Meta Visualization of 10-20 HuBMAP-relevant Kaggle Competitions

Pratap Roy Choudhury
Masters in Data Science
Indiana University, Bloomington, Spring 2023

Source- Google Image

Abstract— *Kaggle is one of the best platforms to host competitions for data science category with good reward amount. It is also a very good resource for dataset, innovative techniques, contributions by the experts for learning purpose. Healthcare domain data is one of the challenging data to analyze and gain insights and to take precautionary measures from the analysis results. Also, the meta data for any competition hosted in Kaggle plays a vital role to understand the people's involvement, participation for the growing challenges in terms of competition to solve problems, rise of the performance tier of contributors etc. This project report presents the results of a metadata visualization study conducted on relevant competitions hosted on Kaggle related to the HuBMAP (Human BioMolecular Atlas Program) project. HuBMAP aims to create a comprehensive map of the human body at a single-cell level to improve our understanding of human health and disease.*

Index Terms— Kaggle, Metadata, Competitions, Statistics, Exploratory Data Analysis, Visualization, Dashboard, Tableau

INTRODUCTION

Kaggle, a popular platform for data science competitions, has hosted several HuBMAP-related competitions that have attracted a diverse set of participants from around the world.

The metadata [1] of these competitions, including participant affiliations, submissions, and competition outcomes, provide valuable insights into the collaborations and competition outcomes in data science competitions. This project report presents the results of a metadata visualization study conducted on relevant competitions hosted on Kaggle related to the HuBMAP project. The study used various visualization techniques, including tree maps and scatter plots, to explore patterns and relationships among the metadata.

The goal of this study was to gain insights into the collaborations and competition outcomes in HuBMAP-related competitions hosted on Kaggle. By visualizing the metadata, this study aimed to identify patterns and relationships among the participants, their affiliations, submissions, and competition outcomes. The study also aimed to evaluate the quality of the submissions and their correlation with the competition outcomes.

The importance of this study lies in its potential to provide valuable insights into the collaborations and competition outcomes in data science competitions related to the HuBMAP project. These insights could help researchers and data scientists to better understand the collaborations and competition outcomes in data science competitions, which could ultimately advance the goals of the HuBMAP project. Furthermore, the visualization techniques used in this study could be applied to other Kaggle competitions, providing insights into the patterns and relationships among the metadata of various data science competitions.

1 Insight Needs

The insight needs of this project include gaining a better understanding of the collaborations and competition outcomes in data science competitions related to the HuBMAP project. Specifically, the study aims to identify patterns and relationships among the metadata including team engagement, collaboration between participants from different tier ranks, number of teams participating, and the user distribution based on the performance tier, number of submissions and quality of the outcomes.

The study also seeks to evaluate the quality of the submissions in terms of evaluation score of the submissions, correlation with the competition outcome, dataset categories used in different competitions, most popular dataset based on the votes, views and downloads, user registration in Kaggle who participated in the competitions etc.

2 Data Description and Analysis

The Kaggle metadata contains 32 entity tables out of which 10 most important tables are used in this study. Competition and Tags are the entity tables that give information about all the different competition id and relevant metadata. This study focuses on 10 most HuBMAP relevant competitions including two HuBMAP competitions. All these competitions are all closely related to technology-powered healthcare solutions in the areas of organ dysfunction, disease prediction or anticipation, and other related areas of research.

The major entity tables that are used in this study are Competitions, which includes information about competition id, start and end date, mapping with tag id. Teams, Team Membership tables have data related to different team ids and users in each team who participated in competitions. Users table gives the user information and their performance tier. Submission table has all the records of submissions made by the team and user along with the evaluation method and precision score. Datasets, Dataset Tags tables have the data of each dataset that are used in all the competitions, their ids, number of votes, views, and downloads of each dataset. The Organization table has the data regarding which organisation hosted the competition, who provides the dataset etc.

The initial exploratory data analysis includes merging relevant data tables in Python to fetch information about the ten selected competitions. Firstly, the ten relevant competitions are filtered from the competitions table and then team details with their public leader board ranking, submission precision scores are filtered from the Teams table. For the selected teams, details of teams and users of those teams are filtered from the Team Membership table. Data analysis on these merged data shows some interesting results such as most of the users in each team belongs to tier 0 which is novice and very minimal amount of master and Grandmasters are in the teams. Performance Tier is a categorical column with 5 categories depicted as Novice, Contributor, Expert, Master, Grandmaster level.

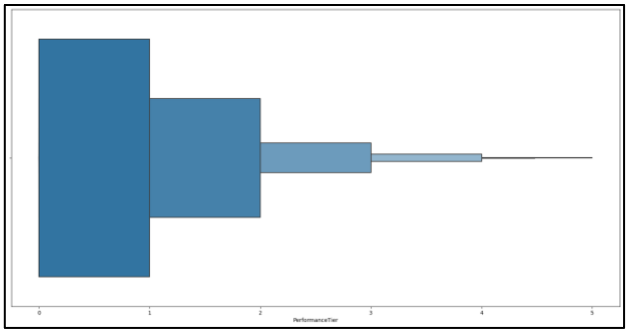


Fig 1. Boxen plot of distribution of users based on performance tier

The competition table has the rewards type and reward amount of each competition from which it has been found that the competitions 22990 (HuBMAP-Hacking the kidney) and 34547 ('HuBMAP + HPA - Hacking the Human Body) have the highest prize money of 60K USD. 3 of the competitions (9318, 9511, 11433) have 0 rewards whereas 5 competitions (13050, 14502, 22995, 27177, 27290) have rewards field NaN in the dataset.

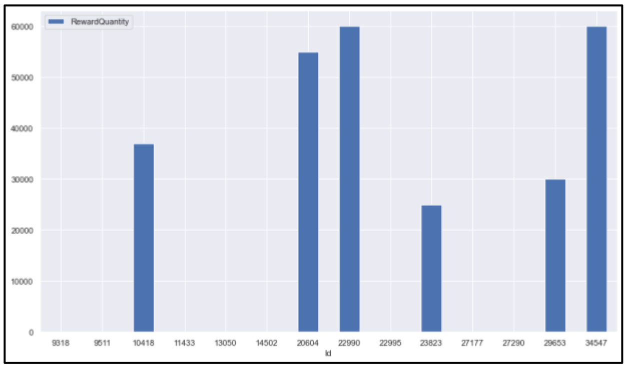


Fig 2. Reward quantity of each competition in USD

The submission table has a column PublicScoreFullPrecision which has a median value of 0.44 and standard deviation of 3.27. If we see a distribution of the values in a box plot, we can see the details clearly. It has a min value of -76 based on the evaluation metric of a competition and max value of 23.8478.

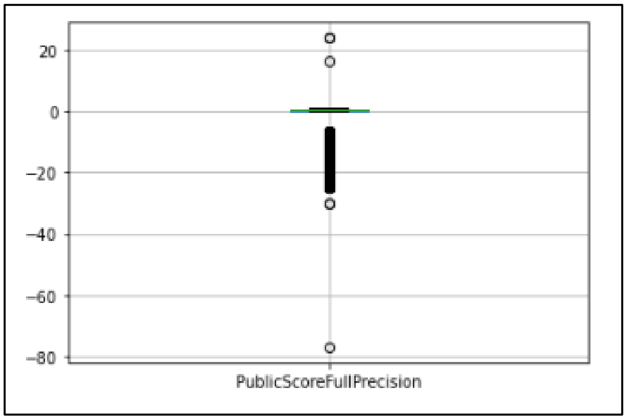


Fig 3. Box plot of public full precision score

The box plot shows that few precision scores are outliers and upon exploration, it has been found that, many competitions used Laplace log likelihood as the evaluation algorithm, which gives better accuracy when the score is highly negative.

