Project Report

On

A STATISTICAL STUDY ON JUVENILE DELINQUENCY IN INDIA

Submitted

in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in

APPLIED STATISTICS AND DATA ANALYTICS

by

THAHSEELA BEEGUM M P

(Reg No. SM23AS017)

(2023-2025)

Under the Supervision of

SMT. JESNA BABU

DEPARTMENT OF MATHEMATICS AND STATISTICS

ST. TERESA'S COLLEGE (AUTONOMOUS)

ERNAKULAM, KOCHI – 682011

APRIL 2025

Scanned by CamScanner

ST. TERESA'S COLLEGE (AUTONOMOUS), ERNAKULAM

CERTIFICATE

This is to certify that the dissertation entitled, **A STATISTICAL STUDY ON JUVENILE DELINQUENCY IN INDIA** is a bonafide record of the work done by Ms. **THAHSEELA BEEGUM M P** under my guidance as partial fulfilment of the award of the degree of **Master of Science in Applied Statistics and Data Analytics** at St. Teresa's College (Autonomous), Ernakulam affiliated to Mahatma Gandhi University, Kottayam. No part of this work has been submitted for any other degree elsewhere.

Date:

Place: Ernakulam

JESNA BABU
Assistant Professor,
Department of Mathematics and Statistics
St. Teresa's College (Autonomous)
Ernakulam.

Nisha Oommen
Assistant Professor & HOD
Department of Mathematics and Statistics
St. Teresa's College (Autonomous)
Ernakulam.

ST. TERESA'S COLLEGE (AUTONOMOUS) DEPARTMENT OF MATHEMATICS AND STATISTICS ERNAKULAM

External Examiners

1. Sangeetha Chandran
30.04.25

2. Anju N B
30.04.25

Scanned by CamScanner

DECLARATION

I hereby declare that the work presented in this project is based on the original work done by me under the guidance of **SMT. JESNA BABU**, Assistant Professor, Department of Mathematics and Statistics, St. Teresa's College (Autonomous), Ernakulam and has not been included in any other project submitted previously for the award of any degree.

Ernakulam

Date:

THAHSEELA BEEGUM M P

SM23AS017

Scanned by CamScanner

ACKNOWLEDGEMENTS

I must mention several individuals who encouraged me to carry out this work. Their continuous invaluable knowledgeable guidance throughout this study helped me to complete the work up to this stage.

I am very grateful to my project guide SMT. Jesna Babu for the immense help during the period of work.

In addition, the very energetic and competitive atmosphere of the Department had much to do with this work. I acknowledge with thanks to the faculty, teaching, and non-teaching staff of the department and Colleagues.

I am also very thankful to HOD Mrs. Nisha Oommen for their valuable suggestions, critical examination of work during the progress.

Ernakulam

Date:

THAHSEELA BEEGUM M P

SM23AS017

Scanned by CamScanner

ABSTRACT

Juvenile delinquency is a critical social issue that requires accurate forecasting to inform policy and intervention strategies. This study analyzes trends in juvenile delinquency rates in India from 1970 to 2022 using time series forecasting techniques. The research employs five key methodologies: ARIMA Modeling, Linear Regression, Exponential Smoothing and Holt-Winters Forecasting to examine historical trends and predict future delinquency rates. The ARIMA model was used to forecast delinquency rates for 2023 to 2027, demonstrating its ability to handle non-seasonal patterns. Linear Regression provided a trend analysis of delinquency over the years, while Exponential Smoothing and the Holt-Winters Model captured both trend and seasonal variations.

A comparative analysis of ARIMA and Holt-Winters Forecasting was conducted using RMSE values to assess predictive performance. The results indicate that the Holt-Winters Model outperforms ARIMA, with a lower RMSE value, making it the preferred model for forecasting juvenile delinquency trends. The study's findings contribute to a deeper understanding of delinquency patterns in India and provide a data-driven foundation for policymakers and law enforcement agencies to develop effective preventive measures.

Keywords: Juvenile Delinquency, Time Series Forecasting, ARIMA, Holt-Winters Model, Linear Regression, Exponential Smoothing, RMSE.

Scanned by CamScanner

ST.TERESA'S COLLEGE (AUTONOMOUS)
ERNAKULAM

Certificate of Plagiarism Check for Dissertation

Author Name	THAHSEELA BEEGUM M P
Course of Study	M.Sc. Applied Statistics & Data Analytics
Name of Guide	Ms. Jesna Babu
Department	P.G. Dept of Mathematics & Statistics
Acceptable Maximum Limit	20
Submitted By	library@teresas.ac.in
Paper Title	A STATISTICAL STUDY ON JUVENILE DELINQUENCY IN INDIA
Similarity	9% AI - 10%
Paper ID	3420921
Total Pages	46
Submission Date	2025-03-21 10:34:57

Signature of Student

Signature of Guide

Checked By
College Librarian

ST. TERESA'S COLLEGE LIBRARY ERNAKULAM

* This report has been generated by DrillBit Anti-Plagiarism Software

TABLE OF CONTENTS

Scanned by CamScanner

Scanned by CamScanner

Chapter 1

INTRODUCTION

Constantly shown that poverty, ignorance, and unstable family surroundings contribute significantly to juvenile delinquency. Children from economically weaker sections are more vulnerable to engaging in unlawful conditioning due to lack of education, employment openings, and social support systems (Acharya, S, 2017). Rajasthan has a high academy powerhouse rate, leading to increased exposure to negative influences and lower openings for legal employment, which can push kids toward crime (Atrey. I & Singh. B, 2023). The analysis of statistical data available at functionary spots indicates that there's an increase of youth involvement in heinous crimes (Dhaka, s. K., 2021). The study involves the data collection, preprocessing, exploratory data analysis, model development and evaluation criteria (Jain, H., & Patel, R., 2024). t is revealed from the study that, no particular reason is responsible for kids' delinquency- a variety of reasons are responsible for this (Haveripet, P., 2013). The Indian Legislature and Judiciary have shown perceptivity and responsiveness in securing the rights of the kids. The "Juvenile Justice Act, 2015" was legislated to insure that the law related to kids keep pace with the changing times (Khan, S., 2022). A responsible society it's a duty to keep an eye on this particular age group people, so that they could be corrected at the right time if set up shamefaced at any point during this period (Hazarika, J., & Goswami, D, 2020). The origins of felonious behaviour in youth are a complex matter; delinquency is often predictable early in some children's lives (Sharma, B. R., Dhillon, S., & Bano, S., 2009)

Juvenile delinquency is a growing concern in India, affecting not only the individuals involved but also society at large. It refers to illegal or antisocial behaviour committed by individuals under the age of 18. Understanding the trends and patterns of juvenile delinquency is pivotal for developing effective programs and preventative measures to check youth crime. The juvenile delinquency rate is told by colourful socio profitable factors similar as poverty, education, family background, peer influence, and government programs. Analysing the literal trends in juvenile crime can give precious perceptivity into the effectiveness of being laws, recuperation programs, and socio- profitable development sweats aimed at reducing youth offenses. This study focuses on examining and vaticinating juvenile delinquency trends in India from 1970 to 2022 using statistical styles.

The dataset consists of two crucial variables

1. Time (1970- 2022) – Representing the timeline of recorded delinquency data.

2. Juvenile Delinquency Rate – The number of offenses committed by kids per time.

The exploration will employ time series analysis ways, including Exponential Smoothing, Holt- Winters Method, ARIMA, and Machine literacy models to assay once trends and prognosticate unborn delinquency rates. These soothsaying models will help policymakers and law enforcement agencies anticipate implicit unborn trends and make data- driven opinions to address juvenile crime effectively. By conducting this statistical study, we aim to give a comprehensive understanding of the patterns and oscillations in juvenile delinquency in India, which can contribute to the development of better crime forestallment strategies, recuperation programs, and social programs to cover and guide the youth towards a positive future.

1.1 Objectives of the Study

1. To forecast future juvenile delinquency rates using ARIMA model.
2. To analysis trends line in juvenile delinquency rates over Linear Regression.
3. To model & forecast the future values of rice price for Ernakulam District using Exponential smoothing.
4. To compare the forecast by ARIMA and Exponential smoothing to find best model

Chapter 2

REVIEW OF LITERATURE

1. Sharma (2009) proposed that Juvenile delinquency is a global issue, intensified in developing countries by rapid socio-economic shifts. Urbanization, industrialization, and migration weaken traditional controls, increasing youth crime. Research links adverse childhood experiences to delinquency, stressing early intervention. Despite extensive studies, gaps remain in India's long-term trends, with most focusing on short-term analyses. This study fills that gap by analyzing juvenile delinquency from 1970 to 2022 using statistical methods. Unlike prior qualitative studies, it employs a quantitative time-series approach, integrating historical, socio-economic, and policy factors for objective, data-driven insights.
2. Agyemang (2012) proposed that Community policing is an effective crime prevention strategy, enhancing police-community collaboration and reducing crime. However, its impact varies based on implementation and socio-economic factors. Existing studies often rely on qualitative or short-term analyses, lacking empirical validation of long-term effects. This study fills that gap using ARIMA intervention analysis to assess Ghana's Community Policing Unit (2000–2011). Findings show a statistically significant yet temporary crime reduction of 16 cases per month, with a long-term decrease of 16.23 cases. Unlike prior qualitative research, this study employs a quantitative time-series approach, ensuring objective, data-driven insights while avoiding redundancy.
3. Haveripet (2013) proposed that Juvenile delinquency is a global issue, with 10.2 offenders per 100,000 population worldwide and 0.9–1% of total crimes in India. Key causes include family conflict, socio-economic conditions, lack of supervision, media influence, and child abuse, which lower self-esteem and increase criminal tendencies. Despite research, gaps remain in understanding long-term trends and socio-economic influences on juvenile delinquency in India. Most studies focus on specific causes rather than comprehensive statistical analysis or the impact of rehabilitation efforts. This study fills these gaps by analyzing juvenile delinquency trends in India (1970–2022) through quantitative time-series analysis, integrating historical and socio-economic factors for objective, data-driven insights.

