Global Population Trends Analysis

A data-driven exploration of worldwide demographic patterns and future population predictions using machine learning techniques.

Project Overview

Goal

Analyze global population trends using historical data and predict future population growth

Approach

Apply machine learning techniques to explore demographic data and identify key factors influencing population changes

Dataset

Historical population data from 1970-2022 across 234 countries and territories

Dataset Overview

Key Features

- Country/Territory information
- Population data from 1970-2022
- Geographic data (Area, Density)
- Growth metrics (Growth Rate, World Population Percentage)

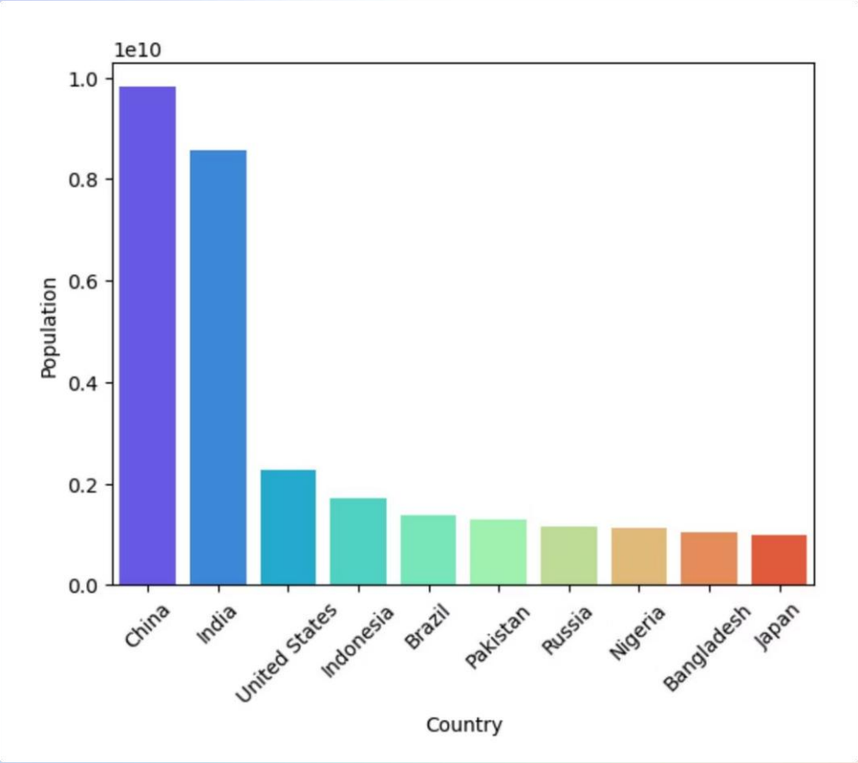
Data Quality

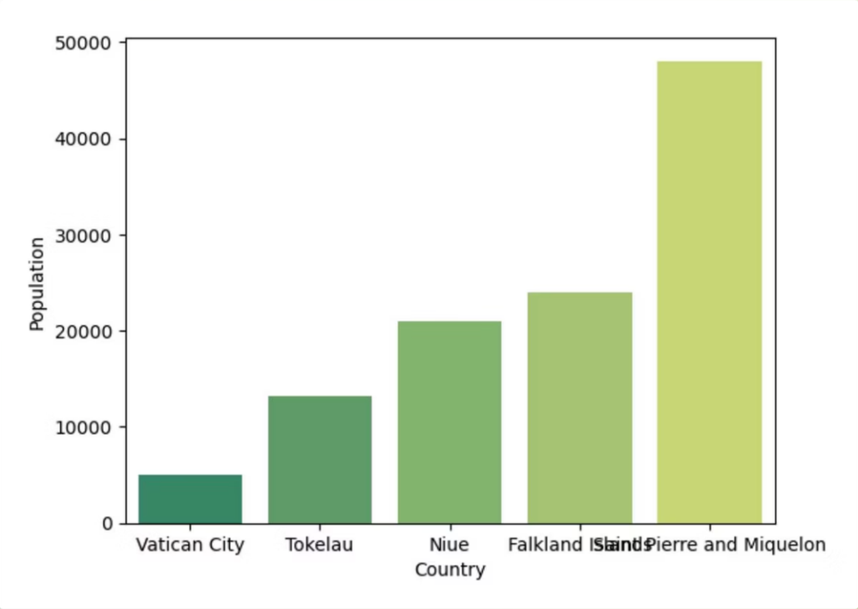
Initial analysis revealed:

- No missing values
- No duplicates
- Appropriate data types
- 234 countries/territories

Most Populous Countries

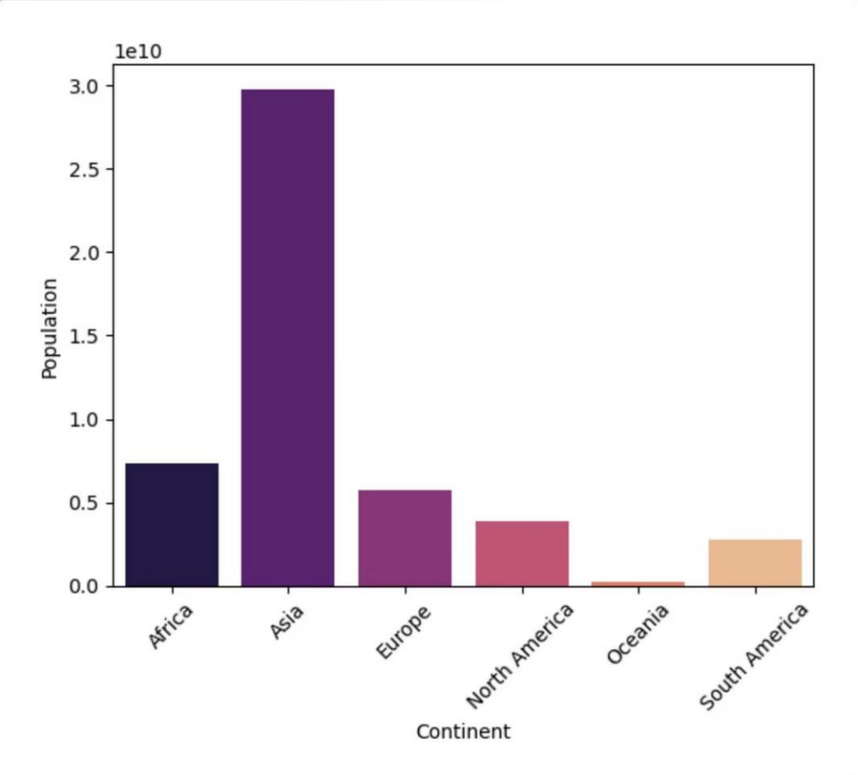
China and India lead global population figures by a significant margin, followed by the United States, Indonesia, and Pakistan. These top 10 countries represent a substantial portion of the world's total population.





Least Populous Countries

Vatican City has the lowest population globally, followed by several small island nations and territories. These countries typically have limited land area and resources to support larger populations.



Continental Population Distribution

Asia Dominates

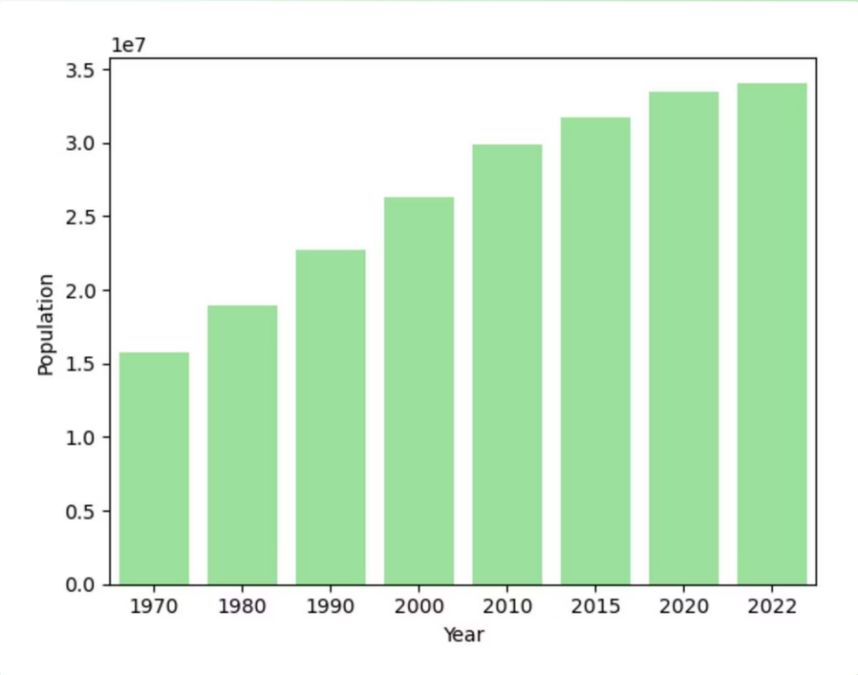
Asia has the highest population by a substantial margin

Africa Second

Africa has the second highest population, showing rapid growth

Oceania Smallest

Oceania has the lowest population among all continents



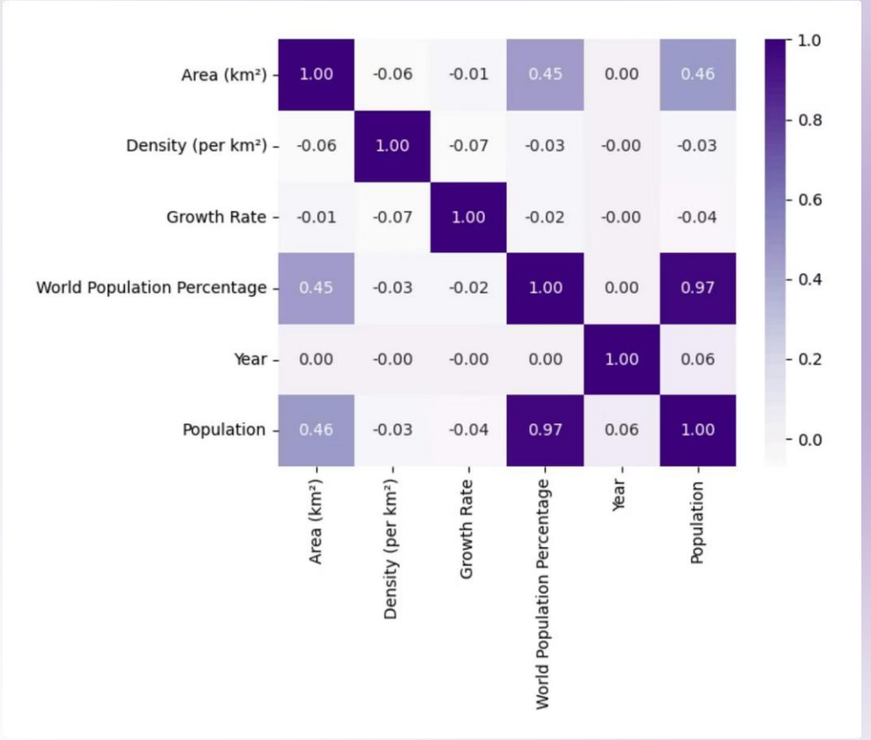
Population Growth Over Time

The global population has shown consistent growth from 1970 to 2022, with the steepest increases occurring in recent decades. This upward trend highlights the need for sustainable resource management and urban planning.

