Theoretical and Applied Economics
Volume XXI (2014), No. 12(601), pp. 3-12

Data mining of transactional data for sales of dairy products

Julian VASILEV
Varna University of Economics, Bulgaria
vasilev@ue-varna.bg

Abstract. *The purpose of this article is to find out the most important factors for sales of dairy products. Transactional data for sales are used. This article is published for the first time. The results of the analysis will show how transactional data may be used for data mining. SPSS is used to analyze the data. Statistical methods (such as independent samples t-test, Chi-square tests, Mann-Whitney U test, one-way ANOVA test) are applied to find dependencies. We suppose that the quantity of sales depends on the year, month, group, item and average price. At the end of the study we found out that "year" and "group" are factors which mostly influence the quantity of a sale. Factors "month", "item" and "average sale price" do not affect the quantity of a single sale. Mathematical formulas are derived to predict the quantity of a future sale on the basis of independent variables such as year and group of stock.*

Keywords: sales analysis, dairy products, SPSS.

JEL Classification: C81.
REL Classification: 10B.

1. Introduction

Sales analysis is a common technique adapted in practice. POS systems generate thousands of transactions monthly. All these transactions may be used to analyze data. The simplest way of analysis is to report sales on a yearly, monthly or quarterly basis. A more detailed view allows drilling down the transactional data. Pivot tables or crosstab tables may be produced when grouping the sales on two factors – for instance month and year. Grouping may be done by customers and products. All these reports use simple SELECT queries for extracting data within database management systems (DBMS).

Select queries may be also used for grouping of data. Grouping of data allows the end user to generate custom aggregated reports. Usually one variable (or more variables) is used for grouping and another variable (variables) is used for summarizing data. The first variable (the first group of variables) may have text values (for instance EAN, customer name, and month). The second one (the second group of variables) must have numeric values in order to make calculations – using the function SUM, AVERAGE, STDEV or another one.

The technique for grouping transactional data allows confidential data to be hidden. Summarized values do not show the values of each transaction. They also do not show the number of transactions.

Data mining is a well-known technique for analyzing datasets. The purpose of the application of data mining is to find dependencies within data which may not be marked or highlighted by simple select queries and sales reports.

2. Recent research on sales analysis

Sales analysis is an interesting theme during periods of crisis. Different authors focus on specific aspects and particular methods. Some articles (Cheng et al., 2013) focus on sales performance. Other articles (Celik et al., 2013) focus on specific methods – for instance panel data analysis. Contemporary instruments such as neural networks are used to predict food demand (Huang, 2013). Sales for the past 6 months are used to predict future sales. A recent article (Azad et al., 2013) focuses on the factor analysis of dairy products. The last cited paper investigates factors such as salesman qualification, motivation, personality and content motivation. Our research does not cover behavioral aspects of sales force automation. Other articles treat another product, group of products or specific type of business.

Factor analysis is a well-known method applied in sales in the sphere of customer loyalty (Hanzaee et al., 2013). Our paper extends the possibilities of the factor analysis. A group of authors (Spadoni et al., 2014) use factor analysis to assess the applicability and impact of private food standards. They classify food companies into groups by a qualitative investigation based on in-depth interviews. Factor analysis is used (Giritlioglu et al., 2014) to develop an instrument to evaluate food and beverage service quality in spa hotels. Customer expectations are evaluated.

Seasonal fluctuations are studied in many papers (Parvathi et al., 2013). Elgin-Stuczynski (2014) surveys dairy farmers' lay knowledge of climate change and the adaptation strategies they have implemented to respond to climatic and economic drivers. All respondents make changes to their dairy business. Few farmers study seasonal fluctuations in temperatures. Most of seasonal fluctuations are caused by climate (Sada et al., 2014). This study is based on a household survey. Rainfall and temperature data are collected from Department of Hydrology and Meteorology within Kathmandu valley and analyzed to understand the climatic trend.

Food industry is widely discussed (Qiao et al., 2013). The quality of economic growth depends mainly on produced foods. Dairy products have an interesting marketing behavior (Ryzhyk, 2013). This study focuses on dairy products in Ukraine. Since we have only data for one Bulgarian company, this paper makes conclusions which are specific for one dairy products distributor. Agribusiness products are vital for humans. That is why some researchers investigate agribusiness producers (Knecht, Srodon, 2013). Holubek et al. (2013) presented an empirical investigation and an economic analysis of low-input farming systems.

