SMART CITIES RESEARCH REPORT

Deliverables

Present your deliverables in a report format. Please see below the report structure

1. Abstract (suggest 150 words)
2. Main text of the report including:
    - Introduction - background to research, overview of topic, aims and objectives
    - Results Discussion with supported arguments
    - Conclusions and recommendations
3. References
4. Appendix or appendices (if applicable)

Note: Your report should be 5,000 words (with 10% discretionary allowance). The word limit applies to the main text area only, as this is where your report is introduced and then carried forward. Your final word count should not include the abstract, references, bibliography or appendices. In-text references are included in the word count. If the report is to contain a considerable number of tables and figures, it may be best to place them in an appendix and use summary tables or charts in the main text only which will assist the reader in following your arguments without necessarily having to go into great detail.

Research Question 1

a) In recent years, the population and visitors of Cyberjaya has grown tremendously. Hence, Cyberview is embarking on a data-driven approach to anticipate and plan for future infrastructure and services. To assist in this endeavour, the town planners would like to get a rough estimate of the current number of visitors. Calculate **average daily subscribers rounded to 2 decimal places** for grid IDs in Cyberjaya listed below for the time period between **2019-07-19 to 2019-08-18** during **0700 to 1500**

   Grid_id: [172456.0, 172457.0, 171754.0, 171755.0, 171756.0, 171052.0, 171053.0, 171054.0, 171058.0, 170356.0]. Kindly also use the shape file provided

b) Part of the growth has been attributed to proliferation of the F&B industry. **Identify all the Grid ID's with F&B POIs** using the datasets provided

c) Boost is also interested in expanding its footprint in Cyberjaya's F&B industry as it has a significant number of userbase in Cyberjaya. The ability to identify the most popular F&B locations during lunch hour (1200 to 1400) allows for Boost to refine its F&B merchant acquisition strategy and run better campaigns. **From the Grid ID's identified in Question 1b), state the top 10 Grid ID's with the highest number of Boost users from 1200 to 1400 and provide the count.**

Research Question 2

a) Currently traffic count is monitored by intelligent cameras at certain main roads in Cyberjaya. Cyberview intends to augment traffic tracking on roads which is currently not served by these intelligent cameras. **Using telco data, develop a method to track road usage / traffic density from 1200 to 1400 on the 6 main routes listed below. Creatively visualise your output.**

N.B. We are looking specifically for mobility / movements. Do account for static subscribers

Route1 grid_id = [173860.0,173861.0,173159.0,173160.0,173161.0,172459.0,172460.0,172461.0,172462.0,172463.0,172464.0]
Route2 grid_id = [173860.0, 173859.0,173157.0,172455.0,171753.0,171051.0,170349.0,169647.0,168945.0,168243.0,167541.0, 166839.0,166137.0,165435.0]
Route3 grid_id = [165435.0,165436.0,165437.0,165438.0,166140.0,166141.0,166843.0,166844.0,166845.0,166846.0,166847.0 ,166848.0,166849.0]
Route4 grid_id = [166847.0,167549.0,167550.0,168252.0,168954.0,169656.0,170358.0,171060.0,171762.0,172464.0]
Route5 grid_id = [170361.0,170360.0,171061.0,170359.0,170358.0,170357.0,170356.0,170355.0,169653.0,169652.0,169651.0, 170353.0,170352.0,170351.0,170350.0,170349.0,170348.0,169645.0,169644.0]
Route6 grid_id = [166844.0,167546.0,168248.0,168247.0,168949.0,169651.0,170353.0,171055.0,171757.0,171756.0,172458.0,173160]

b) Air quality is one of the major concerns among Malaysians and air sensors are deployed across Cyberjaya to measure air quality. Using data from one of the air sensors stored in “air_sensor.json”, **build a model to forecast air quality (pm25) from 2019-07-01 to 2019-07-07 on an hourly interval from 0000 to 2300 for each day.**

Research Question 3

a) In-line with Cyberview's green aspirations, Boost will be launching a Low Carbon Mobility cross-sell campaign to promote the usage of Neuron Electric Scooter and GoCar's fleet of Energy Efficient Vehicles (EEVs).

You are required to build a model to identify customers who are likely to be interested in the campaign as well as their preferred mode of transportation. Based on your model's outcome, the quantity of resources e.g. cars, scooter, manpower, etc. will be allocated. The objective of this campaign is profit maximization rather than just the accuracy of the model as there is a revenue and cost impact for wrong predictions. For example, incorrectly allocating a GoCar results in losses such as cost of financing the car, parking fees, fuel consumption, etc. On the other hand, correctly predicting customer's demand for the right vehicle generates revenue. The cost and revenue of each outcome are summarized below:

		Actual Value		
Predicted Value		Neuron scooter	GoCar	Not Interested
	Neuron scooter	+ RM 70	- RM 40	- RM 10
	GoCar	- RM 110	+ RM 330	- RM 120
	Not Interested	- RM 10	- RM 30	+ RM 20

Build a model that returns the highest profit, so that running this campaign meets this objective. Use the historical result stored in "past_result.csv" to train your prediction model. You are required to submit your model's prediction result for each Customer ID in the holdout sample given in "potential_users.csv".

