New York Air Traffic

Team ?, Abby Castro, Qinyuan Jiang, Angela Wan, Alexandra Labus, Imani Hankison, Julia Sharff

Please load the nycflights13 package and other necessary packages and follow the following steps.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.6     v dplyr   1.0.3
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

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

```
library(nycflights13)
```

Step 1. Describe the main question you are testing.

Are delays in EWR worse than the delays at the NY airports?

Step 2. Identify the variables that are relevant to the question.

- What are the types of those variables? How do you determine that?

```
str(flights)
```

```
## tibble [336,776 x 19] (S3: tbl_df/tbl/data.frame)
##  $ year          : int [1:336776] 2013 2013 2013 2013 2013 2013 2013 2013 2013 2013 ...
##  $ month         : int [1:336776] 1 1 1 1 1 1 1 1 1 1 ...
##  $ day           : int [1:336776] 1 1 1 1 1 1 1 1 1 1 ...
##  $ dep_time      : int [1:336776] 517 533 542 544 554 554 555 557 557 558 ...
##  $ sched_dep_time: int [1:336776] 515 529 540 545 600 558 600 600 600 600 ...
##  $ dep_delay     : num [1:336776] 2 4 2 -1 -6 -4 -5 -3 -3 -2 ...
##  $ arr_time      : int [1:336776] 830 850 923 1004 812 740 913 709 838 753 ...
##  $ sched_arr_time: int [1:336776] 819 830 850 1022 837 728 854 723 846 745 ...
##  $ arr_delay     : num [1:336776] 11 20 33 -18 -25 12 19 -14 -8 8 ...
##  $ carrier       : chr [1:336776] "UA" "UA" "AA" "B6" ...
```

```
##  $ flight        : int [1:336776] 1545 1714 1141 725 461 1696 507 5708 79 301 ...
##  $ tailnum       : chr [1:336776] "N14228" "N24211" "N619AA" "N804JB" ...
##  $ origin        : chr [1:336776] "EWR" "LGA" "JFK" "JFK" ...
##  $ dest          : chr [1:336776] "IAH" "IAH" "MIA" "BQN" ...
##  $ air_time      : num [1:336776] 227 227 160 183 116 150 158 53 140 138 ...
##  $ distance      : num [1:336776] 1400 1416 1089 1576 762 ...
##  $ hour          : num [1:336776] 5 5 5 5 6 5 6 6 6 6 ...
##  $ minute        : num [1:336776] 15 29 40 45 0 58 0 0 0 0 ...
##  $ time_hour     : POSIXct[1:336776], format: "2013-01-01 05:00:00" "2013-01-01 05:00:00" ...
```

```
flights
```

```
## # A tibble: 336,776 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
## # ... with 336,766 more rows, and 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>
```

origin: discrete (chr) dep_delay: continuous(num) arr_delay: continuous(num)

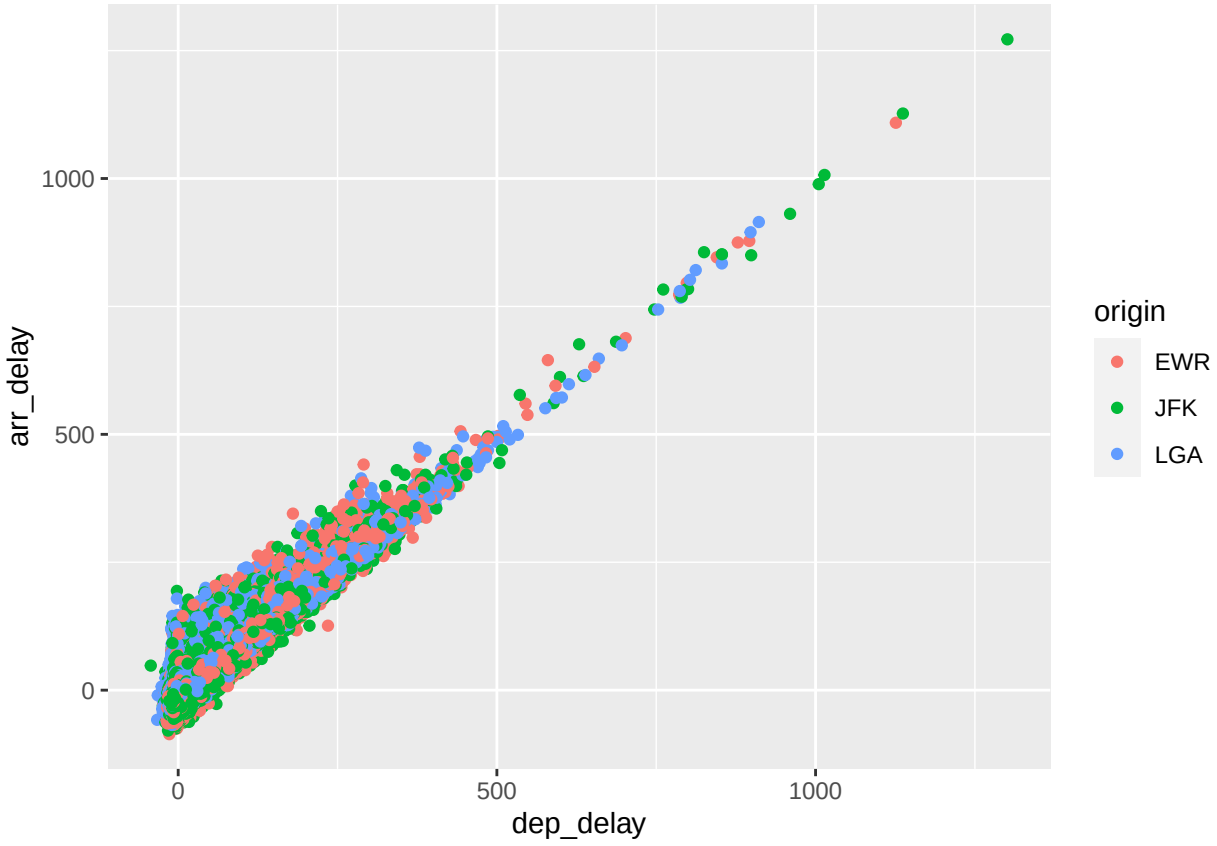
- Describe why those variables may be relevant to this question and why other variables are not relevant
  The origin is relevant to the question because we are testing if the origin is related to the delay. The dep_delay and arr_delay is relevant to the question because that is the measure we will use for comparisons between the origin airports.

Step 3. Search for evidence by visualising, transforming, and modeling your data

(Check RDS 3, 5, 7.3, 7.4, 7.5, 7.6 for ideas and inspiration)

```
ggplot(flights, aes(dep_delay, arr_delay, color = origin)) +
  geom_point()
```

```
## Warning: Removed 9430 rows containing missing values (geom_point).
```



3.1 What type of variation occurs within each variable?

Pick one variable and test the following:

3.1.1 Variable 1 (replace with the name of the variable)

3.1.1.1 Visualising distributions (Barcharts, Histograms, etc.) (RDS 7, RDS 3)

- What chart is appropriate for this variable? Why?
- Which values are the most common? Why?
- Which values are rare? Why? Does that match your expectations?
- Can you see any unusual patterns? What might explain them?

3.1.1.2 Unusual values (RDS 7, 5.2)

- Describe and demonstrate how you determine if there are unusual values in the data. E.g. too large, too small, negative, etc.
- Describe and demonstrate how you determine if they are outliers.
- Show how do your distributions look like with and without the unusual values.
- Discuss whether or not you need to remove unusual values and why.

3.1.1.3 Missing values (RDS 5.2.3)

- Does this variable include missing values? Demonstrate how you determine that.
- Demonstrate and discuss how you handle the missing values. E.g., removing, replacing with a constant value, or a value based on the distribution, etc.
- Show how your data looks in each case after handling missing values. Describe and discuss the distribution.

3.1.1.4 Does converting the type of this variable help exploring the distribution of its values or identifying outliers or missing values? (RDS 7)

- What type can the variable be converted to?
- How will the distribution look? Please demonstrate with appropriate plots.

3.1.1.5 What new variables do you need to create from this? (RDS 5.5, 5.6, 5.7)

- List the variables
- Describe and discuss why they are needed and how you plan to use them.

3.2. What type of covariation occurs between the two variables? (RDS 7)

3.2.1 Between a categorical and continuous variable or between two categorical variables or between two continuous variables

- Describe what type of visualization you can use and why.
- Describe the patterns and relationships you observe. Could the identified patterns be due to coincidence (i.e. random chance)?
- Describe the relationship implied by the pattern? (e.g., positive or negative correlation)
- Calculate the strength of the relationship implied by the pattern (e.g., correlation)
- Discuss how the observed patterns support/reject your hypotheses or answer your questions.

Step 4. Summarize your findings

- Summarize your findings about the questions you asked at the beginning.
- Discuss if you have enough evidence to make a conclusion about your analysis.