SAS Insight on Toys

Module Code: MTH782P - SAS FOR BUSINESS INTELLIGENCE.

Student ID: 210691685

Student Name: Venkata Sri Naga Sailusha Peddibhotla

Question-2:

Before starting your analysis, you should access and investigate the dataset, i. e., understand and be familiar with the data set, including understanding the shape, and possible data issues. Provide detailed instructions on how this is achieved and a detailed description of the dataset.

The dataset given in SAS Drive is on Sale of Toys. The data was viewed by clicking on Data on left side of visual analytics page, which displayed distinct values of various datatypes. It is divided into four data types such as Category, Geography, Hierarchy and Measure. Character and datetime are data items which were treated as categories. Numeric data items were treated as measures. A category data item values were mapped to its respective geographical locations or regions, and it is called geography data items. Hierarchy is a datatype which showed the chain of similar items. These are related to each other by the following items such as Facility, facility state, city, country, and continent.

The SAS Data explorer page helped in investigating the dataset to know the rows and columns in the data. There are three tabs displayed named Sample data, details, and Profile. The sample data provides the content of the data in no of rows required by the user. The details showed the following: type, format and number of rows and columns. The profile has descriptive metrics, pattern distribution and frequency distribution. Descriptive metrics showed summary of data such as mean, median and mode which are useful measures. Pattern Distribution presented repeated patterns in words of the values. It is only helpful when there is a category. The frequency distribution disclosed the distribution of frequency of items in each column in a form of histogram which gave quick overview of data.

The data has fifty-seven columns and 1.4 million rows.

This dataset is about Toy manufacturing organisation and its sales for 34 years from January 1980 to October 2013. The toys are produced in manufacturing facility and sent to facilities located in different continents, countries, states, and cities to store the products. Vendors sell these toys to consumers through their stores. Sales representative help customers place their order and support in tracking no of orders placed. It also contains Target of sales rep, vendor satisfaction. The columns below were given descriptions which helped for better understanding before going to the prepare phase.

Facility- The Unique ID for representing the facility where the toys are stored

Facility City-The Cities where facilities are available

Facility Continent-Continents where facilities are present

Facility Country -Available countries where facilities are located

Facility Country Code-Code used to represent country

E. g: US -United States, IT-Italy

Facility Date Closed -The Date when facility was closed

Facility Date Opened -The Date when Facility was started

Facility State Province-The States where facilities are situated

Manufacturing Batch: The ID used to identify the batch of toys manufactured

Manufacturing Batch SKU: SKU represents Stock Keeping Unit. The Unique Code given by vendors for a particular manufacturing batch in their warehouse. It differs between facilities. Manufacturing Facility: The Place where the products were manufactured.

Order-The Code for representation of order of toy /toys.

Order Note: Information given about a particular order to communicate the issue with the order between sales rep and consumer. These contain null values. As all orders may not require a note.

Product Brand: There are two product brands which shows the product type. The two products are novelty and Toy

Novelty is a precious present /Gift which can be used as a showcase item at houses.

Toys are items used by children to play.

Product Line: The description of product based on how it looks. It is category based on features of product. There are eight lines of Product

Product make: seventy-seven products make, falling into the eight product lines
For example: Superhero, Puzzle, Party Bead etc.

Product SKU: The product code given to each good in warehouses for manufacturing facility which is same for all vendors. 779 product SKUs produced, falling into the various product styles

Product Style: The name of the product identified by public. Category 355 product styles, falling into the various product makes

E.g.: Wonder women, Cars, Tic Toe Puzzle

Sales rep: The ID of representative for the sales

Transaction Date: The Date on which order was placed and transaction took place

Transaction Day of the week: The day of the week on which transaction happened

Transaction Month of Year: Month in which Orders are placed

Transaction Year: The Year in which the sales occurred.

Transaction Month Year: The Month and Year in which people bought products.

Vendor: The Code for the Vendor

Vendor Date Ended: When the vendor contract with the products sales has ended

Vendor date Started: Date on which Vendor relationship with the sales of the products started

Vendor Distance: The Distance between vendor company and facility location.

Vendor Loyalty program: Loyalty towards the customers for sale of products. There are two values. Yes /No

Vendor Satisfaction: The response of the vendor after purchasing the products. There are two responses. Good and Bad.

