**MODULE 4**

**DEALING WITH MESSY DATA**

**(Using Excel)**

**By Aggrey Mutimba – Data Researcher**

**Date: 12th June, 2014**

**DATA JOURNALISM**

Data Visualization

Publishing

Story telling

Data Analysis

Data Cleaning

Data Scrapping

**INTRODUCTION**

Before you can analyze the data, you often need to clean it up. Especially when you have imported the data from a database, text file, or a Web page i.e. (xml, \*.dta, \*.csv, tab, sav, defined formats, dumps.sql).

Data that is not clean could have the following:

1. Missing data.
2. Spaces before and after values.
3. Outliers - data that is beyond rage for example, age of a human being recorded as 879 years.
4. Inconsistency.
5. Unknown characters due to different encoding i.e. UTF-8.

**CONTENTS**

In this lesson, we are going to focus on the following;

1. Removing duplicate rows
2. Finding and replacing text
3. Removing spaces and nonprinting characters from text
4. Fixing numbers and number signs
5. Transforming and rearranging columns and rows
6. What Open Refine can do.
7. Cleaning big data (Crowdsourcing)

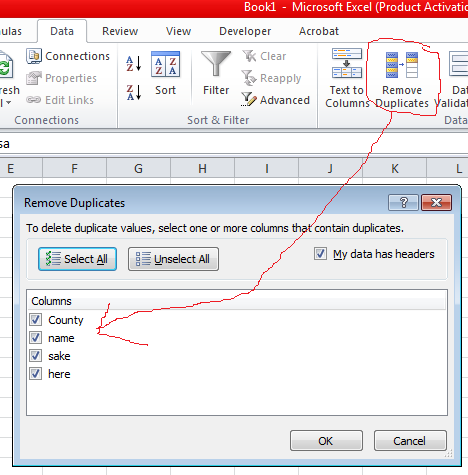
**Always create back-up before you embark on data cleaning**

. [Download data from here. KE\_budget\_Training.xlsx](https://docs.google.com/spreadsheets/d/1MXa6Z2uAvvA8q8K3XIZwfn1b537exj520MMGodUyN7M/edit?usp=sharing)

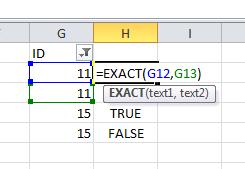
**REMOVE DUPLICATE ROWS**

Duplicate rows are a common problem when you import data. It is a good idea to filter for unique values first to confirm that the results are what you want before you remove duplicate values.

1. Using KE\_budget\_Training.xlsx that you had already downloaded.
2. Click in any cell in the worksheet.
3. Click on “remove duplicates” button as shown below.
4. Check columns that may have duplicates.

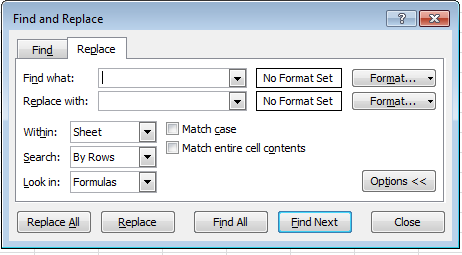


1. You could also use, EXACT function to display the duplicates in a column as shown. *(In the data KE\_budget\_Training.xlsx, find duplicates.)*



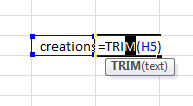
**FIND AND REPLACE**

1. The most common way to find and replace is to press ***Ctrl + F*** on your computer.



**REMOVING SPACES AND NONPRINTING CHARACTERS FROM TEXT**

TRIM, CLEAN



TRIM function removes spaces before and after the words but not between the words.

## FIXING NUMBERS AND NUMBER SIGNS

One of the main issues with numbers that may require you to clean the data is when the number was inadvertently imported as text.

How to change it

## 

## Changing the decimal places

## Highlight the column you want to increase or reduce the decimal places and click on the buttons shown.

## 

## TRANSFORMING AND REARRANGING COLUMNS AND ROWS

Most of the analysis and formatting features in Office Excel assume that the data exists in a single, flat two-dimensional table. Sometimes you may want to make the rows become columns, and the columns become rows.

## TRANSPOSE,

## 

## =TRANSPOSE (ARRAY)

**CROWDSOURCING**

Asking crowd to clean existing data for you;

Always provide clear instructions otherwise data will be messed up! Examples of where this method is being used.

* [Free the Files: Help ProPublica Unlock Political Ad Spending](https://projects.propublica.org/free-the-files/) (ProPublica)
* [Crowdsourcing app for data collection and cleaning](https://projects.propublica.org/free-the-files/sessions/new) - Login required (ProPublica)

**USING GOOGLE REFINE TO CLEAN DATA**