DATA 606 Fall 2022 - Final Exam

Mathew Katz

Part I

Please put the answers for Part I next to the question number (please enter only the letter options; 4 points each):

1. B
2. A
3. D
4. B

5. B
6. E

7. D

8. E

9. B
10. C

Part II

Consider the three datasets, each with two columns (x and y), provided below. Be sure to replace the `NA` with your answer for each part (e.g. assign the mean of `x` for `data1` to the `data1.x.mean` variable). When you Knit your answer document, a table will be generated with all the answers.

For each column, calculate (to four decimal places):

```
data1.x.mean <- mean(data1$x)
data1.y.mean <- mean(data1$y)
data2.x.mean <- mean(data2$x)
data2.y.mean <- mean(data2$y)
data3.x.mean <- mean(data3$x)
data3.y.mean <- mean(data3$y)
```

a. The mean (for x and y separately; 5 pt).

```r
data1.x.median <- median(data1$x)
data1.y.median <- median(data1$y)
data2.x.median <- median(data2$x)
data2.y.median <- median(data2$y)
data3.x.median <- median(data3$x)
data3.y.median <- median(data3$y)
```

b. The median (for x and y separately; 5 pt).

```r
data1.x.sd <- sd(data1$x)
data1.y.sd <- sd(data1$y)
data2.x.sd <- sd(data2$x)
data2.y.sd <- sd(data2$y)
data3.x.sd <- sd(data3$x)
data3.y.sd <- sd(data3$y)
```

c. The standard deviation (for x and y separately; 5 pt).

For each x and y pair, calculate (also to two decimal places):

```r
data1.correlation <- cor(data1$x, data1$y)
data2.correlation <- cor(data2$x, data2$y)
data3.correlation <- cor(data3$x, data3$y)
```

d. The correlation (5 pt).

```r
lm1 = lm(formula = y~x, data = data1)
lm2 = lm(formula = y~x, data = data2)
lm3 = lm(formula = y~x, data = data3)
data1.slope <- lm1$coefficients[2]
data2.slope <- lm2$coefficients[2]
data3.slope <- lm3$coefficients[2]

data1.intercept <- lm1$coefficients[1]
data2.intercept <- lm2$coefficients[1]
data3.intercept <- lm3$coefficients[1]
```

e. Linear regression equation (5 points).

```
data1.rsquared <- data1.correlation ^ 2
data2.rsquared <- data2.correlation ^ 2
data3.rsquared <- data3.correlation ^ 2
```

f. R-Squared (5 points). Summary Table

	Data 1		Data 2		Data 3	
	x	y	x	y	x	y
Mean	54.2633	47.8323	54.2678	47.8359	54.2661	47.8347
Median	53.3333	46.0256	53.1352	46.4013	53.3403	47.5353
SD	16.7651	26.9354	16.7668	26.9361	16.7698	26.9397
r	-0.0645		-0.0690		-0.0641	
Intercept	53.4530		53.8497		53.4251	
Slope	-0.1036		-0.1108		-0.1030	
R-Squared	0.0042		0.0048		0.0041	

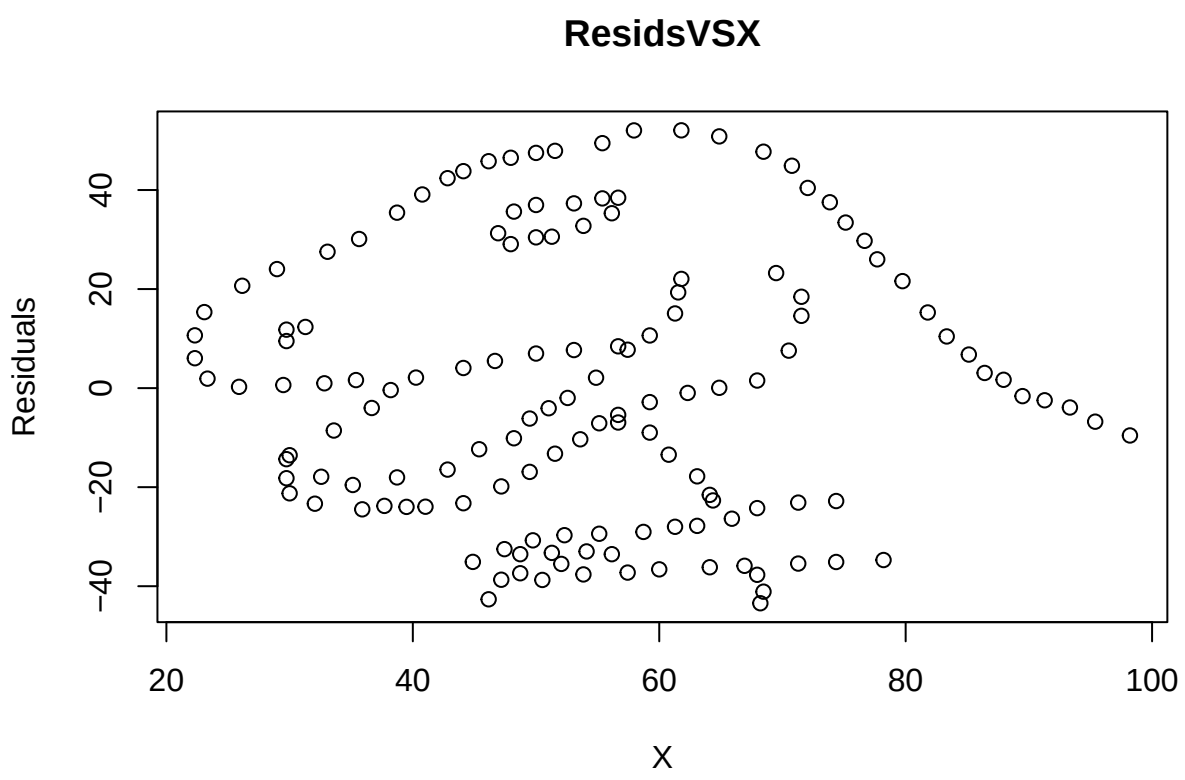
g. For each pair, is it appropriate to estimate a linear regression model? Why or why not? Be specific as to why for each pair and include appropriate plots! (15 points) Linear regression is an analysis that assesses whether one or more predictor variables explain the dependent (criterion) variable. The regression has five key assumptions:

-Linear relationship -Multivariate normality -No or little multicollinearity -No auto-correlation -Homoscedasticity

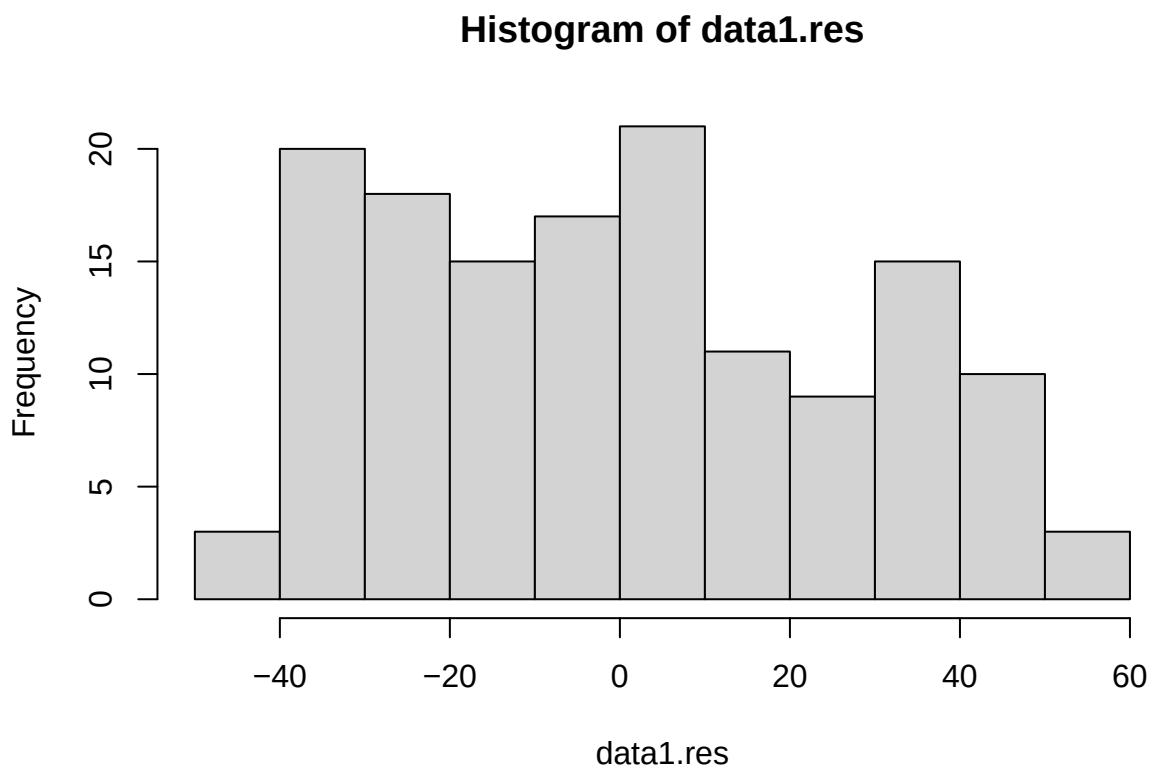
A linear relationship is a statistical term used to describe a straight-line relationship between two variables. All 3 datasets don't have a linear relationship. (See graphs below) A multivariate distribution describes the probabilities for a group of continuous random variables, particularly if the individual variables follow a normal distribution. Each variable has its own mean and variance. In this regard, the strength of the relationship between the variables (correlation) is very important. All 3 datasets have variables that have low correlation and don't follow a normal distribution. (See graphs below) Multicollinearity is a statistical concept where several independent variables in a model are correlated. That is not true for all 3 datasets. Linear regression model assumes that error terms are independent. This means that the error term of one observation is not influenced by the error term of another observation. In case it is not so, it is termed as autocorrelation. Using the Durbin-Watson Test we can find that in the first 2 models (first 2 datasets), the p-value is less than 0.05, we can reject the null hypothesis and conclude that the residuals in this regression model are autocorrelated. The assumption of homoscedasticity, states that the y population corresponding to various X values have the same variance i.e. it neither increases nor decreases as X increases. Using the Breusch-Pagan Test, we see that the p-value is not less than 0.05 in all 3 models, we fail to reject the null hypothesis. We do not have sufficient evidence to say that heteroscedasticity is present in the regression model. For each pair, is it NOT appropriate to estimate a linear regression model.

Checking Data1:

```
data1.res = resid(lm1)
plot(data1$x, data1.res,
     ylab="Residuals", xlab="X",
     main="ResidsVSX")
```




```
hist(data1.res)
```




```
data1.correlation
```

```
## [1] -0.06447
```

```
library(car)
```

```
## Loading required package: carData
```

```
durbinWatsonTest(lm1)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1          0.9442       0.06872       0
##  Alternative hypothesis: rho != 0
```

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

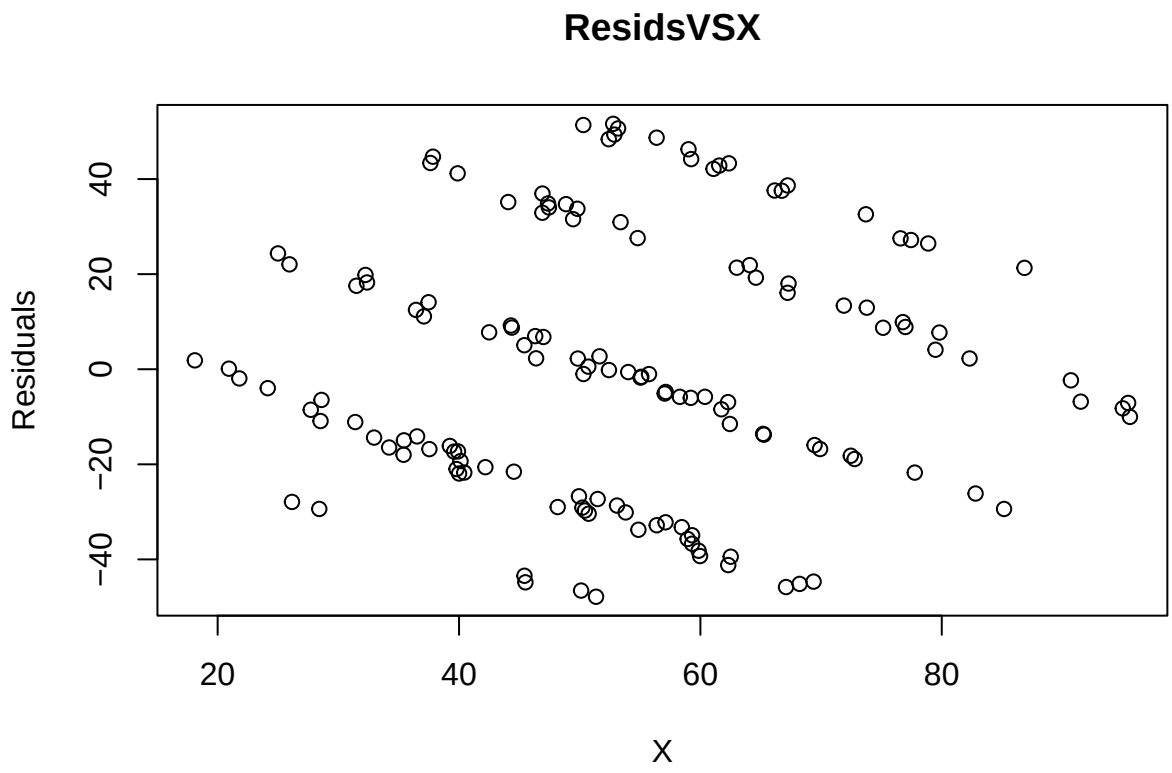
```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
bptest(lm1)
```

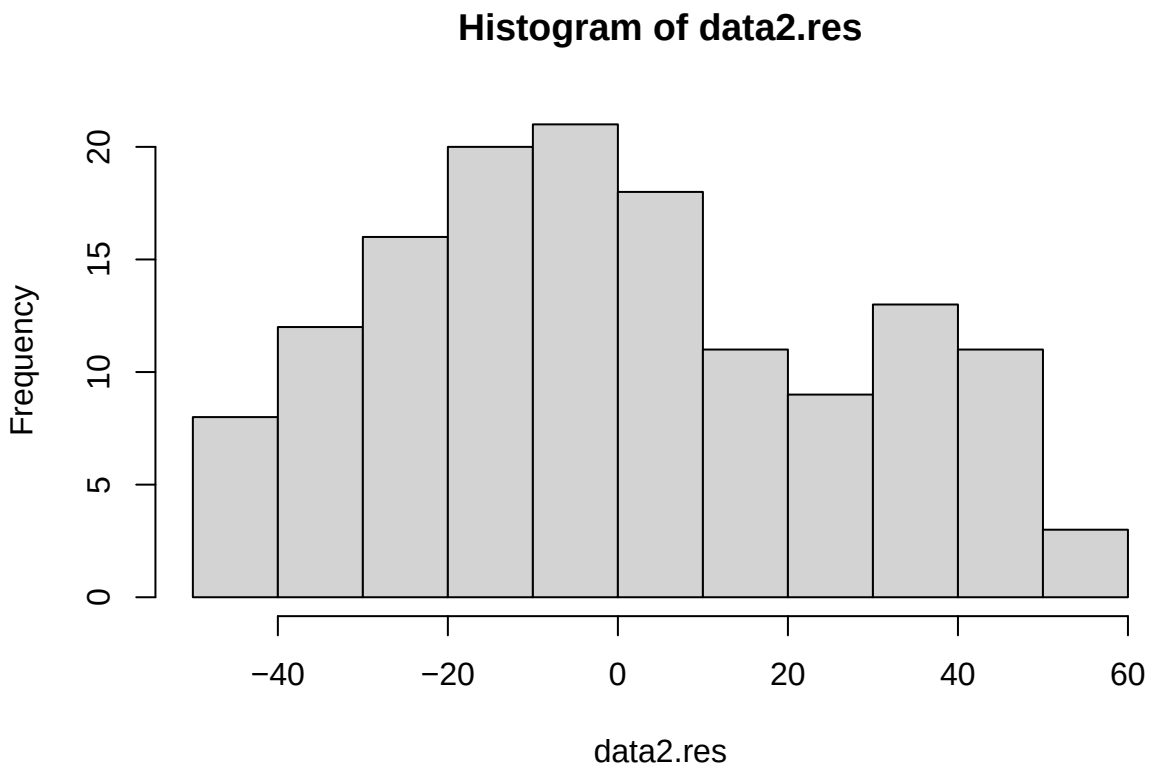
```
##
##  studentized Breusch-Pagan test
##
## data:  lm1
## BP = 0.84, df = 1, p-value = 0.4
```

Checking Data2:

```
data2.res = resid(lm2)
plot(data2$x, data2.res,
      ylab="Residuals", xlab="X",
      main="ResidsVSX")
```




```
hist(data2.res)
```




```
data2.correlation
```

```
## [1] -0.06898
```

```
durbinWatsonTest(lm2)
```

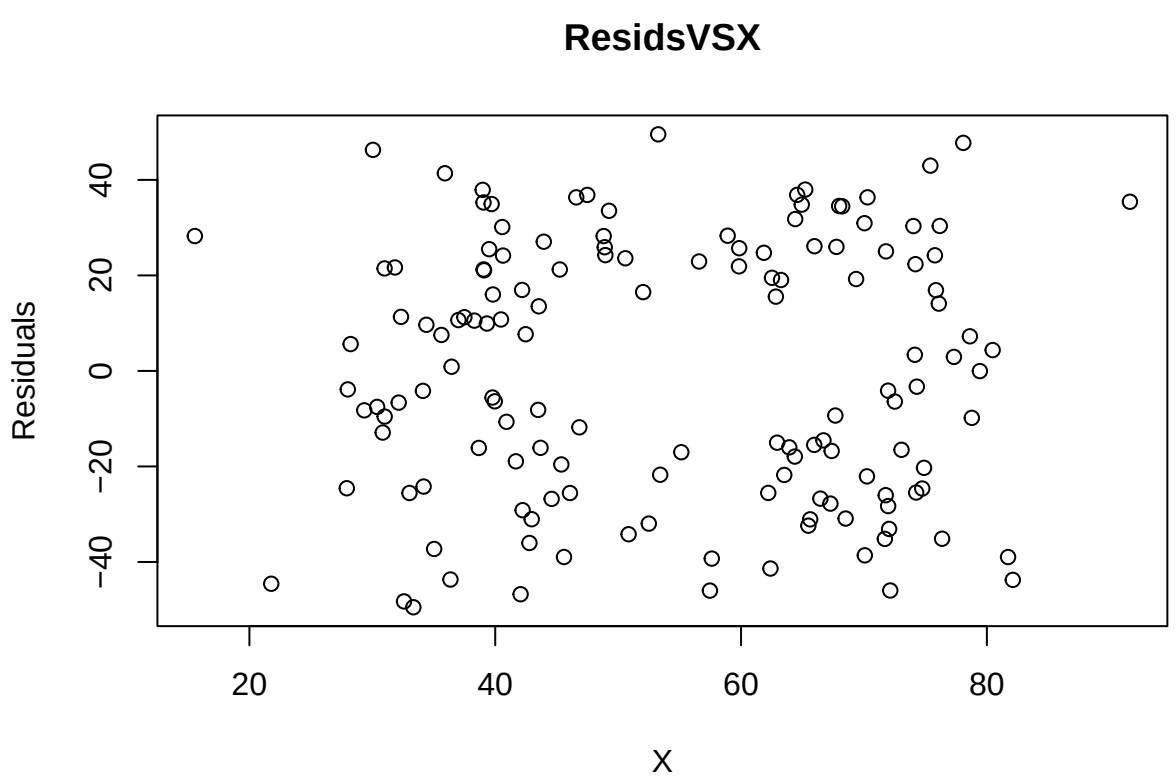
```
##  lag Autocorrelation D-W Statistic p-value
##    1          0.9065        0.1447       0
##  Alternative hypothesis: rho != 0
```

```
bptest(lm2)
```

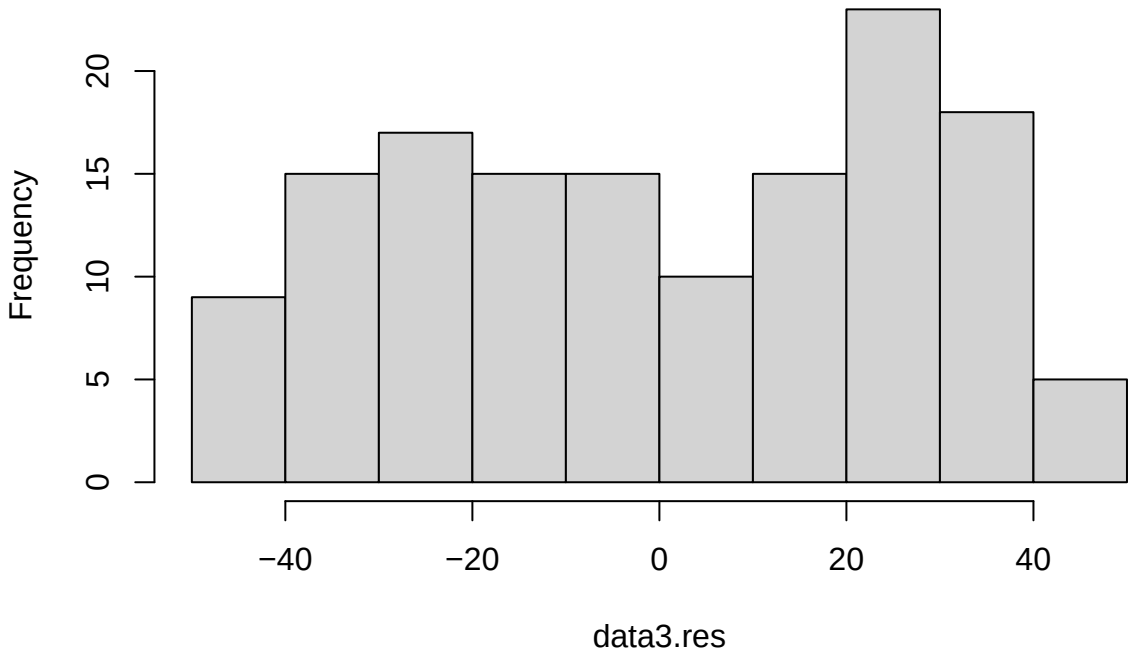
```
##
##  studentized Breusch-Pagan test
##
## data:  lm2
## BP = 0.12, df = 1, p-value = 0.7
```

Checking Data3:

```
data3.res = resid(lm3)
plot(data3$x, data3.res,
      ylab="Residuals", xlab="X",
      main="ResidsVSX")
```




```
hist(data3.res)
```




```
data3.correlation
```

```
## [1] -0.06413
```

```
durbinWatsonTest(lm3)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1        -0.07253         2.142    0.38
##  Alternative hypothesis: rho != 0
```

```
bptest(lm3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lm3
## BP = 0.59, df = 1, p-value = 0.4
```

h. Explain why it is important to include appropriate visualizations when analyzing data. Include any visualization(s) you create. (15 points) Data visualization is another form of visual art that grabs our interest and keeps our eyes on the message. The importance of data visualization is simple: it helps people see, interact with, and better understand data. Whether simple or complex, the right visualization can bring everyone on the same page, regardless of their level of expertise. When we see a

chart, we quickly see trends and outliers. If we can see something, we internalize it quickly. It's storytelling with a purpose. If you've ever stared at a massive spreadsheet of data and couldn't see a trend, you know how much more effective a visualization can be.

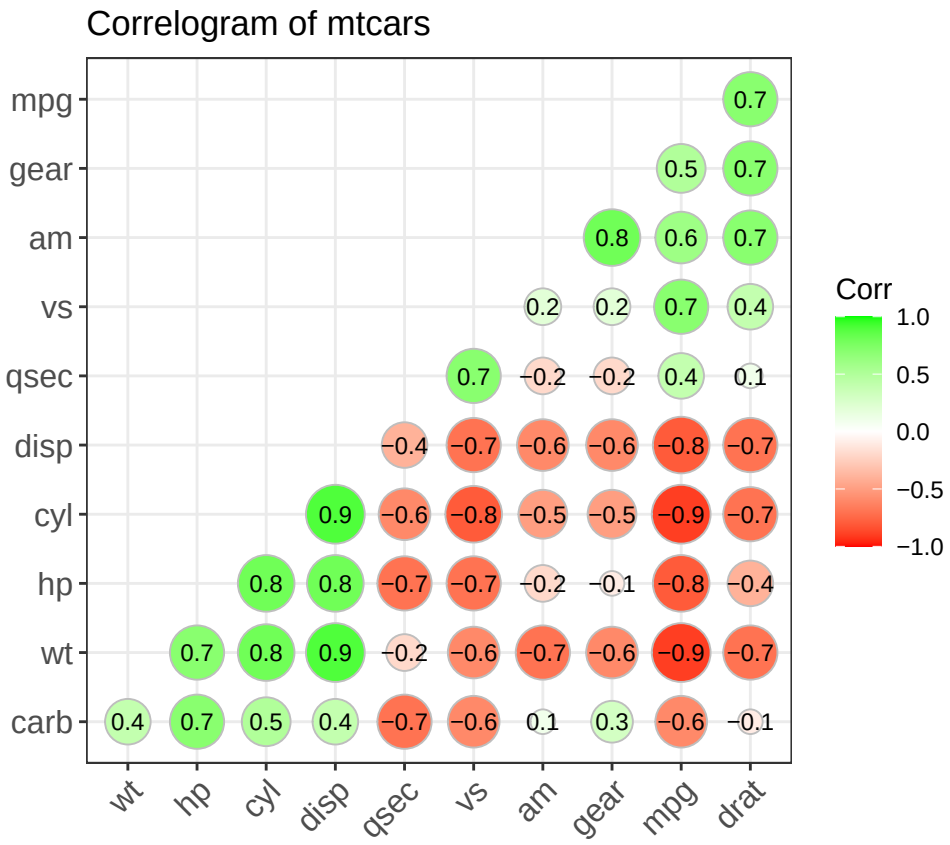
EXAMPLE: Let's look at the correlation between the variables in the dataset, 'mtcars.'

```
data(mtcars)
corr <- round(cor(mtcars), 1)
corr
```

```
##       mpg  cyl disp   hp drat   wt qsec   vs   am gear carb
## mpg   1.0 -0.9 -0.8 -0.8  0.7 -0.9  0.4  0.7  0.6  0.5 -0.6
## cyl  -0.9  1.0  0.9  0.8 -0.7  0.8 -0.6 -0.8 -0.5 -0.5  0.5
## disp -0.8  0.9  1.0  0.8 -0.7  0.9 -0.4 -0.7 -0.6 -0.6  0.4
## hp   -0.8  0.8  0.8  1.0 -0.4  0.7 -0.7 -0.7 -0.2 -0.1  0.7
## drat  0.7 -0.7 -0.7 -0.4  1.0 -0.7  0.1  0.4  0.7  0.7 -0.1
## wt   -0.9  0.8  0.9  0.7 -0.7  1.0 -0.2 -0.6 -0.7 -0.6  0.4
## qsec  0.4 -0.6 -0.4 -0.7  0.1 -0.2  1.0  0.7 -0.2 -0.2 -0.7
## vs    0.7 -0.8 -0.7 -0.7  0.4 -0.6  0.7  1.0  0.2  0.2 -0.6
## am    0.6 -0.5 -0.6 -0.2  0.7 -0.7 -0.2  0.2  1.0  0.8  0.1
## gear  0.5 -0.5 -0.6 -0.1  0.7 -0.6 -0.2  0.2  0.8  1.0  0.3
## carb -0.6  0.5  0.4  0.7 -0.1  0.4 -0.7 -0.6  0.1  0.3  1.0
```

When you look at that, you see a bunch of numbers... None of it really makes sense and you don't really know what's happening. Why use data visualization?? This is why:

```
library(ggplot2)
library(ggcorrplot)
ggcorrplot(corr, hc.order = TRUE,
           type = "lower",
           lab = TRUE,
           lab_size = 3,
           method="circle",
           colors = c("red", "white", "green"),
           title="Correlogram of mtcars",
           ggtheme=theme_bw)
```



It is now MUCH easier to understand the correlation between the different variables in the dataset.