607 Tidying and Transforming Data

Pricilla

2025-09-16

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

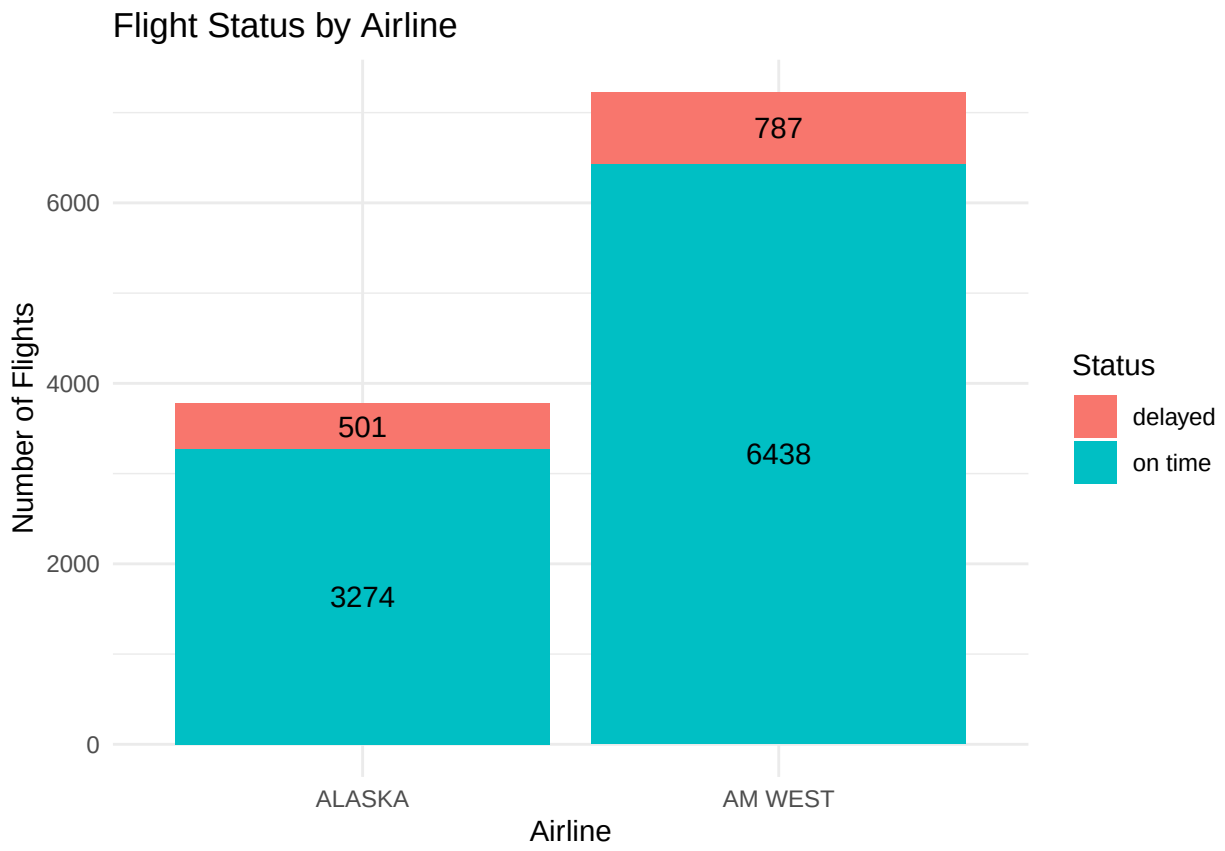
Delay rate = 787 ÷ (6438 + 787) is 10.9%

```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

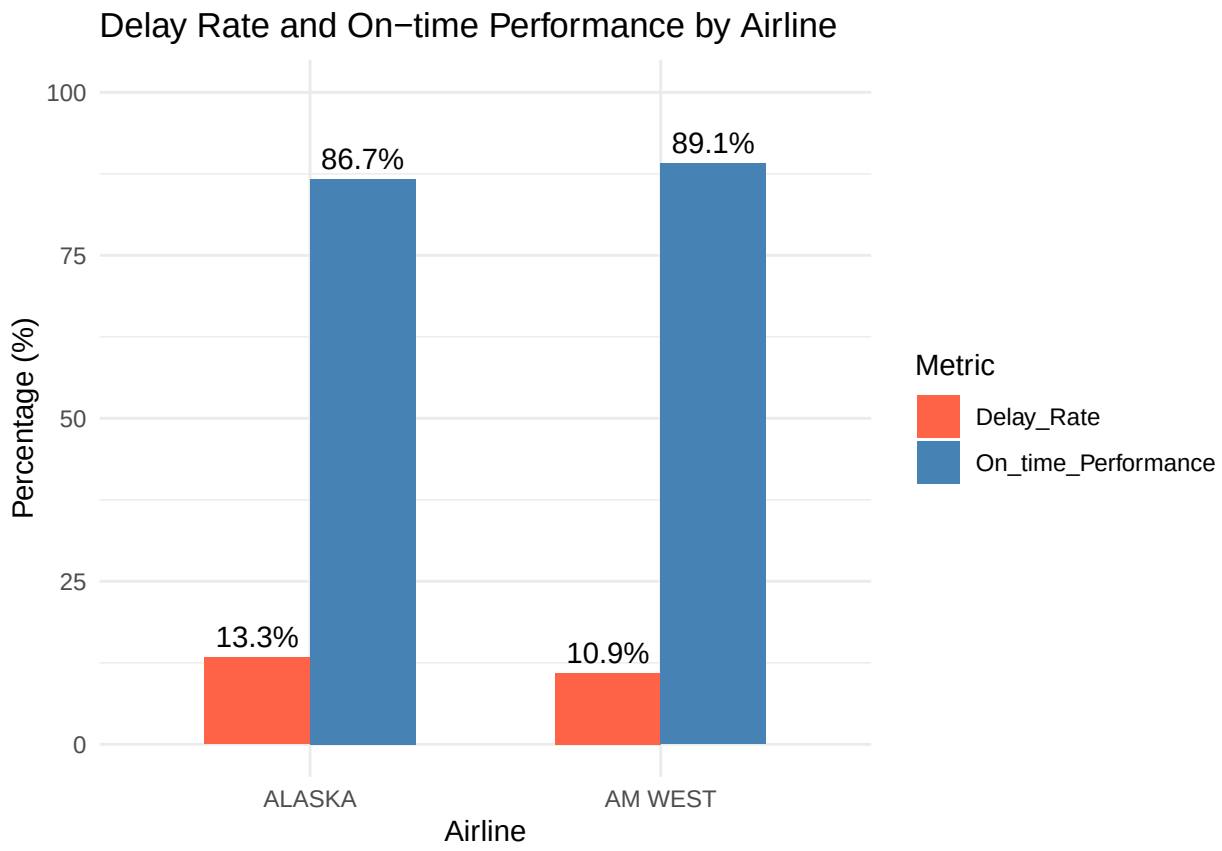
```r
summary_long <- summary_df %>%
  pivot_longer(cols = c("on time", delayed),
               names_to = "Status",
               values_to = "Count")

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.