Code For Good Data Cleaner User Guide

PREPARATION

So you want to check for errors in your data, but wait, before you start, you need to make sure your computer is ready!

First, you need to have Python 3 installed on your computer. These instructions will be for Windows, but it is similar for Mac. Please note that during the installation, there will be a screen where you will have a few check boxes with options for the installation. At the bottom of this screen with options (you may have to scroll down or something), there will be an option to 'add to PATH'. Check this box!! Otherwise, you'll have to Google 'how to add Python to PATH' and do it yourself.

The link to the Python 3 download is here (make sure to download from the box near the top that says 'Download Python 3.x.x', not the one that says '2.x.x'):

https://www.python.org/downloads/

You can make sure it works by opening up Command Prompt (open the search bar at the bottom left of your desktop, type 'cmd' and hit enter). Once you have Command Prompt open, type 'python' and hit enter. It should then say 'Python 3.x.x' (depending on the version of Python 3 you download) and a bunch of other stuff. That means it works, and you can close Command Prompt. If you get an error, it didn't work, and you should seek help!

Next, we have to install a few extra libraries for Python to work with the excel files! Don't worry, if you just installed Python and tested it as above, it's a breeze. We've included a file called 'install_libraries.bat'. Just double click this file to run it, and let it do its thing. When the popup Command Prompt window disappears, it's done. If you aren't on Windows, this may not work, so you have to open up this file as a text file, and run each line in the file individually on your command line.

Take a deep breath, and be proud, because you'll never have to do that again on this computer! If you have other Python versions installed or have tinkered with your installation, these preparatory steps might be a little different, but you probably have a rudimentary understanding of how it translates over.

USE

Now we're ready to use the data cleaner tool! Please make sure the files 'errorTableMaker.py', 'columnTests.py', 'TestSettings.xlsx', and the folder 'surveyData' are all together in the same folder. They're going to look for each other in the folder that they're in, so they need to be kept together!

First, we need to prepare the data to be imported. Open up the .xls Excel file of the data you want to check for errors in Excel, click 'save as', and save it in the 'surveyData' folder. Maybe you want to give it a shorter name or just leave the name as it is, it's up to you. You can put as many files here as you want;

it doesn't have to be one at a time. For some reason (maybe from how the raw data is originally exported to .xls), the code thinks that the original files are corrupt, so that's why we need to re-save the files with Excel.

Next, all we have to do is run the code! Navigate to the folder with the file 'errorTableMaker.py' and double-click it! If you just have Python 3 installed, this should be all you need to do. If you have other versions of Python (you know who you are if so), run it with the proper version from command line to ensure that it works right.

Now, when you open the 'surveyData' folder, you will see all the files you saved there (those haven't changed), and then another copy of the file with '_errors' added to the end of the name. This is the spreadsheet with the errors for that survey data. When you open it, you'll see some of the rows from the original file. There is an additional column at the end of the table, and the values for this column are the names of the fields that our code thinks is wrong for that row. These files are saved as '.xlsx', which is a newer file type than '.xls', so if you're using a very old version of Excel, you may need to install some updates/plugins in order to open it. Look at the rows, and use the information however you wish, as the code does NOT edit the original data. It only reports in a new file what it thinks is wrong.

FUTURE EDITS

This code will ONLY work as long as the survey column titles do NOT change at ALL from the way they were originally sent to us. If they do, there are some things you will need to change. Or maybe you want to only test for a certain kind of error for a little while, or just add a new survey field. If any of those apply, there's a way to do that!

Here's where that 'TestSettings.xlsx' table comes in. If you open it, you'll see a big Excel workbook with survey field titles (as they show up in the raw data) for rows and test names for titles. There are green and red cells. A green cell is non-empty, signifying that the field in that row will be tested by the test in that column. A red cell is empty, signifying that the field in that row will not be tested by the test in that column. To turn off the test for a field, click the green cell and hit backspace, deleting the contents. When you click away and the cell turns red, you know it worked! If you want to enable a test, click a red cell, type something (for instance 'Y'), and click away. If it turns green, then you know that test is enabled. Remember to save the file when you're done! And if the survey field titles change, you need to change them to match in 'TestSettings.xlsx'.

For more information on this file, see the second sheet of the file, called 'Instructions.'

For information on writing code to make your own tests, see the 'README.txt' file, which contains more technical instructions.

Enjoy!