HW_8

GA

2022-11-11

Consider the data in the nycflights13 package.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(nycflights13)
```

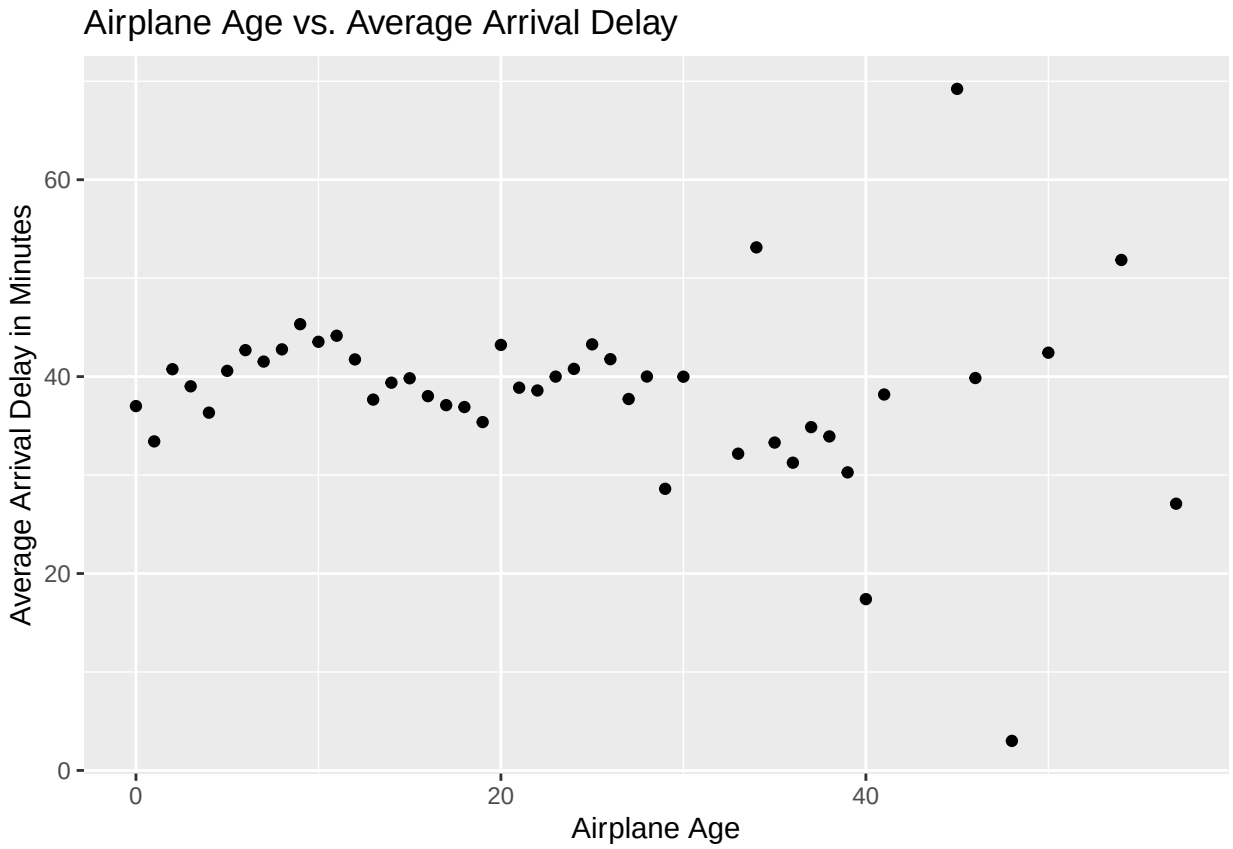
Exercise 1: Is there a relationship between the age of a plane and its delays?

```
#rename variable year in planes to plane_year so it's distinct from year in flights
planes1 <- planes %>%
  select(tailnum:year) %>%
  rename(plane_year = "year")
head(planes1)
```

```
## # A tibble: 6 x 2
##   tailnum plane_year
##   <chr>        <int>
## 1 N10156        2004
## 2 N102UW        1998
## 3 N103US        1999
## 4 N104UW        1999
## 5 N10575        2002
## 6 N105UW        1999
```

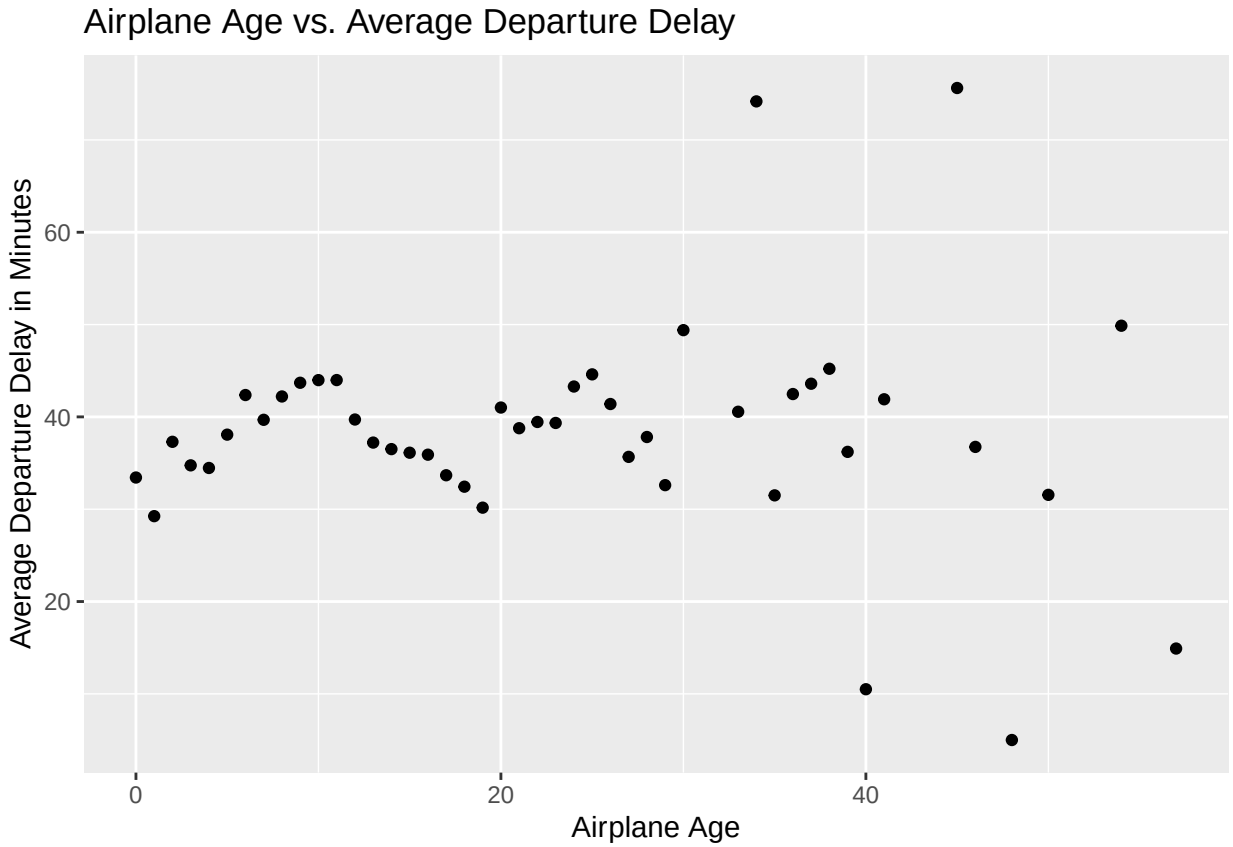
```
flights %>%
  left_join(planes1, by = "tailnum") %>%       #left join flights and planes1
  mutate(plane_age = year - plane_year) %>%    #create a new variable of plane_age
  filter(arr_delay > 0) %>%                    #filter out non-delay entries
  group_by(plane_age) %>%
  summarise(arr_delay_mean = mean(arr_delay)) %>%      #calculate the average arrival delay per plane_a
  ggplot(mapping = aes(x = plane_age, y = arr_delay_mean)) +
  geom_point() +
  labs(x = "Airplane Age", y = "Average Arrival Delay in Minutes", title = "Airplane Age vs. Average Ar
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```




```
flights %>%
  left_join(planes1, by = "tailnum") %>%       #left join flights and planes1
  mutate(plane_age = year - plane_year) %>%    #create a new variable of plane_age
  filter(dep_delay > 0) %>%                    #filter out non-delay entries
  group_by(plane_age) %>%
  summarise(dep_delay_mean = mean(dep_delay)) %>%      #calculate the average departure delay per plane
  ggplot(mapping = aes(x = plane_age, y = dep_delay_mean)) +
  geom_point()+
  labs(x = "Airplane Age", y = "Average Departure Delay in Minutes", title = "Airplane Age vs. Average
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



Thus, there seems (both in arrival and departure delay) some kind of relation, but it is not strong.

Exercise 2: Find the 10 days of the year that have the highest median departure delay, then select all flights from those 10 days.

```
filter(flights, min_rank(desc(dep_delay))<=10)
```

```
## # A tibble: 10 x 19
##     year month   day dep_time sched_de~1 dep_d~2 arr_t~3 sched~4 arr_d~5 carrier
##    <int> <int> <int>    <int>      <int>   <dbl>   <int>   <int>   <dbl> <chr>
##  1  2013     1     9      641        900    1301    1242    1530    1272 HA
##  2  2013     1    10     1121       1635    1126    1239    1810    1109 MQ
##  3  2013    12     5      756       1700     896    1058    2020     878 AA
##  4  2013     3    17     2321        810     911     135    1020     915 DL
##  5  2013     4    10     1100       1900     960    1342    2211     931 DL
##  6  2013     6    15     1432       1935    1137    1607    2120    1127 MQ
##  7  2013     6    27      959       1900     899    1236    2226     850 DL
##  8  2013     7    22      845       1600    1005    1044    1815     989 MQ
##  9  2013     7    22     2257        759     898     121    1026     895 DL
## 10  2013     9    20     1139       1845    1014    1457    2210    1007 AA
## # ... with 9 more variables: flight <int>, tailnum <chr>, origin <chr>,
## #   dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>,
## #   time_hour <dttm>, and abbreviated variable names 1: sched_dep_time,
## #   2: dep_delay, 3: arr_time, 4: sched_arr_time, 5: arr_delay
```

```r
flights %>% top_n(n = 10, wt = dep_delay)
```

```
## # A tibble: 10 x 19
##     year month   day dep_time sched_de~1 dep_d~2 arr_t~3 sched~4 arr_d~5 carrier
##    <int> <int> <int>    <int>      <int>   <dbl>   <int>   <int>   <dbl> <chr>
##  1  2013     1     9      641        900    1301    1242    1530    1272 HA
##  2  2013     1    10     1121       1635    1126    1239    1810    1109 MQ
##  3  2013    12     5      756       1700     896    1058    2020     878 AA
##  4  2013     3    17     2321        810     911     135    1020     915 DL
##  5  2013     4    10     1100       1900     960    1342    2211     931 DL
##  6  2013     6    15     1432       1935    1137    1607    2120    1127 MQ
##  7  2013     6    27      959       1900     899    1236    2226     850 DL
##  8  2013     7    22      845       1600    1005    1044    1815     989 MQ
##  9  2013     7    22     2257        759     898     121    1026     895 DL
## 10  2013     9    20     1139       1845    1014    1457    2210    1007 AA
## # ... with 9 more variables: flight <int>, tailnum <chr>, origin <chr>,
## #   dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>,
## #   time_hour <dttm>, and abbreviated variable names 1: sched_dep_time,
## #   2: dep_delay, 3: arr_time, 4: sched_arr_time, 5: arr_delay
```