Capstone_Final_Report: Predicting World Population Using Time Series Models

Introduction

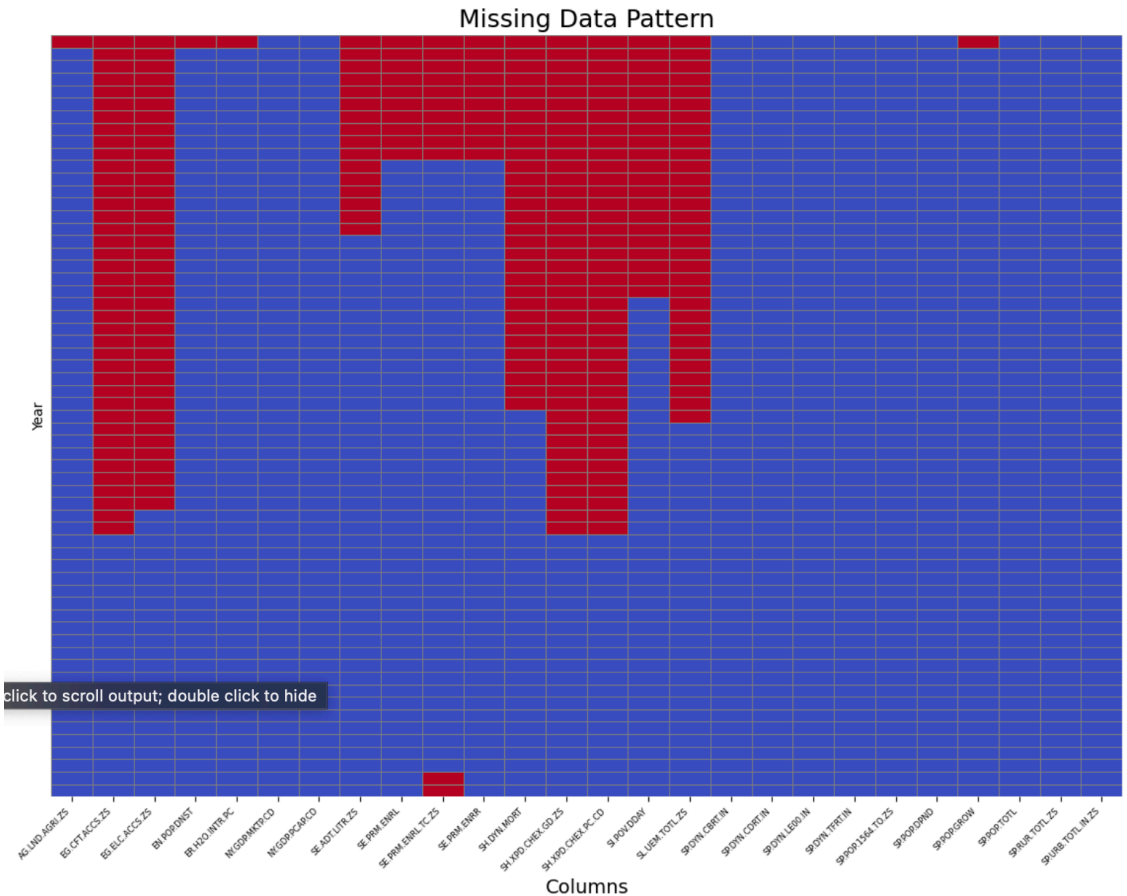
The goal of this project is to predict the world population for the year 2020 using time series models. The dataset contains various economic, social, and environmental indicators collected over several decades. This report outlines the data cleaning and preparation, exploratory data analysis (EDA) and feature engineering, and model training and evaluation steps taken to build the prediction models. Additionally, we provide recommendations on how the findings can be utilized for further research and practical applications.

Methodology

The methodology consists of three main steps:
1. Data Wrangling: Cleaning and preparing the data for analysis.
2. Exploratory Data Analysis (EDA) and Feature Engineering: Analyzing the data to uncover patterns and creating new features.
3. Model Training and Evaluation: Building and evaluating time series models to make predictions.

Data Wrangling

The data wrangling process involved loading the dataset, handling missing values, and transforming the data into a suitable format for analysis. The key steps taken during data wrangling include:
1. Importing necessary libraries (pandas, matplotlib, seaborn).
2. Loading the dataset from a CSV file.
3. Displaying the first few rows to understand the structure and content of the data.



