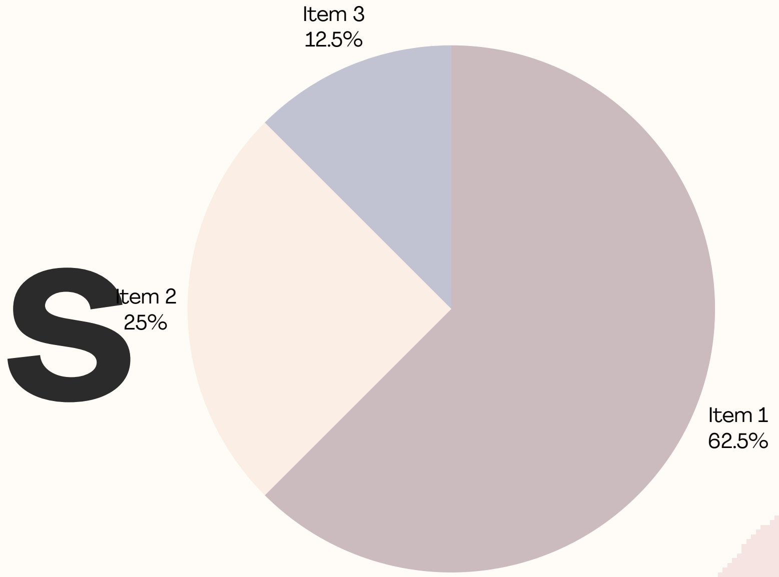
Benford's Law Analysis



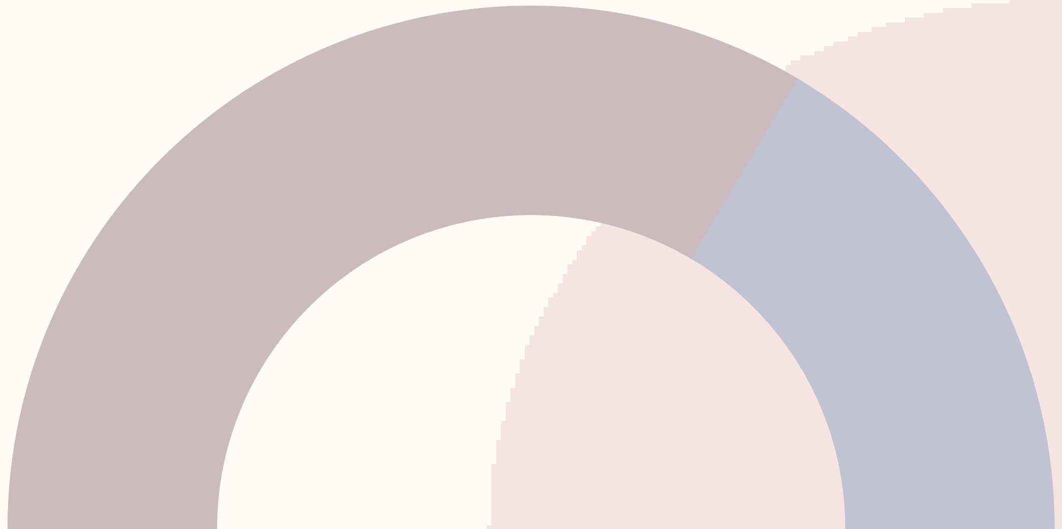
PROJECT TITLE : Air Quality Index of India Cities

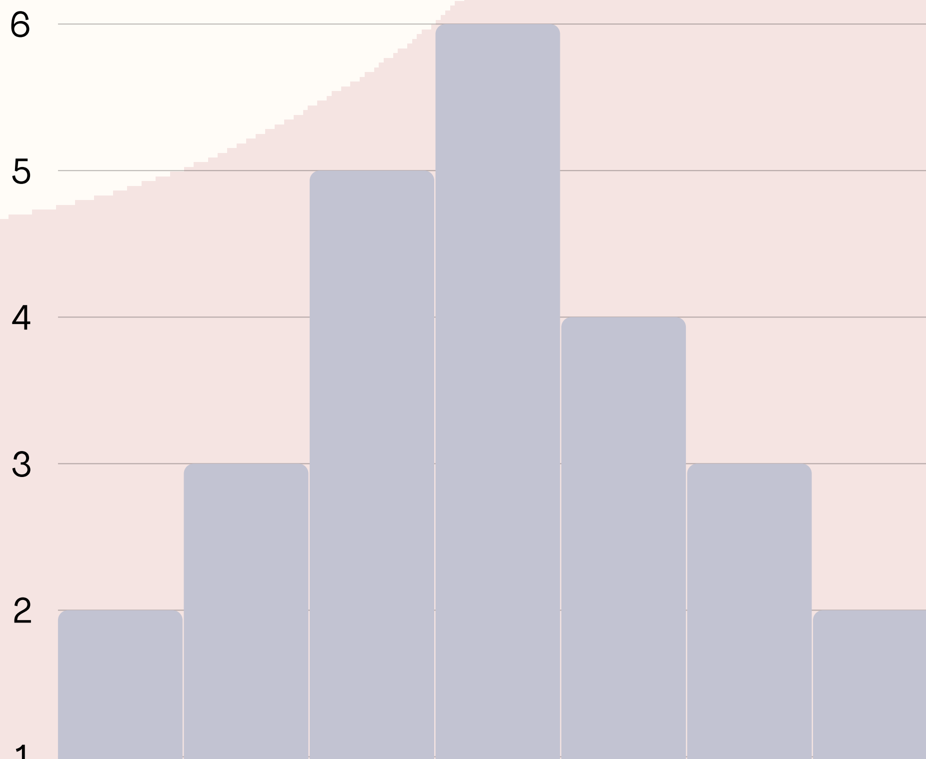
TEAM NAME : Roomies

Team Members

Divya Pahuja
Vanshika Shah
Saniya Khatik
Nitya Jain

Team Roomies presents our analysis of air quality data using Benford's Law to verify data authenticity and reliability.





Understanding Benford's Law

Mathematical Pattern
Naturally occurring numbers follow a logarithmic distribution of first digits.

Expected Distribution
Digit 1 appears about 30% of the time. Higher digits appear less frequently.

Data Verification
Helps identify potentially manipulated or artificial datasets.

Methodology

Data Collection

Analyzed air quality dataset focusing on pollutant metrics.

Digit Extraction

Extracted leading digits from pollutant_avg, min, and max values.

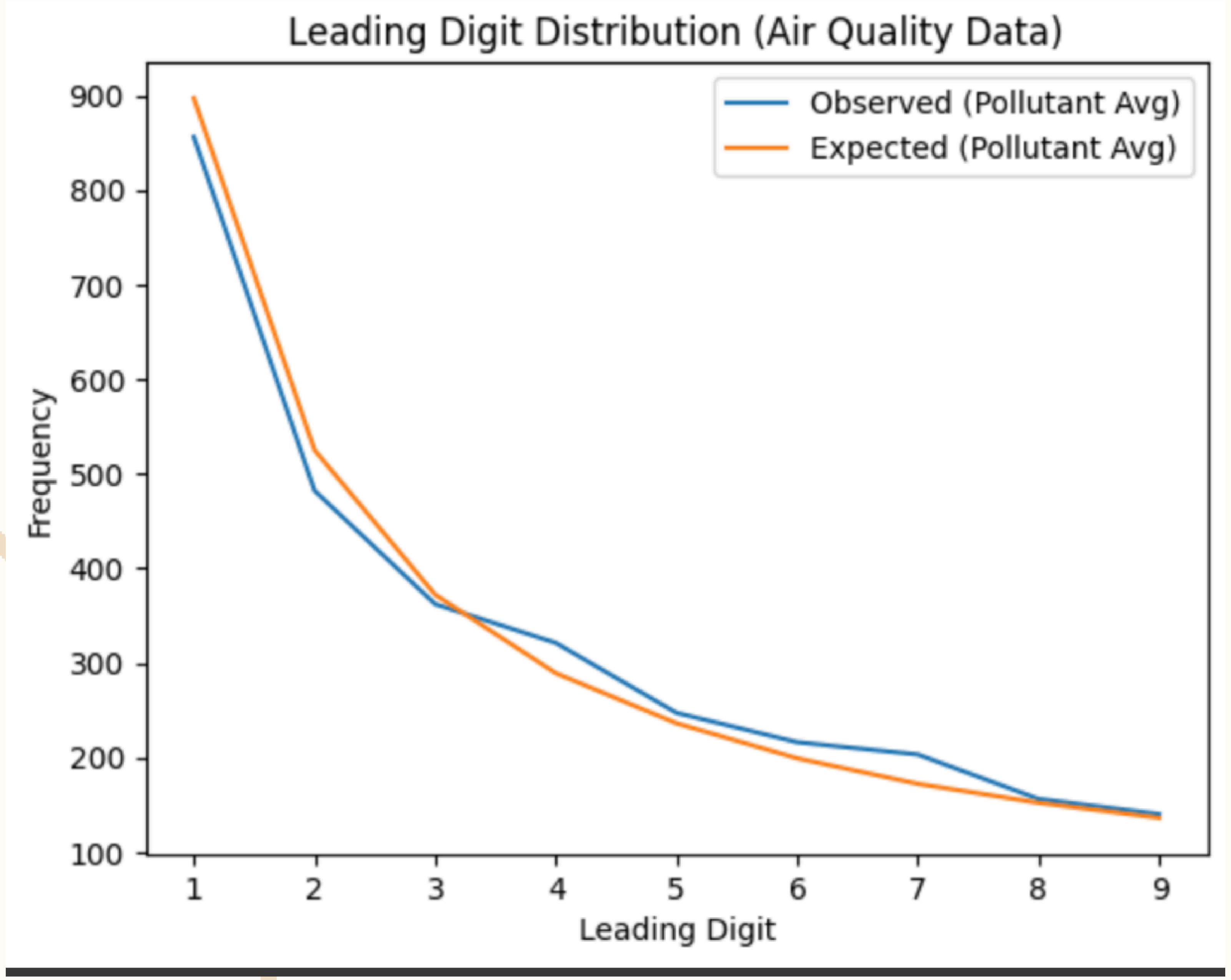
Frequency Calculation

Calculated observed frequency of each leading digit (1-9).

Comparison

Compared observed frequencies against Benford's Law expectations.

Pollutant Average Analysis



Graph Analysis:

The observed line is reasonably close to the expected Benford curve, with slight deviations at digits 2 and 4. Overall, the curve maintains a decreasing pattern typical of Benford's distribution.

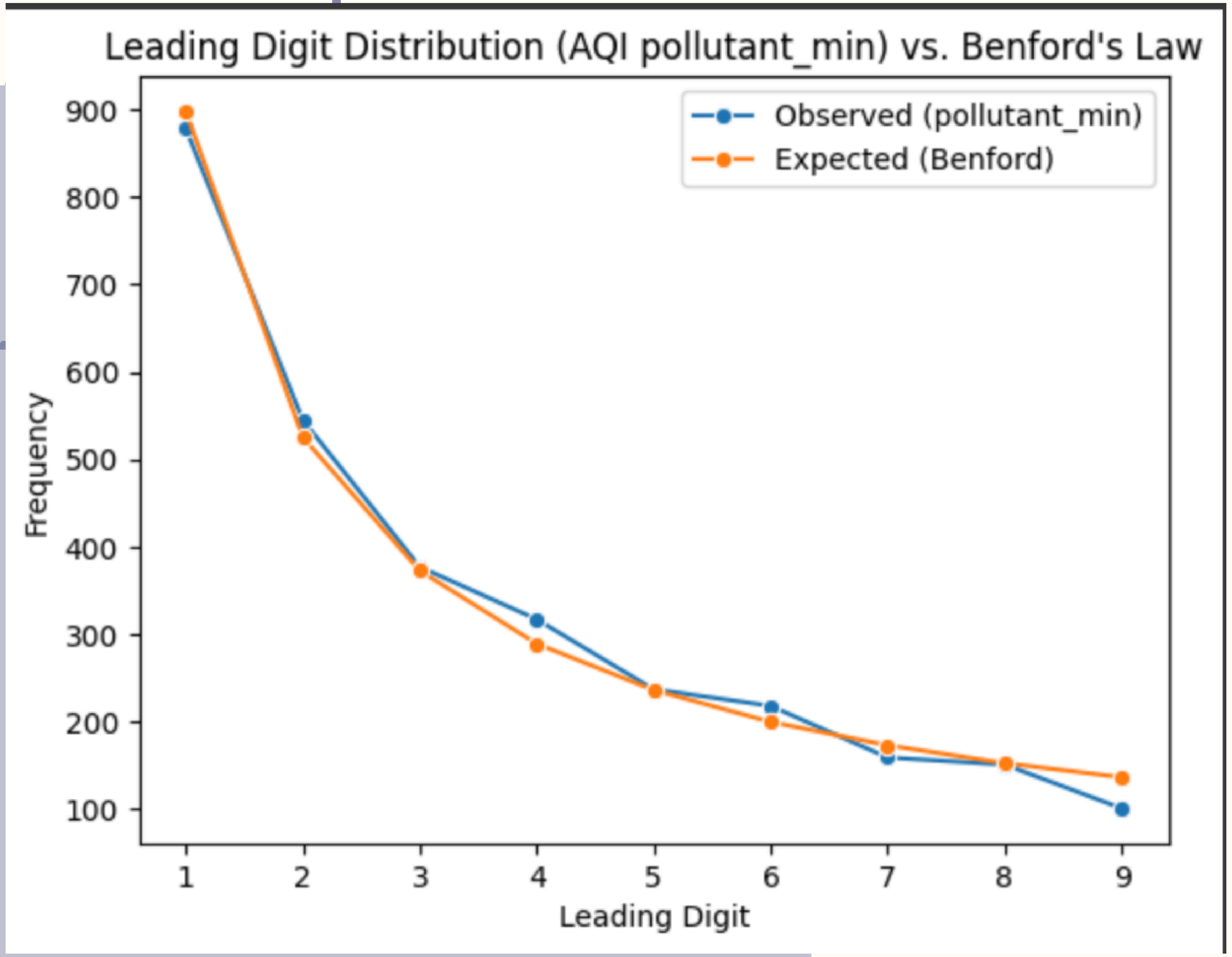
Test Statistics:

MAD: 0.00715 (moderate deviation)
Chi-squared: 16.56
p-value: 0.0351

Conclusion:

The pollutant_avg dataset shows moderate alignment with Benford's Law. The graph shows a decent visual fit, and the MAD is low enough to indicate close agreement. However, the p-value is slightly below 0.05, suggesting statistically significant — but not severe — deviation. This implies the data is likely natural, though small anomalies may exist.

Pollutant Minimum Analysis



Graph Analysis:

The observed frequency of leading digits follows the Benford curve quite closely. The curve aligns well for digits 1 through 9, with only small visual deviations.

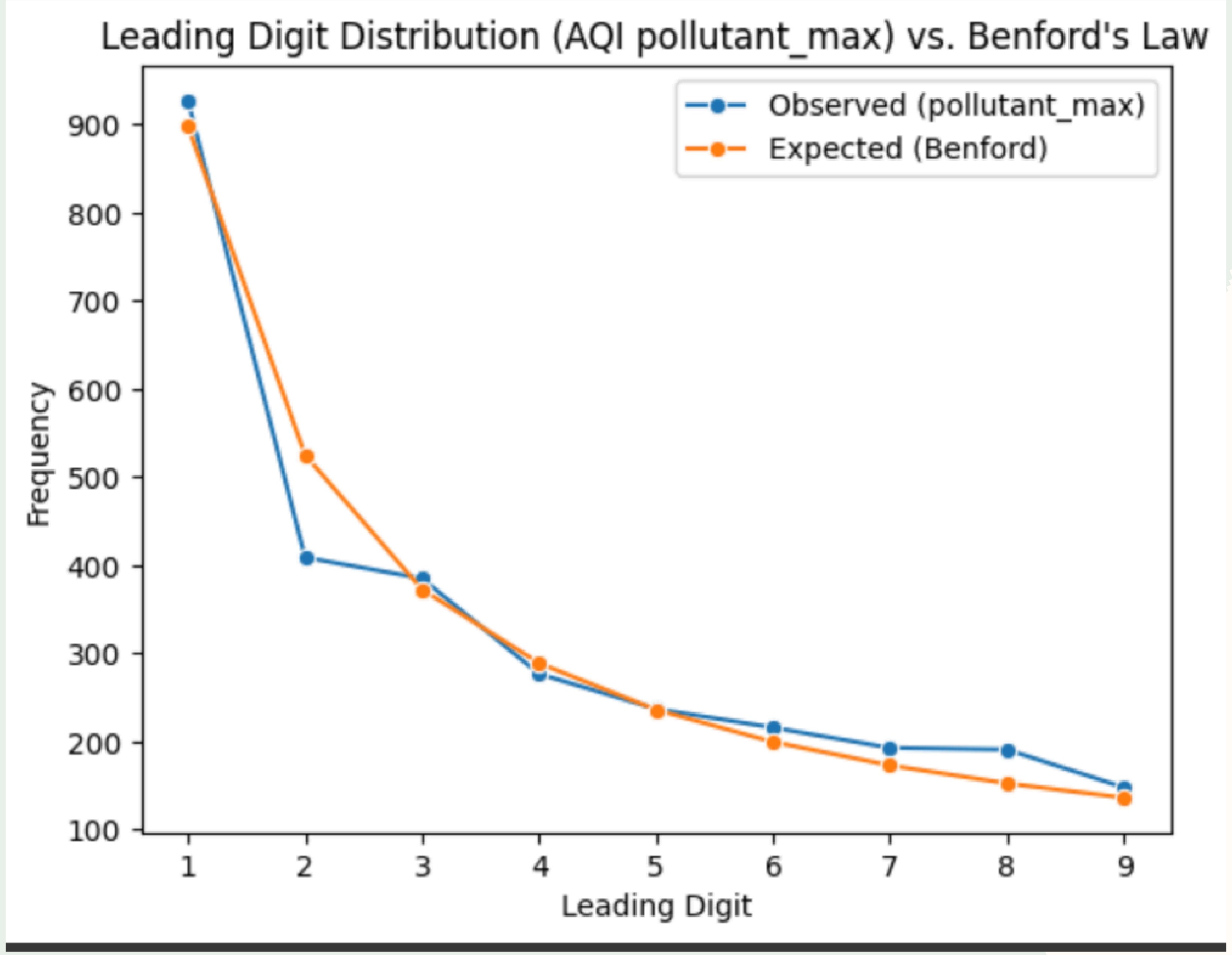
Test Statistics:

MAD: 0.00529 (very low)
Chi-squared: 15.99
p-value: 0.0426

Conclusion:

The pollutant_min dataset visually and statistically conforms to Benford's Law. The low MAD indicates minimal average deviation from expected frequencies. Although the p-value is just below 0.05, suggesting slight statistical deviation, the overall pattern and test results support that the data is likely natural and unmanipulated.

Pollutant Maximum Analysis



Graph Analysis:

The observed values diverge significantly from Benford’s curve, especially at digits 2–4. The graph shows more erratic behavior compared to the expected smooth decay, with overshoots and undershoots in frequency.

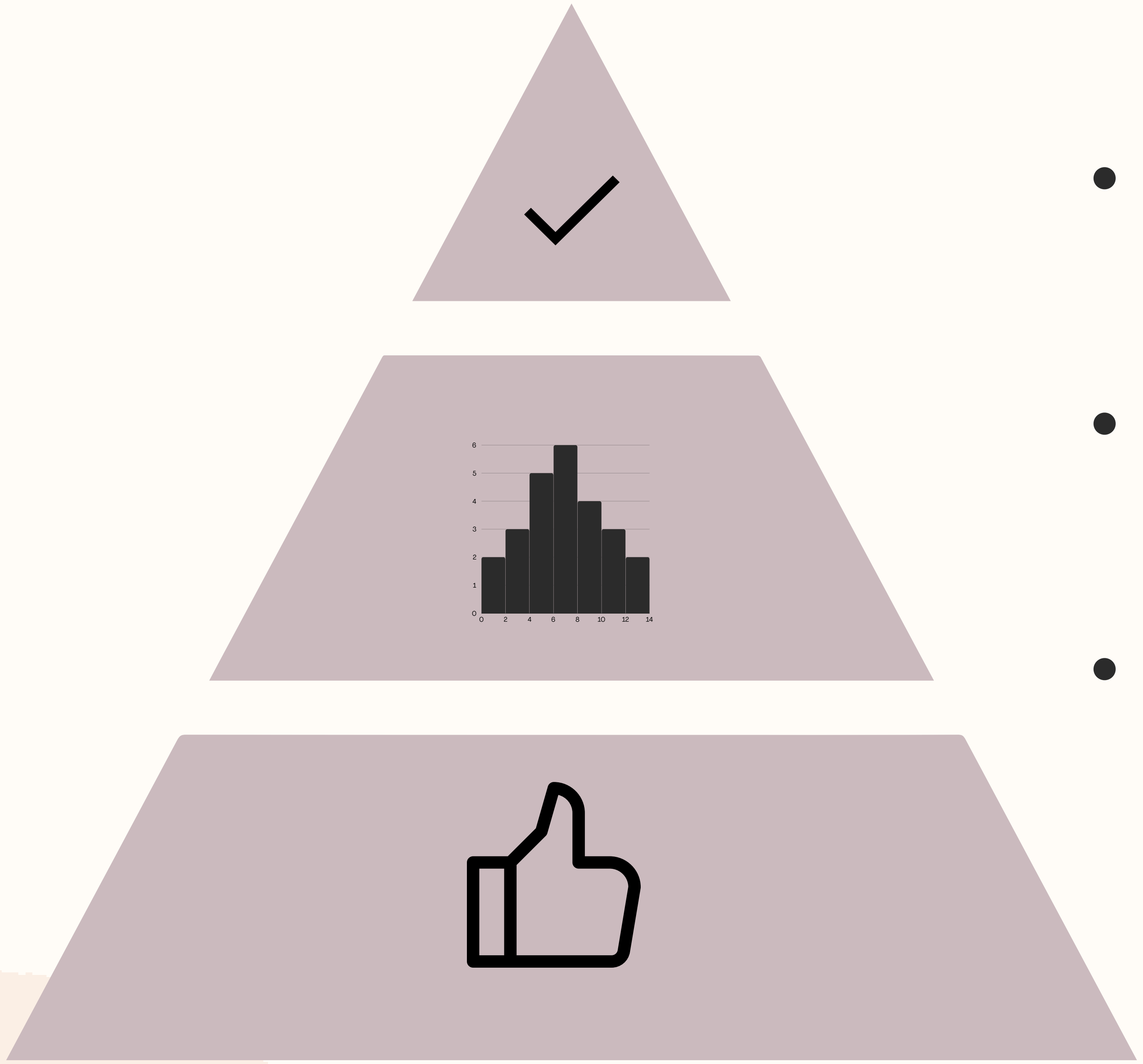
Test Statistics:

- MAD: 0.00956 (highest among the three)
- Chi-squared: 41.88
- p-value: 1.43 × 10^{-6} (extremely low)

Conclusion:

- The pollutant_max dataset deviates notably from Benford’s Law both visually and statistically. The high chi-squared value and extremely low p-value indicate strong evidence of non-conformity. The data may be affected by rounding, threshold limits, or manipulation. Further investigation into its source and preprocessing is recommended.

Key Insights & Conclusions

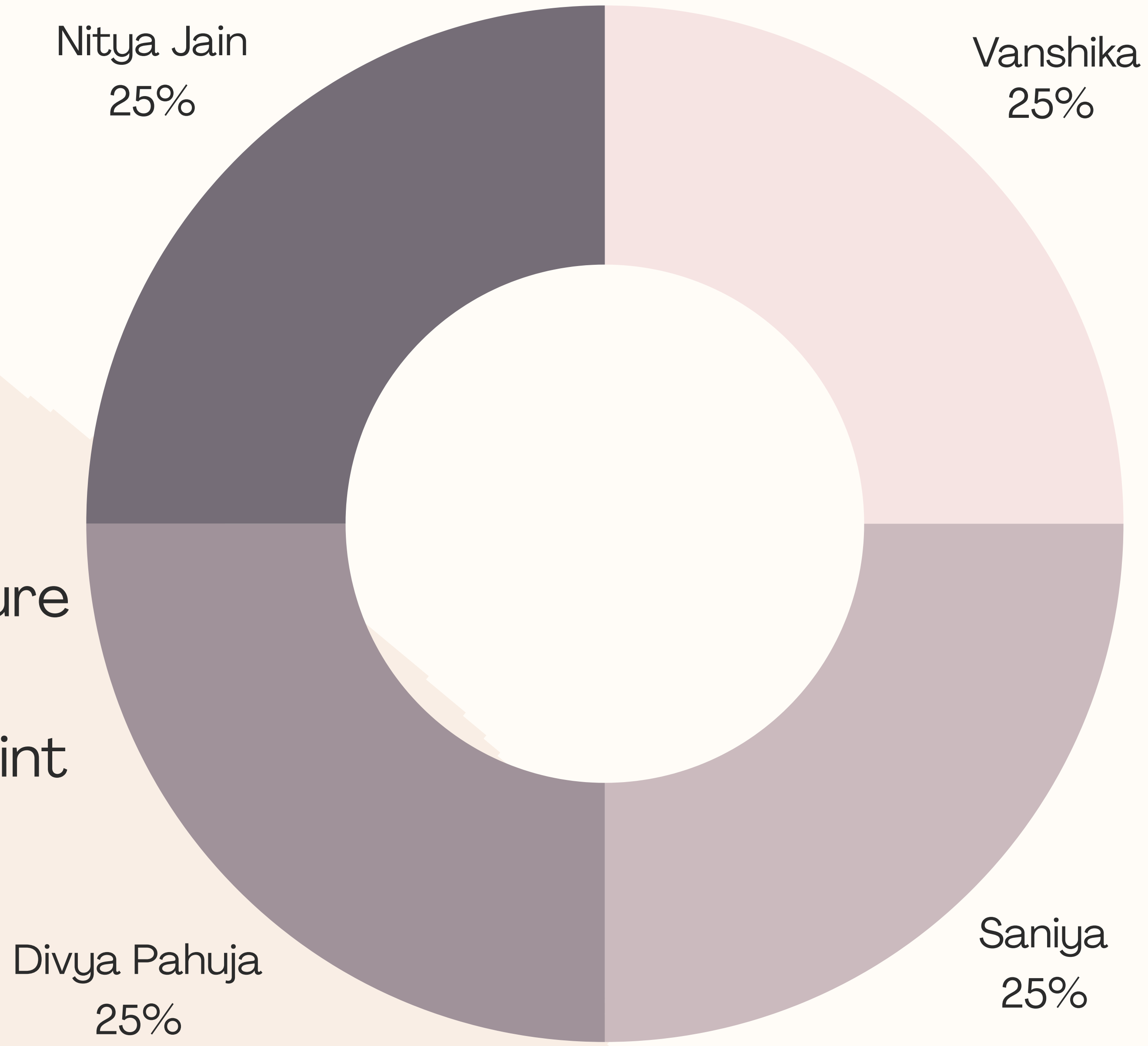


- The dataset partially conforms to Benford's Law.
- Min and Avg values are likely naturally occurring and trustworthy.
- Max values warrant further scrutiny, especially if used for critical decision-making or compliance.

Individual Contributions

Worked on the design and layout of the slides
Ensured all key points and insights were presented effectively and engagingly

Conducted the Chi-Square Test and MAD Test on the dataset
Collaborated with Saniya to interpret graphs
Contributed to writing key insights and conclusions based on the analysis



Summarized statistical findings to ensure clarity and coherence
Designed and structured the PowerPoint presentation

Created graphs and visualizations for each section of the dataset
Assisted Vanshika in interpreting the visual data for meaningful insights

Thank you!