flight times

Luke Yee

8/9/2020

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------------------------------- tidyver
```

```
## v ggplot2 3.3.1     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   1.0.0
## v tidyr   1.1.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------------------------------------ tidyverse_co
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(nycflights13)
```

I am conducting a basic analysis regarding the variation of flight times, demonstrating the extremely useful tidyverse package. The dataset I am using is called flights has roughly 336,000 observations and multiple columns with relevant information pertaining to flight times.

```r
glimpse(flights)
```

```
## Rows: 336,776
## Columns: 19
## $ year           <int> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013...
## $ month          <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ day            <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ dep_time       <int> 517, 533, 542, 544, 554, 554, 555, 557, 557, 558, 55...
## $ sched_dep_time <int> 515, 529, 540, 545, 600, 558, 600, 600, 600, 600, 60...
## $ dep_delay      <dbl> 2, 4, 2, -1, -6, -4, -5, -3, -3, -2, -2, -2, -2, -2,...
## $ arr_time       <int> 830, 850, 923, 1004, 812, 740, 913, 709, 838, 753, 8...
## $ sched_arr_time <int> 819, 830, 850, 1022, 837, 728, 854, 723, 846, 745, 8...
## $ arr_delay      <dbl> 11, 20, 33, -18, -25, 12, 19, -14, -8, 8, -2, -3, 7,...
## $ carrier        <chr> "UA", "UA", "AA", "B6", "DL", "UA", "B6", "EV", "B6"...
## $ flight         <int> 1545, 1714, 1141, 725, 461, 1696, 507, 5708, 79, 301...
## $ tailnum        <chr> "N14228", "N24211", "N619AA", "N804JB", "N668DN", "N...
## $ origin         <chr> "EWR", "LGA", "JFK", "JFK", "LGA", "EWR", "EWR", "LG...
## $ dest           <chr> "IAH", "IAH", "MIA", "BQN", "ATL", "ORD", "FLL", "IA...
## $ air_time       <dbl> 227, 227, 160, 183, 116, 150, 158, 53, 140, 138, 149...
## $ distance       <dbl> 1400, 1416, 1089, 1576, 762, 719, 1065, 229, 944, 73...
## $ hour           <dbl> 5, 5, 5, 5, 6, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 5, 6, 6...
## $ minute         <dbl> 15, 29, 40, 45, 0, 58, 0, 0, 0, 0, 0, 0, 0, 0, 0, 59...
## $ time_hour      <dttm> 2013-01-01 05:00:00, 2013-01-01 05:00:00, 2013-01-0...
```

First, I will add a column that shows the amount of time gained during air (the arrival - the departure delay), then sort the the data by the amount of gain time

```r
flights%>%mutate(gain=arr_delay-dep_delay)
```

```
## # A tibble: 336,776 x 20
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
##  1  2013     1     1      517            515         2      830            819
##  2  2013     1     1      533            529         4      850            830
##  3  2013     1     1      542            540         2      923            850
##  4  2013     1     1      544            545        -1     1004           1022
##  5  2013     1     1      554            600        -6      812            837
##  6  2013     1     1      554            558        -4      740            728
##  7  2013     1     1      555            600        -5      913            854
##  8  2013     1     1      557            600        -3      709            723
##  9  2013     1     1      557            600        -3      838            846
## 10  2013     1     1      558            600        -2      753            745
## # ... with 336,766 more rows, and 12 more variables: arr_delay <dbl>,
## #   carrier <chr>, flight <int>, tailnum <chr>, origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>,
## #   gain <dbl>
```

```r
flights2<-flights%>%mutate(gain=arr_delay-dep_delay)%>%arrange(desc(gain))
```

We can find out if flights lost or gained time on the average

```r
flights2%>%summarize(average=mean(gain,na.rm=TRUE))
```

```
## # A tibble: 1 x 1
##   average
##     <dbl>
## 1   -5.66
```

So on average, flights gained about 5.5 minutes in the air. We can also find the the average amount of time gained by planes coming from or going to a certain airport.

```r
flights2%>%filter(dest=="SEA")%>%summarize(average=mean(gain,na.rm=TRUE))
```

```
## # A tibble: 1 x 1
##   average
##     <dbl>
## 1   -11.7
```

```r
flights2%>%filter(origin=="JFK")%>%summarize(average=mean(gain,na.rm=TRUE))
```

```
## # A tibble: 1 x 1
##   average
##     <dbl>
## 1   -6.47
```

So flights heading to the Seattle-Tacoma airport lose about 11 minutes, and flights coming from JFK lose about 6 minutes. Besides this, we can find key information about specific routes as well.

```r
flights2%>%filter(origin=="JFK",dest=="SEA")%>%summarize(minimum=min(air_time,na.rm=TRUE),maximum=max(ai
```

```
## # A tibble: 1 x 3
##   minimum maximum average
##     <dbl>   <dbl>   <dbl>
## 1     275     389    329.
```

In this case, we see that the minimum flight time from JFK airport to was 275 minutes, the maximum was 389 minutes, and the average was 329 minutes. We can also sort delay times by month of the year

```r
slowest<-flights2%>%group_by(month)%>%summarize(delayance=mean(dep_delay,na.rm=TRUE)) %>% arrange(desc(
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```r
slowest
```

```
## # A tibble: 12 x 2
##    month delayance
##    <int>     <dbl>
##  1     7     21.7
##  2     6     20.8
##  3    12     16.6
##  4     4     13.9
##  5     3     13.2
##  6     5     13.0
##  7     8     12.6
##  8     2     10.8
##  9     1     10.0
## 10     9      6.72
## 11    10      6.24
## 12    11      5.44
```

Here we can see that July has the highest average delay times, while September has the lowest average delay times. Let's also find the airport that usually has the highest average delay times

```r
slowflight<-flights2%>%group_by(dest)%>%summarize(arrivaldelay=mean(arr_delay,na.rm=TRUE))%>%arrange(de
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```r
slowflight %>% head(5)
```

```
## # A tibble: 5 x 2
##   dest  arrivaldelay
##   <chr>        <dbl>
## 1 CAE           41.8
## 2 TUL           33.7
## 3 OKC           30.6
## 4 JAC           28.1
## 5 TYS           24.1
```

We can see Columbia Metropolitan Airport on average has the greatest delay times. Going in another direction, we can find the airports that flown to the fastest by creating another column called speed (distance flown/air time), grouping by destinations, and find the averages of those groups

```r
flights%>%mutate(speed=distance/air_time)%>%group_by(dest)%>%summarize(average=mean(speed,na.rm=TRUE))%
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 5 x 2
##   dest  average
##   <chr>   <dbl>
## 1 ANC      8.17
## 2 BQN      8.12
## 3 SJU      8.09
## 4 HNL      8.06
## 5 PSE      8.01
```

We see that Anchorage International Airport is the destination that is flown to with the greatest speed. Finally, lets try combining relevant data from this dataset with a similar dataset using a "join"

```
#first, using the flights data, we find information that we would like to
#be added to the similar dataset
avg_arrival_delay<-flights%>%group_by(dest)%>%summarize(delayal=mean(arr_delay, na.rm=TRUE))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
#we are combining average delay times by airport with another dataset called airports
#that has information containing latitudes, longitutes, etc. We join by columns that
#exist in both datasets
plane2<-left_join(avg_arrival_delay,airports,by=c("dest"="faa"))
plane2 %>% head(10)
```

```
## # A tibble: 10 x 9
##    dest  delayal name                     lat    lon   alt    tz dst   tzone
##    <chr>   <dbl> <chr>                  <dbl>  <dbl> <dbl> <dbl> <chr> <chr>
##  1 ABQ      4.38 Albuquerque Interna~  35.0 -107.   5355    -7 A     America/De~
##  2 ACK      4.85 Nantucket Mem         41.3  -70.1    48    -5 A     America/Ne~
##  3 ALB     14.4  Albany Intl           42.7  -73.8   285    -5 A     America/Ne~
##  4 ANC     -2.5  Ted Stevens Anchora~  61.2 -150.    152    -9 A     America/An~
##  5 ATL     11.3  Hartsfield Jackson ~  33.6  -84.4  1026    -5 A     America/Ne~
##  6 AUS      6.02 Austin Bergstrom In~  30.2  -97.7   542    -6 A     America/Ch~
##  7 AVL      8.00 Asheville Regional ~  35.4  -82.5  2165    -5 A     America/Ne~
##  8 BDL      7.05 Bradley Intl          41.9  -72.7   173    -5 A     America/Ne~
##  9 BGR      8.03 Bangor Intl           44.8  -68.8   192    -5 A     America/Ne~
## 10 BHM     16.9  Birmingham Intl       33.6  -86.8   644    -6 A     America/Ch~
```