FIT5137 S2 2024 Assignment 3: PTV Assignment Scenario
(Weight = 35%)
Due date: Friday, 25 October 2024, 4:30 PM

Version: 2.0 – 21/08/2024

General Information and Submission

- This is an *individual assignment*.
- *Submission method*: Submission is online through Moodle.
- *Penalty for late submission*: 5% deduction for each day.
- *Assignment FAQ*: There is an Assignment 3 FAQ page set up on the EdStem forum.

Assignment Background

Data Analyst

Public Transport Victoria · Melbourne, AU

Part-time · Mid-Senior level

201-500 employees · Transportation/Trucking/Railroad

You have been hired as a data analyst at Public Transport Victoria (PTV), the Victoria Government authority responsible for public transport in the state. Some of your duties are data extraction, integration and analysis to provide good understanding regarding the public transportation condition in Victoria to the stakeholders.

After the COVID-19 restrictions were lifted, most companies are switching the workstyle from work-from-home to face-to-face. Therefore, transportation infrastructure and network is one of the most important aspects. While some people prefer to drive to work, some other people prefer to use the public transportation network as their main transportation mode. PTV as the sole provider for the public transportation network reduced their services during the lockdown period. Now, PTV has restored the services to cover as many areas as possible in the whole region. However, some questions remained mysteries. How good is the current PTV coverage? Are there any uncovered spots? Which area has the best public transportation options?

Therefore, as a data analyst, your task is to evaluate the data and provide the spatial data analysis to the stakeholders of PTV. The data should be presented in an area level, such as municipality, suburbs or postcode. For example, you may present "The number of bus services in Bundoora" or "The number of Trains or Trams network in Bundoora".

Data

There are two datasets that you have to obtain in this assignment, which are the PTV/GTFS dataset and Australian Boundary data.

The **General Transit Feed Specification (GTFS)** is a data specification that allows public transit agencies to publish their transit data in a format that can be consumed by a wide variety of software applications. Today, the GTFS data format is used by thousands of public transport providers.

GTFS is split into a schedule component that contains schedule, fare, and geographic transit information and a real-time component that contains arrival predictions, vehicle positions and service advisories. A GTFS feed is composed of a series of text files collected in a ZIP file. Each file models a particular aspect of transit information: stops, routes, trips, and other schedule data.
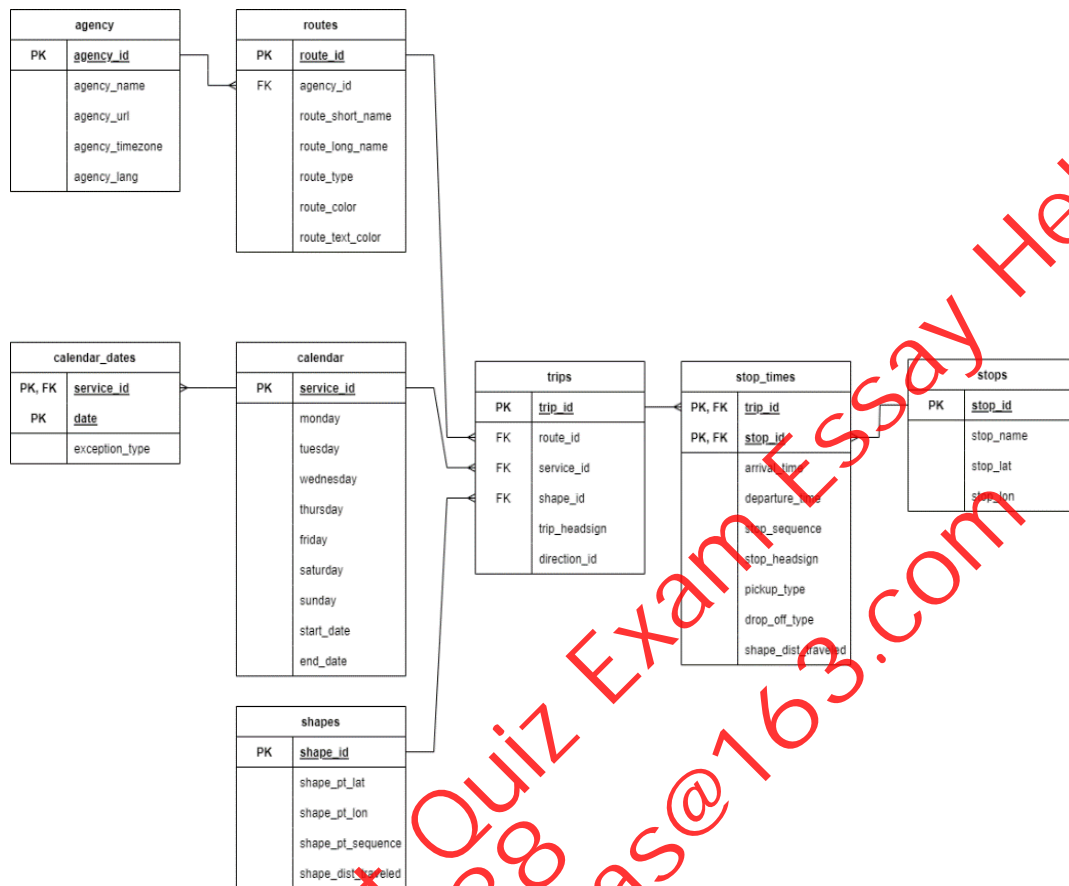
For more detailed information about GTFS, you can refer to the official documentation provided by Google at https://developers.google.com/transit/gtfs. Additionally, You can read further explanation about the PTV-GTFS data from https://transitfeeds.com/p/ptv/497. **For this assignment, we will be using the 17th March 2023 version of the dataset.**

