# Modelling the Ridership data of Amtrak transportation Company- A Time Series Analysis


**M.Sc Statistics ( 2019 - 2021 )**
**Department of Statistics, Kalyani University**

| *NAME* | *REG. No.* | *Roll No.* |
|---|---|---|
| Prateeti De | 100647 of 2019-20 | 96/STA No. 190026 |
| Sattwik Roy | 100648 of 2019-20 | 96/STA No. 190036 |
| Ayan Pal | 100652 of 2019-20 | 96/STA No.190013 |

Supervised by **Aniruddha Chatterjee** & **Daipayan Sinha Roy**

# INDEX

## ACKNOWLEDGEMENT:

A project is a golden opportunity for learning. We consider ourselves very lucky & honoured to be given a chance to carry out a project. We are highly indebted to Aniruddha Chatterjee(CU) and Daipayan Sinha Roy (KU), for helping us in every possible way & providing all the necessary information regarding the project. Our thanks and appreciations also go to our classmates in developing the project & to the people who have willingly helped us out with their abilities .

## 1. INTRODUCTION :

A Time Series data is a series of data points collected, observed or recorded in successive intervals or points of time where the values of the variable are observed chronologically over days, weeks, months, quarters or years. For Example,

a. Population of a country over successive censuses
b. Temperature of a place noted at different days of a week
c. Number of passengers ride a train over last five years

Time Series are used in various fields such as, mathematical finance, manufacturing, event data, IoT data and generally in any domain of applied science and engineering which involves temporal measurements. This type of data generally has some deterministic components such as Trend component, Seasonal component, Cyclical component and a stochastic component consisting of random disturbances.

Time Series DBMS (database management system) are the fastest growing segment in the database industry and can testify to the growing need for Time Series forecasting in the industry.

In this project, our domain of interest is in **Analysing and Forecasting the Time Series Data**.

**Time Series Analysis** is a statistical technique that deals with Time Series data or Trend analysis. It is used to examine how the changes associated with the chosen data point compare to shifts in other variables over the same time period.
On the other hand, **Time Series Forecasting** uses information regarding historical values and associated patterns to predict future activity. Most often, this relates to Trend analysis, Cyclical fluctuation analysis and issues of Seasonality. To carry out the Analysis and Forecasting on the Time Series **R statistical software packages** are used.

## 2. DATA DESCRIPTION :

This project considers a secondary dataset abstracted from the **Amtrak Railway Transport Company,** which provides intercity passenger railway service in the contiguous United States and to nine Canadian cities. The following data represents Monthly Amtrak Train Ridership over the Canadian cities . There are a total of **159 data points** from **January 1991 to April 2003**.
The columns involved in this dataset are as follows -

**Table 1 : Table describing the columns in the dataset**

| | |
|---|---|
| **Months** | Represents month with the corresponding years |
| **Ridership** | Number of Riders travelled |
| **t** | Time period |
| **Seasons** | Shows only names of the months. |

To carry out Time Series analysis, we are considering the t is quite intuitive to split the data into training and test portions, so that the model created can be trained on first and then tested with the testing data. Training a single model is quite straightforward. We split it into 2 dataset -

- Training set for model fitting
- Testing data for estimation of model accuracy and forecasting.

Here , 80% of the data is taken as the training set and remaining data 20% is taken as test data. Hence, we now perform the entire analysis on the 127 time points. viz. from January 1991 to June 2001 and keep the rest of the data model accuracy and prediction purpose.

**Table 2 : A Monthly Amtrak train Ridership Data**

| Month | Ridership | t | Season | Month | Ridership | t | Season | Month | Ridership | t | Season |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 01-01-1991 | 1709 | 1 | Jan | 01-03-1993 | 1837 | 27 | Mar | 01-05-1995 | 1772 | 53 | May |
| 01-02-1991 | 1621 | 2 | Feb | 01-04-1993 | 1957 | 28 | Apr | 01-06-1995 | 1761 | 54 | Jun |
| 01-03-1991 | 1973 | 3 | Mar | 01-05-1993 | 1917 | 29 | May | 01-07-1995 | 1792 | 55 | Jul |
| 01-04-1991 | 1812 | 4 | Apr | 01-06-1993 | 1882 | 30 | Jun | 01-08-1995 | 1875 | 56 | Aug |
| 01-05-1991 | 1975 | 5 | May | 01-07-1993 | 1933 | 31 | Jul | 01-09-1995 | 1571 | 57 | Sep |
| 01-06-1991 | 1862 | 6 | Jun | 01-08-1993 | 1996 | 32 | Aug | 01-10-1995 | 1647 | 58 | Oct |
| 01-07-1991 | 1940 | 7 | Jul | 01-09-1993 | 1673 | 33 | Sep | 01-11-1995 | 1673 | 59 | Nov |
| 01-08-1991 | 2013 | 8 | Aug | 01-10-1993 | 1753 | 34 | Oct | 01-12-1995 | 1657 | 60 | Dec |
| 01-09-1991 | 1596 | 9 | Sep | 01-11-1993 | 1720 | 35 | Nov | 01-01-1996 | 1382 | 61 | Jan |
| 01-10-1991 | 1725 | 10 | Oct | 01-12-1993 | 1734 | 36 | Dec | 01-02-1996 | 1361 | 62 | Feb |
| 01-11-1991 | 1676 | 11 | Nov | 01-01-1994 | 1563 | 37 | Jan | 01-03-1996 | 1559 | 63 | Mar |
| 01-12-1991 | 1814 | 12 | Dec | 01-02-1994 | 1574 | 38 | Feb | 01-04-1996 | 1608 | 64 | Apr |
| 01-01-1992 | 1615 | 13 | Jan | 01-03-1994 | 1903 | 39 | Mar | 01-05-1996 | 1697 | 65 | May |
| 01-02-1992 | 1557 | 14 | Feb | 01-04-1994 | 1834 | 40 | Apr | 01-06-1996 | 1693 | 66 | Jun |
| 01-03-1992 | 1891 | 15 | Mar | 01-05-1994 | 1831 | 41 | May | 01-07-1996 | 1836 | 67 | Jul |
| 01-04-1992 | 1956 | 16 | Apr | 01-06-1994 | 1776 | 42 | Jun | 01-08-1996 | 1943 | 68 | Aug |
| 01-05-1992 | 1885 | 17 | May | 01-07-1994 | 1868 | 43 | Jul | 01-09-1996 | 1551 | 69 | Sep |
| 01-06-1992 | 1623 | 18 | Jun | 01-08-1994 | 1907 | 44 | Aug | 01-10-1996 | 1687 | 70 | Oct |
| 01-07-1992 | 1903 | 19 | Jul | 01-09-1994 | 1686 | 45 | Sep | 01-11-1996 | 1576 | 71 | Nov |
| 01-08-1992 | 1997 | 20 | Aug | 01-10-1994 | 1779 | 46 | Oct | 01-12-1996 | 1700 | 72 | Dec |
| 01-09-1992 | 1704 | 21 | Sep | 01-11-1994 | 1776 | 47 | Nov | 01-01-1997 | 1397 | 73 | Jan |
| 01-10-1992 | 1810 | 22 | Oct | 01-12-1994 | 1783 | 48 | Dec | 01-02-1997 | 1372 | 74 | Feb |
| 01-11-1992 | 1862 | 23 | Nov | 01-01-1995 | 1548 | 49 | Jan | 01-03-1997 | 1708 | 75 | Mar |
| 01-12-1992 | 1875 | 24 | Dec | 01-02-1995 | 1497 | 50 | Feb | 01-04-1997 | 1655 | 76 | Apr |
| 01-01-1993 | 1705 | 25 | Jan | 01-03-1995 | 1798 | 51 | Mar | 01-05-1997 | 1763 | 77 | May |
| 01-02-1993 | 1619 | 26 | Feb | 01-04-1995 | 1733 | 52 | Apr | 01-06-1997 | 1776 | 78 | Jun |

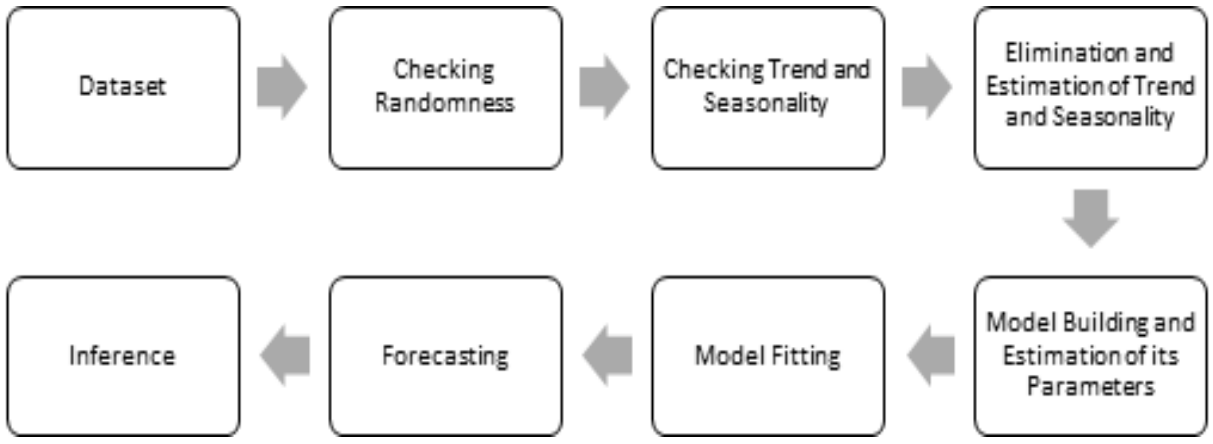
| Month | Ridership | t | Season | Month | Ridership | t | Season | Month |
|---|---|---|---|---|---|---|---|---|
| Jul | 01-10-1999 | 1804 | 106 | Oct | 01-01-2002 | 1760 | 133 | Jan |
| Aug | 01-11-1999 | 1850 | 107 | Nov | 01-02-2002 | 1771 | 134 | Feb |
| Sep | 01-12-1999 | 1836 | 108 | Dec | 01-03-2002 | 2020 | 135 | Mar |
| Oct | 01-01-2000 | 1542 | 109 | Jan | 01-04-2002 | 2048 | 136 | Apr |
| Nov | 01-02-2000 | 1617 | 110 | Feb | 01-05-2002 | 2069 | 137 | May |
| Dec | 01-03-2000 | 1920 | 111 | Mar | 01-06-2002 | 1994 | 138 | Jun |
| Jan | 01-04-2000 | 1971 | 112 | Apr | 01-07-2002 | 2075 | 139 | Jul |
| Feb | 01-05-2000 | 1992 | 113 | May | 01-08-2002 | 2027 | 140 | Aug |
| Mar | 01-06-2000 | 2010 | 114 | Jun | 01-09-2002 | 1734 | 141 | Sep |
| Apr | 01-07-2000 | 2054 | 115 | Jul | 01-10-2002 | 1917 | 142 | Oct |
| May | 01-08-2000 | 2097 | 116 | Aug | 01-11-2002 | 1858 | 143 | Nov |
| Jun | 01-09-2000 | 1824 | 117 | Sep | 01-12-2002 | 1996 | 144 | Dec |
| Jul | 01-10-2000 | 1977 | 118 | Oct | 01-01-2003 | 1778 | 145 | Jan |
| Aug | 01-11-2000 | 1981 | 119 | Nov | 01-02-2003 | 1749 | 146 | Feb |
| Sep | 01-12-2000 | 2000 | 120 | Dec | 01-03-2003 | 2066 | 147 | Mar |
| Oct | 01-01-2001 | 1683 | 121 | Jan | 01-04-2003 | 2099 | 148 | Apr |
| Nov | 01-02-2001 | 1663 | 122 | Feb | 01-05-2003 | 2105 | 149 | May |
| Dec | 01-03-2001 | 2008 | 123 | Mar | 01-06-2003 | 2130 | 150 | Jun |
| Jan | 01-04-2001 | 2024 | 124 | Apr | 01-07-2003 | 2223 | 151 | Jul |
| Feb | 01-05-2001 | 2047 | 125 | May | 01-08-2003 | 2174 | 152 | Aug |
| Mar | 01-06-2001 | 2073 | 126 | Jun | 01-09-2003 | 1931 | 153 | Sep |
| Apr | 01-07-2001 | 2127 | 127 | Jul | 01-10-2003 | 2121 | 154 | Oct |
| May | 01-08-2001 | 2203 | 128 | Aug | 01-11-2003 | 2076 | 155 | Nov |
| Jun | 01-09-2001 | 1708 | 129 | Sep | 01-12-2003 | 2141 | 156 | Dec |
| Jul | 01-10-2001 | 1951 | 130 | Oct | 01-01-2004 | 1832 | 157 | Jan |
| Aug | 01-11-2001 | 1974 | 131 | Nov | 01-02-2004 | 1838 | 158 | Feb |
| Sep | 01-12-2001 | 1985 | 132 | Dec | 01-03-2004 | 2132 | 159 | Mar |

**Link of the dataset : [Amtrak Raw Dataset]**

## 3. OBJECTIVES :

The main objective of this project is to build a time series model that can depict the deterministic components and then using that model we need to ***forecast for the next 12 months of the ridership.***

***Route Map :***



# 4. ANALYSIS :

"*In our view, the first step in any time series investigation always involves careful scrutiny of the recorded data plotted over time. This scrutiny often suggests the method of analysis as well as statistics that will be of use in summarizing the information in the data.*"

*--- Shumay and Stoffer*

This part extracts meaningful statistics and other characteristics of the dataset in order to understand it. Time Series analysis can help to make better predictions, but this is not necessarily the main goal of the analysis. In practice a suitable model is fitted to a given Time Series and the corresponding parameters are estimated using the known data values. The Time Series analysis process comprises methods that attempt to understand the nature of the series and is often useful for future forecasting and simulation. This field of study seeks the "*why*" behind a Time Series dataset.