The Users, Tags, Dataset Tags tables are filtered for the target competition Id and Team Ids that are fetched previously and merged based on the competition and tag id. Similarly, the Organisations table is also filtered for the required data points. Lastly, the data tables are mapped with a mapping dictionary to rename the competition Id with the competition names, Evaluation algorithm names for the submissions and the tier names with the performance tier rank.

3 Visualizations

The best and the most effective way to represent the insights is through tableau visualization. To examine how many submissions are made before and after the deadlines, how many users participated in the competition based on their tier ranks, visualization such as temporal bar graph is very useful. For gaining insights about the hierarchy of the teams and users in each team participating the competitions, competition conducted under different tag and parent tags, Tree map visualization is employed. Also, the scatter plot with circles shows the result of the maximum precision score obtained in each competition submissions both before and after deadline.

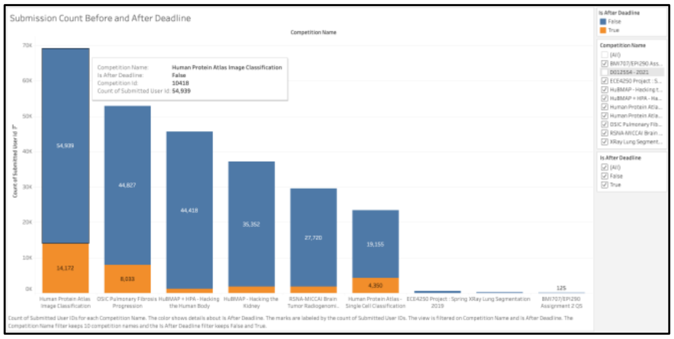


Fig 4. Number of submissions before and after deadline

In Fig 4 above, the number of submissions for each competition are showed as a bar graph stacked with the count for before and after deadline of the competition. All the competitions have greatest number of submissions done before the deadline. In competition id 10418 – Human Protein Atlas Image Classification, there were huge submissions done even after the deadline which shows the importance of the competition. Competition name and True/False selection filters are added for the client to select any competition for its before and after the deadline submission count.

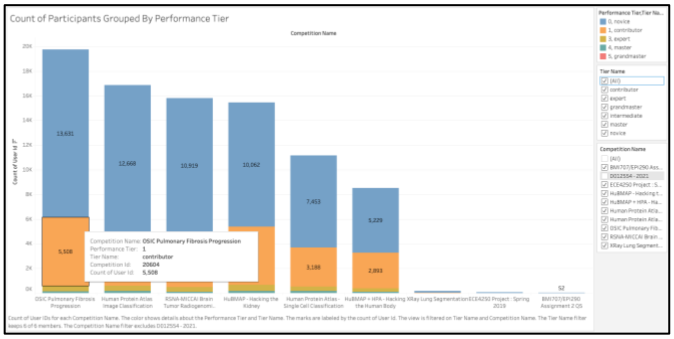


Fig 5. User participation grouped by performance tier

A stacked bar graph in tableau is generated in Fig 5 to show the total user participation across different tiers. Scatter plot was a challenge as the performance tier cannot be used as a measure which is a requirement in tableau. Each colour represents the performance tier of the participants, and the counts are given as label. Competition Name and Tier names are added a s custom filter for the user to get an insight about the participants from different tier for any selected competition. This visualization shows that most of the members are from tier 0- novice and a very minimum count of members belong to tier 4- master and tier 5- grandmaster.

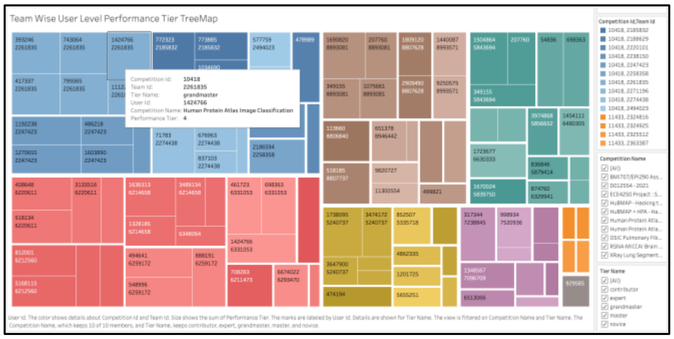


Fig 6. Team wise user participation from different tier

A very interesting insight has been discovered regarding the teams which are in top 10 public leader board ranking in each competition. The team member data has been filtered for each competition with the top 10 teams to check how the user performance tier impacts the team to get the top position. A tree map visualization is created in Fig 6 that shows the distribution of team members in every team that participated in every competition. The hierarchy is designed as the different colour map represent different competitions and the colour saturation in every unique colour map represents the unique team. Every block within the team represents each member and the block size is according to the member's tier rank. Higher the tier – bigger the size. It shows that the team id 2261835 participated in competition id 10418 has 6 members and five of them are the Grandmaster i.e., tier-4 and one of them is a Master i.e., tier-3. Similarly, it shows that the competition id 10418 (blue) and 11433 (Orange) have the greatest number of teams participated with most of the members from master and Grandmaster level.

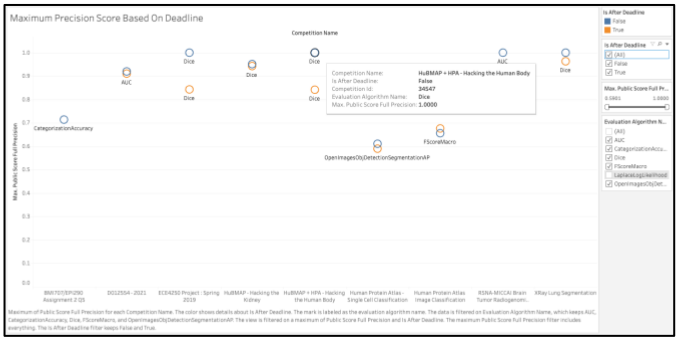


Fig 7. Scatter plot of maximum precision score in each competition

In Fig 7, a scatter plot of maximum precision score for all the competitions generated. It shows the evaluation metric score for the winning submissions in each competition. There is different evaluation algorithm used in these competitions to evaluate the submissions such as Dice, AUC, F-score Macro, Categorization Accuracy. It shows that the precision scores are always better when submitted before the deadline. The evaluation algorithm names and True/false selection for before and after deadline are added as filter. For one of the competition ids 20604 – 'OSIC Pulmonary Fibrosis Progression' has the precision score in negative range -6.787 because it used the evaluation algorithm Laplace Log Likelihood which is a negative score metric. This becomes an outlier in the scatter plot, so this evaluation name has been filtered out by default from the visualization.

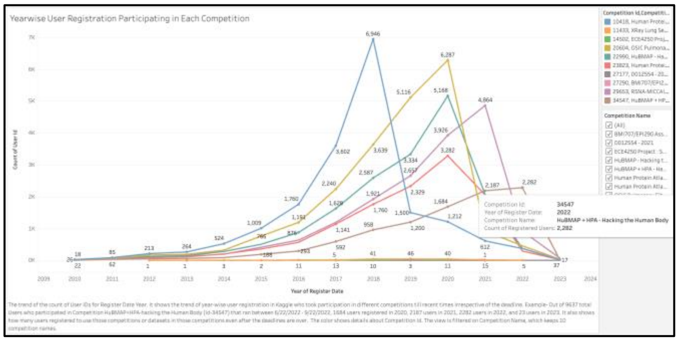
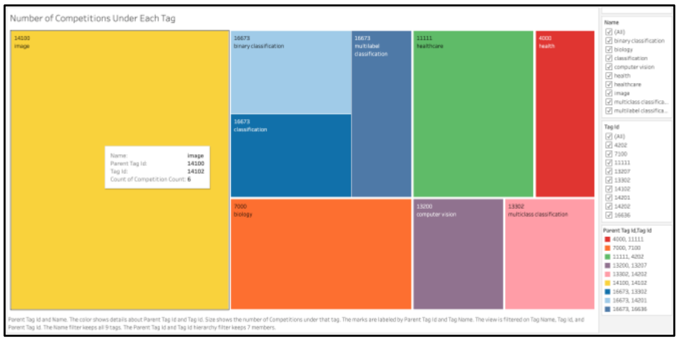


Fig 8. User registration trend who participated in every competition

In Fig 8, a line chart visualization shows the trend of all the user registrations who participated in different competitions when it started. This visualization focuses specially to the users who joined Kaggle in various time. It answers the insight need of how many users joined Kaggle in different years who participated in selected competitions. For an example, a user who registered in Kaggle during 2010, has participated in competition that has been conducted in 2022. Majority of the users joined Kaggle during 2017-2022. In recent time, competition id 34547 – HuBMAP+HPA Hacking the Human Body is finding more user registration i.e., new people are registering and participating in the competition irrespective of the deadline which is expected to rise in 2023.



Parent Tag Id and Name. The color shows details about Parent Tag Id and Tag Id. Size shows the number of Competitions under that tag. The marks are labeled by Parent Tag Id and Tag Name. The view is filtered on Tag Name, Tag Id, and Parent Tag Id. The Name filter keeps all 9 tags. The Parent Tag Id and Tag Id hierarchy filter keeps 7 members.

Fig 9. Number of competitions under each tags

In Fig 9, the tree map shows the distribution of the competition count under each parent tag and tag name. The size of each rectangle in the tree map represents the number of competitions held under each tag. By looking at the size of the rectangles, one can easily identify the most active areas of competition related to HuBMAP. For example, a large rectangle is associated with a 'image' tag, it indicates that there are 6 competitions being held under this tag. The colour of the rectangles provides details about the parent tag and tag ID. The filters applied to the tree map provide further insights into the distribution of competitions within the HuBMAP project.

4 Results and Insights

The results showed that the HuBMAP competitions had a diverse set of participants from various academic and research institutions worldwide. There were also numerous collaborations among participants, indicating the importance of teamwork and collaboration in achieving the goals of the HuBMAP project. The scatter plot revealed that the competition submissions were of high quality with good precision score.

In this project, a comprehensive interactive dashboard has been created that combines all the visualizations created from the metadata of HuBMAP-related competitions hosted on Kaggle. The dashboard includes a range of visualizations, such as stacked bar graph, scatter plots, tree maps, and bubble chart that provide insights into the collaborations and competition outcomes of these competitions.

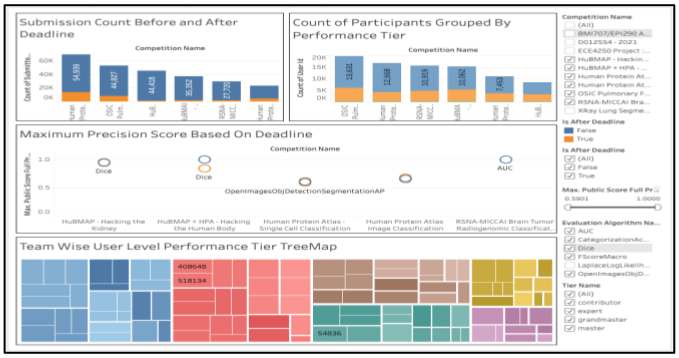
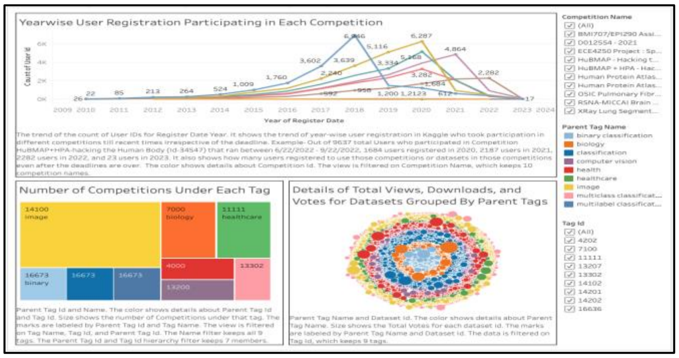


Fig 10. Dashboard-1

Dashboard 1 Link- https://public.tableau.com/app/profile/pratap.roy.choudhury/viz/VizNinja-HuBMAPKaggleMetaDataVisualizationDashboard/Dashboard1

To make the dashboard user-friendly and easily understandable, necessary filters and legends are added for all the visualizations. These filters enable users to focus on specific areas of interest within the metadata, such as participant affiliations, submissions, and competition outcomes. The legend helps users to understand the meaning of the colours and shapes used in the visualizations, making it easier to interpret the data.

For better understandability of the visualization, simple caption texts are added to the visualizations in dashboards as storytelling. These explanations give a high-level summary of the story that are attempted to convey through the visualizations.



The trend of the count of User IDs for Register Date Year. It shows the trend of year-wise user registration in Kaggle who took participation in different competitions till recent times irrespective of the deadline. Example- Out of 9637 total Users who participated in Competition HuBMAP+HPA-hacking the Human Body (Id-34547) that ran between 6/22/2022 - 9/22/2022, 1684 users registered in 2020, 2187 users in 2021, 2282 users in 2022, and 23 users in 2023. It also shows how many users registered to use those competitions or datasets in those competitions even after the deadlines are over. The color shows details about Competition Id. The view is filtered on Competition Name, which keeps 10 competition names.

Parent Tag Id and Name. The color shows details about Parent Tag Id and Tag Id. Size shows the number of Competitions under that tag. The marks are labeled by Parent Tag Id and Tag Name. The view is filtered on Tag Name, Tag Id, and Parent Tag Id. The Name filter keeps all 9 tags. The Parent Tag Id and Tag Id hierarchy filter keeps 7 members.

Parent Tag Name and Dataset Id. The color shows details about Parent Tag Name. Size shows the Total Votes for each dataset Id. The marks are labeled by Parent Tag Name and Dataset Id. The data is filtered on Tag Id, which keeps 9 tags.

Fig 11. Dashboard-2

Dashboard 2 Link- https://public.tableau.com/app/profile/pratap.roy.choudhury/viz/VizNinja-HuBMAPKaggleMetaDataVisualizationDashboard/Dashboard2

5 Conclusion

The study provided valuable insights into the metadata of the HuBMAP-related competitions hosted on Kaggle. The visualization techniques used in this study could be applied to other Kaggle competitions to gain further insights into the patterns and relationships among the metadata. These insights could help researchers and data scientists to better understand the collaborations and competition outcomes in data science competitions and advance the HuBMAP project's goals.

Acknowledgments

We would like to thank Andreas Bueckle, Michael Ginda, and our sponsor Shriya Mandarapu for their constant support and guidance. We are also grateful for the peer reviews given to our project write-up to resolve us the general questions and better designing the visualization and results.

References

[1] https://www.kaggle.com/datasets/kaggle/meta-kaggle

[2] https://github.com/llschers/Envisioning-Kaggle

[3] https://www.kaggle.com/code/steubk/kaggle-grand-masters-map/notebook

[4] https://www.kaggle.com/code/stassl/bms-competition-stats/notebook

[5] https://docs.appspace.com/latest/how-to/create-leaderboard-chart/

[6] https://www.emerald.com/insight/content/doi/10.1108/TPM-03-2019 0024/full/html?skipTracking=true

[7] https://www.kaggle.com/competitions/osic-pulmonary-fibrosis-progression/overview/evaluation