Cheat Sheet for Data Exploration in R

Most Commonly Used R Libraries

> install.packages("package name")

- Outlier Detection - outlier, EVIR
- Feature Selection - Features, RRF
- Data Transformation - plyr, data.table
- Data Visualization - ggplot2, googleVis
- Dimension Reduction - factoMiner, CCP
- Missing Value Imputations - MissForest, MissMDA

How to ? Load a data file(s)

```
# Read CSV file into R
> MyData <- read.csv("C:/TheDataIWantToReadIn.csv", header=TRUE, sep=",")

#Read a Tab seperated file
> Tabseperated <- read.table("c:/TheDataIWantToReadIn.tsv", sep="\t", header=TRUE)
```

How to ? Convert a variable to different data type

```
is.numeric(), is.character(), is.vector(), is.matrix(), is.data.frame()
as.numeric(), as.character(), as.vector(), as.matrix(), as.data.frame()
```

NOTE
Use is.xyz to test for data type xyz. Returns TRUE or FALSE
Use as.xyz to explicitly convert it.

	to one long vector	to matrix	to data frame
from vector	c(x,y)	cbind(x,y) rbind(x,y)	data.frame(x,y)
from matrix	as.vector(mymatrix)		as.data.frame(mymatrix)
from data frame		as.matrix(myframe)	

How to ? Transpose a Data set

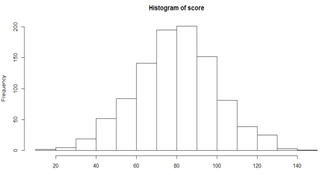
```
# example of melt function
> library(reshape)
> mdata <- melt(mydata, id=c("id","time"))
```

How to ? Sort DataFrame

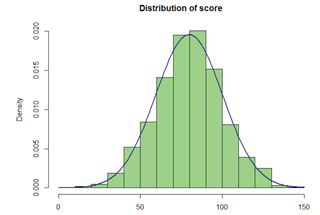
```
# sort by var1
> newdata <- old[order(var1),]

# sort by var1 and var2 (descending)
> newdata2 <- old[order(var1, -var2),]
```

How to ? Create plots (Histogram)

```
> score <-rnorm(n=1000, m=80, sd=20)
> hist(score)
```




```
> histinfo<-hist(score)
> histinfo

> $breaks
 [1] 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150

> $counts
 [1] 2 5 19 52 84 141 195 201 152 81 39 25 3 1

> $density
 [1] 0.0002 0.0005 0.0019 0.0052 0.0084 0.0141 0.0195 0.0201 0.0152
[10] 0.0081 0.0039 0.0025 0.0003 0.0001

> $mids
 [1] 15 25 35 45 55 65 75 85 95 105 115 125 135 145

> $xname
 [1] "score"

> $equidist
 [1] TRUE

> attr(,"class")
 [1] "histogram"


> hist(score, freq=FALSE, xlab="Score", main="Distribution of score", col="lightgreen",
+ xlim=c(0,150), ylim=c(0, 0.02))
curve(dnorm(x, mean=mean(score), sd=sd(score)), add=TRUE, col="darkblue", lwd=2)
```



How to ? Generate frequency tables with R

```
> attach(iris)
> table(iris$Species)

# 2-Way Cross Tabulation
> library(gmodels)
> CrossTable(mydata$myrowvar, mydata$mycolvar)
```

How to ? Sample Data set in R

```
> mysample <- mydata[sample(1:nrow(mydata), 100,replace=FALSE),]
> mysample #check your sample
```

How to ? Remove duplicate values of a variable

```
> set.seed(150)
> x <- round(rnorm(20, 10, 5))
> x
 [1] 2 10 6 8 9 11 14 12 11 6 10 0 10 7 7 20 11 17 12 -1
> unique(x)
 [1] 2 10 6 8 9 11 14 12 0 7 20 17 -1
```

How to ? Find class level count average and sum in R

```
> tapply(iris$Sepal.Length,iris$Species,sum)
 setosa versicolor virginica
 250.3 296.8 329.4

> tapply(iris$Sepal.Length,iris$Species,mean)
 setosa versicolor virginica
 5.006 5.936 6.588
```

How to ? Recognize and treat missing values and outliers

```
> y <- c(4,5,6,NA)
> is.na(y)
[1] FALSE FALSE FALSE TRUE

# and here is a quick fix for the same

> y[is.na(y)] <- mean(y,na.rm=TRUE)
> y
[1] 4 5 6 5
```

How to ? Merge / Join data sets

```
# merge two data frames by ID
> total <- merge(data_frameA,data_frameB,by="ID")

# merge two data frames by ID and Country
> total <- merge(data_frameA,data_frameB,by=c("ID","Country")
> total <- rbind(data_frameA, data_frameB)
```

To view the complete guide on Data Exploration in R
http://bit.ly/1MTSpe0

Analytics Vidhya
Learn Everything About Analytics