データクリーニング

Naoko Ishibashi

2025-02-02

```
# このスクリプトを「YOUR LAST NAME_HW2.R」として保存し、Canvasにアップロードしてください。
# また、レポートとグラフを含むWordドキュメントもアップロードしてください。
# 以下のセクションにコードを入力してください。
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.5.1     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.1
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflic
ts to become errors
```

```
library(nycflights13)
?flights
View(flights)
```

```
#-----------------------------------------------
#  1

not_cancelled <- flights %>%
  filter(!is.na(dep_delay), !is.na(arr_delay))
#-----------------------------------------------
```

“not_cancelled」という新しい変数を作成することで、コードが簡単になります。 なぜなら、「not_cancelled」変数は、実際に出発したフライトのみを含み、”dep_delay”と”arr_delay”をフィルタリングすることで、欠損データを毎回フィルタリングする必要がなくなるからです。

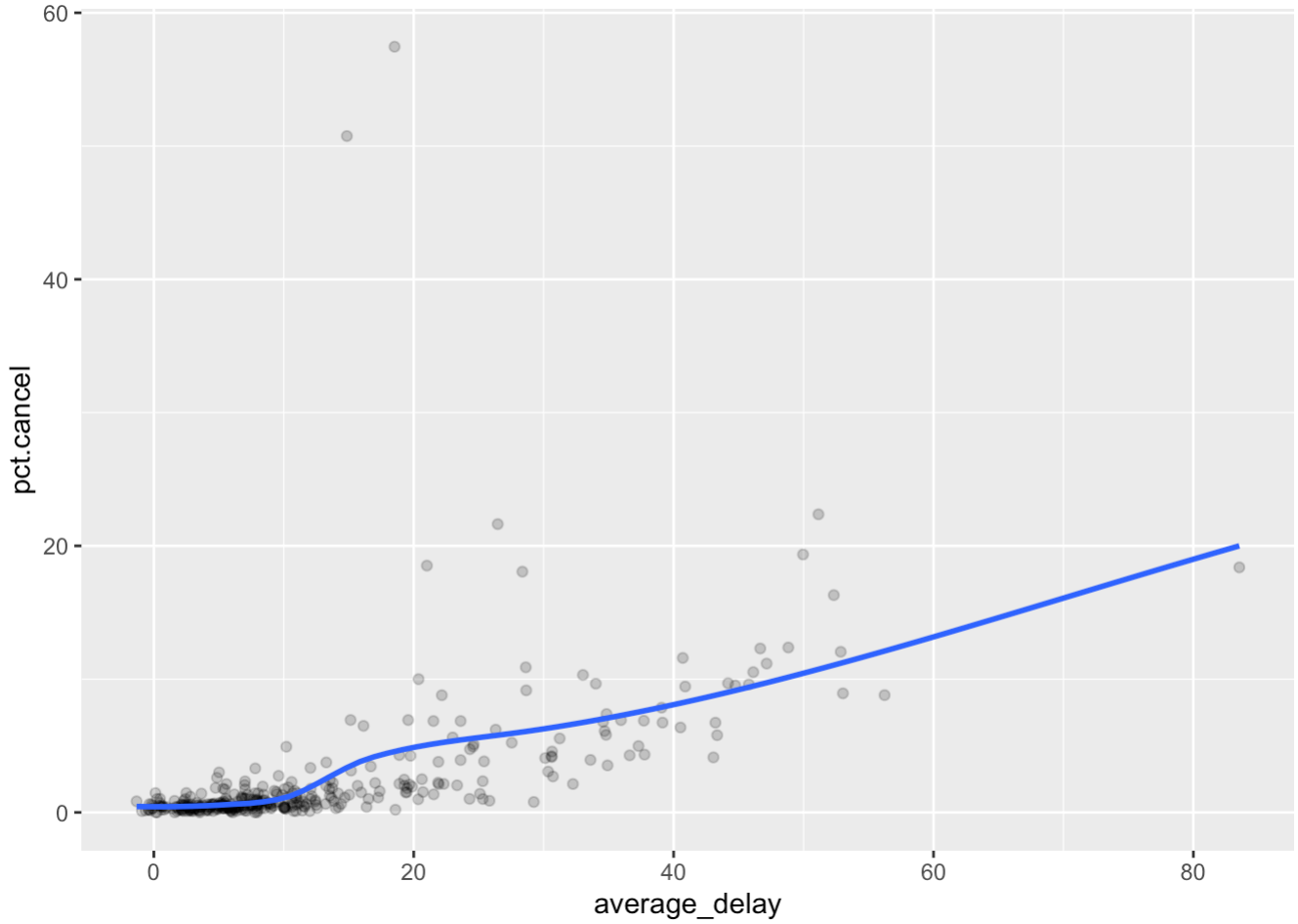
```
#-------------------------------------------
#  2
# 1日の平均遅延とその日にキャンセルされたフライトの割合に関係があるのではないかと疑うかもしれません。
# 例えば、天気が悪ければ、多くのフライトは遅れて出発し、その後キャンセルされるかもしれません。
# この直感をテストしてみましょう。
# 最初に、毎日の平均遅延とキャンセルされたフライトの割合を求めます。
# 次に、それらをプロットし、関係についてコメントします。私たちの直感は正しかったでしょうか？

# ANSWER
# P.32 キャンセルされたフライト
# P.16 平均遅延
relationship <- flights %>%
  mutate(cancelled = is.na(dep_time)) %>%
  group_by(year, month, day) %>%                          # 毎日
  summarise(average_delay = mean(dep_delay, na.rm=TRUE),  # 平均遅延
            pct.cancel = 100*mean(cancelled),             # キャンセルされたフライトの割合
            flights = n())
```

```
## `summarise()` has grouped output by 'year', 'month'. You can override using the
## `.groups` argument.
```

```
ggplot(data = relationship, mapping = aes(x = average_delay, y = pct.cancel)) +  # P.
17
  geom_point(alpha = 1/5) +
  geom_smooth(se = FALSE)
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```




```
#------------------------------------------------
#  3
# 飛行機で遅延を経験するのは誰も好ましくありません。これを避けようとする場合、
# 1日のどの時間帯が出発遅延が最も少ないのか気になるかもしれません。
# その時間帯は何時でしょうか？
# また、各時間帯で定刻または早めに出発するフライトの割合（つまり、見つけたいフライト！）を計算してみましょう。
# どの時間帯がこれらのフライトを最も見つけやすいのでしょうか？

# 3A
## 1日のどの時間帯が出発遅延が最も少ないか。何時でしょうか？
# ページ：18
# 5時が最も出発遅延が少ないようです
not_cancelled %>%
  group_by(hour) %>%      # 1日のどの時間帯か
  summarise(
            n(),          # フライトの総数を取得
            delay = mean(dep_delay, na.rm=T))  # 出発遅延を計算
```

```
## # A tibble: 19 × 3
##     hour `n()`  delay
##    <dbl> <int>  <dbl>
##  1     5  1940  0.689
##  2     6 25447  1.60
##  3     7 22475  1.91
##  4     8 26734  4.11
##  5     9 19931  4.54
##  6    10 16370  6.45
##  7    11 15689  7.15
##  8    12 17744  8.52
##  9    13 19457 11.3
## 10    14 21022 13.7
## 11    15 23082 16.8
## 12    16 22045 18.6
## 13    17 23667 21.0
## 14    18 21072 21.0
## 15    19 20507 24.7
## 16    20 16061 24.2
## 17    21 10503 24.2
## 18    22  2558 18.7
## 19    23  1042 14.0
```

```
# 3B
## 各時間帯で定刻または早めに出発するフライトの割合を計算
# ページ：27
not_cancelled %>%
  group_by(hour) %>%
  summarise(
    n(),                                    # フライトの総数を取得
    hour_perc = 100*mean(dep_delay <= 0) # 定刻または早めに出発したフライトの割合を計算
  ) %>%                                     # 負の遅延は早めの出発、ゼロは定刻出発
arrange(hour_perc)                          # 定刻または早めに出発した割合で並べ替え
```

```
## # A tibble: 19 × 3
##     hour `n()` hour_perc
##    <dbl> <int>     <dbl>
##  1    20 16061      46.4
##  2    21 10503      46.9
##  3    19 20507      47.4
##  4    17 23667      49.0
##  5    18 21072      49.8
##  6    15 23082      51.2
##  7    16 22045      51.8
##  8    22  2558      53.8
##  9    23  1042      56.2
## 10    14 21022      56.4
## 11    13 19457      58.2
## 12    12 17744      64.0
## 13    11 15689      68.1
## 14    10 16370      69.9
## 15     9 19931      73.1
## 16     8 26734      74.7
## 17     5  1940      74.9
## 18     7 22475      78.0
## 19     6 25447      78.7
```

3C 1日のどの時間帯で定刻または早めに出発するフライトが最も多く見つかるか？ 3Bの情報を使って調べてみましょう。 6時が定刻または早めに出発するフライトの割合が高いようです。

```
#-----------------------------------------------
# Question 4
# どの航空会社が出発遅延が30分以上である可能性が高いか？
# ヒント: ifelse()関数を使うと便利かもしれません。

# ANSWER
# P.22 ifelse関数を使用
# SkyWest Airlines Incが最も出発遅延が30分以上である可能性が高い
flights %>%
  mutate(departure_delay = ifelse(dep_delay >= 30, 1, 0)) %>%   # 出発遅延が30分以上か？
  group_by(carrier) %>%                                          # どの航空会社か
  summarise(pct.dep_delay = 100 * mean(departure_delay)) %>%     # 出発遅延の割合
  arrange(desc(pct.dep_delay))                                   # 出発遅延の割合が高い順
に並べ替え
```

```
## # A tibble: 16 × 2
##    carrier pct.dep_delay
##    <chr>           <dbl>
##  1 HA               4.68
##  2 9E              NA
##  3 AA              NA
##  4 AS              NA
##  5 B6              NA
##  6 DL              NA
##  7 EV              NA
##  8 F9              NA
##  9 FL              NA
## 10 MQ              NA
## 11 OO              NA
## 12 UA              NA
## 13 US              NA
## 14 VX              NA
## 15 WN              NA
## 16 YV              NA
```

```
airlines
```

```
## # A tibble: 16 × 2
##    carrier name
##    <chr>   <chr>
##  1 9E      Endeavor Air Inc.
##  2 AA      American Airlines Inc.
##  3 AS      Alaska Airlines Inc.
##  4 B6      JetBlue Airways
##  5 DL      Delta Air Lines Inc.
##  6 EV      ExpressJet Airlines Inc.
##  7 F9      Frontier Airlines Inc.
##  8 FL      AirTran Airways Corporation
##  9 HA      Hawaiian Airlines Inc.
## 10 MQ      Envoy Air
## 11 OO      SkyWest Airlines Inc.
## 12 UA      United Air Lines Inc.
## 13 US      US Airways Inc.
## 14 VX      Virgin America
## 15 WN      Southwest Airlines Co.
## 16 YV      Mesa Airlines Inc.
```

```
#-----------------------------------------
# Question 5
# どの目的地が最も小さい平均到着遅延を持っているか？

# ANSWER: ANC -2.5 (Ted Stevens Anchorage Intl)
# P.31 どの目的地が最も小さい平均到着遅延を持っているか
# P.18 平均到着遅延
flights %>%
  mutate(cancelled = is.na(dep_time)) %>%
  group_by(dest) %>%     # 目的地(dest)でグループ化
  summarise(delay = mean(arr_delay, na.rm=TRUE)) %>%   # 到着遅延の平均を計算
  arrange(delay)         # 最も小さい到着遅延を最初に表示
```

```
## # A tibble: 105 × 2
##    dest     delay
##    <chr>    <dbl>
##  1 LEX   -22
##  2 PSP   -12.7
##  3 SNA    -7.87
##  4 STT    -3.84
##  5 ANC    -2.5
##  6 HNL    -1.37
##  7 SEA    -1.10
##  8 MVY    -0.286
##  9 LGB    -0.0620
## 10 SLC     0.176
## # ℹ 95 more rows
```

```
# ANCはTed Stevens Anchorage Intlの空港コード
filter(airports, faa == "ANC")
```

```
## # A tibble: 1 × 8
##   faa   name                          lat   lon   alt    tz dst   tzone
##   <chr> <chr>                       <dbl> <dbl> <dbl> <dbl> <chr> <chr>
## 1 ANC   Ted Stevens Anchorage Intl  61.2 -150.   152    -9 A     America/Anchor…
```