Data Preprocessing and Visualization

MEGA MARKET

Group Project Report

Group D

15th December 2023

João Capitão | 20221863

Margarida Cruz | 20221929

Luís Mendes | 20221949

Dinis Gaspar | 20221869

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

ABSTRACT

The ability to monitor its business is certainly an advantage to Mega Market in such a competitive area as retail market. Given the necessity of data treatment, this document uses pre-processing techniques to trace a profile for the company's customers. Data containing details of past transactions is analysed. Moreover, the company lacking information on its own activity is addressed.

KEYWORDS

Market; Data; Client Profile; Customer Behaviour

INDEX

1. INTRODUCTION

Mega Market aims at developing different analysis that will help the growth and success of the company. Therefore, understanding the business activity and its customers' shopping behaviour is the right path to be on. This necessary exploratory analysis will very likely lead to an improved customer service and increased client satisfaction. Hence, data preprocessing techniques must be used to reach the required insights.

The work here presented relies on the dataset provided by the company. It reflects the business' situation and contains past customers' transactions, which is vital to outline the profile of the company's customers. Once these are analysed, Mega Market will be more aware of its dealings' reality and, subsequently, better equipped to decide on the next step of its growth strategy.

The present report is expected to significantly contribute to that end, being organized as follows: section two details the methods employed, whereas sections three the results and section four the main conclusions. Finally, section five presents the limitations and suggestions.

2. METHODOLOGY

The steps taken on the used softwares (SAS Enterprise Miner, SAS Guide and PowerBI) are presented in this section of the report. The extracted insights are also exposed by means of graphics, tables or plots.

2.1. Data Pre-Processing

The procedures regarding data preparation were developed in SAS Enterprise Miner. The corresponding diagram can be consulted in **Appendix Fig. 1**.

2.1.1. Variable Definition

The data stored in Mega Market's information systems represents its customers' transactions. Thus, there is not a target variable. In order to determine how each variable in the dataset provided by the company will be used from this point on, the roles and data type definitions were set according to **Fig. 1**[1] on the *File Import* Node.

Name	Role	Level	Report	Order	Drop
Age	Input	Interval	No		No
Channel	Input	Nominal	No		No
CustomerNo	ID	Nominal	No		No
Date	Time ID	Interval	No		No
Gender	Input	Nominal	No		No
Kids	Input	Nominal	No		No
Monthly_Income	Input	Interval	No		No
Nationality	Input	Nominal	No		No
Payment	Input	Nominal	No		No
Product_Category_ID	ID	Nominal	No		No
Product_Category_Name	Input	Nominal	No		No
ProductID	ID	Nominal	No		No
ProductName	Input	Nominal	No		No
Quantity	Input	Interval	No		No
Reviews	Input	Nominal	No		No
Total_payed	Input	Interval	No		No
TransactionNo	ID	Nominal	No		No
Unit_Price	Input	Interval	No		No

Figure 1 *– Initial Variable Configuration*

2.1.2. Descriptive Statistics

From the *StatExplore* Node, it is possible to extract some basic preliminary insights that lightly show the dataset at hand. Regarding Mega Market's customer base, looking at **Figures 2-3** descriptive statistics one denotes that most clients have children, and are predominantly male individuals. Also, almost 91% of the clients are from the United Kingdom, the average age is 54 years old and customers spend, on average, about 23 monetary units on 3 items per transaction.

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	Channel	INPUT	2	0	Store	91.72	Online	8.28
TRAIN	Gender	INPUT	3	0	M	61.95	F	36.61
TRAIN	Kids	INPUT	3	2266	1	75.15	0	22.59
TRAIN	Nationality	INPUT	31	0	United Kingdom	90.81	Germany	2.42
TRAIN	Payment	INPUT	3	0	Paypal	40.07	Credit Card	34.98
TRAIN	ProductName	INPUT	513	0	Alarm Clock Bakelike Red	1.06	Alarm Clock Bakelike Ivory	0.79
TRAIN	Product_Category_Name	INPUT	9	0	Miscellaneous	87.12	Decorative items	8.32
TRAIN	Reviews	INPUT	3	57701	.	57.70	0	21.26

Figure 2 *– Class Variable Summary Statistics*

Variable	Role	Mean	Standard Deviation	Non Missing	Missing
Age	INPUT	54.36123	15.11825	97733	2266
Monthly_Income	INPUT	2506.403	1443.184	97733	2266
Quantity	INPUT	3.0133	1.743847	99999	0
Total_payed	INPUT	22.91824	17.8784	99999	0
Unit_Price	INPUT	7.628366	4.291977	99999	0

Figure 3 *– Interval Variable Summary*

To what regards customer activity, the large majority of transactions are made in-store, with PayPal as the preferred payment method. Moreover, 513 different products grouped into 9 distinct categories were purchased overall. Also, **Fig. 2** allows to state "Miscellaneous" and "Decorative items" as the two most commonly bought categories, accounting for over 95% of the total purchases.

These descriptive statistics also highlight one of the problems to be later addressed in this preprocessing stage – missing values. Variables *Kids*, *Age* and *Monthly_Income* are indicated to have 2266 missing values out of a total of 99999 observations. Using the filtering tool in Excel, it is possible

[1] **NOTE:** customer age (AGE); sales channel name (CHANNEL); customer ID number (CUSTOMERNO); transaction date (DATE); customer gender: M/F/O (GENDER); 1 = customer has kids (KIDS); customer monthly income (MONTHLY_INCOME); customer nationality (NATIONALITY); customer payment method (PAYMENT); product category ID number (PRODUCT_CATEGORY_ID); name of the product category (PRODUCT_CATEGORY_NAME); product ID number (PRODUCTID), name of the product (PRODUCTNAME), number of items bought (QUANTITY), 1 = customer left a review about the product (REVIEWS); amount spent by the customer (TOTAL_PAYED); transaction ID number (TRANSACTIONNO); product unit price (UNIT_PRICE).

to conclude that these are missing simultaneously in the same transaction, meaning that there are 2266 rows without information for these three variables – Missing Not At Random (the customer opted not to share those personal details). On the other hand, variable *Reviews* is refered to have 57701 observations with missing information.

The *Nationality* column takes 31 different nations as its values, one of them being "Unspecified" (**Appendix Fig. 6**). This was not interpreted as a missing value. Instead, it was assumed the customer simply did not provide that information.

Additionally, since *ProductName* presents 503 different levels, it revealed to be necessary to reject this variable to check the *MultiPlot* Node results and later in the following steps.

From the *MultiPlot* Node, the presence of outliers is evident in variables *Age* (a customer with 299 years old), *Monthly_Income* and *Quantity*. **Figures 4-6** illustrate the histograms of those variables. Other histograms can be consulted in the Appendices section.

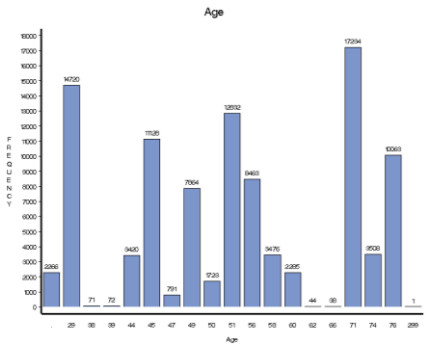

Figure 4 *- Age Histogram*

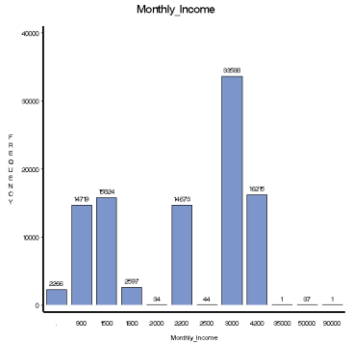

Figure 5 *- Monthly_Income Histogram*

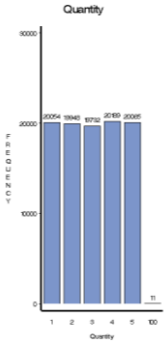

Figure 6 *– Quantity Histogram*

Although the 39 *Monthly_Income* values above 30000 monetary units are not impossible, they were considered extreme compared to the monthly income values of the other clients in Mega Market's customer base. Similarly, *Quantity* variable follows a uniform distribution with the exception for 11 outlier observations with a number of items sold of 100 units (by looking at the histogram it is possible to indicate that the highest quantity below 100 is 5 units).

2.1.3. *Filter* Node

According to the analysis made above, the boundaries stetted to eliminate outliers are stated in **Fig. 7**. This step excluded 50 observations (**Appendix Fig. 12**), which corresponds to about 0.05% of the total number of observations.

Name	Report	Filtering Method	Keep Missing Values	Filter Lower Limit	Filter Upper Limit	Role	Level
Age	No	User Specified	Default	18	90	Input	Interval
Date	No	Default	Default	.	.	Time ID	Interval
Monthly_Income	No	User Specified	Default	0	5000	Input	Interval
Quantity	No	User Specified	Default	0	10	Input	Interval
Total_payed	No	Default	Default	.	.	Input	Interval
Unit_Price	No	Default	Default	.	.	Input	Interval

Figure 7 *– Filter Boundaries Configuration*

2.1.4. Multidimensional Outliers

Being very sensitive to multidimensional outliers, the K-means method used in the *K-means* Node (standardized variables and *Princompt* method to spread the seeds of the clusters in a uniform way) had the purpose of segmenting the observations and identifying a significantly small group of observations which is multidimensionally different enough to be considered a unique cluster group.

Once again, ProductName variable was not used due to its many levels. Since it represents the same information, *ProductID* was also rejected. Also, given that *Product_Category_Name* is already to use, *Product_Category_ID* was not (same information).

The results obtained testing a different number of clusters each time (5 seeds – 10 seeds) showed that there was always a smaller segment, indicating the existence of multidimensional outliers. However, the optimal division shown in **Fig. 8** – the one that produced the smallest cluster - was achieved using 9 seeds. This segment groups 2884 observations (**Appendix Fig. 12**), which corresponds to 2.88% of the total number of observations (to be excluded in SAS Guide).

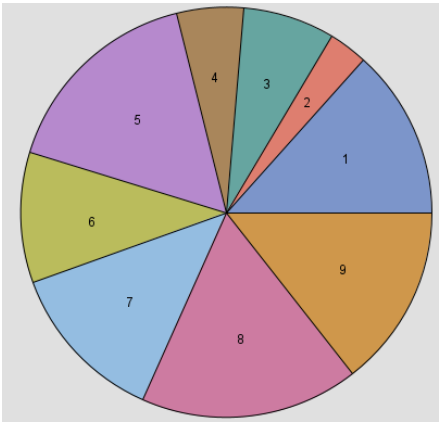

Figure 8 *– K-means Node Results (9 seeds)*

2.1.5. *Impute* Node

With the outliers' subject being solved, this node will address the missing values situation of variables *Kids*, *Age* and *Monthly_Income* using a Tree input method.

The variable edition required to this step is demonstrated in **Fig. 9**. Given that they resulted from the previous node, variables *_SEGMENT_LABEL_* and *Distance* are not from the dataset made available and therefore should not be used. On the other hand, *Product_Name* must be rejected again to avoid errors.

Name	Use	Method	Use Tree	Role	Level
Age	Default	Default	Default	Input	Interval
Channel	Default	Default	Default	Input	Nominal
Distance	No	Default	Default	Rejected	Interval
Gender	Default	Default	Default	Input	Nominal
Kids	Default	Default	Default	Input	Nominal
Monthly_Income	Default	Default	Default	Input	Interval
Nationality	Default	Default	Default	Input	Nominal
Payment	Default	Default	Default	Input	Nominal
ProductName	No	Default	Default	Input	Nominal
Product_Category_Name	Default	Default	Default	Input	Nominal
Quantity	Default	Default	Default	Input	Interval
Reviews	Default	Default	Default	Input	Nominal
Total_payed	Default	Default	Default	Input	Interval
Unit_Price	Default	Default	Default	Input	Interval
_SEGMENT_LABEL_	No	Default	Default	Rejected	Nominal

Figure 9 *– Variable Edition, Impute Node*

As indicated in **Fig. 10** and supporting the previous insights, it is worth noting that variable *Reviews* is refered to have more than half of its values (57.69% of the total observations) missing. For that reason, it was automatically rejected.

Rejected Variables Summary
Number Of Observations

Variable Name	Label	Percent Missing
Reviews	Reviews	57.6934

Figure 10 *– Rejected Variables Summary, Impute Node*

It now becomes interesting to re-analyse the variables which had their missing values computed. The corresponding histograms are exposed bellow in **Fig. 11-13**. Their interpretation allows to conclude that variables *Kids* and *Monthly_Income* (re-named *IMP_Kids* and *IMP_Monthly_Income* respectively) had their missing values successfully imputed, as the variables' distribution remains similar to the previous ones. However, variable *IMP_Age* presents some inconsistencies (to be addressed) as some of the imputed values are not integers.

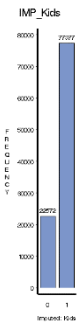

Figure 11 *– IMP_Kids Histogram*

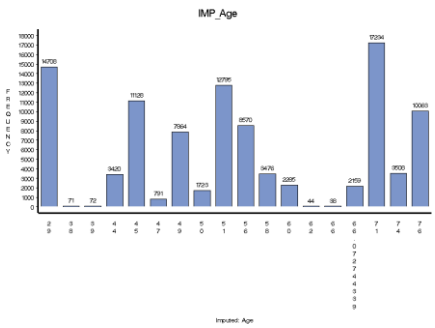

Figure 12 *– IMP_Age Histogram*

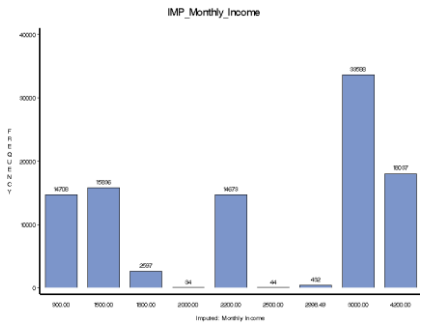

Figure 13 *– IMP_Monthly_Income Histogram*

Since all other variables did not suffer any alteration, there is no need to analyse them again. Moreover, given that the pre-processed transactional table is not meant to be used in clustering, its variables' correlation analysis will not be performed. Nevertheless, it might be consulted in **Appendix Fig. 14**.

2.1.6. *Save Data* Node

With this stage completed, the *Save Data* Node was utilized to export the pre-processed transactional table to an Excel format. This data is referred to as *SAS_exported*.

2.2. Coherence Checking

The processes of identifying and solving the incongruities found in the *SAS_exported* dataset were developed in SAS Guide. This is a crucial step as it prevents derived variables from reflecting existing problems.

Firstly, and in accordance with what was mentioned in portion ***2.1.4. Multidimensional Outliers***, it is necessary to eliminate the observations grouped in the second cluster (where *_SEGMENT_* variable is equal to 2). Also, *_SEGMENT_* and *_WARN_* features should be eliminated.

Secondly, the code presented in ***Annex 1 – Coherence Checking Code*** includes some verifications which allow a less manual process when handling future data. Although the filtered observations up to this moment regard customers over 18 years old, that either do or do not have children and are either male, female or individuals that identify themselves with other gender, with the impossibility of performing a transaction with a non-positive quantity sold, product unitary price or total amount paid; such verifications make sure future datasets also fulfil these conditions. Moreover, guarantying that the *Total_payed* column results indeed from the product between *Quantity* and *Unit_Price* columns also automates a standard verification.

Additionally, when a purchase is made online paying with physical cash is not an option. To note that when a transaction is performed in-store and payment is made via *PayPal* will not be considered an incongruity due to the high number of observations that would be excluded.

Finally, it is good practice to explore if the same *CustomerID* has different information associated to it depending on the transaction. In this case, the clients with more than one gender or monthly income values were removed.

As a result, 3.15% of the total observations of the original dataset were excluded (see **Table 1**).

Table 1 – Excluded Observations Summary

Moment of Exclusion	DATA	Filtered	Excluded	% Excluded
Filter Node	99999	99949	50	0.05%
Multidimensional Outliers (*_SEGMENT_* = 2)	99949	97065	2884	2.88%
Inconsistency: when a purchase is made online is not possible to pay with physical cash	97065	96976	89	0.09%
Inconsistency: customers with different information depending on the transaction	96976	96853	123	0.13%
TOTAL			**3146**	**3.15%**

2.3. Transactional Data Insights

The *transactional_table* dataset – which resulted from the coherence checking stage - will be uploaded to *PowerBI*. This exploration aims to conclude on Mega Market's business situation and to outline a profile of the company's customers. The results will be exposed in section ***3. PowerBI Visualization Results***.

2.4. Analytic-Based Table - ABT

Recurring to SAS Guide (***Annex 2 – Building ABT Code***), new features were created in order to build a customer-signature table, where each observation represents a customer. **Table 2** contains all the variables on the analytical-based table and their respective description.

