Forecasting the Number of Jabodetabek Train Passengers Using ARIMA

Moerpradighta Prayreyka[1], Edwin Setiawan Nugraha[2], Mokhammad Ridwan Yudhanegara[3]

[1,2]President University, North Cikarang, Bekasi, 17550, Indonesia

[3]Mathematics Education Department

email: moerpradighta.prayreyka@student.president.ac.id[1], edwin.nugraha@president.ac.id[2], mridwan.yudhanegara@staff.unsiska.ac.id[3]

Abstrak

This study aims to analyze a suitable forecasting model using ARIMA to help PT. KAI Indonesia in predicting the number of train passengers in Jabodetabek. This study uses the method of identifying forecasting patterns. Model selection is very important in forecasting because forecasting models are beneficial for forecasting using past data in the past. The sample used is the number of Jabodetabek train passengers from January 2014 to December 2016. The results show that the suitable forecasting method to predict the number of Jabodetabek train passengers is the ARIMA method (3,1,6). The results from this analysis can be used for considering to calculate operational costs and business development in the future.

Keywords: ARIMA, Forecasting, Jabodetabek Train Passenger, Time Series.Analysis

INTRODUCTION

Jabodetabek is a metropolitan area consisting of Jakarta, Bogor, Depok, Tangerang, Bekasi. The people in Jabodetabek, in carrying out their daily activities, sometimes need to go from one place to another. Various public transportation facilities, including trains and buses, are available to facilitate this. The train is a practical transportation choice because it will not be affected by traffic jams in Jabodetabek, and the price is economical. This has encouraged many Jabodetabek people to choose trains as the most desirable transportation.

The number of train passengers in Jabodetabek is a reference that can be used to conduct research using the Box-Jenkins Time Series forecasting method, especially the Auto-Regressive Integrated Moving Average (ARIMA) model. Many previous studies used ARIMA model to forecast the number of train passenger. In (Ria & Indrasetianingsih, 2016), forecasting the number of train passenger using ARIMA method. The results show that the best time series model analysis for forecasting the number of Java train passenger is ARIMA model (1,1,0) $(0,1,1)^{12}$, because it has smallest RMSE value dan the MAPE value under 10% which is 9.8% compared with the other model. Also, in (Hidayat, 2019), analyzing forecast the number of Penataran train passenger using ARIMA and Exponential Smoothing. The results show that forecasting with ARIMA can produce the highest forecasting value if compared with Exponential Smoothing-Winter Method, and the number of passengers in previous years. Arima Box Jenkins Method is more suitable to to determining the number of passengers in the future, because the accuracy value is smaller compared with Exponential Smoothing-Winter Method.

Forecasting is an important tool in effective and efficient planning. Perspectives on forecasting may be as diverse as those of any other group of scientific methods. An institution always sets goals and objectives, tries to estimate environmental factors, then determines the actions that are expected to achieve these goals and objectives. This study discusses the forecasting model for the number of Jabodetabek train passengers in 2017 using sample data on the number of Jabodetabek train passengers from January 2014 to December 2016.

This forecasting aims to take strategic steps that need to be done. The choice of the forecasting model is also very important because each type of data has a different model. Based on these conditions, the problem in this study is which forecasting model is most suitable for forecasting data on the number of Jabodetabek train passengers.

METHOD

A. *Introduction of Time Series Models*

The time series method is a forecasting method that analyzes the relationship pattern between the variables to be estimated and the time variable. Forecasting a time series data needs to pay attention to the type or pattern of data. There are four types of time series data patterns, namely horizontal, trend, seasonal, and cyclical (Hanke & Wichers, 2005). Horizontal patterns are unexpected and random events, but their occurrence can affect fluctuations in time series data. The trend pattern is the tendency of the direction of the data in the long term, and it can be in the form of an increase or decrease. Seasonal patterns are fluctuations in data that occur periodically within one year, such as quarterly, quarterly, monthly, weekly, or daily. While the cyclical pattern is a fluctuation of the data for more than one year (Lisnawati, 2012). The method that is often used is the Box-Jenkins ARIMA method which is used to process univariate time series, and the transfer function analysis method is used to process time-series data. Multivariate. To be processed using the Box-Jenkins ARIMA method, a time series data must meet the stationarity requirements (Makridakis, 1999).

B. *Stochastic Process*

The presentation of this section refers to (Cryer & Chan, 2008). In mathematics, especially in probability theory and statistics, a stochastic process is a collection of random variables X_t where t is a parameter of a set of indices (usually corresponding to a discrete-time unit with the set of indexes $\{1, 2, \dots\}$). The stochastic process is one way to quantify the relationship between a set of random events. Therefore, stochastic processes are often used to model a system that changes randomly over time, such as in finance, biology, and others. A stochastic process is generally denoted as $\{X_t\}_{t \in T}$ or $\{X_t\}$. There are several ways to classify a stochastic process, for example, by using the cardinality of its index set and the conditioned space. When the set of indices is interpreted as time and has a finite or calculated cardinality (for example, the set of natural numbers), we call it a discrete-time stochastic process. If the set of indices is an interval of real numbers, we call it a continuous-time stochastic process. There are two examples of stochastic processes:

- Bernoulli Process
  The Bernoulli process is one of the simplest stochastic processes. This process is a collection of identically distributed independent random variables (iid) with a value of 0 or 1 with probabilities p and $1 - p$, respectively. This process can be associated with repeatedly tossing a coin (which may be unfair).

- Markov Process
  A Markov process is a stochastic process that satisfies the Markov condition. Given the situation at the current time, the probability of an event in the future is not affected by additional information regarding past behaviour. Formally,

$$\Pr\{X_{n+1} = j | X_0 = i_0, \ldots, X_{n-1} = i_{n-1}, X_n = i\} = \Pr\{X_{n+1} = j | X_1\} \quad [1]$$

C. *Stationary*

The presentation of this section refers to (Cryer & Chan, 2008). To make statistical conclusions about the structure of a stochastic process based on the observed records, we usually have to make some simplifying (and possibly reasonable) assumptions about that structure. The most important assumption is the assumption of stationarity. The basic idea of stationarity is that the laws of probability that govern the behaviors of processes do not change over time. In a sense, the process is in statistical equilibrium. Specifically, a process $\{Y_t\}$ is said to be completely stationary if the joint distribution $Y_{t_1}, Y_{t_2}, \ldots, Y_{t_n}$ equals the shared distribution $Y_{t_1-k}, Y_{t_2-k}, \ldots, Y_{t_n-k}$ for all time point options t_1 , $t_2, \ldots, t_n$ and all k time lag options.

Thus, when $n = 1$, the (univariate) distribution of Y_t equals the distribution of Y_{t-k} for all t and k; in other words, Y (slightly) is identically distributed. It then follows that $E(Y_t) = E(Y_{t-k})$ for all t and k so that the average function is constant for all time. Moreover, $Var\,(Y_t) = Var\,(Y_{t-k})$ for all t and k so that the variance is also constant over time.

Setting $n = 2$ in the definition of stationarity we see that the bivariate distribution of Y_t and Y_s must equal Y_{t-k} and Y_{s-k} so that $Cov(Y_t, Y_s) = Cov(Y_{t-k}, Y_{s-k})$ for all $t, s,$ and k. Putting $k = s$ and then $k = t$, we get

$$\begin{aligned} \gamma_{t,s} &= Cov(Y_{t-s}, Y_0) \\ &= Cov(Y_0, Y_{s-t}) \\ &= Cov\left(Y_0, Y_{|t-s|}\right) \\ &= \gamma_{0,|t-s|} \end{aligned} \quad [2]$$

The covariance between Yt and Ys is time-dependent only through the time difference |t−s| and not vice versa at the actual time t and s. So, for a stationary process, we can simplify the notation and write it

$$\gamma_k = Cov(Y_t, Y_{t-k}) \qquad \text{and} \qquad \rho_k = Corr(Y_t, Y_{t-k}) \quad [3]$$

also note that

$$\rho_k = \frac{\gamma_k}{\gamma_0} \quad [4]$$

If a process is completely stationary and has a finite variance, the covariance function must depend only on the time lag.

D. *Auto Correlation Function (ACF) and Partial Autocorrelation Function (PACF)*

The presentation of this section refers to (Cryer & Chan, 2008). In the time series method, the primary way to identify a model from a data to forecast is to use the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF). According to (Wei, 2006), from the stationary process of time series data (X_t) obtained $E\ X_t = \mu$ and the variance of $Var\ X_t = E(X_t - \mu)^2 = \sigma^2$ which is constant, and the covariance $Cov(X_t, X_{t+k})$, whose function is only on the time difference $|t-(t-k)|$.

Then, the result can be written as an intermediate covariance X_t and X_(t+k) as follows:

$$\gamma_k = Cov\ X_t, X_{t+k} = E\ X_t - \mu\ X_{t+k} - \mu \quad [5]$$

And the correlation between X_t, X_{t+k}:

$$\rho_k = \frac{Cov(X_t, X_{t+k})}{Var(X_t)\ Var(X_{t+k})} = \frac{\gamma_k}{\gamma_0} \quad [6]$$

Where the notation $Var\ X_t = VarX_{t+k} = \gamma_0$. As a function of $k\ \gamma_k$, where the autocovariance and ρ_k functions are called autocorrelation functions (ACF), in the time series analysis γ_k and ρ_k describe the covariance and correlation between X_t and X_{t+k} from the same process, only separated from lag-k .

