Assignment 1

Bank Marketing Case Study: loading and merging data

Sukanya Mukherjee

128041217

BAN110ZBB

Data Preparation and Handling

Dr. Milad Rezamand

Learning outcomes

1. Load data using input Files in Various Formats to combine information from many data domains and sources

2. Rename columns and convert column types from character to numeric to prepare for merging

3. Merge sas datasets to obtain a Datawarehouse ready for analysis

Introduction

The head of Marketing wants to know which customers have the highest propensity for buying a Certificate of Deposit (CD) from the institution. The goal of this assignment is to create part of an analytical data mart by combining information from many data domains and sources.

Q1. Load data from customer_banking_info_promo.xslx

- define the library name "mylib" and specify its location using libname
- Use proc import DATAFILE to import customer_banking_info_promo.xlsx into a sas dataset named customer_banking_info_promo under mylib
- Print the first five rows of the dataset add (obs=5) at the end of proc print.

Answer: Code

Q1.sas

CODE LOG RESULTS OUTPUT DATA

```
/*define the library name "mylib" and specify its location using libname*/
libname mylib "~";
data mylib.customer_banking_info_promo;
run;
/*Use proc import DATAFILE to import customer_banking_info_promo.xlsx
into a sas dataset named customer_banking_info_promo under mylib*/
proc import
        datafile="/home/u59406283/customer_banking_info_promo.xlsx"
        out=mylib.customer_banking_info_promo
        dbms=xlsx replace;
run;
/*Print the first five rows of the dataset add (obs=5) at the end of proc print*/
proc print data=mylib.customer_banking_info_promo (obs=5) noobs;
run;
```

Results:

Q1.sas

CODE LOG RESULTS OUTPUT DATA

Table of Contents

customer_id2	contact	day	month	duration	campaign	pdays	previous	poutcome	y
122482	cellular	22	aug	229	2	-1	0	unknown	no
119725	cellular	7	aug	125	2	-1	0	unknown	no
103490	unknown	15	may	68	2	-1	0	unknown	no
126218	cellular	19	nov	517	2	187	3	failure	no
104835	unknown	20	may	165	2	-1	0	unknown	no

Q2. Examine the variable Customer ID. Check the type and format.

- Use proc content procedure to examine the variables and their types. This will also print more details.

ref:
http://support.sas.com/documentation/cdl/en/proc/65145/HTML/default/viewer.htm#p120panelmbpren1m0j2n77s9f67.htm

or

https://www.cpc.unc.edu/research/tools/data_analysis/sastopics/contents

Answer: Code

Q2.sas

CODE LOG RESULTS

Line #

```
/*mylib is already defined and is present in Libraries*/
/*Use proc content procedure to examine the variables and their types.
This will also print more details*/
proc contents data=mylib.customer_banking_info_promo;
run;
```

Results:

Q2.sas

CODE LOG RESULTS

Table of Contents

The CONTENTS Procedure

Data Set Name	MYLIB.CUSTOMER_BANKING_INFO_PROMO	**Observations**	10578
Member Type	DATA	**Variables**	10
Engine	V9	**Indexes**	0
Created	29/09/2021 03:47:31	**Observation Length**	72
Last Modified	29/09/2021 03:47:31	**Deleted Observations**	0
Protection		**Compressed**	NO
Data Set Type		**Sorted**	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Engine/Host Dependent Information	
Data Set Page Size	131072
Number of Data Set Pages	6
First Data Page	1
Max Obs per Page	1816
Obs in First Data Page	1773
Number of Data Set Repairs	0
Filename	/home/u59406283/customer_banking_info_promo.sas7bdat
Release Created	9.0401M6
Host Created	Linux
Inode Number	21004182962
Access Permission	rw-r--r--
Owner Name	u59406283
File Size	896KB
File Size (bytes)	917504

Alphabetic List of Variables and Attributes						
#	Variable	Type	Len	Format	Informat	Label
6	campaign	Num	8	BEST.		campaign
2	contact	Char	9	$9.	$9.	contact
1	customer_id2	Char	6	$6.	$6.	customer_id2
3	day	Num	8	BEST.		day
5	duration	Num	8	BEST.		duration
4	month	Char	3	$3.	$3.	month
7	pdays	Num	8	BEST.		pdays
9	poutcome	Char	7	$7.	$7.	poutcome
8	previous	Num	8	BEST.		previous
10	y	Char	3	$3.	$3.	y

Q3. Column deletion/renaming

Look at the description of the different columns here:
https://archive.ics.uci.edu/ml/datasets/bank+marketing

duration:

last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

Within a data step, perform the following:

- keep the output dataset name same as input (customer_banking_info_promo)
- Rename "customer_id2" to customer_id
- drop the column "duration" from the dataset.
- print the first 5 observations in the dataset

References:

rename option: https://newonlinecourses.science.psu.edu/stat481/node/17/

drop option: https://newonlinecourses.science.psu.edu/stat481/node/15/

Answer: Code

Q3.sas

CODE LOG RESULTS OUTPUT DATA

```
data customer_banking_info_promo;
/*keep the output dataset name same as input (customer_banking_info_promo)*/
/*Rename "customer_id2" to customer_id*/
    set mylib.customer_banking_info_promo(rename=(customer_id2=customer_id));
    /*drop the column "duration" from the dataset*/
    drop duration;
run;
/*print the first 5 observations in the dataset*/
proc print data=customer_banking_info_promo (obs=5) noobs;
run;
```

Results:

Q3.sas

CODE LOG RESULTS OUTPUT DATA

Table of Contents

customer_id	contact	day	month	campaign	pdays	previous	poutcome	y
122482	cellular	22	aug	2	-1	0	unknown	no
119725	cellular	7	aug	2	-1	0	unknown	no
103490	unknown	15	may	2	-1	0	unknown	no
126218	cellular	19	nov	2	187	3	failure	no
104835	unknown	20	may	2	-1	0	unknown	no

Q4. Load data from customer_banking_info.csv

- load the data and print the first five rows.

Answer: Code

Q4.sas

CODE LOG RESULTS OUTPUT DATA

```
/*Load data from customer_banking_info.csv*/
proc import datafile="/home/u59406283/customer_banking_info.csv"
            dbms=csv
            out=mylib.customer_banking_info
            replace;
run;
/*print the first five rows*/
proc print data=mylib.customer_banking_info (obs=5) noobs;
run;
```

Results:

Q4.sas

CODE LOG RESULTS OUTPUT DATA

Table of Contents

customer_id1	default	balance	housing	loan
122482	no	347	no	no
119725	no	3462	no	no
103490	no	157	yes	no
126218	no	3689	yes	no
104835	no	0	yes	yes

Q5. Renaming columns

use proc contents to examine the list of variables as before.

You will see that customer_id1 is numerical with len=8.

This is important to check as this column will be used to merge the datasets.

Within a data step, perform the following:

- keep the output dataset name same as the input dataset name (customer_banking_info)
- Rename "customer_id1" as customer_id
- print the first 5 observations in the dataset

Answer: Code

Q5.sas

CODE LOG RESULTS OUTPUT DATA

```
/*use proc contents to examine the list of variables as before.
You will see that customer_id1 is numerical with len=8*/
proc contents data=mylib.customer_banking_info;
run;
/*keep the output dataset name same as the input dataset name (customer_banking_info)*/
data customer_banking_info;
/*Rename "customer_id1" as customer_id*/
    set mylib.customer_banking_info(rename=(customer_id1=customer_id));
run;
/*print the first 5 observations in the dataset*/
proc print data=customer_banking_info (obs=5) noobs;
run;
```

Q5.sas

CODE LOG RESULTS OUTPUT DATA

Table of Contents

The CONTENTS Procedure

Data Set Name	MYLIB.CUSTOMER_BANKING_INFO	**Observations**	10578
Member Type	DATA	**Variables**	5
Engine	V9	**Indexes**	0
Created	29/09/2021 03:58:59	**Observation Length**	32
Last Modified	29/09/2021 03:58:59	**Deleted Observations**	0
Protection		**Compressed**	NO
Data Set Type		**Sorted**	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Engine/Host Dependent Information	
Data Set Page Size	131072
Number of Data Set Pages	3
First Data Page	1
Max Obs per Page	4078
Obs in First Data Page	4002
Number of Data Set Repairs	0
Filename	/home/u59406283/customer_banking_info.sas7bdat
Release Created	9.0401M6
Host Created	Linux
Inode Number	21255903203
Access Permission	rw-r--r--
Owner Name	u59406283
File Size	512KB
File Size (bytes)	524288

Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Informat
3	balance	Num	8	BEST12.	BEST32.
1	customer_id1	Num	8	BEST12.	BEST32.
2	default	Char	3	$3.	$3.
4	housing	Char	3	$3.	$3.
5	loan	Char	3	$3.	$3.

customer_id	default	balance	housing	loan
122482	no	347	no	no
119725	no	3462	no	no
103490	no	157	yes	no
126218	no	3689	yes	no
104835	no	0	yes	yes

Q6. SAS data from customer_demographics.sas7bdat

- print the first 5 rows of customer_demographics.sas7bdat
- use proc contents and examine the list of variables. What is the type of customer_id - **NUM**

Answer: Code

Q6.sas

CODE LOG RESULTS

```
/*print the first 5 rows of customer_demographics.sas7bdat*/
proc print data=mylib.customer_demographics (obs=5) noobs;
run;
/*use proc contents and examine the list of variables. What is the type of customer_id*/
proc contents data=mylib.customer_demographics;
run;
```

Results:

Q6.sas

CODE LOG RESULTS

Table of Contents

Education	customer_id	AGE	marital	JOB
secondary	100103	33	married	entrepreneur
tertiary	100106	35	married	management
primary	100118	57	married	blue-collar
primary	100119	60	married	retired
secondary	100121	28	married	blue-collar

The CONTENTS Procedure

Data Set Name	MYLIB.CUSTOMER_DEMOGRAPHICS	**Observations**	10578
Member Type	DATA	**Variables**	5
Engine	V9	**Indexes**	0
Created	24/01/2019 22:57:12	**Observation Length**	48
Last Modified	24/01/2019 22:57:12	**Deleted Observations**	0
Protection		**Compressed**	NO
Data Set Type		**Sorted**	YES
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Engine/Host Dependent Information	
Data Set Page Size	65536
Number of Data Set Pages	8
First Data Page	1
Max Obs per Page	1360
Obs in First Data Page	1310
Number of Data Set Repairs	0
Filename	/home/u59406283/customer_demographics.sas7bdat
Release Created	9.0401M5
Host Created	Linux
Inode Number	21006274992
Access Permission	rw-r--r--
Owner Name	u59406283
File Size	576KB
File Size (bytes)	589824

Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Label
3	AGE	Num	8	F4.	AGE
1	Education	Char	9	$CHAR9.	Education
5	JOB	Char	14	$CHAR14.	JOB
2	customer_id	Num	8		
4	marital	Char	8	$CHAR8.	marital

Sort Information	
Sortedby	customer_id
Validated	YES
Character Set	ASCII

Q7. Convert from character to numeric type

Before merging multiple datasets, the common column between the datasets should be of the same type.

In customer_banking_info_promo, customer_id is defined as character.

You are given a sample data step code to run:

- the output dataset name customer_banking_info_promocv
- to convert customer_id to numeric variable, we use the input function.
  reference: http://support.sas.com/kb/24/590.html
- check the customer_id variable type again by using proc contents or proc means to see the list of numerical variables

Answer: Code

Q7.sas

CODE LOG RESULTS OUTPUT DATA

```
data customer_banking_info_promocv;
/*the output dataset name customer_banking_info_promocv*/
    set customer_banking_info_promo;
    /*to convert customer_id to numeric variable, we use the input function.*/
    customer_id_new = input(customer_id, 8.);
    drop customer_id;
    rename customer_id_new=customer_id;
run;
/*check the customer_id variable type again by using proc contents or
proc means to see the list of numerical variables*/
proc contents data=customer_banking_info_promocv;
run;
```

Results:

Q7.sas

CODE LOG RESULTS OUTPUT DATA

Table of Contents

The CONTENTS Procedure

Data Set Name	WORK.CUSTOMER_BANKING_INFO_PROMOCV	Observations	10578
Member Type	DATA	Variables	9
Engine	V9	Indexes	0
Created	29/09/2021 04:12:53	Observation Length	64
Last Modified	29/09/2021 04:12:53	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Engine/Host Dependent Information	
Data Set Page Size	131072
Number of Data Set Pages	6
First Data Page	1
Max Obs per Page	2043
Obs in First Data Page	1996
Number of Data Set Repairs	0
Filename	/saswork/SAS_workFE0D0001C031_odaws03-usw2.oda.sas.com/SAS_work6DF60001C031_odaws03-usw2.oda.sas.com/customer_banking_info_promocv.sas7bdat
Release Created	9.0401M6
Host Created	Linux
Inode Number	1074936928
Access Permission	rw-r--r--
Owner Name	u59406283
File Size	896KB
File Size (bytes)	917504

Alphabetic List of Variables and Attributes						
#	Variable	Type	Len	Format	Informat	Label
4	campaign	Num	8	BEST.		campaign
1	contact	Char	9	$9.	$9.	contact
9	customer_id	Num	8			
2	day	Num	8	BEST.		day
3	month	Char	3	$3.	$3.	month
5	pdays	Num	8	BEST.		pdays
7	poutcome	Char	7	$7.	$7.	poutcome
6	previous	Num	8	BEST.		previous
8	y	Char	3	$3.	$3.	y

Q8. Data Merging

Join the three sources of data into a single SAS data set.

- sort each of the datasets by customer_id
- merge the three datasets using the merge function within a data step. name the new dataset as "customer_all"
- print the first five observations.

Refer to https://newonlinecourses.science.psu.edu/stat481/node/28/

Answer: Code

Q8.sas

CODE LOG RESULTS OUTPUT DATA

```
/*sort each of the datasets by customer_id*/
proc sort data=customer_banking_info_promocv out=customer_banking_1_sort;
    by customer_id;
run;

proc sort data=customer_banking_info out=customer_banking_2_sort;
    by customer_id;
run;

proc sort data=mylib.customer_demographics out=customer_banking_3_sort;
    by customer_id;
run;
/*merge the three datasets using the merge function within a data step
name the new dataset as "customer_all"*/
data customer_all;
    merge customer_banking_1_sort customer_banking_2_sort customer_banking_3_sort;
    by customer_id;
run;
/*print the first five observations*/
proc print data=customer_all (obs=5) noobs;
run;
```

Results:

Q8.sas

CODE LOG RESULTS OUTPUT DATA

Table of Contents

contact	day	month	campaign	pdays	previous	poutcome	y	customer_id	default	balance	housing	loan	Education	AGE	marital	JOB
unknown	5	may	1	-1	0	unknown	no	100103	no	2	yes	yes	secondary	33	married	entrepreneur
unknown	5	may	1	-1	0	unknown	no	100106	no	231	yes	no	tertiary	35	married	management
unknown	5	may	1	-1	0	unknown	no	100118	no	52	yes	no	primary	57	married	blue-collar
unknown	5	may	1	-1	0	unknown	no	100119	no	60	yes	no	primary	60	married	retired
unknown	5	may	1	-1	0	unknown	no	100121	no	723	yes	yes	secondary	28	married	blue-collar