**BUS 473 Homework 2 – Due September 26, 2018**

**Again you should write up individually but may work together particularly on R program, please submit as an html file generated by markdown.**

1. For the data set pepsicokes posted on BlackBoard, do the following R operations:

a. Bring the file into R;

b. Create a dummy variable whitecollar ( 1 if white collar, 0 otherwise), representing

stores with white collar percent > 60 and calculate frequencies for this variable; the white collar percentages on the dataset are represented by the variable WhiteCollarPctPen.

c. Do a crosstab of Prizm cluster (the 16 social groups) with Chain; and another crosstab of Prizm cluster with whitecollar (use table() function); the social groups are represented on the dataset by the variable prizm\_cluster.

d. Calculate descriptive statistics for any 5 continuous variables in the dataset. (you should already know from the previous assignment that what function you must use!)

e. Regress Corp\_Pep\_Volume\_per\_\_MM\_ACV (million dollars All Commodity Volume = Total Store Sales) on a set of any 6 independent variables. (use lm() for linear regression). Output the results using summary().

f. Run a Hierarchical Cluster Analysis using the variables Pepsi Volume per MM ACV, Pepsi Price per MM ACV, Coke Volume per MM ACV, Coke Price per MM ACV. Generate a Scree plot and find the kink point to determine the optimal number of clusters. Code for this is available in Kabacoff Chapter 16.4.1)

g. Run a K-means Cluster Analysis using the number of clusters determined in d.

2. Consider the regression below (and on the next page) that was estimated on weekly data over a 2-year period on a sample of Kroger stores for Pepsi carbonated soft drinks. The dependent variable is the log of Pepsi volume per MM ACV. There are 53 stores in the dataset (data were missing for some stores in some weeks). Please answer the following questions about the regression output.

a. Comment on the goodness of fit and significance of the regression, and of individual variables. What does the ANOVA table reveal?

b. Write out the equation and interpret the meaning of each of the parameters.

c. What is the price elasticity? The cross-price elasticity with respect to Coke price? Are these results reasonable? Explain

d. What do the results tell you about the effectiveness of Pepsi and Coke display and advertising?

e. What are the 3 most important variables? Explain how you arrived at this conclusion.

f. What is collinearity? Is collinearity a problem for this regression? Explain. If it is a problem, what action would you take to deal with it?

g. What changes to this regression equation, if any, would you recommend? Explain.

**Model Summary(b)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
| 1 | .869(a) | .754 | .754 | .4120 |

a Predictors: (Constant), Mass stores in trade area, Labor Day dummy, Pepsi advertising days, Store traffic, Memorial Day dummy, Pepsi display days, Coke advertising days, Log of Pepsi price, Coke display days, Log of Coke price

b Dependent Variable: Log of Pepsi volume/MM ACV

**ANOVA(b)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model |  | Sum of Squares | df | Mean Square | F | Sig. |
| 1 | Regression | 2881.089 | 10 | 288.109 | 1697.262 | .000(a) |
|  | Residual | 937.695 | 5524 | .170 |  |  |
|  | Total | 3818.784 | 5534 |  |  |  |

a Predictors: (Constant), Mass stores in trade area, Labor Day dummy, Pepsi advertising days, Store traffic, Memorial Day dummy, Pepsi display days, Coke advertising days, Log of Pepsi price, Coke display days, Log of Coke price

b Dependent Variable: Log of Pepsi volume/MM ACV

**Coefficients(a)**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model |  | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | Collinearity Statistics | |
|  |  | B | Std. Error | Beta |  |  | Tolerance | VIF |
| 1 | (Constant) | 7.79429 | .06249 |  | 124.721 | .000 |  |  |
|  | Log of Pepsi price | -3.34665 | .03483 | -.739 | -96.091 | .000 | .751 | 1.332 |
|  | Log of Coke price | .65877 | .03170 | .181 | 20.784 | .000 | .587 | 1.703 |
|  | Pepsi advertising days | .00173 | .00020 | .065 | 8.644 | .000 | .784 | 1.275 |
|  | Coke advertising days | -.00009 | .00018 | -.004 | -.502 | .616 | .689 | 1.450 |
|  | Pepsi display days | .00011 | .00021 | .004 | .546 | .585 | .656 | 1.525 |
|  | Coke display days | -.00299 | .00020 | -.123 | -14.766 | .000 | .646 | 1.549 |
|  | Labor Day dummy | .27190 | .04167 | .045 | 6.525 | .000 | .923 | 1.083 |
|  | Memorial Day dummy | .21295 | .04269 | .036 | 4.988 | .000 | .834 | 1.199 |
|  | Store traffic | .00000 | .00000 | .023 | 3.367 | .001 | .961 | 1.040 |
|  | Mass stores in trade area | -.00910 | .00026 | -.238 | -35.161 | .000 | .968 | 1.033 |

a Dependent Variable: Log of Pepsi volume/MM ACV

3. The tables on the following 4 pages summarize the output of K-Means clustering for the WNY soft drink file that I used to demonstrate regression.

