**Pedestrian data analysis task process and overall strategy**

**Part 1:**

1. Reading documentation about provided API links which contain pedestrian and sensor data to get overview about what is the context before dive into analysis
2. Extract data via APIs and convert into dataframe with python
3. Run data validation:

* Checked no null values for pedestrian data
* Checked both pedestrian and sensor data, no duplicate rows found
* Few columns are empty for sensor tables. Some sensors’ direction data is missing, but not impacting the required analysis. ‘note’ column for few removed sensors are empty which lead to no remove date info can be extracted from sensor table alone.

1. Process and run analysis to get required output. Details and comments are captured in Jupyter notebook file on github.

(Alternative solution in a full production environment. I would prefer to build incremental load ELT pipeline with Dataflow in GCP and load data into BigQuery. Since BigQuery has auto scale features which will be suitable for this type of analysis with SQL. Also, BigQuery ML could help with building some predictive analysis models for further analysis)

Q1,Q2：

Please find attached csv files

Q3:

Lincoln-Swanston (West) has shown the most decline from 2019 to 2020 with 77.45% decrease

Southern Cross Station has shown the most decline from 2019 to 2021 with 83.12% decrease

Text

Description automatically generated with medium confidence

Q4:

Building 80 RMIT has shown the most growth last year (compared with 2020) with 199.40% increase

Text

Description automatically generated with medium confidence

**Part 2:**

1. Please find excel file in attachment
2. Other metrics:

|  |  |
| --- | --- |
| Metrics | Description |
| Peak hours | Period from day to day with pedestrian volume beyond certain threshold |
| Pedestrian density | To tell decision makers about which area has high volume than others |
| Standard deviation | Calculate standard deviation of sensor locations’ volume, it would help tell whether locations are consistently busy |

Diagram

Description automatically generated

* columns “sensor\_description” and “sensor\_name” are removed from “sensor” table as that information is also stored in “pedestrian” table
* Add another table called “incident” even it is not given by the task. But it could be helpful to analyse correlation between pedestrian volumes and traffic incidents which is a interesting topic to study. And the idea is coming from the geo data provided by “sensor table.”
* Relationship between entities are shown as above ER diagram.

**Challenges:**

1. For the first two questions, the scope is not defined, so I checked the data, and thought about to split the output into different year or maybe create a dashboard with filter on year since many sensors are installed in different years. However, the total volumes despite the year of sensor installation still reflect how busy it is for specific locations. From government’s perspective, without further filtering, it can still derive insights and help make decisions on traffic control based on the volumes of pedestrian and different days. So, I decided to just give top 10 based on all sensors. Similar rules applied to Q2.
2. Another challenge is that sensors removed date is not provided and it might have impact for analysis on Q3 and Q4. So, I applied logic to find out sensors that are removed status and last pedestrian read is before‘2021-12-31’. So those sensors will be removed from the list to query the answer for Q3,Q4 since they might get removed half way through 2021. Even in the end it turns out those removed sensors do not impact the result, but it is always good practice to consider about this kind of factor before analysing data.