Lab 3

Brandon Leslie

Lab 3

Exercise 1

Create a new dataset that only contains flights that do not have a missing departure time. Include the columns year, month, day, dep_time, dep_delay, and dep_delay_hours

```
# A tibble: 328,521 x 6
    year month   day dep_time dep_delay dep_delay_hours
   <int> <int> <int>    <int>     <dbl>           <dbl>
 1  2013     1     9      641      1301            21.7
 2  2013     6    15     1432      1137            19.0
 3  2013     1    10     1121      1126            18.8
 4  2013     9    20     1139      1014            16.9
 5  2013     7    22      845      1005            16.8
 6  2013     4    10     1100       960            16  
 7  2013     3    17     2321       911            15.2
 8  2013     6    27      959       899            15.0
 9  2013     7    22     2257       898            15.0
10  2013    12     5      756       896            14.9
# i 328,511 more rows
```

Exerxise 2

For each airplane (uniquely identified by tailnum), use a group_by() paired with summarize() to find the sample size, mean, and standard deviation of flight distances. Then include only the top 5 and bottom 5 airplanes in terms of mean distance traveled per flight in the final data frame.

Top 5 Rows

```
# A tibble: 5 x 4
  tailnum Distance_Mean Distance_Standard_Deviation Distance_Sample_Size
  <chr>           <dbl>                       <dbl>                <int>
1 D942DN           854.                      107.                      4
2 N0EGMQ           676.                      200.                    371
3 N10156           758.                      332.                    153
4 N102UW           536.                        6.75                   48
5 N103US           535.                        6.62                   46
```

Bottom 5 rows

```
# A tibble: 5 x 4
  tailnum Distance_Mean Distance_Standard_Deviation Distance_Sample_Size
  <chr>           <dbl>                       <dbl>                <int>
1 N103US           535.                        6.62                   46
2 N102UW           536.                        6.75                   48
3 N10156           758.                      332.                    153
4 N0EGMQ           676.                      200.                    371
5 D942DN           854.                      107.                      4
```

Exercise 3

Exercise: Find the maximum arrival delay

```
# A tibble: 1 x 3
  flight arr_delay     n
   <int>     <dbl> <int>
1     51      1272     1
```

Exercise 4

4A

Flights that flew to Portland (PWM)

```
# A tibble: 2,352 x 2
   dest  tailnum
   <chr> <chr>
 1 PWM   N306JB
 2 PWM   N11544
 3 PWM   N216JB
 4 PWM   N11544
 5 PWM   N13988
 6 PWM   N16561
 7 PWM   N279JB
 8 PWM   N14993
 9 PWM   N353JB
10 PWM   N13903
# i 2,342 more rows
```

4B

Flights with arrival delay exceeding 2 hrs.

```
# A tibble: 123,096 x 2
   tailnum arr_delay
   <chr>       <dbl>
 1 N14228         11
 2 N24211         20
 3 N619AA         33
 4 N39463         12
 5 N516JB         19
 6 N3ALAA          8
 7 N29129          7
 8 N3DUAA         31
 9 N542MQ         12
10 N730MQ         16
# i 123,086 more rows
```

4C

Had arrival delay more than 2 hrs, but did not depart late.

```
# A tibble: 34,583 x 3
   tailnum arr_delay dep_delay
   <chr>       <dbl>     <dbl>
```

```
 1 N39463          12         -4
 2 N516JB          19         -5
 3 N3ALAA           8         -2
 4 N29129           7         -2
 5 N3DUAA          31         -1
 6 N542MQ          12          0
 7 N730MQ          16         -3
 8 N807AW           3         -8
 9 N11107          29         -6
10 N518MQ          10         -6
# i 34,573 more rows
```

4D

How many flights have missing arrival time? What other variables are missing?

Missing Arrival Time

```
# A tibble: 1 x 1
      n
  <int>
1  8713
```

Variables With NA values.

```
# A tibble: 8,713 x 19
    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     2016           1930        46       NA           2220
 2  2013     1     1       NA           1630        NA       NA           1815
 3  2013     1     1       NA           1935        NA       NA           2240
 4  2013     1     1       NA           1500        NA       NA           1825
 5  2013     1     1       NA            600        NA       NA            901
 6  2013     1     2     2041           2045        -4       NA           2359
 7  2013     1     2     2145           2129        16       NA             33
 8  2013     1     2       NA           1540        NA       NA           1747
 9  2013     1     2       NA           1620        NA       NA           1746
10  2013     1     2       NA           1355        NA       NA           1459
# i 8,703 more rows
# i 11 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>
```

dep_time, dep_delay, arr_delay, arr_time, air_time, & dep_time. are the variables with NA values. This bit of code: *filter(flights, !is.na(arr_time))* returns all values of arr_time that aren't NA. Here is an example:

```
# A tibble: 328,063 x 19
    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
# i 328,053 more rows
# i 11 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>
```

Exercise 5

Suppose we want to see how much time was gained in the air. We would want to subtract arr_delay-dep_delay. Create a data frame that represents the difference between arrival and departure delays. What is the average gain for all flights?

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

The average gain for all flights is roughly **-5.66.**

5A

Are some airlines better than others regarding delays?

```
# A tibble: 16 x 7
   carrier avg_arrival_delay avg_dep_delay avg_dist_flown avg_arr_time
   <chr>               <dbl>         <dbl>          <dbl>        <dbl>
 1 9E                  7.38          16.7             530.        1639.
 2 AA                  0.364          8.59           1340.        1521.
 3 AS                 -9.93           5.80           2402         1565.
 4 B6                  9.46          13.0            1069.        1406.
 5 DL                  1.64           9.26           1237.        1573.
 6 EV                 15.8           20.0             563.        1488.
 7 F9                 21.9           20.2            1620         1672.
 8 FL                 20.1           18.7             665.        1574.
 9 HA                 -6.92           4.90           4983         1474.
10 MQ                 10.8           10.6             570.        1551.
11 OO                 11.9           12.6             501.        1913.
12 UA                  3.56          12.1            1529.        1509.
13 US                  2.13           3.78            553.        1402.
14 VX                  1.76          12.9            2499.        1523.
15 WN                  9.65          17.7             996.        1443.
16 YV                 15.6           19.0             375.        1761.
# i 2 more variables: avg_dep_time <dbl>, avg_air_time <dbl>
```

We may also want to study whether delay times have anything to do with things like distance flown, destination, etc.

Yes, there are some airlines that have better times. While for some variables like *avg_arr_time* or *avg_dep_time* where there is little difference between times, this difference is seen more in the delays.

Exercise 6

6A

Which airlines have the greatest mean departure and arrival delays?

```
# A tibble: 16 x 3
   carrier mean_dep_delay mean_arr_delay
   <chr>            <dbl>          <dbl>
```

```
 1 F9                20.2          21.9
 2 EV                20.0          15.8
 3 YV                19.0          15.6
 4 FL                18.7          20.1
 5 WN                17.7           9.65
 6 9E                16.7           7.38
 7 B6                13.0           9.46
 8 VX                12.9           1.76
 9 OO                12.6          11.9
10 UA                12.1           3.56
11 MQ                10.6          10.8
12 DL                 9.26          1.64
13 AA                 8.59          0.364
14 AS                 5.80         -9.93
15 HA                 4.90         -6.92
16 US                 3.78          2.13
```

The top 3 airlines with greatest mean delays are *F9*, *EV*, and *YV*.

6B

Which airlines have the most flights?

```
# A tibble: 16 x 2
   carrier flights_Per_Carrier
   <chr>                 <int>
 1 UA                    58665
 2 B6                    54635
 3 EV                    54173
 4 DL                    48110
 5 AA                    32729
 6 MQ                    26397
 7 US                    20536
 8 9E                    18460
 9 WN                    12275
10 VX                     5162
11 FL                     3260
12 AS                      714
13 F9                      685
14 YV                      601
15 HA                      342
16 OO                       32
```

The top 3 airlines with the most flights are *UA*, *B6*, and *EV*.

6C

How many planes are there? Find the total miles flown by each plane.

These are the # of miles flown for each plain, with the total # of planes adding up to exactly **336,776**

```
# A tibble: 1 x 1
  True_Plane_Sum
           <int>
1         336776
```

6D

Find which airlines fly to which destinations and give count of number of flights

I cannot create a graph that represents this clearly, nor do I know how to attempt this question. I tried creating a graph and it turns out to be rubbish, and I've tried creating a table but I can't use it to explain myself.

6E

Find which airlines fly to Honolulu (HNL)? Honalulu and Ankorage?

```
# A tibble: 2 x 3
# Groups:   carrier [2]
  carrier dest  flights_num
  <chr>   <chr>       <int>
1 HA      HNL           342
2 UA      HNL           365
```

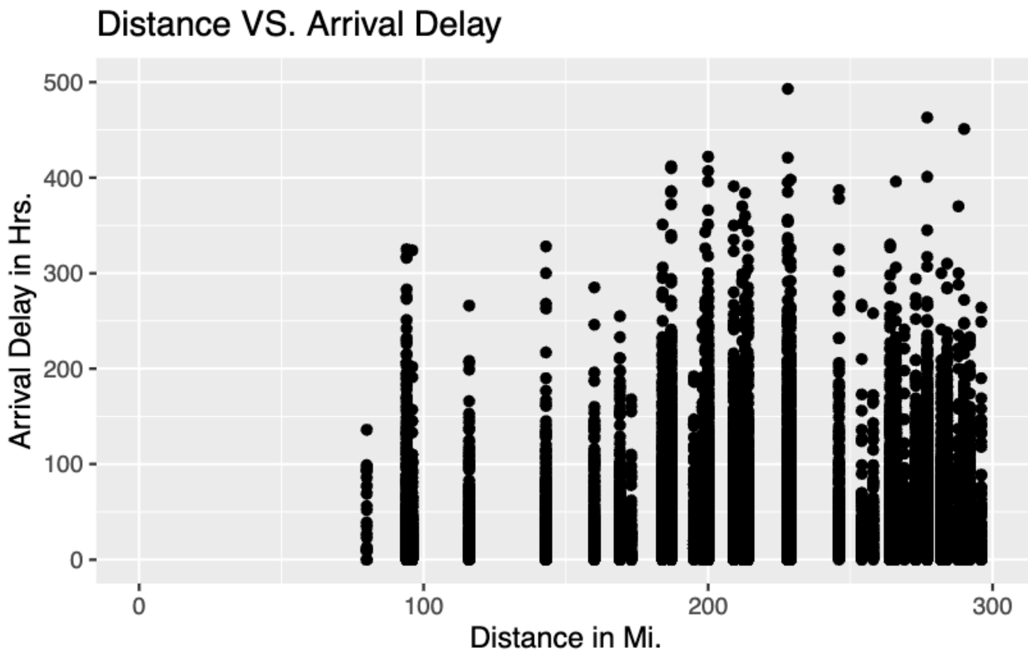
The airlines flying to *Honolulu* are *UA* and *HA*.

```
# A tibble: 1 x 3
# Groups:   carrier [1]
  carrier dest  flights_num
  <chr>   <chr>       <int>
1 UA      ANC             8
```

the airline flying to *Ankorage* is *UA*.

6F

Study the relationship between distance and arrival delay with a suitable plot. We might want to filter out outliers.



The relationship between Distance and Arrival delay is directly positive. As *distance* increases, *arrival delay* increases as well. While this may not seem to be the case at first glance, there are an upwards of 300,000 flights,compared to the handful of outliers present within the plot. This means that the mean and/or median will be widly different when compared to the uneducated first guess.