GROUP 09

CPC351 PROJECT

SOLVING ANALYTICS PROBLEM

NAME	MATRIC NUM
Aliff Farhan	158607
Luqman Azri	158532
Muhammad Luqman	158629
Zahir Hariz	158176

JANUARY 2024

RIDERSHIP_HEADLINE

This dataset contains daily-frequency ridership data for various public transport services across the country (Malaysia)

DATASET 01

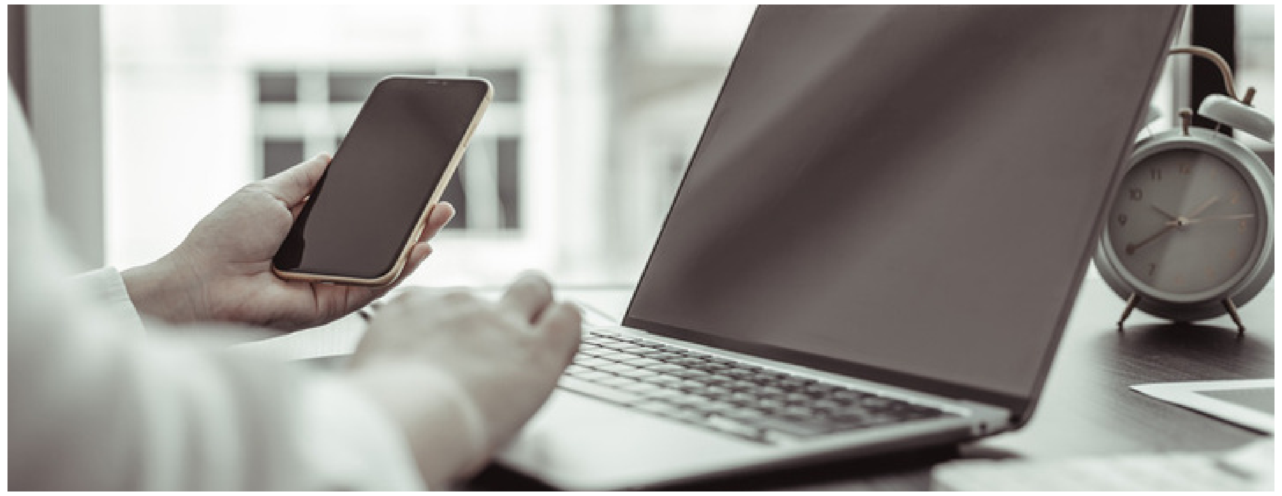
PROBLEM STATEMENT

- Lack of precise and current data regarding the utilization patterns of various public transportation modes can hinders the future planning and management capabilities.

- Making accurate predictions of ridership is essential for transportation agencies and strategists.

- Ridership predictions are crucial due to occurrence of challenges posed by external factors and shifting trends occurs.

OBJECTIVE



01 Provide insights on each type of public transport to be utilized by relevant organizations for future work

02 Build a predictive model for public transportation ridership using historical data.

03 Determine factors affecting quality of predictive model

INITIAL HYPOTHESIS

The most favorable type of transport is LRT followed by MRT, KTM.

The number of people using public transportation depends on a variety of factors, such as the time of day, special occasions, and external situations

DATA PREPARATION

Check Missing Value

Defines a function that prints the count of missing values for each column in the ride dataframe.

Check Missing Value

Defines a function that prints the count of missing values for each column in the ride dataframe.

Changing Date Data Type

Converts the 'date' column to the Date data type using the as.Date function.

Formatting Functions for Millions and Thousands

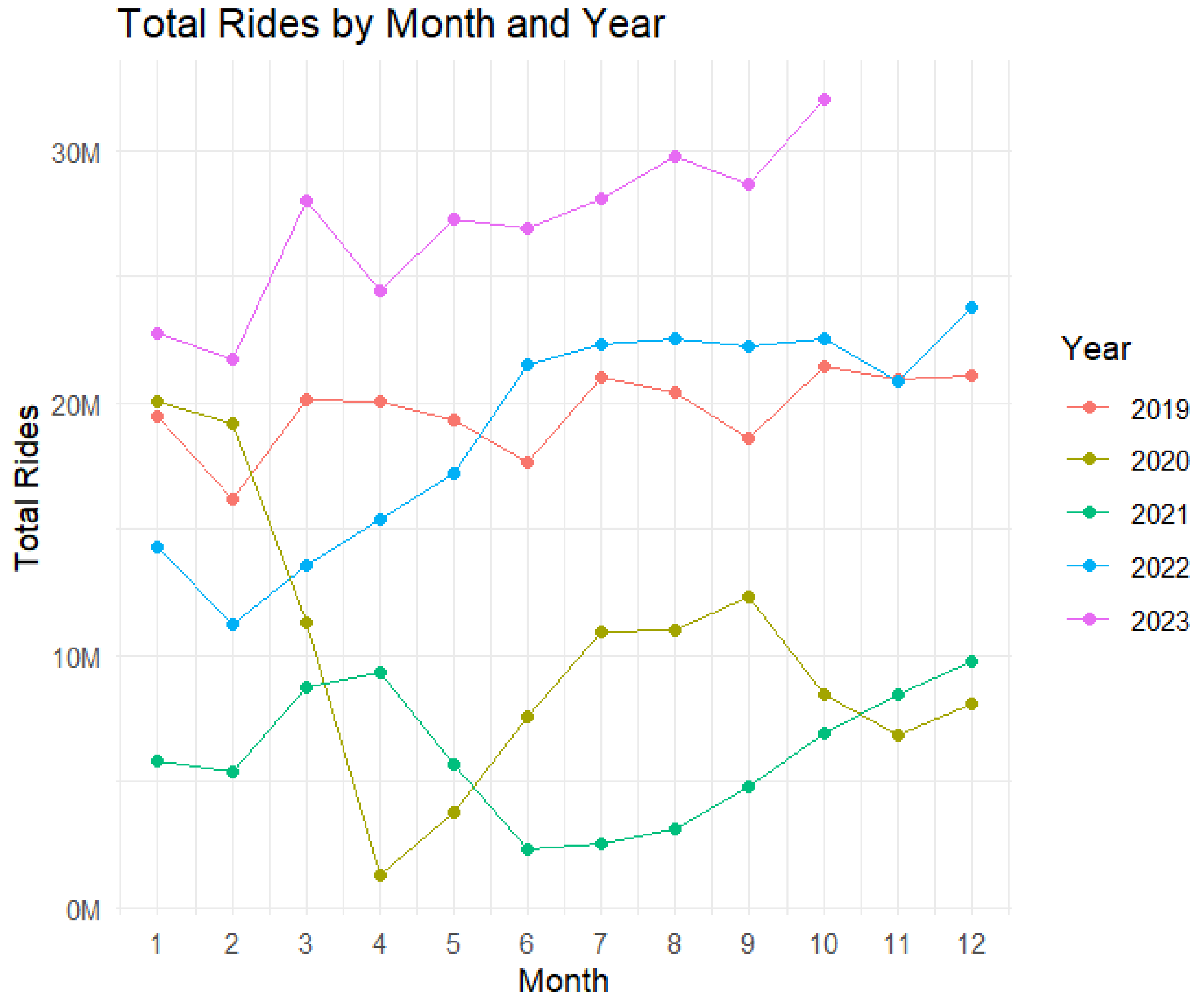
Defines two formatting functions (and to format numerical values in millions and thousands, respectively.