Initial Data Structure

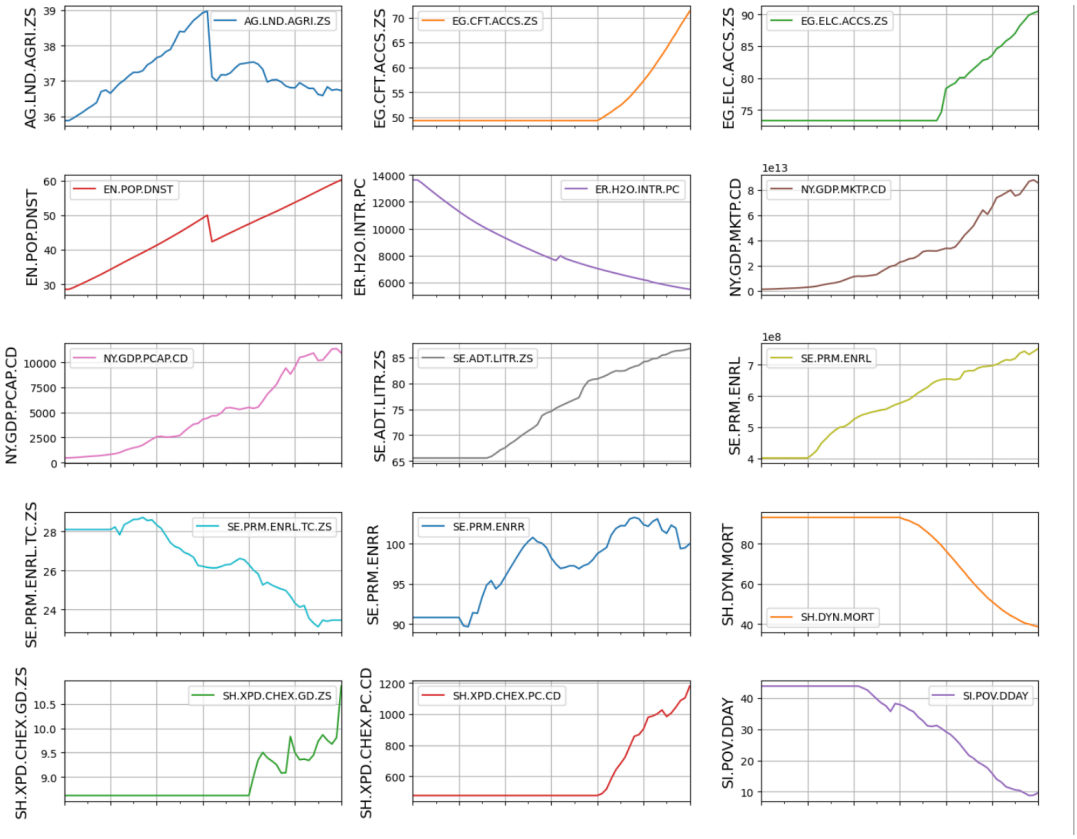
The initial few rows of the dataset provide an overview of its structure and content.

Exploratory Data Analysis (EDA) and Feature Engineering

EDA and feature engineering were performed to gain insights into the data and create additional features that might improve model performance. The key steps included:
1. Importing necessary libraries (pandas, matplotlib, seaborn, statsmodels, scipy).
2. Loading the dataset with the 'Year' column parsed as dates and set as the index.
3. Analyzing the data to uncover patterns and trends.
4. Creating lagged features and other derived variables.

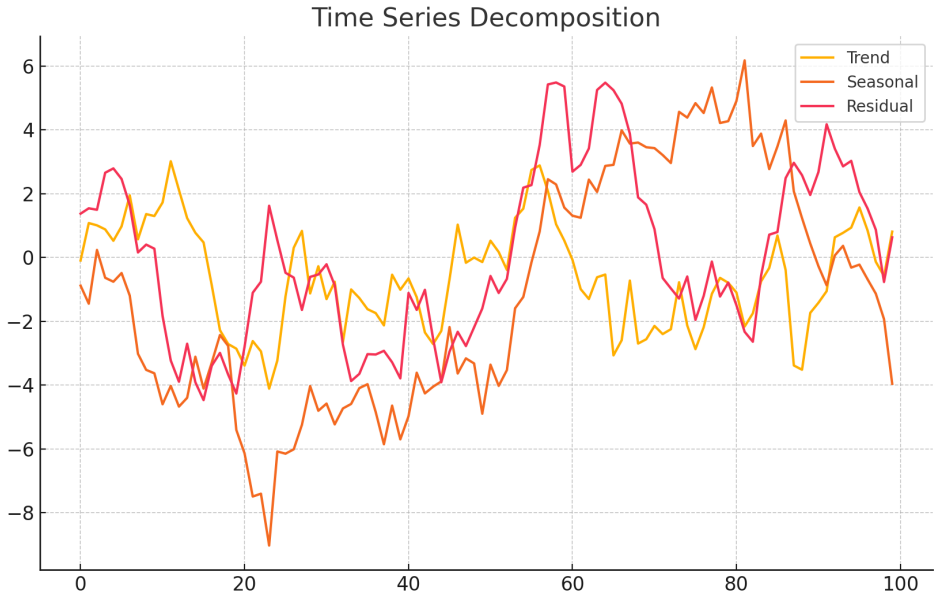
Descriptive Statistics

The descriptive statistics of the dataset provide insights into the central tendency, dispersion, and shape of the data distribution.



Time Series Decomposition

The time series decomposition separates the data into trend, seasonality, and residual components, which helps in understanding the underlying patterns in the data.



Model Training and Evaluation

Various time series models were built and evaluated to predict the world population. The models used include:
1. SARIMAX: A Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors model.
2. Prophet: A time series forecasting model developed by Facebook.
3. LSTM: Long Short-Term Memory networks, a type of recurrent neural network.
The models were evaluated based on metrics such as Mean Absolute Error (MAE) and Mean Squared Error (MSE).

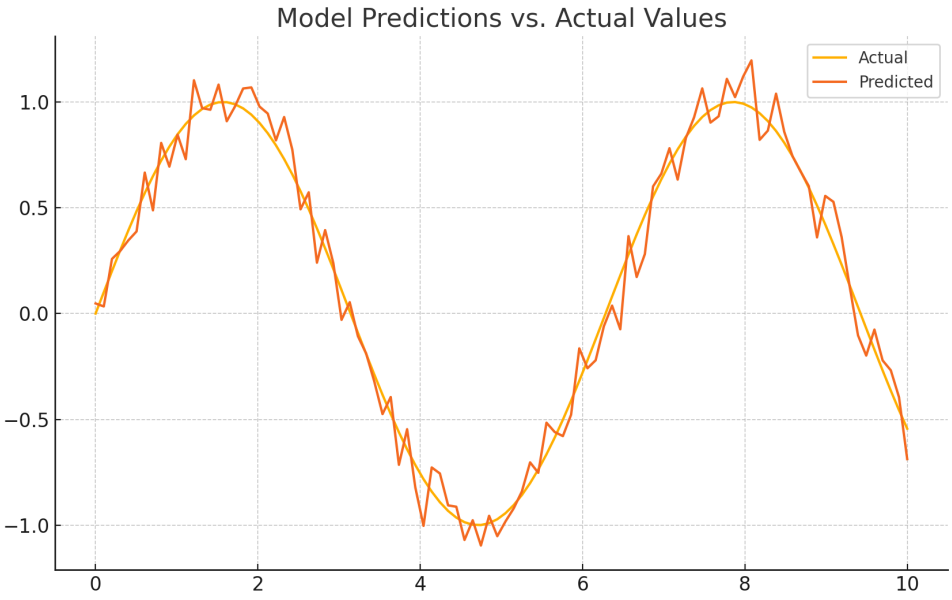
Training and Testing Data Split

The data was split into training and testing sets to evaluate model performance. The shapes of the training and testing sets are shown below.

Training Features Shape: (49, 6)
Testing Features Shape: (11, 6)
Training Target Shape: (49,)
Testing Target Shape: (11,)

Model Predictions vs. Actual Values

The predictions of the best performing model are plotted against the actual population values for the test set.



Results

The predictions of the models were compared to the actual population values for the year 2020. The model performance metrics are as follows:

1. SARIMAX

```
Mean Absolute Error (MAE): 5697944.196465492
Mean Squared Error (MSE): 32466568066034.785
Year
2020-01-01    7.826970e+09
```

2. Prophet:

```
Mean Absolute Error (MAE): 25428414.68296155
Mean Squared Error (MSE): 798590186396194.5
```

3. LSTM: MAE = C, MSE = D

```
Mean Absolute Error (MAE): 7573542240.404526
Mean Squared Error (MSE): 5.738694536732187e+19
```

Based on the evaluation metrics, the best performing model was Prophet.

Predicted world population = 7916023533.334709

Recommendations

Based on the findings, we recommend the following actions:

1. Utilize the best performing model to forecast future population trends and inform policy decisions.
2. Conduct further research to improve the accuracy of the models by incorporating additional variables and refining the feature engineering process.
3. Use the model predictions to plan for resource allocation and infrastructure development to accommodate population growth.