



Script Walkthrough





What each script does

DIFF (Two files)

- Can compare two Excel files
- Highlights differences: Green = new, Red = changed, Grey = removed

Format Check (One file)

- Setup: (only needs to be run once) Uses a sample file as a reference for formatting - Creates json file
- Running: Analyzes given file with json as reference for formatting

Number Check (One file)

- Setup: Uses 3x Standard Deviation from numeric column as default values (Covers ~99.7%)
- Running: Compares values in numeric column with range created by setup



**Be sure to have
Python or Anaconda**



Links to Python Distributions (Pick one)

Anaconda:

<https://www.anaconda.com/distribution/>

Miniconda:

<https://docs.conda.io/en/latest/miniconda.html>

Vanilla Python:

<https://www.python.org/downloads/release/python-374/>



Command Line stuff - Changing Directory (Folder)

Type `cd` followed by a path (Case insensitive) to change directories

Note: If the path contains white space, put it in quotes

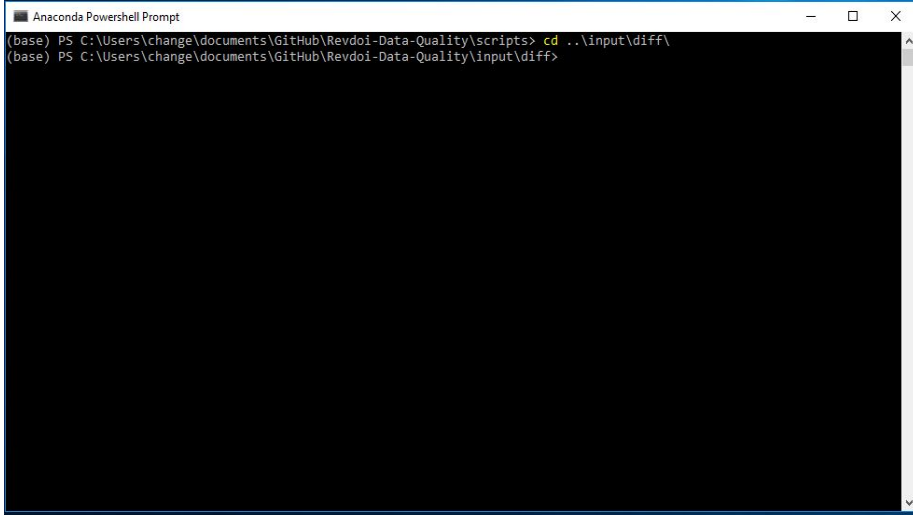
E.g. `cd documents/github/revdoi-data-quality/scripts`



Command Line stuff - Changing Directory (Folder)

To navigate backwards, type `cd ..`

- Can be followed by any path within that directory



```

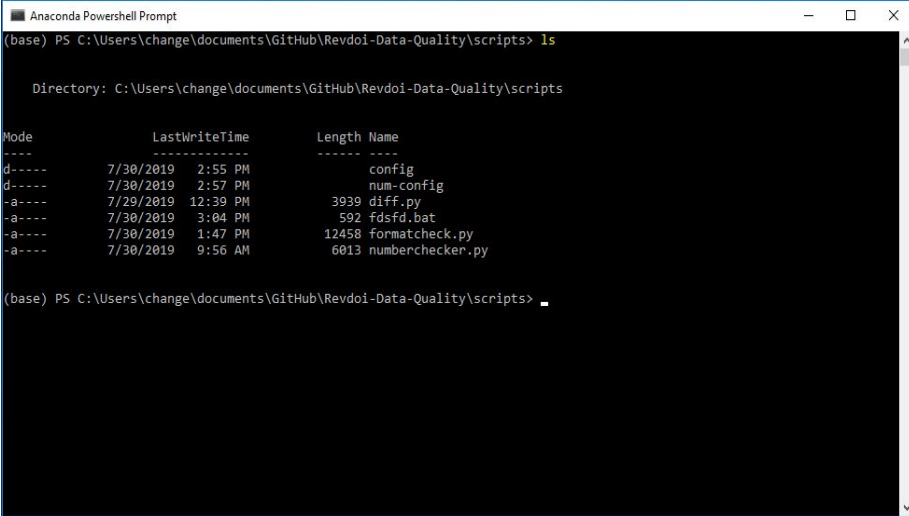
Anaconda PowerShell Prompt
(base) PS C:\Users\change\documents\GitHub\Revdoi-Data-Quality\scripts> cd ..\input\diff\
(base) PS C:\Users\change\documents\GitHub\Revdoi-Data-Quality\input\diff>

```



Command Line stuff - dir / ls

`dir` or `ls` will print out anything listed under the current directory



```

Anaconda PowerShell Prompt
(base) PS C:\Users\change\documents\GitHub\Revdoi-Data-Quality\scripts> ls

Directory: C:\Users\change\documents\GitHub\Revdoi-Data-Quality\scripts

Mode                LastWriteTime         Length Name
----                -
d-----          7/30/2019   2:55 PM              config
d-----          7/30/2019   2:57 PM             num-config
-a-----          7/29/2019  12:39 PM           3939 diff.py
-a-----          7/30/2019   3:04 PM            592 fdsfd.bat
-a-----          7/30/2019   1:47 PM          12458 formatcheck.py
-a-----          7/30/2019   9:56 AM            6013 numberchecker.py

(base) PS C:\Users\change\documents\GitHub\Revdoi-Data-Quality\scripts>

```



Command Line stuff - Running Python scripts

Type python followed by the name of the script

E.g. `python hello.py`



Sample Locations

Diff samples

`../input/diff`

Format Sample

`../input/setup`

`../input/formatTest`

Number Sample

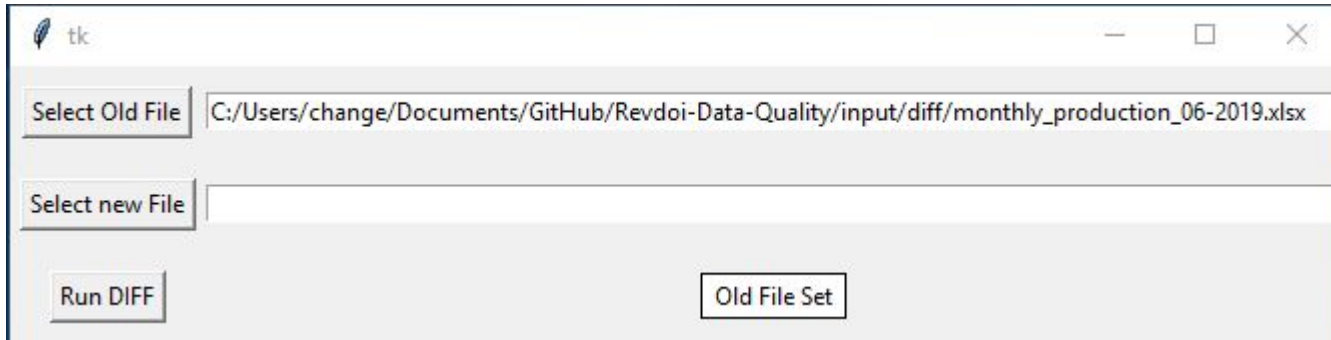
`../input/numTest`

Running the Scripts - diff.py

Cd [someplace]/revdoi-data-quality-master/scripts

```
python diff.py
```

Use the buttons to select files, then click “run diff” to compare the files



The screenshot shows a Tkinter window titled 'tk'. It contains two file selection fields and two buttons. The first field, labeled 'Select Old File', contains the path 'C:/Users/change/Documents/GitHub/Revdoi-Data-Quality/input/diff/monthly_production_06-2019.xlsx'. The second field, labeled 'Select new File', is empty. At the bottom left is a button labeled 'Run DIFF', and at the bottom right is a button labeled 'Old File Set'.

Select Old File	C:/Users/change/Documents/GitHub/Revdoi-Data-Quality/input/diff/monthly_production_06-2019.xlsx
Select new File	
Run DIFF	Old File Set

Running the Scripts - formatcheck.py

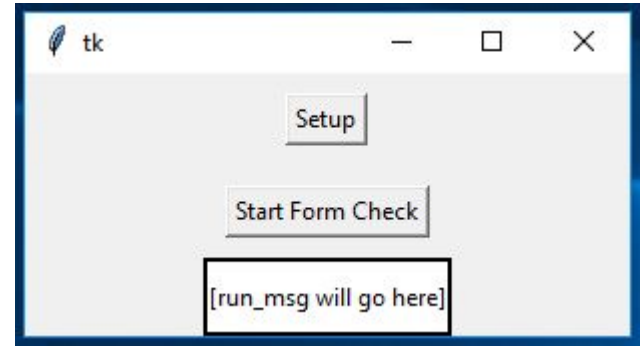
Python formatcheck.py

Run setup if a JSON file has not already been not created for a specific data type

- Found under Revdoi-Data-Quality/scripts/config

Start Form Check

- Prints and Exports Anomalies



Running the Scripts - numberchecker.py

Can work without calling setup

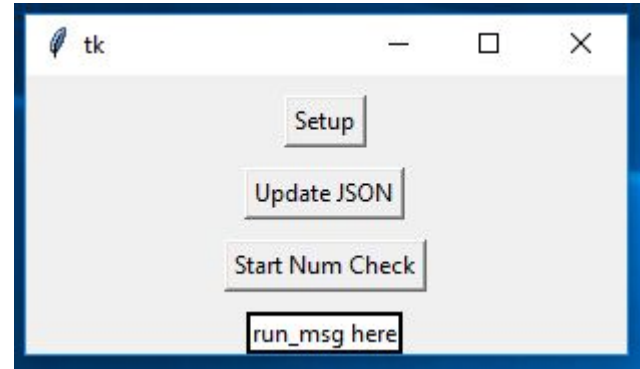
Python numberchecker.py

Setup

- Only need to be run once per data type
- Asks for input on which columns to use for grouping
- Creates a .json file with thresholds based on SD, is editable.

Start Num Check

- Prints out values surpassing certain thresholds
- Will export to output/NumChecked-filename.xlsx





In case I forgot something

Link to Data Script Wiki:

<https://github.com/ONRR/Revdoi-Data-Quality/wiki>

Glossary:

<https://github.com/ONRR/Revdoi-Data-Quality/wiki/Glossary>