

Homework 5

Due: Thursday, February 17th (before 11:59 pm PST) on GauchoSpace.

Instructions: For the General Knowledge Questions, answer as succinctly as possible. For the Programming Assignment, show a partial screenshot of your output. Submit your code separately as a SAS program file.

General Knowledge Questions

1. How many methods are there for combining data sets?
2. Which statement is used to combine data sets with at least one common variable and other unique variables?
3. Which statement is used to combine data sets with the same variables?
4. Suppose the data sets `ds1` and `ds2` are combined as follows: `set ds1 ds2;`
What is the order of the observations in the combined (output) dataset?
5. Which statement is used to specify the variable(s) used to link observations when merging data sets?
6. How is the `IN=` data set option used when merging data sets?
7. When merging data sets, the `RENAME=` option is *typically* used to create ____ variable names.
8. Name the different types of match merges.
9. What is a lookup table?
10. How can a lookup table be used to attach labels to the coded values in another data set?

Programming Assignment

The Daily Show, an award-winning television program, has had a diverse guest lineup in the last 26 seasons. The **Guests2014** data set (posted on GauchoSpace) contains the 2014 lineup of guests on the show, while the **Categories** data set (posted on GauchoSpace) contains the different occupation categories and their abbreviations.

- a) Browse the descriptor portions of the **Guests2014** and **Categories** data sets posted on GauchoSpace.

How many observations and variables are in each data set?

Guests2014 _____

Categories _____

What are the names of the variables in each data set?

Guests2014 _____

Categories _____

- b) Create a new data set called **DailyShow2014** that combines the **Guests2014** and **Categories** data sets. Rename variables, if necessary. This output data set should only contain observations created with information originating from both **Guests2014** and **Categories**. How many observations are in **DailyShow2014**? Is this the same number of observations as in **Guests2014**?
- c) For the **DailyShow2014** data set, sort the observations on occupation categories in the following order: **Acting, Music, Comedy, Media, Politics, Athletics, Government, Academia, and Other**. (Hint: Create a lookup table with a variable called **Order**.)
- d) Create a report using the **DailyShow2014** data set from part C. Display each group of **Order** values on separate pages. Set **Category** to be an ID variable. Suppress observation numbers, but display the number of observations at the end of each group of **Order**. Only display the variables **Guest** and **Occupation** (in this order). Display a title of '*Daily Show Guests (2014)*'.
- e) Browse the descriptor portion of the **DailyShow2014** data set from part C and of the **DailyShow2015** data set posted on Gauchospace. How many observations and variables are in each data set?

DailyShow2014 _____

DailyShow2015 _____

What are the names of the variables in each data set?

DailyShow2014 _____

DailyShow2015 _____

- f) Write a data step that combines **DailyShow2014** and **DailyShow2015** into a new data set called **DailyShowAll** in the **work** library. Rename variables, if necessary. This output data set should contain the variables **year** (the year of the guest appearance), **guestname**, **occupation**, and **category**. (Hint: The name of the input data set contains the **year** of the guests' appearance.)
- g) Browse the descriptor portion of **DailyShowAll**. How many variables and observations are there? Is this what you expected?

- h) Create a report using **DailyShowAll** where the guest names are sorted in alphabetical order. Display the variables in the order of **GuestName**, **Year**, **Occupation**, and **Category**. Add appropriate labels and a title. Suppress observation numbers but display the total number of observations at the end of the report.