Cat 8 Research Project

The optimisation of taxi services in Singapore through data analysis

Group Members:

Tin En Hao 4S326 Leader

Tan Chern Lin, Justin 4S222

Yeoh Yong Jie 4S129

Hwa Chong Institution (High School)

Introduction & Rationale

In Singapore, taxi services have played an important role in transportation with a daily ridership of 785,000 as of 2017. However, based on a survey released by the Public Transport Council, the daily ridership of local taxi companies have tumbled 18% from 2016. This can be attributed to the inefficient system of local taxi companies which has caused commuters to be unsatisfied. Based on a survey conducted by LTA in 201, it was found that 96.4% of the 1200 survey respondents felt that waiting time to be important yet only 86.2% felt satisfied by the waiting time. The 10.2% gap between expectation and satisfaction clearly points out the inefficient system of local taxi services (Adrian, 2015). Another reason for the decline in taxi ridership may be the presence of private-hire car company - Grab in Singapore.

Thus, there is a need to assist local taxi companies as they play a vital role in the field of transportation as well as the economy. Otherwise, local taxi companies may face the threat of being eliminated from the market. This project will make use of open data provided by the Singapore government to find solutions to increase the efficiency of taxi services.

Objectives

1. To create a supply heat map and identify factors which correlate to demand.
2. To identify the shortest path to reach a destination based on the Dijkstra's algorithm.
3. To redistribute the taxis to achieve the ideal allocation of taxis in each region.

Research Questions

1. How to create a supply heat map and identify factors which correlate to demand?
2. How to identify the shortest path to reach a destination based on the Dijkstra's algorithm?
3. How to redistribute the taxis to achieve the ideal allocation of taxis in each region?

Literature Review

Since the late 20th century, extensive studies have been conducted in relation to the taxi sector. The first few studies were focused on studying the profitability of the sector and the necessity of using regulation. Since then, the later studies have evolved to include many other factors such as congestion, elasticity of demand, different market configurations, different user classes etc. to create a more realistic model. The precursor of the first few studies was conducted by Douglas in 1972. His study considered a taxicab market where taxicabs can be engaged anywhere along the city streets, with scheduled (by a regulatory authority) fares, and free entry. He concluded that the maximum revenue to the industry occurs at the point where demand is less than maximum, characterizing social welfare as an efficient but unfeasible (deficit) equilibrium. He also proved that taking into account the social welfare, the points where the number of taxi hours in service is maximized and where demand is max are the same (Josep et al., 2011).

In a recent study, the reasonable analysis of the different time and space residents travel intensity was done and an optimization model was established to achieve the best matching degree of supply and demand. This provides a scientific, reasonable and feasible basis for the management department (Xiaopeng, Xinyuan, Xian, 2016). These studies proposed mathematical formulas for calculating demand and supply, simulating different types of markets and obtaining different results for each regulation scheme. Aggregated models calculated total demand and supply using different parameters. However, these studies only cover the optimal supply of taxis and the ways to generate maximum revenue but not the ideal allocation of taxis across the region. Thus, this project aims to assist in optimizing the allocation of taxis in Singapore.

Dijkstra's algorithm is an algorithm for finding the shortest path from a starting node to a target node in a weighted graph. This algorithm creates a tree of shortest paths from the starting vertex to all the other points. This would generate the shortest distance between each and every node on the weighted graph (Thaddeus, Hannah & Christopher, 2016). The Ford - Fulkerson Maximum Flow algorithm is an algorithm to distribute objects in a weighted graph by maximizing the amount of flow in the network. A maximum flow indicates the most efficient distribution with the minimum cost (Kumar, n.d.). With the help of the Dijkstra's algorithm to

identify the shortest route and the Ford - Fulkerson Maximum Flow algorithm to redistribute the taxis in the most efficient manner, this project aims to identify the ideal positions of the taxis at a given time to optimize the services of the taxi industry.

Methodology

1. Collect data on taxi availability and co-ordinates from Singapore's open data portal
2. Use the data to plot a supply heat map and identify factors which correlate to demand
3. Find the shortest path to reach the destination based on the Dijkstra's algorithm
4. Identify regions in Singapore with an oversupply or undersupply of taxis and apply the Ford - Fulkerson Maximum Flow algorithm to redistribute the taxis

Results

Research Question 1

Methodology:

1. Create an algorithm using R to collect data on taxi availability online
2. Process and clean the data obtained
3. Create a supply heat-map using Tableau
4. Identify factors which correlate to demand

The data for taxi availability used in this project was obtained from `www.data.gov.sg` and it contains the co-ordinates of all available taxis which is refreshed every 30 seconds. The data was collected over a 2 months per minute (1440 files per day). The 10^{th} (Sunday) and 11^{th} (Monday) of June was chosen to create the heat-map (Figure 1) as served as a representation of weekdays and weekends.

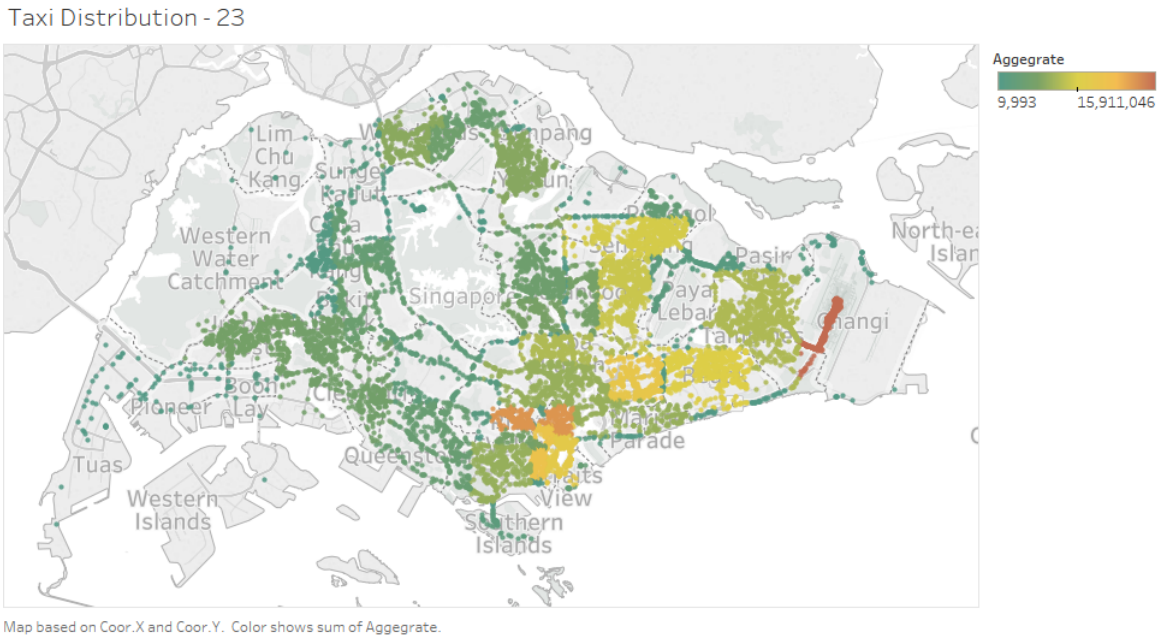

Figure 1: Taxi supply heat-map during 10th June 11pm

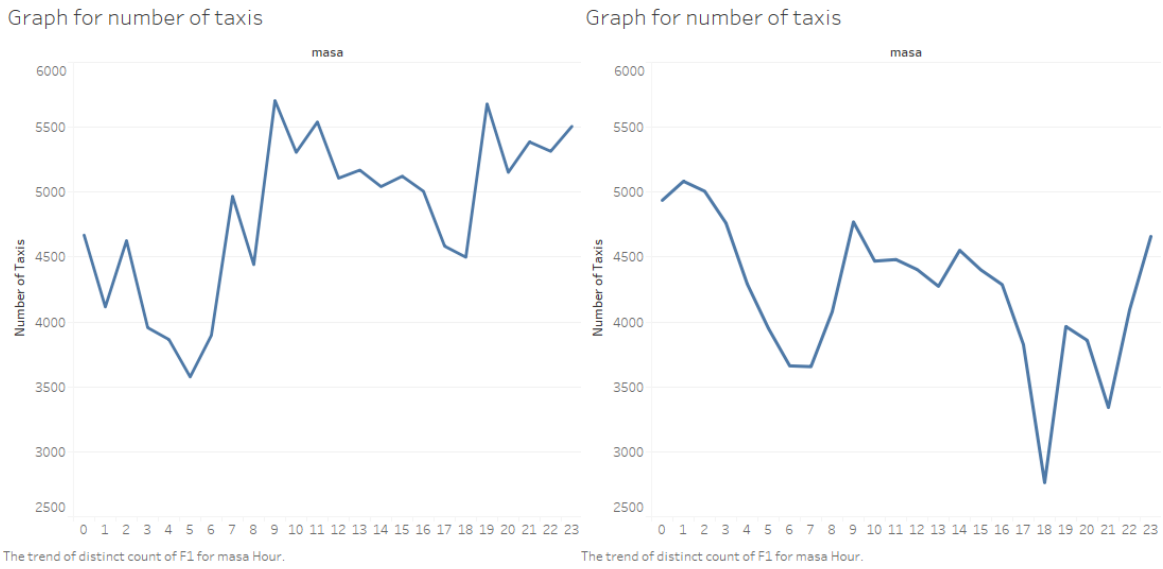

Figure 2: Amount of available taxis over a weekday (left) and a weekend (right)

Drawing a comparison between the graphs (Figure 2) of the amount of available taxis over a weekday and a weekend, it can be observed that there is a higher number of available taxis across all time periods and this can be attributed to the higher amount of taxi drivers working during weekends. Furthermore, it can be observed that both graphs possess similar shape as seen from the same trends in the same time period e.g. sharp decrease in taxi availability during dinner time.

Next, this project made use of the residential population distribution in Singapore obtained from `onemap.sg` and the movement of passengers across the EW line obtained from an analysis

done online as a proxy to identify areas with high demand. However, this can only be achieved by making the assumption that the supply of taxis correlate to the demand. It was predicted that the higher the residential population or the movement of passengers in the region, the higher the demand of taxis.

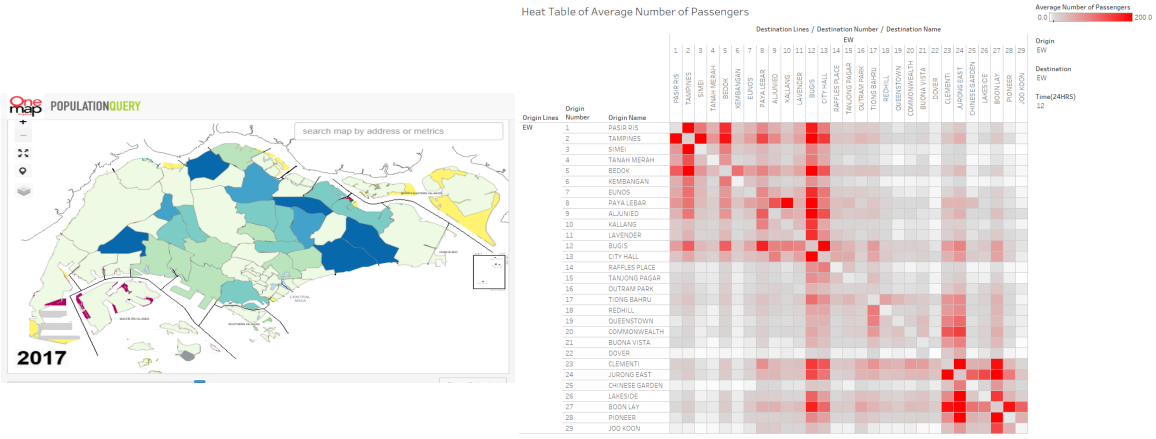

Figure 3: Population distribution (left) Movement of passengers in EW line (right)

Comparing these factors (Figure 3) to the taxi availability heat map, it can be concluded that our hypothesis holds true that the higher the residential population or movement of passengers in the region would result in higher demand for most cases. For instance places like Geylang, Bedok, Tampines etc. with high residential population has a high demand of taxis. However, places like Jurong, Clementi etc. have a high residential population but low demand for taxis. Furthermore, the movement of passengers at Raffles Place, Outram, Tiong Bahru which has a high demand for taxis has a low movement of passengers in the MRT. Thus, it can be concluded there is either an under-supply of taxis or the proxy used is inaccurate. Although the comparison of data between the taxi distribution(1 day) and the movement of passengers in MRT(2 months) may be inaccurate, but this comparison only serves as a proof of concept to determine the feasibility of this comparison. However, a concrete conclusion cannot be made due to the lack of data.

Research Question 2

Dijkstra's algorithm is an algorithm to find the shortest path in an weighted graph by creating a tree of shortest paths.

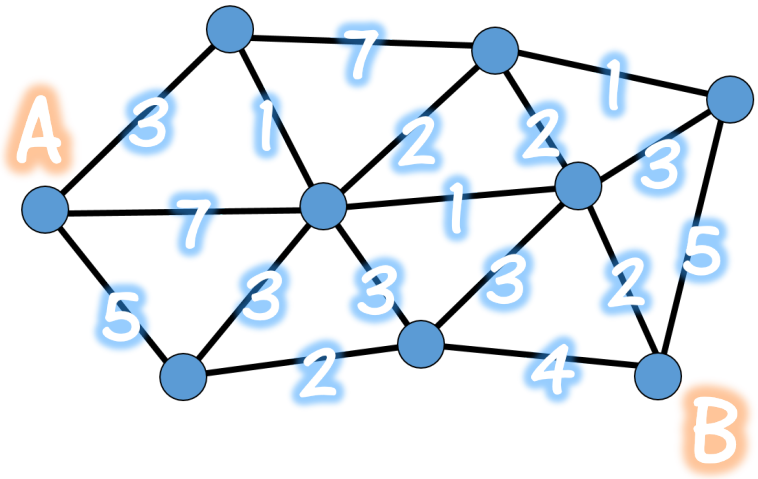

Figure 4: Example of a weighted graph

In the weighted graph (Figure 4), A and B represents the starting point and the destination whereas the black lines denote all roads and the blue circles denote different locations. The numbers denote the length of the roads. Below is the breakdown of the algorithm:

Definition of variables used in the algorithm

Vertices denoted as u or v

Weighted edges that connect 2 nodes; (u, v) denotes an edge & $w(u, v)$ denotes its weight.

Initialization

$dist \rightarrow$ an array of distances from source node A to other nodes in the graph

At the start of the program,

$dist(A) = 0$

All other nodes $v, dist(v) = \infty$

$Q \rightarrow$ a set of all nodes in the graph (at the end, Q will be empty)

$S \rightarrow$ an empty set which indicates which nodes have been visited (at the end, S will contain all nodes)

Algorithm

While Q is not empty, pop the node v

$\rightarrow$ Which is not in S

$\rightarrow$ Smallest $dist(v)$

Add node v to S to indicate it has been visited

Update each adjacent node u of the current node v as such:

$\rightarrow$ If $dist(v) + w(u, v) < dist(u)$, update $dist(u)$ to new minimal distance

Otherwise, no updates

Methodology:

1. Download the Singapore map from open street map
2. Identify all roads in Singapore
3. Use geosphere to calculate the distances of the roads
4. Apply Dijkstra's algorithm on the map

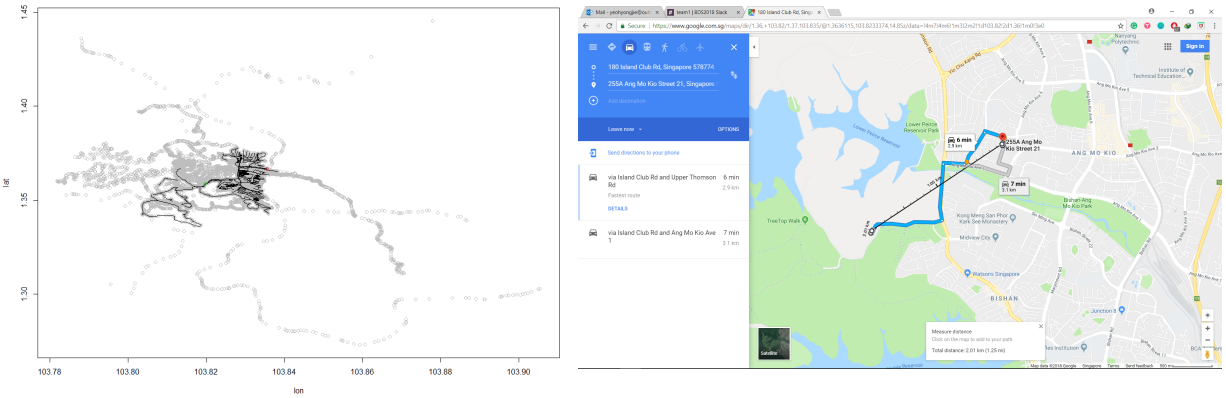

Figure 5: Results for Dijkstra's algorithm (left) Results for Google maps (right)

Both values, the starting point(103.818, 1.3575) and the ending point(103.818, 1.3750) were input into the Dijkstra's algorithm and the result was 2660.855m. Comparing this value with the value from Google Maps which is 2900.000m, it can be concluded that the Dijkstra's algorithm is rather accurate with an error margin of 8.00% which is 239.145m.

Research Question 3

Methodology:

1. Divide Singapore into various regions and find the centre of each region
2. Group the locations into points of oversupply and undersupply computed by comparing the base - case with the current data. The base - case is calculated by the average of all the data sets collected and divided into each hour and region.
3. Find the distance between each point using Dijkstra's algorithm
4. Apply Ford - Fulkerson Maximum Flow algorithm on these locations to redistribute the taxis

Terminology:

Flow Network

- Digraph
- Weights, called capacities on edges
- Source S and Sink T

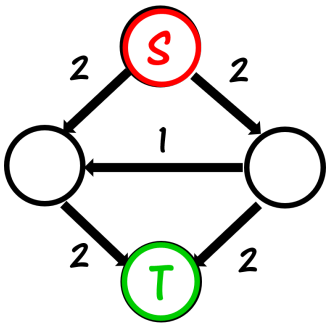

Figure 6: Example of a Flow Network

Flow

Function on network edges

$0 \leq flow \leq capacity$

$flow$ into vertex $= flow$ out of vertex

Value of $flow = flow$ into sink

Maximum $flow \rightarrow$ Maximum Distribution
Minimum Cost

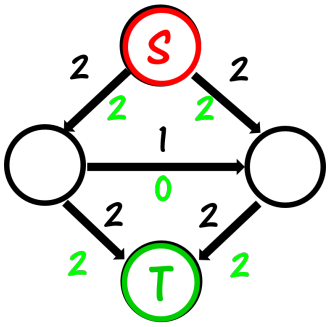

Figure 7: Example of a Flow Network

Ford - Fulkerson Maximum Flow Algorithm

Find an augmenting path → Compute the bottleneck capacity → Augment each edge and total flow

Augmenting path → A path from Source to Sink made of non-full forward edges and non-empty backward edges.

Bottleneck Capacity → The bottleneck capacity of an augmenting path is the minimum residual capacity (capacity−flow) of any edge in the augmenting path

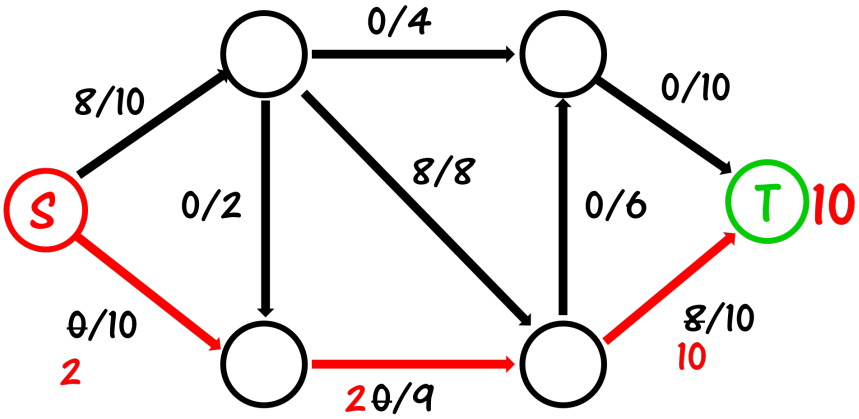

Figure 8: Ford - Fulkerson Algorithm

As seen in Figure 8, the red lines represent the augmenting path. Since the bottleneck capacity is 2, augment flow value of 2 along the augmentation path. Value of flow is updated accordingly.

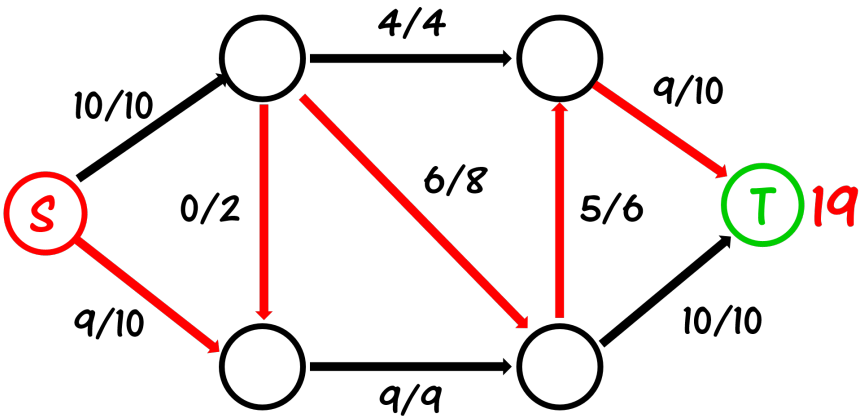

Figure 9: Maximum Flow

It can be observed from Figure 9 that an augmenting path does not exist in the graph, indicating that the Maximum Flow - 19, has been found.

Program algorithm:

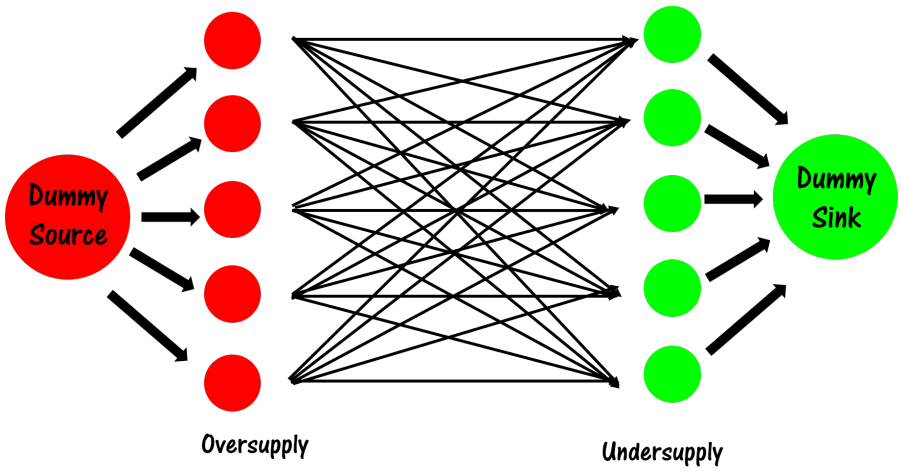

Figure 10: Program Algorithm

1. Group all regions with oversupply and undersupply which will be connected to the Dummy Source and Dummy Sink respectively.

2. The weights on the nodes connecting the Dummy Source and Dummy Sink with the regions would be determined by the amount of over/under supply of taxis

3. Each region with an oversupply would be connected to all the regions with an undersupply and vice versa.

4. Perform a Binary Search to identify the Minimum distance with Maximum Flow. This will reduce the number of times needed to run the Ford - Fulkerson Maximum Flow Algorithm to optimize the speed of the program

5. When the 2 regions with the minimum distance has been identified and has maximum flow, taxis would be distributed through the node. This process would than repeat until all the taxis have been distributed.

6. If the base case has more taxis than the current data, the average amount of taxis per region in the base case would be decreased by a similar percentage across all regions. Similarly, if the current data has more taxis than the base case, the average amount of taxis per region in the base case would be increased by a similar percentage across all regions. This is to ensure that each region has as little over/under supply of taxis as possible.

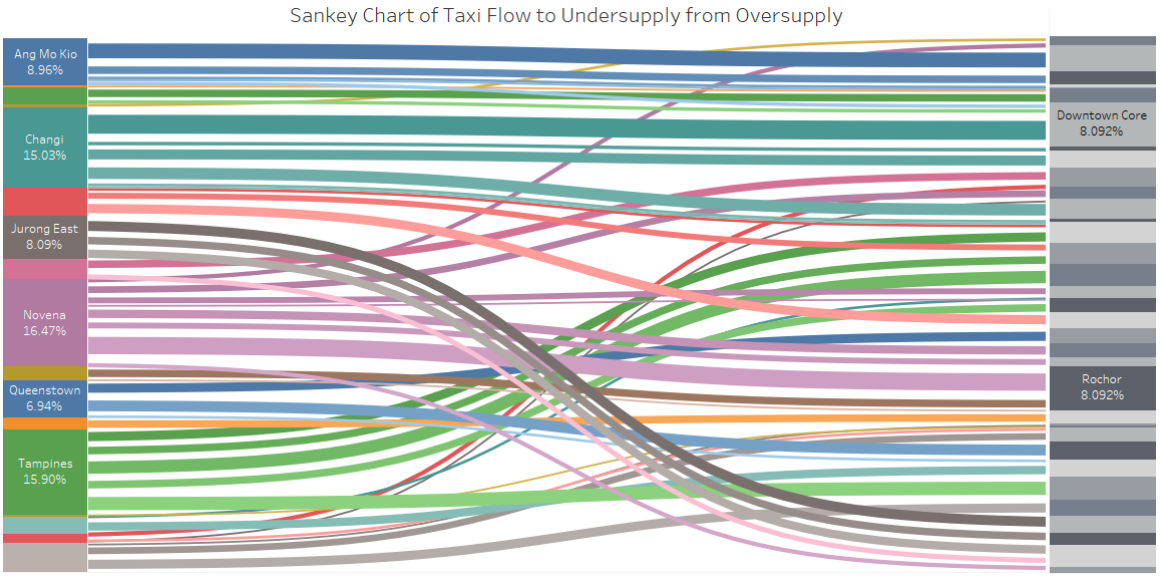

Figure 11: Sankey diagram

The sankey diagram(Figure 11) shows the amount of taxis allocated from the region with oversupply(Left column) to the regions with under-supply(Right column) on 24th July 1400 hours. The thickness of the node indicates the amount of taxis allocated in each node i.e. the thicker the node, the higher the amount of taxis transferred.

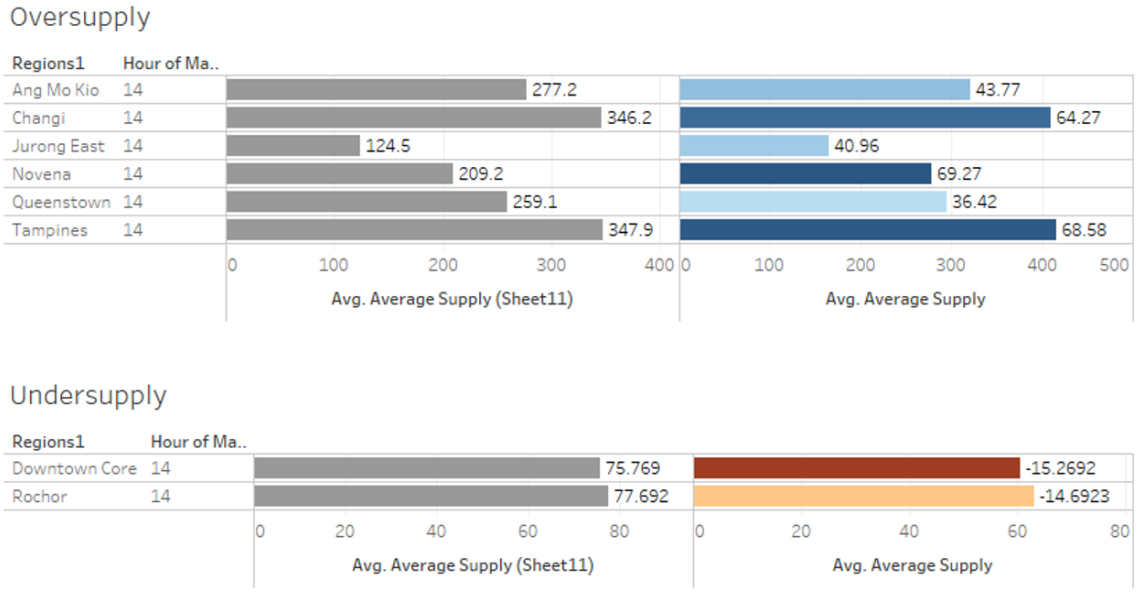

Figure 12: Part of the data collected on 24th July 1400 hours

Based on the Sankey diagram (Figure 11), it can be observed that there are a few regions with high amounts of over/under supply of taxis. Comparing it with the data collected on 24th July 1400 hours (Figure 12), it can be concluded that the program is a success.

Limitation

One limitation of this project is the lack of time to collect data for a long term project. For this model to be accurate months and years of data have to be collected to form the base case. Another limitation of this project is the lack of professional knowledge in the field of data science which has limited the scope and depth of this project and the accuracy of the results. Despite the various limitations, this project can still serve as a model or prototype which can be applied in real life.

Further Extension

The taxis can be examined on an individual level instead of regions. By looking at the taxis on the individual level, better results can be achieved. The allocation of taxis can be more specific as now individual taxis are directed to move instead of a number for each region.

Timeline

Time	Work to be Done
Term 1 Week 9 -10	Complete Project Proposal Start on Powerpoint Slides
March Holidays	Complete Powerpoint Slides
Term 2 Week 1	Submit Proposal
Term 2 Week 2	Project Rehersals
Term 2 Week 3	Proposal Evaluation
Term 2, Week 4 - 10	Start on Research Paper
June Holidays	Complete Mid-Term Evaluation Slides
Term 3 Week 1-2	Project Rehearsals
Term 3 Week 3	Mid-Term Evaluation
Term 3 Week 4-7	Complete Powerpoint Slides Completion of Research Paper
Term 3 Week 8	Final Evaluation

References

[1] Adrian L. (2016). Waiting time for taxis still biggest bugbear: Study. Retrieved April 20, 2018, from https://www.straitstimes.com/singapore/transport/waiting-time-for-taxis-still-biggest-bugbear-study

[2] Kumar, V. (n.d.). Maximum Flow. Retrieved August 3, 2018, from https://www.hackerearth.com/practice/algorithms/graphs/maximum-flow/tutorial/

[3] Josep, M.S., Miquel, E., Georgia, A., & Evangelos, M. (2011). A review of the modeling of taxi services. Retrieved August 6, 2018, from https://www.sciencedirect.com/science/article/pii/S1877042811014005

[4] Thaddeus, A., Hannah, P., & Christopher, W. (n.d.). Dijkstra's Shortest Path Algorithm. Retrieved April 23, 2018, from https://brilliant.org/wiki/dijkstras-short-path-finder/

[5] Wing, Y. & Kai, W. (2017). MRT Closures: What is the impact?. Retrieved June 20, 2018, from https://blog.data.gov.sg/mrt-closure-what-is-the-impact-9472027ccf4e

[6] Xiaopeng L., Xinyuan D. & Xian C. (2016). A Study on Optimal Resource Allocation of Taxis. Retrieved July 15, 2018, from https://www.matec-conferences.org/articles/matecconf/pdf/2016/26/matecconf_mmme2016_04009.pdf