ISSN: 2622-5778 (online)
Advances in Transportation and Logistics Research
http://proceedings.itltrisakti.ac.id/index.php/altr

IMPUTATION PROCESS FOR MISSING DATA IN CARGO TRAIN SERVICES

Siti Nasuha Zubir[1], S.Sarifah Radiah Shariff[2], Wan Mazlina Wan Mohamed[3]
1,2. Centre for Statistics and Decision Science Studies, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Malaysia,
2,3. Malaysia Institute of Transport (MITRANS), Universiti Teknologi MARA Shah Alam, Malaysia
*Corresponding author: shari990@uitm.edu.my

Abstract. Derailments of cargo train have frequently occurred during the last decade. Many factors contribute to this incident, especially its total amount of carried weight. Severe derailments cause damage to both lives and properties every year. If the amount of carried weight of cargo train could be accurately forecasted in advance, then its detrimental effect could be greatly minimized. The major aim of the study is to model and predict the amount of carried weight of cargo train by routes in the rail transportation system, however, the unavailability of complete data is always a big problem. This study shows the process of handling missing values in the data. The data in carried cement in twelve (12) different routes by Keretapi Tanah Melayu Berhad (KTMB) cargo for 2016 -2018 is used as a case study.

Keywords: Carried weight, cargo derailment, data imputation, missing data

Introduction (Include Literature Review)

Cargo refers to goods or product that are transferred of distributed generally for commercial gain. Nowadays, cargo can be transported via water, air or land. Cargo comprises all types of freight, including those goods transported by van, truck, airplane, ship, inter-modal container or train. In order to stabilize and secure carried load, there are many different ways and materials available. During the old-time, cargo was originally a shipload. In this era, ship to carry cargo are still relevant. Container shipping is the one of the most important and necessary means of cargo transportation. However, the ship can only sail when a certain amount of cargo to be transported is met and this result in low frequency. Hence, most widely used to carry cargo is road transport. Road transport has many advantages such that it can be done from door to door delivery on top of having several types of alternative vehicles like trucks, busses, lorry, cars and so on. However, some bulky items like sugar, cements, charcoals that need to be transferred in large volume are moved using train or rail transport.

Other than known as able to carry passengers, train is also capable of transporting large volume of items such as water, cement, steel, wood and coal. Generally, train cargo has a direct route to its destination. Under the right condition, cargo transport by rail is more economic and more productive compared to road transport, especially when transporting items in large volume over long distance. The choice of mode of transportation depends very much on carried weight. Carried weight is an important matter in transport system and need to be considered. In the logistic transportation system, the amount of weight carried is very important to ensure that all the goods arrive safely at the destination in time. In Peninsular Malaysia, Keretapi Tanah Melayu Berhad (KTMB) is the main rail operator that provides rail cargo service. Cargo services provided by KTMB are safe, efficient and trustworthy (KTMB, 2019).

However, in 2017, KTMB has experienced three (3) major derailments. On August 21, 2017, a cargo train crashed at Jalan Kucing, causing delays for a few days (Noor, 2017). On September 23, 2017, KTMB's cargo train snapped electrical cable between Rawang and Kuang stations and forcing KTMB to close all tracks for two (2) days (My Metro, 2017). On November 23, 2017, once again another cargo train accidents occurred when twelve (12) cargo trains travelling southward between the National Bank Station and Kuala Lumpur Station slipped due to heavy weight and oversized loads carried by the cargo trains. As a result, KTM and ETS services were disrupted on several routes around the Klang Valley. One of the major causes of this tragedy is

ISSN: 2622-5778 (online)
Advances in Transportation and Logistics Research
http://proceedings.itltrisakti.ac.id/index.php/altr

the overloading of the cargo train's wagon (Bernama, 2017). In recent accident that occurred on 21 July 2019 cargo train that carried 30 wagons of cement. During the derailment, KTMB needed to relocate all the wagons as soon as possible because all the KTMB's services were affected (Hussin, 2019). Having the amount of carried weight planned to match the track capability can avoid derailment occurrences.

It is very important to reduce the derailment occurrence in the developing country like Malaysia. Injuries not only the result of the derailment but indirectly lead to the waste of social wealth. In fact, on 2018, KTMB recorded losses at almost RM 200, 000 to repair their asset along 1,655 kilometer of KTMB's track (Daim, 2019). Therefore, the cargo train safety should be improved by analysing data that can possibly contribute to the derailment.

It is a common situation to encounter a dataset with the entries of the missing value problem in database. Missing value occurs when no value or data is stored for the variable in an observation. The missing data can occur during the data collection is done improperly or mistake are made in data entry process. The existence of missing value can influence the significant effect on the validity of the conclusion or not provide the accurate results at the end of a research. Therefore, methods for imputing missing value are needed in order to minimize the effect of incomplete dataset in an analysis (Troyanskaya et al., 2000, Patil et al., 2010: using Expectation-Maximization algorithm by Miyahari et al 2020; using Probable values, Peng and Lei, 2005).

Fortunately, missing data is not a big deal for the whole data, especially if it only a small presence of missing data such as 1% of the missing data. Besides, Schafer (1999) claimed that 5% rate of missing value is still negligible and maximum upper threshold for larger datasets. However, Bennet (2001) stated that if the data are missing more than 15%, then the statistical analysis appears to be biased. Meanwhile, statistical guidance articles have stated that biased is likely to occur in analysed with a lack of more than 10% and that if more than 40% of data is missing in important variables, results should be considered as hypothesis only (Dowd et al., 2019; van Ginkel et al, 2020).

Method

Data on nine (9) variables obtained are used in this study, carried weight is the dependent or target variable. While, the other eight (8) variables which are *Date*, *Distance*, *Company*, *Route, Total wagon*, *Labour, Train no.* and *Tonnage/KM* is assumed to be the input variables. The selected input variables are based on past literature (Rahman et al., 2014; Ghoushchi and Rahman, 2016) and the nature of amount of carried weight prediction.

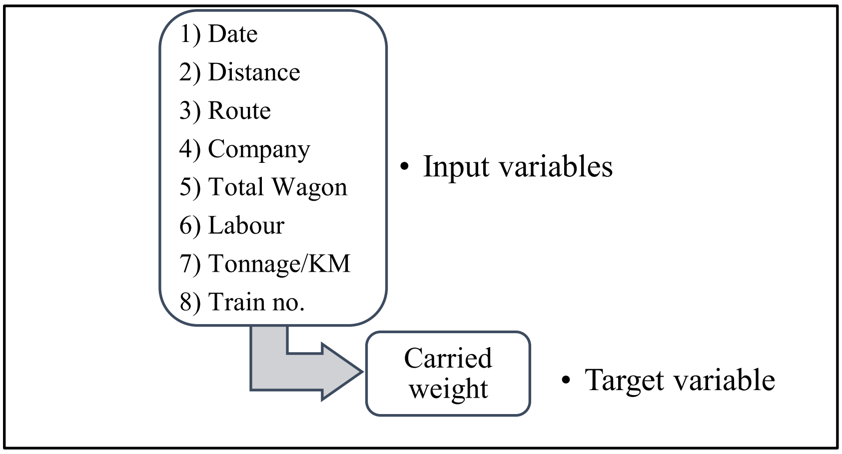


Figure 1: Research Framework (Source: Rahman et al., 2014)

ISSN: 2622-5778 (online)
Advances in Transportation and Logistics Research
http://proceedings.itltrisakti.ac.id/index.php/altr

KTM Cargo provided two (2) types of service which are containerized cargo and conventional cargo as shown in Figure 2. However, as mentioned earlier, this study focused on conventional cargo service that carried cement as a type of commodity. KTMB provided daily data of cargo train to carry cement for three (3) years from 1st January 2016 until 31st December 2018. There are 1096 observation for each route. Therefore, for twelve (12) routes total observation is 13,152.

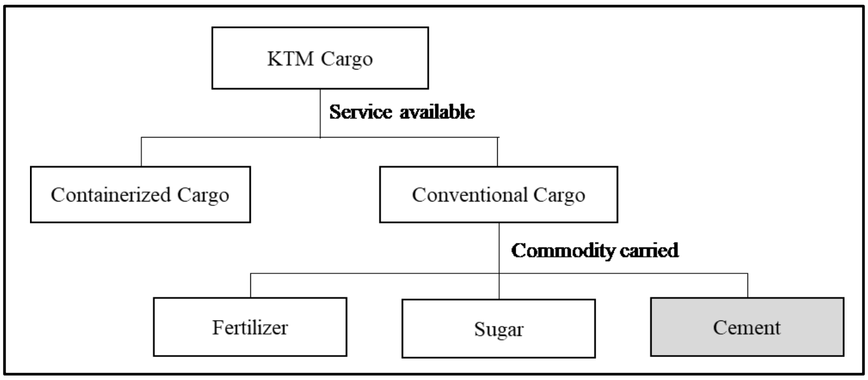


Figure 2. Services Provided by KTMB

There are twelve (12) operating routes: Padang Rengas → Gelang Patah, Tasek → Gelang Patah, Kuang → Tasek, Tasek → Kuang, Kuang → Bukit Ketri, Bukit Ketri → Kuang, Kanthan → Padang Jawa, Padang Jawa → Kathan, Batu Caves → Padang Rengas, Padang Rengas → Batu caves, Gelang Patah → Padang Rengas and Gelang Patah → Tasek. The maps of these routes are illustrated in Figure 3.

The data is a secondary data, focusing on conventional cargo train service that carried cement only in twelve (12) routes. Table 1 summarizes each of the variables used and its detail descriptions. There are missing values occurred in all six (6) routes. For Route 1, 1009 data valid with 8% missing value and Route 2 has 934 valid data with 15% of missing value. From the table, missing values for Route 3, Route 4, Route 5 and Route 6 are quite high with respectively only 686, 364, 126, and 32 valid data available for those routes. Missing values that are more than 15% are considered to be omitted in this study (Bennet, 2001). Therefore, Route 3, Route 4, Route 5 and Route 6 need to be omitted that leaves only two (2) route (Route 1 and Route 2) to be analysed further in this study.

ISSN: 2622-5778 (online)
Advances in Transportation and Logistics Research
http://proceedings.itltrisakti.ac.id/index.php/altr

Figure 3. KTMB Cargo Route Map

The data was stored in Excel files. Then, IBM SPSS Modeler 18.0 software is used to observe the categorical and continuous variables. The variables that have the missing value problem have been identified and ready for treatments. There are three (3) continuous and four (4) categorical variables. For both Route 1 and Route 2, there is no missing value in *Company*, *Train no*. and *Distance* variables. Thus, these categorical variables can be used for further analysis. The valid data for those variables was 1096 observations. However, only one (1) categorical variable has missing value which is *Labour* with only 922 and 844 valid data in Route 1 and Route 2 respectively. While, there are missing values for all continuous variable which are *Total Wagon*, *Tonnage/KM* and *Carried Weight* for both routes. In Route 1, valid data for *Total Wagon* is 905, valid data for *Tonnage/KM* is 1009 and valid data for *Carried Weight* is 1009. While in Route 2, valid data for *Total Wagon* is 900, valid data for *Tonnage/KM* is 934 and valid data for *Carried Weight* is 934. Therefore, imputation are needed for *Labour* and *Total Wagon*. However, *Tonnage/KM* will not undergo imputation process. For target variable, which is *Carried Weight,* all the cases with the missing value are discarded for both routes.

Table 1. Summary Statistics for All Variables before Imputation

Station		**Route 1**		**Route 2**	
Variable	Type	Valid	Missing	Valid	Missing
Train no.	Categorical	1096	0	1096	0
Company	Categorical	1096	0	1096	0
Distance	Categorical	1096	0	1096	0
Labour	Categorical	**922**	174	**844**	252
Total Wagon	Continuous	**905**	191	**900**	196
Tonnage/KM	Continuous	**1009**	87	**934**	162
Carried Weight	Continuous	**1009**	87	**934**	162

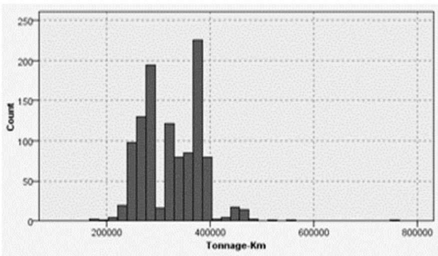
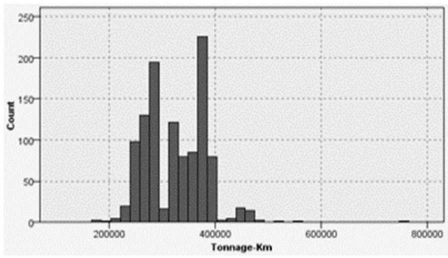
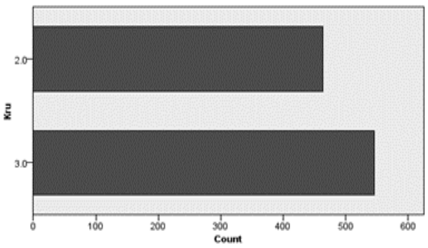
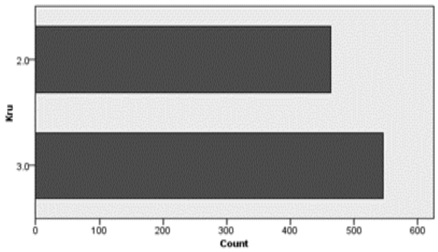
ISSN: 2622-5778 (online)
Advances in Transportation and Logistics Research
http://proceedings.itltrisakti.ac.id/index.php/altr

Discussion and Result

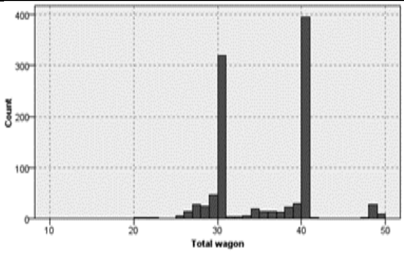
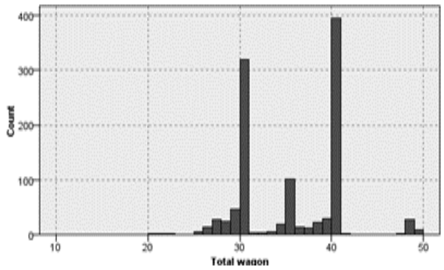
There is no imputation process happened for target variable, which is *Carried Weight*. However, even *Tonnage/KM* is not a target variable, but only when the carried weight exists tonnage per kilometre can be determined. Hence, all the cases with missing value at *Carried Weight* and *Tonnage/KM* are discarded which make only 1009 valid data for all the variables. Table 2 illustrates the continuous and categorical variables before and after imputation for Route 1. For *Labour* variable missing value has been replaced with its mode. For missing value in *Total Wagon* variable are replaced with its mean value which is 35.152 wagon. The standard deviation value for *Total Wagon* decreases from 5.808 to 5.573 and the skewness increase from 0.104 to 0.109.

The same exercise applied for *Carried Weight* and *Tonnage/KM* on Route 2 that results in only 934 valid data for all the variables. Table 3 illustrates the continuous and categorical variables before and after imputation for Route 2.

Table 2. Variables Before and After Data Discard for Route 1

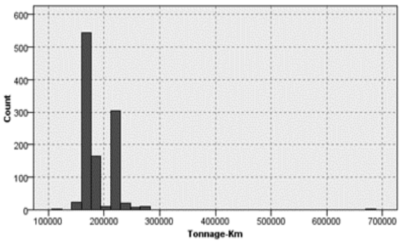
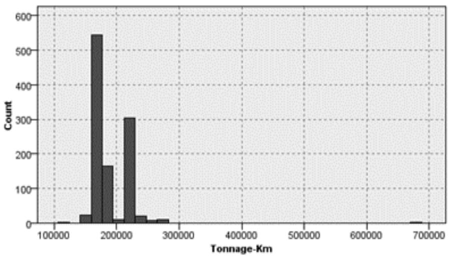
Tonnage/KM	*Tonnage/KM*
	
Min: 166834.16 Max: 764168.560 Mean: 326074.571 Std. Dev: 57504.670 Skewness: 0.583 Valid: 1009	Min: 166834.160 Max: 764168.560 Mean: 326074.571 Std. Dev: 57504.670 Skewness: 0.583 Valid: 1009
Labour	*Labour*
	
Min: 2 Max: 3 Valid: 922	Min: 2 Max: 3 Valid: 1009
Total Wagon	*Total Wagon*

ISSN: 2622-5778 (online)
Advances in Transportation and Logistics Research
http://proceedings.itltrisakti.ac.id/index.php/altr

	
Min: 17	Min: 17
Max: 50	Max: 50
Mean: 35.152	Mean: 35.152
Std. Dev: 5.808	Std. Dev: 5.573
Skewness: 0.104	Skewness: 0.109
Valid: 905	Valid: 1009

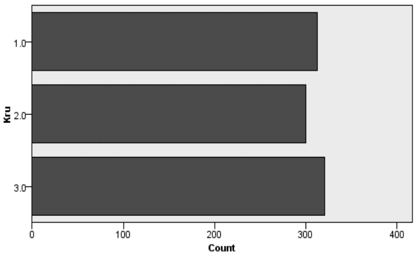
For *Labour* variable missing value has been replaced with its mode. For missing value in *Total Wagon* variable are replaced with its mean value which is 33.953 wagon. The standard deviation value for *Total Wagon* decrease from 5.659 to 5.224 and the skewness increase from 1.001 to 1.084.

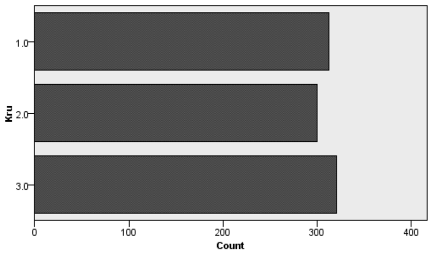
Table 3. Variables Before and After Imputation for Route 2

Before Imputation	**After Imputation**
Route 1	**Route 2**
Tonnage/KM	*Tonnage/KM*
	
Min: 106156.800	Min: 106156.800
Max: 688139.520	Max: 688139.520
Mean: 191879.066	Mean: 191879.066
Std. Dev: 38096.219	Std. Dev: 38096.219
Skewness: 6.150	Skewness: 6.150
Valid: 934	Valid: 934

ISSN: 2622-5778 (online)
Advances in Transportation and Logistics Research
http://proceedings.itltrisakti.ac.id/index.php/altr

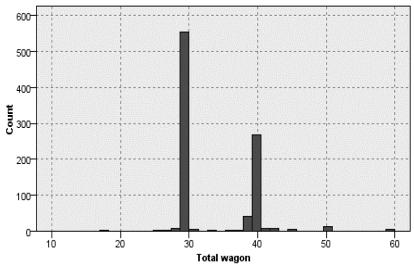
Labour | *Labour*

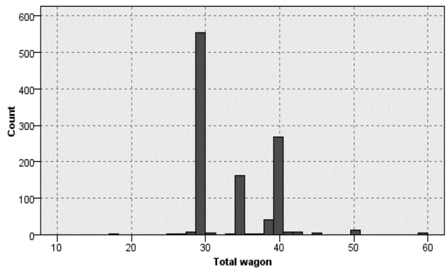




Min: 1	Min: 1
Max: 3	Max: 3
Valid: 844	Valid: 934

Total Wagon | *Total Wagon*





Min: 17	Min: 17
Max: 60	Max: 60
Mean: 33.953	Mean: 33.953
Std. Dev: 5.659	Std. Dev: 5.224
Skewness: 1.001	Skewness: 1.084
Valid: 900	Valid: 934

The amount of carried weight carried along the two (2) routes is not the same every day. Therefore, in Figure 4 and 5 show the line graphs to depict the trend or pattern of amount of carried weight for Route 1 and Route 2 from 2016 to 2018 respectively. The trend lines of the amount of carried weight are constructed and it can be concluded that there is a decrease in amount of carried weight of cargo train each day for both routes by 0.0691 and 0.1178 respectively, this is due to negative relationship. The equation are $y = -0.0691x + 4115.9$ and $y = -0.1178 x + 6052.1$.

Conclusion and Recommendation

There is definitely of strong importance to be able to predict the amount of carried weight for cargo service. Most of the train services providers have been collecting massive amount of data for their daily operations. However, some data can be lost and missed due to several reasons such as black-outs and so on. This study shows the process of handling missing data for cargo services data for a specific case study, which ends with a complete data that can be usable for predicting the data for the upcoming years.

ISSN: 2622-5778 (online)
Advances in Transportation and Logistics Research
http://proceedings.itltrisakti.ac.id/index.php/altr

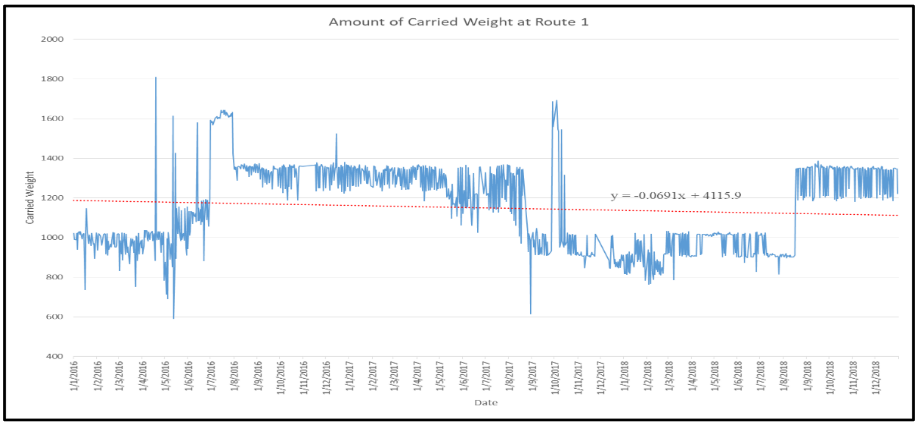


Figure 0. Trend Line of Amount of Carried Weight for Route 1

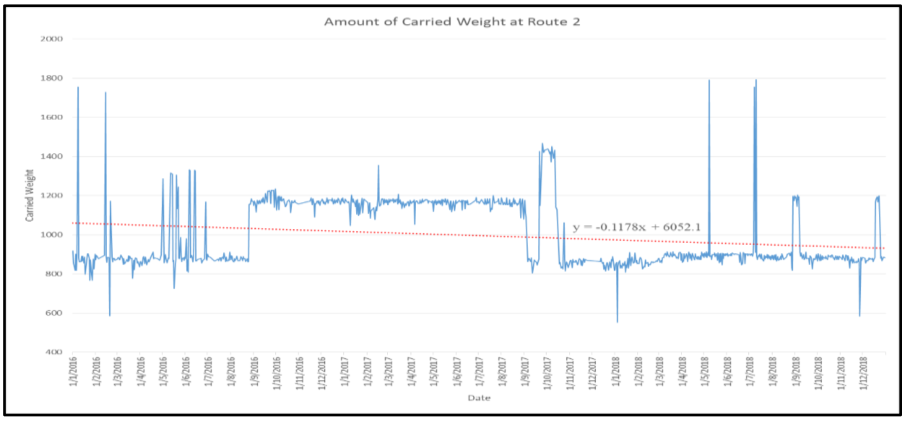


Figure 5: Trend Line of Amount of Carried Weight for Route 2

Acknowledgement

We would like to acknowledge KTMB Kargo Sdn Bhd for data compilation as well as Malaysia Institute of Transport (MITRANS), UiTM Shah Alam for funding this presentation.

References

Bennett, D. A. (2001). How can I deal with missing data in my study? *Australian And New Zealand Journal of Public Health*, 25(5), 464-469.

Bernama. (2017, April 21). KTMB rekod 17 kemalangan maut babitkan pelanggaran tren pada 2018. Asto Awani, Retrieved http://www.astroawani.com/berita-malaysia/ktmb-rekod-17-kemalangan-maut-babitkan-pelanggaran-tren-pada2018 -205358.

Bernama. (2017, November 23). The derailment disrupts KTM rail service. Malaymail, Retrieved https://www.malaymail.com/news/malaysia/2017/11/23/trainderailmen - disrupts-ktm-rail-services/1516919.

