**FIT5137 S2 2023 Assignment 4: PTV Answer Sheet (Weight = 30%)**

**PLEASE SUBMIT ANSWER SHEET IN PDF FORMAT**

**Due date: Wednesday, 25 October 2023, 11:55pm**

Version: 2.0 – 25/09/2023

**Assignment Task list**

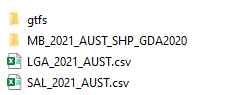
Your assignment consists of several parts. Always read the instruction one by one. Do not move to the step without completing the previous step:

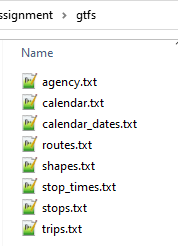
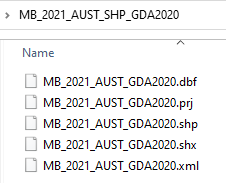
* Task 1: Data Restoration - Restore the data to the database. Monitor the success indicator to ensure successful restoration of the data.
* Task 2: Data Preprocessing - Perform necessary structure maintenance and create result tables for further processing.
* Task 3: Data Analytics - Develop SQL queries to analyze the data and evaluate performance.
* Task 4: Data Visualization - Create visualizations to present the results of the data analytics.

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



Verify your data before the restoration process.



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

**Do not edit or remove any content on the answer sheet, including the questions**. Please write your answers in the answer boxes provided. If necessary, you can adjust the size of the answer box to fit your answer.

## **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 utilize foreign tables, as the data must be stored directly within the PostgreSQL database – updated 27/09/2023
* Make sure all 8 GTFS tables are restored successfully
* Index or constraints can be added to the table after the data has been restored completely

(Note: This index or constraint requirement is **NOT** mandatory in this Task 1 – updated 25/09/2023)

* **No data cleaning required for this assignment**
* For more information, see the FAQ for Assignment 4.

### 1.1 PTV schema

Write the SQL script to create the destination schema named “**ptv**”.

|  |
| --- |
|  |

### 1.2 GTFS

Write the SQL script to restore **ALL** tables in GTFS files.

|  |
| --- |
|  |

### 1.3 ABS Mesh Blocks

Scripts to restore the Mesh Blocks files by using correct dataset file. Restore the file using ogr2ogr into table “**mb2021**”

|  |
| --- |
|  |

### 1.4 ABS Allocation Files

Write the SQL script to restore the LGA2021 Allocation file.

|  |
| --- |
|  |

Write the SQL script to restore the SAL 2021 Allocation file for suburb2021.

|  |
| --- |
|  |

The allocation tables have 1-N relationship with the mb2021 in mb\_code21 – mb\_code\_2021. Although there are no PK – FK defined in the table, the relationship rule still apply.



### 1.5 Data Verification

Verify your restoration by running this script. 1.Do not modify the verification script. 2.Make sure that the table name is consistent with the table we provided.

Output: Attach a screenshot of the results to include all tables you have restored in Task 1.

|  |
| --- |
| **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; |

Screenshot:

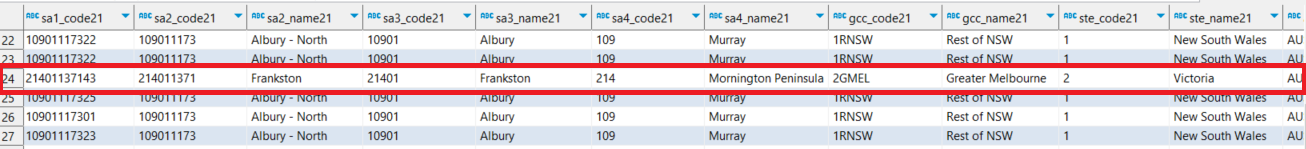
|  |
| --- |
|  |

## **Task 2: Data Preprocessing**

### 2.1 Filter Melbourne Metropolitan area

The mb2021 table contains whole mesh blocks in Australia. To minimise the query cost, we want to ensure that you only use the mesh blocks in Melbourne Metropolitan. 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.

