Data Management Report

Goal A: All Kaggle Dataset

By

Siddharth Bammani (11011885)

Introduction:

The project report presents the data profiling and cleaning of ‘All Kaggle Datasets‘ dataset. The profiling and cleaning the dataset is explained in the report along with the chosen approach. The results are represented in a graphical way, the cleaning of each data column has been documented. The process of data profiling and cleaning based on the quality dimension can be reflected through the report.

Data Profiling:

The given dataset is analyzed of information in order to clarify the structure, content, relationships, and derivation rules of the data. The dataset is the complete collection of all the datasets that are published on Kaggle, in one csv file. For basic understanding of the dataset the column variables are analyzed, and their meaning is evaluated. For profiling of the data, we are making use of the Talend Data Quality which is an open source software to quickly profile and process the data.

Below are the column variables of the given dataset:

categories	dateUpdated	maintainerOrganization
commonFileTypes	diffType	overview
creatorName	downloadCount	ownerName
creatorUrl	isCollaborator	ownerUrl
creatorUserId	isDeleted	ownerUserId
currentDatasetVersionId	isFailed	title
currentDatasetVersionNumber	isFeatured	viewCount
datasetId	isHidden	voteButton
datasetSize	isPrivate	
datasetUrl	isSuperFeatured	

We start by uploading the dataset into Talend Data quality software. The file is comma separated value, also after general exploration of data it is found that there are different types of data format present in the column variable among them three columns that are ‘categories’, ‘maintainerOrganization’ and ‘voteButton’ contain data in json format. So as we uploaded the CSV file the column count gets jumbled up due to comma in the data, hence we removed the two rows and did separate column analysis on them.

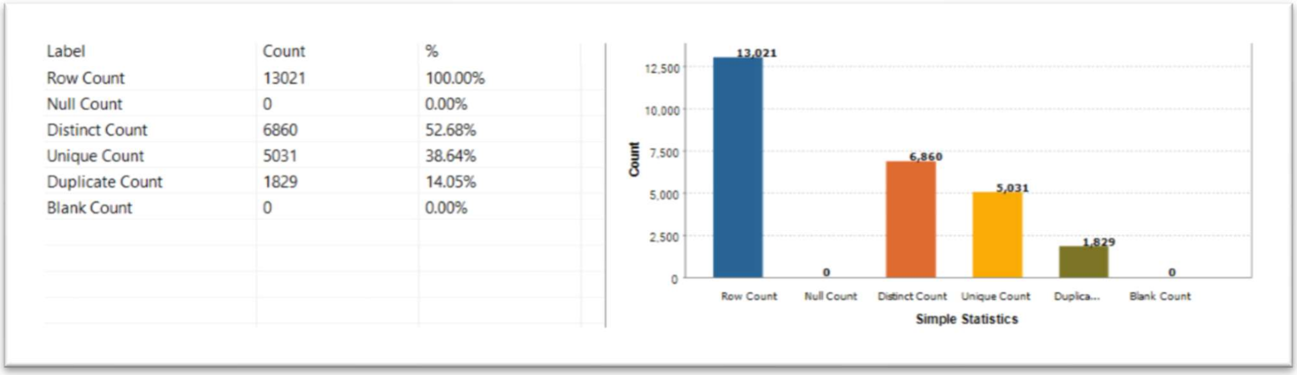
Each individual column is analyzed in Talend data quality software to find out the data errors, inconsistent formatting within a column, missing values, or outliers.

We start the data profiling of the dataset by **Column Analysis**, the method includes task to examine individual, atomic values and determining if they are valid or not. We are carrying out two tasks majorly to profile the data that are Basic Column Analysis which provides us the simple statistics, and another task is Pattern frequency analysis which provides the pattern frequency statistics.

The Basic Column Analysis generates metadata that comprises of various count, such as # of values (size), # of distinct value (cardinality, distinctness), # of non-null values (completeness), maximum & minimum value. Below table shows the simple statistics result.

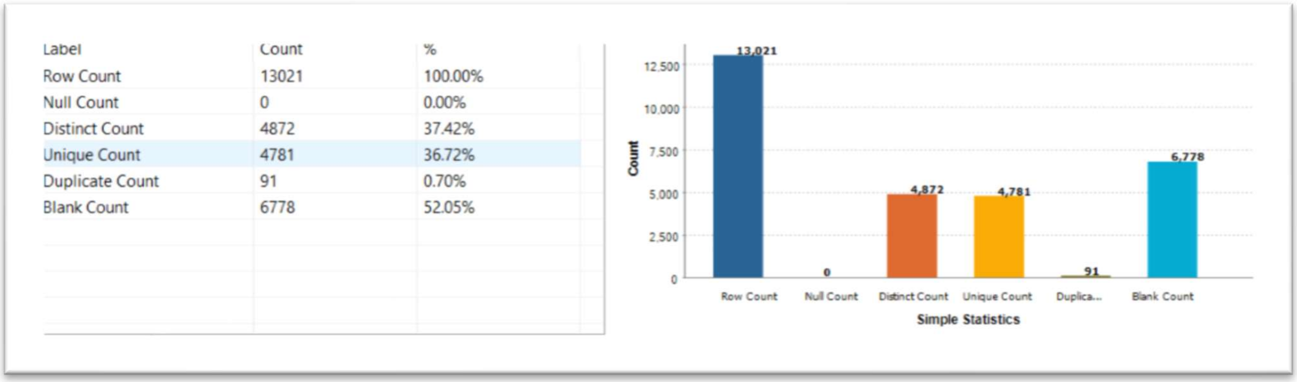
We found that we have in total 13,021 number of rows and 27 number of columns.

