Birth Weight Analysis and Visualization Project - Report

Overview

This report outlines the steps, findings, and insights from a data analysis project focused on birth weights across different countries and years. The project encompasses a complete data pipeline from ingestion and cleaning to visualization and reporting.

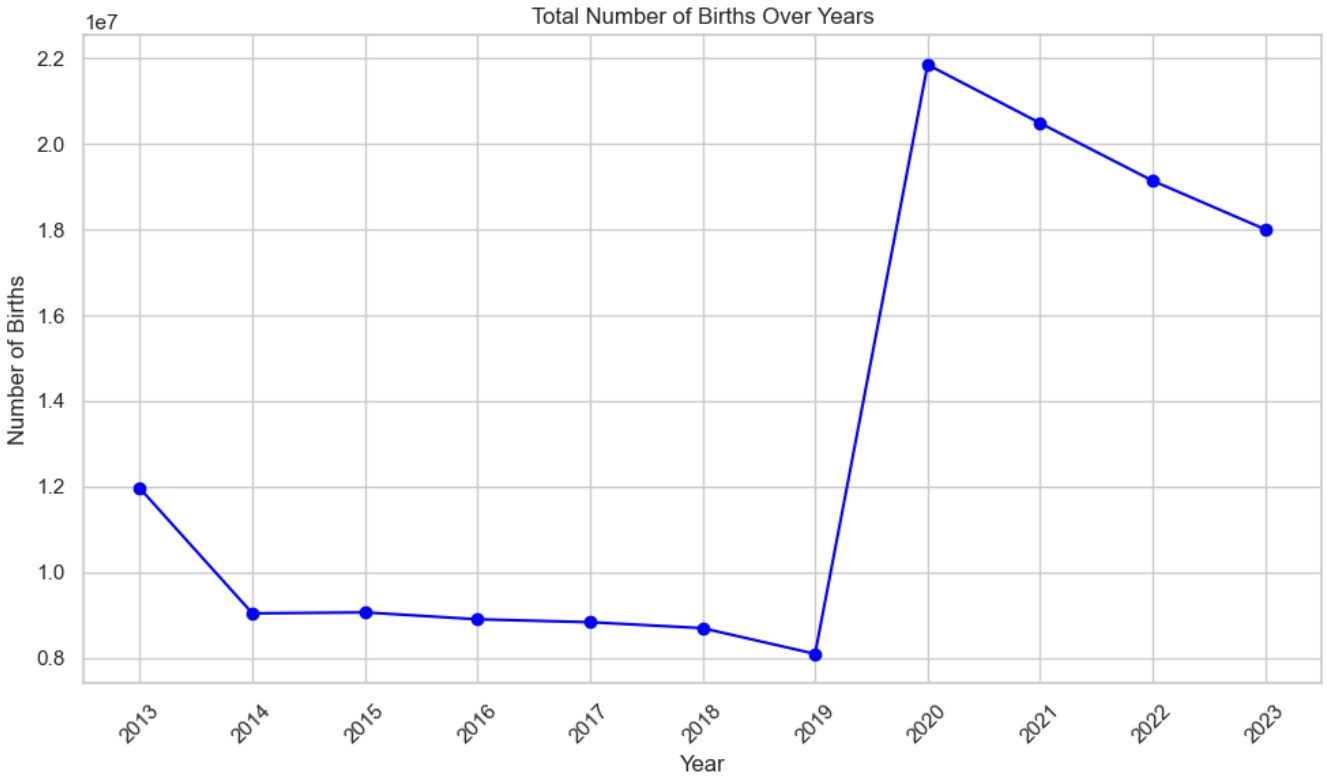
Data Cleaning and Transformation

- **Outlier Removal**: Utilized the **3-sigma rule** to identify and eliminate outliers beyond three standard deviations from the mean.
- **Log Transformation**: Applied `log1p` to the data to reduce skewness and better approximate a normal distribution.

