PM 566 HW 02

AUTHOR
Ziquan 'Harrison' Liu

Packages

```
library(nycflights13)
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(lubridate)
```

```
Attaching package: 'lubridate'

The following objects are masked from 'package:base':

    date, intersect, setdiff, union
```

```
library(stringr)
library(maps)
```

Check on all description of dataset

```
summary(flights)
```

```
      year          month             day           dep_time    sched_dep_time
 Min.   :2013   Min.   : 1.000   Min.   : 1.00   Min.   :   1   Min.   : 106
 1st Qu.:2013   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.: 907   1st Qu.: 906
 Median :2013   Median : 7.000   Median :16.00   Median :1401   Median :1359
 Mean   :2013   Mean   : 6.549   Mean   :15.71   Mean   :1349   Mean   :1344
 3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:1744   3rd Qu.:1729
 Max.   :2013   Max.   :12.000   Max.   :31.00   Max.   :2400   Max.   :2359
```

```
                                          NA's   :8255
  dep_delay          arr_time    sched_arr_time   arr_delay
Min.   : -43.00   Min.   :   1   Min.   :   1   Min.   : -86.000
1st Qu.:  -5.00   1st Qu.:1104   1st Qu.:1124   1st Qu.: -17.000
Median :  -2.00   Median :1535   Median :1556   Median :  -5.000
Mean   :  12.64   Mean   :1502   Mean   :1536   Mean   :   6.895
3rd Qu.:  11.00   3rd Qu.:1940   3rd Qu.:1945   3rd Qu.:  14.000
Max.   :1301.00   Max.   :2400   Max.   :2359   Max.   :1272.000
NA's   :8255      NA's   :8713                  NA's   :9430
  carrier             flight       tailnum             origin
Length:336776      Min.   :   1   Length:336776      Length:336776
Class :character   1st Qu.: 553   Class :character   Class :character
Mode  :character   Median :1496   Mode  :character   Mode  :character
                   Mean   :1972
                   3rd Qu.:3465
                   Max.   :8500

    dest              air_time        distance         hour
Length:336776      Min.   : 20.0   Min.   :  17   Min.   : 1.00
Class :character   1st Qu.: 82.0   1st Qu.: 502   1st Qu.: 9.00
Mode  :character   Median :129.0   Median : 872   Median :13.00
                   Mean   :150.7   Mean   :1040   Mean   :13.18
                   3rd Qu.:192.0   3rd Qu.:1389   3rd Qu.:17.00
                   Max.   :695.0   Max.   :4983   Max.   :23.00
                   NA's   :9430
    minute        time_hour
Min.   : 0.00   Min.   :2013-01-01 05:00:00
1st Qu.: 8.00   1st Qu.:2013-04-04 13:00:00
Median :29.00   Median :2013-07-03 10:00:00
Mean   :26.23   Mean   :2013-07-03 05:22:54
3rd Qu.:44.00   3rd Qu.:2013-10-01 07:00:00
Max.   :59.00   Max.   :2013-12-31 23:00:00
```

```
summary(airlines)
```

```
  carrier             name
Length:16          Length:16
Class :character   Class :character
Mode  :character   Mode  :character
```

```
summary(airports)
```

```
    faa                name               lat             lon
Length:1458        Length:1458        Min.   :19.72   Min.   :-176.65
Class :character   Class :character   1st Qu.:34.26   1st Qu.:-119.19
Mode  :character   Mode  :character   Median :40.09   Median : -94.66
                                      Mean   :41.65   Mean   :-103.39
                                      3rd Qu.:45.07   3rd Qu.: -82.52
```

```
                                  Max.   :72.27   Max.   : 174.11
     alt               tz             dst               tzone
Min.   : -54.00   Min.   :-10.000   Length:1458        Length:1458
1st Qu.:  70.25   1st Qu.: -8.000   Class :character   Class :character
Median : 473.00   Median : -6.000   Mode  :character   Mode  :character
Mean   :1001.42   Mean   : -6.519
3rd Qu.:1062.50   3rd Qu.: -5.000
Max.   :9078.00   Max.   :  8.000
```

```
summary(planes)
```

```
  tailnum              year         type           manufacturer
Length:3322        Min.   :1956   Length:3322        Length:3322
Class :character   1st Qu.:1997   Class :character   Class :character
Mode  :character   Median :2001   Mode  :character   Mode  :character
                   Mean   :2000
                   3rd Qu.:2005
                   Max.   :2013
                   NA's   :70
   model              engines          seats            speed
Length:3322        Min.   :1.000   Min.   :  2.0   Min.   : 90.0
Class :character   1st Qu.:2.000   1st Qu.:140.0   1st Qu.:107.5
Mode  :character   Median :2.000   Median :149.0   Median :162.0
                   Mean   :1.995   Mean   :154.3   Mean   :236.8
                   3rd Qu.:2.000   3rd Qu.:182.0   3rd Qu.:432.0
                   Max.   :4.000   Max.   :450.0   Max.   :432.0
                                                   NA's   :3299
   engine
Length:3322
Class :character
Mode  :character
```

```
summary(weather)
```

```
   origin              year          month             day
Length:26115       Min.   :2013   Min.   : 1.000   Min.   : 1.00
Class :character   1st Qu.:2013   1st Qu.: 4.000   1st Qu.: 8.00
Mode  :character   Median :2013   Median : 7.000   Median :16.00
                   Mean   :2013   Mean   : 6.504   Mean   :15.68
                   3rd Qu.:2013   3rd Qu.: 9.000   3rd Qu.:23.00
                   Max.   :2013   Max.   :12.000   Max.   :31.00

     hour            temp            dewp            humid
Min.   : 0.00   Min.   : 10.94   Min.   :-9.94   Min.   : 12.74
1st Qu.: 6.00   1st Qu.: 39.92   1st Qu.:26.06   1st Qu.: 47.05
```

```
 Median :11.00   Median : 55.40   Median :42.08   Median : 61.79
 Mean   :11.49   Mean   : 55.26   Mean   :41.44   Mean   : 62.53
 3rd Qu.:17.00   3rd Qu.: 69.98   3rd Qu.:57.92   3rd Qu.: 78.79
 Max.   :23.00   Max.   :100.04   Max.   :78.08   Max.   :100.00
                 NA's   :1        NA's   :1       NA's   :1
    wind_dir       wind_speed         wind_gust         precip
 Min.   :  0.0   Min.   :   0.000   Min.   :16.11   Min.   :0.000000
 1st Qu.:120.0   1st Qu.:   6.905   1st Qu.:20.71   1st Qu.:0.000000
 Median :220.0   Median :  10.357   Median :24.17   Median :0.000000
 Mean   :199.8   Mean   :  10.517   Mean   :25.49   Mean   :0.004469
 3rd Qu.:290.0   3rd Qu.:  13.809   3rd Qu.:28.77   3rd Qu.:0.000000
 Max.   :360.0   Max.   :1048.361   Max.   :66.75   Max.   :1.210000
 NA's   :460     NA's   :4          NA's   :20778
    pressure          visib          time_hour
 Min.   : 983.8   Min.   : 0.000   Min.   :2013-01-01 01:00:00
 1st Qu.:1012.9   1st Qu.:10.000   1st Qu.:2013-04-01 21:30:00
 Median :1017.6   Median :10.000   Median :2013-07-01 14:00:00
 Mean   :1017.9   Mean   : 9.255   Mean   :2013-07-01 18:26:37
 3rd Qu.:1023.0   3rd Qu.:10.000   3rd Qu.:2013-09-30 13:00:00
 Max.   :1042.1   Max.   :10.000   Max.   :2013-12-30 18:00:00
 NA's   :2729
```

standardize time

```
# helper: convert HHMM integer time (e.g., 517) to hour-of-day on [0,24)
to_hour <- function(x) ifelse(is.na(x), NA_real_, (x %/% 100) %% 24 + (x %% 100)/60)

# helper: map hour to part-of-day
part_of_day <- function(hour) {
  cut(hour,
      breaks = c(0, 6, 12, 18, 24),
      labels = c("early morning", "morning", "afternoon", "evening"),
      right = FALSE, include.lowest = TRUE)
}
```

Question 1

```
top10_dest <- flights %>%
  count(dest, sort = TRUE, name = "n_flights") %>%
  slice_head(n = 10)
top10_dest
```

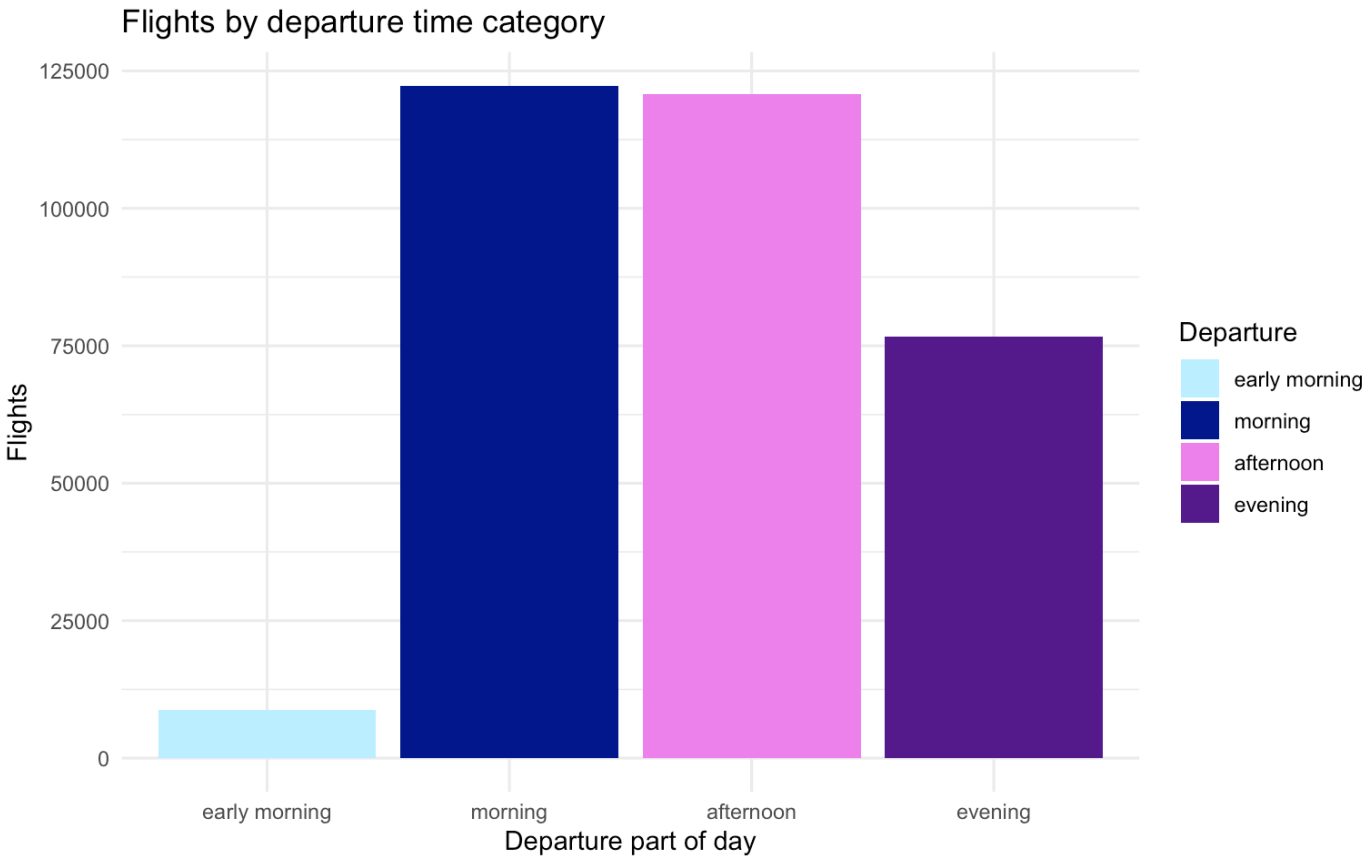
```
# A tibble: 10 × 2
   dest  n_flights
   <chr>     <int>
 1 ORD       17283
 2 ATL       17215
```

```
 3 LAX       16174
 4 BOS       15508
 5 MCO       14082
 6 CLT       14064
 7 SFO       13331
 8 FLL       12055
 9 MIA       11728
10 DCA        9705
```

Question 2

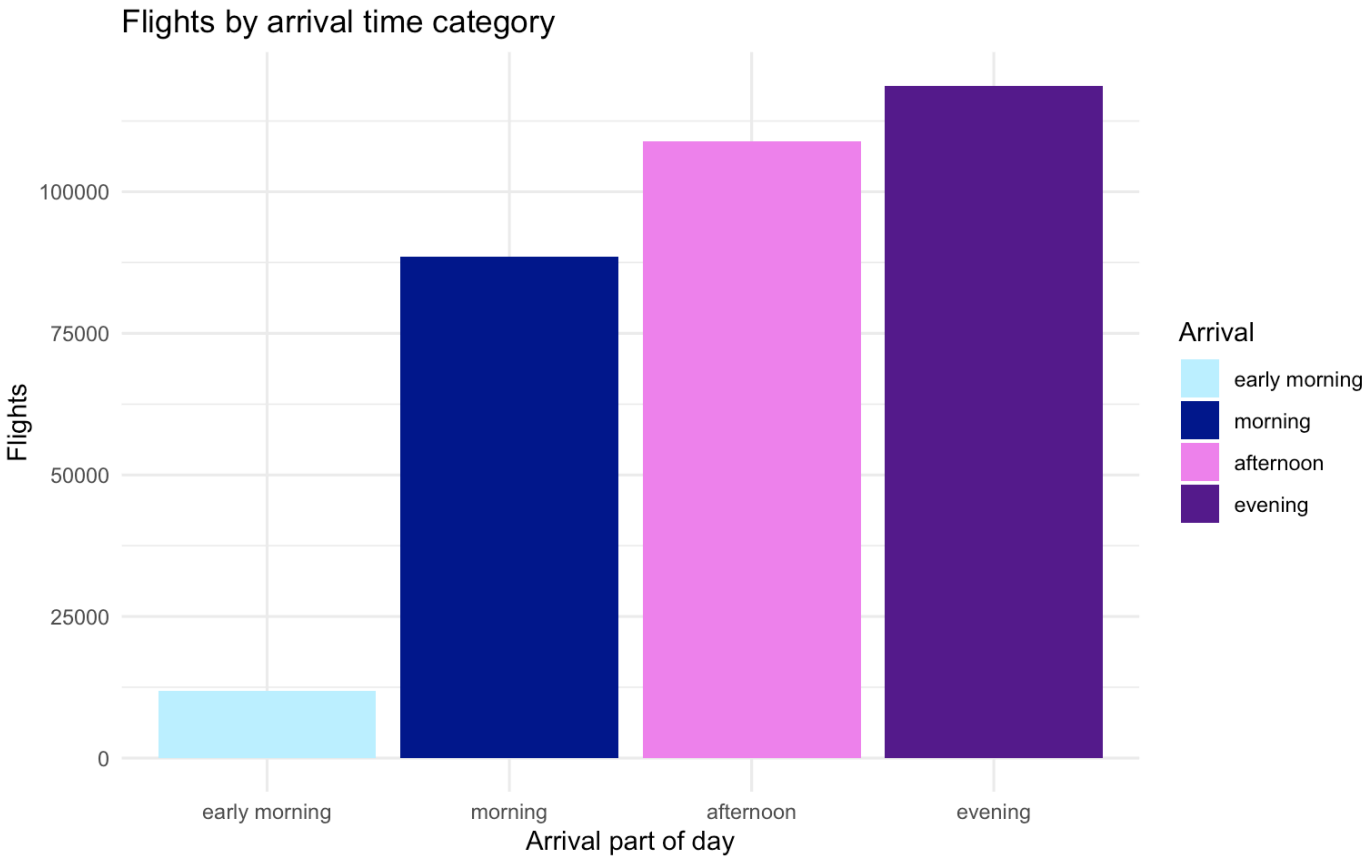
```
flights2 <- flights %>%
  mutate(
    dep_hour = to_hour(dep_time),
    arr_hour = to_hour(arr_time),
    dep_part = part_of_day(dep_hour),
    arr_part = part_of_day(arr_hour)
  )

# barplots
## select corlor
pal <- c(
  "early morning" = "lightblue1",
  "morning"       = "darkblue",
  "afternoon"     = "violet",
  "evening"       = "purple4"
)

ggplot(flights2, aes(x = dep_part, fill = dep_part)) +
  geom_bar() +
  scale_x_discrete(na.translate = FALSE) +         # mute NA category
  scale_fill_manual(values = pal, na.translate = FALSE) +
  labs(x = "Departure part of day", y = "Flights",
       title = "Flights by departure time category") +
  guides(fill = guide_legend(title = "Departure")) +
  theme_minimal()
```

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




```
ggplot(flights2, aes(x = arr_part, fill = arr_part)) +
  geom_bar() +
  scale_x_discrete(na.translate = FALSE) +
  scale_fill_manual(values = pal, na.translate = FALSE) +
  labs(x = "Arrival part of day", y = "Flights",
       title = "Flights by arrival time category") +
  guides(fill = guide_legend(title = "Arrival")) +
  theme_minimal()
```

```
Warning: Removed 8713 rows containing non-finite outside the scale range
(`stat_count()`).
```




```
# % red-eye = depart in afternoon/evening AND arrive in early morning/morning
valid <- flights2 %>% filter(!is.na(dep_part), !is.na(arr_part))
red_eye <- valid %>%
  mutate(is_redeye = dep_part %in% c("afternoon","evening") &
                     arr_part %in% c("early morning","morning")) %>%
  summarise(
    n = n(),
    n_redeye = sum(is_redeye),
    pct_redeye = 100 * n_redeye / n
  )

red_eye
```

```
# A tibble: 1 × 3
      n n_redeye pct_redeye
  <int>    <int>      <dbl>
1 328063    10754       3.28
```

Question 3

```r
# unique tailnum–carrier pairs with airline names
tail_carriers <- flights %>%
  filter(!is.na(tailnum), tailnum != "", !is.na(carrier)) %>%
  distinct(tailnum, carrier) %>%
  left_join(airlines, by = "carrier")

# count distinct carriers per plane, keep those with >1
multi_airline_planes <- tail_carriers %>%
  group_by(tailnum) %>%
  summarise(
    n_airlines = n_distinct(carrier),
    airlines = paste(sort(unique(name)), collapse = ", "),
    .groups = "drop"
  ) %>%
  filter(n_airlines > 1) %>%
  arrange(desc(n_airlines), tailnum)

# how many such planes?
n_multi_planes <- nrow(multi_airline_planes)

n_multi_planes
```

```
[1] 17
```

```r
multi_airline_planes
```

```
# A tibble: 17 × 3
   tailnum n_airlines airlines
   <chr>        <int> <chr>
 1 N146PQ           2 Endeavor Air Inc., ExpressJet Airlines Inc.
 2 N153PQ           2 Endeavor Air Inc., ExpressJet Airlines Inc.
 3 N176PQ           2 Endeavor Air Inc., ExpressJet Airlines Inc.
 4 N181PQ           2 Endeavor Air Inc., ExpressJet Airlines Inc.
 5 N197PQ           2 Endeavor Air Inc., ExpressJet Airlines Inc.
 6 N200PQ           2 Endeavor Air Inc., ExpressJet Airlines Inc.
 7 N228PQ           2 Endeavor Air Inc., ExpressJet Airlines Inc.
 8 N232PQ           2 Endeavor Air Inc., ExpressJet Airlines Inc.
 9 N933AT           2 AirTran Airways Corporation, Delta Air Lines Inc.
10 N935AT           2 AirTran Airways Corporation, Delta Air Lines Inc.
11 N977AT           2 AirTran Airways Corporation, Delta Air Lines Inc.
12 N978AT           2 AirTran Airways Corporation, Delta Air Lines Inc.
13 N979AT           2 AirTran Airways Corporation, Delta Air Lines Inc.
14 N981AT           2 AirTran Airways Corporation, Delta Air Lines Inc.
15 N989AT           2 AirTran Airways Corporation, Delta Air Lines Inc.
16 N990AT           2 AirTran Airways Corporation, Delta Air Lines Inc.
17 N994AT           2 AirTran Airways Corporation, Delta Air Lines Inc.
```

Question 4

```
# The missing key is: weather$origin  <->  airports$faa
# Example join to attach airport metadata to each weather row:
weather_with_airport <- weather %>%
  left_join(airports %>% select(faa, name, lat, lon, tz), by = c("origin" = "faa"))

# peek to confirm the relationship
weather_with_airport %>% select(origin, name, time_hour) %>% slice_head(n = 5)
```

```
# A tibble: 5 × 3
  origin name                time_hour
  <chr>  <chr>               <dttm>
1 EWR    Newark Liberty Intl 2013-01-01 01:00:00
2 EWR    Newark Liberty Intl 2013-01-01 02:00:00
3 EWR    Newark Liberty Intl 2013-01-01 03:00:00
4 EWR    Newark Liberty Intl 2013-01-01 04:00:00
5 EWR    Newark Liberty Intl 2013-01-01 05:00:00
```

Question 5

```
# 5a. Make an hourly key in `weather` and count duplicate keys
weather_keyed <- weather %>%
  mutate(
    hw_key = str_c(year, month, day, hour, origin, sep = "-")
  )

dup_count <- sum(duplicated(weather_keyed$hw_key))
dup_breakdown <- weather_keyed %>%
  count(year, month, day, hour, origin, name = "n") %>%
  arrange(desc(n)) %>%
  filter(n > 1)

dup_count
```

```
[1] 3
```

```
head(dup_breakdown)
```

```
# A tibble: 3 × 6
   year month   day  hour origin     n
  <int> <int> <int> <int> <chr>  <int>
1  2013    11     3     1 EWR        2
```

```
2  2013    11     3     1 JFK        2
3  2013    11     3     1 LGA        2
```

```
# Interpretation:
# The combination (year, month, day, hour, origin) is *usually* unique,
# but duplicates occur because multiple measurements can be recorded
# within the same hour at an airport (e.g., corrections/updates), so we
# occasionally get >1 row per hour per origin.
```

```
# 5b. Merge weather onto each flight by scheduled departure hour & origin
fl_wx <- flights %>%
  select(year, month, day, dep_time, sched_dep_time, dep_delay, arr_delay,
         origin, dest, time_hour, flight, carrier, tailnum) %>%
  left_join(weather %>% select(origin, time_hour, temp, dewp, humid, wind_dir,
                               wind_speed, wind_gust, precip, pressure, visib),
            by = c("origin","time_hour"))

dim(fl_wx); fl_wx %>% glimpse()
```

```
[1] 336776     22

Rows: 336,776
Columns: 22
$ year           <int> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2…
$ month          <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
$ day            <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
$ dep_time       <int> 517, 533, 542, 544, 554, 554, 555, 557, 557, 558, 558, …
$ sched_dep_time <int> 515, 529, 540, 545, 600, 558, 600, 600, 600, 600, 600, …
$ dep_delay      <dbl> 2, 4, 2, -1, -6, -4, -5, -3, -3, -2, -2, -2, -2, -2, -1…
$ arr_delay      <dbl> 11, 20, 33, -18, -25, 12, 19, -14, -8, 8, -2, -3, 7, -1…
$ origin         <chr> "EWR", "LGA", "JFK", "JFK", "LGA", "EWR", "EWR", "LGA",…
$ dest           <chr> "IAH", "IAH", "MIA", "BQN", "ATL", "ORD", "FLL", "IAD",…
$ time_hour      <dttm> 2013-01-01 05:00:00, 2013-01-01 05:00:00, 2013-01-01 0…
$ flight         <int> 1545, 1714, 1141, 725, 461, 1696, 507, 5708, 79, 301, 4…
$ carrier        <chr> "UA", "UA", "AA", "B6", "DL", "UA", "B6", "EV", "B6", "…
$ tailnum        <chr> "N14228", "N24211", "N619AA", "N804JB", "N668DN", "N394…
$ temp           <dbl> 39.02, 39.92, 39.02, 39.02, 39.92, 39.02, 37.94, 39.92,…
$ dewp           <dbl> 28.04, 24.98, 26.96, 26.96, 24.98, 28.04, 28.04, 24.98,…
$ humid          <dbl> 64.43, 54.81, 61.63, 61.63, 54.81, 64.43, 67.21, 54.81,…
$ wind_dir       <dbl> 260, 250, 260, 260, 260, 260, 240, 260, 260, 260, 260, …
$ wind_speed     <dbl> 12.65858, 14.96014, 14.96014, 14.96014, 16.11092, 12.65…
$ wind_gust      <dbl> NA, 21.86482, NA, NA, 23.01560, NA, NA, 23.01560, NA, 2…
$ precip         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
$ pressure       <dbl> 1011.9, 1011.4, 1012.1, 1012.1, 1011.7, 1011.9, 1012.4,…
$ visib          <dbl> 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10,…
```

```
# Each flight now carries the *departure-hour* weather at its origin.
```

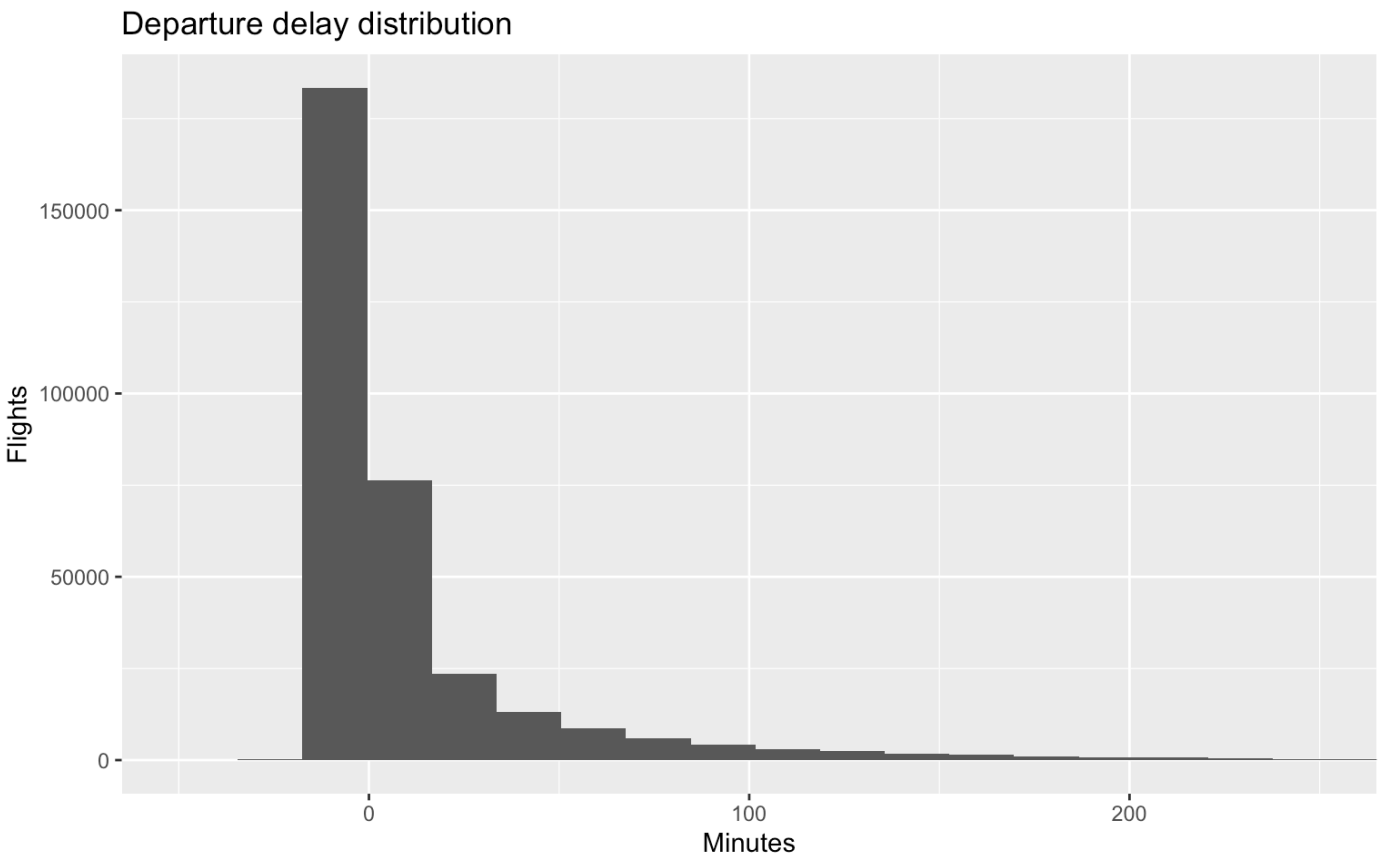
Question 6

```
# Construct a few weather "flags" helpful for EDA
fl_wx2 <- fl_wx %>%
  mutate(
    rain     = precip > 0,
    heavy_r  = precip >= 0.25,        # adjustable threshold (inches)
    low_vis  = visib < 3,             # < 3 miles visibility
    hi_wind  = wind_speed >= 20,      # sustained high wind
    gusty    = !is.na(wind_gust) & wind_gust >= 30,
    cold     = temp <= 32,
    hot      = temp >= 90
  )

# Step 2: What's missing & ranges?
fl_wx2 %>%
  summarise(across(c(dep_delay, precip, visib, wind_speed, wind_gust,
                     temp, humid, pressure),
                   list(miss = ~sum(is.na(.)),
                        min  = ~min(., na.rm = TRUE),
                        p50  = ~median(., na.rm = TRUE),
                        max  = ~max(., na.rm = TRUE)))) %>%
  tidyr::pivot_longer(everything())
```

```
# A tibble: 32 × 2
   name             value
   <chr>            <dbl>
 1 dep_delay_miss 8255
 2 dep_delay_min   -43
 3 dep_delay_p50    -2
 4 dep_delay_max  1301
 5 precip_miss    1556
 6 precip_min        0
 7 precip_p50        0
 8 precip_max        1.21
 9 visib_miss     1556
10 visib_min         0
# ℹ 22 more rows
```

```
# Step 3: Single-variable distributions (delays)
ggplot(filter(fl_wx2, !is.na(dep_delay)), aes(x = dep_delay)) +
  geom_histogram(bins = 80) +
```

```
  coord_cartesian(xlim = c(-50, 250)) +
  labs(title = "Departure delay distribution", x = "Minutes", y = "Flights")
```

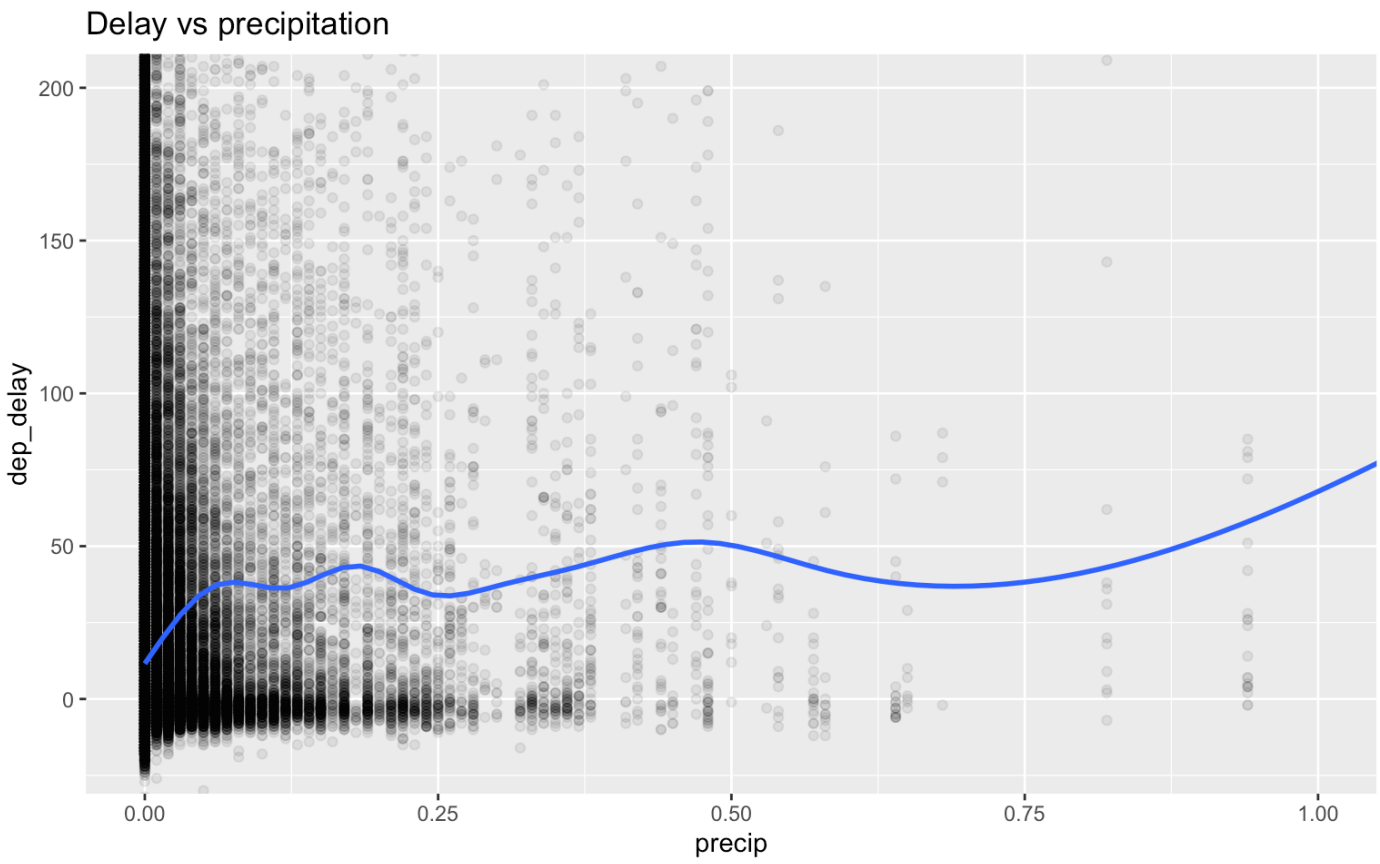



```
# Step 4: Bivariate—delays vs. key weather features (binning for robustness)
# Mean delay by weather flags
by_flags <- fl_wx2 %>%
  filter(!is.na(dep_delay)) %>%
  summarise(
    n         = n(),
    mean_delay = mean(dep_delay, na.rm = TRUE),
    rain      = mean(dep_delay[rain],    na.rm = TRUE),
    norain    = mean(dep_delay[!rain],   na.rm = TRUE),
    heavy_r   = mean(dep_delay[heavy_r], na.rm = TRUE),
    low_vis   = mean(dep_delay[low_vis], na.rm = TRUE),
    hi_wind   = mean(dep_delay[hi_wind], na.rm = TRUE),
    gusty     = mean(dep_delay[gusty],   na.rm = TRUE),
    cold      = mean(dep_delay[cold],    na.rm = TRUE),
    hot       = mean(dep_delay[hot],     na.rm = TRUE)
  )
by_flags
```

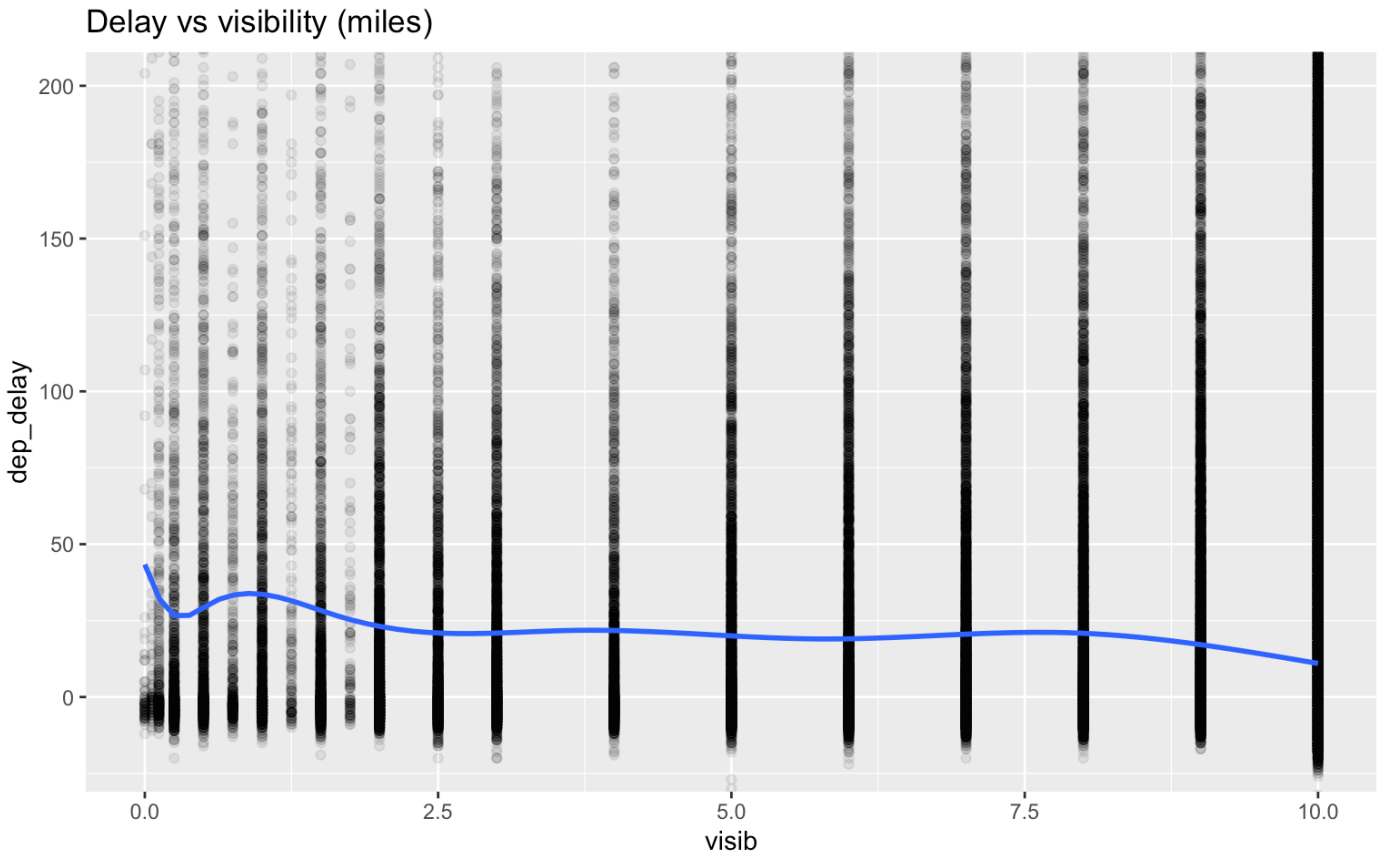
```
# A tibble: 1 × 10
       n mean_delay  rain norain heavy_r low_vis hi_wind gusty  cold   hot
   <int>      <dbl> <dbl>  <dbl>   <dbl>   <dbl>   <dbl> <dbl> <dbl> <dbl>
1 328521       12.6  30.9    NaN    42.0    26.7    16.9  16.4  11.4  18.8
```

```r
# Continuous relationships with smoothing
p1 <- ggplot(fl_wx2, aes(precip, dep_delay)) +
  geom_point(alpha = 0.08) + geom_smooth(se = FALSE) +
  coord_cartesian(xlim = c(0, 1.0), ylim = c(-20, 200)) +
  labs(title = "Delay vs precipitation")

p2 <- ggplot(fl_wx2, aes(visib, dep_delay)) +
  geom_point(alpha = 0.08) + geom_smooth(se = FALSE) +
  coord_cartesian(xlim = c(0, 10), ylim = c(-20, 200)) +
  labs(title = "Delay vs visibility (miles)")

p3 <- ggplot(fl_wx2, aes(wind_speed, dep_delay)) +
  geom_point(alpha = 0.08) + geom_smooth(se = FALSE) +
  coord_cartesian(xlim = c(0, 40), ylim = c(-20, 200)) +
  labs(title = "Delay vs sustained wind (mph)")

p1; p2; p3
```

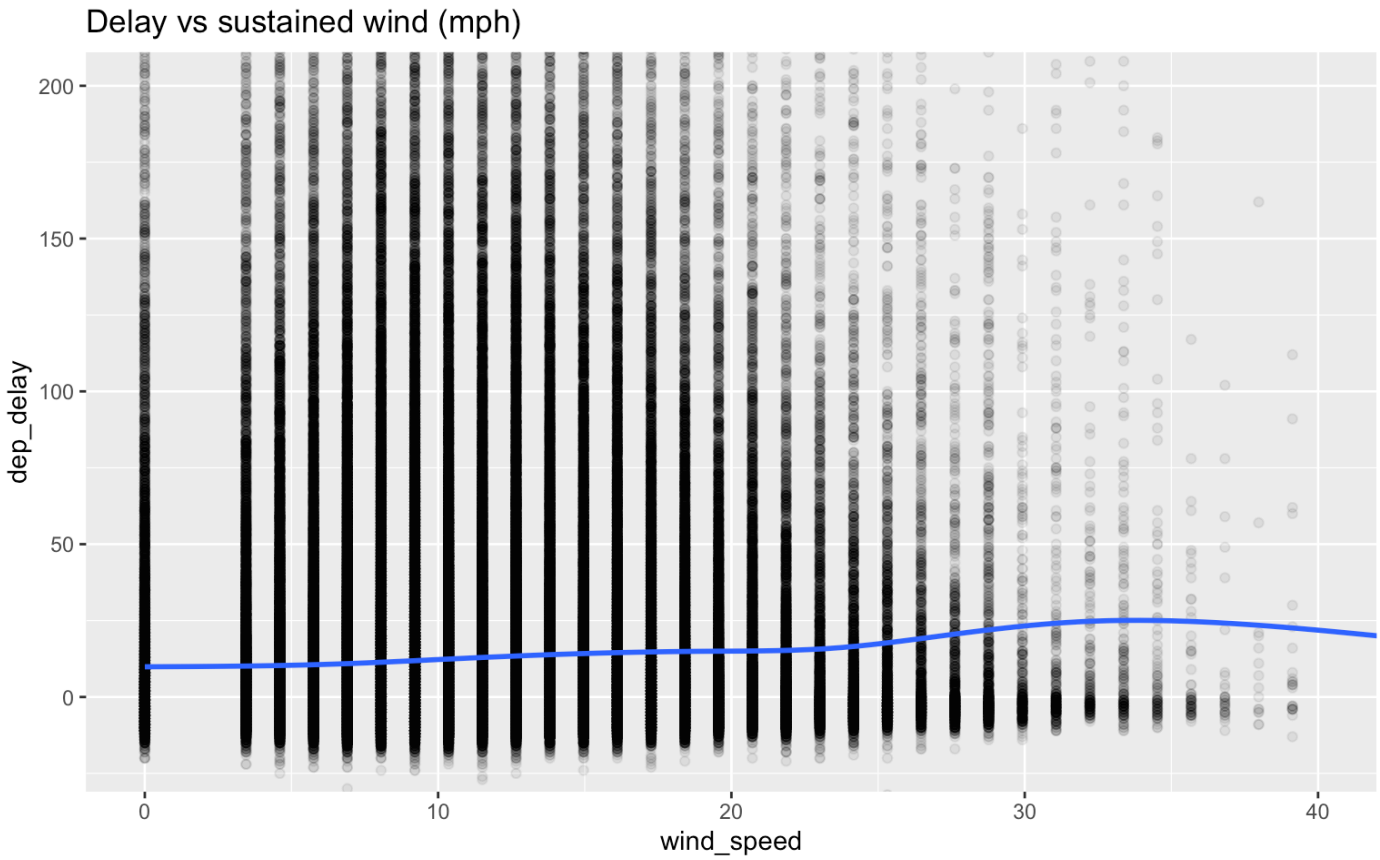
```
`geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'

Warning: Removed 9783 rows containing non-finite outside the scale range
(`stat_smooth()`).

Warning: Removed 9783 rows containing missing values or values outside the scale range
(`geom_point()`).
```



```
`geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'

Warning: Removed 9783 rows containing non-finite outside the scale range
(`stat_smooth()`).
Removed 9783 rows containing missing values or values outside the scale range
(`geom_point()`).
```




```
`geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'

Warning: Removed 9861 rows containing non-finite outside the scale range
(`stat_smooth()`).

Warning: Removed 9861 rows containing missing values or values outside the scale range
(`geom_point()`).
```




```
# Step 5: A quick multivariable check (not causal, just descriptive)
fit_eda <- lm(dep_delay ~ precip + visib + wind_speed + wind_gust +
                humid + temp + pressure, data = fl_wx2)
summary(fit_eda)
```

```
Call:
lm(formula = dep_delay ~ precip + visib + wind_speed + wind_gust +
    humid + temp + pressure, data = fl_wx2)

Residuals:
   Min     1Q Median     3Q    Max
-57.74 -17.19 -11.71  -0.06 753.86

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 220.527462  22.540984   9.783  < 2e-16 ***
precip       -4.661303  11.664874  -0.400   0.6895
visib        -1.827980   0.155435 -11.760  < 2e-16 ***
wind_speed   -0.109122   0.058710  -1.859   0.0631 .
wind_gust     0.326509   0.051161   6.382 1.76e-10 ***
humid         0.212380   0.009987  21.265  < 2e-16 ***
temp          0.121399   0.008247  14.720  < 2e-16 ***
pressure     -0.209755   0.021637  -9.694  < 2e-16 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 37.83 on 73225 degrees of freedom
  (263543 observations deleted due to missingness)
Multiple R-squared:  0.02382,   Adjusted R-squared:  0.02373
F-statistic: 255.2 on 7 and 73225 DF,  p-value: < 2.2e-16
```

Question 7

```
# Helper to keep only flights with a reported dep_delay
fw <- fl_wx2 %>% filter(!is.na(dep_delay))

# 7a. Average departure delay by *day*
daily <- fw %>%
  group_by(year, month, day) %>%
  summarise(avg_dep_delay = mean(dep_delay), n = n(), .groups = "drop") %>%
  arrange(desc(avg_dep_delay))
head(daily, 1)    # worst day
```

```
# A tibble: 1 × 5
   year month   day avg_dep_delay     n
  <int> <int> <int>         <dbl> <int>
1  2013     3     8          83.5   799
```

```
# 7b. By day × origin
daily_org <- fw %>%
  group_by(origin, year, month, day) %>%
  summarise(avg_dep_delay = mean(dep_delay), n = n(), .groups = "drop") %>%
  arrange(desc(avg_dep_delay))
head(daily_org, 1) # worst airport-day
```

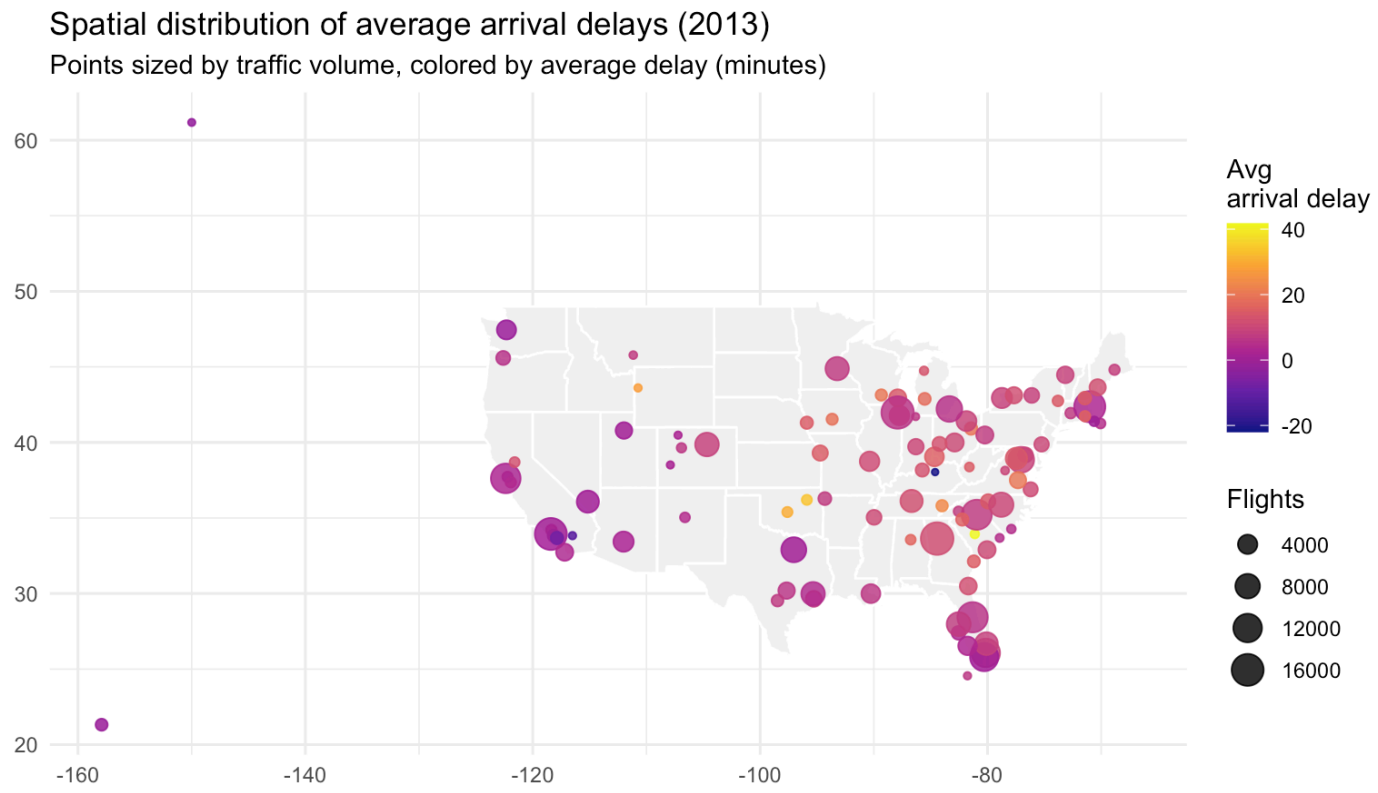
```
# A tibble: 1 × 6
  origin  year month   day avg_dep_delay     n
  <chr>  <int> <int> <int>         <dbl> <int>
1 LGA     2013     3     8          106.   229
```

```
# 7c. By hour × origin
hourly_org <- fw %>%
  mutate(hour = hour(time_hour)) %>%
  group_by(origin, year, month, day, hour) %>%
  summarise(avg_dep_delay = mean(dep_delay), n = n(), .groups = "drop") %>%
  arrange(desc(avg_dep_delay))
head(hourly_org, 1) # worst airport-hour
```

```
# A tibble: 1 × 7
  origin  year month   day  hour avg_dep_delay     n
  <chr>  <int> <int> <int> <int>         <dbl> <int>
1 LGA     2013     7    28    21          280.     3
```

Question 8

```
# Average arrival delay by destination airport (dest)
dest_avgs <- flights %>%
  filter(!is.na(arr_delay)) %>%
  group_by(dest) %>%
  summarise(avg_arr_delay = mean(arr_delay), n = n(), .groups = "drop")

airports_delay <- airports %>%
  inner_join(dest_avgs, by = c("faa" = "dest"))

usa <- map_data("state")

ggplot() +
  geom_polygon(data = usa, aes(long, lat, group = group),
               fill = "grey95", color = "white") +
  geom_point(data = airports_delay,
             aes(lon, lat, color = avg_arr_delay, size = n),
             alpha = 0.85) +
  scale_color_viridis_c(option = "plasma", name = "Avg\narrival delay") +
  scale_size_continuous(range = c(1, 6), name = "Flights") +
  coord_quickmap() +
  labs(title = "Spatial distribution of average arrival delays (2013)",
       subtitle = "Points sized by traffic volume, colored by average delay (minutes)",
       x = NULL, y = NULL) +
  theme_minimal()
```



Question 9

```
# Bin continuous weather into quintiles and compare mean delays between extremes
bin_compare <- function(x, y = fl_wx2$dep_delay, k = 5) {
  q <- quantile(x, probs = seq(0, 1, length.out = k + 1), na.rm = TRUE)
  g <- cut(x, breaks = unique(q), include.lowest = TRUE)
  tibble(g, y) %>%
    group_by(g) %>%
    summarise(mean_delay = mean(y, na.rm = TRUE), .groups = "drop") %>%
    mutate(bin = row_number())
}

res_precip <- bin_compare(fl_wx2$precip)
res_visib  <- bin_compare(fl_wx2$visib)
res_wspd   <- bin_compare(fl_wx2$wind_speed)
res_wgst   <- bin_compare(fl_wx2$wind_gust)

# Rank variables by (top-bin mean − bottom-bin mean)
impact_rank <- tibble(
  variable = c("precip", "visibility", "wind_speed", "wind_gust"),
  diff     = c(diff(range(res_precip$mean_delay, na.rm = TRUE)),
```

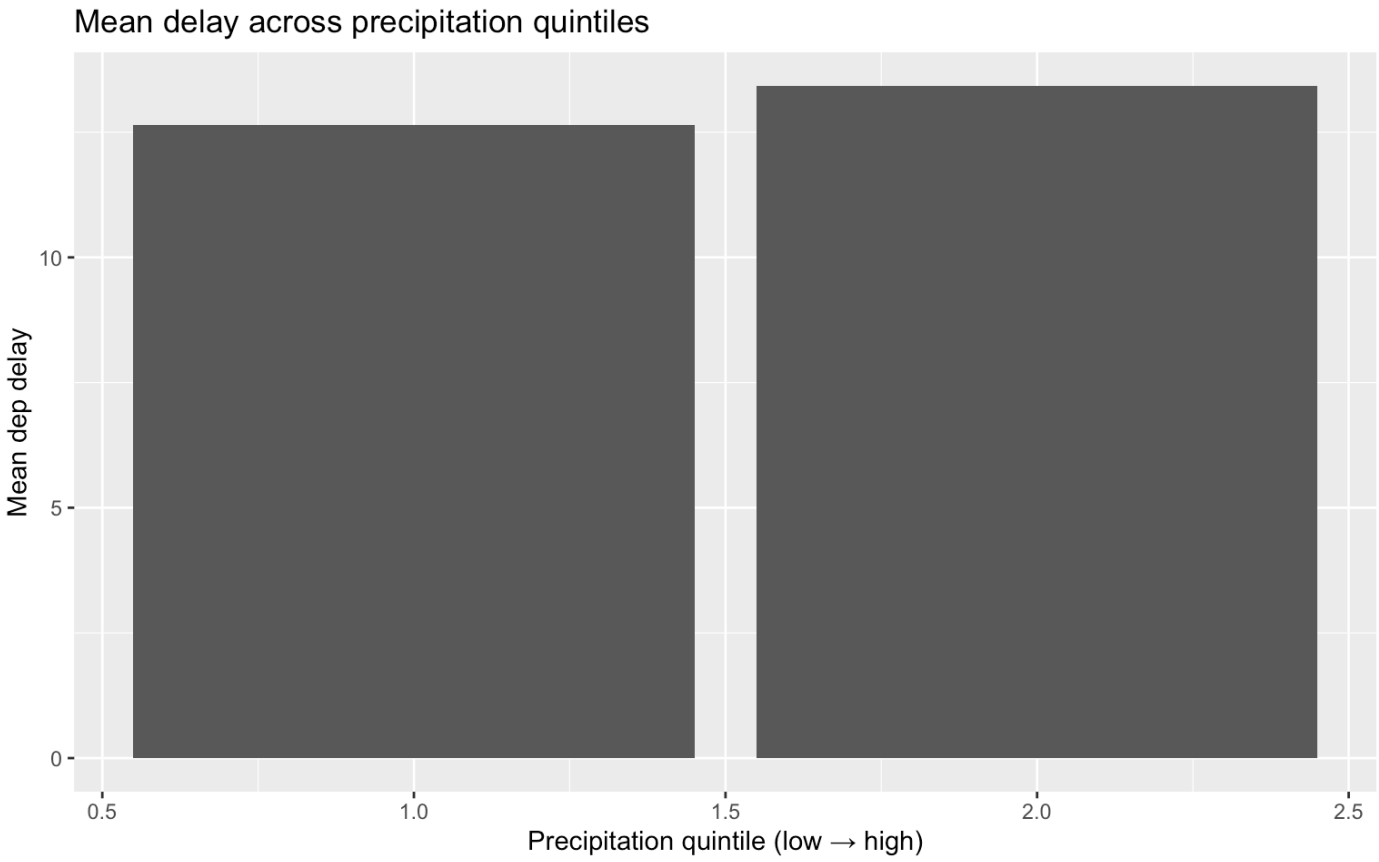
```
                diff(range(res_visib$mean_delay,  na.rm = TRUE)),
                diff(range(res_wspd$mean_delay,   na.rm = TRUE)),
                diff(range(res_wgst$mean_delay,   na.rm = TRUE)))
) %>% arrange(desc(diff))

impact_rank
```

```
# A tibble: 4 × 2
  variable    diff
  <chr>      <dbl>
1 wind_speed 5.25
2 wind_gust  4.29
3 precip     0.791
4 visibility 0.791
```

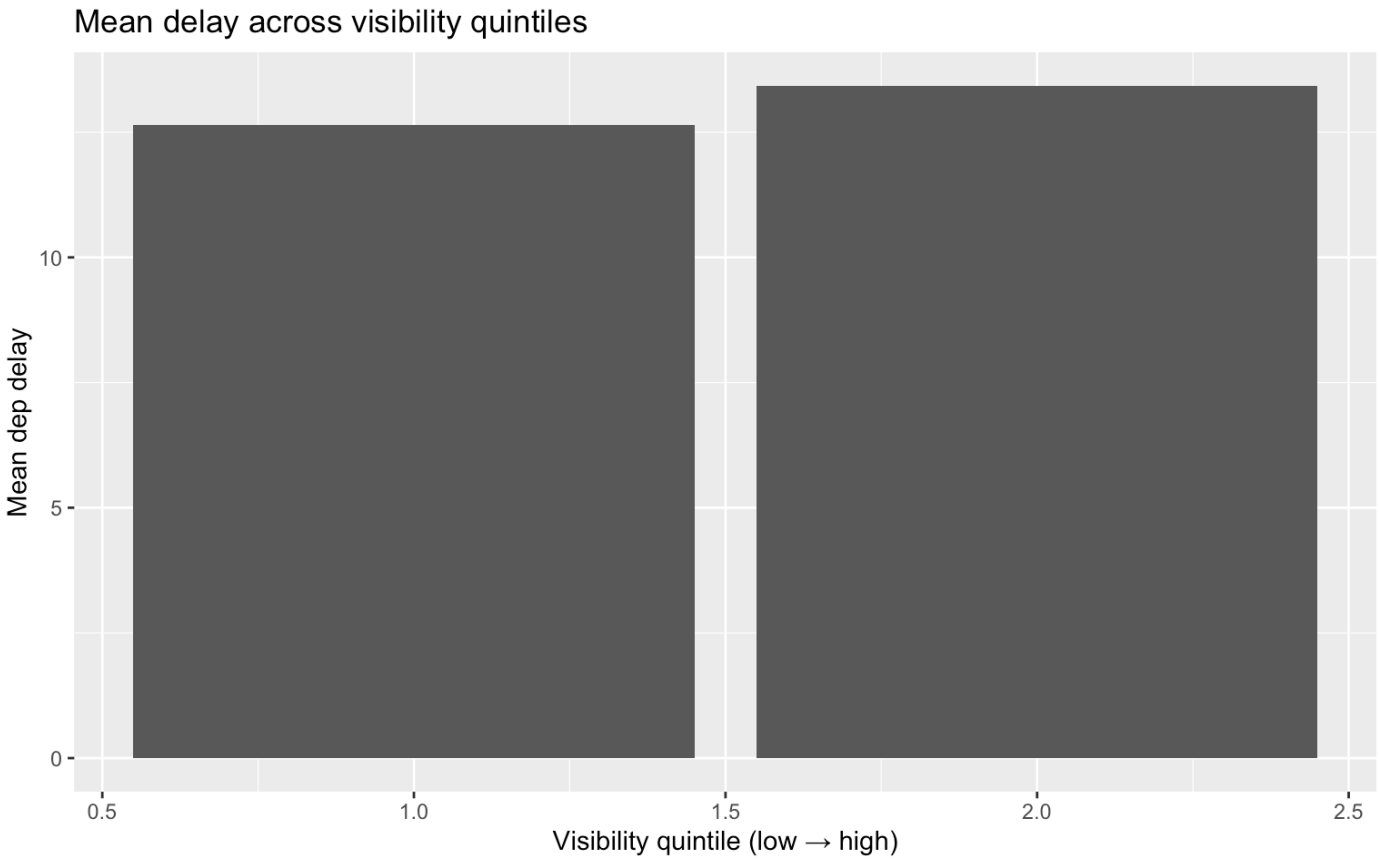
```
# Clear, compact comparisons with flag variables
flag_summary <- fl_wx2 %>%
  filter(!is.na(dep_delay)) %>%
  tidyr::pivot_longer(c(rain, heavy_r, low_vis, hi_wind, gusty, cold, hot),
                      names_to = "condition", values_to = "on") %>%
  group_by(condition, on) %>%
  summarise(mean_delay = mean(dep_delay), n = n(), .groups = "drop") %>%
  tidyr::pivot_wider(names_from = on, values_from = c(mean_delay, n), names_prefix = "on_
  mutate(delta = mean_delay_on_TRUE - mean_delay_on_FALSE) %>%
  arrange(desc(delta))

flag_summary
```

```
# A tibble: 7 × 8
  condition mean_delay_on_FALSE mean_delay_on_TRUE mean_delay_on_NA n_on_FALSE
  <chr>                   <dbl>              <dbl>            <dbl>      <int>
1 heavy_r                  12.5               42.0             13.4     325992
2 rain                     11.4               30.9             13.4     305907
3 low_vis                  12.1               26.7             13.4     314978
4 hot                      12.5               18.8             13.5     321757
5 hi_wind                  12.4               16.9             13.0     306927
6 gusty                    12.5               16.4             NA       315958
7 cold                     12.8               11.4             13.5     297446
# ℹ 3 more variables: n_on_TRUE <int>, n_on_NA <int>, delta <dbl>
```

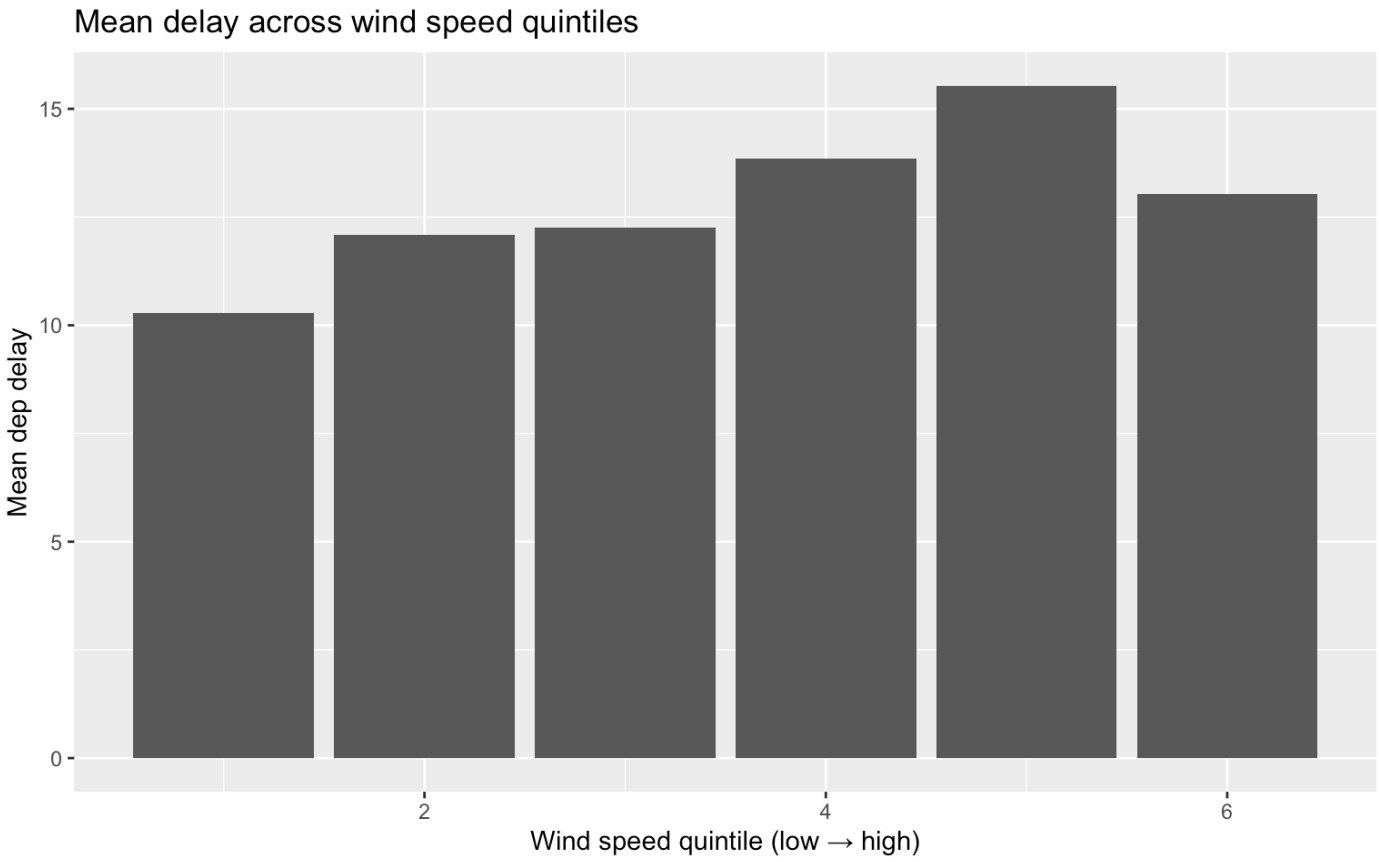
```
# Visual: mean delay by weather bins
ggplot(res_precip, aes(bin, mean_delay)) +
  geom_col() + labs(title = "Mean delay across precipitation quintiles",
                    x = "Precipitation quintile (low → high)", y = "Mean dep delay")
```




```
ggplot(res_visib, aes(bin, mean_delay)) +
  geom_col() + labs(title = "Mean delay across visibility quintiles",
                    x = "Visibility quintile (low → high)", y = "Mean dep delay")
```




```
ggplot(res_wspd, aes(bin, mean_delay)) +
  geom_col() + labs(title = "Mean delay across wind speed quintiles",
                    x = "Wind speed quintile (low → high)", y = "Mean dep delay")
```


Mean delay across wind speed quintiles
Mean dep delay
15
10
5
0
2
4
6
Wind speed quintile (low → high)