# Assignment #1 : Data exploration

*Vishesh Kakarala*

*February 16, 2016*

## Question 1

**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. [link]http://www.hhs.gov/about/agencies/regional-offices/

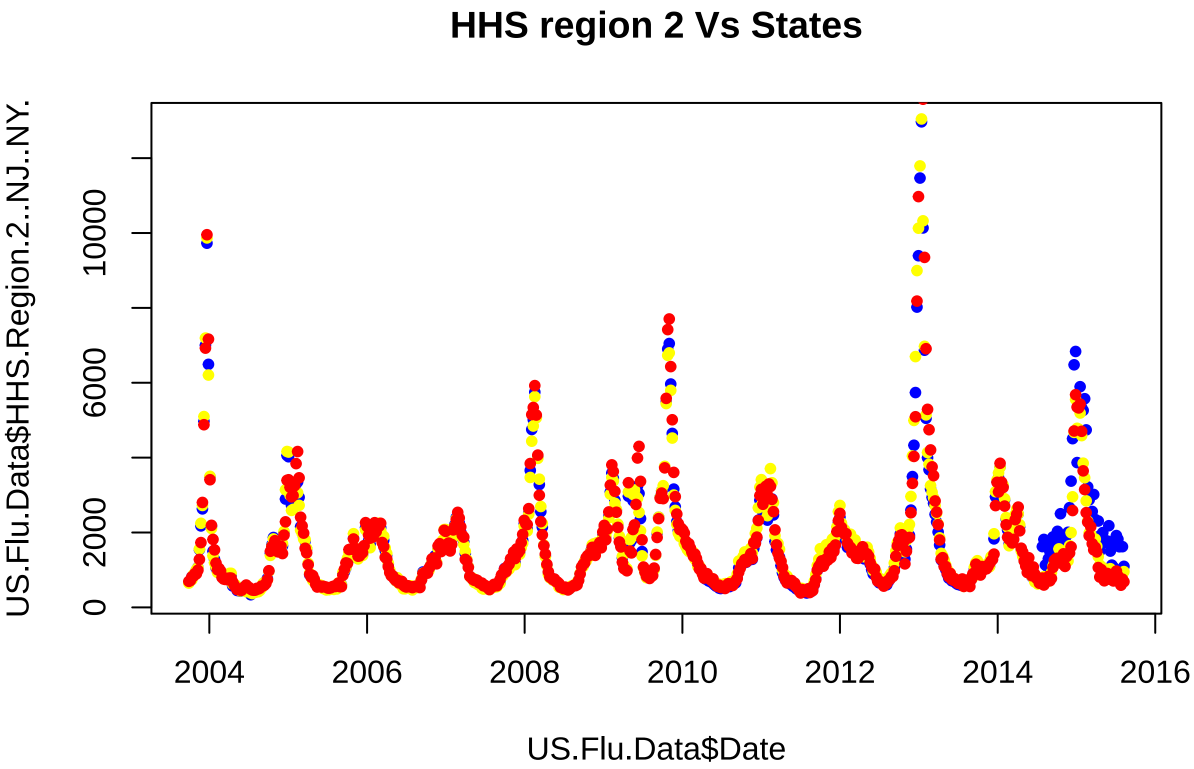
### Part 2: HHS region & states comparison

```
library(XML)
US.Flu.Data <- read.csv("C:/Users/Vishesh Kakarala/Desktop/Foundations of data science/HW 1/US Flu Data
US.Flu.Data$Date = as.Date(US.Flu.Data$Date, format = "%m/%d/%Y")
```

#### 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", c
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)
```

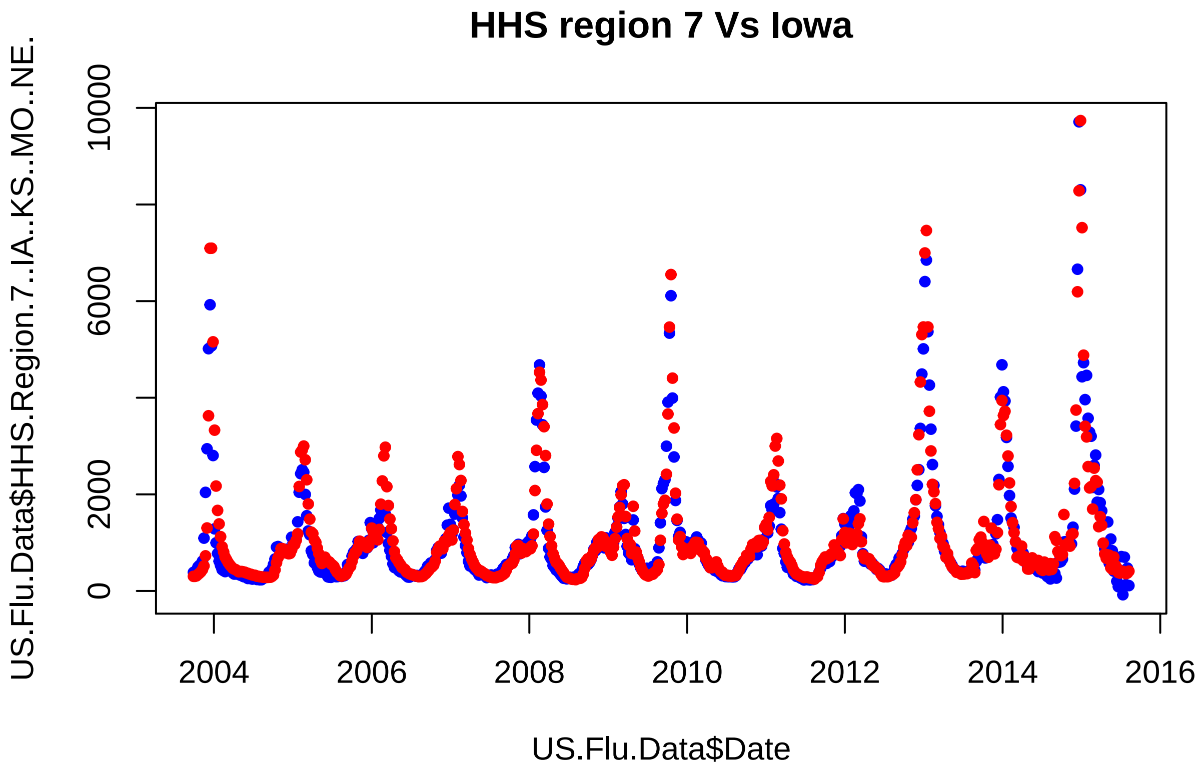


The states data is simlar to the HHS region except in the period from 2014 to 2016 where HHS region shows higher peak value and longer period of activity when compared to the state values.

A similar comparison can be done for another HHS region

**Comparison between HHS region 7 and State

```
plot(US.Flu.Data$Date, US.Flu.Data$HHS.Region.7..IA..KS..MO..NE.,main = "HHS region 7 Vs Iowa", type =
points(US.Flu.Data$Date, US.Flu.Data$Iowa, col = "red", pch = 20)
```



**Part 3: 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
```

For comparing between the cities and states data we take an example of the state of Arizona and perform the comparison

```
summary(US.Flu.Data$Arizona)
```

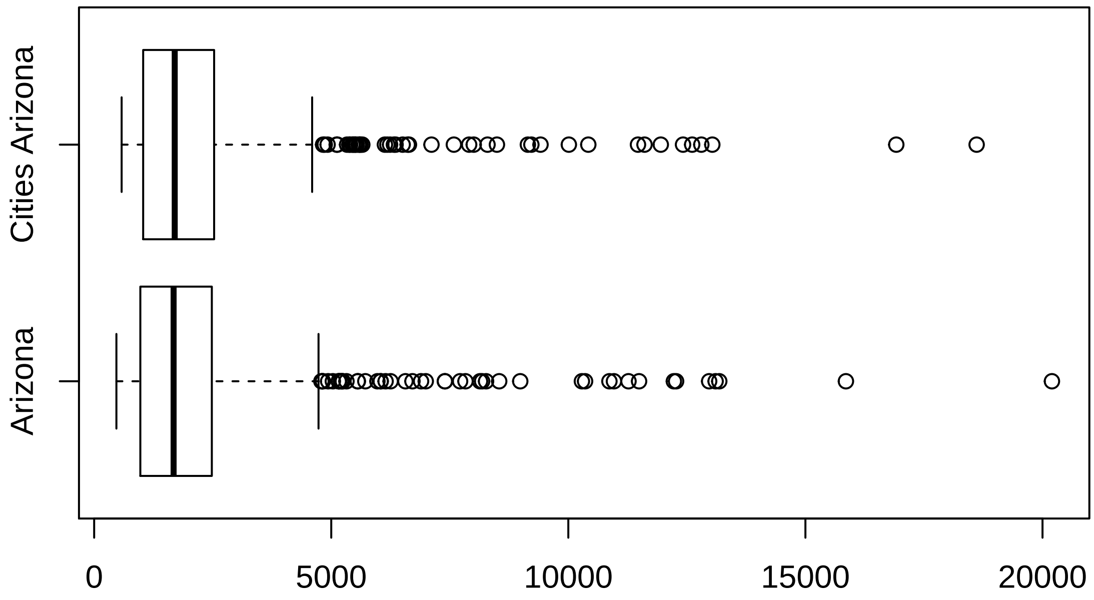
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     469     974    1676    2188    2477   20200
```

```
summary(State_cities$Arizona)
```

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

By comparing the Descriptive statistics we find that there are no missing values, and the data from the two columns have similar central tendency and distribution of data can be visualised using a boxplot

```
boxplot(US.Flu.Data$Arizona,State_cities$Arizona, names = c("Arizona","Cities Arizona"),horizontal = TRU
```



Using a box plot we can visually determine the distribution of data and median values of the two columns

When we compare data from the cities of 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")
US.Flu.Data$Alaska[1:63]<-Alaska_interpol(1:63)

Alaska_cities_interpol <- splinefun(State_cities$Alaska, y = State_cities$date, method = "periodic")
State_cities$Alaska[1:55]<-Alaska_interpol(1:55)
```

Natural Spline Interpolation is used because of the presence of one more cusps in the data,After we interpolate the missing data we can summarise the Data.

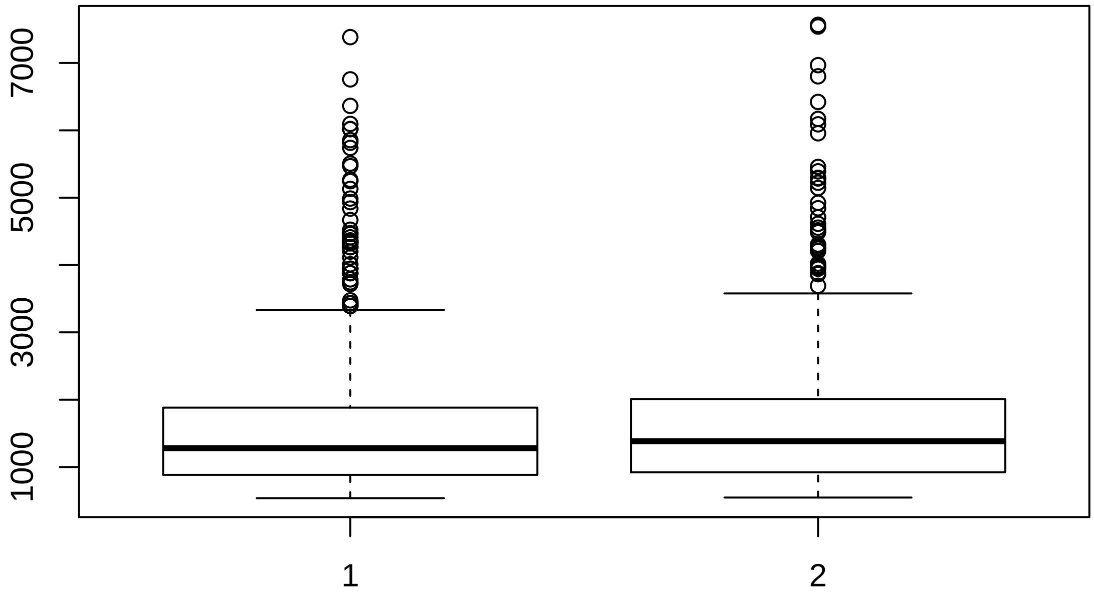
```
summary(US.Flu.Data$Alaska)
```

```
##    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)
```



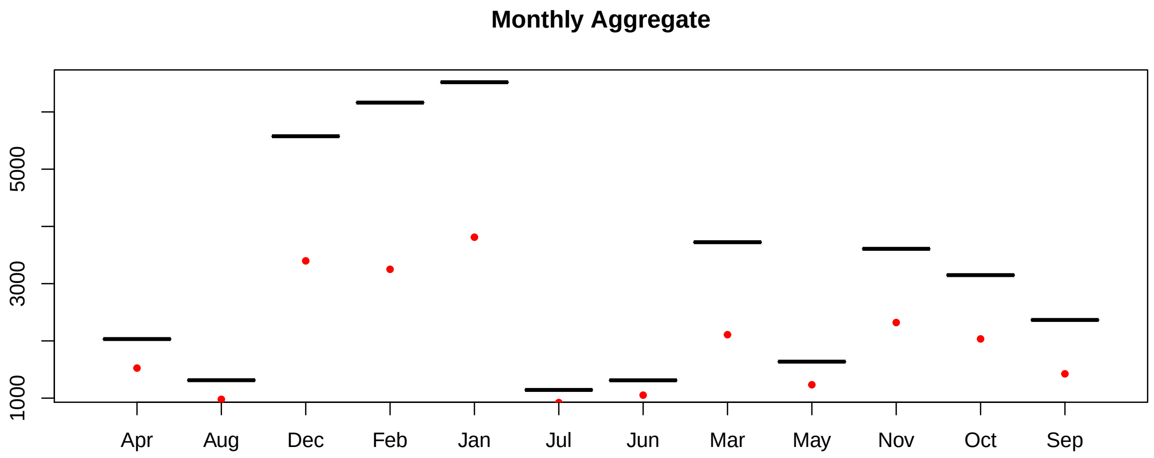
From the summary and the visual represntation we can find that the two columns of data have similar descriptive statistics.

### Part 4: Design Two relevant Metrics

First Metric : *Month Wise average Flu search activity*

Using this metric we can track which month has the highest Flu search activity from the period Dec'03 to Aug'15. This metric will aid in identfying the peak month of Flu activity and help understand seasonality of FLU trends in the geographical region. we can compare this data with the month wise heirarchial aggregates to understand Seasonal trends.

```
metric_1 <- data.frame(format(US.Flu.Data$Date, "%b"))
aggregate_metric <- aggregate(US.Flu.Data[,2:53],list(month=metric_1$format.US.Flu.Data.Date....b..),mea
frame()
plot(aggregate_metric$month, aggregate_metric$Oklahoma,main = "Monthly Aggregate",n= 1500,col = "blue",
points(aggregate_metric$month, aggregate_metric$California, col = "red", pch = 20)
```



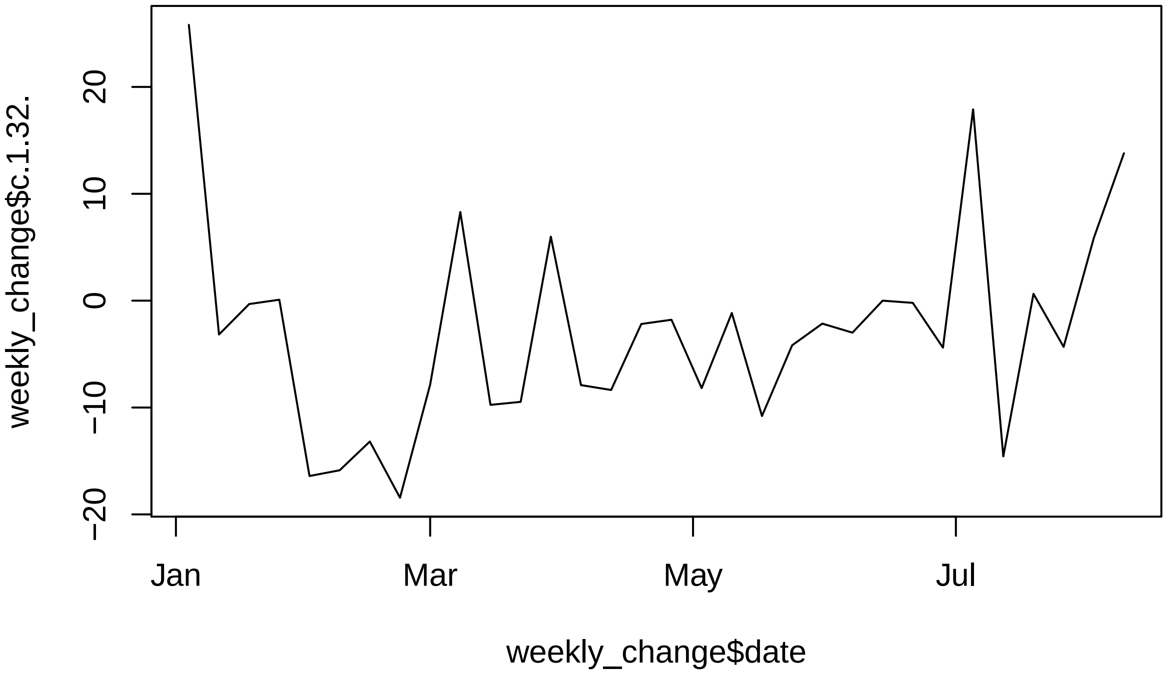
While Jan and Feb are the seasonal months in Oklahoma, it is Dec and Jan in California

Second Metric : *weekly rate of Change*

This metric aids in understanding the movement of Flu trends on a weekly or monthly basis.

Here we calculate weekly rate of change in the latest Year for the state of California

```
weekly_change <- data.frame(c(1:32))
for(i in 1:32)
{

  weekly_change[i,1]<- ((US.Flu.Data$California[i+588]- US.Flu.Data$California[i+588-1])/US.Flu.Data$Cal

}

weekly_change$date <- US.Flu.Data$Date[589:620]
plot(weekly_change$date,weekly_change$c.1.32., type = "l")
```



The plot shows the rate of change by months for the latest year in california, we can also calculate this information in a year wise distribution to understand changing trends.

**Part 5** Population data - [link]http://www.census.gov/popest/data/state/totals/2015/index.html

```
population_data <- read.csv("~/learning test/NST-EST2015-01.csv")
US.Flu.Data.2015 <- US.Flu.Data[589:620,]
flu_max_2015 <- data.frame(apply(US.Flu.Data.2015[,3:53], 2,function(x) max(x)))
flu_max_2015$population <- population_data$Population.estimates.2015
colnames(flu_max_2015) <- c("Peak_2015","Population_2015")
```

For Two Continous variables linear regression test is done to reject or fail to reject a null hypothesis

```
significance.lm <- lm(flu_max_2015$Population_2015 ~ flu_max_2015$Peak_2015, data = flu_max_2015)
summary((significance.lm))
```

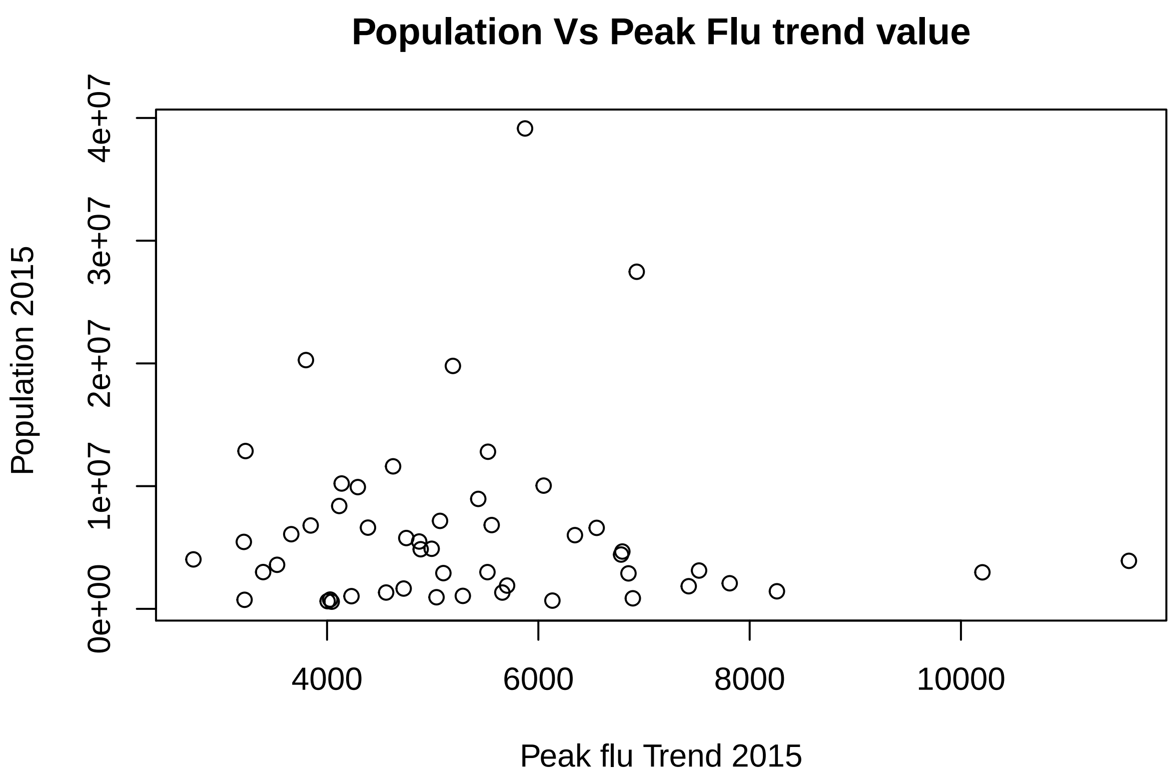
```
##
## Call:
## lm(formula = flu_max_2015$Population_2015 ~ flu_max_2015$Peak_2015,
##     data = flu_max_2015)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5948932 -4359422 -1529357   685357 32928624
##
```

```
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             7239635.1  3342415.6   2.166   0.0352 *
## flu_max_2015$Peak_2015    -174.2      591.8  -0.294   0.7697
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7268000 on 49 degrees of freedom
## Multiple R-squared:  0.001766,   Adjusted R-squared:  -0.01861
## F-statistic: 0.08667 on 1 and 49 DF,  p-value: 0.7697
```

Here since p value is 0.797 , we fail to reject the null hypothesis and the result is statistically not significant

We analyze the scatter plot

```
plot(flu_max_2015$Peak_2015,flu_max_2015$Population_2015, main = "Population Vs Peak Flu trend value",
```



From the scatter plot it is clear that higher population does not mean higher peak Flu trend value

```
cor(flu_max_2015$Peak_2015,flu_max_2015$Population_2015)
```

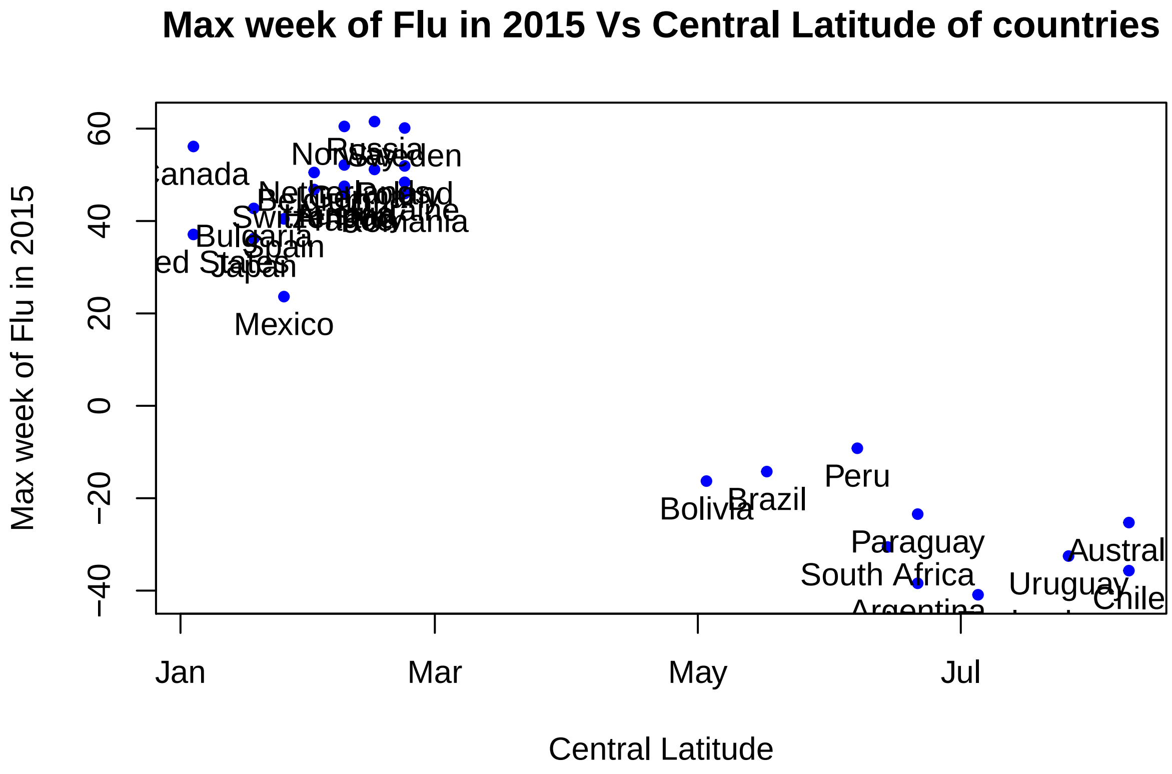
```
## [1] -0.04202087
```

The correlation coefficient is negative i.e Peak flu value does not increase with increasing population

## Question 2

Source for Country wise central latitude [link]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
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 southern hemisphere.

## Question 3

### Part 1: Reading vaccine status Data

```
test_table <- readHTMLTable("http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6401a4.htm?s_cid=mm6401", heade
table1 <- data.frame(test_table$`table-3`)
table1 = table1[-(1:3),]
table1 = table1[-38,]
colnames(table1) <- c("charecteristics","Influenza_positive_no","Influenza_positive_%","Influenza_negat
rownames(table1) <- table1[,1]
table1<-table1[,-1]
```

### Part 2: 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)-[link]http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6450a3.htm?s_cid=mm6450a3_w

```
table_2 <- readHTMLTable("http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6450a3.htm?s_cid=mm6450a3_w")
table_2 <- table_2$`table-4`
```