DS302.3
Advanced Statistics for Data Science

Forecasting Specific Countries' Populations Using ARIMA Model

Group - B

Group Members

Student Name	Student Number
M K I M Rohana	24572
M R K Karunathilaka	24490
O T N Rajapaksha	24882

Module Lecturer: Ms. Kavishka Rajapaksha

Table of Contents

1. Problem Background

- The world's population is undergoing dynamic changes with implications for various aspects of society, including social, economic, and environmental factors.
- Understanding these changes is crucial for informed decision-making and policy formulation.
- However, a comprehensive and in-depth time-series analysis is needed to extract meaningful insights and patterns.

Selected countries and reasons:

Sri Lanka:

- Crucial for cultural and geopolitical insights, Sri Lanka's analysis explores demographic shifts and their impacts.

India:

- As the world's most populous country, India's time-series data is pivotal for identifying growth patterns and addressing demographic challenges.

China:

- With a focus on one of the most populous nations, China's analysis unravels the impacts of policies, social changes, and economic developments.

Ukraine:

- Amid geopolitical challenges, Ukraine's population time-series data reveals the effects of conflict on demographic shifts.

Russia:

- As the largest country, Russia's analysis explores the intricate relationship between size, geography, and population dynamics.

2. Problem Statement

- Do not have enough accuracy forecasting models for the countries that are selected.

- There is a need for a detailed time-series analysis of the countries mentioned above' population data spanning from **1960 to 2022**. This analysis should consider key variables such as gender distribution, migration trends, and territories of origin to uncover patterns, trends, and correlations over time.

3. Project Objectives

Main Objective:

- Develop a sophisticated specific country's population forecasting model that incorporates key demographic factors, regional variations, and dynamic socioeconomic indicators and validate the model using historical data.

Other objectives:

- Exploratory data analysis and descriptive analysis used Python to perform descriptive analysis.
- Provided a user-friendly interface for accessing and visualizing population projections to enable policymakers and researchers to make informed decisions.
- Used the BI (Business Intelligence) Tool to extract valuable insights for End Users according to their requirements.

4. Theoretical Background

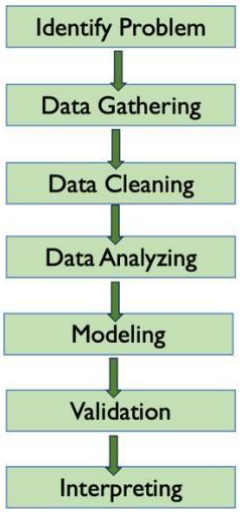
4.1 Time Series Analysis

Time series analysis forms a cornerstone for forecasting models, particularly in understanding patterns and trends within temporal data. Models like Autoregressive Integrated Moving Average (ARIMA) and Seasonal ARIMA have effectively captured the seasonality, trends, and irregularities in time-ordered population data. Incorporating time series analysis into the forecasting methodology will enable the model to account for historical patterns and project them into the future.

4.2 Machine Learning Techniques

Advancements in machine learning have provided new avenues for population forecasting. Algorithms such as Long Short-Term Memory (LSTM) networks and Random Forests have shown promise in capturing non-linear relationships and intricate patterns in demographic data. By leveraging machine learning techniques, the forecasting model will be able to adapt to the complexities of global population dynamics and provide more accurate predictions.

5. Methodologies



5.1 Data Collection Methods

To ensure the reliability and comprehensiveness of our world population forecasting model, a multi-faceted data collection approach was employed. The primary data source will be reputable global demographic databases, such as those of the World Bank, the United Nations, and other authoritative agencies specializing in population statistics. These sources provided historical data, regional breakdowns, and projections that are crucial for training and validating our forecasting model.

Additionally, auxiliary data from social, economic, and environmental indicators was collected to enhance the predictive power of the model. This includes variables such as GDP, education indices, healthcare metrics, and climate data. The integration of these diverse datasets will provide a holistic understanding of the factors influencing population dynamics.

Dataset link: https://nsbm365-my.sharepoint.com/:x:/g/personal/mrkkarunathilaka_students_nsbm_ac_lk/EZyZzOxlQTZNqIgagjrlq_4Bu_amfQWBdrXW8pEJ__GrWQ?e=NNCJLP

5.2 Tools and Technologies

5.2.1 Python

Python was the primary programming language for data pre-processing, analysis, and model development. Python's extensive libraries, including Pandas, NumPy, Matplotlib, and Seaborn, were utilized for data manipulation, exploratory data analysis (EDA), and visualization. Scikit-learn and TensorFlow were employed for machine learning model development, which allowed us to implement robust forecasting algorithms.

5.2.2 Tableau

Tableau will serve as a powerful tool for data visualization and interactive dashboards. This platform was instrumental in convening complex population trends, regional variations, and model outputs in a user-friendly and accessible manner. The visualizations created in Tableau facilitate effective communication of insights to stakeholders and decision-makers.

