Final Project Description

Getting the Data

You need to obtain your own data from the BTS:

https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236

This page allows you to select a lot of columns and customize your data in a number of ways. In these instructions I will not tell you exactly what settings you will need to use. Just read through the questions below and see what kind of data you will need. For questions 1 and 2, you need not worry about the cause of any delay. Just use the total delay in minutes.

Beware! You might be tempted to just get all the data. This is something that even some experienced researchers will sometimes do, just to find out later that it is a very bad idea. Excessively large files not only take forever to download, they also cause all kinds of trouble later on. So be smart and figure out what you really need. It is expected that you will play around with this a bit until you find the settings that best meet your needs. You may even want to get different data sets for the different questions.

Question 1: The worst flight

In the month of September 2017, what domestic flight out of Boston, MA (BOS) had the worst delays and cancellations? (In other words, which flight had the lowest probability of getting you there on time). Print out the flight number and the destination. Keep in mind that not every flight operates every day, so use care when you calculate your averages!

Question 2: The best carrier

If you call New York's JFK airport home, which carrier was most likely to get you to your destination on time during the month of September 2017? Consider both outbound and inbound flights (aka flights to and from JFK).

Question 3: Weather Induced Delays

You may know that often airlines claim that delays are caused by weather. This is sometimes true while other times it provides an easy excuse for carriers to avoid compensating you for delays that actually result from poor fleet or crew scheduling (if you are interested, look at the research on schedule robustness). For this question you will not have to investigate truth or fiction behind "weather" delays, but you will have to visualize the relationship between local weather delays and weather events.

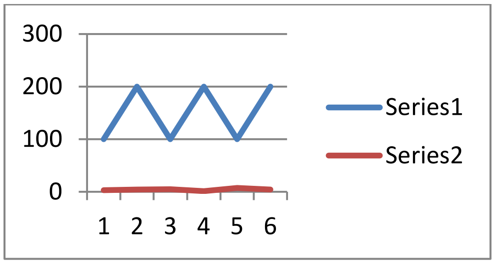
To complete this part, you will need flight data covering flights departing Providence (PVD) during the month of January 2017. You will need to obtain the "WeatherDelay" column. For the graph below, you need to sum up all weather delays of all flights throughout an entire day. You should end up with a dictionary that has the total delay for each day.

You will also again use weather data. Please get a fresh copy of pvdWxJan.csv from the resources section in Sakai.

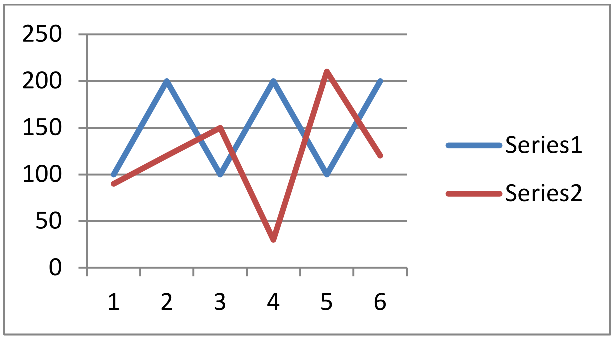
This is a very comprehensive report. We are only interested in the peak wind speed throughout the day. You will notice that there are many observations for each day and that only one of them lists the average and peak wind speeds for each day. This makes sense since we don't know either of these two values until the end of the day.

Your job is to create a line graph with the days of the month (1-31) on the abscissa (fancy Math speak for the x-axis). We want to see two lines – one representing total delay in minutes and one showing the peak wind speeds during the day.

Note that you need to make the graph look appealing, so you must scale one of the graphs. The total delay in minutes will routinely be in the hundreds or higher while peak wind speeds will be much lower. Don't create a graph like this one:



But one like this (In this example I just multiplied each value for series 2 with factor 30) :



Risk Free Extra Credit Assignment:

This extra credit assignment is intended to be quite challenging. It is, however, risk-free. This means that you cannot lose any points for attempting it. If you choose to try it, you can only gain points – up to 3 additional points in fact.

In question 3 we have plotted wind speeds against delays. Now wouldn't it be interesting to know what weather phenomena are most likely to cause flight delays? So let's take a look and try to find correlations (*) between weather delays and the different columns in our weather data.

Write a program that will look at the different data and print out a list of the three weather measurements that are most correlated with flight delays.

(*) We are only looking at correlations. A strong correlation does not, by itself, imply a causal connection, so technically we are heavily cheating here. But let's just see what we find anyway.