Analysis of Plane Crash Fatality Data

Abstract

This project investigates patterns in airplane crashes and corresponding fatalities using a historical dataset of aviation incidents from 1908 to the present. Although air travel is safer than car travel, statistically, public anxiety persists often thanks to media coverage of rare, but high profile, plane crashes. Our team analyzed the crash data post 1990, focusing on how factors like airline operator, aircraft type, and cause of crash related to the proportion of fatalities on each plane. For data cleaning, we applied one-hot encoding and reduced dimensionality through PCA to prepare the data and build supervised regression models to predict the crash severity, measured by fatality rate. Despite our preprocessing, our models yielded low predictive power with an R^2 of around 0.03, leading us to conclude that most fatality variation is driven by unpredictable or unrecorded factors. We performed some analysis to examine which factors contributed the most to our predictive model. We found that hijacking/terrorism, fuel-related issues, and engine issues showed the strongest association with fatal crashes. While hijacking and terrorism are difficult to predict and control at the operational level, fuel and engine issues are more tangible technical factors that could potentially be identified early through improved monitoring. However, overall, our analysis was limited by the nature of our dataset. The structure of our data was not ideal for predictive modeling, as many key variables were recorded as free-text entries with high variability. This made it more challenging to extract consistent patterns from the data. In future work, access to a more standardized dataset with more numerical data would greatly improve the potential to build meaningful regression models that would lead us to actionable insights on the causes of high-fatality plane crashes.

Introduction

There is a well-known statement that a person is more likely to be involved in a car crash than a plane crash. Statistically - this is true, air travel is a safer mode of travel than that of a car. And even though the statistical evidence to back this up is lengthy, this information rarely eases the qualms of an anxious flyer. Turbulence, though handled well by most planes, still makes passengers apprehensive, as they do not have control over the trajectory of the flight. News headlines about plane crashes and disruptions have become a more common occurrence in the last few months, exaggerating the normalcy of the event taking place. This matter piqued our team's interest.

Flights can be split into two major groups, commercial and non-commercial. Commercial flights transport goods and people for the purpose of generating revenue. Non-commercial flights occur for medical reasons, sight-seeing activities, and other recreational uses. Another category, general aviation, like small private planes or hot air balloons, is 82 times more likely to experience a fatal crash than large airline companies (Li & Baker). However, our dataset did not include many of these instances and had more information on commercial and military flights. Because of this fact, along with the wide range of countries and carriers, we wanted to look for more general trends rather than honing in on a specific airline or area. Looking beyond general aviation fit the purpose of our project better as well because analyzing data on rarer forms of flight is not applicable for the general population.

Beyond the raw numbers, are there patterns in plane crash data that could help us better understand the circumstances surrounding plane crashes? Our team found a publicly available dataset listing plane crashes from 1908, recording information about the aircraft itself and other variables that could be vital for determining factors contributing to plane crashes. When examining the dataset at first, it contained a lot of strings (written values instead of raw statistics). When seeing this, we knew it would be more difficult to inspect compared to a dataset filled with just numbers, but we thought there could be something valuable hidden within the strings.

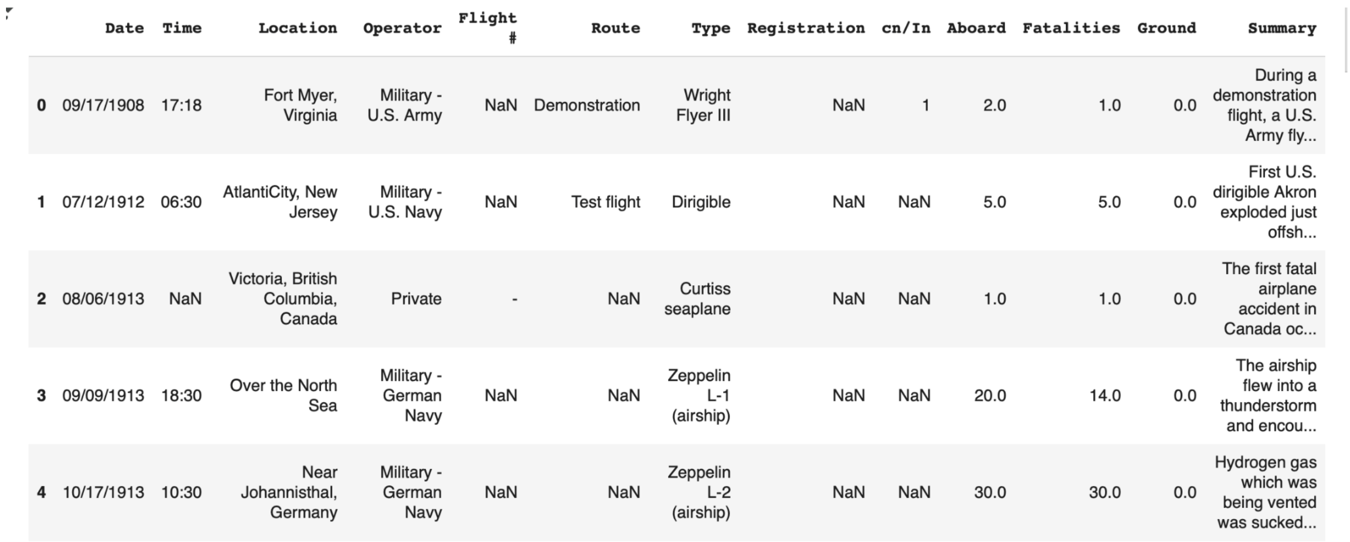
Since this dataset was curated with such an expansive time frame, we decided to narrow down our scope to focus on a more modern era of aviation ie. occurrences after 1990. Our objective was to analyze how factors like the operator of the airline and the summary of the crash are related to the severity of the event, more specifically, the proportion of fatalities on board. Though this analysis will not be able to stop future accidents, our hope is that it could inform future safety measures, further the public's understanding of why these things happen, or promote more research into certain aspects of aviation.

Data

Overview and Brief Data Summary and Expectations

To wrangle the data and perform some exploratory data analysis, we first had to import the necessary tools and libraries, then .head() the data to get a glimpse at the first few rows.

```
import pandas as pd
import matplotlib.pyplot as plt
df = pd.read_csv("Airplane_Crashes_and_Fatalities_Since_1908 2.csv")
# getting an idea of what information out data contains
df.head()
```



	Date	Time	Location	Operator	Flight #	Route	Type	Registration	cn/In	Aboard	Fatalities	Ground	Summary
0	09/17/1908	17:18	Fort Myer, Virginia	Military - U.S. Army	NaN	Demonstration	Wright Flyer III	NaN	1	2.0	1.0	0.0	During a demonstration flight, a U.S. Army fly...
1	07/12/1912	06:30	AtlantiCity, New Jersey	Military - U.S. Navy	NaN	Test flight	Dirigible	NaN	NaN	5.0	5.0	0.0	First U.S. dirigible Akron exploded just offsh...
2	08/06/1913	NaN	Victoria, British Columbia, Canada	Private	-	NaN	Curtiss seaplane	NaN	NaN	1.0	1.0	0.0	The first fatal airplane accident in Canada oc...
3	09/09/1913	18:30	Over the North Sea	Military - German Navy	NaN	NaN	Zeppelin L-1 (airship)	NaN	NaN	20.0	14.0	0.0	The airship flew into a thunderstorm and encou...
4	10/17/1913	10:30	Near Johannisthal, Germany	Military - German Navy	NaN	NaN	Zeppelin L-2 (airship)	NaN	NaN	30.0	30.0	0.0	Hydrogen gas which was being vented was sucked...

We aim to explore the Operator variable to see if certain types of operators are associated with more plane crashes. We will also examine if a particular type of plane results in more fatalities in a crash than others, and also identify the leading causes of crashes. Essentially, we want to explore whether certain flying conditions—such as the location, plane type, number of people aboard, etc.—are more dangerous for flying.

Additionally, a time series graph will be useful to examine if plane crash fatalities have increased or decreased over time. It could also be valuable to create a heatmap to identify any "hotspot" locations for crashes around the world. Some challenges will arise with handling NaN values, especially in the flight number and registration variables. This means we will focus more on variables with more complete data. Identifying the cause of the crash will be tricky, as the Summary column contains free-text descriptions. We anticipate using code to extract certain patterns from the strings in the Summary field to get a sense of the leading causes of crashes. This will likely be the most challenging part, especially to ensure we don't overlook any critical information when filtering.

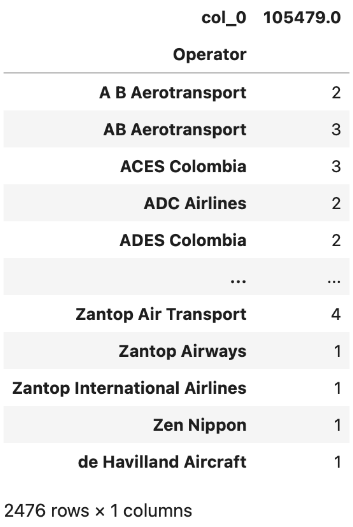
```
df = df.drop(columns=['Flight', 'Route', 'Registration', 'cn/In'])
```

Basic EDA

1.
```
pd.crosstab(df['Operator'], df['Fatalities'].sum())

# there seem to be a lot of operators, will have to clean and
combine certain operators that are similar
```

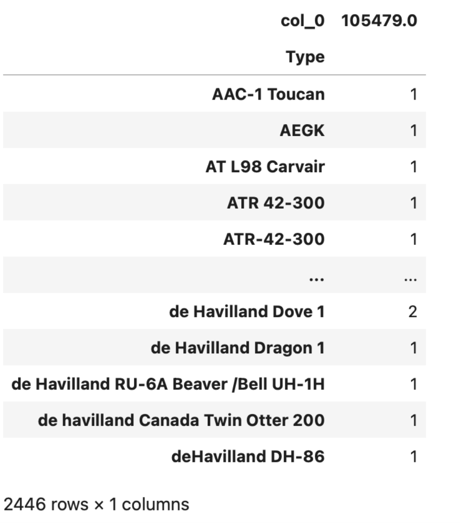
2.
```
pd.crosstab(df['Type'], df['Fatalities'].sum())

# again lots of rows need to be cleaned, also doesn't appear to
have a correlation?
```

3.
```
df['Year'] = pd.to_datetime(df['Date']).dt.year

pd.crosstab(df['Year'], df['Fatalities'].sum())

# want to observe if plane crashes have increased over the years
```



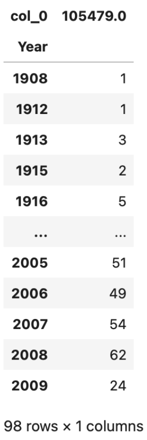
col_0	105479.0
Operator	
A B Aerotransport	2
AB Aerotransport	3
ACES Colombia	3
ADC Airlines	2
ADES Colombia	2
...	...
Zantop Air Transport	4
Zantop Airways	1
Zantop International Airlines	1
Zen Nippon	1
de Havilland Aircraft	1

2476 rows × 1 columns



col_0	105479.0
Type	
AAC-1 Toucan	1
AEGK	1
AT L98 Carvair	1
ATR 42-300	1
ATR-42-300	1
...	...
de Havilland Dove 1	2
de Havilland Dragon 1	1
de Havilland RU-6A Beaver /Bell UH-1H	1
de havilland Canada Twin Otter 200	1
deHavilland DH-86	1

2446 rows × 1 columns

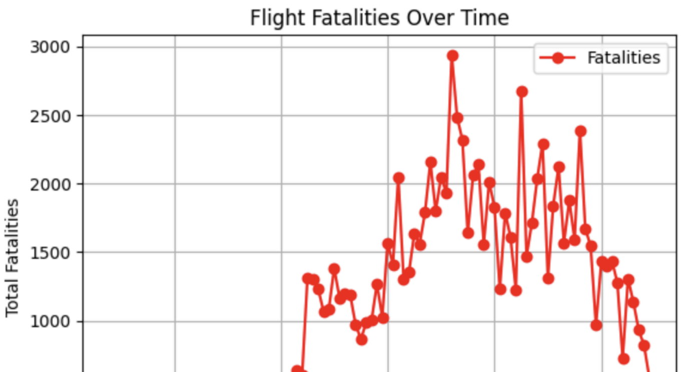


col_0	105479.0
Year	
1908	1
1912	1
1913	3
1915	2
1916	5
...	...
2005	51
2006	49
2007	54
2008	62
2009	24

98 rows × 1 columns

Next, we decided to graph some of our findings.

```
df['Date'] = pd.to_datetime(df['Date'], errors='coerce') # date-time
format

df['Year'] = df['Date'].dt.year # extracting the year

fatalities_by_year = df.groupby('Year')['Fatalities'].sum() # group by
year

plt.plot(fatalities_by_year,
marker='o', linestyle='-',

color='red', label='Fatalities')
# simple graph to show trend
```



```
plt.title('Flight Fatalities Over Time')
plt.xlabel('Year')
plt.ylabel('Total Fatalities')
plt.grid(True)
plt.legend()
plt.show()
```

Overall, we can conclude that a lot of our variables need heavy cleaning and grouping - specifically, the location, operator, plane type, and summary variables. By cleaning and grouping these variables we will be able to draw more significant conclusions and be able to cross-tabulate variables more effectively to draw out potential patterns.

Methods

We are focusing on airplane crashes and fatalities since 1908, looking at a multitude of factors to see what caused the crash. Our guiding question: What factors (such as ‘operator’, ‘type’, and ‘crash summary’) are associated with more deadly plane crashes (a 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, the number of fatalities, etc). Our group observed that there was a spike in plane crashes between 1940 and 1980, which we attribute to the periods of war occurring around the world at 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 we 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.

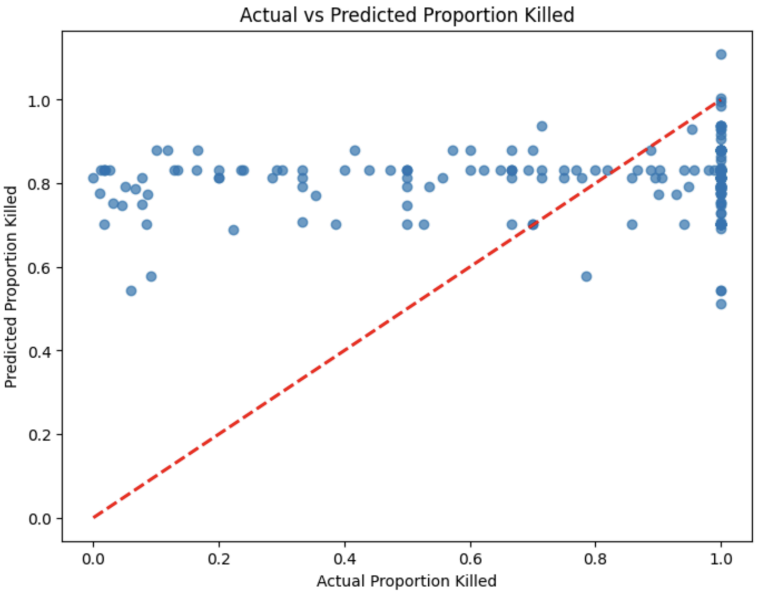
To prepare our data for modeling, we aim to use one-hot encoding and PCA in our regression model to incorporate multiple variables in our predictive model. We will use one-hot encoding for categorical variables like “Operator”, “Type”, and “Cause” to ensure that the model can interpret these values numerically. We will get the “Cause” variable from “Summary” by cleaning up the data so it is grouped by a single word rather than an entire description. 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 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 its 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 considered when looking at plane crashes, the data can get very messy very fast and become more difficult to sort, analyze, and make a model out of.

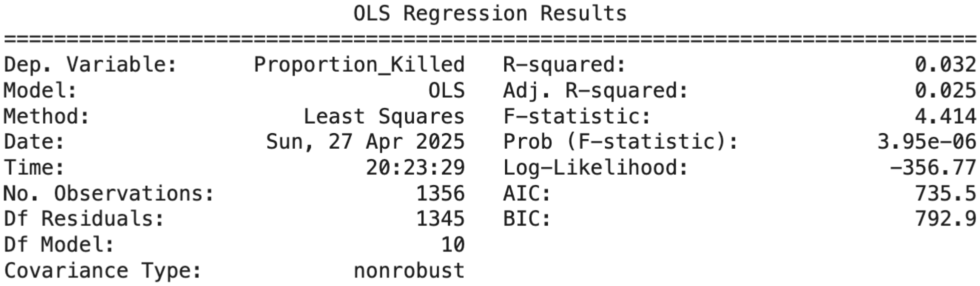
Results

The first model we created was a linear regression that examined the proportion of deaths on the plane relative to the number aboard. Our model presented us with this graph. Our mean squared error is 0.0943, and our R^2 ended up being 0.0140. We acknowledge



that these numbers are extremely low, suggesting no relationship between our variables, and that crash fatality outcomes are highly sporadic and can be influenced by many unpredictable factors not captured in our dataset and model. One possible reason for the poor model performance is that linear regression assumes a linear relationship between each input variable and the response variable, which may not reflect the true structure of our data. In reality, crash severity is likely influenced by a combination of nonlinear and interacting effects—for example, a mechanical failure during bad weather could be more dangerous than either factor on its own. Additionally, linear regression is sensitive to outliers, and our dataset includes incidents with both total fatalities and complete survivals, which could skew the regression coefficients.

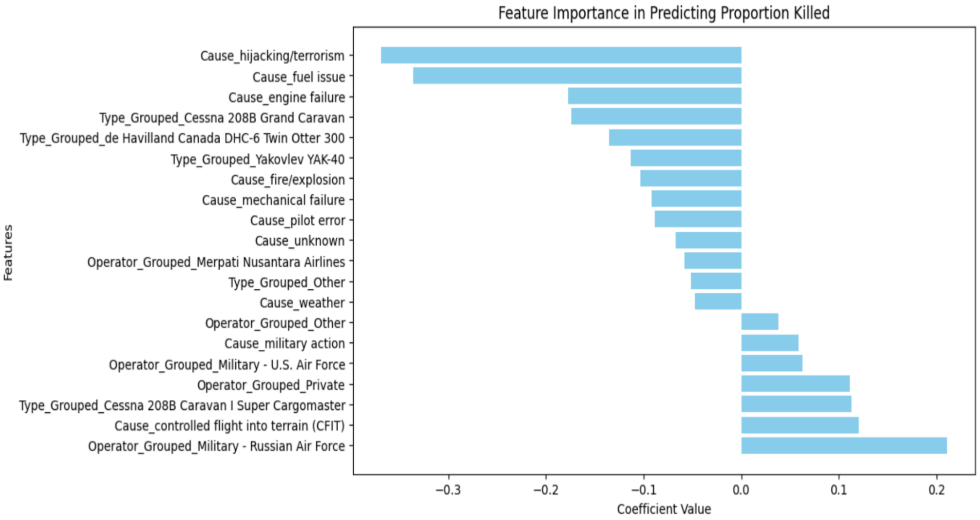
We also noted that some of our categorical variables, once one-hot encoded, resulted in sparse matrices with hundreds of dimensions, making it harder for the linear model to generalize patterns. The sheer number of aircraft types and operators likely diluted any strong signal that could exist in a smaller or more consistent subset. Overall, while linear regression provided a straightforward starting point, the nature of our data likely requires more flexible modeling techniques to uncover meaningful relationships.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:      Proportion_Killed   R-squared:                       0.032
Model:                            OLS   Adj. R-squared:                  0.025
Method:                 Least Squares   F-statistic:                     4.414
Date:                Sun, 27 Apr 2025   Prob (F-statistic):           3.95e-06
Time:                        20:23:29   Log-Likelihood:                -356.77
No. Observations:                1356   AIC:                             735.5
Df Residuals:                    1345   BIC:                             792.9
Df Model:                          10
Covariance Type:            nonrobust
```

We did end up conducting PCA after our linear regression to see if we could improve our R^2; however, after PCA, our MSE was equal to 0.0960, and our R^2 was -0.0035. After seeing the poor results from our first model, we made the Cause variable a one-hot encoded dummy variable and tried to find a relationship between that variable and the proportion killed on the flight. Our model which aims to predict the proportion of fatalities in plane crashes based on several variables such as plane type, operator, etc., is statistically significant which we see based on the Prob (f-score). However, the R-squared value of the model is super low, meaning it is only explaining 3.2% of variation in fatality rates. This means that some factors have an association with crash outcomes. However, the majority of influence for an airplane crash is not captured by the model. This new R^2 is still not very telling of any relationship, but we figured any improvement was something to show, given the difficult nature of our data.

When analyzing the importance of our features, we found that plane crashes caused by hijacking or terrorism had the highest positive impact on the proportion killed, followed closely by fuel issues and engine failures. It seemed like the aircraft type played somewhat of a role, with certain models like the Cessna 208B Grand Caravan and the Havilland Canada DHC-6 Twin Otter 300 showing notable influence on the outcomes. However, there were over two thousand different types of aircraft within the data set, so the variability could make it difficult to detect the relevance of specific models. The operator feature, such as private companies versus military branches, appears to contribute less to the model compared to cause and aircraft type, which explains why reducing the features to only include the cause improved our R^2 value slightly. Even though the improvement



in the value was practically negligible, it was still our most significant discovery and perhaps an observation that could be used in future studies with a better data structure.

Conclusion

After running our regression and analysis, we were unable to accurately predict the fatality of a plane crash using the factors of a given crash. The root cause of this is the actual data at hand; looking back, we were doomed from the start. Each crash had a unique summary, which we grouped into 12 different 'Cause' categories, many of which had very few entries. The second largest cause was 'unknown', which doesn't tell us anything. Despite this, 'Cause' was still our most useful variable in the regression. Other variables, such as 'Operator' and 'Type', similarly had far too many unique values to find a significant correlation between the variables and the outcome. We tried our best to clean and mold the data into a more usable form for the regression, but the data that we started with simply wasn't built to be used in this way. This is disappointing, but it taught us a valuable lesson. Not all data is created equally, especially in the context of analytics. We should have found a data set that had many more numerical variables, so we wouldn't be forced to one-hot encode the 12 unique values of the 'Cause' variable or group the incredibly specific operators and types.

Our overarching goal was to see whether we could find any consistent relationships between features of a crash, such as operator, aircraft type, or cause, and how deadly the crash was. We focused on crashes after 1990 to keep the analysis more relevant to modern aviation. From there, we cleaned the data, grouped categories, and built regression models to predict the proportion of fatalities. Even though our models didn't perform well in terms of predictive power, the process still revealed some useful insights. Certain causes, like hijackings or mechanical failures, were much more closely tied to high fatality rates than other factors. In contrast, features like operator or aircraft type didn't show much influence, likely due to how many unique entries and inconsistencies were in those columns. So while our model didn't give us strong predictions, it did help us see which parts of the data had the most potential, and which ones held us back.

We also have some suggestions for how this dataset could be improved for the future. There should have only been a handful of predetermined causes used to explain a crash instead of the free-form summary that was initially included. This would prevent so many unknown causes and group the entries into clearer categories. Another variable that could have been organized better is location. We were unable to use this variable due to the inconsistency of the entries. If the location were just the name of the country or the continent, it could be used much more easily. These fallbacks of the data reflect a larger problem with data collection in general. We encountered this issue repeatedly throughout our analysis. Variables like aircraft type, operator, and location contained so many unique and inconsistently formatted entries that it became difficult to extract meaningful patterns or use them effectively in our models. Even after applying grouping strategies and data-cleaning techniques, the underlying variability remained a barrier to producing strong results. In hindsight, the greatest limitation was not our modeling approach, but the structure of the dataset itself. What seemed like a straightforward analytical task at first quickly shifted into a more complex data-wrangling effort, emphasizing just how essential clean, standardized inputs are for any kind of predictive analysis. Data collection should always be standardized where possible, meaning that the possible values of any categorical variable are predetermined. Since a lot of the plane crash data is old, it makes sense that there is very little consistency since data wasn't being used in the way that it is today. Moving forward, data should be collected in a more controlled format with analytics in mind.

References/Bibliography

Li, G., & Baker, S. P. (2007). Crash risk in general aviation. *Jama*, *297*(14), 1596-1598.