# Data Cleaning R Plot

*Team Combat*

*12/5/2017*

**Data Cleaning and R Plot - Documentation**

Below are the steps used for data cleaning for this project.

```
# To get the current working directory:
getwd()
```

```
## [1] "C:/Users/user/Desktop/ISQA8086-2 Adrea Wigins/gitrep/Deliverables/RPlotDeliverable"
```

```
#To set the working directory to the desired location:
setwd("C:/Users/user/Desktop/ISQA8086-2 Adrea Wigins/gitrep/RawDataSet/")

#Loading the desired files to the data frame "data":
data<-read.csv("500CitiesLocalDataSetForBetterHealth2013.csv", header = T , na.strings = c("", "NA"))

#To view the loaded data frame:
View(data)

#To removed the unwanted columns from the loaded data frame and making a new data frame:
clean_data<-subset(data, select = -c(6,8,11,12,13,18,19,20))

#Renaming the column "Data_Value" to Data_Value in %, as the entire column values reflect percentages:
colnames(clean_data)[colnames(clean_data)=="Data_Value"]<-"Data_Value(in%)"

#Changing the "NA" values in "PopulationCount" Column to "Unknown" :
clean_data[["PopulationCount"]][is.na(clean_data[["PopulationCount"]])] <- "Unknown"

#To view the top 6 rows of our working dataframe:
head(clean_data)
```

```
##   Year     StateDesc CityName GeographicLevel        Category
## 1 2013 United States     <NA>              US Health Outcomes
## 2 2013 United States     <NA>              US Health Outcomes
## 3 2013 United States     <NA>              US      Prevention
## 4 2013 United States     <NA>              US      Prevention
## 5 2013 United States     <NA>              US      Prevention
## 6 2013 United States     <NA>              US      Prevention
##                                                                                  Measu
## 1                                                    High blood pressure among adults aged >=18 Yea
## 2                                                    High blood pressure among adults aged >=18 Yea
## 3 Taking medicine for high blood pressure control among adults aged >=18 Years with high blood pressu
## 4 Taking medicine for high blood pressure control among adults aged >=18 Years with high blood pressu
## 5                                                  Cholesterol screening among adults aged >=18 Yea
## 6                                                  Cholesterol screening among adults aged >=18 Yea
##           Data_Value_Type Data_Value(in%) Data_Value_Footnote
## 1 Age-adjusted prevalence            30.2                <NA>
## 2        Crude prevalence            32.4                <NA>
## 3 Age-adjusted prevalence            58.2                <NA>
## 4        Crude prevalence            77.1                <NA>
```

```
## 5 Age-adjusted prevalence            74.8                <NA>
## 6        Crude prevalence            76.4                <NA>
##   PopulationCount CategoryID  MeasureId
## 1       308745538    HLTHOUT     BPHIGH
## 2       308745538    HLTHOUT     BPHIGH
## 3       308745538    PREVENT      BPMED
## 4       308745538    PREVENT      BPMED
## 5       308745538    PREVENT CHOLSCREEN
## 6       308745538    PREVENT CHOLSCREEN
```

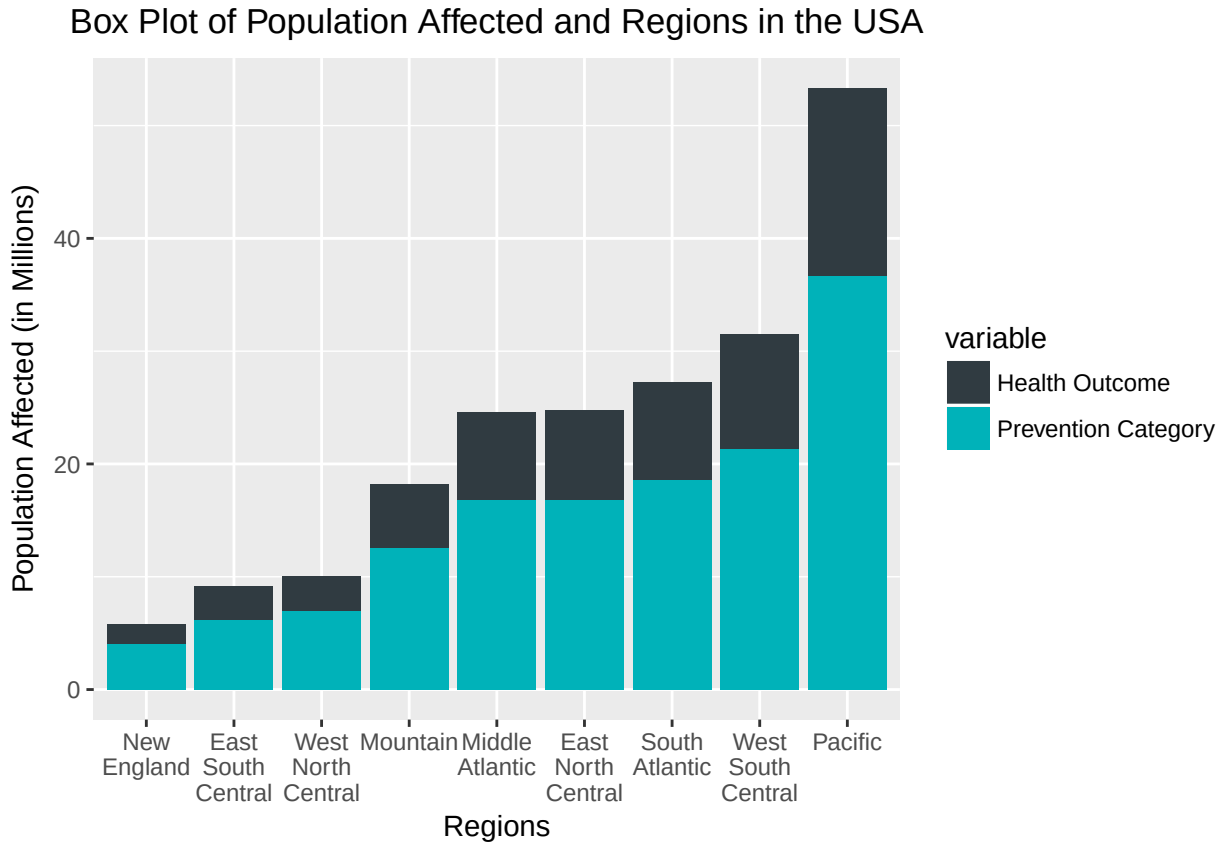
```
#To view the entire dataframe:
View(clean_data)

#To print the required table with set of arguments listed aboove for clean data:
write.csv(clean_data, "500_Cities__Local_Data_for_Better_Health2013_Clean.csv")
```

This document explains the steps in creating an R Plot for data visualization for the Research Questions raised in the previous assignments for 500 Local Cities Health Dataset The R Plot is created based on the R script generated previously as part of Data Preparation.

```
## Warning: package 'readr' was built under R version 3.4.2
```

```
## Warning: package 'stringr' was built under R version 3.4.2
```

```
#Removed unwanted state rowname with the value "USA"
mergedHealthOutcome_Prevention <- mergedHealthOutcome_Prevention[-8,]

#Create 3 temporary data frames for creating an R Plot
a <- data.frame(mergedHealthOutcome_Prevention$Region)
b <- data.frame(mergedHealthOutcome_Prevention$Population_Health_Outcomes)
c <- data.frame(mergedHealthOutcome_Prevention$Population_Prevention_Category)

#Merge all the data frames into one
df <- data.frame(a,b,c)

#Filter the data frame values based on Region
df <- melt(df, id.vars = "mergedHealthOutcome_Prevention.Region")

#Create an R plot using ggplot library
ggplot(df, aes(x=reorder(mergedHealthOutcome_Prevention.Region, value),
               y=value/1000000, fill=variable)) +
  geom_bar(stat = "identity") +
  scale_fill_manual(values = c("#303B41", "#00B2B9"),
                    labels= c("Health Outcome", "Prevention Category")) +
  xlab("Regions") +
  ylab("Population Affected (in Millions)") +
  ggtitle("Box Plot of Population Affected and Regions in the USA") +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_x_discrete(labels = function(x) str_wrap(x, width=5))
```




The above plot intends to provide a visualization on the different measure of Categories by which the 500 Cities Local Health Dataset is based upon. The categories, Health Outcomes and Preventive Measures, are combined to be depicted on a single bar graph divided based on Regions in the USA The visualization depicts that there has been a linear relation with the number of Preventive Measures with respect to Health outcomes in the 9 Regions in the USA

```
#Create a CSV of the cleaned data and R Script created in the previous assignment
state_and_region <- read.csv("Updated_US_States_Regions_Health.csv")

#Create a new data frame creating a relation between Measures and States
measureStateRelation <- aggregate(list(state_and_region$ActualPopulation,
                                       state_and_region$PopulationCount),
                                  by=list(Measures = state_and_region$MeasureId, State =
                                            state_and_region$StateDesc), FUN=sum)

#Rename column name to a meaningful name
colnames(measureStateRelation)[3] <- "Affected_Population"
colnames(measureStateRelation)[4] <- "Total_Population"

measureStateRelation$Percent_Affected_Population <-
  (measureStateRelation$Affected_Population /
     measureStateRelation$Total_Population * 100)

#Create a data frame for High BP Health Outcome
measureBPHigh <- subset(measureStateRelation, Measures == 'BPHIGH')

#Removed unwanted "USA" column
```

```r
measureBPHigh <- measureBPHigh[-45,]
measureBPHigh <- measureBPHigh[,-1]

#Convert column and rownames to lower case for plotting on US Map
colnames(measureBPHigh)[1] <- "state"
levels(measureBPHigh$state) <- tolower(levels(measureBPHigh$state))

#Install required libraries
library(fiftystater)
```

```
## Warning: package 'fiftystater' was built under R version 3.4.3
```

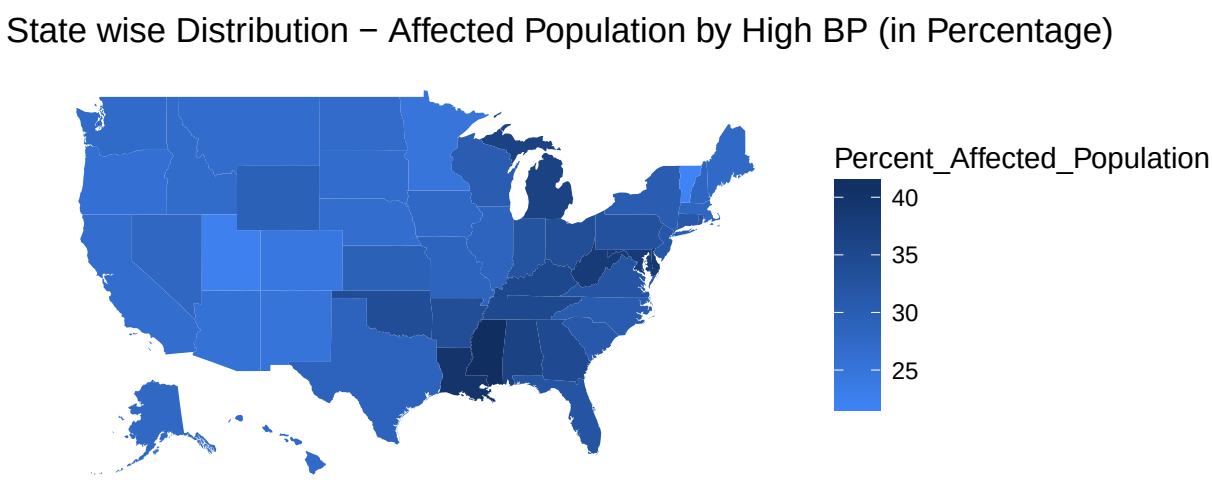
```r
library(mapproj)
```

```
## Warning: package 'mapproj' was built under R version 3.4.2
```

```
## Loading required package: maps
```

```
## Warning: package 'maps' was built under R version 3.4.2
```

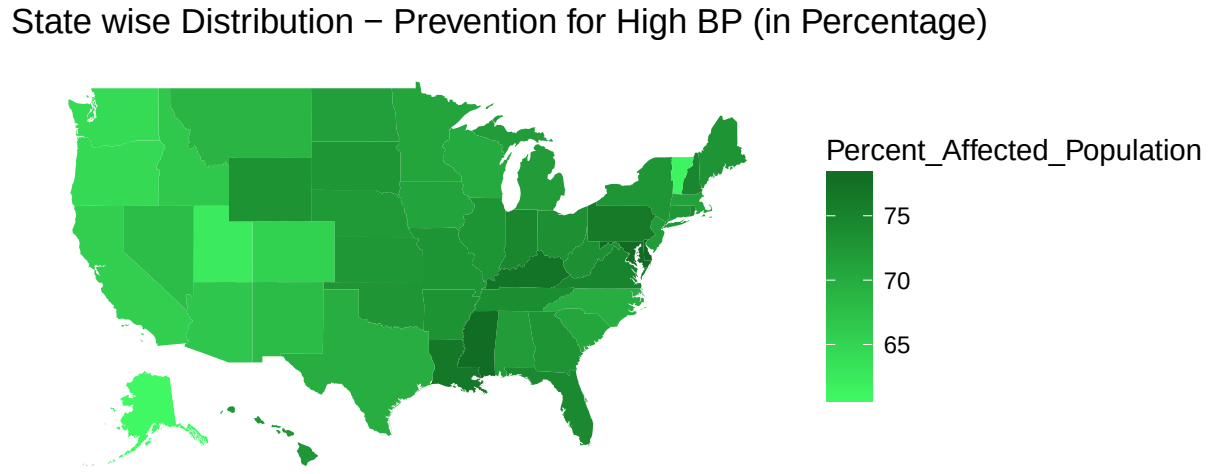
```r
data("fifty_states")

#Create a plot to show State wise Distribution of High BP in the USA
p <- ggplot(measureBPHigh, aes(map_id = state)) +
  geom_map(aes(fill=Percent_Affected_Population),
           map = fifty_states) +
  expand_limits(x = fifty_states$long, y = fifty_states$lat) +
  coord_map() + scale_x_continuous(breaks=NULL) +
  scale_y_continuous(breaks=NULL) +
  labs(x="", y="") +
  theme(panel.background = element_blank())

#Provide a title to the R Plot
p + scale_fill_gradient(low="#3f84f5", high = "#102e60",
                        space = "Lab",
                        guide = "colourbar") +
  ggtitle("State wise Distribution - Affected Population by High BP (in Percentage)") +
  theme(plot.title = element_text(hjust = 0.1))
```




The above plot shows a visualization of the High BP measure (Health Outcome) in all the states in the USA. It can be seen that a state like California has population affected by High BP

```
#Rename column to meaningful name in the main data frame
colnames(measureStateRelation)[2] <- "state"

#Create a data frame for BP Prevention Measure
measureBPMed <- subset(measureStateRelation, Measures == 'BPMED')
measureBPMed <- measureBPMed[,-1]
measureBPMed <- measureBPMed[-45,]

#Change the state names to lower case for mapping on US map
levels(measureBPMed$state) <- tolower(levels(measureBPMed$state))

#Create a R Plot for State Wise Distribution for BP Prevention
vizMeasureBPMed <- ggplot(measureBPMed, aes(map_id = state)) +
  geom_map(aes(fill=Percent_Affected_Population),
           map = fifty_states) +
  expand_limits(x = fifty_states$long, y = fifty_states$lat) +
  coord_map() + scale_x_continuous(breaks=NULL) +
  scale_y_continuous(breaks=NULL) + labs(x="", y="") +
  theme(panel.background = element_blank())

#Provide a Title and colour to the R Plot
vizMeasureBPMed + scale_fill_gradient(low="#3ff863", high = "#106921",
                                      space = "Lab",
                                      guide = "colourbar") +
```

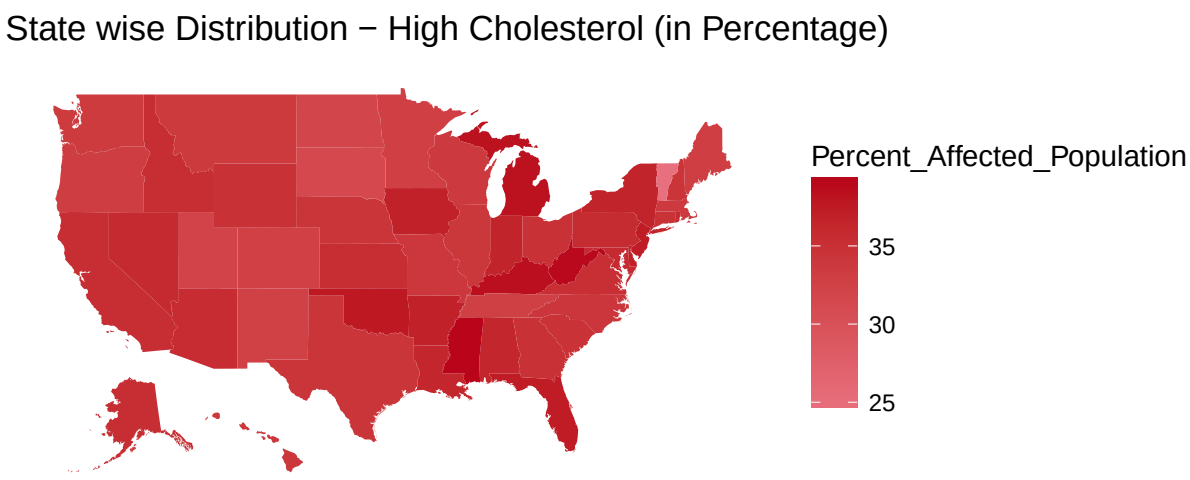
```
  ggtitle("State wise Distribution - Prevention for High BP (in Percentage)") +
  theme(plot.title = element_text(hjust = 0.1))
```




The above plot shows a visualization of a Preventive measure of Medium BP in all the states in the USA. This provides an insight, in comparison with the Health Outcomes of High BP in the previous plot, on how the Preventive measure was carried out

```
#Create a data frame for High Cholesterol Health Outcome in states of USA
measureHighCholesterol <- subset(measureStateRelation, Measures == 'HIGHCHOL')
measureHighCholesterol <- measureHighCholesterol[,-1]
measureHighCholesterol <- measureHighCholesterol[-45,]

#Change rownames to lower case for proper mapping
levels(measureHighCholesterol$state) <- tolower(levels(measureHighCholesterol$state))

#Create a R Plot for State wise distribution of High Cholesterol Health Outcome in the USA
vizMeasureHighCholesterol <- ggplot(measureHighCholesterol, aes(map_id = state)) +
  geom_map(aes(fill=Percent_Affected_Population),
           map = fifty_states) +
  expand_limits(x = fifty_states$long, y = fifty_states$lat) +
  coord_map() + scale_x_continuous(breaks=NULL) +
  scale_y_continuous(breaks=NULL) + labs(x="", y="") +
  theme(panel.background = element_blank())

#Provide a title and colour to the R Plot
vizMeasureHighCholesterol +  scale_fill_gradient(low="#e86f7e", high = "#b9041a",
```

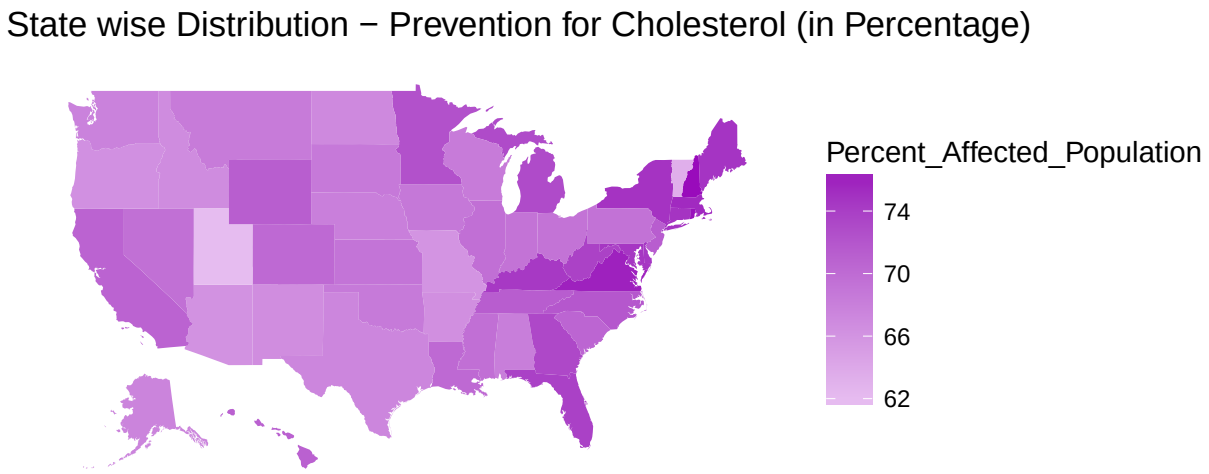
```
                                                space = "Lab",
                                                guide = "colourbar") +
  ggtitle("State wise Distribution - High Cholesterol (in Percentage)") +
  theme(plot.title = element_text(hjust = 0.1))
```




The above plot shows a visualization of the High Cholesterol measure (Health Outcome) in all the states in the USA. It can be seen that a state like California has population affected by High Cholesterol

```
#Create a data frame for Cholesterol Prevention Measure in USA
measureCholesterolScreen <- subset(measureStateRelation, Measures == 'CHOLSCREEN')
measureCholesterolScreen <- measureCholesterolScreen[,-1]
measureCholesterolScreen <- measureCholesterolScreen[-45,]

#Change rownames to lower case for proper mapping
levels(measureCholesterolScreen$state) <- tolower(levels(measureCholesterolScreen$state))

#Create R Plot for State wise distribution of Cholesterol Prevention measure
vizHighCholesterolPrevention <- ggplot(measureCholesterolScreen, aes(map_id = state)) +
  geom_map(aes(fill=Percent_Affected_Population),
           map = fifty_states) +
  expand_limits(x = fifty_states$long, y = fifty_states$lat) +
  coord_map() + scale_x_continuous(breaks=NULL) +
  scale_y_continuous(breaks=NULL) + labs(x="", y="") +
  theme(panel.background = element_blank())

#Provide a title and colour to the R Plot
```

```
vizHighCholesterolPrevention + scale_fill_gradient(low="#e6bcf0", high = "#990bbb",
                                                   space = "Lab",
                                                   guide = "colourbar") +
  ggtitle("State wise Distribution - Prevention for Cholesterol (in Percentage)") +
  theme(plot.title = element_text(hjust = 0.1))
```




Analysis from the R plots:

The above plot shows a visualization of the Cholesterol Screening (Preventive Measure) in all the states in the USA.

The dataset has only one quantitative value by which different states or regions of the USA could be compared. Hence the opportunity to use data and create different kinds of visualization was limited.

However, the current visualization created through R Plots provides an insight on how the Health Outcomes and Preventive Measures are distributed across various Regions in the USA and provides a concentration of each Measure in all the states of USA

The healthy outcomes and prevention category is seen high in Pacific Region with a low in New England region. Over all the prevention category is high in all the regions over the healthoutcomes.

The High Blood Pressure percentages are high in states MS, WV,MI with the least in UT state.

The prevention for high blood pressure are highly visible in the states MS,KY followed by MD and DE. On the contrary the low measures for prevention of high blood pressure were seen in states AK(Pacific), UT, WA.

On close observation we can notice that the cases of High Cholestrol is high among all other

categories in the US. There are many states that are in the high ranking range in this category, the AK, MS, FL, KY, WV,MI has the high percentage/concentration of High cholestrol cases. The least is seenin VT(New England).

The cholestrol prevention measures were seen high in various regions/states. On visual analysis we can see that the prevention measures are high in states NorthEast(VA), NewEngland(NH), Middle Atlantic (NY),East North Central (MI),MidWest(MN) , Florida. The prevention measures for high cholestrol are low in UT.

Over all we can see the prevention meaures are less in High cholestrol and High BP category in UT, and it is the state which is having the least percentages cases of these health issues too.

There are good preventions already taken for the regions from highest health outcome regions like Pacific to the least healthy outcome region like NewEngland.But there is a need for some of the specific countries need special prevention measures for improving the healthy measures percentages. Like AK, MS, FL, KY, WV for high cholestrol and for states like WV,MI for high blood pressure.