New York Air Traffic

Team 6, Qinyuan Jiang, Imani Hankinson, Abigail Castro, Angela Wan, Alexandra Labus, Julia Sharff

```
knitr::opts_chunk$set(echo = TRUE)
```

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

```
library(nycflights13)
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(dplyr)
```

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? origin - discrete (chr); dep_delay - continuous (int); arr_delay - continuous (int).
- Describe why those variables may be relevant to this question and why other variables are not relevant these variables show the relationship between the airports and how their delays may be different from each other.

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) Use a plot to see where each airport's delay is distributed.

3.1 What type of variation occurs within each variable? -Abby

Pick one variable and test the following:

```
flights %>%
    group_by(origin) %>% summarize(mean_dep_delay = mean(dep_delay,na.rm=TRUE),
                                   med_dep_delay = median(dep_delay,na.rm=TRUE),
                                   mean_arr_delay = mean(arr_delay,na.rm=TRUE),
                                   med_arr_delay = median(arr_delay,na.rm=TRUE))
```

```
## # A tibble: 3 x 5
##   origin mean_dep_delay med_dep_delay mean_arr_delay med_arr_delay
## * <chr>           <dbl>         <dbl>          <dbl>         <dbl>
## 1 EWR              15.1            -1           9.11            -4
## 2 JFK              12.1            -1           5.55            -6
## 3 LGA              10.3            -3           5.78            -5
```

From this data, it can be concluded that the delays are skewed right and that EWR has a longer average delay.

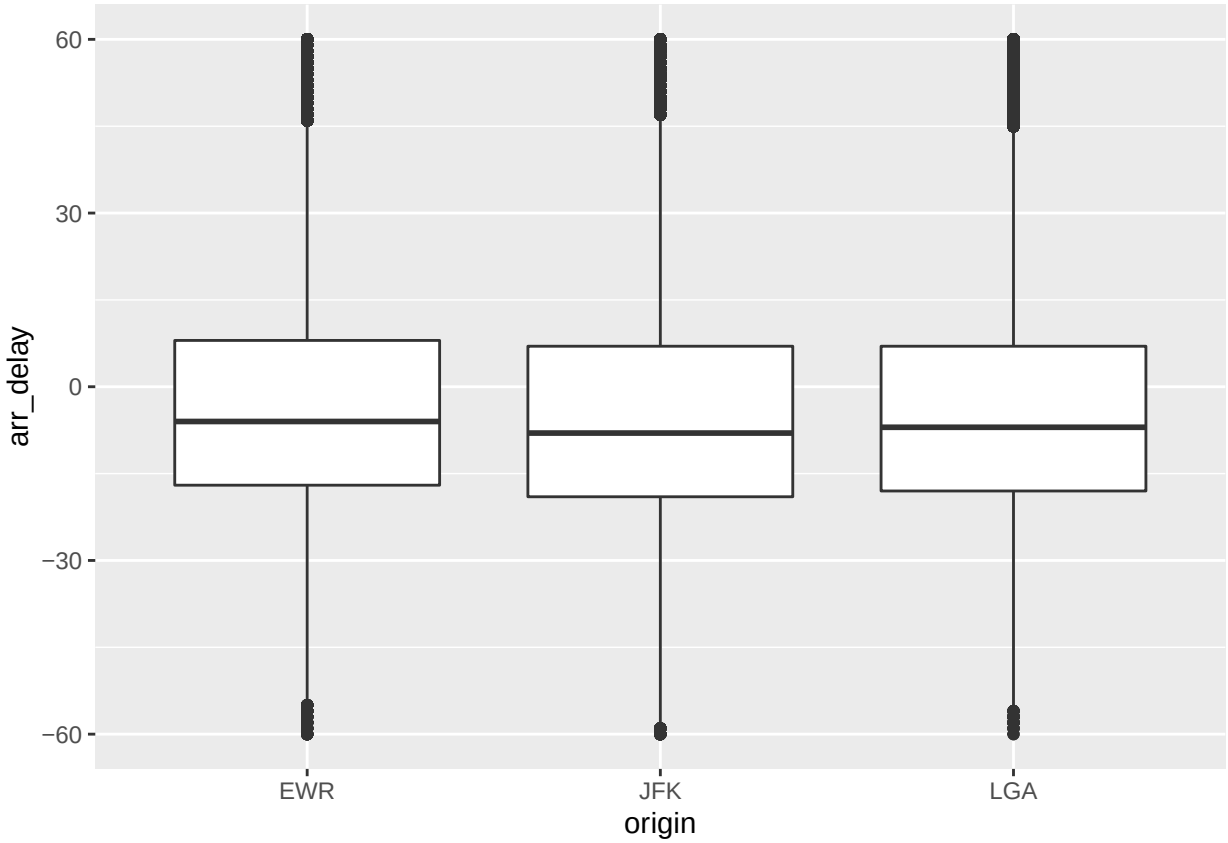
3.1.1 Variable 1 (replace with the name of the variable) - Qinyuan

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

- What chart is appropriate for this variable? Why? Histogram because it shows the variation between different airports and how they might be distributed differently from one another.
- Which values are the most common? Why? Shorter delays are more common and this is expected because flights are more likely to arrive on time or shortly after the expected arrival/departure.
- Which values are rare? Why? Does that match your expectations? Longer delays are rare because more flights are expected to
- Can you see any unusual patterns? What might explain them? Although the data should strictly decrease as the arr_delay and dep_delay increases towards infinity, there are times when the histogram count increases for the next bin.

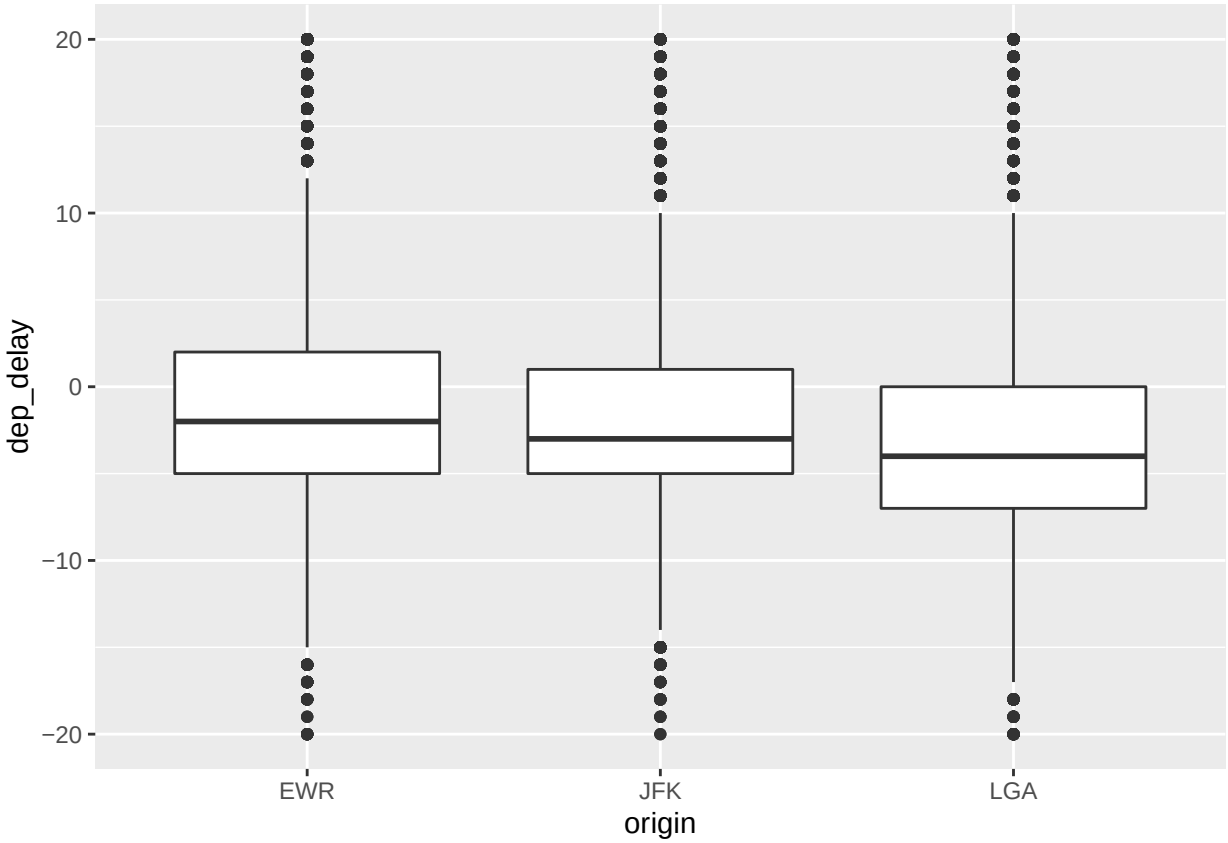
```
ggplot(data = flights, mapping = aes(x = origin, y = arr_delay)) +
  geom_boxplot() +
  ylim(-60,60)
```

```
## Warning: Removed 37418 rows containing non-finite values (stat_boxplot).
```




```
ggplot(data = flights, mapping = aes(x = origin, y = dep_delay)) +
  geom_boxplot() +
  ylim(-20,20)
```

```
## Warning: Removed 69929 rows containing non-finite values (stat_boxplot).
```



3.1.1.2 Unusual values (RDS 7, 5.2) - Angela

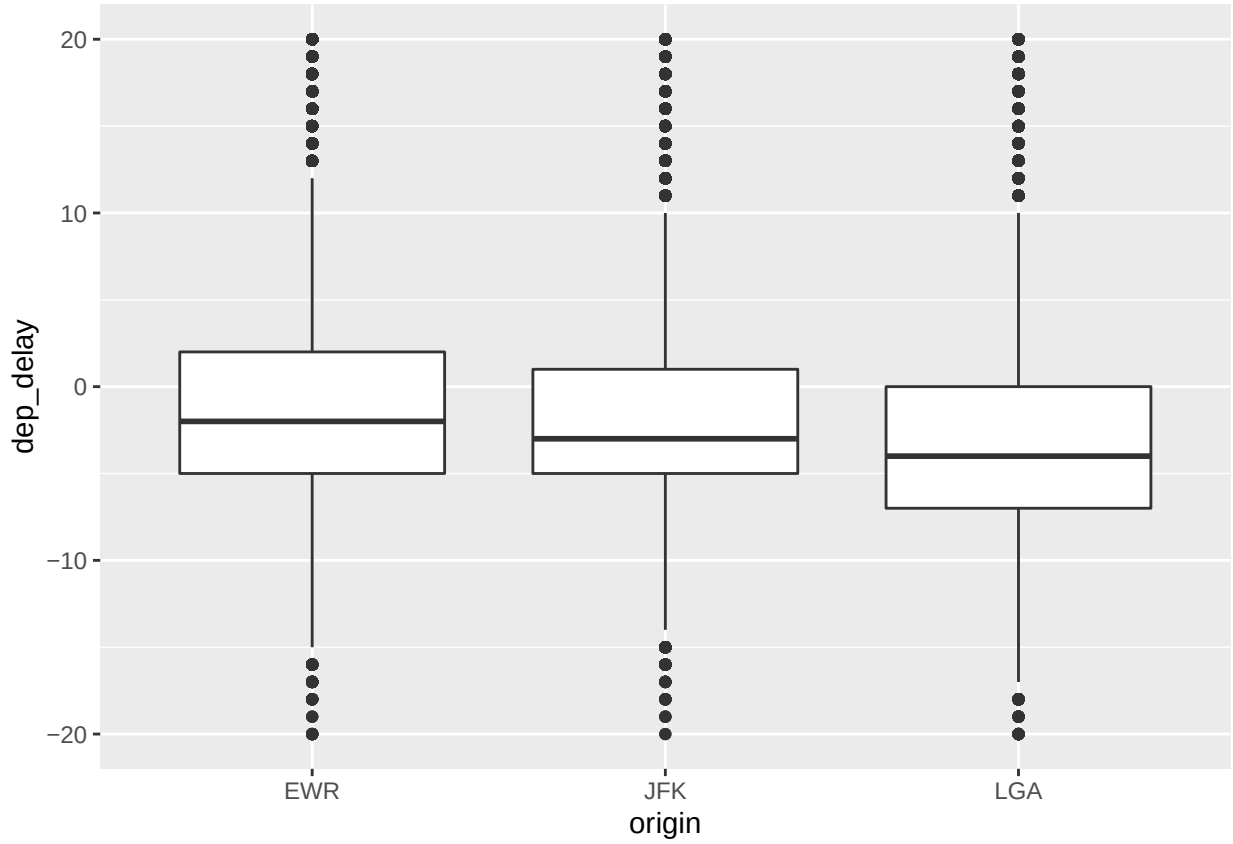
- Describe and demonstrate how you determine if there are unusual values in the data. E.g. too large, too small, negative, etc. Boxplots are very useful to determine these kinds of values; you can see the minimum and maximum of the delays for each airport and which values do not lie within the interquartile by inspecting the boxes. When a data point is unreasonably large or small compared to box's range, you can conclude it is an outlier.
- Describe and demonstrate how you determine if they are outliers. It depends on the data, however since standard deviation and mean are very sensitive to unusual values, it is typically good practice to remove unusual values if they will negatively affect your data and goal during data analysis.
- 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) -Imani

- 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.

```
library(tidyr)
flights2 <- flights %>% drop_na()
ggplot(data = flights, mapping = aes(x = origin, y = dep_delay)) +
  geom_boxplot() +
  ylim(-20,20)
```

```
## Warning: Removed 69929 rows containing non-finite values (stat_boxplot).
```



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) - Julia

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.