Hindawi
Journal of Advanced Transportation
Volume 2022, Article ID 1108105, 14 pages
https://doi.org/10.1155/2022/1108105

Research Article

RFM Model and *K*-Means Clustering Analysis of Transit Traveller Profiles: A Case Study

Angela H. L. Chen[1] **Yun-Chia Liang**[2] **Wan-Ju Chang**[1] **Hsuan-Yuan Siauw**[1] **and Vanny Minanda**[2]

[1]*Department of Industrial and Systems Engineering, Chung Yuan Christian University, Taoyuan 320, Taiwan*
[2]*Department of Industrial Engineering and Management, Yuan Ze University, Taoyuan 320, Taiwan*

Correspondence should be addressed to Angela H. L. Chen; achen@cycu.edu.tw and Yun-Chia Liang; ycliang@saturn.yzu.edu.tw

Received 4 April 2022; Revised 6 July 2022; Accepted 7 July 2022; Published 8 August 2022

Academic Editor: Zhixiang Fang

Public transportation users increase as the population grows. In Taipei, Taiwan, this tendency is observed by analyzing historical data from the Mass Rapid Transit (MRT) and economy-shared bicycle (known as YouBike) riders. While this trend exists, the Taipei City government promotes green transportation by providing discounts to users who transfer from MRT or bus to YouBike within a particular period. Therefore, this study focuses on analyzing the patterns of users in order to identify possible clusters. Clusters of customers can be considered fundamental and competitive factors for the Ministry of Transportation to encourage the use of green transportation and promote a sustainable environment. Based on big data smart card information, this paper proposes using the RFM and *K*-means clustering algorithm to analyze and construct mode-switching traveller profiles on MRT and YouBike riders. As a result, three distinct clusters of MRT-YouBike riders have been identified: potential, vulnerable, and loyal. There are also suggestions regarding the most profitable groups, which customers to focus on, and to whom give special offers or promotions to foster loyalty among transit travellers.

1. Introduction

Public transport, defined as high-capacity vehicle sharing with fixed routes and schedules, will remain an essential engine to economic activities, social connections, and the standard of living. Due to traffic density and the demands on road infrastructure, land, material, energy, and workforce have been invested in providing transport services and developing its infrastructure. As transport demand continues growing, particularly in fast-developing nations, many cities expand their transportation networks and support infrastructure, indicating how vital the transport system is to economies and social welfare. A 2018 McKinsey report[1] concluded that wealthier cities have greater opportunities to build advanced transportation systems, but such prosperity does not guarantee the successful development of such systems. According to the 2020 “Foresight Research Survey,” as many as 81.1% of Taiwanese people have access to private transportation and only 44.4% rely on public transport for their daily commute. In addition, over 80% of those aged 18 or older rely on private transportation, primarily gasoline-powered motorcycles. Evidence has shown that people prefer to travel using their vehicles, which imposes considerable challenges to reducing private transportation dependence and encouraging the use of public transportation. However, since people naturally avoid transit they perceive as incompatible with their demands, such transformations are fraught with difficulty.

Taipei’s Mass Rapid Transit (MRT), the first subway system built in Taiwan, has already become a hallmark of Taipei City. Residents in Taipei welcomed its arrival and viewed it as an example of the city’s bright future. The Taipei Metro, once known as the Taipei Rapid Transit Corporation (TRTC), is a city government public transit operator. The MRT has made commuting more accessible for people in Taipei; however, its annual ridership from 770 million visits in 2018 drops to 690 million in 2020, partially due to the outbreak of COVID-19. In addition, the Taipei Metro launched many promotions to encourage people to take public transportation. For instance, a public transportation monthly pass, a trip discount on mode-switching within an hour, and the first 30 minutes of free YouBike rental. In order to retain existing travellers, discounts and incentives seem appropriate. However, the impact on encouraging new riders is questionable, especially when substantial decreases in operating costs resulting in significant profits come through various government subsidies. As a result, the Taipei Metro must allocate or invest resources in the desired services and travellers. The same action taken by travellers may have resulted in a value destroyer rather than a value creator for the system. Hence, it needs a strong foundation of customer-oriented strategic development.

In this regard, it is necessary to grasp the dynamics and heterogeneity of public transportation users. Li and Schmöcker [2] and Lin et al. [3] used questionnaires to have travellers indicate their reasons for public transport for descriptive analysis of behavioural changes. Tang et al. [4] collected data on users' socioeconomic characteristics, vehicle ownership, public bicycle use, and user satisfaction using online questionnaires. The Unified Theory of Acceptance and Use of Technology (UTAUT) was utilized by Jahanshahi et al. [5] to investigate travellers" opinions and identify factors that influence the adoption of bike-share systems. In the context of the movement of travellers and public transit stations, some studies focused on regional analysis [6–8], route analysis [9–11], site analysis [5, 10, 12, 13], ticketing channel [14–16], mode choice [17–20], and traveller characterization [21, 22]. In particular, Kim et al. [12] used ridership counts of selected intervals to classify the subway stations regarding their diurnal ridership patterns associated with land use. Gan et al. [7] analyzed the daily mobility patterns concerning land use from a station ridership perspective. As interest in people movement and urban mobility has surged in recent years, studies have attempted to predict individual travel-mode choices using traditional random utility models and machine learning approaches [2, 14, 22–32].

In segmentation, the K-means algorithm, an unsupervised learning cluster approach, is still very popular because of its speed, efficiency, and simplicity. The algorithm finds optimal groups (clusters) of customers, transactions, or other behaviours and things with high similarities and characteristics within the clusters. Many applications of cluster analysis have been applied in various industries, such as banking [33–35], energy supply [36], agriculture, food [37–39], health and insurance [40–43], telecom [44–48], postal service [49, 50], transportation [14, 17, 26, 29, 51–56], and retail [38, 57–62]. Furthermore, other researchers focus on customer relationship management (CRM) models and adopt them in K-means clustering. For RFM-based traveller segmentation, Reades et al. [51] used every 15-minute interval of boarding and alighting information, rather than the day of the week. Qian et al. [29] proposed customer segmentation rules of Electronic Toll Collection based on vehicle behavioural characteristics. In Chiang [54]; the concept of RFM was applied to discover valuable airline travellers, and the association rules led to identifying the optimal target markets. Also, based on the insights mentioned above, determining traveller values in each type of transportation study has its characteristics that are not fully met by the same model [56].

To our knowledge, there has been less research into different transfer and transit portfolio modes. Therefore, this study aims to profile the travel patterns of visitors to Taipei solely based on the MRT and the bike-share network ("YouBike"). From longitudinal observations of ridership patterns, travellers can be grouped into specific categories based on the RFM scoring model [63]. Furthermore, such a transit ridership portfolio reveals insights capturing travellers' travel behaviours, assisting public transit operators such as Taipei Metro in developing effective customer relationships and market strategies, and efficiently allocating resources.

2. Research Methodology

This section discusses how to combine the RFM model and clustering for constructing a transit ridership portfolio. Assuming that the transaction data are largely unlabelled, we begin considering a K-means clustering algorithm. We next intend to examine transfer behaviour by considering attributes that reflect travel spending and preferences, namely, RFM indicators. Further information is provided.

2.1. Data Description and Preprocessing. Original data were extracted from a smart card system (called the Easy Card) used by MRT and YouBike, with records spanning 31 months between January 2017 and July 2019. The dataset consists of 11 fields for the MRT service, such as card number, ticket type, entry/exit time, entry/exit code, transaction amount, transfer code, transfer discount, and commuter ticket. The dataset also contains 18 fields for the bike-sharing service, including card number/type, deduction time/amount, borrowing/return time, borrowing/return station code and slot, bicycle number, rental free, mobile phone, rate type, and others. In this study, only trips that included a YouBike-to-MRT transfer were included in the analysis. Data fields for the MRT and YouBike systems contained some inconsistencies, and certain details have been omitted for confidentiality reasons.

Table 1 lists the MRT and YouBike data fields and descriptions. Following the Metro Taipei website, all passengers purchase different passes on which the trip and fare amount are recorded. E-ticket types include standard, students, welfare (seniors and charities), and children. A ride begins and ends when travellers swipe cards to enter and leave the system, and its time is thus recorded and calculated as the travel time. A trip is considered when swiping a card from an original station to a destination station; nevertheless, a transfer made within a time window will be regarded as only a single trip. This study can identify the transfer behaviour between MRT-YouBike modes since both exact boarding and alighting stations are known. Finally, the fare

Table 1: The easy card transaction data format for the MRT and YouBike in the study.

Field	Data type	Field description	MRT	YouBike
Card no.	Character string	Anonymised smart card user identification	+	+
Card type	Nominal	Types of smart cards: EasyCards, iPASSes, iCASH, HappyCash.	+	+
Cardholder status	Nominal	Travel statuses: welfare (senior, disability, and charity), student, children, and standard.	+	+
Entry time	Numeric	Timestamp (date and time) for a traveller aboard a system	+	+
Exit time	Numeric	Timestamp (date and time) for a traveller exits a system	+	+
Entry code	Nominal	Boarding MRT station that a traveller enters the system	+	+
Exit code	Nominal	Alighting the MRT station that a traveller exits the system	+	+

for the trip is calculated and deducted after counting all discounts. A few travellers purchase All-Pass Tickets, which cover nearly all public transportation modes in Taipei and are eligible to rent YouBikes for free within the Taipei area for the first 30 minutes of each rental session.

