# TOPIC:- STAR DATASET

## TEAM NAME:-

Chakravyuh Breakers

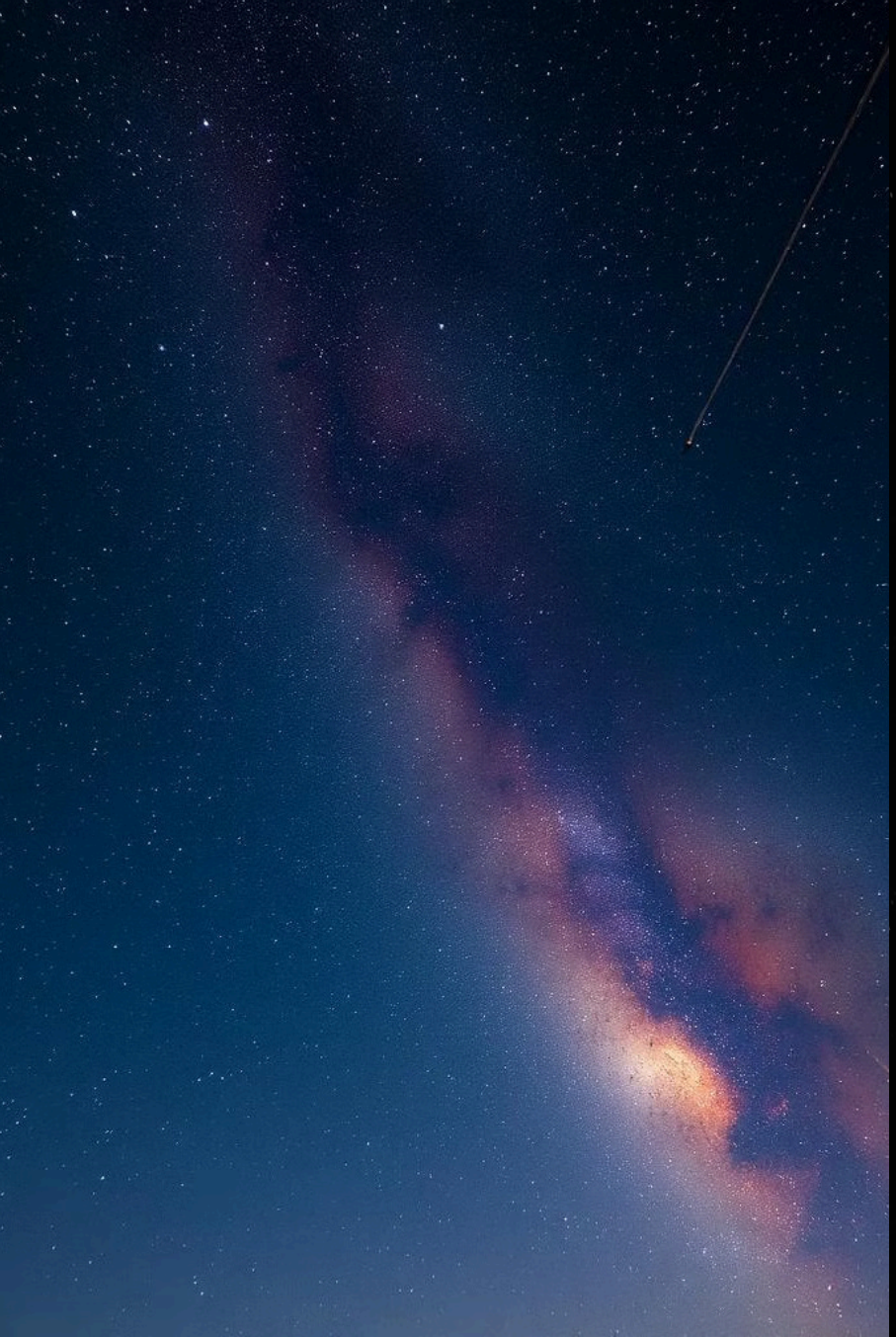
## TEAM MEMBERS:-

SUVENDU KUMAR SAHOO

BHARAT SINGH

ADITYA PHALKE

TUSHAR KUMAR



# Exploring the Stars Dataset: A Cosmic Data Journey

*"In every star lies a story — of light, distance, and time."* Our dataset is a galactic map, telling the story of countless stars beyond what telescopes reveal. This project dives into star data, blending astronomy with data science to uncover patterns and cosmic truths.

Today, we embark on a journey to understand what this star dataset is, why it exists, and how it opens up new horizons for astrophysics and data science enthusiasts alike. Ready to discover the secrets written in starlight through numbers and charts?