Transactional data for sales is sometimes used for the supplier selection criteria (Kar and Pani, 2014). They want to find the critical factors for selecting suppliers. Fuzzy analytic hierarchy processes are used to analyze data from 188 firms. The adoption of e-transaction systems and e-procurement platforms are important evaluation criteria. Sales data usually omit out-of-stocks. A group of authors (Joachim et al., 2014) analyze the effects of retail out-of-stocks from a Service-Dominant (S-D) logic view. They analyze supply chains in retail market. The focus is on fast-moving consumer goods. They calculated the novel costs of an out-of-stock. They give new ideas for business alignment of manufacturers to retailers.

3. Dataset used for data mining

Transactional data are summarized in a table.

Table 1. *A part of the dataset used for data mining of sales data*

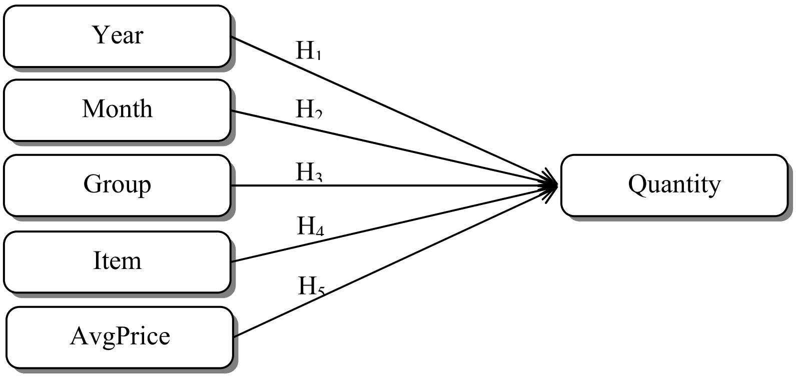
Year	Month	group	item	Quantity	Value	AvgPrice
2012	4	1	7	110	40.29	0.3663
2012	4	1	8	321	117.57	0.3663
2012	4	1	14	75	33.86	0.4515
2012	4	1	15	50	22.57	0.4514
2012	4	1	126	8	7.50	0.9375
2012	4	1	128	1	0.43	0.4300
2012	4	2	101	54	5.11	0.0946
2012	4	2	102	120	43.44	0.3620
2012	4	2	103	10	2.13	0.2130
2012	4	2	106	14	2.56	0.1829
2012	4	2	108	70	17.89	0.2556

Internal article number is used instead of the EAN. 559 records are used for our dataset. The initial transactional dataset consists of more than 40 000 records. By grouping data for year, month, group and item, the sum of quantity, the sum of value is calculated. The average price is calculated by dividing the last two columns. Because of grouping data, 561 records are used for our dataset. The dataset consists of data starting 2012, April and ending January 2014. Before transferring data from Excel to SPSS it is a good idea to change the decimal symbol from point to comma in Control Panel.

4. Defining hypothesis

We suppose that the quantity of sales is the dependent variable. All other variables are independent variables.

Figure 1. *Initial model for data mining transactional data*



Our prediction for the quantity of sales may be formulated in five hypotheses.

H1: The quantity of sales depends on the year.

H2: The quantity of sales depends on the month.

H3: The quantity of sales depends on the group.

H4: The quantity of sales depends on the item.

H5: The quantity of sales depends on the average price of sales.

Statistical methods have to be used for accepting or rejecting the five hypotheses. SPSS is used for analyzing the dataset and testing hypothesis (Gujarati, 2004). All the data in our dataset have numeric values. That is why the dataset is prepared to be transferred from Excel to SPSS. No other transformations or computations have to be made. In SPSS, 7 variables are defined according to each column of the table (containing the dataset). It is very important to give a correct measure for each variable. Variables “group” and “item” are on a nominal scale. Variables “year”, “month”, “quantity”, “value”, “AvgPrice” are on an interval scale.

5. Analyzing sales data

5.1. Testing H1 – does the quantity of sales depends on the year

By using descriptive statistics and Chi-square tests a check for dependency may be made (Analyze/Descriptive Statistics/Crosstab). Since both variables are on an interval scale, the Pearson coefficient may be used. The Chi-square test shows Pearson Chi-square value of 900.899 which is statistically not significant (Asymp. Sig. 2-tailed is 0.039). The comment below the Chi-square test table says that 99.5% of the cells have expected count less than 5. **Now we can reject H1. The quantity of sales does not depend on the year**.

To make another check for the dependency of quantity of sales on the year, the dataset may be filtered by choosing cases for 2012 and 2013 (Data/Select cases; if year<2014). 26 cases are not selected. The selected cases contain data for 2012 and 2013. The independent samples t-test may be applied to check the influence of the year on the quantity of sales. But before that another check has to be made.

Quantities must have a normal distribution. The one-sample Kolmogorov-Smirnov test is executed (Analyze/Nonparametric tests/1-Sample K-S). The one-sample Kolmogorov-Smirnov test shows the Kolmogorov-Smirnov Z value 8.059. It is statistically significant (Asymp. Sig. 2-tailed is 0.000). The values in the column “quantities” have a normal distribution.

Now we may continue with the independent samples t-test (Analyze/Compare Means/Independent-Samples T test). The grouping variable is “year” and the test variable is “quantity”. The confidence interval is set to the default value of 95%. The Levene’s test for equality of variances shows the value of F-test 27.063 with significance of 0.000 (0.000 < 0.05). It means that equal variances are not assumed. The t-test for equality of means shows the value of -3.268 with Sig. (2-tailed) 0.001 (0.001 < 0.05). The t-test shows that there are significant differences between the means of two groups (year 2013 and year 2014). **Now we can accept H1. The quantity of sales does depend on the year**.

A nonparametric test “Two-independent-samples test” may be made. The test type is Mann-Whitney U test. The value of Mann-Whitney U test is 28345. It is statistically significant (Asymp. Sig. 2-tailed is 0.009). The nonparametric test confirms our conclusion – **we may accept H1. The quantity of sales depends on the year**.

The one-way ANOVA test (Analyze/Compare Means/One-Way ANOVA) shows the F value of 8.196. The test has a significance of 0.004 (0.004 < 0.05). It shows that the quantity of sales depends on the year. **We may accept H1**.

Initially we rejected H1. But the three following tests showed that we may accept H1.

5.2. Testing H2 – does the quantity of sales depend on the month

Sales do have seasonal component. So we may expect the sales to depend on the month. Again descriptive statistics is applied. The Chi-square tests shows the value of the Pearson chi-square value of 4420.541. It is not statistically significant (Asymp. Sig. 2-sided is 0.122 > 0.05). The comment below the table says that 100% of the cells have expected count less than 5. **We may reject H2**. **The quantity of sales does not depend on the month**.

5.3. Testing H3 – does the quantity of sales depend on the group

The groups of dairy products are: packed liquid milk and yoghurt. The group of product may have a significant influence on the quantity of sales.

First, a Chi-square test is done. The Pearson Chi-square value is 430.217 with asymp. Sig. 2-sided 0.258 (0.258 > 0.05). The Pearson Chi-square value is not statistically significant. Moreover the comment below the table shows that 99.3% of the cells have expected count less than 5. **We reject H3**. **The quantity of sales does not depend on the group**.

Second, the independent samples t-test is done. Since there are two groups of stock keeping units (SKUs), its application is appropriate. The grouping variable is “group” and the test variable is “quantity”. The confidence interval is set to the default value of 95%. The Levene’s test for equality of variances shows the value of F-test 151.139 with significance of 0.000 (0.000 < 0.05). It means that equal variances are not assumed. The t-test for equality of means shows the value of 9.980 with Sig. (2-tailed) 0.000 (0.000 < 0.05). The t-test shows that there are significant differences between the means of two groups (group 1 “liquid milk” and group 2 “yoghurt”). **Now we can accept H3. The quantity of sales does depend on the group**.

Third, the nonparametric test “Two-independent-samples test” may be made. The test type is Mann-Whitney U test. The value of Mann-Whitney U test is 13593. It is statistically significant. Asymp. Sig. 2-tailed is 0.000 (0.000 < 0.05). The nonparametric test confirms our conclusion – **we may accept H3. The quantity of sales depends on the group**.

Fourth, the one-way ANOVA test shows the F value of 51.598. The test has a significance of 0.000 (0.000 < 0.05). It shows that the quantity of sales depends on the group. **We may accept H3**.

Initially we rejected H3. But the three following tests showed that we may accept H3.

5.4. Testing H4 – does the quantity of sales depend on the item

SKUs are very important for the sales of dairy products. Sometimes some SKUs are sold at minimum profit for the sake of making sales of other dairy products. It may be argued that the appropriate list of SKUs produced and sold may bring maximum profit and maximum customer satisfaction. Since items are nominally scaled only the Chi-square test may be done. The Pearson Chi-square value is 20244.922 with asymp. Sig. 2-sided 1.000 (1.000 > 0.05). The comment below the Chi-square tests says that 100% of the cells have expected count less than 5. So the Pearson Chi-square coefficient is not statistically significant. **We reject H4. The quantity of sales does not depend on the item (SKU)**.

