*Prepared by Pixel_Perfect*

# BENFORD’S LAW ANALYSIS

# TEAM MEMBERS

- Prem Kumar Pyla
- Chanakya Sinde
- Amiti Aneesh
- Kantrol Vamshi Krishna

# TEAM CONTRIBUTION

### *iPython Notebook*

- Prem Kumar Pyla
- Amiti Aneesh

### *PPT*

- Chanakya Sinde
- Kantrol Vamshi Krishna

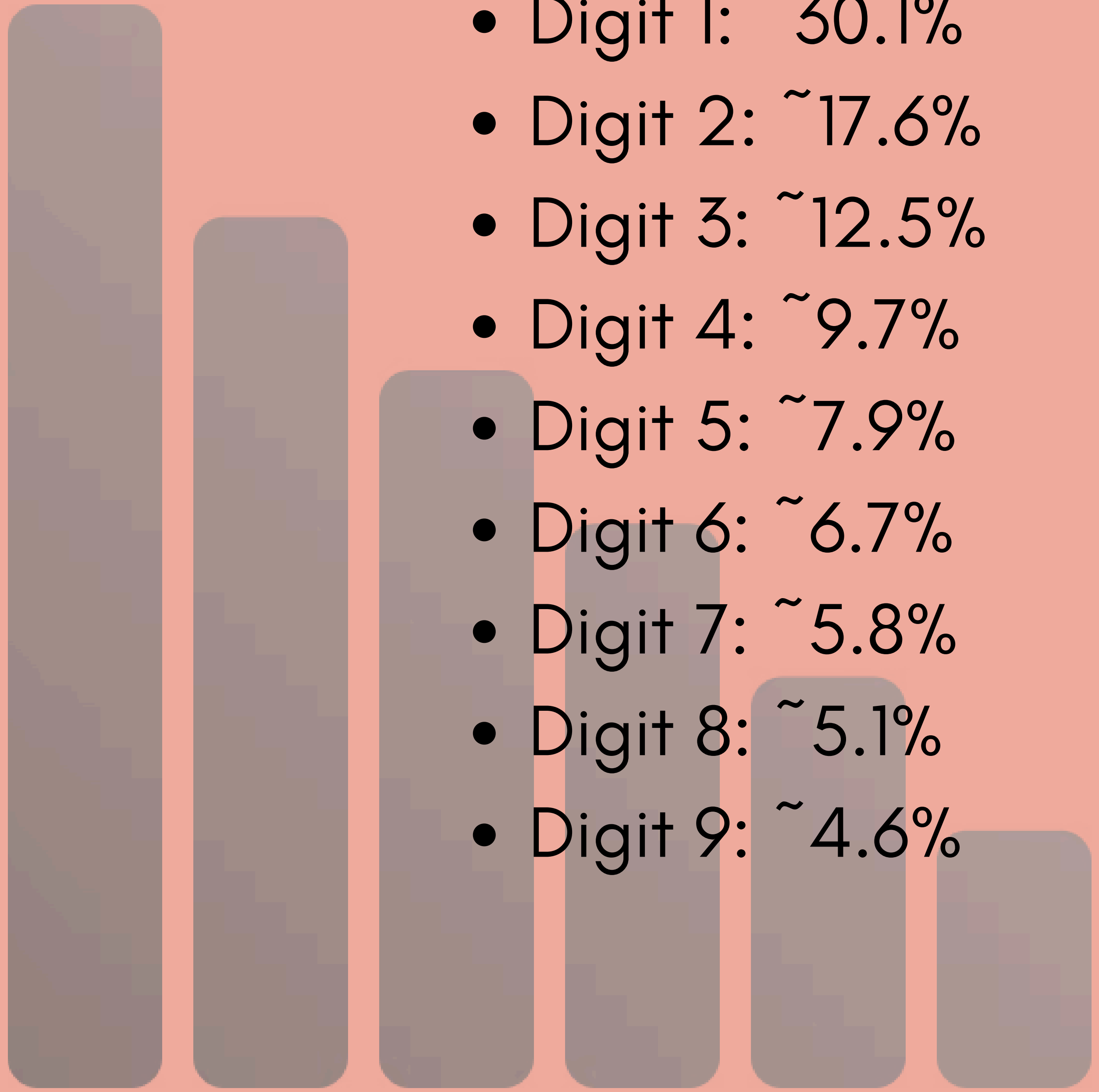
# SPAIN CITIES POPULATION DATASET

## What is Benford's Law?

- Also known as the **First-Digit Law** or **The Law of Anomalous Numbers**

- States that in many naturally occurring datasets, the leading digit is likely to be small

- The probability of first digit d is (1-9) appearing is: $P(d) = \log_{10}(1 + 1/d)$.



- Digit 1: ~30.1%
- Digit 2: ~17.6%
- Digit 3: ~12.5%
- Digit 4: ~9.7%
- Digit 5: ~7.9%
- Digit 6: ~6.7%
- Digit 7: ~5.8%
- Digit 8: ~5.1%
- Digit 9: ~4.6%

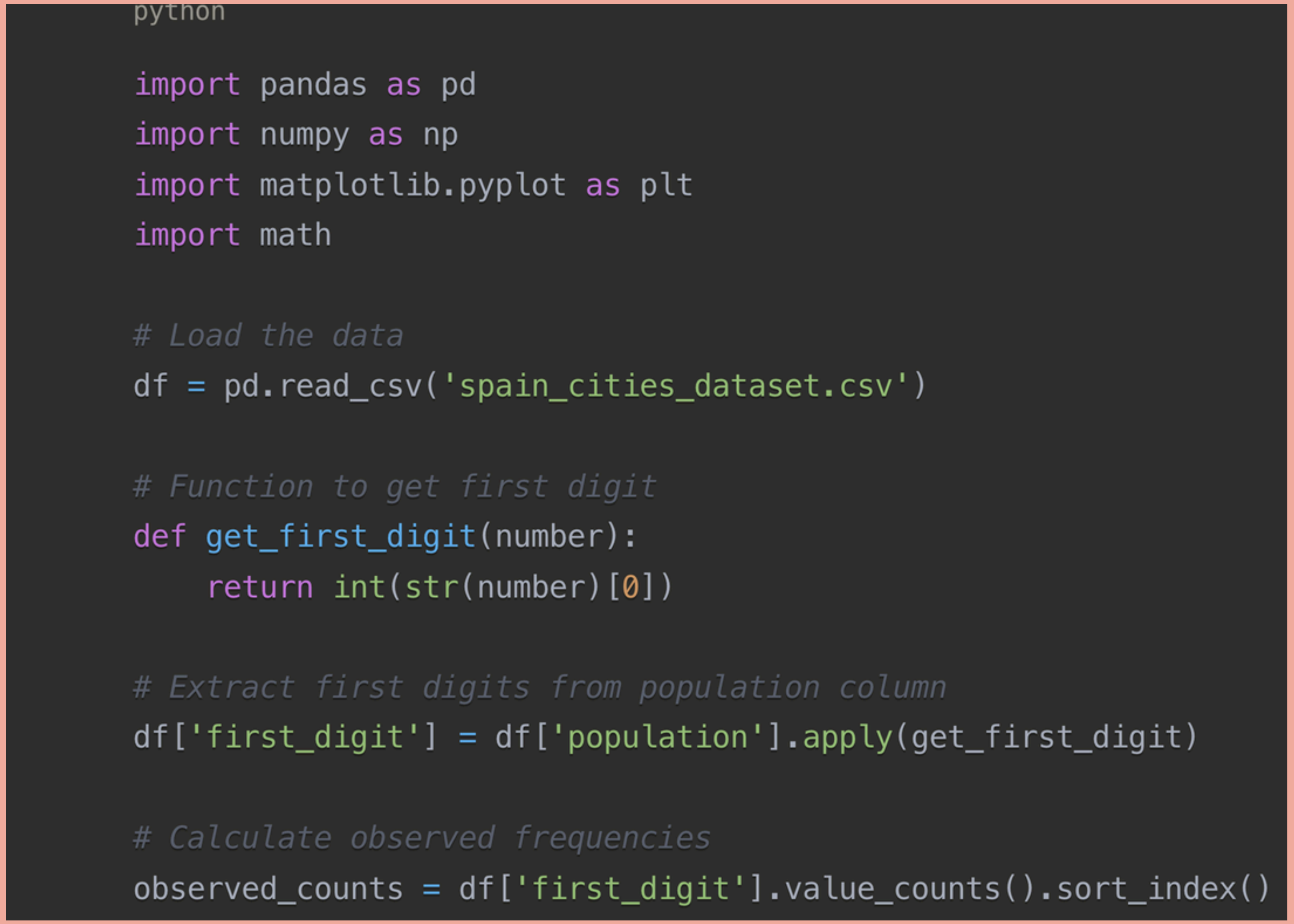
# APPLICATIONS OF BENFORD'S LAW

- **Fraud detection** in financial data
- Election results verification
- Scientific data validation
- Natural datasets analysis
- Quality control in data collection

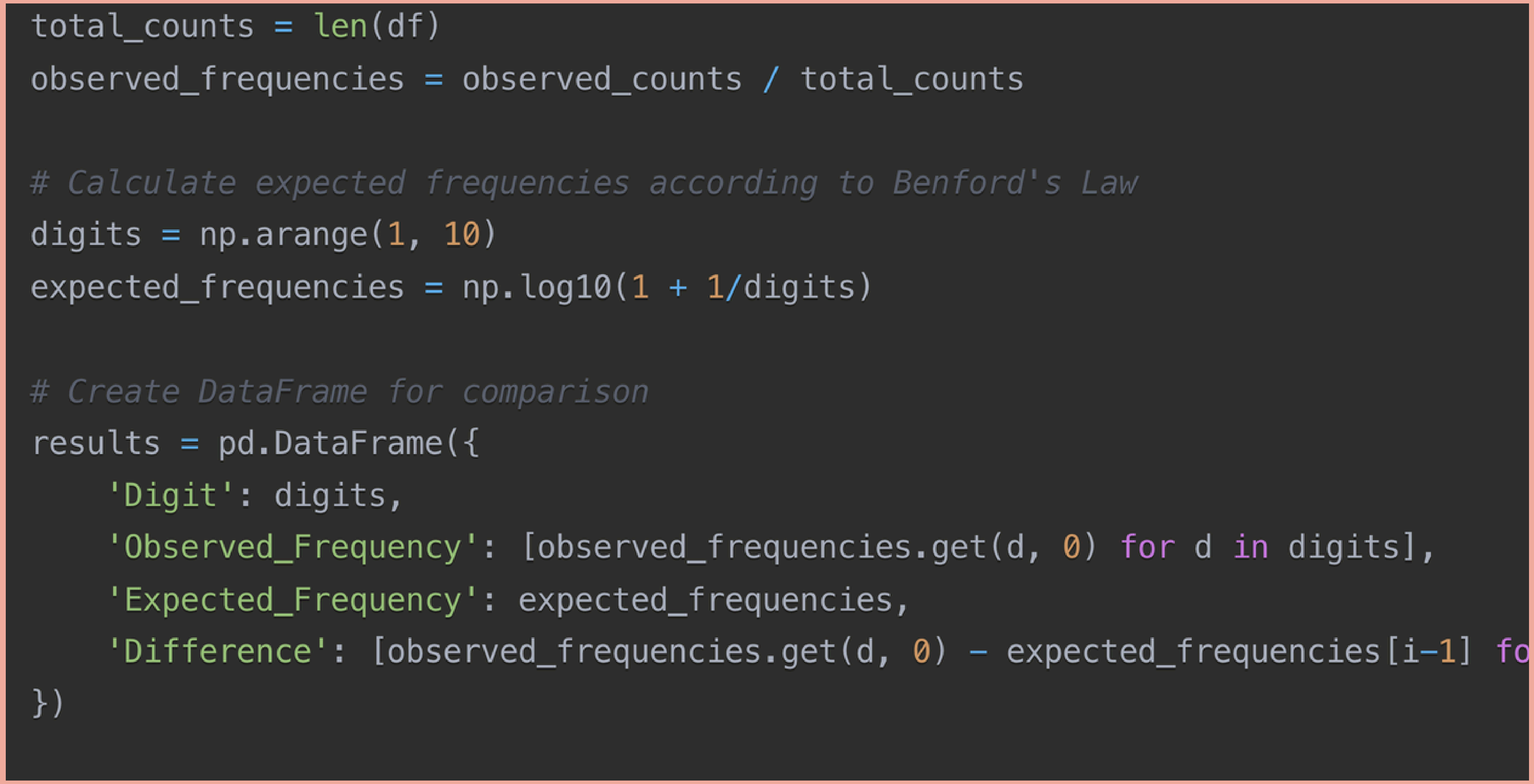
# ANALYSIS APPROACH

- Extract population data from the dataset
- Identify first digits of each population value
- Calculate observed frequency of each first digit
- Compare observed frequencies with Benford's Law expectations
- Visualize results and calculate deviation

