BENFORD'S LAW & ANALYSIS

By : Data-Refiners

Dataset : Spanish Cities

DID YOU KNOW ??

Most natural datasets, the number 1 appears as the first digit six times more often than the number 9? Today we'll explore this fascinating mathematical pattern in Spanish city populations.

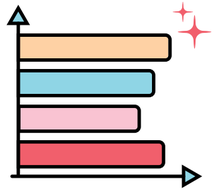
TEAM CONTRIBUTION :

Ipython Notebook

- **Vamshi Krishna Pendyala**
- **Sravanth Varri**

Presentation

- **Uday Kumar Choudhary**
- **R Sai Praneeth Sharma**



WHAT IS BENFORD'S LAW?

- Mathematical phenomenon where the first digits in many real-world numerical datasets follow a predictable pattern
- The digit "1" appears as the first digit about 30.1% of the time
- Each subsequent digit (2-9) appears with decreasing frequency
- The digit "9" appears only about 4.6% of the time

WHY IT MATTERS:

- Acts as a "natural fingerprint" for authentic data
- Deviations can signal potential data manipulation or anomalies
- Applied in fraud detection, data verification, and quality control
- Provides insights into the natural growth patterns of cities and populations

DATASET : SPANISH CITIES

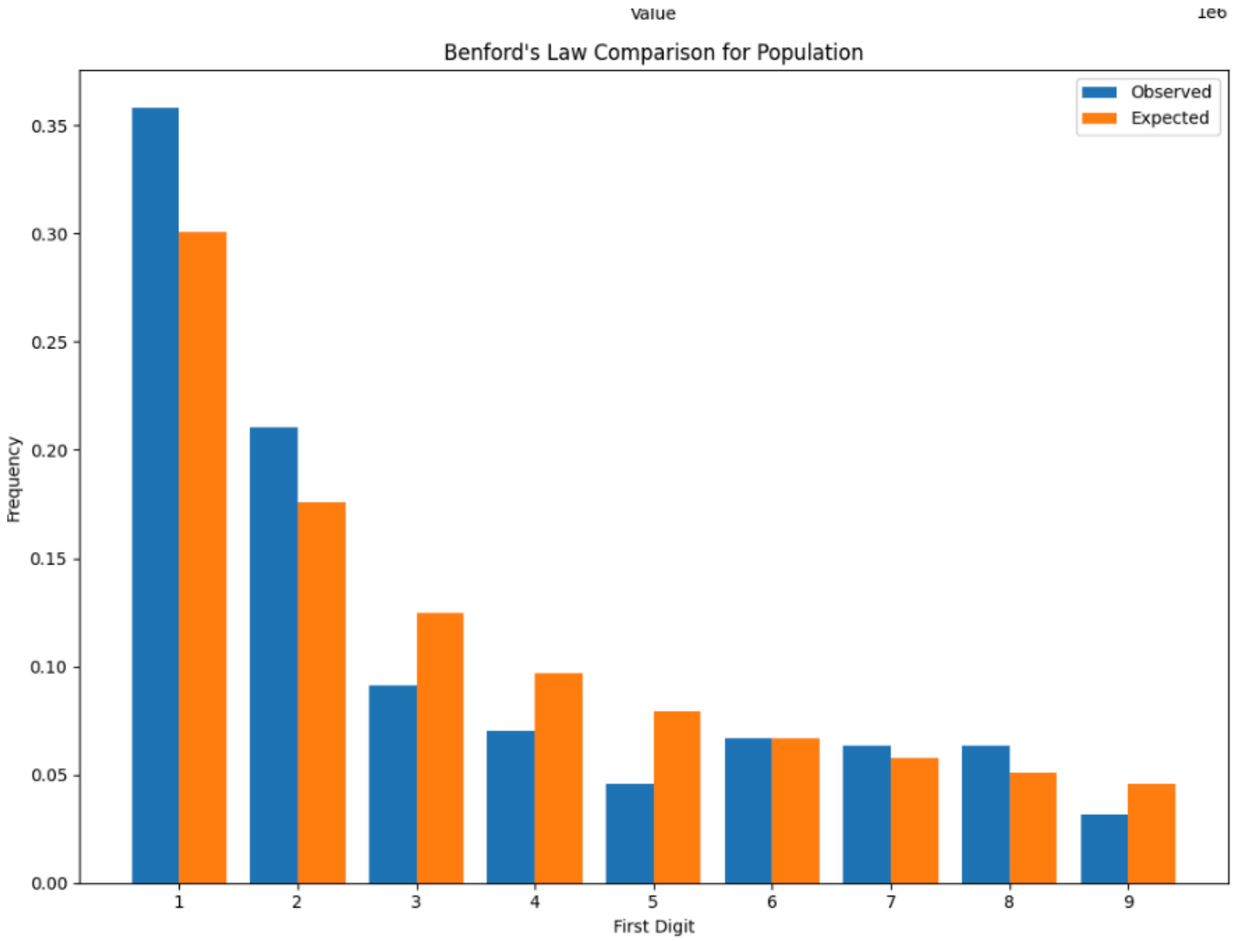
Dataset Specifications:

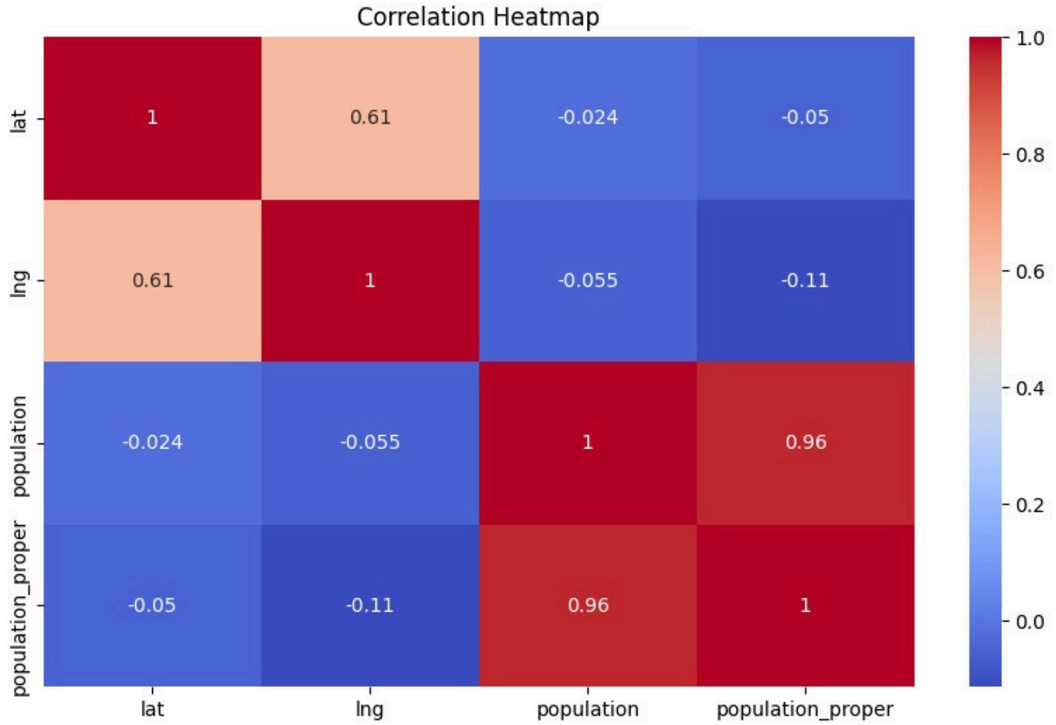
1. Comprehensive collection of 250+ Spanish cities
2. Includes geographic coordinates (latitude/longitude)
3. Administrative information (capital status, region)
4. Two distinct population measurements

Population Metrics Explained:

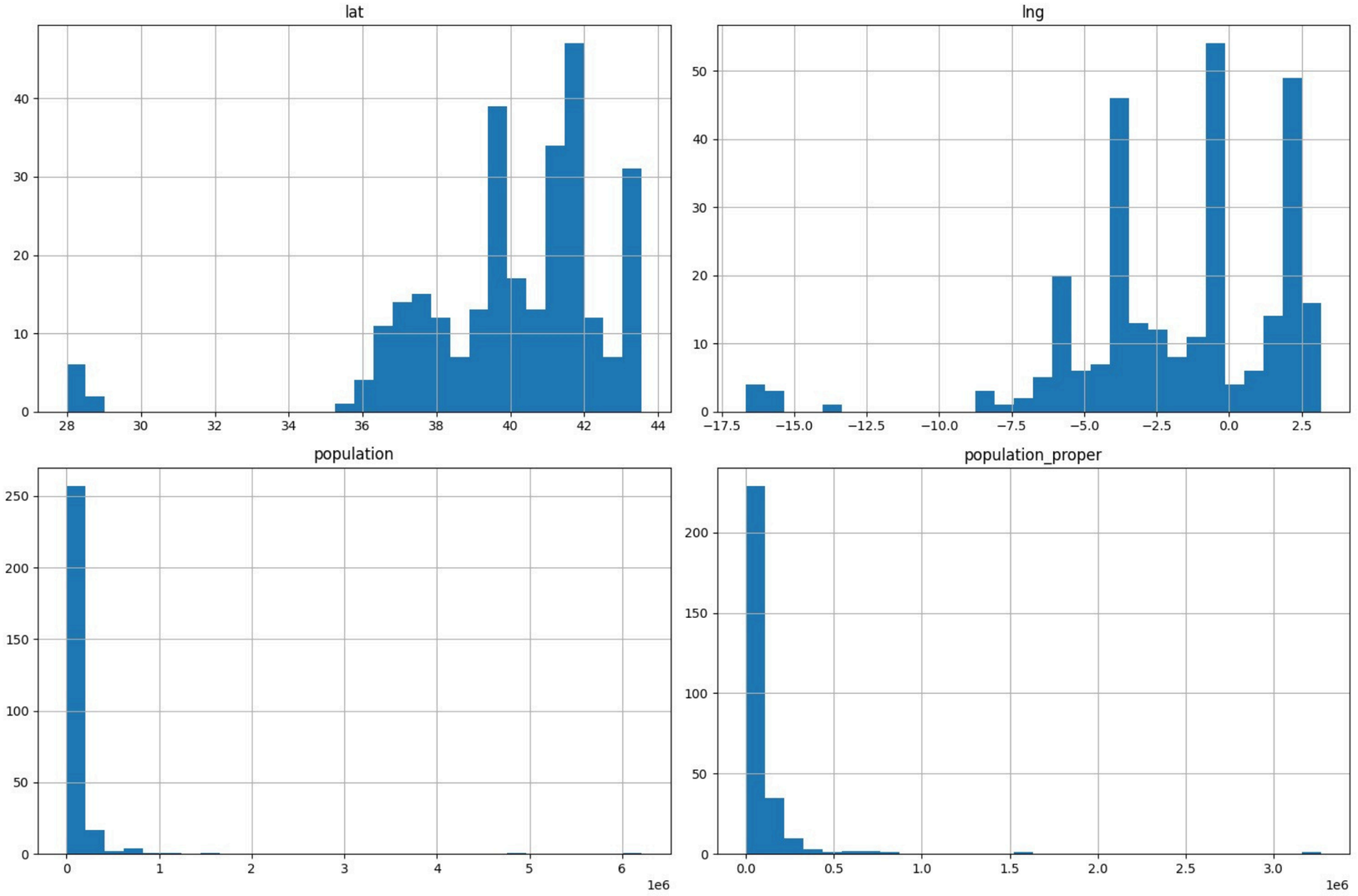
1. Population: Total metropolitan area population (including suburbs)
2. Population_proper: City limits population (administrative boundaries)
3. Range from major metropolises (Madrid: 6.2M) to small towns (under 2,000)
4. Diverse geographic distribution across 17 autonomous communities

RESULTS VISUALIZATION FOR GIVEN DATASET





RESULTS VISUALIZATION FOR GIVEN DATASET



DATA ANALYSIS KEY FINDINGS

- **Missing Capital Data**: The 'capital' column initially had 245 missing values. These were imputed using the mode (most frequent value, which is 'admin').

- **Benford's Law Analysis**: A Chi-squared test comparing the observed first-digit frequencies of the 'population' data against the expected frequencies based on Benford's Law resulted in a p-value of 0.999999986. This indicates that the population data does not significantly deviate from Benford's Law.

- **No significant deviations from Benford's Law were found**: The analysis did not reveal any digits with deviations greater than 0.01 from the expected frequencies based on Benford's Law, further supporting the conclusion of no significant data anomalies.

BENFORD'S LAW ANALYSIS:

- The analysis aimed to determine if the 'population' data in the Spain cities dataset follows Benford's Law, which describes the expected distribution of first digits in naturally occurring datasets.
- The Chi-squared test, used to compare the observed and expected frequencies of first digits, resulted in a high p-value (0.9999999752236586).
- This high p-value indicates that the deviations from Benford's Law for the population data are not statistically significant.
- There is **no** strong evidence of data anomalies or manipulations based on Benford's Law analysis for this specific dataset.

OVERALL INSIGHTS

- The population data in the Spain cities dataset appears to be consistent with what would be expected by random chance, according to Benford's Law.
- This suggests that the data is likely authentic and not subject to significant manipulation or fabrication related to population figures.
- However, further data quality checks might be necessary to ensure overall data integrity, including examining other numeric features for outliers or inconsistencies.
- Exploring other features in the dataset (e.g., latitude, longitude) could reveal additional patterns or insights about the distribution of cities in Spain.

THANK YOU

DATA-REFINERS SIGNING OFF
UNTIL NEXT TIME,
CONTINUE REFINING