Below is the screenshot of the statistical table that displays the null, distinct and unique count of the column ‘title’.



Label	Count	%
Row Count	13021	100.00%
Null Count	0	0.00%
Distinct Count	6860	52.68%
Unique Count	5031	38.64%
Duplicate Count	1829	14.05%
Blank Count	0	0.00%

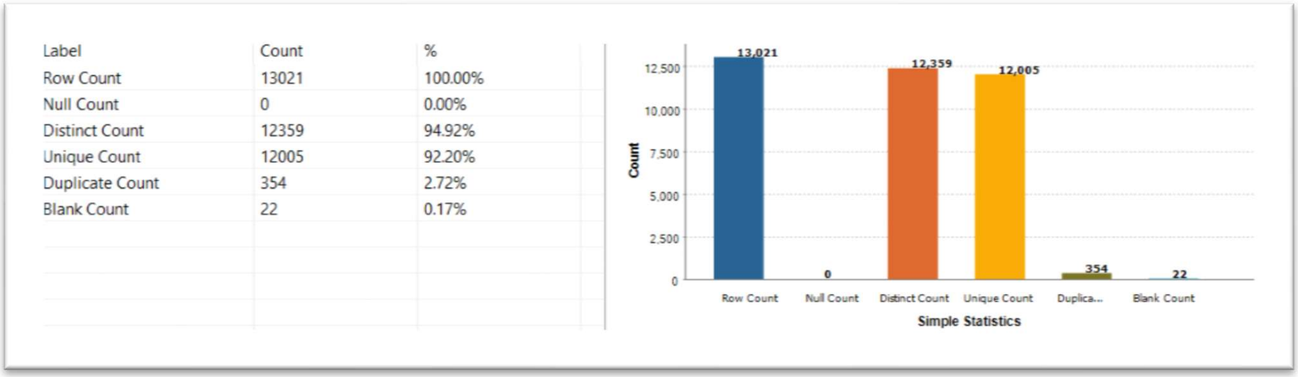
Below is the screenshot of the statistical table that displays the null, distinct and unique count of the column ‘ownerName’.



Label	Count	%
Row Count	13021	100.00%
Null Count	0	0.00%
Distinct Count	4872	37.42%
Unique Count	4781	36.72%
Duplicate Count	91	0.70%
Blank Count	6778	52.05%

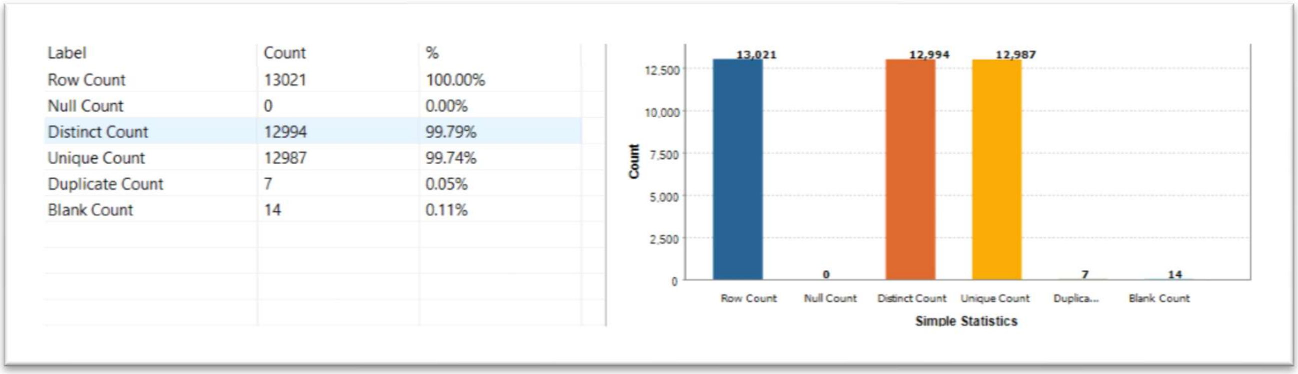
The variables ‘currentDatasetVersionId’, ‘datasetId’, ‘datasetUrl’ and ‘title’ has the maximum number of distinct and unique count. Below are the graph and the table which display the count.

The below graph is for the column ‘title’, here we can see 12,359 rows are distinct and the unique count is around 12,005.



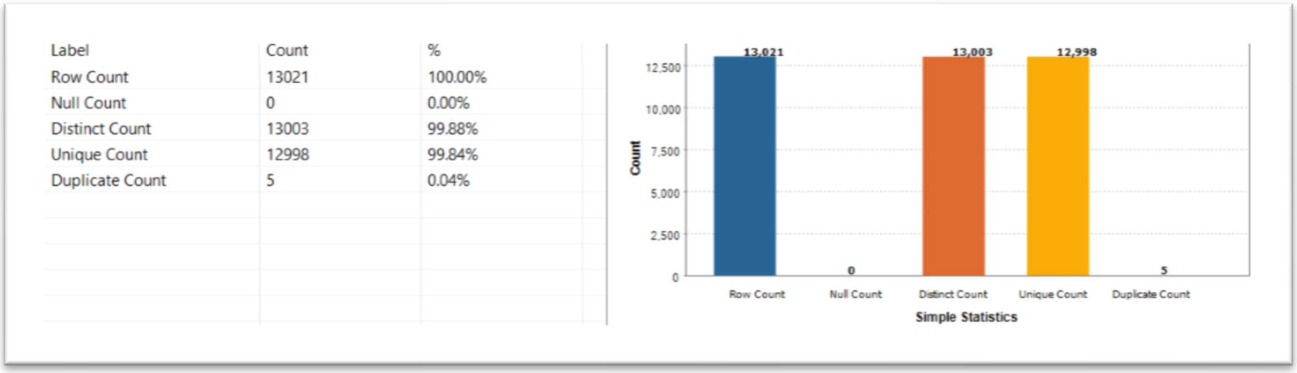
Label	Count	%
Row Count	13021	100.00%
Null Count	0	0.00%
Distinct Count	12359	94.92%
Unique Count	12005	92.20%
Duplicate Count	354	2.72%
Blank Count	22	0.17%

The below graph is for the column ‘currentDatasetVersionId’, here we can see 12,987 rows are unique and the total count is around 13,021 with 7 duplicate value.



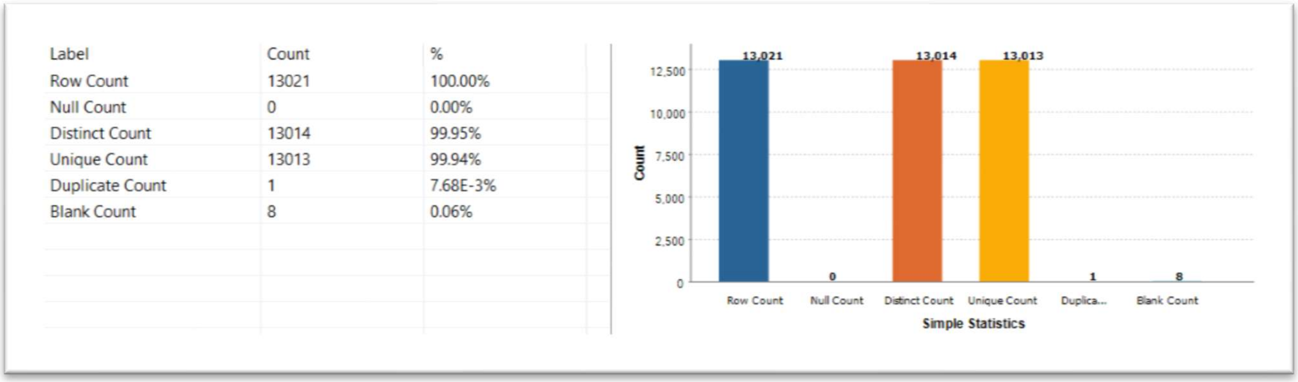
Label	Count	%
Row Count	13021	100.00%
Null Count	0	0.00%
Distinct Count	12994	99.79%
Unique Count	12987	99.74%
Duplicate Count	7	0.05%
Blank Count	14	0.11%

The below graph is for the column ‘datasetId’, here we can see 13,003 rows are distinct and the unique count is 12998.



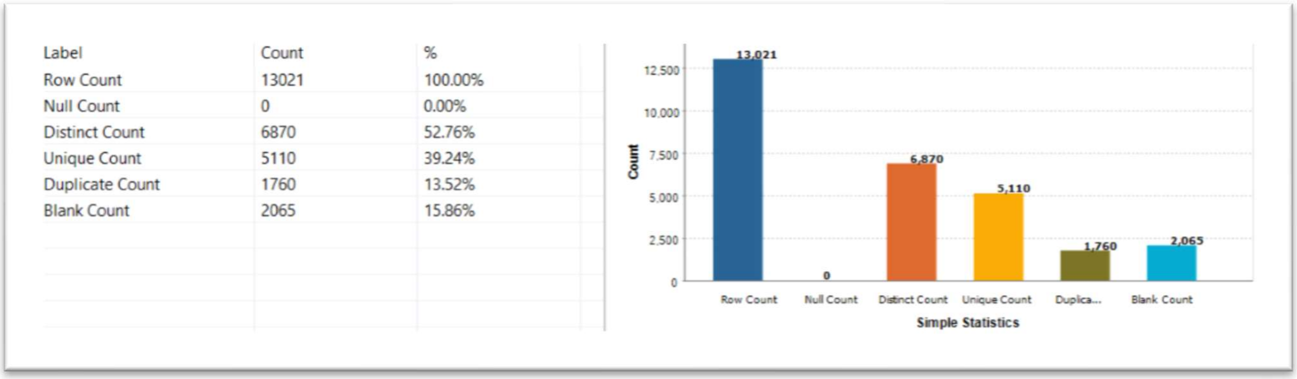
Label	Count	%
Row Count	13021	100.00%
Null Count	0	0.00%
Distinct Count	13003	99.88%
Unique Count	12998	99.84%
Duplicate Count	5	0.04%

The below graph is for the column ‘datasetUrl’, here we can see 13,014 rows are distinct and the unique count is 13,013.



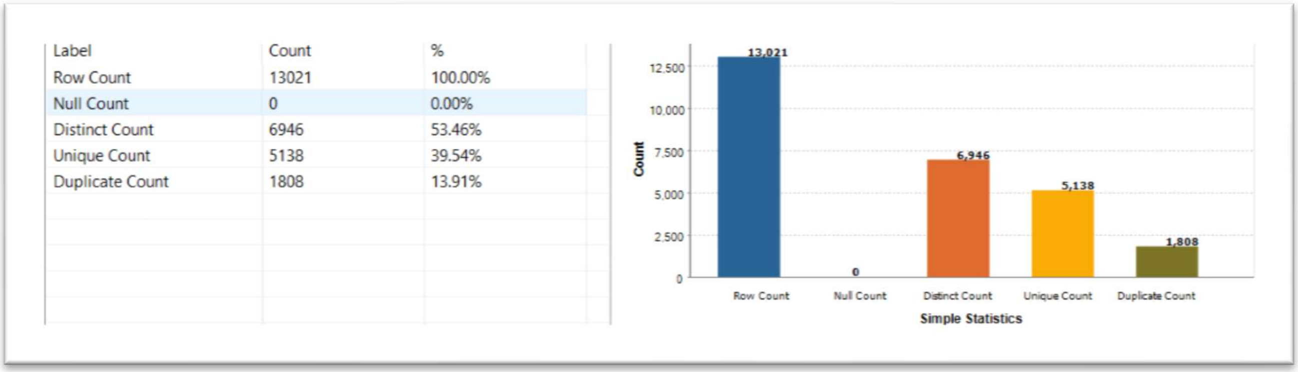
Label	Count	%
Row Count	13021	100.00%
Null Count	0	0.00%
Distinct Count	13014	99.95%
Unique Count	13013	99.94%
Duplicate Count	1	7.68E-3%
Blank Count	8	0.06%

The column ‘maintainerOrganization’ and ‘ownerUserId’ have the maximum number of blank values in them, below is the statistical representation of column ‘ownerUserId’.



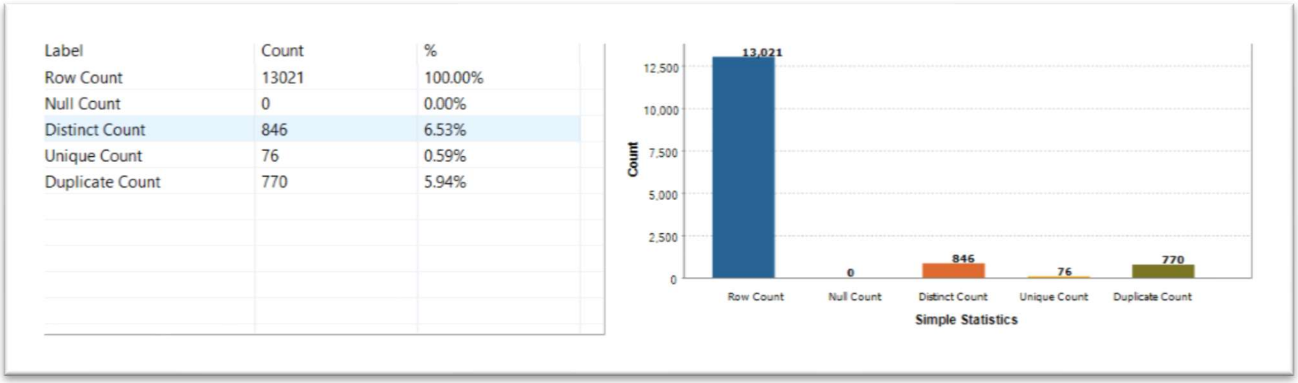
Label	Count	%
Row Count	13021	100.00%
Null Count	0	0.00%
Distinct Count	6870	52.76%
Unique Count	5110	39.24%
Duplicate Count	1760	13.52%
Blank Count	2065	15.86%

The below graph is for the column ‘creatorUserId’ , here we can see 6,946 rows are distinct and the unique count is around 5138 with 1808 duplicate value.



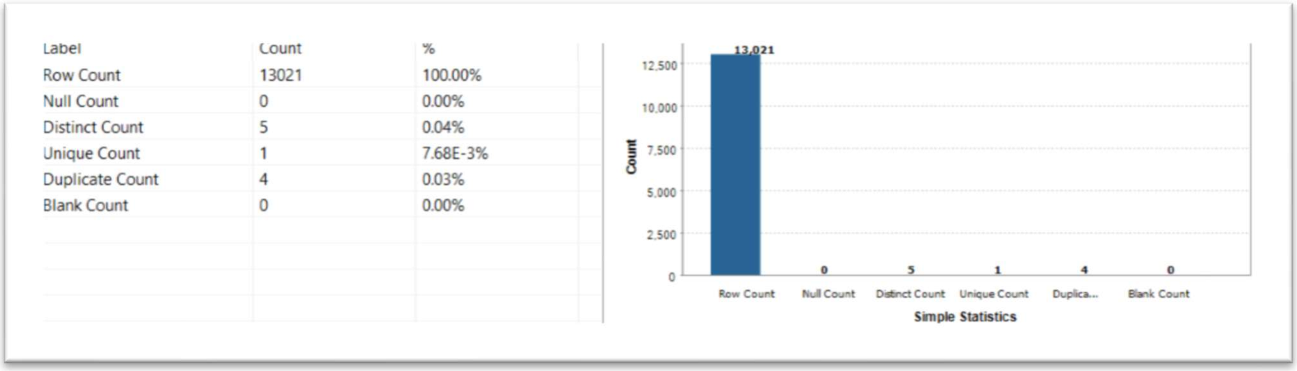
Label	Count	%
Row Count	13021	100.00%
Null Count	0	0.00%
Distinct Count	6946	53.46%
Unique Count	5138	39.54%
Duplicate Count	1808	13.91%

The below graph is for the column ‘dateUpdated’ , here we can see 846 rows are distinct and the unique count is around 76 with 770 duplicate value.



Label	Count	%
Row Count	13021	100.00%
Null Count	0	0.00%
Distinct Count	846	6.53%
Unique Count	76	0.59%
Duplicate Count	770	5.94%

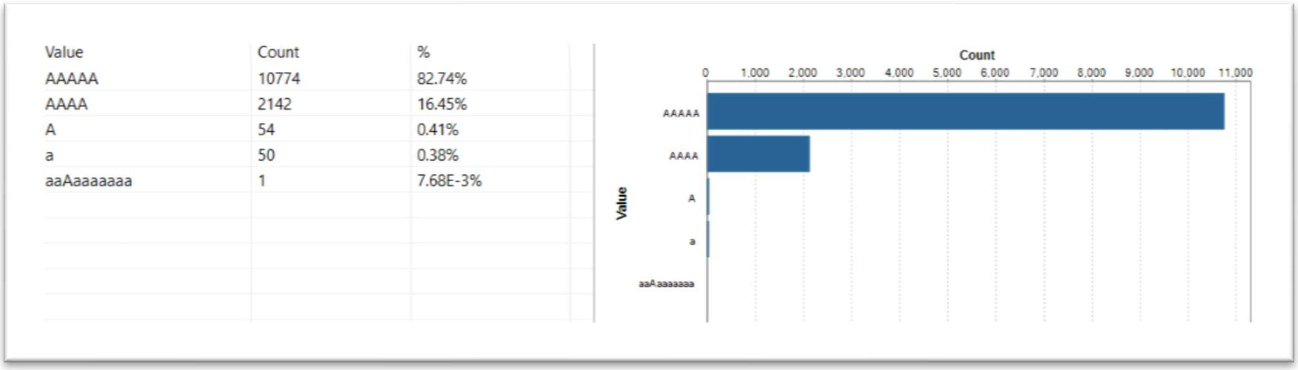
Below is the statistical representation of the column ‘isFeatured’ but most of the columns shows similar counts of indicators.



Label	Count	%
Row Count	13021	100.00%
Null Count	0	0.00%
Distinct Count	5	0.04%
Unique Count	1	7.68E-3%
Duplicate Count	4	0.03%
Blank Count	0	0.00%

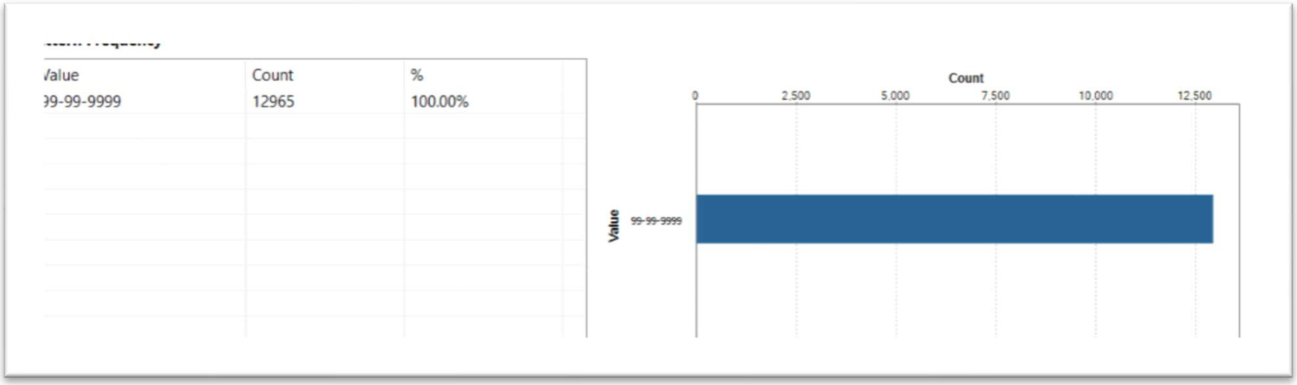
We now got to know the completeness and the uniqueness factor of the data set. The next step is to find the data type and pattern in the data and its frequency, for which we use the task Pattern Frequency Analysis.

We fond that the column variables ‘isCollaborator’, ‘isDeleted’, ‘isFailed’, ‘isFeatured’, ‘isHidden’, ‘isPrivate’, ‘isSuperFeatured’ have same pattern summary.



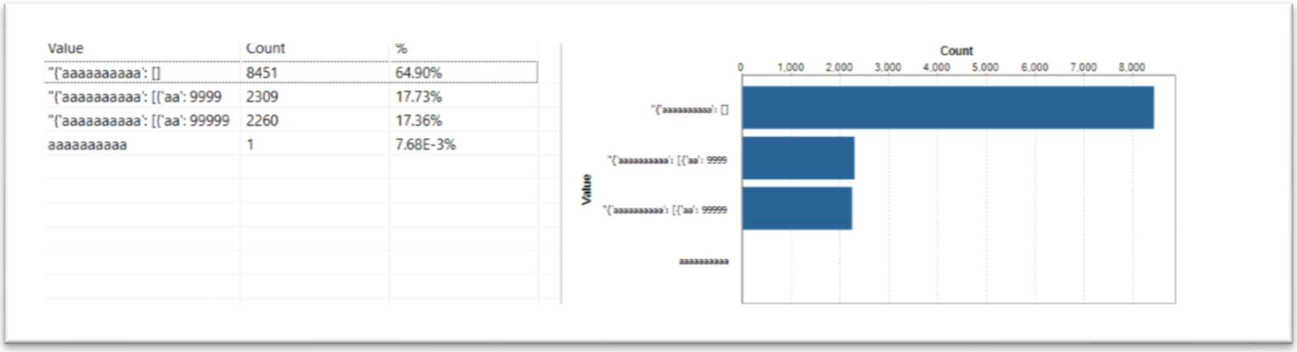
Value	Count	%
AAAAA	10774	82.74%
AAAA	2142	16.45%
A	54	0.41%
a	50	0.38%
aaAaaaaaaa	1	7.68E-3%

Whereas in column ‘dateUpdated’ it displays 12,965 counts have the same date type, but the total count of the column is 13,021 which means 56 row counts are of different date type.



Value	Count	%
99-99-9999	12965	100.00%

Since most of the column is of data type integer and string the patterns are like each other. The other three columns with json data type have the pattern frequency graph like each other, below is the graph that represents ‘category’ column variable.



Value	Count	%
"{'aaaaaaaaaa': []	8451	64.90%
"{'aaaaaaaaaa': [{'aa': 9999	2309	17.73%
"{'aaaaaaaaaa': [{'aa': 99999	2260	17.36%
aaaaaaaaaa	1	7.68E-3%

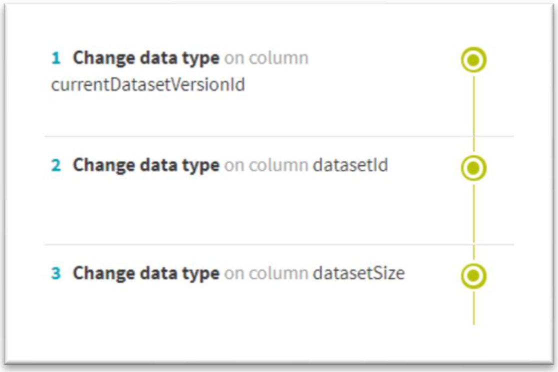
By Profiling the data, we come to know the overall type of data and its quality, now it needs to be cleaned and further we will proceed to data cleaning.

Data Cleaning:

We are making use of Talend Data Preparation for cleaning of the dataset. We are cleaning the data based on the quality dimension.

Start by uploading the data file into the software. The Talend automatically detects the data format and set the data type but due to irregular data pattern the software wrongly sets the data format for columns so first step we start by is checking each columns data format and assigning them the correct format.

Below is the screenshot of the recipe that mentions the column name where the operations were performed.



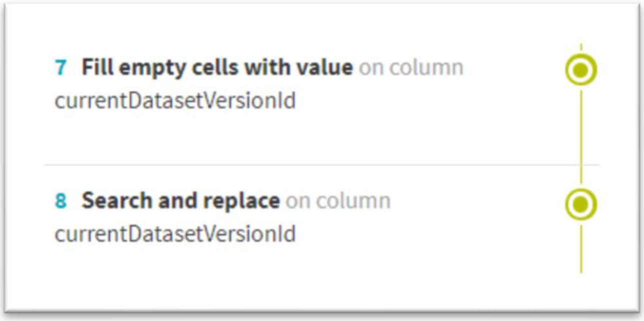
1 **Change data type** on column currentDatasetVersionId

2 **Change data type** on column datasetId

3 **Change data type** on column datasetSize

Talend indicates the invalid, and empty string values present in the column, hence we start by cleaning based on it. We find the column 'creatorUserId' has invalid values, so we check the pattern and come to know that we have non numeric and non-alpha numeric values present in the data, we remove them to fulfill the completeness dimension of the table.

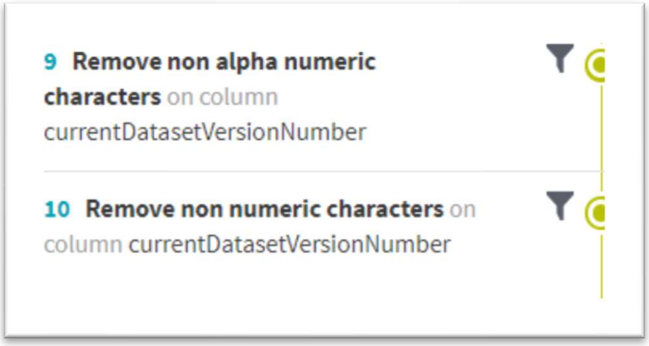
The next column is 'currentDatasetVersionId' it includes very less number of empty values and duplicates for that we are replacing the empty values with NA and removing the duplicates in the column so the data in the column satisfies Validity and Uniqueness dimension.



7 **Fill empty cells with value** on column currentDatasetVersionId

8 **Search and replace** on column currentDatasetVersionId

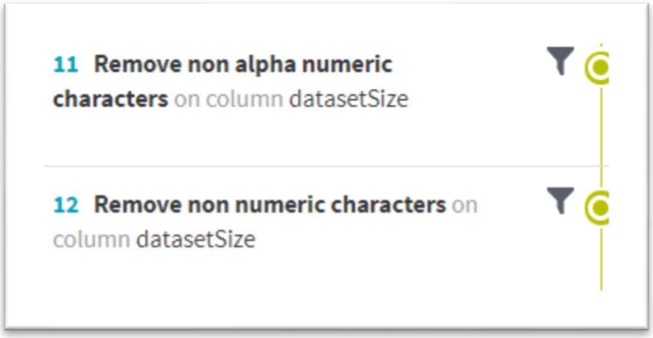
The column ‘currentDatasetVersionNumber’ has zero empty values and has 25 invalid values, we remove the invalid characters present in the data and now the data column fulfills Completeness dimension.



9 **Remove non alpha numeric characters** on column currentDatasetVersionNumber

10 **Remove non numeric characters** on column currentDatasetVersionNumber

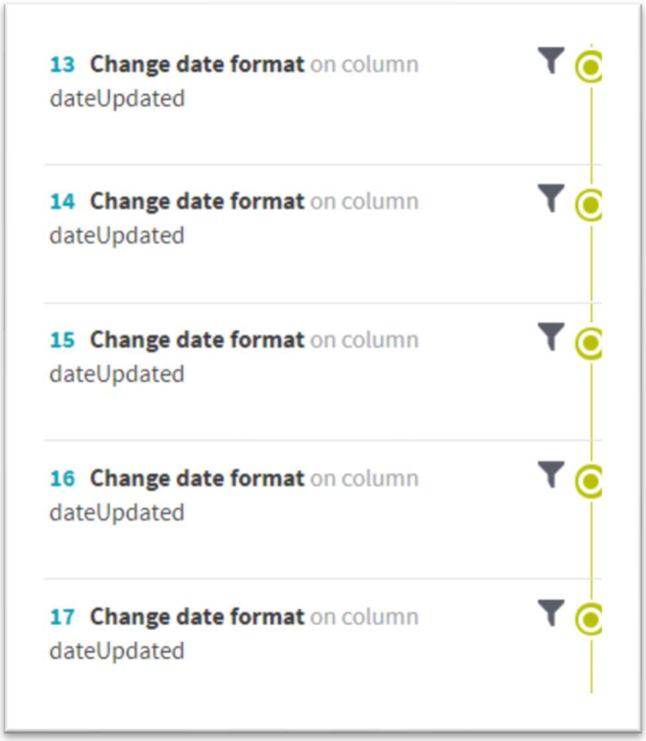
The data column ‘datasetSize’ has 37 invalid values presents, we remove the invalid charaters and now the complete data is valid, this column fulfills the accuracy, completeness, validity dimensions.



11 **Remove non alpha numeric characters** on column datasetSize

12 **Remove non numeric characters** on column datasetSize

The data column ‘dateUpdated’ can be useful data but there are 3335 data rows with invalid values, so to find out the invalid data we check the pattern analysis tab, below we can see the difference in the date format.



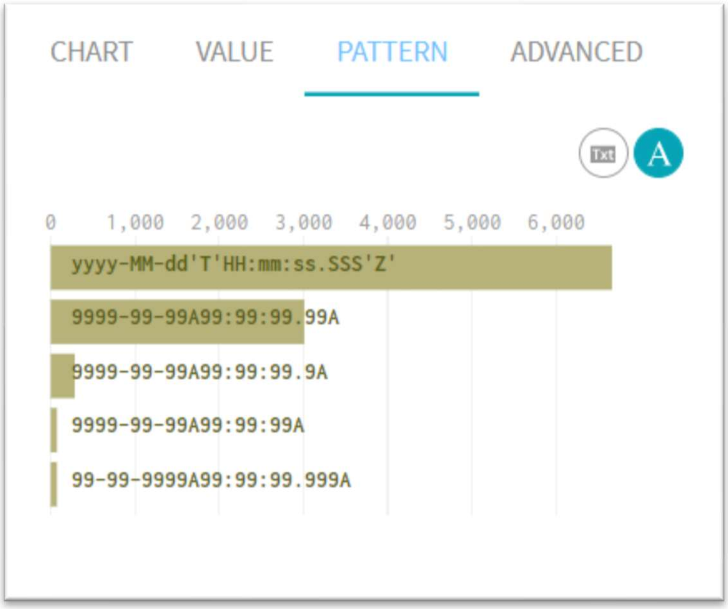
13 **Change date format** on column dateUpdated

14 **Change date format** on column dateUpdated

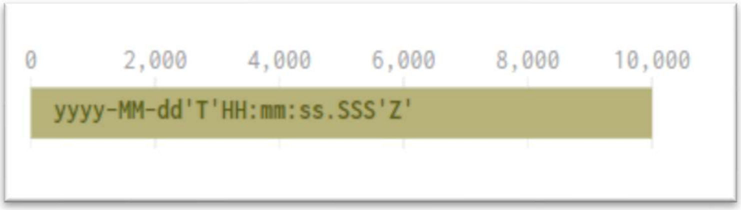
15 **Change date format** on column dateUpdated

16 **Change date format** on column dateUpdated

17 **Change date format** on column dateUpdated



So, we change the date format of invalid values by filtering out the selective pattern and make it consistent throughout. Below we can see now the date format is consistent throughout and there are no invalid or empty values present. The column now fulfills the validity, completeness and timeliness.



The next column is the ‘downloadCount’ , we found there are 55 invalid data containing non numerical values, hence removing them gets us the entire data as valid and complete. We fulfill the completeness, validity and accuracy dimensions.

The columns ‘isCollaborator’, ‘isDeleted’, ‘isFailed’, ‘isFeatured’, ‘isHidden’, ‘isPrivate’, ‘isSuperFeatured’ have the same data format and we have also seen the pattern frequency is similar while data profiling. Hence, we remove the invalid data format and convert them in to true values and now the data fulfills the accuracy, conformity, validity and completeness.

The column ‘diffType’ has two values mainly as ‘versioned’ and ‘unversioned’ but there were rows were the data was mentioned as ‘v’ so we suppose that v is for versioned and replace all the rows containing ‘v’ as ‘versioned’. This helps to fulfill the validity, accuracy and completeness dimension.

The column ‘viewCount’ also have some invalid value, by removing the non-numerical character the data is valid and complete thus it fulfills the completeness, accuracy, validity.

Now the data is cleaned and we can further proceed and try to get insights out of the analyzed data.

Insights gained after analysis of the dataset:

For gaining meaningful insight we analyzed the following columns variables -datasetSize, downloadCount, ownerName, title, viewCount.

Below is the list of insights that were obtained from the data.

1. We analyzed the ‘viewCount’ and can say that out of 13,021 datasets there are only 4 datasets with zero view. The ‘viewCount’ column and the ‘title’ both fulfill the dimension completeness, validity and conformity, so the analysis fit in these quality dimensions.

Title of dataset	View-count
cleanTrain	0
5 Day Data Challenge: Day 1	0
Yelp Reviews 2015	0
out Rdata	0

2. Top 10 dataset that are most viewed.

We analyzed the ‘viewCount’ and can say that out of 13,021 datasets these are the ones with top 10 most viewed datasets. The ‘viewCount’ column and the ‘title’ both fulfill the dimension completeness, validity and conformity, so the analysis fit in these quality dimensions.

Title of dataset	View-count
Credit Card Fraud Detection	997159
TMDB 5000 Movie Dataset	539476
European Soccer Database	527757
2018 Kaggle ML & DS Survey Challenge	274306
Banks data	258962
Global Terrorism Database	258504
Bitcoin Historical Data	246640
Iris Species	215955
Lending Club Loan Data	213503
Breast Cancer Wisconsin (Diagnostic) Data Set	209227
Wine Reviews	187870