the diagnostic plots again until all remaining observations have similar Cook's distance value.

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	94053	31351	27.62	<.0001
Error	434	492654	1135.14678		
Corrected Total	437	586706			

We removed two extreme observations in total. The ANOVA table implies our model is significant, so at least one of the parameters in the model is significant.

Root MSE	33.69194	**R-Square**	0.1603
Dependent Mean	28.63578	**Adj R-Sq**	0.1545
Coeff Var	117.65681		

R-square of our model is 0.1603, meaning 16.03% of variation in frozen food spending can be explained by delicatessen, fresh food, and detergents paper spending in total. Again this is not a high percentage.

Parameter Estimates						
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	1	16.95634	2.59566	6.53	<.0001	0
Delicassen	1	0.28392	0.09823	2.89	0.0040	1.11715
Fresh	1	0.09065	0.01346	6.74	<.0001	1.10321
Detergents_Paper	1	-0.10747	0.03449	-3.12	0.0020	1.04507

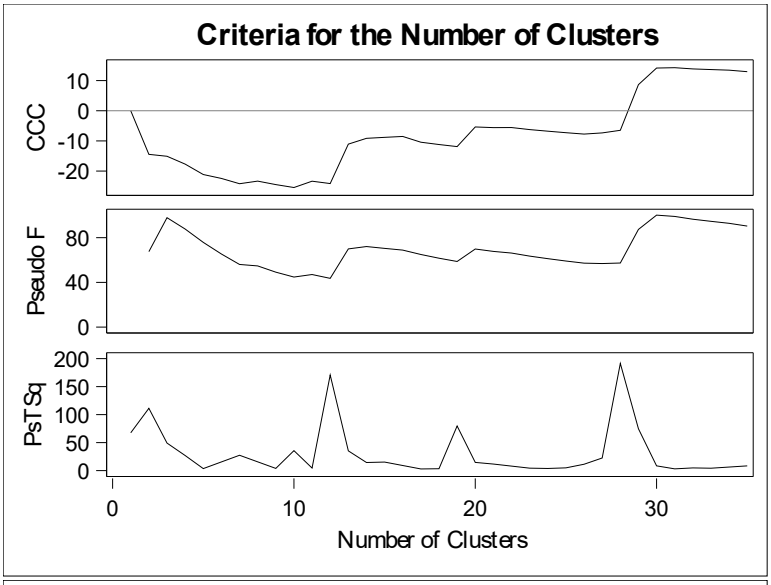
Looking at the parameters estimates table, all the three types of spending are significant. For each additional unit increase in delicatessen spending, an increase of 0.28392 in frozen food spending is expected by the model. For each additional unit increase in fresh food spending, an increase of 0.09065 in frozen food spending is expected by the model. For each additional unit increase in detergents paper spending, a decrease of 0.10747 in frozen food spending is expected by the model. A high variance inflation (VIF) value is an indicator of multicollinearity between predictors, and in the final model all VIFs are pretty small, which means the predictors are nearly independent to each other.

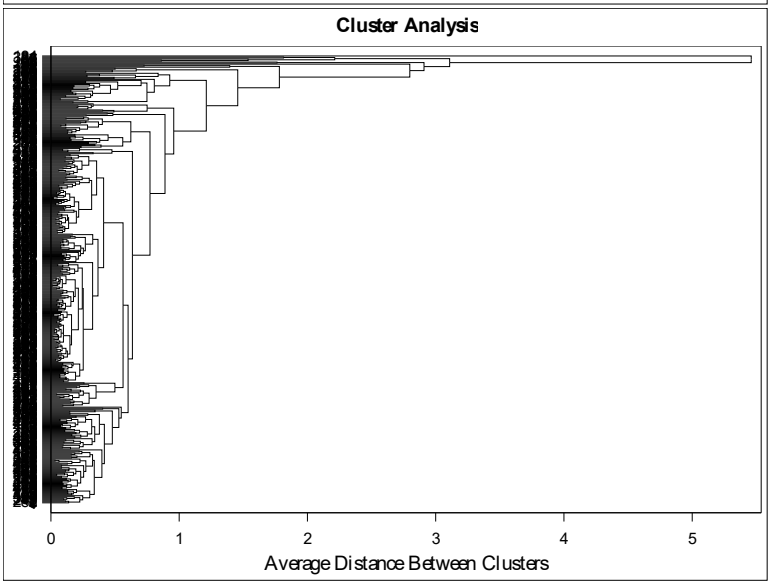
Finally, we did some diagnostic checks, and all the assumptions of a linear regression model were achieved. The explanation of diagnostics plots is presented under Figure 4.2 and Figure 4.3 in Appendix D. Accordingly, we treated this final model as the best model to predict frozen food spending. This also means frozen food, delicatessen and fresh food are positively correlated, while detergents paper and frozen food are negatively correlated.

Question 5:

The distributor is also interested in groupings of observations based on annual spending, and whether those groups are consistent with either the sales channels or the regions. Group the observations based on similarities and differences of annual spending. Identify differences of characteristics for the groups, compare the groups to the channels and regions and explain to the distributor what these results indicate about annual spending characteristics of the channels or regions.

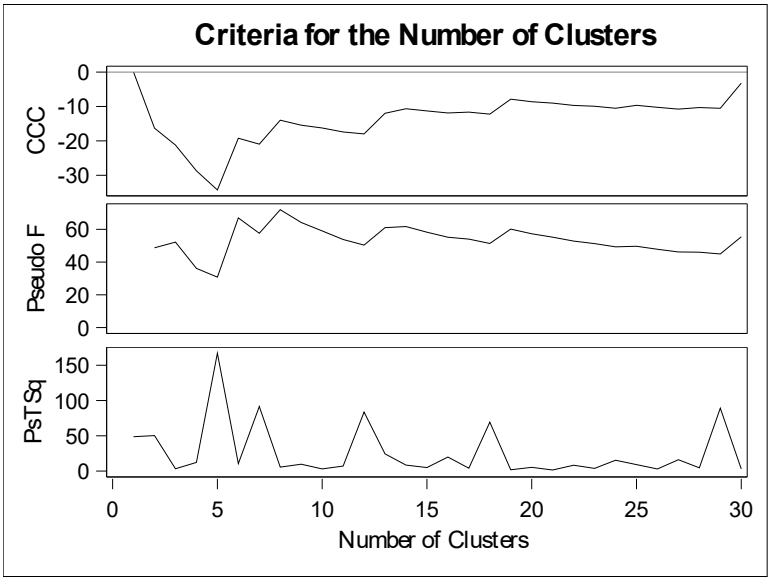
To determine the number of clusters using all types of spending, CCC, Pseudo F and Pseudo T-squared statistics were used. Higher value of CCC and Pseudo F statistic with lower value of Pseudo T-squared statistic are indications of better clustering.

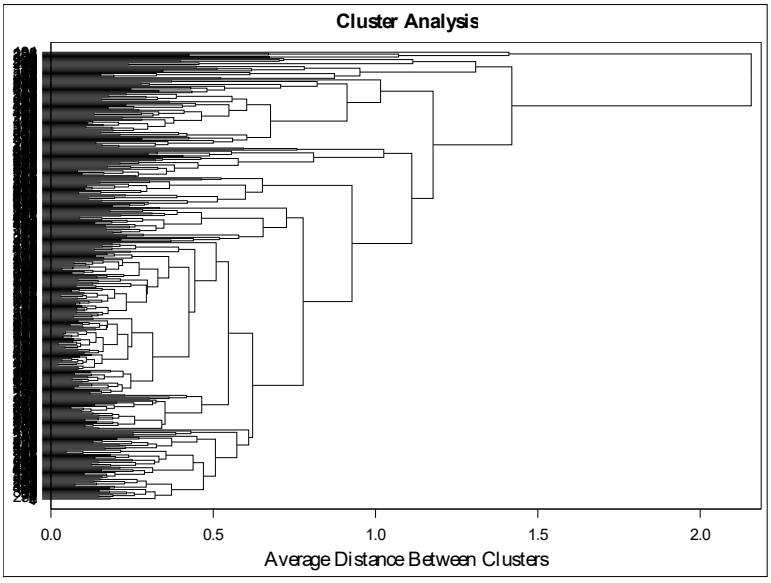




An average linkage was used such that the within cluster distance is minimized and the between cluster distance is maximized. As can be seen from the plots above, CCC achieves its peak value at 30 clusters, and Pseudo F statistic achieves its maximum at 30 clusters. Pseudo T-squared also has a lower value by fitting 30 clusters, so 30 clusters seem to be a good choice based on these criteria. However, we only have 2 channels and 3 regions, so the number of clusters suggested by CCC, Pseudo F and Pseudo T-squared are inconsistent with the truth. Also, since many CCC statistics are negative, there might be some extreme observations that affect the results.

The dendrogram is also helpful to determine the number of clusters. Due to the existence of extreme observations, it is hard to determine how many clusters to be fitted. Even if the clusters that only consists of 1 or 2 observations are removed, the dendrogram still supports more than 3 clusters to be a good choice. The result from dendrogram is also inconsistent with the true number of channels or regions. Therefore, before clustering the data, we removed 19 extreme observations according to the distribution of each type of spending. The removal process is recorded in Appendix E.





Looking at the plots, all of the CCC statistics are negative. This suggests outliers or extreme points still exist, or the average linkage is not appropriate for this study.

The dendrogram still does not have a good separate. The CCC, Pseudo F, Pseudo T-square and the dendrogram suggest about 6 or 8 clusters to work well, but this is a not good clustering comparing to the true number of categories. This means the channels or regions are hard to be clustered based on spending.

To estimate the prediction accuracy and check the consistency of the clusters to the truth, we used 6 clusters for sales channels and regions.

The MEANS Procedure

CLUSTER=1

Variable	N	Mean	Std Dev	Minimum	Maximum
Fresh	331	118.6932931	101.4919313	0.0300000	490.6300000
Milk	331	33.3450755	26.7351881	0.5500000	166.8700000
Grocery	331	44.7260423	34.7739051	0.0300000	169.6600000
Frozen	331	28.1931420	33.0407694	0.2500000	187.1100000
Detergents_Paper	331	12.1417221	15.4726250	0.0300000	72.7100000
Delicassen	331	10.0742296	7.9175701	0.0300000	36.2800000

CLUSTER=2

Variable	N	Mean	Std Dev	Minimum	Maximum
Fresh	64	45.2976562	43.3692854	0.3700000	156.1500000
Milk	64	106.2995313	38.6508619	12.7500000	214.1200000
Grocery	64	179.2779688	61.5024018	80.2500000	364.8600000
Frozen	64	12.8454688	10.6162264	0.3300000	44.2500000
Detergents_Paper	64	81.1868750	32.6941970	8.3600000	171.2000000
Delicassen	64	12.1339063	9.8916360	0.0300000	36.3700000

CLUSTER=3

Variable	N	Mean	Std Dev	Minimum	Maximum
Fresh	13	142.4092308	117.3336453	0.1800000	312.7600000
Milk	13	59.7976923	28.3411492	19.1700000	114.8700000
Grocery	13	66.1315385	40.9334392	16.4100000	152.0500000
Frozen	13	36.6930769	29.9480810	8.3900000	95.1000000
Detergents_Paper	13	17.8130769	15.9821434	2.3500000	47.9700000
Delicassen	13	52.8930769	12.4639917	30.9500000	78.4400000

CLUSTER=4

Variable	N	Mean	Std Dev	Minimum	Maximum
Fresh	7	53.1771429	41.4719635	2.0000000	106.8300000
Milk	7	214.8728571	27.4106662	179.7200000	258.6200000
Grocery	7	146.8728571	88.5184387	16.6000000	259.5700000
Frozen	7	40.1985714	33.5048372	6.5100000	101.5500000
Detergents_Paper	7	39.8928571	40.2303947	2.8200000	87.7300000
Delicassen	7	46.5171429	10.3691541	32.6500000	62.5000000

CLUSTER=5

Variable	N	Mean	Std Dev	Minimum	Maximum
Fresh	5	82.6160000	31.5677182	40.9800000	121.1900000
Milk	5	267.7880000	26.8210770	231.3300000	298.9200000
Grocery	5	299.6500000	82.6789036	176.4500000	396.9400000
Frozen	5	36.9160000	21.4244739	11.2800000	67.4600000
Detergents_Paper	5	174.1160000	28.6220034	124.0800000	194.1000000
Delicassen	5	32.1720000	18.3458352	13.4000000	51.3000000

CLUSTER=6

Variable	N	Mean	Std Dev	Minimum	Maximum
Fresh	1	220.3900000	.	220.3900000	220.3900000
Milk	1	83.8400000	.	83.8400000	83.8400000
Grocery	1	347.9200000	.	347.9200000	347.9200000
Frozen	1	0.4200000	.	0.4200000	0.4200000
Detergents_Paper	1	125.9100000	.	125.9100000	125.9100000
Delicassen	1	44.3000000	.	44.3000000	44.3000000

We first observe that the number of observations in each of 6 clusters is very different. Some clusters have a large sample size, while some clusters have very few samples. The 6th cluster has only 1 data.

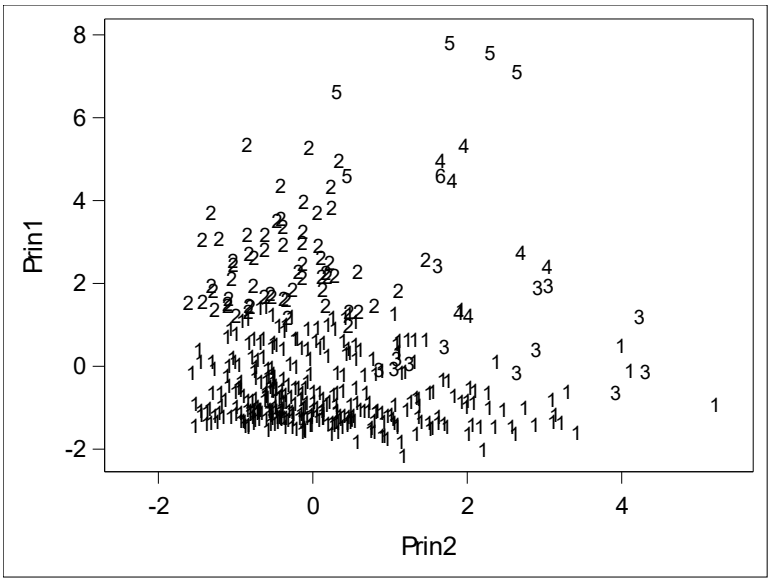
For the remaining 5 clusters, cluster 1 and cluster 3 seems to have more spending on fresh food. Cluster 4 and Cluster 5 tend to have more spending on milk, with cluster 2 the third largest. In terms of grocery spending, cluster 5 tends to be the greatest, with cluster 2 and 4 being in the middle. In terms of frozen food spending, there is not huge difference between the clusters. For detergents paper spending, cluster 5 has the largest amount and cluster 2 has the second largest amount. For delicatessen spending, cluster 3 and cluster 4 seems to have the greatest amount. And overall, the variability in each type of spending for each cluster is pretty high, which is a good indication of why CCC, Pseudo F and Pseudo T-squared with the dendrogram did not indicate easiness in clustering the data.

The PRINCOMP Procedure

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	2.76718884	1.38430141	0.4612	0.4612
2	1.38288743		0.2305	0.6917

Eigenvectors		
	Prin1	Prin2
Fresh	-.170199	0.566941
Milk	0.526947	0.164450
Grocery	0.566321	0.006096
Frozen	-.123790	0.585188
Detergents_Paper	0.551945	-.085593
Delicassen	0.229507	0.549295

For principal component analysis, the first 2 components describe about 69.17% in the variation of spending. The eigenvectors suggest that principal component 1 is contrasting milk, grocery, detergents paper and delicatessen with fresh food and frozen food. Principal component 2 is less clear, but it seems to contrast detergents paper with other type of spending because it is the only variable with a negative value.



Moving on to the plot above, cluster 4 and 5 are high in principal component 1 and 2, so these clusters end to have greater spending in milk, grocery, detergents paper and delicatessen. Cluster 1 is negative in principal component 1 and positive in component 2, so cluster 1 seems to have greater spending in fresh food and frozen food and relatively large spending in grocery, milk and delicatessen. Cluster 2 is high in component 1 and low in component 2, so cluster 2 tends to have more spending on milk, grocery and delicatessen. Finally, cluster 3 is only high in component 2, so cluster 3 tends to have more spending on all variables except for detergents paper.

The FREQ Procedure

Table of CLUSTER by channelname			
CLUSTER	**channelname**		
Frequency	**Horeca**	**Retail**	**Total**
1	270	61	331
2	3	61	64
3	9	4	13
4	4	3	7
5	0	5	5
6	0	1	1
Total	286	135	421

Table of CLUSTER by regionname				
CLUSTER	regionname			
Frequency	Lisbon	Oporto	Other	Total
1	59	34	238	331
2	10	10	44	64
3	2	0	11	13
4	2	0	5	7
5	2	1	2	5
6	0	0	1	1
Total	75	45	301	421

If we check the prediction results, the frequency tables for both channels and regions did not split the groups well. For channels, cluster 1, 2, 3 and 4 mixed Horeca and Retail observations together, which means the clusters are inconsistent with the channels. Similarly, clusters for regions mixed observations for the three regions together, of which still shows inconsistency.

Based on previous analysis, it is hard to cluster channels or regions based on spending. Since cluster 1 almost include all samples, the clustering fails to separate channels or regions well.

3.0 – Conclusion

In this section, we present a summary of our analysis of spending of different hotels/restaurants/cafes and regions.

We first had a brief of summary of the characteristics of two channels. Two channels have different mean for all types of spending based on Wilcoxon two-sample tests. We analyzed the significant types of spending to predict whether a business is Horeca or Retail by applying stepwise selection to the logistic regression model, and we ended up with a model including fresh food and detergents paper spending as significant predictors. We found that both the fresh food and detergents paper spending would make a business more likely to be a Retail business.

