TTSH-HEAL
Data Science Assignment

By: Enoch Mok

Section 1 [Scenario 2]

There is concern that the experience of the current public transport infrastructure varies significantly across HDB towns.

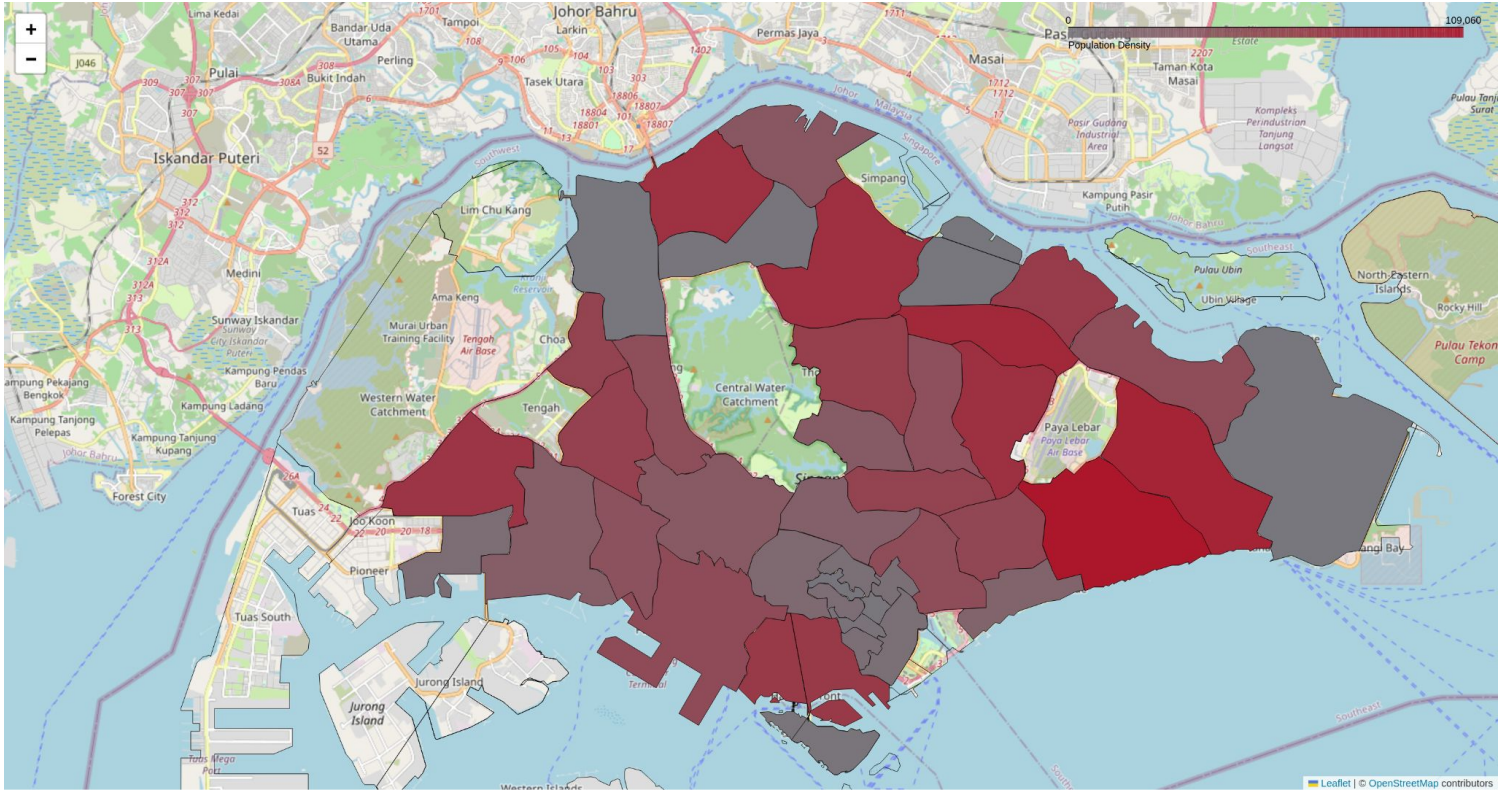
The Land Transport Authority has tasked your team to assess the current infrastructure, and review if there is merit to accelerate the construction of any specific phase of any MRT line that is due to open before 2030.

*The following maps are interactive, and the html files can be found in ./Section1/report folder

HDB Population Density Heatmap

Grey = Less populated

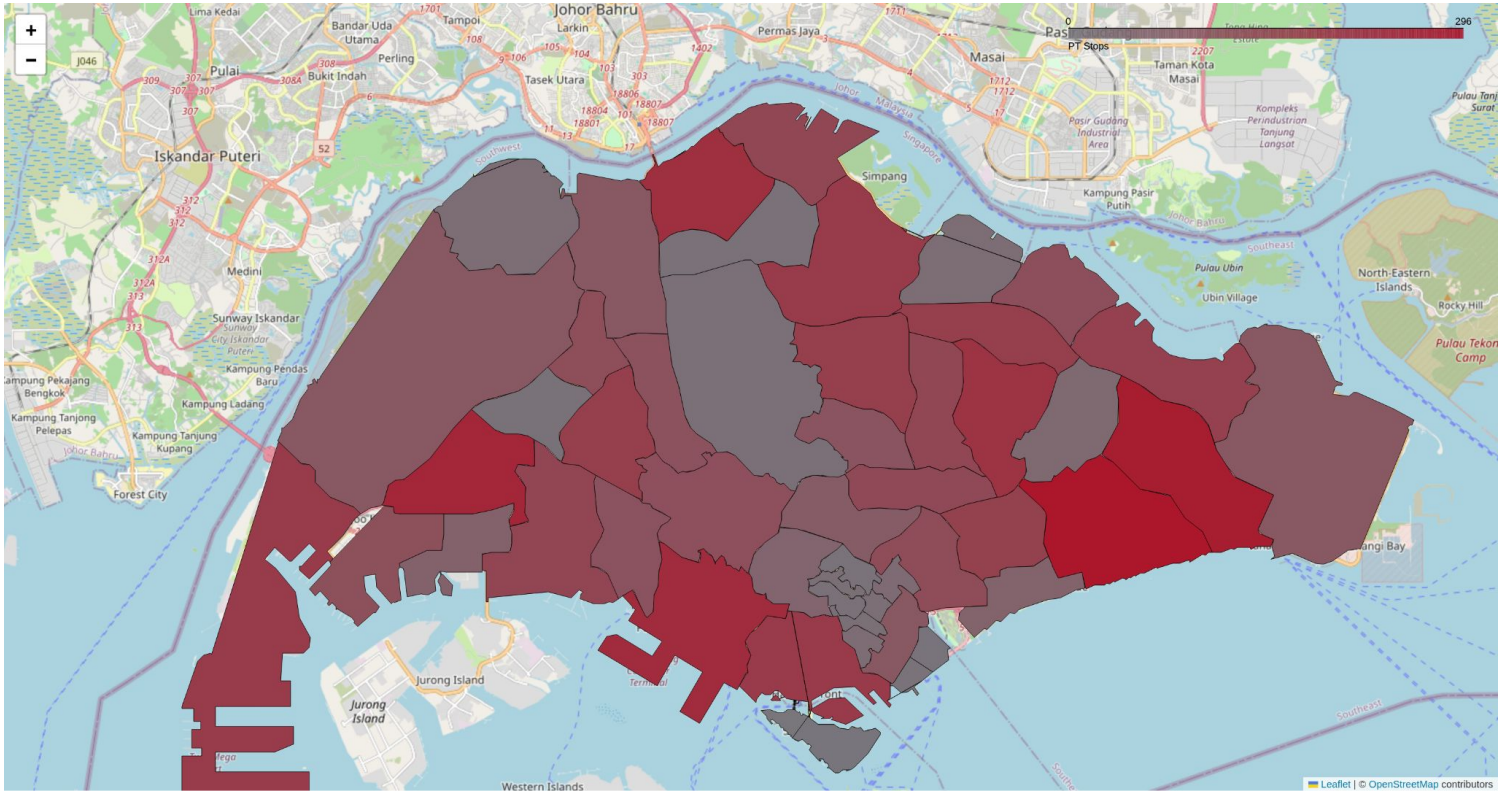
Red = Very populated



Public Transport (PT) Density Heatmap

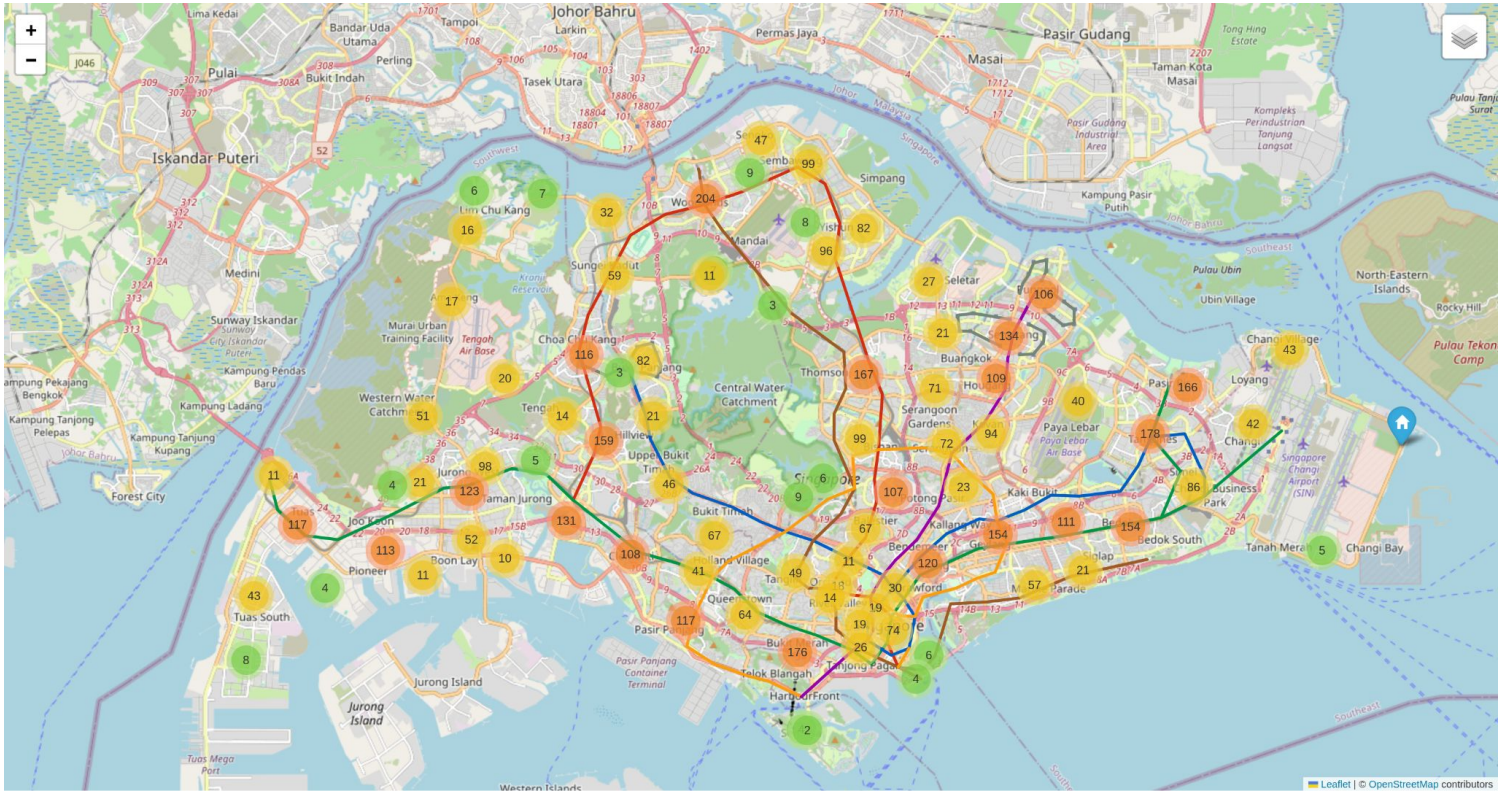
Grey = Regions with not many PT stops

Red = Regions with many PT stops



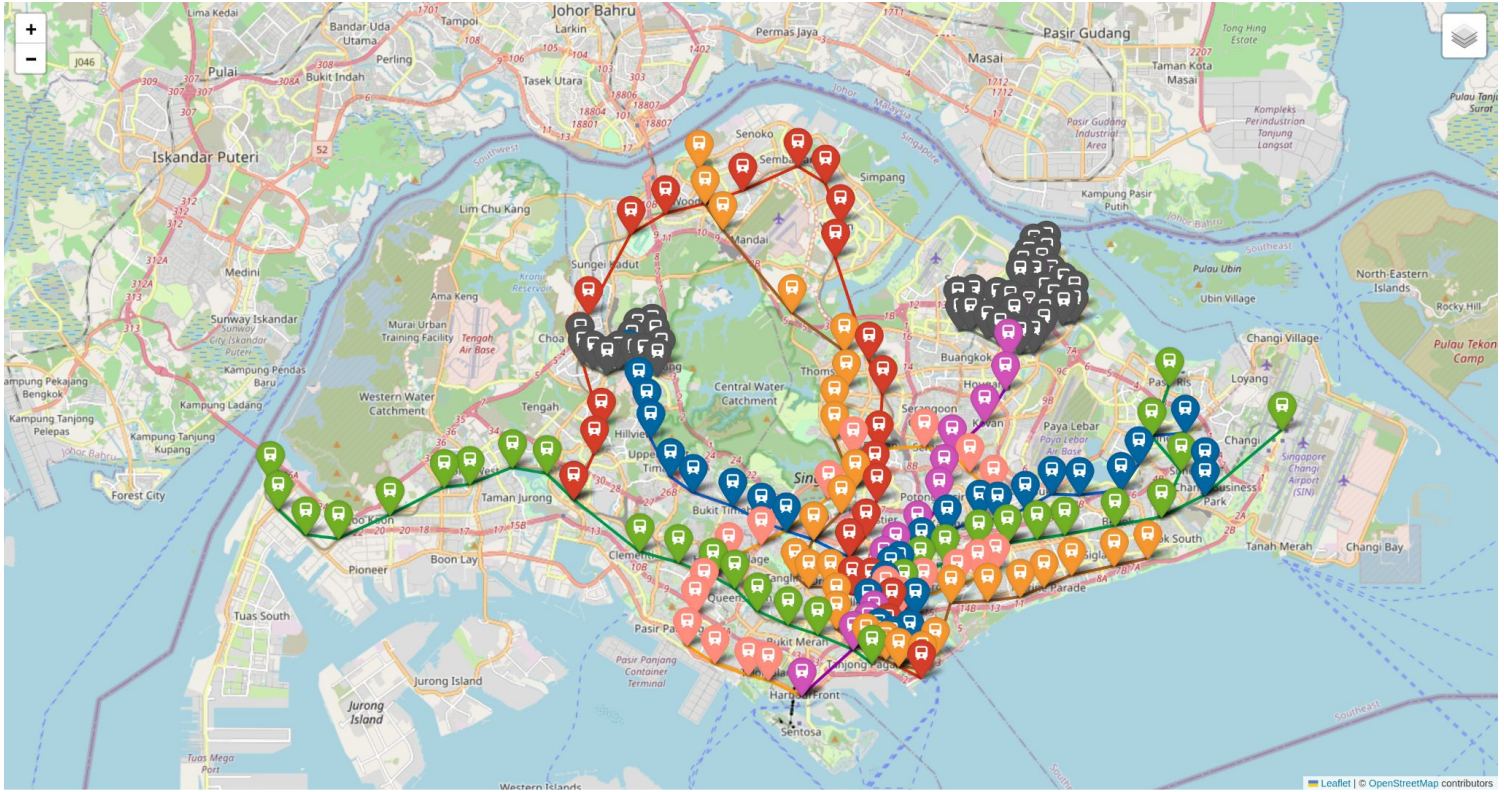
Public Transport Stops (Bus)

The number represents the number of bus stops within that cluster



Public Transport Stops (Train)

Map showing the train stops in Singapore



Public Transport (Bus) Density Discrepancy Heatmap

Below shows the top locations that has the highest demand for bus stops

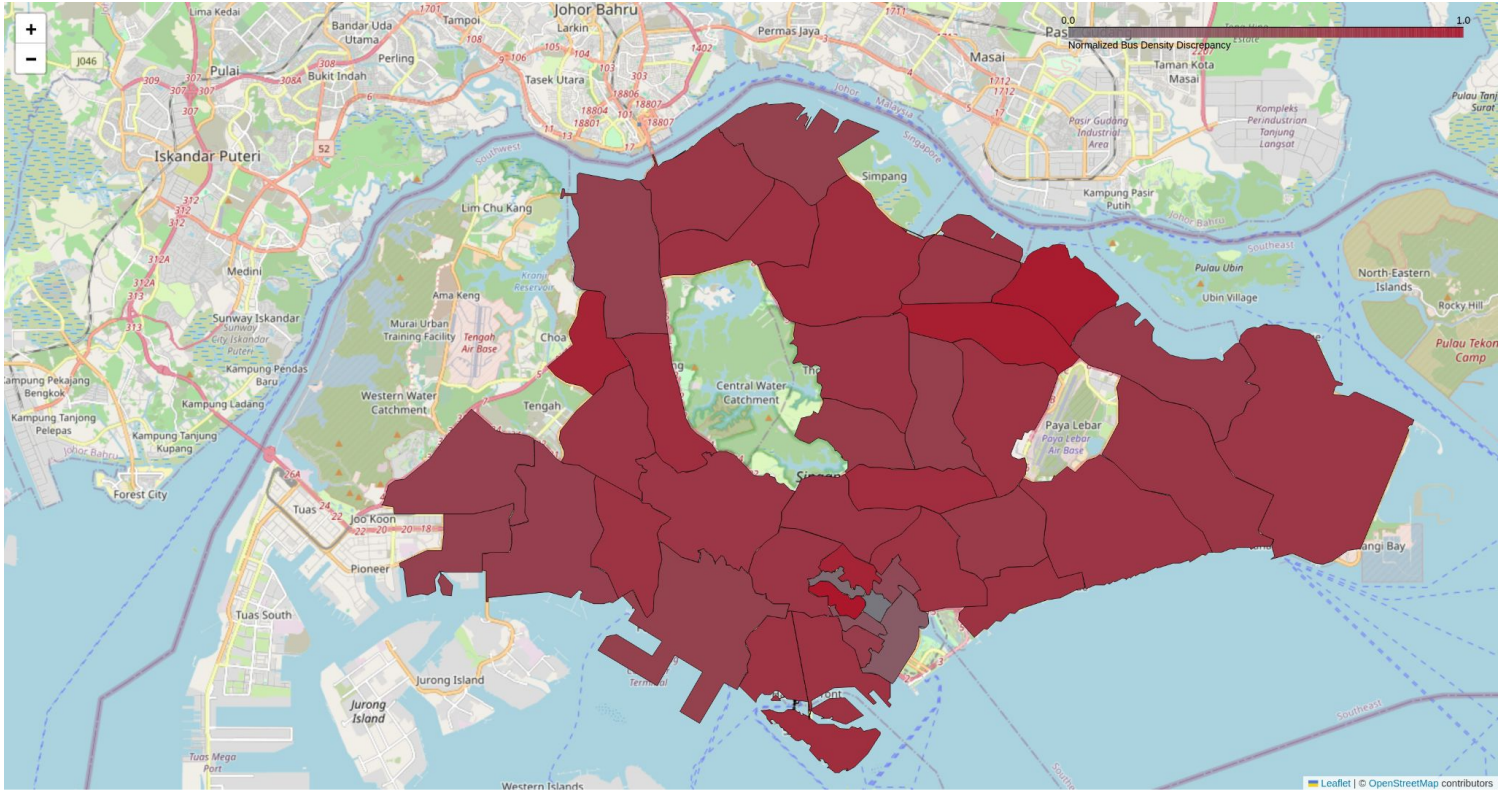
busstop_density_discrepancy = (*normalized population density of area*) - (*normalized bus stop density of area*)

Rank	Area	busstop_density_discrepancy
1	River Valley	0.3414
2	Punggol	0.2707
3	Sengkang	0.2377
4	Choa Chu Kang	0.1777
5	Newton	0.1727

Public Transport (Bus) Density Discrepancy Heatmap

Grey = Regions that do not require more bus stops

Red = Regions requiring more bus stops



Public Transport (Train) Density Discrepancy Heatmap

Below shows the top locations that has the highest demand for train stops

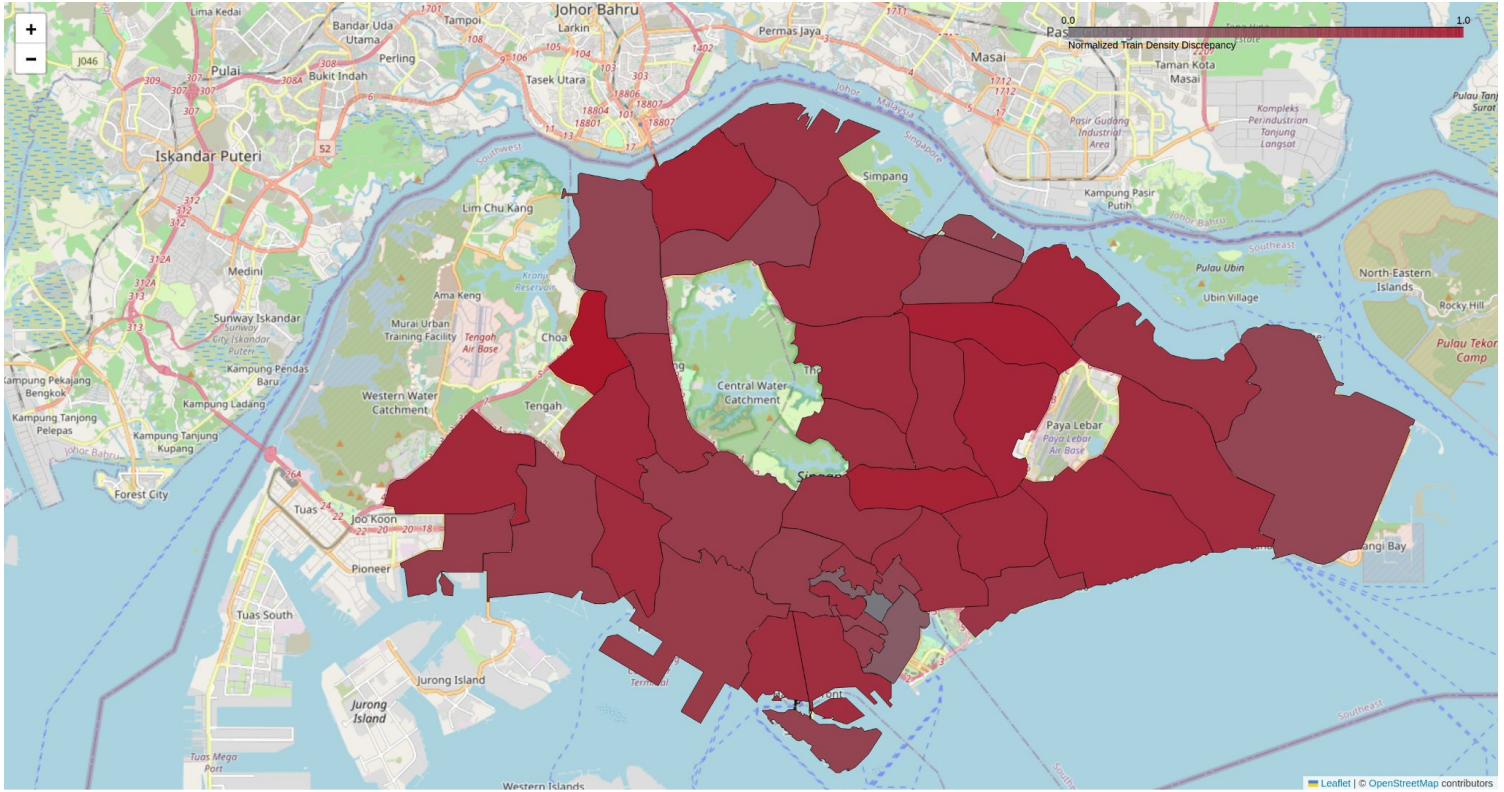
trainstop_density_discrepancy = (*normalized population density of area*) - (*normalized train stop density of area*)

Rank	Area	trainstop_density_discrepancy
1	Choa Chu Kang	0.8190
2	Sengkang	0.6543
3	Toa Payoh	0.6025
4	Hougang	0.5640
5	Woodlands	0.5396

Public Transport (Train) Density Discrepancy Heatmap

Grey = Regions that do not require more train stops

Red = Regions requiring more train stops



Public Transport (All) Density Discrepancy Heatmap

Below shows the top locations that has the highest demand for PT stops

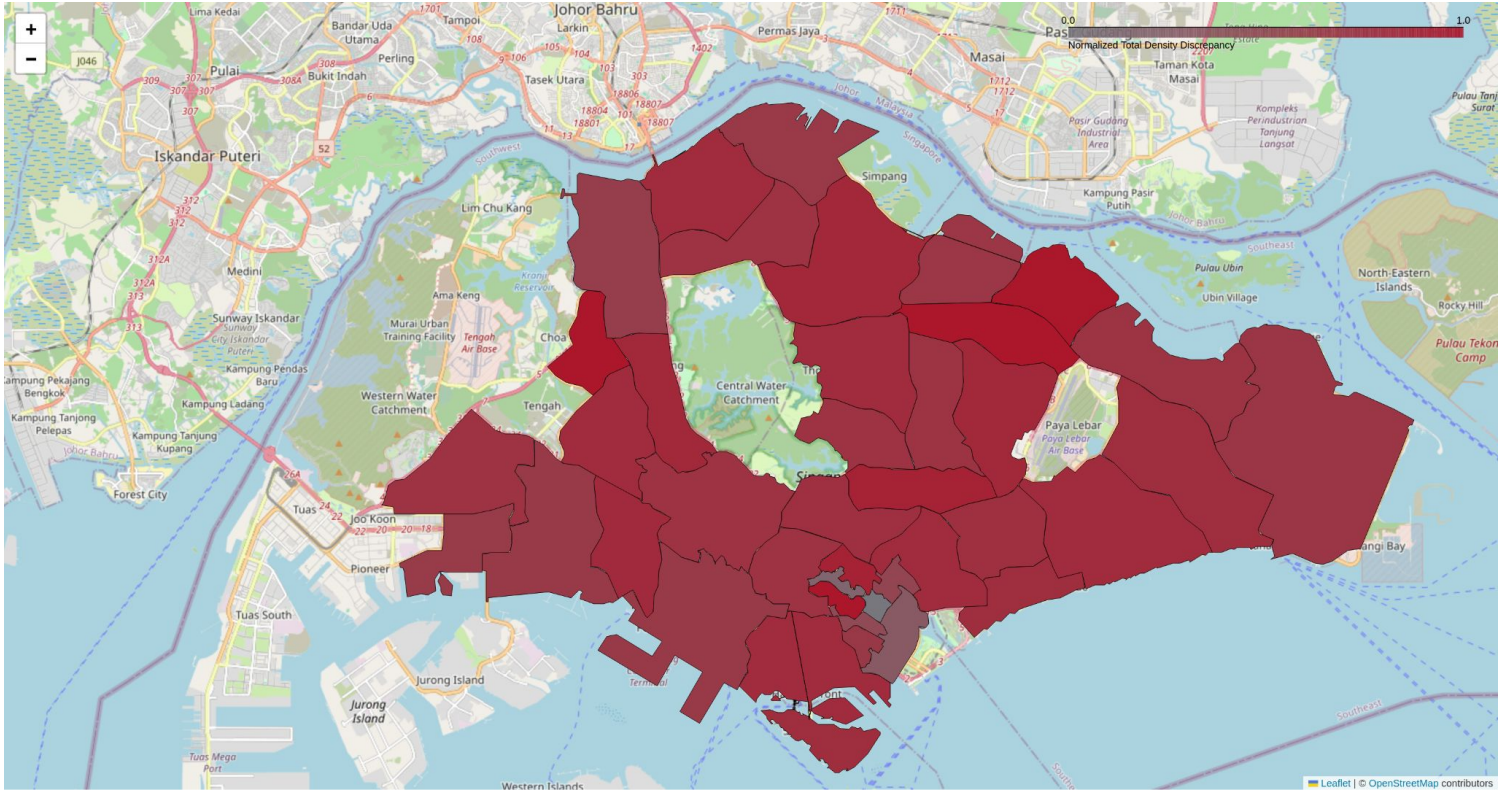
ptstop_density_discrepancy = (*normalized population density of area*) - (*normalized PT stop density of area*)

Rank	Area	ptstop_density_discrepancy
1	River Valley	0.3457
2	Sengkang	0.3405
3	Choa Chu Kang	0.3360
4	Punggol	0.3296
5	Newton	0.1899

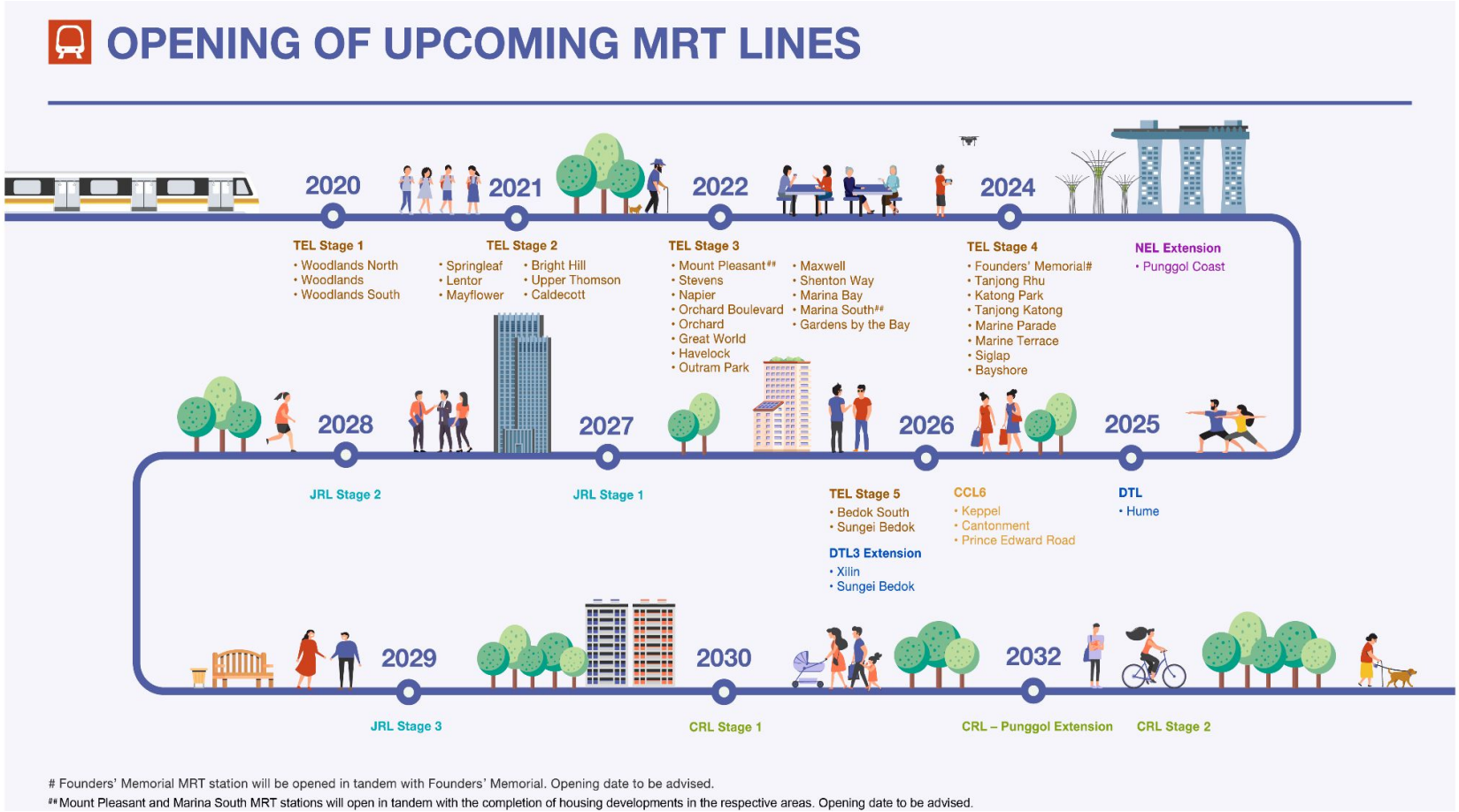
Public Transport (All) Density Discrepancy Heatmap

Grey = Regions that do not require more PT stops

Red = Regions requiring more PT stops



LTA Master Plan 2040



Founders' Memorial MRT station will be opened in tandem with Founders' Memorial. Opening date to be advised.

##Mount Pleasant and Marina South MRT stations will open in tandem with the completion of housing developments in the respective areas. Opening date to be advised.

LTA Master Plan 2040

JRL Phase 1 (2027)	JRL Phase 2 (2028)	JRL Phase 3 (2029)	CRL Phase 1 (2030)	CRL Punggol Extension (2032)	CRL Phase 2 (2032)
JS1/NS4/BP1 Choa Chu Kang	JE1 Tengah Plantation	JS9 Enterprise	CR2 Aviation Park	CP1/CR5/EW1 Pasir Ris	CR14 Turf City
JS2 Choa Chu Kang West	JE2 Tengah Park	JS10 Tukang	CR3 Loyang	CP2 Elias	CR15/DT6 King Albert Park
JS3 Tengah	JE3 Bukit Batok West	JS11 Jurong Hill	CR4 Pasir Ris East	CP3/PE4 Riviera	CR16 Maju
JS4 Hong Kah	JE4 Toh Guan	JS12 Jurong Pier	CR5/EW1 Pasir Ris	CP4/NE17/PTC Punggol	CR17/EW23 Clementi
JS5 Corporation	JE5/EW24/NS1 Jurong East	JW3 Nanyang Gateway	CR6 Tampines North		CR18 West Coast
JS6 Jurong West	JE6 Jurong Town Hall	JW4 Nanyang Crescent	CR7 Defu		CR19 Jurong Lake District
JS7 Bahar Junction	JE7 Pandan Reservoir	JW5 Peng Kang Hill	CR8/NE14 Hougang		
JS8/EW27 Boon Lay			CR9 Serangoon North		
JW1 Gek Poh			CR10 Tavistock		
JW2 Tawas			CR11/NS16 Ang Mo Kio		
			CR12 Teck Ghee		
			CR13/TE7 Bright Hill		

Top 5 areas with highest demand for train stops: Choa Chu Kang, Sengkang, Toa Payoh, Hougang, Woodlands

LTA should therefore expedite the following projects:

- **JRL Phase 1**
- **CRL Phase 1**
- **CRL Punggol Extension**

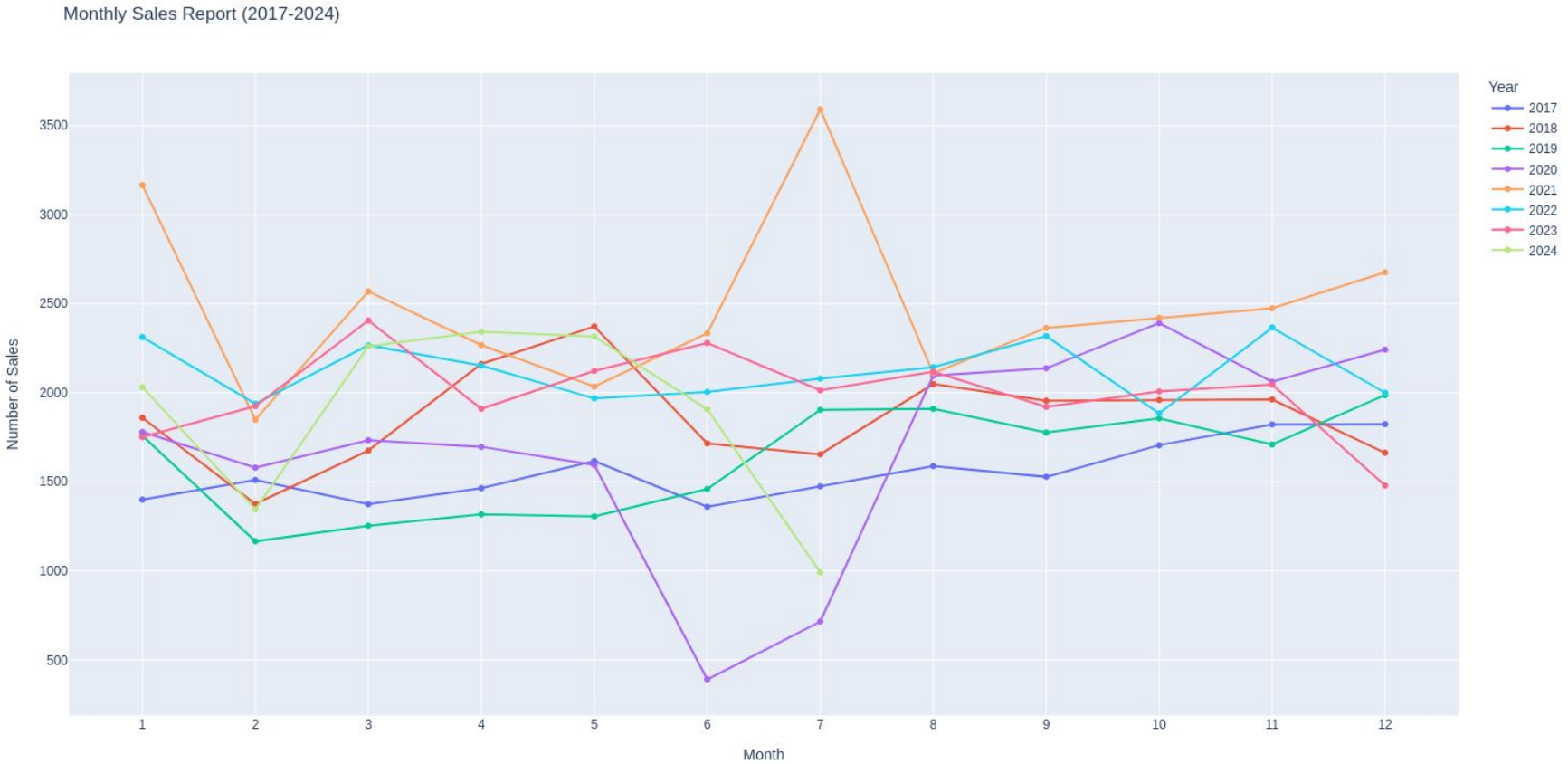
Section 2 [Question 1]

Task: Examine the distribution for number of sales closed by an agent in a year & suggest a probability distribution that may be suitable for modelling this set of values.

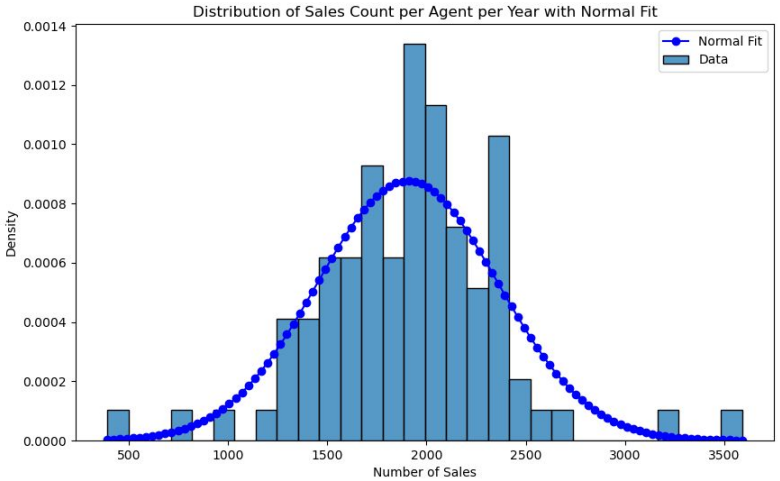
What are some ways in which your suggested distribution is appropriate?

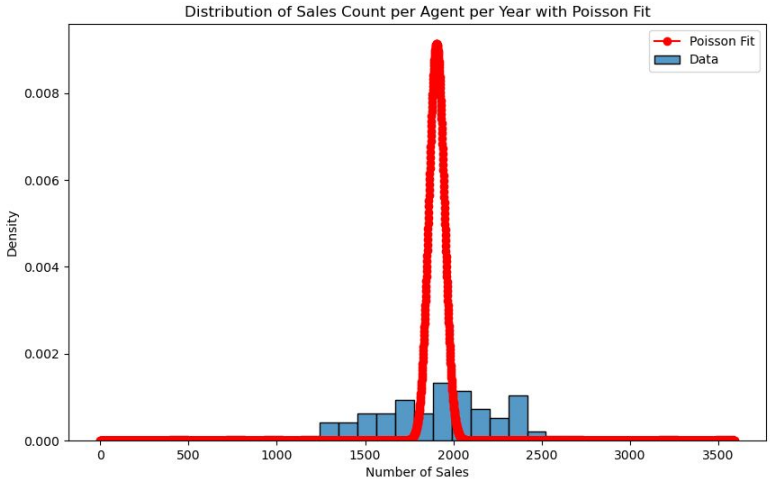
What are some of its limitations?

Section 2 [Question 1]



Section 2 [Question 1]





Based on visual interpretation, distribution of sales count seems to follow a normal distribution better compared to Poisson distribution.

This distribution is appropriate because the number of sales count an agent closes is dependent on multiple independent random variables, such as various client preferences, market conditions and economy at that given point of time.

However, if the data has outliers or is skewed, it will not lead to an accurate normal distribution. This can be seen the outliers at the extreme ends of the left graph.

Section 2 [Question 2]

Task 1: From the table, what are some of the information you can deduce for each hotspot?

Task 2: Due to a system error, the location type column for the last 200 rows of the dataset has become garbled. Using all earlier rows as well as all other columns in the dataset, build a classification model to predict the location type for these hotspots. You may treat the three rarest location types as one category.

Task 3: The information has now been recovered from a backup copy of the file. Compared to the true location types, how good was your model? Be prepared to explain the metrics you use to evaluate your model.

Section 2 [Question 2]

Task 1

From each hotspot, the following information can be extracted:

- Location Coordinates (X-coordinate, Y-coordinate)
- Name
- Location Name
- Location Type
- Postal Code
- Street Address
- Operator Name
- INC_CRC
- FMEL_UPD_D

Section 2 [Question 2]

Task 3

Compared to the true location types, my model was able to achieve a much higher mean CV accuracy (0.7350 vs 0.5400)

However, it is heavily skewed towards F&B location, causing bias towards locations in F&B.
This may be due to the data not being shuffled properly, especially towards the last 200 rows.

Hence, when looking at accuracy, it is better to look at the average cross-validation accuracy instead of the accuracy of one round. Cross-validation ensures that the data within the 200 rows are shuffled properly and that the model is trained and tested on all kinds of data that exists in that 200 rows.

Section 2 [Question 3]

Data Characteristics:

1. There are considerably more students from Group X than Group Y in this course of study.

2. Proportionately more students from Group Y are in jobs unrelated to their course of study.

3. The distribution of students among various industries is considerably different between the two student groups.

4. Students from Group X tend to command higher salaries, for the same type of job & industry.

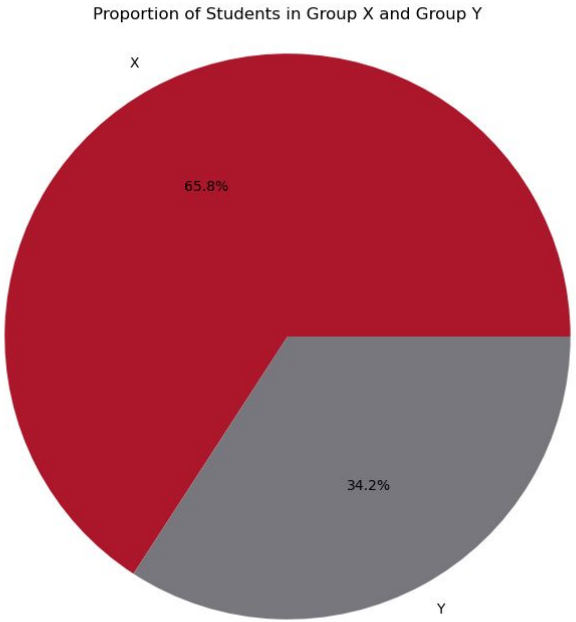
5. The salary differential between the two student groups differs by job nature and industry.

Task: Help your colleague present the insight in an intuitive manner that is easily understood by a non-technical audience, and that reflects as many characteristics in the list as possible. Be prepared to justify any and every aspect of your visualisation (e.g. chart choice, colour palette, labels, orientation, etc.).

Section 2 [Question 3]

Characteristic 1

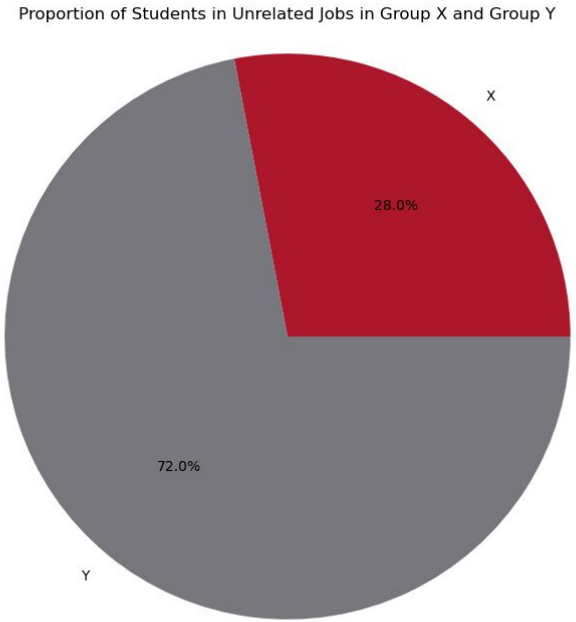
There are considerably more students from Group X than Group Y in this course of study.



Section 2 [Question 3]

Characteristic 2

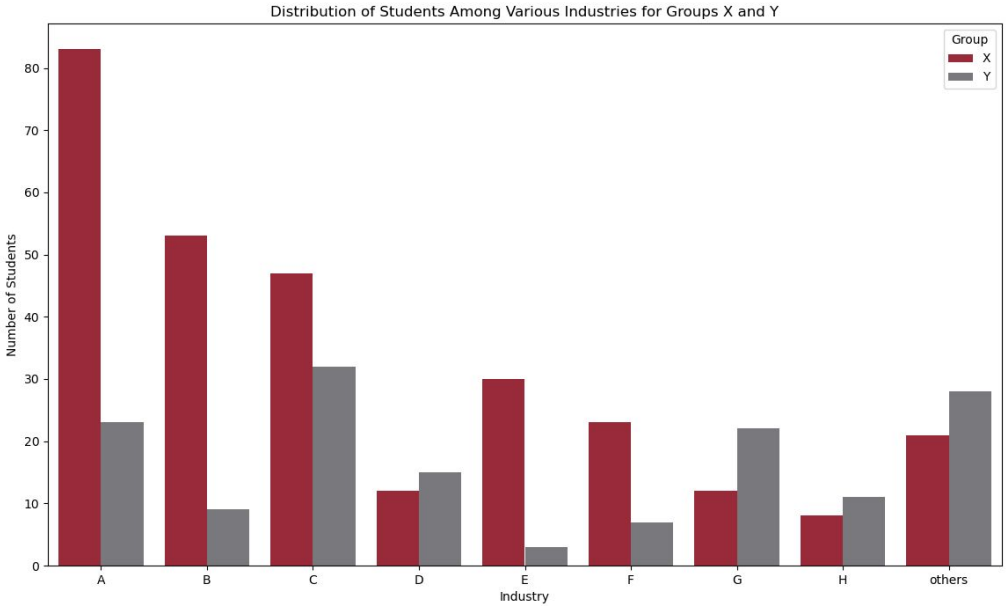
Proportionately more students from Group Y are in jobs unrelated to their course of study.



Section 2 [Question 3]

Characteristic 3

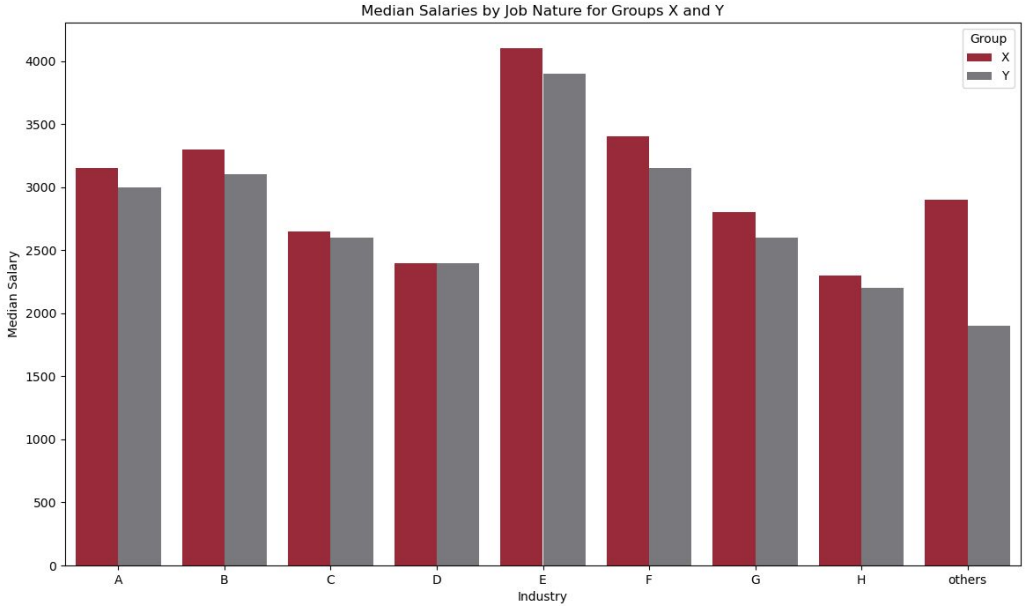
The distribution of students among various industries is considerably different between the two student groups.



Section 2 [Question 3]

Characteristic 4

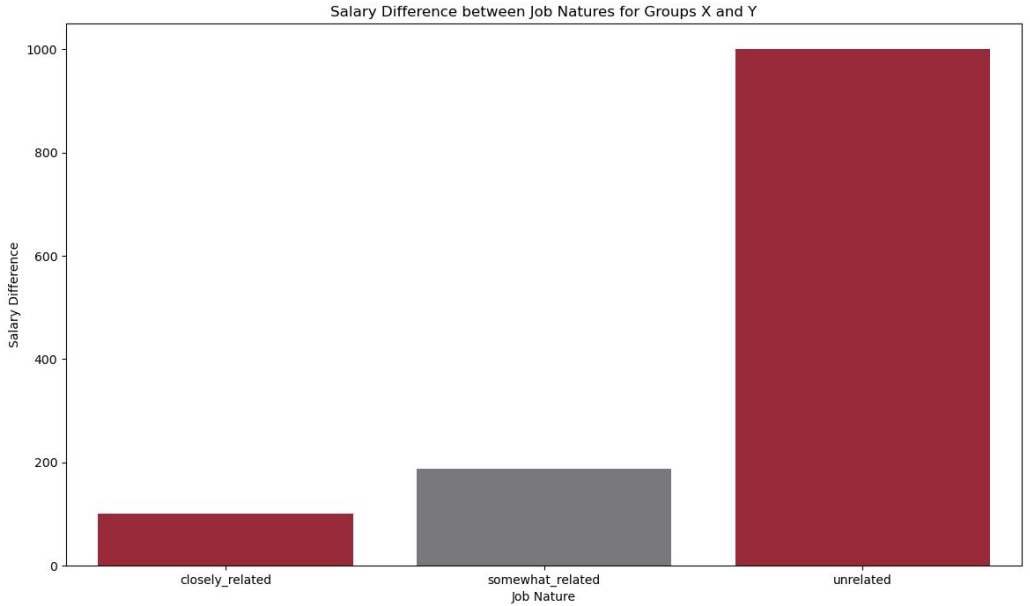
Students from Group X tend to command higher salaries, for the same type of job & industry.



Section 2 [Question 3]

Characteristic 5 (Job Nature)

The salary differential between the two student groups differs by job nature and industry.



Section 2 [Question 3]

Characteristic 5 (Industry)

The salary differential between the two student g
differs by job nature and industry.