Data manipulation assignment

William Ou

5/31/2020

Data set

- load appropriate libraries

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

- install the package "nycflights13" and load it

```
install.packages("nycflights63")
library(nycflights63)
```

- once loaded, you will be able to directly call a dataframe object called *flights*

```
head(flights)
```

```
## # A tibble: 6 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
## # ... with 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>
```

- inspect the data set as you would. You can also read the help file for more details (ie. ?flights)

Use data manipultation functions (or other methods you like) to answer the following questions:

Question 1.

a) Which NYC airport has the most flights?
b) Which NYC airport flies to the most destinations?
c) BONUS: What are the top 3 destinations of each airport

- There are 3 NYC airports : EWR, LGA, JFK

```
unique(flights$origin)
```

```
## [1] "EWR" "LGA" "JFK"
```

- Suggested functions to use: **group_by()**, **summarise()**, **n()**, **distinct()**
- You can always check out the help files of those functions to see how they work (or check the cheatsheet!)
- the function **arrange()** may also be helpful when you want to order your values
- For the bonus section, you may consider using the **pivot_wider()** function so that each airport contains its own column (and then you can call arrange() on each column to get the top 3)

Question 2. For simplicity, let's assume that delay means that there are delays in both arrival and departure (ie. arr_delay>0 & dep_delay>0).

a) Which airport has the "MOST" (ie. frequency) delays?
b) Does the ranking of a) change after dividing by the number of flight for each airport (1a)?
c) On average, which carrier has the "LONGEST" (ie. duration) delays (add arrival and departure delays together)?

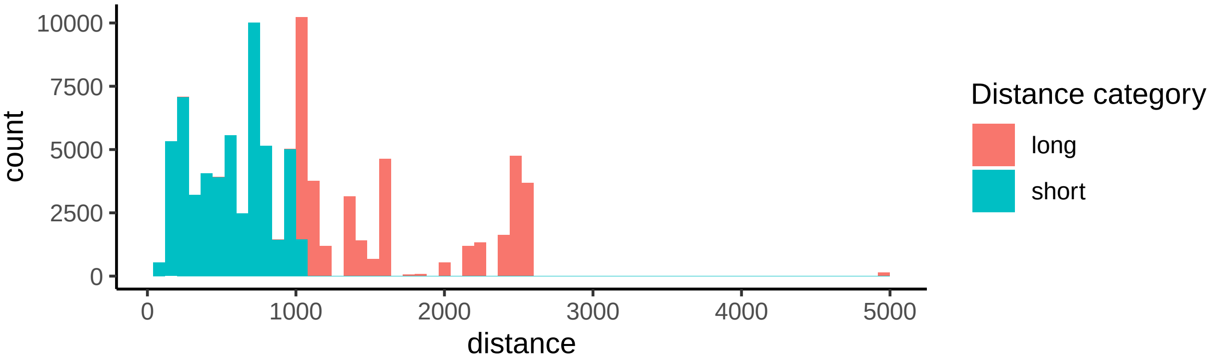
- you can use the **filter()** function to get observations where arr_delay and dep_delay are both >0

- example:

```
flights%>%
filter(arr_delay>0&dep_delay>0)
```

Question 3: Using the overall mean, convert travel distances into 2 distance categories (ie. longer or shorter than average).Do departure or arrival delay times differ betweeen distance categories?

- like in Q2, filter out non-delays observations


count
distance
Distance category
long
short