Analysis of annual sales of a supermarket chain

Piyush Verma

MS Business Analytics, University of Cincinnati

# Introduction

Dataset used in the analysis is from an open dataset hackathon competition (now closed) which is deidentified by the owners. It consists of 3 tables: Transaction table, which has records for each product bought in any of the supermarket’s store, Product table which details the hierarchy of the product according to the Department to which they belong and a Demographic table, which identifies the attributes at the customer level.

# Analysis

Data Cleaning (SQL server): The original retail dataset contained 2.5 million records at customer-store-product-day level comprising sales of 2 years. For the scope of the project and the workability around accessing the data through R and Tableau following steps were taken:

1. Restricting the records to the 1st year sales by filtering the Transaction table
2. Renaming certain obscure column names and changing the datatypes
3. Created a VIEW in SQL-SERVER on top of Transaction, Product and DEMOGRAPHIC table.
4. SALES\_VALUE column ofTransaction table contained certain records with scientific notation like below, which were inhibiting operations like data type change. Hence, all such dirty records were identified by their char length and removed by updating the transaction table

Below table gives information about the columns used in the created view:

|  |  |  |  |
| --- | --- | --- | --- |
| Table Name | Column Name | Data Type | Description |
| [dbo]. [transaction\_data] | H\_KEY | INT | Unique Identifier of a family or household |
| BASKET\_ID | BIGINT | Unique transaction ID or Bill number |
| PRODUCT\_ID | INT | Product identifier |
| QUANTITY | INT | Number of units purchased of a product |
| SALES\_VALUE | NUMERIC | Sales from a product |
| STORE\_ID | INT | Store of transaction was |
| RETAIL\_DISC | NUMERIC | Discount given |
| [dbo].[product] | PRODUCT\_ID | INT | Product identifier |
| DEPARTMENT | VARCHAR | Department of product |
| [dbo].[hh\_demographic] | H\_KEY | INT | Unique Identifier of a family or household |
| AGE\_DESC | VARCHAR | Age of the Family head |
| MARITAL\_STATUS\_CODE | VARCHAR | Marital status of Family head |
| INCOME\_DESC | VARCHAR | Income range of the Family head |
| HOMEOWNER\_DESC | VARCHAR | If the Family head owns a house |
| HH\_COMP\_DESC | VARCHAR | Family category: With/Without kids |
| HOUSEHOLD\_SIZE\_DESC | VARCHAR | Number of family members |
| KID\_CATEGORY\_DESC | VARCHAR | Number of kids in the customer’s family |

Data Manipulation (SQL server): Created 3 new metrics for each house hold key (H\_KEY) which quantifies their spending behaviors. DENSE\_RANK, OVER and PARITION BY functions are used. Following were the 3 metrics:

|  |  |
| --- | --- |
| FAMILY\_TOT\_SALES | How much has a family spent in the entire 1 year in the supermarket |
| FAMILY\_TOT\_VISITS | How many times or how frequently has the family visited the supermarket |
| FAMILY\_VALUE (Basket Value) | How much was spent in each of those visits on an average by a family |

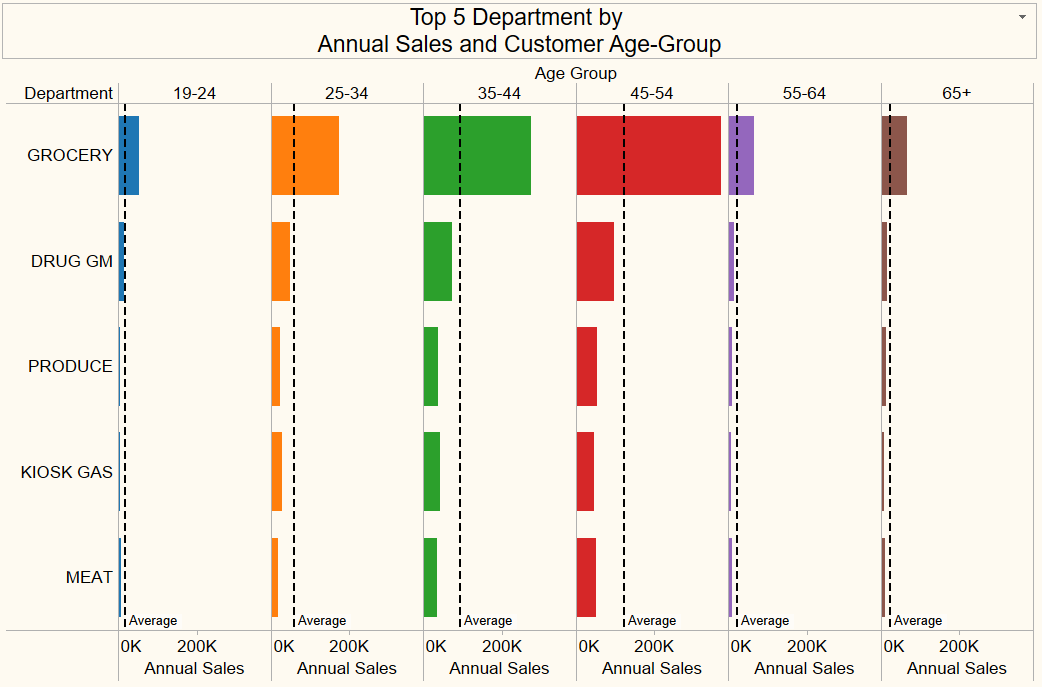
For example: A high value of FAMILY\_VALUE means that whenever the family visits the supermarket they tend to spend more in that visits.

Data Exploration PART-I (Tableau): The above created view is now connected to Tableau to produce quick insights on questions like:

1. Which departments attract the most customers by sale
2. What are the demographics of the customer base
3. Relation between the above two

Insight-1:

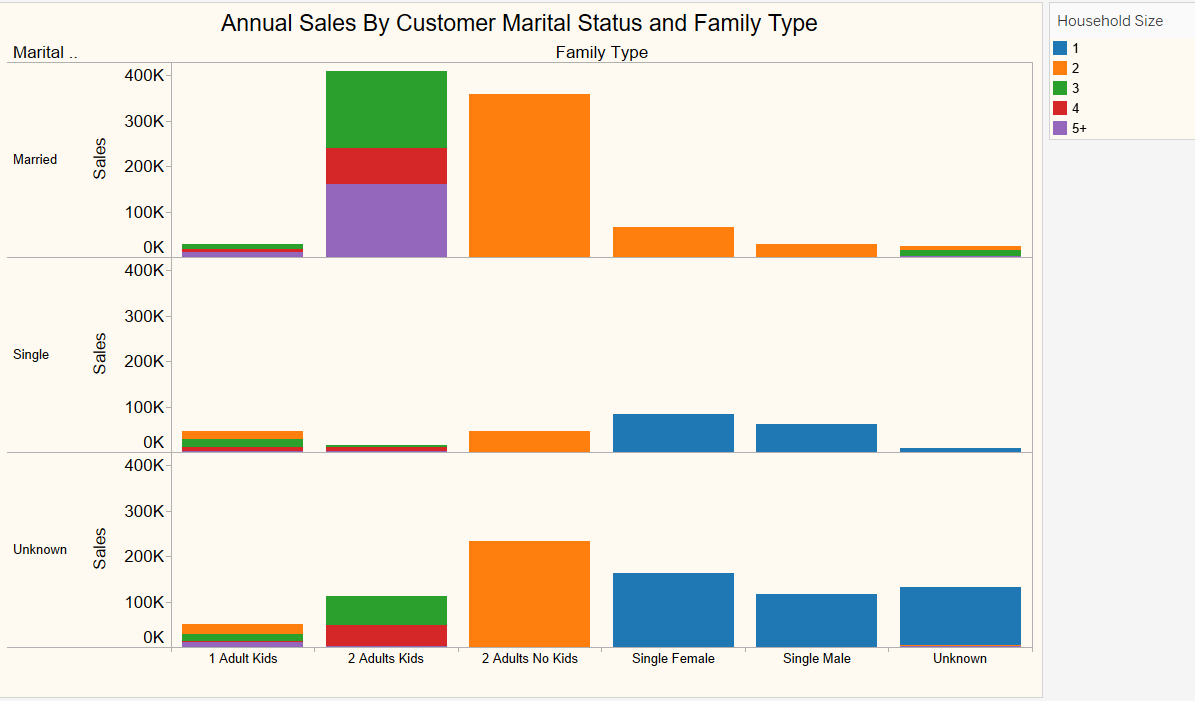
Top 5 department which generate the most revenue are Grocery, Drug, Produce, Kiosk-gas and Meat. Further, most of the customers are aged between 35 – 44 and they mostly buy in Grocery.



