r4ds Ex 5.6.7

MW

2019/05/29

5.6.7

1

> Brainstorm at least 5 different ways to assess the typical delay characteristics of a group of flights. Consider the following scenarios: - A flight is 15 minutes early 50% of the time, and 15 minutes late 50% of the time. - A flight is always 10 minutes late. - A flight is 30 minutes early 50% of the time, and 30 minutes late 50% of the time. - 99% of the time a flight is on time. 1% of the time it's 2 hours late. Which is more important: arrival delay or departure delay?

I think always being late is hard to cost for the passengers because passengers will act depending on its. Most important is the probability, and the next one is the delay time.

2

> Come up with another approach that will give you the same output as `not_cancelled %>% count(dest)` and `not_cancelled %>% count(tailnum, wt = distance)` (without using `count()`).

The former is capable written as follows:

```r
not_cancelled %>% group_by(dest) %>%
    summarize(n=length(dest))
```

```
## # A tibble: 104 x 2
##    dest      n
##    <chr> <int>
##  1 ABQ     254
##  2 ACK     264
##  3 ALB     418
##  4 ANC       8
##  5 ATL   16837
##  6 AUS    2411
##  7 AVL     261
##  8 BDL     412
##  9 BGR     358
## 10 BHM     269
## # ... with 94 more rows
```

The latter is capable written as follows:

```r
not_cancelled %>% group_by(tailnum) %>%
    summarize(n=sum(distance))
```

```
## # A tibble: 4,037 x 2
##    tailnum      n
##    <chr>    <dbl>
##  1 D942DN    3418
##  2 N0EGMQ  239143
##  3 N10156  109664
```

```
##  4 N102UW   25722
##  5 N103US   24619
##  6 N104UW   24616
##  7 N10575  139903
##  8 N105UW   23618
##  9 N107US   21677
## 10 N108UW   32070
## # ... with 4,027 more rows
```

3

> Our definition of cancelled flights (`is.na(dep_delay) | is.na(arr_delay)`) is slightly suboptimal. Why? Which is the most important column?

Logically, `arr_delay < dep_delay` because it may not arrive even if it departed. So more important is `arr_delay`, I think.

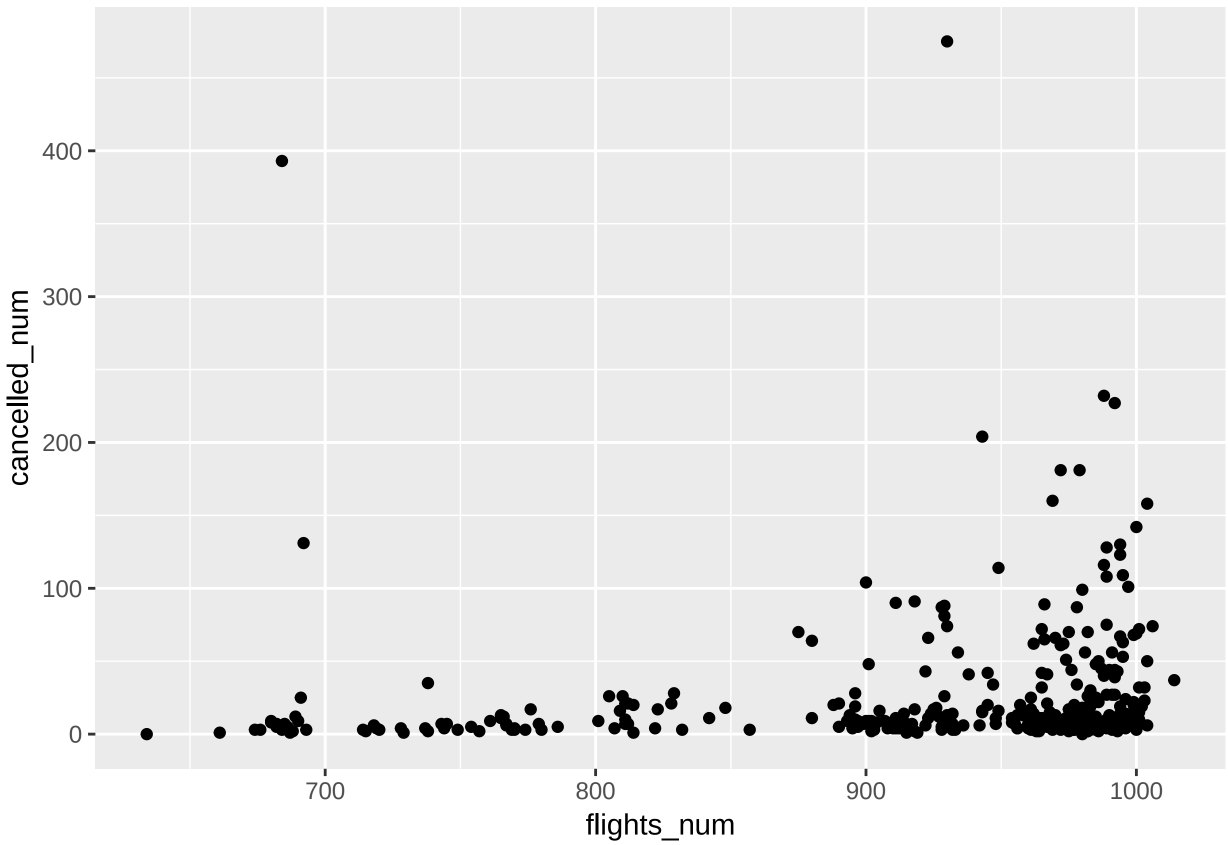
4

> Look at the number of cancelled flights per day. Is there a pattern? Is the proportion of cancelled flights related to the average delay?

```
ex4 <- flights %>%
    group_by(year, month, day) %>%
    summarize(cancelled_num = sum(is.na(arr_delay) | is.na(dep_delay)),
        flights_num = n())
ex4
```

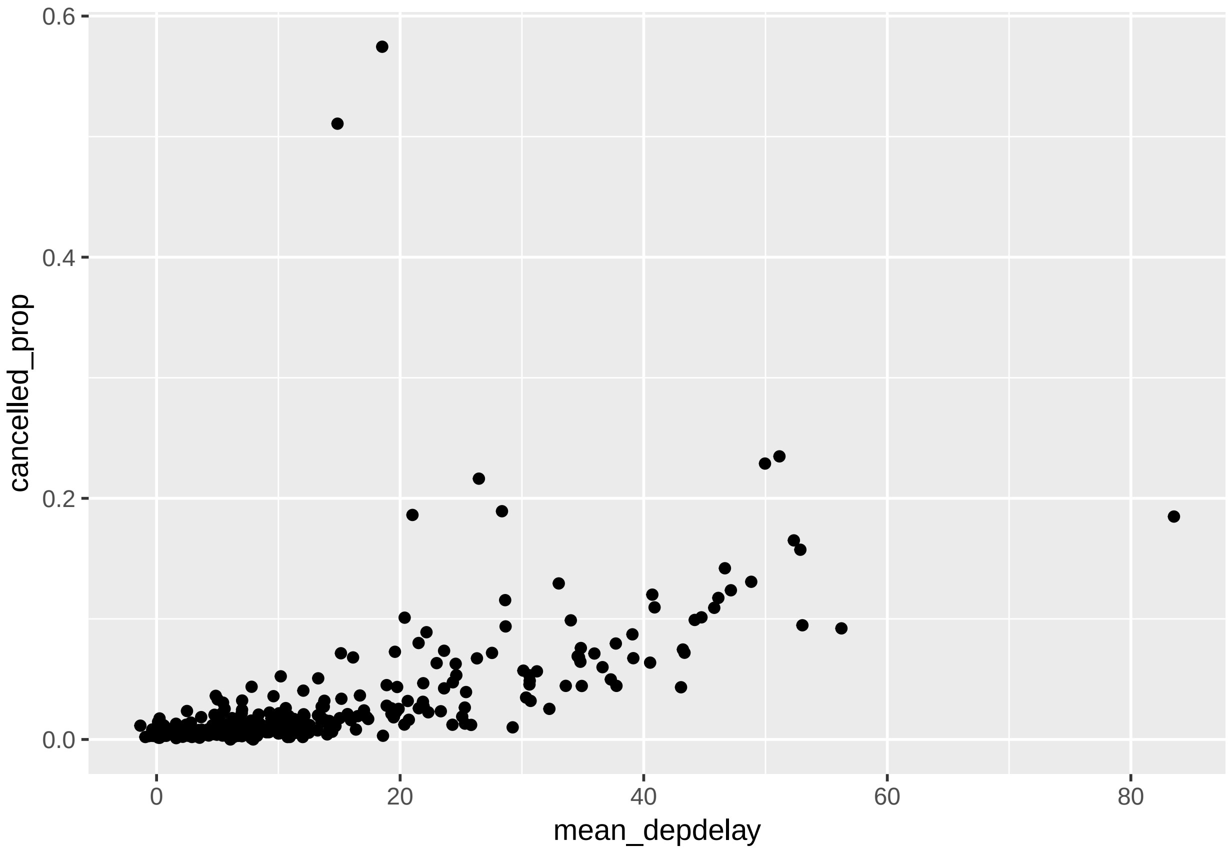
```
## # A tibble: 365 x 5
## # Groups:   year, month [?]
##     year month   day cancelled_num flights_num
##    <int> <int> <int>         <int>       <int>
##  1  2013     1     1            11         842
##  2  2013     1     2            15         943
##  3  2013     1     3            14         914
##  4  2013     1     4             7         915
##  5  2013     1     5             3         720
##  6  2013     1     6             3         832
##  7  2013     1     7             3         933
##  8  2013     1     8             7         899
##  9  2013     1     9             9         902
## 10  2013     1    10             3         932
## # ... with 355 more rows
```

```
ex4 %>% ggplot() +
    geom_point(aes(x=flights_num, y=cancelled_num))
```

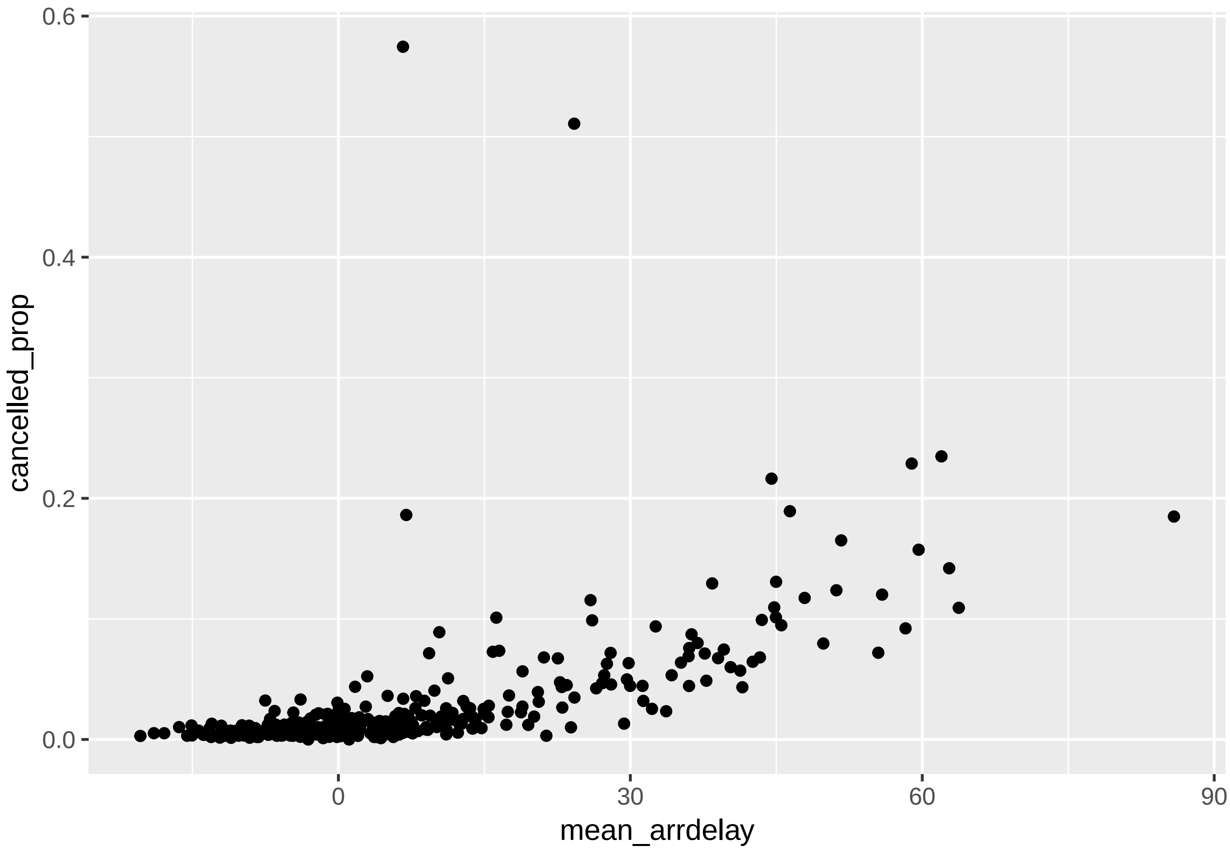



```r
ex4 <- flights %>%
    group_by(year, month, day) %>%
    summarize(cancelled_prop = mean(is.na(arr_delay) | is.na(dep_delay)),
        mean_depdelay = mean(dep_delay, na.rm = TRUE),
        mean_arrdelay = mean(arr_delay, na.rm = TRUE)) %>%
    ungroup()

ex4 %>% ggplot() +
    geom_point(aes(x=mean_depdelay, y=cancelled_prop))
```




```r
ex4 %>% ggplot() +
    geom_point(aes(x=mean_arrdelay, y=cancelled_prop))
```



These figures shows that increasing departure and arrival delay cause cancell.

5

> Which carrier has the worst delays? Challenge: can you disentangle the effects of bad airports vs. bad carriers? Why/why not? (Hint: think about `flights %>% group_by(carrier, dest) %>% summarise(n())`)

```
flights %>%
    group_by(carrier) %>%
    summarize(arr_delay=mean(arr_delay, na.rm=TRUE)) %>%
    arrange(desc(arr_delay))
```

```
## # A tibble: 16 x 2
##    carrier arr_delay
##    <chr>       <dbl>
##  1 F9         21.9
##  2 FL         20.1
##  3 EV         15.8
##  4 YV         15.6
##  5 OO         11.9
##  6 MQ         10.8
##  7 WN          9.65
##  8 B6          9.46
##  9 9E          7.38
## 10 UA          3.56
## 11 US          2.13
## 12 VX          1.76
```

```
## 13 DL           1.64
## 14 AA           0.364
## 15 HA          -6.92
## 16 AS          -9.93
```

`F9` carrier is worst.

I didn't do Challenge…

6

> What does the sort argument to count() do? When might you use it?

```
?dplyr::count
sort: if 'TRUE' will sort output in descending order of 'n'
```