Kaggle Data Community Survey

Data summary/analysis report

Authors: Laura Koesten, Jude Yew, Kathleen Gregory

Abstract:

Having greater access to data leads to many benefits, from advancing science and promoting transparency and accountability in government to boosting innovation. However, merely releasing the data does not make it easy to use; even when the data is openly available online, people may struggle to work with it. We aim to understand what makes some data more reused than other data, through the lens of one of the largest data-sharing platforms worldwide, Kaggle. This report presents summarised findings from an online survey taken by 434 active members of the Kaggle community in February 2021. We identify several factors that demonstrably support data use, which are related to the data itself, but also factors related to how people engage with the data. Key findings highlight the importance of textual descriptions of the data, related to `understandability' which is perceived as a key dimension of data quality. Our insights can inform the design of data platforms, in areas such as community building and user retention, and also support data publishers in prioritising data maintenance work.

1. Introduction:

In this report we describe a study of the online volunteer data-sharing community on Kaggle, currently the biggest, most popular platform of its kind with more than eight million registered users and more than 170,000 datasets (as of 2022). We explore how the community interacts with data that is publicly available through a survey taken by 434 Kaggle community members in February 2021.

Our aim was to conduct a preliminary investigation of data practices on Kaggle and gain insights into what makes a dataset more likely to be used by others. We see two relevant types of factors: some related to the data itself, such as the availability of text descriptions of data values and attributes and data quality conceptions; and other factors related to the means available to engage with the data, such as discussion forums, as well as co-located tools and tasks.

2. About Kaggle:

Founded in 2010, Kaggle has grown to more than 5 million registered users at the time of the survey. The platform initially hosted supervised machine learning competitions before turning into a data science community, where people routinely share datasets, code, and engage in discussions. Users search and publish datasets, explore and build models collaboratively, and enter competitions to solve data science challenges, some with significant cash prizes.

More than 170,000 public datasets of different formats and more than 3000 competitions have been hosted on Kaggle (by 2022). At the time of the survey, any Kaggle user could create a competition, ranging from educational to high-profile public challenges such as the COVID-19 pandemic. The platform offers integrated notebooks, which are a virtual environment that allows users to create and share code, data or text for collaborative data analysis in the cloud. In order to find datasets users can browse and filter the collection of datasets according to file size and type, license, and topical tags, as well as search via a query in a search-box[1].
We consider Kaggle an interesting case study for several reasons:

- It is the largest platform of its kind.
- Equally, it is host to a substantial data science community of practice, beyond competitions.
- It directly supports data work on the platform itself (Wang et al., 2021). Other platforms such as GitHub or Zenodo have a distinct focus i.e. code or data publishing.
- It implements a diverse range of user-centric features, which prior literature link to data use (Birnholtz & Bietz 2003, Koesten & Simperl, 2021).

3. Methodology:

We designed and tested the survey questionnaire iteratively, including feedback from the Kaggle team, and piloted it with 12 participants. We edited the wording of questions after feedback from the pilot to ensure understandability. To assure data validity, we filtered pilot responses from our analysis. Response options were randomised where applicable. The question design draws on the findings of earlier work on data discovery. For instance, answer options for selection criteria and quality dimensions for data are informed by Gregory et al.,2020; Kern & Mathiak, 2015; Koesten et al., 2020a). Overall we targeted insights on participants' self-reported interactions and criteria, rather than attitudes.

The survey was organised in three sections:
1) *General information*, aimed at collecting demographic information about participants as well as their experience with data analysis and Kaggle.

2) *Working with data*, covers motivations to work with Kaggle datasets, different aspects of dataset selection criteria as well as specific interaction behaviours with datasets.

3) *Competitions*, focuses on how people organise during competitions.

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

In total the survey contained 26 questions (half of which were single-choice, 5 multiple-choice, 3 Likert-scales, 4 ranking questions, and 1 open question), and can be seen in the supplemental material. To give participants the opportunity to contribute their own input, all questions include an ``other'' option.

The survey was distributed to a randomised subset of 20k Kaggle users who have been active in the past 3 months.Our survey was scripted and administered using Qualtrics (https://www.qualtrics.com/). It took participants a median of 6.75 minutes to complete. There was no incentive to take part. Participants were recruited via Kaggle internal recruitment with the following inclusion criteria : i) over 16 years old, ii) Kaggle dataset user, and iii) active on Kaggle in the past 6 months. The survey was sent to a randomised subset of 40k Kaggle users with a 1,09% response rate and was taken by 434 people in the second half of 2021 between the 27th of August and the 4th of September 2021.

We note that our results are descriptive statistics of our quantitative data, as our survey was not designed to determine statistical significance among different variables. Where we discuss observed trends, these are not statistically significant or generalizable to a broader population. Our results represent the reported behaviours of the individuals completing the questionnaire.

Ethics: The study was approved by and vetted by the Kaggle team and their legal counsel. Participants were shown an informed consent document at the beginning of the survey. They indicated their consent to participate before beginning the questionnaire.

4. Survey respondents: Demographics and experience:

From our sample of 434 respondents, 26% come from India, followed by the US (12%) and Brazil (5%). Participants were generally young, just over 40% are under 30 and another 28% are between 30 and 40 years old. Over 70% have a bachelor’s degree, 35.4% a MSc and >14% a PhD. The top reported job role that emerged was data scientist (26.76%), followed by student (17.95%) and software engineer (9.49%). More than half of the participants reported that they perform data analysis both for work and in their free time (55%), >16% only for work and <25% in their free time. Age range, nationality, and education level match what we know from the Kaggle Machine Learning and Data Science from the same year, with only a slightly older participant age range which is an indication that the composition of our respondent sample might be similar to larger survey[2] participation on the platform. Our survey results add to Kaggle’s internal survey by giving insights into the data work practices of more than 400 active Kaggle users.

We asked respondents how long they have been actively doing data analysis (n=387). Most reported being relatively junior; more than 60% have been practising data analysis for less than 3 years and only 20% for more than 5 years. The picture painted of the Kaggle community is made up of younger, early-career folk or students. This reflects Kaggle’s own

[2] https://www.kaggle.com/c/kaggle-survey-2019

advertising as a space where one has access to ``all the code and data you need to do your data science work”. From our participants just under 84% have downloaded/used/uploaded a dataset in the last year, another 10% in the last three years, which means the majority are active users of Kaggle datasets.