Daim, N. (2019, April 20). Keep off the rail tracks, KTMB warn trespassers. The News Straits Times. Retrieved https://www.nst.com.my/news/nation/2019/04/481316 /keep-rail-tracks-ktmb-warn-trespassers.

ISSN: 2622-5778 (online)
Advances in Transportation and Logistics Research
http://proceedings.itltrisakti.ac.id/index.php/altr

Ghoushchi, S. J., & Rahman, M. N. A. (2016). Performance study of artificial neural network modelling to predict carried weight in the transportation system. *International Journal of Logistics Systems and Management*, 24(2), 200.

Hussin, M. H. (2019, July 18). Tren dijangka beroperasi semula Isnin. Mymetro, Retrieved https://www.hmetro.com.my/mutakhir/2019/07/476835/tren-dijangka-beroperasi-semula-isnin.

KTMB (2019). Network Service. Retrieved May 23, 2019, from http://www.intranet4.ktmb.com.my/ktmb/index.php.

Madley-Dowd, P., Hughes, R., Tilling, K., & Heron, J. (2019). The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of clinical epidemiology*, *110*, 63-73.

Miyahara, H., Aihara, K., & Lechner, W. (2020). Quantum expectation-maximization algorithm. *Physical Review A*, *101*(1), 012326.

Noor, H. M. (2017, August 28). Semua 'tergelincir' apabila kereta api keluar landasan. Utusan Online, Retrieved https://www.utusan.com.my/rencana/utama/semua-tergelincir-apabila-kereta-api-keluar-landasan-1.519872

Patil, B. M., Joshi, R. C., & Toshniwal, D. (2010, August). Missing value imputation based on k-mean clustering with weighted distance. In *International Conference on Contemporary Computing* (pp. 600-609). Springer, Berlin, Heidelberg.

Peng, L., & Lei, L. (2005). A review of missing data treatment methods. *Intelligent Information Management Systems and Technologies*, *1*(3), 412-419.

Rahman, M. N. A., Jafarzadeh-Ghoushchi, S., Wahab, D. A., & Jafarzadeh-Ghoushji, M. (2014). Artificial Neural Network Modeling Studies to Predict the Amount Of Carried Weight By Iran Khodro Transportation System. *Life Science Journal*, 11(SPEC. ISS. 2), 146-154.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical methods in medical research*, *8*(1), 3-15.

van Ginkel, J. R., Linting, M., Rippe, R. C., & van der Voort, A. (2020). Rebutting existing misconceptions about multiple imputation as a method for handling missing data. *Journal of Personality Assessment*, *102*(3), 297-308.