Deakin University
School of IT
Unit Team: SIT742

Trimester 2, 2025
Assignment 2

SIT742: Modern Data Science

Group Assignment This is a group assignment, and each group should have no more than **3** members. If you want to work on this assignment individually, please contact the unit chair in Week 6 (no later than 5:30pm on 22/08/2025 (Friday)) by email for approval. If you want to form a group, please use the `CloudDeakin` discussion form "`Find a Group Member for Assessment Task 2`" to find group members, and once you have formed a group, please self-enrol your group in the `CloudDeakin` group tool. For those approved individual students, you need to self-enrol in the group in the range $120 - 150$.

Extension Request Students with difficulty in meeting the deadline because of various reasons, must apply for an assignment extension no later than 5:30pm on 15/09/2025 (Monday). Apply via '`CloudDeakin`', the menu item '`Extension Request`' under the '`Assessment`' drop-down menu.

Academic Integrity All assignment will be checked for plagiarism, and any academic misconduct will be reported to unit chair and university.

Generative AI Deakin's Policy and advices on responsible usage of Generative AI in your studies: `https://www.deakin.edu.au/students/study-support/study-resources/artificial-intelligence`

Instructions

Assignment Questions

There are total **2** main parts and **1** optional part in the assessment task:

Part 1 The first part will focus on the data manipulation and pyspark skills which includes the `Data Acquisition`, the `Data Wrangling`, the `EDA` and `Spark`, the `modules and library` from **M03, M04**.

Part 2 The second part focus on more advanced data science skills with particular scenario. This part will require the knowledge covered in **M05**.

Part 3 **(Optional)** This part is **optional** and carries **0** mark.

What to Submit?

Part 1 and Part 2 are compulsory for this assessment task. You (your group) are required to submit the following completed files to the corresponding *Assignment* (Dropbox) in *CloudDeakin*:

`SIT742Task2.ipynb` The completed notebook with `all the run-able code` for all requirements (part 1 and part2).

In general, you (group) need to complete, **save** the results of running, download/export the notebook as a local file, and submit your **notebook** from Python platform such as `Google Colab`. You need to clearly list the answer for each question, and the expected format from your notebook will be like in Figure 1 (**One Notebook** for each group).

Figure 1: Notebook Format

`SIT742Task2Report.pdf` You (group) are also required to write a report with your answer (code) and running results from `SIT742Task2.ipynb` for all the questions (`Part 1` and `Part 2`). You could make screenshot on your answer (code) and running results from `SIT742Task2.ipynb` and paste into the report. Please try to include the code comments, and results including plot images as well in the report, and make sure the code format such as Indentation keeps same as the ipynb notebook. For questions that require analysis in a markdown cell, please ensure those explanations are also included in your report.

In this report (**one for each group**), you will also need to provide a clear explanation on your logic for solving each question (you could write explanation below your solution and results in the report). In the explanation, you will need to cover below parts: 1) why you decide to choose your solution; 2) are there any other solutions that could solve the question; 3) whether your solution is the optimal or not? why? The length of this part of explanation with above for each question is limited below 120 words.

In the end of your report, you (group) also need to discuss below three points in a paragraph:

- How you and your team member collaborate on this assignment?
- What you have learned with your team member from the second assignment.
- What is the contribution of each the team member for finishing the second assignment.

`SIT742Task2Video.avi` A video demonstration between 10 and 15 minutes, and the file format can be any common video format, such as '`MKV`', '`WMV`', '`MOV`' etc.

For your group, one important submission is a **short video** in which each of *You* (group members) orally present the solutions that you provide in the notebook and illustrate the running of code with the used logic (for all the questions (`Part 1` and `Part 2`)). In the video, your group need to work together to discuss the following three points:

- Which question(s) you have worked on and how did you collaborate with other team members.
- What is the logic behind your solution on the question(s)? and is there any alternative optimized ways to resolve the question?

- What is your understanding of `Code collaboration`? How do you collaborate with your group in coding? What are the common tools/platform to support the `Code collaboration`?

Part I
Data Acquisition and Manipulation

There are **8** questions in this part, totalling **60** marks. Each of question is worth **5** marks. Additionally, the quality of your explanation in both the report and video will collectively be worth **20** marks.

You are recommended to use `Google Colab` to finish all the coding in the *code block cell*, and provide sufficient coding comments, and also save the result of running as well.

The (`business_review_submission.zip`) data used for this part could be found in **here**. There are two files in the data. The first one is about the business review submission with many companies. For each of the row, the review submission is provided with relevant information such as `userid`,`time`,`name` and many others. The second one is the meta information of the business and the two data could be joined with `gmap_id`. You will need to use `spark` to **first read the unzipped (csv) review data for starting** and later join the meta review business data on dataframe (pandas or spark). You could find the code on reading csv data with Spark from M04G. In some of the tasks, if the question is not specifically asking to use `spark`, you could use both `pandas` and `numpy`.

Question 1.1

Using PySpark to do some data wrangling process, so that:

1.1.1 For the `none` or `null` in `text` column, change it to `'no review'`.

1.1.2 Process the content in `time` column, and convert the strings from `time` to `yyyy-mm-dd` format in the new column as `newtime` and show the first 5 rows.

Question 1.2

Find out the information for `gmap_id` on the reviews. In order to achieve the above, some wrangling work is required to be done:

1.2.1 Using `pyspark` to calculate the number of reviews per each unique `gmap_id` and save as float format in pyspark dataframe to show the top 5 rows.

1.2.2 Transform the current pyspark dataframe to `pandas` dataframe (named as `df`) and create the column `reivew_time` with the information of review time on hours level. Print your `df` pandas dataframe with top 5 rows after creating the column `review_time`.

1.2.3 Using matplotlib or seaborn to draw some (two or more if possible) visualizations on the relationship between `gmap_id` and `reivew_time`. You could explore for example, what is the time people usually review? How many business is reviewed in the morning time etc. Please also discuss the insights you are finding with your visualizations in the markdown cell. Please also include your findings and visualizations in the report.

Question 1.3

Let's continue to analyze the `reivew_time` with reviews and related `gmap_id`. You need to use another data `meta-business` to join with the current dataframe on `gmap_id`.

1.3.1 Determine which `workday` (day of the week), generates the most reviews (plotting the results in a line chart with workday on averaged submissions).

1.3.2 Identify the names of business (column `name` from data `meta-business`) that has the highest averaged ratings on '`that workday`' (you need to find out from 1.3.1), and find out which category those businesses are from?

1.3.3 Please further explore the data on name of business and find out some more insights by yourself such as which category it is and what are the peak hours etc. Please use visualizations and tables to support your findings and write down the insights in the markdown cell. Please also include your findings and visualizations in the report.

Question 1.4

For the reviews on each of the submissions, work on all the review content and find out the top 30 most common words; Also generate separate word cloud visualizations for different years by grouping the reviews by review year and write down the insights in the markdown cell. Please also include your findings and visualizations in the report.

Question 1.5

Let's do some analysis on the `business_name` and the reviewers. Determine the number of unique reviewers of business and its categories to identify which business / category has attracted the most reviewers (find out the highest distinct count of reviewers on business / category level). Also, analyze the temporal patterns of when reviewers submitted their reviews (you could leverage the workday, year, month, or hours to conduct the analysis) and share your findings and insights in the markdown cell. Please also include your findings and insights (visualizations) in the report.

Question 1.6

As the data scientist, you are required to build a recommendation for the business by using reviews, ratings, and its categories. In this task, you need to:

1.6.1 Write down your strategy of building the recommendation on business for customers in the markdown cell. You could create your own strategy or leverage the provided one here `KNN` on collaborative filtering. Please also include your strategy details in the report.

1.6.2 Could you please try to implement the strategy (code) you have written down for the recommendation system? Please give detailed explanation of your code and the logic in the comments and also interpret the recommendations with examples in the markdown cell. Please also include your implementation details and results in the report.

Question 1.7

Continue work on the review data you have now, for each of the submissions of the review, you will need to explore the rating with other information:

1.7.1 Build visualization to explore the relationships of the rating and business categories. Please write down your insights in the markdown cell and also include your insights and visualizations in the report.

1.7.2 Let's focus on the lower ratings now. Could you please find out the actual reviews on lower ratings and analyze on the reason? (You could use the common used words in lower rating reviews or design your own strategy with reasonable logic). Please also include your analysis details in the report.

Question 1.8

Continue to work on the submission of the reviews, we would like to focus on the reviewer level with all the reviewed business history, but before we actually conduct the programming, we will need to finish few questions for exploration:

1.8.1 Check on the reviewer level reviewed business, sort the review of each business by the review time (`newtime` column) and then save the business name into the list variable `user_business_list` for each reviewer.

