American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS)
ISSN (Print) 2313-4410, ISSN (Online) 2313-4402
© Global Society of Scientific Research and Researchers
http://asrjetsjournal.org/

Mining Metro Data to Determine Taxi Distribution in Peak Hours

Zainab Al Kashari[a], Fatma Al Taheri[b]*

[a]The British University in Dubai, Dubai, United Arab Emirates
[a]Email: zainabalkashri@gmail.com
[b]Email: fatimaaltaheri94@gmail.com

Abstract

Public transportation activities and processes can provide valuable data and information that be useful for public transportation management. As a matter of fact, metro ridership data can be used to assess ridership and traffic flow, which can be useful for building interconnected public transportation system. For instance, metro ridership can be used to understand intermodal relationship with taxi systems, further enabling taxi companies to capitalize on traffic flow and pressure that can be observed in metro ridership datasets. Different data mining and data processing techniques can be of great use for determining taxi distribution based on metro ridership dataset. This study used data mining techniques to determine the traffic and transaction pressures on metro stations in Dubai, further identifying the peaks hours to direct the taxi drivers to the best destination. Dataset was collected from Dubai Pulse, and preprocessed and manipulated. Findings of this study indicated that high traffic pressures in Dubai Metro Red Line zone six, with peak points at three different times. Taxis can use traffic pressures at each station, and calculate arrival time to stations for optimal travel and revenue.

Keywords: Data Mining; Taxi Distribution; Metro Ridership; Path Planning; Trip Planning; Transportation.

1. Introduction

Taxis are considered to be a major on-demand public transportation in several countries, providing passengers with services without the need for owning private vehicles.

----

* Corresponding author

More so, taxis, as a transportation mode, could play significant role in satisfying the transport and travel demands that are unmet by other transportation modes. Taxi networks can be used to supplement other public transportation modes, with multi-modal relationships between taxi systems and other public transport systems (e.g. metro, bus) can be assessed to understand and better manage effective transportation system. Apparently, there are limited number of studies that assessed on the multimodal relationship between taxis and metro systems [4], and taxis, specifically, received less attention in research mainly because taxi services' nature of being non-subsidized, making it difficult to conduct research concerning the cost-effectiveness of taxi services and investments [5]. This research paper presents a study based on a dataset provided by Dubai pulse to understand the peak hours and traffic/transaction pressures at metro stations in Dubai. The main purpose is to determine the traffic and transaction pressures on metro stations in Dubai, further identifying the peaks hours to direct the taxi drivers to the best destination; further fostering intermodal shifts from metro to taxi, and vice versa. The methodology applied in this paper focused on clustering using a data science software tool. The paper is organized as follows: (1) review of related literature, (2) research goals, (3) description of the dataset used, (4) implemented research approach, (5) data preprocessing, (6) methods used, (7) presentation and discussion of findings, and (8) conclusion statement.

2. Related Work

Existing literature on transportation is extensive, with several studies used data mining techniques and tools to understand public transportation ridership, travel patterns, and intermodal relationships between different modes of transportation. For instance, the study explored and analyzed taxi ridership in Shanghai, with the aim of understanding travel patterns of taxis and the relationship between taxi ridership and other external factors affecting taxi system and travel patterns. Using data mining tools, their study found different factors affecting taxi ridership, including population density, car ownership, employment and many others [3]. On the other hand, aimed at understanding the relationship between short-term subway ridership and other external factors using a novel predictive methodology [1]. By collecting data sets from smart cards, the authors used gradient boosting decision tree that allowed effective assessment on the potential factors that could be used to improve short-term subway ridership. There are few studies that explored on the intermodal relationship between taxis and other public transportation systems, specifically subway transits or metro systems. For instance, Jiang and his colleagues (2018) investigated on the intermodal relationship between taxi and subway in China. Collection and analyzing four datasets allowed the authors to: (1) determine the competition between taxi and subway trips, (2) identify tax trips connecting to subway stations, (3) determining transit trips complementing taxi trips, and (4) arrival and departure trips around stations. Similar study by Wang and Ross explored on the multimodal connection between taxi and transit system in New York City. Processing and analyzing collected datasets indicated that transit-extending taxi trips having shorter average trip lengths, but passengers are paying more in comparison to other trip types [5].

3. Goal

The goal of this research is to use rapid miner to determine the pressure of transactions on the metro stations and the peak hours at each station, further aiding in directing taxi services to the best destination for high revenue.

More so, the study can contribute to:

- Avoid the random movement of taxi drivers
- Increase the taxi driver's revenue
- To do map check the purser area
- Go green - fewer gasses in the air
- Decrease the percentage of crowded streets
- Increase happiness level among passengers

4. Dataset

The data used in this research is an actual dataset gathered from Dubai Pulse. Dubai Pulse is a central platform created under the partnership between Smart Dubai Office and du, providing a hub for sharing and accessing different data from a wide range of data sources [2]. This research collected Dubai metro ridership dataset provided by the Rail and Transport Authority, and accessed through Dubai Pulse. The metro ridership dataset consists of over 7,339,880 records.

5. Proposed Approach

Data manipulation will involve clustering as well as Rapid Miner data exploration tools and functions are used to determine the expected distribution of demand or pressure points and times at each zone of the Dubai Metro Red Line. The data manipulation techniques are also used to determine the arrival time of taxis and the benefits for visiting the station at a specific time window. Prior to manipulating the data using Rapid Miner, some pre-processing or normalization of data was conducted, as discussed in the next section of this report.

6. Data Preprocessing

Before further analytical manipulation is conducted on the data set, the data collected are pre-processed. The study conducted normalization of the data. Initially, the dataset collected consisted of 7,339,880 records, and after the normalization, the dataset now consist of 339,627 records. The normalization step involved removing records dating March 30, 2018, and keeping records dating March 31, 2018. Data pre-processing also involves:

- Chose recorded dataset for Dubai Metro Red Line only
- Selected the start location, and removing the end location
- Time format is in hours; removed minutes and seconds
- Renamed the zone to numeric identifiers – e.g. 'Zone1' is denoted as 1
- Added ID number for each entrance
- Selecting only the start zone, and removing the end zone
- Changed the name of the location from the Station name to numeric code (station number) – e.g. 'Airport3' is changed to 25.

Table 1: Pre-processed dataset

Station Number	Zone	Station Name
1	1	UAE Exchange
2	1	Danlube
3	2	Energy
4	2	Ibn Battuta
5	2	Jumirah Lakes
6	2	DAMAC Properties
7	2	Kaheel
8	2	Dubai Internet City
9	2	Sharaf DG
10	2	Mall of the Emirates
11	2	First Abu Dhabi Bank
12	6	Business Bay
13	6	Burj Khalifa/ Dubai Mall
14	6	Financial Center
15	6	Emirates Towers
16	6	World Trade Center
17	6	Al Jafillya
18	6	ADCB
19	6	Burjuman
20	5	Union
21	5	Al Rigga
22	5	Dira City Center
23	5	GGICO
24	5	Airport Terminal 1
25	5	Airport Terminal 3
26	5	Emirates
27	5	Al Rashidiya

7. Methodology

The processing of the data set was conducted using the Rapid Miner. Rapid Miner is a data science software platform that can be used for various processes and activities involved in data science, including data preparation, text and data mining, predictive analytics and many others. Clustering is conducted in Rapid Miner, and data processing involved processing the available information related to the metro location and the number of passengers dropped off at each location. The processing will allow to establish the expected distribution demand for the number of passengers to pick up from a specific Dubai Metro Red Line station.

More so, utility-like process control operations in Rapid Miner are used to help in determining the expected time of arrival of taxis for each metro (I) and arrival time (AT).

8. Findings and Discussions

In order to establish the expected distribution demand for the number of passengers to pick up from each metro station, data processing on the dataset related to the metro location and the number of passenger drop off at each location are conducted. The processing involved dividing the number of working hours of each day into 17 intervals, with the first interval starts at 5 am and the last interval end at mid-night. For each metro station, the study aggregated the number of passengers selecting specific metro stations as drop off location. This will result to a 2 dimensional table, in which each cell represent the number of passenger dropped off at the specific location during the given time window. To simplify the process of calculating the demand, normalization of the aggregation values for each time window are conducted. The highest number of passengers dropping off at the metro station is identified, denoted by HC. Using a visualization tool, TABLEAU, the number of transaction or traffic pressure at each zone are identified and visualized (Figure 2).

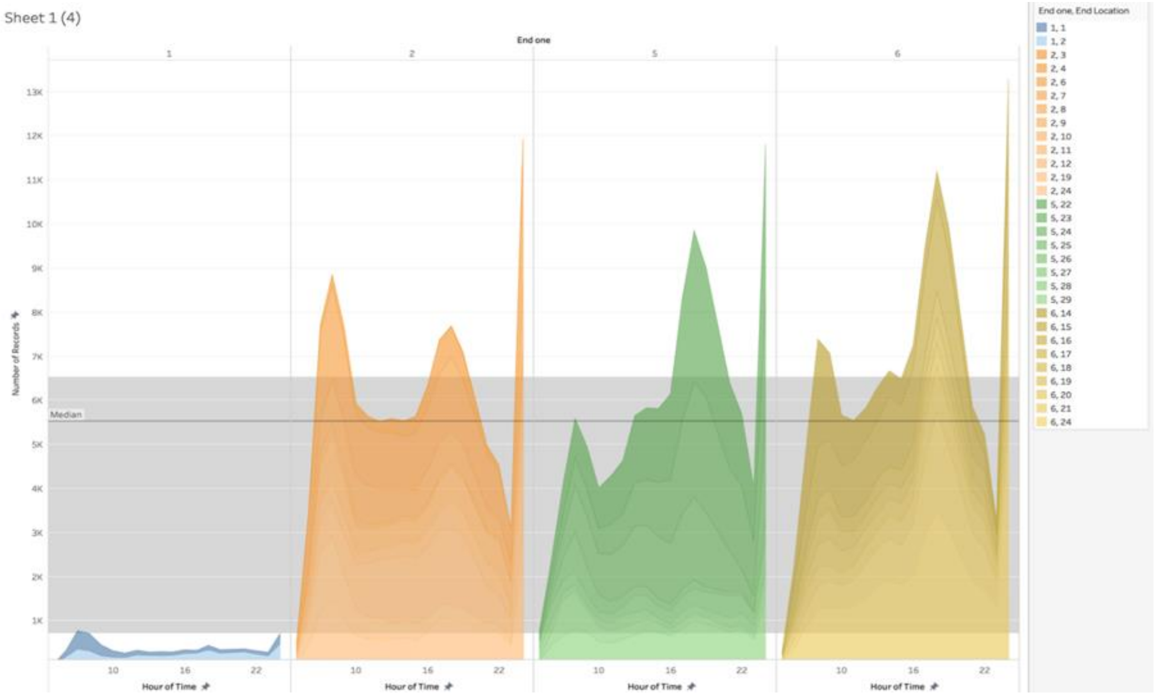

Figure 1: Traffic Pressures at each metro zone

Figure 2 below shows the time pressure in each metro station along the Dubai Metro Red Line. Time pressures are categorized into four colors that represent the four zones. For each station, the time pressure is then divided by the earlier identified highest number of passengers dropping off at the metro station, HC; resulting a value between 0 and 1. The higher the value, the higher the expected demand for the station is.

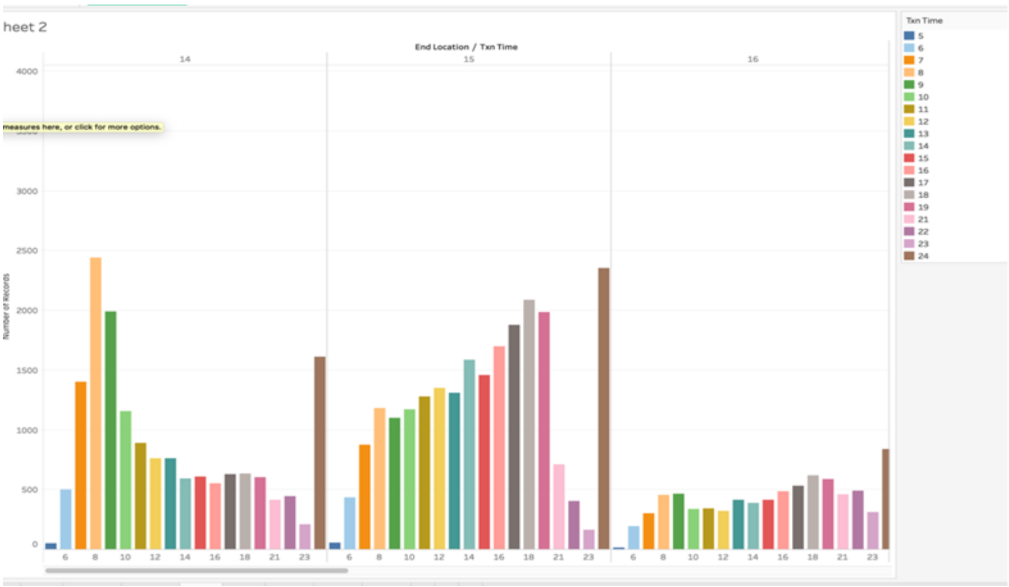

Figure 2: Traffic pressure at each metro station

Having been able to identify the traffic pressure and peak hours at each metro station, taxi drivers can now use these information to decide which metro station would be more optimal to visit and more beneficial in terms of expected demand. The process starts with determining the arrival time for each metro station (I), and arrival time (AT). The arrival time is expressed below:

Arrival Time = Current Time + Expected Time of Arrival

Using the recast function in data miner, the utility function is determined indicating the benefit of visiting each metro station. This utility function is divided into two parts: (1) benefit of visiting a particular station at the current time window, and (2) the expected demand for the next time window. The first part of the utility function involves multiplying the gap between the arrival time and the end of the current window by the normalized weight values for current time window. The utility function will yield a value ranging between 0 and 1. Here, the gap value controls the interest in the demand of the current window. If the gap is too small or close to zero, then it could indicate that passengers had already left the station. Gap, the normalized weight value for the current time window, and the overall first part utility function are expressed below:

Gap = maximum limit of current time window – arrival time at the station

Normalized weight = frequency per hour / maximum frequency per hour

Utility1 = Gap X *Normalized weight*

On the other hand, the second part of the utility function looks at the expected demand in the time window. For instance, if the driver near the station is expected to arrive in the metro close to the end of the current time window, then the next time window demand is taken into consideration. Here, this part of the utility function involves subtracting the time gap derived from the first part of the function from 1, which will yield a value between 0 and 1. Result will then be multiplied by the obtained value by the normalized weight of the current window. The overall utility function represents the sum of the two utility parts. The second part of the utility function is expressed below:

Utility2 = (1 – Gap) X *Normalized weight*

9. Conclusions

The purpose of this research was to explore the traffic pressures and peak hours at Dubai Metro Red Line, and further determine the taxi distribution for optimal benefits in station visit. The data mining techniques used in this research allowed the determination of the traffic/transaction pressures at each of the four zones. The data indicated that traffic/transaction pressure it at its highest in zone six of the Dubai Metro Red Line, which encompasses seven metro stations. More so, the data processing also enabled the determination of the peak hours and the traffic pressures at each metro station across all four zones. More so, having been able to identify the traffic pressure at each metro station, taxi drivers can make use of this information to decide which metro stations can provide more benefits based on their expected arrival time and expected demand at the station. The utility function was used for this purpose. For instance, it was indicated that there is a significant traffic pressure in the World Trade Center metro station between 9 am to 10 pm. Having communicated with this information, the taxi driver can decide whether or not it is best to visit the station. When the expected time to reach the station is at 9:50, then the gap between the end of current window and the expected arrival time is close to 0. It is, therefore, best for the driver to check the traffic pressure for the next time window (10-11am) to save time and increase their revenue.

References

[1] Ding, C., Wang, D., Ma, X. & Li, H., 2016. Predicting Short-Term Subway Ridership and Prioritizing Its Influential Factors Using Gradient Boosting Decision Trees. Sustainability, 8(1100), pp. 1-16.

[2] Dubai Pulse, 2019. About Dubai Pulse. [Online] Available at: https://www.dubaipulse.gov.ae/about [Accessed 2019 22 March].

[3] Ge, W. et al., 2017. Urban Taxi Ridership Analysis in the Emerging Metropolis: Case Study in Shanghai. Transportation Research Procedia, Volume 25, p. 4916–4927.

[4] Jiang, S., Guan, W., He, Z. & Yang, L., 2018. Exploring the Intermodal Relationship between Taxi and Subway in Beijing, China. Journal of Advanced Transportation, 208(3981845), pp. 1-14.

[5] Wang, F. & Ross, C., 2017. New potential for multimodal connection: exploring the relationship between taxi and transit in New York City (NYC). Transportation, Volume 2017, pp. 1-122.