Due: _____/_____/_____

R · I · T	**Rochester Institute of Technology** **Golisano College of Computing and Information Sciences** **School of Information**

Practice Exercise 5 (PE05)
Data Cleansing with SAS

(**Not** a Team Assignment)

Overview

In this exercise, you will have an opportunity to investigate how data analysis techniques can cleanse data in preparation for loading it into a data warehouse/mart.

After completing this exercise, you should be able to:

- Show how data analysis software, in this case, SAS, can be used to check for problems in data.
- Discuss why data cleansing is a complicated but essential process.

For this exercise, you will need:

- Access to the SAS software on one of the Windows servers in the iSchool Database labs.
- A copy of the data file, patients.txt, which is available from the RIT *my*Courses website for this course.

Step #1: *Get the Data*

Download the patient data file onto your machine into the directory: C:\cleaning. Open and investigate the data. Note the number of rows.

The format of the data in the patients.txt[1] file is:

Variable Name	Description	Length	Data Type	Valid Values
patientNo	Patient ID	3	Character	Numbers only; If missing, duplicate or none alpha, assign a unique number;
gender	Patient gender	1	Character	‘M’ or ‘F’

visit	Visit date	8	Character (MMDDYYYY)	Any valid date;
				If missing, 1/1/1900;
				If month>12, 12;
				If day>31, 31;
				If year>1999, 1999;
				If non-digit, 1/1/1900

HR	Heart rate	3	Numeric	≥ 40 and ≤ 100;

[1] Source: “Cody’s Data Cleaning Techniques, Using SAS Software,” Ron Cody, SAS Institute Press, 1999.

				If missing, 40;
				If <40, 40;
				If >100, 100

SBP	Systolic blood pressure	3	Numeric	≥ 80 and ≤ 200;
				If missing, 80;
				If <80, 80;
				If >200, 200

DBP	Diastolic blood pressure	3	Numeric	≥ 60 and ≤ 120;
				If missing, 60;
				If <60, 60;
				If >120, 120

DX	Diagnosis code	3	Character	1- to 3-digit number;
				If missing, 999;
				If non-digit, 999

AE	Adverse event	1	Character (Boolean)	‘0’ or ‘1’; If missing or invalid, 0;

Step #2: *Start the SAS software*

Follow the instructions in the “SAS University Edition: QuickStart Guide for Oracle VirtualBox or VMware Workstation Player.”

On the lab computers,

- Click the SASUniversity icon on the desktop and then select "I moved it". It should start the server then.

Take a few minutes to investigate the SAS interface and to identify all of the functional components discussed in the lecture. Note the startup information displayed in the SAS Log window.

Step #3: *Data Cleansing*

Create a SAS program, PE05 and then clean up the patients.txt file to meet the data specifications given in the Step #1. Export the final clean data to a file called patients_clean.txt. Submit PE05.sas and patients_clean.txt to the MyCourses drop box.

Note: see next page for some helpful hints & included Selected SAS Functions.pdf

Some Helpful Hints:

To run your SAS program, click the 🏃 icon. Check the Log window to see if your program ran successfully. If so, your output will be in the Output window. If you had an error, fix it and rerun.

Note: your run results are saved in the Results window. They will stay there while your SAS session is active; however, they will be deleted at shutdown unless you save them (use Downloads/Print icons while the RESULTS window is active).

Save your SAS program (use Save/Save As icons with the CODE window active).

Did you have any “bad” data? Remember, bad data must be “cleansed” before loading it into your data warehouse/mart.

There are lots of analysis options. Some handy options are:

- If the values in two columns are related, you can determine if they represent the appropriate business rules by cross tabulating them.

  ```
  proc freq;
       title "...";
       tables var_x * var_y / missing;
  run;
  ```

- You can break down a character string with the substring

  function, substr(). `data WORK.SOMENAMEtemp;     *`

  ```
  declare a temporary dataset; set

  SOMELIBRARY.SOMEDATASET (rename = (var1= ...

  ));

  * declare a local variable and assign a value
  to it stringPart =
  substr(variableName,startingPosition,length);
  ```

- Data can be converted between character and numeric formats:
  - Character to numeric: INPUT
  - Numeric to character: PUT

  See the reference in section 11.8 in Chapter 11 of Delwiche and Slaughter, “The Little SAS Book: A Primer, Sixth Edition.” **This can be found in RIT Library.**

- There are many SAS functions which can also be found in “The Little SAS Book”, in Chapter 3.

- Also, chapter 6 of this book is a good section on modifying data

Name: ______________________________

R · I · T	**Rochester Institute of Technology** **Golisano College of Computing and Information Sciences** **School of Information**

PE05: Data Cleansing with SAS

Requirements	**Point Value**	**Points Earned**
Patients_clean.txt:		
- Headers for all the file existed in the cleansed text file	10	
- Data for patientNo, visit, HR, SBP, DBP, DX, AE fields & duplicate record are cleansed.	35	
PE05.sas:		
- The process of cleansing the fields is shown in the SAS file.	35	
- The SAS file can be run	20	
Points Earned	100	
Graded By		

Comments: