DATA MANAGEMENT PROJECT

BHARATH KUMAR NANDA KUMAR

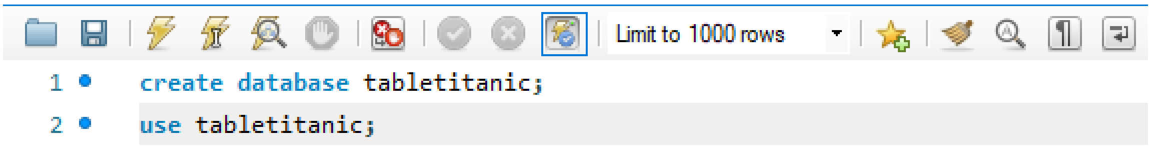
STUDENT ID:11014144

MYSQL WORKBENCH

Usage : Storing and Fetching Data

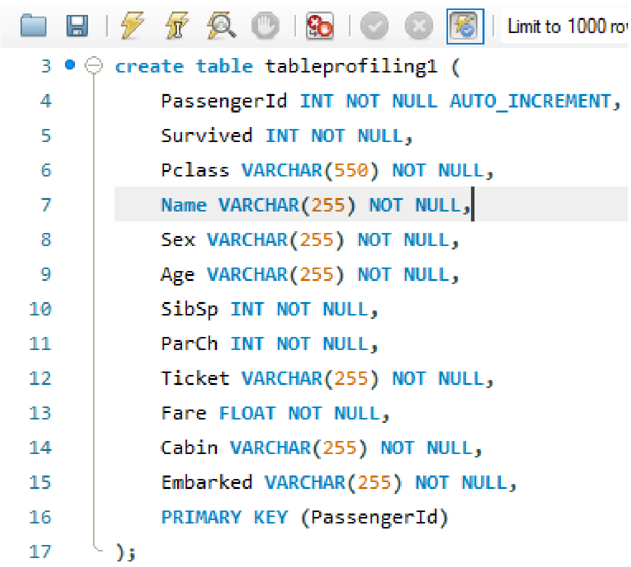
Stepwise Process Used For Storing Data In MYSQL Database:

Step 1: Creating the database tabletitanic and Activate the database tabletitanic



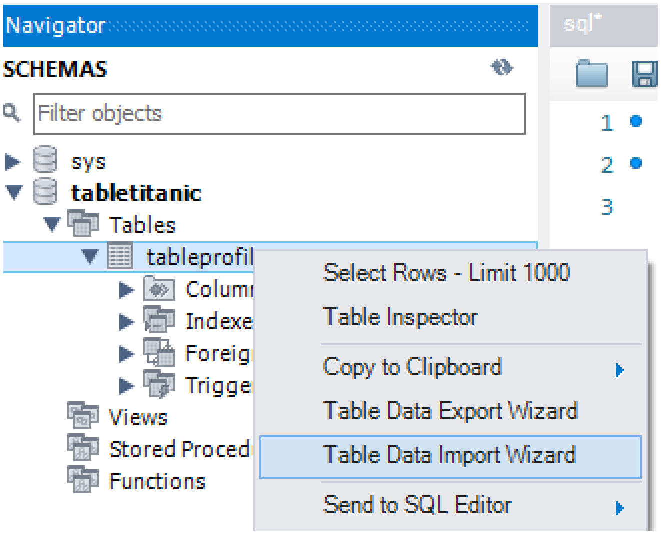
```
create database tabletitanic;
use tabletitanic;
```

Step 2: Creating a table named tableprofiling1 containing the data variables and each variable is defined w.r.t to data type



```
create table tableprofiling1 (
    PassengerId INT NOT NULL AUTO_INCREMENT,
    Survived INT NOT NULL,
    Pclass VARCHAR(550) NOT NULL,
    Name VARCHAR(255) NOT NULL,
    Sex VARCHAR(255) NOT NULL,
    Age VARCHAR(255) NOT NULL,
    SibSp INT NOT NULL,
    ParCh INT NOT NULL,
    Ticket VARCHAR(255) NOT NULL,
    Fare FLOAT NOT NULL,
    Cabin VARCHAR(255) NOT NULL,
    Embarked VARCHAR(255) NOT NULL,
    PRIMARY KEY (PassengerId)
);
```

Step 3: Importing titanic csv data using the existing table we created named tableprofiling1



Data Cleaning

Data cleaning is done for the titanic dataset based on different data quality dimensions using the various data cleaning tools

Tools Used:

- Open refine
- Tableau Prep
- Talend data preparation
- Excel

Stepwise Process of cleaning Titanic CSV dataset:

Totally there are 12 columns in the titanic dataset which are

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	ParCh	Ticket	Fare	Cabin	Embarked

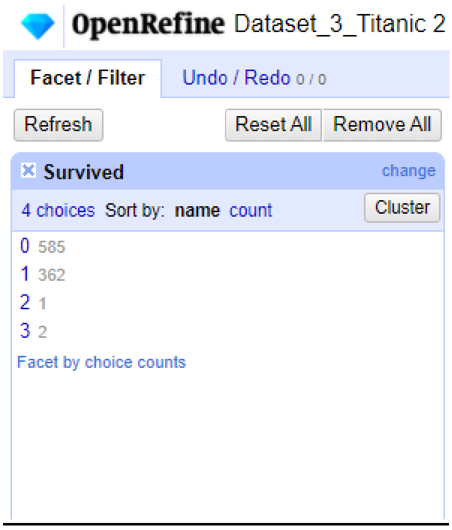
1) PassengerId cleaning based on Data quality dimensions:

From the data profiling done using talend tool, we know that PassengerId column follows: Completeness, conformity, accuracy, consistency, validity by default so no data cleaning is required on this column.

2) Survived cleaning based on Data quality dimension:

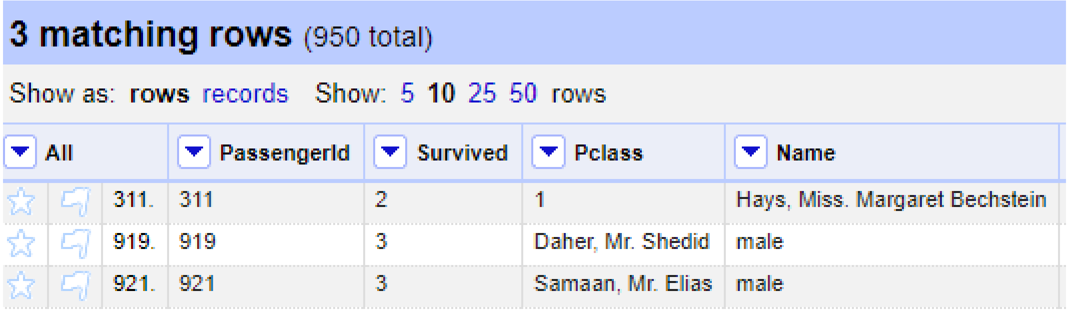
STEP 1: Use text facet for checking the data quality based on its dimensions and we found the integers '2' and '3'. So we came to know that survived column is inconsistent and invalid and those two integers are included for making it consistent and valid.

Output:



Survived	
0	585
1	362
2	1
3	2

STEP 2:
Survived results after including those cells.



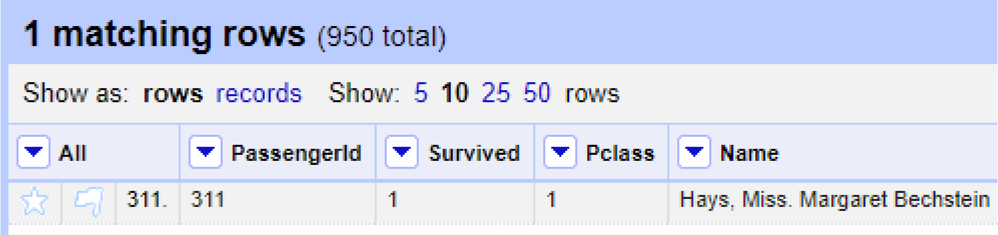
3 matching rows (950 total)

Show as: **rows** records Show: 5 **10** 25 50 rows

All	PassengerId	Survived	Pclass	Name
311.	311	2	1	Hays, Miss. Margaret Bechstein
919.	919	3	Daher, Mr. Shedid	male
921.	921	3	Samaan, Mr. Elias	male

From the external source ‘wikipedia’-survival of these three passengers is identified through their names given on the raw dataset.

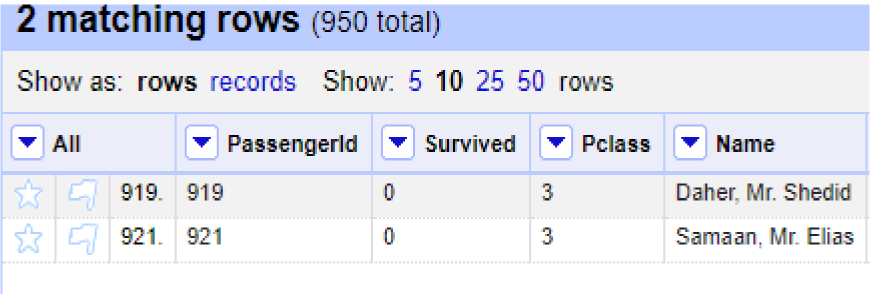
After identified:



1 matching rows (950 total)

Show as: **rows** records Show: 5 **10** 25 50 rows

All	PassengerId	Survived	Pclass	Name
311.	311	1	1	Hays, Miss. Margaret Bechstein



2 matching rows (950 total)

Show as: **rows** records Show: 5 **10** 25 50 rows

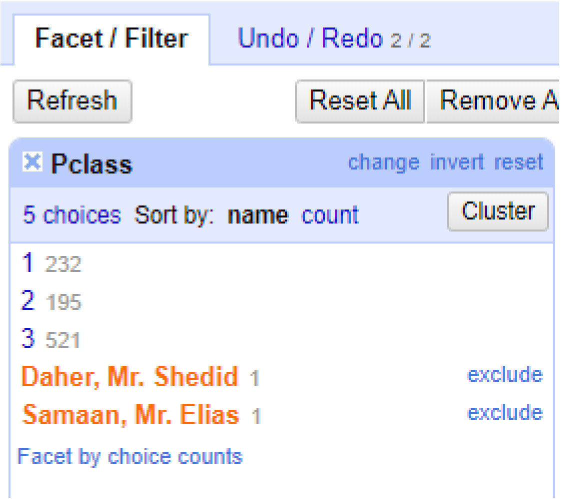
All	PassengerId	Survived	Pclass	Name
919.	919	0	3	Daher, Mr. Shedid
921.	921	0	3	Samaan, Mr. Elias

FINAL OUTCOME:

- So the survived column has been corrected and made consistent, valid.
- There are no blanks so survived column is complete by default and it follows a standard data type ‘INTEGER’ so it has conformity by default.
- Data objects accurately represents the real world values and there are no spelling mistakes or special characters so it accurate by default.

