Tutorial 2

Rebecca Letsinger

February 20, 2024

Step 1: In this step, I installed the packages and opened each library

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(nycflights13)
```

Step 2: In this step, I filtered through the data set to specifically focus on flights within United Airlines.

```
united_flights <- flights %>%
  filter(carrier == "UA")
```

Step 3: In this step, I used the used the same filtered data set using United Airlines flights. The origin function was added to restrict the data to flights that specifically departed from Laguardia. Both sets of code gave me the same information.

```
united_flights <- flights %>%
  filter(carrier== "UA"&(origin=="LGA"))
glimpse(united_flights)
```

```
## Rows: 8,044
## Columns: 19
## $ year           <int> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2~
## $ month          <int> 1, 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, 1~
## $ dep_time       <int> 533, 623, 646, 709, 728, 752, 931, 1028, 1114, 1123, 11~
## $ sched_dep_time <int> 529, 627, 645, 700, 732, 750, 930, 1026, 900, 1110, 120~
## $ dep_delay      <dbl> 4, -4, 1, 9, -4, 2, 1, 2, 134, 13, -2, -2, 2, 0, -4, -6~
## $ arr_time       <int> 850, 933, 910, 852, 1041, 1025, 1121, 1350, 1447, 1410,~
## $ sched_arr_time <int> 830, 932, 916, 832, 1038, 1029, 1108, 1339, 1222, 1336,~
## $ arr_delay      <dbl> 20, 1, -6, 20, 3, -4, 13, 11, 145, 34, 7, 0, 11, -9, 17~
## $ carrier        <chr> "UA", "UA", "UA", "UA", "UA", "UA", "UA", "UA", "UA", "~
## $ flight         <int> 1714, 496, 883, 1092, 473, 477, 255, 1004, 1086, 405, 7~
## $ tailnum        <chr> "N24211", "N459UA", "N569UA", "N26226", "N488UA", "N511~
## $ origin         <chr> "LGA", "LGA", "LGA", "LGA", "LGA", "LGA", "LGA", "LGA",~
## $ dest           <chr> "IAH", "IAH", "DEN", "ORD", "IAH", "DEN", "ORD", "IAH",~
## $ air_time       <dbl> 227, 229, 243, 135, 238, 249, 154, 237, 248, 256, 142, ~
## $ distance       <dbl> 1416, 1416, 1620, 733, 1416, 1620, 733, 1416, 1416, 162~
## $ hour           <dbl> 5, 6, 6, 7, 7, 7, 9, 10, 9, 11, 12, 12, 13, 14, 15, 15,~
## $ minute         <dbl> 29, 27, 45, 0, 32, 50, 30, 26, 0, 10, 0, 50, 54, 30, 0,~
## $ time_hour      <dttm> 2013-01-01 05:00:00, 2013-01-01 06:00:00, 2013-01-01 0~
```

```r
united_flights <- flights %>%
  filter(carrier == "UA") %>%
  filter(origin=="LGA")
glimpse(united_flights)
```

```
## Rows: 8,044
## Columns: 19
## $ year           <int> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2~
## $ month          <int> 1, 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, 1~
## $ dep_time       <int> 533, 623, 646, 709, 728, 752, 931, 1028, 1114, 1123, 11~
## $ sched_dep_time <int> 529, 627, 645, 700, 732, 750, 930, 1026, 900, 1110, 120~
## $ dep_delay      <dbl> 4, -4, 1, 9, -4, 2, 1, 2, 134, 13, -2, -2, 2, 0, -4, -6~
## $ arr_time       <int> 850, 933, 910, 852, 1041, 1025, 1121, 1350, 1447, 1410,~
## $ sched_arr_time <int> 830, 932, 916, 832, 1038, 1029, 1108, 1339, 1222, 1336,~
## $ arr_delay      <dbl> 20, 1, -6, 20, 3, -4, 13, 11, 145, 34, 7, 0, 11, -9, 17~
## $ carrier        <chr> "UA", "UA", "UA", "UA", "UA", "UA", "UA", "UA", "UA", "~
## $ flight         <int> 1714, 496, 883, 1092, 473, 477, 255, 1004, 1086, 405, 7~
## $ tailnum        <chr> "N24211", "N459UA", "N569UA", "N26226", "N488UA", "N511~
## $ origin         <chr> "LGA", "LGA", "LGA", "LGA", "LGA", "LGA", "LGA", "LGA",~
## $ dest           <chr> "IAH", "IAH", "DEN", "ORD", "IAH", "DEN", "ORD", "IAH",~
## $ air_time       <dbl> 227, 229, 243, 135, 238, 249, 154, 237, 248, 256, 142, ~
## $ distance       <dbl> 1416, 1416, 1620, 733, 1416, 1620, 733, 1416, 1416, 162~
## $ hour           <dbl> 5, 6, 6, 7, 7, 7, 9, 10, 9, 11, 12, 12, 13, 14, 15, 15,~
## $ minute         <dbl> 29, 27, 45, 0, 32, 50, 30, 26, 0, 10, 0, 50, 54, 30, 0,~
## $ time_hour      <dttm> 2013-01-01 05:00:00, 2013-01-01 06:00:00, 2013-01-01 0~
```

Step 4: In this step, the data was restricted even more by only including flights that departed from Laguardia and arrived in Orlando or Denver. I used the “or” function to include both arrival cities.

