Week 5Assignment - Tidying and Transforming Data

Peter Gatica

03/06/2021

Load needed libraries

```
library(devtools)
library(tidyverse)
library(RCurl)
library(knitr)
```

Source the untidy data source for cleansing and transformation

```
filename <- getURL("https://raw.githubusercontent.com/audiorunner13/Masters-Coursework/main/DATA607%20Sp
(airline_untidy <- read.delim(text=filename,header=TRUE, sep = ","))
```

```
##   airline arr_status Los.Angeles Phoenix San.Diego San.Franscisco Seattle
## 1  Alaska    on time         497     221       212            503    1841
## 2            delayed          62      12        20            102     305
## 3  Amwest    on time         694    4840       383            320     201
## 4            delayed         117     415        65            129      61
```

Use tidyr gather function to gather the values that are used as columns and make them correctly name the column that those values represent. In this case the values represent the airport.

```
(airline_tidy <- airline_untidy %>%
  gather(‘Los.Angeles‘,‘Phoenix‘,‘San.Diego‘,‘San.Franscisco‘,‘Seattle‘, key = "airport",value = "count
```

```
##    airline arr_status        airport count
## 1   Alaska    on time    Los.Angeles   497
## 2             delayed    Los.Angeles    62
## 3   Amwest    on time    Los.Angeles   694
## 4             delayed    Los.Angeles   117
## 5   Alaska    on time        Phoenix   221
## 6             delayed        Phoenix    12
## 7   Amwest    on time        Phoenix  4840
## 8             delayed        Phoenix   415
## 9   Alaska    on time      San.Diego   212
## 10            delayed      San.Diego    20
## 11  Amwest    on time      San.Diego   383
## 12            delayed      San.Diego    65
## 13  Alaska    on time San.Franscisco   503
## 14            delayed San.Franscisco   102
## 15  Amwest    on time San.Franscisco   320
## 16            delayed San.Franscisco   129
## 17  Alaska    on time        Seattle  1841
## 18            delayed        Seattle   305
## 19  Amwest    on time        Seattle   201
## 20            delayed        Seattle    61
```

Use the str_replace() function to look for the "." in the airport name and replace with a blank space.

```r
(airline_tidy$airport <- str_replace(airline_tidy$airport,"\\.", " "))
```

```
##  [1] "Los Angeles"    "Los Angeles"    "Los Angeles"    "Los Angeles"
##  [5] "Phoenix"        "Phoenix"        "Phoenix"        "Phoenix"
##  [9] "San Diego"      "San Diego"      "San Diego"      "San Diego"
## [13] "San Franscisco" "San Franscisco" "San Franscisco" "San Franscisco"
## [17] "Seattle"        "Seattle"        "Seattle"        "Seattle"
```

Every other row starting at record two is missing the airport value for that record. Use a while loop and if statement to identify those rows that need the airport name added.

```r
x <- 2
while (x < 21){
  if (x == 2 | x == 6 | x == 10 | x == 14 | x == 18){
    airline_tidy$airline[x] = 'Alaska'
  }
  if (x == 4 | x == 8 | x == 12 | x == 16 | x == 20){
    airline_tidy$airline[x] = 'Amwest'
  }
  x <- x + 2
}
airline_tidy
```

```
##    airline arr_status        airport count
## 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 Franscisco   503
## 14  Alaska    delayed San Franscisco   102
## 15  Amwest    on time San Franscisco   320
## 16  Amwest    delayed San Franscisco   129
## 17  Alaska    on time        Seattle  1841
## 18  Alaska    delayed        Seattle   305
## 19  Amwest    on time        Seattle   201
## 20  Amwest    delayed        Seattle    61
```

Use the filter() function to extract only those records with a delayed status.

```
(airline_delays <- airline_tidy %>% filter(arr_status == ’delayed’))
```

```
##    airline arr_status        airport count
## 1   Alaska    delayed    Los Angeles    62
## 2   Amwest    delayed    Los Angeles   117
## 3   Alaska    delayed        Phoenix    12
## 4   Amwest    delayed        Phoenix   415
## 5   Alaska    delayed      San Diego    20
## 6   Amwest    delayed      San Diego    65
## 7   Alaska    delayed San Franscisco   102
## 8   Amwest    delayed San Franscisco   129
## 9   Alaska    delayed        Seattle   305
## 10  Amwest    delayed        Seattle    61
```

Create a data.frame of Alaska airline delayed records to perform some analysis on.

```
(airline.alaska <- airline_delays %>% filter(airline == "Alaska"))
```

```
##   airline arr_status        airport count
## 1  Alaska    delayed    Los Angeles    62
## 2  Alaska    delayed        Phoenix    12
## 3  Alaska    delayed      San Diego    20
## 4  Alaska    delayed San Franscisco   102
## 5  Alaska    delayed        Seattle   305
```

Calculate the total count of delayed flights and the percentage of delayed flights by location for Alaska airlines.

One can see that Alaska Airlines at the Seattle airport experiences the most delayed flights. Once explanation for that may be weather. The Seattle area is known for the high amount of rainfall every year.

```
(airline.alaska <- group_by(airline.alaska, arr_status, sum(count), count / sum(count)))
```

```
## # A tibble: 5 x 6
## # Groups:   arr_status, sum(count), count/sum(count) [5]
##   airline arr_status airport        count ‘sum(count)‘ ‘count/sum(count)‘
##   <chr>   <chr>      <chr>          <dbl>        <dbl>              <dbl>
## 1 Alaska  delayed    Los Angeles       62          501              0.124
## 2 Alaska  delayed    Phoenix           12          501              0.0240
## 3 Alaska  delayed    San Diego         20          501              0.0399
## 4 Alaska  delayed    San Francisco    102          501              0.204
## 5 Alaska  delayed    Seattle          305          501              0.609
```

Use the rename() to tidy up the column names in the data.frame

```
(airline.alaska <- rename(airline.alaska,"Airline"="airline","Status"="arr_status","Location"="airport"
```

```
## # A tibble: 5 x 6
## # Groups:   Status, TotalDelayCount, PercentageDelay [5]
##   Airline Status  Location       DelayedCount TotalDelayCount PercentageDelay
##   <chr>   <chr>   <chr>                 <dbl>           <dbl>           <dbl>
## 1 Alaska  delayed Los Angeles              62             501          0.124
## 2 Alaska  delayed Phoenix                  12             501          0.0240
## 3 Alaska  delayed San Diego                20             501          0.0399
## 4 Alaska  delayed San Franscisco          102             501          0.204
## 5 Alaska  delayed Seattle                 305             501          0.609
```

Calculate the median and mean for delayed Alaska airlines delayed flights. For analytic purposes, I would probably use the median of 62 to determine reliability of Alaska airline arriving on time and from a performance standpoint. Although I would see the Seattle delay count as an outlier because it is 3 times larger than the next largest delay count, I would definitely use that indicator if I am flying into or departing from the Seattle airport.

```
(Delay.mean <- mean(airline.alaska$DelayedCount))
```

```
## [1] 100.2
```

```
(delay.median <- median(airline.alaska$DelayedCount))
```

```
## [1] 62
```

```
summary(airline.alaska)
```

```
##    Airline              Status             Location          DelayedCount
##  Length:5           Length:5           Length:5           Min.   : 12.0
##  Class :character   Class :character   Class :character   1st Qu.: 20.0
##  Mode  :character   Mode  :character   Mode  :character   Median : 62.0
##                                                           Mean   :100.2
##                                                           3rd Qu.:102.0
##                                                           Max.   :305.0
##  TotalDelayCount PercentageDelay
##  Min.   :501     Min.   :0.02395
##  1st Qu.:501     1st Qu.:0.03992
##  Median :501     Median :0.12375
##  Mean   :501     Mean   :0.20000
##  3rd Qu.:501     3rd Qu.:0.20359
##  Max.   :501     Max.   :0.60878
```

Perform the same cleansing, subsetting and calculations for Amwest Airlines.

```
(airline.amwest <- airline_delays %>% filter(airline == "Amwest"))
```

```
##   airline arr_status        airport count
## 1  Amwest    delayed    Los Angeles   117
## 2  Amwest    delayed        Phoenix   415
## 3  Amwest    delayed      San Diego    65
## 4  Amwest    delayed San Franscisco   129
## 5  Amwest    delayed        Seattle    61
```

```
(airline.amwest <- group_by(airline.amwest, arr_status, sum(count), count / sum(count)))
```

```
## # A tibble: 5 x 6
## # Groups:   arr_status, sum(count), count/sum(count) [5]
##   airline arr_status airport        count ‘sum(count)‘ ‘count/sum(count)‘
##   <chr>   <chr>      <chr>          <dbl>        <dbl>              <dbl>
## 1 Amwest  delayed    Los Angeles      117          787             0.149
## 2 Amwest  delayed    Phoenix          415          787             0.527
## 3 Amwest  delayed    San Diego         65          787             0.0826
## 4 Amwest  delayed    San Franscisco   129          787             0.164
## 5 Amwest  delayed    Seattle           61          787             0.0775
```

```
(airline.amwest <- rename(airline.amwest,"Airline"="airline","Status"="arr_status","Location"="airport"
```

```
## # A tibble: 5 x 6
## # Groups:   Status, TotalDelayCount, PercentageDelay [5]
##   Airline Status  Location       DelayedCount TotalDelayCount PercentageDelay
##   <chr>   <chr>   <chr>                 <dbl>           <dbl>           <dbl>
## 1 Amwest  delayed Los Angeles             117             787          0.149
## 2 Amwest  delayed Phoenix                 415             787          0.527
## 3 Amwest  delayed San Diego                65             787          0.0826
## 4 Amwest  delayed San Franscisco          129             787          0.164
## 5 Amwest  delayed Seattle                  61             787          0.0775
```

Calculate the median and mean for delayed Amwest airlines delayed flights. For analytic purposes, I would probably use the median of 62 to determine reliability of Amwest airlines arriving on time and from a performance standpoint. Although I would see the Phoenix delay count as an outlier because it is 3 times larger than the next largest delay count, I would definitely use that indicator if I am flying into or departing from the Phoenix airport. Phoenix is a major hub and a very busy airport.

```
(Delay.mean <- mean(airline.amwest$DelayedCount))
```

```
## [1] 157.4
```

```
(delay.median <- median(airline.amwest$DelayedCount))
```

```
## [1] 117
```

```
summary(airline.amwest)
```

```
##    Airline             Status            Location          DelayedCount
##  Length:5           Length:5           Length:5           Min.   : 61.0
##  Class :character   Class :character   Class :character   1st Qu.: 65.0
##  Mode  :character   Mode  :character   Mode  :character   Median :117.0
##                                                           Mean   :157.4
##                                                           3rd Qu.:129.0
##                                                           Max.   :415.0
##  TotalDelayCount PercentageDelay
##  Min.   :787     Min.   :0.07751
##  1st Qu.:787     1st Qu.:0.08259
##  Median :787     Median :0.14867
##  Mean   :787     Mean   :0.20000
##  3rd Qu.:787     3rd Qu.:0.16391
##  Max.   :787     Max.   :0.52732
```

```
total.alaska <- airline_tidy %>% filter(airline == "Alaska")
max(total.alaska$count)
```

```
## [1] 1841
```

```
total.amwest <- airline_tidy %>% filter(airline == "Amwest")
max(total.amwest$count)
```

```
## [1] 4840
```

Final thought on the performance of both airlines. While comparing the Alaska and Amwest airlines performance it may appear by their respective medians and means that Alaska experiences fewer delays than Amwest. However, when you look at the overall total flights for each airline individually Amwest flew over 2.5 times the number of flights into those locations.