Stat 133 HW07: Working with XML Data

Your name and SID

Introduction

The main purpose of this assignment is to work with XML data and introduce you to the R package "XML" (by Duncan Temple Lang). You'll have the opportunity to:

a) work with RStudio Projects
b) keep practicing regular expressions
c) getting better at data manipulation
d) create more data visualizations (now with maps)
e) last but not least: becoming familiar with some of the data and general aspects for your final project

Submit your assignment to bcourses, specifically turn in your R, Rmd and pdf files.

RStudio Project

Create a new RStudio Project for this assignment: name it `"homework07"`. If you need more information about using RStudio Projects you can check the supporting tutorial (very basic info):

https://support.rstudio.com/hc/en-us/articles/200526207-Using-Projects

---

IBTrACS XML Data

You'll be working with one of the year-data XML files from the *International Best Track Archive for Climate Stewardship* (IBTrACS) located at ftp://eclipse.ncdc.noaa.gov/pub/ibtracs/v03r06/wmo/cxml/year.

Specifically, you'll be working with the file from 2010:

Year.2010.ibtracs_wmo.v03r06.cxml

Download a copy of the file to your RStudio Project (name it `"Year.2010.ibtracs_wmo.v03r06.cxml"`)

Data Processing and Cleaning

Create an R script `"data_cleaning.R"` dedicated to the processing part. Your mission here is to work with the functions from the "XML" package in order to extract various pieces of data from the xml file. Focus on using **XPATH** expressions and the functions `getNodeSet()` and `xpathSApply()`.

The goal is to create a cleaned data.frame with the following variables:

- `name`: name of the storm (e.g. ANJA) as `character`
- `date`: date (e.g. 2009-11-13) as `Date`

- `time`: time (e.g. 06:00:00) as `character`
- `latitude`: latitude (e.g. -9.50) as `numeric`
- `lat_deg`: latitude degrees (e.g. "N") as `character`
- `longitude`: longitude (e.g. 72.50) as `numeric`
- `lon_deg"`: longitude degrees (e.g. "E") as `character`
- `press`: pressure (e.g. 1006.0) as `numeric`
- `wind`: wind speed (e.g. 0.0) as `numeric`

Notice that `-999` is the value used in pressure and wind speed for missing data. This implies that you have convert those values to `NA`.

Remove those storms with names `MISSING` and `INVEST`.

This is what the first and last rows of the cleaned data.frame should look like:

```
##   name       date     time latitude lat_deg longitude lon_deg press wind
## 1 ANJA 2009-11-13 06:00:00     -9.5       N      72.5       E  1006    0
## 2 ANJA 2009-11-13 12:00:00    -10.2       N      71.9       E  1004    0
## 3 ANJA 2009-11-13 18:00:00    -11.1       N      71.4       E  1002   25
## 4 ANJA 2009-11-14 00:00:00    -11.9       N      71.1       E   999   28
## 5 ANJA 2009-11-14 06:00:00    -12.5       N      70.9       E   996   33
```

```
##              name       date     time latitude lat_deg longitude lon_deg
## 2422 OMEKA:TD1219 2010-12-21 18:00:00     29.0       N    -172.6       E
## 2423 OMEKA:TD1219 2010-12-22 00:00:00     30.7       N    -172.2       E
## 2424 OMEKA:TD1219 2010-12-22 06:00:00     32.2       N    -172.4       E
## 2425 OMEKA:TD1219 2010-12-22 12:00:00     33.3       N    -172.3       E
## 2426 OMEKA:TD1219 2010-12-22 18:00:00     34.1       N    -172.0       E
##      press wind
## 2422  1002   35
## 2423  1002   30
## 2424  1004   30
## 2425  1006   25
## 2426  1008   25
```

Export the clean data.frame to your Rstudio Project as a csv file named `"ibtracs_2010.csv"`.

