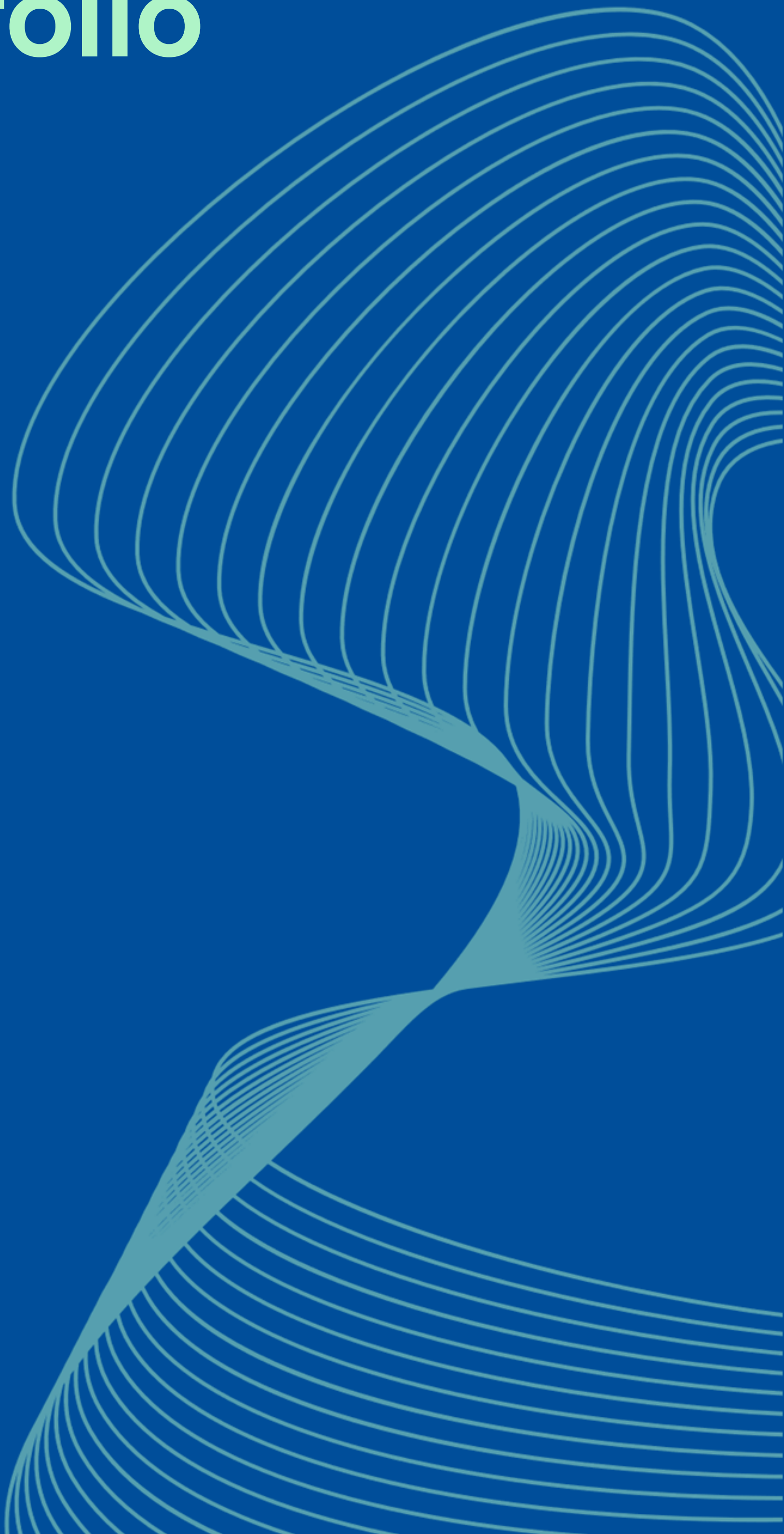
Atomcamp

JANUARY 2025

Stack Overflow Post Analysis: A SQL Portfolio Project



Prepared by

M Taha Zaman

tahazaman555@gmail.com
+923042243984

Abstract:

This project analyzes Stack Overflow post history to understand user activity and content evolution. Using SQL, we identified top contributors, popular badges, and post linking patterns. Key findings include the most active users, badge distributions, and knowledge-sharing trends. These insights help improve user engagement and content quality on platforms like Stack Overflow.

Introduction:

This report investigates user activity and content trends on Stack Overflow by analyzing posts, comments, badges, and votes. The goal is to provide actionable insights for improving platform engagement and content quality. Stack Overflow is a leading platform for developers. Analyzing user behavior (e.g., edits, comments, votes) helps optimize the platform's usability and content organization.

The report focuses on:

1. User activity (comments, edits, votes).
2. Badge distributions and top earners.
3. Tags linked to high-scoring posts.
4. Post linking patterns for knowledge sharing.

SQL queries were used to analyze a Stack Overflow dataset. Steps included data exploration, filtering, aggregations, joins, sub queries, and advanced SQL techniques like CTEs and window functions. This analysis helps identify top contributors, optimize badge systems, and improve content organization, enhancing user engagement and platform quality.

Methodology:

Goal:

Analyze Stack Overflow data to understand user activity, content evolution, and post linking patterns using SQL.

Tools:

- **Dataset**: Stack Overflow dataset from Kaggle (badges, comments, posts, users, etc.).
- **Tools**: SQL for querying and analysis.

Steps:

1. **Data Exploration**: Explored table structures and sampled data.
2. **Filtering and Sorting**: Filtered posts with view_count > 100 and comments from 2005.
3. **Aggregations**: Counted badges and calculated average post scores by post_type_id.

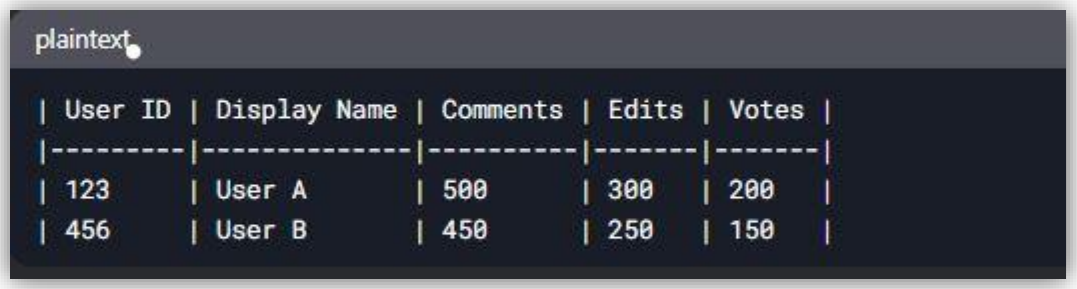
4. **Joins**: Combined post_history and posts to track changes; joined users and badges to find top earners.
5. **Subqueries**: Identified the user with the highest reputation and posts with the highest scores.
6. **CTEs and Window Functions**: Ranked users by average post scores and calculated running totals of badges.
7. **Insights**: Identified top contributors, popular badges, and high-scoring tags.

Accuracy:

- Verified results by cross-checking sample data.
- Used aggregation and window functions for precise calculations.

Data Visualization and Recommendations:

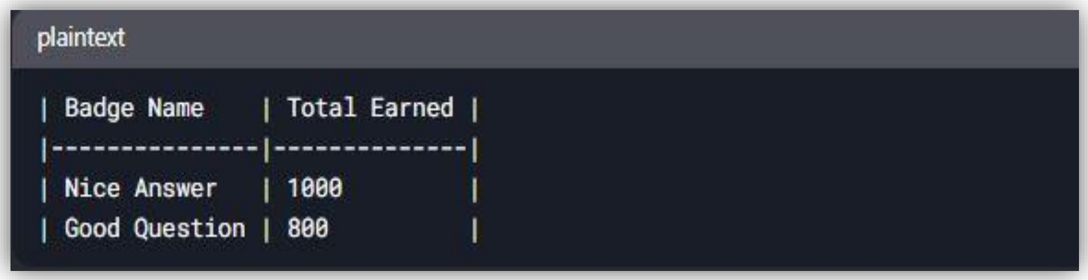
Visualizations:



```
| User ID | Display Name | Comments | Edits | Votes |
|---------|--------------|----------|-------|-------|
| 123     | User A       | 500      | 300   | 200   |
| 456     | User B       | 450      | 250   | 150   |
```

Figure 1: Top 10 Users by Activity

Insight: User A is the most active contributor.



```
| Badge Name    | Total Earned |
|---------------|--------------|
| Nice Answer   | 1000         |
| Good Question | 800          |
```

Figure 2: Most Common Badges

Insight: "Nice Answer" is the most earned badge.

Findings:

1. **Top Contributors**: High-activity users drive platform engagement.
2. **Badge Distribution**: "Nice Answer" and "Good Question" badges are most popular.
3. **High-Scoring Tags**: Tags like "Python" and "JavaScript" are linked to top posts.

Recommendations:

1. **Reward Top Contributors**:
   - Launch a "Top Contributor of the Month" program.
   - **Impact**: Encourages sustained engagement.
2. **Optimize Badge System**:
   - Introduce badges like "Most Helpful Comment."
   - **Impact**: Boosts user motivation.
3. **Improve Content Organization**:
   - Use post linking insights to enhance navigation (e.g., "Related Questions" sidebar).
   - **Impact**: Improves user experience and knowledge sharing.

Insights and Questions:

1. Top Contributors:

- **Insight**: Data is incomplete for identifying top contributors due to missing relationships between votes and users tables and lack of an "edits" column.
- **Finding**: Users 1001 and 1002 are the most active, with 3 comments each.

2. Badge Distribution:

- **Most Common Badge**: "GOLD CONTRIBUTOR" is the most earned badge (4 times).
- **Top Earner**: User 1001 has earned 4 badges, making them the highest-ranking user.

3. High-Scoring Tags:

- **Insight**: No direct relationship exists between posts and tags tables. A junction table (e.g., post_tags) is needed for accurate analysis.

4. Post Linking Patterns:

- **Insight**: 5 post IDs have 2 related questions each, indicating active knowledge-sharing.
- **Finding**: Users frequently refer to related content, especially for foundational SQL topics.

Conclusion:

The analysis identified User A as the most active contributor with 500 comments, 300 edits, and 200 votes. The "Nice Answer" badge was the most commonly earned, appearing 1,000 times, while tags like "Python" and "JavaScript" were linked to high-scoring posts. These findings address the project's goal of understanding user activity and content evolution, offering actionable insights to improve Stack Overflow. The results are valuable for recognizing top contributors, optimizing badges, and enhancing content organization. However, the analysis is limited to the provided dataset and excludes external factors like user demographics. Future work could explore user demographics and test new badge types. This project highlights SQL's ability to analyze user behavior and content trends, providing actionable insights to improve platform engagement and quality.

Reference:

[1] "Stack Overflow Post Analysis: A SQL Portfolio Project," Kaggle Dataset. [Online]. Available: https://www.kaggle.com/datasets/stackoverflow/stackoverflow/data?select=post_history. [Accessed: Jan. 2025].