NAME

AnalyzeTextFilesData.pl - Analyze numerical coulmn data in TextFile(s)

SYNOPSIS

AnalyzeTextFilesData.pl TextFile(s)...

AnalyzeTextFilesData.pl [-c, --colmode colnum | collabel] [--columns "colnum,[colnum,...]" | "collabel,[collabel,...]" | All] [--columnpairs "colnum,colnum,[colnum,colnum]..." | "collabel,collabel,[collabel,collabel]..." | AllPairs] [-d, --detail infolevel] [-f, --fast] [--frequencybins number | "number,number,[number,...]"] [-h, --help] [--indelim comma | semicolon] [--klargest number] [--ksmallest number] [-m, --mode DescriptiveStatisticsBasic | DescriptiveStatisticsAll | All | "function1, [function2,...]"] [-o, --overwrite] [--outdelim comma | tab | semicolon] [-p, --precision number] [-q, --quote yes | no] [-r, --root rootname] [--trimfraction number] [-w, --workingdir dirname] TextFiles(s)...

DESCRIPTION

Anaylze numerical column data in *TextFile(s)* using a combination of various statistical functions; Non-numerical values are simply ignored. For *Correlation, RSquare, and Covariance* analysis, the count of valid values in specifed column pair must be same; otherwise, column pair is ignored. The file names are separated by space. The valid file extensions are *.csv* and *.tsv* for comma/semicolon and tab delimited text files respectively. All other file names are ignored. All the text files in a current directory can be specified by **.csv*, **.tsv*, or the current directory name. The --indelim option determines the format of *TextFile(s)*. Any file which doesn't correspond to the format indicated by --indelim option is ignored.

OPTIONS

-c, --colmode *colnum | collabel*

Specify how columns are identified in TextFile(s): using column number or column label. Possible values: *colnum or collabel*. Default value: *colnum*.

--columns *"colnum,[colnum,...]" | "collabel,[collabel]..." | All*

This value is mode specific. It's a list of comma delimited columns to use for data analysis. Default value: *First column*.

This value is ignored during *Correlation/Pearson Correlation* and *Covariance* data analysis; -coulmnparis option is used instead.

For *colnum* value of -c, --colmode option, input values format is: *colnum,colnum,...*. Example:

```
1,3,5
```

For *collabel* value of -c, --colmode option, input values format is: *collabel,collabel,...* Example:

```
ALogP,MolWeight,EC50
```

--columnpairs *"colnum,colnum,[colnum,colnum,...]" | "collabel,collabel,[collabel,collabel,...]" | AllPairs*

This value is mode specific and is only used for *Correlation, PearsonCorrelation, or Covariance* value of -m, --mode option. It is a comma delimited list of column pairs to use for data analysis during *Correlation* and *Covariance* calculations. Default value: *First column, Second column*.

For *colnum* value of -c, --colmode option, input values format is: *colnum,colnum,[colnum,colnum]...*. Example:

```
1,3,5,6,1,6
```

For *collabel* value of -c, --colmode option, input values format is: *collabel,collabel,[collabel,collabel]...* Example:

```
MolWeight,EC50,NumN+O,PSA
```

For *AllPairs* value of --columnparis option, all column pairs are used for *Correlation* and *Covariance* calculations.

-d, --detail *infolevel*

Level of information to print about column values being ignored. Default: *1*. Possible values: 1, 2, 3, or 4.

-f, --fast

In this mode, all the columns specified for analysis are assumed to contain numerical data and no checking is performed before analysis. By default, only numerical data is used for analysis; other types of column data is ignored.

--frequencybins *number | "number,number,[number,...]"*

Specify number of bins or bin range to use for frequency analysis. Default value: *10*

Number of bins value along with the smallest and largest value for a column is used to group the column values into different groups.

The bin range list is used to group values for a column into different groups; It must contain values in ascending order. Examples:

```
10,20,30
0.1,0.2,0.3,0.4,0.5
```

The frequency value calculated for a specific bin corresponds to all the column values which are greater than the previous bin value and less than or equal to the current bin value.

-h, --help

    Print this help message.

--indelim *comma | semicolon*

    Input delimiter for CSV *TextFile(s)*. Possible values: *comma or semicolon*. Default value: *comma*. For TSV files, this option is ignored and *tab* is used as a delimiter.

--klargest *number*

    Kth largest value to find by *KLargest* function. Default value: *2* Valid values: positive integers.

--ksmallest *number*

    Kth smallest value to find by *KSmallest* function. Default value: *2*. Valid values: positive integers.

-m, --mode *DescriptiveStatisticsBasic | DescriptiveStatisticsAll | All | "function1, [function2,...]"*

    Specify how to analyze data in TextFile(s): calculate basic or all descriptive statistics; or use a comma delimited list of supported statistical functions. Possible values: *DescriptiveStatisticsBasic | DescriptiveStatisticsAll | "function1,[function2]..."*. Default value: *DescriptiveStatisticsBasic*

    DescriptiveStatisticsBasic includes these functions: *Count, Maximum, Minimum, Mean, Median, Sum, StandardDeviation, StandardError, Variance*.

    DescriptiveStatisticsAll, in addition to *DescriptiveStatisticsBasic* functions, includes: *GeometricMean, Frequency, HarmonicMean, KLargest, KSmallest, Kurtosis, Mode, RSquare, Skewness, TrimMean*.

    All uses complete list of supported functions: *Average, AverageDeviation, Correlation, Count, Covariance, GeometricMean, Frequency, HarmonicMean, KLargest, KSmallest, Kurtosis, Maximum, Minimum, Mean, Median, Mode, RSquare, Skewness, Sum, SumOfSquares, StandardDeviation, StandardDeviationN, StandardError, StandardScores, StandardScoresN, TrimMean, Variance, VarianceN*. The function names ending with N calculate corresponding values assuming an entire population instead of a population sample.

    Here are the formulas for these functions:

    Average: See Mean

    AverageDeviation: $\text{SUM}(\text{ABS}(x[i] - X_{mean})) / n$

    Correlation: See Pearson Correlation

    Covariance: $\text{SUM}((x[i] - X_{mean})(y[i] - Y_{mean})) / n$

    GeometricMean: $\text{NthROOT}(\text{PRODUCT}(x[i]))$

    HarmonicMean: $1 / (\text{SUM}(1/x[i]) / n)$

    Mean: $\text{SUM}(x[i]) / n$

    Median: $X_{sorted}[(n - 1)/2 + 1]$ for even values of n; $(X_{sorted}[n/2] + X_{sorted}[n/2 + 1])/2$ for odd values of n.

    Kurtosis: $[\{n(n + 1)/(n - 1)(n - 2)(n - 3)\} \text{SUM}\{ ((x[i] - X_{mean})/STDDEV)^4 \}] - \{3((n - 1)^2)\}/\{(n - 2)(n-3)\}$

    PearsonCorrelation: $\text{SUM}((x[i] - X_{mean})(y[i] - Y_{mean})) / \text{SQRT}(\text{SUM}((x[i] - X_{mean})^2) (\text{SUM}((y[i] - Y_{mean})^2)))$

    RSquare: $PearsonCorrelation^2$

    Skewness: $\{n/(n - 1)(n - 2)\} \text{SUM}\{ ((x[i] - X_{mean})/STDDEV)^3 \}$

    StandardDeviation: $\text{SQRT} (\text{SUM}((x[i] - Mean)^2) / (n - 1))$

    StandardDeviationN: $\text{SQRT} (\text{SUM}((x[i] - Mean)^2) / n)$

    StandardError: $StandardDeviation / \text{SQRT}(n)$

    StandardScore: $(x[i] - Mean) / (n - 1)$

    StandardScoreN: $(x[i] - Mean) / n$

    Variance: $\text{SUM}((x[i] - X_{mean})^2 / (n - 1))$

    VarianceN: $\text{SUM}((x[i] - X_{mean})^2 / n)$

-o, --overwrite

    Overwrite existing files.

--outdelim *comma | tab | semicolon*

    Output text file delimiter. Possible values: *comma, tab, or semicolon* Default value: *comma*.

-p, --precision *number*

    Precision of calculated values in the output file. Default: up to *2* decimal places. Valid values: positive integers.

-q, --quote *yes | no*

    Put quotes around column values in output text file. Possible values: *yes or no*. Default value: *yes*.

-r, --root *rootname*

    New text file name is generated using the root: <Root>.<Ext>. Default new file name: <InitialTextFileName><Mode>.<Ext>. Based on the specified analysis, <Mode> corresponds to one of these values:

DescriptiveStatisticsBasic, DescriptiveStatisticsAll, AllStatistics, SpecifiedStatistics, Covariance, Correlation, Frequency, or StandardScores. The csv, and tsv <Ext> values are used for comma/semicolon, and tab delimited text files respectively. This option is ignored for multiple input files.

--trimfraction *number*

Fraction of data to exclude from the top and bottom of the data set during *TrimMean* calculation. Default value: *0.1*. Valid values: > 0 and < 1.

-w --workingdir *text*

Location of working directory. Default: current directory.

EXAMPLES

To calculate basic statistics for data in first column and generate a NewSample1DescriptiveStatisticsBasic.csv file, type:

```
% AnalyzeTextFilesData.pl -o -r NewSample1 Sample1.csv
```

To calculate basic statistics for data in third column and generate a NewSample1DescriptiveStatisticsBasic.csv file, type:

```
% AnalyzeTextFilesData.pl --columns 3 -o -r NewSample1 Sample1.csv
```

To calculate basic statistics for data in MolWeight column and generate a NewSample1DescriptiveStatisticsBasic.csv file, type:

```
% AnalyzeTextFilesData.pl -colmode collabel --columns MolWeight -o
-r NewSample1 Sample1.csv
```

To calculate all available statistics for data in third column and all column pairs, and generate NewSample1DescriptiveStatisticsAll.csv, NewSample1CorrelationMatrix.csv, NewSample1CorrelationMatrix.csv, and NewSample1MolWeightFrequencyAnalysis.csv files, type:

```
% AnalyzeTextFilesData.pl -m DescriptiveStatisticsAll --columns 3 -o
--columnpairs AllPairs -r NewSample1 Sample1.csv
```

To compute frequency distribution of data in third column into five bins and generate NewSample1MolWeightFrequencyAnalysis.csv, type:

```
% AnalyzeTextFilesData.pl -m Frequency --frequencybins 5 --columns 3
-o -r NewSample1 Sample1.csv
```

To compute frequency distribution of data in third column into specified bin range values, and generate NewSample1MolWeightFrequencyAnalysis.csv, type:

```
% AnalyzeTextFilesData.pl -m Frequency --frequencybins "100,200,400"
--columns 3 -o -r NewSample1 Sample1.csv
```

To calculate all available statistics for data in all columns and column pairs, type:

```
% AnalyzeTextFilesData.pl -m All --columns  All --columnpairs
AllPairs -o -r NewSample1 Sample1.csv
```

AUTHOR

Manish Sud <msud@san.rr.com>

SEE ALSO

JoinTextFiles.pl, MergeTextFilesWithSD.pl, ModifyTextFilesFormat.pl, SplitTextFiles.pl, TextFilesToHTML.pl

COPYRIGHT