Time Series Data Mining: A Retail Application Using SAS Enterprise Miner

The Honors Program
Senior Capstone Project
Student's Name: Daniel Hebert
Faculty Sponsor: Dr. Alan Olinsky
April 2013

Table of Contents

ACKNOWLEDGEMENTS

It is with utmost honor that I acknowledge Dr. Alan Olinsky and Dr. Billie Anderson for their exhaustive efforts throughout the completion of this project. Both professors contributed a substantial amount of time, knowledge, and resources to our research. Their passion and support throughout this project motivated me to achieve my goals and objectives. I also acknowledge Pat Casey and his administrative team for ensuring that I gain access to the technologies required to conduct my research. This group of individuals made certain that I was given access to SAS Enterprise Miner, the data mining software employed throughout this project. Dr. Segovis, the director of the Honors Program at Bryant University, was also helpful and supportive throughout my research. He provided vital motivation when overcoming challenges and obstacles within the project. Finally, I acknowledge my family and close friends, whose support and wisdom inspired me to remain diligent and steadfast in pursuing my research objectives.

ABSTRACT

Modern technologies have allowed for the amassment of data at a rate never encountered before. Organizations are now able to routinely collect and process massive volumes of data. A plethora of regularly collected information can be ordered using an appropriate time interval. The data would thus be developed into a time series. With such data, analytical techniques can be employed to collect information pertaining to historical trends and seasonality. Time series data mining methodology allows users to identify commonalities between sets of time-ordered data. This technique is supported by a variety of algorithms, notably dynamic time warping (DTW). This mathematical technique supports the identification of similarities between numerous time series. The following research aims to provide a practical application of this methodology using SAS Enterprise Miner, an industry-leading software platform for business analytics. Due to the prevalence of time series data in retail settings, a realistic product sales transaction data set was analyzed. This information was provided by dunnhumbyUSA. Interpretations were drawn from output that was generated using “TS nodes” in SAS Enterprise Miner.

INTRODUCTION

Data analysis is a commonly practiced methodology that is largely recognized as a means for gaining powerful insights and knowledge. Researchers and organizations often employ data analysis techniques as a paramount method for improving intelligence in a key area or topic of interest. The process of gleaning insights from records and observations within a data set is widely referred to as data mining. This analysis technique is often invaluable when attempting to interpret information and discern patterns or notable features within data. Organizations and businesses worldwide have adopted the data mining process as a paramount method for supporting informed decision making. By employing applicable analytical techniques, noteworthy insights can be gained. This knowledge can support a better understanding of certain areas of interest and support the development of educated decisions. Accordingly, in business situations, researchers and analysts are often selected according to their skills and talents in the science of data mining. This analytical methodology has become an integral practice for a variety of businesses and organizations.

The prominence and significance of data mining is becoming recognized by researchers and analysts worldwide. The expansion of this science is largely facilitated by substantial advancements in technology. Through an array of sophisticated hardware and software solutions, organizations can amass and interpret data at rapid rates. Such massive data sets can be effectively warehoused and processed using appropriate technologies. In retail and commerce settings, for instance, vast quantities of data are collected during purchase transactions. Companies are able to amass data regarding their customers and product purchase rates. This information can be effectively used to gain a deeper understanding of key markets of interest and notable buying behaviors. Such findings are often uncovered through sophisticated data mining techniques. Valuable information can now be successfully gathered from massive volumes of data. Many organizations worldwide are realizing the value of data mining methodologies. With a plethora of data collected and warehoused, researchers and businesses are searching for novel and effective means of analysis. As the science of data mining continuously grows, advanced technologies and solutions will undoubtedly allow organizations to identify new analytical techniques for gaining valuable insights.

Big Data

With the amassment of large volumes of data and numerous interpretive methods, the notion of "big data" is realized by many organizations and analysts worldwide. According to SAS Institute (2013), a leading provider of software solutions for business intelligence, big data involves the "exponential growth, availability and use of information, both structured and unstructured." This movement has had paramount effects on the analytical capacities of businesses and organizations. Big data involves large volumes of observations and variables, numerous data formats and structures, and rapid analyses of large data sets. The notion of big data is discussed extensively among research communities. In the past several years, this movement has had notable effects on the data mining processes of companies and organizations. The increasing prominence of big data has led to the development of new competitive arenas for businesses. Organizations are competitively seeking new technologies and resources to address the growing importance of big data. Such entities are seeking novel means of gaining insight from massive and previously unmanageable data sets.

The big data movement is defined by several unique characteristics. SAS Institute (2013) identifies three qualities of big data—volume, variety, and velocity. Organizations are amassing data sets that contain numerous observations and an array of variables. Big data possesses volume. Data sets often include a steadily increasing number of records. Variety is also a paramount quality of big data. There are numerous formats and degrees of cleanliness that must be considered by analysts. Accordingly, data must often be prepared and processed prior to the employment of appropriate data mining techniques and software solutions. Big data is also characterized by the velocity in which data and observations are collected and processed. Data sets are warehoused, analyzed, and interpreted at rapid rates. These key qualities of the big data movement are being realized by numerous worldwide organizations and researchers. Sophisticated software techniques are employed to understand and interpret massive volumes of data. Businesses and companies seek new and effective methodologies to gain insights that support competitive efforts. Ultimately, big data has imposed an array of challenges and opportunities when attempting to glean knowledge and information from vast warehouses of data. Organizations can now access large arrays of data that are collected in

numerous formats. The prevalence of big data has undoubtedly provoked a plethora of advancements and developments in the field of data mining.

TIME SERIES DATA MINING

Through the collection of information on a routine and daily basis, organizations are amassing sequentially ordered data. Records and observations in such data sets possess a time factor. Accordingly, information is collected over a period of time. The data is sequentially ordered, and observations within the data set often possess a time variable. This factor indicates the date and time associated with each record. In retail environments, for instance, information regarding sales transactions is often recorded routinely. Such purchase transactions exist as observations within a data set. These records maintain a data and time associated with each transaction. Such time-ordered data presents new challenges for businesses and companies. Sequential data that is collected over a period of time is referred to as time series. When faced with data sets that contain time series, organizations must employ new and sophisticated techniques for analysis. Traditional data mining methods and statistical techniques are often inappropriate when analyzing data that possess a time factor. This dilemma has led to the development and rising importance of time series data mining techniques. In order to address the increasing prevalence of time-ordered data, researchers and analysts have begun seeking new analysis methods that account for information collected over a period of time. In order to support informed decision making and gain useful knowledge, organizations are searching for novel means of understanding and interpreting time series.

Time series data mining has become increasingly important due to the prominence of sequentially ordered data. In their publication, *Time Series Data Mining with SAS Enterprise Miner*, Schubert and Lee (2011) provide background on this field of research. The authors offer information regarding the benefits that can be realized through time series data mining analysis. Many organizations are beginning to experience situations in which this form of analysis is particularly relevant and applicable. Schubert and Lee (2011) specifically provide background on the applications of cluster analysis in time series data mining. This technique produces output that identifies groupings, or clusters, of time series that share related trends

and similar patterns. Accordingly, different time series are effectively compared and grouped according to similarity. The authors explain that time series data mining can be used to detect "similar patterns in historical time series data" (Schubert & Lee, 2011). Their publication provides insight into the overall function and purpose of comparing time series. Through the use of time series data mining techniques, business and organizations can gain insight into time series that may be related within a data set. Patterns and similarities within a sequence of data can be effectively discerned with the employment of such methodologies.

Time series data mining techniques are commonly applied in retail environments. This methodology allows businesses to identify groups of observations and records that share patterns and common seasonality. Kumar, Patel, and Woo (2002) specifically reference time series data mining in the context of the retail industry. Retailers are recognized as primary collectors of time series data. This is largely due to the exorbitant volume of transactional information that is recorded on a routine basis. Data related to purchase dates, items sold, and customer identification numbers are collected daily in many retail settings. With the routine amassment of such time-ordered data, retailers collect a multitude of time series. For instance, transactions of certain products over a period of time may serve as a set of time series. Kumar, Patel, and Woo (2002) identify time series data mining as a means of discerning patterns and similarities in such data. With such methodology, retailers can identify groupings of time series that are related. For instance, transactional data can be used to distinguish products that share similar purchase rates. This application of time series data mining in retail settings can support business decision making. Kumar, Patel, and Woo (2002) explain that promotional campaigns and advertisements can be created to promote certain groupings of items during specific seasons. Nakkeeran, Garla, and Chakraborty (2012) provide further insight into the rising importance of time series data mining in retail settings in their publication regarding time series comparison for retailers. It is found that such organizations are seeking data mining techniques to draw comparisons within a data set and discern groupings of time series to support marketing efforts.

Retail companies and other businesses often employ the use of sophisticated software to analyze data and produce understandable output. In identifying similarities and patterns

among time series, data mining software programs are created using a specific algorithm—dynamic time warping (DTW). In their publication, *Searching and Mining Trillions of Time Series Subsequences under Dynamic Time Warping*, Rakthenmanon, Campana, Mueen, et al. (2012) identify this mathematical formula as the key algorithm in the effective comparison of time series. Many prominent software packages rely on DTW to support pattern recognition and the identification of similar trends between sequenced and time-ordered observations. Prior to the development of the DTW algorithm, most traditional data mining formulas discerned similarities between data points without accounting for a time factor. Accordingly, observations were often compared under static conditions. Under traditional methodologies, the sequential ordering of data over time was not considered. DTW, however, identifies similar trends that may occur over time across multiple arrays of sequenced data. This mathematical formula serves as an effective data mining technique when algorithmically comparing sets of time-ordered data. Dynamic time warping (DTW) offers a means of identifying similar trends across sets of sequenced records and observations.

Different time series may possess common trends that do not necessarily occur simultaneously. For instance, sales transactions of two different products may share similar trends in seasonality over time. However, traditional statistical methodology may not discern such a relationship, as it does not consider time as a factor in the comparison. See Figure 1A (Rakthenmanon, Campana, Mueen, et al., 2012). The similarity between Product A (red) and Product B (blue) is computed using a traditional distance measure. While both products share similar transaction histories over time, a similarity between both time series is undiscerned. The relationship between transaction records of both products is statically compared without accounting for the full time-ordered sequence of data. Accordingly, similarity between both time series would not be discerned if trends in the data occurred at slightly different times. DTW accounts for this time factor. See Figure 1B (Rakthenmanon,

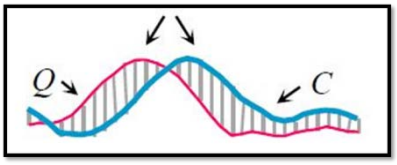


Figure 1A – Similarity between Product A (red) and Product B (blue) time series is undiscerned.

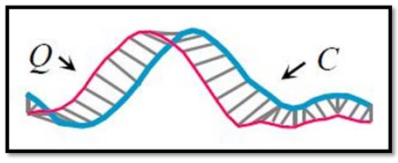


Figure 1B – DTW discerns a similar trend between Product A (red) and Product B (blue) time series.

Campana, Mueen, et al., 2012). The algorithm accounts for the full sequence of data over time. DTW employs a distance measure between the transactional records of Product A and Product B, while considering the factor of time. Both products share similar transactional patterns. There appears to be a similar peak in purchase transactions of each product. Through the use of a dynamic time warping (DTW) algorithm, a relationship can be discerned despite the asynchronous nature of the purchasing trends. Such a mathematical technique is paramount to time series data mining and allows for the effective comparison of sequenced and time-ordered data.

As businesses and organizations continue to routinely amass data, the science of time series data mining will undoubtedly gain prevalence. This form of analysis is particularly applicable in retail environments. Businesses are seeking software solutions that can effectively interpret and discern patterns among time series. Software companies are beginning to develop new analysis tools that allow for the identification of groups of time series that share similar features and trends. Such novel techniques are developed on the fundamental principles of the dynamic time warping (DTW) algorithm. Existing literature related to time series data mining provides highly theoretical and mathematical information. Accordingly, this field of research lacks a practical explanation and description of time series analysis. There appears to be an absence of literature that provides analysts with a discussion and application of new time series software tools. Such research could provide a high-level investigation of the process of performing a time series data mining analysis. Existing literature is highly complex and often lacks a description of the process in effectively comparing time series and recognizing patterns within a sequential data set. This investigation could prove particularly useful for organizations, such as retailers, that aim to understand the overall analysis and interpretation that is involved in time series data mining.

