Homework #8

Univariate Distributions - Derrik Adams

## Concepts

1. Carefully explain how visualizing, transforming, and modelling interact in exploratory data analysis.
2. Identify and explain the ggplot geometric objects that appropriately depict the distributions of
   1. categorical variables
   2. numerical variables

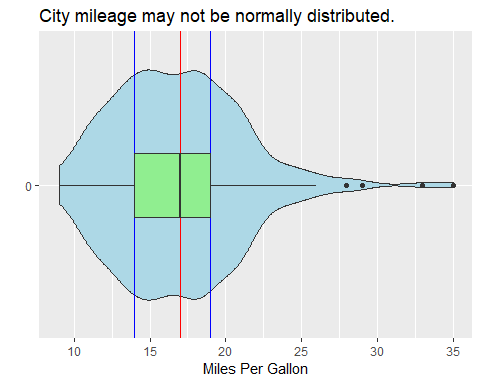
### Concept Answers

1. Visualizing your data using charts or plots allows you to see trends in the data easily, see data that’s missing or outliers, and lots else. Through visualizing your data, you can answer questions you had and find more questions based on the trends that you find. Using those new questions you can beging to transform your data and manipulate it to show other trends that are less obvious, improving your understanding of all the variables at play. As you learn more and more about your data through this, you can begin to model the data and create new models to predict situations involving those variables.
2. Geometric objects for categorical and numerical variables
   1. Categorical - To depict categorical variables, use a bar chart that shows the count or the number of observations for each category
   2. Numerical - To depict numerical variables, use a histogram to depict the count or frequency of observations that occur within specified ranges of the variable.

## Interpretation of code

Interpret the following R code:

Median <- median(mpg$cty)  
lower\_hinge <- quantile(mpg$cty, 0.25)  
upper\_hinge <- quantile(mpg$cty, 0.75)  
  
mpg %>%   
 ggplot(aes(x = factor(0), y = cty)) +  
 geom\_violin(fill = "lightblue") +  
 geom\_hline(yintercept = Median, color = "red") +  
 geom\_hline(yintercept = lower\_hinge, color = "blue") +  
 geom\_hline(yintercept = upper\_hinge, color = "blue") +  
 geom\_boxplot(fill = "lightgreen", width = 0.25) +  
 coord\_flip() +  
 labs(  
 title = "City mileage may not be normally distributed.",  
 x = "",  
 y = "Miles Per Gallon"  
 )



### Code Interpretation

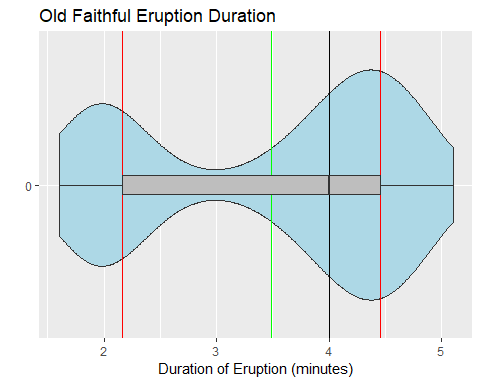
1. Define a new data variable called “Median” that is the median value of MPG for city from the mpg dataset
2. Define a new data variable called “lower\_hinge” that is the MPG value that is higher than 25% of the data and lower than 75%
3. Define a new data variable called “upper\_hing” that is the MPG value that is higher than 75% of the data and lower than 25%
4. Working on the mpg dataset
5. Create a plot where the aesthetic has all 0’s mapped to the x axis and the cty mpg is mapped to the y axis
6. Generate a violin plot that is filled light blue
7. Create a horizontal line whose y-intercept is at the Median value defined above and colored red
8. Create a horizontal line whose y-intercept is at the lower hinge value defined above and colored blue
9. Create a horizontal line whose y-intercept is at the upper hinge value defined above and colored blue
10. Create a boxplot that is filled light green with a width of .25
11. Flip the coordinate axis so that the x-axis is on the y-axis and the y-axis is on the x-axis
12. The rest of the code just labels the title of the plot and the axes

## Write code

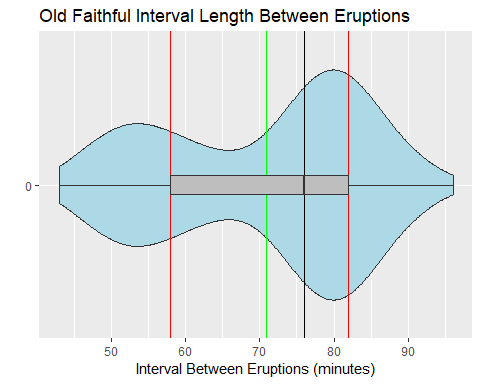
The faithful dataset includes 272 observations that report the length of time that the Old Faithful geyser in Yellowstone National Park erupts and the time interval between the eruptions. Use exploratory data analysis to investigate the distributions of each of the two variables, eruptions and waiting. Your code should generate summary statistics and graphs that give you insights into:

1. Location
2. Scale
3. Symmetry
4. Outliers

data(faithful)  
# Rename the variables so they're more intuitive and names actually make sense  
faithful <- rename(faithful, duration = eruptions, interval = waiting)  
  
#Location or central tendency (mean and median)  
#Scale or spread (sd, range, and interquartile range)  
#Symmetry   
#Outliers   
  
#Location Insight - Create a violin boxplot with quartile ranges like above for the duration and interval variables  
mean\_duration <- mean(faithful$duration)  
median\_duration <- median(faithful$duration)  
lower\_hinge\_duration <- quantile(faithful$duration, .25)  
upper\_hinge\_duration <- quantile(faithful$duration, .75)  
  
mean\_interval <- mean(faithful$interval)  
median\_interval <- median(faithful$interval)  
lower\_hinge\_interval <- quantile(faithful$interval, .25)  
upper\_hinge\_interval <- quantile(faithful$interval, .75)  
  
#Plot for Duration Location  
faithful %>%  
 ggplot(mapping = aes(x = factor(0), y = duration))+  
 geom\_violin(fill = "lightblue")+  
 geom\_hline(yintercept = mean\_duration, color = "green")+  
 geom\_hline(yintercept = median\_duration, color = "black")+  
 geom\_hline(yintercept = lower\_hinge\_duration, color = "red")+  
 geom\_hline(yintercept = upper\_hinge\_duration, color = "red")+  
 geom\_boxplot(fill = "grey", width = .075)+  
 coord\_flip()+  
 labs(  
 x = "",  
 y = "Duration of Eruption (minutes)",  
 title = "Old Faithful Eruption Duration"  
 )



#Plot for Interval Length Location  
faithful %>%  
 ggplot(mapping = aes(x = factor(0), y = interval))+  
 geom\_violin(fill = "lightblue")+  
 geom\_hline(yintercept = mean\_interval, color = "green")+  
 geom\_hline(yintercept = median\_interval, color = "black")+  
 geom\_hline(yintercept = lower\_hinge\_interval, color = "red")+  
 geom\_hline(yintercept = upper\_hinge\_interval, color = "red")+  
 geom\_boxplot(fill = "grey", width = .075)+  
 coord\_flip()+  
 labs(  
 x = "",  
 y = "Interval Between Eruptions (minutes)",  
 title = "Old Faithful Interval Length Between Eruptions"  
 )



#### Duration Data Summary

summary(faithful$duration)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.600 2.163 4.000 3.488 4.454 5.100

#### Interval Data Summary

summary(faithful$interval)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 43.0 58.0 76.0 70.9 82.0 96.0