Research Question 4

b) **Propose a smart city/green solution for the future planning of Cyberjaya which must include a low carbon traffic solution**. For this question, define your role – either as a town planner, consultant, traffic planner, entrepreneur, etc. and present the solution from this perspective.

Points will be given for the following criteria;

- Big picture thinking & how well you have incorporated available solutions and datasets
- Solution technicality; solidly applied data engineering/science methodology and results, creative visualization
- Business model & Go-to-market Strategy
- Other considerations (e.g. Data governance and ethics)

Success Criteria

#	Questions	Success criteria
1a	Calculate **average daily subscribers** for given grid IDs	**Accuracy**
1b	Identify all grid IDs with **F&B POIs**	**Accuracy**
1c	Based on Q1(b), identify top 10 grid IDs with the highest number of Boost users	**Accuracy**
2a	Track subscriber movement on identified roads in Cyberjaya. Provide visualization	**Open ended**
2b	Build a model to forecast air quality	**RMSE**
3	Build profit maximization model for Boost's low carbon mobility campaign	**Profit**
4	Propose a smart city/green solution of the future for Cyberjaya which includes a low carbon traffic solution.	**Open ended**

Files provided by data source type

Source	Type	File Name
A	Telco: Customers' Profile	crm.csv
B	Telco: Customer Location Updates	event_log.csv
C	Boost: Merchant List	Merchant_list.csv
D	Boost: Transaction List	transaction.csv
E	Point of Interest (POI)	poi.csv
F	Partner / Open Source API: GoCar & Neuron Electric Scooter	past_result.csv, potential_users.csv
G	Partner / Open Source API: Transport Location	Transport_location.csv
H	Partner / Open Source API: Air Quality Data	air_sensor.json
I	Shape	Files in folder 'shape'

A. Telco: Customers' Profile

Description

- This datasets includes the customer information such as gender, age, device model and ARPU.

Typical data set for illustration purposes

customer_id	customer_age	arpu	model	gender
135733420730428000	25	86.73	iPhone 6 (A1586)	female
125554154025311000	28	134.04	Sony Xperia 1	male
126170404671792000	19	56.59	SM-G532G DS	female
126283295550923000	34	285.66	nova 4e	male

B. Telco: Customer Location Updates

Description
• Telco operators pick up on subscribers; location and mobility when there is any voice or data usage activity and the subscribers' are in within the towers' serving range. • The table below contains the data of subscribers along with their movements captured by the tower cells over a period of 41 days.

Typical data set for illustration purposes

customer_id	grid_id	date_time
q04645676476400588	171750.0	1563522900.0
d188524437162006696	159123.0	1563522960.0
2585249050450420000	160529.0	1563522960.0

C. Boost: Merchant List

Description
• This table provides a list of Boost Merchants as well as information such as Merchant's Name, Merchant Category and Merchant grid id which provides the grid location of the Boost Merchant.

Typical data set for illustration purposes

MerchantCode	MerchantName	MerchantCategory	grid_id
025cpo2su800	F&B Chain 10	F&B	164735.0
025ct01350	Retail Chain 38	Retail	169654.0
0271cag00ns8	Household Goods and Groceries Chain 6	Household Goods and Groceries	162634.0
0280h2b8id3	Services Individual 54	Services	175968.0

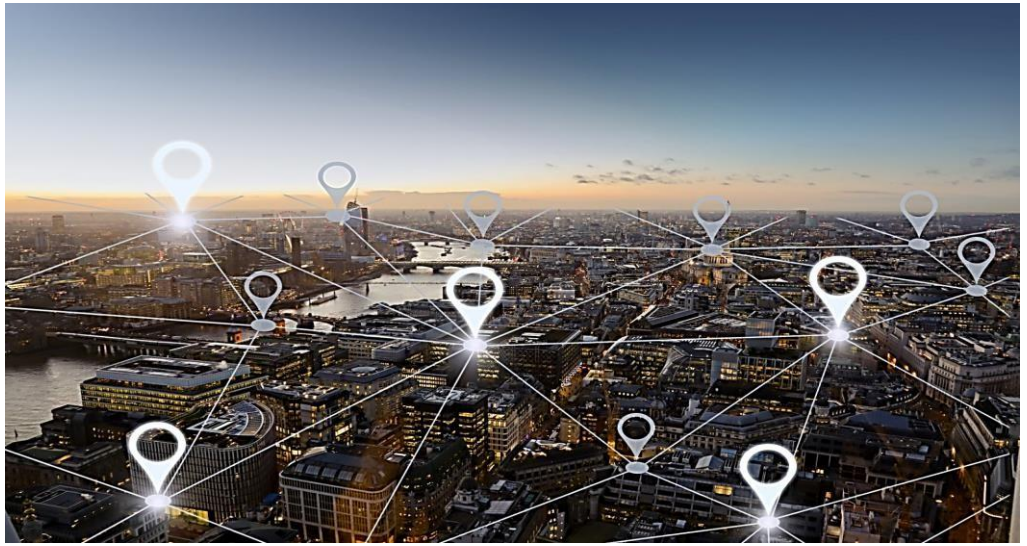
D. Boost: Transaction List

Description
• This table records information on each transaction, which hold customer id, merchant code, datetime and gross transactional value.