Adding Month and Year Columns

Adds three new columns to the ride dataframe extracted from the 'date' column for better data manipulation.

TOTAL RIDES

The graph gives us insights on the usage frequency of public transportation from year 2019-2023.

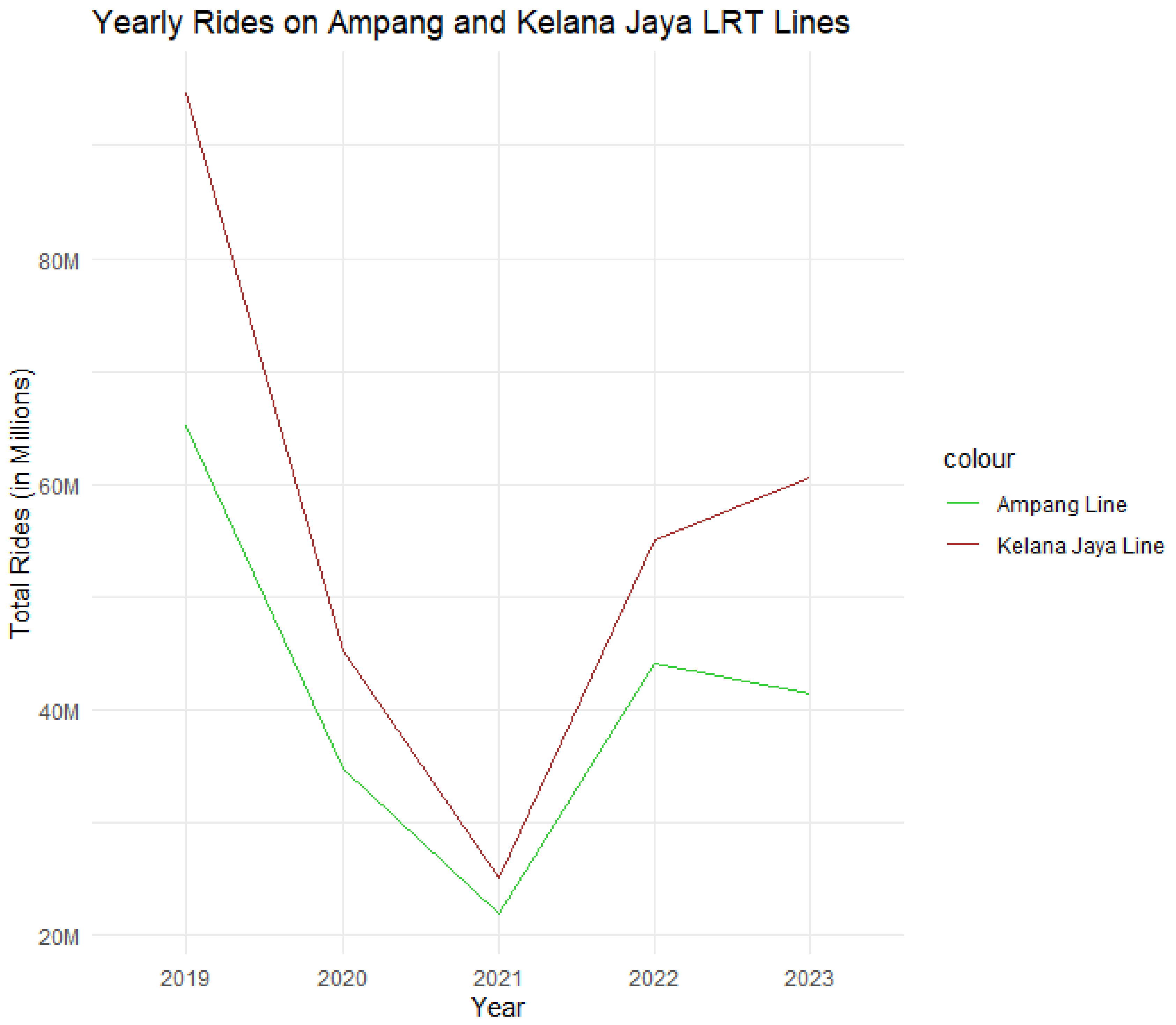


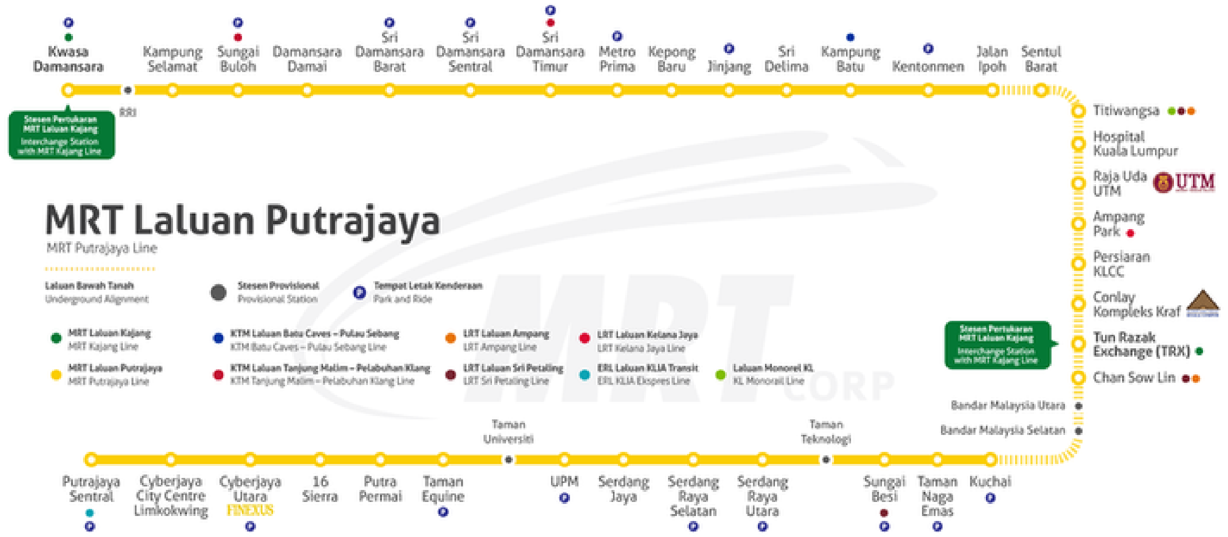
- The graph shows the comparison of total rides for each year starting from 2019 until 2023.
- Slight upward movement during end of every year.
- Major drop on early phase of year 2020.

LRT LINES

From the vizualiation, we can see that LRT Kelana Jaya is more favorable than LRT Ampang, Even after downfall of 2021, it manage to spike higher than Ampang Line. This is due to the number of stations provided by Kelana Jaya (37) is higher than Ampang (19).

Source : https://www.mrt.com.my/lrt_kelana/





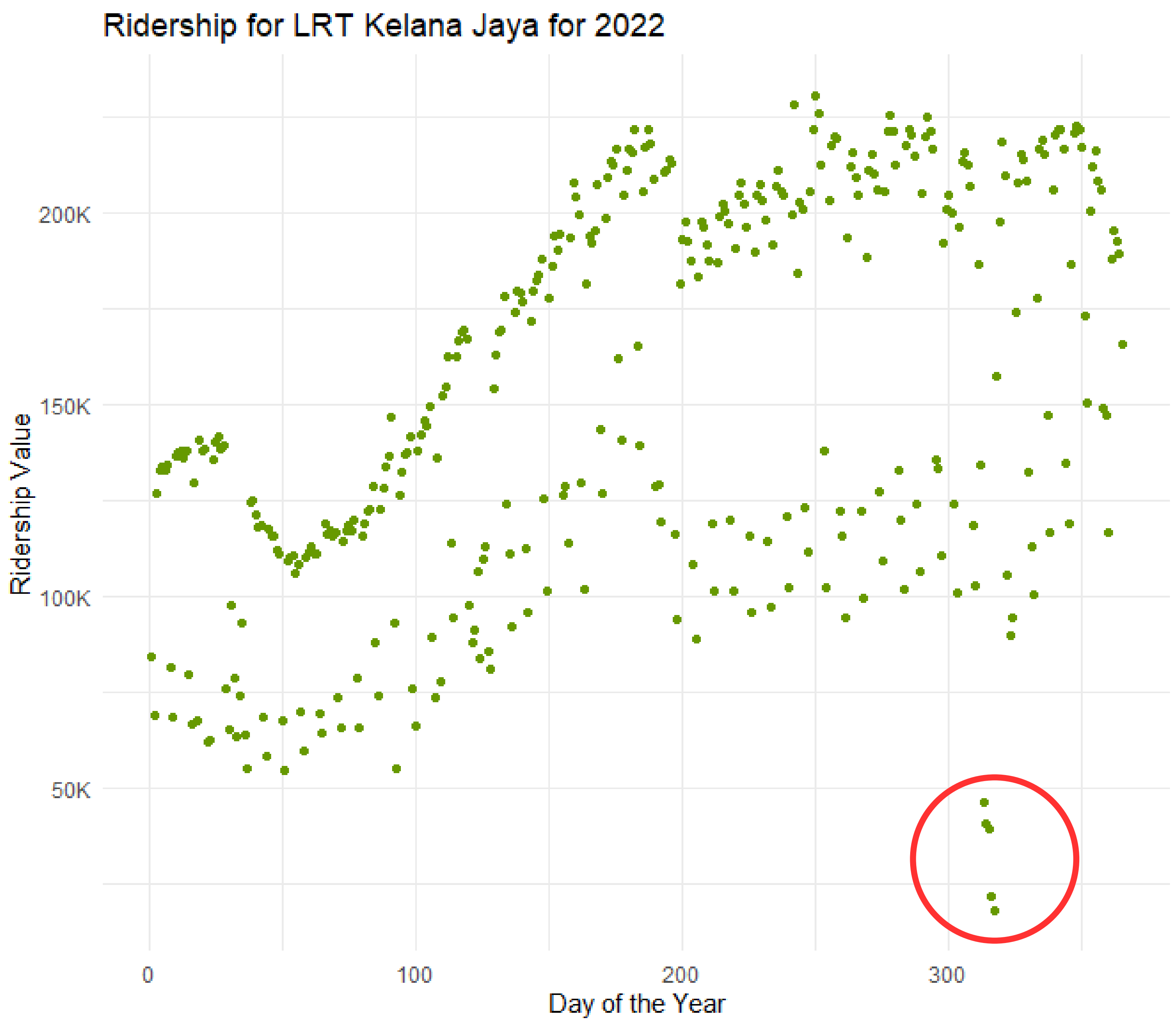
Starting at higher level, both of transportation line hit rock bottom at 2021.

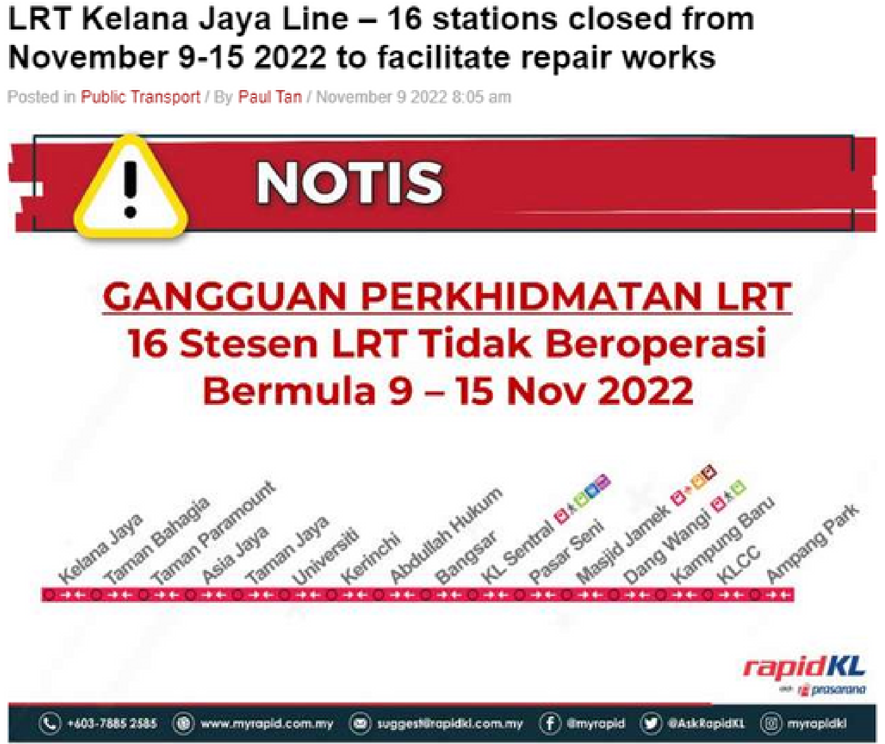
LRT Ampang face slight downfall from 2022 to 2023 due to Opening MRT Putrajaya on March 2023

