The datasets which have been explored are flights dataset for years 1990, 2008 and dataset which was created solely for the cancelled flights out of 22 datasets for the period of years 1987 to 2008.

Initial datasets we downloaded from http://stat-computing.org/dataexpo/2009/the-data.html.

Datasets contains information about date of flights which is represented in few separate columns (day of the week, day of the month, month), scheduled and actual departure and arrival time, delay time and cause of the delay, destination and origin of the flight, carriers information as well as cancelled flights.

As my main interest of this exploration is the cancelled flights, datasets for 1990 and 2008 were modified by selecting only cancelled flights and dropping columns which wouldn't provide relevant information, such as actual departure time or delay information. To be able to print some histograms, entries with NaN values got to deleted, however it was very low amount which won't affect the whole picture. Additional information about total amount of flights per year (within 1987-2008 period) as well as amount of flight per carrier per and and per origin per year (see get_data_function.ipnb file) was collected. In some cases proportions were calculated as absolute amount wouldn't be representative.

From univariate exploration and comparison of two years we can see some relations. Flights more often being cancelled because if weather and there more during winter time. Year 1990 dataset doesn't have information about cancellation reason and thus comparison between two years couldn't be made.

Day of the month or time of the day don't show any specific tendency as well as distance of the flight. Weekends have lower percentage of cancellation for about 20%

Amount of cancellations for specific carrier was checked as well. Comparison of carries which present in datasets of both 1990 and 2008 showed that for 7 out of 8 companies percentage of cancellation has increased.

Size of the airport and the period of time it's been operating might be a valid reason for higher amount of cancelled flights. The exploration showed that the airports with lower total amount of flight have higher percentage of cancellation, however comparison with later year showed that with the increasing amount of flights amount of cancellation has actually decreased. That is the hypothesis which can be investigated more.

Investigation of destinations showed similar to origin correlation which actually makes sense.

For bivariate exploration it was checked how the amount of cancelled flight per month aligns with the reason of cancellation which confirmed the hypothesis of strong weather related reasons for cancellations.

Tendencies within time period ('cancelled.csv' dataset was used) were explored however only those graphs which showed any relations in the exploration for one year datasets.

Cancellations due carrier appeared to be more common reason than the weather for the period of 22 years. However amount of cancellations due to wether during thewinter season is still leading.

The main tendency is that amount of cancellation is increasing. General amount of canceled flights is low and the highest amount of cancellation happened in 2001 which is most likely connected to events of 9/11/2001. also, total amount of flight after decreased but already by 2003 increased again with percentage of cancellations going down.

For the explanation slide deck it was chosen to look at main cancellations reasons, i.e. weather, carrier and NAS. This information is elaborated and supported with graphs showing how weather, carrier and origin of flight affect cancellations.

To bring up important points in the presentation slides and for building up a story, the order of the graphs was changed and accentuation and extra details on them were added (in comparison to how they appear in the exploration file). Colors were also changed in purpose to avoid red and green colors appear in the same graph.