Readme

Workflow example:

In order to produce the data and website created with this project, all files from the *public_html* folder should be hosted on a web server. The data that is displayed on the website is created by files in the *Dataset* folder and its subfolders.

An example of how to update the files if new data was received might go as follows:

- Create a PercentChangesCSV-Measurements.csv from the original data file. This is done by hand and consists of taking columns that are desired to be processed and pasting them into an Excel document and saving this document as a csv.
    - A participants variation of this file can be created by just transposing the data so that the rows are turned into columns, changing the measurement names into numbers, and relabeling each column as Participant1, Participant3, etc.
- From there all python files can be run to generate new data. This includes: 3cols.py, dataset.py, firstCol.py, heatmapCols.py, participantCols.py, and py-notebook.ipynb.
- Running these scripts should create all of the other files of data aside from the CorrCo files and the betaCVTable.csv file. These files are created by taking the 3cols.csv file and opening it in Excel, sorting from greatest to least, and splitting the rows in half into values above zero and values below zero. The betaCVTable.csv file is created from putting values from the output of betaCV.py into an Excel document and saving it as a .csv.
- Once all of the data files have been created, the ones that the files in the *public_html* folder require can be moved into that folder.
- The Output_Matrix.csv file(s) can be copied into the clustering folder which will allow the KMeansClustering.py file to be run.
- Any data that the clustering algorithm produces should be stored in the *Clusters* folder along with betaCV.py. Once all desired cluster files are in the *Clusters* folder the betaCV.py program can be run. This program will output values that can be copied into the betaCVTable.csv file via Excel. betaCVTable.csv should be found in the *Clustering* folder.
- Once all of the clustering data (files for two to ten clusters of participant data and two to fifteen clusters of measurement data) has been generated and placed in the folder and the betaCVTable.csv file has been updated, the *Clusters* folder and betaCVTable.csv should be copied to the *public_html* directory.
- In the *pubic_html* directory, the betaCV.py file can be deleted from the *Clusters* folder.

*** Images of the structure of these folders can be found at the bottom of this readme.**

public_html:

The .php and .css files in this folder are the files that make the web pages appear but they all use data and images from the .csv .jpg and .png files. All files in this folder are required for proper functioning of the website. This includes the files in the subfolder “Clusters” which contains the files that the clustering.php file reads and displays.

about.php - Is a simple about us page that gives a little background but is otherwise unimportant.

home.php - Is the file that creates the clustering page and is set to be the homepage of the interactive website. It reads data from the files in the *Clusters* folder within the *public_html* folder to create the tables and visualization of the clusters. It also reads data from **betaCVTable.csv** to display data in a table at the top of the page. **betaCVTable.csv** should be in the *public_html* folder but not in the *Clusters* folder.

footer.php - Contains the javascript calls and enables bootstrap for formatting tables and page layout. Also contains a footer message with the group's name.

header.php - Is used in all other .php files to provide stylesheets and metadata.

heatmap.php - Reads data from **heatmapCols.csv** and displays interactive heatmap. Also relies on the **Heatmap_output.png** to allow users to download the .png version of the heatmap.

correlation.php - Displays correlation data in tables. The two tables that display correlation values above and below zero read data from the **CorrCoAboveZero.csv** and **CorrCoBelowZero.csv** files. The search functionality of this page relies on **firstCol.csv** to provide autocomplete measurement names and **3cols.csv** to display filtered search results.

nav.php - Similar to footer.php, just allows the navigation to appear at the top of each page which enables the user to click to view different pages. It relies on **logo2.jpg** to display the logo image.

style.css - Provides styling of the .php web pages.

Dataset:

The python files in this folder are used to generate the other files in this folder. All python files should be run with python3 (ex: python3 dataset.py) and some require libraries to be installed on the computer running the code. *Descriptions of how to install these libraries on a windows machine are provided below but resources online also should be helpful in describing how to install these libraries in case the code is going to be run on another operating system. These should be installed to avoid any errors.*

3cols.py - Generates the file **3cols.csv** which is used to create the **CorrCoAboveZero.csv** and **CorrCoBelowZero.csv** files. Those two CorrCo files are not created with any python code however, we just took the data from 3cols.csv and put it into Excel where we sorted it from highest to lowest and split the values above zero and below zero into those two separate files. These three files that 3cols.py produces are responsible for providing the clustering data on the **home.php** webpage.

dataset.py - Reads either **PercentChangeCSV-Measurements.csv** or **PercentChangeCSV-Participants.csv** to produce a **correlation_coefficent_results.txt** file and a **Output_Matrix.csv** file. The Output_Matrix.csv files are read by the clustering algorithm. Explanations of how to change the code to output different named files are in the comments of the dataset.py file.

firstCol.py - Is used to create a file of just measurement names called **firstCol.csv**. firstCol.csv is used by **home.php** to generate autocomplete results for the search function.

heatmapCols.py - Creates a file very similar to 3col.csv that is used by **heatmap.php** to provide data for the heatmap to display. This program creates a file called **heatmapCols.csv** that has values rounded to 3 decimal points.

participantCols.py - Does the same thing 3cols.py does except just for the participant data. It creates a file called **participantCols.csv** that is currently not used anywhere but could be used in similar fashion to the 3cols.csv file.

PercentChangeCSV-Measurements.csv and **PercentChangeCSV-Participants.csv** - Are the two files that we created by hand after receiving the original data from the GHA study. These two files should contain all of the data from the GHA studied that was labeled to be used.

py-notebook.ipynb - Is used to create the **Heatmap_output.png** which is provided as a download on **heatmap.php**.

- To view py-notebook.ipynb, Jupyter Notebook may need to be installed.
    - View steps below to see installation

Clustering:

This is a subfolder of the *Dataset* folder. This contains all the files needed for and produced by the clustering algorithm. The files that the clustering algorithm creates should be stored in the *Clusters* folder.

Clusters folder - Is where to move all .json and .csv files created by **KMeansClustering.py**. This folder should be copied to the *public_html* directory because **clustering.php** reads data from it to make the displays on that page. This folder also includes **betaCV.py** which can be run once the clustering files are placed in the folder. As output it displays values that should be placed in another file in the *Clustering* folder called **betaCVTable.csv**. When copied over to the *public_html* directory, **betaCV.py** can be deleted and **betaCVTable.csv** should be in just the *public_html* directory, not in the *Clusters* folder.

kmeans_clustering_idea.txt - Is the general idea behind the K-means algorithm we wrote. Serves no purpose other than viewing and explanation.

KMeansClustering.py - Does the clustering and creates .json and .csv files that can be stored in the *Clusters* folder and read by the **clustering.php** website file. A copy of the *Clusters* folder should be kept on the website so that those files can be read by the clustering.php file. There are comments in KMeansClustering.py that explain how to change the file names and code in order to produce different files.

Output_Matrix.csv - Is the file that contains the measurement data output by dataset.py

Output_Matrix2.csv - Is the same as the other file except it contains participant data.

HOW TO INSTALL PYTHON LIBRARIES (For Windows):

- pandas

Open Command Terminal & Enter the following:

- python.exe -m pip install pip
- python.exe -m pip install --upgrade pip

    - Run python.exe -m pip install --upgrade pip *one* more time to ensure its installed

- numpy

Open Command Terminal & Enter the following:

- pip install numpy

- matplotlib.pylot

Open Command Terminal & Enter the following:

- pip install matplotlib

- seaborn

Open Command Terminal & Enter the following:

- pip install seaborn

- JupyterLab

Open Command Terminal & Enter the following:

- pip install jupyterlab
- Opening *.ipynb* files: Open the folder containing the .ipynb file and open the terminal in that directory and type the following:
    - Python -m notebook

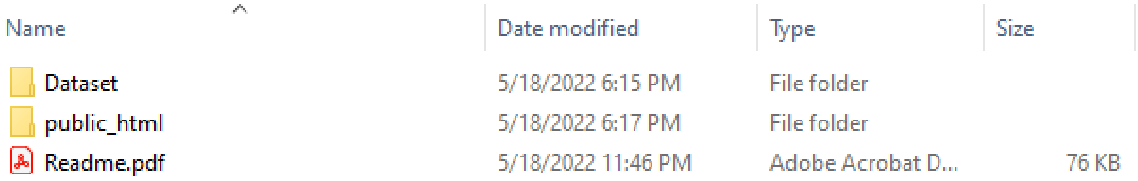
- Statsmodels

Open Command Terminal & Enter the following:

- python -m pip install statsmodels

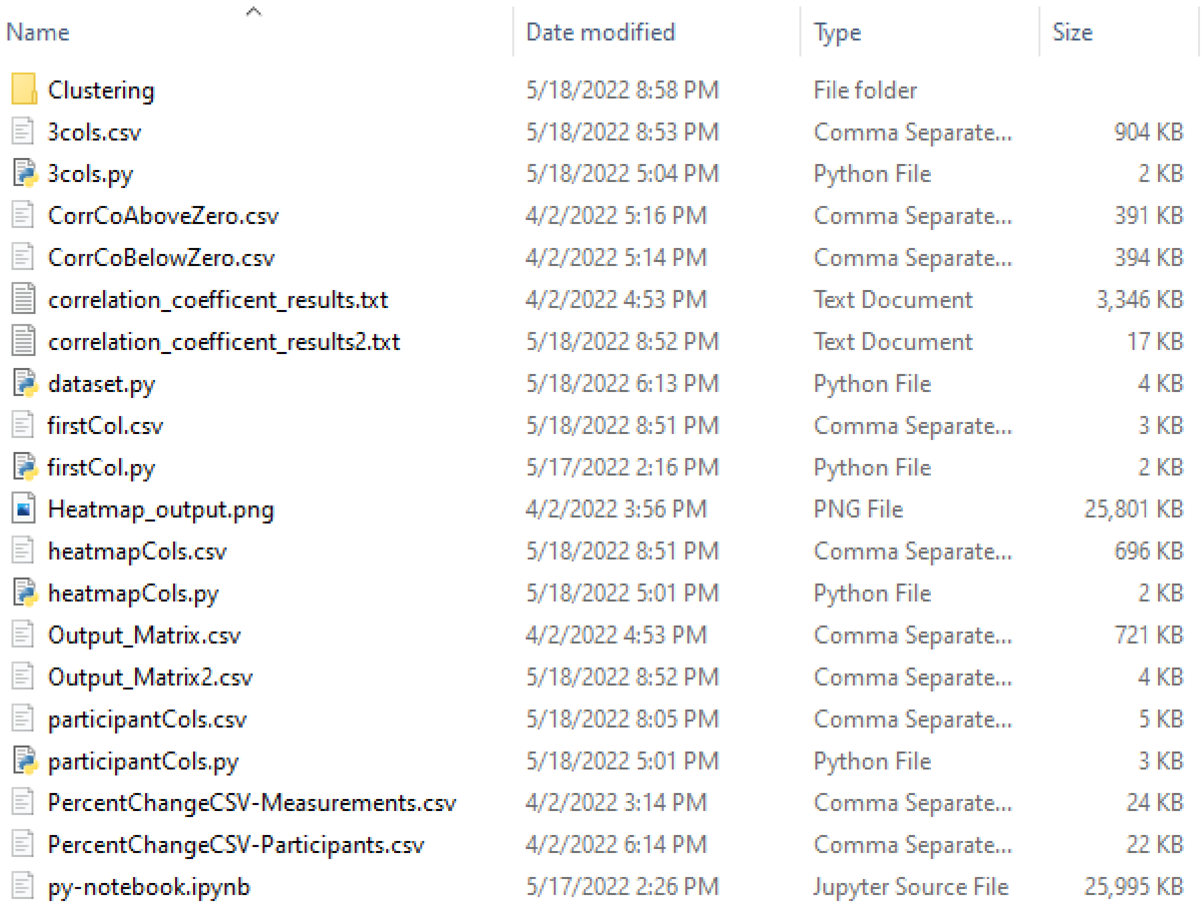
File structures:

Root directory:



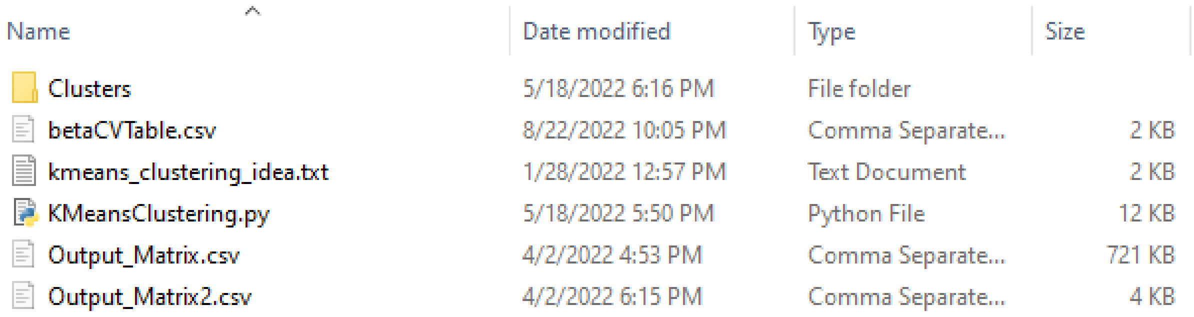
Name	Date modified	Type	Size
Dataset	5/18/2022 6:15 PM	File folder	
public_html	5/18/2022 6:17 PM	File folder	
Readme.pdf	5/18/2022 11:46 PM	Adobe Acrobat D...	76 KB

Dataset:



Name	Date modified	Type	Size
Clustering	5/18/2022 8:58 PM	File folder	
3cols.csv	5/18/2022 8:53 PM	Comma Separate...	904 KB
3cols.py	5/18/2022 5:04 PM	Python File	2 KB
CorrCoAboveZero.csv	4/2/2022 5:16 PM	Comma Separate...	391 KB
CorrCoBelowZero.csv	4/2/2022 5:14 PM	Comma Separate...	394 KB
correlation_coefficent_results.txt	4/2/2022 4:53 PM	Text Document	3,346 KB
correlation_coefficent_results2.txt	5/18/2022 8:52 PM	Text Document	17 KB
dataset.py	5/18/2022 6:13 PM	Python File	4 KB
firstCol.csv	5/18/2022 8:51 PM	Comma Separate...	3 KB
firstCol.py	5/17/2022 2:16 PM	Python File	2 KB
Heatmap_output.png	4/2/2022 3:56 PM	PNG File	25,801 KB
heatmapCols.csv	5/18/2022 8:51 PM	Comma Separate...	696 KB
heatmapCols.py	5/18/2022 5:01 PM	Python File	2 KB
Output_Matrix.csv	4/2/2022 4:53 PM	Comma Separate...	721 KB
Output_Matrix2.csv	5/18/2022 8:52 PM	Comma Separate...	4 KB
participantCols.csv	5/18/2022 8:05 PM	Comma Separate...	5 KB
participantCols.py	5/18/2022 5:01 PM	Python File	3 KB
PercentChangeCSV-Measurements.csv	4/2/2022 3:14 PM	Comma Separate...	24 KB
PercentChangeCSV-Participants.csv	4/2/2022 6:14 PM	Comma Separate...	22 KB
py-notebook.ipynb	5/17/2022 2:26 PM	Jupyter Source File	25,995 KB

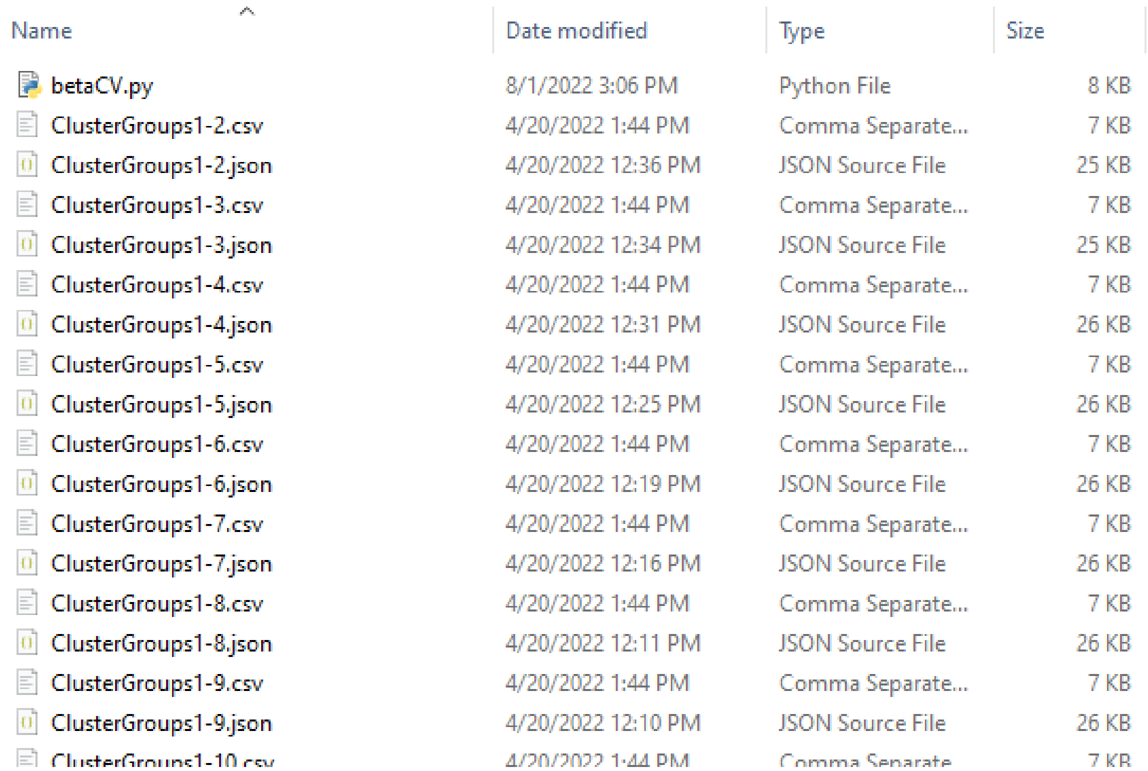
Clustering (within Dataset):



Name	Date modified	Type	Size
Clusters	5/18/2022 6:16 PM	File folder	
betaCVTable.csv	8/22/2022 10:05 PM	Comma Separate...	2 KB
kmeans_clustering_idea.txt	1/28/2022 12:57 PM	Text Document	2 KB
KMeansClustering.py	5/18/2022 5:50 PM	Python File	12 KB
Output_Matrix.csv	4/2/2022 4:53 PM	Comma Separate...	721 KB
Output_Matrix2.csv	4/2/2022 6:15 PM	Comma Separate...	4 KB

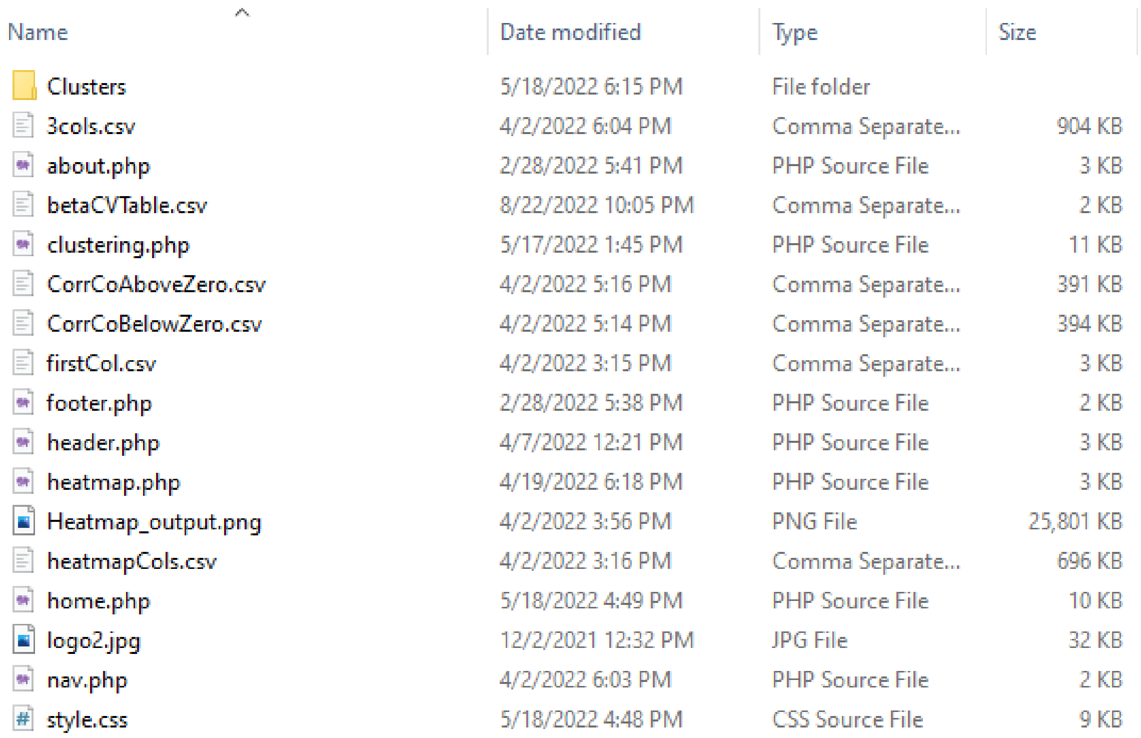
Clusters (within Dataset/Clustering):

This snippet is only part of what is in the folder, there are files for ClusterGroups1-2 to ClusterGroups1-15 and ClusterGroups2-2 to ClusterGroups2-10.



Name	Date modified	Type	Size
betaCV.py	8/1/2022 3:06 PM	Python File	8 KB
ClusterGroups1-2.csv	4/20/2022 1:44 PM	Comma Separate...	7 KB
ClusterGroups1-2.json	4/20/2022 12:36 PM	JSON Source File	25 KB
ClusterGroups1-3.csv	4/20/2022 1:44 PM	Comma Separate...	7 KB
ClusterGroups1-3.json	4/20/2022 12:34 PM	JSON Source File	25 KB
ClusterGroups1-4.csv	4/20/2022 1:44 PM	Comma Separate...	7 KB
ClusterGroups1-4.json	4/20/2022 12:31 PM	JSON Source File	26 KB
ClusterGroups1-5.csv	4/20/2022 1:44 PM	Comma Separate...	7 KB
ClusterGroups1-5.json	4/20/2022 12:25 PM	JSON Source File	26 KB
ClusterGroups1-6.csv	4/20/2022 1:44 PM	Comma Separate...	7 KB
ClusterGroups1-6.json	4/20/2022 12:19 PM	JSON Source File	26 KB
ClusterGroups1-7.csv	4/20/2022 1:44 PM	Comma Separate...	7 KB
ClusterGroups1-7.json	4/20/2022 12:16 PM	JSON Source File	26 KB
ClusterGroups1-8.csv	4/20/2022 1:44 PM	Comma Separate...	7 KB
ClusterGroups1-8.json	4/20/2022 12:11 PM	JSON Source File	26 KB
ClusterGroups1-9.csv	4/20/2022 1:44 PM	Comma Separate...	7 KB
ClusterGroups1-9.json	4/20/2022 12:10 PM	JSON Source File	26 KB
ClusterGroups1-10.csv	4/20/2022 1:44 PM	Comma Separate...	7 KB

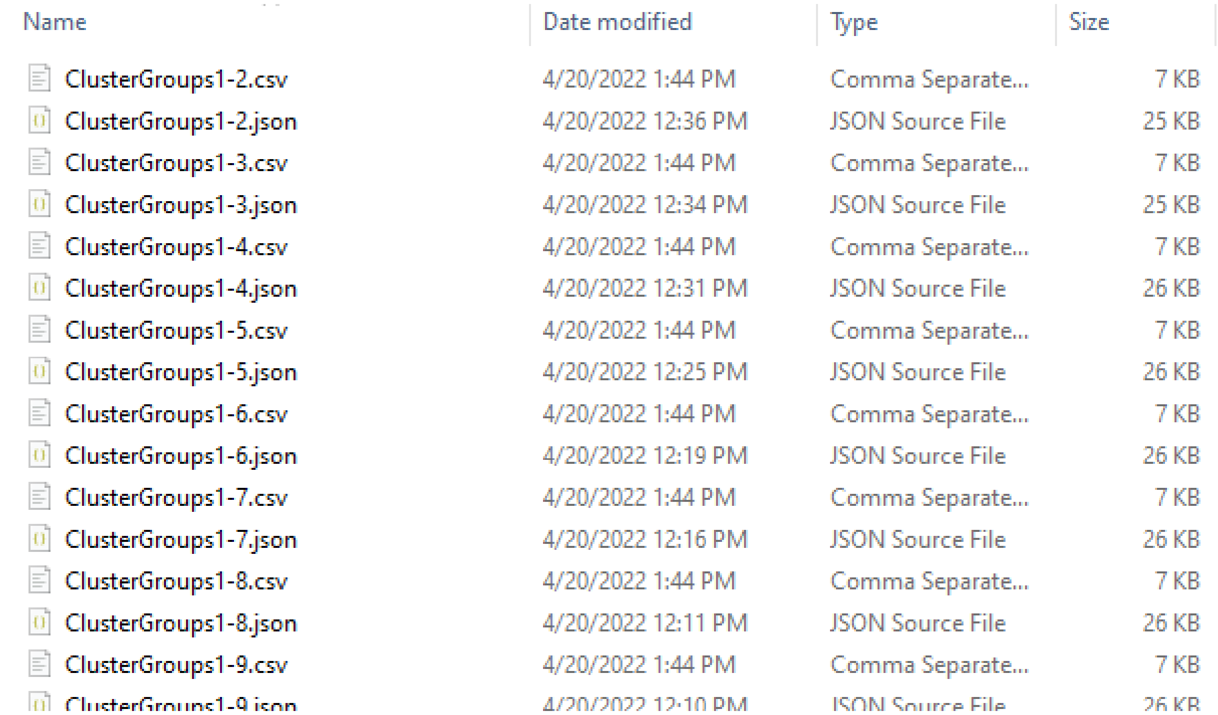
Public_html:



Name	Date modified	Type	Size
Clusters	5/18/2022 6:15 PM	File folder	
3cols.csv	4/2/2022 6:04 PM	Comma Separate...	904 KB
about.php	2/28/2022 5:41 PM	PHP Source File	3 KB
betaCVTable.csv	8/22/2022 10:05 PM	Comma Separate...	2 KB
clustering.php	5/17/2022 1:45 PM	PHP Source File	11 KB
CorrCoAboveZero.csv	4/2/2022 5:16 PM	Comma Separate...	391 KB
CorrCoBelowZero.csv	4/2/2022 5:14 PM	Comma Separate...	394 KB
firstCol.csv	4/2/2022 3:15 PM	Comma Separate...	3 KB
footer.php	2/28/2022 5:38 PM	PHP Source File	2 KB
header.php	4/7/2022 12:21 PM	PHP Source File	3 KB
heatmap.php	4/19/2022 6:18 PM	PHP Source File	3 KB
Heatmap_output.png	4/2/2022 3:56 PM	PNG File	25,801 KB
heatmapCols.csv	4/2/2022 3:16 PM	Comma Separate...	696 KB
home.php	5/18/2022 4:49 PM	PHP Source File	10 KB
logo2.jpg	12/2/2021 12:32 PM	JPG File	32 KB
nav.php	4/2/2022 6:03 PM	PHP Source File	2 KB
style.css	5/18/2022 4:48 PM	CSS Source File	9 KB

Clusters (within public_html):

This snippet is only part of what is in the folder, there are files for ClusterGroups1-2 to ClusterGroups1-15 and ClusterGroups2-2 to ClusterGroups2-10.



Name	Date modified	Type	Size
ClusterGroups1-2.csv	4/20/2022 1:44 PM	Comma Separate...	7 KB
ClusterGroups1-2.json	4/20/2022 12:36 PM	JSON Source File	25 KB
ClusterGroups1-3.csv	4/20/2022 1:44 PM	Comma Separate...	7 KB
ClusterGroups1-3.json	4/20/2022 12:34 PM	JSON Source File	25 KB
ClusterGroups1-4.csv	4/20/2022 1:44 PM	Comma Separate...	7 KB
ClusterGroups1-4.json	4/20/2022 12:31 PM	JSON Source File	26 KB
ClusterGroups1-5.csv	4/20/2022 1:44 PM	Comma Separate...	7 KB
ClusterGroups1-5.json	4/20/2022 12:25 PM	JSON Source File	26 KB
ClusterGroups1-6.csv	4/20/2022 1:44 PM	Comma Separate...	7 KB
ClusterGroups1-6.json	4/20/2022 12:19 PM	JSON Source File	26 KB
ClusterGroups1-7.csv	4/20/2022 1:44 PM	Comma Separate...	7 KB
ClusterGroups1-7.json	4/20/2022 12:16 PM	JSON Source File	26 KB
ClusterGroups1-8.csv	4/20/2022 1:44 PM	Comma Separate...	7 KB
ClusterGroups1-8.json	4/20/2022 12:11 PM	JSON Source File	26 KB
ClusterGroups1-9.csv	4/20/2022 1:44 PM	Comma Separate...	7 KB
ClusterGroups1-9.json	4/20/2022 12:10 PM	JSON Source File	26 KB