CO4759: Enterprise Data Management

Assignment 1

Date Issued: 29/10/2020
Hand in Date: 03/12/2020

IMPORTANT

- As work is submitted on-line, **the deadline is midnight on the hand in date**
- **Read the marking scheme carefully**
- **This is an individual project** and no group work is permitted
- This project is part of the required procedure for obtaining the **SAS Joint Certificate in Business Intelligence and Data Mining**
- **The assignment accounts for 40% towards your final grade.**

SAS Milestone Project Requirements

To be eligible for automatically passing the Enterprise Data Management part of the SAS Milestone project (see Course Handbook), the **student must achieve a grade of over 70%** in this assignment.

Learning Outcomes

- Prepare and Load Data
- Query Data
- Produce Reports

Assignment Purpose and Overview

The purpose of this assignment is to allow students to apply techniques for accessing, processing, managing and reporting of real-world data and to provide solutions to business problems that today's organizations face through the use of SAS Enterprise Guide. In order to accomplish the above objectives, you are provided with a set of real-world Point of Sale (POS) data that are related to the operations of a retail company RealPOS. You are asked to prepare and query the provided data through SAS Enterprise Guide and to create a comprehensive analysis report that will be presented to the management team of RealPOS.

In particular, in this assignment, you will:
- Import data to a data management system
- Pre-process the data to improve their quality
- Create and Manipulate SQL queries to retrieve data from multiple files
- Prepare basic descriptive statistics reports
- Prepare advanced analysis reports to gain insight about different aspects of the company

Requirements

RealPOS is a retail company based in Brazil that sells equipment and accessories for outdoor sports. The company has hired you to analyse their sales and customers and provide key insights that will be utilized to refine their marketing and sales strategies. To this end, the company has provided with a number of datasets (see Appendix Dataset Information) that were exported by their individual ERP systems.

Your required to import and process the data and later on prepare a number of reports to deliver key insights to the company. The individual tasks that you need to perform are outlined below. All tasks should be prepared in a single project but in individual process flows.

A. Data Pre-processing

1. **Import Raw Data**
   Import all raw files inside resources and create SAS Datasets for each file.
   Important: In order to avoid errors when transforming data sets to SAS format, all variables (e.g., SKU, BasketID) that will not be used in statistics or similar numerical operations should be read using the string type. Additionally, all the new datasets to be produced in SAS Enterprise Guide during the project should be stored in the SASUSER library.

2. **Create Project Library**
   Prepare a library, named SASMS, which will contain all prepared SAS Datasets in step "Import Raw Data".
   Important: The library should be used for all tasks.

3. **Sales vs. Returns**
   Divide the observations of the table "Invoice" into two new tables where one stores the Sales and the other one stores the returns. This division must be done using the variable "Operation". The two tables' structure should be identical to the invoices tables besides the variable "Operation".
   What was the level of Sales and Returns? Create a bar chart with the monetary values to answer the question.

4. **Customer Age**
   Calculate the customer's age and store it as new integer variable, named "Customer Age". Assume that today's date is 01/01/2019. Additionally, you should check if the data are rightfully stored by the system according to the GDPR regulation. In particular, the company is not allowed to keep data for under-age persons (i.e., less than 18 years old). A conditional report should be generated in the cases where such customers exist.

5. **Recoding**
   A. Based on the "Customer Age" variable, create a new variable named "Age Group" that is calculated as follows:

Condition (years)	Recoded Label
1 - 18	"Under 18"
18 - 25	"Very Young"
26 - 35	"Young"
36 - 50	"Middle Aged"
51 - 65	"Mature"

66 - 75   "Senior"
> 75   "Very Senior"

B. The supplier code (Supplier_ID) of each product is contained in the 9th digit its SKU code. For example, for SKU code 58720393**4**50301, the supplier code is 4. Use appropriate functions to store the supplier code as a new column named Supplier_ID.

B. Basic Analysis

1. **Invoice Total Items**
   Create a report with the Total Items for every invoice.
   The report structure should resemble the following (partial results):

	Invoice_ID	InvoiceNo	InvoiceDate	customer_id	Last_name	First_name	Payment Met...	Total Items
1	1	539730	12/21/2010	1	Johnson	Stanley	Debit Card	9
2	2	552969	05/12/2011	1	Johnson	Stanley	Debit Card	4
3	3	577382	11/18/2011	1	Johnson	Stanley	Cash	23
4	4	580363	12/02/2011	1	Johnson	Stanley	Debit Card	6
5	5	550837	04/20/2011	2	Cramer	Henry	Debit Card	1

2. **Invoice Total Value**
   Create a report with the Total Value (i.e., Quantity x Price) for every invoice. 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.
   The report structure should resemble the following (partial results):

	Invoice_ID	InvoiceNo	InvoiceDate	customer_id	Last_name	First_name	Payment Met...	Total Value
1	1	539730	12/21/2010	1	Johnson	Stanley	Debit Card	270.62
2	2	552969	05/12/2011	1	Johnson	Stanley	Debit Card	122.8
3	3	577382	11/18/2011	1	Johnson	Stanley	Cash	637.99
4	4	580363	12/02/2011	1	Johnson	Stanley	Debit Card	217.02
5	5	550837	04/20/2011	2	Cramer	Henry	Debit Card	33.31

3. **Region Analysis**
   Create a report with appropriate graphs to demonstrate the contribution to the company's revenues of each region of the country. For the top region found show the contribution to the company's revenues per gender.

4. **Basket Analysis**
   Create a report with appropriate graphs that describe the average basket. The report should include at least the number of SKU's, total monetary value, etc. Comment on your findings.

5. **Demographic Analysis**
   Create a report with the demographic analysis of the customers. Your report should include the analysis (tables and graphs) of appropriate variables and should be performed at a national and regional level. The report should include a pie chart and a frequency table with the percentages of customers that belong to each age group.

C. Advanced Analysis

1. **Product Rankings**
   Create a report that shows the top products per product type with respect to total sales value. The report should include the ranking of each product. Furthermore, the report should also include the subtotal sales of each product type at the beginning.

2. **Behavioural Characteristics**
   Create a report that analyzes the behavioral characteristics of each age group (visits to the stores, number of distinct SKU's purchased, total cost of purchases). Augment your analysis by providing pie charts for the behavioral characteristics for each age group.

3. **Promotions Analysis**
   Create a report with appropriate graphs to show what is the percentage of products that are sold with and without promotion. Create a format to display the 0% promotion as "No Promotion" and the 10%, 20% and 30% as "Promotion". Additionally, create a pie chart 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.

4. **Daily Sales**
   Create a report with appropriate graphs to answer the following managerial question: "Is there any difference among the various days with respect to the number of distinct SKU's per invoice."

5. **Supplier Analysis**
   Create a report, which includes appropriate charts, to show an analysis of supplier sales. The report should include the percentage of products that each supplier supplies, the total revenue generated by these products and the weight of that revenue (i.e., percentage of that revenue against all the sales).
   Finally, create a cross tabulation table, using the Summary Tables Wizard task to show the total revenue of the company with respect to the country 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. Add the total revenue in the middle of the cross tabulation, the origin in the rows and the suppliers in the columns.

IMPORTANT: The requirements provided in the previous section may not be sufficiently defined. You will need to record your assumptions and how these have influenced your analysis in the cases that are required.

Deliverables and File naming scheme

The deliverable consists of a compressed folder using the zip format, named as "202021.CO4759.A2.<GNumber>.zip" (e.g., 202021.CO4759.A2.G1234567.zip), with the following files:
   1. 1x Enterprise Guide project (.egp) named as 202021.CO4759.A2.G1234567.egp.
   2. 1x Document (Microsoft Word format) consisting of the documentation of all parts (i.e., A, B, C) and the screenshots of the results (partial results) and charts/tables generated through the analysis.

Grading Criteria

Marks will be awarded based on the following criteria. Within each part, aim to complete the work for each section before moving on to the next as you will not get the full credit for later sections if there are significant defects in an earlier section.

In assessing the work within a section, factors such as simplicity, quality and appropriateness of comments will be considered.

Part	Description	Range	Marking	Detailed Marking
A	Data Pre-processing	0-5	0-1 marks for each task	• 0 – No attempt or task completed with deficiencies • 1 – Task completed correctly with no deficiencies
B	Basic Analysis	0-15	0-3 marks for each task	• 0 – No attempt or serious deficiencies • 1 – Task completed correctly with minor deficiencies
C	Advanced Analysis	0-15	0-3 marks for each task	• 2 – Task completed correctly with no deficiencies, but no key insights were provided • 3 – Task completed correctly with no deficiencies, and key insights were provided
D	Presentation	0-5		

Submission of assignment work
- Anonymous marking is being used. Apart from your University ID number ("G2…"), avoid doing anything that would allow you to be identified from your work.
- *Keep a complete copy of the work you hand in.*
- Avoid submitting work at the last minute, but if there is a technical problem uploading to Blackboard, email the zip file to me before the deadline and upload the work when Blackboard is available.

Extenuating circumstances, extensions and late work
Except where an extension of the hand-in deadline date has been approved (see https://www.uclan.ac.uk/students/cyprus/study/extenuating_circumstances.php), work that is handed in up to 5 days late will be capped to 50%. After this, it will receive a mark of 0%.

Cheating
The consequences of cheating in assessments are serious. Cheating is using or attempting to use unfair means to enhance performance. This includes plagiarism (presenting someone else's work as if it was your own), collusion (working with others on an individual assignment), taking prohibited material into examinations and allowing other students to access your work. Make sure that you do not give someone the opportunity to steal your work (e.g. *by asking them to print it out for you*). We tell students about cheating both during induction and in your student handbook, but if you have any doubt about what cheating is or how to reference material properly, please ask a tutor. We recommend that you use the Harvard system for referencing.

The University operates an electronic plagiarism detection service where your work may be uploaded, stored and cross-referenced against other material. The software searches the World Wide Web and extensive databases of reference material to identify duplication.

For more information about plagiarism, please see the University Academic Regulations and the Assessment Handbook (http://www.uclan.ac.uk/aqasu/academic_regulations.php). See the Student Union Website (http://www.uclansu.co.uk/academicmatters/unfairmeans)

Reassessment and Revision

Reassessment in written examinations and coursework is at the discretion of the Course Assessment Board and is dealt with in accordance with University policy and procedures.

Appendix - Dataset Information

Customer table

CustID	LastName	FirstName	Address	Country	Post	City	Region	Gender	...
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

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