**Stat 6740, Autumn Semester 2015**

**Homework 1**

1. In 1995, the American Statistical Association (ASA) held a data analysis exposition sponsored by the Statistical Graphics Section. For this exposition, they created two data sets regarding colleges and universities and asked statisticians to demonstrate techniques for data analysis and visualization using these two data sets. The two data sets were drawn from

* The US News and World Report 1994 census of US colleges and universities; and
* The American Association of University Professors (AAUP) 1994 salary and compensation survey.

The US News data set contains data from US colleges and university regarding student admission and enrollment, college costs, and graduation rates. The AAUP data set contains information about average salaries and total compensation across several categories of faculty. Additional details about the content and format of these data can be found at

* <http://www.amstat.org/publications/jse/datasets/usnews.txt> and
* <http://www.amstat.org/publications/jse/datasets/aaup.txt>;

Text files containing the data can be found on the class Carmen site under the names:

* aaup\_dat.txt and
* usnews\_dat.txt.

The format for the AAUP data is as follows:

|  |  |  |
| --- | --- | --- |
| **Line** | **Columns** | **Variable** |
| 1 | 1 - 5 | FICE (Federal ID number) |
| 1 | 7 - 37 | College name |
| 1 | 38 - 39 | State (postal code) |
| 1 | 40 - 43 | Type (I, IIA, or IIB) |
| 1 | 44 - 48 | Average salary - full professors |
| 1 | 49 - 52 | Average salary - associate professors |
| 1 | 53 - 56 | Average salary - assistant professors |
| 1 | 57 – 60 | Average salary - all ranks |
| 1 | 61 - 65 | Average compensation - full professors |
| 1 | 66 - 69 | Average compensation - associate professors |
| 1 | 70 – 73 | Average compensation - assistant professors |
| 1 | 74 – 78 | Average compensation - all ranks |
| 2 | 1 - 4 | Number of full professors |
| 2 | 5 - 8 | Number of associate professors |
| 2 | 9 - 12 | Number of assistant professors |
| 2 | 13 - 16 | Number of instructors |
| 2 | 17 - 21 | Number of faculty - all rank |

Write a SAS program that performs the following:

1. Read the AAUP data into SAS and create a local dataset called “salary.” Use appropriate variable names.
2. The Department of Education divides the states (and territories) into 10 regions as follows:

|  |  |
| --- | --- |
| **Region** | **States / Territories** |
| I | Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont |
| II | New Jersey, New York, Puerto Rico, Virgin Islands |
| III | Delaware, District of Columbia, Maryland, Pennsylvania, Virginia, West Virginia |
| IV | Alabama, Florida, Georgia, Kentucky, Mississippi, North Carolina, South Carolina, Tennessee |
| V | Illinois, Indiana, Michigan, Minnesota, Ohio, Wisconsin |
| VI | Arkansas, Louisiana, New Mexico, Oklahoma, Texas |
| VII | Iowa, Kansas, Missouri, Nebraska |
| VIII | Colorado, Montana, North Dakota, South Dakota, Utah, Wyoming |
| IX | Arizona, California, Hawaii, Nevada, American Samoa, Guam, Northern Mariana Islands |
| X | Alaska, Idaho, Oregon, Washington |

Create a new local dataset based on your name and place the dataset “salary” into it. Create a new variable in this new dataset that contains the region where the university is located (use IF/THEN/ELSE statements).

1. In the same local dataset created in Step 3, perform the following actions:
   1. Calculate a new variable that is the total salary paid to all full professors.
   2. Calculate new variables that are the percentage of faculty of each rank (full, associate, assistant, instructors).
   3. Calculate a variable that sums the counts of each faculty rank (that can be compared to the total faculty count variable.
   4. Calculate a new variable that is the ratio of average assistant salary to average associate salary.
2. Using the “WHERE” statement in PROC PRINT, print the following (separately):
   1. All variables for Region VII;
   2. Average compensation for each faculty rank for all universities of Type I;
   3. Total salary paid to all full professors for universities in Region X;
   4. Total faculty count and calculated total faculty count for every college in California (with college name and type); and
   5. All variables (original and calculated) for all universities in the state of Ohio that are of Type IIA.
3. Ashley Hart, a former graduate student researcher from The Department of Human Nutrition at The Ohio State University, examined the effects of dietary fiber on the processing of carotenoids in the human body. She designed a laboratory experiment that used a human digestive system simulation. She prepared samples combining fiber at several levels with several different foods and different enzyme strength, simulated the digestive process on each sample, and performed measurements on the sample after the simulated digestion. She used six different fibers (control = none, FOS, GOS, Pectin, RS2, RS3) at three levels (0, 2, 4), with four foods (no-food control, yogurt, raw bananas, ripe bananas), and two enzyme levels (1X, 2X). She repeated various combinations of these factors several times, with the number of replicates for the combinations varying between 3 and 11. Her measured variables of interest were the percent micelleration for each of three carotenoids (alpha-carotenoid – AC, beta-carotenoid – BC, and lutein).

The data from this study have been placed in an Excel spreadsheet named “Carotenoid Data.xlsx” on the Stat 6740 Carmen site.

Write a SAS program to do the following:

1. Read the data into SAS (choose your own dataset name).
2. Print a summary of the contents of the dataset.
3. Create a new local dataset by setting the first data set and do the following:
   1. Calculate differences in the carotenoid concentrations between each pair of the first three replicates (i.e., Rep 1 vs Rep 2, Rep 1 vs Rep 3, and Rep 2 vs Rep 3);
   2. Create a new variable that indicates which replicate has the highest concentration.
4. Using the “WHERE” statement, print the following:
   1. All non-concentration variables (i.e., variables that do not start with “Rep”) for all control samples (all variables);
   2. All lutein observations with “1X” enzyme level (all variables except enzyme);
   3. All replicate differences (from 3a) for samples that are not no-food controls (include fiber, level, food, enzyme, carotenoid, and differences);
   4. All observations where the maximum concentration appear after Rep 3 (include fiber, enzyme, level, food, carotenoid, and max replicate); and
   5. All observations with fiber = GOS and with more than 3 replicates (all variables).
5. The Ohio Department of Health is performing an analysis to determine risk factors for a lead childhood poisoning. They have collected disease data from a number of children throughout Ohio, but they do not have any individual risk data for each person. Using data from the U.S. Census Bureau, they can collect risk factor data within census tracts, which are clearly defined geographic areas within counties. In Franklin County, there are 284 census tracts.

Risk factor data for Franklin County have been obtained and placed into four SAS datasets named modelvars1.sas7bdat – modelvars4.sas7bdat that are located on the Carmen site for the class.

Write a SAS program that performs the following:

1. Import all four SAS datasets.
2. Print a summary of the contents of each dataset.
3. Create a new local dataset named “modelvars1\_ext’ and set modelvars1. Perform the following calculations:
   1. Create a variable that is the ratio of the number of persons less than or equal to 5 years old to the number of persons greater than or equal to 65 year old.
   2. Create a variable that is equal to the percentage of persons who are white (equal to 100 – the sum of the percentages for black, Asian, mixed, and Hispanic).
   3. Create a new variable that identifies the race with the largest percentage (use IF/THEN/ELSE).
   4. Calculate variables that are equal to the number of persons of each race.
4. Using “WHERE” print the following:
   1. The entire data sets for modelvars2, modelvars3 and modelvars4;
   2. All observations in modelvars1\_ext where the percentage of white people is less than 50 (include GEO\_id2, population count, percent mixed race, and number of mixed race);
   3. All observations where the percent of Hispanics is greater than the percent of blacks (include GEO\_id2, population count, percent Hispanic and percent black);
   4. All observations where either the percent of the population that is less than 5 or greater than 65 is greater than 20 (include GEO\_ID, percent 5 or younger, percent 65 and older, population size); and
   5. All observations where the percentage of children 5 or younger is greater than 10 and where the race with the largest percentage is black (all variables).