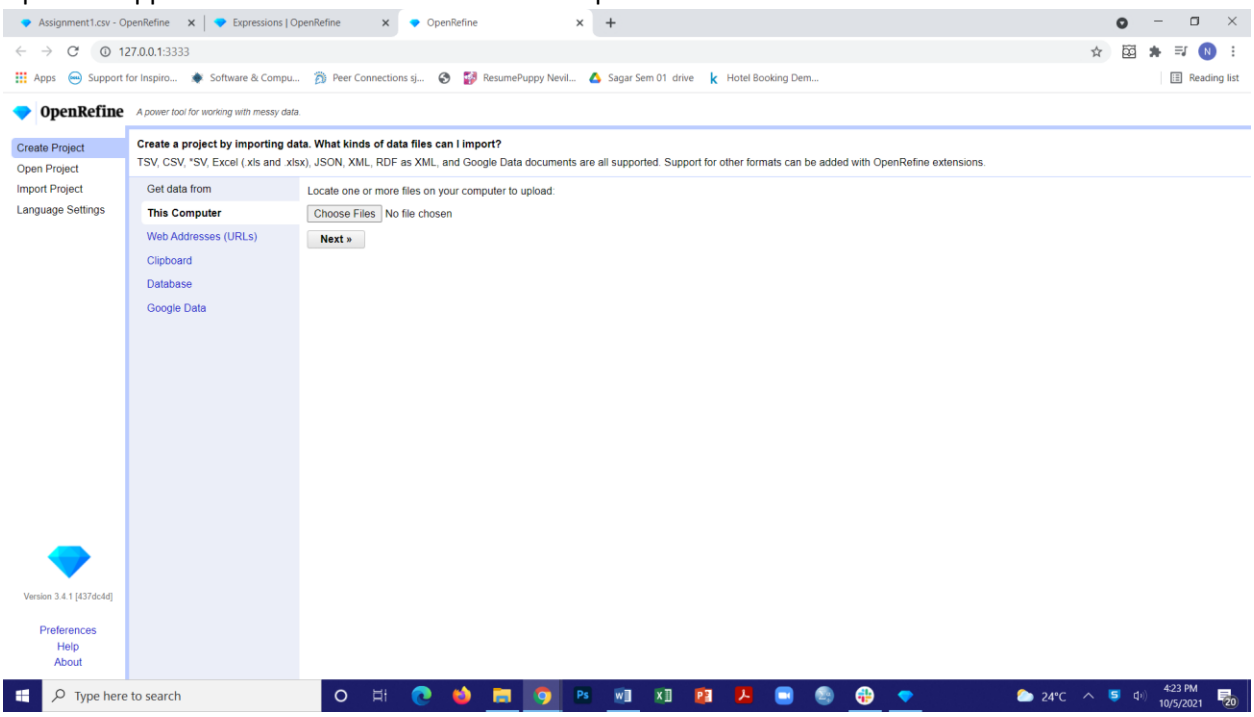
# Data Cleaning using Open-Refine

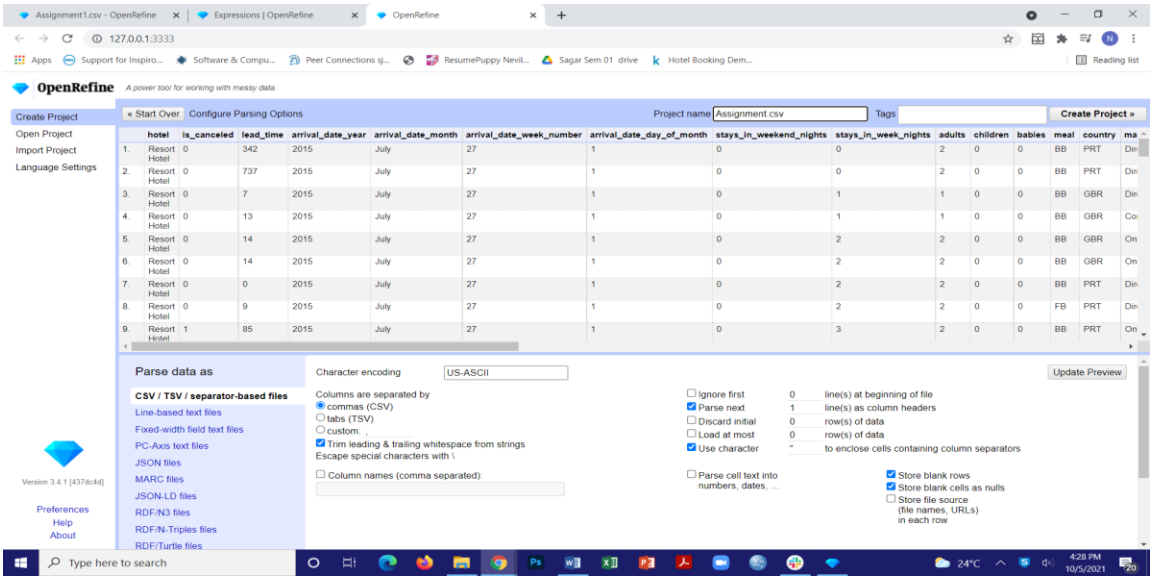
## Step 1: Install the OpenRefine Application

