DataAssignment\_8

Md. Saifur Rahman

3/14/2021

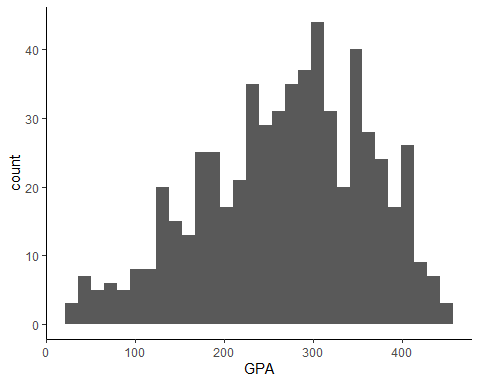
# Task-1: Load the Student Performance 04 dataset.

library(readxl)  
Data <- read\_excel("Student Performance 04.xlsx",   
 na = "NA")  
View(Data)  
attach(Data)

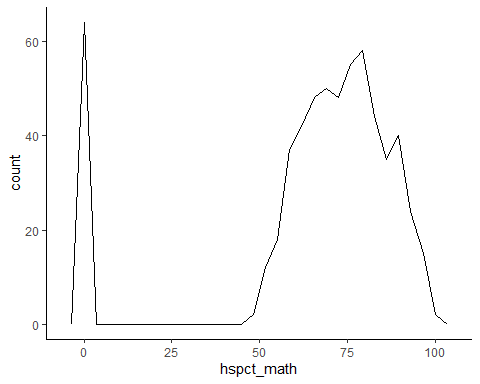
# Task-2: Select three different variables from the dataset and create three unique plots of their distributions (i.e., make one plot for each variable): histogram, frequency polygon, and a density plot.Use best practices as covered in the lesson to modify the plots appearance (color, fill, bin number, bin width, bandwidth, etc.) – EXPLAIN what you did and RATIONALIZE your choices!

## Draw Histogram, Frequency polygon, and Density plot with three diferent variable.

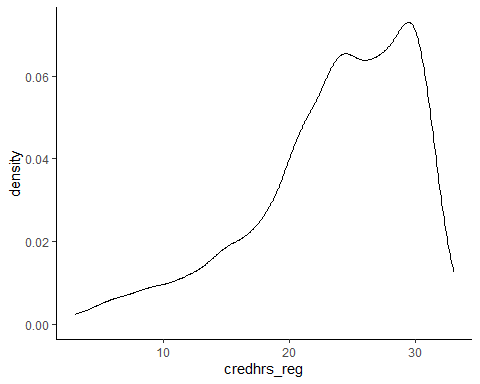
# Histogram for the continuous variable "GPA"   
ggplot(Data, aes(GPA))+geom\_histogram()



# Frequency Polygon for the continuous variable "hspct\_math" (high school percent math!)  
ggplot(Data, aes(hspct\_math))+geom\_freqpoly()

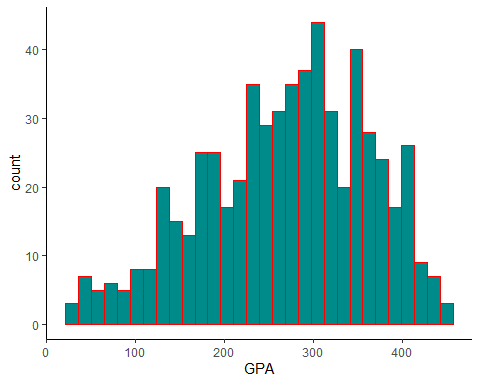


# Density plot for the continuous variable "credhrs\_reg" (total registered credit hours!)  
ggplot(Data, aes(credhrs\_reg))+geom\_density()

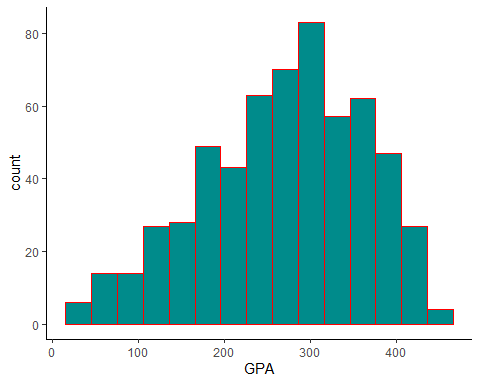


## *Modify plot appearance for Histogram:* In this section, I am using attributes to modify the apprearance of the previous(original) histogram to make this more appealing for the viewers.(i) color=“red”, draw outline color as red; (ii) fill=“cyan4”, fill each histogram bar with the cyan color; (iii) bins=15, limit total number of bins to draw this histogram at 15 (default is 30);(iv) I am not sure about the value of GPA in this excel file! This ranges from 26.28 to 447. Setting binwidth = 42 allow only 10+ or 11 bins to draw this histogram.Also, Using alpha to control transparency.

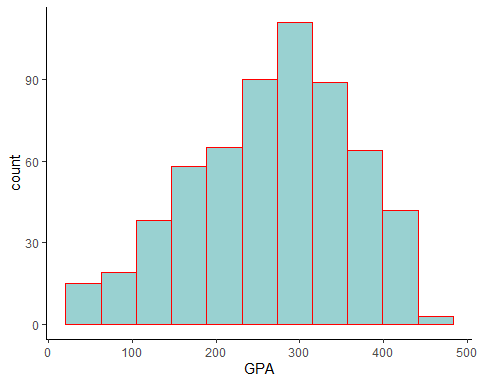
# Histogram for the continuous variable "GPA"   
# Add color and fill attributes to make this histogram more attractive  
ggplot(Data, aes(GPA))+geom\_histogram(color="red", fill="cyan4")



# Add "bins" attribute   
ggplot(Data, aes(GPA))+geom\_histogram(color="red", fill="cyan4", bins=15)

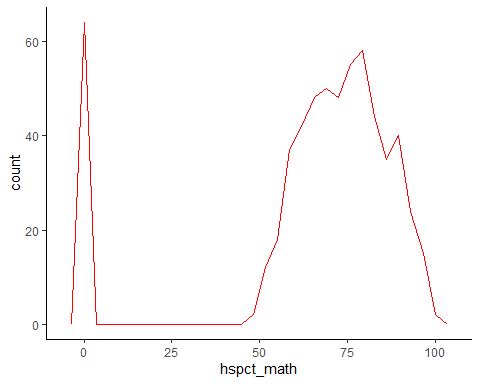


# Set binwidth=42 to reduce number of bins. Using alpha to control transparency  
ggplot(Data, aes(GPA))+geom\_histogram(color="red", fill="cyan4", binwidth=42, alpha=0.4)

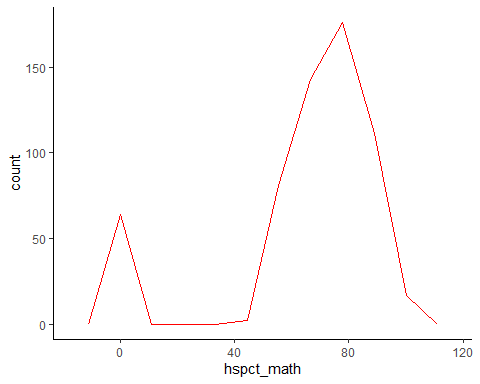


## *Modify plot appearance for Frequency polygon:* In this section, I am using attributes to modify the apprearance of the previous(original) frequency polygon to make this more appealing for the viewers.(i) color=“red”, convert the default line color to red; (ii) bins=10, make this frequency polygon more user friendly and smooth, (iii) another way to make this polygon easy to understand is using binwidth=42.

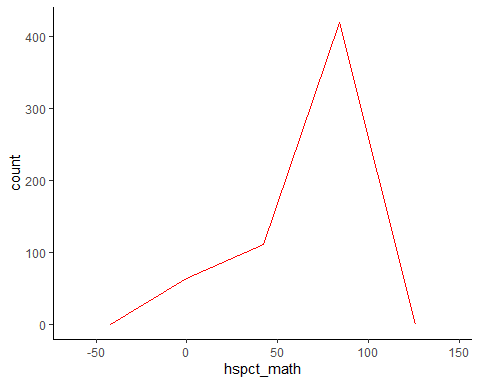
# Frequency Polygon for the continuous variable "hspct\_math" (high school percent math!)  
# Add red color in line.  
ggplot(Data, aes(hspct\_math))+geom\_freqpoly(color="red")



# Add bins number to make this line smoother  
ggplot(Data, aes(hspct\_math))+geom\_freqpoly(color="red", bins=10)

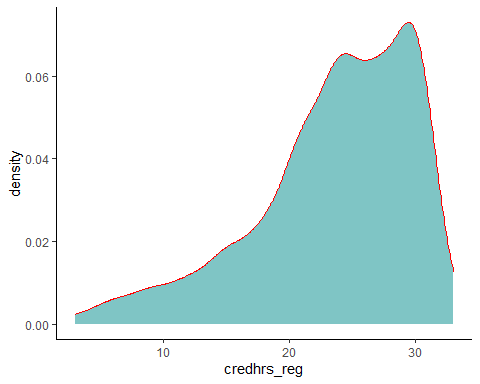


# Add binwidth=42 to reduce the number of bins  
ggplot(Data, aes(hspct\_math))+geom\_freqpoly(color="red", binwidth=42)

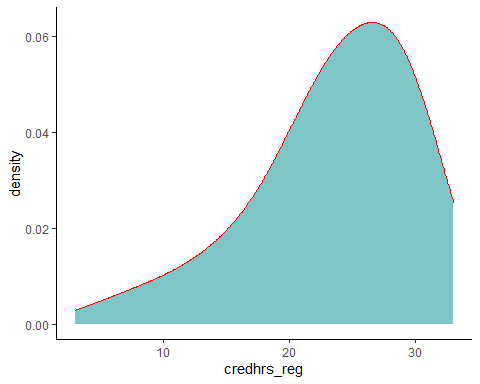


## *Modify plot appearance for Density plot:* In this section, I am using attributes to modify the apprearance of the previous(original) density plot to make this more appealing for the viewers.(i) (color=“red”, fill=“cyan4”, alpha=0.5), use outline color, fill color and transparency to increase acceptance of this density plot to the viewers; (ii) adjust= 2, adjust the shape of this plot and make it smoother; (iii) kernel= “cosine”, this doesn’t make much difference though!

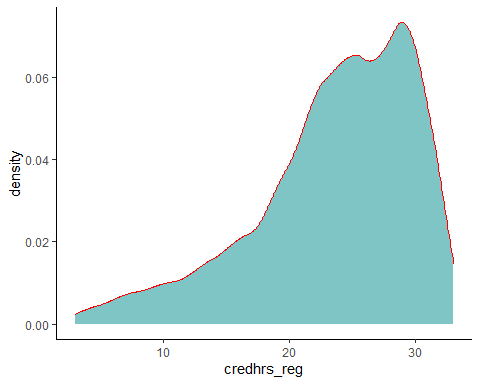
# Density plot for the continuous variable "credhrs\_reg" (total registered credit hours!)  
# add outline color, fill color and transparency  
ggplot(Data, aes(credhrs\_reg))+geom\_density(color="red", fill="cyan4", alpha=0.5)



# Adjust the shape  
ggplot(Data, aes(credhrs\_reg))+geom\_density(adjust= 2,color="red", fill="cyan4", alpha=0.5)



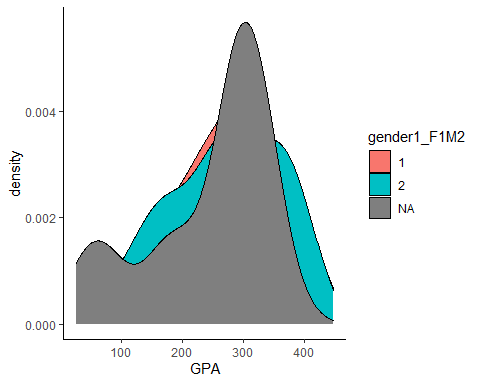
# Use cosine kernel only to check what this look like!  
ggplot(Data, aes(credhrs\_reg))+geom\_density(kernel= "cosine",color="red", fill="cyan4", alpha=0.5)



# Task-3: Create three different plots using variables of your choosing that compare multiple distributions: density plots, boxplots, violin plots. Use best practices as covered in the lesson to modify the plots appearance (color, fill, bandwidth, order, etc.) – EXPLAIN what you did and RATIONALIZE your choices!

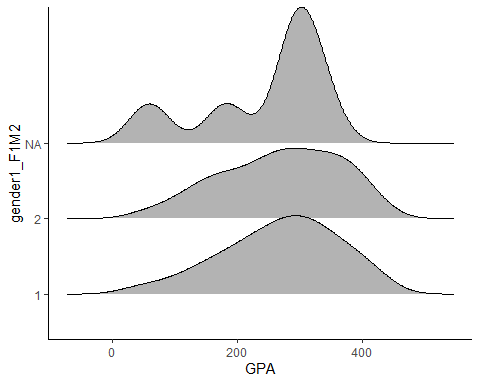
## *Plotting multiple distribution: Density plot, Boxplot, Violin plot:* There is no categorical variable in this dataset. I am seeing the gender variable but in numaric format. So, first convert this gender varaiable from numeric to factor(categorical) variable. I need categorical variable to draw plot for multiple variable distribution.

# Density plots  
# x axis is continuous variable and y axis (fill) is categorical variable. First, convert the gender(gender1\_F1M2) variable from numeric to factor.   
Data$gender1\_F1M2 <- as.factor(Data$gender1\_F1M2)  
# Use geom\_density()  
ggplot(Data, aes(GPA, fill=gender1\_F1M2))+ geom\_density()

