# Aircraft Accident Analysis

IST718 Big Data Analytics | Group Project – Final Report Spring 2019

BreeAna B. Baker

Editt Gonen-Friedman

Joseph R. Hartman

Hongseok Kim

##### **Executive Summary of Findings and Recommendations**

Improving aircraft safety and minimizing risk has always been an important task in the aviation field. Usually, accidents are examined on some sort of periodic basis, however this might leave out potential insights about what factors into aviation accidents on a wholistic level. To accomplish this, we’ve examined aircraft accident data from the NTSB (National Transportation Safety Board) from the 1980s until recently to try and answer these questions using visualizations and informative features from training random forest (RF) models. Despite the observation of the general trend over time that the number of accidents has decreased every year, we still noted that the majority of accidents tend to happen around general aviation types of flights (private transport and leisure) and our RF models deemed factors around general aviation and more specifically the aircraft used in general aviation (such as types of planes, purpose of flight, number of engines etc.) to be important in determining accident severity. Our RF models also determined factors like weather conditions and the landing phase were also important features. Overall, we recommend that the NTSB focus improving safety in general aviation, possibly though stricter pilot’s license requirements, better training and access to weather information, and setting better standards for personal aircrafts in terms of instruments.

##### **Specification**

**Problem**

Team West seeks to help stakeholders such as aircraft manufacturers, insurers, the FAA (Federal Aviation Administration) and the NTSB (National Transportation Safety Board) with risk assessment and analysis related to aircraft crashes. This is important as recent Boeing 737 crashes highlighted a problem in aviation: despite extensive automation in modern aircraft, aircraft still crash, bringing on heavy losses of life and monetary losses.

One might ask, “Isn’t this redundant? Aren’t aircraft crashes already investigated?” The NTSB investigates every incident in the US and many incidents abroad, and then makes safety recommendations to FAA and other authorities. The analysis is done one incident at a time, with some annual summaries. We want to look at the data as a whole, look for trends, discover factors that might affect fatalities and damage, and see if insight gained from such incident data analysis can be used to make recommendations.

**Hypothesis**

We hypothesize that weather conditions, the model of the airplane, and the location play some role in the severity of accidents. The problem is complex, and looking at the factors one at a time might not suffice. We posit that we can determine what factors lends themselves to determine the severity of an accident by using Random Forest (RF) models and pulling the important features that the model uses to predict values.

**The Data**

The data is composed of 83,044 accident report entries that are described by 32 variables. This dataset contains information about incidents that have been investigated; information includes incident location, data on the aircraft involved, flight details, and fatalities or damage tinformation. The information is categorical. For example, damage has categories such as ‘destroyed’ or ‘substantial’; Location information includes country / state / city and coordinates. The accompanying data dictionary which we have enclosed as an appendix, can be found at <https://www.ntsb.gov/_layouts/ntsb.aviation/index.aspx>.

**Data manipulation and pre-processing:**

* The accident dates were split into corresponding month, day, and year
* The “InjurySeverity” attribute was broken up into “Fatal\_Bool” and “Fatal\_Counts” so the number of fatalities were put in its own column
* There was some data cleaning done for the “AircraftCarrier” (carriers like Delta would appear as Delta Airlines and Delta Airlines Inc.) so the visualizations could accurately reflect the data
* Location information was broken down to get state information, for observations by state
* Weather codes were broken down to visibility variables

Upon first glance of the data, it is very evident that not all fields are always reported for each accident. For the columns that seemed most likely to hold significance, we looked at the possible responses as well as the total number of responses that the column had. The worst offender was InjurySeverity, for which we cleaned up by converting the data to a fatalities count. Here are some examples:

InjurySeverity: Possible Responses

['Unavailable' 'Non-Fatal' 'Fatal(1)' 'Fatal(6)' 'Fatal(2)' 'Fatal(3)'

'Fatal(4)' 'Incident' 'Fatal(157)' 'Fatal(14)' 'Fatal(5)' 'Fatal(9)'

'Fatal(10)' 'Fatal(112)' 'Fatal(71)' 'Fatal(12)' 'Fatal(11)' 'Fatal(7)'

'Fatal(39)' 'Fatal(8)' 'Fatal(16)' 'Fatal(66)' 'Fatal(62)' 'Fatal(23)'

'Fatal(224)' 'Fatal(150)' 'Fatal(43)' 'Fatal(162)' 'Fatal(58)'

'Fatal(295)' 'Fatal(239)' 'Fatal(33)' 'Fatal(50)' 'Fatal(21)' 'Fatal(19)'

'Fatal(153)' 'Fatal(127)' 'Fatal(28)' 'Fatal(77)' 'Fatal(42)'

'Fatal(158)' 'Fatal(103)' 'Fatal(89)' 'Fatal(90)' 'Fatal(152)'

'Fatal(228)' 'Fatal(17)' 'Fatal(13)' 'Fatal(24)' 'Fatal(88)' 'Fatal(65)'

'Fatal(154)' 'Fatal(30)' 'Fatal(20)' 'Fatal(40)' 'Fatal(57)' 'Fatal(199)'

'Fatal(113)' 'Fatal(107)' 'Fatal(117)' 'Fatal(145)' 'Fatal(45)'

'Fatal(160)' 'Fatal(121)' 'Fatal(15)' 'Fatal(104)' 'Fatal(25)'

'Fatal(55)' 'Fatal(46)' 'Fatal(141)' 'Fatal(115)' 'Fatal(75)'

'Fatal(206)' 'Fatal(138)' 'Fatal(92)' 'Fatal(26)' 'Fatal(265)'

'Fatal(118)' 'Fatal(44)' 'Fatal(64)' 'Fatal(18)' 'Fatal(83)' 'Fatal(143)'

'Fatal(60)' 'Fatal(131)' 'Fatal(169)' 'Fatal(217)' 'Fatal(80)'

'Fatal(229)' 'Fatal(87)' 'Fatal(52)' 'Fatal(97)' 'Fatal(35)' 'Fatal(29)'

'Fatal(125)' 'Fatal(349)' 'Fatal(34)' 'Fatal(70)' 'Fatal(230)'

'Fatal(110)' 'Fatal(123)' 'Fatal(189)' 'Fatal(72)' 'Fatal(54)'

'Fatal(68)' 'Fatal(132)' 'Fatal(37)' 'Fatal(56)' 'Fatal(47)' 'Fatal(27)'

'Fatal(73)' 'Fatal(111)' 'Fatal(174)' 'Fatal(144)' 'Fatal(270)'

'Fatal(156)' 'Fatal(82)' 'Fatal(256)' 'Fatal(31)' 'Fatal(135)']

Number of responses: 83044

-------------------------------------------

AircraftDamage: Possible Responses

['Substantial' 'Destroyed' 'Minor']

Number of responses: 80415

-------------------------------------------

AircraftCategory: Possible Responses

['Airplane' 'Helicopter' 'Balloon' 'Powered Parachute' 'Weight-Shift'

'Unknown' 'Gyroplane' 'Glider' 'Ultralight' 'Blimp' 'Powered-Lift'

'Gyrocraft' 'Rocket']

Number of responses: 26306

-------------------------------------------

NumberofEngines: Possible Responses

['1' '2' '0' '3' '4']

Number of responses: 78352

-------------------------------------------

FARDescription: Possible Responses

['Part 91: General Aviation' 'Part 137: Agricultural'

'Part 135: Air Taxi & Commuter' 'Non-U.S., Non-Commercial'

'Public Aircraft' 'Part 121: Air Carrier' 'Non-U.S., Commercial'

'Part 129: Foreign' 'Part 133: Rotorcraft Ext. Load'

'Armed Forces' 'Part 103: Ultralight' 'Part 437: Commercial Space Flight'

'Part 125: 20+ Pax,6000+ lbs' 'Part 91F: Special Flt Ops.']

Number of responses: 25974

-------------------------------------------

BroadPhaseofFlight: Possible Responses

['MANEUVERING' 'CRUISE' 'APPROACH' 'TAKEOFF' 'LANDING' 'CLIMB' 'TAXI'

'DESCENT' 'STANDING' 'UNKNOWN' 'GO-AROUND' 'OTHER']

Number of responses: 76481

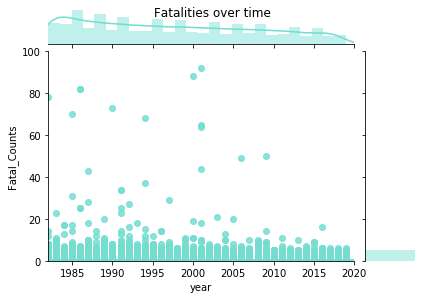
-------------------------------------------

##### **Observation**

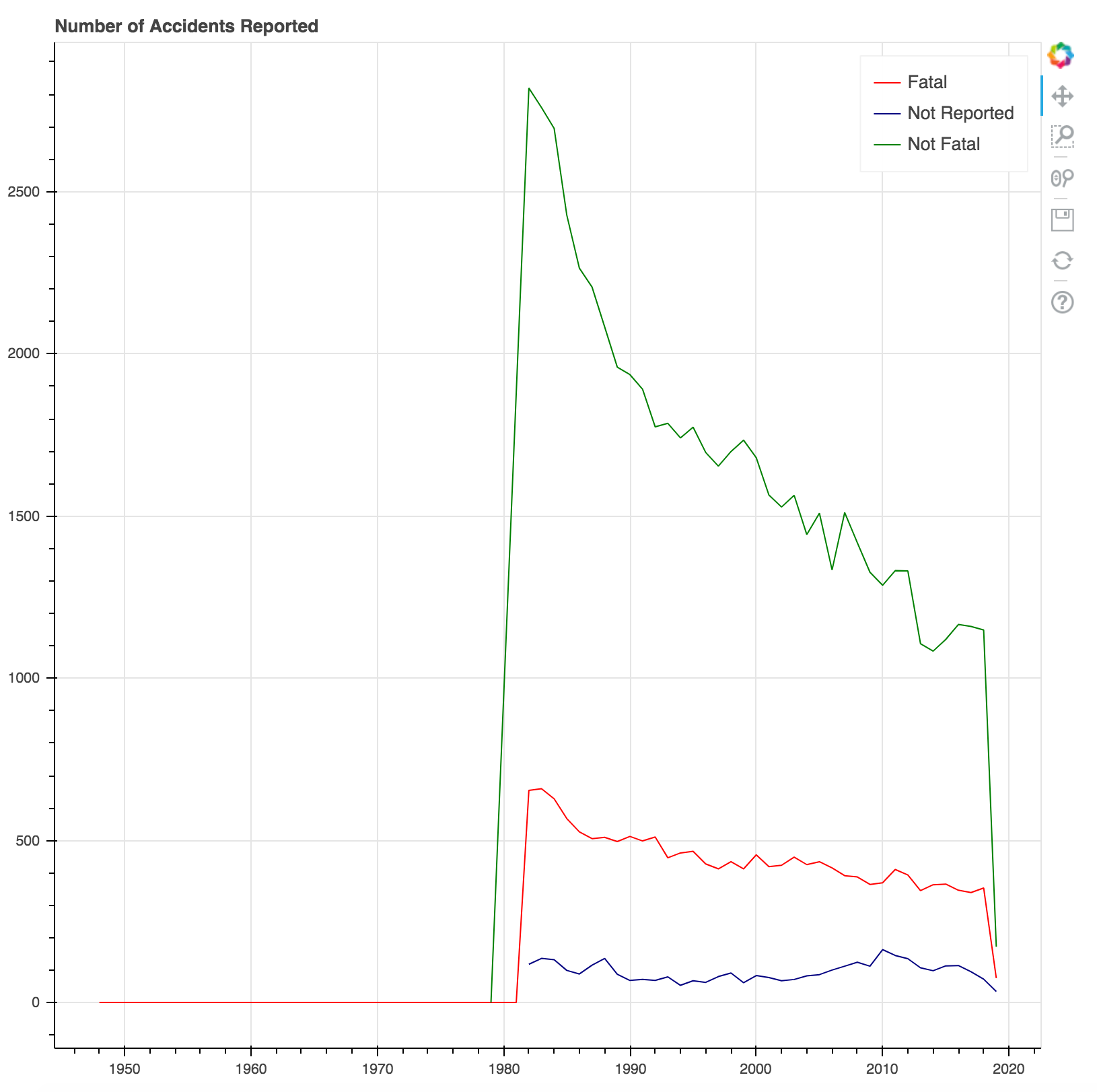
**Exploration, visualization, observations from the data:**

**Changes Over Time**

The first thing we looked at was correlations and changes over time. The fatalities occur mostly in small numbers each time, it’s more rare to have incidents with a very high number of deaths, but both frequencies and counts have a clear downward trend in this graph.



This is a different way to look at the change over time.

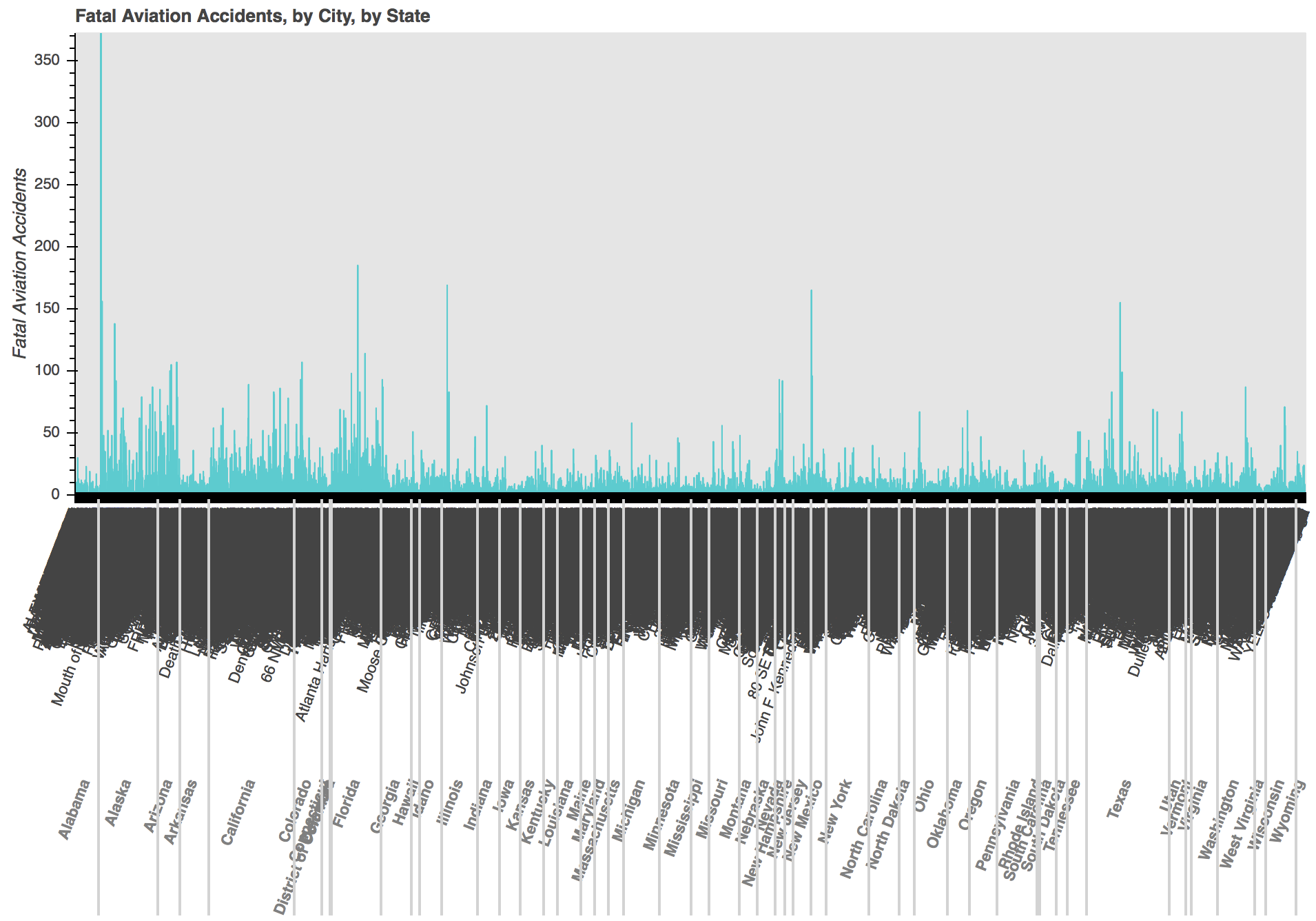


With the vast improvement in technology, we would expect that aircraft are getting safer more and more. The number of accidents overall is definitely decreasing, however the number of accidents that have at least one fatality is not decreasing as rapidly, but it is decreasing. The graph appears to drop suddenly in the end, but that is only due to the fact that we have only parts of 2019 data.

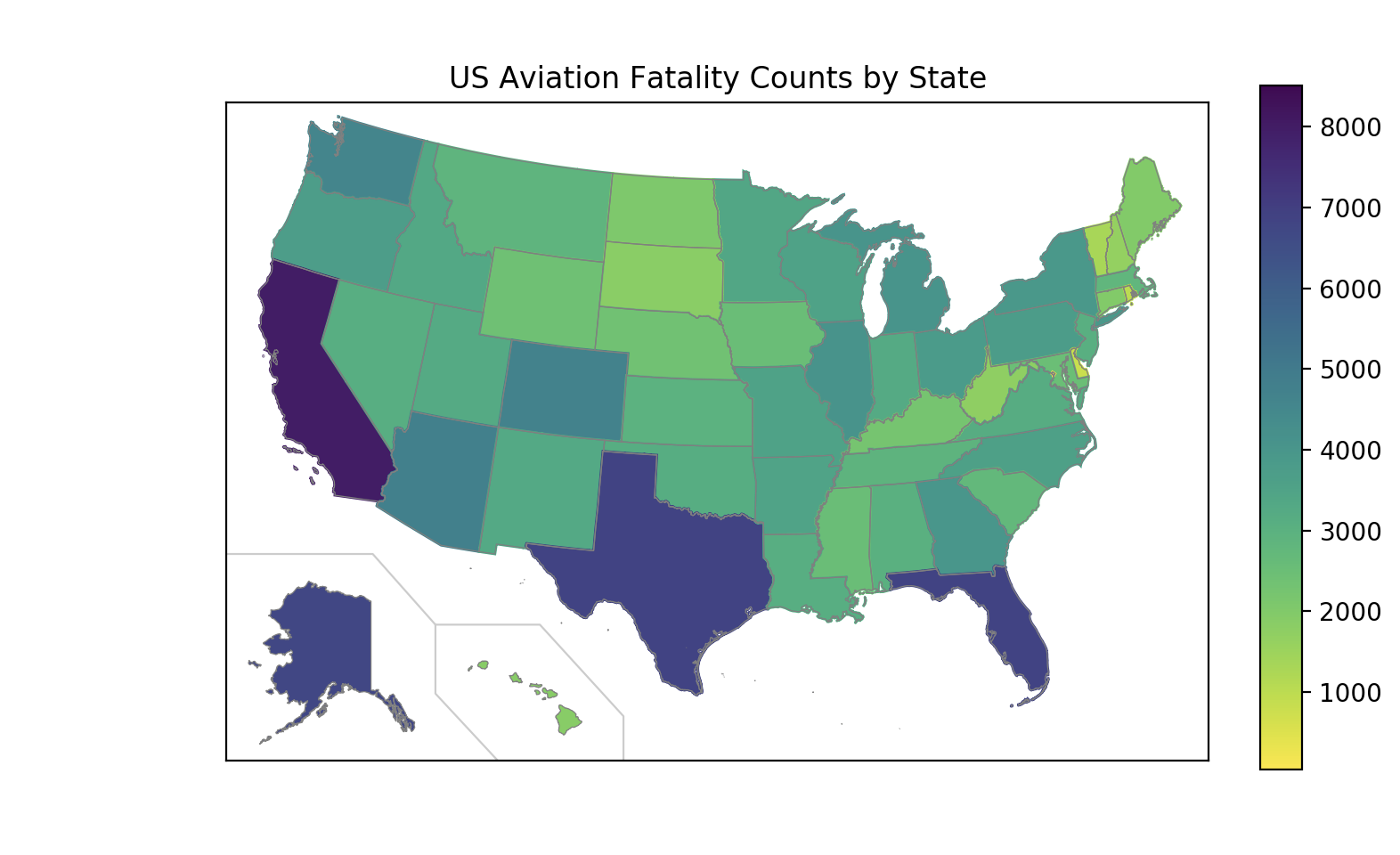
Apparently, there is a big jump in the 1980’s. After reviewing the data, we could see that the earliest accidents reported were all personal flights, with only 5 flights taken into consideration between 1948 – 1981. It would be expected that the accidents reported jump as the commercial aviation industry grew, so the jump in the data is not alarming.

**Location**

One thing we were curious about is if location is a driver. In the following graph you can see that there are spikes in specific states. However, we found that those states also have a denser population and more air traffic, like California, Texas, and Florida, so that is the simplest explanation for what is very apparent. Another factor these states have in common is the weather, it’s generally comfortable to warm, which means more pilots are out flying. Alaska also stands out: air traffic is a very common way to cover the great distances, and it’s hard to know if that factor is more important than ice that might cover the wings.



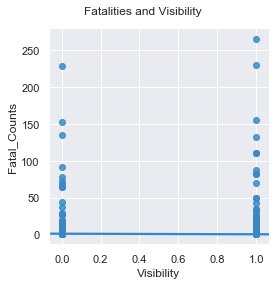
We can also review this in a US map:



This map highlights the states where the fatalities are concentrated: Florida, Texas, California, Alaska. The fact that there are more aircraft flying in these states means that there are more accidents, statistically, so if the percent is stable, this means a higher count of fatalities in the more populated states. If you look at specific cities or airports with maybe one exception we didn’t find any glaring trend.

**Weather**

Weather is a second factor that we suspected to have an impact.



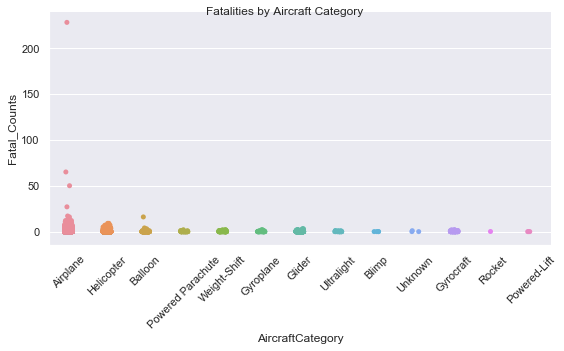
The weather variable uses codes:

1. VMC= Visual meteorological conditions. Visibility is good, this code coincides with good weather.
2. IMC= Instrument meteorological conditions. This means visibility is too low to fly while relying on your eyes, the pilots must rely on instruments, which requires extra training and certification. Pilots who fly under these conditions are more experienced, and the weather is not favorable.
3. UNK= unknown

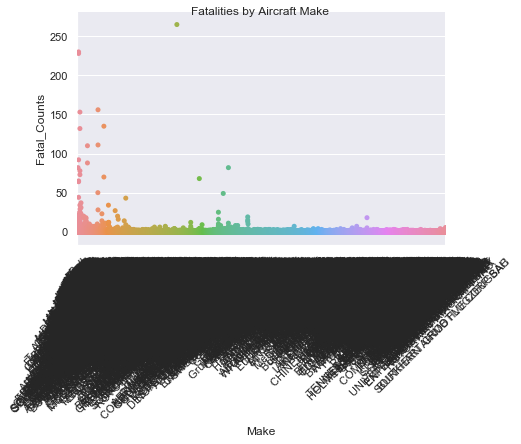
To be able to work with the codes we used one-hot-encoding and created visibility and non-visibility variables. The resulting graph shows more correlation with good visibility conditions. Since more people fly in good conditions, we conclude that in itself is not a major factor for fatalities.

**Aircraft-Related Factors: Aircraft Category**

There are more fatalities with airplanes than other categories; this only makes sense- more airplanes are flown, than aircraft from other categories such as balloons or gyrocraft.

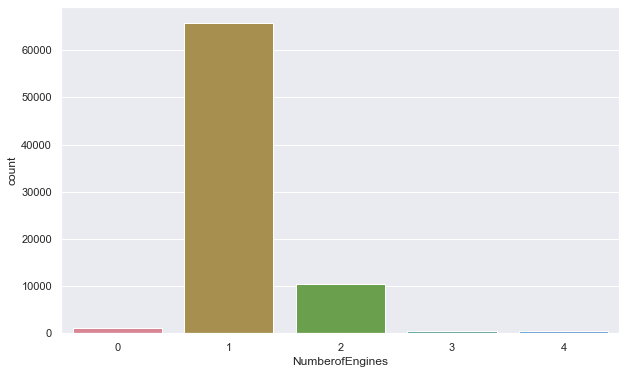


**Aircraft Make**



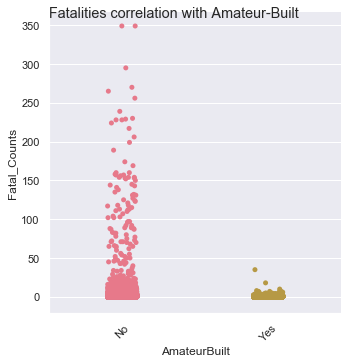
It’s hard to decipher what is seen in the aircraft make graph, but this came into play in later analysis.

**Number of Engines**



This graph shows the prevailing number of engines. As it happens, 1 engine corresponds to personal flights on small private non-commercial aircraft.

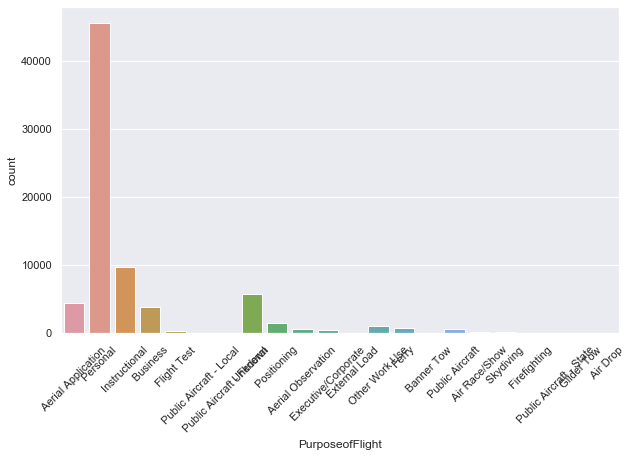
**Amateur-Built**



Amateur built aircraft are not lnked more to fatalities, but recall there are a lot less of them.

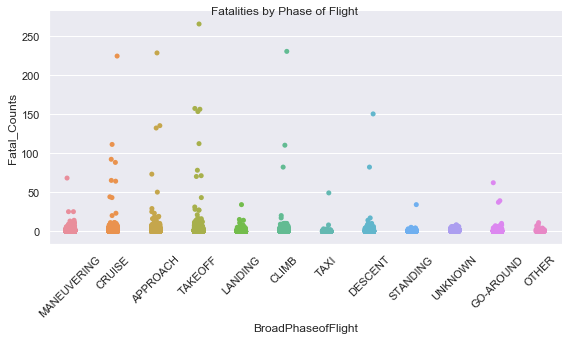
**Flight-Related Factors: Purpose of Flight**

We looked at the Purpose of Flight, and realized the number one in count was personal flights:



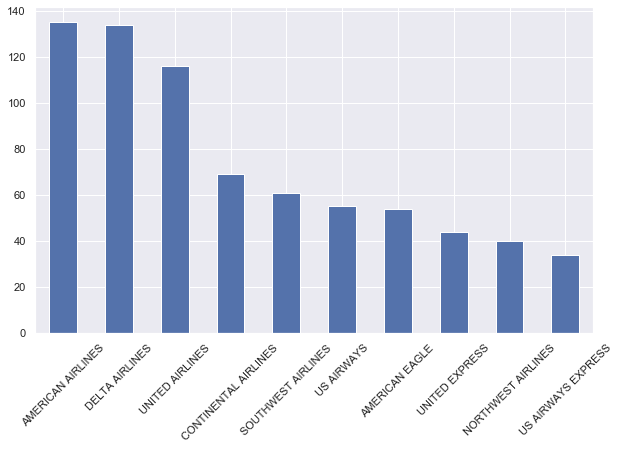
**Phase of Flight**

What about the phase of flight? Looks like there’s more on takeoff or approach, the pre-cursor for landing:



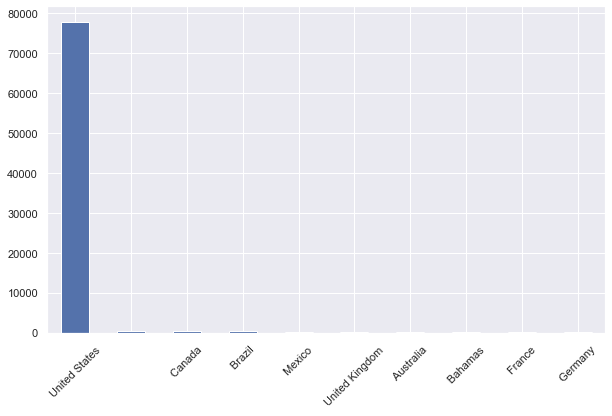
**Airliners**

A portion of aviation accidents occur with commercial airliners involved. This is the distribution of accidents per airliner in our dataset covering the years 1982-2018:



Some are more prone to fatalities than others, but we have to remember they are altogether a smaller portion of fatalities, and the numbers have been going down and changing over the years.

Our dataset covers accidents that have occurred outside the US, investigated by the U.S. NTSB(National Transportation Safety Board). We have decided to drop the foreign events based on the insignificance of the number of events abroad in our data:

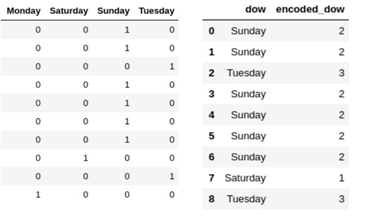


##### **Analysis**

For our analysis we’re specifically looking at US accidents and those that accidents where we know the type of aircraft that was used (airplane, helicopter, balloon, etc.). To try and determine which of our variables (aka columns/features) should be used when differentiating accidents between levels of fatality and aircraft damage, we built two random forest models and pulled the most important features the random forest (RF) model uses to make its conclusions (aka what the model thinks are the most important factors for predicting an outcome). The two models are:

* An RF regression model that can predict how many (aka number of) fatalities might occur if an accident were to happen
* An RF classifier that can predict the damage to an aircraft in an accident were to happen (minor, substantial, or destroyed)

All the data preprocessing training and testing was done with “pandas” and “scikit-learn/sklearn” in Python 3. It is important to state before further analysis explanation that unlike R or Weka, sklearn is incompatible with string elements in our input variables, and thus, must be encoded. The method we used to encode our discrete elements was “One Hot Encoding”. This method takes all the unique elements of a variable and turns it into a series of logic (True/False or 1/0) elements showing the presence of that discrete element in an observation. An example of one hot encoding is shown below using days of the week.



There are several reasons we chose to use RF models for this task, some important ones to note are:

* Random forest models do well with models with data that has high dimensionality (large number of input variables). Our data has 32 columns and with one hot encoding that number can jump well over 20,000 for our particular problem.
* It’s great at handling missing data or if there are large sections of missing data aka null values
* There is infrastructure in place to handle unbalanced classes
* Most importantly, we can pull informative and important features from the model for interpretability.

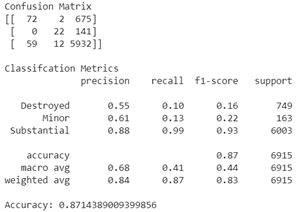
We decided not to use all of the variables in the dataset. Some of the variables like “TotalUninjured” or “InjurySeverity” would be using future data to predict the past, and intuitively, variables like “EventID” or“ReportStatus” seemed inconsequential for predicting accident severity. Also there were simply some variables that had too much missing data (greater than 90% null elements). A full list of the variables removed can be found in the accompanying notebook.

For our RF regression model, we used 25 trees with a total training time of 32.3 minutes. Compared to our classification model, we used fewer trees as the computation for RF regression is more intensive and we blew through our RAM limit in Google Colab with larger number of trees. Performance metrics are below:

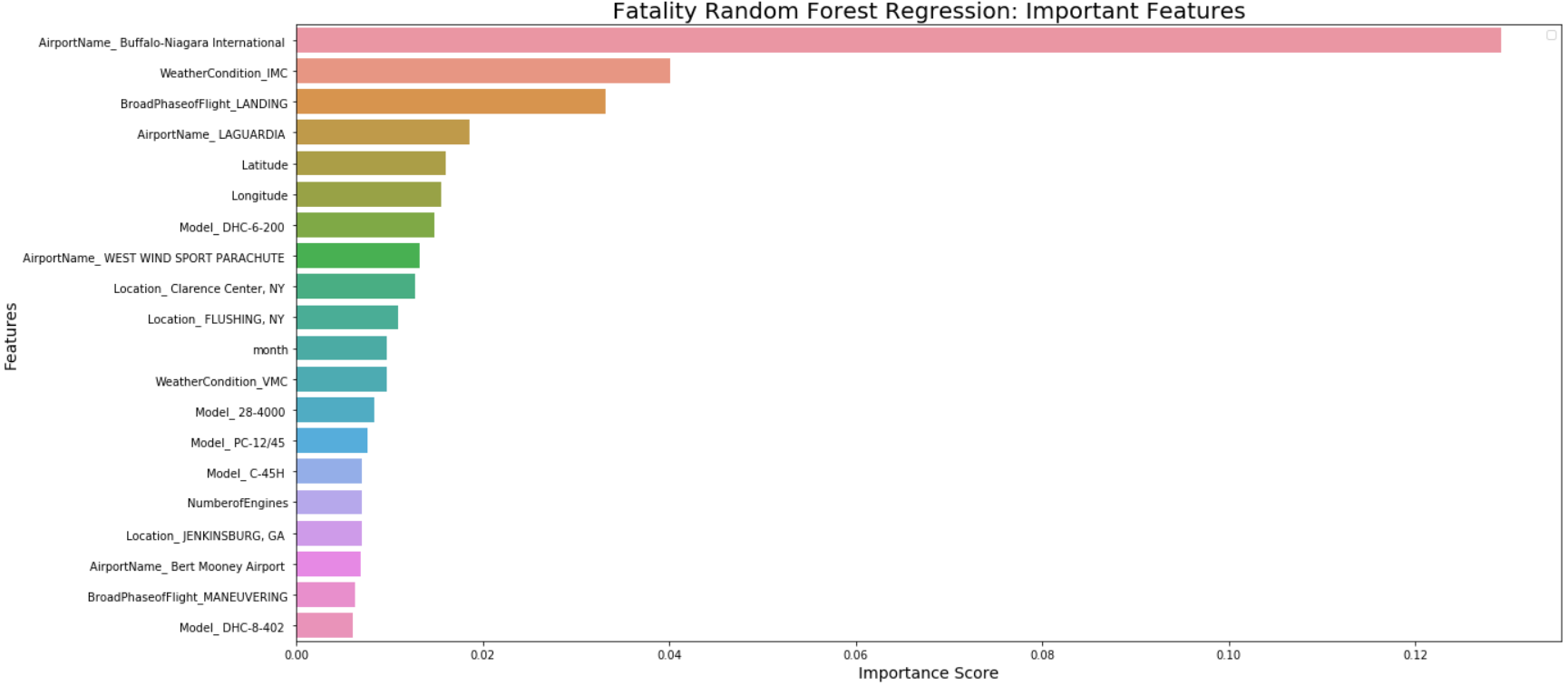
|  |  |
| --- | --- |
| **Metric** | **Value** |
| Range | 228 |
| Mean Absolute Error | 0.364 |
| Root Mean Squared Error (RMSE) | 2.98 |

