

Question 1 (Data Exploration)

a)

HHS stands for Department of Health and Human Services. There are 10 different regions that make up the US HHS. Each region consists of various states. For eg, States of New York and New Jersey make up HHS region 2. Each region is responsible to perform various duties ranging from Health Insurance to Prevention and Cure of diseases.

b)

Regression test has been used to analyze and compare the region and state values for each region.

Region2 Analysis:

```
> model<-lm(HHS.Region.2..NJ..NY.~New.York+New.Jersey)
> summary(model)
```

Call:

```
lm(formula = HHS.Region.2..NJ..NY. ~ New.York + New.Jersey)
```

Residuals:

| Min | 1Q | Median | 3Q | Max |
|---------|---------|--------|--------|---------|
| -908.18 | -117.94 | -45.84 | -12.10 | 2138.06 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|----------|------------|---------|-------------|
| (Intercept) | 62.38849 | 19.48924 | 3.201 | 0.00144 ** |
| New.York | 0.92488 | 0.03341 | 27.685 | < 2e-16 *** |
| New.Jersey | 0.05422 | 0.02774 | 1.954 | 0.05111 . |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 301 on 617 degrees of freedom
Multiple R-squared: 0.9571, Adjusted R-squared: 0.957
F-statistic: 6890 on 2 and 617 DF, p-value: < 2.2e-16

Region10 Analysis:

```
> 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
```

Region6 Analysis:

```
> model<-lm(HHS.Region.6..AR..LA..NM..OK..TX.~Arkansas+Louisiana+New.Mexico+O  
klahoma+Texas)  
> summary(model)
```

Call:

```
lm(formula = HHS.Region.6..AR..LA..NM..OK..TX. ~ Arkansas + Louisiana +  
    New.Mexico + Oklahoma + Texas)
```

Residuals:

| Min | 1Q | Median | 3Q | Max |
|----------|--------|--------|-------|---------|
| -1897.25 | -97.88 | -14.34 | 86.92 | 3056.86 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|------------|------------|---------|--------------|
| (Intercept) | -20.332632 | 26.870907 | -0.757 | 0.44956 |
| Arkansas | 0.055739 | 0.019471 | 2.863 | 0.00436 ** |
| Louisiana | 0.002419 | 0.015865 | 0.152 | 0.87888 |
| New.Mexico | 0.084057 | 0.020822 | 4.037 | 6.17e-05 *** |
| Oklahoma | 0.092971 | 0.019789 | 4.698 | 3.31e-06 *** |
| Texas | 0.761098 | 0.022111 | 34.421 | < 2e-16 *** |

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 319.1 on 556 degrees of freedom  
(58 observations deleted due to missingness)
```

```
Multiple R-squared:  0.982,    Adjusted R-squared:  0.9818
```

```
F-statistic:  6062 on 5 and 556 DF,  p-value: < 2.2e-16
```

Conclusion:

Based on the Multiple R-Squared and Adjusted R-Squared values, it can be concluded that there exists a very strong relation between the HHS region values and the corresponding state values. This can be said as both these values are nearly 1 which means that the region's value and the corresponding states' value is strongly related.

c)

Again regression can be used to study how much related 2 different attributes are.

```
> model<-lm(California~Berkeley..CA+Fresno..CA+Irvine..CA+Los.Angeles..CA+Oak  
land..CA+Sacramento..CA+San.Diego..CA+San.Francisco..CA+San.Jose..CA+Santa.Cl  
ara..CA+Sunnyvale..CA)  
> summary(model)
```

```
Call:  
lm(formula = California ~ Berkeley..CA + Fresno..CA + Irvine..CA +  
  Los.Angeles..CA + Oakland..CA + Sacramento..CA + San.Diego..CA +  
  San.Francisco..CA + San.Jose..CA + Santa.Clara..CA + Sunnyvale..CA)
```

```
Residuals:  
      Min       1Q   Median       3Q      Max  
-1779.59   -94.16   -20.33    89.81  1021.17
```

```
Coefficients:  
              Estimate Std. Error t value Pr(>|t|)  
(Intercept)   88.01344    16.57126   5.311 1.58e-07 ***  
Berkeley..CA  -0.10826     0.01696  -6.382 3.70e-10 ***  
Fresno..CA     0.03236     0.01594   2.030 0.042873 *  
Irvine..CA     0.10293     0.02112   4.874 1.43e-06 ***  
Los.Angeles..CA 0.35338     0.02530  13.965 < 2e-16 ***  
Oakland..CA    -0.05853     0.01162  -5.035 6.46e-07 ***  
Sacramento..CA  0.15235     0.02107   7.231 1.60e-12 ***  
San.Diego..CA   0.24279     0.02368  10.255 < 2e-16 ***  
San.Francisco..CA 0.11032     0.01548   7.127 3.21e-12 ***  
San.Jose..CA    0.05345     0.01329   4.021 6.59e-05 ***  
Santa.Clara..CA  0.03711     0.01413   2.625 0.008892 **  
Sunnyvale..CA   0.05660     0.01508   3.753 0.000193 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 197.9 on 555 degrees of freedom  
(53 observations deleted due to missingness)  
Multiple R-squared:  0.9836,    Adjusted R-squared:  0.9833  
F-statistic: 3032 on 11 and 555 DF,  p-value: < 2.2e-16
```

Hence, based on the Multiple R-Squared and Adjusted R-Squared values, it is clearly observable that the state and city Flu values are strongly co-related.

Another example:

```
> model<-lm(Arizona~Mesa..AZ+Phoenix..AZ+Scottsdale..AZ+Tempe..AZ+Tucson..AZ)  
> summary(model)
```

```
Call:  
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:

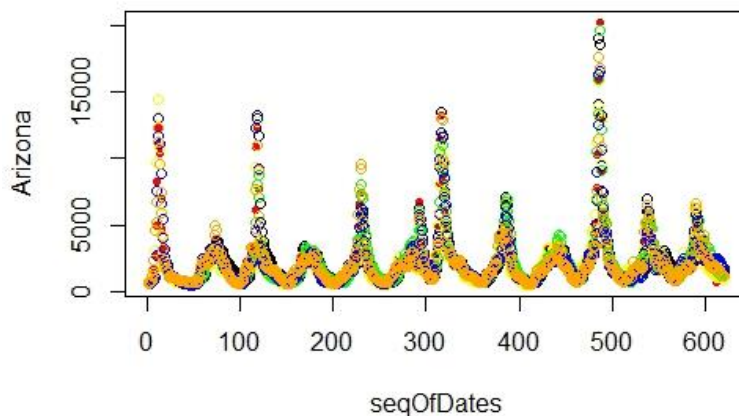
| | Estimate | Std. Error | t value | Pr(> t) |
|----------------|----------|------------|---------|--------------|
| (Intercept) | 15.08731 | 17.88802 | 0.843 | 0.39935 |
| Mesa..AZ | -0.16575 | 0.02140 | -7.745 | 4.55e-14 *** |
| Phoenix..AZ | 0.68452 | 0.01873 | 36.555 | < 2e-16 *** |
| Scottsdale..AZ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

Graphical proof:

```
> plot(seqOfDates,Arizona,col="red",pch=20)  
> points(seqOfDates,Mesa..AZ,col="black")  
> points(seqOfDates,Phoenix..AZ,col="yellow")  
> points(seqOfDates,Scottsdale..AZ,col="green")  
> points(seqOfDates,Tempe..AZ,col="blue")  
> points(seqOfDates,Tucson..AZ,col="orange")
```



Conclusion:

Based on the above discussion where near to 1 R-Squared value signify the strong co-relation between the values of city and state, the graph re-iterates the stand taken. From the graph, it is clearly visible that the peak and the bottom of the city and state values happen together. Hence, it can be concluded that the state values would have definitely been derived from the city values grouped together.