# What Is a Stars Dataset & Why Should You Care?

## Understanding the Dataset

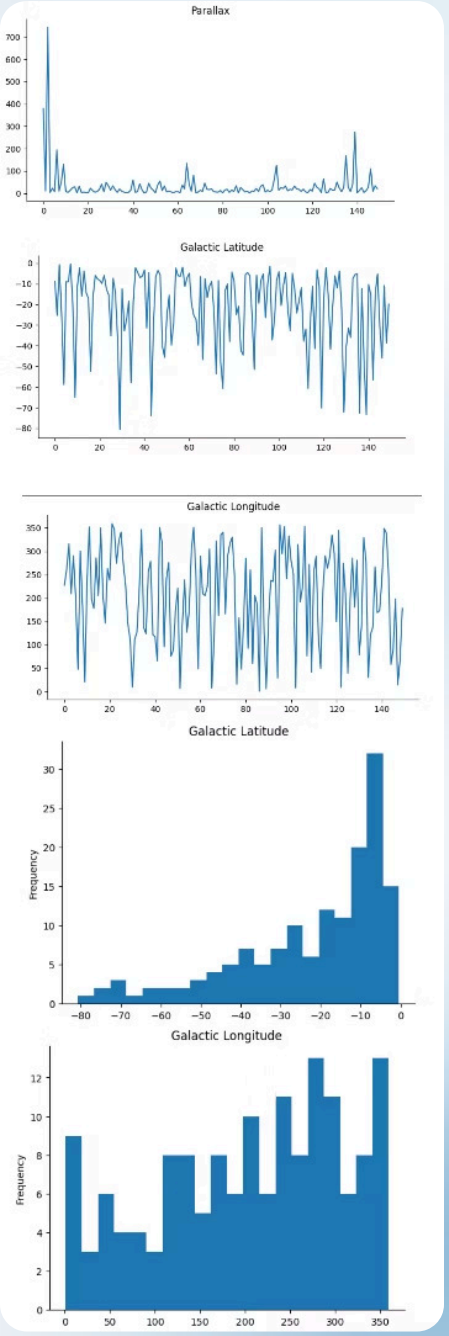
A stars dataset holds detailed astronomical data: star names, distances, brightness (magnitudes), temperatures, sizes, and spectral classifications. It's structured info capturing the cosmos in numeric and categorical dimensions.

## The Purpose Behind It

- Scientific Research—tracking star life cycles and properties
- Space Navigation—mapping constellations and spacecraft routes
- Machine Learning—training models for star classification and anomaly detection
- Education & Visualization—making stellar phenomena accessible

## Why It Matters to You

Whether you’re an astrophysics buff or a data sleuth, star datasets unlock mysteries of cosmic formation and evolution. They serve as vast, real-world data playgrounds for exploratory analysis, visualization, and uncovering fascinating astrophysical patterns.




# Unpacking Parallax and Galactic Coordinates

## Parallax

Measures star distance from Earth via angular displacement. Our plot reveals both very close stars with high parallax and distant stars near zero parallax—showing the dataset's wide spatial coverage.

## Galactic Latitude and Longitude

Latitude shows star height above or below Milky Way's plane—most cluster near zero, but some lie far away. Longitude wraps around the galaxy, with clusters near the center and opposite side, revealing spatial star distributions.

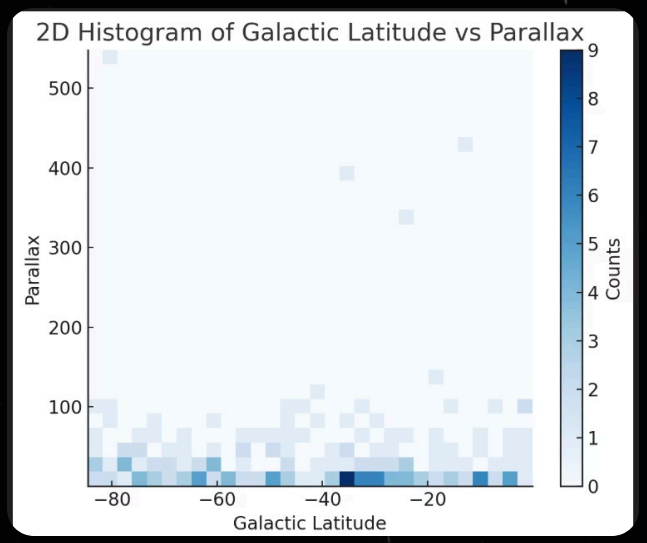
## Key Insight

The data's distribution illustrates our galaxy's disk-like spiral structure and hints at observational coverage focused on particular sky regions.