We also analyzed the significant types of parameters to determine where a business comes from. The quadratic discriminant analysis was used for clustering and prediction, and due to the insignificance of variables at lower threshold of p-value for stepwise selection, we increase the threshold up to 0.33, and the best model includes fresh food, frozen food, detergents paper and milk spending as significant variables. According to the high and uneven classification error using cross-validation, we concluded that the regions could not be distinguished well based on spending alone.

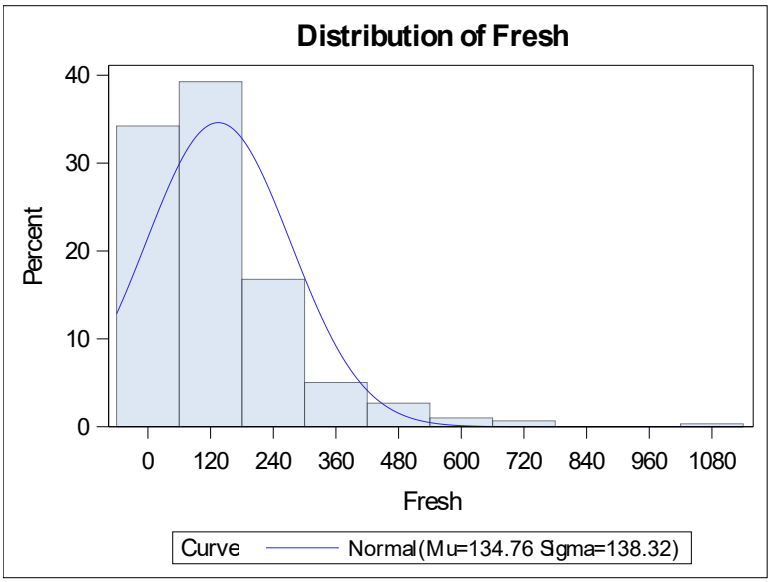
To avoid multicollinearity between spending, we performed model selection on linear regression to predict frozen food spending with other types of spending as predictors. The stepwise, forward and backward selection result in the same final model, which includes fresh food, delicatessen and detergents paper spending as significant variables. The diagnostics plots and VIF value were checked such that all the assumptions were satisfied and multicollinearity issues were removed. Using the linear regression output, we concluded that delicatessen and fresh food have a positive linear relationship to frozen food spending, and detergents paper has a negative linear relationship to frozen food spending.

Finally, we explored the characteristics of channels and regions through clustering designs. The data was hard to cluster based on spending even after removing extreme observations, and we selected 6 clusters according to the CCC, Pseudo F and Pseudo T-square statistics with trends shown in the dendrogram. Additionally, by taking principal component analysis, we found that cluster 1 has greater spending in fresh food and frozen food and relatively large spending in grocery, milk and delicatessen. Cluster 2 tends to have more spending on milk, grocery and delicatessen. Cluster 3 tends to have more spending on all variables except for detergents paper. Cluster 4 and 5 tend to have greater spending in milk, grocery, detergents paper and delicatessen. Cluster 6 only includes one observation so we did not take it into consideration. Finally, since the clustering results in prediction mixed the observations from different channels or regions together, we concluded that the clustering results are inconsistent to the true channels or regions. The clustering fails to separate channels or region well, which suggests the channels or regions might have similar characteristics in terms of spending.

4.0 – Appendix

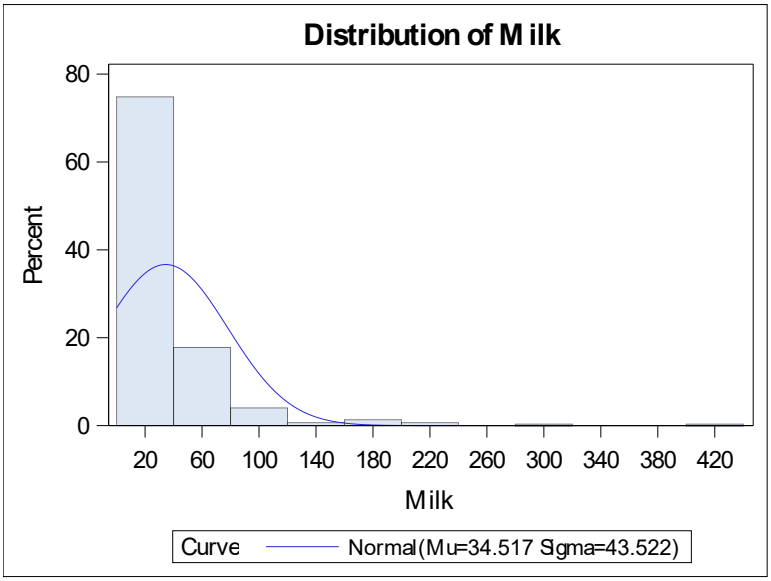
Appendix A

Tests of normality and histogram for each type of spending:



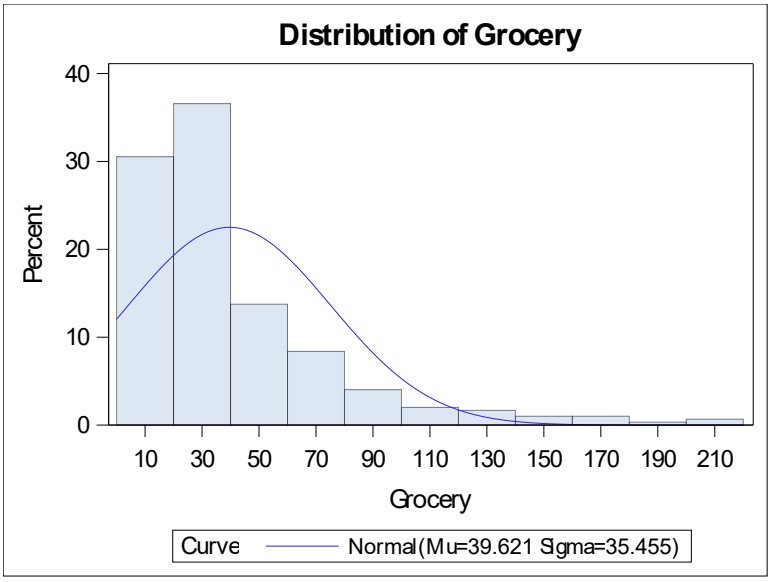
The UNIVARIATE Procedure
Variable:
Milk

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.584983	**Pr < W**	<0.0001
Kolmogorov-Smirnov	**D**	0.218075	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	5.236625	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	28.97205	**Pr > A-Sq**	<0.0050



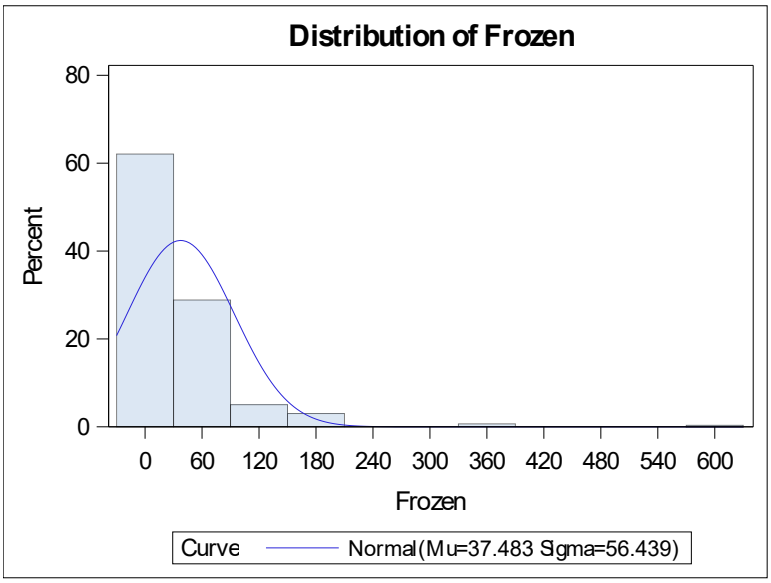
The UNIVARIATE Procedure
Variable:
Grocery

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.782827	**Pr < W**	<0.0001
Kolmogorov-Smirnov	**D**	0.177945	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	3.251138	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	18.47584	**Pr > A-Sq**	<0.0050



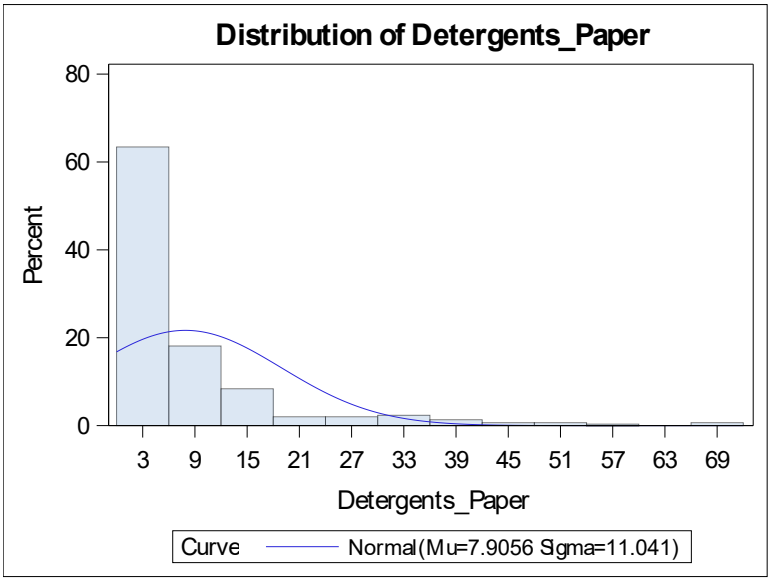
The UNIVARIATE Procedure
Variable:
Frozen

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.558747	**Pr < W**	<0.0001
Kolmogorov-Smirnov	**D**	0.254725	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	5.650255	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	30.3912	**Pr > A-Sq**	<0.0050



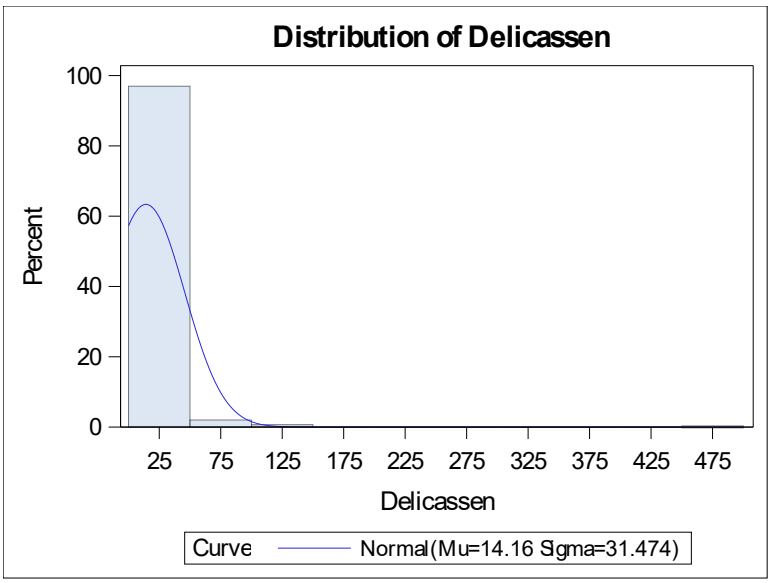
The UNIVARIATE Procedure
Variable:
Detergents_Paper

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.645859	**Pr < W**	<0.0001
Kolmogorov-Smirnov	**D**	0.237827	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	6.261785	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	33.59951	**Pr > A-Sq**	<0.0050

