What a desaster!

Claudius

2025-05-02

Packages used

```
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   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     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 becom
```

```
library(DataScienceExercises)
library(knitr)
```

Exploring flight data

In this short text we explore the following data set on flights departing from New York.

```
base_data <- DataScienceExercises::nycflights21_small[1:200, ]
data.frame(head(DataScienceExercises::nycflights21_small, 50))
```

```
   arr_delay dep_delay month carrier distance
1        -39        -4     4      DL     2248
2        -22        -4    12      AA     1389
3          0        -4     1      B6     1076
4         -8        -1     7      UA     1608
5         -7        -4     3      DL     1035
6        -17       -10    11      YX      335
7        -50        -3     6      9E      425
8        -29        -5     1      DL     1969
9        -46        -9     5      DL     1035
10       112        92     6      UA     1605
11        50        69     4      DL     1020
12        -3        13    12      B6     1417
13       -35        -9     1      YX      264
14        -7         6     3      B6     1065
15       -14        -4     8      DL      488
16       239       266     4      AA      529
17        -9         0    11      UA     1085
18       -17        -4    12      9E      288
19         0        12     3      B6     1089
20       -46       -11     7      DL     1020
21        -6        -1     9      9E      431
22       -14        -1    11      UA     2454
23        48        54    11      YX      799
24       -20        -4    11      YX      502
25        26        28    11      DL     1598
26       263       284    10      UA     2565
27       108        43     2      B6      944
28       -13       -10    12      YX     1107
29       -35        -1     5      AA     1372
30        -6        -7     9      YX      544
31        17        -5     7      UA      997
32       129       153    11      DL      431
33       -14        -5     3      NK      550
34       -11        -3     8      UA     2454
35        -5        -2     5      UA      997
36       -11         0    10      DL     1010
37         0        -8     9      YX      214
38        13        19     5      B6     1041
39        13         2    11      DL     1990
40       -21       -10    12      YX      288
41        -9        -5     9      YX      708
42       -19        -1     8      DL      502
```

```
43         8        -3    12      YX      541
44       -26        -4    11      DL     1010
45       -11         2     8      DL     2475
46       -20        -6    11      B6     1626
47       -24        -6     6      YX      636
48       -25        -7     6      9E      764
49        -6         9     6      YX      184
50       -13        -5     9      YX      184
```

To have a first look on the relationship of the variables, consider the following scatter plots:

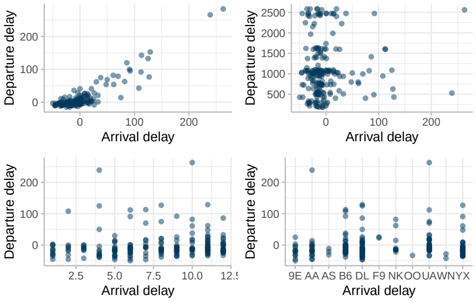
```r
arrival_dep <- ggplot(data = base_data) +
  geom_point(mapping = aes(x=arr_delay, y=dep_delay),
             alpha=0.5, color="#00395B") +
  ggplot2::theme_bw() +
  labs(x="Arrival delay", y="Departure delay") +
  theme(
    legend.position = "bottom",
    legend.title = ggplot2::element_blank(),
    panel.border = ggplot2::element_blank(),
    axis.line = ggplot2::element_line(colour = "grey"),
    axis.ticks = ggplot2::element_line(colour = "grey")
  )

arrival_dist <- ggplot(data = base_data) +
  geom_point(mapping = aes(x=arr_delay, y=distance),
             alpha=0.5, color="#00395B") +
  ggplot2::theme_bw() +
  labs(x="Arrival delay", y="Departure delay") +
  theme(
    legend.position = "bottom",
    legend.title = ggplot2::element_blank(),
    panel.border = ggplot2::element_blank(),
    axis.line = ggplot2::element_line(colour = "grey"),
    axis.ticks = ggplot2::element_line(colour = "grey")
  )

arrival_month <- ggplot(data = base_data) +
  geom_point(mapping = aes(y=arr_delay, x=month),
             alpha=0.5, color="#00395B") +
  ggplot2::theme_bw() +
  labs(x="Arrival delay", y="Departure delay") +
  theme(
    legend.position = "bottom",
    legend.title = ggplot2::element_blank(),
    panel.border = ggplot2::element_blank(),
    axis.line = ggplot2::element_line(colour = "grey"),
    axis.ticks = ggplot2::element_line(colour = "grey")
  )

arrival_carrier <- ggplot(data = base_data) +
  geom_point(mapping = aes(y=arr_delay, x=carrier),
             alpha=0.5, color="#00395B") +
  ggplot2::theme_bw() +
  labs(x="Arrival delay", y="Departure delay") +
  theme(
    legend.position = "bottom",
    legend.title = ggplot2::element_blank(),
    panel.border = ggplot2::element_blank(),
    axis.line = ggplot2::element_line(colour = "grey"),
    axis.ticks = ggplot2::element_line(colour = "grey")
  )

ggpubr::ggarrange(
  arrival_dep, arrival_dist,
  arrival_month, arrival_carrier,
  ncol = 2, nrow = 2)
```



This suggests that there is a strong correlation between departure and arrival delay. To compute the correlation we might use the following R code:

```
[1] 0.9114122
```

There is indeed a very strong correlation. But is it significant? Lets check it using the Pearson correlation test:

```
cor.test(base_data$arr_delay, base_data$dep_delay, method = "pearson")
```

```

	Pearson's product-moment correlation

data:  base_data$arr_delay and base_data$dep_delay
t = 31.166, df = 198, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8845188 0.9322677
sample estimates:
      cor 
0.9114122 
```

Of course, these are just preliminary results, from a methodological point of view there is still much to do…