The sample of the autocovariance function and the sample of the autocorrelation function can be written as:

$$\gamma_k = \frac{1}{T} \begin{matrix} T-k \\ t=1 \end{matrix} X_t - X\ X_{t+k} - X \quad [7]$$

and

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{\begin{matrix} T-k \\ t=1 \end{matrix} X_t - X\ \ X_{t+k} - X}{X_t - X^2}, k = 0,1,2, \ldots. \quad [8]$$

with

$$X = \frac{1}{T} \begin{matrix} T \\ t=1 \end{matrix}\ X_t \quad [9]$$

The autocovariance function γ_k and the autocorrelation function ρ_k have the following characteristics:

$$\gamma_0 = Var\ X_t;\ \rho_0 = 1$$

1) $|\gamma_k \leq \gamma_0; |\rho_k \leq 1$ [10]
2) $\gamma_k = \gamma_{-k}$ and $\rho_k = \rho_{-k}$ for all k, γ_k and ρ_k in the function $lag\ k = 0$ are the same and symmetrical. This property is obtained from the time difference between X_t and X_{t+k}. Therefore, the autocorrelation function is often only plotted for non-negative lags. Such plots are sometimes called correlograms.

E. Model of Time-Series

The presentation of this section refers to (Cryer & Chan, 2008).

- Model Autoregressive or AR(p)

  AR(p) is the most basic linear model for stationary processes. This model can be interpreted as a process of regression results itself. Mathematically it is given by

$$X_t = \phi_0 + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + a_t \quad [11]$$

  where X_t is data at time *t*; *t* = 1, 2, 3, …, n. X_{t-i} is data at time *t-i*, *i* = 1, 2, 3, …, p, a_t is error on time *t*, ϕ_0 is a constant, ϕ_i is AR coefficient; *i* = 1, 2, 3, …, p

- Model Moving Average or MA(*q*)

  The general form of a q-level or MA(q) moving average model is defined as:

$$X_t = \theta_0 + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q} \quad [12]$$

  where X_t is data at time *t* with *t* = 1, 2, 3, …, n, a_{t-i} is error at time *t-i* with *i* = 1, 2, 3, …, q, θ_0 is a constant, θ_i is MA coefficient with *i* = 1, 2, 3, …, q.

- Model Autoregressive Moving Average or ARMA(*p*,*q*)

  This model is a combination of AR(p) and MA(q). It can be expressed as ARMA(p,q) with the general form:

$$X_t = \phi_0 + \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q} \quad [13]$$

  where X_t is data at time *t* with *t* = 1, 2, 3, …, n, $X_{t-i}$is data at time *t-i* with *i* = 1, 2, 3, …, p, a_{t-i} is error on period *t-i* with *i* = 1, 2, 3, …, q, θ_0 is a constant, $\phi_i$is AR coefficient with *i* = 1, 2, 3, …, p, θ_i is MA coefficient with *i* = 1, 2, 3, …, q.

- Model Autoregressive Integrated Moving Average or ARIMA(*p*,*d*,*q*)

  The ARMA(p,q) involving the differencing process with degree d will give us ARIMA(p,d,q). The general formula is written as follows:

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots + \phi_p Z_{t-p} + \cdots + dZ_{t-p-d}\varepsilon_t \quad [14]$$

The summary of all process to find the best ARIMA model is given by Figure 5.

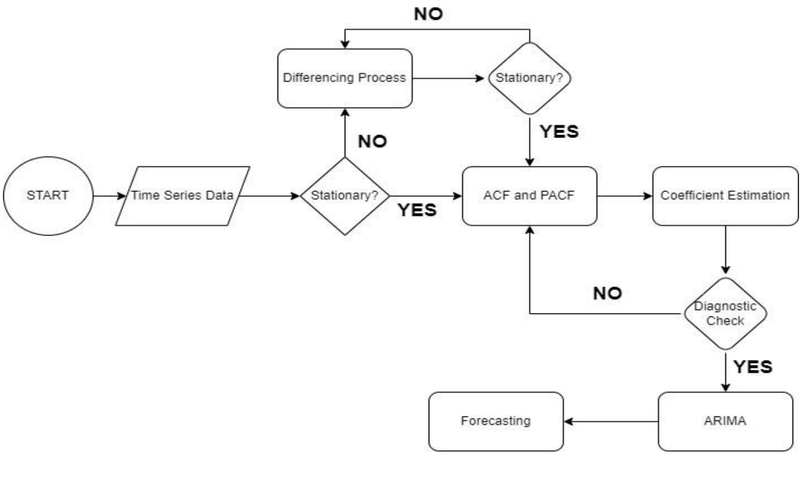

Figure 5. Box-Jenkins Method

RESULT AND DISCUSSION

A. *Data Preparation*

In this study, we use the secondary data of the number of of Jabodetabek train passengers of PT Kereta Api Indonesia from January 2014 until December 2016. This data is obtained from website of Badan Pusat Statistik Republik Indonesia (Badan Pusat Statistik, 2022) and presented at Table 1 below. All processing this data is helped by statistical software R.

Table 1. Number of Jabodetabek Train Passengers January 2014 – December 2016

Date	Passenger	Date	Passenger	Date	Passenger
01 January 2014	15176.00	01 April 2015	21171.00	01 August 2016	23923.00
01 February 2014	14856.00	01 May 2015	22177.00	01 September 2016	23570.00
01 March 2014	17471.00	01 June 2015	22207.00	01 October 2016	24533.00
01 April 2014	16671.00	01 July 2015	21171.00	01 November 2016	24104.00
01 May 2014	16781.00	01 August 2015	22295.00	01 December 2016	24841.00
01 June 2014	17848.00	01 September 2015	22021.00		
01 July 2014	16585.00	01 October 2015	22964.00		
01 August 2014	17091.00	01 November 2015	22355.00		
01 September 2014	18253.00	01 December 2015	22996.00		
01 October 2014	19079.00	01 January 2016	22238.00		
01 November 2014	18605.00	01 February 2016	21229.00		
01 December 2014	20080.00	01 March 2016	23206.00		
01 January 2015	19244.00	01 April 2016	23149.00		
01 February 2015	17640.00	01 May 2016	24401.00		
01 March 2015	21290.00	01 June 2016	23821.00		

B. *Stationary Check*

The plot of the data is presented by Figure 8. From the figure, we found that there is fluctuation and increasing over time. Figure 9 show the plot data after first differencing process. The next step, we need to stationarity checking by using the ADF test. The condition for stationarity of data is that the p-value is less than 0.05. The result of checking stationarity using R studio give us the p-value of ADF test is 0.5009, which means that this data is not stationary. The next step is differencing the data. After this process, we obtain p-value = 0.01. Since this value is less than 0.05, we conclude that the data is already stationary. From this, we also get that the degree of differencing is d=1.

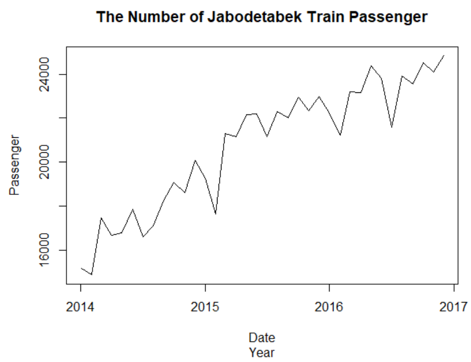

Figure 8. Plot Before Differencing

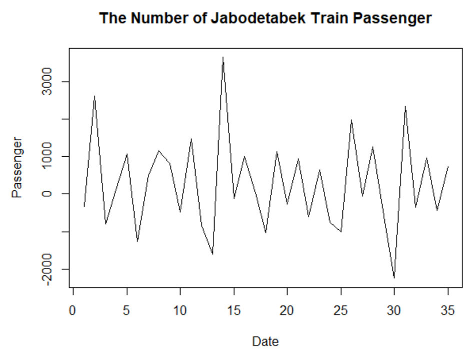

Figure 9. Plot After First Differencing

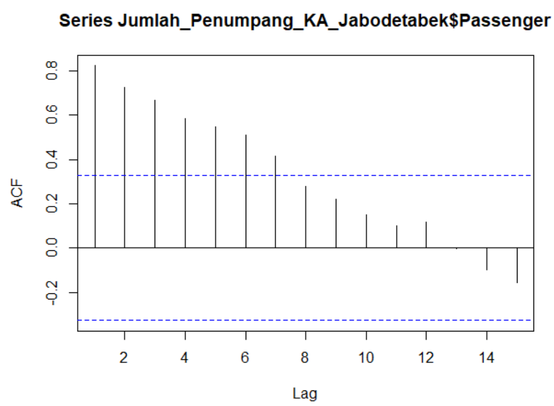

Figure 10. ACF Before Differencing

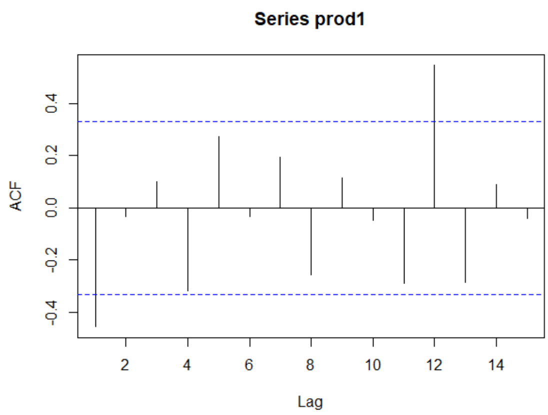

Figure 11. ACF After First Differencing

This is the P Plot (ACF). According to Figure 11, we can see that the numbers traversed by the blue dotted line are number 1 and number 12 or we can say it's P.

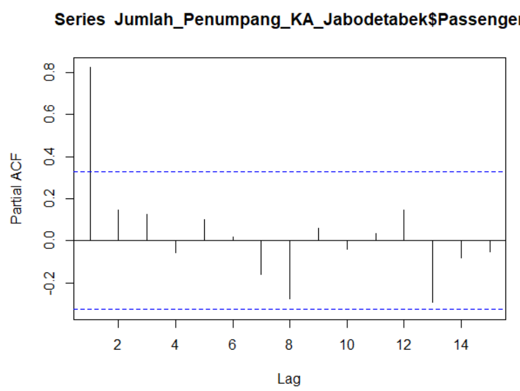

Figure 12. Before Differencing

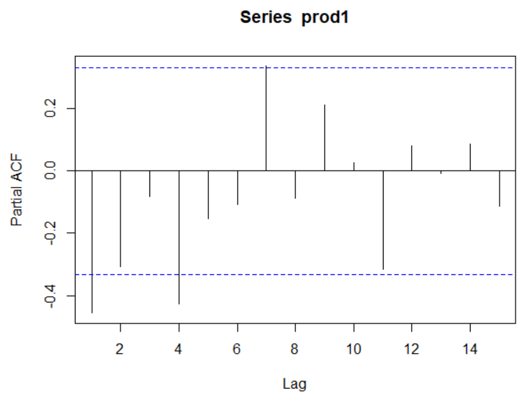

Figure 13. After First Differencing

This is the Q Plot (PACF). On figure 13, we can see that the numbers traversed by the blue dotted line are 1, 4, and 7, and for this we can say it's Q.

C. *Model Spesification*

In this analysis, we choose (p, d, q) = (4, 1, 12). Table 2 below present some random numbers so that 35 different models are formed.

Table 2. ARIMA model data specifications

Model ARIMA	P	D	Q
ARIMA (4,1,12)	4	1	12
ARIMA (4,1,11)	4	1	11
ARIMA (4,1,10)	4	1	10
ARIMA (4,1,9)	4	1	9
ARIMA (4,1,8)	4	1	8
ARIMA (4,1,7)	4	1	7
ARIMA (4,1,6)	4	1	6
ARIMA (3,1,12)	3	1	12
ARIMA (3,1,11)	3	1	11
ARIMA (3,1,10)	3	1	10
ARIMA (3,1,9)	3	1	9
ARIMA (3,1,8)	3	1	8
ARIMA (3,1,7)	3	1	7
ARIMA (3,1,6)	3	1	6
ARIMA (2,1,12)	2	1	12
ARIMA (2,1,11)	2	1	11
ARIMA (2,1,10)	2	1	10
ARIMA (2,1,9)	2	1	9
ARIMA (2,1,8)	2	1	8
ARIMA (2,1,7)	2	1	7
ARIMA (2,1,6)	2	1	6
ARIMA (1,1,12)	1	1	12
ARIMA (1,1,11)	1	1	11
ARIMA (1,1,10)	1	1	10
ARIMA (1,1,9)	1	1	9
ARIMA (1,1,8)	1	1	8
ARIMA (1,1,7)	1	1	7
ARIMA (1,1,6)	1	1	6
ARIMA (0,1,12)	0	1	12
ARIMA (0,1,11)	0	1	11
ARIMA (0,1,10)	0	1	10
ARIMA (0,1,9)	0	1	9
ARIMA (0,1,8)	0	1	8
ARIMA (0,1,7)	0	1	7
ARIMA (0,1,6)	0	1	6

D. *Parameter Estimation*

The following are parameter estimates for all ARIMA models. After knowing the results of the ARIMA model, we can determine the estimated coefficients consisting of AR1, AR2, MA1, Mean Square Error (MSE), Log-likelihood, AIC, and MAPE, which will be considered for forecasting later. AR1, AR2, MA1, Log-Likelihood, and AIC were calculated using Rstudio. Meanwhile, MSE and MAPE were calculated using excel.

Table 3. ARIMA Parameter Estimation

MODEL ARIMA	COEFFICIENT OF ESTIMATION RESULT						
	AR1	**AR2**	**MA1**	**MSE**	**LOG LIKEHOOD**	**AIC**	**MAPE**
ARIMA (4,1,12)	-0.5335	-0.2907	-0.0259	4657594.167	-281.59	595,17	91.13%
ARIMA (4,1,11)	-0.2275	0.0672	-0.1969	3651702.783	-285.37	600,74	79.69%
ARIMA (4,1,10)	0.1480	-0.2051	-0.6455	2700658.822	-286.56	601,13	67.77%
ARIMA (4,1,9)	0.0788	-0.2113	-0.5772	2589202.631	-286.58	599,17	66.61%
ARIMA (4,1,8)	0.4084	0.6345	-0.9301	2050198.4	-288.1	600,2	60.98%
ARIMA (4,1,7)	-0.2973	0.6417	-0.2703	1731700.909	-288.23	598,46	56.61%
ARIMA (4,1,6)	0.6749	-0.0493	-1.2517	1726787.605	-288.24	596,47	56.29%
ARIMA (3,1,12)	-0.2229	0.0024	-0.2565	3028617,129	-283.88	597,76	75.18%
ARIMA (3,1,11)	0.0773	0.0671	-0.5129	4500508,813	-285.38	598,77	88.00%
ARIMA (3,1,10)	0.0316	0.2321	-0.4659	1778827,176	-285.74	597,48	57.64%
ARIMA (3,1,9)	0.0125	0.2481	0.7324	1869933,946	-285.76	595,53	59.38%
ARIMA (3,1,8)	0.3618	0.6101	-0.8812	2164273,908	-288.11	598,23	63.08%
ARIMA (3,1,7)	-0.0991	-0.4209	-0.3493	4360389,729	-289.83	599,67	88,75%
ARIMA (3,1,6)	0.6939	-0.0468	-1.2646	1739139,115	-288.25	594,49	56.61%
ARIMA (2,1,12)	-0.2189	0.0085	-0.2633	3084159,354	-283.88	595,77	75.91%
ARIMA (2,1,11)	-1.7098	-0.9747	1.6430	2757235,702	-284.12	594,23	72.75%

ARIMA (2,1,10)	-0.8425	-0.1827	0.3337	1821906,591	-287.27	598,54	57.24%
ARIMA (2,1,9)	0.0427	-0.0554	-0.6008	2020651,973	-288.41	598,81	59.18%
ARIMA (2,1,8)	0.5262	0.4722	-1.1087	1809428,287	-288.12	596,25	58.00%
ARIMA (2,1,7)	-0.6973	-0.1441	0.2499	4623771,887	-290	598	93.02%
ARIMA (2,1,6)	-0.4125	-0.2849	-0.0443	4630850,201	-289.99	595,98	92.15%
ARIMA (1,1,12)	-0.2206	-	-0.2645	3069722,254	-283.88	593,77	75.69%
ARIMA (1,1,11)	-0.6546	-	0.1683	1823115,846	-287.36	598,73	56.93%
ARIMA (1,1,10)	-0.7490	-	0.3142	1770402,224	-287.53	597,06	54.96%
ARIMA (1,1,9)	0.0434	-	-0.6008	1951288,328	-288.44	596,87	57.93%
ARIMA (1,1,8)	-0.9809	-	0.579	4714624,634	-290.28	598,56	93.69%
ARIMA (1,1,7)	-0.5662	-	0.1274	4659555,009	-290.04	596,08	93.52%
ARIMA (1,1,6)	-0.7738	-	0.4096	3942582,704	-290.55	595,1	85.21%
ARIMA (0,1,12)	-	-	-0.4726	2819883,809	-284.21	592,43	72.36%
ARIMA (0,1,11)	-	-	-0.5763	2937816,019	-288.61	599,21	69.64%
ARIMA (0,1,10)	-	-	-0.5536	1950539,218	-288.43	596,87	57.91%
ARIMA (0,1,9)	-	-	-0.5813	1974631,82	-288.45	594,91	58.43%
ARIMA (0,1,8)	-	-	-0.4554	3967501,543	-289.85	595,69	85.53%
ARIMA (0,1,7)	-	-	-0.4233	4764266,701	-290.32	594,65	93.85%
ARIMA (0,1,6)	-	-	-0.4494	4518725,384	-290.66	593,31	90.48%

E. *Residual Analysis*

In residual analysis, we can determine which model is best for our data using the Shapiro and Ljung's tests. The basic requirement needed to become the best model is to pass both tests, with the p-value criteria being more than 0.05.