## 4.1 PLOTS :

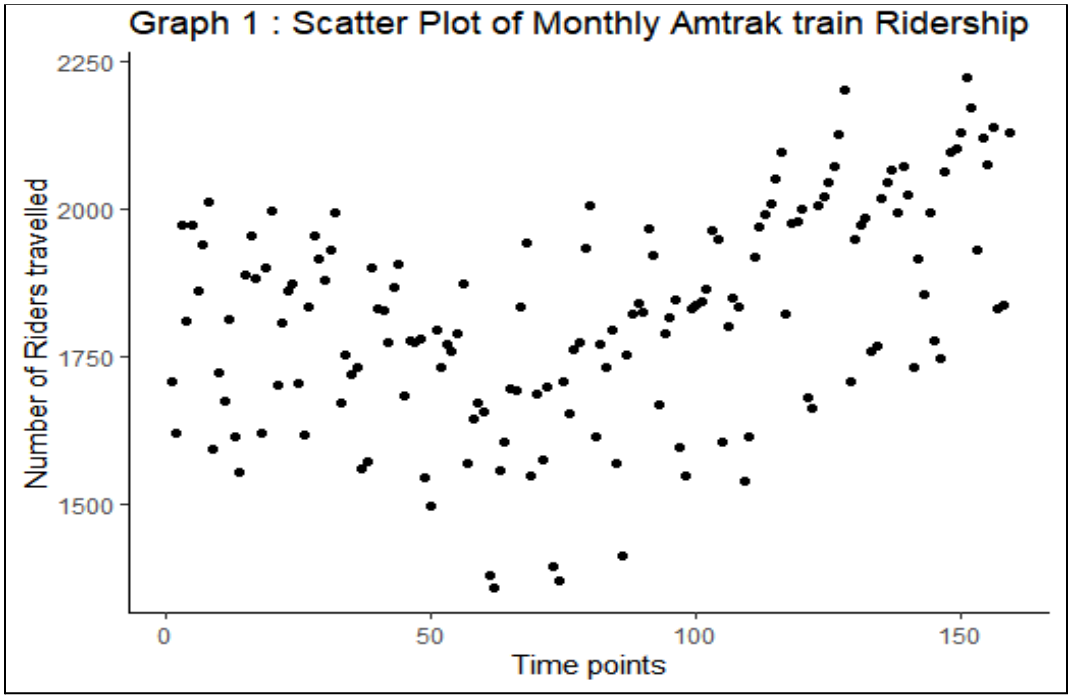



*fig 1*

**Comment :**The Scatter Plot of the data is showing a **slightly upward trend**.To study such characteristics of the data with more details, we will carry out the following Analysis.

**Graph 2 :**

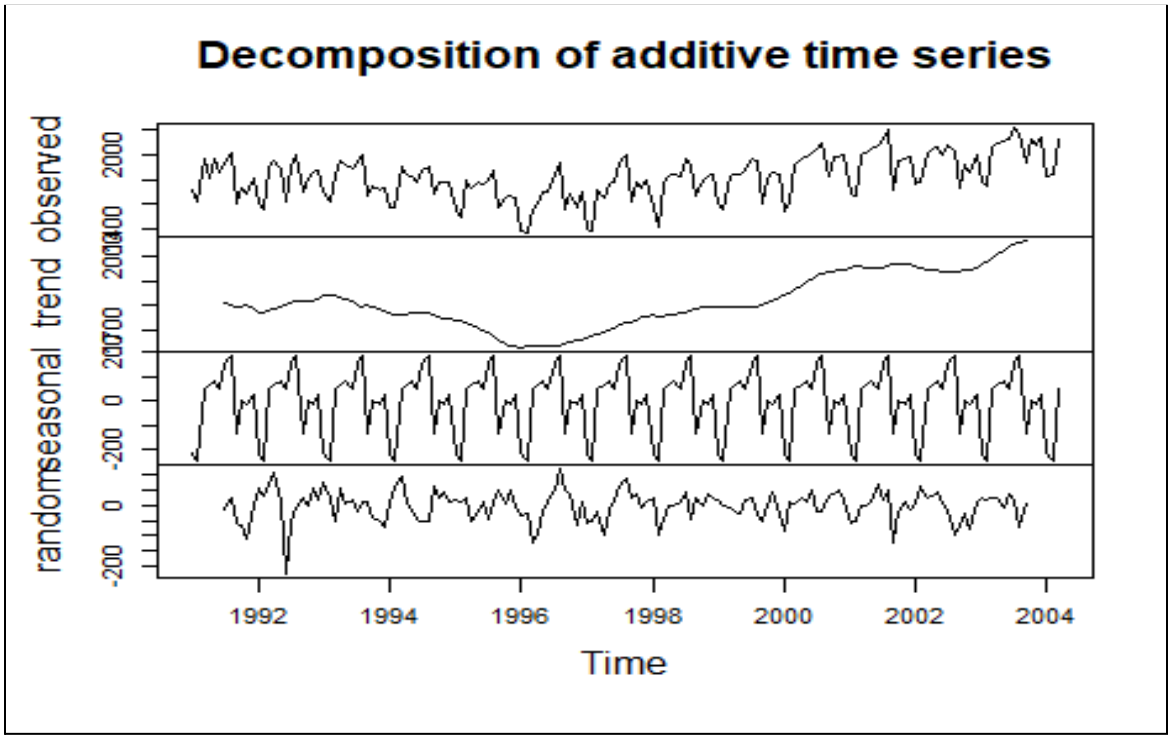



*fig 2*

**Comment :** To further analyze trends in the data, we can decompose the data into seasonal, and trend components. **Decomposition** is the foundation for building the ARIMA model by providing insight towards the changes in the number of riders.

- The **observed data** section is a reproduction of the data to understand the comparison.
- The **trend component** describes the overall pattern of the series over the entire range of time, taking into account increases and decreases in riderships. From the plot above, the trend is **overall increasing**.
- The **seasonal component** describes the fluctuations in the number of Riders based on the calendar From the plot above, the peak in the number of Riders occurs every year at Q4 (July and August) and the trough in the number of Riders occurs every year at Q2 (January and February), with clear oscillating functions in between.
- The **random section** describes the trends that cannot be explained by trend or seasonal components. Statistically, Random error is particularly important for this project because a statistical model can only be fit if the residuals are independent and independently distributed**.**

## 4.2 TEST OF RANDOMNESS :

Now we check if our data is purely random or some deterministic part is present in it. For this purpose, we here perform the **Turning Point Test**.

The **Turning Point Test** helps us to understand whether a set of random variables are **independently and identically distributed or not**.

Suppose we have '$n$' number of Time Series data points, then consider three consecutive numbers among these '$n$' points. If the middle one is or less than the other two points, then we can say that it is a **Turning Point**, i.e. $y_i$ is defined to be a Turning Point if $y_i > y_{i-1}$ and $y_i > y_{i+1}$ or $y_i < y_{i-1}$ and $y_i < y_{i+1}$.

On the other hand, if in a set of observations every observation is little greater than the next, then it would fail to have any Turning Points.

Here our Null Hypothesis and Alternative Hypothesis are:

$H_0$: **The data is purely random** *against* $H_1$: **Not** $H_0$

Here the appropriate Test Statistic for testing $H_0$ is given by,

$$Z = \frac{(U - E(U))}{SD(U)}$$ follows N(0,1) under $H_0$

where $U$ is the total number of Turning Points with $E(U) = \frac{(2n-4)}{3}$ and $V(U) = \frac{(16n-29)}{90}$.

*Critical Region :* In case of two-tailed test, we will reject the null hypothesis $H_0$ iff $|Z| > Z_{\frac{\alpha}{2}}$, at level of significance $\alpha$.

### Findings from dataset:

We see our observed $|Z| = 2.61 > 1.96$, which is the tabulated Standard Normal value at $\alpha = 0.05$. Hence we reject $H_0$ at $5\%$ level of significance and conclude that **the data is not purely random, i.e. it has deterministic components.**

## 4.3 TEST OF EXISTENCE OF TREND :

From **graph 2**, we can see that the Trend component is present. So, now we aim to estimate and eliminate the deterministic components.

First, we perform the **Relative Ordering Test** to check for the existence of Trend.

It is a non-parametric measure of a relationship between columns of sequential data and the Time-Series is sequential. Hence we can use $\tau$ to check the relationship between time and

the dependent variable. For our dataset, the dependent variable is Monthly number of riders travelled and time, here, is months for different years.

If they are **highly correlated**, then we can say a Trend exists, as the timestamps are always increasing. Consequently, if they are **positively correlated**, an **increasing Trend exists**. Whereas, if they are **negatively correlated**, a **decreasing Trend exists**.

Here our Null Hypothesis and Alternative Hypothesis are:

$H_0$: **There is no presence of Trend** *against* $H_1$: **Not** $H_0$

Here the appropriate Test Statistic for testing $H_0$ is given by,

$$Z = \frac{(\tau - E(\tau))}{SD(\tau)}$$ follows N(0,1) under $H_0$

where, 'τ' is **Kendall's Rank Correlation Coefficient**, which is related to '$Q$' (the number of discordances).

$$\tau = 1 - \frac{4Q}{n(n-1)}$$

*Critical Region :* In case of two-tailed test, we will reject the null hypothesis $H_0$ if $|Z| > Z_{\frac{\alpha}{2}}$, at level of significance $\alpha$.

## Findings from dataset:

We see our observed $|Z| = 1.98 > 1.96$, which is tabulated standard normal value at $\alpha = 0.05$. Hence we reject $H_0$ at $5\%$ level of significance and conclude that the **Trend component is present in our data.**

## 4.4 ELIMINATION AND ESTIMATION OF TREND :

First we obtain a rough estimate of the trend component using **Moving Average Method** of period equal to the period of seasonality, so that the effect of seasonality can be avoided. The estimate of trend is given by:

$$\widehat{m_t} = \frac{1}{2q}\left(0.5y_{t-q} + y_{t-q+1} + \ldots + y_t + \ldots + 0.5y_{t+q}\right)$$

Where 2q=12 is the period of seasonality,

$m_t$ is the trend component corresponding to $t^{th}$ time point.

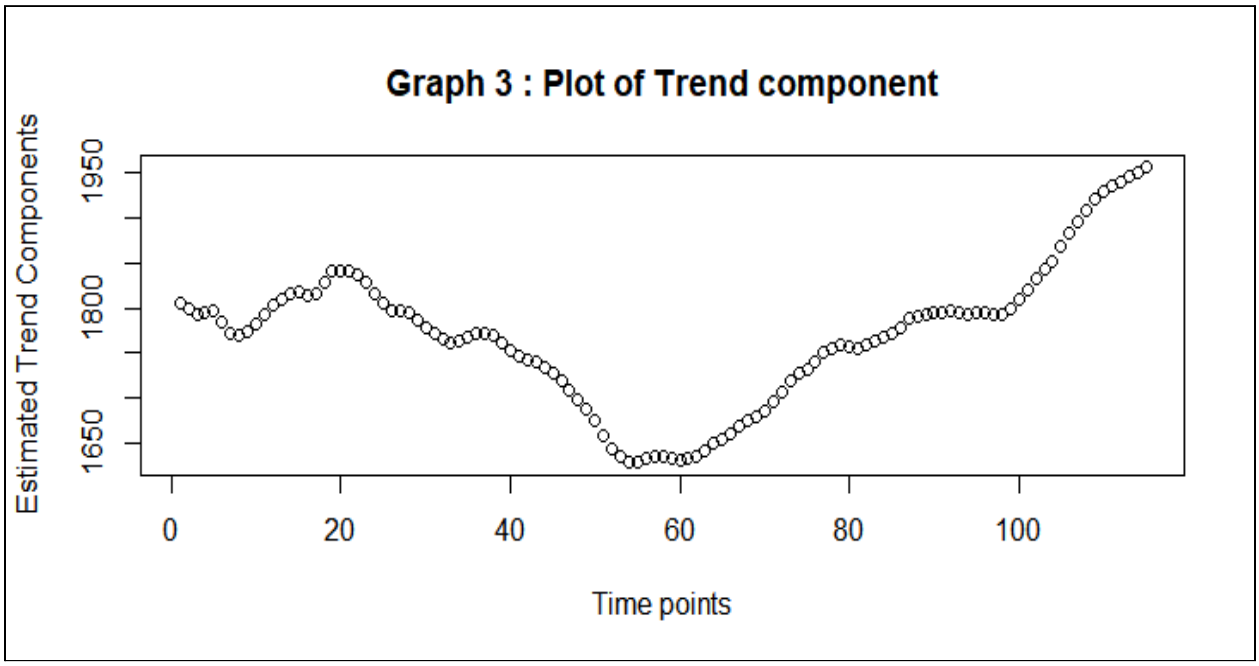



*fig 3*

Now, we subtract the estimated trend values from the original data. Thus the trend component is eliminated and we are left with the **detrended data**.