PTV GTFS

| ID / Code / Name | Search Routes | Search Stops |

| Download Latest 262.3 MB | Routes 2,62 | Stops 27,77 |

Date	Size	Routes	Status
5 May 2023	168.3 MB	2,395	View Download
27 April 2023	177.5 MB	2,511	View Download
25 April 2023	147.0 MB	2,679	View Download
21 April 2023	165.6 MB	2,863	View Download
14 April 2023	183.2 MB	2,983	View Download
11 April 2023	240.0 MB	4,020	View Download
30 March 2023	256.4 MB	3,932	View Download
24 March 2023	237.2 MB	3,514	View Download
17 March 2023	223.0 MB	3,052	View Download
9 March 2023	226.1 MB	3,381	View Download

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | ... | 47 |

The GTFS data structure is shown below:

agency
PK	agency_id
	agency_name
	agency_url
	agency_timezone
	agency_lang

routes
PK	route_id
FK	agency_id
	route_short_name
	route_long_name
	route_type
	route_color
	route_text_color

calendar_dates
PK, FK	service_id
PK	date
	exception_type

calendar
PK	service_id
	monday
	tuesday
	wednesday
	thursday
	friday
	saturday
	sunday
	start_date
	end_date

trips
PK	trip_id
FK	route_id
FK	service_id
FK	shape_id
	trip_headsign
	direction_id

stop_times
PK, FK	trip_id
PK, FK	stop_id
	arrival_time
	departure_time
	stop_sequence
	stop_headsign
	pickup_type
	drop_off_type
	shape_dist_traveled

stops
PK	stop_id
	stop_name
	stop_lat
	stop_lon

shapes
PK	shape_id
	shape_pt_lat
	shape_pt_lon
	shape_pt_sequence
	shape_dist_traveled

The **Australian digital boundary** is defined by the Australian Bureau of Statistics using the **Australian Statistical Geography Standard (ASGS)**. The ASGS is a classification of Australia into a hierarchy of statistical areas. It is a social geography, developed to reflect the location of people and communities. It is used for the publication and analysis of official statistics and other data. The ASGS is updated every 5 years to account for growth and change in Australia's population, economy and infrastructure. For the 2021 release, the ASGS will be re-named to the Australian Statistical Geography Standard (ASGS) Edition 3.

