# Statistical Learning and Data Analysis 2021 - 52525
# Lab 1 - Flights data

Due April 6nd before 4:30pm

**Hand In Procedure:** Please prepare a file with a writeup and code (the writeup can be in Hebrew or English). Labs can be handed in alone or in pairs (no more than 2 per lab!). Please upload two versions into Moodle; one as an Rmd file, and the other compiled without code. Please make sure the reports are under 6 pages in the pdf.

### Background

We will analyse data for flights out of three terminals of New York City during 2013. This flight data is collected by the oversight agency to supervise the airports. [1] We are particularly interested in the following questions:

1. **Flight Schedule** What can we say about the recurrent flight schedule? What are recurring patterns of variation? Identifying changes or deviance from the recurring behavior.
2. **Flight Delays** What are the patterns of flight delays (both short and long)? Can we find potential causes for further investigation?

To get the data, type in R:
```
>install.packages('nycflights13')
>library('nycflights13')
```

The main data set is `flights`. Additional information includes weather patterns `weather`, destination airport information `airport`, and information about the planes in `planes`. Before you begin, take some time to get to know the data; main variables, how they relate to each other, drastic outliers, etc.

## 1 Graph Critique

I've uploaded two graphics into Moodle from the winning posters. Please discuss in brief the following:

1. What questions / stories the graphic is trying to answer?
2. Do they answer successfully?
3. Do they raise new questions not addressed?
4. Please suggest one way in which these figures can be improved.

[1] Every other year, the American Statistical Association holds the ata Exposition special poster session on visualizing interesting large data sets. In 2009, the data set included on-time information for all US domestic flight information between 1987 and 2008. Details about the competition are in http://stat-computing.org/dataexpo/2009/, and the winning poster (Wicklin and Allison) as well as other posters are found here: http://stat-computing.org/dataexpo/2009/posters/. We will use a smaller but similar dataset collected in 2013.

## 2 Reproducing these analyses

For each of the two graphics from part 1, try to reproduce the analyses using the 2013 NYC flights data as closely as possible. That is produce:

1. A graphic summarizing the flight volume and flights delayed, broken by day and showing weekly cycles.
2. A graphic summarizing the percent of flights delayed, broken by destination Airport.

## 3 Freestyle analysis

Now, explore the data on your own. Produce 1-2 graphical summaries showing interesting things you found. (You can look at the winning posters for inspiration, or try things on your own). For each, prepare a caption that explains what is shown in the graph, and what can be learned about the data. Make sure graphs have labels, are clear. Think about the zoom, about colors, display, labels, etc.

## 4 Graphical "Lineup

Here, we would like to prove (to ourselves) that our indings not due to chance variation. The idea is to use a Graphical Lineup (as in the paper by Wickham et al 2010 in the Moodle). We will check whether delayed-departure has a seasonal pattern. Our null hypothesis is:

- The proportion of delayed flights per month is independent across months.

For example, according to the null hypothesis the proportion of delays in April is not associated with delays in May.

1. Produce a graphic that tries to answer this question for the real data.
2. Produce simulated data-sets based on the null hypothesis, and produce a graphic for each of them.
3. Is it easy to tell apart the real data from the simulated ones? How is it different? What have we learned?