Good indicates to continue the production of the products without any change. Bad represents change Quality is required in the product. If it is greater than 50% it is good otherwise bad.

Vendor Type: The Type of vendor categorized as five values with integers as 1,2,3,4,5.

There may be distinguished based on the size of the vendor. Small scale, Middle scale, and large-scale companies. It depends on Number of products purchased by vendor.

Market Penetration: The Popularity of the product in the market.

Order distribution Cost: The Cost spent on product distribution during shipping and logistics.

Order Marketing Cost: The amount of money used on promotion of the goods.

Order product cost: The Expenses on the making of product including manufacturing cost

Order Sales cost: The expenditure required for the sales of the order.

Order margin: The profit acquired after removing cost from the sales revenue generated

Order Total: This is the total revenue generated by the orders

Sales Rep Orders: Number of orders assigned to the sales representative

Sales Rep Actual: The amount of money received by sales of orders.

Sales Rep Target: The Target amount to be achieved by the sales rep by selling the orders

Sales rep % of Target: The Percentage amount of the target to be achieved

Sales Rep Vendors: Number of vendors sales rep is working

Sales Rep Vendor Base: The potential no of vendors a sales rep can have

Vendor rating: The rating given to the vendor by the customers for the products sold

The other columns represent facility co-ordinates in latitude and longitude. There are three ways of analysing data to gain better understanding before starting to fix the data issues This is achieved by list tables, automatic explanation, and cross tabs. These provide information about which factors are important for the item we are interested in. Automatic explanation of a category provides quick overview of the data. For example, it shows Product make has been common value of Athlete. This is dependent on many underlying factors like. The potential issues in dataset are null values. The column Facility data Ended does not have data, this means the facility was opened but not closed for all the facilities. The ‘Vendor Rating’, ‘Vendor date ended’, and ‘Vendor satisfaction’ also has null values.

Question-2: In the preparing phase of Visual Analytics, we create the tables needed for analysis, correct data quality issues and modify or create new data items. Provide a detailed description of the data issues in the dataset, how you resolved them, and any new or modified data items you have used in your report.

In preparing phase of analytics, Tables can be created depending on the requirement. If there are a greater number of tables, it can be combined using Join. The data issues were identified by profile tab in data explorer.

Column	Unique	Null	Blank	Pattern Count	Mean	Median	Mode	Standa...	S
SalesRepVen...	0.05% (671)				85.07	36.00	18.00	132.41	
TransactionDate	0.62% (8,826)				15,55...	16,04...	18,78...	2,813.46	
TransactionDa...	0.62% (8,826)				15,55...	16,04...	18,78...	2,813.46	
TransactionM...	0.62% (8,826)				15,55...	16,04...	18,78...	2,813.46	
Vendor	10.49% (148,...			8			00001...		
VendorDateE...	0.07% (923)	10.65% (150,...			17,47...	18,07...	19,62...	2,003.86	
VendorDateSt...	0.03% (488)				15,39...	15,88...	15,97...	2,809.19	
VendorDistance	10.49% (148,...				10.79	8.22	13.88	8.48	
VendorLat	10.49% (148,...				33.81	37.68	35.79	16.31	
VendorLon	10.49% (148,...				-77.61	-85.72	-78.39	37.24	
VendorLoyalt...	<0.01% (2)			1			Y		
VendorRating	0.10% (1,373)	0.37% (5,246)			0.53	0.55	0.77	0.30	
VendorSatisfa...	29.60% (...	0.37% (5,246)			0.57	0.55	0.66	0.12	

Figure 1:Unique and null values in dataset

The columns that had null values were ‘Facility Date Closed’, ‘Vendor date ended’, ‘order note’, ‘Vendor rating’ and ‘vendor satisfaction’. The potential reason could be that vendor has recently opened the business and hence no rating is provided. The satisfaction of vendor is not present as many don’t fill the survey forms and customers are not interested to provide feedback. Another reason for not providing feedback might be, they don’t want their identity to be revealed.

Figure 2:New calculated item

The Data issues were resolved by applying filters for not including null data from vendor rating, vendor satisfaction by using operators. From Figure-2, it can be understood on how to create a new data item. I created one new data item called **Manufacturing cost** from given data items. It is obtained by removing others costs like distribution, marketing, sales from product cost.

