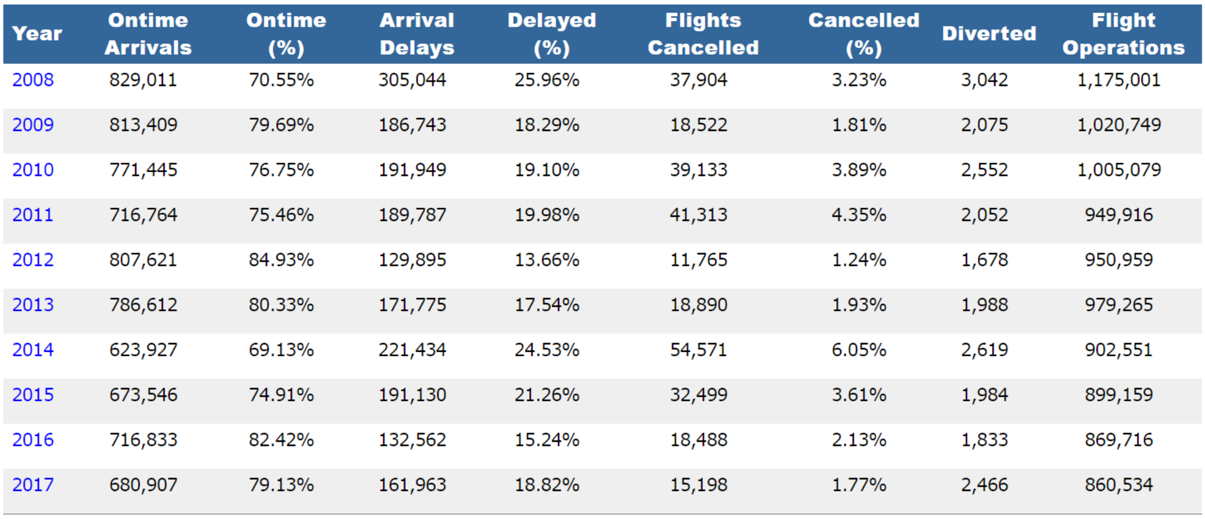
Capstone project proposal

The problem:

Air transport delays are not only a painful inconvenience for travelers, they also induce large costs, for the airlines, the passengers and the economy of the countries. In the USA, every year around 20% of airline flights are delayed or cancelled (see figure below), resulting in significant costs to airlines and passengers. A study commissioned by the U.S. Federal Aviation Administration in 2010 estimated that flight delays cost the airline industry $8 billion a year, much of it due to increased spending on crews, fuel and maintenance. Delays cost passengers even more, nearly $17 billion. In addition to these direct costs for the airline industry and its customers, flight delays have indirect effects on the economy of the countries. The same study estimated that air transportation delays reduced the 2007 US gross domestic product (GDP) by $4 billion.



Year	Ontime Arrivals	Ontime (%)	Arrival Delays	Delayed (%)	Flights Cancelled	Cancelled (%)	Diverted	Flight Operations
2008	829,011	70.55%	305,044	25.96%	37,904	3.23%	3,042	1,175,001
2009	813,409	79.69%	186,743	18.29%	18,522	1.81%	2,075	1,020,749
2010	771,445	76.75%	191,949	19.10%	39,133	3.89%	2,552	1,005,079
2011	716,764	75.46%	189,787	19.98%	41,313	4.35%	2,052	949,916
2012	807,621	84.93%	129,895	13.66%	11,765	1.24%	1,678	950,959
2013	786,612	80.33%	171,775	17.54%	18,890	1.93%	1,988	979,265
2014	623,927	69.13%	221,434	24.53%	54,571	6.05%	2,619	902,551
2015	673,546	74.91%	191,130	21.26%	32,499	3.61%	1,984	899,159
2016	716,833	82.42%	132,562	15.24%	18,488	2.13%	1,833	869,716
2017	680,907	79.13%	161,963	18.82%	15,198	1.77%	2,466	860,534

Report from the United States Department of Transportation.

These high costs have motivated the analysis and prediction of air traffic delays, and the development of better delay management mechanisms (Rebollo and Balakrishnan 2014). In 2014 General Electric (GE) even sponsored a challenge with an award of $250,000 to the team who can most accurately predict flight delays (https://www.kaggle.com/c/flight2-final).

The goals of this capstone project are to use exploratory data analysis to provide insights about main sources of flight delays, and to develop machine learning models to predict airline's departure and arrival delays.

Potential clients:

The results from this study could provide airlines, airports and aviation administration departments with insights to improve their operations and policies. Examples of this possible uses are:

- airlines could identify opportunities that could reduce fuel consumption,
- airports could better manage their air traffic by anticipating periods with higher probability of flight delays,
- travelers may be able choose the most suitable time / destination /carrier for their next trip.

Data:
Data are from U.S. Department of Transportation:
https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time.

The dataset used for this project includes details about flights in the US from January to December 2017. It includes the following variables:

- Dates: day of week, date, month, year
- Arrival and departure times: actual and scheduled
- Flight times: actual and scheduled
- Origin and destination: airport code, latitude, longitude
- Carrier: American, Aloha Air, United, US Air, …

Approach:
The analysis and prediction of flight delays have been the topic of several previous efforts. Examples of previous works are Bratu and Barnhart 2005, AhmadBeygi et al. 2008, Jetzki 2009, Sridhar and Chen 2009, Hastie et al. 2009, Klein et al. 2010, Pyrgiotis et al. 2013, Rebollo and Balakrishnan 2014, Jin Kim 2016.
In a first step the proposed approach will use feature selection and reduction techniques to identify the main aspects related to flight delays. The aim is to provide insights about the following points:

- Main sources of flight delays.
- When is the best time of day/day of week/time of year to fly to minimize delays?
- Do older planes suffer more delays?
- How well does weather affect plane delays?
- Is it possible to detect cascading failures as delays in one airport create delays in others?

In a second step, machine learning models to predict airline's departure and arrival delays will be proposed. Two types of prediction mechanisms will be considered:

- Classification, where the output is a binary prediction of whether the delay is more or less than a predefined threshold. It is important to note that Gradient Boosting and Random Forests methods have been used with success in previous works ()).

- Regression to estimate delay duration.

Finally, storytelling techniques will be used to present the main insights and the final results of the study.

Deliverables:

- Jupyter notebook with code,
- slide deck,
- report.