5.3 Data Pre-processing and Exploratory Data Analysis (EDA)

Before feeding the data into the forecasting model, thorough pre-processing will be conducted. This includes handling missing values, normalizing data, and addressing outliers. Time series decomposition techniques will extract trend, seasonality, and residual components, enhancing the model's ability to capture underlying patterns.

Exploratory Data Analysis (EDA) – (Use this Python links**)**

Sri Lanka - Sri lanka.ipynb

China - China.ipynb

Russia - Russia.ipynb

Ukraine - Ukraine.ipynb

All Countries - All Countries.ipynb

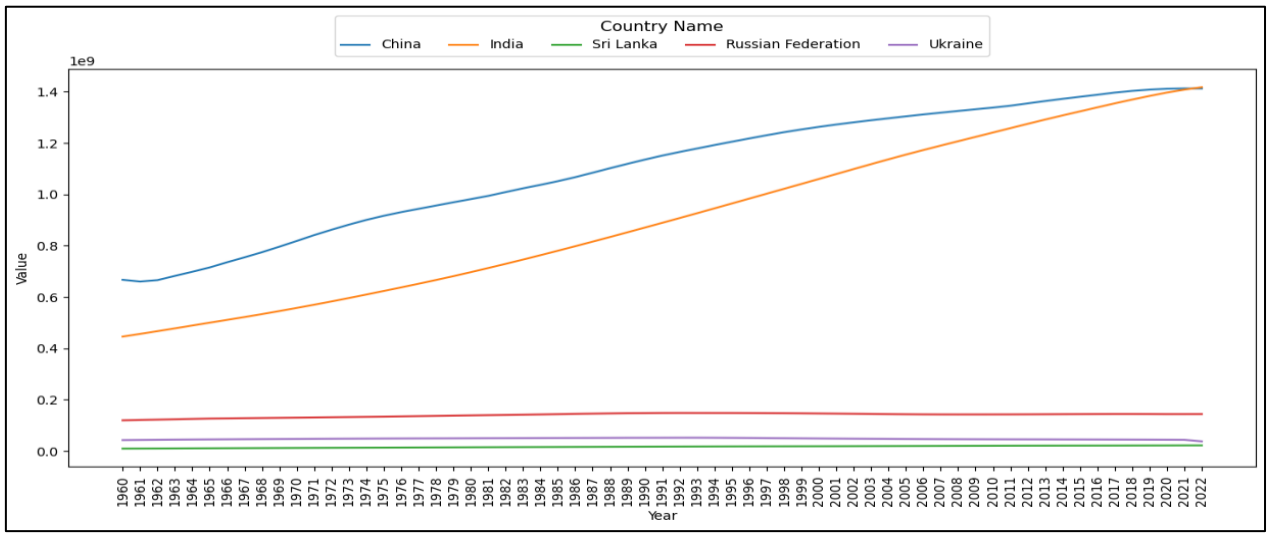
(Interpreting visualizations include in the **6.1 Data Visualising**)

6. Developing steps of the model

6.1 Data Visualising

- The main aim of the data visualization is to identify whether it has a seasonal component or not.
- If it has a seasonal component, we should use the SARIMA (Seasonal ARIMA) model.
- If not, we should use the ARIMA model.

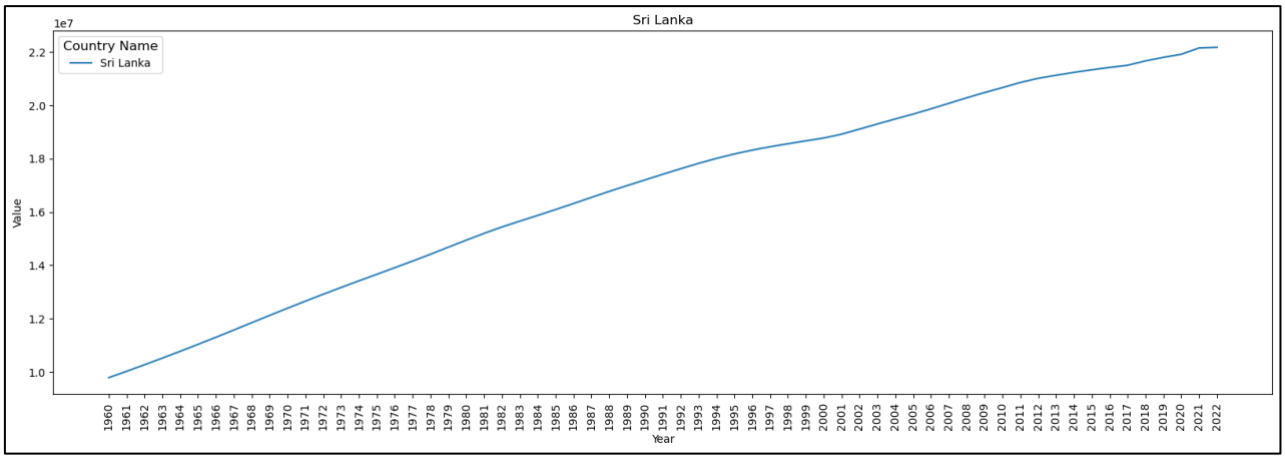
❖ The population of all selected countries from 1960 to 2022.



The time series shows the total population of several countries over time, from 1960 to 2022. Here are some observations from the chart:

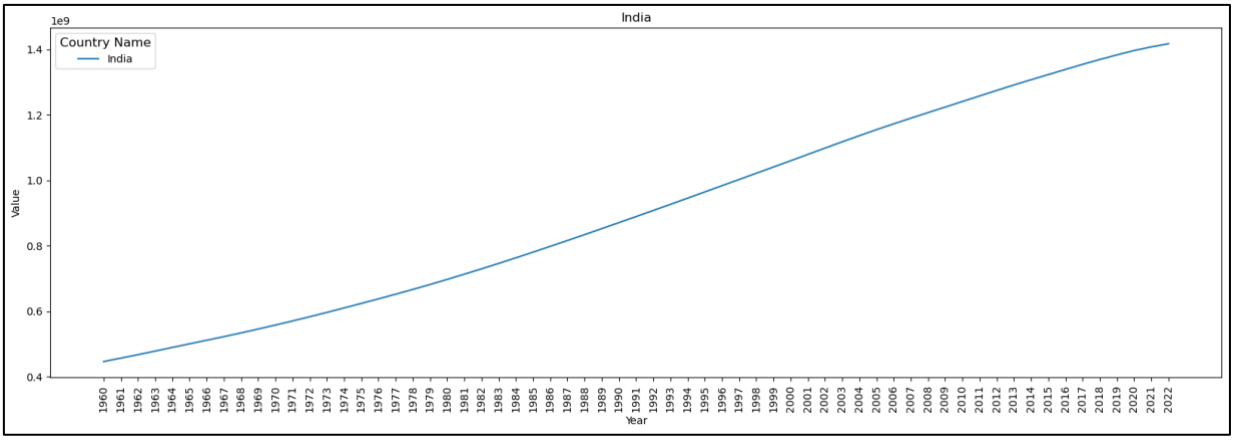
- The y-axis shows the value of the population, in billions.
- The x-axis shows the year.
- The lines for each country show a general upward trend over time, which means the population has increased.
- China has the highest population overall, followed by India.
- But when comes to 2022 India was the most populated country.

❖ The population of Sri Lanka from 1960 to 2022.



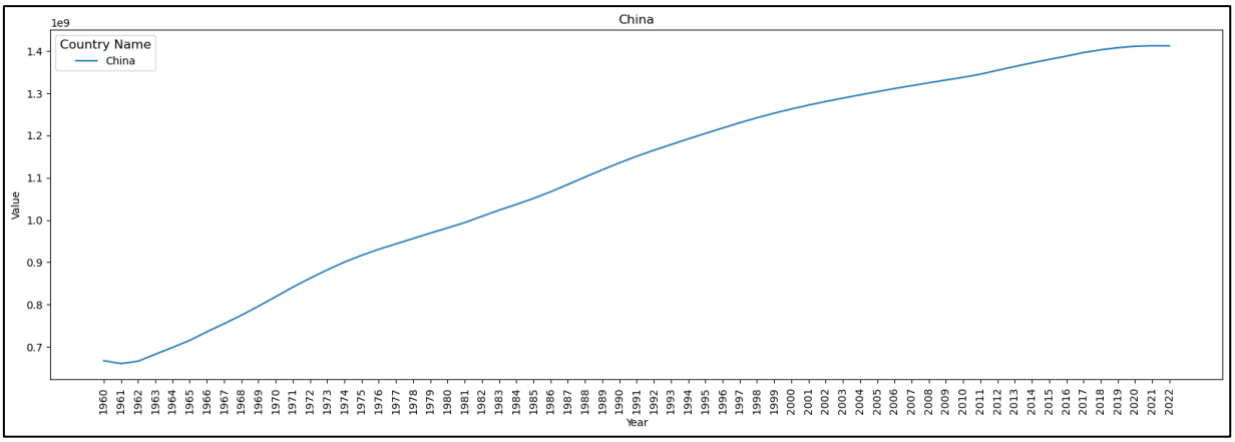
- The y-axis shows the population in millions.
- The x-axis shows the year.
- The blue line shows a steady increase in Sri Lanka's population from 1960 to 2022.

❖ The population of India from 1960 to 2022.



