

A Project Report on

Data Management

Under supervision of Prof. Dr. Ajinkya Prabhune,

Professor, SRH university

Prof. Dr. Barbara Sprick

Professor, SRH university



Submitted by

Medisetti Narendra, Matrikel Number: 11012094

M.Sc in Big data & Business Analytics, SRH University, Heidelberg

Data profiling

Data profiling is the process of investigating source data, understanding structure. It's given a clear idea flow of data like data Entity, data Attribute interrelationships and data about data. Data profiling tools give a better understanding on missing data, simple statistics, the real-world object your data and pattern.

Data profiling Tools:

- Talend open studio Data quality

Data cleaning Tools:

- Talend data preparation
- Open refine
- Trifacta

software we used in this project: Talend

Talend is an open source platform. It provides various services for data management, data quality, data cleaning.

Dataset:

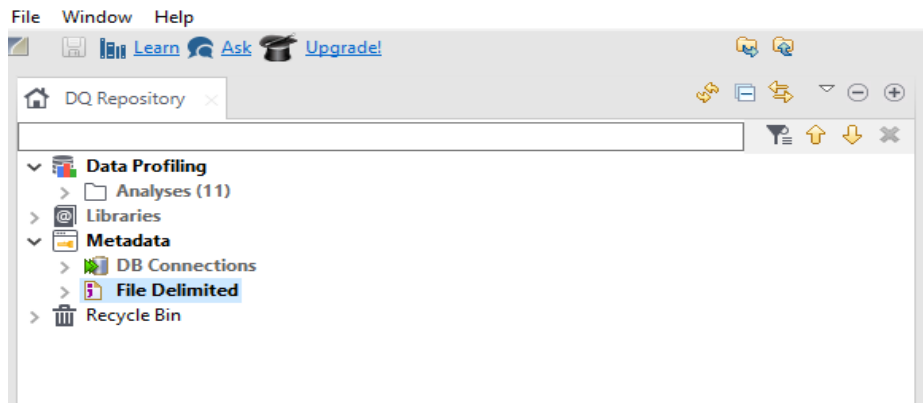
This dataset tells how many users downloads a dataset in Kaggle website and I do a bit of analysis on. The dataset contains contain 28 fields and 13020 rows. These data set contain many duplicate, unique data and many rows are repeated. So, to clean data we must know the data quality.

Profiling:

To do data profiling. I made used of a Talend open studio for data quality. After installing and running a Talend Data quality. Firstly, we need to import a dataset.

File Delimited

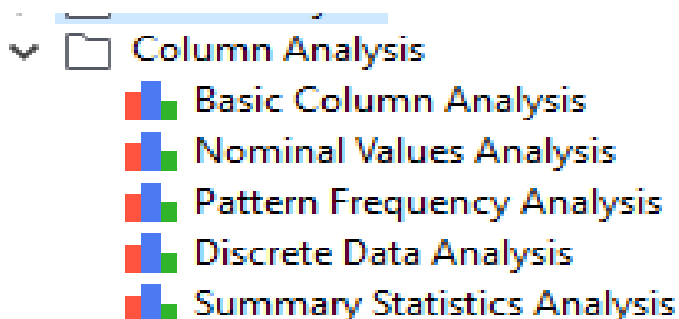
To import the csv file, select the File Delimited right click and create file delimited connection add the file name on next browse your csv file click on next, keep the field separator comm, setheading row as column name. After this is ready to have analysis on dataset. The complete datat is stored in Metadata.

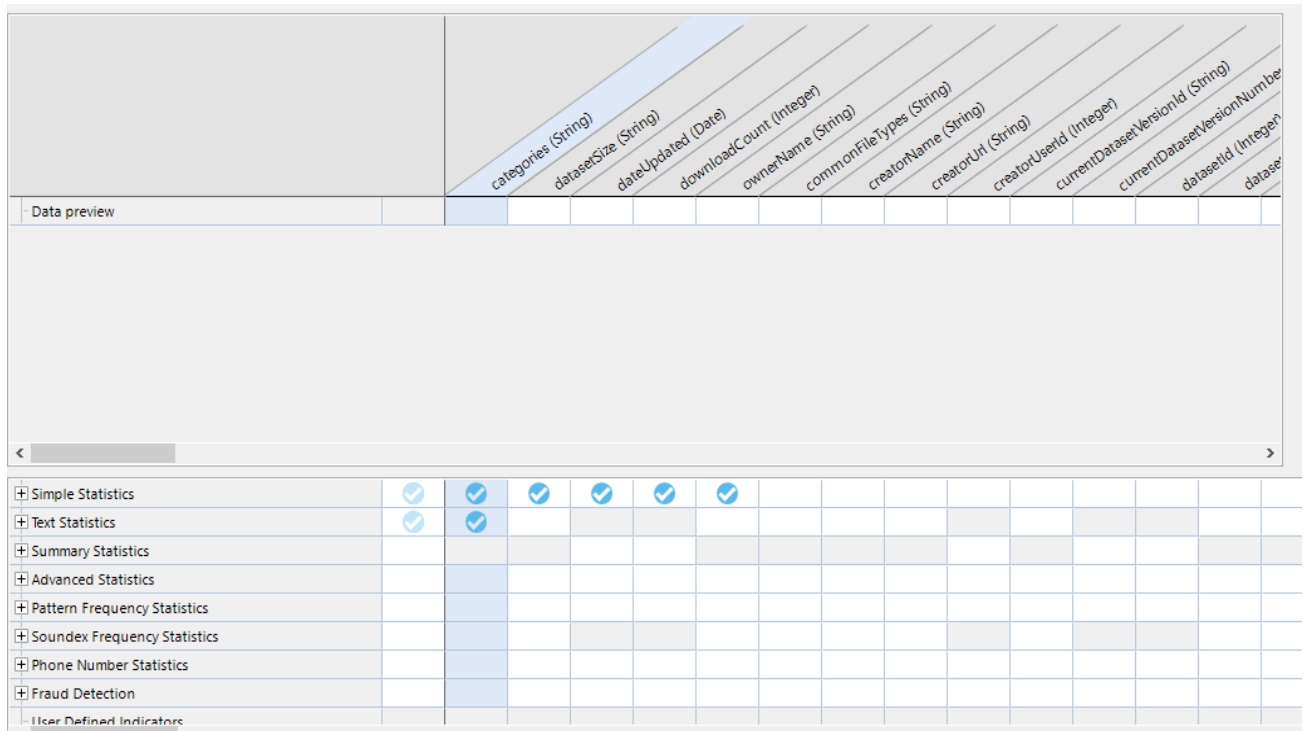


Column Analysis:

The analyses in this envelop empower you to profile the data on a column basis.

- All indicators are computed against on several columns, but each column is analyzed separately and independently. The Columns Analysis mainly focus on a column and analyze each cell of the column.
- This gives basic information about statistic and advance statistic, patterns and frequency, example number of the distinct values in the specific column, Number of the null values in the column which refers to the completeness.





Above figure show the option of selecting indicator according to the column

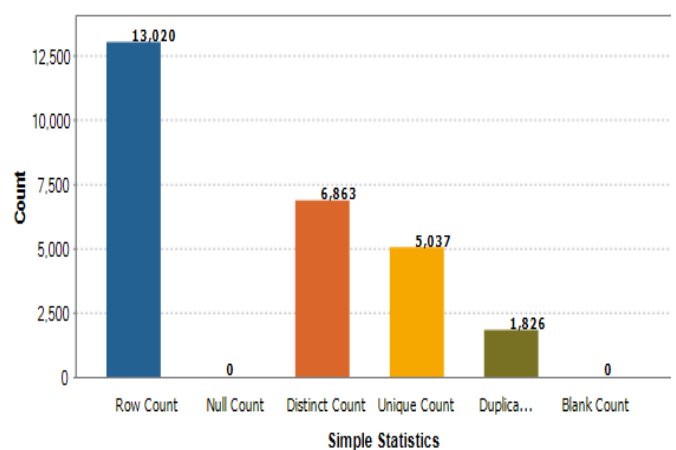
Basic statistics

Creator Name:

Column: ALLKAGGLE.creatorName

Simple Statistics

Label	Count	%
Row Count	13020	100.00%
Null Count	0	0.00%
Distinct Count	6863	52.71%
Unique Count	5037	38.69%
Duplicate Count	1826	14.02%
Blank Count	0	0.00%



Above graph depicts simple statistics about creator name column of the dataset. Row count, Null count, Distinct count, Unique count and Duplicate count. There is no null count it reflect the completeness

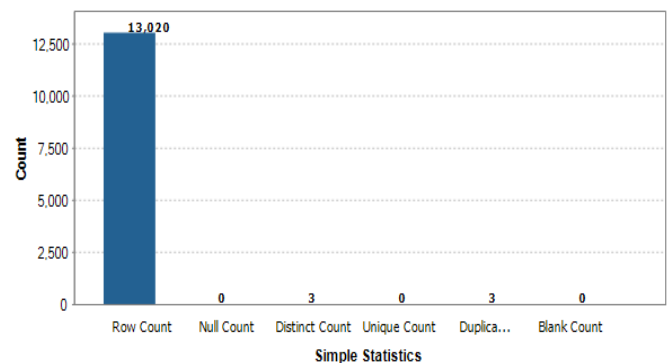
but there are duplicate values means uniqueness dimension is the affected in this column

Is.Private:

▼ Column: ALLKAGGLE.isPrivate

▼ Simple Statistics

Label	Count	%
Row Count	13020	100.00%
Null Count	0	0.00%
Distinct Count	3	0.02%
Unique Count	0	0.00%
Duplicate Count	3	0.02%
Blank Count	0	0.00%



Above statistics reflect there is indication of completeness but there is factor of consistency.

Discrete Data Analysis:

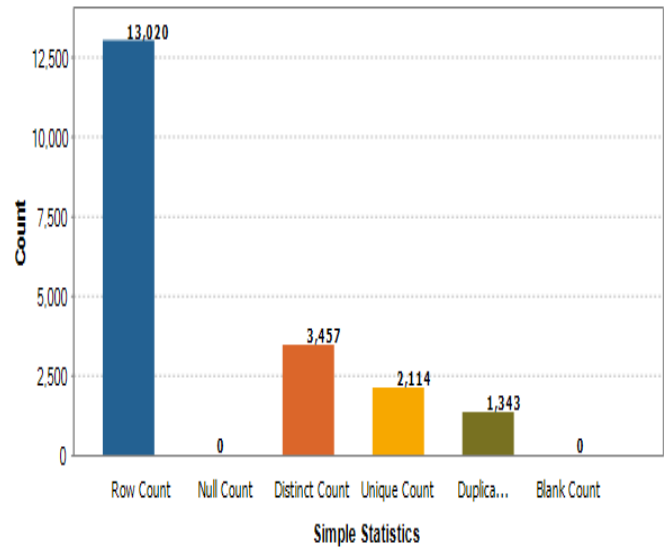
This analysis enables you to analyze numerical data. It creates a column analysis in which indicators, appropriate for numeric data.

Creation of discrete analysis automatically assigns the frequency indicators for numerical columns.

View Count:



▼ Simple Statistics

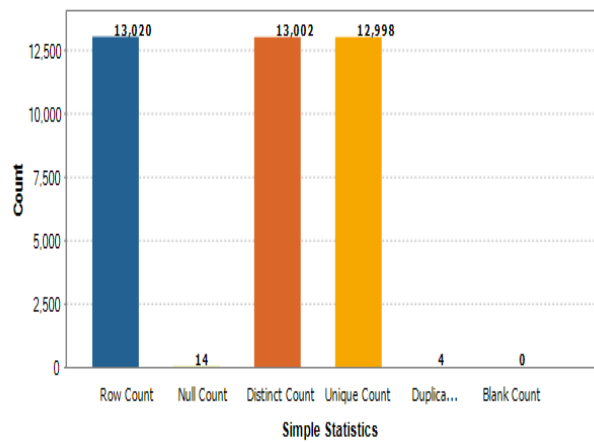
[illegible]

Above statistics show that this column also reflects the completeness but not uniqueness dimension as it has repeated values.

Current Dataset Version ID:



▼ Simple Statistics

[illegible]

Above statistics show that this column also reflects the completeness and unique as it has more row count, distinct count and unique count.

Pattern Analysis:

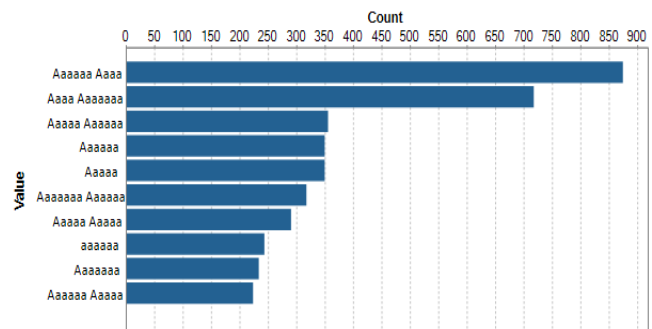
In Pattern analysis discover pattern in the data and frequency

Creator Name:

▼ Column: metadata.creatorName

▼ Pattern Frequency

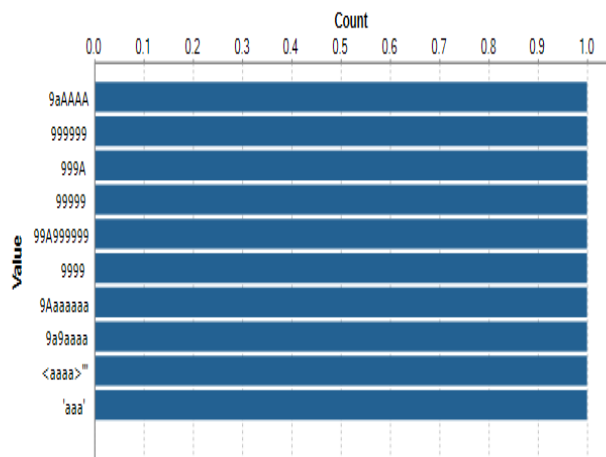
Value	Count	%
Aaaaa Aaaa	875	N/A
Aaaa Aaaaaa	718	N/A
Aaaaa Aaaaaa	356	N/A
Aaaaaa	350	N/A
Aaaaa	350	N/A
Aaaaaa Aaaaaa	318	N/A
Aaaaa Aaaaa	291	N/A
aaaaaa	244	N/A
Aaaaaa	234	N/A
Aaaaaa Aaaaa	224	N/A



In the above graph describe the which string size in a column has repeated or count. This tells the data and frequency .

▼ Pattern Low Frequency

Value	Count	%
9aAAAA	1	N/A
999999	1	N/A
999A	1	N/A
99999	1	N/A
99A999999	1	N/A
9999	1	N/A
9Aaaaaaa	1	N/A
9a9aaaa	1	N/A
<aaaa>'''	1	N/A
'aaa'	1	N/A



In the above graph describe the which string size in a column has repeated or count less. This tells the data and low frequency .

Text Analysis:

This analyze the characteristics of the textual fields including maximum, minimum and average length .Text analysis is the part of column analysis which gives the information about the column which has

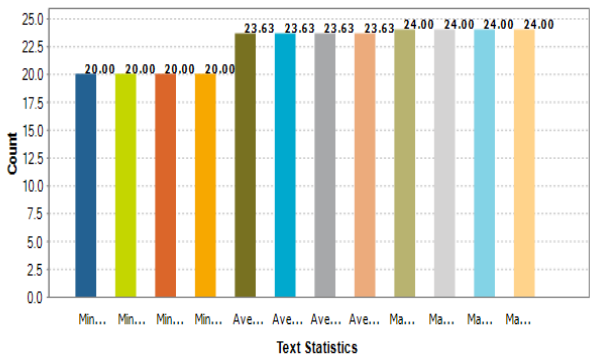
got the strings. It shows the pattern of the text distributed among the column. It is basically textual representation of the column.

Date Updated:

▼ Column: ALLKAGGLE.dateUpdated

▼ Text Statistics

Label	Value
Minimal Length With Blank and N...	20.00
Minimal Length With Blank	20.00
Minimal Length With Null	20.00
Minimal Length	20.00
Average Length With Blank and N...	23.63
Average Length With Blank	23.63
Average Length With Null	23.63
Average Length	23.63
Maximal Length With Blank and N...	24.00
Maximal Length With Blank	24.00
Maximal Length With Null	24.00
Maximal Length	24.00



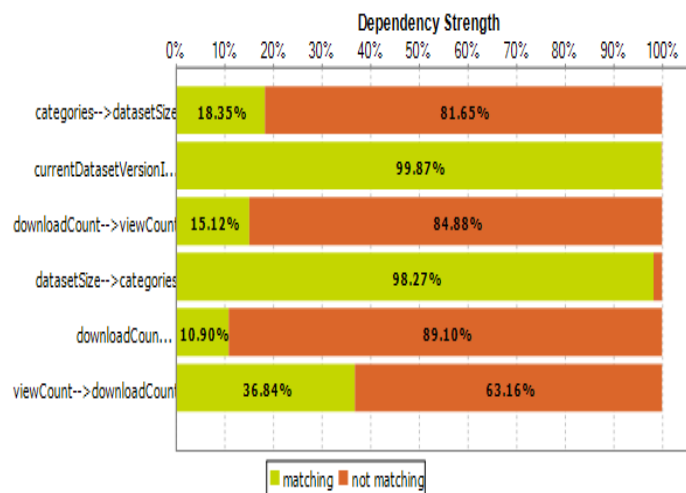
Above graph shows the consistency in the text , there is no mismatch with different text.

Table Analysis:

Function dependency:

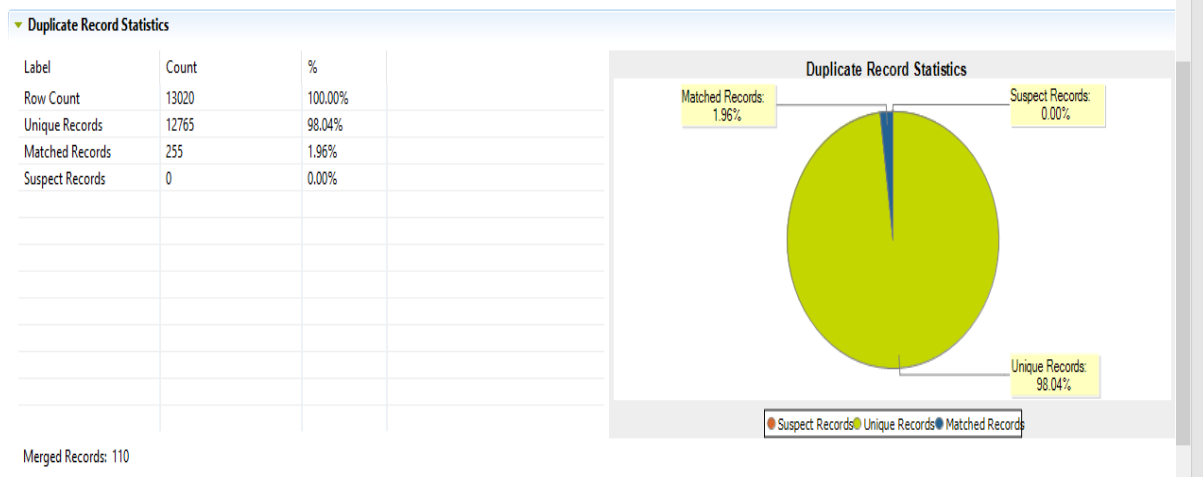
This analysis help to find the column dependency. It determines the that which extent the value of column A depend on the value of column B . This relation is denoted as A-> B .

Dependency	#Match	%Match	#row
categories-->datasetSize	2179	18.35%	11876
currentDatasetVersionId-->downl...	13001	99.87%	13018
downloadCount-->viewCount	1419	15.12%	9383
datasetSize-->categories	11670	98.27%	11876
downloadCount-->currentDataset...	1419	10.90%	13018
viewCount-->downloadCount	3457	36.84%	9383



Above graph give the idea of how the column categories depend on the dataset size . Value of categories less match with value of dataset size but dataset size value match categories . It means value of party has more dataset size on the value of categories. Value of download count less column match with value of view count and dataset size higher than the categories. The download count columns also less than currentdataset size but view count less match to the download count columns.

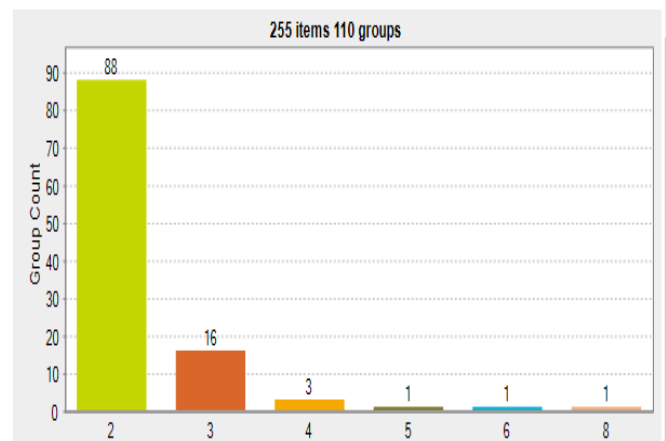
Math Analysis:



Above diagram represent the duplicate record statistics tells that how less are duplicated. By this we can say that this columns has uniqueness, completeness and consistency dimension.

▼ Group Statistics

Group Size	Group Count	Record Count	% Records
1	12765	12765	98.04%
2	88	176	1.35%
3	16	48	0.37%
4	3	12	0.09%
5	1	5	0.04%
6	1	6	0.05%
8	1	8	0.06%



Above figure tell that group count and Record Count and percentage of record.

Structural Analysis:

Structural Analysis determine the schema rule in meta data. how columns relate to each other to form tables and how tables relate to each other and to form a market analysis.

▼ Statistical Information

Schema	#rows	#tables	#rows/table	#views	#rows/view	#keys	#indexes
	13020	1	13020.00	0	NaN	0	0

Table	#rows	#keys	#indexes
ALLKAGGLE	13020	0	0

View	#rows

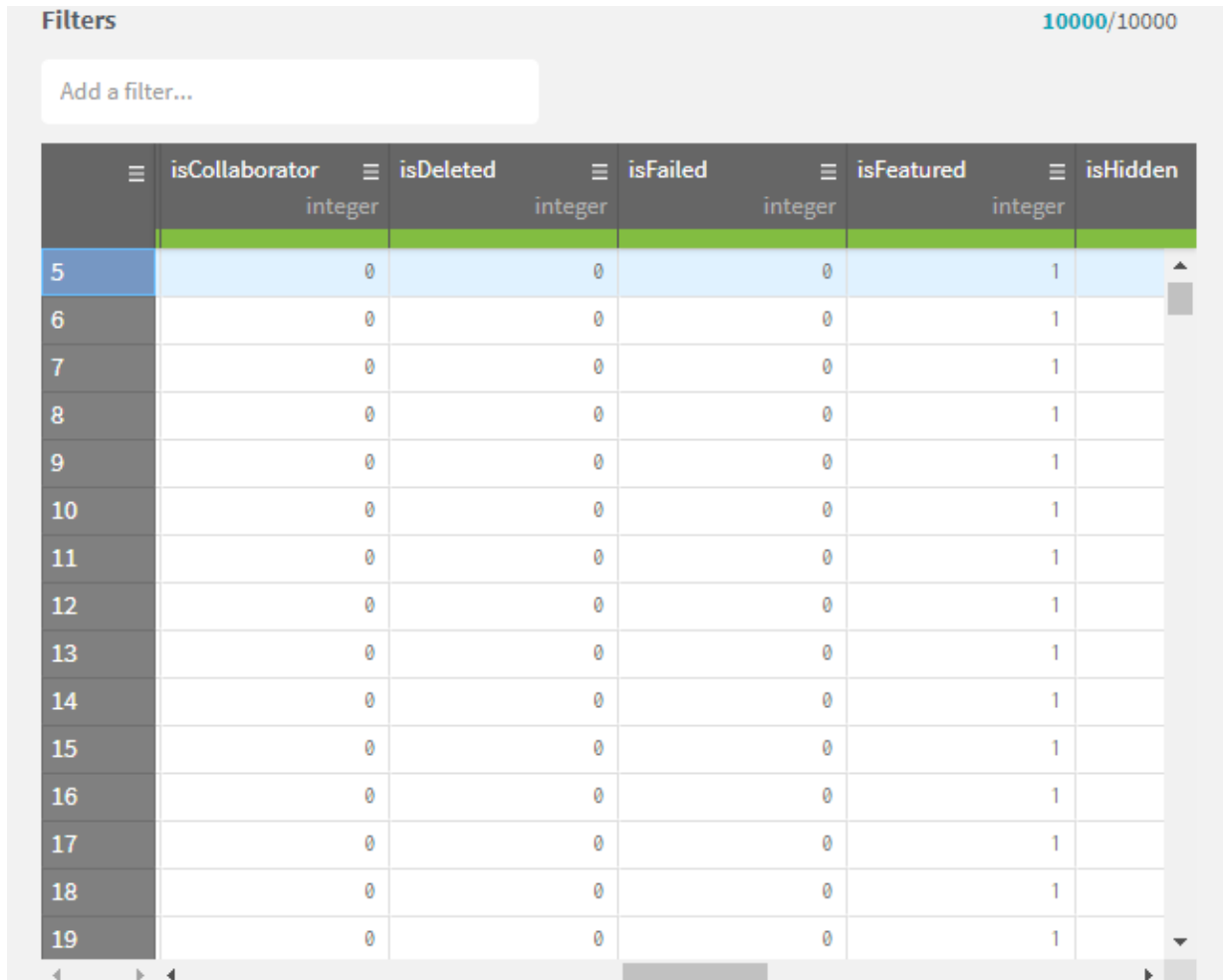
Above figure tell that data table stored in meta data and gives views of the rows.

Data cleaning:

Data cleaning suggest to the way toward distinguishing and expelling invalid data focuses from a dataset. This includes looking at the data for extraordinary incorrect information focuses that may inclination the consequences of your exploration. To guarantee that no data cooking happened, information cleaning methods must be done before the factual examination of study results begins.

Talend data Preparation:

Firstly we should import data in talend data preparation tool and we should begin the preparation



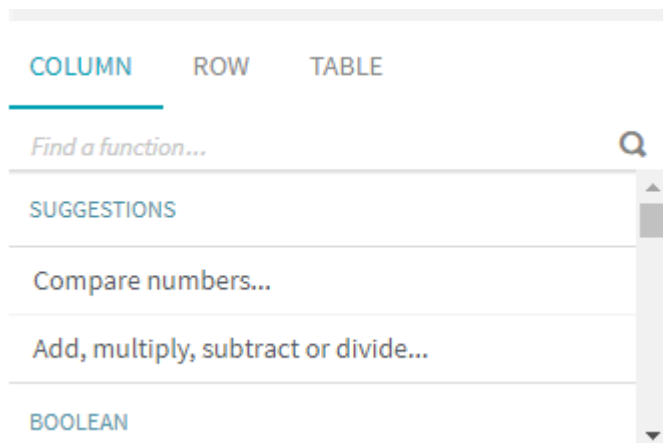
The screenshot displays the Talend data preparation tool interface. At the top, there is a 'Filters' section with a search bar labeled 'Add a filter...'. Below this, a table is shown with the following columns: **isCollaborator** (integer), **isDeleted** (integer), **isFailed** (integer), **isFeatured** (integer), and **isHidden**. The table contains 15 rows of data, with the first row (row 5) highlighted in blue. The values for the first four columns are 0, and the value for the fifth column is 1. The interface also includes a '10000/10000' indicator in the top right corner and a scroll bar on the right side of the table.

	isCollaborator integer	isDeleted integer	isFailed integer	isFeatured integer	isHidden
5	0	0	0	1	
6	0	0	0	1	
7	0	0	0	1	
8	0	0	0	1	
9	0	0	0	1	
10	0	0	0	1	
11	0	0	0	1	
12	0	0	0	1	
13	0	0	0	1	
14	0	0	0	1	
15	0	0	0	1	
16	0	0	0	1	
17	0	0	0	1	
18	0	0	0	1	
19	0	0	0	1	

Above figure view the talend data preparation tool.

