Lab 01 - Hello R!

due January 10, 2021 at 11:59 PM

Dav King

1/5/22

Load Packages

```
library(tidyverse)
library(datasauRus)
```

Exercise 1

Type your answer to exercise #1 here. Note this exercise doesn't require any R code.

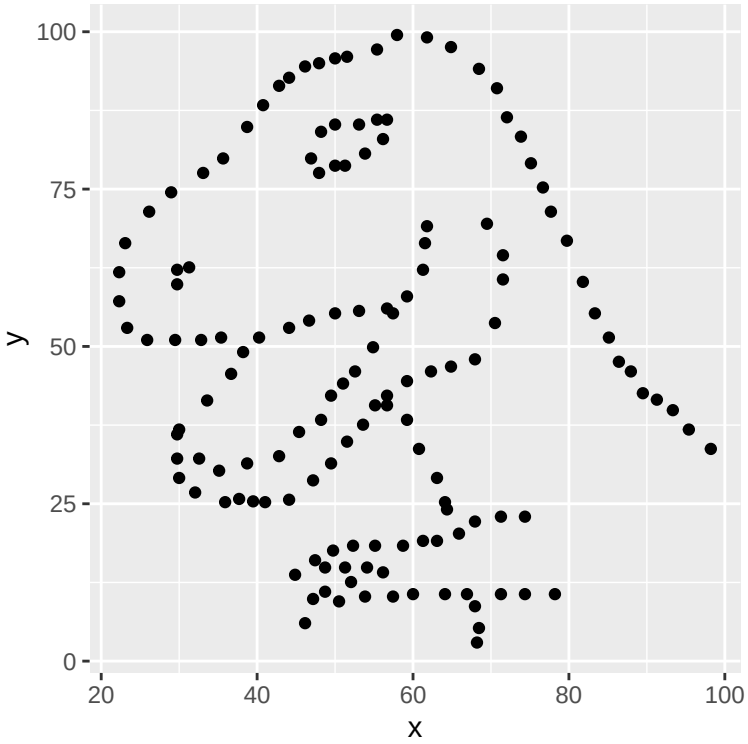
Note also in this lab we are telling you exactly when to stage, commit, and push.

the data frame has 1846 rows and 3 variables (columns). Variables: `dataset` (indicates which dataset the data are from), `x` (x-values), and `y` (y-values)

Exercise 2

This filters the data set to only values where dataset has the value "dino", and plots the x-y dot plot of that filter data set.

```
dino_data <- datasaurus_dozen %>%
  filter(dataset == "dino")
ggplot(data = dino_data, mapping = aes(x = x, y = y)) +
  geom_point()
```



Thus calculates the correlation coefficient, r = -0.0645

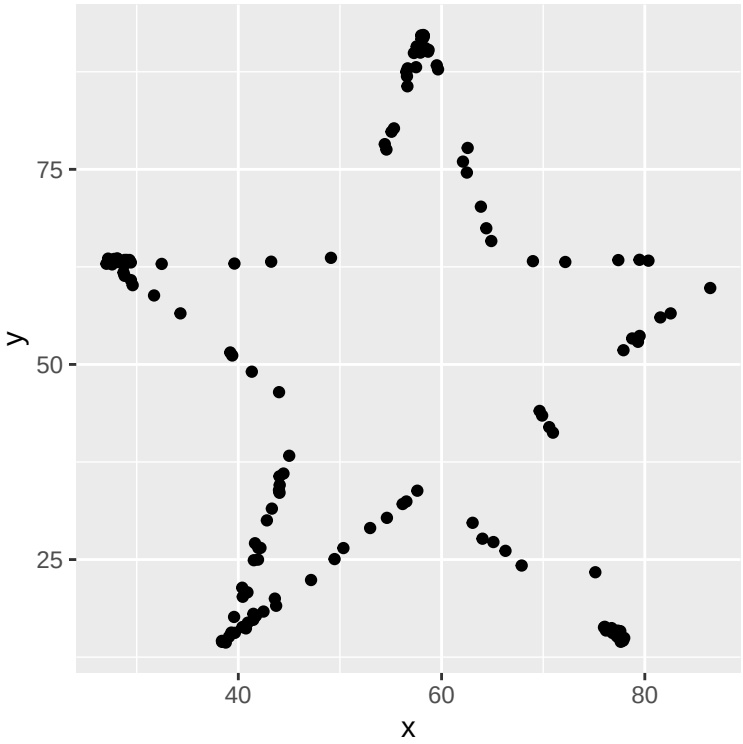
```
dino_data %>%
  summarize(r = cor(x, y))
```

```
## # A tibble: 1 x 1
##         r
##     <dbl>
## 1 -0.0645
```

Exercise 3

This filters the data frame to only values where dataset == “star”, and plots y against x for that dataset.

```
star_data <- datasaurus_dozen %>%
  filter(dataset == "star")
ggplot(data = star_data, mapping = aes(x = x, y = y)) +
  geom_point()
```



Note: the instructions say star on sakai but circle in the rmd file; to use circle with as few changes as possible the second line of this code would be changed to " filter(dataset == "circle")", which I am not denoting in R code because it renders a massive, unnecessary data table.

This calculates the correlation coefficient of the star data set.

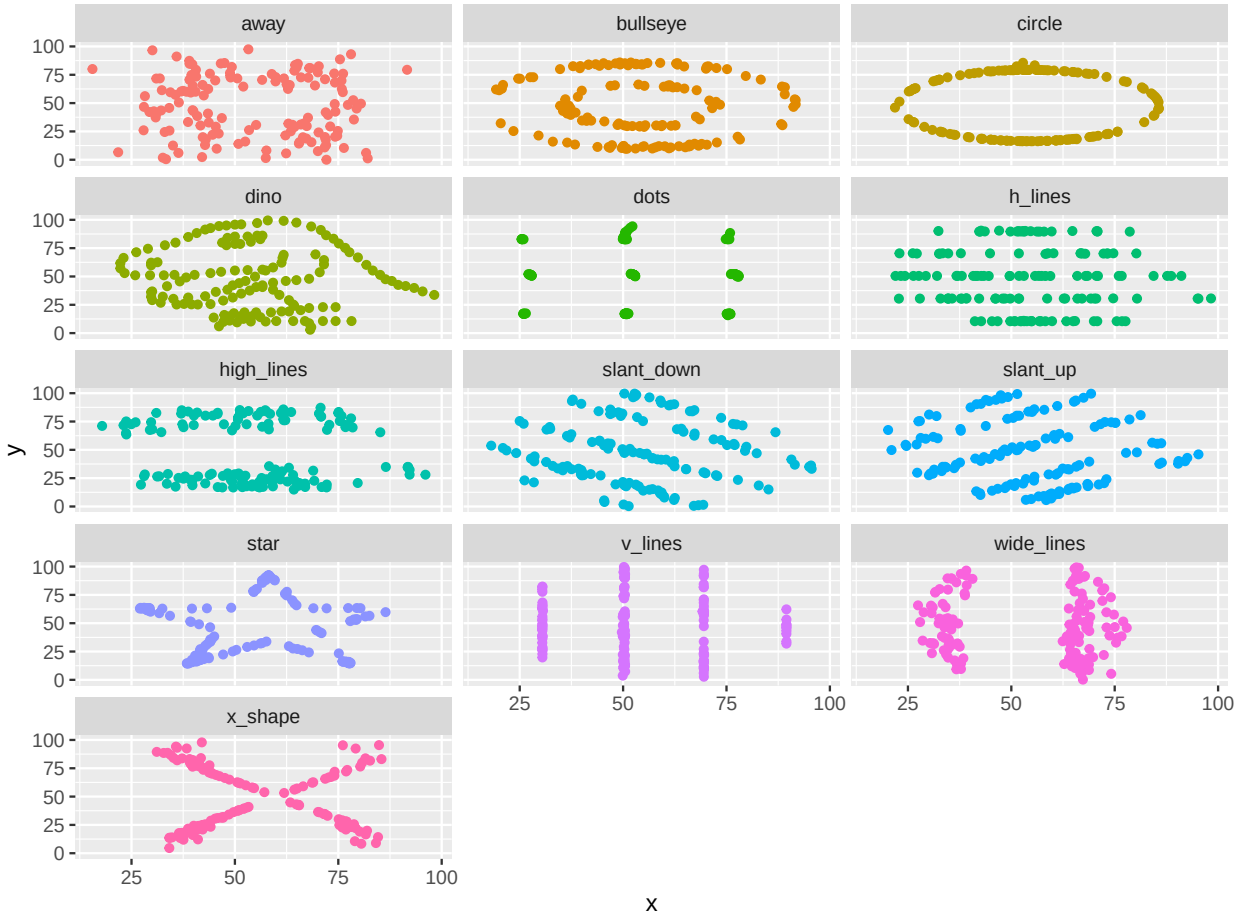
```
star_data %>%
  summarize(r = cor(x,y))
```

```
## # A tibble: 1 x 1
##         r
##     <dbl>
## 1 -0.0630
```

The correlation is -0.0630, which is a slightly weaker negative correlation than for the dino set but once again largely meaningless as a linear relationship.

Exercise 4

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



All plots have a distinct shape, based upon their according names.

```
datasaurus_dozen %>%
  group_by(dataset) %>%
  summarize(r = cor(x, y))
```

```
## # A tibble: 13 x 2
##    dataset          r
##    <chr>        <dbl>
##  1 away       -0.0641
##  2 bullseye   -0.0686
##  3 circle     -0.0683
##  4 dino       -0.0645
##  5 dots       -0.0603
##  6 h_lines    -0.0617
##  7 high_lines -0.0685
##  8 slant_down -0.0690
##  9 slant_up   -0.0686
## 10 star       -0.0630
## 11 v_lines    -0.0694
## 12 wide_lines -0.0666
## 13 x_shape    -0.0656
```

Despite a wide variety in the shapes of these data sets, they all have a correlation of -0.07 $<$ r $<$ -0.06.

This is especially noteworthy, as these data are so unlike one another (in fact, the slant_down and slant_up datasets are almost complete opposites conceptually), and r-values this similar should not occur randomly - suggesting that the data were designed for this principle, and perhaps explaining why only the dino file is particularly well drawn and free of errors in its image compared to the rest of the datasets.