DATA 606 Fall 2023 - Final Exam

```
options(digits = 4)
```

Part I

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

```
# This workspace exists to get answers for the below questions
# Q3
n1 <- 52
n2 <- 88
s1 <- 13
s2 <- 11
p1 <- s1 / n1
p2 <- s2 / n2

q3_se <- sqrt(
  (
    (p1 - p1^2) / (n1)
  ) +
    (
      (p2 - p2^2) / (n2)
    )
)

print(q3_se)
```

```
## [1] 0.06963
```

```
# Q5/6
# MOE = (z*)*(Standard Error = SE)
# SE = sqrt(
#  (p(1-p)) # nolint
#  /
#  n
#)
# CI = p +- 1.96 * SE # nolint
```

1. D
2. A
3. D
4. D
5. D

6. E
7. C
8. C
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 <- sprintf("%.4f", mean(data1$x))
data1.y.mean <- sprintf("%.4f", mean(data1$y))
data2.x.mean <- sprintf("%.4f", mean(data2$x))
data2.y.mean <- sprintf("%.4f", mean(data2$y))
data3.x.mean <- sprintf("%.4f", mean(data3$x))
data3.y.mean <- sprintf("%.4f", mean(data3$y))
```

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

```
data1.x.median <- sprintf("%.4f", median(data1$x))
data1.y.median <- sprintf("%.4f", median(data1$y))
data2.x.median <- sprintf("%.4f", median(data2$x))
data2.y.median <- sprintf("%.4f", median(data2$y))
data3.x.median <- sprintf("%.4f", median(data3$x))
data3.y.median <- sprintf("%.4f", median(data3$y))
```

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

```
data1.x.sd <- sprintf("%.4f", sd(data1$x))
data1.y.sd <- sprintf("%.4f", sd(data1$y))
data2.x.sd <- sprintf("%.4f", sd(data2$x))
data2.y.sd <- sprintf("%.4f", sd(data2$y))
data3.x.sd <- sprintf("%.4f", sd(data3$x))
data3.y.sd <- sprintf("%.4f", sd(data3$y))
```

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

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

```r
data1.correlation <- sprintf("%.4f", cor(data1$x, data1$y))
data2.correlation <- sprintf("%.4f", cor(data2$x, data2$y))
data3.correlation <- sprintf("%.4f", cor(data3$x, data3$y))
```

d. The correlation (5 pt).

```r
model1 <- lm(y ~ x, data = data1)
model2 <- lm(y ~ x, data = data2)
model3 <- lm(y ~ x, data = data3)

data1.slope <- sprintf("%.4f", coef(model1)["x"])
data2.slope <- sprintf("%.4f", coef(model2)["x"])
data3.slope <- sprintf("%.4f", coef(model3)["x"])

data1.intercept <- sprintf("%.4f", coef(model1)["(Intercept)"])
data2.intercept <- sprintf("%.4f", coef(model2)["(Intercept)"])
data3.intercept <- sprintf("%.4f", coef(model3)["(Intercept)"])
```

e. Linear regression equation (5 points).

```r
data1.rsquared <- sprintf("%.4f", summary(model1)$r.squared)
data2.rsquared <- sprintf("%.4f", summary(model2)$r.squared)
data3.rsquared <- sprintf("%.4f", summary(model3)$r.squared)
```

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

```
## Warning: package 'kableExtra' was built under R version 4.3.3
```

	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	

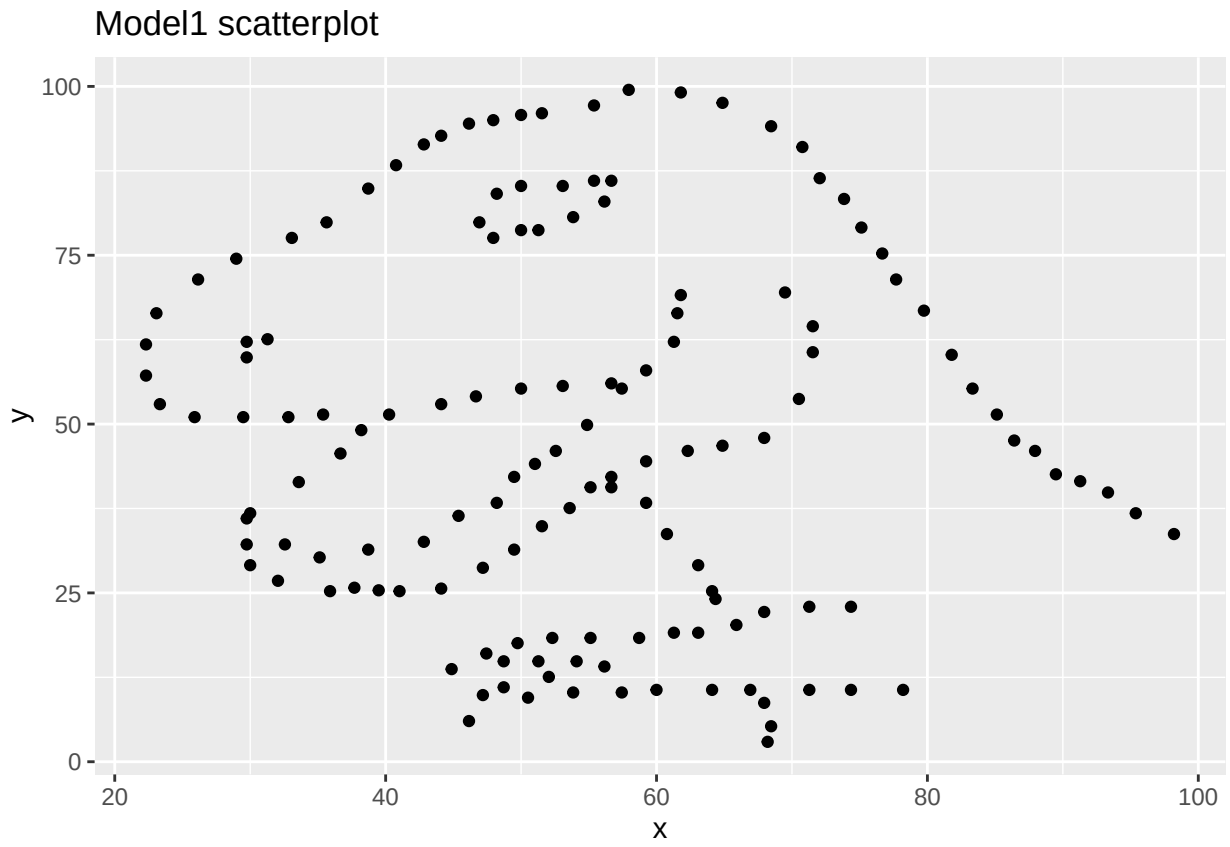
g. For each pair, is it appropriate to estimate a linear regression model? Why or why not? Be specific as to why for each pair and include appropriate plots! (15 points)

Data set 1 Yes or No

Why?

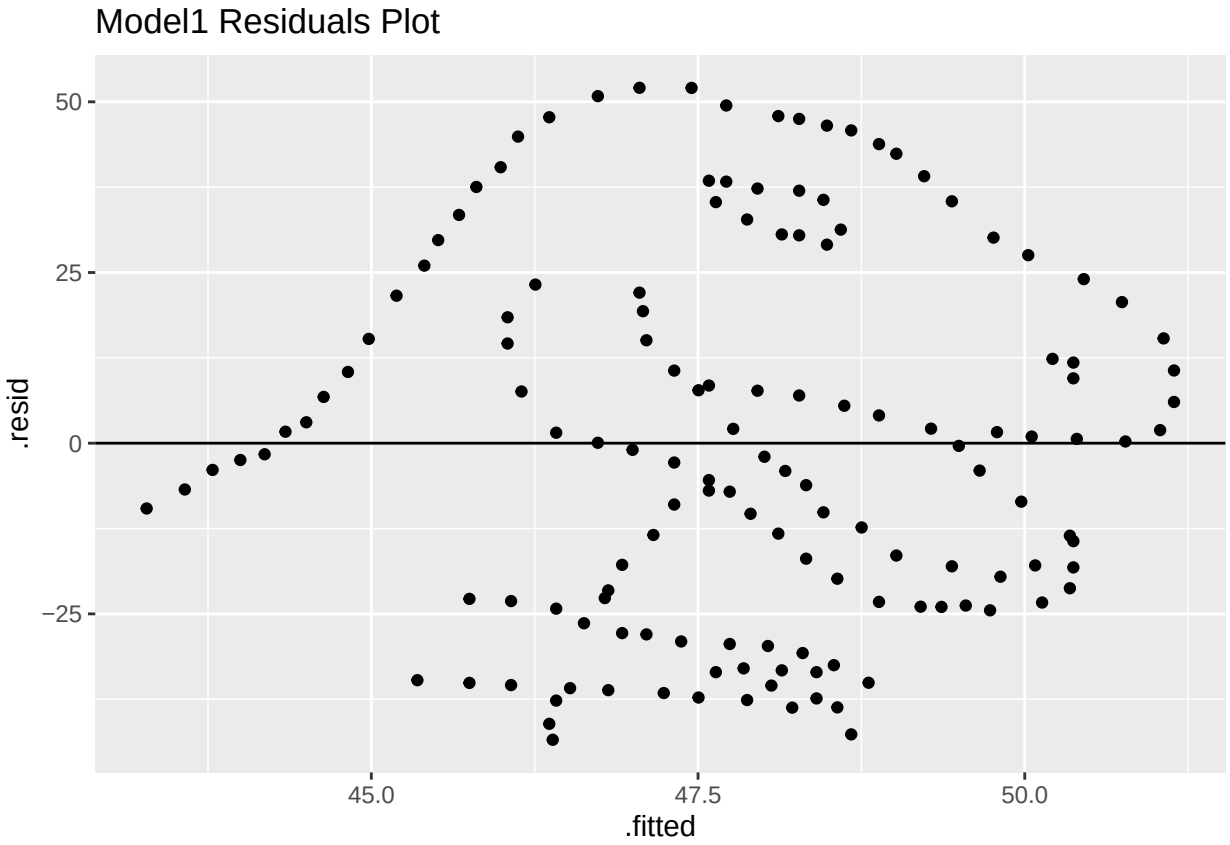
```
library(ggplot2)

ggplot(
  data1,
  aes(
    x = x,
    y = y
  )
) +
  geom_point() +
  labs(
    title = "Model1 scatterplot"
  )
```




```
ggplot(
  aes(
    .fitted,
    .resid
  ),
  data = model1
) +
  geom_point() +
```

```
  geom_hline(
    yintercept =  0
  ) +
  labs(
    title = "Model1 Residuals Plot"
  )
```




```
print(data1.correlation)
```

```
## [1] "-0.0645"
```

Looking visually, the data looks (like a dinosaur) as though there is no linear correlation here. We can also see that with a correlation coefficient of -0.0645 which is very close to 0. A score close to 0 would mean that there isn't linear relationship between the data.

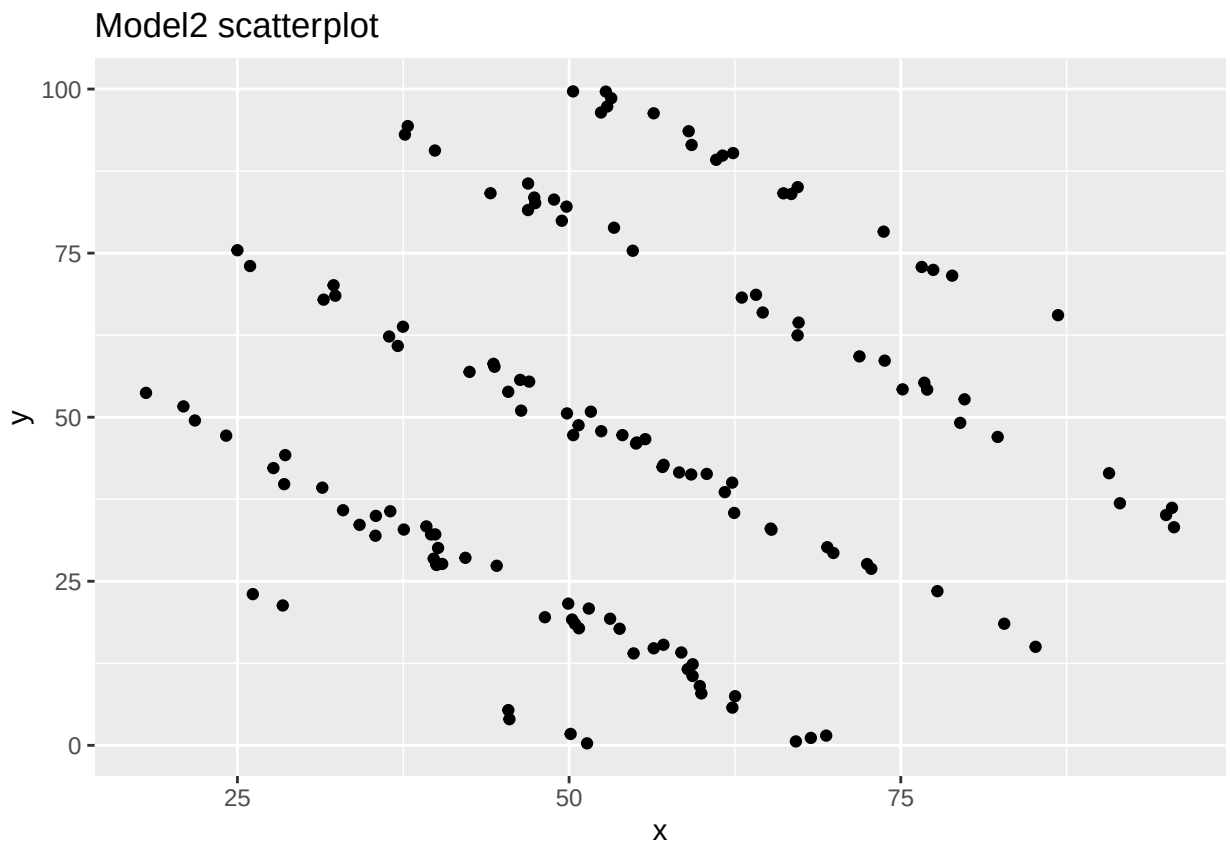
Additionally, because the slope of this graph is so close to 0 (-0.1036) the residuals resemble the scatterplot.

Data set 2 Yes or No

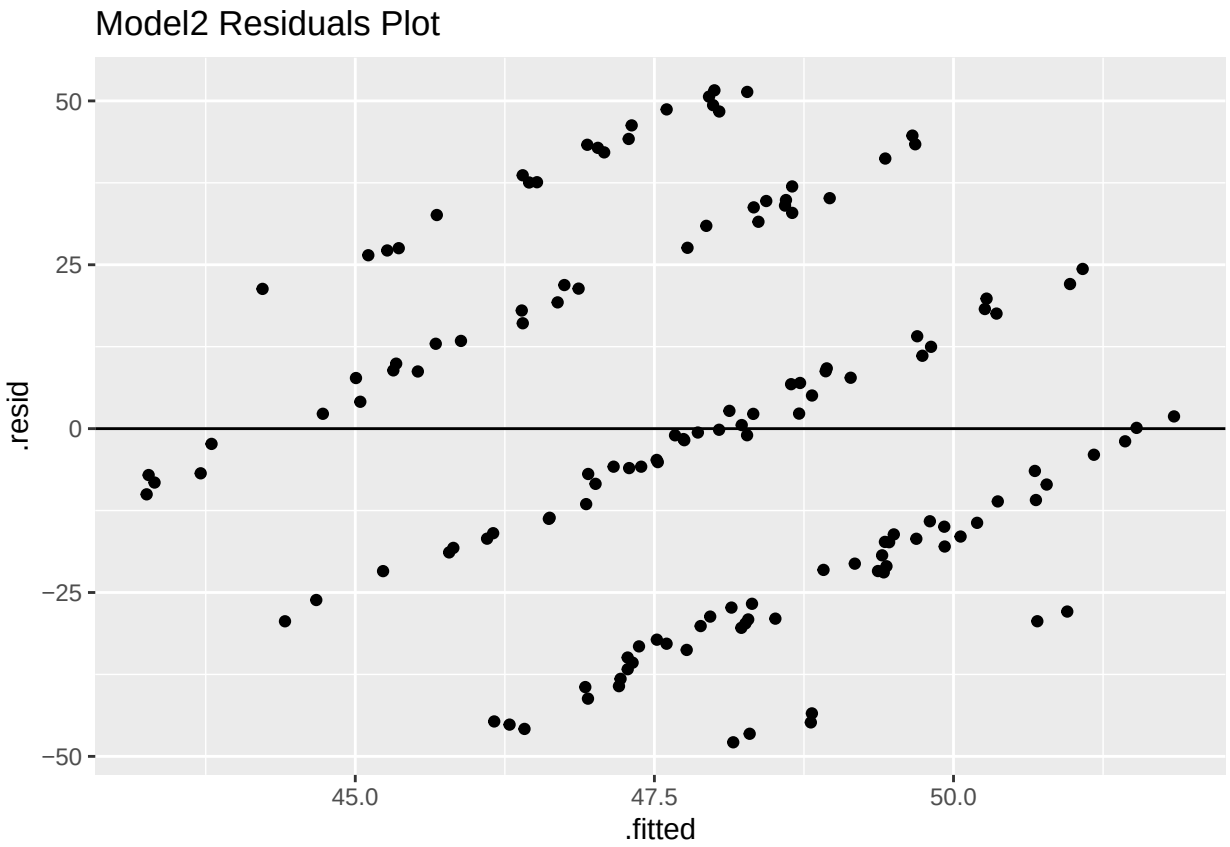
Why?

```
ggplot(
  data2,
  aes(
```

```
    x = x,
    y = y
  )
) +
  geom_point() +
  labs(
    title = "Model2 scatterplot"
  )
```




```
ggplot(
  aes(
    .fitted,
    .resid
  ),
  data = model2
) +
  geom_point() +
  geom_hline(
    yintercept =  0
  ) +
  labs(
    title = "Model2 Residuals Plot"
  )
```




```
print(data2.correlation)
```

```
## [1] "-0.0690"
```

Looking visually at the data, we can see that there seems to be a strong linear correlation between multiple bands of categories of data. Since we only have the x and y, it's likely not a good candidate for linear regression since any given x could have a wide array of potential y values. This can also be seen with the correlation coefficient of -0.0690 which is also very close to 0 suggesting there is there isn't linear relationship between the data.

Moving onto the residuals plot, the slope of each "categories" line seems to have reversed which helps illustrate how the dataset is not very linear as the residuals are still very far from the regression line. This further provides evidence that a linear model isn't appropriate here.

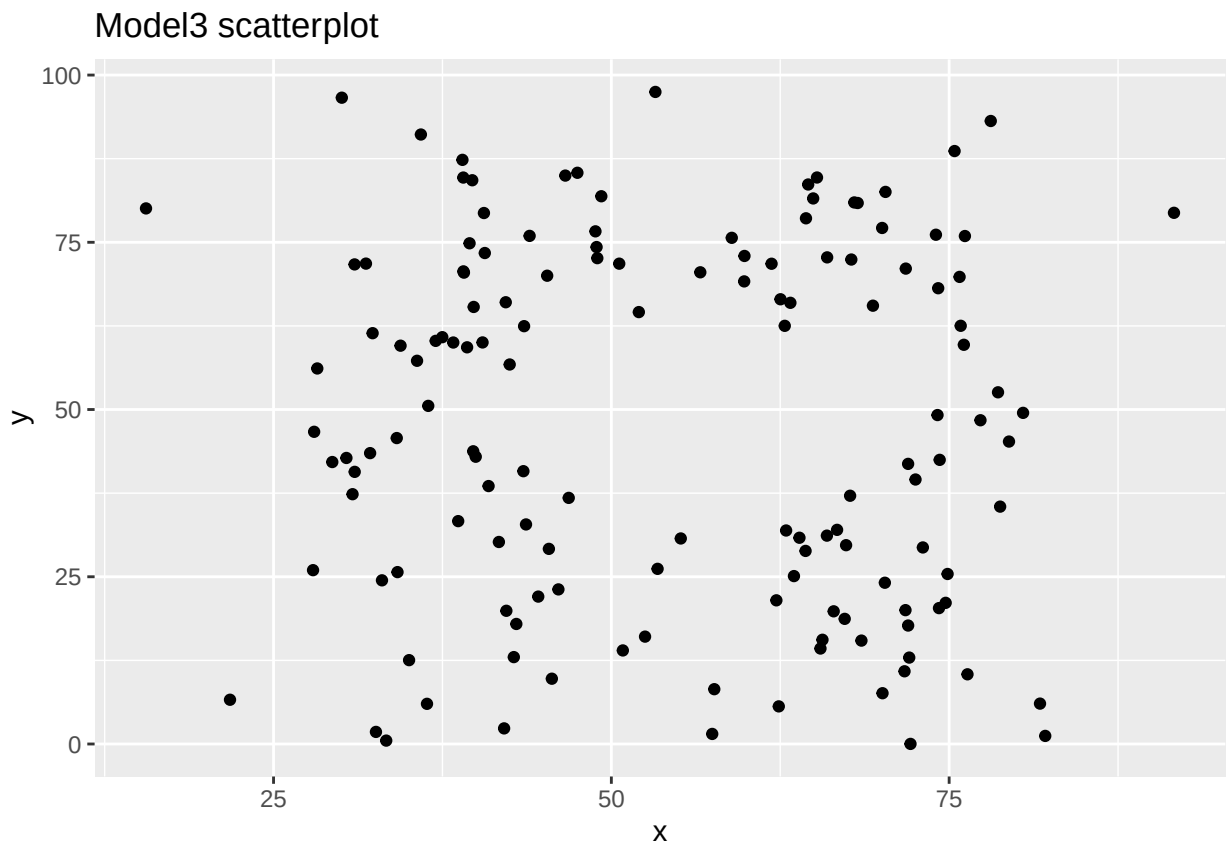
If we were able to have insight into the categories, then x could be a good predictor of y.

Data set 3 Yes or No

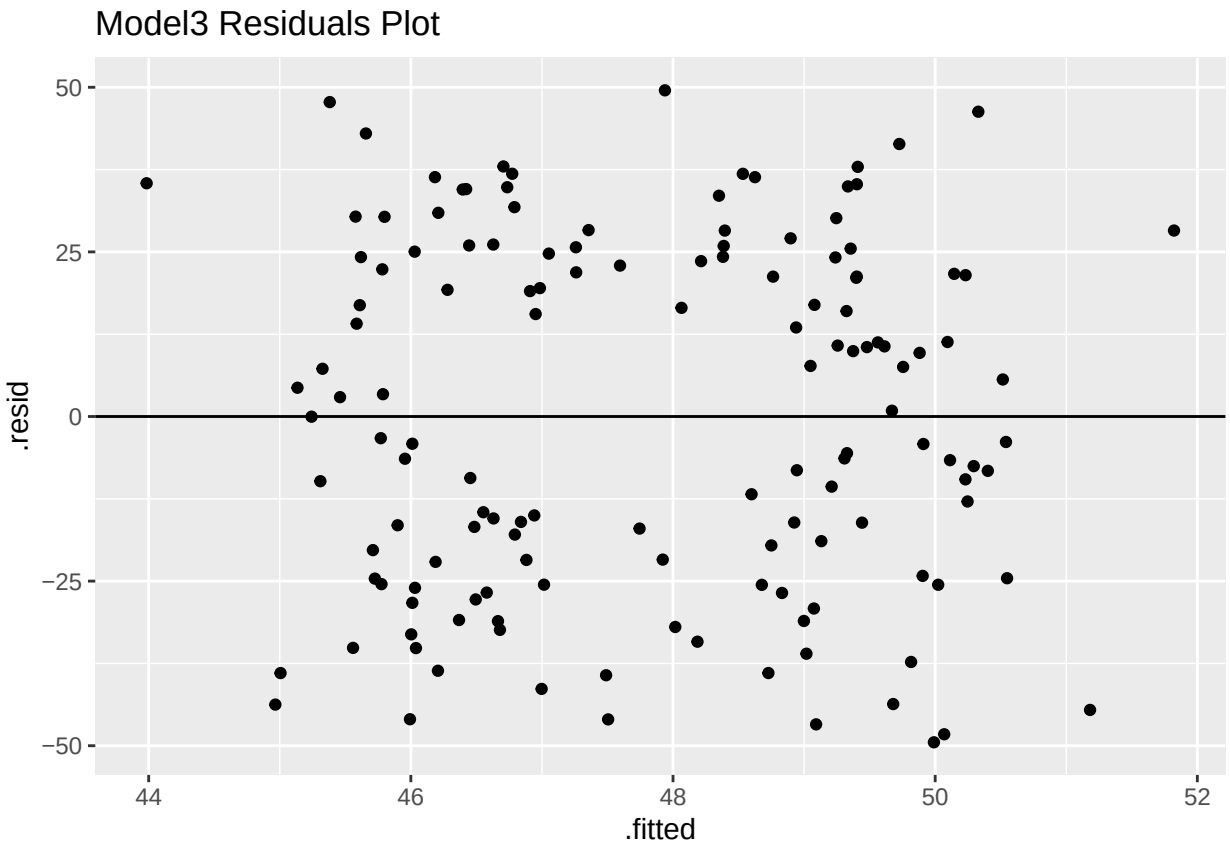
Why?

```
ggplot(
  data3,
  aes(
    x = x,
    y = y
  )
```

```
) +
  geom_point() +
  labs(
    title = "Model3 scatterplot"
  )
```




```
ggplot(
  aes(
    .fitted,
    .resid
  ),
  data = model3
) +
  geom_point() +
  geom_hline(
    yintercept =  0
  ) +
  labs(
    title = "Model3 Residuals Plot"
  )
```




```
print(data3.correlation)
```

```
## [1] "-0.0641"
```

Here we are in a similar situation as model1 and model2. Both the scatterplot and residuals plot provide evidence that a linear model isn't appropriate. Additionally, a close to 0 correlation coefficient of -0.0641 also provides evidence that a linear model here isn't appropriate.

h. Why it is important to include appropriate visualizations when analyzing data? Be sure to ground your reasoning in the context of the analyses completed above. Include any visualization(s) you create. (15 points) Visualizations are important becasue intuitive and easily interpretable method for understanding the data you're working with and they help in communication. Intuition really helped with the first dataset as we were able to see that it was a collection of points that resemble a connect-the-dots drawable. By seeing that, you can easily determine the intent of the data you're working with and determining what analytical techniques can be applied to the data.

Regarding the second dataset, it was easy to discover that the data we're working with was perhaps not the entire story. This was easily seen with the bands of linear data which could have helped an analyst visually determine that there is perhaps more data that can be used to assist in building a model.

Finally, the third dataset seemed like noise. Aside from the fact that there doesn't seem to be any points at the center, there was no linear pattern that we could observe.

Appropriate visualizations are like clues to the detective work of analysis. They help provide clues and highlight patterns that the data has which can inform the following steps in the analysis. For example, with the third dataset, I would be interested to know why there are no other data points near the center and it could perhaps inform me to use KNN or another cluster based model.