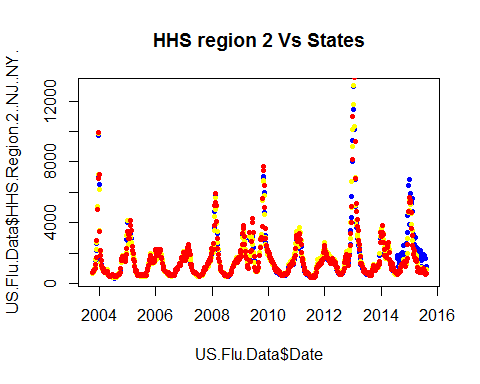
HW\_1.R

Vishesh Kakarala

Tue Feb 16 02:16:30 2016

library(XML)  
  
#Q1   
#Part 1 :  
#The Office of Intergovernmental and External Affairs hosts ten Regional Offices that directly serve state and local organizations. Each Regional Director ensures the Department maintains close contact with state, local, and tribal partners and addresses the needs of communities and individuals served through HHS programs and policies.  
  
#Part 2:  
US.Flu.Data <- read.csv("C:/Users/Vishesh Kakarala/Desktop/Foundations of data science/HW 1/US Flu Data.csv")  
US.Flu.Data$Date = as.Date(US.Flu.Data$Date, format = "%m/%d/%Y")  
  
#HHS region & states comparison  
  
# comparison between HHS region 2 and states -New york & New Jersey  
  
# In order to compare, we plot the data from the HHS region and the states data.  
plot(US.Flu.Data$Date, US.Flu.Data$HHS.Region.2..NJ..NY.,main = "HHS region 2 Vs States", type = "p", col = "blue", pch = 20)  
points(US.Flu.Data$Date, US.Flu.Data$New.York, col = "yellow", pch = 20)  
points(US.Flu.Data$Date, US.Flu.Data$New.Jersey, col = "red", pch = 20)



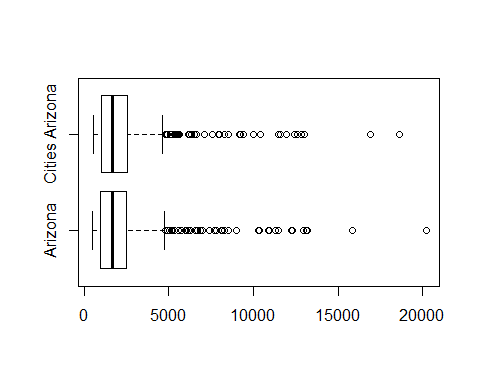
# The states data is simlar to the HHS region except in the period from 2014 to 2016 where HHS region shows higher peak value and longer period of activity when compared to the state values  
  
#part3  
  
#grouping cities as States   
  
State\_cities <- data.frame(US.Flu.Data$Date)  
colnames(State\_cities) <- "Date"  
State\_cities$Date = as.Date(State\_cities$Date, format = "%m/%d/%Y")  
  
State\_cities$Alaska <- US.Flu.Data$Anchorage..AK  
State\_cities$Alabama <- US.Flu.Data$Birmingham..AL  
State\_cities$Arkansas <- US.Flu.Data$Little.Rock..AR  
State\_cities$Arizona <- apply(US.Flu.Data[,67:71],1,function(x) mean(x,na.rm = TRUE))  
State\_cities$California <- apply(US.Flu.Data[,72:82],1,function(x) mean(x,na.rm = TRUE))  
State\_cities$Colorado <- apply(US.Flu.Data[,83:84],1,function(x) mean(x,na.rm = TRUE))  
State\_cities$Florida <- apply(US.Flu.Data[,86:90],1,function(x) mean(x,na.rm = TRUE))  
State\_cities$Georgia <- apply(US.Flu.Data[,91:92],1,function(x) mean(x,na.rm = TRUE))  
State\_cities$Hawaii <- US.Flu.Data$Honolulu..HI  
State\_cities$Iowa <- US.Flu.Data$Des.Moines..IA  
State\_cities$Idaho <- US.Flu.Data$Boise..ID  
State\_cities$Illinois <- US.Flu.Data$Chicago..IL  
State\_cities$Indiana <- US.Flu.Data$Indianapolis..IN  
State\_cities$Kansas <- US.Flu.Data$Wichita..KS  
State\_cities$Kentucky <- US.Flu.Data$Lexington..KY  
State\_cities$Louisiana <- US.Flu.Data$Baton.Rouge..LA  
  
#to understand the comparison between the cities and states data we take an example  
  
#comparing cities of Arizona with overall data from the state of Arizona  
summary(US.Flu.Data$Arizona)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 469 974 1676 2188 2477 20200

summary(State\_cities$Arizona)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 579.3 1034.0 1699.0 2211.0 2528.0 18610.0

#There are no Missing values and like to like comparison can be donw  
  
boxplot(US.Flu.Data$Arizona,State\_cities$Arizona, names = c("Arizona","Cities Arizona"),horizontal = TRUE)



#Using a box plot we can visually determine the distribution of data and central tenedencies are similar for the state and cities grouping  
  
# Comparing cities from Alaska and the data of the entire State of Alaska  
  
summary(US.Flu.Data$Alaska)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 538 908 1297 1619 1896 7384 63

summary(State\_cities$Alaska)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 547 938 1403 1683 2010 7566 55

# we find that there is missing data in both the columns.  
  
#We can populate missing data using interpolation, here Spline Interpolation is used to predict the missing values  
  
Alaska\_interpol <- splinefun(US.Flu.Data$Alaska, y = US.Flu.Data$date, method = "periodic")

## Warning in splinefun(US.Flu.Data$Alaska, y = US.Flu.Data$date, method =  
## "periodic"): spline: first and last y values differ - using y[1L] for both

US.Flu.Data$Alaska[1:63]<-Alaska\_interpol(1:63)  
  
Alaska\_cities\_interpol <- splinefun(State\_cities$Alaska, y = State\_cities$date, method = "periodic")

## Warning in splinefun(State\_cities$Alaska, y = State\_cities$date, method =  
## "periodic"): spline: first and last y values differ - using y[1L] for both

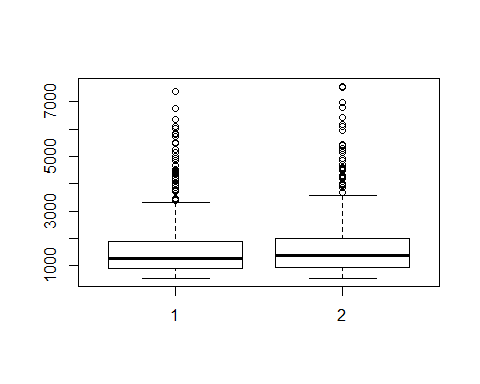
State\_cities$Alaska[1:55]<-Alaska\_interpol(1:55)  
  
# Now we can compare the two columns of data  
summary(US.Flu.Data$Alaska)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 538 884 1281 1601 1879 7384

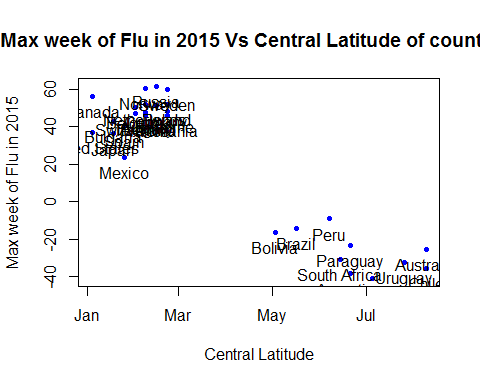
summary(State\_cities$Alaska)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 547.0 922.8 1384.0 1671.0 2010.0 7566.0

boxplot(US.Flu.Data$Alaska,State\_cities$Alaska)



# here we find the two data sets to be similar.  
  
  
#q2  
  
#https://developers.google.com/public-data/docs/canonical/countries\_csv  
  
world\_data <- read.csv("~/learning test/world\_data.csv")  
world\_data\_2015 <- world\_data[628:659,]  
flu\_max <- data.frame(apply(world\_data\_2015[,2:30], 2,function(x) which.max(x)))  
rownames(flu\_max) <- colnames(world\_data\_2015[2:30])  
colnames(flu\_max) <- "Max\_flu"  
flu\_max$Max\_flu <- world\_data\_2015[c(flu\_max$Max\_flu),1]  
flu\_max$Max\_flu = as.Date(flu\_max$Max\_flu, format = "%m/%d/%Y")  
  
world\_lat <- read.csv("~/learning test/world\_lat.csv")  
  
rownames(world\_lat) <- world\_lat$name  
  
rownames(flu\_max)[16]<- "New Zealand"  
rownames(flu\_max)[28]<- "United States"  
rownames(flu\_max)[23]<- "South Africa"  
  
flu\_max$latitude <- world\_lat[match(rownames(flu\_max),rownames(world\_lat)),]$latitude  
  
plot(flu\_max$Max\_flu,flu\_max$latitude, main = "Max week of Flu in 2015 Vs Central Latitude of countries",xlab = "Central Latitude", ylab = "Max week of Flu in 2015", type = "p", col = "blue", pch = 20)  
text(flu\_max$Max\_flu,flu\_max$latitude,labels = row.names(flu\_max),pos = 1)



# It is clear from the graph that Flu search trends are seasonal depending on the position of the countries latitude above or below the equator. With flu search trends increasing in the peak winter season in the northern hemisphere and in the peak summer months in the souther hemisphere.  
#Q3 - Web Scraping  
# part a reading vaccine status Data  
test\_table <- readHTMLTable("http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6401a4.htm?s\_cid=mm6401", header = TRUE, trim = TRUE)  
table1 <- data.frame(test\_table$`table-3`)  
table1 = table1[-(1:3),]  
table1 = table1[-38,]  
colnames(table1) <- c("charecteristics","Influenza\_positive\_no","Influenza\_positive\_%","Influenza\_negative\_no","Influenza\_negative\_%","Test\_result\_P\_value","vaccinated\_status\_no","vaccinated\_status\_Total","vaccinated\_status\_%","vaccinated\_status\_pvalue")  
rownames(table1) <- table1[,1]  
table1<-table1[,-1]  
  
# Part B - example of a similar table- table showing Drug overdose and deaths by by sex, age, race and Hispanic origin, Census region, and state —United States, 2013 and 2014  
# from CDC's morbidity and mortality report(MMWR)  
table\_2 <- readHTMLTable("http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6450a3.htm?s\_cid=mm6450a3\_w")  
table\_2 <- table\_2$`table-4`