Open the application and this window would be opened.




## Step 2:- Inserting the Dataset

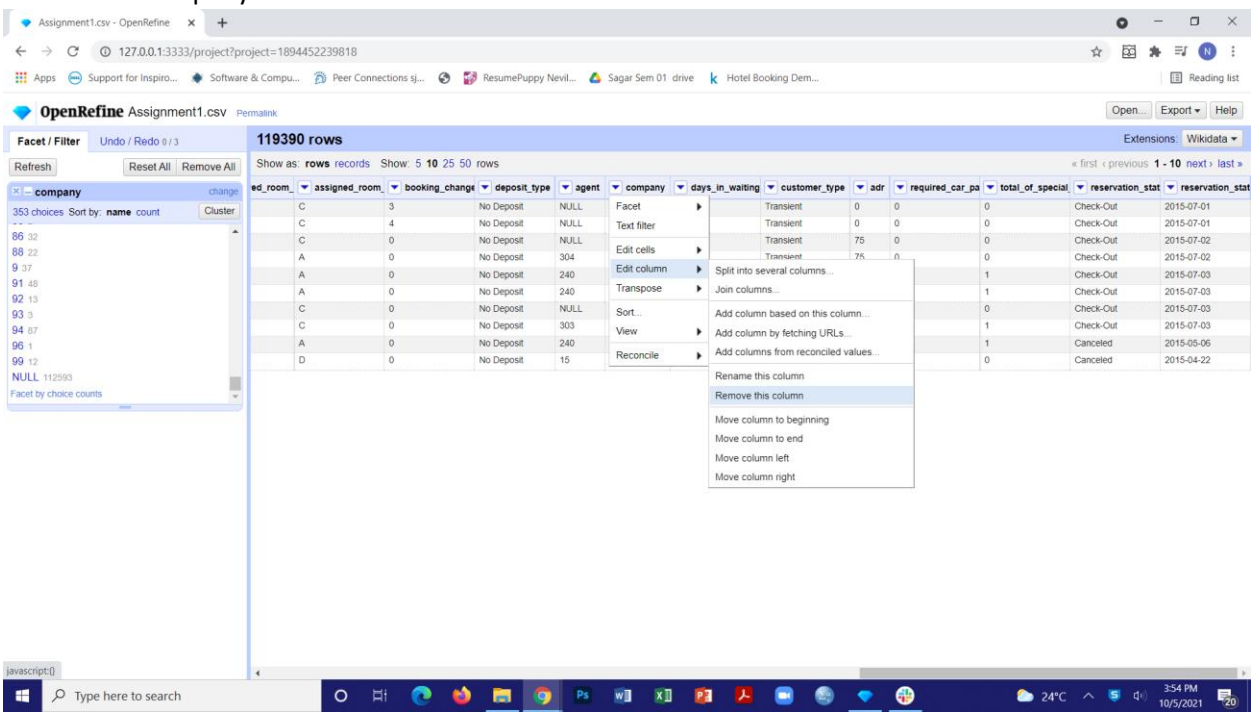
In the choose file button, Upload the Dataset and then, Click on Create Project.



## Step 3: Data Cleaning

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

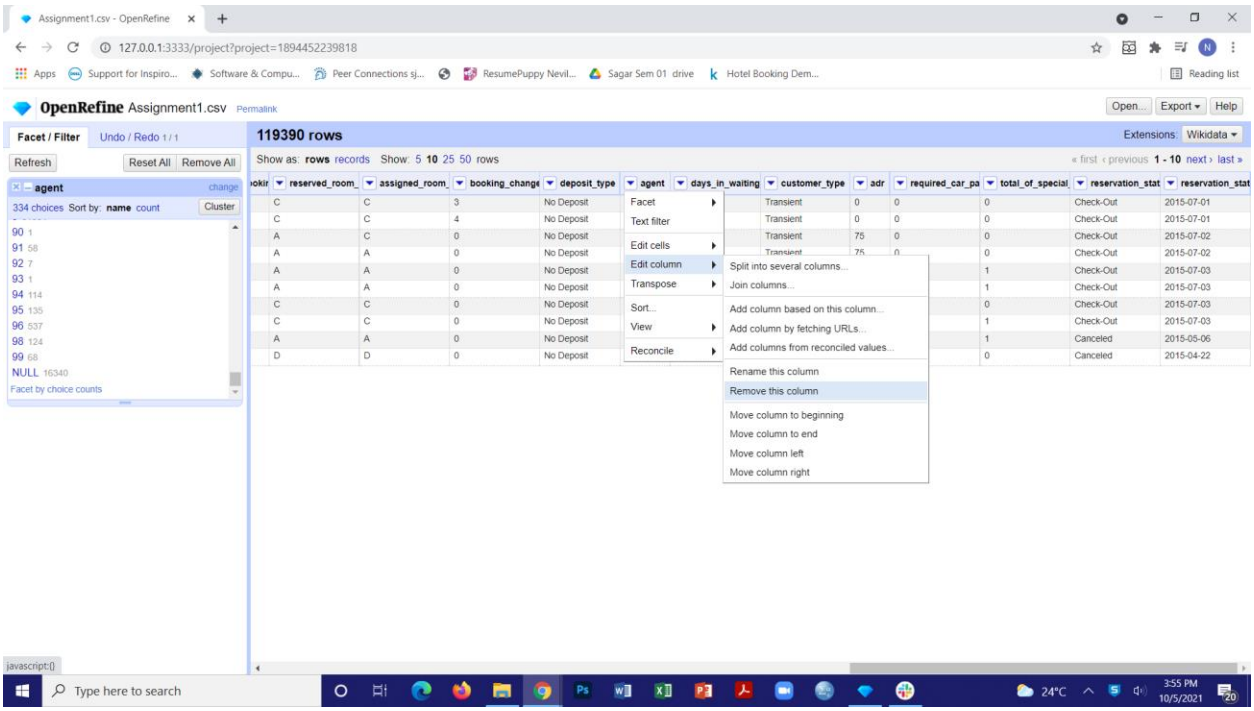
1) Column == Company



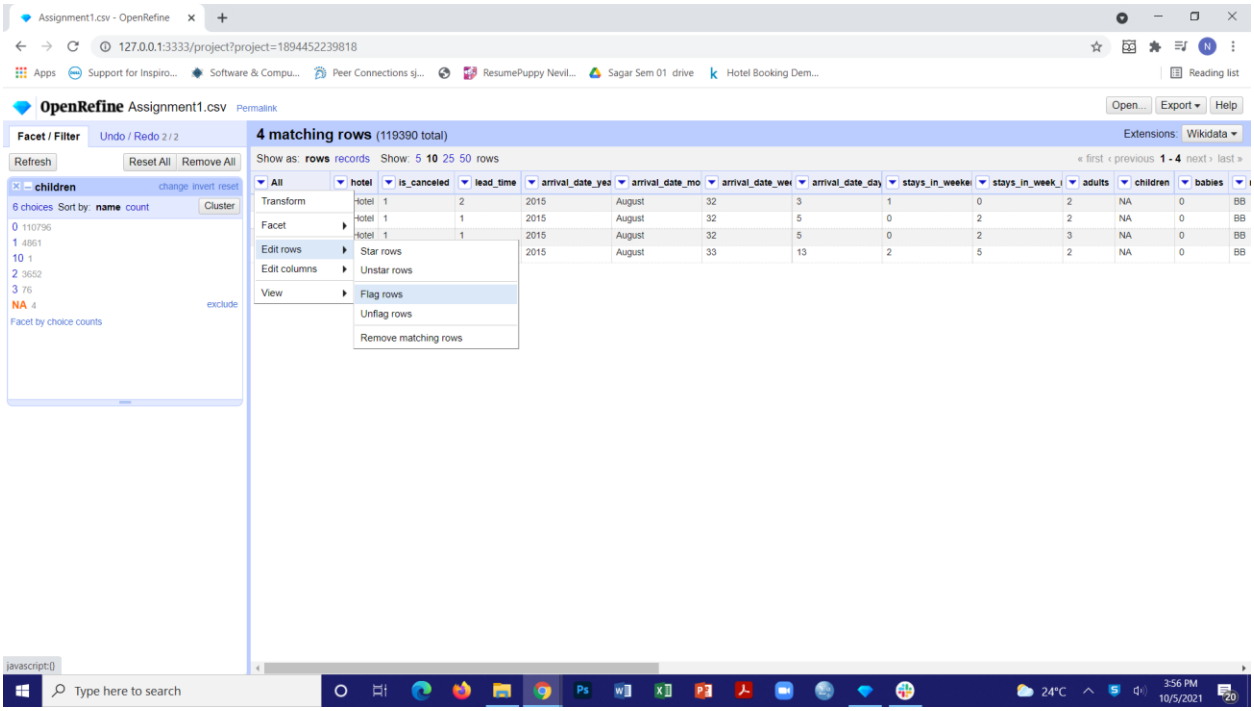

As we can see from the above image, about 112k values in this column are Null , hence , we will remove this column

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.




3) Column == ‘Children’

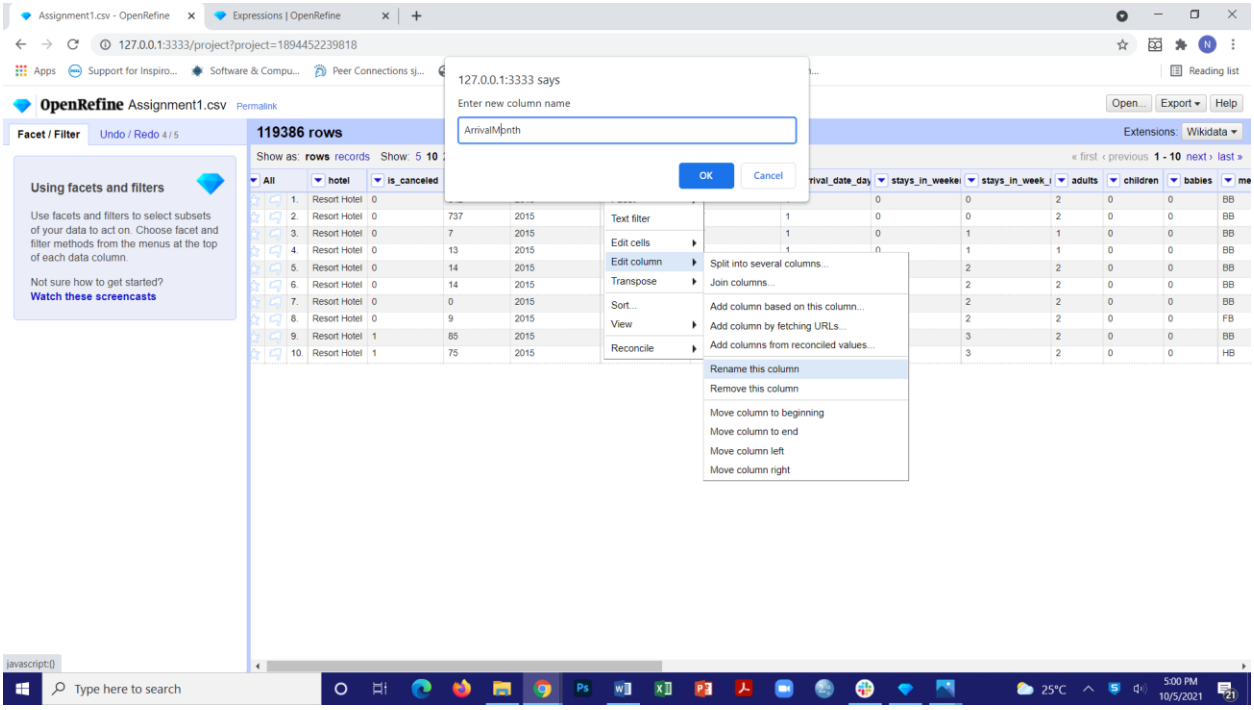


As seen from the image above, there are 4 records in the 'Children' column which has null values. Hence, we will remove the 4 records.

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.

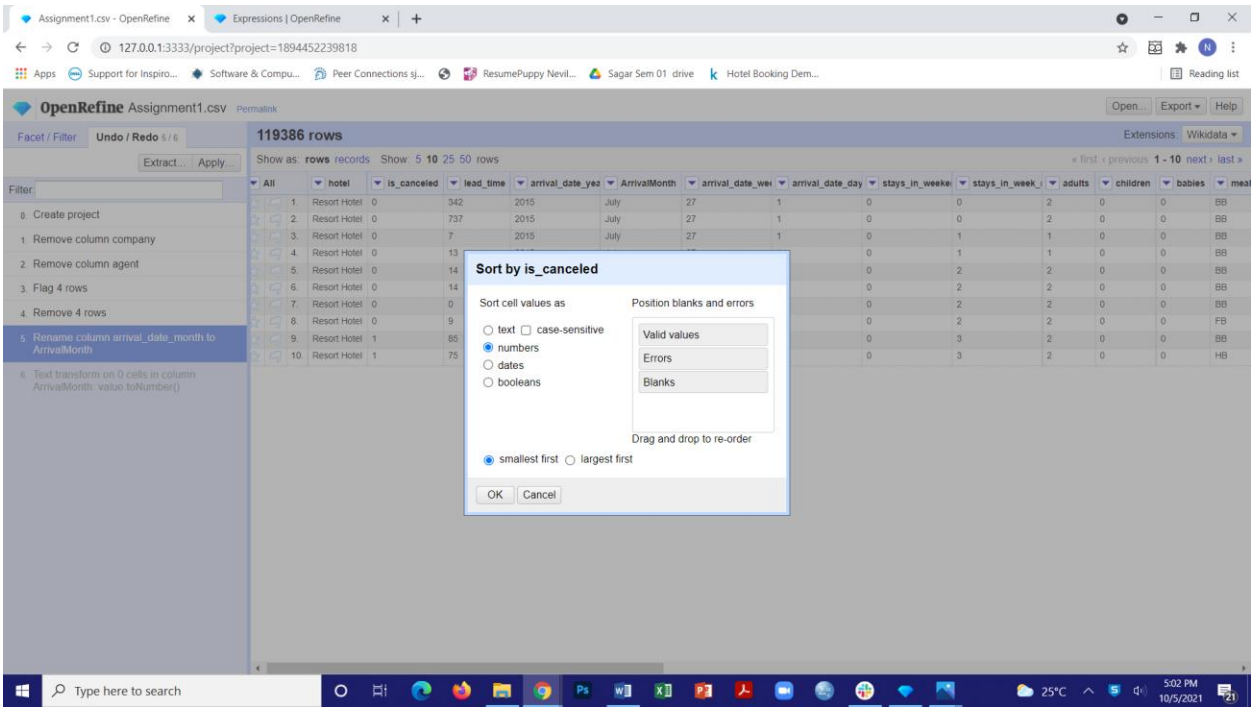
Step 2:- A Dialog box will be opened at the top, write the new name of the column and Click 'ok'.

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 one.

Step 4:- Select ‘Ok’.

The column ‘is_cancelled’ is now sorted in ascending order. In this way, we can sort a column in OpenRefine.

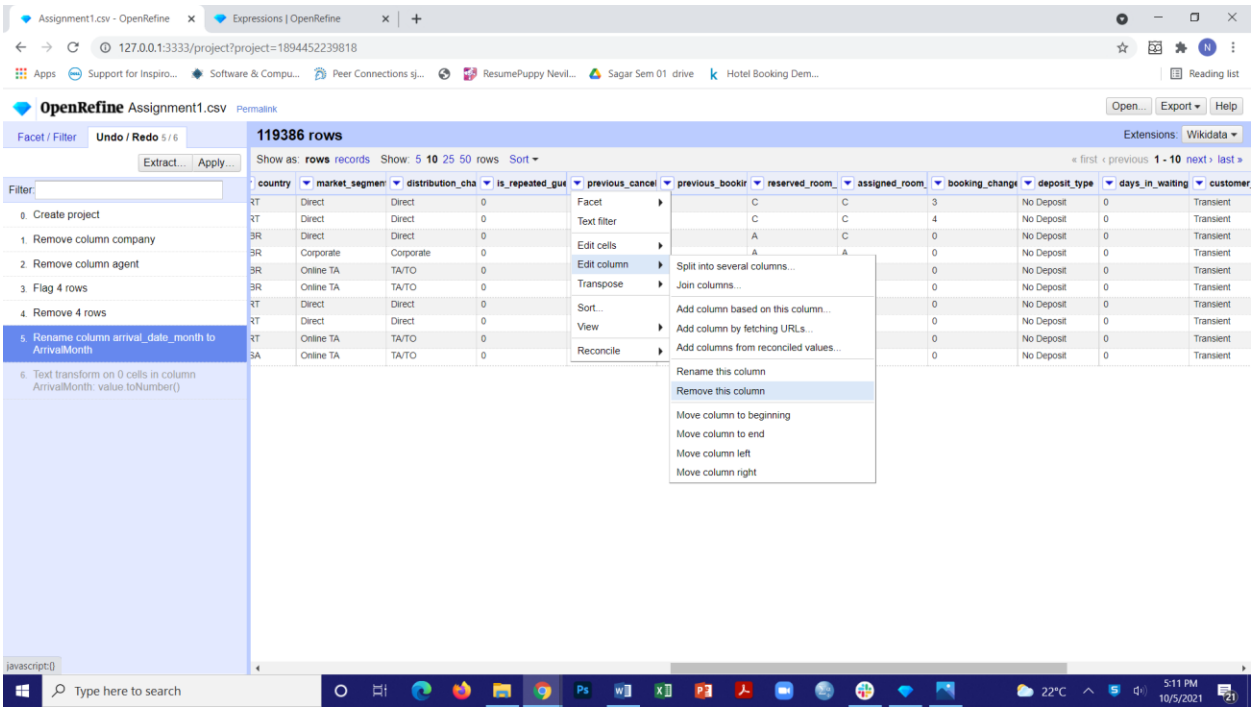
## Deleting a Column:-

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.

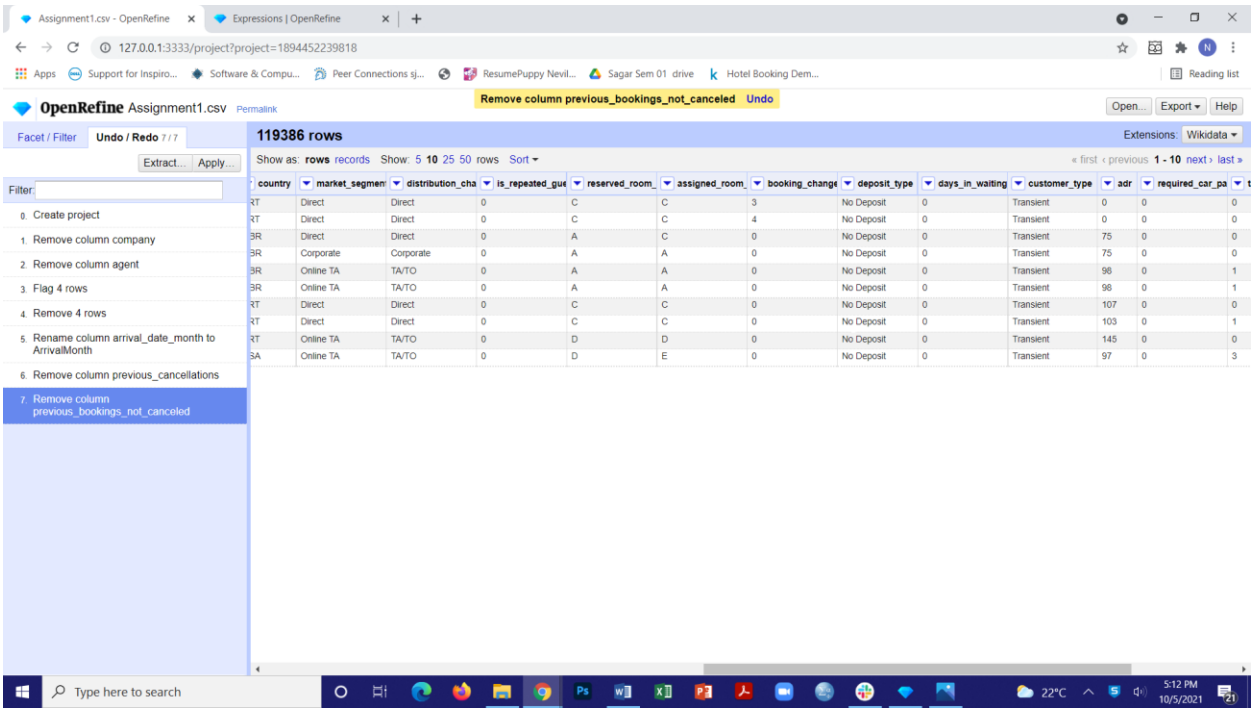
Steps to remove a Column:-

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




Now, repeat the above 2 steps for the column ‘previous_bookings’ and this column will also be deleted thereafter.




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

3) https://thomaspadilla.org/dataprep/

4)https://www.youtube.com/watch?v=wGVtycv3SS0&ab_channel=UniversityofIdahoLibraryDigitalInitiatives