STAT3612 Assignment 1: Data Exploration

Date: February 16, 2017

Sumbit in HTML or PDF format by email to Jason You on or before February 28, 2017.

Today there are plenty of open datasets available on the Internet; see e.g. the Kaggle Datasets at `https://www.kaggle.com/datasets`. In this assignment you are required to explore the Times Higher Education World University Ranking dataset, with data description provided by the website `https://www.kaggle.com/mylesoneill/world-university-rankings`.

Step 1. (30%) Download the "`timesData.csv`" file. Select the subdata with top 200 world-ranked universities. Identify all the numeric types of variables that can be represented by floating point numbers, then summarize them by `R:summary()`.

`Hints:` The numerical values formatted as character strings can be converted to floating point numbers with R commands like `as.numeric(as.character(x))`, `as.numeric(sub(",", "", x))`, `substring(x, start, stop)`.

Step 2. (40%) a) Draw the histogram for the variable `total_score` for all years; b) Draw the side-by-side boxplots of `total_score` for different years; c) For year 2016 data, draw a pairwise plot of the following four variables: `total_score, teaching, research` and `citations`; d) compute the correlation matrix of these four variables.

Step 3. (30%) Perform data subsetting with conditions `country == "Hong Kong"` and `year > 2011`, followed by a graphical method to show the ranking changes of local universities over the years.

`Hints:` You may either use base plots (upon necessary data formation) or `dplyr` and `ggplot2` packages.