Table 2 – Customer-Signature Table Variables Description	
Variable	**Description**
CustomerNo	Customer ID
Age	Customer's age (integer value)
Gender	Customer's gender (M/F/O)
Nationality	Customer's nationality
Kids	Customer has children (1 = Yes; 0 = No)
Montlhy_Income	Customer's monthly income
Freq_*<product category>*	Number of transactions per *product category*
Mon_*<product category>*	Amount spent per *product category*
Date_First_Purchase	Date of customer's first transaction
Date_Last_Purchase	Date of customer's last transaction
Time_Since_Fisrt_Purchase	Number of days between first transaction and Dec 31st 2019
Time_Since_Last_Purchase	Number of days between last transaction and Dec 31st 2019
Favourite_Weekday	Customer's favourite weekday to shop
Favourite_Month	Customer's favourite month to shop
Total_Nr_Purchases	Customer's total number of transactions
Pct_Paypal	Proportion of customer's transactions paid via PayPal
Pct_Credit Card	Proportion of customer's transactions paid with credit card
Pct_Cash	Proportion of customer's transactions paid in cash
Pct_Store	Proportion of customer's transactions made in-store
Pct_Online	Proportion of customer's transactions made online
Total_Amt_Spent	Customer's total amount spent
Largest_Amt_Spent	Customer's highest transaction
Avg_Amt_Spent	Customer's average amount spent
Smallest_Amt_Spent	Customer's smallest transaction
Rate_of_Income	Proportion of customer's monthly income spent
Category	Customer's category (Gold/Silver/Bronze)

Notes:

- Freq_Miscellaneous + Freq_Office supplies + Freq_Candles & Lights + Freq_Decorative items + Freq_Kitchenware + Freq_Entryway items + Freq_Socks + Freq_Beauty & Accessories + Freq_Sombrero = Total_Nr_Purchases
- Mon_Miscellaneous + Mon_Office supplies + Mon_Candles & Lights + Mon_Decorative items + Mon_Kitchenware + Mon_Entryway items + Mon_Socks + Mon_Beauty & Accessories + Mon_Sombrero = Total_Amt_Spent
- Pct_Paypal + Pct_Credit Card + Pct_Store = 100%
- Pct_Store + Pct_Online = 100%

Considering what was mentioned in section ***2.1.5. Impute Node***, it is worth clarifying that *IMP_Age* values were converted to integer values in order to create ABT's variable *Age*.

Regarding the frequency per product category, a table with a count of the number of transactions per each product category purchased by each client is firstly produced. When transposed, it gives a count of the number of transactions for all product categories for each customer. In cases when a customer did not purchase a certain category, it is necessary to fill its frequency with the value 0. Following this line of thinking, the same happens with the monetary value per product category. Moreover, the same reasoning was applied to get each client's proportion of transactions by payment method and by channel.

Given that the dataset provided by Mega Market encompasses transactions made between August 2019 and December 2019, the number of days since the first and last transactions was measured with the last day of 2019 as a reference.

In order to find each customer's favourite weekday to shop, it was necessary to extract the weekdays from the transactions' dates. Afterwards, by calculating the number of transactions per weekday, it was possible to get the most frequent weekday for each customer. In case a client had more than one favourite weekday it was decided to consider it as "NoneExistent". The same reasoning was applied to figure out each customer's favourite month to shop.

Finally, the division of customers into categories was made with variable *Avg_Amt_Spent* as reference. Computing its quartiles, a "Gold" customer was identified as the one who spends an average amount per transaction higher than the mean between its Median and 3rd Quartile; whereas a "Bronze" spends, on average, a smaller amount than the mean between its 1st Quartile and Median.

With the final analytic-based table (*abt_final*) exported to Excel format, it is pertinent to analyse its variables' correlations since this is the dataset to be used for customer segmentation (clustering). Although the decision on which course to pursue relying on the next team, and as shown in **Appendix Fig. 15**, there are a lot of highly correlated variables – as expected. For instance, some of the most positively correlated variables are *Total_Nr_Purchases* and *Freq_Miscellaneous*, *Total_Amt_Spent* and *Mon_Miscellaneous*, and *Freq_Sombrero* and *Mon_Sombrero*.

3. POWERBI VISUALIZATION RESULTS

The exploration analysis performed on *PowerBI* was divided into three major parts: business overview, sales analysis and customers insights. This section exposes the main insights and explains some of the interactivity tools so that Mega Market can make the best use of them.

3.1. Business Overview Dashboard

This first dashboard – with a static representation in **Fig. 14** – is intended for one to know Mega Market's business reality.

Meeting some of the insights already mentioned on this document in a more easily readable way, the measures that we included at the top of the page provide the following general information:

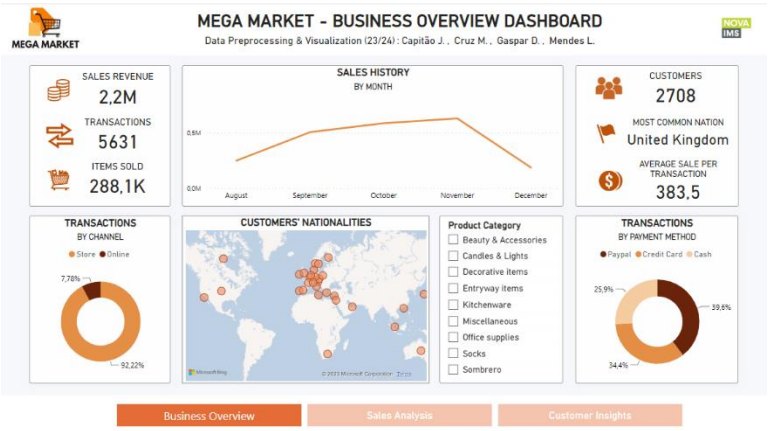

***Figure 14** – Business Overview Dashboard*

a) Mega Market had a total sales revenue of 2.2 million monetary units, as a result of 5631 transactions with a total of 288,1 thousand items being sold;
b) During the time period considered, the company had a total of 2708 customers, most of them from the United Kingdom;
c) The average sale per transaction corresponds to 383.5 monetary units.

Focusing on the *Sales Distributions by Month*, one can state that sales have increased from August to December in a notably linear trajectory.

As expected, the great majority of the transactions were made at the physical store and a small minority were made online. The distribution of transactions across the three available payment methods is comparable, with a slightly higher prevalence observed in PayPal transactions.

From the map, it is possible to understand that Mega Market has reached people from numerous countries around Europe, while also having a connection with people from 5 different continents.

The slicer tool allows to filter the visualizations by product category enabling, for instance, to check which nationality is more prone to buy a certain category of products. It is also possible to verify that all categories had their peak sales in November, except for 'Beauty and Accessories', 'Entryway items' and 'Socks'. Furthermore, most categories reach people from all around the world. However, both 'Socks' and 'Sombrero' categories only reach one additional nationality apart from British.

3.2. Sales Analysis Dashboard

The dashboard statically represented in **Fig. 15** allows to draw conclusions on how Mega Market sales evolve according to other criteria.

Overall, the greatest sales' value per transaction was 3885 monetary units, the average one almost 384 monetary units and the smallest one of 1. The 5 most sold items are presented in descending order in the matrix. The most sold product was 'Paper Chain Kit 50'S Christmas'.

The bar chart describes the sales for each product category. 'Miscellaneous' items completely disperse from the others, summing 1.885 million in sales. In second place, 'Decorative Items' - although with much less sales than the previous - also has disproportionally more sales than the others. Hence, for a better visualization of the proportion of the remaining categories in regard to the total sales, the button "Exclude TOP 2 Categories" should be clicked on. Furthermore, by hovering in each bar, the 3 best-selling items for each category are shown.

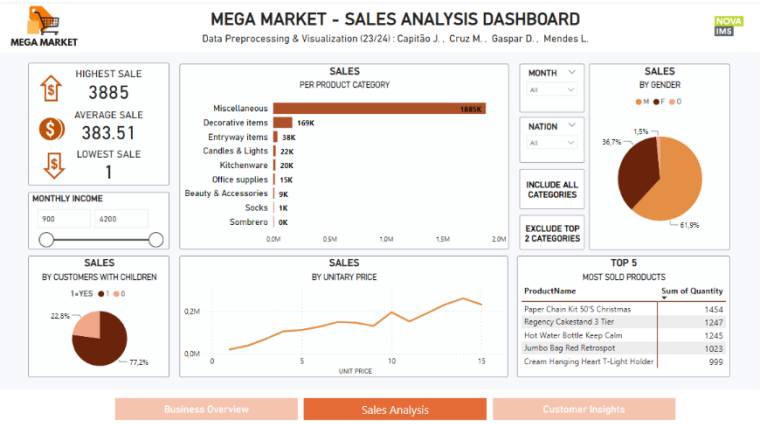

Figure 15 *– Sales Analysis Dashboard*

Through the *Sales by Unit_Price* line chart, a positive correlation is identified – meaning that the products from which Mega Market has the greater revenue are the most expensive ones (alternatively, the items which translate to less sales are the cheapest ones).

An analysis of the *Gender* pie chart reveals that most of the company's sales are derived from male customers (61.9%), 36.7% from female customers and 1.5% from customers that identify themselves as *Other*. Apart from that, one can conclude that 77.2% of the sales resulted from customers who have at least one child.

Utilizing the slicers allows to filter the month of transactions, customer's nationality and customer's monthly income, enabling a deeper understanding of the data by examining the filtered plots. An interesting insight is that the customers with a lower monthly income do more expensive transactions (by filtering the monthly income to smaller values, the average sale per transaction is greater than the average amount spent by clients with higher monthly income values). In addition, the 'Miscellaneous' category is the one with the most sales, regardless of the month or the customer's nationality.

3.3. Customer Insights Dashboard

The third and final dashboard (**Fig. 16**) combines variables from both *transactional_table* and *abt_final* datasets in order to contribute to the analysis of Mega Market's customers' shopping behaviour.

From the 2708 total customers, 1712 are male, 953 are female and 43 identify as *Other*. This information might be discerned by filtering through the *Gender* slicer. The average age corresponds to 55 years old and the average monthly income is 2600 monetary units.

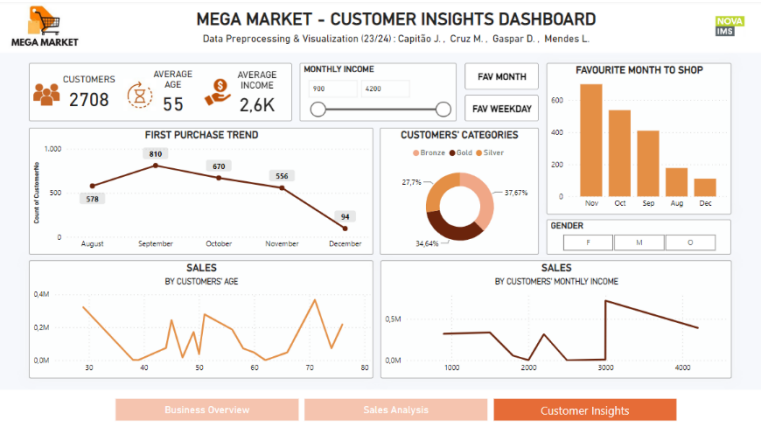

Figure 16 *– Customer Insights Dashboard*

Moreover, male customers are older (*Age* >= 38) and have a higher monthly income (*Monthly_Income* >= 2200). In opposition, female customers have a wider age range, including younger customers (29 <= *Age* <= 76). Besides, their monthly income is much lower (900 <= *Monthly_Income* <= 2200).

From the *First Purchase Trend* line plot, it is evident that the number of new customers is decreasing over time. The month with the highest number of new customers was September (810 new customers).

The *Sales by Customer's Monthly Income* line plot shows that Mega Market's customer's monthly income ranges from 900 to 4200. The ones that generate the most revenue to the company have monthly income greater than 3000 (having its peak at exactly 3000).

Clients prefer to shop at the weekend, especially on Sundays. The least favourite day to shop is on Wednesdays. By clicking on the button 'FAV MONTH', the bar chart changes to show customer's favourite month to shop. As expected, it corresponds to November. It is worth noting that these values concern solely to five months of a single year.

Finally, and as seen before, the customer division into categories based on their average amount spend on the shop is fairly even across all three categories, with the highest concentration observed in the "Silver" category.

4. CONCLUSIONS

Posterior to this analysis, it is not possible to identify a straightforward pattern regarding customer's age and how much they spend. The oldest customer is identified with 76 years old and the youngest with 29 years old, meaning that Mega Market has not yet reached the younger generations.

Going deeper on customers' shopping behaviour, the fact that Tuesdays are absent from the *Favourite Weekday* bar chart could potentially be attributed to the store being closed on that particular day.

Furthermore, the decrease in sales revenue from November to December might be explained by the fact the dataset only contains data until Dec 9th. In addition, the records from August only start at Aug 8th which can clarify the more accentuated increase from August to September (since in August has less days with records).

5. LIMITATIONS AND SUGGESTIONS

This final section introduces limiting circumstances to the analysis at hand as well as suggestions for both the company and the next team of consultants.

5.1. Limitations

It is important to mention that the sample studied is not representative of all Mega Market's customers. The data can be generalized; however, different customers can perform different actions than the corresponding to the data made available.

Not only that, but the dataset provided comprised transactions performed in a very restricted time slot (only transactions that took place in the last five months of 2019 were represented).

5.2. Suggestions for MEGA MARKET

It is recommended to continue to invest in business monitorization and looking for deeper insights that would eventually lead to more generalizable conclusions and Mega Market's growth.

5.3. Suggestions for future work

As further work, it is suggest using more time-wise diversified data to figure out whether the current business' information changes. Another interesting possibility is to perform an analysis targeting the outlier observations.

To the next team of consultants to perform customer segmentation, it is advised to elaborate on the variable correlation analysis of the *abt_final* features.

APPENDIX[2]

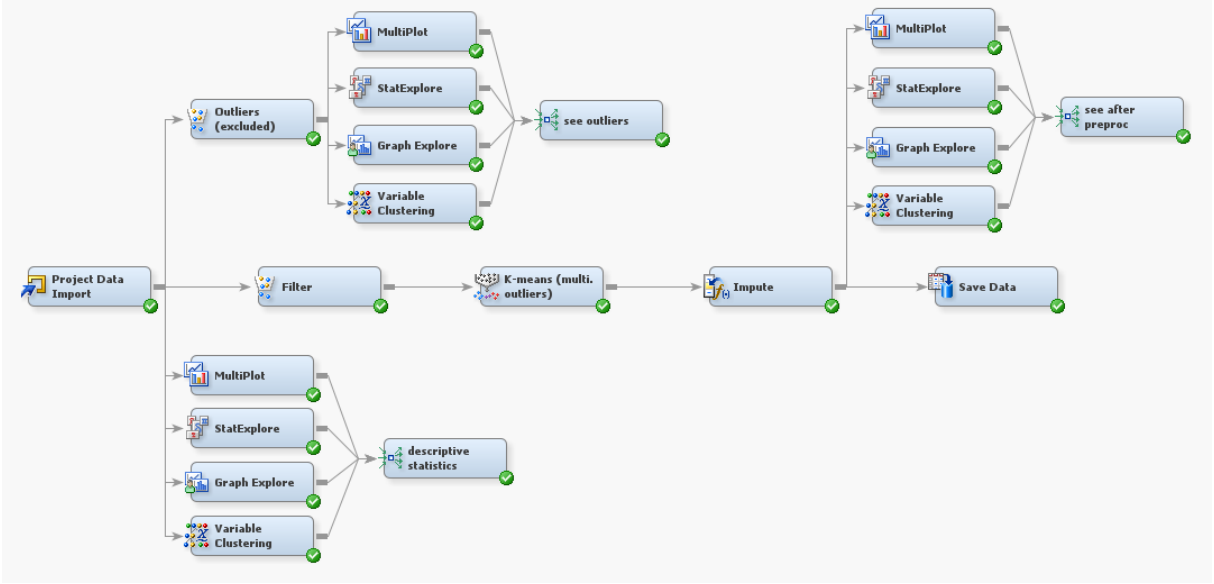

Figure 1 – SAS Enterprise Miner Diagram

Ordered Inputs	Data Role	Variable	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label	Abs C.V.	Coefficient of Variation	Sign
1	TRAIN	Total_payed	18	0	99999	1	77	22.91824	17.8784	0.981472	0.184304	INPUT	Total_payed	0.780095	0.780095	+
2	TRAIN	Quantity	3	0	99999	1	100	3.0133	1.743847	18.90988	1050.286	INPUT	Quantity	0.578717	0.578717	+
3	TRAIN	Monthly_Income	3000	2266	97733	900	90000	2506.403	1443.184	15.90465	582.5151	INPUT	Monthly Income	0.575799	0.575799	+
4	TRAIN	Unit_Price	7	0	99999	1	15	7.628366	4.291977	0.155836	-1.1938	INPUT	Unit Price	0.562634	0.562634	+
5	TRAIN	Age	51	2266	97733	29	299	54.36123	15.11825	-0.09631	-0.28342	INPUT	Age	0.278107	0.278107	+

