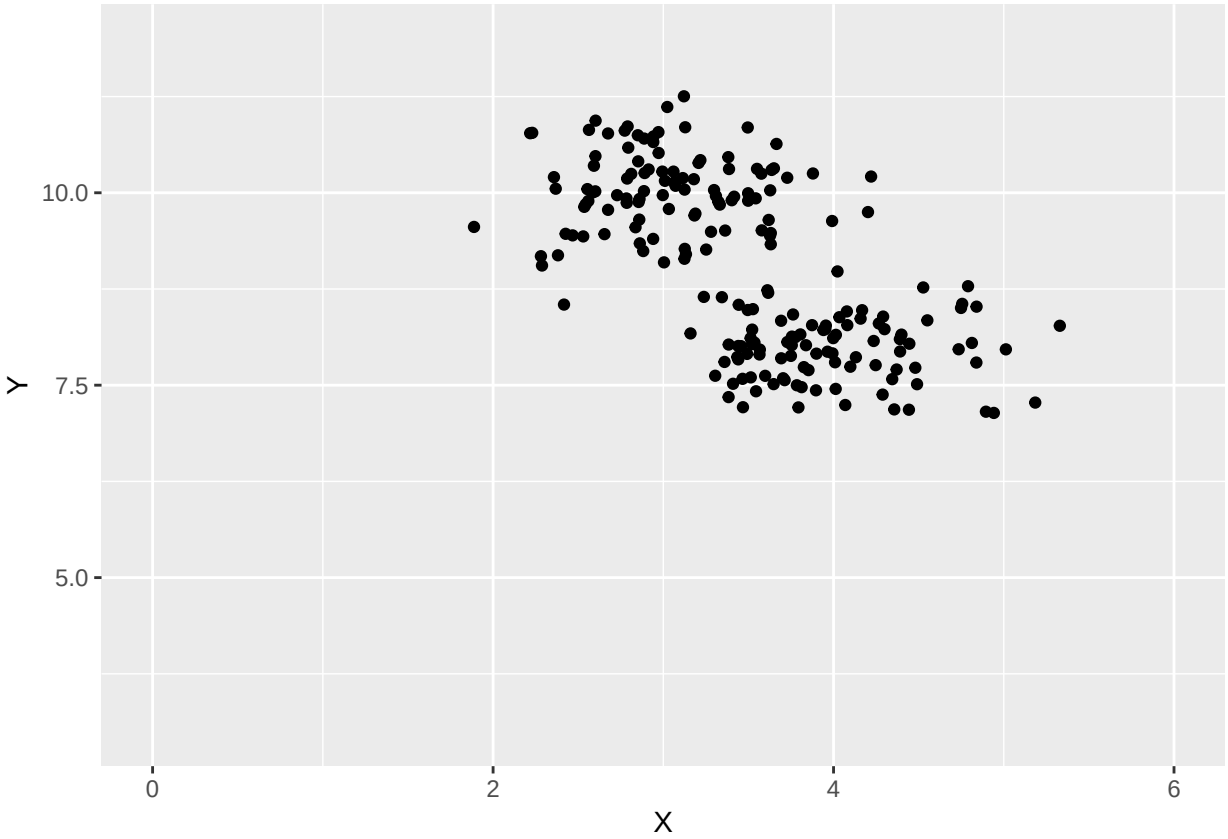
Clustering Example 1

Tim

3/19/2019

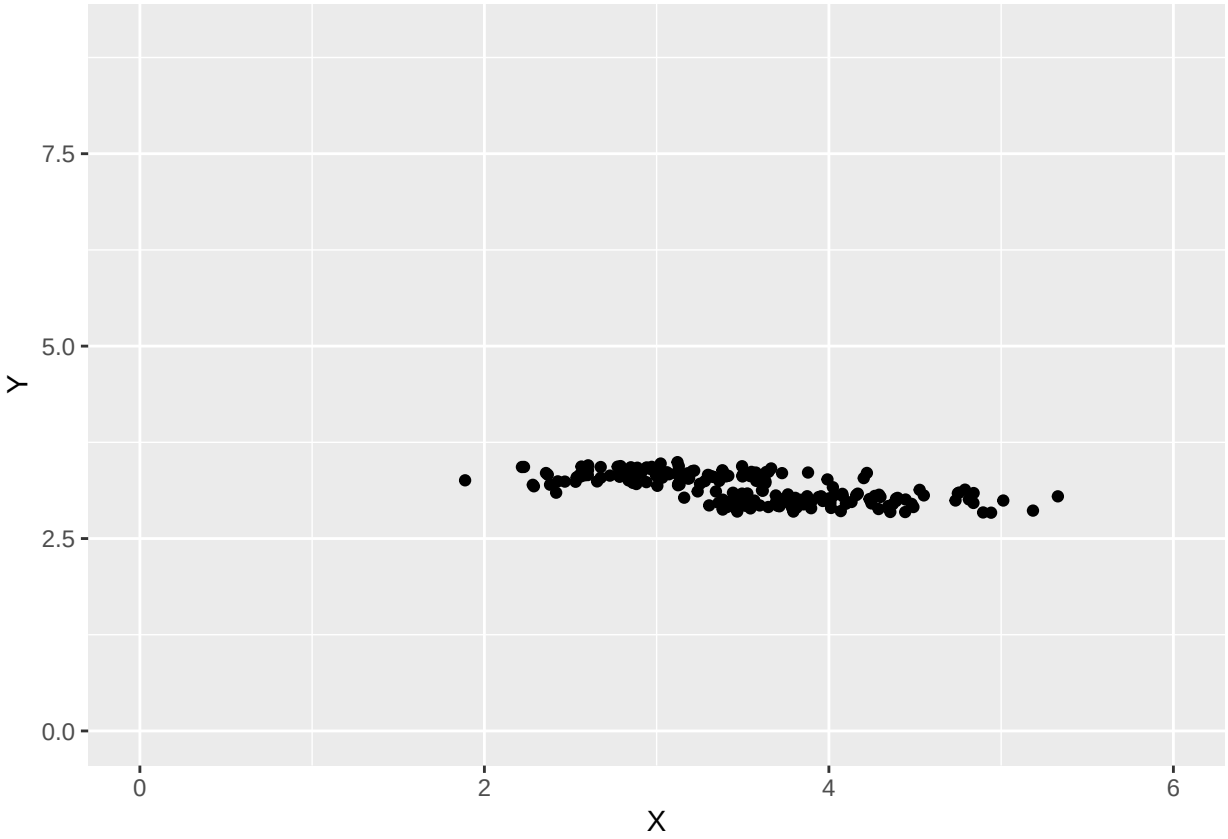
```
set.seed(0)
# Generate the data. We will, by definition, create 2 clusters.
# Create 2 sets of normally distributed X's, and Y's, then combine them into 2D space.
x1 = rnorm(100, mean=3, sd=0.5)
x2 = rnorm(100, mean=4, sd=0.5)

y1 = rnorm(100, mean=10, sd=0.5)
y2 = rnorm(100, mean=8, sd=0.5)

X = c(x1, x2)
Y = c(y1, y2)

df = data.frame(X, Y)
ggplot(df, aes(x=X, y=Y)) + geom_point() +
  xlim(0,6) + ylim(3,12)
```



I think most people would say there are 2 visual clusters here. By definition of how I constructed the problem (2 sets of x's, and y's), there mathematically is.

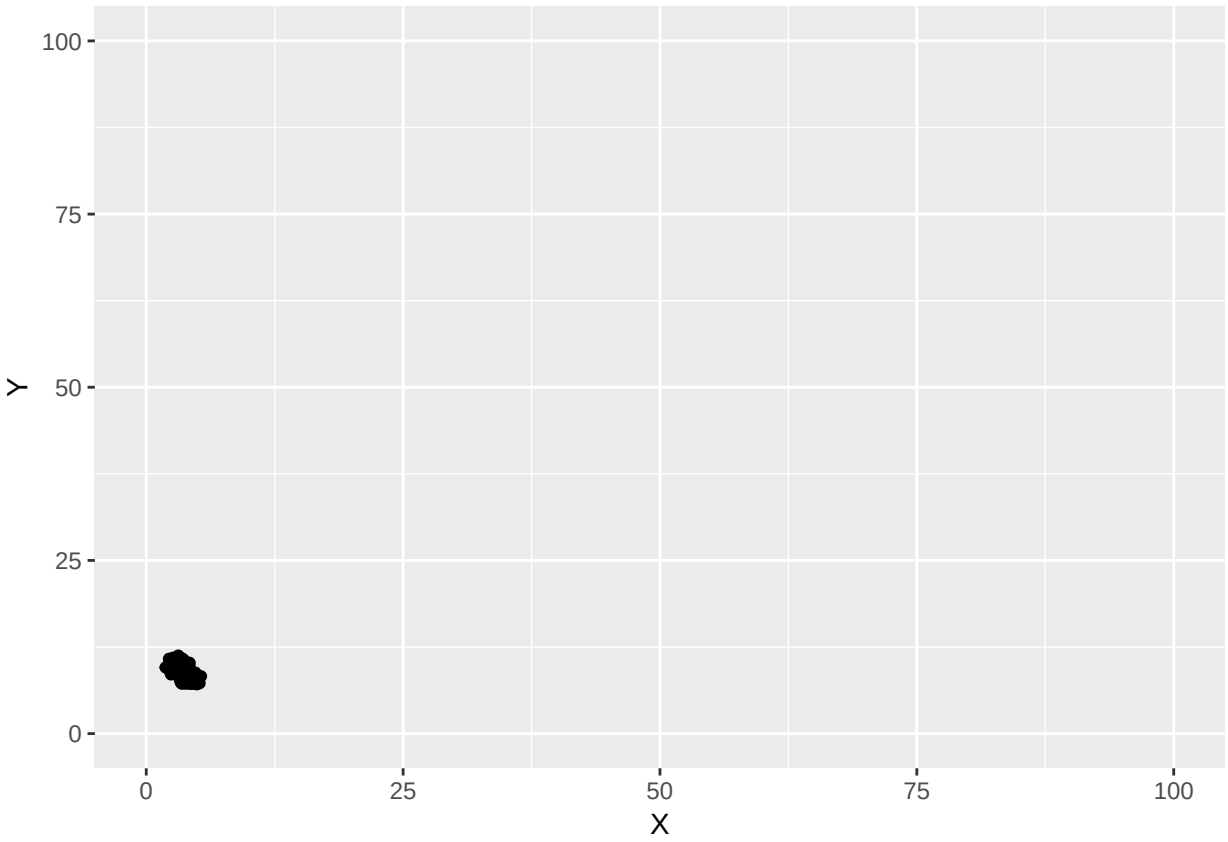
But what about if we log transform. . . ?

```
df2 = df
df2$Y = log2(df2$Y)

ggplot(df2, aes(x=X, y=Y)) + geom_point() +
  xlim(0,6) + ylim(0,9)
```



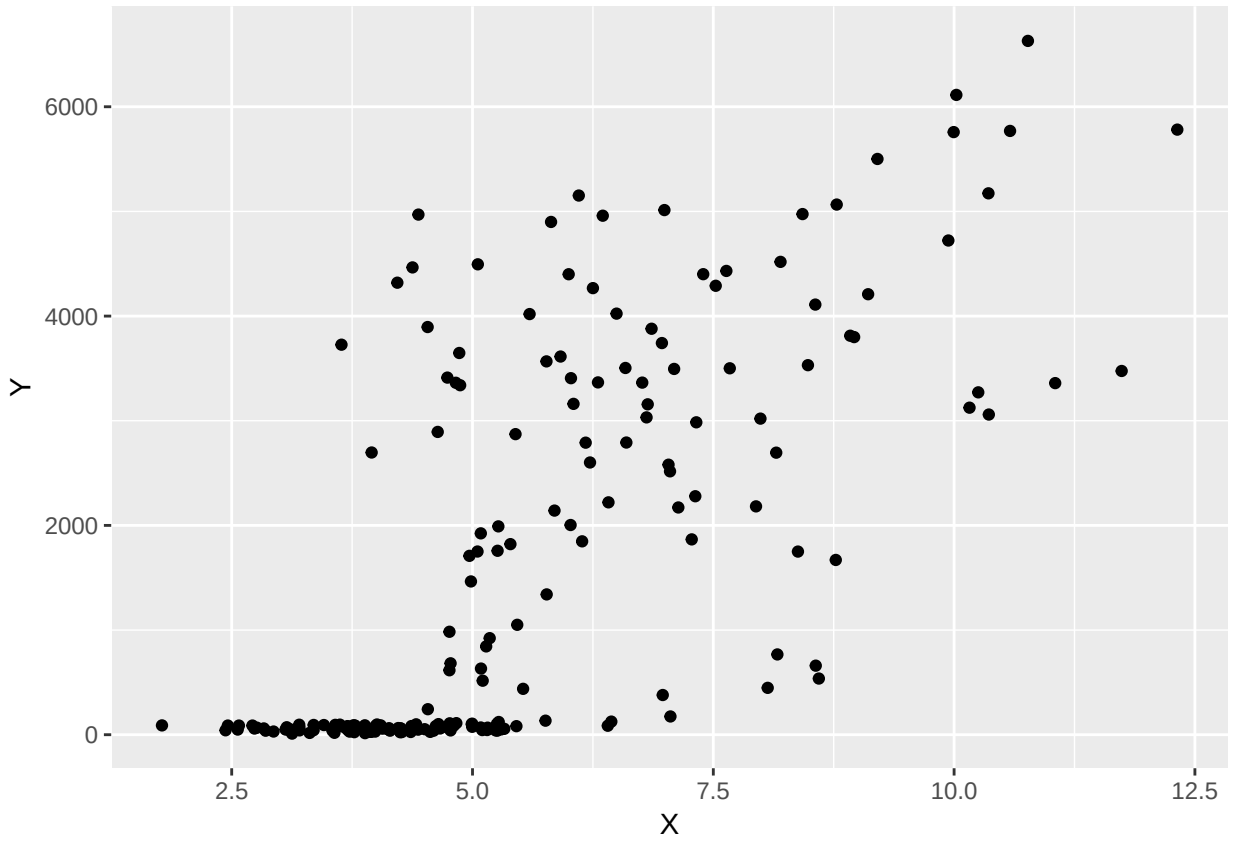
Suddenly things become less clear. Even though the same relationship between X's and Y's exist.

```
# Small note on misleading graphs - visual clustering can be very deceptive.
# What if we just change the view window of the original data?
# Observe the xlim and ylim values have changed

df = data.frame(X, Y)
ggplot(df, aes(x=X, y=Y)) + geom_point() +
  xlim(0,100) + ylim(0,100)
```

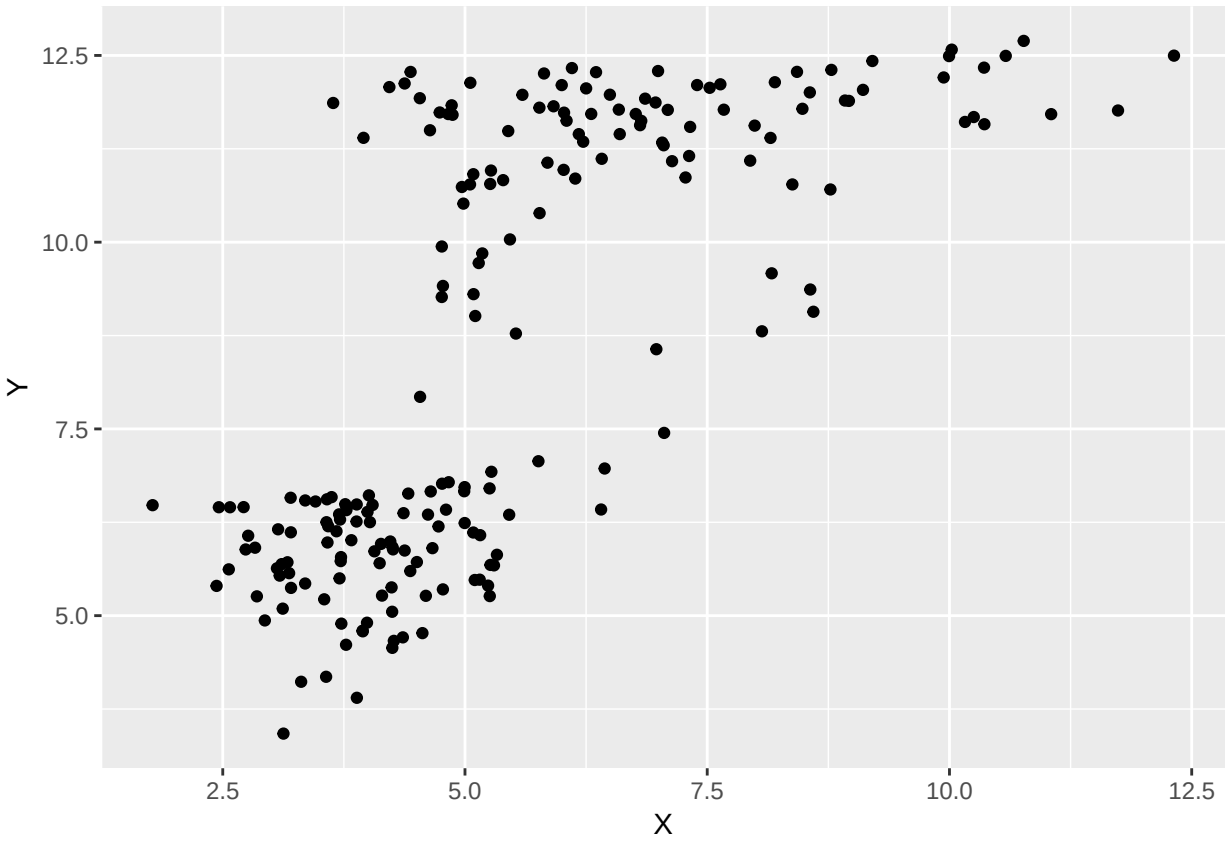


What about if we have data in exponential space? What will happen if we do a log transform then?

```
set.seed(0)
x1 = rnorm(100, mean=4, sd=1)
x2 = rnorm(100, mean=7, sd=2)

y1 = 2**x1+runif(length(x1), 0, 2**max(x1))
y2 = 2**x2+runif(length(x2), 0, 2**max(x2))

X = c(x1, x2)
Y = c(y1, y2)

df = data.frame(X, Y)
ggplot(df, aes(x=X, y=Y)) + geom_point() +
  xlim(min(X),max(X)) + ylim(min(Y),max(Y))
```




```
df2 = df
df2$Y = log2(df2$Y)

ggplot(df2, aes(x=X, y=Y)) + geom_point() +
  xlim(min(df2$X),max(df2$X)) + ylim(min(df2$Y),max(df2$Y))
```



How we prepare our data is important. Non-linear transformations (one example is log transformation) can have dramatic effects on our data.