```r
united_flights <- flights %>%
  filter(carrier == "UA") %>%
  filter(origin=="LGA") %>%
  filter(dest == "ORD" | dest == "DEN")
```

Step 5: In this step, I used the similar format to the previous question to include all 4 cities.

```r
many_airports <- flights %>%
  filter(carrier == "UA") %>%
  filter(origin=="LGA") %>%
  filter(dest %in% c("IAH", "CLE", "ORD", "DEN"))
```

Step 6: This step summarized the mean and standard deviation for the arrival delays all the flight data. The second step omitted the missing observations to see the correct results.

```
summary_airports <- flights %>%
  summarize(mean = mean(arr_delay), std_dev = sd(arr_delay))
summary_airports
```

```
## # A tibble: 1 x 2
##    mean std_dev
##   <dbl>   <dbl>
## 1    NA      NA
```

```
summary_airports <- flights %>%
  summarize(mean = mean(arr_delay, na.rm = TRUE),   std_dev = sd(arr_delay, na.rm = TRUE))
summary_airports
```

```
## # A tibble: 1 x 2
##    mean std_dev
##   <dbl>   <dbl>
## 1  6.90    44.6
```

```
library(knitr)
kable(summary_airports)
```

mean	std_dev
6.895377	44.63329

Step 7: This step puts all of our flight delay information into one clear table.

```
summary_airports <- flights %>%
  summarize(mean = mean(arr_delay, na.rm = TRUE),
  std_dev = sd(arr_delay, na.rm = TRUE),
  count= n())
summary_airports
```

```
## # A tibble: 1 x 3
##    mean std_dev  count
##   <dbl>   <dbl>  <int>
## 1  6.90    44.6 336776
```

```
kable(summary_airports)
```

mean	std_dev	count
6.895377	44.63329	336776

Group 8: This step takes out delay data set and breaks it down by month

```
by_month <- flights %>%
  group_by(month) %>%
  summarize(mean = mean(arr_delay, na.rm = TRUE),
  std_dev = sd(arr_delay, na.rm = TRUE),
  count= n())
by_month
```

```
## # A tibble: 12 x 4
##    month   mean std_dev count
##    <int>  <dbl>   <dbl> <int>
##  1     1  6.13     40.4 27004
##  2     2  5.61     39.5 24951
##  3     3  5.81     44.1 28834
##  4     4 11.2      47.5 28330
##  5     5  3.52     44.2 28796
##  6     6 16.5      56.1 28243
##  7     7 16.7      57.1 29425
##  8     8  6.04     42.6 29327
##  9     9 -4.02     39.7 27574
## 10    10 -0.167    32.6 28889
## 11    11  0.461    31.4 27268
## 12    12 14.9      46.1 28135
```

```
kable(by_month)
```

month	mean	std_dev	count
1	6.1299720	40.42390	27004
2	5.6130194	39.52862	24951
3	5.8075765	44.11919	28834
4	11.1760630	47.49115	28330
5	3.5215088	44.23761	28796
6	16.4813296	56.13087	28243
7	16.7113067	57.11709	29425
8	6.0406524	42.59514	29327
9	-4.0183636	39.71031	27574
10	-0.1670627	32.64986	28889
11	0.4613474	31.38741	27268
12	14.8703553	46.13311	28135

Step 9: This step breaks the data down by month and adds additional variables

```
by_origin_month <- flights %>%
  group_by(origin, month) %>%
  summarize(count = n())
```

```
## `summarise()` has grouped output by 'origin'. You can override using the
## `.groups` argument.
```

```
by_origin_month
```

```
## # A tibble: 36 x 3
## # Groups:   origin [3]
##    origin month count
##    <chr>  <int> <int>
##  1 EWR        1  9893
##  2 EWR        2  9107
##  3 EWR        3 10420
##  4 EWR        4 10531
##  5 EWR        5 10592
##  6 EWR        6 10175
##  7 EWR        7 10475
##  8 EWR        8 10359
##  9 EWR        9  9550
## 10 EWR       10 10104
## # i 26 more rows
```

```
kable(by_origin_month)
```

origin	month	count
EWR	1	9893
EWR	2	9107
EWR	3	10420
EWR	4	10531
EWR	5	10592
EWR	6	10175
EWR	7	10475
EWR	8	10359
EWR	9	9550
EWR	10	10104
EWR	11	9707
EWR	12	9922
JFK	1	9161
JFK	2	8421
JFK	3	9697
JFK	4	9218
JFK	5	9397
JFK	6	9472
JFK	7	10023
JFK	8	9983
JFK	9	8908
JFK	10	9143
JFK	11	8710
JFK	12	9146
LGA	1	7950
LGA	2	7423
LGA	3	8717
LGA	4	8581
LGA	5	8807
LGA	6	8596

origin	month	count
LGA	7	8927
LGA	8	8985
LGA	9	9116
LGA	10	9642
LGA	11	8851
LGA	12	9067