Package 'SmartEDA'

July 25, 2018

Type Package

Title Summarize and Explore the Data

Version 0.3.0

Author Dayanand Ubrangala; Kiran R; Ravi Prasad Kondapalli

Maintainer Dayanand Ubrangala<daya6489@gmail.com>

Depends R (>= 3.3.0)

Imports ggplot2(>= 2.2.1),scales(>= 0.5.0),utils,rmarkdown(>= 1.9),ISLR(>= 1.2),data.table(>= 1.10.4-3),stringi(>= 1.1.7),gridExtra(>= 2.3),sampling(>= 2.8),GGally(>= 1.4.0)

Description Exploratory analysis on any input data describing the structure and the relationships present in the data. The package automatically select the variable and does related descriptive statistics. Analyzing information value, weight of evidence, custom tables, summary statistics, graphical techniques will be performed for both numeric and categorical predictors.

License MIT + file LICENSE

Suggests psych, Hmisc, smbinning,testthat,knitr, covr

Encoding UTF-8

LazyData true

Repository CRAN

RoxygenNote 6.0.1

VignetteBuilder knitr

NeedsCompilation no

Date/Publication 2018-07-25 08:00:03 UTC

R topics documented:

`ExpCatStat` *Function provides summary statistics for all character or categorical columns in the dataframe*

Description

This function combines results from weight of evidence, information value and summary statistics.

Usage

```
ExpCatStat(data,Target=NULL,Label=NULL,result=c("Stat","IV"),clim=10,nlim=10,Pclass=NULL)
```

Arguments

`data`	dataframe or matrix
`Target`	target variable
`Label`	target variable label (not mandatory)
`result`	"Stat" - summary statistics, "IV" - information value
`clim`	maximum unique levles for categorical variable. Variables will be dropped if unique levels is higher than clim for class factor/character variable
`nlim`	maximum unique values for numeric variable.
`Pclass`	reference category of target variable

Details

Criteria used for categorical variable predictive power classification are

If information value is < 0.03 then predictive power = "Not Predictive"

If information value is 0.3 to 0.1 then predictive power = "Somewhat Predictive"

If information value is 0.1 to 0.3 then predictive power = "Meidum Predictive"

If information value is >0.3 then predictive power = "Highly Predictive"

Value

This function provides summary statistics for categorical variable

Stat-Summary statistics includes Chi square test scores, p value, Information values, Cramers V and Degree if association

IV- Weight of evidence and Information values

Columns description:

• Variable – variable name

• Target - Target variable label

• class – name of bin (variable value otherwise)

• out0 – number of good observations

• out1 – number of bad observations

• Total – Total values for each category

• pct1 – good observations / total good observations

• pct0 – bad observations / total bad observations

• odds – pct1/pct0

• woe – Weight of Evidence – calculated as ln(odds)

• iv – Information Value - ln(odds) * (pct0 – pct1)

Author(s)

dubrangala

Examples

```
# Example 1
## Read mtcars data
# Target variable "am" - Transmission (0 = automatic, 1 = manual)
# Summary statistics
ExpCatStat(mtcars,Target="am",Label="Transmission",result = "Stat",clim=10,nlim=5,Pclass=1)
# Inforamtion value for categorical Independent variables
ExpCatStat(mtcars,Target="am",Label="Transmission",result = "IV",clim=10,nlim=5,Pclass=1)
```

ExpCatViz *Distributions of categorical variables*

Description

This function automatically scans through each variable and creates bar plot for categorical variable.

Usage

```
ExpCatViz(data,gp=NULL,fname=NULL,clim=10,col=NULL,
margin=1,Page=NULL,Flip=F,sample=NULL,rdata=FALSE,value=NULL)
```

Arguments

data	dataframe or matrix
gp	target variable. This is not a mandatory field.
fname	output file name. Output will be generated in PDF format
clim	maximum categories to be considered to include in bar graphs.
col	define the colors to fill the bars, default it will take sample colours.
margin	index, 1 for row based proportions and 2 for column based proportions
Page	output pattern. if Page=c(3,2), It will generate 6 plots with 3 rows and 2 columns
Flip	default vertical bars. It will be used to flip the axis vertical to horizontal
sample	random selection of categorical variable
rdata	to plot bar graph for frequency/aggregated table
value	value coloumn name. This is mandatory if 'rdata' is TRUE

Value

This function returns collated graphs in grid format in PDF or JPEG format. All the files will be stored in the working directory

Bar graph - for raw data(this function will dynamically pick all the categorical variable and plot the bar chart)

Bar graph - aggregated data

Stacked Bar graph by target variable

See Also

geom_bar

Examples

```
ExpCatViz(data=mtcars,gp=NULL,fname=file.path(tempdir(),"Cat_1"),clim=10,margin=1,Page = c(2,2))
## Generate Bar graph for all the descrete data with column based proportions - random colors
set.seed(1234)
ExpCatViz(data=mtcars,gp="gear",fname=file.path(tempdir(),"Cat_2"),clim=10,margin=2,Page = c(2,2))
## Bar graph for specified variable
mtdata <- mtcars
mtdata$carname <- rownames(mtcars)
ExpCatViz(data=mtdata,gp="carname",col="blue",rdata=TRUE,value="mpg")
```

ExpCTable *Function to create frequency and custom tables*

Description

this function will automatically select categorical variables and generate frequency or cross tables based on the user inputs. Output includes counts, percentages, row total and column total.

Usage

```
ExpCTable(data,Target=NULL,margin=1,clim=10,nlim=NULL,round=2,bin=NULL,per=FALSE)
```

Arguments

data	dataframe or matrix
Target	target variable (dependent variable) if any. Default NULL
margin	margin of index, 1 for row based proportions and 2 for column based proportions
clim	maximum categories to be considered for frequency/custom table. Variables will be dropped if unique levels are higher than 'clim' for class factor/character variable. Default value is 10.
nlim	numeric variable unique limits. Default 'nlim' values is 3, table excludes the numeric variables which is having greater than 'nlim' unique values
round	round off
bin	number of cuts for continuous target variable
per	percentage values. Default table will give counts.

Details

this function provides both frequency and custom tables for all categorical features. And ouput will be generated in data frame

Value

Frequency tables, Cross tables

Columns description for frequency tables:

• Variable – Variable name

• Valid – Variable values

• Frequency – Frequency

• Percent – Relative frequency

• CumPercent – Cumulative sum of relative frequency

Columns description for custom tables:

• Variable – Variable name

- Category - Variable values
- Count – Number of counts
- Per – Percentages
- Total – Total count

Examples

```
# Frequency table
ExpCTable(mtcars,Target=NULL,margin=1,clim=10,nlim=3,bin=NULL,per=FALSE)
# Crosstbale for Mtcars data
ExpCTable(mtcars,Target="gear",margin=1,clim=10,nlim=3,bin=NULL,per=FALSE)
```

`ExpCustomStat` *Customized summary statistics*

Description

Table of descriptive statistics. Output returns matrix object containing descriptive information on all input variables for each level or combination of levels in categorical/group variable. Also while running the analysis user can filter out the data by individual variable level or across data level.

Usage

```
ExpCustomStat(data,Cvar=NULL,Nvar=NULL,stat=NULL,gpby=TRUE,filt=NULL,dcast=FALSE,
value=NULL)
```

Arguments

`data`	dataframe or Matrix
`Cvar`	qualitative variables on which to stratify / subgroup or run categorical summaries
`Nvar`	quantitative variables on which to run summary statistics for.
`stat`	descriptive statistics. Sepecify which summary statistics required (Included all base stat functions like 'mean','medain','max','min','sum','IQR','sd','var',quantile like P0.1, P0.2 etc'). Also added two more stat here are 'PS' - percentage of shares and 'Prop' - column percentage
`gpby`	default value is True. Group level summary will be created based on list of categorical variable. If summary required at each categorical variable level then keep this option as FALSE
`filt`	filter out data while running the summary statistics. Filter can apply accross data or individual variable level using filt option. If there are multiple filters, seperate the conditons by using '^'. Ex: Nvar = c("X1","X2","X3","X4"), let say we need to exclude data X1>900 for X1 variable, X2==10 for X2 variable, Gender !='Male' for X3 variable and all data for X4 then filt should be, filt = c("X1>900"^"X2==10"^"Gender!='Male'"^all) or c("X1>900"^"X2==10"^"Gender!='Male'"^ ^). in case if you want to keep all data for some of the variable listed in Nvar, then specify inside the filt like ^all^ or ^ ^(single space)

dcast	fast dcast from data.table
value	If dcast is TRUE, pass the variable name which needs to come on column

Details

Filter unique value from all the numeric variables

This will be usefull when we need to exclude some unique imputed or outliers values like '999' or '9999' or '-9' or '-1111', or '888' etc from each selected variables.

Eg:dat = data.frame(x = c(23,24,34,999,12,12,23,999,45), y = c(1,3,4,999,0,999,0,8,999,0)

Exclude 999: x = c(23,24,34,12,12,23,45) y = c(1,3,4,0,0,8,0)

The complete functionality of 'ExpCustomStat' function is detailed in vignette help page with example code.

Value

summary statistics as dataframe. Usage of this function is detailed in user guide vignettes document.

Examples

```
ExpCustomStat(mtcars,Cvar=c("vs","am","gear"),Nvar=c("disp","mpg"),
stat=c("Count","sum","PS"),gpby=TRUE,filt =NULL)
```

ExpData *Function to generate the overview of a data frame*

Description

This function used to produce summaries of data structure and overview of the data frame.

Usage

```
ExpData(data,type=1)
```

Arguments

data	a data frame
type	Type 1 is overview of the data; Type 2 is structure of the data

Details

This function provides overview and structure of the data frames.

IF Type=1,overeview of the data (column names are "Descriptions", "Obs")

If Type=2,structure of the data (column names are "S.no","Variable Name","Variable Type","

Examples

```
# Overview of the data
ExpData(data=mtcars,type=1)
# Structure of the data
ExpData(data=mtcars,type=2)
```

`ExpInfoValue` *Information value*

Description

Provides information value for each categorical variable (X) against target variable (Y)

Usage

```
ExpInfoValue(X, Y, valueOfGood = NULL)
```

Arguments

`X`	Independent categorical variable.
`Y`	Binary response variable, it can take values of either 1 or 0.
`valueOfGood`	Value of Y that is used as reference category.

Details

Information value is one of the most useful technique to select important variables in a predictive model. It helps to rank variables on the basis of their importance. The IV is calculated using the following formula

IV = (Percentage of Good event - Percentage of Bad event) * WOE, where WOE is weight of evidence

WOE = log(Percentage of Good event - Percentage of Bad event)

Here is what the values of IV mean according to Siddiqi (2006)

If information value is < 0.03 then predictive power = "Not Predictive"

If information value is 0.03 to 0.1 then predictive power = "Somewhat Predictive"

If information value is 0.1 to 0.3 then predictive power = "Meidum Predictive"

If information value is >0.3 then predictive power = "Highly Predictive"

Value

Information value (iv) and Predictive power class

information value

predictive class

See Also

IV

Examples

```
X = mtcars$gear
Y = mtcars$am
ExpInfoValue(X,Y,valueOfGood = 1)
```

ExpKurtosis *Measures of Shape - Kurtosis*

Description

Measures of shape to give a detailed evaluation of data. Explains the amount and direction of skew. Kurotsis explains how tall and sharp the central peak is. Skewness has no units: but a number, like a z score

Usage

```
ExpKurtosis(x,type)
```

Arguments

x	A numeric object or data.frame
type	a character which specifies the method of computation. Options are "moment" or "excess"

Value

ExpKurtosis returns Kurtosis values

Author(s)

dubrangala

Examples

```
ExpKurtosis(mtcars$hp,type="excess")
ExpKurtosis(mtcars$carb,type="moment")
ExpKurtosis(mtcars,type="excess")
```

ExpNumStat *Summary statistics for numerical variables*

Description

Function provides summary statistics for all numerical variable. This function automatically scans through each variable and select only numeric/integer variables. Also if we know the target variable, function will generate relationship between target variable and each independent variable.

Usage

```
ExpNumStat(data,by=NULL,gp=NULL,Qnt=NULL,Nlim=10,MesofShape=2,
Outlier=FALSE,round=3,dcast=FALSE,val=NULL)
```

Arguments

data	dataframe or matrix
by	group by A (summary statistics by All), G (summary statistics by group), GA (summary statistics by group and Overall)
gp	target variable if any, default NULL
Qnt	default NULL. Specified quantiles [c(.25,0.75) will find 25th and 75th percentiles]
Nlim	numeric variable limit (default value is 10 which means it will only consider those variable having more than 10 unique values and variable type is numeric/integer)
MesofShape	Measures of shapes (Skewness and kurtosis).
Outlier	Calculate the lower hinge, upper hinge and number of outliers
round	round off
dcast	fast dcast from data.table
val	Name of the column whose values will be filled to cast (see Detials sections for list of column names)

Details

Summary by – overall

Summary by – group (target variable)

Summary by – overall and group (target variable)

coloumn descriptions

- Vname – Variable name
- Group – Target variable
- TN – Total sample (inculded NA observations)
- nNeg – Total negative observations
- nZero – Total zero observations

- nPos – Total positive observations
- NegInf – Negative infinite count
- PosInf – Positive infinite count
- NA_value – Not Applicable count
- Per_of_Missing – Percentage of missings
- Min – minimum value
- Max – maximum value
- Mean – average value
- Median – median value
- SD – Standard deviation
- CV – coefficient of variations (SD/mean)*100
- IQR – Inter quartile range
- Qnt – Specified quantiles
- MesofShape – Skewness and Kurtosis
- Outlier – Number of outliers
- Cor – Correlation b/w target and independent variables

Value

summary statistics for numeric independent variables

Author(s)

dubrangala

See Also

describe.by

Examples

```
## Descriptive summary of numeric variables - Summary by Target variables
ExpNumStat(mtcars,by="G",gp="gear",Qnt=c(0.1,0.2),MesofShape=2,
Outlier=TRUE,round=3)
## Descriptive summary of numeric variables - Summary by Overall
ExpNumStat(mtcars,by="A",gp="gear",Qnt=c(0.1,0.2),MesofShape=2,
Outlier=TRUE,round=3)
## Descriptive summary of numeric variables - Summary by Overall and Group
ExpNumStat(mtcars,by="GA",gp="gear",Qnt=seq(0,1,.1),MesofShape=1,
Outlier=TRUE,round=2)
## Summary by specific statistics for all numeric variables
ExpNumStat(mtcars,by="GA",gp="gear",Qnt=c(0.1,0.2),MesofShape=2,
Outlier=FALSE,round=2,dcast = TRUE,val = "IQR")
```

ExpNumViz *Distributions of numeric variables*

Description

This function automatically scans through each variable and creates density plot, scatter plot and box plot for continuous variable.

Usage

```
ExpNumViz (data,gp=NULL,type=1,nlim=NULL,fname=NULL,col=NULL,Page=NULL,sample=NULL)
```

Arguments

data	dataframe or matrix
gp	target variable
type	1 (boxplot by category and overall), 2 (boxplot by category only), 3 (boxplot for overall)
nlim	numeric variable unique limit. Default nlim is 3, graph will exclude the numeric variable which is having less than 'nlim' unique value
fname	output file name
col	define the fill color for box plot. Number of color should be equal to number of categories in target variable
Page	output pattern. if Page=c(3,2), It will generate 6 plots with 3 rows and 2 columns
sample	random selection of plots

Details

This function automatically scan each variables and generate a graph based on the user inputs. Graphical representation includes scatter plot, box plot and density plots. If input "gp" is continuous then output is scatter plots

If input "gp" is categorical then output is box plot.

If input "gp" is NULL, means there is no target variable and this will generate density plot for all numeric features

Value

returns collated graphs in PDF or JPEG format

Scatter plot for numeric data

Density plot for numeric data

Boxplot – by overall

Boxplot – by group (target variable)

Boxplot – by overall and group (target variable)

See Also

geom_boxplot

Examples

```
## Generate Boxplot by category
ExpNumViz(mtcars,gp="gear",type=2,nlim=25,fname = file.path(tempdir(),"Mtcars2"),Page = c(2,2))
## Generate Density plot
ExpNumViz(mtcars,gp=NULL,type=3,nlim=25,fname = file.path(tempdir(),"Mtcars3"),Page = c(2,2))
## Generate Scatter plot
ExpNumViz(mtcars,gp="carb",type=3,nlim=25,fname = file.path(tempdir(),"Mtcars4"),Page = c(2,2))
```

ExpOutQQ *Quantile-Quantile Plots*

Description

This function automatically scans through each variable and creates normal QQ plot also adds a line to a normal quantile-quantile plot.

Usage

```
ExpOutQQ(data,nlim=NULL,fname=NULL,Page=NULL,sample=NULL)
```

Arguments

data	Input dataframe or data.table
nlim	numeric variable limit
fname	output file name. Output will be generated in PDF format
Page	output pattern. if Page=c(3,2), It will generate 6 plots with 3 rows and 2 columns
sample	random number of plots

Value

Normal quantile-quantile plot

See Also

geom_qq

Examples

```
CData = ISLR::Carseats
ExpOutQQ(CData,nlim=10,fname=NULL,Page=c(2,2),sample=4)
```

ExpParcoord *Parallel Co-ordinate plots*

Description

This function creates parallel Co-ordinate plots

Usage

```
ExpParcoord (data,Group=NULL,Stsize=NULL,Nvar=NULL,Cvar=NULL,scale=NULL)
```

Arguments

data	Input dataframe or data.table
Group	stratification variables
Stsize	vector of startum sample sizes
Nvar	vector of numerice variables, default it will consider all the numeric variable from data
Cvar	vector of categorical variables, default it will consider all the categorical variable
scale	scale the variables in the parallel coordinate plot (Default normailized with minimum of the variable is zero and maximum of the variable is one) (see ggparcoord details for more scale options)

Details

The Parallel Co-ordinate plots having the functionalities of visulization for sample rows if data size large. Also data can be stratified basis of Target or group variables. It will normalize all numeric variables between 0 and 1 also having other standardization options. It will automatically make dummy (1,0) variables for categorical variables

Value

Parallel Co-ordinate plots

See Also

ggparcoord

Examples

```
CData = ISLR::Carseats
# Defualt ExpParcoord funciton
ExpParcoord(CData,Group=NULL,Stsize=NULL,
Nvar=c("Price","Income","Advertising","Population","Age","Education"))
# With Stratified rows and selected columns only
ExpParcoord(CData,Group="ShelveLoc",Stsize=c(10,15,20),
Nvar=c("Price","Income"),Cvar=c("Urban","US"))
```

```
# Without stratification
ExpParcoord(CData,Group="ShelveLoc",Nvar=c("Price","Income"),
Cvar=c("Urban","US"),scale=NULL)
# Scale changed std: univariately, subtract mean and divide by standard deviation
ExpParcoord(CData,Group="US",Nvar=c("Price","Income"),
Cvar=c("ShelveLoc"),scale="std")
# Selected numeric variables
ExpParcoord(CData,Group="ShelveLoc",Stsize=c(10,15,20),
Nvar=c("Price","Income","Advertising","Population","Age","Education"))
```

ExpReport *Function to create HTML EDA report*

Description

Create a exploratory data analysis report in HTML format

Usage

```
ExpReport(data,Template=NULL,Target=NULL,label=NULL,op_file=NULL,
op_dir=getwd(),sc=NULL,sn=NULL,Rc=NULL)
```

Arguments

data	a data frame
Template	R markdown template (.rmd file)
Target	dependent variable. If there is no defined target variable then keep as it is NULL.
label	target variable descriptions, not a mandatory field
op_file	output file name (.html)
op_dir	output path
sc	sample number of plots for categorical variable. User can decide how many number of plots to depict in html report.
sn	sample number of plots for numerical variable. User can decide how many number of plots to depict in html report.
Rc	reference category of target variable. If Target is categorical then Pclass value is mandatory and which should not be NULL

Details

The "ExpReport" function will generate a HTML report for any R data frames. If the markdown template is ready, we can use that template to generate the HTML report else It will generate three different types of HTML report based on the Target field

IF Target = NULL, means there is no defined dependent varaible then it will genreate general EDA report at overall level

IF Target = continuous, then it will genreate EDA report including univariate and multivarite summary statistics with correlation.

IF Target = categorical, then it will genreate EDA report including univariate and multivarite summary statistics with chi-square, Information values.

See Also

create_report

Examples

```
# Overview of the data
ExpReport(mtcars,Template=NULL,Target=NULL,label=NULL,op_file="Myreport.html",
op_dir=getwd(),sc=2,sn=2,Rc=NULL)
```

ExpSkew *Measures of Shape - Skewness*

Description

Measures of shape to give a detailed evaluation of data. Explains the amount and direction of skew. Kurotsis explains how tall and sharp the central peak is. Skewness has no units: but a number, like a z score

Usage

```
ExpSkew(x,type)
```

Arguments

x	A numeric object or data.frame
type	a character which specifies the method of computation. Options are "moment" or "sample"

Value

ExpSkew returns Skewness values

Author(s)

dubrangala

Examples

```
ExpSkew(mtcars,type="moment")
ExpSkew(mtcars,type="sample")
```

`ExpStat` *Function provides summary statistics for individual categorical predictors*

Description

Provides bivariate summary statistics for all the categorical predictors against target variables. Output includes chi - square value, degrees of freedom, information value, p-value

Usage

```
ExpStat(X,Y,valueOfGood = NULL)
```

Arguments

`X`	Independent categorical variable.
`Y`	Binary response variable, it can take values of either 1 or 0.
`valueOfGood`	Value of Y that is used as reference category.

Details

For a given binary Y variable and X categorical variables, the summary statistics are computed. Summary statistics included Pearson's Chi-squared Test for Count Data, "chisq.test" which performs chi-squared contingency table tests and goodness-of-fit tests. If any NA value present in X or Y variable, which will be considered as NA as in category while computing the contingency table.

Also added unique levels for each X categorical variables and degrees of freedom

Value

The function provides summary statistics like

Unique levels

Chi square statistics

P value

Degrees of freedom

Information value

Predictive class

See Also

`chisq.test`

Examples

```
X = mtcars$carb
Y = mtcars$am
ExpStat(X,Y,valueOfGood = 1)
```

ExpWoeTable *Function provides summary statistics with weight of evidence*

Description

Weight of evidence for categorical(X-independent) variable against Target variable (Y)

Usage

```
ExpWoeTable(X, Y, valueOfGood = NULL,print=FALSE)
```

Arguments

X	Independent categorical variable.
Y	Binary response variable, it can take values of either 1 or 0.
valueOfGood	Value of Y that is used as reference category.
print	print results

Details

The weight of evidence tells the predictive power of an independent variable in relation to the dependent variable

Value

Weight of evidance summary table

See Also

WOETable

Examples

```
X = mtcars$gear
Y = mtcars$am
WOE = ExpWoeTable(X,Y,valueOfGood = 1)
```

Index