DATA CLEANING TECHNIQUES

Data forms the backbone of any data analytics you do. Regarding data, there are many things to go wrong – be it the construction, arrangement, formatting, spellings, duplication, extra spaces, and so on. To perform the data analytics properly we need various data cleaning techniques so that our data is ready for analysis. It has commonly said that,

"Data scientists spend 80% of their time cleaning and manipulating data and only 20% of their time actually analyzing it."

Thus, it is important to grow accustomed to the process of data cleaning and all of the tools that relates to this process. This post introduces data cleansing techniques in Excel.

This post covers the following data cleaning steps in Excel along with data cleansing examples:

1. Get Rid of Extra Spaces
2. Select and Treat All Blank Cells
3. Convert Numbers Stored as Text into Numbers
4. Remove Duplicates
5. Highlight Errors
6. Change Text to Lower/Upper/Proper Case
7. Spell Check
8. Delete all Formatting

What is Data Cleaning?

Data cleansing or data cleaning is the process of identifying and removing (or correcting) inaccurate records from a dataset, table, or database and refers to recognizing unfinished, unreliable, inaccurate or non-relevant parts of the data and then restoring, remodeling, or removing the dirty or crude data. Data cleaning might performed as batch processing through scripting or interactively with data wrangling tools.

After cleaning, a dataset should be uniform with other related datasets in the operation. The discrepancies identified or eliminated may have be caused by user entry mistakes, by corruption in storage or transmission, or by various data dictionary descriptions of similar items in various stores.

DATA CLEANING TECHNIQUES

Are Data Cleaning Techniques Essential?

Data cleaning is not only an essential part of the data science process – it is also the most time-consuming part. As the New York Times reported in a 2014 article called “For Big-Data Scientists, ‘Janitor Work’ Is Key Hurdle to Insights”,

“Data scientists … spend from 50 per cent to 80 per cent of their time mired in this more mundane labour of collecting and preparing unruly digital data, before it can be explored for useful nuggets.”

Unfortunately, Data Cleaning is generally not spoken about in the media nor is it taught in most intro data science courses because it is not as important as training a neural network or identifying images but to perform those things data cleaning plays a very important role. Without the data cleaning the neural networks and image identification modules will not be as efficient as we want them to be.

With the rise of big data, data cleaning has become more important than ever before. Every industry – banking, healthcare, retail, hospitality, education – is now navigating in a large ocean of data. In addition, as the data pool is getting bigger, the variables of things going wrong too are getting larger. Each fault becomes difficult to find when you cannot just look at the whole dataset in a spreadsheet on your computer. In fact, this could be true for a variety of reasons.

Data Cleansing Examples and Data Cleaning Methods in Excel

In this post, I will show you various ways to clean data in Excel.

1. Get Rid of Extra Spaces

Here I have the text **Welcome to Digital Vidya** written in four different ways.

Welcome to digital vidya

Welcome   to digital vidya

Welcome   to digital vidya

Welcome   to   digital   vidya

First one is the regular way with only one space between words, in the second case I have more than one space between words, in a third case I have some leading spaces along with a couple of spaces between words and in the fourth case I have trailing spaces, you can see there are a couple of space after the last word. Now, this could typically be the case if you get this data from a colleague or you get it from a text file or imported from a database. So to clean this data and get rid of these extra spaces you can use the function *trim*.

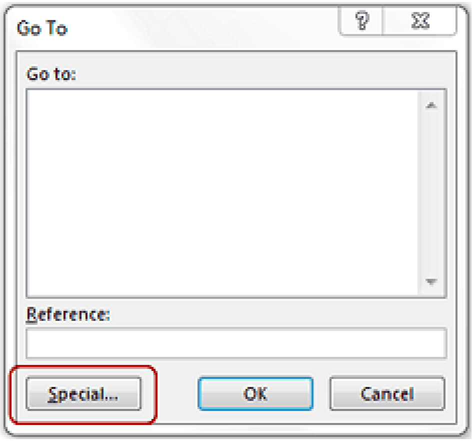
Syntax: =TRIM(Text)

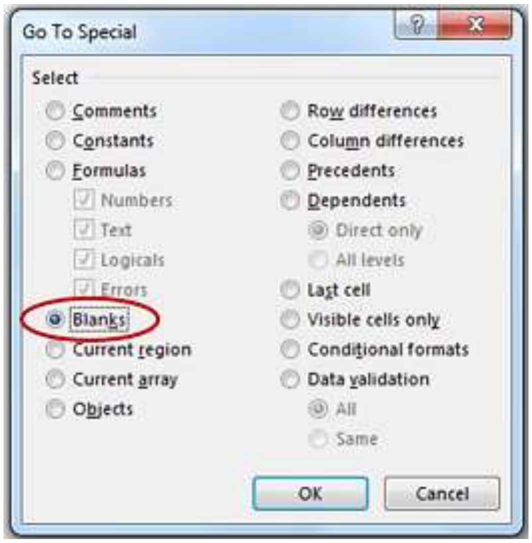
Trim function takes one single argument which could either be the text which you type manually or it could be the cell reference, in this case, I will take the cell reference *A1* and what this function does is it would remove all the leading spaces and trailing spaces and extra spaces between words except one single space that is allowed. So if I drag this down you would see that it has corrected all these texts. It has removed the extra space here between welcome into it has removed the leading spaces and trailing spaces.

2. Select and Treat All Blank Cells

If you just need to use the text you can convert it into values by using *paste special* I have student names here and their marks in three subjects. You can see that there are gaps in this dataset which could be because the student could not appear in the exam. now you may not want to leave this data set with blanks, you may want to type *not appear* in all these cells which are blank. So to do that you can either go and select each cell manually and type *not appear.* But if you have a huge data set that because this could be very tiresome. So to do it at one go,

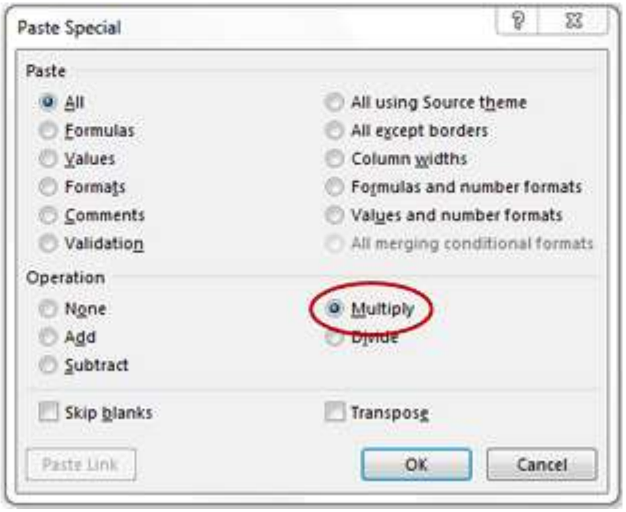
- Select the entire data set,
- Go to *find and select* and select this option *Go to Special* this opens the go-to special dialog box. You can also use the keyboard shortcut *F5* and when you do this it opens the go-to *dialog box* here you have *special button*, click on it and it again opens it equal to special dialogue box.
- Click *blanks* and click *okay*, this would select all the blank cells in your data set at the same time.





- So now you have these cells in grey and the first cell is in white because this is the active cell so to type *not appear* in all these cells just start typing *not appear* and hit *ctrl+enter* and as soon as you hit *ctrl+enter* this gets entered in all the cells.

DATA CLEANING TECHNIQUES



Paste Special

Paste

● All
○ Formulas
○ Values
○ Formats
○ Comments
○ Validation
○ All using Source theme
○ All except borders
○ Column widths
○ Formulas and number formats
○ Values and number formats
○ All merging conditional formats

Operation

○ None
○ Add
○ Subtract
● Multiply
○ Divide

□ Skip blanks
□ Transpose

Convert Numbers Stored as Text into Numbers

Here I have this number entered in three different ways,

```
   123
123
'123
```

In the first case, it is a number as you can see it is aligned to the right of the cell numbers are always aligned to the right while text gets aligned to the left of a cell and in the other two cases, you can see these are text format because these are aligned to the left. now to convert all these three back into numbers. The first one is already a number but to convert these two back into a number there are two ways to do it. The first one is I would go to the formatting box and I would type *general* and when I hit *enter* the second one gets converted back into a number because in this case it was merely in the text format but the third case is a little more difficult because it has been entered by using a leading apostrophe and a lot of people do this, a lot of people enter numbers for starting with an apostrophe so that it gets converted into a text and this could create some problems. So to take care of this let me delete this to take care of this. A very foolproof method is type in any of the blank cells go to the cell and copy this now select these cells go to *paste* -> *paste special* and this opens the paste special dialog box. Here you have *operation* category within this select *multiply* and click okay so what this does is this multiplies this number with one and any number multiplied by one is unchanged but this also takes care of the apostrophe so now all these three numbers get converted back into a number format.

4. Remove Duplicates

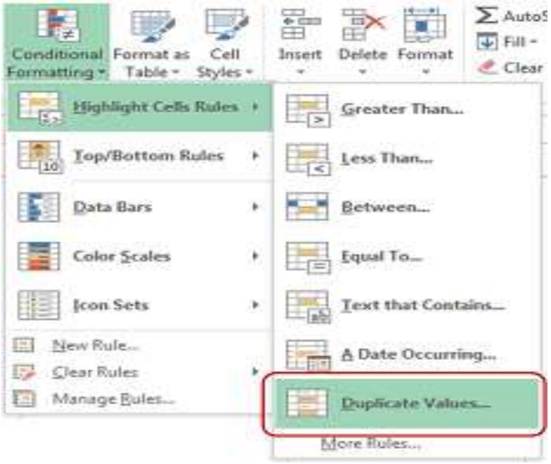
Here I have a data set of students and their marks in three subjects and there are duplicates in this data so you can see there is a duplicate for Bill and duplicate for Phil now if you want to remove these duplicate values there are two ways to do it first is using **conditional formatting**,

- So you can select the data set
- Go to home -> conditional formatting

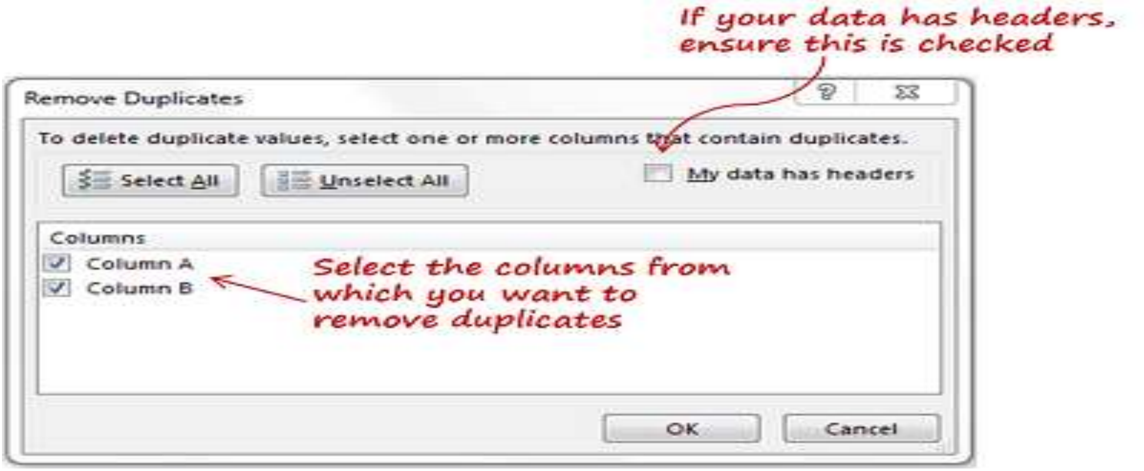
DATA CLEANING TECHNIQUES

- *Highlight cell rules* -> *duplicate values* and as soon as you select this it gives you the option to highlight duplicates and the formatting. I will keep the formatting as a red fill with dark red text and when I hit OK you can see that this has been highlighted and all those numbers and names that appear more than once it highlighted in red.

I can manually see that Phil repeats twice and Bill repeats twice so I can select this data and manually delete this.



The other way to remove duplicates is by selecting the entire data set going to data and here I have the option remove duplicates I click on this option and it opens the remove duplicates dialog box here make sure that if your data has headers which in this case it has this option is selected if this is not selected then this is also counted as a part of your data it should not be the case when you have selected this option these names are the names of the columns so I can see that there is a student column math column physics column in chemistry column I can then select ok to remove all those rows or all the data set which is duplicate but in this case it would not remove a number which repeats again rather the entire row has to be an exact duplicate so, for example, Jack and these three marks have to meet exactly the numbers and name here Jack and these three marks and if that is the case then this row gets deleted similar is the case with Jill and these numbers so now when I click OK it says one duplicate values found and removed unique values remain the reason being that in case of Jill you can see that the marks do not match so this entire row is not an exact match and hence it remains but since in case of Bill it was an exact match then the row was removed.



If your data has headers, ensure this is checked

To delete duplicate values, select one or more columns that contain duplicates.

□ My data has headers

Columns

☑ Column A

☑ Column B

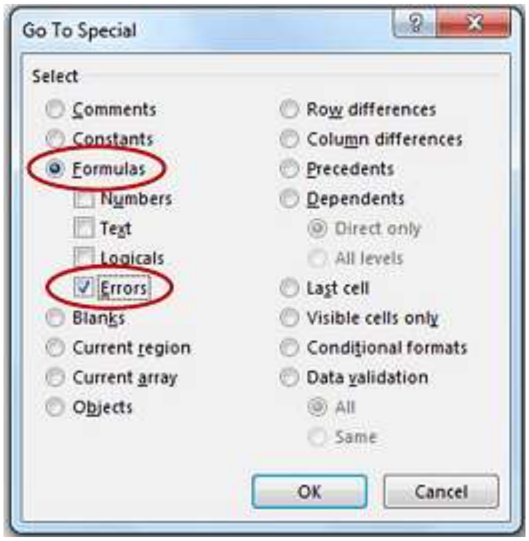
Select the columns from which you want to remove duplicates

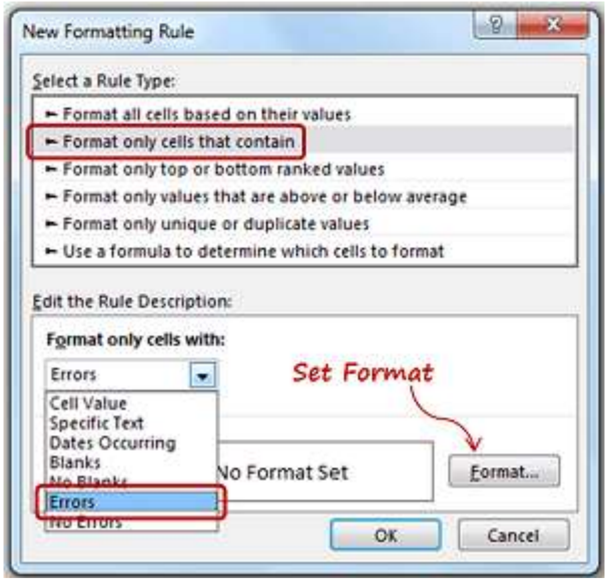
DATA CLEANING TECHNIQUES

5. Highlight Errors

Here I have a dataset for five companies. I have their revenue number for three years and net income numbers for three years and using these numbers I have calculated the net income margin which is net income by revenue. Now you can see that there are errors in data for Company X and Company Z, the reason being that there is no revenue number for these companies in and hence I get a division error because I try to divide their net income by nothing. Now this is a small data set and you can visually spot these errors but if you have a huge data set these errors could be difficult to spot so to do that you can use two methods first is using **conditional formatting,**

So select this entire dataset go-to *home -> conditional formatting* and select *new rule* within *new formatting rule dialog box* select *format only cells that contain* and from this drop-down select *errors* when you select errors you would get the option to format the cells which have error, in this case, let me select *red* and I click OK and as soon as I do this all the cells that have errors in it get highlighted in red




New Formatting Rule
Select a Rule Type:
► Format all cells based on their values
► Format only cells that contain
► Format only top or bottom ranked values
► Format only values that are above or below average
► Format only unique or duplicate values
► Use a formula to determine which cells to format
Edit the Rule Description:
Format only cells with:
Errors
Cell Value
Specific Text
Dates Occurring
Blanks
No Blanks
Errors
No Errors
Set Format
No Format Set
Format...
OK
Cancel

We *control Z* to go back the other way to do this would be to select those cells which have errors and you can do this by using the go-to special dialog box so to do that press ***F5*** this opens the go-to dialog box here you have the *special*button, click on it, this opens the *go-to special* dialog box. Here select *formulas*and within *formulas* as soon as you selected all these four options get available, deselect the first three options and only keep the *errors* option selected and now click *OK.* When you do this all those cells which have error in it get selected now you can manually either delete all these cells or type something like *not available*and hit *control enter* so that it gets entered in all the cells which have error in.

6. Change Text to Lower/Upper/Proper Case

Here I have names written in different ways you can see either it could be all caps, it could be all lowercase and in some cases, it's a mix-and-match of uppercase lowercase so to make it all consistent you can use one of these three formulas,

DATA CLEANING TECHNIQUES

SYNTAX:

LOWER() – Converts all text into Lower Case.

ex. mary jane

UPPER() – Converts all text into Upper Case.

ex. MARY JANE

PROPER() – Converts all Text into Proper Case.

ex. Mary Jane

LOWER() formula takes one argument, it could be either the text that you type in or you can use a cell reference in this case if I'll use the cell reference A1 and when I hit *ctrl enter* this gives me *mary jane* the name but all the alphabets have been converted into lowercase and when I drag this down this is the case for all these names all these names now look consistent in lowercase you may want these all in the uppercase so in that case you can use the formula *UPPER()* and you can see this these are all in uppercase now as I drag this down the most used way is proper case because it would keep the first alphabet of your name as in capital and the rest all would be in the lowercase may show you I will select *PROPER()* and I hit *ctrl enter* and you can see *M* of M*ary* and J of *Jane* is in caps and rest all the alphabets are in the lowercase and now I drag it down so these are three formulas that can very quickly make your text consistent this could be the case when you are sharing a worksheet or you get it from a text file where a lot of people enter it in different ways these formulas can quickly make these consistent.

7. Spell Check

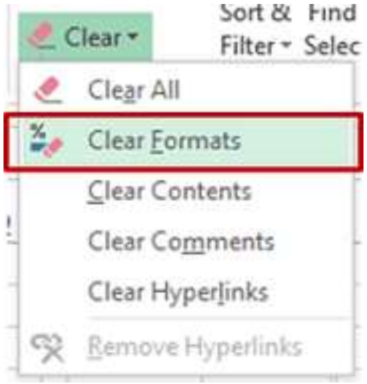
If you have huge data set and you want to only extract a part of it while Microsoft PowerPoint and Microsoft Word have a feature where it would underline if there are any errors grammatical errors or spelling errors Microsoft Excel does not have that feature however you can still a run spellcheck and correct these errors. So to do that select the data and press **F7** and when you do that it runs the spellcheck for you and it is the same thing that you see in Microsoft Word or PowerPoint it will show you the text that it thinks is a spelling error and it will show you the suggestions as well so you can change these and once it is done it will show you that spellcheck is complete and you are good to go.

8. Delete all Formatting

If you have a worksheet where there is a lot of formatting and you need to clear all the formatting, you can quickly do that by,

- Selecting the entire data
- Go to Home –> Clear –> Clear Formats
- You can also use *clear all* this would remove everything from your sheet including the content you can only clear the content would remain the formatting would remain intact you can clear the comments and the hyperlinks.

DATA CLEANING TECHNIQUES



Data Cleansing Tools

Here are some interesting tools relating to cleaning, analysis and modelling of data,

JASP – Open Source statistical software similar to SPSS with support of COS

Rattle – GUI for user-friendly machine learning with R

Rapid Miner – Another point and click machine learning package

Orange – Open Source GUI for user-friendly machine learning with Python

Talend data preparation – Data cleaning, preparation tool with smarts

Trifacta Wrangler – Data cleaning, preparation tool with match by example feature

They are all open source, or have free versions focusing on cleaning, analysing and modelling data.

Conclusion

Data cleaning is an inherent part of the data science process. In simple terms, you might divide this process down into four stages: collecting the data, cleaning the data, analyzing/modelling the data, and publishing the results to the relevant audience. If you try to skip the data cleaning steps, you will often run into problems getting the raw data to work with traditional tools for analysis in, say, R or Python.