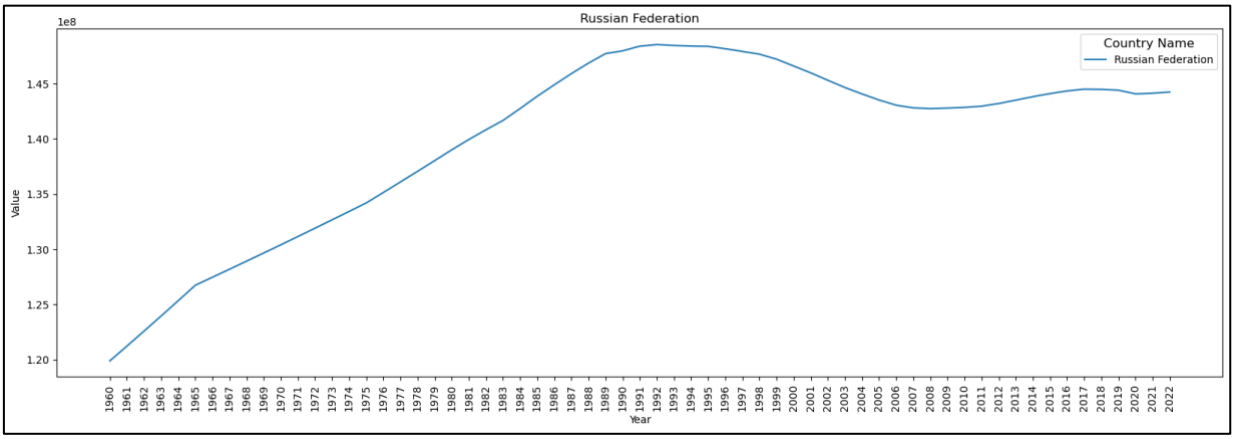
- The y-axis shows the value of the population, in billions.
- The x-axis shows the year.
- The line shows an overall increasing trend in India's population from 1960 to 2022.

❖ The population of China from 1960 to 2022.



- The y-axis shows the value of the population, in billions.
- The x-axis shows the year.
- The blue line shows an overall increasing trend in China's population from 1960 to 2022.

❖ The population of Russia from 1960 to 2022.

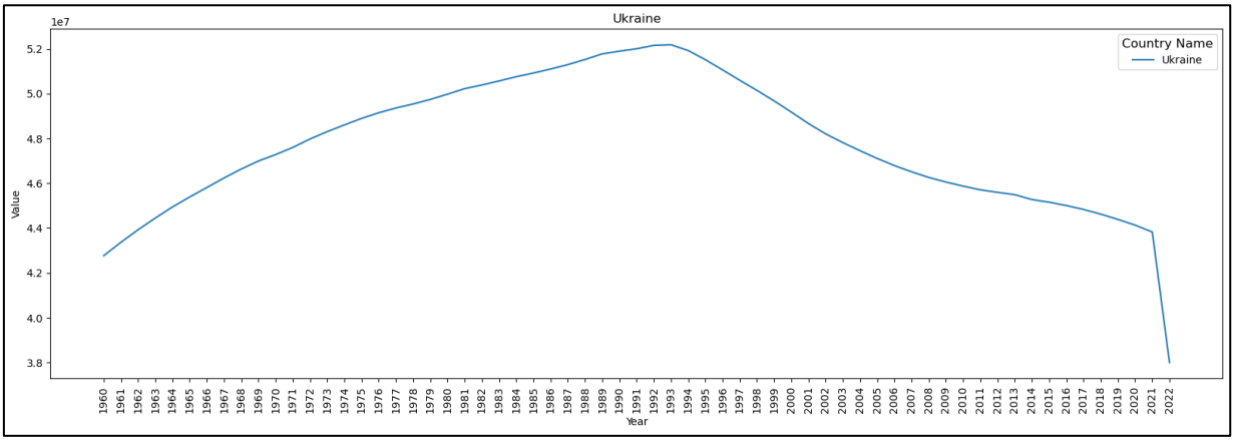


- The y-axis shows the value of the population.
- The x-axis shows the year.
- The blue line shows an overall increase in Russia's population from 1960 to 1991, followed by a decrease from 1992 to 2022.

- It is important to note that this chart doesn't tell us why Russia's population might have declined after 1991. These are some possible reasons for this,

    - Emigration: People moving out of Russia
    - Lower birth rates: Russia might have had fewer births than deaths during this time period.

❖ The population of Ukraine from 1960 to 2022.



- The y-axis shows the value, in millions.
- The x-axis shows the year.
- The blue line shows a general increase in Ukraine's population from 1960 to 1993, followed by a decrease from 1994 to 2022.
- From 2021 to 2022 there is a huge decrease in the population line.
- It is important to note that this chart doesn't tell us why Ukraine's population might have declined. These are some possible reasons for this,

    - Emigration: People moving out of Russia
    - A state of war prevailed.

- According to the above time series charts, we can say there are no seasonal components.
- Hence, we can use the ARIMA model for that.
- Also, the ARIMA model is built with main three components.
- They are,

    - ✓ AR (Auto Regression) → p (Lagged Observations)
    - ✓ I (Integration) → d (Differencing)
    - ✓ MA (Moving Average) → q (Lagged Residuals)

6.2 Checking whether the data set is Stationary or Nonstationary.

- Stationary simple mean if its mean and variance do not change over time.
- For that, we can use different tests such as ADF, KPSS, etc.
- In ADF,
    - Null Hypothesis (H0): The data set is non-stationary.
    - Alternative Hypothesis (H1): The data set is stationary.
    - Also, here If the test statistic is less than the critical value, we reject the null hypothesis and conclude that the series is stationary.
- In KPSS,
    - Null Hypothesis (H0): The data set is stationary.
    - Alternative Hypothesis (H1): The data set is not stationary.
    - Also, here If the statistic is greater than the critical value, the null hypothesis is rejected, suggesting that the series may indeed be non-stationary.

- In our project, we had to take two differencing to dataset and make stationery. That means finding the **d value** of the ARIMA model.
- Therefore, the **d value** of our model is **2.**

6.3 Finding the p-value of the model (No of lagged observations).

- For that use PACF (Partial Auto Correlation Function).
- Also, we used the auto_arima function.
- In our project, we found these components.

    - **p-value** as **0** for the **Sri Lanka** country dataset.
    - **p-value** as **0** for the **China** country dataset.
    - **p-value** as **0** for the **Russia** country dataset.
    - **p-value** as **0** for the **Ukraine** country dataset.

6.4 Finding the q value of the model (No of lagged residuals).

- For that use ACF (Auto Correlation Function).
- Also, we used the auto_arima function.
- In our project, we found these components.

    - **q-value** as **0** for the **Sri Lanka** country dataset.
    - **q-value** as **2** for the **China** country dataset.
    - **q-value** as **0** for the **Russia** country dataset.
    - **q-value** as **0** for the **Ukraine** country dataset.

6.5 Building the ARIMA model.

Sri Lanka

SARIMAX Results

Dep. Variable:	Population	No. Observations:	53
Model:	ARIMA(0, 2, 0)	Log Likelihood	-555.143
Date:	Sat, 06 Apr 2024	AIC	1112.286
Time:	12:32:38	BIC	1114.218
Sample:	01-01-1960 - 01-01-2012	HQIC	1113.024
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
sigma2	1.604e+08	1.87e+07	8.569	0.000	1.24e+08	1.97e+08

Ljung-Box (L1) (Q):	13.29	Jarque-Bera (JB):	43.64
Prob(Q):	0.00	Prob(JB):	0.00
Heteroskedasticity (H):	15.47	Skew:	1.08
Prob(H) (two-sided):	0.00	Kurtosis:	6.98

China

SARIMAX Results

Dep. Variable:	Population	No. Observations:	50
Model:	ARIMA(0, 2, 2)	Log Likelihood	-740.949
Date:	Sat, 06 Apr 2024	AIC	1487.899
Time:	10:38:55	BIC	1493.513
Sample:	01-01-1960 - 01-01-2009	HQIC	1490.020
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.0361	0.024	-1.512	0.130	-0.083	0.011
ma.L2	-0.0170	0.015	-1.119	0.263	-0.047	0.013
sigma2	6.502e+11	2.37e-15	2.74e+26	0.000	6.5e+11	6.5e+11

Ljung-Box (L1) (Q):	10.35	Jarque-Bera (JB):	10.76
Prob(Q):	0.00	Prob(JB):	0.00
Heteroskedasticity (H):	0.13	Skew:	0.86
Prob(H) (two-sided):	0.00	Kurtosis:	4.55

Russia

SARIMAX Results

Dep. Variable:	Population	No. Observations:	50
Model:	ARIMA(0, 2, 0)	Log Likelihood	-645.163
Date:	Sat, 06 Apr 2024	AIC	1292.325
Time:	12:06:04	BIC	1294.196
Sample:	01-01-1960 - 01-01-2009	HQIC	1293.032
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
sigma2	2.655e+10	2.42e+09	10.956	0.000	2.18e+10	3.13e+10

Ljung-Box (L1) (Q):	0.03	Jarque-Bera (JB):	113.32
Prob(Q):	0.87	Prob(JB):	0.00
Heteroskedasticity (H):	0.44	Skew:	-2.12
Prob(H) (two-sided):	0.11	Kurtosis:	9.22

Ukraine

SARIMAX Results

Dep. Variable:	Population	No. Observations:	50
Model:	ARIMA(0, 2, 0)	Log Likelihood	-599.080
Date:	Sat, 27 Apr 2024	AIC	1200.161
Time:	10:40:53	BIC	1202.032
Sample:	01-01-1960 - 01-01-2009	HQIC	1200.868
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
sigma2	3.892e+09	3.58e+08	10.882	0.000	3.19e+09	4.59e+09