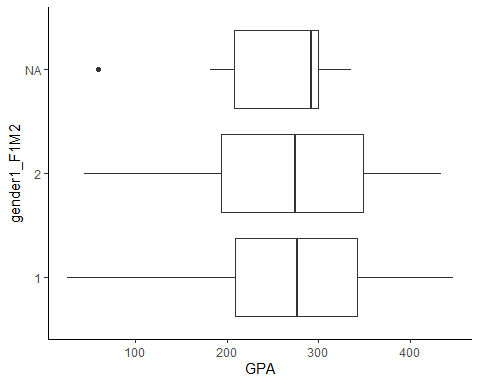


# Use geom\_density\_ridges()  
ggplot(Data, aes(GPA,gender1\_F1M2))+ geom\_density\_ridges()

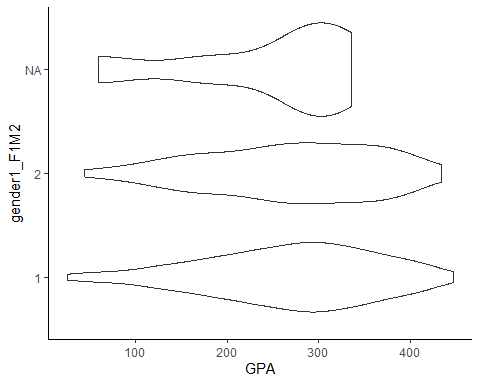
## Picking joint bandwidth of 32.6



# Boxplot   
ggplot(Data, aes(GPA,gender1\_F1M2))+ geom\_boxplot()

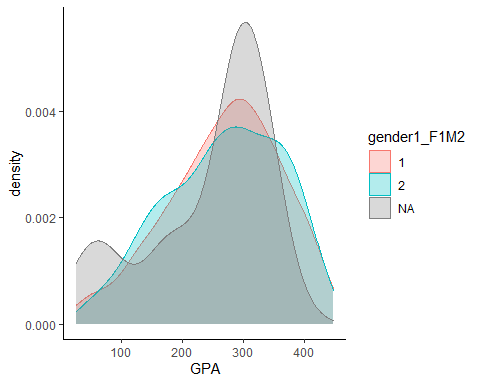


# Violin plots  
ggplot(Data, aes(GPA,gender1\_F1M2))+ geom\_violin()



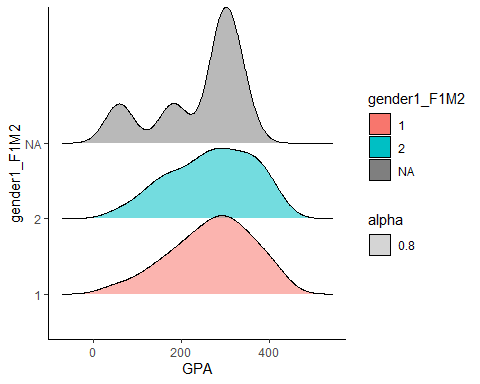
## *Modify plot appearance for Density plots:* In this section, I am using attributes to modify the apprearance of the previous(original) density plot to make this more appealing for the viewers. (i) Add outline color, fill color and transparency. Adjust=1, make some adjustment in the plot. I am considering “NA” i.e mising gender because that’s also a large group in this dataset!!

# Add outline color, fill color and transparency. Adjust=1, make some adjustment in the plot. I am considering "NA" i.e missing gender because that's also a large group in this excel file!!  
ggplot(Data, aes(GPA, fill=gender1\_F1M2, color=gender1\_F1M2))+ geom\_density(adjust=1, alpha=0.3)



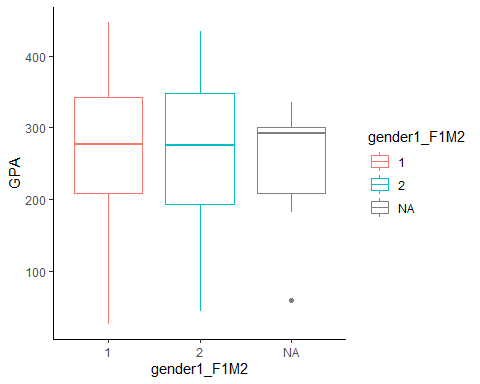
# geom\_density\_ridges()  
ggplot(Data, aes(GPA,gender1\_F1M2,fill=gender1\_F1M2, alpha=0.8))+ geom\_density\_ridges()

## Picking joint bandwidth of 32.6



## *Modify plot appearance for Boxplot:* In this section, I am using attributes to modify the apprearance of the previous(original) boxplot to make this more appealing for the viewers.(i) I interchange variable between x-axis and y-axis to rotate the boxplot. This makes the boxplot verticle which looks good, (ii) add outline color to each box to make this eye catchy.

# Add outline color in each box   
ggplot(Data, aes(gender1\_F1M2, GPA,color=gender1\_F1M2))+ geom\_boxplot()



## *Modify plot appearance for Violin plot:* In this section, I am using attributes to modify the apprearance of the previous(original) violin plot to make this more appealing for the viewers.(i) fill=gender1\_F1M2, fill each violin plot based on gender; (ii) color=“green”, draw outline color for each violin plot as green.

# Add outline color and fill color to increase the appealing  
ggplot(Data, aes(gender1\_F1M2,GPA, fill=gender1\_F1M2))+ geom\_violin(color="green")

