JAMES COOK UNIVERSITY
AUSTRALIA

MA5831 Advanced Data Processing and Analysis using SAS

Assignment 2 Content

Name:	Data quality profiling and standardising
Type:	Case study
Issued:	8:00 PM AEST Monday of Week 1
Due:	11:59 PM AEST Sunday of Week 4
Weight:	20%
Length:	1500 words +/-10%

Submission method: PDF or MS Word submission to LearnJCU

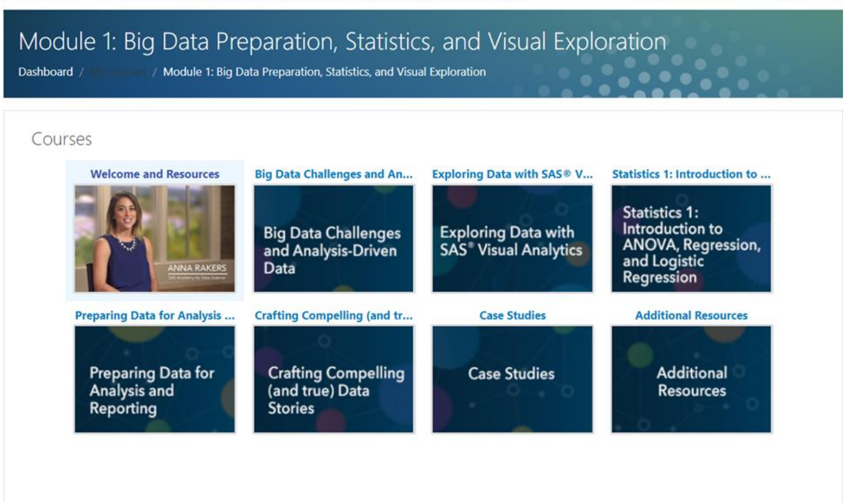
Overview

This case study reinforces the concepts that you learned in Weeks 1, 2 and 3, and the practical work you carried out in the SAS course ‘Preparing Data for Analysis and Reporting’. Specifically, it reinforces profiling data for anomalies, viewing the results of the profiling, creating standardisation schemes to correct the data and creating a job to execute the standardisation schemes.

It will use the DataFlux Data Management Platform in the SAS Academy for Data Science virtual learning environment (VLE) and a dataset relating to the performance of first-year students in the College of Engineering at a particular college in the United States. The student data can be used to determine whether there are students in certain demographics that are at risk for failure. You are asked to determine the usefulness of the data for this analysis.

The case study, along with a video tutorial on data setup, can be found on the SAS Academy for Data Science in the Case Studies section of Module 1: Big Data Preparation, Statistics, and Visual Exploration:

- Case Study: Preparing Data for Analysis and Reporting (pdf)


Module 1: Big Data Preparation, Statistics, and Visual Exploration
Dashboard / My courses / Module 1: Big Data Preparation, Statistics, and Visual Exploration
Courses
Welcome and Resources
ANNA RAKERS
Big Data Challenges and An...
Big Data Challenges and Analysis-Driven Data
Exploring Data with SAS® V...
Exploring Data with SAS® Visual Analytics
Statistics 1: Introduction to ...
Statistics 1: Introduction to ANOVA, Regression, and Logistic Regression
Preparing Data for Analysis ...
Preparing Data for Analysis and Reporting
Crafting Compelling (and tr...
Crafting Compelling (and true) Data Stories
Case Studies
Case Studies
Additional Resources
Additional Resources

Learning outcomes

This assessment evaluates your ability to:

1. Critically appreciate the causes of data quality
2. Define and create a data collection, a data exploration, data profiles and standardisation schemes
3. Create and execute the basic options for data jobs including monitoring of events.

Work-based skills

This assessment relates to the following work-based skills:

1. Ability to work with standard data: preparing it, cleaning it and transforming it for analytics and data science using SAS DataFlux
2. Critically appreciate the use of standardisation schemes and their application in a real-world scenario
3. Ensuring that the quality, and the consistency, and the accuracy, and the completeness of the data to be used are sufficient for data science projects.

The case study

The problem

The chancellor of a U.S. college has asked for analysis from the data science team that summarises the performance of first-year students in the College of Engineering. The data provided can be used to determine whether there are students in certain demographics that are at risk for failure. An attempt will be made to assist those students with the resources that are necessary for the successful completion of the degree program. As a member of the data quality team, you will use the DataFlux Data Management Studio to discover inconsistencies in the data, and create DataFlux jobs to improve the consistency and reliability of the data that is to be used for analysis and reporting.

Desired outcome

The cohort of interest in this study is the first-year students within the college, both new students (freshmen) and new transfers. You need to ensure that the data is appropriate for the reports to be delivered to the chancellor. Specifically, report on the following items:

- The demographics of the first-year students with a GPA of 2.5 or lower
- The quality and consistency of the data to be used for future reports and analysis

In addition, you need to ensure that you have all necessary information for the analysis.

Your analysis should include data quality metrics on any and all fields that might be used for these reports, including the key or ID field to be used for the report or analysis. You should create data cleansing jobs to ensure the consistency of the data for reporting and, if necessary, create new data elements for analysis.

The data

There is a step-by-step video guide to connecting to the data in this case study section overview in the SAS Academy for Data Science using the DataFlux Data Management Platform. The College of Engineering collected data about first-year students for the fall semester of 2009. This data is located in the following location on your server:

- D:\Workshop\CASESTUDY

The data table contains the following details:

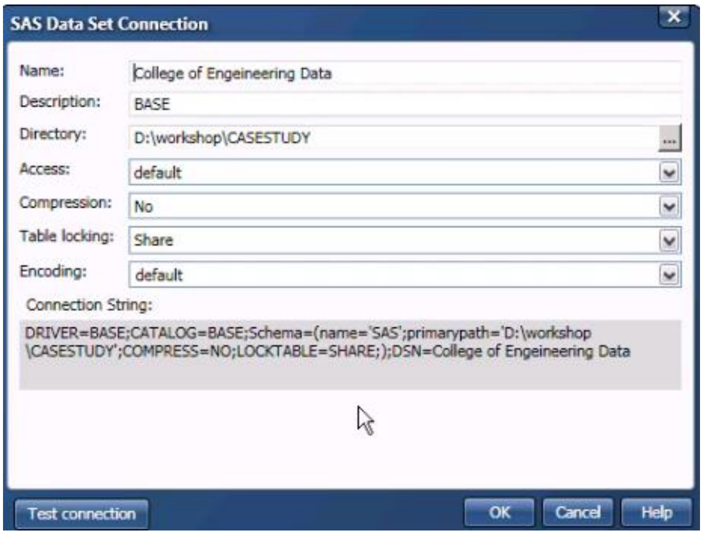
- Demographic information about each student (gender, student type, residency status and geographical location)
- Information about their high school careers (rank, test scores and classes taken)
- Information about their college careers (GPA, academic program and academic standing) at two time points:
    - The beginning of the semester (census on the 11th day of the semester)
    - The end of the term.

More information about the variables in the FIRST_YR_COE table can be found in the *COE_Data Dictionary.xlsx* spreadsheet, which is in the same directory as the data tables in the DataFlux Data Management Platform.

The tasks

Technical preparation and setup

The case study assignment will be completed using the SAS Academy for Data Science with the preloaded dataset available in the VLE (D:\Workshop\CASESTUDY). Connect to this SAS dataset using the DataFlux Data Management Platform.



Name: College of Engeineering Data

Description: BASE

Directory: D:\workshop\CASESTUDY

Access: default

Compression: No

Table locking: Share

Encoding: default

Connection String:

```
DRIVER=BASE;CATALOG=BASE;Schema=(name='SAS';primarypath='D:\workshop
\CASESTUDY';COMPRESS=NO;LOCKTABLE=SHARE;);DSN=College of Engeineering Data
```

For further details about the SAS Virtual Lab, please refer to 'Accessing SAS® Software Using the Virtual Lab Reservation System' (pdf) in the SAS Academy for Data Science.

Data processing tasks

Use DataFlux Data Management Studio to perform the following tasks:

1. Explore the data to become familiar with the data elements.
2. Profile the data to ensure that the data is appropriate for analysis and reporting.
3. Identify what types of data cleansing techniques could be used to correct any inconsistencies that are observed in the data.

Specifically, you need to perform the following steps:

1. Register data sources for the two data tables in DataFlux Data Management Studio
2. Create an exploration to identify the types of data contained in the table, as well as create any collections that might be useful in the future.
3. Create a profile that analyses the demographic and high school information for at-risk first-year students, to determine whether it is appropriate for analysis:
    a. You might want to create a filter on only the necessary data for the profile prior to processing, because the profile could fill up the memory on your Virtual Lab machine and fail.
    b. Pay close attention to the metrics that you calculate for each column, because some are very process intensive and some metrics are not applicable to certain data elements.
4. Create charts and graphs from the profile report that outline the overall cleanliness of the data.
5. If time permits, build jobs to improve the quality and consistency of the data. Be sure to save the target tables for later use.

Some helpful tips:

1. Run an exploration of the data table in DataFlux Data Management Studio:
    a. Perform a column-name analysis for any columns that you do not recognise based on the name.

    b. Perform a sample data analysis on the first 500 rows of the data columns that you want to use in reports and analysis.
2. Run a profile of the data to ensure that the data is appropriate for analysis and reporting:
    a. Use the appropriate metrics based on the type of data in the column that is analysed.

    b. If you build jobs to improve the quality of the data, rerun the profile report so that you have a history of the metrics (for reporting purposes).
3. Identify the types of processes that can be used to improve the quality and consistency of the data.

Reporting tasks

Your report needs to answer the following questions based on the outcomes of your data analysis:

1. What is the overall quality of the available data?
2. What data elements need to be cleansed before you proceed with analysis and reporting?
3. What outliers exist in the numeric fields?
4. Do the columns of demographic data contain the expected data values?
5. Are the fields containing demographic data suitable for analysis and reporting? If not, what changes need to be made to the data?
6. How many students are identified as at-risk?
7. Are there any notable demographic patterns that exist among the at-risk students?
8. What processes should be used to improve the quality of the data to be used in reports?
9. Which columns require data standardisation?
10. Which columns need the case changed for consistency?
11. Are there any records where entity resolution might help?
12. Are there any applications where parsing can add to the value of the data?
13. Do you have full data for the relevant variables for analysis and reporting? If not, what additional information might be helpful for analysing the data?

Assessment criteria

Written responses, with supporting output evidence, for the 12 questions in the Reporting tasks section submitted to LearnJCU. See marking rubric for further details (20%).

Examples of supporting evidence, including but not limited to:

- Evidence of creation of visual exploration of the data (screenshots of DataFlux)
- Evidence of creation of data cleansing jobs to ensure the consistency of the data for analysis and reporting (screenshots of DataFlux)
- Evidence provided that new data elements where created as part of the data preparation (screenshots of DataFlux)
- Clear identification of the types of processes that can be used to improve the quality and consistency of the data.

Submission guidelines

Your submission for Assessment 2 should be uploaded to LearnJCU as two (2) files:

1. The task cover sheet
2. Completed answers for the 13 questions in the Reporting tasks section and saved in the following format A2_questions_firstname_lastname (PDF or DOCX format)
    - Length: 1500 words +/-10%
    - Screenshots of the assessment profiles used to answer the questions with labelling can be included in the submission if they help with the overall answer to the related question.

Additional information

The Case Studies in the SAS Academy for Data Science are different from the exercises in the courses.

- The case studies are selected because they ***do not have a clean solution***.
- The datasets are ***messy*** (and real!), and the business problem is framed by a domain expert who knows ***very little*** about data management, analytics or programming.
- This enables you to experience ***true client interactions*** where you, the data scientist, are the technical expert and the client ***lacks the vocabulary*** to guide you to appropriate analyses.

If you find the case studies frustrating to work on, take heart in knowing that you are doing ***realistic*** data science work. Use the case study review sessions to observe the thought process of an experienced analyst.

In data science, solving problems often requires taking a novel approach, making assumptions, engineering new inputs or combining analyses in unconventional ways.

- It is rare that one solution is uniformly correct!
- It is rare that a problem has a clean solution.

Seeing the approaches of different analysts helps you find new ways to solve problems in the future.

Marking criteria: MA5831 Assessment 2 – Case study: Data quality profiling and standardising

Criteria	High Distinction	Distinction	Credit	Pass	Fail
Criterion 1: Student has an understanding of the overall data quality, cleansing, consistency, outliers, demographics, standardisation and suitability for analysis and reporting.	Student demonstrates an excellent understanding of the overall data quality, cleansing, consistency, outliers, demographics, standardisation and suitability for analysis and reporting in their answers to the case study questions.	Student demonstrates a strong understanding of the overall data quality, cleansing, consistency, outliers, demographics, standardisation and suitability for analysis and reporting in their answers to the case study questions.	Student demonstrates a developing understanding of the overall data quality, cleansing, consistency, outliers, demographics, standardisation and suitability for analysis and reporting in their answers to the case study questions.	Student demonstrates a basic understanding of the overall data quality, cleansing, consistency, outliers, demographics, standardisation and suitability for analysis and reporting in their answers to the case study questions.	Student fails to demonstrate an understanding of the overall data quality, cleansing, consistency, outliers, demographics, standardisation and suitability for analysis and reporting in their answers to the case study questions.
Criterion 2: Student applies and integrates DataFlux Data Management Studio effectively.	Demonstrates independent investigation, examines in detail and provides strong evidence to support their answer for each case study question using DataFlux Data Management Studio.	Evidence of independent investigation, examines in detail and provides evidence to support their answer for each case study question using DataFlux Data Management Studio.	Some evidence of independent investigation, examines and provides evidence to support their answer for each case study question using DataFlux Data Management Studio.	Very limited evidence of independent investigation and provides limited evidence to support their answer for each case study question using DataFlux Data Management Studio.	No evidence of independent investigation and no evidence to support their answer for each case study question using DataFlux Data Management Studio.
Criterion 3: Student demonstrates quality analytical writing skills using appropriate conventions.	Demonstrates sophisticated analytical writing and synthesis skills in answers to the case study questions, which are logically structured and written with clarity. No grammatical errors, good use of paragraphs to structure answers and document.	Demonstrates competent analytical writing in answers to the case study questions, which are logically structured and written with clarity. Few grammatical errors, good use of paragraphs to structure answers and document.	Evidence of developing analytical writing in answers to the case study questions with the beginnings of a logically structured flow and written with clarity. Some grammatical errors, poor sentence structure and poor use of paragraphs to structure document.	Very limited evidence of analytical writing in answers to the case study questions with poor structure and flow. Some grammatical errors, poor sentence structure and poor use of paragraphs to structure documents.	No evidence of analytical writing in answers to the case study questions, which include grammatical errors, poor sentence and poor use of paragraphs to structure documents. Answers are poorly structured, do not flow easily and do not address the case study question. Does not exhibit good writing skills.