Homework 5

OMER OZEREN

March 02, 2019

Table of Contents

Library definition 1
(1) Read information from .CSV file into R. 2
(2) Renaming Column headers 2
(3) Eliminating Empty rows with "NA" values 2
(4) Adding missing Airline name to "delayed" row 3
(5) Analysis 4
(a) Flights Status by airlines 5
(b) Flight Status by City 6
(c) Joining tables with horizontal probabilities 10

---

(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.

Library definition

```
library(stringr)
library(tidyr)
library(dplyr)
library(knitr)
library(ggplot2)
```

(1) Read information from .CSV file into R.

For simplicity and reproducibility reasons, I have posted this file on my GitHub repository as follows:

GitHub URL

```
url <-
"https://raw.githubusercontent.com/omerozeren/DATA607/master/HMW_5/airlines.c
sv"

raw_data <- read.csv(url, header=FALSE, sep=",", stringsAsFactors=FALSE)
raw_data <- data.frame(raw_data)
raw_data

##         V1      V2          V3      V4        V5            V6     V7
## 1                  Los Angeles Phoenix San Diego San Francisco Seatle
## 2  ALASKA on time         497     221       212           503   1841
## 3          delayed          62      12        20           102    305
## 4
## 5 AM WEST on time         694    4840       383           320    201
## 6          delayed         117     415        65           129     61
```

(2) Renaming Column headers

```
# Adding "Missing" titles from original file onto the Row #1
raw_data$V1[1] <- "Airline"
raw_data$V2[1] <- "Status"
# Assigning all the values from the row #1 as the Column Headers
names(raw_data) <- raw_data[1,]
# Need to eliminate Row #1 in order to keep data consistency.
raw_data <-raw_data[-c(1), ]
```

Table displaying correct column titles.

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

(3) Eliminating Empty rows with “NA” values

For this, I have to transform our data as follows:

```
## 'data.frame':    5 obs. of  7 variables:
##  $ Airline      : chr  "ALASKA" "" "" "AM WEST" ...
##  $ Status       : chr  "on time" "delayed" "" "on time" ...
##  $ Los Angeles  : chr  "497" "62" "" "694" ...
##  $ Phoenix      : chr  "221" "12" "" "4840" ...
##  $ San Diego    : chr  "212" "20" "" "383" ...
##  $ San Francisco: chr  "503" "102" "" "320" ...
##  $ Seatle       : chr  "1841" "305" "" "201" ...
```

Procedure to transform values into integers

```
for (i in 3:dim(raw_data)[2]){
  raw_data[,i] <- as.integer(raw_data[,i])
}
```

Preview of data after transformation

```
## 'data.frame':    5 obs. of  7 variables:
##  $ Airline      : chr  "ALASKA" "" "" "AM WEST" ...
##  $ Status       : chr  "on time" "delayed" "" "on time" ...
##  $ Los Angeles  : int  497 62 NA 694 117
##  $ Phoenix      : int  221 12 NA 4840 415
##  $ San Diego    : int  212 20 NA 383 65
##  $ San Francisco: int  503 102 NA 320 129
##  $ Seatle       : int  1841 305 NA 201 61
```

Procedure to eliminate all the **NA** lines from our original file by employing **drop_na()**

```
raw_data <- raw_data %>% drop_na()
```

	Airline	Status	Los Angeles	Phoenix	San Diego	San Francisco	Seatle
2	ALASKA	on time	497	221	212	503	1841
3		delayed	62	12	20	102	305
5	AM WEST	on time	694	4840	383	320	201
6		delayed	117	415	65	129	61

(4) Adding missing Airline name to “delayed” row

```
for (i in 1:dim(raw_data)[1]){
  if (i %% 2 == 0){
      raw_data$Airline[i] <-  raw_data$Airline[i-1]
  }
}
```

Final completed table in order to start employing **tidy** transformations for further analysis.

	Airline	Status	Los Angeles	Phoenix	San Diego	San Francisco	Seatle
2	ALASKA	on time	497	221	212	503	1841
3	ALASKA	delayed	62	12	20	102	305

5	AM WEST	on time	694	4840	383	320	201
6	AM WEST	delayed	117	415	65	129	61

(5) Analysis

First: we need to transform our table by employing **gather()** from **tidyr** library.

```
# Tidy table by having 4 variables (Airline, Status, City, number of flights)
flight <- gather(raw_data, City, Flight_Count, 3:7)
```

Airline	Status	City	Flight_Count
ALASKA	on time	Los Angeles	497
ALASKA	delayed	Los Angeles	62
AM WEST	on time	Los Angeles	694
AM WEST	delayed	Los Angeles	117
ALASKA	on time	Phoenix	221
ALASKA	delayed	Phoenix	12
AM WEST	on time	Phoenix	4840
AM WEST	delayed	Phoenix	415
ALASKA	on time	San Diego	212
ALASKA	delayed	San Diego	20
AM WEST	on time	San Diego	383
AM WEST	delayed	San Diego	65
ALASKA	on time	San Francisco	503
ALASKA	delayed	San Francisco	102
AM WEST	on time	San Francisco	320
AM WEST	delayed	San Francisco	129
ALASKA	on time	Seatle	1841
ALASKA	delayed	Seatle	305
AM WEST	on time	Seatle	201
AM WEST	delayed	Seatle	61

```
# grouping by flights
total_A <- flight %>% group_by(Airline) %>% summarise(Total_Flights =
sum(Flight_Count))
kable(total_A)
```

Airline	Total_Flights
ALASKA	3775
AM WEST	7225

(a) Flights Status by airlines

```
# Total of flights from each airline that were on time
on_time_airline <- flight %>% group_by(Airline) %>% filter(Status == 'on
time') %>% summarise(Flights_On_Time = sum(Flight_Count))
kable(on_time_airline)
```

Airline	Flights_On_Time
ALASKA	3274
AM WEST	6438

```
# Total of flights from each airline that were delayed.
delayed_airline <- flight %>% group_by(Airline) %>% filter(Status ==
'delayed') %>% summarise(Flights_Delayed = sum(Flight_Count))
kable(delayed_airline)
```

Airline	Flights_Delayed
ALASKA	501
AM WEST	787

Combine delayed and on_time data sets

```
# Now will combine all the data set information (including new columns) into
data.frame flights.summary
flights_summary_airline <- cbind(on_time_airline, Flights_Delayed =
delayed_airline$Flights_Delayed, Total_Flights = total_A$Total_Flights)
flights_summary_airline <- flights_summary_airline %>%
mutate(Percent_On_Time_airline = Flights_On_Time/Total_Flights,
Percent_Delayed_airline = Flights_Delayed/Total_Flights)

kable(flights_summary_airline)
```

Airline	Flights_On_Time	Flights_Delayed	Total_Flights	Percent_On_Time_airline	Percent_Delayed_airline
ALASKA	3274	501	3775	0.8672848	0.1327152
AM WEST	6438	787	7225	0.8910727	0.1089273

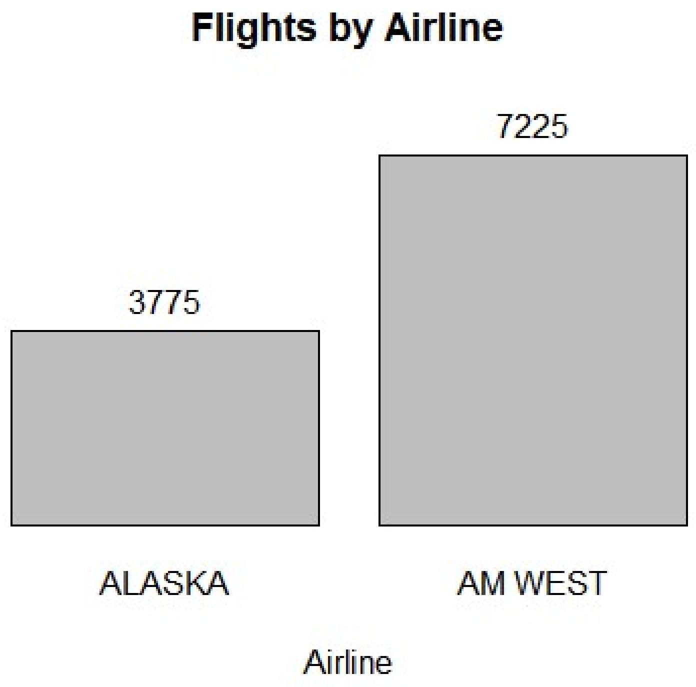
Overall, it appears that AM_West seems to be doing slightly a better job of staying on time. And not to mention, AM West flew more flights than Alaska.

Now I will create two data.frames where one is Alaska Airlines, and the other is AM_West Airlines.

```
Alaska <- flight %>% filter(Airline == 'ALASKA')
AM_West <- flight %>% filter(Airline == 'AM WEST')
```

Plot:

```
my.plot <- barplot(flights_summary_airline$Total_Flights, main="Flights by
Airline", xlab="Airline", names.arg=flights_summary_airline$Airline,
axes=FALSE, ylim = c(0, max(flights_summary_airline$Total_Flights)+1000))
# Placing values on top of bars
text(my.plot, flights_summary_airline$Total_Flights, labels =
flights_summary_airline$Total_Flights, pos = 3)
```



(b) Flight Status by City

grouping by flights

```
total_C <- flight %>% group_by(City) %>% summarise(Total_Flights =
sum(Flight_Count))
kable(total_C)
```

City	Total_Flights
Los Angeles	1370
Phoenix	5488
San Diego	680
San Francisco	1054
Seatle	2408

```
# Total of flights from each airline that were on time
on_time_city <- flight %>% group_by(City) %>% filter(Status == 'on time') %>%
summarise(Flights_On_Time = sum(Flight_Count))
kable(on_time_city)
```

City	Flights_On_Time
Los Angeles	1191
Phoenix	5061
San Diego	595
San Francisco	823
Seatle	2042

```r
# Total of flights from each airline that were delayed.
delayed_city <- flight %>% group_by(City) %>% filter(Status == 'delayed') %>%
summarise(Flights_Delayed = sum(Flight_Count))
kable(delayed_city)
```

City	Flights_Delayed
Los Angeles	179
Phoenix	427
San Diego	85
San Francisco	231
Seatle	366

Now will combine all the data set information (including new columns) into data.frame flights.summary

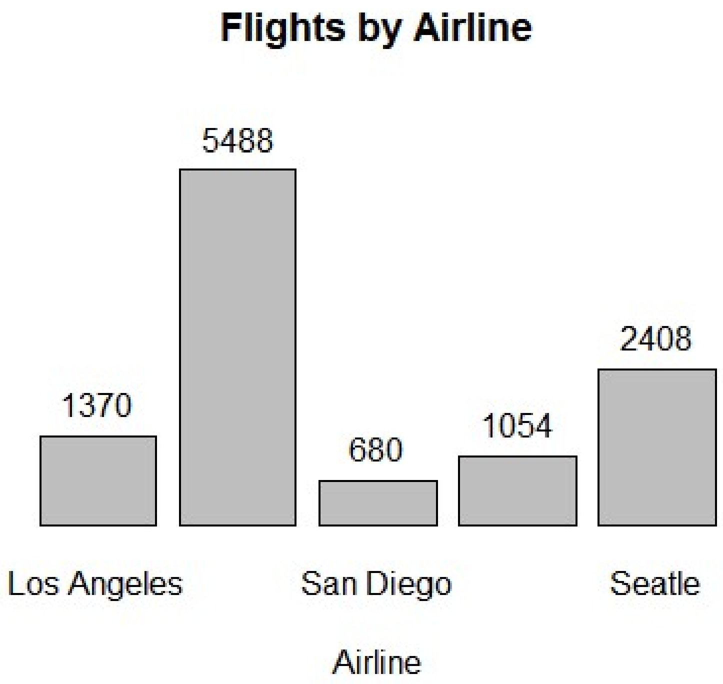
```r
flights_summary_city<- cbind(on_time_city, Flights_Delayed =
delayed_city$Flights_Delayed, Total_Flights = total_C$Total_Flights)
flights_summary_city <- flights_summary_city %>% mutate(Percent_On_Time_city
= Flights_On_Time/Total_Flights, Percent_Delayed_city =
Flights_Delayed/Total_Flights)

kable(flights_summary_city)
```

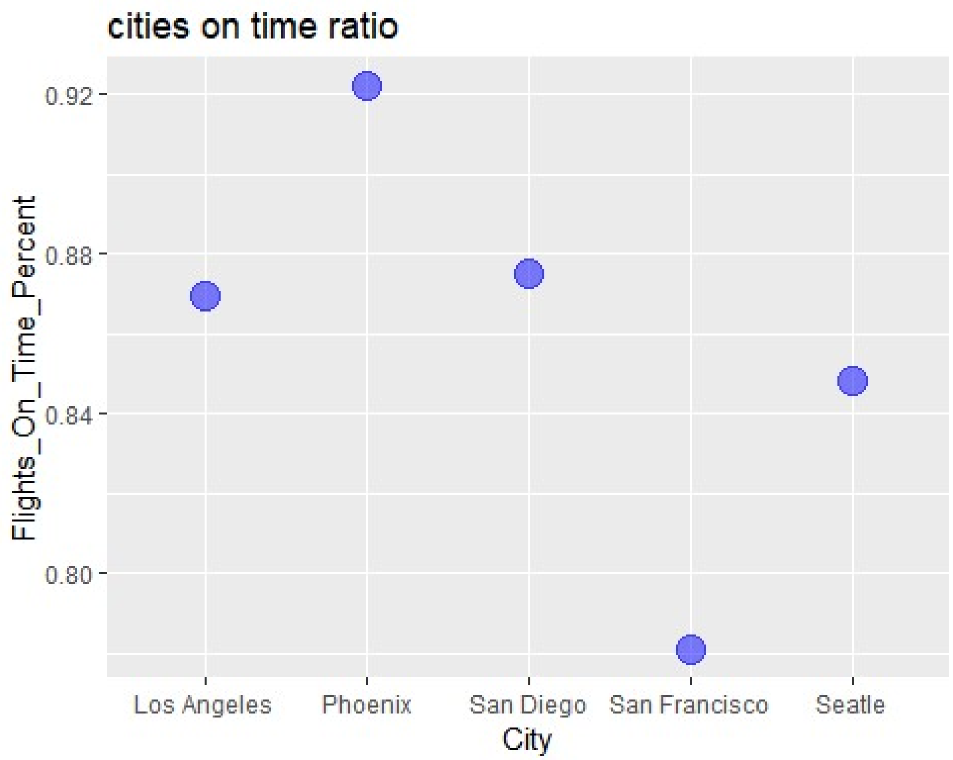
City	Flights_On_Time	Flights_Delayed	Total_Flights	Percent_On_Time_city	Percent_Delayed_city
Los Angeles	1191	179	1370	0.8693431	0.1306569
Phoenix	5061	427	5488	0.9221939	0.0778061
San Diego	595	85	680	0.8750000	0.1250000
San Francisco	823	231	1054	0.7808349	0.2191651
Seatle	2042	366	2408	0.8480066	0.1519934

Plot:

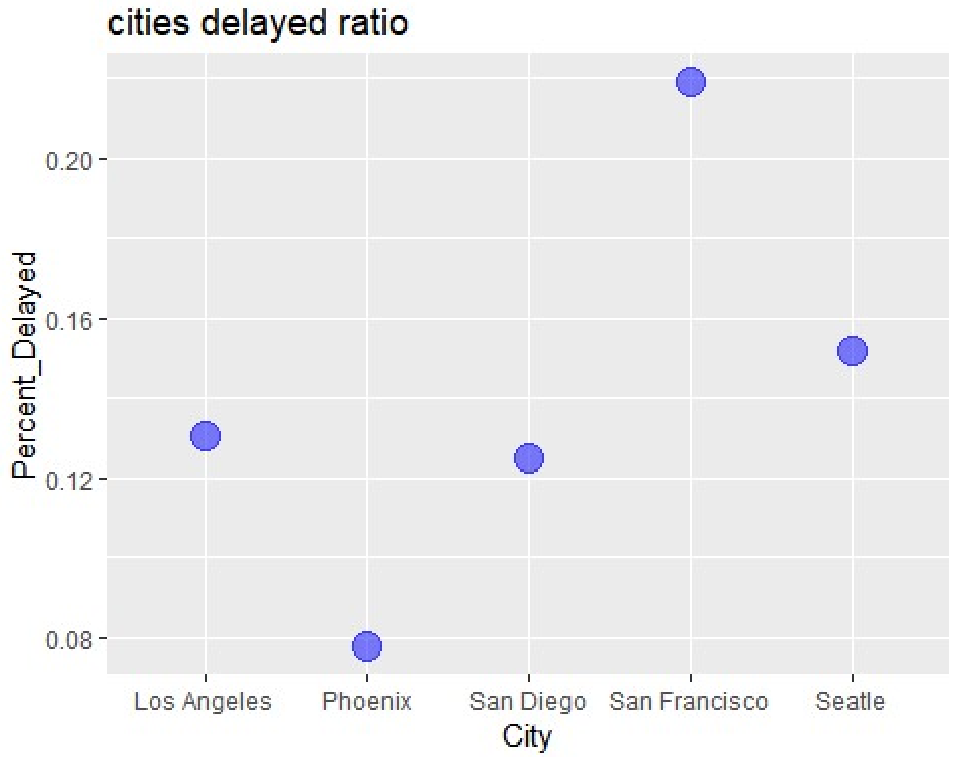
```
my.plot <- barplot(flights_summary_city$Total_Flights, main="Flights by
Airline", xlab="Airline", names.arg=flights_summary_city$City, axes=FALSE,
ylim = c(0, max(flights_summary_city$Total_Flights)+1000))
# Placing values on top of bars
text(my.plot, flights_summary_city$Total_Flights, labels =
flights_summary_city$Total_Flights, pos = 3)
```



```
ggplot(flights_summary_city, aes(x = City, y = Percent_On_Time_city)) +
geom_point(alpha = 0.5, size = 5, color = 'blue') +  labs(title ="cities on
time ratio", x = "City", y = "Flights_On_Time_Percent")
```



```
ggplot(flights_summary_city, aes(x = City, y = Percent_Delayed_city)) +
geom_point(alpha = 0.5, size = 5, color = 'blue') +  labs(title ="cities
delayed ratio", x = "City", y = "Percent_Delayed")
```



(c) Joining tables with horizontal probabilities

```r
spread_data <-  flight %>% spread(Status, `Flight_Count`)

main_table <- spread_data %>% subset(select=c(Airline, City))
airline_table <-  flights_summary_airline %>% subset(select=c(Airline,
Percent_Delayed_airline,Percent_On_Time_airline))
city_table <-  flights_summary_city %>% subset(select=c(City,
Percent_Delayed_city,Percent_On_Time_city))
main_table <-  main_table %>% inner_join(airline_table, by="Airline" )
main_table <-  main_table %>% inner_join(city_table, by="City" )
kable(main_table)
```

Airline	City	Percent_Delayed_airline	Percent_On_Time_airline	Percent_Delayed_city	Percent_On_Time_city
ALASKA	Los Angeles	0.1327152	0.8672848	0.1306569	0.8693431
ALASKA	Phoenix	0.1327152	0.8672848	0.0778061	0.9221939
ALASKA	San Diego	0.1327152	0.8672848	0.1250000	0.8750000
ALASKA	San Francisco	0.1327152	0.8672848	0.2191651	0.7808349
ALASKA	Seatle	0.1327152	0.8672848	0.1519934	0.8480066
AM WEST	Los Angeles	0.1089273	0.8910727	0.1306569	0.8693431
AM WEST	Phoenix	0.1089273	0.8910727	0.0778061	0.9221939
AM WEST	San Diego	0.1089273	0.8910727	0.1250000	0.8750000
AM WEST	San Francisco	0.1089273	0.8910727	0.2191651	0.7808349
AM WEST	Seatle	0.1089273	0.8910727	0.1519934	0.8480066