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.

patients_clean.txt :

```
patientNo;gender;visit;HR;SBP;DPB;DX;AE
0;M;11111998;88;140;80;1;0
1;F;11131998;84;120;78;999;0
2;M;10211998;68;190;100;3;1
3;F;01011999;100;200;120;5;0
4;M;05071998;68;120;80;1;0
5;F;06151999;72;102;68;6;1
6;M;08311998;88;148;102;999;0
7;M;11111998;90;190;100;999;0
8;F;08081998;100;80;60;7;0
9;M;09251999;86;200;120;4;1
10;F;10191999;40;80;120;1;0
11;M;12131998;68;200;60;4;1
12;M;101298;60;122;74;999;0
13;F;08231999;74;108;64;1;0
14;M;02021999;40;130;90;999;1
15;F;11131998;84;120;78;999;0
16;M;11121999;58;112;74;999;0
17;F;01011900;82;148;88;3;1
18;F;04051999;100;80;84;2;0
19;M;06071999;58;118;70;999;0
20;M;12121999;60;80;60;1;0
21;F;01011900;100;200;120;5;1
22;F;12311999;40;80;60;999;0
23;M;10101999;48;114;82;2;1
24;F;12311998;40;80;78;999;0
25;F;11091998;76;120;80;1;0
26;M;01011999;74;102;68;5;1
27;F;01011900;40;166;106;7;0
28;F;03281998;66;150;90;3;0
29;M;05151998;40;80;60;4;1
30;F;07071999;82;148;84;1;0
```

```
* Data Warehousing PE05 Ellie Parobek;

* Read in data;
libname MyLib 'folders/myfolders';
data patients;
	infile '/folders/myfolders/cleaning/patients.txt' dsd delimiter=';' firstobs=1;
	input patientNo $ gender $ visit $ HR $ SBP $ DPB $ DX $ AE;
run;

data patients;
	modify patients;
	retain counter 0;

	* Fix patientNo;
	patientNo=counter;
	counter+1;

	* Fix visit;
	if missing(visit) then visit="01011900";
	if substr(visit, 1, 2)>12 then substr(visit, 1, 2)="12";
	if substr(visit, 3, 2)>31 then substr(visit, 3, 2)="31";
	if substr(visit, 5, 4)>1999 then substr(visit, 5, 4)="1999";
	if anyalpha(visit) then visit="01011900";

	* Fix HR;
	if missing(HR) then HR=40;
	if HR<40 then HR=40;
	if HR>100 then HR=100;

	* Fix SBP;
	if missing(SBP) then SBP=80;
	if SBP<80 then SBP=80;
	if SBP>200 then SBP=200;

	* Fix DPB;
	if missing(DPB) then DPB=60;
	if DPB<60 then DPB=60;
	if DPB>120 then DPB=120;

	* Fix DX;
	if missing(DX) then DX=999;
	if anyalpha(DX) then DX=999;

	* Fix AE;
	if missing(AE) then AE=0;
	if (AE ne 1 & AE ne 0) then AE=0;
run;

* Export data;
proc export data=patients outfile='/folders/myfolders/cleaning/patients_clean.txt' dbms=dlm
replace;
	delimiter=';';
run;
```

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: