Assignment 5

Alice Ding

2023-02-21

Overview

This data set includes information from two airlines: ALASKA and AM WEST. These two airlines have 6 destinations that they operate with and this data includes a count of flights that were on time and delayed for each of the 6 cities.

```
data <- read.csv("https://raw.githubusercontent.com/addsding/data607/main/assignment5/data.csv")
head(data)
```

```
##   Airline Destination On.Time Delayed
## 1  ALASKA Los Angeles     497      62
## 2 AM WEST Los Angeles     694     117
## 3  ALASKA     Phoenix     221      12
## 4 AM WEST     Phoenix    4840     415
## 5  ALASKA   San Diego     212      20
## 6 AM WEST   San Diego     383      65
```

Tidying the Data

Tidy data is data where:

- Every columns is variable
- Every row is an observation
- Every cell is a single value

We see that to start with, every row here contains two observations: `On Time` and `Delayed`. We first want to separate this into two rows in order to have clean data.

```
data_pivoted <- pivot_longer(data, cols=3:4, names_to="arrival_type", values_to="flight_count")
head(data_pivoted)
```

```
## # A tibble: 6 x 4
##   Airline Destination arrival_type flight_count
##   <chr>   <chr>       <chr>               <int>
## 1 ALASKA  Los Angeles On.Time               497
## 2 ALASKA  Los Angeles Delayed                62
## 3 AM WEST Los Angeles On.Time               694
## 4 AM WEST Los Angeles Delayed              117
## 5 ALASKA  Phoenix     On.Time               221
## 6 ALASKA  Phoenix     Delayed                12
```

Looks good! Next, it looks like there's a `.` in `On.Time` – I'll replace those with a space instead.

```
data_pivoted$arrival_type <- gsub("\\.", " ", data_pivoted$arrival_type)
head(data_pivoted)
```

```
## # A tibble: 6 x 4
##   Airline Destination arrival_type flight_count
##   <chr>   <chr>       <chr>               <int>
## 1 ALASKA  Los Angeles On Time               497
## 2 ALASKA  Los Angeles Delayed                62
## 3 AM WEST Los Angeles On Time               694
## 4 AM WEST Los Angeles Delayed               117
## 5 ALASKA  Phoenix     On Time               221
## 6 ALASKA  Phoenix     Delayed                12
```

The data is now clean for analysis!

Analysis

First, let's see just nominally how many flights were delayed for each airline.

```
total_delayed <- data_pivoted |>
  group_by(Airline) |>
  summarise(total_delayed = sum(flight_count[arrival_type=="Delayed"]), .groups = 'drop') |>
  arrange(total_delayed)

head(total_delayed)
```

```
## # A tibble: 2 x 2
##   Airline total_delayed
##   <chr>           <int>
## 1 ALASKA            501
## 2 AM WEST           787
```

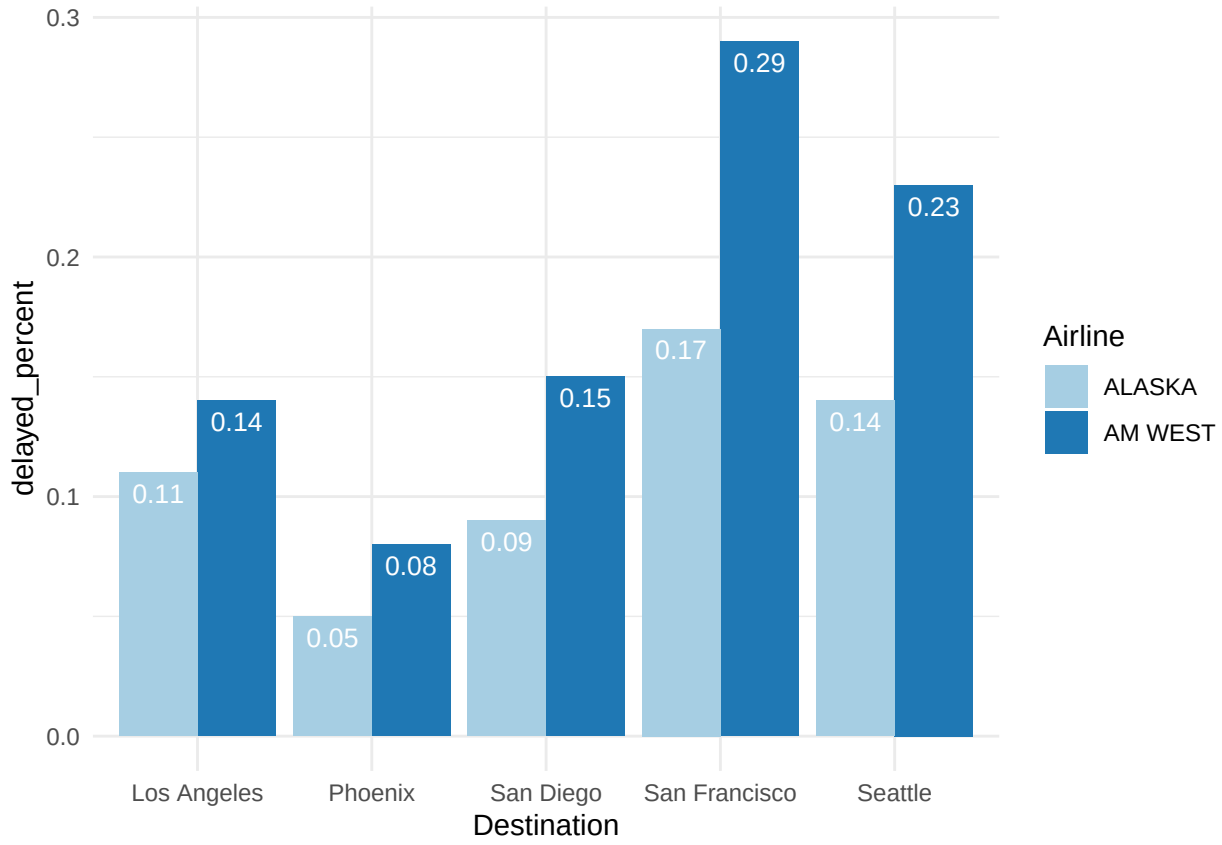
We can see here that nominally, AM WEST has more delays. What percent of flights is that though?

```
delayed_percentage <- data_pivoted |>
  group_by(Airline) |>
  summarise(total_delayed = sum(flight_count[arrival_type=="Delayed"]),
            total_flights = sum(flight_count),
            delayed_percent = sum(flight_count[arrival_type=="Delayed"])
              /sum(flight_count),
            .groups = 'drop') |>
  arrange(total_delayed, total_flights, delayed_percent)

head(delayed_percentage)
```

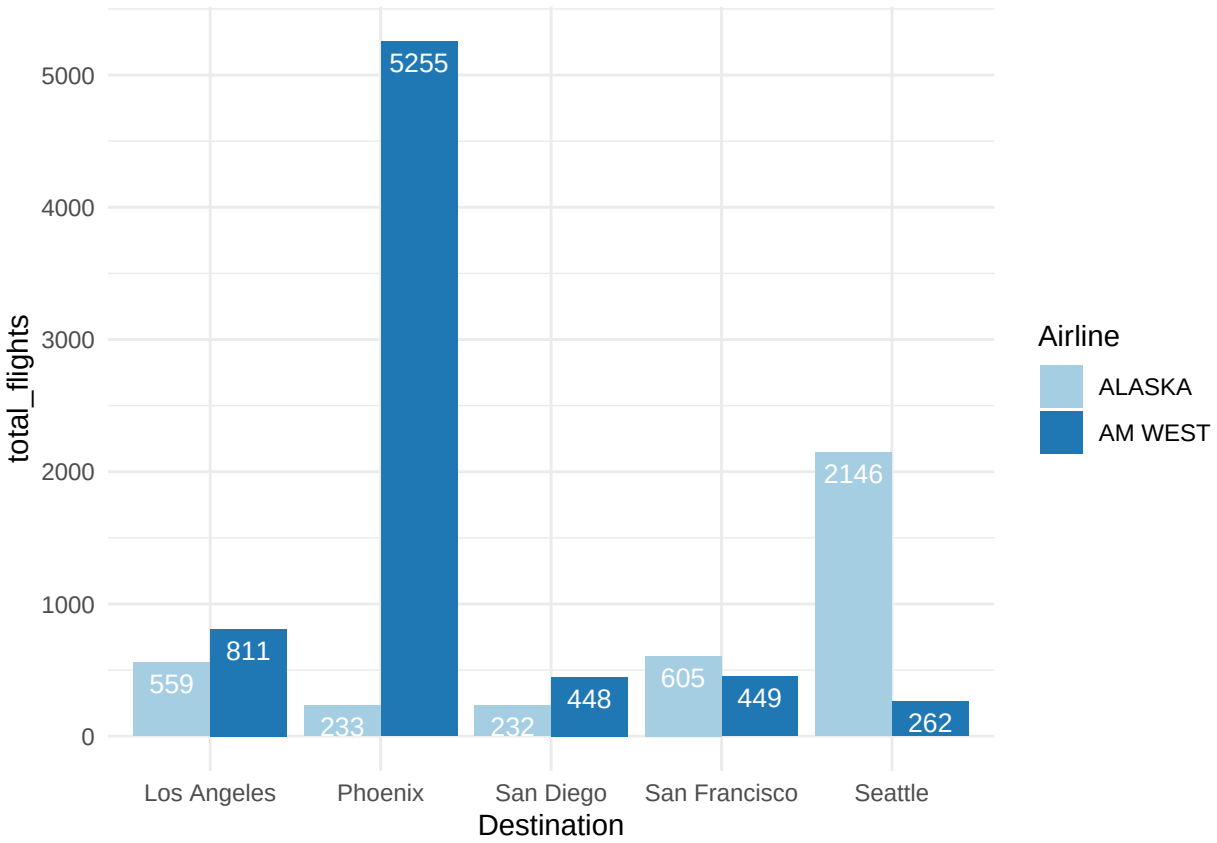
```
## # A tibble: 2 x 4
##   Airline total_delayed total_flights delayed_percent
##   <chr>           <int>         <int>           <dbl>
## 1 ALASKA            501          3775           0.133
## 2 AM WEST           787          7225           0.109
```

Despite having more delays nominally, ALASKA actually has a higher delay rate (13.3% vs. 10.9%). Do we see any city that perhaps is driving up delays for both or either of the airlines?

```
by_city <- data_pivoted |>
  group_by(Airline, Destination) |>
  summarise(total_delayed = sum(flight_count[arrival_type=="Delayed"]),
            total_flights = sum(flight_count),
            delayed_percent = sum(flight_count[arrival_type=="Delayed"])
              /sum(flight_count),
            .groups = 'drop') |>
  arrange(desc(delayed_percent), total_delayed, total_flights)

# round delayed_percent
by_city <- by_city |> mutate(across(delayed_percent, round, 2))

# bar chart for delay rates
delay_bar <-ggplot(data=by_city
                   , aes(x=Destination
                         , y=delayed_percent
                         , fill=Airline)) +
  geom_bar(stat="identity", position=position_dodge()) +
  geom_text(aes(label=delayed_percent), vjust=1.6, color="white",
            position = position_dodge(0.9), size=3.5)+
  scale_fill_brewer(palette="Paired")+
  theme_minimal()

# bar chart for counts of flights
flight_bar <-ggplot(data=by_city
                   , aes(x=Destination
                         , y=total_flights
                         , fill=Airline)) +
  geom_bar(stat="identity", position=position_dodge()) +
  geom_text(aes(label=total_flights), vjust=1.6, color="white",
            position = position_dodge(0.9), size=3.5)+
  scale_fill_brewer(palette="Paired")+
  theme_minimal()


delay_bar
```


0.3
0.2
0.1
0.0
delayed_percent
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
Destination
Airline
ALASKA
AM WEST

```
flight_bar
```



We can see here that for AM WEST, San Francisco and Seattle are huge outliers with delay rates higher than 20%. The average for AM WEST is 10% though – it seems like Phoenix really brings the rate down as only 8% of flights to that destination are delayed and a vast majority of flights go there (5k).

For ALASKA, San Francisco is the most extreme at 17% and the only other one above the average delay rate (13.3%) is Seattle at 14%. ALASKA's flights mainly seem to go to Seattle (2k) so the overall average being close to Seattle's average makes sense.

Overall, AM WEST seems to have a wider range of delay rates across the different destinations (min of 8, max of 29) while ALASKA's numbers are less diverse (min of 5, max of 17). The variability of AM WEST seems to rely heavily on what cities the flight is going to and it may be due to less experience flying to that region as a majority of flights go towards Phoenix and that's its least delayed destination. ALASKA on the other hand still has Seattle (its most popular destination) as its second most delayed flight which signals that there probably be some sort of improvement. Both airlines seem to struggle the most with Seattle and San Francisco.

In the end, AM WEST is more delayed for every destination – if I were to pick one of these airlines, I would likely pick AM WEST for less of a chance of being delayed.

Conclusion

The tidying and analysis of this data using `tidyr` and `dplyr` was overall made very simple with how extensive and easy-to-use each library is. For future analysis ideas, having this data broken down by any sort of time period (weeks, months, years, etc.) or adding weather conditions could help narrow down as to why certain flights were delayed or not.