Version 2 City by city Delay comparison

Pricilla

2025-10-09

Overview

The assignment is tidying and transforming data.

Loading the Libraries

```
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   3.5.2     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
## -- 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
```

```
library(tidyr)
library(dplyr)
library(ggplot2)
```

Read the Data

(1) Create a .CSV file (or optionally, a MySQL database!) that includes all of the information above. You're encouraged to use a "wide" structure similar to how the information appears above, so that you can practice tidying and transformations as described below.

I created a CSV in Github

```
Flightdata <- read.csv("https://raw.githubusercontent.com/prnakyazze94/Data_607/refs/heads/main/Airlinel
print(Flightdata)
```

```
##         X     X.1 Los.Angeles Phoenix San.Diego San.Francisco Seattle
## 1  ALASKA on time         497     221       212           503    1841
## 2         delayed          62      12        20           102     305
```

```
## 3                          NA      NA        NA            NA      NA
## 4 AM WEST on time         694    4840       383           320     201
## 5         delayed         117     415        65           129      61
```

(2) Read AirlineData.CSV file into R, and use tidyr and dplyr as needed to tidy and transform your data.

Assign header names to columns X and X.1 columns.

```
names(Flightdata) = c("Airline", "On_time_Delayed", "Los Angeles", "Phoenix", "San Diego", "San Francis
print(Flightdata)
```

```
##   Airline On_time_Delayed Los Angeles Phoenix San Diego San Francisco Seattle
## 1  ALASKA         on time         497     221       212           503    1841
## 2                 delayed          62      12        20           102     305
## 3                                  NA      NA        NA            NA      NA
## 4 AM WEST         on time         694    4840       383           320     201
## 5                 delayed         117     415        65           129      61
```

Fill in Airline name for delayed rows.

```
Flightdata[2,1] = "ALASKA"
Flightdata[5, 1] = "AM WEST"
print(Flightdata)
```

```
##   Airline On_time_Delayed Los Angeles Phoenix San Diego San Francisco Seattle
## 1  ALASKA         on time         497     221       212           503    1841
## 2  ALASKA         delayed          62      12        20           102     305
## 3                                  NA      NA        NA            NA      NA
## 4 AM WEST         on time         694    4840       383           320     201
## 5 AM WEST         delayed         117     415        65           129      61
```

(3) Perform analysis to compare the arrival delays for the two airlines

Fill in NUll Values in Airline and On_time_Delayed with NA so it's possible to do numeric calculations. I used position of values but there should be a better way incase there is a lot of data to handle.

```
Flightdata[3, 1] <- NA
Flightdata[3, 2] <- NA
```

Summarize total on time vs delayed for each airline

Alaska Airlines

On time: 497 + 221 + 212 + 503 + 1841 = 3,274 flights

Delayed: 62 + 12 + 20 + 102 + 305 = 501 flights

Delay rate = 501 ÷ (3274 + 501) is 13.3%

AM West Airlines

On time: 694 + 4840 + 383 + 320 + 201 = 6,438 flights

Delayed: 117 + 415 + 65 + 129 + 61 = 787 flights

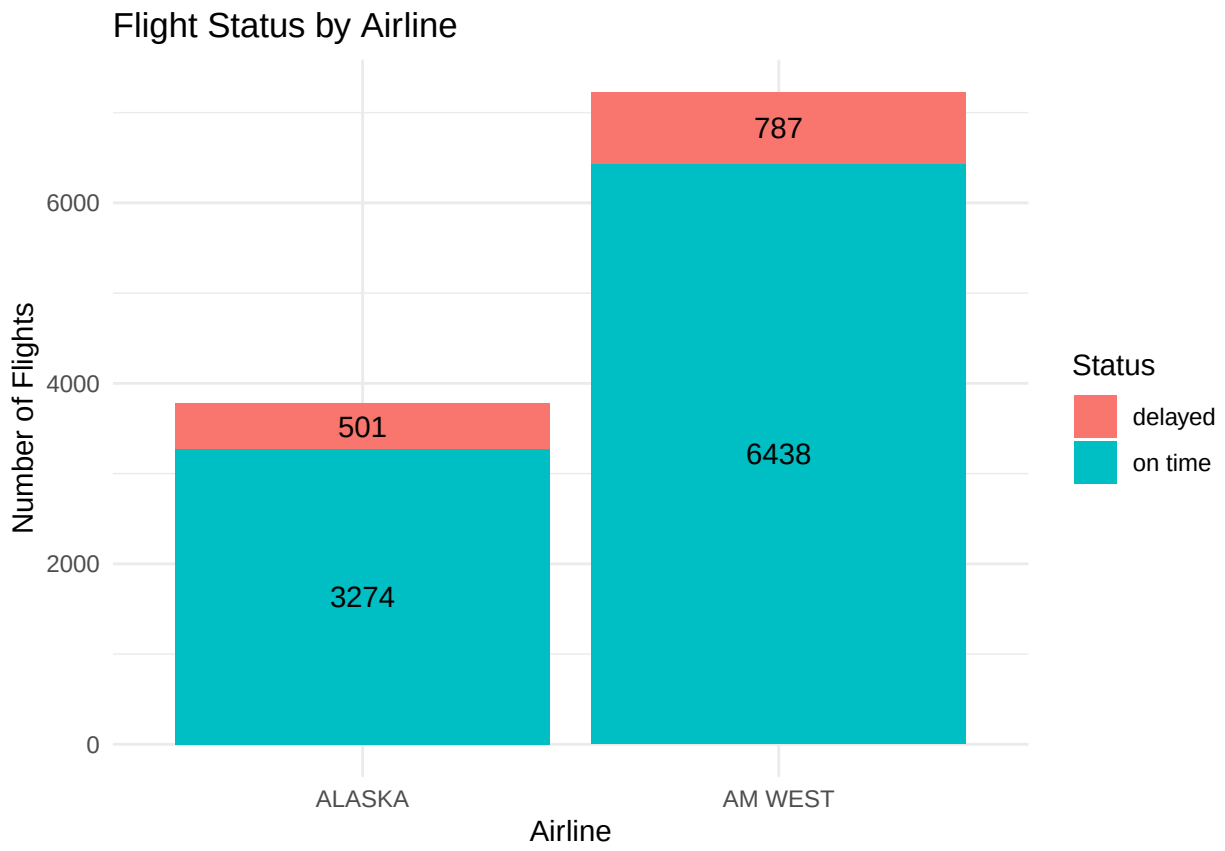
Delay rate = 787 ÷ (6438 + 787) is 10.9%

```r
# First, remove completely empty rows
Flightdata <- Flightdata %>%
  filter(!(is.na(Airline) & is.na(On_time_Delayed)))

# Then calculate summary
summary_df <- Flightdata %>%
  rowwise() %>%
  mutate(Total = sum(c_across(where(is.numeric)), na.rm = TRUE)) %>%
  ungroup() %>%
  select(Airline, On_time_Delayed, Total) %>%
  pivot_wider(
    names_from = On_time_Delayed,
    values_from = Total
  ) %>%
  mutate(
    Total_Flights = `on time` + delayed,
    Delay_Rate = round(delayed / Total_Flights * 100, 1),
    On_time_Performance = round(`on time` / Total_Flights * 100, 1)
  )

print(summary_df)
```

```
## # A tibble: 2 x 6
##   Airline `on time` delayed Total_Flights Delay_Rate On_time_Performance
##   <chr>       <int>   <int>         <int>      <dbl>               <dbl>
## 1 ALASKA       3274     501          3775       13.3                86.7
## 2 AM WEST      6438     787          7225       10.9                89.1
```

Plot on-time vs delayed as stacked bar chart

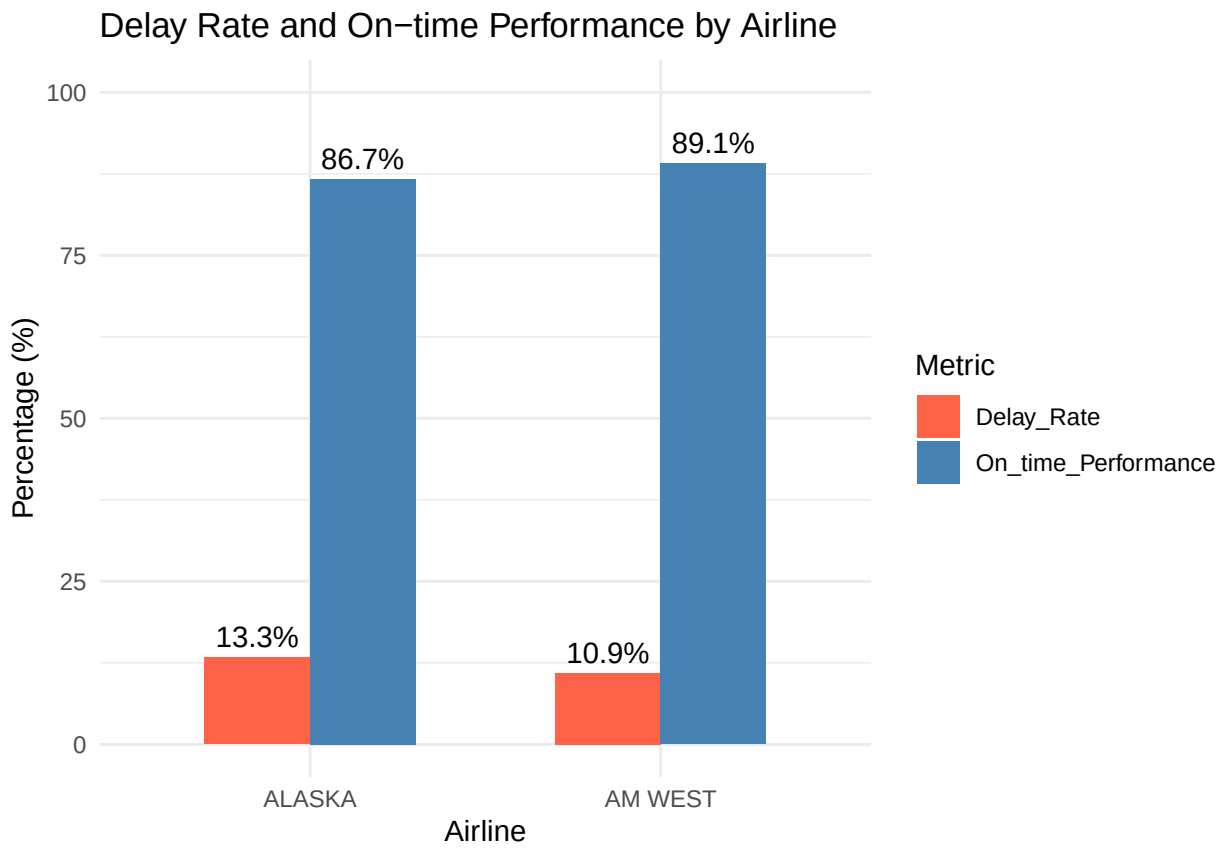
```r
summary_long <- summary_df %>%
  pivot_longer(cols = c("on time", delayed),
               names_to = "Status",
               values_to = "Count")

ggplot(summary_long, aes(x = Airline, y = Count, fill = Status)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = Count),
            position = position_stack(vjust = 0.5), size = 4) +
  labs(title = "Flight Status by Airline",
       y = "Number of Flights",
       x = "Airline") +
  theme_minimal()
```



AM West handled nearly twice as many flights (7,225) as Alaska (3,775).

Plot of Delay Rate and On-time Performance by Airline

```
# Convert to long format
plot_df <- summary_df %>%
  select(Airline, Delay_Rate, On_time_Performance) %>%
  pivot_longer(cols = c(Delay_Rate, On_time_Performance),
               names_to = "Metric", values_to = "Percentage")

# Plot with grouped bars
ggplot(plot_df, aes(x = Airline, y = Percentage, fill = Metric)) +
  geom_col(position = "dodge", width = 0.6) +
  geom_text(aes(label = paste0(Percentage, "%")),
            position = position_dodge(width = 0.6),
            vjust = -0.5, size = 4) +
  labs(
    title = "Delay Rate and On-time Performance by Airline",
    y = "Percentage (%)",
    x = "Airline"
  ) +
  ylim(0, 100) + # keep percentage scale
  scale_fill_manual(values = c("Delay_Rate" = "tomato", "On_time_Performance" = "steelblue")) +
  theme_minimal()
```



COMPARISON

AM West handled nearly twice as many flights (7,225) as Alaska (3,775).

On-time performance was calculated by using on time / Total_Flights * 100 For example Alaska: 86.7% on time 3274/3775*100 = 86.7

AM West: 89.1% on time

Delays

Alaska had a slightly higher proportion of delays (13.3%) compared to AM West (10.9%).

Even though Alaska's absolute delay numbers are lower (501 vs 787), that's because they operated fewer flights overall.

CONCLUSION

AM West performed better overall in terms of arrival delays, with a lower delay rate of(11%) compared to Alaska (13%).

Alaska still maintained strong on time performance, but its flights were slightly more likely to be delayed relative to AM West.

In tidy data:

Each column is a variable. Each row is an observation. Each cell is a single value.

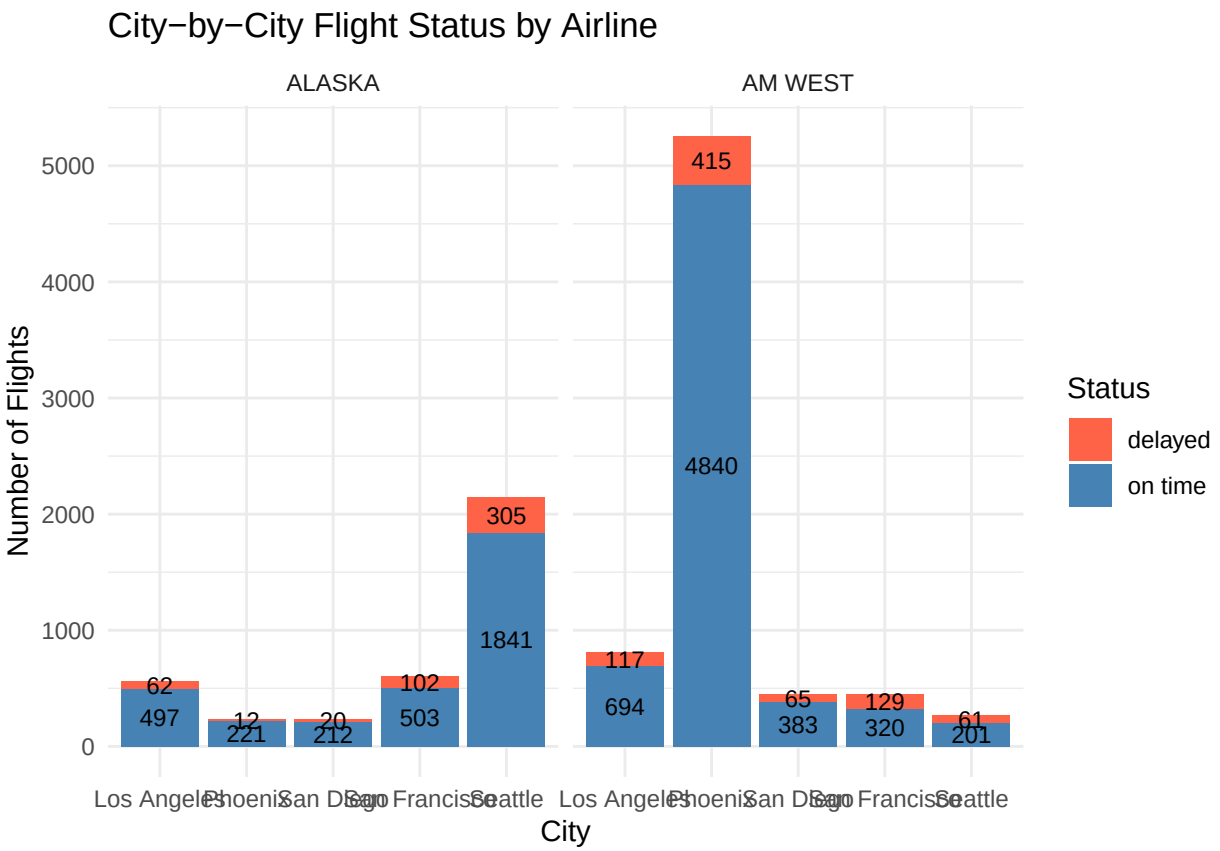
City by city delay analysis

```
# Convert Flightdata to long format for city-by-city analysis
flight_long <- Flightdata %>%
```

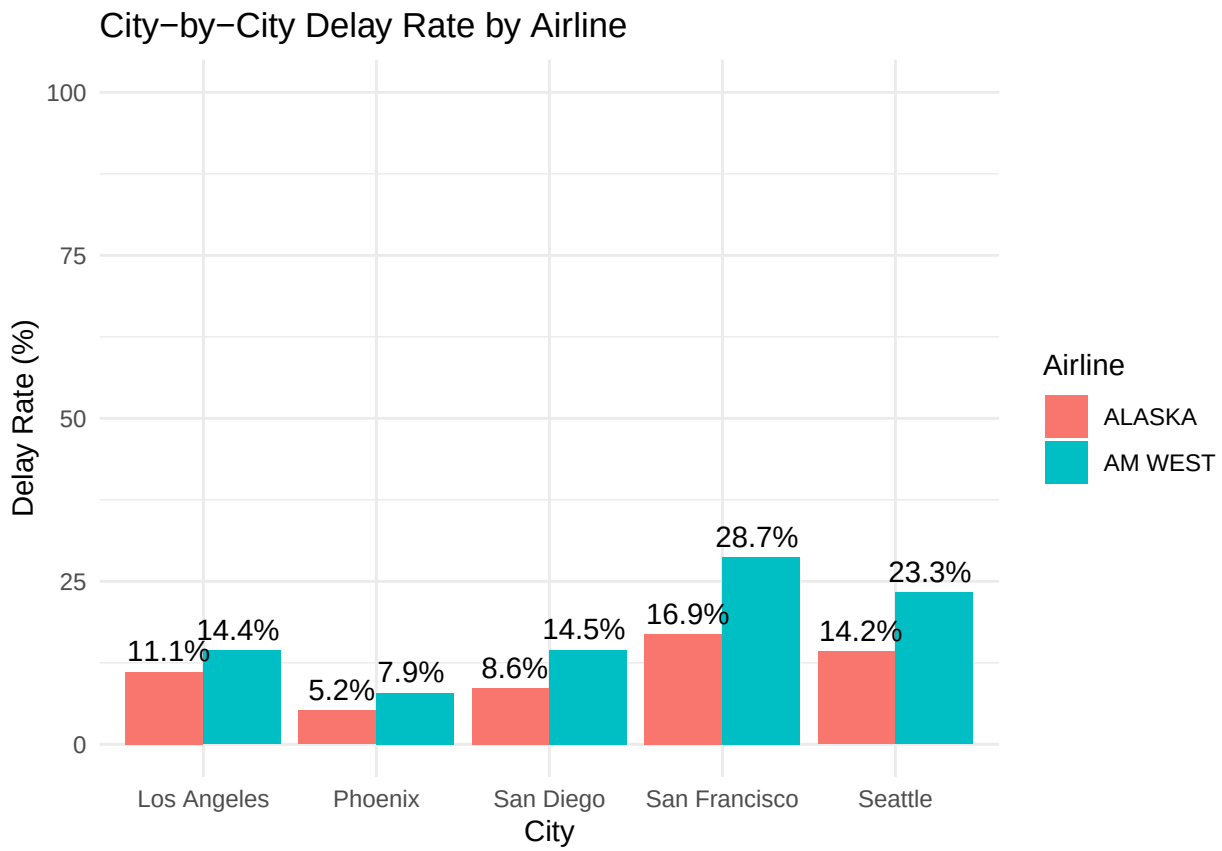
```r
  pivot_longer(
    cols = c("Los Angeles", "Phoenix", "San Diego", "San Francisco", "Seattle"),
    names_to = "City",
    values_to = "Flights"
  ) %>%
  filter(!is.na(Flights))  # remove NA rows

# Separate 'On_time_Delayed' into a more readable column
flight_long <- flight_long %>%
  mutate(Status = On_time_Delayed)

# Calculate total flights and delay rate per airline per city
city_summary <- flight_long %>%
  group_by(Airline, City, Status) %>%
  summarise(Total = sum(Flights), .groups = "drop") %>%
  pivot_wider(names_from = Status, values_from = Total, values_fill = 0) %>%
  mutate(
    Total_Flights = `on time` + delayed,
    Delay_Rate = round(delayed / Total_Flights * 100, 1),
    On_time_Performance = round(`on time` / Total_Flights * 100, 1)
  )

print(city_summary)
```

```
## # A tibble: 10 x 7
##    Airline City   delayed `on time` Total_Flights Delay_Rate On_time_Performance
##    <chr>   <chr>    <int>     <int>         <int>      <dbl>               <dbl>
##  1 ALASKA  Los A~      62       497           559       11.1                88.9
##  2 ALASKA  Phoen~      12       221           233        5.2                94.8
##  3 ALASKA  San D~      20       212           232        8.6                91.4
##  4 ALASKA  San F~     102       503           605       16.9                83.1
##  5 ALASKA  Seatt~     305      1841          2146       14.2                85.8
##  6 AM WEST Los A~     117       694           811       14.4                85.6
##  7 AM WEST Phoen~     415      4840          5255        7.9                92.1
##  8 AM WEST San D~      65       383           448       14.5                85.5
##  9 AM WEST San F~     129       320           449       28.7                71.3
## 10 AM WEST Seatt~      61       201           262       23.3                76.7
```

```r
# Convert to long format for plotting
city_plot_df <- city_summary %>%
  select(Airline, City, `on time`, delayed) %>%
  pivot_longer(cols = c("on time", delayed), names_to = "Status", values_to = "Count")

# Stacked bar chart by city
ggplot(city_plot_df, aes(x = City, y = Count, fill = Status)) +
  geom_bar(stat = "identity") +
  facet_wrap(~Airline) +
  geom_text(aes(label = Count), position = position_stack(vjust = 0.5), size = 3) +
  labs(title = "City-by-City Flight Status by Airline",
       y = "Number of Flights",
       x = "City") +
  scale_fill_manual(values = c("on time" = "steelblue", "delayed" = "tomato")) +
  theme_minimal()
```




```
# Plot Delay Rate by city for each airline
ggplot(city_summary, aes(x = City, y = Delay_Rate, fill = Airline)) +
  geom_col(position = "dodge") +
  geom_text(aes(label = paste0(Delay_Rate, "%")), position = position_dodge(width = 0.8), vjust = -0.5)
  labs(title = "City-by-City Delay Rate by Airline",
       y = "Delay Rate (%)",
       x = "City") +
  ylim(0, 100) +
  theme_minimal()
```



Alaska Airlines has lower delay rates across all five cities.

AM WEST has significant delays in San Francisco and Seattle. AM West had more flight than Alaska though which could have a compounding effect of the delays. The most delays are seen in those two cities as well.

The difference in delays is smallest in Phoenix and Los Angeles, but more pronounced in San Francisco and Seattle.