607 - Data Cleaning Operations

Karim Hammoud

2020-09-25

Please find the Rpubs here and the Github link

Tidying and Transforming Data

		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

The chart above describes arrival delays for two airlines across five destinations. Your task is to:

Import the libraries

```
library(tidyr)
library(dplyr)
library(RMySQL)
library(ggplot2)
```

Create the CSV file with the arrival delays data

```
data <- rbind(c("Airline", "Status", "Los Angeles", "Phoenix", "San Diego", "San Francisco", "Seattle")
              c("ALASKA", "On Time", 497, 221, 212, 503, 1841),
```

```
             c("ALASKA", "Delayed", 62, 12, 20, 102, 305),
             c("AM WEST", "On Time", 694, 4840, 383, 320, 201),
             c("AM WEST", "Delayed", 117, 415, 65, 129, 61))
write.table(data, file = "/Users/karimh/Documents/Google Drive/607 - 2020 Fall- Data Acquisition and Ma
```

Read the data from Github

```
url <- "https://raw.githubusercontent.com/akarimhammoud/CUNY-SPS/master/607-Data-Acquisition-and-Manage
delays <- read.csv(url, sep = ",")
delays
```

```
##   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
## 3 AM WEST On Time         694    4840       383           320     201
## 4 AM WEST Delayed         117     415        65           129      61
```

Connect to MySql on Google Cloud host

```
conn <- dbConnect (MySQL(),
                   user="root", password= "dav",
                   dbname="data607", host= "35.188.162.1")
conn
```

```
## <MySQLConnection:0,0>
```

Creating Tables

```
dbWriteTable (conn, 'delays', delays, overwrite = TRUE)
```

```
## [1] TRUE
```

Reading from the database

```
delays1 <- dbGetQuery(conn, 'select * from delays')
delays <- delays1 [,-1]
delays
```

```
##   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
## 3 AM WEST On Time         694    4840       383           320     201
## 4 AM WEST Delayed         117     415        65           129      61
```

Rearrange the data as a table and sort by Status.

```
delays_table <- delays %>%
 gather("Destination", "Flights", 3:7) %>%
 arrange(Airline, desc(Status), Destination)
delays_table
```

```
##    Airline  Status   Destination Flights
## 1   ALASKA On Time   Los.Angeles     497
## 2   ALASKA On Time       Phoenix     221
## 3   ALASKA On Time     San.Diego     212
## 4   ALASKA On Time San.Francisco     503
## 5   ALASKA On Time       Seattle    1841
## 6   ALASKA Delayed   Los.Angeles      62
## 7   ALASKA Delayed       Phoenix      12
## 8   ALASKA Delayed     San.Diego      20
## 9   ALASKA Delayed San.Francisco     102
## 10  ALASKA Delayed       Seattle     305
## 11 AM WEST On Time   Los.Angeles     694
## 12 AM WEST On Time       Phoenix    4840
## 13 AM WEST On Time     San.Diego     383
## 14 AM WEST On Time San.Francisco     320
## 15 AM WEST On Time       Seattle     201
## 16 AM WEST Delayed   Los.Angeles     117
## 17 AM WEST Delayed       Phoenix     415
## 18 AM WEST Delayed     San.Diego      65
## 19 AM WEST Delayed San.Francisco     129
## 20 AM WEST Delayed       Seattle      61
```

Analyze the arrival delays for the two airlines.

filtering and creating new data frame

```
on_time <- filter(delays_table, Status == "On Time")
on_time
```

```
##    Airline  Status   Destination Flights
## 1   ALASKA On Time   Los.Angeles     497
## 2   ALASKA On Time       Phoenix     221
## 3   ALASKA On Time     San.Diego     212
## 4   ALASKA On Time San.Francisco     503
## 5   ALASKA On Time       Seattle    1841
## 6  AM WEST On Time   Los.Angeles     694
## 7  AM WEST On Time       Phoenix    4840
## 8  AM WEST On Time     San.Diego     383
## 9  AM WEST On Time San.Francisco     320
## 10 AM WEST On Time       Seattle     201
```

```
delayed <- filter(delays_table, Status == "Delayed")
delayed
```

```
##    Airline  Status   Destination Flights
```

```
## 1   ALASKA Delayed   Los.Angeles      62
## 2   ALASKA Delayed       Phoenix      12
## 3   ALASKA Delayed     San.Diego      20
## 4   ALASKA Delayed San.Francisco     102
## 5   ALASKA Delayed       Seattle     305
## 6  AM WEST Delayed   Los.Angeles     117
## 7  AM WEST Delayed       Phoenix     415
## 8  AM WEST Delayed     San.Diego      65
## 9  AM WEST Delayed San.Francisco     129
## 10 AM WEST Delayed       Seattle      61
```

```
my_frame <- data.frame(on_time, delayed)
my_frame
```

```
##    Airline  Status   Destination Flights Airline.1 Status.1 Destination.1
## 1   ALASKA On Time   Los.Angeles     497    ALASKA  Delayed   Los.Angeles
## 2   ALASKA On Time       Phoenix     221    ALASKA  Delayed       Phoenix
## 3   ALASKA On Time     San.Diego     212    ALASKA  Delayed     San.Diego
## 4   ALASKA On Time San.Francisco     503    ALASKA  Delayed San.Francisco
## 5   ALASKA On Time       Seattle    1841    ALASKA  Delayed       Seattle
## 6  AM WEST On Time   Los.Angeles     694   AM WEST  Delayed   Los.Angeles
## 7  AM WEST On Time       Phoenix    4840   AM WEST  Delayed       Phoenix
## 8  AM WEST On Time     San.Diego     383   AM WEST  Delayed     San.Diego
## 9  AM WEST On Time San.Francisco     320   AM WEST  Delayed San.Francisco
## 10 AM WEST On Time       Seattle     201   AM WEST  Delayed       Seattle
##    Flights.1
## 1         62
## 2         12
## 3         20
## 4        102
## 5        305
## 6        117
## 7        415
## 8         65
## 9        129
## 10        61
```

The percentage of differences per city and airlines

```
my_frame$Diffrences <- my_frame$Flights / (my_frame$Flights + my_frame$Flights.1)
my_frame$Diffrences
```

```
##  [1] 0.8890877 0.9484979 0.9137931 0.8314050 0.8578751 0.8557337 0.9210276
##  [8] 0.8549107 0.7126949 0.7671756
```

creating new data frame

```
my_frame <- data.frame(my_frame$Airline, my_frame$Status, my_frame$Destination, my_frame$Diffrences)
my_frame
```

```
##    my_frame.Airline my_frame.Status my_frame.Destination my_frame.Diffrences
```

```
## 1            ALASKA         On Time          Los.Angeles           0.8890877
## 2            ALASKA         On Time              Phoenix           0.9484979
## 3            ALASKA         On Time            San.Diego           0.9137931
## 4            ALASKA         On Time        San.Francisco           0.8314050
## 5            ALASKA         On Time              Seattle           0.8578751
## 6           AM WEST         On Time          Los.Angeles           0.8557337
## 7           AM WEST         On Time              Phoenix           0.9210276
## 8           AM WEST         On Time            San.Diego           0.8549107
## 9           AM WEST         On Time        San.Francisco           0.7126949
## 10          AM WEST         On Time              Seattle           0.7671756
```

Percentage of delays and ontime per Airline

```
 Status_details <- delays_table %>%
  group_by(Airline) %>%
  mutate(Total_Airline = sum(Flights)) %>%
  group_by(Airline, Status) %>%
  mutate(Total_Airline_Status = sum(Flights), Status_Percentage  = Total_Airline_Status / Total_Airline

Final_percentage <- data.frame(Status_details[c(1,10,11,20), c(1,2,7)])
```

Percentage of delays and ontime per Ditenation for each Airline

```
Destination_details <- delays_table %>%
 group_by(Airline, Destination) %>%
 mutate(Total_destination = sum(Flights), Percentage_distintaion = Flights / Total_destination)

head(Destination_details)
```

```
## # A tibble: 6 x 6
## # Groups:   Airline, Destination [5]
##   Airline Status  Destination   Flights Total_destination Percentage_distintaion
##   <chr>   <chr>   <chr>           <dbl>             <dbl>                  <dbl>
## 1 ALASKA  On Time Los.Angeles       497               559                  0.889
## 2 ALASKA  On Time Phoenix           221               233                  0.948
## 3 ALASKA  On Time San.Diego         212               232                  0.914
## 4 ALASKA  On Time San.Francisco     503               605                  0.831
## 5 ALASKA  On Time Seattle          1841              2146                  0.858
## 6 ALASKA  Delayed Los.Angeles        62               559                  0.111
```

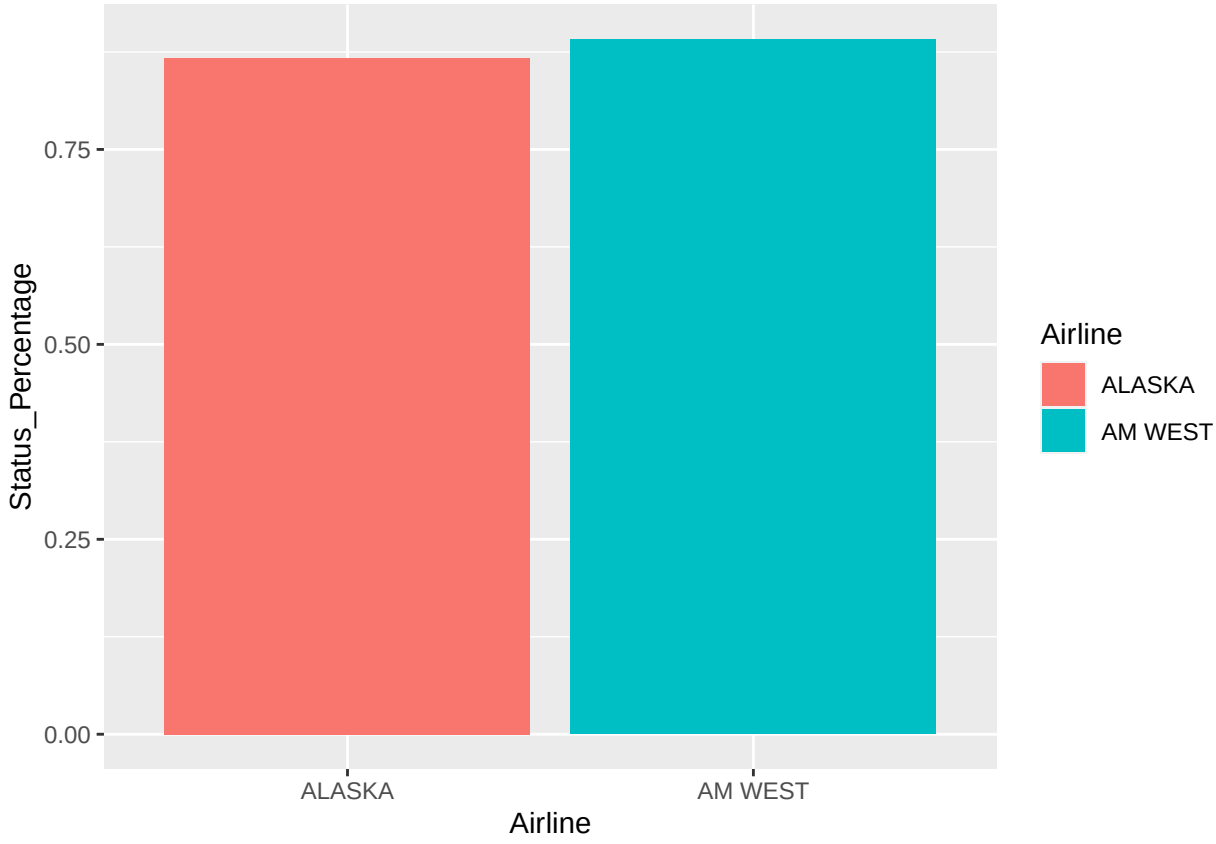
Filter only for the ONTIME and create the plot

```
Final_percentage
```

```
##   Airline  Status Status_Percentage
## 1  ALASKA On Time         0.8672848
## 2  ALASKA Delayed         0.1327152
## 3 AM WEST On Time         0.8910727
## 4 AM WEST Delayed         0.1089273
```

```r
on_time <- filter(Final_percentage, Status == "On Time")

ggplot(on_time , mapping = aes(x=Airline, y=Status_Percentage, fill=Airline)) +
  geom_bar(stat="identity",)
```



Conclusion

From the analysis above it look like the AM West has over 89% of its flights on time while Alaska has over 86%, both airlines have a close percentages of on time flights but in this example AM West has a higher number of flights.