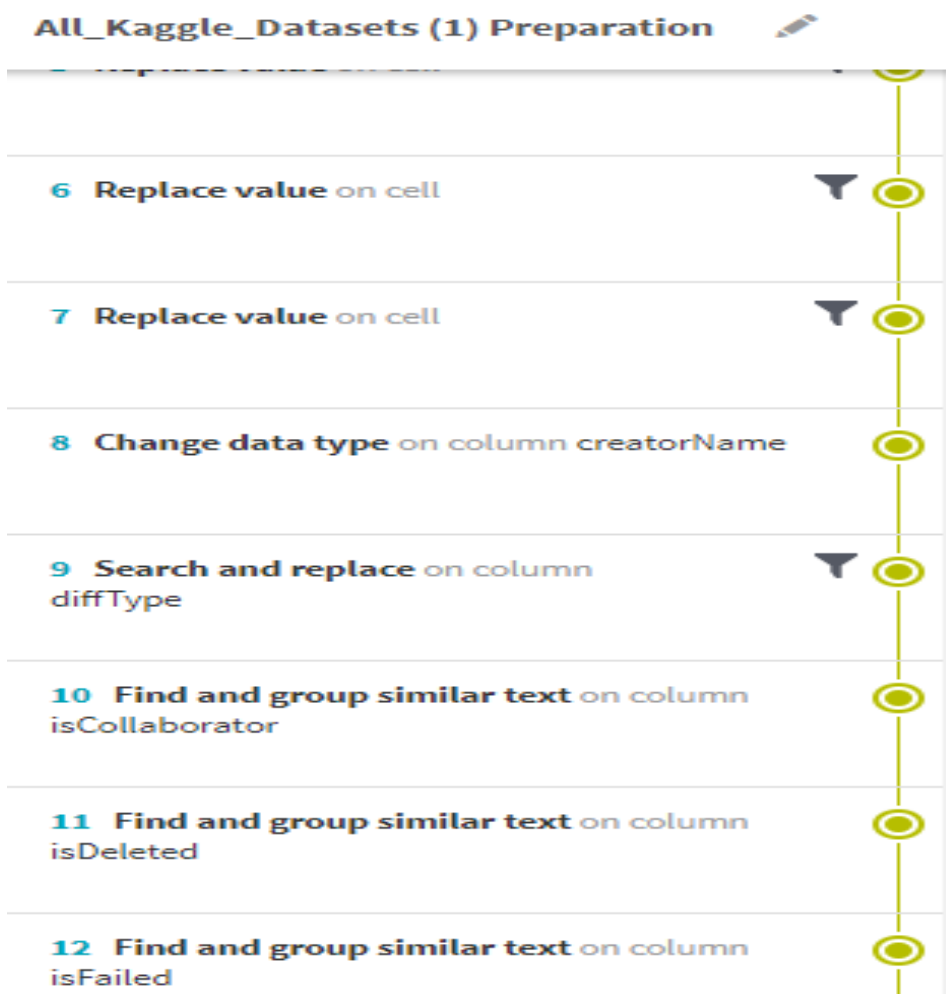
Search Function:

The find function is used to search command to do cleaning operation in talend tool.



Recipie:

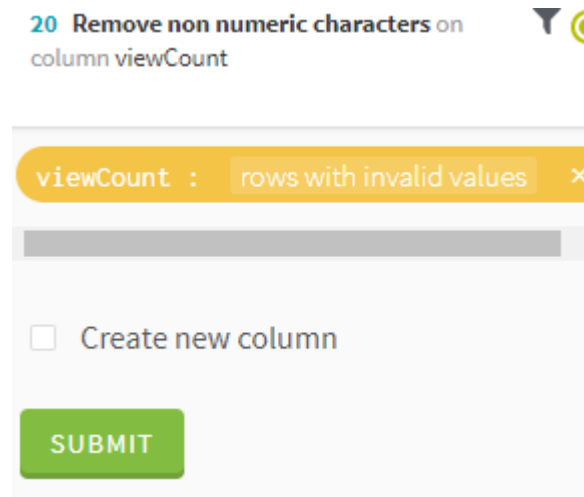
what function we used for data cleaning and function steps are show in recipie.



Dimension:

Accuracy:

In order to get accuracy we need to find which has any junk values. So remove this we used function called removing non numeric character. These operation are done in these columns



Completeness:

In order to get completeness columns should not have null, empty values. These columns comes under datasetId.

Conformity:

In order to get conformity we change the type of columns. So that the column has same datatype.

Validity:

In order to get the validity columns should be same type. In Kaggle dataset these columns are is collaborate, is delete, is feature, is hidden, is private, diff type, is failed columns are active validity dimension. By using a function called search and replaced.

9 Search and replace on column diffType



diffType =

☐ Create new column

Search for:

Replace with:

☐ Overwrite entire cell

Consistency:

In order to get Consistency, the columns should be uniqueness, semantic consistency, format should be same. In my dataset creator name, diff Type, overview and owner Name columns. In these columns we used function call change to title case and removing the white space in columns using function called remove trailing and leading characters. we have remove some junks in these columns function called remove non numeric character.

34 Change to title case on column creatorName

☐ Create new column

Uniqueness:

In order to get uniqueness we have some columns are unique, the columns are dataset id, user ID.

Bussiness question:

1. by seeing the data qualities insights we you gain from dataset.
2. completeness, accuracy, consistence, validity data quality dimension does fits my analysis
3. By using the user and owner url we can find there details. We can known highest dataset vesion comparing view count.