LRT KELANA JAYA

It is detected the outliers are due to external factors which is 16 stations closed at Kelana Jaya line. This is due to maintenance of the stations

source : https://paultan.org/2022/11/09/lrt-kelana-jaya-line-16-stations-closed-from-november-9-15-2022-to-facilitate-repair-works/





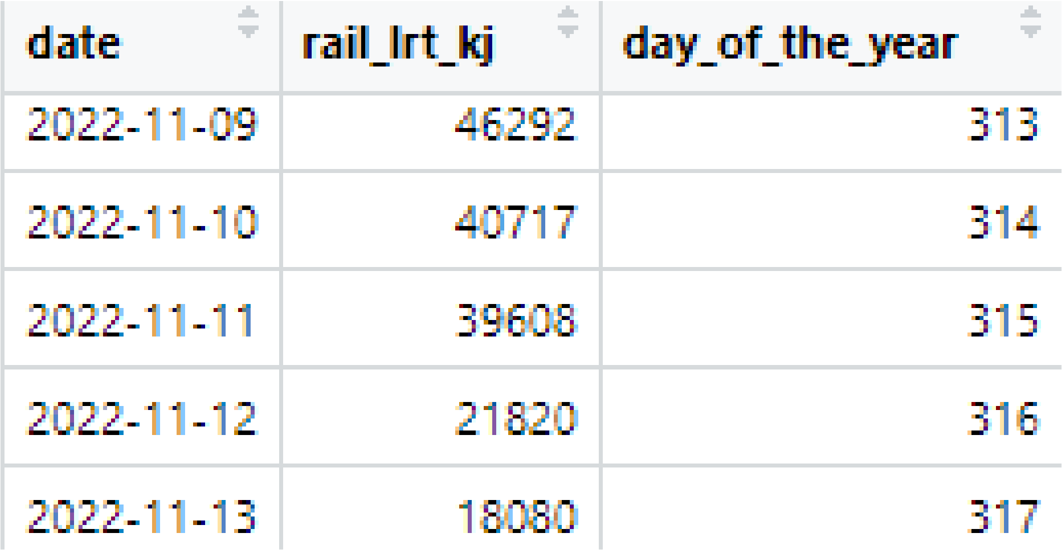
LRT Kelana Jaya Line – 16 stations closed from November 9-15 2022 to facilitate repair works

Posted in Public Transport / By Paul Tan / November 9 2022 8:05 am

NOTIS

GANGGUAN PERKHIDMATAN LRT
16 Stesen LRT Tidak Beroperasi
Bermula 9 – 15 Nov 2022

+603-7885 2585 www.myrapid.com.my suggest@rapidkl.com.my @myrapid @AskRapidKL myrapidkl



date	rail_lrt_kj	day_of_the_year
2022-11-09	46292	313
2022-11-10	40717	314
2022-11-11	39608	315
2022-11-12	21820	316
2022-11-13	18080	317

● The Scatter plot shows the amount of ride for each day throughout 2022

● Several Outliers detected during day 313 (9/11) until 317 (13/11)