Insight-2:

Top customers are generally married with 2 or more kids in their family size as indicated by the below stacked bar.

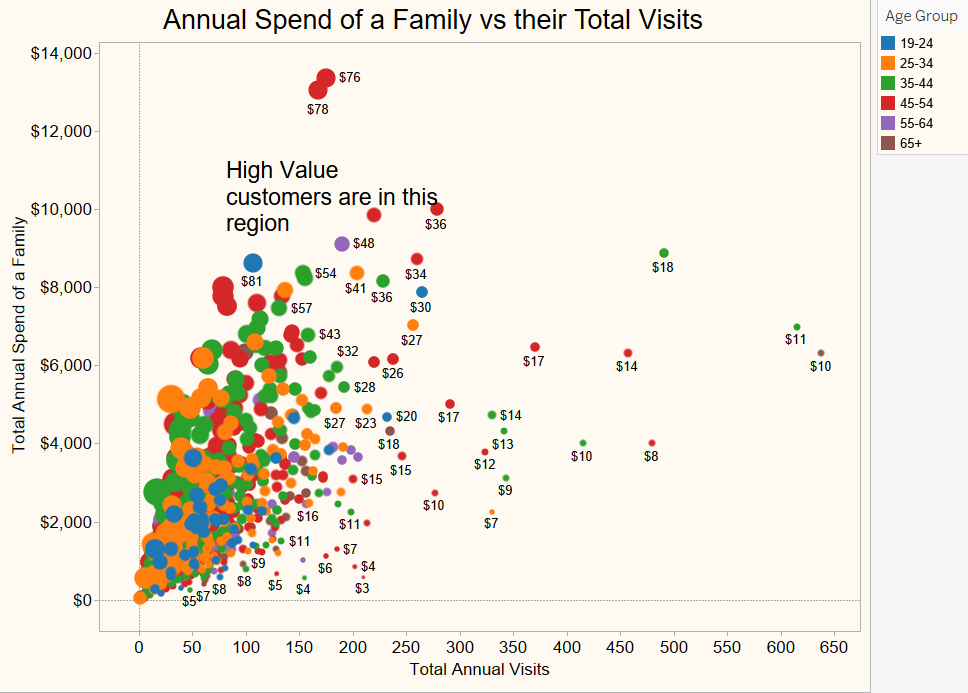
Possible Data Discrepancy: If we look carefully for “***Married***” families we can observe that certain sales are coming from families which are marked as “***Single Female***”, “***Single Male***” but yet are identified by MARITAL\_CODE\_STATUS as “***Married***”. This seems to be a Data discrepancy which can be brought o attention to the data owners.



Insight-3:

Generally, one can have a notion that “***MOST FREQUENT CUSTOMERS SPEND THE MOST”*** dollars at a supermarket. BUT, below scatter plot reveals that it is “***NOT TRUE***”. It reveals that there are certain families who despite of going lesser than the most frequent shoppers spend way lot more. These customers have been termed as HIGH-VALUE customers because they have the potential of shopping big in the supermarket. The annotated region in the plot below envelops customers who are HIGH-VALUE. If we look closely we can observe:

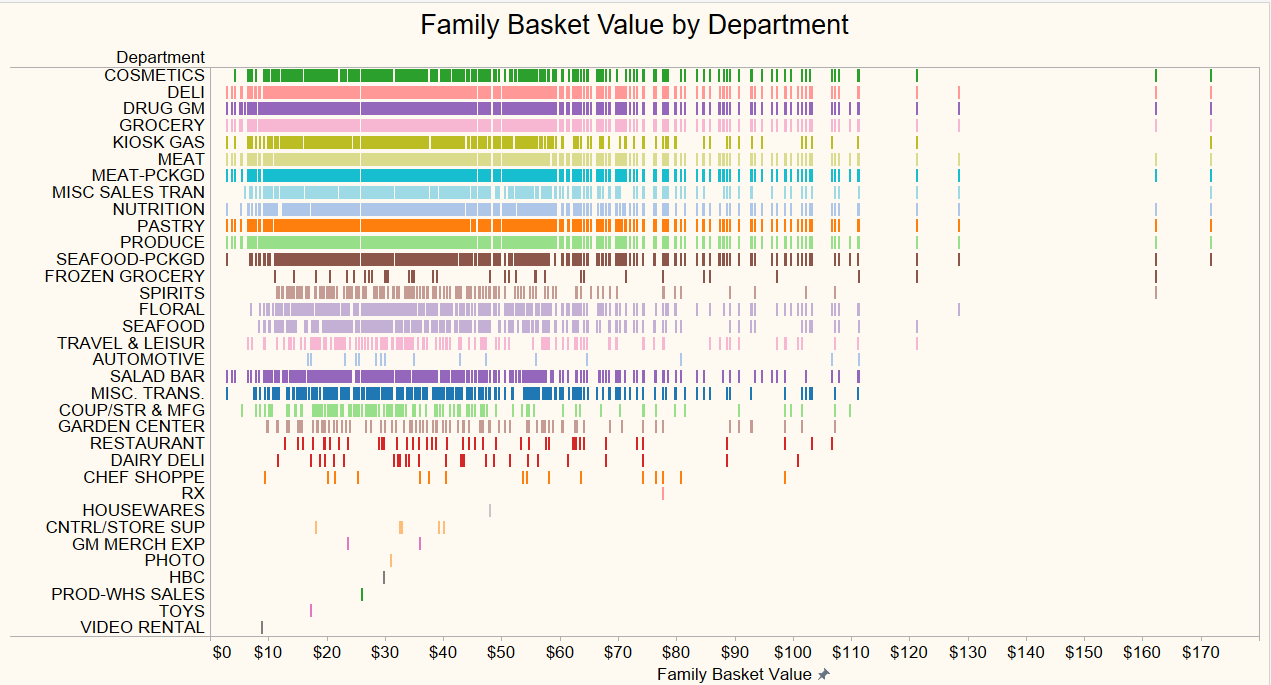
1. These HIGH-VALUE customers shops anywhere between 150 – 200 times in a year
2. Moreover, the basket-value i.e. Dollars spent per supermarket transaction is generally around and above $50.
3. Their annual expenditure is anywhere around $8000 - $13000
4. Also, there is a good number of Red dots which are somehow away from the bottom dense could. These red outliers as depicted are from customers who are aged between 45 – 54.



Insight-4:

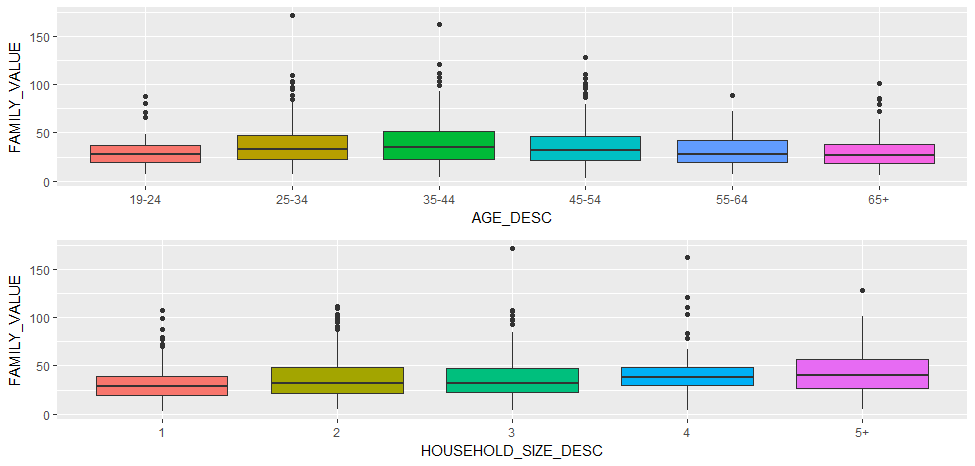
Below plot is a GANTT VIEW. Here each strip denotes 1 transaction in a Department against the size of the Basket Value in Dollars in which they were shopped. High value baskets are of importance of supermarkets. By looking carefully to the plot below we can observe that:

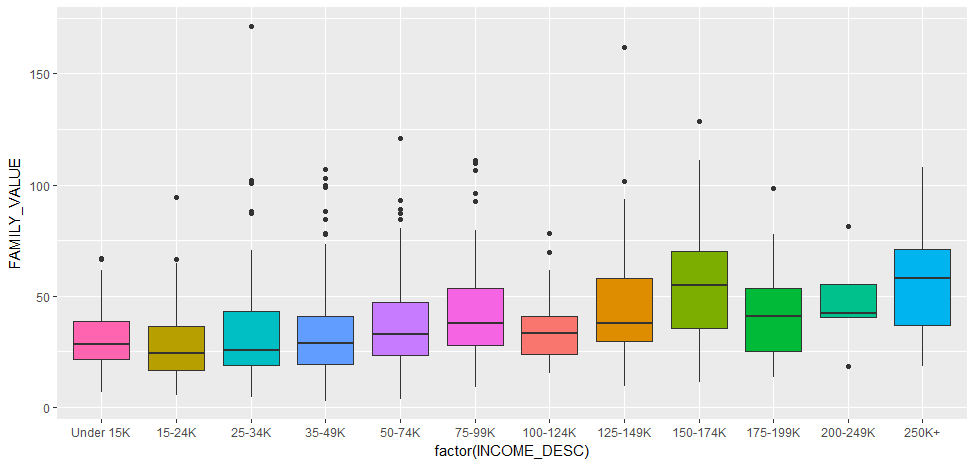
1. the “HIGH-VALUE” baskets contain only certain Departmental products (like Cosmetics, Deli, Drug, Grocery) and doesn’t involve certain other Departments (like Video Rental, Housewares, Toys, Travel & Leisure products)
2. This plot is important in the sense that a marketing campaign can be designed on those low-performing departments to drive the sales towards the “HIGH-VALUE BASKETS”
3. Moreover, the strips are denser around more frequently sized basket-values. With that analogy we can say that majority of the baskets have size ranging roughly between $20 - $40 almost for all Departments. But again, certain good performing Departments like Cosmetics, Meat and Sea-Food have a good number of baskets around size between $80-$100. These are the Departments whose quality need to be consistent or improved further to drive more sales.



Data Exploration PART-II (in R using GGPLOT2 package): Additional data exploration is done in R software using ggplot2 package. The main idea here is to capture the distribution of the Basket Value with each of the categorical attributes defining the family like Age-group, House hold size and Income. Below are the results:

Insight-5: As we can see, Basket Value Size increases with the Age-group till 44 and then it decreases. Basket Value Size consistently increases with the house-hold size (which make sense: big the family more the expenditures). With the income group as well, the Basket Value Size increases with higher income groups.

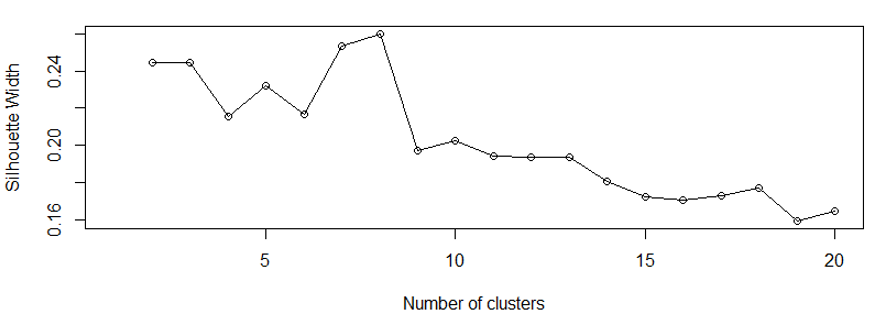




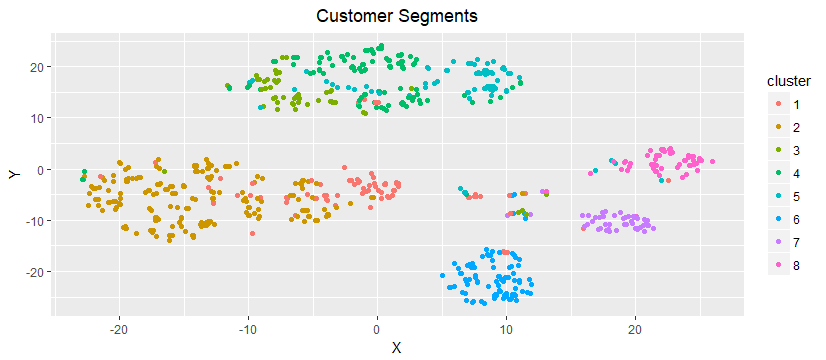
Advanced Data Analysis: (in R):Proceeding with the analysis done above, if a supermarket wants to target customers selectively, they need to tag customers based on their value to them. A supermarket won’t find it useful sending high-end electronic equipment sale offers to a customer who often buy only Groceries in small basket sizes. Hence the clustering of customers should be the next step.

