DATA 606 Lab 2

Magnus Skonberg

2020-09-06

To load nyc flight data and view names of variables.

```r
data(nycflights)
```

```r
names(nycflights)
```

Exercise 1

Let's vary the bin widths on these three histograms and observe. How do they compare? Are features revealed in one that are obscured in another?

1st histogram (default bin width of 30): 6 bins, max value significantly above 20000. 2nd histogram (bin width of 15): 13 bins, max value just above 20000. 3rd histogram (bin width of 150): 3 bins, max value just above 30000.

The 2nd histogram, that with the narrowest bin width, gave the most accurate breakdown of the number of flights that fell into each delay category while the 3rd histogram, that with the widest bin width, was the most obscure since there were only 3 bins to represent 32735 observations worth of flight delay data and thus each bin represented a much larger range in delay.

```r
ggplot(data = nycflights, aes(x = dep_delay)) +
  geom_histogram()
```

```r
ggplot(data = nycflights, aes(x = dep_delay)) +
  geom_histogram(binwidth = 15)
```

```r
ggplot(data = nycflights, aes(x = dep_delay)) +
  geom_histogram(binwidth = 150)
```

Exercise 2

Create a new data frame that includes flights headed to SFO in February, and save this data frame as sfo__feb__flights. How many flights meet these criteria?

68 flights meet this criteria.

```r
sfo_feb_flights <- nycflights %>%
  filter(dest == "SFO", month == 2)

count(sfo_feb_flights)
```

Exercise 3

Describe the distribution of the arrival delays of these flights using a histogram and appropriate summary statistics. Hint: The summary statistics you use should depend on the shape of the distribution.

```
ggplot(data = sfo_feb_flights, aes(x = arr_delay)) +
  geom_histogram(binwidth = 15)
```

```
sfo_feb_flights %>%
  summarise(
            min_ad = min(arr_delay),
            max_ad = max(arr_delay),
            median_ad = median(arr_delay),
            IQR_ad = IQR(arr_delay)
            )
```

Exercise 4

Calculate the median and interquartile range for arr_delays of flights in in the sfo_feb_flights data frame, grouped by carrier. Which carrier has the most variable arrival delays?

Delta (DL) and United Airlines (UA) have the largest IQR values (Q3 - Q1) and thus the most variable arrival delays.

```
sfo_feb_flights %>%
  group_by(carrier) %>%
  summarise(median_ad = median(arr_delay), iqrad = IQR(arr_delay), n_flights = n())
```

Exercise 5

Suppose you really dislike departure delays and you want to schedule your travel in a month that minimizes your potential departure delay leaving NYC. One option is to choose the month with the lowest mean departure delay. Another option is to choose the month with the lowest median departure delay. What are the pros and cons of these two choices?

The lowest mean would give the average departure delay while the median would give the observation that sits at the middle of the data set. The lowest mean accounts for outlier values and thus would be the better choice when accounting for delays that are well-above average while the median would not account for outlier values and thus give a better idea of departure delays without accounting for the extraneous cases. For this occasion, the mean could inform a better decision.

July (mo 7) is the month with the highest average departure delay.

```
nycflights %>%
  group_by(month) %>%
  summarise(mean_dd = mean(dep_delay)) %>%
  arrange(desc(mean_dd))
```

Exercise 6

If you were selecting an airport simply based on on time departure percentage, which NYC airport would you choose to fly out of?

LaGuardia (LGA) airport.

```
nycflights <- nycflights %>%
  mutate(dep_type = ifelse(dep_delay < 5, "on time", "delayed"))
```

```
nycflights %>%
  group_by(origin) %>%
  summarise(ot_dep_rate = sum(dep_type == "on time") / n()) %>%
  arrange(desc(ot_dep_rate))
```

Exercise 7

Mutate the data frame so that it includes a new variable that contains the average speed, `avg_speed` traveled by the plane for each flight (in mph). Hint: Average speed can be calculated as distance divided by number of hours of travel, and note that `air_time` is given in minutes.

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

Exercise 8

Make a scatterplot of avg_speed vs. distance. Describe the relationship between average speed and distance. Hint: Use geom_point().

The average speed increases until the distance is in the range of 1000-1500 (miles) and then it stabilizes.

```
ggplot(data = nycflights, mapping = aes(x = distance, y = avg_speed)) +
        geom_point()
```

Exercise 9

Replicate the following plot. Hint: The data frame plotted only contains flights from American Airlines, Delta Airlines, and United Airlines, and the points are colored by carrier. Once you replicate the plot, determine (roughly) what the cutoff point is for departure delays where you can still expect to get to your destination on time.

```
c_delay <- nycflights %>%
  filter(carrier == 'AA' | carrier == 'DL' | carrier == 'UA')
ggplot(c_delay, aes(dep_delay, arr_delay, color = carrier)) + geom_point()
```

Based on the below plot, the cutoff point is approximately 20mins. After this point, as departure delays climb so do arrival delays.

```
ggplot(c_delay, aes(dep_delay, arr_delay, color = carrier)) +
        xlim(-10, 60) +
        ylim(-10, 60) +
        geom_point()
```

```
## Warning: Removed 7108 rows containing missing values (geom_point).
```