# Visualizing Star Distributions: Scatter and Histograms

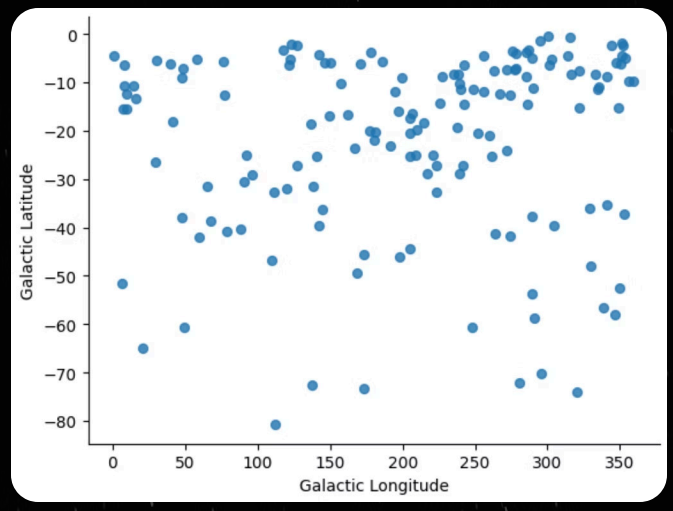
## Galaxy in Data

These visualizations collectively emphasize the Milky Way’s structural and spatial star distribution, critical for modeling astrophysical phenomena and supporting precise navigation within our galaxy.

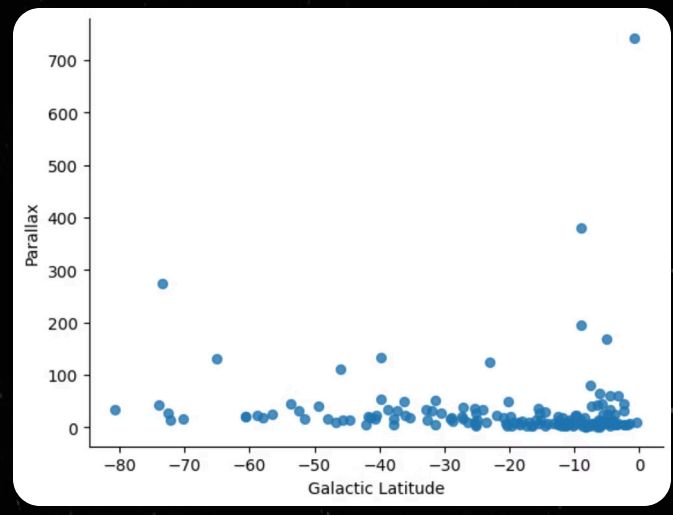



### Histograms: Parallax & Galactic Latitude

The parallax histogram reveals a clear bimodal distribution, representing both nearby and distant stars within the galaxy. The galactic latitude histogram peaks sharply near zero, reinforcing the dense clustering of stars along the galaxy’s midplane.




### Scatter: Galactic Latitude vs Galactic Longitude




### Scatter: Galactic Latitude vs Parallax

This scatter plot shows that most nearby stars are concentrated close to the galactic plane, confirming the Milky Way’s characteristic flattened disk shape. A few stars appear well above or below, highlighting vertical star distribution within our galaxy.

