Final Project - Data Management Competition

STAT 440 - Statistical Data Management
Summer 2018

The focus of the final exam, which is a final project consisting of a written report and SAS program, is applying data management approaches we have covered during the semester to prepare a thoroughly clean and validated dataset for competition. These approaches include reading raw data files checking data for errors, validating and cleaning data, creating formats and labels, deriving new variables, array creation, iterative (DO-loop) processing, subsetting data, joining/merging/appending, and conditional output. Each week of the semester (excluding Week 8), the Instructor will provide a dataset that is related to the IRI drug store sales data. It is up to you (the student) to work with these datasets becoming familiar with them and investigating how they may be used together for the competition. Check the Data Competition folder for more information about the IRI drug stores sales data and for the weekly added datasets. Each student will manage the IRI drug store sales data set to compete in one of the following categories:

1. **Most Relevant Variables**: the prepared dataset that has the largest number of relevant variables. Relevant variables are those that are deemed appropriate, related to the subject of the observations, in correspondence with the other variables/measurements, and sensible given the nature of the original data.

2. **Least Missing Values**: the prepared dataset that has the smallest number of missing values among all variables in the dataset. False, student-defined recoding, or imputed values are not allowed.

3. **Largest Number of Consistent Stores**: the prepared dataset that has the most stores that appear every week of the year 2001 along with the most relevant variables for those observations.

4. **Overall Best Prepared Dataset**: the prepared dataset that utilizes all of the Instructor's provided datasets as well as other datasets, which is the most data-rich. This data set will look clean, be easy to query, and contain the most useful information related to college students' spending.

For each category, there will be two winners (undergraduate and graduate student) for a total of 8 winners. You can only select one of these categories to compete in. The final project has a maximum total of 50 points, for which all students in the class have the

ability to earn 50 points on the final project. This means that each student could earn $(50/50) \cdot 20 = 20$ points contributed to your final grade. These 8 students that win the competition's categories will receive 5 bonus points earning them $(55/50) \cdot 20 = 22$ points contributed to your final grade.

Each student's final report should be between 3-5 pages double-spaced. You must write in complete sentences and pay attention to grammar, spelling, and readability. If you include a table or chart, make sure you say something about it. Do not place charts or figures in your report without discussing them. You need to provide thorough description and narration about the steps you are taking, the subsequent decisions you made during the data management process, an explanation of your results, and a concise conclusion.

Each student will submit the final report (.pdf) and SAS program (.sas) in Compass by **11:50 PM on August 4, 2018** and will count for 20% of the final grade. *Late submissions will not be accepted.* The final report should at least include the following:

1. Title of the project

2. Category of the data competition you want to be considered for

3. Introduction section
   - Background information and description of the Instructor-proved datasets and any other datasets you used
   - Explain why you are considering the category for the competition

4. Methods section
   - Description/explanation of the guidelines used to validate the data
   - Description/explanation of the issues that needed cleaning and you did the that
   - Description/explanation of additional data preparation that you performed (e.g. merging, joining, subsetting)

5. Results section
   - Charts and tables pertaining to validation and cleaning.
   - Summary of the results while noting important information from the charts and tables

6. Conclusion section
   - Persuade the audience to believe why your prepared dataset is going to win the category of the competition

Final Project Grading Rubric

The grading criteria for the final project are listed below. The maximum total points for the final project is 50 points.

- SAS program (.sas)

  – 10 points: the efficacy of the whole SAS program

  – 10 points: the data management aspect of the code

  – 5 points: the organization and logical flow of the code

- Summary Report (.pdf)

  – 10 points: the data preparation techniques

  – 5 points: the output and results in the report

  – 5 points: the value of the information in the report and how that information aligns with the data management process

  – 5 points: the organization and readability of the report

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder