Crime Data Analytics

Capstone Project II

MBA702

by

Karan Jain (210491)

Under the Supervision of

Prof. Deepu Philip

DEPARTMENT OF MANAGEMENT SCIENCES

INDIAN INSTITUTE OF TECHNOLOGY KANPUR

April 2025

Certificate

This is to certify that the project entitled “Crime Data Analytics” was submitted to the Department of Management Sciences, IIT Kanpur, as a Capstone Project. It is work done by Karan Jain, under my supervision and guidance.

Prof. Deepu Philip
Professor
Department of Management Sciences
Indian Institute of Technology Kanpur

Declaration

This is to certify that the report titled “Crime Data Analytics” has been authored by me. It presents the Capstone project conducted by me under the supervision of Prof. Deepu Philip.

Karan Jain
Roll No. 210491
MSE Department
Indian Institute of Technology Kanpur

Abstract

This report presents a comprehensive district-level crime study based on National Crime Records Bureau (NCRB) data (2001–2012) and eight-year forecasts (2013–2020) generated through Holt-Winters Exponential Smoothing (additive trend, no seasonality). A reproducible workflow—combining Excel dashboarding and Python-assisted validation—was developed. Key outputs include (i) crime statistics district-wise, (ii) trend-slope ranking of districts, and (iii) forecast accuracy assessment. Actionable recommendations are offered for resource allocation and proactive policing.

Acknowledgements

I would like to express my sincere gratitude to **Prof. Deepu Philip** for his invaluable guidance, support, and mentorship throughout the course of this capstone project. His deep expertise in Data Analytics as well as his insightful feedback and discussions have been instrumental in the successful completion of this work.

Contents

1. Introduction

Crime analysis plays an important role in understanding social challenges and in guiding policy decisions. The Census of India and the National Crime Records Bureau (NCRB) have been publishing detailed crime data till the year 2012, listing district-wise incidents across different categories such as murder, theft, kidnapping, arson, and riots.

While this information reflects the situation till 2012, there is a growing need to forecast future crime trends to assist law enforcement agencies and policymakers in better planning. In this project, an attempt has been made to **forecast district-wise crime figures for the years 2013 to 2020**, using historical data available till 2012.

Before undertaking the analysis, the raw NCRB data was thoroughly cleaned and organised using **Excel tools like Pivot Tables and Power Query Editor**. After ensuring that the dataset was consistent and analysis-ready, forecasting was carried out using **Python**, employing exponential smoothing techniques.

2. Objectives and Scope

The objectives of this work are outlined below:

- To clean and prepare district-wise crime data from NCRB reports for the years 2001–2012, ensuring it is in a suitable format for analysis.
- To apply exponential smoothing methods to forecast the crime figures for each district from 2013 to 2020.
- To validate the forecasts by comparing them with the actual (cleaned) data available for 2013 and 2014.
- To identify districts that are projected to experience significant increases in crime and require focused intervention using statistics.
- To provide a foundation for further development of predictive crime models using more advanced techniques in the future.

The scope of this analysis is limited to the overall number of reported crimes per district. It does not separately model each crime category. However, the methodology adopted here can easily be extended to individual categories if required.

3. Forecasting

3.1 Data Preparation

The initial step involved meticulous data preparation. The available NCRB data was not ready for direct analysis, as it contained inconsistencies across districts. Therefore, using Excel's Pivot Tables and Power Query Editor, the data was carefully cleaned, standardised, and reshaped into a year-wise district-wise format.

This cleaned dataset was then exported as a CSV file and used for further processing in Python.

3.2 Application of **Exponential Smoothing**

The forecasting model used was the Holt-Winters Exponential Smoothing method, excluding seasonality since we were dealing with annual data. Using Python's statsmodels library, different combinations of smoothing parameters (alpha and beta) were tried to achieve the best fit for each district's historical data.

3.3 Validation on 2013–2014

The forecasted values for the years 2013 and 2014 were compared with the actual crime counts (which were separately cleaned from NCRB reports). The forecasting model showed strong accuracy:
- The mean absolute percentage error (MAPE) across districts was around 8.7%.
- More than 85% of districts had a MAPE below 10%.
- Forecast deviations were mainly seen in districts with very low or highly volatile crime numbers.

3.4 Forecasts for 2015–2020

The forecasts from 2015 to 2020 revealed important observations:
- Urban districts such as Bengaluru Urban, Thane, and Pune showed consistently rising trends.
- Districts in Kerala and Tamil Nadu reflected stabilising or slightly declining trends.
- Emerging hotspots were seen in Telangana and parts of North-East India.

3.5 Key Findings

- Districts with high predicted crime growth need urgent attention.
- Flat or declining trend districts can serve as models for best practices.
- Smaller districts with high volatility require dynamic, responsive policing.

3.6 Limitations

- Seasonality was not considered.
- External socio-economic influences were not factored.
- Data quality issues like under-reporting and district boundary changes may affect accuracy

4. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a statistical method used to simplify complex datasets by reducing the number of variables while retaining most of the original information. In this case, where we have district-wise mean values of over 20 crime categories, PCA helps us summarise these into a smaller number of uncorrelated indicators, known as principal components.

This technique is particularly useful because it:
- Reduces dimensionality, making analysis more manageable.
- Eliminates correlation among variables, which helps improve the stability of statistical models.
- Highlights latent patterns that might not be obvious when examining raw figures.

4.1 Importance of PCA in Crime Analysis

For district-wise crime analysis, PCA is highly beneficial because:
- It allows us to focus on a few core indicators instead of working with 20+ separate crime categories.
- It helps us understand broader trends—for example, whether a district generally has a high crime load or is affected by specific crime types.
- It supports better forecasting and clustering by removing redundancy and making the data easier to interpret.

4.2 Interpretation of the Three Principal Components

We extracted three principal components from the dataset. Together, they account for nearly 98% of the total variation observed in district-wise crime patterns. Each component represents a different underlying dimension of crime.

Component	% Variance Explained	Key Contributing Crime Types	Interpretation
PC1	~86%	Total IPC Crimes, Rape, Murder, Assault on Women, Other IPC Crimes	Reflects the overall crime burden in a district. A high PC1 score indicates a district with high crime figures across most categories.
PC2	~9%	Theft, Counterfeiting, Burglary, Robbery, Culpable Homicide	Highlights the mix between property and violent crimes. Districts with high PC2 scores are more affected by property-related crimes.
PC3	~3%	Arson, Grievous Hurt, Kidnapping, Importation of Girls	Captures less frequent or niche crimes that are still significant in certain districts.

4.3 Practical Implications

- Districts with a high PC1 score are likely to require broad crime-reduction strategies across multiple domains.
- Districts that score high on PC2 may benefit more from targeted efforts to curb theft and similar offences.
- A high PC3 score might indicate the need for specialised interventions—such as anti-trafficking units or arson investigation teams.

Thus, PCA has helped us compress complex, multi-dimensional crime data into three meaningful indicators, which in turn make analysis, planning, and forecasting more efficient and impactful.

5. Understanding the Trend Slope

The Trend Slope refers to the rate at which crime incidents have been increasing or decreasing in a district over a period of time. In this study, the trend slope for each district was calculated by fitting a straight line through its annual crime data from 2001 to 2012, using the method of least squares. The slope of this line indicates, on average, how much the number of crimes is rising or falling each year.

5.1 Importance of Trend Slope in District-Wise Crime Analysis

In the context of district-wise crime analysis, the trend slope plays a crucial role for several reasons:

- Identifying hotspots and improvements: Districts with a high positive trend slope are experiencing a continuous rise in crime and need immediate attention, while those with a negative slope show improvement.
- Better allocation of resources: Resources can be prioritised to areas where crime is increasing sharply.
- Enhancing forecasting models: Including trend slope improves the accuracy of predicting future crime rates.
- Guiding policy making and planning: Policymakers can quickly grasp district-wise crime trends and plan interventions accordingly.
- Simplifying district comparisons: Trend slopes offer a simple yet powerful number to compare districts across the country.

Thus, the trend slope serves as a powerful tool to condense years of crime data into a single, actionable figure, making it easier for law enforcement agencies and policymakers to identify rising concerns and take timely corrective measures.

5.2 Results and Discussions

The trend slope was calculated individually for each district based on its annual crime data spanning the period from 2001 to 2012.

By fitting a straight line through the yearly crime counts, the slope provides a clear measure of whether crime incidents have been rising or falling over time in that district.

6. Conclusion

This project demonstrates that even with relatively simple models like exponential smoothing, reliable and actionable crime forecasts can be produced. By starting with proper data cleaning using Excel, and then applying Python-based time series modelling, it was possible to project crime trends district-wise up to the year 2020 with good accuracy.

Key achievements from the study include:

- Forecasting district-wise total crime incidents for the years 2013 to 2020 using Exponential Smoothing.

- Calculating Principal Component Analysis (PCA) to summarise crime patterns and reduce dimensionality.

- Computing the mean and standard deviation for each district across various crime categories to understand average crime levels and volatility.

- Estimating the trend slope for each district to identify whether crime was generally rising or falling over the period 2001–2012.

Going forward, this approach can be further strengthened by including monthly or quarterly data, incorporating external variables, and adopting advanced machine learning models.

This work lays a strong foundation for predictive crime analytics in India, combining simplicity, transparency, and effectiveness.

Bibliography

1. National Crime Records Bureau (NCRB). *Crime in India Reports (2001–2012)*. Ministry of Home Affairs, Government of India. Available at: https://ncrb.gov.in/en/crime-india

2. Jolliffe, I. T. (2002). *Principal Component Analysis (2nd ed.).* Springer Series in Statistics. Springer-Verlag, New York.
   [A foundational text on the statistical theory and practical applications of PCA.]

3. Hotelling, H. (1933). *Analysis of a complex of statistical variables into principal components.* Journal of Educational Psychology, 24(6), 417–441.
   [The original work that introduced PCA in multivariate statistics.]

4. Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (1998). *Forecasting: Methods and Applications (3rd ed.).*Wiley.