

Homework #7-Reading Raw Data Files; Data Transformaton Part I

Directions: Please submit one program file, one output file, and one log file for the entire assignment. Use comment statements to separate your answers. For questions that do not require a SAS program use comment statements. For example:

```
/*
```

```
Question #1d: my answer
```

```
Question #2a: my answer
```

```
*/
```

```
/*Question #4b: */
```

```
--SAS program--
```

```
/*Question #5*/
```

Please make sure the log and output file contain only one run. For example, clear the screen for the log and output file and submit your program one last time before you upload your solutions to **Blackboard**. See lab 1 for the instructions on how to clear your output and log files.

Part I: Reading Raw Data Files

1. Using Formatted Input and the Subsetting IF Statement

The raw data file **sales1.dat** has employee information for the Australian and U.S. sales staff. The record layout is shown in the table below.

Layout for **sales1.dat**

	Field Description	Starting Column	Length of Field	Data Type
➡	Employee ID	1	6	Numeric
	First Name	8	12	Character
➡	Last Name	21	18	Character
	Gender	40	1	Character
➡	Job Title	43	20	Character
➡	Salary	64	8	Numeric \$100,000
➡	Country	73	2	Character 'AU' or 'US'
	Birth Date	76	10	Numeric mm/dd/yyyy
➡	Hire Date	87	10	Numeric mm/dd/yyyy

- Create two SAS data sets from the raw data file, and base them on the country of the trainee.

- Name the data sets **US_trainees** and **AU_trainees**. For this exercise, a trainee is anyone that has the job title of Sales Rep. I
- Each data set should contain the fields indicated by arrows in the layout table.
- Write only U.S. trainees to the **US_trainees** data set and only Australian trainees to the **AU_trainees** data set. Do **not** keep the **Country** variable in the output data sets.

b. Print both of the data sets with appropriate titles.

Partial **work.AU_trainees** (21 Total Observations)

Australian Trainees					
Employee_ID	Last_Name	Job_Title	Salary	Hire_Date	
120123	Hotstone	Sales Rep. I	26190	18901	
120124	Daymond	Sales Rep. I	26480	18687	
120130	Lyon	Sales Rep. I	26955	18748	
120131	Surawski	Sales Rep. I	26910	18628	
120136	Leyden	Sales Rep. I	26605	18659	

Partial **work.US_trainees** (42 Total Observations)

US Trainees					
Employee_ID	Last_Name	Job_Title	Salary	Hire_Date	
121023	Fuller	Sales Rep. I	26010	18748	
121028	Smades	Sales Rep. I	26585	18932	
121029	Mcelwee	Sales Rep. I	27225	18962	
121030	Areu	Sales Rep. I	26745	18659	
121036	Mesley	Sales Rep. I	25965	18901	

2. Working with Mixed Record Types

- The raw data file **sales3.dat** has employee information for the Australian and U.S. sales staff.
- Information for each employee is in two lines of raw data.
- The record layouts are shown below.

Line 1 layout

	Field Description	Starting Column	Length of Field	Data Type
➡	Employee ID	1	6	Numeric
	First Name	8	12	Character
➡	Last Name	21	18	Character
	Gender	40	1	Character
➡	Job Title	43	20	Character

Line 2 layout for Australian employees

	Field Description	Starting Column	Length of Field	Data Type
➡	Salary	1	8	Numeric \$100.000
➡	Country	10	2	Character
	Birth Date	13	10	Numeric dd/mm/yyyy

➡	Hire Date	24	10	Numeric dd/mm/yyyy
---	-----------	----	----	-----------------------

Line 2 layout for U.S. employees

	Field Description	Starting Column	Length of Field	Data Type
➡	Salary	1	8	Numeric \$100,000
➡	Country	10	2	Character
	Birth Date	13	10	Numeric mm/dd/yyyy
➡	Hire Date	24	10	Numeric mm/dd/yyyy

- a. Create two new SAS data sets, **US_sales** and **AU_sales**, that contain the fields indicated by arrows in the layout table. Write only U.S. employees to the **US_sales** data set and only Australian employees to the **AU_sales** data set. Do **not** include the **Country** variable in the output data sets.



The salary and hire date values are different for Australian and U.S. employees. Be sure to use the correct informats in each INPUT statement.

- b. Print both of the data sets with appropriate titles.

Partial **work.AU_sales** (63 Total Observations)

Australian Sales Staff					
Employee_ ID	Last_Name	Job_Title	Salary	Hire_ Date	
120102	Zhou	Sales Manager	108255	12205	
120103	Dawes	Sales Manager	87975	6575	
120121	Elvish	Sales Rep. II	26600	6575	
120122	Ngan	Sales Rep. II	27475	8217	
120123	Hotstone	Sales Rep. I	26190	18901	
120124	Daymond	Sales Rep. I	26480	18687	
120125	Hofmeister	Sales Rep. IV	32040	8460	

Partial **work.US_sales** (102 Total Observations)

US Sales Staff					
Employee_ ID	Last_Name	Job_Title	Salary	Hire_ Date	
120261	Highpoint	Chief Sales Officer	243190	11535	
121018	Magolan	Sales Rep. II	27560	6575	
121019	Desanctis	Sales Rep. IV	31320	17684	
121020	Ridley	Sales Rep. IV	31750	16922	
121021	Farren	Sales Rep. IV	32985	13939	
121022	Stevens	Sales Rep. IV	32210	16833	
121023	Fuller	Sales Rep. I	26010	18748	
121024	Westlund	Sales Rep. II	26600	17653	

Part II-Data Transformation Part I

1. Extracting Characters Based on Position

The data set **orion.newcompetitors** has data on competing retail stores that recently opened near existing Orion Star locations.

orion.newcompetitors

ID	City	Code	Postal_
AU15301W	PERTH	6002	
AU12217E	SYDNEY	2000	
CA 150	Toronto	M5V 3C6	
CA 238	Edmonton	T5T 2B2	
US 356NC	charlotte	28203	
US1013CO	denver	80201	
US 12CA	San diego	92139	

- Orion Star would like a data set containing only the small retail stores from these observations.
 - Create a new variable, **Country**, that contains the first two characters of **ID**.
 - Create a new variable, **Store_Code**, that contains the other characters from the value in **ID**. Left-justify the value so that there are no leading blanks.
 - The first character in the value of **Store_Code** indicates the size of the store, and **1** is the code for a small retail store.
 - Write a program to output only the small retail store observations.
Hint: You might need to use a SUBSTR functions as part of a subsetting IF statement
 - Make sure that the **City** values appear in proper case (as displayed below).

- Print your results with an appropriate title.

Show these columns only once: **Store_Code**, **Country**, **City**, and **Postal_Code**.

PROC PRINT output (5 Total Observations)

New Small-Store Competitors				
Store_ Code	Country	City	Postal_ Code	
15301W	AU	Perth	6002	
12217E	AU	Sydney	2000	
150	CA	Toronto	M5V 3C6	
1013CO	US	Denver	80201	
12CA	US	San Diego	92139	

2. Searching Character Values and Explicit Output


- The data set **orion.employee_donations** contains information about charity contributions from Orion Star employees.
- Each employee is allowed to list either one or two charities, which are shown in the **Recipients** variable.

Partial **orion.employee_donations** (124 Total Observations, 7 Total Variables)

Employee_ID	Recipients
120265	Mitleid International 90%, Save the Baby Animals 10%
120267	Disaster Assist, Inc. 80%, Cancer Cures, Inc. 20%
120269	Cancer Cures, Inc. 10%, Cuidadores Ltd. 90%
120270	AquaMissions International 10%, Child Survivors 90%
120271	Cuidadores Ltd. 80%, Mitleid International 20%

 Some charity names have a comma in them.

- Use explicit output to create a data set named **work.split**.
 - The data set has one observation for each combination of employee and charity to which he donated.
 - Some employees made two contributions. Therefore, they have two observations in the output data set. These employees contain a % character in the value of **Recipients**.

 Store the position where the % character is found in a variable named **PctLoc**. This can make subsequent coding easier.

- Create a variable named **Charity** with the name and percent contribution of the appropriate charity.
- Read only the first 10 observations from **orion.employee_donations** to test your program.

b. Modify the program to read the entire **orion.employee_donations** data set.

- Print only the columns **Employee_ID** and **Charity**.
- Add an appropriate title.

Partial PROC PRINT Output (212 Total Observations)

Charity Contributions for each Employee	
Employee_ID	Charity
120265	Mitleid International 90%
120265	Save the Baby Animals 10%
120267	Disaster Assist, Inc. 80%
120267	Cancer Cures, Inc. 20%
120269	Cancer Cures, Inc. 10%

Part III- Supplemental exercises for STAT 625 and Honors credit

1. Using a Text String with Column Pointer Controls

- The raw data file **seminar.dat** contains comments and ratings from participants at a seminar given to Orion Star sales staff.
- The data file contains one line for each participant:
 - The first 15 characters are reserved for the name of the participant (if given).
 - There can be a comment of up to 60 characters.
 - The text **Rating:** is followed immediately by a numeric score from 1 to 5.

Listing of **seminar.dat**

J. Mitchell	Very Well done! Rating:5
Amy Jung	Rating:4
Carl Heisman	Rating:4
Linda Deal	Not enough give aways Rating:3
Gabrielle Heron	Nice! Rating:4
	Not helpful at all Rating:2
Kyle Patterson	Very good. Need more like it Rating:5

- Create a new SAS data set named **seminar_ratings** that contains the names of the participants and the ratings that were given.
- Print the data set and give it an appropriate title.

PROC PRINT Output

Names and Ratings		
Obs	Name	Rating
1	J. Mitchell	5
2	Amy Jung	4
3	Carl Heisman	4
4	Linda Deal	3
5	Gabrielle Heron	4
6		2
7	Kyle Patterson	5

2. Converting U.S. Postal Codes to State Names

The data set **orion.contacts** contains a list of contacts for the U.S. charities that Orion Star donates to.

Partial **orion.contacts**

ID	Title	Name	Address1	Address2
AQI	Ms.	Farr,Sue	15 Harvey Rd.	Macon, GA 31298
CCI	Dr.	Cox,Kay B.	163 McNeil Pl.	Kern, CA 93280
CNI	Mr.	Mason,Ron	442 Glen Ave.	Miami, FL 33054
CS	Ms.	Ruth,G. H.	2491 Brady St.	Munger, MI 48747
CU	Prof.	Florentino,Helen-Ashe H.	PO Box 2253	Washington, DC 20018

- Create a new data set named **states** that includes the variables **ID** and **Name** as well as a new variable named **Location** that shows the full name in proper case for the state that the contact is based in.

Hint: **Address2** is 24 characters long and the last item in **Address2** is always the ZIP code. Look in the online Help for character functions that use ZIP codes as arguments.
- Print your results.

Partial PROC PRINT output (12 Total Observations)

ID	Name	Location
AQI	Farr,Sue	Georgia
CCI	Cox,Kay B.	California
CNI	Mason,Ron	Florida
CS	Ruth,G. H.	Michigan

CU	Florentino,Helen-Ashe H.	District of Columbia
----	--------------------------	----------------------