**DATA ANALYSIS**

Data analysis, is a process of inspecting, cleansing, transforming and modelling data. It is used to discover useful information, suggest conclusions, and support decision-making. It has numerous faces and approaches which results in variety of methods under diverse names, in variant domains such as science, social science and businesses.

In it’s statistical applications it can be divided into:

* Descriptive statistics
* Exploratory data analysis (discovering new features in the data)
* Confirmatory data analysis (confirming or falsifying existing hypotheses)

**DATA SCIENCE**

It is an interdisciplinary field about scientific methods, processes and systems. It is used either on structured or unstructured data to get insights.

It unifies statistics, data analysis and the methods related to them with the aim to understand and analyze actual facts. The theories and techniques used in it are harassed from numerous fields of statistics, mathematics, computer science and information science.

When given a challenging question, it initiates with exploration, They try to understand the pattern or characteristics within the data by investigating leads.

In order  to get a level deeper with the intent to frame together a forensic view of what data is saying, quantitative techniques can be applied by data scientists. e.g. inferential models, segmentation analysis, time series forecasting, synthetic control experiments, etc.

Data scientist guides business stakeholders on how to takes decisions i.e they act as consultants as this data-driven insight is central to providing strategic guidance.]

**DATA SCIENCE VS DATA ANALYSIS**

|  |  |
| --- | --- |
| **Data science** | **Data analytics** |
| It is a field that comprises of everything that is related to data cleansing, preparation and analysis. | It involves automating into a certain dataset and as well as supposes the usage of queries and data aggregation procedure. |
| Data science algorithms are used in industries such as internet searches, digital advertisements, search recommenders etc. | Data analytics is used in industries such as healthcare, gaming, energy management, travel etc. |
| Skills required:   * In depth knowledge in SAS and/or R. * Python coding * Hadoop platform * SQL database * Working with unstructured data | Skills required:   * Programming skills * Statistical skills * Mathematics * Machine learning skills * Data wrangling * Communication and data visualisation  skills * DAta intitution |

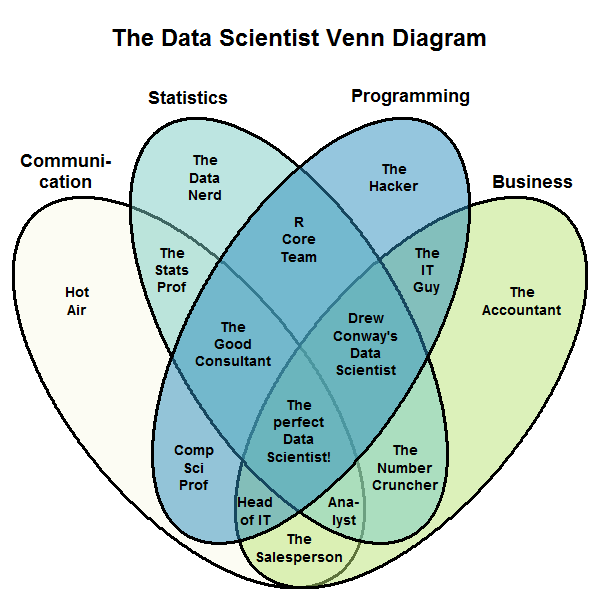


Fig 1: *Venn diagram showing skills required to be indulged in a specific role.*

**R/R Studio**

**R**

* The R statistical programming language is a free open source package based on the S language developed by Bell Labs.
* It is an interpreted language.
* It is very powerful for writing programs.
* It has many statistical functions built in.
* It is free.

**R Studio**

It is an IDE which allows the user to run R in more user  friendly manner. It was founded by JJ Allaire. IT is available in two editions: RStudio Desktop and RStudio Server. It is written in C++ and for graphical framework it uses Qt framework. Development of it was started at the end of 2010. It’s first beta version(v0.92) was released in February, 2011. On 1 November,2016 Version 1.0 was released.

It has four components:

* File editor
* Console
* Workspace

It aids users by providing features such as Code completion(it predicts the possible arguments, functions, braces etc.), Command history search(it  gives you the liberty to look for previous commands), Command history to R script/file, Function extraction from Rscript etc.

**Data Preparation**

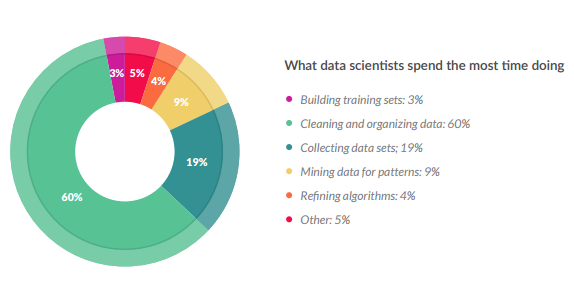
Data scientists roughly spend 50-80% of their time in gathering and cleaning data i.e trying to get it in the form on which they can perform desired operations upon. What makes it so important to devote such amount of time to this activity is the diverse nature of data that they encounter from their size to formats to ordering.

Data cleaning provides the direction to our research. It is extent to which we cleaned the data which decides how fruitful our algorithms will be. This quite clearly tells that cleaning the data is as significant to any firm as making good algorithms is. More often than not this turns out to be the point of differentiation between great enterprise data scientist and moderate enterprise data scientist.

https://www.dezyre.com/article/why-data-preparation-is-an-important-part-of-data-science/242

According to the survey which was perforemed on 80 data scientists San Francisco based crwodsourcing and data mining company ‘CrowdFlower’:

http://visit.crowdflower.com/rs/416-ZBE-142/images/CrowdFlower\_DataScienceReport\_2016.pdf

****

Even though it is the one of the least glamourous tasks but most of the time a data scientist is spent in data cleaning which we also call data munging and data wrangling. The amount of time needed to prepare data (clean it) is dependent on the fact that how healthy data is i.e how complete it is, how many inconsistencies it has, how many missing values are there and many such things.

Data cleaning is essential because given data can have discrepancies in the names or codes, it may also have outliers or errors, to summarize the initial dataset is not qualitative rather it is quantitative.

Steps involved in data preparation are:

* Data cleaning:

This step include dealing with missing and inconsistent values and thereby smoothing out noisy data. The missing values can either be treated by removing them or by filling them with appropriate values depending upon the situation. Noisy data is either is tackled manually or using various regression and clustering techniques.

* Data integration:

It involves things like data conflict resolving, handling of redundancies in data if there are any & schema integration

* Data transformation:

In this step noise is removed from the data if there is any  and it also does normalization, aggregation and generalization.

* Data reduction:

This step is done with the motive to reduce the size of data keeping but at the same time maintaining it’s utility i.e. making sure that it has the same statistical inference qualities as it used to have. Depending the requirement we can use either of the three approaches: dimensionality reduction, data cube aggregation and numerosity reduction.

* Data discretization:

This step is mainly used for the algorithms which only accept categorical attributes. This step helps data scientists to divide continuous data into intervals and thereby also assisting in reducing the size of the data in hand.

**EDA: Exploratory Data Analysis**

It is the approach to analyze data with the motive to summarize the main characteristics of the data and get vital insights. Mostly graphical methods like scatter plots, histograms, bar graphs, line plots etc. are used but it does have certain quantitative techniques as well. EDA’s main motive is to help us:

* Select the appropriate tool and technique that should be used.
* Design hypothesis tests.
* Evaluate the assumptions which will be the basis of our statistical inferences.
* Assess the requirement of data collection that has to be done in future in order to make proper conclusions or may even predict things.

**Example1**

I tried to perform EDA on numerous datasets but to begin with I will try to elaborate the process with the ‘iris’ dataset which is available in ‘datasets’ library. This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica. iris is a data frame with 150 cases (rows) and 5 variables (columns) named Sepal.Length, Sepal.Width, Petal.Length, Petal.Width, and Species.

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) The New S Language. Wadsworth & Brooks/Cole. (has iris3 as iris.)

**Report generated with R Markdown**

**analysis\_on\_iris.R**

Mrinal Jhamb

Wed Jun 21 13:17:12 2017

**library**('datasets')  
**data**()  
**data**(iris)  
**View**(iris)  
**head**(iris)

## Sepal.Length Sepal.Width Petal.Length Petal.Width Species  
## 1 5.1 3.5 1.4 0.2 setosa  
## 2 4.9 3.0 1.4 0.2 setosa  
## 3 4.7 3.2 1.3 0.2 setosa  
## 4 4.6 3.1 1.5 0.2 setosa  
## 5 5.0 3.6 1.4 0.2 setosa  
## 6 5.4 3.9 1.7 0.4 setosa

**summary**(iris)

## Sepal.Length Sepal.Width Petal.Length Petal.Width   
## Min. :4.300 Min. :2.000 Min. :1.000 Min. :0.100   
## 1st Qu.:5.100 1st Qu.:2.800 1st Qu.:1.600 1st Qu.:0.300   
## Median :5.800 Median :3.000 Median :4.350 Median :1.300   
## Mean :5.843 Mean :3.057 Mean :3.758 Mean :1.199   
## 3rd Qu.:6.400 3rd Qu.:3.300 3rd Qu.:5.100 3rd Qu.:1.800   
## Max. :7.900 Max. :4.400 Max. :6.900 Max. :2.500   
## Species   
## setosa :50   
## versicolor:50   
## virginica :50   
##   
##   
##

*# What are the names of different columns in IRIS data ?*  
**names**(iris)

## [1] "Sepal.Length" "Sepal.Width" "Petal.Length" "Petal.Width"   
## [5] "Species"

*#Identify the means of relevant columns in IRIS data ?*  
lst=iris[,1:4]  
**View**(lst)  
?sapply

## starting httpd help server ...

## done

**sapply**(lst,mean)

## Sepal.Length Sepal.Width Petal.Length Petal.Width   
## 5.843333 3.057333 3.758000 1.199333

*#Identify the means and standard deviation (sd) of different columns for each specie separately.*  
specie1=iris[iris$Species=="setosa",1:4]  
**sapply**(specie1,mean)

## Sepal.Length Sepal.Width Petal.Length Petal.Width   
## 5.006 3.428 1.462 0.246

**sapply**(specie1,sd)

## Sepal.Length Sepal.Width Petal.Length Petal.Width   
## 0.3524897 0.3790644 0.1736640 0.1053856

specie2=iris[iris$Species=="virginica",1:4]  
**sapply**(specie2,mean)

## Sepal.Length Sepal.Width Petal.Length Petal.Width   
## 6.588 2.974 5.552 2.026

**sapply**(specie2,sd)

## Sepal.Length Sepal.Width Petal.Length Petal.Width   
## 0.6358796 0.3224966 0.5518947 0.2746501

specie3=iris[iris$Species=="versicolor",1:4]  
**sapply**(specie3,mean)

## Sepal.Length Sepal.Width Petal.Length Petal.Width   
## 5.936 2.770 4.260 1.326

**sapply**(specie3,sd)

## Sepal.Length Sepal.Width Petal.Length Petal.Width   
## 0.5161711 0.3137983 0.4699110 0.1977527

**summary**(specie1)

## Sepal.Length Sepal.Width Petal.Length Petal.Width   
## Min. :4.300 Min. :2.300 Min. :1.000 Min. :0.100   
## 1st Qu.:4.800 1st Qu.:3.200 1st Qu.:1.400 1st Qu.:0.200   
## Median :5.000 Median :3.400 Median :1.500 Median :0.200   
## Mean :5.006 Mean :3.428 Mean :1.462 Mean :0.246   
## 3rd Qu.:5.200 3rd Qu.:3.675 3rd Qu.:1.575 3rd Qu.:0.300   
## Max. :5.800 Max. :4.400 Max. :1.900 Max. :0.600

**summary**(specie2)

## Sepal.Length Sepal.Width Petal.Length Petal.Width   
## Min. :4.900 Min. :2.200 Min. :4.500 Min. :1.400   
## 1st Qu.:6.225 1st Qu.:2.800 1st Qu.:5.100 1st Qu.:1.800   
## Median :6.500 Median :3.000 Median :5.550 Median :2.000   
## Mean :6.588 Mean :2.974 Mean :5.552 Mean :2.026   
## 3rd Qu.:6.900 3rd Qu.:3.175 3rd Qu.:5.875 3rd Qu.:2.300   
## Max. :7.900 Max. :3.800 Max. :6.900 Max. :2.500

**summary**(specie3)

## Sepal.Length Sepal.Width Petal.Length Petal.Width   
## Min. :4.900 Min. :2.000 Min. :3.00 Min. :1.000   
## 1st Qu.:5.600 1st Qu.:2.525 1st Qu.:4.00 1st Qu.:1.200   
## Median :5.900 Median :2.800 Median :4.35 Median :1.300   
## Mean :5.936 Mean :2.770 Mean :4.26 Mean :1.326   
## 3rd Qu.:6.300 3rd Qu.:3.000 3rd Qu.:4.60 3rd Qu.:1.500   
## Max. :7.000 Max. :3.400 Max. :5.10 Max. :1.800

?iris

**Conclusions drawn**

* On an avg. mean size of petals is the greatest in virginica and it is the least in setosa.
* Size of the petals are most diverse in virginica.
* On an avg. sepals of setosa are expected to be the most circular or the least elliptical(because of the least difference between the mean length and width of setosa).
* Versicolor has the sepal with the least width whereas width is the largest in setosa.
* Setosa has the sepal with the least length whereas width is the least in versicolor.
* Virginica has the petal with the largest width and it also has the one with the largest length.
* Setosa has the petal with the least width and it also has the one with the least length.
* From the exploratory data analysis we found that the shape &size of flowers are different.

**Example 2**

Second EDA is on the ‘women’ dataset which is available in ‘datasets’ library. This data set gives the average heights and weights for American women aged 30–39. Data has been taken from the ‘American Society of Actuaries Build and Blood Pressure Study’.This is a data frame with 15 observations on 2 variables: height (in) & weight (lbs).

McNeil, D. R. (1977) Interactive Data Analysis. Wiley.

**Report generated with R Markdown**

**analysis\_on\_women.R**

Mrinal Jhamb

Wed Jun 21 13:16:13 2017

**library**(datasets)  
**data**()  
**data**(women)  
**View**(women)  
**sapply**(women,mean)

## height weight   
## 65.0000 136.7333

**sapply**(women,min)

## height weight   
## 58 115

**sapply**(women,max)

## height weight   
## 72 164

**sapply**(women,sd)

## height weight   
## 4.472136 15.498694

**sapply**(women,median)

## height weight   
## 65 135

136.7333-15.48694

## [1] 121.2464

136.7333+15.48694

## [1] 152.2202

65+4.472136

## [1] 69.47214

65-4.472136

## [1] 60.52786

lst1=women[women$height>"69.47214",]  
**length**(lst1$height)

## [1] 3

lst2=women[women$height<"60.52786",]  
**length**(lst2$height)

## [1] 3

lst3=women[women$height>"60.52786" & women$height<"69.47214",]  
**length**(lst3$height)

## [1] 9

lst4=women[women$wieght>"152.2202",]  
**length**(lst1$height)

## [1] 3

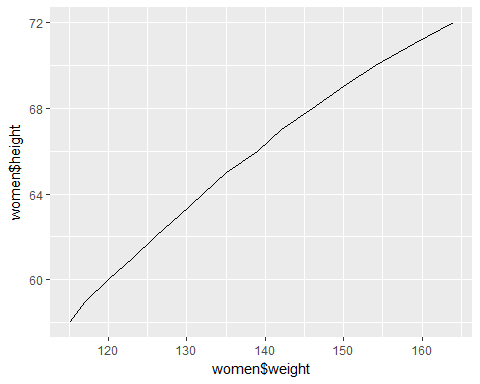
lst5=women[women$weight<"121.2464",]  
**length**(lst2$height)

## [1] 3

lst6=women[women$weight>"121.2464" & women$height<"152.2202",]  
**length**(lst3$height)

## [1] 9

**library**(ggplot2)  
**qplot**(women$weight,women$height,data=women,geom="line")



bmi=**data.frame**(703\*women[[2]]/(women[[1]]\*women[[1]]))  
**colnames**(bmi)=**c**("bmi")  
**View**(bmi)  
**length**(bmi[bmi<18.5])

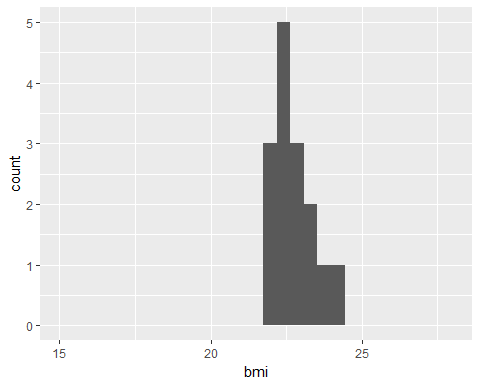
## [1] 0

**length**(bmi[bmi>18.5 & bmi<24.9])

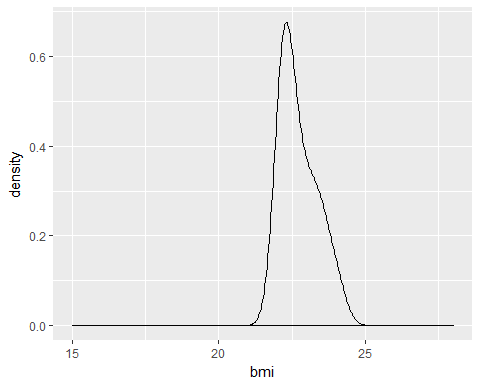
## [1] 15

**qplot**(bmi,data=bmi,geom="histogram",bins=30,xlim=**c**(15,28))

## Warning: Removed 1 rows containing missing values (geom\_bar).



**qplot**(bmi,data=bmi,geom="density",xlim=**c**(15,28))



**Conclusions drawn**

* On the basis of standard deviation the women can be categorised on the basis of their weight and height.
* Categorisation:
  + Weight:
    - Women with weight between 121.2464 & 152.2202: Normal Count:9
    - Women with weight less than 121.2464: Under Weight Count:3
    - Women with weight greater than 152.2202: Over Weight Count:3
  + Height:
    - Women with height between 60.52786 & 69.47214: Normal Heighted Count:9
    - Women with height less than 60.52786: Short Heighted Count:3
    - Women with height greater than 69.47214: Tall Heighted Count:3
* The trend quite clearly leads us to conclusion that women who are 'Tall Heighted' also lie in the category of 'Over Weight' & those who are 'Short Heighted' lie in the 'Under Weight'.
* So by looking at the trend over judgement of people being 'Over Weight' & 'Under Weight' is wrong. Since height also has the role to play. That's why we have BMI.
* BMI:
  + underweight (BMI less than 18.5) count:0
  + normal weight (BMI between 18.5 & 24.9) count:15
  + overweight (BMI between 25.0 & 29.9) count:0
  + obese (BMI 30.0 and above) count:0

These turns out to be the excellent example of how EDA can extremely useful to add new dimensions to our analysis. In this case it did assist us with the fact that only by looking at the weight one can’t judge someone to be underweight or overweight as height also has the role to play which got to know by looking at the line plot between the height and the weight.

**Example 3**

Third EDA is on the ‘chickweight’ dataset which is available in ‘datasets’ library. The ChickWeight data frame has 578 rows and 4 columns from an experiment on the effect of diet on early growth of chicks.The body weights of the chicks were measured at birth and every second day thereafter until day 20. They were also measured on day 21. There were four groups on chicks on different protein diets.

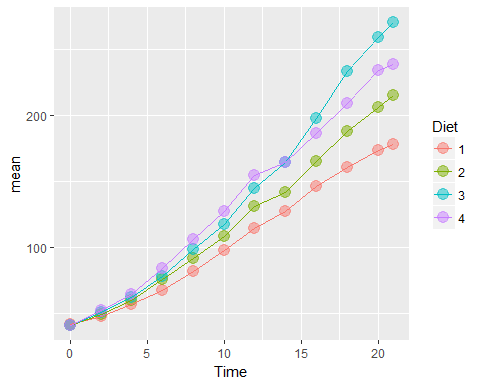
**Report generated with R Markdown**

**analysis\_on\_chick.R**

Mrinal Jhamb

Wed Jun 21 13:12:20 2017

**library**(datasets)  
**data**("ChickWeight")  
**View**(ChickWeight)  
  
cw=ChickWeight  
count=**data.frame**(**table**(cw$Chick))  
**View**(count)  
defective=count[**which**(count$Freq<12),]  
**View**(defective)  
clean=cw[-**which**(cw$Chick %in% defective$Var1),]  
**View**(clean)  
**library**(plyr)  
x=**ddply**(clean[,**c**(1,2,4)],**c**('Time','Diet'),function(df) **mean**(df$weight))  
**colnames**(x)[3]='mean'  
**View**(x)  
**library**(ggplot2)  
p=**ggplot**(x, **aes**(Time,mean))  
p+**geom\_point**(**aes**(color=Diet),size=4,alpha=1/2)+**geom\_line**(**aes**(color=Diet))



**Conclusions drawn**