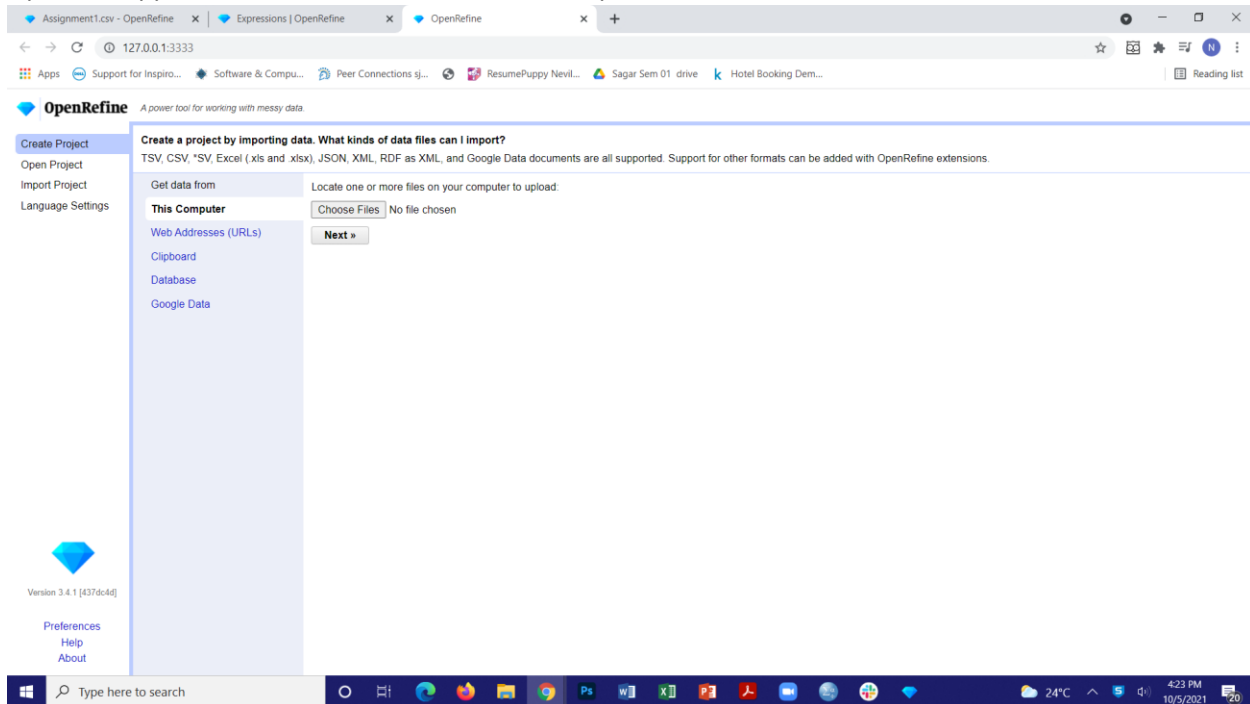


## Data Cleaning using OpenRefine

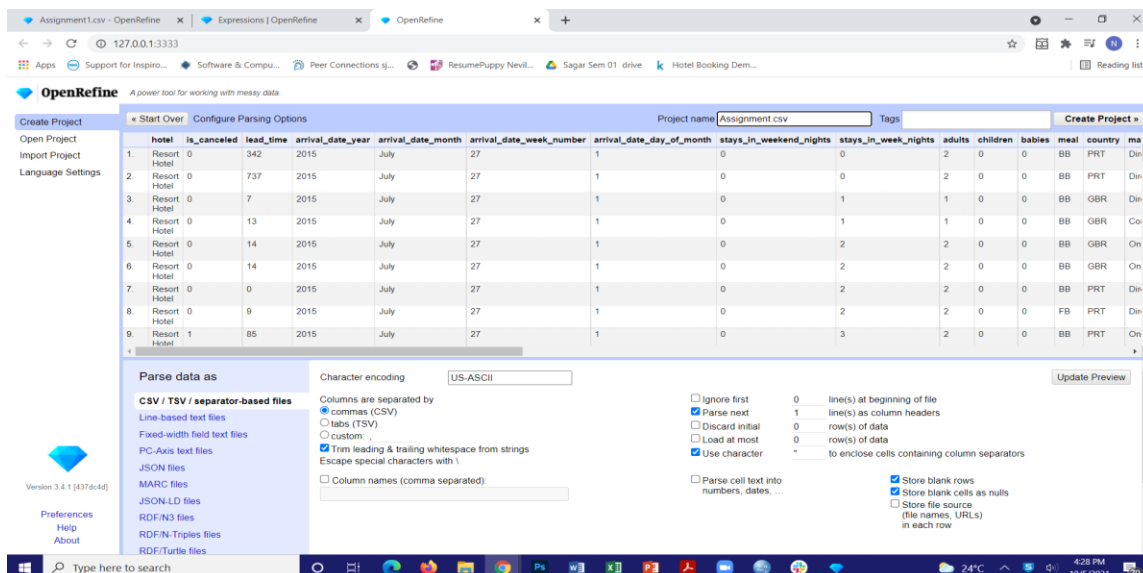
### Step 1: Install the OpenRefine Application

Open the application and this window would be opened.



### Step 2:- Inserting the Dataset

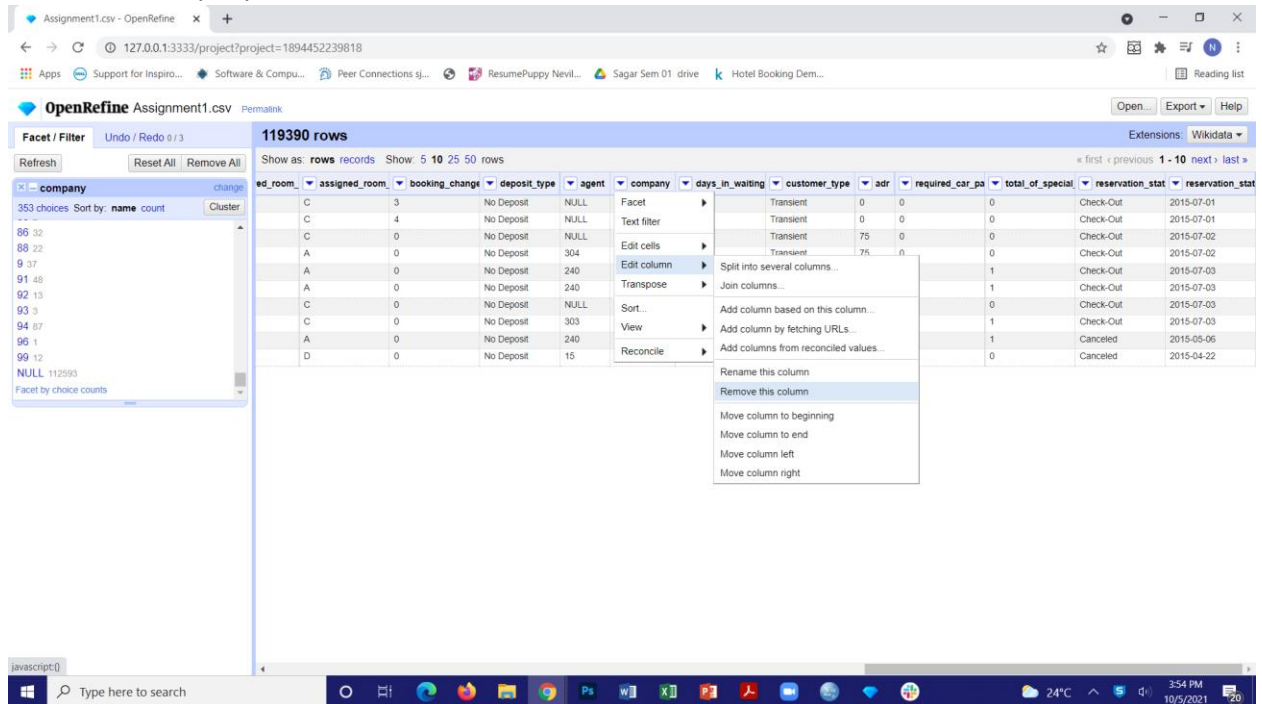
In the choose file button, Upload the Dataset and then, Click on Create Project.



### Step 3: Data Cleaning

Now, we need to process the null values present in the columns in the dataset in the correct manner.

#### 1) Column == Company



As we can see from the above image, about 112k values in this column are Null , hence , we will remove this column

#### 2) Column == 'Agent'

There are about 16k null values in this column. Further , this column will not be used in further analysis , hence , we will drop the entire column.

OpenRefine Assignment1.csv

Facet / Filter Undo / Redo 1 / 1

agent

119390 rows

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

Extensions: Wikidata

« first < previous 1 - 10 next > last »

agent reserved\_room assigned\_room booking\_change deposit\_type agent days\_in\_waiting customer\_type adr required\_car\_pa total\_of\_special reservation\_stat reservation\_stat

Facet by choice counts

javascript:0

Type here to search

24°C 3:55 PM 10/5/2021

### 3) Column == 'Children'

OpenRefine Assignment1.csv

Facet / Filter Undo / Redo 2 / 2

children

4 matching rows (119390 total)

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

Extensions: Wikidata

« first < previous 1 - 4 next > last »

All hotel is\_canceled lead\_time arrival\_date\_year arrival\_date\_month arrival\_date\_week arrival\_date\_day stays\_in\_week stays\_in\_week\_adults children babies

Transform Hotel 1 2 2015 August 32 3 1 0 2 NA 0 BB

Facet Hotel 1 1 2015 August 32 5 0 2 2 NA 0 BB

Edit rows Star rows

Edit columns Unstar rows

View Flag rows

Unflag rows

Remove matching rows

Facet by choice counts

javascript:0

Type here to search

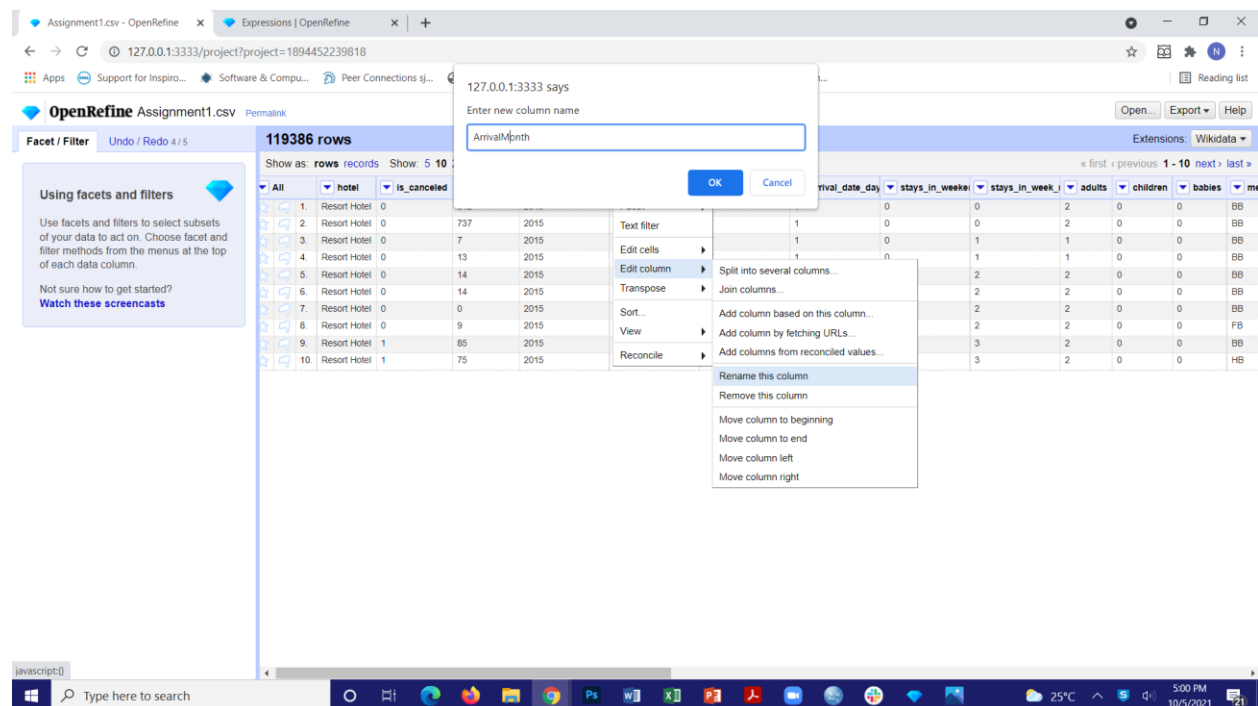
24°C 3:56 PM 10/5/2021

As seen from the image above, there are 4 records in the 'Children' column which has null values. Hence, we will remove the 4 records.

Steps to do so:

- 1) After displaying the 4 records as shown in the image, go to 'Edit rows' and select 'Flag rows' option.
- 2) Then, go to 'Edit rows' and select 'Remove Matching rows' option. This step will remove all the 4 records in the "Children" column that contained Null values.

How to rename a Column in OpenRefine?



Step 1:- Go into 'Edit column' and select 'Rename the Column' option.

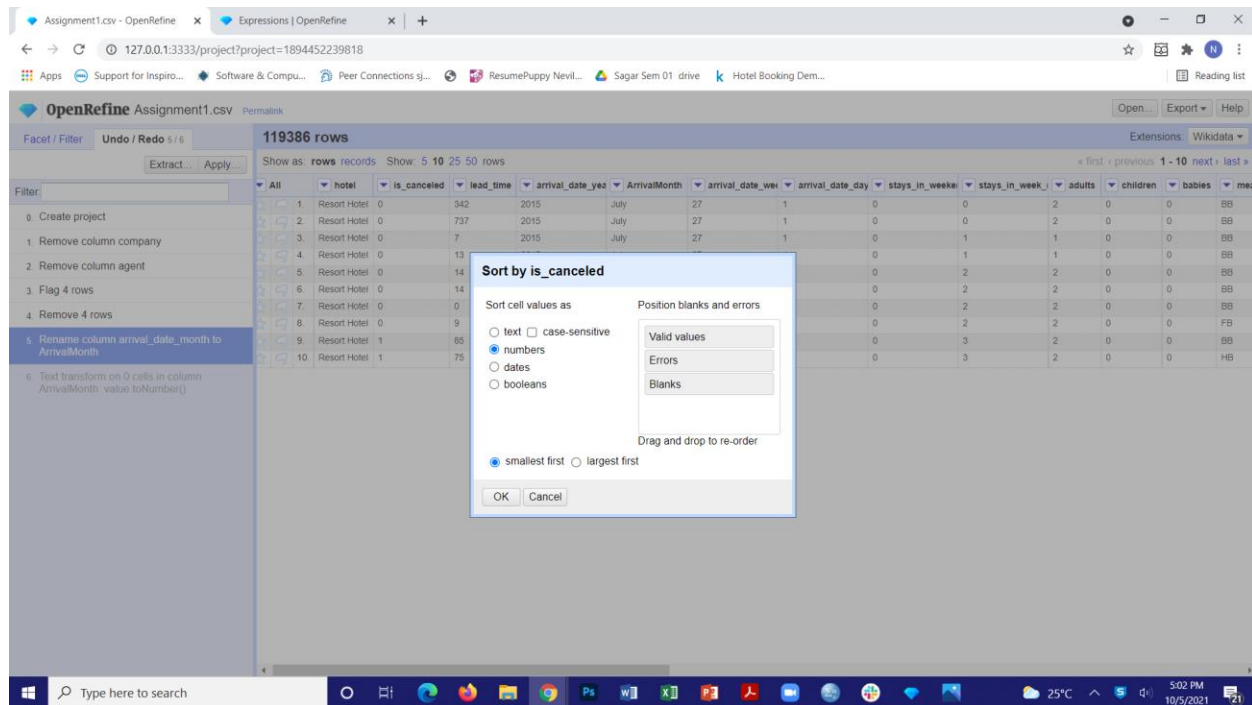
Step 2:- A Dialog box will be opened at the top, write the new name of the column and Click 'ok'.

Now, the Column 'arrival\_by\_month' is replaced by 'ArrivalMonth'.

## Sorting a column:-

Here, we will sort the 'is\_cancelled' column.

Step 1:- Select the 'Sort' option and a dialog box like this will be opened.



Step 2:- Select the radio button according to which you want to sort the values.

Step 3:- Click 'Smallest first' if you want to sort in the ascending order and 'largest first' for the descending order.

Step 4:- Select 'OK'.

The column 'is\_cancelled' is now sorted in ascending order. In this way, we can sort a column in OpenRefine.

## Deleting a Column:-

Here, we will perform 'remove the column' operation on 2 columns namely- 'previous\_booking' and 'previous\_cancel' as they will not be used very effectively in the further process.

Steps to remove a Column:-

Step 1:- Go to 'Edit column' as shown in the figure below and Select 'Remove this column'' option.

Step 2:- The column 'previous\_cancel' will be deleted after this process

The screenshot shows the OpenRefine web interface with a CSV file named 'Assignment1.csv' open. The interface displays 119,386 rows. A context menu is open over the 'previous\_bookings' column, showing options like 'Facet', 'Text filter', 'Edit cells', 'Edit column', 'Transpose', 'Sort...', 'View', 'Reconcile', 'Split into several columns...', 'Join columns...', 'Add column based on this column...', 'Add column by fetching URLs...', 'Add columns from reconciled values...', 'Rename this column', 'Remove this column', 'Move column to beginning', 'Move column to end', 'Move column left', and 'Move column right'. The 'Remove this column' option is highlighted.

Now, repeat the above 2 steps for the column 'previous\_bookings' and this column will also be deleted thereafter.

The screenshot shows the OpenRefine web interface after the 'previous\_bookings' column has been removed. A yellow banner at the top of the interface reads 'Remove column previous\_bookings\_not\_canceled Undo'. The interface still displays 119,386 rows. The 'previous\_bookings' column is no longer visible in the table. The 'Edit column' context menu is no longer open.

In this way, we can perform Data Cleaning in the Google OpenRefine.

References:-

- 1) <https://openrefine.org/documentation.html>
- 2) [https://www.youtube.com/watch?v=nORS7STbLyk&ab\\_channel=WebScraper](https://www.youtube.com/watch?v=nORS7STbLyk&ab_channel=WebScraper)
- 3) <https://thomaspadilla.org/dataprep/>
- 4) [https://www.youtube.com/watch?v=wGVtycv3SS0&ab\\_channel=UniversityofIdahoLibraryDigitalInitiatives](https://www.youtube.com/watch?v=wGVtycv3SS0&ab_channel=UniversityofIdahoLibraryDigitalInitiatives)