



# FIT5196 Data wrangling S2 2017

Dashboard ► Faculty of Information Technology ► Active Category ► 2017 ► FIT S2 2017 ►  
FIT5196-S2-2017 ► My Assessments ► Assessment 3: Data Integration and Reshaping

## Assessment 3: Data Integration and Reshaping

Please carefully review all the requirements below to ensure you have a good understanding of what is required for your assessment.

1. **Due Date**
2. **Instructions & Brief**
3. **Assessment Resources**
4. **Assessment Criteria**
  1. **Grading Rubric**
  2. **Penalties**
5. **How to Submit**

### 1. Due Date

This specific due date and time can be **viewed below in the Grading Summary**.

### 2. Assessment description

This assessment focuses on data integration and reshaping. You will be required to integrate data that might be collected from different sources. You will need to resolve different levels of conflicts in integration according to what we have discussed in the lectures. The output of this assessment should be an integrated dataset and a JuPyteR Notebook containing information about your designed global schema and all the Python scripts used in integrating the data. The detailed tasks are as follows:

#### Integration between S1 and S2:

In this part, you are required to write Python code to integrate data from two different data sources, indicated by S1 and S2 respectively, and generated one unified table (i.e., one Pandas DataFrame). The two datasets are

- S1 (data\_s1\_ass4.csv): contains 4,600 records
- S2 (data\_s2\_ass4.xml): contains 430 records

In order to finish the integration task, you will need to resolve schema conflicts (For example, structure conflicts, naming conflicts, and/or entity resolution conflicts) and data conflicts. **The output of the task should be a unified data stored in a global schema that should be justified with a clear explanation of the semantic mapping between local schemas** used in S1 and S2 and the global schema. While integrating the two datasets, you might need to add new columns, delete columns and merge columns. All the methods you used to identify and resolve the conflicts must be justified in your notebook.

#### Data Normalization

In this task, you will need to apply z-score normalization, Min-Max normalization and log transformation to the prices, and analyze how they affect the distribution of the data.

### 3. Assessment Resources

Before you start writing your code, you will need to read the following materials:

- While using Jupyter Notebook, you should consider the use of different cells to make notes as you go. However, be precise and concise, please do not include things that are not relevant to the tasks in your notebook.
- Standards for commenting code are available online (e.g. <https://www.python.org/dev/peps/pep-0008/#comments>). You must ensure that you can clearly explain what your code does with clear comments.
- The work required to finish this assessment should be your own. If you use resources elsewhere, make sure that you properly acknowledge them in your notebook. You may need to review the FIT citation style tutorial to make you're familiar with appropriate citing and referencing for this assessment. Also, review the demystifying citing and referencing for help.

---

### 4. Assessment Criteria

The following outlines the criteria which you will be assessed against.

#### 4.1 Grading Rubric

Specific grading details for this assessment are:

- Programming and results: (80%)
  - The submitted code should work without any errors and produces the required results correctly.
  - All the errors should be identified and resolved.
- Quality of code, commenting and notebook: (20%)
  - The code should be well structured and properly commented. A high-quality code is clean, explicit, consistent, efficient, easy to read and **well use of existing libraries**
  - The methods used in each task should be well justified in the Jupyter notebook.
  - The notebook should be structured in a logical way so that it clearly shows how students finish the tasks in the assessment.

Detailed rubric can be found [here](#)

#### 4.2 Penalties

- Late submission: for all assessment items handed in after the official due date, and without an approved extension, a 5% penalty applies to the student's mark for each day after the due date (including weekends, and public holidays) for up to 10 days. **Assessment items handed in after 10 days will not be considered!**
- Submission: please do follow **Section 5 How to Submit** to submit your assignment. Otherwise, a 5% penalty will be applied.

---

### 5. How to Submit

Once you have completed your work, take the following steps to submit your work.

1. Only two files to be submitted:
  1. A **Jupyter Notebook file**, whose extension must be “.ipynb”. The file name must be in the following format: **surname\_studentID\_ass4.ipynb**.
  2. The **CSV file** generated by your jupyter notebook, which contains the cleaned data. The file name should be in the following format: **surname\_studentID\_ass4\_data.csv**

3. Two files should be uploaded individually, rather than in a zip file.
2. Click the **Add Submission** button below to submit and upload your assignment. **Please do remember to accept the submission statement! Only the submitted assignments will be marked. Those shown as a draft won't be marked!**

If you need further guidance on how to submit an assessment item, please review the **Submitting an assessment** overview. If you need further assistance, please go to **Help & Support**.

## Submission status

Attempt number	This is attempt 1.
Submission status	No attempt
Grading status	Not graded
Due date	Sunday, 29 October 2017, 11:55 PM
Time remaining	13 days 14 hours
Last modified	-
Submission comments	► Comments (0)

Add submission

Make changes to your submission

### NAVIGATION



Dashboard

■ Site home

Site pages

Current unit

FIT5196-S2-2017

Participants

Unit Updates

Unit Information

Consultation Times


Forums

My Assessments

 Academic Integrity Tutorials

 Student Academic Integrity Policy

 Assessment 1: Text Preprocessing

 Assessment 2: Parsing and Cleansing Raw Data


 **Assessment 3: Data Integration and Reshaping**

 Assessment 1 Patent Data

 Assessment 2 Raw Data

 Assessment 3 Data Set 1

 Assessment 3 Data Set 2

 [How to submit an assignment in Moodle](#)

 [Notebook template for assessment 1](#)

 [Assessment 2 Rubric](#)

 [Assessment 3 Rubric](#)

 [Assessments Summary](#)

[Week 1 \(24 July - 30 July\) Introduction to Data Wr...](#)

[Week 2 \(31 July - 6 Aug\) Introduction to Regular E...](#)

[Week 3 \(7 Aug - 13 Aug\) Text Data Preprocessing](#)

[Week 4 \(14 Aug - 20 Aug\) Text Data Preprocessing](#)

[Week 5 \(21 Aug - 27 Aug\) Handling Raw Data in Diff...](#)

[Week 6 \(28 Aug - 3 Sept\) Data Cleansing](#)

[Week 7 \(4 Sept - 10 Sept\) Data Cleansing](#)

[Week 8 \(11 Sept - 17 Sept\) Data Cleansing](#)

[Week 9 \(18 Sept - 24 Sept\) Data Integration](#)

[Mid-Semester Break \(25 Sept - 1 Oct\)](#)

[Week 10 \(2 Oct - 8 Oct\) Data Integration](#)

[Week 11 \(9 Oct - 15 Oct\) Data Enrichment-reshapin...](#)

[Week 12 \(16 Oct - 22 Oct\) Data Enrichment-reshapin...](#)

[My units](#)



## Jump to Content



## My.Monash



## IT Academic Integrity



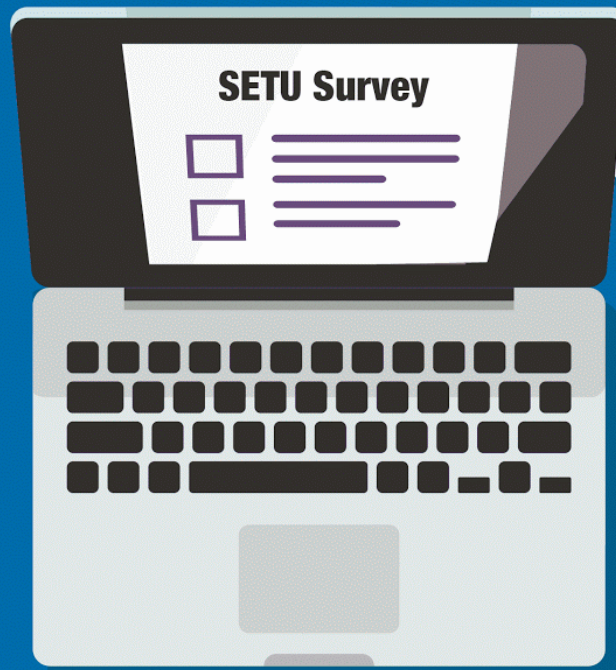
## IT Student Portal

### ADMINISTRATION



[Unit administration](#)

**Click here**



**Complete your  
SETU Survey by  
Sunday, 29th Oct  
to influence  
change!**

## LOGGED IN USER



**Chao-Kai Hsu**

Country: Australia

City/town: Melbourne

chsu0002@student.monash.edu

Copyright © 2014 Monash University - These course materials are for your research and study only. Further reproduction, transmission or sale without permission is prohibited - Privacy - CRICOS Provider Number: 00008C

You are logged in as Chao-Kai Hsu (Log out)

FIT5196-S2-2017