Benford's Law Analysis on COVID-19 Data

A data-driven approach to detect patterns in reported total COVID-19 cases.

Team Name:
The Data Detectives

Team Members:
Mahi Sawner | Divyanjali Gopisetty | Sai Sri Spruha Perumalla | Gopi Raman Thakur

What is Benford's Law?

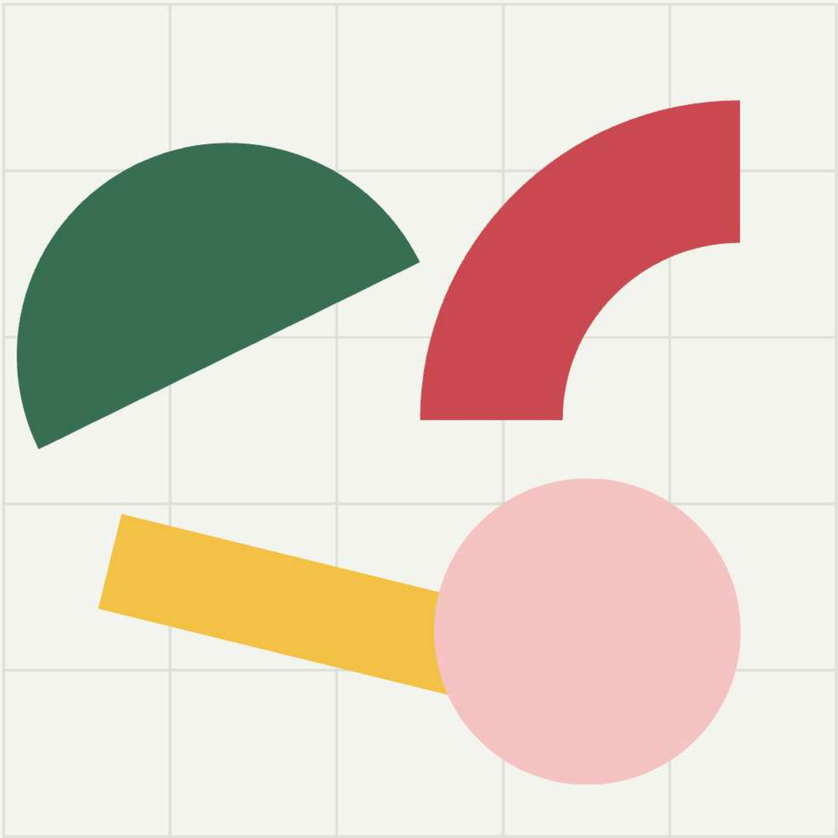
- Benford’s Law is a mathematical rule that predicts how often each digit (1 through 9) appears as the first digit in many naturally occurring datasets.

Key Idea:

- The number 1 appears as the leading digit about 30.1% of the time.
- The number 2 appears about 17.6%, and the probability keeps decreasing up to 9.

Formula:

- $P(d)=\log10(1+1/d)$
- Where d is the first digit (1 to 9)



Dataset Overview

Dataset Source:

- Global COVID-19 statistics containing “Total Cases” for different regions or countries.

Step 1: Uploading the Data

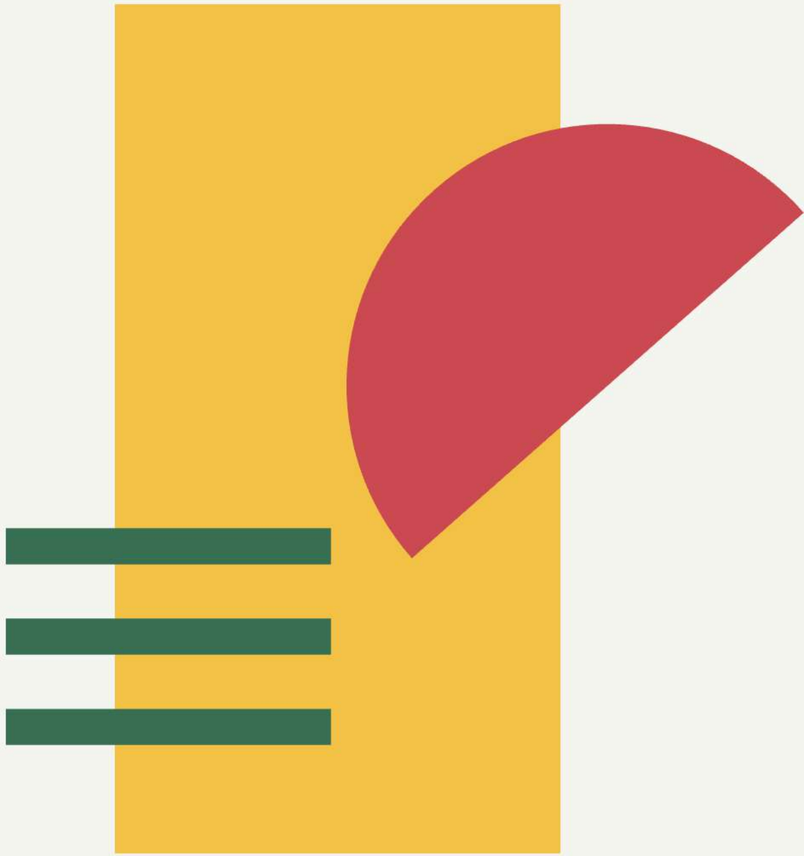
- Used files.upload() in Google Colab to upload the dataset.

Step 2: Loading the Data

- Loaded dataset using Pandas to enable further processing.

Purpose:

- Focused on analyzing the "Total Cases" column for conformity to Benford’s Law.



Data Cleaning and Preprocessing

Objective:

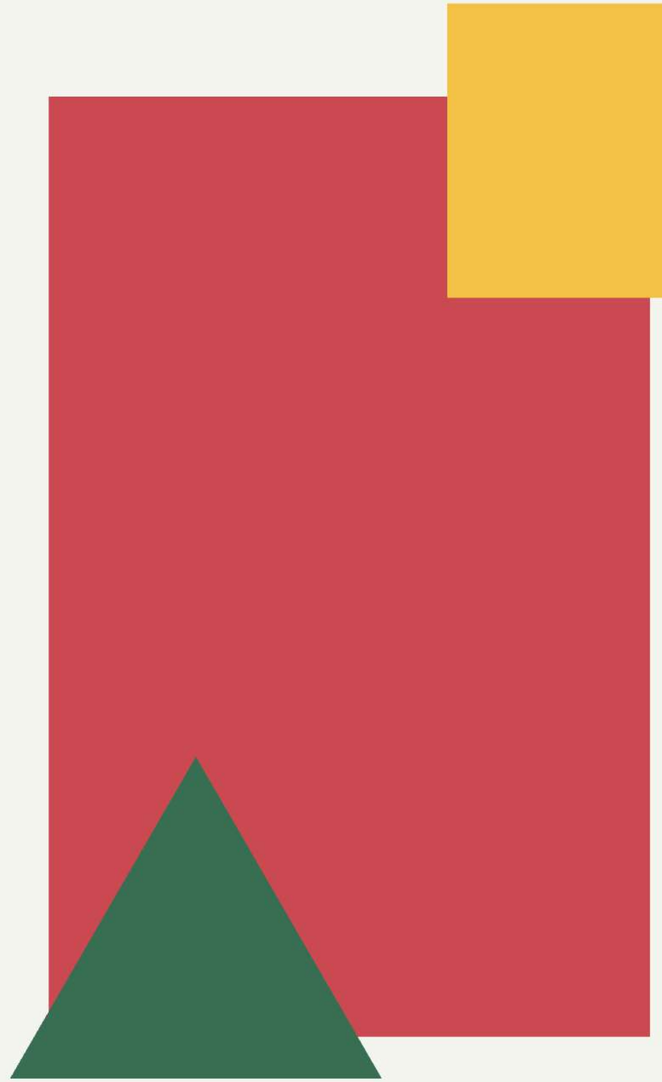
- Ensure “Total Cases” column is in the correct numeric format.

Steps Taken:

- Removed commas from the “Total Cases” column.
- Filtered out non-numeric values.
- Converted values to integer type for analysis.

Code Snippet:

```
df['Total Cases'] = df['Total Cases'].str.replace(',', '').astype(str)
df = df[df['Total Cases'].str.isnumeric()]
df['Total Cases'] = df['Total Cases'].astype(int)
```



Extracting Leading Digit & Calculating Frequencies

Step 4: Extract First Digit

- Extracted the first digit of each "Total Cases" entry using:
- df['First Digit'] = df['Total Cases'].astype(str).str[0].astype(int)

Step 5: Calculate Frequencies

- Counted and normalized how often each digit appears:
- observed_count = df['FirstDigit'].value_counts(normalize=True).sort_index()

Comparison:

- Computed using logarithmic formula from Benford's Law.

Visualization of Results

Tool Used:

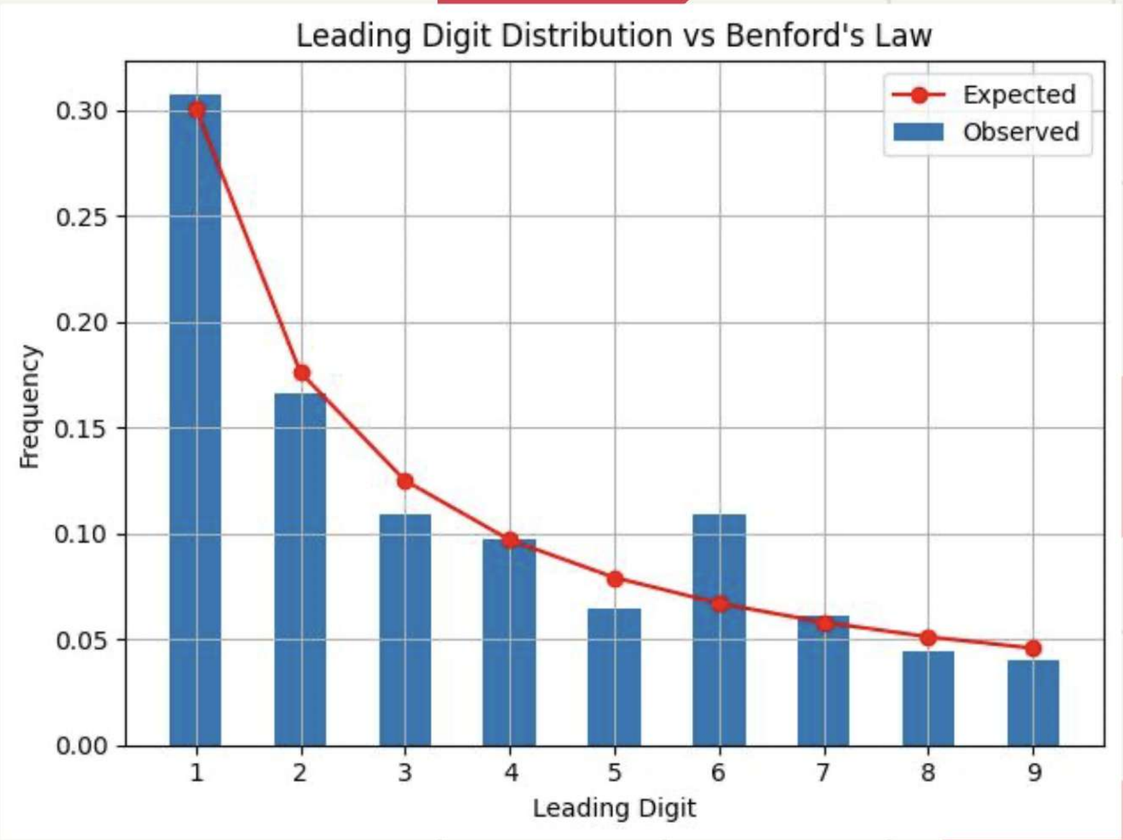
- Matplotlib

Chart:

- Bar chart for observed vs. expected frequency

Insights:

- Visual comparison makes it easier to spot any deviation from Benford’s Law.



Conclusion & Insights

Key Findings:

- COVID-19 "Total Cases" data showed [insert pattern: e.g., good, moderate, or poor] alignment with Benford’s Law.
- Deviations might indicate reporting inconsistencies or regional anomalies.

Limitations:

- Dataset source and completeness can affect accuracy.
- Benford's Law applies best to large, non-truncated datasets.

Next Steps:

- Apply the same method to other COVID metrics (deaths, recoveries).
- Analyze by country or time-series trends.

Individual Contributions

Sai Sri Spruha Perumalla

Uploaded and cleaned dataset, removed commas, converted to numeric

Gopi Raman Thakur

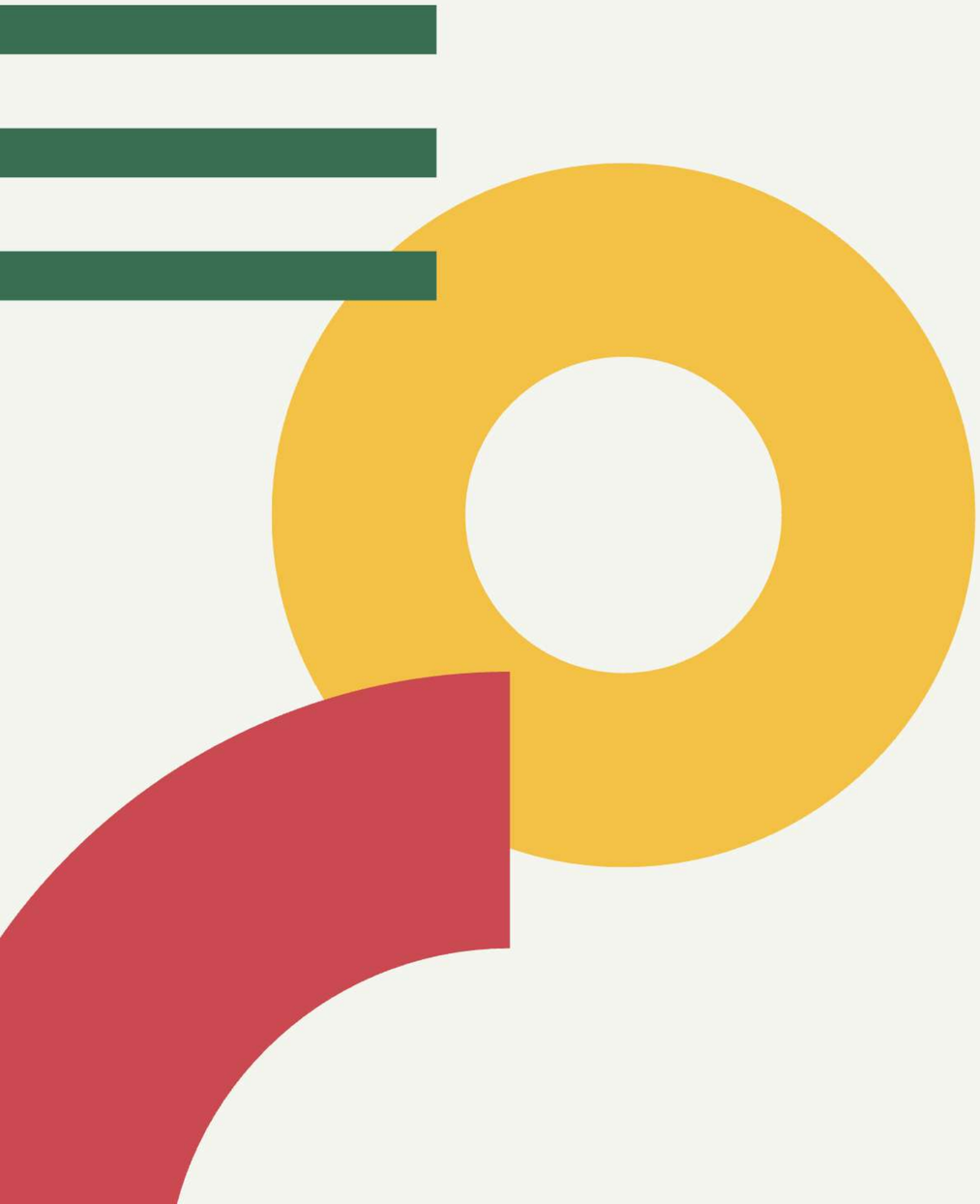
Extracted leading digit, added column for first digit, gave final insights

Mahi Sawner

Computed observed and expected frequencies using Benford's formula

Divyanjali Gopisetty

Visualized results using Matplotlib, styled and designed the plot



Thank You