DATA607_Assignment5

Pei-Ming Chen

2023-02-26

Create a .CSV file that includes all of the information above. It is encouraged to use a "wide" structure similar to how the information appears above, so that I can practice tidying and transformations as described below. Read the information from your.CSV file in to R,and use tidyr and dplyr as needed to tidy and transform the data.Perform analysis to compare the arrival delays for the two airlines.The code should be in an R Markdown file, posted to rpubs.com, and should include narrative descriptions of the data cleanup work, analysis, and conclusions.

Packages download

```
library(tidyr)
library(dplyr)
```

FLights data CSV. file was uploaded to my Github account.

```
flights <- "https://raw.githubusercontent.com/PMCformosa/Assignment-5/main/Data_Flights.csv"

flights_Data <- read.table(flights,header = TRUE, sep=",", na.strings = c("","NA"))

flights_Data
```

```
##      X..       X Los.Angeles Phoenix San.Diego San.Francisco Seattle
## 1 ALASKA on time         497     221       212           503    1841
## 2   <NA> delayed          62      12        20           102     305
## 3   <NA>    <NA>          NA      NA        NA            NA      NA
## 4 AMWEST on time         694    4840       383           320     201
## 5   <NA> delayed         117     415        65           129      61
```

Tidy up the data

```
flights_Data[2,1] <- flights_Data[1,1]
flights_Data[3,] <- flights_Data[4,]
flights_Data[5,1] <- flights_Data[4,1]
flights_Data[4,] <- flights_Data[5,]
flights_Data <- flights_Data[-5,]

flights_Data
```

```
##      X..       X 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 AMWEST on time         694    4840       383           320     201
## 4 AMWEST delayed         117     415        65           129      61
```

Rename the first and second column head

```
colnames (flights_Data) [1] <- "Airline"
colnames (flights_Data) [2] <- "Arrival_status"

flights_Data
```

```
##   Airline Arrival_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  AMWEST        on time         694    4840       383           320     201
## 4  AMWEST        delayed         117     415        65           129      61
```

Convert wide data format into a long one by using gather function of tidyr package

```
flights_LongData <- gather(flights_Data, "city", "n", 3:7)

flights_LongData
```

```
##    Airline Arrival_status          city    n
## 1   ALASKA        on time   Los.Angeles  497
## 2   ALASKA        delayed   Los.Angeles   62
## 3   AMWEST        on time   Los.Angeles  694
## 4   AMWEST        delayed   Los.Angeles  117
## 5   ALASKA        on time       Phoenix  221
## 6   ALASKA        delayed       Phoenix   12
## 7   AMWEST        on time       Phoenix 4840
## 8   AMWEST        delayed       Phoenix  415
## 9   ALASKA        on time     San.Diego  212
## 10  ALASKA        delayed     San.Diego   20
## 11  AMWEST        on time     San.Diego  383
## 12  AMWEST        delayed     San.Diego   65
## 13  ALASKA        on time San.Francisco  503
## 14  ALASKA        delayed San.Francisco  102
## 15  AMWEST        on time San.Francisco  320
## 16  AMWEST        delayed San.Francisco  129
## 17  ALASKA        on time       Seattle 1841
## 18  ALASKA        delayed       Seattle  305
## 19  AMWEST        on time       Seattle  201
## 20  AMWEST        delayed       Seattle   61
```

```
dplyr::glimpse(flights_LongData)
```

```
## Rows: 20
## Columns: 4
## $ Airline        <chr> "ALASKA", "ALASKA", "AMWEST", "AMWEST", "ALASKA", "ALAS~
## $ Arrival_status <chr> "on time", "delayed", "on time", "delayed", "on time", ~
## $ city           <chr> "Los.Angeles", "Los.Angeles", "Los.Angeles", "Los.Angel~
## $ n              <int> 497, 62, 694, 117, 221, 12, 4840, 415, 212, 20, 383, 65~
```

Spread the elements of the Arrival_status column into two separate columns names "delayed" and "on time" by using the spread() function of the dplyr package

```
Airline_data2 <- flights_LongData %>% spread(Arrival_status, n)
Airline_data2
```

```
##    Airline          city delayed on time
## 1   ALASKA   Los.Angeles      62     497
## 2   ALASKA       Phoenix      12     221
## 3   ALASKA     San.Diego      20     212
## 4   ALASKA San.Francisco     102     503
## 5   ALASKA       Seattle     305    1841
## 6   AMWEST   Los.Angeles     117     694
## 7   AMWEST       Phoenix     415    4840
## 8   AMWEST     San.Diego      65     383
## 9   AMWEST San.Francisco     129     320
## 10  AMWEST       Seattle      61     201
```

Rename the fourth column

```
colnames (Airline_data2)[4] <- "on_time"
Airline_data2
```

```
##    Airline          city delayed on_time
## 1   ALASKA   Los.Angeles      62     497
## 2   ALASKA       Phoenix      12     221
## 3   ALASKA     San.Diego      20     212
## 4   ALASKA San.Francisco     102     503
## 5   ALASKA       Seattle     305    1841
## 6   AMWEST   Los.Angeles     117     694
## 7   AMWEST       Phoenix     415    4840
## 8   AMWEST     San.Diego      65     383
## 9   AMWEST San.Francisco     129     320
## 10  AMWEST       Seattle      61     201
```

Use the pipe operator to obtain mean and median values of delayed or on_time numbers

The result below showed that the mean number of delayed flights are 128.8 with a median number 83.5. And the mean number of on_time flights are 971.2 with a median 440.

```
Airline_data2 %>% summarise(mean = mean(delayed), median = median(delayed), n = n())
```

```
##    mean median  n
## 1 128.8   83.5 10
```

```
Airline_data2 %>% summarise(mean = mean(on_time), median = median(on_time), n = n())
```

```
##    mean median  n
## 1 971.2    440 10
```

Use the pipe operator to obtain the rate of on_time flights

The highest on_time rate is observed in ALASKA airline flying Phoenix. And the lowest on_time rate happened in AMWEST airline flying San.Francisco.

```
Airline_data2 <- mutate(Airline_data2, rate_on_time = on_time/(on_time+delayed))
Airline_data3 <- mutate(Airline_data2, rate_delayed = delayed/(on_time+delayed))
Airline_data3
```

```
##    Airline          city delayed on_time rate_on_time rate_delayed
## 1   ALASKA   Los.Angeles      62     497    0.8890877   0.11091234
## 2   ALASKA       Phoenix      12     221    0.9484979   0.05150215
## 3   ALASKA     San.Diego      20     212    0.9137931   0.08620690
## 4   ALASKA San.Francisco     102     503    0.8314050   0.16859504
## 5   ALASKA       Seattle     305    1841    0.8578751   0.14212488
## 6   AMWEST   Los.Angeles     117     694    0.8557337   0.14426634
## 7   AMWEST       Phoenix     415    4840    0.9210276   0.07897241
## 8   AMWEST     San.Diego      65     383    0.8549107   0.14508929
## 9   AMWEST San.Francisco     129     320    0.7126949   0.28730512
## 10  AMWEST       Seattle      61     201    0.7671756   0.23282443
```

```
Airline_data2 %>%
   group_by(Airline) %>%
        dplyr::summarise(max = max(delayed), min=min(delayed),
        mean=mean(delayed), median=median(delayed))
```

```
## # A tibble: 2 x 5
##   Airline   max   min  mean median
##   <chr>   <int> <int> <dbl>  <int>
## 1 ALASKA    305    12  100.     62
## 2 AMWEST    415    61  157.    117
```

Conclusion

From the data cleaning and analysis above, we can see that AMWEST airline had higher mean and median delayed flight numbers, compared to the numbers of ALASKA. AMWEST airline flying San.Francisco also had the lowest on_time rate.