Homework 5
STAT 4410/8416 Section 001
Fall 2014
Due: November 25, 2014 by midnight

1. **Scrapping HTML data:** We often obtain data from Wikipedia. This exercise will guide us to collect some data about the native speakers of some common languages. The information can be obtained from the following link.

   http://en.Wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers

   Now answer the questions below.

   (a) Read all the HTML tables available in the above link and store the result in an object called `tables`.

   (b) Now notice that the table in this link does not have any ID to specifically get the data from. But if you examine the source of the page, the first table is the data table. Thus we pick first table in the list of tables as our data. Store the data of first table in an object called `datRaw`.

   (c) We are particularly interested about columns 2 and 3 of `datRaw`. Subset columns 2 and 3 of `datRaw` and store the data in `dat`. Give the column names as language, nativeSpeaker. Display some data from `dat`.

   (d) Notice that the data is not clean. We have a '*' in the number of native speaker. Also for Arabic and the Spanish, the number just messed up. We have to carefully review the data before we can use it. But first let us remove the '*' from column 2 and make the column numeric. Also notice in the first column we have some non-english character that we don't want to print. Clean the data and store it in `cleanDat`. Display some cleaned data.

   (e) Now let us plot the data to show language wise ranks and their relative position. For this we plan to select only top 20 languages based on number of speakers. Generate a bar chart showing the top 20 languages. Order the bars according to the number of speakers.

2. **Working with databases:** For this exercise we will use `MySQL` database available in the data science lab or the `datascienceVM`. Answer the following questions.

   (a) Write down the connection string that would establish a connection to the `MySQL` database `trainingDB`.

   (b) Write down a `SQL` command to select pclass, sex, survived and their average age from the titanic table. Store the selected data in data frame `avgAge` and display all the aggregated data.

   (c) Now generate a line plot showing average age vs pclass colored by survived and faceted by sex.

   (d) Use the package `dplyr` to obtain the same result as you did in question 2b. Display the results and the underlying `SQL` command used by `dplyr`.

   (e) Find the name, age, sex and pclass of the 5 oldest and 5 youngest persons who died. Remove the people whose age information are not available for this computation.

3. **Exploring data:** Explore the crime data by downloading it from the blackboard. Provide nice tables and some plots that explain some important features revealed from the data. Discuss what you have found.