Since clusters should reflect consumer and firm behavior for the focal product category, I used the following bases for clustering: Pepsi Volume per MM ACV, Pepsi Price per MM ACV, Coke Volume per MM ACV, Coke Price per MM ACV. These variables were all standardized to mean 0, standard deviation 1 (Zscores) before clustering.

A 4-cluster solution, which I think provides a good description of this market, is provided below. Output consists of averages of standard scores (Zscores) for each cluster, ANOVA tests of significant differences between these averages across clusters, and number of stores in each cluster.

Descriptor variables for each cluster were taken to be: Currentpop, MedianHHincome,

MedianYrsSchool, WhiteCollarPctPen, Farm\_Forest\_FishPctPen, BlueCollarPctPen,

MedianHomeValue, WhitePopPctPen, BlackPopPctPen, Groc\_Miles,

Mass\_Miles. Descriptives and ANOVA tests are presented for the continuous variables. Cross-tabs are presented for chain and prizm social group.

a. Interpret the results of the clustering. Name each cluster, and describe what it stands for. Explain your choices.

b. What can you say from the results about the market for soft drinks, and about the apparent strategies of Tops and Wegmans, and Pepsi and Coke?

c. Develop a strategy for targeting each of the four segments for Pepsi.

**K-Means Clustering of WNY Data**

**Bases**

**Final Cluster Centers**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Corp\_Pepsi\_ |  | Corp\_Coke\_ |  |
|  | Volume\_per\_\_ | Corp\_Pepsi\_ | Volume\_per\_\_ | Corp\_Coke\_ |
| Cluster | MM\_ACV | Price | MM\_ACV | Price |
| 1 | -0.9092 | 1.1475 | 0.8399 | -0.8676 |
| 2 | 0.2813 | -0.5680 | -1.2003 | -0.6806 |
| 3 | -0.1705 | -0.0227 | -0.1792 | 1.2173 |
| 4 | 1.5359 | -1.1635 | 0.6072 | 0.1488 |
| Means are for standardized variables - scale = mean 0, stdev = 1 | | | | |

**ANOVA**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Cluster | | Error | | F | Sig. |  |
| Variable | Mean Square | df | Mean Square | df | R-Square |
| Corp\_Pep\_Volume\_per\_\_MM\_ACV | 29.3050 | 3 | 0.3567 | 132 | 87.92 | 0.0000 | 0.6512 |
| Corp\_Pep\_Price | 30.4930 | 3 | 0.3297 | 132 | 91.48 | 0.0000 | 0.6776 |
| Corp\_Coke\_Volume\_per\_\_MM\_ACV | 27.7033 | 3 | 0.3931 | 132 | 83.11 | 0.0000 | 0.6156 |
| Corp\_Coke\_Price | 36.1290 | 3 | 0.2016 | 132 | 108.39 | 0.0000 | 0.8029 |

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

**Cluster Summary and Number of Cases in each Cluster**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Cluster | Frequency in Cluster | RMS Std Deviation | Maximum Distance from Seed to Observation | Nearest Cluster | Distance Between Cluster Centroids |
|
|
| 1 | 39 | 0.4396 | 1.7172 | 3 | 2.7020 |
| 2 | 32 | 0.4881 | 2.6067 | 3 | 2.2685 |
| 3 | 43 | 0.5985 | 2.8373 | 2 | 2.2685 |
| 4 | 22 | 0.7716 | 2.2536 | 2 | 2.4256 |

**Descriptors**

**Descriptives**

|  |  |  |  |
| --- | --- | --- | --- |
| Variable and Cluster | N | Mean | Std Dev |
| Currentpop |  |  |  |
| 1 | 39 | 44707.74 | 17134.62 |
| 2 | 32 | 39931.44 | 15379.06 |
| 3 | 43 | 54964.19 | 23737.98 |
| 4 | 22 | 14805.00 | 14126.24 |
| Overall | 136 | 41989.54 | 22814.25 |
| MedianHHincome |  |  |  |
| 1 | 39 | 48707.69 | 14214.63 |
| 2 | 32 | 44103.84 | 13361.84 |
| 3 | 43 | 43920.47 | 12139.50 |
| 4 | 22 | 36287.27 | 4185.07 |
| Overall | 136 | 44101.64 | 12748.94 |
| MedianYrsSchool |  |  |  |
| 1 | 39 | 13.48 | 0.82 |
| 2 | 32 | 13.15 | 0.63 |
| 3 | 43 | 13.23 | 0.74 |
| 4 | 22 | 12.74 | 0.23 |
| Overall | 136 | 13.20 | 0.72 |
| WhiteCollarPctPen |  |  |  |
| 1 | 39 | 61.64 | 9.29 |
| 2 | 32 | 57.73 | 9.61 |
| 3 | 43 | 57.25 | 9.15 |
| 4 | 22 | 48.23 | 5.09 |
| Overall | 136 | 57.16 | 9.72 |
| Farm\_Forest\_FishPctPen |  |  |  |
| 1 | 39 | 1.36 | 1.26 |
| 2 | 32 | 1.42 | 1.74 |
| 3 | 43 | 1.93 | 1.71 |
| 4 | 22 | 4.15 | 2.95 |
| Overall | 136 | 2.00 | 2.09 |
| BlueCollarPctPen |  |  |  |
| 1 | 39 | 24.25 | 6.78 |
| 2 | 32 | 26.32 | 6.31 |
| 3 | 43 | 26.59 | 6.53 |
| 4 | 22 | 29.42 | 5.01 |
| Overall | 136 | 26.31 | 6.49 |

