Assignment 5

Rachel Greenlee

9/21/2020

Introduction

I start with a screenshot of a small dataset and the goal is to put it into a tidy data format and then perform analysis to compare the arrival delays for the two airlines.

Step 1 - Reproduce data in MySQL and import to R

For the sake of practice, I'll create two separate tables in MySQL, one for each airline. In MySQL I create the empty tables with the variables in such a way that the data will be in a long structure per the grading rubric. I then input the handful of rows of data.

```
1  CREATE TABLE alaska_fl (
2      destination VARCHAR(255),
3      on_time NUMERIC(10),
4      delay NUMERIC(10)
5  );
6
7  INSERT INTO alaska_fl (destination, on_time, delay)
8  VALUES ('Los Angeles', 497, 62),
9         ('Phoenix', 221, 12),
10        ('San Diego', 212, 20),
11        ('San Francisco', 503, 102),
12        ('Seattle', 1841, 305);
13
14 CREATE TABLE amwest_fl (
15     destination VARCHAR(255),
16     on_time NUMERIC(10),
17     delay NUMERIC(10)
18 );
19
20 INSERT INTO amwest_fl (destination, on_time, delay)
21 VALUES  ('Los Angeles', 694, 117),
22        ('Phoenix', 4840, 415),
23        ('San Diego', 383, 65),
24        ('San Francisco', 320, 129),
25        ('Seattle', 201, 61);
```

Figure 1: Image of MySQL Code.

Step 2 - Connect R to MySQL to create dataframe

First I load the RMySQL package in order to connect with MySQL. I also load the keyring package so I can access my stored password and keep it hidden from the world. Let's load kable too for tables.

```
library('keyring')
library("RMySQL")
library("kableExtra")
```

Now we must access the datasets from MySQL.And let's preview them quick as well.

```
mysqlpass <- key_get('mysql', 'root')

rachdb = dbConnect(MySQL(), user='root', password=mysqlpass,
  dbname='607assign5_flights', host='localhost')

amwest_fl <- dbGetQuery(rachdb, "select * from amwest_fl")
alaska_fl <- dbGetQuery(rachdb, "select * from alaska_fl")

kable(amwest_fl, format = "markdown")
```

destination	on_time	delay
Los Angeles	694	117
Phoenix	4840	415
San Diego	383	65
San Francisco	320	129
Seattle	201	61

```
kable(alaska_fl, format = "markdown")
```

destination	on_time	delay
Los Angeles	497	62
Phoenix	221	12
San Diego	212	20
San Francisco	503	102
Seattle	1841	305

Step 3 - Use tidyr and dplyr as needed to tidy and transform data

Using dplyr we need to combine our two datasets into one and add a column to identify which airline the data is from.

```
library(dplyr)
```

```
#add airline variable with correct airline for each dataset
alaska_fl <- mutate(alaska_fl, airline = "ALASKA")
amwest_fl <- mutate(amwest_fl, airline = "AMWEST")
```

```
#bind the two datasets together vertically then arrange by destination
flights <- bind_rows(alaska_fl, amwest_fl)
flights <- arrange(flights, destination)
kable(flights, format = "markdown")
```

destination	on_time	delay	airline
Los Angeles	497	62	ALASKA
Los Angeles	694	117	AMWEST
Phoenix	221	12	ALASKA
Phoenix	4840	415	AMWEST
San Diego	212	20	ALASKA
San Diego	383	65	AMWEST
San Francisco	503	102	ALASKA
San Francisco	320	129	AMWEST
Seattle	1841	305	ALASKA
Seattle	201	61	AMWEST

Step 4 - Perform analysis to compare arrival delays for the two airlines

First by looking at a summary of the full dataset we see that the median frequency of on time flights is 440 and the median frequency of delayed flights is 83.5. One airline, we don't know which yet, has the max of 415 delayed flights to a certain destination.

```
kable((summary(flights)), format = "markdown")
```

destination	on_time	delay	airline
Length:10	Min. : 201.0	Min. : 12.00	Length:10
Class :character	1st Qu.: 245.8	1st Qu.: 61.25	Class :character
Mode :character	Median : 440.0	Median : 83.50	Mode :character
NA	Mean : 971.2	Mean :128.80	NA
NA	3rd Qu.: 646.2	3rd Qu.:126.00	NA
NA	Max. :4840.0	Max. :415.00	NA

Looking by airline we can see a comparison between the two. ALASKA has a much higher median on-time frequency across the destinations at 597 compared to AMWEST's 383. AMWEST has a larger IQR suggesting more variance in their on-time rates by destination.

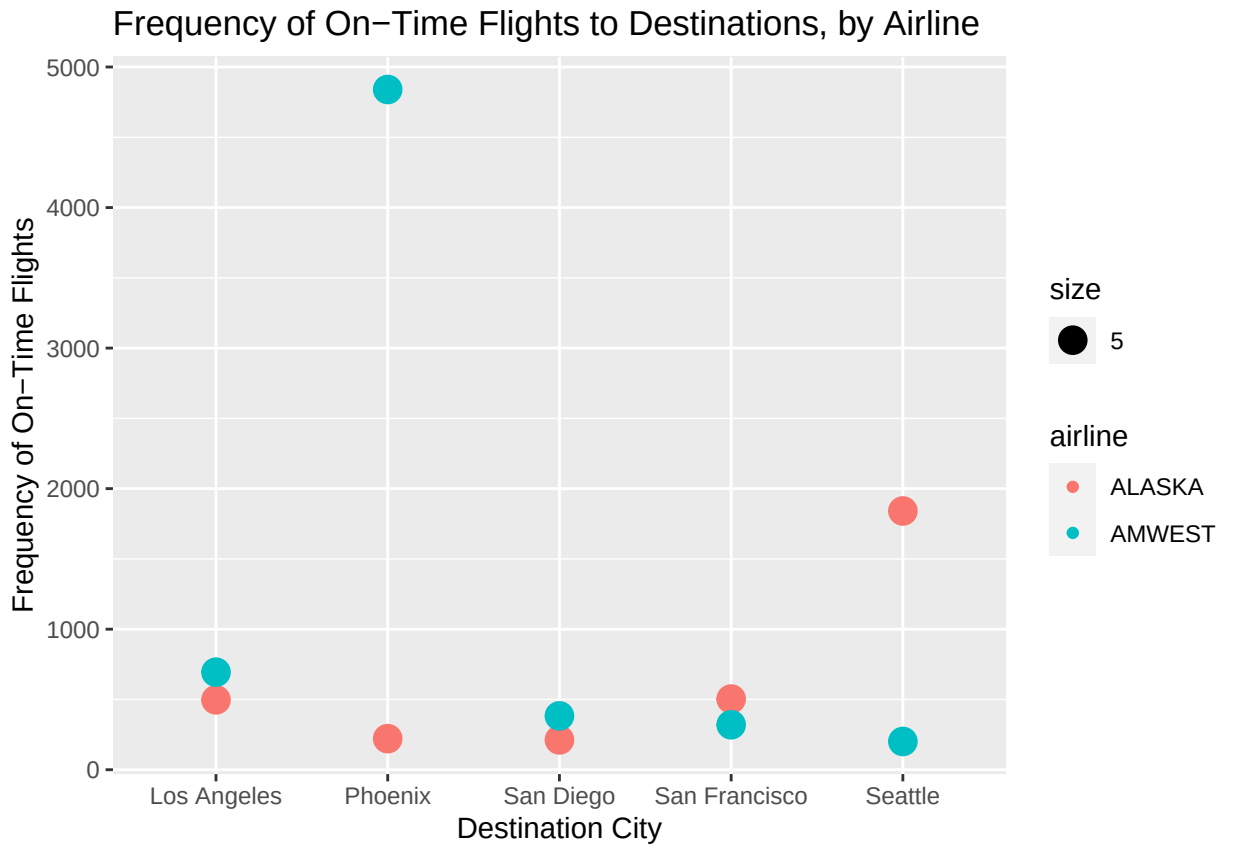
```
kable((flights %>%
        group_by(airline) %>%
        summarise(median_on_time = median(on_time), iqr_on_time = IQR(on_time),
        min_on_time = min(on_time), max_on_time = max(on_time))),
  format = "markdown")
```

airline	median_on_time	iqr_on_time	min_on_time	max_on_time
ALASKA	497	282	212	1841
AMWEST	383	374	201	4840

Since this data is already aggregated, a nice way to visualize it might be this first dot plot below. The number of on-time flights for Los Angeles, San Diego, and San Francisco look similar, but there are large difference in the other two cities. In Phoenix AMWEST has nearly 5,000 on-time flights while ALASKA only has 250. In a flip, ALASKA has the greater amount of on-time flights to Seattle, around 2000, with AMWEST only have around 200.

```
library(ggplot2)
```

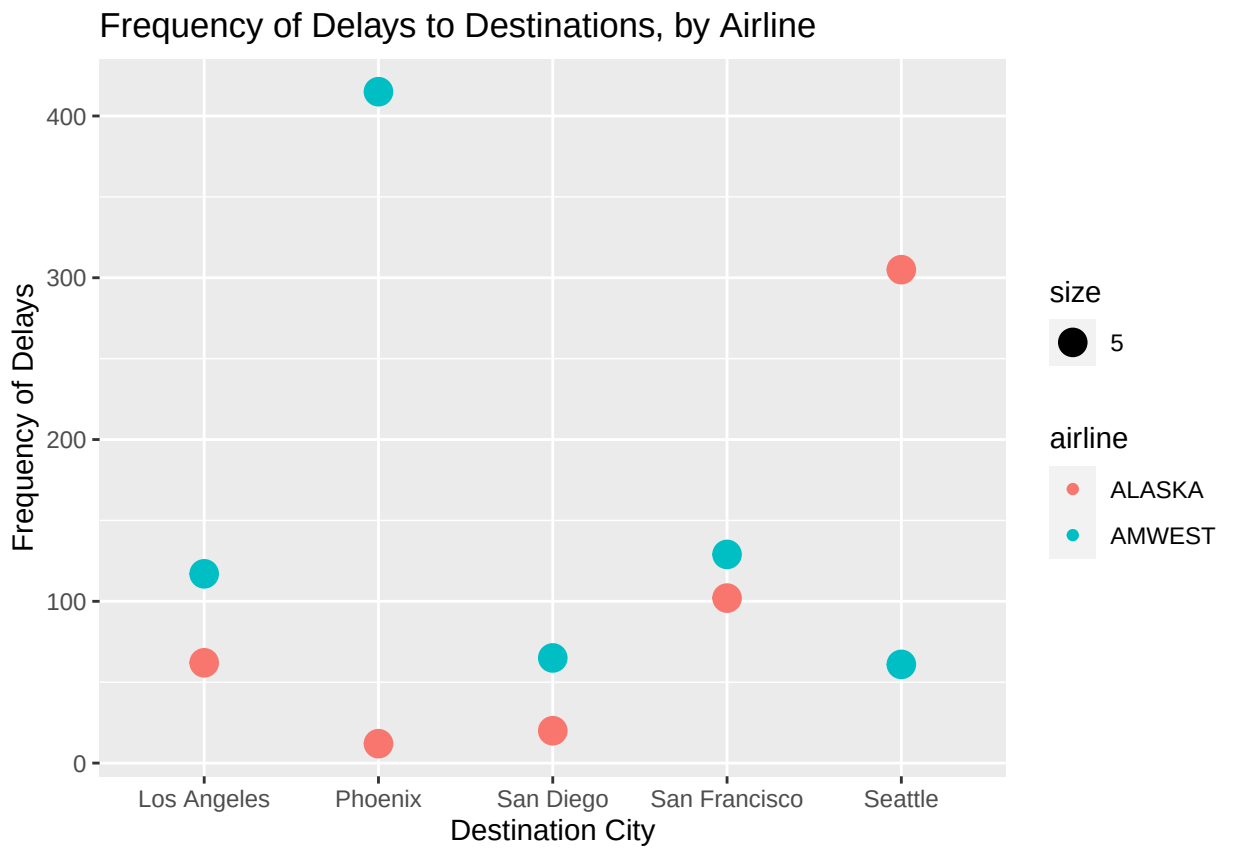
```
ggplot(data = flights, aes(x = on_time, y = destination)) +
  geom_point(aes(color = airline, size = 5)) +
  ylab('Destination City') +
  xlab('Frequency of On-Time Flights') +
  labs(title="Frequency of On-Time Flights to Destinations, by Airline") +
  coord_flip()
```



Below, in looking at the delayed flights variable we can see that across all destinations, except for Seattle, AMWEST has more delays. The largest discrepancies are found in Seattle and Phoenix. Seattle is the only destination where ALASKA has more delays, at over 300 compared to AMWEST's around 60. Phoenix is the inverse of that, ALASKA has what looks to be around 10 delays while that is the worst amount of delays in the entire dataset for AMWEST, that 415 maximum delays.

From this view one would conclude if you were flying to these destinations you'd want to fly ALASKA, unless you're going to Seattle, then it's AMWEST. However...

```
ggplot(data = flights, aes(x = delay, y = destination)) +
  geom_point(aes(color = airline, size = 5)) +
  ylab('Destination City') +
  xlab('Frequency of Delays') +
  labs(title="Frequency of Delays to Destinations, by Airline") +
  coord_flip()
```

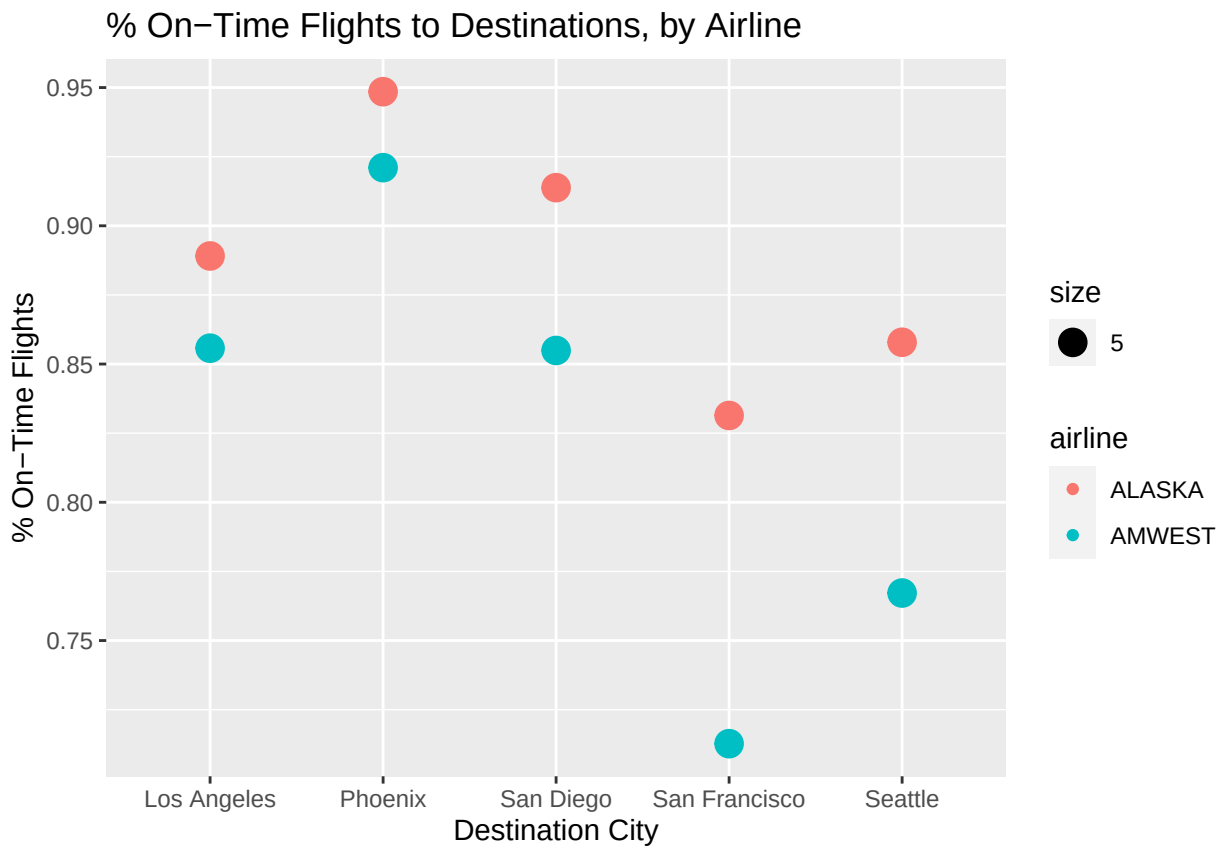


It might be more valuable to look at these not as just counts, but as percents out of the total flights to that destination for that airline. In this final dot plot below, now we see that for every destination, regardless of the raw number of flights, ALASKA has a greater percentage of on-time flights.

This plot also shows us that AMWEST has a particularly hard time being on-time in San Francisco, it's the lowest dot on the chart near 71%.

```
flights <- mutate(flights, on_time_perc = flights$on_time/(flights$on_time + flights$delay))

ggplot(data = flights, aes(x = on_time_perc, y = destination)) +
  geom_point(aes(color = airline, size = 5)) +
  ylab('Destination City') +
  xlab('% On-Time Flights') +
  labs(title="% On-Time Flights to Destinations, by Airline") +
  coord_flip()
```



Conclusion

This process shows how important it is to get the data in the correct format for analysis and graphing, and the importance of not just plotting the variables you have but taking time to create a more logical measure, such as the percent of on-time flights as opposed to the frequency. While the number of times an airline flies on time to a destination certainly holds merit, the percentage allows us to compare between the two airlines better as in some cases one simply didn't fly to a certain destination as often.