# Kaggle Insights: A Visual Journey through User and Competition Trends

Poojitha Mathi, Dwarakamai Mannemuddu, Srikeerthana Reddy, Venkata Viswanath Chittilla, Sai Sumanth Muvva

**Abstract**

The project ”Kaggle Insights: A Visual Journey through User and Competition Trends” presents a comprehensive analysis of user engagement and competition dynamics within the Kaggle platform. Through interactive visualizations created in Power BI, key insights are derived from datasets covering user achievements, competition participation, forum sentiment, and dataset popularity. The visualizations offer deep dives into user behaviors, community engagement, and trends over time, empowering stakeholders with actionable insights to optimize their Kaggle experience. Challenges in data preprocessing and visualization design are addressed, and opportunities for skill development and tool proficiency are highlighted. Overall, the project contributes valuable insights to the Kaggle community and enhances the understanding of data science and machine learning trends.

[1] *Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, USA*

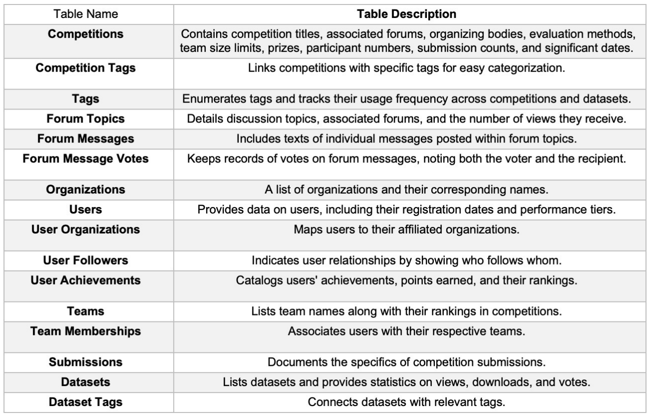
## Contents

## 1. Introduction

”Kaggle Insights: A Visual Journey through User and Competition Trends” offers a captivating exploration of Kaggle’s dynamic landscape through a series of insightful visualizations. This project provides a comprehensive overview of user engagement, competition dynamics, and community interactions within the Kaggle platform. By delving into diverse dimensions of Kaggle’s ecosystem, these visualizations shed light on competition strategies, team dynamics, and the global distribution of users across continents. Analyzing the dominance of top 50 users in competitions and exploring evolving participation trends, this visual journey uncovers valuable insights into Kaggle’s bustling community. Additionally, the project dives into forum sentiment and dataset dynamics, offering a holistic view of community interactions beyond competition realms. Through this visual exploration, ”Kaggle Insights” serves as an essential resource for understanding trends and dynamics within the thriving world of data science competitions on Kaggle.

## 2. Simple statistics of the data

The total files in the Meta Kaggle data set is 32. Among the 32 files, we have gone through all the data and finalized the below tables to analyze for our initial analysis.
The tables and their contents:



| Table Name | **Table Description** |
|---|---|
| **Competitions** | Contains competition titles, associated forums, organizing bodies, evaluation methods, team size limits, prizes, participant numbers, submission counts, and significant dates. |
| **Competition Tags** | Links competitions with specific tags for easy categorization. |
| **Tags** | Enumerates tags and tracks their usage frequency across competitions and datasets. |
| **Forum Topics** | Details discussion topics, associated forums, and the number of views they receive. |
| **Forum Messages** | Includes texts of individual messages posted within forum topics. |
| **Forum Message Votes** | Keeps records of votes on forum messages, noting both the voter and the recipient. |
| **Organizations** | A list of organizations and their corresponding names. |
| **Users** | Provides data on users, including their registration dates and performance tiers. |
| **User Organizations** | Maps users to their affiliated organizations. |
| **User Followers** | Indicates user relationships by showing who follows whom. |
| **User Achievements** | Catalogs users' achievements, points earned, and their rankings. |
| **Teams** | Lists team names along with their rankings in competitions. |
| **Team Memberships** | Associates users with their respective teams. |
| **Submissions** | Documents the specifics of competition submissions. |
| **Datasets** | Lists datasets and provides statistics on views, downloads, and votes. |
| **Dataset Tags** | Connects datasets with relevant tags. |

**Figure 1.** Meta Kaggle Dataset Tables

**Dataset Links:** Dataset 1, Dataset 2

## 3. Data Preprocessing & Exploratory Data Analysis

The Meta Kaggle dataset underwent streamlined processing to focus on essential data for analysis. Excel was used to remove duplicate data, and Power BI filled in missing values with ’N/A’ for categories or averages for numerical data, ensuring consistency across the dataset. Using pgAdmin, tables from Postgres were integrated and joined to establish relationships, enabling complex queries and multifaceted data exploration. As an example of preprocessing, the Competitions.csv dataset was loaded into Power BI, where missing values were replaced with ”N/A.” Additionally, erroneous rows with data conversion issues were removed to enhance data quality. Column profiling tools in Power BI were employed to analyze the dataset’s distribution and statistics, such as the ”Maxteamsize” column, which revealed insights into team composition, with 12 distinct values and a team size ranging from 1 to 20 members, among other valuable findings.

**Preprocessing codes in python:**Click here
**Tables creation in postgresql:**Click here

## 4. Existing work

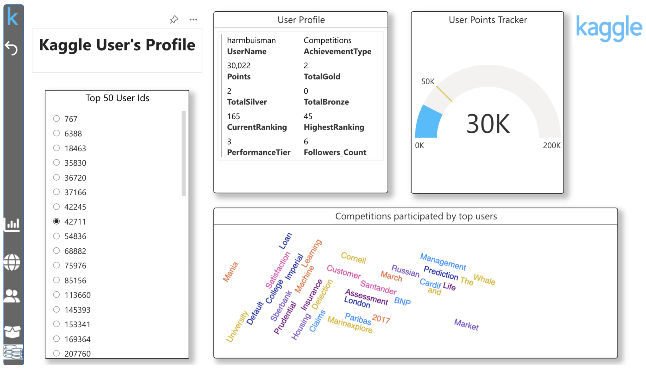
There is a python notebook created by Yoku Ishizaki[2] which presents a line chart visualization depicting the counts of competitions over a series of years, offering insights into the evolutionary trajectory of competitions within the Kaggle ecosystem over time. The visualization allows users to discern patterns, trends, and fluctuations in competition activity across different years, thereby facilitating a comprehensive understanding of the platform’s dynamics and growth over a specific time series.

There is a python notebook created by Jonathan Bouchet[1] showcases a histogram visualization of the top owners of the most upvoted datasets, providing a meaningful way to recognize prolific dataset creators in the community. This visualization not only acknowledges contributors but also inspires others to achieve similar success. It serves as a valuable resource for identifying collaborators, benchmarking performance, and gaining insights into effective dataset creation strategies. Ultimately, this visualization promotes community engagement, encourages dataset sharing, and enriches the collective knowledge of the Kaggle community.

## 5. Data Visualizations & Key Insights

**Project Deployment Url:** Click here

**Visualization 1: ”Kaggle Titans: Analyzing Top 50 Users and Their Competition Dominance”**
In this visualization, we focus on UserAchievements, Teams, TeamMemberships, and Competitions by filtering and selecting essential attributes such as userid, currentranking, achievementtype, teamid, competitionid, and competitiontitle. Through SQL preprocessing, we consolidated competition information of the top 50 users into a single table. The Power BI visualization utilizes a word cloud to highlight competitions where top users have participated. Users can interact with the visualization by selecting a specific user ID from a slicer, which then displays competition-related terms extracted from the titles associated with that user’s participation.



Kaggle User's Profile

Top 50 User Ids
○ 767
○ 6388
○ 18463
○ 35830
○ 36720
○ 37166
○ 42245
● 42711
○ 54836
○ 68882
○ 75976
○ 85156
○ 113660
○ 145393
○ 153341
○ 169364
○ 207760

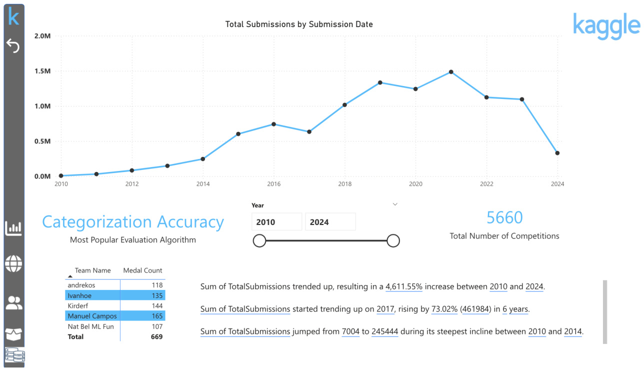
User Profile

| | |
|---|---|
| harmbuisman<br>**UserName** | Competitions<br>**AchievementType** |
| 30,022<br>**Points** | 2<br>**TotalGold** |
| 2<br>**TotalSilver** | 0<br>**TotalBronze** |
| 165<br>**CurrentRanking** | 45<br>**HighestRanking** |
| 3<br>**PerformanceTier** | 6<br>**Followers_Count** |

**Figure 2.** User Profile

**Insights gained:**
Analyzing the competitions in which top-ranked users(based on current ranking) participated provided insights into changes in community interests, the emergence of modern technologies, and evolving challenges. This visualization revealed that a considerable proportion of the top users have engaged in competitions centered around tasks such as classification, prediction, and forecasting.

**Visualization 2: ”Exploring Participation Trends: Insights from Competition Submissions”**
This visualization in Power BI combines data from Competitions and Teams tables to analyze total submissions and submission dates across competitions from 2010 to 2015. Using a line chart, it depicts submission trends over the years and includes interactive features such as a slicer for selecting specific years and a drill-down option to explore submissions at different time intervals. Additionally, two card visuals provide insights into the total number of competitions within the selected timeframe and the most popular evaluation algorithm based on occurrence counts.



Total Submissions by Submission Date
2.0M
1.5M
1.0M
0.5M
0.0M
2010
2012
2014
2016
2018
2020
2022
2024
kaggle
Categorization Accuracy
Most Popular Evaluation Algorithm
Year
2010
2024
5660
Total Number of Competitions

| Team Name | Medal Count |
|---|---|
| andrekos | 118 |
| Ivanhoe | 135 |
| Kirderf | 144 |
| Manuel Campos | 165 |
| Nat Bel ML Fun | 107 |
| **Total** | **669** |

Sum of TotalSubmissions trended up, resulting in a 4,611.55% increase between 2010 and 2024.

Sum of TotalSubmissions started trending up on 2017, rising by 73.02% (461984) in 6 years.

Sum of TotalSubmissions jumped from 7004 to 245444 during its steepest incline between 2010 and 2014.

**Figure 3.** Submission Analysis

**Insights gained:**

Competition submission analysis highlights a surge in submissions during the second quarter, followed by a decline in the third quarter. This trend could be due to competitions primarily occurring in the second quarter or reduced user interest in the third quarter when fewer competitions are available.
This visualization can help competition organizers understand what makes a competition successful and how to improve future events.

**Visualization 3: ”Kaggle Chronicles: Unveiling Competition Strategies and Team Dynamics”**
The visualization leverages data from a competition table, emphasizing attributes like HostSegmentTitle, CompetitorsCount, EvaluationAlgorithmsbyAbbrevation, EnabledDate, and HasKernel. In Power BI, two measures, KernelTrue and KernelFalse, are created based on the HasKernel attribute to tally counts for each category. This analysis provides insights into Kaggle competition dynamics, exploring reward types, host segments, and evolving evaluation algorithms. The visualization includes a year slicer to display competition distribution by reward types and host segments upon selection of a specific year. Additionally, it visualizes changes in competition evaluation metrics over the years, offering valuable perspectives on participant motivations and competitive landscapes across different industries or fields.

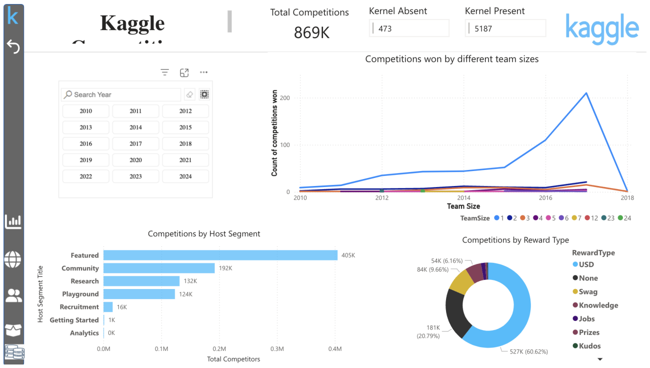


**Figure 4.** Competitions Analysis

**Insights gained:**
The bar graph indicates more wins by teams of size 1 over the years, suggesting that individual participants working alone have been successful in Kaggle competitions. This reflects the presence of talented individuals capable of producing high-quality solutions independently.
Solo participants dominate the wins, but occasional victories by small teams (size 2 or 3) show that collaboration can also lead to success, leveraging complementary skills and expertise.
Identifying which types of rewards are most common in competitions reveals trends in what motivates participants. From our visualization, USD was the most common reward type.
During COVID-19, we saw more competitions, many of which were community-focused and lacked rewards, indicating a shift away from monetary incentives during this period.

**Visualization 4: ”Meta Kaggle Insights: Exploring Forum Sentiment and Dataset Dynamics”**
We created a sentiment analysis visualization by analyzing the ”Forum Messages” dataset using the VADER sentiment analysis library. This process involved assigning polarity scores to each message and storing them in a new ”sentiment” column to track sentiment trends. In Power BI, we designed an area chart to display sentiment counts across years, categorized as positive, negative, and neutral. Users can interactively select sentiment categories and explore sentiment distribution over time.
Additionally, our visualization includes a bubble chart using data from the Datasets and Datasetversions tables, showcasing relationships between tags across dataset components. The bubble chart highlights the top 20 datasets based on total downloads, with bubble size representing the number of views, providing a clear view of dataset visibility within the Meta Kaggle community.

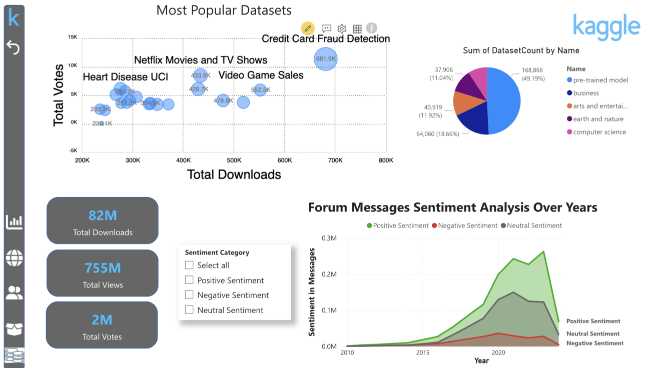
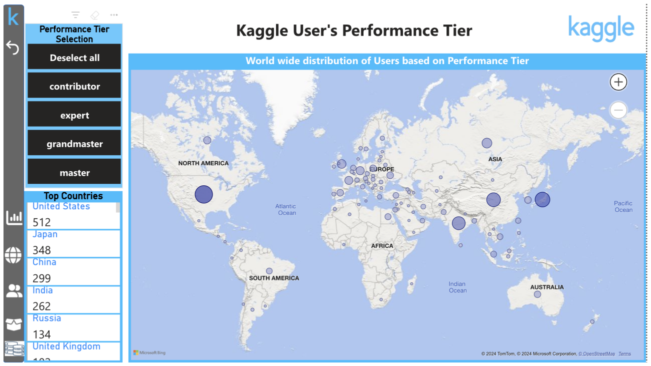


**Figure 5.** Dataset & Sentiment Analysis

**Insights gained:**
The sentiment analysis area charts reveal trends in forum communication over time, showing a peak in positive sentiment around 2020 followed by a decline, alongside a less pronounced pattern in negative sentiment and consistent levels of neutral dialogue.
The decline post-2020 in sentiments highlights evolving user behaviors or external events affecting communication trends, guiding content moderation and engagement strategies.
The high download count of the ”Credit Card Fraud Detection” dataset on Kaggle underscores the community’s focus on addressing financial transaction fraud, highlighting its importance in the industry.

**Visualization 5: “Kaggle Atlas: Visualizing User Distributions Across the Globe”**
This visualization integrates data from the MasterProfiles and MasterAchievements tables, expanding the dataset with additional Kaggle sources. Key attributes used include username, location, and competitions. The geospatial visualization show-

cases the distribution of Kaggle users worldwide based on performance tiers, from Novice to Grandmaster. This interactive tool offers a captivating glimpse into the global spread of Kaggle's diverse community, highlighting user concentrations and performance levels. With features like zooming and hovering for detailed insights, it provides a dynamic exploration of the Kaggle landscape, offering valuable insights into the platform's reach and dynamics.


Performance Tier Selection
Deselect all
contributor
expert
grandmaster
master
Kaggle User's Performance Tier
kaggle
World wide distribution of Users based on Performance Tier
NORTH AMERICA
EUROPE
ASIA
AFRICA
SOUTH AMERICA
AUSTRALIA
Atlantic Ocean
Pacific Ocean
Indian Ocean
Microsoft Bing
© 2024 TomTom, © 2024 Microsoft Corporation, © OpenStreetMap Terms

| Top Countries |
|---|
| United States<br>512 |
| Japan<br>348 |
| China<br>299 |
| India<br>262 |
| Russia<br>134 |
| United Kingdom |

**Figure 6.** Global User Performance Analysis

**Insights gained:**
We observed that Japan tops in grandmasters, US tops in masters and India tops in experts count.

## 6. Validation & Redesign

We improved the accuracy of our visualizations by addressing dataset issues with blanks and empty values using Python scripts, Kaggle Jupyter Notebooks, and Power BI. This process involved merging files and thorough data cleansing. Additionally, in the initial stages of our visualization project, we strategically selected the most effective graph types for each question through discussions and trials within Power BI, refining our approach to enhance clarity and relevance.
In ”Total Submissions by Submission Date,” we improved the visualization by switching to a line graph that clearly depicted submission trends over time. Adding an interactive slider enhanced user interactivity. In ”Unlocking Dataset Dynamics: Meta Kaggle's Most Popular Datasets,” we enhanced our approach with a bubble chart integrating downloads, views, and votes, offering deeper insights into dataset attributes and overcoming initial design challenges.

## 7. Challenges and Opportunities

**Challenges:**
Cleaning and reducing large datasets for Power BI presents challenges in data integrity, performance, and efficient modeling. Ensuring data accuracy post-cleansing demands meticulous planning, especially given Power BI's 1 GB dataset limit, which requires optimizing data size through transformation and aggregation. Creating accurate measures and defining relationships to avoid issues like ambiguity or many-to-many relationships also pose challenges. Addressing these complexities involves strategic use of Python, Power Query, DAX, and thoughtful schema designs to optimize data models for analysis and reporting.

**Opportunities:**
Preparing large datasets for Power BI enhances data preprocessing skills crucial for data science and fosters expertise in tools like Power Query and Databricks for versatile data manipulation. Mastering Data Analysis Expressions (DAX) improves proficiency in advanced analysis within Power BI, emphasizing effective data modeling for visualization and deeper understanding of data relationships. This process encourages creativity in designing impactful visualizations and innovative data summarization techniques to optimize efficiency.

## 8. Acknowledgements

We would like to extend our heartfelt gratitude to Professor Andreas Bueckle and our fellow students for their invaluable feedback throughout the development process. Their constructive insights and suggestions significantly enhanced the quality of this project. We also wish to express our appreciation to our project sponsor, Yashvardhan Jain, for providing valuable inputs that contributed to the project's outcomes.

## 9. Summary and Conclusions

In conclusion, the visualizations presented in this paper offer a multifaceted exploration of the Kaggle ecosystem, illuminating key aspects such as competition strategies, team dynamics, and the impact of top users. The analysis of user distributions across the globe highlights the widespread nature of Kaggle's community, underscoring its global presence. By uncovering patterns in user engagement and competition activity over time, these visualizations provide valuable insights into the evolving dynamics within Kaggle competitions. Additionally, the examination of forum sentiment and dataset dynamics enriches our understanding of community interactions and the evolving landscape of Kaggle's data resources. Together, these visualizations provide a detailed and actionable depiction of Kaggle's vibrant community, serving as a valuable resource for participants, researchers, and data enthusiasts seeking to navigate and comprehend the evolving trends in Kaggle competitions.

## References

[1]Jonathan Bouchet. “Kaggle metadata analysis”. In: (2018).
[2]Yuko Ishizaki. “Meta Kaggle Analysis”. In: (2019).
[3]Carl Mcbride Ellis. “Kaggle in Numbers”. In: (2024).
[4]Steubk. “Meta Kaggle-Master Achievements Snapshot”. In: (2024).