Ljung-Box (L1) (Q):	10.81	Jarque-Bera (JB):	104.45
Prob(Q):	0.00	Prob(JB):	0.00
Heteroskedasticity (H):	5.06	Skew:	-2.01
Prob(H) (two-sided):	0.00	Kurtosis:	9.00

6.6 Model Validation

- For the model validation process, we decided to use three methods.
- They are,
    - MSE (Mean Squared Error)
    - RMSE (Root Mean Squared Error)
    - MAPE (Mean Absolute Percentage Error)

6.7 Model outputs.

- Sri Lanka
    - MSE: 72906271831.5
    - RMSE: 270011.61
    - MAPE: 1.1367%

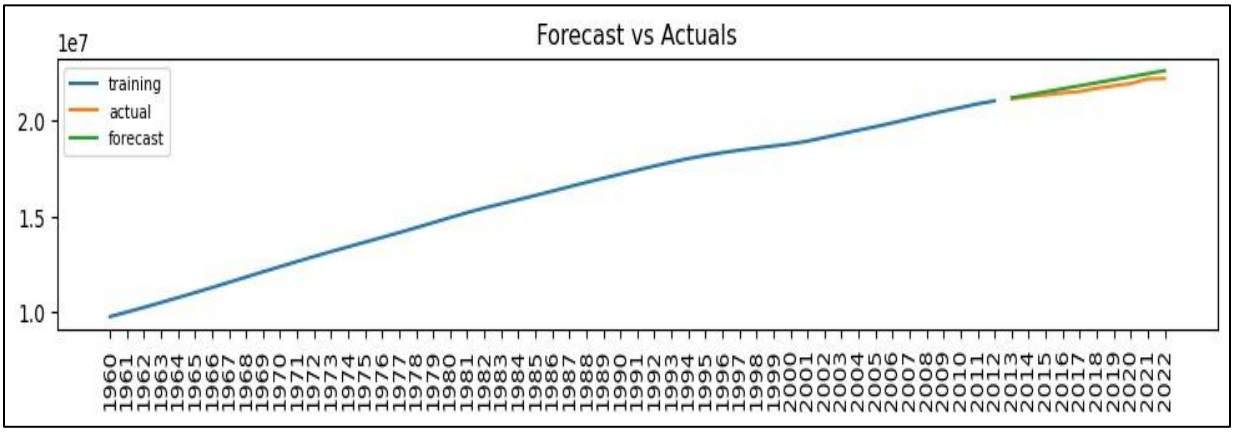
- China
    - MSE: 57958583727564.04
    - RMSE: 7613053.50871
    - MAPE: 0.4678%

- Russia
    - MSE: 838109002491.4615
    - RMSE: 915482.9340
    - MAPE: 0.56016%

- Ukraine
    - MSE: 2317743297106.117
    - RMSE: 1522413.641920
    - MAPE: 1.48895%

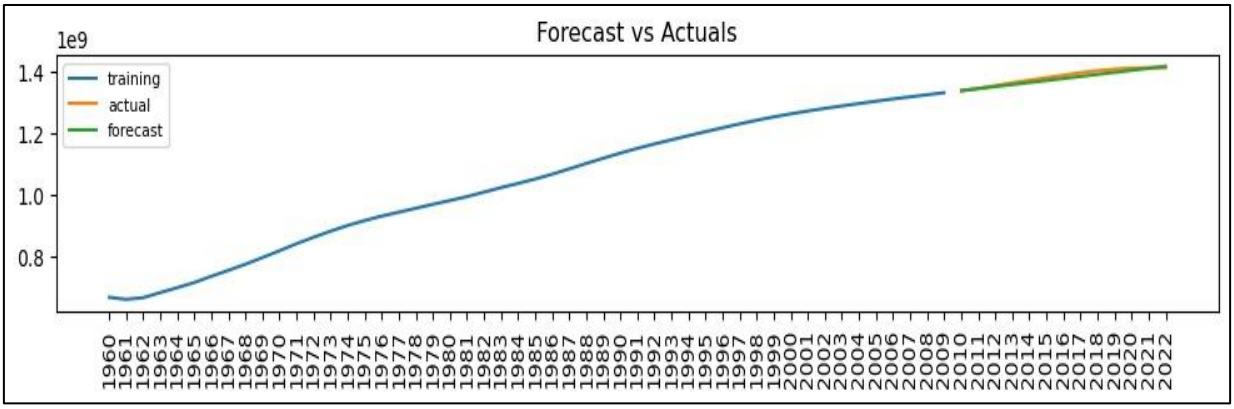
6.8 Predictions

- Next six years' prediction of the Sri Lanka Population.



