ENGM 182 – Data Analytics - Homework #1

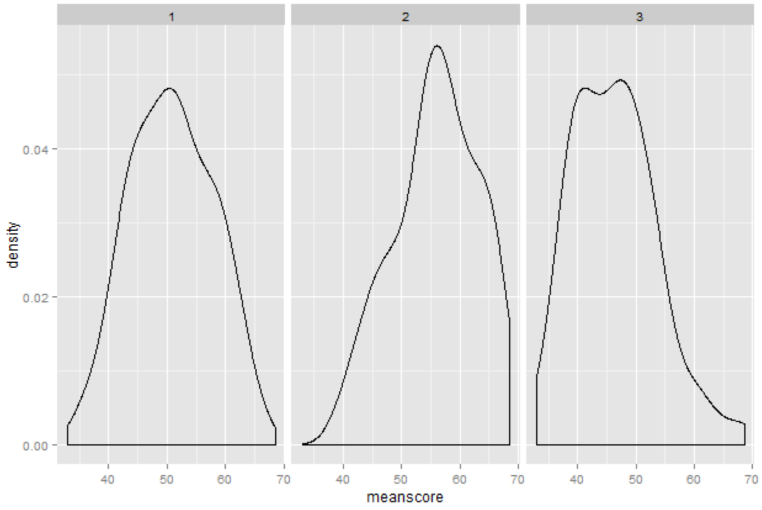
Due 5pm Tuesday April 2, 2019. Note – you are encouraged to work together to figure problems. However, each student should upload their own solutions to Canvas, preferably as a pdf file. Please identify who you worked with as part of your submission.

# **Useful resources for ggplot2 library in general:**
https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf
http://shinyapps.stat.ubc.ca/r-graph-catalog/

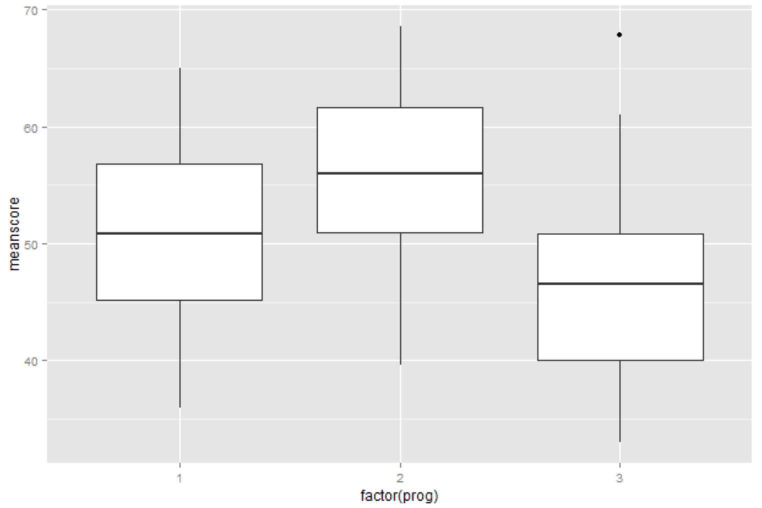
# **Install the hsb2.csv test data set from the UCLA tutorial used in class**

```
# Note, this dataset can be downloaded from Canvas.

# Q 1.1
# Create a new variable titled “meanscore” and add it to the data frame. This
should produce the # average score of all five tests for each observation
(row).

# See book chapter 4, part 2 (4.2) for hints if needed.

# Q 1.2
#Create a new variable titled “meancat” and add it to the data frame. Use the
following criteria:

meanscore<45 = “Low”
45<meanscore<60 = “Middle
meanscore>60 = “High”


# Q 1.3
#Sort the new data set from highest mean score to lowest call this “newdata”


# Q 1.4
#Notice you will have entries with NA in the meancat for scores of exactly 45
and 60. Delete these observations from your dataset. Call this “newdata2”

# Hint: the function na.omit(dataset) is of use for this.

# Q 1.5
# Convert “newdata2” to a set called “newdata3” that just includes the test
scores and the two new variables (meanscore and meancat) that you created for
each observation.
```

# Q 2.1
# From the data set d which should now include 13 variables create a density plot for the meanscores by program type. Your result should look like the following:

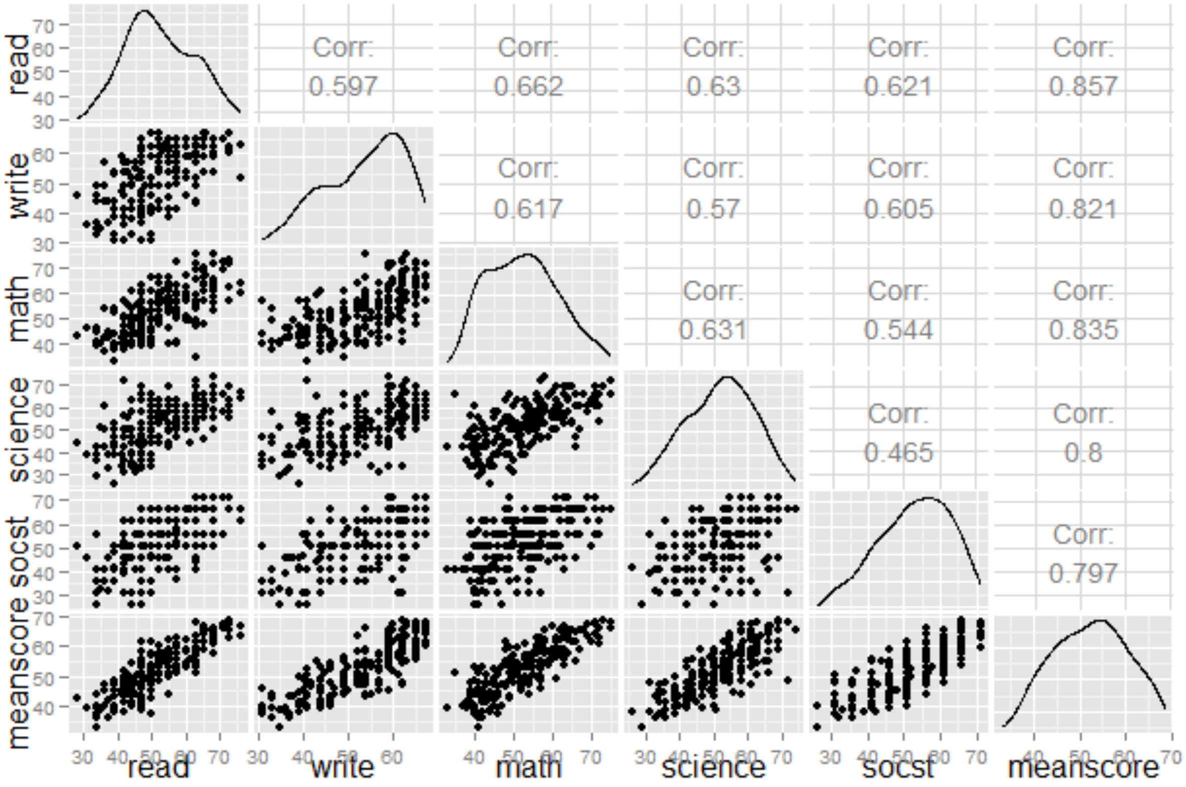


# Q2.2
# Create a boxplot from the same data for meanscore by program type. Your result should look like the following:



# Q2.3
# How many students make up the "High" "Middle" and "Low" categories? Hint these numbers will not total 200

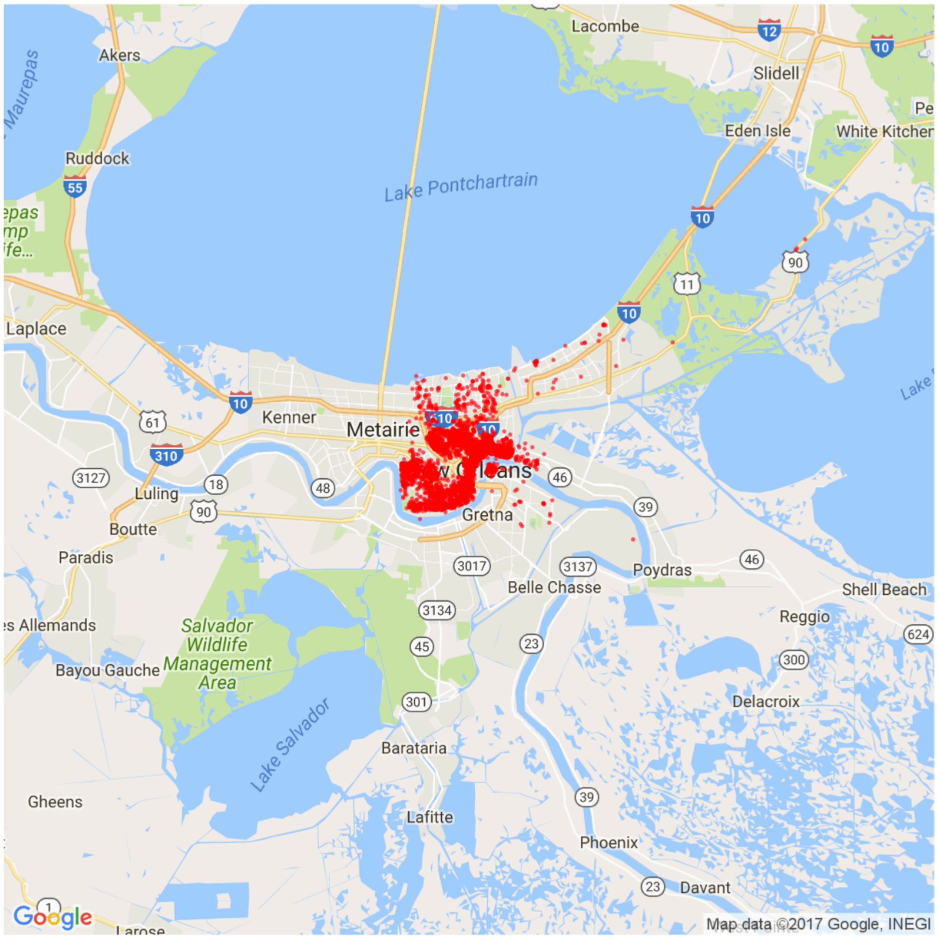
```
# Q2.4
# Recreate the following image:
```



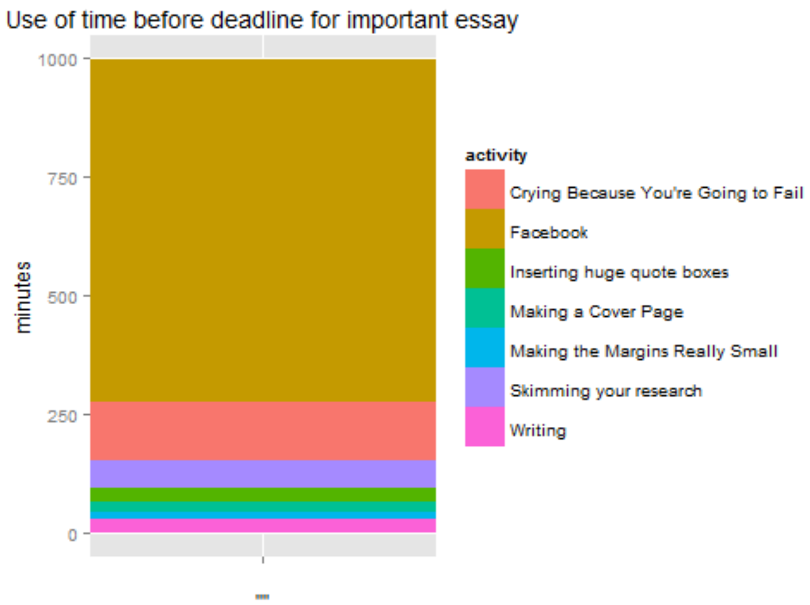
```
# Recall that you need to load GGally

library(GGally)
```

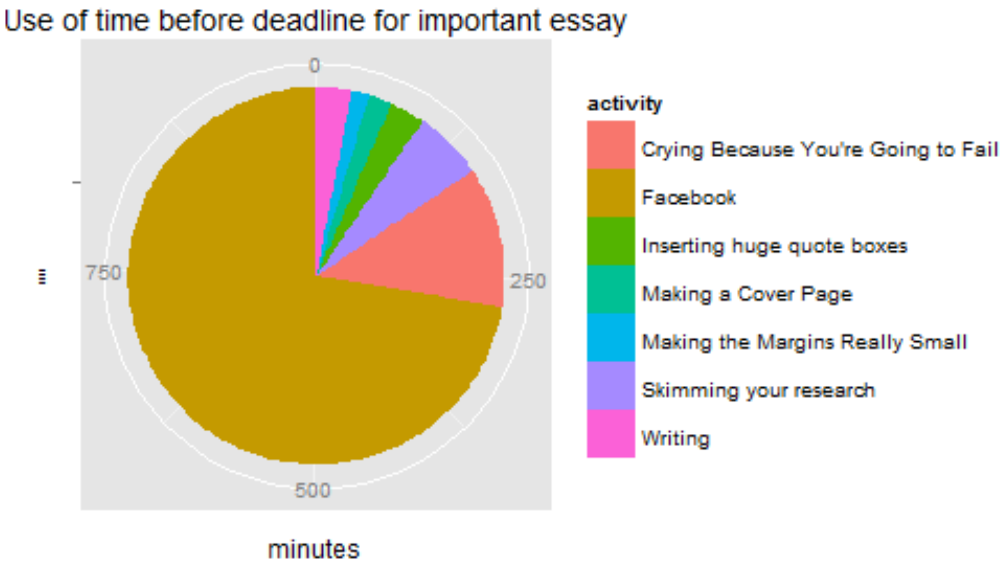
```
# download the file "NOLAlistingsJune2016.csv" from Canvas and get it into R.
# This file contains information on the Airbnb listings in New Orleans as of
June 2016 (source: http://insideairbnb.com/get-the-data.html).

#### Basic histogram and x,y plots ####

# Q3.1 Using the New Orleans Airbnb data, what is the mean, median, min, max
of price? Does this give you any information about how a chart will look?



# Q3.2 create a histogram of the prices with 10 intervals

# for example

hist(NOLAlistingsJune2016$price,10)

# what happened? Did you get a useful graphic?

# Q3.3 try increasing the number of intervals to 100 to see what you get.
# Any better?


# Q3.4 - If there is a long tail that is making the graph fail,
# then try dropping all of the prices above $1000 and redoing
# the plot with 20 intervals


# Q3.5 Now use the Airbnb longitude and latitude data to make a raw x,y plot
# of each listing.

# For example

plot(NOLAlistingsJune2016$longitude,NOLAlistingsJune2016$latitude)

# Q3.6 – Is there an outlier that is making the graph unappealing? If so,
drop it and redo the x,y graph.


# Q3.7 Now plot this data on a map. Read the following documents for an
overview of the ggmap package. (Credit to Professor Horiuchi in the
Government Department for flagging these.)

#for a quick summary:
#https://www.nceas.ucsb.edu/%7Efrazier/RSpatialGuides/ggmap/ggmapCheatsheet.p
df

#for more depth on how to use ggplot in mapping see the following:
```

```
#http://stat405.had.co.nz/ggmap.pdf

#Using these, try and recreate the something like the following image.
Experiment with different map sources/types. To remove axes, use:

…+theme(line = element_blank(),
        text = element_blank(),
        line = element_blank(),
        title = element_blank())
```



```
# Q4 - Reproduce these charts, using the same data for the bar chart and pie
chart (code to create the charts is on the next page):
```





```
# Code for the charts:

# you will need the graphics library ggplot2. Install and load it.

install.packages("ggplot2")
library(ggplot2)

#
df <- data.frame(activity = c("Writing", "Making the Margins Really Small", "Making a Cover Page", "Inserting huge quote boxes", "Skimming your research", "Crying Because You're Going to Fail", "Facebook"), minutes = c(30, 15, 20, 30, 60, 120, 720))

bp <- ggplot(df, aes(x="", y=minutes, fill=activity)) + geom_bar(width = 1, stat = "identity") + ggtitle("Use of time before deadline for important essay")

bp

pie <- bp + coord_polar("y", start=0)

pie
```

# Q 4.1 Change the colors of the bar and pie chart

# Q 5 Challenge Question

Imagine you work for a company which, for a current project, needs to analyze a shipment of diamonds which was recently delivered. Use the following commands in R to download the dataset, attach the dataset, and use the ‘head’ and ‘view’ commands to simply observe the data.

```
library(ggplot2)
data(diamonds)
attach(diamonds)
head(diamonds)
```

You can also research this dataset here:
https://ggplot2.tidyverse.org/reference/diamonds.html

Now imaging you are charged with the task of delivering a brief presentation to your team describing the diamond delivery. How do some variables change with others? What are the best visual ways to express the structure of this data? (Maybe a bar graph, maybe a density plot?) Use the ggplot package to generate a few plots to help you describe the new diamond delivery to your team.

A good place to start is simply by google searching “ggplot examples”!