**STAT 6740**

**Homework 3**

1. Recall the colleges and university data project from HW1. Do the following for the
2. Read the AAUP and US News data, combine the two datasets, and add the Department of Education region to the entire combined dataset. Add the institution type from the HW2.
3. Create a cross-tabulation of college type by region, displaying the count and the percent of each type within region. Output the table results to a local dataset and print that dataset.
4. Create a bar chart that shows the number of missing values for lecturer salary for each type of institution.
5. Produce a scatterplot of average salary across all ranks versus the total undergraduate enrollment. Use different symbols for public and private schools.
6. Create a graphics-level histogram of the student-to-faculty ratio for each combination of institution type (use only U and C) and region, overlay a normal distribution curve, and place the mean, standard deviation, normality test statistic, and normality test p-value in an inset box.
7. For each region, calculate the mean institutional expenditure per student and combine it with every school in the region. Print this value, along with the region, school name, state, and actual institutional expenditure per student for all “N”-type institutions.
8. Submit the SAS program, SAS log, and SAS output.
9. Recall Ashley Hart’s data from Homework 1.
10. Read the data into SAS and change the value of the food variable for observations where it is blank to “No food”.
11. Create a separate observation for each replicate. Discard any observations where the percent micellization is missing.
12. Combine observations together with the same food/fiber/level/enzyme by creating new variables for AC, BC, and lutein percent micellizations.
13. Repeat steps 2 and 3 using PROC TRANSPOSE and confirm they have the same number of observations by performing a PROC MEANS for both datasets and displaying the number of observations.
14. Prepare scatterplots of the percent micellizations for every pair of carotenoids. Use a different symbol for each food.
15. Produce a chart that shows the number of observations for each fiber.
16. Produce basic summary statistics (MEANS) including mean, minimum, maximum, standard deviation, and count for each enzyme level for each carotenoid.
17. Submit your SAS program, SAS log, and SAS output.
18. Recall the Ohio Department of Health data from Homeworks 1and 2.
19. Import the four SAS datasets and merge the datasets by census tract number.
20. High levels of risk have been defined based on 4 variables: percent black (pct\_black), percent with a female head of household (pct\_f\_head), percent unemployed (pct\_unemploy), and percent of houses built before 1950 (pct\_pre\_50). Calculate and print summary statistics for each of these variables.
21. Create a scatterplot for each pair of these four risk factors.
22. A census tract is at high risk if the percent black is greater than 75%, the percent of households with female heads is greater than 25%, the percent unemployed is greater than 20%, and percent of houses built before1950 is greater than 33%. Create a new dataset with variables that indicate if each census tract is high risk based on each of the criteria (one variable for each criterion).
23. Do the following tasks with the data set created in Question 4:
    1. Print frequency tables for each risk indicator,
    2. Print cross-tabulations for each pair of risk factors,
    3. Create a variable that counts the total number of risk factors for each census tract, and create a frequency chart that shows the numbers of tracts with each risk factor count.
    4. Print out all data for any census tract with 3 or more risk factors.
24. Submit the SAS program, SAS log, and SAS output.