Version Date: April 5th, 2021

Instruction Manual

NIH Database Pipeline Application
(Version 2.0)

Funded by NIMH R01MH116156-01A1

Principal Investigator: Jessica L. Nielson, PhD
Developed by: Thomas Kirsh, BS

Table of Contents

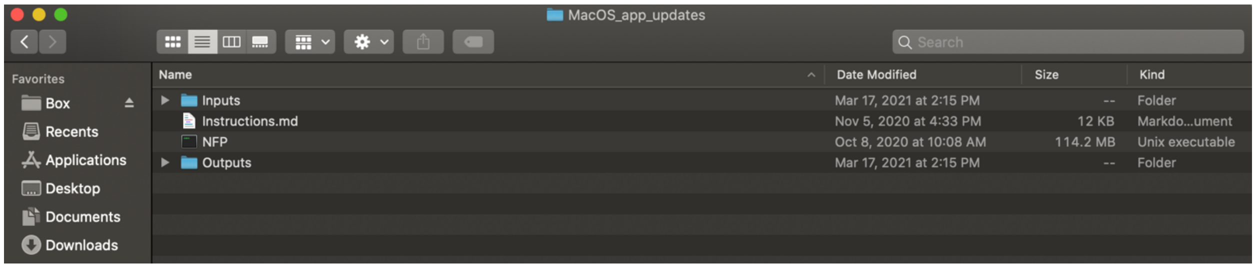
Downloading/Installing the Pipeline

GitHub Repository

Download Instructions

- Location: https://github.com/Nielson-Lab/NIH-database-pipeline
- Click on "Releases" (underneath the "About" section) on the right side of the home page
- Find the most recent MacOS and Windows versions of the application, then click on "Assets" underneath those headers
- Only download the application ZIP file that is compatible with your system

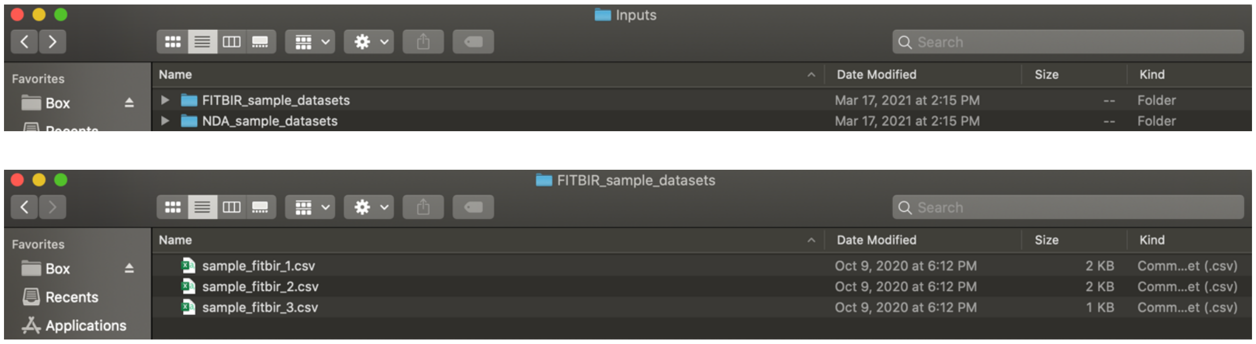
Using the Pipeline



Name	Date Modified	Size	Kind
Inputs	Mar 17, 2021 at 2:15 PM	--	Folder
Instructions.md	Nov 5, 2020 at 4:33 PM	12 KB	Markdo...ument
NFP	Oct 8, 2020 at 10:08 AM	114.2 MB	Unix executable
Outputs	Mar 17, 2021 at 2:15 PM	--	Folder

Figure 1. The main folder for the application. This folder gets downloaded from Github. Please note that although this screenshot is for the Mac version of the application, the Windows version has the same files.

The downloaded file should contain folders for *Inputs* and *Outputs,* as well as the application (NFP) and an *Instructions.md* file. The user **should not** move the application out of the *MacOS_app* folder.



Name	Date Modified	Size	Kind
FITBIR_sample_datasets	Mar 17, 2021 at 2:15 PM	--	Folder
NDA_sample_datasets	Mar 17, 2021 at 2:15 PM	--	Folder

Name	Date Modified	Size	Kind
sample_fitbir_1.csv	Oct 9, 2020 at 6:12 PM	2 KB	Comm...et (.csv)
sample_fitbir_2.csv	Oct 9, 2020 at 6:12 PM	2 KB	Comm...et (.csv)
sample_fitbir_3.csv	Oct 9, 2020 at 6:12 PM	1 KB	Comm...et (.csv)

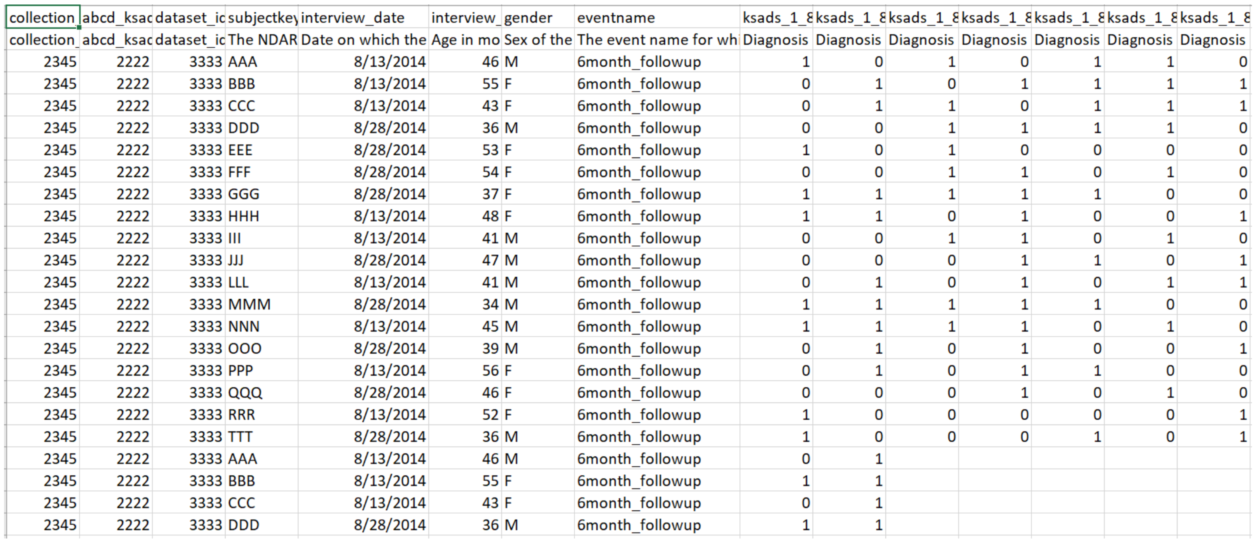
Figure 2. The *Inputs* folder in the application. Datasets that will be used in the application are uploaded here. The application only reads ".csv" or ".txt" files, so files stored in folders will not be read. Those files will need to be moved from the folder.

The application will automatically look for the files the user wants to work on in the *Inputs* folder. Therefore, the user should put the files they want to work with or on in the *Inputs* folder. The files **must** either be ".csv" or ".txt" files; ".xlsx" files are not supported at this time. Those files can easily be converted to either of the accepted formats. Having separate folders in the *Inputs* folder is fine because the application will ignore them when looking for files.

Fake Datasets for Practice

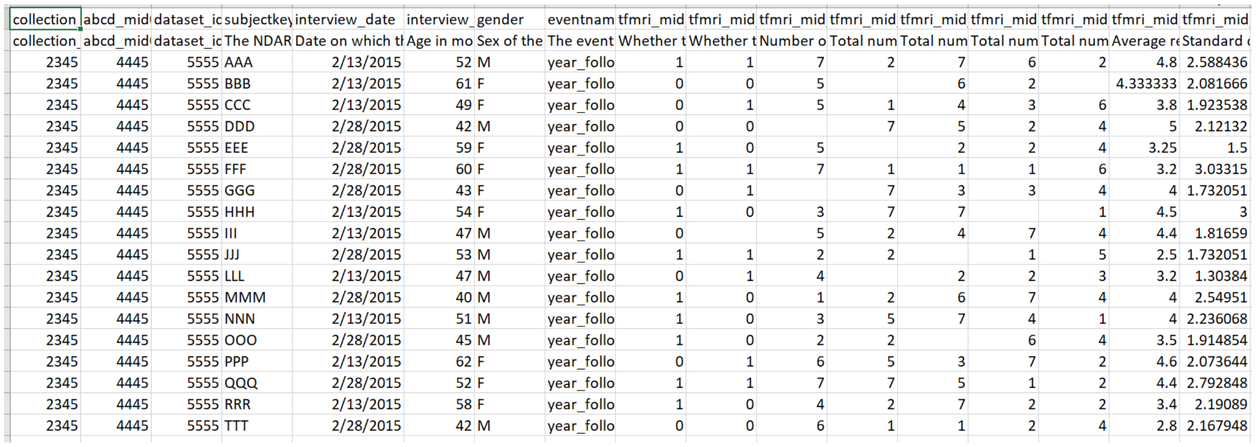
The application comes with some sample FITBIR and NDA datasets. These datasets were generated by us to reflect the unique aspects about the data files from each database. All the continuous variables were randomly generated using either the RAND() function or the RANDBETWEEN() function in Excel. Categorical and date variables were assigned random variables that reflect values in a similar format to the format used by the databases. Certain time intervals were modeled after the "3 months" and "6 months" values found inside the TRACK-TBI dataset in FITBIR; however, the data in the sample files do not match any person inside the TRACK-TBI dataset.

Legitimate names of forms in NDA, and data elements in both NDA and FITBIR are used and necessary for the user to be able to play with the data dictionary scraping features.



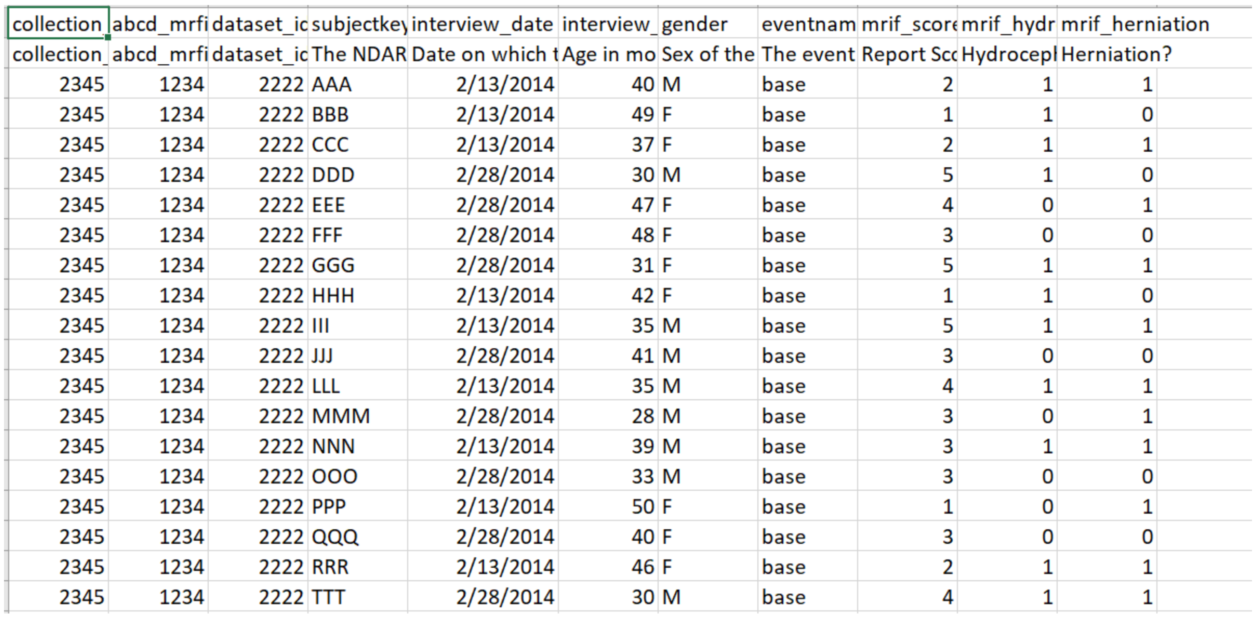
collection_	abcd_ksac	dataset_ic	subjectkey	interview_date	interview_	gender	eventname	ksads_1_8	ksads_1_8	ksads_1_8	ksads_1_8	ksads_1_8	ksads_1_8	ksads_1_8
collection_	abcd_ksac	dataset_ic	The NDAR	Date on which the	Age in mo	Sex of the	The event name for wh	Diagnosis	Diagnosis	Diagnosis	Diagnosis	Diagnosis	Diagnosis	Diagnosis
2345	2222	3333	AAA	8/13/2014	46	M	6month_followup	1	0	1	0	1	1	0
2345	2222	3333	BBB	8/13/2014	55	F	6month_followup	0	1	0	1	1	1	1
2345	2222	3333	CCC	8/13/2014	43	F	6month_followup	0	1	1	0	1	1	1
2345	2222	3333	DDD	8/28/2014	36	M	6month_followup	0	0	1	1	1	1	0
2345	2222	3333	EEE	8/28/2014	53	F	6month_followup	1	0	1	0	0	0	0
2345	2222	3333	FFF	8/28/2014	54	F	6month_followup	0	0	1	1	0	1	0
2345	2222	3333	GGG	8/28/2014	37	F	6month_followup	1	1	1	1	1	0	0
2345	2222	3333	HHH	8/13/2014	48	F	6month_followup	1	1	0	1	0	0	1
2345	2222	3333	III	8/13/2014	41	M	6month_followup	0	0	1	1	0	1	0
2345	2222	3333	JJJ	8/28/2014	47	M	6month_followup	0	0	0	1	1	0	1
2345	2222	3333	LLL	8/13/2014	41	M	6month_followup	0	1	0	1	0	1	1
2345	2222	3333	MMM	8/28/2014	34	M	6month_followup	1	1	1	1	1	0	0
2345	2222	3333	NNN	8/13/2014	45	M	6month_followup	1	1	1	1	0	1	0
2345	2222	3333	OOO	8/28/2014	39	M	6month_followup	0	1	0	1	0	0	1
2345	2222	3333	PPP	8/13/2014	56	F	6month_followup	0	1	0	1	1	0	0
2345	2222	3333	QQQ	8/28/2014	46	F	6month_followup	0	0	0	1	0	1	0
2345	2222	3333	RRR	8/13/2014	52	F	6month_followup	1	0	0	0	0	0	1
2345	2222	3333	TTT	8/28/2014	36	M	6month_followup	1	0	0	0	1	0	1
2345	2222	3333	AAA	8/13/2014	46	M	6month_followup	0	1					
2345	2222	3333	BBB	8/13/2014	55	F	6month_followup	1	1					
2345	2222	3333	CCC	8/13/2014	43	F	6month_followup	0	1					
2345	2222	3333	DDD	8/28/2014	36	M	6month_followup	1	1					

Figure 3. Sample NDA dataset called "abcd_ksads01.txt" in order for users to be able to test the web scraping functionality. Please see the above description for how the datasets were generated.



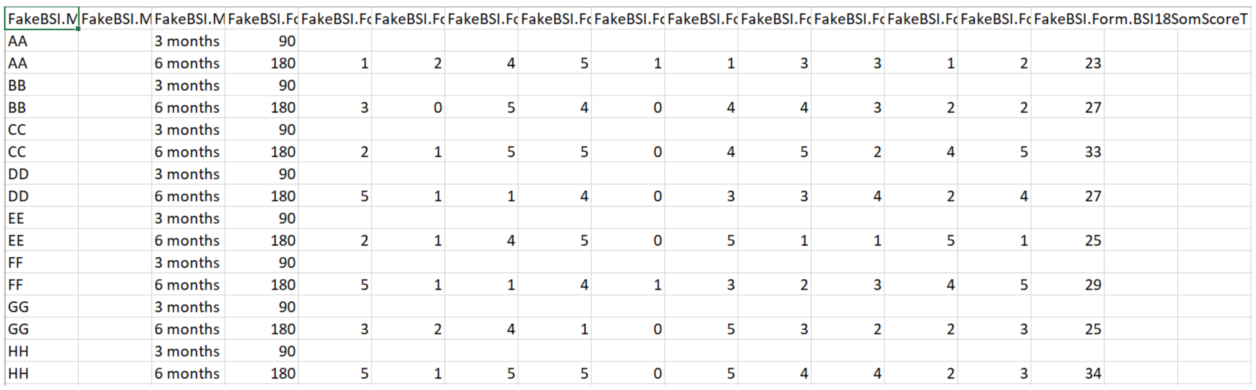
collection_	abcd_mid	dataset_ic	subjectkey	interview_date	interview_	gender	eventnam	tfmri_mid	tfmri_mid	tfmri_mid	tfmri_mid	tfmri_mid	tfmri_mid	tfmri_mid	tfmri_mid	tfmri_mid
collection_	abcd_mid	dataset_ic	The NDAR	Date on which th	Age in mo	Sex of the	The event	Whether t	Whether t	Number o	Total num	Total num	Total num	Total num	Average re	Standard
2345	4445	5555	AAA	2/13/2015	52	M	year_follo	1	1	7	2	7	6	2	4.8	2.588436
2345	4445	5555	BBB	2/13/2015	61	F	year_follo	0	0	5		6	2		4.333333	2.081666
2345	4445	5555	CCC	2/13/2015	49	F	year_follo	0	1	5	1	4	3	6	3.8	1.923538
2345	4445	5555	DDD	2/28/2015	42	M	year_follo	0	0		7	5	2	4	5	2.12132
2345	4445	5555	EEE	2/28/2015	59	F	year_follo	1	0	5		2	2	4	3.25	1.5
2345	4445	5555	FFF	2/28/2015	60	F	year_follo	1	1	7	1	1	1	6	3.2	3.03315
2345	4445	5555	GGG	2/28/2015	43	F	year_follo	0	1		7	3	3	4	4	1.732051
2345	4445	5555	HHH	2/13/2015	54	F	year_follo	1	0	3	7	7		1	4.5	3
2345	4445	5555	III	2/13/2015	47	M	year_follo	0		5	2	4	7	4	4.4	1.81659
2345	4445	5555	JJJ	2/28/2015	53	M	year_follo	1	1	2	2		1	5	2.5	1.732051
2345	4445	5555	LLL	2/13/2015	47	M	year_follo	0	1	4		2	2	3	3.2	1.30384
2345	4445	5555	MMM	2/28/2015	40	M	year_follo	1	0	1	2	6	7	4	4	2.54951
2345	4445	5555	NNN	2/13/2015	51	M	year_follo	1	0	3	5	7	4	1	4	2.236068
2345	4445	5555	OOO	2/28/2015	45	M	year_follo	1	0	2	2		6	4	3.5	1.914854
2345	4445	5555	PPP	2/13/2015	62	F	year_follo	0	1	6	5	3	7	2	4.6	2.073644
2345	4445	5555	QQQ	2/28/2015	52	F	year_follo	1	1	7	7	5	1	2	4.4	2.792848
2345	4445	5555	RRR	2/13/2015	58	F	year_follo	1	0	4	2	7	2	2	3.4	2.19089
2345	4445	5555	TTT	2/28/2015	42	M	year_follo	0	0	6	1	1	2	4	2.8	2.167948

Figure 4. Sample NDA dataset called "abcd_mid02.txt" to match the name of a data dictionary users could scrape. Please see the above description for how the datasets were generated.



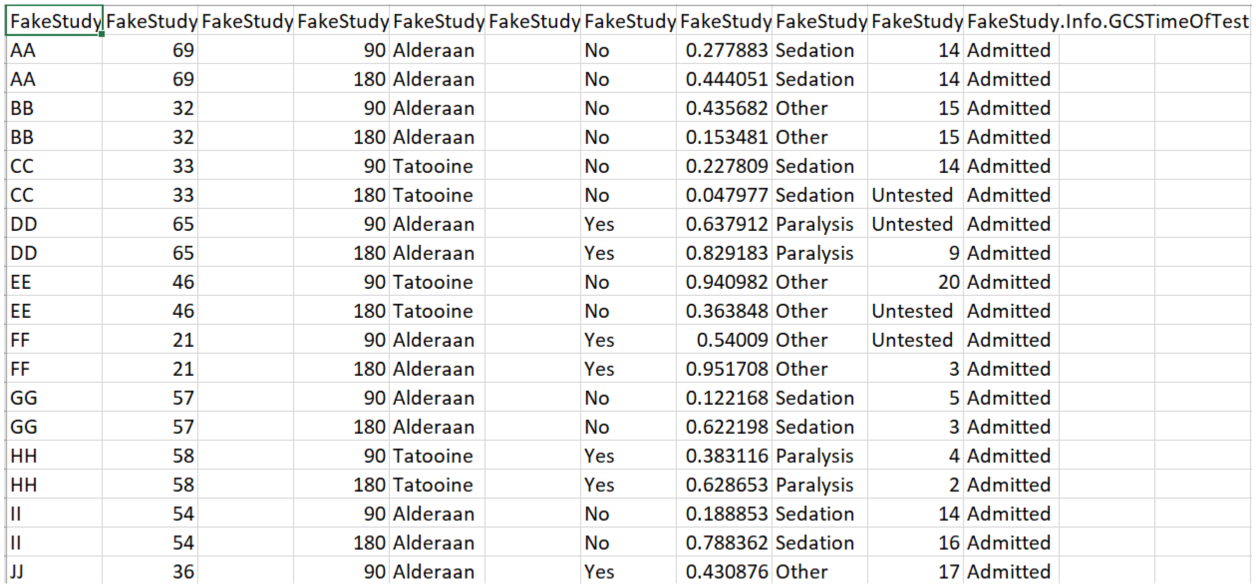
collection	abcd_mrfi	dataset_ic	subjectkey	interview_date	interview_	gender	eventnam	mrif_score	mrif_hydr	mrif_herniation
collection_	abcd_mrfi	dataset_ic	The NDAR	Date on which t	Age in mo	Sex of the	The event	Report Sc	Hydrocepl	Herniation?
2345	1234	2222	AAA	2/13/2014	40	M	base	2	1	1
2345	1234	2222	BBB	2/13/2014	49	F	base	1	1	0
2345	1234	2222	CCC	2/13/2014	37	F	base	2	1	1
2345	1234	2222	DDD	2/28/2014	30	M	base	5	1	0
2345	1234	2222	EEE	2/28/2014	47	F	base	4	0	1
2345	1234	2222	FFF	2/28/2014	48	F	base	3	0	0
2345	1234	2222	GGG	2/28/2014	31	F	base	5	1	1
2345	1234	2222	HHH	2/13/2014	42	F	base	1	1	0
2345	1234	2222	III	2/13/2014	35	M	base	5	1	1
2345	1234	2222	JJJ	2/28/2014	41	M	base	3	0	0
2345	1234	2222	LLL	2/13/2014	35	M	base	4	1	1
2345	1234	2222	MMM	2/28/2014	28	M	base	3	0	1
2345	1234	2222	NNN	2/13/2014	39	M	base	3	1	1
2345	1234	2222	OOO	2/28/2014	33	M	base	3	0	0
2345	1234	2222	PPP	2/13/2014	50	F	base	1	0	1
2345	1234	2222	QQQ	2/28/2014	40	F	base	3	0	0
2345	1234	2222	RRR	2/13/2014	46	F	base	2	1	1
2345	1234	2222	TTT	2/28/2014	30	M	base	4	1	1

Figure 5. Sample dataset titled "abcd_mrifindings01.txt". Please see the above description for how the dataset was generated.



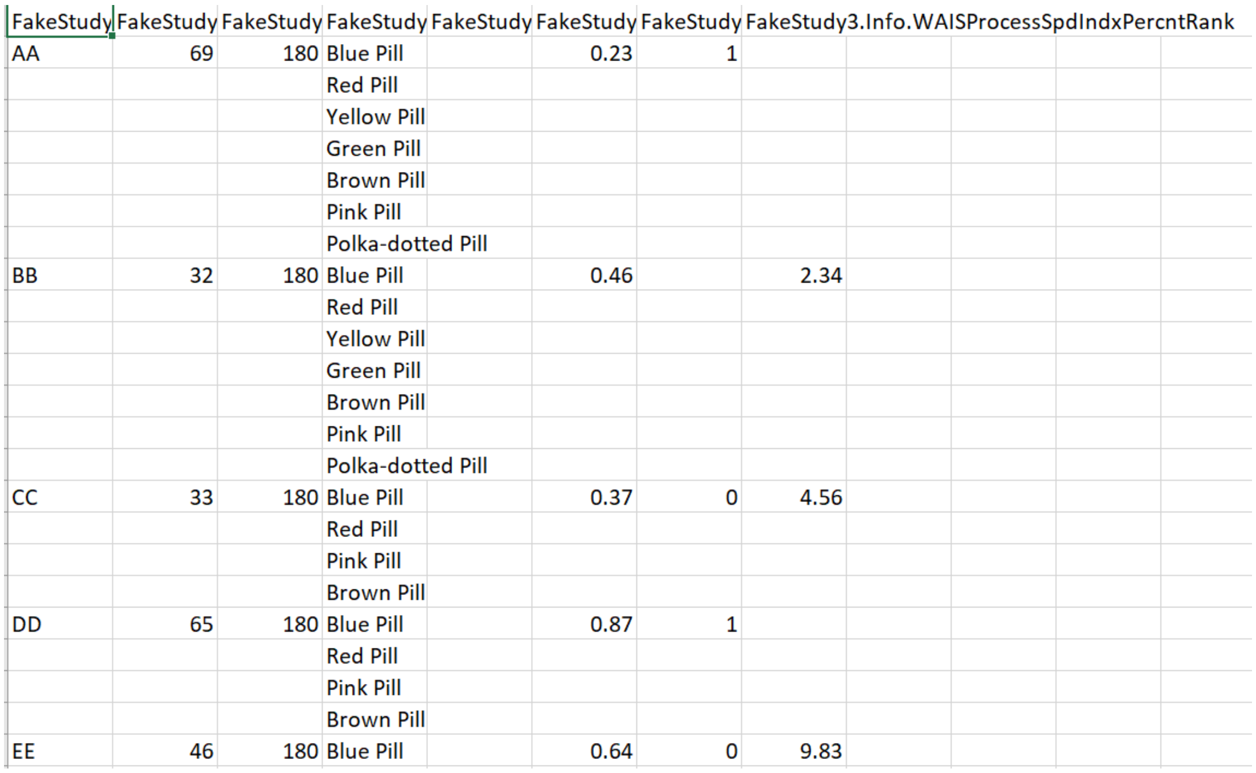
FakeBSI.M	FakeBSI.M	FakeBSI.M	FakeBSI.Fc	FakeBSI.Fc	FakeBSI.Fc	FakeBSI.Fc	FakeBSI.Fc	FakeBSI.Fc	FakeBSI.Fc	FakeBSI.Fc	FakeBSI.Fc	FakeBSI.Fc	FakeBSI.Fc	FakeBSI.Form.BSI18SomScoreT
AA		3 months	90											
AA		6 months	180	1	2	4	5	1	1	3	3	1	2	23
BB		3 months	90											
BB		6 months	180	3	0	5	4	0	4	4	3	2	2	27
CC		3 months	90											
CC		6 months	180	2	1	5	5	0	4	5	2	4	5	33
DD		3 months	90											
DD		6 months	180	5	1	1	4	0	3	3	4	2	4	27
EE		3 months	90											
EE		6 months	180	2	1	4	5	0	5	1	1	5	1	25
FF		3 months	90											
FF		6 months	180	5	1	1	4	1	3	2	3	4	5	29
GG		3 months	90											
GG		6 months	180	3	2	4	1	0	5	3	2	2	3	25
HH		3 months	90											
HH		6 months	180	5	1	5	5	0	5	4	4	2	3	34

Figure 6. Sample dataset from FITBIR. The column names contain parts separated by periods. The "time" column here contains strings as time points.



FakeStudy	FakeStudy	FakeStudy	FakeStudy	FakeStudy	FakeStudy	FakeStudy	FakeStudy	FakeStudy	FakeStudy	FakeStudy.Info.GCSTimeOfTest
AA	69		90	Alderaan		No	0.277883	Sedation	14	Admitted
AA	69		180	Alderaan		No	0.444051	Sedation	14	Admitted
BB	32		90	Alderaan		No	0.435682	Other	15	Admitted
BB	32		180	Alderaan		No	0.153481	Other	15	Admitted
CC	33		90	Tatooine		No	0.227809	Sedation	14	Admitted
CC	33		180	Tatooine		No	0.047977	Sedation	Untested	Admitted
DD	65		90	Alderaan		Yes	0.637912	Paralysis	Untested	Admitted
DD	65		180	Alderaan		Yes	0.829183	Paralysis	9	Admitted
EE	46		90	Tatooine		No	0.940982	Other	20	Admitted
EE	46		180	Tatooine		No	0.363848	Other	Untested	Admitted
FF	21		90	Alderaan		Yes	0.54009	Other	Untested	Admitted
FF	21		180	Alderaan		Yes	0.951708	Other	3	Admitted
GG	57		90	Alderaan		No	0.122168	Sedation	5	Admitted
GG	57		180	Alderaan		No	0.622198	Sedation	3	Admitted
HH	58		90	Tatooine		Yes	0.383116	Paralysis	4	Admitted
HH	58		180	Tatooine		Yes	0.628653	Paralysis	2	Admitted
II	54		90	Alderaan		No	0.188853	Sedation	14	Admitted
II	54		180	Alderaan		No	0.788362	Sedation	16	Admitted
JJ	36		90	Alderaan		Yes	0.430876	Other	17	Admitted

Figure 7. A second FITBIR sample dataset.



FakeStudy	FakeStudy	FakeStudy	FakeStudy	FakeStudy	FakeStudy	FakeStudy	FakeStudy3.Info.WAISProcessSpdIndxPercntRank
AA	69	180	Blue Pill		0.23	1	
			Red Pill				
			Yellow Pill				
			Green Pill				
			Brown Pill				
			Pink Pill				
			Polka-dotted Pill				
BB	32	180	Blue Pill		0.46		2.34
			Red Pill				
			Yellow Pill				
			Green Pill				
			Brown Pill				
			Pink Pill				
			Polka-dotted Pill				
CC	33	180	Blue Pill		0.37	0	4.56
			Red Pill				
			Pink Pill				
			Brown Pill				
DD	65	180	Blue Pill		0.87	1	
			Red Pill				
			Pink Pill				
			Brown Pill				
EE	46	180	Blue Pill		0.64	0	9.83

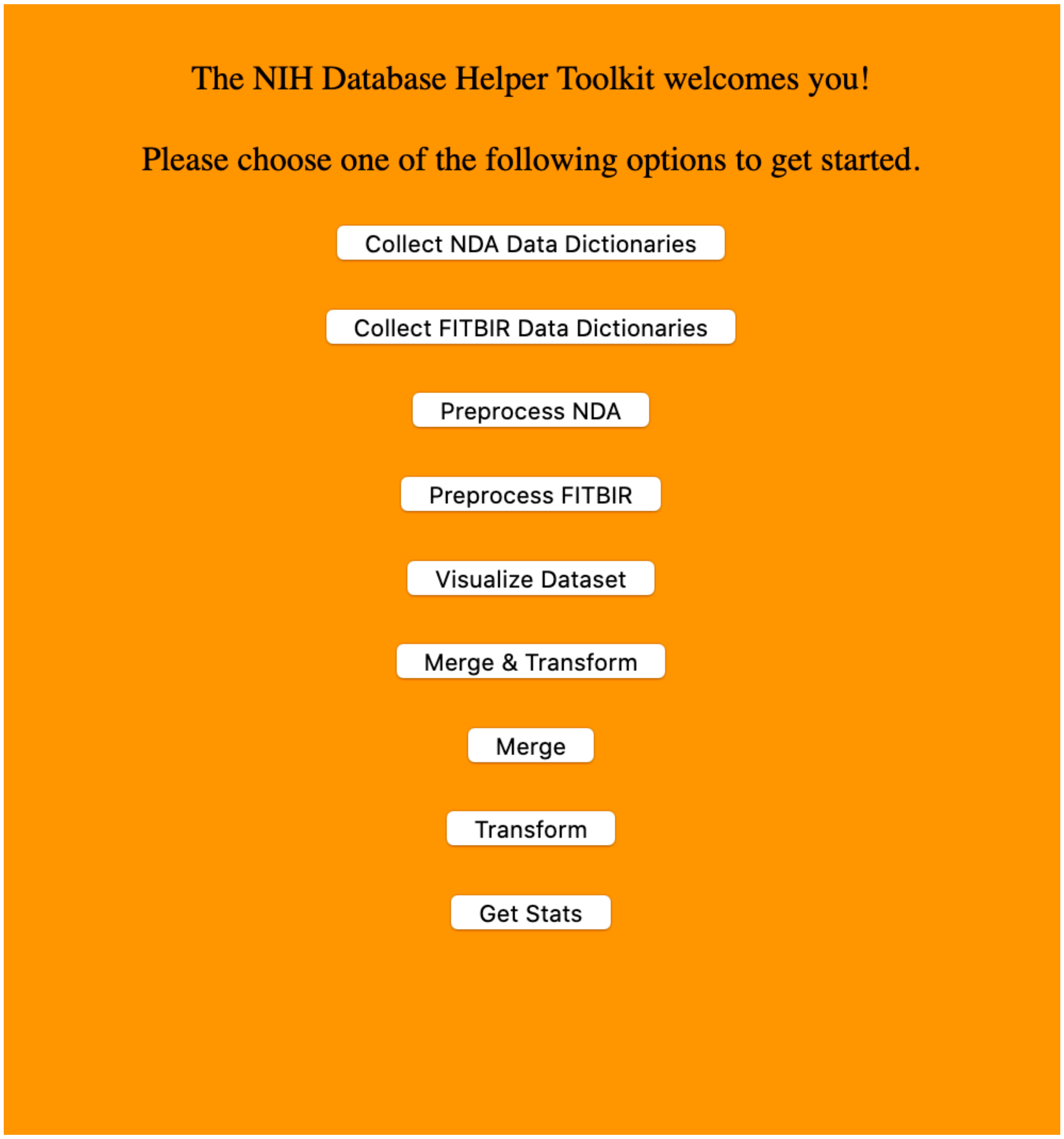
Figure 8. An example of a fake FITBIR dataset with a similar format to unflattened CSV files. The last part of the column names match an actual data element in FITBIR so the user can practice scraping data dictionaries from FITBIR.

All files that the application processes and returns will be saved to the *Outputs* folder.

Features of the Application

All files downloaded from NDA and FITBIR contain the name of the form to which they correspond. The Data Dictionary collection pages of this application rely on the downloaded dataset names to know which dictionaries to collect! Do not change the names until after you are done collecting the relevant data dictionaries.

Main Page of the Application

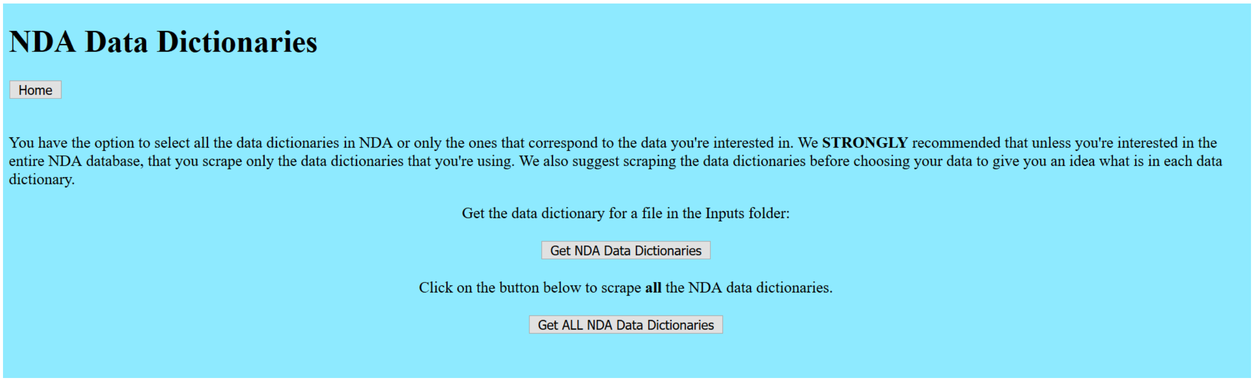

The NIH Database Helper Toolkit welcomes you!

Please choose one of the following options to get started.

Figure 9. The main page for the web application. Each button takes the user to a different function in the application. The buttons are displayed in the order we recommend users to use the options.

The main page displays all the options the users have for working with their data. Each option describes its function. The top-down order of the options should be similar to the order the user would expect to work with their data.

Collect NDA Data Dictionaries



NDA Data Dictionaries

You have the option to select all the data dictionaries in NDA or only the ones that correspond to the data you're interested in. We **STRONGLY** recommended that unless you're interested in the entire NDA database, that you scrape only the data dictionaries that you're using. We also suggest scraping the data dictionaries before choosing your data to give you an idea what is in each data dictionary.

Get the data dictionary for a file in the Inputs folder:

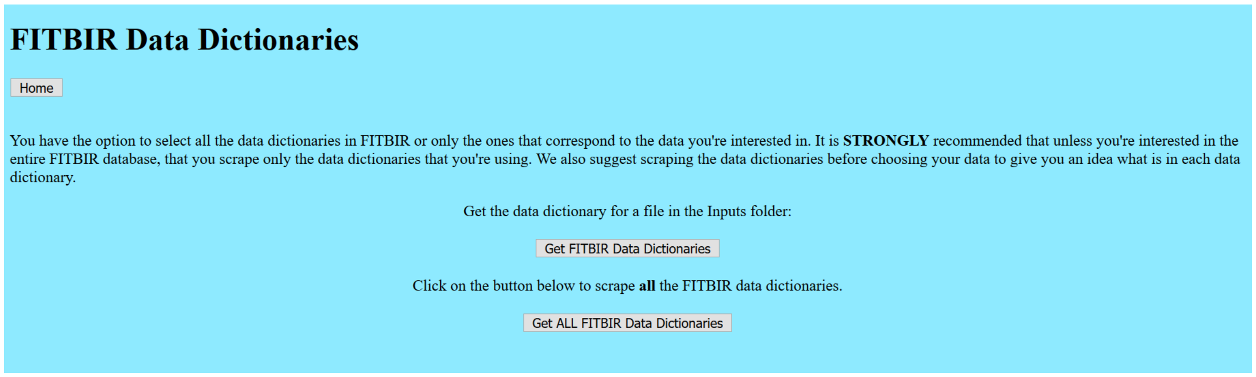
Click on the button below to scrape **all** the NDA data dictionaries.

Figure 10. The page for the collecting NDA data dictionaries for all files in the *Inputs* folder. There are options to collect only the data dictionaries that correspond to the files in the *Inputs* folder and an option to collect all the data dictionaries in NDA. The second option takes more time and is useful if the user is studying the whole of NDA.

The underlying script here takes advantage of the Python API that NDA provides. Using that API, we scrape the data dictionaries, which are publicly available, that match the names of the data text files. These text files are the forms used in a study (e.g. “Beck Symptoms Inventory”). The variables are the individual *data elements* in each form (e.g. “feels sad”). From there, the script combines all the data dictionaries into one big data dictionary, with information about the variable name, its description, type, and possible values. The possible values are the values that the standard data element can have, as agreed upon by NDA, but do not necessarily reflect the actual values in your dataset. For example, the data dictionary could say that the values [1,2,3,4,5] are possible, but your dataset only has people who have values [1,2,4].

In the application, we have tried to account for when some datasets are not public. If a data dictionary cannot be found, the application will return an error. Check to make sure your files can be found in NDA’s Data Dictionary search tool.

Collect FITBIR Data Dictionaries


FITBIR Data Dictionaries

Home

You have the option to select all the data dictionaries in FITBIR or only the ones that correspond to the data you're interested in. It is **STRONGLY** recommended that unless you're interested in the entire FITBIR database, that you scrape only the data dictionaries that you're using. We also suggest scraping the data dictionaries before choosing your data to give you an idea what is in each data dictionary.

Get the data dictionary for a file in the Inputs folder:

Get FITBIR Data Dictionaries

Click on the button below to scrape **all** the FITBIR data dictionaries.

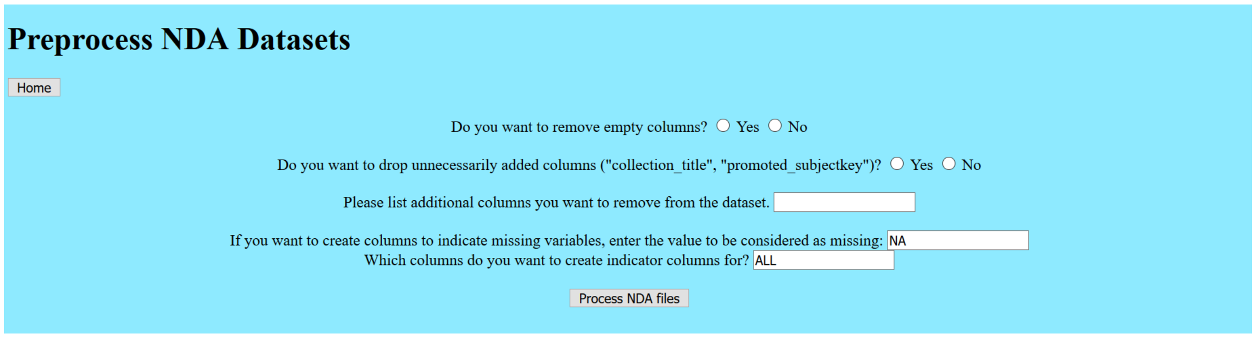
Get ALL FITBIR Data Dictionaries

Figure 11. The page for the collecting FITBIR data dictionaries for all files in the *Inputs* folder. There are options to collect only the data dictionaries that correspond to the files in the *Inputs* folder and an option to collect all the data dictionaries in FITBIR. The second option takes more time and is useful if the user is studying the whole of FITBIR.

In principle, the underlying script is the same as the one that collects NDA data dictionaries. The big difference is that FITBIR does not have a Python API that facilitates web scraping. Additionally, FITBIR data dictionaries (like NDA dictionaries) do not display statistics for each study in which a form was used in. This means that the value range given in FITBIR data dictionaries denotes the *possible values* that could be in this data element. A quick note about terminology: *data files* refers to the collection of forms used in a study. *forms* refer to the different assessments used in a study (i.e. Beck's Depression Inventory, PTSD Checklist, PHQ-9, etc.) while *data elements* refer to the variables measured in each assessment (i.e. "feels sad", "feels alone", etc.).

The data dictionaries can be merged with the "Get Stats" output file to create a more complete data dictionary for the user's specific dataset.

Preprocess NDA


Preprocess NDA Datasets

Home

Do you want to remove empty columns? ○ Yes ○ No

Do you want to drop unnecessarily added columns ("collection_title", "promoted_subjectkey")? ○ Yes ○ No

Please list additional columns you want to remove from the dataset.

If you want to create columns to indicate missing variables, enter the value to be considered as missing: NA

Which columns do you want to create indicator columns for? ALL

Process NDA files

Figure 12. The page for the user to decide how to process their NDA datasets.

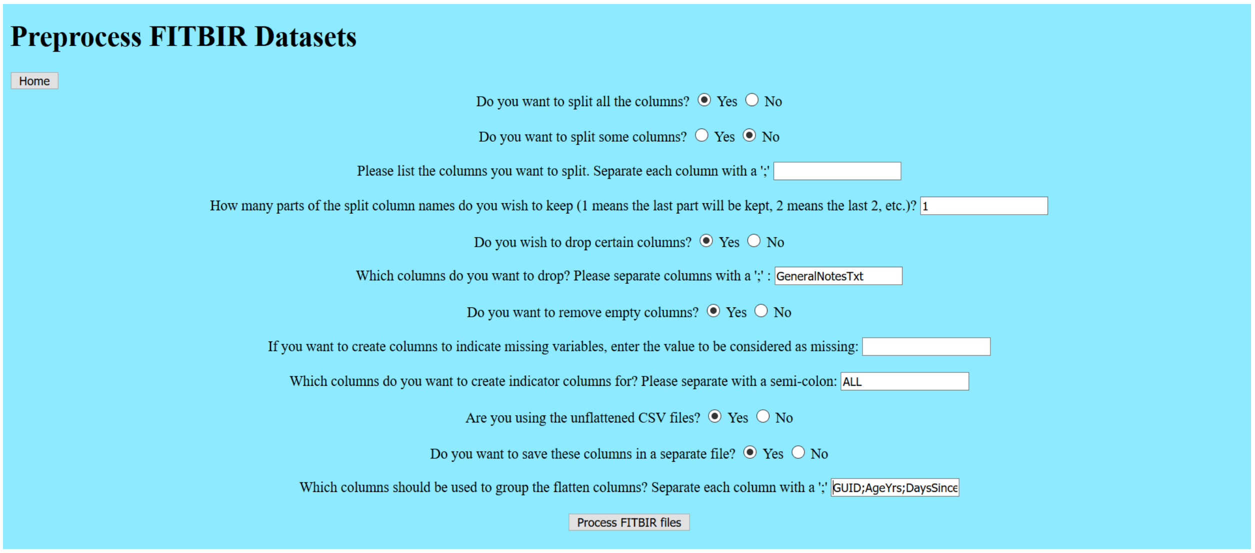
Options to process NDA datafiles are:

1. Remove empty columns
2. Drop spurious columns and other columns
3. Create missing data indicator columns and for which columns.

The default missing data indicator value is an empty string (or an empty cell in Excel). If you want to tell the application that another value indicates missingness, enter that value or a list of values separated by a ‘;’. Examples of other values that indicate missingness are -777, -999, NA, NaN (as a string), etc.

Because NDA datafiles have the first row of each file as metadata, the script automatically removes them and saves those to separate files in the *Outputs* folder.

Preprocess FITBIR


Preprocess FITBIR Datasets

Home

Do you want to split all the columns? [✓ Yes] No

Do you want to split some columns? Yes [✓ No]

Please list the columns you want to split. Separate each column with a ';'

How many parts of the split column names do you wish to keep (1 means the last part will be kept, 2 means the last 2, etc.)? 1

Do you wish to drop certain columns? [✓ Yes] No

Which columns do you want to drop? Please separate columns with a ';' : GeneralNotesTxt

Do you want to remove empty columns? [✓ Yes] No

If you want to create columns to indicate missing variables, enter the value to be considered as missing:

Which columns do you want to create indicator columns for? Please separate with a semi-colon: ALL

Are you using the unflattened CSV files? [✓ Yes] No

Do you want to save these columns in a separate file? [✓ Yes] No

Which columns should be used to group the flatten columns? Separate each column with a ';' GUID;AgeYrs;DaysSince

Process FITBIR files

Figure 13. The page for the user to decide how they want to process their FITBIR files.

Options to process FITBIR files are:

1. Split the column names by period and which portion of the column name should be retained
2. Dropping specified columns. These are separated by a ‘;’.
3. Remove all empty columns.
4. Create missing value indicator columns for specific variables.
5. Fix unflattened files (by either removing the offending columns and saving them to a separate file, or by joining their values into one cell separated by a ‘;’.

Splitting the column names means separating the column names by period, so each name will be split into three parts. The option for how many parts of the name to keep tells the application to keep the last *n* parts. For example, if one part is requested to be kept in the column name “FakeStudy.Info.GUID” , the new column name will be “GUID”. If two parts are requested, the new column name will be “Info.GUID”. If zero or more than 3 parts are requested, the new column name will be “FakeStudy.Info.GUID”.

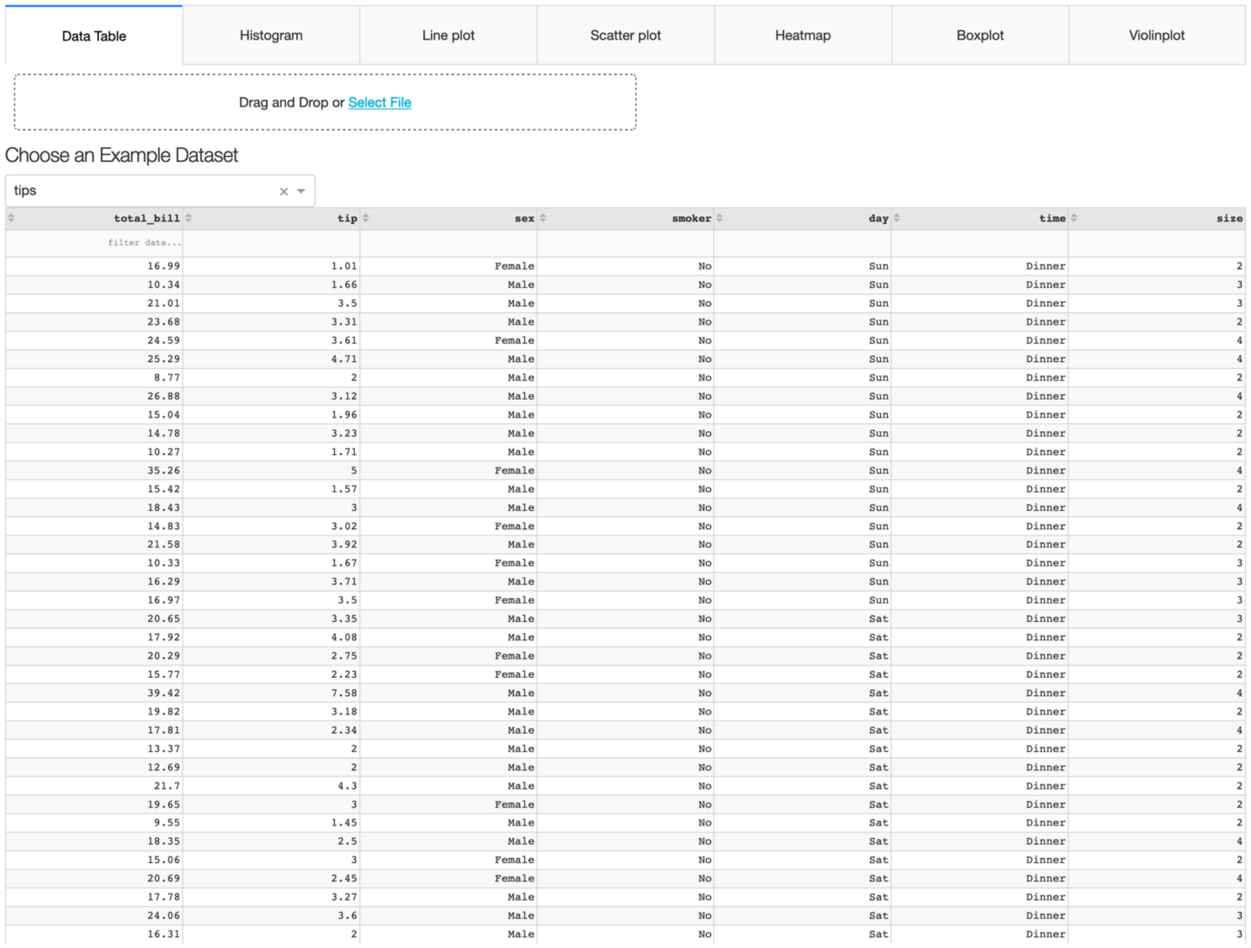
When downloading the datasets from FITBIR, users have the option to download the files as “flattened” or “unflattened”. We **strongly** recommend you always download the files as “flattened”. This option creates binary variables for each value in the list columns. It is a much cleaner method for working with the data. An example of an unflattened file can be found in Figure 6.

If, however, you are given an unflattened file, you can handle it in two ways:

1. Remove them and copy the GUID column and store them in a separate file.
2. Merge all the cells by group and combine the values in the list column into one string separating values by a ';'.

The "group columns by" columns should be columns that don't have more than one value per row (like the list column). A good example is "GUID;AgeYrs;GeneralNotesTxt", etc.

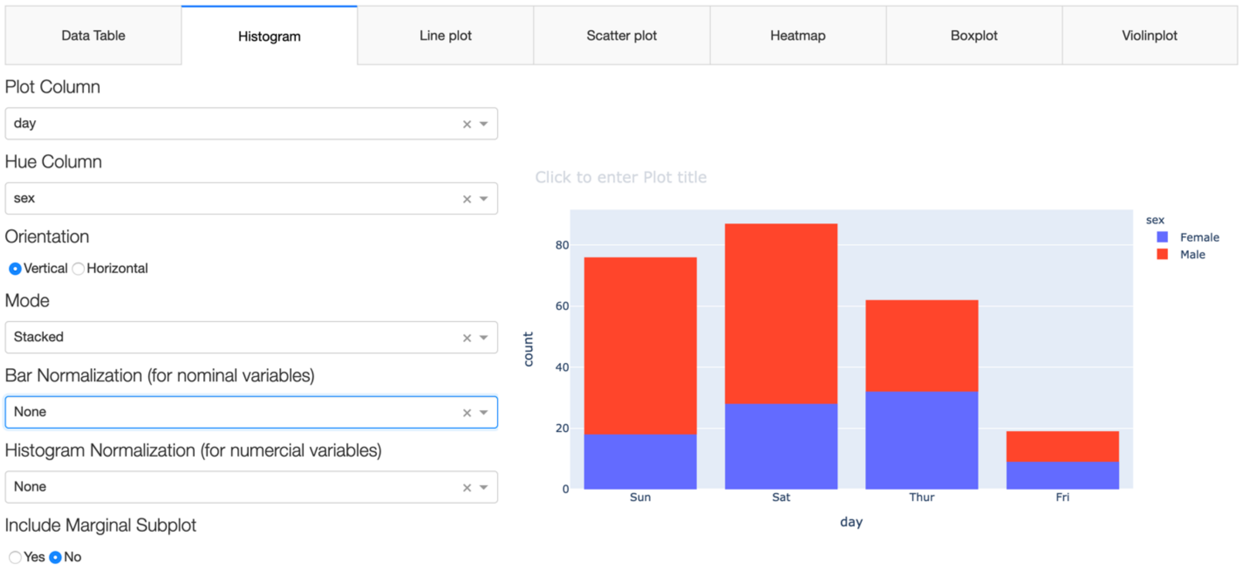
Plotting the Datasets (Visualize Dataset)



total_bill	tip	sex	smoker	day	time	size
16.99	1.01	Female	No	Sun	Dinner	2
10.34	1.66	Male	No	Sun	Dinner	3
21.01	3.5	Male	No	Sun	Dinner	3
23.68	3.31	Male	No	Sun	Dinner	2
24.59	3.61	Female	No	Sun	Dinner	4
25.29	4.71	Male	No	Sun	Dinner	4
8.77	2	Male	No	Sun	Dinner	2
26.88	3.12	Male	No	Sun	Dinner	4
15.04	1.96	Male	No	Sun	Dinner	2
14.78	3.23	Male	No	Sun	Dinner	2
10.27	1.71	Male	No	Sun	Dinner	2
35.26	5	Female	No	Sun	Dinner	4
15.42	1.57	Male	No	Sun	Dinner	2
18.43	3	Male	No	Sun	Dinner	4
14.83	3.02	Female	No	Sun	Dinner	2
21.58	3.92	Male	No	Sun	Dinner	2
10.33	1.67	Female	No	Sun	Dinner	3
16.29	3.71	Male	No	Sun	Dinner	3
16.97	3.5	Female	No	Sun	Dinner	3
20.65	3.35	Male	No	Sat	Dinner	3
17.92	4.08	Male	No	Sat	Dinner	2
20.29	2.75	Female	No	Sat	Dinner	2
15.77	2.23	Female	No	Sat	Dinner	2
39.42	7.58	Male	No	Sat	Dinner	4
19.82	3.18	Male	No	Sat	Dinner	2
17.81	2.34	Male	No	Sat	Dinner	4
13.37	2	Male	No	Sat	Dinner	2
12.69	2	Male	No	Sat	Dinner	2
21.7	4.3	Male	No	Sat	Dinner	2
19.65	3	Female	No	Sat	Dinner	2
9.55	1.45	Male	No	Sat	Dinner	2
18.35	2.5	Male	No	Sat	Dinner	4
15.06	3	Female	No	Sat	Dinner	2
20.69	2.45	Female	No	Sat	Dinner	4
17.78	3.27	Male	No	Sat	Dinner	2
24.06	3.6	Male	No	Sat	Dinner	3
16.31	2	Male	No	Sat	Dinner	3

Figure 14. First page displaying the datasets in tabular form. The application provides 2 default datasets for the user to familiarize themselves with the capabilities of the visualization. The visualization page of the application no longer requires that your dataset be located in the *Inputs* folder. Instead, you can click on the “Select File” or drag and drop the file you want to look at into the dashed box. The row underneath the column names is for filtering the data. Text, numbers, and dates can be filtered. For example, you can only look at data from women by typing “Female” underneath “sex”. Using the command “= Female” (without quotes) also works. The command “contains” searches for all text values

that contain the specified substring (e.g. “contains Fem”). All columns can be filtered using “=, >, <, != (not equal), >=, <=”. Datetime columns take the format of YYYY-MM-DD (e.g. “< 2020-01” or “= 2020-01-01”). Filtering the data table does not change the dataset that is used for plots.



Plot Column

day

Hue Column

sex

Orientation

[✓ Vertical] Horizontal

Mode

Stacked

Bar Normalization (for nominal variables)

None

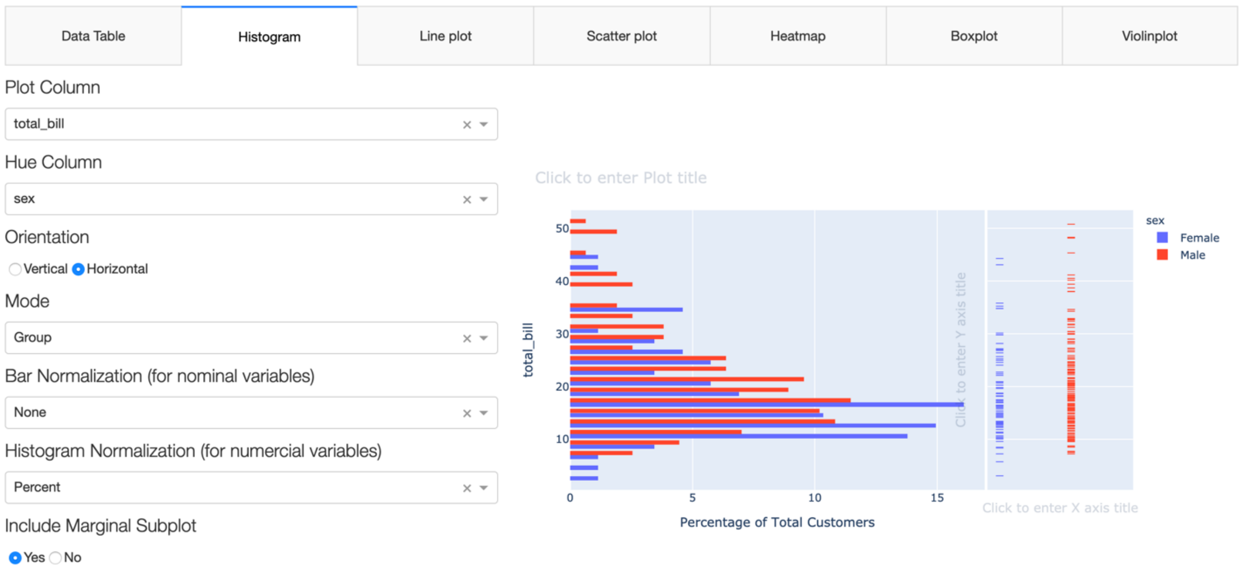
Histogram Normalization (for numercial variables)

None

Include Marginal Subplot

Yes [✓ No]

Figure 15. Example of a histogram for a nominal/categorical variable, plotting the count of the days that people tipped on, colored by “sex”. The application can produce vertical (shown here) or horizontal histograms/barplots. The counts are stacked by color (Mode can be stacked, overlay (default), or group) and not normalized (Bar Normalization can be none, fraction or percent).



Data Table
Histogram
Line plot
Scatter plot
Heatmap
Boxplot
Violinplot
Click to enter Plot title
total_bill
Percentage of Total Customers
Click to enter Y axis title
Click to enter X axis title
0
5
10
15
10
20
30
40
50
sex
🟦 Female
🟥 Male

Plot Column

total_bill

Hue Column

sex

Orientation

Vertical [✓ Horizontal]

Mode

Group

Bar Normalization (for nominal variables)

None

Histogram Normalization (for numercial variables)

Percent

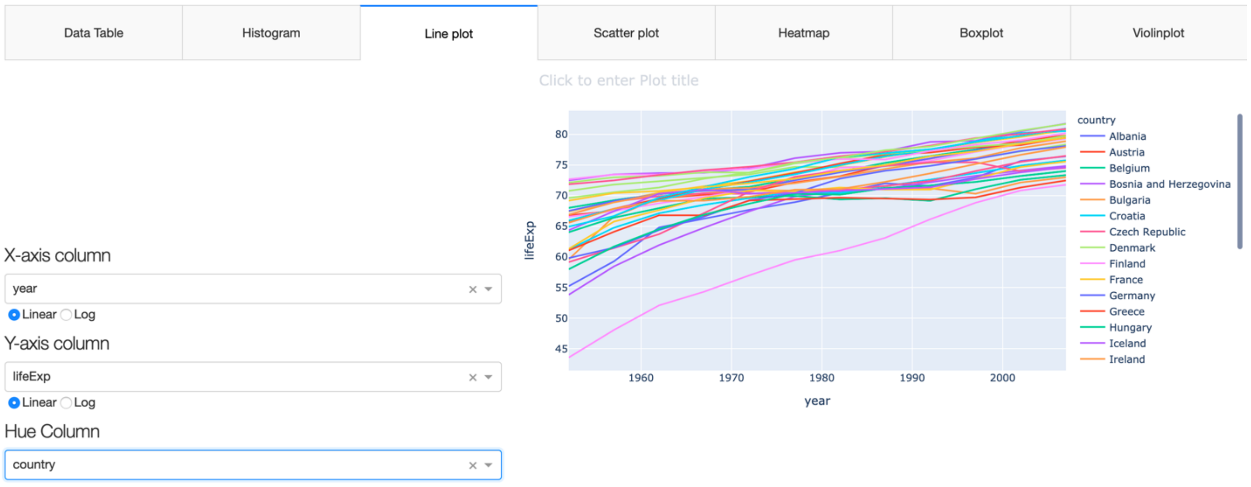
Include Marginal Subplot

[✓ Yes] No

Figure 16. Example of a histogram for a continuous variable. Here the “total_bill” distribution is plotted and grouped by “sex”. Instead of plotting the count for each bin, which are automatically calculated, the percent is chosen. A marginal subplot (also called a “rug plot”) is plotted above, using marks to show the

distribution of the data. The user also has the option to rename the axes labels and the title of the plot by clicking on the current text.

The histograms will only plot non-missing data. If your data has a column that has all its values missing, an error will be thrown.



X-axis column: year

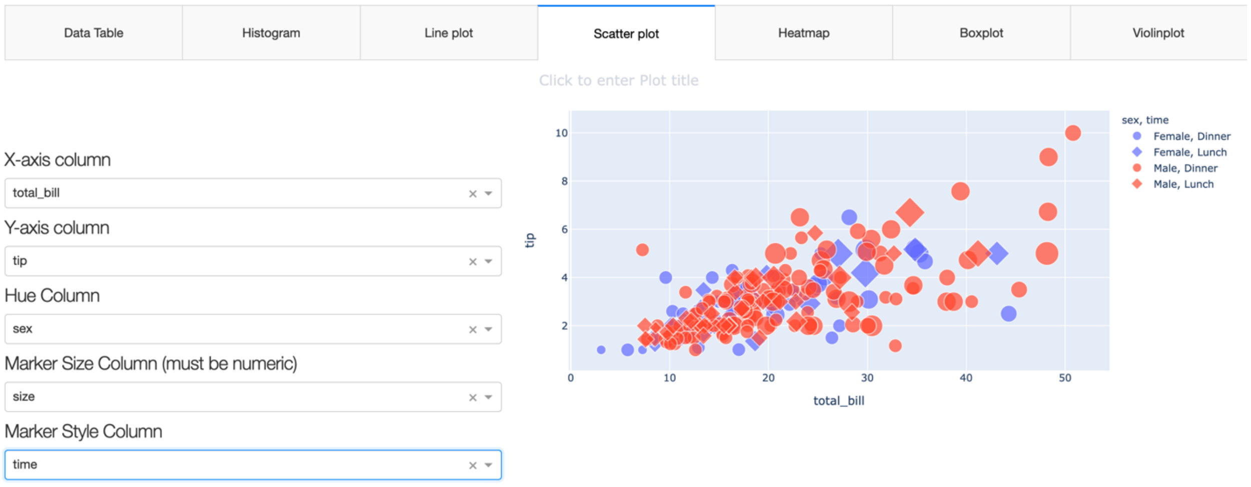
[✓ Linear] Log

Y-axis column: lifeExp

[✓ Linear] Log

Hue Column: country

Figure 17. Example of a lineplot using the “gapminder” dataset (see the Data Table example datasets). Plots two continuous variables to search for trends.



X-axis column: total_bill

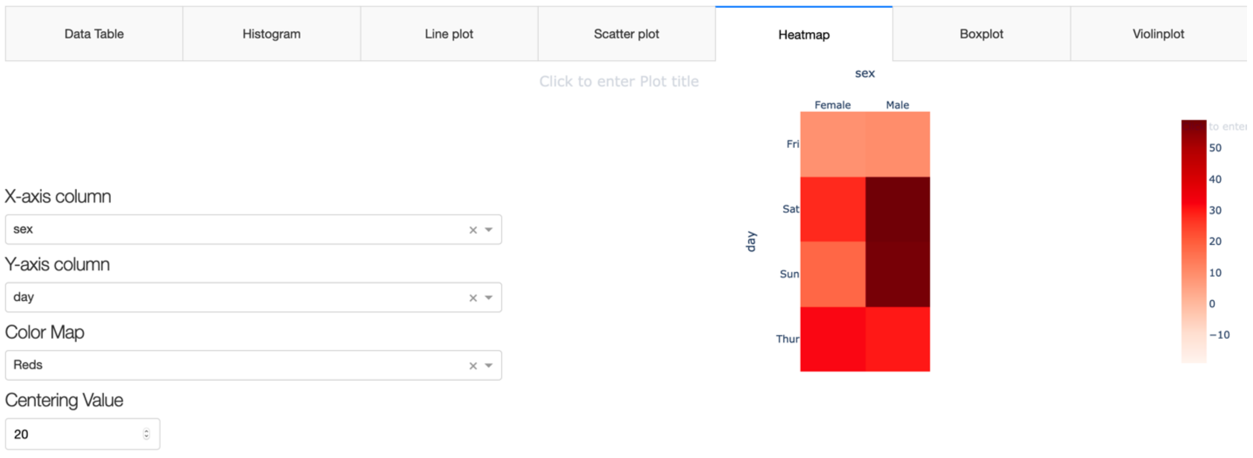
Y-axis column: tip

Hue Column: sex

Marker Size Column (must be numeric): size

Marker Style Column: time

Figure 18. Example of a scatter plot. Markers can be colored by the hue column, vary in sizes specified in the dataset, and stylized.



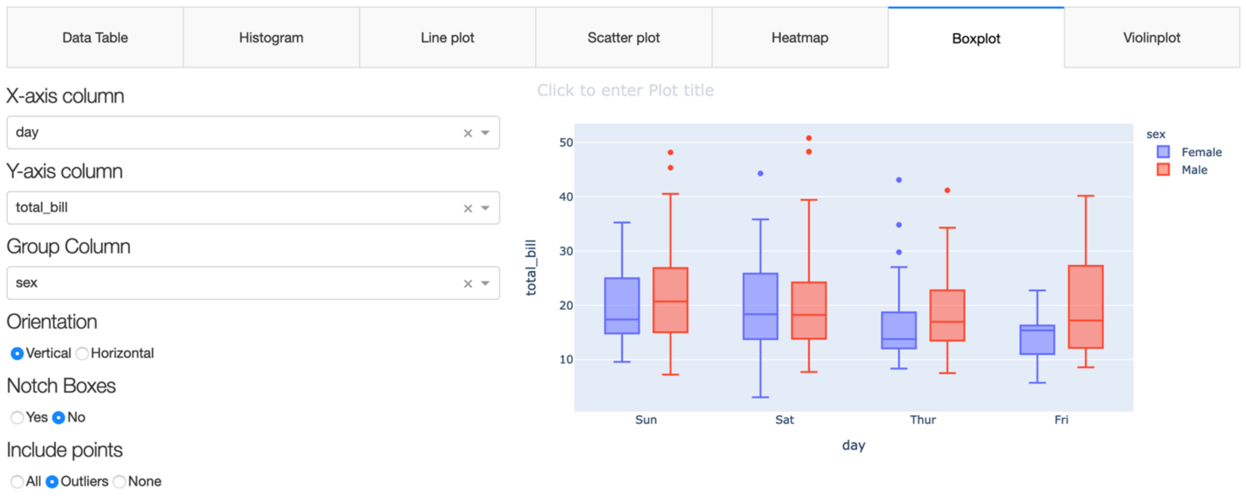
X-axis column: sex

Y-axis column: day

Color Map: Reds

Centering Value: 20

Figure 19. Example of a heatmap. Best for plotting categorical variables but numerical variables can be used as well. Multiple coloring options are included and the option to center the color scale is called "Centering Value".



X-axis column: day

Y-axis column: total_bill

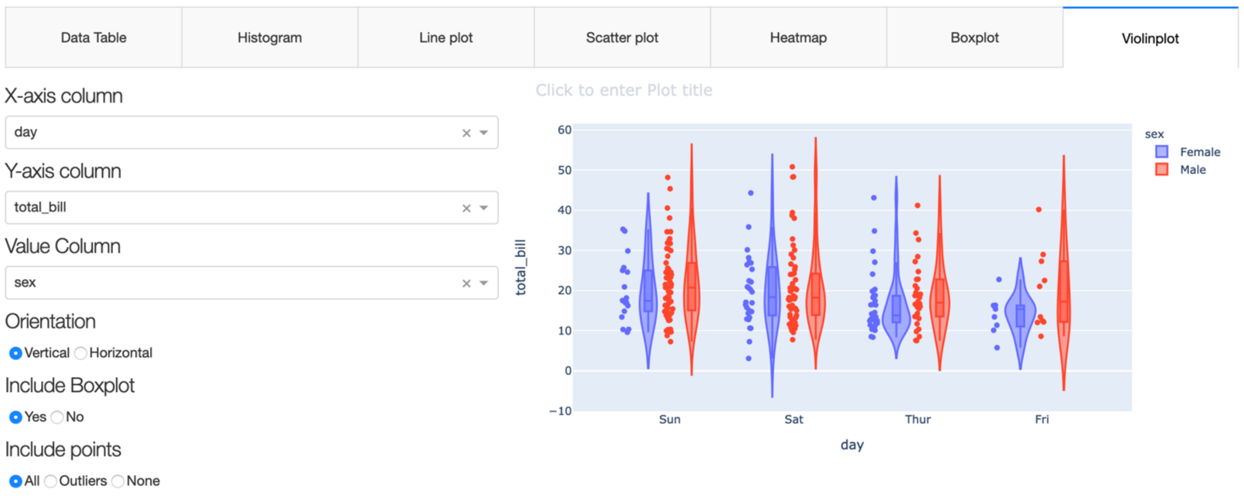
Group Column: sex

Orientation: [✓ Vertical] Horizontal

Notch Boxes: Yes [✓ No]

Include points: All [✓ Outliers] None

Figure 20. Example of the boxplot feature for visualizing data. These can be oriented vertically (shown) or horizontally. The columns will need to be adjusted appropriately. There is also the option to exclude outliers, show outliers, or show all data as points in addition to the boxplot.



X-axis column: day

Y-axis column: total_bill

Value Column: sex

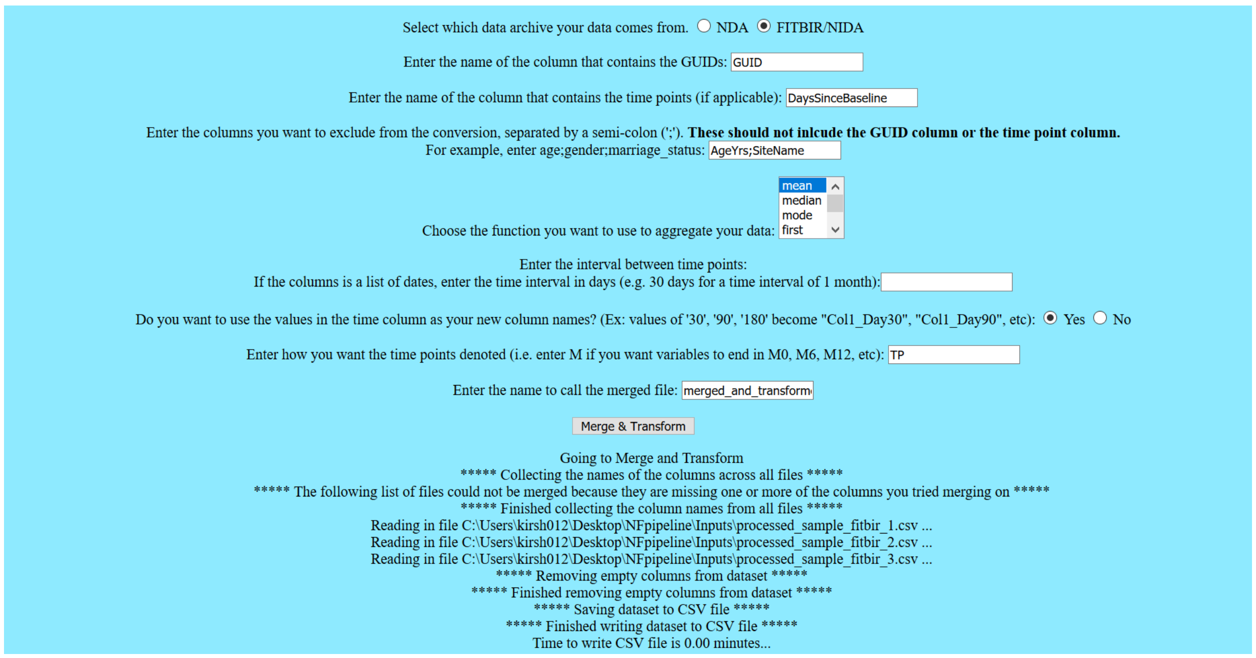
Orientation: [✓ Vertical] Horizontal

Include Boxplot: [✓ Yes] No

Include points: [✓ All] Outliers None

Figure 21. Example of the violin plot feature for visualizing data. These can be oriented vertically (shown) or horizontally. The columns will need to be adjusted appropriately. There is also the option to exclude outliers, show outliers, or show all data as points in addition to a boxplot.

Merge & Transform



Select which data archive your data comes from. NDA [✓ FITBIR/NIDA]

Enter the name of the column that contains the GUIDs: GUID

Enter the name of the column that contains the time points (if applicable): DaysSinceBaseline

Enter the columns you want to exclude from the conversion, separated by a semi-colon (';'). **These should not inlcude the GUID column or the time point column.**
For example, enter age;gender;marriage_status: AgeYrs;SiteName

Choose the function you want to use to aggregate your data: [✓ mean] median mode first

Enter the interval between time points:
If the columns is a list of dates, enter the time interval in days (e.g. 30 days for a time interval of 1 month):

Do you want to use the values in the time column as your new column names? (Ex: values of '30', '90', '180' become "Col1_Day30", "Col1_Day90", etc): [✓ Yes] No

Enter how you want the time points denoted (i.e. enter M if you want variables to end in M0, M6, M12, etc): TP

Enter the name to call the merged file: merged_and_transform

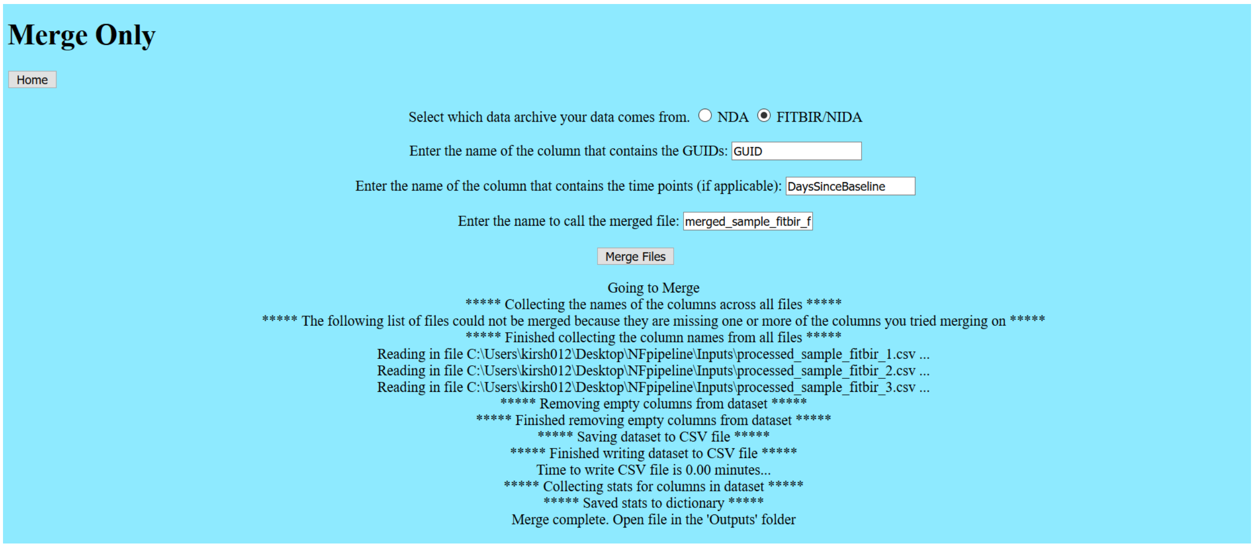
```
Going to Merge and Transform
***** Collecting the names of the columns across all files *****
***** The following list of files could not be merged because they are missing one or more of the columns you tried merging on *****
***** Finished collecting the column names from all files *****
Reading in file C:\Users\kirsh012\Desktop\NFpipeline\Inputs\processed_sample_fitbir_1.csv ...
Reading in file C:\Users\kirsh012\Desktop\NFpipeline\Inputs\processed_sample_fitbir_2.csv ...
Reading in file C:\Users\kirsh012\Desktop\NFpipeline\Inputs\processed_sample_fitbir_3.csv ...
***** Removing empty columns from dataset *****
***** Finished removing empty columns from dataset *****
***** Saving dataset to CSV file *****
***** Finished writing dataset to CSV file *****
Time to write CSV file is 0.00 minutes...
```

Figure 22. The page that prompts the user for the various inputs necessary to merge and convert files from longitudinal to wide format.

The merge and transform page merges all the files in the *Inputs* folder and then converts the resultant dataset from longitudinal to wide format. This is a streamlined process for users who know that they want to merge and transform all their data files and how they want to do it. However, each time the files are merged first and then transformed. If the user wants to transform first and then merge, they can run the ‘Transform’ option first from the main page and then run the ‘Merge’ option.

See the ‘Merge’ and ‘Transform’ options for more details.

Merge



Select which data archive your data comes from. NDA [● FITBIR/NIDA]

Enter the name of the column that contains the GUIDs: GUID

Enter the name of the column that contains the time points (if applicable): DaysSinceBaseline

Enter the name to call the merged file: merged_sample_fitbir_f

```
Going to Merge
***** Collecting the names of the columns across all files *****
***** The following list of files could not be merged because they are missing one or more of the columns you tried merging on *****
***** Finished collecting the column names from all files *****
Reading in file C:\Users\kirsh012\Desktop\NFpipeline\Inputs\processed_sample_fitbir_1.csv ...
Reading in file C:\Users\kirsh012\Desktop\NFpipeline\Inputs\processed_sample_fitbir_2.csv ...
Reading in file C:\Users\kirsh012\Desktop\NFpipeline\Inputs\processed_sample_fitbir_3.csv ...
***** Removing empty columns from dataset *****
***** Finished removing empty columns from dataset *****
***** Saving dataset to CSV file *****
***** Finished writing dataset to CSV file *****
Time to write CSV file is 0.00 minutes...
***** Collecting stats for columns in dataset *****
***** Saved stats to dictionary *****
Merge complete. Open file in the 'Outputs' folder
```

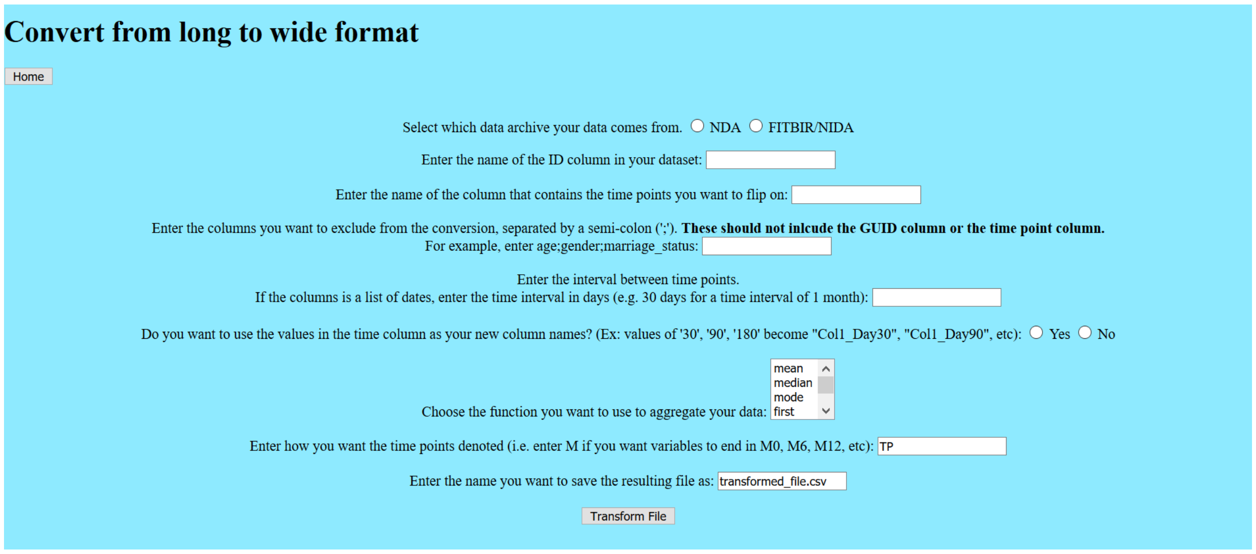
Figure 23. The page prompts the user for the inputs necessary to merge files.

The merge option runs the same merging script that “merge and transform” uses, the only difference is that this script will output the file merged file. The files are automatically read into the script from the *Inputs* folder; the user does not need to select the files in the application.

The ‘time’ column does not have to be a ‘Time’ column (likewise the ‘GUID’ column doesn’t need to be a ‘GUID’ columns), it can be all column that you want to use to direct the merge.

If your time column has empty strings either: remove the rows with empty strings, find another column to use as your time column, or figure out what dates those empty strings indicate, or just merge on the GUID column and leave the time column option to the default value. The default value tells the application to just merge on the first column.

Transform



Convert from long to wide format

Home

Select which data archive your data comes from. ○ NDA ○ FITBIR/NIDA

Enter the name of the ID column in your dataset:

Enter the name of the column that contains the time points you want to flip on:

Enter the columns you want to exclude from the conversion, separated by a semi-colon (';'). **These should not inlcude the GUID column or the time point column.**
For example, enter age;gender;marriage_status:

Enter the interval between time points.
If the columns is a list of dates, enter the time interval in days (e.g. 30 days for a time interval of 1 month):

Do you want to use the values in the time column as your new column names? (Ex: values of '30', '90', '180' become "Col1_Day30", "Col1_Day90", etc): ○ Yes ○ No

Choose the function you want to use to aggregate your data: mean / median / mode / first

Enter how you want the time points denoted (i.e. enter M if you want variables to end in M0, M6, M12, etc): TP

Enter the name you want to save the resulting file as: transformed_file.csv

Transform File

Figure 24. The page that prompts the user for the necessary inputs to convert a longitudinal dataset to a wide dataset (only for NDA and FITBIR. NIDA is another beast).

The conversion from longitudinal to wide format requires that a column that contains the time points will be used. In NDA and FITBIR (and NIDA), there are three types of ‘time’ columns: dates (like 4/3/2017), strings (‘3 months’, ‘6 months’), or numbers (30, 90, 180). The application can handle all of these. For the dates, the application converts to days from the earliest time point by default. If you have a time column with specific dates (e.g. ‘03/14/15’), sort this column from earliest to latest.

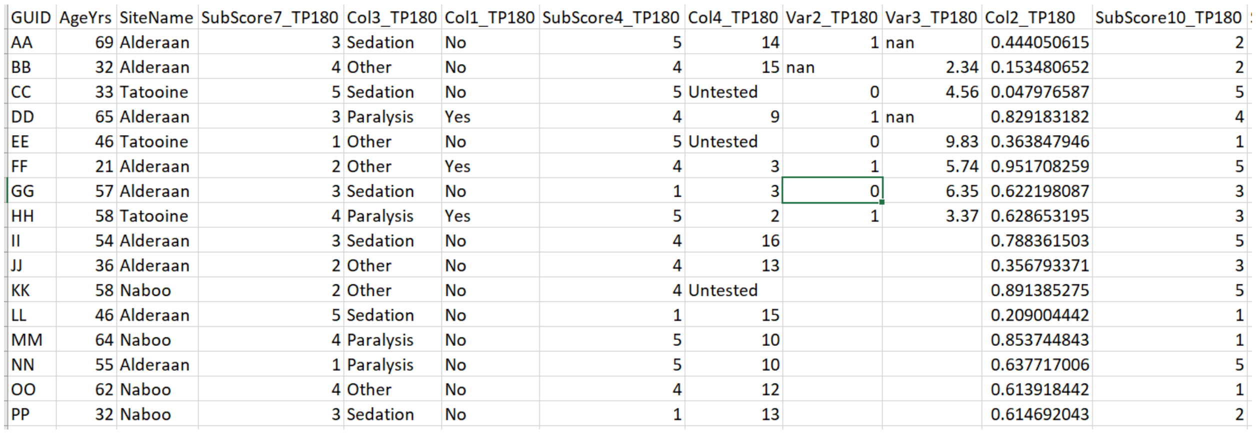
If you want the output to be in months or year or weeks, you can enter a number for the interval between time points in days (i.e. 30 days for intervals of a month), and the application will divide the raw days by the interval to convert times. To illustrate, imagine your starting date is 3/14/15 and the second time point is 6/14/15. The time interval is 3 months. If you enter the time interval as 30 (for 30 days) and a prefix of “M” (for month), the application will add “M3” to the column names. If you instead enter 90, the application will add “M1” to the column names at the second time point, because these times are the “M1” times, times before them are “M0”.

For columns with strings and/or numbers, the application can use those values as the new times. Note that this requires the user to know these things about their data prior to converting.

If your time column has empty strings either: remove the rows with empty strings, find another column to use as your time column, or figure out what dates those empty strings indicate. The application will not transform the way you expect if your column has empty strings.

Because there could be multiple measurements made at the same time point (for whatever reason), the application can aggregate over these measurements in various standard ways: using the mean, median, mode, first value, last value, or no aggregation (use only if you have one measurement per time point). The aggregation will be used to return one value per GUID.

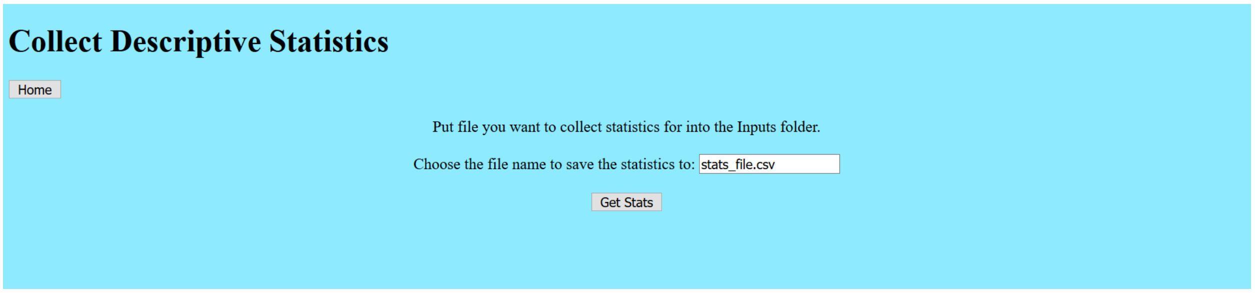
Finally, the user can enter a prefix to denote the different time points. The default prefix is “TP”, so “TP” plus the number indicating the time will be added on to the end of each column name.



GUID	AgeYrs	SiteName	SubScore7_TP180	Col3_TP180	Col1_TP180	SubScore4_TP180	Col4_TP180	Var2_TP180	Var3_TP180	Col2_TP180	SubScore10_TP180
AA	69	Alderaan	3	Sedation	No	5	14	1	nan	0.444050615	2
BB	32	Alderaan	4	Other	No	4	15	nan	2.34	0.153480652	2
CC	33	Tatooine	5	Sedation	No	5	Untested	0	4.56	0.047976587	5
DD	65	Alderaan	3	Paralysis	Yes	4	9	1	nan	0.829183182	4
EE	46	Tatooine	1	Other	No	5	Untested	0	9.83	0.363847946	1
FF	21	Alderaan	2	Other	Yes	4	3	1	5.74	0.951708259	5
GG	57	Alderaan	3	Sedation	No	1	3	0	6.35	0.622198087	3
HH	58	Tatooine	4	Paralysis	Yes	5	2	1	3.37	0.628653195	3
II	54	Alderaan	3	Sedation	No	4	16			0.788361503	5
JJ	36	Alderaan	2	Other	No	4	13			0.356793371	3
KK	58	Naboo	2	Other	No	4	Untested			0.891385275	5
LL	46	Alderaan	5	Sedation	No	1	15			0.209004442	1
MM	64	Naboo	4	Paralysis	No	5	10			0.853744843	1
NN	55	Alderaan	1	Paralysis	No	5	10			0.637717006	5
OO	62	Naboo	4	Other	No	4	12			0.613918442	1
PP	32	Naboo	3	Sedation	No	1	13			0.614692043	2

Figure 25. Example output after using the transform options in the previous figure and the merged dataset from the ‘merge’ example. The example prefix is the default “TP”. The columns “AgeYrs” and “SiteName” were excluded from the transformation because those variables do not change over time.

Get Stats



Collect Descriptive Statistics

Home

Put file you want to collect statistics for into the Inputs folder.

Choose the file name to save the statistics to: stats_file.csv

Get Stats

Figure 26. The page that prompts the user to decide what they want to call the file that collects various statistics for a single dataset in the *Inputs* folder.

The statistics that are calculated per column are:

- % missing
- # unique values
- Mean
- Median
- Min
- Max
- Mode
- Variance
- Standard Deviation
- 5th percentile
- 95th percentile
- Skewness
- Kurtosis

- Value Range (for up to 10 unique values)

Metrics

Source	Method	No. of Files	File Sizes	Time
Sample FITBIR datasets	Merge	3	3 KB, 1 KB, 1 KB	< 1 second
Sample FITBIR datasets	Transform	1	6 KB	< 1 second
Sample NDA datasets	Merge	3	2 KB, 5 KB, 8 KB	< 1 second
Sample NDA datasets	Transform	1	15 KB	<1 second
ABCD datasets	Merge	3	90 MB, 40 MB, 40 MB	51 minutes
Suicidality datasets	Merge	11	113 KB - 34 MB	19 minutes
Suicidality datasets	Transform	1	267 MB	13 minutes (output file size 63 KB)

Upcoming Features

- Concatenate files in the "Merge" section
- The bar, box, and violin plots will sorted by the count, alphabetically, and (for the box and violin plots, mean and median) of the categories is ascending and descending order
- Transform from wide format to longitudinal format

Known Bugs

- Data Table can't read comma-separated .txt files
- File names displayed in the "Merge", "Transform", and "Merge & Transform" sections are too long
- The MacOS version needs to be on Catalina 10.15.7 to use the application
- PC runs on Windows 10

Other Notes:

This application is not intended for use with imaging data.

You can merge using 1-2 columns as identifiers at a time. While the purpose of this application is to facilitate data mining and processing data from NDA, FITBIR, and NIDA, you can use datasets from other sources and use this application on them in the same way. For other datasets (or even NDA, FITBIR, or NIDA), your "GUID" column doesn't *have* to be an actual GUID column, it can be whatever column you want to merge on. Same with the "time" column for merging. For transforming, your "time" column does need to actually be a "time" column and match formatting with NDA or FITBIR.

If you use a dataset that is not from NDA, FITBIR, or NIDA, that is comma-separated (CSV), enter the information into the application as though it were a FITBIR/NIDA dataset.