Flight Delay Case - week 1

Ali Ladha

2024-09-26

```
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 conflicts t
o become errors
```

```
fd <- read.csv("/Users/aliladha/Documents/Files/School Work/College/Graduate/R files/fligh
t_delay_clean (1).csv")
str(fd)
```

```
## 'data.frame':    360 obs. of  13 variables:
##  $ airline                : chr  "RegionEx" "RegionEx" "RegionEx" "RegionEx" ...
##  $ departure_date         : chr  "2008-09-01" "2008-09-01" "2008-09-01" "2008-09-02"
...
##  $ origin                 : chr  "DFW" "DFW" "DFW" "DFW" ...
##  $ destination            : chr  "MSY" "MSY" "MSY" "MSY" ...
##  $ route_code             : chr  "DFW/MSY" "DFW/MSY" "DFW/MSY" "DFW/MSY" ...
##  $ scheduled_departure    : chr  "09:10:00" "13:10:00" "18:10:00" "09:10:00" ...
##  $ scheduled_arrival      : chr  "10:40:00" "14:40:00" "19:40:00" "10:40:00" ...
##  $ actual_arrival         : chr  "11:00:00" "15:00:00" "19:58:00" "10:50:00" ...
##  $ scheduled_flight_length: int  90 90 90 90 90 90 90 90 90 90 ...
##  $ actual_flight_length   : int  110 110 108 100 101 100 99 99 99 100 ...
##  $ delay                  : int  20 20 18 10 11 10 9 9 9 10 ...
##  $ delay_indicator        : int  1 1 1 0 0 0 0 0 0 0 ...
##  $ day_of_week            : int  2 2 2 3 3 3 4 4 4 5 ...
```

Q1

```
#Compute the mean, median, 90th percentile, and standard deviation of arrival delay minute
s for RegionEx flights. Do the same for MDA flights. Contractual obligations aside, which
measure of central tendency would be most appropriate for comparing airline performance?

fd %>% group_by(airline) %>%
  summarize(mean_delay = mean(delay) %>% round(1),
            median_delay = median(delay),
            perc_90 = quantile(delay, probs = .90),
            sd_delay = sd(delay) %>% round(1))
```

```
## # A tibble: 2 × 5
##   airline  mean_delay median_delay perc_90 sd_delay
##   <chr>         <dbl>        <dbl>   <dbl>    <dbl>
## 1 MDA            10.9           13    16.1      6.3
## 2 RegionEx       15.7            9    21       27.7
```

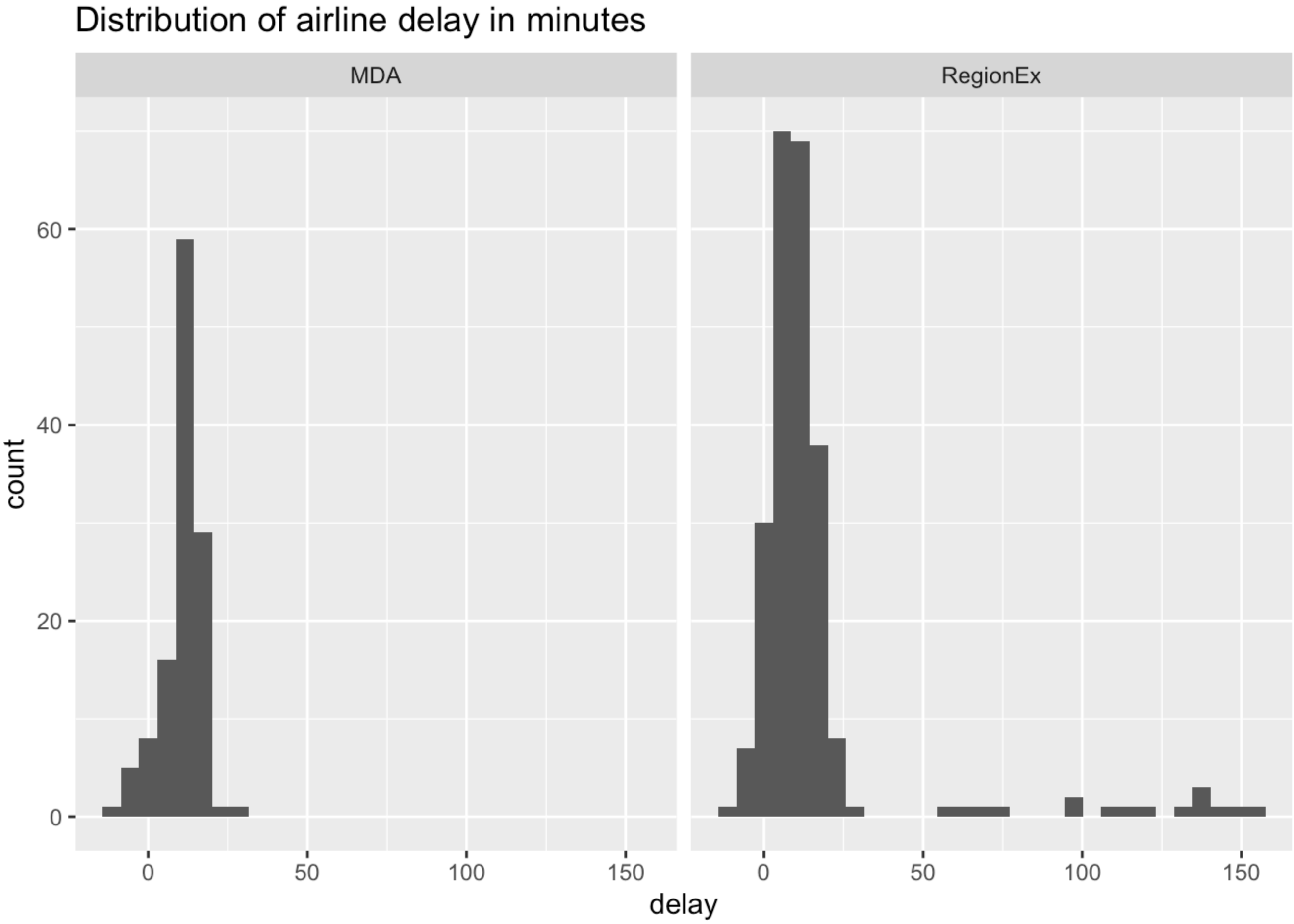
The median would be more suitable as a measure of central tendency due to not being affected by outliers. This is due to the median looking a middle value/50th percentile. Whereas the mean is sensitive to outliers since it calculates an overall average.

Q2

```
#Inspect the distribution of RegionEx’s arrival delays by constructing a histogram of the
number of arrival delay minutes of RegionEx’s flights. Do the same for MDA’s flights. Hin
t: use facet_wrap().
#How do these two distributions compare?

ggplot(data = fd, aes(x = delay)) + geom_histogram() + facet_wrap(~airline) + labs(title =
'Distribution of airline delay in minutes')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



RegionEx distribution is more spread out than MDA for its delays. RegionEx also has delays upto 50 minutes and above. But those delays are rare as some (such as the delay of 50 minutes and above) can be evaluated as outliers. Furthermore, based on the histogram, RegionEx also has a higher count for flights than MDA.

Q3

```
#So far you have considered airline performance in terms of minutes delayed. However, the
performance metrics, as noted in the case description, also include the percentage of dela
yed flights. Let’s verify that MDA’s COO is correct: does RegionEx have a higher percentag
e of delayed flights?"

#Here is code to answer that question:"

# Create a summary table of percent delayed by airline.

fd %>%
  group_by(airline) %>%
  summarize(n = n(),
            percent_delay = (mean(delay_indicator) * 100) %>% round(1))
```

```
## # A tibble: 2 × 3
##   airline      n percent_delay
##   <chr>    <int>         <dbl>
## 1 MDA        120          25.8
## 2 RegionEx   240          26.2
```

```
#Note that because delay_indicator is numeric (a binary 0/1 variable) calculating the mean of the vector returns the proportion of 1s, which, multiplied by 100, is equivalent to the percentage of delayed flights.

#Write your own code to create a table summarizing the percentage of delayed flights by airline and route.


#Notice that these tables—percent delayed by airline vs. percent delayed by airline and route— contain conflicting information. How should you answer the question of whether RegionEx has a higher percentage of delayed flights? Is the the COO correct? And, if not, why not?"
```

```
fd %>% group_by(airline, route_code) %>%
  summarize(percentage_delay = (mean(delay_indicator) * 100) %>%
              round(1))
```

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

```
## # A tibble: 8 × 3
## # Groups:   airline [2]
##   airline  route_code percentage_delay
##   <chr>    <chr>                 <dbl>
## 1 MDA      DFW/MSY                26.7
## 2 MDA      MSY/DFW                30  
## 3 MDA      MSY/PNS                20  
## 4 MDA      PNS/MSY                26.7
## 5 RegionEx DFW/MSY                25.6
## 6 RegionEx MSY/DFW                28.9
## 7 RegionEx MSY/PNS                20  
## 8 RegionEx PNS/MSY                26.7
```

The COO is incorrect: When calculating based on airline and the route_code, RegionEx is performing better than MDA in routes: DFW/MSY & MSY/DFW. Whereas it is matching MDAs delay percentage for routes: MSY/PNS & PNS/MSY.

Q4

```
#Compare the scheduled flight durations for the two airlines on each of their four routes. Also compare the actual flight durations for the two airlines. What do you notice? If the two airlines had the same scheduled duration, what impact would this have on their delay records?

fd %>% group_by(airline, route_code) %>%
  summarize(
            median_act_length = median(actual_flight_length),
            median_sch_length = median(scheduled_flight_length), n = n(),
            delay_proportion = ((median_act_length - median_sch_length) / median_sch_lengt
h * 100))
```

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

```
## # A tibble: 8 × 6
## # Groups:   airline [2]
##   airline  route_code median_act_length median_sch_length     n delay_proportion
##   <chr>    <chr>                  <dbl>             <dbl> <int>            <dbl>
## 1 MDA      DFW/MSY                114.                100    30            13.5 
## 2 MDA      MSY/DFW                113                 100    30            13   
## 3 MDA      MSY/PNS                 86                  75    30            14.7 
## 4 MDA      PNS/MSY                 79.5                75    30             6   
## 5 RegionEx DFW/MSY                100                  90    90            11.1 
## 6 RegionEx MSY/DFW                 99                  90    90            10   
## 7 RegionEx MSY/PNS                 76.5                70    30             9.29
## 8 RegionEx PNS/MSY                 77.5                70    30            10.7
```

MDA has a higher delay proportion than RegionEx. The highest delay proportion for MDA was 14.7 minutes, whereas for RegionEx, the highest delay proportion was 11.1 minutes. MDA has a better delay_proportion for the PNS/MSY route. However, MDA has a longer scheduled flight length. For routes: DFW/MSY and MSY/DFW, MDA has 10 extra minutes. For Routes MSY/PNS, PNS/MSY, MDA has 5 extra minutes. If RegionEx scheduled lengths were equal to MDA, it would outperform MDA even more since the calculated delay would be less. Furthermore, RegionEx also has more flights than MDA too.

Q5

```
#Does the data support the claim that the on-time performance of RegionEx is worse than th
at of MDA? Write a paragraph in which you argue a position. In your answer, please incorpo
rate quantitative evidence from the earlier questions.
```

```
(summary_table <- fd |>
  group_by(airline,route_code) |>
  summarize(n = n(),
            percent_delay = (mean(delay_indicator) * 100) |> round(1)) |>
  group_by(route_code) |>
  mutate(total_n = sum(n),
         prop_n = round(total_n/360,3)))
```

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

```
## # A tibble: 8 × 6
## # Groups:   route_code [4]
##   airline  route_code     n percent_delay total_n prop_n
##   <chr>    <chr>      <int>         <dbl>   <int>  <dbl>
## 1 MDA      DFW/MSY       30          26.7     120  0.333
## 2 MDA      MSY/DFW       30          30       120  0.333
## 3 MDA      MSY/PNS       30          20        60  0.167
## 4 MDA      PNS/MSY       30          26.7      60  0.167
## 5 RegionEx DFW/MSY       90          25.6     120  0.333
## 6 RegionEx MSY/DFW       90          28.9     120  0.333
## 7 RegionEx MSY/PNS       30          20        60  0.167
## 8 RegionEx PNS/MSY       30          26.7      60  0.167
```

```
#MDA weighted average
weighted.mean(x = summary_table$percent_delay[1:4],
              w = summary_table$prop_n[1:4])
```

```
## [1] 26.68
```

```
#regionEx weighted average
weighted.mean(x = summary_table$percent_delay[5:8],
              w = summary_table$prop_n[5:8])
```

```
## [1] 25.9474
```

The data does not support the claim that the on-time performance of RegionEx is worse than MDA. RegionEx has 240 flights whereas MDA has 120 flights, so a weighted mean should be used for the analysis instead of a regular mean. The weighted mean for RegionEx delays is: 25.95%, the weighted mean for MDA delays is: 26.68%. Furthermore, the median for MDA delay is 13 minutes, whereas for Regional it is 9 minutes. Additionally MDA has a longer scheduled flight length. For routes: DFW/MSY and MSY/DFW, MDA has 10 extra minutes. For Routes MSY/PNS, PNS/MSY, MDA has 5 extra minutes. Despite the additional scheduled time allotted, RegionEx is still outperforming MDA.