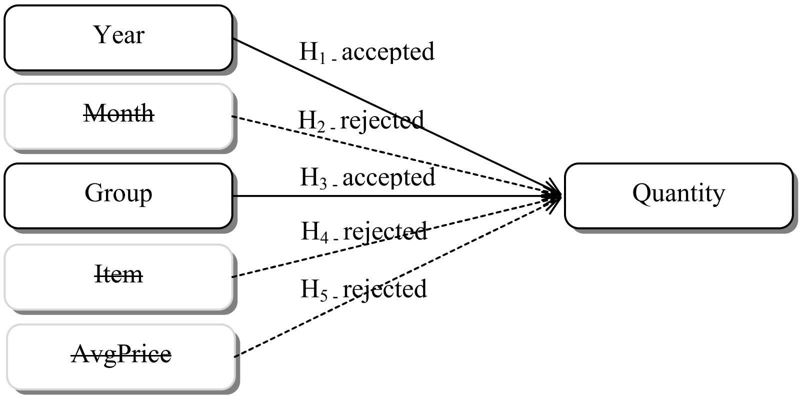
5.5. Testing H5 – does the quantity of sales depend on the average price

The average price is calculated by dividing the total sum of sales on the total quantity of sales. The microeconomic theory says that with the increase of the price quantities also increase. This postulate is in the basis of the well-known supply and demand curves model. It is an adequate reflection of the supply curve. Goods with low elasticity (such as dairy products) react differently. Sometimes an increase or decrease of the price does not affect sales. All these sentences are only predictions. They have to be checked by applying statistical methods. The average price is on an interval scale. So the appropriate statistical methods will be applied. The Pearson Chi-square is 163682.985. Asymp.sig. 2-sided is 0.000. The comment below the Chi-square tests table is that 100% of the cells have expected count less than 5. So the conclusion is the following. **We may reject H5. The quantity of sales does not depend on the average price**.

6. Final remarks. Discussion

The statistical tests on the dataset of sales show that hypotheses H1 and H3 are accepted and hypotheses H2, H4 and H5 are rejected. The quantity of sales does not depend on the month, item and average price. The quantity of sales depends mainly on the year and the group. The quantity of sales surely depends on other factors, not included in our model. The final view of our model is the following.

Figure 2. *Final model for factor estimation for sales of dairy products*



Our research continues with checking all available one-factor models (linear, logarithmic, inverse, quadratic, cubic, compound, power, growth, exponential and logical). The dependent variable is “quantity”. The independent variable is “year”. The constant is included in the equation (Analyze/regression/Curve estimation). The highest R-square value is 0.016. It is very low. ANOVA tables show that some models are adequate, others – not. So the conclusion is that there is not a functional dependency between the year and the quantity of sales. The dependency may not be expressed by a mathematical formula. The check for the one-factor dependency may continue by excluding the constant in equation. The power and the exponential function have the highest R-square value of 0.836. The two models are adequate (according to the ANOVA tables). The unstandardized coefficient B (in both models) is statistically significant.

The first equation (the power function) for predicating quantity of sales is the following:

$$\text{Predicted quantity} = \text{Year} \wedge 0.758 \qquad (1)$$

For year 2014, the predicted quantity of sales is 2014^0.758 = 319.50.

Since the quantity of sales varies from 1 to 40 000, the predicted value by the power function is nearer the lower bound.

The second equation (the exponential function) for predicating quantity of sales is the following:

Predicted quantity = exp(1) ^ (0.758 * Year) (2)

For year 2014, the predicted quantity of sales is 2.71828 ^ (0.758*2014) = 420.73.

Since the quantity of sales varies from 1 to 40 000, the predicted value by the exponential function is nearer the lower bound.

Different products have different diapason of sales (upper and lower interval for quantity of sales). These two predictions are quite rough and they are not relevant to a specific SKU. These two predictions have an error of 17.4%. It means that other factors affect the quantity of sales.

Since the group of items has an influence on the quantity of sales similar one-factor models may be constructed and tested. The independent variable is “group”. Models are checked with constant in equation and excluding constant in equation. ANOVA tables are generated. By including the constant in equation the highest value of R-square is 0.249 for the power, growth, logistic and exponential function. By excluding the constant in the equation the S function has the highest value of R-square 0.873. The ANOVA test shows that the model is adequate. The unstandardized coefficient B is 6.826. It is statistically significant. The comment below the coefficients table is that the dependent variable is ln(Quantity).

