DATA607_HW5_RKASA

Renida Kasa

10/8/2023

Assignment – Tidying and Transforming Data

Your task is to:

(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.

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

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

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- 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(dplyr)
library(ggplot2)
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
##
## The following object is masked from 'package:tidyr':
##
##     smiths
```

I started by entering all of the data for this assignment into a data frame.

```
Week5<-data.frame("Airline"=c("Alaska","Alaska","AM West","AM West"),
           "Time"=c("on time","delayed","on time","delayed"),
           "Los_Angeles"=c(497,62,694,117),
           "Phoenix"=c(221,12,4840,415),
           "San_Diego"=c(212,20,383,65),
           "San_Francisco"=c(503,102,320,129),
           "Seattle"=c(1841,305,201,61))
Week5
```

```
##   Airline    Time 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 AM West on time         694    4840       383           320     201
## 4 AM West delayed         117     415        65           129      61
```

In this data frame, we see flights for two different airlines, Alaska and AM West, arriving into 5 different destinations: Los Angeles, Phoenix, San Diego, San Francisco, and Seattle. The counts of flights are featured for each, as well as whether they were on time or delayed. It looks like AM West is delayed more often than Alaska, however, they also experience more flights in some cases. Just by looking at this, I would be more likely to take the Alaska airlines to any of the destinations, except for Seattle where it experienced more delays than AM West. However, to determine which flight is better overall, we would have to dig deeper!

I converted this to a csv file to easily access the data.

```
write.csv(Week5, "airline_delays.csv", row.names = FALSE)
```

I calculated the mean delay for each of the airlines.

```
airline_data <- read.csv("airline_delays.csv")

summary_stats <- airline_data %>%
  group_by(Airline) %>%
  summarise(avg_delay = mean(Los_Angeles + Phoenix + San_Diego + San_Francisco + Seattle),total_delay =

summary_stats
```

```
## # A tibble: 2 x 3
##   Airline avg_delay total_delay
##   <chr>       <dbl>       <int>
## 1 AM West     3612.        7225
## 2 Alaska      1888.        3775
```

I started by calculating the average overall flight delays for each of the airlines. Even though AM West experienced more delays than Alaska, I wanted to know the proportions of delays and on time arrivals for each flight. This would help to better determine which flight has a lower risk of being delayed. We are unable to tell at this point because the flights for each of the airlines at each destination are not equal, so the mean does not help much here.

I then calculated the proportion of on time and delayed arrivals:

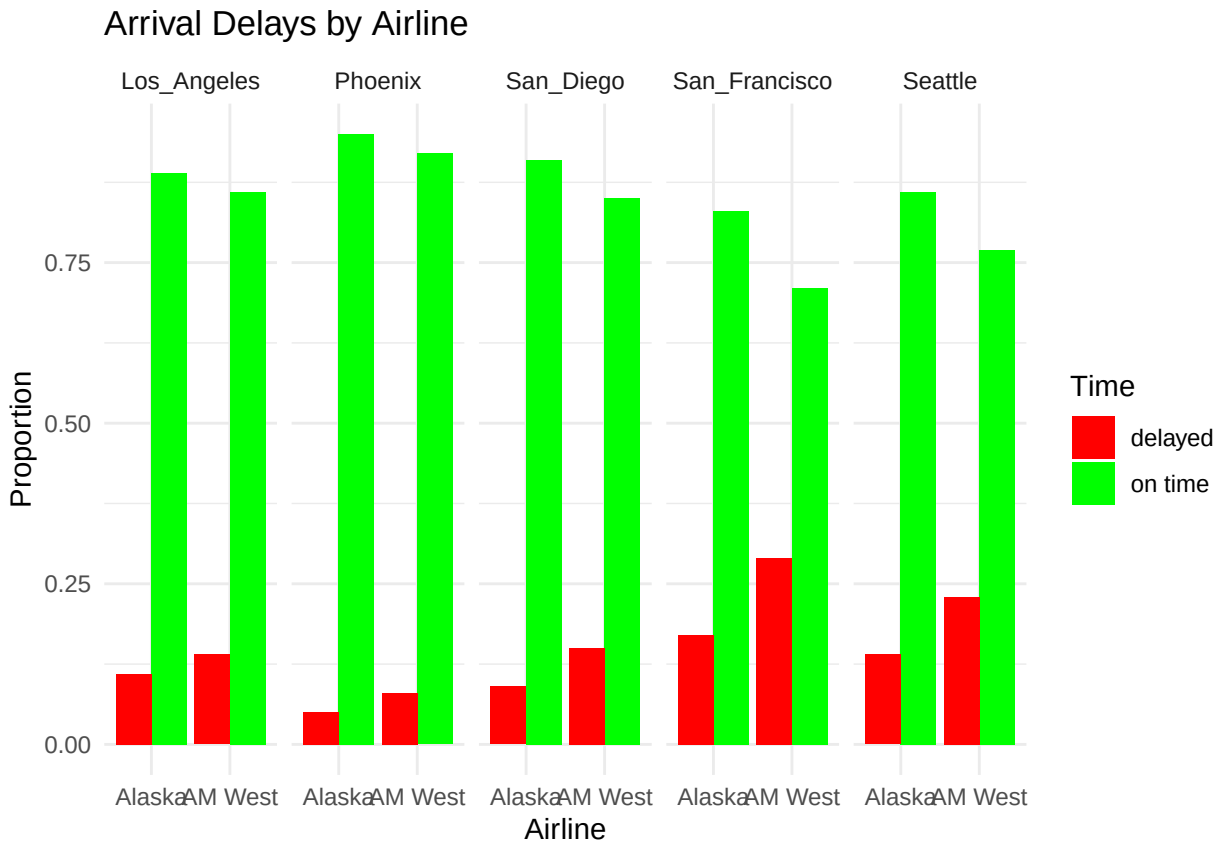
```r
arrival_prop <- data.frame(
  "Airline" = c("Alaska", "Alaska", "AM West", "AM West"),
  "Time" = c("on time", "delayed", "on time", "delayed"),
  "Los_Angeles" = c(round(497 / (497 + 62),2), round(62 / (497 + 62),2), round(694 / (694 + 117),2), rou
  "Phoenix" = c(221 / (221 + 12), 12 / (221 + 12), 4840 / (4840 + 415), 415 / (4840 + 415)),
  "San_Diego" = c(212 / (212 + 20), 20 / (212 + 20), 383 / (383 + 65), 65 / (383 + 65)),
  "San_Francisco" = c(503 / (503 + 102), 102 / (503 + 102), 320 / (320 + 129), 129 / (320 + 129)),
  "Seattle" = c(1841 / (1841 + 305), 305 / (1841 + 305), 201 / (201 + 61), 61 / (201 + 61))
)

arrival_prop <- arrival_prop %>%
  mutate_if(is.numeric, ~round(., 2))

print(arrival_prop)
```

```
##   Airline    Time Los_Angeles Phoenix San_Diego San_Francisco Seattle
## 1  Alaska on time        0.89    0.95      0.91          0.83    0.86
## 2  Alaska delayed        0.11    0.05      0.09          0.17    0.14
## 3 AM West on time        0.86    0.92      0.85          0.71    0.77
## 4 AM West delayed        0.14    0.08      0.15          0.29    0.23
```

The proportion of on time and delayed flights varies for both airlines. We now have a clearer pictue of what to expect if we were to travel with each one to any of the 5 locations, improving our judgement when deciding on an airline.

Here is a graph of the proportions:

```r
arrival_prop_long <- melt(arrival_prop, id.vars = c("Airline", "Time"))
ggplot(arrival_prop_long, aes(x = Airline, y = value, fill = Time)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Arrival Delays by Airline", x = "Airline", y = "Proportion") +
  scale_fill_manual(values = c("on time" = "green", "delayed" = "red")) +
  theme_minimal() +
  facet_wrap(~ variable, nrow = 1)
```



Here we see that Alaska airlines is generally on time or experiences less delays when compared to AM West. This is information that we would not have gathered had we just looked at the table without analyzing any data. Previously, I stated that I would rather take AM West over Alaska airlines if I were to fly to Seattle, simply because it experienced fewer delays. Even though it experienced fewer delays, we can also see in the graph above that it experienced fewer flights. 23% of its flights were delayed, comared to 12% for Alaska airlines. To conclude, it looks like in every case, Alaska airlines has fewer delays and is more likely to be on time!