Group Project

The group project is based on the **airlines** data set from the US Department of Transportation's Bureau of Transportation Statistics (BTS). This is the data we looked at in Chapter 5 of **Data Science in R**. To get the data for certain years, we will use the same script file for downloading particular years.

The **airlinesauxiliaryfiles.zip** file in the compass course space includes a **ReadMe** file describing the full data set, details on how to download particular years of data, and descriptions of the auxiliary files included in the zip file. The auxiliary files include airport codes, carrier information and plane information which might be useful to combine with the raw data.

Each group will be assigned two years of data to analyze. The assigned years are as follows:

Group	Years
1	1999 and 2003
2	1998 and 2006
3	1996 and 2005
4	1997 and 2002
5	1997 and 2004
6	2000 and 2007
7	1999 and 2004
8	2000 and 2006
9	1998 and 2002

The final project will count for **30% of your final grade**.

Each group will be working with a little over 1 GB of data. Processing and analysis of the data must make use of multiple software applications (including Hadoop-based applications) we have used in the virtual machine.

The general goal is to extract interesting information, trends, and comparisons about flights in the years your group is assigned. You may need to deal with missing values in variables and perhaps variables that are entirely missing.

Some possibilities you might consider for your two years:

- Cancellation trends
- Trends in delayed arrival or departure
- Trends in number of flights
- Other trends you think might be interesting

You should consider these types of analysis and any others you find interesting to give a broad overview and comparison of travel in your two years. Analysis should focus on comparing between the two years

and with respect to other classifications. For instance, you might consider trend or overall analysis by region, airport, carrier, plane manufacturer or anything else you think might be interesting. Models as a function of categorical factors (region, carrier, etc.) might also be useful for exploration and analysis.

Interesting features in time should be explored for historical significance. For instance, if there is a noticeable difference in cancellation rates or delays on a particular day or week, was there a major storm or other historical event that might be related?

Interesting visualizations must also be provided. This should include visualization beyond the usual scatter plots and bar charts and might include heat maps, tree maps, network visualizations (e.g. to see connectivity and/or identify carrier hubs), or other advanced visualizations. Some of these types of visualizations will be discussed in class, but you are not limited to visualizations discussed in class.

Final reports (and supporting code files) will be due at **11:59pm on Wednesday May 1** and must be submitted via the course website. The project is to be done as a group and submitted as a group. The report must also contain a section explaining each individual's contributions to the project.

Note: If there are any concerns about conflicts within a group or individuals' contributions throughout the project, the group should first attempt to address and resolve the issue among its members. If the group is unable to resolve issues on its own, or if individuals are not responsive to email or other attempts to communicate, this should be discussed with the instructor for assistance in resolving the issue. If instructor assistance is required, the instructor should be asked for assistance early enough in the semester to allow adequate time for meeting with the group, resolving the issue, and completion of the project by the group as a group after resolution.