Lab 01 - Hello R!

due January 10, 2021 at 11:59 PM

Ayush Jain

Load Packages

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

Exercise 1

datasauRus_dozen is a data frame with 1846 rows and 3 variables (columns) which are:

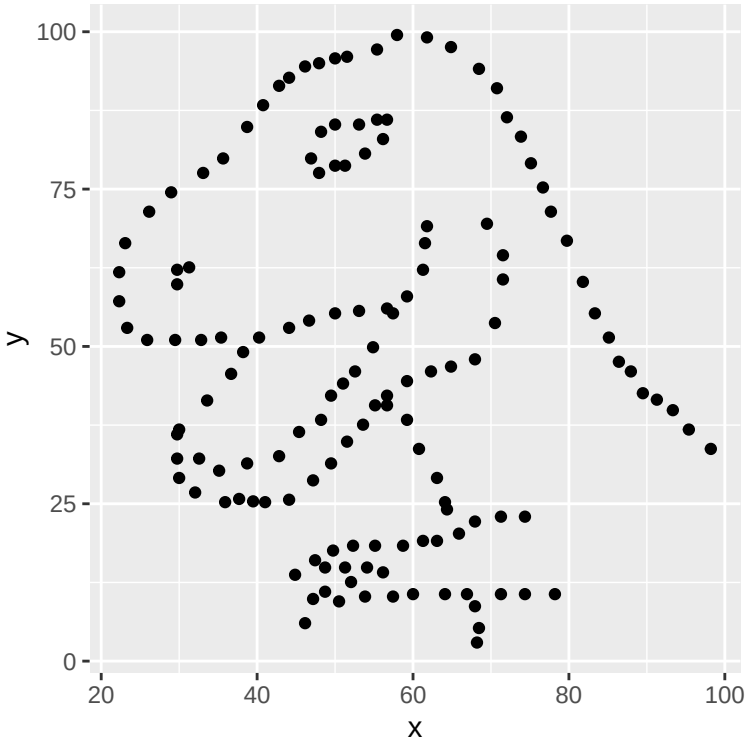
dataset: indicates which dataset the data are from

x: x-values

y: y-values

Exercise 2

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

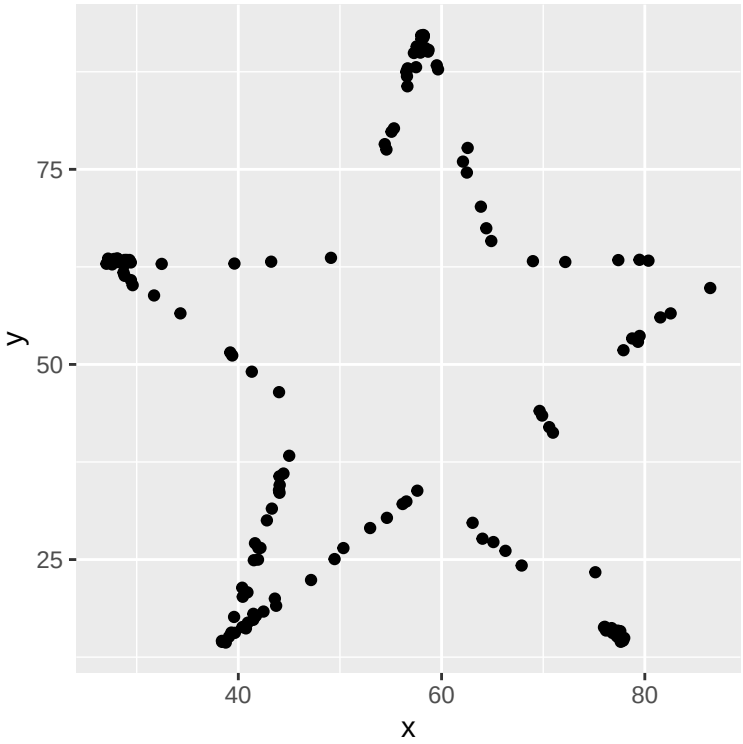



```
dino_data %>%
  summarize(r = cor(x, y))
```

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

Exercise 3

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



The code above is the same as Ex2 but the new data frame is star_data. The shape is that of a star

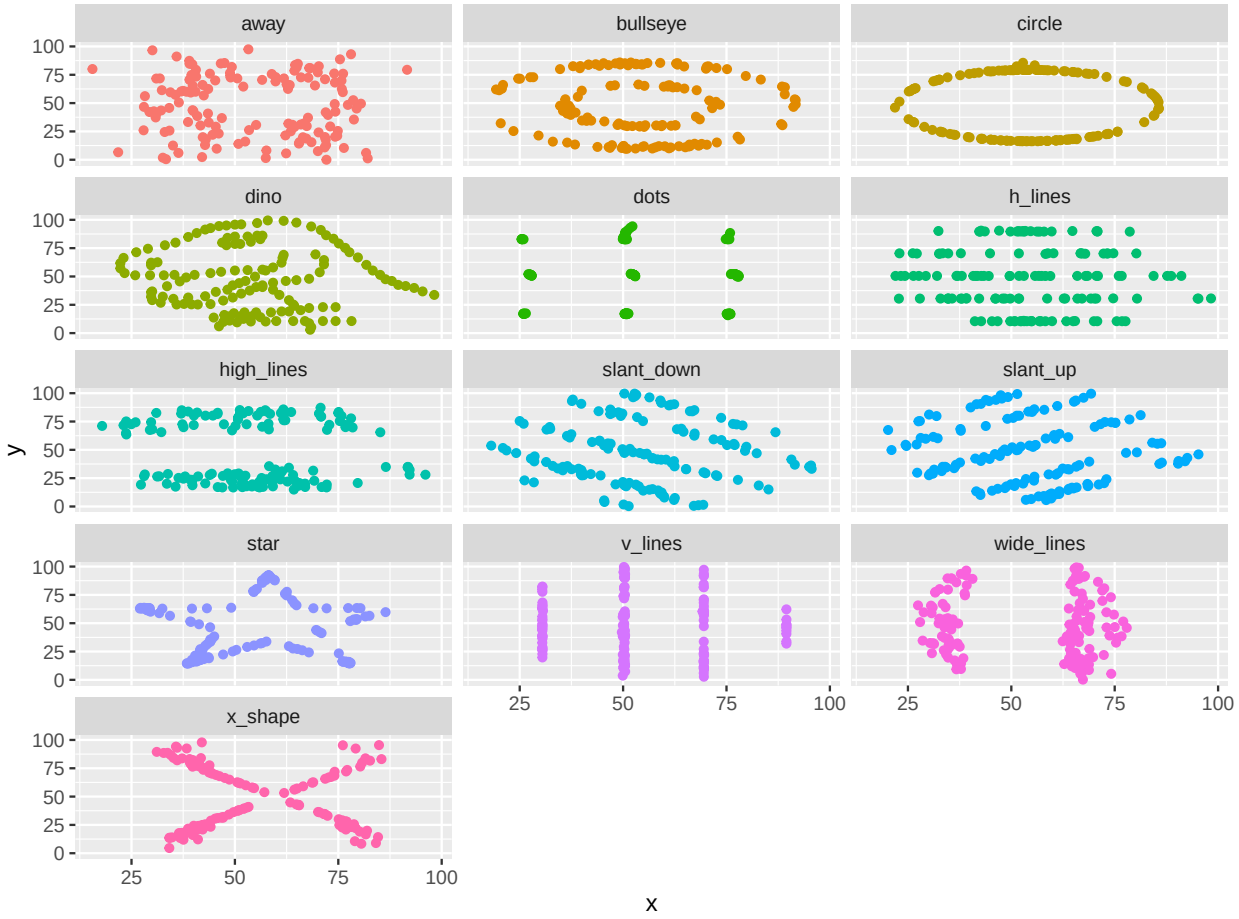
```
star_data %>%
  summarize(r = cor(x, y))
```

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

The r value is nearly the same as that of dino (-0.0630). I find that interesting because the shapes are very different

Exercise 4

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



The above chunk plots everything in the dataframe using facet wrap. It is evident that every dataset plots vastly different shapes, many of which seem to have no clear correlation

```
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
```

The above chunk gives the key summary correlation coefficients, grouped by dataset. It is interesting that all of them have nearly the same r value despite being so different. This illustrates that numerical statistics can often be misleading and do not reveal everything about a dataset.