R Plot

Team Combat

11/8/2017

R Plot - Documentation

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.

```r
#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)) +
  labs(fill="Measurement Category")
```
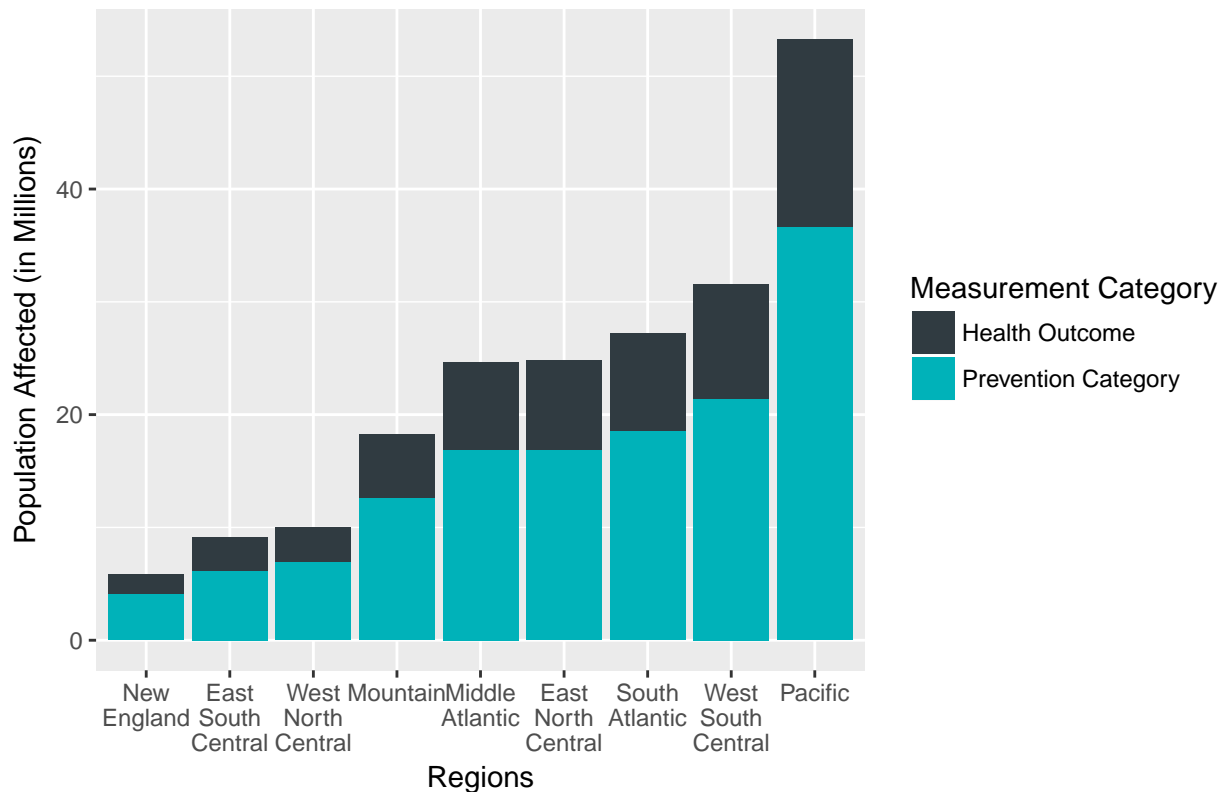
Box Plot of Population Affected and Regions in the USA

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 5 Regions in the USA

```r
#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
```

```
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)
library(mapproj)
```

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

```
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="#86b2f9", high = "#002266",
                        space = "Lab",
                        guide = "colourbar") +
  ggtitle("State wise Distribution - Affected Population by High BP (in Percentage)") +
  theme(plot.title = element_text(hjust = 0.1)) +
  labs(fill="Population Percentage (%)")
```

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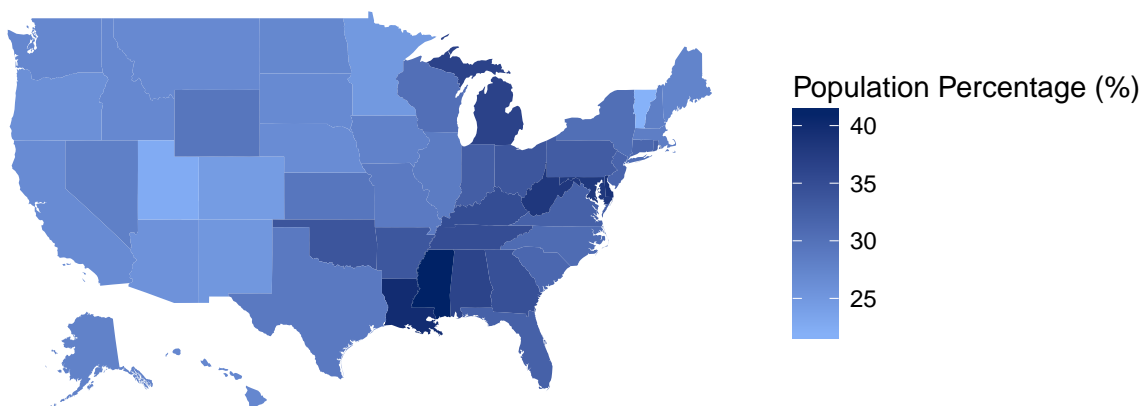
The above plot shows a visualization of the High BP measure (Health Outcome) in all the states in the USA. It can be seen that a state like California has population affected by High BP

```
#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="#9dfbb0", high = "#006622",
                                      space = "Lab",
                                      guide = "colourbar") +
  ggtitle("State wise Distribution - Prevention for High BP (in Percentage)") +
  theme(plot.title = element_text(hjust = 0.1)) +
  labs(fill="Population Percentage (%)")
```
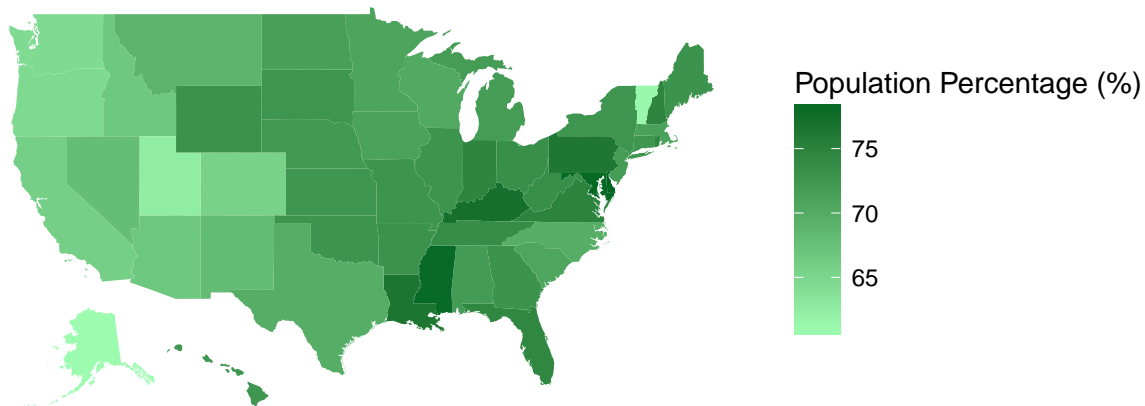
State wise Distribution – Prevention for High BP (in Percentage)

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="#f1a7b0", high = "#af0418",
                                                 space = "Lab",
                                                 guide = "colourbar") +
  ggtitle("State wise Distribution – High Cholesterol (in Percentage)") +
  theme(plot.title = element_text(hjust = 0.1)) +
  labs(fill="Population Percentage (%)")
```

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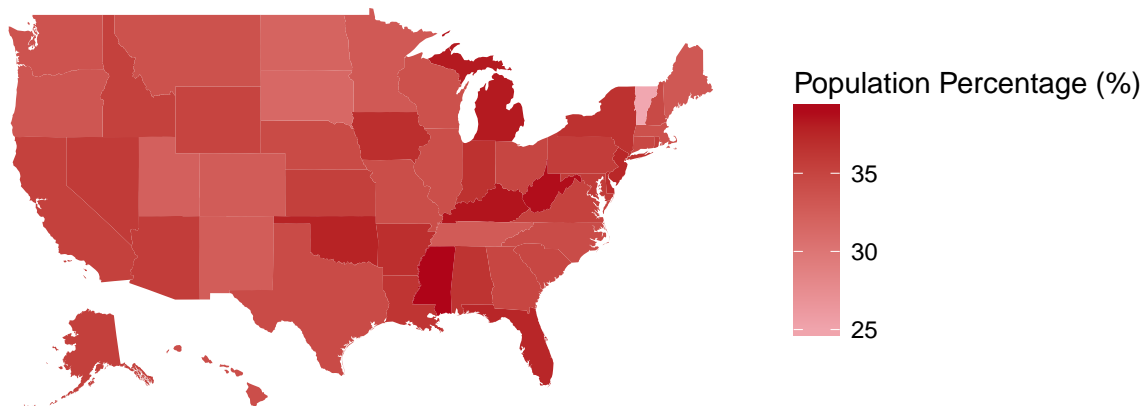
The above plot shows a visualization of the High Cholesterol measure (Health Outcome) in all the states in the USA. It can be seen that a state like California has population affected by High Cholesterol

```
#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",
```
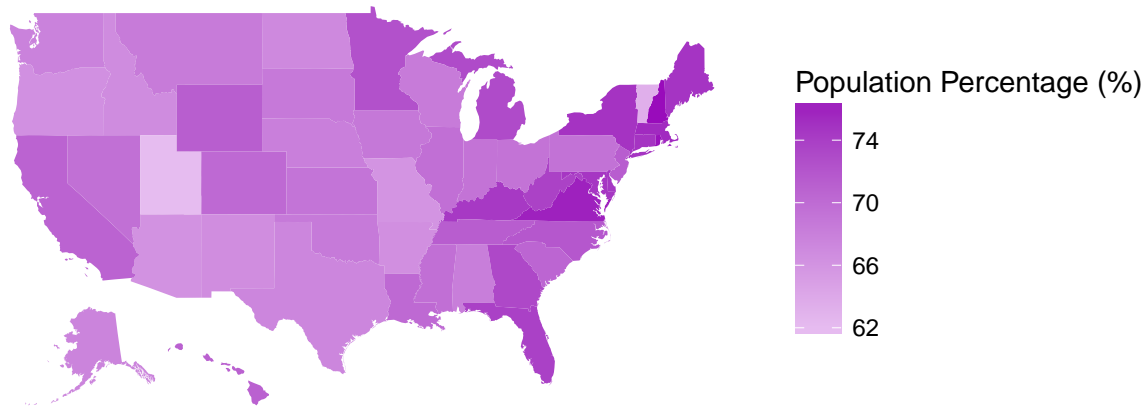
```
                                              guide = "colourbar") +
ggtitle("State wise Distribution - Prevention for Cholesterol (in Percentage)") +
theme(plot.title = element_text(hjust = 0.1)) +
labs(fill="Population Percentage (%)")
```

State wise Distribution – Prevention for Cholesterol (in Percentage)

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