Note on missing values: The missing values have not been touched and simply neglected. This was done for a couple of reasons: 1) As compared to the enormous data set, a few missing values would not have made a difference especially when we were considering various

different cities. 2) There was no good way to approximate the value of missing places. For example, one method of approximating the value is by using the values of 2 weeks on either sides of the missing value. But as the data for entire months and year were missing, this approach would have not have been feasible or possible. And simply approximating from the state value would have had no impact on the comparison analysis that we aimed to do. Therefore, the best approach was to simply neglect the missing values and use the available data.

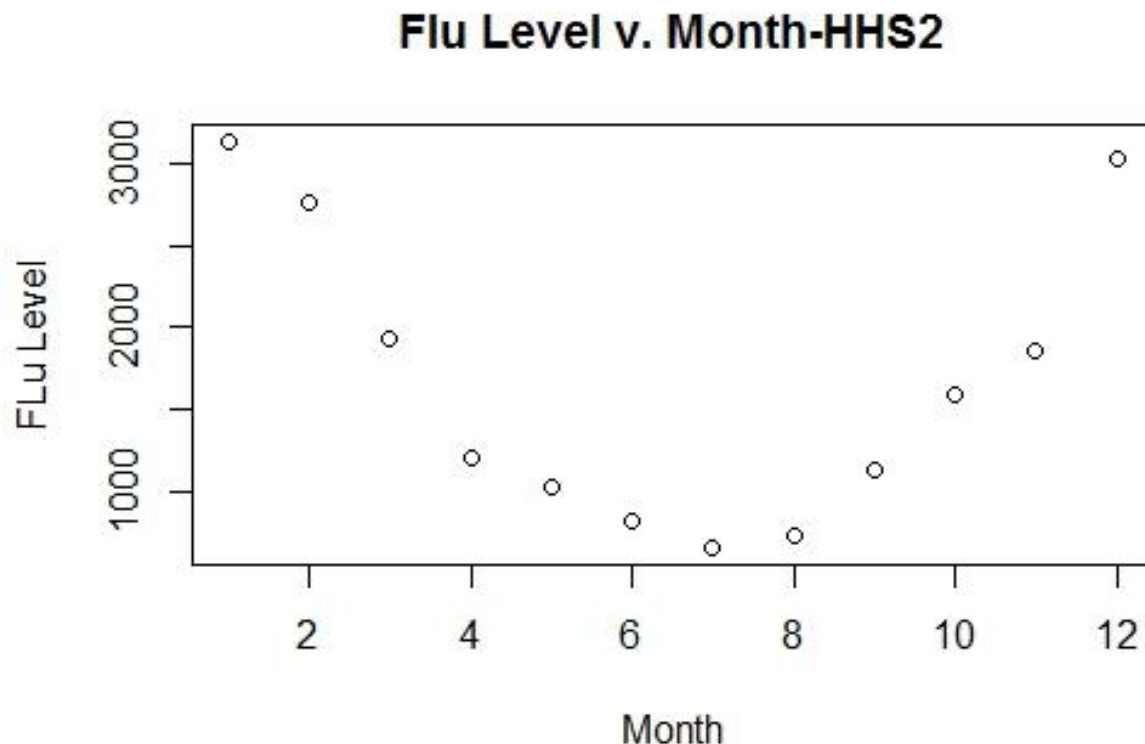
d)

Metric 1 - Mean:

The most appropriate metric that can be used here is finding the monthly mean of HHS regions and subsequently its states in order to track the Flu outbreak over the period of years. What mean will help us do is that it will help us analyze the months when the outbreak of Flu is maximum in certain regions. Hence, this information can help the authorities predict the months when Flu is likely to hit and help them take precautionary measures and also alert the citizens to be cautious.

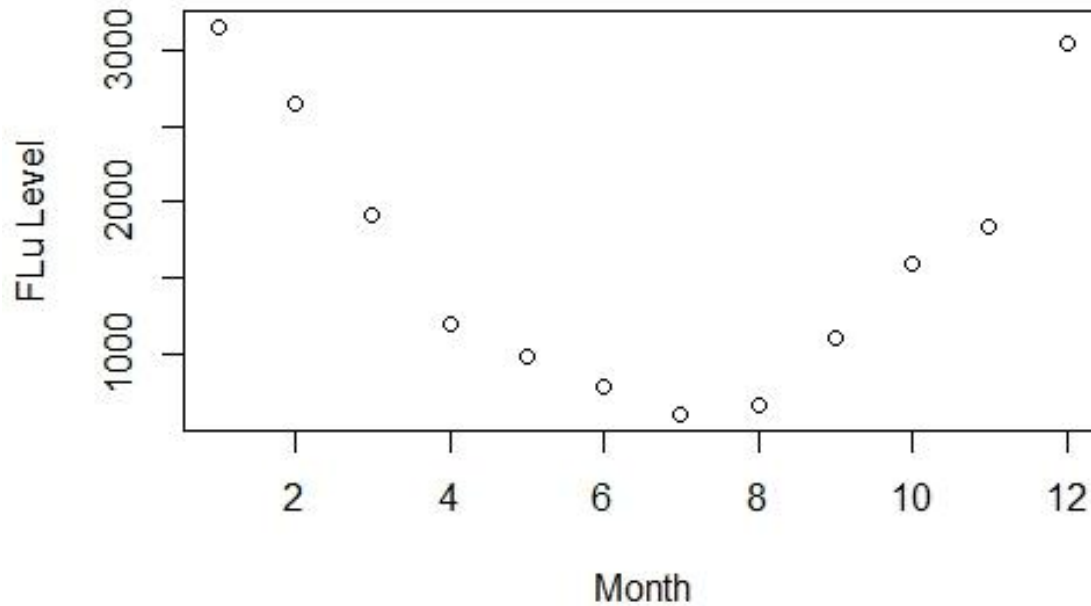
Code

```
> install.packages("lubridate")
> library(lubridate)
> period<-paste(month(data$Date))
> temp<-data.frame(aggregate(HHS.Region.2..NJ..NY.,list(period),mean))
> plot(temp$Group.1,temp$x,xlab="Month",ylab="FLu Level",main="Flu Level v. Month-HHS2")
```



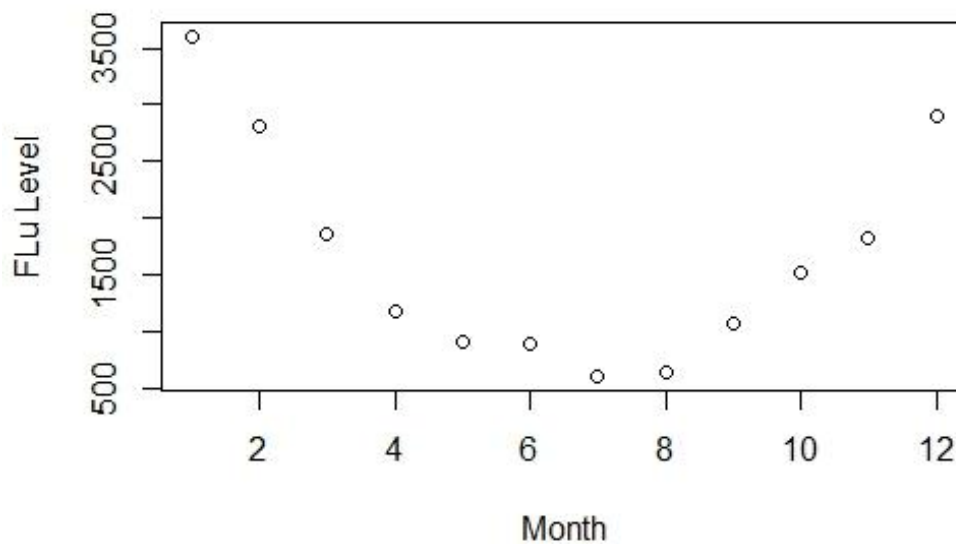
```
> temp<-data.frame(aggregate(New.York,list(period),mean))
> plot(temp$Group.1,temp$x,xlab="Month",ylab="FLu Level",main="Flu Level v. Month-New York")
```

Flu Level v. Month-New York



```
> temp<-data.frame(aggregate(New.Jersey,list(period),mean))  
> plot(temp$Group.1,temp$x,xlab="Month",ylab="FLU Level",main="Flu Level v. M  
onth-New Jersey")
```

Flu Level v. Month-New Jersey



Conclusion for choosing Mean

Hence from the above sample, we can clearly infer that over the past 12 years, the winter months, especially December and January, have been the worst months for HHS region2 as far as Flu trends are concerned. The 12-year average for winter is extremely high as seen in the graph.

