data607_assignment5A

Joao De Oliveira

2025-09-27

Overview

This report uses the charted airline arrival delays in 5 cities. The provided data was used to create a CSV file with that data. I tidied into a long format and analyzed the data by comparing the percentage of delayed flights per city and among all cities. I created a table and plots to visualize these comparisons.

```r
library(tidyverse)
```

Load Libraries and Data

```
## -- 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   3.5.2     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- 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 error
```

```r
library(readr)

# load data
raw_url <- "https://raw.githubusercontent.com/JDO-MSDS/DATA-607/refs/heads/main/Assignment5A/arrivals%2

wide_data <- readr::read_csv(raw_url, show_col_types = FALSE)
```

```
## New names:
## * `` -> `...1`
## * `` -> `...2`
```

```r
wide_data
```

```
## # A tibble: 5 x 7
##   ...1    ...2    `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 <NA>    delayed          62     12         20             102     305
## 3 <NA>    <NA>             NA     NA         NA              NA      NA
## 4 AM WEST on time         694   4840        383             320     201
## 5 <NA>    delayed         117    415         65             129      61
```

```
nm <- names(wide_data)
stopifnot(length(nm) >= 2)

df <- wide_data %>%
  setNames(replace(nm, 1:2, c("airline", "status"))) %>%
  tidyr::fill(airline, .direction = "down") %>%
  # get rid of the empty row
  filter(!(is.na(status) & if_all(-c(airline, status), is.na))) %>%
  mutate(status = tolower(status),
         status = ifelse(status %in% c("on time","on-time"), "on_time", status)
  )

# Reformat to long format
city_columns <- setdiff(names(df), c("airline","status"))
long <- df %>%
  pivot_longer(all_of(city_columns), names_to = "city", values_to = "count") %>%
  mutate(count = as.numeric(count))

# percentage of delayed flights per city
per_city <- long %>%
  group_by(airline, city, status) %>%
  summarise(count = sum(count, na.rm = TRUE), .groups = "drop") %>%
  tidyr::pivot_wider(names_from = status, values_from = count, values_fill = 0) %>%
  mutate(
    total_city = delayed + on_time,
    pct_delayed = ifelse(total_city > 0, 100 * delayed / total_city, NA_real_)
    )

# percentage of total delayed flights
all_city <- per_city %>%
  group_by(airline) %>%
  summarise(
    delayed = sum(delayed), on_time = sum(on_time),
    total = delayed + on_time,
    pct_delayed = 100 * delayed / total,
    .groups = "drop"
  )

# tables
print(per_city %>% select(city, airline, pct_delayed) %>% arrange(city, airline))
```

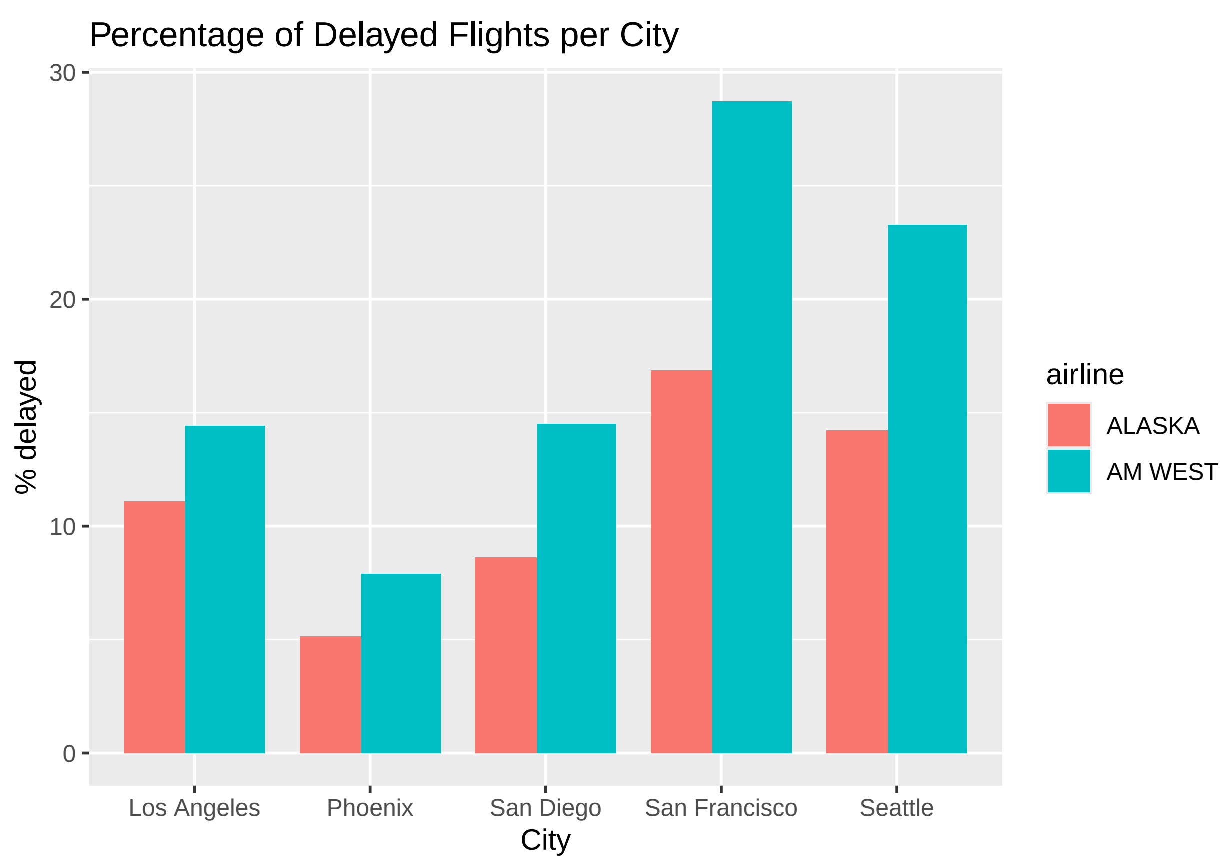
```
## # A tibble: 10 x 3
##    city          airline pct_delayed
##    <chr>         <chr>         <dbl>
##  1 Los Angeles   ALASKA        11.1
##  2 Los Angeles   AM WEST       14.4
##  3 Phoenix       ALASKA         5.15
##  4 Phoenix       AM WEST        7.90
```

```
##  5 San Diego     ALASKA         8.62
##  6 San Diego     AM WEST       14.5
##  7 San Francisco ALASKA        16.9
##  8 San Francisco AM WEST       28.7
##  9 Seattle       ALASKA        14.2
## 10 Seattle       AM WEST       23.3
```

```
print(all_city %>% select(airline, pct_delayed) %>% arrange(airline))
```

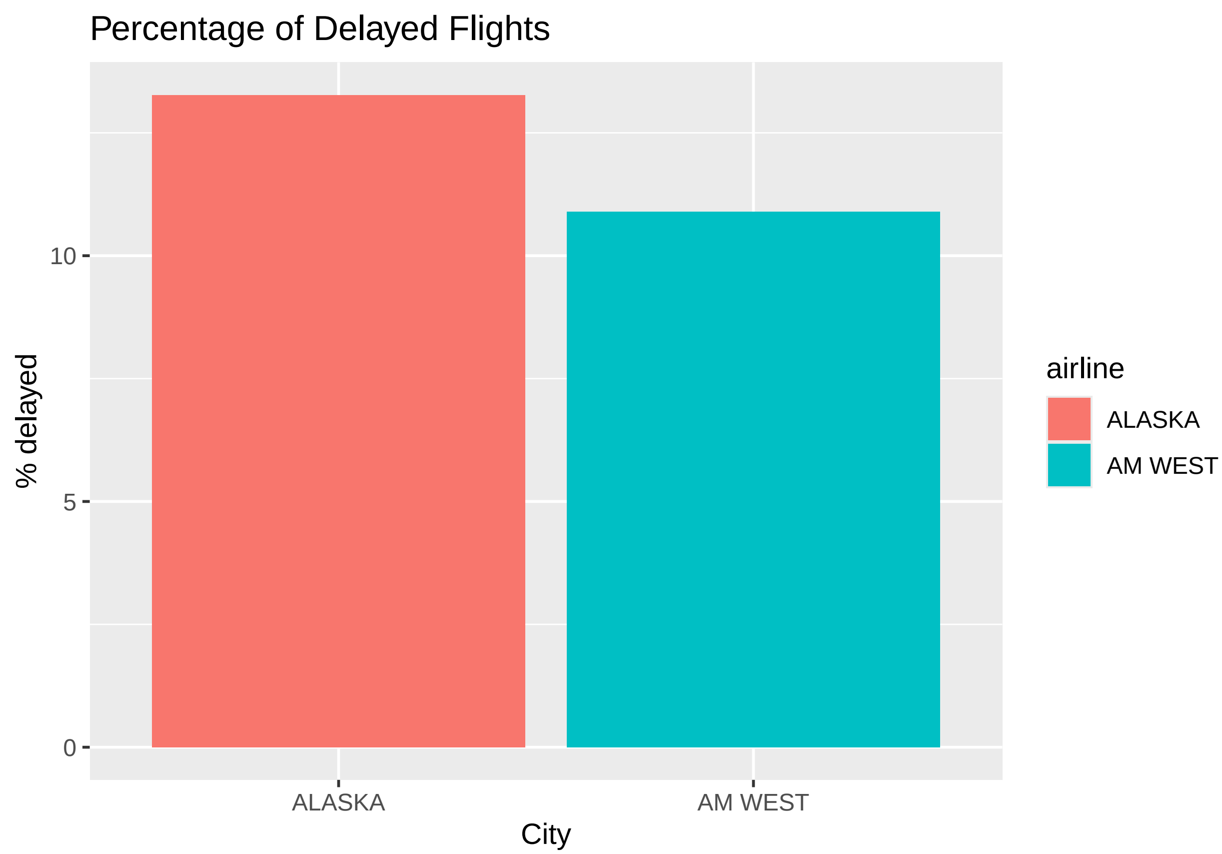
```
## # A tibble: 2 x 2
##   airline pct_delayed
##   <chr>         <dbl>
## 1 ALASKA         13.3
## 2 AM WEST        10.9
```

```
ggplot(per_city, aes(x = city, y = pct_delayed, fill = airline)) +
  geom_col(position = position_dodge(width = 0.7)) +
  labs(title = "Percentage of Delayed Flights per City", x = "City", y = "% delayed")
```



Plots

```
ggplot(all_city, aes(airline, pct_delayed, fill = airline)) +
  geom_col() +
  labs(title = "Percentage of Delayed Flights", x = "City", y = "% delayed")
```



Conclusion

By looking at the plot with the percentage of delayed flights per city, we can see that AM WEST has a higher percentage of delayed flights in every single city. However, the plot representing the percentage of delayed flights in all cities combined shows that ALASKA has an overall higher percentage of delayed flights when compared with AM WEST. This discrepancy comes from the fact that the percentage of combined cities might be impacted by a lot of flights in a single airport that usually has more delayed flights.