4. Acharya (2017) proposed that Juvenile delinquency is driven by psychological factors, with poverty, illiteracy, and family disturbances as key causes. Perceived abuse leads to crime, while community rehabilitation is favoured over harsh punishment. Despite media focus, most juvenile crimes are non-violent. Gaps remain in quantifying long-term psychological and socio-economic influences. Research largely focuses on case studies rather than statistical trends, with limited empirical analysis of rehabilitation. This study fills these gaps by analyzing juvenile delinquency trends in India (1970–2022) using statistical time-series analysis, integrating psychological and socio-economic factors for objective insights into crime patterns and prevention.
5. Khairuddin et al (2019) proposed that Crime forecasting has evolved from statistical models to artificial intelligence (AI)-based models, with AI proving more effective in handling nonlinear crime patterns. While statistical models work well for linear and small datasets, AI techniques, particularly deep learning models, offer greater accuracy by adapting to complex crime distributions. However, gaps remain in long-term crime trend analysis and integrating socio-economic factors such as unemployment and GDP into forecasting models. Most studies focus on short-term predictions without considering broader influences on crime trends. This study addresses these gaps by applying time-series forecasting to juvenile delinquency trends in India (1970–2022) using AI-based models like RNN-LSTM. By incorporating multivariate analysis, it improves crime trend predictions and provides data-driven insights for policy interventions.
6. Devi and Kavitha (2021) proposed that Crime forecasting using time-series analysis and machine learning is crucial for predictive policing. Studies highlight deep learning models, especially Recurrent Neural Networks (RNN), for accurate crime trend predictions. The N-Beats RNN model enhances forecasting through data preprocessing and hyperparameter tuning. However, research mainly focuses on short-term predictions, lacking long-term trend analysis and integration of socio-economic factors. This study fills these gaps by applying RNN-LSTM models to forecast juvenile delinquency trends in India (1970–2022), ensuring high accuracy and a data-driven approach to crime analysis and prevention.
7. Dhaka (2021) proposed that Juvenile delinquency is rising globally, with India witnessing increasing youth involvement in crimes. The Indian legal system has responded by amending the Juvenile Justice Act, allowing juveniles involved in

heinous crimes to be tried as adults under Juvenile Justice Board supervision. Despite legal reforms, gaps remain in understanding long-term trends and the effectiveness of policy changes. Most studies analyze legal provisions rather than statistical patterns of juvenile crime over time. This study bridges these gaps by analyzing juvenile delinquency trends in India (1970–2022) through time-series analysis, offering empirical insights into policy impact and crime patterns.

8. Khan (2022) proposed that Juvenile delinquency in India, driven by socio-economic factors, psychological distress, and systemic challenges, continues to rise despite legal reforms like the Juvenile Justice Act, 2015. The Act, influenced by the Nirbhaya case, introduced stricter provisions, including trial of 16-year-olds as adults for heinous crimes. However, gaps remain in understanding socio-economic drivers and assessing legal interventions over time. Existing studies focus on judicial responses but lack long-term statistical analysis of crime patterns. This study fills the gap by analyzing juvenile delinquency trends (1970–2022) using time-series analysis, integrating sociological and legal perspectives to evaluate policy impact and intervention effectiveness.
9. Atrey and Singh (2023) proposed that Juvenile delinquency in Rajasthan is driven by family conflict, peer influence, poverty, child marriage, and school-related factors, with serious social consequences. Studies show that poor parental supervision, low academic performance, and mental health issues significantly contribute to juvenile crime. Existing research highlights the prevalence of delinquency and the legal framework but lacks long-term statistical analysis of trends and the effectiveness of preventive measures. This study bridges these gaps by analyzing juvenile delinquency trends (1970–2022) using time-series analysis, integrating sociological, psychological, and legal perspectives. By evaluating risk factors and policy interventions, it provides data-driven insights for more effective prevention strategies in Rajasthan.
10. Muthamizharasan and Ponnusamy (2024) proposed that Crime forecasting is crucial for law enforcement and public safety, with ARIMA and LSTM as key models. ARIMA predicts linear patterns but struggles with complex trends, while LSTM captures long-term dependencies, making it better for dynamic crime trends. Research confirms LSTM's superiority, but comparative studies on long-term juvenile delinquency trends remain scarce. Few studies explore hybrid models (CNN-LSTM) or socio-economic factors in crime forecasting. This study fills these

gaps by applying LSTM-based forecasting to juvenile delinquency in India (1970–2022), comparing it with ARIMA. By merging time-series analysis and machine learning, it improves predictive accuracy and informs policy decisions.

Chapter 3

MATERIALS AND METHODS

3.1 DATA SOURCE

The data correspond of reported incidents of cases about juvenile delinquency in India 1970 – 2022

The dataset contains 2 variables. It consists of reported incidents of juvenile delinquency cases in India from 1970 to 2022. Below are the crucial features of the data

1. Year: This variable represents the specific time for which the juvenile delinquency data has been recorded. The dataset follows a chronological order from 1970 to 2022.

2. Delinquency Rate: This variable indicates the number of reported juvenile delinquency cases per a specified population for each time. It helps in assaying trends over time and understanding the oscillations in juvenile crime rates.

This dataset allows for statistical analysis to examine long- term trends in juvenile delinquency, identify implicit factors impacting changes, and make data- driven policy recommendations.

3.2 PYTHON PROGRAMMING LANGUAGE

In this project, Python was extensively used for data analysis, statistical modeling, and time series forecasting of juvenile delinquency rates in India from 1970 to 2022. Python's powerful libraries, including Pandas and NumPy, were utilized for data preprocessing and manipulation, while Matplotlib and Seaborn were employed for visualizing trends and patterns. Time series forecasting techniques, such as ARIMA, Linear Regression, Exponential Smoothing, and the Holt-Winters Model, were implemented using the Statistical models and Scikit-learn libraries. RMSE values were calculated to evaluate model performance, ensuring accurate comparisons. Python's efficiency in handling large datasets and automation capabilities streamlined the analysis, making it an essential tool for deriving meaningful insights and predictions in this study.

3.3 MICROSOFT EXCEL

In this project, Microsoft Excel was extensively used for organizing, analyzing, and visualizing juvenile delinquency data in India from 1970 to 2022. Excel's functions and formulas were applied for data preprocessing, trend analysis, and statistical calculations. PivotTables and charts were utilized to summarize and present delinquency trends effectively. Time series forecasting techniques, including ARIMA, Linear Regression, Exponential Smoothing, and the Holt-Winters Model, were implemented using Excel's built-in tools and add-ins. RMSE values were calculated to compare model performance. Additionally, Excel's automation features, such as Macros, streamlined repetitive tasks, enhancing efficiency in data handling and analysis.

3.4 TIME SERIES

A time series is a sequence of data points recorded at consistent time intervals such as daily, monthly, or yearly, used to track changes over time. It is widely applied in fields like finance, economics, meteorology, and social sciences. Examples include daily stock market prices, annual rainfall data, and monthly sales revenue of a company. Understanding time series is crucial for analyzing trends, identifying patterns, and making future predictions.

1. Time Series Analysis

Time series analysis involves examining historical data to identify trends, patterns, and relationships over time. This analysis is used to forecast future values and make informed decisions. It includes statistical techniques to study underlying behaviors in data and derive insights. For example, a company analyzing past sales data to predict future demand for its products or a government analyzing unemployment rates to assess economic policies.

2. Secular Trend (Tt)

A secular trend represents the long-term upward or downward movement in a time series over an extended period. It reflects the overall direction of data while ignoring short-term fluctuations.

Upward Trend: Population growth, increasing e-commerce sales, rising global temperatures.

Downward Trend: Declining birth rates, decreasing landline telephone usage, reduced illiteracy rates.

For instance, the long-term increase in internet usage worldwide represents an upward trend, while the decline in the use of printed newspapers shows a downward trend.

3. Seasonal Fluctuations (St)

Seasonal fluctuations are repetitive patterns that occur at fixed intervals, such as annually, quarterly, or monthly, due to seasonal factors like weather, holidays, or cultural events. These patterns repeat over time and influence business and economic activities.

4. Cyclic Fluctuations (Ct)

Cyclic fluctuations are variations that occur over a period longer than a year and follow economic or business cycles, which consist of phases like expansion, recession, depression, and recovery. Unlike seasonal trends, cyclic changes do not have a fixed periodicity.

5. Irregular Fluctuations (It)

Irregular fluctuations are unpredictable and sudden changes in a time series caused by unexpected events such as natural disasters, political instability, pandemics, or financial crises. These fluctuations do not follow any specific pattern and are usually temporary.

Understanding these components of a time series helps in making more accurate forecasts, planning business strategies, and improving decision-making in various industries.

3.4.1 Simple Moving Average

The Simple Moving Average (SMA) is a widely used time series analysis technique that smooths data by calculating the average of a fixed number of consecutive values over a specific period. This method helps to reduce short-term fluctuations, making it easier to identify underlying trends and patterns in the data. By continuously updating as new data points are added, SMA provides a clearer view of long-term movements, filtering out random noise and volatility. It is extensively applied in financial markets for stock price analysis, where investors use it to detect trends and potential buy/sell signals. Similarly, businesses use SMA for sales forecasting, inventory management, and demand prediction, ensuring better decision-making and resource planning. The choice of the period length

(e.g., 10-day, 50-day, or 200-day SMA) depends on the level of trend analysis required—short-term or long-term. Despite its simplicity, SMA is a powerful tool for analyzing historical data trends, improving forecasting accuracy, and supporting strategic planning in various fields.

3.4.2 Auto-Regressive (AR) Process

An Autoregressive (AR) process is a fundamental concept in time series analysis, representing a type of stochastic process where current values in a series are linearly dependent on their past values plus a random error term. It is commonly used for modeling time-dependent data, where the assumption is that past observations influence future values. The general form of an autoregressive process of order p, denoted as AR(p), expresses a time series value as a linear combination of its p previous values, along with a white noise error term to account for random variations. Mathematically, it is represented as:

$$X_t = \emptyset_1 X_{t-1} + \emptyset_2 X_{t-2} + \ldots + \emptyset_p X_{t-p} + \varepsilon_t$$

Where:

X_t is the value at time t.

$\emptyset_1, \emptyset_2, \ldots, \emptyset_p$ are the finite set of weight parameters.

ε_t is the white noise term or errors at time t.

The AR process is widely used in various fields such as finance, economics, climatology, and engineering for modeling and forecasting time-dependent data. For example, in financial markets, AR models help predict stock prices, exchange rates, and interest rates by analyzing past trends. In meteorology, they are used to model temperature fluctuations and weather patterns. The choice of order p determines how many past values influence the current observation, and this order is typically selected using criteria like the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC). The AR process is a key component of more advanced models like Autoregressive Integrated Moving Average (ARIMA) and is essential for understanding time series dynamics and making reliable future predictions.

3.4.3 Moving Average (MA) Process

A Moving Average (MA) process is a fundamental stochastic process used in time series analysis to model a sequence of random variables, where each value in the series is expressed as a linear combination of current and past white noise (error) terms. Unlike an Autoregressive (AR) process, which relies on past values of the time series itself, an MA process depends solely on past forecast errors, making it useful for capturing short-term dependencies and smoothing fluctuations in time series data.

The general form of a Moving Average process of order q denoted as MA(q), can be expressed as:

$$X_t = \Theta_1 \varepsilon_{t-1} + \Theta_2 \varepsilon_{t-2} + \ldots\ldots + \Theta_q \varepsilon_{t-q} + \varepsilon_t$$

where:

X_t is the value at time t.

$\Theta_1, \Theta_2, \ldots\ldots, \Theta_q$ are the moving average coefficients.

q is the order of the MA process.

ε_t is the white noise term.

The MA process is particularly useful for modeling time series data that exhibit short-term dependencies but no long-term trend, such as financial returns, economic indicators, and temperature anomalies. It helps smooth out random fluctuations by averaging past errors, making it valuable for forecasting, signal processing, and economic modeling.

A key advantage of the MA model is its ability to capture shocks or sudden changes in data since each observation is directly influenced by recent random disturbances. It is widely used in combination with AR models in more advanced frameworks like the Autoregressive Moving Average (ARMA) model and the Autoregressive Integrated Moving Average (ARIMA) model, which are essential for predictive analytics and trend analysis. The order q of the MA process is usually determined using statistical techniques such as the Autocorrelation Function (ACF) and model selection criteria like the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC).

3.4.4 Auto-Regressive Integrated Moving Average (ARIMA) Model

The Autoregressive Integrated Moving Average (ARIMA) model is one of the most widely used methods for time series forecasting, combining autoregressive (AR), differencing (I), and moving average (MA) components to model and predict complex time-dependent data. It is particularly useful for data that exhibit trends or seasonality, making it a fundamental tool in fields such as economics, finance, weather forecasting, and demand forecasting.

Structure of ARIMA (p, d, q):

ARIMA is represented as ARIMA (p, d, q), where:

p (Autoregressive order) – Specifies the number of past values used to predict future values. The AR component assumes that past observations influence the current value through a linear relationship.

d (Degree of differencing) – Represents the number of times the data needs to be differenced to make the time series stationary. Differencing helps remove trends and stabilize the mean, ensuring that the statistical properties of the series do not change over time.

q (Moving average order) – Defines how many past forecast errors are used to correct future predictions. The MA component captures the impact of random shocks or noise in the data.

Key Features and Assumptions of ARIMA:

1. Stationarity – ARIMA assumes that the input data is stationary, meaning its mean, variance, and autocorrelation remain constant over time. If a dataset is non-stationary, it must be transformed through differencing or other techniques.
2. Univariate Data – ARIMA is designed for single-variable (univariate) time series forecasting, meaning it works best when analyzing a single dependent variable over time without external influences.
3. Past Dependency – The model assumes that future values depend on past values (AR component) and past forecast errors (MA component).

4. No Seasonality – ARIMA does not inherently account for seasonal patterns; however, an extended version called SARIMA (Seasonal ARIMA) is used for datasets with seasonal variations.

Model Selection and Evaluation:

The optimal values of p, d, and q are determined using statistical tools such as:

Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) to identify p and q.

Augmented Dickey-Fuller (ADF) Test to check for stationarity.

Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) to compare different ARIMA models and select the best fit.

Due to its ability to model linear relationships in time series data, ARIMA remains one of the most powerful forecasting techniques. However, for complex datasets with multiple influencing variables, advanced models like SARIMA, VAR (Vector Autoregression), and machine learning-based time series models may be more effective.

3.4.5 Auto Correlation Function

The Autocorrelation Function (ACF) is a fundamental statistical tool used in time series analysis to measure the correlation between a time series and its lagged values at different time points. It helps in understanding how past observations influence current values, making it a crucial component in forecasting models such as Autoregressive (AR), Moving Average (MA), ARMA, and ARIMA models.

Definition and Importance of ACF:

Autocorrelation refers to the degree of similarity between a given time series and a lagged version of itself over successive time intervals. The ACF function calculates the correlation coefficient between a time series and its past values at different lags. If the autocorrelation is high at a particular lag, it indicates that past values strongly influence future values, which is useful for trend analysis, pattern recognition, and forecasting.

Mathematically, the autocorrelation at lag k is given by:

$$\text{ACF(k)} = \frac{\sum_{t=1}^{N-k}(X_t-\bar{X})-(X_{t+k}-\bar{X})}{\sum_{t=1}^{N}(X_t-\bar{X})^2}$$

where:

X_t represents the time series values at time t.

$\overline{X}$ is the mean of the series.

k is the time lag.

N is the total number of observations.

Uses of ACF in Time Series Analysis:

1. Determining the Order of Moving Average (MA) Models:

   ACF is particularly useful in identifying the appropriate order (q) of an MA(q) model in ARIMA modeling. A significant autocorrelation at a certain lag indicates that past forecast errors at that lag influence current values.

2. Detecting Seasonality and Cyclical Patterns:

   If the ACF shows periodic spikes at regular intervals, it suggests a seasonal or cyclical pattern in the time series. This is essential in selecting SARIMA (Seasonal ARIMA) models for forecasting seasonal data.

3. Checking for Stationarity:

   A slowly declining or high autocorrelation at higher lags suggests that the time series may be non-stationary. In such cases, differencing the data can help remove trends and stabilize variance.

4. Evaluating Forecast Accuracy:

   ACF is used to analyze the residuals of a forecasting model to check whether they exhibit any pattern. If residuals still show significant autocorrelation, it indicates that the model may need improvement.

Interpreting the ACF Plot:

An ACF plot is a graphical representation of the correlation coefficients at various lags. The interpretation is as follows:

If the ACF decreases gradually with increasing lags, it suggests a strong trend component in the data.

If the ACF has significant spikes at regular intervals, it indicates seasonality.

If the ACF drops sharply to near zero after a few lags, it suggests that an MA(q) model may be appropriate for the data.

3.4.6 Partial Auto Correlation Function

The Partial Autocorrelation Function (PACF) is a key statistical tool in time series analysis that measures the direct correlation between a time series and its lagged values while removing the effects of intermediate lags. Unlike the Autocorrelation Function (ACF), which considers the cumulative influence of all previous observations, PACF isolates the pure relationship between a time series and a specific lag, making it an essential tool for model selection in forecasting.

Definition and Importance of PACF

PACF helps in identifying the true relationship between past observations and the present value of a time series by eliminating indirect influences. This is particularly useful in selecting the appropriate order (p) of an Autoregressive (AR) model in ARIMA modeling.

Mathematically, the PACF at lag k represents the correlation between X_t and X_{t-k} after removing the contributions of all intermediate lags (X_{t-1},X_{t-2},…..,X_{t-k+1}). It is computed using regression techniques where each lagged value is regressed on all previous lags, and the residual correlation is taken as the partial autocorrelation.

Uses of PACF in Time Series Analysis

1. Determining the Order of Autoregressive (AR) Models

   PACF is primarily used to identify the order (p) of an AR(p) model in ARIMA modeling. If the PACF shows significant correlation at a particular lag p but drops to near zero afterward, it indicates that the data follows an AR(p) process.

2. Distinguishing Between AR and MA Processes

   If the ACF gradually declines while PACF cuts off sharply after a few lags, it suggests an AR process.

If the PACF declines gradually while ACF cuts off after a few lags, it suggests a Moving Average (MA) process.

3. Identifying Stationarity in a Time Series

   A slowly decaying PACF at increasing lags may indicate a non-stationary time series. Differencing the data can help transform it into a stationary series before applying ARIMA modeling.

4. Selecting the Correct Forecasting Model

   PACF is crucial in determining whether a time series follows an ARIMA (p, d, q) or MA(q) model, ensuring accurate forecasting.

Interpreting the PACF Plot

A PACF plot is a graphical representation where:

Significant spikes at a particular lag p followed by a sharp decline suggest an AR(p) process.

A gradual decline in PACF but a sharp cutoff in ACF suggests an MA(q) process.

Significant spikes at multiple lags suggest seasonality or a mix of AR and MA processes.

3.4.7 Augmented Dickey-Fuller test

The Augmented Dickey-Fuller (ADF) test is a widely used statistical test in time series analysis to determine whether a given time series is stationary or non-stationary. Stationarity is a key assumption in many forecasting models, including ARIMA, as non-stationary data can produce unreliable predictions and misleading results. The ADF test is an extension of the Dickey-Fuller test, incorporating lagged differences of the time series to account for autocorrelation, making it more robust for real-world data.

Concept of Stationarity and Unit Root

A time series is stationary if its statistical properties, such as mean, variance, and autocorrelation, remain constant over time. Conversely, a non-stationary series exhibits trends, seasonality, or changing variance. The ADF test checks for the presence of a unit root, a

characteristic of non-stationary time series, to determine whether differencing is required before applying forecasting models.

Hypotheses of the ADF Test

Null Hypothesis (H_0): The series is non-stationary (has a unit root)

If the null hypothesis is not rejected, the data exhibits trends, and differencing is required to make it stationary.

Alternative Hypothesis (H_1): The series is stationary (no unit root)

If the null hypothesis is rejected, the data is already stationary, meaning its statistical properties remain stable over time.

ADF Test Formula

The ADF test is based on the regression equation:

$$\Delta Y_t = \alpha + \beta t + \gamma Y_{t-1} + \sum \delta_i \Delta Y_{t-i} + \epsilon_t$$

Where:

Y_t is the time series at time t

$\Delta Y_t = Y_t - Y_{t-1}$ represents the first difference of the series

α is a constant (intercept)

βt represents a trend component (optional)

γY_{t-1} tests whether the series has a unit root

δ_i are the coefficients of the lagged differences

ϵ_t is white noise (random error)

Interpreting ADF Test Results

The test produces a test statistic, which is compared against critical values at significance levels (1%, 5%, and 10%):

If the test statistic is less than the critical value, the null hypothesis is rejected, indicating stationarity.

If the test statistic is greater than the critical value, the null hypothesis is not rejected, meaning the series is non-stationary and needs differencing.

3.4.8 Seasonal data

A time series is considered seasonal when it exhibits repeating patterns at regular intervals due to factors like weather, holidays, or business cycles. These variations occur daily, monthly, or annually and help in understanding trends for accurate forecasting.

For example, retail sales peak during festive seasons, tourism rises in summer, and electricity consumption fluctuates with seasons. Identifying seasonality aids in business planning, inventory management, and financial forecasting.

Statistical methods like STL decomposition and forecasting models such as SARIMA and Holt-Winters help analyze and predict seasonal effects, ensuring better decision-making in time series forecasting.

3.4.9 Akaike Information Criterion (AIC)

The Akaike Information Criterion (AIC) is a widely used statistical measure in model selection, particularly in time series analysis and machine learning, to determine the best-fitting model among a set of competing models. It evaluates models based on their goodness of fit while also penalizing excessive complexity to prevent overfitting. The AIC is designed to strike a balance between model accuracy and simplicity, ensuring that the chosen model is both effective and generalizable to new data.

AIC can be calculated using the following formula:

$$AIC = -2 * \ln(L) + 2 * k$$

Where:

L is the maximized value of the likelihood function of the model.

k is the number of parameters in the model.

A lower AIC value indicates a better model, meaning it achieves a good fit with fewer parameters. When comparing multiple models, the one with the smallest AIC is preferred. However, AIC values are relative, meaning they are only useful when comparing models fitted to the same dataset.

AIC is widely applied in autoregressive models (AR), moving average models (MA), ARIMA models, and regression analysis to identify the most efficient model for forecasting. It helps researchers and analysts make data-driven decisions by selecting models that provide accurate predictions without unnecessary complexity. However, AIC does not directly assess model accuracy but rather provides a trade-off between fit and complexity, making it a crucial tool in model evaluation and selection.

3.4.10 Model Fitting

Model fitting is the process of estimating a model's parameters to ensure that it accurately represents the underlying patterns and relationships in a given dataset. The primary goal of model fitting is to identify a model that provides the best balance between accuracy and generalization, meaning it should perform well not only on the training data but also on unseen data. A well-fitted model captures the essential trends and patterns in the data without overfitting (capturing noise) or underfitting (failing to capture important details).

Model fitting involves several steps, including selecting an appropriate model type, estimating its parameters, and evaluating its performance using statistical measures such as mean squared error (MSE), root mean squared error (RMSE), R-squared (R^2), and information criteria like AIC or BIC. Techniques like maximum likelihood estimation (MLE), least squares regression, and gradient descent are commonly used to optimize model parameters.

In time series analysis, model fitting plays a crucial role in forecasting, where models like ARIMA, Exponential Smoothing, and Holt-Winters are trained to capture temporal patterns. In machine learning, fitting refers to training algorithms like linear regression, decision trees, and neural networks to minimize error and improve predictive performance. Proper model fitting ensures that the selected model is both accurate and efficient, making it a critical step in statistical analysis, forecasting, and predictive modeling.

3.4.11 Forecasting

Forecasting is the process of estimating future trends, values, or events based on historical data, statistical techniques, and mathematical models. It plays a crucial role in decision-making across various domains, including finance, economics, business, weather prediction, and supply chain management. Forecasting helps organizations anticipate market demand, resource allocation, revenue projections, and risk assessment, enabling them to plan effectively and make data-driven decisions.

In time series forecasting, past observations are analyzed to predict future values. This approach includes methods like moving averages, exponential smoothing, and ARIMA models, which identify patterns such as trends, seasonality, and cyclic behavior. Qualitative forecasting relies on expert judgment, market research, and scenario analysis, making it useful in cases where historical data is limited or unreliable, such as new product launches or geopolitical events. Quantitative forecasting, on the other hand, uses statistical models, regression analysis, and machine learning algorithms to generate precise and data-driven predictions.

Forecasting can be classified based on the time horizon: short-term forecasting (days to months) is used for inventory management, staffing, and daily operations, while long-term forecasting (years to decades) aids in strategic planning, investment decisions, and economic policy formulation. Effective forecasting improves efficiency, reduces uncertainty, and enhances business competitiveness, making it an essential tool in analytics and predictive modelling.

3.4.12 Residual analysis

Residual analysis is a crucial step in model validation, used to evaluate how well a model captures patterns in data by examining the residuals—the differences between actual observed values and model-predicted values. It helps assess model accuracy, reliability, and assumptions, ensuring that predictions are unbiased and errors follow expected statistical properties. Residual analysis is widely applied in regression analysis, time series forecasting, and machine learning to diagnose model performance and identify potential improvements.

A well-fitted model should have residuals that are randomly distributed with zero mean and constant variance (homoscedasticity). If residuals show non-random patterns, such as systematic trends, increasing or decreasing variance, or autocorrelation, it may indicate model

deficiencies like underfitting (failing to capture essential patterns), overfitting (too closely modeling noise), or incorrect assumptions (such as non-linearity or omitted variables). Statistical tests like the Durbin-Watson test (for autocorrelation), Breusch-Pagan test (for heteroscedasticity), and normality checks (e.g., Q-Q plots, Shapiro-Wilk test) are commonly used to assess residual behavior.

By analyzing residuals, data scientists and analysts can refine models, select better features, and ensure that the assumptions underlying statistical models hold true. Effective residual analysis enhances forecasting accuracy, improves decision-making, and increases confidence in predictive modeling outcomes across various fields, including finance, economics, healthcare, and engineering.

3.4.13 Model evaluation

Model accuracy is assessed using various error metrics to evaluate how well a model's predictions align with actual values. Two widely used metrics in regression and time series forecasting are Mean Squared Error (MSE) and Root Mean Squared Error (RMSE).

MSE calculates the average squared difference between predicted and observed values, emphasizing larger errors due to squaring. It is given by:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where y_i represents actual values, $\hat{y}_i$ represents predicted values, and nn is the number of observations. A lower MSE indicates a better model fit.

RMSE is derived by taking the square root of MSE, making it easier to interpret since it is in the same units as the original data:

$$\text{RMSE} = \sqrt{MSE}$$

RMSE provides an intuitive measure of prediction error, with lower values indicating higher model accuracy.

Both MSE and RMSE are essential for comparing models, fine-tuning parameters, and ensuring reliable forecasting in various fields like finance, healthcare, and engineering. However, since they are sensitive to large errors, they are often complemented by other metrics

like Mean Absolute Error (MAE) or Mean Absolute Percentage Error (MAPE) for a comprehensive evaluation.

3.5 REGRESSION

Regression analysis is a powerful statistical technique used to model relationships between a dependent variable (response variable) and one or more independent variables (predictors or regressors). It helps identify the strength, direction, and significance of these relationships, making it a fundamental tool in forecasting, trend analysis, and decision-making across various fields like finance, economics, healthcare, and social sciences.

3.5.1 Linear Regression Model

Steps for Determining the Trend Line Using Linear Regression

1. Define Variables: Identify the independent variable (Year) and the dependent variable (Delinquency Rate) to detect trends over time.
2. Collect Data: Gather historical data covering multiple years to ensure accurate trend detection.
3. Preprocess Data: Check for missing values or inconsistencies and clean the dataset to ensure reliability.
4. Apply Linear Regression: Fit a Simple Linear Regression model to the data using the equation: $Y= \beta_0+ \beta_1 *X + \varepsilon$, where Y is the delinquency rate, X is the year, β_0 is the intercept, β_1 is the slope, and ε is the error term.
5. Interpret the Trend Line: Analyze the slope (β_1) to determine whether the delinquency rate is increasing, decreasing, or stable over time. A positive slope indicates an upward trend, while a negative slope suggests a downward trend.

3.6 SMOOTHING

In terms of Data analytics and Statistics smoothing refers to technique in which we can reduce the variability in data. By adopting smoothing techniques we can create a smooth representation of data. In this method, we usually calculate the average of nearby data points or use some mathematical model to the data trend. Smoothing methods are generally applied to those data sets which we are unable or difficult to find trend and patterns which may affect our prediction. These methods of smoothing are used in signal processing, time series analysis, machine learning etc

3.6.1 Exponential Smoothing

Exponential Smoothing is a widely used method in time series analysis for smoothing data and forecasting future values. It is particularly effective for datasets that exhibit trends or seasonality, making it applicable to real-world problems where such patterns are common. The method assigns exponentially decreasing weights to past observations, giving more importance to recent data points.

Steps in Exponential Smoothing Model

1. Initialize the Level Estimate: Choose an initial value for the level, which can be the first observation in the dataset or determined using statistical measures like the mean or mode.
2. Apply the Smoothing Equation: For each observation, compute the smoothed value using the formula:

$$S_t = \alpha Y_t + (1-\alpha)\, S_{t-1}$$

Where:

S_t = Smoothed value at time t

Y_t = Actual observation at time t

S_{t-1} = Smoothed value of the previous time period

α = Smoothing constant, controlling the level of smoothing

Update the Level Estimate: The newly computed smoothed value becomes the level estimate for the next observation.

3. Repeat the Process for All Observations: Continue applying the smoothing equation for all observations until the last data point is reached.
4. Forecast Future Values: Use the last smoothed value to estimate future data points based on the exponential smoothing formula.
5. Evaluate Forecast Accuracy: Compute error metrics such as:

   Mean Absolute Error (MAE)

   Mean Squared Error (MSE)

   Root Mean Squared Error (RMSE)

3.6.2 Holt-Winters Forecasting

Holt-Winters Exponential Smoothing is an extension of exponential smoothing that accounts for both trends and seasonality. It is commonly used for forecasting time series data where patterns repeat over fixed periods.

Steps in Holt-Winters Model

1. Initialize Components: Set initial values for the level (L), trend (T), and seasonal (S) components based on historical data.
2. Update Level, Trend, and Seasonality: Compute the smoothed values using the following equations:
   Level:

$$L_t = \alpha\frac{Y_t}{S_{t-m}} + (1-\alpha)\,(L_{t-1}+T_{t-1})$$

   Trend:

$$T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1}$$

   Seasonality:

$$S_t = \gamma\frac{Y_t}{L_t} + (1-\gamma)S_{t-m}$$

Where:

L_t = Level component at time t

T_t = Trend component at time t

S_t = Seasonal component at time t

m = Length of the seasonal cycle

α, β, γ = Smoothing constants controlling how much weight is given to recent data

3. Repeat for All Observations: Update L, T, and S for each time step using the above formulas.
4. Forecast Future Values: Use the last computed values of L, T, and S to forecast future data points using: $\widehat{Y_{t+h}} = (L_t + hT_t)S_{t+h-m}$,where h is the forecasting horizon.
5. Evaluate Forecast Accuracy: Measure performance using MAE, MSE, and RMSE to assess the accuracy of the forecast.

