Unsupervised ML HW#1 - LR

1. Obtain a dataset (preferably of substantive interest/domain expertise).

PSAData <- read.csv(file.path('/Users/lillyreich/Desktop/PSAData.csv'),header = TRUE)  
summary(PSAData)

## Year State Percent Numerator   
## Min. :2014 AK : 4 Min. : 7.00 Min. : 1.00   
## 1st Qu.:2015 AL : 4 1st Qu.:39.75 1st Qu.: 15.75   
## Median :2016 AR : 4 Median :56.00 Median : 36.50   
## Mean :2016 AZ : 4 Mean :55.03 Mean : 51.77   
## 3rd Qu.:2016 CA : 4 3rd Qu.:71.00 3rd Qu.: 64.75   
## Max. :2017 CO : 4 Max. :96.00 Max. :346.00   
## (Other):184   
## Denominator   
## Min. : 8.00   
## 1st Qu.: 32.00   
## Median : 71.00   
## Mean : 88.78   
## 3rd Qu.:113.25   
## Max. :435.00   
##

2.Choose a visual technique to illustrate your data (e.g., barplot, histogram, scatterplot).

PSAData <- read.csv(file.path('/Users/lillyreich/Desktop/PSAData.csv'),header = TRUE)  
newdata <- subset(PSAData,Year == 2014)  
mean(newdata$Percent)

## [1] 37.80769

var(newdata$Percent)

## [1] 172.276

sd(newdata$Percent)

## [1] 13.1254

newdata <- subset(PSAData, Year == 2015)  
mean(newdata$Percent)

## [1] 44.92308

var(newdata$Percent)

## [1] 218.2293

sd(newdata$Percent)

## [1] 14.77258

newdata <- subset(PSAData, Year == 2016)  
mean(newdata$Percent)

## [1] 62.51923

var(newdata$Percent)

## [1] 189.431

sd(newdata$Percent)

## [1] 13.76339

newdata <- subset(PSAData, Year == 2017)  
mean(newdata$Percent)

## [1] 74.86538

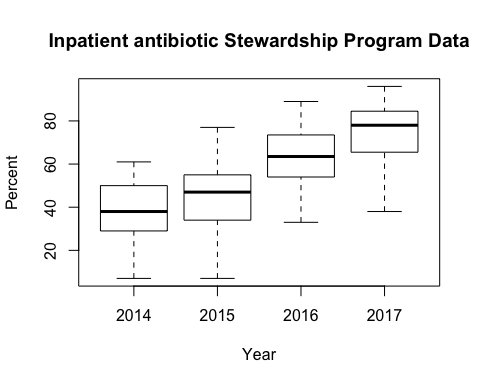
var(newdata$Percent)

## [1] 166.7854

sd(newdata$Percent)

## [1] 12.91454

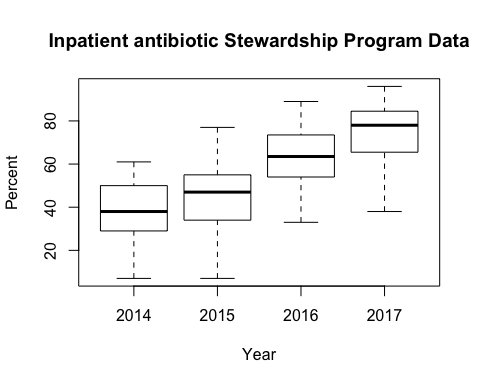
#boxplot(newdata$Percent,mpg~cyl,Year,data=PSAData)  
boxplot(Percent ~ Year, data = PSAData, main = "Inpatient antibiotic Stewardship Program Data", xlab = "Year", ylab = "Percent")



## R Markdown

1. Now generate and present the visualization and describe what you see.

boxplot(Percent ~ Year, data = PSAData, main = "Inpatient antibiotic Stewardship Program Data", xlab = "Year", ylab = "Percent")



In the Inpatient antibiotic Stewardship Program Data, the mean per year reveals a trend of the adoption of the program throughout the U.S. The variance shows there is a wide differnce in adoption practices in different states since some are higher than others.

1. Calculate the common measures of central tendency and variation, and then display your results.

print('Year 2014')

## [1] "Year 2014"

newdata <- subset(PSAData,Year == 2014)  
mean(newdata$Percent)

## [1] 37.80769

median(newdata$Percent)

## [1] 38

var(newdata$Percent)

## [1] 172.276

sd(newdata$Percent)

## [1] 13.1254

print('Year 2015')

## [1] "Year 2015"

newdata <- subset(PSAData, Year == 2015)  
mean(newdata$Percent)

## [1] 44.92308

median(newdata$Percent)

## [1] 47

var(newdata$Percent)

## [1] 218.2293

sd(newdata$Percent)

## [1] 14.77258

print('Year 2016')

## [1] "Year 2016"

newdata <- subset(PSAData, Year == 2016)  
mean(newdata$Percent)

## [1] 62.51923

median(newdata$Percent)

## [1] 63.5

var(newdata$Percent)

## [1] 189.431

sd(newdata$Percent)

## [1] 13.76339

print('Year 2017')

## [1] "Year 2017"

newdata <- subset(PSAData, Year == 2017)  
mean(newdata$Percent)

## [1] 74.86538

median(newdata$Percent)

## [1] 78

var(newdata$Percent)

## [1] 166.7854

sd(newdata$Percent)

## [1] 12.91454

##5. Describe the numeric output in substantive terms, e.g.,

##a. What do these numeric descriptions of data reveal?

The data reveals that states are increasingly adopting the stewardship program from 2014-2017. The average for 2014 was 37.80% and in 2017 was 74.86%, which is nearly double the adoption rate.

##b. Why is this important?

We are able to see that states are taking more precaution with antibacterial resistance in an inpatient hospital setting. Overall, the safety measures for antibacterial resistance in the U.S. are being adopted.

##c. What might you infer about the distribution or spread of the data? Why?

The distribution becomes negatively skewed, particularly in 2017, as the adoption rates increase. The median is 78%, which is higher than the mean.

##d. Etc.

In 2015, as indicated by the whiskers, there was a larger range of percentage in hospitals in states adopting antibacterial resistance.

##Critical Thinking

##1. Describe the different information contained in/revealed by visual versus numeric exploratory data analysis. (Hint: Think of different examples of each and then what we might be looking for when leveraging a given technique).

Numeric exploratory data analysis provides a descriptive statistical result that is a straightforward way to make comparisons across measures.

Visual exploratory data analysis will allow more information about the individual data points that were involved in the descriptive statistical analyis, whereas a descriptive statistic will summarize all of the provided data points into single scores.

If the measure is normally distributed, then a visual wouldn’t be required.  
On the other hand, for non-normally distributed measures, we could have similar numeric results (means and standard deviations, for example) for two distributions that may appear different and without visuals we would not know about their different shapes.

Numeric exploratory data analysis is sometimes more economical since we are provided with less information.

##2.Find (and include) two examples of “bad” visualizations and tell me precisely why they’re bad.

##Alien Abductions:Figure 1 

The visualization is considered poor for reasons, such as: 1) Clearly, an x and y axis are not provided with labels. 2) The values do not have any meaning since the percentages on the y-axis are not labeled. The x-axis likely indicates years, but is not carefully labeled.

##Bush Tax Cuts:Figure 2

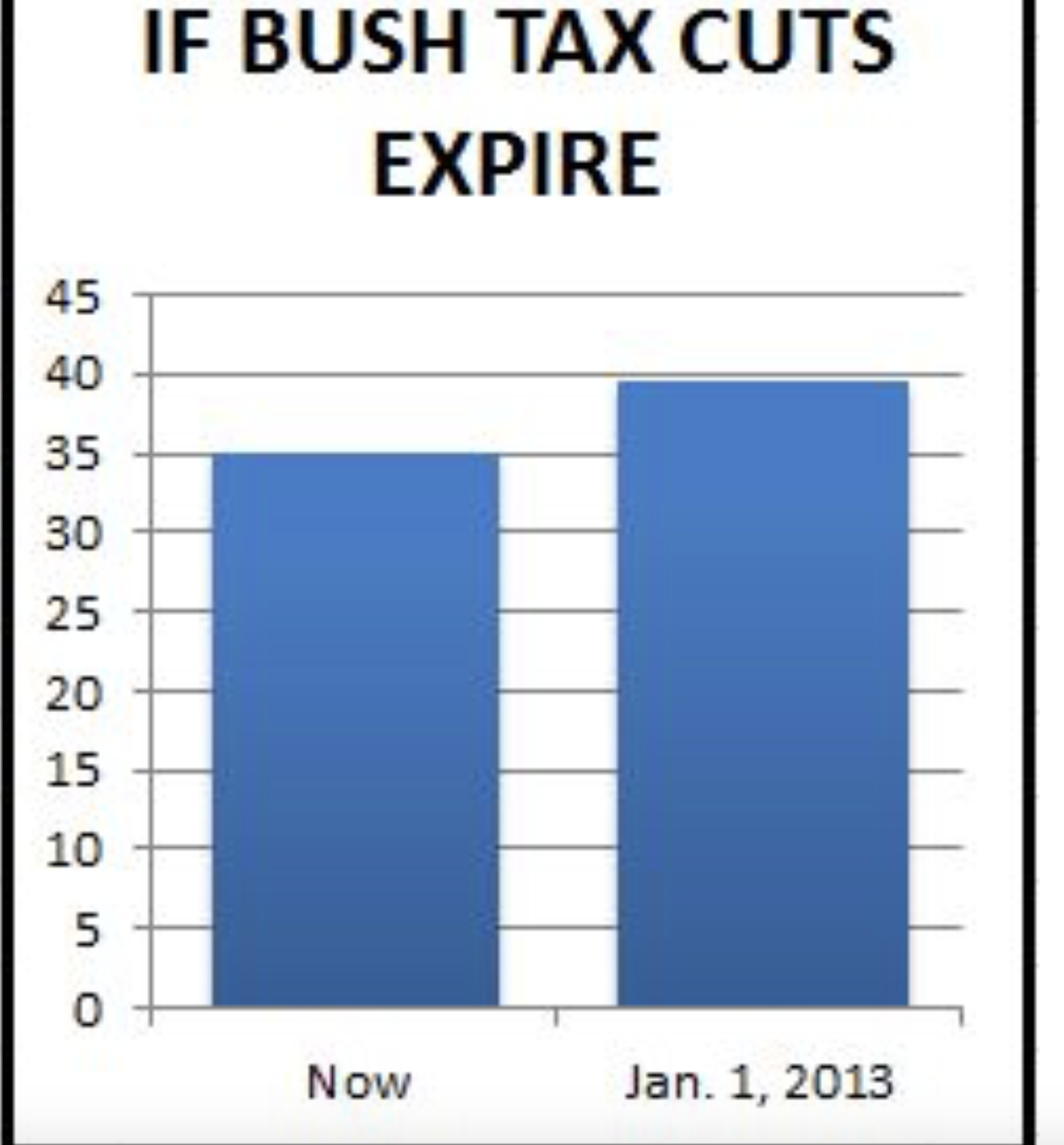


Figure 2.

The Bush Tax Cuts visualization is also poor for several reasons: Firstly, the x and y axis are not labeled; therefore, we are unable to make sound judgements of comparisons. Secondly, we do not have any description of the y-axis values. As a result, it was difficult to conclude with any meaningful information of the bar plot. More importantly, the Bush Tax Cuts visualization is unecessary and could have been represented by displaying two numeric values.

3.Find (and include) two examples of “good” visualizations and tell me precisely why they’re good.

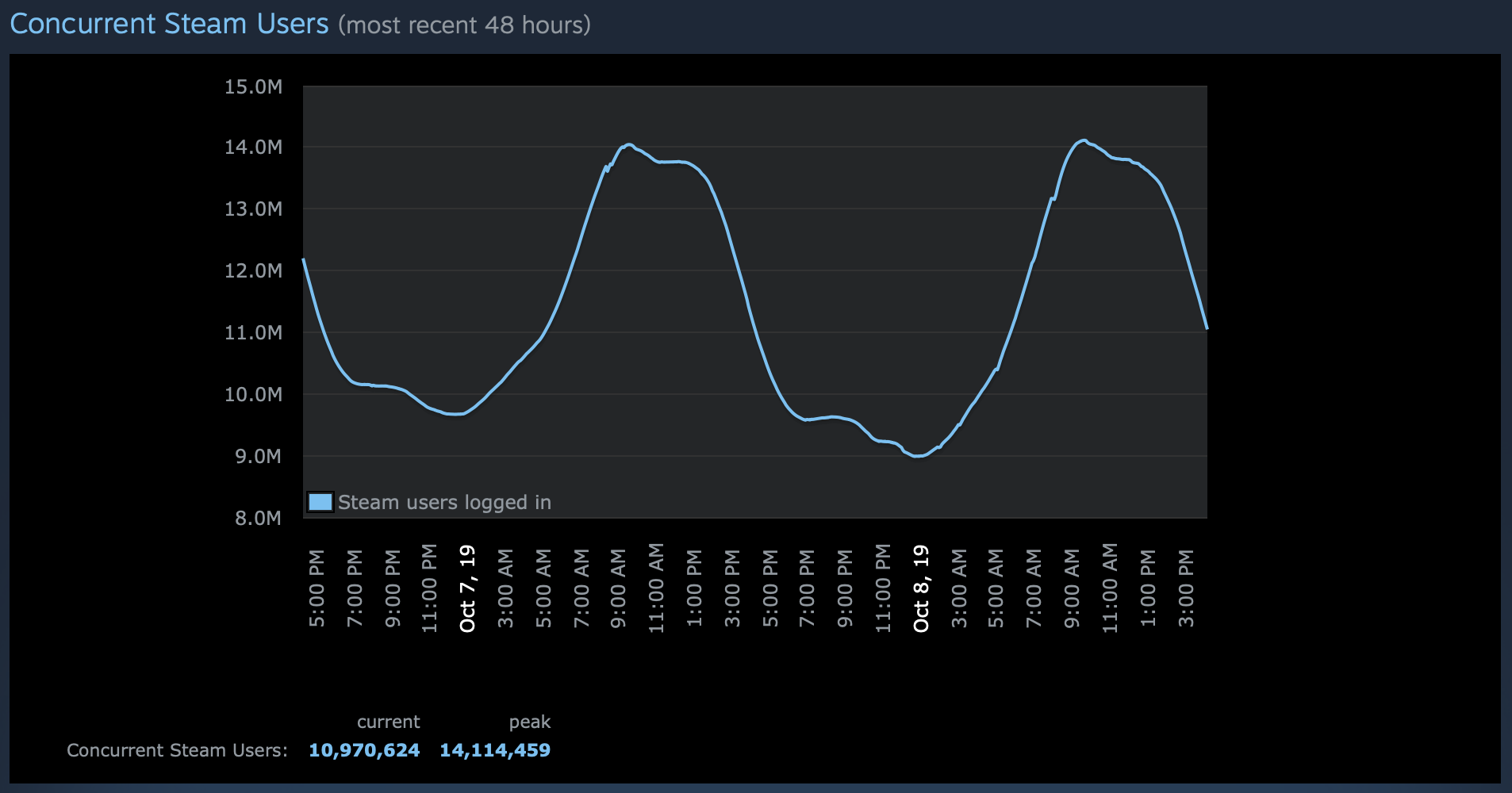


Figure 3.

The Gaming service, titled, Steam shows a clear visualization of the number of customers they gather over a 48 hour period. Firstly, the x and y-axis shows a clear relationship of times and users over a specific time period. Secondly, a legend labeled: Steam users logged in, displays clarity in the x and y-axis comparison. Thirdly, the figure displays a brief description of the current Steam users and a peak value. However, the figure is lacking a label for the axis, but the x-axis clearly displays time and the y-axis displays count. Overall, the graph clearly displays the times customers are active on Steam.

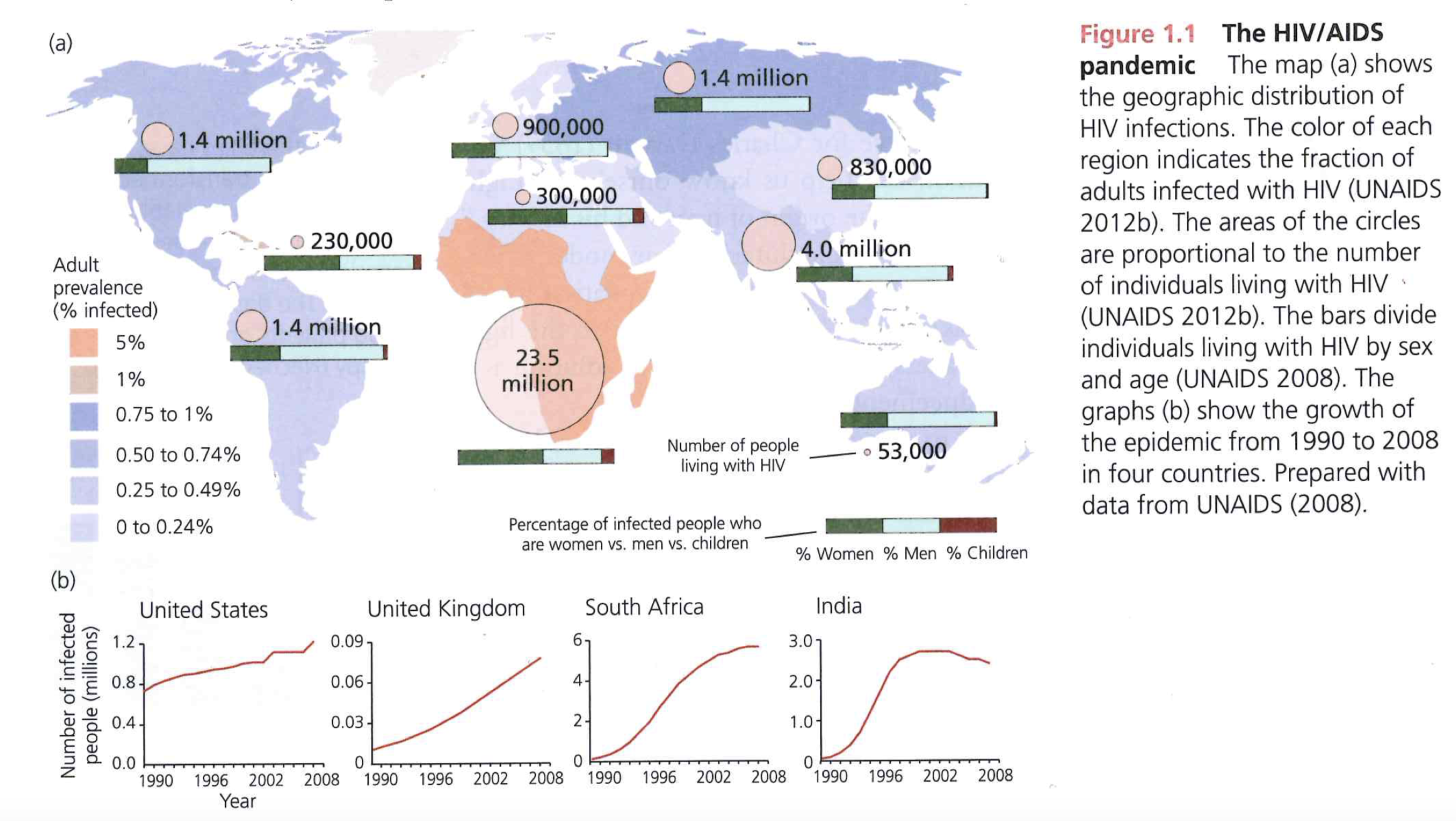


Figure 4.

The map of the HIV/Aids pandemic provides clarity of the disease spread from 1990-2008. Firstly, a color legend is provided on the left side of Figure 1.1a with a title and brief description of the value’s being represented. Secondly, Figure 1.1a provides a choropleth map of the number of individuals affected by HIV/AIDs. Thirdly, in Figure 1.1b), there are four line graphs that have clear labels of both x and y-axis. As a result, the reader is able to understand the HIV/AIDs geographic distribution in terms of prevalance rates as well as the changes over time by certain geographic regions from part 1.1b.

1. When might we use EDA and why/how does it help the research process?

If we are looking at a large dataset to formulate an initial hyothesis, we might also consider using Exploratory Data Analysis (EDA) to assist in understanding the need for more data or additional experimentation. Therefore, EDA will help us formulate more questions for the research process.

1. What did John Tukey mean by “confirmatory” versus “exploratory”? Give me an example for each.

John Tukey described confirmatory analysis as statistical hypothesis testing, which he disagreed with in pedagogical approaches for introductory statistics students. To exemplify, a confirmatory data analysis may consist of a linear regression analysis of the 2019 presidential approval rating with respect to growth of the economy.Exploratory data analysis encourages statisticians to closely explore the data and devise new hypothesis and more data collection. For example, a scatterplot that displays a shape in a relationship or contains outliers.

```