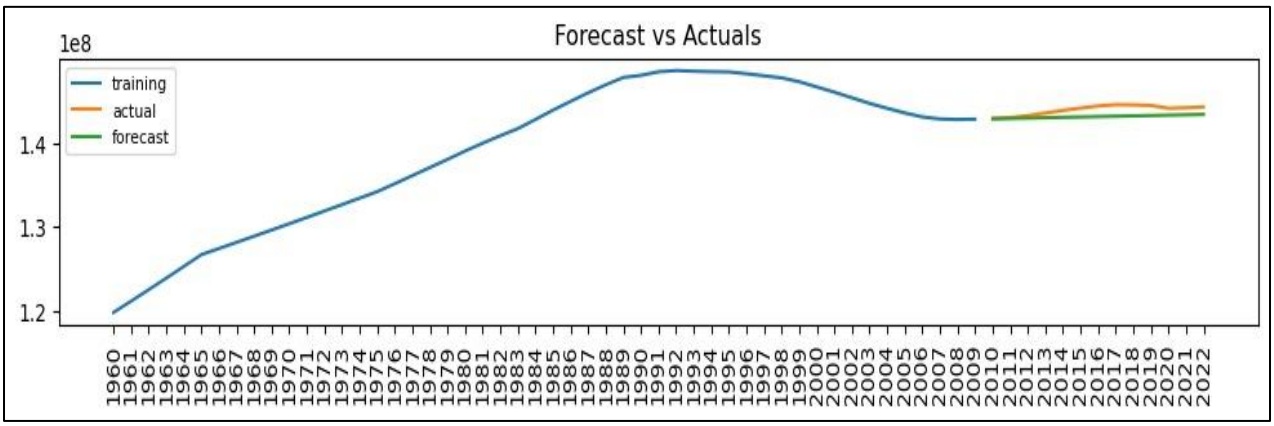
  - 2025 --- 23063399
  - 2026 --- 23220803
  - 2027 --- 23378207
  - 2028 --- 23535611
  - 2029 --- 23693015
  - 2030 --- 23850419

- Next six years' prediction of the China Population.



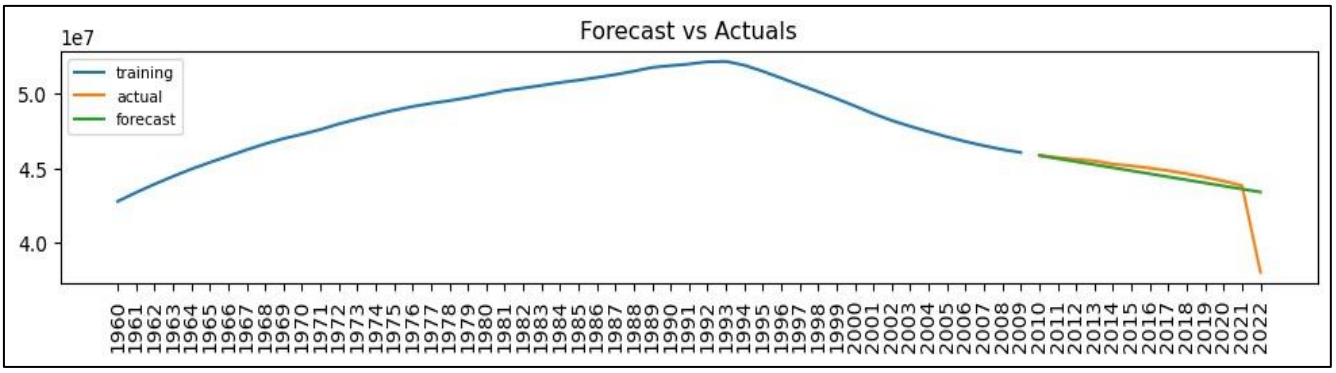
  - 2025 --- 1.437119e+09
  - 2026 --- 1.443736e+09
  - 2027 --- 1.450352e+09
  - 2028 --- 1.456969e+09
  - 2029 --- 1.463585e+09
  - 2030 --- 1.470201e+09

- Next six years' prediction of the Russia Population.



  - 2025 --- 143473077.0
  - 2026 --- 143516060.0
  - 2027 --- 143559043.0
  - 2028 --- 143602026.0
  - 2029 --- 143645009.0
  - 2030 --- 143687992.0

- Next six years' prediction of the Ukraine Population.

