* ANALYTICS: The science of using data to build models that lead to better decisions that add value to individuals, to companies and to institutions.
* The first example : IBM WATSON challenges
  + Deep Blue ===> a computer to compete against best human chess players to restrict the people thought.
  + Blue Gene ===> a computer to map human genome to challenge the computer speed and performance.
  + A computer that can compete at JEOPARDY, a popular game show.
* WATSON is a supercomputer with 3000 processors and massive database and is used in many applications such as finding best treatment for cancer.
* Second example : eHARMONY
  + A popular online dating site in US that focuses on long term relations
  + It first predicts whether the user is compatible using “29 different dimensions of personality”.
  + It finds a high quality match for every one using linear regression.
  + It collects data from users using 436 questionnaire.
* Third Example: FRAMINGHAM HEART STUDY
  + A modern medical study that is ongoing study of Framingham residents.
  + Most of the common knowledge we know about heart disease is from this
    - High BP
    - Clogged arteries are not normal
    - Smoking; all these can cause heart disease.
  + The data is collected from people with questionnaire for every 2 years about physical, behavioral characteristics.
  + Used regression to predicts whether the patient can develop heart disease in next ten years or not.
  + It helped in:
    - BP drug development
    - Many other clinical prediction rules
    - Predictions for patients.
* Fourth example: D2HAWKEYE
  + A medical software company that combines dat with DA to improve quality and cost management in healthcare.
  + It took data from health care industries and doctors insight regarding risk factors and other gender and age factors.
  + Predicts future health care costs to identify patients who need more attention.
* What is R:
  + A s/w environment for data analytics, statistical computing and graphics.
  + Originated from S programming language and developed by Robert Gentleman and Ross Ihaka in University of Auckland in 1990s.
* Why R?
  + Free open source, widely used
  + Can rerun previous work, Nice visualizations and graphics
* R console:
  + It shows a pointer(“> “) that you can perform operations here.
  + You can directly give expressions like 2+6, 5^4 and some basic math functions like sqrt, abs that works same as basic calculator and returns the output n the new line.
  + We can assign values to variables using “=” or “<-” symbols.
  + ls() helps to get all the variables in the console till now.
  + Vectors: A series of characters or numbers that refers to same object.
    - c(2,3,4,5); c for combine.
    - C[1] ==> 2, c[3] ==>4
    - seq(0,10,2) ⇒ 0,2,4,6,8,10
  + Data frame is collections of data objects into single.
    - Country = c(“abc”, “qwe” , “xyz”)
    - Value = c(1,2,3)
    - D = data.frame(country,value) ⇒ country value

Abc 1

Qwe 2

Xyz 3

* + - To add another column in data frame
      * D$population = c(987,9876,9987) or
      * cbind(D, population)
    - To combine 2 dataframes
      * rbind(dataframe1, dataframe2)
  + getwd() give the folder path that contains the dataset that have been selected.
  + To add the data in the dataset file to a dataframe
    - WHO = read.csv(“WHO.CSV”)
  + To view the data
    - str(WHO) that gives the structure of the data
    - summary(WHO) that gives numerical summary of each variable.
    - subset(WHO, Region == “Europe”) give the data of europe region.
  + To write into the csv file
    - write.csv(dataframe\_name, “csv file name”)
  + To remove a variable
    - rm(variable\_name)
  + To acces the variable in the data frame
    - Dataframe\_name $variable\_name
  + To get min value in the dataframe
    - which.min(dataframe\_name$ variable\_name) ⇒ returns row number
  + To get a graph
    - plot(variable1, variable2)
  + To print histograms
    - hist(dataframe\_name)
  + Script files are used to save the commands that we run on the R console.
  + “#” symbol is used to write comments in the script file.
  + names(dataframe) gives all the names of the variables in it.
  + nrow(datframe) gives number of rows in it.
  + To get the match of a string in a dataframe
    - match(“string\_to\_ be\_matched”, dataframe\_name)
  + as.numeric(dataframe) converts the dataframe datatype to numeric.
  + Tapply is used to perform grouping on tables and it talkes 3 arguments
    - tapply(USDA$iron, USDA$Protein, mean)
      * Here USDA$iron is grouped into different groups based on different types of USDA$protein and then calculate mean for each group.
      * I.e.; if proteins levels re of 2 types then iron dataframe is categoriesed into 2 groups and mean is calculated for eqach group seperately.