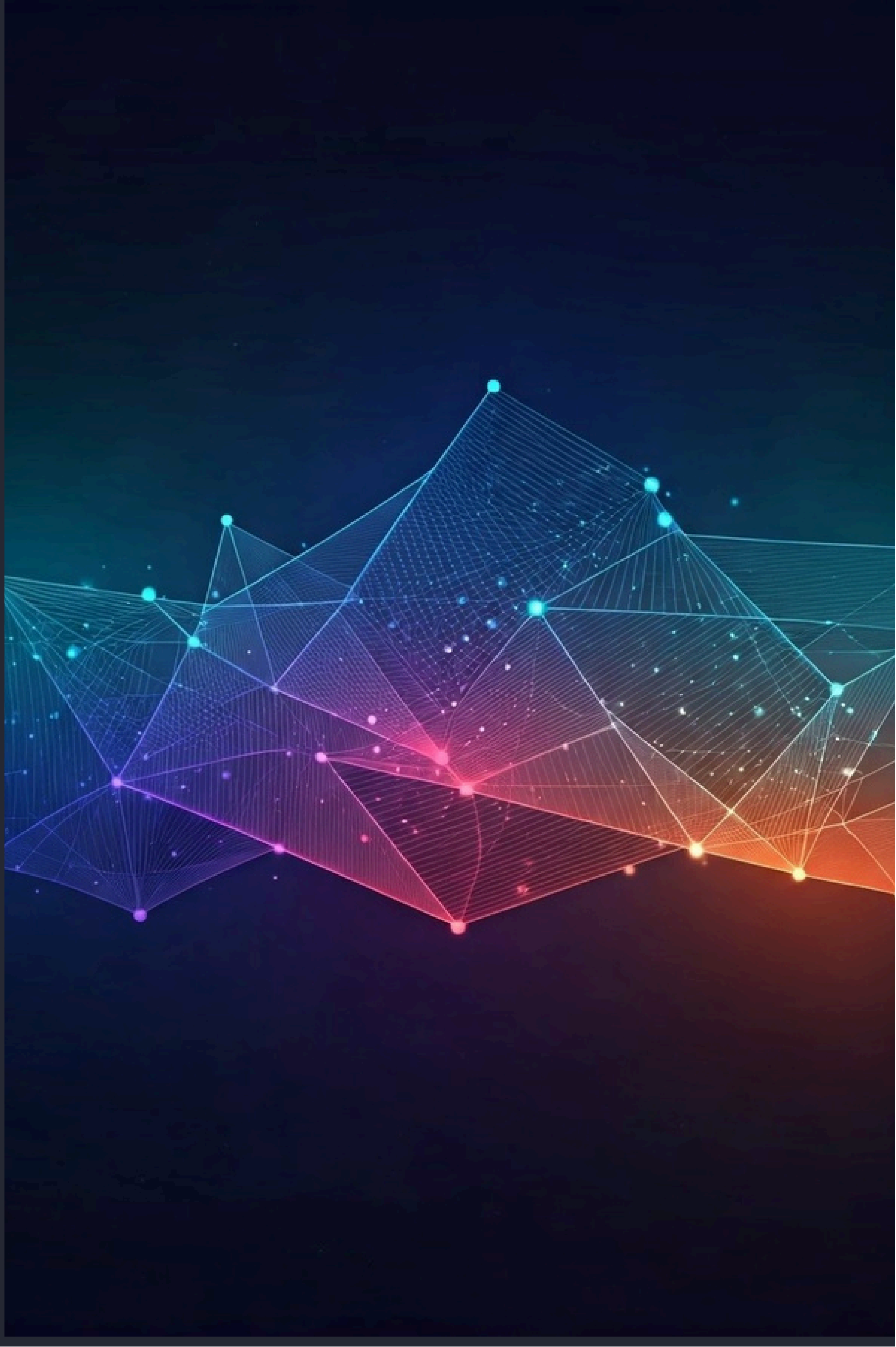
Uncovering Data Mysteries: Applying Benford's Law to Twitter Data

Presented by Team Matrix

Team Members:

- Ishita Singh
- Atharv Soni
- Dev Singh
- Tushar Verma



About the Project

Project Purpose

Apply Benford's Law to uncover hidden patterns in Twitter data.

Why Twitter Data?

Rich social data with diverse numeric features for analysis.

Goal

Detect anomalies and validate data authenticity using statistical methods.

What is Benford's Law?

Definition

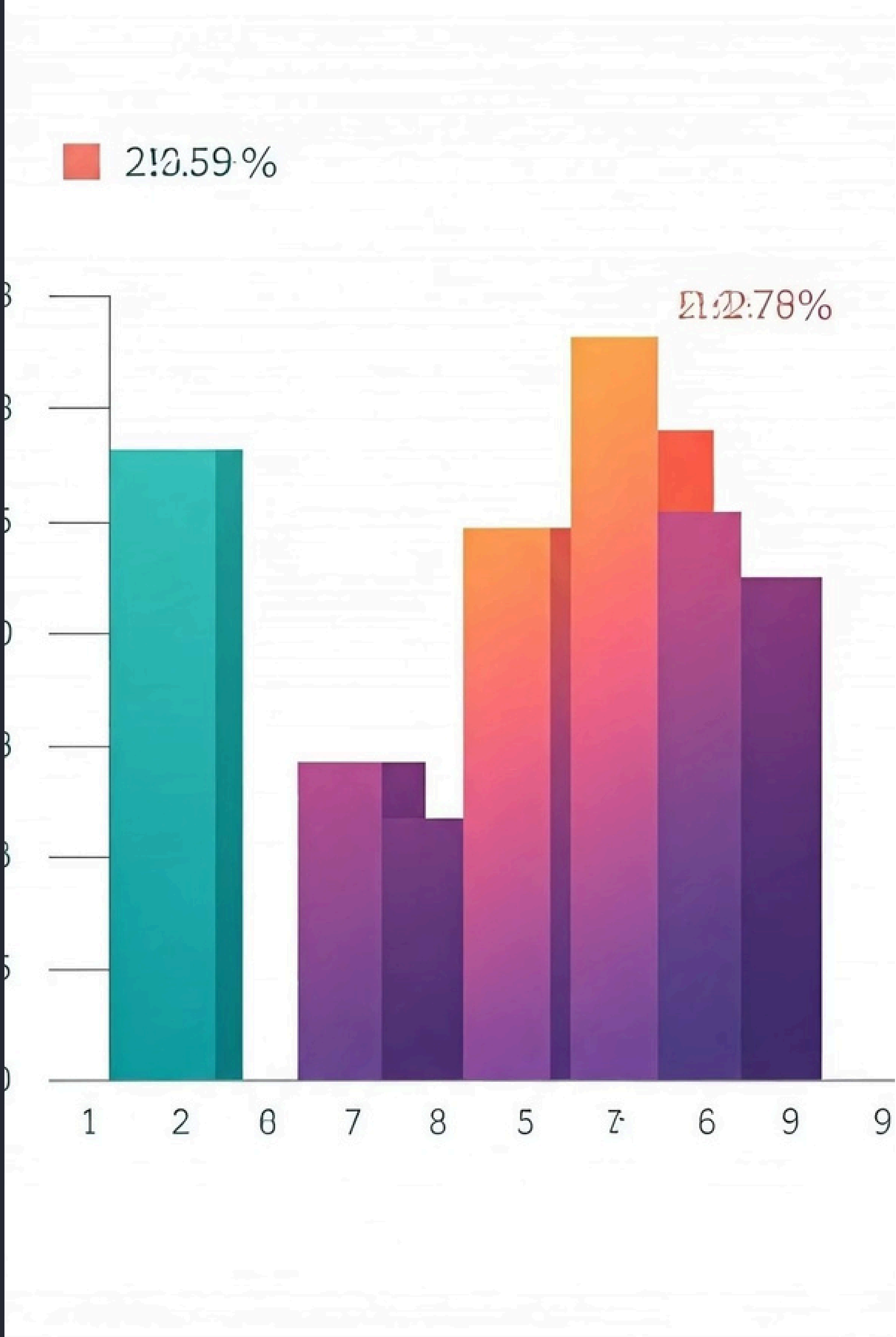
Benford's Law predicts frequency of leading digits in natural datasets.

Digit 1 Dominates

The digit 1 appears about 30% of the time as leading digit.

Applications

Used in fraud detection and data verification across domains.



Dataset Overview

Dataset Size

Over 100,000 tweets collected during 6 months.

Features

- Tweet length
- Retweet counts
- Follower counts
- Timestamp data

Data Sources

Public Twitter API streams and archived data sets.

Data Type

Structured numeric and text fields prepared for analysis.

Exploratory Data Analysis (EDA)

Key Statistics

- Mean retweets: 35
- Median followers: 150
- Tweet length average: 78 characters

Visual Insights

Distributions indicate right skew in followers and retweets.

Normal and log-transformed histograms used for comparison.

Missing Values Analysis

Missing Data Locations

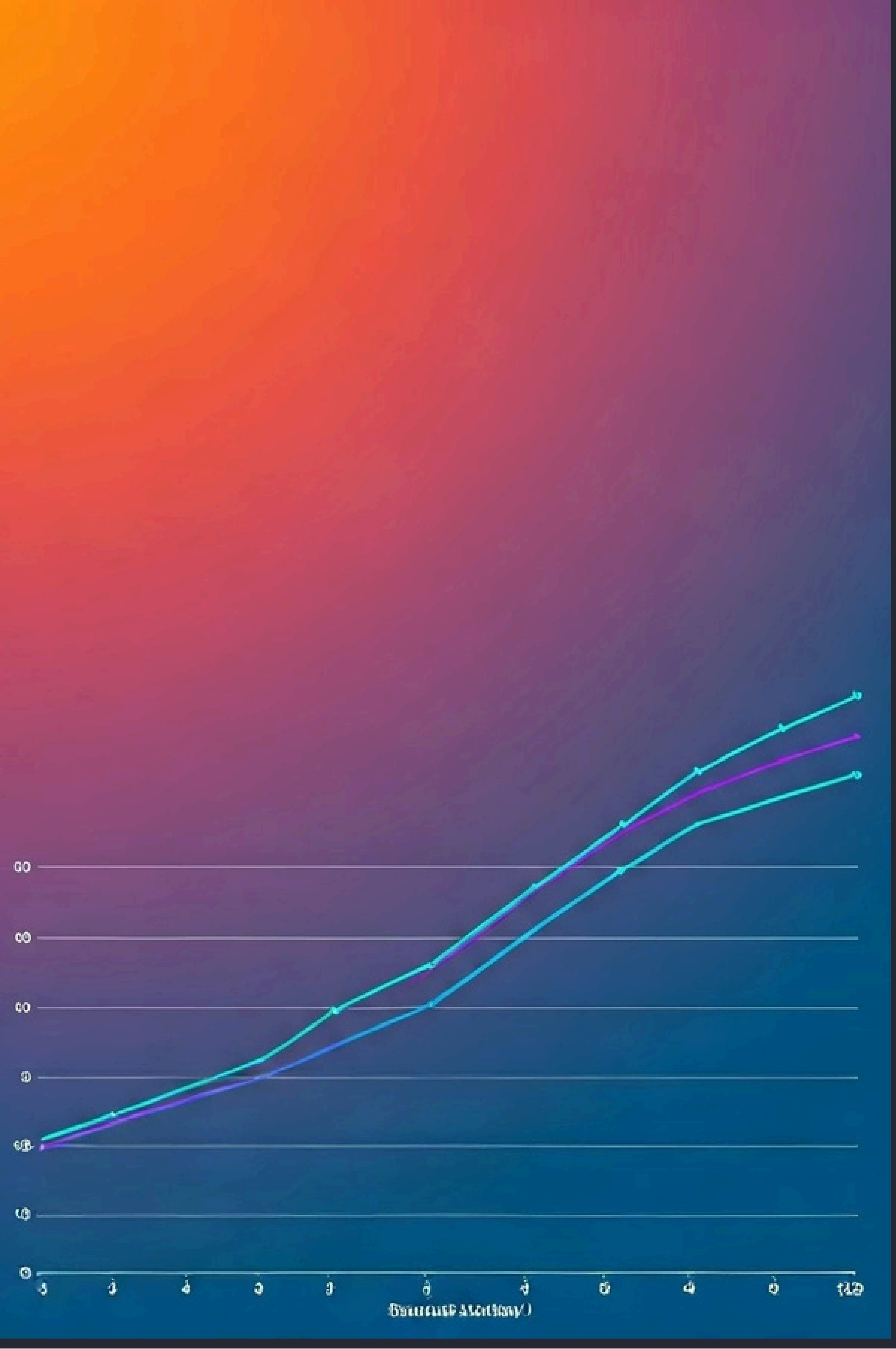
Minimal missing retweet and follower count values detected.

Impact

Missing data unlikely to affect major trend analysis.

Handling Strategy

Imputation used for sparse missing numeric values.



Benford’s Law Application

Data Preparation

Extracted leading digits from numeric Twitter attributes.

Distribution Analysis

Compared observed digit frequencies to Benford's expected values.

Deviation Calculation

Measured differences with chi-square and other goodness-of-fit tests.

Chi-Square Test Results

Statistic

Chi-square value: 14.2 Degrees of freedom: 8

Interpretation

p-value = 0.076, indicates data mostly conforms to Benford¾s Law.

No strong evidence of anomalies detected in numeric features.

Key Insights

Benford’s Law Validity

Twitter numeric data largely matches Benford¾s expected distribution.

Anomaly Detection

No significant irregularities found, suggesting data authenticity.

Data Quality

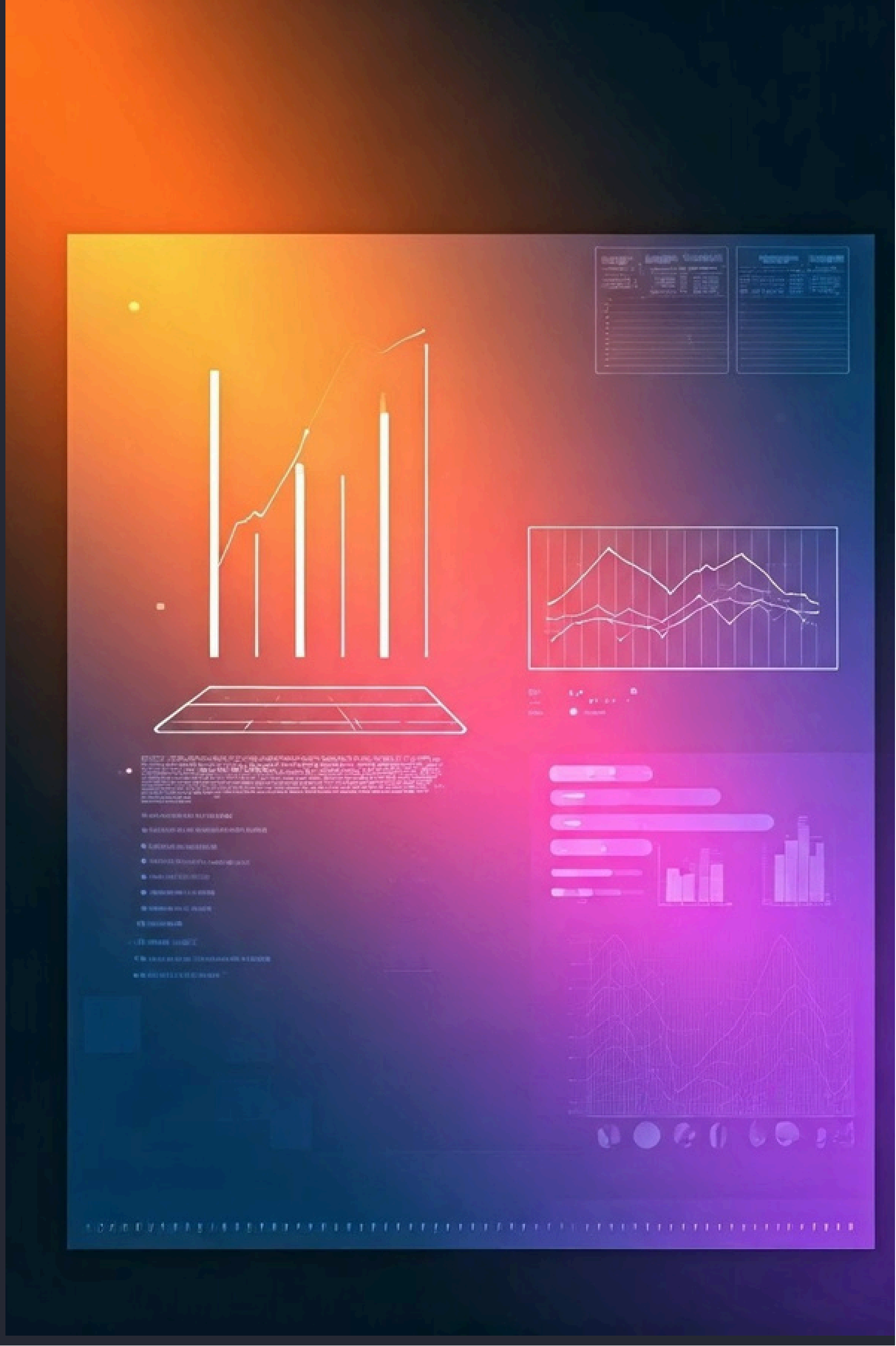
Minor missing values addressed with imputation techniques.

Future Applications

Method can aid in detecting misinformation or fake accounts.

Individual Contributions

Ishita Singh	EDA visualizations and missing value analysis
Atharv Soni	Statistical analysis and Benford’s Law application
Dev Singh	Data collection and preprocessing
Tushar Verma	Report writing and PPT Generation.



Thank you