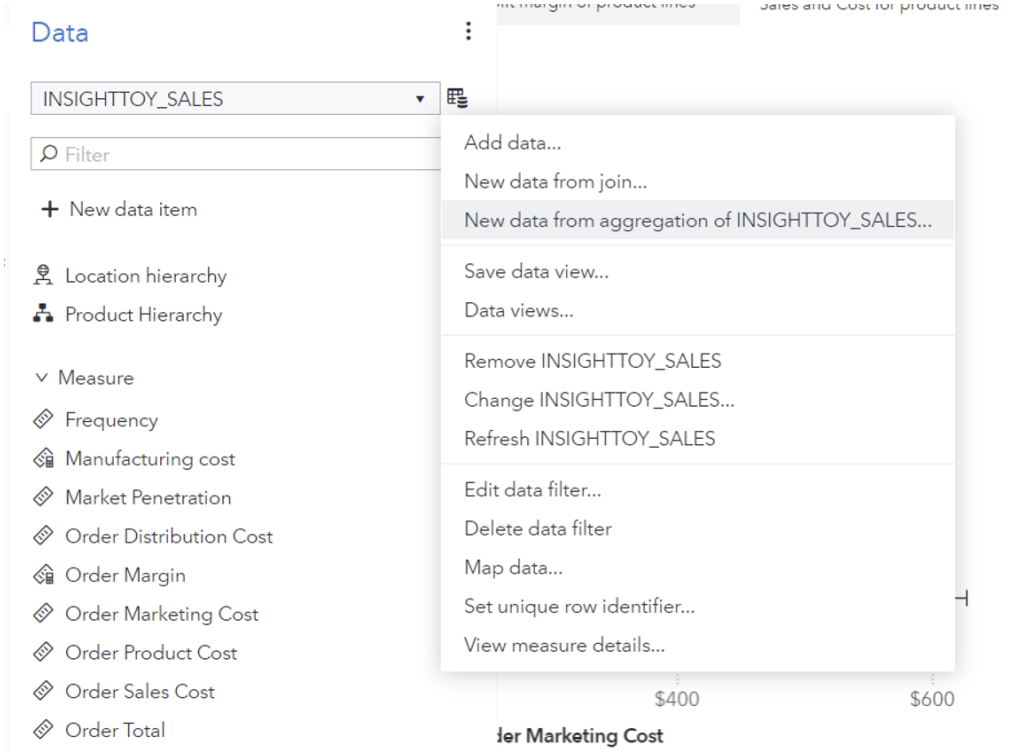

Figure 3:New subset of data

Figure-3 showed the steps to create a new data from existing data. A subset of data was created by selecting the data pane for new aggregated data from the current dataset. Graphs and charts are built on new subset of data.

Question-3:

Provide detailed explanations on how you chose your story line, what alternative story lines you considered and why you decided not to include them in your report. (15 points)

Based on the preparation phase, it was clear that analyzing all columns is tedious and not efficient. Hence, considered few columns and examined the sales of products based on product lines.

Storyline was decided to conduct analysis on various product line sales over years as it had more scope to explore how product lines influence sales. This is helpful in prediction of future sales. Products were categorized based on product make, style, product brand and product line. There are two product brands named Toy and Novelty. There were eight distinct values for product lines such as Figure, Game, Plush in Toy and Bead, Gift, Promo, Thrift, and Beach in Novelty.

