Final Project - Ryan Winston

MSBA 320

19 April, 2023

Table of Contents

Introduction……3

Data Collection……..…….……...4

Results & Interpretation…….……5

Conclusion……...16

References……17

Introduction

I performed an analysis on aircraft accidents in the United States from 1/1/2012 - 12/31/2021. I chose this topic because flying is always relevant and the consequences of flight accidents are obviously devastating. In addition, the topic I chose was partially a result of me simply wanting to use a reliable data set. The data set I obtained is from the National Transportation Safety Board (NTSB), an independent US federal agency. On the NTSB website, they have an interactive PowerBI dashboard. When I saw this dashboard, this indicated to me that there must be a robust and useful data set the dashboard pulls from. I eventually found the source data near the bottom of the page and used it for my analysis

The aspect of the data set I am most interested in is the fatality rate. The data set has attributes called FatalInjuries and InjuryLevel. The FatalInjuries attribute shows how many fatalities were in an accident, and the InjuryLevel attribute is a binary attribute which shows simply whether or not an accident contained at least one fatality. My goal for this analysis was to discover which attributes contribute most to fatalities. In order to conduct the analysis, I ended up using a few of the statistical methods practiced in the MSBA 320 course. The methods I used were descriptive statistics, time series analysis, and binomial regression analysis. I executed these methods in Jupyter Notebook using the Python programming language.

Before proceeding, I want to note some of the limitations of the data set. All records in the data set are records of US non-military aircraft accidents from 2012 - 2021. The data set does not contain records of flights that did not result in an accident. The data set also does not contain information on how many total passengers were on each flight.

Data Collection

The data collection process did not present any overwhelming challenges. The data set is available as a CSV file on the NTSB website. The file downloads as 3 separate sheets on a single Excel workbook. Jupyter Notebook does not easily handle pulling in more than 1 sheet, so I saved each sheet as its own CSV file to make data loading into Jupyter Notebook smoother. I was not able to use one of the sheets called “findings” in the analysis even though it loaded into Jupyter Notebook successfully. This is because my goal was to join the 3 sheets into a single dataframe using Python. However, the data from the “findings” sheet did not match up well 1:1 with the other 2 sheets. Fortunately, the “findings” sheet did not contain much useful data for the analysis anyway. The 2 other sheets, “accidents” and “aircraft” were joinable with Python on a common attribute within the 2 sheets called “MKey”. Each accident is identifiable with a unique MKey. The only issues that arose from this were situations where an accident contained more than 1 aircraft. In those situations, the “aircraft” and “accidents” sheets did not match up 1:1 due to there being more total aircraft than the total number of accidents. However, these situations were very uncommon and would not have a significant detriment to the quality of the analysis. I am not certain how Python performs dataframe joins exactly, but using the pd.merge() function seemed to fill in the blanks for the accidents with multiple aircraft, thus bringing the total records up to the number of aircraft involved in accidents as found on the “aircraft” CSV file. The exact code is as follows:

aviation = pd.merge(aircraft, accidents, on=["MKey"])

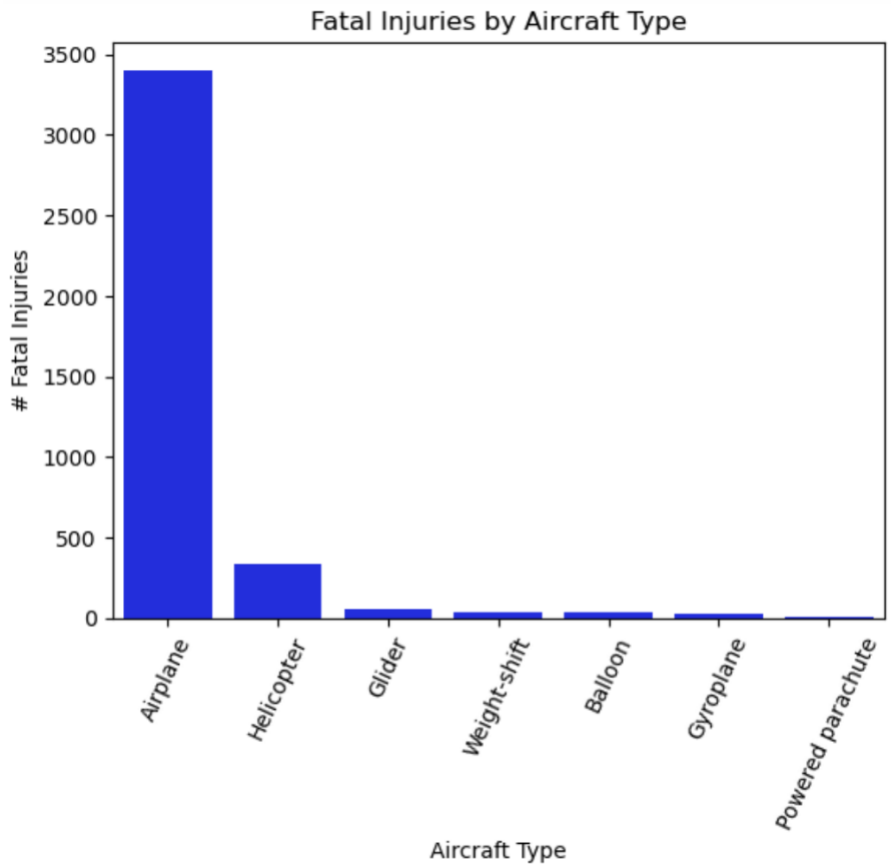
Out of 12,368 documented accidents, there were 12,506 aircraft involved. I subsequently felt comfortable proceeding with the analysis.

Results & Interpretation

The first insight I found was when ranking the different categories of aircraft by total number of fatalities. The table is as follows:

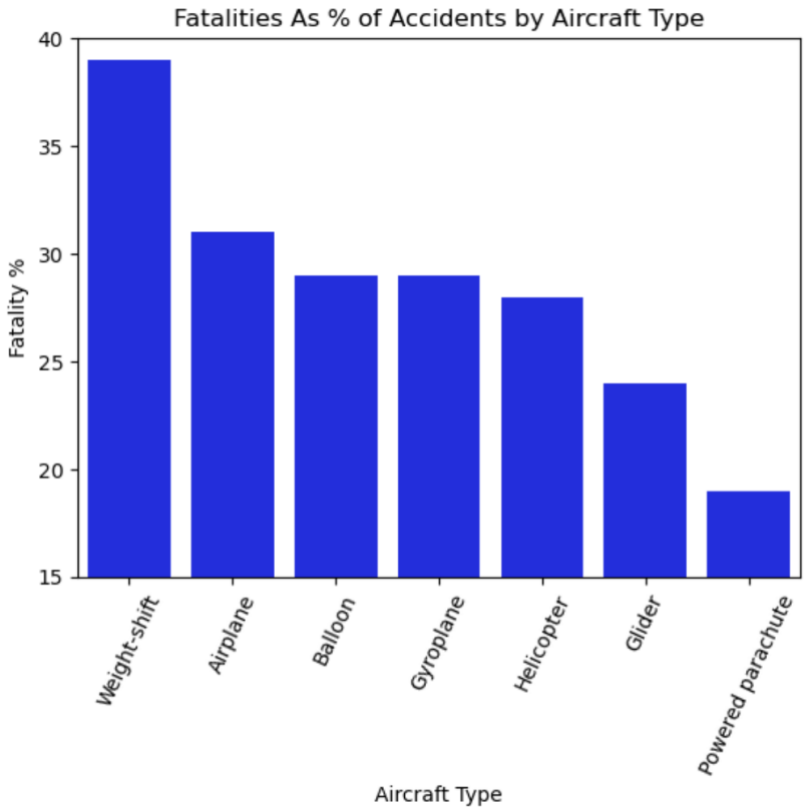
	AircraftCategory	AircraftNumber	FatalInjuries	SeriousInjuries
0	Airplane	10869	3403.0	1779.0
4	Helicopter	1202	337.0	242.0
2	Glider	227	55.0	55.0
9	Weight-shift	89	35.0	16.0
1	Balloon	111	32.0	94.0
3	Gyroplane	94	27.0	18.0
5	Powered parachute	42	8.0	35.0

In this table, the AircraftCategory is the type of aircraft involved in accidents, AircraftNumber is the number of aircraft involved in accidents. FatalInjuries is the number of deaths involved in accidents, and SeriousInjuries is the number of serious non-fatal injuries involved in accidents. According to the table, airplane accidents caused the most fatalities, followed by helicopter accidents. This table can be better visualized with the following bar chart:



With this visualization, we can see more precisely how prevalent airplane accident fatalities were compared to other aircraft accidents. The only other significant number of fatalities came from helicopter accidents. After seeing this, I decided to focus the analysis mostly on airplanes and helicopters.

Before moving on to the airplane and helicopter analysis, I was curious to see how deadly different accidents were by using fatalities as a percentage of the number of accidents for each aircraft type. I won't show the table for the sake of brevity, but a rank order bar chart displays an interesting result:

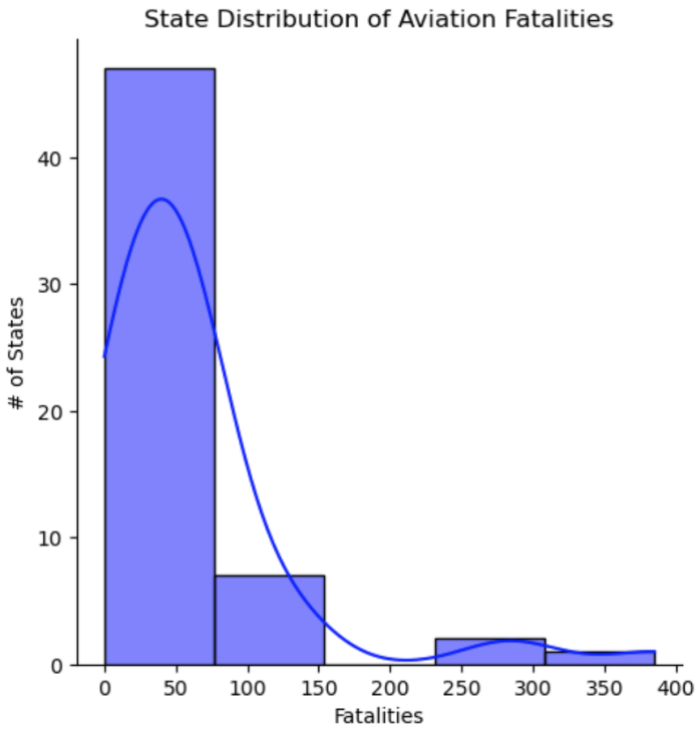


Weight-shift aircraft accidents were significantly more fatal than the rest of the aircraft types as a percentage of each category's total accidents. It is also worth noting that airplane accidents and helicopter accidents led to a similar rate of fatalities as a percentage of accidents of their respective aircraft types. I had no idea what a weight-shift aircraft was before this project, so I did a quick Google/Wikipedia search to get a clue:

https://upload.wikimedia.org/wikipedia/commons/thumb/c/cf/AirBorne_XT912_Tourer_microlight.jpg/800px-AirBorne_XT912_Tourer_microlight.jpg

It is not hard to see why an accident in this type of aircraft is so much more likely to lead to a fatality than other aircraft types. There is almost nothing surrounding the person to absorb any impact.

The next sub-analysis related to the location of aircraft accidents. I performed this analysis only for airplanes and helicopters, as those aircraft types were by far the most prevalent in accidents. The distribution of plane and helicopter fatalities by state is as follows:

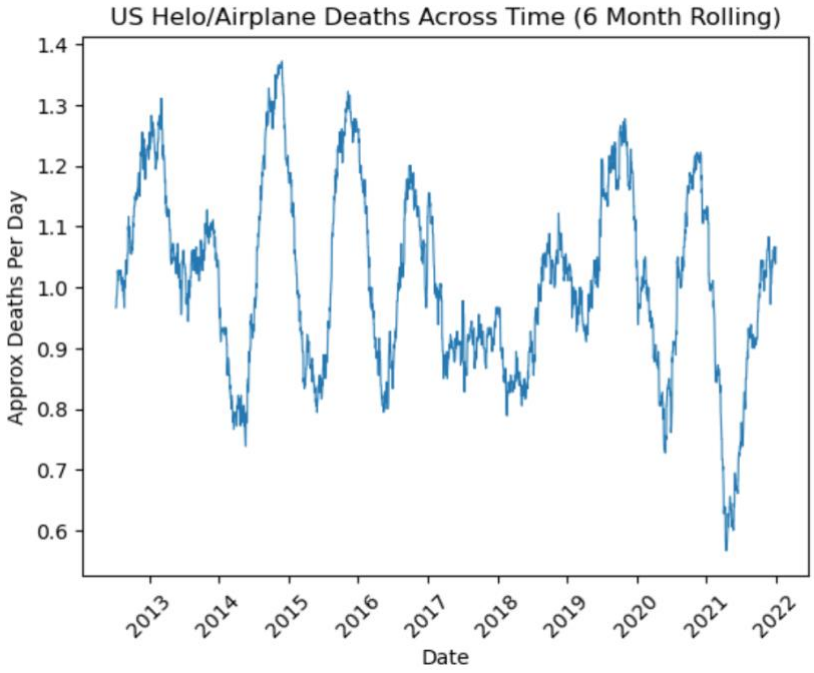


This histogram shows that almost all states had less than 150 airplane and helicopter fatalities, while a small number of states had 250 fatalities or more. In fact, according to the table this histogram is pulling from, only 3 states had more than 250 plane or helicopter accidents:

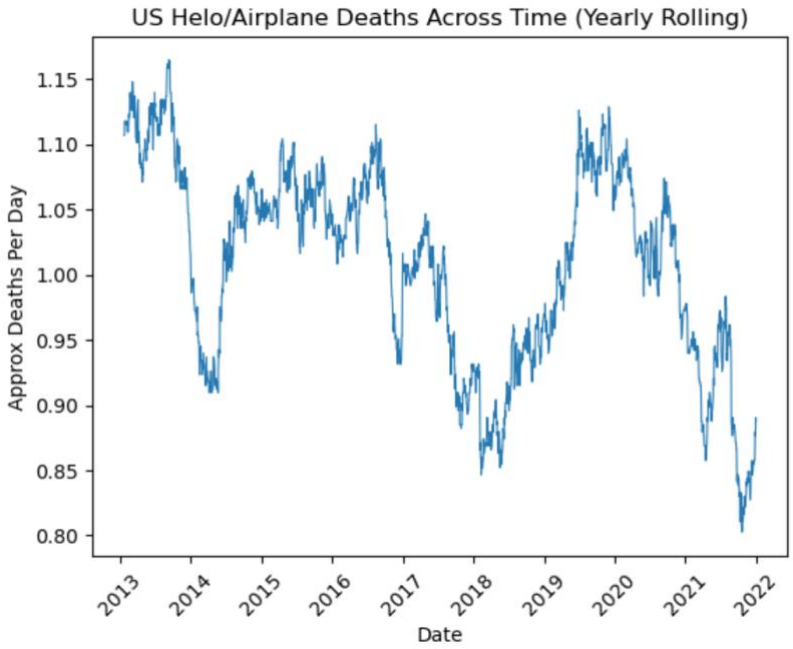
StateOrRegion	AircraftNumber	FatalInjuries	SeriousInjuries
California	1144	385.0	182.0
Texas	1031	295.0	246.0
Florida	924	273.0	167.0
Georgia	328	140.0	52.0
Alaska	724	131.0	80.0
Arizona	443	124.0	69.0
Colorado	369	114.0	52.0
New York	256	95.0	40.0

This table showing the top 8 states ranked by number of plane/helicopter fatalities shows that only California, Texas, and Florida had more than 250 fatalities. All other states had less than 140 fatalities.

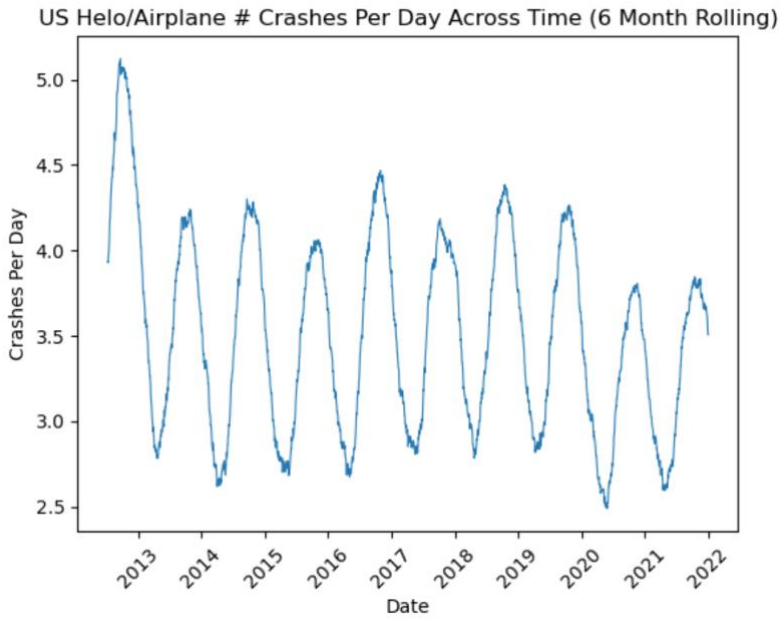
Though I considered analyzing the top 3 states further, I decided that pivoting the analysis toward time series analysis would be much more interesting since the data set covers 10 years of data. With such a long timeframe and many categorical attributes, there were many ways to break down the time series analysis. But I managed to break the analysis down into a few time series plots. I began by creating time series plots based on the average number of deaths per accident (fatality rate) across time, as this was the most readily available analysis. The plots weren't exactly what I was looking for, so I took some extra time to create plots based on plane/helicopter deaths per day using a rolling 6-month average:



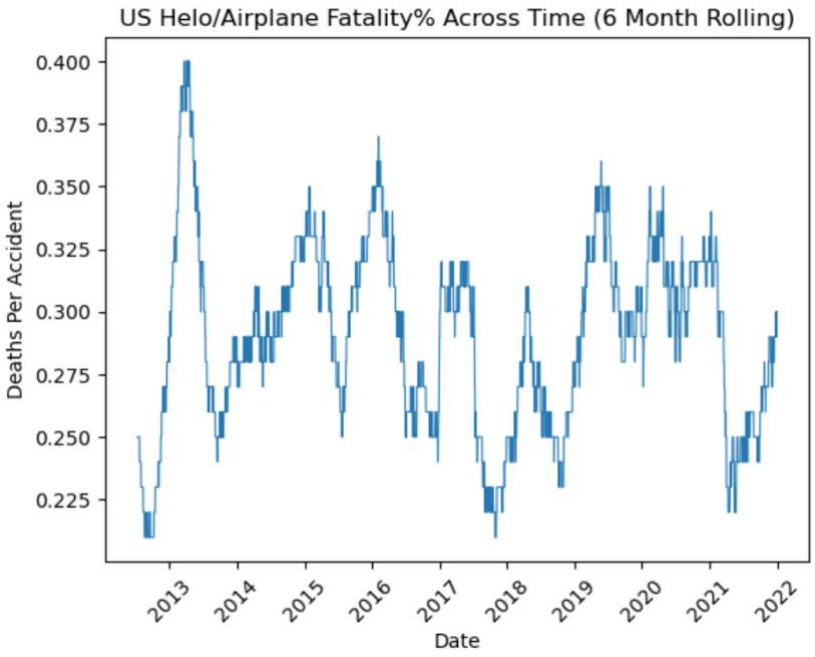
Stunning! This plot shows that the average US helicopter/plane deaths per day has a seasonal component/pattern. It appears that the deaths per day spikes annually around the holidays or the new year, but plummets during the Summer. We also see a local minimum in 2021 that is far lower than the other local minima across the timeframe. Could this be Covid-19 related? We can re-visualize the time series to view the longer-term trend:



With the yearly rolling average for deaths per day, we can see a slight general downward trend across the 10 year span. We see an especially drastic decline in deaths per day around the time Covid-19 started in 2020. However, Covid-19 cannot explain the large dip in deaths per day in 2018. I don't know what could have caused this dip. And since the overall trend is slightly downward, does this mean aircraft are generally becoming safer even in the event of an accident? Or are there simply less accidents? The following time series plots may hint at an answer, though not indisputably conclusive:



Going back to the 6 month rolling average, the plot above shows the 6 month rolling average for accidents per day rather than the deaths per day that the previous 6 month rolling average showed. The accidents per day doesn't conclusively show a general downward trend in the number of accidents over time. The exception may be in 2013 versus the Covid years. But 2014 - 2019 don't show anything other than a seasonal component. We can't really say that the general downward trend in flight accident fatalities is due to less overall crashes. But does this mean that the crashes that are happening are simply becoming less fatal? Let's look at the fatality rate over time:



Using the same 6 month rolling period, the deaths per accident looks a little bit different from the accidents per day. The local minima and local maxima appear to decline over time, indicating that aircraft safety features are more likely to be the cause of less flight deaths than any sort of general decline in overall accidents! However, this exact conclusion can't necessarily be confirmed based on this analysis. We can only infer this as a possible conclusion.

This leaves one more glaring question: what are the most common factors leading to fatal accidents? I decided to do a binomial GLM analysis to help determine the most predictive factors. I performed the binomial GLM analysis using the InjuryLevel attribute as a binomial variable. The InjuryLevel attribute has only 2 values: Fatal or Non-Fatal. After converting the Fatal values to 1 and the Non-Fatal values to 0, the binomial GLM analysis was straightforward. Unfortunately, due to the excessive amount of values in other attributes, a full GLM analysis did not work so well in Jupyter Notebook as it could not run the GLM function on too many variables. I ran the binomial GLM for the InjuryLevel in relation to the PurposeOfFlight and PhaseOfFlight only. This is a suboptimal analysis because I wanted to include

AircraftMake, AircraftModel, and DefiningEvent in the model. Hopefully in a real-world setting there is a more powerful platform than Jupyter Notebook to run Python for inferential/predictive analysis. Regardless, I ran the following code to reveal the following binomial GLM:

model = smf.glm('InjuryLevel ~ PurposeOfFlight + PhaseOfFlight', data=planeUStime, family=sm.families.Binomial()).fit()

model.summary()

Generalized Linear Model Regression Results

Dep. Variable:	InjuryLevel	**No. Observations:**	10530
Model:	GLM	**Df Residuals:**	10500
Model Family:	Binomial	**Df Model:**	29
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-3845.9
Date:	Wed, 19 Apr 2023	**Deviance:**	7691.8
Time:	16:44:57	**Pearson chi2:**	1.05e+04
No. Iterations:	21	**Pseudo R-squ. (CS):**	0.1834
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-2.1828	0.162	-13.439	0.000	-2.501	-1.864
PurposeOfFlight[T.Aerial observation]	0.8011	0.302	2.655	0.008	0.210	1.393
PurposeOfFlight[T.Air drop]	-20.3781	2.42e+04	-0.001	0.999	-4.75e+04	4.75e+04
PurposeOfFlight[T.Air race/show]	0.9730	0.369	2.636	0.008	0.249	1.697
PurposeOfFlight[T.Banner tow]	0.2450	0.427	0.574	0.566	-0.592	1.082
PurposeOfFlight[T.Business]	1.6738	0.226	7.405	0.000	1.231	2.117
PurposeOfFlight[T.Executive/Corporate]	1.5570	0.530	2.940	0.003	0.519	2.595
PurposeOfFlight[T.Ferry]	0.8975	0.435	2.062	0.039	0.044	1.751
PurposeOfFlight[T.Firefighting]	2.5777	0.786	3.281	0.001	1.038	4.117
PurposeOfFlight[T.Flight test]	0.8623	0.307	2.808	0.005	0.260	1.464
PurposeOfFlight[T.Glider tow]	0.7851	0.686	1.145	0.252	-0.559	2.129
PurposeOfFlight[T.Instructional]	0.6526	0.173	3.780	0.000	0.314	0.991
PurposeOfFlight[T.Other work use]	1.3782	0.358	3.845	0.000	0.676	2.081
PurposeOfFlight[T.Other/Unknown]	2.7253	0.597	4.569	0.000	1.556	3.894
PurposeOfFlight[T.Personal]	1.2892	0.150	8.574	0.000	0.994	1.584
PurposeOfFlight[T.Positioning]	1.2628	0.274	4.613	0.000	0.726	1.799
PurposeOfFlight[T.Public aircraft]	1.0971	0.386	2.841	0.005	0.340	1.854
PurposeOfFlight[T.Skydiving]	0.2662	0.441	0.604	0.546	-0.598	1.130
PurposeOfFlight[T.Unknown]	0.5794	0.647	0.895	0.371	-0.689	1.848
PhaseOfFlight[T.Emergency Descent]	0.5512	0.330	1.670	0.095	-0.096	1.198
PhaseOfFlight[T.Enroute]	0.0371	0.085	0.437	0.662	-0.129	0.203
PhaseOfFlight[T.Initial Climb]	0.1740	0.094	1.855	0.064	-0.010	0.358
PhaseOfFlight[T.Landing]	-3.2418	0.154	-21.029	0.000	-3.544	-2.940
PhaseOfFlight[T.Maneuvering]	1.0644	0.097	10.970	0.000	0.874	1.255
PhaseOfFlight[T.Post-Impact]	-21.5758	2.4e+04	-0.001	0.999	-4.71e+04	4.71e+04
PhaseOfFlight[T.Standing]	-1.0892	0.269	-4.054	0.000	-1.616	-0.563
PhaseOfFlight[T.Takeoff]	-1.2455	0.116	-10.714	0.000	-1.473	-1.018
PhaseOfFlight[T.Taxi]	-4.6768	1.004	-4.656	0.000	-6.646	-2.708
PhaseOfFlight[T.Uncontrolled Descent]	1.8755	0.629	2.982	0.003	0.643	3.108
PhaseOfFlight[T.Unknown]	1.0259	0.214	4.797	0.000	0.607	1.445

The binomial GLM shows that according to coefficients and P-values, the most reliable predictors of a plane crash being fatal are:

1. If the purpose of the flight is for firefighting or if the purpose is unknown
2. If the defining accident-causing event occurs during an Uncontrolled Descent phase

According to coefficients and P-values, the most reliable predictors of a plane crash being non-fatal are:

1. If the phase of the flight is in the taxi stage or the landing stage

In my opinion, the biggest surprise here is that the landing stage of the flight is a reliable predictor of a non-fatal accident. From the perspective of an average US citizen, my initial hypothesis would have been that the deadliest phase of the flight is the landing phase. However, the binomial GLM suggests otherwise.

Conclusion

My analysis of the NTSB aviation accident data set covered many angles. So what can be concluded when tying all of this together? There are several interesting suggestions/inferences from the data even if there aren't many rock-solid conclusions.

1. Airplane accidents are the leading cause of non-military aviation fatalities in the US
2. California, Texas, and Florida have significantly more aviation fatalities than the other US states
3. The daily average amount of airplane & helicopter fatalities has a seasonal component, with the holiday season & new year consistently delivering a spike in fatalities
4. There was an unsurprising decline in accidents/fatalities during the Covid-19 years
5. Even disregarding a decline in accidents during the Covid-19 years, there was still a slight downward trend in the deadliness of airplane/helicopter accidents over time
6. Firefighting missions increase the likelihood of an accident being fatal
7. Fatalities are generally less likely to occur during aircraft landings than during other flight phases

References

National Transportation Safety Board. General Aviation Accident Dashboard: 2012-2021. https://www.ntsb.gov/safety/data/Pages/GeneralAviationDashboard.aspx#AVSpreadsheet

Wikipedia. (2022.) Ultralight trike. https://en.wikipedia.org/wiki/Ultralight_trike#/media/File:AirBorne_XT912_Tourer_microlight.jpg