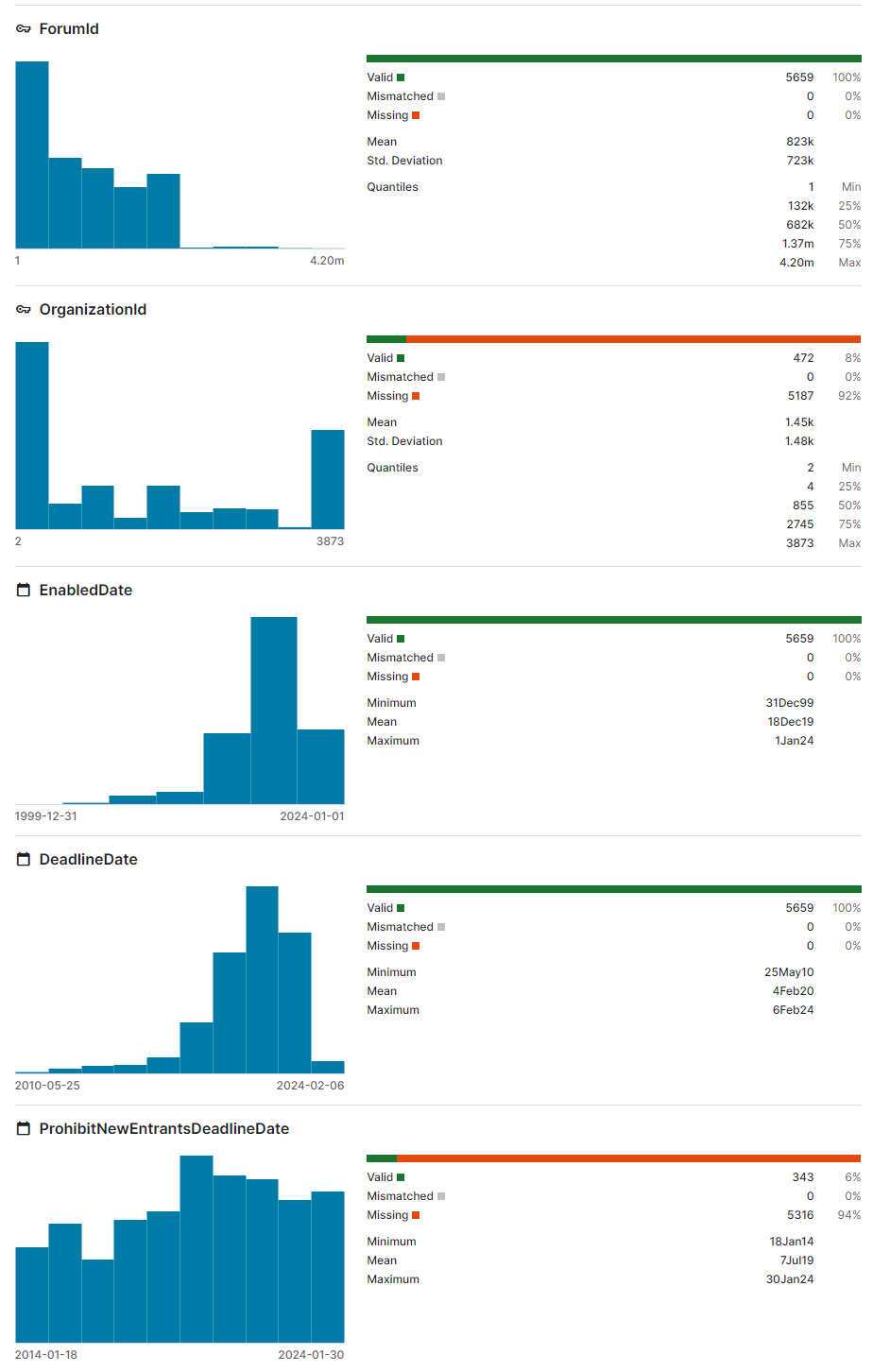
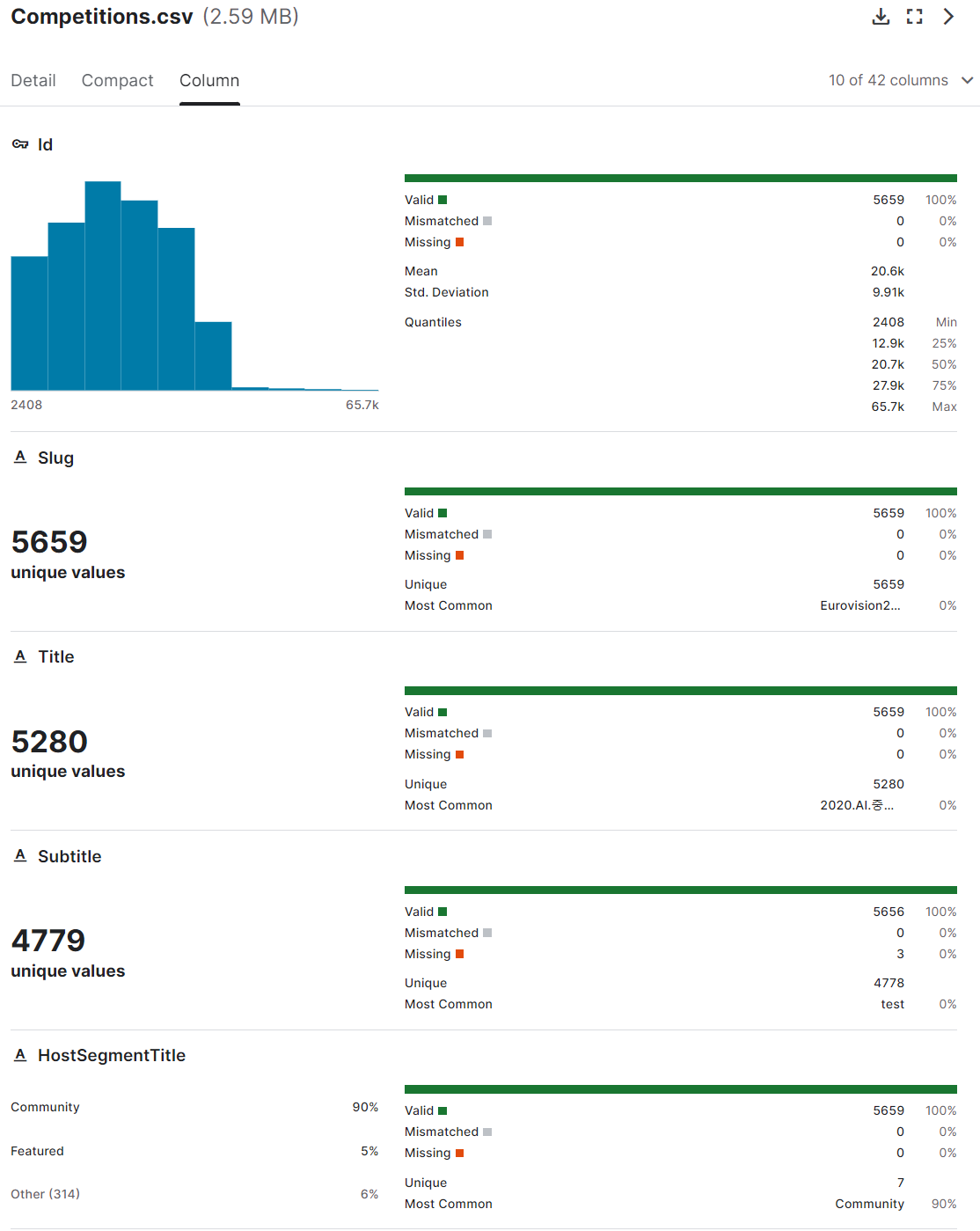
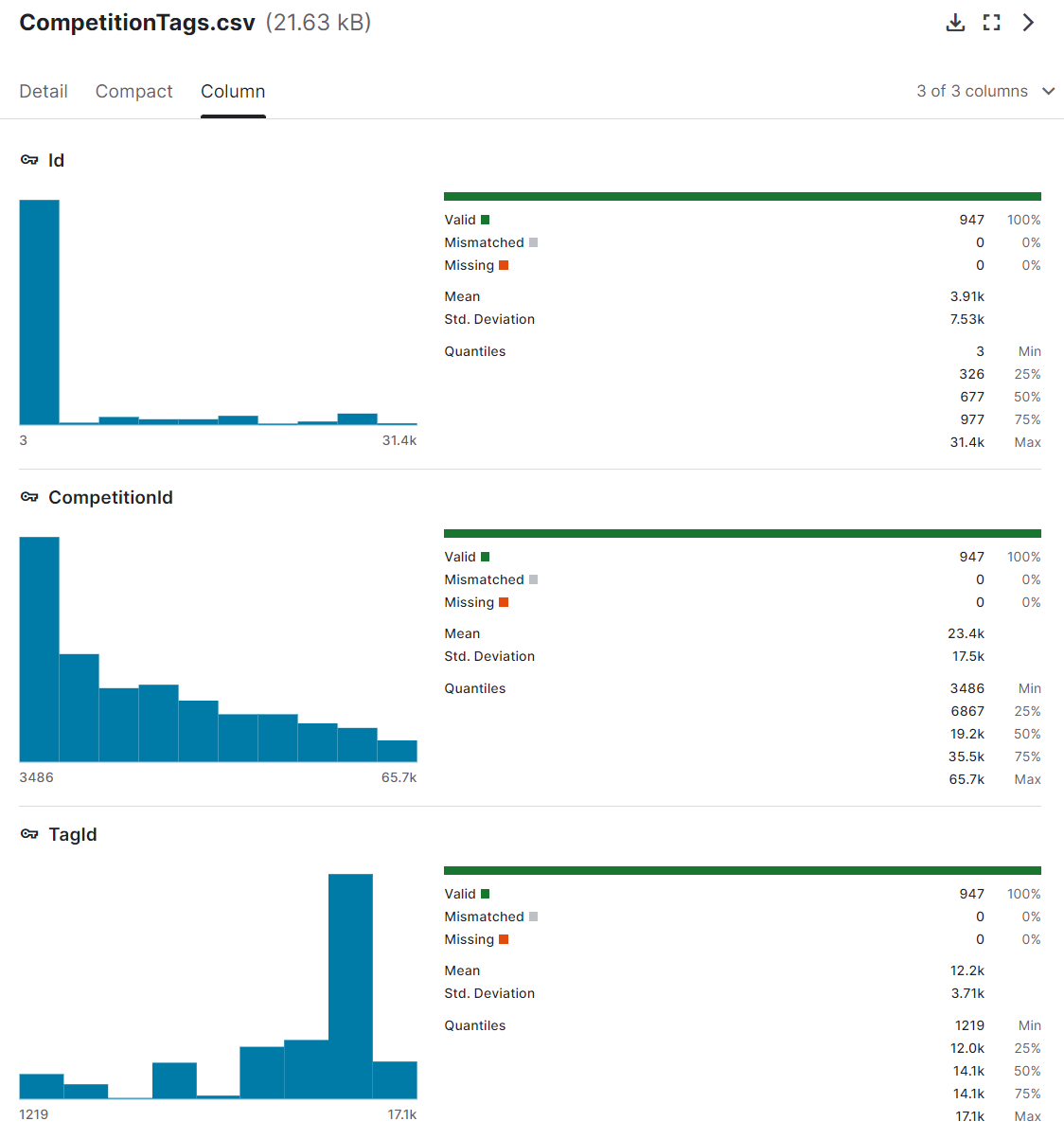
# Competitions.csv



Notes

* 5659 rows of data
* Id: min=2408, max=65711. This id uniquely identifies each competition in Kaggle; can be used for joining with other data sets.
* Slug: URL slug
* Title: project title. A small number of rows are gibberish, so I think we can delete them when visualizing.
* Subtitle: brief project description.
* HostSegmentTitle: Community, Featured, and Research are the major categories. Featured and Research projects allow cash prizes.

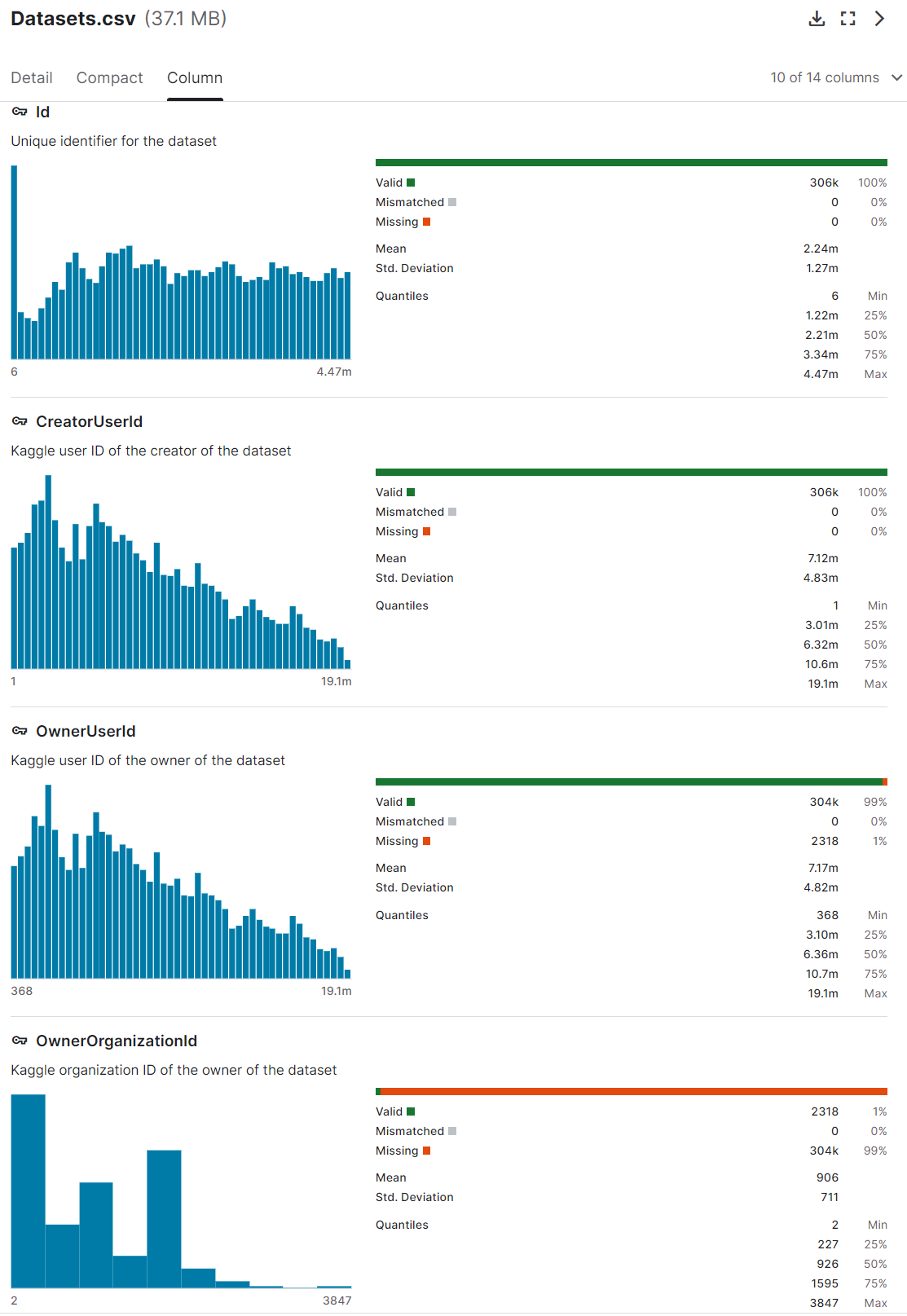
# CompetitionTags.csv

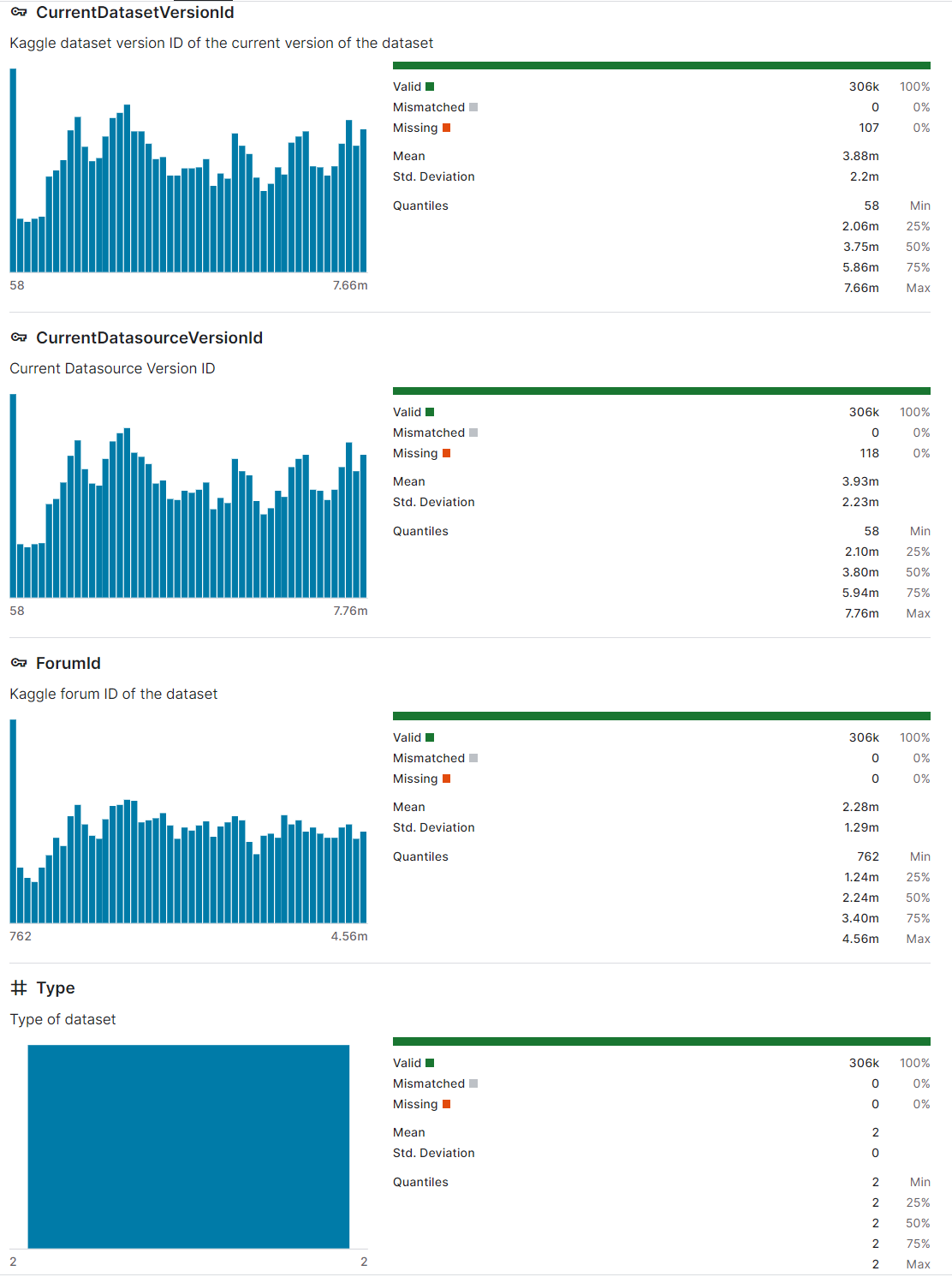


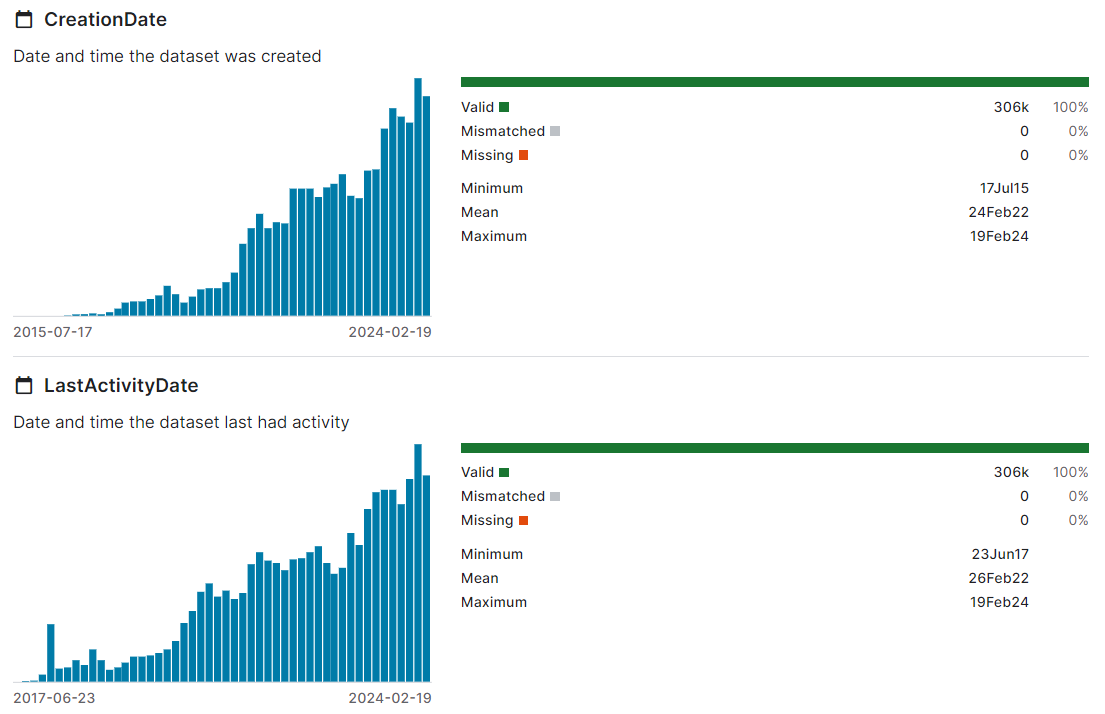
* Id: I think this is just a column of primary keys for this data set.
* CompetitionId: These ID’s are associated with ID’s in Competition.csv, although not all of them are present in this data set, probably because not all competitions are tagged to a category.
* TagId: Assigns each competition to a tag. This number corresponds to the Id and ParentId columns in the Tags.csv dataset.

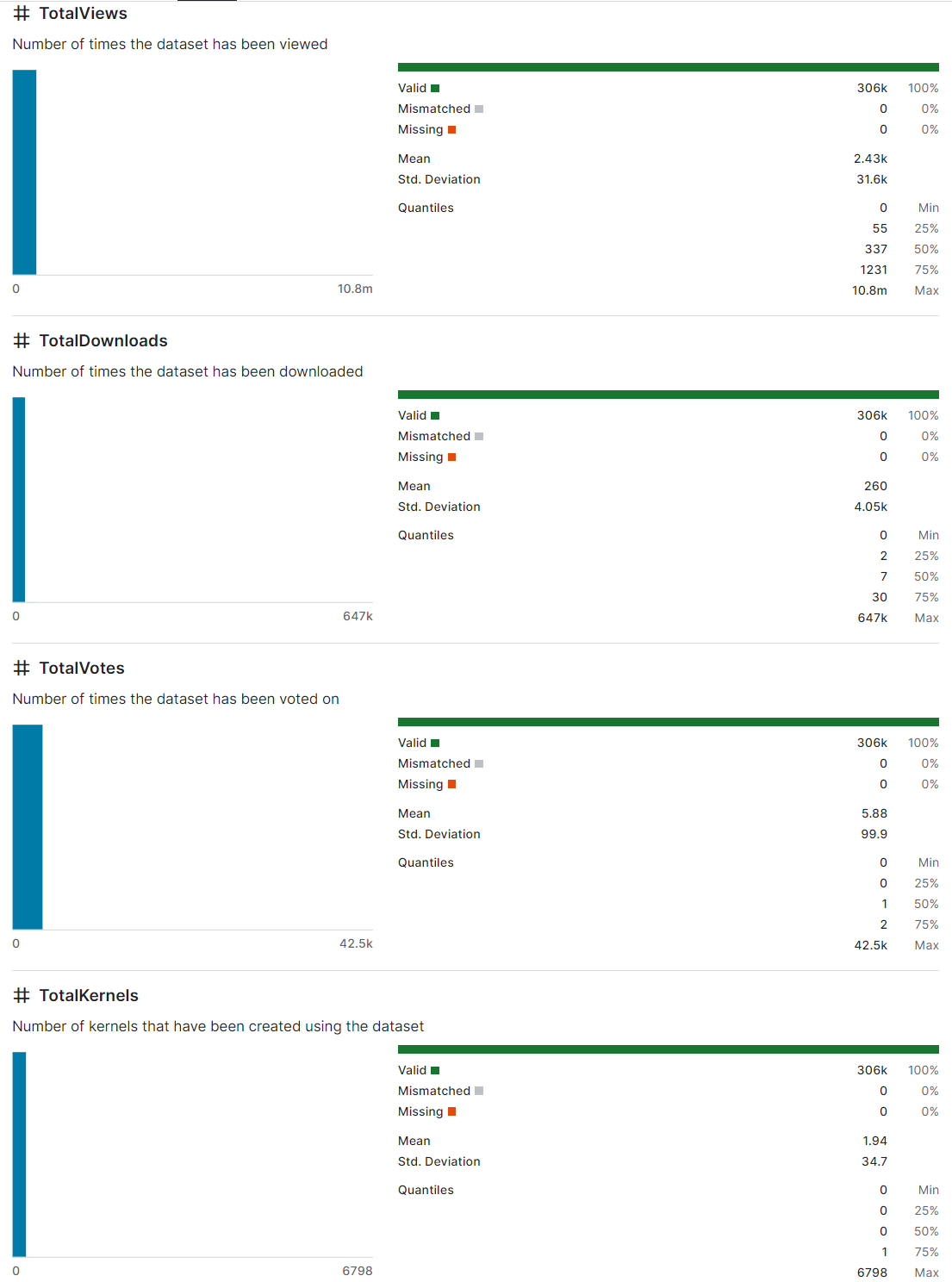
# Datasets.csv

Kaggle describes the variables for this data set.

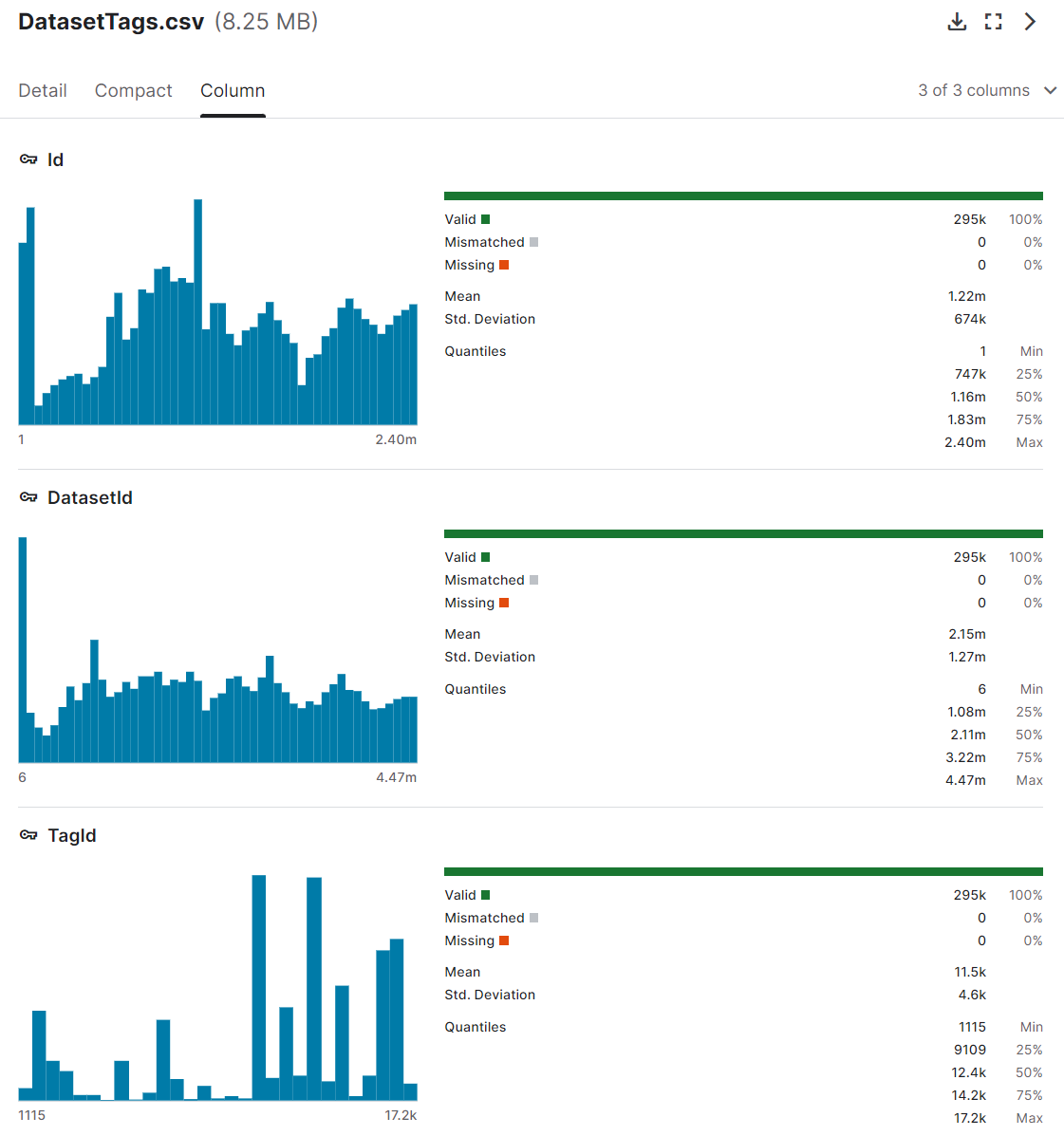








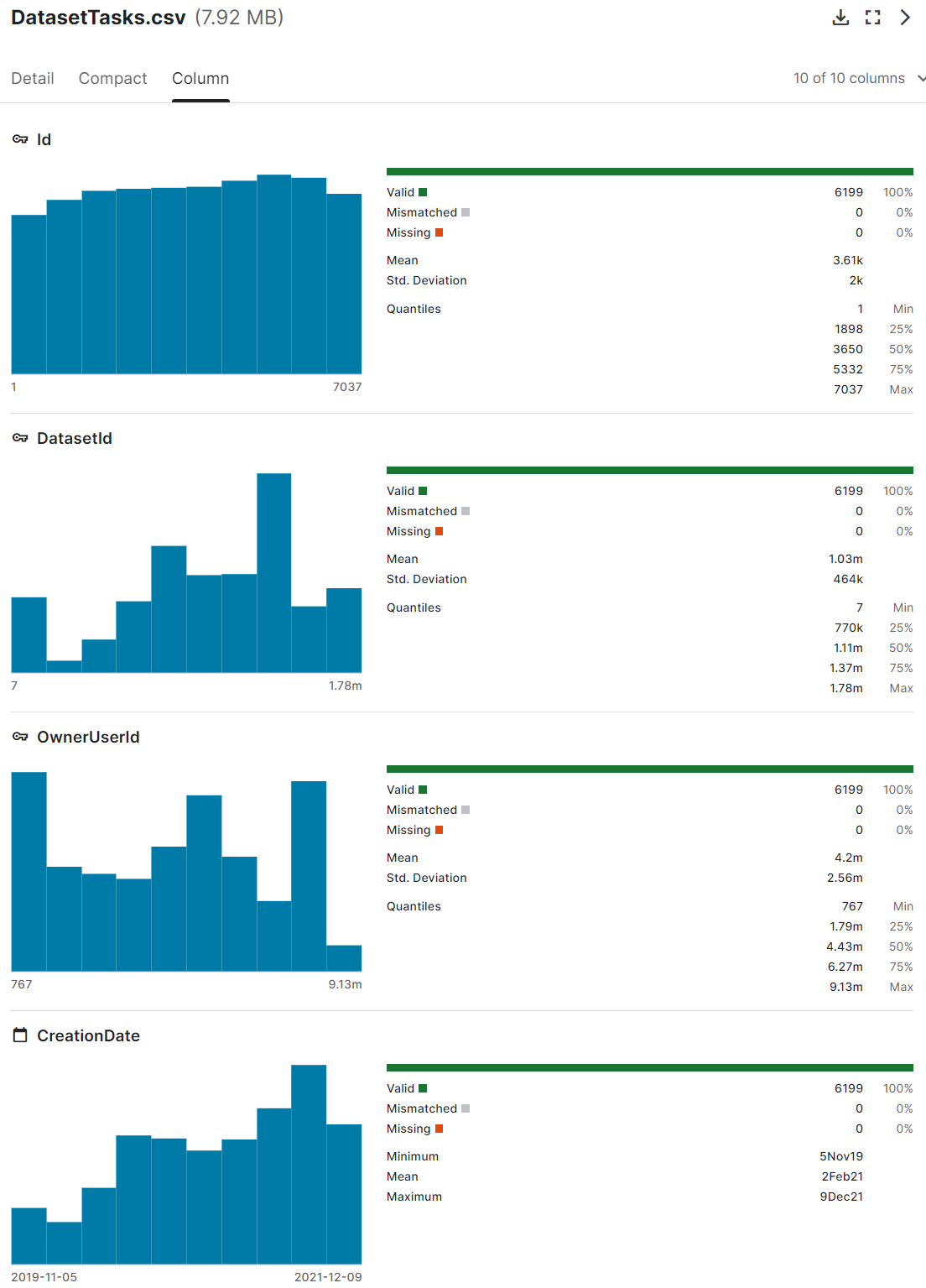
# DatasetTags.csv

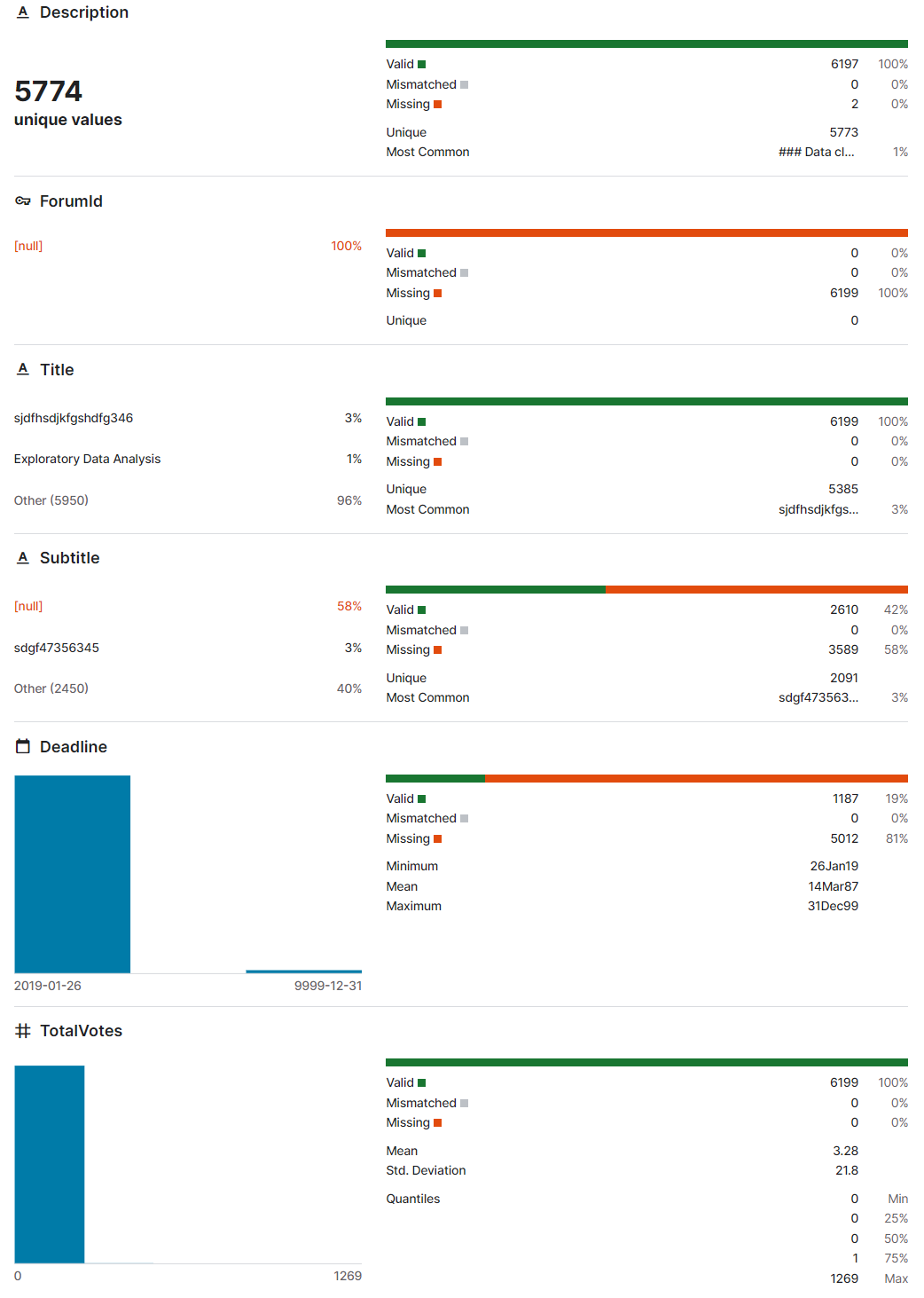


* Id: a unique identifier for this data set
* DatasetId: These Id’s can join on “Id” column in datasets.csv
* TagId: can join on “TagId” in Tags.csv

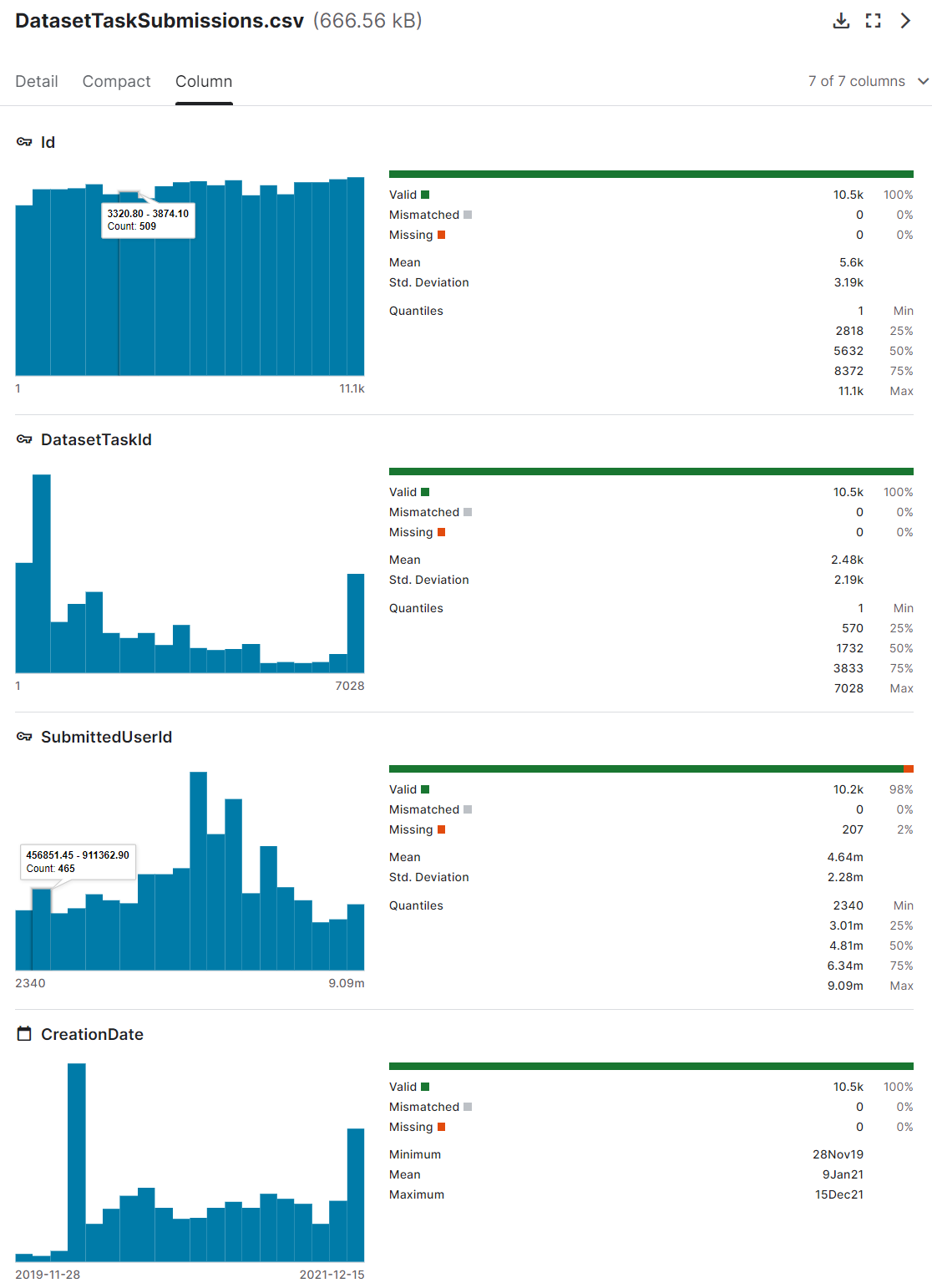
# DatasetTasks.csv

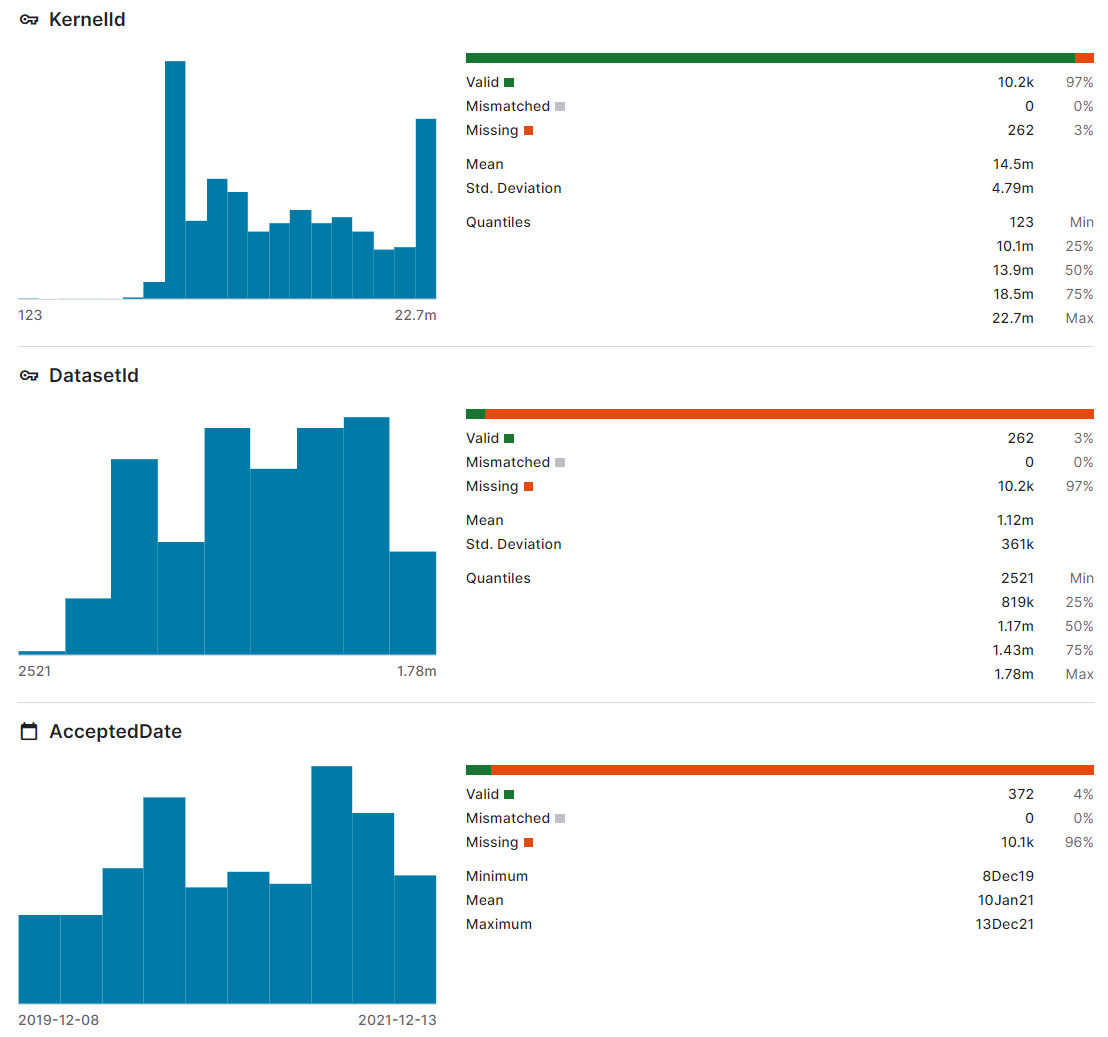
This data set contains information about the tasks (e.g., EDA, visualization, ML model, etc.) related to the data sets. So, “DatasetId” and “OwnerUserId” correspond to columns of the same name in Datasets.csv.





# DatasetTaskSubmissions.csv

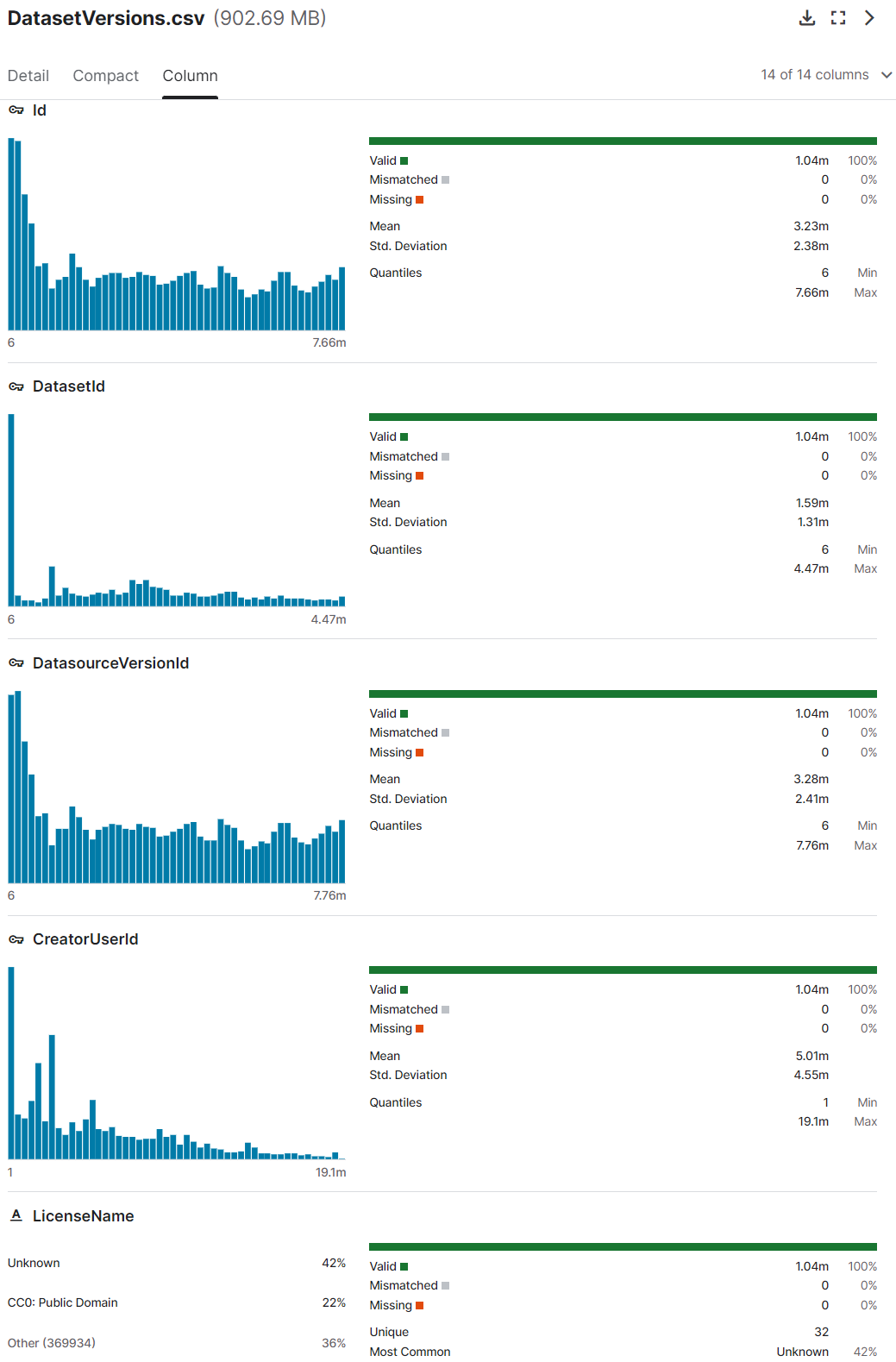


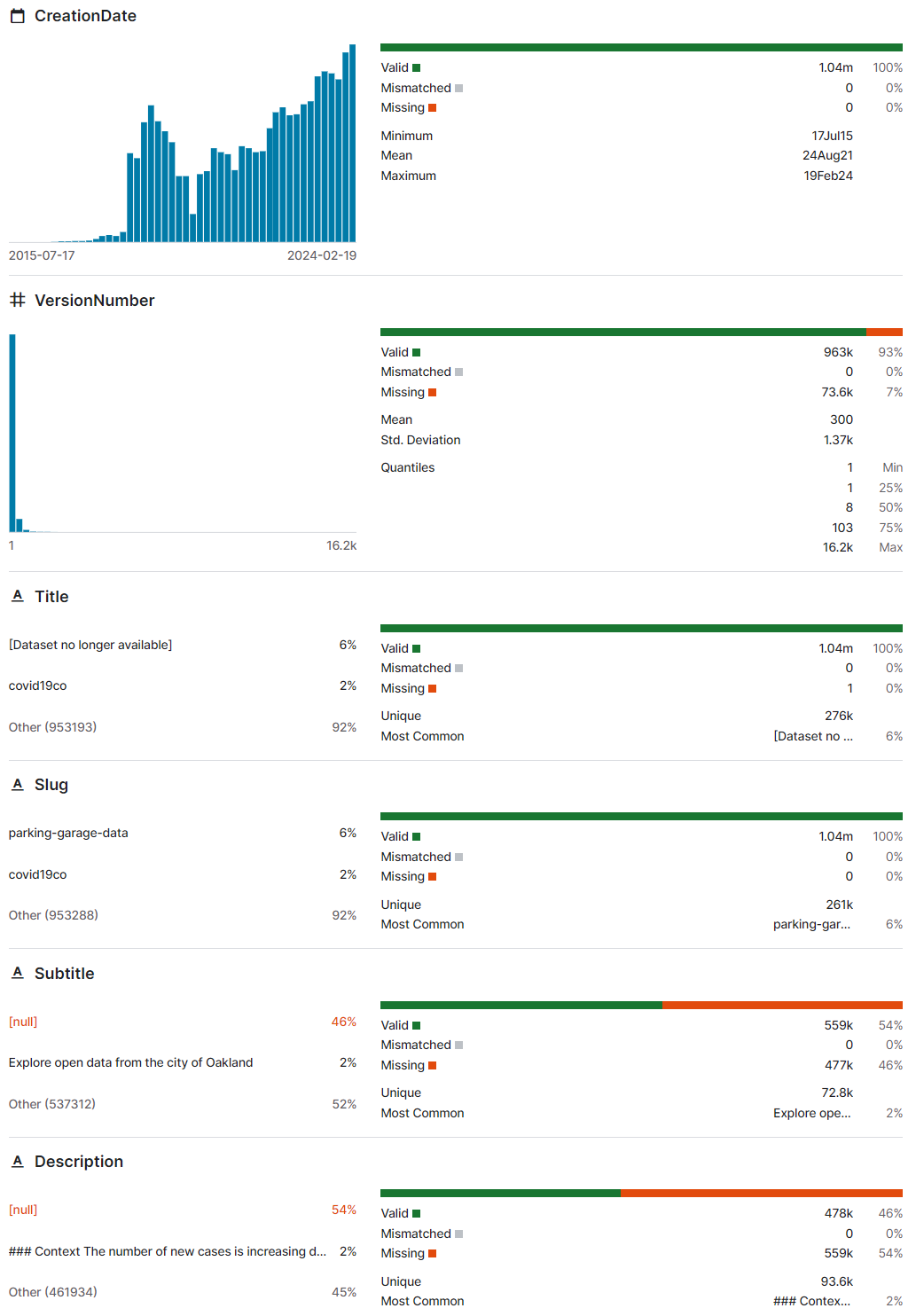


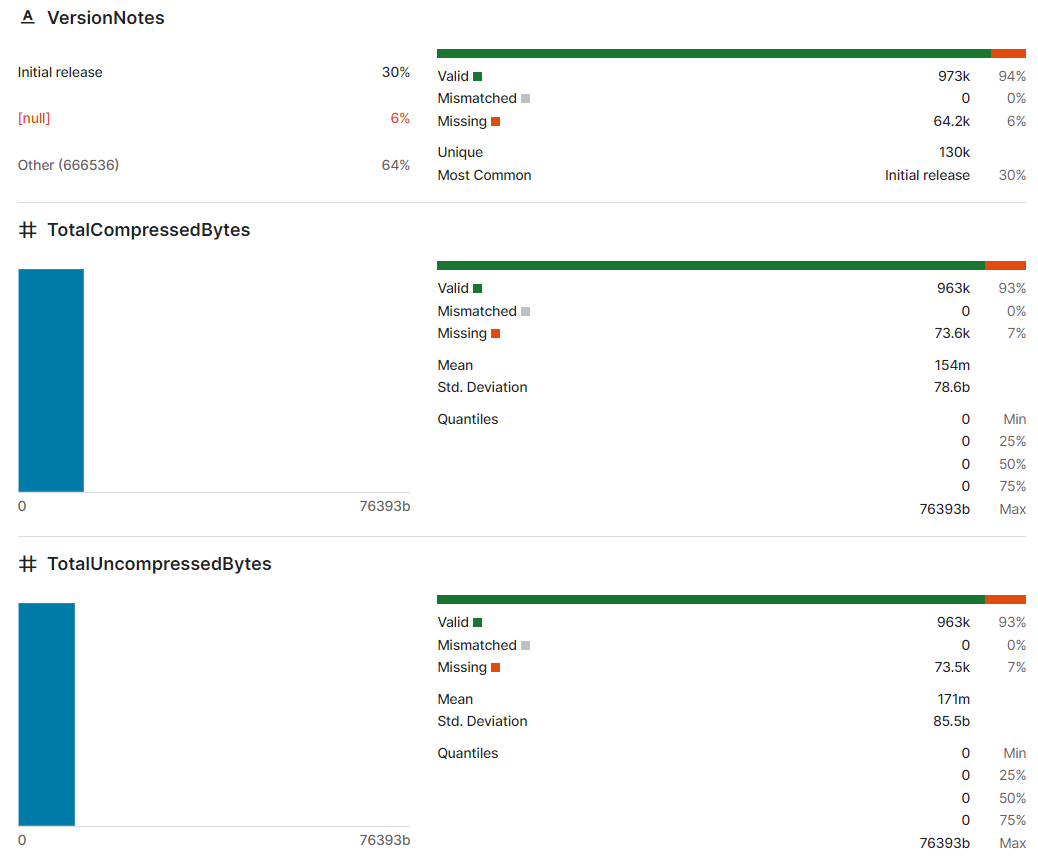
* DatasetTaskId can join on “Id” in DatasetTasks.csv
* I have realized that all user ID’s, like “SubmittedUserId”, “OwnerUserId”, and “CreatorUserId” share the same keys from “Id” column in Users.csv. These kinds of IDs will pop up in other data sets.
* KernelId: Id’s associated with the different kernels that have been submitted by users for this data set of submitted tasks.
* “DatasetId” can join on “Id” in Datasets.csv.

# DatasetVersions.csv

Some data sets have multiple versions (revisions or updates).

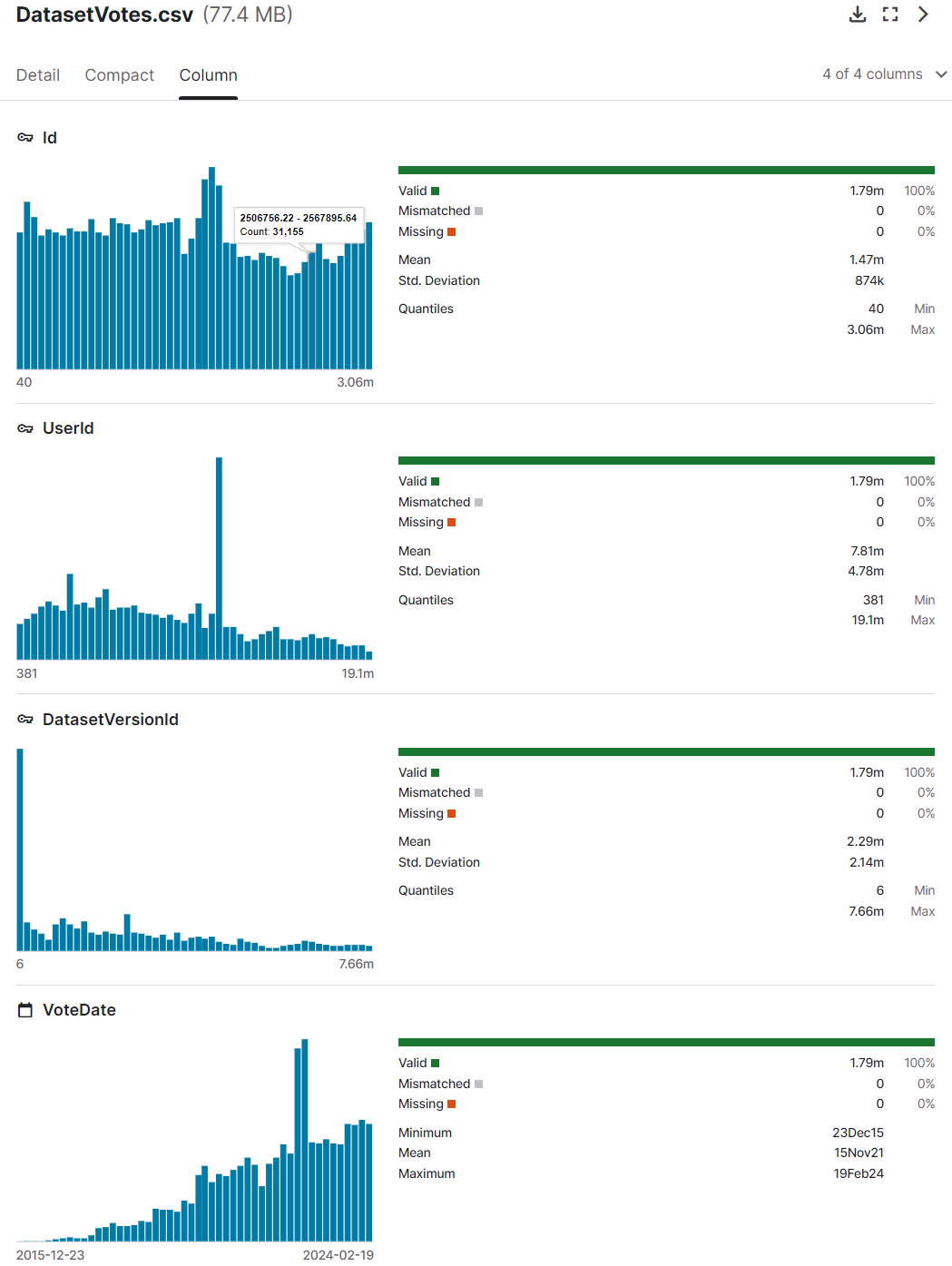






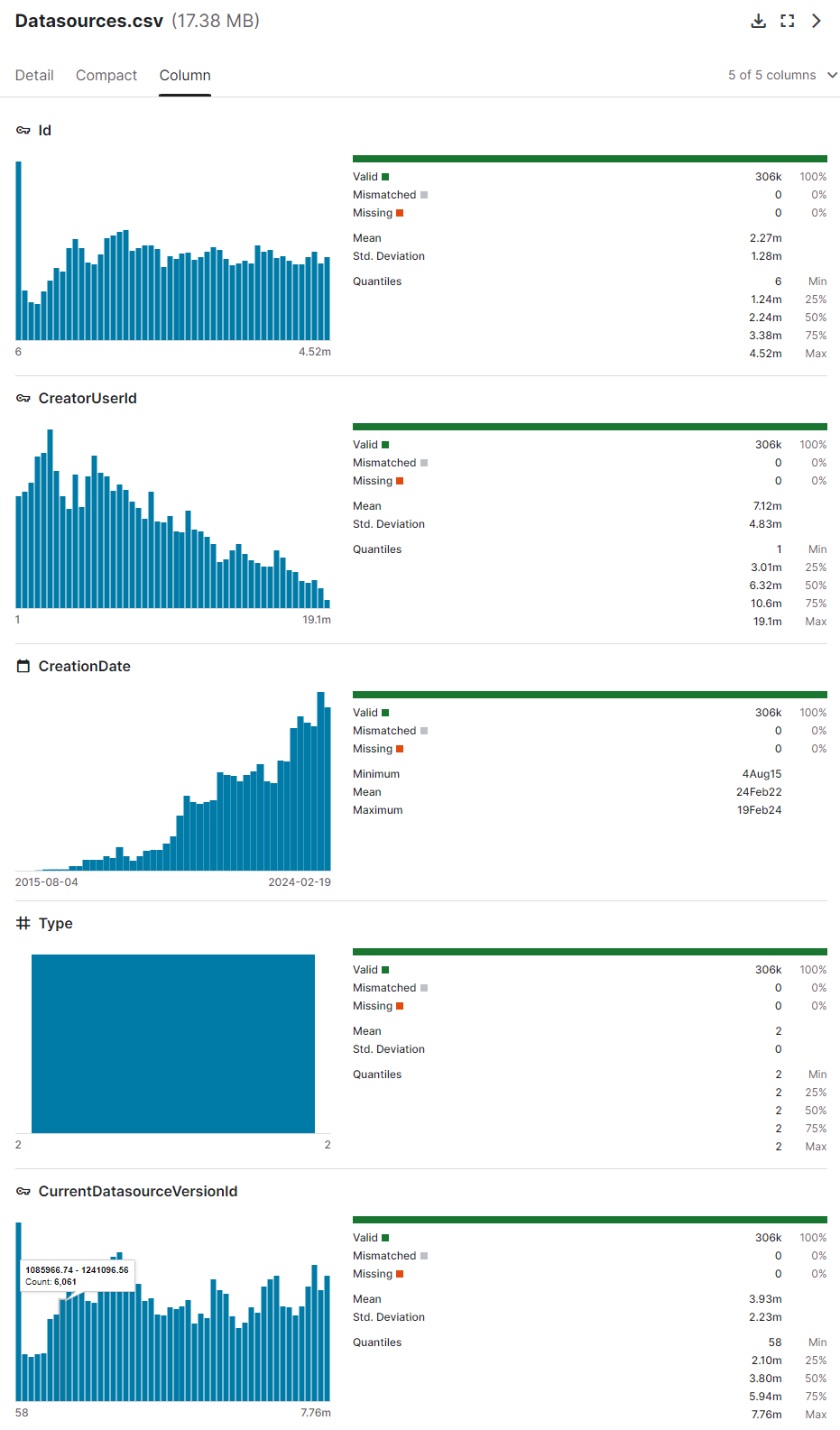
* There are some rows that are different from the rest and only contain 1-3 columns. May need to delete these rows.
* “DatasetId” can join on “Id” in Datasets.csv

# DatasetVotes.csv



* UserId is ID of user that cast vote
* DatasetVersionId can join on “Id” in DatasetVersion.csv

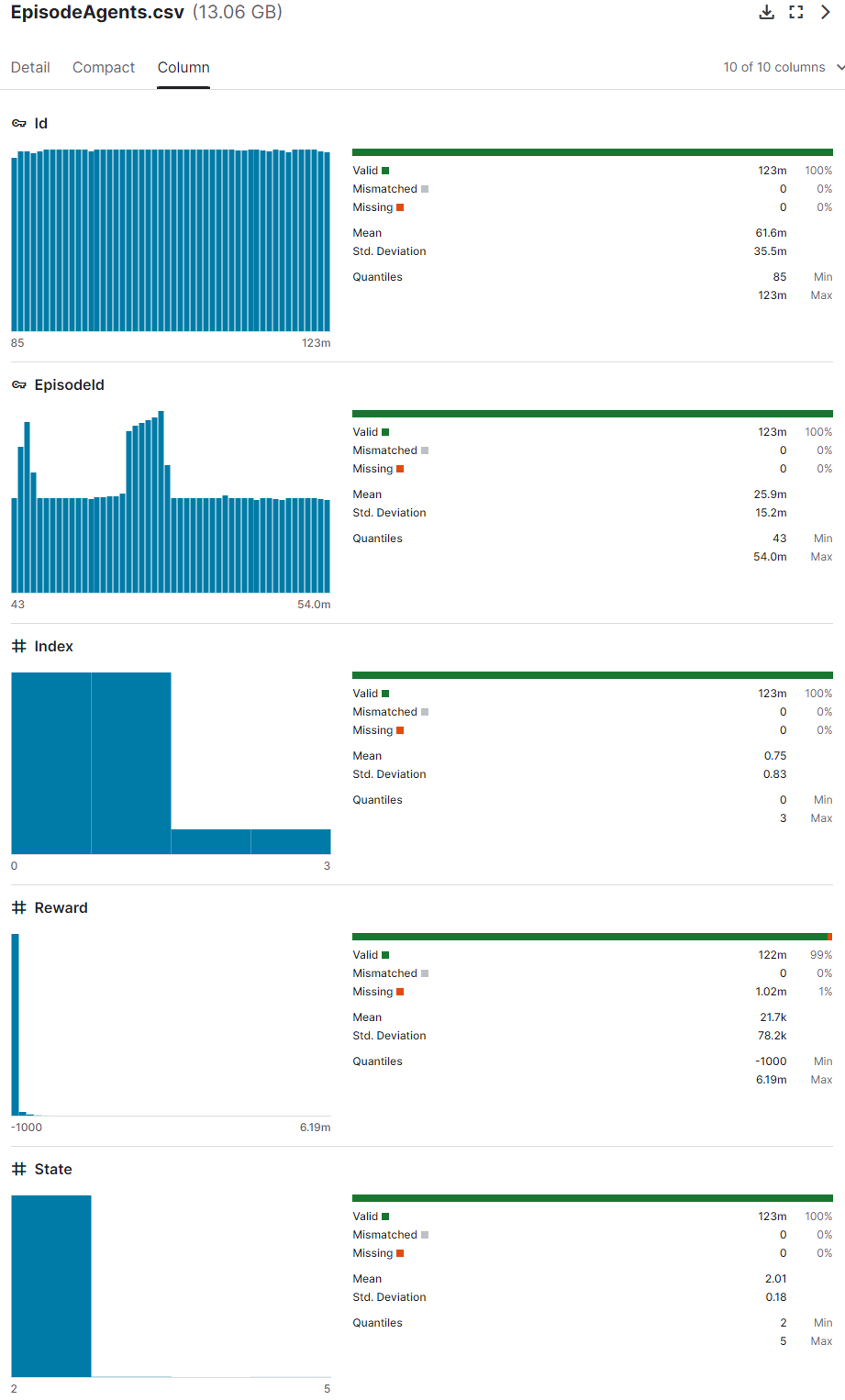
# Datasources.csv

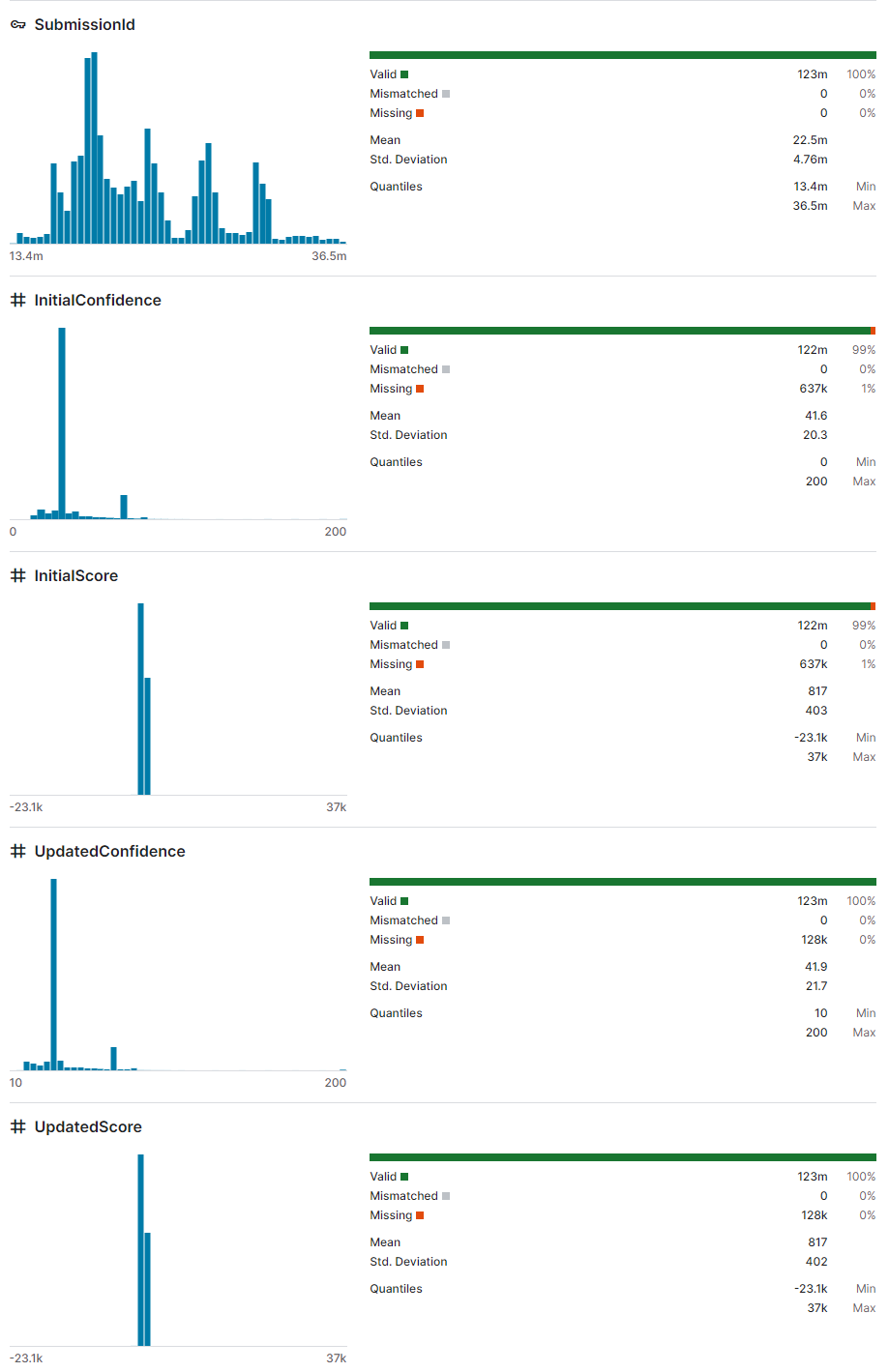


* “CurrentDatasourceVersionId” can join on “CurrentDatasourceVersionId” in Datasets.csv

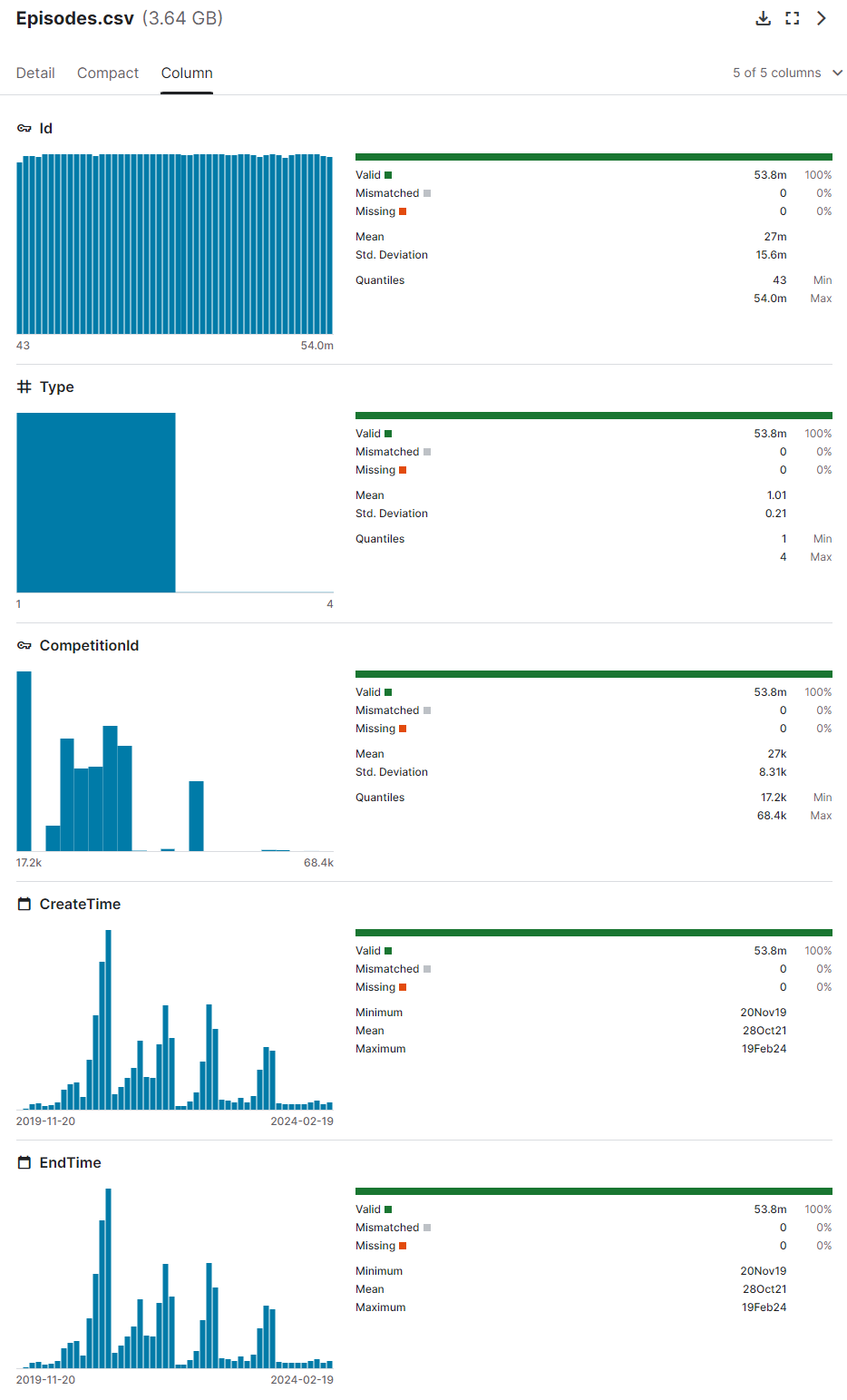
# EpisodeAgents.csv

EpisodesAgent.csv and Episodes.csv are data sets pertaining to reinforcement learning scores/competitions. Not sure if we will need to use these.





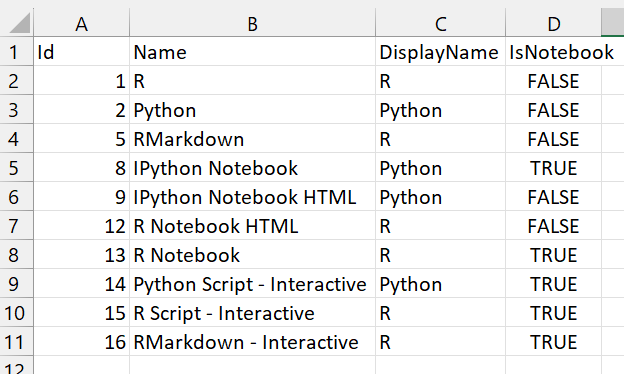
# Episodes.csv



# Forum Datasets

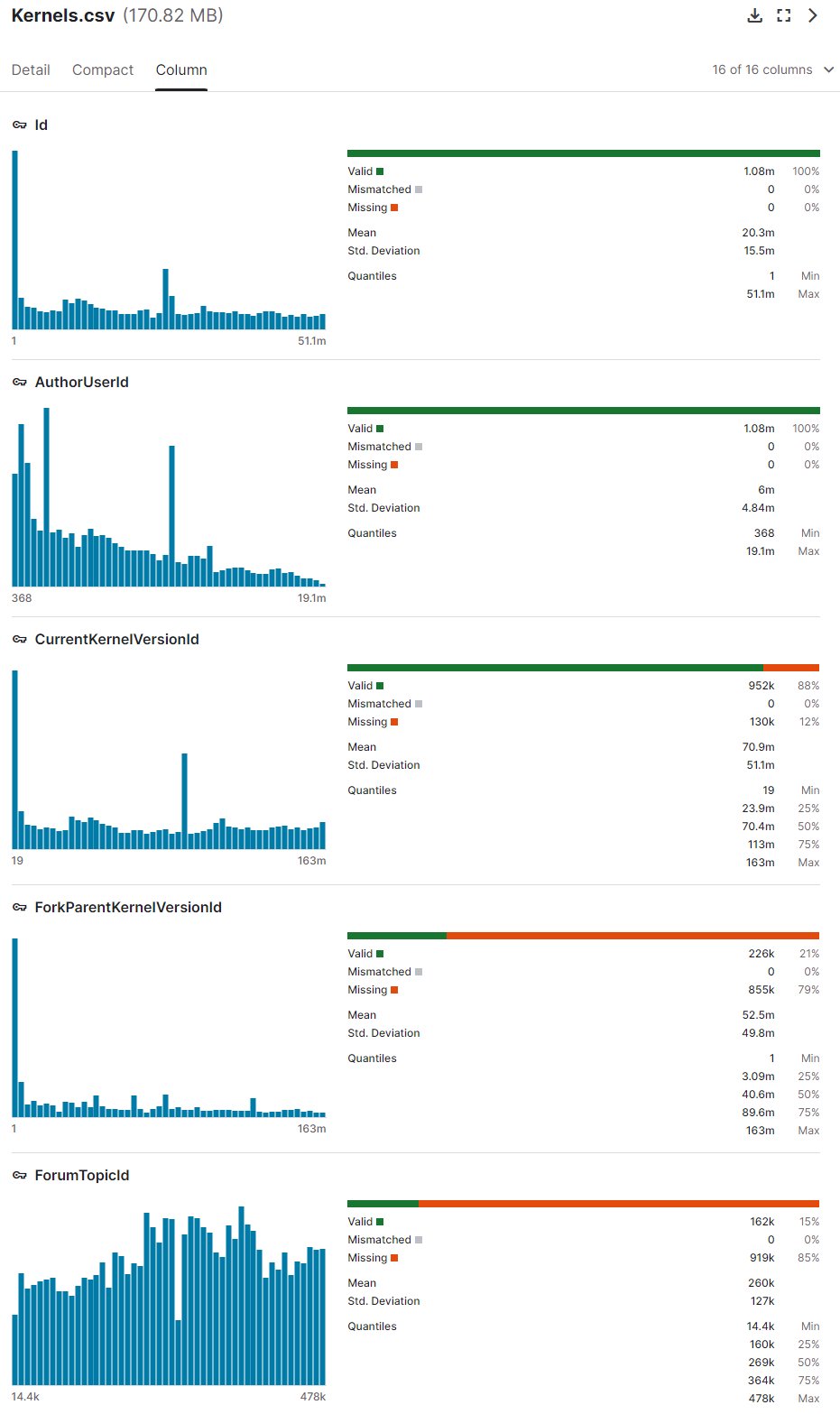
I think we can work with data involving user ID’s and the votes users gave to each other’s forum message; and medals earned by posts.

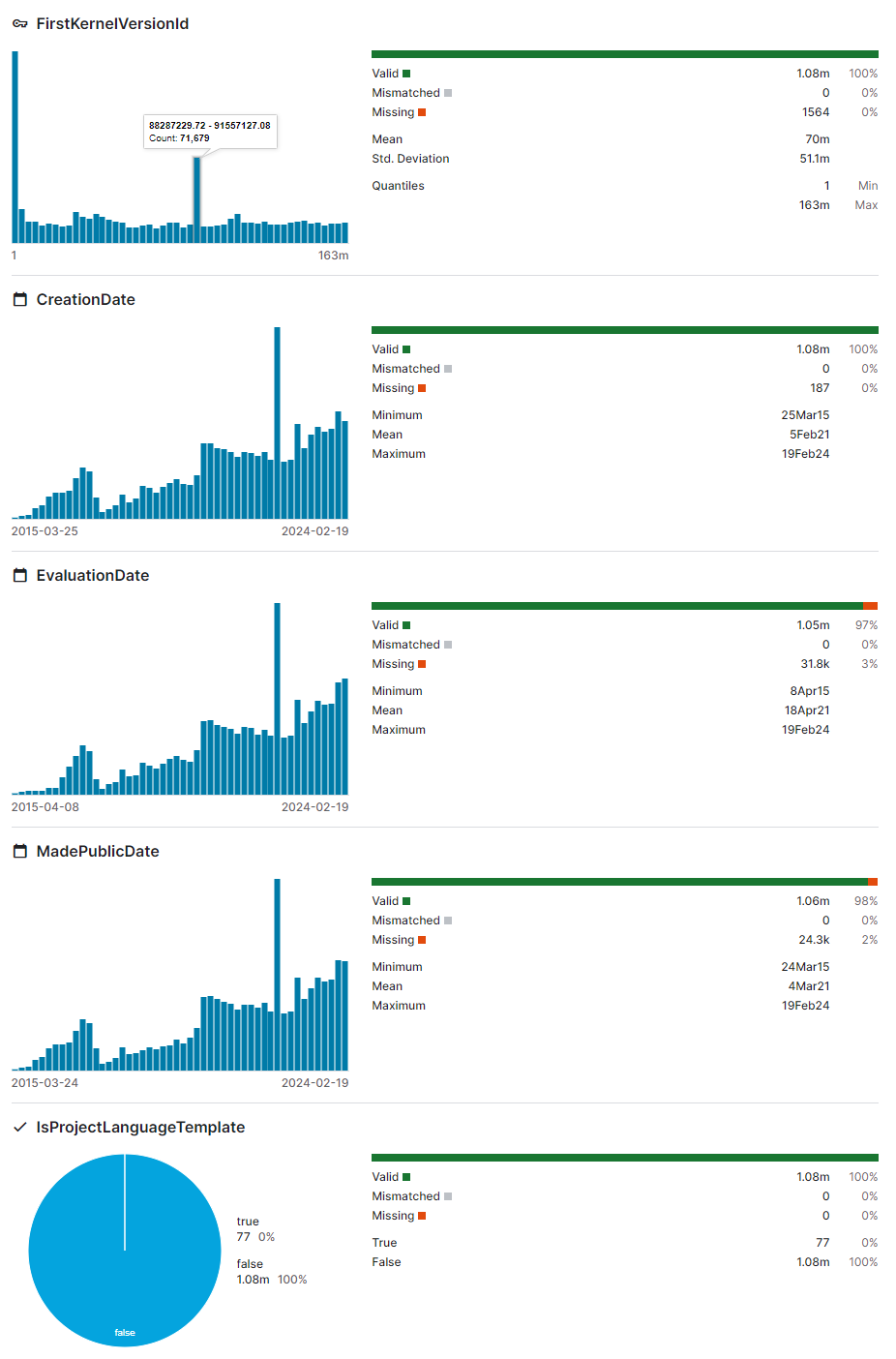
# KernelLanguages.csv

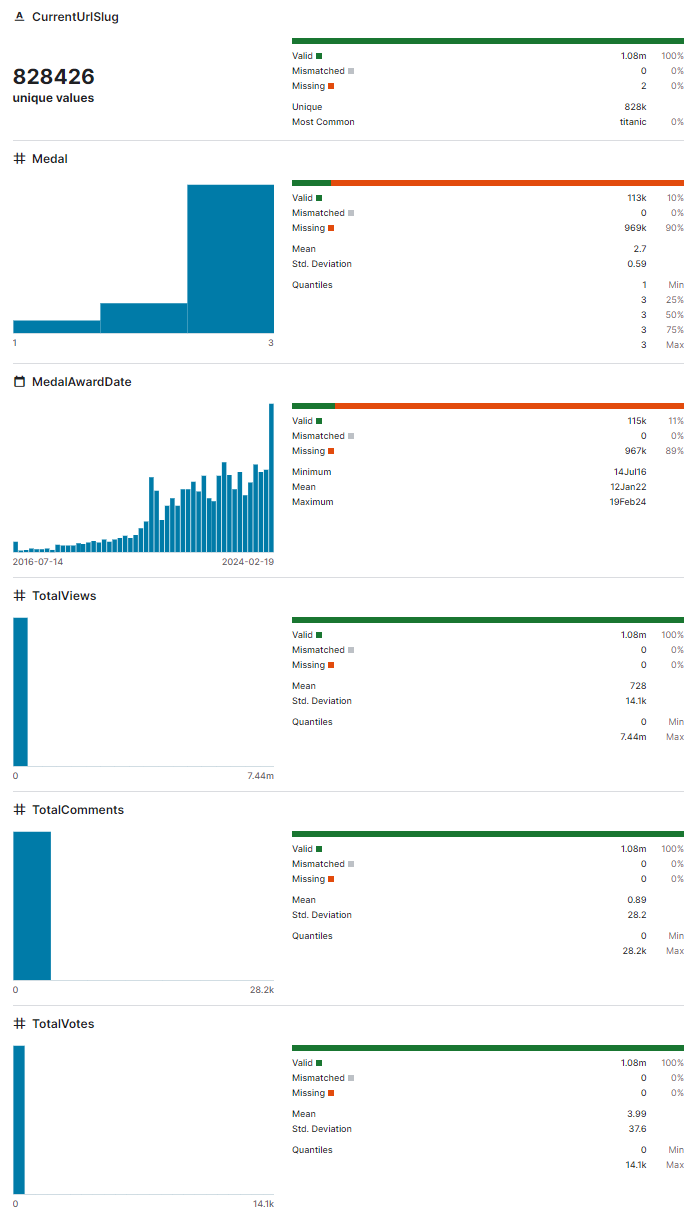


# Kernels.csv

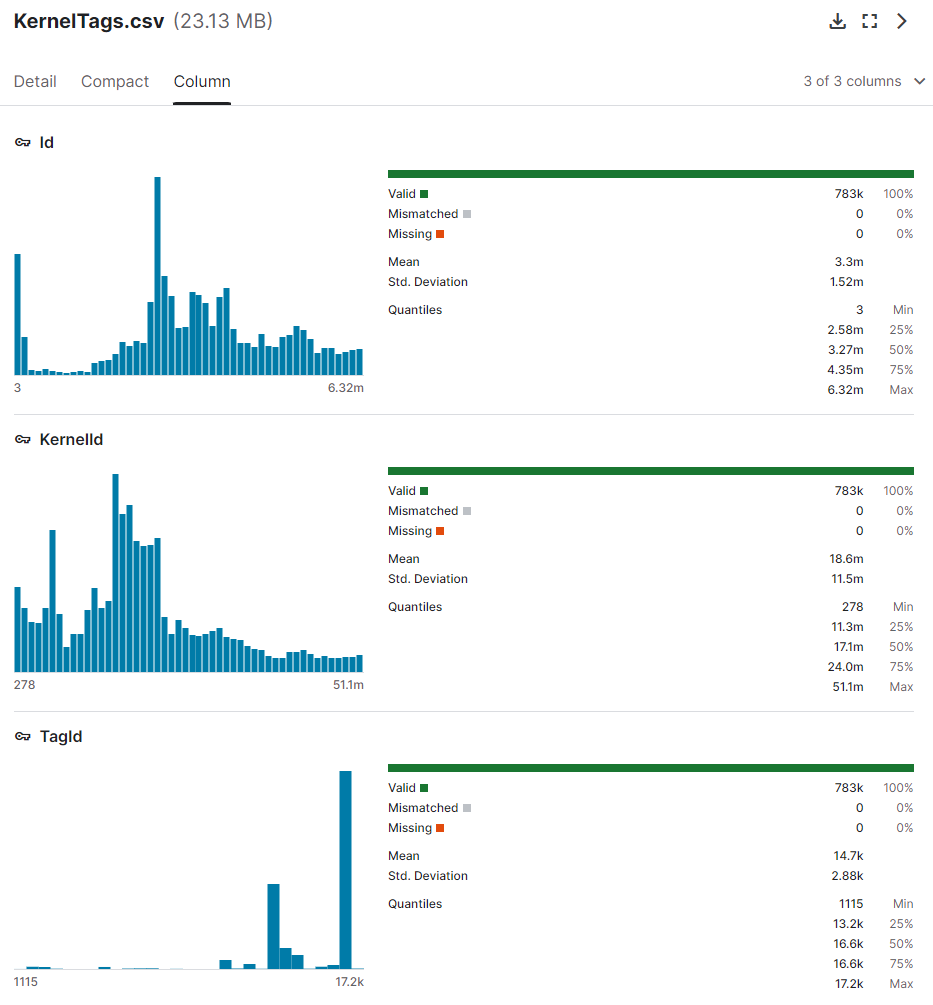
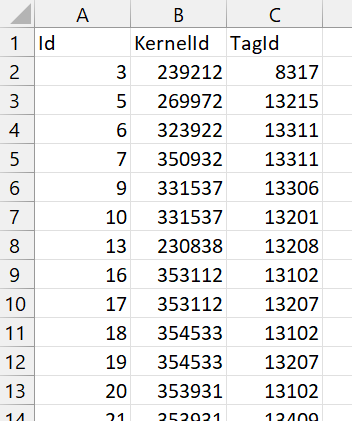
This csv contains 3 columns of kernel Id’s (CurrentKernelVersionId, ForkParentKernelVersionId, and FirstKernelVersionId), which correspond to “KernelId” column in DatasetTaskSubmission.csv. There are much fewer unique values in “KernelId” because the kernels used for submissions are much fewer than the total number of kernels that exist.





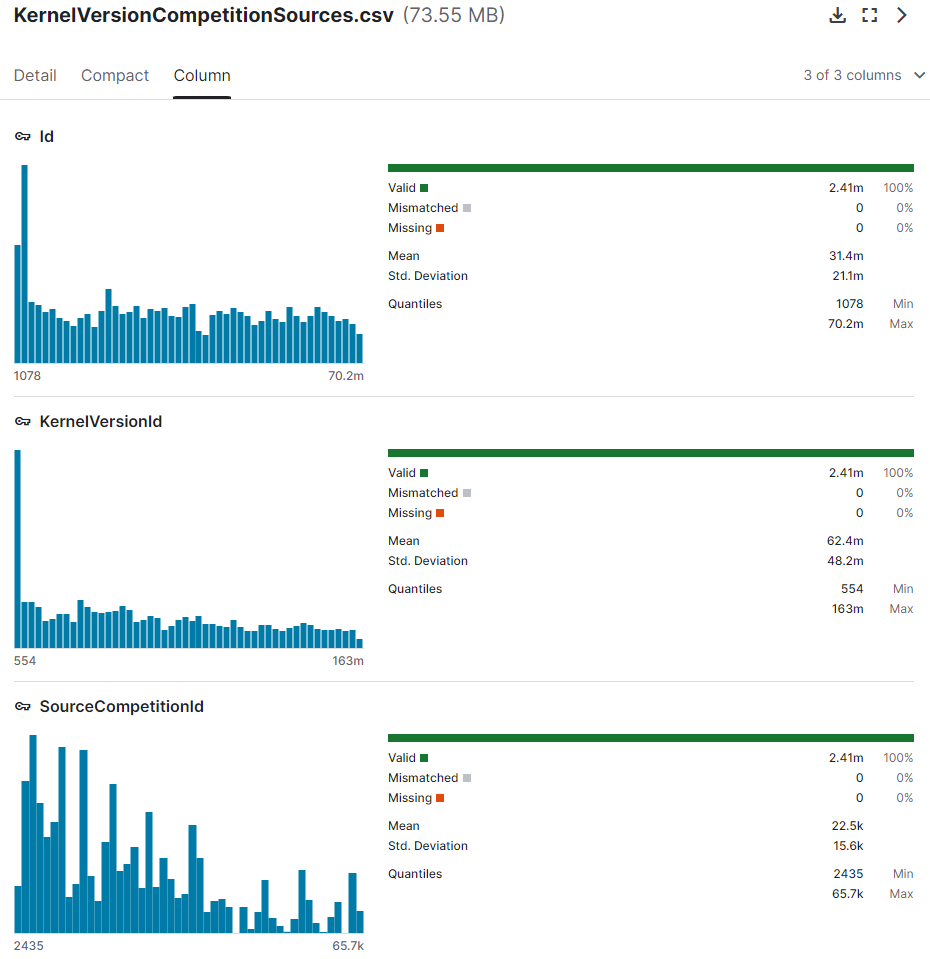
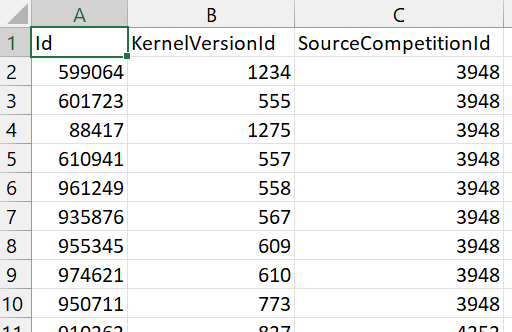


# KernelTags.csv



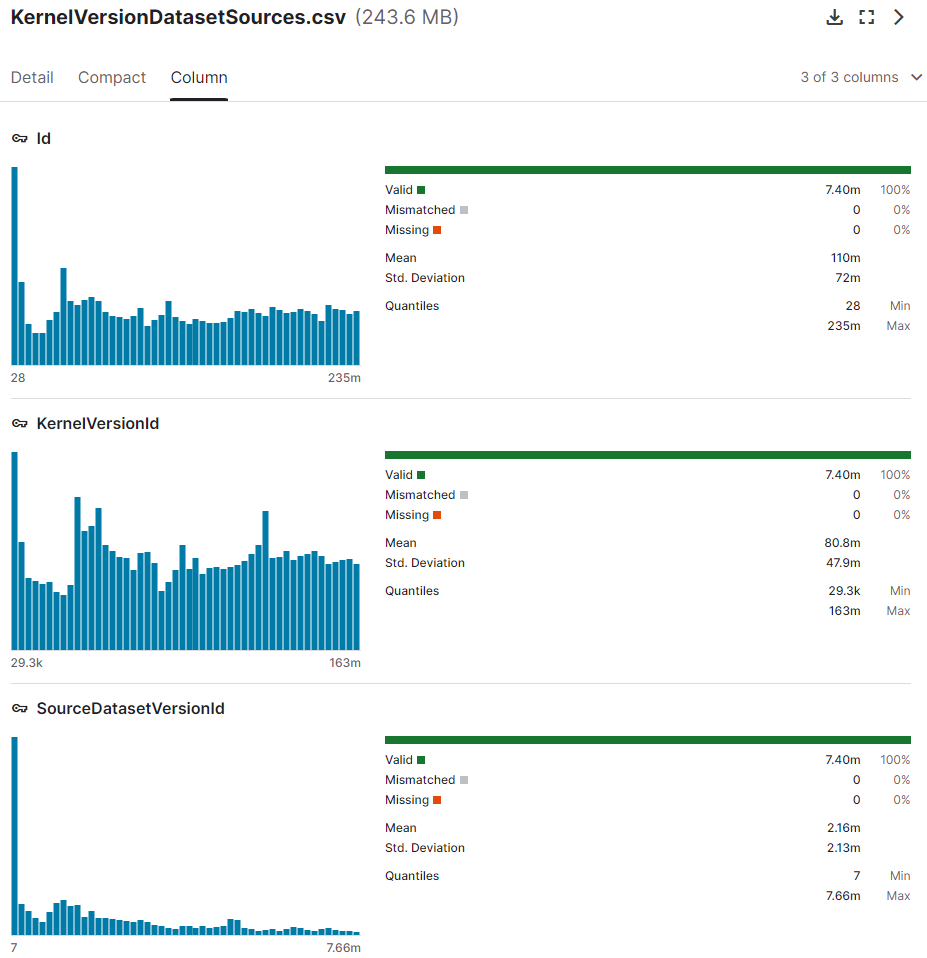
* KernelId corresponds to “Id” in Kernels.csv
* TagId corresponds to “TagId” in DatasetTags.csv

# KernelVersionCompetitionSources.csv



* KernelVersionId corresponds to (CurrentKernelVersionId, ForkParentKernelVersionId, FirstKernelVersionId) in Kernels.csv
* SourceCompetitionId can join on “Id” in Competitions.csv

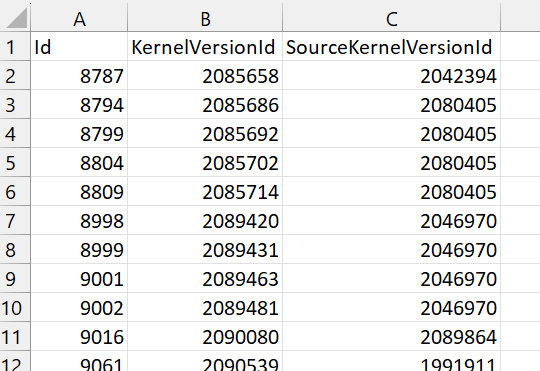
# KernelVersionDatasetSources.csv

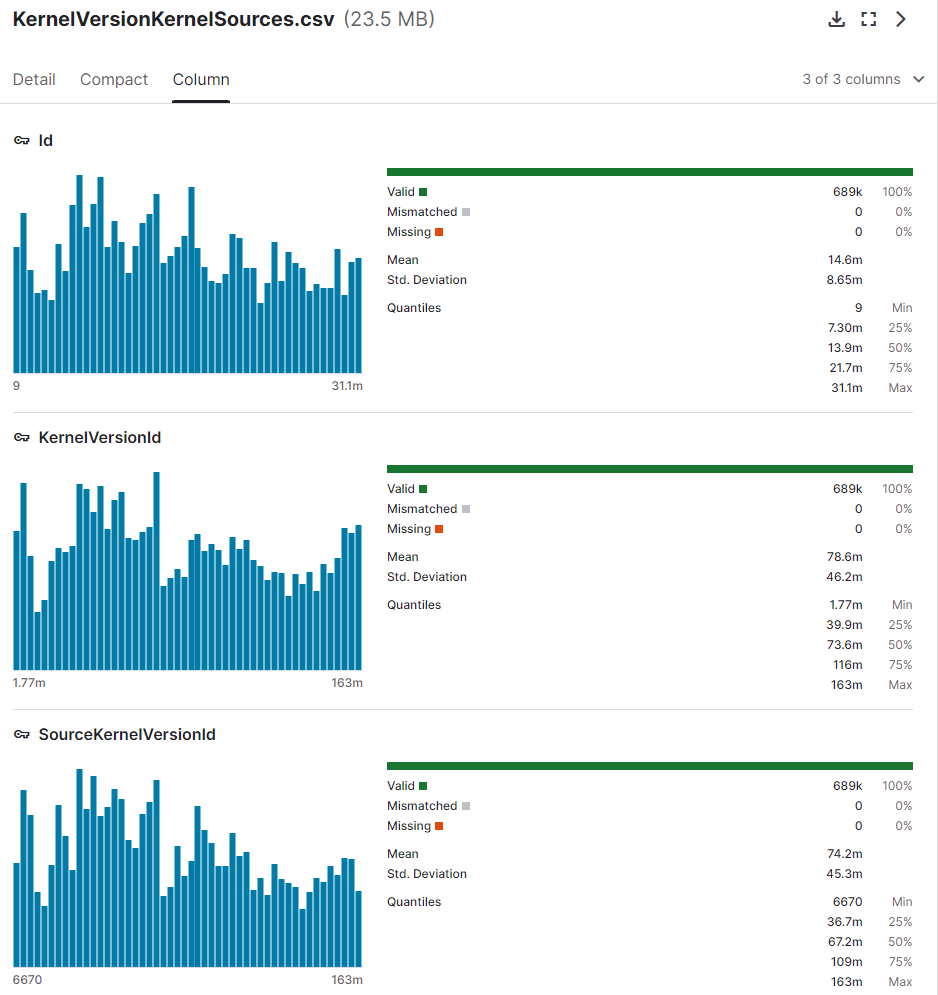


* KernelVersionId corresponds to (CurrentKernelVersionId, ForkParentKernelVersionId, FirstKernelVersionId) in Kernels.csv
* SourceDatasetVersionId corresponds to Id in DatasetVersions.csv

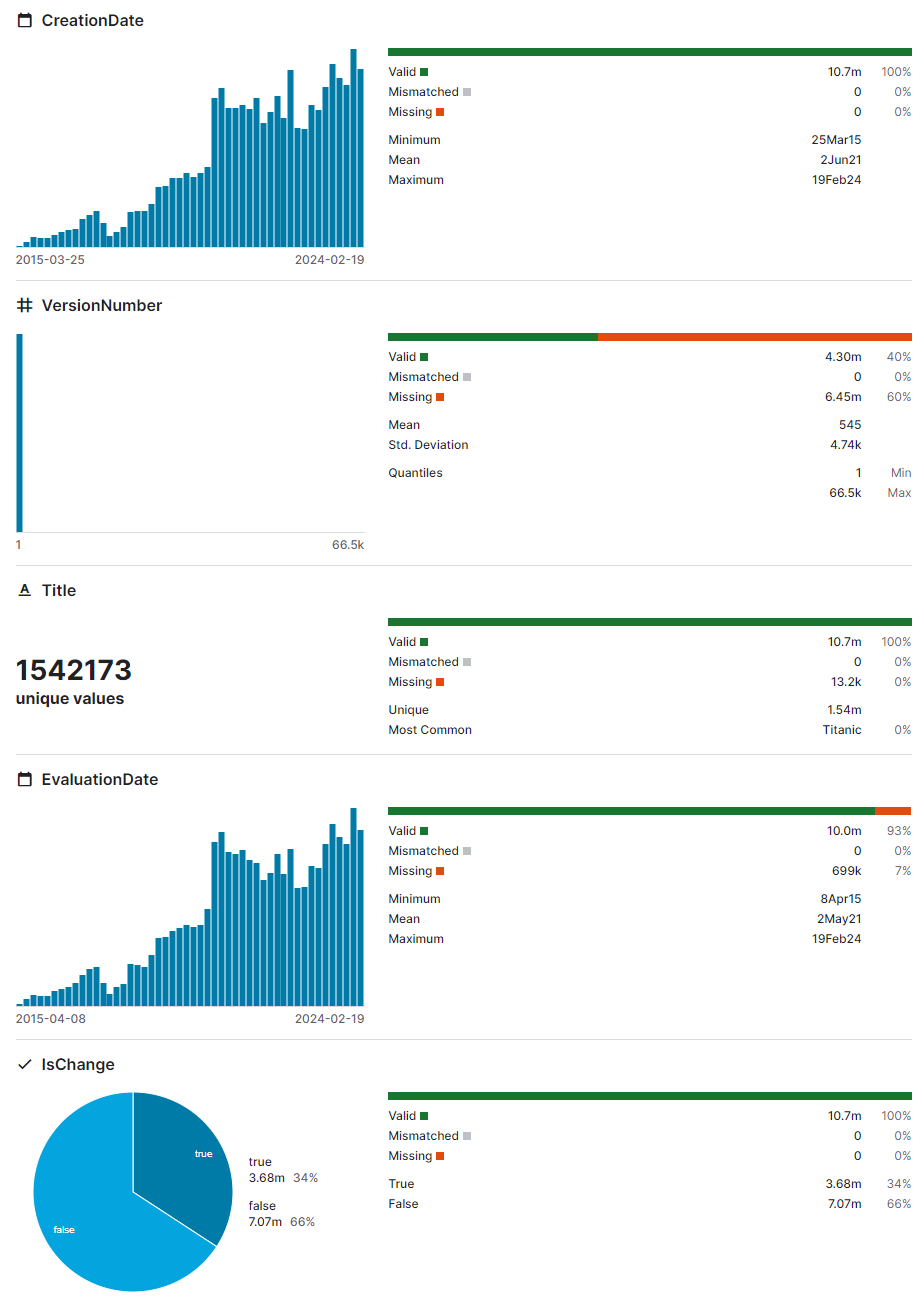
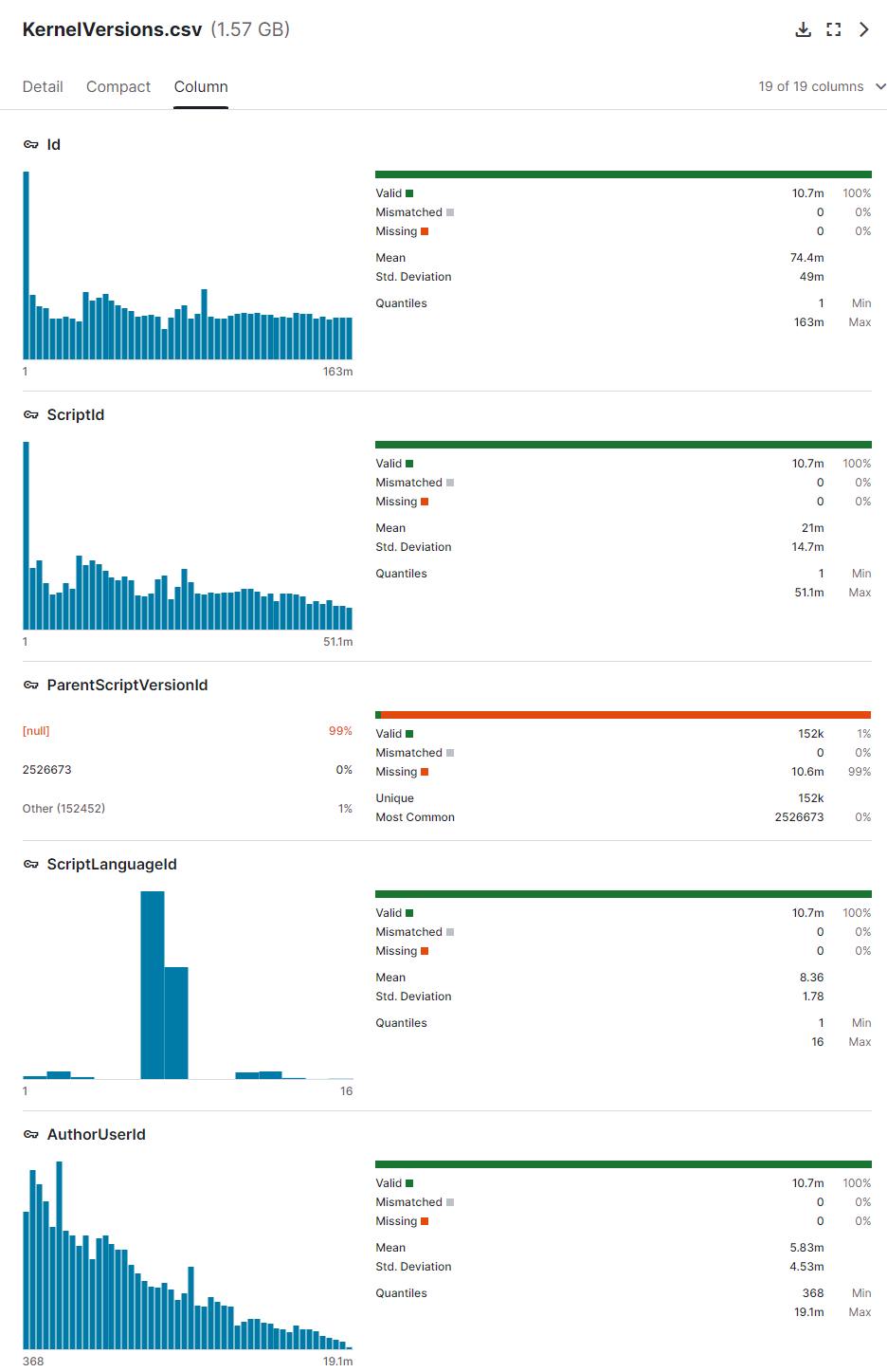
# KernelVersionKernelSources.csv

This data set pairs every kernel version ID with its associated source kernel version ID.



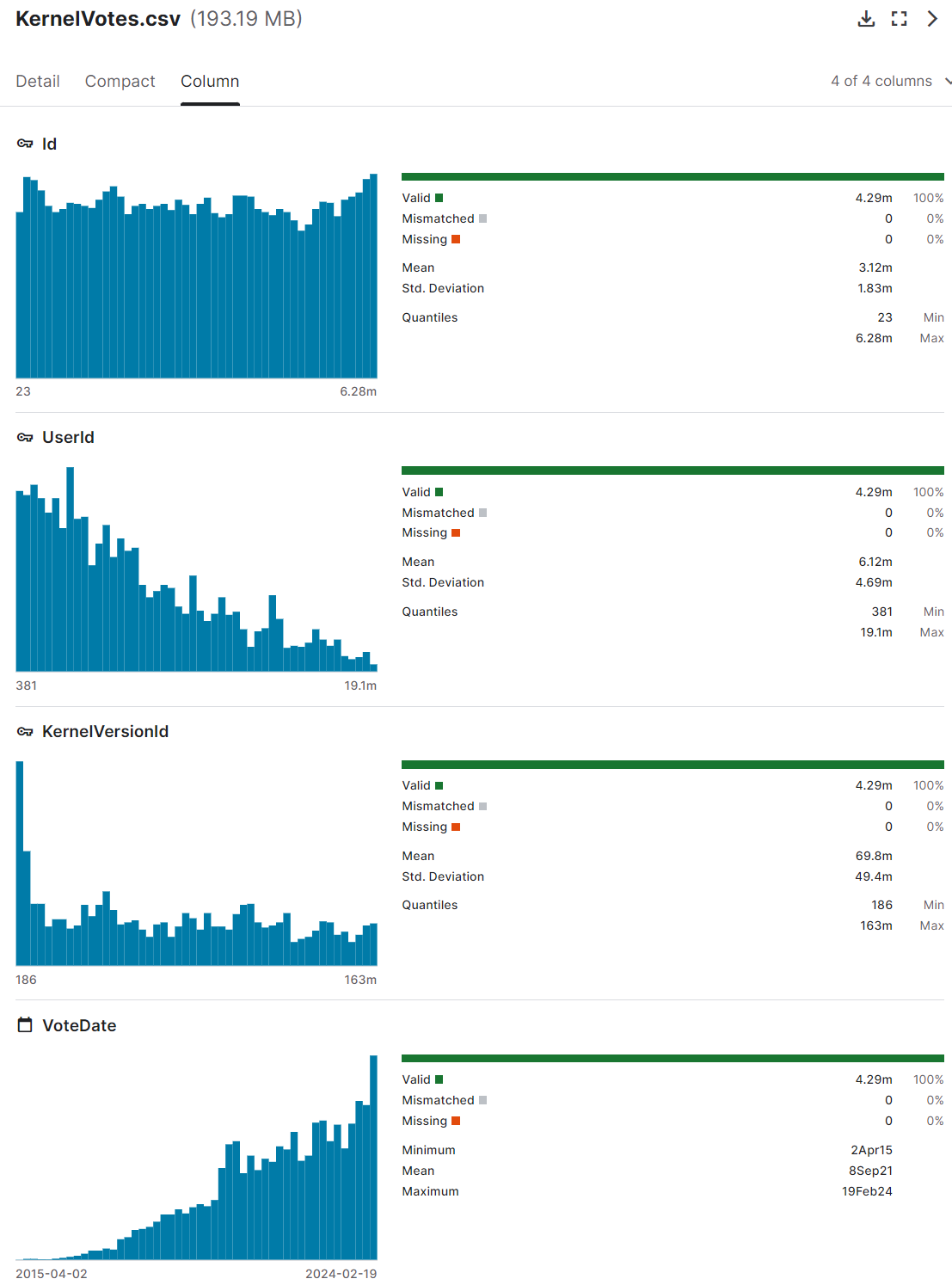
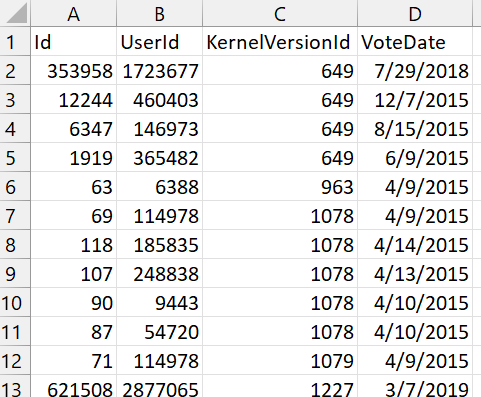


# KernelVersions.csv



* I did not include a snip of the rest of variables (seem pretty straightforward)
* “Id” column here can join on all the different kernel version ID columns in other data sets, like (CurrentKernelVersionId, ForkParentKernelVersionId, FirstKernelVersionId).
* ScriptId: a script is code written in an editing interface vs notebook has cells. It seems that one script can be associated with multiple kernel versions. Not sure how we will use it for our project.
* ScriptLanguageId can join on “Id” in KernalLanguages.csv

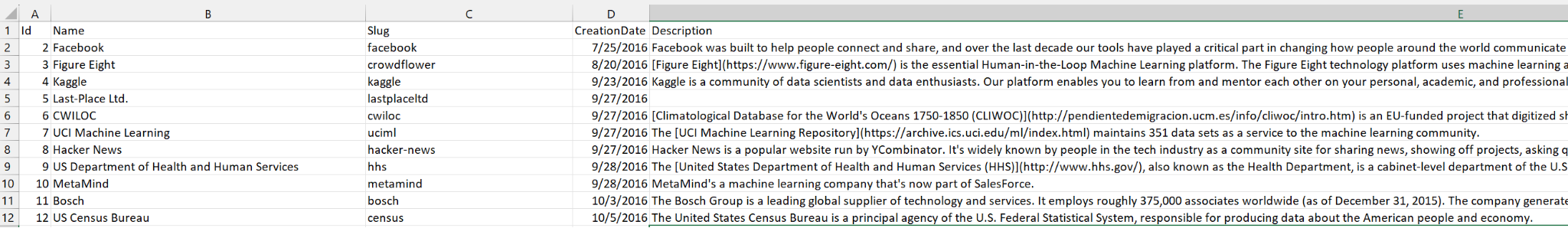
# KernelVotes.csv

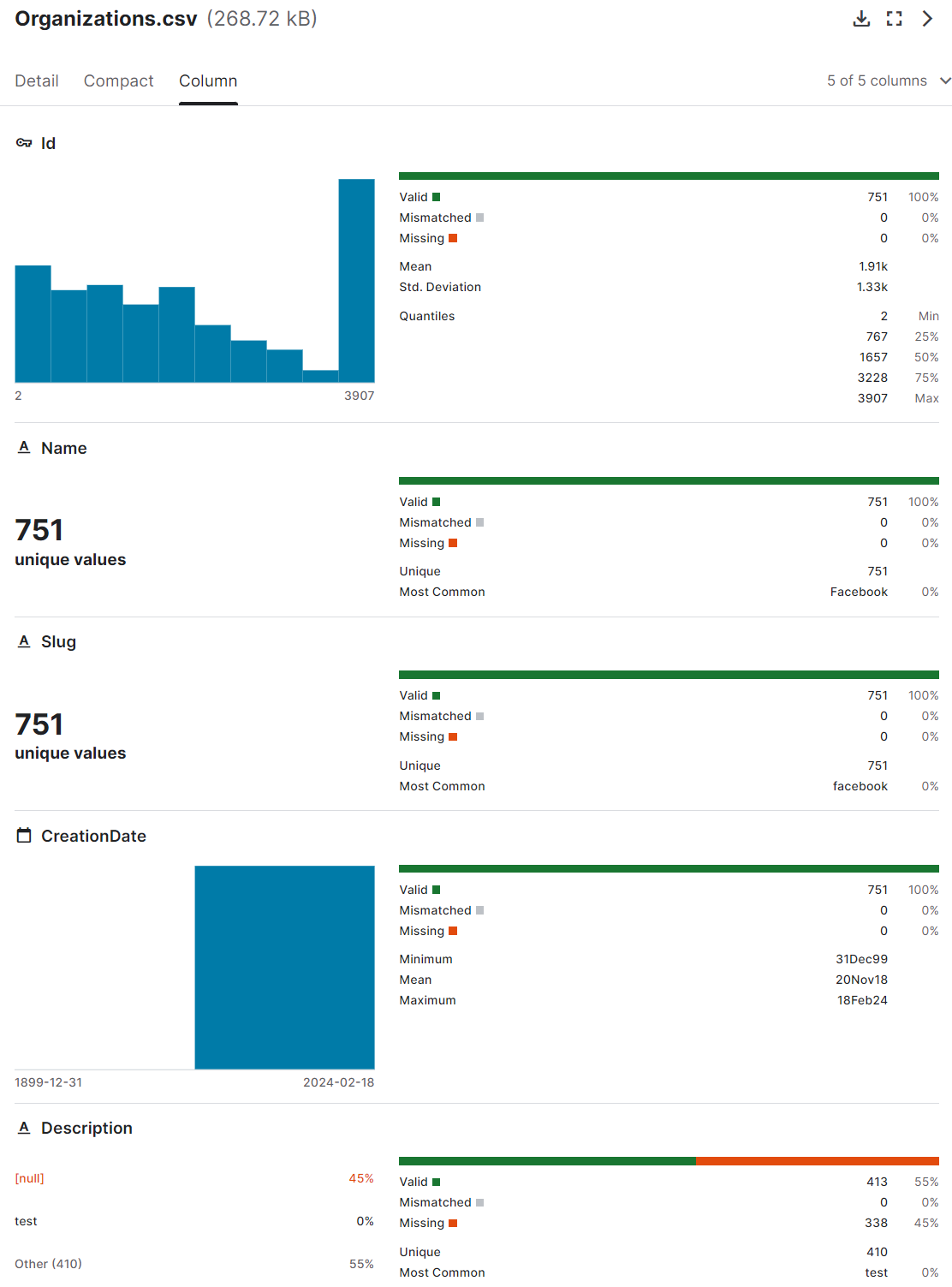


* UserId is associated with all other User ID variables in other data sets.
* KernelVersionId can join on “Id” in KernelVersions.csv

# Organizations.csv

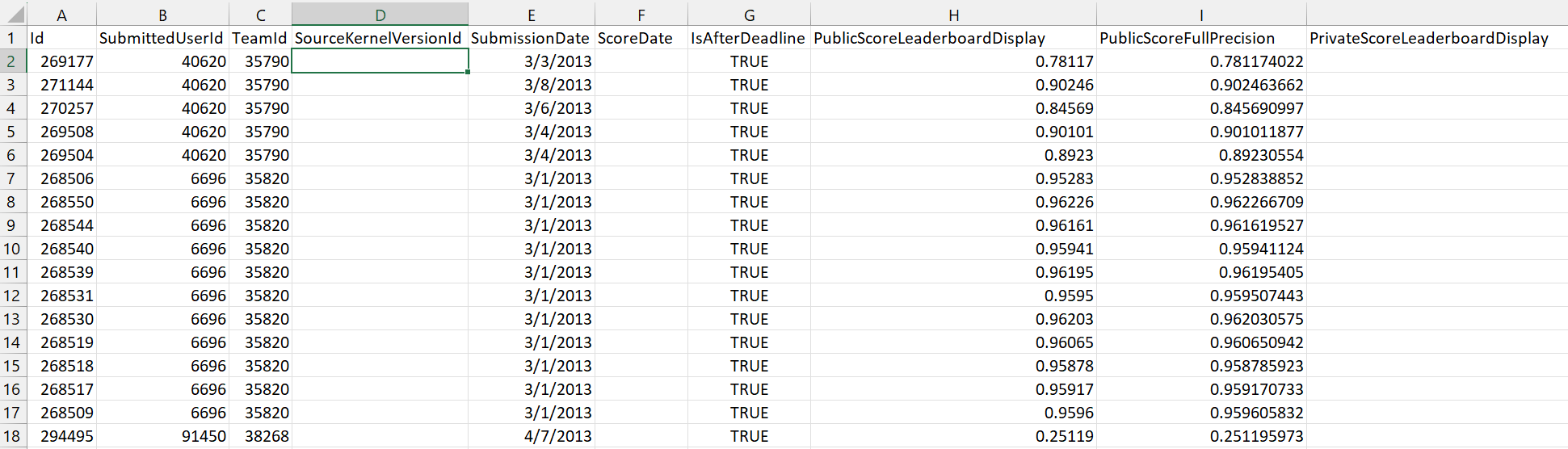
List of organizations associated with Kaggle.





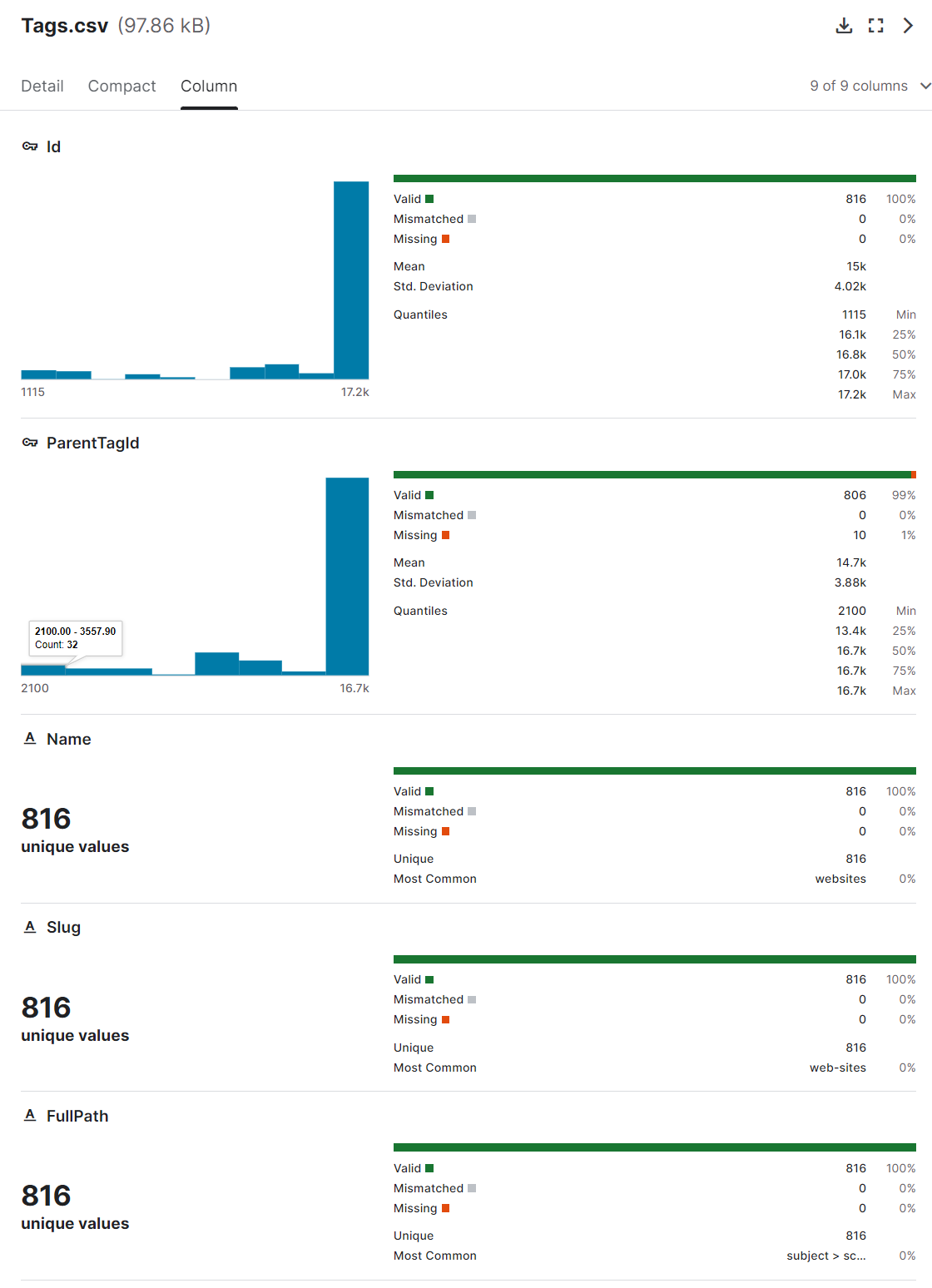
* “Id” can join on “OrganizationId” in Competitions.csv

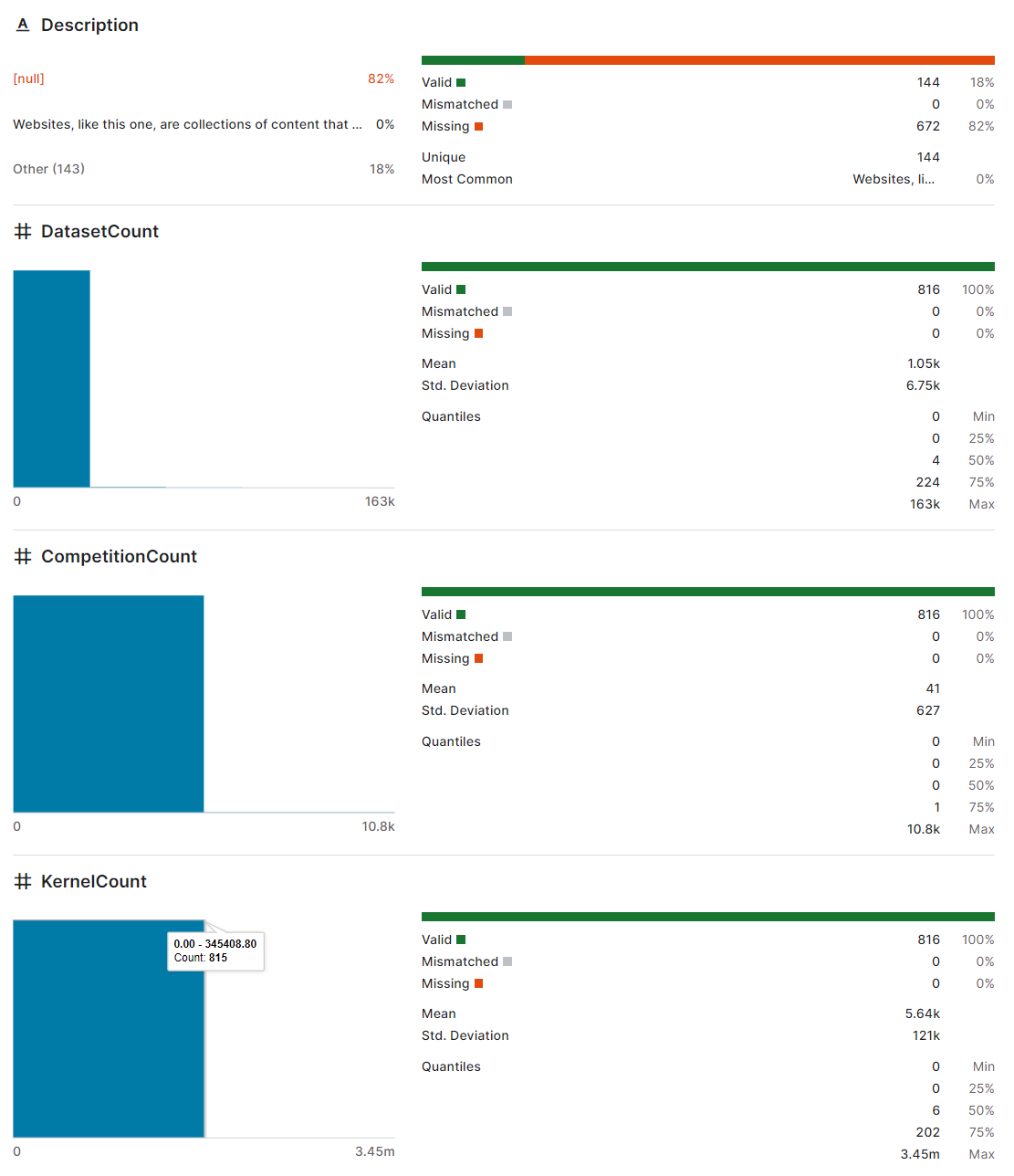
# Submissions.csv



We don’t know what kind of evaluation metric was used for the scores listed here. To determine the metric, we can join this table with Teams.csv (on Submission.TeamId = Teams.Id). Then join this with Competitions.csv (on Team.CompetitionId = Competitions.Id). Competitions.csv contains evaluation metrics data.

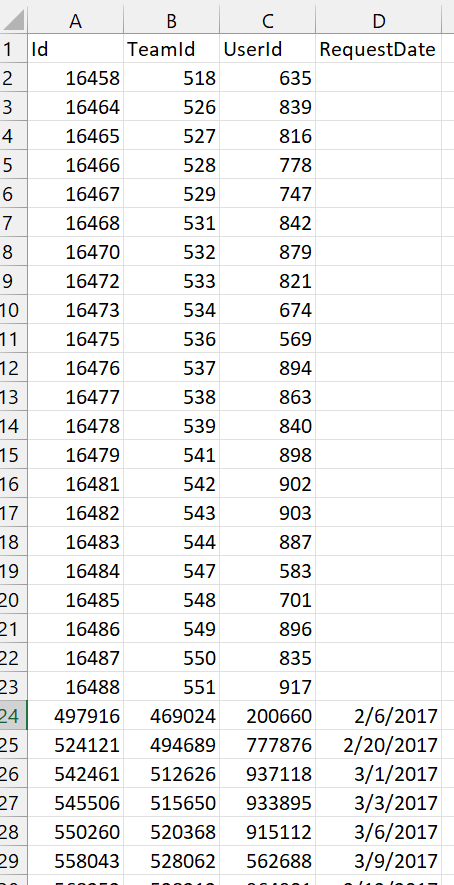
# Tags.csv





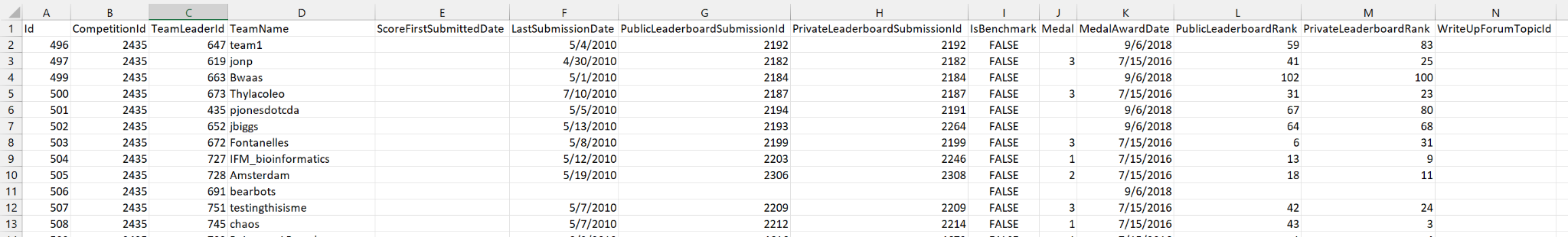
* Id: can join on “TagId” in CompetitionTags.csv
* ParentId: This csv’s Id column is the subcategory of its associated ParentId.
* Name: tag (or category) name, e.g., art, football, healthcare, united states, etc.

# TeamMembership.csv



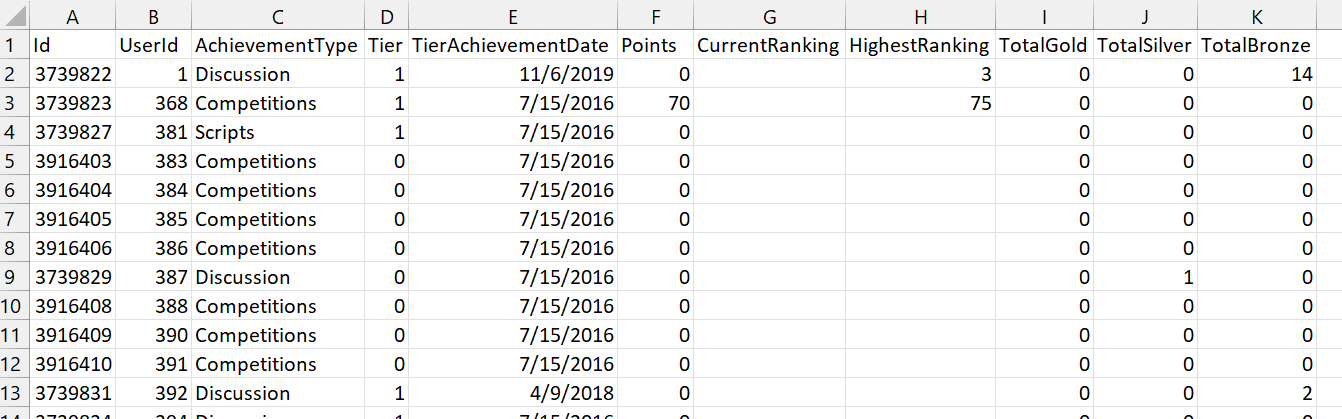
I’m not sure what exactly this data set will be used for.

# Teams.csv



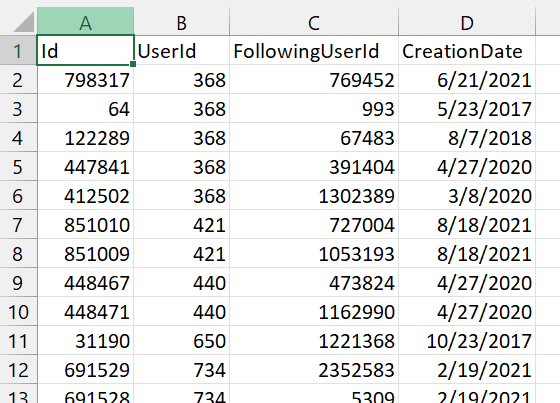
* Id can join on “TeamsId” in Submissions.csv
* CompetitionId can join on “Id” in Competitions.csv
* TeamLeaderId can join on all the user ID variables.

# UserAchievements.csv



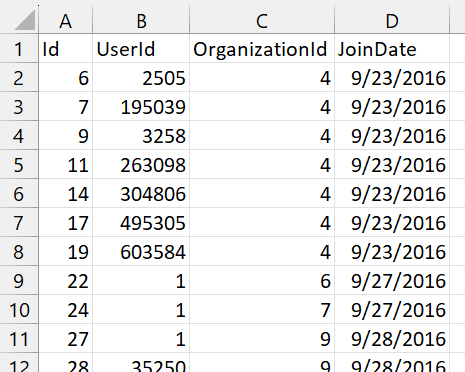
UserId can be joined on any of the user-ID-based columns in other data sets.

# UserFollowers.csv



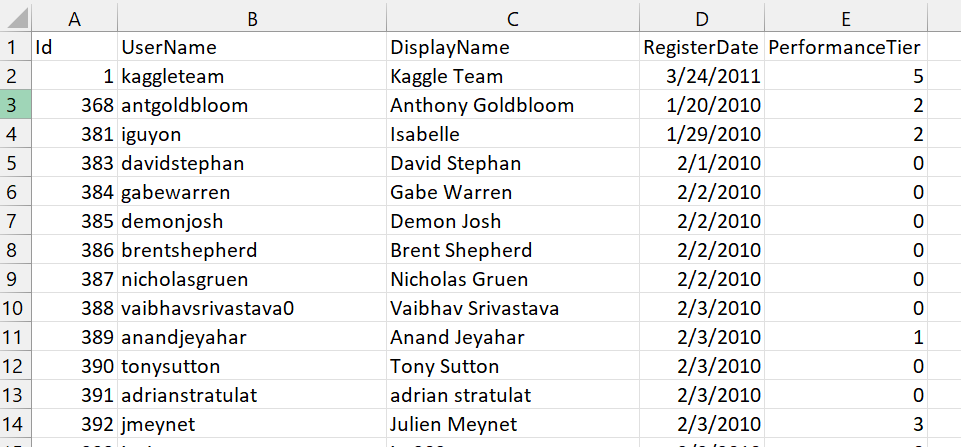
* “UserId” and “FollowingUserId” are user IDs.

# UserOrganizations.csv



* OrganizationId joins with “Id” in Organizations.csv

# Users.csv



* The Id column here must be the primary keys for all user ID’s.