Assignment, 12 Manipulating Data part 2

**Exercise 1**. Extracting Characters Based on Position

The data set au\_salesforce has employee data for the Sales branch in Australia. (63 total observations, 9 total variables) it is in the STA5066 sub directory.

1. Use a PROC PRINT step to examine the data on au\_salesforce

2. Use the data set au\_salesforce in a DATA step to create a new data set work.codes. Work.codes will contain user ID codes for logging onto the intranet site. Each user ID will consist of the first letter of the first name, the final letter of the first name, and the first four letters of the last name. All these letters should be in lowercase and there should be no spaces in the user id.

Create four new variables FCode1, FCode2, LCode and User\_ID. These variables should contain:

Variable Name Value

FCode1 First letter of First\_Name

FCode2 Final letter of First\_Name

LCode First four letters of Last\_Name

User\_ID The new intranet ID (As described above)

3. Use a PROC PRINT step, printing only the variables First\_Name, FCode1, FCode2, Last\_Name, LCode, and User\_ID to verify that the results are correct.

**Exercise 2**. Extracting Characters Based on Position 2

The data set newcompetitors has data on competing retail stores that have recently opened up near existing locations of a company’s stores. It is in the STA5066 sub directory. A manager would like a data set containing just the small retail stores from these observations.

1. Use a PROC PRINT step to examine the data set newcompetitors.

Note that the first numeral in the value of ID indicates the size of the store a 1 indicates a small retail store.

The first numeral occurs at different positions in the ID, but will never occur in positions 1 or 2.

2. Use a DATA step to create a data set work.smallstores that contains only small retail stores. The values for the variable city should be in proper case. Hint: You might need to use two SUBSTR functions to extract the first numeral in ID.

3. Use a PROC PRINT step to examine the data in work.smallstores and validate your results.

**Exercise 3**. Converting US Postal Codes to State Names

The data set contacts contains a listing of contacts for U.S. charities a company donates to. The data set is in the STA5066 subdirectory.

1. Use a PROC CONTENTS step to examine the structure of the data set contacts. Note that Address2 is 24 characters long.

2. Use a PROC PRINT step to examine the data on the data set contacts. Note that the last item in the variable Address2 is always the zip code.

3. Use a DATA step to create a new data set named states that includes the variables ID and Name as well as a new variable Location that shows the full name in proper case for the state the contact is based in. Hint: Read the documentation for the zipname function.

4. Use a PROC PRINT step to examine and validate your results.

**Exercise 4**. Cleaning Up Text Data

The data set customers\_ex5 is in the STA5066 sub directory.

1. Use a PROC CONTENTS step to examine the structure of the data set customers\_ex5.

2. Use a PROC PRINT step to examine the first 15 observations on the data set customers\_ex5.

3. Use a DATA step to read customers\_ex5 and create a new data set work.names.

(a) The names data set should contain only three variables: New Name, Name, and Gender.

(b) The variable New Name should contain the customers’ full name in the following format:

Mr. John B. Smith

Ms. Jane Doe

(c) Male names should be preceded with the title Mr.

(d) Female names should be preceded with the title Ms.

4. Use a PROC PRINT step to examine and verify your results.

**Exercise 5**. Searching for and Replacing Character Values

The data set customers\_ex5 is in the STA5066 sub directory.

1. Use a PROC PRINT step to examine the data in the data set customers\_ex5.

Customers on the data set customers\_ex5 who are frequent purchasers have been tagged as Silver, Gold, or Platinum. This designation appears at the beginning of their Customer ID value. Due to updates in the way the company designates Customer ID values, the existing values need to be modified.

2. Use a DATA step to create three output data sets, work.silver, work.gold, and work.platinum.

(a) Any four digit string -00- in Customer\_ID should be replaced by -15- in the output data sets.

(b) Search Customer\_ID for the values Silver, Gold, and Platinum and output to the respective data set when found. (There should be 17 observations in work.silver, 2 in work.gold, and 5 in work.platinum.)

(c) Keep only the variables Customer\_ID, Name, and Country in all data sets.

3. Use PROC PRINT steps to print the three new data sets and verify that they are correct.

**Exercise 6**. Searching Character Values and Explicit Output

The data set employee\_donations contains information on charity contributions from a company’s employees. It is in the STA5066 sub directory.

1. Use a PROC PRINT step to examine the data in employee\_donations. Note that each employee has either one or two charities listed in the Recipients variable. Note that some charity names have a comma in them.

2. Use a DATA step reading employee\_donations to create a data set work.split.

(a) The data set work.split will have one observation for each combination of employee and charity to which they donated.

(b) Some employees made two contributions; therefore, they have two observations in the input data set. These employees will have a % character in the value of Recipients.

(c) Hint: Store the position where the % character was found in a variable PctLoc. This can make subsequent coding easier.

(d) Create a variable Charity with the name and percent contribution of the appropriate charity.

3. Use a PROC PRINT step to examine the contents of work.split and verify your results.

**Exercise 7**. Calculating Statistics and Rounding

The data set orders\_midyear (in the STA5066 sub directory) contains an observation for each customer, with the total retail value of the customer’s monthly orders for the first half of the year.

1. Use a PROC PRINT step to examine the data set orders\_midyear.

2. Use a DATA step to read orders midyear and create a data set work.sale\_stats. Use SAS descriptive statistics functions to create three new variables that summarize all months in which the customer placed an order.

(a) MonthAvg should contain the average, rounded to the nearest integer.

(b) MonthMax should contain the maximum.

(c) The variable MonthSum should contain the sum of values.

3. Use a PROC PRINT step to print the variables Customer\_ID, MonthAvg, MonthMax, and MonthSum on work.sale\_stats and verify your results.

**Exercise 8**. Calculating Statistics for Missing Median and Highest Values

The data set orders\_midyear (in the STA5066 sub directory) contains an observation for each customer, with the total retail value of the customer’s monthly orders for the first half of the year. The manager wants to look at information on the median order and the top two months orders, but only for frequent customers. Frequent customers are defined to be those who placed an order in at least five of the six months.

1. Use a PROC PRINT step to examine the data in orders midyear

2. Use a DATA step that reads the data set orders midyear and creates a data set work.freqcustomers containing the statistics requested.

3. Use a PROC PRINT step to examine the dataset work.freqcustomers and verify your results.

**Exercise 9**. Using the PUT and INPUT Functions

The data set shipped (in the STA5066 sub directory) contains details about each product shipped to one of a company’s retail outlets in 2007.

1. Use a PROC CONTENTS step to examine the descriptor portion of the data set shipped.

2. Use a PROC PRINT step to examine the data on the data set shipped.

3. An analyst at the company wrote the following SAS program to calculate the total price of the items shipped and create a comment that includes the ship date.

data shipping\_notes;

set prg2.shipped;

length Comment $ 21.;

Comment = cat(’Shipped on ’,Ship\_Date);

Total = Quantity \* Price;

run;

proc print data=shipping\_notes noobs;

format Total dollar7.2;

run;

(a) Run the program to verify that it does not give the correct values.

(b) Debug the program, making the changes necessary to give the correct results

**Exercise 10**. Changing a Variable Data Type

The data set US\_newhire (in the STA5066 subdirectory) contains information about newly hired employees.

1. Use a PROC CONTENTS step to examine the descriptor portion of US\_newhire

2. Use a PROC PRINT step to examine the data on US\_newhire.

3. Use a DATA step to read US\_newhire and create a new data set work.US\_converted that has the following changes:

(a) Remove the embedded dashes in ID.

(b) Convert ID to a numeric value.

(c) Convert Telephone to character and place a - between the third and fourth digits.

(d) Convert Birthday to a SAS date value.

4. Use a PROC CONTENTS step with work.US\_converted to verify that the variable types are correct.

5. Use a PROC PRINT step to verify that the work.US\_converted data set contains the correct information (incorporating the required changes).