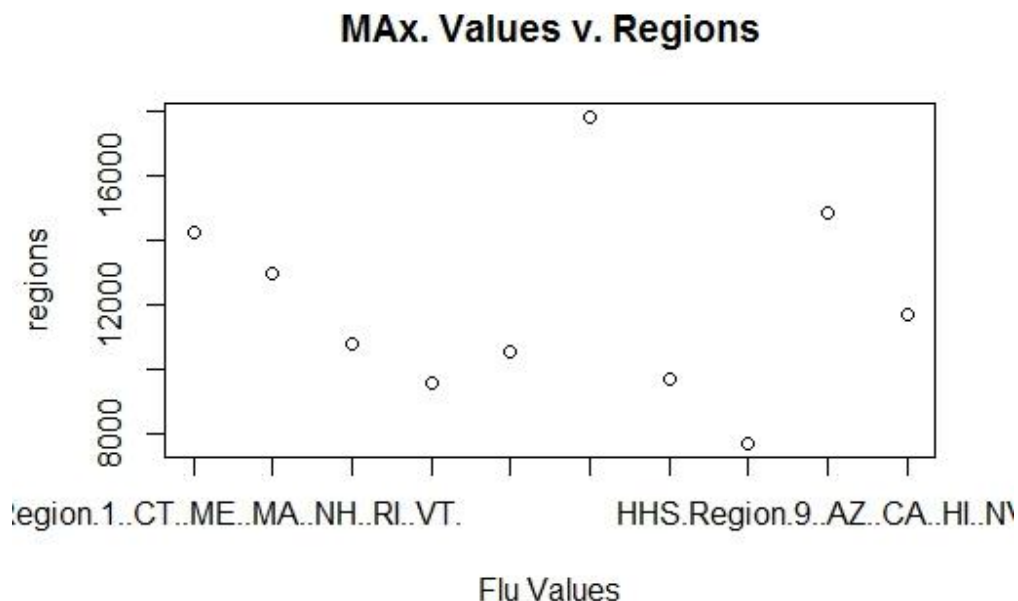
Similarly, we can plot the graph of various geographical regions as per requirement to know the trend of Flu over the past 12 years. Logically too we know that Flu outbreak is imminent and common in winter which is confirmed by this statistical analysis.

Metric 2 - Max:

The max will help the authorities track the levels of outbreak. A decreased Max over the years will be a positive sign for the authorities and they can be happy that they are on the right track. Also Max will help the authorities zero in on the regions where the outbreak is pre-dominant.

Code

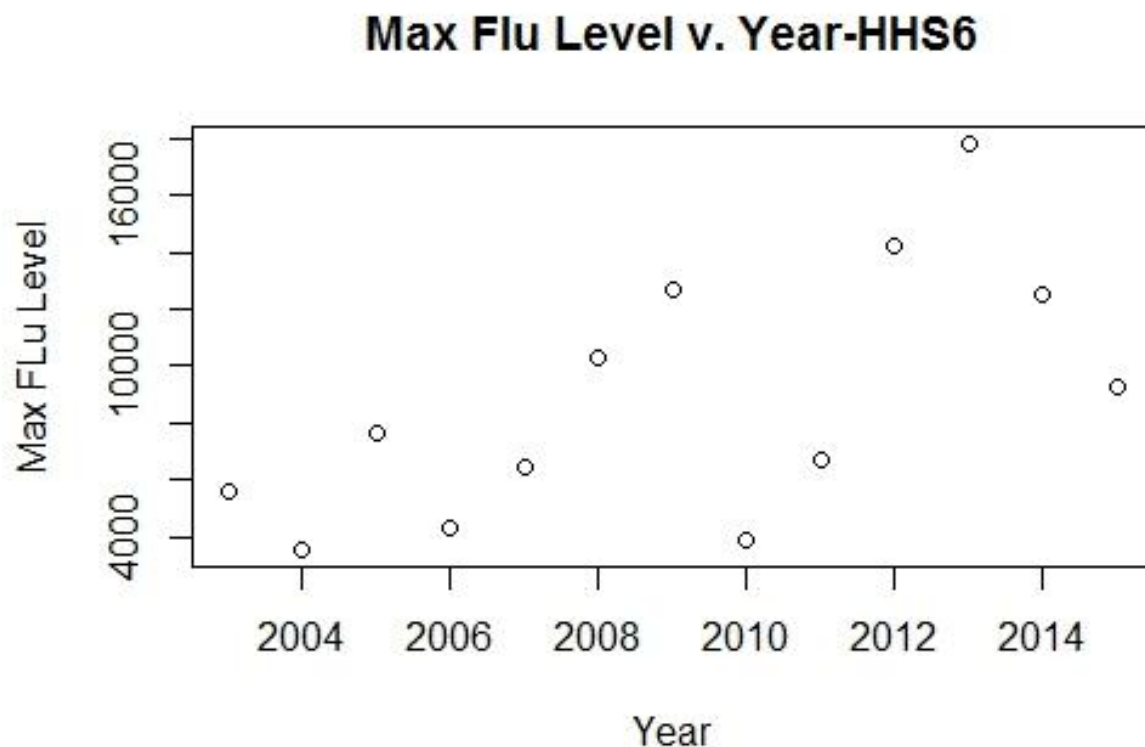
```
> tempdata<-data[54:63]
> temp<-data.frame(apply(tempdata, 2, function(x) max(x, na.rm = TRUE)))
> tempDF<-data.frame(temp)
> tempDF$names<-rownames(tempDF)
> colnames(tempDF)[2]<-"region"
> colnames(tempDF)[1]<-"value"
> plot(tempDF$value,axes=FALSE,xlab="Flu Values",ylab="regions",main="Max. Values v. Regions")
> axis(2)
> axis(1,at=seq_along(tempDF$value),labels=as.character((tempDF$region),las=2))
> box()
```



The above graph will help authorities realize the fact that region 6 has had the maximum outbreak ever. Hence, they can zero in on this region to see whether this region has a continuous high Flu or the Max over the years has changed.

Code

```
> period<-paste(year(data$Date))  
> temp<-data.frame(aggregate(HHS.Region.6..AR..LA..NM..OK..TX.,list(period),m  
ax))  
> plot(temp$Group.1,temp$x,xlab="Year",ylab="Max FLu Level",main="Max Flu Lev  
el v. Year-HHS6")
```



This trend may be a worrying thing. Over the years, Max have changed drastically but more off, after a stiff fall, it has risen steadily over a period of few years in HHS6 region.

Conclusion for choosing Max

Hence from the above sample, authorities can study the max levels of Flu of a region over the past 12 years and analyze how bad or how good is the situation in a given a region.

If the Max has been continuously increasing over the years, then the authorities need to evaluate the cause for the increasing Flu and find a way to stop this climb and bring the max Flu levels down.

Conclusion for choosing this combination – Mean and Max

Only mean or only max would have definitely been useful for the authorities. But the combined analysis that can be derived from using the two metrics can help us make a clear conclusion of the Flu situation. For example, if a region has been showing a record Max over the years with a relatively lower mean, this means that it is just that during certain periods the Flu hits the region very badly and rest of the times it is not so severe. Further if periodical mean is also very low, we can conclude that Flu hits the regions very rarely each year but whenever it does, it is sort of an epidemic and spreads very quickly on a large scale.

e)

Source of population data:

<https://www.census.gov/popest/data/national/totals/2015/files/NST-EST2015-alldata.csv>

Code

Fetch the data for 2015 for just the states of USA. Once we have this data, find the maximum for each state and create a data frame having state name and maximum value:

```
> fluData<-read.csv(file.choose(),head=TRUE,sep=" ",skip=588)
> header <- read.csv(file.choose(), nrows = 1, header = FALSE, sep = ' ', stringsAsFactors = FALSE)
> colnames(fluData) <- unlist(header)
> fluData<-fluData[3:53]
> MaxFlu<-data.frame(apply(fluData, 2, function(x) max(x, na.rm = TRUE)))
> colnames(MaxFlu)[1]<-"value"
```

Based on the data fetched from above mentioned URL, create a Data frame consisting all the states and union territories along with the estimated 2015 population:

```
> Population<-read.csv(file.choose(),head=TRUE,sep=" ")
> PopulationDF<-data.frame(Population[, 5])
> colnames(PopulationDF)[1]<-"name"
> PopulationDF[, '2015Pop']<-data.frame(Population[,13])
```

Finally, create a new dataframe mergeState which merges MaxFlu and Population DF based on the column names, ie, only the country names.

```
> mergeState<-data.frame(merge(MaxFlu,PopulationDF, by.x=c("names"),by.y=c("name")))
> plot(mergeState$value,mergeState$X2015Pop,xlab="Peak State Values", ylab="State Population", main="Population v. Peak Flu")
```

Output

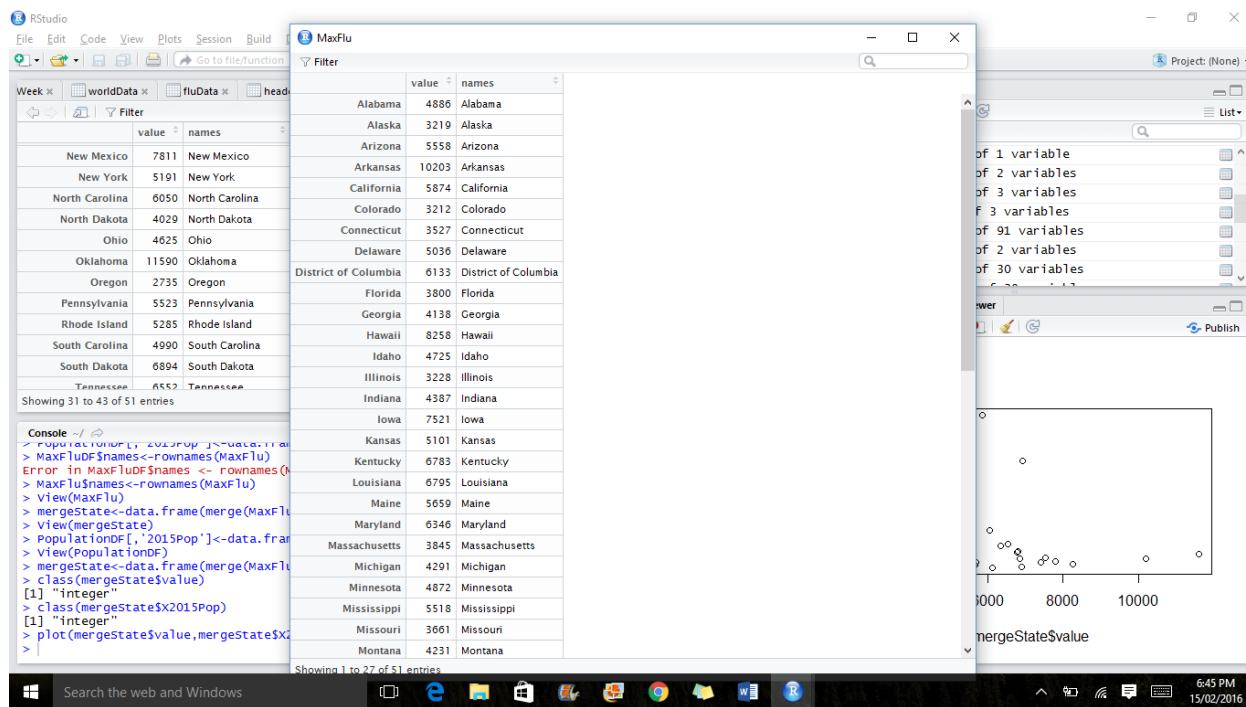


Fig: MaxFlu Dataframe

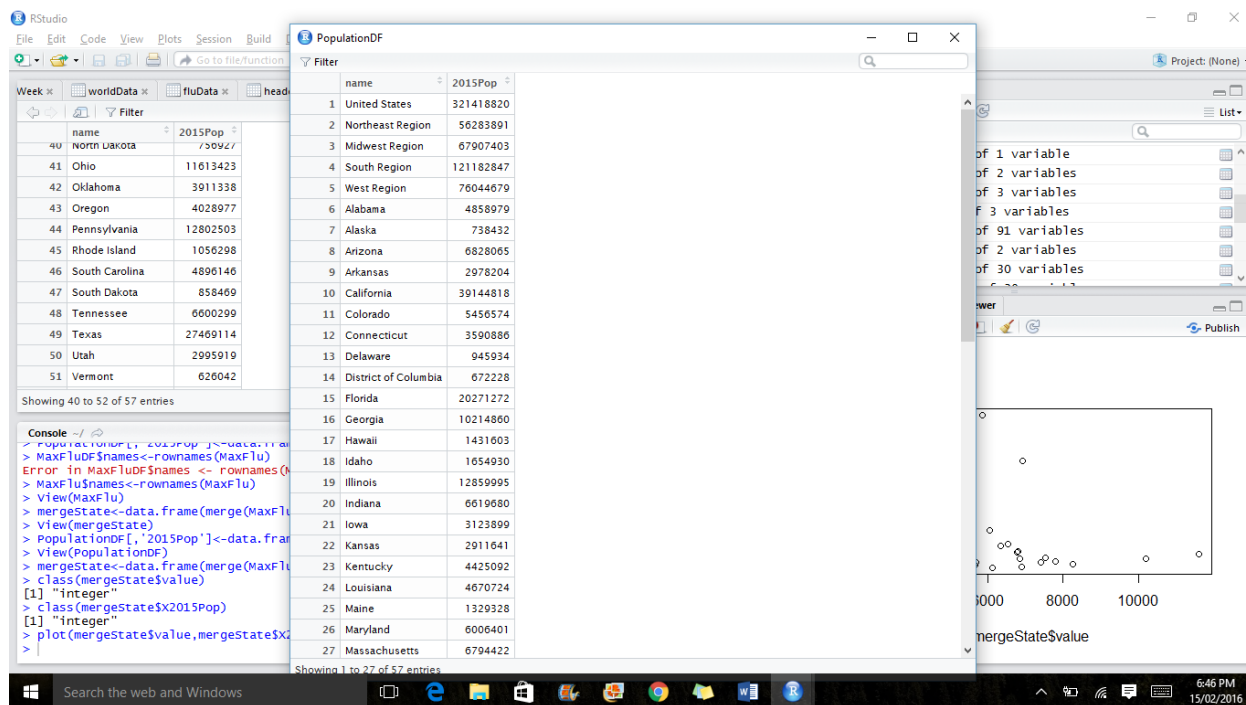


Fig: PopulationDF Dataframe

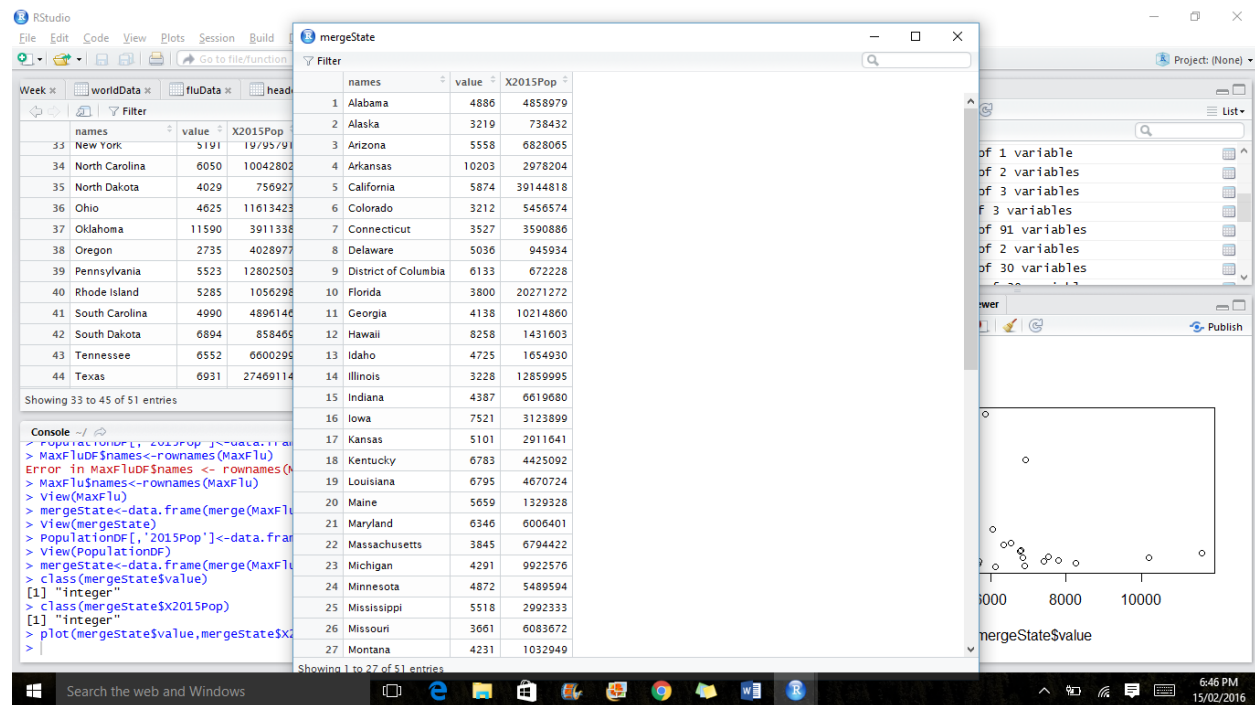
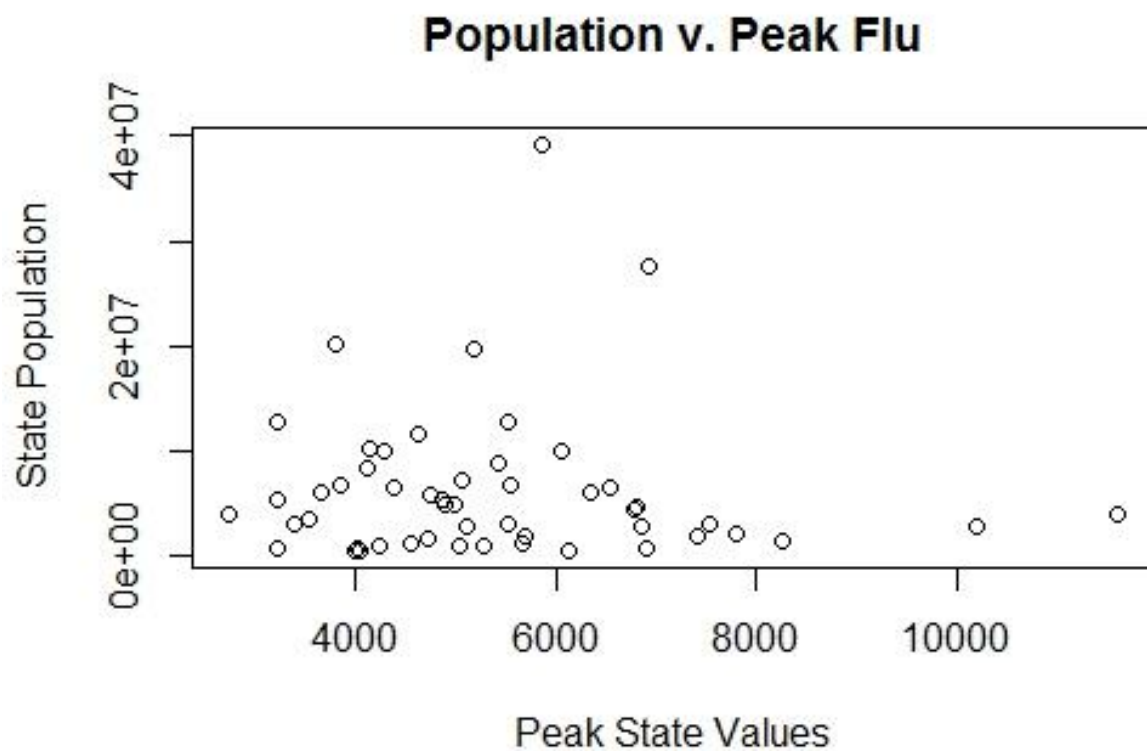


Fig: mergeState Dataframe



Conclusion

Based on the population and the Peak Values of each state, I decided to plot a graph between the two to analyze the relationship between them. But as the above graph suggests, there is no direct relation between the two attributes. The population of a state and its peak value in 2015 have NO direct relation. We can clearly see that the highest Flu is of a state having one of the lowest populations. Also states having similar populations are seen to have Flu values ranging from 4000 to 10000.

The above conclusion is further supported by means of regression.

```
> model<-lm(mergeState$value~mergeState$X2015Pop)
> summary(model)
```

```
Call:
lm(formula = mergeState$value ~ mergeState$X2015Pop)
```

```
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 ***
mergeState$X2015Pop -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
```

Based on the Multiple R-Squared and Adjusted R-Squared values, we can conclude that there is no relation between the Peak Flu values and the corresponding state populations. This can be said as both these values are nearly 0.1% (0.001) which means that the Max Flu value and the corresponding states' populations have no relation.

Question 2 (Simple Data Analysis)

Source for latitude: https://developers.google.com/public-data/docs/canonical/countries_csv

Code

From the data provided, fetch the data only of 2015, then fetch the maximum value for each country. This value is stored in data frame MaxWeekDF

```
> worldData<-read.csv(file.choose(),head=TRUE,sep=" ",skip=627)
> header <- read.csv(file.choose(), nrows = 1, header = FALSE, sep = ' ', stringsAsFactors = FALSE)
> colnames(worldData) <- unlist(header)
> MaxWeek<-data.frame(apply(worldData, 2, function(x) max(x, na.rm = TRUE)))
> colnames(MaxWeek)[1]<-"value"
> MaxWeekDF<-data.frame(MaxWeek)
> MaxWeekDF$names<-rownames(MaxWeekDF)
> MaxWeekDF<-MaxWeekDF[2:length(MaxWeekDF$names),]
```

From the URL mentioned above, fetch all the data then create a new data frame LatitudeDF which just contains the country names and latitude. (the row names of the data frame have been converted to a column)

```
> install.packages("XML")
> install.packages("RCurl")
> library("XML")
> library("RCurl")
> URL<-getURL("https://developers.google.com/public-data/docs/canonical/countries_csv")
> htmlPage<-data.frame(readHTMLTable(URL,header = TRUE,as.data.frame = TRUE,width=1))
> LatitudeDF<-data.frame(htmlPage$name)
> colnames(LatitudeDF)[1]<-"name"
> LatitudeDF[, 'latitude']<-data.frame(htmlPage$latitude)
```

Finally, create a new dataframe merge which merges MaxWeekDF and LatitudeDF based on the column names, ie, only the country names.

```
> merge<-data.frame(merge(MaxWeekDF, LatitudeDF, by.x=c("names"), by.y=c("name")))
> merge$value = as.numeric(as.character(merge$value))
> merge$latitude<-as.numeric(as.character(merge$latitude))

> plot(merge$value, merge$latitude, ylab = "Latitude", xlab = "Peak week of Flu", main = "Latitude v. Flu count")
```

Output

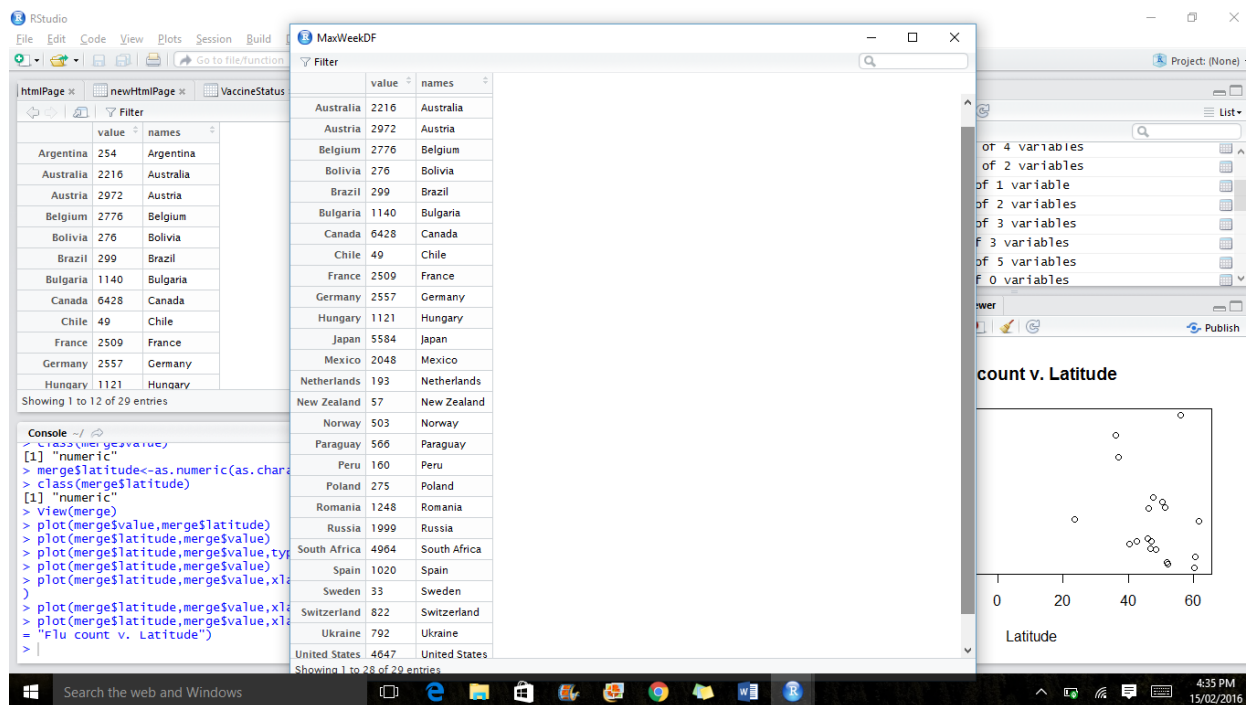


Fig:Data Frame : MaxWeekDF

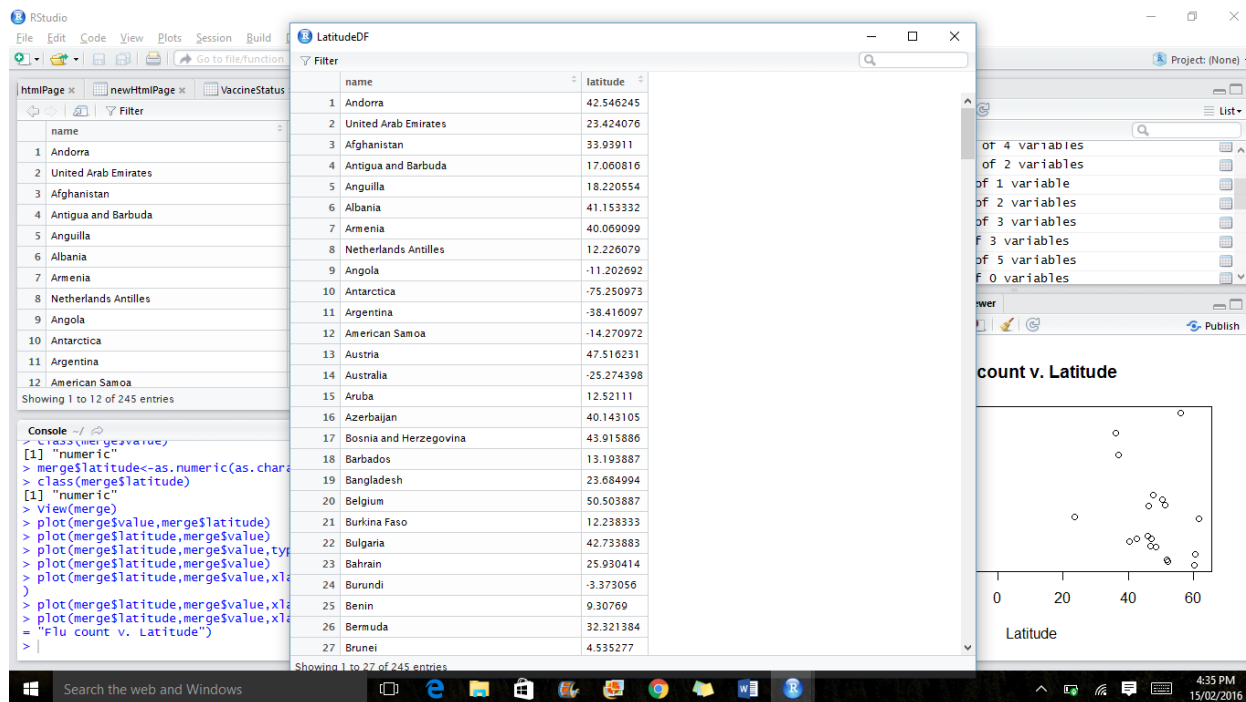


Fig:Data Frame : LatitudeDF

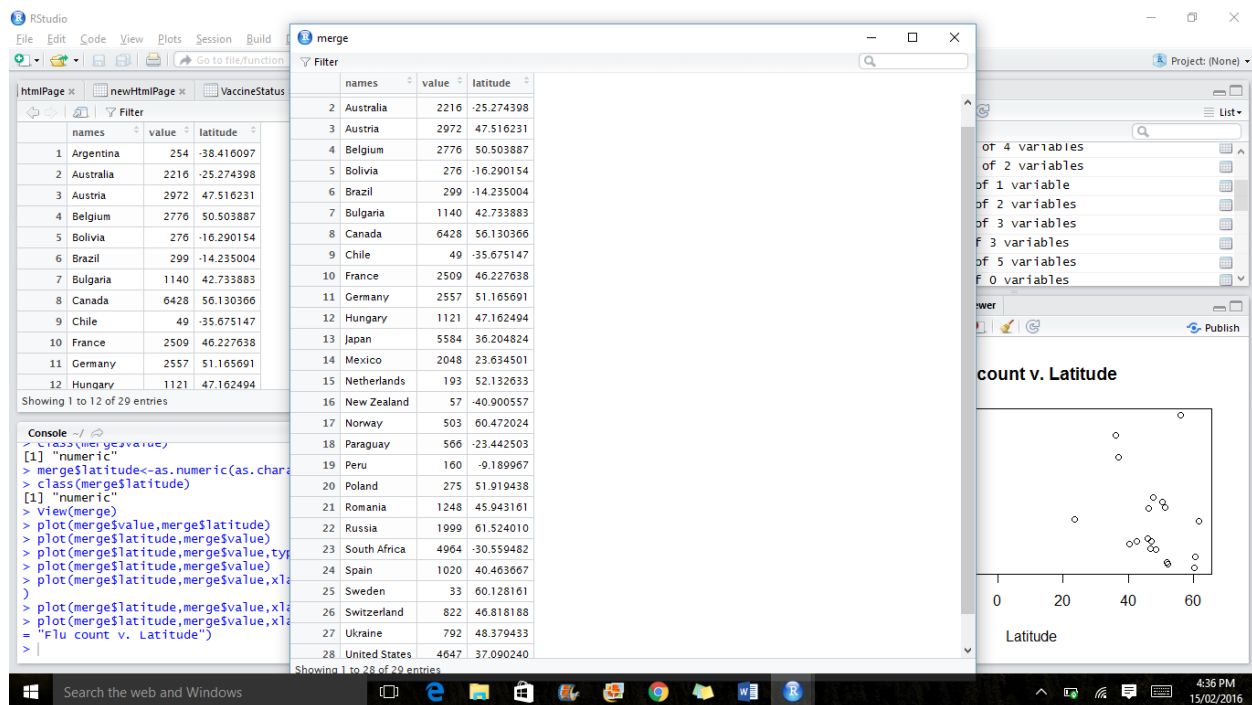
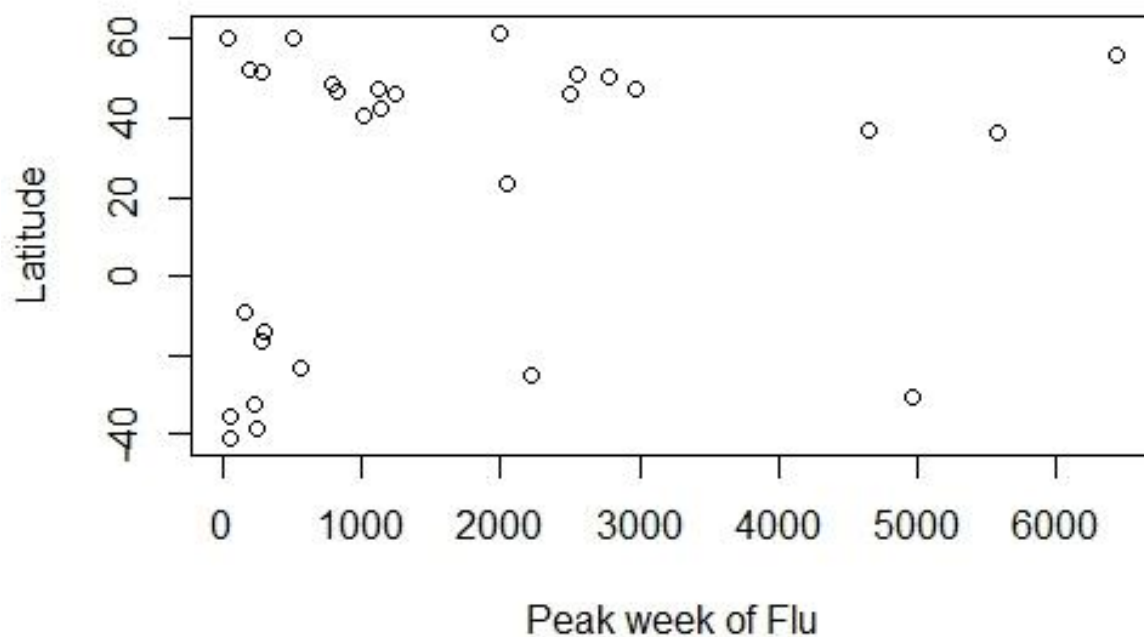


Fig:Data Frame : Merge

Latitude v. Flu count



Conclusion

Based on the graph we got, we can deduce that there exists no direct relation between the latitude of the country and Flu trends. We can hence conclude that the geographical position of a country has no effect on the maximum Flu value. This can be said because between latitudes 40 & 60, we have a maximum flu value ranging from nearly 0 to 6000.

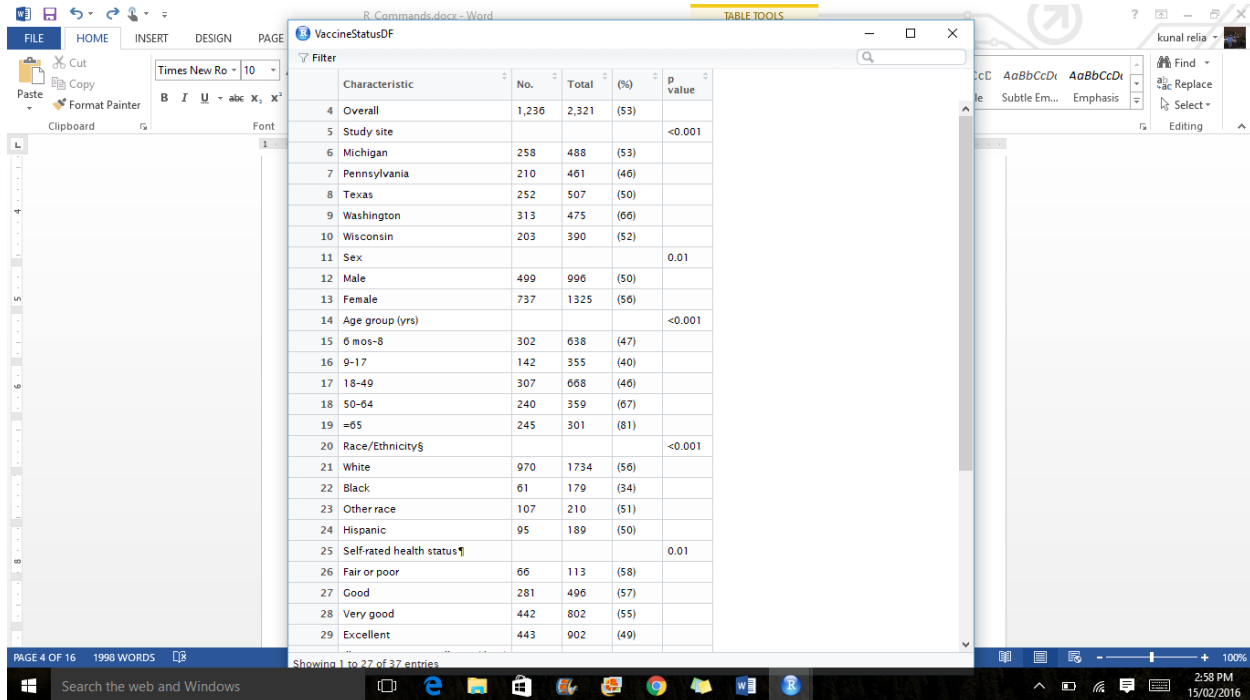
Question 3 (web Scrapping)

a) vaccine status data from the given URL

Code

```
>URL <-  
getURL("http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6401a4.htm?s_cid=mm6401a4_w")  
>htmlPage<-data.frame(readHTMLTable(URL,header = TRUE,as.data.frame = TRUE,  
skip.rows=41, which=1, stringsAsFactors=FALSE))  
>VaccineStatusDF<-data.frame(htmlPage$V1)  
>colnames(VaccineStatusDF)[1]<- 'Characteristic'  
>VaccineStatusDF[, 'No. ']<-data.frame(htmlPage$V7)  
>VaccineStatusDF[, 'Total']<-data.frame(htmlPage$V8)  
>VaccineStatusDF[, '(%)']<-data.frame(htmlPage$V9)  
>VaccineStatusDF[, 'p value']<-data.frame(htmlPage$V10)  
>VaccineStatusDF<-VaccineStatusDF[4:length(VaccineStatusDF$Characteristic),]  
>View(VaccineStatusDF)
```

Output



| | Characteristic | No. | Total | (%) | p value |
|----|---------------------------|-------|-------|------|---------|
| 4 | Overall | 1,236 | 2,321 | (53) | |
| 5 | Study site | | | | <0.001 |
| 6 | Michigan | 258 | 488 | (53) | |
| 7 | Pennsylvania | 210 | 461 | (46) | |
| 8 | Texas | 252 | 507 | (50) | |
| 9 | Washington | 313 | 475 | (66) | |
| 10 | Wisconsin | 203 | 390 | (52) | |
| 11 | Sex | | | | 0.01 |
| 12 | Male | 499 | 996 | (50) | |
| 13 | Female | 737 | 1325 | (56) | |
| 14 | Age group (yrs) | | | | <0.001 |
| 15 | 0 mos-8 | 302 | 638 | (47) | |
| 16 | 9-17 | 142 | 355 | (40) | |
| 17 | 18-49 | 307 | 668 | (46) | |
| 18 | 50-64 | 240 | 359 | (67) | |
| 19 | =65 | 245 | 301 | (81) | |
| 20 | Race/Ethnicity§ | | | | <0.001 |
| 21 | White | 970 | 1734 | (56) | |
| 22 | Black | 61 | 179 | (34) | |
| 23 | Other race | 107 | 210 | (51) | |
| 24 | Hispanic | 95 | 189 | (50) | |
| 25 | Self-rated health status¶ | | | | 0.01 |
| 26 | Fair or poor | 66 | 113 | (58) | |
| 27 | Good | 281 | 496 | (57) | |
| 28 | Very good | 442 | 802 | (55) | |
| 29 | Excellent | 443 | 902 | (49) | |

| | | | | | |
|---|----------------|-------|-------|------|---------|
| 4 | Characteristic | No. | Total | (%) | p value |
| 5 | Overall | 1,236 | 2,321 | (53) | |
| | Study site | | | | <0.001 |

| | | | | | |
|----|------------------------------------|-----|-------|------|--------|
| 6 | Michigan | 258 | 488 | (53) | |
| 7 | Pennsylvania | 210 | 461 | (46) | |
| 8 | Texas | 252 | 507 | (50) | |
| 9 | Washington | 313 | 475 | (66) | |
| 10 | Wisconsin | 203 | 390 | (52) | |
| 11 | Sex | | | | 0.01 |
| 12 | Male | 499 | 996 | (50) | |
| 13 | Female | 737 | 1325 | (56) | |
| 14 | Age group (yrs) | | | | <0.001 |
| 15 | 6 mos-8 | 302 | 638 | (47) | |
| 16 | 9-17 | 142 | 355 | (40) | |
| 17 | 18-49 | 307 | 668 | (46) | |
| 18 | 50-64 | 240 | 359 | (67) | |
| 19 | =65 | 245 | 301 | (81) | |
| 20 | Race/Ethnicity§ | | | | <0.001 |
| 21 | White | 970 | 1734 | (56) | |
| 22 | Black | 61 | 179 | (34) | |
| 23 | Other race | 107 | 210 | (51) | |
| 24 | Hispanic | 95 | 189 | (50) | |
| 25 | Self-rated health status¶ | | | | 0.01 |
| 26 | Fair or poor | 66 | 113 | (58) | |
| 27 | Good | 281 | 496 | (57) | |
| 28 | Very good | 442 | 802 | (55) | |
| 29 | Excellent | 443 | 902 | (49) | |
| 30 | Illness onset to enrollment (days) | | | | 0.15 |
| 31 | <3 | 420 | 805 | (52) | |
| 32 | 3-4 | 458 | 883 | (52) | |
| 33 | 5-7 | 358 | 633 | (57) | |
| 34 | Influenza test result | | | | |
| 35 | Negative | 771 | 1,371 | (56) | |
| 36 | Influenza B positive** | 17 | 35 | (49) | |
| 37 | Influenza A positive** | 448 | 916 | (49) | |
| 38 | A (H1N1)pdm09 | 0 | 0 | (0) | |
| 39 | A (H3N2) | 407 | 842 | (48) | |
| 40 | A subtype pending | 41 | 74 | (55) | |

b)Table from the some random URL

URL used: https://developers.google.com/public-data/docs/canonical/countries_csv

Code

```
>URL<-getURL("https://developers.google.com/public-data/docs/canonical/countries_csv")
>htmlPage<-data.frame(readHTMLTable(URL,header = TRUE,as.data.frame = TRUE,which=1))
>View(htmlPage)
```

Output

The screenshot shows the RStudio interface. The 'htmlPage' data frame is displayed in a table view, showing columns for country, latitude, longitude, and name. The console shows the execution of the code, including the URL and the readHTMLTable function. The environment pane on the right shows the 'htmlPage' object with 245 entries.

| | country | latitude | longitude | name |
|---|---------|-----------|------------|----------------------|
| 1 | AD | 42.546245 | 1.601554 | Andorra |
| 2 | AE | 23.424076 | 53.847818 | United Arab Emirates |
| 3 | AF | 33.93911 | 67.709953 | Afghanistan |
| 4 | AG | 17.060816 | -61.796428 | Antigua and Barbuda |
| 5 | AI | 18.220554 | -63.068615 | Anguilla |
| 6 | AL | 41.153332 | 20.168331 | Albania |
| 7 | AM | 40.069099 | 45.038189 | Armenia |

| | country | latitude | longitude | name |
|---|---------|-----------|------------|----------------------|
| 1 | AD | 42.546245 | 1.601554 | Andorra |
| 2 | AE | 23.424076 | 53.847818 | United Arab Emirates |
| 3 | AF | 33.93911 | 67.709953 | Afghanistan |
| 4 | AG | 17.060816 | -61.796428 | Antigua and Barbuda |
| 5 | AI | 18.220554 | -63.068615 | Anguilla |
| 6 | AL | 41.153332 | 20.168331 | Albania |
| 7 | AM | 40.069099 | 45.038189 | Armenia |

Further Entries till 245 not shown here...

Group Work:

Some of the questions of this assignment were discussed with Fenil Tailor. Only the idea of solving a few questions were discussed by us followed by individual application of the ideas.