SAS Exercise – Homework 1

1. The World Health Organization (WHO) collected data in countries across the world regarding the outbreak of swine flu cases and deaths in 2009. The data in the file SwineFlu2009.dat include counts per country by month during the epidemic. There are many variables in the raw data file with the following descriptions:

- By date: ID for sorting by first case date
- By continent: ID (X.YY) for sorting by first case date within a continent where X represents continent X, and YY represents the YYth country with the next first case
- Country
- Date of first case reported
- Number of cumulative cases reported on the first day of the month for April, May, June, July, and August (across the columns, respectively)
- Last reported cumulative number of cases reported to WHO as of August 9, 2009
- By date: ID for sorting by first death date
- By continent: ID (X.YY) for sorting by first death date within a continent where X represents continent X, and YY represents the YYth country with the next first death
- Date of first death
- Number of cumulative deaths reported on the first day of the month for May, June, July, August, September, October, November, and December (across the columns, respectively)

Your task is to create a corresponding SAS dataset by reading in this data file.

a. First, examine the raw data file SwineFlu2009.dat using Notepad.
   To open in notepad Method 1: Right-click on file, choose Open with → Other programs → Notepad.

To open in notepad Method 2: Open Notepad, From the menu: File → Open (change from .txt to all files to display all files).

b. Write a DATA step to read the file into SAS. Make sure that each variable is assigned a unique (and descriptive) name and is of the correct type – character or numeric.

c. After the dataset is created, locate the file in the work library through the explorer window in SAS. Double click on the dataset to view the data.

d. In SAS, dates can be stored as a special type of numeric data. Modify the DATA step to make sure that the dates are read in the correct SAS date format (not as character). (HINT: Use the correct "informat" in the input statement – e.g., informat firstdate ddmmyy8.;)

e. Create a permanent label for each variable based on the preceding descriptions. (HINT: Use the "LABEL" statement).

f. Print a report that describes the contents of the data set including the labels that you've created and other attributes of the variables. (HINT: Use PROC CONTENTS).

(Hint: Read pages 40-50 from the book "The Little SAS Book". In particular, you need to know common INFORMAT structures to answer this question. You don't need to buy the book. There are many websites where you can download the pdf version.)

2. A gourmet pizza restaurant is considering adding new toppings to its menu. Each month they survey 10 customers about their preferences for three different toppings. They want data on several different toppings, so they don't always ask about the same three toppings. Customers rate each topping on a scale of 1(would never order) to 5 (would order often). The restaurant wants to compute average ratings for all toppings, so the ratings variables need to be numeric. The raw data file Pizza.csv has variables for the respondent's ID, and the ratings for five different toppings: arugula, pine nuts, roasted butternut squash, shrimp, and grilled eggplant. The first two digits in the ID correspond to the month of the survey.

a. Examine the raw data file Pizza.csv and read it into SAS using the IMPORT procedure.

b. Print the data set (on the results screen). Print a report that describes the contents of the data set to make sure all the variables are the correct type.

c. Open the raw data file in a simple editor like WordPad and compare the data values to the output from part b) to make sure that they were read correctly into SAS. In a comment in your report, identify any problems with the SAS data set that cannot be resolved using the IMPORT procedure. Explain what is causing the problem.

   (Hint: You need to make sure the type of each variable is read correctly. This is explained in details in lecture video 1 of SAS Basics.)

d. Read the same raw data file, Pizza.csv, this time using a DATA step (instead of the IMPORT procedure). Be sure to resolve any issues identified above.

   (Hint: Using the hint given for part c, it should be clear why a DATA step is able to resolve the problem)

e. Create a new dataset with the average ratings for each topping.

   (Hint: Read pages 118-121 from the book "The Little SAS Book" regarding PROC MEANS.)

3. The new management of a local hotel decided to update their recently acquired (and very outdated) property by installing wireless Internet service for their guests. They are also considering updating their billing system because the method used by the previous owner seems faulty. In order to conduct a billing analysis, they would like some calculations about the guests who stayed with them during the first part of February (this was the first month after the change of ownership). The raw data file Hotel.dat contains variables with information on room number, number of guests, check-in month, day, year, check-out month, day, year, use of wireless Internet service, number of days of Internet use, room type, and room rate.

a. Examine the raw data file Hotel.dat and read it into SAS.

b. Create date variables for the check-in and check-out dates, and format them to display as readable dates.

(Hint

Step1: You need to combine three columns into one that looks like a date: for example 2 /7 /2014. You can use "CATX" function.

Step2: If you do step 1, you have a column that has the date. Now you need to let SAS know that this is actually a date. Note that SAS does not realize this by itself, as you have seen in question 1. You can combine "INPUT" function and a proper INFORMAT to let SAS recognize the date.

Step3: Dates are saved as numbers in SAS. To display them as dates in printed output, use a proper FORMAT structure.

c. Create a variable that calculates the subtotal as the room rate times the number of days in the stay, plus a per person rate ($10 per day for each person beyond one guest), plus an Internet service fee ($9.95 for a one-time activation and $4.95 per day of use).

(Hint1: You can subtract dates if they have been stored as SAS dates

Hint2: Since the per person and internet service rates are different for different observations, you can use IF-THEN statement to do the job efficiently.)

d. Create a variable that calculates the grand total as the subtotal plus sales tax at 7.75%. The result should be rounded to two decimal places.
e. View the resulting data set. In a comment in your report, state the value for the grand total for room 211.