**Introduction**

**-Question**

Flight delays, especially those over 30 minutes, usually cause some losses. Until recently, people still cannot accurately predict the flight delay, and they usually estimate it by simple sense such as weather and different airlines. We believe that there are more factors have the influence on the length of flight delays, and we want to find them by processing this research. Although we are not building a prediction model, this research will help us to have a deeper understanding on the impactful variables on flight delays, which is the first step to a more accurate prediction and a more efficient future travel.

**-Hypothesis**

To start with simplicity, we only pick two airports, DFW from Dallas and MSP from twin cities. We want to develop the models on both two airports and compare their factors. We choose these two airports because although they are in different locations with various weather patterns, economic levels, and even living styles, they have similar size and the same time zone that we will not be distracted by transforming the time variables. What we estimated is that the factors of the two models and their coefficients will be different. For example, snowing days are so common in Minnesota and MSP is so experienced to deal with this kind of bad weather, so the effect of the snow event on MSP’s flight delays should be much smaller than on DFW’s since Dallas has seldom snows around the year.

**-Variables**

Some of the data are purely picked from our data, but the other ones have been processed by us to meet our specific analysis.

Created variables:

* has\_events: It is a categorical variable that indicates if there is any weather event around the airport district.
* airlines\_class: A categorical factor to divide the airlines into two classes, Major for top 20 airlines and Minor for the others.
* is\_weekend: all the Friday, Saturday, and Sunday are set as weekend and the others are weekday.
* time\_of\_day: the categorical variable to specify the morning, afternoon, evening and night of the day.
* season: our season definition does not divide the months evenly. Summer has four month(6-9), while Fall has only two(10,11).

Original Variables:

* Weather factors:
  + Temperature: We use minimum temperature of the day because we think lower temperature will have more influence on the delays. It has a mean of 39.7F and SD of 23.05F
  + Cloud cover: Degree from 0-8, has a mean of 5.102 and SD of 2.
  + Visibility: We use minimum visibility of the day because we think lower visibility will have more influence on the delays. It has a mean of 5.75 miles and SD of 3.69 miles.
  + Wind speed: We use mean wind speed of the day because we want to know the whole wind status of that day. It has a mean of 9.91 mph and SD of 3.59 mph.
  + Humidity: Range from 12 to 92, has a mean of 51.35 and SD of 15.48.
  + Wind direction degree: Range from 2 to 359, has a mean of 195.5 and SD of 96.5.
* Other factors:
  + Distance: We use log() to process it because it has a much larger range than delay time.