Typical data set for illustration purposes
(detailed Datathon data set walkthrough to follow)

customer_id	MerchantCode	MerchantCategory	date_time	GTV
r768997424647556924	0p0035r51kc	Household Goods and Groceries	1563536880.0	13.76
1218092905714920000	c4202cier008	F&B	1563536940.0	73.7
43257509490323300000	ni14000ehi37	Services	1563537000.0	123.56

E. Point of Interest(POI)



Description
• POI is a term where it is most often used on a map or in a guidebook, to indicate an attraction that might be of interest to visitors.

Typical data set for illustration purposes

Category	SubCategory	Latitude	Longitude	POIName
amenity	cafe	2.924346	101.639315	Poolside Cafe
office	company	2.921199	101.657462	Dell Asia Pacific Sdn. Bhd.
shop	supermarket	2.924815	101.6362535	A&C Grocery

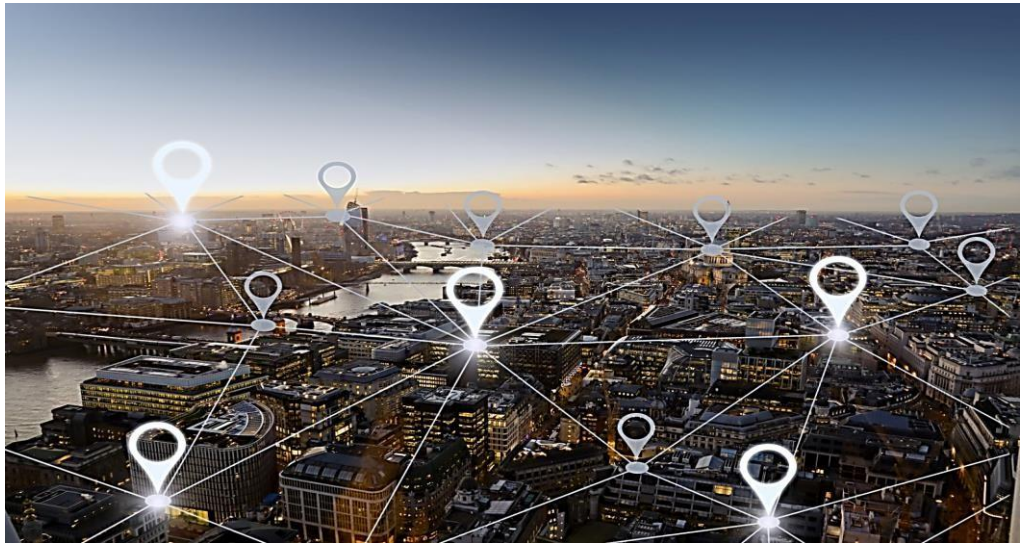
F. Partner / Open Source API: GoCar & Neuron Electric Scooter

Description
• past_result.csv - This table provides the campaign outcome (successful / not interested) when campaigns are targeted at the specific list of users. • potential_users.csv – Table contain the list of customer for targeted marketing campaign recommendation.

Typical data set for illustration purposes

customer_id	campaign_outcome
k482906940645133770	Not Interested
1592070382506620000	Neuron electric scooter
8013861800980100000	GoCar
3062964217448050000	Neuron electric scooter

G. Partner / Open Source API: Transport Location



Description
• This table provides a list of transport stations (e.g. Gocar Station, Neuron electric scooter, bus stop) and their latitude and longitude.

Typical data set for illustration purposes

POI Name	Latitude	Longitude
Gocar Station 1	2.99384	101.23452
Gocar Station 2	2.99803	101.35253
Gocar Station 3	2.99821	101.324324

H. Partner / Open Source API: Air Quality Data

Description

- This datasets includes measurements and records from sensors in Cyberjaya
- Air sensor information is important in predicting the air quality at a specific time period.
- To ensure accurate air quality prediction, the dataset provides multiple perspectives to the current air quality.

Typical data set for illustration purposes

Data

{ "CreatedDateTime": 1558483742, "DeviceDataTypeId": 5, "DeviceId": 96, "DisplayName": "pm25", "FabrickName": "pm25", "Name": "pm25", "Status": "Normal", "Unit": "ug/m3", "Value": 35.0, "_id": { "$oid": "5ce4935d31b728505c55b058" } }

I. Shape

Description

- This shape folder contains files of grid id with the corresponding Polygon based on latitude and longitude.

Typical data set for illustration purposes

Grid_id	geometry
159117	POLYGON ((101.6391197006636 2.88681056372882, 101.6391197006636 2.88681056372882, 101.6391197006636 2.88681056372882))
159118	POLYGON ((101.6391197006636 2.88681056372882, 101.6391197006636 2.88681056372882, 101.6391197006636 2.88681056372882))
159120	POLYGON ((101.6391197006636 2.88681056372882, 101.6391197006636 2.88681056372882, 101.6391197006636 2.88681056372882))