Datasaurus or Why Visualizations Are Important

Stephen Kaluzny

12 May, 2017

This note uses a group of datasets that were recently released to show the importance of visualizations when doing a data analysis. It illustrates how relying on significant bivariate correlations can lead one astray.

The datasets are a larger and more modern version of what is know as Anscombe's Quartet. Anscombe's quartet consisted of 4 bivariate datasets with the same mean, standard deviation and correlation but with distinctly different scatterplots.

The Data

The data is now available as an R package, `datasauRus`, on CRAN.

```
if(!suppressWarnings(require("datasauRus", quietly=TRUE, character.only=TRUE))) {
  install.packages("datasauRus")
  library("datasauRus")
}
```

There are 13 datasets consisting of x, y, variable pairs.

We will use the combined (stacked), 3 column version of the datasets (called `datasaurus_dozen`) here to easily compute statistics and plots for all the datasets at once using the `tidyverse`.

The columns have roughly the same univariate statistics:

```
datasaurus_dozen %>%
  group_by(dataset) %>%
  summarise(MeanX=mean(x), SdX=sd(x), MeanY=mean(y), SdY=sd(y))
```

```
## # A tibble: 13 × 5
##       dataset    MeanX      SdX    MeanY      SdY
##         <chr>    <dbl>    <dbl>    <dbl>    <dbl>
## 1        away 54.26610 16.76982 47.83472 26.93974
## 2    bullseye 54.26873 16.76924 47.83082 26.93573
## 3      circle 54.26732 16.76001 47.83772 26.93004
## 4        dino 54.26327 16.76514 47.83225 26.93540
## 5        dots 54.26030 16.76774 47.83983 26.93019
## 6     h_lines 54.26144 16.76590 47.83025 26.93988
## 7  high_lines 54.26881 16.76670 47.83545 26.94000
## 8  slant_down 54.26785 16.76676 47.83590 26.93610
## 9    slant_up 54.26588 16.76885 47.83150 26.93861
## 10       star 54.26734 16.76896 47.83955 26.93027
## 11    v_lines 54.26993 16.76996 47.83699 26.93768
## 12 wide_lines 54.26692 16.77000 47.83160 26.93790
## 13    x_shape 54.26015 16.76996 47.83972 26.93000
```

The (pearson) correlations between the pairs are also roughly the same (-.06) and none of the correlations are considered significant when tested with `cor.test`

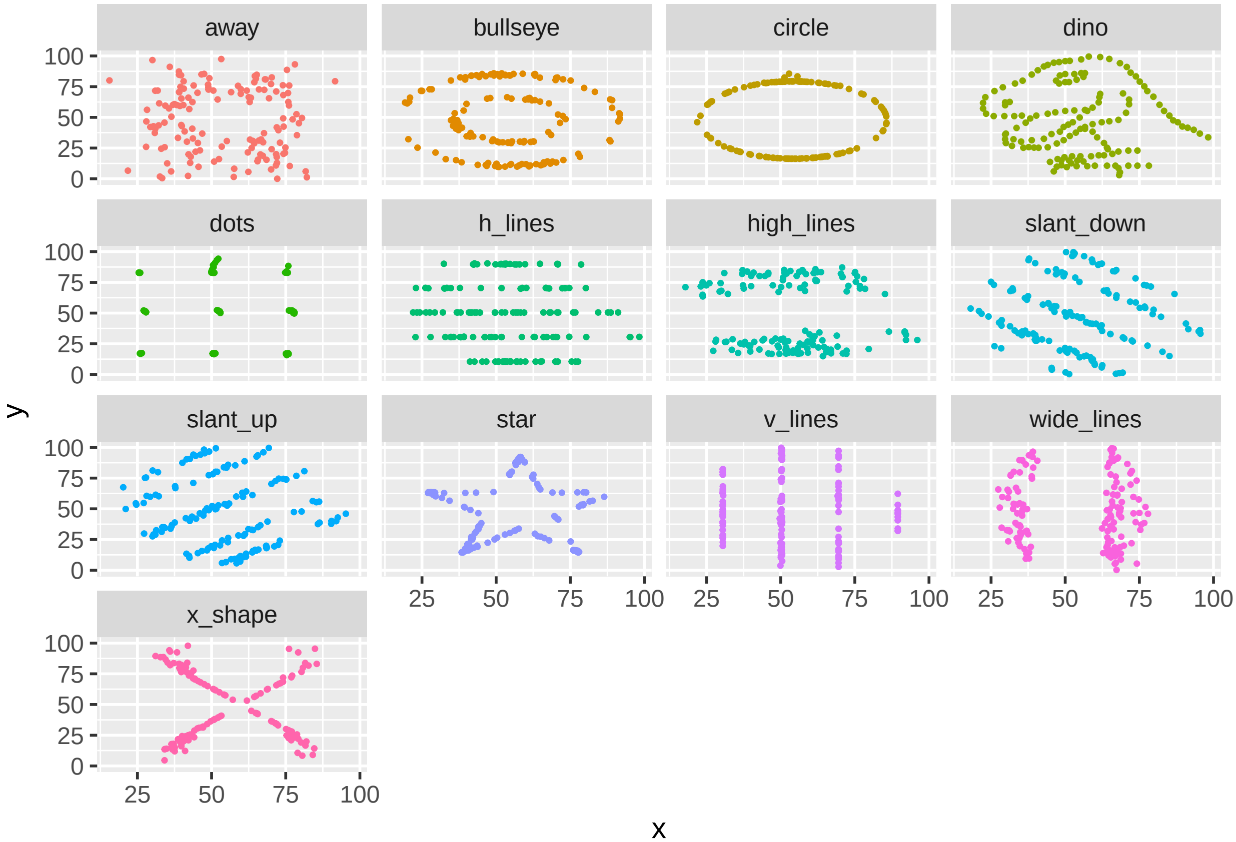
```
datasaurus_dozen %>%
  group_by(dataset) %>%
  summarise(Corr=cor(x, y), Pvalue=cor.test(x, y)$p.value)
```

```
## # A tibble: 13 × 3
##       dataset        Corr    Pvalue
##         <chr>       <dbl>     <dbl>
## 1        away -0.06412835 0.4483288
## 2    bullseye -0.06858639 0.4173467
## 3      circle -0.06834336 0.4190029
## 4        dino -0.06447185 0.4458966
## 5        dots -0.06034144 0.4756316
## 6     h_lines -0.06171484 0.4656268
## 7  high_lines -0.06850422 0.4179063
## 8  slant_down -0.06897974 0.4146744
## 9    slant_up -0.06860921 0.4171915
## 10       star -0.06296110 0.4566492
## 11    v_lines -0.06944557 0.4115226
## 12 wide_lines -0.06657523 0.4311664
## 13    x_shape -0.06558334 0.4380777
```

Visualization

The surprising result are the scatterplots for each dataset. Clearly, there is some strong structure in each dataset yet the bivariate correlations gave no indication of this.

```
datasaurus_dozen %>%
  ggplot(aes(x=x, y=y, color=dataset)) +
    geom_point(size=0.5) +
    facet_wrap(~ dataset) +
    guides(color=FALSE)
```



Appendix

The the `datasauRus` package was originally posted to Github. To install that version in R:

```
devtools::install_github("stephlocke/datasauRus")
```

The package contains information and references to how the datasets were created.