Report Name: INF558 Homework 4 submission

Submitted By: Anand Prashar

Date: 04 October 2017

1. What are the fields in your data that require cleaning and what kind of data do these fields contain?

Ans: I downloaded some messy data from internet as my project did not have any. This is a tabular data on Africa'a provincial and district hospitals. My subset has 5 fields: Name, Owner, Category, District, Province.

3 fields needed cleaning:

Name (type: string data, some Unicode characters in hospital names, many whitespaces)

Owner (type: string data, different names for same entities in many fields)

Category (type: string data, uneven representation of letter casing, null cells)

2. Describe which cleaning operations should be performed on these fields?

Ans:

On Name field-

1. Need to remove unnecessary numbering labels .
2. Whitespaces varies in between words too. Need to have at max 1 whitespace in between.
3. Some names had weird characters. Need to map to Unicode to correct their state.

On Owner field-

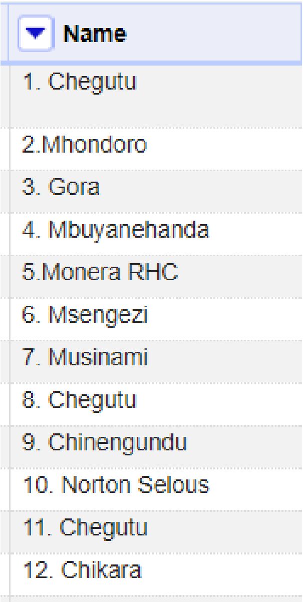
1. Same words had . or abbreviation causing to show different instances in facets. Need clustering to resolve this.
2. In some cases, entries from Owner category has spread into adjacent column. Need to fix that with flagging and reference correction.

On Category field-

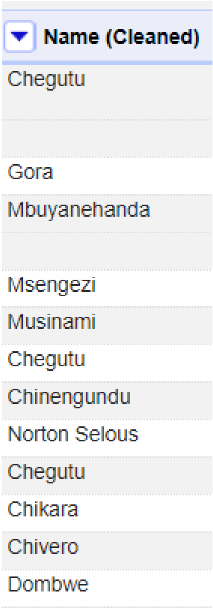
1. Same words are represented in uppercase and titlecases. Need to cluster them, thus reducing facet classes.
2. There are many null cells, need to at least replace with uniform whitespace.

Screenshots

Name field (Before and After)

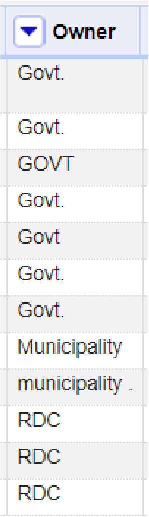


Name
1. Chegutu
2.Mhondoro
3. Gora
4. Mbuyanehanda
5.Monera RHC
6. Msengezi
7. Musinami
8. Chegutu
9. Chinengundu
10. Norton Selous
11. Chegutu
12. Chikara

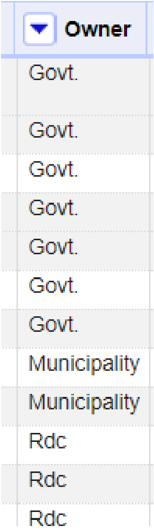


Name (Cleaned)
Chegutu
Gora
Mbuyanehanda
Msengezi
Musinami
Chegutu
Chinengundu
Norton Selous
Chegutu
Chikara
Chivero
Dombwe

Owner field (Before and After)

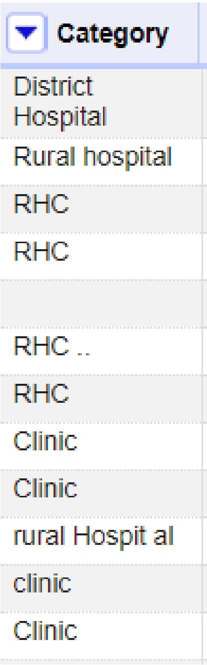


Owner
Govt.
Govt.
GOVT
Govt.
Govt
Govt.
Govt.
Municipality
municipality .
RDC
RDC
RDC

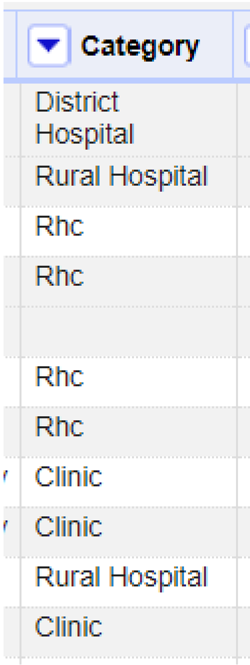


Owner
Govt.
Govt.
Govt.
Govt.
Govt.
Govt.
Govt.
Municipality
Municipality
Rdc
Rdc
Rdc

Category field (Before and After)



Category
District Hospital
Rural hospital
RHC
RHC
RHC ..
RHC
Clinic
Clinic
rural Hospit al
clinic
Clinic



Category
District Hospital
Rural Hospital
Rhc
Rhc
Rhc
Rhc
Clinic
Clinic
Rural Hospital
Clinic