Descriptive and Visual Analysis of GitHub Trending Repositories

Name: Aakash Desai

SRN: PES1UG24CS006

Section: A(3rd Semester)

Department: Computer Science and Engineering(CSE)

1. Descriptive Statistics Results:

```
2]: print("\nDescriptive Statistics for Stars:")
    print(df['stars'].describe())
```

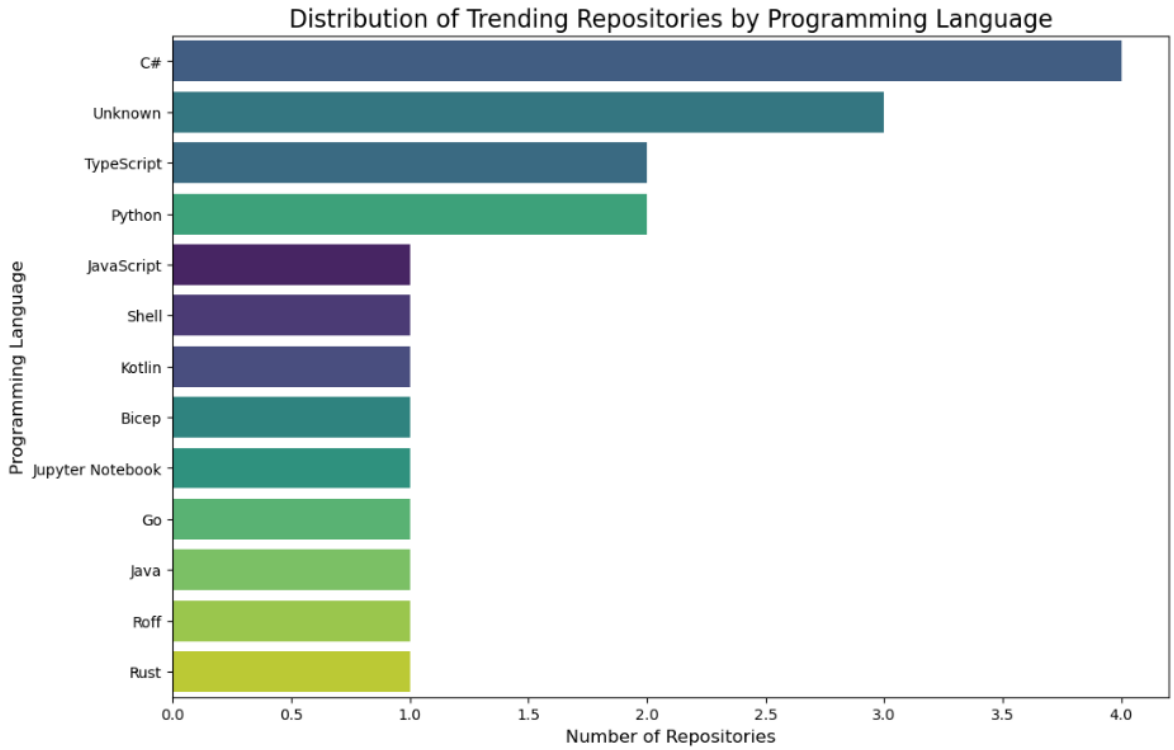
```
Descriptive Statistics for Stars:
count        20.000000
mean      20336.050000
std       27354.761399
min         803.000000
25%        5704.750000
50%       11845.000000
75%       19499.000000
max      122871.000000
Name: stars, dtype: float64
```

Descriptive Data that can be concluded include:

- There are 20 GitHub repositories present on the GitHub Trending page as of 03-09-2025.
- The mean of the data, measured in terms of stars, was approximately 20,000(20336.05) with a large standard deviation of approximately 27,000(27354.761399).
- The no of stars provided for the repositories on the GitHub Trending page as of 03-09-2025 range from 803 to 122871.Thus the range is found out to be 122068.
  The median(Q2)is approximately 11,900(11845) stating that half the repositories have star rating of under this value while the remaining half are over it. The corresponding logics can be further extended to Q1, Q3(with 25% of repositories have star rating less than the value specified against Q2, 75% of the repositories have star rating less than the value specified against Q3)

2. Visualization Results

Figure 1: Distribution of Trending Repositories by Programming Language



The above bar graph shows which languages dominate trending repos

Figure 2: Distribution of Stars by Programming Language (Log Scale)

The below boxplot shows spread, outliers, and comparison between languages.

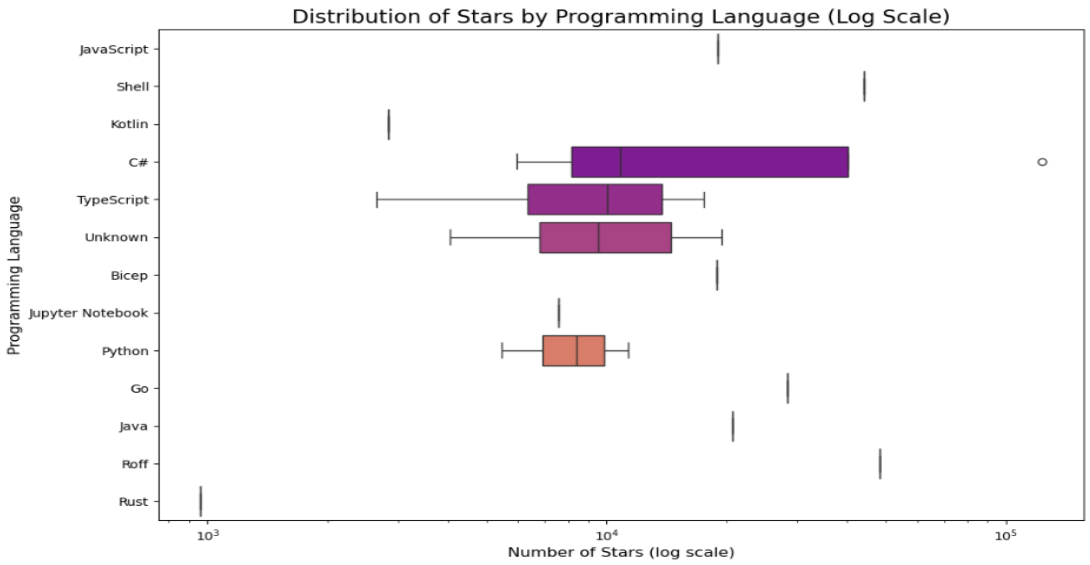
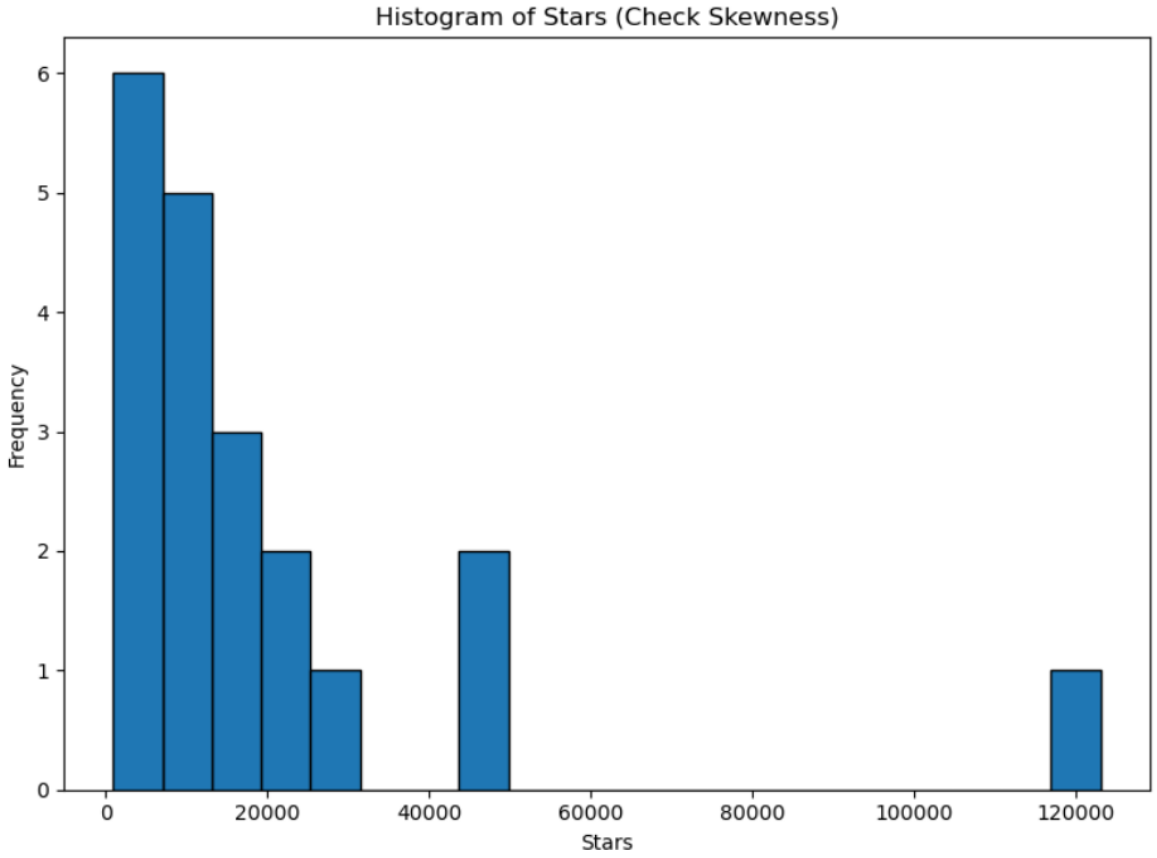


Figure 3: Histogram of Stars (Check Skewness)



Some observations

- A long tail exists: several languages appear only once or twice.
- The dominance of a few languages suggests community preference and ecosystem maturity on GitHub.
- The boxplot reveals that most repositories have stars concentrated in a lower range, but there are outliers with exceptionally high stars.
- Outliers represent **exceptionally popular projects**, which skew averages upwards.
- The histogram confirms a right-skewed distribution: many repositories have relatively few stars, while only a handful have very high stars.

- The frequency drops steeply after a certain star threshold, reinforcing that GitHub trending is dominated by a small number of “superstar” repositories.

3. Inferences Drawn

1. **Language Popularity:**
    - Certain languages like **Python, JavaScript, and TypeScript** appear multiple times, suggesting they dominate the trending ecosystem.
    - Niche languages (e.g., **Rust, Bicep**) appear less frequently.

2. **Skewed Star Distribution:**
    - The **mean stars (~20k)** is much higher than the **median (~11.8k)** → showing that a few repositories with very high stars are pulling the average up.
    - This indicates **positive skew/right skew** in the data.

3. **Presence of Outliers:**
    - Boxplots reveal several outliers with **exceptionally high star counts** (>100k).
    - These represent massively popular projects (frameworks, tools, or libraries).

4. **Emerging / Unknown Projects:**
    - A few repositories are tagged with **“Unknown”** language → these could be multi-language repos or those where GitHub doesn’t detect a primary language.
    - Indicates that not all trending projects are tied strongly to one programming language.

5. **Log Scale Insights:**
    - Using a log scale made it clear that most repos are in the **1k–20k star range**, while only a few break into **100k+ stars**.
    - Without the log scale, these differences would have been hidden.