Documentation for csvtool: for indexing, slicing, analysing, splitting and joining CSV

This program will be used by us to analyse the CSV/TSV files like mean, avg, median, sum, max/min, cardinality and other stats. Also will allow us to modify the CSV/TSV files like re-ordering the columns, and rows, slicing, joining, splitting etc.

Statistics: Sample statistics config file

```
# Sample statistics config file for CSV/TSV statistics
# Mean, Max, Min, Median, Average, Sum, max_length, min_length, Mode, Cardinality etc.

# Configuration type
# Defines the type of this config file.
# WARNING: Please don't use different type with different config files
# As it may cause the program to not work properly
# Use the type 'statistics' to obtain stats of CSV
TYPE="statistics"

# Path of the CSV/TSV file
PATH_TO_CSV="/mnt/z/Intern_proj/iits_bsp/influxdb_py/sample_data.csv"

# When set false, the first row will NOT be interpreted
# as column names. i.e., They will be included in statistics.
HEADERS=true

# The field delimiter for reading CSV data.
# Must be a single character. (default: ,)
DELIMITER=','

# Show all statistics available.
# If this set true, cardinality, mode, median, etc will be enabled as well.
EVERYTHING=true

# Show the cardinality, This requires storing all CSV data in memory.
CARDINALITY=false

# Show the mode, This requires storing all CSV data in memory.
MODE=false

# Show the median, This requires storing all CSV data in memory.
MEDIAN=false

# Include NULLs in the population size for computing mean and standard deviation.
NULLS=false

# The number of jobs to run in parallel.
# This works better when the given CSV data has
# an index already created. Note that a file handle
# is opened for each job.
# When set to '0', the number of jobs is set to the
# number of CPUs detected. [default: 0]
JOBS=

# If defined, Save the output result as text file.
PATH_TO_OUT="/mnt/z/Intern_proj/iits_bsp/tool/out.txt"
```

Fields

- TYPE: A String which denotes the work which needs to be done by the program, please don’t use different types with different defined work.
- PATH_TO_CSV: A path to the CSV/TSV file which will be used to generate stats.
- HEADERS: Takes a Boolean value, if set false then the first row will be interpreted as statistics, not as columns.
- DELIMITER: A char value which is the field delimiter for the CSV/TSV data.
- EVERYTHING: Show all statistics available. If this is set true, cardinality, mode, median, etc. will be enabled as well.
- CARDINALITY: Show the cardinality, this requires storing all CSV data in memory.
- MODE: Show the mode, this requires storing all CSV data in memory.
- MEDIAN: Show the median, this requires storing all CSV data in memory.
- NULLS: Include NULLs in the population size for computing mean and standard deviation.
- JOBS: The number of jobs to run in parallel. This works better when the given CSV data has an index already created. Note that a file handle is opened for each job. When set to '0', the number of jobs is set to the number of CPUs detected. [Default: 0]
- PATH_TO_OUT: If defined, Save the output result in a text file.

Selection: Sample select config file

```
# Sample Select config file for CSV/TSV statistics
# lets you manipulate the columns in CSV data. You can re-order
# them, duplicate them or drop them. Columns can be referenced by index or by
# name if there is a header row (duplicate column names can be disambiguated with
# more indexing). Finally, column ranges can be specified.

# Configuration type
# Defines the type of this config file.
# WARNING: Please don't use different type with different config files
# As it may cause the program to not work properly
# Use the type 'slicing' to re-order, duplicate them or drop them.
TYPE="select"

# Path of the CSV/TSV file
PATH_TO_CSV="/mnt/z/Intern_proj/iits_bsp/influxdb_py/sample_data.csv"

# When set false, the first row will NOT be interpreted
# as column names. i.e., They will be included in statistics.
HEADERS=true

# The field delimiter for reading CSV data.
# Must be a single character. (default: ,)
DELIMITER=','

# Select the columns to add in new csv
#
# Examples
#
# Select the first 4 columns (by index and by name):
# 1-4 or Header1-Header4
#
# Ignore the first 2 columns (by range and by omission):
# 3- or '!1-2'
#
# Select the third column named 'Foo':
# 'Foo[2]'
#
COLUMNS_TO_USE="1,2"

# If defined, Save the output result as text file.
PATH_TO_OUT="/mnt/z/Intern_proj/iits_bsp/tool/out.csv"
```

Fields

- TYPE: A String which denotes the work which needs to be done by the program, please don't use different types with different defined work.
- PATH_TO_CSV: A path to the CSV/TSV file which will be used to generate stats.
- HEADERS: Takes a Boolean value, if set false then the first row will be interpreted as statistics, not as columns.
- DELIMITER: A char value which is the field delimiter for the CSV/TSV data.
- COLUMNS_TO_USE: Selects the columns of the file which should be used.
    - Select the first 4 columns (by index and by name): 1-4 or Header1-Header4
    - Ignore the first 2 columns (by the range and by omission): 3- or '!1-2'
    - Select the third column named 'Foo': 'Foo[2]'
- PATH_TO_OUT: If defined, Save the output result as a new CSV file.

Note: Please refer to the online version of the documentation as I'll keep updating and adding new params in the sources and the docs will also be updated as per the same.

Link: https://github.com/radcolor/iit_bsp_intern_files/blob/master/documentation_for_csvtool_prog.markdown