The equation is the following:

Predicted quantity for group = exp(1)^(6.826/group) (3)

The predicted quantity for group 1 is: 2.71828 ^ (6.826/1) = 921.50.

The predicted quantity for group 2 is: 2.71828 ^ (6.826/2) = 30.36.

So we may conclude that we are 87% sure that the average quantity of sales for SKUs in group 1 “liquid milk” is 922 and for group 2 “yoghurt” is 30. We are not absolutely sure because other factors are influencing the quantity of sales.

Conclusions

Sales of dairy products do not depend on the average price of products. With the increase and decrease of price, sold quantities do not change. Adding new items to the SKU list or removing some of them does not affect sales. The consumption of dairy products is not influenced by the month. We may expect comparatively constant sales with low levels of standard deviation. The quantity of sales is mainly affected by the year and the group of stocks. Future research may focus on time series analysis of sales using the same initial dataset consisting of transactional data for sales.

References

Azad, N., Karimi, O., Tabar, A. (2013). An empirical investigation on factors influencing sales force. *Management Science Letters*, 3(6), pp. 1671-1676

Celik, M., Kok, D. (2013). The Validity of Cost Stickiness in Turkey: A Panel Data Analysis in Istanbul Stock Exchange (ISE). *Business and Economics Research Journal*, 4(4), pp. 37-48

Ehrenthal, J.C., Gruen, T.W., Hofstetter, J.S. (2014). Value attenuation and retail out-of-stocks–a Service-Dominant logic perspective. *A Service-Dominant Logic View of Retail On-Shelf Availability*, p. 179

Elgin-Stuczynski, I.R., Batterbury, S. (2014). Perceptions of climate variability and dairy farmer adaptations in Corangamite Shire, Victoria, Australia. *International Journal of Climate Change Strategies and Management*, 6(1), p. 7

Giritlioglu, I., Jones, E., Avcikurt, C. (2014). Measuring food and beverage service quality in spa hotels: A case study in Balikesir, Turkey. *International Journal of Contemporary Hospitality Management*, 26(2), pp. 183-204

Gujarati, B. (2003). *Basic Econometrics,* 6th edition, New York: McGraw Hill Book Co.

Hanzaee, S., Eisapour, F., Azizi, B., Asgari, H., Bagheri, H. (2013). An empirical investigation on factors influencing on customer loyalty: A case study of Shahrvand food chain in Tehran. *Management Science Letters*, 3(6), pp. 1665-1670

Holubek, I. (2013). Production and Economic Analysis of Mountain Grasslands in Low-input Farming System. *Journal of Central European Agriculture*, 14(3), pp. 331-346

Huang, H.C. (2013). Construction of a Health Food Demand Prediction Model Using a Back Propagation Neural Network. *Advance Journal of Food Science and Technology*, 5(7), pp. 896-899

Kar, A.K., Pani, A.K. (2014). Exploring the importance of different supplier selection criteria. *Management Research Review*, 37(1), pp. 89-105

Knecht, D., Srodon, S. (2013). Analiza działalności grupy producentów trzody chlewnej na przykładzie Zrzeszenia Producentów Rolnych Gminy Biała. *Journal of Agribusiness and Rural Development*, (01 [27])

Parvathi, S., Avadi, C., Ramanarayanan, R., Srinivasaragavan, S. (2013). General Analysis of Two Products Inventory System with SCBZ Seasonal Production and Sales. *Applied Mathematical Sciences*, 7(72), pp. 3579-3590

Qiao, C., Xiao, R., Yan, L. (2013). Green Food Industry and Quality of Economic Growth in China: The Positive Analysis Based on Granger Causality Test and Variance Decomposition. *Advance Journal of Food Science & Technology*, 5(8), pp. 1059-1063

Ryzhyk, I.O. The prospects of marketing dairy cooperatives development in Ukraine. *Marketing ì Menedžment Innovacìj*, 4(3), pp. 264-271

Sada, R., Shrestha, A., Shukla, A.K., Melsen, L.A. (2014). People's Experience and Facts of Changing Climate: Impacts and Responses. *International Journal of Climate Change Strategies and Management*, 6(1), p. 5

Spadoni, R., Lombardi, P., Canavari, M., Hingley, M. (2013). Private food standard certification: analysis of the BRC standard in Italian agri-food. *British Food Journal*, 116(1), p. 10

Thomson, I.S.I. (2013). Chih-Min Ma, Cheng-Tao Yu and Bor-Wen Cheng. *Journal of Applied Sciences*, 13(8), pp. 1177-1184