A PROJECT REPORT

ON

"A STATISTICAL ANALYSIS ON RAILWAY ELECTRIFICATION & FREIGHT MOVEMENT"

UNDER

SDG 9 – "INDUSTRY, INNOVATION AND INFRASTRUCTURE"

SUBMITTED TO

THE DEPARTMENT OF STATISTICS

VEER NARMAD SOUTH GUJARAT UNIVERSITY, SURAT, GUJARAT

IN PARTIAL FULLFILLMENT OF DEGREE OF

MASTER OF SCIENCE IN APPLIED STATISTICS

SUBMITTED BY

MANDAVIYA PRACHI HIMATBHAI	[SEM-IV, ROLL NO: 18]
PAREKH PRASHANT CHETANKUMAR	[SEM-IV, ROLL NO: 22]
TADHAWALA DWARIKESH KIRANKUMAR	[SEM-IV, ROLL NO: 33]
UCHADIYA POOJA PRAVINBHAI	[SEM-IV, ROLL NO: 35]

UNDER THE GUIDENCE OF

Mr. MAHAMMADZUBERAZA S. PATEL

APRIL – 2024

Re-Accredited 'B++' 2.86 CGPA by NAAC
VEER NARMAD SOUTH GUJARAT UNIVERSITY
University Campus, Udhna-Magdalla Road, SURAT - 395 007, Gujarat, India.
વીર નર્મદ દક્ષિણ ગુજરાત યુનિવર્સિટી
યુનિવર્સિટી કેમ્પસ, ઉધના-મગદલ્લા રોડ, સુરત - ૩૯૫ ૦૦૭, ગુજરાત, ભારત.
Tel : +91 - 261 - 2227141 to 2227146, Toll Free : 1800 2333 011, Digital Helpline No.- 0261 2388888
E-mail : info@vnsgu.ac.in, Website : www.vnsgu.ac.in

Certificate

DEPARTMENT OF STATISTICS
VEER NARMAD SOUTH GUJARAT UNIVERSITY, SURAT

This is to certify that the project report entitled "A STATISTICAL ANALYSIS ON RAILWAY ELECTRIFICATION & FREIGHT MOVEMENT", under ***SDG Goal 9: Industry, Innovation & Infrastructure****, submitted to the Department of Statistics, V.N.S.G.Uni., Surat, Gujarat, India, in partial fulfilment of the degree of M. Sc. (Applied Statistics) is a record of work carried out by "MANDAVIYA PRACHIBEN HIMATBHAI (Roll No. 18), PAREKH PRASHANT CHETANKUMAR (Roll No. 22), TADHAWALA DWARIKESH KIRANKUMAR (Roll No. 33), UCHADIYA POOJA PRAVINBHAI (Roll No. 35)" students of M. Sc. Applied Statistics (Semester-IV) for the academic year 2023-24 under my supervision and guidance.*

All sources of information/data have been duly acknowledged. No part of their analysis work has been submitted elsewhere for the award of any other degree.

Place: - Surat
Date:- 23 - 03 - 2024

(Mahammadzuber S. Patel)
Supervising Teacher
Department of Statistics,
V.N.S.G.Uni., Surat

(A. J. Rajyaguru)
Professor and Head
Department of Statistics,
V.N.S.G.Uni., Surat

DEPARTMENT OF STATISTICS * SURAT * Veer Narmad South Gujarat University

THIS PROJECT IS

DEDICATED TO

OUR FAMILY,

OUR DEPARTMENT OF

STATISTICS,

OUR GUIDE,

ALL OUR PROFESSORS,

OUR GROUP

AND

OUR FRIENDS

ACKNOWLEDGEMENT

We would like to thank our head of the Department of Statistics & all the faculties of the Department of Statistics, VNSGU, Surat. The department has helped us grow in so many ways & giving us an opportunity to work on this project was a different journey altogether, it gave us an insight to the world outside our textbooks. It is our great pleasure to acknowledge our thanks to every intellectual & professional personnel who helped us for our completion of project successfully.

We feel proud to acknowledge the kind support & essential guidance imparted by Dr. Arti J. Rajyaguru, Mr. Mahammadzuberaza.S. Patel & also our classmates. We specially would like to thank them, who are such a knowledgeable & experienced personality in this field, for spending time for our project.

- MANDAVIYA PRACHI H. [ROLL NO : 18]
- PAREKH PRASHANT C. [ROLL NO : 22]
- TADHAWALA DWARIKESH K. [ROLL NO : 33]
- UCHADIYA POOJA P. [ROLL NO : 35]

DECLARATION

We, the students of M.Sc. Applied Statistics Sem-IV, Department of Statistics, VEER NARMAD SOUTH GUJARAT UNIVERSITY, SURAT, hereby declare that we have completed our project, entitled "A statistical analysis on railway electrification and freight movement" Under SDG 9 - "INDUSTRY, INNOVATION AND INFRASTRUCTURE" in the academic year 2022-2024. The information here is true & original to the best of our knowledge. We also declare that the report submitted here will not be utilized for any degree

- MANDAVIYA PRACHI H. _____ [ROLL NO : 18]
- PAREKH PRASHANT C. _____ [ROLL NO : 22]
- TADHAWALA DWARIKESH K. _____ [ROLL NO : 33]
- UCHADIYA POOJA P. _____ [ROLL NO : 35]

INDEX


Introduction

Introduction

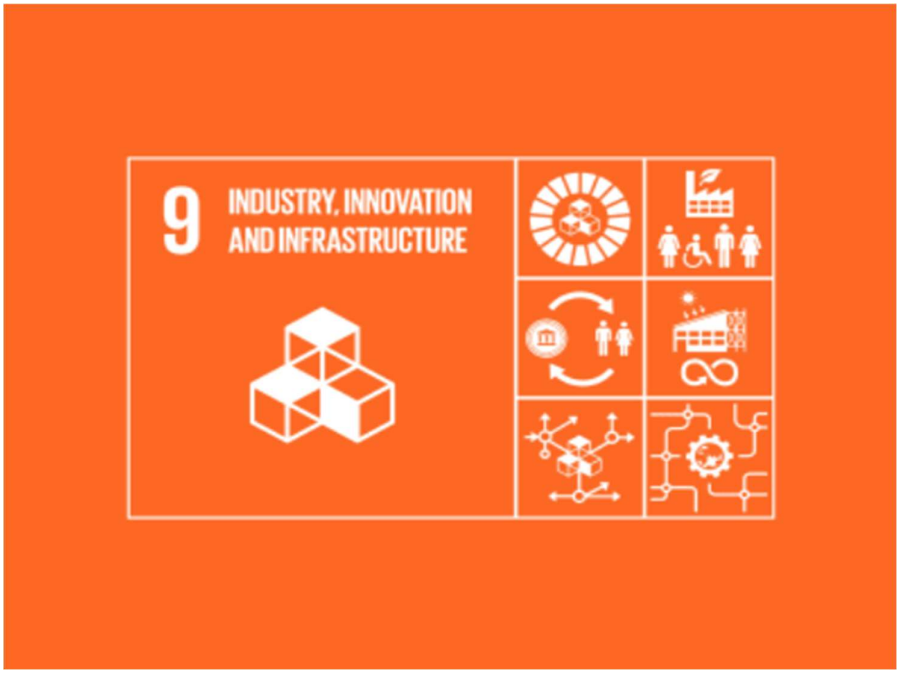
- The 2012 United Nations Conference on Sustainable Development in Rio de Janeiro, Member states decided to launch a process to develop a set of Sustainable Development Goals (SDG's).

- On 25th Sep 2015,193 member nations at the United Nations general assembly set up a collection of 17 Goals known as Sustainable Development Goals (SDGs) or simply Global Goals.

- This came into force on 1st Jan 2016. The 17 SDGs & their 169 targets from the core of "Agenda 2030".These goals are to be achieved by all member states by the Year 2030.

- While the SDGs are not legally binding Governments are expected to take ownership Ans establish national frame for the achievement of the 17 goals They are a Universal call to action to end poverty, protect the planet, & ensure all people enjoy peace & prosperity.

- The Goals are Broad, Independent & cover Social economic issues & making them work is everyone's responsibility.



Build Resilient Infrastructure, Promote Inclusive & Sustainable Industrialization & Foster Innovation

- Economic growth, social development & climate action are heavily dependent on investments in infrastructure, sustainable industrial development & technological progress. In the face of a rapidly changing global economic landscape & increasing inequalities, sustained growth must include industrialization that first of all, makes opportunities accessible to all people, & second, is supported by innovation & resilient infrastructure

- Goal 9 seeks to build resilient infrastructure, promote sustainable industrialization & foster innovation.

- India may aim to enhance regional connectivity by investing in infrastructure projects that promote cross-border trade, transportation, and communication. This could involve the development of transport corridors, energy grids, and digital networks that facilitate seamless connectivity with neighbouring countries and support regional economic integration.

- To promote goal 9 & to analyse infrastructure in Indian Railway domain, here, we analysed total Electrified & Non-Electrified Railway Roots in India.

- The Indian railway network, a vital component of the country's transportation infrastructure, is undergoing significant transformations to align with Sustainable Development Goal 9.

- One of the key initiatives is the electrification of railway routes, which not only reduces greenhouse gas emissions but also enhances operational efficiency and reduces dependence on fossil fuels.

- This project explores the electrification progress of railway routes, the development of passenger and freight networks, and the distribution of route kilometers across different zones, aiming to assess the sustainability and contribution of Indian railways to SDG 9.

TRANSPORTATION SECTOR

The transport sector in India is a crucial component of the country's economy & infrastructure, playing a vital role in facilitating trade, mobility, & connectivity. It encompasses various modes of transportation, including road, rail, air, & waterways. Here's an overview of each:

Road Transport:

- Road transport is the dominant mode of transportation in India, catering to the majority of passenger & freight movement
- India has one of the largest road networks globally, with a total length exceeding 5.8 million kilometres, including national highways, state highways, & rural roads. The sector includes various types of vehicles, ranging from two-wheelers, cars, buses to trucks & commercial vehicles.
- Road transport is often the quickest mode for short to medium-distance shipments within India. Delivery times can vary significantly depending on factors such as distance, road conditions, traffic congestion, & the efficiency of the transport service provider.

Air Transport:

- India's civil aviation sector has witnessed significant growth in recent years, with the emergence of low-cost carriers & increased air travel demand. The country has a vast network of domestic & international airports, with major hubs in cities like Delhi, Mumbai, Bangalore, & Hyderabad.
- Air transport is the fastest mode of freight transport but also the most expensive. It is typically used for high-value, time-sensitive, or perishable goods.

Waterways:

- India has an extensive network of rivers & coastal areas, offering potential for inland water transport.
- Water transport, including coastal shipping & inland waterways, is generally slower than road or rail transport but can be highly cost-effective for bulk cargo over long distances

Rail Transport:

- Indian Railways is one of the largest rail networks in the world, spanning over 67,000 kilometres & carrying millions of passengers & tons of freight daily
- It serves as a lifeline for both urban & rural areas, connecting remote regions to major cities & facilitating the movement of goods & people across the country.
- The government has been investing in modernizing rail infrastructure, including the introduction of high-speed trains, electrification, & the expansion of dedicated freight corridors
- Indian Railways operates an extensive network of freight trains, carrying bulk commodities such as coal, iron ore, cement, fertilizers, grains, & petroleum products. Under construction, DFCs (Dedicated Freight Corridors) are dedicated rail lines aimed at improving the efficiency & capacity of freight transportation between key economic.
- Rail transport is generally slower than road transport but is more cost-effective for long-distance freight movements over land.

In a nutshell there are several ways for freight transportation, but there are some limitations over them. But in Rail Transport there is only one limitation i.e. Rail Transport is slower than road transport, but government is working on that limitation to convert it into beneficial way with the use of projects on the infrastructure in rail transport like [Electrification of Railways, Introduction of High-Speed Freight Trains, Digitalization & Automation, Dedicated Freight Corridors].

TERMS USED IN STUDY:

- **Freight movement:**

  Freight movement refers to the goods or commodities that are transported from one location to another. It encompasses a wide range of products, including raw materials, finished goods, machinery and even livestock.

- **Passenger movement:**

  passenger transportation focuses on the movement of people from one location to another.

- **Passenger and freight tonne-kilometres**:

  Passenger and freight volumes are respectively measured in passenger-kilometres and tonne-kilometres, and broken down by mode of transport. These measures refer to the number of passengers or tonnes, multiplied by how many kilometres they were carried.

- **Route kilometre:**

  "Route kilometre" refers to the total length of railway lines or tracks on a particular route or section of the railway network. It measures the distance covered by the railway tracks from the starting point to the end point of a specific route.

- **Route kilometre electrified:**

  "Route kilometre electrified" refers to the total length of railway tracks or routes that have been electrified within the Indian Railways network. It represents the distance covered by railway lines where electric traction is used to power locomotives instead of traditional diesel locomotives. refers to the total length of railway tracks or routes that have been electrified within the Indian Railways network. It represents the distance covered by railway lines where electric traction is used to power locomotives instead of traditional diesel locomotives.

INDIAN RAILWAY

Indian Railways is one of the largest & busiest railway networks in the world. It is a state-owned enterprise operated by the Ministry of Railways, Government of India. The system spans the entire country, covering a vast network of tracks that connect major cities, towns, & even remote areas. Indian Railways operates both passenger & freight services, playing a crucial role in mass transportation & the movement of goods across the country. Indian Railway known for its diverse range of trains, from high-speed & luxury trains to local & suburban services. Indian Railways is a major employer, providing direct & indirect employment to a large number of people. The network has contributed significantly to the economic development & connectivity of India, serving as a lifeline for millions of passengers & supporting various industries through the transportation of goods. & because of this Indian Railway is Backbone of INDIA. To purify air of India & make India healthy we have to convert Non electrified Indian Railway into Electrified Indian Railway. There are Several advantages if we convert railway, it into electrified.

Transforming the Indian railway system into an electrical green energy-based network offers several advantages. Shifting to electric trains powered by green energy sources, such as solar or wind power, helps reduce the environmental impact associated with traditional diesel locomotives. Electric trains are generally more energy-efficient & produce lower emissions, contributing to India's efforts to combat air pollution & climate change.

Transforming the Indian railway system into an electrical green energy-based network offers several advantages. Shifting to electric trains powered by green energy sources, such as solar or wind power, helps reduce the environmental impact associated with traditional diesel locomotives. Electric trains are generally more energy-efficient & produce lower emissions, contributing to India's efforts to combat air pollution & climate change.

The adoption of green energy for railways aligns with global trends toward sustainable & eco-friendly transportation. It can significantly decrease the carbon footprint of the railway sector, making it a more environmentally friendly mode of transportation. This transition also contributes to India's commitment to promoting renewable energy & achieving a more sustainable & greener future.

Abbreviation name	Name of zone	Headquarters	Year of Establishment	Number of Divisions	States Covered
CR	Central Railway	Mumbai CST	1951	5	Maharashtra, parts of MP
ER	Eastern Railway	Kolkata	1952	4	West Bengal, Bihar
ECR	East Central Railway	Haji pur	1996	5	Bihar, Jharkhand
ECOR	East Coast Railway	Bhubaneswar	2003	3	Odisha, parts of AP
NR	Northern Railway	Delhi	1952	5	Delhi, UP, Haryana, Punjab, J&K, HP
NCR	North Central Railway	Prayagraj	2003	3	UP and MP

NER	North Eastern Railway	Gorakhpur	1952	3	Assam, Arunachal Pradesh, Nagaland, Manipur, Mizoram, Tripura
NWR	North Western Railway	Jaipur	2002	3	Rajasthan and parts of Gujarat
NFR	Northeast Frontier Rly	Guwahati	1958	5	Assam, Arunachal Pradesh, Nagaland, Manipur, Mizoram, Tripura, Sikkim
SR	Southern Railway	Chennai	1951	6	Tamil Nadu, Kerala, Karnataka, AP
SCR	South Central Railway	Secunderabad	1966	6	Andhra Pradesh, Telangana
SER	South Eastern Railway	Kolkata	1955	4	West Bengal, Jharkhand, Odisha
SECR	South East Central Rly	Bilaspur	1998	5	Chhattisgarh, parts of MP
SWR	South Western Railway	Hubballi	2003	3	Karnataka, parts of AP, Goa
WR	Western Railway	Mumbai Churchgate	1951	6	Maharashtra, Gujarat, MP
WCR	West Central Railway	Jabalpur	2003	3	Madhya Pradesh, parts of UP
SCR	South Coast Railway	Visakhapatnam	Yet to be notified	–	Andhra Pradesh, Odisha, Pondicherry, Tamil Nadu
METRO	Metro Railway Kolkata	Kolkata	1984	1	Urban rail in Kolkata

Objectives

- **To visualize the proportion of CO2 emissions generated by the transport sector relative to other sectors in India.**

- **To determine the average speed of diesel and electric trains.**

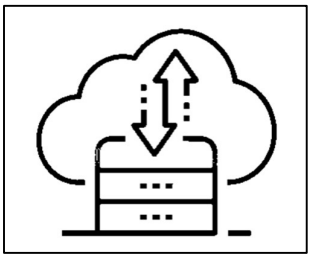
- **To check the significant difference between different railway zones regarding the level of electrification in different years.**

- **To check the significant difference between freight trains and passenger trains.**

- **To determine the progress of infrastructure in Indian railways.**
  - **To forecast total route kilo meters up to 2030.**
  - **To forecast electrified route kilo meters up to 2030.**
  - **To forecast non-electrified route kilo meters up to 2030.**

- **To determine the share of freight movement by different modes of transport.**
  - **Forecasting freight movement by railway up to the year 2030.**
  - **Forecasting freight movement by road up to the year 2030.**
  - **Forecasting freight movement by air up to the year 2030.**

Data collection



We have used Secondary Sources as our data for the Statistical Analysis, which are mentioned below:

- **https://indianrailways.gov.in/**
- **https://www.worldbank.org/en/home**
- **https://www.indiastat.com/**
- **https://www.iea.org/**

Methodology

"Statistics is the study of collecting, analysing, interpreting, presenting, & organizing data to make informed decisions or draw conclusions about a population."

Here, the brief introduction of statistical concepts of method which we used to analyse our data.

Descriptive Statistics

Descriptive statistics is a branch of statistics that involves the collection, summarization, & presentation of data in a meaningful & informative way, typically through measures such as central tendency, variability, & visual representations, to describe, understand & the main features of a dataset. We have used some descriptive statistics like Mean, Variance, Median, Standard Deviation, Shapiro wilks Statistics, P-Value, Charts.

Arithmetic Mean

The arithmetic mean, often simply called the mean, is a measure of central tendency that represents the average of a set of values. It is calculated by adding up all the values in a dataset & then dividing the sum by the total number of values.

For a dataset with n values x_1, $x_{2....}$, x_n the Arithmetic Mean is given by…

$$\bar{x} = \frac{x_1+x_2+...x_n}{n} = \frac{\sum_{i=1}^{n} xi}{n}$$

Chart

Once the data have been collected, the crucial problem becomes learning, whatever we can, from the data. Graph is a powerful tool of describing the dataset. A large dataset is required to be presented in graphical form that can capture the structure of underlying data. A quick glance at the picture elucidates the point easily then does a page filled with words & numbers.

The term charts as a visual representation of data have multiple meaning. A graphical representation is very useful to understand & data before any statistical analysis.

Here we plot Histogram & Box-plot to check Distribution of data & is there any outliers in the data or not.

- **Bar Charts:**
  In this chart type, data is represented in the form of bars. Height of bars represents the value or frequency of the class Distance between two bars and width of bars should remain constant in this chart type.

- **Pie Charts:**
  Pie chart is also called 'angular charts. A circle divided into portions that represent the relative frequencies or percentages of different categories or classes. This chart represents the value of the variable in the relative form of 360o. The area of 360o is divided into slices.

- **Column Charts:**
  A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally. A vertical bar chart is sometimes called a column chart.

- **Line charts**
  : A line chart or line plot or line graph or curve chart is a type of chart which displays information as a series of data points called 'markers' connected by straight line segments. It is a basic type of chart common in many fields

Normality Test

Shapiro wilks Statistics:

The Shapiro-wilk test is a test of normality. It was published in 1965 by Samuel Sanford Shapiro & Martin Wilk. The Shapiro-wilk test is used to check the normality of a random sample. i.e., to check if a random sample comes from a normal distribution or not. The test gives a W value; small values indicate that sample is not normally distributed.

Hypothesis:

H_0: Data is normal.
H_1: Data is not normal.

Test Statistics:

$$w = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Where, x(i) is the ordered sample values ai are constants generated from the means, variances & covariances of the sample of size n from a normally distributed sample.

Interpretation:

If the p-value is less than the chosen alpha level, then the null hypothesis is rejected at α % level of significance & there is evidence that the data is not normally distributed.

Exponential Smoothing

The construction of forecast function based on discounted past observations is most commonly carried out by exponential procedures. These procedures are attractive. That they allow the forecast function to be updated very easily every time a new observation become available. Forecasting procedures based on exponential smoothing have become popular, Since they are easy to implement & can be quite effective, they are implemented without using to a properly defined statistical model Suppose we wish to estimating the level of forecasting future observations, It is easily done by to put more weight on the most recent observations.

A new estimate is the combination of the estimate for the present tme period plus a portion of the random error $(x_t - \hat{x}_t)$ generated in the present time period

$(\widehat{x_{t+1}} = \hat{x}_t + \alpha(e_t)$ **……..(1)**

This equation is usually written as $S_t = S_{t-1} + \alpha(x_t - S_{t-1})$**…….(2)**

Where S_t= The new estimated or forecasted value for the next time period

S_{t-1} =The estimated or forecasted value for the present time period .

x_t = The actual data point in the present time period.

$(x_t - S_{t-1})$ = estimating or forecasting error for the present time period.

α = A weight or discount.

The smoothing equation can be written as

$\hat{x}_{t-1} = S_t = \alpha x_t + (1-\alpha) S_{t-1}$**…….(3)**

Or in more general form.

Next period forecast = weight(present period observation)+(1-weight)(present period forecast)

The smoothing equation is based on averaging past values of a series in decreasing manner. The observations are weighted, with more weight given to the more recent observations. The weights used are α for the most recent observation, α (1- α) for the next more recent α (1- α)^2 for the next & so on. At each time, the weighed observation along with the weighted estimate for the present period are combine to produce a new period forecast.

$\boldsymbol{\hat{x}_{t-1} = S_t = \alpha x_t + \alpha(1-\alpha)x_{t-1} + \alpha(1-\alpha)^2 x_{t-2} + \alpha(1-\alpha)^3 x_{t-3} + \ldots \alpha(1-\alpha)^{t-1} x_1}$ **……..(4)**

Brown's Exponential Smoothing Method

Brown's exponential smoothing method, also known as Brown's linear exponential smoothing or Brown's method, is a variation of exponential smoothing developed by Charles C. Holt and Peter Winters. This method is an extension of simple exponential smoothing that incorporates a trend component.

If time series contains linear trend with linear regression equation $\widehat{x_t} = a + bt$

where the estimates a & b represents intercept & slope of the model & t=1,2,3,4,….,n, the correct procedure would be Double Exponential Smoothing. The argument and techniques for the Double Exponential Smoothing method are similar in nature to that of Single Exponential Smoothing

If we let $\hat{x}_t = a_t + b_t T$

Represent the updated forecast, then $a_t = 2s_t^1 - s_t^2$ ……(5)

$$b_t = \frac{\alpha}{1-\alpha}(s_t^1 - s_t^2)$$

T = number of the time period ahead. The s_t^1 & s_t^2 are the single & double smoothing statistics found by applying then smoothing equation. $s_t^1 = \alpha x_t + (1-\alpha)s_{t-1}^1$ ……(7)

&

$$s_t^2 = \alpha x_t 0 + (1-\alpha)s_{t-1}^2 \quad \text{……(8)}$$

Once the single & double smoothed statistics are computed foe a while period. The values may be substituted into the updating formula for a=the intercept & slope. To start the double smoothing process, initial values of the smoothed estimates must be obtained, this can be done by substituting the values for the estimated intercept & the slope for the linear regression analysis into the following equations.

$s_0^1 = a - \left[\frac{1-\alpha}{\alpha}\right](b)$ ……..(9)

& $s_0^2 = a - \left[\frac{1-\alpha}{\alpha}\right](b)$ ……..(10)

DIAGNOSTIC CHECKING

Time Series modeling is an iterative procedure. To start model identification and parameter estimation, after estimation of parameters we have the assess model adequacy by detecting whether the model assumptions are satisfied. The basic assumption is that $\{e_t\}$ are white noise, that is $\{e_t\}$ are uncorrelated random shock with zero mean and constant variance. For any estimated model, the residual e_t's are estimates of uncorrelated white noise e_t's. Hence model diagnostic checking is accomplished through careful analysis of residual series $\{e_t\}$. Because this residual series is the product of parameter estimation. To check whether the errors are normally distributed, one can construct a histogram of standardized residual $\frac{\hat{Z}t}{\hat{\sigma}Z}$ and compare with the standard normal distribution using the chi-squared ô. goodness of fit test. To check whether the variance is constant, we can examine the plot of residuals, to check whether residuals are white noise, we can compute sample ACF and sample PACF of the residuals to see whether they do not form any pattern and all are statistically insignificant with in confidence limits. The sample ACFS, $\widehat{\rho k}$, of residual of required model, lie within these limits

$- Z_{\alpha/2} \frac{1}{\sqrt{T}} Z \leq \rho_k Z_{\alpha/2} \frac{1}{\sqrt{T}}$

Another useful test is the Portmanteau lack of fit test. This test uses all the residual sample ACF's a zero to check the Null Hypothesis

H0: $\rho_1 = \rho_2 = \rho_3 = \ldots, \rho_k - = 0$

H1: not so

with test statistic $Q = n(n+2)\sum_{j=1}^{K}(n-j)^{-1}\hat{\rho}_j{}^2$

This test statistic is the modified Q statistic originally proposed by Box and Pierce (1970). Under the Null hypothesis of model adequacy, Ljung and Box (1978) and Ansely and Newbold (1974) show that Q statistic approximately follows the $x^2{}_{K-m}$ chi-square distribution based on K-m degrees of freedom, where m is the number of parameters estimated in the model $Q \approx x^2{}_{K-m}$

H0 reject if $Q > x^2{}_{\propto,K-m}$ with a level of significance, based on the results of these residual analysis, if the present models can be derived.

Kruskal Wallis Test

The most widely used non-parametric technique for testing the null hypothesis that several samples have been drawn from the same or identical population is the Kruskal Wallis one way analysis of variance. The Kruskal Wallis test uses more information than the median test. As a consequence, The Kruskal Wallis test is usually more powerful, & is preferred when the available data measured on at least the ordinal scale.

Assumptions:

1) The data for analysis consist of K random samples of sizes $n_1,n_2,\ldots,n_k$
2) The observations are independent both within & among samples.
3) The variable of interest is continuous.
4) The measurement scale is at least ordinal
5) The populations are identical except for a possible difference in location for at least one population

Hypothesis:

H_0 The k population distribution functions are identical
H_1 The k populations do not have the same median

Test Statistics:

We may display the data available for analysis in a table & replace each original observation by its rank relative to all the observations in the k samples

$$\boldsymbol{H = \frac{H}{N(N+1)} \sum_{I=1}^{K} \frac{1}{n_i} \left[R_i - \frac{n_i(N+1)}{2} \right]^2}$$

If we let $\sum_{i=1}^{k} n_i$, be the total number of observations in the & samples, we assign the rank I to the smallest of these, the rank 2 to the next in size, and so on to the largest, which is given the rank N. In case of ties we assign the tied observations the average of the ranks that would be assigned if there were no ties.

If the null hypothesis is true, we expect the distribution of ranks over the groups to be a matter of chance, so that either the small ranks or the large ranks do not tend to be concentrated in one sample.

Therefore, if the null hypothesis is true, we expect the k sums of ranks to be about equal when adjusted for unequal sample sizes.

We Write the Kruskal Wallis Test statistics as given above where R, is the sum of the ranks assigned to observations in the i^{th} sample, and n(N+1)/2 is the expected sum of ranks for the ith treatment under H_0

Decision Rule:

We reject the H_0 if value of H >5% level of significance.

POST HOC TEST

Post hoc tests are used in statistics to compare multiple groups after a significant result has been obtained in an analysis of variance (ANOVA) or a similar test. These tests help identify specific differences between groups when the omnibus test indicates a significant difference but does not specify which groups differ from each other.

BONFERRONI TEST

Bonferroni based procedure is recommended when data is not continuous because this procedure have no distributional assumptions. The Bonferroni method applies to both continuous and discrete data. This method is flexible because it controls the FWE for tests of joint hypotheses about any subset of m separate hypotheses (including individual contracts). The procedure will reject a joint hypotheses H0 if any p – value for the individual hypotheses included in H0 is less than 0.05. Bonferroni method however, yields conservative bounds on Type I error hence it has low power. This procedure controls the FWE at without any further assumption on the dependence structure of the p – value.

Mann -Whitney Wilcoxon Test

A procedure for testing the null hypothesis of equal population location parameters was proposed by Mann& Whitney. The test is usually referred to as the Mann-Whitney test. The test is sometimes also referred to as the Mann-Whitney-Wilcoxon test. Mann& Whitney, who seem to have been the first to treat the case of unequal sample sizes, point out the relationship between their test statistic, as given below, & that of Wilcoxon.

Assumptions:

A) The data consist of a random sample of observations $x_1, x_2,\ldots,x_{n1}$, from population 1 with unknown median M_x , & another random sample of observations $y_1,y_2,\ldots,y_{n2}$ from population 2 with unknown median M_y.
B) The two samples are independent.
C) The variable observed is a continuous random variable.
D) The measurement scale employed is at least ordinal.
E) The distribution functions of the two populations differ only with respect to location, if they differ at all.

Hypothesis:

These hypotheses are appropriate only when assumption E is met. Either one of the following null hypotheses may be tested against the corresponding alternative.

A. (T wo-sided)

H_0: $M_x = M_y$
H_1: $M_x \neq M_y$

B. (One-sided)

H_0: $M_x \geq M_y$
H_1: $M_x < M_y$

C. (One-sided)

H_0: $M_x \leq M_y$
H_1: $M_x > M_y$

Test Statistic:

To compute the observed value of the test statistic, we combine the two samples& rank all sample observations from smallest to largest. We assign tied observations the mean of the rank positions they would have occupied had there been no ties 3. Then we sum the ranks of the observations from the population 1, If the location parameter of population 1 is smaller than the location parameter of population 2.

we expect (for equal sample sizes) the sum of the ranks of the sample observations from population 1 to be smaller than the sum of the ranks of the sample observations from population 2.

Similarly, if the location parameter of population 1 is larger than the location parameter of population 2, we expect just the reverse to be true.

The test statistic is based on this rationale in such a way that. depending on the null hypothesis, either a sufficiently small or a sufficiently large sum of ranks assigned to sample observations from population I causes us to reject the null hypothesis.

$$\boldsymbol{T = S - \frac{n_1(n_1 + 1)}{2}}$$

Where S is the sum of the ranks assigned to the sample observations from population 1.

Decision Rule:

The choice of a decision rule depends on the sample observations from population 1

A) We will reject H_0 if the computed value is less than $W\frac{\alpha}{2}$ or greater than $W1 - \frac{\alpha}{2}$

$$\boldsymbol{W_{1-\frac{\alpha}{2}} = n_1 n_2 - w_{\frac{\alpha}{2}}}$$

B) Reject H_0 if the computed T is less than $W\alpha$
c) Reject H_0 if the computed T is Greater than $W\alpha$

$$\boldsymbol{W_{1-\alpha} = n_1 n_2 - w_{\alpha}}$$

Time Series Analysis

A time series is a set of observations taken at specific times. In other words, a series of observation recorded over time is known as time series.

Utility of Time Series Analysis:

1) It gives a general description of the past behaviour of the series.
2) It helps in Forecasting the future behaviour on the basis of past behaviour.
3) It facilitates comparison.
4) It helps in the evaluation of current accomplishments.

ARIMA MODELS

ARIMA stands for Auto Regressive Integrated Moving Average.

The method has 3 variables to account for…

• p = Periods to lag for e.g.: (if P= 3 then we will use the three previous periods of our time series in the autoregressive portion of the calculation) P helps adjust the line that is being fitted to forecast the series Purely autoregressive models. resemble a linear regression where the predictive variables are P number of previous periods.

• d= In an ARIMA model we transform a time series into stationary one (series without trend or seasonality) using differencing. D refers to the number of differencing transformations required by the time series to get stationary.

• Stationary time series is when the mean & variance are constant over time. It is easier to predict when the series is stationary.

• Differencing is a method of transforming a non-stationary time series into a stationary one. This is an important step in preparing data to be used in an ARIMA model.

- The first differencing value is the difference between the current time period & the previous time period. If these values fail to revolve around a constant mean & variance then we find the second differencing using the values of the first differencing. We repeat this until we get a stationary series
- The best way to determine whether or not the series is sufficiently differenced is to plot the differenced series & check to see if there is a constant mean & variance.

• q= This variable denotes the lag of the error component, where error component is a part of the time series not explained by trend or seasonality.

Autocorrelation function plot (ACF):

Autocorrelation refers to how correlated a time series is with its past values whereas the ACF is the plot used to see the correlation between the points, up to & including the lag unit. In ACF, the correlation coefficient is in the x-axis whereas the number of lags is shown in the y-axis.

Partial Autocorrelation Function plots (PACF):

A partial autocorrelation is a summary of the relationship between an observation in a time series with observations at prior time steps with the relationships of intervening observations removed.

• The partial autocorrelation at lag k is the correlation that results after removing the effect of any correlations due to the terms at shorter lags.

• If the PACF plot drops off at lag n, then use an AR(n) model & if the drop in PACF is more gradual than we use the MA (n)term. & in both ACF & PACF lag is significant than we fit ARIMA model with the significant lag.

Final steps:

Step 1. Identification: That is, find out the appropriate values of p, d, & q. We will show shortly how the correlogram & partial correlogram aid in this task.

Step 2. Estimation: Having identified the appropriate p & q values, the next stage is to estimate the parameters of the autoregressive & moving average terms included in the model. Sometimes this calculation can be done by simple least squares but sometimes we will have to resort to nonlinear (in parameter) estimation methods. Since this task is now routinely handled by several statistical packages, we do not have to worry about the actual mathematics of estimation; the enterprising student may consult the references on that.

Step 3. Diagnostic checking: Having chosen a particular ARIMA model, & having estimated its parameters, we next see whether the chosen model fits the data reasonably well, for it is possible that another ARIMA model might do the job as well. This is why Box–Jenkins ARIMA modelling is more an art than a science; considerable skill is required to choose the right ARIMA model. One simple test of the chosen model is to see if the residuals estimated from this model, are white noise; if they are, we can accept the particular fit; if not, we must start over. Thus, the BJ methodology is an iterative process.

Step 4. Forecasting: One of the reasons for the popularity of the ARIMA modelling is its success in forecasting. In many cases, the forecasts obtained by this method are more reliable than those obtained from the traditional econometric modelling, particularly for short-term forecasts. Of course, each case must be checked.

Akaike Information Criterion (AIC):

Akaike Information Criterion (AIC) is a statistical measure that provides a way to compare the goodness of fit between different models for a given set of data. It is a relative measure of the quality of a statistical model that is used to select the best model among a set of competing models.

The AIC value is based on the likelihood function & the number of parameters in the model. The AIC value is defined as:

$$\mathbf{AIC = -2 * \log(L) + 2 * k}$$

where L is the maximum likelihood estimate of the model, & k is the number of parameters in the model. The AIC value penalizes the number of parameters in the model, so a model with fewer parameters will have a lower AIC value, indicating a better fit to the data.

The AIC criterion can be used to compare different models & select the model that has the lowest AIC value. The model with the lowest AIC value is considered the best model among the competing models, & is chosen as the final model.

AIC is widely used in statistical modelling & machine learning, particularly in linear regression, logistic regression, & time series analysis. It is a popular tool for model selection & helps to prevent overfitting by balancing the goodness of fit with the complexity of the model.

Regression Analysis

Regression analysis is concerned with the study of the dependence of one variable, the dependent variable, on one or more other variables, the explanatory variables, with a view to estimating &/or predicting the (population) mean or average value of the former in terms of the known or fixed (in repeated sampling) values of the latter.

Regression Model:

A regression model is a statistical technique used to examine the relationship between a dependent variable and one or more independent variables. In the book, Gujarati discusses the linear regression model extensively, where the relationship between the variables is assumed to be linear. Simple linear regression is the simplest form of regression analysis where there is only one independent variable and one dependent variable. The relationship between the two variables is expressed by a straight line

Multiple Regression:

In multiple regression, there are two or more independent variables that are used to predict the dependent variable. The relationship is expressed by a linear equation with multiple coefficients.

Assumptions of Regression

1. The variance of the error term ε is constant for all values of the independent variables $\text{Var}(\varepsilon|X1,X2,\ldots,Xk)=\sigma^2$ & it indicates Homoscedasticity. If this assumption violates than it indicates Heteroscedasticity.

2. The expected value of the error term ε conditional on the values of the independent variables is zero:
   $E(\varepsilon|X1,X2,\ldots,Xk)=0$ This shows there is no multicollinearity. If this assumption violates than it indicates Multicollinearity.

3. The error terms εi are uncorrelated across observations, meaning that there is no correlation between the error terms for different observations: $\text{Cov}(\varepsilon i,\varepsilon j)=0$ for all $i\neq j$.

   There are remedies to remove these problems from the regression analysis

AUTOCORRELATION:

Detection methods:

1. Durbin Watson d Test:

The most celebrated test for detecting serial correlation is that developed by statisticians Durbin and Watson. It is popularly known as the Durbin– Watson d statistic, which is defined as

$$d = \frac{\sum_{t=2}^{n} (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{n} \hat{u}_t^2}$$

This is simply the ratio of the sum of squared differences in successive residuals to the RSS. Note that in the numerator of the d statistic the number of observations is n− 1 because one observation is lost in taking successive differences. A great advantage of the d statistic is that it is based on the estimated residuals, which are routinely computed in regression analysis. Because of this advantage, it is now a common practice to report the Durbin–Watson d along with summary measures, such as R^2 , adjusted R^2, t, and F. Although it is now routinely used, it is important to note the assumptions underlying the d statistic.

1. The regression model should include the intercept term. If it is not present, as in the case of the regression through the origin, it is essential to rerun the regression including the intercept term to obtain the RSS.
2. The explanatory variables, the X's, are non-stochastic, or fixed in repeated sampling.
3. The disturbances u_t are generated by the first-order autoregressive scheme: $u_t = \rho\, u_{t-1} + \varepsilon_t$. Therefore, it cannot be used to detect higher order autoregressive schemes.
4. The error term u_t is assumed to be normally distributed.
5. The regression model should not include the lagged value(s) of the dependent variable as one of the explanatory variables. Thus, the test is inapplicable for models of the following type:
   $Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \cdots + \beta_k X_{kt} + \gamma Y_{t-1} + u_t$, where Y_{t-1} is the one period lagged value of Y. Such models are known as autoregressive models.
6. There are no missing observations in the data. Thus, in our regression for the period 1959–1998, if observations for, say, 1978 and 1982 were missing for some reason, the d statistic makes no allowance for such missing observations.

The exact sampling or probability distribution of the d statistic is difficult to derive because, as Durbin and Watson have shown, it depends in a complicated way on the X values present in a given sample. This difficulty should be understandable because d is computed from u_t , which are, of course, dependent on the given X's. Therefore, unlike the t, F, or χ^2 tests, there is no unique critical value that will lead to the rejection or the acceptance of the null hypothesis that there is no first-order serial correlation in the disturbances u_t. However, Durbin and Watson were successful in deriving a lower bound dL and an upper bound d_U such that if the computed d is outside these critical values, a decision can be made regarding the presence of positive or negative serial correlation. Moreover, these limits depend only on the number of observations n and the number of explanatory variables and do not depend on the values taken by these explanatory variables. These limits, for n going from 6 to 200 and up to 20 explanatory variables, have been tabulated by Durbin and Watson and are reproduced in Appendix D, Table D.5 of D. Gujarati's book.

The actual test procedure can be explained better with the aid of Figure 5, which shows that the limits of d are 0 and 4. These can be established as follows. Expand the formula of d to obtain

$$d = \frac{\sum_{t=2}^{n} \hat{u}_t^2 + \sum_{t=2}^{n} \hat{u}_{t-1}^2 - 2\sum_{t=2}^{n} \hat{u}_t \hat{u}_{t-1}}{\sum_{t=1}^{n} \hat{u}_t^2}$$

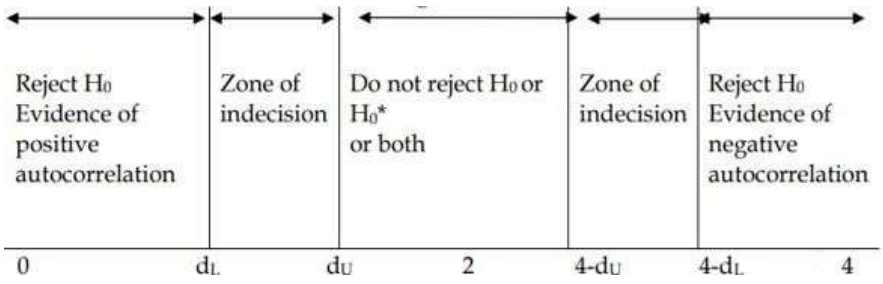
Since $\sum_{t=2}^{n} \hat{u}_t^2$ **and** $\sum_{t=2}^{n} \hat{u}_{t-1}^2$ differ in only one observation, they are approximately equal. Therefore, setting $\sum_{t=2}^{n} \hat{u}_t^2 \approx \sum_{t=2}^{n} \hat{u}_{t-1}^2$, **we get,**

$$d \approx 2\left(1 - \frac{\sum_{t=2}^{n} \hat{u}_t \hat{u}_{t-1}}{\sum_{t=1}^{n} \hat{u}_t^2}\right)$$

The mechanics of the Durbin–Watson test are as follows, assuming that the assumptions underlying the test are fulfilled:

1. Run the OLS regression and obtain the residuals.
2. Compute d.
3. For the given sample size and given number of explanatory variables, find out the critical d_L and d_U values.
4. Now follow the decision rules given in Table 5 for ease of reference, these decision rules are also depicted in the following Figure 5. Here we test H_0: No

positive autocorrelation i.e. H_0: □ = 0 against H^*_0: No negative autocorrelation i.e. H_1 : □ □ 0.



A General Test of Autocorrelation:

2. The Breusch– Godfrey (BG) Test:

To avoid some of the pitfalls of the Durbin–Watson d test of autocorrelation, statisticians Breusch and Godfrey have developed a test of autocorrelation that is general in the sense that it allows for (1) non stochastic regressors, such as the lagged values of the regress and; (2) higher-order autoregressive schemes, such as AR(1), AR(2), etc.; and (3) simple or higher order moving averages of white noise error terms, such as ε_t, $u_t = \rho ut{-}1 + \varepsilon_t$, -1 < □ <1. The **BG test**, which is also known as the **LM** test, proceeds as follows: We use the two-variable regression model to illustrate the test, although many regressors can be added to the model. Also, lagged values of the regressand can be added to the model.

Let Yt= β1+β2X + μt

Assume that the error term μt follows the p^{th}-order autoregressive, AR(p), scheme as follows:

ut = ρ1μt-1 + ρ2μt-2 + …. + ρpμt-p +εt

Here ε_t is a white noise error term. As you will recognize, this is simply the extension of the AR(1) scheme. The null hypothesis H0 to be tested is that

H0: ρ1=ρ2=…… .= ρp=0

That is, there is no serial correlation of any order. The BG test involves the following steps:

1. Estimate by OLS and obtain the residuals $\hat{u}_t$.

2. Regress $\hat{u}_t$ on the original X_t (if there is more than one X variable in the original model, include them also) and $\hat{u}_{t-1}, \hat{u}_{t-2}, \ldots, \hat{u}_{t-p}$, where the later are the lagged values of the estimated residuals in step 1. Thus, if p = 4, we will introduce four lagged values of the residuals as additional regressor in the

model. Note that to run this regression we will have only (n − p) observations. In short, run the following regression:

$$\hat{u}_t = \alpha_1 + \alpha_2 X_t + \hat{\rho}_1 \hat{u}_{t-1} + \hat{\rho}_2 \hat{u}_{t-2} + \cdots + \hat{\rho}_p \hat{u}_{t-p} + \varepsilon_t$$

and obtain R2 from this (auxiliary) regression.

3. If the sample size is large (technically, infinite), Breusch and Godfrey have shown that

$$(n-p)R^2 \sim \chi^2_p$$

That is, asymptotically, n−p times the R^2 value obtained from the auxiliary regression follows the chi- square distribution with p df. If in an application, (n −p) R^2 exceeds the critical chi-square value at the chosen level of significance, we reject the null hypothesis, in which case at least one ρ is statistically significantly different from zero.

The following practical points about the BG test may be noted:

1. The regressors included in the regression model may contain lagged values of the regress and Y, which is, Yt−1, Yt−2, etc., may appear as explanatory variables. Contrast this model with the Durbin– Watson test restriction that there be no lagged values of the regress and among the regressors.
2. As noted earlier, the BG test is applicable even if the disturbances follow a pth-order moving average (MA) process, that is, the ut are generated as follows: $u_t = \varepsilon_t + \lambda_1\varepsilon_{t-1} + \lambda_2\varepsilon_{t-2} + \cdots + \lambda_p\varepsilon_{t-p}$ Here ε_t is a white noise error term, that is, the error term that satisfies all the classical assumptions.
3. If p = 1, meaning first-order autoregression, then the BG test is known as Durbin's M test.

Autocorrelation function plot (ACF):

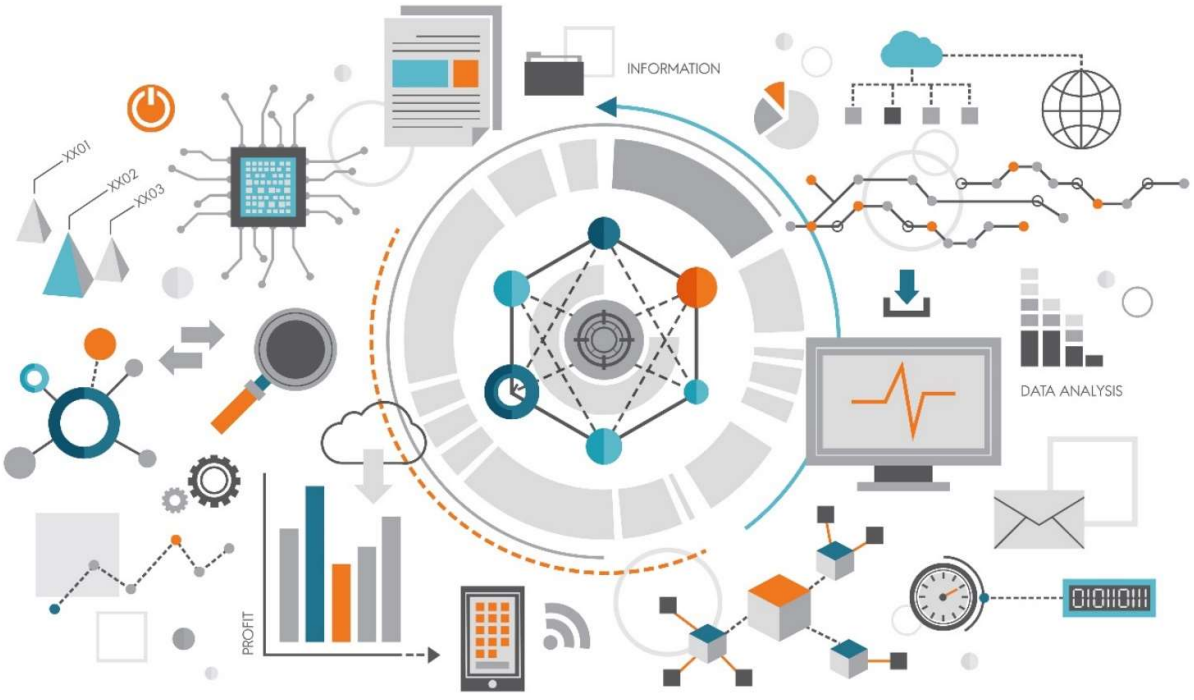
Autocorrelation refers to how correlated a time series is with its past values whereas the ACF is the plot used to see the correlation between the points, up to and including the lag unit. In ACF, the correlation coefficient is in the x-axis whereas the number of lags is shown in the y-axis.

▪ The Autocorrelation function plot will let you know how the given time series is correlated with itself **Partial Autocorrelation Function plots (PACF):**

A partial autocorrelation is a summary of the relationship between an observation in a time series with observations at prior time steps with the relationships of intervening observations removed.

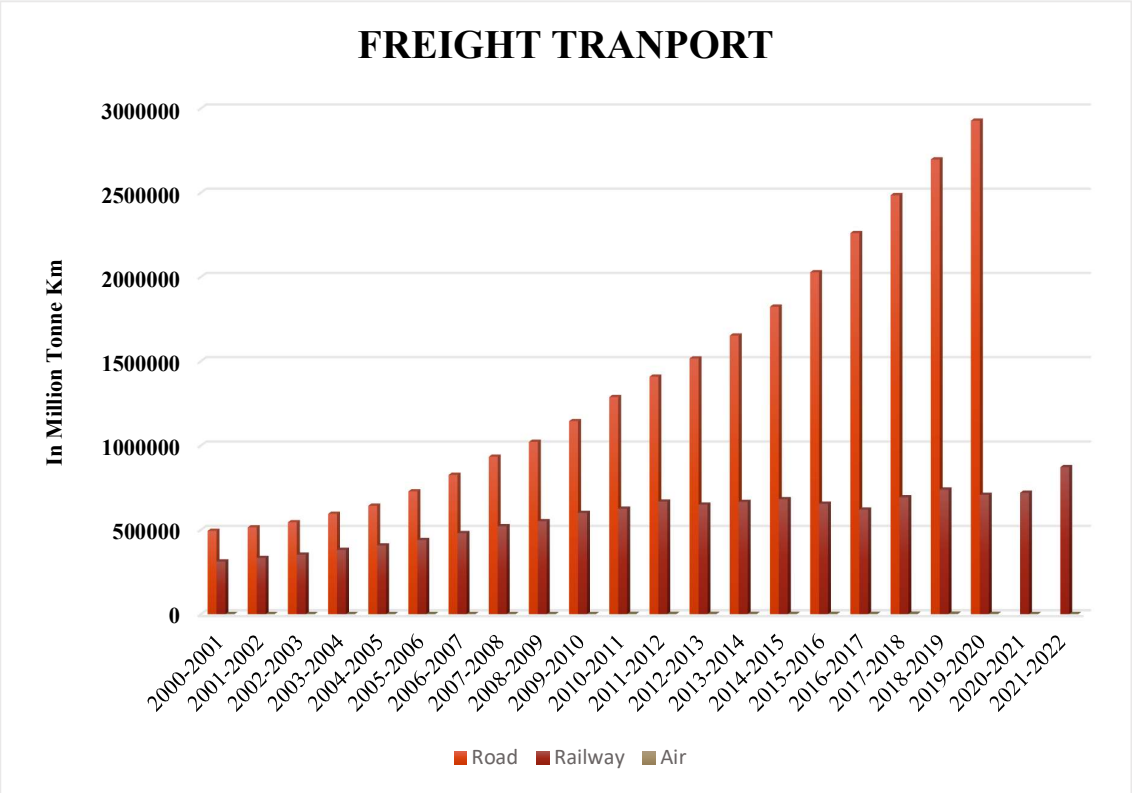
- The partial autocorrelation at lag k is the correlation that results after removing the effect of any correlations due to the terms at shorter lags.

Analysis
&
Interpretation



Graphical Representation

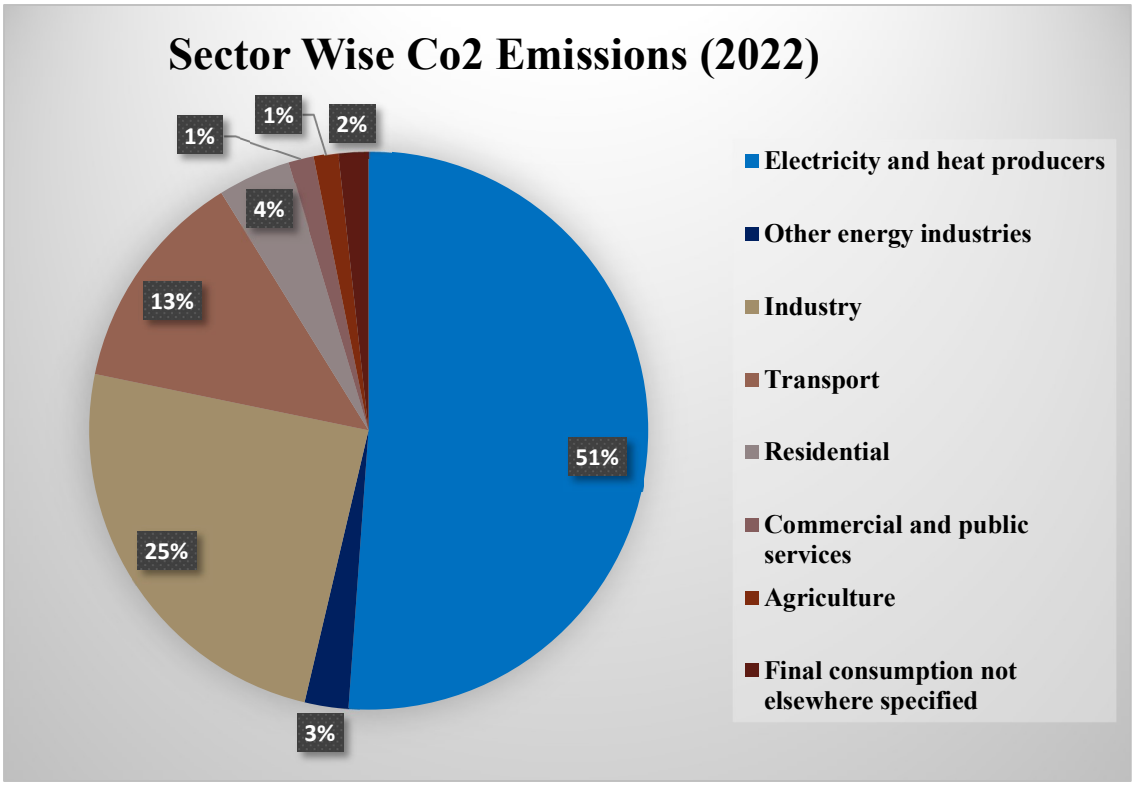
Freight transport by different modes of transport:



- **Interpretation:**

  From the above graph we can see that freight transport by road is constantly increases whereas freight transport by railway is nearer to constant & freight transport by air is very much low as compare to road & rail.

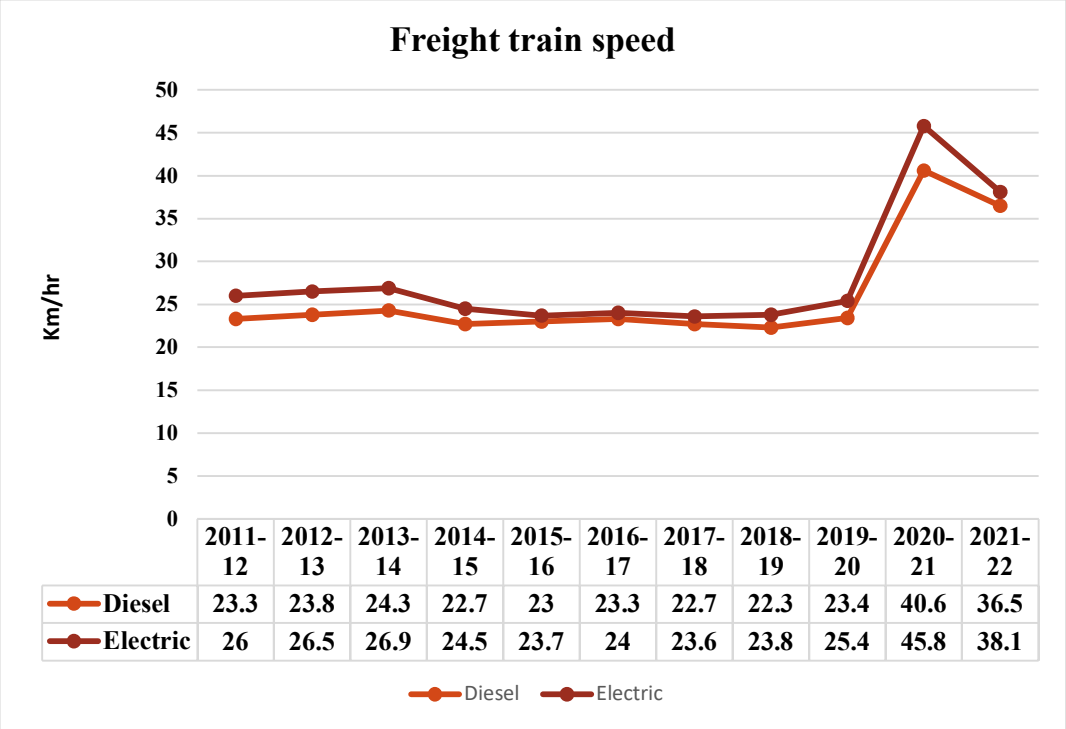
Sector wise co2 Emissions:



- **Interpretation**

  From the above given Pie Chart, we can say that Electricity & Heat producers produce highest Co2 emission (51%) and agriculture contribute lowest Co2 emission (1%) and transport sector produce 13% Co2 emission .

Average speed of diesel and electric locomotives :

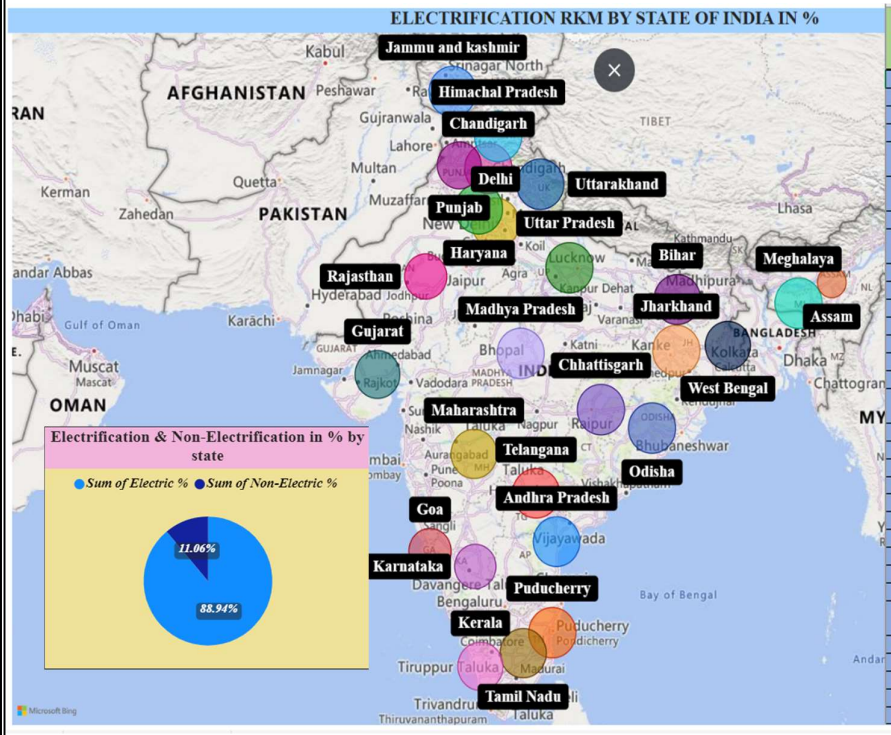


	2011-12	2012-13	2013-14	2014-15	2015-16	2016-17	2017-18	2018-19	2019-20	2020-21	2021-22
Diesel	23.3	23.8	24.3	22.7	23	23.3	22.7	22.3	23.4	40.6	36.5
Electric	26	26.5	26.9	24.5	23.7	24	23.6	23.8	25.4	45.8	38.1

Interpretation:

From the above graph we can say that the on an average speed of Electrified Freight train is higher than the diesel Freight train.

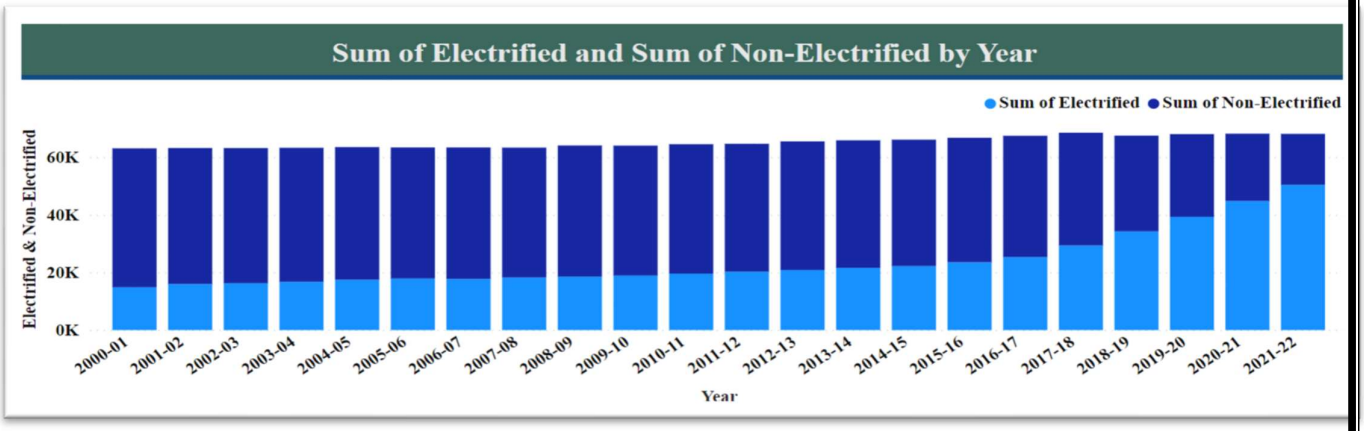
State wise electrified route kilometres (%):



S.N.	State	Total BG RKM	Electrification in %	Non-Electrification in %
1	Chandigarh	16	100	0
2	Chhattisgarh	1199	100	0
3	Delhi	183	100	0
4	Haryana	1701	100	0
5	Himachal Pradesh	67	100	0
6	Jammu and kashmir	298	100	0
7	Jharkhand	2558	100	0
8	Madhya Pradesh	4822	100	0
9	Meghalaya	9	100	0
10	Odisha	2822	100	0
11	Puducherry	21	100	0
12	Telangana	1858	100	0
13	Uttar Pradesh	8482	100	0
14	Uttarakhand	347	100	0
15	Bihar	3710	97.41239892	2.5876
16	Andhra Pradesh	3965	96.87263556	3.12736
17	Tamil Nadu	3854	94.94032174	5.05968
18	Maharashtra	5760	94.46180556	5.53819
19	West Bengal	4047	90.98097356	9.01903
20	Kerala	1047	90.44890162	9.5511
21	Gujarat	3862	88.94355256	11.0564
22	Punjab	2253	84.99778074	15.0022
23	Rajasthan	5623	78.01885115	21.9811
24	Goa	189	77.77777778	22.2222
25	Karnataka	3836	74.13972888	25.8603
26	Assam	2518	31.81096108	68.189
27	Arunachal Pradesh	12	0	0
28	Manipur	13	0	0
29	Mizoram	2	0	0
30	Nagaland	11	0	0
31	Sikkim	-	0	0
32	Tripura	265	0	0

- **Interpretation:**

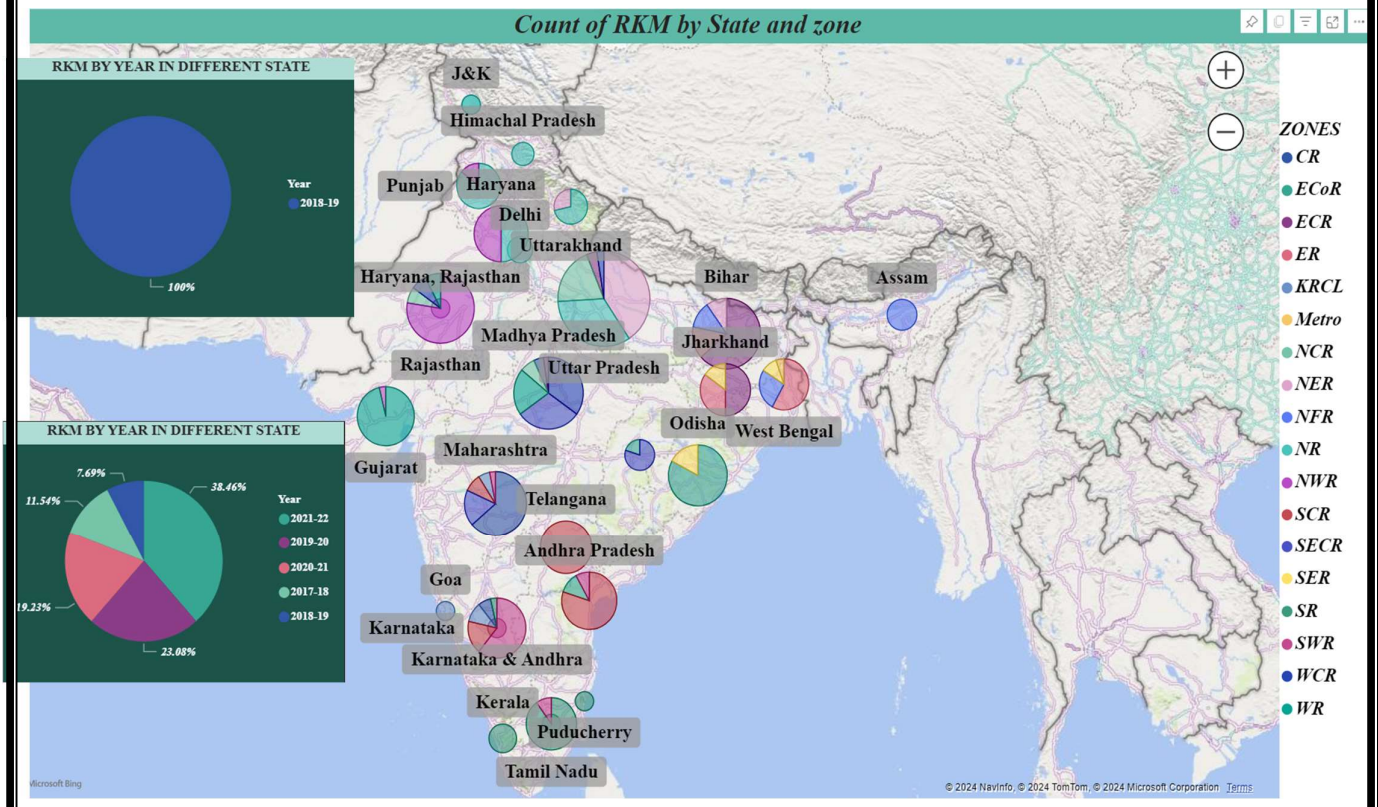
From the Bubble Map, we can observe which states have completely achieved the target of electrified railway routes and which states have not achieved that target by 30-06-2023. In the table above, states like Chandigarh, Chhattisgarh, and Delhi have achieved the 100% electrification target. Additionally, Bihar, Andhra Pradesh, and Maharashtra are leading in completing electrification in the railway sector. Some states are progressing slowly in electrification in the railway sector, such as Gujarat, Karnataka, Goa, and Assam.
The Pie Chart indicates that 21.06% of non-electrification remains, while 88.94% of electrification has been completed by the Indian government in the Gujarat Railway Sector, as shown in the Bubble Map above.



Interpretation:

From the above chart, we can see that electrification of route kilometers is increasing year by year.

Zone wise electrified route kilometres:Interpretation



Interpretation:

From the above pie map we can see the electrification of different zones in different years .

Statistical Analysis

To Study The Difference Between Different Railway Zones

Normality Test

- **Hypothesis:**

H_0:Data follows normal distribution.
H_1:Data does not follow normal distribution.

Tests of Normality

	Shapiro-Wilk		
	Statistic	Df	Sig.
ECOR	.887	7	.258
WCR	.929	7	.539
METRO	.453	7	.000
SER	.810	7	.052
SECR	.941	7	.651
ER	.865	7	.167
NCR	.961	7	.824
ECR	.929	7	.542
CR	.779	7	.025
NR	.952	7	.750
NER	.783	7	.027
SR	.680	7	.002
WR	.804	7	.044
SCR	.873	7	.199
KRCL	.738	7	.009
NWR	.866	7	.172
SWR	.879	7	.222
NFR	.864	7	.163

* This is a lower bound of the true significance.
a Lilliefors Significance Correction

- **Conclusion:**

Here, for some variables p-value < 0.05 so we do reject H_0 at 5 % level of significance. Therefore, we conclude that majority of variables does not follows normal distribution.

Kruskal Wallis Test

- **Hypothesis:**

H_0: There is no significant difference between zones regarding electrification level.

H_1: At least one zone is differed from other zones.

Ranks

	Id	N	Mean Rank
Zones	ECOR	7	59.43
	WCR	7	60.00
	METRO	7	13.14
	SER	7	32.00
	SECR	7	51.86
	ER	7	61.00
	NCR	7	62.64
	ECR	7	78.50
	CR	7	77.43
	NR	7	106.64
	NER	7	86.79
	SR	7	69.43
	WR	7	70.86
	SCR	7	80.57
	KRCL	7	37.93
	NWR	7	87.21
	SWR	7	60.93
	NFR	7	46.64
	Total	126	

Test Statistics(a,b)

	Rate
Chi-Square	44.383
df	17
Asymp. Sig.	.000

a Kruskal Wallis Test

- **Conclusion:**

Here, p-value < 0.05 so we reject H_0 at 5 % level of significance & conclude that at least one zone is differed from the other.

Post-hoc Test: Bonferroni

- **Hypothesis:**

  H_0: There is no significant difference between any two zones.
  H_1: There is significant difference between any two zones.

METRO	
ECOR	1.0000
WCR	1.0000
SER	1.0000
SECR	1.0000
ER	1.0000
NCR	1.0000
ECR	0.4980
CR	1.0000
NR	**0.0001**
NER	**0.0481**
SR	1.0000
WR	0.2530
SCR	0.0561
KRCL	1.0000
NWR	**0.0083**
SWR	1.0000
NFR	1.0000

NR	
ECOR	0.1022
WCR	0.1869
METRO	**0.0001**
SER	**0.0009**
SECR	**0.0197**
ER	0.0556
NCR	0.1327
ECR	1.0000
CR	0.7174
NER	1.0000
SR	0.4353
WR	1.0000
SCR	1.0000
KRCL	**0.0045**
NWR	1.0000
SWR	0.2495
NFR	**0.0163**

- **Interpretation:**

  From the above p-value we can say that there is significant different between METRO & NR or METRO & NER or METRO & NER or SER & NR or SECR & NR or NR & KRCL or NR & NFR.

To Study the Difference Between Freight Train & Passenger Train

Normality Test

- **Hypothesis:**

H_0: Data follows normal distribution.
H_1: Data does not follow normal distribution.

Tests of Normality

	Shapiro-Wilk		
	Statistic	df	Sig.
Passenger	.898	22	.027
Freight	.941	22	.208

a Lilliefors Significance Correction

- **Conclusion:**

Here, for passenger p-value < 0.05 so we do not accept at 5 % level of significance.

Mann–Whitney U test

- **Hypothesis:**

H_0: There is insignificant difference between volume of passengers and freight train.

H_1: There is significant difference between volume of passengers and freight train.

Ranks

	Index1	N	Mean Rank	Sum of Ranks
trans1	passenger	22	27.82	612.00
	freight	22	17.18	378.00
	Total	44		

Test Statistics(a)

	trans1
Mann-Whitney U	125.000
Wilcoxon W	378.000
Z	-2.746
Asymp. Sig. (2-tailed)	.006

- **Conclusion:**

Here, P -value < 0.05 so we reject at 5 % level of significance and conclude that There is significant difference between Volume of passenger train and freight train.

Regression Analysis

Railway Total Route Km

- **Model 1**
- **Descriptive Statistics**

Descriptives

			Statistic	Std. Error
total	Mean		65192.7727	421.95501
	95% Confidence Interval for Mean	Lower Bound	64315.2692	
		Upper Bound	66070.2762	
	5% Trimmed Mean		65133.7677	
	Median		64530.0000	
	Variance		3917012.755	
	Std. Deviation		1979.14445	
	Minimum		63028.00	
	Maximum		68442.00	
	Range		5414.00	
	Interquartile Range		4065.50	
	Skewness		.432	.491
	Kurtosis		-1.482	.953

- **Interpretation:**

From the above table we observed that,
S.E. is high, Skewness is not near to 0(Zero) & Kurtosis is not 3.
So, we may say that, there are very less chance that data will follow normal distribution.

Normality Test

Hypothesis:

H0: Data follows Normal Distribution.
H1: Data does not follow Normal Distribution.

Tests of Normality

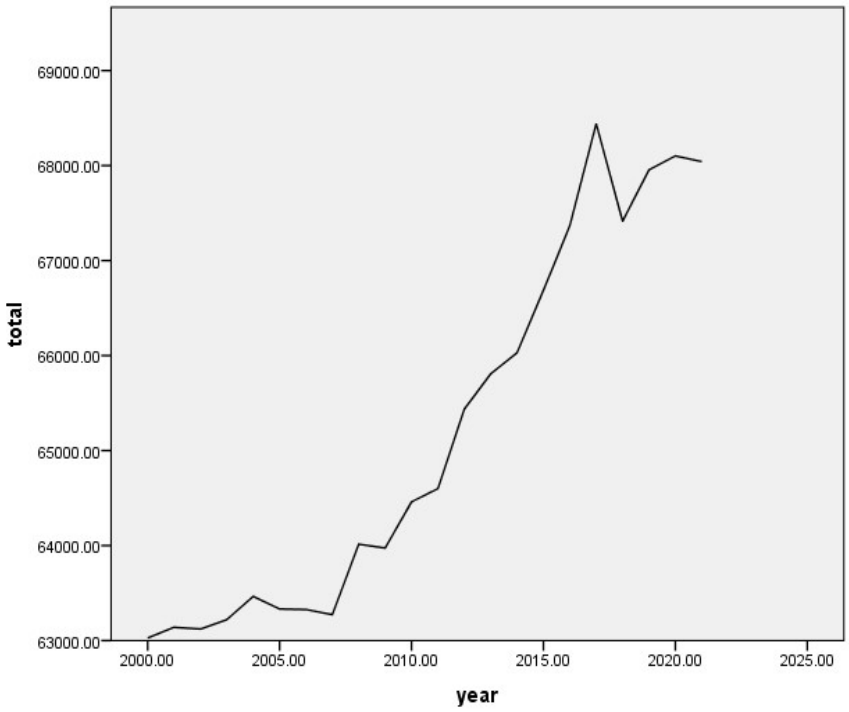
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
total	.179	22	.066	.860	22	.005

a. Lilliefors Significance Correction

- **Conclusion:**

  Here, p-Value < 0.05 so, we reject H_0 at 5% level of significance. So, we can say that the data does not follow normal distribution.

Railway Total Route Km



Model Fitting

Study Variable - Y = Ln Total Electrified

Explanatory variable – X = Year

- Model = $\boldsymbol{L_{n-y} = \beta_0 + \beta_1 x + \varepsilon}$

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.959[a]	.919	.915	.00879	.594

a. Predictors: (Constant), year

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.018	1	.018	227.975	.000[b]
	Residual	.002	20	.000		
	Total	.019	21			

a. Dependent Variable: ln_total

b. Predictors: (Constant), year

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.122	.594		3.574	.002
	year	.004	.000	.959	15.099	.000

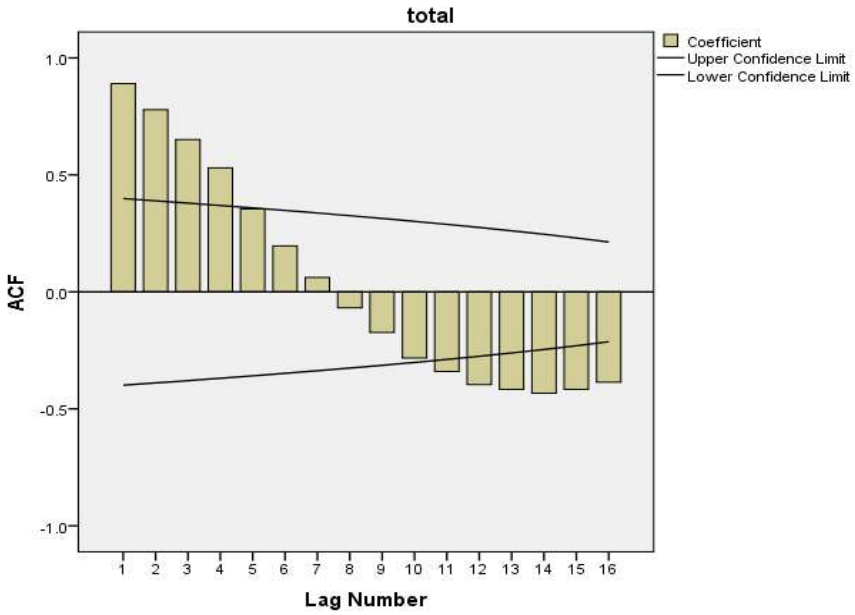
Dependent Variable: ln_total

Interpretation:

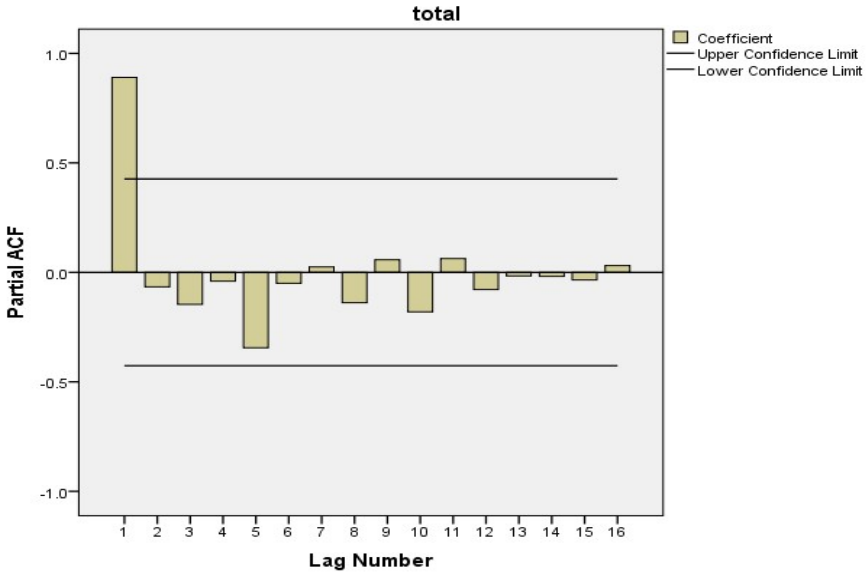
Here, 91.5% variation is explained by fitted model, S.E of model and coefficients is low and model and coefficients both are significant But Durbin-Watson Statistics - .594

Its far away from value 2. So, it indicates 1st order Autocorrelation problem in the data.

ACF and PACF plots



.



- **Interpretation:**

  From the above graph we observe that 1st lag is the out of confidence limit. So, there is problem of 1st order autocorrelation.

  For conformation, We do B.G. Test

Detection of Autocorrelation

- **B-G Test:**

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-.001	.002		-.560	.584
	lag1	.557	.257	.568	2.165	.047
	lag2	.234	.287	.251	.815	.428
	lag3	-.174	.241	-.198	-.721	.482

a. Dependent Variable: Unstandardized Residual

- **Interpretation:**

From the above coefficient table, we can say that, only 1st lag residual is significant, which is 0.47.And it is less than 0.5.
So there is only 1st order autocorrelation problem in our data.

➢ Remedial Measure:

General Regression Technique

Model = $\boldsymbol{L_{n-y} = \beta_0 + \beta_1 x + \beta_2 Lag1 + \varepsilon}$

Here as Remedial Measure we introduce Lag 1 as another Explanatory variable.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.981[a]	.961	.957	.00620	2.255

a. Predictors: (Constant), lag1, year

- **Interpretation:**

From the above table we observed that the Durbin Watson Test Statistic is near to 2. We can say that the problem of autocorrelation is removed .

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.017	2	.009	224.485	.000[b]
	Residual	.001	18	.000		
	Total	.018	20			

a. Dependent Variable: ln_total

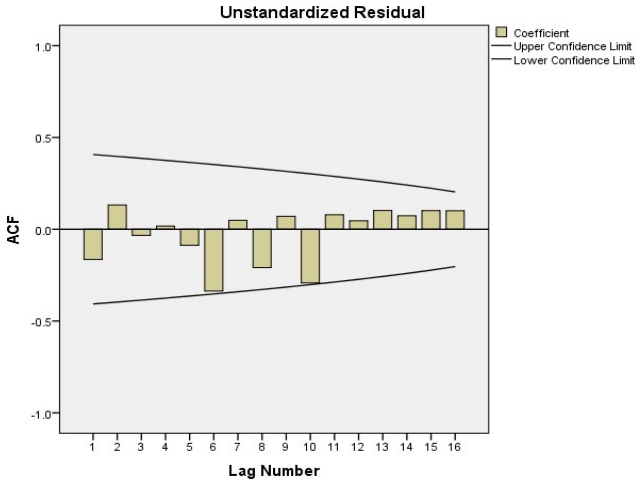
b. Predictors: (Constant), lag1, year

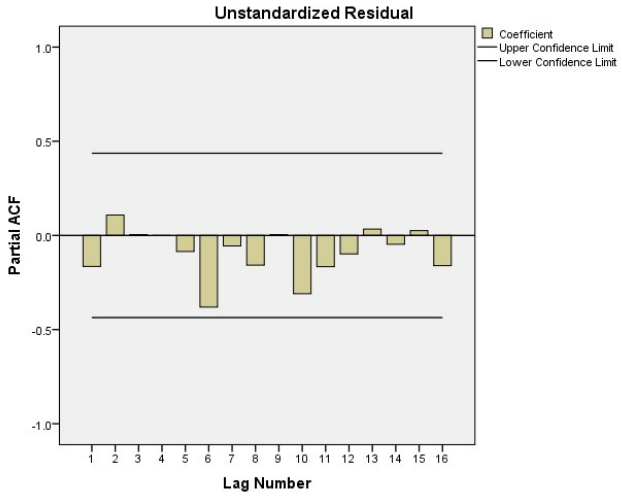
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.800	.450		4.000	.001
	year	.005	.000	.956	20.636	.000
	lag1	.643	.159	.188	4.052	.001

a. Dependent Variable: ln_total

ACF and PACF plots after removing problem of autocorrelation





- **Interpretation:**

  From the above graph, we conclude that, we removed the problem of Autocorrelation successfully.

Detection of Heteroscedasticity

Now to check is their problem of Heteroscedasticity or not, we'll run Glejser Test.

- **Glejser Test**

- **Hypothesis:**

  H_0: The residuals are Homoscedastic
  H_1: The residuals are Heteroscedastic

To check is their Heteroscedasticity exist or not in the model we take Dependent Variable = |ui| & Independent Variable = Year, Lag 1

Coefficients

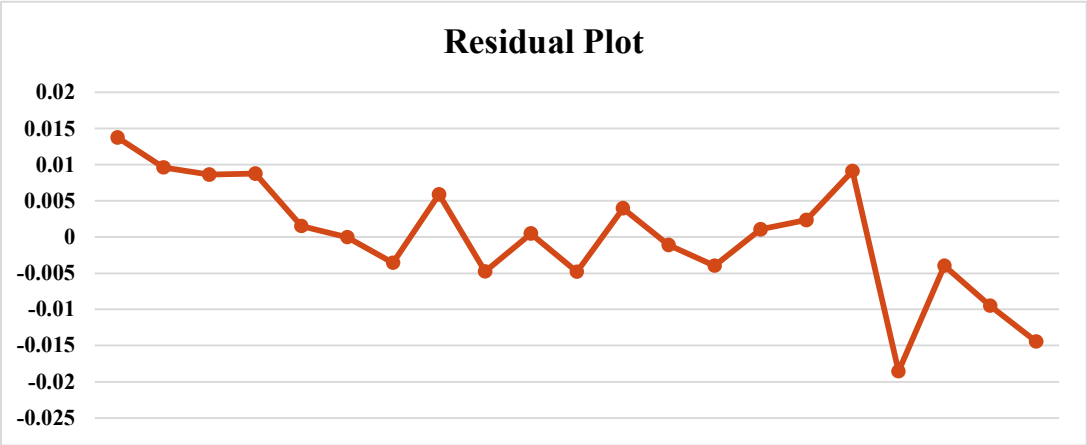
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-.379	.242		-1.567	.135
	year	.000	.000	.328	1.585	.130
	lag1	.140	.085	.340	1.644	.117

a. Dependent Variable:|ui|

- **Interpretation:**

From the above table we can say that there is no Problem of Heteroscedasticity.

Residual Plot

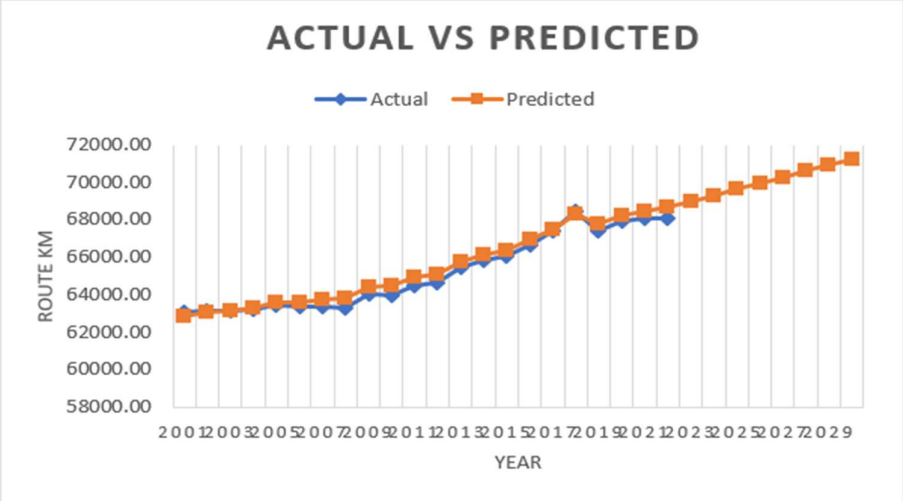


- **Interpretation:**

From the above given residual plot we observed that the residuals are random.

Forecasting of Total KM

Year	ln(Y)	Y(RKM)
2022	11.13694	**68661.23**
2023	11.14156	**68979.18**
2024	11.14617	**69297.91**
2025	11.15079	**69618.81**
2026	11.15541	**69941.19**
2027	11.16003	**70265.06**
2028	11.16464	**70589.73**
2029	11.16926	**70916.61**
2030	11.17388	**71245.01**

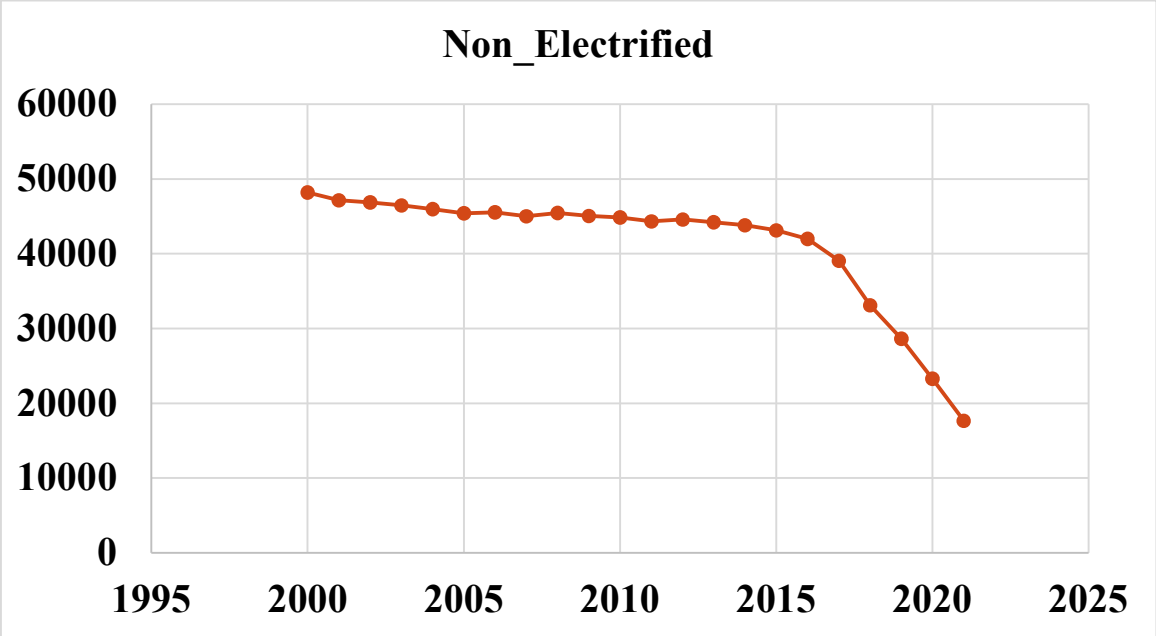


- **Interpretation:**

  From the above graph we can observed that then actual & fitted trend line are closed to each other which indicates that the fitted model is significant.

Time Series Analysis

Non-Electrified Route Km



- **Interpretation:**

  The above graph indicates downward trend of Non-Electrified RKM of Indian Railway by years.

To check the trend in the data:

- **Hypothesis:**

  H_0: There is no trend.
  H_1: There is trend.

- **Output (R-Studio):**

  Mann-Kendall Test

  tau = -0.957, 2-sided p-value = 5.5205e-10

- **Conclusion:**

Here, the P-Value < 0.05. So, we reject H_0 at 5% level of significance. hence, we may conclude that there is trend.