3.7 TOOLS USED FOR COMPARISON

The present study uses Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) for comparison of models.

Mean Squared Error (MSE) is the average value of squared difference between actual and predicted values.

Root Mean Squared Error (RMSE) is the square root of Mean Squared Error (MSE).

Both Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) can be used to compare the difference model performance on certain data.

Chapter 4

RESULTS AND ANALYSIS

4.1 DATA DESCRIPTION

For the purpose of the study, the data was collected from the National Crime Records Bureau official website (https://ncrb.gov.in). The data consists of yearly records from 1970 to 2022. This dataset provides a comprehensive overview of juvenile delinquency rates across the specified years. For analysis, the data has been organized into a structured format to facilitate insights and comparisons over the study period.

4.2 ARIMA MODELLING

To analyze and predict juvenile delinquency trends in India, the ARIMA model was used to forecast delinquency rates for the years 2023 to 2027. ARIMA is a powerful tool for time series analysis, as it accounts for non-seasonal patterns by utilizing historical trends and assumes the data is stationary and univariate.

4.2.1 Time series plot

Figure 4.1 is the time series plot of the juvenile delinquency rates data from 1970 to 2022.

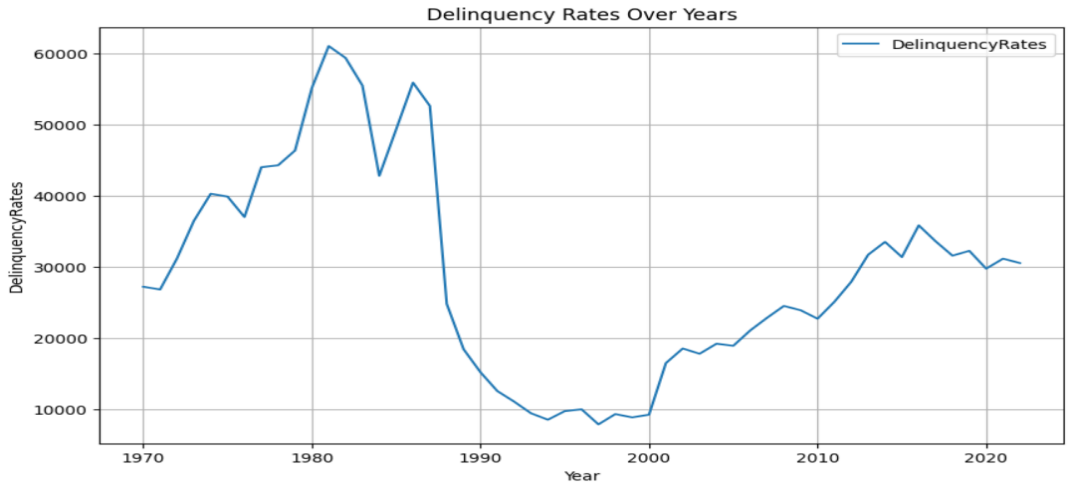

Figure 4.1 Time series plot of the juvenile delinquency rates

4.2.2 Decomposition of time

The second step perform seasonal decomposition to capture the trend, seasonal and random components of time series. Figure 4.2 depict the seasonal plot.

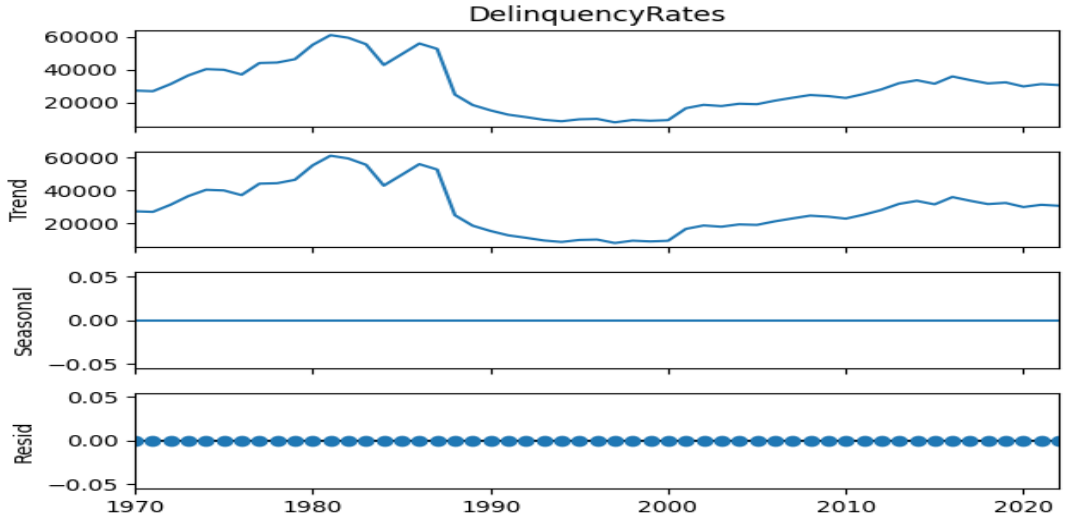

Figure 4.2 Seasonal decomposition plot

4.2.3 Stationarity check using Augmented Dickey- Fuller Test

To test the time series data for stationarity using ADF test, follows a hypothesis testing approach. The null hypothesis H0 is given by,

H0: The data is non stationary.

The alternative hypothesis H1 is given by,

H1: The data is stationary.

The outcome achieved, ADF test statistic = -2.4675596535198423, Lag order = 6, p-value = 0.12356821958177372

The ADF test gives the p-value 0.12356821958177372, which is greater than 0.05, so accept the null hypothesis i.e.; the data is non stationary, hence we perform n order differencing until we get time series stationary We perform differencing with n = 1 Now we again check stationarity using ADF test.

Here we test the hypothesis,

H0: The data is non-stationary.

Against

H1: The data is stationary.

ADF test statistic =-5.351559, Lag order = 6, p-value= 0.000004 The ADF test gives the p-value0.000004, which is smaller than 0.05, so reject the null hypothesis H0 and Hence, we can conclude that data is stationary; Figure 4.3 shows the differenced delinquency rates.

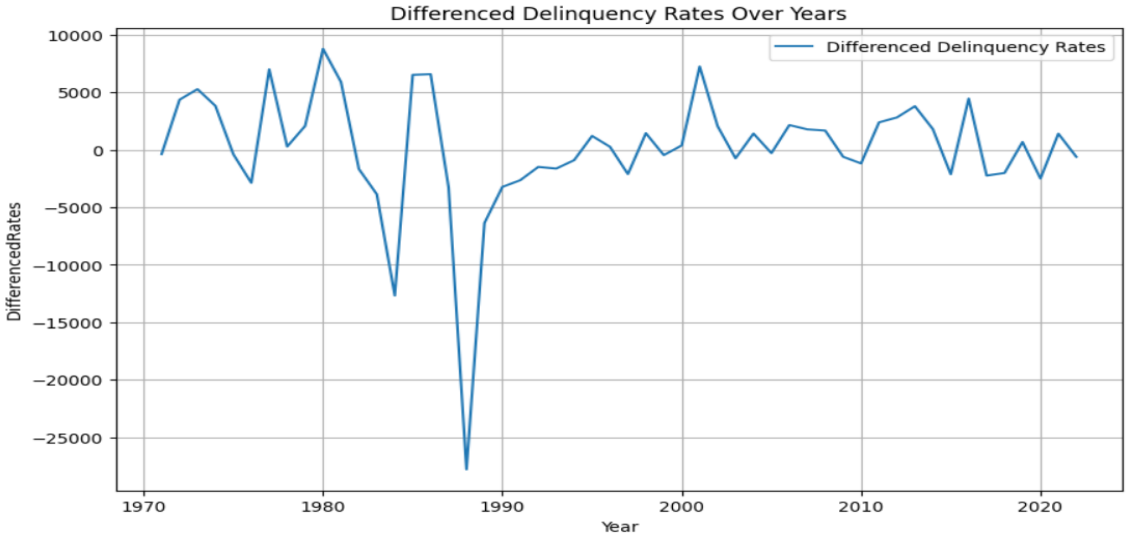

Figure 4.3 Plot of differenced delinquency rates

4.2.4 Autocorrelation and Partial Autocorrelation Function

Next step in Time Series Analysis is to plot and examine Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF). ACF &PACF Plot is given in Figure 4.4.

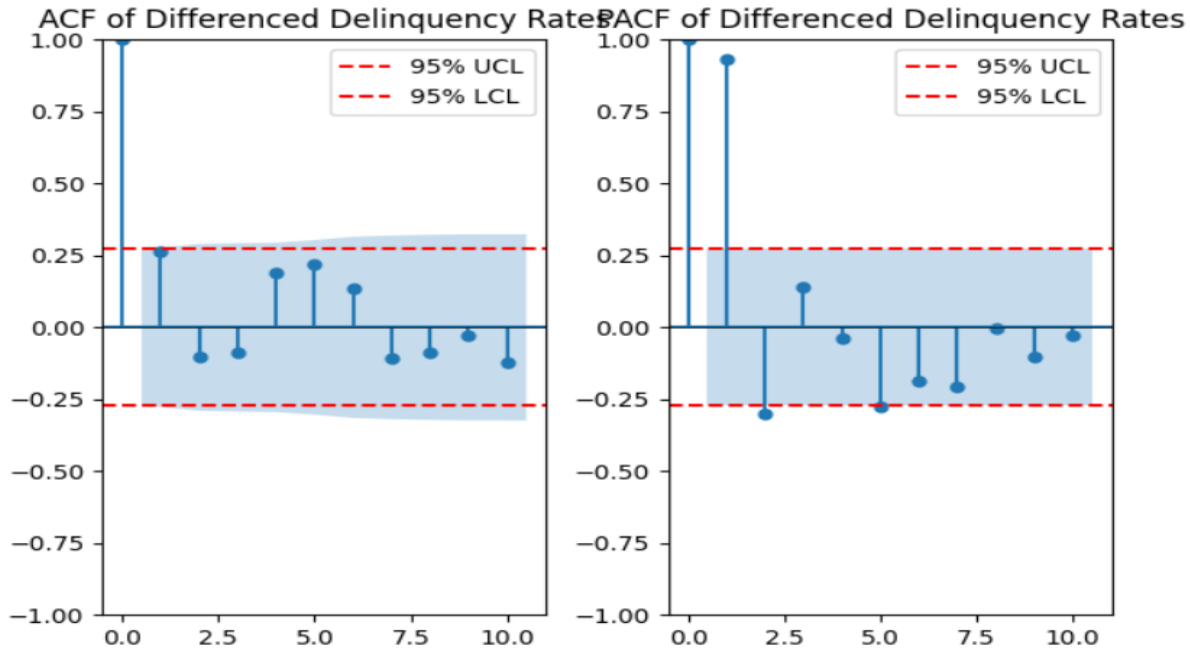

Figure 4.4 Autocorrelation Function & Partial Autocorrelation Function Plot

4.2.5 ARIMA Model for delinquency rates

In this step we choose the best model for forecasting the values. It is done by choosing one model from all possible models according to Akaike Information Criterion (AIC). The model with lowest AIC value is chosen as the best model. Below Table 4.1 shows the possible models with their AIC values.

Table 4.1 The model with AIC value

SL NO.	MODEL ARIMA (p, d, q) x (P, D, Q)	AIC
1	ARIMA (0, 1, 0) x (0, 0, 0)	1004.0658
2	ARIMA (0, 1, 1) x (0, 0, 0)	981.8864
3	ARIMA (0, 1, 2) x (0, 0, 0)	964.6589
4	ARIMA (0, 1, 3) x (0, 0, 0)	946.6425
5	ARIMA (0, 1, 4) x (0, 0, 0)	928.1296
6	ARIMA (0, 1, 5) x (0, 0, 0)	907.7286
7	ARIMA (1, 1, 0) x (0, 0, 0)	1002.4053
8	ARIMA (1, 1, 1) x (0, 0, 0)	983.8744
9	ARIMA (1, 1, 2) x (0, 0, 0)	966.5088
10	ARIMA (1, 1, 3) x (0, 0, 0)	948.5292
11	ARIMA (1, 1, 3) x (0, 0, 0)	927.9114
12	ARIMA (1, 1, 5) x (0, 0, 0)	908.7660
13	ARIMA (2, 1, 0) x (0, 0, 0)	983.0972
14	ARIMA (2, 1, 1) x (0, 0, 0)	985.1166
15	ARIMA (2, 1, 2) x (0, 0, 0)	967.2678
16	ARIMA (2, 1, 3) x (0, 0, 0)	944.6161
17	ARIMA (2, 1, 4) x (0, 0, 0)	925.0720
18	ARIMA (2, 1, 5) x (0, 0, 0)	907.4081
19	ARIMA (3, 1, 0) x (0, 0, 0)	965.8004
20	ARIMA (3, 1, 1) x (0, 0, 0)	965.8859
21	ARIMA (3, 1, 2) x (0, 0, 0)	965.4729
22	ARIMA (3, 1, 3) x (0, 0, 0)	939.9363
23	ARIMA (3, 1, 4) x (0, 0, 0)	919.5848
24	ARIMA (3, 1, 5) x (0, 0, 0)	909.4331
25	ARIMA (4, 1, 0) x (0, 0, 0)	946.4052
26	ARIMA (4, 1, 1) x (0, 0, 0)	947.9586
27	ARIMA (4, 1, 2) x (0, 0, 0)	944.5428
28	ARIMA (4, 1, 3) x (0, 0, 0)	940.6651
29	ARIMA (4, 1, 4) x (0, 0, 0)	921.2991
30	**ARIMA (4, 1, 5) x (0, 0, 0)**	**905.3694**
31	ARIMA (5, 1, 0) x (0, 0, 0)	928.7157
32	ARIMA (5, 1, 1) x (0, 0, 0)	930.3933
33	ARIMA (5, 1, 2) x (0, 0, 0)	926.9935

34	ARIMA (5, 1, 3) x (0, 0, 0)	923.2703
35	ARIMA (5, 1, 4) x (0, 0, 0)	924.5933
36	ARIMA (5, 1, 5) x (0, 0, 0)	907.3538

Here the best model is ARIMA (4, 1, 5) x (0, 0, 0) with AIC value 905.3694

Coefficients:

	ar.L1	ar.L2	ar.L3	ar.L4	ma.L1	ma.L2	ma.L3	ma.L4	ma.L5
	0.8331	-0.8835	0.6520	-0.2426	-1.1904	0.9380	-0.6797	0.7338	0.6324
std err	0.206	0.272	0.266	0.162	0.301	0.481	0.424	0.444	0.432

4.2.6 Diagnostic checking

Diagnostics checking is performed for confirming the validity, effectiveness and reliability of statistical models. The main objective of it is to choose the right and best model. Below figure 4.5 shows the diagnostic plot.

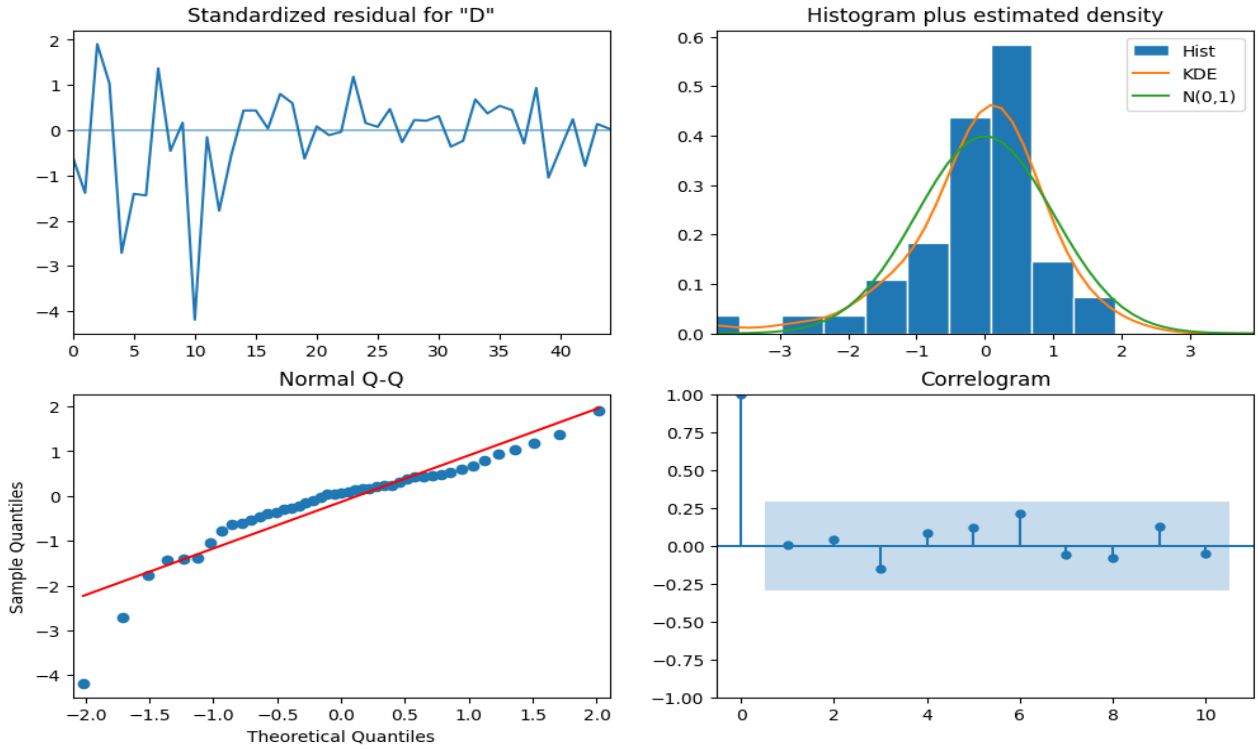

Figure 4.5 Diagnostic plot

From Q-Q plot, it is clear that most of the residuals are on the same line and standard residual are normally fitted

4.2.7 Forecasting the Sample

Forecasting the future and predicting the actual data points or the training data points. Here evaluate model performance on training data. Figure 4.6 is the plot of actual and predicted values.

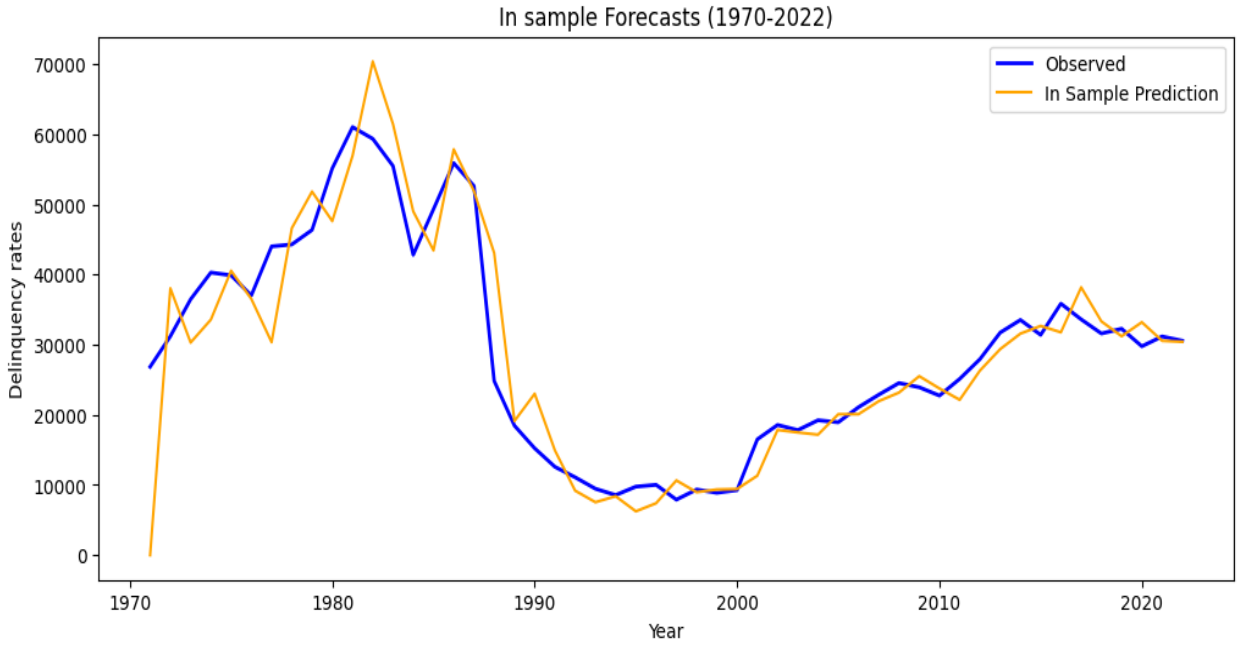

Figure 4.6 Plot of actual and predicted values

4.2.8 Forecasting the Future Values

The forecasted juvenile delinquency rates data from 2023 to 2027 is given in Table 4.2.

Table 4.2 Forecasted juvenile delinquency rates

Year	Forecasted Values	LCL	UCL
2023	40089.361223	20439.884961	37532.837485
2024	41617.647742	15617.635356	42817.660129
2025	43119.399848	11747.873881	46290.925816
2026	44491.022597	8265.235837	47716.809356
2027	46065.160297	6084.528473	49685.792120

The Figure 4.7 shown below is the graph of forecasted juvenile delinquency rates data from 2023 to 2027.

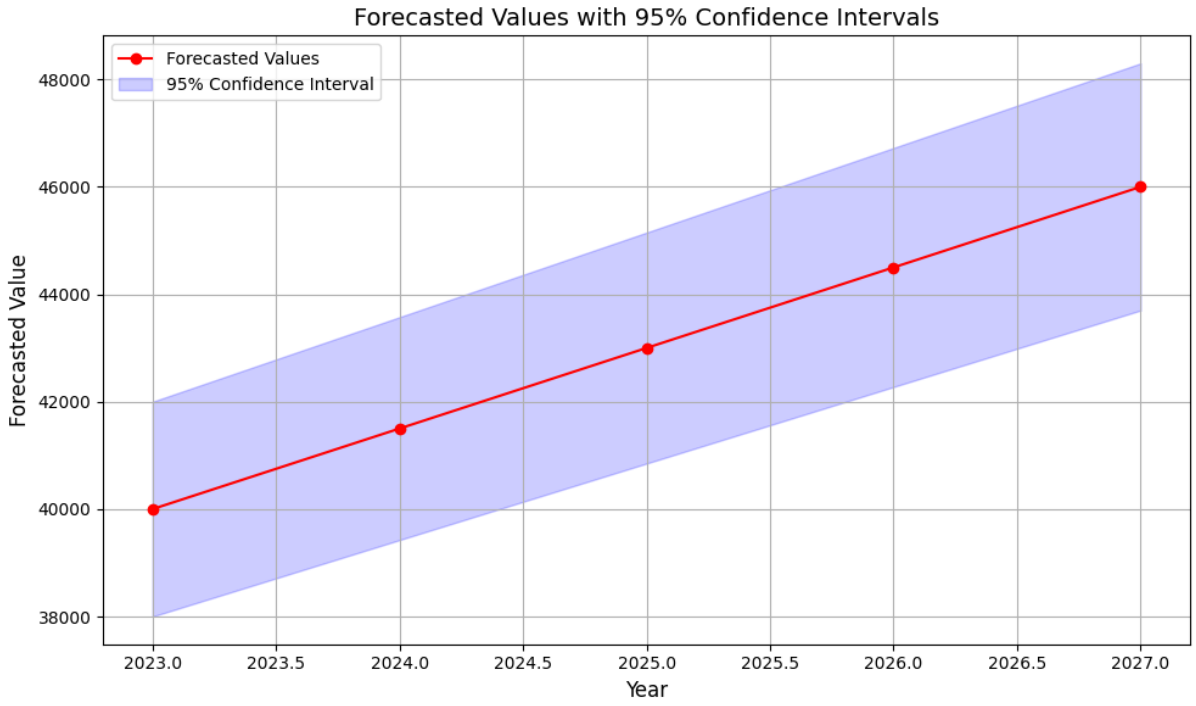

Figure 4.7 forecasted juvenile delinquency rates plot

4.3 REGRESSION ANALYSIS MODEL

Linear Regression was applied to analyze the trend line, as it effectively studies the relationship between variables under the assumption that the dependent variable changes proportionally with the independent variable.

4.3.1 LINEAR REGRESSION MODEL

Figure 4.8 is the plot of actual Juvenile Delinquency Rate Over Years and Figure 4.9 is the plot of linear trend line of Delinquency Rate

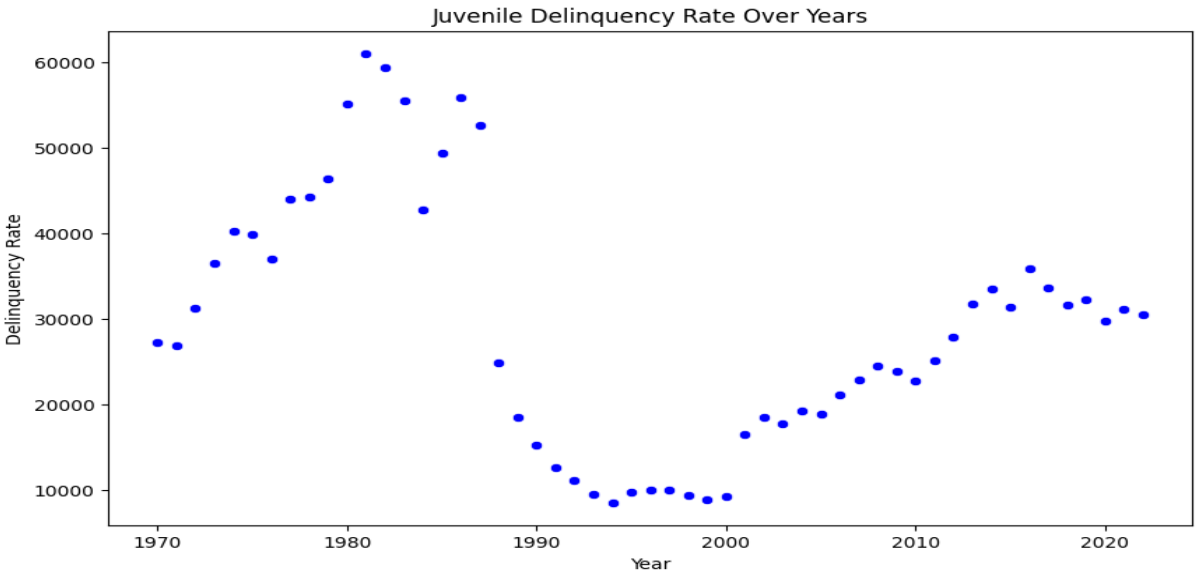

Figure 4.8 plot of actual Juvenile Delinquency Rate Over Years using linear regression

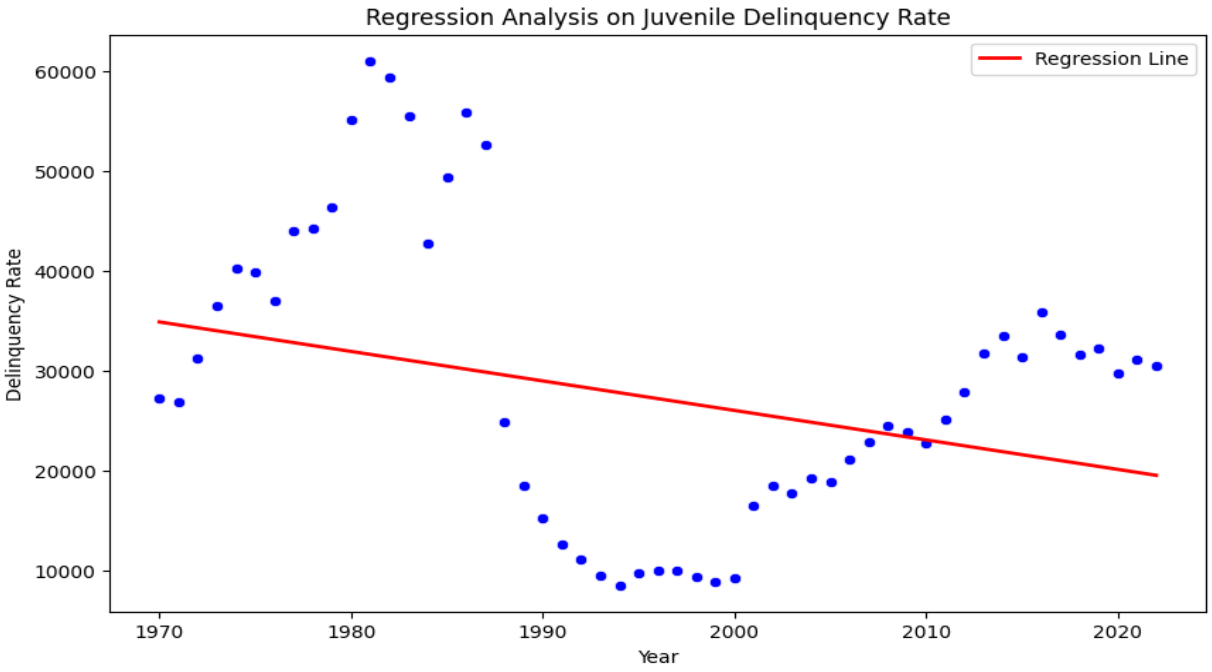

Figure 4.9 Plot of linear trend line of Delinquency Rate

4.4 EXPONENTIAL SMOOTHING MODEL

In this study, the smoothing parameter (alpha) is a key constant chosen based on the characteristics of the data. A smaller alpha value results in greater smoothing, reducing noise and fluctuations, while a larger value makes the model more responsive to recent changes. To effectively capture trends and patterns in juvenile delinquency rates, an alpha value of 0.1 was selected. This choice ensures a balanced smoothing effect, helping to remove short-term variations while preserving the overall trend, making it suitable for forecasting and trend analysis using Exponential Smoothing and Holt-Winters Forecasting.

4.4.1 Forecasting future values

Using Python and the equations for Single Exponential Smoothing and Holt-Winters Forecasting, we forecasted the juvenile delinquency rate for the years 2023 to 2032. The table 4.3 shown below presents the forecasted values obtained through these methods, providing insights into future trends in juvenile delinquency based on historical patterns.