MODEL PLANNING

BUILD A PREDICTIVE MODEL FOR PUBLIC TRANSPORTATION RIDERSHIP USING WHOLE AVAILABLE DATA

80% TRAIN

20% TEST

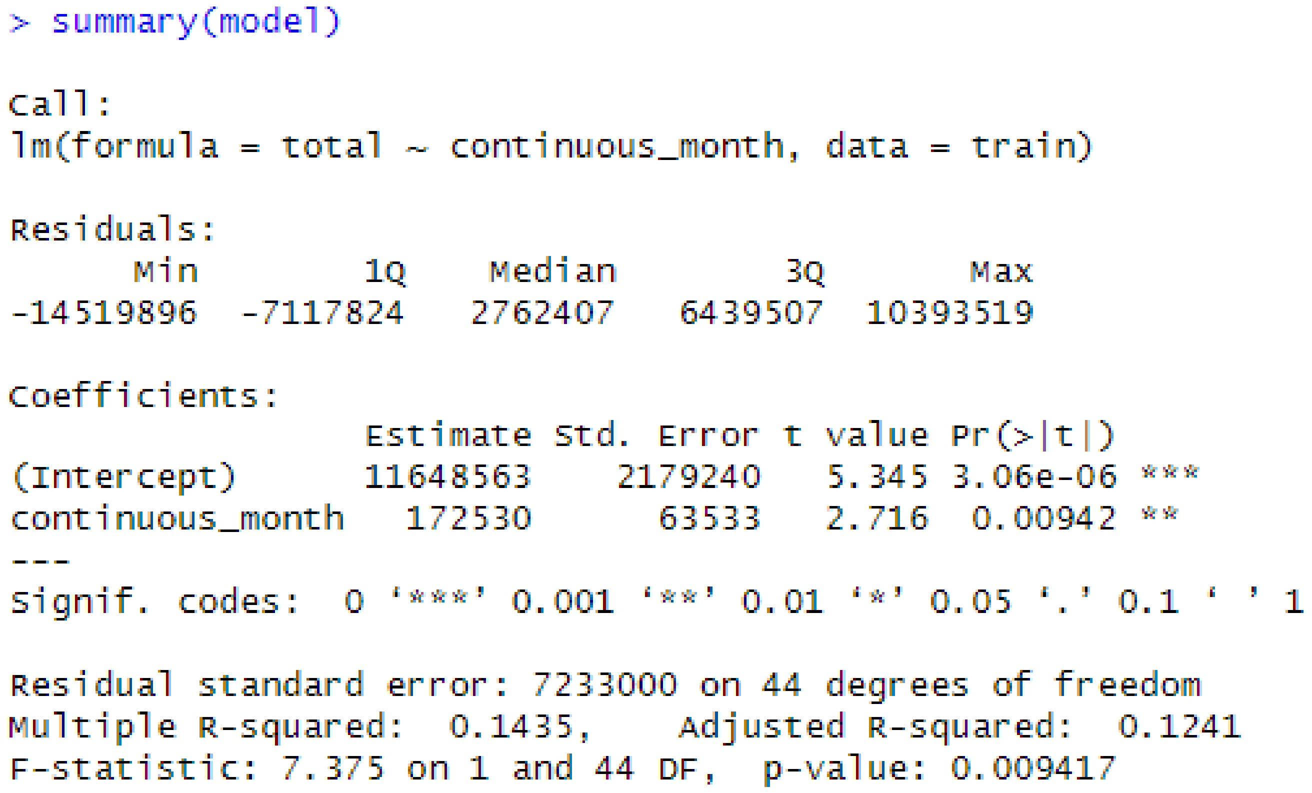
BUILD ANOTHER PREDICTIVE MODEL THAT CAN PERFORM BETTER THAN THE FIRST MODEL

MODEL : LINEAR REGRESSION

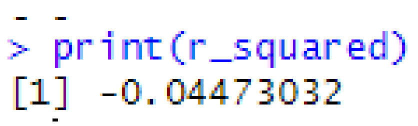
UNDERSTAND HOW VARIOUS FACTORS INFLUENCE RIDERSHIP NUMBERS

PREDICTIVE MODEL 1

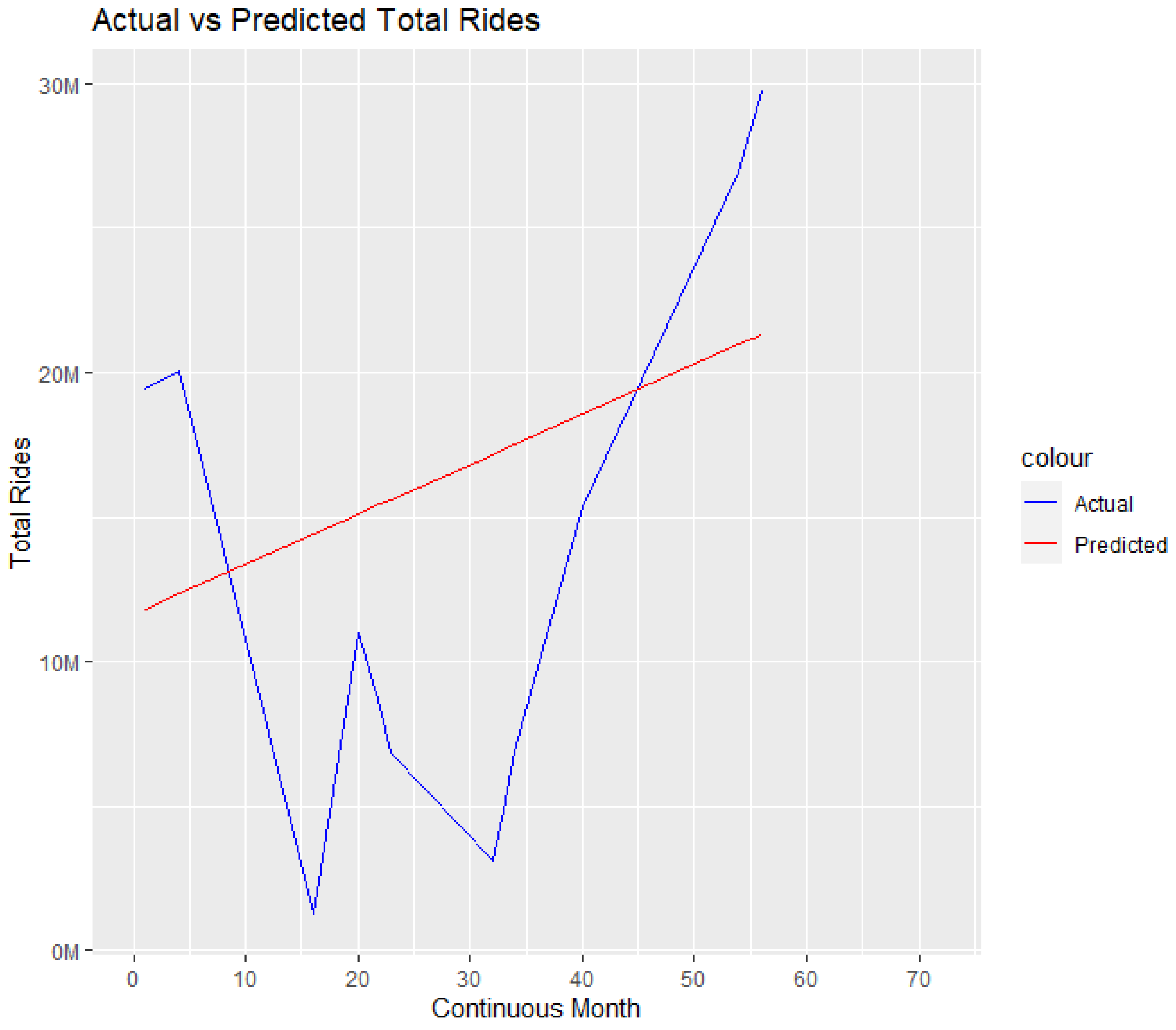
The predictive model was build up without by using every month available in the dataset.



```
> summary(model)

Call:
lm(formula = total ~ continuous_month, data = train)

Residuals:
      Min        1Q    Median        3Q       Max 
-14519896  -7117824   2762407   6439507  10393519 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      11648563    2179240   5.345 3.06e-06 ***
continuous_month   172530      63533   2.716  0.00942 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7233000 on 44 degrees of freedom
Multiple R-squared:  0.1435,	Adjusted R-squared:  0.1241 
F-statistic: 7.375 on 1 and 44 DF,  p-value: 0.009417
```



```
> print(r_squared)
[1] -0.04473032
```