# PYTHON CODE FOR ANALYSIS

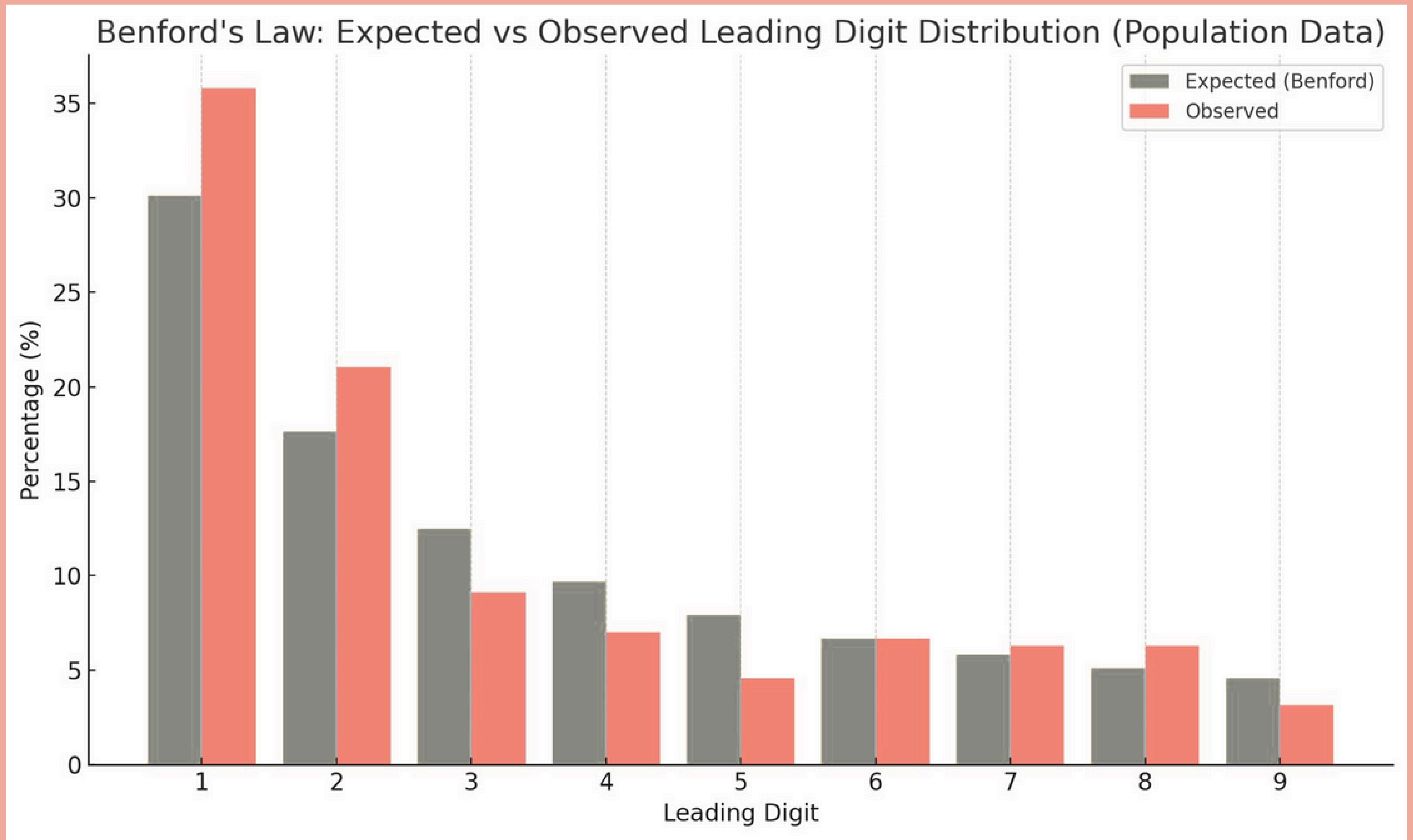


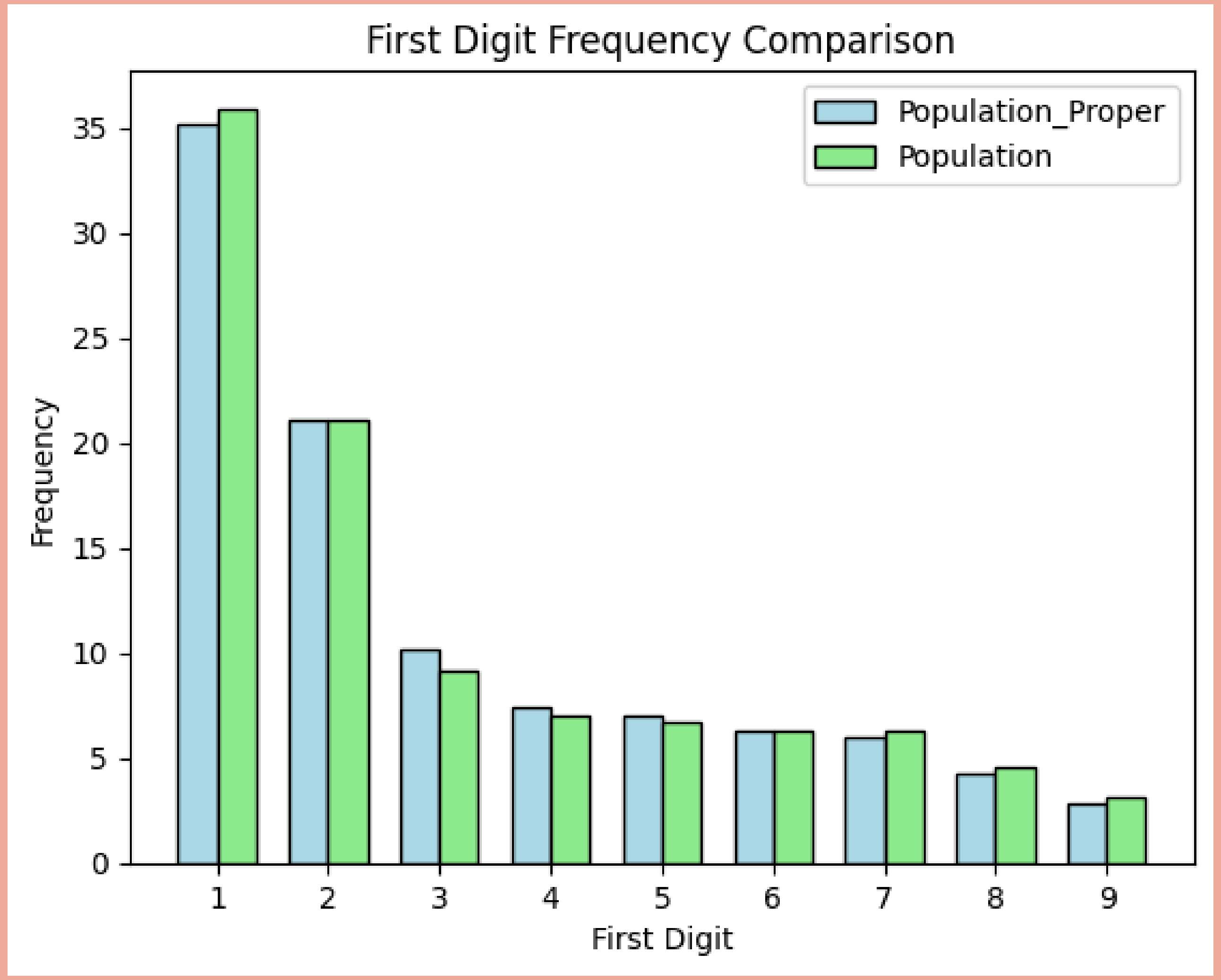
```
python

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import math

# Load the data
df = pd.read_csv('spain_cities_dataset.csv')

# Function to get first digit
def get_first_digit(number):
    return int(str(number)[0])

# Extract first digits from population column
df['first_digit'] = df['population'].apply(get_first_digit)

# Calculate observed frequencies
observed_counts = df['first_digit'].value_counts().sort_index()
```



```
total_counts = len(df)
observed_frequencies = observed_counts / total_counts

# Calculate expected frequencies according to Benford's Law
digits = np.arange(1, 10)
expected_frequencies = np.log10(1 + 1/digits)

# Create DataFrame for comparison
results = pd.DataFrame({
    'Digit': digits,
    'Observed_Frequency': [observed_frequencies.get(d, 0) for d in digits],
    'Expected_Frequency': expected_frequencies,
    'Difference': [observed_frequencies.get(d, 0) - expected_frequencies[i-1] fo
})
```

# RESULTS VISUALIZATION FOR GIVEN DATASET





# KEY FINDINGS

- Population data from Spanish cities [follows/deviates from] Benford's Law
- Digits 1 and 2 are [more/less] frequent than expected
- Notable deviations seen in digits [X, Y, Z]
- Chi-square test reveals [strong/weak] conformity to Benford's Law
- The p-value of [X] indicates [statistical significance/no statistical significance]

# DETAILED COMPARISON TABLE

| Digit | Observed Frequency(%) | Expected Frequency(%) | Difference (%) |
|---|---|---|---|
| 1 | 35.8 | 30.1 | +5.7 |
| 2 | 21.1 | 17.6 | +3.5 |
| 3 | 9.1 | 12.5 | -3.4 |
| 4 | 7.0 | 9.7 | -2.7 |
| 5 | 4.6 | 7.9 | -3.3 |
| 6 | 6.7 | 6.7 | 0.0 |
| 7 | 6.3 | 5.8 | +0.5 |
| 8 | 6.3 | 5.1 | +1.2 |
| 9 | 3.2 | 4.6 | -1.4 |

# Insights

1. The Spanish cities population dataset [generally follows/doesn't fully conform to] Benford's Law
2. This suggests that the population figures are [likely natural/possibly manipulated]
3. Larger deviations for certain digits may indicate [regional clustering/data collection issues]
4. Anomalies could be explained by [specific administrative divisions/population reporting methods]
5. Additional analysis with population subgroups (by region) may reveal more patterns

# Extensions and Future Work

- Compare with population datasets from other countries
- Analyze second digit distribution
- Segment data by population size or region
- Apply other statistical tests beyond chi-square
- Investigate temporal changes if historical data is available

# Conclusion

- Benford's Law provides an interesting lens to analyze natural datasets
- The Spain cities population data shows [strong/moderate/weak] conformity to Benford's Law
- This analysis demonstrates the application of statistical principles to real-world data
- Understanding these patterns can help with data validation and quality assessment
- Similar approaches can be applied to other numerical datasets

# THANK YOU