RETAIL APPLICATION

With the growing prominence of time series data mining, new analytical tools are being developed to discern similarities and notable trends within sequenced and time-ordered data. Due to its prevalence in commerce settings, an investigation of time series data mining analysis is conducted in the context of retail product transaction data. These realistic time series are investigated using tools in a widely recognized business intelligence software package. This time series data mining analysis fundamentally employs a dynamic time warping (DTW) algorithm to produce output. The resulting information is interpreted and explored. Through this practical research, the impact of time series data mining on retail marketing analytics is examined. Specifically, through a retail application, the research offers an exploration of how time series can be effectively compared in transactional data sets. Using a leading software package for data mining and business intelligence, a realistic set of time series are compared using sophisticated analysis tools. Groupings of products are identified that share similar purchase histories over a period of time. Ultimately, such output can be interpreted and employed to support marketing materials and effective promotional campaigns.

SAS Enterprise Miner

In the analysis of massive data sets, retailers and other businesses often seek software resources. Such technology provides a platform for analyzing data and producing useful output for interpretation. SAS Enterprise Miner, a product of SAS Institute, is a leading data mining software package and business intelligence solution. The software provides a powerful, industry-grade platform for data analytics. SAS Institute software solutions are used by most Fortune 500 companies, including retailers. The organization is recognized globally as a "pioneer of sophisticated data-analysis tools" (Lohr, 2012). SAS Institute's prominent data mining software, SAS Enterprise Miner, possesses numerous programs and mathematical functions that can aid in the analysis and interpretation of massive volumes of data. Accordingly, this software package served as the primary platform for conducting time series data mining analysis. It serves as a powerful tool in the effective comparison of sequenced

data. Due to its popularity among large retailers and other businesses, SAS Enterprise Miner serves as a practical resource for applying time series methodology.

Retail Data Set

Through a rich and realistic data set, an investigation of time series data mining is conducted. Using SAS Enterprise Miner, time series are grouped and compared in order to discern similar patterns and relationships within the data. Through discussions with company representatives, dunnhumbyUSA provided retail data for research purposes (Perry, 2013). This organization is a joint venture of The Kroger Company and the London-based dunnhumby. The company supports retail clients with business analytics and data-driven marketing insights. Since 2003, dunnhumbyUSA has conducted data analysis of over 400 million retail customers in 28 countries (dunnhumby, 2013). Through its active and successful role in data analysis, this firm served as a reputable provider of previously aggregated data for research purposes. Specifically, dunnhumbyUSA offered retail data that contains time series. Appropriate methodologies and software solutions can thus be applied in order to produce fruitful time series data mining output.

The data set retrieved from dunnhumbyUSA for research purposes is titled “The Complete Journey”. The data contained observations and records that were amassed by an undisclosed retail store. The data set contains the purchase histories of approximately 2,500 households over a two year time period. Information related to product categories and purchase dates were recorded. Accordingly, product transactions made by each of 2,500 households were recorded over two years. Due to the inherent factor of time, this data serve as time series. Transactions are recorded for each product category. Such records are order sequentially according to dates and times of purchase. The time series associated with each product category serves as the fundamental data that is analyzed using SAS Enterprise Miner. Using new time series techniques, similarities in purchasing trends are identified across time series. Accordingly, the analysis produces output that groups product categories according to their similar purchase transaction histories over a two year time period.

Due to the breadth and complexity of the data set, the data was preprocessed, cleaned, and prepared for a time series analysis in SAS Enterprise Miner. The data set contains product transaction history for 43 distinct departments within a retail store. Upon further analysis, the "grocery" department appeared to possess the largest number of product transactions over a two year time period. There were 1,646,076 purchases made in this area of the retail store. Due to this sufficiently large number of observations, the data was subset to only include transactions for product categories in the grocery department. A portion of the resulting data set is provided in Figure 1 of Appendix A. In addition to this isolation of certain records, data merging was a necessary step in preparing for an effective time series analysis. "The Complete Journey" contained several related data sets. In conducting a fruitful time series analysis, certain data sets with vital information were chosen. The data set, product_id, possessed information regarding the product categories in the grocery department and the various identifiers, or SKU numbers, associated with these items. Another data set, transaction_data, contained information regarding purchase dates, quantities sold, and household identification. This data set also assigned SKU numbers to each product category. Using the common SKU identification variable, transaction_data and product_id data sets were merged to create a single set of time series for analysis.

Data Analysis

Without the employment of appropriate time series data mining methodology, it is unlikely that certain relationships and similarities can be discerned within the data set. Due to the expansive nature of "The Complete Journey" data set, it is extremely difficult to compare time series without appropriate analytical tools. The transactions associated with 94 product categories were collected over a two year time period and developed into time series. Please refer to Figure 2 of Appendix A. Each line represents the sales transactions of a certain product category throughout a two year period of time. This representation illustrates the complexity involved when attempting to compare time series. In order to effectively discern grocery products with similar purchase histories, necessary software tools are paramount. Accordingly, the new time series features in SAS Enterprise Miner were employed in order to effectively analyze the data set.

Analytical processes are conducted in SAS Enterprise Miner through the use of "nodes". Such blocks represent steps in the analysis. Nodes are developed with sophisticated algorithms and mathematical formulas. SAS Institute recently developed a set of new nodes for time series data mining purposes. These novel tools employ a dynamic time warping algorithm to effectively compare time series within a data set. Below, Figure 2 identifies the primary nodes that were used when analyzing "The Complete Journey" data set.

Figure 2 – New time series nodes provided in SAS Enterprise Miner.