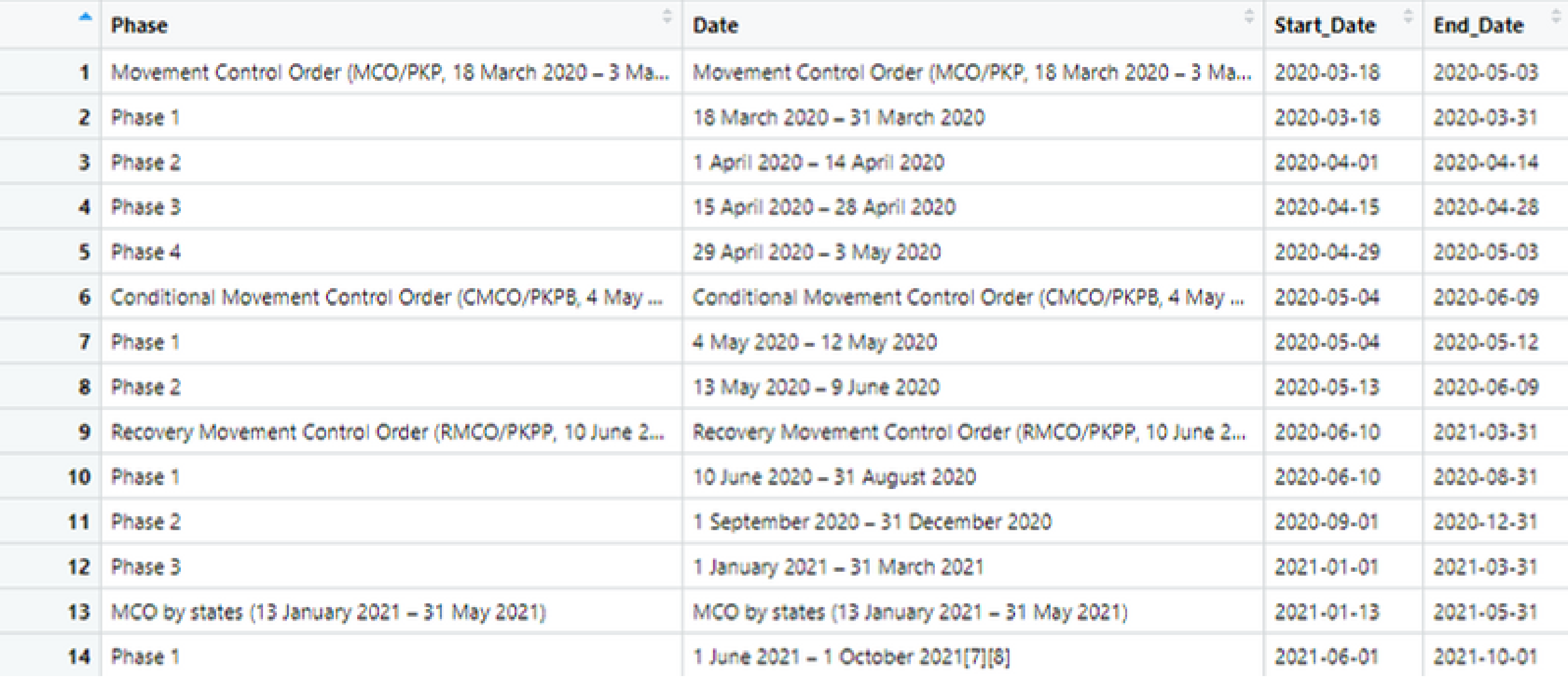


The graph shows the relationship between actual and the predicted outcome from the predictive model.

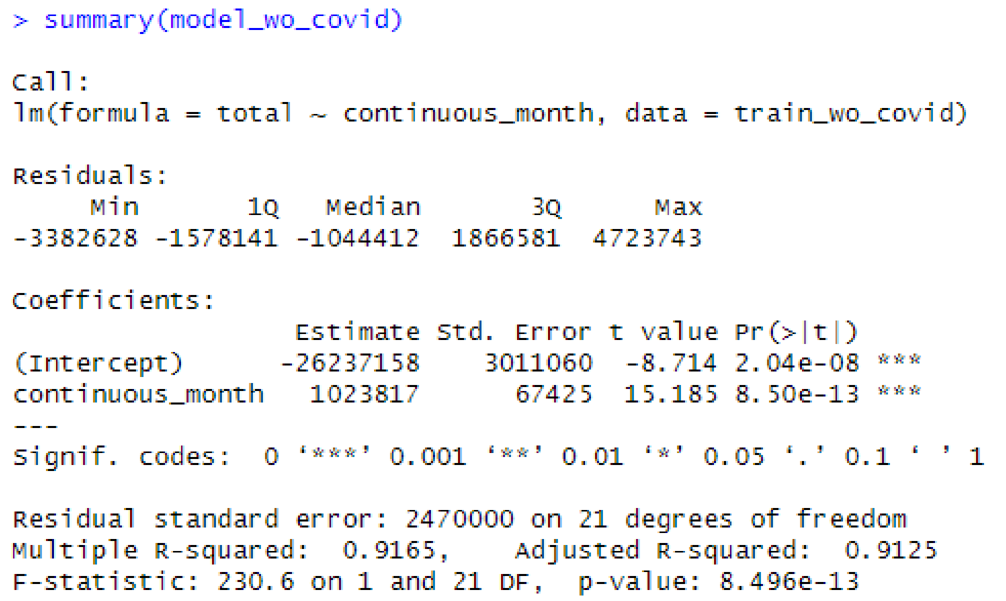
The R-squared value indicates that the predictive model is not well fitted to the actual dataset

PREDICTIVE MODEL 2

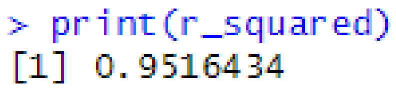
The model was developed by using data that exclude dates that are related with Covid-19 (MCO).



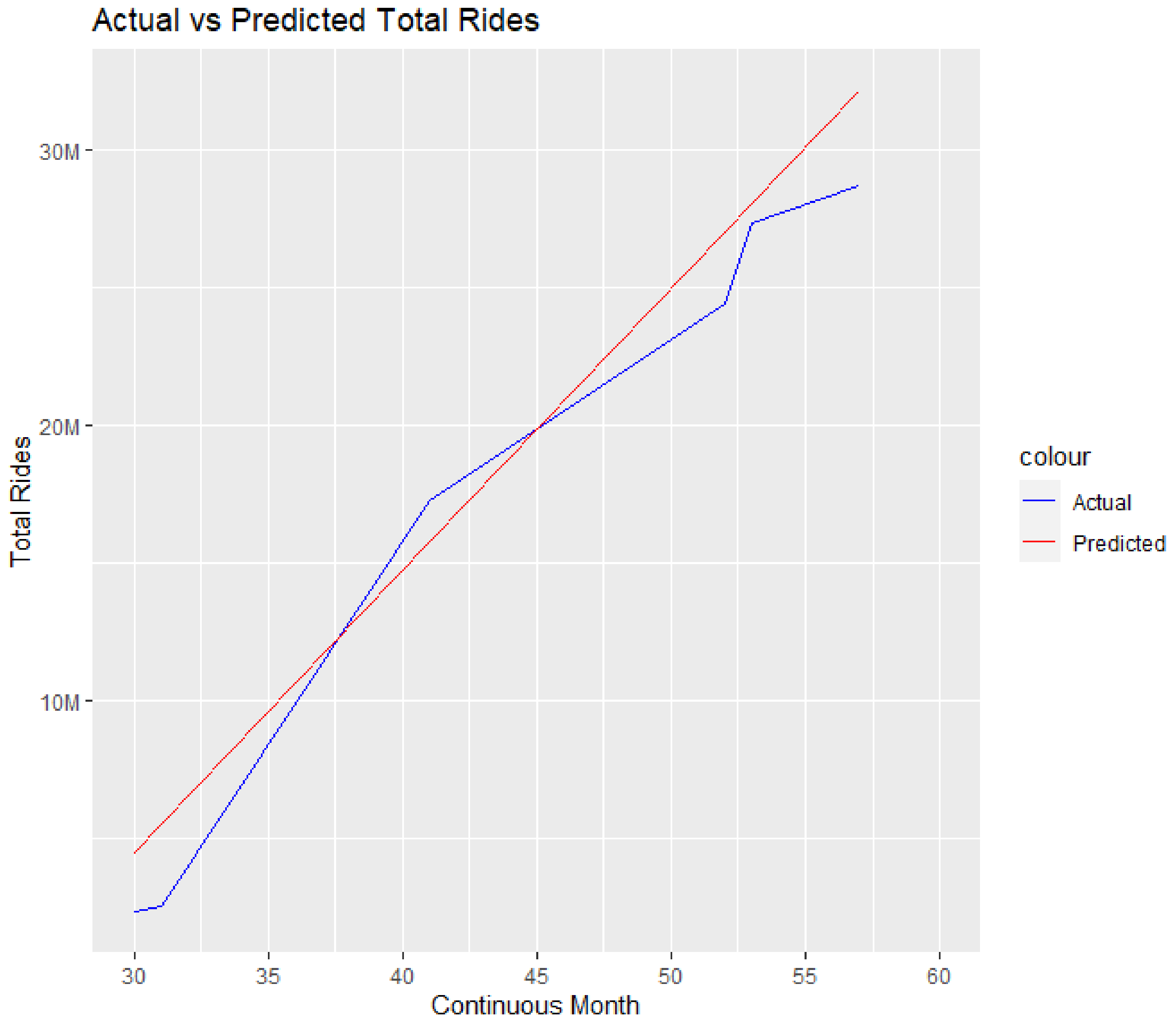
	Phase	Date	Start_Date	End_Date
1	Movement Control Order (MCO/PKP, 18 March 2020 – 3 Ma...	Movement Control Order (MCO/PKP, 18 March 2020 – 3 Ma...	2020-03-18	2020-05-03
2	Phase 1	18 March 2020 – 31 March 2020	2020-03-18	2020-03-31
3	Phase 2	1 April 2020 – 14 April 2020	2020-04-01	2020-04-14
4	Phase 3	15 April 2020 – 28 April 2020	2020-04-15	2020-04-28
5	Phase 4	29 April 2020 – 3 May 2020	2020-04-29	2020-05-03
6	Conditional Movement Control Order (CMCO/PKPB, 4 May ...	Conditional Movement Control Order (CMCO/PKPB, 4 May ...	2020-05-04	2020-06-09
7	Phase 1	4 May 2020 – 12 May 2020	2020-05-04	2020-05-12
8	Phase 2	13 May 2020 – 9 June 2020	2020-05-13	2020-06-09
9	Recovery Movement Control Order (RMCO/PKPP, 10 June 2...	Recovery Movement Control Order (RMCO/PKPP, 10 June 2...	2020-06-10	2021-03-31
10	Phase 1	10 June 2020 – 31 August 2020	2020-06-10	2020-08-31
11	Phase 2	1 September 2020 – 31 December 2020	2020-09-01	2020-12-31
12	Phase 3	1 January 2021 – 31 March 2021	2021-01-01	2021-03-31
13	MCO by states (13 January 2021 – 31 May 2021)	MCO by states (13 January 2021 – 31 May 2021)	2021-01-13	2021-05-31
14	Phase 1	1 June 2021 – 1 October 2021[7][8]	2021-06-01	2021-10-01



```
> summary(model_wo_covid)

Call:
lm(formula = total ~ continuous_month, data = train_wo_covid)

Residuals:
     Min       1Q   Median       3Q      Max 
-3382628 -1578141 -1044412  1866581  4723743 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      -26237158    3011060  -8.714 2.04e-08 ***
continuous_month   1023817      67425  15.185 8.50e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2470000 on 21 degrees of freedom
Multiple R-squared:  0.9165,	Adjusted R-squared:  0.9125 
F-statistic: 230.6 on 1 and 21 DF,  p-value: 8.496e-13
```



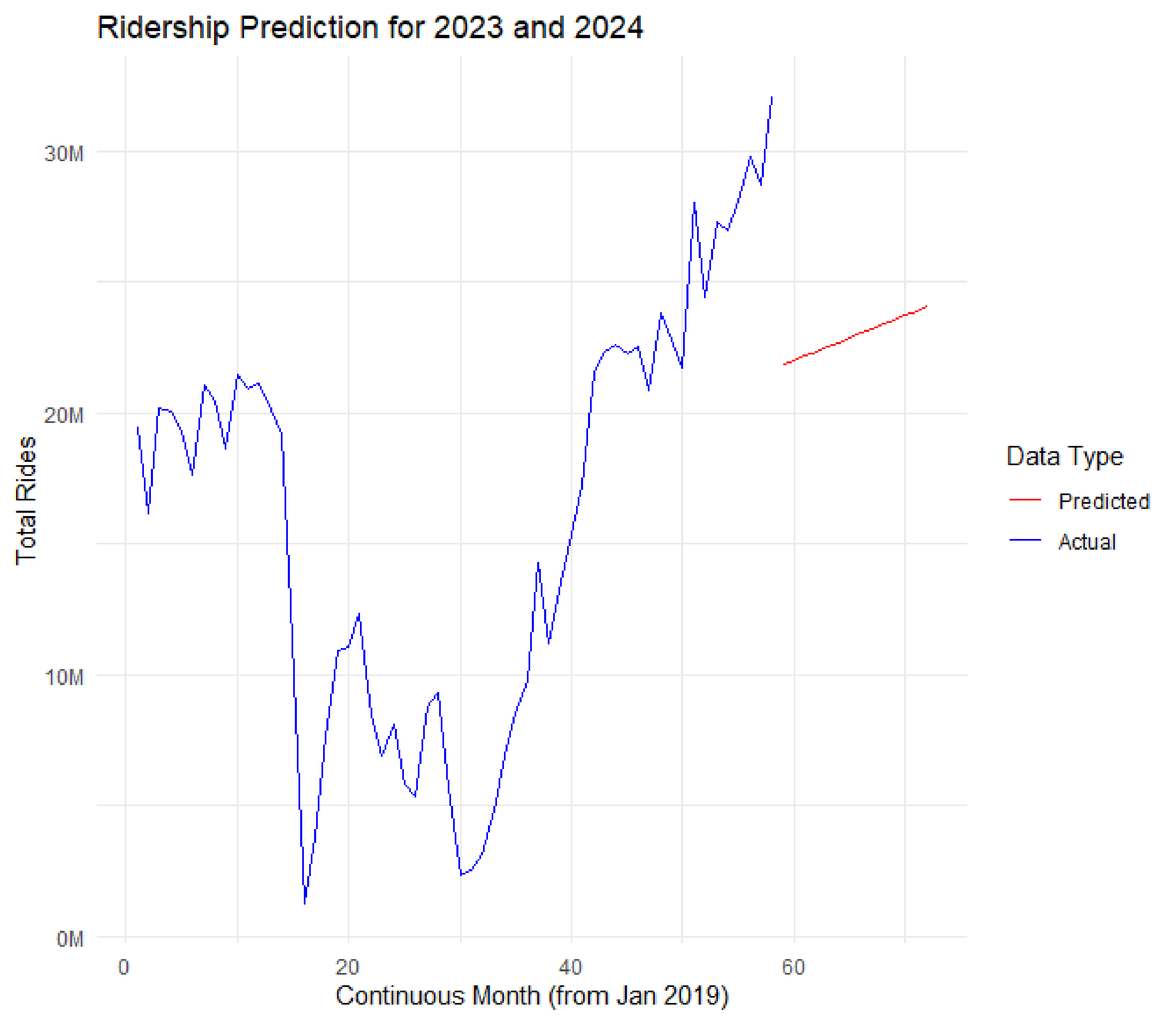
```
> print(r_squared)
[1] 0.9516434
```



