Datasaurus Analysis

Prabuddha Deore

2025-05-09

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.4.3
```

```
## -- 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   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.2
## -- 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 errors
```

```
library(datasauRus)
```

```
## Warning: package 'datasauRus' was built under R version 4.4.3
```

```
#?datasaurus_dozen
```

```
# Count the rows and columns and list the variables
dim(datasaurus_dozen)
```

```
## [1] 1846    3
```

```
names(datasaurus_dozen)
```

```
## [1] "dataset" "x"       "y"
```

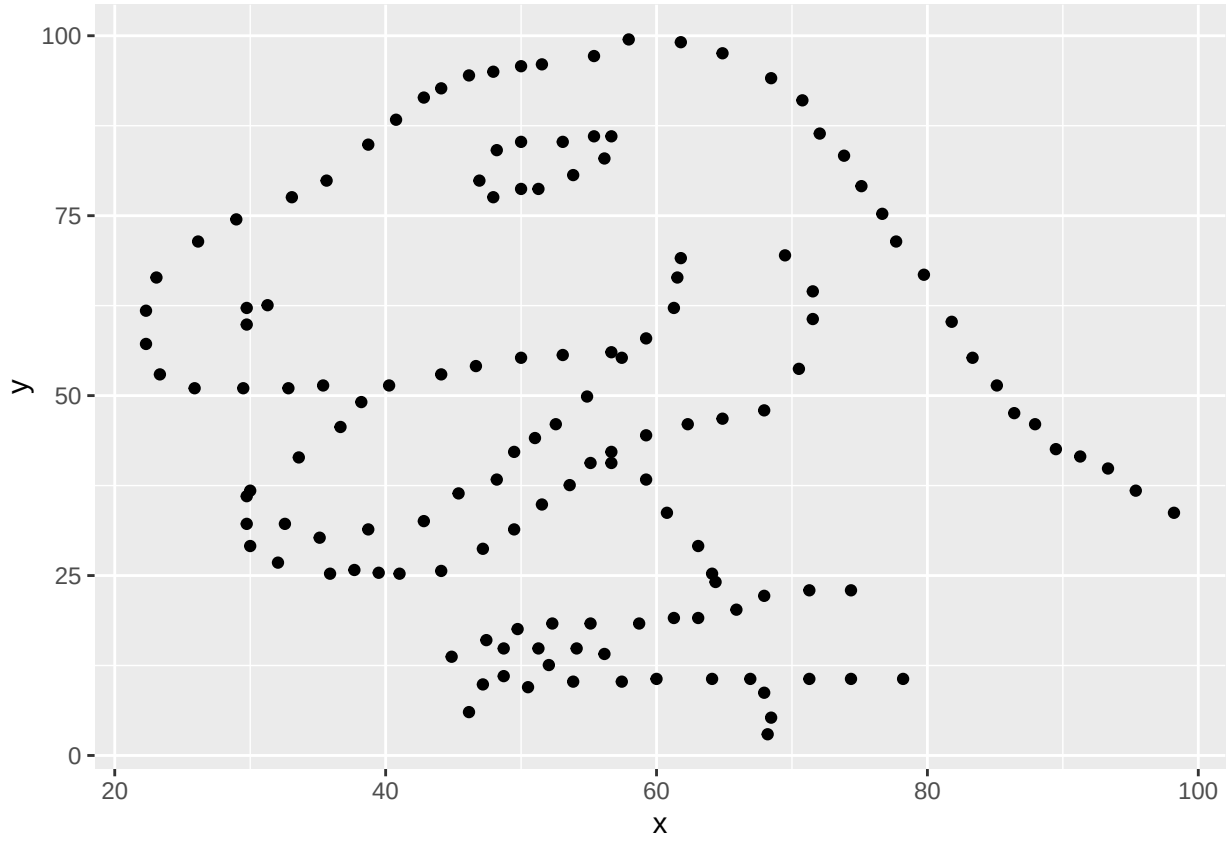
```
# Filter for the 'dino' dataset
dino_data <- datasaurus_dozen %>%
  filter(dataset == "dino")

# Preview the first few rows
head(dino_data)
```

```
## # A tibble: 6 x 3
##   dataset     x     y
##   <chr>   <dbl> <dbl>
## 1 dino     55.4  97.2
## 2 dino     51.5  96.0
## 3 dino     46.2  94.5
## 4 dino     42.8  91.4
## 5 dino     40.8  88.3
## 6 dino     38.7  84.9
```

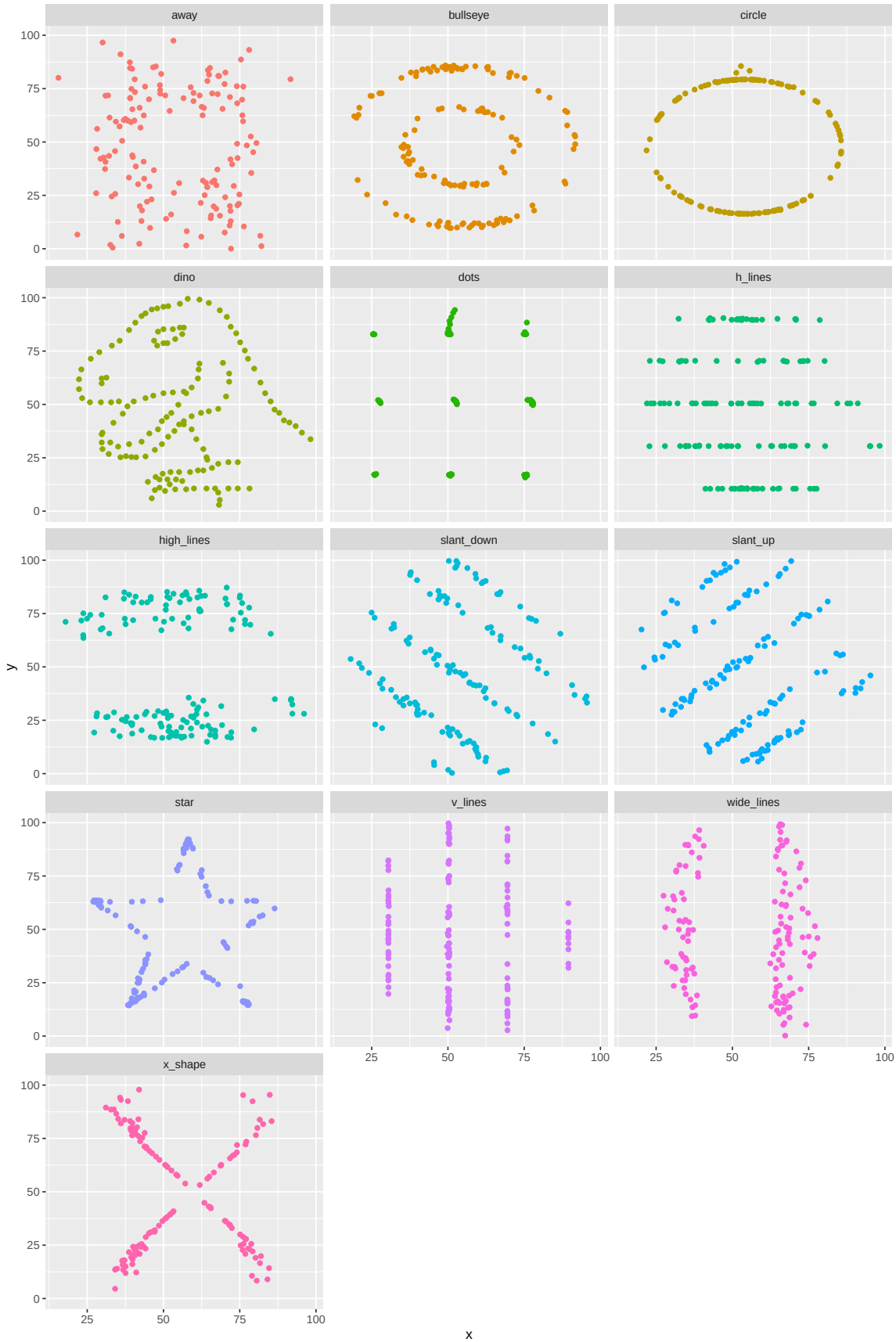
```r
# Compute summary statistics
summary_stats <- dino_data %>%
  summarise(
    mean_x = mean(x),
    mean_y = mean(y),
    sd_x = sd(x),
    sd_y = sd(y),
    correlation = cor(x, y)
  )

# Print the summary
summary_stats
```

```
## # A tibble: 1 x 5
##   mean_x mean_y  sd_x  sd_y correlation
##    <dbl>  <dbl> <dbl> <dbl>       <dbl>
## 1   54.3   47.8  16.8  26.9     -0.0645
```

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




```r
ggplot(datasaurus_dozen, aes(x = x, y = y, color = dataset))+
  geom_point(size= 1.5)+
  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
x
0
25
50
75
100

Summary

The Datasaurus Dozen data set highlights the importance of data visualization. While the summary statistics, such as mean, standard deviation, and correlation, are very similar for these data sets, their visual forms show very different shapes and patterns. This fact demonstrates that using only statistical measures can lead to incorrect interpretations and highlights the importance of using visualizations to identify underlying structures or anomalies in the data.