The first node represents the transactional data of 94 different grocery products in the retail store. A TS Data Preparation node ensured that the data was processed prior to the generation of output. The data was transposed according to the purchase dates indicated in each observation. This ensures that the data is properly sequenced according to a variable that indicates time. The resulting partial table of observations may be seen in Figure 3 of Appendix A. Following this preparation, the new TS Similarity node in SAS Enterprise Miner was used to generate output for the data set. This analytical step was developed using a dynamic time warping (DTW) algorithm. Accordingly, time series are grouped, or clustered. Each grouping indicated in the resulting output contains time series that are related and share similar trends over time. The algorithm accounts for the slight differences in timing among purchasing trends and seasonal patterns. With the addition of a TS Similarity node, the analysis was ready to be performed. All necessary preprocessing and setting selections were made.

Output and Key Plots

The new time series analysis tools in SAS Enterprise Miner produced an array of output in the form of plots and graphs. Such information offered a comparison of time series within the data set. In this retail application, grocery product categories were identified that share similar transaction histories over a period of two years. Specifically the output generated by the TS

Similarity node was created with a dynamic time warping (DTW) algorithm. Accordingly, when identifying similarities between combinations of time series, the algorithm accounted for the factor of time. Transactions were collected over a two year time period. Trends and seasonality that occur over time must be considered when discerning similarities between grocery products in "The Complete Journey". The TS Similarity node provides output that considers the sequential ordering of data points over time. While trends in sales transactions may not occur in perfect unison, SAS Enterprise Miner is able to discern similar qualities among time series.

Through the aforementioned time series data mining analysis, various plots were generated that are particularly applicable for retail marketing research. These graphs provide a visual representation of groups of time series that share similar features and historical patterns over a defined time period. Such plots include cluster dendrograms and cluster constellation plots. These graphs are may be intuitively understood by researchers and analysts. They prove particularly effective in clearly identifying clusters of related products in retail settings. Accordingly, marketing analysts can produce this graphical output to intuitively understand patterns among product categories. Ultimately, transactional trends can be compared while taking into account a period of time in which data was collected. Through novel data mining techniques in SAS Enterprise Miner, cluster dendrograms and constellation plots can be generated to effectively compare time series. Such output can support marketing initiatives and business intelligence among retail users.

A cluster dendrogram provides a statistically-oriented visualization of similar time series. Please refer to Figure 3A below. The graph employs the use of a semi-partial R-squared. This is represented on the horizontal axis of the plot. This serves as a measure of similarity between time series in the data set. When applied to "The Complete Journey" data set, this plot identifies the similarity between product categories according to their transactional histories over two years. The vertical axis of the plot contains Sales_Value listings. Each value represents grocery product categories. The bars seen in the dendrogram are referred to as clades (Drout & Smith, 2012). The arrangement of clades identifies combinations of time series that share similar purchasing trends. The length of each bar indicates the degree of

similarity between grocery products. As clades grow to incorporate large groupings of related products, the degree of similarity within the cluster diminishes. Please refer to Figure 3B below. Clades grow longer as they incorporate more time series and smaller groupings of related product categories, or Sales_Value points. This plot provides an effective means of understanding the similarity between different combinations and groupings of time series. Retail marketers can use this intuitive plot to gain insight into the degree of similarity between product category transactions.

Figure 3A – Cluster dendrogram indicates the degree of similarity between combinations and groups of product transaction time series.

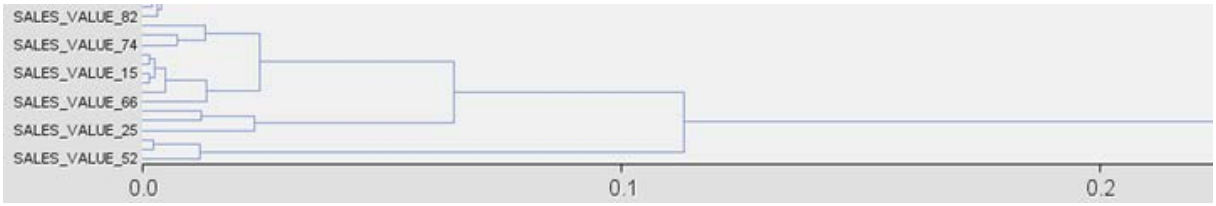


Figure 3B – Clades grow longer as they incorporate more product time series and smaller clusters of related items.

A cluster constellation plot provides marketing analysts with a more visually intuitive illustration of similar time series. Please refer to Figure 4A. The plot contains an array of points that are visibly arranged in connected clusters and groupings. In regards to "The Complete Journey", such groups clearly indicate products that share similar transaction

histories over a two year time period. Distinct points within the cluster constellation plot are visible after enlarging the graph. Please see Figure 4B and 4C. Two types of points are present in the plot—CL and Sales_Value. CL points represent cluster identifications. Accordingly, subsequent connected points are assumed to be grouped and related in some way. Sales_Value points represent time series associated with distinct grocery products. When connected to CL points, or cluster identifiers, such product time series maintain comparable trends and similar patterns over time.

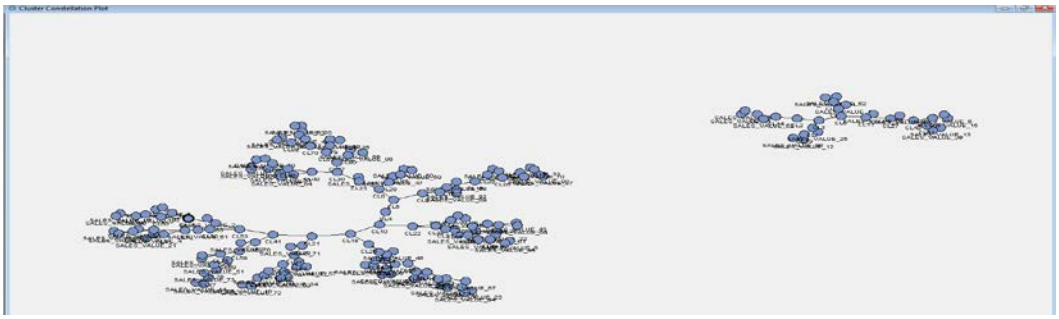


Figure 4A – Cluster constellation plot clearly identifies groupings of related product transaction time series.

