ASSIGNMENTS FOR DATA SCIENCE GROUP

ANALYTIXLABS
www.analytixlabs.co.in

First Floor, SCO 394, Sector 29, Gurgaon
info@analytixlabs.co.in
(+91)95552-19007

Assignment 1

1. Use Proc Import to import Car_sales data (given in XLS format) into SAS permanent library.
2. Clean the data set by removing the observations which have missing values in resale or price variable.
3. Create 5 datasets from the Car_sales dataset using the following price ranges: Less than or equal to USD 15K, 15k - 20K, 20k - 30K, 30k - 40k and more than USD 40 K.
4. Make a data set containing only following 4 variables: Manufacturer, Model, Sales and Price
5. (Bonus question!) Separate out data for passenger vehicles launched after 1-October-2014.

Assignment 2 – Proc Format, Proc Freq and Proc Means

Download Grocery_coupons file

1. Use Proc Import to prepare SAS data set from 'Grocery_Coupons' excel file and provide proper variable and value labels to all the variables, provided on the second tab of the excel file.
2. Generate a table to show the distribution of coupon values and shopping style. Create 2 separate tables for these 1) for overall data 2) separately for gender. *Hint: ProcFreq and by statement.*
3. Generate a table to show the distribution of Store size by store organization. Create two separate tables, one with only cell frequencies and other with only cell percentages. *Hint: Use ProcFreq to create cross-tabulation.*
4. Calculate average, min, max, variance and total amount spent at overall level, by different store size and store organization. *Hint: Use Proc Means and class statement.*

Assignment 3 – Appending and Merging

1. Using select-when statements, define a new variable to provide the country of origin for each model based on the manufacturer. *Hint: Run procfreq to see the lift of unique manufacturers in the data and then bit of googling to see country of origin.*
2. Create a unique ID in the data set using model and manufacturer variables. *Hint: Use trim and concatenate function or double pipe function (||).*
3. Make 2 separate data sets, containing following variables: **1)** Unique ID, Manufacturer, Model, launch data, Sales, Resale and Price **2)** Unique ID, and remaining technical variables. *Hint: Use Keep and drop statements.*
4. Using Data Instream option create a new data file "Hyundai" with following data and again create a unique ID based on manufacturer and model:

Manufacturer	Model	Sales in thousands	4-year resale value	Latest Launch
Hyundai	Tuscon	16.919	16.36	2Feb12
Hyundai	i45	39.384	19.875	3Jun11
Hyundai	Verna	14.114	18.225	4Jan12
Hyundai	Terracan	8.588	29.725	10Mar11

5. Create a new file "Total_sales" by appending data file "Hyundai" with the file first file created in problem 3.
6. Create a new data set after merging Total_sales file with the second file created in problem 3. *Hint: Use Proc sort and merge statement (by unique ID).*
7. Create a new data set after merging Total_sales file with the second file created in problem 3 but the new file should only have common records from both the files. *Hint: Use Proc sort and merge statement (by unique ID) with IN option.*

Assignment 4 - Loops

1. On January 1 of each year, $5,000 is invested in an account. Complete the DATA step with DO LOOP to determine the value of the account after 15 years i) if a constant annual interest rate of 10% is expected ii) if a compounding annual interest rate of 10% is expected.
2. A car delivers a mileage of 20 miles per gallon. Write a program so that the program stops generating observations when Distance reaches 250 miles or when 10 gallons of fuel have been used. *Hint: Miles=gallon x mpg.*
3. In a fixed term deposit of 25 years calculate the total amount at the end of term with initial amount of $5,00,000 and annual interest rate of 7% i) compounded annually ii) compounded monthly. Show the amount accrued at monthly level.

Assignment 5 - Functions

Download Grocery_coupons file

1. Calculate the months, weeks and days separately since the coupon expiry date (from current date and 31-Mar-14). *Hint: Use INTCK function.*
2. Calculate a field for coupon issuance date, assuming each coupon is valid for 3 months. *Hint: Use INTNX function.*
3. Using Datedif function, calculate the number of days between coupon expiry and 30-Sep-2014 (assuming 30 days in each month of the year).
4. The retail chain is planning to introduce faster and convenient billing system, in which the customers will be charged whole number dollar value rather than current system of charging exact amount. Do an estimation which method would be more profitable and by how much margin; charging based on the integer value of the bill amount or rounding off to nearest dollar value.

Download Department file

5. Create a new variable last name from name variable. *Hint: Use SCAN function.*
6. Find the starting position of first name. *Hint: Use INDEX OR FIND function.*
7. Create new variable first name from name variable using the positions found in previous question.*Hint: Use SUBSTR function*

Assignment 6 (PROC SQL)

SMB data provides the landline usage information for 33333 small and medium businesses (SMB) for months of April-September during 3 different years (2009-2011). Hence each customer will appear 3 times for each respective year, with extensive usage information across different kind of split in various columns.

Churners and labels file has two sheets:

1. Churners – which provides the month of churning for all product_ids that have churned
2. Labels – which provides detailed variables labels

Churners are the customers in service industry who quit/ stop using the services of a particular service provider and may switch to another service provider (example in case of telecom, internet, banking and insurance).

1. Import SMB usage data into SAS and provide labels from 'Churners and labels' file. Be patient with this data it's huge with 99999 observations and more than 400 variables. You may use excel to make label syntax to save time. (Please read the data description provided along with the files.)

2. In the SMB usage data create another column to identify churners by using product_ID as key variable (check if the Key variable is unique ID in this case). The list of churners is given in the 'Churners and labels' along with churn month. You may use left join or right join as per your convenience.

3. From SMB usage data select Product_ID, Age on Network fields and calculate average, min and max Age of network, using PROC SQL.

4. You may notice that there is discrepancy in the Age on network variable i.e. for same line number different values are provided in different years. After checking with client we got to know that maximum Age on Network out of the 3 given value per line number should be deem as correct. Using PROC SQL pick the correct AON for each line and rectify in the data. (Hint: Make a data set having Max Age on network by line number and join with the main data).

5. Calculate the average Call duration, count of local calls, and total usage charge for Churners and Non-Churners for all the 6 months, only for years 2010 and 2011, and where average call duration is less than 60 seconds.

6. Create a sub-set of the SMB usage data only with usage variables for July, August, and September rather than all the 6 months.

7. In the Grocery_coupons data, categorize the customers into 3 categories based on total spending (below 100, 100-200, above 200) and display a count of customers for each of these categories

8. (Bonus question!) Refer to the Airlines_grouping file. The data provided in the first tab has to be summarized like in the output tab: i) Airlines to be clubbed as 2 entities Group 1: Jet Airways and Air India and Group2: Others ii) Top 15 clients to be identified location-wise based upon the total volume and the airline-group share.

Assignment 7 (SAS Macros - OPTIONAL)

Question-1:
Write a macro to sort & print the data by satisfying the following conditions. (Use Models data set)

1. Macro should able to take the request for ascending or descending order
2. It should be able to give option to sort on specific variable
3. It should print the data set with the titles of information like sorted by ascending price.

Question-2:
Data: Orders data set
The company maintains a file with information about every order they receive. For each order, the data includes the customerID number, date the order was placed, model name, and quantity ordered.

Business Problem:
Every Monday the president of the company wants to see all the current orders. On Friday, the president wants a report summarized by customer.

Write a macro as one program that satisfies both conditions by using conditional logic with the automatic variable &SYSDAY. The report prints all current orders if you run it on Monday, and a customer-wise summary report if you run it on Friday (no other days)

Question-3:

/*Iterative processing
Write a macro to create the following dataset
var_1 var_2 ..... var_9 Var_10
1 4 ..... 81 100
1 8 ..... 729 1000
*/

Question-4:
/*Conditional iteration
Identify words in a list supplied as macro parameter.
Start with the first word and continue as long as more words are found.
If no further word is found, display a message.*/

Question-5:
Write a macro to create a two-way table using proc freq with the cross-list option. Include the variables to be included as macro parameters.

Use the dataset car_sales focusing on categorical variables – Manufacturer, Model and Vehicle Type.

a. write macro by using positional parameters

b. write macro by use keyword parameters */