


## Data Analysis Automatic Fare Collection Light Rail Transit Jakarta using the Cluster Method

Mario Hagi<sup>1</sup>, Heri Suroyo<sup>2</sup>, Alek Wijaya<sup>3</sup>, Ahmad Syazili<sup>4</sup>  
<sup>1,2,3,4</sup>Faculty of Sains Teknologi, Bina Darma University, Indonesia

### ABSTRACT

The use of public transportation such as LRT (Light Rail Transit) can reduce the level of congestion in an area and DKI Jakarta is a province with a high level of congestion. The purpose of this study was to determine the general preferences of passengers and to classify stations based on the number of passengers entering-station-exit pairs of the Jakarta LRT by utilizing the cluster analysis method. The data analyzed using the Tableau Desktop application was obtained from the AFC (Automatic Fare Collection) LRT Jakarta. The data contains seven fields, namely PAYMENT METHOD, DATE, TIME OUT, RANGE 60', RANGE 15', STATION OUT, STATION IN for 15 days from 1 January 2023 to 15 January 2023. The results of the study are in the form of data visualization, descriptive statistics, clustering results, and dashboard. Based on the results of the analysis, three clusters were formed with cluster 3 being filled by VEL-BVU and BVU-VEL by controlling 14038 passengers or 40.8% of the total passengers, cluster 2 being filled by VEL-DPD, VEL-BVS, DPD-VEL, and BVS -VEL controlling 10,053 passengers or 29.2% of the total passengers, and the remaining 27 items are in cluster 1 controlling 10,317 passengers or 30.0% of the total passengers.

**Keyword : AFC; LRT Jakarta; general preferences of passengers; Cluster Analysis.**

 This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

#### Corresponding Author:

Mario Hagi,  
Faculty of Sains Teknologi  
Bina Darma University  
Jl. Jenderal A. Yani No. 3 Palembang, Indonesia, 20238.  
Email : hagi864@gmail.com

#### Article history:

Received Aug 28, 2023  
Revised Aug 31, 2023  
Accepted Sep 2, 2023

### 1. INTRODUCTION

In 2019, the population density of DKI Jakarta reached 15,900 people per square kilometer, and will increase to 15,978 people per square kilometer in 2021. This figure far exceeds the average population density in Indonesia, which is only around 140 people per square kilometer in 2019 and 142 inhabitants per square kilometer in 2021 (Badan Pusat Statistik, 2021).

The impact of this high level of population density is the frequent occurrence of traffic jams on the streets of DKI Jakarta, especially during the hours of going to and returning from work around 8 am and 5 pm. Before Covid-19 and the work from home (WFH) policy, in 2019, the level of congestion in DKI Jakarta reached around 55%, it could even reach more than 60% during working hours. This means that traveling on DKI Jakarta roads takes 60% longer than traveling on roads that are not jammed. For example, a trip that normally takes 1 hour on a non-trafficked road can take around 1 hour and 36 minutes on a DKI Jakarta road during business hours (TomTom).

To overcome the congestion problem, the government has built various modes of public transportation, one of which is the Jakarta LRT which is being built in stages. However, so that people want to switch to using the Jakarta LRT from private vehicles, attention is needed on a number of things, including the quality of service and promotion. Promotion and service quality have a simultaneous effect on purchasing decisions for rail transportation services (Gumaeri et al., 2022). Service quality has a significant effect on passenger loyalty at PT Kereta Api Indonesia (Utari, 2019).

To ensure the right quality of service and promotion, LRT Jakarta needs to evaluate, improve and eliminate elements in its operational activities. This step can be achieved by collecting information about the general preferences of passengers and grouping stations based on the number of passengers in pairs of entry-exit stations through analysis of available passenger travel data. AFC became a valuable data source for this analysis. Therefore, the authors will conduct research by analyzing AFC data using the cluster method.

By paying attention to the congestion problems caused by high population density, as well as by improving the quality of services and appropriate promotions, it is hoped that the Jakarta LRT can become an alternative transportation that is in demand by the public. With a deep understanding of passenger needs and effective strategy implementation, LRT Jakarta has the potential to gain stronger passenger trust and loyalty in the long term.

## 2. LITERATURE REVIEW

Many studies have utilized AFC data and conducted research from several perspectives, including passenger statistics and performance indicators, to support operational analysis, identification of travel patterns, and network analysis for service planning and behavior analysis to facilitate long-term public transport planning (Cevallos et al., 2021). One of them is research conducted by Otero Niño and Julián Darío who conducted research using AFC data to find out patterns of passenger travel behavior which means the same as what the authors are currently researching, only the scope of the AFC data they studied is wider because it includes an integrated system. The final result of their research is to find out the purpose of passenger travel, therefore they also utilize a different analysis method, namely the trip chaining method. While the final results that the authors achieved in this study were to determine the general preferences of passengers and group pairs of incoming-station outgoing stations based on the number of passengers by utilizing the cluster method. The tools or applications used for research are also different, they use the R programming language while the authors use the Tableau Desktop application (Niño and Darío, 2022).

Research conducted by M. W. Talakua et al used the k-means method in conducting cluster analysis with the aim of classifying regencies/cities in Maluku province based on the 2014 human development index indicators. They used SPSS as a data processor (Talakua et al., 2017), in contrast to the authors who used Tableau Desktop, which in SPSS uses the Euclidian distance to measure the distance of the data to the centroid, but Tableau Desktop uses the squared Euclidian distance.

Tableau was also used by Dandie Triyanto et al in conducting his research entitled "Implementation of Business Intelligence Using Tableau to Visualize Data on the Impact of Flood Disasters in Indonesia" (Triyanto et al., 2023). They display a lot of data visualization with dashboards as the end result with descriptive statistical methods in explaining the data.

## 3. RESEARCH METHOD

### A. Research Approach

The research method is the main method used by researchers to achieve goals and determine answers to the problems posed. Research will use quantitative methods (Arikunto and Suharsimi, 2019 & Sukandarrumidi, 2012). Quantitative research is a type of research that produces discoveries that can be achieved (obtained) using statistical procedures or other means of quantification (measurement) (Sujarweni, 2014). The quantitative method was chosen because the main data to be analyzed is numerical data. Quantitative data is a research method that is based on positivistic (concrete data), research data is in the form of numbers that will be measured using statistics as a calculation test tool, related to the problem under study to produce a conclusion. Positivistic philosophy is used in certain populations or samples (Sugiyono, 2018).

### B. Data Collection

Secondary data are data sources that do not directly provide data to data collectors, for example through other people or through documents. In this study the data to be analyzed is data taken from the AFC LRT Jakarta (Sugiyono, 2018).

Population is a generalized area consisting of objects or subjects that have certain qualities and characteristics set by researchers to study and then conclusions are drawn (Sugiyono, 2018). The population in this study is data collected by the Jakarta AFC LRT system. The sample is part of the number and characteristics possessed by the population. While sample size is a step to determine the size of the sample taken in carrying out a study (Sugiyono, 2018). Using the purposive sampling method, researchers will use sample data collected by the AFC LRT Jakarta system for 15 days, namely from 1 January 2023 to 15 January 2023. Purposive sampling is sampling using certain considerations according to the desired criteria to be able to determine the amount sample to be studied (Sugiyono, 2018). The data was selected with the following criteria:

1. Data for 15 consecutive days.
2. The latest data at the time of collection.

3. The maximum number of days off from the amount of data retrieved.

### C. Data Processing

Explaining Summary Cards are available on the Show/Hide Cards toolbar menu, providing a quick view of information about a selection or an entire data source. By default, the Summary Card displays the Sum, Average (mean), Minimum, Maximum, and Median values for the data in view (Tableau, 2022).

These features are descriptive statistical analysis because according to what Sugiyono said, descriptive statistical analysis is a data analysis technique to explain data in general or generalizations, by calculating the minimum value, maximum value, average value (mean), and standard deviation (standard deviation) (Sugiyono, 2017).

Clustering is the process of making groupings so that all members of each partition have similarities based on a certain matrix. Cluster analysis or group analysis is a data analysis technique that aims to classify individuals or objects into several groups that have different characteristics between groups, so that individuals or objects that are located in one group will have relatively homogeneous properties. The purpose of cluster analysis is to group these objects (Talakua et al., 2017). Cluster analysis will be used by the author to group starting stations - destination stations based on the number of daily passengers and other things that will be useful later if they are grouped.

Cluster analysis partitions the signs in the view into clusters, where the signs in each cluster are more similar to each other than the signs in other clusters. Tableau distinguishes clusters using colors. Tableau provides cluster analysis using the k-means algorithm. For a specified number of k clusters, the algorithm partitions the data into k clusters. Each cluster has a center (centroid) which is the average value of all points in the cluster. K-means locates the center through an iterative procedure that minimizes the distance between individual points in a cluster and the cluster center. In Tableau, users can specify the desired number of clusters, or have Tableau test different values of k and suggest the optimal number of clusters.