Create a table named “**mb2021\_mel**” that contains ONLY the mesh blocks in Melbourne Metropolitan.



Write the SQL script to do this.

|  |
| --- |
|  |

Attach a screenshot of the Spatial Map results.

Screenshot:

|  |
| --- |
|  |

### 2.2 Melbourne Metropolitan Boundary

Since the working area will be Melbourne Metropolitan, it is important to have a polygon for the boundary of our working area. Create a table, named “**melbourne**” for Melbourne Metropolitan boundary.   
Hint: aggregate all mesh blocks polygon to create one large polygon for Melbourne Metropolitan boundary.

Write the SQL script to do this.

|  |
| --- |
|  |

Attach a screenshot of the Spatial Map results.

|  |
| --- |
|  |

### 2.3 Add Geometry column to Stops table

Stops table does not have any geometry column. Add a geometry column by using the latitude and longitude value that are available in the table. Make sure you use GDA2020 (SRID:7844) for this column.

Write the SQL script to do this.

|  |
| --- |
|  |

Attach a screenshot of the Spatial Map results.

|  |
| --- |
|  |

### 2.4 Denormalise GTFS structure

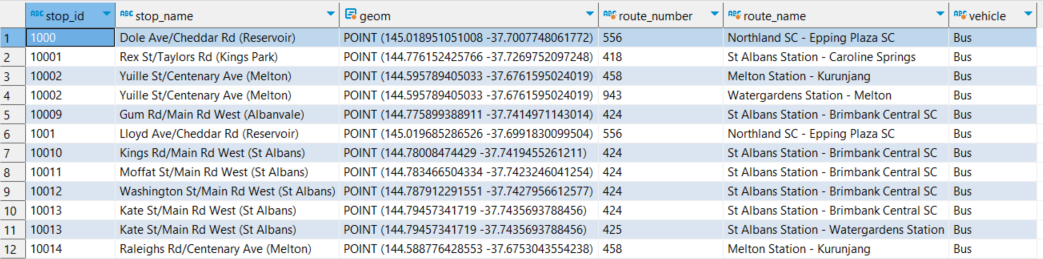
The ptv.stops table does not show direct information regarding the vehicle types, routes\_short\_name and routes\_long\_name. These information are stored in the routes table.

Create a table called "stops\_routes\_mel" to encompass the following attributes: stop\_id, stop\_name, coordinates, route number (derived from routes\_short\_name), route name (derived from routes\_long\_name), and vehicle type. This data set should encompass all stops within the Melbourne Metropolitan area.

The vehicle type is determined by the corresponding route type, where:

* 0 corresponds to tram
* 2 corresponds to train
* 3 corresponds to bus
* Any other route type is labeled as 'Unknown'.

Use this figure as an example of expected result. (Note:Data value is for demonstration purposes only.)



Make sure you remove any duplications in your result.

Write your SQL query here

|  |
| --- |
|  |

Please complete the following statistics from **stops\_routes\_mel** table:

**Question 2.4.1**:How many rows do you have in the stops\_routes\_mel table?

Write the SQL script to do this and attach a screenshot of the query

|  |
| --- |
|  |

Screenshot

|  |
| --- |
|  |

**Question 2.4.2**:How many unique stop\_ids do you have in the stops\_routes\_mel table?

Write the SQL script to do this and attach a screenshot of the query

|  |
| --- |
|  |

Screenshot

|  |
| --- |
|  |

## **Task 3: Data Analytics**

### 3.1 Suburbs Accessibility

Create an SQL query to identify the **number of bus stops** in each Suburb. Your result should have the suburb name and the total bus stops in it.

Hint :

* identify the mesh block location of a bus stop. Then, aggregate the number in Suburb level.
* One suburb consists of multiple mesh blocks

Write your SQL query here

|  |
| --- |
|  |

Provide the screenshot of your **execution plan** and the real **execution time** for this query. You can get the real execution time under your result set.