Table 4.3 Forecasted juvenile delinquency rates using Holt-Winters Forecasting

Year	Holt Winters Forecast
2023	30712.529571
2024	30865.623619
2025	31018.717667
2026	31171.811716
2027	31324.905764
2028	31477.999812
2029	31631.093860
2030	31784.187908
2031	31937.281956
2032	32090.376004

4.4.2 Forecasting the Sample

Forecasting the future involves predicting the actual data points based on historical trends. In this study, model performance is evaluated using the training data. Figure 4.10 presents the plot of actual versus predicted values, showcasing how well the model fits the historical data. Additionally, displays the forecasted juvenile delinquency rates from 2023 to 2032, using the Holt-Winters Forecasting Model to capture both trend and seasonal variations for better accuracy.

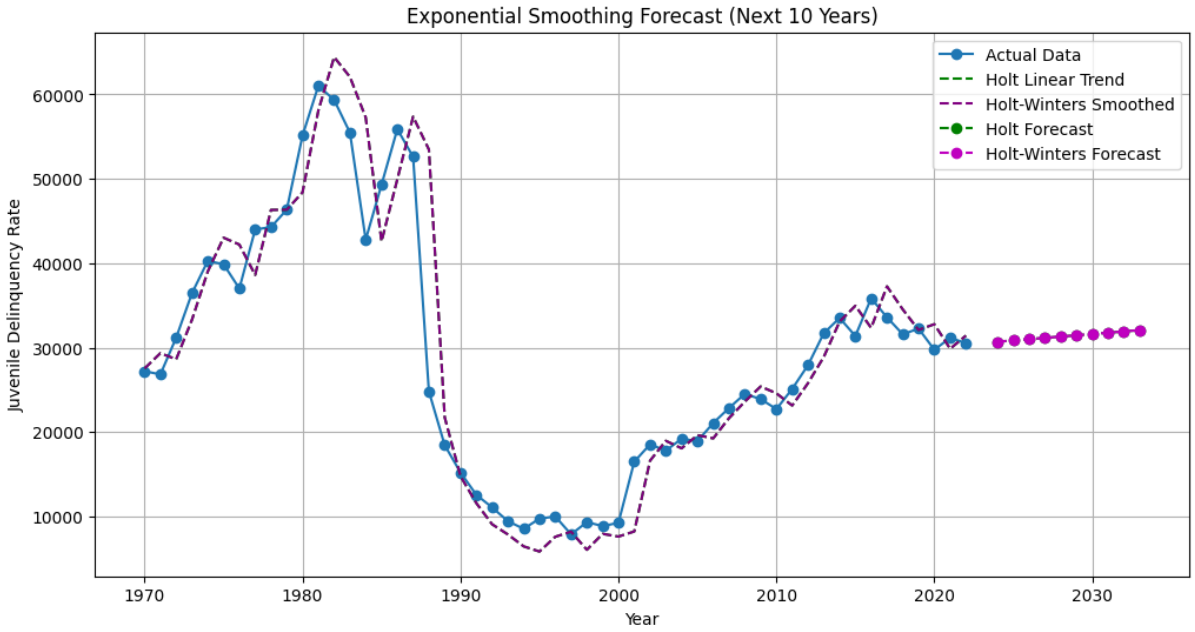

Figure 4.10 plot of actual versus predicted values and forecasted juvenile delinquency rates

4.5 COMPARISON OF RMSE AND MSE VALUES

To determine the best model between ARIMA and the Holt-Winters Forecasting Model, a comparison of their RMSE values is conducted. Mean Squared Error (MSE) represents the average squared difference between actual and predicted values, while Root Mean Squared Error (RMSE) is the square root of MSE, providing a more interpretable measure of error. The comparison of RMSE values for the ARIMA model and Holt-Winters Forecasting Model is presented in Table 4.4, highlighting their forecasting accuracy.

Table 4.4 The comparison of RMSE values for the ARIMA model and Holt-Winters Forecasting Model

	ARIMA model	**Exponential smoothing model**
RMSE	6093.8429	5443.08

From Table 4.4, it is evident that the Holt-Winters Forecasting Model has a smaller RMSE value compared to the ARIMA model. Based on the RMSE values, it can be concluded that the Holt-Winters Forecasting Model performs better than the ARIMA model and can be considered the best model for this study in terms of forecasting accuracy.

To enhance prediction accuracy, Exponential Smoothing and Holt-Winters Forecasting were also implemented. Exponential Smoothing is a widely used forecasting technique that assigns

exponentially decreasing weights to past observations, making it useful for smoothing data and predicting future trends. The Holt-Winters Model, an extension of Exponential Smoothing, was applied to capture both trend and seasonal components in the data, improving forecasting precision.

Lastly, a comparative analysis was conducted between ARIMA and Holt-Winters Forecasting to evaluate their predictive performance. By assessing different forecasting techniques, this study ensures a comprehensive understanding of juvenile delinquency trends in India, allowing for more informed policy decisions.

Chapter 5
CONCLUSION

This study aimed to analyze and forecast juvenile delinquency trends in India using historical data from 1970 to 2022. Through the application of various time series forecasting methodologies, including ARIMA, Linear Regression, Exponential Smoothing, and Holt-Winters Forecasting, a comprehensive understanding of the patterns and trends in juvenile delinquency rates was developed.

The analysis began with data visualization and pre-processing steps, including stationarity checks and decomposition of time series components to identify trends and seasonality. The Augmented Dickey-Fuller (ADF) test was employed to assess the stationarity of the data. The initial test confirmed the presence of non-stationarity, leading to first-order differencing to achieve stationarity, which was essential for further time series modeling.

The ARIMA model was extensively tested by evaluating various parameter combinations to select the best-fitting model based on the Akaike Information Criterion (AIC). The best-performing ARIMA model was identified as ARIMA (4,1,5) with an AIC value of 905.3694. Diagnostic checks, including residual analysis, were conducted to validate the model's reliability. The ARIMA model was then used to forecast juvenile delinquency rates for the period 2023 to 2027.

Additionally, a Linear Regression model was employed to analyze long-term trends in delinquency rates. The linear trend analysis provided valuable insights into the overall direction of juvenile delinquency rates over time but lacked the capability to capture non-linearity and seasonal variations in the data.

To further enhance forecasting accuracy, Exponential Smoothing and the Holt-Winters Forecasting Model were applied. The Exponential Smoothing model effectively reduced short-term fluctuations by assigning exponentially decreasing weights to past observations, ensuring a smoother trend representation. The Holt-Winters method extended this approach by incorporating trend and seasonal components, leading to more precise long-term forecasts. The forecasted juvenile delinquency rates using the Holt-Winters method covered the period from 2023 to 2032, offering an extended predictive outlook.

A comparative analysis was conducted between the ARIMA model and the Holt-Winters Forecasting Model using Root Mean Squared Error (RMSE) as a performance metric. The RMSE values for both models indicated that the Holt-Winters Forecasting Model (RMSE = 5443.08) outperformed the ARIMA model (RMSE = 6093.8429) in terms of predictive accuracy. The lower RMSE value of the Holt-Winters model signifies its superior ability to capture underlying patterns in juvenile delinquency trends.

Key Findings:

1. The juvenile delinquency rate data exhibited non-stationarity, which was addressed through differencing techniques.

2. The ARIMA model was effective in short-term forecasting, with the ARIMA(4,1,5) model identified as the best fit.
3. Linear regression provided insights into the overall trend but was limited in capturing complex variations in delinquency rates.
4. The Holt-Winters Forecasting Model effectively captured both trend and seasonal variations, resulting in more accurate long-term forecasts.
5. The comparative RMSE analysis confirmed that the Holt-Winters model outperformed the ARIMA model, making it the preferred choice for forecasting juvenile delinquency rates.

Implications and Future Scope: The findings of this study have significant implications for policymakers, law enforcement agencies, and social organizations aiming to mitigate juvenile delinquency. The use of advanced forecasting techniques can aid in formulating effective intervention strategies by predicting future delinquency trends and enabling proactive measures.

Future research could focus on incorporating additional socio-economic variables, such as education levels, unemployment rates, and family structures, to improve the predictive capabilities of the models. Additionally, machine learning techniques like Long Short-Term Memory (LSTM) networks and hybrid models could be explored for further enhancements in forecasting accuracy.

Overall, this study provides a robust framework for understanding and predicting juvenile delinquency trends in India, contributing valuable insights for effective policy formulation and crime prevention strategies.

REFERENCES

1. Acharya, S. (2017). The study of juvenile delinquency with reference to psychological perspectives in the juvenile homes of Delhi. *International Journal of Advanced Research in Management and Social Sciences, 6*(11), 13–27.
2. Agyemang, B. (2012). *Autoregressive integrated moving average (ARIMA) intervention analysis model for the major crimes in Ghana (The case of the Eastern Region)* (Doctoral dissertation). [Institution Name, if available].
3. Atrey, I., & Singh, B. (2023). An analysis of juvenile delinquency in Rajasthan: Risk factors and protective factors. *Journal of Research Administration - SRA International, 5*(2).
4. Devi, J. V., & Kavitha, K. (2021). Automating time series forecasting on crime data using RNN-LSTM. *International Journal of Advanced Computer Science and Applications, 12*(10), 458–463.
5. Dhaka, S. K. (2021). Juvenile delinquency in Delhi (India): Latest trends and new amendments in juvenile justice. *Research Journey, 102*.
6. Haveripet, P. (2013). Causes and consequences of juvenile delinquency in India. *Recent Research in Science and Technology, 5*(3).
7. Khairuddin, A. R., Alwee, R., & Haron, H. (2019, August). A review on applied statistical and artificial intelligence techniques in crime forecasting. *IOP Conference Series: Materials Science and Engineering, 551*(1), 012030. IOP Publishing.
8. Khan, S. (2022). A study on juvenile delinquency in India: Sociological aspect and judicial response. *Supremo Amicus, 28*, 478.
9. Muthamizharasan, M., & Ponnusamy, R. (2024). A comparative study of crime event forecasting using ARIMA versus LSTM model. *Journal of Theoretical and Applied Information Technology, 102*, 2162–2171.
10. Sharma, B. R., Dhillon, S., & Bano, S. (2009). Juvenile delinquency in India—a cause for concern. *Journal of Indian Academy of Forensic Medicine, 31*(1), 68–72.