To perform grouping, “***K-Means Clustering***” technique is generally used, but since K-means can only be used for quantitative variables, we can’t use it in this case. “***K-mode***” algorithm is one algorithm which can handle the categorical variables. It is implemented in this problem utilizing the “***cluster package***”. Below scatter plot identifies the families based on their newly defined “***Customer-Category***”:

First, we needed to find an optimal number of clusters in which the customers can be segmented. For this we used the “silhouette-width” method, in which we look for the number of clusters for which the “silhouette-width” is maximum. We found that it was maximum for eight clusters.

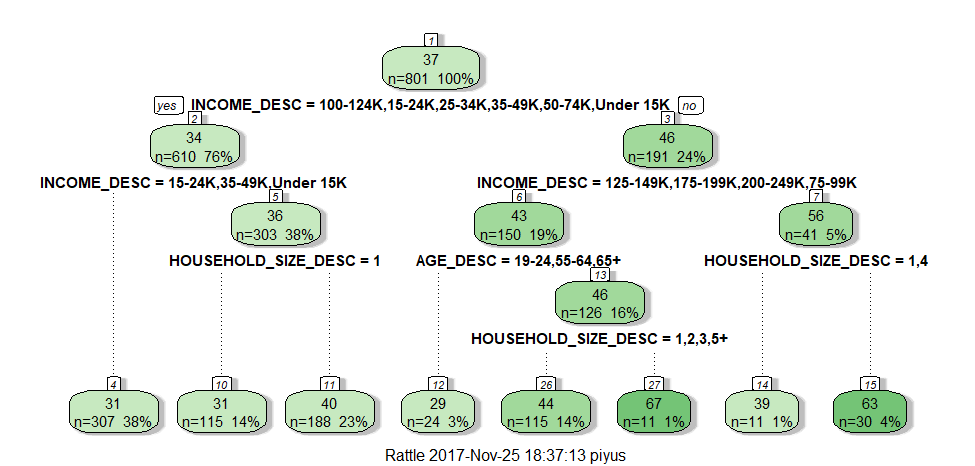


Finally, 8 clusters were made based upon above graph. Following are the clusters visualized in 2-D:



Analysis was finally completed by building a random forest model predicting the FAMILY\_VALUE / Basket Value Size associated with a family. Below is the generated decision tree:

1. Darker the green, higher the predicted Basket Value Size. This again reveals the information that if a family has a higher income and has big family size (4 or 5+), they generally tend to have higher expenditure.



Summary: From all the above analysis with the limited dataset, from a high-level view we can say that:

1. “*Mid-Age bigger families who shops a lot in groceries are the major customer base for the given supermarket. And for the future marketing campaigns, Supermarket can focus on this customer segment and their preferred Top 5 Departments: Groceries, Drug, Produce, Kiosk-Gas and Meat*”.
2. “*There are 8 different segments of customers present”*

Challenge: The biggest challenge was the difficulty in identifying the dirty transaction records which were not letting the data type to change (at least for 2 columns: SALES\_VALUE and RETAIL\_DISC). The next challenge was exploring the data in Tableau which was time consuming because of the various inter-related categorical variables, and it took time to separate and decide which categories are influencing sales (like Income group, House hold size, Departments) and which one don’t add any additional information (like Brand, Manufacturers etc.)

Future Scope: Store geographic location data could also have added insights to locate high performing stores so that the supermarket can optimize inventories in those over / under utilized stores.

Code and Tableau Dashboard:

Tableau: The analysis done in this report is available in Tableau public[here](https://public.tableau.com/profile/piyush.verma#!/vizhome/AnalysisofaSupermarketChain/Final?publish=yes)**.**

Github: