Practical 0

Cassidy Singh

2025-05-15

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

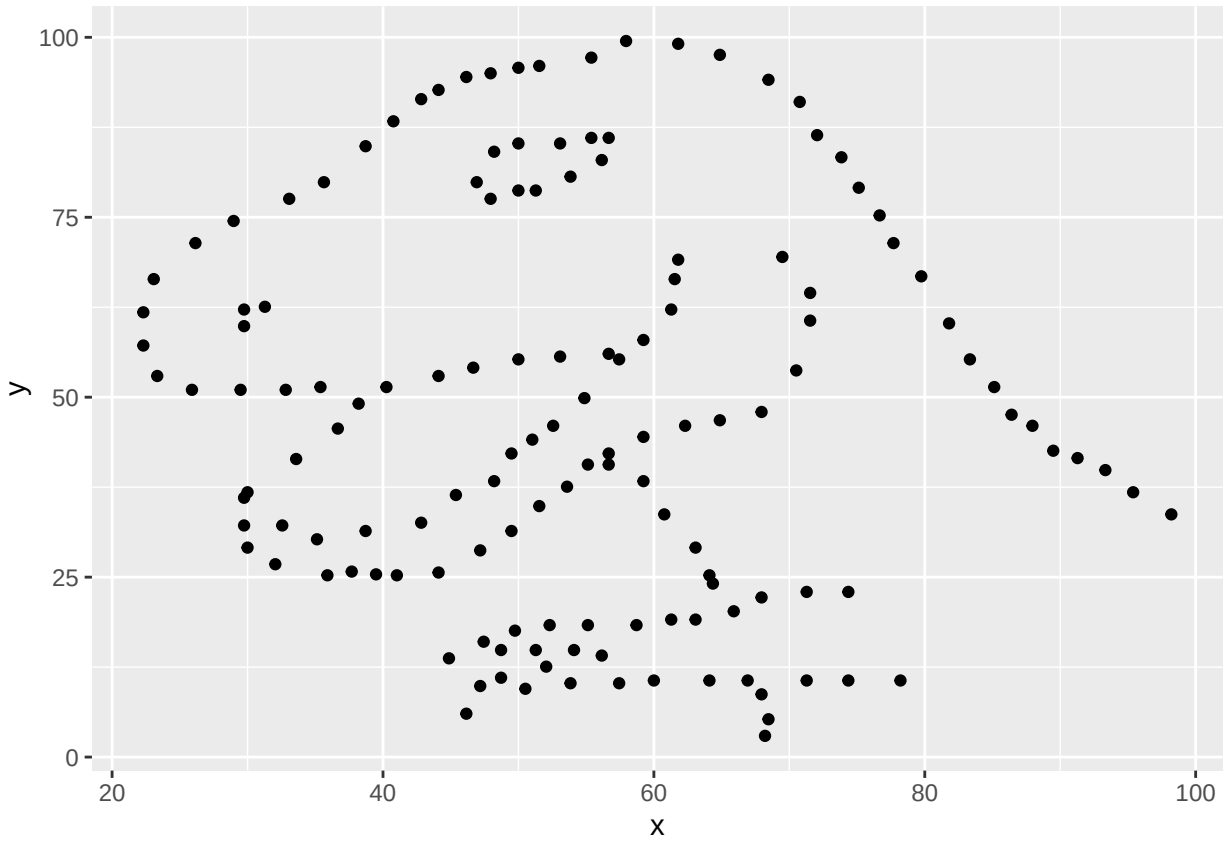
```
library(datasauRus)
?datasaurus_dozen
```

1. Based on the help file, how many rows and how many columns does the datasaurus_dozen file have? What are the variables included in the data frame? (this can be hardcoded)

there are 1846 rows and 3 columns. the variables included in the data frame are dataset, x, and y.

```
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))
```

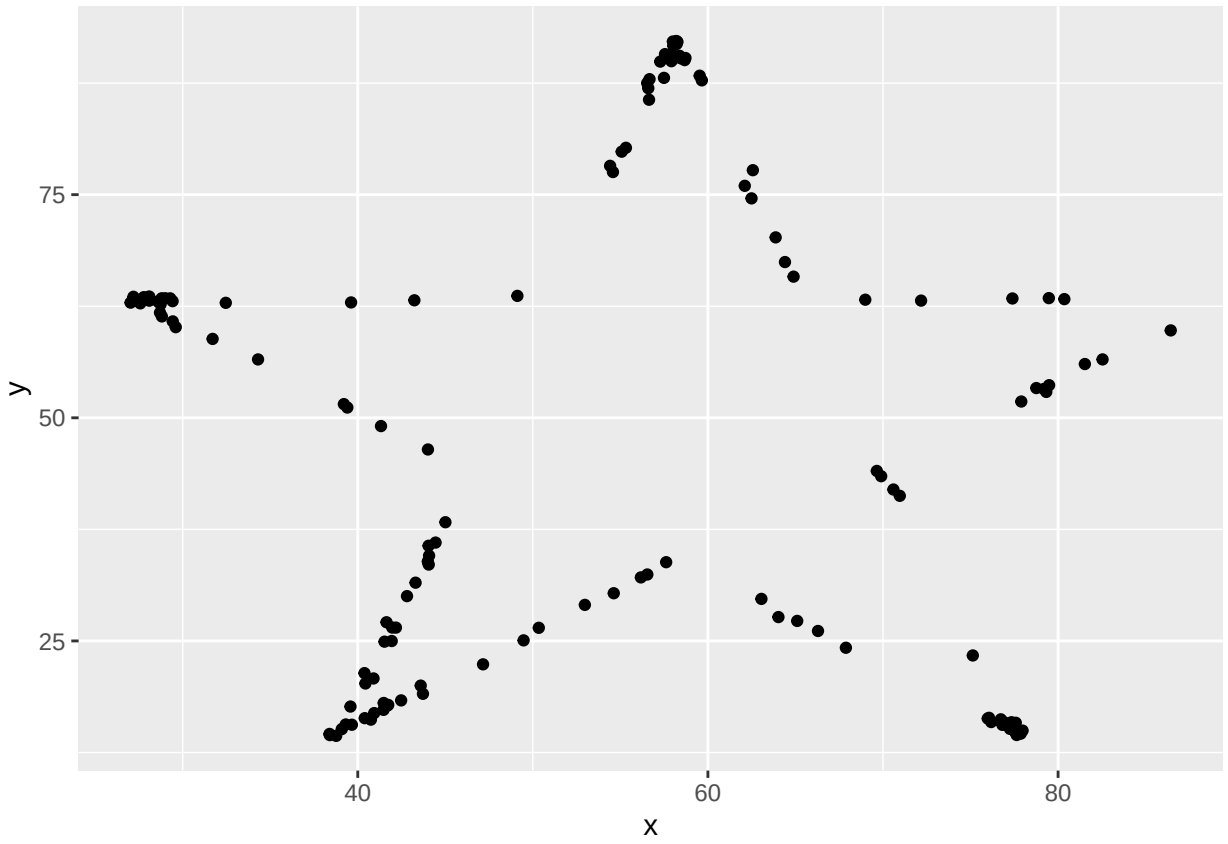
```
## # A tibble: 1 x 1
##         r
##     <dbl>
## 1 -0.0645
```