- The graph shows the relationship between actual and the predicted outcome from the predictive model.
- Selective month outside of MCO dates is used to develop the predictive model.
- R-squared value indicates the predicted model fitted the actual dataset by 95%.

COMPARISON

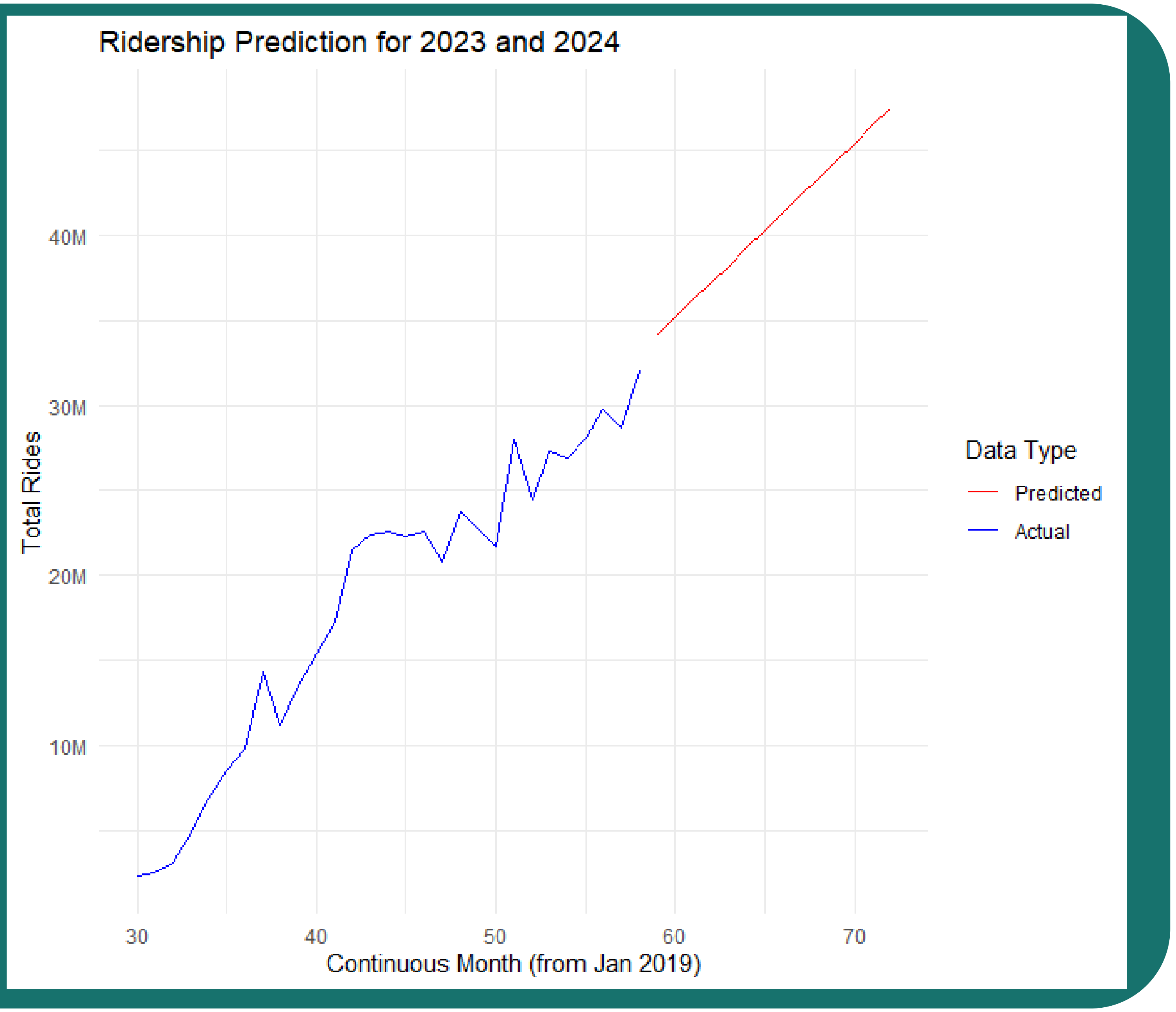
For Prediction Total Ridership Until 2024



With Covid

VS

Without Covid



Based on the graph, the predicted line is out of reach from the actual line. Most likely will not occur.

Predicted line is in the area continuation of actual line. Without any major external factor, most likely to land near the predicted line.

FINDINGS/CONCLUSION

THE AMOUNT OF TOTAL RIDES ARE MOST LIKELY TO DIFFER WHEN MAJOR EXTERNAL EVENT OCCURS

PREDICTIVE MODEL 2 > PREDICTIVE MODEL 1

PREDICTIVE MODEL 2 > PREDICTIVE MODEL 1

RELATED AGENCIES CAN MAKE PREPARATION ON EXPECTED INCREASED IN TOTAL RIDES OF PUBLIC TRANSPORTATION