Figure 2 – Interval Variables Summary Statistics, *StatExplore*

Data Role	Variable Name	Level	CODE	Frequency Count	Type	Percent	Level Index	Role	Label	Plot
TRAIN	Channel	Store	0	91723	C	91.72392	2	INPUT	Channel	1
TRAIN	Channel	Online	1	8276	C	8.276083	1	INPUT	Channel	1
TRAIN	Gender	M	1	61954	C	61.95462	2	INPUT	Gender	1
TRAIN	Gender	F	0	36806	C	36.80637	1	INPUT	Gender	1
TRAIN	Gender	O	2	1439	C	1.439014	3	INPUT	Gender	1
TRAIN	Kids	1	1	75148	N	75.14875	3	INPUT	Kids	1
TRAIN	Kids	0	0	22585	N	22.58523	2	INPUT	Kids	1
TRAIN	Kids		2	2266	N	2.266023	1	INPUT	Kids	1
TRAIN	Nationality	United Kingdom	0	90810	C	90.81091	30	INPUT	Nationality	1
TRAIN	Nationality	Germany	1	2420	C	2.420024	12	INPUT	Nationality	1
TRAIN	Nationality	France	2	2108	C	2.108021	11	INPUT	Nationality	1
TRAIN	Nationality	Belgium	4	675	C	0.675007	3	INPUT	Nationality	1
TRAIN	Nationality	Spain	5	556	C	0.556006	25	INPUT	Nationality	1
TRAIN	Nationality	Switzerland	15	483	C	0.483005	27	INPUT	Nationality	1
TRAIN	Nationality	Portugal	6	386	C	0.386004	22	INPUT	Nationality	1
TRAIN	Nationality	Norway	3	382	C	0.382004	20	INPUT	Nationality	1
TRAIN	Nationality	Italy	7	282	C	0.282003	16	INPUT	Nationality	1
TRAIN	Nationality	EIRE	13	231	C	0.231002	9	INPUT	Nationality	1
TRAIN	Nationality	Cyprus	18	228	C	0.228002	6	INPUT	Nationality	1
TRAIN	Nationality	Finland	12	190	C	0.190002	10	INPUT	Nationality	1
TRAIN	Nationality	USA	10	162	C	0.162002	28	INPUT	Nationality	1
TRAIN	Nationality	Israel	27	147	C	0.147001	15	INPUT	Nationality	1
TRAIN	Nationality	Australia	16	133	C	0.133001	1	INPUT	Nationality	1
TRAIN	Nationality	Austria	20	132	C	0.132001	2	INPUT	Nationality	1
TRAIN	Nationality	Channel Islands	9	87	C	0.087001	5	INPUT	Nationality	1
TRAIN	Nationality	Unspecified	28	79	C	0.079001	31	INPUT	Nationality	1
TRAIN	Nationality	Denmark	11	75	C	0.075001	8	INPUT	Nationality	1
TRAIN	Nationality	Sweden	17	69	C	0.069001	26	INPUT	Nationality	1
TRAIN	Nationality	Malta	19	59	C	0.059001	18	INPUT	Nationality	1
TRAIN	Nationality	RSA	25	57	C	0.057001	23	INPUT	Nationality	1
TRAIN	Nationality	Hong Kong	23	48	C	0.048	14	INPUT	Nationality	1
TRAIN	Nationality	Netherlands	14	47	C	0.047	19	INPUT	Nationality	1
TRAIN	Nationality	Poland	21	46	C	0.046	21	INPUT	Nationality	1
TRAIN	Nationality	Greece	8	35	C	0.035	13	INPUT	Nationality	1
TRAIN	Nationality	United Arab Emirates	29	29	C	0.029	29	INPUT	Nationality	1
TRAIN	Nationality	Singapore	24	26	C	0.026	24	INPUT	Nationality	1
TRAIN	Nationality	Japan	22	12	C	0.012	17	INPUT	Nationality	1
TRAIN	Nationality	Canada	30	4	C	0.004	4	INPUT	Nationality	1
TRAIN	Nationality	Czech Republic	26	1	C	0.001	7	INPUT	Nationality	1
TRAIN	Payment	Paypal	2	40065	C	40.0654	3	INPUT	Payment	1
TRAIN	Payment	Credit Card	1	34981	C	34.98135	2	INPUT	Paymnent	1
TRAIN	Payment	Cash	0	24953	C	24.95325	1	INPUT	Payment	1
TRAIN	Product_Category_Name	Miscellaneous	2	87116	C	87.11687	6	INPUT	Product Category Name	1
TRAIN	Product_Category_Name	Decorative items	0	8319	C	8.319083	3	INPUT	Product Category Name	1
TRAIN	Product_Category_Name	Entryway items	3	1696	C	1.696017	4	INPUT	Product Category Name	1
TRAIN	Product_Category_Name	Candles & Lights	5	1009	C	1.00901	2	INPUT	Product Category Name	1
TRAIN	Product_Category_Name	Kitchenware	1	772	C	0.772008	5	INPUT	Product Category Name	1
TRAIN	Product_Category_Name	Office supplies	4	620	C	0.620006	7	INPUT	Product Category Name	1
TRAIN	Product_Category_Name	Beauty & Accessories	6	375	C	0.375004	1	INPUT	Product Category Name	1
TRAIN	Product_Category_Name	Socks	8	82	C	0.082001	8	INPUT	Product Category Name	1
TRAIN	Product_Category_Name	Sombrero	7	10	C	0.01	9	INPUT	Product Category Name	1
TRAIN	Reviews		2	57701	N	57.70158	1	INPUT	Reviews	1
TRAIN	Reviews	0	1	21255	N	21.25521	2	INPUT	Reviews	1
TRAIN	Reviews	1	0	21043	N	21.04321	3	INPUT	Reviews	1

Figure 3 – Class Variables Summary Statistics (*Product_Name* was omitted given the wide variety of product sold), *StatExplore*

[2] The reader is kindly asked to zoom in when any of the appendix figures presents itself illegible with the default zoom percentage.

