Lab 4 Solution: Flights

STAT 20 Spring 2022

Part I: Understanding the Context of the Data

1. What is the unit of observation in the data frame on the handout?

Each unit of observation is a flight record departing from SFO in 2020.

2. Which variables are categorical?

"year", "month", "day", "carrier", "flight", "tailnum", "Origin", "dest", "hour", "minute", "time hour".

3. Which variables are numerical?

"dep_time", "sched_dep_time", "dep_delay", "arr_time", "sched_arr_time", "arr_delay", "air_time", "distance".

4. Do any of the variable have ambiguous data types to you?

Month, and date could be some confusing data type, that appears to be numerical to many students. Flight number, without careful examination, could also be mistreated as numerical.

5. Is there any discernible pattern to the manner in which the rows are ordered?

sched_dep_time are not in order.

6. What is your guess for the units/format used to record the departure time?

integers.

7. What filter would you use to extract the flights that left in the springtime?

filter(flights, month >= 3 & month <= 6)

or

flights %>% filter(month >= 3 & month <= 6)

Part II: Computing on the Data

1. **filter()**: Filter the data set to contain only the flights that went to Portland, Oregon and print the first few rows of the data frame. How many were there in 2020?

flights %>% filter(dest == "PDX") %>% head(10)

```
flights %>% filter(dest == "PDX") %>% summarise(count = n())
```

```
## # A tibble: 1 x 1
##   count
##   <int>
## 1  3882
```

1. **mutate()**: Mutate a new variable called `avg_speed` that is the average speed of the plane during the flight, measured in miles per hour. (Look through the column names or the help file to find variables that can be used to calculate this.)

flights %>% mutate(avg_speed = distance / air_time * 60)

1. **arrange()**: Arrange the data set to figure out: which flight holds the record for longest departure delay (in hrs) and what was its destination? What was the destination and delay time (in hrs) for the flight that was least delayed, i.e. that left the most ahead of schedule?

```
flights %>% filter(dep_delay == max(select(flights, dep_delay), na.rm = TRUE)) %>%
  mutate(delay_in_hr = dep_delay/60) %>% select(dest, delay_in_hr)
```

```
## # A tibble: 1 x 2
##   dest  delay_in_hr
##   <chr>       <dbl>
## 1 PHX            29
```

```
flights %>% filter(dep_delay == min(select(flights, dep_delay), na.rm = TRUE)) %>%
  mutate(delay_in_hr = dep_delay/60) %>% select(dest, delay_in_hr)
```

```
## # A tibble: 1 x 2
##   dest  delay_in_hr
##   <chr>       <dbl>
## 1 GEG        -0.667
```

1. **summarize()**: Confirm the records for departure delay from the question above by summarizing that variable by its maximum and its minimum value.

```
flights %>% summarize(max_delay_hr = max(dep_delay/60, na.rm = TRUE), min_delay_hr = min(dep_delay/60, r
```

```
## # A tibble: 1 x 2
##   max_delay_hr min_delay_hr
##          <dbl>        <dbl>
## 1           29       -0.667
```

6. How many flights left SFO during March 2020?

```
flights %>% filter(month == 3) %>% summarize(num_flights  = n())
```

```
## # A tibble: 1 x 1
##   num_flights
##         <int>
## 1       18415
```

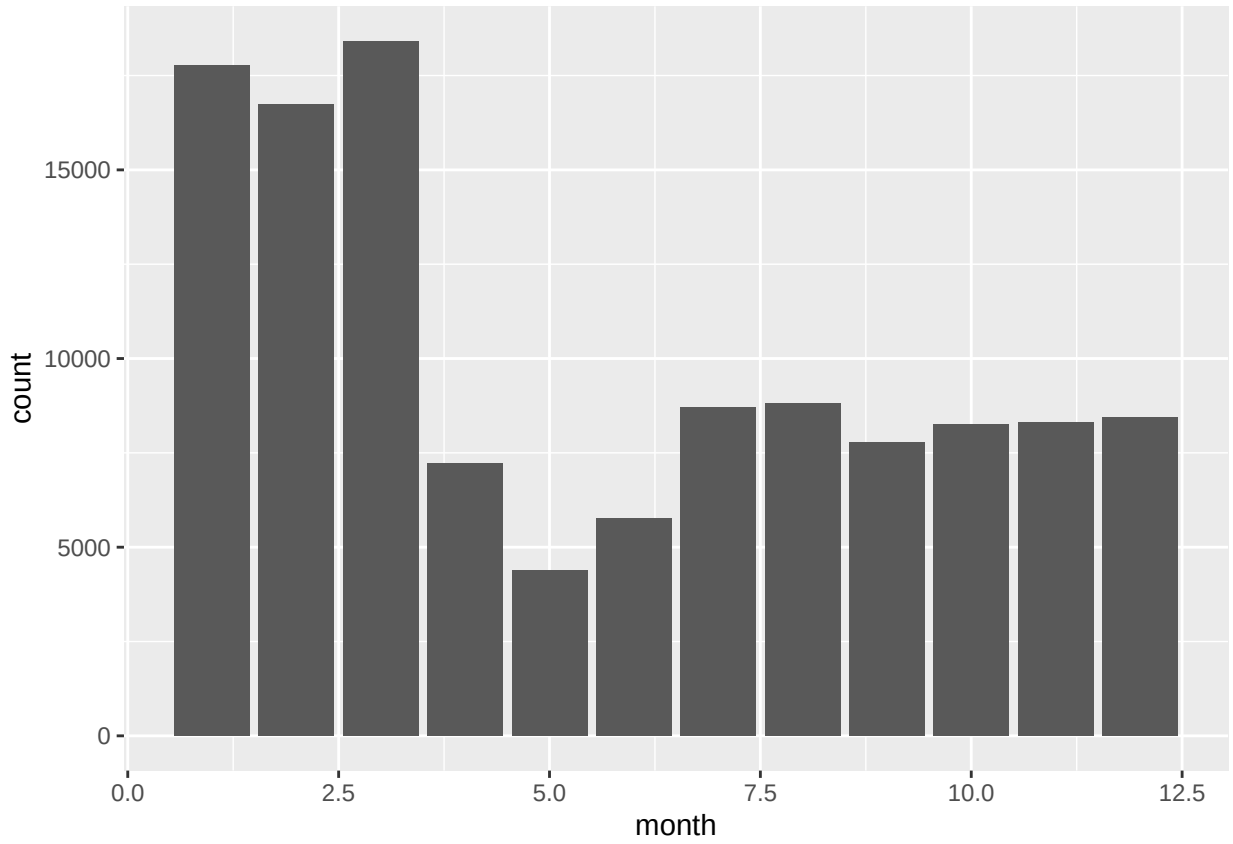
7. How many flights left SFO during April 2020?

```
flights %>% filter(month == 4) %>% summarize(num_flights  = n())
```

```
## # A tibble: 1 x 1
##   num_flights
##         <int>
## 1        7224
```

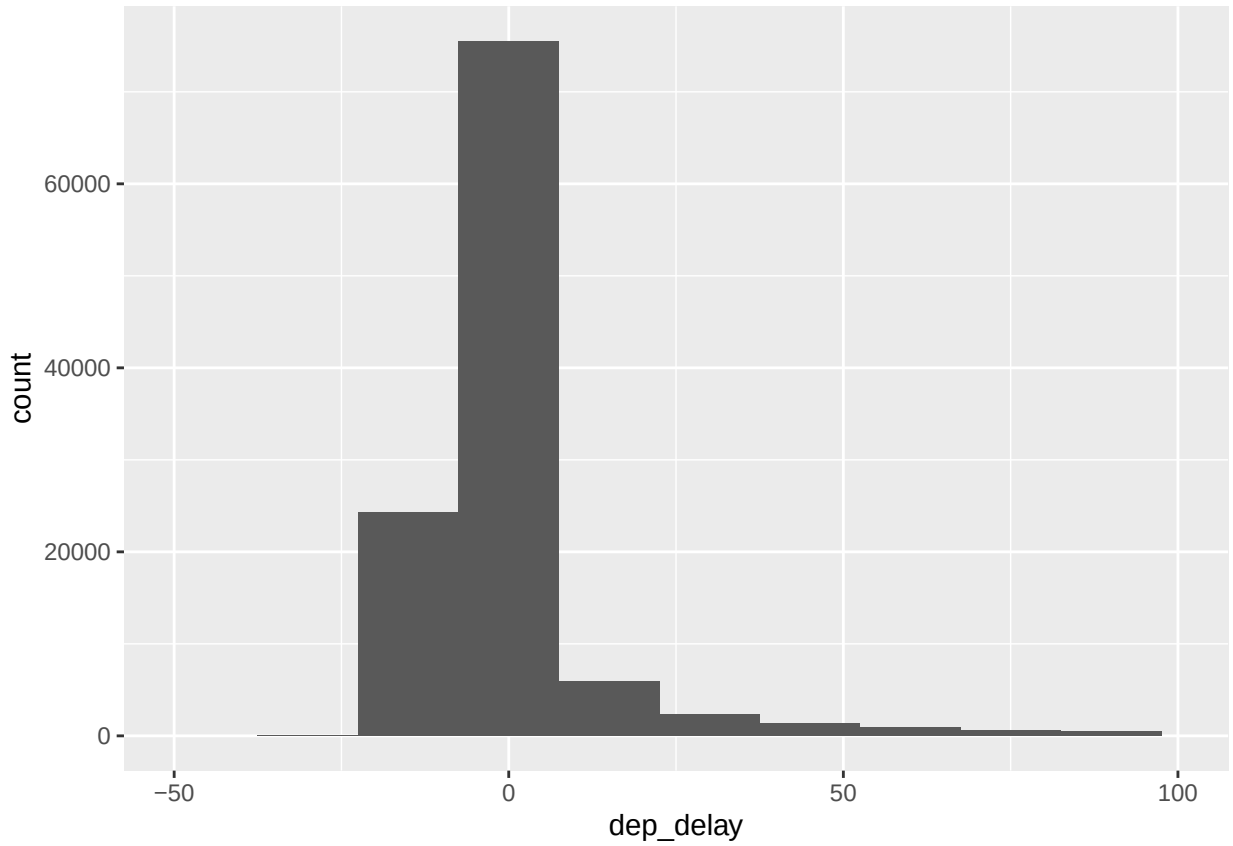
8. Create a bar chart that shows the distribution all of the flights leaving the Bay Area. Do you any sign of an effect of the pandemic?

```
flights  %>% ggplot(aes(x = month)) + geom_bar()
```



9. Create a histogram showing the distribution of departure delays for all flights. Describe in words the shape and modality of the distribution and, using numerical summaries, (i.e. summary statistics) its center and spread. Be sure to use measures of center and spread that are most appropriate for this type of distribution. Also set the limits of the x-axis to focus on where most of the data lie.

```
flights %>%
  ggplot(aes(x = dep_delay)) +
  geom_histogram(binwidth = 15) +
  xlim(c(-50, 100))
```



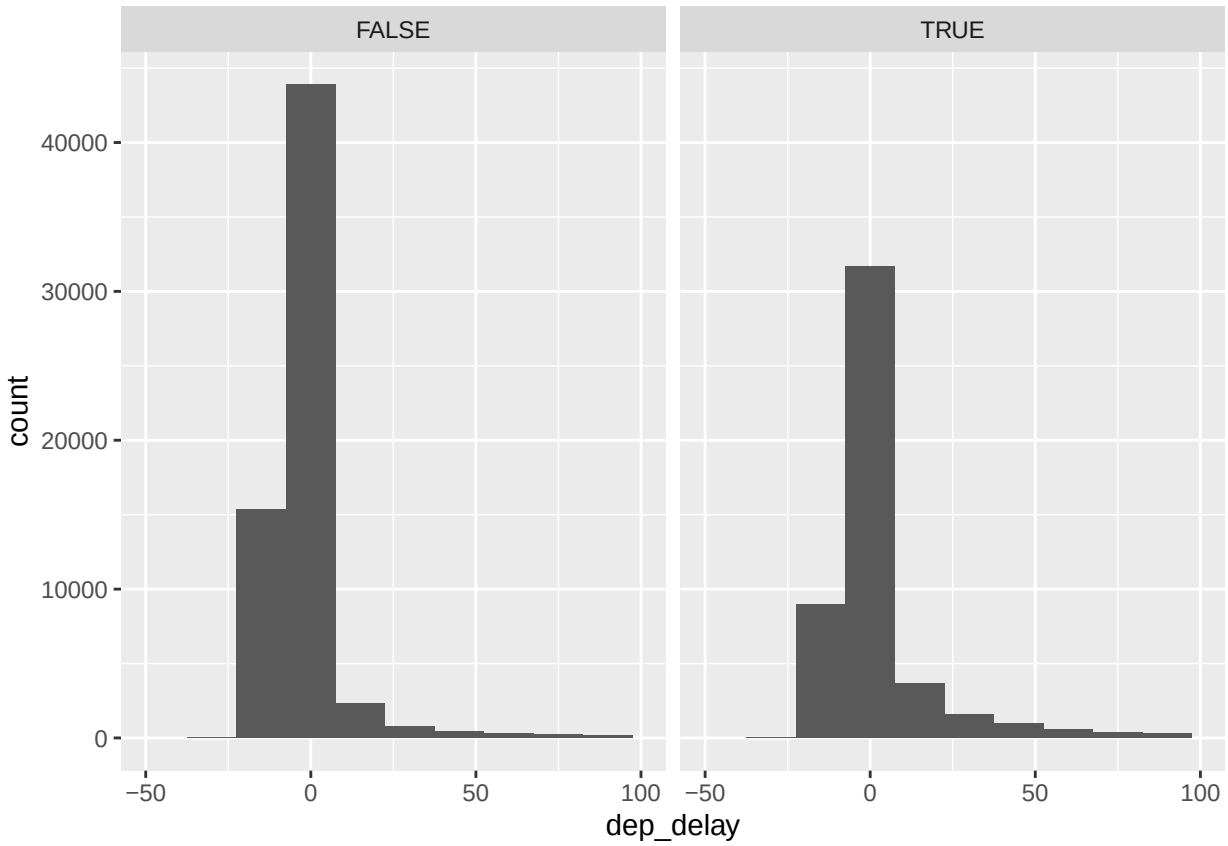

```
flights %>%
  summarise(med_delay = median(dep_delay, na.rm = T),
            iqr_delay = IQR(dep_delay, na.rm = T))
```

```
## # A tibble: 1 x 2
##   med_delay iqr_delay
##       <dbl>     <dbl>
## 1        -4         6
```

Encourage them to spiff up this plot by zooming in on the x-axis. The statistical summaries should be median and IQR since it's a skewed distribution. The shape they should describe as unimodal and right skewed.

10. Add a new column to your data frame called `before_times` that takes values of `TRUE` and `FALSE` indicating whether the flight took place up through the end of March or after April 1st, respectively. Remake the histograms above, but now separated into two subplots: one with the departure delays from the before times, the other with the flights from afterwards.

```
flights %>%
  mutate(before_times = month <= 3) %>%
  ggplot(aes(x = dep_delay)) +
  geom_histogram(binwidth = 15) +
  facet_wrap(vars(before_times)) +
  xlim(c(-50, 100))
```



Can you visually detect any difference in the distribution of departure delays?

12. If you flew out of OAK or SFO during this time period, what is the tail number of the plane that you were on? If you did not fly in this period, find the tail number of the plane that flew JetBlue flight 40 to New York's JFK Airport from SFO on May 1st.

```
flights %>%
  filter(carrier == "B6" & flight == 40 &
           dest == "JFK" & month == 5 & day == 1) %>%
  select(tailnum)
```

```
## # A tibble: 1 x 1
##   tailnum
##   <chr>
## 1 N982JB
```

13. What proportion of the flights left on or ahead of schedule?

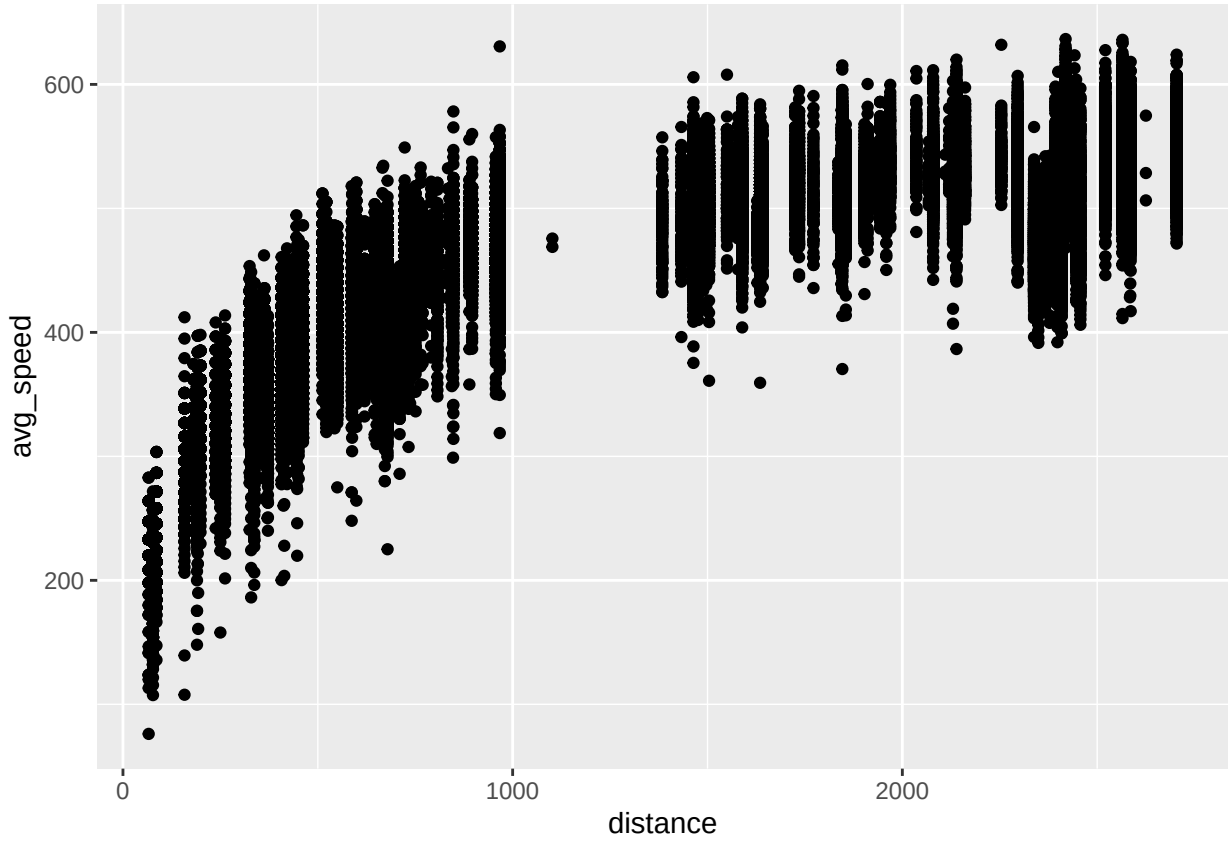
```
flights %>% summarise(ratio = mean(dep_delay <= 0, na.rm=TRUE))
```

```
## # A tibble: 1 x 1
##   ratio
##   <dbl>
## 1 0.811
```

14. Create a plot that captures the relationship of average speed vs. distance and describe the shape that you see. What phenomena related to taking flights from the Bay Area might explain this structure? Make any modifications needed to help you see all of the data.

```
flights <- flights %>%
  mutate(avg_speed = distance/(air_time/60))

ggplot(flights, aes(x = distance, y = avg_speed)) +
  geom_point()
```



As the distance of a flight increases, so too does the average speed. The increase is roughly quadradratic in shape and levels out at an average maximum around 500 mph. Good job if they add alpha transparency.

15. What is the most common destination of the flights from the Bay Area? The most distant destination?

```
# common destination
flights %>%
  filter(origin == "SFO") %>%
  group_by(dest) %>% # can also use count()
  summarise(n = n()) %>%
  arrange(desc(n))
```

```
## # A tibble: 85 x 2
##    dest      n
##    <chr> <int>
##  1 LAX    8513
##  2 SEA    5000
##  3 SAN    4442
##  4 LAS    4422
##  5 DEN    3709
##  6 JFK    3388
##  7 ORD    3197
```

```
##  8 PHX    2854
##  9 SLC    2792
## 10 EWR    2643
## # ... with 75 more rows
```

```
# most distant dest
flights %>%
  filter(origin == "SFO") %>%
  arrange(desc(distance))
```

```
## # A tibble: 89,394 x 20
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <dbl> <dbl> <dbl>    <dbl>          <dbl>     <dbl>    <dbl>          <dbl>
##  1  2020     1     1      718            720        -2     1558           1600
##  2  2020     1     1      812            820        -8     1629           1652
##  3  2020     1     1      830            830         0     1712           1703
##  4  2020     1     1     1042           1040         2     1910           1915
##  5  2020     1     1     1410           1414        -4     2230           2247
##  6  2020     1     1     1436           1440        -4     2257           2309
##  7  2020     1     1     2036           2050       -14      449            517
##  8  2020     1     1     2152           2200        -8      613            625
##  9  2020     1     1     2217           2220        -3      746            705
## 10  2020     1     1     2347           2350        -3      814            819
## # ... with 89,384 more rows, and 12 more variables: arr_delay <dbl>,
## #   carrier <chr>, flight <dbl>, tailnum <chr>, origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>,
## #   avg_speed <dbl>
```

There are some variants on this code that are probably ok. Their answer should also provide full sentence answers, including saying the actual name of the airports.

16. For OAK and SFO separately, what proportion of the flights left on or ahead of schedule?

```
flights %>% filter(origin == "SFO") %>%
  summarise(proportion = mean(dep_delay <= 0, na.rm = TRUE))
```

```
## # A tibble: 1 x 1
##   proportion
##        <dbl>
## 1      0.818
```

```
flights %>% filter(origin == "OAK") %>%
  summarise(proportion = mean(dep_delay <= 0, na.rm = TRUE))
```

```
## # A tibble: 1 x 1
##   proportion
##        <dbl>
## 1      0.792
```

17. Create a data frame that contains the median and interquartile range for departure delays, grouped by carrier. Which carrier has the lowest typical departure delay? Which one has the least variable departure delays?

```
flights %>%
  group_by(carrier) %>%
  summarise(med_delay = median(dep_delay, na.rm = TRUE),
            iqr_delay = IQR(dep_delay, na.rm = TRUE)) %>%
  arrange(iqr_delay)
```

```
## # A tibble: 12 x 3
##    carrier med_delay iqr_delay
##    <chr>       <dbl>     <dbl>
##  1 DL             -5         5
##  2 WN             -3         5
##  3 AA             -6         6
##  4 NK             -4         6
##  5 UA             -5         6
##  6 B6             -8         7
##  7 OO             -5         7
##  8 AS             -7         9
##  9 F9             -5         9
## 10 HA             -5         9
## 11 YV             -7        11
## 12 G4             -7      13.8
```

Hawaiian Airlines has the least variable departure delays with an IQR of 7 minutes.

Part III: Extensions

18. For flights leaving SFO, which month has the highest average departure delay? What about the highest median departure delay? Which of these measures is more reliable for deciding which month(s) to avoid flying if you really dislike delayed flights?

```
flights %>%
  filter(origin == "SFO") %>%
  group_by(month) %>%
  summarize(mean_delay = mean(dep_delay, na.rm = TRUE),
            med_delay = median(dep_delay, na.rm = TRUE))
```

```
## # A tibble: 12 x 3
##    month mean_delay med_delay
##    <dbl>      <dbl>     <dbl>
##  1     1      9.05         -3
##  2     2      8.06         -3
##  3     3      1.15         -5
##  4     4     -3.81         -7
##  5     5     -2.68         -6
##  6     6     -1.86         -5
##  7     7     -1.56         -6
##  8     8     -1.80         -5
##  9     9     -1.75         -5
## 10    10     -0.155        -5
## 11    11     -1.91         -6
## 12    12     -0.970        -5
```

Answers should be in full sentences. Median is the more useful metric since delays are heavily right skewed. They might point out that the winter months do have much higher means, suggesting the presence of high outliers.

19. Each individual airplane can be uniquely identified by its tailnumber in the same way that people can be by their social security numbers. Which airplane flew the farthest during this year for which we have data? How many times around the planet does that translate to?

```
flights %>%
  group_by(tailnum) %>%
```

```
  summarize(total_dist = sum(distance)) %>%
  arrange(desc(total_dist))
```

```
## # A tibble: 3,774 x 2
##    tailnum total_dist
##    <chr>        <dbl>
##  1 <NA>       4570452
##  2 N705TW      245670
##  3 N980JT      243490
##  4 N969JT      242297
##  5 N986JB      242229
##  6 N984JB      238697
##  7 N983JT      238624
##  8 N968JT      234069
##  9 N989JT      231144
## 10 N977JE      229138
## # ... with 3,764 more rows
```

Students might have grabbed a different number for the denominator. It's fine if it's slightly different, but if it looks like it is in km, then its wrong because the numerator is in miles.

20. What is the tailnumber of the fastest plane in the dataset? What type of plane is it (google it!)? Be sure to be clear how you're defining fastest.

```
flights %>%
  mutate(speed = distance / air_time) %>%
  group_by(tailnum) %>%
  summarize(avg_speed = mean(speed),
            cnt = n()) %>%
  arrange(desc(cnt))
```

```
## # A tibble: 3,774 x 3
##    tailnum avg_speed   cnt
##    <chr>       <dbl> <int>
##  1 <NA>         NA     4454
##  2 N184SY       NA      293
##  3 N400SY        6.33   288
##  4 N192SY       NA      286
##  5 N179SY       NA      283
##  6 N191SY       NA      266
##  7 N402SY       NA      258
##  8 N174SY       NA      256
##  9 N183SY       NA      256
## 10 N187SY       NA      255
## # ... with 3,764 more rows
```

There are several reasonable ways to define fastest, including fastest average time and fastest single flights.

21. Using the airport nearest your hometown, which day of the week and which airline seems best for flying there from San Francisco (if you're from near SFO or OAK or from abroad, use Chicago as your hometown)? Be clear on how you're defining *best*. (note that there is no explicit weekday column in this data set, but there is sufficient information to piece it together. The following line of code can be added to your pipeline to create that new column. It uses functions in the `lubridate` package, so be sure to load it in at the start of this exercise).

```
library(lubridate)
flights %>% mutate(day_of_week = wday(ymd(paste(year, month, day, set = "-")), label = T)) %>% head()
```

```
## # A tibble: 6 x 21
##    year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##   <dbl> <dbl> <dbl>    <dbl>          <dbl>     <dbl>    <dbl>          <dbl>
## 1  2020     1     1        8           2359         9      528            532
## 2  2020     1     1       29             39       -10      356            420
## 3  2020     1     1       37             40        -3      846            856
## 4  2020     1     1       41             45        -4      908            913
## 5  2020     1     1       44           2300       104      834            709
## 6  2020     1     1       48             56        -8      641            658
## # ... with 13 more variables: arr_delay <dbl>, carrier <chr>, flight <dbl>,
## #   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>,
## #   hour <dbl>, minute <dbl>, time_hour <dttm>, avg_speed <dbl>,
## #   day_of_week <ord>
```

```
flights %>%
  mutate(day_of_week = wday(ymd(paste(year, month, day, set = "-")), label = T)) %>%
  filter(origin == "SFO", dest == "LAX") %>%
  group_by(day_of_week, carrier) %>%
  summarise(med_delay = median(dep_delay, na.rm = T)) %>%
  arrange(med_delay)
```

```
## # A tibble: 49 x 3
## # Groups:   day_of_week [7]
##    day_of_week carrier med_delay
##    <ord>       <chr>       <dbl>
##  1 Wed         B6            -11
##  2 Sun         B6            -10
##  3 Tue         B6            -10
##  4 Thu         B6            -10
##  5 Tue         AS             -9
##  6 Fri         B6             -9
##  7 Mon         AS             -8
##  8 Mon         B6             -8
##  9 Tue         AA             -8
## 10 Wed         AA             -7
## # ... with 39 more rows
```

This is the answer for Los Angeles - answers will vary. They may also come up with a different notion of "best" (here I used median departure delay) and they may do the grouping differently: decide on best first for one category, then separately for the other category.

22. The plot below displays the relationship between the mean arrival delay and the mean distance traveled by every plane in the data set. It also shows the total number of flights made by each plane by the size of the plotted circle. Please form a single chain that will create this plot, starting with the raw data set. You will also want to exclude the edge cases from your analysis, so focus on the planes that have logged more than 20 flights and flown an average distance of less than 2000 miles.

```
flights %>%
  group_by(tailnum) %>%
  summarize(avg_delay = mean(arr_delay, na.rm = TRUE),
            avg_dist = mean(distance, na.rm = TRUE),
            n = n()) %>%
  filter(n > 20,
         n < 500,
         avg_dist < 2000) %>%
  ggplot(aes(x = avg_dist, y = avg_delay)) +
```

```
  geom_point(alpha = .1, aes(size = n)) +
  labs(x = "average distance",
       y = "average delay",
       size = "number of flights")
```