Heterogeneous Data Visual Analytics

Professor Georges Grinstein

*Student ID:01526533*

*Student Name: Janakiram Sundaraneedi*

**Final Report**

Contents:

* Introduction
* Presentation 1
* Presentation 2
* Presentation 3
* List of Data Science Packages
* Course Work Reference Websites and books
* Projects
* Overall Conclusion

**Introduction**

When I began this course I wasn’t well versed in presentations and data science visualization. My Engineering foundations had only given me a brief picture of data visualization. During the entirety of this course, I used various online sources and course work lectures to teach myself programming in R, filtering data through complex SQL queries and also how to present myself before an audience. I expected to work with extremely large datasets and fixing data loops holes, for instance data having NA or null values. I considered these projects which I had not tried before.

**Presentation 1**

**Dataset Source:** <https://www.ncdc.noaa.gov/stormevents/>

Considering a raw dataset these are minimum knowledge one should know

Size, Dimension, number of rows and Columns of any dataset.

**Data Cleaning**: Using is.na () function we can find out how many not applicable keys are present in each column. In some cases, instead of NA values, there will be ‘?’ values or other values which we need to implement algorithm to filter those values and replace it.

In some cases, removing NA value is appropriate but for some cases it may not be appropriate.

For values in continuous attribute we can replace the NA value with mean of that column.

For nominal or categorical attribute: Use Bayesian techniques to fill NA values.

Using these tasks, complexity will be higher.

**Data Manipulation:** Filtering out some of the columns for our requirement using package dplyr.

**How to plot:**

Basic 2-dimensional Plotting Method

Plot(data\_set\_name, aes(x\_attribute(continuous, nominal, categorical), y\_attribute(continuous, nominal, categorical), color\_by=some\_attribute(ranges))

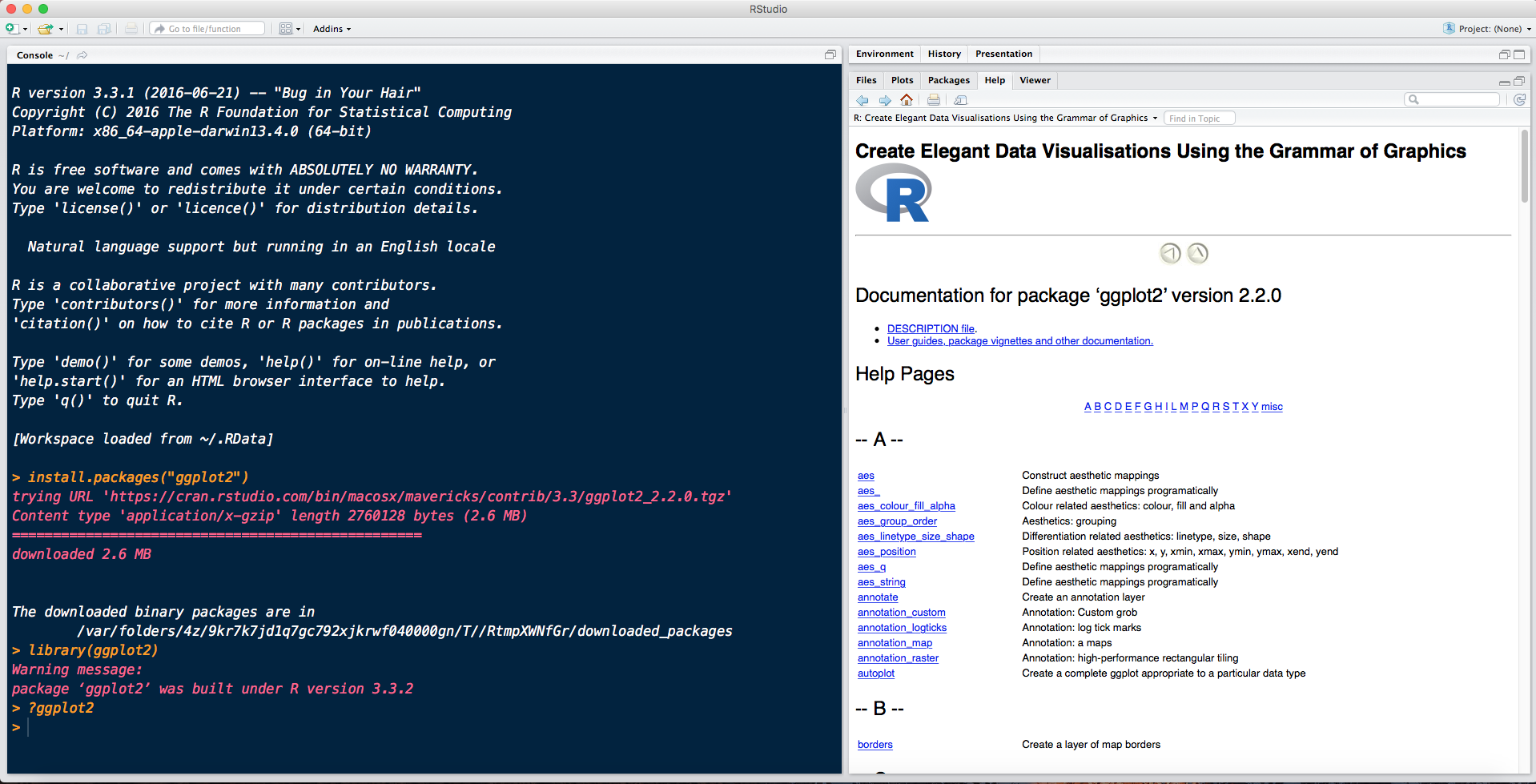
Above plot functions as set of argument like data set name, and its attributes assigning to x-axis and y-axis and more arguments which we can customize the plots according to our needs.

**Installing a package:**

Step1:install.packages("ggplot2")

Step2:library(ggplot2)

Step3:?ggplot2 #you can see documentation of any package on right side of window.



**Data Visualization: Plotly Package**

**Data transformation:**

Sometimes raw data will be skewed, concentrated on a particular area in a plot and it is difficult to come to conclusion. For this, we use transformation techniques like log, sqrt and 1/column\_name transformation to particular keys to better view it.

**Bootstrapping:**

Datasets are costly, bootstrapping random sampling technique help overcome this problem.

For instance, If we have dataset for 30 days’ period, Analyzing the 30 days’ dataset won’t be sufficient to come to conclusion. we can get other samples for 11 months using bootstrapping method. So that we can cut down costs of buying data.

**Presentation 2:**

Link: <http://rpubs.com/janakiram/datavis6>

**Data Merging:** Basic intuition taking data from different web API sources, and use the merging techniques to get the required data set for further processes.

Here we use ‘R’ functions:

Cbind: Column binding means adding new column to present dataset and number of rows should be equal while adding.

Rbind: Row binding means adding same attributes one by one.

Outer Join, Left Outer, Right Outer, Cross Join

But here we have created sample using random sampling technique, with sequence of numbers, ordinal and categorical attributes for 4 datasets. Sometimes we need to filter column names from different datasets and merge it depending upon the application we are dealing with it.

For the four data sets that I created, I used the above data merging techniques.

**Understanding Basics of plotting:** Before plotting, we should see how the data looks using str() function, and then choose which plot type is required and we should check what are continuous, Discrete, nominal and ordinal attributes.

**Data Visualization: ggplot2 package**

Examples of plots:

* Bar graph
* Density graph

**Case study:** Using Stents to prevent stokes:

Question: Does the use of stents reduce the risk of stroke?

Data Set Source: <https://www.openintro.org/stat/textbook.php>

Dataset dimension: 451 rows, 2 columns.

Treatment group: Received stent and medical management (medications, management of risk factors, and lifestyle modification)

Control group: Only medical management

Conclusion: It is proved 20% are likely to get strokes after 365 days from treatment group and 12% are likely to get from control group, which means 8% more people tend to get more heart attacks than the control group. Another assumption, samples are about 451 patients which is less, it is better that we should test on more samples.

**Note:** Used sqldf package for counting and filtering the number of patients in each group and concluding it.

**Reference:** https://drive.google.com/file/d/0B-DHaDEbiOGkc1RycUtIcUtIelE/view

**How to plot in Spatial Visualization:**

Rpubs link: http://rpubs.com/janakiram/leaflet

Reference: <https://rstudio.github.io/leaflet/>

**Presentation 3**

As you said in the class, try to implement neural networks for dataset

[https://www.youtube.com/watch?v=xbYgKoG4x2g (Watched](https://www.youtube.com/watch?v=xbYgKoG4x2g%20%20%20(Watched) 3 classes)

<https://ocw.mit.edu/courses/brain-and-cognitive-sciences/9-641j-introduction-to-neural-networks-spring-2005/index.htm>

https://github.com/ram978/ANN-Notes

**Basic approach for any packages to be tested in rstudio Software:**

Step1:install.packages("package\_name") #installing package

Step2:library(package\_name) # initializing package calling library function

Step3:?package\_name #after this you can see documentation of any package on the sub window

**Packages Learned and Used:**

**Sqldf:**

Reference:

<https://cran.r-project.org/web/packages/sqldf/sqldf.pdf>

**dplyr:**

Reference:

https://cran.r-project.org/web/packages/dplyr/dplyr.pdf

**tidy:**

Reference:

<https://cran.r-project.org/web/packages/tidyr/tidyr.pdf>

**ggplot2:**

**Reference:**

<https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf>

**googleVis**

Reference: <https://cran.r-project.org/web/packages/googleVis/vignettes/googleVis.pdf>

**Rcharts:**

Reference: <http://ramnathv.github.io/rCharts/>

**Leaflet:**

Reference: <https://cran.r-project.org/web/packages/leaflet/leaflet.pdf>

**Highchart:**

Reference: <http://jkunst.com/highcharter/>

**Swirl**: (Statistical learning package)

Reference: Install.package(“swirl”)

**Data Visualization Software**

**Tableau:**

Reference:<http://www.tableau.com/>

**Iweave**

Reference:<http://www.iweave.com/>

Shiny app

Reference: https://shiny.rstudio.com/

**Lecture Work Reference Websites and books:**

<http://www.terrafly.com>

<https://www.spotify.com>

<http://www.iweave.com>

<https://www.treemap.com>

[www.buckets.com](http://www.buckets.com)

<http://finviz.com>

<http://flowingdata.com>

<http://visualoop.com/>

<https://venngage.com/>

**Books:**

Data mining algorithms : explained using R

Data mining for business analytics : concepts, techniques, and applications with XLMINER

Information Visualization

Open IntroStatistics

**Projects:**

Project 1:

Fitbit Application:

https://github.com/ram978/RepData\_PeerAssessment1-1

http://rpubs.com/janakiram/Activity\_monitoring

Project 2:

Hospital Ranking:

https://github.com/ram978/Programming-week4project

Project 3:

Consumption Utilities:

https://github.com/ram978/ExData\_Plotting1

Project 4:

Data Cleaning and Merging

https://github.com/ram978/Getting-and-Cleaniing-Data/tree/master/Desktop/dataclean

Project 5:

Storm data analysis :

<http://rpubs.com/janakiram/storm>

Project 6:

Air Pollution Analysis

<https://www.coursera.org/learn/exploratory-data-analysis/peer/b5Ecl/course-project-2/submit>

Username:jsundara@cs.uml.edu

Password:basketball9676

**Video Conferences:**

https://vimeo.com/42529425

https://www.youtube.com/watch?v=TXUOtQwUL4c

https://www.youtube.com/watch?v=uMWASyT7DMA

What I am doing right now:

Using R programming to understand statistics better.

Statistics with R:

Introduction to Probability and Data

Inferential Statistics

Linear Regression and Modeling

Bayesian Statistics

Statistics with R capstone.

**Overall Conclusion**: I am 100% confident that I can handle any type of data and plot it. Furthermore, make use of machine learning and data mining techniques. I learned how to be better speaker.