Meta Kaggle Client Project Details

**Author: Yashvardhan Jain**

# Meta Kaggle and Trends in Data Science and Machine Learning

Status: online

Data: online

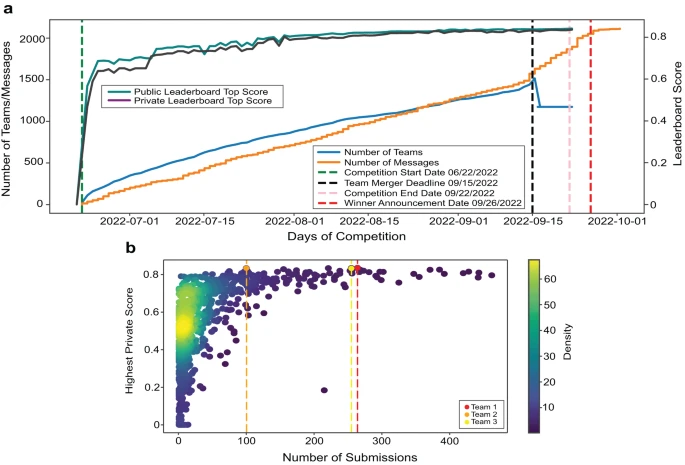


Image Source: <https://www.nature.com/articles/s41467-023-40291-0>

### Project sponsor: Yashvardhan Jain

### Sponsor email address: [yashjain@iu.edu](mailto:yashjain@iu.edu)

### Project description:

Each year, many Kaggle competitions are held on topics such as “Passenger Screening Algorithm Challenge”1, “Zillow Prize: Zillow’s Home Value Prediction (Zestimate)”2, and “Ultrasound Nerve Segmentation”3. Thousands of teams from around the world participate and compete for prizes totaling as much as $1,500,00 USD1 for a single competition. The results of some of these competitions have been documented in scholarly papers4,5,11. In addition to advancing AI-ready data development and algorithm innovation via competition, Kaggle challenges are places for extensive collaboration, learning, and mentoring among the data science community.

Different types of data visualizations can be created from Kaggle’s “Meta Kaggle” dataset6 to explore and communicate the structural and temporal dynamics of the evolving networks of competitors and their impact on success—in terms of team performance scores and expertise medals. Additionally, the Meta Kaggle and Meta Kaggle Code Datasets provide snapshots of how data science and machine learning has evolved over the past decade.

Your task is to utilize the **Meta Kaggle** and **Meta Kaggle Code** datasets to explore and gain insights on team performance, collaboration, and **trends** in data science over the past decade. Some visualizations for the “HuBMAP - Hacking the Kidney”8 and “HuBMAP - Hacking the Human Body”9 competition have been provided for inspiration but you are highly encouraged to explore further and go beyond these visualizations and be more creative. Ideally, you should ask 10 specific questions that you would like to explore using these datasets which would lead to 10 visualizations.

### Information on dataset(s) to be used:

The **Meta Kaggle** dataset6 provides detailed information on Kaggle competition (e.g., title, host name, deadlines, max. team size, evaluation algorithm, reward, number of teams/competitors/submissions), datasets (e.g., data source, versions, total views/downloads/votes), kernels (e.g., author, title, parent kernel, total views/comments/votes), teams (e.g., team members, leaderboard rank, medals), forum topics (e.g., name of discussion topic, first and last message dates, total messages/replies), and more. Tables were initiated in 2015 and are updated daily with information on closed competitions. Additionally, the **Meta Kaggle Code**10 dataset provides more insight into these competitions in terms of the actual code notebooks created.

Note: You will be required to have a [Kaggle.com](https://www.kaggle.com/) user account to access the data.

### Link to dataset:

* <https://www.kaggle.com/kaggle/meta-kaggle>
* <https://www.kaggle.com/datasets/kaggle/meta-kaggle-code>

### Relevant publications, websites, etc.:

Initial analysis and visualizations: [https://github.https://github.com/cns-iu/ccf-research-kaggle-2022/tree/main/utils/kaggle\_meta\_visualizationcom/cns-iu/ccf-research-kaggle-2022/tree/main/utils/kaggle\_meta\_visualization](https://github.com/cns-iu/ccf-research-kaggle-2022/tree/main/utils/kaggle_meta_visualization)

1. Passenger Screening Algorithm Challenge. <https://kaggle.com/c/passenger-screening-algorithm-challenge>.
2. Zillow Prize: Zillow’s Home Value Prediction (Zestimate). <https://kaggle.com/c/zillow-prize-1>.
3. Ultrasound Nerve Segmentation. <https://kaggle.com/c/ultrasound-nerve-segmentation>.
4. Ouyang, W. et al. Analysis of the Human Protein Atlas Image Classification competition. Nat. Methods 16, 1254–1261 (2019).
5. Jain, Y., Godwin, L.L., Ju, Y. *et al.* Segmentation of human functional tissue units in support of a Human Reference Atlas. *Commun Biol* 6, 717 (2023). <https://doi.org/10.1038/s42003-023-04848-5>
6. Meta Kaggle Dataset. <https://kaggle.com/kaggle/meta-kaggle>.
7. About – HuBMAP Consortium. <https://hubmapconsortium.org/about/>.
8. HuBMAP - Hacking the Kidney. <https://kaggle.com/c/hubmap-kidney-segmentation>.
9. HuBMAP - Hacking the Human Body. <https://www.kaggle.com/competitions/hubmap-organ-segmentation>
10. Meta Kaggle Code Dataset: <https://www.kaggle.com/datasets/kaggle/meta-kaggle-code>
11. Jain, Y., Godwin, L.L., Joshi, S. *et al.* Segmenting functional tissue units across human organs using community-driven development of generalizable machine learning algorithms. *Nat Commun* 14, 4656 (2023). <https://doi.org/10.1038/s41467-023-40291-0>

### Publication notes:

Software developed for this project should be made open source. Publications are encouraged, but should be coordinated with the project sponsor.

### Project sponsor biography:

<https://cns.iu.edu/current_team/bio/yashvardhan-jain.html>

# More Information

The primary objective of the project is to gain interesting insights from the publicly available data on Meta Kaggle and Meta Kaggle Code. This project is quite flexible and we are interested in what kind of insights you can gain from these datasets. Looking at how competitions “unfold” over time, team dynamics and correlation between public and private leaderboards, code sharing and discussions etc. are some of the things we could look at. Additionally, since the data repository contains competition information from the past ten years, what kind of insights can we gain from that in relation to how the ML and data science landscape has evolved in the past decade. Some things one could look at are the evolution of different packages and libraries people have been using. Once again, these are merely some examples and I would encourage you to dive deeper into the datasets and come up with ~ten or more interesting questions that you can answer from these datasets. More interesting and unique visualizations are also encouraged.

Looking at things like team dynamics, leaderboard correlation, how active discussion forums for competitions are and their correlation with number of participants and/or leaderboard scores, are some more examples.

Primary stakeholders for this project: Our lab has organized 4 Kaggle competitions in the past 3 years and we plan to organize more. We are interested in learning how effective these competitions are in terms of creating innovative solutions, and how this effectiveness changes across different competitions. This is also a Meta study on the Kaggle platform itself so we can look at interesting trends across Kaggle competitions.

I would encourage diving deep into all the different tables and properties available across both datasets and looking at how the two can be connected for interesting insights.

Meta Kaggle Code is fairly new (compared to Meta Kaggle) and is underexplored, especially how much code is actually available for different competitions and what can be extracted from that.

In terms of what specific visualizations to create, this is an area we would highly encourage you to explore and innovate in. While we have created some visualizations for specific competitions such as Figure 4 in<https://www.nature.com/articles/s41467-023-40291-0>, how can you expand on those to incorporate information from many different competitions. How do you look at a decade worth of Kaggle competitions?

Primarily, we want to showcase how visualization can “show” interesting trends for Kaggle competitions, such that they can be improved in the future. Additionally, the Kaggle competitions (in terms of models they use, libraries they use, types of datasets etc. ) give a snapshot of the landscape of ML and data science in the past ten years.

As mentioned before, we think the competitions and code on Kaggle can provide insight into how the field of ML and Data Science has evolved in the past decade. Beyond Kaggle, such competitions have become very common and there are a lot of researchers looking at benchmarking competitions as well. Meta Kaggle and Meta Kaggle Code, together, provide a huge repository of competition data over the past ten years or more. This can provide insights into different competition dynamics and effectiveness of such competitions.

Once again, this project is highly flexible and we encourage you to showcase your creativity in terms of what questions you ask and how you visualize your answers and insights.

I would also suggest reading<https://www.nature.com/articles/s41467-018-07619-7>.