

Data Analysis Project Using R

Adam Tirado

For this project I have chosen to grab weather data from a buoy at the Santa Monica Pier, the data gathered encompasses wind speeds and temperatures and directions, water temperatures, and pressure readings. The goal is to transform the data so as to gather the change in these metrics over time in the specific area and graph the result to better understand the climate. This program can be edited to gather data from other buoys that are stationed all over the world.

Data Set

<http://www.ndbc.noaa.gov/data/realtime2/ICAC1.txt>

Data Source

http://www.ndbc.noaa.gov/station_realtime.php?station=icac1

About the Data

The data is in a .txt file, and separated arbitrarily by spaces to create a visually appealing data set. Missing data is denoted by the string “MM”, and for this specific buoy there are 8 columns that have no data collected, the reason why is uncertain. The data set is updated every 6 minutes, which makes for 10 entries every hour, and therefore 240 entries in a day (useful information to keep on hand when graphing). The data set is set to the GMT time zone, which is 8 hours ahead of our PST time zone here in California. The measurements for temperature are in Celsius, speeds in meters per second, and pressure in hectopascals (hPa, a unit of barometric pressure).

The raw code for the project is included in the .txt file, and can be copied all in at once. Below is a detailed explanation of each line used.

Code that works & Explanation

Loads and saves the data-set under the name "data1", sets certain strings to not be Factors (messes up the data if done raw), and labels all data that is of the string "MM" to be an NA object in the data frame as the data does not come with actual blanks for missing data and requires being manually told what is an NA object

```
data1<-  
read.table("http://www.ndbc.noaa.gov/data/realtime2/ICAC1.txt",stringsAsFactors=FALSE,na.strings="MM")
```

Shows the "header" of the data-set, the first 6 rows of the set for each column, good for recognizing what the data looks like before manipulation.

```
head(data1)
```

Prepares the named data set of the individual columns where it is known that there is no usable data. Requires that you know exactly which columns are empty before entering the code.

```
data1$V9 <- data1$V10 <- data1$V11 <- data1$V12 <- data1$V16 <- data1$V17 <- data1$V18  
<- data1$V19 <- NULL
```

Gives basic descriptive statistics for each column in the data-set.

```
summary(data1)
```

Renames column number 1 to be "Year"

```
names(data1)[1] <- "Year"
```

Plots the 11th column, which we know as Water Temperature, as a line graph of relative line thickness level 2, with a y-range between 16.5 and 18 which are degrees Celsius, and an x-range of 240 entries. If we study the data we see that an entry is made every 6 minutes, with 10 entries per hour, making a 24-hour period equal to 240 entries, or an x-range of 240. The graph is titled "Water Temperature over Time", y-axis is titled "Water Temp.", the x-axis is labeled "Entries", and the line is colored Red. Changes can be made to see graphs of the other items.

```
plot(data1[,11], type="l", lwd=2, col="red", xlab="Entries", xlim=c(0,240), ylab="Water  
Temp.", main="Water Temperature over Time", ylim=c(16.5,18) )
```