2. Plot y vs. x for the dino dataset. Then, calculate the correlation coefficient between x and y for this dataset.

##-0.0645

```
dino_data <- datasaurus_dozen %>%
  filter(dataset == "star")

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




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

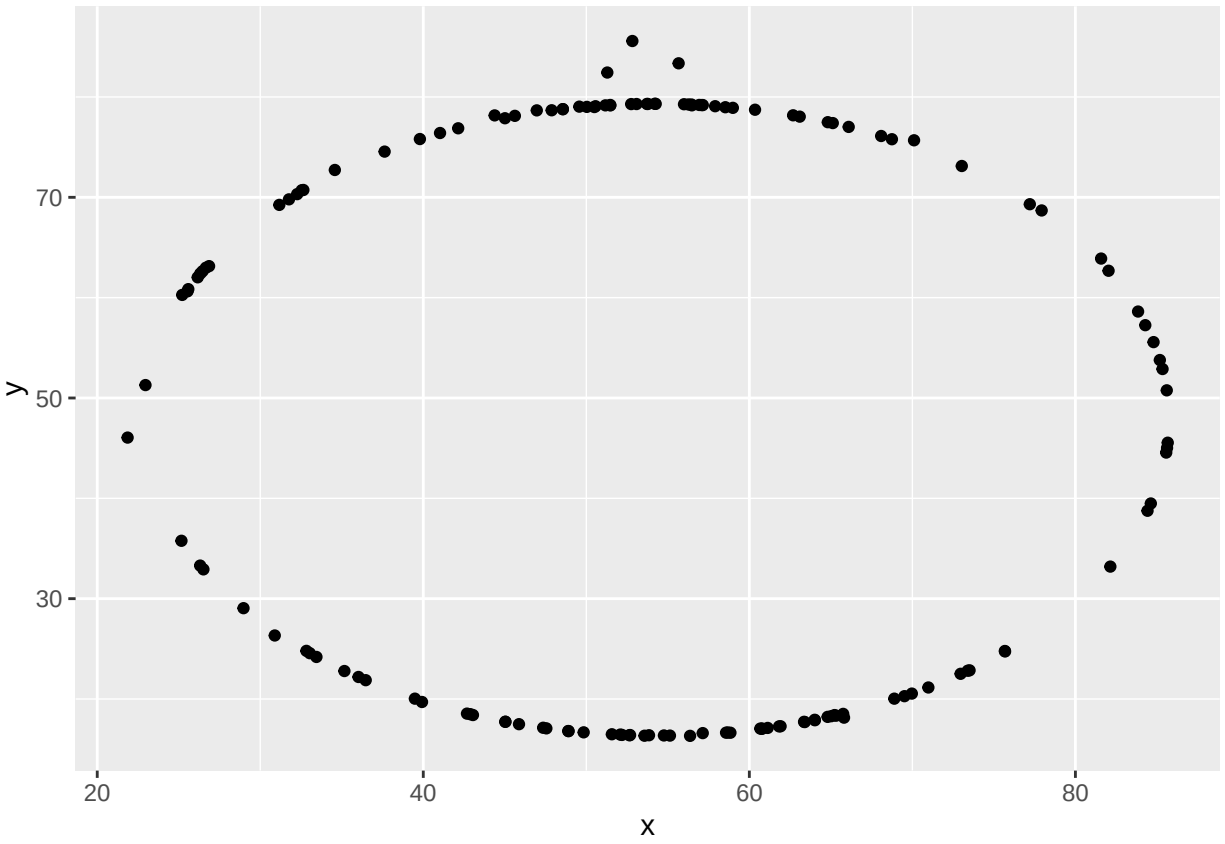
```
## # A tibble: 1 x 1
##         r
##     <dbl>
## 1 -0.0630
```

3. Plot y vs. x for the star dataset. You can (and should) reuse code we introduced above, just replace the dataset name with the desired dataset. Then, calculate the correlation coefficient between x and y for this dataset. How does this value compare to the r of dino?

the correlation coef is -0.0630, which is slightly less than the correlation coef of the dino.

```
dino_data <- datasaurus_dozen %>%
  filter(dataset == "circle")

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



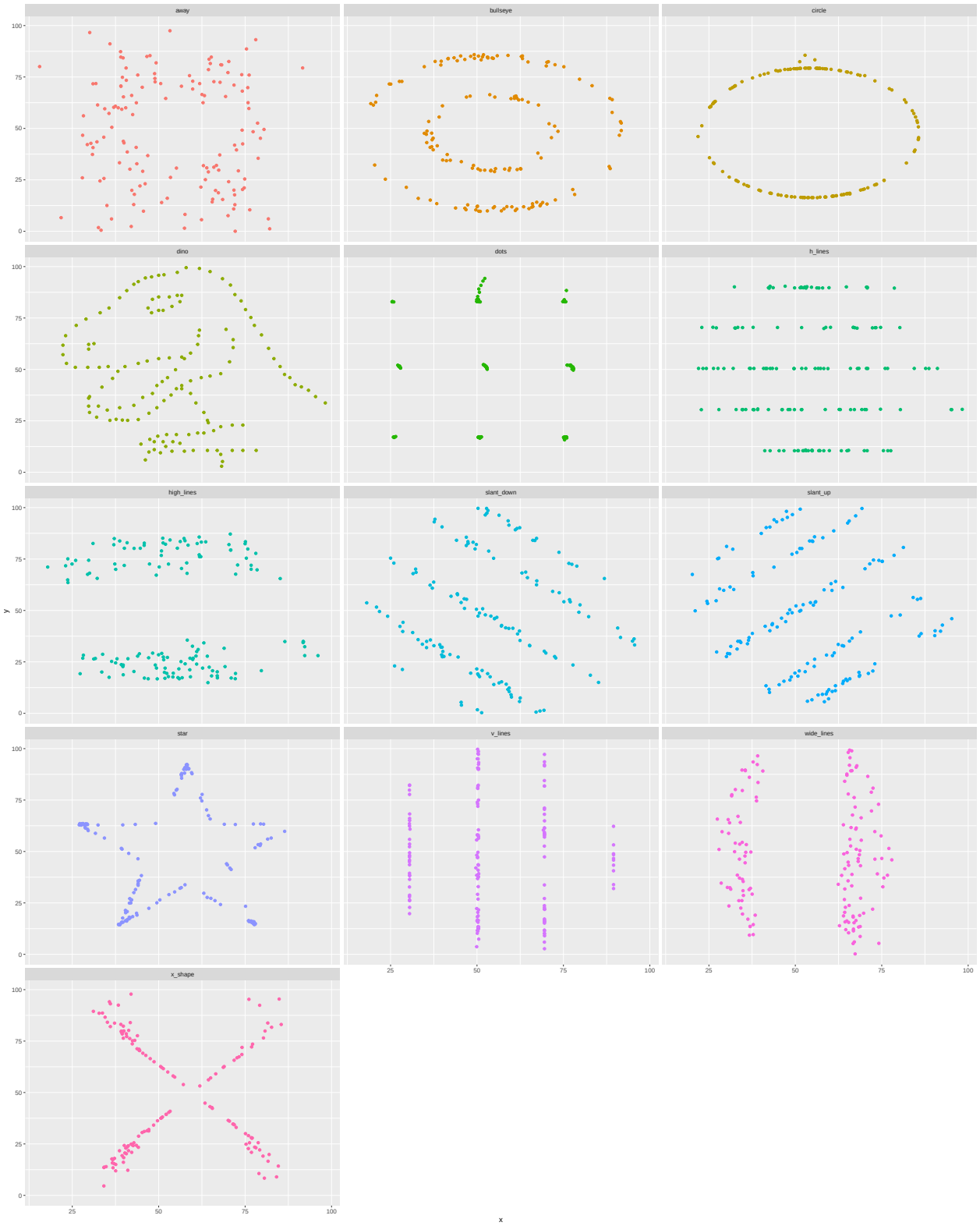

```
dino_data %>%
  summarize(r = cor(x, y))
```

```
## # A tibble: 1 x 1
##         r
##     <dbl>
## 1 -0.0683
```

4. Plot y vs. x for the circle dataset. You can (and should) reuse code we introduced above, just replace the dataset name with the desired dataset. Then, calculate the correlation coefficient between x and y for this dataset. How does this value compare to the r of dino?

the correlation coef for the circle dataset is -0.0683, which is higher than the dino correlation coef

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




```
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
```

5. Finally, let's plot all datasets at once. In order to do this we will make use of facetting.

```
##i didnt mean to do this and idk how to delete it :()
```

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

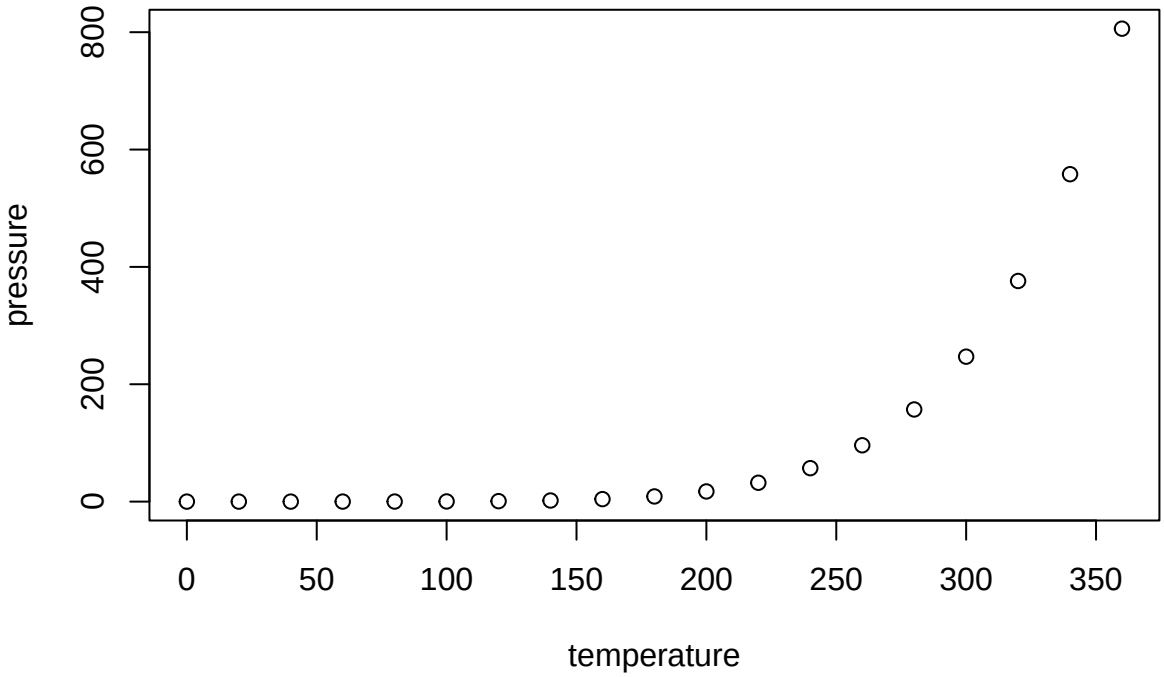
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.