Lab 11: Databases

MinJae Jo

2025-11-21

Lab report

Exercise 1

```
library(DBI)
library(RSQLite)
library(dbplyr)

con <- dbConnect(SQLite(), "nycflights13.sqlite")
```

Exercise 2

```
library(dbplyr)
flights_tbl <- tbl(con, "flights")
```

Exercise 3

```
flights_query <- flights_tbl %>%
  select(year, month, day, hour, dep_delay, origin)
```

Exercise 4

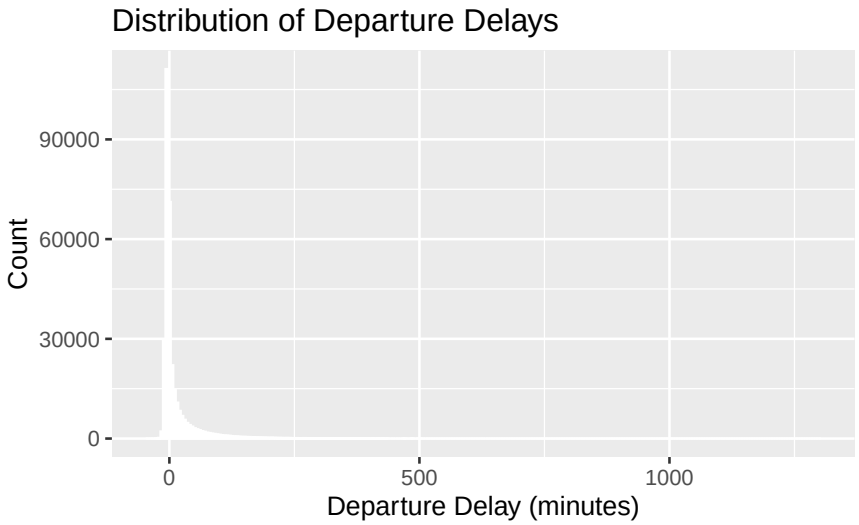
```
## <SQL>
## SELECT `year`, `month`, `day`, `hour`, `dep_delay`, `origin`
## FROM `flights`
```

Exercise 5

```r
flights_df <- flights_query %>%
  collect()
```

Exercise 6

```
## Warning: Removed 8255 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



- The average is much higher than the median, meaning there are many very long delays that drive the average up. While most flights don't experience significant delays, some have very long delays.

Exercise 7

i. Each row represents a weather observation record of a specific date and time at an airport in one of the JFK/LGA/EWRs.

ii. The precip column contains information related to rain or precipitation.

iii. The origin column indicates at which airport (JFK, LGA, EWR) the meteorological measurements were made.

Exercise 8

```
## <SQL>
## SELECT `origin`, `year`, `month`, `day`, `hour`, `temp`, `wind_speed`, `precip`
## FROM `weather`
```

Exercise 9

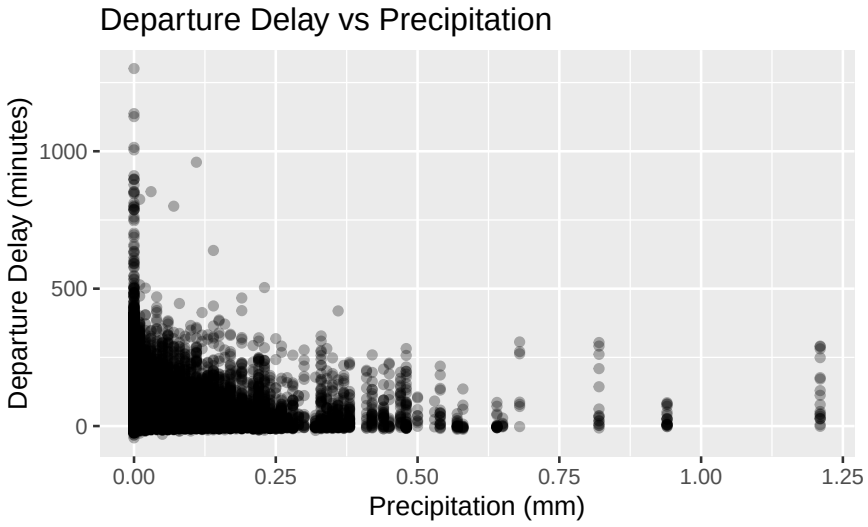
```
## <SQL>
## SELECT
##   `flights`.`year` AS `year`,
##   `flights`.`month` AS `month`,
##   `flights`.`day` AS `day`,
##   `flights`.`hour` AS `hour`,
##   `dep_delay`,
##   `flights`.`origin` AS `origin`,
##   `temp`,
##   `wind_speed`,
##   `precip`
## FROM `flights`
## LEFT JOIN `weather`
##   ON (
##     `flights`.`origin` = `weather`.`origin` AND
##     `flights`.`year` = `weather`.`year` AND
##     `flights`.`month` = `weather`.`month` AND
##     `flights`.`day` = `weather`.`day` AND
##     `flights`.`hour` = `weather`.`hour`
##   )
```

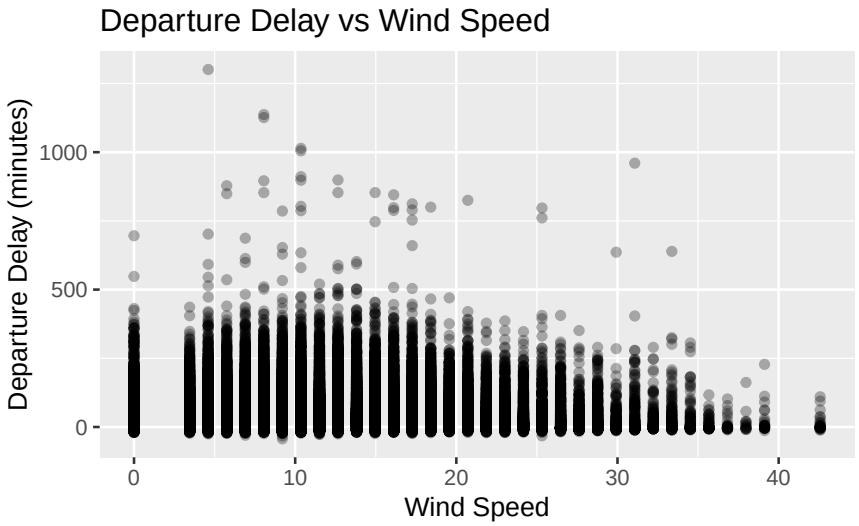
Exercise 10

- Each row shows one flight with weather conditions at sa airport.

```
joined_df <- joined_query %>% collect()
```

Exercise 11





-As it rains more, dep_delay tends to increase slightly. However, it is not a very strong correlation because the dots are spread out. Wind_speed seems to have little apparent relationship with delay. ### Exercise 12 - When I looked at the two scatterplots, the correlation between bad weather and delay in departure did not seem very strong. There were a few times when the delay was longer when it was raining heavily or windy, but most of the points were scattered without any obvious shape. Therefore, the overall pattern appeared weaker than expected. But on more thought, this dataset only includes flights that actually left. Flights canceled due to bad weather are not included in the data at all. This creates a kind of survival bias, as it only analyzes flights that "survived" from bad weather. Including canceled flights, the impact of bad weather on delays would seem greater. Therefore, the weak pattern in the graph may be due to the worst cases not appearing in the dataset at all.