Airlines across five destinations

CUNY MSDA - DATA607 - Homework 5

Completed by: Duubar Villalobos Jimenez mydvtech@gmail.com

March 5, 2017

		Los Angeles	Phoenix	San Diego	San Francisco	Seattle
ALASKA	on time	497	221	212	503	1,841
	delayed	62	12	20	102	305
AM WEST	on time	694	4,840	383	320	201
	delayed	117	415	65	129	61

Source: *Numbersense*, Kaiser Fung, McGraw Hill, 2013

Figure 1:

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.

PROCEDURE

Library definition

```
# Need to employ kable
library(knitr)
```

```
# Need to employ stringr for Regular Expressions
library(stringr)
# Need to employ to use tidy data functions
library(tidyr)
library(dplyr)
```

(1) Create .CSV file

I have created a .CSV file named: "Villalobos-airlines.csv"

		Los Angeles	Phoenix	San Diego	San Francisco	Seattle
ALASKA	on time	497	221	212	503	1,841
	delayed	62	12	20	102	305
AM WEST	on time	694	4,840	383	320	201
	delayed	117	415	65	129	61

Source: *Numbersense*, Kaiser Fung, McGraw Hill, 2013

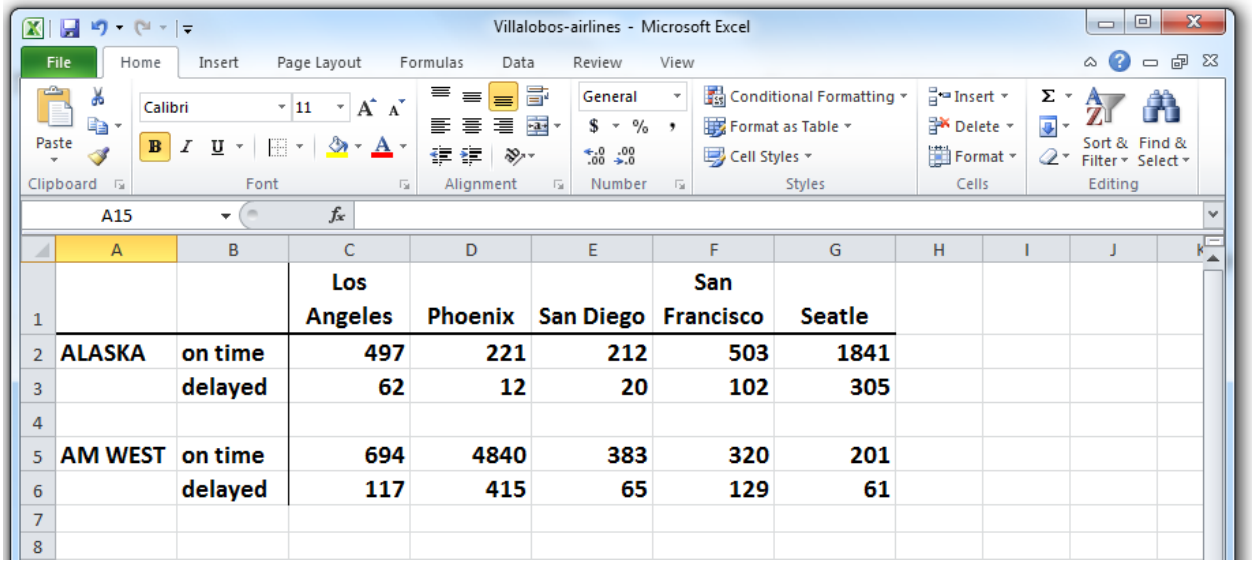


Figure 2: Villalobos-airlines.csv

(2) 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/dvillalobos/MSDA/master/607/Homework/Homework5/Villalobos-airl
```

Read .csv from url by employing read.csv()

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

Imported file display

```
##        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
```

Data transformation

Renaming Column headers

```
# Adding "Missing" titles from original file onto the Row #1
my.data$V1[1] <- "Airline"
my.data$V2[1] <- "Status"

# Assigning all the values from the row #1 as the Column Headers
names(my.data) <- my.data[1,]

# Need to eliminate Row #1 in order to keep data consistency.
my.data <- my.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

Eliminating Empty rows with "NA" values by employing drop_na() from the tidy library.

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(my.data)[2]){
  my.data[,i] <- as.integer(my.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()**

```
my.data <- my.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

Adding missing Airline name to "delayed" row

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

(3) 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)
my.tidy.data <- my.data %>% gather("City","n flights", 3:7)
```

Airline	Status	City	n flights
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

Second: Now, I will separate the values **"on time"** and **"delayed"** from the **Status** column into two different columns by employing the **spread()** function from **tidyr** library.

Please note that these values can be considered as two different variables.

```
my.tidy.data <-  my.tidy.data %>%
                 spread(Status, `n flights`)
```

Final Tidy Table

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

a) Total of flights sorted ascending.

Airline	City	delayed	on time	Total
ALASKA	San Diego	20	212	232
ALASKA	Phoenix	12	221	233
AM WEST	Seatle	61	201	262
AM WEST	San Diego	65	383	448
AM WEST	San Francisco	129	320	449
ALASKA	Los Angeles	62	497	559
ALASKA	San Francisco	102	503	605
AM WEST	Los Angeles	117	694	811
ALASKA	Seatle	305	1841	2146
AM WEST	Phoenix	415	4840	5255

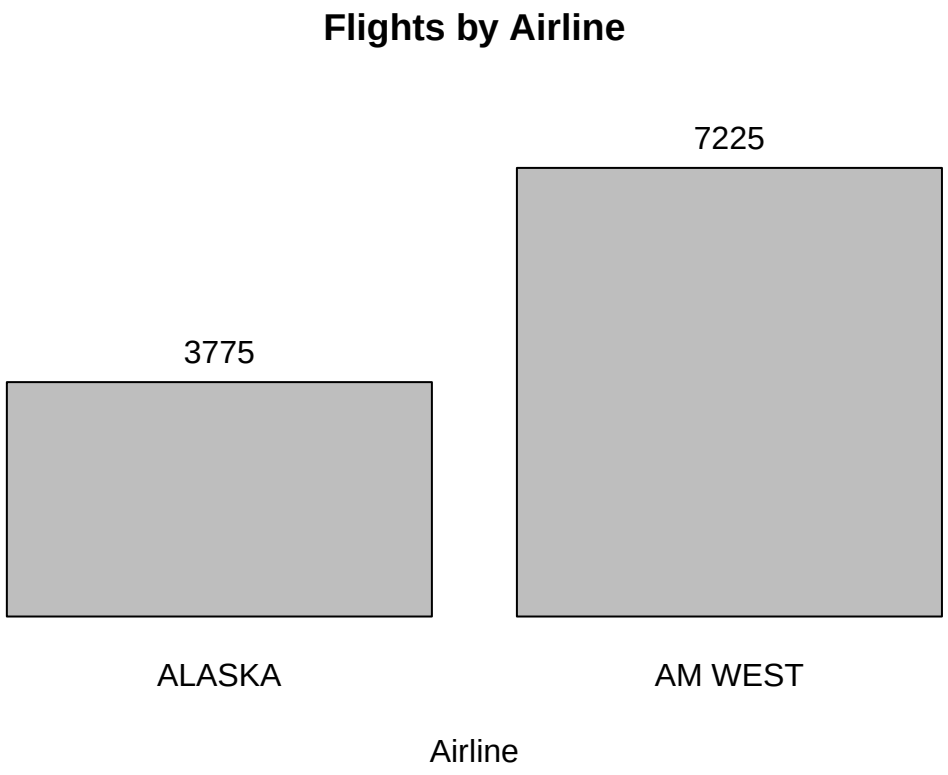
b) Ratio of "delayed" flights vs "on time" flights sorted ascending.

Airline	City	delayed	on time	Total	ratio
ALASKA	Phoenix	12	221	233	0.0542986
AM WEST	Phoenix	415	4840	5255	0.0857438
ALASKA	San Diego	20	212	232	0.0943396
ALASKA	Los Angeles	62	497	559	0.1247485
ALASKA	Seatle	305	1841	2146	0.1656708
AM WEST	Los Angeles	117	694	811	0.1685879
AM WEST	San Diego	65	383	448	0.1697128
ALASKA	San Francisco	102	503	605	0.2027833
AM WEST	Seatle	61	201	262	0.3034826
AM WEST	San Francisco	129	320	449	0.4031250

Noticed how the order changed!

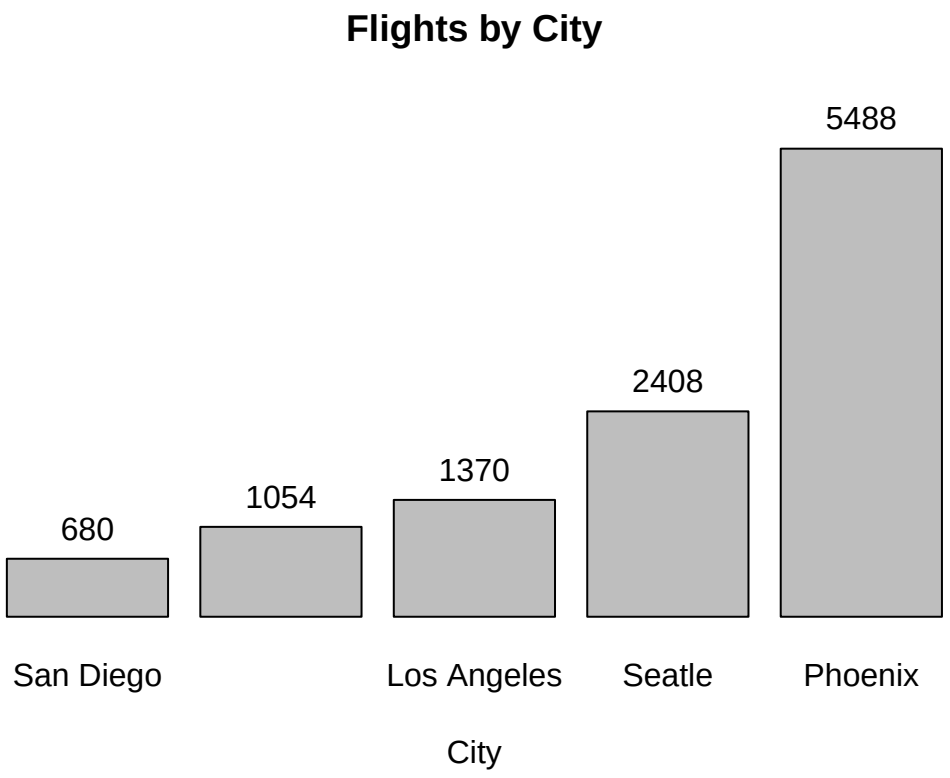
c) Total of flights by Airline sorted ascending.

Airline	delayed	on time	Total
ALASKA	501	3274	3775
AM WEST	787	6438	7225



d) Total of flights by City sorted ascending.

City	delayed	on time	Total
San Diego	85	595	680
San Francisco	231	823	1054
Los Angeles	179	1191	1370
Seatle	366	2042	2408
Phoenix	427	5061	5488



Some Horizontal Probabilities

From the previous **"Final Tidy Table"** we can find some horizontal probabilities, those can be found as follows:

Airline	City	delayed	on time	Total
ALASKA	San Diego	20	212	232
ALASKA	Phoenix	12	221	233
AM WEST	Seatle	61	201	262
AM WEST	San Diego	65	383	448
AM WEST	San Francisco	129	320	449
ALASKA	Los Angeles	62	497	559
ALASKA	San Francisco	102	503	605
AM WEST	Los Angeles	117	694	811

Airline	City	delayed	on time	Total
ALASKA	Seatle	305	1841	2146
AM WEST	Phoenix	415	4840	5255

a) Horizontal probabilities for "delayed" and "on time" flights by Airline and City.

Airline	City	delayed	on time	Total	P(delayed)	P(on time)
ALASKA	San Diego	20	212	232	0.0862069	0.9137931
ALASKA	Phoenix	12	221	233	0.0515021	0.9484979
AM WEST	Seatle	61	201	262	0.2328244	0.7671756
AM WEST	San Diego	65	383	448	0.1450893	0.8549107
AM WEST	San Francisco	129	320	449	0.2873051	0.7126949
ALASKA	Los Angeles	62	497	559	0.1109123	0.8890877
ALASKA	San Francisco	102	503	605	0.1685950	0.8314050
AM WEST	Los Angeles	117	694	811	0.1442663	0.8557337
ALASKA	Seatle	305	1841	2146	0.1421249	0.8578751
AM WEST	Phoenix	415	4840	5255	0.0789724	0.9210276

b) Horizontal probabilities for "delayed" and "on time" flights by Airline only.

Airline	delayed	on time	Total
ALASKA	501	3274	3775
AM WEST	787	6438	7225

Airline	delayed	on time	Total	P(A delayed)	P(A on time)
ALASKA	501	3274	3775	0.1327152	0.8672848
AM WEST	787	6438	7225	0.1089273	0.8910727

c) Horizontal probability for "delayed" and "on time" flights by City only.

City	delayed	on time	Total
San Diego	85	595	680
San Francisco	231	823	1054
Los Angeles	179	1191	1370
Seatle	366	2042	2408
Phoenix	427	5061	5488

City	delayed	on time	Total	P(C delayed)	P(C on time)
San Diego	85	595	680	0.1250000	0.8750000
San Francisco	231	823	1054	0.2191651	0.7808349
Los Angeles	179	1191	1370	0.1306569	0.8693431
Seatle	366	2042	2408	0.1519934	0.8480066
Phoenix	427	5061	5488	0.0778061	0.9221939

Joining tables with horizontal probabilities

For this, I will join the **Final Tidy Table** with the respective probabilities tables in order to create a **Final Horizontal Probability Table** by employing **inner_join()** from the **dplyr** library.

Resulting Horizontal Probability table:

Airline	City	P(A delayed)	P(A on time)	P(C delayed)	P(C on time)
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

By comparing with the below table, we noticed that the values are different. That is that so far I have calculated only horizontal probabilities and further analysis can be performed.

Airline	City	delayed	on time	Total	P(delayed)	P(on time)
ALASKA	San Diego	20	212	232	0.0862069	0.9137931
ALASKA	Phoenix	12	221	233	0.0515021	0.9484979
AM WEST	Seatle	61	201	262	0.2328244	0.7671756
AM WEST	San Diego	65	383	448	0.1450893	0.8549107
AM WEST	San Francisco	129	320	449	0.2873051	0.7126949
ALASKA	Los Angeles	62	497	559	0.1109123	0.8890877
ALASKA	San Francisco	102	503	605	0.1685950	0.8314050
AM WEST	Los Angeles	117	694	811	0.1442663	0.8557337
ALASKA	Seatle	305	1841	2146	0.1421249	0.8578751
AM WEST	Phoenix	415	4840	5255	0.0789724	0.9210276

Other probabilities

Other probabilities that can be found will be **Vertical Probabilities** or/and **Total Probabilities** by taking into consideration the total flights ("delayed" + "on time").

Once those probabilities are found we can then start answering questions like:

a) What's the probability that a randonmly selected flight will be delayed?

b) What's the probability that a randonly selected flight will be from Alaska Airlines and it's destination will be Seatle?

c) What's the probability that out of 5 randonmly selected flights will be 2 delayed and 3 on time?