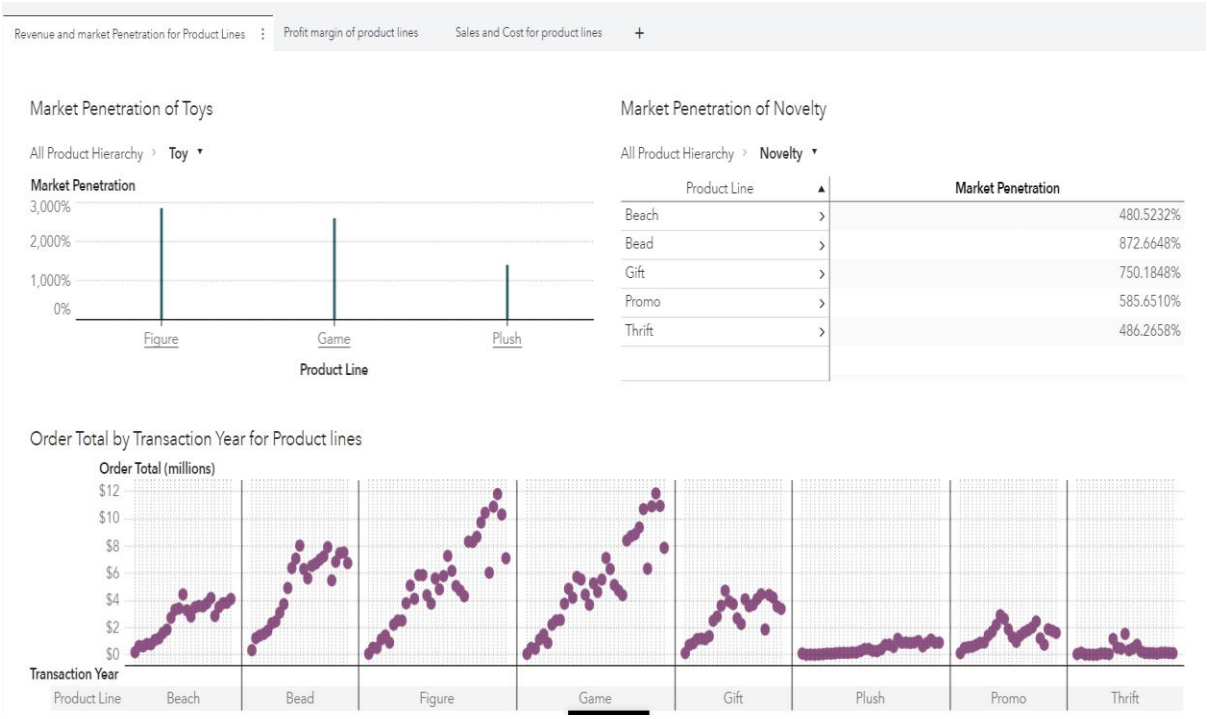

Figure 4:Revenue and market penetration for product lines

Report consists of three pages. Figure-4 showed first page consisted of Revenue and Market penetration for product lines. After Looking at the visualizations, it was clear from needle plot and cross tab that Toy has more market penetration compared to novelty. The market penetration was higher which explained the popularity in the industry. The Figure and Game were most demanding. Plush also had more demand compared to novelty products. The bead was popular product among novelty. The aim was to reduce the costs required for each of product lines and improve the sales and market value. In Dot plot Revenue generated by game and figure had an increasing curve which showed positive response from customers over the years as type of figures change from generation to generation. Bead, beach, and gift also gained sales over years. Other products were improved but very slowly. It would be helpful to marketize the other products to increase the revenue.

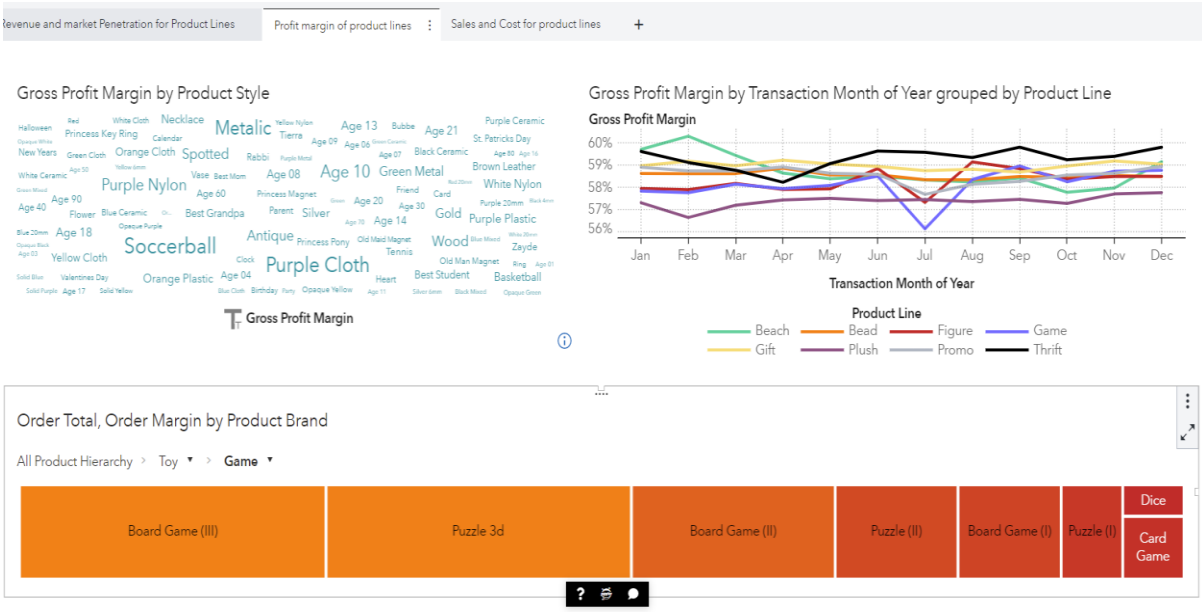

Figure 5:Profits and Order total, order margin for product lines

Figure 5 showed Profit margin of product lines, style and make during the months of a year which helps in focusing on sales of all products in different months. Gross Profit margin by product style. The most profitable product style is Soccer ball in novelty which is maximum in July due to summer Holidays for children. Another product style named purple cloth which is backpack also higher profits during new term of the school. Tree Map was useful to see many categories in one map with assorted colors and sizes. The Board games III was the most selling product make in game type. Dice was the least selling one in Toy.

Plush sales were high during summer and dropped in February. It is profitable to focus on marketing and manufacturing of the goods in other months to meet the demand. The product line Game had rapid decrease during July as parent take children on vacation due to holidays. Increase in Gift profits during February and December might be due to valentine's day and new year. The profit is constant over the entire year for gift. Thrift has improved from January to December and highest in December. Plush seemed to have least profit, among others. Promo and figure also had a dip in July due to summer in few countries where children prefer to play outside.

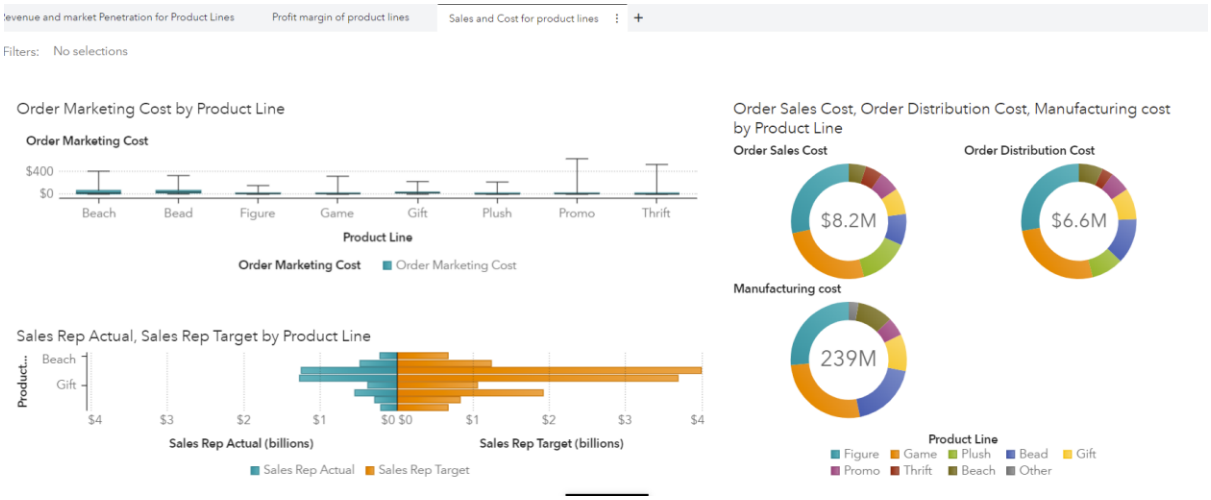

Figure 6:Sales Rep Actual /Target and Various cost of product lines

From Figure-6, it provided Sales representative and marketing cost for product lines page, marketing cost spent on each of the products were observed. The promo had wider range of distribution compared to other products. Plush had very less marketing cost which might be the reason for the sales rep actual not meeting the target of consumers. Figure and game marketing cost were not enough to meet the demands. It also might be due to higher requirement than supply. Figure and Game production should be planned in July for winter months where sale increases. There are different costs which contributes to making a product. Some products may have more manufacturing cost than marketing. It is important to reduce and manage cost of each product line. Three pie charts created for manufacturing, distribution, and sales cost. All costs for figure and game were higher than others. Whereas plush had more sales cost among novelty and bead had higher manufacturing cost.

