Week 4 Assignment Resubmit

Jiaxin Zheng

2024-09-27

```
library(readr)
library(tidyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

```
#read the CSV file
data <- read_csv("https://raw.githubusercontent.com/Jennyjjxxzz/Data607_Week4Assignment-/refs/heads/mai
```

```
## New names:
## Rows: 5 Columns: 7
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (2): ...1, ...2 dbl (5): Los Angeles, Phoenix, San Diego, San Francisco,
## Seattle
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
## * `` -> `...2`
```

```
#rename the columns
colnames(data) <- c("Airline", "Status", "Los Angeles", "Phoenix", "San Diego", "San Francisco", "Seattl

#try to fill in the missing values in column 1 "Airline"
data <- data %>%
  fill(Airline, .direction = "down")

#remove one row
data <- data[-c(3),]
```

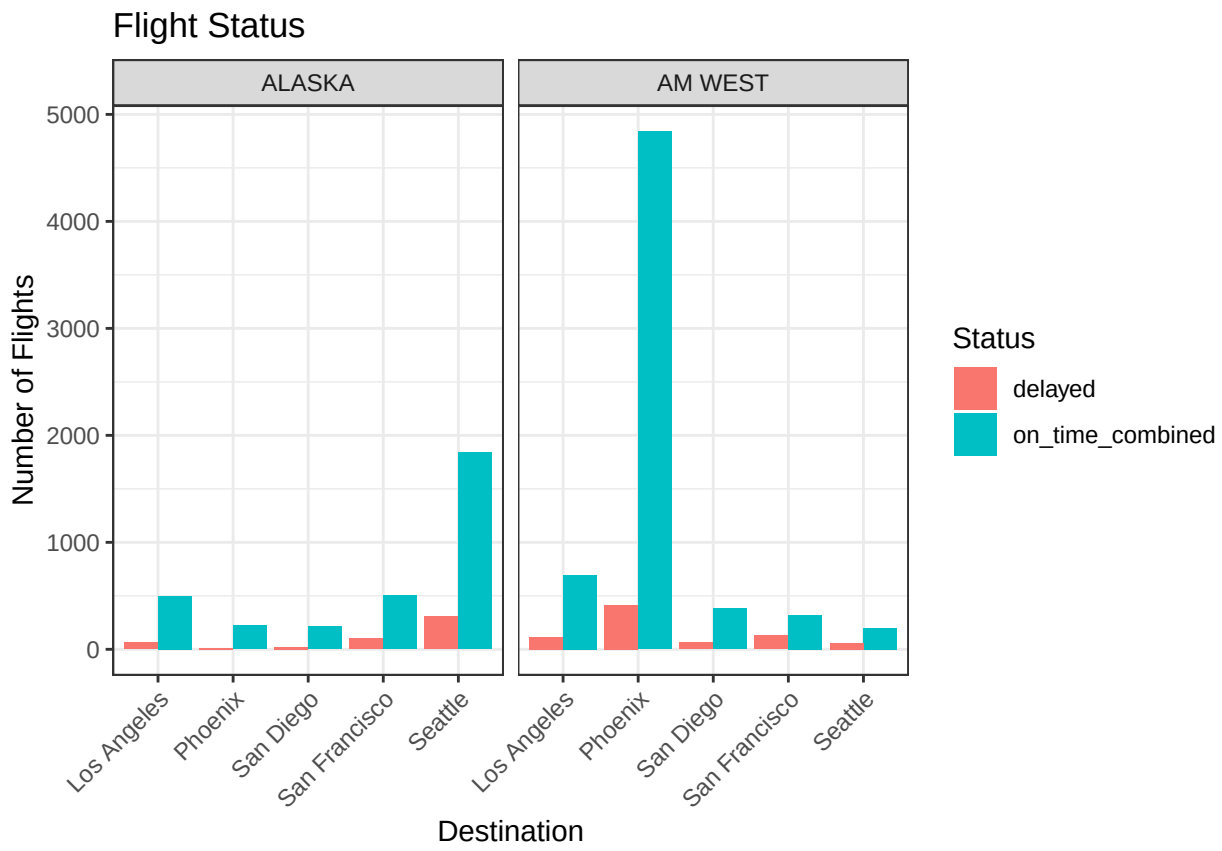
```r
data_longer <- data |>
  pivot_longer(cols = 3:7,
               names_to = "Destination",
               values_to = "Count",
               values_drop_na = TRUE)
```

```r
data_wider <- data_longer |>
  pivot_wider(names_from = Status,
              values_from = Count)
colnames(data_wider) <- c("Airline", "Destination", "On_Time_1", "delayed", "On_Time_2")
```

```r
#combine the "on time" column
data_wider <- data_wider |>
  mutate(on_time_combined = coalesce(On_Time_1, On_Time_2))
```

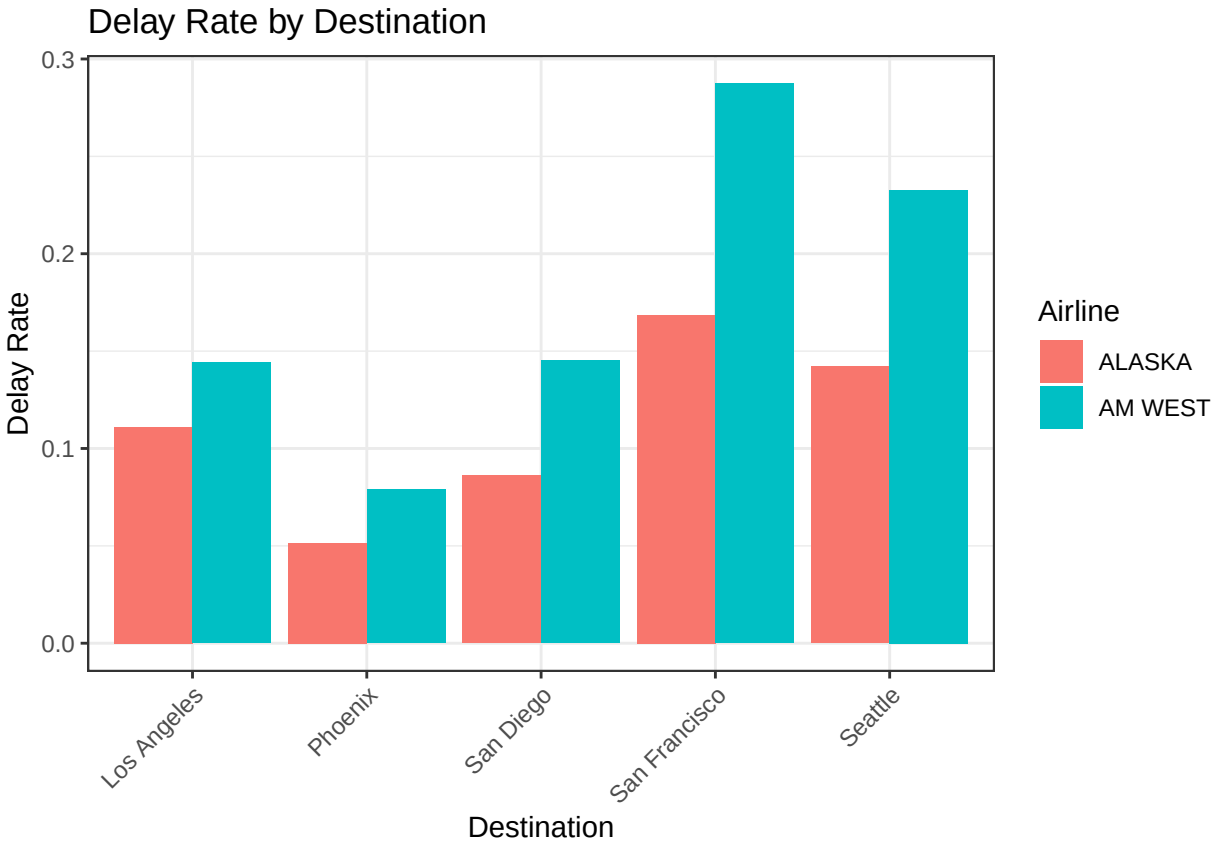
```r
data_plot <- data_wider |>
  select(Airline, Destination, on_time_combined, delayed) |>
  pivot_longer(cols = c(on_time_combined, delayed),
               names_to = "Status",
               values_to = "Count")

ggplot(data_plot, aes(x = Destination, y = Count, fill = Status)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_wrap(~ Airline) +
  labs(title = "Flight Status",
       x = "Destination",
       y = "Number of Flights",
       fill = "Status") +
  theme_bw()+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
# Calculate delay rate
data_wider <- data_wider |>
  mutate(total_flights = on_time_combined + delayed,
         delay_rate = delayed / total_flights)


ggplot(data_wider, aes(x = Destination, y = delay_rate, fill = Airline)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Delay Rate by Destination",
       x = "Destination",
       y = "Delay Rate",
       fill = "Airline") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



In conclusion: As we can see the plots depend on the delay rate for each city. Alaska Air line has lower delay percentages in each city than American West. But the overall plot(first plot), American West in Phoenix holds lowest delay rate.