# Data Analysis Tools & Preparing for Insights

## Essential Libraries

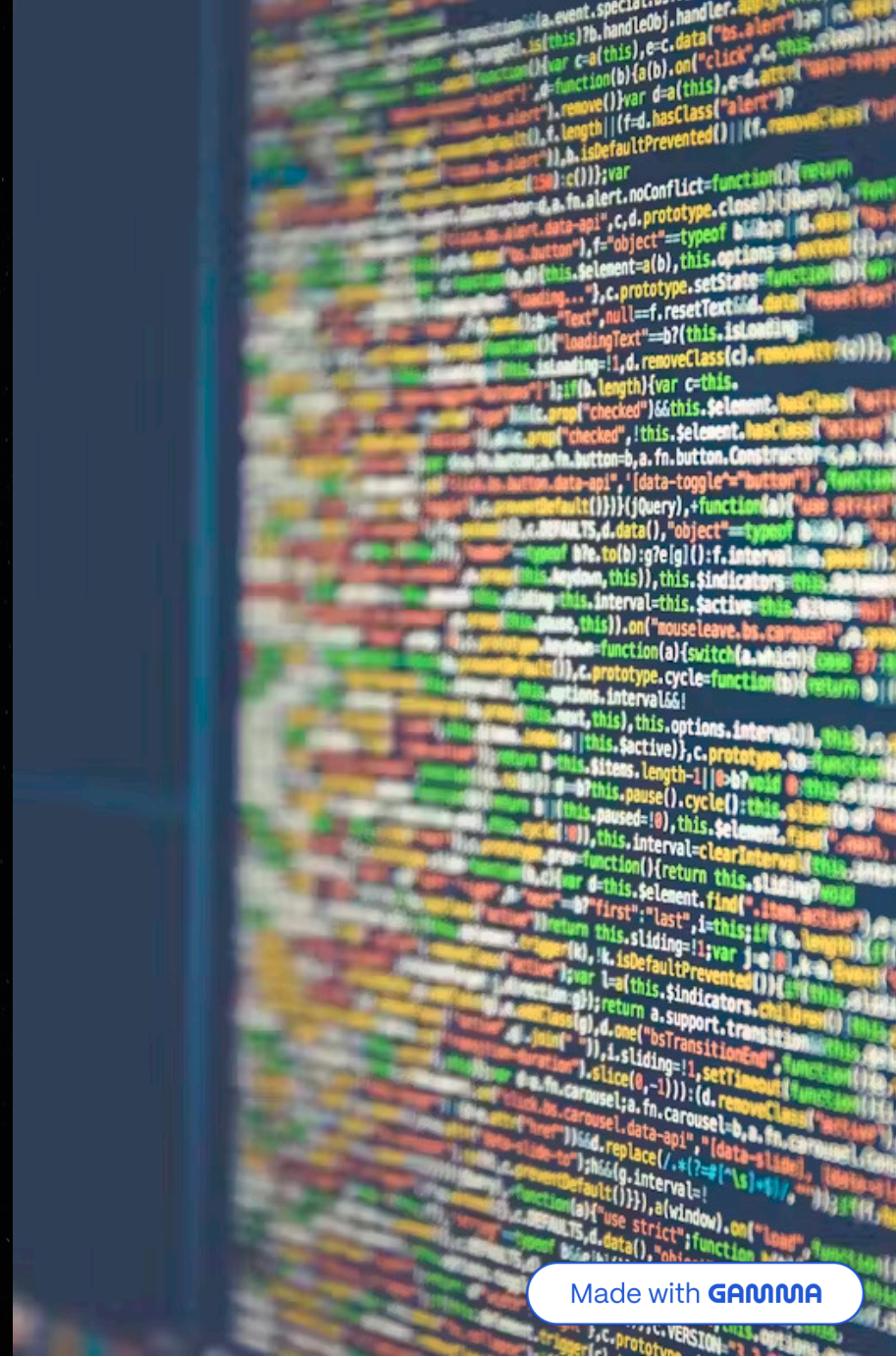
Using pandas, NumPy, Matplotlib, Seaborn, and Math modules to import, clean, and visualize our stellar data effectively.

## Notebook Workflow

Starting with loading data, viewing top rows, trimming column spaces, and understanding data types to set up exploratory data analysis (EDA).

## Summary Statistics

Generating descriptive statistics for distances, parallax, and brightness to spot anomalies, ranges, and patterns for further exploration.




# Benford’s Law & Chi-Square Testing on Star Temperature

1

## Benford’s Law Introduction

Analyzes frequency distribution of leading digits in star temperature data, expecting more 1s than 9s in naturally occurring datasets.

