Molly Gallagher, PhD: Interviewing for the Infectious Disease Modeler Position, June 2020
Data Scientist Exercise 02: I have written a non-comprehensive overview of the National Transportation Safety Board (NTSB) data, as well as a short paragraph and accompanying figure that could be provided to a client with a fear of flying.

Overview of the data, and some key limitations: The NTSB website notes that the database contains information dating from 1962 and later. However, prior to 1982 there only 6 entries, one of which is from 1948. There are only 3 entries without event dates, two of which are still in the preliminary phase. There are a total of 77,257 entries with 31 different fields. Each entry has an event ID number and an accident number. The accident numbers are unique, but some event ID numbers occur more than once because some events involved multiple aircraft: for example, rows 6 and 7 both refer to the same event, which involved a Dassault Falcon 2000 and a Beech C90. Entries include both accidents and 'selected' incidents, but the selection criteria are not specified on the NTSB website. The vast majority of the entries are about accidents; less than 4% of the entries refer to incidents. About 4.6% of the entries have a report status listed as 'foreign'. The majority of events with 'foreign' report status did occur outside the United States, although 28 of them list the US in the 'country' field. Based on the 'country' field, 94.6% of the entries occurred in the United States. About 0.66% of the entries do not have a country listed, but the majority of these do list a 'location' over water, such as the Atlantic Ocean or Caribbean Sea. Across all entries there are almost 25,000 unique locations, including blank entries and entries such as 'Unknown, UN'. Many fields are entered manually when logging an entry on the website, which allows for greater flexibility in reporting but also increases the amount of error and variation. For example, there are thousands of entries in the 'make' field for both 'CESSNA' and 'Cessna', as well as one for 'Cesna'.

It is often important to understand how variables change over time, but when I examined the data to see how the frequency of different aircraft types changes over the years, I ran into a problem. The majority of the entries in the database do not contain information about the aircraft type: see the line labeled "no data" in Fig. 1. And this reporting behavior changes abruptly over time: every entry from about 1986 to 2013 is missing the aircraft type, but that information is sometimes recorded in years prior to 1986, and also in the years since 2013. It is difficult to say much about the differences in accident rates by aircraft type over time, then, because we can't expect the relative proportion of airplanes versus helicopters or balloons to have remained constant across the decades.

There are additional limitations to what we can infer from these data. For example, we might be inclined to report that accidents involving airplanes are more deadly at the individual level than accidents involving other types of aircraft. But airplanes typically carry a greater number of passengers than most aircraft, such as gliders or helicopters, and of course they are far more common as well. So based on these accident and incident reports alone, we can't verify that certain factors are relevant in causing accidents. To better understand risk and causal relationships, we should also use comparative data and consider flight plans where accidents or incidents did not take place.

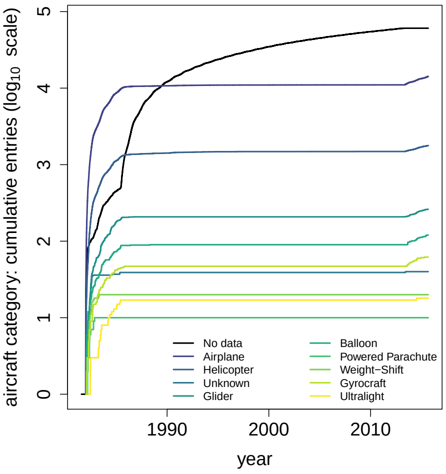

Figure 1: The way information is reported changes non-randomly over time.

Notes for the nervous fliers: If you must be in an accident in an aircraft, make sure that you are in a professionally built machine. It is true that across all the available data, there are more fatal accidents in professionally built aircraft than in amateur ones: almost 21% of fatal accidents occur in professional builds, while less than 3% of fatal accidents occur in amateur aircraft. But these statistics are misleading! It is more common to have a fatal accident in a professionally built aircraft simply because professional aircraft are much more common than amateur ones. Once we control for those differences, we see that about 23% of accidents in professional aircraft result in fatalities, but that number jumps to 29% for amateur aircraft[1]. It becomes clear that while flying, you will be safest in the hands of a professional.

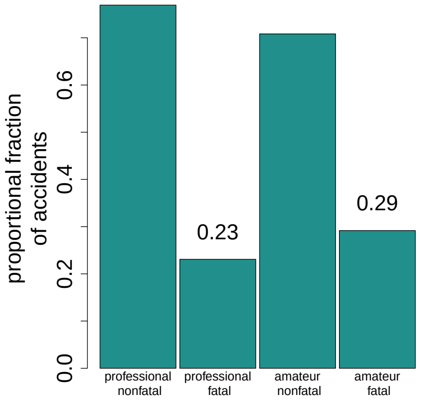

Figure 2: For maximum safety, we recommend traveling in a professionally built aircraft.

[1] This difference was statistically validated by chi-square test in R: $p << 0.001$. Note that we do not account here for flights in which no accident occurred, nor for the number of fatalities that occur in different accidents, merely the severity of injury that occurs in given accident: fatal versus non-fatal.