The ASGS is split into two parts, the ABS and Non ABS Structures. The **ABS Structures** are geographies that the ABS designs specifically for the release and analysis of statistics. This means that the statistical areas are designed to meet the requirements of statistical collections as well as geographic concepts relevant to those statistics. This helps to ensure the confidentiality, accuracy and relevance of ABS data.The **Non ABS Structures** generally represent administrative regions which are not defined or maintained by the ABS, but for which the ABS is committed to directly providing a range of statistics.

The Main Structure is developed by the ABS and is used to release and analyse a broad range of social, demographic and economic statistics. It is a nested hierarchy of geographies, and each level directly aggregates to the next level. **Mesh Blocks (MBs) are the smallest geographic areas defined by the ABS and form the building blocks for the larger regions of the ASGS. Most Mesh Blocks contain 30 to 60 dwellings.**

Below is the simplified ABS and Non ABS Structure. You can read further explanation about the structure here

https://www.abs.gov.au/statistics/standards/australian-statistical-geography-standard-asgs-edition-3/jul2021-jun2026#overview
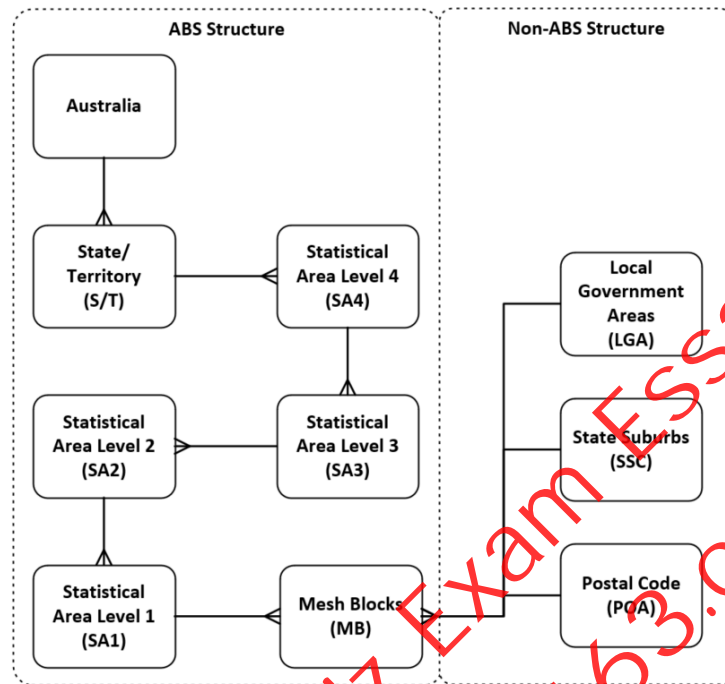
The Digital boundary files that you have to get is the Mesh Blocks dataset. The Mesh Blocks dataset is available as Shape file. You can read further explanation about the Mesh Blocks dataset here

https://www.abs.gov.au/statistics/standards/australian-statistical-geography-standard-asgs-edition-3/jul2021-jun2026/access-and-downloads/digital-boundary-files

Allocation files are non-spatial representations of how each geography is aggregated from their building block geography. You can also read further explanation about the Allocation files dataset here

https://www.abs.gov.au/statistics/standards/australian-statistical-geography-standard-asgs-edition-3/jul2021-jun2026/access-and-downloads/allocation-files
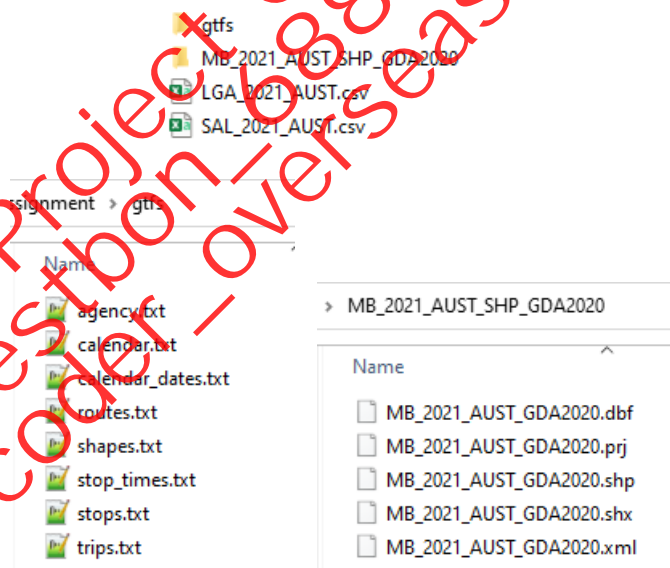
Assignment Task list

Your assignment consists of several parts. Always read the instruction one by one. Do not move to the step without completing the previous step:

- Task 1: Data Restoration - Restore the data to the database. Monitor the success indicator to ensure successful restoration of the data.
- Task 2: Data Preprocessing - Perform necessary structure maintenance and create result tables for further processing.
- Task 3: Data Analytics and Visualization - Develop SQL queries to analyze the data and evaluate performance & Create visualizations to present the results of the data analytics.
- **No data cleaning required for this assignment.**
- **For more information, see the FAQ for Assignment 3.**

For simplicity, all the data required for this assignment is readily available in the PostGIS Docker container. You can access these datasets within the container by navigating to the /data/adata folder. If you don't know how to do it, refer to the labs 10 activities.

```
root@db94d38b7162:/home/student# ls /data/adata
gtfs  LGA_2021_AUST.csv  MB_2021_AUST_SHP_GDA2020  SAL_2021_AUST.csv
```

Verify your data before the restoration process.

As a data analyst, it is your responsibility to understand and explore these publicly available data.

Assignment Task

Task 1: Data Restoration

Before you can start the data analytic processes, the first thing you have to do is to restore the external data to your database. Make sure you prepare a destination schema to restore your data. The destination schema for your assignment is "**ptv**".

Note:

- Before initiating the data restoration process, **it is essential to thoroughly explore the dataset**. This exploration involves identifying appropriate data types, determining field lengths, and making other relevant considerations that will inform the creation of the table structure.

- **Ensure that you restore the data into the PTV schema using regular (local) tables. Do not utilise foreign tables, as the data must be stored directly within the PostgreSQL database.**

- **Ensure that all tables are successfully restored, including 8 tables from GTFS and 3 tables for MB_2021, LGA_2021 and SAL_2021 respectively.**

The outputs of this task for **Report** are:

a) Attach a screenshot of the results to include all the tables you restored in Task 1, including the number of rows for each table you restored by using following code:

```
with tbl as
(select table_schema, TABLE_NAME
 from information_schema.tables
 where table_schema in ('ptv'))
select table_schema, TABLE_NAME,
(xpath('/row/c/text()', query_to_xml(format('select count(*) as c from %I.%I', table_schema,
TABLE_NAME), FALSE, TRUE, '')))[1]::text::int AS rows_n
from tbl
order by table_name;
```

Task 2: Data Preprocessing for Melbourne Metropolitan area

The purpose of this section is to manipulate the data into a suitable format for the following task analysis. This task has two parts: Mandatory requirement and Optional requirements.

Mandatory requirement
[You must meet the mandatory requirements described in this section.]

In this assignment, we aim to explore the **transportation accessibility [Topic of report]** of the **Melbourne Metropolitan area exclusively [Scope of report]**. The *mb_2021* table

contains mesh blocks for the entire country of Australia. To minimise query costs, ensure that you only use the mesh blocks within the Melbourne Metropolitan area for this assignment. The Melbourne Metropolitan's mesh blocks can be identified from the gcc_name21. If the column contains "Greater Melbourne", this mesh block is located in Melbourne Metropolitan. As a result, you need to create a table called "mb2021_mel" that contains ONLY the mesh blocks in Melbourne Metropolitan.

	sa1_code21	sa2_code21	sa2_name21	sa3_code21	sa3_name21	sa4_code21	sa4_name21	gcc_code21	gcc_name21	ste_code21	ste_name21	ste
22	10901117322	109011173	Albury - North	10901	Albury	109	Murray	1RNSW	Rest of NSW	1	New South Wales	AU
23	10901117322	109011173	Albury - North	10901	Albury	109	Murray	1RNSW	Rest of NSW	1	New South Wales	AU
24	21401137143	214011371	Frankston	21401	Frankston	214	Mornington Peninsula	2GMEL	Greater Melbourne	2	Victoria	AU
25	10901117325	109011173	Albury - North	10901	Albury	109	Murray	1RNSW	Rest of NSW	1	New South Wales	AU
26	10901117301	109011173	Albury - North	10901	Albury	109	Murray	1RNSW	Rest of NSW	1	New South Wales	AU
27	10901117323	109011173	Albury - North	10901	Albury	109	Murray	1RNSW	Rest of NSW	1	New South Wales	AU

Optional requirements:

[You are free to explore and manipulate the data creatively within the mandatory requirements, which are limited to Melbourne Metropolitan for the topic of transport accessibility.]

For *optional requirements* can be selected based on your specific data analysis needs. Make sure to include a detailed explanation of your rationale in the report for optional requirements you choose.

Question:
Do I have to answer at least one of the optional requirements?

Answer:
No, you are free to explore and manipulate the data creatively as long as the data is analysed in Melbourne Metropolitan for the topic of transport accessibility.

The following suggestion may useful for data exploration and analysis transportation accessibility of Melbourne Metropolitan area:

1.Since the working area will be Melbourne Metropolitan, it is important to have a polygon for the boundary of our working area. Hint: aggregate all mesh blocks polygon to create one large polygon for Melbourne Metropolitan boundary.

2.*Stops* table does not have any geometry column. It might be useful to add a geometry column, using the latitude and longitude values available in the table. Make sure you use GDA2020 (SRID:7844) for this column.

3.The *Stops* table does not show direct information regarding the vehicle types, routes_short_name and routes_long_name. These informations are stored in the *routes* table.

4.If you want to explore the transportation situation for different vehicle types, such as tram, train, or bus, the vehicle type is determined by the corresponding route type in the routes table, where:

- 0 corresponds to tram
- 2 corresponds to train
- 3 corresponds to bus

● Any other route type is labelled as 'Unknown'.

The outputs of this task for **Report** are:

b) Attach a screenshot of SQL script for creating a table named "**mb2021_mel**" that contains ONLY the mesh blocks in Melbourne Metropolitan.

c) Provide a detailed explanation of the **remaining** data processing steps you have conducted, including screenshots of the SQL scripts and the rationale behind your choices in the report.

Task 3: Data Analytics and Visualisation

In this section you will need to perform data analysis on the tables you have restored, focusing on transport accessibility in metropolitan Melbourne. Use the techniques you have learned in the spatial database part to carry out your analysis. You are free to choose any specific perspectives or aspects of data analysis relevant to your dataset, but ensure that your analysis relates to the **main topic: transport accessibility in metropolitan Melbourne.**

This could include exploring different statistical measures or carrying out other relevant analyses. Present your findings clearly and concisely, demonstrating your understanding of the dataset and highlighting any notable observations or patterns.

As part of this data visualisation, you will also need to create at least one map-based headmap using QGIS to present your findings related to the main topics. These visualisations will be used in the next section of the assignment, the summary report. To support your analysis, you can include screenshots of the visualisations directly in the report.