5. Survey Results:

We present the results from our online survey in three sections structured according to our research questions: i. interaction with datasets; ii. criteria for dataset selection; and iii. competitions on Kaggle.

5.1 How do Kaggle users interact with datasets?

5.1.1 Data users on Kaggle

Kaggle represents a unique online community of practice, made of people who are interested in data science. The platform has grown significantly over the past years and is the largest of its kind with over five million registered users at the time.

More than half of the participants (55.3%) reported that they perform data analysis both for work and in their free time, with 16.1% using Kaggle only for work and 24.8% only in their free time. We asked people how long they have actively been doing data analysis (n=387); most respondents reported being relatively junior: more than 60% have been practising for less than three years, while only 22% for more than five years (as can be seen in Figure 1b).

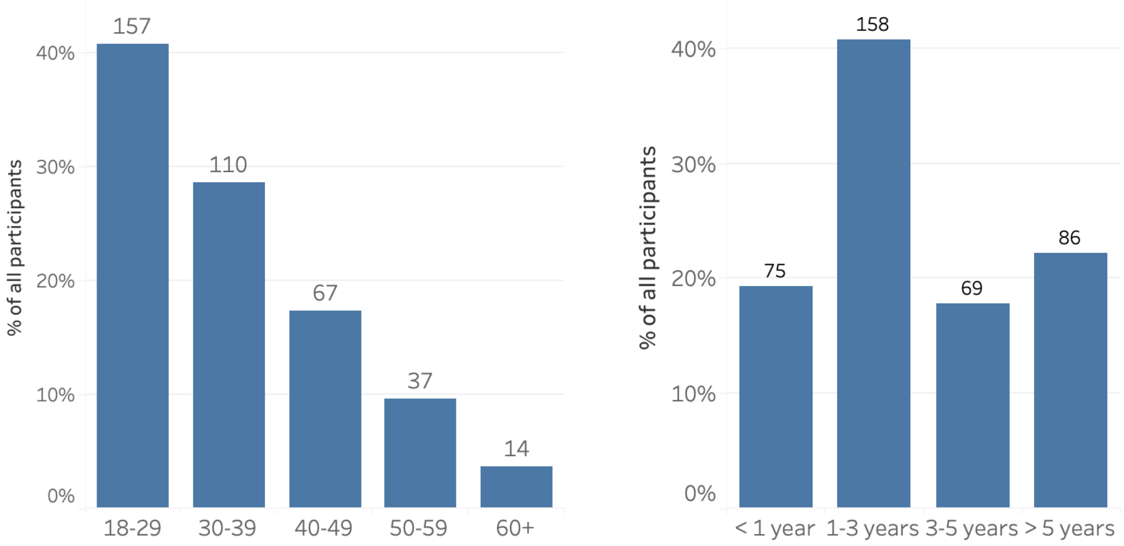


Fig.1a Participant sample. (left) Age range (n=434);
Fig. 1b (right) Experience with data analysis in years (n=388)[3]

[3] The number of survey responses varied with each question, as we did not mandate that all questions are compulsory to answer (except for demographic questions).

The top reported job role was data scientist (26.7%), followed by student (18%) and software engineer (9.5%). Over 85.3% have a bachelor's degree, of those have 35.4% a MSc and 14.5% a PhD. They are generally young, 40.3% are under 30 years old and another 28.2% are between 30 and 40 years old.

The main motivations for working with Kaggle datasets reported by our respondents in a multiple choice question were: learning new skills or methods (83%), solving open (research) problems (52.1%), taking part in competitions (50.2%), followed by fun (45.4%), as shown in Figure. 2.

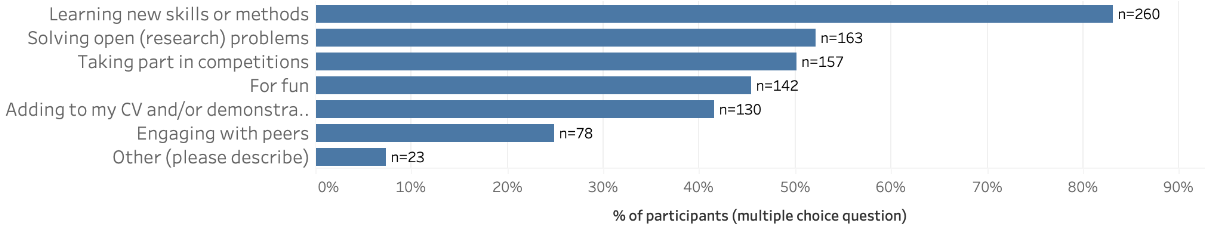


Fig. 2 Motivations for using Kaggle

Adding to their CV or demonstrating skills was mentioned as a motivator to work with Kaggle datasets by 41.5% of the respondents. Those with less than three years of experience chose this option more frequently (73%), which matches our assumption that younger members of the data science community want to gain experience. We also provided respondents with an ``other" choice where they could write free text entries. The common topics of responses (n=23) included: teaching and training, finding datasets as a source, and problem solving.

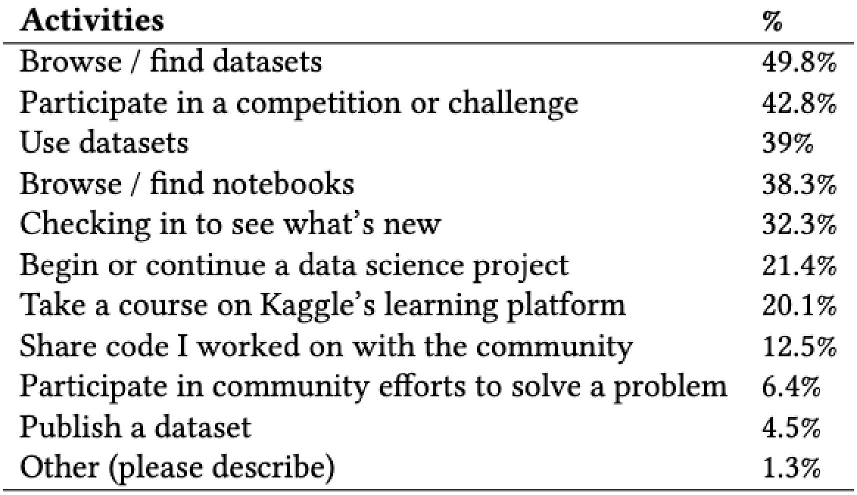
Learning new skills or methods was most prominent in the responses. Respondents also mentioned being able to contribute to the learning experience of others as a motivation, rather than just individual skill improvement, as articulated in these free-text answers:

- “Posting data and notebooks for others to learn”.
- “Demonstrating a concept/method to colleagues on an open dataset”

The desire to help others learn is mentioned in the literature, Hung et al. (2011) also emphasise the importance of reputation feedback in knowledge sharing, which is likely reflected here in the motivator ``demonstrating skills" as mentioned above. This makes Kaggle attractive for use in educational settings and also explains its attraction towards younger, early-career learners.

5.1.2 User activities on Kaggle

When asked about their main activities when using Kaggle, just under half of the survey participants reported browsing or finding datasets. This was followed by participating in a competition or challenge, using datasets and browsing or finding notebooks all mentioned by around 40% of our participants. Checking in to see what’s new was reported by over 30%. (Detailed values can be seen in Table 1)



Activities	%
Browse / find datasets	49.8%
Participate in a competition or challenge	42.8%
Use datasets	39%
Browse / find notebooks	38.3%
Checking in to see what's new	32.3%
Begin or continue a data science project	21.4%
Take a course on Kaggle's learning platform	20.1%
Share code I worked on with the community	12.5%
Participate in community efforts to solve a problem	6.4%
Publish a dataset	4.5%
Other (please describe)	1.3%

Table 1. Three most frequent activities on Kaggle (% of all responses for this question (n=312))
Respondents were able to select more than one option for this question

In the words of a survey respondent, people are on Kaggle:

> "To learn, to explore, to win."

Another respondent highlighted learning explicitly as an activity: to ``see how others solve a problem" in a free text response.

This mirrors the reported main motivation for using Kaggle datasets: to learn new skills, and matches our overall relatively junior participant sample.

5.1.3 Contributions to Kaggle

Over 70% of our participants reported having previously contributed data or code to Kaggle;, however only 9.5% of those stated that they contributed often. More than a quarter indicated never contributing data or code to Kaggle (29.7%), as can be seen in Figure 2. We assume they reflect a user segment that perceives Kaggle mostly as a resource for learning resources, including curated data. This aligns with what we found about user activities earlier. At the same time, it also points to a healthy peer-production community, where a large share of the participants make some contribution to the ecosystem rather than merely using resources created by others (Cheliotis & Yew, 2009).

When asked about downloading datasets (n=306), 66.7% reported downloading datasets either sometimes, often, or a lot before deciding to work with them. Fewer participants (12.4%) stated to *never* download datasets before deciding to work with them.

A minority of participants is actively giving back to the community by re-publishing new versions of data and code after making changes to existing resources. 13.7% of our participants reported to do so often or very often (as shown in Figure 2 (left)). Another 21.9% mentioned to do this sometimes, and a majority of participants stated to rarely or never publish new versions of data or their code on Kaggle (64.4%). When asked whether they share their code elsewhere, respondents dropped GitHub, Collab, and git as examples of other tools, but also friends or colleagues, as free text answers. This suggests the willingness to share data or code via personal connections and hence the importance of community building to establish a prolific data use environment.

The contribution to Kaggle's discussion forums was reported to be similar to publishing data and code. A majority stated to never or rarely take part in discussions (69.6%), 20.3% sometimes, and only 10.1% often or a lot (Figure 2 (right)).

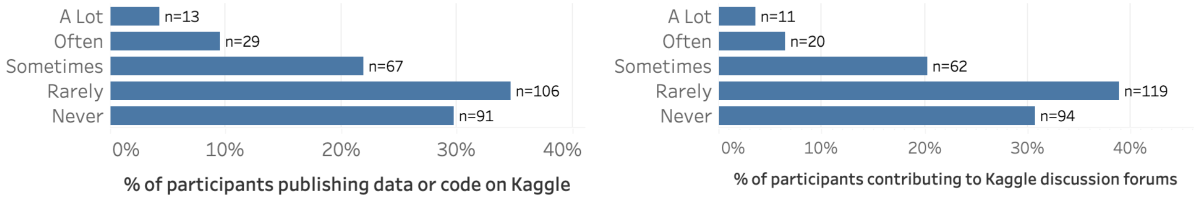


Fig. 2: Re-publishing data or code (n=306); (right) Contributions to Kaggle's discussion forums (n=306)

Overall participation patterns are not dissimilar from other online communities of practice and peer-production systems. A substantial share of the community does more than just consuming resources created by others, though only some go beyond standard activities like competing and sharing data and code by actively improving the resources or discussing with others (Cheliotis & Yew, 2009).

In summary, our results suggest that Kaggle is used primarily for finding and using datasets, alongside competitions and staying up-to-date with data science. More community-based functionalities such as publishing data, code, or discussion comments are used by a segment of the participants only.

5.2 How do Kaggle users decide to use datasets?

5.2.1 Data discovery

Finding relevant data is the first step in using data. In order to find datasets, most participants reported browsing Kaggle (81.9%), while more than half also use search engines (55%), and 28.7% mentioned recommendations from other people as search strategies. Less experienced Kaggle users reported more often relying on peer recommendations; 33% of those with less than 1 year experience mentioned recommendations from other people as a way to find data, opposed to 21% in those with more experience.

We also provided respondents with an ``other" choice where papers and research were mentioned most frequently as resources to find datasets (n=14).

This resonates with literature on dataset search (Koesten et al., 2017; Gregory et al., 2020). The high prevalence of browsing strategies on the platform is likely due to the variety of facets offered to find datasets on the platform, which was also noted in another data-sharing context beyond Kaggle (Ibáñez et al., 2020).

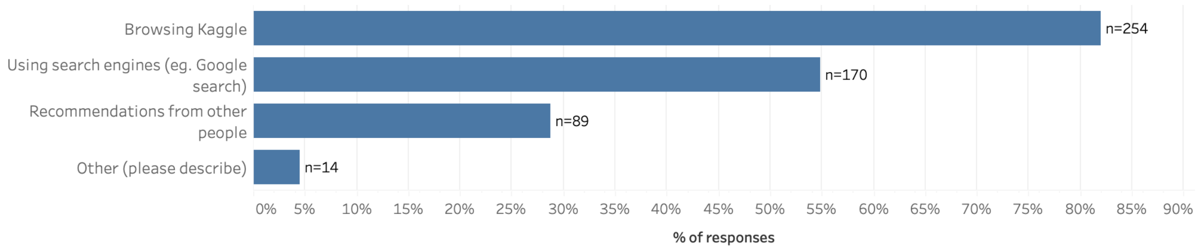


Fig. 3: Mechanisms for finding and choosing datasets (n=434)

On Kaggle, datasets are ranked by aggregated engagement metrics, tagged with topic labels, and displayed by recency. They can be discovered via notebooks that make use of them, via competitions or even via output types such as visualisations. Facets, which are filtering options, afford pivot browsing (Millen & Feinberg, 2006) of datasets on the platform itself, using various criteria to narrow down datasets to work with. For instance, being able to browse for datasets with high numbers of votes or access to notebooks authored by highly rated community members play a role in the selection of datasets. These facets function not only allow easy access and findability, but also become signals for trust and reliability of the datasets. For instance, higher reuse metrics or the usage of particular datasets by popular community members could encourage greater engagement and usage. This has been discussed in the context of other types of online content by Gantayat et al. (2015) and Borges, Hora, & Valente (2016). Similar to these examples, we see the ‘preferential attachment’ (Cheliotis & Yew, 2009) phenomenon in Kaggle where the ability to identify more popular datasets & contributors leads to these datasets & contributors being reused more.

Thinking how to increase votes for a dataset might be a way to actively facilitate reuse via dataset promotion avenues. As seen in other online content-sharing communities the phenomena of preferential attachment, or the ``rich get richer'' or power law effect, results in a few highly popular content that get engaged with the most (Lu et al, 2012).

5.2.2 Criteria for choosing data: Factors related to dataset content

Prior research has shown a variety of factors that users consider when deciding to work with a dataset, such as data provenance, methodology, ease of access, recency of activity, amongst others (Gregory et al. 2020; Wang & Strong 1996).

Our survey asked about dataset selection criteria in different ways: (i) we asked what characteristics of a dataset were important during dataset selection (see Fig. 4) and (ii) we asked whether community engagement indicators (e.g. views, downloads) played a role in deciding to use data (see Fig. 6). We also asked (iii) how respondents determine a dataset’s quality (Fig. 5). The answer options in these questions were informed by research on dataset discovery and data quality (Gregory et al., 2020; Koesten et al., 2020a; Wang & Strong, 1996). For this set of questions, we asked the respondents to rank each option by whether they considered their selection to be ``critically important'', ``important'', or ``not important'', including an additional ``other'' option. Results from these ranking questions are a combined weighted score (not important: 0, important: 1, critically important: 2) and ranked.

Looking at the characteristics of a dataset that are considered important for selecting it (Fig. 4), data quality is seen as highly important for choosing datasets (97.7%). Out of all responses concerning data quality specifically, only 1.7% rated it as not important. As mentioned earlier, we explore the specifics of how data quality is conceptualised in a separate question.

This factor was followed by information about the dataset content, such as a description of the data (92.3%) and of the column headers or variables (87.1%). The next two factors in the ranked list relate to more technical concerns, the size of the data (82.1%), as well as the format or data type (81.9%).

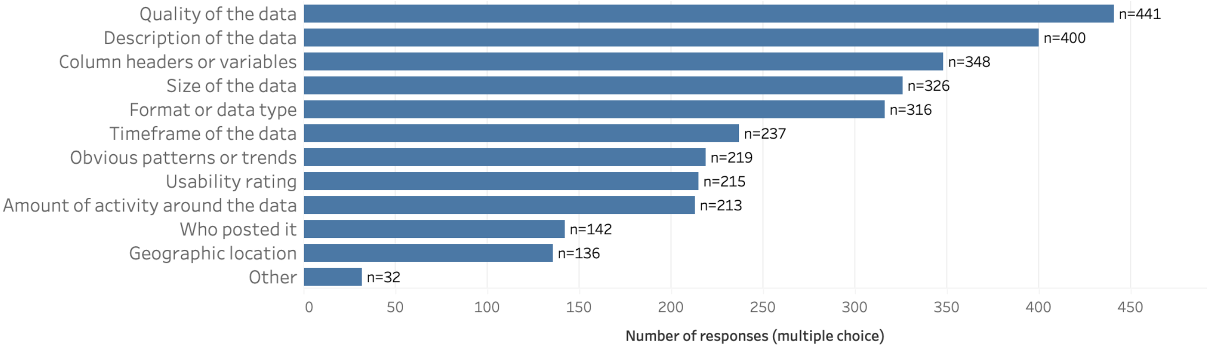


Fig.4: Factors important for choosing datasets (n=293)

The ``other" option (n=32) included topic areas such as openness and licensing, usage ideas, caveats, dataset topic, as well as social importance. One response to the ``other" category mentions EDA (exploratory data analysis) notebooks as a factor, which typically produce statistical and graphical summaries of the dataset.

This emphasises one more time the focus on data descriptions, which we can see reflected in the survey responses regarding criteria for choosing datasets. A free text response mentioned task-specific quality concepts, such as the importance of vocabulary and semantics for NLP tasks. The importance of data documentation has been widely discussed in recent years, moving from a long-standing acknowledgement in information sciences and data archiving to a recent push towards more accountable and user-centred documentation practices (e.g., Sambasivan et al., 2021 & Holland et al., 2018). We observed that the free text responses regarding selection criteria related to the dataset itself, either information that can be derived directly from the dataset, or descriptive aspects about the datasets content.

5.2.3 Criteria for choosing data: Factors related to the perceived quality of a dataset

As noted earlier, in our survey, data quality was considered highly important for choosing a dataset to work with. As the term “quality” encompasses different dimensions we asked specifically how participants determine the quality. The most frequent category considered important by our respondents was understandability (94%), followed by consistency and believability (over 85%), and completeness (76%) (Figure 5). We defined completeness as “not many missing values”. Interestingly, the provenance of the data (the individual or organisation who published the data) ranked last in the context of data quality, in contrast to literature (Herschel, Diestelkämper & Lahma, 2017; Koesten et al., 2020a).

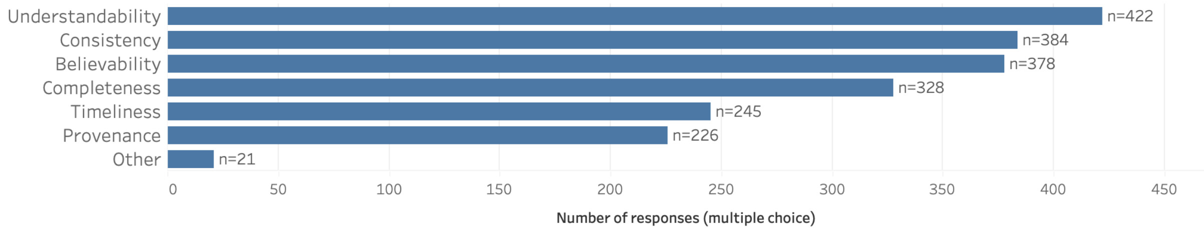


Fig. 5: Dataset quality criteria (n=287), Ranking based on combined weighted score (not important: 0, important:1, critically important:2)

High-quality data is perceived as ``understandable" data, as well as ``consistent" and ``believable" data. A participant described their expectations of data quality in a free text response as:

> “A well written and complete dataset description with links to the data source and to relevant papers.”

Perceptions of data quality, referenced in the free text responses, are often influenced by factors beyond just the data -- with participants paying attention to ``metadata" such as data descriptions. Additionally links between the data and discussions and usage examples, help participants assess the quality of the dataset being shared.

The ``understandability" of data is seen by our respondents as an aspect of a datasets' quality. This has repercussions for how we make data ready for reuse, considering accessible documentation or supply of contextual information, e.g. through access to the community via discussion forums and usage examples in public notebooks.

Data provenance did not show to be a strong factor for selecting datasets by our respondents. This is in contrast to literature on data quality and selection criteria that shows provenance can give an indication of the authoritativeness, trustworthiness, context and purpose of a dataset (Moreau & Groth, 2013). One reason for this might be due to the learning focus of Kaggle users, where aspects such as provenance in the sense of trustworthiness are naturally not as important. Another reason might simply be the availability of other quality signals, such as popularity metrics indicating reuse, discussions, etc; which are not always present on other data platforms.

The definition of provenance also varies in the literature, often not just including information about who published it and its origin (as defined in our survey), but also contact points or a dataset's traceability, which describes workflow provenance through version indicators and version histories (Missier et al., 2012; Kim et al, 2008) or more broadly systematic access to the entire lifecycle of a dataset including sociotechnical context (Victorell et al., 2020; Koesten et al., 2021). Furthermore, every dataset in Kaggle is associated with an ``owner", giving access to their previous activities and interactions, and ultimately reputation. Perhaps changing perceptions of trust and authority also play a role here -- especially when all interaction with the data happen in the cloud. If one does not have to download the dataset, then worrying about corrupted data would not be foremost on the users' mind.

We note that our earlier depiction of Kaggle as a place for learning and browsing suggests that data is used as material for data science tasks, where the focus is on the performance of e.g. a machine learning model without having to worry about real-world implications. This implies that data quality conception on Kaggle might be different from how data quality is judged in real-world scenarios with potential social impact.

However, this highlights the importance of adopting a user-centric view of data quality, moving beyond measurable indicators such as completeness, duplicates, or consistency (e.g. Zhang et al., 2014) which have dominated technical quality frameworks from the literature.

5.2.4 Criteria for choosing data: Factors related to the data community

For each dataset, Kaggle shows a number of engagement metrics generated by the activity of the community, for instance votes, downloads, number of ongoing discussions, status of the top contributors or tasks described together with the data (Figure 6). We asked our survey respondents whether these make a difference when choosing a dataset. In general, they felt that clear tasks or questions described together with the dataset were most critical for choosing a dataset in 73% of the responses, followed by the number of notebooks using the dataset, number of votes a dataset has (61%) and number of ongoing discussions (topics and comments) (58%). The other option %was chosen relatively often (n=23) referred mostly to criteria aside from community engagement, covered in other sections.

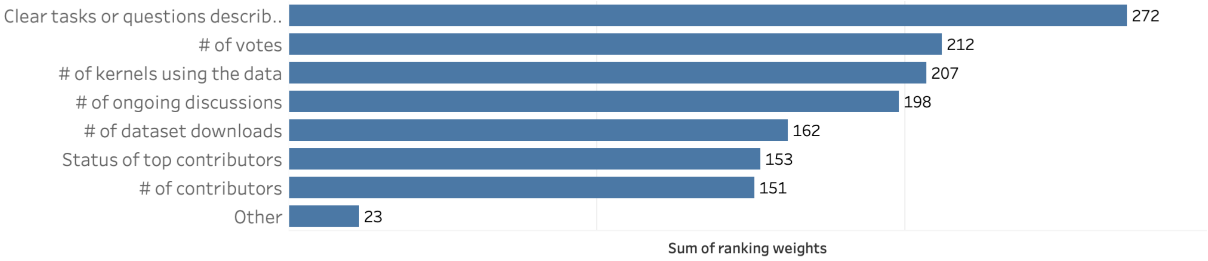


Fig. 6: Community engagement metrics important for choosing datasets (n=434)

Our results so far highlight the importance of textual descriptions, understandability as well as of clear tasks or questions around the data, amongst other factors. We envision the automated creation of text dataset descriptions, based on information needs, as suggested in Koesten et al. (2020a). Furthermore, textual descriptions of columns can serve to aid the understandability of datasets. Kaggle displays statistical column summaries, which could be extended access routes to facilitate even more direct interactive data exploration. Another avenue for adding descriptions to data would be a connected exploratory data analysis (EDA) notebook together with the dataset, as mentioned by one respondent. We discuss the use of notebooks in the next section.

Our findings confirm the use of various and additional communication tools (e.g. private messaging apps) during Kaggle competitions, which point to a potential design space for formalising informal communication in data-centric conversations. Insights related to online discussions can be applied in a data work context depending on the scale, e.g. as discussed by Zhang, Muller & Wang (2020), who mention summarization tools to make sense of long discussions or annotation tools to situate conversations in context. At the same time, questions about differentiated access are relevant in data-centric teamwork to mirror the different communication and trust levels that play a role in informal conversations during such work. Highly used datasets spark discussion, which in turn can be seen as a signal for new users that they can learn from others and will be able to find or call for advice when needed. The importance of contact points and social interaction around data has been mentioned in literature before (e.g. Birnholtz & Bietz, 2003) our results confirm these in the example of dataset reuse on Kaggle.

5.2.5 Criteria for choosing data: Factors related to data tools

Co-locating tools with datasets has been shown to facilitate data use (Koesten et al., 2020b). In Kaggle, tools are deployed in Notebooks, a virtual notebook environment similar to Jupyter notebooks. This allows users to create and share code, text as well as visualisations. Datasets are directly loaded into a notebook hosted directly on the platform, removing the need to download the dataset and possibly install and configure a local data science programming environment.

A survey participant mentioned the usefulness of data and notebooks as a motivation for using Kaggle and participating in competitions:

> “I work in academia and I am interested in publishing. Databases are hard to collect and Kaggle offers them freely. Sometimes, one finds interesting kernels (notebooks) as well. Kernels (notebooks) offer a baseline for comparison.”

Notebooks are also used to describe datasets, for instance, in the case of exploratory data analysis, as pointed out by participants:

> “One kernel (= notebook) that briefly demos how to use the data - a starter kernel.”

Being able to see the data in action adds to its understandability, which is the factor ranking highest in our participants' concept of data quality, as reported earlier.

Notebooks are becoming the de facto standard tool for data collaborations, especially in early stages of a project focusing on exploration and prototyping rather than production-ready services and applications. Notebooks combine code, visuals and text in one place, which allows people to reproduce what is happening to the data and hence increase data accessibility (Rule, Tabard & Hollan., 2018; Kery et al., 2018).

6. Limitations

Our data and analyses are descriptive, not predictive. They only represent the practices of our respondents -- a group of data-aware people, working with data to varying degrees, and confident in their ability to respond to an English-language survey.

As with every survey, there is a self-selection bias, which excludes information about non-respondents. The results presented here depict the behaviours and attitudes of the 400+ survey participants. Survey data does not offer the same richness as qualitative methods, and responses to single or multiple-choice questions are shaped by the options provided. We attempted to counter these issues by piloting the survey, and by including a free-text response option for each question, which are also reported in the results.

This work focuses on Kaggle as a case study. While users of Kaggle share some commonalities with the rest of the data science community, Kaggle remains, for a range of reasons, a unique platform, which has come about as a place to host and participate in data science competitions. Learning and data-related hackathons, challenges and competitions are fairly common in other contexts, including within organisations. However, without further studies, we cannot be certain about how much our findings apply elsewhere. There are also limitations in terms of types of data available on Kaggle and related analyses. For example, qualitative data is underrepresented on Kaggle, and hence, user assessment of fitness of use might look very different in that case. The same applies to affordances around co-located tools and other resources for qualitative data.

7. Bibliography

1. Jeremy P. Birnholtz and Matthew J. Bietz. 2003. Data at work: supporting sharing in science and engineering. In *Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work, GROUP 2003, Sanibel Island, Florida, USA, November 9-12, 2003*, Kjeld Schmidt, Mark Pendergast, Marilyn Tremaine, and Carla Simone (Eds.). ACM, 339–348. https://doi.org/10.1145/958160.958215
2. Hudson Borges, André C. Hora, and Marco Tulio Valente. 2016. Understanding the Factors That Impact the Popularity of GitHub Repositories. In *2016 IEEE International Conference on Software Maintenance and Evolution, ICSME 2016, Raleigh, NC, USA, October 2-7, 2016*. IEEE Computer Society, 334–344. https://doi.org/10.1109/ICSME.2016.31
3. Giorgos Cheliotis and Jude Yew. 2009. An analysis of the social structure of remix culture. In Proceedings of the Fourth International Conference on Communities and Technologies, C&T 2009, University Park, PA, USA, June 25-27, 2009, John M. Carroll (Ed.). ACM, 165–174. https://doi.org/10.1145/1556460.1556485
4. Luis-Daniel Ibáñez, Laura Koesten, Emilia Kacprzak, Elena Simperl,2020. Analytical Report 18: Characterising Dataset Search on the European Data Portal: An Analysis of Search Logs. European Commission, European Data Portal https://data.europa.eu/sites/default/files/analytical_report_18-characterising_data_search_edp.pdf
5. Neelamadhav Gantayat, Pankaj Dhoolia, Rohan Padhye, Senthil Mani, and Vibha Singhal Sinha. 2015. The Synergy between Voting and Acceptance of Answers on StackOverflow - Or the Lack Thereof. In 12th IEEE/ACM Working Conference on Mining Software Repositories, MSR 2015, Florence, Italy, May 16-17, 2015, Massimiliano Di Penta, Martin Pinzger, and Romain Robbes (Eds.). IEEE Computer Society, 406–409. https://doi.org/10.1109/MSR.2015.50
6. Alyssa Goodman, Alberto Pepe, Alexander W. Blocker, Christine L. Borgman, Kyle Cranmer, Mercè Crosas, Rosanne Di Stefano, Yolanda Gil, Paul T. Groth, Margaret Hedstrom, David W. Hogg, Vinay Kashyap, Ashish Mahabal, Aneta Siemiginowska, and Aleksandra Slavkovic.

2014. Ten Simple Rules for the Care and Feeding of Scientific Data. PLoS Computational Biology 10, 4 (2014). https://doi.org/10.1371/journal.pcbi.1003542

7. Kathleen Gregory, Paul Groth, Andrea Scharnhorst, and Sally Wyatt. 2020. Lost or Found? Discovering Data Needed for Research. Harvard Data Science Review 2, 2 (30 4 2020). https://doi.org/10.1162/99608f92.e38165eb https://hdsr.mitpress.mit.edu/pub/gw3r97ht.
8. Melanie Herschel, Ralf Diestelkämper, and Houssem Ben Lahmar. 2017. A survey on provenance: What for? What form? What from? *VLDB J.* 26, 6 (2017), 881–906. https://doi.org/10.1007/s00778-017-0486-1
9. Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The dataset nutrition label: A framework to drive higher data quality standards. *arXiv preprint arXiv:1805.03677* (2018).
10. Shin-Yuan Hung, Alexandra Durcikova, Hui-Min Lai, and Wan-Mei Lin. 2011. The influence of intrinsic and extrinsic motivation on individuals' knowledge sharing behavior. *International journal of human-computer studies* 69, 6 (2011), 415–427.
11. Dagmar Kern and Brigitte Mathiak. 2015. Are There Any Differences in Data Set Retrieval Compared to Well-Known Literature Retrieval?. In *Research and Advanced Technology for Digital Libraries - 19th International Conference on Theory and Practice of Digital Libraries, TPDL 2015, Poznań, Poland, September 14-18, 2015. Proceedings (Lecture Notes in Computer Science, Vol. 9316)*, Sarantos Kapidakis, Cezary Mazurek, and Marcin Werla (Eds.). Springer, 197–208. https://doi.org/10.1007/978-3-319-24592-8_15
12. Mary Beth Kery, Marissa Radensky, Mahima Arya, Bonnie E. John, and Brad A. Myers. 2018. The Story in the Notebook: Exploratory Data Science using a Literate Programming Tool. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018*, Regan L. Mandryk, Mark Hancock, Mark Perry, and Anna L. Cox (Eds.). ACM, 174. https://doi.org/10.1145/3173574.3173748
13. Jihie Kim, Ewa Deelman, Yolanda Gil, Gaurang Mehta, and Varun Ratnakar. 2008. Provenance trails in the Wings/Pegasus system. *Concurrency and Computation: Practice and Experience* 20, 5 (2008), 587–597. https//doi.org/10.1002/cpe.1228
14. Laura Koesten, Kathleen Gregory, Paul Groth, and Elena Simperl. 2021. Talking datasets–understanding data sense- making behaviours. *International Journal of Human-Computer Studies* 146 (2021), 102562.
15. Laura Koesten, Emilia Kacprzak, Jeni Tennison, and Elena Simperl. 2019. Collaborative Practices with Structured Data: Do Tools Support What Users Need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*, Stephen A. Brewster, Geraldine Fitzpatrick, Anna L. Cox, and Vassilis Kostakos (Eds.). ACM, 100. https://doi.org/10.1145/3290605.3300330
16. Laura Koesten and Elena Simperl. 2021. UX of data: making data available doesn't make it usable. *Interactions* 28, 2 (2021), 97–99.
17. Laura Koesten, Elena Simperl, Tom Blount, Emilia Kacprzak, and Jeni Tennison. 2020a. Everything you always wanted to know about a dataset: Studies in data summarisation. *International Journal of Human-Computer Studies* 135 (2020). https://doi.org/10.1016/j.ijhcs.2019.10.004
18. Laura Koesten, Pavlos Vougiouklis, Elena Simperl, and Paul Groth. 2020b. Dataset Reuse: Toward Translating Principles to Practice. *Patterns* 1, 8 (2020), 100136. https://doi.org/10.1016/j.patter.2020.100136
19. David R Millen and Jonathan Feinberg. 2006. Using social tagging to improve social navigation. In Workshop on the Social Navigation and Community based Adaptation Technologies. Citeseer.
20. Paolo Missier, Bertram Ludäscher, Saumen C. Dey, Michael Wang, Timothy M. McPhillips, Shawn Bowers, Michael Agun, and Ilkay Altintas. 2012. Golden Trail: Retrieving the Data History that Matters from a Comprehensive Provenance Repository. IJDC 7, 1 (2012), 139–150. https://doi.org/10.2218/ijdc.v7i1.221

21. Adam Rule, Aurélien Tabard, and James D. Hollan. 2018. Exploration and Explanation in Computational Notebooks. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018, Regan L. Mandryk, Mark Hancock, Mark Perry, and Anna L. Cox (Eds.). ACM, 32. https://doi.org/10. 1145/3173574.3173606
22. Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021, Yoshifumi Kitamura, Aaron Quigley, Katherine Isbister, Takeo Igarashi, Pernille Bjørn, and Steven Mark Drucker (Eds.). ACM, 39:1–39:15. https://doi.org/10.1145/3411764.3445518
23. Sebastian Schelter, Dustin Lange, Philipp Schmidt, Meltem Celikel, Felix Bießmann, and Andreas Grafberger. 2018. Automating Large-Scale Data Quality Verification. *Proc. VLDB Endow.* 11, 12 (2018), 1781–1794. https://doi.org/10. 14778/3229863.3229867
24. April Yi Wang, Dakuo Wang, Jaimie Drozdal, Xuye Liu, Soya Park, Steve Oney, and Christopher Brooks. 2021. What Makes a Well-Documented Notebook? A Case Study of Data Scientists' Documentation Practices in Kaggle. In *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama Japan, May 8-13, 2021, Extended Abstracts*, Yoshifumi Kitamura, Aaron Quigley, Katherine Isbister, and Takeo Igarashi (Eds.). ACM, 445:1–445:7. https://doi.org/10.1145/3411763.3451617
25. Richard Y. Wang and Diane M. Strong. 1996. Beyond Accuracy: What Data Quality Means to Data Consumers. *J. Manag. Inf. Syst.* 12, 4 (1996), 5–33. http://www.jmis-web.org/articles/1002
26. Amy X. Zhang, Michael J. Muller, and Dakuo Wang. 2020. How do Data Science Workers Collaborate? Roles, Workflows, and Tools. *Proc. ACM Hum. Comput. Interact.* 4, CSCW (2020), 022:1–022:23. https://doi.org/10.1145/3392826

8. Appendix:

Questions and response options for a Kaggle community survey on data work practices.

Intro, privacy notice, researcher contact, and consent note are omitted for publication.

Question type:
SC = single choice, MC = multiple choice, L = Likert Scale, R = Rank

A. Demographics and experience

1. **What is your age range? [SC]**
   a. <18 / 18-29 / 30-39 / 40-49 / 50-59 / 60+

2. **What best describes your current job role? [SC]**
   a. *Business Analyst*
   b. *Data Analyst*
   c. *Data Engineer*
   d. *Data Scientist*
   e. *DBA / Database Engineer*
   f. *Educator*
   g. *Product / Project Manager*

   h. *Researcher*
   i. *Software Engineer*
   j. *Statistician*
   k. *Student*
   l. *Employed in a non technology related role*
   m. *Not Employed*
   n. *Retired*
   o. *Other, please specify...*

3. **In which country do you currently reside? [SC]**
   a. *[dropdown of countries]*

4. **What is the highest degree or level of school you have completed? [SC]**
   a. Less than high school degree
   b. High school degree or equivalent
   c. Professional degree (for a specific profession)
   d. Some college/university study without earning a bachelor's degree
   e. Bachelor's degree
   f. Master's degree
   g. Doctoral degree
   h. Other: (please describe)

5. **Do you perform data analysis as part of your daily work or in your free time? [SC]**
   a. For work
   b. Outside of work
   c. Both
   d. Other, please describe *[free text]*
   e. None

6. **How long have you been practicing data analysis? [SC]**
   a. Less than a year / 1-3 years / 3-5 years / More than 5 years

7. **When was the last time you used datasets on Kaggle (e.g. downloaded / used a dataset, or uploaded your own dataset [SC]**
   a. Less than a year / 1-3 years / 3-5 years / More than 5 years/ Have not used Kaggle datasets before (->branch to terminate survey)

B. Working with data

8. **What are some of your motivations or reasons for working with Kaggle datasets? [MC]**
   a. Learning new skills or methods
   b. Taking part in competitions
   c. Engaging with peers
   d. Adding to my CV / demonstrate my skills
   e. Solving open (research) problems
   f. For fun
   g. Other, please describe (free text)

9. **What are the three most frequent things you do when going to Kaggle? Please select up to 3 choices that apply. [MC]**
   a. Browse / find datasets
   b. Browse / find kernels
   c. Use datasets
   d. Share code I worked on with the community
   e. Publish a dataset

f. Begin or continue a data science project
g. Participate in a competition or challenge
h. Participate in community efforts to solve a problem (e.g. COVID-19 datasets)
i. Checking in to see what's new
j. Take a course on Kaggle's Learn platform
k. Other, please describe [free text]

10. How do you find and choose datasets to work with? [MC]

a. Using search engines (e.g. Google search)
b. Browsing Kaggle
c. Recommendations from other people
d. Other, please describe *[free text]*

11. When you choose a dataset to work with, how important are the following characteristics? Please select the choices according to their importance to you. [R]

a. Format or data type
b. Who posted it (person / institution)
c. Column headers or variables
d. Description of the data
e. Geographic location of data
f. Timeframe of the data
g. Quality of the data
h. Size of the dataset
i. Obvious patterns or trends in the data
j. Amount of activity around the dataset
k. Usability rating
l. Other, please describe *[free text]*

12. How important are the following features provided by Kaggle for you to choose a dataset? Please select the choices according to their importance to you. [R]

a. Tags
b. Overview description
c. File description
d. Column description
e. License
f. File format
g. Provenance / origin
h. Update frequency specified
i. Has a public kernel

13. How do you know a dataset is good quality? Please drag and drop, and rank these in order of importance. [R]

a. Completeness (not many missing values)
b. Timeliness (is the data recent?)
c. Believability (is the data believable?)
d. Understandability (is the data understandable?)
e. Consistency (is the data consistent, e.g. formatting?)
f. Provenance / origin (who published the data?)
g. Other *[free text]*

14. When you choose a dataset, how important are the following characteristics? Please select the choices according to their importance to you. [R]

a. Number of ongoing discussions (topics and comments)
b. Number dataset downloads
c. Number of votes a dataset has
d. Number of contributors

e. Status of top contributors
f. Number of Kernels using the dataset
g. Clear tasks or questions described together with the dataset
h. Other, please describe *[free text]*

15. Do you contribute to Kaggle by publishing data or code? [L]

a. Never, Rarely, Sometimes, Often, A Lot

16. Do you contribute to Kaggle's discussion forums? [L]

a. Never, Rarely, Sometimes, Often, A Lot

17. Do you download a dataset before deciding to work with it? [L]

a. Never, Rarely, Sometimes, Often, A Lot

18. When I have used a Kaggle dataset I publish my code back on Kaggle [SC]

a. Yes
b. Sometimes
c. No
d. Not applicable

20.1 *If yes or sometimes:* **What motivates you to post your code back on Kaggle?** [free text]
20.1 *If no:* **Do you share the code elsewhere, or with others?** Please specify: [free text]

19. When I make changes to a Kaggle dataset I publish the new dataset version back on Kaggle? [SC]

a. Yes
b. Sometimes
c. No
d. Not applicable

C. Competitions

20. Have you participated in a competition on Kaggle in the last 2 years? [SC]

a. Yes.
b. No *[survey termination]*

21. How many people were in your team (including yourself)? [SC]

a. Just me / 2 / 3 / 4 / 5 / more than 5

22. What are typical roles of team members during a competition? [MC]

a. Data Exploration
b. Data Visualization
c. Data Modelling
d. Data Wrangling
e. Data Cleaning
f. Project management (organizing)
g. Other, please describe *[free text]*

23. Did you know any of the people in your team before? [SC]

a. No
b. Yes, all of them
c. Yes, some of them
d. I was a 1 person team

If yes, did you know them in person or online?

In person / Online / Both

24. How do you communicate with your teammates during a competition? [MC]

a. On Kaggle
b. Email
c. Messaging app (eg. WhatsApp / WeChat)
d. Video conferencing (e.g. Zoom, Hangouts, etc)
e. In a notebook
f. Other, please describe *[free text]*

25. How did you make decisions during a competition? [SC]

a. Alone
b. Mostly alone
c. Mostly as a team
d. As a team

26. Why do you take part in competitions on Kaggle?

a. *[free text]*

Thank you very much for taking the time to complete this survey!
If there is anything you would like to comment on please do it here: *[free text]*