607 Week 5

Lu Beyer

Overview:

I am working to transform this data on Alasaka and AM West airline flight delays to more easily compare the airlines based on their on-time and delayed flights.

Load Data

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Make all blank cells "NA"

```
x <- read.csv("607_airline_dirty.csv", na.strings = c("", "NA"))
```

Consolidated destinations into a single destination column

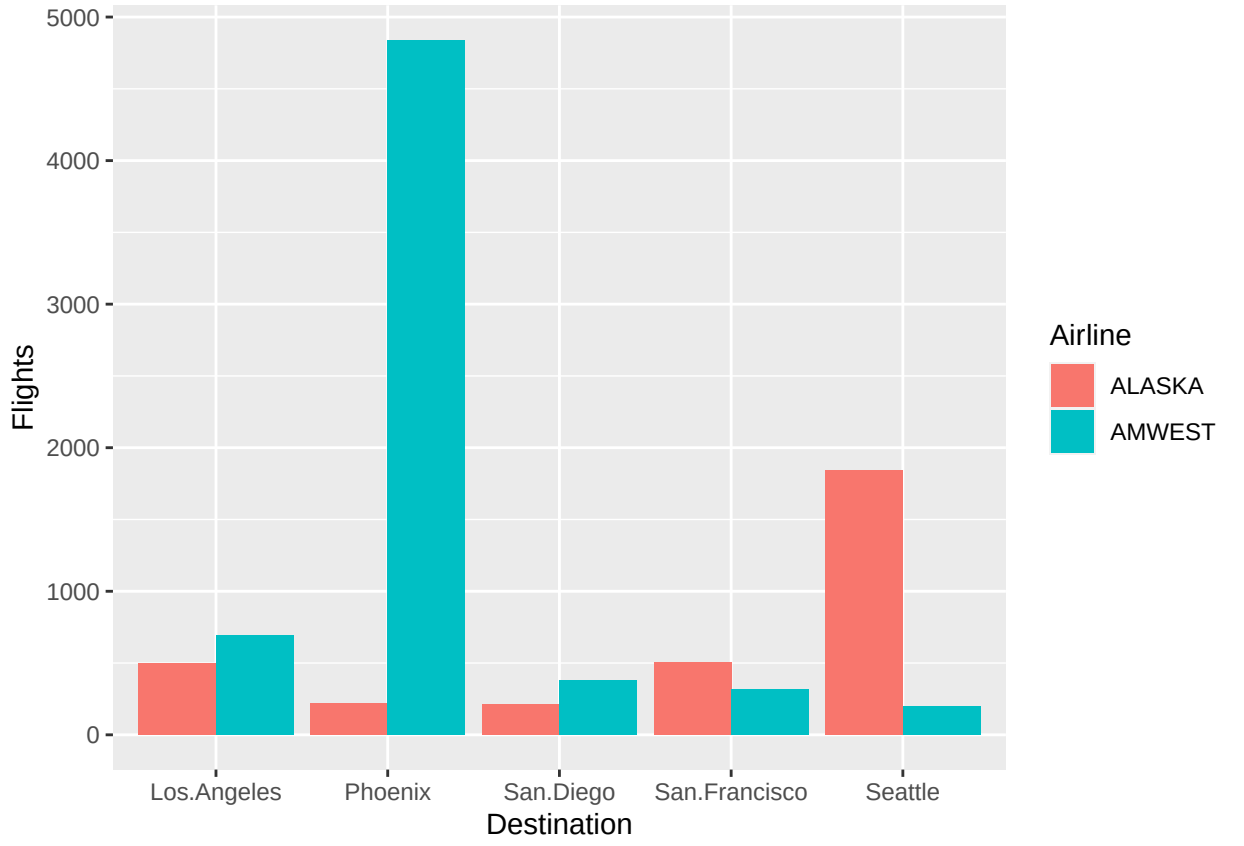
```
y <- x %>%
  pivot_longer(
    cols = c(Los.Angeles, Phoenix, San.Diego, San.Francisco, Seattle),
    names_to = "Destination",
    values_to = "Flights",
    values_drop_na =  TRUE
  ) %>%
  rename(Airline = X, Delay_status = X.1) %>%
  fill(Airline, .direction = c("down"))
```

I did not need to do this, but created a df of on time-flights

```
on_time <- y %>%
  filter(Delay_status == "on time")
```

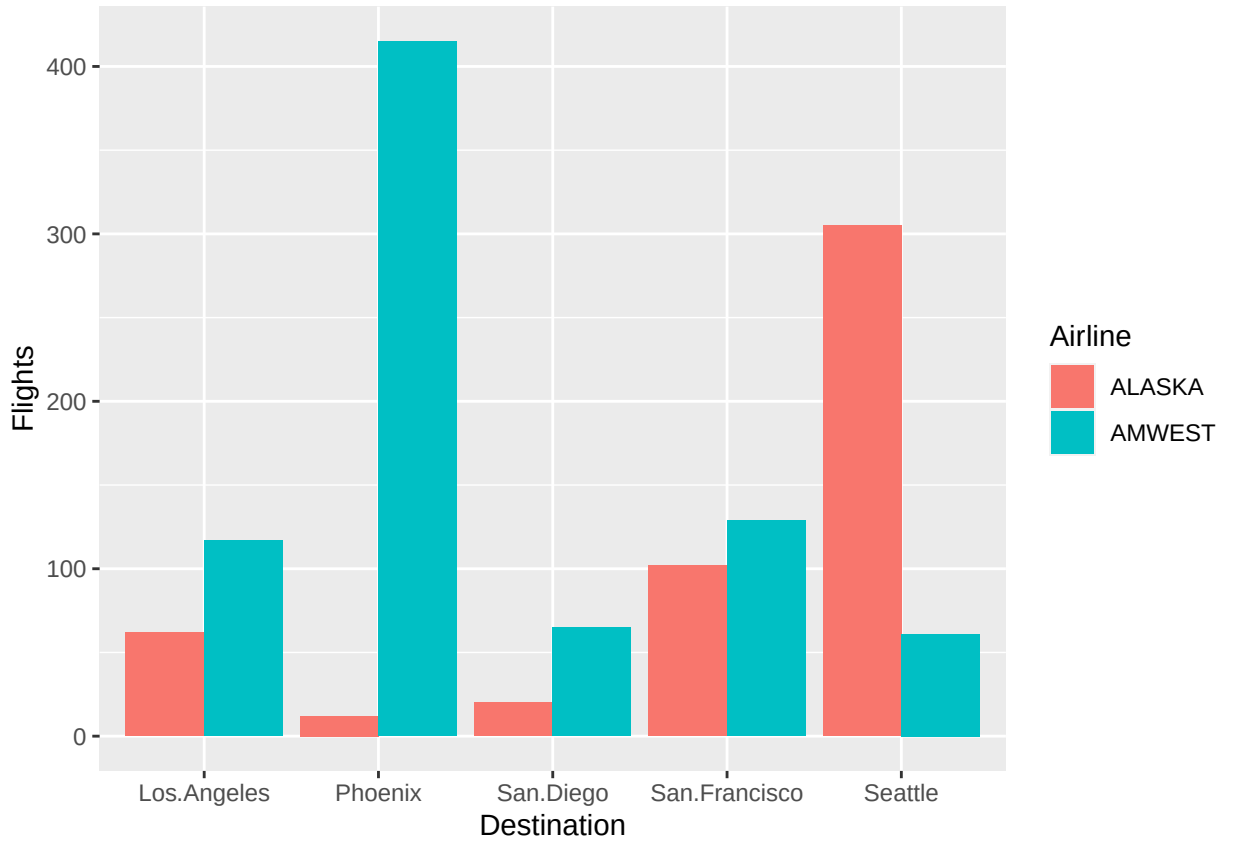
I tried to evaluate the data based on on-time flights, but realized there was a big difference in overall flight totals between both airlines

```
ggplot(data = on_time, aes(x = Destination, y = Flights, fill = Airline)) +
  geom_bar(position = "dodge", stat = "identity")
```




```
delayed <- y %>%
  filter(Delay_status == "delayed")
```

```
ggplot(data = delayed, aes(x = Destination, y = Flights, fill = Airline)) +
  geom_bar(position = "dodge", stat = "identity")
```



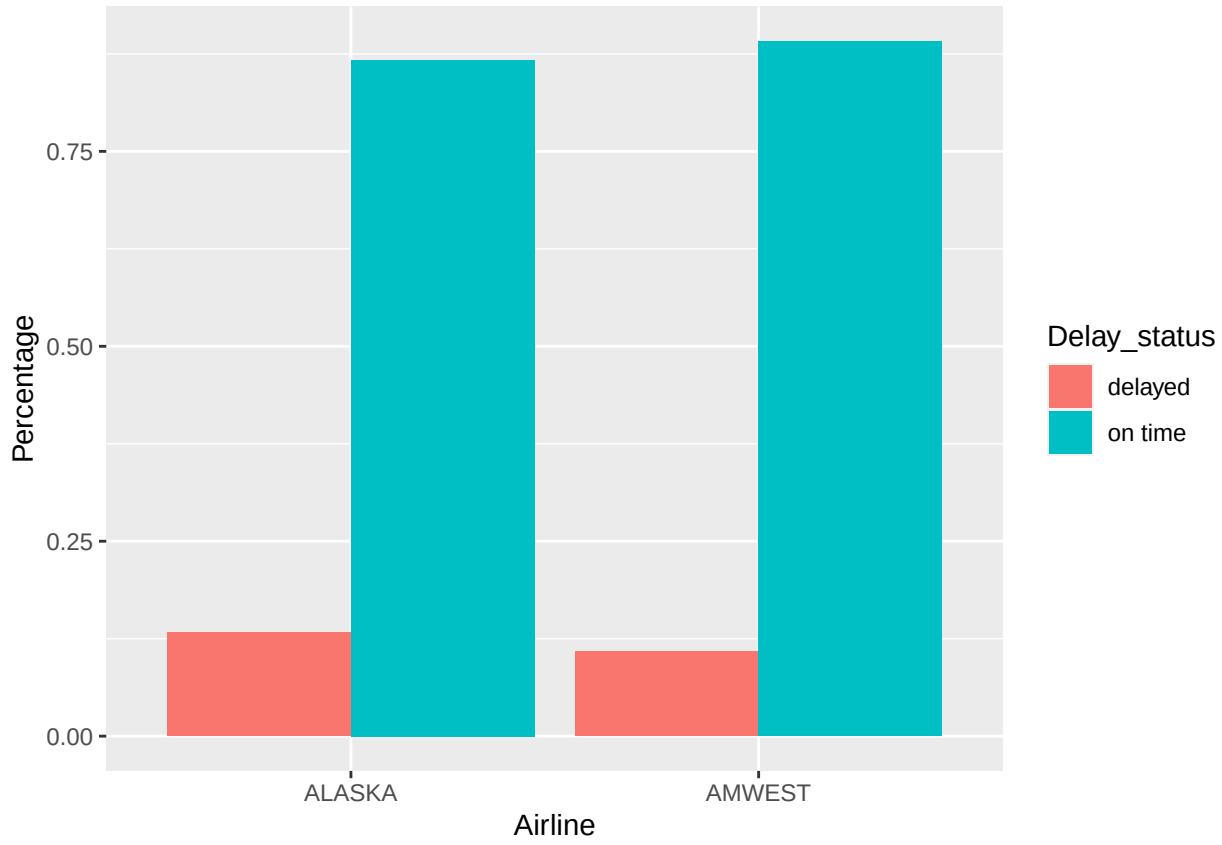
I decided to try to evaluate the data based on flight delay status percentages, I grouped by Airline and Delay status to get a total flight count, and then calculated the flight delay percentages based on those totals.

```
z <- y %>%
  group_by(Airline, Delay_status) %>%
  summarise(Flights_count = sum(Flights)) %>%
  ungroup() %>%
  group_by(Airline) %>%
  mutate(Total_flights = sum(Flights_count)) %>%
  mutate(Percentage = Flights_count/Total_flights)
```

```
## `summarise()` has grouped output by 'Airline'. You can override using the
## `.groups` argument.
```

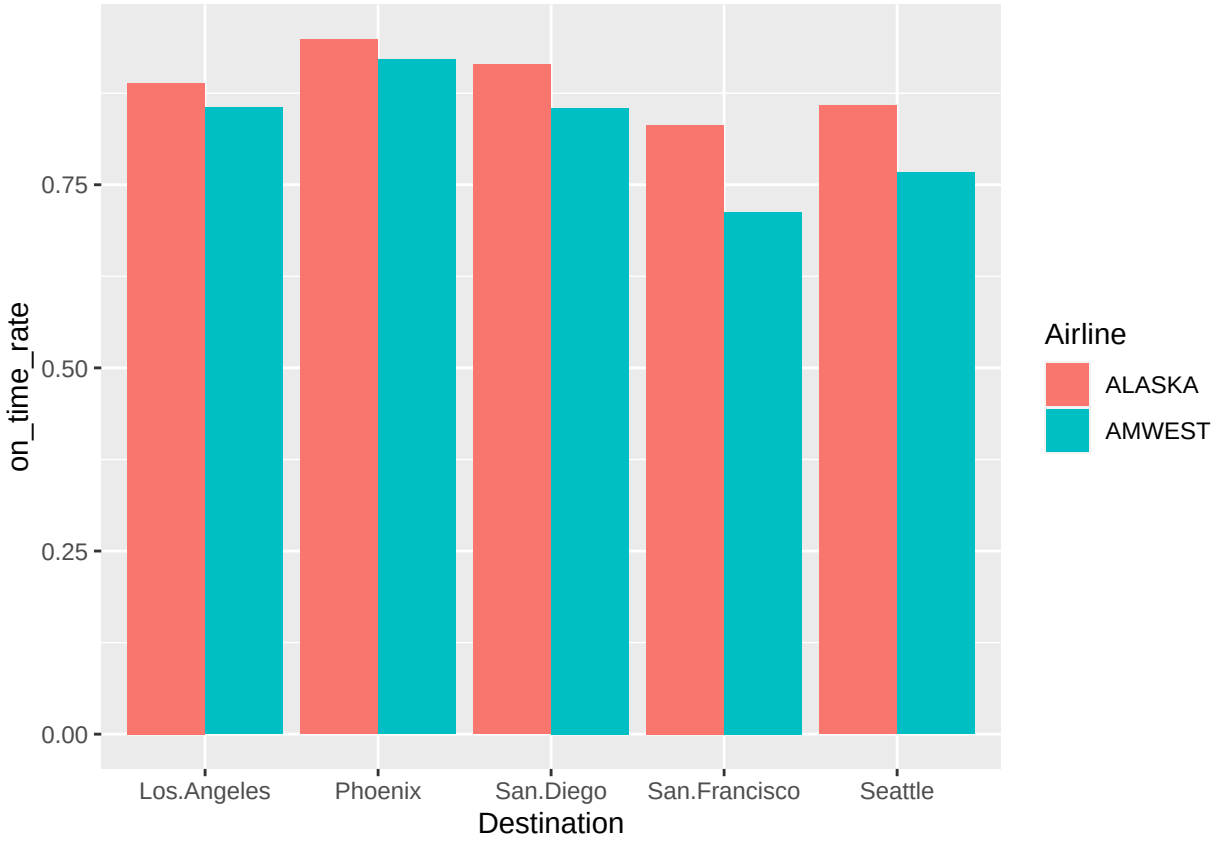
I created a bar graph to help visualize those percentages of on-time and delayed flights between both airlines

```
ggplot(data = z, aes(x = Airline, y = Percentage, fill = Delay_status)) +
  geom_bar(position = "dodge", stat = "identity")
```

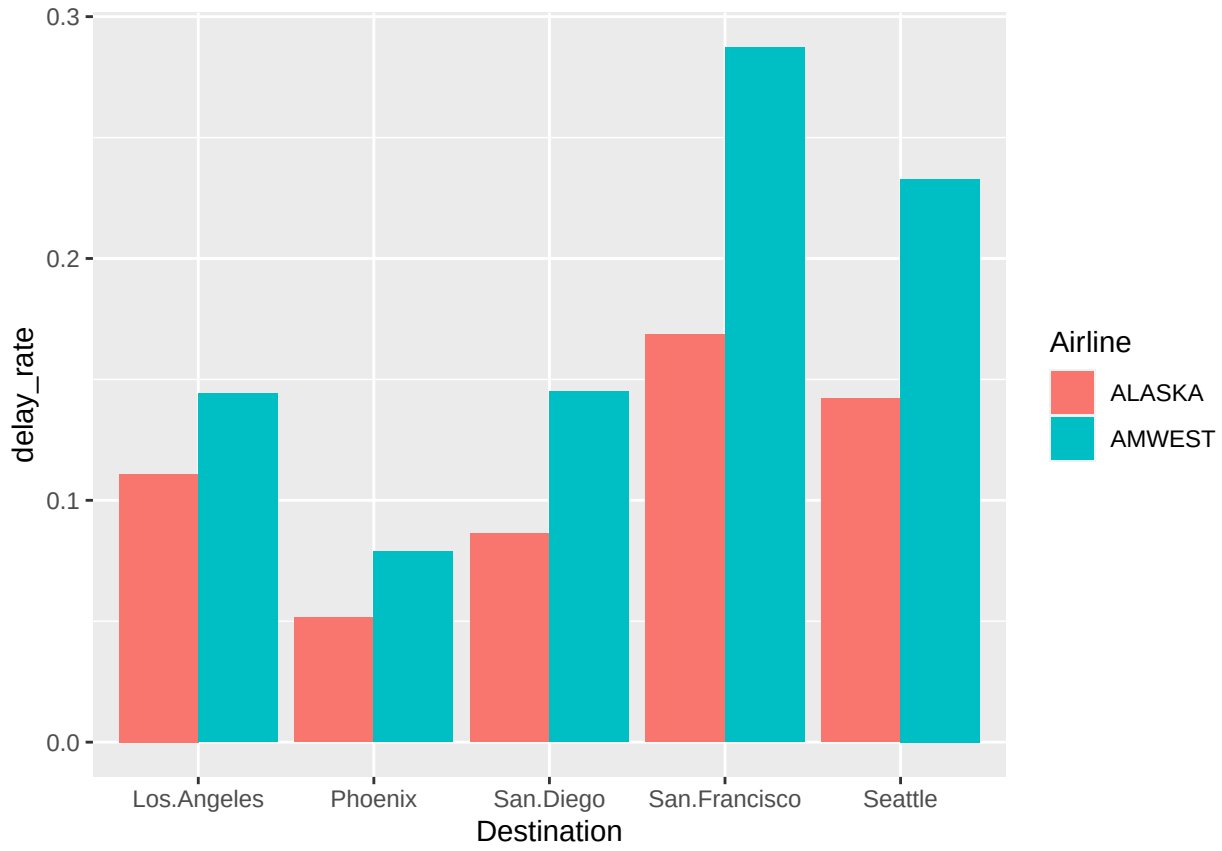



```
wider_y <- y %>%
  pivot_wider(names_from = Delay_status, values_from = Flights) %>%
   rename(on_time = `on time`) %>%
  group_by(Airline, Destination) %>%
  mutate(total_flights = sum(on_time, delayed)) %>%
  mutate(on_time_rate = on_time / total_flights) %>%
  mutate(delay_rate = delayed / total_flights)
```

```
ggplot(data = wider_y, aes(x = Destination, y = on_time_rate, fill = Airline )) +
  geom_bar(position = "dodge", stat = "identity")
```




```
ggplot(data = wider_y, aes(x = Destination, y = delay_rate, fill = Airline )) +
  geom_bar(position = "dodge", stat = "identity")
```



Conclusion:

Based on the plot showing the overall flights status of both airlines, it seems that AM west is a better airline to fly out of because they have more flights and an overall lower percentage of delayed flights and a higher percentage o online flights, however, when separating the data by destination from each airline, we see AM west has a higher percentage of delayed flights out of each destination than Alaska. So while they both tend to run on-time more than they are delayed, out of each destination, Alaska Airlines have a lower percentage of their flights being delayed than AM West.