1.8.2 Check on the `user_business_list`, could you observe some repeated business names for the same user? If so, could you remove those duplicated business names under same user? Please print out the number of element in the `user_business_list` for each reviewer before removing the duplicated business name and after removing the duplicated business name.

1.8.3 Check on the `user_business_list`, could you find the user similarities according to their past reviewed business ? You are free to design your own strategy and give sufficient explanation in markdown cell and code implementation together. Please also include your strategy details and implementation in the report.

Hint: you might consider to use `encoding` for each of the business names and then calculate the difference of the users.

Part II
Submission Prediction

There are **3** questions in this part, totaling **40** marks. Each question is worth **10** marks. Additionally, the quality of your explanation in both the report and video will collectively be worth **10** marks.

You are required to use `Google Colab` to finish all the coding in the *code block cell*, and provide sufficient coding comments, and also save the result of running as well.

Question 2.1

In this question, we will focus only on two information: total `reviews` per day with `review time` (`newtime` from the dataframe) to form the review volume time series. You are required to explore the review time series. There are some days not available in the review time series. Please add those days into the review time series with default number of review with the mean value of the number of review per day in the whole data (without any filtering on reviews). After that, decompose the submission review time series with addictive mode and analyses on the results to find if there is any seasonality pattern (you could leverage the `M05A` material from lab session with default setting in `seasonal_decompose` function). Please also include your analysis details and implementation in the report.

Question 2.2

We will try to use time series model `ARIMA` for forecasting the future. You need to find the best model with different parameters on `ARIMA` model. The parameter range for p,d,q are all from `[0, 1, 2]`. In total, you need to find out the best model with lowest `Mean Absolute Error` from 27 choices (you might need to split the time series to train and test with yourself with grid search according to the `M05B` material). Also, you are required to discuss with your group member on exploring the deep learning time series forecasting methods such as `LSTM` and `RNN`. Please write down your discussion around the necessary data wrangling and modeling steps (steps on how to achieve, not actual code). Also please give the reference of the deep learning time series forecasting models you are using. Please also include your discussion details and implementation in the report.

Question 2.3

In this question, you are provided with the PDF file by *Universities Australia* via Indigenous Strategy annual report. You are required to critically analyze this report using your data science skills.

Data Extraction Carefully review the PDF and identify all relevant quantitative data, tables, and figures that can be extracted or digitized; Present any extracted data in a structured format (e.g., CSV, Excel table, or DataFrame);

Data Analysis Utilize your data analytics skills to discover common patterns or trends from the report; Where possible, compare trends over multiple years, between institutions, or across different Indigenous strategy metrics.

Insights Provide a clear and concise summary of the main patterns, trends, or correlations discovered from your analysis; Interpret what these findings reveal about the progress and challenges of Indigenous strategies in Australian universities.

You may use any data analytics tools or libraries you are comfortable with. All steps, from extraction to insights, should be clearly documented in your `SIT742Task2Report.pdf`, and source code should be in `SIT742Task2Code.ipynb`.

Part III

Optional: Questionnaire on *Integrating Indigenous Perspectives into DS Education*

Note: This part of the assignment is **optional** and carries **no mark**.

In modern data-driven environments, collaboration often involves individuals from diverse cultural and social backgrounds. As part of our ongoing efforts to develop inclusive and socially responsible unit, we are exploring ways to meaningfully integrate Australian Indigenous perspectives into Data Science education, specifically in the context of the unit *SIT742: Modern Data Science*. We invite you to share your thoughts and feedback on this important topic. Your input will help us improve the quality and inclusiveness of both the unit and its associated learning materials.

If you are willing to assist, please consider completing a short questionnaire titled *Integrating Indigenous Perspectives into Data Science Education*, which can be accessed at the following link:

- POST-Survey Questionnaire

(Please interpret its references to "`ICT`" as also including *Data Science*.)

Related Information

Human Research Ethics Application ID: `2025/HE000518`

Approval Date: `28/04/2025`

Estimated Time to Complete: Approximately 10 minutes

Confidentiality: All responses are strictly confidential and will be used solely for research and educational improvement purposes.

University Policy Deakin Indigenous Strategy 2023-2028