Estimating:

Exponential Model	R-squared	RMSE	MAPE	MAE	Significant
Holt's	0.984	0.03344	0.195028	0.020205	No
Brown's	0.984	0.032629	0.195116	0.020215	Yes

- **Interpretation:**
  From the above table we observe that , Brown's exponential smoothing model is significant.

Model

$$\hat{x}_t = a_t + b_t(t)$$

Significancy of co-efficient:

- **Hypothesis:**

  H_0: Co-efficient is insignificant.
  H_1: Co-efficient is significant.

Exponential Smoothing Model Parameters

Model			Estimate	SE	t	Sig.
Ln_electrified-Model_1	No Transformation	Alpha (Level and Trend)	1.000	.118	8.447	.000

- **Conclusion:**
  Here P-value < 0.05 , so we do reject H_0 at5% level of significant.so we can say that the Co-efficient is significant.

Diagnostic Checking:

➢ **Ljung-Box test**

- **Hypothesis:**

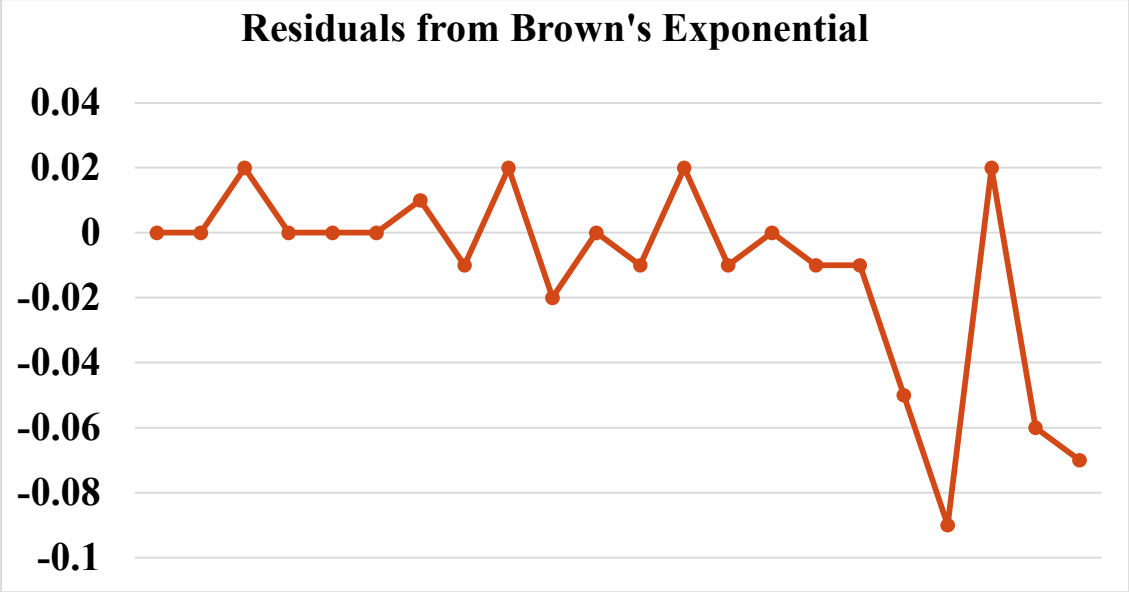
  H_0: The residuals are random.

  H_1: The residuals exhibit serial correlation.

Ljung-Box Q(18)		
Statistics	DF	Sig.
17.635	17	.412

- **Conclusion:**

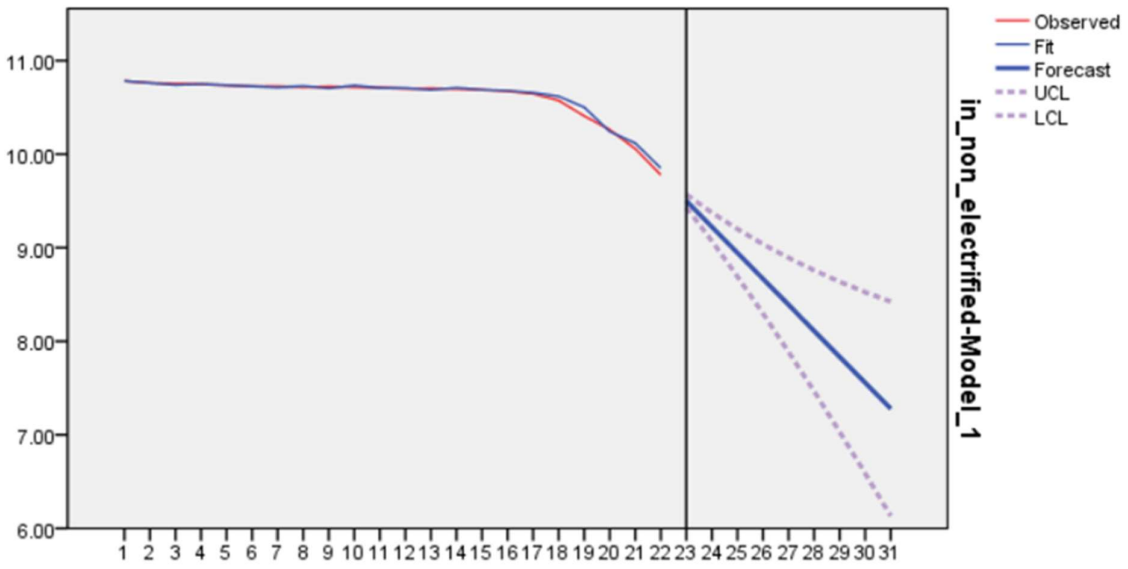
  Here the p value is >0.05 so do not reject H_0 at 5% level of significant, hence we can say that the residual are random.



- **Interpretation:**

  From the above graph, we observed that the residual are randoms.

- **Forecasting of Non-Electrified RKM:**

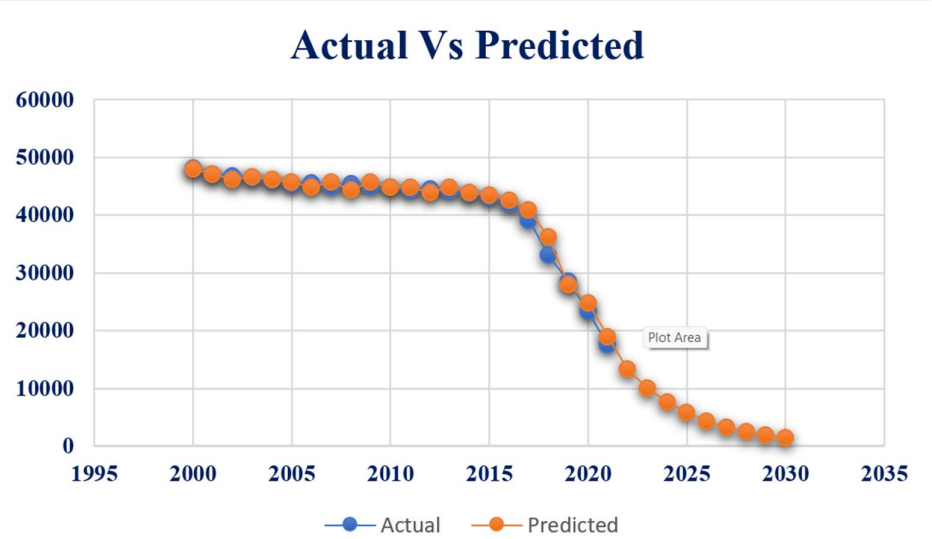


- **Interpretation:**

In the picture above, the two shaded zones of forecast represent the 95% (lower and upper side) projection of prediction intervals.

Forecasting of Non-Electrified RKM

Year	ln(Y)	Y(RKM)	LCL(ln(Y)	LCL(Y)	UCL (ln(Y)	UCL(Y)
2022	9.5	**13359.7268**	9.43	**12456.5267**	9.57	**14328.41632**
2023	9.22	**10097.0643**	9.07	**8690.62381**	9.37	**11731.11509**
2024	8.94	**7631.19706**	8.69	**5943.18224**	9.2	**9897.129059**
2025	8.67	**5825.49935**	8.3	**4023.87239**	9.04	**8433.777056**
2026	8.39	**4402.81769**	7.89	**2670.44392**	8.89	**7259.019183**
2027	8.11	**3327.57803**	7.46	**1737.14806**	8.76	**6374.111578**
2028	7.83	**2514.92937**	7.03	**1130.03061**	8.64	**5653.329824**
2029	7.56	**1919.84551**	6.59	**727.78087**	8.52	**5014.053757**
2030	7.28	**1450.98803**	6.13	**459.436161**	8.42	**4536.903455**



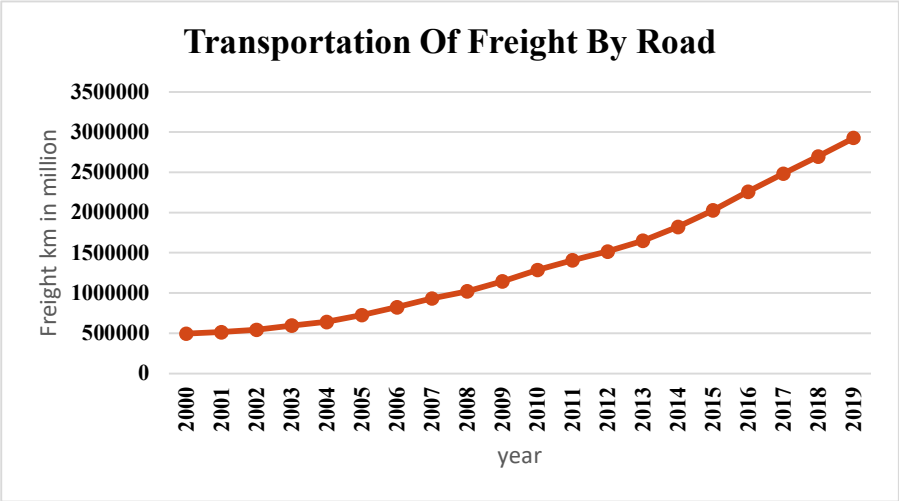
- **Interpretation:**
  From the plot it can be observed that actual and fitted trend line are close to each other which indicates that the fitted model is significant.

Percentage Of Electrification

Year	Total Route Km.	Non-Electrified Route Km.	Electrified Route Km.	Percentage of Non-Electrification	Percentage of Electrification
2022	68661.23129	13359.7268	55301.50449	19%	81%
2023	68979.18007	10097.0643	58882.11577	15%	85%
2024	69297.9082	7631.19706	61666.71114	11%	89%
2025	69618.80524	5825.49935	63793.30589	8%	92%
2026	69941.18825	4402.81769	65538.37056	6%	94%
2027	70265.06411	3327.57803	66937.48608	5%	95%
2028	70589.73385	2514.92937	68074.80448	4%	96%
2029	70916.61293	1919.84551	68996.76742	3%	97%
2030	71245.00568	1450.98803	69794.01765	**2%**	**98%**

Time Series Analysis

Transportation Of Freight By Road



To check the trend in the data:

- **Hypothesis:**

  H_0: There is no trend.
  H_1: There is trend.

- **Output (R-Studio):**

  Mann-Kendall Test

  tau = 1, 2-sided p-value =< 2.22e-16

- **Conclusion:**
  Here, The p-value < 0.05. So, we reject H0 at 5% level of signifi cance, Hence, we may conclude that there is trend.

- **Estimating:**

Exponential Model	R-squared	RMSE	MAPE	MAE	Significant
Holt's	0.998938	0.019583	0.10106	0.014038	No
Brown's	0.998939	0.019054	0.09987	0.041395	Yes

- **Interpretation:**
  From the above table we observe that, Brown's exponential smoothing model is significant.

- **Model:**

$$\hat{x}_t = a_t + b_t(t)$$

- **Significancy of co-efficient :**
- **Hypothesis:**

  H_0: Co-efficient is insignificant.

  H_1: Co-efficient is significant.

Exponential Smoothing Model Parameters

Model			Estimate	SE	t	Sig.
ln_road_freight-Model_1	No Transformation	Alpha (Level and Trend)	.984	.098	10.030	.000

- **Conclusion:**
  Here P-value < 0.05 , so we do reject H_0 at 5% level of significance .so we can say that the Co-efficient is significant.

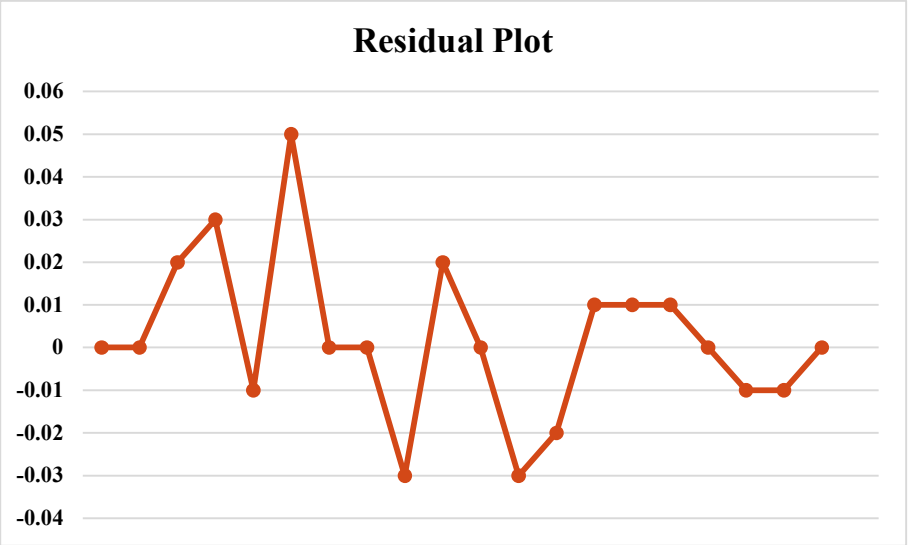
Diagnostic Checking:

- **Ljung-Box test:**

- **Hypothesis:**
  H_0: The residuals are random.
  H_1: The residuals exhibit serial correlation.

Ljung-Box Q(18)		
Statistics	DF	Sig.
6.830	17	.986

- **Conclusion:**
  Here, the p value is > 0.05 so do not reject H_0 at 5% level of significance, hence we can say that the residuals are random.

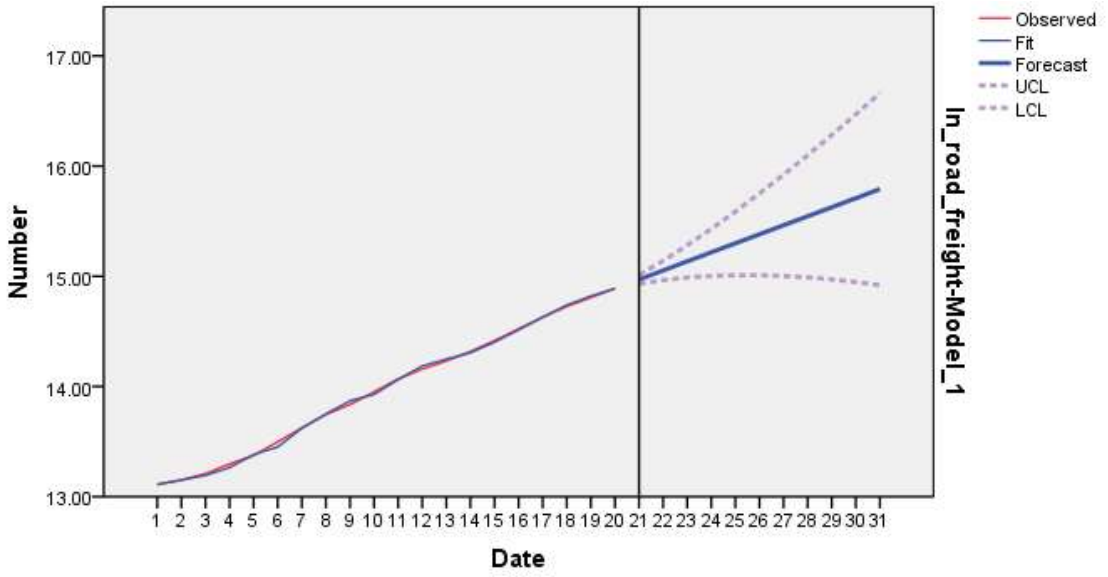


- **Interpretation:**

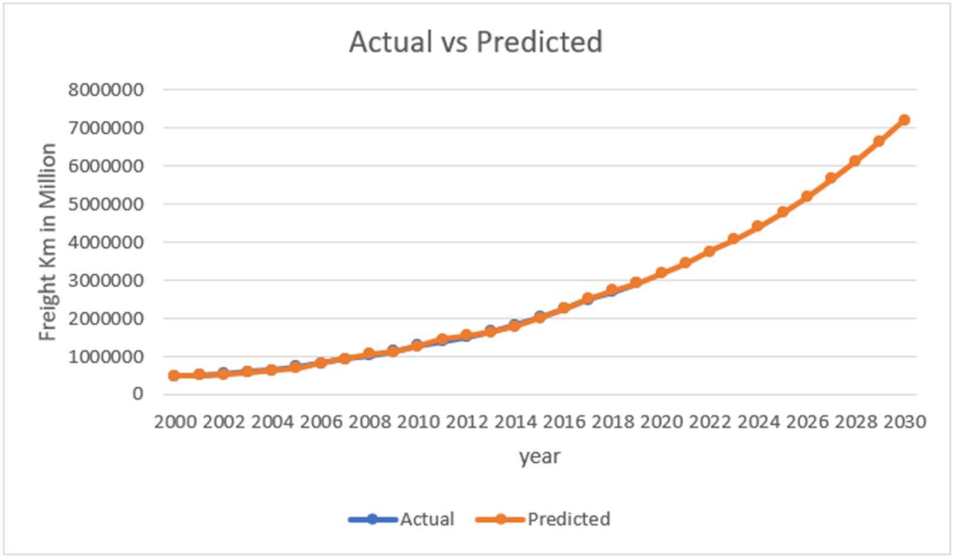
  From the above graph, we observed that the residuals are random.

Forecasting Of Freight By Road:

Year	ln(Y)	Y	LCL (ln(Y)	LCL(Y)	UCL (ln(Y)	UCL(Y)
2020	14.97	**3172403.31**	14.93	3053072.251	15.01	3306562
2021	15.05	**3436623.48**	14.97	3157990.738	15.14	3766016
2022	15.14	**3760265.03**	14.99	3233004.236	15.28	4333771
2023	15.22	**4073446.48**	15.00	3280693.689	15.43	5031364
2024	15.3	**4412711.89**	15.01	3303140.999	15.59	5887129
2025	15.38	**4780233.73**	15.01	3302238.522	15.75	6937476
2026	15.46	**5178365.38**	15.00	3279842.7	15.92	8228793
2027	15.55	**5666034.23**	14.99	3237834.419	16.10	9820051
2028	15.63	**6137941.6**	14.97	3178134.472	16.28	11786241
2029	15.71	**6649152.76**	14.95	3102696.843	16.47	14222878
2030	15.79	**7202941.2**	14.92	3013491.069	16.66	17251868

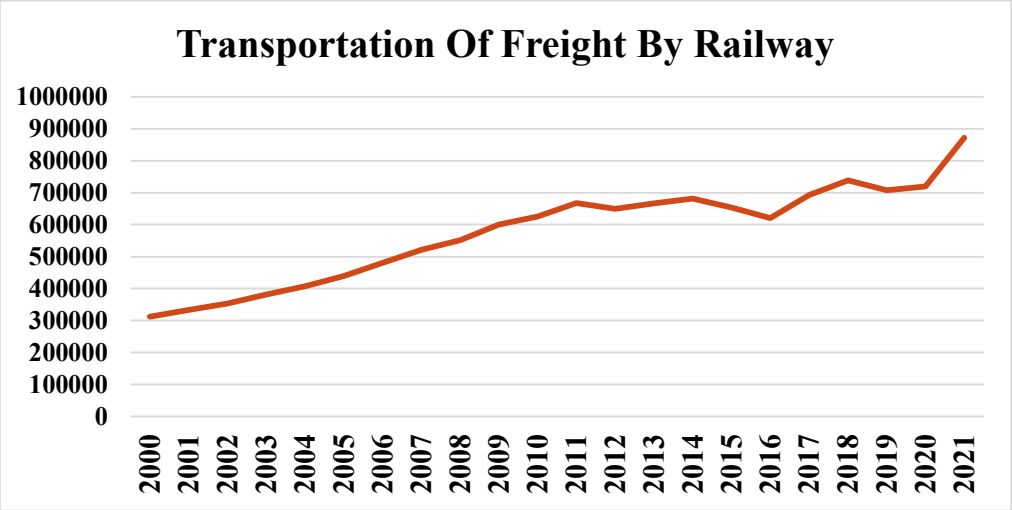


- **Interpretation:**
  In the picture above, the two shaded zones of forecast represent the 95% (lower and upper side) projection of prediction intervals.



- **Interpretation:**
  From the plot it can be observed that actual and fitted trend line are close to each other which indicates that the fitted model is significant.

Transportation Of Freight by Railway



To check the trend in the data:

- **Hypothesis**

  H_0: There is no trend.
  H_1: There is trend.

- **Output (R-Studio):**

  Mann-Kendall Test
  tau = 0.887, 2-sided p-value > 2.22e-16

- **Conclusion:**
  Here The p-value < 0.05. So, we reject H0 at 5% level of significance, He nce, we may conclude that there is trend.

Estimating:

Exponential Model	R-squared		RMSE	MAPE	MAE	Significant
Holt's	0.962071		0.057874	0.31769	0.04238	No
Brown's	0.952481		0.063218	0.310189	0.041395	Yes

- **Interpretation:**
  From the above table we observe that, Brown's exponential smoothing model is significant.

- **Model:**

$$\hat{x}_t = a_t + b_t(t)$$

- **Significancy of co-efficient:**
- **Hypothesis:**

H_0: Co-efficient is insignificant.
H_1: Co-efficient is significant.

Exponential Smoothing Model Parameters

Model			Estimate	SE	t	Sig.
ln_Freight-Model_1	No Transformation	Alpha (Level and Trend)	.665	.115	5.775	.000

- **Conclusion:**
  Here P-value < 0.05 , so we do reject H_0 at 5% level of significant.so we can say that the Co-efficient is significant.

Diagnostic Checking:

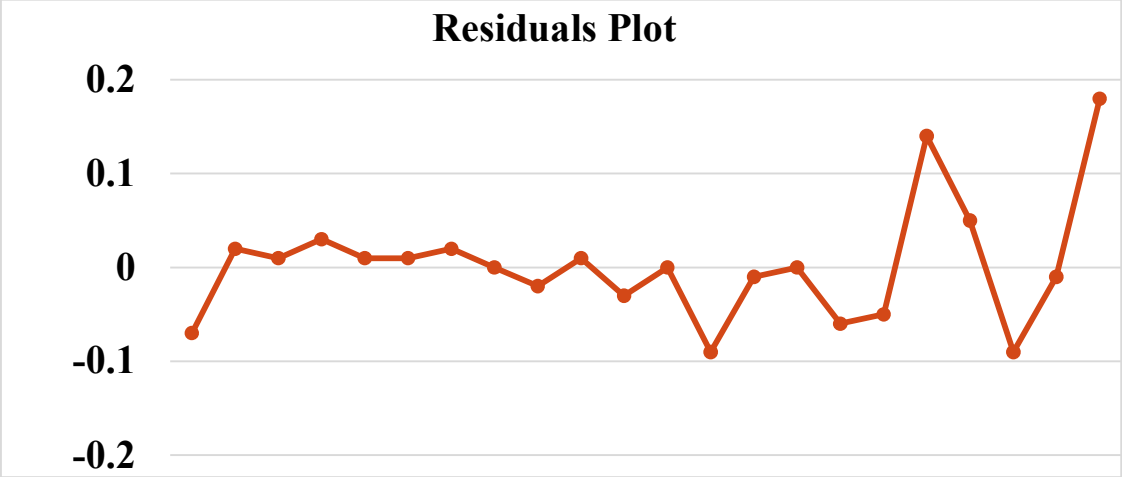
➢ Ljung-Box test

- **Hypothesis**

H_0: The residuals are random.
H_1: The residuals exhibit serial correlation.

Ljung-Box Q(18)		
Statistics	DF	Sig.
20.393	17	.255

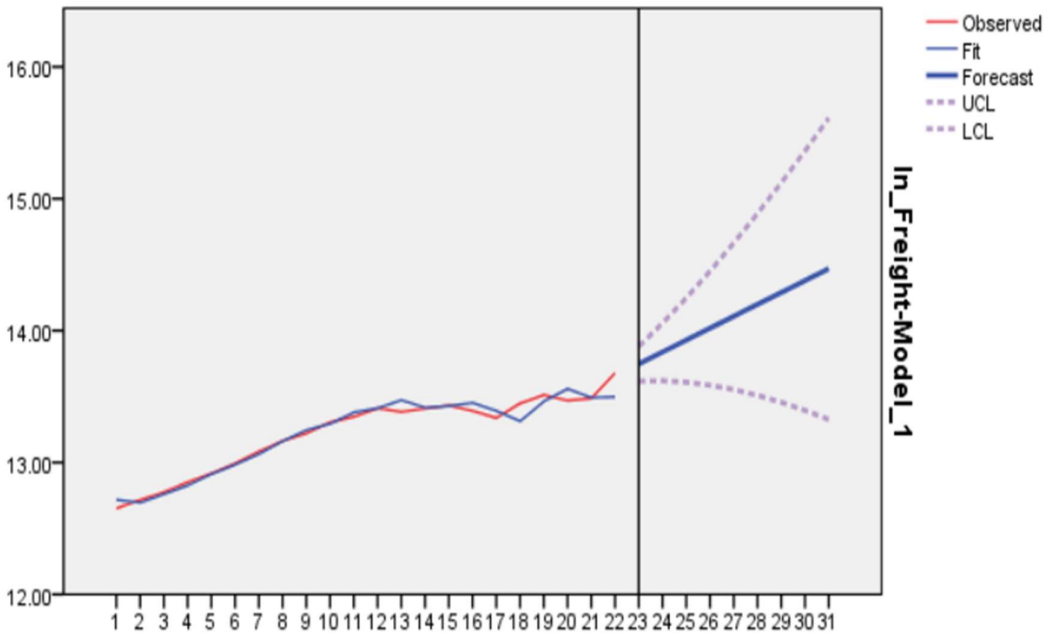
- **Interpretation:**
  Here the p value is > 0.05 so do not reject H0 at5% level of significant, hence we can say that the residuals are random.



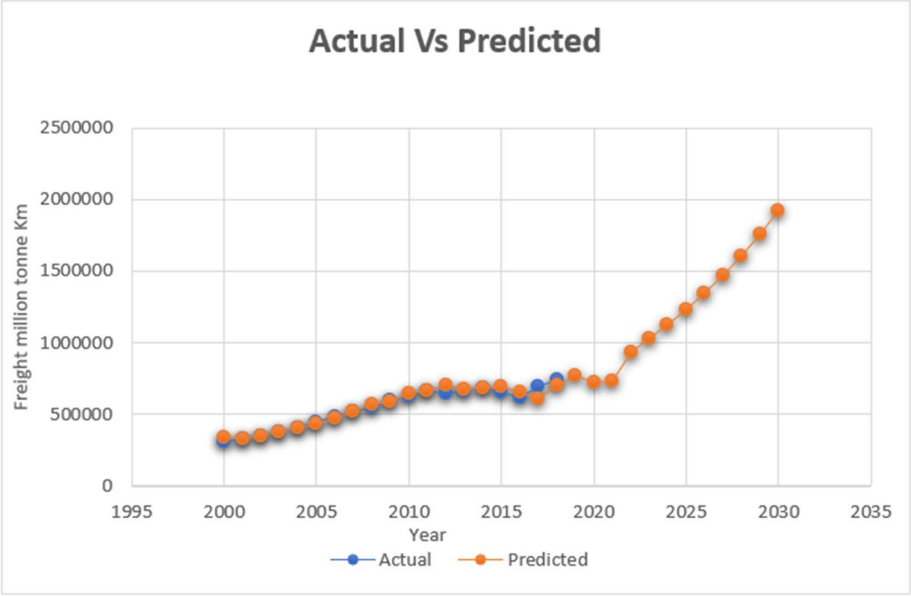
- **Interpretation:**
  From the above graph, we observed that the residual are random.

Forecasting of Freight By Railway:

Year	ln(Y)	Y(RKM)	LCL(ln(Y)	LCL(Y)	UCL (ln(Y)	UCL(Y)
2022	13.75	936589.2	13.62	822414.7	13.88	1066614
2023	13.84	1024792	13.62	822414.7	14.06	1276969
2024	13.93	1121301	13.61	814231.5	14.25	1544174
2025	14.02	1226899	13.59	798108.6	14.45	1886059
2026	14.11	1342441	13.55	766814.3	14.67	2350174
2027	14.2	1468864	13.51	736747.1	14.89	2928497
2028	14.29	1607193	13.46	700815.5	15.12	3685807
2029	14.38	1758550	13.4	660003.2	15.36	4685579
2030	14.47	1924160	13.33	615382.9	15.61	6016402

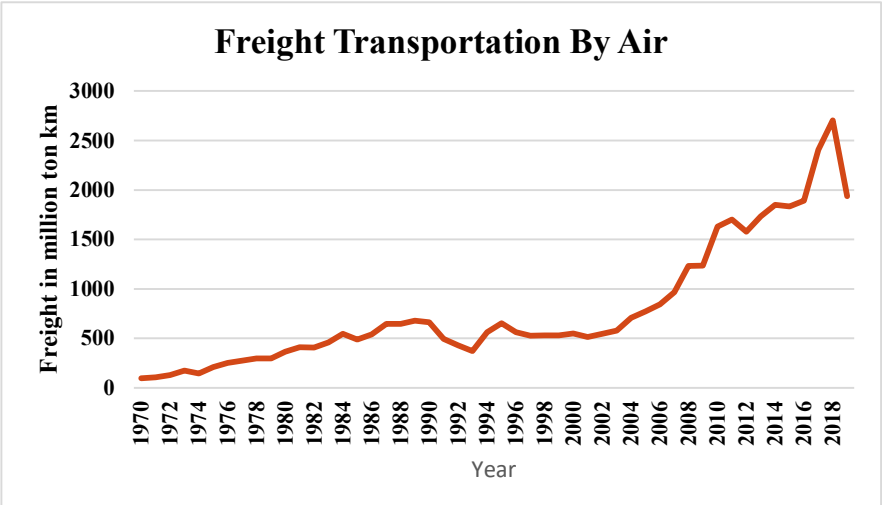


- **Interpretation:**
  In the picture above, the two shaded zones of forecast represent the 95% (lower and upper side) projection of prediction interval



- **Interpretation:**
  From the plot it can be observed that actual and fitted trend line are close to each other which indicates that the fitted model is significant.

Transportation Of Freight by Air



- **To check the trend in the data:**

- **Hypothesis:**

  H_0: There is no trend.
  H_1: There is trend.

- **Output (R-Studio):**

  Mann-Kendall Test

  ```
  tau = 0.822, 2-sided p-value =< 2.22e-16
  ```

- **Conclusion:**
  Here The p-value < 0.05. So, we reject H_0 at 5% level of significance, He nce, we may conclude that there is trend.

- **Test for stationarity check:**
- **Hypothesis**:

H_0: Data is not stationary.
H_1: Data is stationary.

- **Output (R-Studio):**

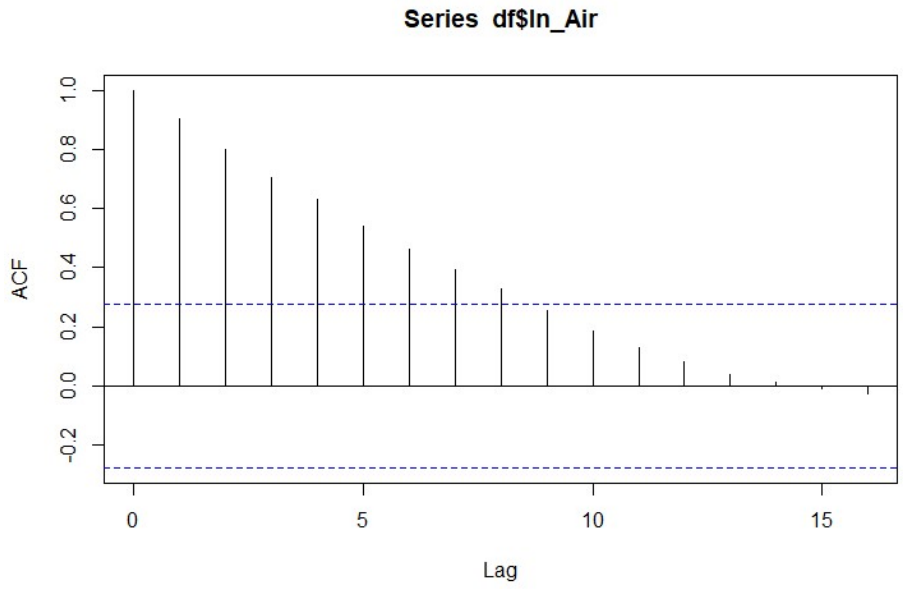
Augmented Dickey-Fuller Test

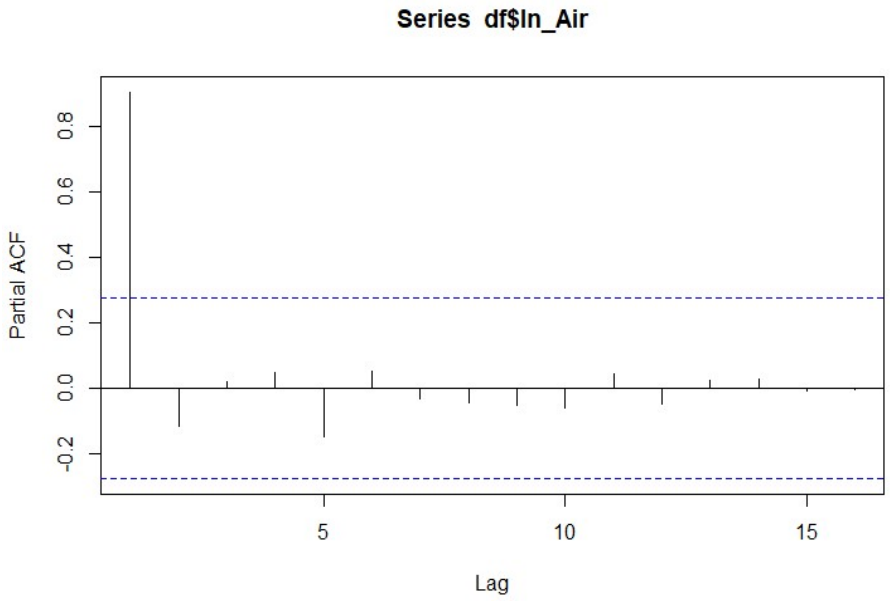
Dickey-Fuller = -1.8648, Lag order = 3, p-value = 0.6285

- **Conclusion:**

Here, p-value > 0.05. So, we accept H_0 at 5% level of significance. Hence, we may conclude that data is not stationary so we have to take first difference.

- **<u>ACF and PACF plots:</u>**





- **Interpretation:**

  From the Above graph we can see that ACF is decreasing exponentially which shows that the data is not stationary. In order to make our data stationary we first take the log transformation. After reading the graph it was found that data is still not stationary, so now we will take the second difference for the same.

- **Test For Stationary after taking 2nd Difference.**

- **Hypothesis:**

  H_0: Data is not stationary.
  H_1: Data is stationary.

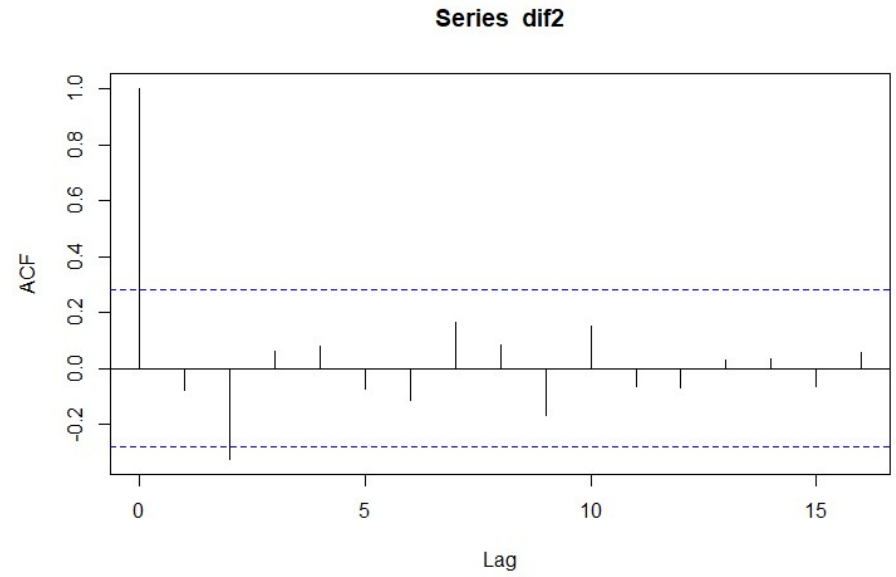
- **Output (R-Studio):**

  Augmented Dickey-Fuller Test
  data: dif2
  Dickey-Fuller = -4.8482, Lag order =3, p-value = 0.01
  alternative hypothesis: stationary

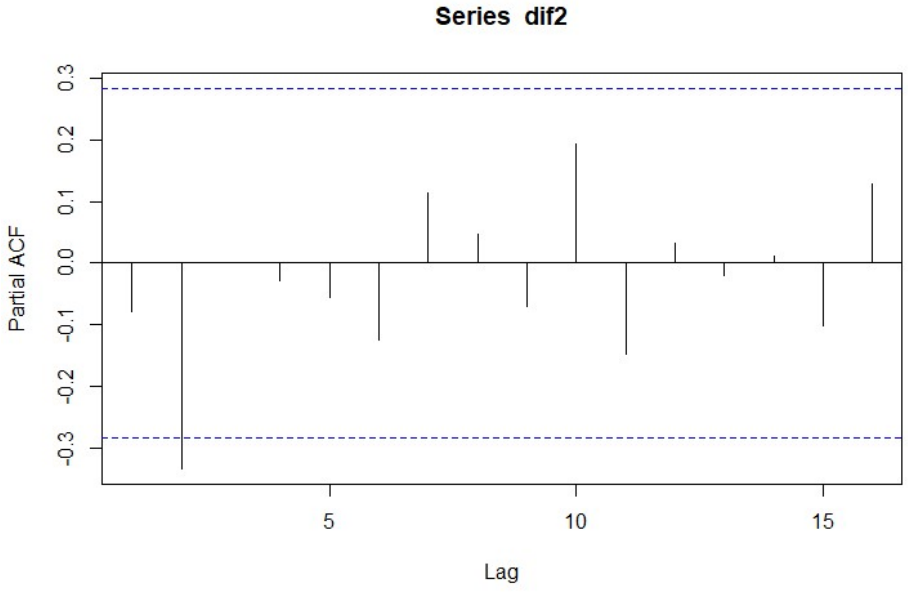
- **Conclusion:**
  Here, p-value < 0.05. So, we reject H_0 at 5% level of significance.Hence we may conclude that data is stationary after taking the second difference.

Identification:

- **Autocorrelation function 2nd difference**



- **Partial autocorrelation at 2nd difference:**



- **Interpretation:**

From The ACF Graph 2nd lag differ significantly which shows MA(2) and PACF 2nd Lag differ Significantly which shows AR(2) Process.

ARIMA Model	Sigma^2	RMSE	MAPE	AIC	Significant
(2,2,2)	0.02063	0.1407324	1.682803	-34.37	No
(1,2,2)	0.02329	0.1495249	1.707074	-33.16	No
(2,2,1)	0.0237	0.1508296	1.73814	-33.75	No
(1,2,1)	0.02348	0.1501346	1.721448	-34.88	No
(1,2,0)	0.0352	0.1838151	2.086119	-20.23	Yes
(2,2,0)	0.02962	0.1686213	1.817632	-26.12	Yes

Estimating:

So, we will fit ARIMA (2,2,0) in this model, coefficients are significant and AIC is also lower as compared to ARIMA(1,2,0) model.

- **Model:** $$\Delta y_t = \beta_1 \Delta y_{t-1} + \beta_2 \Delta y_{t-2} + \varepsilon_t$$ $$where, \Delta y_t = y_t - y_{t-1}$$

- **Significance of Co-efficient:**

- **Hypothesis:**

  H_0: Co-efficient is insignificant.
  H_1: Co-efficient is significant.

z test of coefficients:

```
    Estimate Std. Error z value  Pr(>|z|)
ar1 -0.62767    0.14254 -4.4035 1.065e-05 ***
ar2 -0.41551    0.14034 -2.9608  0.003068 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- **Conclusion:**

Here, p-value < 0.05 for both the co-efficient. So, we reject H_0 at 5% level of significance for all the co-efficient. Hence, we may conclude that coefficients are significant.

- **MODEL SUMMARY:**

arima(x = df$ln_Air, order = c(2, 2, 0))

Coefficients:

	ar1	ar2
	-0.6277	-0.4155
s.e.	0.1425	0.1403

sigma^2 estimated as 0.02962: log likelihood = 16.06, aic = -26.12

Training set error measures:

	ME	RMSE	MAE	MPE	MAPE
Training set	-0.008958454	0.1686213	0.1141749	-0.1414599	1.817632

	MASE	ACF1
Training set	0.8994482	-0.1054187

Diagnostic Checking:

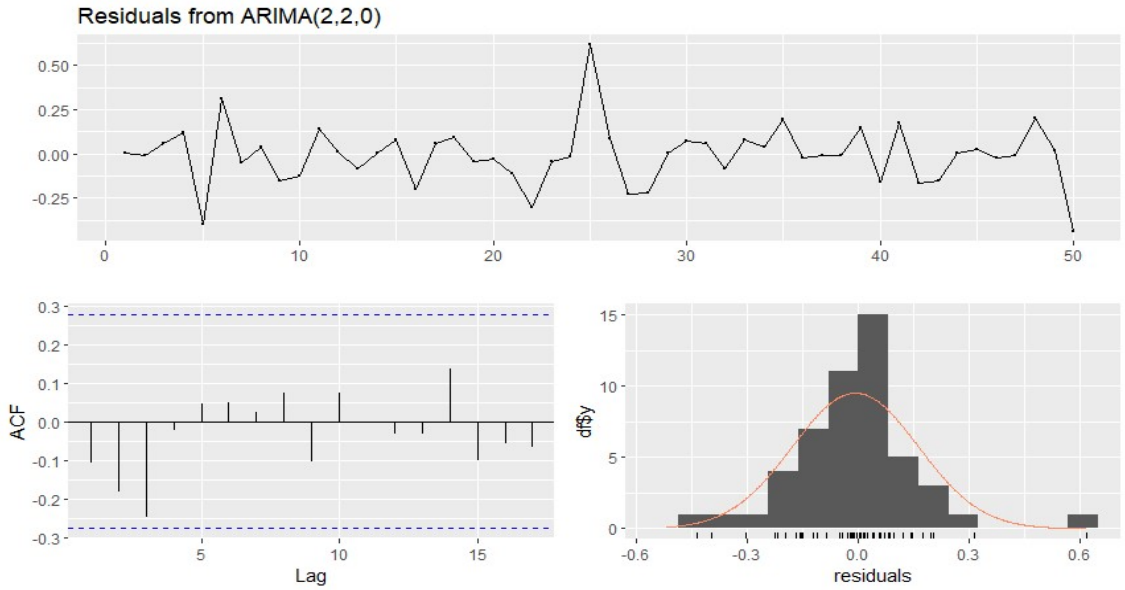
- **Output (R-Studio):**
  - **Ljung-Box test:**

data: Residuals from ARIMA (2,2,0)
Q* = 7.4735, df = 8, p-value = 0.4865

Model df: 2. Total lags used: 10

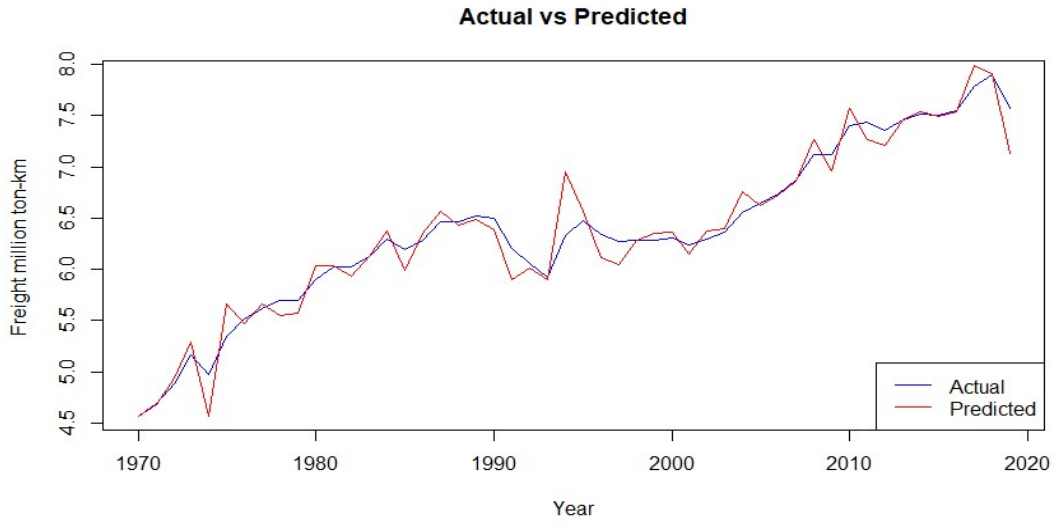
- **Conclusion:**

Here, the p-value > 0.05. So, we do not reject H_0 at 5% level of significance. Hence, we may conclude that residuals are random.



- **Interpretation:**

  Here, we observe that the variation in residuals are random, all the lags in the ACF graph are under the limits, also from the output we can observe that the standard error of the model is near to zero and RMSE, AIC and BIC are also low. So, the model is a good fit.

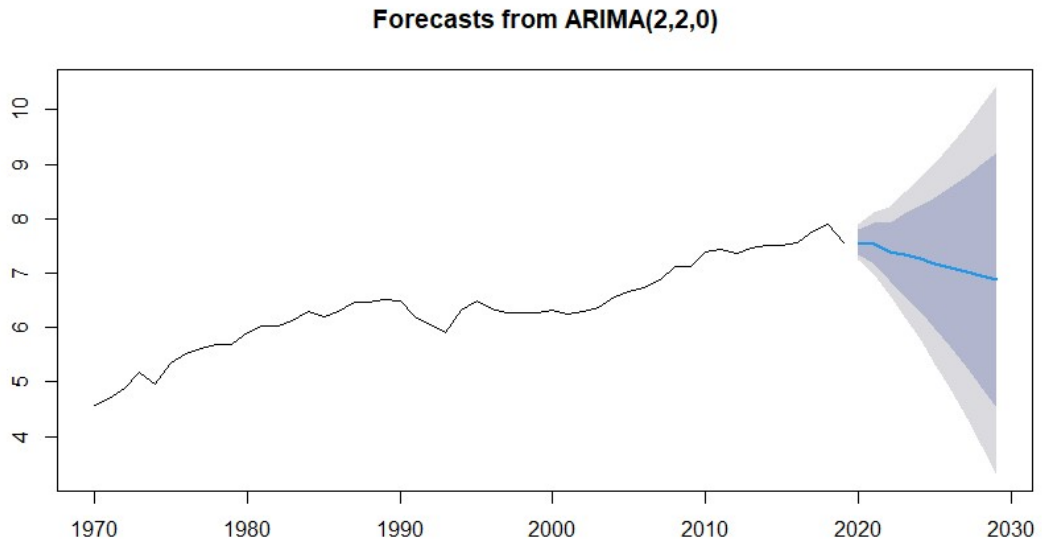


- **Interpretation:**

  From the plot it can be observed that actual and fitted trend line are close to each other which indicates that the fitted model is significant.

Forecasting Of freight By Air

Year	ln(Y)	Y	LCL (ln(Y)	LCL(Y)	UCL (ln(Y)	UCL(Y)
2020	7.570192	**1939.512632**	7.349642	1555.639509	7.79074	2418.111155
2021	7.545956	**1893.071649**	7.171457	1301.739857	7.92046	2753.023386
2022	7.399854	**1635.745594**	6.866305	959.3970364	7.9334	2788.904095
2023	7.340394	**1541.319274**	6.596823	732.7634977	8.08397	3242.065806
2024	7.277188	**1446.913578**	6.309194	549.6017912	8.24518	3809.228675
2025	7.180333	**1313.345529**	5.974469	393.259225	8.3862	4386.109963
2026	7.106155	**1219.449736**	5.639427	281.3014865	8.57288	5286.348384
2027	7.031725	**1131.981595**	5.288811	198.1077352	8.77464	6468.108529
2028	6.948031	**1041.097787**	4.914794	136.2912309	8.98127	7952.701808
2029	6.870256	**963.1951122**	4.530831	92.83567548	9.20968	9993.39846
2030	6.792615	**891.2411127**	4.133212	62.37795951	9.45202	12733.83623



- **Interpretation:**

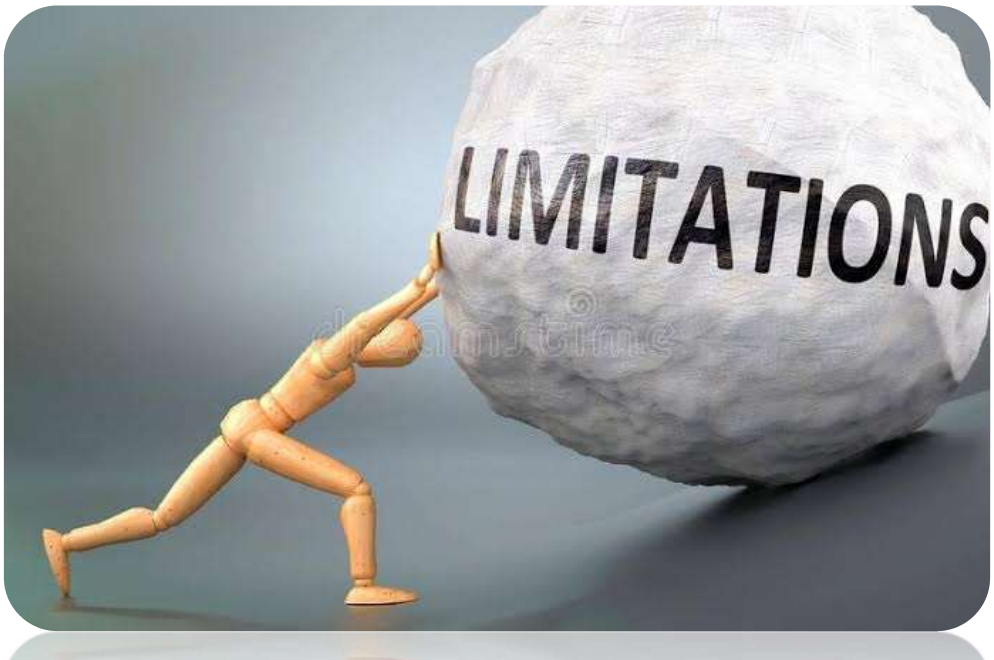
  In the picture above, the two shaded zones of forecast represent the 80% and 95% (lower and upper side) projection of prediction intervals.

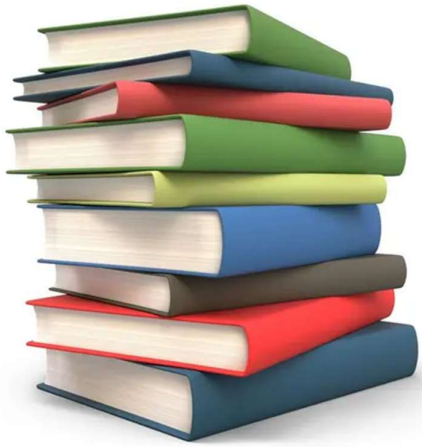
Findings

- Freight movement by road has constantly increased from 2001 to 2021, whereas freight movement by railway has remained relatively constant during the same period. Freight transport by air is significantly less compared to road and rail.
- CO2 emissions from the transport sector account for 13% of the overall CO2 emissions in India.
- The average speed of electric locomotives is higher than that of diesel locomotives.
- There is a significant difference between different zones; for example, the METRO zone is more electrified than the NER zone, NWR zone, and NR zone.
- There is a significant difference between passenger trains and freight trains.
- In the year 2000, the total route kilometres were 63,028 km, which will increase to 71,245.01 kilo meters by 2030.
- In the year 2000, non-electrified route kilometres were 48,172 km (76.42%), which will decrease to 1,450.99 km (2%) by 2030.
- In the year 2000, electrified route kilometres were 14,856 km (23.57%), which will increase to 69,794.02 km (98%) by 2030.
- In the year 2000, 312,371 million tonnes km of freight were transferred by railway, which will increase to 1,924,160 million tonnes km by 2030.
- In the year 2000, 494,000 million tonnes km of freight were transferred by road, which will increase to 7,202,941.2 million tonnes km by 2030.
- In the year 2000, 547.65 million tonnes km of freight were transferred by air, which will increase to 891.24 million tonnes km by 2030.

Limitations

- If same situation prevails then our prediction will be same as mentioned above.
- we have very limited data for use because we could not find more data because of time limitation.
- We have no data for other modes of transport such as waterways, pipelines so we can't say that the prediction is totally accurate.





Bibliography

& Webliography

BIBLOGRAPHY:

- Fundamentals of Statistics – S.C. Gupta
- Applied Nonparametric Statistics (Second Edition) – Wayne W Denial
- Basic Econometrics – Damodar N. Gujarati and Dawan C. Porter
- Applied Time Series Analysis and Forecasting – T.M.J.A Cooray

WEBLIOGRAPHY

- **https://indianrailways.gov.in/**
- **https://www.worldbank.org/en/home**
- **https://www.indiastat.com/**
- **https://www.iea.org/**

SOFTWARE:

- MS EXCEL
- SPSS
- R-STUDIO
- POWER BI