**#1: Meta-Kaggle and Trends in Data Science and Machine Learning**

**#2: Visual Analysis of Meta-Kaggle Data to Uncover Trends in the Data Science Landscape**

**#3: Authorship**

**Kabir Chaturvedi**

Indiana University

[kschatur@iu.edu](mailto:kschatur@iu.edu)

**Yeon-Soo Chung**

Indiana University

[chungye@iu.edu](mailto:chungye@iu.edu)

**Bhoomika Pathapati**

Indiana University

[pbhoo@iu.edu](mailto:pbhoo@iu.edu)

**Monisha Patro**

Indiana University

[monpatro@iu.edu](mailto:monpatro@iu.edu)

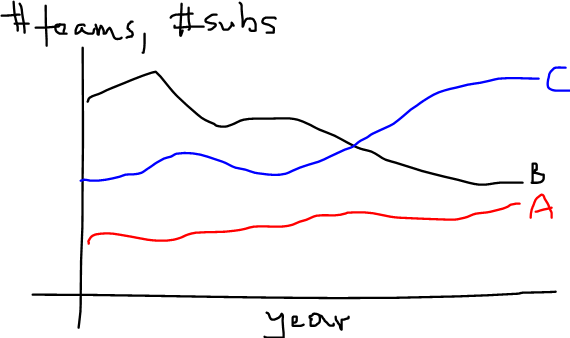
**Ishika Thakur** Indiana University [isthakur@iu.edu](mailto:isthakur@iu.edu)



**#4: Description of visualization goal/need, hand-sketch of the envisioned visualization, and discussion why this project is important.**

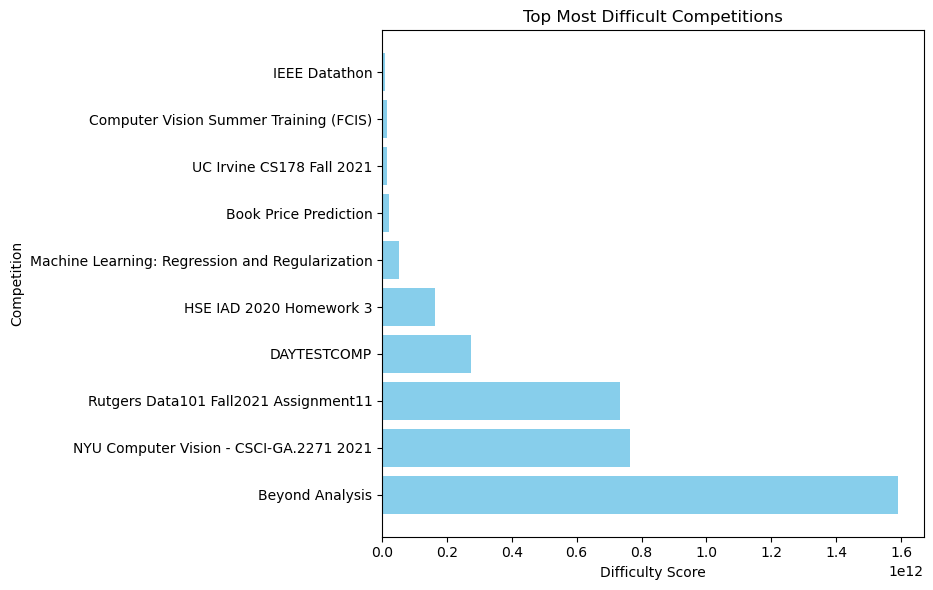
Below are our preliminary visualizations and the questions we aim to answer with them for this project. These will likely be updated and enhanced as we continue to explore and analyze the data.

* **Which competitions were most popular over time? Which competition tags were most popular over time? Let’s add a time component to these plots (e.g., line graph) and look for any temporal trends. Replace tag ID with tag name. Bhoomika.**



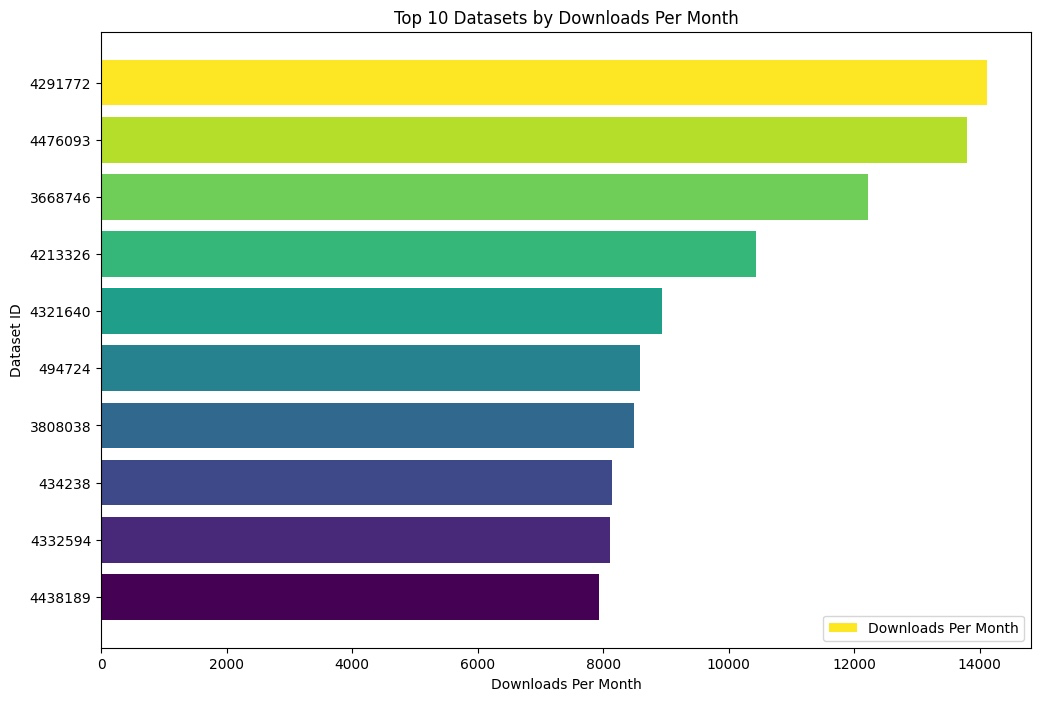
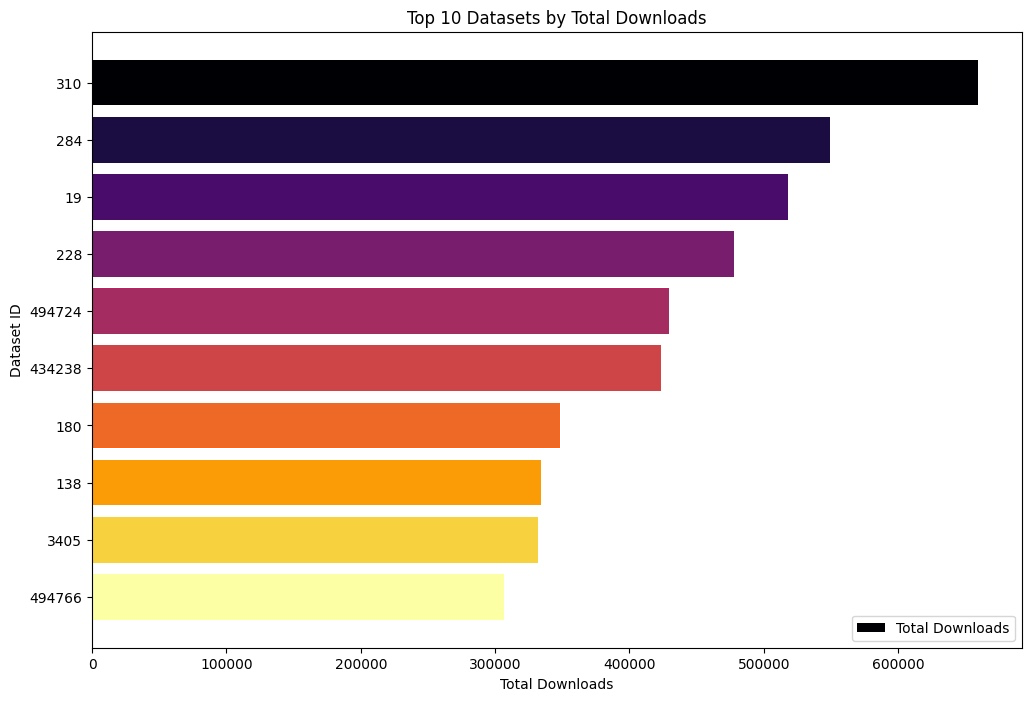
From the above visualizations, we can see the 10 most popular competitions based on the number of submissions and number of competitors. Also we are displaying the competition tags that were most popular over time.

* **Which competitions were most difficult (measured by lowest scores, least submissions, etc.)?**



This visualization showcases the competitions rated as the most challenging, highlighting considerable variation in difficulty scores among them. Certain competitions stand out with notably higher difficulty levels compared to others, indicating a diverse range of challenges within the dataset.

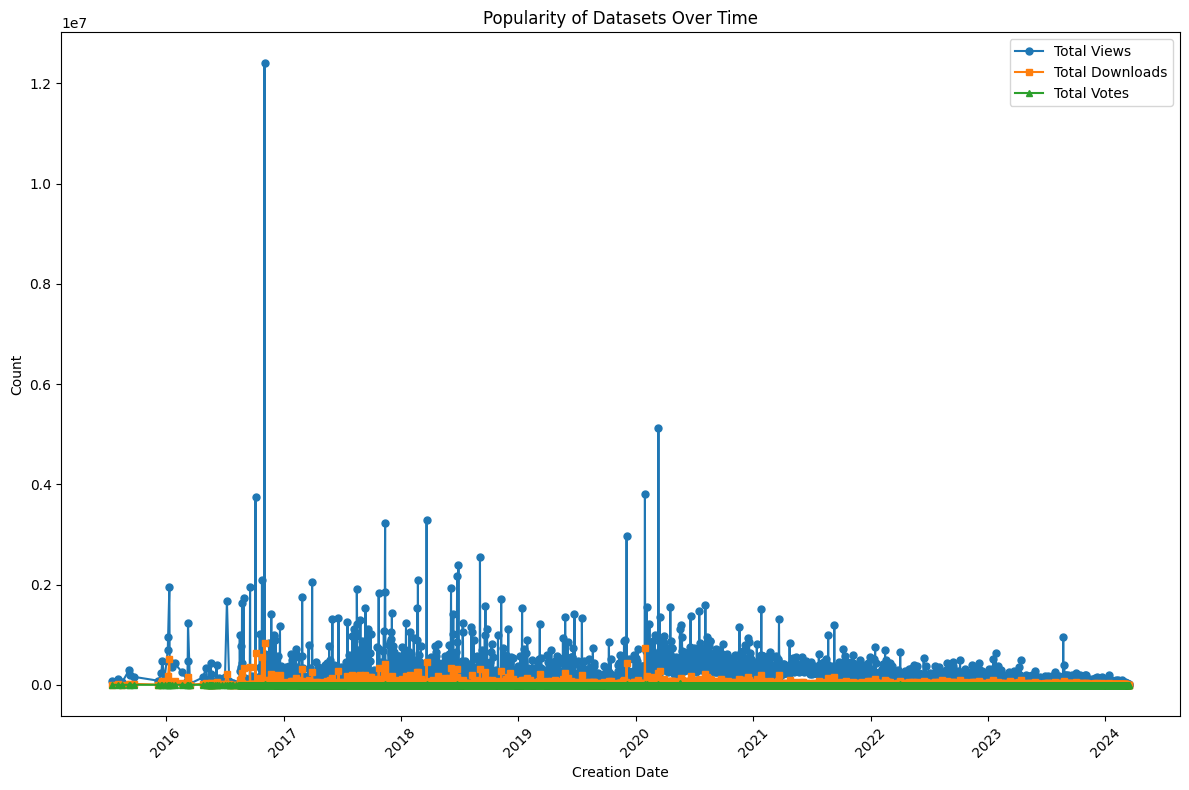
* **What are the top 10 datasets in terms of total downloads and average monthly downloads? - Monisha**



The two images present horizontal bar charts displaying the top 10 datasets based on different download metrics. The first chart is titled "Top 10 Datasets by Total Downloads" and it orders the datasets by their total number of downloads. Each dataset is identified by its unique ID on the y-axis, and the total download count is represented by the length of the bars on the x-axis, with different colors for visual distinction. The most downloaded dataset has close to 600,000 downloads, while the least downloaded within the top 10 has noticeably fewer.

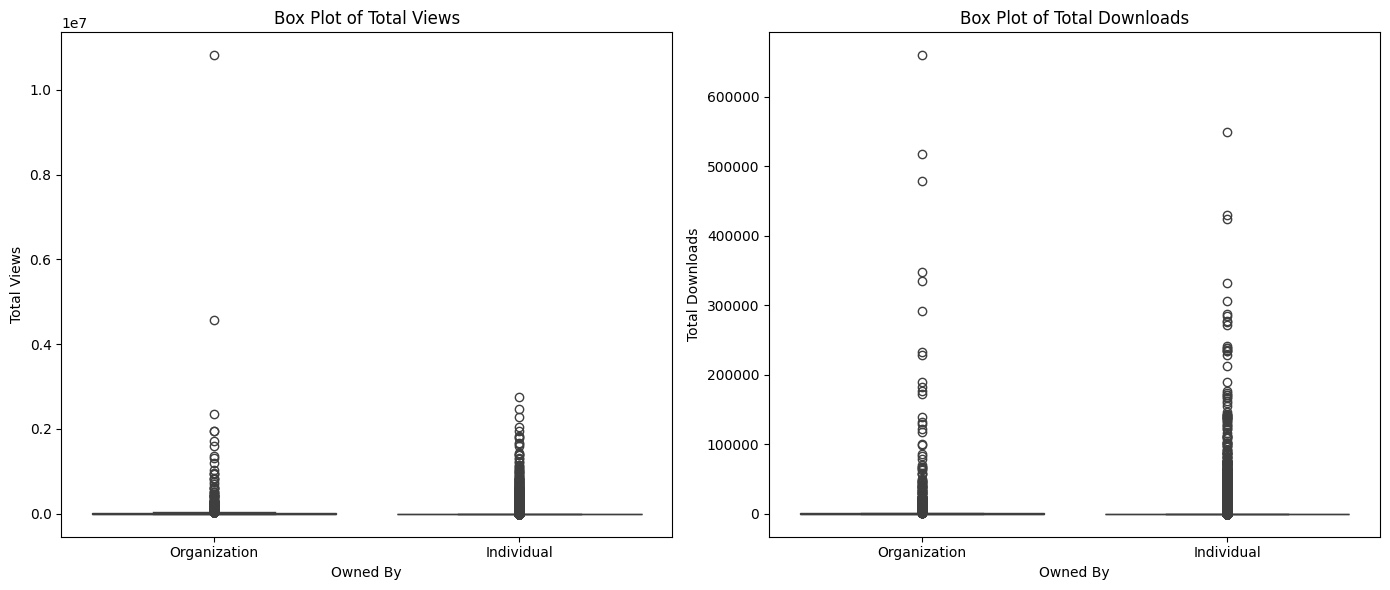
The second image, titled "Top 10 Datasets by Downloads Per Month", shows a similar layout but focuses on the average monthly download numbers for the top datasets. Here again, each dataset is represented by an ID, with the monthly downloads reflected by the bar lengths. The scale for monthly downloads is much smaller, suggesting a different period or a calculation of average downloads over time. While both charts include datasets identified by their IDs, they may not necessarily feature the same datasets or ranking order, as one measures total downloads and the other average monthly downloads.

* **How has the popularity of datasets changed over time? (measured by views, downloads and votes) - Monisha**

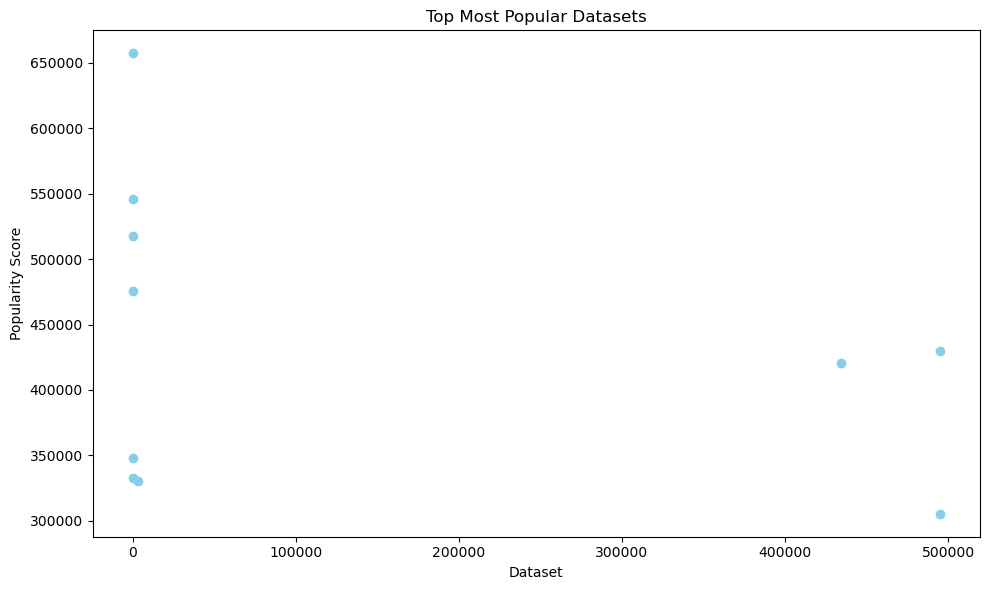


This image shows a scatter plot titled "Popularity of Datasets Over Time," which tracks three different metrics: total views (blue), total downloads (orange), and total votes (green) for datasets from 2015 to 2024. The creation date of each dataset is on the horizontal axis, and the count for each metric is on the vertical axis, going up to 10 million. The plot points indicate that while views are the highest and most variable over time, with some datasets experiencing particularly high viewership, downloads and votes are significantly lower and have fewer standout peaks, suggesting that while many people may view datasets, fewer take the step to download or vote on them.

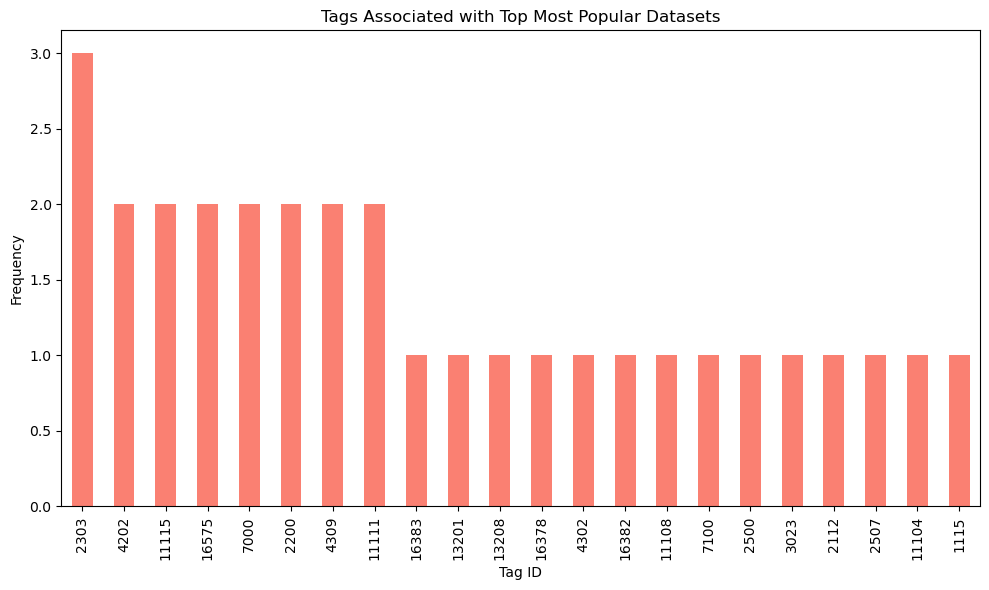
* **Do datasets owned by organizations receive more views/downloads compared to those owned by individuals? - Monisha**



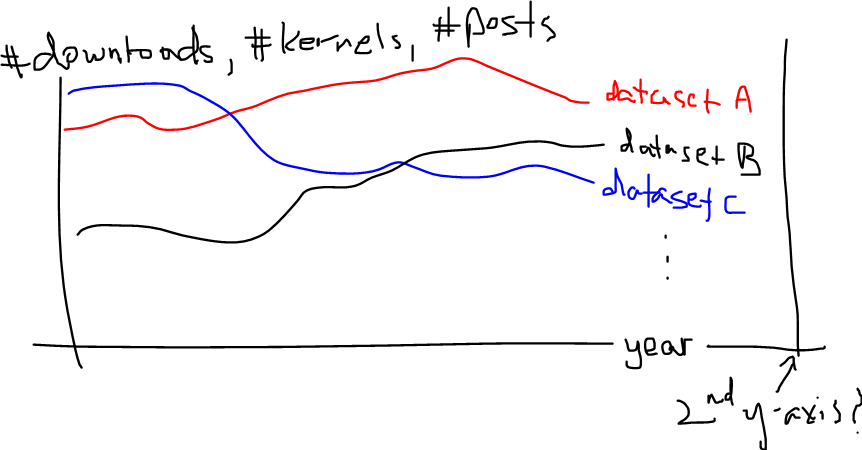
The image presents two box plots comparing the total views and total downloads of datasets owned by organizations versus those owned by individuals. In both plots, the x-axis categorizes the data by owner type, while the y-axis shows the number of views or downloads. The box plots illustrate the spread and outliers of the data. They indicate that both views and downloads have a wide range but with many outliers suggesting that a few datasets, whether owned by organizations or individuals, have exceptionally high views or downloads compared to most. The plots also appear to show that individual-owned datasets have a slightly higher spread in the number of downloads, hinting at greater variability compared to organization-owned datasets.



**^Is x-axis data set ID? Please replace it with name of data set. How was popularity score determined?**

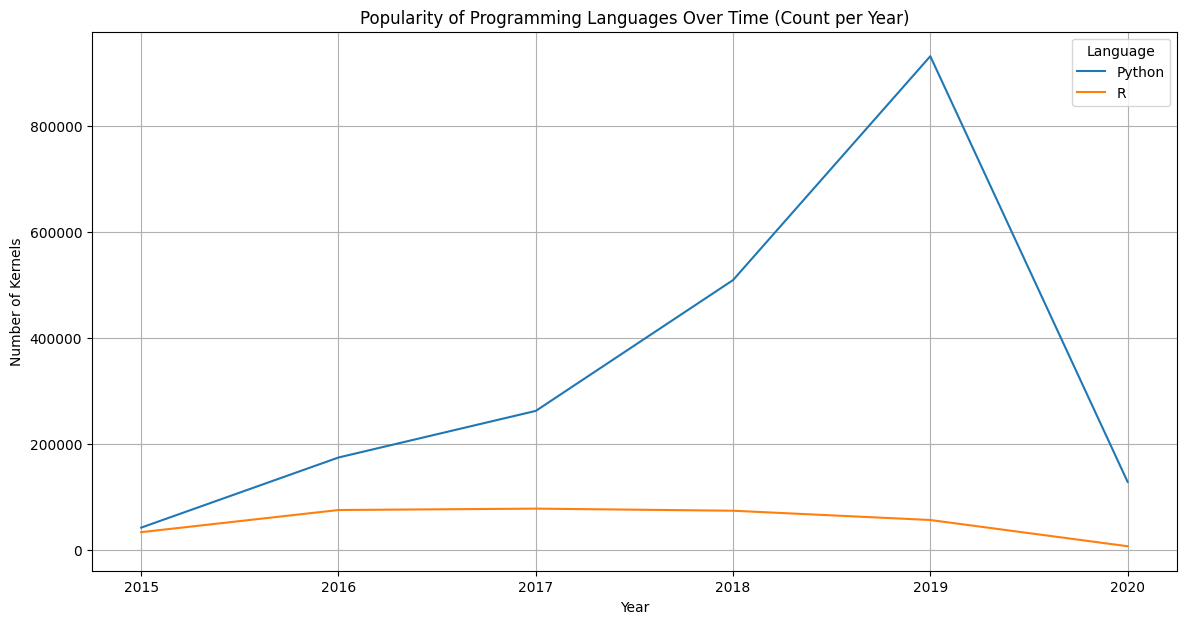


**^Replace tag ID with tag name. It seems like this plot is showing some data sets that were downloaded only 1, 2, or 3 times. I think we can replace the above two plots with one plot that has y-axis being the number of downloads, and also add a temporal dimension (line graph):**

****

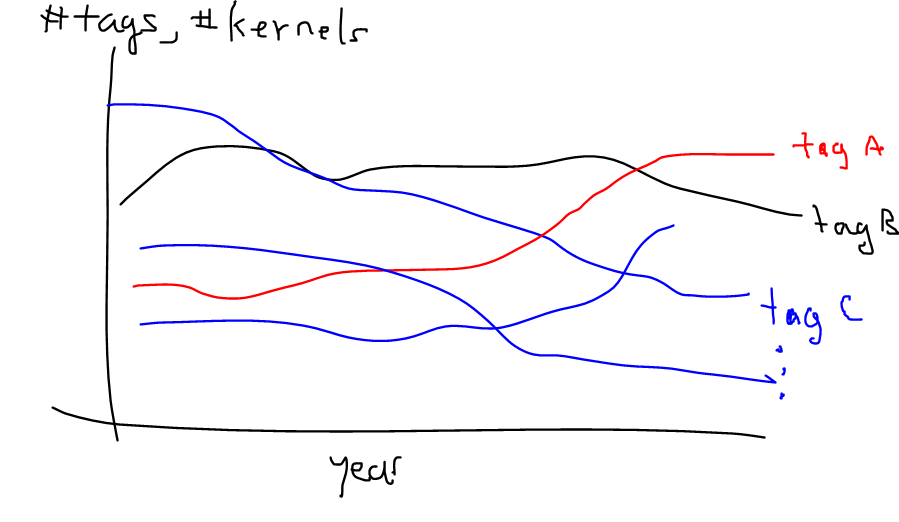
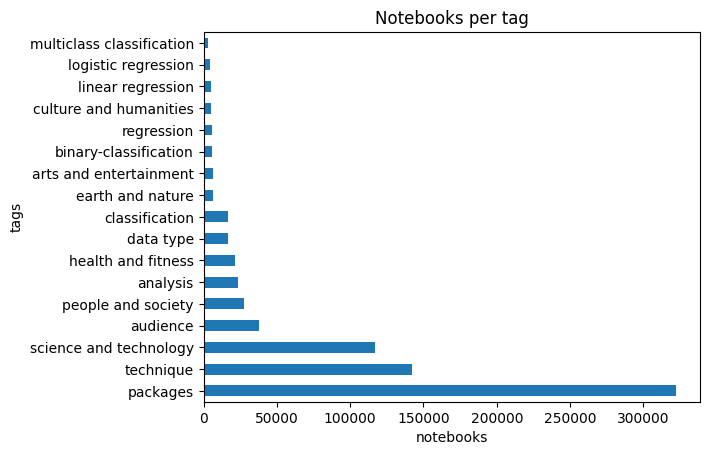
The first visualization depicts the most popular datasets over time in the form of a scatter plot whereas the second visualization displays the tags associated with the most popular datasets using a bar chart.

* **How has the popularity of Python and R changed over time? Seems strange to have this steep drop. Do we only have data from part of 2020? If so, maybe we can remove 2020 data. Kabir**

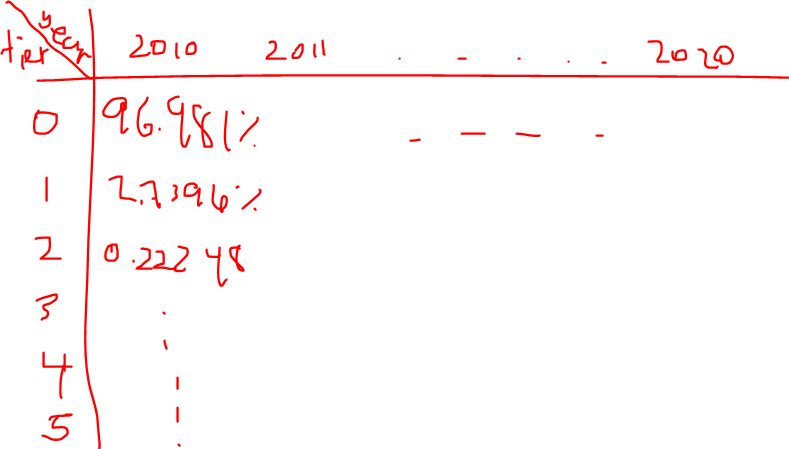


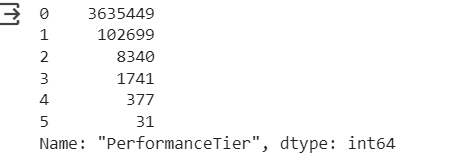
Python's popularity has shown a significant increase over the years, reaching a peak in 2019 before a sharp decline. In contrast, the popularity of R has remained relatively stable and significantly lower than Python throughout the same time period.

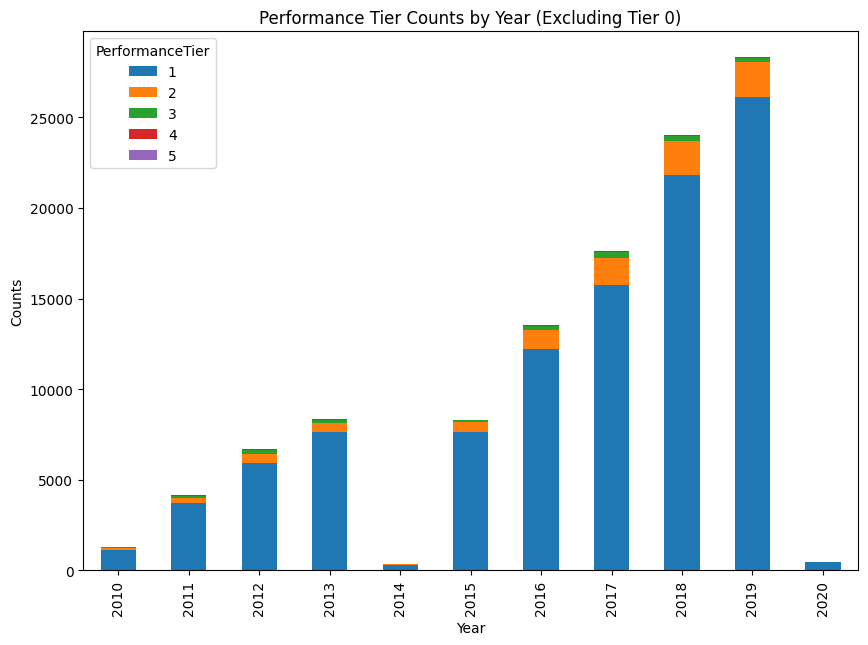
* **Which notebook tags were most popular over time? Let’s add temporal dimension: line graph with years on x-axis. Also add plot for which kernel tags were most popular over time. Kabir**



The “packages" tag is the most popular tag, with over 300,000 notebooks associated with it, followed by "technique" and "science and technology," which also have a substantial number of notebooks tagged.

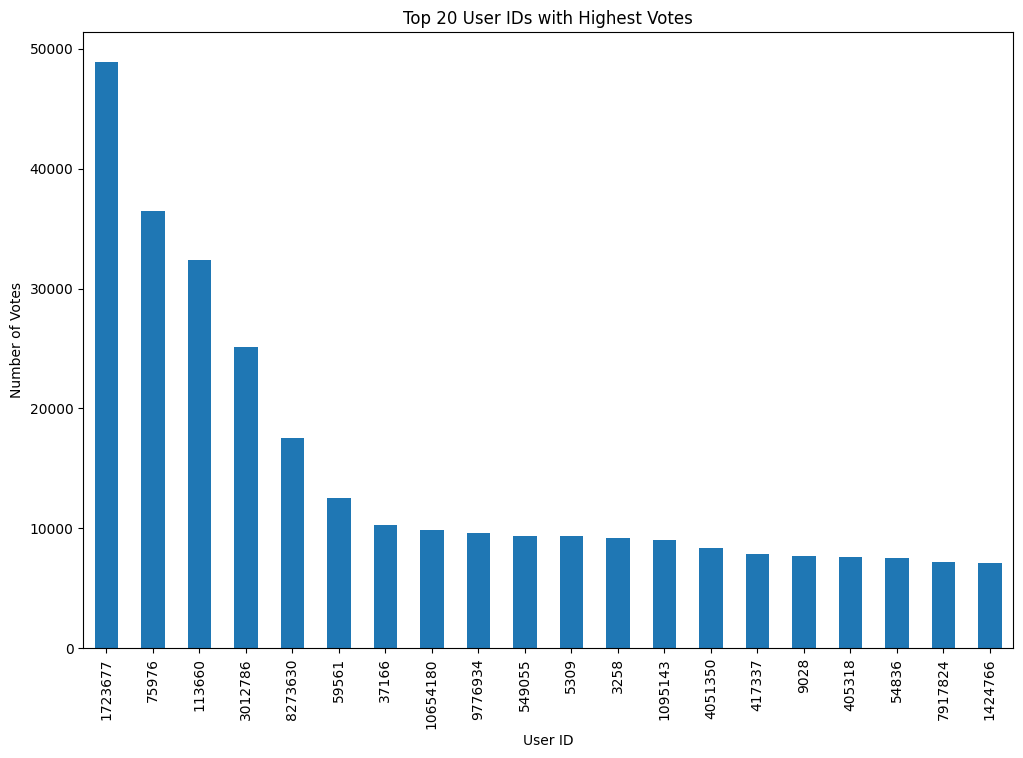
* **How did the distribution of performance tiers change over time? I think a table of percentages might be a better visualization for this. It would be interesting to show if the percentage breakdown of the tiers stays relatively consistent over the years. Also, please check the anomalies at 2014 and 2020. Bhoomika  
  **

****

****

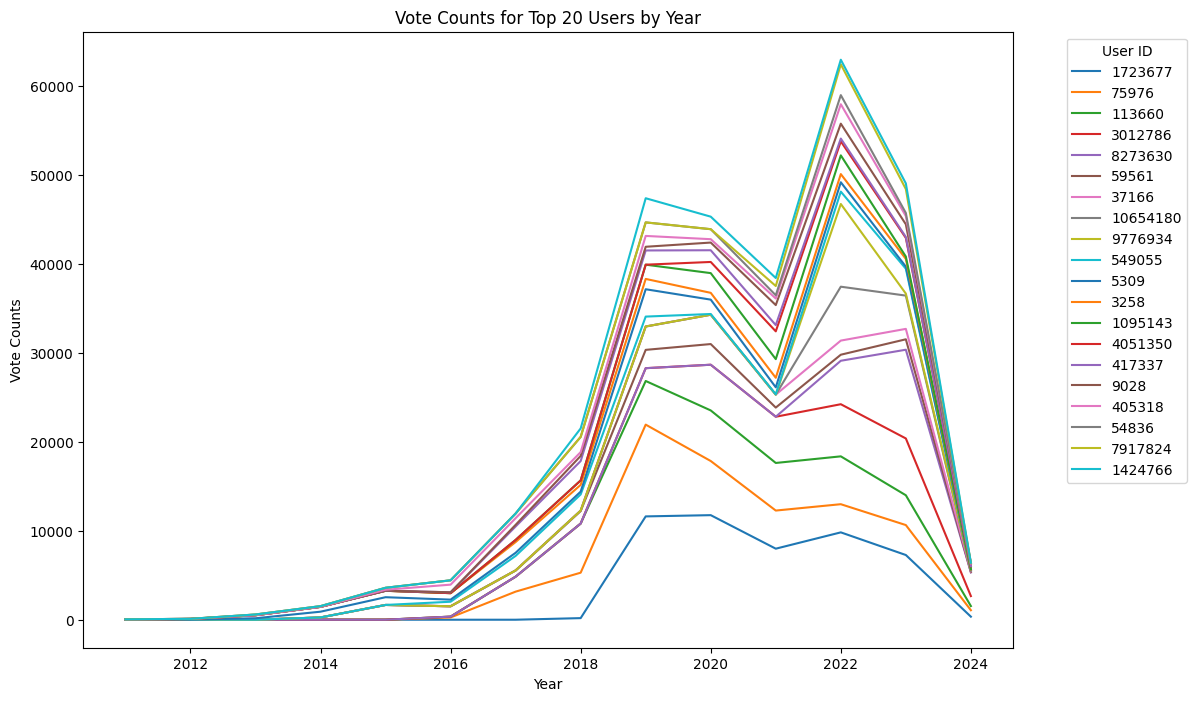
The graph shows the performance tier count distribution during the years 2010-2020.Since most of the column values were 0 I removed those values to focus more on the rest of the tiers distribution.

* **Which users received the most votes in forums and what is its distribution? For the next three plots, I was interested in scatter plots of those variables plotted against each other (e.g., number of votes vs number of posts). Add temporal dimension, like add colors to the scatter plots which represent year. YSC will do this.**



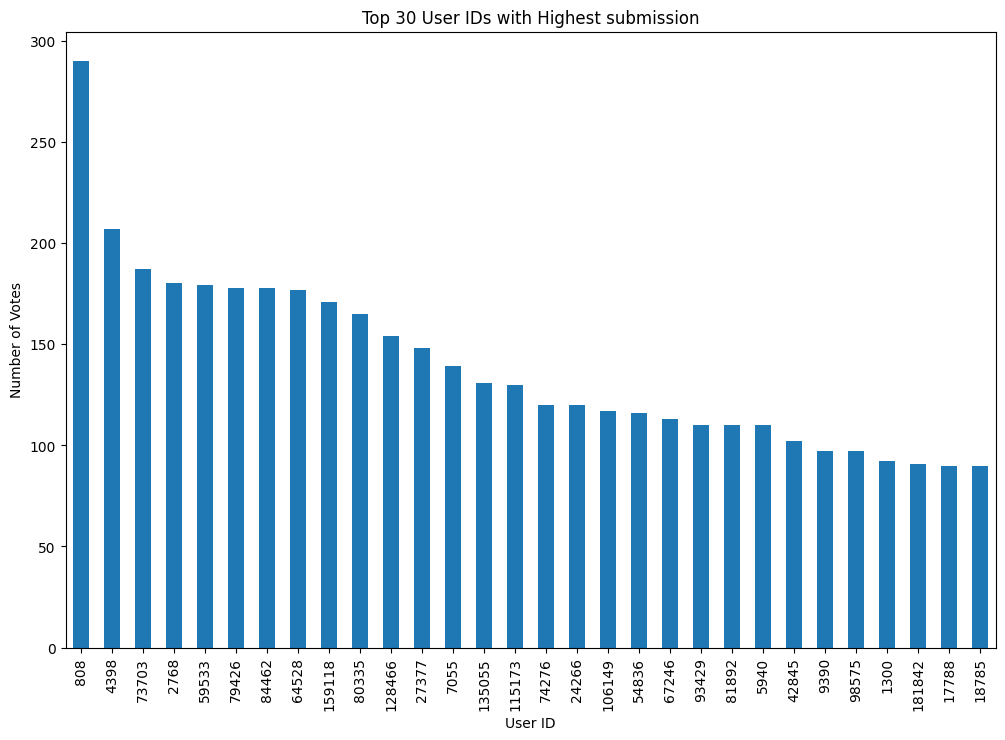
The graph shows the top 20 user id that received the highest votes in the forums. This is the count of votes all-time in the data, but we will explore how this distribution changes conditioned on other factors.

* **How have the highest vote counts changed over time? YSC**



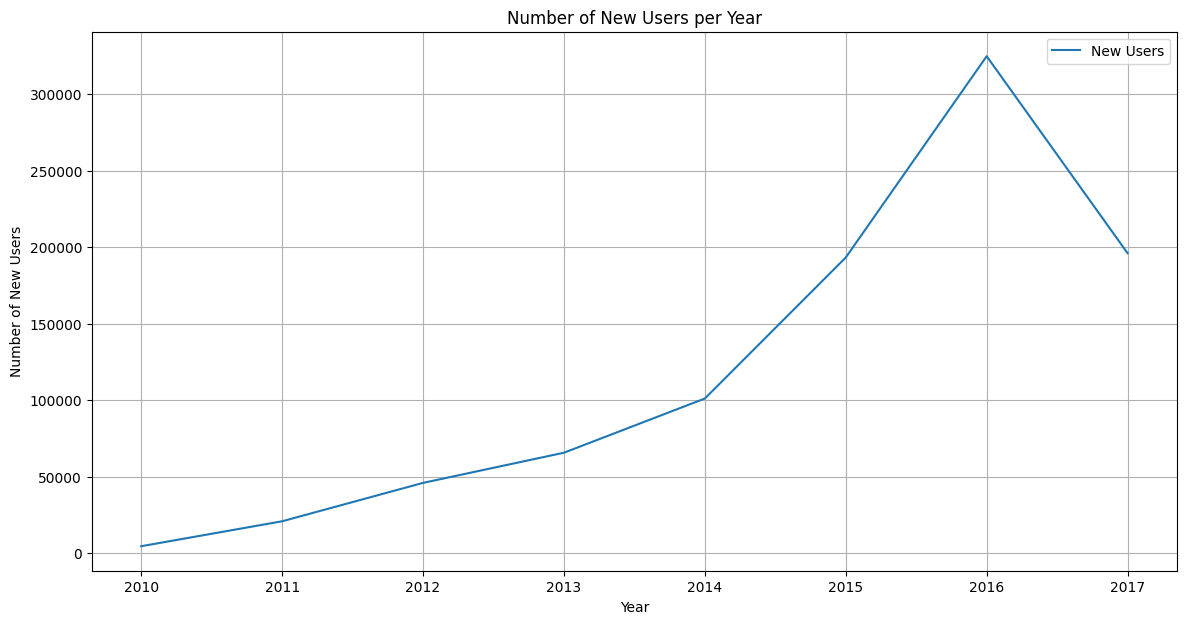
The graph shows the top 20 user id that received votes and how their relationship changes over time. The graph shows that the total number of votes peaked at around 2022.The graph shows only the top 20 user id but is representative of the overall trend observed in the dataset.

* **Which users made the most submissions and what is its distribution? YSC**



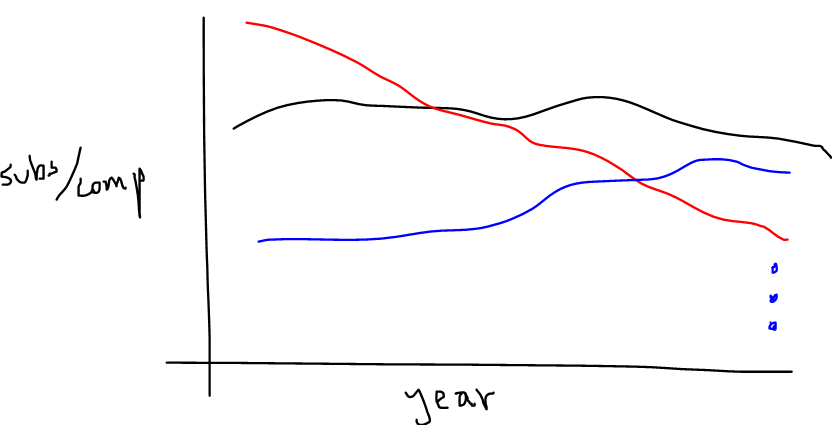
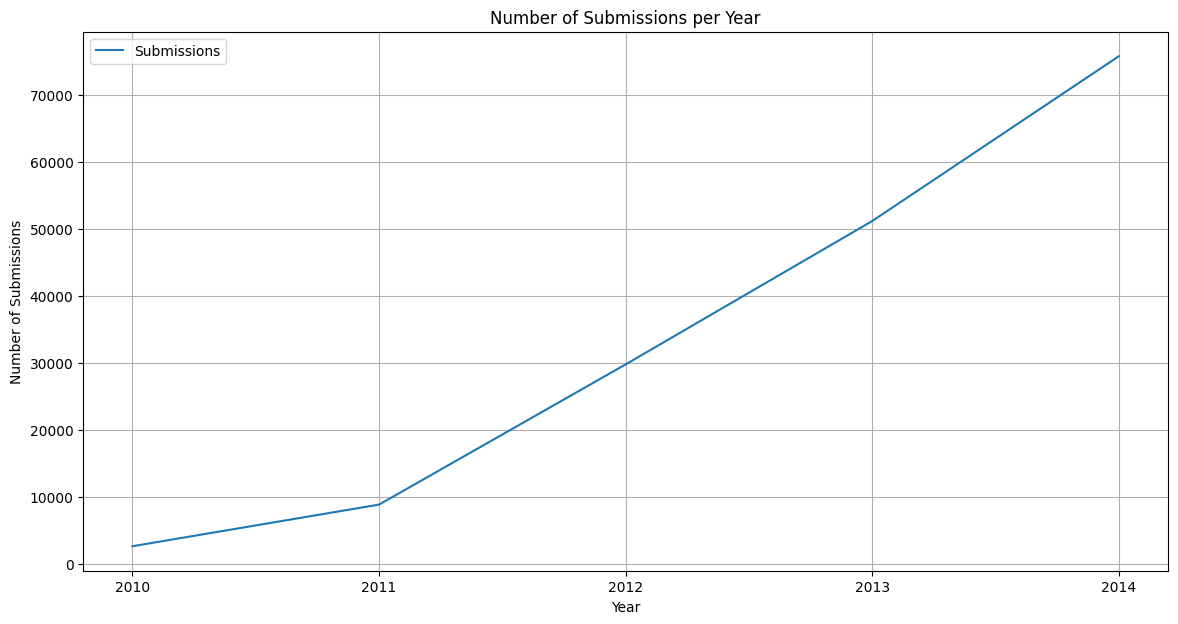
The graph shows the user id that made the most submissions irrespective of the deadline arranged in descending order of count.

* **How do the number of new users change per year? This looks OK. Kabir**



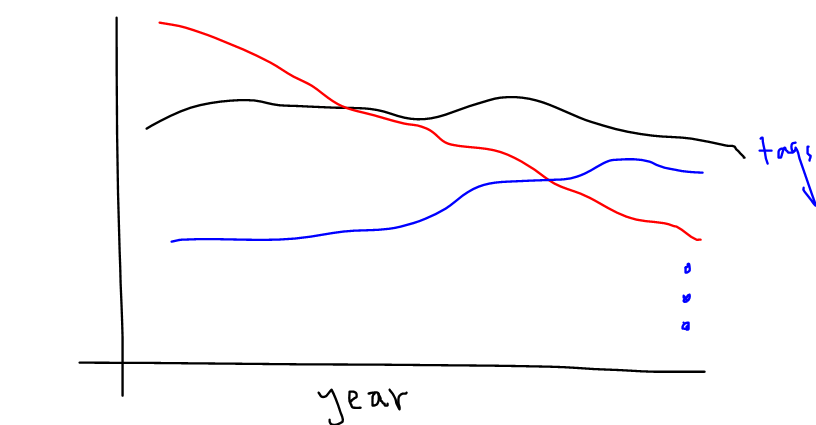
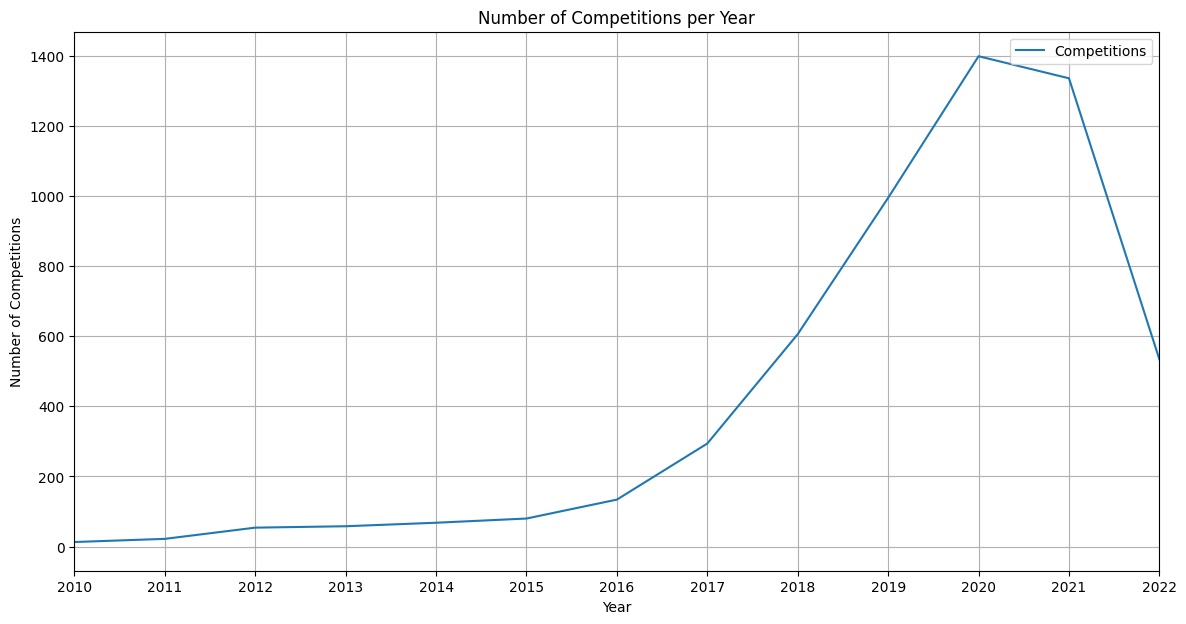
The number of new users steadily increased per year with a sharp increase in 2016, followed by a small dip.

* **How do the number of submissions change per year? Let’s plot the number of submissions per competition over the years for the top X competitions (line graph). Ishika.**

****

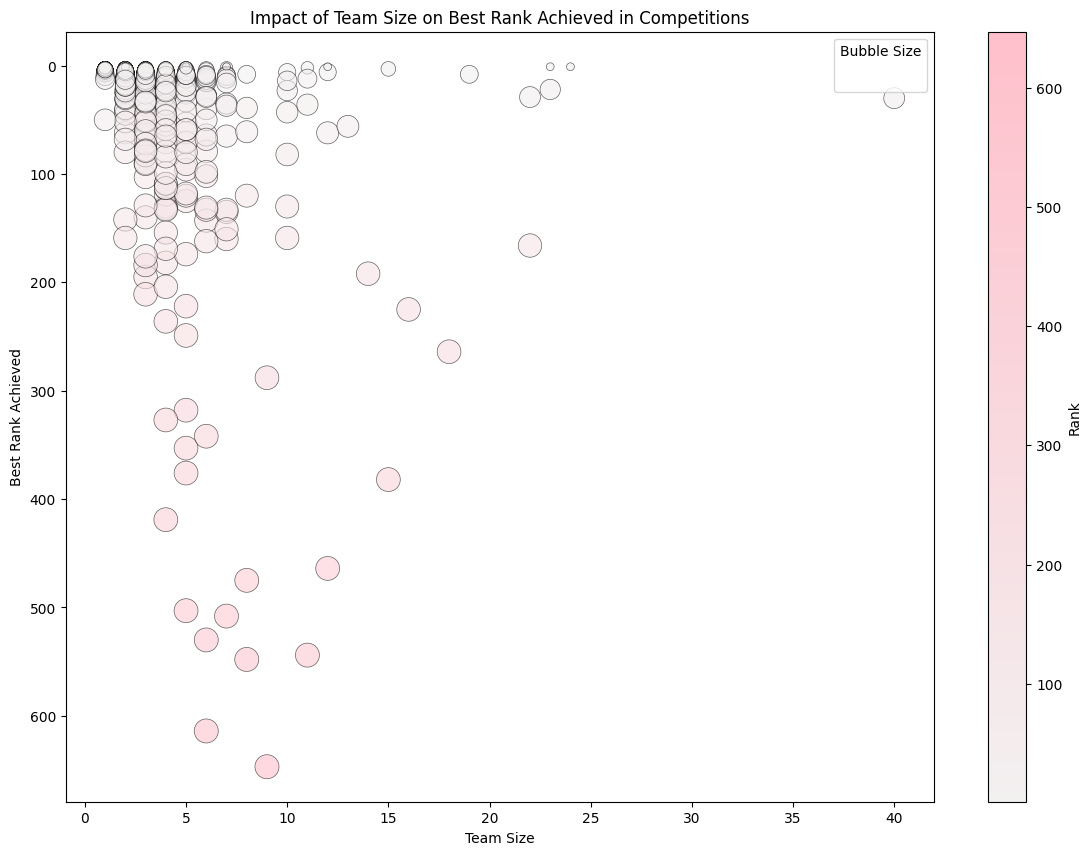
We observe that the number of submissions increase per year at a rapid rate. We can also infer that there is a growing interest or popularity in Kaggle competitions over the years.

* **How does the trend of the number of competitions created per year change? Let’s create a line graph that plots this for the top X competition tags. Are there certain tags that got more popular over the years? Kabir**



The trend in the number of competitions created per year shows a significant increase from 2010 to 2019, indicating a growing interest and possibly the expansion of the organization or platform hosting these competitions. However, there is a sharp decline in the number of competitions from 2019 to 2021.

* **How does collaborating on competitions (team size) impact the likelihood of winning or achieving top rankings?**



Smaller teams are more frequently achieving higher ranks in competitions, as evidenced by the cluster of bubbles at the top of the chart for team sizes less than 10. As team sizes increase, the likelihood of achieving top ranks decreases, shown by the placement of bubbles towards the bottom of the chart for larger teams. The chart suggests that in this data set, there is a trend where smaller teams are more likely to win or achieve top rankings.

**Importance of this project :**

This project holds significant importance in the realm of data science and machine learning for several reasons. First of all, it provides a thorough examination of how Kaggle competitions have changed over the previous ten years, as well as the dynamics of the communities that have been impacted. Analysis of team dynamics, dataset popularity and competition trends can yield important insights into how data science techniques and practices have evolved. Additionally, the project advances algorithmic innovation and AI-ready data development by fostering a deeper understanding of cooperative efforts, learning patterns, and winning strategies within the Kaggle community. The results of this project may also influence the layout of upcoming contests, promote knowledge sharing, and stimulate fresh approaches to data science research.

Overall, by delving into the Meta Kaggle and Meta Kaggle Code datasets, this project provides an essential investigation into the history, present, and future of data science and machine learning on Kaggle, providing insightful information for scholars, professionals, and enthusiasts alike.

**#5: Discussion of related work. Here you will have to do a literature/web search to find out what approaches/visualizations already exist. You don’t want to re-invent but to build on existing work.**

Below are visualizations from three papers that we found to be not only interesting and relevant but also to have inspired us to think of new approaches to visualizing the meta-Kaggle data set.

# **Collaborative Problem Solving on a Data Platform Kaggle (2021)**

# Link: <https://arxiv.org/abs/2107.11929> Below is Figure 1 in this paper, which plots the average number of Kaggle followers by user ranking, also known as “PerformanceTier” in the data set, specifically the User.csv file. This is not to be confused with {Novice, Contributor, Expert, Master, Grandmaster} tiers, which confusingly are also called performance tiers. Public documentation on how a user’s tier {0, 1, 2, 3, 4, 5} is determined could not be found.

# 

# This can be enhanced by visualizing how this distribution changes over time; or by visualizing the relationships of other variables against user ranking.

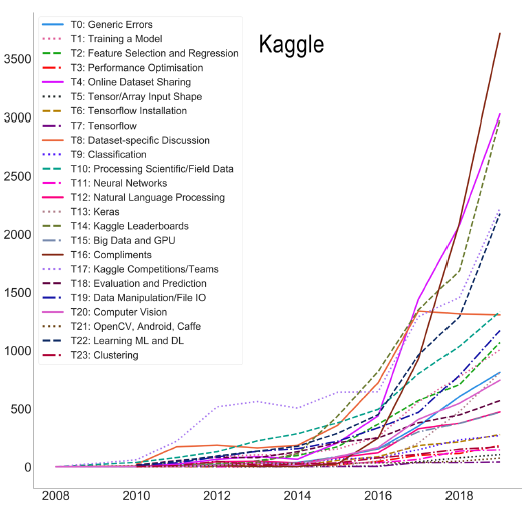
# Below is Figure 2 in the paper, which plots the top-10 frequently used Kaggle data sets.

# 

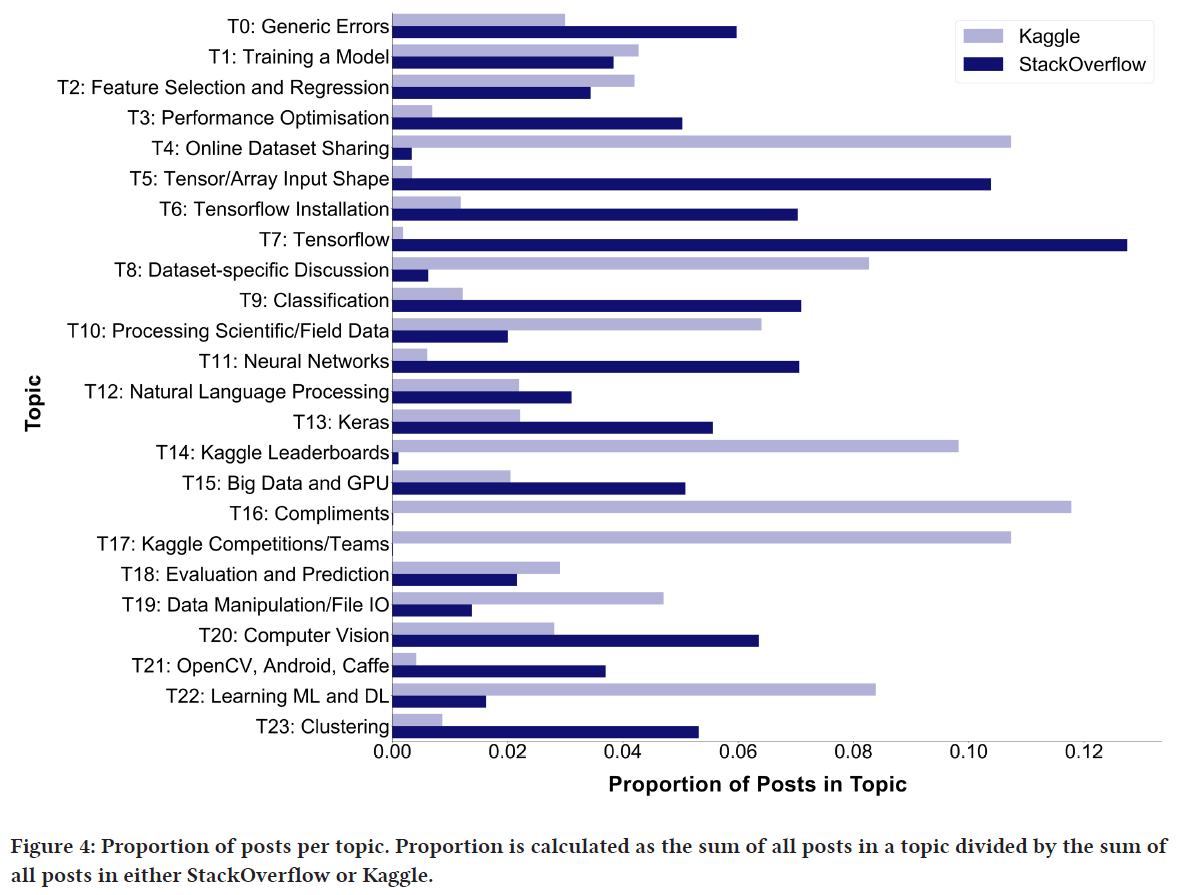
# Prior to viewing this paper, this was one of our visualization ideas, but we are also considering ways to build upon this by plotting the most popular competitions and data related to forums, which are associated with competitions.

1. **StackOverflow vs Kaggle: A Study of Developer Discussions About Data Science (2020)**

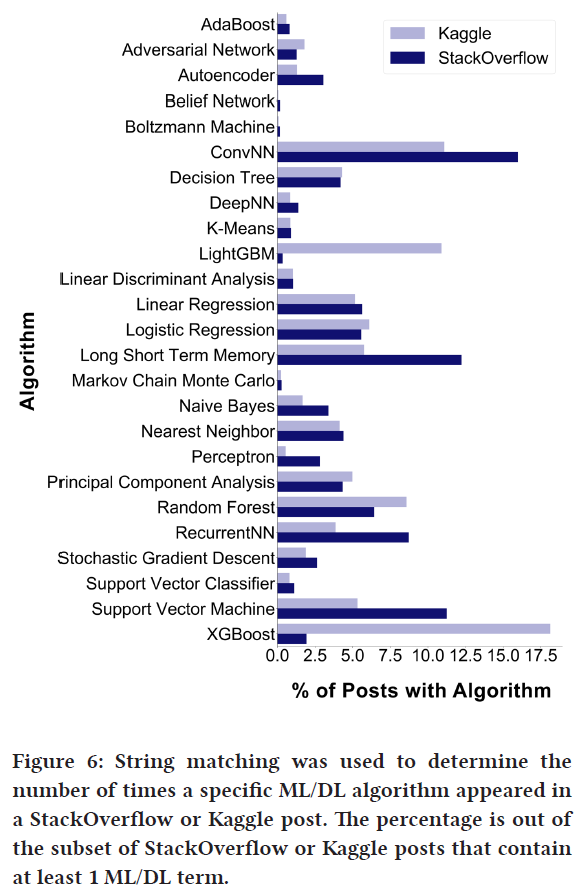
Link: <https://arxiv.org/abs/2006.08334>  
 This paper contains some interesting visualizations regarding the frequency of the most common discussion topics in Kaggle and StackOverflow. Below is a figure in the paper that plots the temporal trend of the topics mentioned in Kaggle (y-axis is number of posts, x-axis is years).



Below is another figure that essentially normalizes the visualization above and plots the proportion of posts associated with each topic.



Similarly, this paper also has the figure below which plots the percentage of posts that mention various machine learning algorithms.

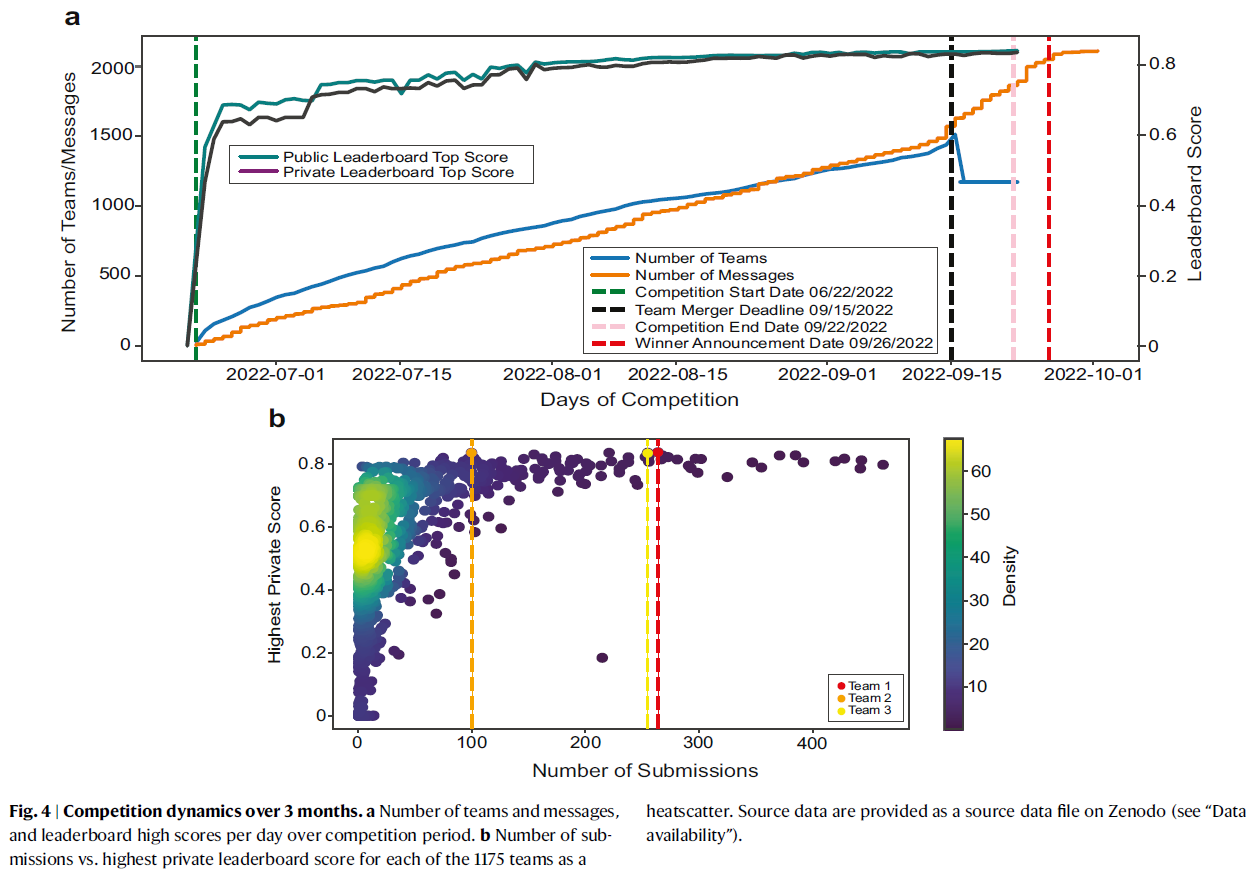


The visualizations in this paper provide us with a new approach to take on the meta-Kaggle data set, which is to visually compare the frequency of key terms related to the data science landscape. We are interested in how these frequencies change with time and use that to compare the popularity of different algorithms, programming languages, libraries/packages, and frameworks. For example, comparing the popularity of TensorFlow vs Pytorch vs Keras over time, or over different user attributes such as tiers.

1. **Segmenting Functional Tissue Units Across Human Organs Using Community-Driven Development of Generalizable Machine Learning Algorithms**

Link: <https://www.nature.com/articles/s41467-023-40291-0>  
 Our project sponsor and Prof. Katy Borner are among the authors of this paper, which analyzes the results of a Kaggle competition pertaining to medical image segmentation.

Below is a figure from the paper that visualizes temporal trends of the quantity of teams and messages; and the leaderboard score throughout the duration of the competition.



The top figure’s use of two x-axis is an effective method of conveying a lot of information. The curves for the leaderboard score gets very close to the maximum (around 0.8) about halfway into the competition. It would be interesting to analyze how long the highest scoring teams took to reach their highest scores along with their score progressions. The bottom figure plots the top three teams’ highest private score against their submission quantities.