Methods

We are focusing on airplane crashes and fatalities since 1908, looking at many factors to see what caused the crash and when it happened. Our guiding question: What factors (operator, type, and summary) are associated with more deadly plane crashes (larger proportion of deaths)? Observations in our study are each row in our dataset, which represents a single plane crash incident with features describing it (date, time, location, # of fatalities, etc). As far as an observation our team made when understanding that data, there was a spike in plane crashes seen between 1940 and 1980, which we understand as having to do with a significant amount of war occurring around the world during this time. Because of this, we decided to hone the investigation of our study to crashes occurring after 1990, as that year signifies a shift in commercial flights and a movement towards what humans understand as the modern airplane today. Focusing on only commercial flights and those that ran their course within the last 35 years will have a more meaningful and topical impact on those reading our study.

We will be using supervised learning for our model because we have the label data, specifically the number of fatalities and people aboard, and supervised learning can use this data to predict relationships with these known outputs. Regression will be used because we are predicting a continuous proportion, comparing the number of people who died to the total number of people on the plane.

To prepare our data for modeling, we aim to use one-hot encoding and PCA in our regression model to incorporate multiple variables in our predictive model. We will use one-hot encoding to categorical variables like “Operator”, “Type”, and “Cause” to ensure that the model can interpret these values numerically. Since one-hot encoding may significantly increase the number of features, we will apply PCA to reduce dimensionality. These two methods will allow us to keep the most important parts while simplifying everything for the model.

Since our model seeks to predict how fatal a plane crash will be, our success will be defined by how accurately the model can do this. We will use an 80-20 train-test split on our data so that we can test the accuracy of the model with the test data. To measure how accurate the model is, we will look at the mean squared error (MSE) and R-squared. MSE will indicate how far our data is from the model's predicted values. We are aiming for an MSE of under 0.05. R-squared will tell us how well the model explains the variability of the true data. A good R-squared that we will aim for is 0.6. This would indicate a strong relationship that can give us some useful predictions. Combined, these metrics will give us an idea of how strong of a predictor the regression is.

Two potential weaknesses we anticipate in our analysis are multicollinearity and overfitting. Multicollinearity can occur when predictor variables are highly correlated, making it difficult to determine the individual effect of each variable. For instance, Operator (airline) and Flight # are likely correlated, since each airline has a unique numbering system for their flights. Similarly, Date and Time could be correlated if crashes tend to happen more frequently during certain seasons or times of the day. To address multicollinearity, we will calculate a correlation matrix and remove or combine highly correlated features, such as using only Operator or Flight #, but not both. Overfitting, on the other hand, happens when the model captures noise or overly specific patterns in the data, such as memorizing specific Route or Location combinations. If the approach fails, we may learn that plane crashes are an unpredictable experience or can not be 'placed into boxes' based on similarities. We may also learn that since so many factors can be taken into account when looking at plane crashes, the data can get very messy very fast and become more difficult to spare, analyze, and make a model out of.