DataAssignment\_10

Md. Saifur Rahman

3/28/2021

# Task-1: Load the Tophat2 dataset.

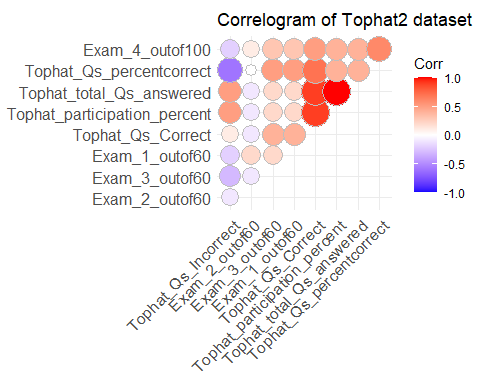
library(readxl)  
Data <- read\_excel("C:/Users/mdsaifur.rahman.1/Desktop/Spring 2021/EFR 535/Code R/Tophat data2 - April 15 2019.xls",   
 na = "NA")  
View(Data)  
attach(Data)

# Task-2: Create a correlogram to examine the correlations among the following Tophat variables: Tophat\_participation\_percent, Tophat\_total\_Qs\_answered, Tophat\_Qs\_Correct, Tophat\_Qs\_Incorrect, Tophat\_Qs\_percentcorrect, Exam\_1\_outof60, Exam\_2\_outof60, Exam\_3\_outof60, Exam\_4\_outof100. Identify two correlations (1 meduim/large positive, 1 medium/large negative) to investigate further.

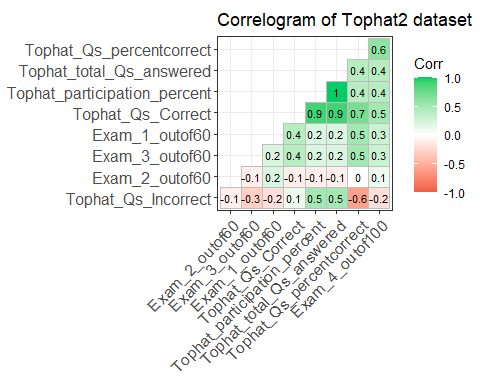
## Identify two correlations (1 meduim/large positive, 1 medium/large negative) to investigate further.

## Answer: From the square shaped correlogram it has been found that there is a strong positive correlation (0.9) in between “Tophat\_Qs\_Correct” and “Tophat\_total\_Qs\_answered”. So, most of the answer is correct according to this dataset. On the other hand, because of same reason there is a large negative correlation in between “Tophat\_Qs\_Incorrect” and “Tophat\_total\_Qs\_answered”.

# Create subset with continuous variable form the original dataset  
Data\_cor <- subset(Data, select = c("Tophat\_participation\_percent", "Tophat\_total\_Qs\_answered", "Tophat\_Qs\_Correct", "Tophat\_Qs\_Incorrect", "Tophat\_Qs\_percentcorrect", "Exam\_1\_outof60", "Exam\_2\_outof60", "Exam\_3\_outof60", "Exam\_4\_outof100"))  
  
# Round the data and consider 1 value after decimal point for correlation  
Data\_cor\_round <- round(cor(Data\_cor),1)  
  
# Draw Correlogram with circle shape  
ggcorrplot(Data\_cor\_round, hc.order = TRUE, type = "upper", lab = FALSE, lab\_size = 3, method = "circle", title = "Correlogram of Tophat2 dataset")



# Draw Correlogram with square shape. Show the label values.  
ggcorrplot(Data\_cor\_round, hc.order = TRUE, type = "lower", lab = TRUE, lab\_size = 3, method = "square", title = "Correlogram of Tophat2 dataset", colors = c("tomato2","white","springgreen3"),ggtheme = theme\_bw)

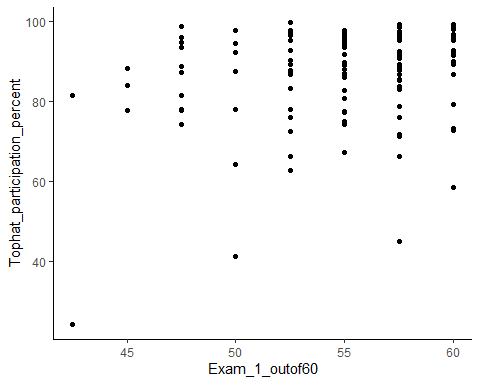


# Task-3: Examine the relationship between Tophat\_participation\_percent and Exam\_1\_outof60 with a scatterplot. What issue is this plot having and how can you resolve it to better visualize the relationship between these variables?

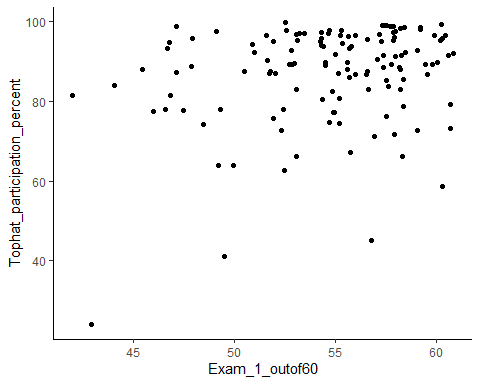
## What issue is this plot having and how can you resolve it to better visualize the relationship between these variables?

## Answer: There is a lots of overlap in this scatterplot. To better understand this scatterplot I am using jittter option to jitter the position of the data points but still need more work. Then, I color those data points exam points (Exam\_1\_outof60) which make this more easy to understand. Add a title to dercribe the graph.

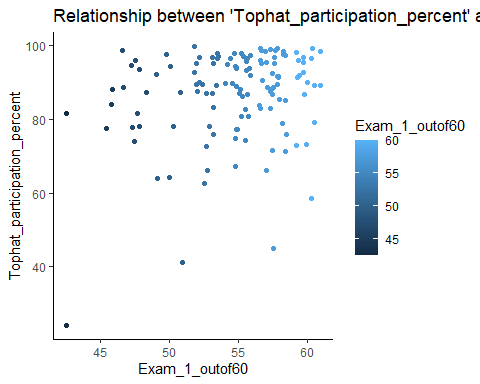
# Examine the relationship between Tophat\_participation\_percent and Exam\_1\_outof60 with a scatterplot.  
ggplot(Data\_cor, aes(Exam\_1\_outof60,Tophat\_participation\_percent))+geom\_point()



# jitter the data help to better visualize the points but still, need some work!  
ggplot(Data\_cor, aes(Exam\_1\_outof60,Tophat\_participation\_percent))+geom\_point(position = "jitter")



# Add "color" attributes to make the visualization more easy to visualize. Add a title.  
ggplot(Data\_cor, aes(Exam\_1\_outof60,Tophat\_participation\_percent, color=Exam\_1\_outof60))+geom\_point(position = "jitter") + labs(title = "Relationship between 'Tophat\_participation\_percent' and 'Exam\_1\_outof60'")



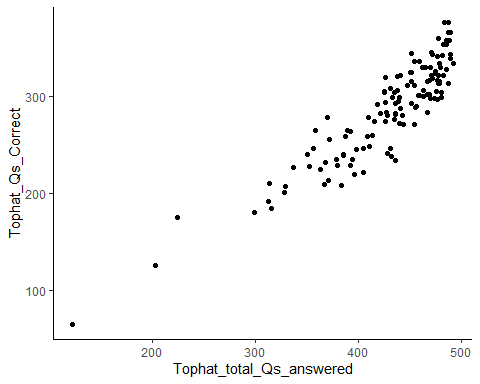
# Try this but not so good for this data!  
#ggplot(Data\_cor, aes(Exam\_1\_outof60,Tophat\_participation\_percent))+geom\_count(color="blue")

# Task-4: Using your identified variables that are positively correlated, create a scatterplot. Include in your plot a regression line, standard error confidence band, and correlation coefficient. Make this into an object and create a marginal density plot. What did you learn about these variables and their relationship that you didn’t know before creating the plot?

## What did you learn about these variables and their relationship that you didn’t know before creating the plot?

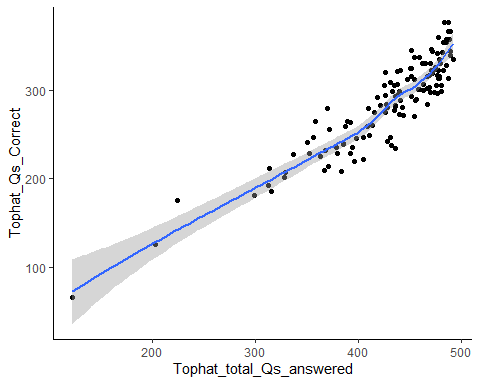
## Answer: I come to know that total number of correct answers(Tophat\_Qs\_Correct) increases with total number of answers (Tophat\_total\_Qs\_answered) given by students.

# Using your identified variables that are positively correlated, create a scatterplot.  
ggplot(Data\_cor, aes(Tophat\_total\_Qs\_answered,Tophat\_Qs\_Correct))+geom\_point()



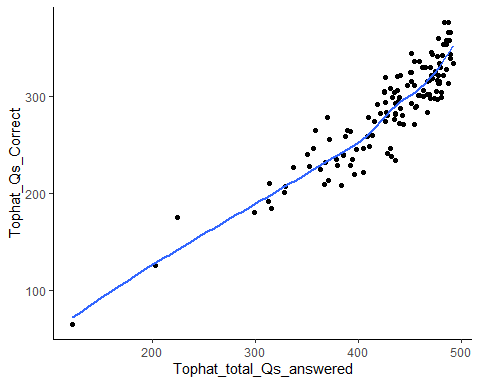
# Add a regression line.  
ggplot(Data\_cor, aes(Tophat\_total\_Qs\_answered,Tophat\_Qs\_Correct))+geom\_point()+ geom\_smooth()

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



# Remove standard error confidence band by setting se = FALSE.  
ggplot(Data\_cor, aes(Tophat\_total\_Qs\_answered,Tophat\_Qs\_Correct))+geom\_point()+ geom\_smooth( se = FALSE)

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

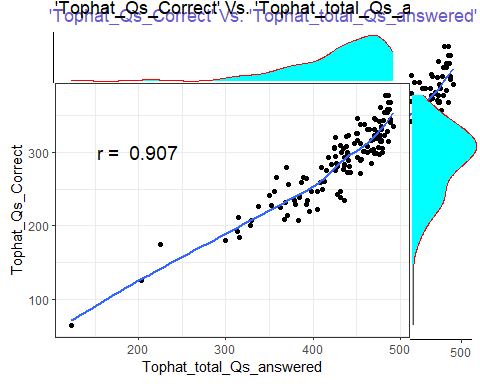


# Add correlation coefficient. Add a title , center the title and change color.   
ggplot(Data\_cor, aes(Tophat\_total\_Qs\_answered,Tophat\_Qs\_Correct))+geom\_point()+geom\_smooth(se = FALSE)+ annotate(200,300, label=paste("r = ", round(cor(Data\_cor$Tophat\_total\_Qs\_answered,Data\_cor$Tophat\_Qs\_Correct),3)),geom = "text", size=5) + ggtitle("'Tophat\_Qs\_Correct' Vs. 'Tophat\_total\_Qs\_answered'")+theme(plot.title = element\_text(hjust = 0.5, color = "slateblue"))

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

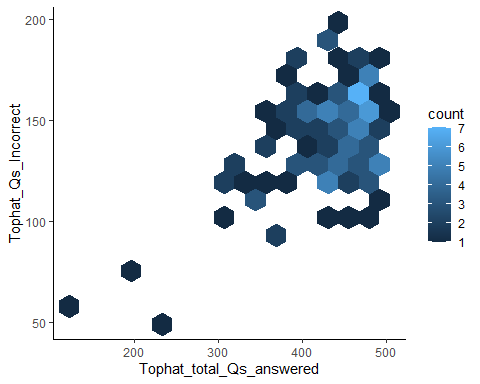
# Make this into an object and create a marginal density plot.  
# Create an object to reuse in creating marginal density plot.  
Plot2object <- ggplot(Data\_cor, aes(Tophat\_total\_Qs\_answered,Tophat\_Qs\_Correct))+geom\_point()+geom\_smooth(se = FALSE)+ annotate(200,300, label=paste("r = ", round(cor(Data\_cor$Tophat\_total\_Qs\_answered,Data\_cor$Tophat\_Qs\_Correct),3)),geom = "text", size=5)+ ggtitle("'Tophat\_Qs\_Correct' Vs. 'Tophat\_total\_Qs\_answered'")+theme(plot.title = element\_text(hjust = 0.5, color = "slateblue"))+theme\_bw()  
  
# Draw marginal density plot using the 'ggMargin()' function from 'ggExtra' package.  
ggMarginal(Plot2object,type="density", color= "red", fill = "cyan")

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

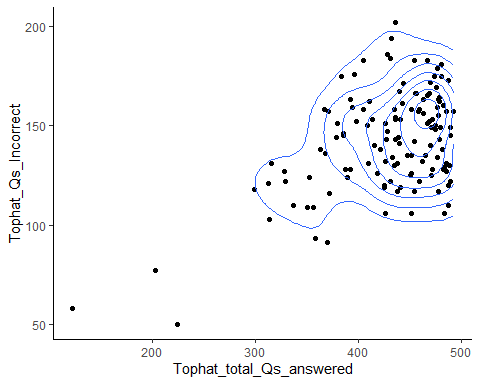


# Task-5: Create 2 different charts to show the associations among your two negatively correlated variables: Count chart, binned scatterplot square or hex, countours, 2d density.

# Binned scatterplot with hex shape  
ggplot(Data\_cor, aes(Tophat\_total\_Qs\_answered,Tophat\_Qs\_Incorrect))+ geom\_hex(bins = 15)

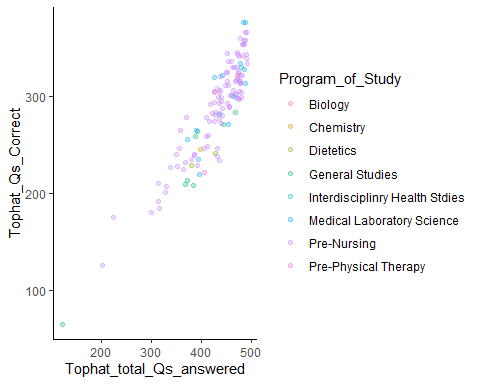


# Scatterplot with 2d density  
ggplot(Data\_cor, aes(Tophat\_total\_Qs\_answered,Tophat\_Qs\_Incorrect))+geom\_point()+geom\_density2d()



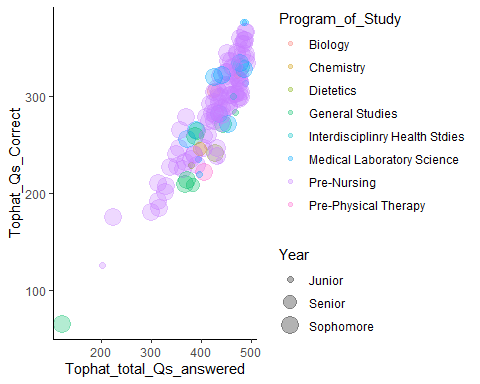
# Task-6: Use the variables Program\_of\_Study2 and/or Year, create 2 plots that cross numeric variables while also identifying student groups according to two aesthetics or plots of your choosing: color, shape, transparency, size, matrix, ellipses, bubbles.

# Bubble chart  
# (3 variables) Consider coloring by the program of study group (color=Program\_of\_Study)  
ggplot(Data, aes(Tophat\_total\_Qs\_answered,Tophat\_Qs\_Correct,color=Program\_of\_Study))+geom\_point(alpha=0.3)

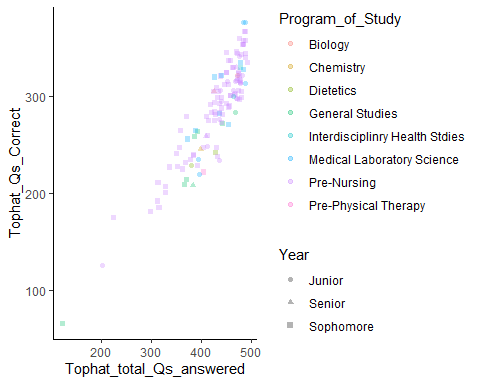


# (4 variables) Bubble chart with two continuous variable "Tophat\_total\_Qs\_answered" and "Tophat\_Qs\_Correct" while grouping by two categorical variables "Program\_of\_Study" for coloring and "Year" for diferent size.   
ggplot(Data, aes(Tophat\_total\_Qs\_answered,Tophat\_Qs\_Correct,color=Program\_of\_Study,size=Year))+geom\_point(alpha=0.3)

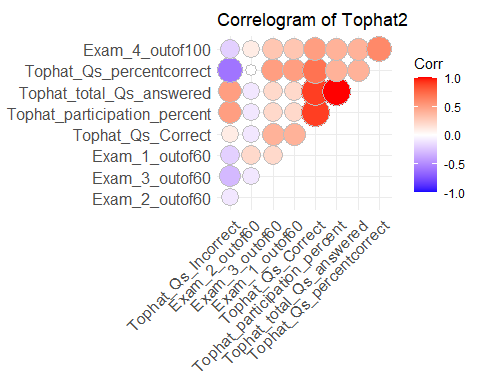
## Warning: Using size for a discrete variable is not advised.



# (4 variables) Bubble chart with two continuous variable "Tophat\_total\_Qs\_answered" and "Tophat\_Qs\_Correct" while grouping by two categorical variables "Program\_of\_Study" for coloring and "Year" for diferent shape.   
ggplot(Data, aes(Tophat\_total\_Qs\_answered,Tophat\_Qs\_Correct,color=Program\_of\_Study,shape=Year))+geom\_point(alpha=0.3)



# (Many variables) Correlation among many variable. Here, color range means positivity or negativity of correlation.   
ggcorrplot(Data\_cor\_round, hc.order = TRUE, type = "upper", lab = FALSE, lab\_size = 3, method = "circle", title = "Correlogram of Tophat2")



# Correlation among several variable. 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   
## 1. Tophat\_participation\_percent 0.50 0.39   
##   
## 2. Tophat\_total\_Qs\_answered 0.50 0.39 1.00\*\*   
## [1.00, 1.00]   
##   
## 3. Tophat\_Qs\_Correct 0.53 0.38 .85\*\* .85\*\*   
## [.42, .97] [.42, .97]   
##   
## 4. Tophat\_Qs\_Incorrect 0.08 0.50 .53 .53   
## [-.20, .88] [-.20, .88]   
##   
## 5. Tophat\_Qs\_percentcorrect 0.39 0.46 .15 .15   
## [-.57, .74] [-.57, .74]   
##   
## 6. Exam\_1\_outof60 0.31 0.32 -.20 -.20   
## [-.76, .54] [-.76, .54]   
##   
## 7. Exam\_2\_outof60 0.09 0.36 -.70\* -.70\*   
## [-.93, -.06] [-.93, -.06]  
##   
## 8. Exam\_3\_outof60 0.27 0.37 -.02 -.02   
## [-.67, .65] [-.67, .65]   
##   
## 9. Exam\_4\_outof100 0.38 0.33 .19 .19   
## [-.54, .76] [-.54, .76]   
##   
## 3 4 5 6 7 8   
##   
##   
##   
##   
##   
##   
## .01   
## [-.66, .67]   
##   
## .65 -.75\*   
## [-.03, .92] [-.94, -.18]   
##   
## .21 -.64 .64   
## [-.53, .77] [-.91, .05] [-.04, .91]   
##   
## -.67\* -.28 -.22 .11   
## [-.92, -.01] [-.80, .47] [-.77, .52] [-.60, .72]   
##   
## .40 -.64 .72\* .29 -.37   
## [-.36, .84] [-.91, .05] [.11, .94] [-.46, .80] [-.83, .39]   
##   
## .54 -.56 .80\*\* .38 -.19 .48   
## [-.20, .89] [-.89, .17] [.29, .96] [-.38, .83] [-.76, .54] [-.27, .87]  
##   
##   
## 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.  
##