Data Analysis

Create an R script `"data_analysis.R"` dedicated to the analysis part. Read in the clean data (contained in `"ibtracs_2010.csv"`) in R as a data.frame named `storms`.

Perform an exploratory analysis of the pressure, wind speed, and duration of storms. Get descriptive summaries (min, max, quartiles, mean, std deviation, etc), as well as histograms and/or boxplots.

In addition, answer the following questions (you'll also have to include them in the report):

- Total number of storms (in 2010)
- Number of storms with winds $\geq$ 35 knots (tropical storms)
- Number of storms with winds $\geq$ 64 knots (hurricanes)
- Number of storms with winds $\geq$ 96 knots (major hurricanes)

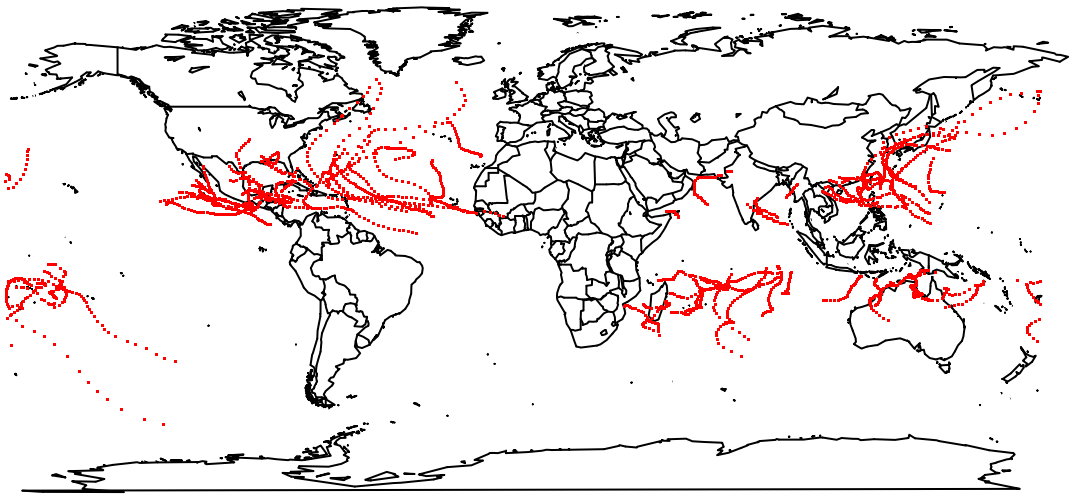
- Number of storms per hemisphere (north and south)
- Frequency table with data points per month (month in words)
- Frequency table with data points per month (month in words) in northern hemisphere
- Frequency table with data points per month (month in words) in southern hemisphere
- Name of storm that lasted more days
- Name of storm with maximum wind speed (and speed value)
- Name of storm with minimum pressure (and pressure value)

Data Visualization

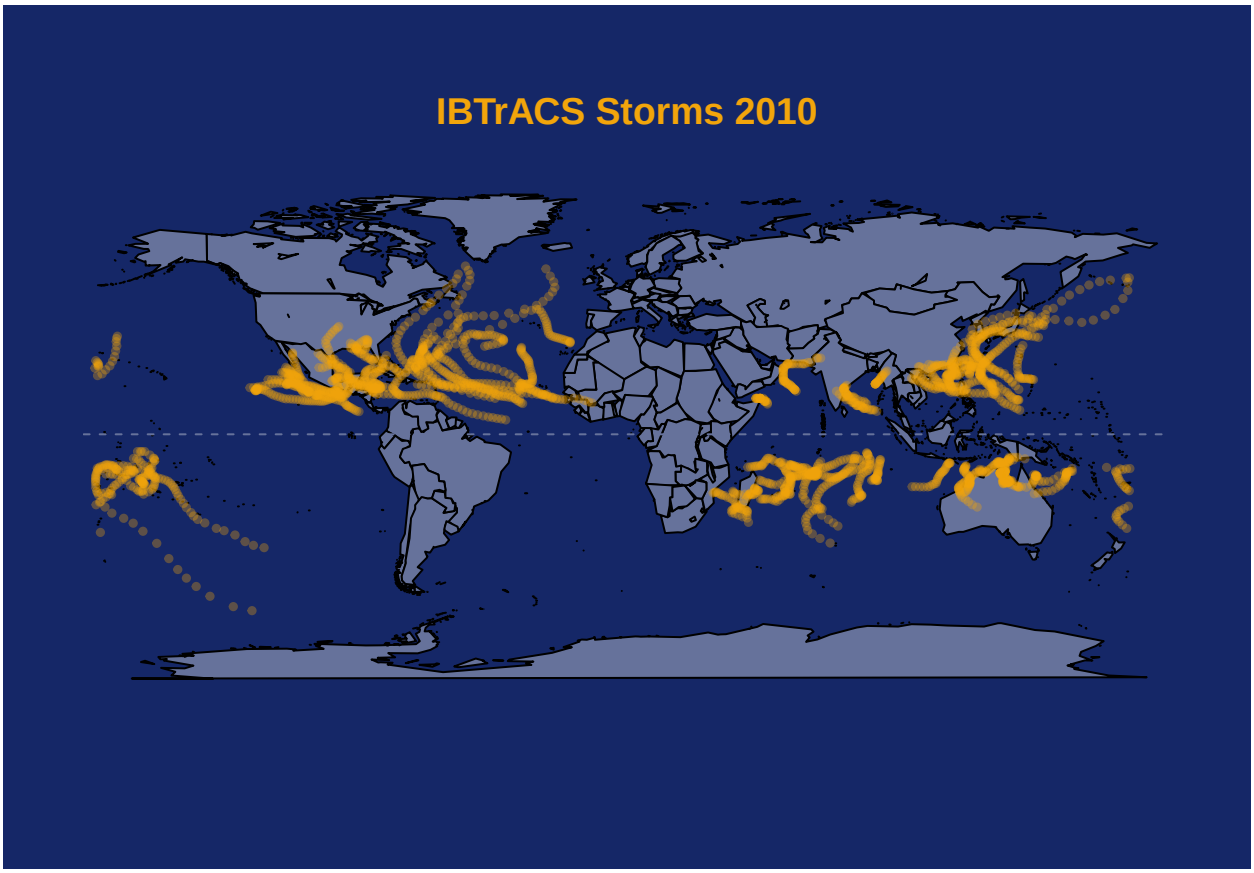
For the visualization part you'll be working with the package "maps". This package allows you to get basic maps to visualize geographic data.

You can get a simple chart to see all the storms using `map()` and plotting the storm coordinates (longitude and latitude) with `points()`:

```
library(maps)
map()
with(storms,
     points(longitude, latitude, pch = '.', col = 'red'))
```



Your mission is to obtain a nicer map by changing the fill, background, and points colors. DO NOT use color names, instead use colors in hexadecimal notation. For instance, something like the map below (notice the dotted line for the Equator). Feel free to choose your own colors, plotting symbol, symbol sizes:



REPORT

You'll have to write an `.Rmd` file, in the form of an *executive summary*. Pretend that you'll turn in this report to some journalist that will use the information based on your analysis.

Provide a short description of the assignment, the data source, the map, and the following information:

- Total number of storms (in 2010)
- Number of storms with winds $\geq$ 35 knots (tropical storms)
- Number of storms with winds $\geq$ 64 knots (hurricanes)
- Number of storms with winds $\geq$ 96 knots (major hurricanes)
- Number of storms per hemisphere (north and south)
- Frequency table with data points per month (month in words)
- Frequency table with data points per month (month in words) in northern hemisphere
- Frequency table with data points per month (month in words) in southern hemisphere
- Name of storm that lasted more days
- Name of storm with maximum wind speed (and speed value)
- Name of storm with minimum pressure (and pressure value)