Assignment 5

Matthew Tillmawitz

2024-09-25

Read in data

The data is read in from the csv that was created and can be viewed in the github repository this project resides in. Some filtering is done to remove empty rows in the csv and fill the airline name in.

```
raw_data <- read_csv("flights.csv")
marshal <- raw_data |>
  filter(!if_all(names(raw_data), ~ is.na(.))) |>
  fill(...1) |>
  rename(airline = ...1, status = ...2)
marshal
```

```
## # A tibble: 4 x 7
##   airline status  `Los Angeles` Phoenix `San Diego` `San Francisco` Seattle
##   <chr>   <chr>           <dbl>   <dbl>       <dbl>           <dbl>   <dbl>
## 1 ALASKA  on time           497     221         212             503    1841
## 2 ALASKA  delayed           62      12          20             102     305
## 3 AM WEST on time           694    4840         383             320     201
## 4 AM WEST delayed           117     415          65             129      61
```

Cleaning Up the Data

In order to make analysis easier, the destinations are collapsed into a single column named "destination" and the number of flights in each cell is mapped to a column named "flights". The "status" column is then broken out into two columns, "on_time" and "delayed" with the value of the "flights" column mapped to the corresponding status. This data format makes calculating the rate of delayed flights simple for each airline and destination. Delayed percentage was chosen as it will be easier to see differences between airlines and destinations when plotting the data due to most flights being on time.

```
pivoted <- marshal |>
  pivot_longer(
    cols = `Los Angeles`:Seattle,
    names_to = "destination",
    values_to = "flights"
  ) |>
  pivot_wider(
    names_from = status,
    values_from = flights
  ) |>
  rename(on_time = `on time`) |>
```

```
  mutate(total_flights = on_time + delayed, delayed_percentage = delayed / total_flights)

pivoted
```

```
## # A tibble: 10 x 6
##    airline destination   on_time delayed total_flights delayed_percentage
##    <chr>   <chr>           <dbl>   <dbl>         <dbl>              <dbl>
##  1 ALASKA  Los Angeles       497      62           559              0.111
##  2 ALASKA  Phoenix           221      12           233              0.0515
##  3 ALASKA  San Diego         212      20           232              0.0862
##  4 ALASKA  San Francisco     503     102           605              0.169
##  5 ALASKA  Seattle          1841     305          2146              0.142
##  6 AM WEST Los Angeles       694     117           811              0.144
##  7 AM WEST Phoenix          4840     415          5255              0.0790
##  8 AM WEST San Diego         383      65           448              0.145
##  9 AM WEST San Francisco     320     129           449              0.287
## 10 AM WEST Seattle           201      61           262              0.233
```

Plotting the Data

Aggregating the on time and delayed flights by airline allows us to get an idea of how the airlines compare overall. The difference in overall on time rates between the two airlines is relatively small at just 2%, but worth noting is that Am West has almost double the total number of flights and a lower overall delayed percentage, initially indicating strong performance by the airline.

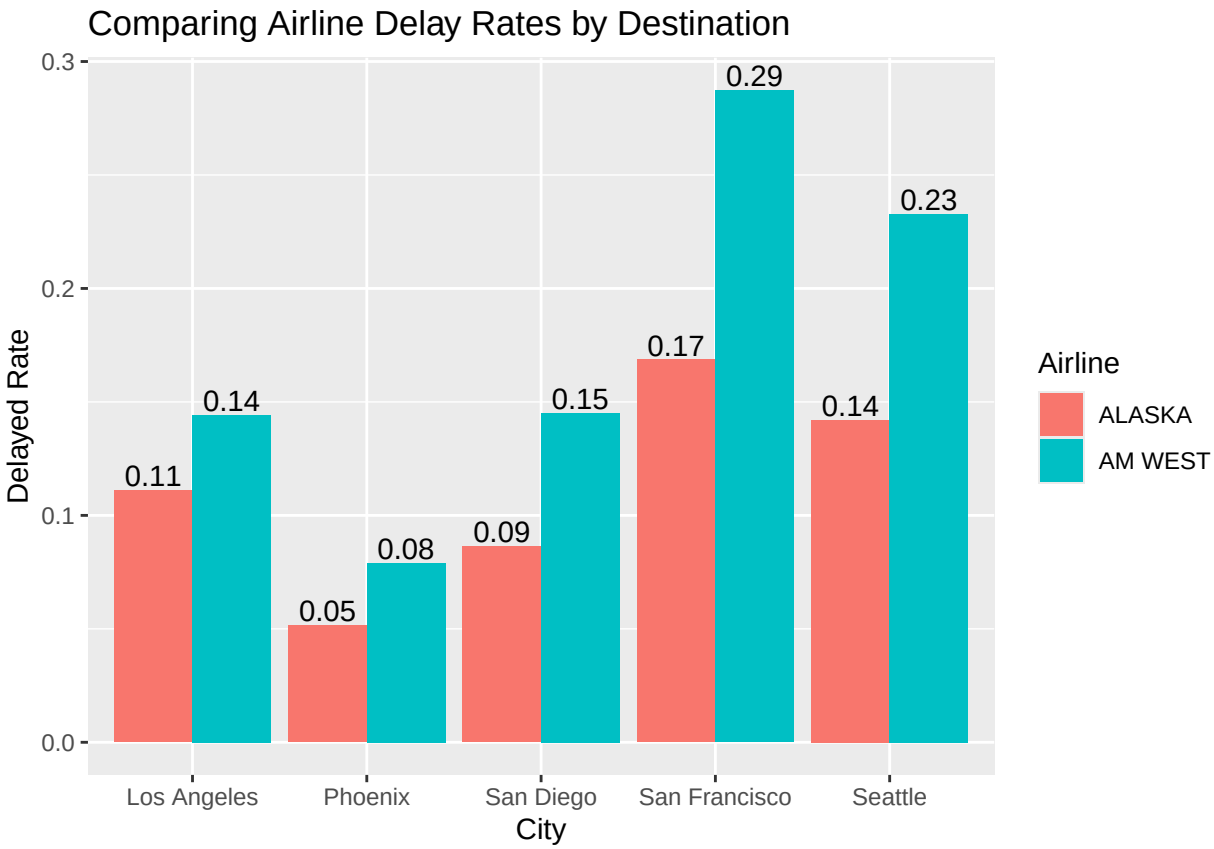
```
agg_airline <- pivoted |>
  group_by(airline) |>
  summarise(total_on_time = sum(on_time), total_delayed = sum(delayed)) |>
  mutate(total_flights = total_on_time + total_delayed, delayed_percentage = total_delayed / (total_dela

agg_airline
```

```
## # A tibble: 2 x 5
##   airline total_on_time total_delayed total_flights delayed_percentage
##   <chr>           <dbl>         <dbl>         <dbl>              <dbl>
## 1 ALASKA           3274           501          3775              0.133
## 2 AM WEST          6438           787          7225              0.109
```

By plotting the rate of delayed flights by destination the data starts to become more interesting. It can be seen that Alaska has a consistently higher rate of on time flights for each individual destination. The difference is most notable for San Francisco and Seattle, with both having around a 10% difference in the rate of on time flights. This appears to run counter to our initial analysis that Am West was the better performing airline. This indicates that Alaska is actually the better performing airline.

```
pivoted |>
  ggplot(aes(x = destination, y = delayed_percentage, fill = airline)) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(aes(label = round(delayed_percentage,2)), vjust = -0.2, position = position_dodge(width = .9
  labs(x = "City", y = "Delayed Rate", fill = "Airline", title = "Comparing Airline Delay Rates by Desti
```


Comparing Airline Delay Rates by Destination
Delayed Rate
0.3
0.2
0.1
0.0
0.11
0.14
0.05
0.08
0.09
0.15
0.17
0.29
0.14
0.23
Los Angeles
Phoenix
San Diego
San Francisco
Seattle
City
Airline
ALASKA
AM WEST