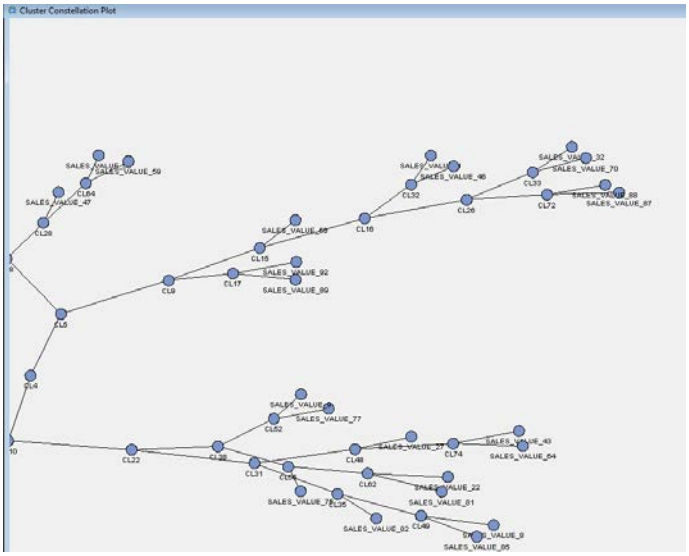


Figure 4B – CL points, or cluster identifiers, groups of subsequent time series that share similarities. Sales_Value points represent distinct time series for grocery product categories.

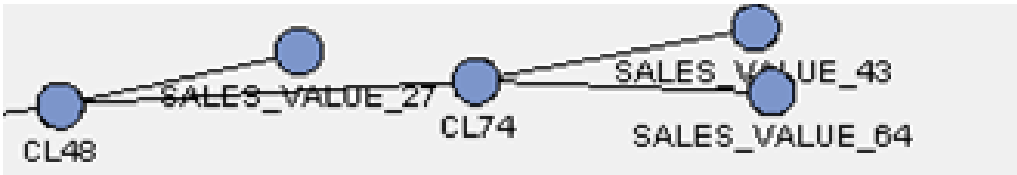


Figure 4C – CL and Sales_Value points are clearly seen.

INTERPRETATION OF OUTPUT

Due to its clarity and simplicity, the cluster constellation plot was used as the primary means of interpreting "The Complete Journey" data set. This graph provides an apparent identification of products that share similar purchasing trends over a two year period. Subsequent time series that are connected to CL points, or cluster identifiers, are presumed to share similarities. Sales_Value points represent the distinct time series for each grocery product category. In order to effectively interpret this information, Sales_Value points must be defined. Please see Figure 5 below. A graph is provided that identifies the product category associated with each Sales_Value. This allows for an understanding of the specific items that share related purchase transactions over a two year time period. Retail marketers can then use this information to craft appropriate promotions for product groups and design tailored marketing materials for certain items.

TSID Map Table

NAMEID	TSID	Category of product
SALES_VALUE_1	1	AIR CARE
SALES_VALUE_2	2	BAG SNACKS
SALES_VALUE_3	3	BAKED BREAD/BUNS/ROLLS
SALES_VALUE_4	4	BAKED SWEET GOODS
SALES_VALUE_5	5	BAKING MIXES
SALES_VALUE_6	6	BAKING NEEDS
SALES_VALUE_7	7	BATH TISSUES
SALES_VALUE_8	8	BEANS - CANNED GLASS & ...
SALES_VALUE_9	9	BEERS/ALES
SALES_VALUE_10	10	BIRD SEED
SALES_VALUE_11	11	BLEACH
SALES_VALUE_12	12	BOTTLE DEPOSITS
SALES_VALUE_13	13	BUTTER
SALES_VALUE_14	14	CANNED JUICES
SALES_VALUE_15	15	CANNED MILK
SALES_VALUE_16	16	CAT FOOD
SALES_VALUE_17	17	CAT LITTER
SALES_VALUE_18	18	CHEESE
SALES_VALUE_19	19	COCOA MIXES
SALES_VALUE_20	20	COFFEE

Figure 5 - Sales_Value points are defined.

With this descriptive information, grouped points in the cluster constellation plot may now be defined and interpreted. An overall indication of patterns between product categories is illustrated. A plethora of relationships are identified between grocery items. For instance, Sales_Value_9 and Sales_Value_77 are connected to CL52. This indicates that both time series share similar transaction trends over a two year time period. Please refer to Figure 6A. Through an identification of these Sales_Value points, it appears that beer and Crystal Light share similar purchase histories within the retail store. Both product categories are beverages. Accordingly, perhaps these items are related due to seasonality. They are presumably purchased more frequently during warmer seasons. Customers would likely by higher quantities of such beverages during when temperatures are warm. Conversely, beer and Crystal Light may be purchased less frequently during the winter season, as refreshments are not less demanded.

Figure 6A – The similarity between time series of beer and Crystal Light is identified.

In further interpreting the cluster constellation plot generated by SAS Enterprise Miner, there appears to be a relationship between the transaction histories of meat and frozen pizza. This is seen apparent in CL74. Sales_Value_43 and Sales_Value_64 are connected to this cluster identifier. Please see Figure 6B. SAS Enterprise Miner discerned similar trends between the time series of both product categories. When isolating both product transaction time series, it appears that the cluster plot accurately identified this relationship. Please refer to Figure 6C below. The time series appear to fluctuate in relative unison. Certain trends appear to be common between both meat and frozen pizza. Perhaps customers purchase these items as high sources of protein. Other shoppers may seek to purchase meat and frozen pizza, as they are both appropriate items for storage in freezers.

Figure 6B - The similarity between time series of meat and frozen pizza is identified.

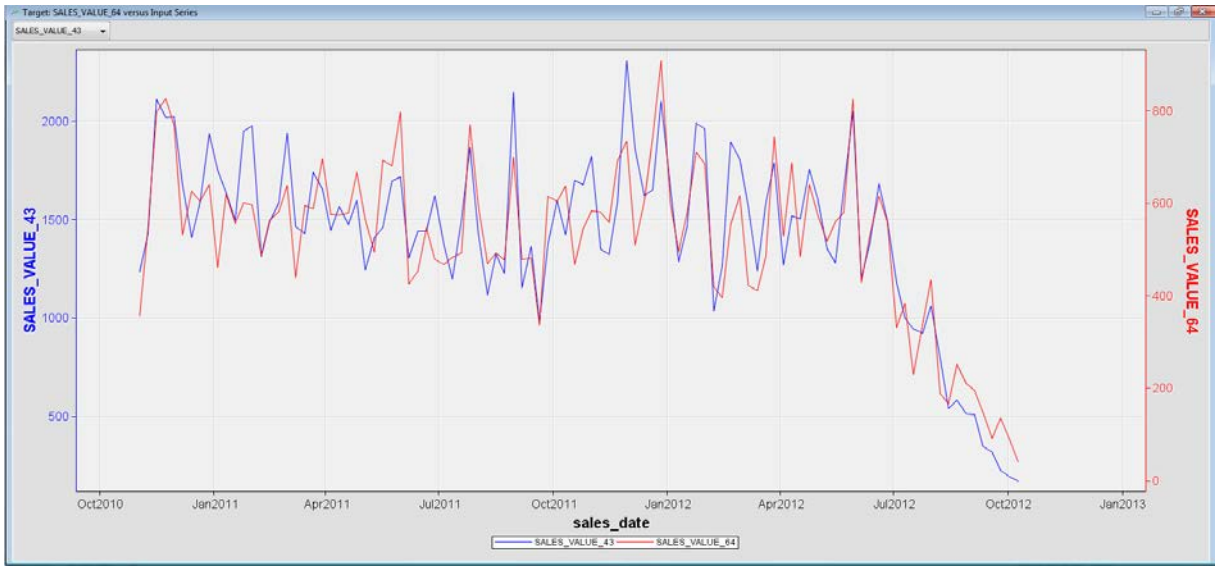


Figure 6C – Time series of meat and frozen pizza transactions are very similar.

Through new analysis tools in SAS Enterprise Miner, output was generated that effectively compared time series and discerned similarities in "The Complete Journey" data set. While relationships were clearly identified among a variety of product categories, it must be noted that the analysis does not provide insight into the causation of such patterns and commonalities within the data. According marketing analysts and researchers are tasked with developing an appropriate assessment for the reasoning behind certain findings. Assumptions must often be made regarding the causation of relationships between time series.

KEY IMPLICATIONS AND FUTURE RESEARCH

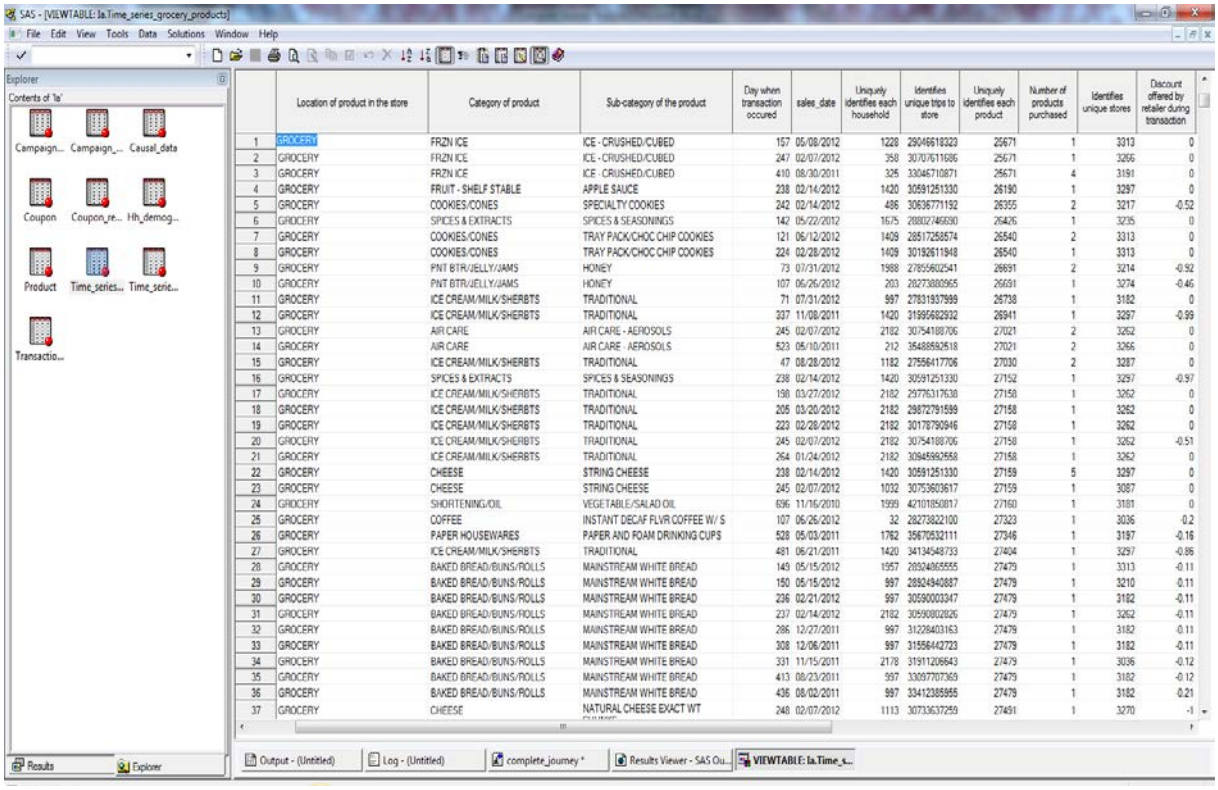
Retailers collect and amass large data sets on a routine and daily basis. Due to this process of warehousing sequential data, such businesses are able to effectively create time series. SAS Enterprise Miner proves to offer novel means of effective comparing this form of data. Retailers can largely benefit from the time series data mining tools provided in this software platform. Through an identification of products that share similar purchase histories and similarities regarding transactions, marketers in retail environments can make informed decisions regarding promotional campaigns and advertisement materials. For instance, after determining a group of products that share similar purchasing trends, perhaps retailers arrange store layouts to place these items together. Marketers can also design campaigns that promote

a grouping of products simultaneously. For instance, in regards to an aforementioned product relationship, perhaps promotional discounts can be offered when meat and frozen pizza are purchased at once.

Future research can be conducted to identify appropriate responses to a time series data mining analysis. Perhaps researchers can isolate forms of promotion that are particularly applicable after conducting a comparison of time series within a retail data set. This literature would provide marketers and businesses with insight into how this data mining technique can lead to the creation of distinct types of promotional campaigns and marketing materials. Due to the global popularity of SAS Enterprise Miner, many organizations possess the analytical tools to effective compare time series. Perhaps future research could identify marketing practices that are most applicable after identifying groups of related product time series. Such literature would be widely accepted by many analysts and researchers as time series data mining becomes increasingly recognized and applied in retail settings.

APPENDIX A

Figure 1



	Location of product in the store	Category of product	Sub-category of the product	Day when transaction occured	sales_date	Uniquely identifies each household	Identifies unique trips to store	Uniquely identifies each product	Number of products purchased	Identifies unique stores	Discount offered by retailer during transaction
1	GROCERY	FRZN ICE	ICE - CRUSHED/CUBED	157	05/08/2012	1228	29046618323	25671	1	3313	0
2	GROCERY	FRZN ICE	ICE - CRUSHED/CUBED	247	02/07/2012	358	30707611686	25671	1	3266	0
3	GROCERY	FRZN ICE	ICE - CRUSHED/CUBED	410	08/30/2011	326	33046710871	25671	4	3191	0
4	GROCERY	FRUIT - SHELF STABLE	APPLE SAUCE	238	02/14/2012	1420	30591251330	26190	1	3297	0
5	GROCERY	COOKIES/CONES	SPECIALTY COOKIES	242	02/14/2012	486	30636771192	26355	2	3217	-0.52
6	GROCERY	SPICES & EXTRACTS	SPICES & SEASONINGS	142	05/22/2012	1675	28802746690	26426	1	3235	0
7	GROCERY	COOKIES/CONES	TRAY PACK/CHOC CHIP COOKIES	121	06/12/2012	1409	28517258574	26540	2	3313	0
8	GROCERY	COOKIES/CONES	TRAY PACK/CHOC CHIP COOKIES	224	02/28/2012	1409	30192611948	26540	1	3313	0
9	GROCERY	PNT BTR/JELLY/JAMS	HONEY	73	07/31/2012	1988	27855602541	26691	2	3214	-0.92
10	GROCERY	PNT BTR/JELLY/JAMS	HONEY	107	06/26/2012	203	28273880965	26691	1	3274	-0.46
11	GROCERY	ICE CREAM/MILK/SHERBTS	TRADITIONAL	71	07/31/2012	997	27831937999	26738	1	3182	0
12	GROCERY	ICE CREAM/MILK/SHERBTS	TRADITIONAL	337	11/08/2011	1420	31995682932	26941	1	3297	-0.99
13	GROCERY	AIR CARE	AIR CARE - AEROSOLS	245	02/07/2012	2182	30754188706	27021	2	3262	0
14	GROCERY	AIR CARE	AIR CARE - AEROSOLS	523	05/10/2011	212	35488592618	27021	2	3266	0
15	GROCERY	ICE CREAM/MILK/SHERBTS	TRADITIONAL	47	08/28/2012	1182	27556417706	27030	2	3287	0
16	GROCERY	SPICES & EXTRACTS	SPICES & SEASONINGS	238	02/14/2012	1420	30591251330	27152	1	3297	-0.97
17	GROCERY	ICE CREAM/MILK/SHERBTS	TRADITIONAL	198	03/27/2012	2182	29776317638	27158	1	3262	0
18	GROCERY	ICE CREAM/MILK/SHERBTS	TRADITIONAL	205	03/20/2012	2182	29872791599	27158	1	3262	0
19	GROCERY	ICE CREAM/MILK/SHERBTS	TRADITIONAL	223	02/28/2012	2182	30178790946	27158	1	3262	0
20	GROCERY	ICE CREAM/MILK/SHERBTS	TRADITIONAL	245	02/07/2012	2182	30754188706	27158	1	3262	-0.51
21	GROCERY	ICE CREAM/MILK/SHERBTS	TRADITIONAL	264	01/24/2012	2182	30945992658	27158	1	3262	0
22	GROCERY	CHEESE	STRING CHEESE	238	02/14/2012	1420	30591251330	27159	5	3297	0
23	GROCERY	CHEESE	STRING CHEESE	245	02/07/2012	1032	30753603617	27159	1	3087	0
24	GROCERY	SHORTENING/OIL	VEGETABLE/SALAD OIL	696	11/16/2010	1999	42101850817	27160	1	3181	0
25	GROCERY	COFFEE	INSTANT DECAF FLVR COFFEE W/ S	107	06/26/2012	32	28273822100	27323	1	3036	0.2
26	GROCERY	PAPER HOUSEWARES	PAPER AND FOAM DRINKING CUPS	528	05/03/2011	1762	35670532111	27346	1	3197	-0.16
27	GROCERY	ICE CREAM/MILK/SHERBTS	TRADITIONAL	491	06/21/2011	1420	34134548733	27404	1	3297	-0.86
28	GROCERY	BAKED BREAD/BUNS/ROLLS	MAINSTREAM WHITE BREAD	149	05/15/2012	1957	28924065555	27479	1	3313	-0.11
29	GROCERY	BAKED BREAD/BUNS/ROLLS	MAINSTREAM WHITE BREAD	150	05/15/2012	997	28824940887	27479	1	3210	-0.11
30	GROCERY	BAKED BREAD/BUNS/ROLLS	MAINSTREAM WHITE BREAD	236	02/21/2012	997	30590003347	27479	1	3182	-0.11
31	GROCERY	BAKED BREAD/BUNS/ROLLS	MAINSTREAM WHITE BREAD	237	02/14/2012	2182	30590802826	27479	1	3262	-0.11
32	GROCERY	BAKED BREAD/BUNS/ROLLS	MAINSTREAM WHITE BREAD	286	12/27/2011	997	31228403163	27479	1	3182	-0.11
33	GROCERY	BAKED BREAD/BUNS/ROLLS	MAINSTREAM WHITE BREAD	308	12/06/2011	997	31556442723	27479	1	3182	-0.11
34	GROCERY	BAKED BREAD/BUNS/ROLLS	MAINSTREAM WHITE BREAD	331	11/15/2011	2178	31911206643	27479	1	3036	-0.12
35	GROCERY	BAKED BREAD/BUNS/ROLLS	MAINSTREAM WHITE BREAD	413	08/23/2011	997	33097707369	27479	1	3182	-0.12
36	GROCERY	BAKED BREAD/BUNS/ROLLS	MAINSTREAM WHITE BREAD	436	08/02/2011	997	33412365955	27479	1	3182	-0.21
37	GROCERY	CHEESE	NATURAL CHEESE EXACT WT	248	02/07/2012	1113	30713637259	27491	1	3270	-1

Figure 2

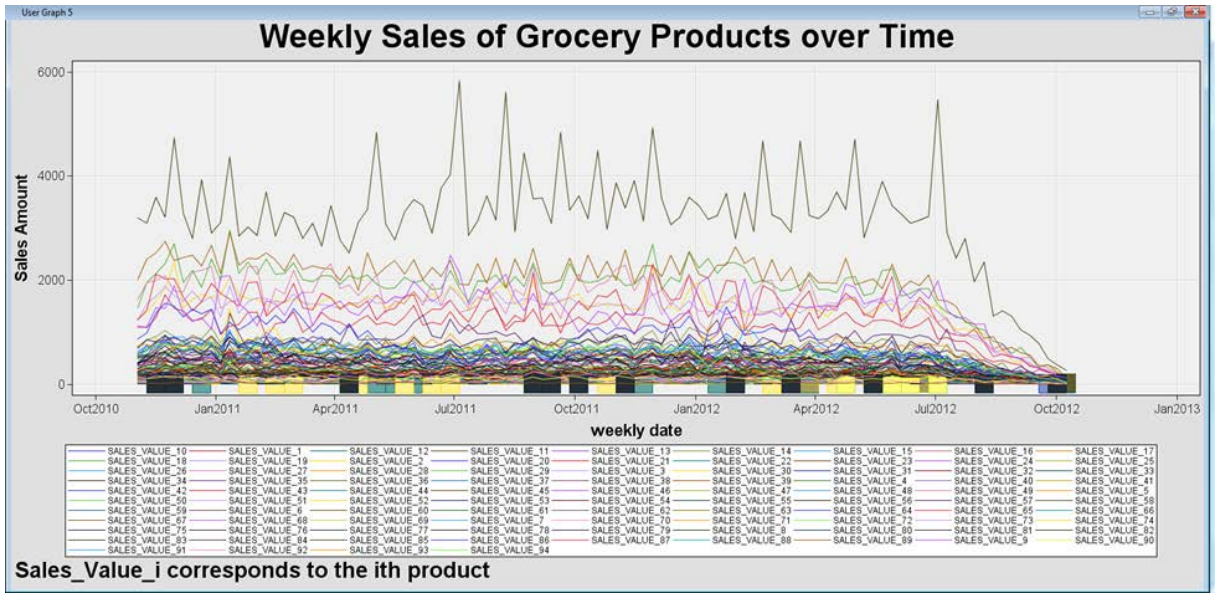


Figure 3

EMWS1.TSDP3_TRAIN

Obs #	sales_date	SALES_VALUE_1	SALES_VALUE_2	SALES_VALUE_3	SALES_VALUE_5	SALES_...	SALES_...
1	11/02/2010	148.16	1268.41	1257.03	206.42	530.43	218.61
2	11/09/2010	162.34	1380.37	1549.94	281.48	539.43	281.8
3	11/16/2010	167.76	1601.85	1689.5	402.01	666.33	332.36
4	11/23/2010	235.48	1838.63	1886.15	425.89	809.39	202.34
5	11/30/2010	174.84	2349.66	1701.23	407.15	738.96	272.76
6	12/07/2010	162.26	1468.18	1597.11	343.23	624.3	297.61
7	12/14/2010	167.19	1625.67	1793.52	291.45	758.05	200.49
8	12/21/2010	227.38	1715.28	1588.93	311.76	643.29	302.1
9	12/28/2010	166.44	1673.56	1675.83	240.11	646.16	212.71
10	01/04/2011	212.81	1719.85	1293.54	224.95	407.34	283.88
11	01/11/2011	295.16	2050.53	1893.51	622.74	581.49	1369.7
12	01/18/2011	259.7	1724.12	1368.08	354.38	622.1	970.13
13	01/25/2011	151.79	1651.15	1585.97	416.4	689.5	652.29
14	02/01/2011	256.77	1626.84	1661.74	297.67	598.5	503.64

REFERENCES

Dunnhumby (2013). What we do. *dunnhumbyUSA.* Retrieved from http://www.dunnhumby.com/us/about-us-what-we-do

Drout, M. & Smith, L. (2012). How to read a dendrogram. *Wheaton College.*

Kumar, M., Patel, N., & Woo, J. (2002). Clustering seasonality patterns in the presence of errors. Center for E-Business at MIT.

Lohr, S. (2012). SAS makes its bid to democratize data analysis. *The New York Times.* Retrieved from http://bits.blogs.nytimes.com/2012/03/22/sas-makes-its-bid-to-democratize-data-analysis/

Nakkeeran, K., Garla, S., & Chakraborty, G. (2012). Applications of time series clustering using SAS Enterprise Miner for a retail chain, SAS Global Forum 2012. SAS Global Forum.

Perry, R. (2013). "The Complete Journey" data set. *dunnhumbyUSA.*

Rakthanmanon, T., Campana, B., Mueen, A., Batista, G., Westover, B., Zhu, Q., Zakaria, J., & Keogh, E. (2012). Searching and mining trillions of time series subsequences under dynamic time warping. *KDD.*

SAS Institute (2013). Big data – what is it?. SAS Institute. Retrieved from http://www.sas.com/big-data/

SAS Institute (2013). SAS Enterprise Miner. *SAS Institute.*

Schubert, S., & Lee, T. (2011). Time series data mining with SAS Enterprise Miner. SAS Global Forum.