Visual Analysis of Meta-Kaggle Data to Uncover Trends in the Data Science Landscape

Yeon-Soo Chung (Team Leader), Kabir Chaturvedi, Bhoomika Pathapati, Monisha Patro, Ishika Thakur

Abstract—Kaggle is a massively popular platform where data scientists of all levels can share knowledge, collaborate on projects and datasets, and enter competitions, some which award prize money. Therefore, general trends within the Kaggle landscape can be viewed as mirroring those in the wider data science profession. In this project, our team visually analyzes the meta-Kaggle dataset [1], which contains data on the major entities in the Kaggle landscape, such as competitions, submissions, and their categories (called tags); users, teams, kernel (code notebooks that users can create, run, and share on Kaggle), forums, and more. With visualizations, we attempt to uncover trends and insights in this dataset. For example, we examine how competitions, submissions, and kernel tags have changed over the past ten or more years. We also analyze the popularity of widely used data science packages and techniques over time via how often they are mentioned in forums and by counting kernel tags. In addition to providing a window into trends among more seasoned data professionals, the results of this project can provide pedagogical insights, as Kaggle is where many people start their data science journeys.

Introduction

We began this project with a literature review to study similar work done on the meta-Kaggle dataset and their relevant visualizations. A paper that studies data science-related discussions among developers on StackOverflow and Kaggle produced some interesting visualization [2]. Below is a figure from this paper that plots the temporal trends of common topics mentioned in Kaggle from 2008-2019. The authors of this paper determined the topics by implementing a Latent Dirichlet Allocation procedure.

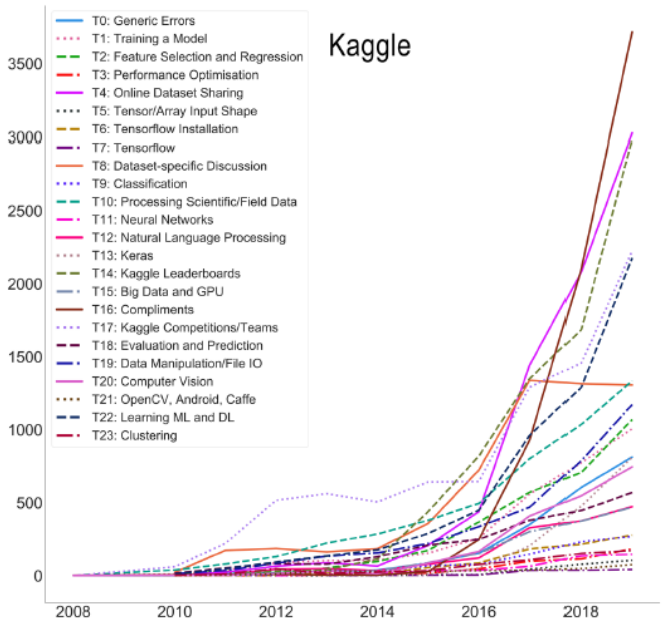

Fig. 1. The number of times major topics were mentioned (y-axis) on Kaggle each year (x-axis).

This paper also produced the following two figures. The first one plots the proportion of posts for each topic. The other one plots the percentage of posts that mention various machine learning algorithms (both are based on 2008-2019 timeframe).

Fig. 1-3 inspired us to take similar approaches in our project. Most of our visualizations convey temporal information on how certain frequencies changed over the years. We analyze the temporal counts of variables such as submissions made to competitions, mentions of key terms in forum posts, and kernels with specific tags.

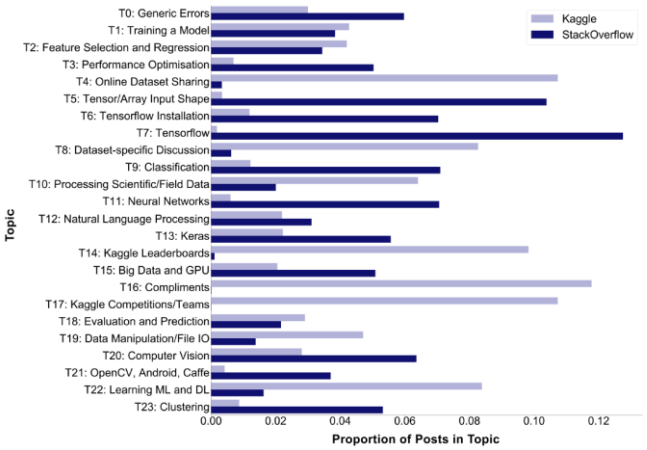

Fig. 2. Proportion of posts per topic. Proportion is calculated as the sum of all posts in a topic divided by the sum of all posts in either StackOverflow of Kaggle.

