CS548 - Homework 4

Name: Apurv Upasani USC ID: 4839-3102-30 Email: aupasani@usc.edu

1. "I, _Apurv Upasani_, declare that the submitted work is original and adheres to all University policies and acknowledge the consequences that may result from a violation of those rules"

2. Following 3 fields were selected from the dataset:

   i) Artist Name
   ii) Painting Date
   iii) Painting Material

   In addition to these, basic cleaning was performed on Painting name, Artist Birth/Death dates and Artist Birth/Death country and place

   Artist Name

   - Initially, all the artist names were trimmed for leading and trailing whitespaces.
   - Some fields were manually cleaned using Text Facets (see section 3)
   - Finally, Artist Name was split into 2 fields – Artist First Name and Artist Last Name

   Painting Date

   - Some painting dates contained only finish date while many contained both start and end date
   - Initially, performed manual cleaning on certain types of dates (see section 3)
   - After that, dates having format (from –to) were manipulated using the following expression:

```
if(value.split("–").length()>1,value.split("–")[0]+"–"+(value.substring(0,2)+(value.split("–")[1])),value)
```

   - Painting dates were then separated as Painting start date and painting end date based on the hyphen (–).
   - Painting end dates having null values after separation were changed to N/A using Text Facet
   - Finally, fields having N/A were replaced with start date using following expression:

```
if(value=="N/A",cells["Painting_Start_Date"].value,value)
```

Painting Material:

- Initially , all the fields having blank values were replaced with “Not Available” using Text Facet.
- After that, the cluster mechanism was used to merge materials having similar names (see section 3).
- Finally, erroneous records were updated manually using Text Facets. Wherever the record was beyond repair, it was replaced with “Not Available”.

3. Examples

Field	Example	Before	After
Artist Name	1	Afro (Afro Basadela)	• After manual cleaning – Afro Basadela • After column splitting – Artist_First_Name ⟶ Afro – Artist_Last_Name ⟶ Basadela
Painting Date	1	ca. 1960	• After manual cleaning – 1960
	2	1959-60	• After evaluation of expression 1 – 1959 -1960 • After column splitting – Painting_Start_Date ⟶ 1959 – Painting_End_Date ⟶ 1960
	3	1960	• After evaluation of expression 1 – 1960 • After column splitting – Painting_Start_Date ⟶ 1960 – Painting_End_Date ⟶ null • After updating end date column – Painting_Start_Date ⟶ 1960 – Painting_End_Date ⟶ N/A • After evaluation of expression 2 – Painting_Start_Date ⟶ 1960 – Painting_End_Date ⟶ 1960
Painting Material	1	Oil and Acrylic on Canvas Acrylic and oil on canvas Oil and acrylic on canvas	• After clustering – Oil and acrylic on canvas
	2	1959-1960	• Invalid painting material (Manual cleaning) – After cleaning ⟶ Not Available