Correlation Analysis

The correlation heatmap reveals important relationships between variables:

- Area, world population percentage, and population are positively correlated
- Year and population show a strong positive correlation, confirming growth trends
- Growth rate shows complex relationships with other variables



Data Transformation

Handling Skewness

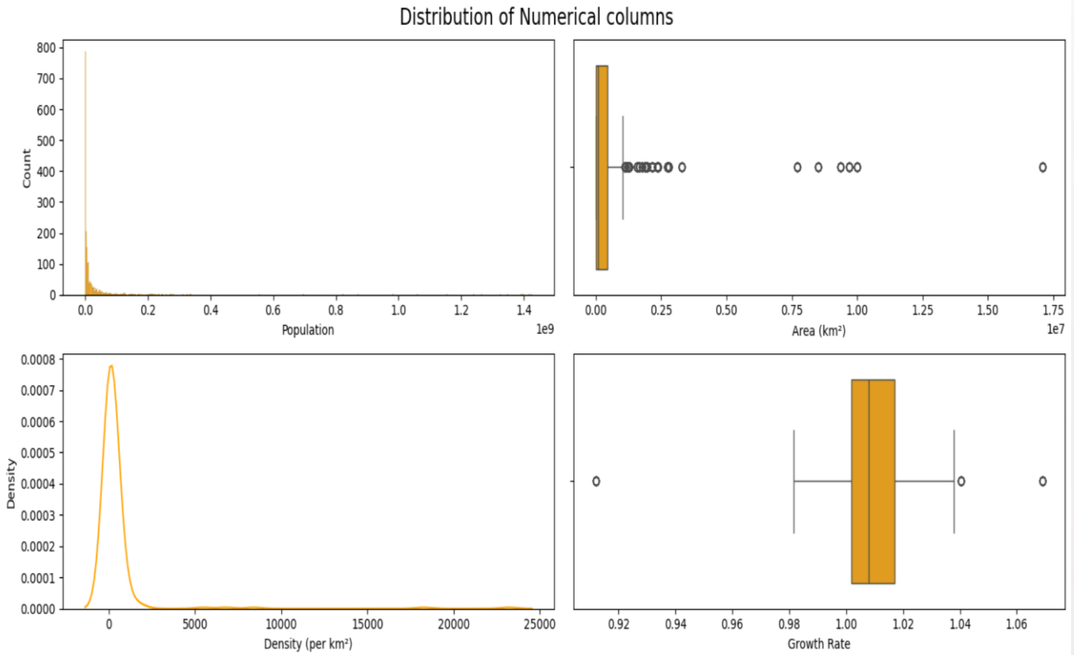
Applied log transformations to address positive skewness in:

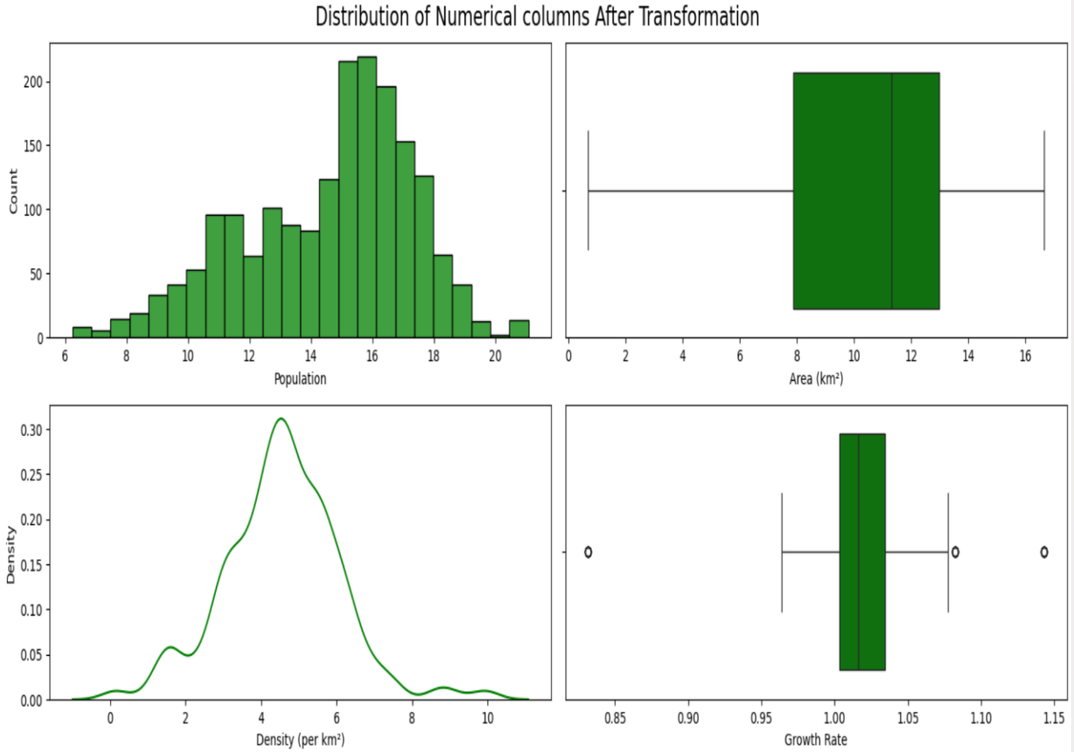
- Area (km^2)
- Density (per km^2)
- World Population Percentage
- Population

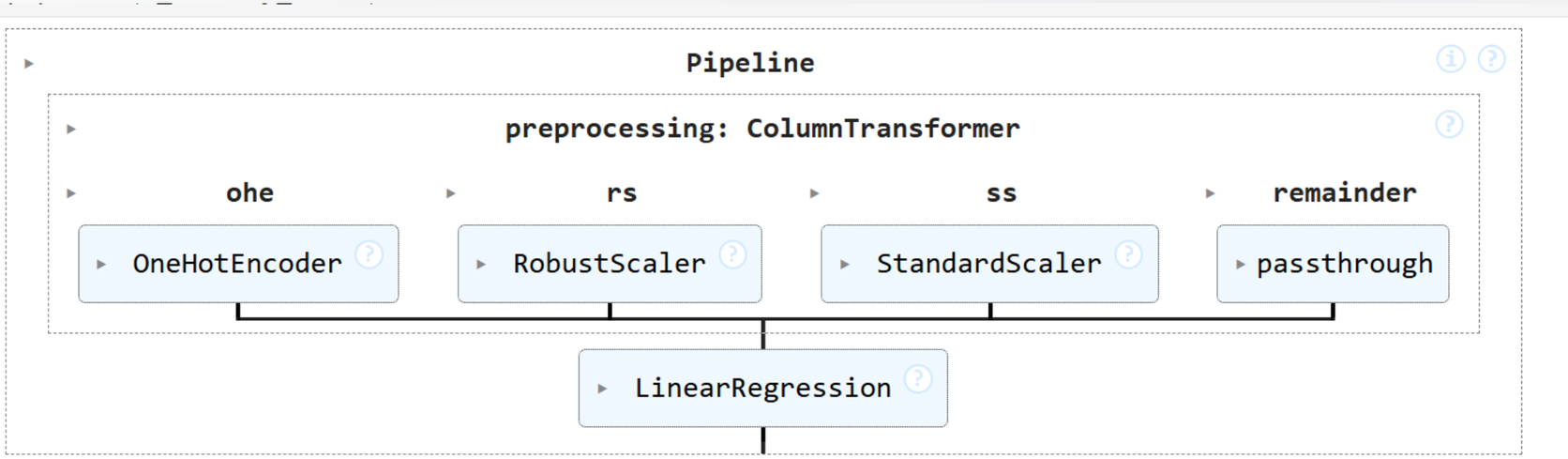
Results

Transformations significantly improved data distribution:

- Reduced skewness in most variables
- Prepared data for more accurate modeling
- World Population Percentage still showed some skewness







Modeling Approach

Data Preprocessing

Split data into training (80%) and testing (20%) sets

Feature Engineering

Applied OneHotEncoder for categorical variables (Country, Continent)

Used RobustScaler for skewed variables

Applied StandardScaler for normally distributed variables

Model Selection

Implemented Linear Regression as the baseline predictive model

Created a pipeline to streamline the modeling process

Conclusion & Key Insights

- Global population has shown consistent growth, especially in Asia, demanding focused regional planning.
- China and India lead in total population; Vatican City has the lowest—highlighting global disparities.
- Applied log and square transformations to correct skewness and improve model accuracy.
- Linear Regression model achieved $R^2 = 0.989$, indicating high prediction reliability.
- Insights support policy-making in infrastructure, healthcare, and resource allocation.