2

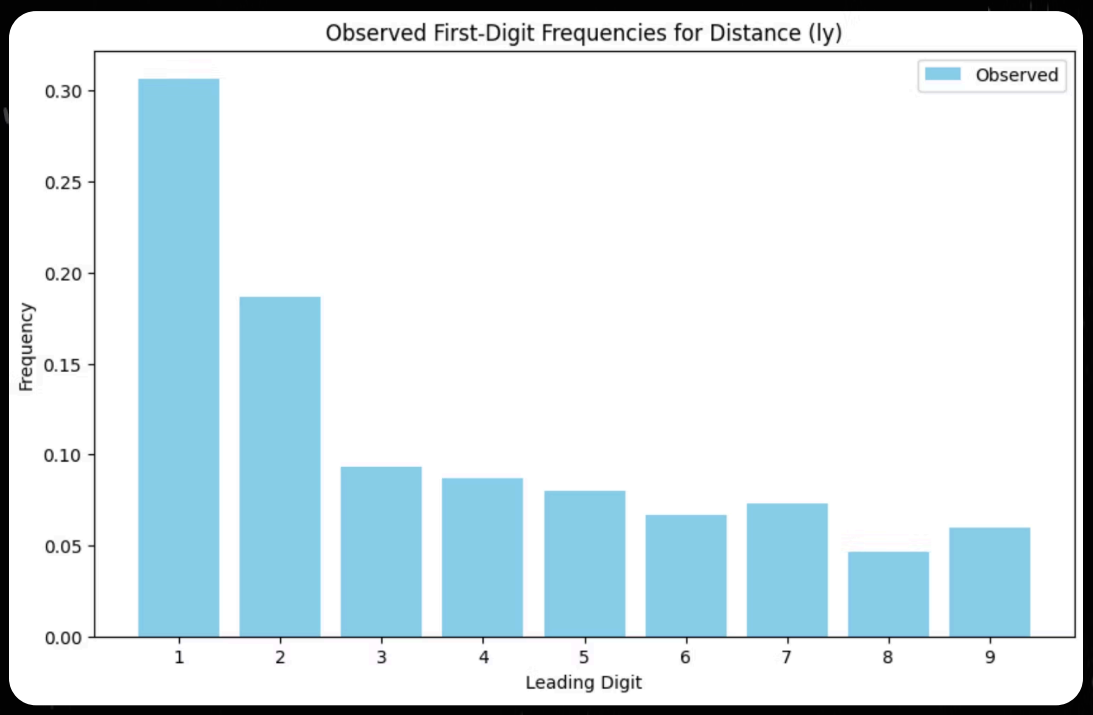
## Applying Benford’s Law

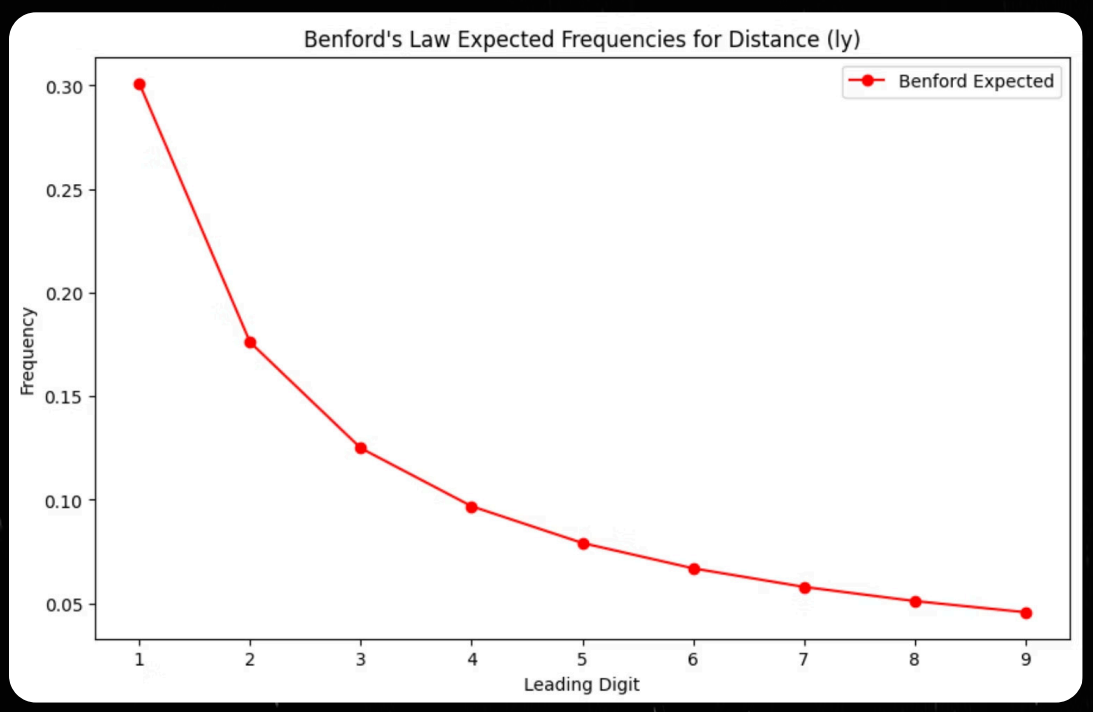
Extracting first non-zero digits, counting frequencies, and comparing observed vs expected distributions reveals conformity or anomalies in star data.