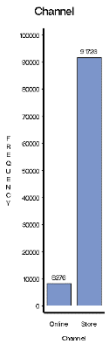

Figure 4 – *Channel* Histogram, *MultiPlot*

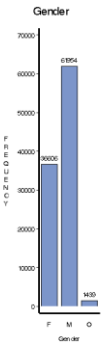

Figure 5 – *Gender* Histogram, *MultiPlot*

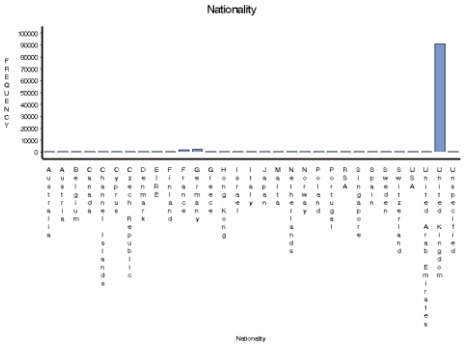

Figure 6 – *Nationality* Histogram, *MultiPlot*

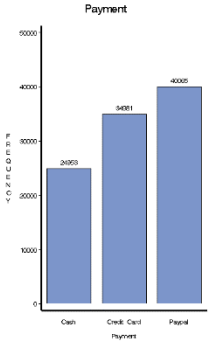

Figure 7 – *Payment* Histogram, *MultiPlot*

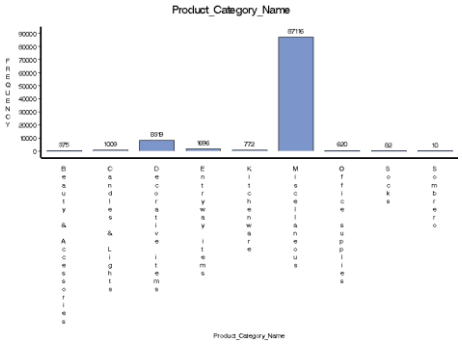

Figure 8 – *Product_Category_Name* Histogram, *MultiPlot*

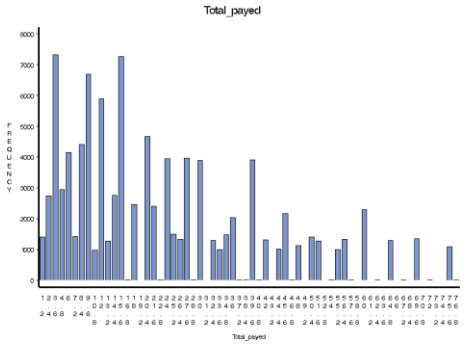

Figure 9 – *Total_payed* Histogram, *MultiPlot*

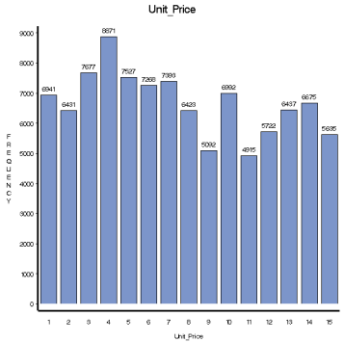

Figure 10 – *Unit_Price* Histogram, *MultiPlot*

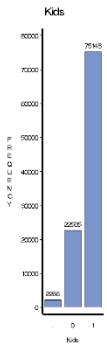

Figure 11 – *Kids* Histogram, *MultiPlot*

TransactionNo	Date	ProductID	ProductName	Quantity	Total_payed	CustomerNo	Nationality	Gender	Age	Kids	Reviews	Payment	Channel	Product_Category_ID	Product_Category_Name	Unit_Price	Monthly_Income
580173	12/02/2019	22699	Roses Regency Teacup And Sau...	100	9	18282	United Kingdom	O	29	0	0	Cash	Store	2	Miscellaneous	3	900
580173	12/02/2019	23174	Regency Sugar Bowl Green	5	40	18282	United Kingdom	O	299	0	1	Cash	Store	2	Miscellaneous	8	900
578849	11/27/2019	23579	Snack Tray I Love London	4	36	13232	United Kingdom	M	51	1	1	Cash	Store	2	Miscellaneous	9	50000
578849	11/27/2019	20914	Set/5 Red Retrospot Lid Glass Bo...	3	3	13232	United Kingdom	M	51	1	1	Cash	Store	2	Miscellaneous	1	50000
578849	11/27/2019	21080	Set/20 Red Retrospot Paper Napk...	4	44	13232	United Kingdom	M	51	1	1	Cash	Store	2	Miscellaneous	11	50000
578849	11/27/2019	21210	Set Of 72 Retrospot Paper Doilies	5	30	13232	United Kingdom	M	51	1	1	Cash	Store	2	Miscellaneous	6	50000
578849	11/27/2019	22666	Recipe Box Pantry Yellow Design	4	16	13232	United Kingdom	M	51	1	1	Cash	Store	2	Miscellaneous	4	50000
578849	11/27/2019	23503	Playing Cards Keep Calm & Carry...	4	20	13232	United Kingdom	M	51	1	1	Cash	Store	2	Miscellaneous	5	50000
578849	11/27/2019	23505	Playing Cards I Love London	4	48	13232	United Kingdom	M	51	1	1	Cash	Store	2	Miscellaneous	12	50000
578849	11/27/2019	22910	Paper Chain Kit Vintage Christmas	3	18	13232	United Kingdom	M	51	1	1	Cash	Store	2	Miscellaneous	6	50000
578849	11/27/2019	22086	Paper Chain Kit 50'S Christmas	2	8	13232	United Kingdom	M	51	1	1	Cash	Store	2	Miscellaneous	4	50000
576025	11/13/2019	22722	Set Of 6 Spice Tins Pantry Design	3	39	13232	United Kingdom	M	51	1	1	Credit Card	Online	2	Miscellaneous	13	50000
576025	11/13/2019	23350	Roll Wrap Vintage Spot	3	15	13232	United Kingdom	M	51	1	1	Credit Card	Online	2	Miscellaneous	5	50000
576025	11/13/2019	23349	Roll Wrap Vintage Christmas	5	70	13232	United Kingdom	M	51	1	1	Credit Card	Online	2	Miscellaneous	14	50000
576025	11/13/2019	21563	Red Heart Shape Love Bucket	5	60	13232	United Kingdom	M	51	1	1	Credit Card	Online	2	Miscellaneous	12	50000
576025	11/13/2019	20655	Queen Of Skies Luggage Tag	1	8	13232	United Kingdom	M	51	1	1	Credit Card	Online	2	Miscellaneous	8	50000
576025	11/13/2019	22203	Milk Pan Red Retrospot	1	9	13232	United Kingdom	M	51	1	1	Credit Card	Online	2	Miscellaneous	9	50000
576025	11/13/2019	22970	London Bus Coffee Mug	1	6	13232	United Kingdom	M	51	1	1	Credit Card	Online	2	Miscellaneous	6	50000
576025	11/13/2019	23344	Jumbo Bag 50'S Christmas	1	10	13232	United Kingdom	M	51	1	1	Credit Card	Online	2	Miscellaneous	10	50000
576025	11/13/2019	22441	Grow Your Own Basil In Enamel ...	4	32	13232	United Kingdom	M	51	1	1	Credit Card	Online	2	Miscellaneous	8	50000
576025	11/13/2019	21261	Green Goose Feather Christmas ...	3	15	13232	United Kingdom	M	51	1	1	Credit Card	Online	2	Miscellaneous	5	50000
576025	11/13/2019	21706	Folding Umbrella Red/White Polka...	4	36	13232	United Kingdom	M	51	1	1	Credit Card	Online	2	Miscellaneous	9	50000
572990	10/27/2019	22865	Hand Warmer Owl Design	4	16	18276	United Kingdom	O	29	0	1	Credit Card	Store	2	Miscellaneous	4	35000
570715	10/12/2019	23272	Tree T-Light Holder Willie Winkie	100	4	18287	United Kingdom	O	29	0	0	Paypal	Store	2	Miscellaneous	4	90000
570715	10/12/2019	23274	Star T-Light Holder Willie Winkie	100	4	18287	United Kingdom	O	29	0	0	Paypal	Store	2	Miscellaneous	4	900
570715	10/12/2019	23264	Set Of 3 Wooden Sleigh Decorati...	100	60	18287	United Kingdom	O	29	0	1	Paypal	Store	2	Miscellaneous	15	900
570715	10/12/2019	22419	Lipstick Pen Red	100	10	18287	United Kingdom	O	29	0	1	Paypal	Store	2	Miscellaneous	2	900
570715	10/12/2019	22421	Lipstick Pen Fuschia	100	32	18287	United Kingdom	O	29	0	0	Paypal	Store	2	Miscellaneous	8	900
570715	10/12/2019	22114	Hot Water Bottle Tea And Sympat...	100	10	18287	United Kingdom	O	29	0	1	Paypal	Store	2	Miscellaneous	2	900
570715	10/12/2019	22866	Hand Warmer Scotty Dog Design ...	100	20	18287	United Kingdom	O	29	0	1	Paypal	Store	2	Miscellaneous	5	900
570715	10/12/2019	22865	Hand Warmer Owl Design	100	20	18287	United Kingdom	O	29	0	1	Paypal	Store	2	Miscellaneous	4	900
570715	10/12/2019	21481	Fawn Blue Hot Water Bottle	100	30	18287	United Kingdom	O	29	0	0	Paypal	Store	2	Miscellaneous	15	900
570715	10/12/2019	22144	Christmas Craft Little Friends	100	15	18287	United Kingdom	O	29	0	0	Paypal	Store	2	Miscellaneous	15	900
565324	09/02/2019	21210	Set Of 72 Retrospot Paper Doilies	5	30	13232	United Kingdom	M	51	1	1	Credit Card	Store	2	Miscellaneous	6	50000
565324	09/02/2019	22909	Set Of 20 Vintage Christmas Napk...	2	12	13232	United Kingdom	M	51	1	1	Credit Card	Store	2	Miscellaneous	6	50000
565324	09/02/2019	21216	Set 3 Retrospot Tea/Coffee/Sugar	3	12	13232	United Kingdom	M	51	1	1	Credit Card	Store	2	Miscellaneous	4	50000
565324	09/02/2019	23350	Roll Wrap Vintage Spot	2	10	13232	United Kingdom	M	51	1	1	Credit Card	Store	2	Miscellaneous	5	50000
565324	09/02/2019	23349	Roll Wrap Vintage Christmas	4	56	13232	United Kingdom	M	51	1	1	Credit Card	Store	2	Miscellaneous	14	50000
565324	09/02/2019	22193	Red Diner Wall Clock	5	65	13232	United Kingdom	M	51	1	1	Credit Card	Store	2	Miscellaneous	13	50000
565324	09/02/2019	23341	Pink Diner Wall Clock	1	8	13232	United Kingdom	M	51	1	1	Credit Card	Store	2	Miscellaneous	8	50000
565324	09/02/2019	22910	Paper Chain Kit Vintage Christmas	3	18	13232	United Kingdom	M	51	1	1	Credit Card	Store	2	Miscellaneous	6	50000
565324	09/02/2019	22910	Paper Chain Kit Vintage Christmas	3	18	13232	United Kingdom	M	51	1	1	Credit Card	Store	2	Miscellaneous	6	50000
565324	09/02/2019	22086	Paper Chain Kit 50'S Christmas	5	20	13232	United Kingdom	M	51	1	1	Credit Card	Store	2	Miscellaneous	4	50000
565324	09/02/2019	22086	Paper Chain Kit 50'S Christmas	2	8	13232	United Kingdom	M	51	1	1	Credit Card	Store	2	Miscellaneous	4	50000
565324	09/02/2019	23198	Pantry Magnetic Shopping List	4	12	13232	United Kingdom	M	51	1	1	Credit Card	Store	2	Miscellaneous	3	50000
565324	09/02/2019	20725	Lunch Bag Red Retrospot	1	3	13232	United Kingdom	M	51	1	1	Credit Card	Store	2	Miscellaneous	3	50000
565324	09/02/2019	23343	Jumbo Bag Vintage Christmas	5	35	13232	United Kingdom	M	51	1	1	Credit Card	Store	2	Miscellaneous	7	50000
565324	09/02/2019	22961	Jam Making Set Printed	1	8	13232	United Kingdom	M	51	1	1	Credit Card	Store	2	Miscellaneous	8	50000
565324	09/02/2019	22131	Food Container Set 3 Love Heart	2	4	13232	United Kingdom	M	51	1	1	Credit Card	Store	2	Miscellaneous	2	50000
565324	09/02/2019	23338	Egg Frying Pan Red	2	14	13232	United Kingdom	M	51	1	1	Credit Card	Store	2	Miscellaneous	7	50000

Figure 12 – Excluded observations

Segment Id	Frequency of Cluster
1	13415
2	2884
3	7308
4	5381
5	16285
6	10308
7	12694
8	17310
9	14364

Figure 13 – *K-means* Node Results (9 seeds)

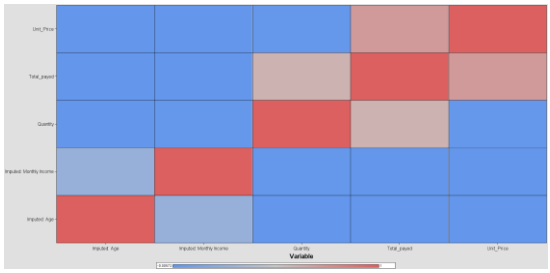

Figure 14 – Variable Correlation, *Variable Clustering* Node (after imputation of missing values)

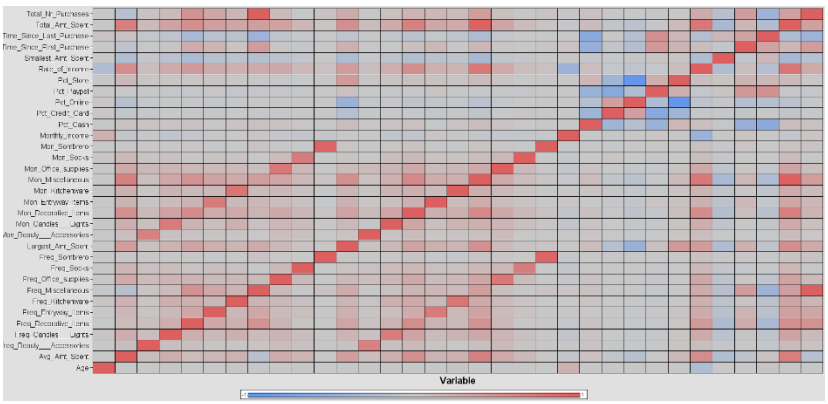

Figure 15 – Variable Correlation, *Variable Clustering* Node (referent to *abt_final* dataset)

ANNEX 1 – COHERENCE CHECKING CODE

```
DATA transactional_table;
SET work.SAS_exported;
/* 99949 rows */

/* exclude observations that have _SEGMENT_ = 2 (multidimensional outliers) */
if _SEGMENT_ = '2' then do;
delete;
end;
/* 97065 rows */

/* when a purchase is made online, it's not possible to pay with physical cash */
if (Payment = 'Cash') and (Channel ='Online') then do;
delete;
end;
/* 96976 rows */

/* verifications just to make sure (for future data) */
if (IMP_Age<18) then do;
delete;
end;

if (Quantity<0) then do;
delete;
end;

if (Unit_Price<0) then do;
delete;
end;

if (Total_payed<0) then do;
delete;
end;

if (IMP_Kids ne 0) and (IMP_Kids ne 1) then do;
delete;
end;

if (Gender ne 'M') and (Gender ne 'F') and (Gender ne 'O') then do;
delete;
end;

/*if (Payment = 'Paypal') and (Channel ='Store') then do;
delete;
end;
the observations that result from this query will not be considered
as an incongruity given being too many */

/* correct the inconsistency associated to Total_payed column */
if (Unit_Price*Quantity ne Total_payed) then do;
Total_payed = Unit_Price*Quantity;
end;
```

```
/* drop the _SEGMENT_ and _WARN_ columns (from SAS Miner) */
PROC SQL;
ALTER TABLE transactional_table
DROP _WARN_, _SEGMENT_;

/* check if there are customers with different information depending on the transaction */
PROC SQL;
CREATE TABLE different_info AS
SELECT CustomerNo,
        count(distinct IMP_Age) as n_unique_age,
        count(distinct Nationality) as n_unique_nationality,
        count(distinct Gender) as n_unique_gender,
        count(distinct IMP_Kids) as n_unique_kids,
        count(distinct IMP_Monthly_Income) as n_unique_mon_inc
FROM transactional_table
GROUP BY CustomerNo
HAVING n_unique_age > 1
        or n_unique_nationality > 1
        or n_unique_gender > 1
        or n_unique_kids > 1
        or n_unique_mon_inc > 1;
RUN;

/* delete rows from transactional_table where customers have different information */
PROC SQL;
DELETE FROM transactional_table
WHERE CustomerNo IN (SELECT CustomerNo FROM different_info);
RUN;
/* 96853 rows */


/* export TRANSACTIONAL_TABLE (to upload in PowerBI) */
PROC EXPORT DATA=transactional_table
  OUTFILE='/home/u63618385/PROJECT/transactional_table.xlsx'
  DBMS=xlsx
  REPLACE;
  SHEET="Transactional_Table";
RUN;
```

ANNEX 2 – BUILDING ABT CODE

```
/* get the basics: age, gender, nationality, kids, income */
PROC SQL;
CREATE TABLE basics_abt AS
SELECT CustomerNo,
          round(min(IMP_Age)) as Age,
          min(Gender) as Gender,
          min(Nationality) as Nationality,
          min(IMP_Kids) as Kids,
          min(imp_monthly_income) as Monthly_Income
FROM transactional_table
GROUP BY CustomerNo;
RUN;


/* number of transactions per product category (frequency) */
PROC SQL;
CREATE TABLE frequency_table as
SELECT CustomerNo,  Product_Category_Name, count(distinct TransactionNo) as frequency
FROM transactional_table
GROUP BY CustomerNo, Product_Category_Name;
RUN;
/* lets sort... */
PROC SORT DATA=frequency_table;
        BY CustomerNo;
RUN;
/* ...and transpose the table */
PROC TRANSPOSE DATA=frequency_table
        OUT=frequency_abt
        PREFIX=Freq_;
        ID Product_Category_Name;
        BY CustomerNo;
RUN;


/* amount spent per product category (monetary) */
PROC SQL;
CREATE TABLE monetary_table as
SELECT CustomerNo, Product_Category_Name, sum(Total_payed) as monetary
FROM transactional_table
GROUP BY CustomerNo, Product_Category_Name;
run;
/* lets sort... */
PROC SORT DATA=monetary_table;
        BY CustomerNo;
RUN;
/* ...and transpose the table */
PROC TRANSPOSE DATA=monetary_table
        OUT=monetary_abt
   PREFIX=Mon_;
        ID Product_Category_Name;
        BY CustomerNo;
RUN;
```

```
/* merge into one single table */
DATA abt_1;
	MERGE basics_abt frequency_abt monetary_abt;
	BY CustomerNo;
	DROP _NAME_;
RUN;


/* fill the missing values as 0 (not actually missing) */
DATA abt_2;
SET abt_1;
ARRAY change _numeric_;
	DO OVER change;
		IF change=. THEN change=0;
	END;
RUN;


/* date of the first and last transactions */
PROC SQL;
CREATE TABLE first_last_dates AS
SELECT distinct CustomerNo,
	min(Date) as Date_First_Purchase,
	max(Date) as Date_Last_Purchase
FROM transactional_table
GROUP BY CustomerNo;
RUN;


/* days since first and last transactions */
DATA days_since;
SET first_last_dates;
/* 21914 is the number of days between 1 JAN 1960 to 31 DEC 2019 */
Time_Since_First_Purchase = 21914 - Date_First_Purchase;
Time_Since_Last_Purchase = 21914 - Date_Last_Purchase;
RUN;


/* format both 'Date_First_Purchase' and 'Date_Last_Purchase' */
DATA dates_abt;
SET days_since;
FORMAT Date_First_Purchase date9.;
FORMAT Date_Last_Purchase date9.;
RUN;


/* merge again into one single table */
DATA abt_3;
	MERGE abt_2 dates_abt ;
	BY CustomerNo;
RUN;
```

```
/* favourite weekday to shop */
/* 1) get the weekdays from the dates and number of transactions per weekday (there is no mode function)
*/
PROC SQL;
CREATE TABLE weekday_freq_table AS
SELECT CustomerNo,
        put(Date, dowName.) as weekday,
        count(distinct(TransactionNo)) as weekday_freq
FROM transactional_table
GROUP BY CustomerNo, weekday;
RUN;
/* 2) get the most frequent weekday for each customer */
PROC SQL;
CREATE TABLE weekday_most_freq AS
SELECT CustomerNo, weekday as most_freq_weekday
FROM weekday_freq_table
GROUP BY CustomerNo
HAVING weekday_freq = max(weekday_freq);
RUN;
/* 3) check if there are equally frequent 'most_freq_weekday' for each customer */
PROC SQL;
CREATE TABLE weekday_equal_freq AS
SELECT CustomerNo,
        most_freq_weekday as Favourite_Weekday,
        count(most_freq_weekday) as nr_weekday_equal_freq
FROM weekday_most_freq
GROUP BY CustomerNo;
RUN;
/* 4) if there are, then those customers won't have a favourite weekday to shop */
DATA weekday_fav_table;
SET weekday_equal_freq;
if (nr_weekday_equal_freq = 1) then Favourite_Weekday = Favourite_Weekday;
if (nr_weekday_equal_freq > 1) then Favourite_Weekday = 'NoneExistent';
RUN;
/* 5) finally get the favourite weekday for each customer */
PROC SQL;
CREATE TABLE weekday_fav_abt AS
SELECT distinct(CustomerNo), Favourite_Weekday
FROM weekday_fav_table
GROUP BY CustomerNo;
RUN;


/* favourite month to shop */
/* 1) get the months from the dates and number of transactions per month (there is no mode function) */
PROC SQL;
CREATE TABLE month_freq_table AS
SELECT CustomerNo,
        put(Date, monname3.) as month,
        count(distinct(TransactionNo)) as month_freq
FROM transactional_table
GROUP BY CustomerNo, month;
RUN;
/* 2) get the most frequent month for each customer */
PROC SQL;
CREATE TABLE month_most_freq AS
SELECT CustomerNo, month as most_freq_month
```

```
FROM month_freq_table
GROUP BY CustomerNo
HAVING month_freq = max(month_freq);
RUN;
/* 3) check if there are equally frequent 'most_freq_month' for each customer */
PROC SQL;
CREATE TABLE month_equal_freq AS
SELECT CustomerNo,
        most_freq_month as Favourite_Month,
        count(most_freq_month) as nr_month_equal_freq
FROM month_most_freq
GROUP BY CustomerNo;
RUN;
/* 4) if there are, then those customers won't have a favourite month to shop */
DATA month_fav_table;
SET month_equal_freq;
if (nr_month_equal_freq = 1) then Favourite_Month = Favourite_Month;
if (nr_month_equal_freq > 1) then Favourite_Month = 'NoneExistent';
RUN;
/* 5) finally get the favourite month for each customer */
PROC SQL;
CREATE TABLE month_fav_abt AS
SELECT distinct(CustomerNo), Favourite_Month
FROM month_fav_table
GROUP BY CustomerNo;
RUN;


/* merge again into one single table */
DATA abt_4;
      MERGE abt_3 weekday_fav_abt month_fav_abt;
      BY CustomerNo;
RUN;


/* total number of transactions */
PROC SQL;
CREATE TABLE total_purchases_abt AS
SELECT CustomerNo, count(distinct TransactionNo) as Total_Nr_Purchases
FROM transactional_table
GROUP BY CustomerNo;
RUN;


/* proportion of transactions by payment method */
PROC SQL;
CREATE TABLE payment_pct_table AS
SELECT t1.CustomerNo,
        t1.Payment,
        count(distinct t1.TransactionNo)*100/(SELECT count(distinct t2.TransactionNo)

       FROM transactional_table t2

       WHERE t1.CustomerNo = t2.CustomerNo) as payment_pct
FROM transactional_table t1
GROUP BY CustomerNo, Payment;
RUN;
```

```
/* lets sort... */
PROC SORT DATA=payment_pct_table;
	BY CustomerNo;
RUN;
/* ...and transpose the table */
PROC TRANSPOSE DATA=payment_pct_table
	OUT=payment_pct_abt
	PREFIX=Pct_;
	ID Payment;
	BY CustomerNo;
RUN;


/* proportion of transactions by channel */
PROC SQL;
CREATE TABLE channel_pct_table AS
SELECT t1.CustomerNo,
	 t1.Channel,
	 count(distinct t1.TransactionNo)*100/(SELECT count(distinct t2.TransactionNo)

	FROM transactional_table t2

	WHERE t1.CustomerNo = t2.CustomerNo) as channel_pct
FROM transactional_table t1
GROUP BY CustomerNo, Channel;
RUN;
/* lets sort... */
PROC SORT DATA=channel_pct_table;
	BY CustomerNo;
RUN;
/* ...and transpose the table */
PROC TRANSPOSE DATA=channel_pct_table
	OUT=channel_pct_abt
	PREFIX=Pct_;
	ID Channel;
	BY CustomerNo;
RUN;


/* merge again into one single table */
DATA abt_5;
	MERGE abt_4 total_purchases_abt payment_pct_abt channel_pct_abt;
	BY CustomerNo;
	DROP _NAME_;
RUN;


/* fill the missing values as 0 (they are not actually missing) */
DATA abt_6;
SET abt_5;
ARRAY change _numeric_;
	DO OVER change;
		IF change=. THEN change=0;
	END;
RUN;
```

```
/* total amount spent */
PROC SQL;
CREATE TABLE amt_spent_abt AS
SELECT CustomerNo, sum(Total_payed) as Total_Amt_Spent
FROM transactional_table
GROUP BY CustomerNo;
RUN;


/* largest transaction (highest amount spent) */
PROC SQL;
CREATE TABLE largest_amt_abt AS
SELECT CustomerNo, max(Total_payed) as Largest_Amt_Spent
FROM transactional_table
GROUP BY CustomerNo;
RUN;


/* merge again into one single table */
DATA abt_7;
	MERGE abt_6 amt_spent_abt largest_amt_abt;
	BY CustomerNo;
RUN;


/* average amount spent */
PROC SQL;
CREATE TABLE avg_amt_abt AS
SELECT CustomerNo, Total_Amt_Spent/Total_Nr_Purchases as Avg_Amt_Spent
FROM abt_7
RUN;


/* smallest transaction (smallest amount spent)  */
PROC SQL;
CREATE TABLE smallest_amt_abt AS
SELECT CustomerNo, min(Total_payed) as Smallest_Amt_Spent
FROM transactional_table
GROUP BY CustomerNo;
RUN;


/* merge again into one single table */
DATA abt_8;
	MERGE abt_7 avg_amt_abt smallest_amt_abt;
	BY CustomerNo;
RUN;


/* proportion of the customer's monthly income spent on the Mega Market */
PROC SQL;
CREATE TABLE rate_income_abt AS
SELECT CustomerNo, Monthly_Income, Total_Amt_Spent, (Total_Amt_Spent*100/Monthly_Income) as
Rate_of_Income
FROM abt_8
GROUP BY CustomerNo;
RUN;
```

```
/* merge again into one single table */
DATA abt_9;
	MERGE abt_8 rate_income_abt;
	BY CustomerNo;
RUN;


/* divide customers into categories (gold, silver, bronze) */
/* step 1) compute quartiles to know limits to define for customer avg amount spent */
PROC UNIVARIATE DATA=abt_8;
   VAR Avg_Amt_Spent;
   OUTPUT OUT=quartiles_Avg_Amt_Spent
   PCTLPTS = 25 50 75
   PCTLPRE = Q_;
RUN;
/* step 2) divide into groups */
PROC SQL;
CREATE TABLE categories_abt AS
SELECT CustomerNo, CASE
            WHEN Avg_Amt_Spent > 436  THEN 'Gold'
            WHEN Avg_Amt_Spent > 231  and Avg_Amt_Spent < 436 THEN 'Silver'
            ELSE 'Bronze'
            END AS Category
FROM abt_8;
RUN;
/* step 3) check number of customers in each category */
PROC SQL;
SELECT Category, count(*)
FROM categories_abt
GROUP BY Category;
RUN;

/* merge again to crate final ABT */
DATA abt_final;
	MERGE abt_9 categories_abt;
	BY CustomerNo;
RUN;


/* export ABT_FINAL */
PROC EXPORT DATA=abt_final
  OUTFILE='/home/u63618385/PROJECT/abt_final.xlsx'
  DBMS=xlsx
  REPLACE;
  SHEET="abt_final";
RUN;
```

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa