SAS and MSc Business Analytics

Athens University of Economics and Business

Academic Specialization in

SAS Programming and Machine Learning

Milestone Project

Deadline: 03/09/2023

A. Objective of the project

This Milestone Project is part of the required procedure for obtaining the SAS Academic Specialization in SAS Programming and Machine Learning.

The objective of the project is to apply techniques for accessing, processing, managing and mining of real world data and to provide solutions to business problems that today's organizations face through the use of Base SAS Programming, SAS Visual Analytics and SAS Visual Data Mining and Machine Learning on SAS Viya.

In order to accomplish the above objectives you are given a set of real world POS data that are related to sales of a retail company along with other related data that are presented and described in section E of this document.

You are asked to analyze the given data through the use of Base SAS, SAS VA and SAS VDMML and to write a relevant report (deliverable) to be handed to the management team of the organization by answering the question that follow. You are asked to analyze the given data through the use of SAS Viya and to write a relevant ***business*** report (deliverable) to be handed to the management team of the organization by answering the question that follow. The business format of the report means that, as it was explained in class, it should contain advice and propositions on ***how the organization should act*** to become more efficient and more effective in their operation, based on the descriptive and predictive output of the data management, reporting and analytics work done using the data provided.

It should be underlined that this is an individual project and the deadline for submission is 02/04/2023. The deliverable should be sent in pdf and docx format to Andreas.Zaras@gmail.com and to the secretariat of the program with title 'FirstName_LastName_SAS', where FirstName and LastName is the first and last name of each student respectively. In the first page of the report the credentials used to access SAS Viya software should also be included.

B. Base SAS Programming Using SAS Studio on SAS Viya

The following tasks require the use of Base SAS. Please take into account the following:

- The data sets should be transformed to SAS format with the use of the data step or through the Import facility (right click import data, beware to choose the correct delimiter for raw data files).
- Proc sql can be used only in answering questions where it is explicitly mentioned, where as in any other case it is obligatory to use only the data step or any other procedure except proc sql (e.g. proc means).
- In order to create the graphs you can use either SAS Studio or SAS Visual Analytics.

Attention: In order to avoid errors when transforming data sets to SAS format, read the variables that will not be used as numbers (e.g. SKU, BasketID) in string type.

1. Data pre – processing:

    - For every invoice calculate the number of SKU's that are related to it *'Invoice total items'*. Save the output in a new SAS data set and print the first 10 observations of it. Only the data step can be used for merging data sets. Proc sql can be used for the statistics. It is suggested to use noprint option in proc sql because the new data set will be large.

- For every invoice calculate the total value of the SKU's that are related to it *'Invoice total value'*. Beware that there exist price discounts that can be seen in the promotions data set. Take into account all the invoices no matter if they are Sales or Returns. Save the output in a new SAS data set. For this task use the proc means with the output statement.

- Divide the observations of the table 'Invoice' into two new tables where in the one the Sales transactions will be stored where as in the second the Returns transactions will be stored. This division must be done using the variable 'Operation'. This action since it is not stated differently should be completed using the data step.
- Calculate the customer's age based on the fact that today's date is 01/01/2019 and store it into a new variable (check the validity of the dates e.g. birth year greater than 1910 and less than 2001). Show integer values of the age without decimals.

2. Describe and explain using graphs who is your customer. What is the profile of the audience to which the company's products are targeted?

   - What are the demographic characteristics i.e. age, gender and region of the company's customers?

   - Based on the age variable, create a new variable entitled Age_Range that takes the following values:

     <18 -- > "Under 18"

     18 - 25 -- > "Very Young"

     26 - 35 -- > "Young"

     36 - 50 -- > "Middle Age"

     51 - 65 -- > "Mature"

     66 – 75 -- > "Old"

     >= 76 -- > "Very Old"

     (Attention: do not format the values of the existing variable but create a new variable entitled Age _Range).

   - What are the behavioral characteristics of each age group? (visits to the stores, number of distinct SKU's purchased, total cost of purchases). The merging of the data sets must be done using exclusively the data step but the calculation of the statistics e.g. visits, total cost of purchases etc can be done using proc sql. Create a pie chart and a frequency table with the percentages of customers that belong to each age group. Augment your analysis by providing pie charts for the behavioral characteristics for each age group.

3. Exploration and understanding of sales:
   - What was the level of Sales and Returns? Create a bar chart with the monetary values.
   - Create graphs for the average basket size i.e. number of SKU's, total monetary value, etc and comment on your findings.
   - Create a report that shows the top products per product line and product type with respect to sales value in descending order. Show also the subtotal sales of each product type.
   - Use graphs to show the contribution to the company's revenues of each region of the country.
   - For the top region found in the previous question show the contribution to the company's revenues per gender.

Proc sql can be used only for the calculation of the statistics e.g. of the average basket and not e.g. for merging data sets (for this data step should be used).

4. Zoom into the promotional activities by answering the following questions:
   - Use graphs to show what is the percentage of products that are sold without promotion and what is the percentage of products sold with promotion. Create a format to display the 0% promotion as "No Promotion" and the 10%, 20% and 30% as "Promotion".
   - Create pie charts to show the percentage of products that are sold on each promotion type (use the description of the promotion and not its code). Do not include the products sold without promotion.
   - What is the distribution of sales per day of the week? Is there any difference among the various days with respect to the number of distinct SKU's per invoice. In order to find the day of the week when the sale takes place use the weekday function.

5. It should be also mentioned that the SKU of each product contains “hidden” information. The ninth (9th) digit indicates the company that supplied the product (supplier). In order to unhide this piece of information use relevant functions and then store it to a new column. If we assume that an SKU is 58720443**4**50301, then the supplier code is 4.

- Create a frequency report and a relevant chart to show the percentage of products sold by each supplier (use the name of the supplier and not its code). Weight the frequency of the SKU by the quantity sold. This will show the supplier with the highest demand.
- Create graphs to show the percentage and actual revenues of products sold by each supplier (use the name of the supplier and not its code).
- Create a cross tabulation table to show the total revenue of the company with respect to the origins of the products sold by each supplier (Use the names of the suppliers and the names of the countries of origins and not their codes. Put the total revenue in the middle of the cross tabulation, the origin in the rows and the suppliers in the columns). For this task you have to use proc tabulate (find relevant instructions in the web or in sas help).

6. The company wants to profile its customers based on their importance so as to offer them personalized services and products. The customer segmentation is asked to be done based on the three parameters of the RFM model. Before the application of the RFM model the RFM data set should be created. It is reminded that the RFM model is based on the following three parameters:

Recency - How recently did the customer purchase?

Frequency - How often do they purchase?

MonetaryValue - How much do they spend?

For this task proc sql can be used. For the calculation of R, F, M the following functions will be useful: max, sum, count and intck (For the intck use the argument week and the argument 16/12/2011 for today’s date).

For the creation of the variable Monetary, the price, quantity and promotion variables should be used.

C. SAS Visual Data Mining and machine Learning (In some questions Base SAS Programming and SAS Visual Analytics should also be used)

7. Create the clusters by analyzing the RFM data set using SAS Visual Data Mining and machine Learning and the three parameters of the RFM model. Then export the RFM data set with the newly created cluster column in a library of SAS and then by using SAS Programming describe the demographic data of the two most important clusters created.

8. The company is interested to change internally the store based on the products that tend to be bought together. In order to apply this initiative the company must be sure about the associations among the product categories. You are asked to find which product categories are bought together (associations of product categories) in the whole data set. Then find the associations among product categories in the two most important clusters previously identified so if a customer is found to belong in one of them to receive the most suitable/ best proposals/ offers. For this task you used to filter the customers that belong to the two most important clusters, create the two relevant data sets and then these data sets to be analyzed using association rules through SAS Studio.

D. Instructions

- It is underlined that the answers to the above questions should be addressed to business people so they should be written accordingly to be understandable and aid in the decision making process.
- Charts and tables that document the answers should be included in the main deliverable.
- Screenshots of technical details about the software and about how it was used to produce the results should be included in the appendix of the report. The SAS code should also be included in the appendix.

E. Datasets description

The datasets consist of POS data from a retail store.

The available data are included in the following tables. The first one of them is related to data about customers and is entitled Customer, the second and the third are related to POS data and are entitled Invoice & Basket respectively, the fourth contains the coding of the payment method done and is entitled Payment_Method, the fifth contains the coding of the promotional activities running and is entitled Promotions, the sixth contains the coding of the suppliers and is entitled Suppliers and finally the seventh contains the coding of the product origin and is entitled Product_Origin.

Customer table

CustID	LastName	FirstName	Address	Country	Postal_Code	City	Region	Gender	Birth Info
201	Johnson	Stanley	…	Brazil	14409	Franca	SP	M	…
202	Cramer	Henry	…	Brazil	9790	Sao paulo	PR	F	..
203	Hoover	Terry	…	Brazil	1151	Pacaja	MG	M	..

This table is related to the data about the customers and contains the following columns:

- **CustomerID:** Customer ID, (unique for every customer)
- **LastName:** The surname of the customer
- **FirstName:** The first name of the customer
- **Address:** The street address and number of the customer
- **Country:** The country of origin of each customer
- **Postal_code:** The postal code of the customer
- **City:** The city where the customer resides.
- **Region:** The region where the customer resides
- **Gender:** The gender of the customer
- **Day_Of_Birth:** The day when the customer was born
- **Month_Of_Birth:** The month when the customer was born
- **Year_Of_Birth:** The year when the customer was born

Invoice table

InvoiceID	InvoiceNo	InvoiceDate	CustomerID	Payment_Method	Operation
125	536365	12/1/2010	250	2	Sale
126	536365	12/1/2010	1008	2	Sale
127	536365	12/1/2010	5	2	Return

This table contains data about the issued invoice (sale or return) and contains the following columns:

- **InvoiceID:** The ID of the invoice (unique for every invoice)
- **InvoiceNo:** The issue number of the invoice (unique for every invoice)
- **InvoiceDate:** The date when the invoice was issued
- **CustomerID:** Customer ID, (unique for every customer)
- **Payment_Method:** The code of the payment method
- **Operation:** Denotes whether the invoice is related to Sales or Return

We make the assumption that an invoice can be paid with more than one payment methods.

Basket table

InvoiceID	ProductID	PromotionID	Quantity
1	**32**	**1**	**2**
1	**126**	**1**	**1**
1	**120**	**1**	**2**

This table contains the following columns:

- **InvoiceID:** The ID of the invoice (unique for every invoice)
- **ProductID:** The ID of the product (unique for every product)
- **PromotionID:** The promotion id
- **Quantity:** The quantity of the product sold

Products table

Product ID	Product Line	Product Type	Product	SKU	Product Origin	Product Price
135	Camping Equipment	Cooking Gear	TrailChef Water Bag	29720443050301	1	17.34
136	Camping Equipment	Cooking Gear	TrailChef Canteen	58720443053456	2	29.45
137	Camping Equipment	Cooking Gear	TrailChef Kitchen Kit	68720443054908	3	35.67

This table contains the following columns:

- **ProductID:** The ID of the product (unique for every product)
- **ProductLine:** The upper level of the product hierarchy
- **ProductType:** The middle level of the product hierarchy
- **Product:** The name of the product (lowest level of the product hierarchy)
- **SKU:** The stock keeping unit of the product.
- **ProductOrigin:** The ID of the origin of the product
- **ProductPrice:** The price of the product

Promotions table

Promotion_ID	Promotion
1	0%
2	10% Off
3	20% Off
4	30% Off

This table contains the following columns:

- **Promotion_ID:** The ID of the promotion
- **Promotion:** The % discount on the product price

Product Origin table

Code	Country
1	US
2	China
3	Turkey
4	Spain
5	India

This table contains the following columns:

- **Country:** The country of origin of the product
- **Code:** The code of the country of origin of the product

Suppliers table

SupplierID	SupplierName
1	Dragon SA
2	Fabulo Ltd
3	Carper & Sons
4	Maestri & Maestri
5	Elegance SA

This table contains the following columns:

- **SupplierID:** The ID of the supplier (unique for every supplier)
- **SupplierName:** The name of the supplier