Lab 1

AUTHOR
Olivia Roberts

```
library(datasauRus)
```

```
Warning: package 'datasauRus' was built under R version 4.5.2
```

```
library(ggplot2)
```

```
Warning: package 'ggplot2' was built under R version 4.5.1
```

Question 1

```
?datasaurus_dozen
```

```
starting httpd help server ... done
```

```
table(datasaurus_dozen$dataset)
```

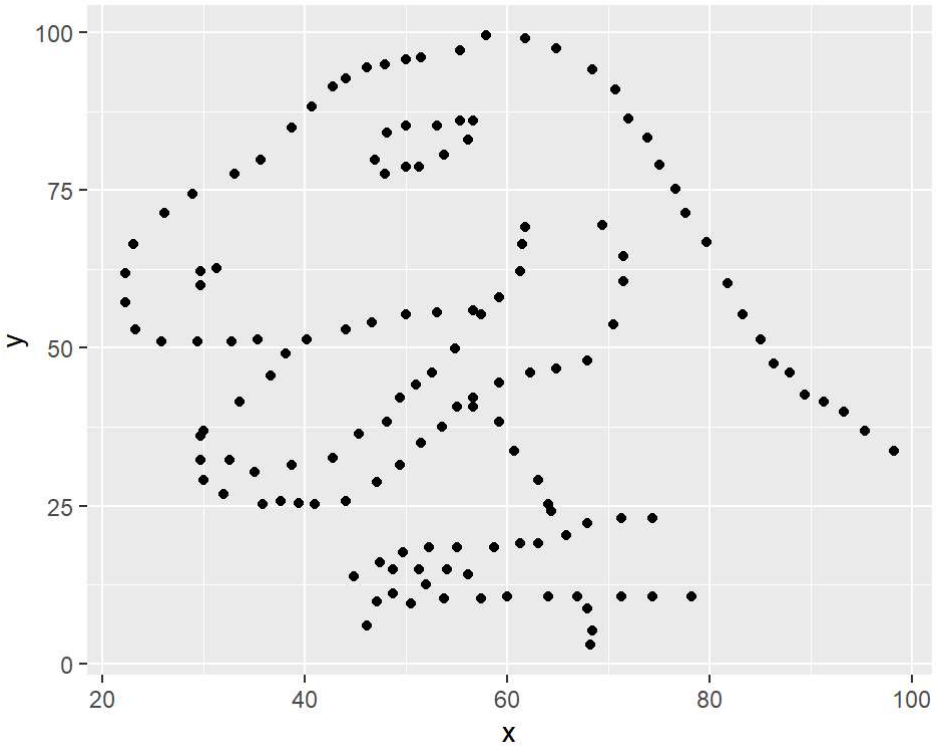
```
      away   bullseye     circle       dino       dots    h_lines high_lines 
       142        142        142        142        142        142        142 
slant_down   slant_up       star    v_lines wide_lines    x_shape 
       142        142        142        142        142        142 
```

Answer: The datasaurus_dozen file contains 1846 rows of 3 variables. The 3 variables are a variable called “dataset” that indicates the dataset the data is from, and x-values variable, and a y-values variable. By accessing the dataset column, we can see a list of the 13 datasets inclued in datasaurus_dozen and that they have 142 rows each.

Question 2

```
dino_data <- datasaurus_dozen[datasaurus_dozen$dataset == 'dino',]

ggplot(data = dino_data, mapping = aes(x = x, y = y)) +
  geom_point()
```



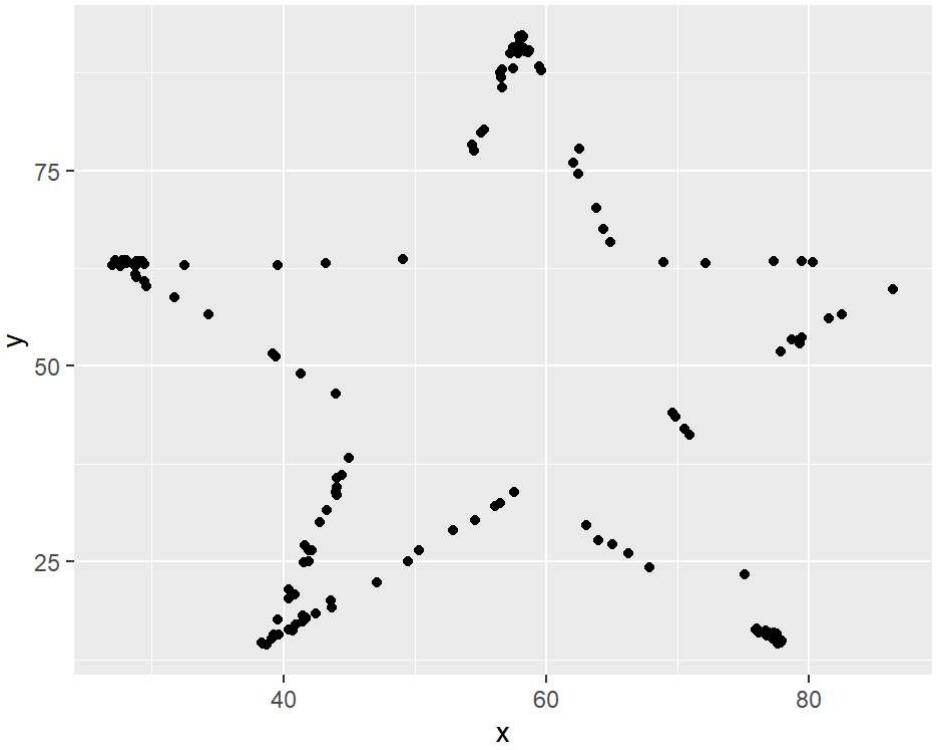

```
cor(dino_data$x, dino_data$y)
```

```
[1] -0.06447185
```

Question 3

```
star_data <- datasaurus_dozen[datasaurus_dozen$dataset == 'star',]

ggplot(data = star_data, mapping = aes(x = x, y = y)) +
  geom_point()
```




```
cor(star_data$x, star_data$y)
```

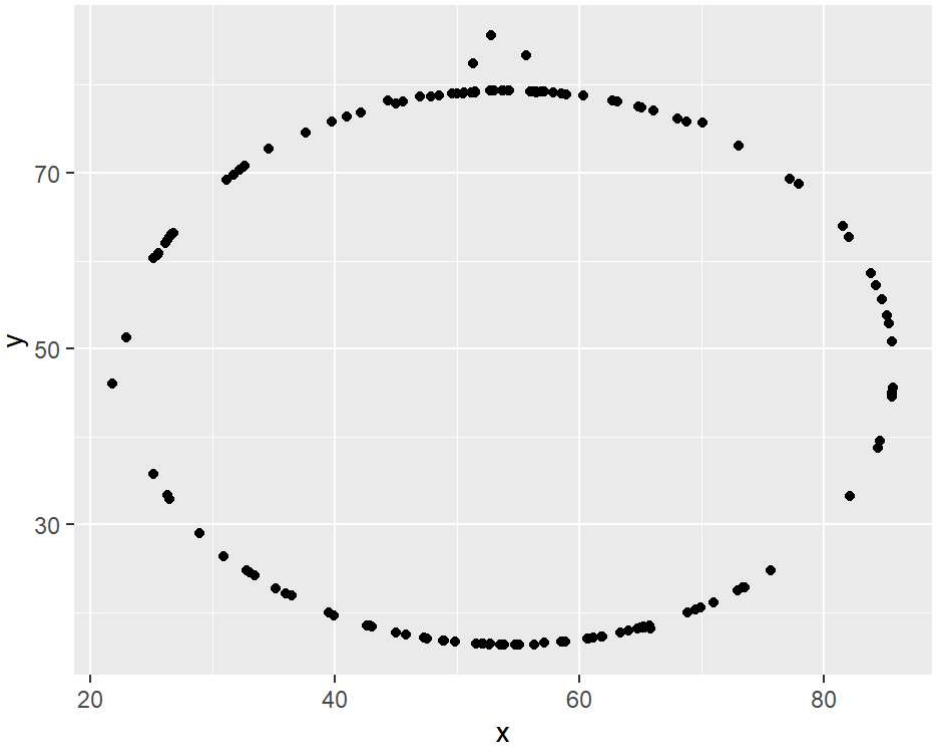
```
[1] -0.0629611
```

Answer: The correlation coefficient for the dino and star datasets are very similar; they are -0.064 and -0.063 respectively. They are both negative, so in both datasets, as x increases, y decreases. The dino dataset has a slightly stronger negative correlation, but neither of the datasets have a very strong negative correlation since the points are very scattered and the magnitude of the correlation is not very large. Despite having similar correlations, the two datasets look very different.

Question 4

```
circle_data <- datasaurus_dozen[datasaurus_dozen$dataset == 'circle',]

ggplot(data = circle_data, mapping = aes(x = x, y = y)) +
  geom_point()
```



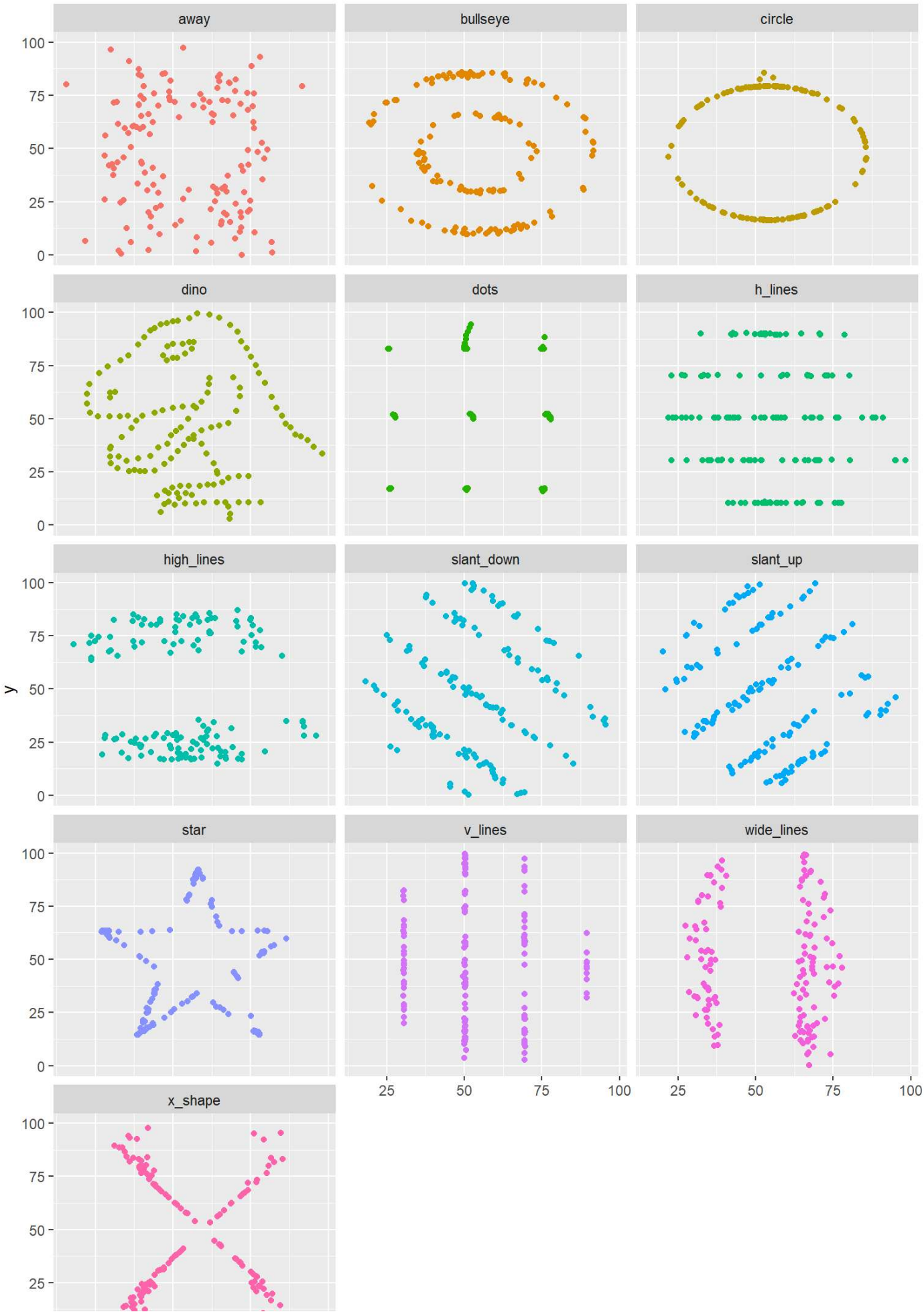

```
cor(circle_data$x, circle_data$y)
```

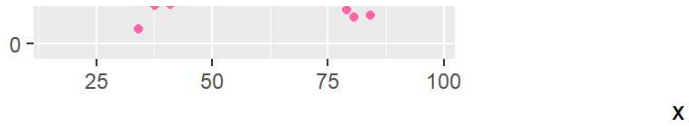
```
[1] -0.06834336
```

Answer: The correlation coefficient of the dino dataset is -0.064, while the correlation coefficient for the circle dataset is -0.068. Again, both datasets have a negative correlation (as x increases, y decreases overall), but the circle dataset now has a slightly stronger magnitude of correlation. However, both datasets still have a very similar magnitude of correlation despite looking very different.

Question 5

```
ggplot(datasaurus_dozen, aes(x = x, y = y, color = dataset))+
  geom_point()+
  facet_wrap(~ dataset, ncol = 3) +
  theme(legend.position = "none")
```


away
bullseye
circle
dino
dots
h_lines
high_lines
slant_down
slant_up
star
v_lines
wide_lines
x_shape
y
0
25
50
75
100



Question 6

```
sapply(unique(datasaurus_dozen$dataset), function(name){
    subset <- datasaurus_dozen[datasaurus_dozen$dataset == name, ]
    return(cor(subset$x, subset$y))
})
```

```
      dino        away     h_lines     v_lines     x_shape        star 
-0.06447185 -0.06412835 -0.06171484 -0.06944557 -0.06558334 -0.06296110 
 high_lines        dots      circle    bullseye    slant_up  slant_down 
-0.06850422 -0.06034144 -0.06834336 -0.06858639 -0.06860921 -0.06897974 
 wide_lines 
-0.06657523 
```