Tableau uses Lloyd's algorithm with squared Euclidian Distance to calculate the k-means grouping for each k. Combined with the splitting procedure to determine the initial center for any  $k > 1$ , the resulting clustering is deterministic, with the results depending only on the number of clusters (Tableau, 2022).

The steps in the k-means clustering algorithm are:

1. Determine the number of clusters
2. Determine the centroid value

In determining the centroid value for the initial iteration, the initial centroid value is done randomly. Meanwhile, if you determine the centroid value which is the stage of the iteration, then the following formula is used.

$$\bar{v}_{ij} = \frac{1}{N_i} \sum_{k=0}^{N_i} x_{kj} , \quad (1)$$

Where:

$v_{ij}$  = centroid/ average of the i-th cluster for the j-th variable

$N_i$  = the amount of data that is a member of the i-th cluster

$i, k$  = index of cluster

$j$  = index of variable

$x_{kj}$  = the k-th data value in the cluster for the j-variable

3. Calculate the distance between the centroid point and the point of each object. To calculate these distances, Tableau uses the squared Euclidean Distance, ie

$$D_e = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2} , \quad (2)$$

Where :

$D_e$  = Euclidean Distance

$i$  = the number of objects,

$(x,y)$  = object coordinates and

$(s,t)$  = centroid coordinates.

4. Grouping objects

To determine the cluster members is to take into account the minimum distance of objects. The value obtained in the data membership in the distance matrix is 0 or 1, where the value is 1 for the data allocated to the cluster and the value is 0 for the data allocated to other clusters.

5. Return to step 2, repeat until the resulting centroid value is fixed and cluster members do not move to other clusters [16].

### 3. RESULTS AND DISCUSSION

In AFC there are two types of payment methods, namely QR Code and Prepaid. QR Code is a type of payment that uses a digital wallet while Prepaid is a type of payment that uses an electronic card.

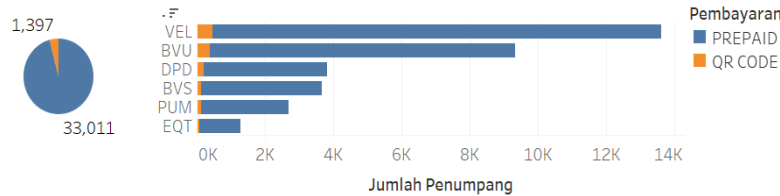


Figure 1. Payment Method

Based on the stacked horizontal bars graph above, the Prepaid payment method is much more widely used with a percentage of 95.5% compared to the QR Code payment method which is only used by 4.5%. The number of uses of the two payment methods is also spread evenly according to the number of passengers at each station.

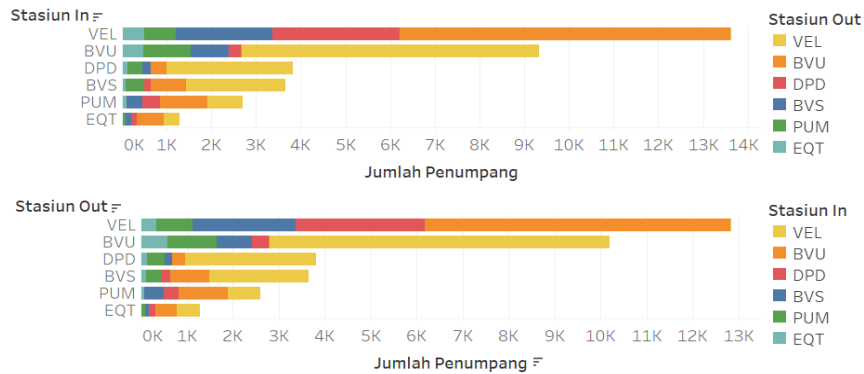


Figure 2. Distribution of passengers by station

Based on the stacked horizontal bars graph above, it can be seen that the distribution of passengers by out/out station and in/in station is almost identical. The number of passengers was dominated by VEL stations and BVU stations, both of which controlled 66.6% of the total passengers at the six LRT Jakarta stations.

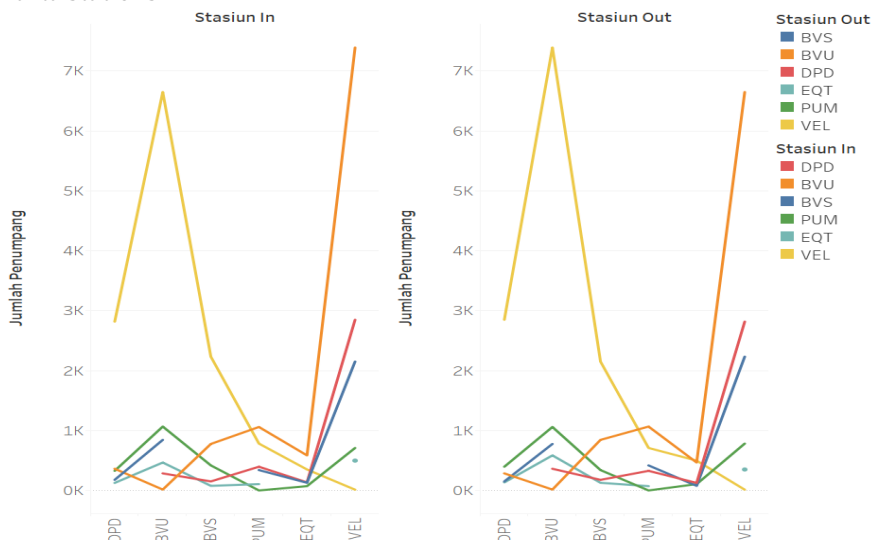


Figure 3. Distribution of passengers by station (2)

From the continuous lines graph above, it can be confirmed that the difference in the distribution of passengers based on exit and entry stations is quite identical, it's just that the relationship between VEL and BVU is slightly different in that more passengers enter the VEL station to leave at the BVU station than vice versa. In addition, the difference between the graphs of the exit stations and incoming stations does not exist or is almost the same, which means that the relationship between incoming and outgoing stations is very strong.

The distribution of passengers from the VEL station itself is quite large, apart from relying on BVU stations, VEL stations can also be seen relying on DPD and BVS stations, compared to BVU stations which only heavily rely on VEL stations.

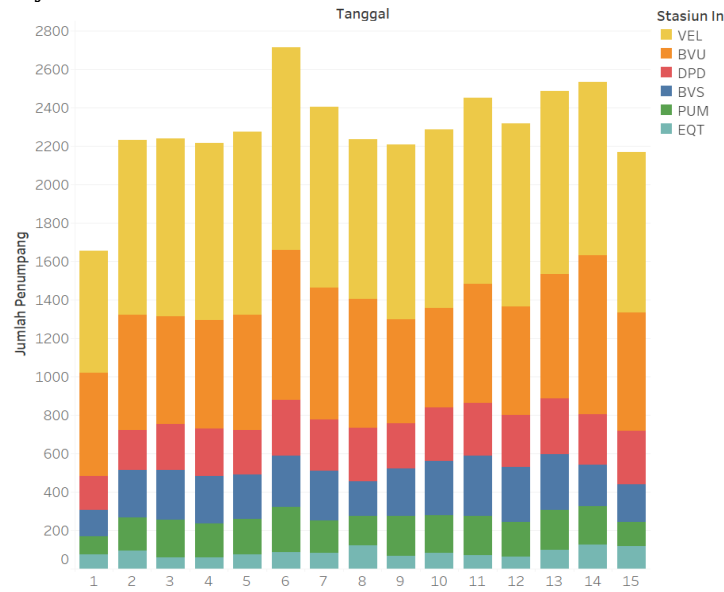


Figure 4. Distribution of passengers by day

The number of passengers also does not differ much every day except on the 1st which is the New Year's red date. For 15 days, the number of passengers was 34408 with a mean of 2294 passengers per day, a minimum of 1653 and a maximum of 2714 and the median was 2274.

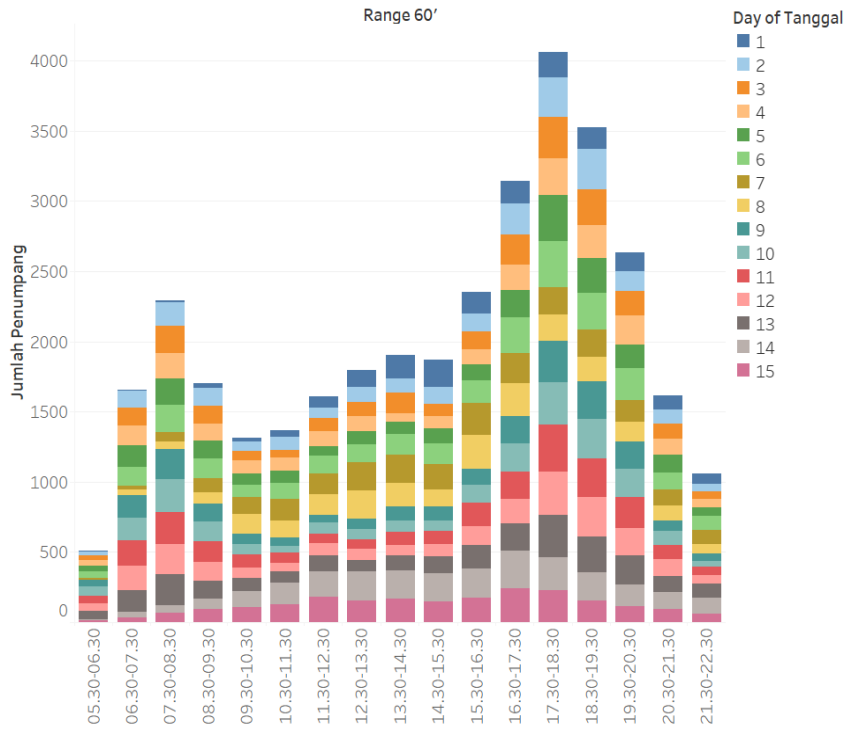


Figure 5. Distribution of passengers by hour

The number of passengers is quite prominent during the hours going to work, then slightly decreased. After that in the afternoon the number of passengers increased dramatically. If the graph is divided into two categories, namely morning hours (05:30-14:30) and afternoon hours (13:30-22:30) with a note that 13:30-14:30 is included in both categories with the reason that the ratio is 50: 50, morning hours have 14152 passengers (41.1%) and afternoon hours have 22162 passengers (64.4%). Overall, the average number of passengers per hour from 17 hours of operation is 2024, a minimum of 512, a maximum of 4059, and a median of 1794.

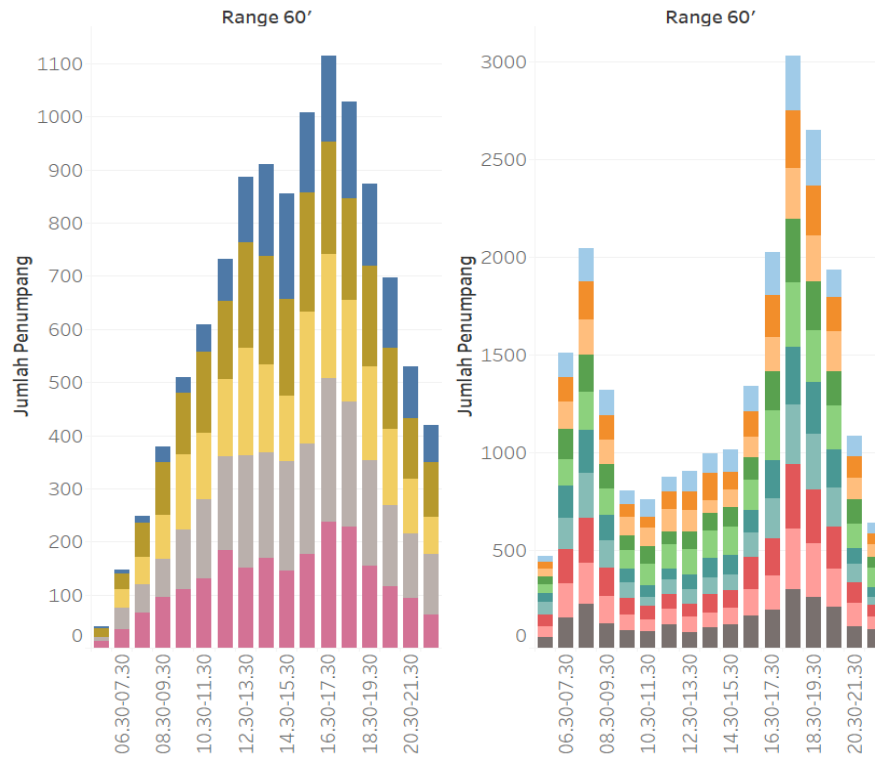


Figure 6. Distribution of passengers by hour (2)

On holidays (left graph), Saturday and Sunday, namely on the 1st, 7th, 8th, 14th and 15th, the number of passengers has increased starting from the start of operating hours 05:30-06:30 with a total of 41 passengers until the peak at 16:30-17:30 with a total of 1114 passengers, after that the number of passengers decreased until the end of operating hours 21:30-22:10 with a total of 419 passengers.

On weekdays (chart on the right), Monday to Friday, namely on the 2nd, 3rd, 4th, 5th, 6th, 9th, 10th, 11th, 12th, and 13th the number of passengers stands out around the hours of departure and return from work, namely around 8 o'clock :00 and 18:00 with afternoon hours which are still more crowded than morning hours.

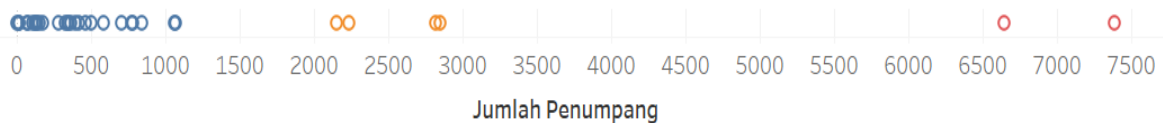


Figure 7. Clustering result

The number of clusters is 3 which is determined automatically, the number of points is 33, Between-group sum of squares is 1.683, within-group sum of squares is 0.06043, and total sum of squares is 1.7435.

Table 1. Statistics of Formed Clusters

Clusters	Jumlah Member	Centroid
1	27	382.11
2	4	2513.2
3	2	7019.0
Not Clustered	0	

Stasiun Out	DPD	BVU	BVS	PUM	EQT	VEL
DPD		○	○	○	○	○
BVU	○	○	○	○	○	○
BVS	○	○		○	○	○
PUM	○	○	○	○	○	○
EQT	○	○	○	○		○
VEL	○	○	○	○	○	

Figure 8. Table of incoming-outgoing station pairs based on clusters

Cluster 3 is filled by VEL-BVU and BVU-VEL by controlling 14,038 passengers or 40.8% of the total passengers. Cluster 2 is filled by VEL-DPD, VEL-BVS, DPD-VEL, and BVS-VEL by controlling 10,053 passengers or 29.2% of the total passengers. The rest are in cluster 3 controlling 10,317 passengers or 30.0% of the total passengers. From the table above, it can be seen that there were several passengers whose entry and exit stations were the same as 34 passengers or only 0.1% of the total number of passengers.

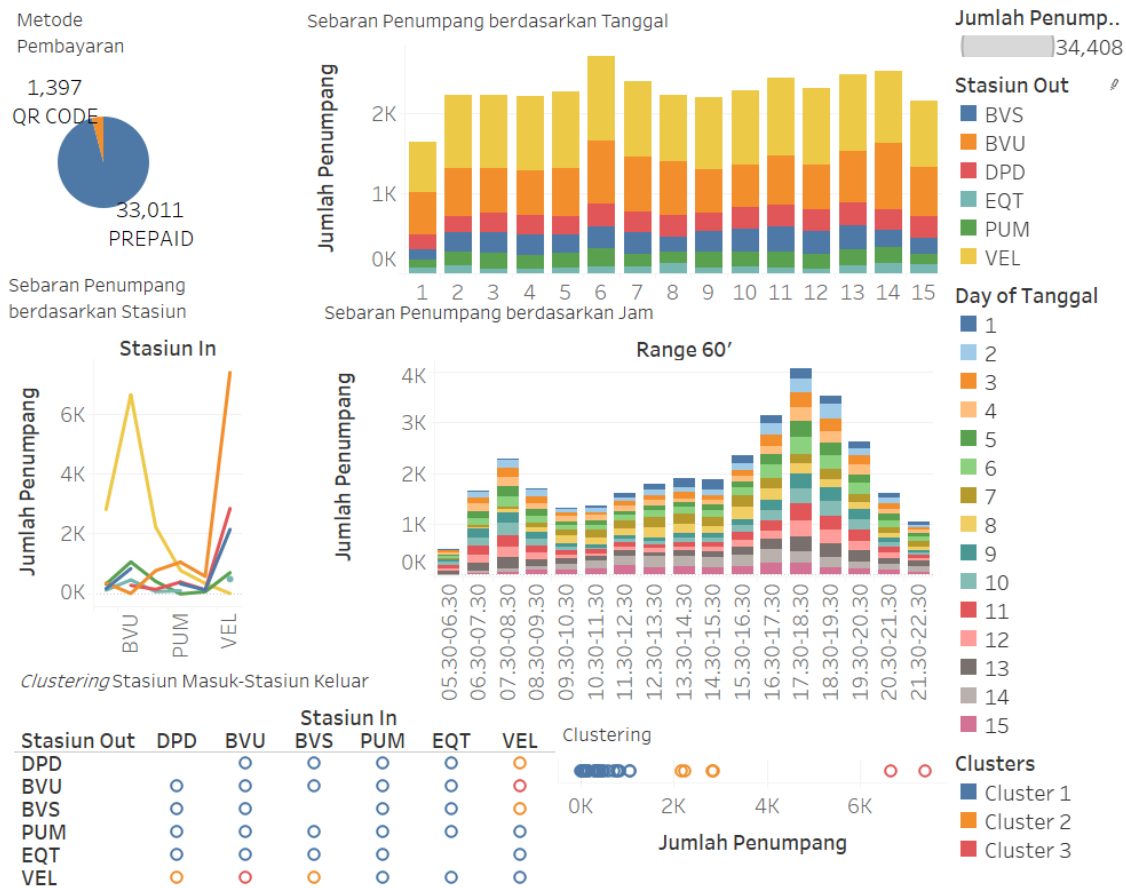


Figure 9. Dashboard

The image above is the dashboard of the overall results of the Jakarta LRT AFC data analysis which includes various visualizations such as passenger distribution, payment methods, and clustering results. With its components, namely pie charts, stacked bars, lines (continuous), and tables. Dashboard is a display that visually presents data, numbers, and metrics that are useful for providing information and making it easier for users to make decisions quickly and accurately based on available data [8].

#### 4. CONCLUSION

## A. Conclusion

1. The prepaid payment method is much more frequently used at all stations with a percentage of 95.5%.
2. The relationship between incoming-station-exit stations is very strong, which means that if the DPD-VEL has many passengers, the VEL-DPD will also have many passengers.
3. The number of passengers is dominated by VEL stations and BVU stations, both of which control as much as 66.6% of the number of passengers at the six LRT Jakarta stations
4. More passengers enter the VEL station to exit at the BVU station than vice versa.
5. The distribution of passengers to and from the VEL station itself is quite large, besides relying on BVU stations, VEL stations also rely on DPD and BVS stations, compared to BVU stations which only rely heavily on VEL stations.
6. There is no significant difference in the number of passengers per day
7. The number of passengers for 15 days was 34408 with a mean daily passenger of 2294 with a minimum of 1653 and a maximum of 2714 and a median of 2274.
8. The number of passengers is greater in the afternoon than in the morning and in both categories of hours, the peak is at the time of going to work and returning from work.
9. Based on the number of passengers from the entry station to the exit station, three clusters are formed. Cluster 3, there are two clusters with the most number of passengers, namely VEL-BVU and BVU-VEL. There are four Cluster 2 namely VEL-DPD, VEL-BVS, DPD-VEL, and BVS-VEL. The remaining 27 are in cluster 1.

## B. Recommendation

1. Optimization of Payment Methods: Expand prepaid payment methods because their use is very significant but there is no need to remove the QR Code payment method.
2. Service Improvement at VEL and BVU Stations: Recognizing the dominance of the number of passengers at VEL and BVU stations, an event can be held there to increase revenue and attractiveness, especially in the afternoon, given that the station's crowd level is busier at that hour.
3. Increasing Connectivity Between Stations: Taking into account the strong connections between several stations such as VEL-DPD and VEL-BVS, if there are more stations later, optimization of train trips can be carried out.
4. Continuing Data Collection and Analysis: It is important to continue to collect data and continue with more in-depth analysis in the long term. This can provide more complete insight into long-term trends and aid in better strategic.

## REFERENCES

- Badan Pusat Statistik. (2021). Kepadatan Penduduk menurut Provinsi (jiwa/km<sup>2</sup>) 2019-2021. *Badan Pusat Statistik*. Available: <https://www.bps.go.id/indicator/12/141/1/kepadatan-penduduk-menurut-provinsi.html>. [Accessed 08-Feb-2023].
- TomTom. (2023). Jakarta traffic report. *TomTom*. Available: <https://www.tomtom.com/traffic-index/jakarta-traffic/>. [Accessed 08-Feb-2023].
- Gumaeri, F. and Hendriyani, R. M. (2022). Pengaruh Promosi Dan Kualitas Layanan Terhadap Keputusan Pembelian Pada Jasa Transportasi Kereta Api Di Tengah Pandemi Studi Kasus Penumpang Kereta Api Stasiun Cikampek. *Journal of Management*, Volume 5, Issue 3, Pages 64-75.
- Utari, S. A. (2019). Pengaruh Kualitas Pelayanan dan Kepuasan Konsumen terhadap Loyalitas Penumpang Pengguna Jasa Kereta Api Bisnis pada PT. Kereta Api Indonesia (Persero). *Muhammadiyah Sumatera Utara University*. Available: <http://repository.umsu.ac.id/bitstream/handle/123456789/3017/Pengaruh%20Kualitas%20Pelayanan%20dan%20Kepuasan%20Konsumen%20Terhadap%20Loyalitas%20Penumpang%20Pengguna%20Jasa%20Kereta%20Api%20Bisnis%20Pada%20PT.%20Kereta%20Api%20Indonesia%20%28Persero%29.pdf>. [Accessed 19-Feb-2023].
- Cevallos, F., Torino, L. and Jin, X. (2021). A Synthesis on Data Mining Methods and Applications for Automated Fare Collection (AFC) Data. *Bureau of Transportation Statistics*. Available: <https://rosap.ntl.bts.gov/view/dot/61846>. [Accessed 19-Feb-2023].
- Niño, O. and Darío, J. (2022). Using automatic fare collection data to reveal travel behaviour patterns: The case of Bogotá's transit system. *Research Gate*. Available: [https://www.researchgate.net/publication/358711685\\_Using\\_automatic\\_fare\\_collection\\_data\\_to\\_reveal\\_travel\\_behaviour\\_patterns\\_The\\_case\\_of\\_Bogota's\\_transit\\_system](https://www.researchgate.net/publication/358711685_Using_automatic_fare_collection_data_to_reveal_travel_behaviour_patterns_The_case_of_Bogota's_transit_system). [Accessed 12-Aug-2023].
- Talakua, M. W., Leleury, Z. A. and Taluta, A. W. (2017). ANALISIS CLUSTER DENGAN MENGGUNAKAN METODE K-MEANS UNTUK PENGELOMPOKKAN KABUPATEN/KOTA DI PROVINSI MALUKU BERDASARKAN INDIKATOR



- INDEKS PEMBANGUNAN MANUSIA TAHUN 2014. *BAREKENG: Jurnal Ilmu Matematika Dan Terapan*, 11(2), 119-128.
- Triyanto, D., Sholeh, M. and Hasan, F. N. (2023). Implementasi Business Intelligence Menggunakan Tableau Untuk Visualisasi Data Dampak Bencana Banjir di Indonesia. *KLIK: KAJIAN ILMIAH INFORMATIKA DAN KOMPUTER*, Vol. 3, No. 6.
- Arikunto and Suharsimi. (2019). *Prosedur Penelitian Suatu Pendekatan Praktik*. Jakarta: Rineka Cipta.
- Sukandarrumidi. (2012). *Metodologi Penelitian: Petunjuk Praktis untuk Peneliti Pemula*. Yogyakarta: Gajah Mada University Press.
- Sujarweni, V. W. (2014). *Metode Penelitian: Lengkap, Praktis, dan Mudah Dipahami*. Yogyakarta: Pustaka Baru Press.
- Sugiyono. (2018). *Metode Penelitian Kuantitatif, Kualitatif, dan R&D*. Bandung: Alfabeta.
- Tableau. (2022). Tableau Desktop and Web Authoring Help. *Tableau*. Available: [https://help.tableau.com/current/pro/desktop/en-us/inspectdata\\_summary.htm](https://help.tableau.com/current/pro/desktop/en-us/inspectdata_summary.htm). [Accessed 10-Aug-2023].
- Sugiyono. (2017). *Metode Penelitian Kuantitatif, Kualitatif, dan R&D*. Bandung: Alfabeta.
- Tableau. (2022). Tableau Desktop and Web Authoring Help. *Tableau*. Available: <https://help.tableau.com/current/pro/desktop/en-us/clustering.htm>. [Accessed 29-Jul-2023].
- Wahidah, N. CLUSTERING MENGGUNAKAN K-MEANS ALGORITHM. *Jurnal Transformatika*, 8(1), 33-39, 2010.