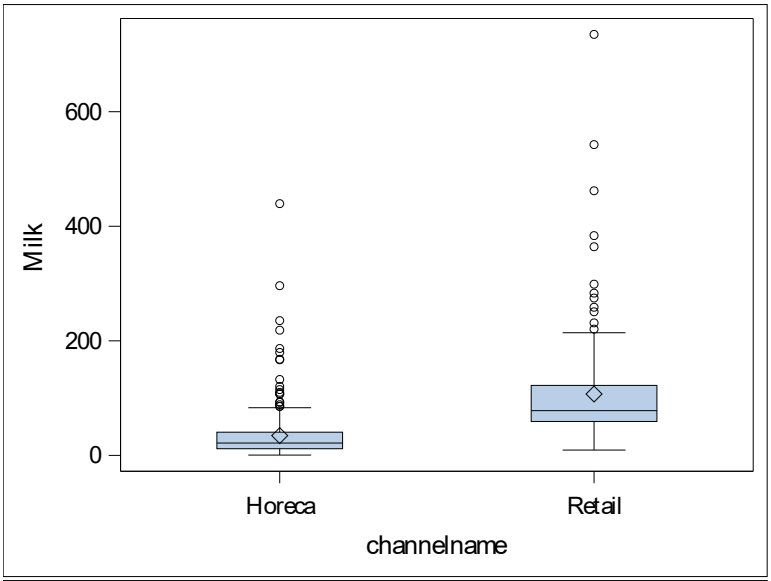


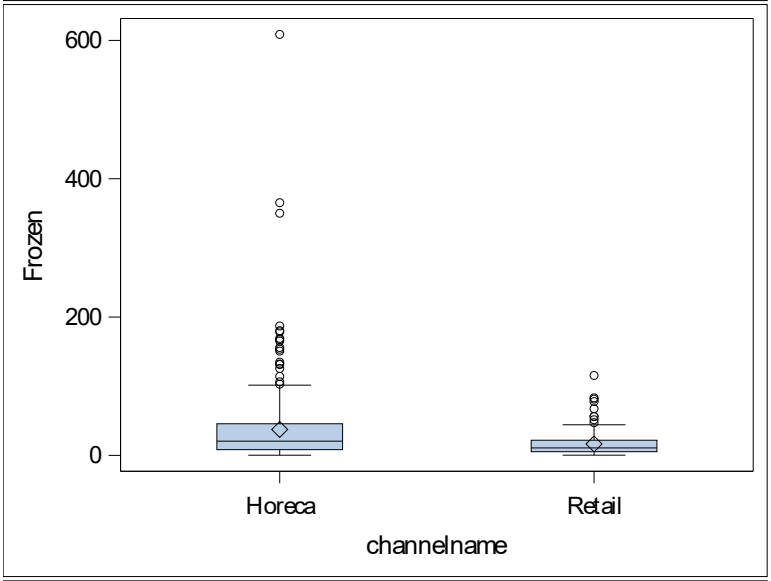
The UNIVARIATE Procedure
Variable:
Delicassen

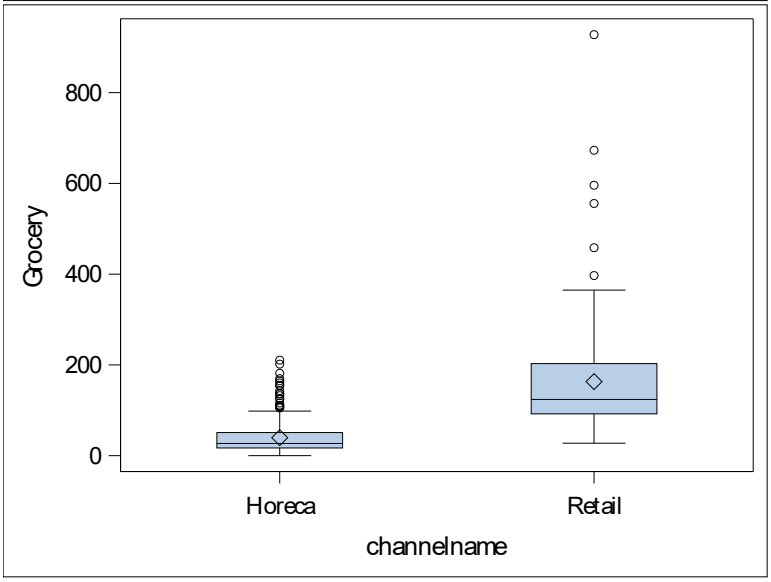
Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.288902	**Pr < W**	<0.0001
Kolmogorov-Smirnov	**D**	0.326743	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	9.894898	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	51.40086	**Pr > A-Sq**	<0.0050