Visual Analysis

1. Total Births Over Years

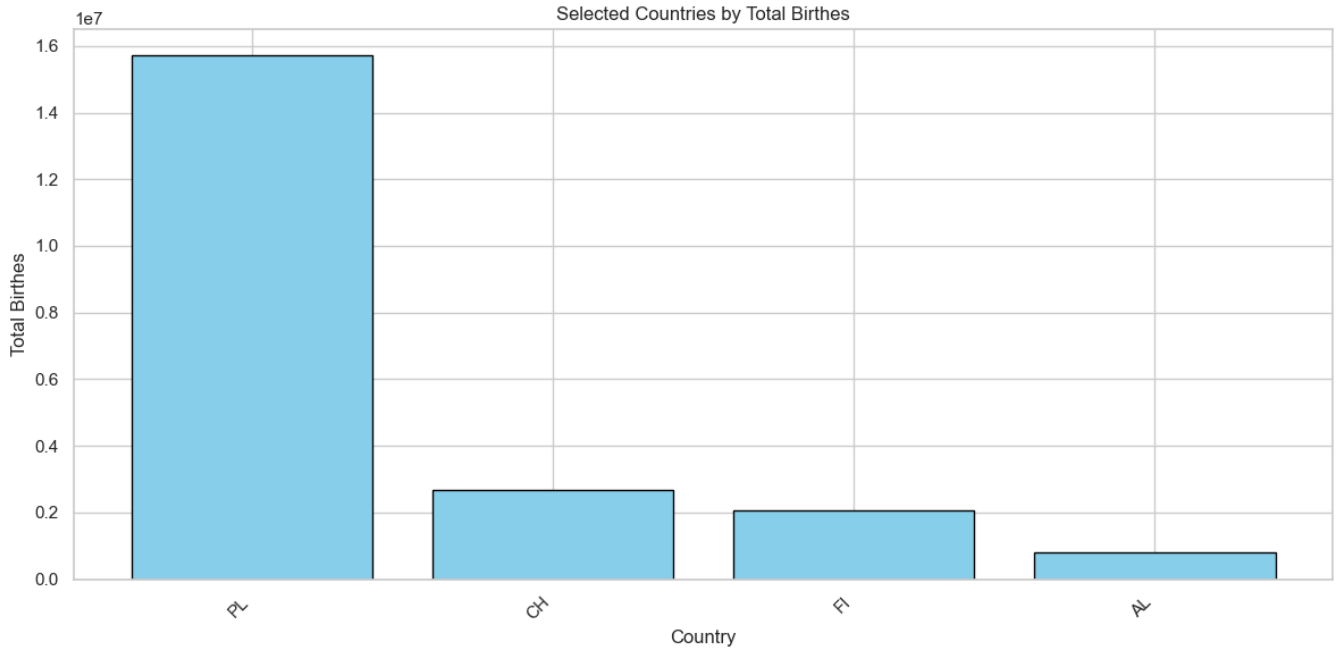
- A line chart visualizes trends in total birth counts over time.
- Key observations include periodic increases and decreases, possibly reflecting policy, health, or social changes.



---

2. Top 10 Countries by Birth Count

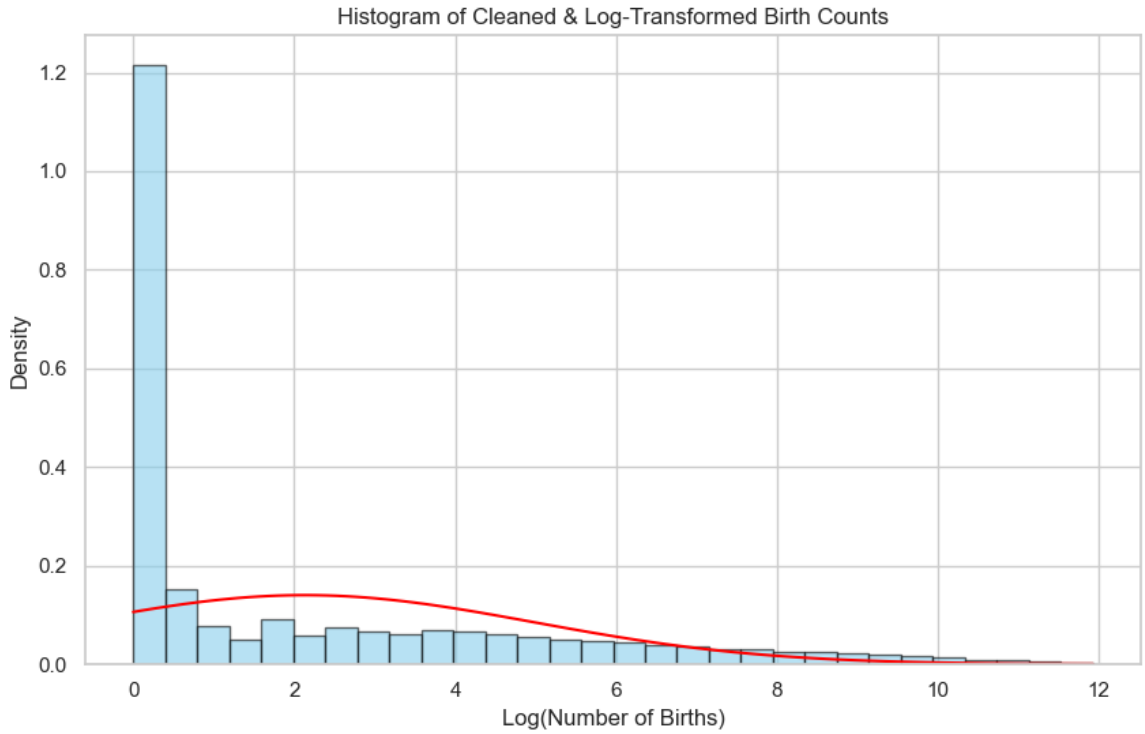
- Bar chart displaying the ten countries with the highest cumulative birth counts.
- Identifies regions with consistently high birth rates.



---

3. Countries by Total Amount of Births Comparison

- Comparative visualizations show the distribution of birth counts across all countries.
- Enables a macro-level understanding of birth distribution globally and identifies disparities or unusual trends between nations.



Statistical Summary

- **Raw Data**: Calculated initial mean and standard deviation for comparison.
- **Cleaned Data**: Recomputed summary statistics post-cleaning.
- **Insights**: Transformation significantly improved data normality, making it more suitable for statistical inference under the **Central Limit Theorem (CLT)**.

Interactive CLI

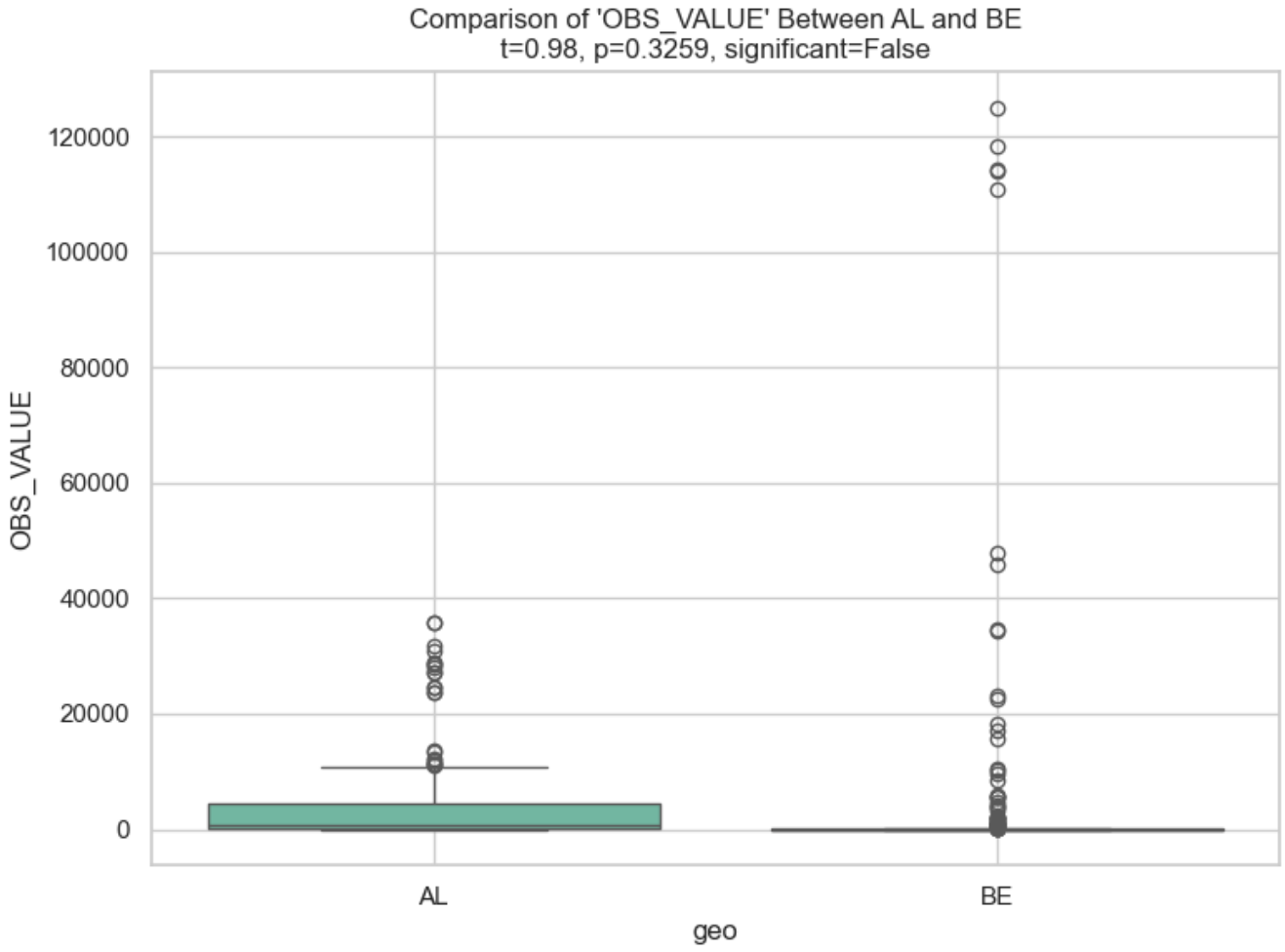
- Implemented using `InquirerPy`.
- Users can select from a menu of reports and generate new graphs without editing the source code.
- Enhances user accessibility and usability of the data pipeline.

Application of Central Limit Theorem

- By reducing outliers and applying normalization, the data better satisfies the assumptions of the CLT.
- This facilitates more accurate inference and hypothesis testing on birth data.

Hypothesis Proving System

- A modular subsystem to test hypotheses based on user-defined filters (e.g., year range, country selection).
- Provides p-values and confidence intervals as output.



Technologies Used

- **Python 3.9+**
- `pandas` — Data wrangling
- `matplotlib` — Plotting
- `numpy` — Math operations
- `scipy.stats` — Distribution fitting
- `InquirerPy` — Interactive CLI