Leadership

ANALYTICS PROGRAMMING - R

Naveen Kumar

Agenda

- Week 1 Summary Presentation
- Catapult Game Analysis and Discussion
- Mega Trends and Digital Future
- Introduction to R Studio
- Key R Operations
- Naming Conventions
- Summary and Conclusions

Megatrends That Shape The Digital Future

Big Data

- Big Data:
  - High Volume (Lots of it)
  - High Velocity (Accrues quickly)
  - High Variety (Different kinds)

- New technologies and techniques required to capture, store, and analyze big data

Cloud Computing

- Use the Internet as the platform for applications and data

- Applications that use to be installed on individual computers are increasingly kept in the cloud
  - e.g., Gmail, Google Docs, Google Calendar

- Can enable advanced analytics of massive amounts of Big Data

Mobile Devices

- Many believe that we're living in a post-PC era
- In the developing world mobile devices often leapfrog traditional PC's
- Implications:
  - Consumerization of IT
  - Bring Your Own Device (BYO D) to work is a major concern
  - Security concerns

Social Media

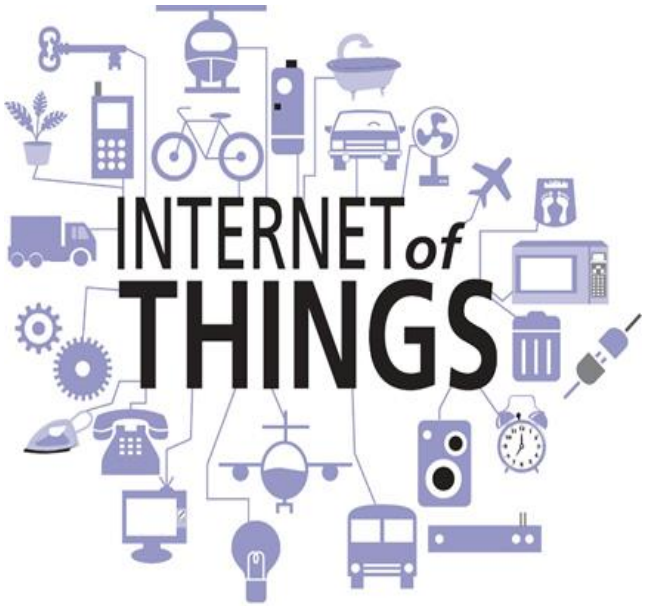
- Over 4.6 billion (and growing) Facebook users share status updates or pictures with friends and family

- Companies harness the power of the crowd by using social media to get people to participate in innovation and other activities

- Organizations use social media to encourage employee collaboration

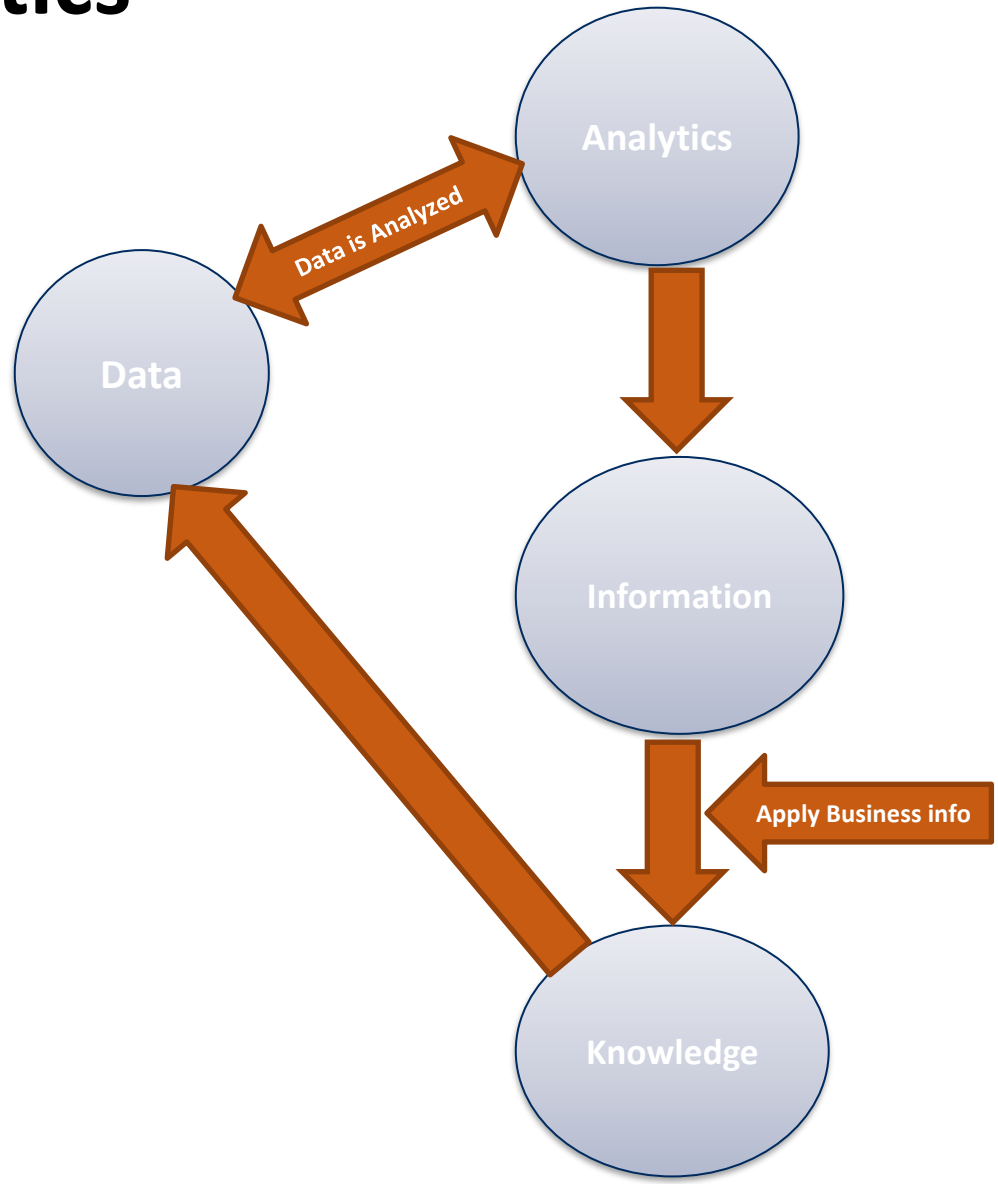
Internet of Things

- A broad range of physical objects that can automatically share data over the Internet
- The Industrial Internet of Things (IOT) enables the convergence of IT and operations technology to enable mass-produced customized products
- The Internet of **everything**?



Data Analytics

- Data Analytics
  - Helps to derive meaningful insights or information and subsequently knowledge from data and make critical business decisions
  - Data can be in various forms, either structured or unstructured



Covid-19 Data and Analytics

- To summarize COVID-19 literature using NLP
  - Roughly 28,000 papers have been published since the start of the outbreak
  - Makes it easy to catch up on recent trends
  - Provides summary on trends in the Covid-19 literature including
    - Most popular research areas
    - Number of new publications per week
    - Most proliferate authors, etc.

Covid-19 Data and Analytics

Predictive Analytics

- To predict COVID-19 infections and fatalities for various regions
  - Helps in planning for hospitalizations and ICUs needed to respond to the crisis.

- To predict increased number of infections
  - Due to a spread of the disease
  - Due to a lack/increase of our testing capabilities.

Basic Operations

- The three basic types of operations used in R GUI (also in R Studio) are:
  - Arithmetic operations
  - Relational operations
  - Logical operations

Built-in Functions

- mean(4,7,19)
  - Will return 4 for answer
- When calculating mean
  - First create a vector
  - Then calculate the mean of the vector
- Vector
  - Simplest type of data structure in R
  - It is a sequence of data elements of the same basic type
- A vector containing three elements: 4, 7, and 19
  - c(4,7,19) **NOTE**: c Must be in lowercase
- mean(c(4,7,19))
  - Will return 10 for answer
  - Note: lower case “m” (R is case sensitive)

Variables or Named Objects

- R works on named objects or assigning variables

- Assign a name to a variable or calculation to create or overwrite the named object.
  - Assigned using leftward

    > add.numbers <- 2+3

- If a name is specified, the result is not shown:

    > ans <- 13 + 11 + (17 - 4/7)

  Type the object name to see the result:

    > ans

    [1] 40.428

- > x <- c(4,7,19)

  > mean(x)

  It will return 10 for answer

Variables or Named Objects

- Data can be assigned to a named object:

```
> data <- c(3, 5, 7, 9)
```

- Text (character) data are surrounded by quotes:

```
> day <- c('Mon', 'Tue')
```

- You can use single or double quotes as long as each pair matches.

- Text values are reported within quotes:

```
> day
[1] "Mon" "Tue"
```

- Each line begins with an index value.

R Variables Naming Conventions

- Variable is used to hold value of a specific object or number
- There are some rules for valid variable names:
  - Can start with letter (lowercase or uppercase)
  - Can start with a period (.) (Should be followed by letter only)
  - Can be composed of letters, numbers, underscores or periods
  - Reserved words: R has some words which cannot be used for a variable name as they have an independent function in R

R Variables Naming Convention

- Variables should be as descriptive as possible to make sure others can make sense of the name

- Best practice – Use all lowercase letters and words separated with dots

  Example - first.variable

Google's R style Guide for Good Code

It gives ideas about how to write good R code (https://google.github.io/styleguide/Rguide.xml)

1. File Names: end in .R
2. Identifiers: variable.name (or variableName), F
3. Line Length: maximum 80 characters
4. Indentation: two spaces, no tabs
5. Spacing
6. Curly Braces: first on same line, last on own line
7. else: Surround else with braces
8. Assignment: use <-, not =
9. Semicolons: don't use them
10. General Layout and Ordering
11. Commenting Guidelines: all comments begin with
12. Function Definitions and Calls
13. Function Documentation
14. Example Function
15. TODO Style: TODO(username)

1. attach: avoid using it
2. Functions: errors should be raised using stop()
3. Objects and Methods: avoid S4 objects and metho

Logical Operations

- Logical operators are used to carry out Boolean operations (AND (&), OR (|) etc.)

- Operators ”&” and ”|”
  - Performs element-wise operation
  - Results match length of the longer operand

```
> x<-c(1,2,3)
> y<-c(1,2,3)
> z<-c(1,2,3,4)
> (x==y)&(x<4)
[1] TRUE TRUE TRUE
> (x==y)&(x<3)
[1]  TRUE  TRUE FALSE
> (x==y)|(x<3)
[1] TRUE TRUE TRUE
> (x==y)&(x==z)
[1]  TRUE  TRUE  TRUE FALSE
```

Limitations of R GUI

- One of the biggest challenges that GUI users face is being able to reproduce their work
  - Reproducibility is re-running everything on the same dataset if you find a data entry error

R Studio Over R GUI

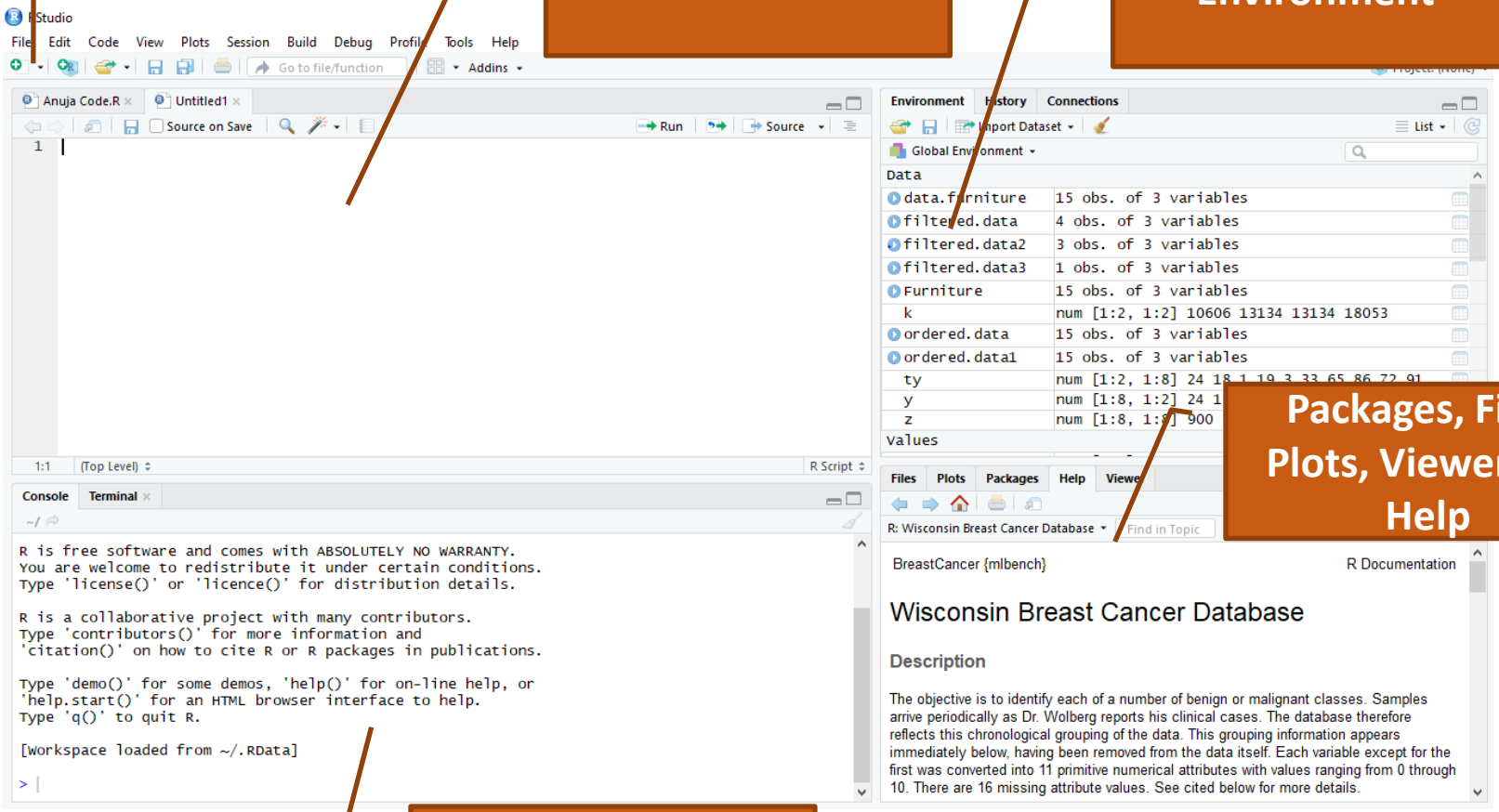
- R Studio is an integrated development environment (IDE) to support the development of R code
- All the variables created are displayed in the Global Environment
- It shows entry history
- R script can be saved and run

R Studio

Create an "R Script"
Ctrl+Shift+N

Source Code

Environment



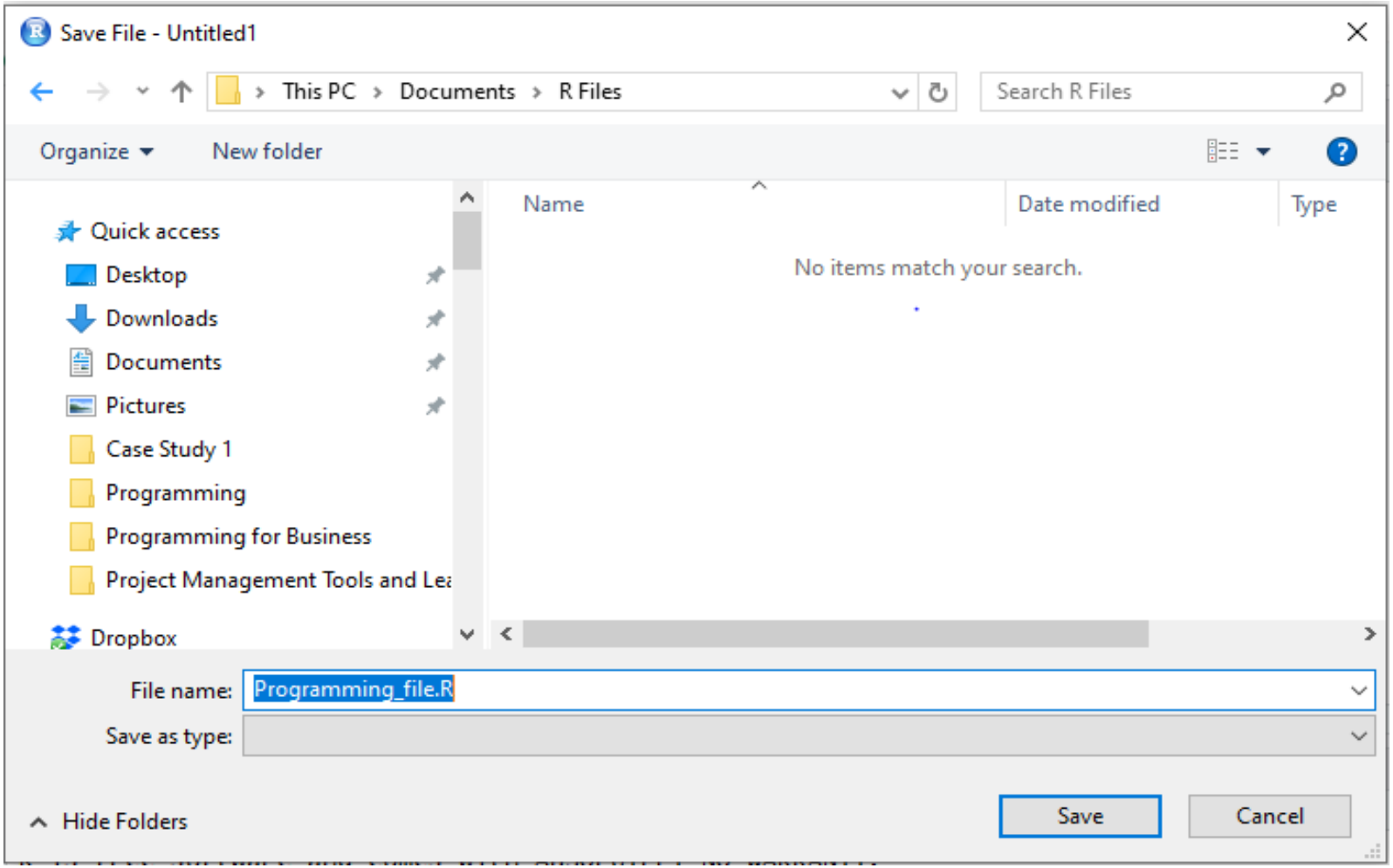
Packages, Files, Plots, Viewer and Help

R CONSOLE

.R File

- An .r file is a script written in R
- It contains code that can be executed within the R software environment
- R files may include commands that create objects (functions, values, etc.)

Saving a .R file



Importing Data (csv) File Into R

- Open R- script

- Right click on the .csv file you want to import and copy the location
- **How to set a working directory**
  - *setwd stands for "Set Working Directory"*
  - setwd (file location)
  - E.g. setwd ("c:\\users\\desktop\\documents")
    - * use 2 backslashes or 2 forward slashes on Windows machine
      setwd ("/Users/desktop/5032")
      * use 1 or 2 forward slashes on Mac

- **How to Open CSV file in R**
  - beer.data <- read.csv (file = “BeerDataExample.csv”, sep = ", ", header = TRUE)
    - *file= File containing the data*
    - *sep = Define Separators*
    - *header = It will not read first line as a data line*

Getting Familiar With the Dataset

dim(): Determines the number of dimensions (rows and columns) in the data.

summary(): Summary function returns the basic descriptive statistics of the vector, matrix or dataframe (like Minimum, 1st Quartile, Mean, Median, 3rd Quartile and Maximum)

dim(med.data)

class(med.data)

summary(med.data)

```
> dim(med.data)
[1] 20  9
> class(med.data)
[1] "data.frame"
> summary(med.data)
       ID          AgeInYears    Gender    Opinion    ChargesInDollars VisitTimeInMin
 Min.   : 1.00   Min.   :21.00   F:11   Min.   :1.0   Min.   : 24.00   Min.   : 31.00
 1st Qu.: 5.75   1st Qu.:32.75   M: 9   1st Qu.:1.0   1st Qu.: 45.75   1st Qu.: 55.50
 Median :10.50   Median :52.00          Median :2.0   Median : 58.50   Median : 66.50
 Mean   :10.50   Mean   :46.30          Mean   :2.6   Mean   : 65.90   Mean   : 72.95
 3rd Qu.:15.25   3rd Qu.:60.25          3rd Qu.:4.0   3rd Qu.: 92.75   3rd Qu.:102.50
 Max.   :20.00   Max.   :65.00          Max.   :5.0   Max.   :114.00   Max.   :120.00

   Insurance  PriorVisits         Date
 BCBS    :5   Min.   : 31.00   1/1/14 : 1
 Medicaid:4   1st Qu.: 55.50   1/15/14: 1
 Private :7   Median : 66.50   1/25/14: 1
 Self Pay:4   Mean   : 72.95   1/5/14 : 1
              3rd Qu.:102.50   2/13/14: 1
              Max.   :120.00   2/14/14: 1
                               (Other):14
```

Getting Familiar With the Dataset

#looking into the dataset

head(): Used to obtain first several rows (like 5 rows/10 rows/100 rows/..) of a vector, matrix or a dataframe

head(med.data)

```
> head(med.data)
  ID AgeInYears Gender Opinion ChargesInDollars VisitTimeInMin Insurance PriorVisits
1  1          33      M       5               58             64  Self Pay          64
2  2          21      F       2               59             69  Medicaid          69
3  3          56      F       1               78             81  Medicaid          81
4  4          53      M       2               24             31      BCBS          31
5  5          51      F       5               46             48   Private          48
6  6          22      M       1               30             38      BCBS          38
     Date
1  1/1/14
2  1/5/14
3 1/15/14
4 1/25/14
5  2/5/14
6 2/13/14
```

Summary and Questions