**STAT 6740**

**Homework 2**

1. Recall the colleges and university data project from HW1. The original format for the US News data is shown in the table below. This dataset has been converted to a CSV file named usnews.csv.

|  |  |  |
| --- | --- | --- |
| **Line** | **Columns** | **Variable** |
| 1 | 1 – 5 | FICE (Federal ID number) |
| 1 | 7 – 51 | College name |
| 1 | 53 – 54 | State (postal code) |
| 2 | 1 – 2 | Public/private indicator (public=1, private=2) |
| 2 | 3 – 6 | Average Math SAT score |
| 2 | 7 – 10 | Average Verbal SAT score |
| 2 | 11 – 15 | Average Combined SAT score |
| 2 | 16 – 18 | Average ACT score |
| 2 | 19 – 22 | First quartile - Math SAT |
| 2 | 23 – 26 | Third quartile - Math SAT |
| 2 | 27 – 30 | First quartile - Verbal SAT |
| 2 | 31 – 34 | Third quartile - Verbal SAT |
| 2 | 35 – 37 | First quartile - ACT |
| 2 | 38 – 40 | Third quartile - ACT |
| 2 | 41 – 46 | Number of applications received |
| 2 | 47 – 52 | Number of applicants accepted |
| 2 | 53 – 57 | Number of new students enrolled |
| 2 | 58 – 61 | Pct. new students from top 10% of H.S. class |
| 2 | 62 – 65 | Pct. new students from top 25% of H.S. class |
| 3 | 1 – 6 | Number of full-time undergraduates |
| 3 | 7 – 12 | Number of part-time undergraduates |
| 3 | 13 – 18 | In-state tuition |
| 3 | 19 – 24 | Out-of-state tuition |
| 3 | 25 – 29 | Room and board costs |
| 3 | 30 – 34 | Room costs |
| 3 | 35 – 39 | Board costs |
| 3 | 40 – 44 | Additional fees |
| 3 | 45 – 49 | Estimated book costs |
| 3 | 50 – 54 | Estimated personal spending |
| 3 | 55 – 58 | Pct. of faculty with Ph.D.'s |
| 3 | 59 – 62 | Pct. of faculty with terminal degree |
| 3 | 62 – 67 | Student/faculty ratio |
| 3 | 68 – 70 | Pct. alumni who donate |
| 3 | 71 – 76 | Instructional expenditure per student |
| 3 | 77 - 80 | Graduation rate |

1. Read the AAUP data and US News data into separate datasets. Add the Department of Education region to the AAUP data.
2. Combine the two datasets from Step 1 into a single dataset with one record per institution. Note that the names of the institutions may differ between the two datasets and take that into consideration when you are combining the data.
3. Create a new variable called “Inst\_type” that takes the value “C” if the name of the institution contains “College” or “Coll”, “U” if the name of the institution contains “University” or “Univ”, and “N” if it contains neither. Print a list of all the “N” institutions.
4. Create 6 new datasets that contain the results for each combination of college type (Type I, Type IIA, and Type IIB) and public/private indicator, one dataset for each combination. Remove any variables that are quartile measurements of SAT or ACT scores and alumni donations from all six datasets.
5. Run PROC CONTENTS for the dataset for public Type IIA institutions.
6. For each dataset, calculate the following:
   1. the percentage of applicant that were accepted;
   2. the percentage of accepted applicants that enrolled;
   3. the difference between total cost of attendance for in-state students (which includes in-state tuition, room and board, additional fees, and book costs) and instructional expenditure per student.
7. For private colleges of Types I and IIB, print the DOE region, state, college name, number of full-time undergraduates, total number of faculty (all ranks), percentage of applicants accepted, percentage of accepted applicant enrolled, difference between cost of attendance and expenditures, and graduation rate.
8. Output the public Type IIB data to an external SAS dataset. Include only the following variables: region, state, college name, average salary (all ranks), total faculty (all ranks), total full-time attendance, student/faculty ratio.
9. Output the data from public Type I colleges to a CSV file. Keep only the observations from the region that includes The Ohio State University.
10. Submit the SAS program, SAS log, SAS output, SAS dataset, and CSV file.
11. Recall Ashley Hart’s data from Homework 1.
12. Read the data into SAS.
13. Change the value of the food variable for observations where it is blank to “No food”.
14. Sort the data by food, then fiber, and finally by level.
15. Combine fiber and food into a single variable that has the format “fiber – food”.
16. Define a new variable that represents a random replicate using a random uniform number generator. Make the probabilities for each replicate be 50% for Rep 1, 30% for Rep 2, and 20% for Rep 3.
17. Define a new variable that is equal to the percent micelleration associated with the random replicate selected in the previous step.
18. Print the following:
    1. All micelleration values for control – raw banana samples;
    2. All data from observations where the random replicate selected in Step 5 is Rep 3;
    3. All data from observations where the percent micelleration for the random replicate from Step 6 is less than 25;
    4. All data from non-food for carotenoids whose percent micelleration values for Rep 1 are less than the random percent micelleration from Item 6.
19. Output a separate SAS dataset for each carotenoid.
20. Submit your SAS program, SAS log, SAS output, and all SAS datasets created in Step 7.

1. Recall the Ohio Department of Health data from HW 1.

1. Import the four SAS datasets.
2. Print a summary of the contents of each dataset.
3. Merge the datasets by census tract number (geo\_id2).
4. The geographic identifier, geo\_id2, is an 11-digit number where the first two digits are the state FIPS number, the next 3 digits are the county FIPS number, and the last 6 digits are the census tract number within the county. Create separate variables for state FIPS, county FIPS, and census tract number using geo\_id2.
5. Calculate the following variables:
   1. Ratio of percent Hispanic to percent unemployed;
   2. Difference between the percent of people with high-school and college educations;
   3. Number of vacant houses;
   4. Number of houses built after 1950.
6. Print the following data:
   1. All data for census tracts with numbers ≥ 10000;
   2. All data for census tracts where the percent of vacant homes is < 5%;
   3. All data for census tracts where the percent of households that include a married couple is less than 10%;
   4. All data for census tracts where the percent of people of Asian heritage is greater than the percent of people of mixed heritage.
7. Create a permanent SAS dataset that has the following variables: state FIPS, county FIPS, census tract number, percent black, percent with a HS education, pct unemployed, and percent vacant homes.
8. The file delaware4.sas7bdat contains the same variables as model4vars for Delaware County. Combine these two datasets together.
9. Output a STATA dataset with the data from Step 9.
10. Submit the SAS program, SAS log, SAS output, and STATA dataset.