Ch 15 from Modern Data Science with R

Database querying using SQL

Section 15.5: Extended example: FiveThirtyEight flights

At FiveThirtyEight, Nate Silver wrote an article about airline delays using the same Bureau of Transportation Statistics data that we have in our database. We can use this article as an exercise in querying our airlines database.

The article makes a number of claims, including:

- In 2014, the 6 million domestic flights the U.S. government tracked required an extra 80 million minutes to reach their destinations. (This is a net number. It consists of about 115 million minutes of delays minus 35 million minutes saved from early arrivals.)
- The majority of flights (54%) arrived ahead of schedule in 2014.

We can begin by computing the total number of flights, the percentage of those that were on time and ahead of schedule, and the total number of minutes of delays.

```
SELECT
  SUM(1) AS numFlights,
  SUM(IF(arr_delay < 15, 1, 0)) / SUM(1) AS ontimePct,
  SUM(IF(arr_delay < 0, 1, 0)) / SUM(1) AS earlyPct,
  SUM(arr_delay) / 1e6 AS netMinLate,
  SUM(IF(arr_delay > 0, arr_delay, 0)) / 1e6 AS minLate,
  SUM(IF(arr_delay < 0, arr_delay, 0)) / 1e6 AS minEarly
FROM flights AS o
WHERE year = 2014
LIMIT 0, 6;
```

```
mydataframe
```

```
##   numFlights ontimePct earlyPct netMinLate  minLate  minEarly
## 1    5819811    0.7868   0.5424   41.61157 77.61574 -36.00417
```

We see the right number of flights (about 6 million), and the percentage of flights that were early (about 54%) is also about right. The total number of minutes early (about 36 million) is also about right. However, the total number of minutes late is way off (about 78 million vs. 115 million), and as a consequence, so is the net number of minutes late (about 42 million vs. 80 million).

In this case, you have to read the fine print. A description of the methodology used in this analysis contains some information about the "estimates" of the arrival delay for cancelled flights. The problem is that cancelled flights have an arr_delay value of 0, yet in the real-world experience of travelers, the practical delay is much longer. The FiveThirtyEight data scientists concocted an estimate of the actual delay experienced by travelers due to cancelled flights; they determined a quick-and-dirty answer that cancelled flights are

associated with a delay of four or five hours, on average. However, the calculation varies based on the particular circumstances of each flight.

As a result, reproducing the estimates made by FiveThirtyEight is likely impossible, and certainly beyond the scope of what we can accomplish here. Since we only care about the aggregate number of minutes, we can amend our computation to add, say, 270 minutes of delay time for each cancelled flight to get a ballpark answer.

```
SELECT
  SUM(1) AS numFlights,
  SUM(IF(arr_delay < 15, 1, 0)) / SUM(1) AS ontimePct,
  SUM(IF(arr_delay < 0, 1, 0)) / SUM(1) AS earlyPct,
  SUM(IF(cancelled = 1, 270, arr_delay)) / 1e6 AS netMinLate,
  SUM(
    IF(cancelled = 1, 270, IF(arr_delay > 0, arr_delay, 0))
  ) / 1e6 AS minLate,
  SUM(IF(arr_delay < 0, arr_delay, 0)) / 1e6 AS minEarly
FROM flights AS o
WHERE year = 2014
LIMIT 0, 6;
```

```
mydataframe
```

```
##   numFlights ontimePct earlyPct netMinLate  minLate  minEarly
## 1    5819811    0.7868   0.5424   75.89725 111.9014 -36.00417
```

This correction puts us in the neighborhood of the estimates from the article. One has to read the fine print to properly vet these estimates. The problem is not that the estimates reported by Silver are inaccurate – on the contrary, they seem plausible and are certainly better than not correcting for cancelled flights at all. However, it is not immediately clear from reading the article (you have to read the separate methodology article) that these estimates – which account for roughly 25% of the total minutes late reported – are in fact estimates and not hard data.

Later in the article, Silver presents a figure that breaks down the percentage of flights that were on time, had a delay of 15 to 119 minutes, or were delayed longer than 2 hours (this is the first figure shown). We will attempt to reproduce this figure.

```
SELECT o.carrier, c.name,
  SUM(1) AS numFlights,
  SUM(IF(arr_delay > 15 AND arr_delay <= 119, 1, 0)) AS shortDelay,
  SUM(
    IF(arr_delay >= 120 OR cancelled = 1 OR diverted = 1, 1, 0)
  ) AS longDelay
FROM
  flights AS o
LEFT JOIN
  carriers c ON o.carrier = c.carrier
WHERE year = 2014
GROUP BY carrier
ORDER BY shortDelay DESC
```

After pulling relevant data into R, we begin by pruning less informative labels from the carriers.

```
res <- res |>
  as_tibble() |>
  mutate(
    name = str_remove_all(name, "Air(lines|ways| Lines)"),
    name = str_remove_all(name, "(Inc\\.|Co\\.|Corporation)"),
    name = str_remove_all(name, "\\(.*\\)"),
    name = str_remove_all(name, " *$")
  )
res |>
  pull(name)
```

```
##  [1] "Southwest"      "ExpressJet"     "SkyWest"        "Delta"
##  [5] "American"       "United"         "Envoy Air"      "US"
##  [9] "JetBlue"        "Frontier"       "Alaska"         "AirTran"
## [13] "Virgin America" "Hawaiian"
```

Next, it is now clear that FiveThirtyEight has considered airline mergers and regional carriers that are not captured in our data. Specifically: "We classify all remaining AirTran flights as Southwest flights," and Envoy Air serves American Airlines. However, there is a bewildering network of alliances among the other regional carriers. Greatly complicating matters, ExpressJet and SkyWest serve multiple national carriers (primarily United, American, and Delta) under different flight numbers. FiveThirtyEight provides a footnote detailing how they have assigned flights carried by these regional carriers, but we have chosen to ignore that here and include ExpressJet and SkyWest as independent carriers. Thus, the data in our figure does not exactly match the figure from FiveThirtyEight.

```
carriers_2014 <- res |>
  mutate(
    groupName = case_when(
      name %in% c("Envoy Air", "American Eagle") ~ "American",
      name == "AirTran" ~ "Southwest",
      TRUE ~ name
    )
  ) |>
  group_by(groupName) |>
  summarize(
    numFlights = sum(numFlights),
    wShortDelay = sum(shortDelay),
    wLongDelay = sum(longDelay)
  ) |>
  mutate(
    wShortDelayPct = wShortDelay / numFlights,
    wLongDelayPct = wLongDelay / numFlights,
    delayed = wShortDelayPct + wLongDelayPct,
    ontime = 1 - delayed
  )
carriers_2014
```

```
## # A tibble: 12 x 8
##    groupName      numFlights wShortDelay wLongDelay wShortDelayPct wLongDelayPct
##    <chr>               <dbl>       <dbl>      <dbl>          <dbl>         <dbl>
##  1 Alaska             160257       18366       2613          0.115        0.0163
##  2 American           930398      191071      53641          0.205        0.0577
```

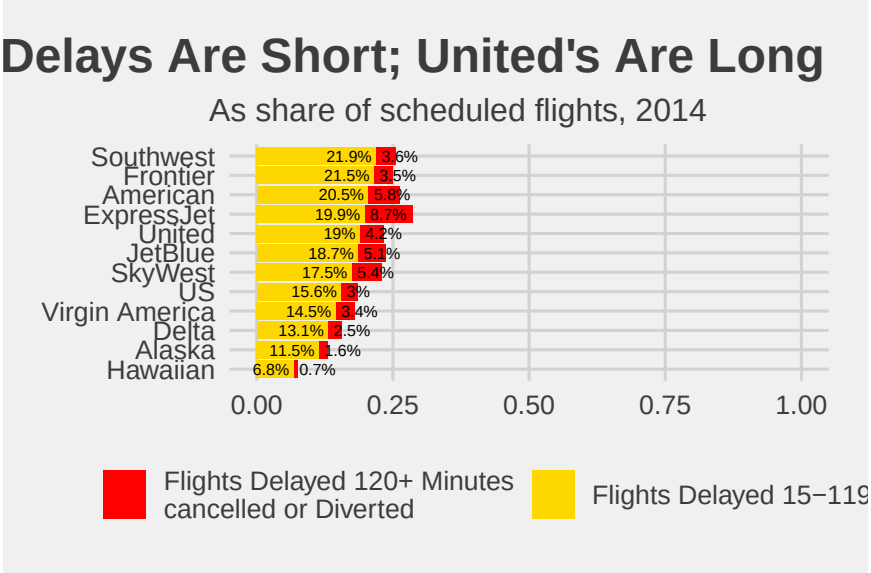
```
##  3 Delta              800375      105194      19818         0.131        0.0248
##  4 ExpressJet         686021      136207      59663         0.199        0.0870
##  5 Frontier            85474       18410       2959         0.215        0.0346
##  6 Hawaiian            74732        5098        514         0.0682       0.00688
##  7 JetBlue            249693       46618      12789         0.187        0.0512
##  8 SkyWest            613030      107192      33114         0.175        0.0540
##  9 Southwest         1254128      275155      44907         0.219        0.0358
## 10 US                 414665       64505      12328         0.156        0.0297
## 11 United             493528       93721      20923         0.190        0.0424
## 12 Virgin America      57510        8356       1976         0.145        0.0344
## # i 2 more variables: delayed <dbl>, ontime <dbl>
```

After tidying this data frame using the pivot_longer() function (see Chapter 6), we can draw the figure as a stacked bar chart.

```
carriers_tidy <- carriers_2014 |>
  select(groupName, wShortDelayPct, wLongDelayPct, delayed) |>
  pivot_longer(
    -c(groupName, delayed),
    names_to = "delay_type",
    values_to = "pct"
  )
delay_chart <- ggplot(
  data = carriers_tidy,
  aes(x = reorder(groupName, pct, max), y = pct)
) +
  geom_col(aes(fill = delay_type)) +
  scale_fill_manual(
    name = NULL,
    values = c("red", "gold"),
    labels = c(
      "Flights Delayed 120+ Minutes\ncancelled or Diverted",
      "Flights Delayed 15-119 Minutes"
    )
  ) +
  scale_y_continuous(limits = c(0, 1)) +
  coord_flip() +
  labs(
    title = "Southwest's Delays Are Short; United's Are Long",
    subtitle = "As share of scheduled flights, 2014"
  ) +
  ylab(NULL) +
  xlab(NULL) +
  ggthemes::theme_fivethirtyeight() +
  theme(
    plot.title = element_text(hjust = 1),
    plot.subtitle = element_text(hjust = -0.2)
  )
```

Getting the right text labels in the right places to mimic the display requires additional wrangling. In fact, by comparing our best attempt to the figure in FiveThirtyEight, it becomes clear that many of the long delays suffered by United and American passengers occur on flights operated by ExpressJet and SkyWest.

```
delay_chart +
  geom_text(
    data = filter(carriers_tidy, delay_type == "wShortDelayPct"),
    aes(label = paste0(round(pct * 100, 1), "% ")),
    hjust = "right",
    size = 2
  ) +
  geom_text(
    data = filter(carriers_tidy, delay_type == "wLongDelayPct"),
    aes(y = delayed - pct, label = paste0(round(pct * 100, 1), "% ")),
    hjust = "left",
    nudge_y = 0.01,
    size = 2
  )
```



[continued]