2) ‘Pclass’ cleaning based on Data quality dimension:

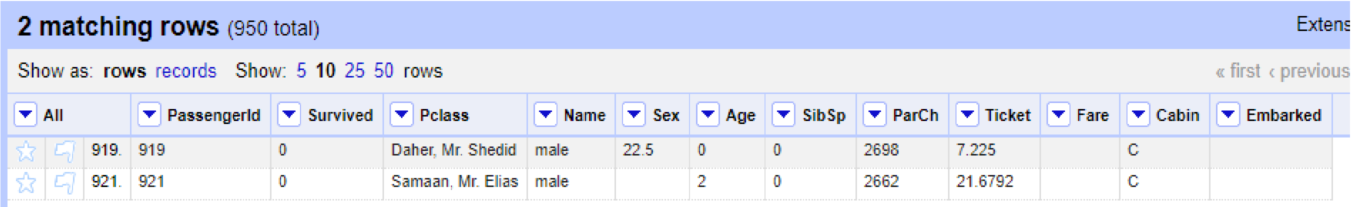
STEP1: Used text facet and found invalid two names so this column has (Invalidity, No conformity)



1	232
2	195
3	521
Daher, Mr. Shedid	1
Samaan, Mr. Elias	1

STEP 2

After including those cells we see:

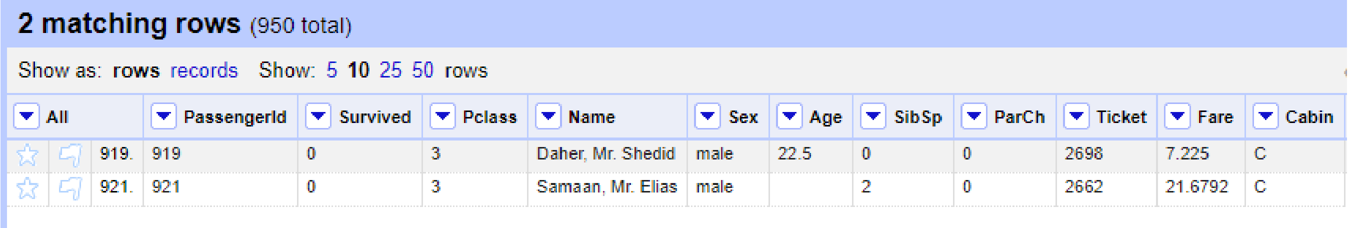


2 matching rows (950 total)

All		PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	ParCh	Ticket	Fare	Cabin	Embarked
☆	919.	919	0	Daher, Mr. Shedid	male	22.5	0	0	2698	7.225		C	
☆	921.	921	0	Samaan, Mr. Elias	male		2	0	2662	21.6792		C	

We could see the datas inside each cells on these two rows has been mistakenly typed on different cells .this can be corrected by swapping the datas into correct columns.

After Swapping:



2 matching rows (950 total)

All		PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	ParCh	Ticket	Fare	Cabin
☆	919.	919	0	3	Daher, Mr. Shedid	male	22.5	0	0	2698	7.225	C
☆	921.	921	0	3	Samaan, Mr. Elias	male		2	0	2662	21.6792	C

So now the Pclass has been corrected. So to conclude

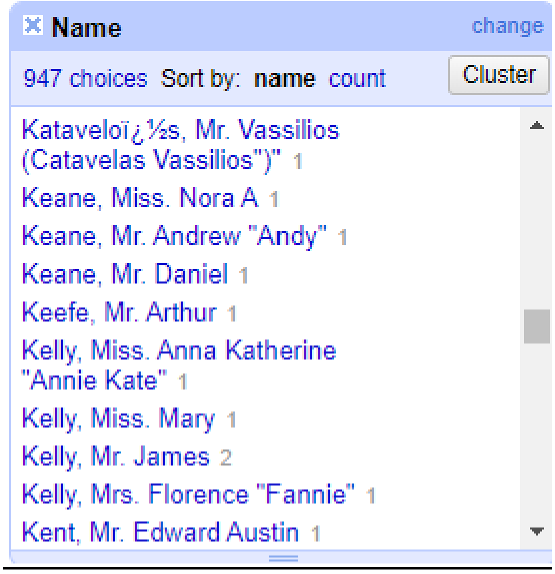
FINAL OUTCOME:

- It has the standard data type after swapping so it achieves conformity.
- It has Valid values after swapping so it achieves validity.
- It has no blank values by default so it has completeness, its consistent and it has no typos so its accurate as well.

3) ‘Name’ cleaning based on Data quality dimension:

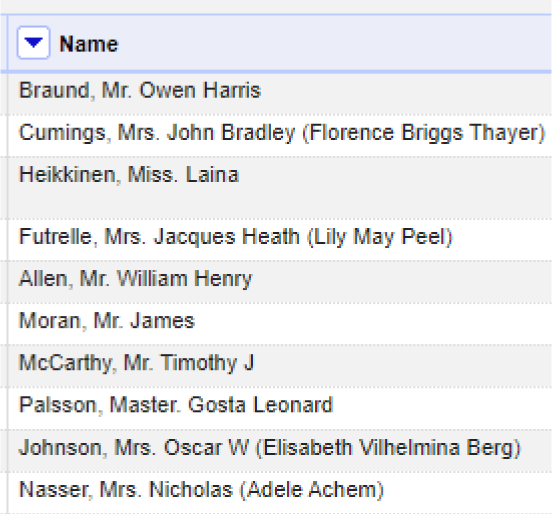
STEP1

Using text facet, we found that there are many Junk values and Typos in value so its inaccurate, few names contains nick names on parenthesis which makes the other names inconsistent from that, order of the name is not correct.



Kataveloï¿½s, Mr. Vassilios (Catavelas Vassilios")" 1
Keane, Miss. Nora A 1
Keane, Mr. Andrew "Andy" 1
Keane, Mr. Daniel 1
Keefe, Mr. Arthur 1
Kelly, Miss. Anna Katherine "Annie Kate" 1
Kelly, Miss. Mary 1
Kelly, Mr. James 2
Kelly, Mrs. Florence "Fannie" 1
Kent, Mr. Edward Austin 1

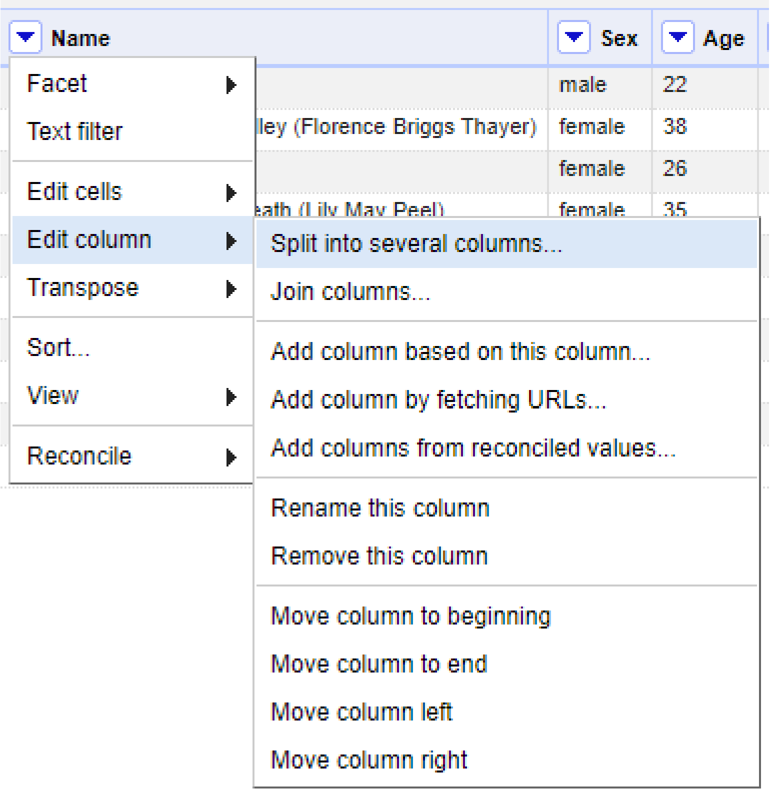
Sample Name Column Data

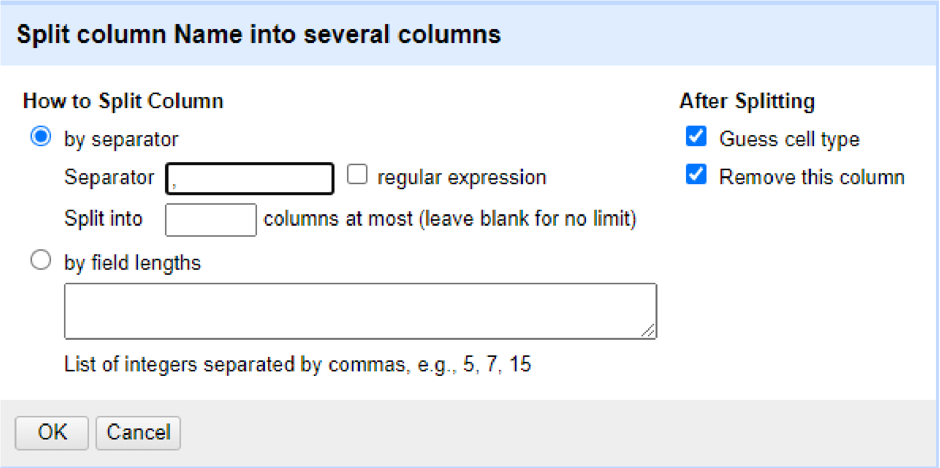


Name
Braund, Mr. Owen Harris
Cumings, Mrs. John Bradley (Florence Briggs Thayer)
Heikkinen, Miss. Laina
Futrelle, Mrs. Jacques Heath (Lily May Peel)
Allen, Mr. William Henry
Moran, Mr. James
McCarthy, Mr. Timothy J
Palsson, Master. Gosta Leonard
Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)
Nasser, Mrs. Nicholas (Adele Achem)

STEP 3:

Extract the last name in a separate column by splitting that using a comma seperator





Split column Name into several columns

How to Split Column

● by separator

Separator [,] □ regular expression

Split into [ ] columns at most (leave blank for no limit)

○ by field lengths

List of integers separated by commas, e.g., 5, 7, 15

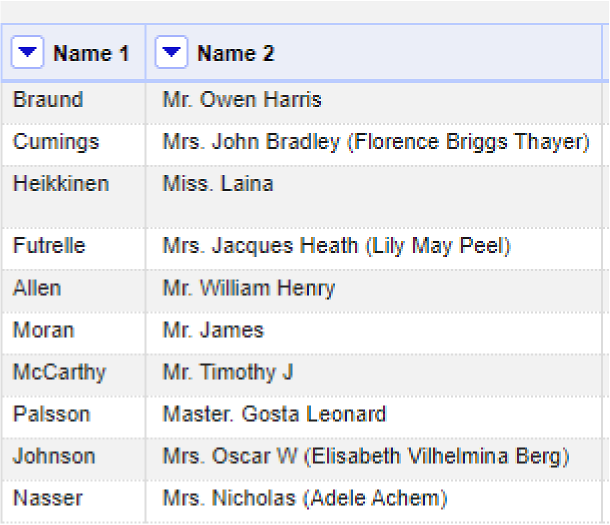
After Splitting

☑ Guess cell type

☑ Remove this column

OK Cancel

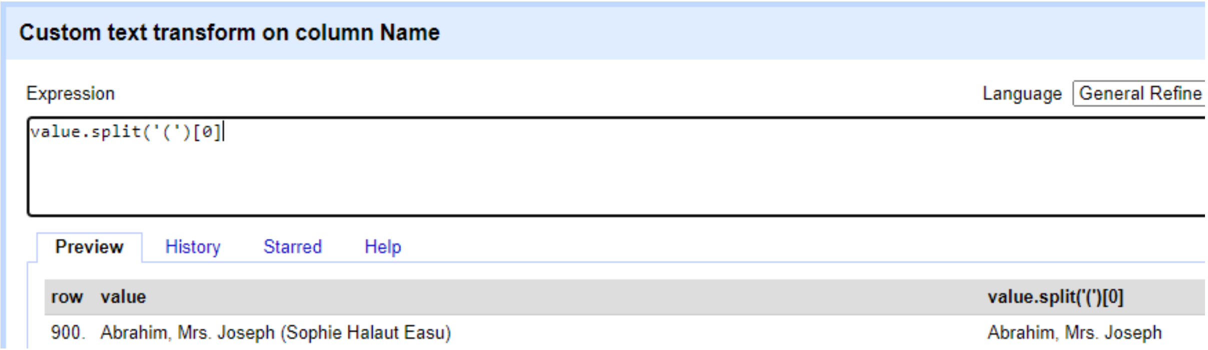
FINAL OUTCOME:



Name 1	Name 2
Braund	Mr. Owen Harris
Cumings	Mrs. John Bradley (Florence Briggs Thayer)
Heikkinen	Miss. Laina
Futrelle	Mrs. Jacques Heath (Lily May Peel)
Allen	Mr. William Henry
Moran	Mr. James
McCarthy	Mr. Timothy J
Palsson	Master. Gosta Leonard
Johnson	Mrs. Oscar W (Elisabeth Vilhelmina Berg)
Nasser	Mrs. Nicholas (Adele Achem)

STEP 4:

Removing the datas inside parenthesis so that consistency can be achieved and that can be done using GREL function:



Custom text transform on column Name

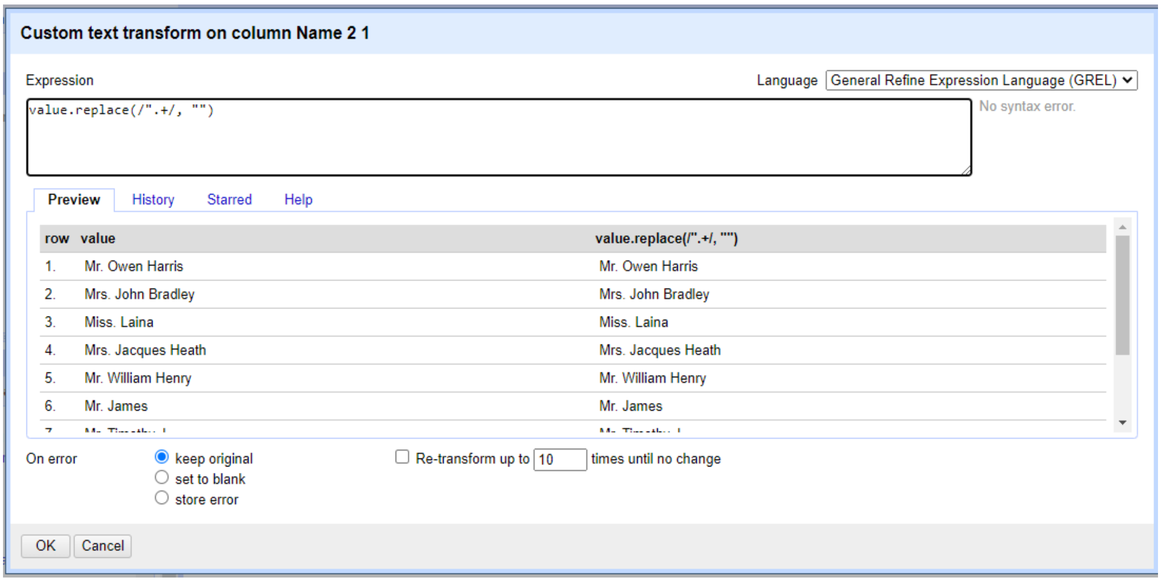
Expression

Language: General Refine

```
value.split('(')[0]
```

row	value	value.split('(')[0]
900.	Abrahim, Mrs. Joseph (Sophie Halaut Easu)	Abrahim, Mrs. Joseph

STEP 5: Removing the datas inside double quotes because it consist of invalid and inaccurate values and those validity and accuracy can be achieved by using GREL function:



Custom text transform on column Name 2 1

Expression

Language: General Refine Expression Language (GREL)

```
value.replace(/".+/, "")
```

No syntax error.

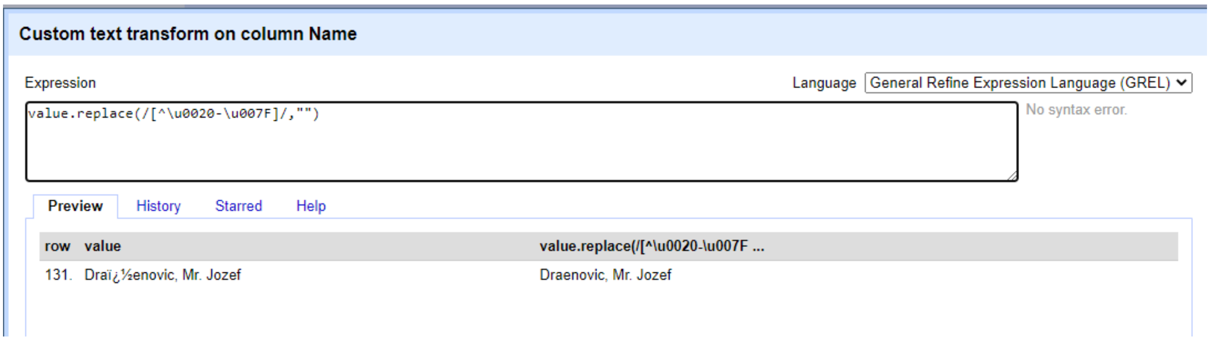
row	value	value.replace(/".+/, "")
1.	Mr. Owen Harris	Mr. Owen Harris
2.	Mrs. John Bradley	Mrs. John Bradley
3.	Miss. Laina	Miss. Laina
4.	Mrs. Jacques Heath	Mrs. Jacques Heath
5.	Mr. William Henry	Mr. William Henry
6.	Mr. James	Mr. James
7.	Mr. Timothy J	Mr. Timothy J

On error: ● keep original ○ set to blank ○ store error

□ Re-transform up to 10 times until no change

STEP 6:

Name column consist of junk values/special characters which comes under inaccuracy data quality dimensions which can be corrected by using GREL Function:



Custom text transform on column Name

Expression

Language: General Refine Expression Language (GREL)

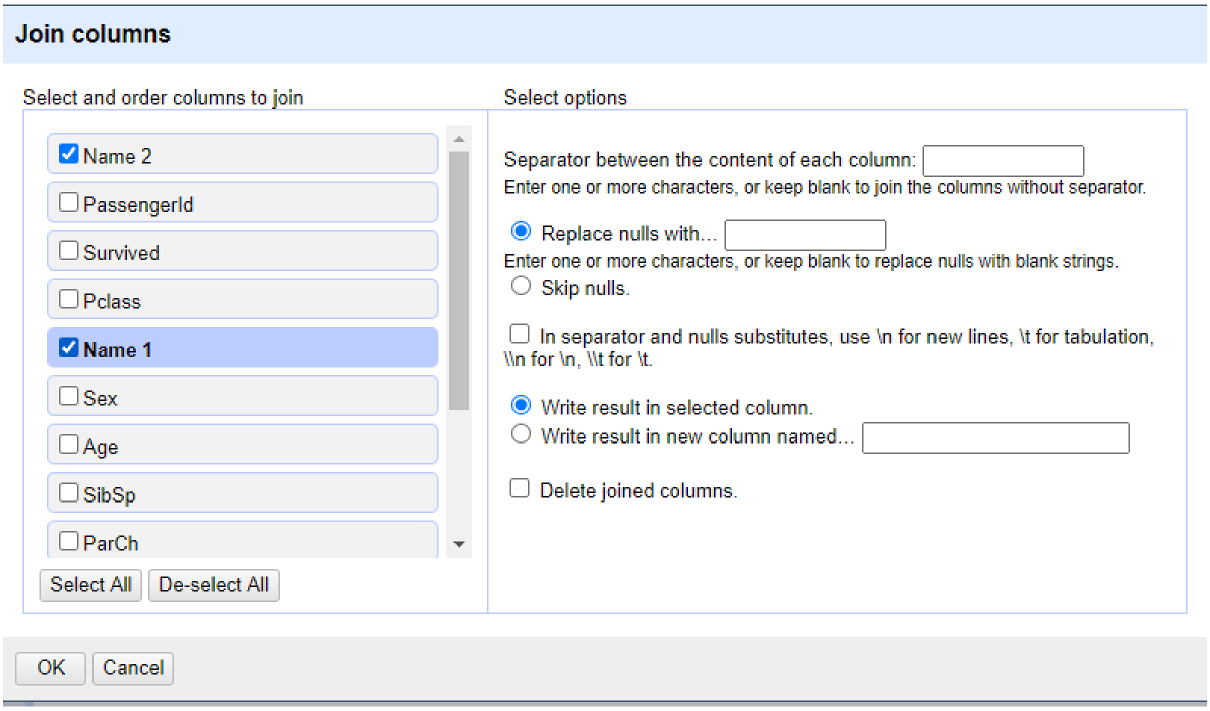
```
value.replace(/[^\u0020-\u007F]/,"")
```

No syntax error.

row	value	value.replace(/[^\u0020-\u007F ...
131.	Draï¿½enovic, Mr. Jozef	Draenovic, Mr. Jozef

STEP 7:

Order of the name is in the improper format so the conformity can be achieved by using splitting and joining the columns shown below



Join columns

Select and order columns to join

☑ Name 2
☐ PassengerId
☐ Survived
☐ Pclass
☑ Name 1
☐ Sex
☐ Age
☐ SibSp
☐ ParCh

Select options

Separator between the content of each column:
Enter one or more characters, or keep blank to join the columns without separator.

● Replace nulls with...
Enter one or more characters, or keep blank to replace nulls with blank strings.
○ Skip nulls.

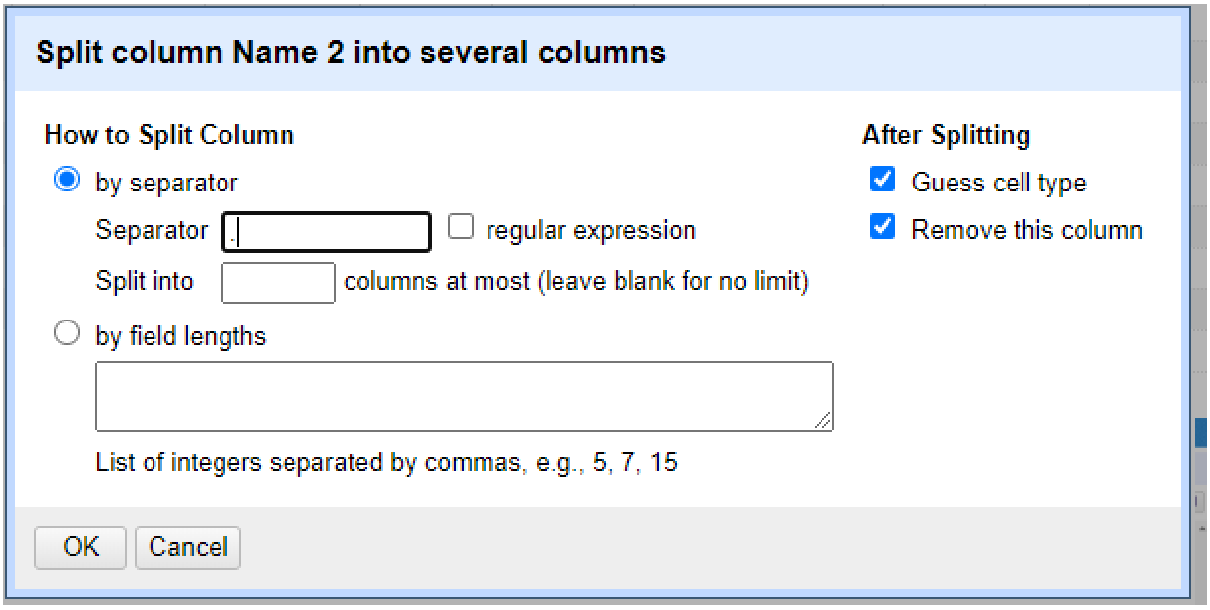
☐ In separator and nulls substitutes, use \n for new lines, \t for tabulation, \\n for \n, \\t for \t.

● Write result in selected column.
○ Write result in new column named...

☐ Delete joined columns.

STEP 8:

Extract the title from the name column by splitting using dot operator



Split column Name 2 into several columns

How to Split Column

● by separator

Separator .   ☐ regular expression

Split into columns at most (leave blank for no limit)

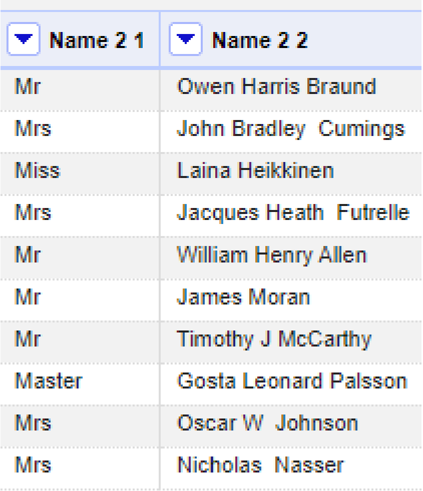
○ by field lengths

List of integers separated by commas, e.g., 5, 7, 15

After Splitting

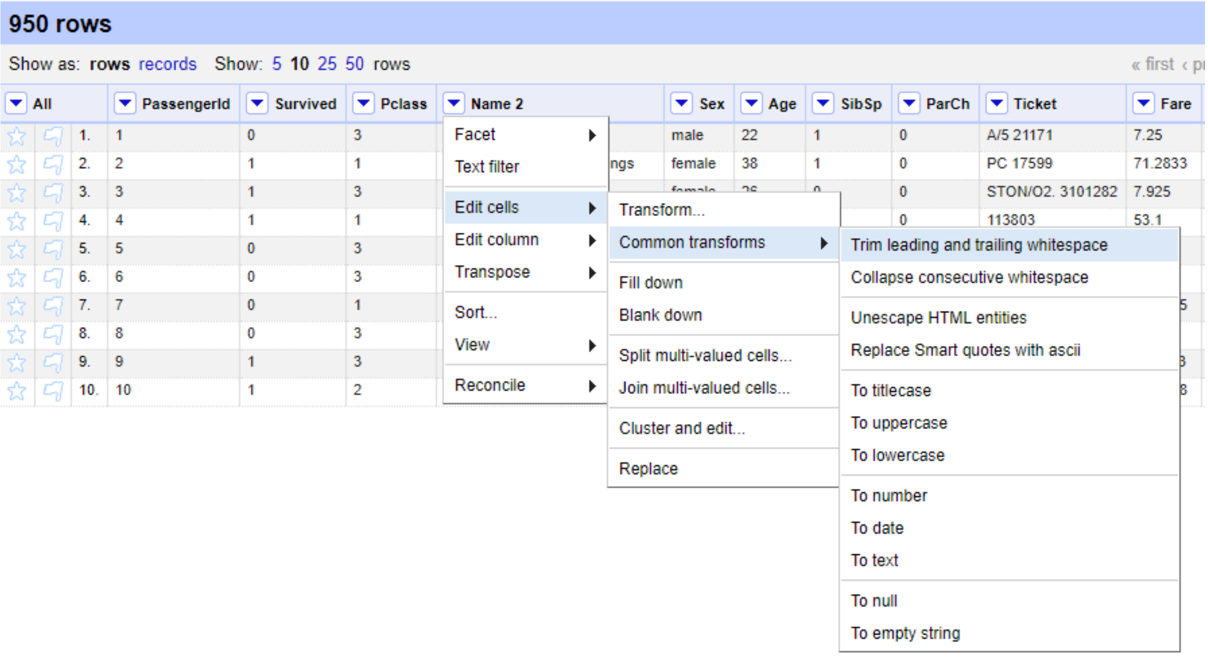
☑ Guess cell type
☑ Remove this column

Outcome of title column:



Name 2 1	Name 2 2
Mr	Owen Harris Braund
Mrs	John Bradley Cumings
Miss	Laina Heikkinen
Mrs	Jacques Heath Futrelle
Mr	William Henry Allen
Mr	James Moran
Mr	Timothy J McCarthy
Master	Gosta Leonard Palsson
Mrs	Oscar W Johnson
Mrs	Nicholas Nasser

STEP 9: Trim leading and trailing white space



950 rows

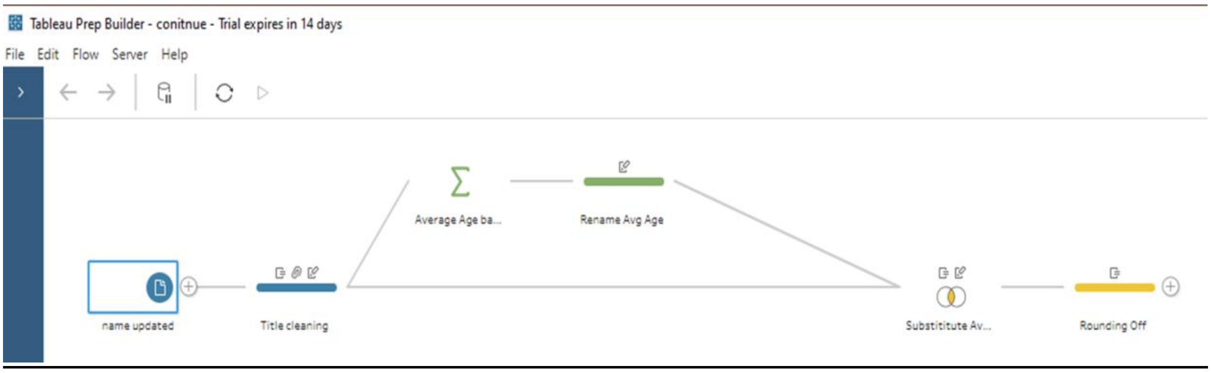
Show as: rows records   Show: 5 10 25 50 rows   « first ‹ pr

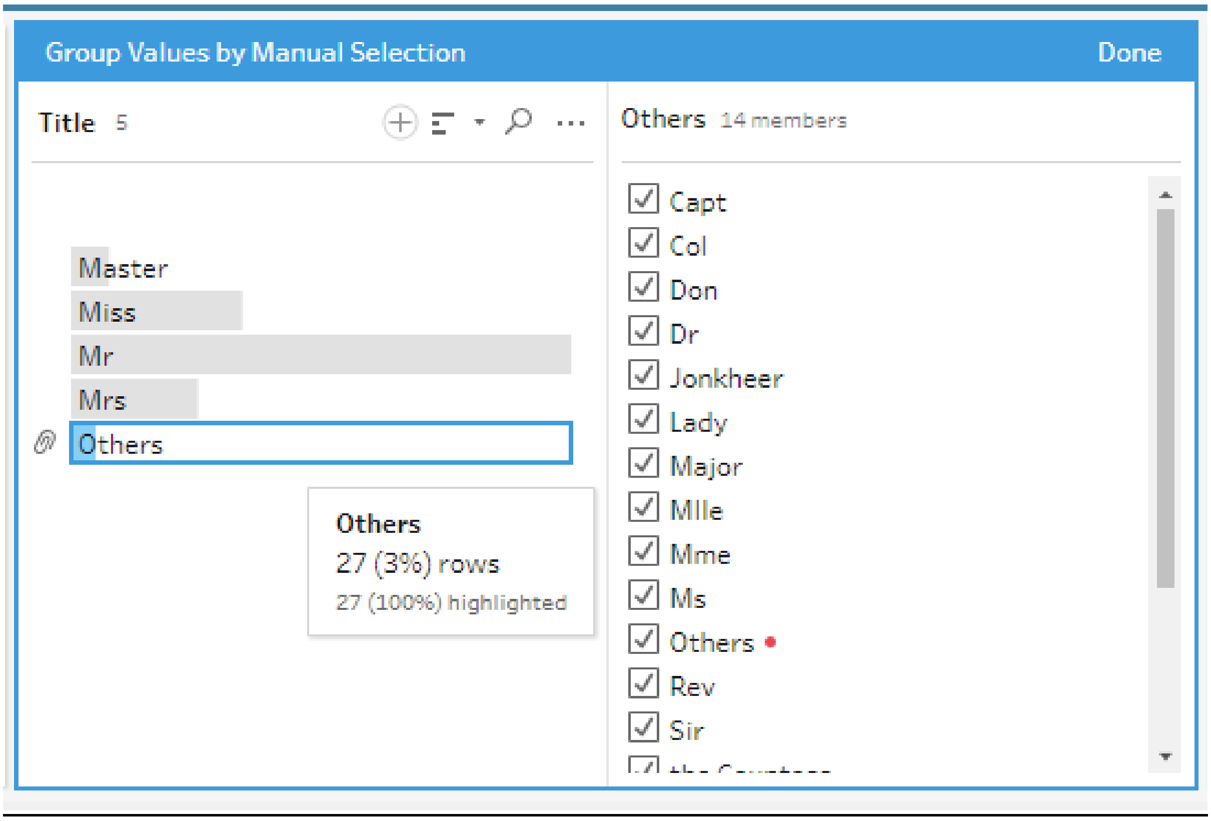
All	PassengerId	Survived	Pclass	Name 2	Sex	Age	SibSp	ParCh	Ticket	Fare
1.	1	0	3		male	22	1	0	A/5 21171	7.25
2.	2	1	1	ngs	female	38	1	0	PC 17599	71.2833
3.	3	1	3		female	26	0	0	STON/O2. 3101282	7.925
4.	4	1	1					0	113803	53.1
5.	5	0	3							
6.	6	0	3							
7.	7	0	1							5
8.	8	0	3							
9.	9	1	3							3
10.	10	1	2							8

Title Column cleaning(Name cleaning part):

STEP 1:

Groupling values by manual selection of titles which are valid (Mr,Mrs,Miss,Master) and all other invalid values are replaced/filled as (others). So validity is achieved. It doesn't have any spelling mistakes so accuracy is there by default, conformity is present because all were of same data type. consistency was achieved because it follows important relationship linkages.





Group Values by Manual Selection   Done

Title 5

Master
Miss
Mr
Mrs
Others

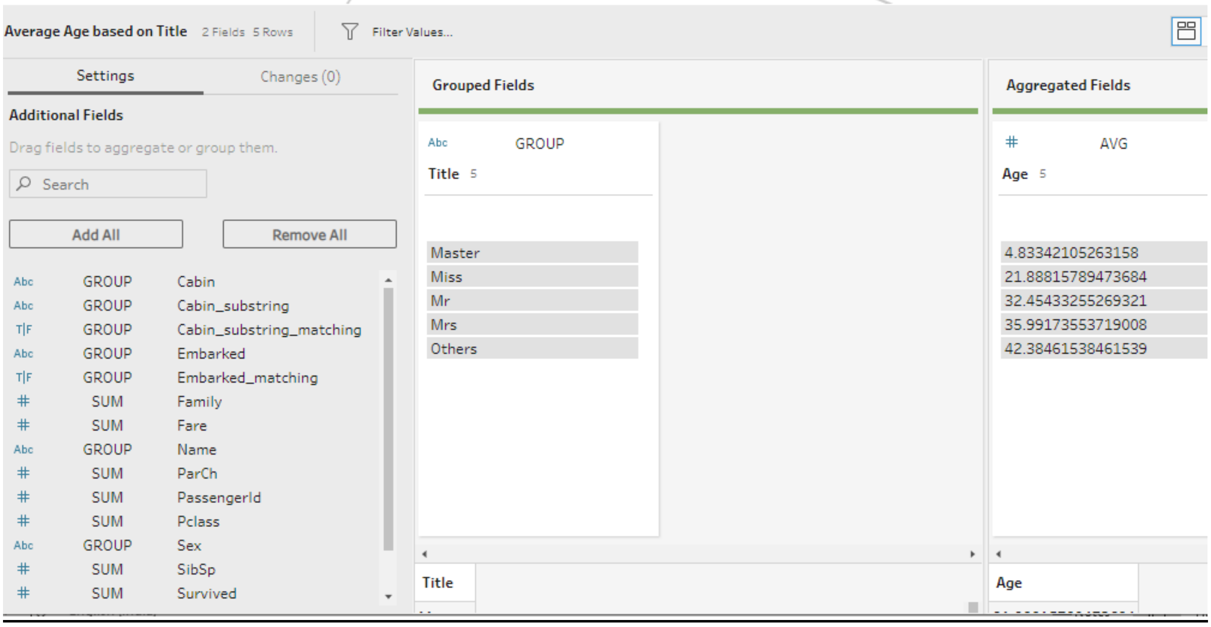
Others
27 (3%) rows
27 (100%) highlighted

Others 14 members

☑ Capt
☑ Col
☑ Don
☑ Dr
☑ Jonkheer
☑ Lady
☑ Major
☑ Mlle
☑ Mme
☑ Ms
☑ Others
☑ Rev
☑ Sir
☑ the Countess

Age cleaning:

STEP 1: Many missing values present on Age column which has lack of completeness. So completeness is achieved by filling the missing values on Age column by taking average based Tite column .For eg. All age who have title ‘Mr’ has an average 32.454332 and that has been filled by ‘GROUPING FIELDS and AGGREGATING’ using Tableau prep.



Average Age based on Title 2 Fields 5 Rows

Additional Fields

Drag fields to aggregate or group them.

Abc	GROUP	Cabin
Abc	GROUP	Cabin_substring
T|F	GROUP	Cabin_substring_matching
Abc	GROUP	Embarked
T|F	GROUP	Embarked_matching
#	SUM	Family
#	SUM	Fare
Abc	GROUP	Name
#	SUM	ParCh
#	SUM	PassengerId
#	SUM	Pclass
Abc	GROUP	Sex
#	SUM	SibSp
#	SUM	Survived

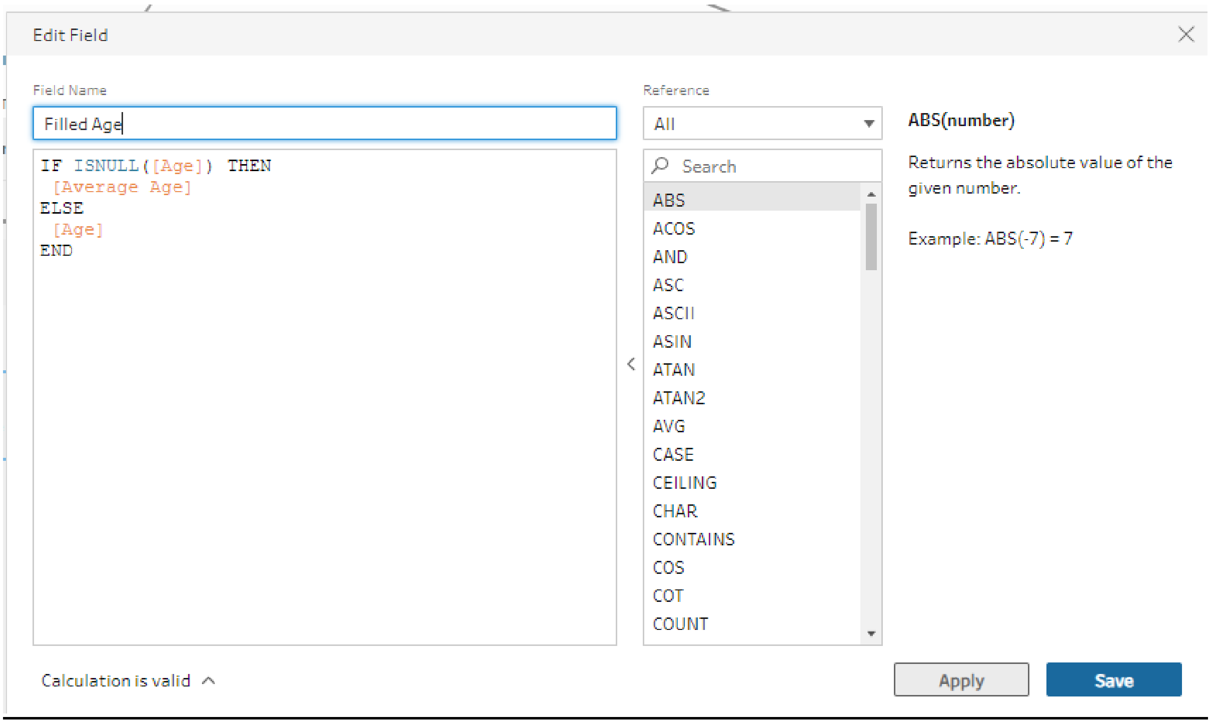
Grouped Fields

Aggregated Fields

Abc GROUP Title 5	# AVG Age 5
Master	4.83342105263158
Miss	21.88815789473684
Mr	32.45433255269321
Mrs	35.99173553719008
Others	42.38461538461539

STEP2:

Averages found based on each title above has been filled on the blank fields on Age column by using command below and created a new column with ‘Filled Age’ .so to conclude completeness has been achieved.



Edit Field

Field Name: Filled Age

```
IF ISNULL([Age]) THEN
 [Average Age]
ELSE
 [Age]
END
```

Reference: All

ABS(number)

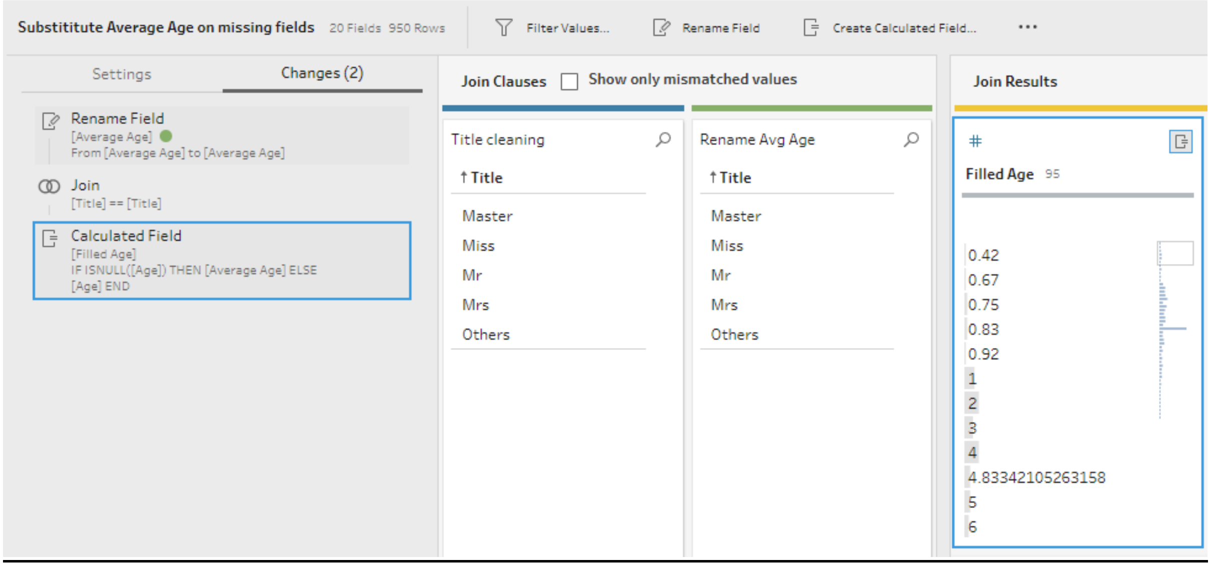
Returns the absolute value of the given number.

Example: ABS(-7) = 7

Calculation is valid

STEP3:

Rounding off Ages by using command below .to achive consistency



Substititute Average Age on missing fields 20 Fields 950 Rows

Rename Field
[Average Age]
From [Average Age] to [Average Age]

Join
[Title] == [Title]

Calculated Field
[Filled Age]
IF ISNULL([Age]) THEN [Average Age] ELSE [Age] END

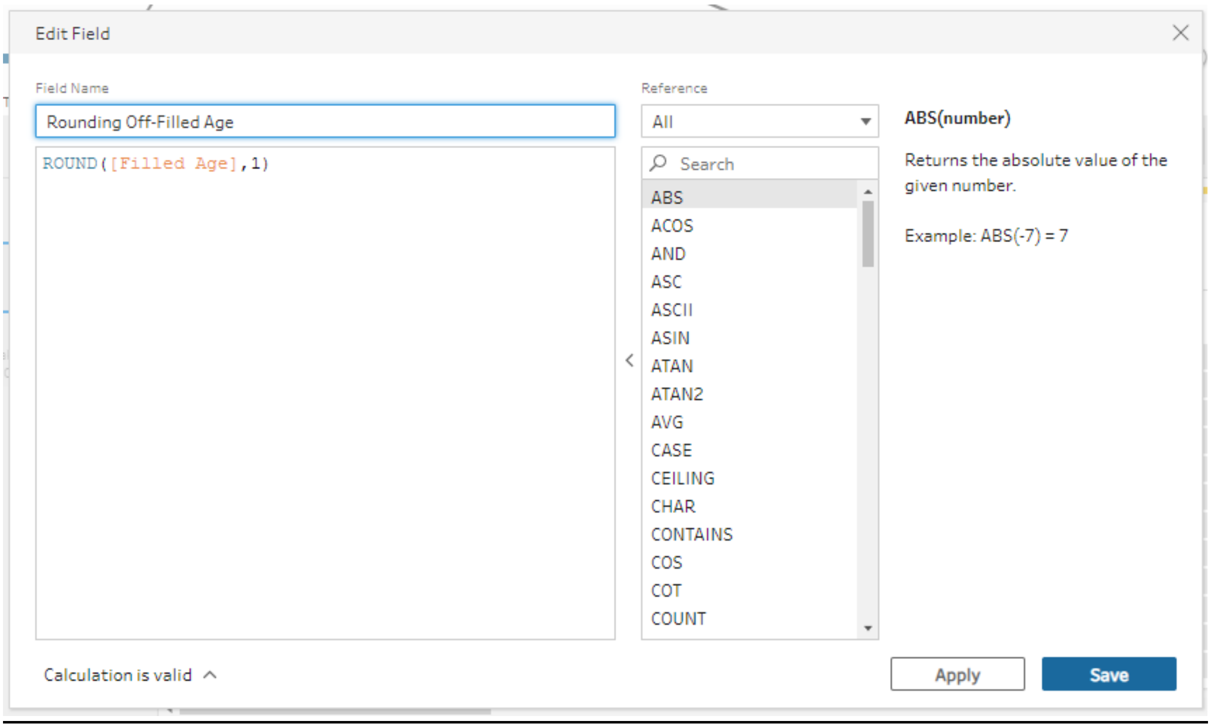
Join Clauses □ Show only mismatched values

Title cleaning	Rename Avg Age
Title	Title
Master	Master
Miss	Miss
Mr	Mr
Mrs	Mrs
Others	Others

Join Results

Filled Age 95

0.42
0.67
0.75
0.83
0.92
1
2
3
4
4.83342105263158
5
6



Edit Field

Field Name: Rounding Off-Filled Age

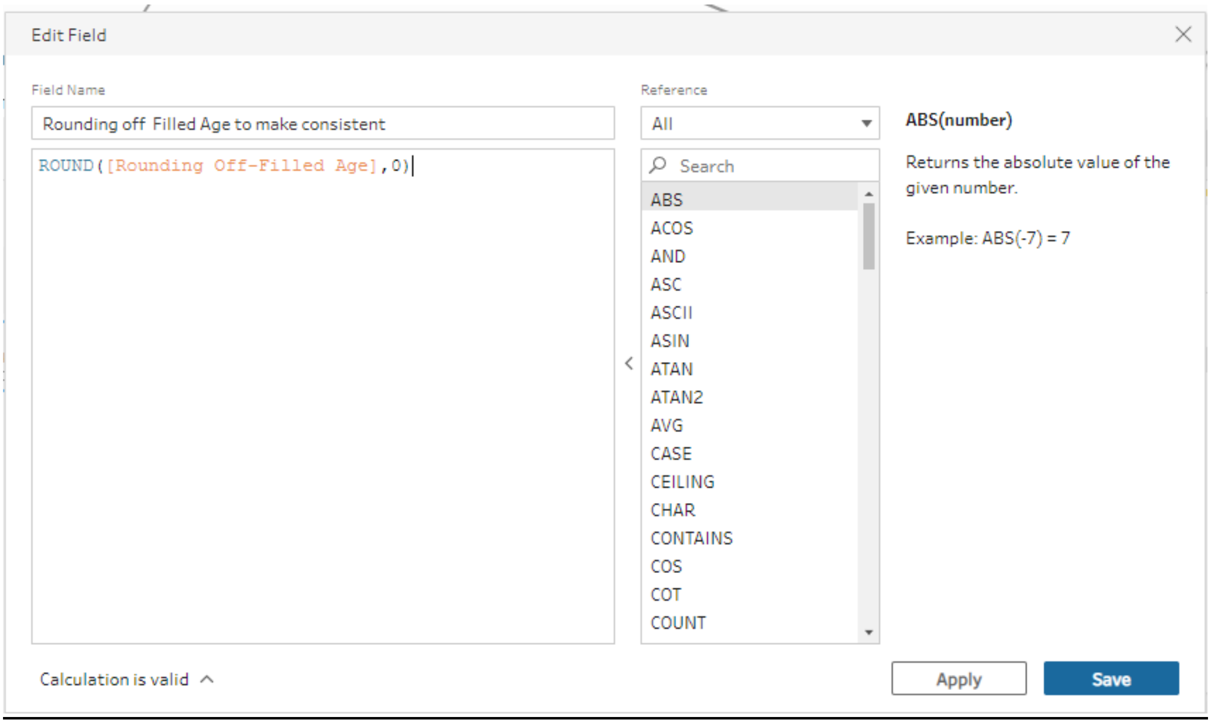
```
ROUND([Filled Age],1)
```

ABS(number)

Returns the absolute value of the given number.

Example: ABS(-7) = 7

Calculation is valid



Edit Field

Field Name

Rounding off Filled Age to make consistent

```
ROUND([Rounding Off-Filled Age],0)
```

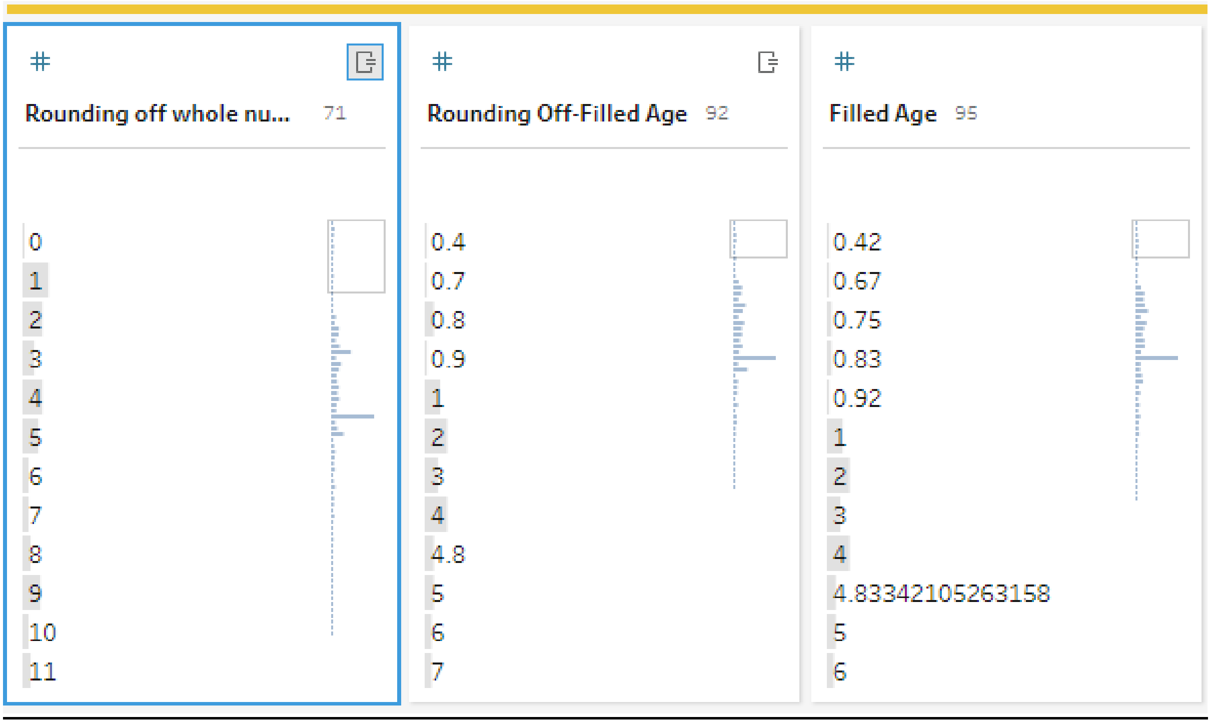
Reference: All

ABS(number)

Returns the absolute value of the given number.

Example: ABS(-7) = 7

Calculation is valid

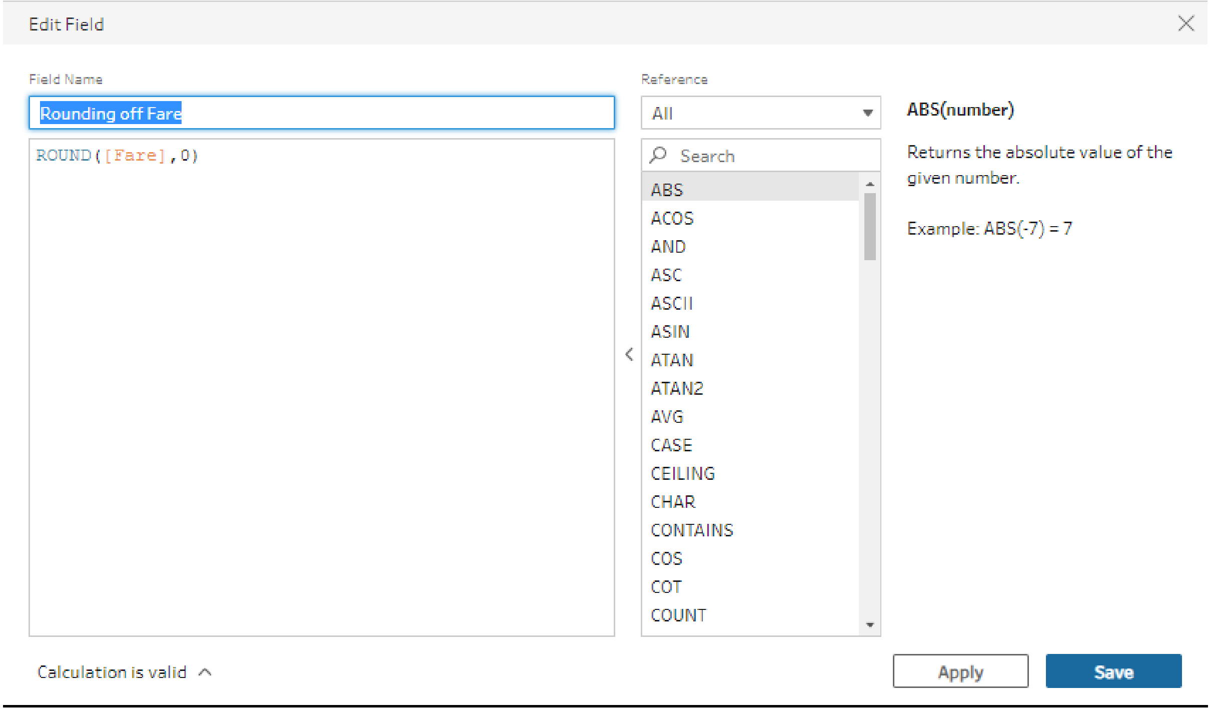


# Rounding off whole nu... 71	# Rounding Off-Filled Age 92	# Filled Age 95
0	0.4	0.42
1	0.7	0.67
2	0.8	0.75
3	0.9	0.83
4	1	0.92
5	2	1
6	3	2
7	4	3
8	4.8	4
9	5	4.83342105263158
10	6	5
11	7	6

Fare:

STEP1:

Fare column has been inconsistent due to decimal values present on few fields so to conclude consistency is achieved after rounding off .all other quality dimensions is present by default.



Field Name: Rounding off Fare

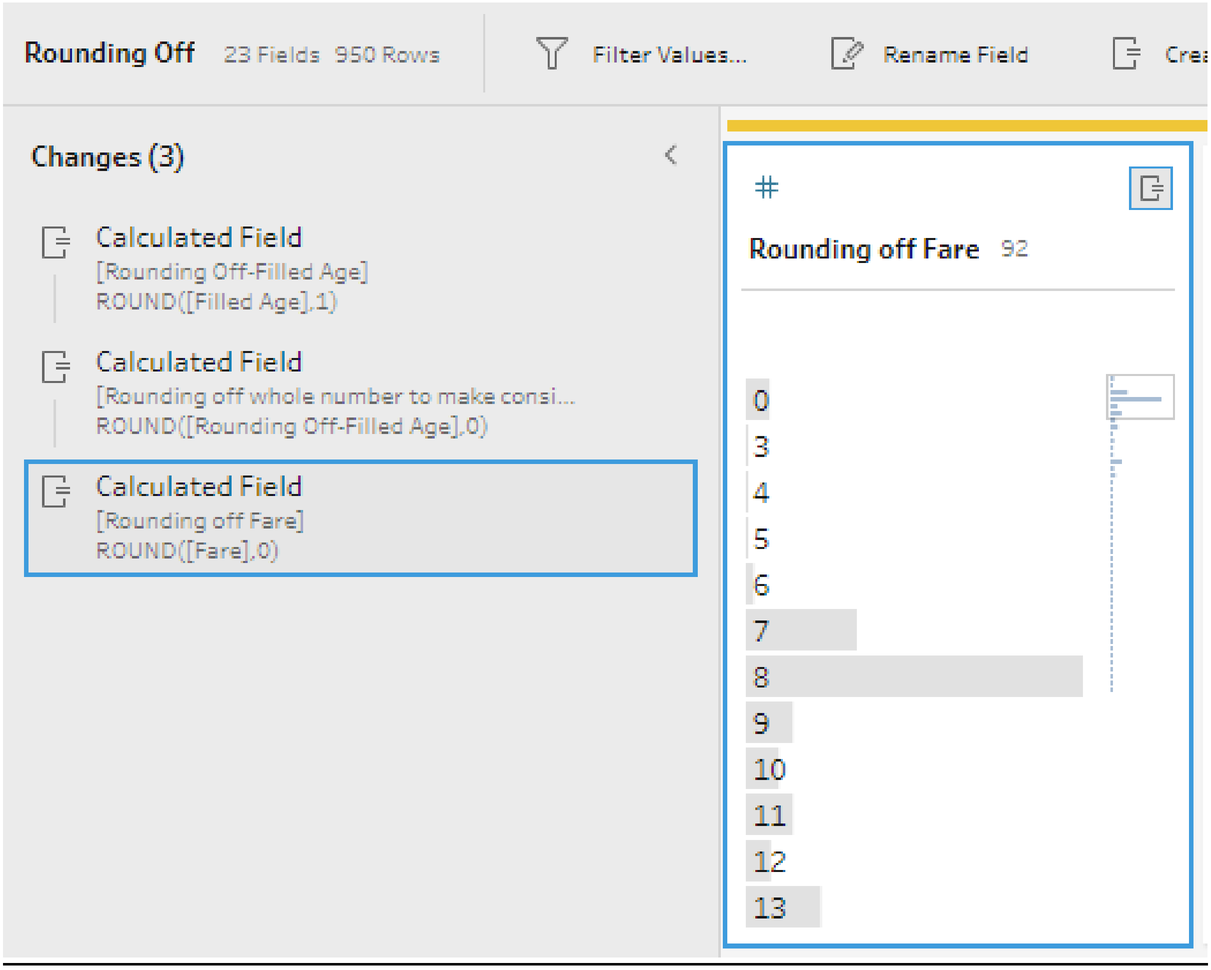
```
ROUND([Fare],0)
```

ABS(number)

Returns the absolute value of the given number.

Example: ABS(-7) = 7

Calculation is valid



Changes (3)

Calculated Field
[Rounding Off-Filled Age]
ROUND([Filled Age],1)

Calculated Field
[Rounding off whole number to make consi...
ROUND([Rounding Off-Filled Age],0)

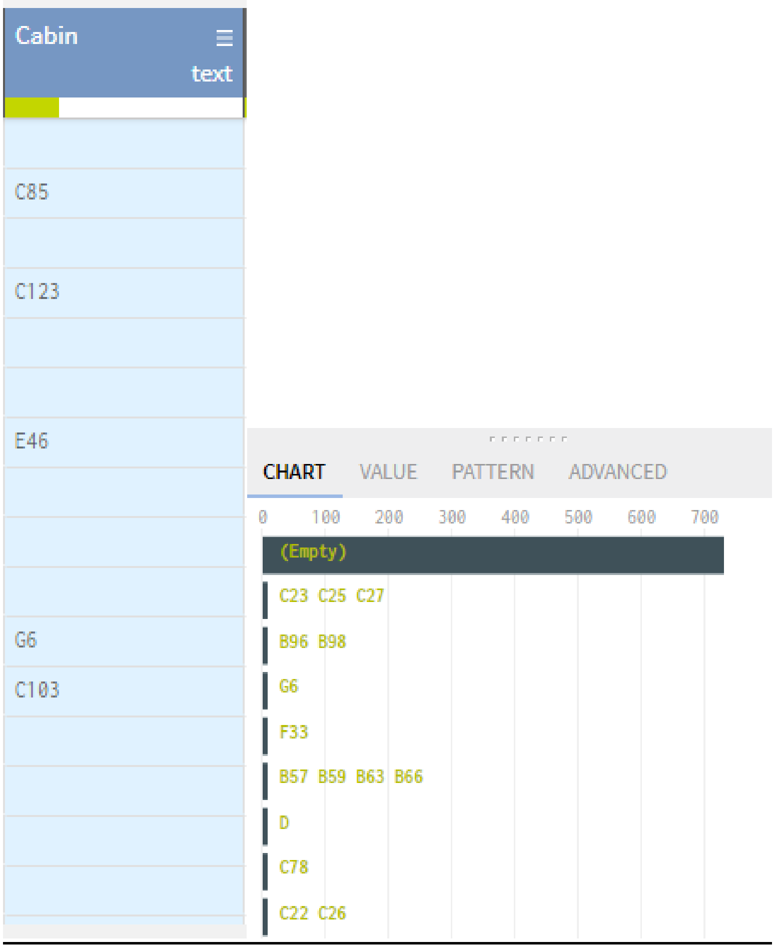
Calculated Field
[Rounding off Fare]
ROUND([Fare],0)

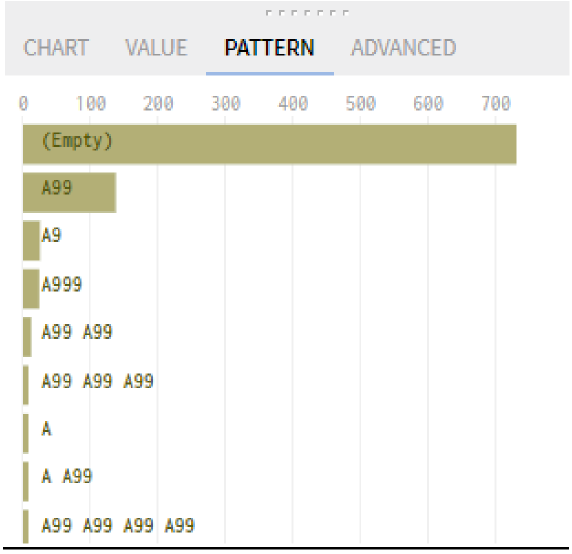
Rounding off Fare 92

Cabin

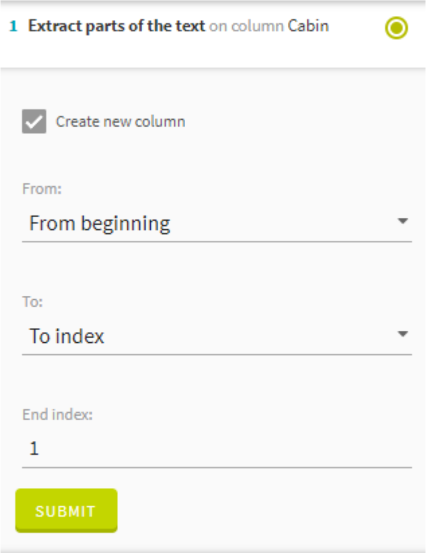
STEP 1:

Many missing values and cabin values has different data formats ,so conformity, consistency and completeness is missing.To achieve those,





First value alone has been extracted throughout by using the function 'Extract part of the text' because many values has only first value on this column .so consistency and conformity has been achied through this step.



1 **Extract parts of the text** on column Cabin

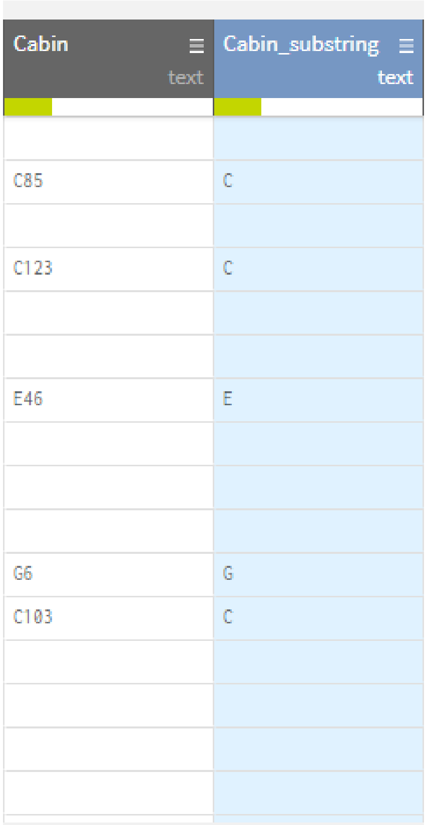
☑ Create new column

From: From beginning

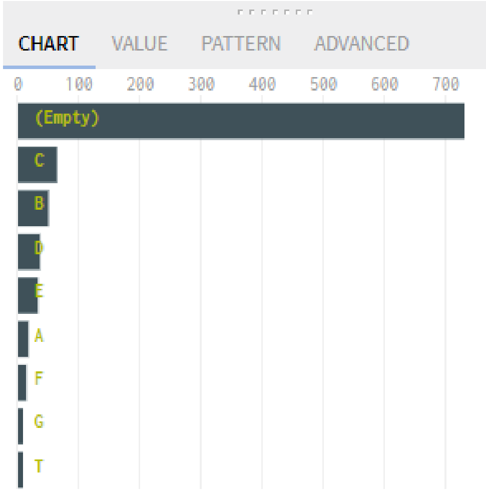
To: To index

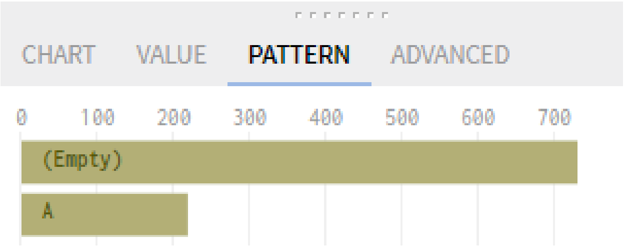
End index: 1

<u>Output of step 1</u>:



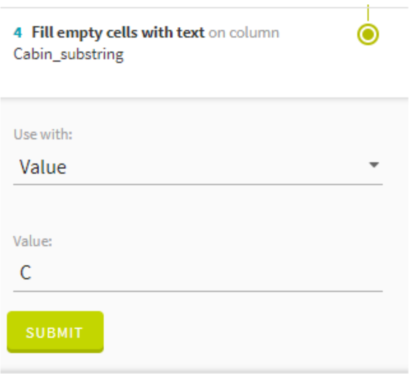
Cabin	Cabin_substring
text	text
C85	C
C123	C
E46	E
G6	G
C103	C





STEP 2:

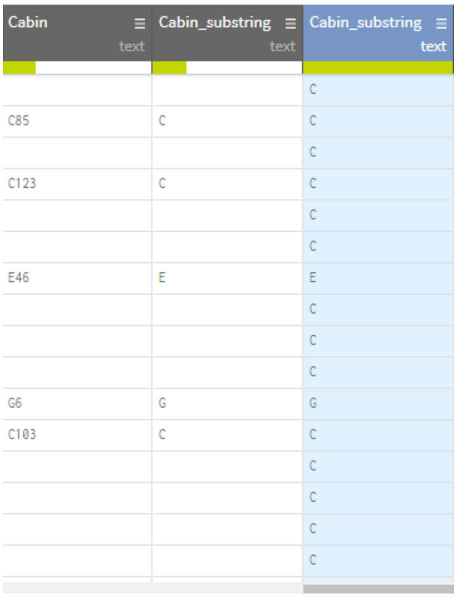
Completeness achieved by using the function below by filling all values using C because on average ,C is the most occurred cabin value so that has been replaced on the blank values.



4 **Fill empty cells with text** on column Cabin_substring

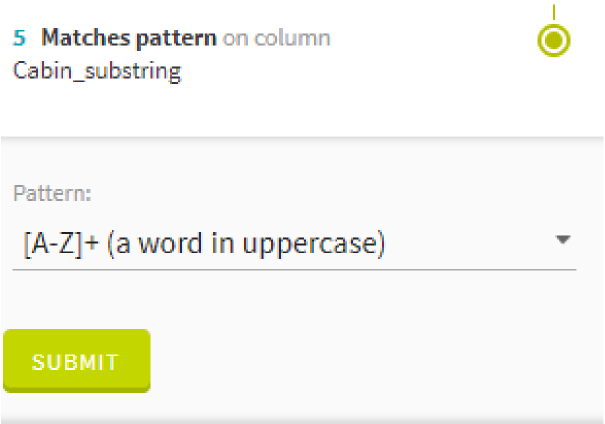
Use with: Value

Value: C



Cabin text	Cabin_substring text	Cabin_substring text
		C
C85	C	C
		C
C123	C	C
		C
		C
E46	E	E
		C
		C
		C
G6	G	G
C103	C	C
		C
		C
		C
		C

STEP 3:values changed to upper case to achieve consistency

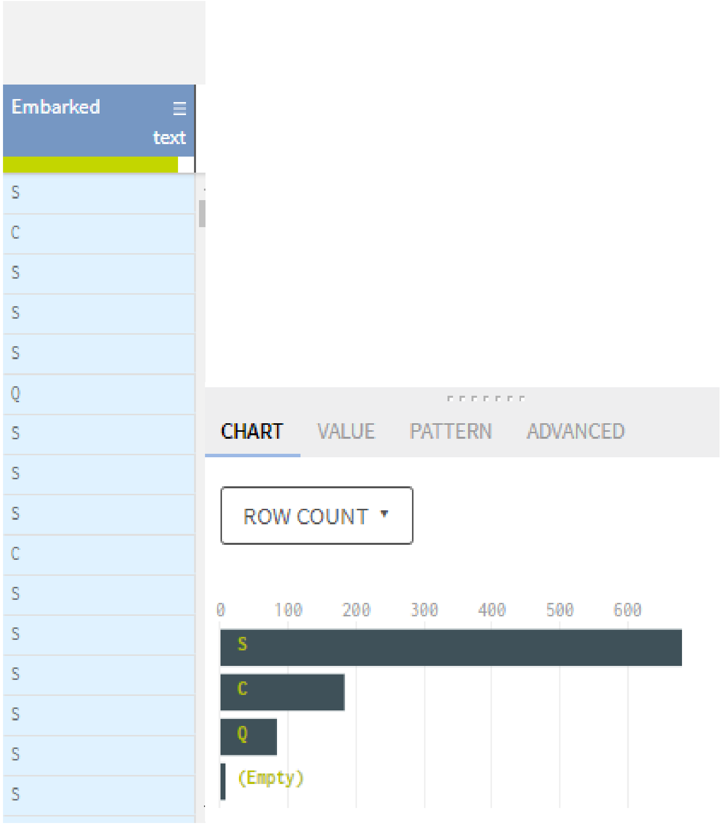


5 Matches pattern on column Cabin_substring

Pattern: [A-Z]+ (a word in uppercase)

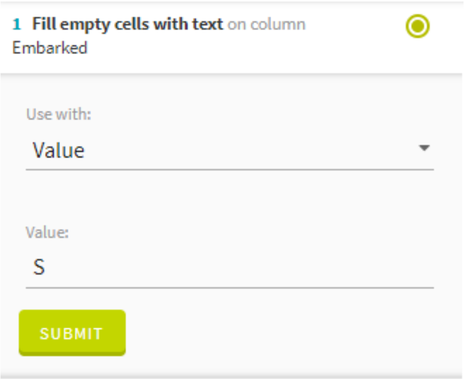
Embarked:

Completeness on missing fields is achieved by replacing is with the S(southhampton) based on average which has been occurred most frequently.



Embarked text

S
C
S
S
S
Q
S
S
S
C
S
S
S
S
S
S



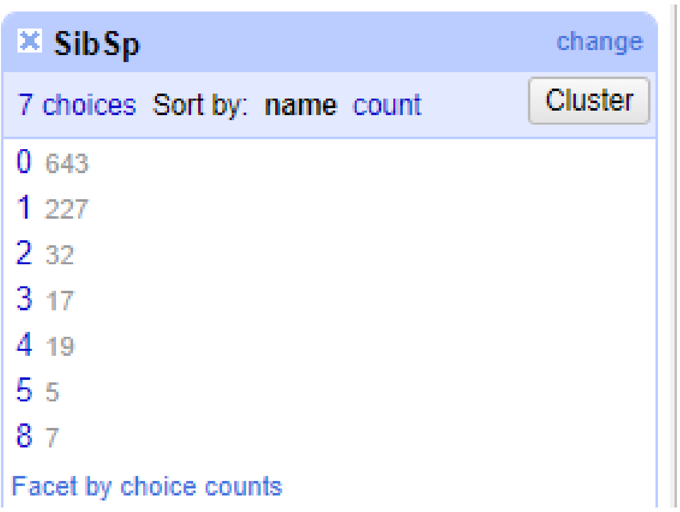
1 Fill empty cells with text on column Embarked

Use with: Value

Value: S

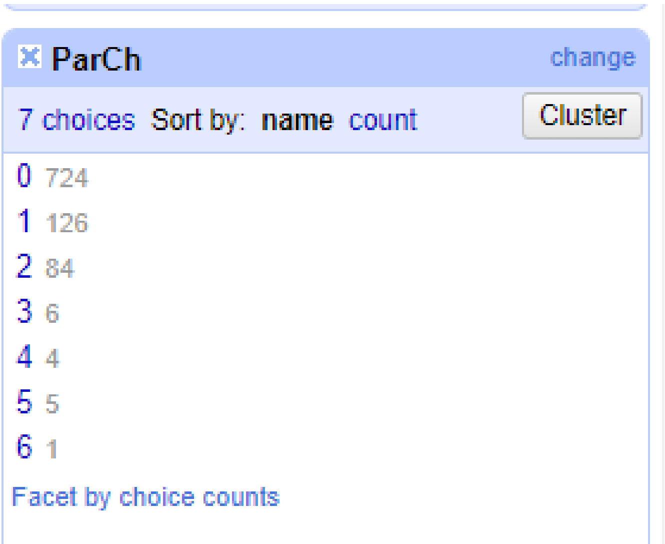
SIBLINGS SPOUSE , PARENT CHILD AND SEX COLUMNS:

Quality of data based on many dimensions has already been present by default.



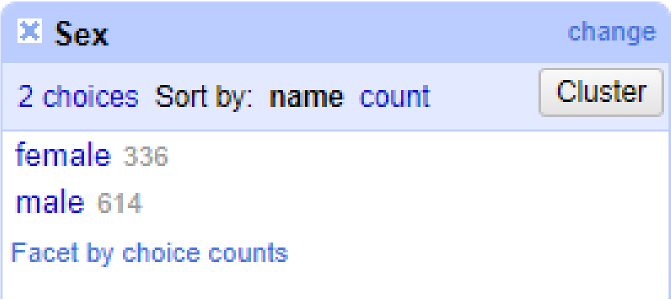
SibSp

0	643
1	227
2	32
3	17
4	19
5	5
8	7



ParCh

0	724
1	126
2	84
3	6
4	4
5	5
6	1



female 336

male 614

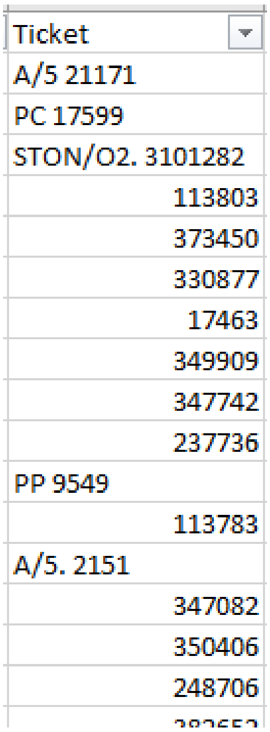
Ticket:

Ticket column has inconsistency and conformity issues.so this has been achieved by removing the / and . using

COMMANDS:

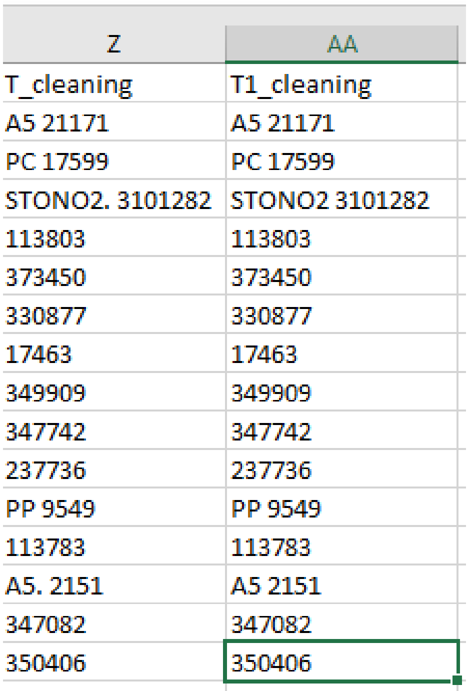
```
=SUBSTITUTE(O2,”/”,””)

=SUBSTITUTE(Z2,”.”,””)
```



Ticket
A/5 21171
PC 17599
STON/O2. 3101282
113803
373450
330877
17463
349909
347742
237736
PP 9549
113783
A/5. 2151
347082
350406
248706

Output:



Z	AA
T_cleaning	T1_cleaning
A5 21171	A5 21171
PC 17599	PC 17599
STONO2. 3101282	STONO2 3101282
113803	113803
373450	373450
330877	330877
17463	17463
349909	349909
347742	347742
237736	237736
PP 9549	PP 9549
113783	113783
A5. 2151	A5 2151
347082	347082
350406	350406