Assignment 4: Due on Thursday 3rd December before 3.00 pm (PDT)

Note: For each question, produce separate PDF files for SAS code and output. Those PDF files should be uploaded to the Crowdmark under each question.

Q1) Use the provided FISH data set and do the followings.

a) Sort the data by lake type and dam status (LT and DAM). You won't need the latitude or longitude variables for this assignment, so you can drop them.

b) Use PROC MEANS to calculate the mean and median for each of the remaining numeric variables (all nine of them) separately for each combination of LT and DAM. Use ODS to save the results to a data set named STATZ that keeps only the necessary variables.

c) Combine these summary statistics with the original dataset (FISH) using an appropriate join technique to create a data set named ALL.

d) After the data sets are joined, but in the same data step, do the following. You must use arrays to get credit for this portion of the assignment.

- For each of the nine variables that have a mean and median computed, you also need to correctly compute the following for each of the 9 variables. (That means no warnings or errors - syntax or otherwise!. Hint: you may have to use if conditions in this computation for missing values.)
  - Difference from the mean [variable - mean.of.variable]
  - Percent difference from the mean [(variable - mean.of.variable)/mean.of.variable]
  - Difference from the median [variable - median.of.variable]
  - Percent difference from the median [(variable - median.of.variable)/median.of.variable]
- Apply reasonable formats and labels to all the variables that weren't originally in the FISH data set.
- Drop any irrelevant variables.

e) Sort your ALL data set by NAME.

Q2) Create Frequency Reports using SAS dataset called ‘sanfran’ as follows.

a) Use PROC FREQ to create a report using the **sanfran** data set that displays the frequency count for each **Destinatation**, and a separate frequency count for each **DepartDay**. Add an appropriate title to each frequency table.
b) Modify the program from part ‘a’ above to repeat the frequency tables, but without the cumulative frequencies.
c) Use PROC FREQ to create a two-way frequency table using the **sanfran** data set that displays the frequency count for each **Destination** by **DepartDay** combination.

Q3)

a) **Validating Data with PROC FREQ**: PROC FREQ is useful in checking the validity and completeness of data (i.e., invalid values stand out). Use PROC FREQ to check the validity of the variables **Gender** and **JobCode** in the **mechanics** data set.

   i. What do you notice about the values of the variable **Gender**? (Comment)
   ii. What do you notice about the values of the variable **JobCode**? (Comment)

   Modify the previous report to display the frequency count for each **Gender** by **JobCode**. What are the **JobCode** values for the invalid values of **Gender**?

b) **Creating Basic Summary Reports:**

   i. Generate a PROC MEANS report using the **sanfran** data set as input to display statistics for the variables **CargoRev** and **TotPassCap** only. Remove any titles currently in effect.
   ii. Modify the previous report to display the data for each **Destination.** Include the following statistics (number of observations, mean, median, mode, range and standard deviation. Limit all output to two decimal places.

Q4)

a) Datasets **ONE** and **TWO** are shown here. Use PROC SQL to create a new, temporary SAS data set (Both) containing Subj, Height, Weight, and Salary. Do these three ways: first, include only those subjects who are in both data sets, second, include all subjects from both data sets, and third, include only those subjects who are in data set **ONE**.

Dataset **ONE**:

Subj	Height	Weight
001	68	155
002	75	20
003	65	99
005	79	266
006	70	190
009	61	122

Dataset **TWO**:

Subj	Salary
001	$46,000
003	$67,900
004	$28,200
005	$98,202
006	$88,000
007	$57,200

b) Use PROC SQL to create a new temporary SAS data set (Percentages) from the Blood dataset (attached), containing the variables Subj, RBC, WBC, MeanRBC, MeanWBC, Percent_RBC, and Percent_WBC. The first few observations in the output shou look like this:

Subj	RBC	WBC	Mean RBC	Mean WBC	Percent_ RBC	Percent_ WBC
1	7.40	7710	5.884	6532.5	125.765	118.025
2	4.70	6560	5.884	6532.5	79.878	100.421
3	7.53	5690	5.884	6532.5	127.974	87.103
4	6.85	6680	5.884	6532.5	116.417	102.258
5	7.72	.	5.884	6532.5	131.203	.