Problem 10.1

Chen Bo Calvin Zhang

02/01/2021

Let us first load and visualise the data set.

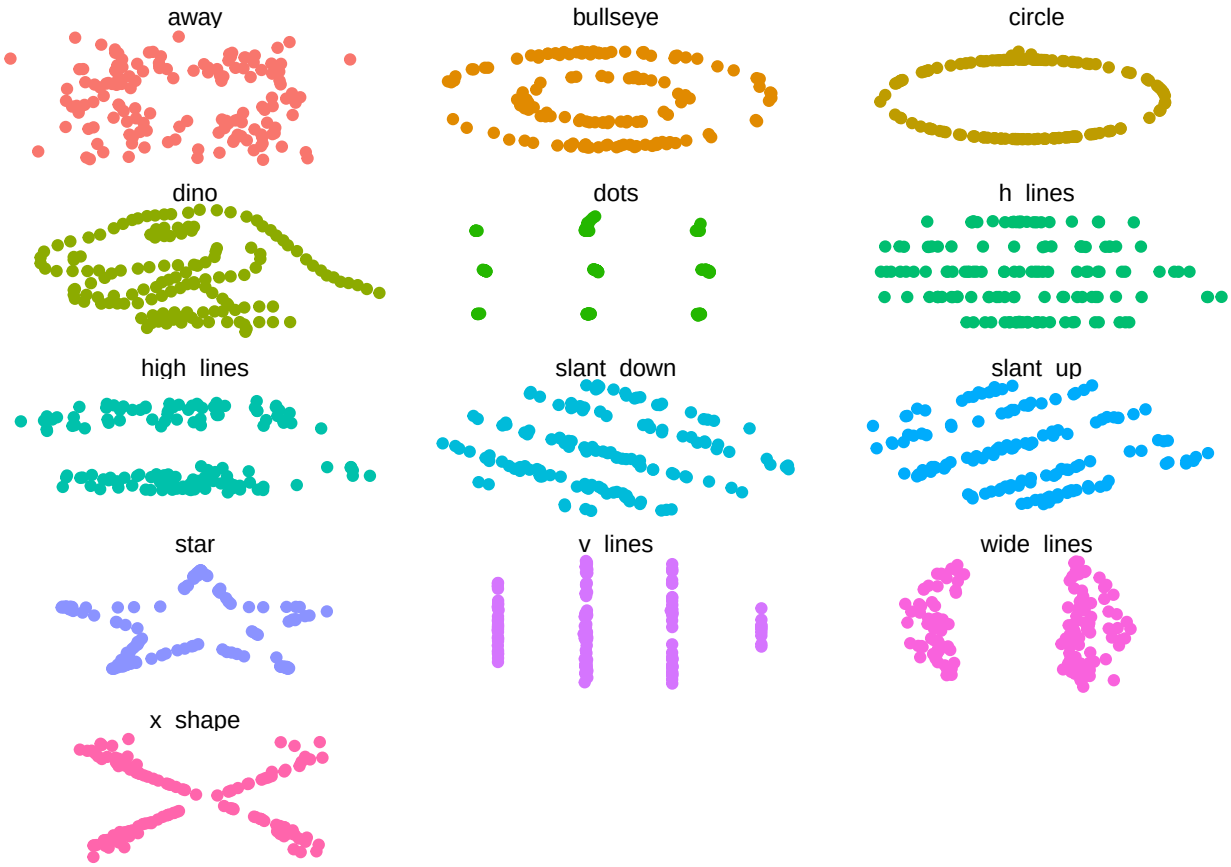
```
# load data set
library("datasauRus")
print(dim(datasaurus_dozen)) # 1846 3
```

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

```
dsname = factor(unlist(datasaurus_dozen[,1]))
print(levels(dsname))
```

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

```
# plot data sets
library("ggplot2")
ggplot(datasaurus_dozen,
  aes(x = x, y = y, colour = dataset)) +
  geom_point() +
  theme_void() +
  theme(legend.position = "none") +
  facet_wrap( ~ dataset, ncol = 3)
```



Now, let us fit a linear model thourgh the data and find some statistics.

```
statistics = matrix(0, ncol=5, nrow=13)

for (i in 1:13)
{
  idx = (dsname == levels(dsname)[i])

  x = unlist(datasaurus_dozen[idx, 2])
  y = unlist(datasaurus_dozen[idx, 3])

  statistics[i, 1] = mean(x)
  statistics[i, 2] = mean(y)
  statistics[i, 3] = cor(x, y)
  statistics[i, 4:5] = coefficients(lm(y~x))
}

print(statistics)
```

```
##           [,1]     [,2]        [,3]     [,4]        [,5]
##  [1,] 54.26610 47.83472 -0.06412835 53.42513 -0.10301845
##  [2,] 54.26873 47.83082 -0.06858639 53.80947 -0.11016745
##  [3,] 54.26732 47.83772 -0.06834336 53.79704 -0.10981430
##  [4,] 54.26327 47.83225 -0.06447185 53.45298 -0.10358250
##  [5,] 54.26030 47.83983 -0.06034144 53.09834 -0.09691270
##  [6,] 54.26144 47.83025 -0.06171484 53.21109 -0.09916499
```

```
##  [7,] 54.26881 47.83545 -0.06850422 53.80879 -0.11006955
##  [8,] 54.26785 47.83590 -0.06897974 53.84971 -0.11081721
##  [9,] 54.26588 47.83150 -0.06860921 53.81260 -0.11021842
## [10,] 54.26734 47.83955 -0.06296110 53.32668 -0.10111300
## [11,] 54.26993 47.83699 -0.06944557 53.89084 -0.11155083
## [12,] 54.26692 47.83160 -0.06657523 53.63495 -0.10694079
## [13,] 54.26015 47.83972 -0.06558334 53.55423 -0.10531687
```

We can observe that all the data sets have very similar means and correlations, and roughly the same coefficient and intercept for the linear model, despite the data looking very different.