# DATA 606 Fall 2022 - Final Exam

## Part I

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

1. C
2. A
3. D
4. E
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).**

```
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).**

```
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):**

```
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).**

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

data1.intercept <- lm(y ~ x, data = data1)$coefficients[[1]]
data2.intercept <- lm(y ~ x, data = data2)$coefficients[[1]]
data3.intercept <- lm(y ~ x, data = data3)$coefficients[[1]]
```

**e. Linear regression equation (5 points).**

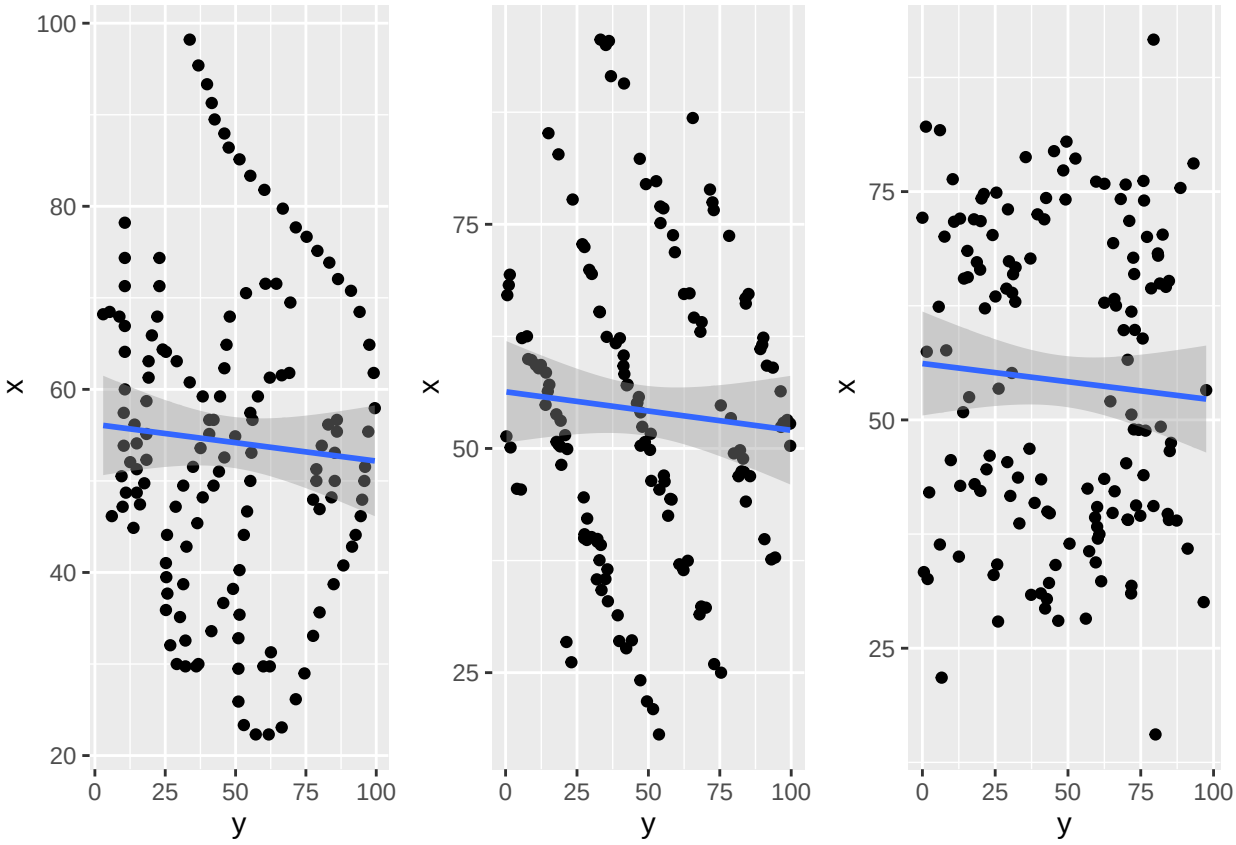
```
data1.rsquared <- summary(lm(y ~ x, data = data1))$r.squared
data2.rsquared <- summary(lm(y ~ x, data = data2))$r.squared
data3.rsquared <- summary(lm(y ~ x, data = data3))$r.squared
```

**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 | |

```
library(ggplot2)

p1 <- ggplot(data1, aes(y,x)) +
  geom_point() +
  geom_smooth(method = lm, formula = y ~ x)

p2 <- ggplot(data2, aes(y,x)) +
  geom_point() +
  geom_smooth(method = lm, formula = y ~ x)

p3 <- ggplot(data3, aes(y,x)) +
  geom_point() +
  geom_smooth(method = lm, formula = y ~ x)

cowplot::plot_grid(p1, p2, p3, ncol = 3)
```

**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 appears inappropriate for all three plots. In all three cases, the plots indicate that there is no linear relationship between x and y.

The first appears to be a dinosaur (hah!), and there is clearly no linear relationship.

The second appears as a series of lines. The pattern indicates that a number of linear relationships may exist between certain segments of the data. Depending on what the data represent, it may be appropriate

to segement the dataset and fit a series of linear models.

The third appears to be randomly scattered around zero, with no apparent pattern. I'm not aware of any model that might appropraitely represent such a relationship.

**h. Explain why it is important to include appropriate visualizations when analyzing data. Include any visualization(s) you create. (15 points)** Visualizations are an important part of confirming the validity of inferred conclusions or modeled relationships. They serve as a "sanity" check to reveal potential issues that may not be apparent from statistical tests / summary statistics alone.

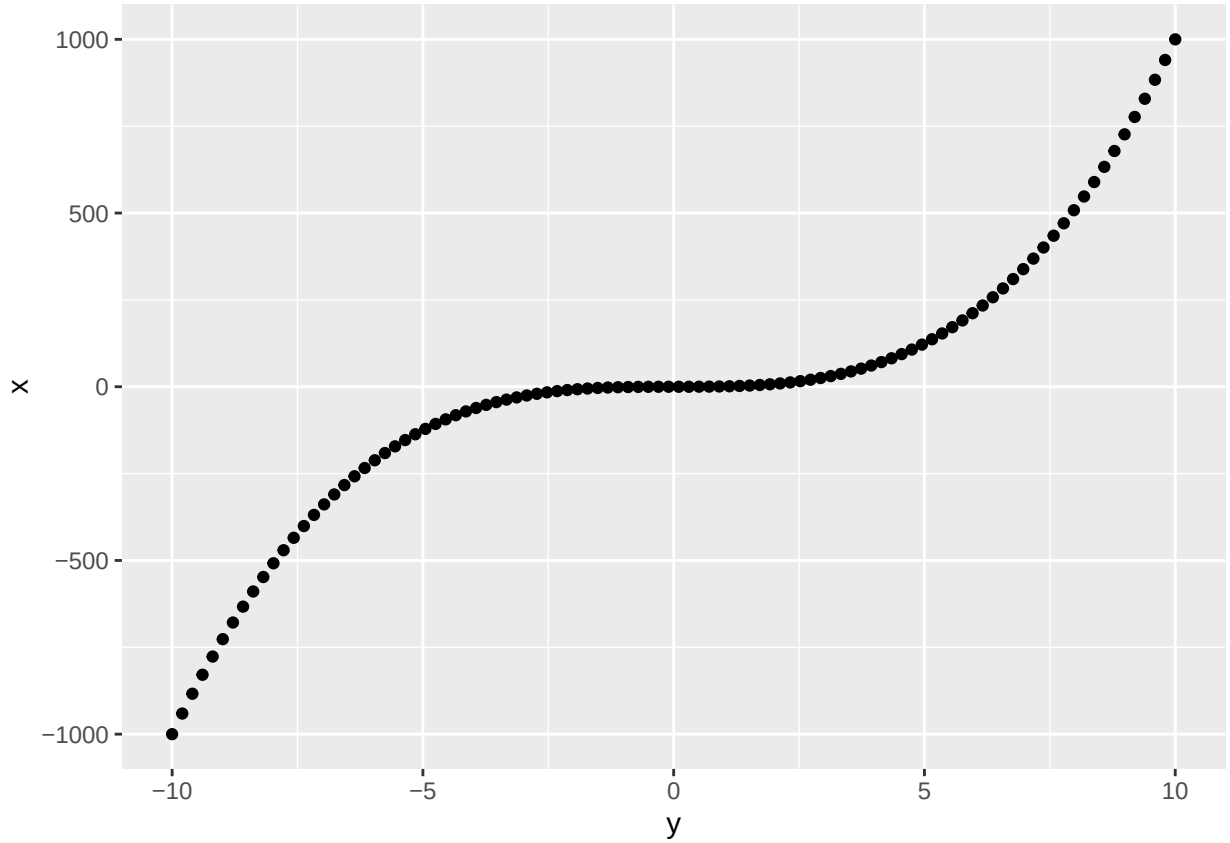
For example, if we use a cubic function to relate x to y, our linear regression results may appear to indicate a solid linear relationship.

```
y <- seq(-10, 10, length.out = 100)
x <- y^3
df <- data.frame(y = y, x = x)

summary(lm(y~x, df))
```

```
##
## Call:
## lm(formula = y ~ x, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -3.73  -2.31   0.00   2.31   3.73
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.59e-16   2.36e-01     0.0        1
## x           1.37e-02   6.05e-04    22.7   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.36 on 98 degrees of freedom
## Multiple R-squared:  0.84,   Adjusted R-squared:  0.838
## F-statistic:  515 on 1 and 98 DF,  p-value: <2e-16
```

The results above indicate a statistically significant relationship with a relatively high r-squared. If we plot that relationship, however, it becomes obvious that a regression equation with an exponential term would model the relationship much better.

```
ggplot(df, aes(y,x)) +
  geom_point()
```


1000
500
0
−500
−1000
x
−10
−5
0
5
10
y