Statistical Learning and Data Analysis 2020 - 52525
Lab 1 - Flights data

Yuval Benjamini

Due April 21nd before 4:00pm

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: a compiled version, and the RMD version with the code. The written answers don't have to be long, but they need to be outside of the code components. Please make sure the reports are under 6 pages in the pdf.

Background

We will analyze 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`. You can use the `left_join` or `right_join` commands to add information from those tables. 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 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?

[1] Every other year, the American Statistical Association holds the "Data Exposition" a 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.

4. Please suggest one way in which these figures can be improved.

2 Reproducing these analyses

For each of the two graphics from part 1, reproduce the analyses using the 2013 NYC flights data. 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.

Please explain briefly what steps you needed to do to prepare these dataset, including transformation of variables, removal of outliers, subsetting, etc.

3 Freestyle analysis

Now, explore the data on your own. Produce 1-2 graphical summaries showing interesting things you found. For each, prepare a caption that explains what is shown in the graph, and what can be learned about the data. Think about the subset of data you chose, any outliers or exclusions, the best chart-type, the colors, labels, etc. A good answer should have more insight than a plot of one variable against the other.

4 Graphical Lineup - Misdar Zihui

Here, we would like to prove (to ourselves) that our "finding" is 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 that the average flight delay may change per month, but varies randomly across the year (i.e. the number 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?