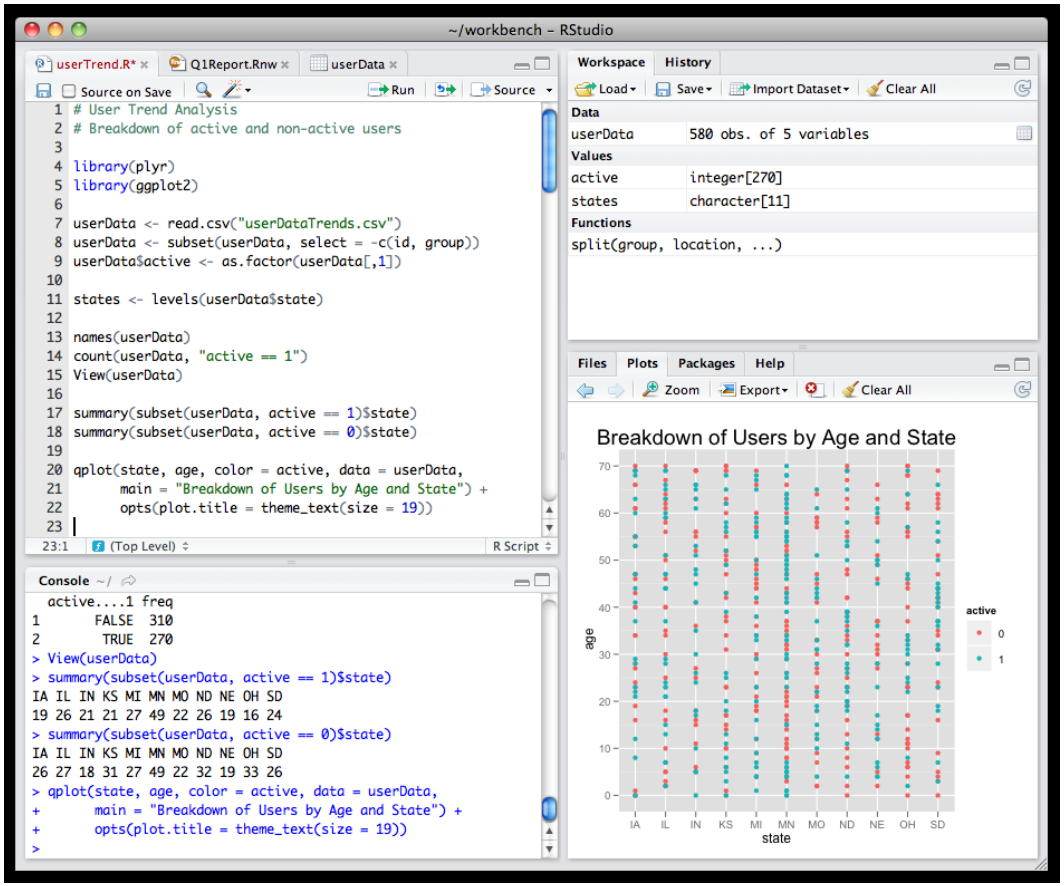
# CSCI446/946 Big Data Analytics

## Week 3
## Tutorial on Data Analytic Methods Using R

School of Computing and Information Technology
University of Wollongong Australia

# Task One

- Launch RStudio and become familiar with it



# Task One

- Run the first example with Rstudio
  - Understand what each function does
  - The .csv file can be downloaded from Moodle

```
# import a csv file of the total annual sales for each
customer
sales <- read.csv("./yearly_sales.csv")

# examine the imported dataset
head(sales)
summary(sales)

# plot num_of_orders vs. sales
plot(sales$num_of_orders,sales$sales_total,
     main="Number of Orders vs. Sales")
```

# Task One

```
# perform a statistical analysis (fit a linear
regression model)
results <- lm(sales$sales_total ~ sales$num_of_orders)
results
summary(results)

# perform some diagnostics on the fitted model
# plot histogram of the residuals
hist(results$residuals, breaks = 800)
```

- Can you create a R script to run all the commands in one shot?

# Task Two

- Data Import and Export

```
sales <- read.csv(“./yearly_sales.csv")

# add a column for the average sales per order
sales$per_order <- sales$sales_total/sales$num_of_orders

# export data as tab delimited without the row names
write.table(sales,"sales_modified.txt", sep="\t",
row.names=FALSE)
```

# Task Two

- Automatically save plots

```
# export a histogram to a jpeg
jpeg(file="c:/data/sales_hist.jpeg") # create a new jpeg
file
hist(sales$num_of_orders) # export histogram to jpeg
dev.off() # shut off the graphic device
```

# Task Two

- Save the workspace environment
  - `save.image()` function to create .Rdata file
  - `load()` function to load .Rdata file

- If you have more time, try more code in the lecture notes…

# Tips

- # Check the current directory (Where I am)
- `getwd()`

- # Set your working directory, for example
- `setwd("G:/CSCI446_Big_Data_Analytics/Tutorial/Week3/Chapter 3/")`

- # List all the files on the path returned by getwd().
- `list.files()`

- # View the content of a variable (Note R is case-sensitive)
- `View(sales)`

# Tips

- # List all the objects in the current workspace environment
- `ls()`

- # Clear an object "sales" from your workspace environment
- `rm(sales)`

- # Clear your whole workspace environment
- `rm(list=ls())`

- # Clear all the display in Console (clear screen)
- `Ctrl+L`

# Tips

- In Rstudio you can run the entire script by pressing `ctrl+shift+enter` without selecting any code
- In addition, there is a shortcut to source the current script file (ctrl+shift+s), which runs the script without echoing each line.

# Q&A


**Images Courtesy of Google Image**