|  |  |  |  |
| --- | --- | --- | --- |
| Variable and Cluster | N | Mean | Std Dev |
| MedianHomeValue |  |  |  |
| 1 | 39 | 104816.67 | 28068.36 |
| 2 | 32 | 92660.03 | 26336.36 |
| 3 | 43 | 96370.60 | 29053.05 |
| 4 | 22 | 85279.36 | 15407.20 |
| Overall | 136 | 96125.39 | 26915.04 |
| WhitePopPctPen |  |  |  |
| 1 | 39 | 83.84 | 16.49 |
| 2 | 32 | 86.13 | 17.39 |
| 3 | 43 | 85.58 | 18.46 |
| 4 | 22 | 95.20 | 3.95 |
| Overall | 136 | 86.77 | 16.40 |
| BlackPopPctPen |  |  |  |
| 1 | 39 | 9.56 | 12.73 |
| 2 | 32 | 9.38 | 16.09 |
| 3 | 43 | 8.67 | 13.90 |
| 4 | 22 | 1.88 | 2.20 |
| Overall | 136 | 7.99 | 13.16 |
| Groc\_Miles |  |  |  |
| 1 | 39 | 2.90 | 0.69 |
| 2 | 32 | 3.20 | 1.60 |
| 3 | 43 | 2.61 | 1.60 |
| 4 | 22 | 5.54 | 4.27 |
| Overall | 136 | 3.31 | 2.32 |
| Mass\_Miles |  |  |  |
| 1 | 39 | 2.49 | 0.90 |
| 2 | 32 | 2.52 | 1.01 |
| 3 | 43 | 2.30 | 1.67 |
| 4 | 22 | 4.87 | 4.81 |
| Overall | 136 | 2.82 | 2.40 |

**ANOVA Tests of Equality of Group Means**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | DF Cluster | DF Error | F | PROB |
| Currentpop | 3 | 132 | 22.7096 | 0.00000 |
| MedianHHincome | 3 | 132 | 4.8345 | 0.00317 |
| MedianYrsSchool | 3 | 132 | 5.6417 | 0.00114 |
| WhiteCollarPctPen | 3 | 132 | 10.9795 | 0.00000 |
| Farm\_Forest\_FishPctPen | 3 | 132 | 12.2661 | 0.00000 |
| BlueCollarPctPen | 3 | 132 | 3.1651 | 0.02670 |
| MedianHomeValue | 3 | 132 | 2.8355 | 0.04066 |
| WhitePopPctPen | 3 | 132 | 2.5265 | 0.06025 |
| BlackPopPctPen | 3 | 132 | 1.9613 | 0.12291 |
| Groc\_Miles | 3 | 132 | 10.2798 | 0.00000 |
| Mass\_Miles | 3 | 132 | 7.3739 | 0.00013 |

**Chain \* Cluster Number of Case Crosstabulation**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Frequency | Cluster | | | |  |
| Row Pct | 1 | 2 | 3 | 4 | Total |
| TOPS | 0 | 31 | 32 | 22 | 85 |
|  | 0 | 36.47 | 37.65 | 25.88 |  |
| WEGMANS | 39 | 1 | 11 | 0 | 51 |
|  | 76.47 | 1.96 | 21.57 | 0 |  |
| Total | 39 | 32 | 43 | 22 | 136 |

**Chi-Square Tests for Chain \* Cluster Number of Case Crosstabulation**

Statistic DF Value Prob

ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ

Chi-Square 3 96.9395 <.0001

Likelihood Ratio Chi-Square 3 122.1429 <.0001

Mantel-Haenszel Chi-Square 1 58.5519 <.0001

**Prizm\_Cluster Crosstabulation Cluster Number of Case \***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Cluster Number of Cases | | | |
| Cluster | 1 | 2 | 3 | 4 |
| C1 | 2 | 0 | 0 | 1 |
| C2 | 3 | 4 | 6 | 0 |
| C3 | 1 | 4 | 5 | 0 |
| S2 | 7 | 1 | 6 | 0 |
| S3 | 11 | 7 | 7 | 0 |
| S4 | 2 | 4 | 2 | 0 |
| T1 | 2 | 4 | 1 | 0 |
| T2 | 4 | 2 | 3 | 1 |
| T3 | 0 | 2 | 3 | 13 |
| T4 | 4 | 0 | 5 | 7 |
| U1 | 0 | 1 | 0 | 0 |
| U2 | 1 | 0 | 0 | 0 |
| U3 | 2 | 3 | 5 | 0 |