2.2. Data Extraction. As soon as the data are preprocessed, the first step is to group travellers. In customer segmentation, three common indicators are recency (i.e., the most recent transfer), frequency (the number of transfers), and money (the total amount spent by the traveller). Figure 1 illustrates the notations for RFM data for a transit traveller. SD and ED are the start and end dates of the study period. TD_t denotes the date when a transfer occurs in the period t, and traveller transactions (x_t) are monitored till the end of period T.

The recency value can be determined by the number of days between the last trip date and the end of the analyzing period. The closer the last transaction date is to the end of the analysis period, the greater its value. Therefore, a traveller's recency (R-value) is determined as follows:

$$R = \mathrm{ED} - \max_{\mathrm{SD} \le t \le \mathrm{ED}} \mathrm{TD}_t. \quad (1)$$

Secondly, the number of transactions (x_t) made by a transit traveller at period t during the analysis period T shall be considered as the frequency (the F-value). (2) calculates the F-value as follows:

$$F = \sum_{\mathrm{SD}}^{\mathrm{ED}} x_t. \quad (2)$$

The frequency count is F in this case since travellers can transfer from one route to another by different modes to complete the same one-way trip. Therefore, the more frequent the traveller travels, the more valuable and loyal they are.

Finally, y_t denotes the monetary value a traveller has paid at the end of the trip for period t. Its total amount is directly related to the number of transactions in the public transportation system, set as M in (3). The higher the value, the more profit will be generated.

$$M = \sum_{\mathrm{SD}}^{\mathrm{ED}} y_t. \quad (3)$$

2.3. The K-Means Algorithm. A K-means clustering algorithm, proposed by [64], is used here to group travellers transferring between YouBikes and MRTs. Initially, the algorithm divides each object (or observation) into an arbitrarily determined number of clusters (k) based on the minimum distance between the object and its centroids. If a set of objects $(o_1, o_2, \ldots, o_n)$ contains a u-dimensional index (i.e., the RFM index), then the k value should be an optimal number effective for clustering. Accordingly, a high degree of similarity homogeneity (i.e., compactness) and a high degree of heterogeneity (i.e., separation) must be apparent between different groups must be observed (as illustrated in Figure 2). So, a common method for validating the appropriate size of clusters is the elbow method. The k value range, in this case, is set so that the results can be compared with the RFM analysis later on. We then perform the K-means clustering for each k value.

Clustering and the average distance are determined as follows:

(1) Select the number of k partitions in which the objects will be clustered

(2) Partition the object (O_i) into k subsets in a u-dimensional feature space

(3) Choose k random points from the partitioning sets as the initial cluster centroids (C_k)

(4) Calculate the distance between the data point (O_i) and the initial cluster centroids for each cluster (C_k) using Euclidian distance measure E as follows:

$$d(O_i, C_k) = \sqrt{\sum_{u=1}^{3} (o_{iu} - c_{ku})^2}. \quad (4)$$

(5) Assign objects to the group with the shortest distance

(6) Identify the new cluster centroid by recalculating the positions of all objects assigned to that cluster

(7) Repeat steps 3 and 6 until convergence or reach a fixed number of iterations, and confirm that the object has the shortest Euclidean distance from the cluster centroid as defined in the following equation:

$$d(O_i, C) = \min_{k \in \{1,2,..K\}} \{d(O_i, C_k)\}. \quad (5)$$

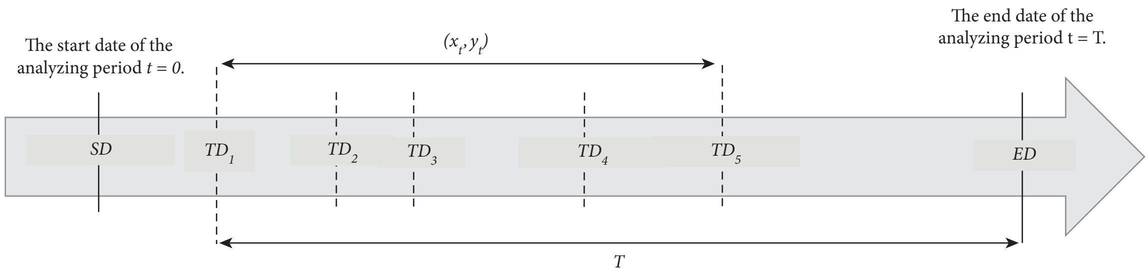

Figure 1: Notations for converting to RFM data in the study.

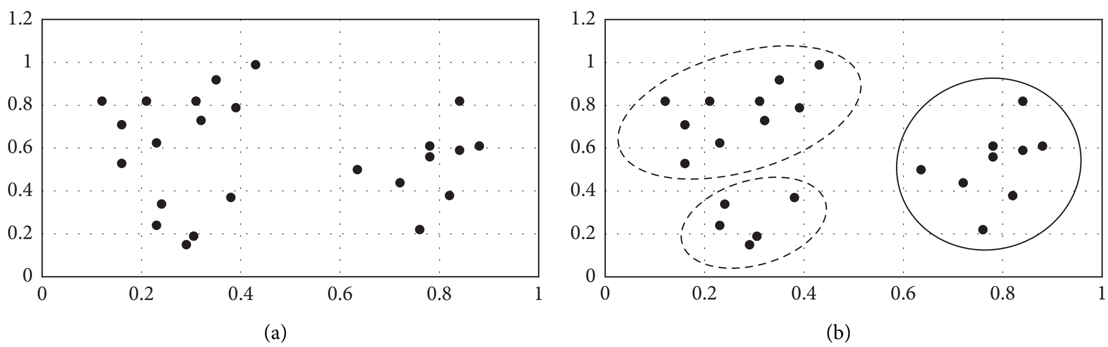

Figure 2: The basis of K-means algorithmic clustering.

(8) Calculate the average dissimilarity $\overline{D}$ of the cluster, where n is the total number of objects in the RFM index dataset, using the following formula:

$$\overline{D} = \frac{1}{n}\sum_{i=1}^{n} d(O_i, C). \tag{6}$$

3. Results and Discussion

This section first introduces the datasets used in the case study. Data are from the transit authority of Taipei Metro and the YouBike company in Taiwan. Millions of transactions have been transformed and preprocessed before being compiled into the dataset for the *K*-means clustering. The results of this implementation are then discussed.

3.1. Data Set Description. Figures 3 and 4 illustrate data for MRT and YouBike. There may be inconsistencies, irrelevant, and abnormal transaction information (e.g., entry and exit times). The data were cleaned using Python, EmEditor, and Spotfire. To display transfer patterns, MRT and YouBike data need to be combined. Consequently, we extracted the data by matching the cardholder ID and the transfer code from the MRT transaction data. Afterwards, we concatenated the matching data with the YouBike data. Following the Taipei Metro policy, there is a limit to the transfer duration, namely, one hour. Therefore, transfer behaviour is analyzed in this study based on the time spent riding a YouBike to the MRT at location A or vice versa. Thus, if a traveller returns a bicycle to a YouBike station and enters an MRT station within one hour, the system considers this a transfer behaviour. Figure 5 presents information regarding the transfer behaviours between YouBike and MRT.

Table 2 summarizes the final data set, consisting of 5,023,808 records of transfer transactions from January 2018 through July 2019. By transforming the EasyCard usage of individual travellers into time-dependent transit frequency in Table 3, the average daily transfers increased from Sunday to the middle of the week before plateauing (or slightly decreasing) till Saturday. Figure 6 indicates that most users are either standard or student cardholders. The number of transactions during the weekdays remains higher than during the weekend, despite various cardholder types. This suggests that most transfers are likely to be made by daily commuters.

The data were then transformed into RFM features for each traveller. Some illustrative examples are shown in Table 4. In this study, all the model features are given equal weight (i.e., are equally important), and the results of the RFM data conversion are used to perform a nonhierarchical cluster analysis. The RFM values differ due to scale differences, which in turn affects the clustering analysis. Accordingly, we standardized the RFM values using the simple z-score method.

3.2. Cluster Analysis. Several methods in the literature are used to determine the number of optimal clusters. However, according to Horvat et al. [65] and Raza et al. [66], clustering algorithms can be discriminatory, making it difficult to evaluate the results objectively. In addition, the categories that emerge from this process can take on different meanings depending on their context. For our study, since no ground-truth label of data exists for our problem, we validate the

26F8438B11AD	255	1	2019/7/8	1900/1/0 08:24	2019/7/8	1900/1/0 08:40	111	11	00:16:17	977	Normal	Different	20	1501	0	74	2019/6/26	1900/1/0
45ED0062CD43	255	1	2019/7/17	1900/1/0 07:54	2019/7/17	1900/1/0 08:48	122	30	00:54:35	3275	Normal	Different	40	1501	0	74	2019/6/20	1900/1/0
6B233806DE3B69	255	1	2019/7/8	1900/1/0 08:22	2019/7/8	1900/1/0 08:40	31	24	00:17:21	1041	Normal	Different	20	1501	0	74	2019/6/10	1900/1/0
4D7D23E03652E	255	1	2019/7/8	1900/1/0 08:20	2019/7/8	1900/1/0 08:40	132	100	00:19:20	1160	Normal	Different	20	1501	0	74	2019/6/16	1900/1/0
B42CF2BBF2EA	255	1	2019/7/8	1900/1/0 08:30	2019/7/8	1900/1/0 08:40	45	134	00:09:21	561	Normal	Different	16	1501	0	74	2019/6/20	1900/1/0
451681DF39C00A	255	1	2019/7/8	1900/1/0 07:52	2019/7/8	1900/1/0 08:40	174	82	00:48:07	2887	Normal	Different	32	1501	0	74	2019/6/12	1900/1/0
D55CA929D82	255	1	2019/7/17	1900/1/0 08:18	2019/7/17	1900/1/0 08:48	103	80	00:29:57	1797	Normal	Different	28	1501	0	74	2019/7/10	1900/1/0
802A6E3379FE0011	255	1	2019/7/8	1900/1/0 08:19	2019/7/8	1900/1/0 08:40	70	61	00:21:07	1267	Normal	Different	24	1501	0	74	2019/7/5	1900/1/0
43B33709859FFB	255	1	2019/7/8	1900/1/0 08:19	2019/7/8	1900/1/0 08:40	80	88	00:21:16	1276	Normal	Different	24	1501	0	74	2019/7/3	1900/1/0
3DE75699873B	255	1	2019/7/17	1900/1/0 08:14	2019/7/17	1900/1/0 08:48	59	99	00:34:42	2082	Normal	Different	23	1501	5	74	2019/6/20	1900/1/0
967F1F6093043C	255	1	2019/7/8	1900/1/0 08:15	2019/7/8	1900/1/0 08:40	48	100	00:25:17	1517	Normal	Different	24	1501	0	74	2019/6/26	1900/1/0
37F20587AC7D	255	1	2019/7/8	1900/1/0 08:07	2019/7/8	1900/1/0 08:40	79	91	00:32:46	1966	Normal	Different	32	1501	0	74	2019/6/12	1900/1/0
286EAEFD972B	255	1	2019/7/8	1900/1/0 08:01	2019/7/8	1900/1/0 08:40	176	13	00:38:40	2320	Normal	Different	28	1501	0	74	2019/6/11	1900/1/0
3267EC6E941C	255	1	2019/7/8	1900/1/0 07:56	2019/7/8	1900/1/0 08:40	175	35	00:43:26	2606	Normal	Different	32	1501	0	74	2019/6/11	1900/1/0
DC2F9EBAA2C3	255	1	2019/7/17	1900/1/0 08:14	2019/7/17	1900/1/0 08:48	180	89	00:34:43	2083	Normal	Different	32	1501	0	74	2019/7/16	1900/1/0
CE510134A1DEE	255	1	2019/7/8	1900/1/0 08:14	2019/7/8	1900/1/0 08:40	14	23	00:25:42	1542	Normal	Different	28	1501	0	74	2019/7/8	1900/1/0
F4EC735C6A887	255	1	2019/7/26	1900/1/0 08:12	2019/7/26	1900/1/0 08:41	83	111	00:28:31	1711	Normal	Different	28	1501	0	74	2019/7/25	1900/1/0
D108953D0C6C1	255	1	2019/7/8	1900/1/0 08:11	2019/7/8	1900/1/0 08:40	77	88	00:28:31	1711	Normal	Different	32	1501	0	74	2019/6/10	1900/1/0
075184728630A	255	1	2019/7/17	1900/1/0 08:09	2019/7/17	1900/1/0 08:48	82	63	00:38:56	2336	Normal	Different	36	1501	0	74	2019/7/17	1900/1/0
B799FC2A940877	255	1	2019/7/17	1900/1/0 08:19	2019/7/17	1900/1/0 08:48	82	37	00:29:28	1768	Normal	Different	28	1501	0	74	2019/7/11	1900/1/0
051C74A252DB	255	1	2019/7/7	1900/1/0 11:03	2019/7/7	1900/1/0 11:45	124	100	00:42:02	2522	Normal	Different	28	1501	0	74	2019/6/22	1900/1/0

Figure 3: An example of raw data for MRT passengers.

313	F7DD9530D7BAB	2019/7/5	1900/1/0 20:09	0	2019/6/21	1900/1/0 21:40	47	MRT Zhongxiao Xinsheng Station (Exit 4)	017A	00050011F8	2019/6/22	1900/1/0 20:48	363	Entrance of Lane 340, Section 1 of Fuxing South Road	005A	23:07:39	83259	Hight	Different	63941F36	1	L01	97	ECC
421	5A4298D0E44C02	2019/7/17	1900/1/0 19:58	0	2019/7/15	1900/1/0 19:59	394	Songde Hulin Street	013A	3216	2019/7/16	1900/1/0 19:28	394	Songde Hulin Street	013A	23:29:03	84543	Hight	same	1AD5203A	1	L01	95	ECC
103	DF1832809CC4D2	2019/7/22	1900/1/0 12:29	1515	2019/7/4	1900/1/0 08:26	105	Emei parking lot	002A	0005000A9A	2019/7/5	1900/1/0 08:19	105	Emei parking lot	004A	23:53:51	86031	Hight	same	3.66E+43	1	L01	96	ECC
1539	976734461E238784	2019/7/5	1900/1/0 12:12	0	2019/7/2	1900/1/0 20:42	83	MRT Daan Forest Park Station	019B	5002196	2019/7/3	1900/1/0 20:23	20	MRT Technology Building Station	021B	23:41:01	85261	Hight	Different	31E7FD38	1	L01	95	ECC
413	1F46C92BD87D	2019/7/3	1900/1/0 21:36	0	2019/6/28	1900/1/0 15:12	45	MRT Gongguan Station (Exit 2)	009A	0000003E0E	2019/6/29	1900/1/0 14:50	30	Keelung Changxing Road	026B	23:37:28	85048	Hight	Different	62F26336	1	L01	96	ECC
1632	BE5C3D5BC9B2F1B	2019/7/17	1900/1/0 13:47	0	2019/7/13	1900/1/0 15:24	70	MRT Taipei 101/World Trade Center	019A	0007001A9C	2019/7/14	1900/1/0 14:21	113	Renai Yixian Road	018A	22:57:25	82645	Hight	Different	06CB7D3A	1	L01	96	ECC
1597	643FE9A07FC53D	2019/7/1	1900/1/0 19:05	0	2019/6/26	1900/1/0 20:02	85	Xinyi Dunhua Road	021A	00050023C8	2019/6/27	1900/1/0 18:59	85	Xinyi Dunhua Road	021A	22:57:42	82662	Hight	same	050ADA3A	1	L01	94	ECC
1302	142D57BE75C4	2019/7/12	1900/1/0 16:10	1355	2019/7/11	1900/1/0 18:22	340	Nanjing Guangfu Road	003A	0005002BB9	2019/7/12	1900/1/0 16:06	320	MRT Nanjing Sanmin Station (Exit 3)	001A	21:44:20	78260	Hight	Different	450BF136	1	L01	97	ECC
551	8D4963568ACB	2019/7/16	1900/1/0 22:16	1315	2019/7/16	1900/1/0 00:16	57	Xinsheng Heping Road	012B	5004395	2019/7/16	1900/1/0 21:24	105	Emei parking lot	014A	21:08:40	76120	Hight	Different	AD3C0438	1	L01	96	ECC
194	9DE09303EAA96	2019/7/1	1900/1/0 12:50	0	2019/6/25	1900/1/0 10:35	99	MRT Shuanglian Station (Exit 2)	014A	00070016D8	2019/6/26	1900/1/0 08:06	99	MRT Shuanglian Station (Exit 2)	006A	21:31:16	77476	Hight	same	C737EF3A	1	G03	96	ECC
402	C6872F3FD14EE8C	2019/7/12	1900/1/0 21:36	1315	2019/7/9	1900/1/0 21:33	73	Jiuquan Yanping Road	015A	5003403	2019/7/10	1900/1/0 18:53	78	MRT Yuanshan Station (Exit 2)	026A	21:19:46	76786	Hight	Different	7F3AD53A	1	L01	97	ECC
1215	9F09650EBE67BA	2019/7/22	1900/1/0 15:11	1355	2019/2/27	1900/1/0 16:53	255	Entrance of Lane 22, Guangfu South Road	002A	5000148	2019/2/28	1900/1/0 14:44	49	Longmen Square	018A	21:50:19	78619	Hight	Different	A767C938	1	L01	95	ECC
31	D8AD85545C4F1	2019/7/5	1900/1/0 21:38	0	2019/6/2	1900/1/0 19:38	15	Raohe Night Market	007B	00070013F9	2019/6/3	1900/1/0 16:01	255	Entrance of Lane 22, Guangfu South Road	016B	20:23:33	73413	Hight	Different	CF94D43A	1	L01	95	ECC
170	F9975383E68CCA	2019/7/12	1900/1/0 16:52	1155	2019/7/8	1900/1/0 15:01	345	Zhongzheng Sports Center	013A	5002681	2019/7/9	1900/1/0 10:05	195	Xinyi Hangzhou Road (Chunghwa Telecom Corporation)	025A	19:03:56	68636	Hight	Different	9B529C39	1	L01	96	ECC
739	9805D1AD2195AD	2019/7/3	1900/1/0 22:22	0	2019/6/27	1900/1/0 20:01	173	MRT Songjiang Nanjing Station (Exit 7)	027B	5001418	2019/6/28	1900/1/0 15:35	121	Longjiang Nanjing Road	001A	19:33:38	70418	Hight	Different	67794236	1	L01	96	ECC
1411	25AF4D34AE4D4	2019/7/3	1900/1/0 08:36	0	2019/4/20	1900/1/0 19:10	126	Dunhua Keelung Road	011B	0007000FB9	2019/4/21	1900/1/0 14:12	31	Xinhai Xinsheng Road	015B	19:01:47	68507	Hight	Different	65A3D43A	1	L01	96	ECC
955	675C74486CC462B	2019/7/22	1900/1/0 21:46	0	2019/3/23	1900/1/0 15:56	42	MRT Houshanpi Station (Exit 1)	018B	0005001DA1	2019/3/24	1900/1/0 11:25	42	MRT Houshanpi Station (Exit 1)	018B	19:27:53	70130	Hight	same	F219E136	1	L01	90	ECC
768	EE55DB89CA52C3	2019/7/17	1900/1/0 11:33	0	2019/7/11	1900/1/0 16:56	237	Wenshan Second Administrative Center	013A	0005000C4C	2019/7/12	1900/1/0 11:47	206	MRT Xinhai Station	002B	18:50:59	67859	Hight	Different	9006FB39	1	L01	96	ECC
317	16E02423C52AD2E	2019/7/30	1900/1/0 15:44	1075	2019/7/25	1900/1/0 23:34	52	Jianguo Changchun Road	001A	5002453	2019/7/26	1900/1/0 17:39	45	MRT Gongguan Station (Exit 2)	007A	18:04:55	65095	Hight	Different	3DD39F37	1	L01	96	ECC

Figure 4: An example of raw data for YouBike riders.

id	ubtime_out	mrttime_in	difference	ubstation_out_no	mrtstation_in_no	ubtime_in	ubstation_in_no	mrttime_out	mrtstation_out_no	cardtype	mrtprice	ubprice	discount
08BF88351	2019/1/5 21:00	2019/1/5 21:05	00:05:13	281	59	2019/1/5 20:53	133	2019/1/5 21:17	55	1	11	5	5
344888C935	2019/1/5 21:00	2019/1/5 21:04	00:04:16	281	56	2019/1/5 20:55	349	2019/1/5 21:15	52	1	11	5	5
C9E3616125	2019/1/5 21:00	2019/1/5 21:01	00:01:28	281	58	2019/1/5 20:52	280	2019/1/5 21:14	62	129	0	0	5
95B94A9D3	2019/1/5 21:00	2019/1/5 21:02	00:01:27	281	57	2019/1/5 20:45	190	2019/1/5 21:23	64	1	20	0	0
3D8B7FE517	2019/1/5 21:00	2019/1/5 21:01	00:00:27	281	59	2019/1/5 20:50	232	2019/1/5 21:30	132	1	15	5	5
D02BFE8650	2019/1/5 21:00	2019/1/5 21:01	00:00:27	281	59	2019/1/5 20:55	114	2019/1/5 21:22	68	1	19	5	5
79C265D1DF	2019/1/5 21:00	2019/1/5 21:01	00:01:01	281	96	2019/1/5 20:58	87	2019/1/5 21:22	11	1	15	0	5
22E5282F38	2019/1/5 21:01	2019/1/5 21:02	00:01:47	281	100	2019/1/5 20:56	12	2019/1/5 21:26	51	1	15	5	5
5533833F205	2019/1/5 21:01	2019/1/5 21:02	00:01:30	281	86	2019/1/5 20:00	82	2019/1/5 21:28	79	1	19	15	5
277B88E59	2019/1/5 21:01	2019/1/5 21:04	00:02:38	281	64	2019/1/5 20:53	137	2019/1/5 21:44	100	6	32	5	0

Figure 5: An example of merged data for YouBike-MRT transfers.

Table 2: Number of rides and transfers after preprocessing.

Year	MRT rides	YouBike rides	MRT-YouBike transfers
2018	66,38,98,743	2,27,43,710	27,58,014
2019*	39,34,28,370	1,43,29,967	22,65,794

*Available data: 01 January to 31 July only.

Table 3: Daily transfer distribution from January 2018 to July 2019 in this study.

Day of a week	2018	%	2019	%	Total	%
Sunday	3,23,293	11.7	2,55,183	11.3	5,78,476	11.5
Monday	3,38,425	12.3	2,62,121	11.6	6,00,546	12.0
Tuesday	4,17,228	15.1	3,52,129	15.5	7,69,357	15.3
Wednesday	4,21,460	15.3	3,61,059	15.9	7,82,519	15.6
Thursday	4,25,731	15.4	3,40,225	15.0	7,65,956	15.2
Friday	4,15,675	15.1	3,46,320	15.3	7,61,995	15.2
Saturday	4,16,202	15.1	3,48,757	15.4	7,64,959	15.2

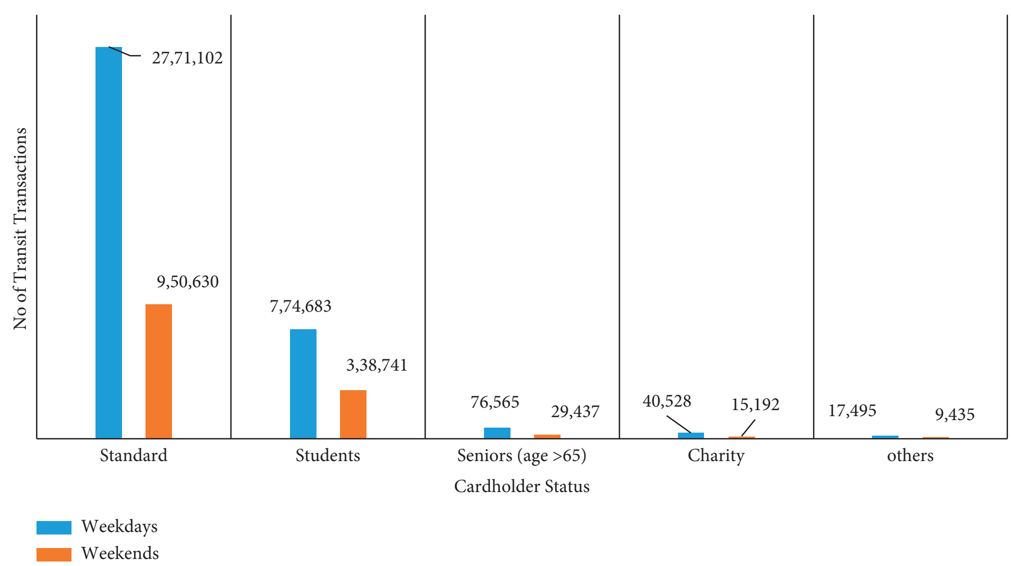

Figure 6: Transfer distribution of easy cardholders for weekdays and weekends from January 2018 to July 2019.

Table 4: Examples of values of the RFM model via data transformation in the study.

Easy card ID	Recency	Frequency	Monetary
$ED0F********80B6D$	1	14	12,467
E6AF ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ACDE7	3	10	7,392
F551 ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ BB99B	2	25	7,047
C14E ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ 17C08	1	4	5,838
DEAA ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ C5AC5	109	8	5,803
AF13 ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ 868F5	4	3	5,744
B61D ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ 386B9	4	3	5,581
EB94 ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ BB99B	2	13	5,477
F2B2 ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ 83BAD	53	20	5,179
DA7A ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ D4A30	3	7	5,168

number of clusters using the elbow method, a type of internal clustering validation. In the elbow method, K-means clustering is performed on the dataset for a range of values of k, and the sum of the square of the distance between each point and its closest centroid, also known as the inertia, is calculated. We represent the inertia as the mean distortion in our graph, while others may represent it as the sum of squared errors (SSEs), the Within Cluster Sum of Squares (WCSSs), etc.

We start with $k = 2$ and increment by 1 until $k = 10$. Upon reaching a certain value of k, the cost of training (i.e., the diminishing return) will drop dramatically and eventually reach a plateau as the k value increases further (see Table 5). The diminishing return is greatest when $k = 1$. From $k = 4$ onward, the change rate becomes indistinguishable; the movement is almost parallel to the X-axis. Figures 7 and 8 show that both distortions decline rapidly as k increases from 1 to 3, their diminishing returns hit at $k = 3$ and slow down after $k = 4$. Therefore, $k = 3$ is the optimal number of clusters.

Tables 6 and 7 provide details regarding three clusters in 2018 and 2019. In 2018, there were 526,697 travellers, while in 2019, there were 426,717. It is important to note that 2019 begins in January, i.e., for seven months, while 2018 is for the entire year. Based on its proportionality, we can conclude that the transit ridership is increasing, providing a promising outlook for the use of public transportation. In the three clusters in 2018, cluster 1 has 116,154 travellers with average recency of 27.73 days, a frequency of 2.43, and an average monetary value of NTD50.16. Cluster 2 travellers have average recency of 165.17 days, a frequency of 1.62, and an average monetary value of NTD37.29; cluster 3 travellers have average recency of 54.93 days, a frequency of 23.66, and an average monetary value of NTD523.33. The results for 2019 are similar to those of 2018, with RFM values slightly lower than in 2018.

The average RFM values of each cluster are compared with the overall average RFM values to determine the RFM score tendency. An upward symbol (↑) is placed if the average R, F, and M value is greater than the total average; otherwise, a downward symbol (↓) is used. Table 8 summarizes transit traveller profiles for 2018 and 2019. For instance, a traveller in cluster 2 ($R\uparrow, F\downarrow, M\downarrow$), with recency values of 165.17 and 122.85, higher than its overall averages of 82.61 and 59.18, respectively, is likely to be a vulnerable customer who has not used transit for a long time. The percentage of travellers in this group decreased slightly from 62.35% in 2018 to 60.59% in 2019. In addition, a traveller in cluster 1 ($R\downarrow, F\downarrow, M\downarrow$) may be a potential customer who has just begun travelling by transit. This group represents 22.05% of all consumers in 2018 and 24.66% in 2019. Moreover, travellers in cluster 3 ($R\downarrow, F\uparrow, M\uparrow$) tend to be loyal (regular commuters), incurring significant travel expenses and frequent use of transit. The number of travellers was 15.60% in 2018 and 14.75% in 2019.

Cluster 2 is a vulnerable segment of the travel market since customers are likely not using public transit. By contrast, cluster 1 customers tend to be transit-savvy. Lastly, cluster 3 customers tend to be loyal travellers with monthly or weekly passes. For this reason, we further break down the data into quarters to understand the transit behaviour, as shown in Tables 9–14. Recency gaps between clusters 1 and 2 have narrowed from 12 to 2 days on average quarterly. Likewise, the difference between these two clusters has been relatively large in most quarters, except for the first quarter of 2018, when Taipei's city government launched a promotional scheme for two-way transfers among YouBike, the MRT, and the bus. Table 15 indicates the market sizes for all clusters have remained steady across all quarters. Cluster 3 is the most profitable among these three clusters, with the highest frequency and monetary value.

Table 5: The cost of training as the number of clusters increases.

Number of clusters (k)	The diminishing return	
	2018	2019
1	29,000	40,000
2	7,500	9,500
3	3,300	4,000
4	2,500	2,500
5	2,000	1,800
6	1,500	1,300
7	1,000	1,000
8	750	750
9	650	650
10	500	500

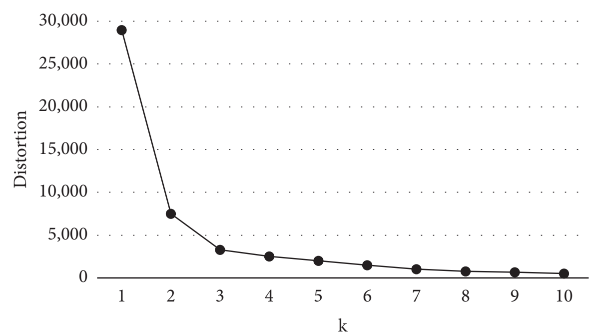

Figure 7: Distortion based on different (k) values for the 2018 dataset.

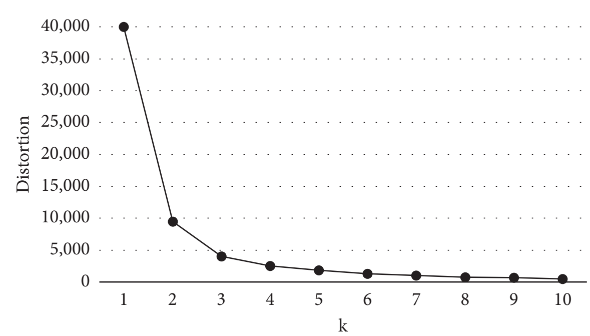

Figure 8: Distortion based on different (k) values for the 2019 dataset.

As shown in Figure 9, we first exclude the frequency data from our analysis to gain a deeper understanding of this target segment. Consequently, the points illustrating the recency coordinates mostly fall between 10 and 20 days. Afterwards, target segments are analyzed based on the number of transit users and their monetary value. Based on Figure 10, the points illustrating the Frequency-Monetary values are clearly divided into two distinct groups: one group with fewer than three transits and monetary values less than 50, and a second group with more than 15 transits and monetary values at least NTD300. Due to this fact, even though one category is more profitable, the two categories can still be promoted differently to maximize profits. In Figure 11, the target segments are visually displayed as three

Table 6: Clustering results by K-means for 2018.

Cluster (K)	$\overline{R}_k$ (days)	$\overline{F}_k$ (counts)	$\overline{M}_k$ (NTD)	No. of travellers (people)
(1) Potential	27.73	2.43	50.16	1,16,154
(2) Vulnerable	165.17	1.62	37.29	3,28,375
(3) Loyal	54.93	23.66	523.33	82,168
Average	82.61	9.24	203.60	Total 526,697

Table 7: Clustering results by K-means for 2019.

Cluster (K)	$\overline{R}_k$ (days)	$\overline{F}_k$ (counts)	$\overline{M}_k$ (NTD)	No. of travellers (people)
(1) Potential	21.89	2.80	56.22	1,05,240
(2) Vulnerable	122.85	1.56	33.34	2,58,535
(3) Loyal	32.81	24.93	539.59	62,942
Average	59.18	9.76	209.72	Total 426,717

Table 8: Profiles of transit travellers in 2018 and 2019.

Cluster (K)	RFM scores	2018 (%)	2019 (%)
(1) Potential	$\downarrow R$ $\downarrow F$ $\downarrow M$	22.05	24.66
(2) Vulnerable	$\uparrow R$ $\downarrow F$ $\downarrow M$	62.35	60.59
(3) Loyal	$\downarrow R$ $\uparrow F$ $\uparrow M$	15.60	14.75

Table 9: Clustering results by K-means for 2018 $Q1$.

Cluster (K)	$\overline{R}_k$ (days)	$\overline{F}_k$ (counts)	$\overline{M}_k$ (NTD)	No. of travellers (people)
(1) Potential	10.09	1.01	30.04	759
(2) Vulnerable	51.52	1.01	28.71	2,018
(3) Loyal	22.02	3.05	82.36	321
Average	27.88	1.69	47.04	Total 3,098

Table 10: Clustering results by K-means for 2018 $Q2$.

Cluster (K)	$\overline{R}_k$ (days)	$\overline{F}_k$ (counts)	$\overline{M}_k$ (NTD)	No. of travellers (people)
(1) Potential	9.80	2.06	44.20	57,689
(2) Vulnerable	53.90	1.40	32.74	1,51,145
(3) Loyal	13.21	14.89	333.83	35,671
Average	25.63	6.12	136.92	Total 244,505

Table 11: Clustering results by K-means for 2018 $Q3$.

Cluster (K)	$\overline{R}_k$ (days)	$\overline{F}_k$ (counts)	$\overline{M}_k$ (NTD)	No. of travellers (people)
(1) Potential	9.80	2.06	44.20	57,689
(2) Vulnerable	53.90	1.40	32.74	1,51,145
(3) Loyal	13.21	14.89	333.83	35,671
Average	25.63	6.12	136.92	Total 244,505

clusters, with the upper left cluster representing the loyal customer ($K = 3$). These findings allow Public-transit operators such as Taipei Metro and YouBike companies to conduct microtargeted campaigns offering incentives to each segment and promoting transit options and special fare subsidies. In addition, public transportation operators may

Table 12: Clustering results by K-means for 2018 $Q4$.

Cluster (K)	$\overline{R}_k$ (days)	$\overline{F}_k$ (counts)	$\overline{M}_k$ (NTD)	No. of travellers (people)
(1) Potential	11.74	2.17	44.09	58,867
(2) Vulnerable	54.96	1.45	31.47	1,37,940
(3) Loyal	13.87	17.27	379.12	36,025
Average	26.86	6.97	151.56	Total 232,832

Table 13: Clustering results by K-means for 2019 $Q1$.

Cluster (K)	$\overline{R}_k$ (days)	$\overline{F}_k$ (counts)	$\overline{M}_k$ (NTD)	No. of travellers (people)
(1) Potential	9.30	2.13	42.85	58,903
(2) Vulnerable	54.22	1.45	31.24	1,43,236
(3) Loyal	11.55	16.18	352.38	35,685
Average	25.02	6.59	142.16	Total 237,824

Table 14: Clustering results by K-means for 2019 $Q2$.

Cluster (K)	$\overline{R}_k$ (days)	$\overline{F}_k$ (counts)	$\overline{M}_k$ (NTD)	No. of travellers (people)
(1) Potential	9.39	2.25	44.53	56,586
(2) Vulnerable	53.17	1.49	31.69	1,46,512
(3) Loyal	13.00	16.82	365.63	38,395
Average	25.19	6.85	147.28	Total 241,493

Table 15: Seasonal profiles of transit travellers in 2018 and 2019.

