Data-Visulizations.R

Data Visualization Using R: Problem Statement

Population Data for MIDWEST

2021-04-26

library(readxl)  
midwest <- read\_excel("~/New folder/Data Visulization/midwest.xlsx")  
View(midwest)  
  
library(ggplot2)

library(treemapify)

library(tidyr)  
library(dplyr)

library(ggalt)

library(scatterplot3d)

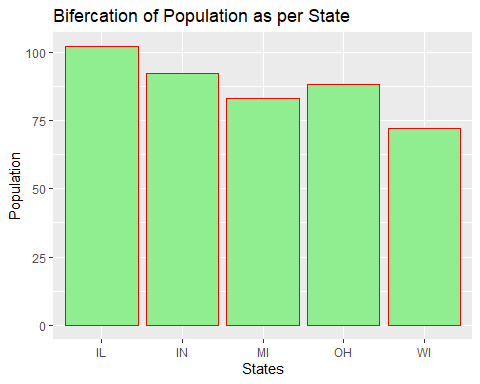
library(factoextra)

library(superheat)

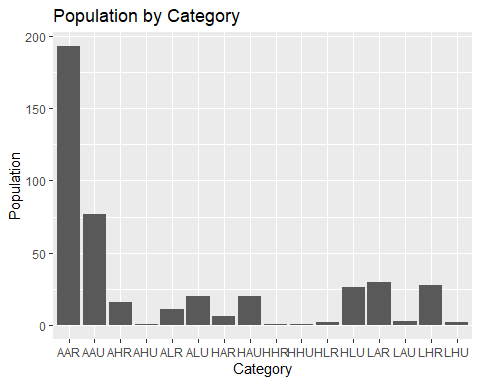
##   
## Attaching all package

## The following object is masked \_by\_ '.GlobalEnv':  
##   
## midwest

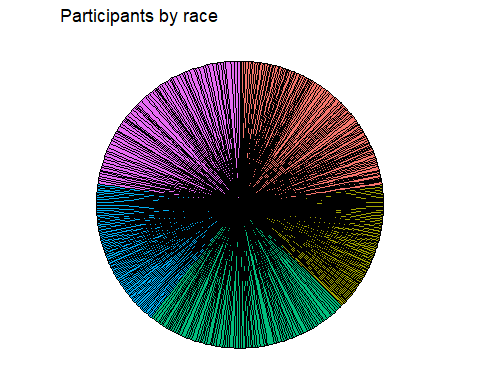
# plot the distribution of Population as per State  
ggplot(midwest, aes(x = state)) +   
 geom\_bar(fill= "lightgreen",color="red")+ labs(x="States",y="Population",  
 title = "Bifercation of Population as per State")



#Seems that IL state has maximum number of Population  
  
# basic bar chart with overlapping labels  
ggplot(midwest, aes(x = category)) +   
 geom\_bar() +  
 labs(x = "Category",  
 y = "Population",  
 title = "Population by Category")



#Seems that AAR Catefory has maximum number of population  
  
  
ggplot(midwest,   
 aes(x = "",   
 y = area,   
 fill = state)) +  
 geom\_bar(width = 1,   
 stat = "identity",   
 color = "black") +  
 coord\_polar("y",   
 start = 0,   
 direction = -1) +  
 theme\_void()+ theme(legend.position = "FALSE") + labs(title = "Participants by race")

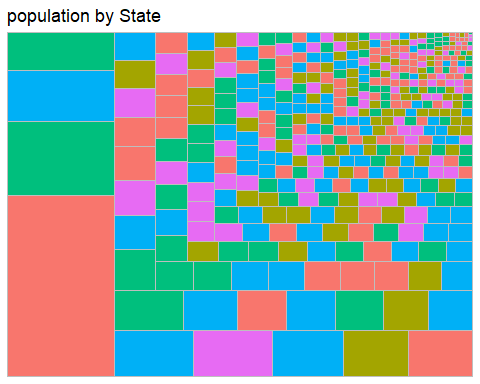


#it’s very Difficult to understant hence we will use Treemapify

library(treemapify)

## Warning: package 'treemapify' was built under R version 4.0.5

# create a treemap of Midwest  
  
  
ggplot(midwest,   
 aes(fill = state,   
 area = poptotal)) +  
 geom\_treemap() +   
 labs(title = "population by State")+theme(legend.position = "none")

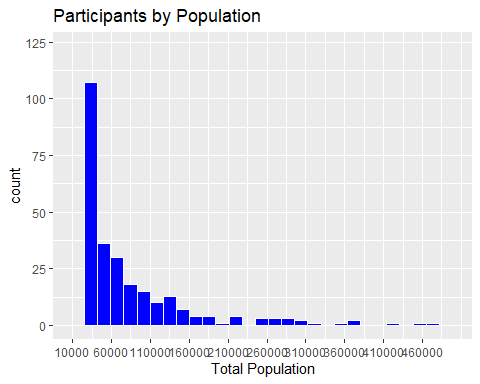


##now as per the colours we can identify that population is high or low at which state  
  
library(ggplot2)  
  
# plot the Population distribution using a histogram  
ggplot(midwest, aes(x = poptotal)) +  
 geom\_histogram(fill="blue",color="white") +scale\_x\_continuous(breaks = seq(10000, 500000, 50000),   
 limits=c(10000, 500000)) +  
 labs(title = "Participants by Population",  
 x = "Total Population")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

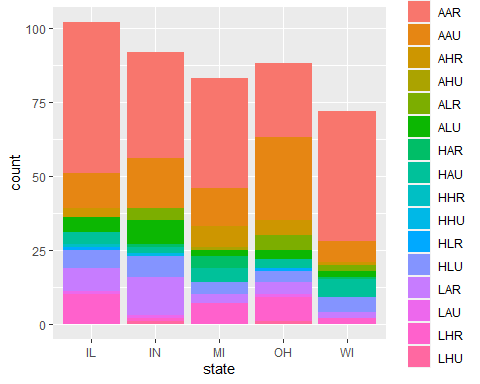
## Warning: Removed 47 rows containing non-finite values (stat\_bin).

## Warning: Removed 2 rows containing missing values (geom\_bar).

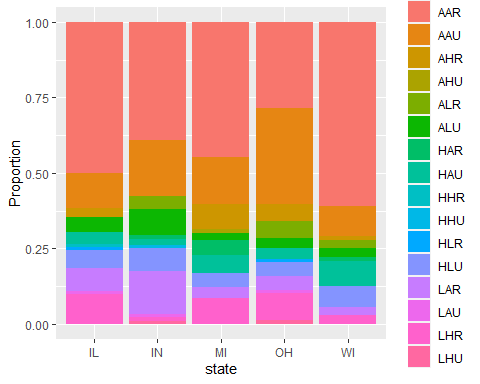


#this is Normal distribution of population

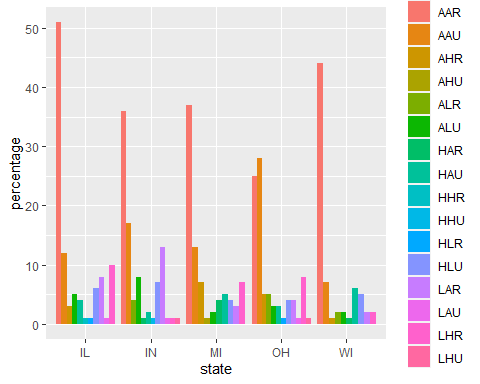
library(ggplot2)  
  
# stacked bar chart  
ggplot(midwest,   
 aes(x = state,   
 fill = category)) +   
 geom\_bar(position = "stack")



##Seems that AAR Category is mostly present is each state,LAR highly present OH state.  
  
# bar plot, with each bar representing 100%  
ggplot(midwest,   
 aes(x = state,   
 fill = category)) +   
 geom\_bar(position = "fill") +  
 labs(y = "Proportion")

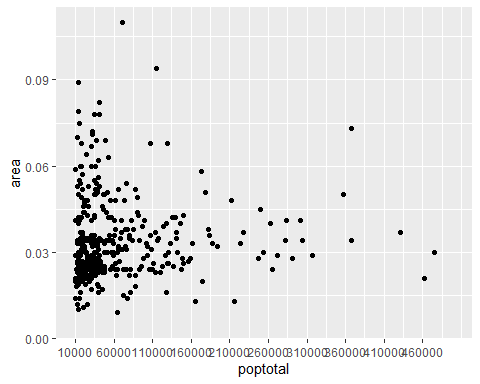


# stacked bar chart made easy to understand the category as per state  
#lets try in dodge position  
ggplot(midwest,   
 aes(x = state,   
 fill = category)) +   
 geom\_bar(position = "dodge")+labs(y = "percentage")



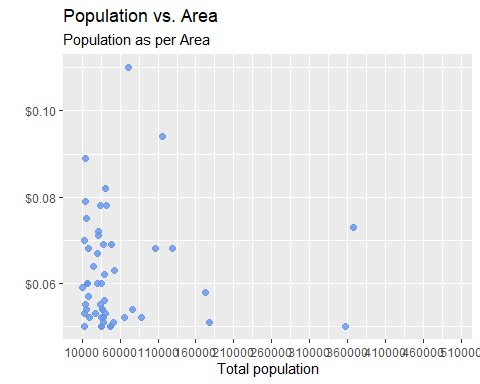
##More clearly visible as per percentage that which state has more or less no of Category  
  
  
  
  
library(ggplot2)  
  
# simple scatterplot  
ggplot(midwest,   
 aes(x = poptotal,   
 y = area)) +scale\_x\_continuous(breaks = seq(10000, 500000, 50000),   
 limits=c(10000, 500000))+  
 geom\_point()

## Warning: Removed 47 rows containing missing values (geom\_point).



# enhanced scatter plot  
ggplot(midwest,   
 aes(x = poptotal,   
 y = area)) +  
 geom\_point(color="cornflowerblue",   
 size = 2,   
 alpha=.8) +  
 scale\_y\_continuous(label = scales::dollar,   
 limits = c(0.05, 0.11)) +  
 scale\_x\_continuous(breaks = seq(10000, 1000000, 50000),   
 limits=c(10000, 500000)) +   
 labs(x = "Total population",  
 y = "",  
 title = "Population vs. Area",  
 subtitle = "Population as per Area")

## Warning: Removed 385 rows containing missing values (geom\_point).

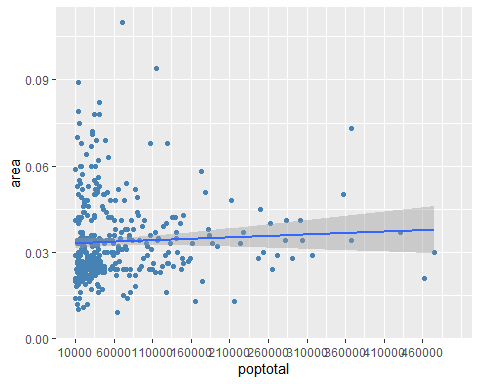


##It is visible that we have very less area as per population  
ggplot(midwest,  
 aes(x = poptotal,   
 y = area)) +scale\_x\_continuous(breaks = seq(10000, 500000, 50000),   
 limits=c(10000, 500000))+  
 geom\_point(color= "steelblue") +  
 geom\_smooth(method = "lm")

## `geom\_smooth()` using formula 'y ~ x'

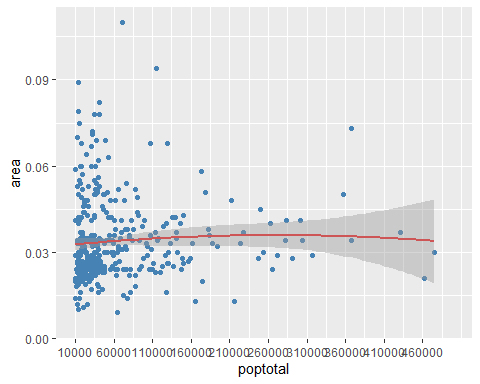
## Warning: Removed 47 rows containing non-finite values (stat\_smooth).

## Warning: Removed 47 rows containing missing values (geom\_point).



#Clearly, Population increases However, there seems to be a dip at the right end - Increase of a population, result lower area  
#A straight line does not capture this non-linear effect. A line with a bend will fit better here.  
# scatterplot with quadratic line of best fit  
ggplot(midwest,   
 aes(x = poptotal,   
 y = area)) +scale\_x\_continuous(breaks = seq(10000, 500000, 50000),   
 limits=c(10000, 500000))+  
 geom\_point(color= "steelblue") +  
 geom\_smooth(method = "lm",   
 formula = y ~ poly(x, 2),   
 color = "indianred3")

## Warning: Removed 47 rows containing non-finite values (stat\_smooth).  
  
## Warning: Removed 47 rows containing missing values (geom\_point).



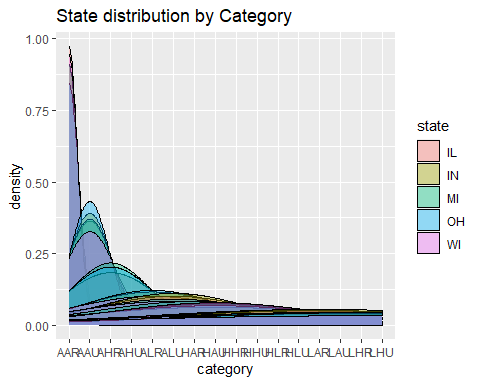
#Finally, a smoothed nonparametric fit line can often provide a good picture of the relationship.   
  
  
  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

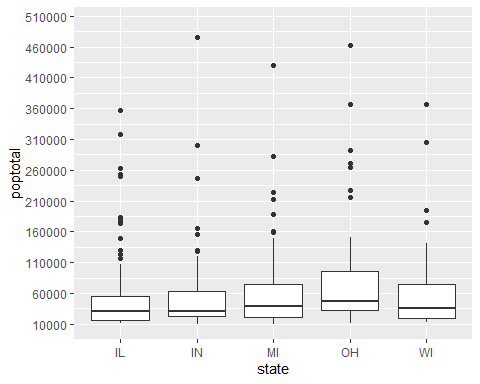
## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

# plot the distribution of midwest   
ggplot(midwest,   
 aes(x = category,   
 fill = state)) +  
 geom\_density(alpha = 0.4) +  
 labs(title = "State distribution by Category")



# plot the distribution of Total population by state using boxplots  
ggplot(midwest,   
 aes(x = state,   
 y = poptotal)) +  
 geom\_boxplot() +  
 scale\_y\_continuous(breaks = seq(10000, 1000000, 50000),   
 limits=c(10000, 500000))

## Warning: Removed 47 rows containing non-finite values (stat\_boxplot).



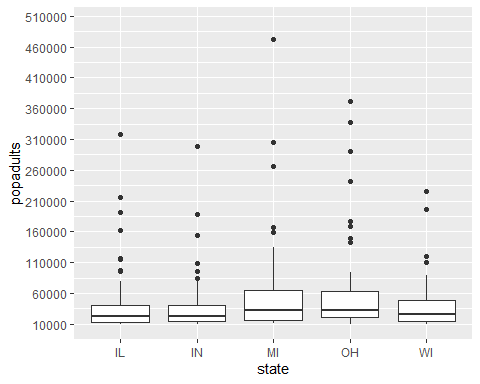
# Distribution of the total population as per state~ STATE MI has Maximum population.

labs(title = "population by state")

## $title  
## [1] "population by state"  
##   
## attr(,"class")  
## [1] "labels"

# plot the distribution of Adult population by state using boxplots  
 ggplot(midwest,   
 aes(x = state,   
 y = popadults)) +  
 geom\_boxplot() +  
 scale\_y\_continuous(breaks = seq(10000, 1000000, 50000),   
 limits=c(10000, 500000))

## Warning: Removed 80 rows containing non-finite values (stat\_boxplot).

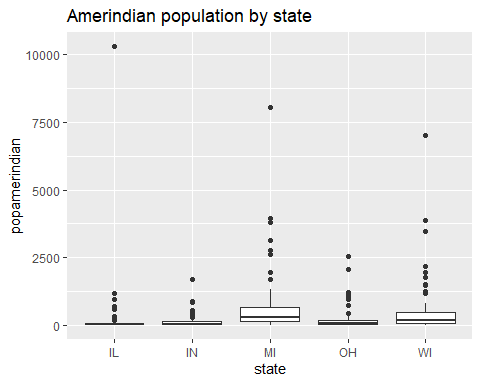


# Distribution of the Adult population as per state, again STATE-MI has max.

labs(title = "Adult population by state")

## $title  
## [1] "Adult population by state"  
##   
## attr(,"class")  
## [1] "labels"

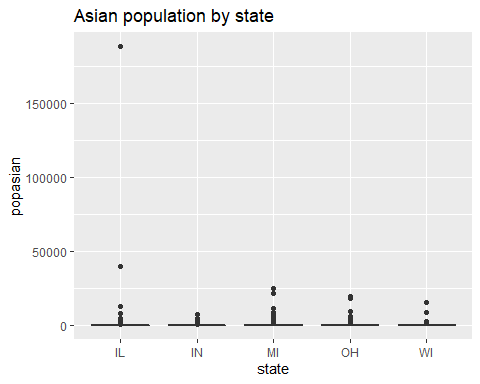
# plot the distribution of Amerindian population by state using boxplots  
 ggplot(midwest,   
 aes(x = state,   
 y = popamerindian)) +  
 geom\_boxplot() +  
 labs(title = "Amerindian population by state")



# Distribution of the American population as per state,State- IL, IN & OH has very less American population.

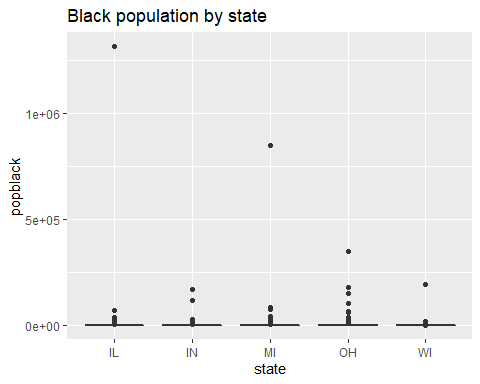
# plot the distribution of Asian population by state using boxplots  
 ggplot(midwest,   
 aes(x = state,   
 y = popasian)) +  
 geom\_boxplot() +  
 labs(title = "Asian population by state")

# Distribution of the total population as per state



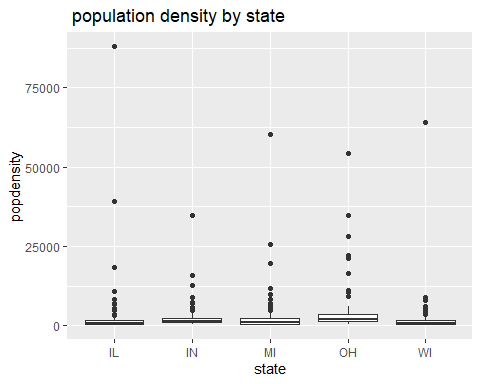
# Distribution of the Asian population as per state, overall it’s too low

# plot the distribution of Black population by state using boxplots  
 ggplot(midwest,   
 aes(x = state,   
 y = popblack)) +  
 geom\_boxplot() +  
 labs(title = "Black population by state")



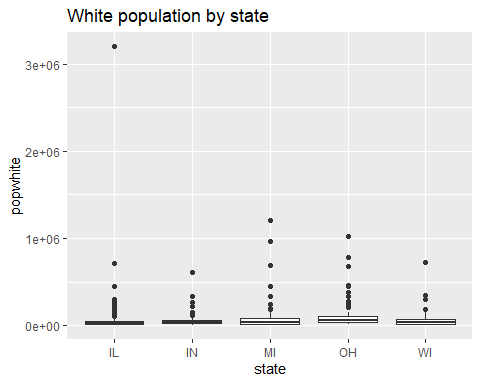
# Distribution of the Black population as per state, overall it’s too low

# plot the distribution of population density by state using boxplots  
 ggplot(midwest,   
 aes(x = state,   
 y = popdensity)) +  
 geom\_boxplot() +  
 labs(title = " population density by state")



# Distribution of the Density population as per state, overall it’s too low, but seems like state MI has some of populations

# plot the distribution of White population by state using boxplots  
 ggplot(midwest,   
 aes(x = state,   
 y = popwhite)) +  
 geom\_boxplot() +  
 labs(title = "White population by state")

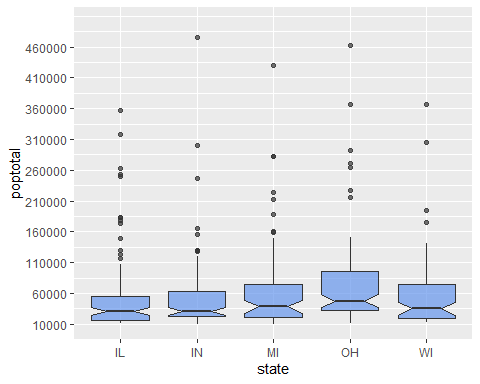


# Distribution of the Asian population as per state, overall it’s too low

But seems like State MI, OH & WI has some population.

# Lets try some more attractive   
 ggplot(midwest, aes(x = state,   
 y = poptotal)) +  
 geom\_boxplot(notch = TRUE,   
 fill = "cornflowerblue",   
 alpha = .7) +scale\_y\_continuous(breaks = seq(10000, 500000, 50000),   
 limits=c(10000, 500000))

## Warning: Removed 47 rows containing non-finite values (stat\_boxplot).

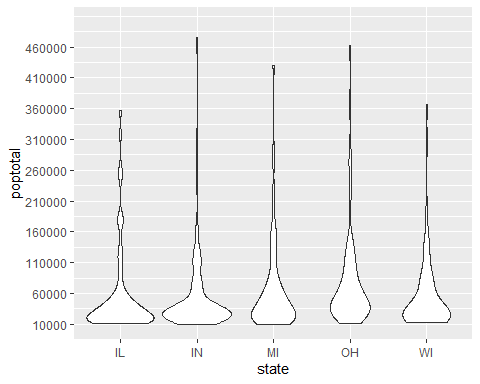


labs(title = "population by state")

## $title  
## [1] "population by state"  
##   
## attr(,"class")  
## [1] "labels"

# by state using violin plots  
 ggplot(midwest,   
 aes(x = state,  
 y = poptotal)) +  
 geom\_violin() +scale\_y\_continuous(breaks = seq(10000, 500000, 50000),   
 limits=c(10000, 500000))

## Warning: Removed 47 rows containing non-finite values (stat\_ydensity).



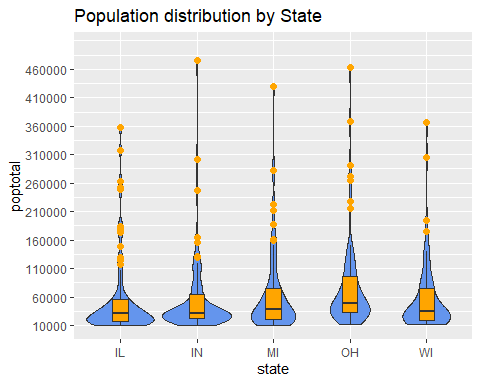
labs(title = "population by state")

## $title  
## [1] "population by state"  
##   
## attr(,"class")  
## [1] "labels"

# plot the distribution using violin and boxplots  
 ggplot(midwest,   
 aes(x = state,   
 y = poptotal)) +  
 geom\_violin(fill = "cornflowerblue") +  
 geom\_boxplot(width = .2,   
 fill = "orange",  
 outlier.color = "orange",  
 outlier.size = 2) + scale\_y\_continuous(breaks = seq(10000, 500000, 50000),   
 limits=c(10000, 500000))+  
 labs(title = "Population distribution by State")

## Warning: Removed 47 rows containing non-finite values (stat\_ydensity).

## Warning: Removed 47 rows containing non-finite values (stat\_boxplot).



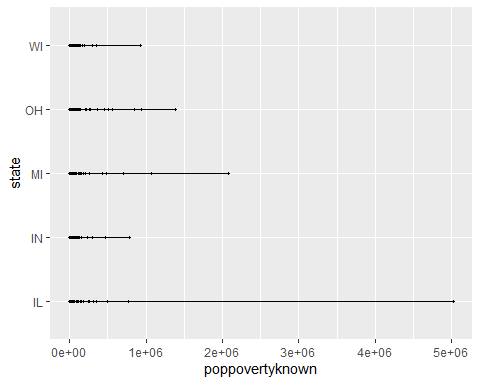
#using this plots to understand better pitcher  
   
   
   
   
   
 library(tidyr)  
 library(dplyr)  
 library(ggalt)

## Warning: package 'ggalt' was built under R version 4.0.5

## Registered S3 methods overwritten by 'ggalt':  
## method from   
## grid.draw.absoluteGrob ggplot2  
## grobHeight.absoluteGrob ggplot2  
## grobWidth.absoluteGrob ggplot2  
## grobX.absoluteGrob ggplot2  
## grobY.absoluteGrob ggplot2

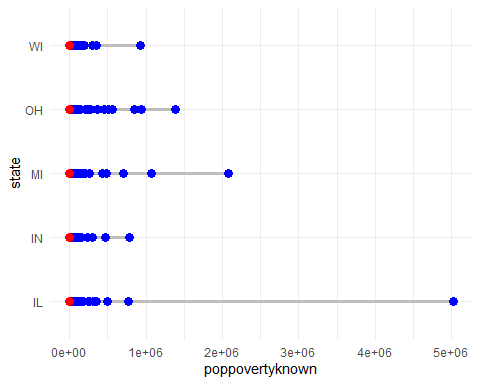
# create dumbbell plot of Number of poor people as per state

ggplot(midwest, aes(y = state,  
 x = poppovertyknown,  
 xend = 2007)) +   
 geom\_dumbbell()

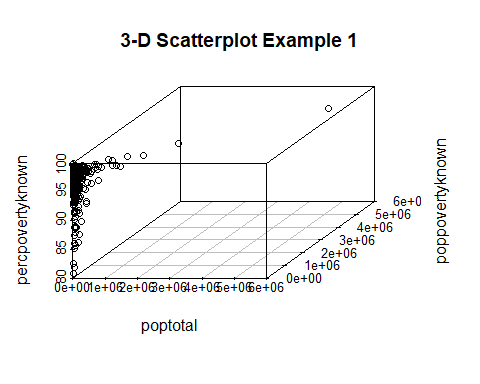


# Dumbbell plot is helping to understand, poverty over the state

# create dumbbell plot  
 ggplot(midwest,   
 aes(y = state,  
 x = poppovertyknown,  
 xend = 2007)) +   
 geom\_dumbbell(size = 1.2,  
 size\_x = 3,   
 size\_xend = 3,  
 colour = "grey",   
 colour\_x = "blue",   
 colour\_xend = "red") +  
 theme\_minimal()



##Clearly,It is easier to discern patterns here,number of poor people are seen as per blue dots   
   
 # basic 3-D scatterplot  
 library(scatterplot3d)  
 with(midwest, {  
 scatterplot3d(x = poptotal,  
 y = poppovertyknown,   
 z = percpovertyknown,  
 main="3-D Scatterplot Example 1")})



##Clearly,Increase in Population has increased poverty   
   
 library(readxl)  
 Tral <- read\_excel("Tral.xlsx")  
 View(Tral)  
   
 # create a biplot  
   
 # fit a principal components model  
 fit <- prcomp(x = Tral,  
 center = TRUE,   
 scale = TRUE)  
   
 # plot the results  
 library(factoextra)

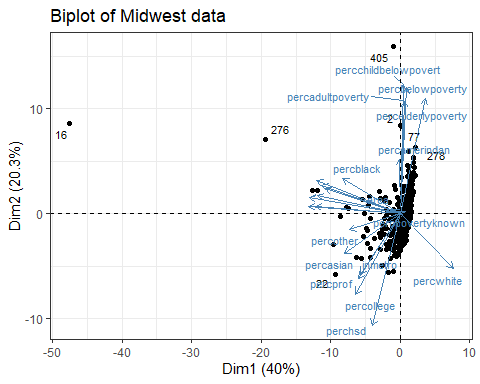
## Warning: package 'factoextra' was built under R version 4.0.5

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

fviz\_pca(fit,   
 repel = TRUE,   
 labelsize = 3) +   
 theme\_bw() +  
 labs(title = "Biplot of Midwest data")

## Warning: ggrepel: 430 unlabeled data points (too many overlaps). Consider  
## increasing max.overlaps

## Warning: ggrepel: 9 unlabeled data points (too many overlaps). Consider  
## increasing max.overlaps



##very difficult to understand

##Conclusion: This Data consist on Population of data for different regions as per state and category, I have applied many grapes and generated plot to portray results (which are given below the plots), also generated visualizations that depicts meaningful insights

Purvi Bishya

26 April 2021