Data607_Assignment5

Introduction

Which airline city has the best overall performance of on-time flight arrival?

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:

```
library(tidyr)
library(dplyr,warn.conflicts = FALSE)
options(dplyr.summarise.inform = FALSE)
library(stringr)
```

Step 1: Create csv file and upload to github

```
flight.data <- read.csv("https://raw.githubusercontent.com/szx868/data607/master/Assignment5/flight_data
flight.data[2,1] <- flight.data[1,1]
flight.data[5,1] <- flight.data[4,1]
flight.data[,2] <- sapply(flight.data[,2], str_replace, " ", ".")
flight.data
```

Step 2: Import csv file from github

```
##         X     X.1 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                          NA      NA        NA            NA      NA
## 4 AM WEST on.time         694    4840       383           320     201
## 5 AM WEST delayed         117     415        65           129      61
```

```
tidy.data <- flight.data %>%
  na.omit() %>%
  rename(airline = X, arrival.type = X.1) %>%
  gather("arrival.city", "n", 3:7) %>%
  spread(arrival.type, "n") %>%
  mutate(total.arrivals = delayed + on.time, on.time.rate.percent = on.time / total.arrivals*100) %>%
  arrange(desc(total.arrivals))

tidy.data[,2] <- sapply(tidy.data[,2], str_replace, "\\.", " ")
tidy.data
```

Step2.1 Tidy up data

```
##    airline  arrival.city delayed on.time total.arrivals on.time.rate.percent
## 1  AM WEST       Phoenix     415    4840           5255             92.10276
## 2   ALASKA       Seattle     305    1841           2146             85.78751
## 3  AM WEST   Los Angeles     117     694            811             85.57337
## 4   ALASKA San Francisco     102     503            605             83.14050
## 5   ALASKA   Los Angeles      62     497            559             88.90877
## 6  AM WEST San Francisco     129     320            449             71.26949
## 7  AM WEST     San Diego      65     383            448             85.49107
## 8  AM WEST       Seattle      61     201            262             76.71756
## 9   ALASKA       Phoenix      12     221            233             94.84979
## 10  ALASKA     San Diego      20     212            232             91.37931
```

Step 3: airline analysis

- The best on-time arrival rate of Arrival city

```
best.airlinecity <-
tidy.data %>%
  filter(on.time.rate.percent == max(on.time.rate.percent))
best.airlinecity
```

```
##   airline arrival.city delayed on.time total.arrivals on.time.rate.percent
## 1  ALASKA      Phoenix      12     221            233             94.84979
```

- The airline that has best on-time arrival rate

```
bestairline <- tidy.data %>%
  group_by(airline) %>%
  summarise(airline.on.time.rate.perecent = sum(on.time) / sum(total.arrivals)*100) %>%
  filter(airline.on.time.rate.perecent == max(airline.on.time.rate.perecent))
bestairline
```

```
## # A tibble: 1 x 2
##   airline airline.on.time.rate.perecent
##   <chr>                           <dbl>
## 1 AM WEST                          89.1
```

- Rank their performances from highest to lowest.

```
performances <- tidy.data %>%
  group_by(arrival.city) %>%
  summarise(city.on.time.rate.percent = sum(on.time) / sum(total.arrivals)*100) %>%
  mutate(on.time.ranking = min_rank(desc(city.on.time.rate.percent))) %>%
  arrange(on.time.ranking)
performances
```

```
## # A tibble: 5 x 3
##   arrival.city  city.on.time.rate.percent on.time.ranking
##   <chr>                             <dbl>           <int>
## 1 Phoenix                            92.2               1
## 2 San Diego                          87.5               2
## 3 Los Angeles                        86.9               3
## 4 Seattle                            84.8               4
## 5 San Francisco                      78.1               5
```

Inconclusion

It looks like City Phoenix has best overall on-time arrival rates with 92%