Name-Fenil Tailor Nid-N18730085 NetId-fst216

Assignment-1

Question 1-1)HHS Region

The Office of Intergovernmental and External Affairs hosts ten Regional Offices that directly serve state and local organizations. A President-appointed Regional Director leads each office.

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.

Reference-http://www.hhs.gov/about/agencies/regional-offices/

Question 1-2)

Reference link for US flu trends- https://www.google.org/flutrends/about/data/flu/us/data.txt

Code

For Region 10

```
US_data=read.csv(file.choose(),header=T,sep=',')
Attach(US_data)
model=lm(HHS.Region.10..AK..ID..OR..WA.~Alaska+Idaho+Oregon+Washington)
summary(model)
```

```
Call:
lm(formula = HHS.Region.10..AK..ID..OR..WA. ~ Alaska + Idaho + 
    Oregon + Washington)

Residuals:
     Min       1Q   Median       3Q      Max 
-1144.27   -63.93    21.51    94.42   509.88 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 12.86149   15.47178   0.831    0.406    
Alaska       0.01028    0.01470   0.699    0.485    
Idaho        0.09505    0.01517   6.264 7.54e-10 ***
Oregon       0.41998    0.01748  24.023  < 2e-16 ***
Washington   0.40644    0.01940  20.953  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 200.2 on 552 degrees of freedom
  (63 observations deleted due to missingness)
Multiple R-squared:  0.9804,	Adjusted R-squared:  0.9802 
F-statistic:  6894 on 4 and 552 DF,  p-value: < 2.2e-16
```

For Region 7

model=lm(HHS.Region.7..IA..KS..MO..NE.~Iowa+Kansas+Missouri+Nebraska)
summary(model)

```
Call:
lm(formula = HHS.Region.7..IA..KS..MO..NE. ~ Iowa + Kansas + 
    Missouri + Nebraska)

Residuals:
     Min       1Q   Median       3Q      Max 
-1146.57   -51.89   -10.59    34.32  1976.25 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -29.82978   12.59192  -2.369  0.01815 *  
Iowa          0.31228    0.02231  14.000  < 2e-16 ***
Kansas        0.32881    0.03349   9.818  < 2e-16 ***
Missouri      0.48102    0.02530  19.010  < 2e-16 ***
Nebraska     -0.08777    0.02516  -3.489  0.00052 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 220.7 on 609 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.9615,	Adjusted R-squared:  0.9612 
F-statistic:  3802 on 4 and 609 DF,  p-value: < 2.2e-16
```

*(**Q (1-2) in assignment_1.rmd file attached)**

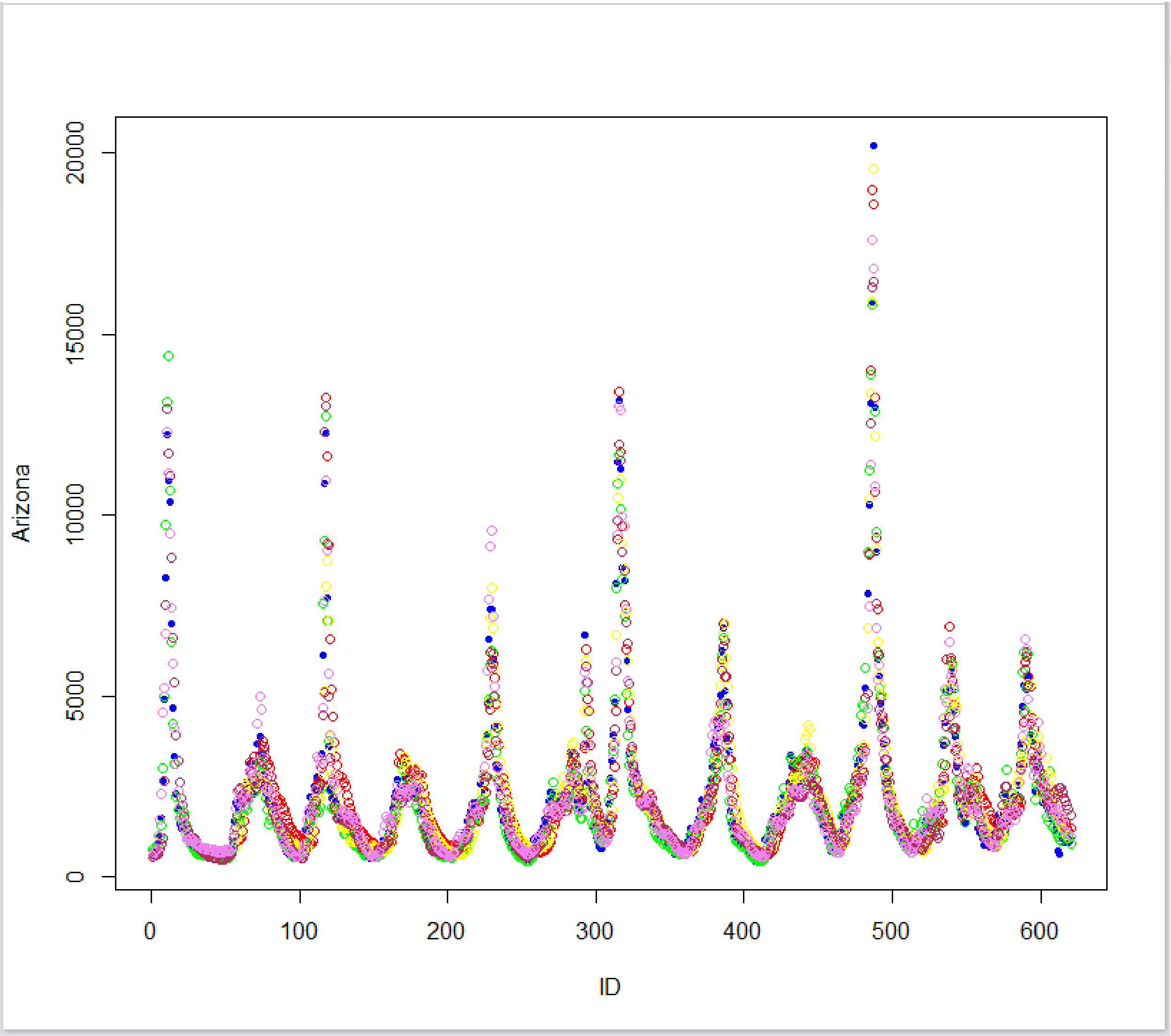
Inference

By seeing Multiple R-squared and Adjust R-squared value in the summary which is nearly 1 ,we can say that flu trends value of HHS region and the states including in that respective HHS region are strongly related.

Question 1-3)

Code

```
US_data<-read.csv(file.choose(),header=T,sep=”,”)

US_data["ID"]<-seq(from=1,to=length(Date),by=1)

plot(ID,Arizona,col="blue",pch=20)

points(ID,Mesa..AZ,col="red")

points(ID,Phoenix..AZ,col="green")

points(ID,Scottsdale..AZ,col="yellow")
points(ID,Tempe..AZ,col="maroon")

points(ID,Tucson..AZ,col="violet")

model=lm(Arizona~Mesa..AZ+Phoenix..AZ+Scottsdale..AZ+Tempe..AZ+Tucson..AZ)

summary(model)
```



```
lm(formula = Arizona ~ Mesa..AZ + Phoenix..AZ + Scottsdale..AZ + 
    Tempe..AZ + Tucson..AZ)

Residuals:
    Min      1Q  Median      3Q     Max 
-950.47 -131.39  -17.07  136.40 1727.28 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)    15.08731   17.88802   0.843  0.39935    
Mesa..AZ       -0.16575    0.02140  -7.745 4.55e-14 ***
Phoenix..AZ     0.68452    0.01873  36.555  < 2e-16 ***
Scottsdale..AZ  0.07328    0.02399   3.055  0.00236 ** 
```

```
Tempe..AZ       0.14609    0.02508   5.824 9.71e-09 ***
Tucson..AZ      0.29414    0.02173  13.534  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 271.8 on 556 degrees of freedom
  (58 observations deleted due to missingness)
Multiple R-squared:  0.9826,  Adjusted R-squared:  0.9824
F-statistic:  6278 on 5 and 556 DF,  p-value: < 2.2e-16
```

***(Q (1-3) in assignment_1.rmd file attached)**

Inference

As we can see from the graph plot and summary of the model that both the values of flu trend count in state and values in the cities in the states are strongly correlated as their multiple R-squared value and adjust R-squared value are nearly 1

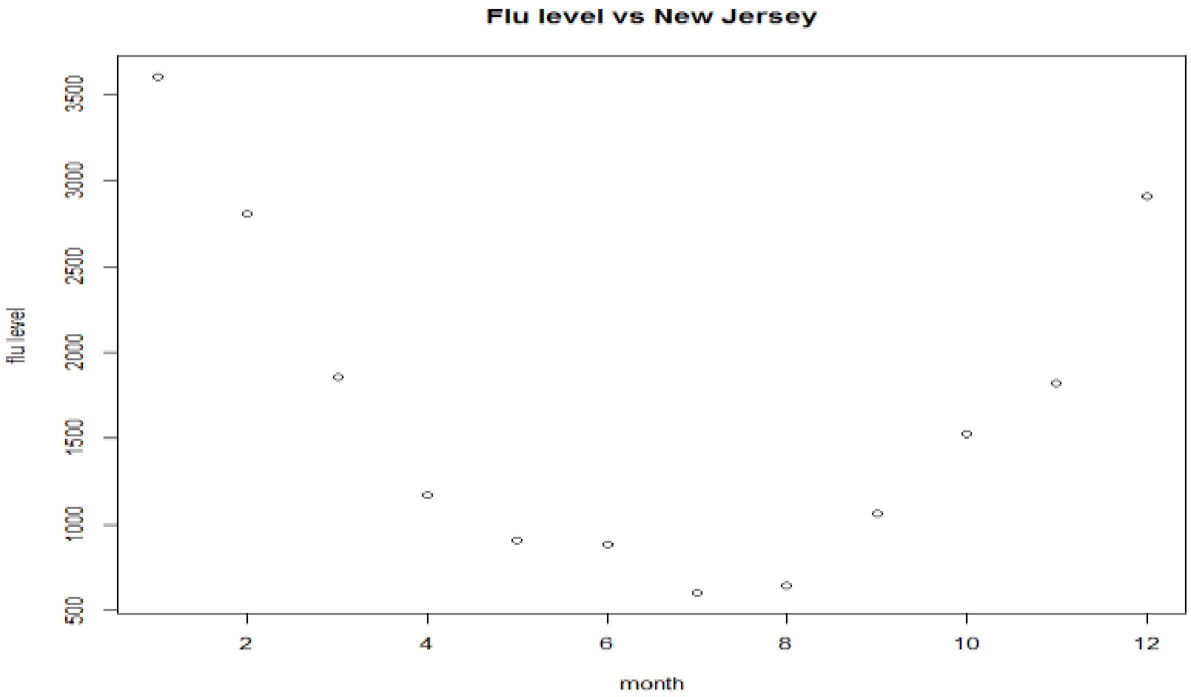
I handle the missing data by just ignoring the tuple values as it won’t Count in mean values and our result would be as accurate as possible

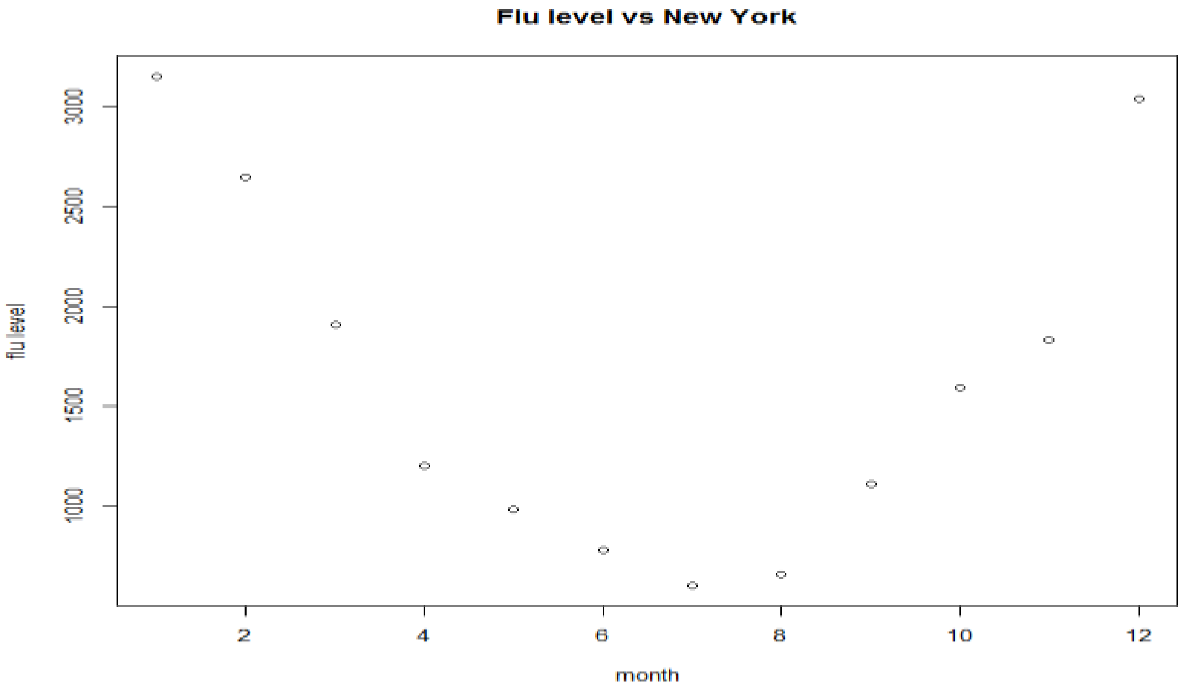
Question 1-4

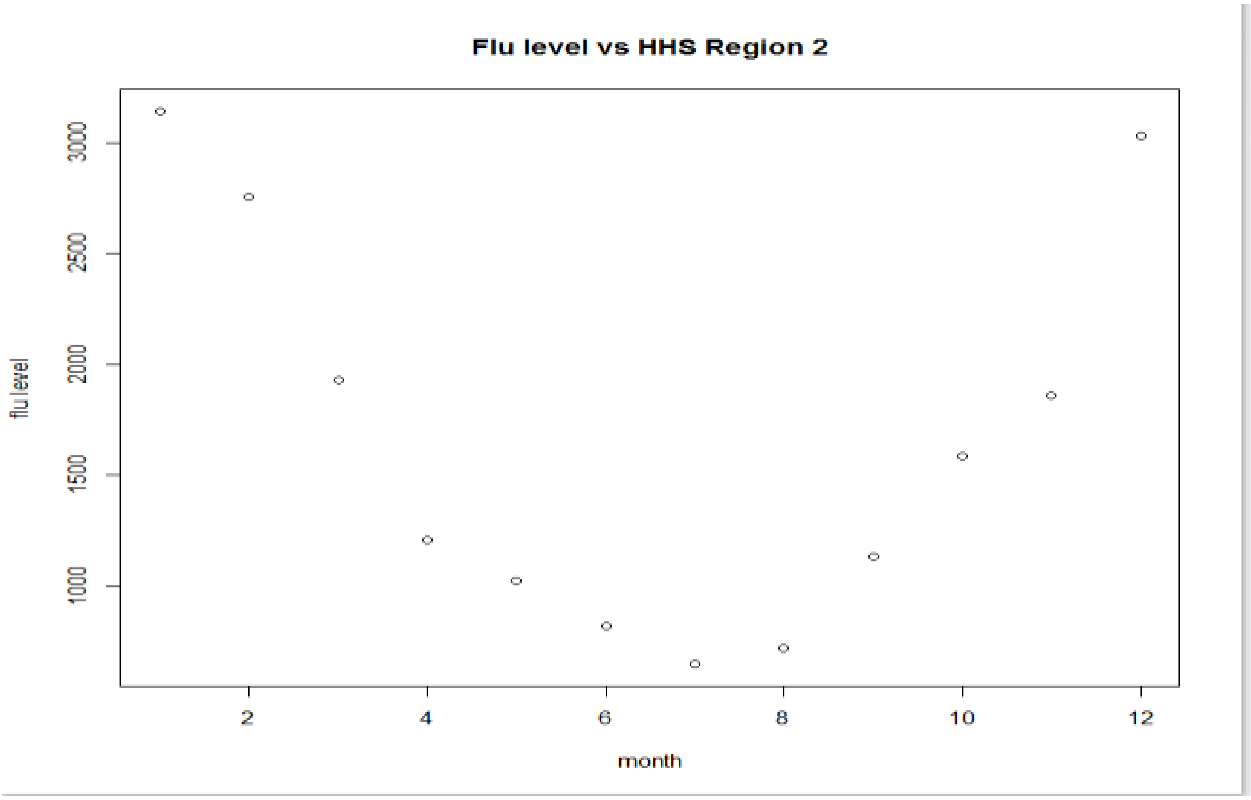
We could use mean as the most helpful metric over various geographical region over period of months for specific year. If in some month it’s crosses or reaches up to the expected threshold expected by government then in next month government can take necessary steps to control the flu counts.

Another information that we can deduce is that as we can see that HHS region and their respective states in it has approximately same median over the period of times.So,if we have missing value for one Of the states from HHS region we can approximately guess the missing value by using all other values.

Name-Fenil Tailor Nid-N18730085 NetId-fst216







Code

```
#Have changed the date format to yyyy-dd-mm in the csv file
Reference:- https://www.google.org/flutrends/about/data/flu/us/data.txt

US_data_1=read.csv(file.choose(),header=T,sep=",")
install.packages("lubridate")
library(lubridate)
period<-paste(month(US_data_1$Date))
var<-data.frame
var<-data.frame(aggregate(HHS.Region.2..NJ..NY.,list(period),mean))
plot(var$Group.1,var$x,xlab="month",ylab="flu level",main="Flu level vs HHS region 2")

var<-data.frame(aggregate(US_data_1$New.York,list(period),mean))
plot(var$Group.1,var$x,xlab="month",ylab="flu level",main="Flu level vs New York")
```

```
var<-data.frame(aggregate(US_data_1$New.York,list(period),mean))

plot(var$Group.1,var$x,xlab="month",ylab="flu level",main="Flu level vs New York")


var<-data.frame(aggregate(US_data_1$New.Jersey,list(period),mean))

plot(var$Group.1,var$x,xlab="month",ylab="flu level",main="Flu level vs Jersey")
```

Inference

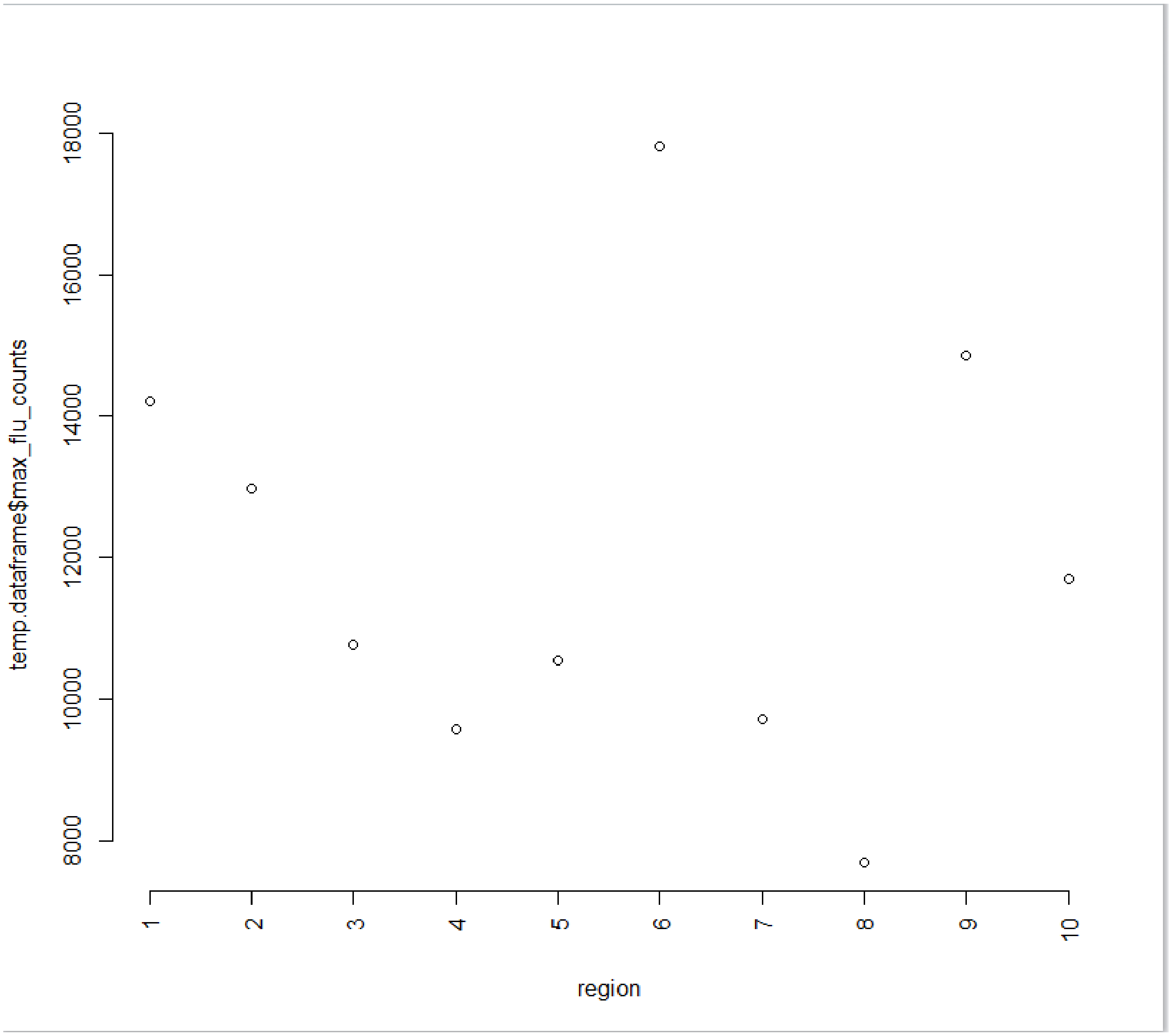
As we can see that flu levels are higher in winter season during November-February for all of the HSS Region and the respective states that are in it. By this information government can take sufficient precautions to manage the flu level.

Second matric that we can use is maximum flu count value in the entire time period of the data which is given to us. As by using this information and the conditions at that time when there was a maximum flu counts, government authorities can take necessary steps to avoid that situation if they think that conditions are similar to that time when we had maximum flu counts in the past.

Code

```
temp<-US_data_1[,54:63]

temp.dataframe<-data.frame(apply(temp,2,function(x) max(x,na.rm = TRUE)))

temp.dataframe$region<-row.names(temp.dataframe)

colnames(temp.dataframe)[1]<-"max_flu_counts"

plot(temp.dataframe$max_flu_counts,axes=FALSE,xlab="region")

axis(2)

axis(1,at=seq_along(temp.dataframe$max_flu_counts),lables=as.character(temp.dataframe$region),las=2)
```

*(**Q (1-4) in assignment_1.rmd file attached)**



Question 1-5)

Reference link for population- https://www.census.gov/popest/data/national/totals/2015/files/NST-EST2015-alldata.csv

Reference links for flu trends- https://www.google.org/flutrends/about/data/flu/us/data.txt

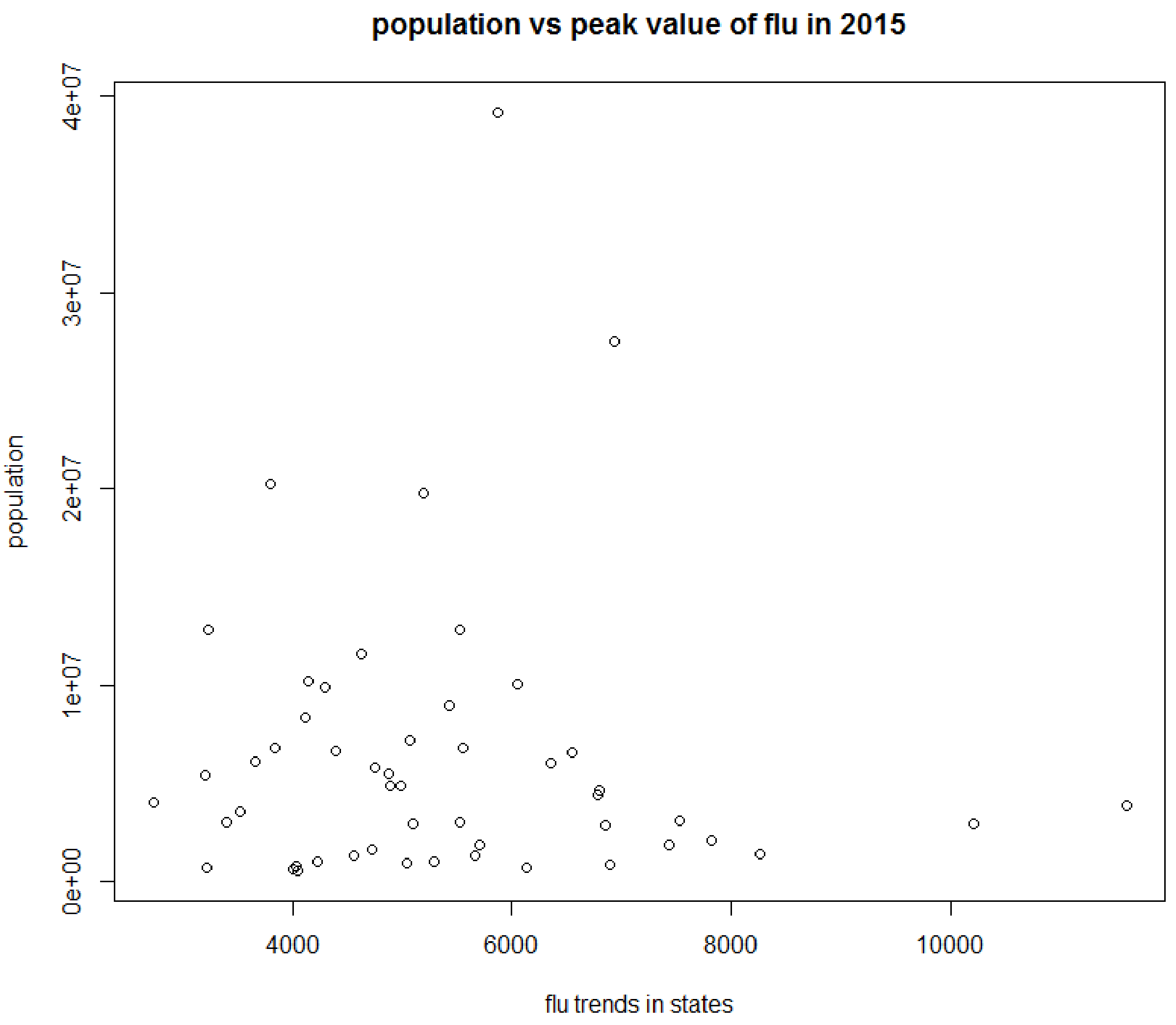
Code

```
#Read the file US_data

US_data=read.csv(file.choose(),header=T,sep=',')


us_data_state=data.frame(US_data)
```

```
us_data_state=us_data_state[c(589:620),c(3:53)]

View(us_data_state)

datastate<-apply(us_data_state,2,function(x) max(x,na.rm=TRUE))

dataset1<-data.frame(data state)

new_df <- dataset1[ order(row.names(dataset1)), ]

new_df=data.frame(new_df)


state_pop<-read.table(file.choose(), sep = ",", header = T)

state_popdf<-data.frame(state_pop)

View(state_popdf1)

state_popdf2<-state_popdf1[order(state_popdf1[,1]),]


View(state_popdf2)

datasetplot2<-sqldf("select * from state_popdf2 where NAME not in('Midwest Region','Northeast
Region','Puerto Rico','South Region','West Region','United States')")


View(datasetplot2)


plot(new_df$new_df,datasetplot2$POPESTIMATE2015,ylab="population",xlab="flu
trends",main="population vs peak value of flu in 2015")
```

---

*(**Q (1-5) in assignment_1.rmd file attached)**



Calculations

1)Taken the maximum(peak) flu trend value in year 2015 for all states

2)After importing the csv file for population from the link, we sort the data by state name and not selecting that extra state which are not their in our US_data file which is given

3)As given file data is already in sorted order of the state and we have sorted the population data by state name, now we can match and plot the graph between population of the state and peak flu value of that state.

Inference

```
model=lm(new_df$new_df~datasetplot2$POPULATION)
```

```
summary()
```

```
Call:
lm(formula = new_df$new_df ~ datasetplot2$POPULATION)
```

```
Residuals:
    Min      1Q  Median      3Q     Max 
-2667.7 -1222.8  -313.0   895.3  6186.1 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)              5.443e+03  3.276e+02  16.617   <2e-16 ***
datasetplot2$POPULATION -1.013e-05  3.442e-05  -0.294     0.77    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1753 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
```

We can see by the summary of the model that Multiple R-squared and Adjust R-squared value is around 0.1%. So we can conclude that there is not strong relation between population and max flu trends in the states.

Question 2:

Reference link for world data - https://www.google.org/flutrends/about/data/flu/data.txt

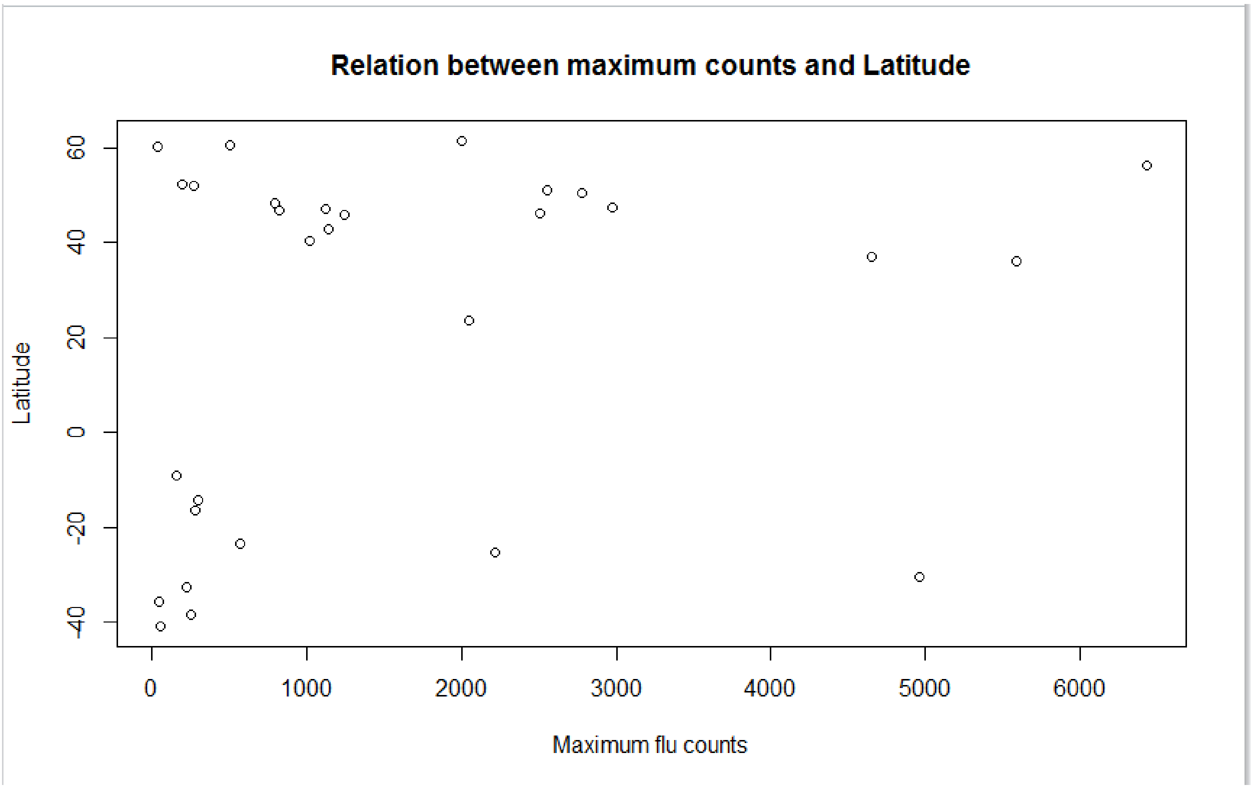
Longitude Ref link is- http://dev.maxmind.com/geoip/legacy/codes/country_latlon/

Code

```
#Read the file from World_Data

World_data1<-read.table(file.choose(),header = T,skip = 627,sep=",")


#Read the file from World_Data

header=read.table(file.choose(),nrows=1,header=F,sep=",")


colnames(World_data1)<-unlist(header)

apply1<-apply(World_data1,2,function(x) max(x,na.rm=TRUE))

df1=data.frame(apply1)

df1=data.frame(df1[-1,])

library(sqldf)

#Read data from statepop.csv file which attached

countrieslglt=read.csv.sql(file.choose(),'select Latitude from file where Name
in(\"Argentina\",\"Australia\",\"Austria\",\"Belgium\",\"Bolivia\",\"Brazil\",\"Bulgaria\",\"Canada\",\"Ch
ile\",\"France\",\"Germany\",\"Hungary\",\"Japan\",\"Mexico\",\"Netherlands\",\"New
Zealand\",\"Norway\",\"Paraguay\",\"Peru\",\"Poland\",\"Romania\",\"Russia\",\"South
Africa\",\"Spain\",\"Sweden\",\"Switzerland\",\"Ukraine\",\"United States\",\"Uruguay\")')

df1$df1..1...<-as.numeric(as.character(df1$df1..1...))

View(df1)

countrieslglt$Latitude<-as.double(as.numeric(countrieslglt$Latitude))

View(countrieslglt)

plot(df1$df1..1...,countrieslglt$Latitude,xlab="Maximum flu counts",ylab="Latitude",main="Relation bet
ween maximum counts and Latitude")
```

*(**Q (2) in assignment_1.rmd file attached)**

Output



Inference:

We can say that there are more and higher peak flu counts values for the countries which have latitude between 40 and 60 than the same values for the countries which has latitude between -40 to 0.

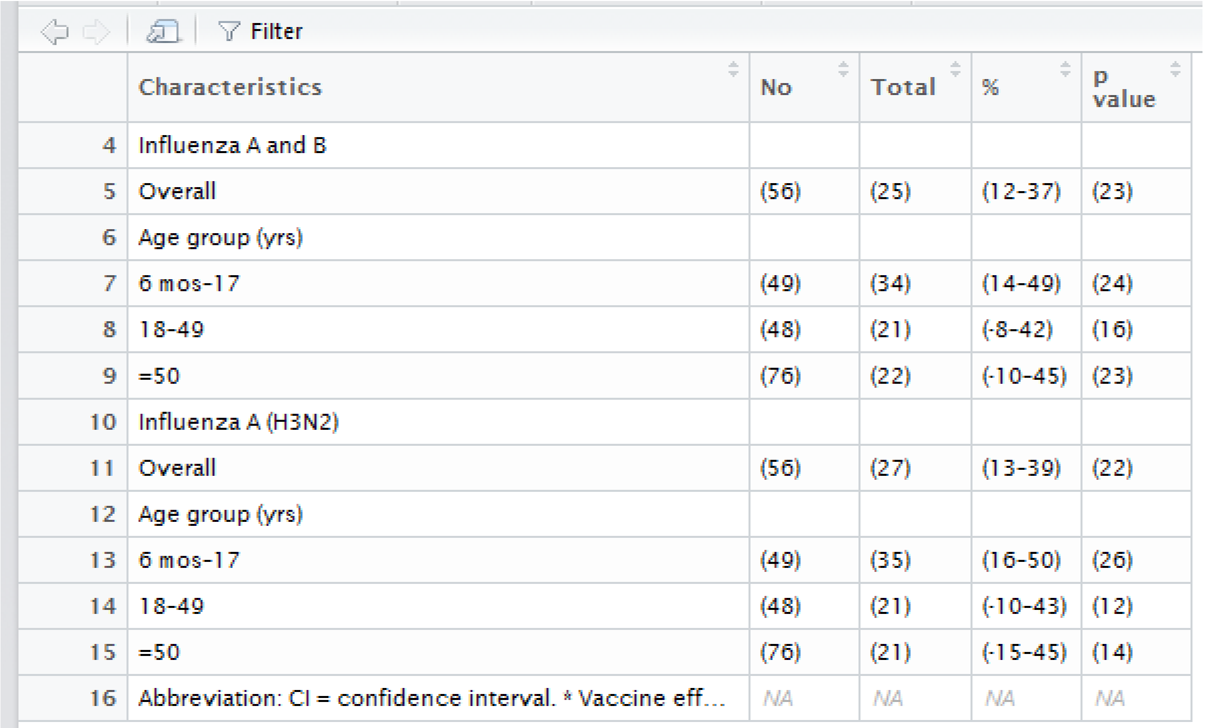
Question 3:

A) Read Vaccination status data from the second table in the webpage. Here, I’ve read sub columns named (No.,Total,%) under column name Vaccinated and (p value) column

Code

```
url="http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6401a4.htm?s_cid=mm6401a4_w"

library(XML)

url.table = readHTMLTable(url, header=T, which=1,stringsAsFactors=F)

View(url.table)

Vaccination_status<-data.frame(url.table$V1)

Vaccination_status[,"No"]<-data.frame(url.table$V7)

Vaccination_status[,"Total"]<-data.frame(url.table$V8)

Vaccination_status[,"%"]<-data.frame(url.table$V9)

Vaccination_status[,"p value"]<-data.frame(url.table$V10)

Vaccination_status<-Vaccination_status[4:length(Vaccination_status$No),]

colnames(Vaccination_status)[1]<-"Characteristics"

View(Vaccination_status)
```

*(**Q (3-1) in assignment_1.rmd file attached)**



	Characteristics	No	Total	%	p value
4	Influenza A and B				
5	Overall	(56)	(25)	(12-37)	(23)
6	Age group (yrs)				
7	6 mos-17	(49)	(34)	(14-49)	(24)
8	18-49	(48)	(21)	(-8-42)	(16)
9	=50	(76)	(22)	(-10-45)	(23)
10	Influenza A (H3N2)				
11	Overall	(56)	(27)	(13-39)	(22)
12	Age group (yrs)				
13	6 mos-17	(49)	(35)	(16-50)	(26)
14	18-49	(48)	(21)	(-10-43)	(12)
15	=50	(76)	(21)	(-15-45)	(14)
16	Abbreviation: CI = confidence interval. * Vaccine eff...	NA	NA	NA	NA

B) Found the table in below url

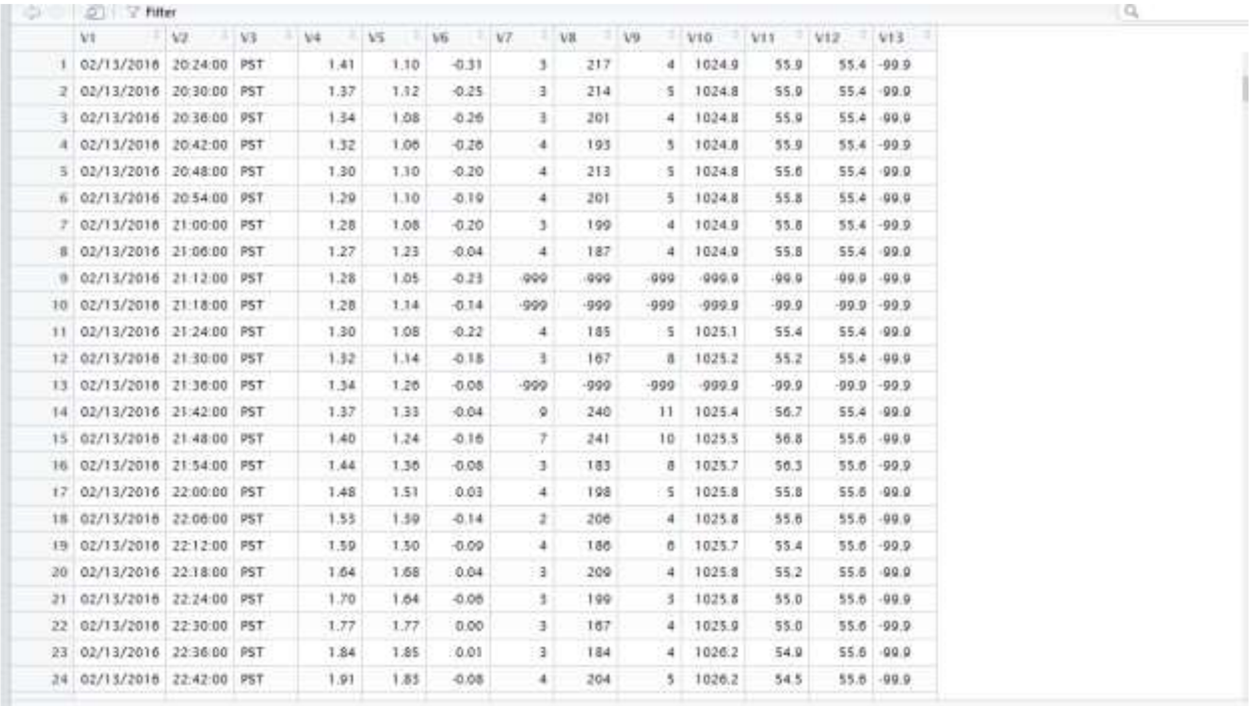
url=" http://tidesonline.nos.noaa.gov/data_read.shtml?station_info=9414290+San+Francisco,+CA"

Code

```
u = "http://tidesonline.nos.noaa.gov/data_read.shtml?station_info=9414290+San+Francisco,+CA"

h = htmlParse(u)

p = getNodeSet(h, "//pre")

con = textConnection(xmlValue(p[[2]]))

tides = read.table(con)

View(tides)
```

***(Q (3-2) in assignment_1.rmd file attached)**

Name-Fenil Tailor Nid-N18730085 NetId-fst216



	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13
1	02/13/2016	20:24:00	PST	1.41	1.10	-0.31	3	217	4	1024.9	55.9	55.4	-99.9
2	02/13/2016	20:30:00	PST	1.37	1.12	-0.25	3	214	5	1024.8	55.9	55.4	-99.9
3	02/13/2016	20:36:00	PST	1.34	1.08	-0.26	3	201	4	1024.8	55.9	55.4	-99.9
4	02/13/2016	20:42:00	PST	1.32	1.06	-0.26	4	193	5	1024.8	55.9	55.4	-99.9
5	02/13/2016	20:48:00	PST	1.30	1.10	-0.20	4	213	5	1024.8	55.6	55.4	-99.9
6	02/13/2016	20:54:00	PST	1.29	1.10	-0.19	4	201	5	1024.8	55.8	55.4	-99.9
7	02/13/2016	21:00:00	PST	1.28	1.08	-0.20	3	199	4	1024.9	55.8	55.4	-99.9
8	02/13/2016	21:06:00	PST	1.27	1.23	-0.04	4	187	4	1024.9	55.8	55.4	-99.9
9	02/13/2016	21:12:00	PST	1.28	1.05	-0.23	-999	-999	-999	-999.9	-99.9	-99.9	-99.9
10	02/13/2016	21:18:00	PST	1.28	1.14	-0.14	-999	-999	-999	-999.9	-99.9	-99.9	-99.9
11	02/13/2016	21:24:00	PST	1.30	1.08	-0.22	4	185	5	1025.1	55.4	55.4	-99.9
12	02/13/2016	21:30:00	PST	1.32	1.14	-0.18	3	167	8	1025.2	55.2	55.4	-99.9
13	02/13/2016	21:36:00	PST	1.34	1.26	-0.08	-999	-999	-999	-999.9	-99.9	-99.9	-99.9
14	02/13/2016	21:42:00	PST	1.37	1.33	-0.04	9	240	11	1025.4	56.7	55.4	-99.9
15	02/13/2016	21:48:00	PST	1.40	1.24	-0.16	7	241	10	1025.5	56.8	55.6	-99.9
16	02/13/2016	21:54:00	PST	1.44	1.36	-0.08	3	183	8	1025.7	56.3	55.6	-99.9
17	02/13/2016	22:00:00	PST	1.48	1.51	0.03	4	198	5	1025.8	55.8	55.6	-99.9
18	02/13/2016	22:06:00	PST	1.53	1.39	-0.14	2	206	4	1025.8	55.6	55.6	-99.9
19	02/13/2016	22:12:00	PST	1.59	1.50	-0.09	4	186	6	1025.7	55.4	55.6	-99.9
20	02/13/2016	22:18:00	PST	1.64	1.68	0.04	3	209	4	1025.8	55.2	55.6	-99.9
21	02/13/2016	22:24:00	PST	1.70	1.64	-0.06	3	199	3	1025.8	55.0	55.6	-99.9
22	02/13/2016	22:30:00	PST	1.77	1.77	0.00	3	167	4	1025.9	55.0	55.6	-99.9
23	02/13/2016	22:36:00	PST	1.84	1.85	0.01	3	184	4	1026.2	54.9	55.6	-99.9
24	02/13/2016	22:42:00	PST	1.91	1.83	-0.08	4	204	5	1026.2	54.5	55.6	-99.9

****Some of the Questions were discussed with Kunal Relia. Only the ideas are discussed followed by individual applications in doing codes.**