STAT 6430: Statistical Computing for Data Science

Project Instructions

Introduction: Part of the duties of the Analytical Consulting Lab, a unit of at an unspecified university, is to handle small- to moderate-sized consulting projects, which arrive continually throughout the year. There are three consultants who work for the Lab, Ms. Smith, Mr. Jones, and Ms. Brown; these are the people who actually meet with clients, offer advice, and, for larger projects, may help to analyze data, to plan and implement a research study, or to develop internet resources or other software. There is also a Lab Coordinator (LC), whose duties are to assign projects to individual consultants as they come in, and to keep track of the time the consultants spent on each project. The purpose of this assignment is to help the LC to compile workload information from this past year and produce a summary report of the consultants' activity.

Record-keeping in the Lab operates as follows. Every work day, each consultant keeps track of the amount of time he or she spends working on a project. At the end of every work day, each consultant fills out a work form for the projects on which he or she worked that day, and submits the form to the LC. The work form requests several pieces of information: the current day's date, the total number of hours spent on the project that day (within fifteen minutes), the project number, the type of activity, and an indicator of whether the project is completed. The project number is a number that the LC assigns to a project at the time that the initial consulting request was made by the client. The type of activity is recorded as a numerical code for one of five categories describing the typical stages of a project. These activity classifications are coded so that:

- Stage 1 is an "initial consultation" with the client;
- Stage 2 represents the "planning" stage of a research study or other task, which may involve numerous meeting with clients and others involved in the project;
- Stage 3 is the "implementation and analysis" stage, during which the main task of the project is carried out;
- Stage 4 is an "interpretation and reporting" stage, which may be as extensive as writing a final report, or as simple as meeting with the client to go over results.
- The project is "closed" once the "interpretation and reporting" stage is completed, but sometimes the client requests a "follow-up" appointment. Stage 5 represents the time spent on such an appointment. (There can be at most one follow-up appointment, as any additional time needed would be treated as a new project and assigned a new project number.)

The activity classifications are treated somewhat cavalierly, in the sense that the LC requests that each entire day's work on a project be classified into one category or another. (*I.e.*, a consultant will never submit two forms for the same project on the same day, representing two different stages of the project). Also, the sequence of stages of any project is always the same: stage 1 is followed by 2, which is followed by, 3, *etc*. Either or both of stages 2 and 3 may be skipped, however, and stage 5 may not occur.

In addition, the LC has been experimenting in an attempt to classify entire projects into one of four "types," each of which suggests a typical course of a project (*e.g.*, the work hours required and duration) and the type of tasks the consultant will be asked to carry out. The four project types are:

- "Study coordination," the most extensive type of project, in which the consultant in essence coordinates an entire experiment, survey, or research study;
- "Information technology and web development," another extensive project type, in which the consultant directs the development of analytical software;
- "Data analysis," in which the consultant carries out basic data analysis of the results of a research study already carried out; and
- "Advising," which is a catch-all project type for the smallest-scale projects, and may require only a consultant's expert advice on a topic.

Data Sets: Five SAS data sets have been provided by the LC for this project. To access them, the files provided should be placed in a folder associated with a SAS Library.

- The SAS Data Set **`master`** is a master record of the consultants' activity since the beginning of the calendar year. Each observation corresponds to an individual work form, and records values of the following variables: DATE records the date that the work was carried out and the form submitted; PROJNUM and ASSIGN records the number of the project and the name of the consultant assigned to it; HOURS records the number of hours spent on the project that day; ACTIVITY records the activity code for the work carried out that day; and CLOSED records a 1 if the project has been completed or if the type of activity is a "follow up" (in which case the activity code is 5), and 0 otherwise. Note that several work forms may be submitted on the same day, even by the same consultant, but not for the same project.

  This information in this data set becomes fairly sparse after September 1, as the LC has not inserted the information from the work forms submitted since then, which has only just recently been put in an electronic format.

  This data set is provided in the file "master. sas7bdat."

- The SAS Data Set **`newforms`** stores the information from new work forms that have been submitted since September 1 and do not already appear in the **`master`** data set. The layout of this data set is the same as that of the **`master`** data set, except **`newforms`** does not include the variable ASSIGN, which records the consultant-to-project assignments.

  This data set is provided in the file "newforms. sas7bdat."

- The SAS Data Set **`correct`** stores a list of corrections, which are to be applied to the work form information stored in the data sets **`master`** and **`newforms`**. The format matches that of **`master`** and **`newforms`**, but includes only the variables DATE, PROJNUM, HOURS, and ACTIVITY. Corrections are provided only for the variables HOURS and ACTIVITY. (One should assume that the remaining variables DATE, PROJNUM, ASSIGN, and CLOSED are error-free, in this and the other data sets.) Missing values in the data set **`correct`** are used only as place holders, and should never be treated as a “correction” to another value.

Note: The data set **`correct`** does not account for every error in the **`master`** and **`newforms`** data sets. For instance, even after the corrections are applied, you may notice some missing values still in the variables HOURS and ACTIVITY.

This data set is provided in the file “correct. sas7bdat.”

- The SAS Data Set **`assign`** stores the list of consultant-to-project assignments made since September 1. (The remaining assignments are indicated in the data set **`master`**.) Each observation corresponds to a single project (with a unique project number) and stores the variables PROJNUM and ASSIGN, which are described above.

  This data set is provided in the file “assign. sas7bdat.”

- The SAS Data Set **`type`** stores the list of project types, as determined by the LC. As with the data set **`assign`**, each observation corresponds to a single project, but here the variables PROJNUM and TYPE are stored. The variable PROJNUM is as in the other data sets. The variable TYPE is a character variable that stores brief descriptions of the project type, of which there are four possible values.

  This data set is provided in the file “type. sas7bdat.”

Part A: The first task is to combine the information stored in the five data sets provided by the LC into a single, new master data set. This data set is to be completely up to date, error-free to the extent that corrections are available, and is to include the available project-type information.

To do this you will need to combine the **`master`** and **`newforms`** data sets, somehow filling in the correct consultant assignments (indicated in the **`master`** and **`assign`** data sets). Similarly, the project-type information in the **`type`** data set is to be included in the new master data set, and the corrections noted in the data set **`correct`** are to be applied. (The steps you take to complete this task need not follow the order just stated.)

Notes and additional instructions for Part A:

- To keep track of corrections, you should include in the new master data set a variable, named CORRECTED, which takes the value 1 if any value of that observation has been corrected and 0 otherwise.

- Name the new master data set **`newmaster`**, and save it as a permanent SAS Data Set, which you are to turn in. It should store the variables DATE, PROJNUM, ASSIGN, HOURS, ACTIVITY, CLOSED, TYPE, and CORRECTED, which are defined as in the data sets provided and are to have the same length, format, and label associations. The one exception is the variable CORRECTED, which does not appear in any of the data sets provided; it should be assigned the label "Data Corrected?"

Part B: The next task is to generate several reports summarizing the consulting activity of the Lab at the current date. (The "current date" should be taken as November 4, 2010, the last date listed in the data set **`newmaster`**, which you created in **Part A**.) For reference (and to clarify the correct labels and formatting), printouts of the SAS Output Window from a similar report done at around this time in a previous year are provided in the file "ACL Report NOV09.txt."

(i.) The first report is to be a list of projects that remain ongoing. By "ongoing," this means that the indicator variable CLOSED has not taken the value 1 as of the current date.

To generate this report, first compile the list of project numbers of ongoing projects into a SAS Data Set, which is to store just one variable, PROJNUM. (This variable is defined as in the date sets provided and is to have the same length, format, and label associations.) This data set should be named **`ongoing`**, and should be stored as a permanent SAS Data Set, which you are to turn in.

Next, use PROC PRINT to list the information stored in that data set in the SAS Results Viewer (declaring the LABEL option so that the labels appear), and use a TITLE statement to give the report an appropriate heading.

(ii.) The next report is to summarize the consulting activity of each consultant on each project since the start of the year. You are actually to generate three reports, one for each consultant, each having the same format and listing the same type of information. The report of an individual consultant should list the project numbers of all the projects to which he or she has been assigned (identifying the rows of a table); with each project number there should appear (as columns in the table) the project type, the total hours spent on the project, an indicator of whether the project is complete or ongoing, the start date of the project, and the "finish" data of the project, the latter of which interpreted as the date at which the last work form associated with the project was turned in. These quantities are all to be calculated as of the current date.

To generate these reports, first compile the information of each consultant in a separate SAS Data Set, which is to store the variables PROJNUM, TYPE, TOTHOURS, ONGOING, STDATE, and FNDATE. Each observation is to correspond to one of the projects assigned to the consultant. The variables

PROJNUM and TYPE are defined as in the data sets provided and are to have the same length, format, and label associations. The variable TOTHOURS is to store the total hours spent on the project. The variable ONGOING is to take the (numeric) value 1 if a project is ongoing, and 0 otherwise. The variables STDATE and FNDATE are to store the start and finish dates of a project, and should be associated with the "DATE9." format. These last variables should be respectively assigned with the labels "Total Hours," "Ongoing?," "Start Date," and "Finish Date" (Formats and labels should be made permanent as part of the "descriptor" portion of the data set.)

The data set compiling Ms. Smith's activity should be named **`smith`**, that compiling Mr. Jones's activity should be named **`jones`**, and that compiling Ms. Brown's activity should be named **`brown`**. Each should be stored as a permanent SAS Data Set, which you are to turn in.

