FinalProject

Md. Saifur Rahman

4/29/2021

# Task-1: Review the data file. Tidy and manage the data (e.g., variable names) as needed to prepare for analysis.Briefly explain what you did and why.

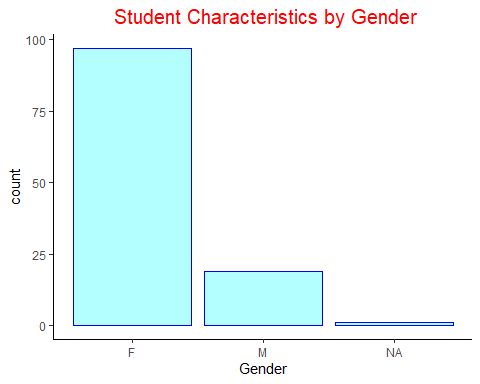
## Answer: I review the data from this file and make this dataset tidy by doing the following. (i) Removing unnecessary data from this file (e.g. delete ‘lab\_sec’, ‘Team’, etc. column with its data), (ii) Rename almost each variable(column) name in this file to make each variable name more understandable (e.g. Exam\_1(35), Mastery\_Chemistry\_HW\_total(100), Mastering\_Chemistry\_Quiz\_Total(42) and so on), (iii) Create a few new variables(column) from the existing data to answer given queries and show visual presenataion (e.g. MC\_Performance(154), Exam\_Review(45), Peer\_Evaluation(30), etc.), (iv) Reorder some variables to improve data organization. Besides, I remove some variables and data from this dataset because I have found those as optional for this project work.

library(readxl)  
Data <- read\_excel("Tidy Chemistry Data.xlsx",   
 na = "NA")  
View(Data)  
attach(Data)

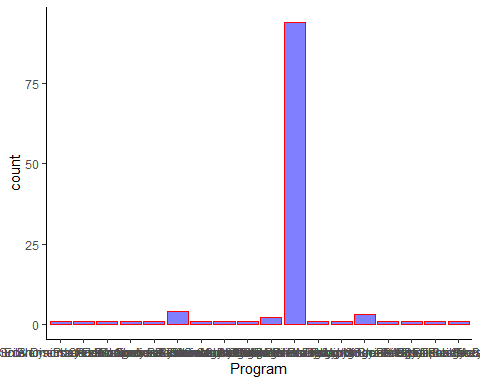
# Task-2: Who were the students in this class? Visualize student characteristics by gender, program, and level.

## Answer: (i) According to gender, highest number of students are female in this class. (ii) According to the program, “Nursing/Pre-Nursing” is the largest group in this class. There is a diversity in the students’ backgroung pogram. (iii) According to level, the largest group of students is in their “Freshman” year. At first, I get 6 levels while having two group in the freshman and sophomore. It’s a capital case issue. A few students insert sophomore starting with lowercase “s” which should be always uppercase “S”. Again, the same occurs in the group of freshman level. That’s why get 6 level instead of 4. I need to convert first letter of all values from this column into uppercase to solve this issue. The complete visualization is given in the following.

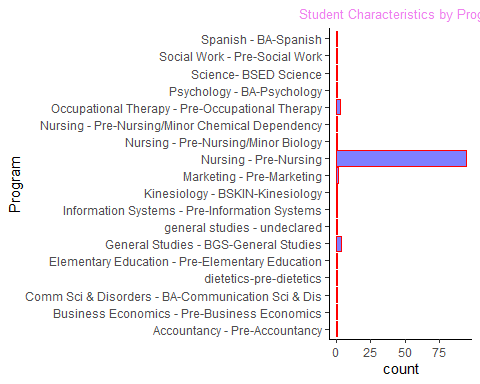
# (i) Visualize student characteristics by gender.  
  
# Visualize student characteristics by gender. I get that a few students didn't insert gender identity in the "Gender" column. Hence, visualizing NA too! Modify color . Add tittle.  
ggplot(Data, aes(Gender))+geom\_bar(fill="cyan", alpha = 0.3, color="blue") + labs(title=" Student Characteristics by Gender") + theme(plot.title = element\_text(hjust = 0.5,size = 15, color = "red"))



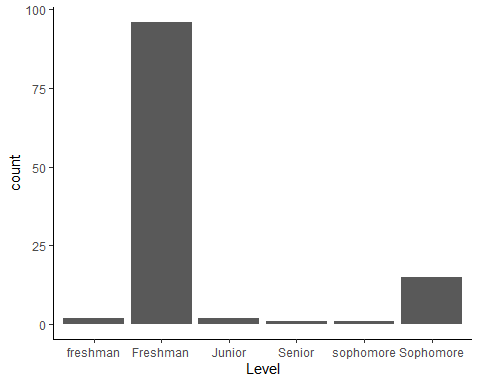
# (ii) Visualize student characteristics by program.  
  
# Visualize student characteristics by program. It's hard to read the labels of x-axis.I need to flip the co-ordinate of this plot.  
ggplot(Data, aes(Program))+geom\_bar(fill="blue", color="red", alpha = 0.5)



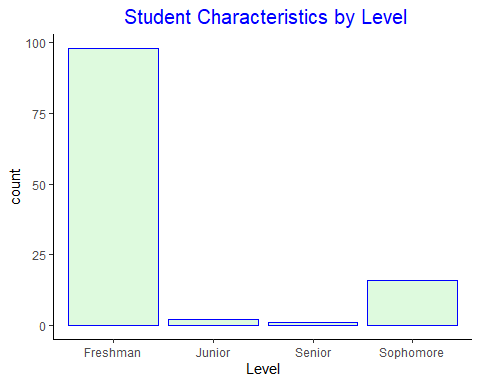
# Flip the plot and draw a horizontal bar chart. Add a title of the graph.  
ggplot(Data, aes(Program))+geom\_bar(fill="blue", color="red", alpha = 0.5)+ coord\_flip()+ labs(title=" Student Characteristics by Program") + theme(plot.title = element\_text(hjust = 0.5,size = 10, color = "violet"))



# (iii) Visualize student characteristics by level.  
  
  
# Visualize student characteristics by level. I get 6 levels while having two group in the freshman and sophomore. It's a capital case issue. A few students insert sophomore starting with lowercase "s" which should be always uppercase "S". Again, the same occurs in the group of freshman level. That's why get 6 level instead of 4. I need to convert first letter of all values from this column into uppercase to solve this issue.   
ggplot(Data, aes(Level))+geom\_bar()



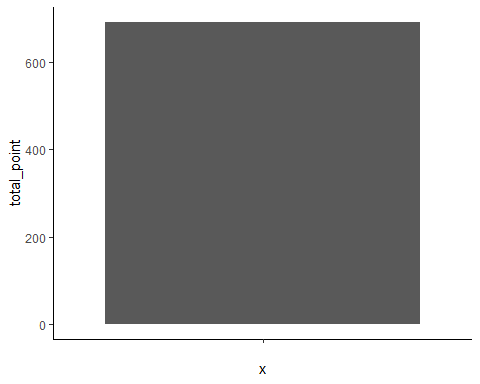
# Capitalize first letter of each value from the "Level" to avoid "Sophomore/sophomore", "Freshman/freshman" letter case issue.  
#?toupper  
capFirst <- function(s) {paste(toupper(substring(s, 1, 1)), substring(s, 2), sep = "")}  
Data$Level <- capFirst(Data$Level)  
  
# Again, Visualize student characteristics by level. Now, this works as expected. Add a title.  
ggplot(Data, aes(Level))+geom\_bar(fill="lightgreen", alpha = 0.3, color="blue") + labs(title=" Student Characteristics by Level") + theme(plot.title = element\_text(hjust = 0.5,size = 15, color = "blue"))



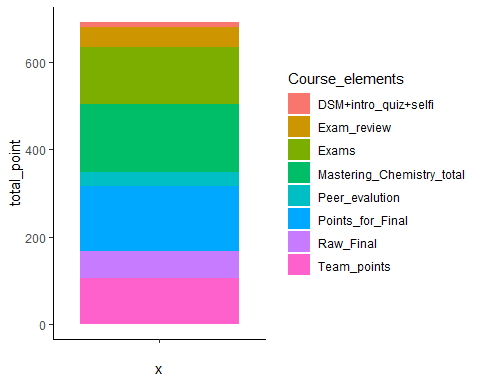
# Task-3: How were the students evaluated in this class? Visualize the composition of the final course grade: exams, online “Mastering Chemistry” (MC), in class “Team”, “Exam Review” sessions, peer evaluations, etc.

## Answer: The given data file is ambiguous in terms of totals course points. It shows total points for this course is 882 but when I add all individual points elements, I get 692 as a total! So, I decided to be consistent with the total getting from individual elements. Elements are as follows, (1) Points\_for\_Final = 150, (2) Raw\_Final = 60, (3) Exams (1+2+3+4) = 131, (4) Mastering\_Chemistry\_total(Mastering\_Chemistry\_module+Mastering\_Chemistry\_HW+Mastering\_Chemistry\_quiz) = 157, (5) Team\_points(Team\_report\_total+Team\_quiz\_total+Team\_survey) = 106, (6) Exam\_review = 45, (7) Peer-evalution (Peer\_evalution+ peer\_feedback) = 31, (8) DSM+intro\_quiz+selfi = 12. I am following this setup for visualizaing the composition of the final course grade.

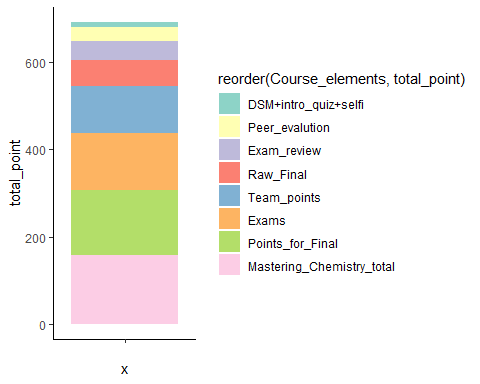
# Create a new data frame with these eight elements.   
Data\_eval <- data.frame( Course\_elements = c("Points\_for\_Final", "Raw\_Final", "Exams", "Mastering\_Chemistry\_total","Team\_points",   
 "Exam\_review", "Peer\_evalution", "DSM+intro\_quiz+selfi"),  
 total\_point = c(150,60,131,157,106,45,31,12))  
#view(Data\_eval)  
  
# Draw stacked barplot. without color stacked barplot is impossible to read.  
ggplot(Data\_eval, aes(x="", y=total\_point))+ geom\_bar(stat = "identity")



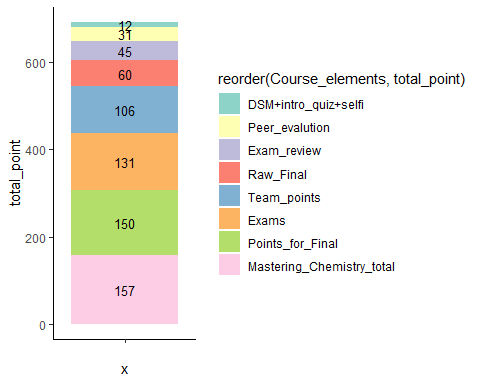
# Add color to understand stacked barplot  
ggplot(Data\_eval, aes(x="", y=total\_point, fill=Course\_elements))+ geom\_bar(stat = "identity")



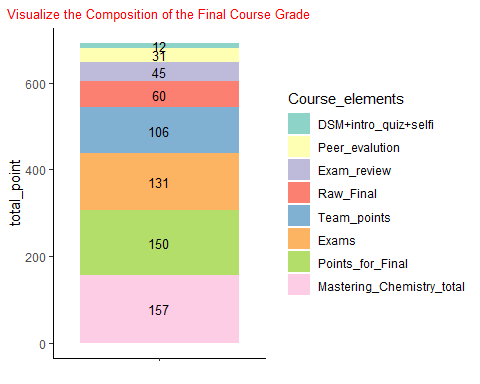
# Change colors of the bars. I am using 'scale\_fill\_brewer()' to change the color of bars.Use 'reorder()' function to reorder the stacks in the bar.  
ggplot(Data\_eval, aes(x="", y=total\_point, fill=reorder(Course\_elements,total\_point)))+ geom\_bar(stat = "identity")+ scale\_fill\_brewer(palette = "Set3")



# Add data labels to them. Use 'geom\_text()' to add value labels, modify color, modify size, and position of the labels.  
ggplot(Data\_eval, aes(x="", y=total\_point, fill=reorder(Course\_elements,total\_point)))+ geom\_bar(stat = "identity")+ scale\_fill\_brewer(palette = "Set3") + geom\_text(aes(label=total\_point),size=3.5, color="black", position = position\_stack(vjust = 0.5))



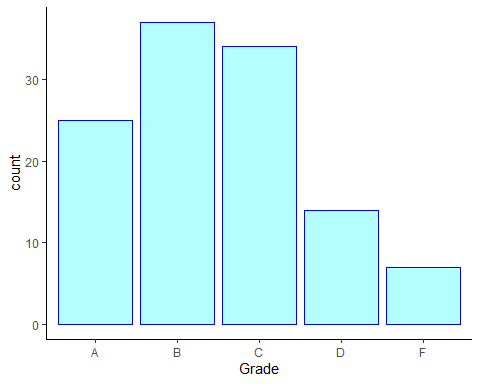
# Add a title for the chart and center it. Use 'ggtitle()' function to add a title. Use 'theme()' function to modify size of the title, change title color, and center the title in this plot.Use "labs()" to rename the legend. Remove text from x-axis because it's not necessary. Add title.  
ggplot(Data\_eval, aes(x="", y=total\_point, fill=reorder(Course\_elements,total\_point)))+ geom\_bar(stat = "identity")+ scale\_fill\_brewer(palette = "Set3") + geom\_text(aes(label=total\_point),size=3.5, color="black", position = position\_stack(vjust = 0.5))+ ggtitle("Visualize the Composition of the Final Course Grade") + theme(axis.title.x = element\_blank(),plot.title = element\_text(hjust = 0.5,size = 10, color = "red")) + labs(fill="Course\_elements")



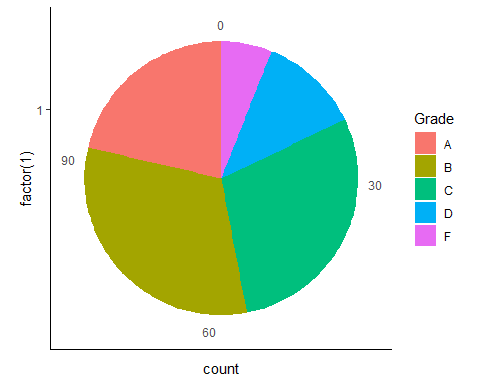
# Task-4: How did the students perform in the class overall? Visualize the proportion of the letter grades and the distribution of the total course points.

## Answer: (i) According to the pie chart presented below, the proportion of the students get “B”, and “C” are almost similar. 32% of the total student get B in this class, while 29% of them get C grade. 21% of them achieve grade A. Unfortunately, 6% students from this class fail in this course.(ii) The least total score is around 390 and the highest total score is around 870. Most of the student get total points distributed in between 595 to 750 (approximate!).

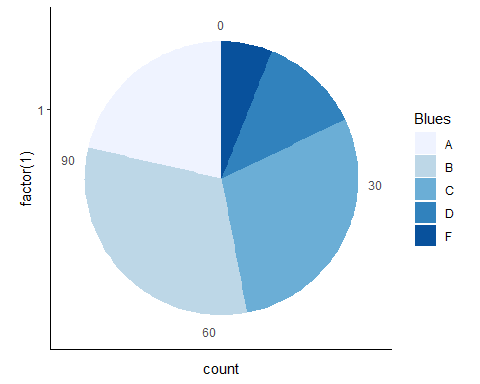
# (i) Visualize the proportion of the letter grades  
  
# Visualize the proportion of the letter grades using bar chart. Bar chart can not present the real proportion. Need pie chart to show real proportion.   
ggplot(Data, aes(Grade))+geom\_bar(fill="cyan", alpha = 0.3, color="blue")



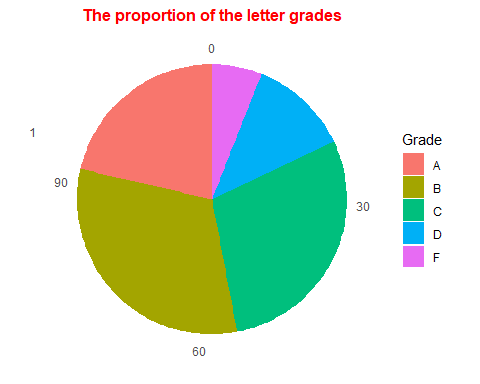
# Visualize the proportion of the letter grades using pie chart  
ggplot(Data, aes(factor(1), fill= Grade))+geom\_bar(width = 1) + coord\_polar("y")



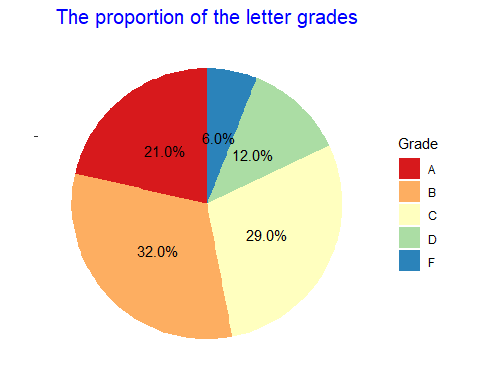
ggplot(Data, aes(factor(1), fill= Grade))+geom\_bar(width = 1) + coord\_polar("y")+ scale\_fill\_brewer("Blues")



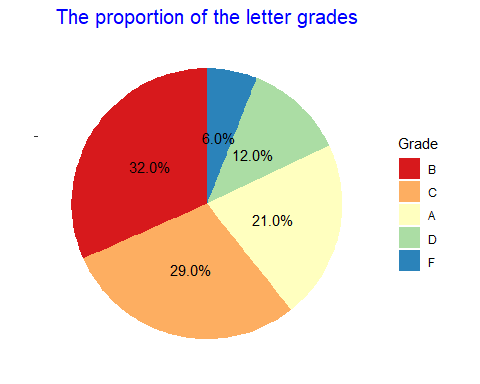
# Modify the pie chart. Add a title using ggtitle(). Use theme() to remove x-axis title, y-axis title, axis line and some modification in the plot title.  
ggplot(Data, aes(factor(1), fill= Grade))+geom\_bar(width = 1) + coord\_polar("y") + ggtitle("The proportion of the letter grades") + theme(axis.title.x = element\_blank(),axis.title.y = element\_blank(), axis.line = element\_blank(),axis.ticks = element\_blank(), plot.title=element\_text( size=12, face="bold", color = "red", hjust = 0.5))



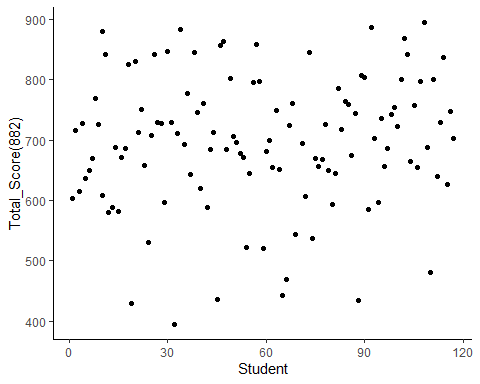
# Add percentage/label to each pie slices.  
# Create a new variable "Grade\_count" to count each group in the grade column. Use "group\_by()" and "summarize()" function to count the number in each grade.Then,create another variable "Grade\_percent" to insert percent value in the "Grade\_count" data frame. Use 'percent()' function to convert percent value into percent presentation. Use 'round()' function to select 3 values after decimal point.  
  
Grade\_count <- group\_by(Data,Grade) %>% summarise( number\_rows = n())   
#view(Grade\_count)  
Grade\_count$Grade\_percent <- round((Grade\_count$number\_rows)/sum(Grade\_count$number\_rows),2)  
  
# Now, draw the same pie chart again but this time use newly created "Grade\_count" data frame to show the percent label. Use "geom\_text()" and "percent()" function to show the percent label.   
ggplot(Grade\_count, aes(x="", y= number\_rows, fill= Grade))+ geom\_bar(stat = "identity")+ scale\_fill\_brewer(palette = "Spectral")+coord\_polar(theta = "y", start = 0) + labs(x=NULL,y=NULL,title="The proportion of the letter grades") + theme(axis.line = element\_blank(),axis.text = element\_blank(),plot.title = element\_text(hjust = 0.5,size = 15,color = "blue"))+ geom\_text(aes(label=scales::percent(round(Grade\_percent,3))), position = position\_stack(vjust = 0.5),size=4)



# Reorder the slices from largest to smallest starting at 12 o’clock. Use 'reorder()' function to reorder the slices. Rename legend title using labs( fill = "Grade")  
ggplot(Grade\_count, aes(x="", y= number\_rows, fill= reorder(Grade,-Grade\_percent)))+ geom\_bar(stat = "identity")+ scale\_fill\_brewer(palette = "Spectral")+coord\_polar(theta = "y", start = 0) + labs(x=NULL,y=NULL,title="The proportion of the letter grades") + theme(axis.line = element\_blank(),axis.text = element\_blank(),plot.title = element\_text(hjust = 0.5,size = 15, color = "blue"))+ geom\_text(aes(label=scales::percent(round(Grade\_percent,3))), position = position\_stack(vjust = 0.5),size=4) + labs(fill= "Grade")

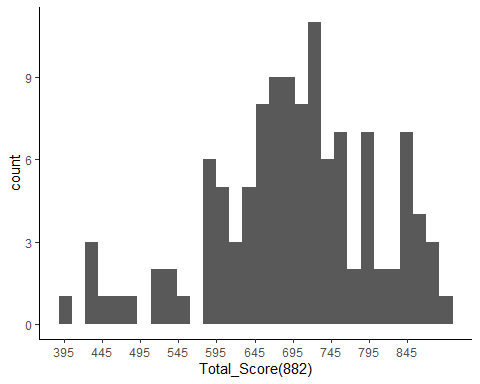


# (ii) Visualize the distribution of the total course points  
  
  
  
# Visualize the distribution of the total course points using scatter plot. Scatter plot doesn't give a good visualization.   
ggplot(Data, aes(Student,`Total\_Score(882)`))+geom\_point()

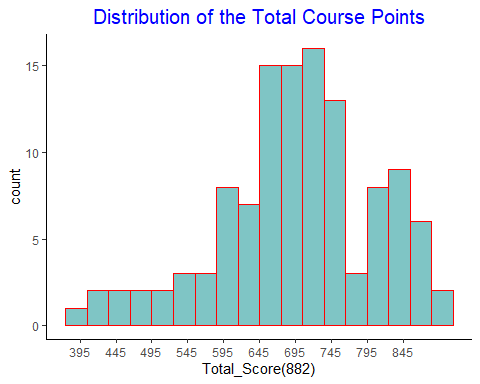


# Visualize the distribution of the total course points using histogram. Modify x-axis ticks using "scale\_x\_continuous()" function.   
ggplot(Data, aes(`Total\_Score(882)`))+geom\_histogram()+ scale\_x\_continuous(breaks = round(seq(min(Data$`Total\_Score(882)`), max(Data$`Total\_Score(882)`), by = 50)))

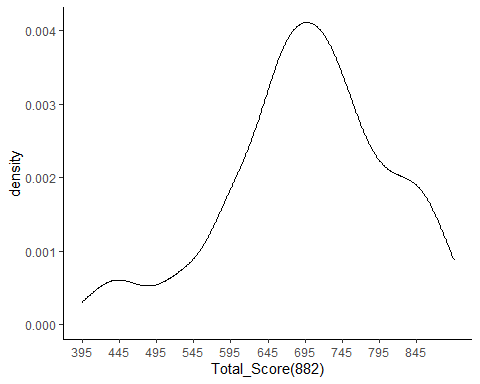
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



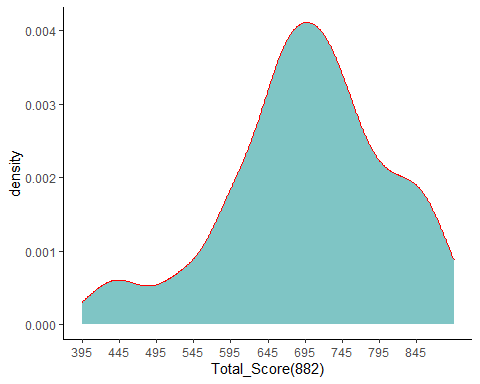
# Add color and fill attributes to make this histogram more attractive. Set binwidth=30 to reduce number of bins. Using alpha to control transparency. Add a title.   
ggplot(Data, aes(`Total\_Score(882)`))+geom\_histogram(color="red", fill="cyan4", binwidth=30, alpha=0.5)+ scale\_x\_continuous(breaks = round(seq(min(Data$`Total\_Score(882)`), max(Data$`Total\_Score(882)`), by = 50))) + labs(title="Distribution of the Total Course Points") + theme(plot.title = element\_text(hjust = 0.5,size = 15, color = "blue"))



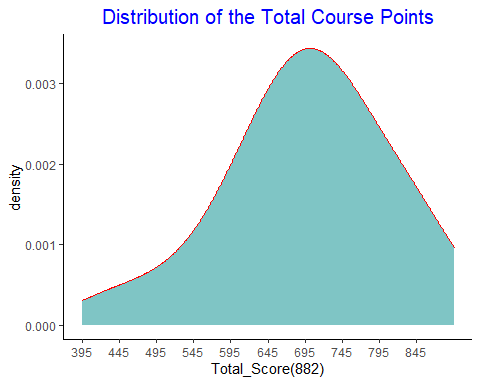
# Visualize the distribution of the total course points using density plot. Modify x-axis ticks using "scale\_x\_continuous()" function.  
ggplot(Data, aes(`Total\_Score(882)`))+geom\_density() + scale\_x\_continuous(breaks = round(seq(min(Data$`Total\_Score(882)`), max(Data$`Total\_Score(882)`), by = 50)))



# Modiy color, fill the shape, color the outline etc.  
ggplot(Data, aes(`Total\_Score(882)`))+geom\_density(color="red", fill="cyan4", alpha=0.5) + scale\_x\_continuous(breaks = round(seq(min(Data$`Total\_Score(882)`), max(Data$`Total\_Score(882)`), by = 50)))



# Adjust the shape and add a title  
ggplot(Data, aes(`Total\_Score(882)`))+geom\_density(adjust = 2,color="red", fill="cyan4", alpha=0.5) + scale\_x\_continuous(breaks = round(seq(min(Data$`Total\_Score(882)`), max(Data$`Total\_Score(882)`), by = 50)))+ labs(title="Distribution of the Total Course Points") + theme(plot.title = element\_text(hjust = 0.5,size = 15, color = "blue"))

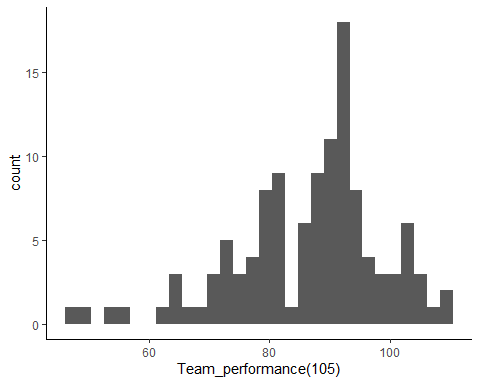


# Task-5: How effectively did students participate in class during “Team” activities (daily group activities in class among three students, includes quizzes and reports)? Visualize the distribution of student Team work. How did Team activity performance relate to exam performance?

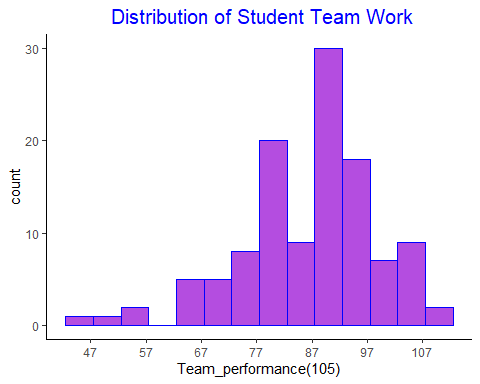
## Answer: (i) I draw histogram and frequency polygon to visualize the distribution of students team activities (Report+Quiz). I add a new varaiable named “Team\_performance(105)” which has summed value of “Team\_Quiz\_total(47)”, and “Team\_Report\_total(58)” (ii) To Visualize the relation between Team activity performance and exam performance, I am considering “Total\_Score(882)” as exam performace. Because I feel “Team\_performance(105)” is more related to“Total\_Score(882)” than only Exam score (4 exams). From the correlogram, I get a strong relation in between “Team\_performance(105)” and “Total\_Score(882)” which label value is 0.8. From the scatter plot and smooth line I get the same trends and that is total score increases with the increase in the team performance.

# (i) Visualize the distribution of student Team work  
  
# Visualize the distribution of student Team work (Report+Quiz) using histogram.   
ggplot(Data, aes(`Team\_performance(105)`))+geom\_histogram()

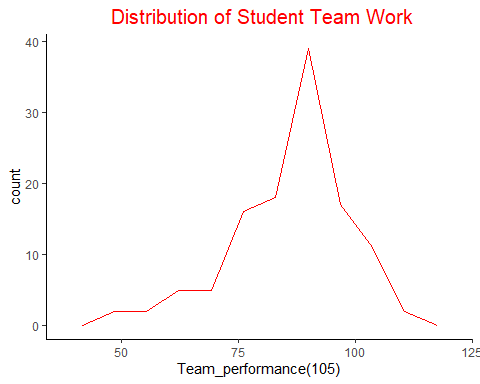
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



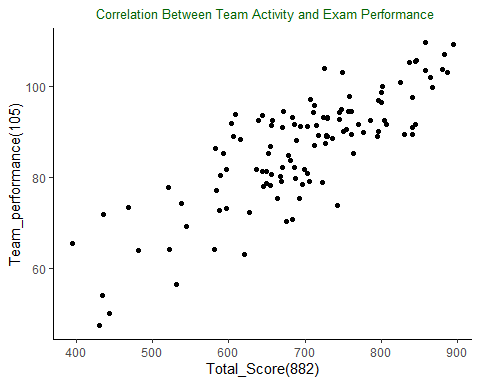
# Add color and fill attributes to make this histogram more attractive. Set binwidth=5 to reduce number of bins. Using alpha to control transparency. Modify x-axis ticks using "scale\_x\_continuous()" function. Add a title.  
ggplot(Data, aes(`Team\_performance(105)`))+geom\_histogram(color="blue", fill="darkviolet", binwidth=5, alpha = 0.7)+ scale\_x\_continuous(breaks = round(seq(min(Data$`Team\_performance(105)`), max(Data$`Team\_performance(105)`), by = 10)))+ labs(title=" Distribution of Student Team Work") + theme(plot.title = element\_text(hjust = 0.5,size = 15, color = "blue"))



# Visualize the distribution of student Team work (Report+Quiz) using frequency polygon. Add red color in line.Add bins number to make this line smoother  
ggplot(Data, aes(`Team\_performance(105)`))+geom\_freqpoly(color="red", bins=10) + labs(title=" Distribution of Student Team Work") + theme(plot.title = element\_text(hjust = 0.5,size = 15, color = "blue"))+ labs(title=" Distribution of Student Team Work") + theme(plot.title = element\_text(hjust = 0.5,size = 15, color = "red"))

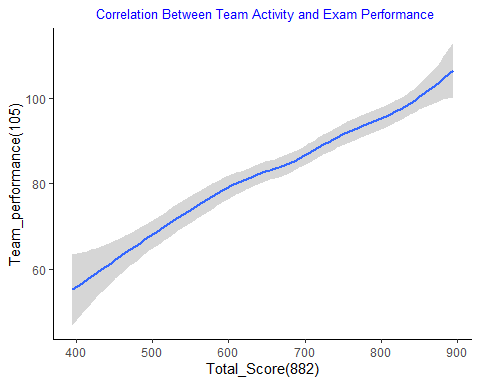


# (ii) How did Team activity performance relate to exam performance?  
  
# Visualize the relation between Team activity performance and exam performance using scatter plot  
ggplot(Data, aes(`Total\_Score(882)`,`Team\_performance(105)`))+geom\_point()+ labs(title=" Correlation Between Team Activity and Exam Performance") + theme(plot.title = element\_text(hjust = 0.5,size = 10, color = "darkgreen"))

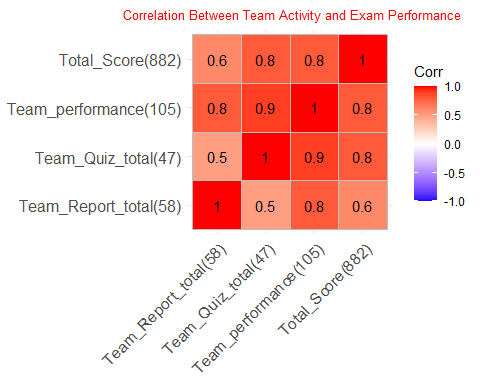


# Visualize the relation between Team activity performance and exam performance using a smooth line  
ggplot(Data, aes(`Total\_Score(882)`,`Team\_performance(105)`))+geom\_smooth()+ labs(title=" Correlation Between Team Activity and Exam Performance") + theme(plot.title = element\_text(hjust = 0.5,size = 10, color = "blue"))

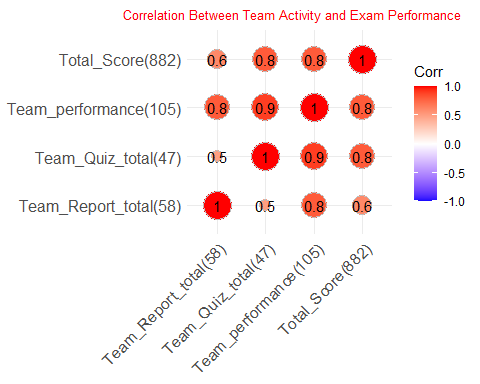
## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



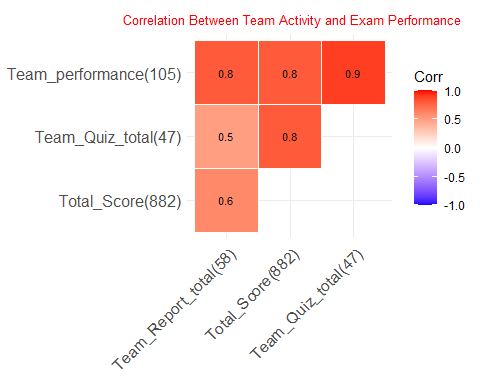
# Correlation graph. Create subset with continuous variable form the original dataset.  
Data\_cor <- subset(Data, select = c("Team\_Report\_total(58)","Team\_Quiz\_total(47)", "Team\_performance(105)", "Total\_Score(882)"))  
  
# Round the data and consider 1 value after decimal point for correlation  
Data\_cor\_round <- round(cor(Data\_cor),1)  
  
# Visualize the correlation matrix. method = "square" (default)Add Correlation label. Add a title.  
ggcorrplot(Data\_cor\_round, lab = TRUE)+ labs(title=" Correlation Between Team Activity and Exam Performance") + theme(plot.title = element\_text(hjust = 0.5,size = 10, color = "red"))



# method = "circle"  
ggcorrplot(Data\_cor\_round, method = "circle", lab = TRUE)+ labs(title=" Correlation Between Team Activity and Exam Performance") + theme(plot.title = element\_text(hjust = 0.5,size = 10, color = "red"))



# Use hierarchical clustering.  
ggcorrplot(Data\_cor\_round, hc.order = TRUE, type = "upper", lab=TRUE, lab\_size = 3, outline.col = "white") + labs(title=" Correlation Between Team Activity and Exam Performance") + theme(plot.title = element\_text(hjust = 0.5,size = 10, color = "red"))



# Correlation Using APA format from apaTables package.  
apa.cor.table(Data\_cor\_round, filename = NA,table.number = NA, show.conf.interval = FALSE,landscape = TRUE)

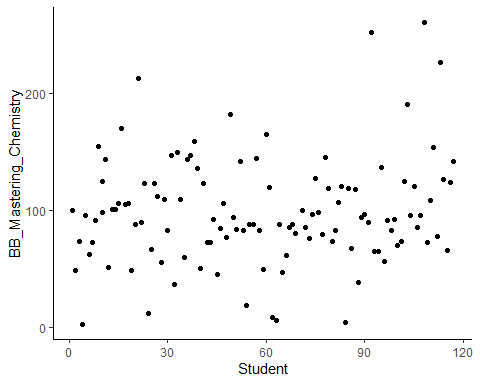
## The ability to suppress reporting of reporting confidence intervals has been deprecated in this version.  
## The function argument show.conf.interval will be removed in a later version.

##   
##   
## Means, standard deviations, and correlations with confidence intervals  
##   
##   
## Variable M SD 1 2 3   
## 1. Team\_Report\_total(58) 0.72 0.22   
##   
## 2. Team\_Quiz\_total(47) 0.80 0.22 -.84   
## [-1.00, .64]   
##   
## 3. Team\_performance(105) 0.88 0.10 -.12 .64   
## [-.97, .95] [-.83, .99]   
##   
## 4. Total\_Score(882) 0.80 0.16 -.74 .57 .00   
## [-.99, .77] [-.87, .99] [-.96, .96]  
##   
##   
## Note. M and SD are used to represent mean and standard deviation, respectively.  
## Values in square brackets indicate the 95% confidence interval.  
## The confidence interval is a plausible range of population correlations   
## that could have caused the sample correlation (Cumming, 2014).  
## \* indicates p < .05. \*\* indicates p < .01.  
##

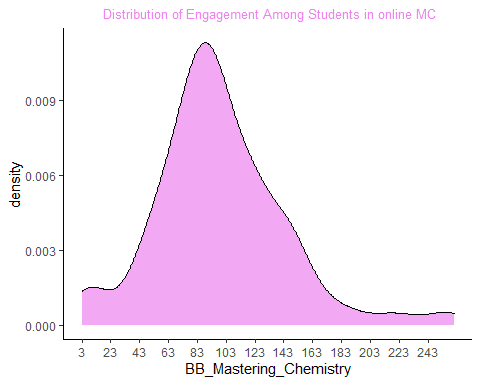
# Task-6: Visualize the distribution of engagement among students outside of class in the online “Mastering Chemistry” (MC). Is greater MC performance related to better student performance on exams?

## Answer: (i) I assume that student engagement in the online “Mastering Chemistry” means the “BB\_Mastering\_Chemistry” data. (ii) I create and add a new variable to serve the purpose in this section. That is “Exam\_performance(131)” which is a summation of “Exam\_1(35)”, “Exam\_2(31)”, “Exam\_3(31)”, and “Exam\_4(34)”. I draw correlation in between the BB\_Mastering\_Chemistry" and “Exam\_performance(131)”. It seems there is a weak relationship in between these with matrix value 0.4.

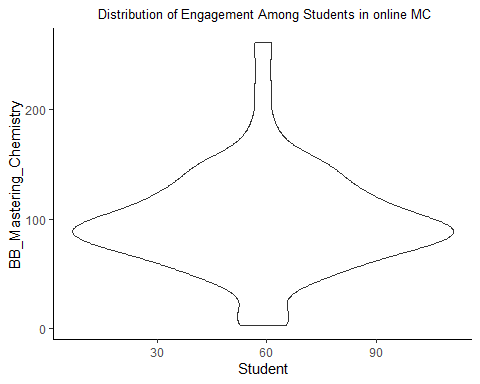
#(i) Visualize the distribution of engagement among students outside of class in the online “Mastering Chemistry” (MC)  
  
  
# Visualize the distribution among students in the online “Mastering Chemistry” (MC) using scatter plot.   
ggplot(Data, aes(Student, BB\_Mastering\_Chemistry))+geom\_point()



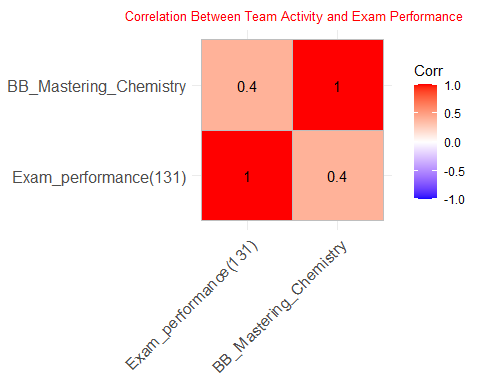
# Visualize the distribution among students in the online “Mastering Chemistry” using density plot. Modify x-axis ticks using "scale\_x\_continuous()" function. Modify outline color, fill color and transparency of the graph. Add a title.  
ggplot(Data, aes(BB\_Mastering\_Chemistry))+geom\_density(color="black", fill="violet", alpha=0.7) + scale\_x\_continuous(breaks = round(seq(min(Data$BB\_Mastering\_Chemistry), max(Data$BB\_Mastering\_Chemistry), by = 20)))+ labs(title=" Distribution of Engagement Among Students in online MC") + theme(plot.title = element\_text(hjust = 0.5,size = 10, color = "violet"))



# Violin plot to see the distribution.   
ggplot(Data, aes(Student,BB\_Mastering\_Chemistry)) + geom\_violin()+ labs(title=" Distribution of Engagement Among Students in online MC") + theme(plot.title = element\_text(hjust = 0.5,size = 10, color = "black"))



#(ii) Is greater MC performance related to better student performance on exams?  
  
# Is greater MC performance related to better student performance on exams?  
# Correlation graph. Create subset with continuous variable form the original dataset.  
Data\_cor\_xmMC <- subset(Data, select = c("Exam\_performance(131)","BB\_Mastering\_Chemistry"))  
  
# Round the data and consider 1 value after decimal point for correlation  
Data\_cor\_round\_xmMC <- round(cor(Data\_cor\_xmMC),1)  
  
# Visualize the correlation matrix. method = "square" (default)Add Correlation label. Add a title.  
ggcorrplot(Data\_cor\_round\_xmMC, lab = TRUE)+ labs(title=" Correlation Between Team Activity and Exam Performance") + theme(plot.title = element\_text(hjust = 0.5,size = 10, color = "red"))



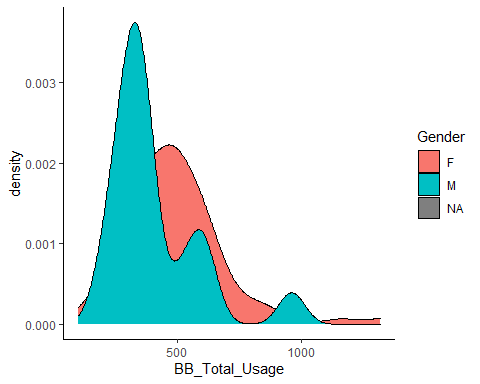
# Task-7: Were any types of students more/less active on Blackboard? Was Blackboard MC activity related to Homework Points, Quiz Points, or Final Course grade?

## Answer: (i) Considering the type of students, I generate density plot based on gender and level. I also draw a violin plot based on gender to visualize student activities on BB (ii) Final course grade has categorical values. So, it’s not posible to make correlation with this variable with others. Instead I am considering the “Total\_Score(882)” in replace of the “Grade” to draw the correlation matrix among “BB\_Mastering\_Chemistry”,“Mastery\_Chemistry\_HW\_total(100)”, “Total\_Score(882)”, and “Mastering\_Chemistry\_Quiz\_Total(42)”. Yes, Blackboard MC activity related to Homework Points, Quiz Points, and final points in the the same scale. The label of the matrix is 0.5 which presents a medium relation of MC activity with therse three varaiables.

# (i) Were any types of students more/less active on Blackboard?  
  
  
# Plotting multiple distribution using density plot (gender).x axis is continuous variable and y axis (fill) is categorical variable.   
ggplot(Data, aes(BB\_Total\_Usage, fill= Gender)) + geom\_density()

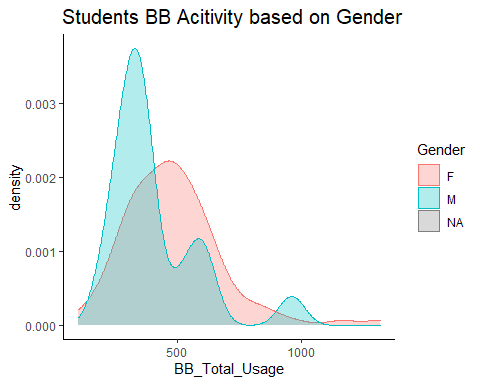
## Warning: Groups with fewer than two data points have been dropped.

## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning -  
## Inf

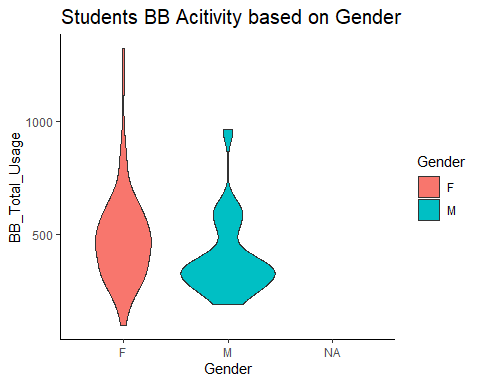


# Modify plot appearance for Density plots. (i) Add outline color, fill color and transparency. Adjust=1, make some adjustment in the plot. I am considering "NA" i.e missing gender because that's also a large group in this dataset!! Add a title.  
ggplot(Data, aes(BB\_Total\_Usage, fill= Gender, color=Gender)) + geom\_density(adjust=1, alpha=0.3)+ labs(title=" Students BB Acitivity based on Gender") + theme(plot.title = element\_text(hjust = 0.5,size = 15, color = "black"))

## Warning: Groups with fewer than two data points have been dropped.  
  
## Warning: no non-missing arguments to max; returning -Inf

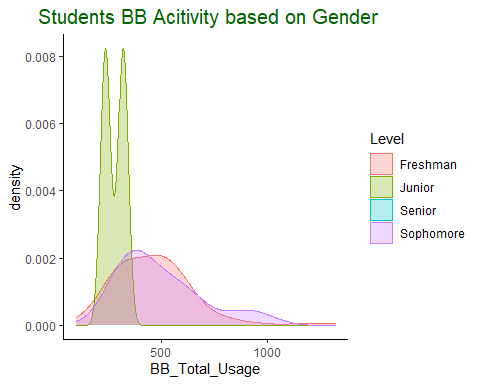


# Violin plot for the same. Modify plot appearance for Violin plot.(i) fill=Gender, fill each violin plot based on gender; (ii) color=“green”, draw outline color for each violin plot as green. Add a title.  
ggplot(Data, aes(Gender,BB\_Total\_Usage,fill=Gender)) + geom\_violin()+ labs(title=" Students BB Acitivity based on Gender") + theme(plot.title = element\_text(hjust = 0.5,size = 15, color = "black"))

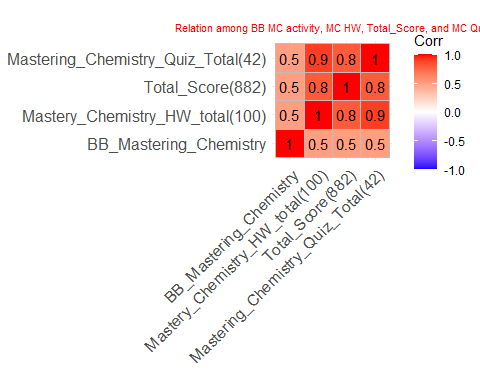


# Same task based on level. (i) Add outline color, fill color and transparency. Adjust=1, make some adjustment in the plot. I am considering "NA" i.e missing gender because that's also a large group in this dataset!! Add a title.  
ggplot(Data, aes(BB\_Total\_Usage, fill= Level, color=Level)) + geom\_density(adjust=1, alpha=0.3)+ labs(title=" Students BB Acitivity based on Gender") + theme(plot.title = element\_text(hjust = 0.5,size = 15, color = "darkgreen"))

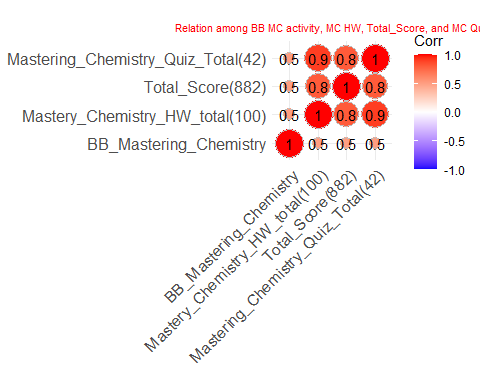
## Warning: Groups with fewer than two data points have been dropped.  
  
## Warning: no non-missing arguments to max; returning -Inf



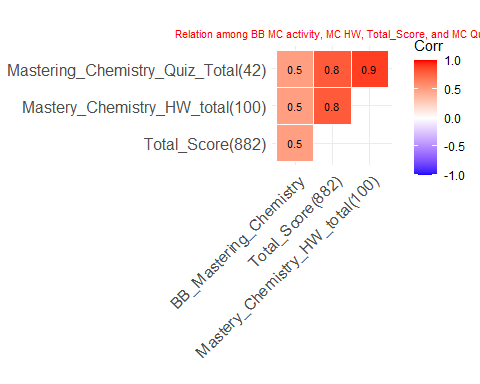
# (ii) Was Blackboard MC activity related to Homework Points, Quiz Points, or Final Course grade?  
  
# Correlation graph. Create subset with continuous variable form the original dataset.Final course grade has categorical value. So, it's not posible to make correlation with this variable with others. Instead I am considering the "Total\_Score(882)" in replace of the "Grade" to draw the correlation matrix  
Data\_corBB <- subset(Data, select = c("BB\_Mastering\_Chemistry","Mastery\_Chemistry\_HW\_total(100)", "Total\_Score(882)", "Mastering\_Chemistry\_Quiz\_Total(42)"))  
  
# Round the data and consider 1 value after decimal point for correlation  
Data\_cor\_round\_BB <- round(cor(Data\_corBB),1)  
  
# Visualize the correlation matrix. method = "square" (default). Add Correlation label. Add a title.  
ggcorrplot(Data\_cor\_round\_BB, lab = TRUE)+ labs(title=" Relation among BB MC activity, MC HW, Total\_Score, and MC Quiz") + theme(plot.title = element\_text(hjust = 0.5,size = 8, color = "red"))



# method = "circle"  
ggcorrplot(Data\_cor\_round\_BB, method = "circle", lab = TRUE)+ labs(title=" Relation among BB MC activity, MC HW, Total\_Score, and MC Quiz") + theme(plot.title = element\_text(hjust = 0.5,size = 8, color = "red"))



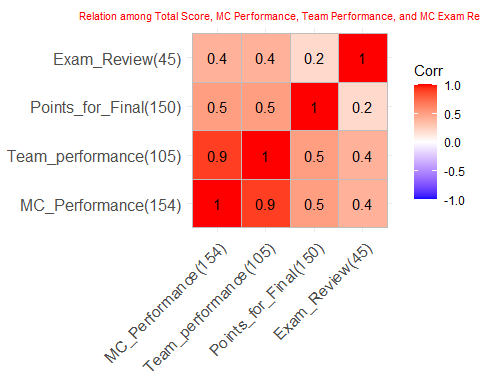
# Use hierarchical clustering.  
ggcorrplot(Data\_cor\_round\_BB, hc.order = TRUE, type = "upper", lab=TRUE, lab\_size = 3, outline.col = "white")+ labs(title=" Relation among BB MC activity, MC HW, Total\_Score, and MC Quiz") + theme(plot.title = element\_text(hjust = 0.5,size = 8, color = "red"))



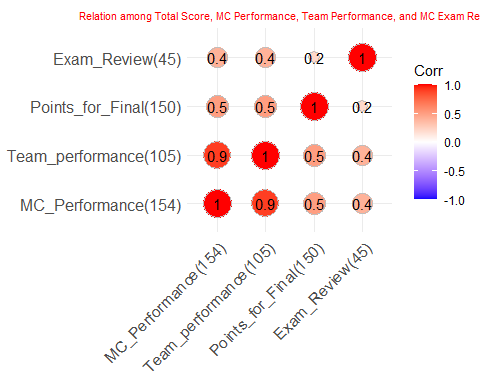
# Task-8: What related more strongly to final exam score: online MC performance, in class Team performance, Exam Review sessions? Visualize these relationships and report on their correlations. Fun bonus: What are some other variables, if any, that relate to final exam performance other than those listed?

## Answer: I ceate a separete “MC\_Performance(154)” varibale to serve the purpose in this section. The variable “MC\_Performance(154)” is the summation of “Mastering\_Chemistry\_Quiz\_Total(42)” , “Mastering\_Chemistry\_Modules(12)”, and “Mastery\_Chemistry\_HW\_total(100)”. I consider the “MC\_Performance(154)” as online MC instead of BB MC. From the correlation matrix, I have found that the “Team\_performance(105)” and “MC\_Performance(154)” are equally correlated with the final exam score (Points\_for\_Final(150)). The corellation among these are 0.5. On the other hand “Exam\_Review(45)” shows 0.2 correlation value which means a weak correlation with the final exam score.

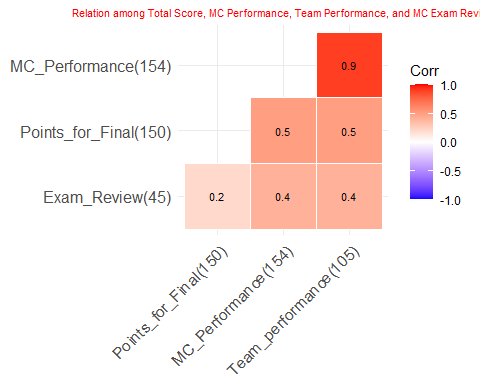
# Correlation graph. Create subset with continuous variable form the original dataset.  
Data\_corExam <- subset(Data, select = c("MC\_Performance(154)","Team\_performance(105)", "Points\_for\_Final(150)", "Exam\_Review(45)"))  
  
# Round the data and consider 1 value after decimal point for correlation  
Data\_cor\_round\_Exam <- round(cor(Data\_corExam),1)  
  
  
# Visualize the correlation matrix. method = "square" (default).Add Correlation label. Add a title  
ggcorrplot(Data\_cor\_round\_Exam, lab = TRUE)+ labs(title=" Relation among Total Score, MC Performance, Team Performance, and MC Exam Review") + theme(plot.title = element\_text(hjust = 0.5,size = 8, color = "red"))



# method = "circle"  
ggcorrplot(Data\_cor\_round\_Exam, method = "circle", lab = TRUE)+ labs(title=" Relation among Total Score, MC Performance, Team Performance, and MC Exam Review") + theme(plot.title = element\_text(hjust = 0.5,size = 8, color = "red"))



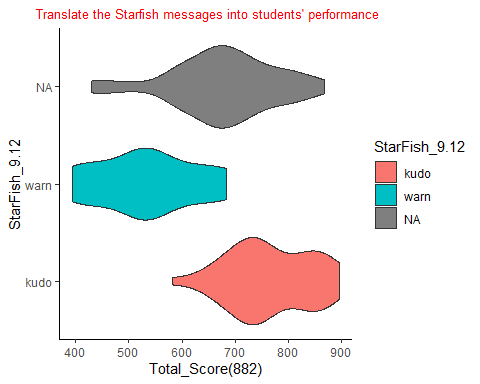
# Use hierarchical clustering.  
ggcorrplot(Data\_cor\_round\_Exam, hc.order = TRUE, type = "lower", lab=TRUE, lab\_size = 3, outline.col = "white")+ labs(title=" Relation among Total Score, MC Performance, Team Performance, and MC Exam Review") + theme(plot.title = element\_text(hjust = 0.5,size = 8, color = "red"))



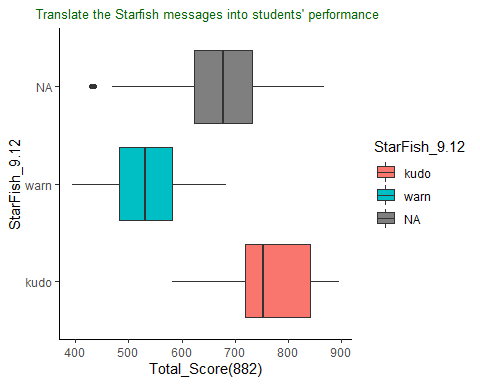
# Task-9: How did the Starfish messages translate into student performance in the course overall? Fun bonus: What are some variables the instructor could use to identify students early on who struggling and performed poorly in the course in the end?

## Answer: I use violin plot and box plot to translate the Starfish messages(StarFish\_9.12 & Starfish\_10.5) into students’ overall performance in the course. Besides ploting the message, I have found lots of missing values in the starfish messages. Violin plot for the StarFish\_9.12 shows a warning for the students with total score renging from around 400 to 650 which is the indication that these students need more care than others. On the other hand, Kudo in starfish means students performance are better than others. Their total score range is greater than 650 (approximate). By this visualization instructor can identify performance of any students at early stage.

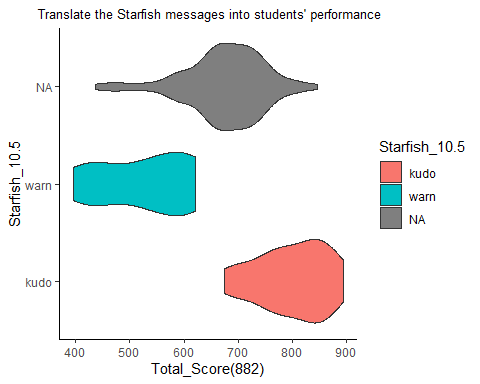
# How did the Starfish messages translate into student performance in the course overall?  
  
# Translate the Starfish messages(StarFish\_9.12) into students' overall performance in the course.   
ggplot(Data, aes(`Total\_Score(882)`,StarFish\_9.12, fill= StarFish\_9.12)) + geom\_violin()+ labs(title=" Translate the Starfish messages into students' performance") + theme(plot.title = element\_text(hjust = 0.5,size = 10, color = "red"))



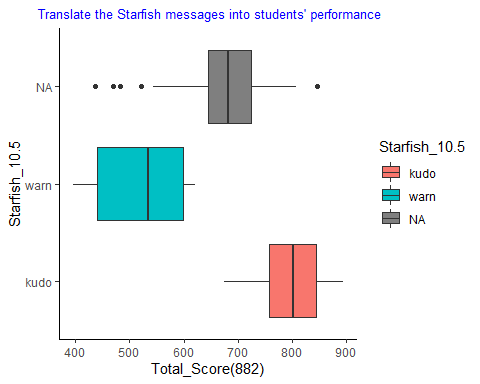
ggplot(Data, aes(`Total\_Score(882)`,StarFish\_9.12, fill= StarFish\_9.12)) + geom\_boxplot()+ labs(title=" Translate the Starfish messages into students' performance") + theme(plot.title = element\_text(hjust = 0.5,size = 10, color = "darkgreen"))



# Translate the Starfish messages(Starfish\_10.5) into students' overall performance in the course.   
ggplot(Data, aes(`Total\_Score(882)`,Starfish\_10.5, fill= Starfish\_10.5)) + geom\_violin()+ labs(title=" Translate the Starfish messages into students' performance") + theme(plot.title = element\_text(hjust = 0.5,size = 10, color = "black"))



ggplot(Data, aes(`Total\_Score(882)`,Starfish\_10.5, fill= Starfish\_10.5)) + geom\_boxplot()+ labs(title=" Translate the Starfish messages into students' performance") + theme(plot.title = element\_text(hjust = 0.5,size = 10, color = "blue"))



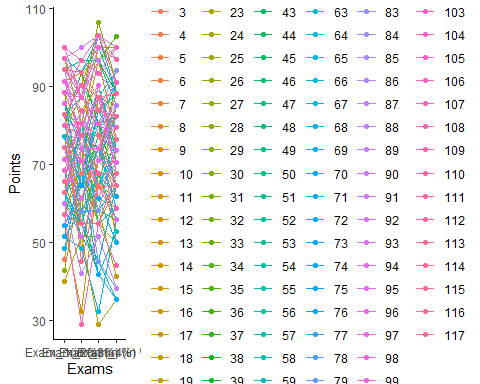
# Task-10: How did students’ performance change over time, if at all? Visualize any two of the following variables over time: (a) Team reports, (b) MC quizzes, (c) MC homework, and (d) exams. Fun bonus: Did students who completed the SELFI perform better or worse in the course than students who did not complete the SELFI?

## Answer: To visualize student’s performance change over time I cosider to use exams, and MC quizzes. (i) Exams: Instead of raw score I am considering each exam score in the percentage value to compare exam score in the same scale. This helps in true comparison. I create a new dara frame with all students and all four exams. All 117 students performance change in this line graph make this line graph visually ugly! I need to reduce students number to draw a more understandable chart. So, I create another data frame but this time consider only 8 students using filter() option. Now, the line graph make more sense and shows the change on exams over time. (ii) MC Quizzes: Draw a line graph to see students(8) performance change over time on the all MC quizzes. This graph is ugly with 8 student points over time. It’s even complex with only 2 students!! So let’s stick with only one student in this case to visualize performance on MC quizzes over time.

# (i) students’ performance change over time with respect to exams  
  
# Create a new data frame with all students and all four exams. I am considering each exam score in the percentage value to compare exam score in the same scale. This helps in true comparison.  
Data\_perform\_all <- select(Data, Student,`Exam\_1(in %)`, `Exam\_2(in %)`, `Exam\_3(in %)`,`Exam\_4(in %)`)  
#view(Data\_perform\_all)  
  
# Draw a line graph to see students(all) performance change over time on the all 4 exams. This chart is ugly!, I need to reduce students number to draw a more understandable chart.   
Data\_perform\_all %>%  
 mutate(row\_number = as.factor(row\_number())) %>%  
 gather(Exams, Points, `Exam\_1(in %)`:`Exam\_4(in %)`) %>%  
 ggplot(aes(x = Exams, y = Points, group = row\_number, color = row\_number)) +  
 geom\_line() + geom\_point() + labs(fill = "Student")

## Warning: Removed 1 row(s) containing missing values (geom\_path).

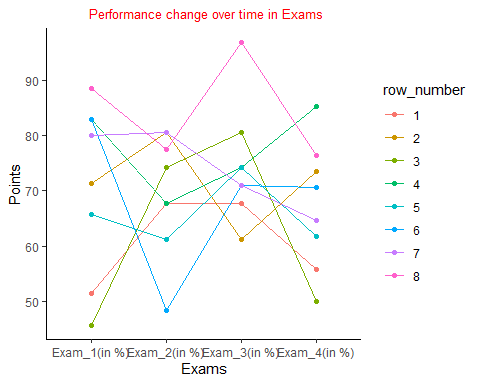
## Warning: Removed 2 rows containing missing values (geom\_point).



# Create a new data frame with only first 10 students and all four exams.  
Data\_perform\_eight <- filter(Data, Student== c(1:8)) %>% select(Student,`Exam\_1(in %)`, `Exam\_2(in %)`, `Exam\_3(in %)`,`Exam\_4(in %)`)

## Warning in Student == c(1:8): longer object length is not a multiple of shorter  
## object length

#view(Data\_perform\_ten)  
  
# Draw a line graph to see students(8) performance change over time on the all 4 exams. This make sense!  
Data\_perform\_eight %>%  
 mutate(row\_number = as.factor(row\_number())) %>%  
 gather(Exams, Points, `Exam\_1(in %)`:`Exam\_4(in %)`) %>%  
 ggplot(aes(x = Exams, y = Points, group = row\_number, color = row\_number)) +  
 geom\_line() + geom\_point() + labs(fill = "Student")+ labs(title=" Performance change over time in Exams") + theme(plot.title = element\_text(hjust = 0.5,size = 10, color = "red"))



# (ii) students’ performance change over time with respect to MC Quizzes  
  
# Create a new data frame with only first 10 students and all MC Quizzes.  
Data\_perform\_one\_Quiz <- filter(Data, Student== c(1:1)) %>% select(Student, `Quiz1(1.5)`:`Quiz27(1.5)`)  
#view(Data\_perform\_one\_Quiz)  
  
# Draw a line graph to see students(10) performance change over time on the all MC quizzes. This graph is ugly with 10 student points over time. It's even complex with only 2 students!! So let's stick with only one student in this case! Change the angle of the x-axis text.  
Data\_perform\_one\_Quiz %>%  
 mutate(row\_number = as.factor(row\_number())) %>%  
 gather(MC\_Quizzes, Points, `Quiz1(1.5)`:`Quiz27(1.5)`) %>%  
 ggplot(aes(x = MC\_Quizzes, y = Points, group = row\_number, color = row\_number)) +  
 geom\_line() + geom\_point() + labs(fill = "Student")+ labs(title=" Performance Change Over Time in MC Quizzes") + theme(plot.title = element\_text(hjust = 0.5,size = 10, color = "blue"),axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1))