Be sure to include the script or code used for data analysis and data visualisation in the appendix of your report. The script should provide clear instructions on how the analysis was performed and any necessary calculations or transformations applied to the data. This will ensure that your analysis can be reproduced and verified. Remember to include appropriate labels, titles, and legends in your visualisations to make them easy to understand. The visualisations should be of sufficient quality and clarity to effectively convey your analysis findings.
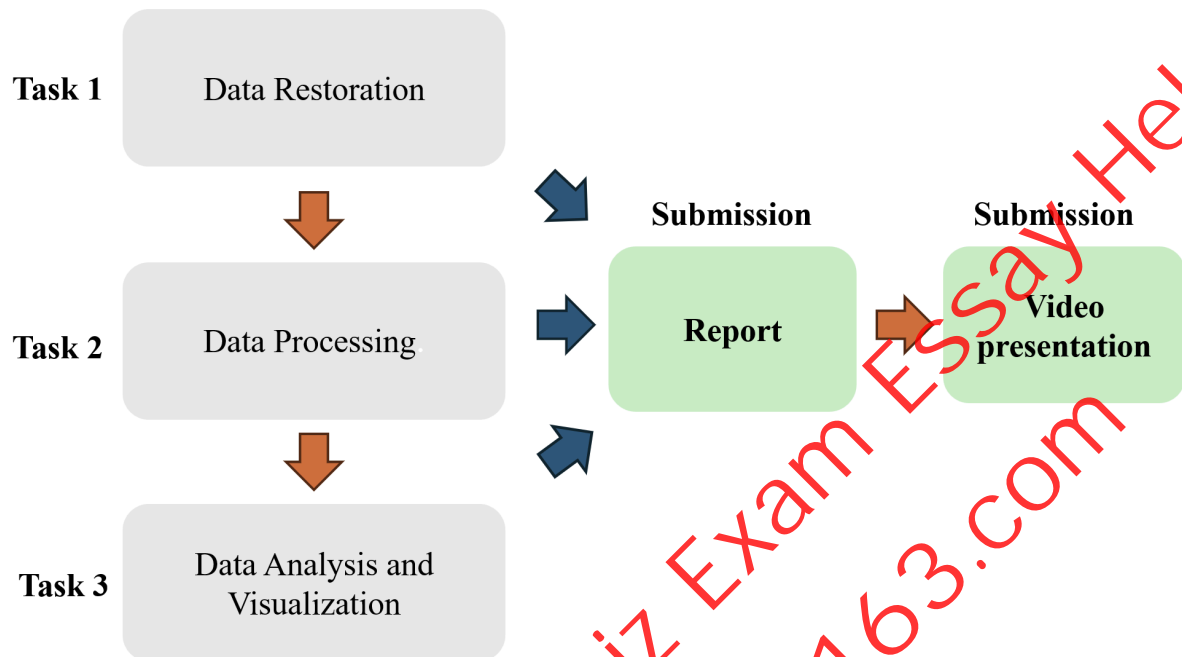
Note:

● Use SQL queries to investigate the restored tables.
● Conduct a thorough descriptive analysis to uncover insights within the data.
● Summarise and Visualise your findings clearly and concisely.
● Highlight key observations and patterns discovered during the analysis.
● Ensure your findings reflect a deep understanding of the data.

The outputs of this task for **Report** are:

d) Data analysis and visualisation, including the screenshot of SQL script and visualisation. For the visualisation, it must contain at least one map base figure.

Submission Checklist

Assignment Flowchart

Task 1 | Data Restoration

Task 2 | Data Processing

Task 3 | Data Analysis and Visualization

Submission

Report

Submission

Video presentation

Summary Report for Task 1 to 3

As a professional data analyst, your task is to consolidate all the previous tasks, including data restoration, processing, analysis, and visualisations, into a comprehensive written report. The report should adhere to a word limit of 2000 words and follow a structured format, consisting of an introduction, methodology, results, conclusion, and appendix. **Please note that a question-and-answer format is not acceptable for this assignment, and marks will be deducted for using such a format.**

Please ensure that the report adheres to the given word limit and is well-organised, concise, and coherent. The sample report should be formatted as follows:

Title: Write your title here on a separate page, (Note: Abstract is not required)

1. Introduction, such as

- Briefly explain the purpose of the report and what you aim to achieve with your analysis.
- Highlight the key questions you want to investigate through your analysis.

2. Methodology

This section should provide a clear explanation of the different stages of your work.

- **Dataset Overview, such as**

Provide an overview of the data and its source.

- **Data Restoration and Preprocessing, such as**

  Explain how you imported and initially explored the data. Include the software and libraries used.

  Provide a detailed explanation of the **remaining** data processing steps you have conducted, and the rationale behind your choices in the report.

- **Data Analysis and Visualization**

  Describe which area of transport accessibility in metropolitan Melbourne you are primarily exploring.

  Describe the analysis you conducted and the types of visualisations you chose to use, and why you felt they would effectively represent your data and findings. What software or libraries were used to create these visualisations?

3. Results:

Present the results of your in-depth investigations. Explain what these results mean and how they answer your initial questions.

For the visualisation, it must contain at least one map base figure.

4. Discussion

Discuss your findings and their implications.

- Restate the main findings of your descriptive and advanced analyses.
- Discuss how these findings answer your initial questions or hypotheses.
- Reflect on the process and any limitations or challenges you faced during your analysis.

5. References [Excluded from the 2000-word limit]

If you have used external resources, don't forget to cite them properly according to the chosen style guide (APA 7th edition).

6. Appendix [Excluded from the 2000-word limit]

The screenshots of the following tasks:
- Attach a screenshot of the results to include all the tables you restored in Task 1, including the number of rows for each table you restored.

- Attach screenshots of the SQL scripts used in Task 2, including the SQL scripts for creating a table called "mb2021_mel" and screenshots of the SQL scripts you used for the remaining data processing steps.
- The SQL script for Data analysis and visualisation

Video presentation

A five minute video presentation in mp4 format save as: YourstudentID_A3_video.mp4

Based on the report you have created, present your design and findings in a **five-minute video presentation**. Ensure you thoroughly understand both the dataset and the report to effectively extract and communicate the key points.

Assignment Submission

1. A combined pdf file save as: YourstudentID_A3_report.pdf, containing all of the above tasks 1 to 3.
2. A five minute video presentation in mp4 format save as: YourstudentID_A3_video.mp4

Zip all above files from step 1 to 3, and name the ZIP folder as **A3_YourstudentID.zip**.

- The submission of this assignment must be in the form of **a single ZIP file**. Only PDF and .mp4 files will be accepted within the zip file. No other formats will be accepted.
- You must ensure that you have all the files listed in this checklist before submitting your assignment to Moodle. Failure to submit a complete list of files will lead to mark penalties.
- It's important to note that our support hours are limited, and we don't have the capacity to address submission issues outside of working hours. You must ensure that you have all the files listed in this checklist before submitting your assignment to Moodle. Failure to submit a complete list of files will result in a mark penalty.
- Penalty for late submission: 5% deduction for each day, including weekends
- Submission cut-off time: Friday, 1 November 2024, 4:30PM. Submissions will not be accepted after this time unless there are special considerations.

Authorship

This assignment is an **individual assignment** and the final submission must be identifiably your own work. Breaches of this requirement will result in an assignment not being accepted for assessment and may result in disciplinary action.

Late Penalty

Late assignments submitted without an approved extension may be accepted up to a maximum of **seven days** with the approval of the Chief Examiner and/or Lecturer but will be **penalised at the rate of 5% per day (including weekends and public holidays)**. Assignments submitted more than seven days after the due date will receive a zero mark for that assignment and may **not receive any feedback**.

Please note(late penalty and extension）:

1. An inability to manage your time or computing resources will not be accepted as a valid excuse. (Several assignments being due at the same time are a fact of university life.)

2. Hardware failures, whether of personal or university equipment, are not normally recognised as valid excuses. Failure to back up assignment files is also not recognised as a valid excuse.

Special Consideration

Students no longer seek extensions from chief examiner/teaching team. All extensions / special considerations will now be handled by the central Spec Con team. **Please do not email teaching staff to request an extension or special consideration.**

Extensions and other individual alterations to the assessment regime will only be considered using the University Special Consideration Policy. Students should carefully read the Special Consideration website, especially the details about what formal documentation is required.

All special consideration requests should be made using the Special Consideration Application.

Please do not assume that submission of a Special Consideration application guarantees that it will be granted – you must receive an official confirmation that it has been granted.

Getting help and support

What can you get help for?

- *Consultations with the Teaching Team*
Talk to the Teaching Team:
https://learning.monash.edu/course/view.php?id=19675&section=5
- *English language skills*
Talk to English Connect: https://www.monash.edu/english-connect
- *Study skills*
Talk to a learning skills advisor: https://www.monash.edu/library/skills/contacts

- ***Counselling***

Talk to a counsellor: https://www.monash.edu/health/counselling/appointments

Plagiarism and Collusion:

Monash University is committed to upholding standards and academic integrity and honesty. Please take the time to view these links.

Academic Integrity Module

Student Academic Integrity Policy

Test your knowledge, collusion (FIT No Collusion Module)

All the best for your Assignment!