



CIS 5270: Business Intelligence

Spring 2018

Analysis of U.S. Chronic Disease Indicators (CDI) by state using R

Submitted to: Dr. Shilpa Balan

Team Members: Tanmai Aurangabadkar CIN: 306605265

Era Singh Kajal

CIN: 306605200

A. Dataset URL

<https://catalog.data.gov/dataset/u-s-chronic-disease-indicators-cdi-e50c9>

US Chronic Diseases Indicators dataset comes from CDC's Division of Population Health that provides cross-cutting set of 124 indicators that were developed by consensus and that allows states and territories and large metropolitan areas to uniformly define, collect, and report chronic disease data that are important to public health practice and available for states, territories and large metropolitan areas. In addition to providing access to state specific indicator data, the CDI web site serves as a gateway to additional information and data resources. The CDI website enables public health professionals and policymakers to retrieve uniformly defined state-level and selected metropolitan-level data for chronic diseases and risk factors that have a substantial impact on public health. These indicators are essential for surveillance, prioritization, and evaluation of public health interventions for chronic disease.

B. Data Cleaning:

1. Data set contained more than 20 columns so, we removed unwanted columns such as “Response”, “DataValueUnit” etc.

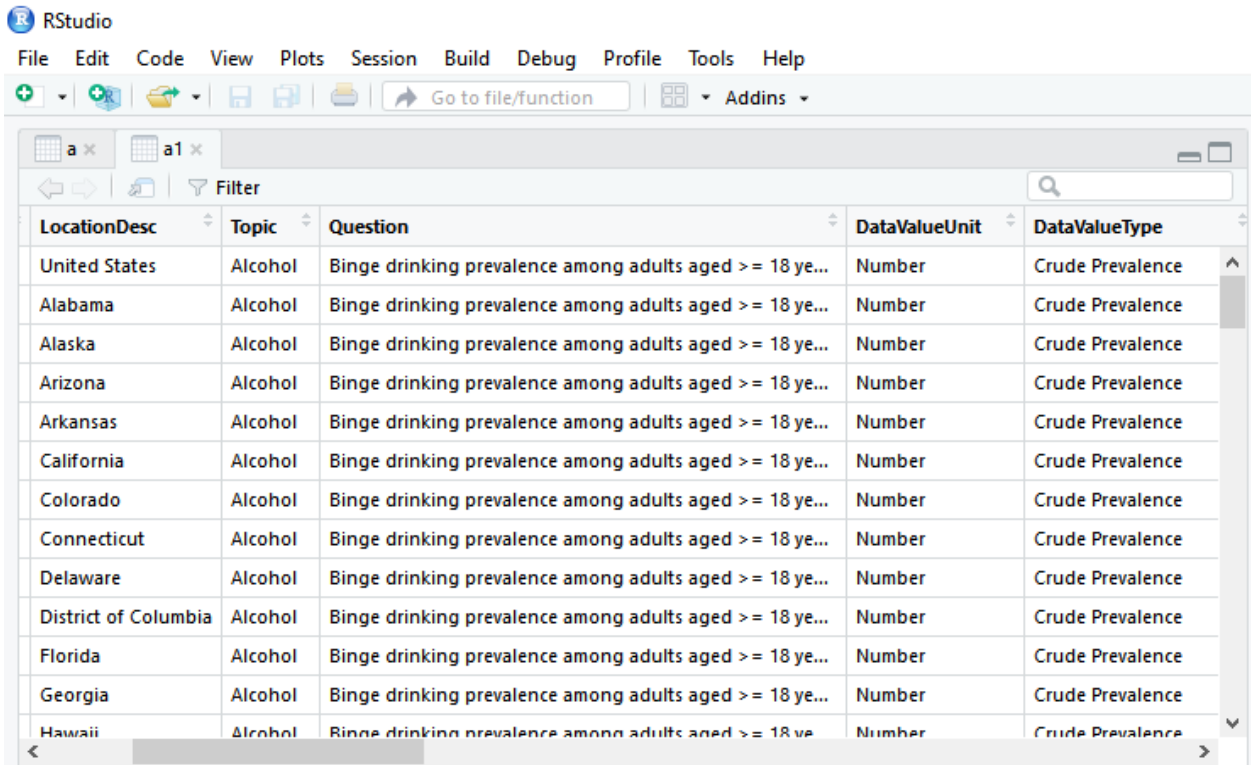
Before cleaning:

[illegible]

Code:

```
a1<-a[c(1,2,3,4,5,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22)]  
View(a1)
```

After Cleaning:



RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

LocationDesc	Topic	Question	DataValueUnit	DataValueType
United States	Alcohol	Binge drinking prevalence among adults aged >= 18 ye...	Number	Crude Prevalence
Alabama	Alcohol	Binge drinking prevalence among adults aged >= 18 ye...	Number	Crude Prevalence
Alaska	Alcohol	Binge drinking prevalence among adults aged >= 18 ye...	Number	Crude Prevalence
Arizona	Alcohol	Binge drinking prevalence among adults aged >= 18 ye...	Number	Crude Prevalence
Arkansas	Alcohol	Binge drinking prevalence among adults aged >= 18 ye...	Number	Crude Prevalence
California	Alcohol	Binge drinking prevalence among adults aged >= 18 ye...	Number	Crude Prevalence
Colorado	Alcohol	Binge drinking prevalence among adults aged >= 18 ye...	Number	Crude Prevalence
Connecticut	Alcohol	Binge drinking prevalence among adults aged >= 18 ye...	Number	Crude Prevalence
Delaware	Alcohol	Binge drinking prevalence among adults aged >= 18 ye...	Number	Crude Prevalence
District of Columbia	Alcohol	Binge drinking prevalence among adults aged >= 18 ye...	Number	Crude Prevalence
Florida	Alcohol	Binge drinking prevalence among adults aged >= 18 ye...	Number	Crude Prevalence
Georgia	Alcohol	Binge drinking prevalence among adults aged >= 18 ye...	Number	Crude Prevalence
Hawaii	Alcohol	Binge drinking prevalence among adults aged >= 18 ye...	Number	Crude Prevalence

2. Data Value Unit column has different name DataValueUnit.1 instead of DataValueUnit.

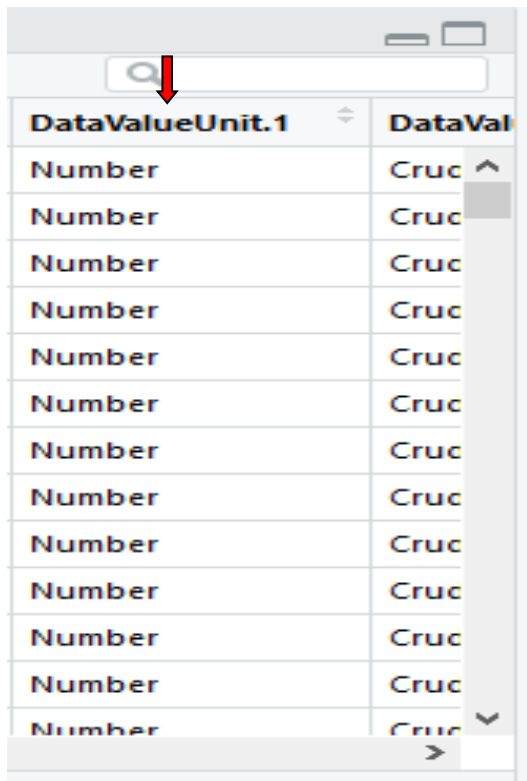
We changed the column name with the help of following code.

Code:

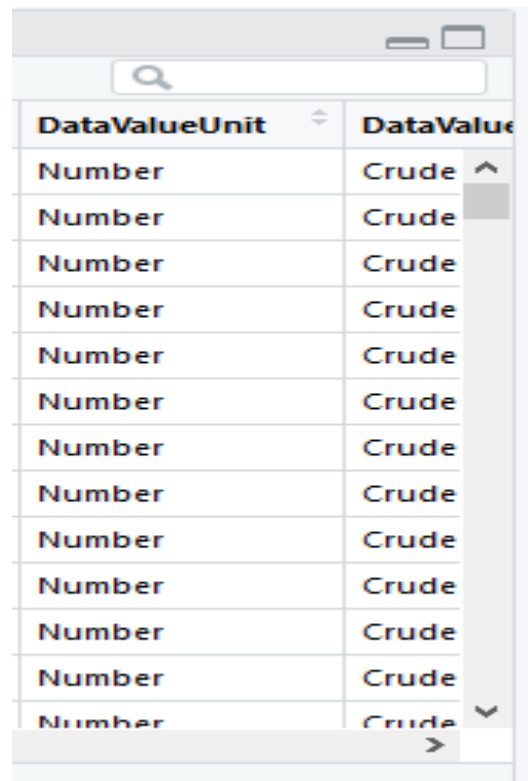
```
names(a1)[5] <- "DataValueUnit"
```

Before cleaning

After cleaning



DataValueUnit.1	DataValue
Number	Crude
Number	Crude
Number	Crude
Number	Crude
Number	Crude
Number	Crude
Number	Crude
Number	Crude
Number	Crude
Number	Crude
Number	Crude
Number	Crude
Number	Crude

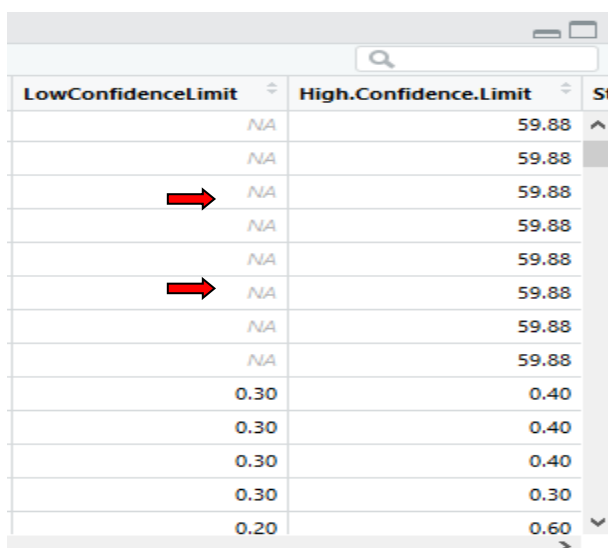


DataValueUnit	DataValue
Number	Crude
Number	Crude
Number	Crude
Number	Crude
Number	Crude
Number	Crude
Number	Crude
Number	Crude
Number	Crude
Number	Crude
Number	Crude
Number	Crude
Number	Crude

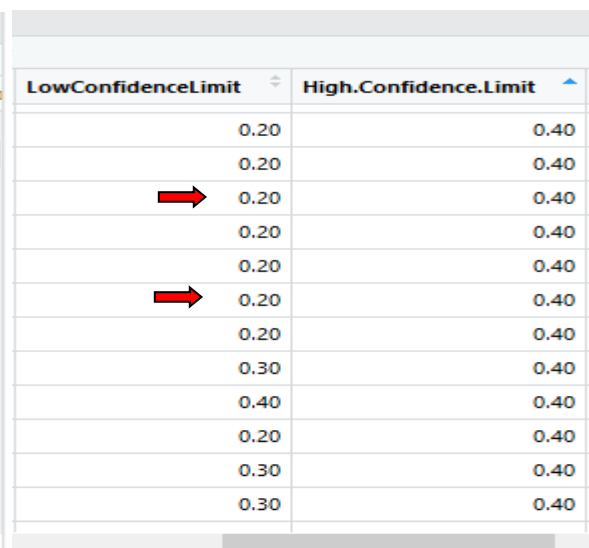
- Low Confidence Limit column had many NA values. We replaced NA values with average of the low confidence limit column values.

Before Cleaning

After Cleaning



LowConfidenceLimit	High.Confidence.Limit	St
NA	59.88	
NA	59.88	
NA	59.88	
NA	59.88	
NA	59.88	
NA	59.88	
NA	59.88	
NA	59.88	
0.30	0.40	
0.30	0.40	
0.30	0.40	
0.30	0.30	
0.20	0.60	



LowConfidenceLimit	High.Confidence.Limit
0.20	0.40
0.20	0.40
0.20	0.40
0.20	0.40
0.20	0.40
0.20	0.40
0.20	0.40
0.20	0.40
0.30	0.40
0.40	0.40
0.20	0.40
0.30	0.40
0.30	0.40

Code:

```
Low_limit<-a1$LowConfidenceLimit
```

```
ave(Low_limit,na.rm=TRUE)
```

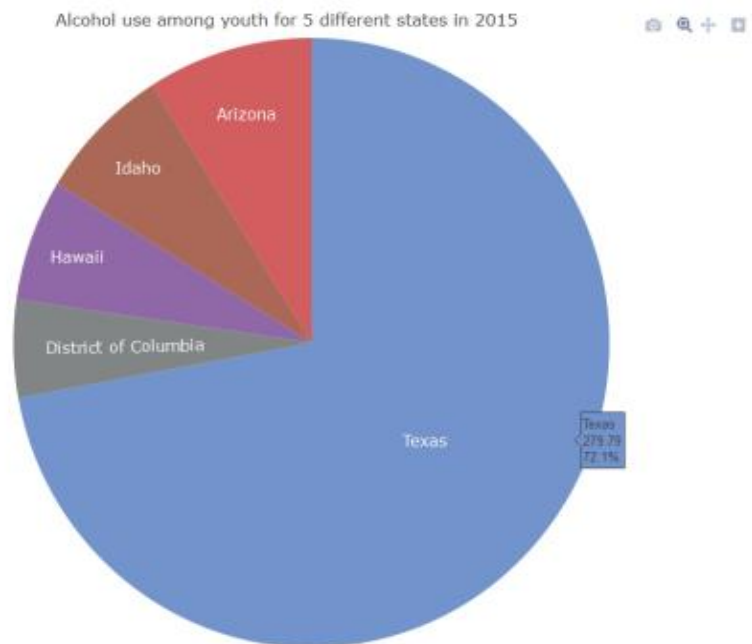
```
Low_limit [is.na(Low_limit)]=ave(Low_limit, na.rm = TRUE)
```

```
a1$ LowConfidenceLimit <- Low_limit
```

```
View(a1)
```

C. Data Visualization:

1. What is the percentage of alcohol use among the youth for 5 different states in year 2015?

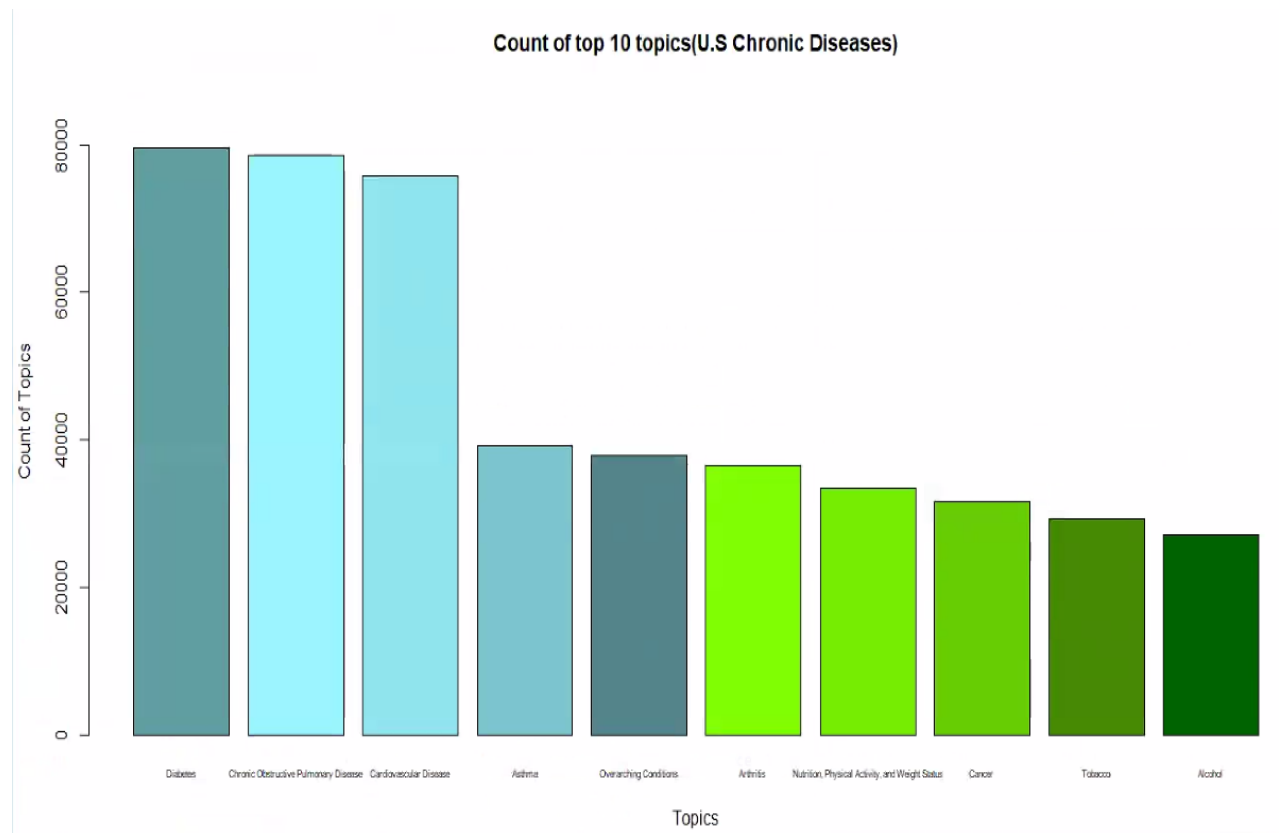


(Highlights from R script - dplyr, plotly package, pie chart)

Each state has a percentage and data value associated with it. It is important to find out that which state gives largest percentage alcohol use among youth of United States. Above pie chart states the

alcohol use among youth for 5 different states of United States in year 2015. It also shows percentage of alcohol use among youth for each state and different colors make it easier to understand. Highest percentage of alcohol use among youth is seen for Texas and lowest percentage of alcohol use among youth is seen for District of Columbia.

2. Display the top 10 US chronic diseases on a bar plot.

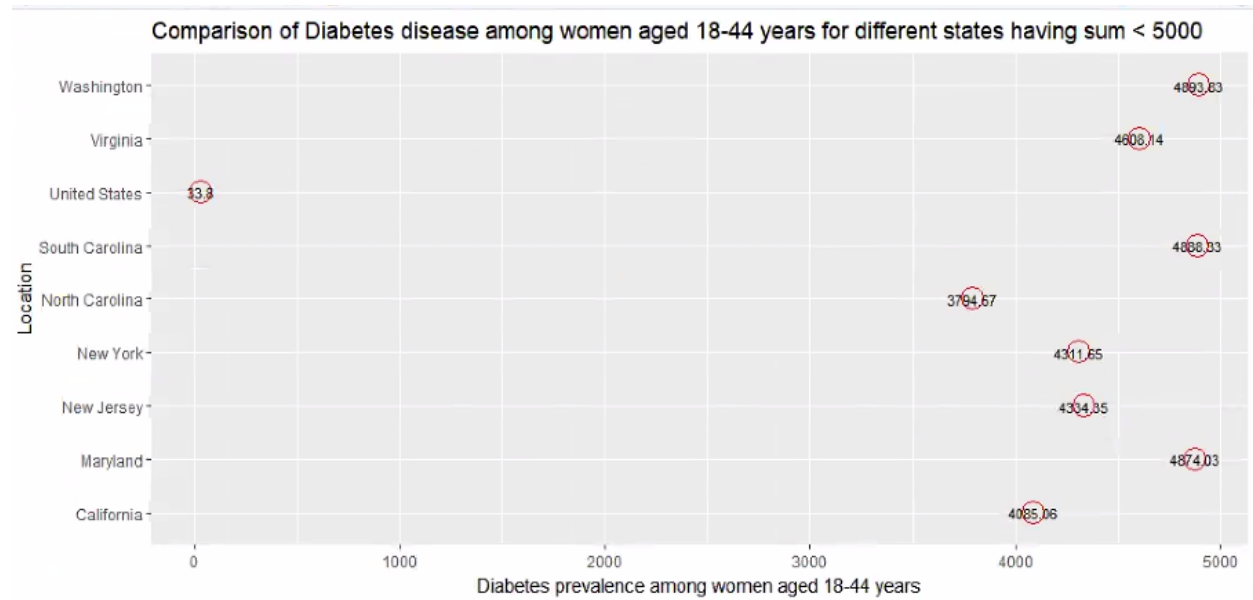


(Highlights from R script – dplyr, bar plot, bar chart, different colors, user defined function)

From the above Bar chart, it represents the number of Top 10 U.S. chronic diseases which depicts that Diabetes, Chronic Obstructive Pulmonary Diseases followed by Cardiovascular Diseases have received high count. It can be recommended that Diabetes should be given special attention by finding the reason behind for highest number seen among all the topics. This can be done by

analyzing the prevalence of diabetes disease among different age groups. To get deeper insights, it can be predicted which question shows that the highest count of diabetes by filtering the data.

3. Show the comparison of diabetes among women aged 18-44 years for different states having sum<5000.



(Highlights from R script – sqldf, ggplot2 package, sum, mean, Scatter Plot, Statistical Analysis)

As seen from the above scatter plot, diabetes prevalence among women aged 18-44 years where the data value has sum less than 5000 are very high in the state Washington having sum 4893.83. This plot shows that United States has the lowest diabetes prevalence among women aged 18-44 years which has sum 33.8. Also, the visualization exhibits that among 9 different locations having sum less than 5000, 6 locations are having sum ranging between 4000-5000 that shows diabetes prevalence among women aged 18-44 years. Mean and standard deviation for diabetes among

women aged 18-44 years for different states having sum less than 5000 are 166.5616 and 37.436 respectively.

R code for Analysis and Visualization

#Data cleaning code in R

```
setwd("C:/Users/Era Kajal/Desktop")

a<-read.csv("disease.csv",header=T,sep=",")

View(a)

install.packages("dplyr")

library("dplyr")

a1<-a[c(1,2,3,4,5,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22)]

View(a1)

Low_limit<-a1$LowConfidenceLimit

ave(Low_limit,na.rm=TRUE)

Low_limit [is.na(Low_limit)]=ave(Low_limit, na.rm = TRUE)

a1$ LowConfidenceLimit <- Low_limit

View(a1)

names(a1)[5] <- "DataValueUnit"

View(a1)
```

Visualization#1 Pie chart for alcohol use among the youth for 5 different states in year 2015

```
install.packages("dplyr")

library(dplyr)

install.packages("plotly")
```



```

library(plotly)

States <- c("District of Columbia", "Hawaii", "Idaho", "Arizona", "Texas")

Topic1 <-c("Alcohol use among youth")

Period <-c("2015")

ValueStore<-group_by(a,LocationDesc,Question,Year) %>% filter(LocationDesc %in% States,

Question %in% Topic1,Year %in% Period)

View(ValueStore)

colors <- c('rgb(211,94,96)', 'rgb(128,133,133)', 'rgb(144,103,167)', 'rgb(171,104,87)',

'rgb(114,147,203)')

plot_ly(ValueStore, labels = ~LocationDesc, values = ~DataValue, type = 'pie',textposition =

'inside',textinfo = 'label', insidetextfont = list(color = '#FFFFFF',size = 16),marker = list(colors =

colors)) %>% layout(title = 'Alcohol use among youth for 5 different states in 2015', xaxis =

list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE),yaxis = list(showgrid =

FALSE, zeroline = FALSE, showticklabels = FALSE))

```

Visualization #2: Bar chart for count of top 10 US chronic diseases

Script

```

library(dplyr)

install.packages("plotly")

library(plotly)

```

```

Cnt_Topics<-group_by(a,Topic) %>% summarise(cnt = n())

Count_Topics <-data.frame(Cnt_Topics)

View(Count_Topics)

sorted_topics<- Count_Topics[order(Count_Topics$cnt,decreasing = T),]

top10_topics<-head(sorted_topics,10)

barplot(top10_topics$cnt, ylim=c(0,1.1*max(top10_topics$cnt)), main="Count of top 10 topics(
US Chronic Diseases)", xlab="Topics",ylab = "Count of Topics", names.arg =
top10_topics$Topic, cex.names=0.5,col=c("cadetblue","cadetblue1", "cadetblue2", "cadetblue3",
" cadetblue4","chartreuse", "chartreuse2", "chartreuse3", "chartreuse4", "darkgreen"))

```

Console

```
source('barplot.R')
```

Visualization #3: Scatter Plot for comparison of Diabetes among women aged 18-44 years for different states having sum<5000

```
install.packages("sqldf")
```

```
library(sqldf)
```

Console:

```

output1<-sqldf("select Topic,Question,LocationDesc as Location,sum(DataValue) as sum from a
group by Topic,Question,LocationDesc") %>% filter(Topic== "Diabetes" & Question==
"Diabetes prevalence among women aged 18-44 years" & sum<5000)

p<-ggplot(output1, aes(x=sum, y=Location, color="red")) + geom_point(size =6, shape=1, color
="red") + geom_text(label = output1$sum, size = 3, color = "black")

```

```
p + labs(title="Comparison of Diabetes disease among women aged 18-44 years for different  
states having sum < 5000", x = "Diabetes prevalence among women aged 18-44 years", y =  
"Location")
```

```
b<-c(output1$sum)
```

```
mean(b)
```

```
[1] 166.5616
```

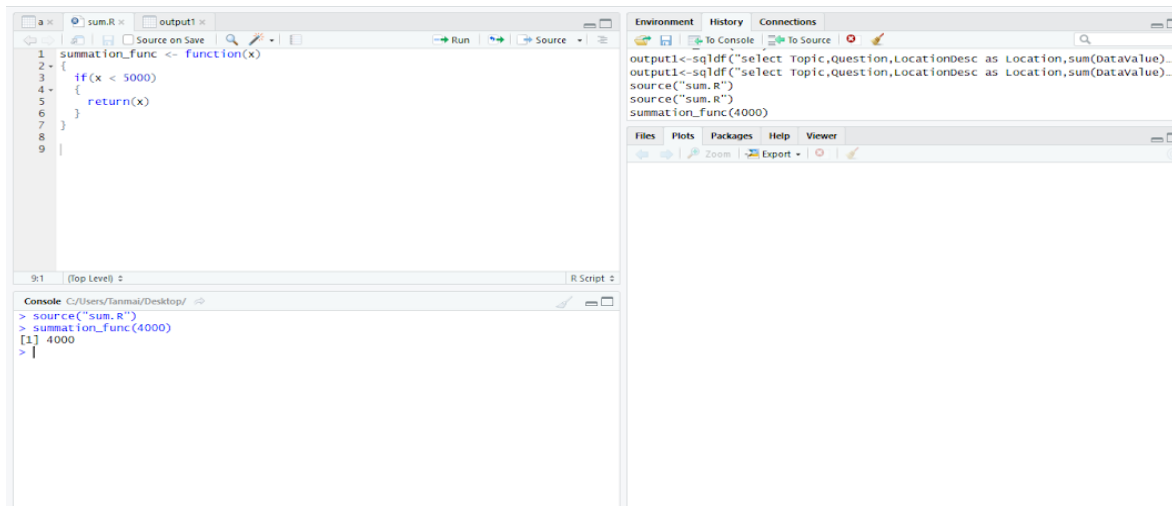
```
sd(b)
```

```
[1] 37.436
```

User Defined Function

```
summation_func <- function(x)
```

```
{  
  if(x < 5000)  
  {  
    print(x)  
  }  
}
```



Works Cited

<https://www.cdc.gov/mmwr/pdf/rr/rr6401.pdf>

<https://www.cdc.gov/mmwr/volumes/66/wr/mm6644a2.htm>

<https://plot.ly/r/>

<https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf>