## Data Description

**Source of data:** <http://www.transtats.bts.gov/>

The data is obtained from the United States Department of Transportation (Bureau of Transportation Statistics) for a period of 12 months from January 2015 to December 2015. The corresponding weather data for each departure date is obtained from <https://www.wunderground.com/history/>. The weather data for the previous day is obtained by shifting the dates in the initial data by 1. The flight and final weather data (current and previous) are then combined for each date and city, which is used as the final database. The data is then filtered for origin city ‘MSP’ (and ‘DFW’) and rows with missing values are dropped.

A brief description of the relevant fields is given below:

* The has\_events flag indicates if a particular day has any weather- related events. It is set to 1 if there is an event or 0 otherwise. The is\_weekend flag is set to 1 if the departure day is Friday, Saturday or Sunday; or 0 otherwise.
* Airline codes are looked up for American, Delta, Southwest, United, Air Canada, JetBlue, Alaska, WestJet, AeroMexico, Spirit, Hawaiian, Porter, Sun Country and Frontier Airlines, Volaris, Allegiant Air, Virgin America, Interjet, Air Transat and Vivaaerobus (20 top airlines in North America1) and the corresponding airline\_class flags are set to 1.
* Time of flight departure is categorized as Night (Flag:1, 11pm-3am), Morning (Flag:2, 4am-10am), Afternoon (Flag:3, 11am-3pm) and Evening (Flag:4, 4pm-10pm).
* The available weather factors are temperatures and dew points in ℉, percentage humidity, Sea level pressure and precipitation in inches (precipitation for days which are too small to measure, indicated by ‘T’ in the data, are set to 0), visibility in miles, cloud cover and wind and gust speeds in mph.

**Analyses**

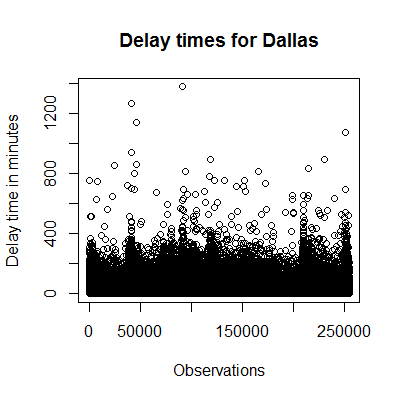
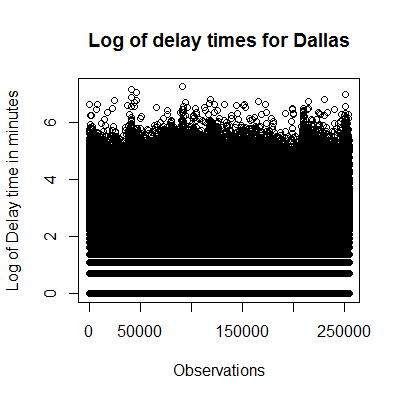
The analysis was started out by examining the delay times of various flights departing from Dallas/Fortworth (DFW)and from Minneapolis(MSP). The delay times for flights which depart early are set to 0, as the scope of the project is to analyze the flights which depart later than the scheduled time. We use the natural log of the delay times as the response variable to create a better fit and to linearize the model.

Separate models are run for flights departing from DFW and MSP, as this gives an accurate idea about the various factors affecting delays in each of the locations. Given the number of predictor variables, the best fitting model was arrived at by backward selection rather than by forward substitution. We start with all possible, logically intuitive predictors in our model which includes logistic, weather variables for the current as well as for the previous day. We remove the predictors that are insignificant and check the effect of removing these predictors from the model by analyzing R squared, R squared adjusted and the values of standard error. The best fit model was arrived at when the removal of extra predictors did not have a positive impact on R squared or failed to reduce the standard error.

**Delays for flights departing from Dallas(DFW)**

It is observed that for Dallas/ Fort Worth, the mean and the median delay times are 13.52 minutes and 0 respectively. 151,722 out of 254,305 (59.6%) observations have a delay time of 0, with the maximum delay time being 1377 minutes. All the observations were used for the analysis in order to analyze accurately the factors which affect the delays.

The graph below shows the distribution for flight delay times for flights departing from Dallas/ Fort Worth.

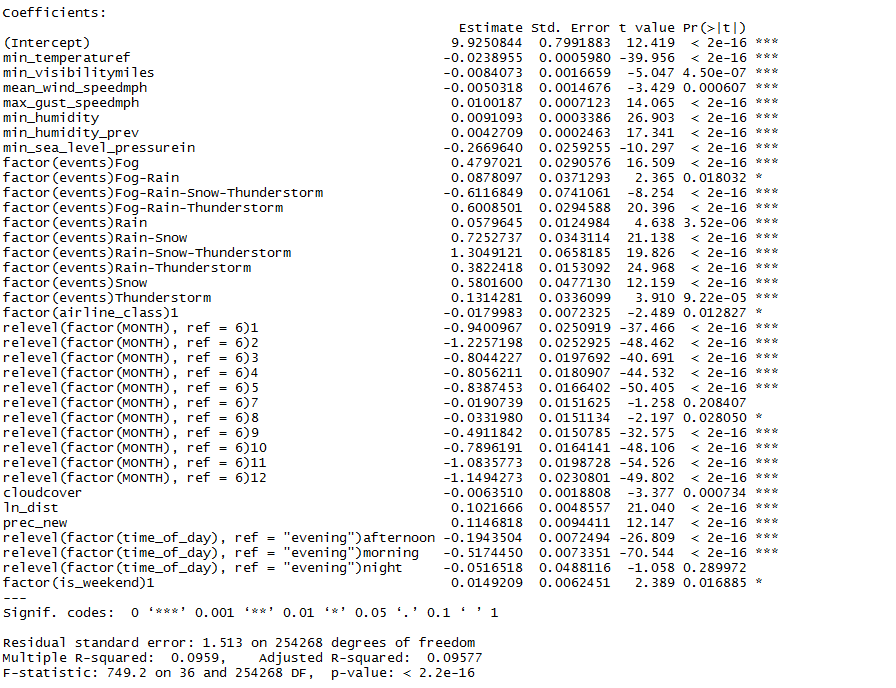
 

**Best Fit model for DFW**

The best fit model to explain delays for DFW includes the following predictors:

**ln(Delay time (in minutes))= β0 + β1 (Minimum Temperature) + β2 (Minimum Visibility) + β3 (Mean Wind speed) + β4 (Max Gust speed) + β5 (Minimum Humidity) + β6 (Minimum Humidity of previous day) + β7 (Minimum sea level pressure) + β8 (Weather related events) + β9 (Class of airline)+ β10 (Month of the year) + β11 (Cloud cover) + β12(ln(distance)) + β13 (Precipitation)+ β14 (Time of the day)+ β15(Weekend or Weekday)**

The beta coefficients of the various predictors are as follows:



It is seen that the weather factor that affects delays the most is the minimum sea level pressure. *Each individual factor is interpreted by keeping all other factors constant*. Keeping all other factors constant, an increase of one inch in the sea level pressure decreases the average delay time by 26.7%. This is followed by precipitation, an increase in one inch of which increases delay by 11.5%. It is noted that humidity of the current day, and that of the previous day as well when increased by one unit, affects delay by 0.9% and 0.4% respectively. Other factors such as minimum visibility miles, the mean wind speed and maximum gust speeds of the day and cloud cover contribute to delay as well, albeit only in a small way.

As the distance between the source and the destination increases by 10%, the average delay time in minutes increases by 1%. Compared to a day with no significant weather related events, it is seen that on a day which has rain, snow and thunderstorm, the average delay time increases by a large amount (130%). The next significant events are found to be rain and snow in conjunction, followed by snow, which increase the delay by 73% and 58% respectively compared to a day with no events.

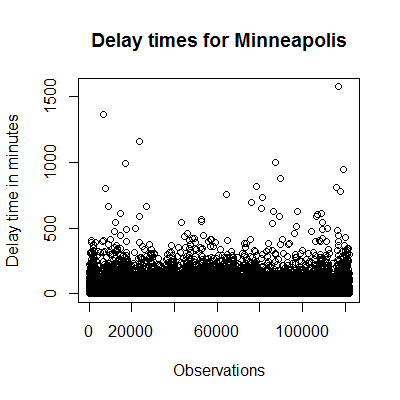
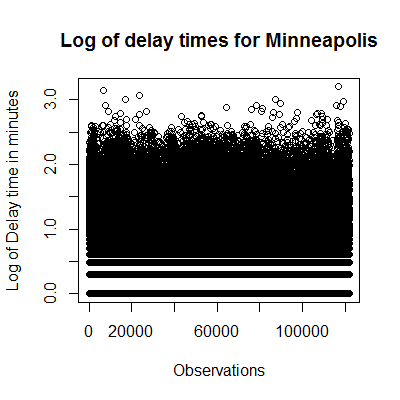
Major airlines seem to be better at handling delays compared to minor airlines (average delay time decreases by 1.8% for minor airlines compared to major ones). It is also observed that delay increases by 1.5% from a weekday to a weekend.

Month of the flight departure is a major factor affecting delays. June is the month where the average delay time is maximum, and July is not significantly different from June as indicated by the high p value of 0.208. November and December are the months where delays are minimum (delays decrease by 108% and 115% respectively compared to June).