Next, use PROC PRINT to list the information stored in each data set in the SAS Results Viewer (declaring the LABEL option so that the labels appear), and use a TITLE statement to give each report an appropriate heading. The variables are to be listed out in the following order: PROJNUM, TYPE, TOTHOURS, STDATE, FNDATE, and ONGOING.

(iii.) The last report is to summarize the overall consulting activity of every consultant since the start of the year. This report is to list three rows of a table, one for each consultant. Within each row there should appear the total number of projects the consultant was assigned, the total number of hours spent on all projects (whether complete or ongoing), the average number of hours spent on all projects that are completed as of the current date (*i.e.*, ONGOING = 0 in the relevant previous data set), and the minimum and maximum number of hours spent on any single project that was completed as of the current date.

To generate this report, first compile the information in a separate date set, named **`overall`**, which should be stored as a permanent SAS Data Set that you are to turn in. The data set should store the variables ASSIGN, TOTPROJ, TOTHOURS, AVGHOURS, MINHOURS, and MAXHOURS. Each observation is to correspond to one of the consultants. The variable ASSIGN is defined as in the data sets provided, and is to have the same length, format, and label association. The variable TOTPROJ is to store the total number of projects assigned to the consultant, and the variables AVGHOURS, MINHOURS, and MAXHOURS are to store the average, minimum, and maximum hours spent on the **completed** projects. The new variables should be respectively assigned with the labels "Total Projects," "Average Hours," "Minimum Hours," and "Maximum Hours" (which should be made permanent as part of the "descriptor" portion of the data set).

Next, use PROC PRINT to list the information stored in this data set in the SAS Results Viewer (declaring the LABEL option so that the labels appear), and use a TITLE statement to give the report an appropriate heading. The variables are to be listed out in the following order: ASSIGN, TOTPROJ, TOTHOURS, AVGHOURS, MINHOURS, and MAXHOURS.

Notes and additional instructions for the general project:

- *Working in groups:* You are to work on this project groups of three or four students. If you choose to work as a group, turn in one complete solution, prepared by all members of the group. Each member will receive the same grade. Also, you should turn in a (hard copy) written statement describing the contributions of each member of the group, signed by both members.

- *Allowed resources:* The project is open-book open-notes. You are not to talk to anyone outside of your group about the project, except the professor. You should not use any resources other than those that have been made available by the professor. The allowed resources include the current course notes, problem set instructions and solutions, the suggested references (which are listed on the course policy statement), and the online documentation that is built in to the SAS user interface. The allowed resources do not include materials from other courses, materials from past versions of this course, or materials downloaded from the internet.

- *The basic task:* The task of this project is to write a SAS Program. Write this program in SAS's Enhanced Editor Window, placing all parts in the same document and in the correct order. If you have completed the project correctly, it should be possible to load this document into the Enhanced Editor, submit it as a single batch of code, and it should run with no errors (assuming that the required external files are stored in the correct directory).

- *Let SAS do the work:* Avoid manually typing in any of the information provided in the data sets. For instance, in **Part A**, the consultant assignments and project types are to be looked up and filled in automatically by your program. Some reasonable exceptions are allowed. For instance if upper or lower bounds on the subscript of an array depend on values in the data, it is OK to look up the required bounds and manually type them into your program. (As a general rule, if you think you can get SAS to do something automatically, you should try hard to do that.)

- *Titles and comments:* Use comments to create a header for your program, which should indicate your name (or the names of those in your group), the date, that the program is your solution to the STAT 6430 Project, and any other comments you want to make. Also, include appropriate TITLE statements and comments throughout the program so as to annotate your output and help organize your work.

- *External directory references:* **Part A** requires you to assign a SAS Library to an external directory and read data from existing SAS Data Sets. This assignment is made with a LIBNAME statement, and you should place that statement somewhere near the very beginning of your program.

- *What to turn in:* Turn in both electronic and hard copies of your program (as a ".sas" file), an electronic version of the SAS Results Viewer listing of activity reports generated by PROC PRINT from **Part B** (in HTML or text format, as can be specified using the "Save As" menu item), and final versions of the SAS Data Sets **`newmaster`** from **Part A** and **`ongoing`**, **`smith`**, **`jones`**, **`brown`**, and **`overall`**, from **Part B** (as ".sas7bdat" files).

  Put all of your files into a single compressed (.zip) folder, and upload it to the "Assignments" tool in UVaCollab. (Click the assignment "Project," and follow the instructions.) Name the folder in such a way that identifies the names of people in your group. For instance, if I am working with Jeff Holt (jjh2b) and Ange Yin (hy4dr), I might name it "Project_DJS4Y_JJH2B_HY4DR.zip." Also print out a hard copy of your program and turn it in. I only need a hard copy of your program. Do not print out anything that appears in the Results Viewer or Log Window when you run your program.

- The data created for this project were generated randomly. Resemblance of any patterns in the data to real-world phenomena is completely coincidental.