Note:

* Execution plan can be found in SQL Editor -> Explain Execution Plan
* The execution time is shown in the screenshot below.(Note: the screenshot is for demonstration purposes only)  
  

|  |
| --- |
| Execution plan : |

|  |
| --- |
| Execution time : |

You are now tasked with devising an approach to enhance your query execution and minimize execution time. Provide a comprehensive explanation of your strategy, accompanied by the SQL script outlining the measures you've taken to optimize query performance. Additionally, include a screenshot showcasing the execution plan and execution time, effectively visualizing the enhancements achieved following the optimizatio.

|  |
| --- |
| Strategy and SQL script: |

|  |
| --- |
| Improved Execution plan : |

|  |
| --- |
| Improved Execution Time : |

**Question 3.1.1**: Provide a list of the five suburbs with the lowest count of stops. In case multiple suburbs share the same minimum number of stops in your findings, arrange them in ascending order based on their suburb names.

Write the SQL script to do this and attach a screenshot of the query

|  |
| --- |
|  |

Screenshot

|  |
| --- |
|  |

**Question 3.1.2**: Average number of distinct stops in suburb

Write the SQL script to do this and attach a screenshot of the query

|  |
| --- |
|  |

Screenshot

|  |
| --- |
|  |

### 3.2 LGA Blankspot

The next step is to evaluate the **residential area** **without Bus Stops**. The residential area without any Bus Stops in it is considered as the **Blankspot**. Each mesh block has a distinct category. The category is defined in “**mb\_cat21**” column, mb2021\_mel table. In this task, your duty is to identified the percentage of blankspot in every city council or LGA. Below is the blankspot example in Kingsbury.



Let B as the number of residential blankspot, R as the total number of Residential Mesh Blocks, where B ⊆ R. The percentage of blankspot X in every LGA can be calculated using the following formula

Display the LGA name, total number of Residential Mesh Blocks, total number of residential blankspot, percentage of blankspot in Melbourne Metropolitan. Sort results in ascending order by total number of Residential Mesh Blocks.

Write the SQL script to do this and attach a screenshot of the query

|  |
| --- |
|  |

Screenshot

|  |
| --- |
|  |

**Question 3.2.1**: Complete the following statistical data based on the result.

Note:

* The query and screenshot are not required for this section. You can write down your results directly
* If more than one LGA has the same percentage of blankspots, sort them by LGA name in ascending order.(Updated 22/10/23)

|  |  |
| --- | --- |
| Criteria | Answer |
| Top 5 LGAs with the highest % of blankspots |  |
| Top 5 LGAs with the lowest % of blankspots |  |
| Average % of blankspots |  |

## **Task 4: Data Visualisation**

In this task, you are required to incorporate a heatmap visualization.

### 4.1 LGA Blankspot Analysis

In this task, you will put the blankspot percentage in the heatmap. Provide the segmentation as follow:

* + X ≤ 20 %
  + 20% < X ≤ 40%
  + 40% < X ≤ 60%
  + 60% < X ≤ 80%
  + X > 80%

Write an SQL query to create a table named 'lga\_blankspot' containing the blankspot percentages categorised by the LGA from the previous task 3.2 and sufficient spatial data. And attach a screenshot of the table contents.

Note: Ensure that this table is structured to facilitate the creation of a visual heat map specifically for the Melbourne region.

|  |
| --- |
|  |

Screenshot

|  |
| --- |
|  |

Provide the screenshot for the heatmap by using QGIS. Please note that the heatmap is map-based and visually represents the distribution of blankspot percentages across different LGAs in the Melbourne region.

Remember to include appropriate labels, titles, and legends in your visualizations to make them easy to understand (Updated 05/10/2023)

|  |
| --- |
|  |

\*\*\*End\*\*\*