CSC 220: Exploring Global Health Crisis Data  
FALL 2022 - Week 10

Research Assistant: Joori Almarzouki

jalmarzouki@uri.edu

horizontal line

# 9 November 2022

Data Cleaning

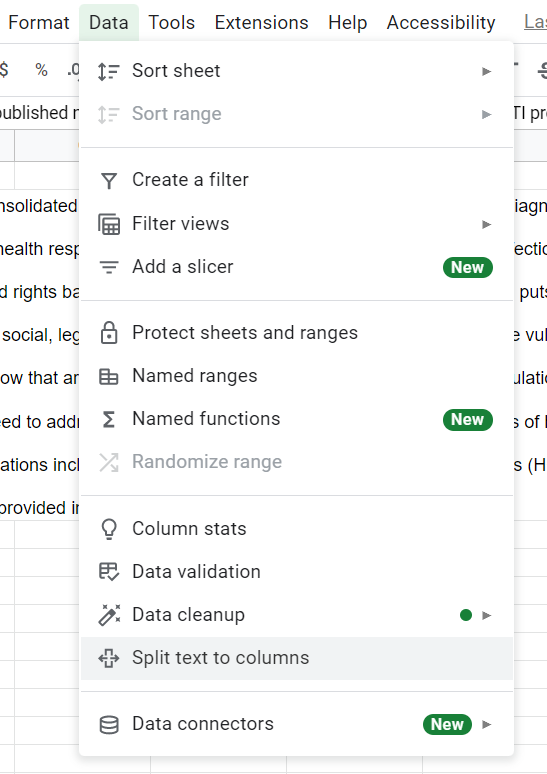
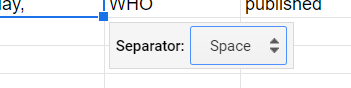
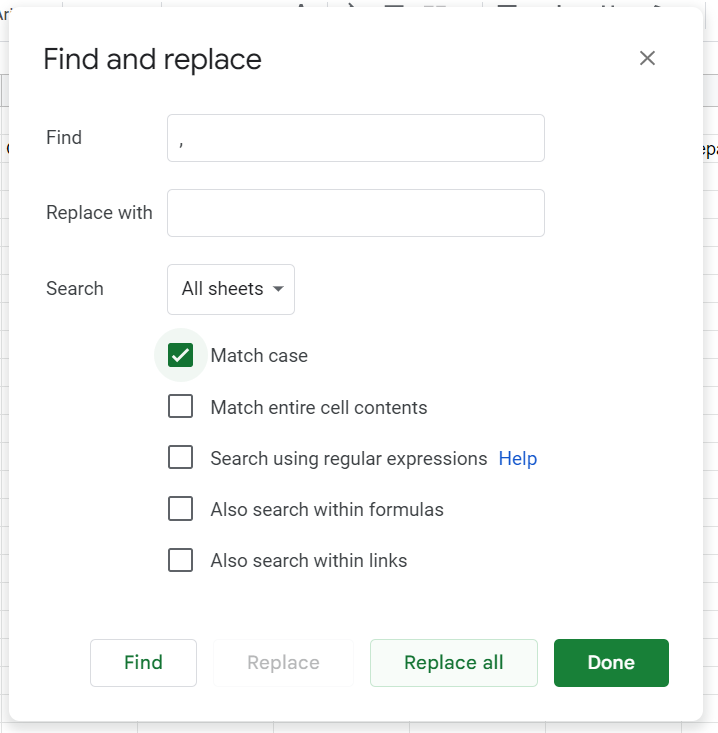
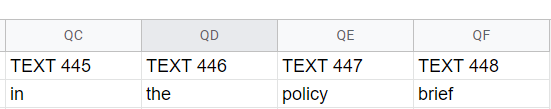
Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.

**Text Preprocessing Using Google sheets Tableau Desktop:**

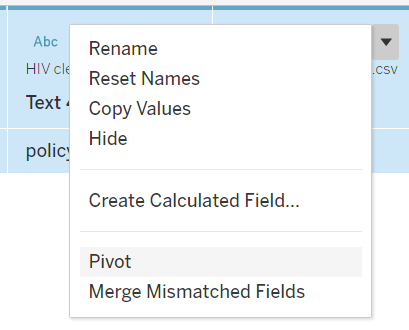
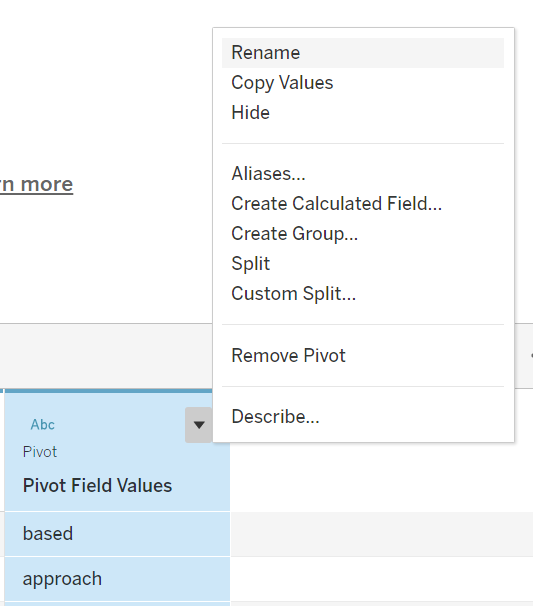
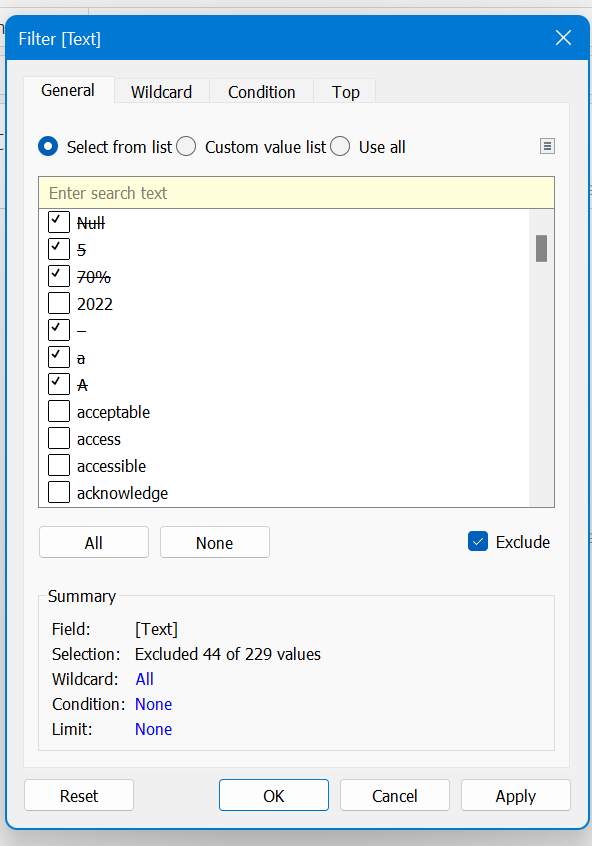
* **In Google Sheets:**

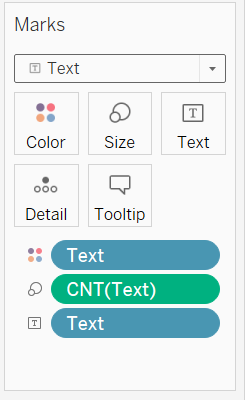
1. Download the [HIV text file](https://docs.google.com/spreadsheets/d/19XvU1zSWWTO0fF9L1tgaVoyq4jsbaDCkheuiJMexCQY/edit?usp=sharing).

* This was created by creating a header called Text, then copying [this article](https://www.who.int/news/item/29-07-2022-who-publishes-new-guidelines-on-hiv--hepatitis-and-sti-for-key-populations) and pasting it into one cell under the Text column.

1. We want to split the text, so that each word becomes a feature.
2. Click on **Data**, then **Split text to columns** and choose the **separator** to be spaces.
3. Now we will do some simple cleaning by removing functions. **Control + F to** find special characters (i.e , . ; : ( ) “ ” ) and replace **them with a space**. Then, check **match case** and **replace all.** 
4. Now we will add a header to each feature (word) by writing TEXT 1 in the original column and dragging it all the way to the last word and will see that we have 448 words.
5. Download the clean data as a CSV file.

* **In Tableau Desktop:**

1. Open the CSV file in Tableau Desktop.
2. Select the entire data set by selecting the first column, holding shift, scrolling to the end then click the last column then right-clicking on **Pivot**. 
3. Rename “Pivot Field Values” to “Text”.
4. Create a new sheet and **filter Text** by **Excluding** Null values and values that do not add much information to the text (like articles, prepositions, pronouns, conjunctions, etc). Examples of a few stop words in English are “the”, “a”, “an”, “so”, “what”. 



1. Drop “Text“ to **Text** under Marks.
2. Drop “Text“ to **size** under Marks, then select the measure to **count**.
3. Drop “Text“ to **Color** under Marks.
4. Change from Automatic to Text (from the drop down menu).
5. The results should be something like this.

