

Updated March 7, 2019

The George Washington University
Stat 6197 – Spring 2019
Take-Home Assignment 1

Due: March 22, 2019

Introduction

For this programming assignment, you will use the rodent data file that [THE CARPENTRIES](#) use for teaching their data management and programming workshops. This is a simplified version of the *Portal Project Database*. This new version is designed to teach many aspects of data management and analysis. The original data were used in various publications that documented “the effects of the experimental manipulations as well as the response of populations and communities to long-term changes in climate and habitat”. See the abstract of the [article](#) by Ernst et al, 2009.

Data

Click the “Download all” button” on the top of [this page](#). Unzip the downloaded file, and you will find eleven files in it. For this assignment, you will need only three files - the rodent data file (**surveys.csv**) and two auxiliary files (**species.csv**, and **plots.csv**).

The rodent data file includes the individual-level data for each animal caught in 24 plots in the study area. The data fields are briefly described below:

Data field	Description
Record_ID	Unique ID for the observation
Month	Month of observation
Day	Day of observation
Year	Year of observation
Plot_ID	ID of a particular plot (See the code description in plots.csv). Add suffix 1 through 8 to the plot type Control (i.e., Control 1 through Control 8)
Species_ID	2-letter code (See the code description in species.csv)
Sex	Sex of animal (“M”, “F”)
Hindfoot_length	Length of the hindfoot in mm
Weight	Weight of the animal in grams

Programming tasks:

Total possible points: 100

- 1) (_1Read_Data.sas): Write and run the SAS code to do the following:
 - a. Read the raw data file **surveys.csv** into SAS, create a permanent SAS data set (say **surveys.sas7bdat**), and then store it in a SAS library.
 - b. Run an appropriate PROC STEP on **surveys.sas7bdat** and generate 1 X 2 tabular output for the total number of observations and the total number of variables.
 - c. Run an appropriate PROC STEP on **surveys.sas7bdat** to generate the list of variables and their attributes.
- 2) (_2Create_Format_plot.sas): Create a user-defined, permanent format for the **Plot_ID** variable based on the information included in the **Plots.csv** data file. Hints: Use the VALUE statement in PROC FORMAT.
- 3) (_3Create_Format_species.sas): Create a user-defined, permanent format for the **Species_ID** variable based on a control data set (i.e., a SAS data set to be created from the **Species.csv** data file). Hints: Use an appropriate option in PROC FORMAT.
- 4) (_4Apply_Format_Function.sas): Use an appropriate SAS procedure and formats (you created for items 2 and 3) or a SAS function to generate the following counts:
 - a. number of records in the **surveys.sas7bdat** data set with missing values for each variable except record_id;
 - b. number of records with nonmissing values for all variables in the **surveys.sas7bdat** data set
- 5) Write SAS code to produce means for the **Hindfoot_length** variable, overall and by the formatted **plot_id** variable, using PROC SUMMARY as well as PROC SQL.