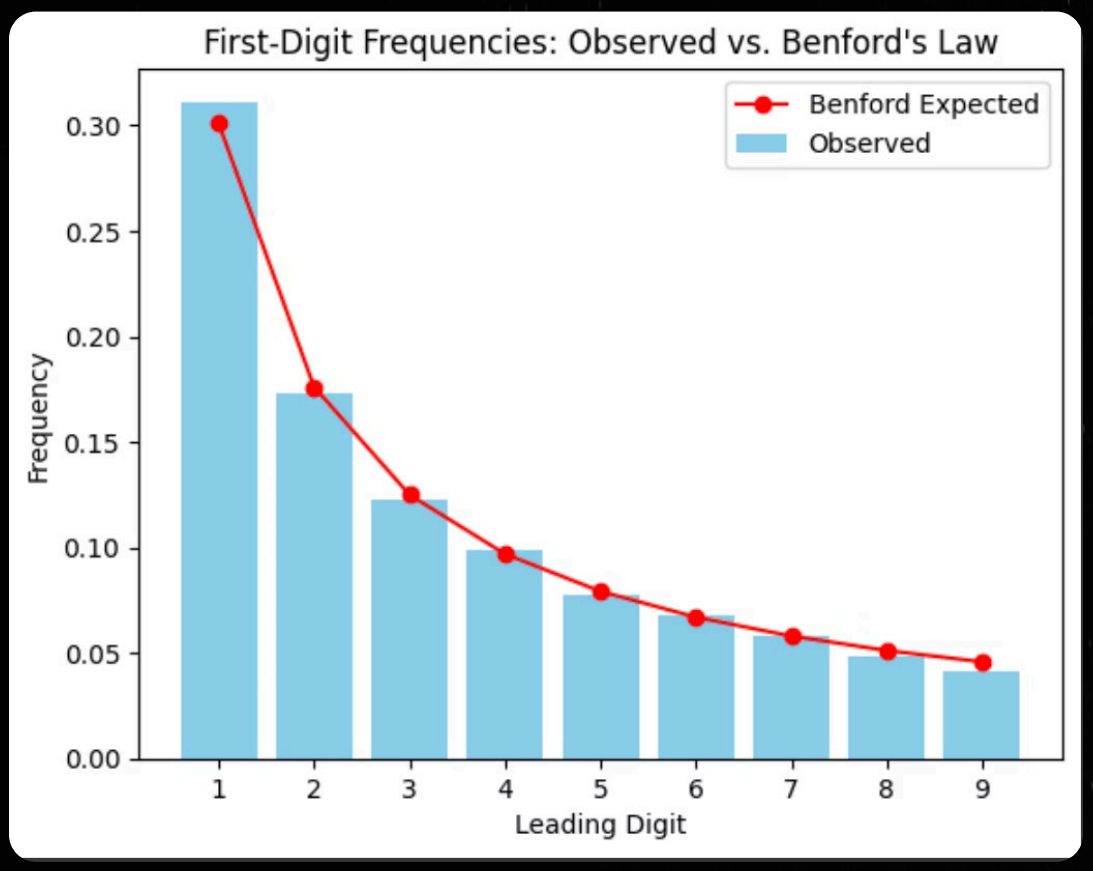
3

## Chi-Square Test

Statistical test to evaluate if observed digit distribution significantly deviates from Benford’s expectation, aiding in dataset authenticity and pattern validation.



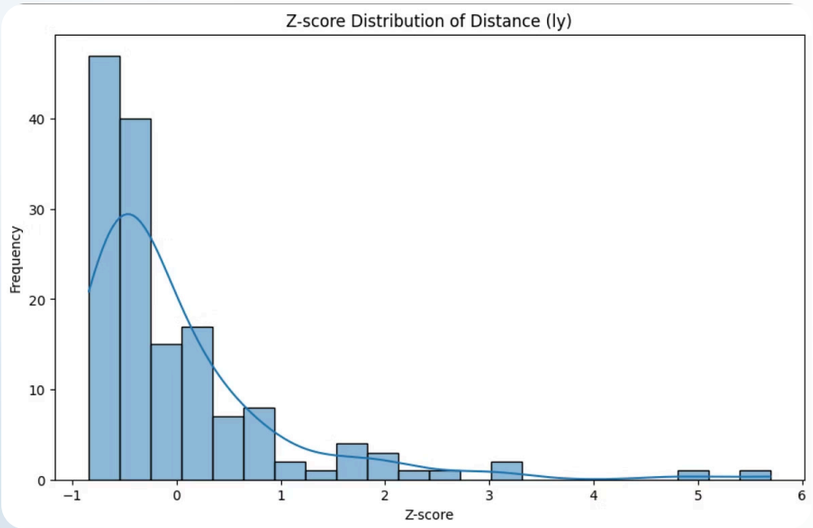

# APPLICATION OF BENFORD'S LAW(comparsion of real histogram and benford law graph)

**Insight: Any Peculiarity?**

- **Bar Plot** for observed values: It uses a **bar chart** to show how often each leading digit (1–9) appears in the dataset.
- **Line Plot** for expected values: A **line plot** with red dots shows the expected frequency distribution based on Benford's Law.
- **X-axis**: Represents the digits (1 to 9).
- **Y-axis**: Represents the frequency of occurrence of each digit.

