1. Briefly describe the selected fields (15 points)
Artist Name : This field contains the complete name of the artist.

Art Dimension : This field contains art dimensions. It contains dimensions in both feet and centimeters.

Art On View : This fields indicates whether the art is currently on display or not.

Art Date : This field indicate the year of creation of the art work.

2. Explain the cleaning operations performed for each field

Cleaning procedure for Artist Name :

This field was selected to illustrate the splitting column functionality available in OpenRefine. The original field is of the format “First_Name” “Second_Name”. Using “ “ as delimiter, the column is separated into two new columns named First Name and Second Name.

Cleaning procedure for Art on View

This field is chosen to illustrate the usage of facet functionality. There are two categories of arts in the database. Art works can be categorised into arts which are currently on display and the others which are not. This field can take value “on view:” -> indicates that the art is presently on view or “not on view” -> indicates that the art is not on view.

- Now using facet and appropriate regular expression a new column was created with values “true” for arts on view and “false” for arts not on view.
- The “match” string function is used for dividing the arts into two categories.

Cleaning procedure for Artist Dimension :

- The field has the dimensions in two metrics(feet and centimeters), which I thought are kind of redundant.
- So I separated the values into two different columns, one for feet and one more centimeters.
- Finally, I retained the column with feet values.
- Now, I divided the value into length, width and height columns.
- *Using replace(value,/[A-Za-z]+/,””)* regular expressions all the letters were removed.
- Now for art works which are “not in view” the dimensions field is empty. I wanted to fill “Not available” string for all the cells with empty values. But the “fill down” option which is precisely for this, is some how not functioning as expected. A conditional operator is used to accomplish this task.

if(isBlank(value),'Not Avaiable',value)

Cleaning procedure for Art Date:

This field contains the year in which the art was created. Most of the fields are in the same format of “year”, but few values of have a different format of “year-year”. So, to maintain consistency across the column, only the first year is retained throughout the column.

Example record:

Before Cleaning:

<Art>
<Date value="1981-82"/>
<Title value="Untitled"/>
<Artist value="Thomas Nozkowski"/>
<Medium value="Oil on canvas board"/>
<Dimension value="15 7/8 x 20" (40.2 x 50.8 cm)"/>
<Image value="http://www.moma.org/collection_images/resized/933/w500h420/CRI_208933.jpg"/>
<Status value="On view"/>
<Credit value="Acquired through the Richard D. Brixey Bequest"/>
<Category value="Painting"/>
</Art>

After Cleaning

{
"Art - Date - value 1" : 1981,
"Art - Title - value" : "Untitled",
"Artist First Name" : "Thomas",
"Artist Last Name" : "Nozkowski",
"On view" : true,
"Art - Credit - value" : "Acquired through the Richard D. Brixey Bequest",
"Art - Medium - value" : "Oil on canvas board",
"Art - Category - value" : "Painting",
"Art - Dimension - Length" : "15 7/8 ",
"Art - Dimension - Width" : 20,
"Art - Dimension - Height" : "Not Applicable",
"Art - Image - value" : "http://www.moma.org/collection_images/resized/933/w500h420/CRI_208933.jpg"
}

The entire database is exported into two different json datasets. One for art works which are presently on view and another one for the art works which are not.

Submission :
The submission contains three files :
assignment5_true.txt->the dataset containing all the art works which are presently in view
assignment5_false.txt->the dataset containing all the art works which are not in view.
assignment5.pdf->description about the assignment.