STAT 33A Homework 3

CJ HINES (3034590053)

Oct 8, 2020

This homework is due **Oct 8, 2020** by 11:59pm PT.

Homeworks are graded for correctness.

As you work, write your answers in this notebook. Answer questions with complete sentences, and put code in code chunks. You can make as many new code chunks as you like.

Please do not delete the exercises already in this notebook, because it may interfere with our grading tools.

You need to submit your work in two places:

- Submit this Rmd file with your edits on bCourses.
- Knit and submit the generated PDF file on Gradescope.

If you have any last-minute trouble knitting, **DON'T PANIC**. Submit your Rmd file on time and follow up in office hours or on Piazza to sort out the PDF.

Exercise 1

For this assignment, you'll use the Datasaurus Dozen data set, which is available on the bCourse (`DatasaurusDozen.tsv`).

Load the Datasaurus Dozen data set and assign it to a variable named `dsaur`.

YOUR ANSWER GOES HERE:

```r
dsaur = read.delim("DatasaurusDozen.tsv")
```

Exercise 2

Now that you've loaded the data set, print out summary information, including:

- Number of columns
- Number of rows
- Classes of the columns
- Levels in the `dataset` column
- The range of the `x` column
- The range of the `y` column
- Number of missing values in each column

YOUR ANSWER GOES HERE:

```
summary(dsaur)
```

```
##    dataset                  x               y
##  Length:1846        Min.   :15.56   Min.   : 0.01512
##  Class :character   1st Qu.:41.07   1st Qu.:22.56107
##  Mode  :character   Median :52.59   Median :47.59445
##                     Mean   :54.27   Mean   :47.83510
##                     3rd Qu.:67.28   3rd Qu.:71.81078
##                     Max.   :98.29   Max.   :99.69468
```

Number of columns: 3

```
ncol(dsaur)
```

```
## [1] 3
```

Number of rows: 1846

```
nrow(dsaur)
```

```
## [1] 1846
```

Classes of the columns: Character

```
class(colnames(dsaur))
```

```
## [1] "character"
```

Levels in the `dataset` column: “away”, “bullseye”, “circle”, “dino”, “dots”, “h_lines”, “high_lines”, “slant_down”, “slant_up”, “star”, “v_lines”, “wide_lines”, “x_shape”

```
dsaur$dataset = as.factor(dsaur$dataset)
levels(dsaur$dataset)
```

```
##  [1] "away"       "bullseye"   "circle"     "dino"       "dots"
##  [6] "h_lines"    "high_lines" "slant_down" "slant_up"   "star"
## [11] "v_lines"    "wide_lines" "x_shape"
```

The range of the `x` column: 82.72737

```
xr = range(dsaur$x, na.rm = TRUE)
xr[2] - xr[1]
```

```
## [1] 82.72737
```

The range of the `y` column: 99.67956

```r
yr = range(dsaur$y, na.rm = TRUE)
yr[2] - yr[1]
```

```
## [1] 99.67956
```

Number of missing values in each column: 0 for all columns

```r
sum(is.na(dsaur$dataset))
```

```
## [1] 0
```

```r
sum(is.na(dsaur$x))
```

```
## [1] 0
```

```r
sum(is.na(dsaur$y))
```

```
## [1] 0
```

Exercise 3

The Datasaurus Dozen is actually a collection of 12 data sets stacked together. The `dataset` column indicates which data set each row comes from.

1. Use subsetting to extract only the rows in the `dino` data set. Assign those rows to the `dino` variable.
2. Compute the mean and standard deviation for the `x` and `y` columns in the `dino` data set.
3. Repeat part 3.1 and 3.2 for the `star` dataset.

   Based on the statistics, are the two data sets similar?

YOUR ANSWER GOES HERE:

1.

```r
dino = dsaur[which(dsaur$dataset == "dino"),]
```

2.

```r
mean(dino$y)
```

```
## [1] 47.83225
```

```r
sd(dino$y)
```

```
## [1] 26.9354
```

```r
mean(dino$x)
```

```
## [1] 54.26327
```

```r
sd(dino$x)
```

```
## [1] 16.76514
```

3.

```r
star = dsaur[which(dsaur$dataset == "star"),]
mean(star$y)
```

```
## [1] 47.83955
```

```r
sd(star$y)
```

```
## [1] 26.93027
```

```r
mean(star$x)
```

```
## [1] 54.26734
```

```r
sd(star$x)
```

```
## [1] 16.76896
```

Yes they have similarities. Dino and star's mean x's and y's are similar. Their medians are slightly off, but still close.

Exercise 4

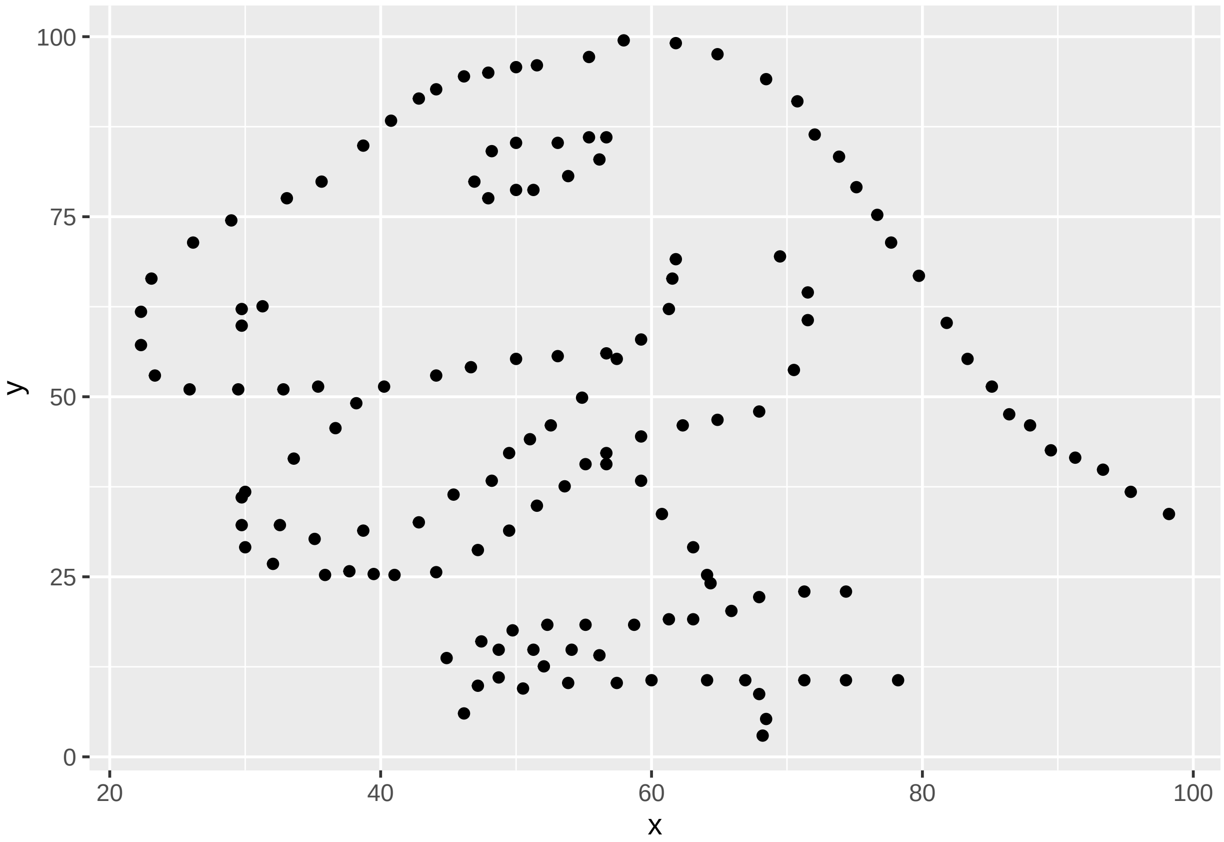
1. Use `ggplot2` to make a scatter plot of `x` versus `y` for the `dino` data set. Make sure your plot includes a title.
2. Repeat for the `star` data set.

   Based on these plots, are the two data sets similar?

YOUR ANSWER GOES HERE:

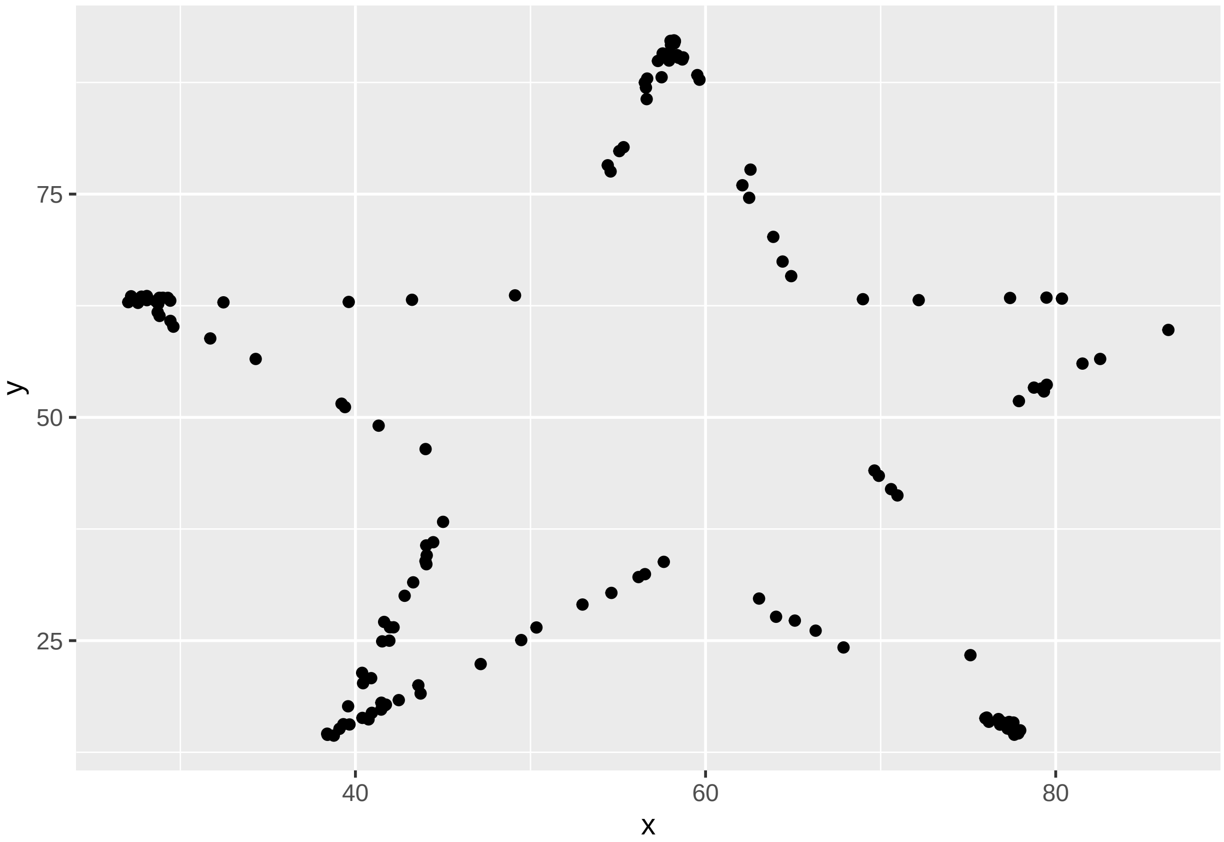
1.

```r
library(ggplot2)
ggplot(dino, aes(x = x, y = y)) + geom_point()
```



2.

```
ggplot(star, aes(x = x, y = y)) + geom_point()
```



Statistically speaking, the two data sets don't seem very similar based on these plots. They are both cool drawings though!

Exercise 5

What do the results for Exercise 3 and 4 suggest about the value of point statistics like the mean and standard deviation, especially in comparison to plots? Explain in 2-5 sentences.

YOUR ANSWER GOES HERE:

> Although the means and standard deviations were extremely similar, their actual plots are very different. This suggests that only relying on point statistics like mean and standard deviation limits our understanding of the data, what questions we're able to answer, and what values we're able to predict.