Cluster (K)	RFM scores	2018				2019	
		Q1	Q2	Q3	Q4	Q1	Q2
(1) Potential	$R\downarrow, F\downarrow, M\downarrow$	24.50	23.59	23.59	25.28	24.77	23.43
(2) Vulnerable	$R\uparrow, F\downarrow, M\downarrow$	65.14	61.82	60.84	59.24	60.23	60.67
(3) Loyal	$R\downarrow, F\uparrow, M\downarrow$	10.36	14.59	15.58	15.47	15.00	15.90

Note. The seasonal profiles are shown as a percentage.

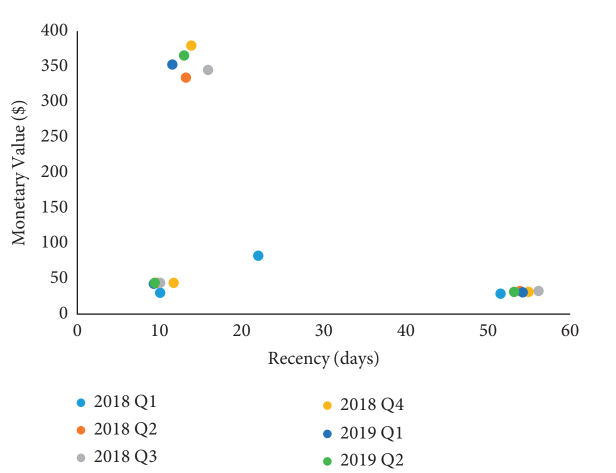

Figure 9: Target segment identification based on transit travellers' recency (R) and monetary value (M). Points represent clusters 1, 2, and 3 for the 2018 and 2019 quarterly.

approach large employers in some areas to encourage employees to use public transportation. Together, public-transit operators and employers can facilitate, and employer-sponsored passes can be beneficial.

Tables 16 and 17 summarize the distribution of each group (potential, vulnerable, loyal) over the week. The highest percentages of transfers are different in 2018 and 2019, from mainly the vulnerable segment shifting to the potential segments. Loyal customers are mainly found from Wednesday to Friday, which indicates that most regular passengers are students and employees. Conversely, the transfer of vulnerable passengers varies from 2018 to 2019. The vulnerable group's participation rate has increased on Saturday and Sunday. Therefore, it can be concluded that the number of casual users has increased over the weekend.

Table 18 presents the annual MRT revenue contributions among the three segments. The "11~20" represents the amount the passenger pays between NTD11 and NTD20 for this MRT ride. The total percentages of the three segments illustrated different contributions in 2018 and 2019, with 51.43% in 2018 and 78.81% in 2019. It is pertinent to note that the base charge in Taipei MRT is NTD20 once the passenger leaves the boarding station by MRT. We also found that revenue contributions for all segments were generated through the expense of NTD11~20 per trip, lower than the base charge. In fact, by adding up all three segments, travellers spending NTD11~20 per ride accounted for approximately 58.94% of the 2018 revenue and 61.10.% of 2019,

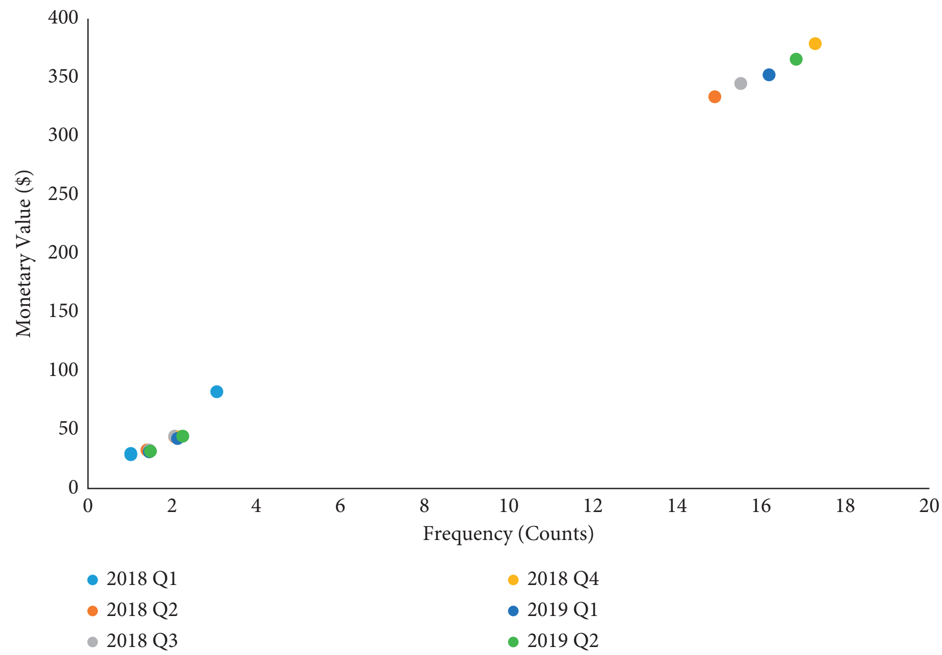

Figure 10: Target segment identification based on transit travellers' relative frequency (F) and monetary value (M). Points represent clusters 1, 2, and 3 for the 2018 and 2019 quarterly.

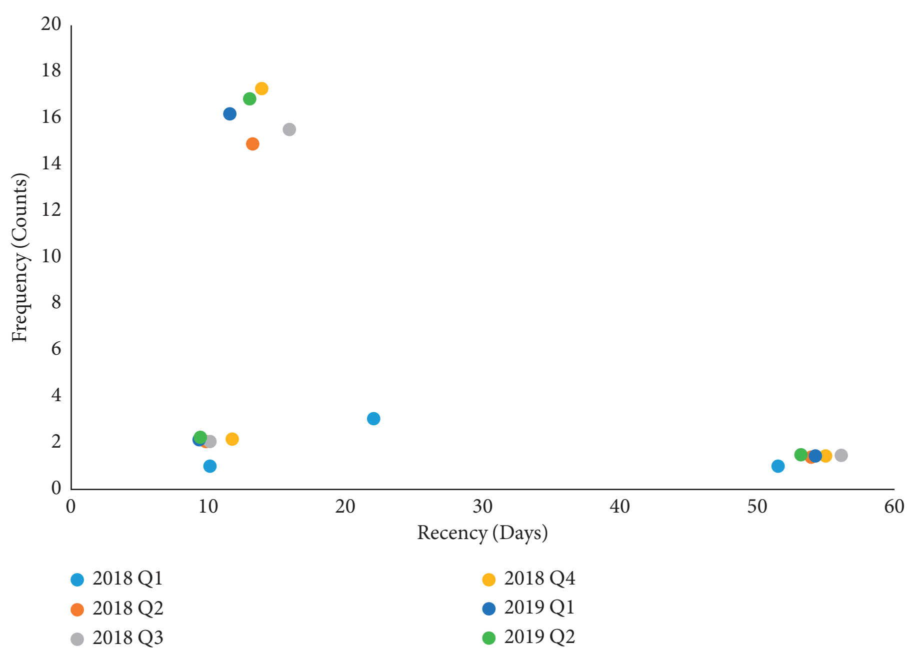

Figure 11: Target segment identification based on transit travellers' relative recency (R) and frequency value (F). Points represent clusters 1, 2, and 3 for the 2018 and 2019 quarterly.

and the traveller segment changed from the vulnerable to the potential in 2019.

In Table 19, the highest YouBike annual revenue percentages in 2018 and 2019 were 50.34% and 45.73%, respectively. In 2018, 68.85% of total revenue came from passengers who spent no more than NTD10 on an excursion, which increased to 76.25 percent in 2019. As such, more travellers are using YouBike as their first and last-mile mode of transportation since it charges NTD10 per 30 minutes, and most customers return their bikes within a half hour. Therefore, while it was encouraging to see more recent travellers using the MRT and YouBike, there seems to be a need to encourage more frequent transfers since a fare increase seems implausible.

Table 16: % daily transfer distribution among target segments ($K = 3$) in 2018.

Day	1-potential	2-vulnerable	3-loyal	Total transfers
Sunday	1.65	5.37	4.7	11.72
Monday	1.39	6.95	3.93	12.27
Tuesday	1.38	7.98	5.77	15.13
Wednesday	1.36	7.85	6.07	15.28
Thursday	1.39	8.02	6.03	15.44
Friday	1.41	8.01	5.65	15.07
Saturday	1.78	7.04	6.27	15.09
Total transfers	10.36	51.22	38.42	100

Table 17: % daily transfer distribution among target segments ($K = 3$) in 2019.

Day	1-potential	2-vulnerable	3-loyal	Total
Sunday	4.73	1.48	5.05	11.26
Monday	5.04	1.18	5.35	11.57
Tuesday	6.89	1.32	7.33	15.54
Wednesday	7.36	1.39	7.19	15.94
Thursday	6.79	1.34	6.89	15.02
Friday	6.85	1.43	7	15.28
Saturday	7.84	1.72	5.83	15.39
Total transfers	45.51	9.85	44.64	100

Table 18: % annual MRT revenue generated among different customer groups in 2018 and 2019.

Fare charge (NTD)	2018 contribution			2019 contribution		
	1-potential	2-vulnerable	3-loyal	1-potential	2-vulnerable	3-loyal
0~10	0.06	0.35	0.21	0.53	0.11	0
11~20	6.48	30.26	22.2	45.16	10.45	5.49
21~30	1.94	14.07	11.05	21.98	3.2	0.52
31~40	0.8	6.71	5.78	11.12	1.35	0.05
41~50	0.02	0.04	0.03	0.02	0.01	0
Total	9.3	51.43	39.27	78.81	15.12	6.06

Table 19: % annual YouBike revenue generated among different customer groups in 2018 and 2019.

Fare charge (NTD)	2018 contribution			2019 contribution		
	1-potential	2-vulnerable	3-loyal	1-potential	2-vulnerable	3-loyal
0~10	6.15	36.75	25.96	35.39	6.15	34.71
11~20	2.7	5.5	4.42	5.18	2.57	4.1
21~30	1.92	3.76	3.07	3.56	1.74	2.73
31~50	1.35	2.62	2.17	1.49	0.93	1.13
51~70	0.3	0.8	0.64	0	0	0
71~1670	0.27	0.91	0.71	0.1	0.05	0.15
Total	12.69	50.34	36.97	45.73	11.45	42.82

4. Conclusions

In order to gain back riders, especially in this period of the pandemic, it is essential to focus on first- and last-mile issues. In this study, we examine the travel patterns of visitors to Taipei exclusively using the MRT and the bike-sharing network ("YouBike"). According to longitudinal observations of ridership patterns, travellers can be categorized into distinct groups within distinctive profiles associated with stations and their areas of influence.

This study contributes to the stream of research. Ridership categorization is derived from the Recency-Frequency-Monetary (RFM) scoring model [63], capturing travel patterns. In transit travel, the number of days since the last transfer is used as the recency parameter, the number of transfers made in a given period as the frequency parameter, and the amount of profit generated over the number of transfers made by travellers in that period as the monetary parameter. Using the K-means approach, a transit ridership portfolio reveals interesting relationships between the local environments of metro stations and urban mobility patterns that differentiate their contributions to the system.

Passes popular before the pandemic, such as monthly or weekly passes, are unlikely to appeal to workers who have switched to hybrid work schedules or riders wary of using public transportation. As a result, agencies can eliminate fares for a limited period, such as a few weeks after the summer holidays, to encourage riders to make transit part of their new routine. It is also possible for agencies to offer deep discounts during off-peak hours to reduce crowding. In addition, agencies may replace monthly or weekly unlimited passes with more flexible arrangements, such as fare capping and the option to buy one-way tickets between one origin and one destination at a specified discount valid for a certain time frame.

This research is a preliminary exploration of the riders' patterns to identify possible clusters. Utilizing data extracted from contactless smart cards has its limitations as factors affecting transit ridership in a large metropolitan setting involve more than just smart cards' POS (point of sale) records. Other critical factors such as socioeconomic characteristics, technology (e.g., the effect of intelligent transportation information systems), energy price, urbanization, and automobile dependence should not be overlooked. An examination of the aforementioned factors, combined with the formation of rider clusters, may provide better policy implications for combating congestion and carbon emissions in the near future. [67–71]

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper."

References

[1] S. M. Knupfer, V. Pokotilo, and J. Woetzel, *Elements of success: Urban Transportation Systems of 24 Global Cities*, McKinsey & Company, New York, NY, USA, 2018.

[2] Y. T. Li and J. D. Schmöcker, "Adaptation patterns to high speed rail usage in Taiwan and China," *Transportation*, vol. 44, no. 4, pp. 807–830, 2017.

[3] X. Lin, P. Wells, and B. K. Sovacool, "Benign mobility? Electric bicycles, sustainable transport consumption behaviour and socio-technical transitions in Nanjing, China," *Transportation Research Part A: Policy and Practice*, vol. 103, pp. 223–234, 2017.

[4] Y. Tang, H. Pan, and Y. Fei, "Research on users' frequency of ride in Shanghai minhang bike-sharing system," *Transportation Research Procedia*, vol. 25, pp. 4979–4987, 2017.

[5] D. Jahanshahi, Z. Tabibi, and B. Van Wee, "Factors influencing the acceptance and use of a bicycle sharing system: applying an extended Unified Theory of Acceptance and Use of Technology (UTAUT)," *Case Studies on Transport Policy*, vol. 8, no. 4, pp. 1212–1223, 2020.

[6] D. Wang and M. Zhou, "The built environment and travel behavior in urban China: a literature review," *Transportation Research Part D: Transport and Environment*, vol. 52, pp. 574–585, 2017.

[7] Z. Gan, M. Yang, T. Feng, and H. Timmermans, "Understanding urban mobility patterns from a spatiotemporal perspective: daily ridership profiles of metro stations," *Transportation*, vol. 47, no. 1, pp. 315–336, 2020.

[8] L. Yang, K. W. Chau, W. Y. Szeto, X. Cui, and X. Wang, "Accessibility to transit, by transit, and property prices: spatially varying relationships," *Transportation Research Part D: Transport and Environment*, vol. 85, Article ID 102387, 2020.

[9] Z. Chen and W. Fan, "Data analytics approach for travel time reliability pattern analysis and prediction," *Journal of Modern Transportation*, vol. 27, no. 4, pp. 250–265, 2019.

[10] A. Fan, X. Chen, and T. Wan, "How have travelers changed mode choices for first/last mile trips after the introduction of bicycle-sharing systems: an empirical study in Beijing, China," *Journal of Advanced Transportation*, vol. 2019, Article ID 5426080, 16 pages, 2019.

[11] P. Shang, Y. Yao, L. Yang, L. Meng, and P. Mo, "Integrated model for timetabling and circulation planning on an urban rail transit line: a coupled network-based flow formulation," *Networks and Spatial Economics*, vol. 21, no. 2, pp. 331–364, 2021.

[12] M. K. Kim, S. P. Kim, J. Heo, and H. G. Sohn, "Ridership patterns at subway stations of Seoul capital area and characteristics of station influence area," *KSCE Journal of Civil Engineering*, vol. 21, no. 3, pp. 964–975, 2017.

[13] G. Gao, Z. Wang, X. Liu, Q. Li, W. Wang, and J. Zhang, "Travel behavior analysis using 2016 Qingdao's household traffic surveys and Baidu electric map API data," *Journal of Advanced Transportation*, vol. 2019, Article ID 6383097, 18 pages, 2019.

[14] Y. H. Cheng and T. Y. Huang, "High speed rail passenger segmentation and ticketing channel preference," *Transportation Research Part A: Policy and Practice*, vol. 66, pp. 127–143, 2014.

[15] Y. Guan, B. Wu, and J. Jia, "Does online ticket booking system make people better off? An empirical study on railway service," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 73, pp. 143–154, 2020.

[16] X. Li, J. Tang, X. Hu, and W. Wang, "Assessing intercity multimodal choice behavior in a touristy city: a factor analysis," *Journal of Transport Geography*, vol. 86, Article ID 102776, 2020.

[17] C. Ding, D. Wang, X. Ma, and H. Li, "Predicting short-term subway ridership and prioritizing its influential factors using gradient boosting decision trees," *Sustainability*, vol. 8, no. 11, p. 1100, 2016.

[18] L. Böcker, P. van Amen, and M. Helbich, "Elderly travel frequencies and transport mode choices in Greater Rotterdam, The Netherlands," *Transportation*, vol. 44, no. 4, pp. 831–852, 2017.

[19] C. Kumar and A. Ganguly, "Travelling together but differently: comparing variations in public transit user mode choice attributes across New Delhi and New York," *Theoretical and Empirical Researches in Urban Management*, vol. 13, no. 3, pp. 54–73, 2018.

[20] D. J. Reck, H. Haitao, S. Guidon, and K. W. Axhausen, "Explaining shared micromobility usage, competition and mode choice by modelling empirical data from Zurich,

Switzerland," *Transportation Research Part C: Emerging Technologies*, vol. 124, Article ID 102947, 2021.
[21] S. Sun and Z. Duan, "Modeling passengers' loyalty to public transit in a two-dimensional framework: a case study in Xiamen, China," *Transportation Research Part A: Policy and Practice*, vol. 124, pp. 295–309, 2019.
[22] B. R. Sperry and T. Dye, "Impact of new passenger rail stations on ridership demand and passenger characteristics: hiawatha service case study," *Case Studies on Transport Policy*, vol. 8, no. 4, pp. 1158–1169, 2020.
[23] M. Bordagaray, L. Dell'Olio, A. Fonzone, and Á. Ibeas, "Capturing the conditions that introduce systematic variation in bike-sharing travel behavior using data mining techniques," *Transportation Research Part C: Emerging Technologies*, vol. 71, pp. 231–248, 2016.
[24] C. Chen, J. Ma, Y. Susilo, Y. Liu, and M. Wang, "The promises of big data and small data for travel behavior (aka human mobility) analysis," *Transportation Research Part C: Emerging Technologies*, vol. 68, pp. 285–299, 2016.
[25] J. De Vos, P. L. Mokhtarian, T. Schwanen, V. Van Acker, and F. Witlox, "Travel mode choice and travel satisfaction: bridging the gap between decision utility and experienced utility," *Transportation*, vol. 43, no. 5, pp. 771–796, 2016.
[26] L. Duan, J. Rezaei, L. Tavasszy, and C. Chorus, "Heterogeneous valuation of quality dimensions of railway freight service by Chinese shippers: choice-based conjoint analysis," *Transportation Research Record*, vol. 2546, no. 1, pp. 9–16, 2016.
[27] J. Hagenauer and M. Helbich, "A comparative study of machine learning classifiers for modeling travel mode choice," *Expert Systems with Applications*, vol. 78, pp. 273–282, 2017.
[28] Y. Ma and Z. Zhang, "Travel mode choice prediction using deep neural networks with entity embeddings," *IEEE Access*, vol. 8, Article ID 64959, 2020.
[29] C. Qian, M. Yang, P. Li, and S. Li, "Application of customer segmentation for electronic toll collection: a case study," *Journal of Advanced Transportation*, vol. 2018, Article ID 3635107, 9 pages, 2018.
[30] X. Yan, X. Liu, and X. Zhao, "Using machine learning for direct demand modeling of ridesourcing services in Chicago," *Journal of Transport Geography*, vol. 83, Article ID 102661, 2020.
[31] X. Zhao, X. Yan, A. Yu, and P. Van Hentenryck, "Prediction and behavioral analysis of travel mode choice: a comparison of machine learning and logit models," *Travel Behaviour and Society*, vol. 20, pp. 22–35, 2020.
[32] S. Wang, B. Mo, S. Hess, and J. Zhao, "Comparing Hundreds of Machine Learning Classifiers and Discrete Choice Models in Predicting Travel Behavior: An Empirical Benchmark," 2021, https://arxiv.org/abs/2102.01130.
[33] E. Nikumanesh and A. Albadvi, "Customer's life–time value using the RFM model in the banking industry: a case study," *International Journal of Electronic Customer Relationship Management*, vol. 8, no. 1/2/3, pp. 15–30, 2014.
[34] J. Zhou, D. Huang, and H. Wang, "RFMS method for credit scoring based on bank card transaction data," *Statistica Sinica*, vol. 28, no. 4, pp. 2903–2919, 2018.
[35] T. Kovács, A. Ko, and A. Asemi, "Exploration of the investment patterns of potential retail banking customers using two-stage cluster analysis," *Journal of Big Data*, vol. 8, no. 1, pp. 141–225, 2021.
[36] I. Benítez, A. Quijano, J. L. Díez, and I. Delgado, "Dynamic clustering segmentation applied to load profiles of energy consumption from Spanish customers," *International Journal of Electrical Power & Energy Systems*, vol. 55, pp. 437–448, 2014.
[37] P. A. Sarvari, A. Ustundag, and H. Takci, "Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis," *Kybernetes*, vol. 45, no. 7, pp. 1129–1157, 2016.
[38] S. Peker, A. Kocyigit, and P. E. Eren, "LRFMP model for customer segmentation in the grocery retail industry: a case study," *Marketing Intelligence & Planning*, vol. 35, no. 4, pp. 544–559, 2017.
[39] C. Maione, D. R. Nelson, and R. M. Barbosa, "Research on social data by means of cluster analysis," *Applied Computing and Informatics*, vol. 15, no. 2, pp. 153–162, 2019.
[40] D. Arunachalam and N. Kumar, "Benefit-based consumer segmentation and performance evaluation of clustering approaches: an evidence of data-driven decision-making," *Expert Systems with Applications*, vol. 111, pp. 11–34, 2018.
[41] K. Khalili-Damghani, F. Abdi, and S. Abolmakarem, "Hybrid soft computing approach based on clustering, rule mining, and decision tree analysis for customer segmentation problem: real case of customer-centric industries," *Applied Soft Computing*, vol. 73, pp. 816–828, 2018.
[42] C. Yan, H. Sun, W. Liu, and J. Chen, "An integrated method based on hesitant fuzzy theory and RFM model to insurance customers' segmentation and lifetime value determination," *Journal of Intelligent and Fuzzy Systems*, vol. 35, no. 1, pp. 159–169, 2018.
[43] S. Shokohyar, A. Rezaeian, and A. Boroufar, "Identifying the customer behavior model in life insurance Sector using data mining," *Management Research in Iran*, vol. 20, no. 4, pp. 65–94, 2021.
[44] H. I. Arumawadu, R. M. K. T. Rathnayaka, and S. K. Illangarathne, "Mining profitability of telecommunication customers using K-means clustering," *Journal of Data Analysis and Information Processing*, vol. 3, no. 3, pp. 63–71, 2015.
[45] V. Babaiyan and S. A. Sarfarazi, "Analyzing customers of South Khorasan telecommunication company with expansion of RFM to LRFM model," *Journal of AI and Data Mining*, vol. 7, no. 2, pp. 331–340, 2019.
[46] H. Chen, L. Zhang, X. Chu, and B. Yan, "Smartphone customer segmentation based on the usage pattern," *Advanced Engineering Informatics*, vol. 42, Article ID 101000, 2019.
[47] J. Li, A. D. Nigatu, B. Yu, Q. Gu, and Y. Yang, "Understanding customer value in the mobile internet era," *Discrete Dynamics in Nature and Society*, vol. 2021, Article ID 9585743, 14 pages, 2021.
[48] Y. Xiao, M. Lin, Y. Huang et al., "Using improved RFM model to classify consumer in big data environment," *International Journal of Embedded Systems*, vol. 14, no. 1, pp. 54–64, 2021.
[49] R. Madlenak and L. Madlenakova, "Comparison of regional postal transportation networks in Zilina region," in *Proceedings of the 19th International Scientific Conference on Transport Means*, pp. 22-23, Lithuania, October 2015.
[50] P. Kolarovszki, J. Tengler, and M. Majerčáková, "The new model of customer segmentation in postal enterprises," *Procedia-Social and Behavioral Sciences*, vol. 230, pp. 121–127, 2016.
[51] J. Reades, C. Zhong, E. D. Manley, R. Milton, and M. Batty, "Finding pearls in London's oysters," *Built Environment*, vol. 42, no. 3, pp. 365–381, 2016.
[52] B. Zhang and Q. Y. Peng, "Railway freight customer segmentation based on KFAV model," *Journal of Transportation*

Systems Engineering and Information Technology, vol. 17, no. 3, p. 235, 2017.
[53] M. DehghaniZadeh, M. Fathian, and M. Gholamian, "LDcFR: a new model to determine value of airline passengers," *Tourism and Hospitality Research*, vol. 18, no. 3, pp. 357–366, 2018.
[54] W. Y. Chiang, "Establishing high value markets for data-driven customer relationship management systems: an empirical case study," *Kybernetes*, vol. 48, no. 3, pp. 650–662, 2019.
[55] Z. Wei and X. Shan, "RFM high-speed railway customer value classification model based on spark IOP Conference Series: materials Science and Engineering," *IOP Conference Series: Materials Science and Engineering*, vol. 563, no. 5, Article ID 52081, 2019, July.
[56] U. Demi̇r Alan and D. Birant, "Passenger scoring for free-pass promotion in public transportation," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 28, no. 4, pp. 2110–2127, 2020.
[57] X. Fu, X. Chen, Y. T. Shi, I. Bose, and S. Cai, "User segmentation for retention management in online social games," *Decision Support Systems*, vol. 101, pp. 51–68, 2017.
[58] V. Holý, O. Sokol, and M. Černý, "Clustering retail products based on customer behaviour," *Applied Soft Computing*, vol. 60, pp. 752–762, 2017.
[59] W. Y. Chiang, "Applying data mining for online CRM marketing strategy: an empirical case of coffee shop industry in Taiwan," *British Food Journal*, vol. 120, no. 3, pp. 665–675, 2018.
[60] D. H. Nguyen, S. de Leeuw, and W. E. Dullaert, "Consumer behaviour and order fulfilment in online retailing: a systematic review," *International Journal of Management Reviews*, vol. 20, no. 2, pp. 255–276, 2018.
[61] P. Anitha and M. M. Patil, "RFM Model for Customer purchase Behavior Using K-Means Algorithm," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, 2019 in press.
[62] A. J. Christy, A. Umamakeswari, L. Priyatharsini, and A. Neyaa, "RFM ranking–An effective approach to customer segmentation," *Journal of King Saud University-Computer and Information Sciences*, vol. 33, no. 10, pp. 1251–1257, 2021.
[63] G. J. Cullinan, "Picking Them by Their Batting Averages' Recency-Frequency-Monetary Method of Controlling Circulation," *Manual Release*, vol. 2103, 1977.
[64] J. MacQueen, "Some methods for classification and analysis of multivariate observations," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1pp. 281–29714, Berkeley, CA, USA, June 1967.
[65] M. Horvat, A. Jović, and K. Burnik, "Assessing the robustness of cluster solutions in emotionally-annotated pictures using monte-carlo simulation stabilized K-means algorithm," *Machine Learning and Knowledge Extraction*, vol. 3, no. 2, pp. 435–452, 2021.
[66] R. Raza, S. Saeed, and R. Sharif, "Ink Detection Using K-Mean Clustering in Hyperspectral Document," Preprints, Article ID 2020070002, 2020.