Tidying and Transforming Data

Z. Palmore

9/25/2020

## Assignment 4

library(tidyverse)

## -- Attaching packages --------------------------------------------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2 v purrr 0.3.4  
## v tibble 3.0.3 v dplyr 1.0.2  
## v tidyr 1.1.1 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.5.0

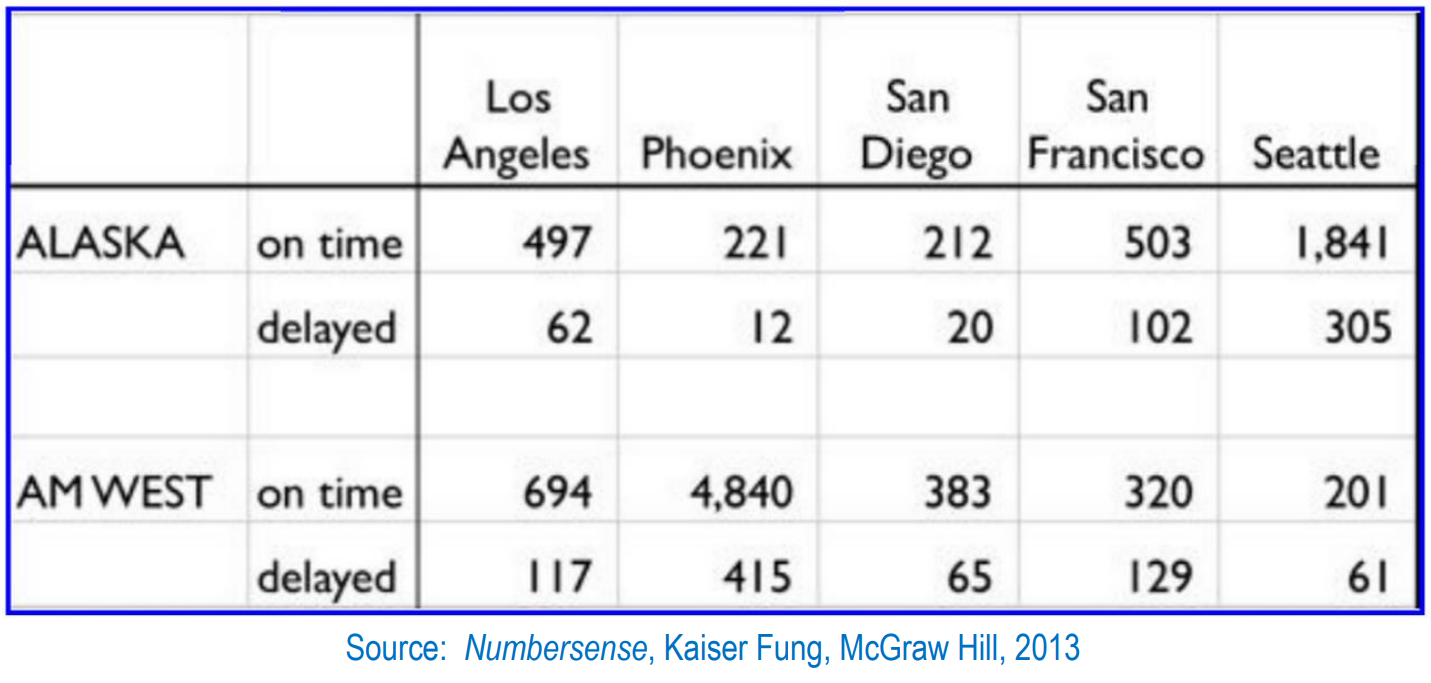
## -- Conflicts ------------------------------------------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(readr)  
library(reshape2)

##   
## Attaching package: 'reshape2'

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

### Directions



Alt text

The chart above describes arrival delays for two airlines across five destinations. 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.
4. Your code should be in an R Markdown file, posted to rpubs.com, and should include narrative descriptions of your data cleanup work, analysis, and conclusions.

Please include in your homework submission:

* The URL to the .Rmd file in your GitHub repository. and
* The URL for your rpubs.com web page.

# Step 1: Creating the CSV

A .csv file was created by manually entering the records from the chart into the spreadsheet. The file includes all data from the chart as it was formatted. It can be viewed in a separate window by running this chunk.

setwd("C:/Users/Owner/Documents/GitHub/msdsdata607/Assignments/Assignment 4")  
sampleflights <- read\_csv("Sample\_Flights.csv")

## Warning: Missing column names filled in: 'X1' [1], 'X2' [2]

# sampleflights <- read.csv("GitHub/msdsdata607/Assignments/Assignment 4/Sample\_Flights.csv")  
View(sampleflights)

### Step 2: Importing and Tidying

With data written into the .csv file and imported exactly as was in the chart, there are many NA values spread throughout the data. These should be removed and the data cleaned to make analysis easier.

In this data, there are presently

I would like to:

* Add Column Names for:
  + Airline
  + Status
* Add the Airline names for
  + delayed status of
    - Alaska
    - AM West
* Remove the 3rd Row
  + Containing all NA values

This was accomplished using the following steps.

sampleflights <- as.data.frame(sampleflights)  
# Specify the column names needed  
colnames(sampleflights) <- c("Airline",   
 "Status",   
 "Los Angeles",   
 "Phoenix",  
 "San Diego",  
 "San Francisco",  
 "Seattle")  
# Remove row 3 since it is all NA's  
sampleflights <- sampleflights[-3,]  
# Add the character string Alaska where is said NA  
sampleflights[2,1] <- ("Alaska")  
# Add the character string AM West where is said NA  
sampleflights[4,1] <- ("AM West")  
# Review the data frame  
sampleflights

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

### Step 3: Arrival Delay Analysis

There are a few ways to compare the two airlines arrival delays. The first is to combine all arrival delays for each location. Another is to find the mean of each row across those same locations. One could also use these to find the probability of delay for each airline. All will be performed for comparison.

# This code sums the rows using the all data from columns 3 through 7  
# Then it creates a new column in the data frame  
# which contains the sums of each row called "Total"  
  
sampleflights$Total <- rowSums(sampleflights[,3:7])  
sampleflights$Means <- rowMeans(sampleflights[,3:7])  
  
# Review the data frame  
sampleflights

## Airline Status Los Angeles Phoenix San Diego San Francisco Seattle Total  
## 1 Alaska on time 497 221 212 503 1841 3274  
## 2 Alaska delayed 62 12 20 102 305 501  
## 4 AM West on time 694 4840 383 320 201 6438  
## 5 AM West delayed 117 415 65 129 61 787  
## Means  
## 1 654.8  
## 2 100.2  
## 4 1287.6  
## 5 157.4

Based on this analysis, the airline AM West has the largest quantity of arrival delays in both its total delayed and its average delayed. Given that there were only two airlines, this makes Alaska’s airline the least delayed by quantity of the data (since we decided the other airline has the longest delayed time). If minutes were used to measure and record delays, then it is clear both airlines had average delays greater than 100 minutes.

Other statistics show the minimum delay occurred with Alaska when arriving in the location of Phoenix at 12. The maximum arrival delay also occurred in Phoenix but with the Airline AM West. This is shown here using the Minima and Maxima columns that were create for this purpose.

# Applying the minimum function over the rows (by using the margin of 1) in the data frame   
# then storing results in a new column called "Minima"  
sampleflights$Minima <- apply(sampleflights[3:7], 1, FUN = min)  
  
# Selecting the variables to display the minimum flight, its status and the airline  
sampleflights[2,c(1,2,4)]

## Airline Status Phoenix  
## 2 Alaska delayed 12

This displayed the minimum flight delay for the study.

# Repeat the same function for the maximum of flights and calling the column "Maxima"  
sampleflights$Maxima <- apply(sampleflights[,3:7], 1, FUN = max)  
  
# Selecting the variables to display the maximum flight, its status and the airline  
sampleflights[3,c(1,2,4)]

## Airline Status Phoenix  
## 4 AM West on time 4840

While this displayed the maximum delay.

The total on-time flights with Alaska and AM West were 3274 and 6438 respectively. If the records directly indicated the quantity of flights that were on-time or delayed, then we can create a ratio of delayed to on-time flights as shown here;

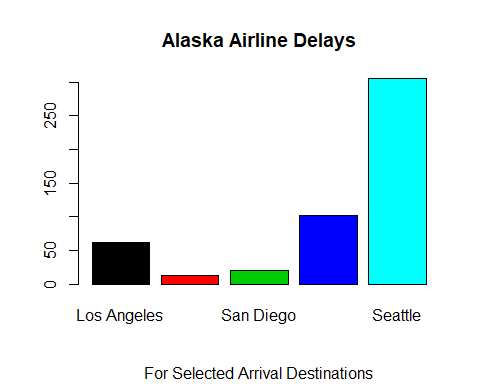
# Probability of delay for airline Alaska  
AK\_ratio <- sampleflights[2,8] / sampleflights[1,8]  
# Probability of delay for airline AM West  
AM\_ratio <- sampleflights[4,8] / sampleflights[3,8]  
# Rounding AK\_ratio to the correct number of significant figures, 3  
PAK <- signif(AK\_ratio, digits = 3)  
# Rounding AK\_ratio to the correct number of significant figures, 3  
PAM <- signif(AM\_ratio, digits = 3)

This calculation results in the probability of being delayed for each airline based on the selected flights for this dataset. We can see that the probability of delay on airline Alaska is 0.153 or about 15.3%. For AM West, the probability of delay is 0.122 or about 12.2%. Using this method of analysis, AW West has a slightly lower change of being delayed based on the ratio of it total delays to its total number of flights. It may may also be helpful to graph the variables and demonstrate the differences.

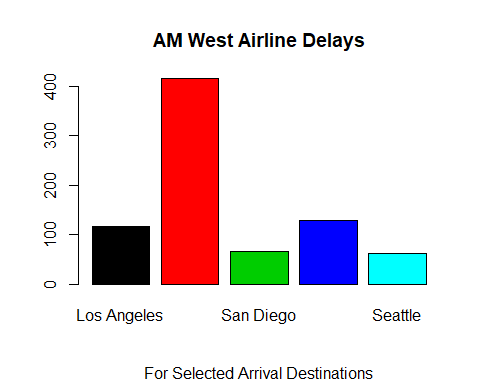
# Create a new data frame with columns for plotting in ggplot  
someflights <- melt(sampleflights)

## Using Airline, Status as id variables

delays <- someflights %>%  
 filter(Status == "delayed")  
# drop the statitics that are not needed by selecting what is  
delays <- delays[1:10,]  
# Creating a data frame with delays for Alaska   
AK\_delays <- delays %>%   
 filter(Airline == "Alaska")  
# Creating a data frame with delays for AW West   
AM\_delays <- delays %>%   
 filter(Airline == "AM West")  
# plotting delays for each location  
barplot(AK\_delays$value, horiz = FALSE, names.arg = AK\_delays$variable, main = "Alaska Airline Delays", sub = "For Selected Arrival Destinations", col = AK\_delays$variable)

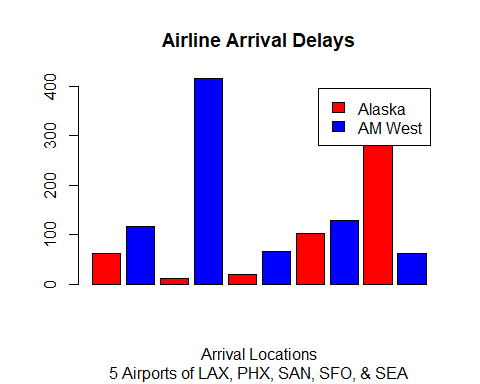


barplot(AM\_delays$value, horiz = FALSE, names.arg = AM\_delays$variable, main = "AM West Airline Delays", sub = "For Selected Arrival Destinations", col = AM\_delays$variable)



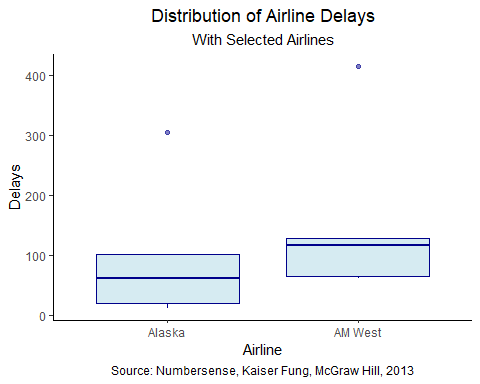
These bar charts show the difference in arrival delays for each airline by location. It is worth noting the difference in the x-axis where AM West has enough airline to increase its range displayed.

barplot(delays$value, col = c("Red", "Blue"), main = "Airline Arrival Delays", xlab = "Arrival Locations", legend = c("Alaska","AM West"), beside = TRUE, sub = "5 Airports of LAX, PHX, SAN, SFO, & SEA")



This bar chart demonstrates that difference in total number of flights a little better.Across the chart we can see AM West almost has consistently more flights than Alaska. It is also worth noting that both seem to be focusing on arrivals at one location. AM West has most of its flights arriving in second blue bar (which happens to be Phoenix and the largest bar), while Alaska concentrates its flights in the last red bar (of Seattle). The difference in distributions however, can be shown in the boxplots below.

ggplot(data = delays, aes(x = Airline, y = value) ) +   
 geom\_boxplot(color="Dark Blue", fill="Light Blue", alpha=0.5) +   
 labs(title = "Distribution of Airline Delays", subtitle = "With Selected Airlines",   
 y ="Delays", caption = "Source: Numbersense, Kaiser Fung, McGraw Hill, 2013") +   
 theme\_classic() +  
 theme(plot.title = element\_text(hjust = 0.5),   
 plot.caption = element\_text(hjust = 0.5),   
 plot.subtitle = element\_text(hjust = 0.5))



There are two points that are skewing the data. One at about 300 on the y-axis for airline Alaska and another around 400 for AM West. Their interquartile ranges are visibly different, with Alaska’s being lager and centered closer to its mean. However, AM West is not centered, it is heavily skewed towards higher delays and thus its mean is much higher in its “box.” These distributions may help explain why we can run two different methods of analysis (one using raw statistics and another using calculated probability) and get two different results.

### Submission

This document has been submitted to rpubs and is also contained within this GitHub repository.