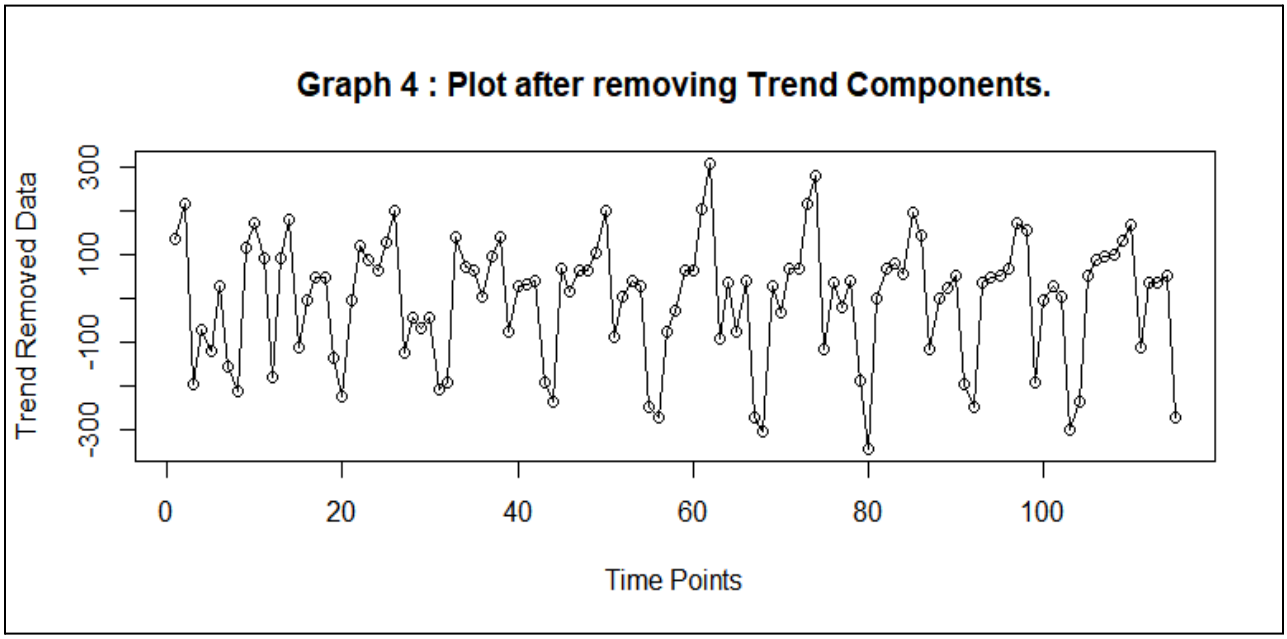



*fig 4*

## 4.5 TEST OF SEASONALITY :

Since trend is present, we detrended the data and now we check for seasonality using **Friedman's Test**.
**F-test** or **Friedman's Test** is a test for the presence of seasonality based on a one-way analysis of variance of SI ratios.Thus the test is performed on the detrended time series adjusted for prior factors.

$H_0$**: No stable seasonality against** $H_1$**: there is a stable seasonality**

From the plot of the data the period of seasonality is guessed to be 12 and hence the **test statistic** is the following:

$$S = \frac{12\sum_{i=1}^{12}\left(M_i - \frac{c(r+1)}{2}\right)^2}{(cr(r+1))}$$

Where c is the number of years, r is the number of months, i.e, 12, $M_i$ is the $i^{th}$ monthly total.

*Critical Region :* In this Chi-Sq test we reject our null hypothesis if $S > \chi^2_{r-1,\alpha}$ at $\alpha$level of significance.

### Findings from dataset:

This statistic follows a Chi-square distribution with (r-1) = 11 degrees of freedom under $H_0$. The value of our **test statistic is 106.90** greater than the **tabulated chi-square value 19.67** at 5% level of significance, so we conclude that **our data shows presence of seasonality.**

## 4.6 ELIMINATION AND ESTIMATION OF SEASONALITY :

The seasonal component of the data is estimated by **method of differencing**. We apply backward differencing of order same as the period of seasonality on the data. Then applying the test for seasonality as before we see that the data has been deseasonalized. So it is also indicated by the plot.

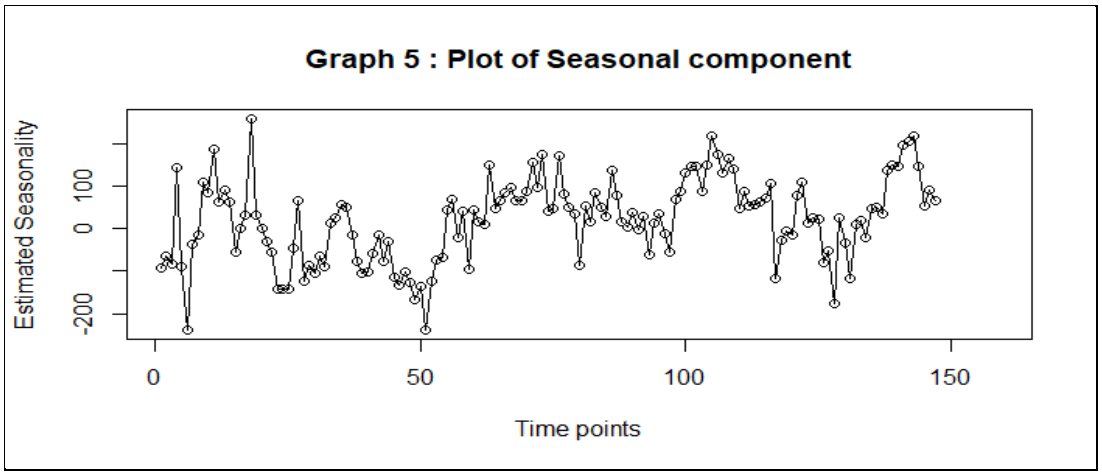



*fig 5*

Seasonal effects can be eliminated by a simple linear filter called **Seasonal Differencing.** For example, in case of monthly data one can use the operator:

$$y_t = \nabla_{12} X_t = X_t - X_{t-12}$$

Now, after elimination of seasonal components, we again check the presence of trend in the data by the relative ordering test as mentioned above. We hereby see that the null hypothesis of no trend is being accepted. **Thus now we have a detrend and deseasonalized data.**

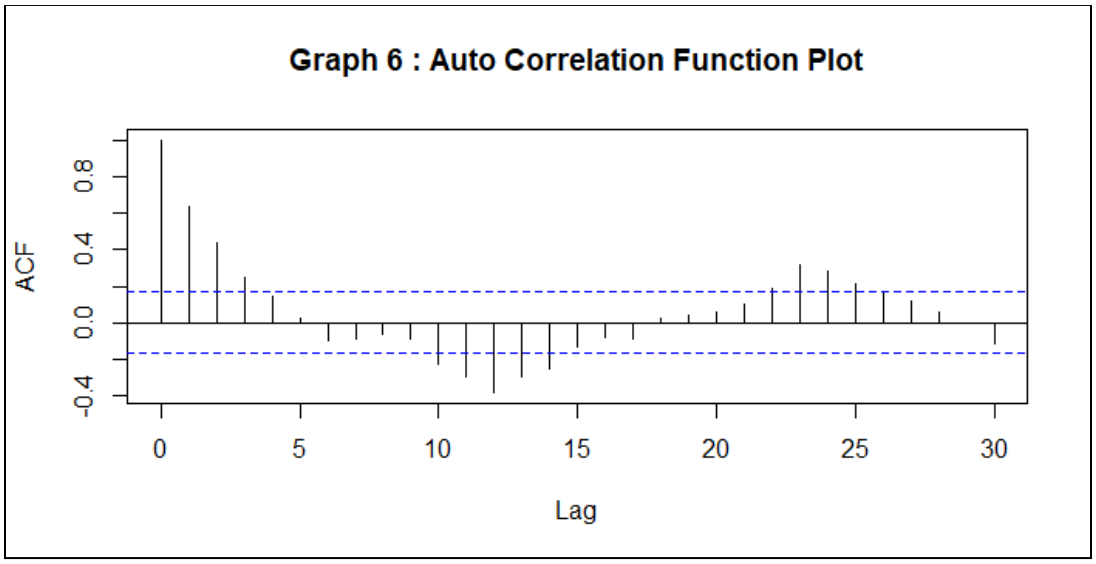



*fig 6*

From the above pictures we can definitely say that **desesnlizd_detrndd_dt** is a **stationary data**.

# 5. FORECASTING MODELS :

## 5.1. SARIMA:

**The SARIMA** model provides an approach to time series forecasting. This model may possibly include AR terms, MA terms and differencing operations with seasonal AR, MA and differencing components with a period of seasonality.
We can evaluate the orders of AR, MA and differencing by plotting the ACF and PACF.

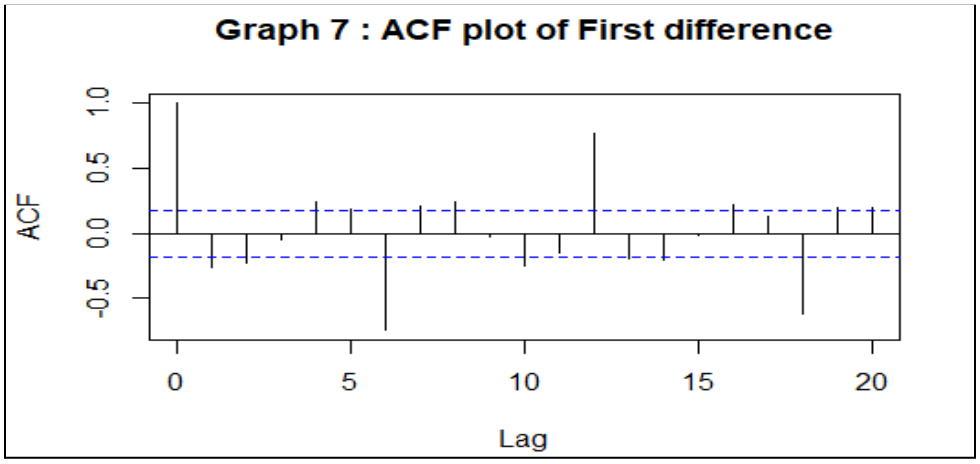



*fig 7*

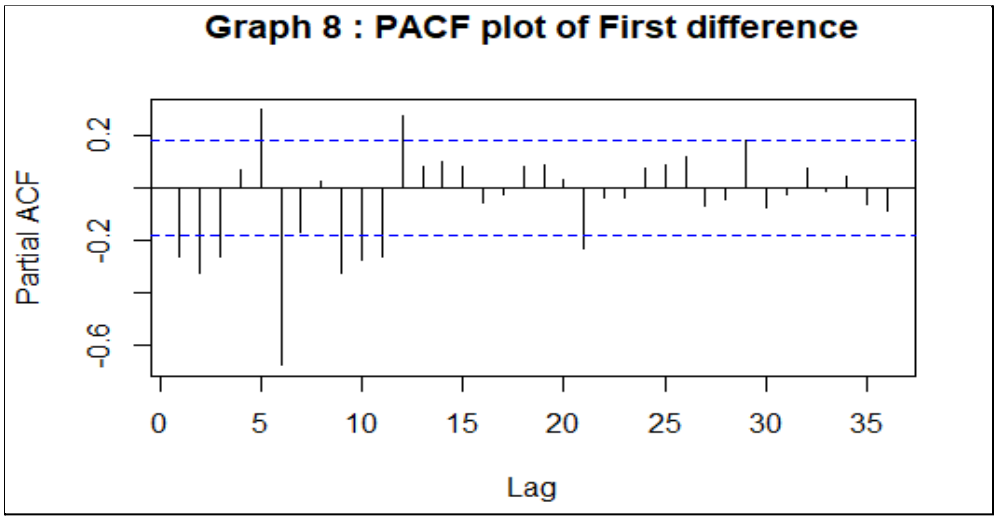



*fig 8*

**Comment :** After inspecting the ACF and PACF plots properly we can conclude that the orders must be *(6,1,6)* with seasonality. And the order of the seasonal component is *(1,0,1)*. The period of the seasonal component can be derived from the PACF plot. As we can see from that plot there is a cyclical nature after each three lags so the period is 3.

Finally, our model looks like ***Seasonal ARIMA (6,1,6)*(1,0,1)S=3.***

***Coefficients of SARIMA :***

| ar1 | ar2 | ar3 | ar4 | ar5 | ar6 |
|---|---|---|---|---|---|
| -0.0080 | -0.0022 | 0.0187 | 0.0199 | 0.0030 | -0.9776 |
| ma1 | ma2 | ma3 | ma4 | ma5 | ma6 |
| -0.2613 | 0.079 | -0.1209 | -0.2437 | 0.0829 | 0.6819 |
| sar1 | sma1 | | | | |
| -0.9923 | 0.9182 | | | | |

***Fitting of SARIMA model on train and test data :***

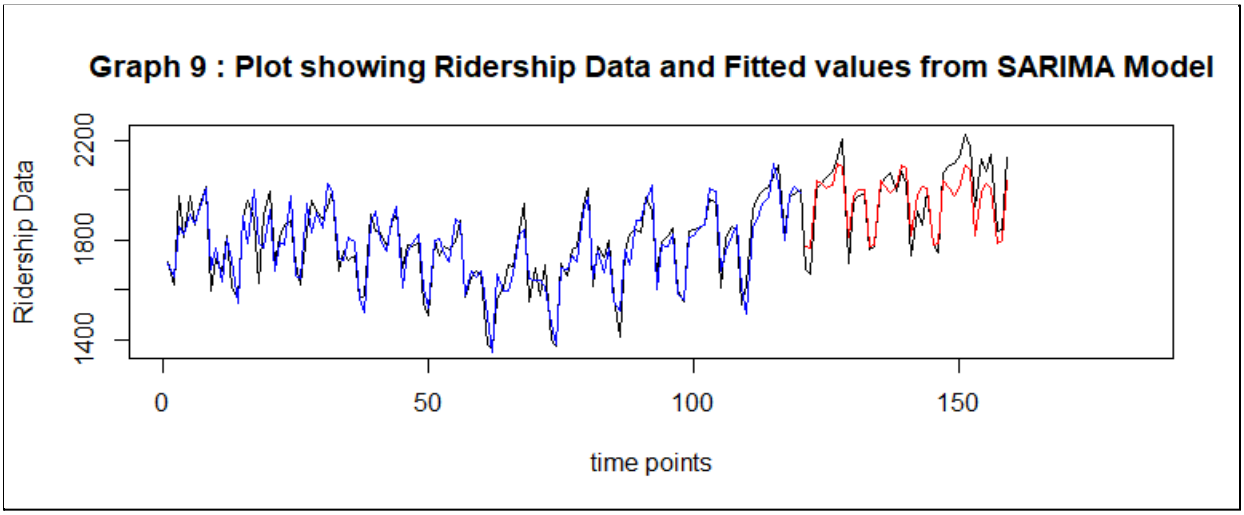



*fig 9*

**Comment** : In the above plot, though the model fits well for the train dataset but it does not predict satisfactorily for the test dataset. Hence we are considering a further advance model for better prediction and forecasting.

## 5.2. SARIMAX:

*Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors*, or SARIMAX, is an extension of the ARIMA class of models. Intuitively, ARIMA models compose 2

parts: the autoregressive term (AR) and the moving-average term (MA). The former views the value at one time just as a weighted sum of past values. The latter model has the same value also as a weighted sum but of past residuals. There is also an integrated term (I) to differentiate the time series.

**SARIMAX extends on this framework just by adding the capability to handle exogenous variables.** Here we take *Season as an exogenous variable*. Season consists of twelve months. We create dummy columns for each month except April. By doing this we achieve the interpretability of these dummies.

- *Here we encoded the season column with dummies. That means we create 12 dummies for each month starting from January to December, and then drop one column that is the column for April to regain the interpretability of the dummies.*

| SeasonAu | SeasonDe | SeasonFe | SeasonJa | SeasonJu | SeasonJu | SeasonMa | SeasonMa | SeasonNo | SeasonOc | SeasonSe |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |

### *Coefficients of SARIMAX:*

| **ar1** | **ar2** | **ar3** | **ar4** | **ar5** | **ar6** |
|---|---|---|---|---|---|
| 0.2004 | 0.0725 | 0.0831 | 0.4538 | -0.6486 | -0.1788 |
| **ma1** | **ma2** | **ma3** | **ma4** | **ma5** | **ma6** |
| -0.5800 | -0.0727 | -0.1524 | -0.4950 | 1.0024 | -0.1623 |
| **sar1** | **sma1** | **SeasonAug** | **SeasonDec** | **SeasonFeb** | **SeasonJan** |

| | | | | | |
|---|---|---|---|---|---|
| -0.0651 | 0.1193 | 146.2292 | -21.4393 | -295.9179 | -253.3127 |
| **SeasonJul** | **SeasonJun** | **SeasonMar** | **SeasonMay** | **SeasonNov** | **SeasonOct** |
| 95.6581 | -10.2331 | 1.1095 | 34.0913 | -63.5591 | -49.0934 |
| **SeasonSep** | | | | | |
| -174.5044 | | | | | |

## 6. IDENTIFYING THE BEST MODEL:

Now we compare our forecasts using the Root Mean Square Error metric.

***RMSEs for test data:***
Here we are comparing the RMSEs to find the best model among these two. The data we are using to calculate this metric is the test data.

| SARIMA Projection | SARIMAX Projection |
|---|---|
| 76.45000 | 75.68301 |

As we can clearly see that the **SARIMAX model has the minimum RMSE**, we can conclude that this is the best model for forecasting the Rider's data.

***RMSEs for train data:***
For more convenience we even checked the RMSEs for train sets.

| SARIMA Projection | SARIMAX Projection |
|---|---|
| 60.89841 | 52.33472 |

Here we also find that the **SARIMAX** model gives the **minimum RMSE**.

- As we found that SARIMAX is the best model therefore we use this model to forecast as well as for the comparison with the test-data.

## 7. FORECASTING:

Forecasting time series data, the aim is to estimate how the sequence of observations will continue into the future. Here we forecast 12 observations i.e. **April 2001 to March 2002.**

We use our Sarimax model to forecast these time points as it is the best model.

***SARIMAX Forecasts:***

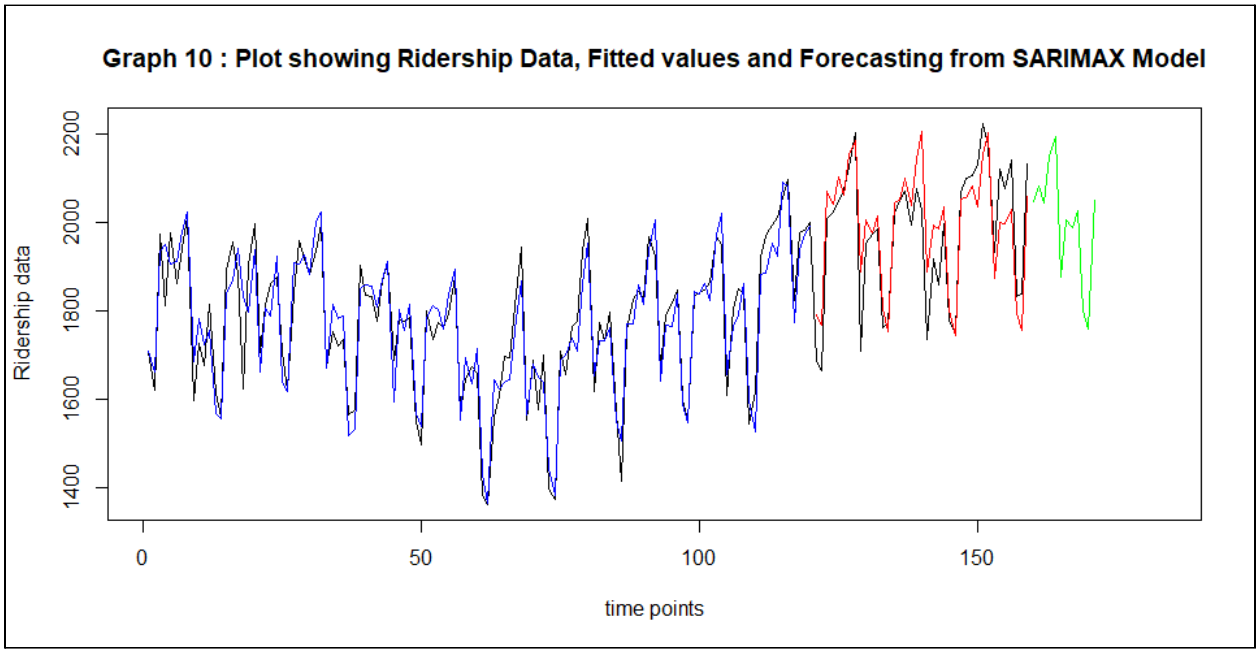
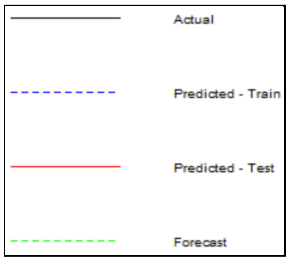



*fig 10*




**Comment :** Here we consider an **exogenous variable (Season)** along with the seasonal components that's why we get the best forecast.

## 7.1. COMPARISON WITH TEST DATA :

| *MONTH* | *RIDERSHIP* | *SARIMAX_PREDICTIONS* |
|---|---|---|
| 1/1/2001 | 1683 | 1789.172929 |
| 2/1/2001 | 1663 | 1765.369591 |
| 3/1/2001 | 2008 | 2068.808612 |
| 4/1/2001 | 2024 | 2041.096347 |
| 5/1/2001 | 2047 | 2101.42751 |
| 6/1/2001 | 2073 | 2060.23361 |
| 7/1/2001 | 2127 | 2153.301656 |
| 8/1/2001 | 2203 | 2184.419595 |
| 9/1/2001 | 1708 | 1886.95629 |

| | | |
|---|---|---|
| 10/1/2001 | 1951 | 2003.719844 |
| 11/1/2001 | 1974 | 1975.030682 |
| 12/1/2001 | 1985 | 2014.547312 |
| 1/2/2001 | 1760 | 1805.871945 |
| 2/2/2001 | 1771 | 1751.008274 |
| 3/2/2001 | 2020 | 2042.037016 |
| 4/2/2001 | 2048 | 2050.354258 |
| 5/2/2001 | 2069 | 2099.640047 |
| 6/2/2001 | 1994 | 2038.402836 |
| 7/2/2001 | 2075 | 2143.869762 |
| 8/2/2001 | 2027 | 2204.752522 |
| 9/2/2001 | 1734 | 1886.50312 |
| 10/2/2001 | 1917 | 1993.908948 |
| 11/2/2001 | 1858 | 1984.931492 |
| 12/2/2001 | 1996 | 2035.030799 |
| 1/3/2001 | 1778 | 1798.173731 |
| 2/3/2001 | 1749 | 1743.978605 |
| 3/3/2001 | 2066 | 2052.70911 |
| 4/3/2001 | 2099 | 2055.97168 |
| 5/3/2001 | 2105 | 2082.406436 |
| 6/3/2001 | 2130 | 2034.38313 |
| 7/3/2001 | 2223 | 2153.060057 |
| 8/3/2001 | 2174 | 2201.75528 |
| 9/3/2001 | 1931 | 1872.862599 |
| 10/3/2001 | 2121 | 2000.070393 |
| 11/3/2001 | 2076 | 1994.78751 |
| 12/3/2001 | 2141 | 2029.717451 |
| 1/4/2001 | 1832 | 1792.446176 |
| 2/4/2001 | 1838 | 1755.443466 |
| 3/4/2001 | 2132 | 2057.065563 |

## 7.2. FORECASTED  DATASET :

Here we are forecasting next 12 months Ridership data of AMTRAK COMPANY i.e., from **April 2001** to **March 2002**.

| Month | Sarimax_forecasting |
|---|---|
| 04-04-2001 | 2047.294792 |
| 05-04-2001 | 2081.020608 |
| 06-04-2001 | 2043.705778 |
| 07-04-2001 | 2149.672113 |
| 08-04-2001 | 2192.82395 |
| 09-04-2001 | 1875.821412 |
| 10-04-2001 | 2006.415183 |
| 11-04-2001 | 1988.195362 |
| 12-04-2001 | 2025.579575 |
| 01-04-2002 | 1799.407344 |
| 02-04-2002 | 1758.548046 |
| 03-04-2002 | 2050.213024 |

## 10. References:

**Forecasting: Principles and Practice**
*Rob J Hyndman and George Athanasopoulos*
Monash University, Australia
https://otexts.com/fpp2/
This book gives a good overview of different components of a time series analysis using practical examples and R codes.

**RDocumentation**
https://www.rdocumentation.org/packages/forecast/versions/8.15/topics/Arima
This online RDocumentation is very handy to know the nitty-gritty details of different R packages.