Table 4. ARIMA Analysis Residual

Model ARIMA	**Shapiro Test p-value**	**Ljung-Box p-value**	**AIC**	**MAPE**
ARIMA (4,1,12)	0,8845	0,6526	595,17	91.13%
ARIMA (4,1,11)	0,3488	0,9468	600,74	79.69%
ARIMA (4,1,10)	0,9479	0,5204	601,13	67.77%
ARIMA (4,1,9)	0,9287	0,4906	599,17	66.61%
ARIMA (4,1,8)	0,9901	0,9854	600,2	60.98%
ARIMA (4,1,7)	0,8727	0,9445	598,46	56.61%
ARIMA (4,1,6)	0,8026	0,9647	596,47	56.29%
ARIMA (3,1,12)	0,7049	0,6345	597,76	75.18%
ARIMA (3,1,11)	0,4745	0,9497	598,77	88.00%
ARIMA (3,1,10)	0,5333	0,9467	597,48	57.64%
ARIMA (3,1,9)	0,5548	0,9024	595,53	59.38%
ARIMA (3,1,8)	0,995	0,995	598,23	63.08%
ARIMA (3,1,7)	0,4829	0,5287	599,67	88,75%
ARIMA (3,1,6)	0,8464	0,942	594,49	56.61%
ARIMA (2,1,12)	0,705	0,6388	595,77	75.91%
ARIMA (2,1,11)	0,9346	0,7208	594,23	72.75%
ARIMA (2,1,10)	0,7813	0,7316	598,54	57.24%
ARIMA (2,1,9)	0,6465	0,6599	598,81	59.18%
ARIMA (2,1,8)	0,907	0,9861	596,25	58.00%
ARIMA (2,1,7)	0,4895	0,5096	598	93.02%
ARIMA (2,1,6)	0,5435	0,5055	595,98	92.15%
ARIMA (1,1,12)	0,714	0,6404	593,77	75.69%
ARIMA (1,1,11)	0,8174	0,5982	598,73	56.93%
ARIMA (1,1,10)	0,8255	0,4033	597,06	54.96%
ARIMA (1,1,9)	0,6154	0,6743	596,87	57.93%
ARIMA (1,1,8)	0,6076	0,4316	598,56	93.69%
ARIMA (1,1,7)	0,5096	0,4747	596,08	93.52%
ARIMA (1,1,6)	0,745	0,312	595,1	85.21%
ARIMA (0,1,12)	0,5605	0,3405	592,43	72.36%
ARIMA (0,1,11)	0,4784	0,9591	599,21	69.64%
ARIMA (0,1,10)	0,6293	0,6583	596,87	57.91%
ARIMA (0,1,9)	0,5398	0,7855	594,91	58.43%
ARIMA (0,1,8)	0,4865	0,571	595,69	85.53%
ARIMA (0,1,7)	0,6366	0,413	594,65	93.85%
ARIMA (0,1,6)	0,69	0,63	593,31	90.48%

However, in this table, all data passed the Saphiro and Ljung-Box tests. In this case, we will then use the smallest MAPE value of the smallest AICs. The line that I marked in blue, or what we can call the ARIMA 14 model (3, 1, 6), is the best model. The model has equation

$$W_t = 0.6939_1 W_{t-1} - 0.0468_2 W_{t-2} + 0.3525_3 W_{t-3} + e_t + 1.2646_1 e_{t-1} - 0.3943_2 e_{t-2} + 0.2190_3 e_{t-3} + 0.3633_4 e_{t-4} - 1.1709_5 e_{t-5} + 0.7062_6 e_{t-6} \quad [15]$$

where $W_t = Y_t - 3Y_{t-1} + 3Y_{t-2} - Y_{t-3}$

F. *Error of The Forecasting*

In this part, we will calculate the error of the forecasting from the best model ARIMA (3,1,6). Here we define the error, squared error, percentage, MSE, RMSE, MAE, and MAPE.

Table 5. Calculate Error Best Model

Point Forecast	**Actual Data**	**Error**	**Error (Kuadrat)**	**Percentage**
25326,72	24185	1303524,558	1141,72	4,72%
24895,32	21743	9937121,382	3152,32	14,50%
25902,03	25775	16136,6209	127,03	0,49%
24998,14	25411	170453,3796	412,86	1,62%
26130,75	27385	1573143,063	1254,25	4,58%
26499,42	24432	4274225,456	2067,42	8,46%
26383,6	27016	399929,76	632,4	2,34%
26685,18	27679	987678,1924	993,82	3,59%
27029,83	26158	760087,5489	871,83	3,33%
27214,02	28765	2405538,96	1550,98	5,39%
27431,99	28246	662612,2801	814,01	2,88%
27696,1	29059	1857496,41	1362,9	4,69%

After calculation we obtain of the MSE, RMSE, MAE, and MAPE are 1739139,115, 1318,764238, and 56,61%, respectively.

G. *Forecasting*

Then we enter into the forecasting stage. What will be predicted in this experiment is the number of Jabodetabek train passengers from January 2017 to December 2017 using ARIMA (3, 1, 6) and a confidence interval of 95%.

Table 9. Forecasting Data Januari – Desember 2017

Date	Actual Data	Point Forecast	Lower Limit	Upper Limit
01 Januari 2017	24185	25326.72	23413.72	27239.73
01 Februari 2017	21743	24895.32	22810.02	26980.62
01 Maret 2017	25775	25902.03	23709.41	28094.65
01 April 2017	25411	24998.14	22582.57	27413.71
01 Mei 2017	27385	26130.75	23714.85	28546.65
01 Juni 2017	24432	26499.42	23586.82	29412.03
01 Juli 2017	27016	26383.60	23096.04	29671.16
01 Agustus 2017	27679	26685.18	23224.97	30145.39
01 September 2017	26158	27029.83	23334.68	30724.97
01 Oktober 2017	28765	27214.02	23251.52	31176.52
01 November 2017	28246	27431.99	23244.23	31619.75
01 Desember 2017	29059	27696.10	23289.30	32102.89

Figure 18 shows the results of forecasting the number of Jabodetabek train passengers in every month in 2017. The black line shows the sample data used to perform this analysis. The red line shows the graph of ARIMA used for forecasting, while the blue line shows the data prediction.

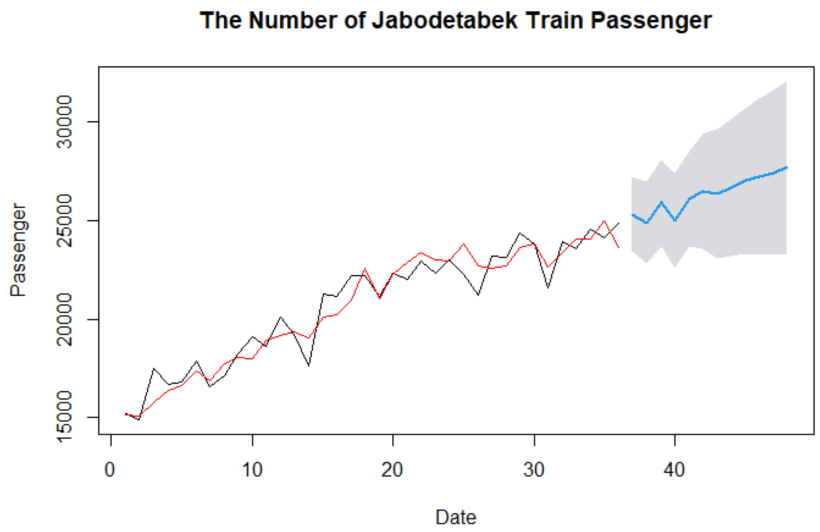

Figure 18. Forecasting of the Passengers in 2017

Conclusion
In the work, we have forecast the the number of of Jabodetabek train passengers of PT Kereta Api Indonesia from January 2017until December 2017. The analysis result show that ARIMA (3,1,6) is the best model The comparison between the predicted data and the actual data shows that the predicted data is not much different from the actual data. The total estimated number of passengers in 2017 is 316193 people. Meanwhile, according to actual data, the total number of passengers in 2017 was 315854. Therefore, forecasting using the ARIMA method can be pretty accurate and effective when correctly choosing the best ARIMA model. This analysis is expected to be helpful in knowing the number of passengers on trains in Jabodetabek. Based on these predictions, PT Kereta Api Indonesia is able to prepare and anticipate if there is a surge in passengers in the future. This will help the company to make a business plan to improve services to passengers.

For further research, we will forecast Jabodatabek train passengers using the latest data and consider residual variance as well. In the case of variance of residual exist, the methods used in forecasting are ARCH and GARCH.

References

Badan Pusat Statistik. (2022). *Jumlah Penumpang Kereta Api 2014-2016*. Diambil kembali dari https://www.bps.go.id/indicator/17/72/6/jumlah-penumpang-kereta-api.html

Cryer, J., & Chan, K.-S. (2008). *Time Series Analysis With Application in R.* Springer.

Hanke, J. E., & Wichers, D. W. (2005). *Business Forecasting Eight Edition.* New Jersey: Pearson Prentice hall.

Hidayat, R. (2019). Analisis Peramalan Jumlah Penumpang Kereta Api Penataran dengan Metode ARIMA Box Jenkins dan Exponential Smoothing. *Jurnal Ilmiah Mahasiswa Universitas Brawijaya*, 1-18.

Isfahani, H., & Listianti, L. (2019). *Peramalan Jumlah Penumpang Kereta Api.* Diambil kembali dari https://www.coursehero.com/file/57533769/Tugas-Arima-Penumoang-Kereta-Apidocx/

Lisnawati. (2012). *Model Exponential Smoothing Holt-Winter dan Model Sarima Untuk Peramalan Tingkat Hunian di Provinsi DIY.* Yogyakarta. Diambil kembali dari https://eprints.uny.ac.id/8326/3/BAB2-06305149010.pdf

Makridakis. (1999). *Metode dan Aplikasi Peramalan, Jilid 1.* Jakarta: Erlangga.

Nurhayati, Paramu, H., & Fadhli, M. R. (2014). *Forecasting Model Berbasis Data Time Series Pada Harga Saham Perusahaan Perbankan Yang Terpilih.* Diambil kembali dari https://repository.unej.ac.id/bitstream/handle/123456789/64319/Rizki%20Maulana%20Fadhli.pdf?sequence=1&isAllowed=y

Ria, M. L., & Indrasetianingsih, A. (2016). Prediksi Jumlah Penumpang Kereta Api dengan Menggunakan Metode ARIMA.

Wei, W. W. (2006). *Time Series Analysis: Univariate and Multivariate Methods Second Edition.* New Jersey: Pearson Prentice Hall.