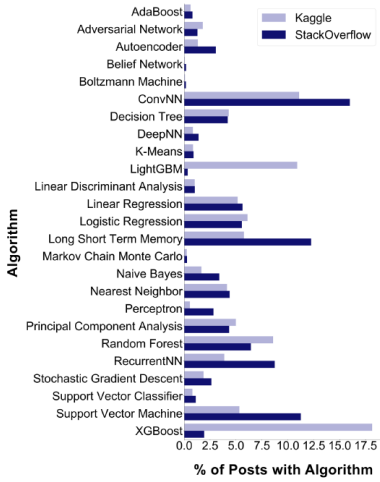

Fig. 3. String matching was used to determine the number of times a specific ML/DL algorithm appeared in a StackOverflow or Kaggle post. The percentage is out of the subset of StackOverflow or Kaggle posts that contain at least one ML/DL term.

1 Background

The Kaggle download contains 32 tables (csv files), or entities, though we did not use all of them. Below is a summary of the data that we used:

- There are 5,662 competitions recorded. Its major attributes are competition ID, title, forum ID, Organization ID (organizer of competition), Enabled Date, Deadline Date, Reward Type, Reward Quantity, Total Teams, Total Competitors, and Total Submissions.
- There are 13,752,749 datasets recorded. Major attributes are Dataset ID, Forum ID, Creation Date, Total Downloads, Total Votes, and Total Kernels (kernels are notebooks that allow users to write, run, and publicly share their code on Kaggle).
- There are 319,073 forums; 2,169,750 forum messages; 2,921,254 forum message votes; and 379,761 forum topics. Users can create discussion forums for various topics or reasons, such as questions related to competitions or datasets, technical advice or opinions, feedback on using Kaggle, etc.
- There are 1,081,172 kernels.
- There are 13,752,749 submissions to competitions. Major attributes are submission ID, Team ID, submission Date, and whether or not it was made after the competition deadline.
- There are 816 tags, which are essentially categories and subcategories that competitions and data sets fall into.
- There are 7,034,741 teams. Its major attribute is Team ID, which is also an attribute in the submissions table; and competition ID, which is an attribute in the competitions table. These identifiers can be used to join tables, resulting in a merged table conducive to creating certain visualizations.
- There are 17,005,031 users. Major attributes are user ID, date of registration, and performance tier. User ID is used in other tables and can be used to join or merge tables.

Our visualization tool of choice was Python, specifically its Matplotlib and Seaborn libraries. For the most part, the raw data was in a clean format. There were some missing values, but they are a very small minority, and deleting those rows from the tables had negligible impact on our analyses. Please see our supplemental Python script for the code that processes our dataset and creates our visualizations.

2 Results and Analysis

Due to the nature of our project and data, our findings are mostly visualized as bar plots along with some line charts and tables. Thus, our main graphic variable is the positioning of our graphic symbols, which are mainly points, lines, and (bar) areas. Text and numerals were also used to describe and quantify the visual information. Bar plots were chosen due to their simplicity and ease of visual perception. We have color-coded them by what the y-axis counts, introducing a graphic variable for aesthetic reasons. Line charts were created to visualize temporal trends.

2.1 Competition Submissions

We began with a bar plot of the top ten competitions by the "total submissions" column in the Competitions data table (Fig. 4). These quantities are the number of submissions made to competitions **before** their deadlines.

Unsurprisingly, all of these competitions were awarded monetary prizes, resulting in them having the most pre-deadline submissions. It is interesting that four of these ten have to do with financial predictions such as predicting risk of defaults and fraud. One reason for this may be that the organizations behind these competitions (financial companies and banks) deem it worthwhile to award prize money in exchange for models that improve their performance. In that industry, even a percentage point improvement can have a large impact in their business. Another reason could be that they have a lot of data they are willing to make publicly available for competitions.

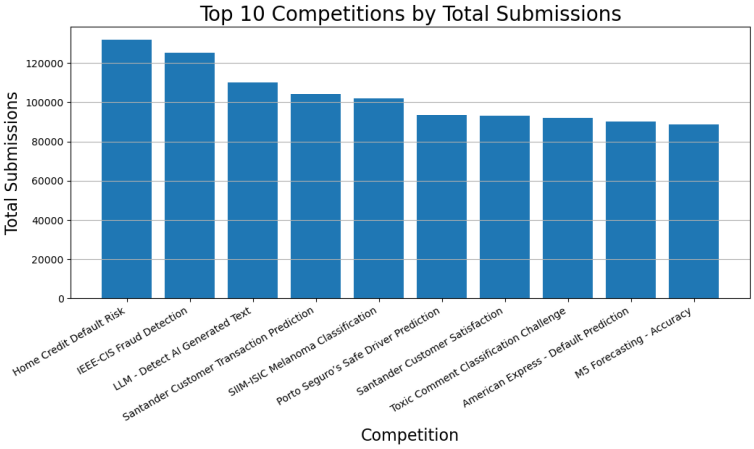

Fig. 4. Bar graph showing the top 10 competitions by total submissions.

We went a step further by adding a temporal dimension to the above result. Fig. 5 visualizes the number of total competition submissions **(before and after deadlines)** each year based on the Submissions table.

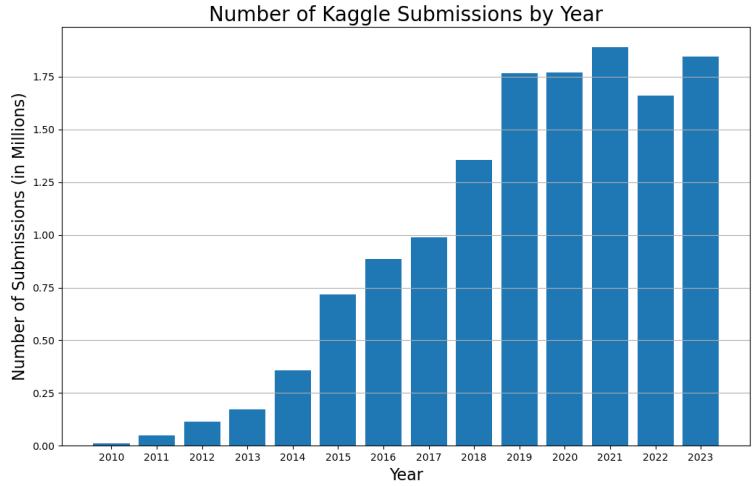

Fig. 5. Bar graph of the number of total submissions to competitions each year.

Many Kaggle users submit their results to expired competitions for learning purposes. The plateauing behavior of the graph above that starts at 2019 is interesting, so it led us to investigate further: how do these numbers break down into pre- and post-deadline submissions? Fig. 6 visualizes this breakdown.

Interestingly, Fig. 6 indicates that the number of submissions made within competitions' open dates began to decrease since 2021, but the number of submissions made after competitions expired plateaued at 2019 but saw an increase in 2023. The aggregation of the two quantities results in Fig. 5's plateau starting in 2019. Despite the small discrepancy between the two trends in Fig. 6, the number of pre- and post-deadline submissions over the years are highly correlated (correlation coefficient of 0.9366).

Fig. 7 plots the number of competitions created each year. It is striking to see how the number of new competitions decreased so much in recent years, yet the number of submissions in Fig. 5 and Fig. 6 did not decrease accordingly. This suggests that, compared to older ones, recent competitions are seeing many more submissions before they expire. This, and the growth of post-deadline submissions shown

in Fig. 3, indicate that interest in data science has exploded in recent years.

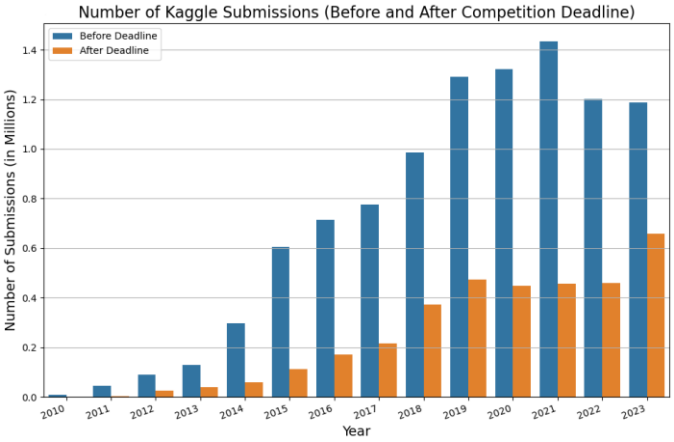

Fig. 6. Clustered bar plot that breaks down the total submissions bars in Fig. 5 into submission made pre- and post-competition deadlines.

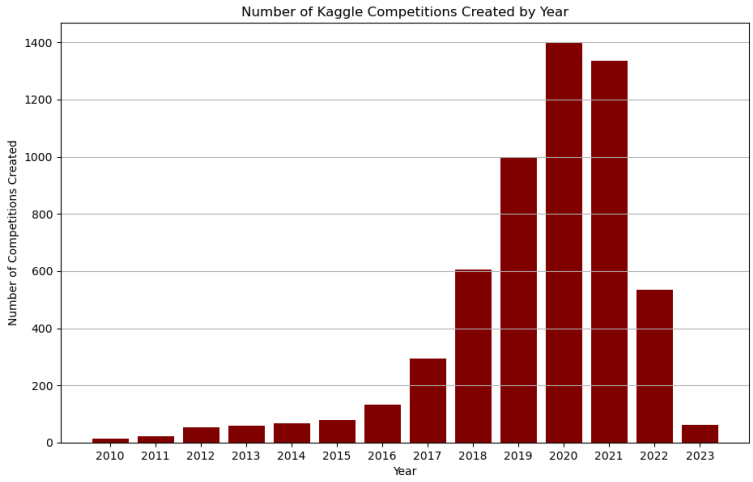

Fig. 7. Bar plot of the number of new Kaggle competitions created each year.

An interesting insight can be gained from plotting the top 10 competitions ranked by post-deadline submissions (Fig. 8).

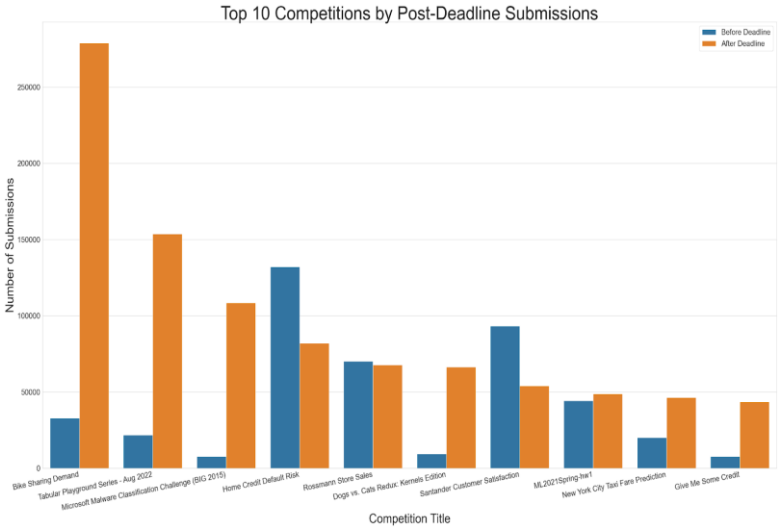

Fig. 8. Clustered bar plot of competitions with ten most post-deadline submissions. For reference, the pre-deadline submissions are also plotted. Please see our Python script for larger image.

In Fig. 8, the top two competitions are the "Bike Sharing Demand" and the "Tabular Playground Series – Aug 2022" competitions. Based on our experiences studying data science, the bike sharing project is often used in classrooms and online tutorials, but to see how imbalanced the pre- and post-submission quantities are is striking, and it demonstrates its value as a pedagogical tool. It is interesting to note that there are two competitions with more post-deadline submission than pre, but their imbalance is much greater than competitions with more pre-deadline submissions, which are more common.

2.2 Popularity of Python and R; Deep Learning Frameworks; and R Packages

We analyzed the popularity of Python and R among Kaggle users by counting the number of kernels that were run on each language every year. Fig. 9 is a line chart of these quantities.

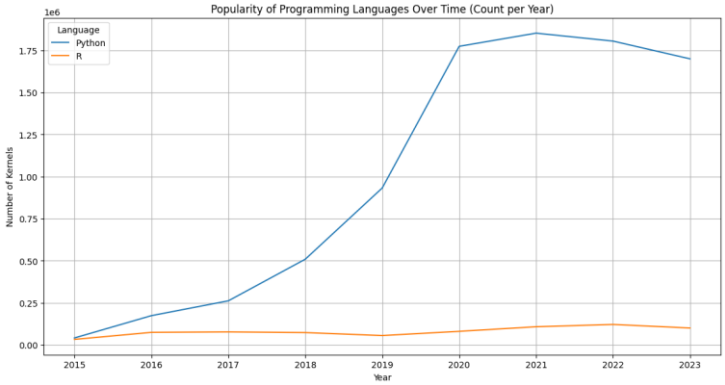

Fig. 9. Line chart that plots the number of Kaggle kernels based on Python and R.

Not surprisingly, Python is a much more popular language than R in the Kaggle landscape most likely due to its machine learning capabilities, while R is much more commonly used for statistical analysis. R's kernel usage is quite consistent over the years, while Python's popularity grew rapidly from 2015 to 2020, before levelling off and decreasing slightly since 2021. Note that Python and R are the only languages supported by Kaggle's kernels.

Deep learning has been gaining popularity over the past decade due to its cutting-edge nature and its applications increasingly becoming a part of everyday life. Keras, TensorFlow, and Pytorch are the most widely used deep learning libraries or frameworks today. The visualization below plots the number of forum posts that mention each library each year.

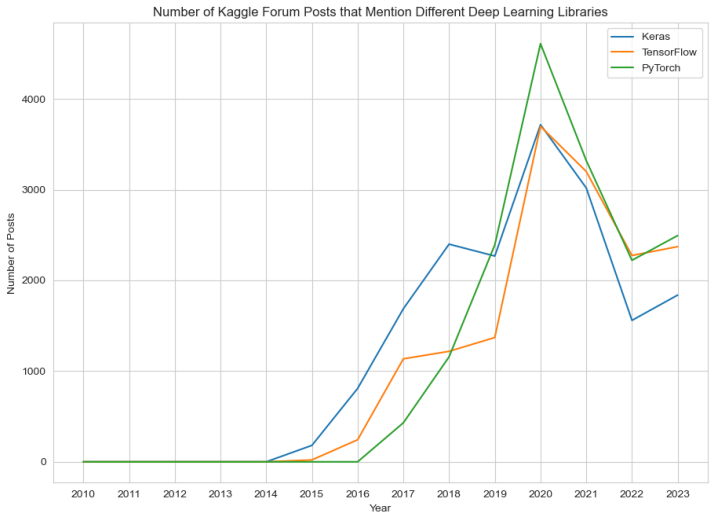

Fig. 10. Line chart of the number of forum posts that mention Keras, TensorFlow, and PyTorch.

In the context of forum activity, Keras and TensorFlow were the most popular libraries from 2014 to 2018, but PyTorch rapidly surpassed them in 2019. The spike at 2020 for all three libraries may have to do with the Covid-19 pandemic - people, including Kaggle users, spent much more time at home, so they were probably more active on Kaggle and its forums. For some reason, PyTorch was mentioned more in 2020, but since 2021, PyTorch and TensorFlow are in a dead heat.

We applied a similar word-counting algorithm to tidyverse, ggplot2, and dplyr; which are three of the most commonly used R packages.

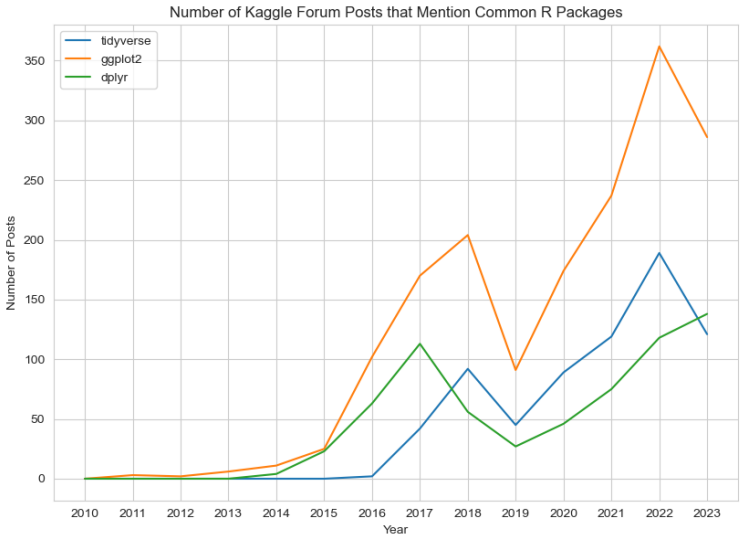

Fig. 11. Line chart of the number of forum posts that mention tidyverse, ggplot2, and dplyr.

Tidyverse is a package that enables a user to transform data tables between different formats, such as pivoting and unpivoting them. Ggplot2 is a visualization package, and dplyr is used for data manipulation. Ggplot2 is the most mentioned, and the other two are mentioned at similar rates. It is interesting to see that all of them, but more so tidyverse and ggplot2, share similar trends and spikes. The dip in 2019 corresponds to the very small dip in Fig. 9's curve for R.

2.3 Data Science Techniques

There are many tags in the meta-Kaggle dataset. Recall that tags are categories which many competitions, datasets, and kernels are tagged with. The tags can be categorized into parent tags and sub-tags. The parent tags, in descending order of how many sub-tags they are associated with, are architecture, language, task, packages, technique, admin, sports, environment, geography and places, and data type. We were interested in exploring the technique parent tag, which has 37 sub-tags, such as data visualization, feature engineering, NLP, random forest, time series analysis, SVM, AutoML, etc. To measure the popularity of these techniques over time, we found the number of kernels tagged with these techniques, for each year. We plotted this in Fig. 12, which is on the last page of this report so that it can be enlarged to be as wide as a page.

Most techniques have similar trends that increase sharply in 2020 and 2021 before decreasing. We believe that a major factor for this is the Covid-19 pandemic, leading to higher activity on Kaggle. It is interesting to note how the techniques' rankings in 2023 differ from the overall rankings. In 2023, NLP, random forest, computer vision, and feature engineering rounded out the top four techniques by kernel count in Kaggle. Feature engineering is generally considered to be a crucial step in the machine learning pipeline. NLP and computer vision are two of the most popular machine learning applications. Random forests appear to be the most commonly used non-deep-learning machine learning algorithm, suggesting it to be one of the main algorithms of choice among data science practitioners.

AutoML, which stands for automated machine learning, is a process that automates the time-consuming steps of the machine learning pipeline, such as data cleaning, hyperparameter tuning, model selection, etc. [3]. It aims to make machine learning more applicable for non-experts. There are various tools and services available for this, such as H2O AutoML, Auto-sklearn, Auto-PyTorch, MLBoX, and more [3]. In Fig. 12, AutoML deviates from the common trend by increasing sharply from 2021-2022, before sharply decreasing in 2023. We speculate that when data scientists first became aware of it, there was a lot of activity in experimenting and exchanging ideas with it, but then, after 2022; and after learning about its capabilities and limitations, they may have decided that, at the moment, AutoML generally does not offer sufficient value to prefer it over traditional (more manual) ML approaches. The increase in 2022 (post-Covid lockdown) followed by the huge drop in 2023 leads us to suggest this reasoning. Thus, these temporal trends also suggest techniques and practices that cause short-term excitement. It would be interesting to monitor how these trends continue in the future.

3 Conclusion

In this project, we focused on temporal analyses of the meta-Kaggle dataset through various approaches: competition submissions; and popularity of programming languages, common data science packages, and data science techniques.

We visualized the explosive growth of interest in data science while uncovering some interesting trends. Although the number of competitions created in 2023 dropped to its lowest level since 2015 (Fig. 7), the number of submissions surrounding them has not dropped significantly from its peak (Fig. 5-6). Among the three major deep learning frameworks, PyTorch and TensorFlow are evenly mentioned in Kaggle forum posts in 2023 after a few changes in rankings (Fig. 10). Although Python remains the language of choice among data scientists, R sees consistent usage, and three of their most widely used packages are seeing a general upward trend in forum post mentions (Fig. 11).

In addition to visualizing continued growth, we found that trends can suggest fad-like behavior, such as the use of AutoML (Fig. 12). It appears to have experienced extensive usage in Kaggle before dropping sharply, indicating its short-term hype. It would be interesting to see how such trends evolve next.

Due to its massive popularity and widespread use, general trends within the Kaggle landscape can be viewed as mirroring those in the wider data science profession. Therefore, we believe that some of our findings hold value in terms of the most salient languages, packages, and techniques being used today; as well as what kind of competitions organizations should create to maximize participation. Please reference our supplemental materials (Python script and full-length report in the course drive) for more visualizations and comments.

Acknowledgments

We wish to thank our project sponsor, Yashvardhan Jain, for the opportunity to work on this project. We also appreciated the feedback provided to us by our peer reviews.

References

[1] Kaggle. (n.d.). *Meta Kaggle*. https://www.kaggle.com/datasets/kaggle/meta-kaggle

[2] Hin, D., (2020, June 6). *Stack Overflow vs Kaggle: A Study of Developer Discussions About Data Science.* arXiv.org. https://arxiv.org/abs/2006.08334

[3] AutoML. (n.d.). https://www.automl.org/automl/

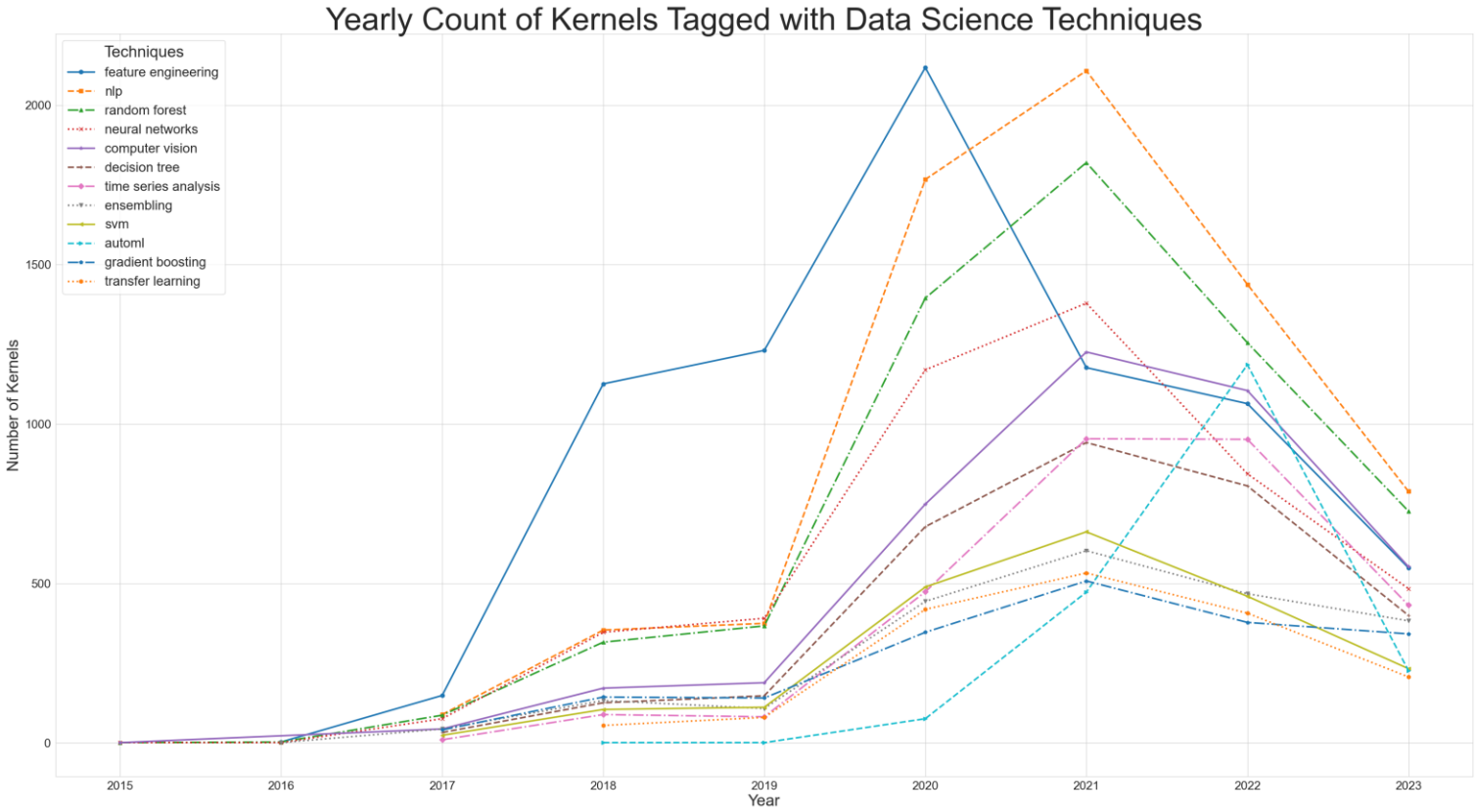

Fig. 12. Number of kernels tagged with common data science techniques each year.