ISE 5103 Intelligent Data Analytics
Homework #2

Instructor: Charles Nicholson

See course website for due date

Learning objective: Explore and visualize data.

Submission notes:

1. Clearly identify each problem (e.g. Problem 1a, Problem 2b, etc.)
2. All *relevant* computer output should be provided unless noted otherwise.
3. The R code itself is part of your solution – make sure to *provide comments* on what your code is doing. Keep it clean and clear!
4. You will submit your complete R script. Note: include `library` commands to load *all* packages that are used in the completion of the assignment. Place these statements at the top of your script.
5. You may use "R Markdown" to *help* with your submission. However, please edit the final submission to clearly and concisely respond to the questions.
6. Please note, going forward it is almost never appropriate to "print out" lots of data excerpts as part of your homework submissions. I have tried to note a few places you certainly do *not* need to print out large amounts of data for submission. *The goal is to limit your complete homework submission to 10 pages.*
7. Do not zip your files for submission. Submit exactly two files. Name the files "LastName-HW1" with the appropriate file extension (that is, .R, .pdf, .docx, or .html)

1 Concordance and Discordance

Given the following two vectors, $x = (3, 4, 2, 1, 7, 6, 5)$ and $y = (4, 3, 7, 6, 5, 2, 1)$, calculate the number of concordant and discordant pairs.

2 Generating data and advanced density plots

To complete this section you may need to do some research on new functions in R. Hints are provided.

(a) Create a data frame `df` with 500 rows and 4 variables: `a, b, c` and, `d` Each variable should contain data generated randomly from a *different type* of distribution (e.g. `rnorm` generates normally distributed data randomly; there are several other similar commands available).

The data frame will look something like the following example:

a	b	c	d
1.2	0.4	-0.1	4.9
0.9	1.3	0.9	-0.7
		...	

Create data frame `df2` from the data frame `df` data by "reshaping" the data into two columns: `groupVar` and `value`.

The variable `value` will contain all of the random values from `df`. The variable `groupVar` will contain the original associated variable name. The new data frame will have 2,000 rows. (Hint: checkout the `reshape2` package and the `melt` command). Example data frame would look something like the following:

groupVar	value
a	1.2
a	0.9
b	0.4
b	1.3
...	

Note: please do not "print out the data" as a part of your homework submission, you may use the "head" function to show a small excerpt if desired; otherwise, code is sufficient.

(b) Using `df2`, plot the densities of each distribution overlaid on each other on one plot. Each density should have some level of transparency and be colored differently. (Hint: the reshaping of the data you completed in a will work very well with the `ggplot2` package and functions `qplot` or `ggplot`.)

3 Shark Attacks

For this problem you will access some real-world data associated with global shark attacks from the site, "Shark Attack Data" (`http://www.sharkattackdata.com`).

The link to the incident log is found here: `http://www.sharkattackfile.net/incidentlog.htm`. This data exists to "provide current and historical data on shark/human interactions for those who seek accurate and meaningful information and verifiable references."

In this problem, you will perform some basic explorations of the data. A CSV (comma separated values) text file of the incident log is available in the course website. Please download the **ISE 5103 GFAF5.csv** file for this problem.

(a) The data contains historical information ranging from early dates till August, 2015. IF you are interested in looking at trends in attack frequency (e.g. `http://www.sharkattackdata.com/country-overview/united_states_of_america`). What issues, if any, might impact your evaluation of the timeliness question of data quality?

(b) The data is not "clean", but at least the columns are at least descriptive. Let's limit our analysis to relatively recent incidents. Create a new data frame, `GSAFdata`, which contains incidents occurring on or after the year 2000 (i.e., rows 4070 through the end of the data frame) *Note: please do not "print out the data" as a part of your homework submission; code is sufficient.*

(c) The `Date` field is currently stored as character field and listed as a "factor". Use the `as.Date` command to create a new variable in the data frame which converts the factor to an R date type. (See `http://www.ats.ucla.edu/stat/r/faq/string_dates.htm` for hints if needed.)

(d) What percentage of the new date field is missing? (Note: should be less than 10%) Why is the data missing?

(e) Delete all of the records in `GSAFdata` that have missing values for the new date field. *Note: please do not "print out the data" as a part of your homework submission; code is sufficient*

(f) It has been said that shark attacks occur as Poisson process[1], and thus that the time between attacks is exponentially distributed. Such an analysis should occur based on the distribution of "days between" attacks.

[1] `http://www.sciencedaily.com/releases/2001/08/010823084028.htm`

i. Use the `diff` command to help you create a vector `daysBetween` with days between attacks. Notice that the vector `daysBetween` will have one less value than the number of rows in `GSAFdata`. Add a missing value as the first element of `daysBetween` and add the revised vector as a new variable in `GSAFdata`. *Note: please do not "print out the data" as a part of your homework submission; code is sufficient*

ii. Run and comment on the results from `boxplot` and `adjbox` for `GSAFdata$daysBetween`.

iii. Is the Grubb's test, the Generalized ESD test, both, or neither appropriate for this data? Support your answer.

(g) Use the `qqplot` and `qqline` commands to help you visually evaluate the claim that time (in days) between shark attacks is exponentially distributed. See the example provided in the documentation of `qqplot` to help you get started.

(h) Use the package `fitdistrplus` and the commands such as `cdfcomp, denscomp, ppcomp` and `qqcomp`, as well as `gofstat` to evaluate the fit of the `GSAFdata$daysBetween` to the exponential distribution.

(i) How do you respond to the claim that shark attacks occur as a Poission process? That is, what is your conclusion? is there an obvious answer? are there any issues with the statistical approaches, visualizations, or data?

4 Missing Data

The `freetrade` data frame from the `Amelia` package has economic and political data on nine developing countries in Asia from 1980 to 1999. The 9 variables include year, country, average tariff rates, Polity IV score, total population, gross domestic product per capita, gross international reserves, a "dummy" variable for if the country had signed an IMF agreement in that year, a measure of financial openness, and a measure of US hegemony. Unfortunately, this data has missing values.

(a) Explore the "missingness" in the `freetrade` using your choice of methods, e.g. from packages `VIM`, `mice`, `Amelia`, and/or others.

(b) Implement your own statistical test (e.g. ANOVA, t-test, chi-square, etc.) to determine if the missingness in the `tariff` variable is independent with the `country` variable. Does your answer change if you remove Nepal or if you remove the Philippines? Discuss why. (Note: a short description of using the chi-square goodness of fit test is available in the course website.)

5 Principal Component Analysis

For this problem you will perform principal component analysis on a variety of data sets.

(a) Mathematics of principal components

i. Using the data `mtcars`, create the correlation matrix of all the attributes and store the results in a new object `corMat`.

ii. Compute the eigenvalues and eigenvectors of `corMat`.

iii. Use `prcomp` to compute the principal components of the `mtcars` attributes (make sure to use the `scale` option).

iv. Compare the results from (ii) and (iii) – Are they the same? Different? Why?

v. Using R demonstrate that principal components 1 and 2 from (iii) are orthogonal. (Hint: the inner product between two vectors is useful in determining the angle between the two vectors)

(b) The `HSAUR2` package contains the data `heptathlon` which are the results of the women's olympic heptathlon competition in Seoul, Korea from 1988. A scoring system is used to assign points to the results from each of the seven events and the winner is the woman who accumulates the most points over the two days.

i. Look at histograms of each numerical variable using `apply(heptathlon[,1:8],2,hist)` (note: these are not labeled well, but that is okay for now since you just want to take a quick look at the distributions). From this quick inspection, are the distributions reasonably normal?

ii. Examine the event results using the Grubb's test. According to this test there is one competitor who is an outlier multiple events: Who is the competitor? And for which events is there statistical evidence that she is an outlier? Remove her from the data.

iii. As is, some event results are "good" if the values are large (e.g. highjump), but some are "bad" if the value is large (e.g. time to run the 200 meter dash). Transform the running events (`hurdles, run200m, run800m`) so that large values are good. An easy way to do this is to subtract values from the max value for the event, i.e. $x_i \leftarrow x_{\max} - x_i$.

iv. Perform a principal component analysis on the 7 event results and save the results of the `prcomp` function to a new variable `Hpca`.

v. Use `ggibiplot` to visualize the first two principal components. Provide a concise interpretation of the results.

vi. The PCA projections onto principal components $1, 2, 3, \ldots$ for each competitor can now be accessed as `Hpca$x[,1]`, `Hpca$x[,2]`, `Hpca$x[,3]`, ... Plot the heptathlon `score` against the principal component 1 projections. Briefly discuss these results.

(c) Handwriting Analysis

Handwritten digits, automatically scanned from envelopes by the U.S. Postal Service have been deslanted and size normalized, resulting in 16 x 16 grayscale images (Le Cun et al., 1990).

The image data has been transformed into text data: 256 variables associated with the grayscale value of every pixel for each image. Each record represents someone's handwriting of a single digit 0 thru 9.

These data were kindly made publicly available by the neural network group at AT&T research labs. There is one file available per digit on the course website. The actual image files are available there as well. (Note: the images are interesting artifacts and not necessary for this assignment.)

i. Choose three different digit data sets to download and analyze using PCA. Please include and comment on the screeplots to justify your choice about the number of principal components that best represent the data in lower dimensional space.

ii. PCA in general is very useful in image analysis (e.g. handwriting and face recognition). Explain why conceptually PCA would be particularly well-suited for image analysis.

6 Kaggle.com

To complete this problem you need to join Kaggle.com – this is site where "the world's largest community of data scientists compete to solve your most valuable problems." Kaggle.com hosts analytics competitions sponsored by various organizations for significant cash awards.

(a) Explore the site, competitions, and data – choose one data set to download from a competition to download. Provide the url and a *brief* description of the data (one sentence is fine!).

(b) Perform an initial basic exploratory analysis of the data which includes at minimum: the number of rows, number of variables, descriptive statistics, a selection of visualizations, information on missing value counts, and some form of outlier labeling/detection.

Note: Installation of the `ggbiplot` is slightly more involved than many R packages.
The following steps should help:

```
install.packages("devtools")
library(devtools)
install_github("ggbiplot", "vqv")
library(ggbiplot)
```