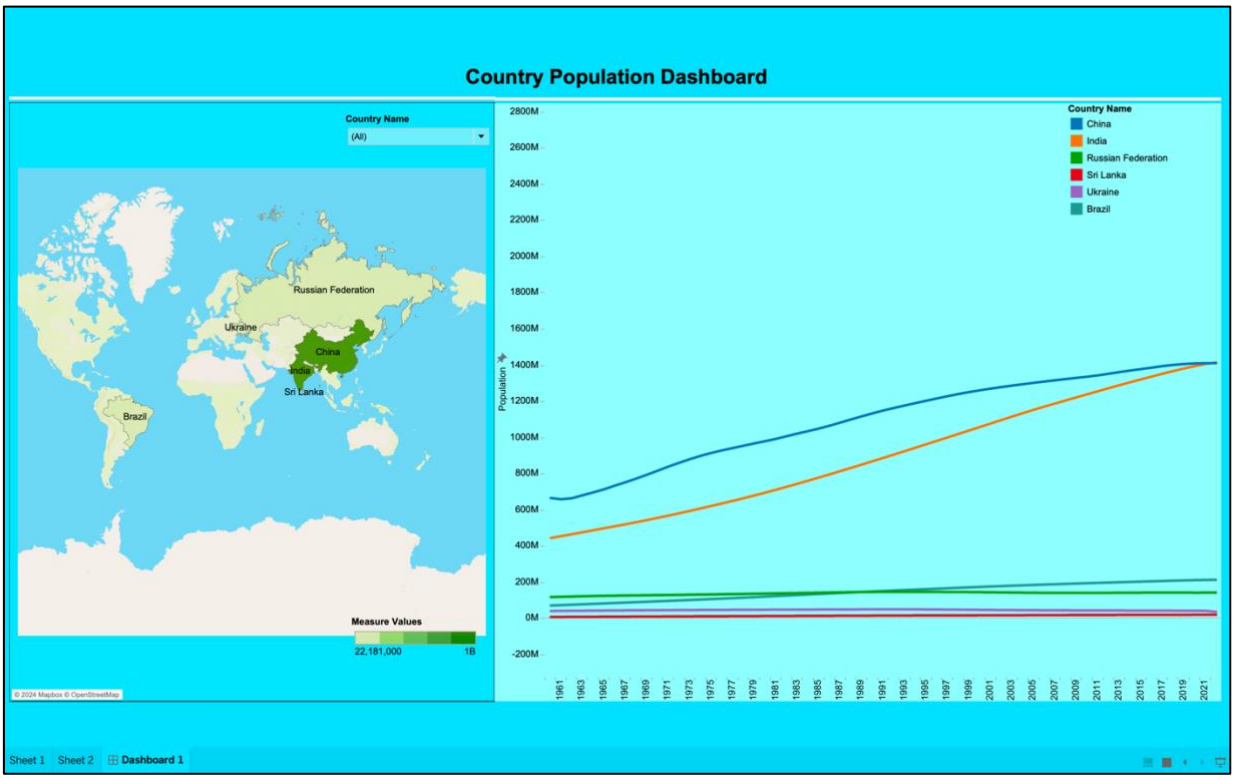


  - 2025 --- 42775603.0
  - 2026 --- 42570745.0
  - 2027 --- 42365887.0
  - 2028 --- 42161029.0
  - 2029 --- 41956171.0
  - 2030 --- 41751313.0

6.9 Dashboard

Population Dashboard:

https://public.tableau.com/views/word_Population/Dashboard1?:language=en-GB&publish=yes&:sid=&:display_count=n&:origin=viz_share_link



7. Conclusion

The report presents a comprehensive analysis and forecasting model for the population dynamics of selected countries: Sri Lanka, India, China, Ukraine, and Russia. The problem statement underscores the need for accurate forecasting models to address demographic challenges, with a focus on understanding historical trends and future projections. The project objectives aim to develop a sophisticated forecasting model incorporating key demographic factors and socioeconomic indicators.

Theoretical background sections highlight the importance of time series analysis and machine learning techniques in population forecasting. Methodologies encompass data collection from reputable sources, data pre-processing, and machine learning model building utilizing Python and Tableau. The model development process involves data visualization, identifying stationary components, determining model parameters (p, d, q), building ARIMA models, and validating model outputs.

Model validation employs metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). Predictions for the next six years are provided for each country, offering insights into future population trends.

The report outlines a rigorous methodology for population forecasting, leveraging advanced analytical techniques to provide valuable insights for policymakers and researchers. The developed forecasting model offers a robust framework for understanding and addressing demographic challenges in the selected countries.

8. References

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994). Time Series Analysis: Forecasting and Control. Prentice Hall. Retrieved from https://www.sciencedirect.com/science/article/pii/S0307904X03000799

Brownlee, J. (2018). Introduction to Time Series Forecasting with Python. Machine Learning Mastery. Retrieved from https://machinelearningmastery.com/start-here/#algorithms

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: principles and practice (2nd ed.). OTexts. Retrieved from https://otexts.com/fpp3/

McKinney, W. (2017). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. O'Reilly Media.

World Bank. (n.d.). World Development Indicators. Retrieved from https://data.worldbank.org/indicator/SP.POP.TOTL

9. Project Timeline

Tasks	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10
Problem Statement, Data Collection, Proposal Development, Review & Approval										
Data Preprocessing & EDA Part										
Model Development & Validation										
Creating a dashboard for visualization										
Final Report										

10. Group Members' Contribution

Student Name	Student Number	Contribution
M K I M Rohana	24572	Time series model development
M R K Karunathilaka	24490	Exploratory Data Analysis (EDA) and Report
O T N Rajapaksha	24882	Dashboard Creation