Cleaning Data in R

Hands on

Motivation

- Why Clean Data?
  - Data preparation is an essential part of any kinds of analysis
- Why R
  - Free / Open source
  - Lots of visualization and statistical packages to use

Objective

- Not about teaching R as a programing language
- Immediate hands-on to clean “dirty data”
  - Cover just enough R knowledge for the workshop

Setup

- R software
  - R Project

    http://www.r-project.org
- Integrated Development Environment for R
  - Rstudio http://www.rstudio.com/products/RStudio/

Quick Introduction to R language

- Assignment
  - Assign a value 1 to a name x
    - x <- 1
    - x = 1
- Basic Types
  - Numeric (1, 2, 3, …)
  - Integer (1L, 2L, 3L, …)
  - Characters
  - ...

Quick Introduction to R language

- Data Structures
  - Data frame
    - storing data tables
- Need help/information?
  - Just type ?<function name>

Quick Introduction to R language

- Getting data in
  - **comma delimited files**
    - **read.csv()**
  - xml
  - json
  - database
- Writing data out
  - **comma delimited files**
    - **write.csv()**

Exercise 1

1. Read in a comma delimited flat file
    - contains median housing prices of 5 room Ang Mo Kio (a region) flats in Singapore
2. Clean the data
    - Carry out some data manipulation functions
3. Visualize data

Instructions

- Open up ../CleaningDataInR/ex1/ex1.pdf
- Codes are provided
  - Copy and paste :)
  - Understand what each functions does

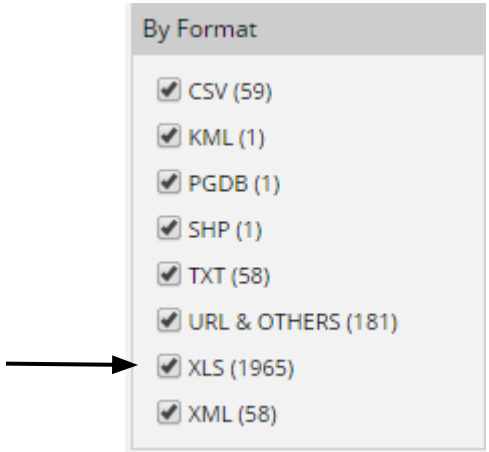
What's Next?

- The data previously used actually came from an Excel file from Housing Development board
  - Median resale prices by town, flat type per quarter

    http://data.gov.sg

What's Next?

- Excel is the de-facto standard in enterprise
  - Most datasets in data.gov.sg are in excel :(
  - R provides various packages for you to read it :)



Exercise 2

- Read an excel file
  - median housing prices from data.gov.sg
- Extract a particular data
  - only need to use the 5 room Ang Mo Kio flats from the 29 sheets of the workbook
- Write it to csv

Instructions

- Open up ../CleaningDataInR/ex2/ex2.pdf
- In view of time we will only do a walkthrough of the codes
  - You can do this as homework

Exercise 3

- Read in another flat file
  - a list of tweets
- Remove words
  - Remove common dictionary words
- Visualize the data
  - word cloud

Instructions

- Open up ../CleaningDataInR/ex3/ex3.pdf
- You will need to download some external R libraries for this exercise

Getting Data from Twitter

- The tweets were harvested using Twitter API
- R provides various package for you to search hashtag and analyse tweets

Exercise 4

1. Authenticate yourself with Twitter
2. Search using your own #hashtag
3. Write out to flat file

Instructions

- Open up ../CleaningDataInR/ex4/ex4.pdf
- You will need
  - internet access
  - twitter account
  - external R libraries
- You can do this during your own free time :)

Thank You

Resources

- A free course on how to get and clean data in R
  - https://www.coursera.org/course/getdata
- Ebook on Data Cleaning in R
  - http://cran.r-project.org/doc/contrib/de_Jonge+van_der_Loo-Introduction_to_data_cleaning_with_R.pdf
- Tutorial Reference
  - http://www.rdatamining.com/examples/text-mining
  - http://theminingbook.blogspot.com/2014/03/r-oauth-for-twitter.html