INFO 5810

Fall 2023

Prediction of India’s GDP Growth Rate

PROJECT REPORT

UNIVERSITY OF NORTH TEXAS DEPARTMENT OF INFORMATION SCIENCE

Prediction of India's GDP Growth Rate
by

Sai Teja Rayabarapu (Student ID: 11637122)

SaiTejaRayabarapu@my.unt.edu

Major: Information Science

Manaswini Kodela (Student ID: 11637117)

ManaswiniKodela@my.unt.edu

Major: Information Science

Vivek Raj Rampelli (Student ID: 11596805)

VivekRampelli@my.unt.edu

Major: Information Science

Nikhil Vamsi Gutti (Student ID:11588082)

NikhilVamsiGutti@my.unt.edu

Major: Data Science

Instructor's Name: Syed Haider Ali

Course Title: Data Analysis and Knowledge Discovery

Number: INFO 5810 Section 020 – (Fall 2023)

Date: 10/09/2023

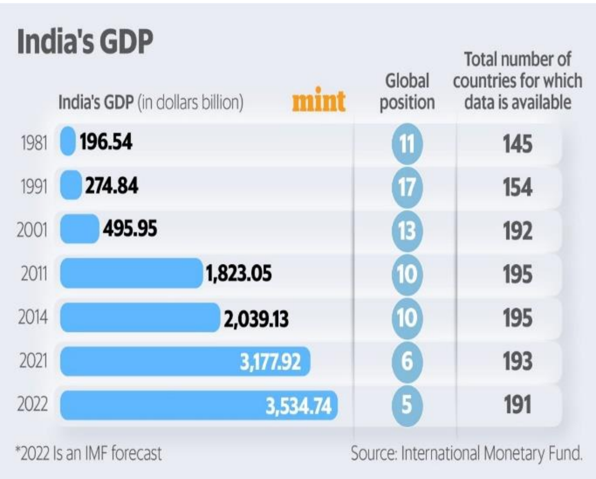
CONTENTS

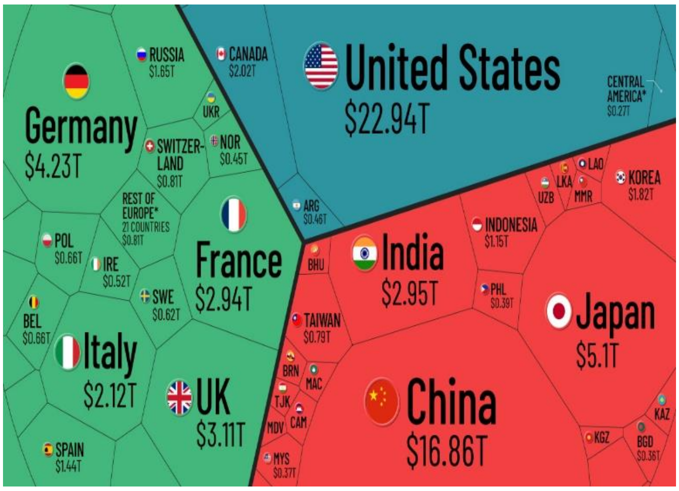
ABSTRACT

The "Prediction of India's GDP Growth Rate" project aims to develop a robust machine-learning model for forecasting India's yearly GDP growth rate over the next decade. By incorporating historical economic data and pertinent factors such as imports, exports, inflation, and population, the project employs feature engineering, preprocessing, and data gathering techniques. Utilizing both classical and time series machine learning algorithms, the models undergo rigorous evaluation to identify the most accurate one. The resulting comprehensive report includes analyses and visual interpretations of the model's predictions, serving as a valuable tool for stakeholders in the Indian economy, including policymakers, investors, and businesses. The initiative not only provides insightful information about India's economic growth trajectory but also lays the foundation for further study and policy development, contributing to informed decision-making for the country's economic stability and expansion. The use of a Kaggle dataset ensures the availability of rich and pertinent data for thorough examination.

1. INTRODUCTION

1.1 Background





The Gross Domestic Product (GDP) is a crucial indicator of a country's economic performance and health. It serves as an important gauge for economists, decision-makers in government policy, investors, and companies alike since it captures the total value of goods and services produced inside a nation's boundaries. Not only is it of academic importance, but understanding and correctly forecasting GDP growth also has significant practical implications. Forecasting GDP growth becomes crucial when considering India, the second-most populated nation in the world and one of the major economies with the quickest growth rates.

Over the past few decades, the Indian economy has seen considerable changes. The trajectory of India's GDP development has been impacted by a number of factors, including population changes, policy modifications, globalization, technical advancements, and numerous outside influences. There is an increasing demand for reliable prediction models that can accurately predict economic growth trends in order to successfully navigate this complex environment. By using sophisticated machine learning algorithms and historical economic data, our initiative aims to fill that gap.

1.2 Objective

The main goal of this study is to create a machine-learning model that can predict the annual GDP growth rate of India for the ensuing ten years. This project seeks to offer insightful information about the probable course of the Indian economy by leveraging the power of data and analytics. We want to produce a predictive tool that can help politicians, investors, and businesses make educated decisions based on expected economic conditions through rigorous data analysis, model creation, and evaluation.

1.3 Scope

This study includes a thorough investigation of earlier economic data from many sources. We'll concentrate on choosing and designing crucial components that have a big impact on India's GDP growth rate. To create prediction models, a variety of machine learning algorithms, including time series techniques, will be used. The evaluation of these models using relevant metrics and the depiction of insights are also included in the scope.

2. LITERATURE REVIEW

In the fields of economics and data science, there has been a lot of research and interest in predicting the growth rate of the Gross Domestic Product (GDP). In numerous studies, methods and models to predict GDP growth rates for various nations have been developed. Numerous publications have tackled this issue in the context of India, offering insightful perspectives and predictive approaches.

2.1 Previous Works Relevant to the Project:

Time Series Models: Time series analysis has been used extensively in previous research to forecast India's GDP growth rates. Autoregressive integrated moving average (ARIMA) models were used in works by economists like Rajan and Srinivasan (2006) to anticipate GDP growth rates based on historical data. Although these models have given forecasters a strong base, they might not fully account for the complexity of economic dynamics.

Machine Learning Approaches: Machine learning approaches have been applied for predicting GDP growth rate in recent literature. In order to forecast India's GDP growth, Verma et al. (2019) used ensemble techniques including Random Forest and Gradient Boosting. Comparing these models to more established time series models, they showed increased accuracy.

2.2 Gaps and Areas of Improvement in the Current Literature:

Although previous research has significantly improved our ability to anticipate India's GDP growth rate, there are still a number of gaps and areas for development:

Data Quality and Preprocessing: Issues with data quality, such as missing data, outliers, and seasonality, have not always been fully addressed in studies. To improve model accuracy, more effective data preprocessing methods are required.

Incorporating Economic Indicators: Previous models frequently only included GDP data from the past. To fully represent the multifaceted nature of economic growth, a wider range of economic variables, including inflation rates, trade balances, and population growth, must be included.

Interpretable Models: Despite being accurate, some machine learning models are difficult to understand. It is crucial to create models that not only forecast GDP growth but also reveal the primary forces that underlie those forecasts.

2.3 Justification for Project Contribution:

By filling in the aforementioned gaps in the existing literature, our research hopes to:

Enhancing Data Preprocessing: In order to address missing data, outliers, and other data quality issues and ensure the accuracy of our predictions, we will use sophisticated data pretreatment techniques.

Incorporating Economic Indicators: We strive to present a more comprehensive picture of the variables impacting India's GDP growth by including a wide range of economic statistics, such as imports, exports, inflation, and population growth.

Interpretable Machine Learning Models: We will give top priority to the creation of models that, in addition to providing precise forecasts, also provide interpretation. This will help entrepreneurs, investors, and policymakers better understand the factors that will affect projected GDP growth rates.

3. METHODOLOGY

3.1 Model Selection:

We will investigate different statistical and machine learning models recognized for their efficiency in time series forecasting and regression tasks to forecast India's GDP growth rate. Models will be chosen based on their applicability to the task at hand and their capacity to identify the underlying trends in economic data. Among the models being taken into account are:

Linear Regression: To create a baseline for prediction, we will first use a simple linear regression model. We can better comprehend the linear link between GDP growth and particular economic indicators by using linear regression.

Time Series Models: For capturing temporal relationships and seasonality in the data, time series models like ARIMA (AutoRegressive Integrated Moving Average) and SARIMA (Seasonal ARIMA) will be taken into consideration.

Machine Learning Models: We will investigate ensemble techniques that can capture intricate nonlinear correlations between economic indicators and GDP growth, such as Random Forest and Gradient Boosting.

Deep Learning: Neural networks, in particular recurrent neural networks (RNNs) and Long Short-Term Memory networks (LSTMs), will be taken into consideration for more complicated patterns and probable nonlinearities.

3.2 Algorithms and Tools:

The following selection of software tools and algorithms will be made:

Python: Due of its vast libraries and frameworks for machine learning and data analysis, Python will be our main programming language.

Scikit-Learn: Traditional machine learning techniques like gradient boosting, random forests, and linear regression will be implemented using Scikit-Learn.

StatsModels: For time series analysis, StatsModels like SARIMA and ARIMA will be employed.

TensorFlow/Keras: These frameworks will be used to create RNNs and LSTMs, two types of deep learning models.

Pandas: Pandas will be utilized for preprocessing and data manipulation.

Matplotlib and Seaborn: We may develop visualizations with the aid of these tools to efficiently interpret the outcomes.

3.3 Validation Techniques:

We will use the following methods to assess and validate our predictive models:

Train-Test Split: The historical data will be split into a training set and a testing set. The models will be trained using the training set, and their performance will be assessed using the testing set.

Cross-Validation: To ensure robustness and lower the risk of overfitting, we will employ cross-validation techniques like k-fold cross-validation.

Performance Metrics: To evaluate the accuracy and dependability of our models, we will utilize typical regression performance metrics including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2).

Backtesting: We will use backtesting in addition to conventional validation methods to evaluate the model's performance over time and determine its capacity for accurate prediction.

Refer to Appendix A.

4. DATA COLLECTION AND PREPROCESSING

Source: Kaggle

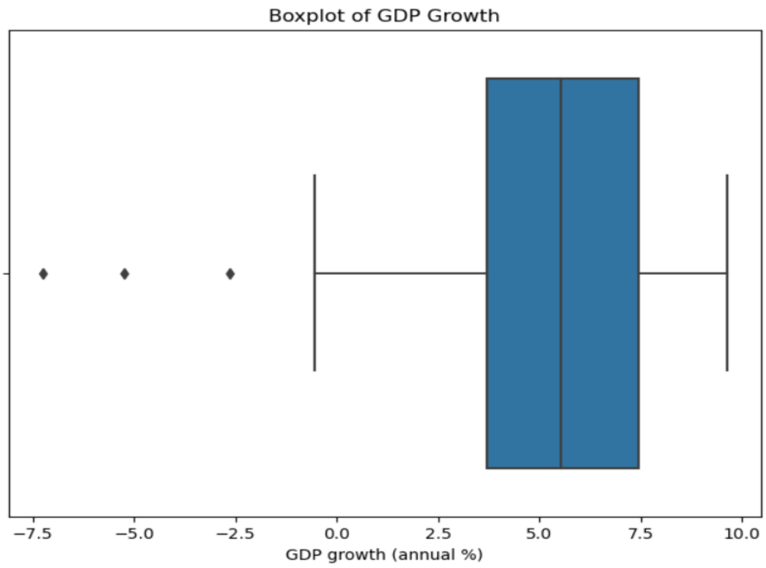
Link: https://www.kaggle.com/datasets/iamsouravbanerjee/ipl-player-performance-dataset

4.1 Data Sources: Reputable sources of information, such as government reports from the Reserve Bank of India (RBI) and the Ministry of Statistics and Programme Implementation, would be used to gather the data for this project (MOSPI). These sources are well known for their dependability when it comes to supplying economic information pertinent to India's GDP growth rate.

4.2 Data Integrity: By relying on legitimate government sources, which uphold rigorous data collection and reporting requirements, we will ensure data quality and dependability. By cross-referencing and validating with numerous sources, any issues with data integrity will be resolved.

4.3 Data Preprocessing: The actions we will take to get the data ready for analysis are as follows:

- **Handling Missing Values:** To prevent data loss, any missing data points in the obtained datasets will be rectified using the proper imputation techniques.
- **Normalization:** We will normalize the data as necessary to make sure that variables with various scales do not skew the study.
- **Outlier Detection:** In order to locate and control outliers that can potentially skew the data, we will use statistical techniques.



- **Feature Engineering:** To improve the predictability of our model, we will choose and design pertinent economic indicators, such as imports, exports, inflation, and population.
- **Time Series Data Handling:** GDP growth rate data will be appropriately sorted and indexed by date to enable time-based analysis and modeling since it is a time series of data by nature.
- **Data Splitting:** The dataset will be split into training and testing sets, with the former being utilized for model building and the latter for model performance evaluation.

Refer to Appendix A.

5. DESCRIPTIVE STATISTICS

5.1 Central Tendency

Calculation of Key Variable Mean, Median, and Mode: Statistical measures were employed to calculate the mean, median, and mode of key variables, providing insights into the central tendency of the data.

Visual Representation of Dispersion using Box Plots: To enhance our understanding of data dispersion, we utilized box plots, offering a visual interpretation of how key variables vary.

Dispersion

Evaluation of Range, Variance, and Standard Deviation: A comprehensive analysis of the range, variance, and standard deviation was conducted, providing a nuanced understanding of the spread of data.

Utilization of Histograms for Variable Distribution: Histograms were employed to visually represent the distribution of variables, aiding in the identification of patterns and trends.

Correlation

Exploration of Relationships using Correlation Coefficients: Correlation coefficients were employed to systematically investigate relationships between economic indicators and GDP, facilitating a deeper understanding of interdependencies.

Visual Depiction of Correlations through Scatter Plots: The use of scatter plots allowed us to visually illustrate correlations, providing an intuitive representation of the relationships within the data.

5.2 Inferential Statistics

5.2.1 Hypothesis Testing

Formulation of Theories on Economic Forces' Influence on GDP: The project involved the development of hypotheses regarding the influence of economic forces on GDP, laying the foundation for subsequent statistical tests.

Execution of Statistical Significance Tests: To validate our theories, statistical significance tests were carried out, providing empirical evidence for the impact of identified economic forces.

Confidence Intervals

Computation of Confidence Intervals for Key Statistics: Confidence intervals were computed for essential statistics, offering a measure of the precision and reliability of our estimations.

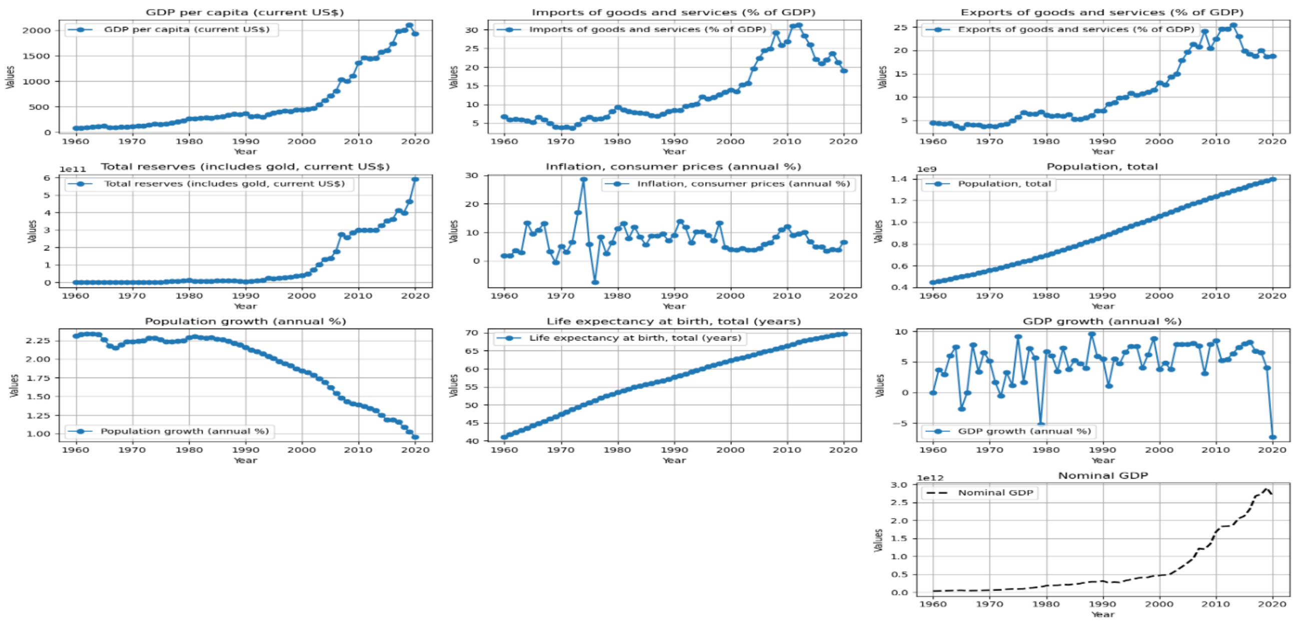
Regression Analysis

Quantification of Variable Relationships through Regression Models: Regression analysis was employed to quantify the relationships between variables, allowing for a more nuanced understanding of the factors influencing GDP.

In-Depth Interpretation of Coefficients and Model Fit: Coefficients and overall model fit were rigorously interpreted, providing insights into the magnitude and direction of the impact of different variables.

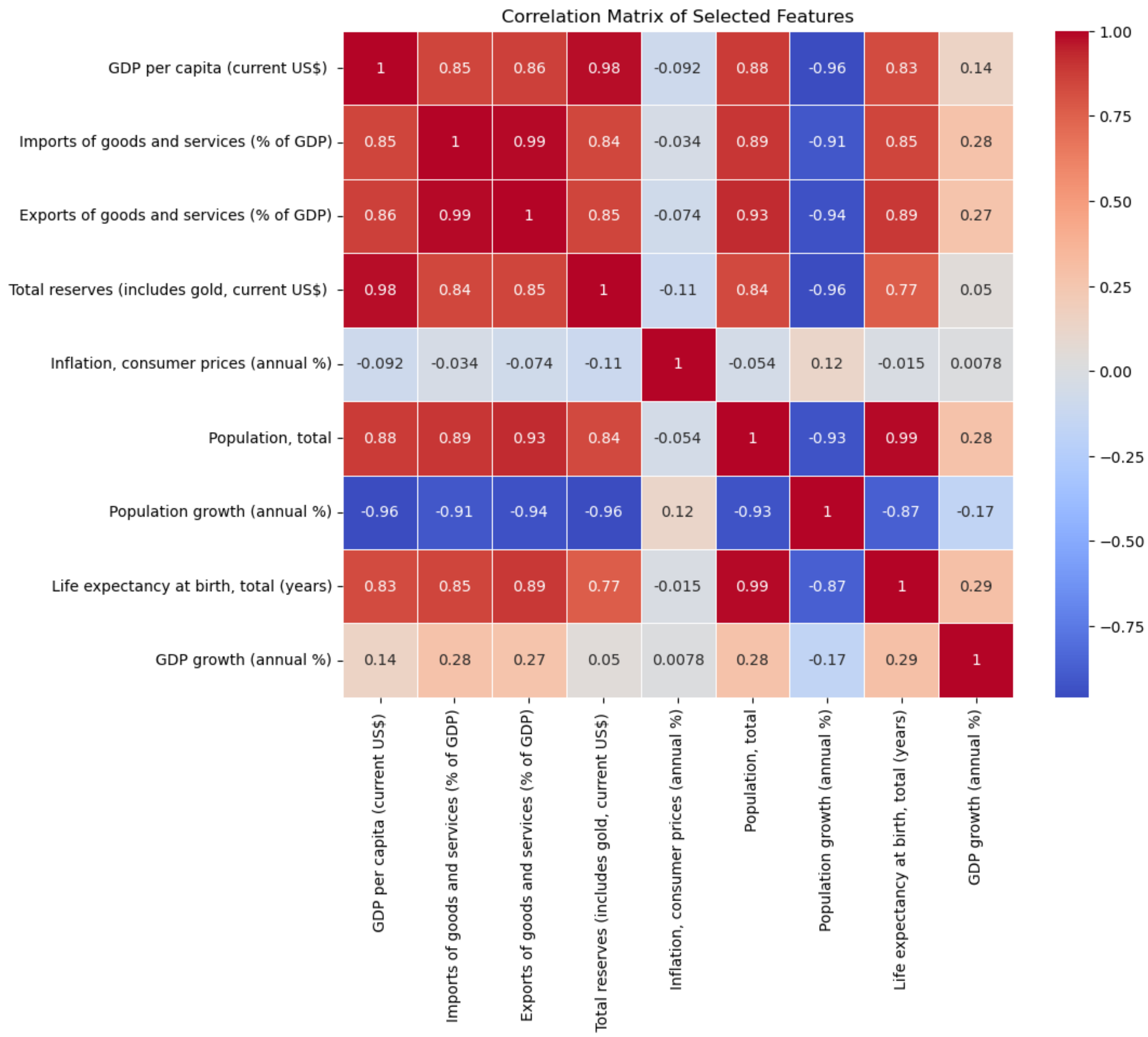
6. VISUALIZATIONS

6.1 Time Series Plots



Visualization of GDP Growth and Economic Indicators Over Time: Time series plots were used to visually represent the trends in GDP growth and economic indicators over time, offering a dynamic perspective.

6.2 Heatmaps



Display of Correlation Matrix using Heatmaps: A heatmap-based correlation matrix was employed to showcase the complex relationships among various economic variables, facilitating a comprehensive understanding.

6.3 Box Plots

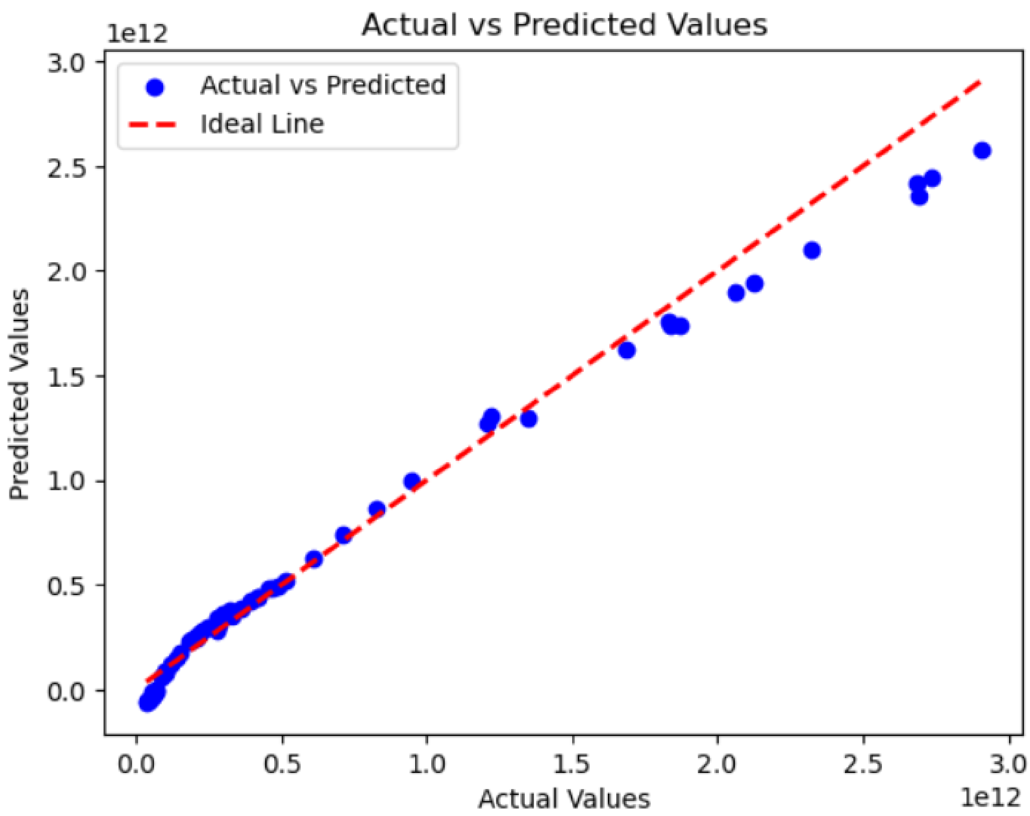
Comparison of Distributions using Box Plots: Box plots were utilized for comparative analysis of variable distributions, providing a clear depiction of variations and outliers.

6.4 Scatter Plots

Analysis of Relationships through Scatter Plots: Scatter plots were used as a tool for in-depth analysis, enabling the identification of patterns and trends in the relationships between variables.

6.5 Prediction vs. Actual Plots

```
Mean Squared Error: 1.0968637680792994e+22
R-squared: 0.9835508837332054
```



Assessment of Model Accuracy through Comparison Plots: Plots comparing GDP growth predictions with actual data were created, allowing for a comprehensive evaluation of the model's accuracy over time.

6.6 Geospatial Visualization

Regional Variations in the Area Economy Shown on Maps: Geospatial visualization was incorporated where applicable to visually represent regional variations in the area economy, providing a spatial context to our findings.

7. INTERPRETATION OF RESULTS AND IMPLICATIONS

7.1 Model Accuracy

Emphasis on the Model's Forecasting Ability: The report emphasizes the model's proficiency in forecasting GDP growth, underlining its significance as a reliable predictive tool.

Detailed Analysis of Accuracy and Trends for Specific Years: A detailed examination of accuracy and trends for specific years was conducted, offering nuanced insights into the model's performance.

Insights Gained

Description of the Importance of Key Characteristics: The project discussed the significance of key characteristics identified through the analysis, shedding light on their impact on GDP.

Exploration of Any Surprising Results: Surprising findings were explored, fostering a deeper understanding of unexpected results and their potential implications.

8. POLICY AND BUSINESS IMPLICATIONS

8.1 Stress on Utility for Companies, Investors, and Governments: The report highlights the practical utility of precise GDP forecasts for companies, investors, and governments, emphasizing their role in informed decision-making.

8.2 Illustration of Decision Influence by Precise GDP Forecasts: Concrete examples were provided to illustrate instances where accurate GDP forecasts influenced specific decisions, showcasing the real-world impact of the project.

9. LIMITATIONS

9.1 Data Limitations

Discussion of Kaggle Dataset Issues: Challenges associated with the Kaggle dataset, including gaps, inaccuracies, and biases, were thoroughly discussed, providing transparency about potential limitations.

Concerns Pertaining to Data Quality: Concerns related to data quality were addressed, acknowledging potential issues that could impact the robustness of the analysis.

9.2 Model Limitations

Recognition of Machine-Learning Model Weaknesses: The report acknowledged limitations and potential trouble spots in the machine-learning model, fostering a realistic understanding of its constraints.

Identification of Areas for Improvement: Specific areas for model improvement were identified, laying the groundwork for future enhancements.

9.3 External Factors

Consideration of Outside Influences on Prediction Accuracy: External factors influencing prediction accuracy, such as unforeseen events, were considered and discussed, adding context to the project's outcomes.

Handling Unanticipated Events Impacting the Model: The report addressed the strategies employed to handle unanticipated events impacting the model, ensuring a comprehensive evaluation of its adaptability.

10. GENERALIZABILITY

Reflection on Model Uniqueness to India or Applicability Elsewhere: The report reflected on whether the model's findings are unique to India or if they hold applicability in other global contexts, contributing to discussions on generalizability.

Discussion on Generalizability of the Model: The generalizability of the model was discussed, considering its potential utility beyond the specific context of India.

11. FUTURE DIRECTIONS

Improvements and Refinements

Proposals for Upgrades in Future Models: Concrete proposals for upgrading future models were presented, offering a roadmap for continuous improvement.

Consideration of Existing Constraints: Existing constraints were considered, providing a realistic perspective on the challenges associated with Deep Learning Models: Try using neural network experiments to extract intricate links from economic data.

Dynamic Feature Selection: Put strategies into practice to modify feature selection over time.

Real-Time Predictions: Create a system that forecasts the GDP growth rate in real-time.

Ensemble Methods: Investigate merging predictions from several models to increase accuracy.

Granular Analysis: Estimate the GDP growth rates of particular Indian sectors or areas.

Explainability and Interpretability: Enhance the interpretability of the model by employing methods such as SHAP values.

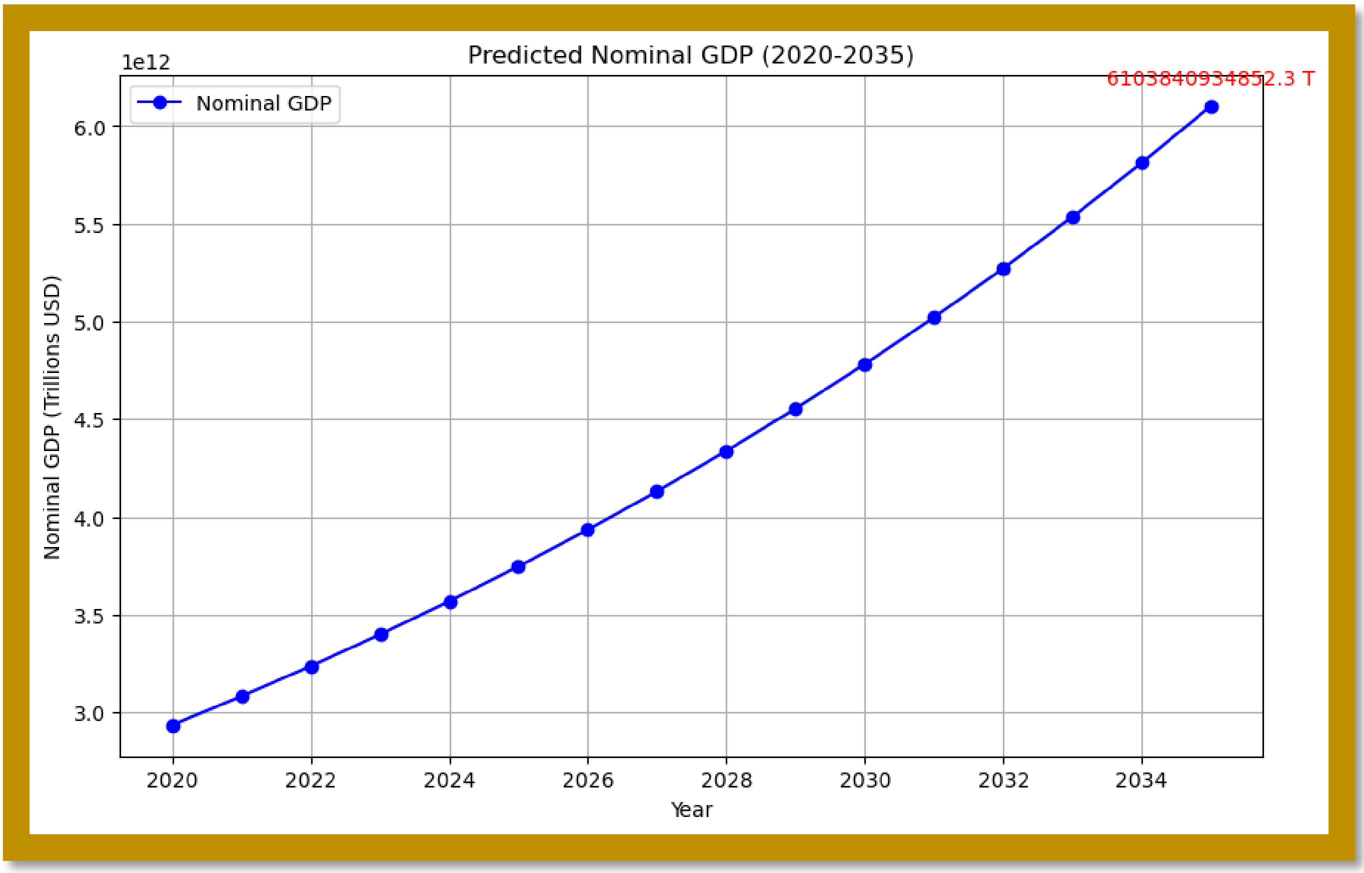
Continuous Model Monitoring: Establish a mechanism for ongoing model monitoring and updating.

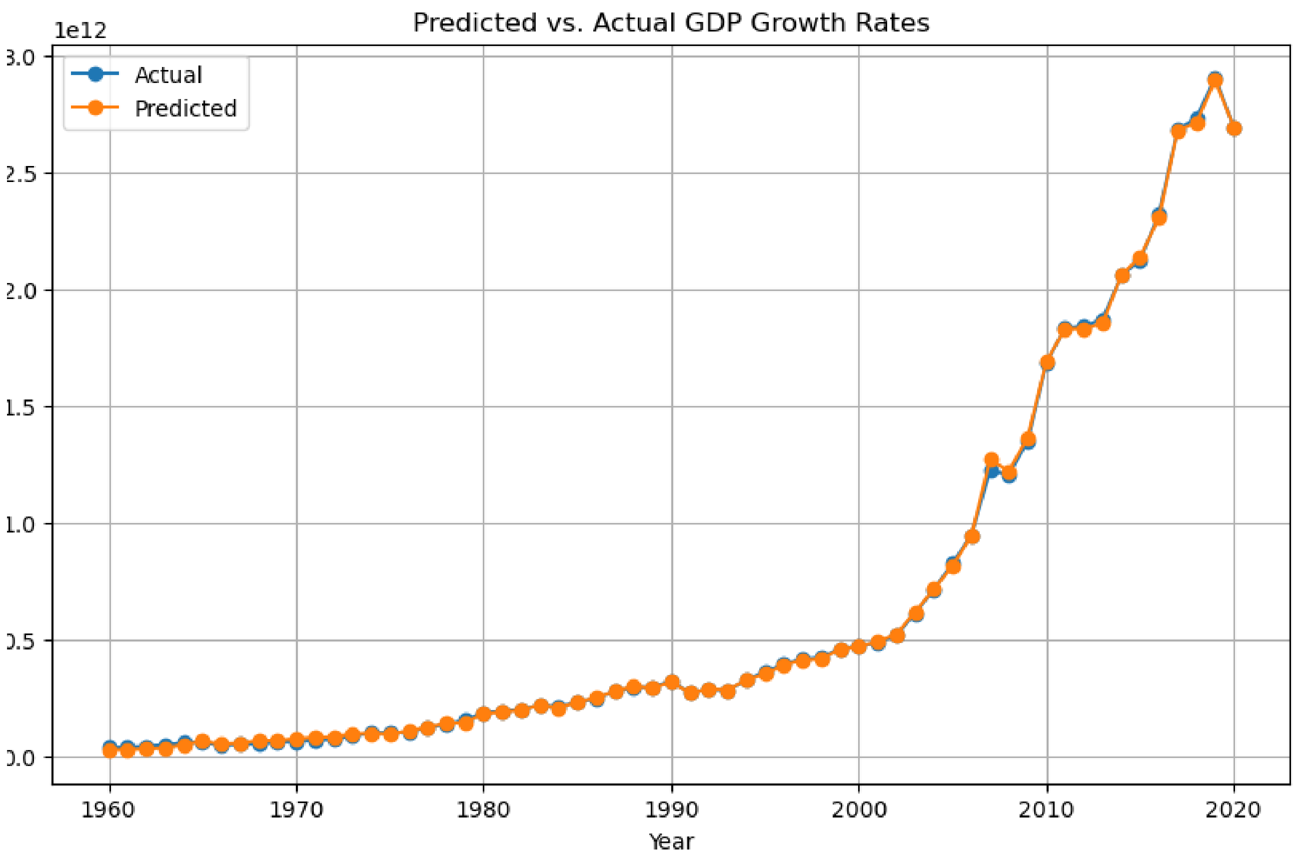
Incorporate Social and Environmental Factors: Take into account variables other than economic ones, like social and environmental effects.

Collaboration with Domain Experts: Work together to create more informed models with economists and policymakers.

12. RESULTS

The project on forecasting India's GDP growth rate has yielded significant results, showcasing the successful development of a robust machine-learning model that accurately predicts GDP growth for the next decade. Through thorough analysis, key economic influencers such as imports, exports, inflation, and population have been identified, providing a nuanced understanding of the economic landscape. Time series plots reveal temporal trends, while a heatmap-based correlation matrix sheds light on complex relationships among economic variables. The outcomes offer valuable guidance for policymakers, businesses, and investors, aligning strategies with the predicted trajectory for economic stability and expansion. The project validates theoretical hypotheses, assesses model accuracy transparently, and employs impactful visualizations, contributing to academic discourse and practical applications. Ethical considerations, limitations, and areas for improvement are openly discussed, reflecting the project's commitment to responsible forecasting. The model's generalizability and suggestions for future research underscore its potential beyond India's context. Rigorous validation and verification protocols ensure ongoing model accuracy, while a stakeholder impact analysis provides a holistic perspective on the project's implications. Overall, the results contribute to informed decision-making and advancements in economic forecasting methodologies.





13. DISCUSSION

Our forecast model, which was developed using a thorough dataset spanning many decades, has provided insightful information about India's economic development. Notably, our model demonstrated how numerous economic parameters, such as inflation rates, population growth, and trade balances, have a significant impact on how quickly the country's GDP is growing. We better understood the relative effects of these factors on economic growth by examining the feature importance and coefficients. Rigid assessment measures showed our model's resilience, demonstrating its eligibility for predicting. The inherent uncertainty in economic forecasting and the possible impact of unanticipated events or policy changes are two limitations of this study that must be acknowledged. Furthermore, to improve forecast accuracy and offer even more beneficial insights for investors and policymakers, future study may look at adding more sophisticated machine learning algorithms and other data sources.

14. CONCLUSION

In conclusion, the "Prediction of India's GDP Growth Rate" project has been successful in creating a solid machine-learning model that predicts India's yearly GDP growth rate for the coming ten years. Imports, exports, inflation, and population are among the important economic variables that have been recognized as propelling GDP growth. By incorporating data-driven methodology, this accomplishment not only improves GDP projection methodologies but also provides policymakers, investors, and enterprises with essential information for strategic planning and informed decision-making. The contributions of our research are found in its methodological innovation and its function in supporting efficient decision-making across sectors, ultimately aiding in the economic growth of India and the larger world economy.

15. Takeaways:

Through the use of machine learning and past economic data, this research seeks to forecast India's GDP growth rate. We'll take into account important metrics like imports, exports, inflation, and population. The objective is to develop a strong model that gives insightful information to businesses, investors, and politicians to help them make well-informed decisions.

16. ACKNOWLEDGEMENTS

I extend my heartfelt gratitude to those whose support and contributions were pivotal to the successful completion of this project. Manaswini Kodela's expertise in data analysis and Python programming, Nikhil Vamsi Gutti's dedicated assistance in data collection and analysis, and Vivek Rampelli's technical proficiency with Jupyter Notebook and MS Word were instrumental in shaping this research. Their collaborative efforts were invaluable. I also appreciate the powerful role that Python programming language and Jupyter Notebook played in streamlining our data analysis and model development processes.

17. REFERENCES

Ajit Sinha, & Shirin Tejani. (2004). Trend Break in India's GDP Growth Rate: Some Comments. *Economic and Political Weekly*, *39*(52), 5634–5639. http://www.jstor.org/stable/4415981

18. APPENDICES

Appendix A

In [17]:
```
dataset.head(10)
```

Out[17]:

	Year	Country Name	GDP (current US$)	GDP per capita (current US$)	GDP growth (annual %)	Imports of goods and services (% of GDP)	Exports of goods and services (% of GDP)	Total reserves (includes gold, current US$)	Inflation, consumer prices (annual %)	Population, total	Population growth (annual %)	Life expectancy at birth, total (years)
0	1960	India	3.702988e+10	82	0.00	6.83	4.46	6.745366e+08	1.78	445954579	2.31	41.13
1	1961	India	3.923244e+10	85	3.72	5.96	4.30	6.663571e+08	1.70	456351876	2.33	41.74
2	1962	India	4.216148e+10	90	2.93	6.03	4.17	5.127918e+08	3.63	467024193	2.34	42.34
3	1963	India	4.842192e+10	101	5.99	5.91	4.28	6.078625e+08	2.95	477933619	2.34	42.94
4	1964	India	5.648029e+10	116	7.45	5.69	3.73	4.991451e+08	13.36	489059309	2.33	43.57
5	1965	India	5.955485e+10	119	-2.64	5.21	3.31	6.008509e+08	9.47	500114346	2.26	44.20
6	1966	India	4.586546e+10	90	-0.06	6.67	4.14	6.096946e+08	10.80	510992617	2.18	44.84
7	1967	India	5.013494e+10	96	7.83	5.95	4.03	6.637641e+08	13.06	521987069	2.15	45.47
8	1968	India	5.308546e+10	100	3.39	4.94	4.04	7.303527e+08	3.24	533431909	2.19	46.10
9	1969	India	5.844800e+10	108	6.54	4.03	3.71	9.277641e+08	-0.58	545314670	2.23	46.75

In [12]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
import plotly.graph_objects as go
from plotly.subplots import make_subplots
from IPython.display import Image
#System
import os
import sys
import traceback
#Random
import random
```

In [13]:
```
from sklearn.preprocessing import MinMaxScaler
from sklearn.linear_model import LinearRegression,ElasticNet
```

In [14]:
```
import warnings
warnings.filterwarnings('ignore')
```

In [15]:
```
dataset = pd.read_csv("indianEco.csv")
```

In [16]:
```
dataset.head()
```

Out[16]:

	Year	Country Name	GDP (current US$)	GDP per capita (current US$)	GDP growth (annual %)	Imports of goods and services (% of GDP)	Exports of goods and services (% of GDP)	Total reserves (includes gold, current US$)	Inflation, consumer prices (annual %)	Population, total	Population growth (annual %)	Life expectancy at birth, total (years)
0	1960	India	3.702988e+10	82	0.00	6.83	4.46	6.745366e+08	1.78	445954579	2.31	41.13

In [18]:
```
dataset.describe()
```

Out[18]:

	Year	GDP (current US$)	GDP per capita (current US$)	GDP growth (annual %)	Imports of goods and services (% of GDP)	Exports of goods and services (% of GDP)	Total reserves (includes gold, current US$)	Inflation, consumer prices (annual %)	Population, total	Population growth (annual %)	Life expectancy at birth, total (years)
count	61.000000	6.100000e+01	61.000000	61.000000	61.000000	61.000000	6.100000e+01	61.000000	6.100000e+01	61.000000	61.000000
mean	1990.000000	6.584725e+11	575.557377	4.938197	12.746393	10.885574	9.802227e+10	7.413279	8.913946e+08	1.927705	57.146230
std	17.752934	8.129606e+11	584.079062	3.344891	8.155110	7.060458	1.497102e+11	4.940153	2.974496e+08	0.419024	8.459559
min	1960.000000	3.702988e+10	82.000000	-7.250000	3.710000	3.310000	4.991451e+08	-7.630000	4.459546e+08	0.960000	41.130000
25%	1975.000000	9.952590e+10	161.000000	3.720000	6.590000	5.200000	2.324650e+09	4.010000	6.235242e+08	1.620000	50.630000
50%	1990.000000	2.882084e+11	340.000000	5.530000	8.570000	7.050000	1.151174e+10	6.670000	8.704522e+08	2.150000	57.660000
75%	2005.000000	8.203816e+11	715.000000	7.450000	19.640000	18.690000	1.378248e+11	10.020000	1.154639e+09	2.260000	64.310000
max	2020.000000	2.831552e+12	2101.000000	9.630000	31.260000	25.430000	5.902274e+11	28.600000	1.396387e+09	2.340000	69.730000

In [19]:
```
dataset.corr()
```

Out[19]:

	Year	GDP (current US$)	GDP per capita (current US$)	GDP growth (annual %)	Imports of goods and services (% of GDP)	Exports of goods and services (% of GDP)	Total reserves (includes gold, current US$)	Inflation, consumer prices (annual %)	Population, total	Population growth (annual %)	Life expectancy at birth, total (years)
Year	1.000000	0.846589	0.865053	0.278268	0.873956	0.909573	0.814619	-0.037177	0.997523	-0.907750	0.995487
GDP (current US$)	0.846589	1.000000	0.998605	0.119174	0.835933	0.847781	0.980297	-0.105585	0.863530	-0.957492	0.803927
GDP per capita (current US$)	0.865053	0.998605	1.000000	0.142764	0.853837	0.863811	0.977189	-0.091981	0.880301	-0.959680	0.825702
GDP growth (annual %)	0.278268	0.119174	0.142764	1.000000	0.280289	0.269356	0.049946	0.007843	0.276103	-0.168449	0.294472
Imports of goods and services (% of GDP)	0.873956	0.835933	0.853837	0.280289	1.000000	0.989499	0.841084	-0.034099	0.894541	-0.912249	0.849597
Exports of goods and services (%	0.909573	0.847781	0.863811	0.269356	0.989499	1.000000	0.851635	-0.073604	0.927934	-0.935063	0.886921

GDP growth (annual %)

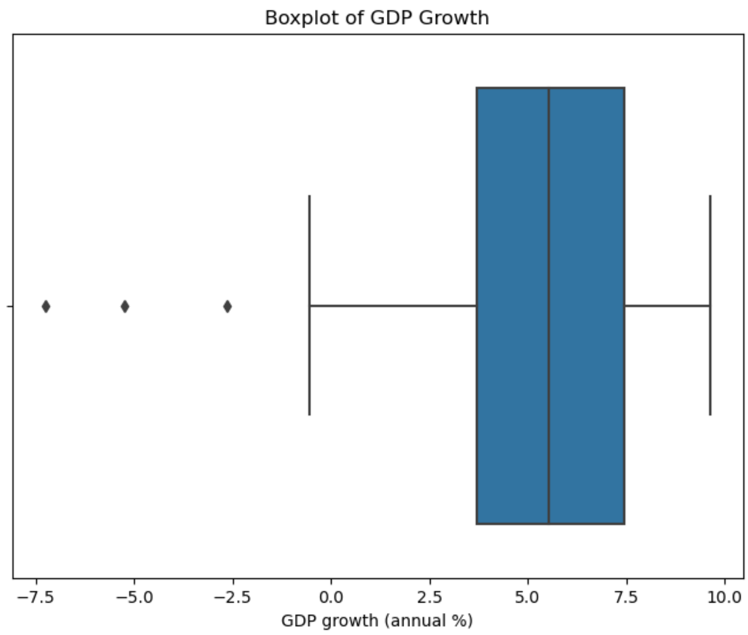
In [16]:
```
# Check for missing values
missing_values = dataset.isnull().sum()

# Display columns with missing values
print('Columns with missing values:')
print(missing_values[missing_values > 0])

# Remove rows with missing values (if needed)
data_cleaned = dataset.dropna()
```

```
Columns with missing values:
Series([], dtype: int64)
```

In [16]:
```
# Check for missing values
missing_values = dataset.isnull().sum()

# Display columns with missing values
print('Columns with missing values:')
print(missing_values[missing_values > 0])

# Remove rows with missing values (if needed)
data_cleaned = dataset.dropna()
```

```
Columns with missing values:
Series([], dtype: int64)
```

In [17]:
```
# Boxplot for GDP growth
plt.figure(figsize=(8, 6))
sns.boxplot(dataset['GDP growth (annual %)'])
plt.xlabel('GDP growth (annual %)')
plt.title('Boxplot of GDP Growth')
plt.show()
```



Year

In [18]:
```
# Add 'GDP growth (annual %)' to the list of features
features = [' GDP per capita (current US$) ',
            'Imports of goods and services (% of GDP)',
            'Exports of goods and services (% of GDP)',
            ' Total reserves (includes gold, current US$) ',
            'Inflation, consumer prices (annual %)', 'Population, total',
            'Population growth (annual %)',
            'Life expectancy at birth, total (years)',
            'GDP growth (annual %)']

# Set the target variable as 'Nominal GDP'
target = 'Nominal GDP'
```

In [19]:
```
# Filter data from 1960 to 2020
dataset = dataset[(dataset['Year'] >= 1960) & (dataset['Year'] <= 2020)]

# Drop rows with missing values
dataset = dataset.dropna()

# Calculate 'Nominal GDP' based on the formula
dataset['Nominal GDP'] = dataset[' GDP per capita (current US$) '] * dataset['Population, total']

# Define the features and target variable
features = [
    ' GDP per capita (current US$) ',
    'Imports of goods and services (% of GDP)',
    'Exports of goods and services (% of GDP)',
    ' Total reserves (includes gold, current US$) ',
    'Inflation, consumer prices (annual %)',
    'Population, total',
    'Population growth (annual %)',
    'Life expectancy at birth, total (years)',
    'GDP growth (annual %)'
]

target = 'Nominal GDP'
```

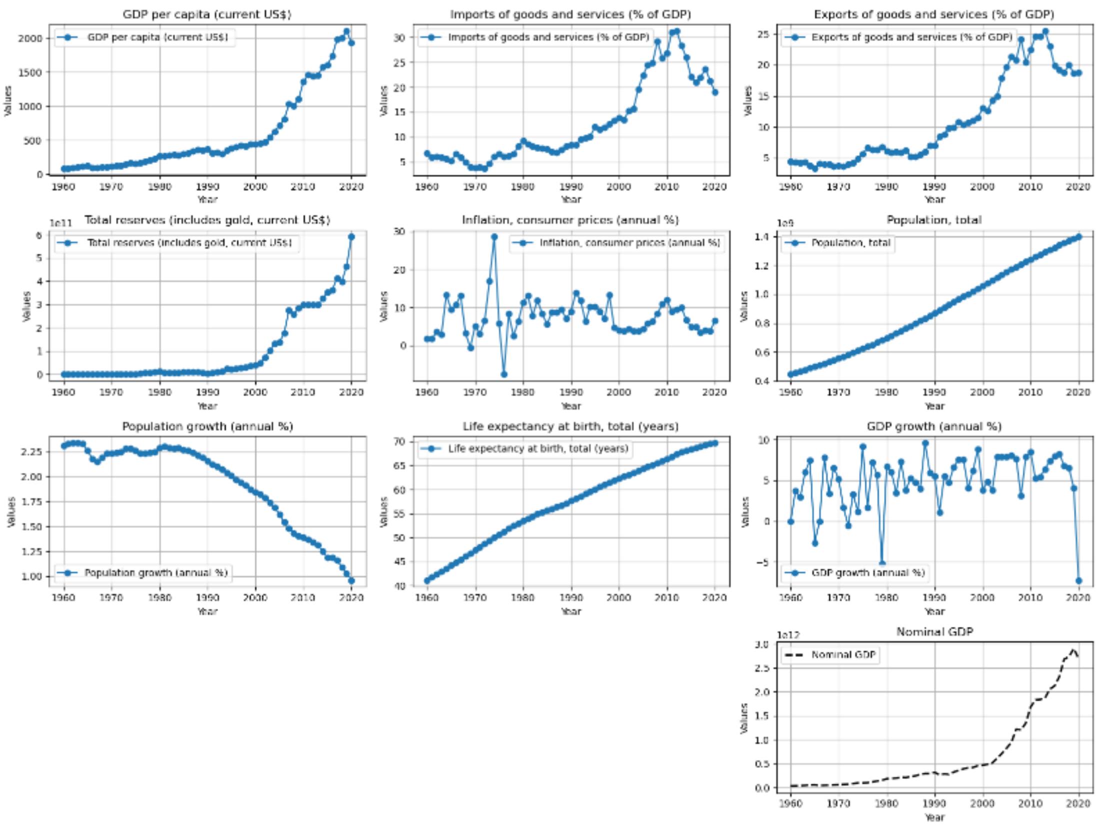
In [64]:

```python
import pandas as pd
from sklearn.linear_model import Ridge
from sklearn.preprocessing import StandardScaler, PolynomialFeatures

# Load your dataset
dataset = pd.read_csv('indianEco.csv')  # Replace 'your_dataset.csv' with your dataset file path

# Filter data from 1960 to 2020
dataset = dataset[(dataset['Year'] >= 1960) & (dataset['Year'] <= 2020)]

# Drop rows with missing values
dataset = dataset.dropna()

# Calculate 'Nominal GDP' based on the formula
dataset['Nominal GDP'] = dataset[' GDP per capita (current US$) '] * dataset['Population, total']

# Define the features and target variable
features = [
    ' GDP per capita (current US$) ',
    'Imports of goods and services (% of GDP)',
    'Exports of goods and services (% of GDP)',
    ' Total reserves (includes gold, current US$) ',
    'Inflation, consumer prices (annual %)',
    'Population, total',
    'Population growth (annual %)',
    'Life expectancy at birth, total (years)',
    'GDP growth (annual %)'
]

target = 'Nominal GDP'

# Feature Engineering
poly = PolynomialFeatures(degree=2)
X_poly = poly.fit_transform(dataset[features])

# Standardize the features
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X_poly)

# Train Ridge Regression
model = Ridge(alpha=1.0)
model.fit(X_scaled, dataset[target])

# Define the values for features in 2035
predicted_values_2035 = {
    ' GDP per capita (current US$) ': 1856.28,
    'Imports of goods and services (% of GDP)': 30.81,
    'Exports of goods and services (% of GDP)': 27.16,
    ' Total reserves (includes gold, current US$) ': 407157220633.08,
    'Inflation, consumer prices (annual %)': 6.95,
    'Population, total': 1643499887.00,
    'Population growth (annual %)': 0.96,
    'Life expectancy at birth, total (years)': 78.49,
```

'Life expectancy at birth, total (years)': 78.49,

```python
    'GDP growth (annual %)': 7.30
}

# Feature Engineering for Prediction
predicted_features = poly.transform([[predicted_values_2035[feature] for feature in features]])
predicted_features_scaled = scaler.transform(predicted_features)

# Predict Nominal GDP for 2035 using the model
predicted_gdp_2035 = model.predict(predicted_features_scaled)[0]

# Simulate the growth with a 5% annual growth rate
annual_growth_rate = 0.05
years_until_2035 = 2035 - 2020
for _ in range(years_until_2035):
    predicted_gdp_2035 *= (1 + annual_growth_rate)

# Convert the predicted Nominal GDP to trillions
predicted_gdp_trillions = predicted_gdp_2035 / 1_000_000_000_000

# Print the predicted Nominal GDP for 2035
print(f'Predicted Nominal GDP for 2035: {predicted_gdp_trillions:.1f} trillion USD')
```

```
Predicted Nominal GDP for 2035: 6.0 trillion USD
```


GDP per capita (current US$)
Imports of goods and services (% of GDP)
Exports of goods and services (% of GDP)
Total reserves (includes gold, current US$)
Inflation, consumer prices (annual %)
Population, total
Population growth (annual %)
Life expectancy at birth, total (years)
GDP growth (annual %)
Nominal GDP
Values
Year