Alternate story lines could be analyzing customer satisfaction over years. It is qualitative data which makes it difficult to analyze the vendor satisfaction as it depends on many factors making it subjective. Another possibility was analysis of sites, countries that were doing well in products, how sales have improved in different countries. How costs were improved due to recession which affects overall business of the organization is another storyline. The chosen story line is due to modesty in understanding the data clearly from visualization without knowledge about the market before. It is important to interpret the data clearly to make any decision and sales are the important attribute for success in competing market.

Question-4:

During your report design process, you must decide on the appropriate visualization tool/type to use based on the data you choose, and information you intend to portray. How did you choose the objects in your report (charts, tables, etc)? How will the charts be perceived by a non-technical user? What questions may he/she ask and what answer(s) could she get with it?

Graphs are taken depending on comparison, relationship, and composition for the chosen story line. It was decided to go with needle and cross tab for market penetration as there are not many categories which provides clear view of where there is less and more. Selected dot plot for the revenue generation as other chart is not providing clear distribution for each product line over years. Line chart explains increase and decrease over months pinpointed. Having a scatter or

dot plot may not give the upward and downward direction clearly. To express many measures and categories. Box plot provided distribution for each category which shows where the costs can be reduced. Butterfly chart provides a difference of sales rep and actual. They both are comparable quantities. Actions were created for third page to see the change is one graph one object was selected in a chart. Having them on same axis makes it more clean for non-technical users. Actions created for third page made the change in one graph was reflect in another chart.

Initially chosen line chart plot to show different product make which is not great as they overlap on each other. Tree Map clearly explains using product hierarchy for size and colour to show the order total and order margin for product make. The questions they might ask is why pie chart and Cross tab. I used bar chart to show the market penetration instead of cross tab. The data labels were not shown in bar chart. Pie chart is useful when there are many measures and one category. Word Cloud was used to find the most profit generating product style.

Question-5:

Include step-by-step explanations of how the report is built (e. g., how filters are built in the report, how various charts and objects are built, etc.) and how key information could be accessed.

A chart was created by choosing appropriate objects for the data and added it to the current page. Assigned the data with categories and measures needed for the graph. On the right side of the page, there are options, roles, actions, filters, and ranks. Based on the requirement, they were be selected. Added one filter for considering only non-missing values for vendor satisfaction and vendor rating as they contain null values. In the roles tab, the measures and categories were edited, removed, or added. In options, the chart was changed suitable to the chart and made it aesthetic to look. Name of the charts and title can be renamed. The colours on the page can be changed to make it more meaningful and cleaner. The choice of colours is also important as it may change the way we look at graphs. Preferred contrast colours to make it easier to understand the graph. Implemented different graphs such as Needle plot, Crosstab, Dot plot, Line chart, Pie chart, Heatmap, Box plot and Butterfly chart.

Needle plot is useful when it is required to show the height and not width. Also, if there are many values in one category. Product Hierarchy is considered as x-axis and y-axis as market penetration. Cross tab made it simpler to read with less rows and columns with product hierarchy and measure as market penetration. The values are accurate and not approximate. Dot plot is like scatter plot, it shows distribution of the plot. It is helpful if the pattern is important in the data not the actual values, when there many data points with category as transaction yean, measure as order total and lattice columns as product line. Line chart displayed the flow of profit across years which enabled understanding of peak and fall in each month. It has transaction month of year as category and gross profit margin as measure and group as product line.

Tree Map is used when there are several categories in data. Applied product hierarchy to explore each product make, line and style by clicking on each part on Tree map. Here the map presented both size and colour with different measures order total and order margin and tile as product hierarchy. Pie chart had category as product line and allowed more than one measure such as order distribution, manufacturing, and sales cost. Word cloud was created by assigning

word as product style and size as gross profit margin. Box plot has category of product line and measure as marketing cost. Butterfly chart displayed sales rep actual and target as measure for category all product lines.

.