Finally, factoring in the time of the day, it is found that evenings have the most delays (with night not being significantly different than evenings). Going from evenings to mornings and afternoons, delay decreases by 52% and 20% respectively. The model explains 9.6% of the average delay times.

**Delays for flights departing from Minneapolis(MSP)**

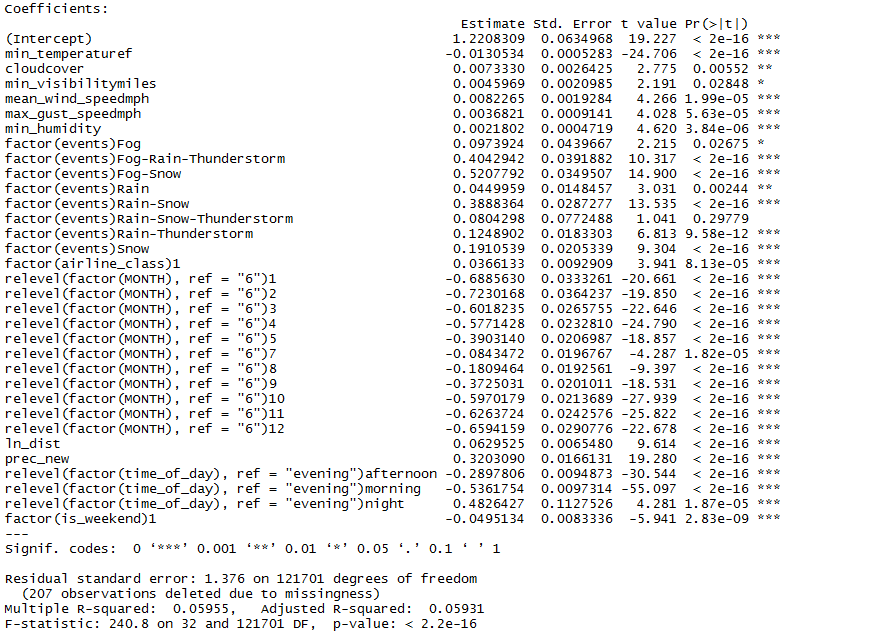
For MSP, the mean delay time is found to be 9.373 minutes, with the maximum being 1576 minutes. The number of observations having a delay equal to or less than 0 are significant (69%) out of 121,941 total observations. Taking into account all the observations resulted in a model with the best fit.

**Best Fit Model for MSP**

The best fit model to explain delays for MSP captures 6% of the delays:

**ln(Delay time (in minutes))= β0 + β1 (Minimum Temperature) + β2 (Cloud cover) + β3 (Minimum visibility miles) + β4 (Mean wind speed) + β5 (Maximum gust speed) + β6 (Minimum humidity) + β7 (Weather related events) + β8 (Class of airline)+ β9 (Month of the year) + β10 (ln(distance)) + β11(Precipitation) + β12 (Time of the day) + β13 (Weekend or Weekday)**



The weather factor which has the most effect on delays is precipitation. An increase of one inch in precipitation causes an increase of 32% in average delay times, keeping all else constant. Increase in temperature by 1℉ causes a decrease in delay by 1.3%. *The beta values of all individual predictors are interpreted by keeping the other predictors constant*.

Increase in minimum visibility miles, mean wind speed and maximum gust speed can also be marginally attributed to the increase in delays. Increase in minimum humidity level by 1 percentage increases delay minutes by 0.2%. Cloud cover also contributes in a small way to delay times. As cloud cover increases by one unit, delay increases by 0.7%.

An increase in the flight distance by 10% increases delay by approximately 0.7%, while keeping all other logistic and weather factors constant.

Compared to a day with no significant weather related events, a day when fog and snow are present together tend to increase delay times by 52%. On the other hand, when fog, rain and thunderstorm occur simultaneously, delay increases by 40.5% compared to the base level (no significant events). Snow comparatively increases delay times by 19%.

Similar to Dallas, evening is the time of the day which is found to have the greatest average delay times in Minneapolis. Relative to evening, delay decreases by 54%, 29% and 48% for mornings, afternoons and nights respectively.

On the contrary, it is observed that minor airlines are better than major airlines at handling delays by 3.6%. Also, surprisingly, weekends are found to have less delay times than weekdays by 5%.

The effect of seasonality in MSP is found to be similar to DFW. June, followed by July, are the months with greatest average delay times. Moving from June to December and January, the average delay times decrease by 69% and 66% respectively.

References:

1. <https://en.wikipedia.org/wiki/List_of_largest_airlines_in_North_America>