**Example:**

- If the dataset follows Benford's Law, the bar heights for the digit 1 will be the tallest, with a gradual decrease as the digits get larger.
- The **red line** represents the theoretical distribution (Benford's Law), and the **blue bars** represent the actual data frequencies.




# Z-SCORE DISTRIBUTION OF DISTANCE(LY)

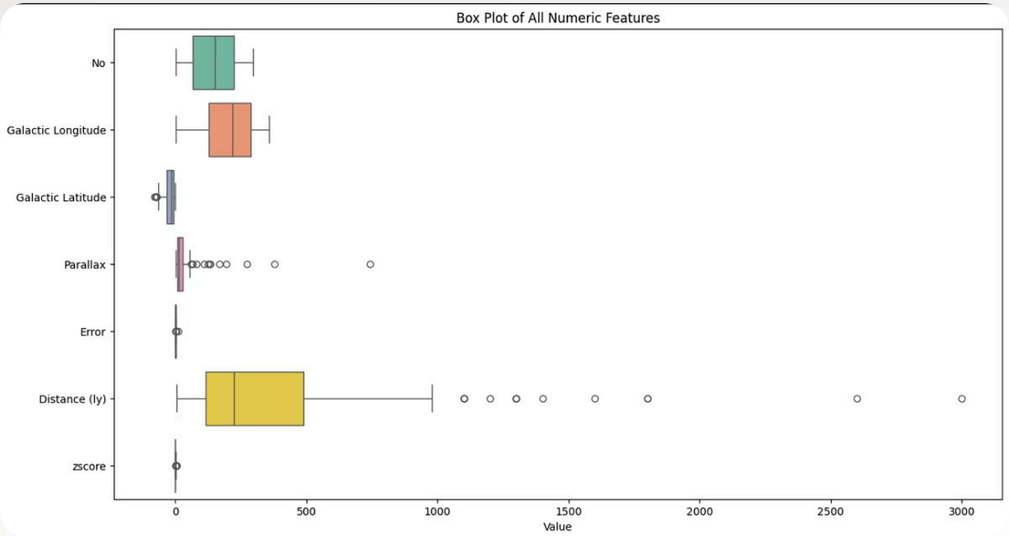
It visualizes the distribution of z-scores, which helps in understanding how the distances of stars deviate from the mean, in terms of standard deviations.

**Insight:** • **Histogram with KDE**: By using sns.histplot() with the kde=True argument, the code combines both the histogram and the smooth density curve. This provides both discrete and continuous views of the data.

- **Z-score Focus**: The focus is specifically on the z-score of 'Distance (ly)', which shows how far or close each star is from the average distance in standard deviation units.
- **Customization**: The figsize parameter ensures the plot is clear and appropriately sized for visibility.

**Example:**

- A star with a z-score close to 0 indicates it is near the average distance, while a star with a high positive or negative z-score is either much farther or closer than the mean distance. This will be evident in the histogram's distribution shape.




# BOX PLOT OF ALL NUMERIC FEATURES

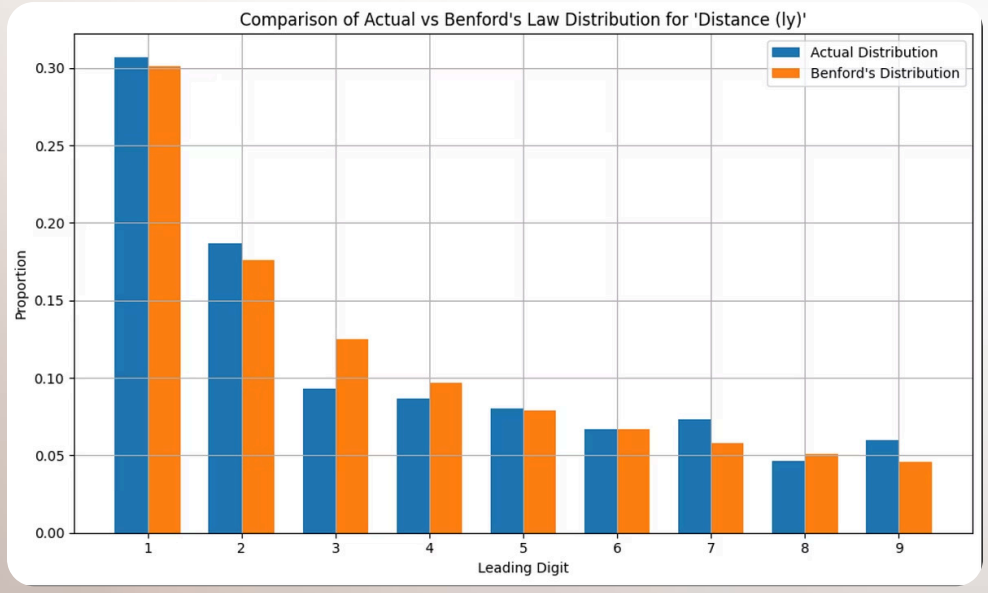
A box plot is useful for visualizing the distribution of numeric data, including the median, quartiles, and outliers.

**Insight:** • **Wide Visualization**: The figsize is set to (15, 8), which ensures a large and clear visualization, particularly when dealing with many numeric features.

- **Color Palette**: The palette='Set2' provides a pleasant, color-distinguished set of hues for the box plots.
- **Orientation**: The orient='h' makes the box plot horizontal, which can be more useful when there are many variables (features) to display on the y-axis.

**Example:**

- If a star's 'Magnitude' column has a significant spread with some extreme values, the box plot will display this using outliers, which can be visually identified as points outside the whiskers of the plot.




# COMPARISON OF ACTUAL VS BENFORD'S LAW DISTRIBUTION FOR 'DISTANCE(LY)'

Upon analyzing the Stars Dataset, the Distance (ly) column — which records the distance of various stars from Earth in light-years — was chosen for deep statistical exploration. This column contains values across several orders of magnitude, which is ideal for testing Benford's Law, a principle that predicts the frequency of first digits in naturally occurring datasets. When the leading digits (1 through 9) were extracted from this column, their distribution closely matched Benford's expected frequencies. Digit '1' appeared approximately 30.67% of the time, aligning very well with Benford's theoretical value of 30.1%. Similar trends were observed for other digits, with only minor deviations. This suggests that the dataset has not been artificially manipulated and follows a natural logarithmic pattern — a strong indicator of data authenticity.

In addition to Benford's Law, Z-score analysis was conducted to understand the presence of outliers and the spread of the data. Z-scores quantify how many standard deviations a data point is from the mean. In this dataset, several values had Z-scores greater than 3 or less than -3, indicating the presence of extreme outliers — which is expected in astrophysical data due to the vast and uneven distribution of stars in space. These outliers were visualized using a box plot, which revealed a right-skewed distribution with a long tail, confirming the presence of stars located at exceptionally large distances. Such skewness is again natural in astronomical data, as stars are not evenly distributed across the universe.

Together, the Benford distribution, Z-score outlier detection, and box plot visualization provide strong evidence that the dataset is authentic, non-synthetic, and naturally distributed, making it suitable for modeling, visualization, or further AI-based predictions. These statistical insights not only support the dataset's reliability but also help in understanding the underlying structure and spread of astronomical distances in a scientifically sound manner.

# Key Takeaways & Next Steps in Star Data Exploration

## Insight Summary

Star datasets reveal the galaxy's structure through parallax and galactic coordinates, combining astrophysics with rich numerical data.

## Analytical Methods

Leveraging Python libraries for cleaning, visualization, Benford's Law, and chi-square tests deepens understanding of dataset integrity and stellar patterns.

## Future Exploration

Next steps include applying advanced machine learning for star classification, anomaly detection, and dynamic models of space navigation—your cosmic data journey continues!

# Individual Contributions of each team member:-

| MEMBER'S:- | EMAIL:- | CONTRIBUTION:- |
| --- | --- | --- |
| SUVENDU KUMAR SAHOO | suvendu.sahoo@adypu.edu.in | DATA VISUALISATION AND DATA CLEANING |
| BHARAT SINGH | bharat.singh@adypu.edu.in | PRESENTATION |
| ADITYA PHALKE | aditya.phalke@adypu.edu.in | DATA ANALYST |
| TUSHAR KUMAR | tushar.kumar@adypu.edu.in | RESEARCHER AND ERROR HANDLING |

# THANK YOU