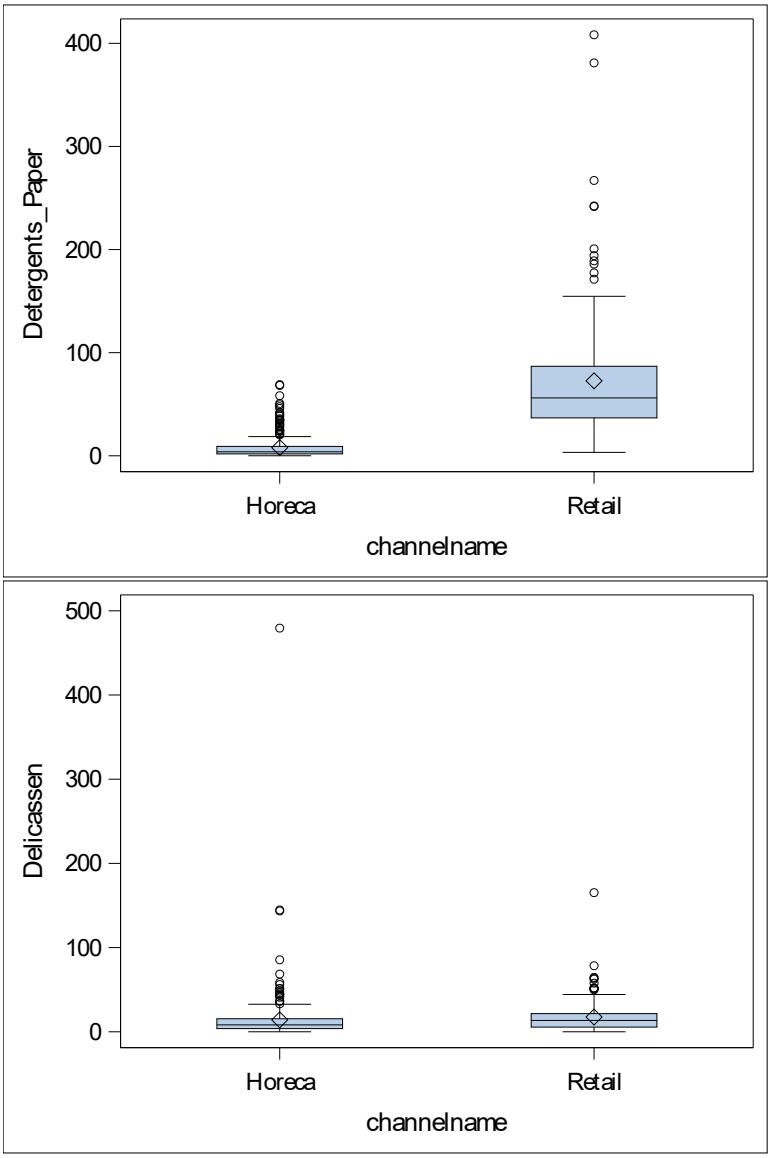


Boxplot for spending comparisons between channels (fresh food is presented under Question 1 for section 2.0):









Appendix B

Global test result of logistic regression model:

The LOGISTIC Procedure

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	342.1563	2	<.0001
Score	179.2924	2	<.0001
Wald	103.9829	2	<.0001

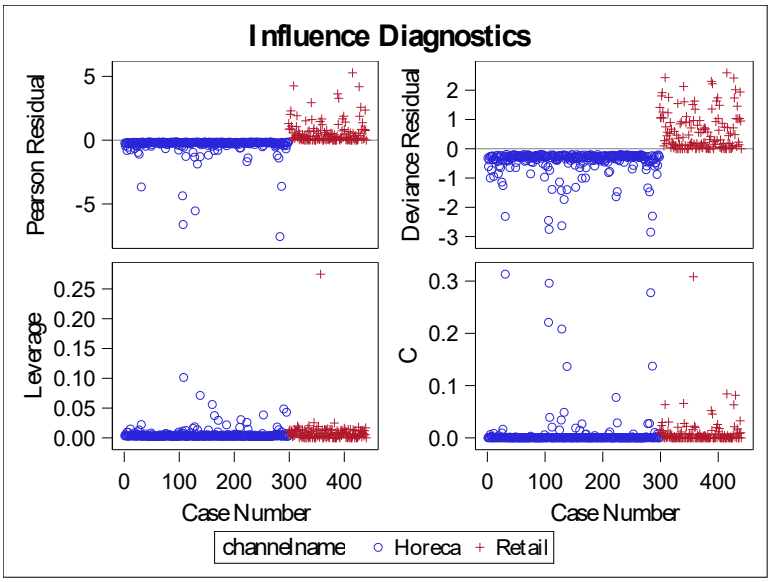
Forward selection:

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-3.9880	0.3749	113.1395	<.0001
Grocery	1	0.0129	0.00454	8.0901	0.0045
Detergents_Paper	1	0.0928	0.0123	57.1691	<.0001

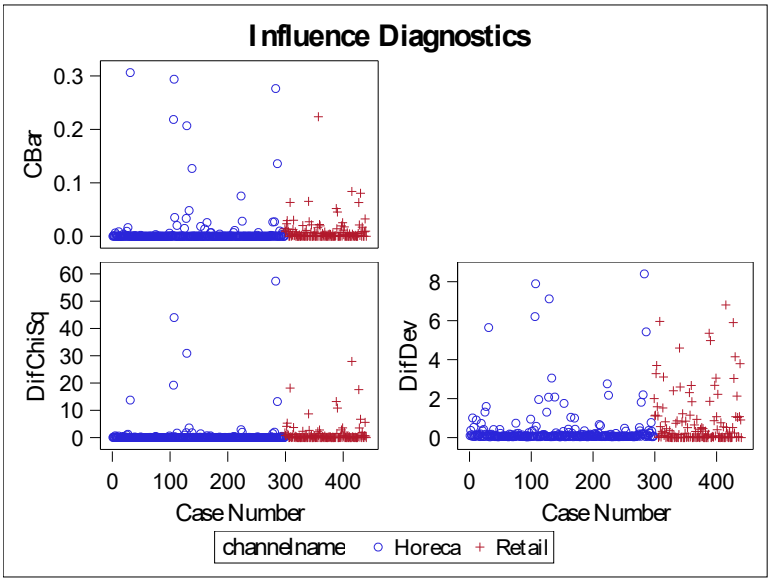
Backward selection:

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-3.9880	0.3749	113.1395	<.0001
Grocery	1	0.0129	0.00454	8.0901	0.0045
Detergents_Paper	1	0.0928	0.0123	57.1691	<.0001

Diagnostic plots:



The residuals and deviance residuals do not have any obvious pattern, indicating homoscedasticity assumption is satisfied. The leverage only has one high leverage value, and hence this potential influential point does not affect the model too much.



Appendix D

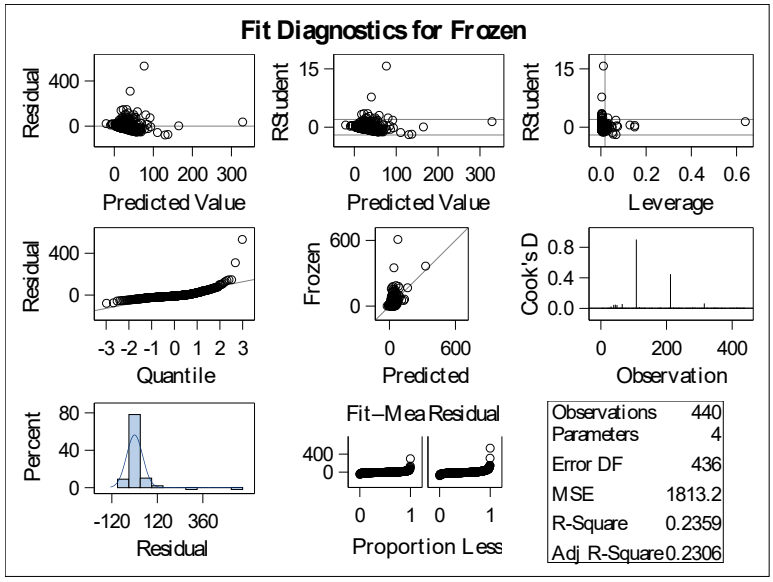


Figure 4.1

Two observations have very large Cook's distance. We removed them and checked the diagnostic plots in Figure 4.2.

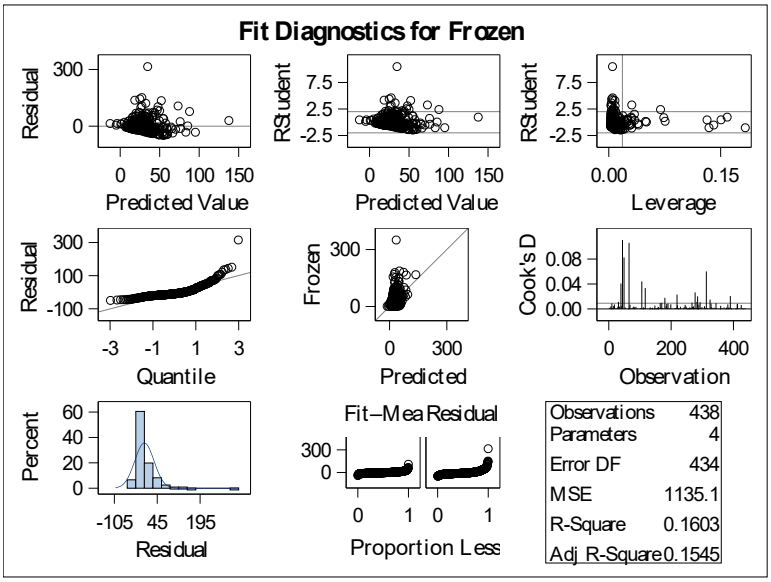
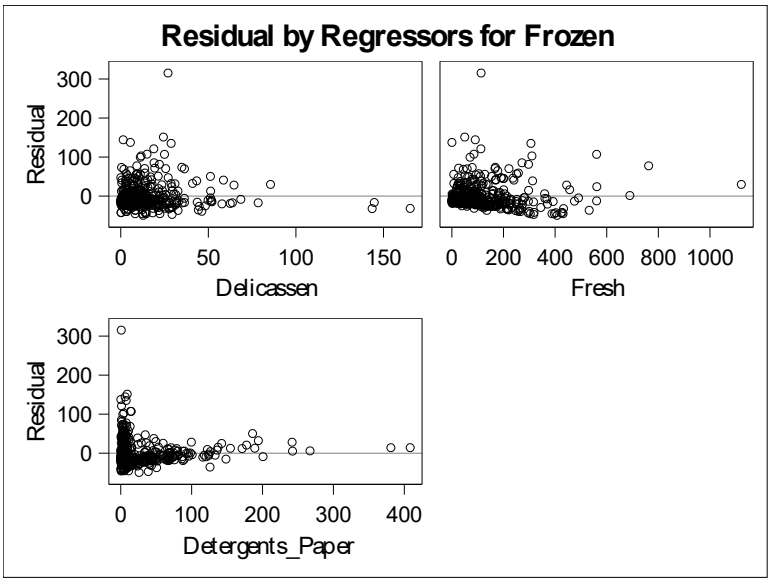


Figure 4.2

After removing data with large Cook's distance, the samples' Cook's distance has similar value and are below 0.1, which means influential points or potential outliers were removed. The constant variance assumption is satisfied because there is no pattern in residuals vs fitted plot, but normality assumption might be violated since the histogram does not have a bell shape. The points on Residuals vs Predicted value plot are almost on the fitted line, so normality assumption should be satisfied.



Appendix E

19 extreme observations were removed. Their observation number are "182","126","285","87","48","326","86","334","184","62","94", "66","24","88","72","40","259","104","260").

The removed observations are based on the tables of extreme observations below:

Fresh

Extreme Observations					
Lowest			Highest		
Value	ID	Obs	Value	ID	Obs
0.03	339	339	560.83	259	259
0.03	96	96	561.59	40	40
0.09	67	67	689.51	285	285
0.18	219	219	762.37	126	126
0.23	97	97	1121.51	182	182

Variable:
Milk

Extreme Observations					
Lowest			Highest		
Value	ID	Obs	Value	ID	Obs
0.55	155	155	383.69	62	62
1.12	99	99	439.50	184	184
1.34	357	357	461.97	86	86
2.01	123	123	542.59	48	48
2.54	98	98	734.98	87	87

Grocery

Extreme Observations					
Lowest			Highest		
Value	ID	Obs	Value	ID	Obs
0.03	76	76	458.28	66	66
1.37	155	155	555.71	48	48
2.18	357	357	595.98	62	62
2.23	276	276	672.98	334	334
2.45	123	123	927.80	86	86

Frozen

Extreme Observations					
Lowest			Highest		
Value	ID	Obs	Value	ID	Obs
0.25	421	421	180.28	104	104
0.33	39	39	187.11	197	197
0.36	66	66	350.09	94	94
0.38	58	58	365.34	184	184
0.42	146	146	608.69	326	326

Detergents_Paper

Extreme Observations					
Lowest			Highest		
Value	ID	Obs	Value	ID	Obs
0.03	162	162	241.71	48	48
0.03	76	76	242.31	66	66
0.05	205	205	267.01	62	62
0.07	155	155	381.02	334	334
0.09	357	357	408.27	86	86

Delicassen

Extreme Observations					
Lowest			Highest		
Value	ID	Obs	Value	ID	Obs
0.03	188	188	85.50	182	182
0.03	143	143	143.51	88	88
0.03	129	129	144.72	72	72
0.03	110	110	165.23	24	24
0.07	234	234	479.43	184	184