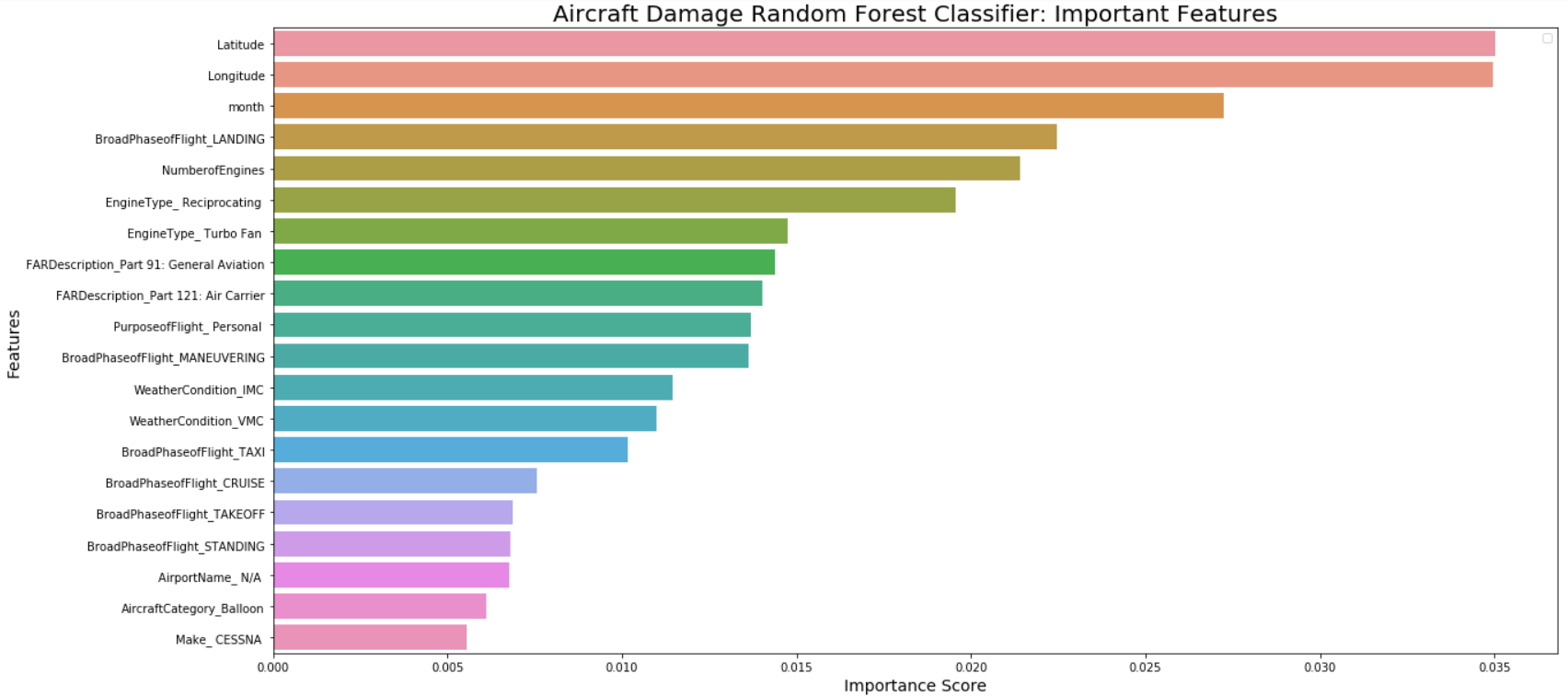
Our Regression model seems perform satisfactory though could use improvement. Unfortunately, with RF models, performance can only be improved so much by increasing the number of estimators.

For our RF classification model, we used 100 trees and told the model to weigh the observations while bootstrapping so that in the bootstrapped subsets, classes with smaller number of observations weigh more than classes with bigger number of observations. This was done to reduce overfitting, which is a common problem with RF models. Total training time was 4.1 minutes, which was faster than training our regression model as training a model to classify amongst three labels is much faster than to assign a number on a continuous scale. Performance metrics are below:



It’s clear that our model is experiencing a large amount of bias which is probably because many of the observations fall under the “AircraftDamage” == “Substantial” class. If we didn’t introduce the weighting mechanism, the model would have probably been more skewed. A future preprocessing step would be to bootstrap the data itself to capture more observations with classes that aren’t as numerous. Nevertheless, the classification model still has some use in determining our important features.

These are the features that our RF models deemed important. Here we show the top 20 for each model.

For features labeled like “BroadPhaseofFlight\_LANDING”, the RF model is saying that whether or not the aircraft was in the landing phase played an important role in determining aircraft damage. This is a result of the “One Hot Encoding”. It’s important to note that just because the RF models deem features as important does not mean that certain values of these features lead to certain conclusions, i.e. just because the Fatality Random Forest Regressor says that Buffalo-Niagra International is important in determining levels of fatality does not mean that that particular airport is a high risk area for fatal aircraft accidents. However, that isn’t to say that intuitive thought and posits can be made from some of the important features.

##### **Recommendation**

Some recommendations based off our analysis

* Our RF models say that the landing phase for aircrafts play a huge role in determining accident severity. An article in the 2002 Aviation Safety Magazine( found here: <http://www.aviationsafetymagazine.com/airplane/landing-accidents-and-runway.html>) said that landing account for more than a third of general aviation accidents. The article attributes this to lack of aircraft control by pilots in adverse conditions. Thus we recommend that pilots in training spend more time practicing landings in a variety of conditions. This conclusion feeds into our next point.
* Weather seems to play an impact in determining accident severity. IMC stands for Instrument Meteorological Conditions meaning that the weather where the pilot must use his/her instruments as the primary references rather than visual cues such as cloudy weather. This doesn’t pose a huge threat to commercial airliners and personal jets that have state of the art instruments, but does pose a threat to those personal airplanes and other aircraft, such as helicopters and balloons, where relying on visual reference is needed more. We recommend improving access to real time weather reports, especially for personal aviation enthusiasts as well as pilots of smaller craft.
* Speaking of personal aircraft, the DH-6 Twin Otter Airplane (a personal sized airplane) had one of the highest crash rates from 2003-2012 (<https://www.prnewswire.com/news-releases/best-and-worst-aircraft-crash-rates-revealed-212040761.html>) The fact that the RF model deems that those flying under the authority of “Part 91: General Aviation” is important and that flying for general aviation possibly private transport or recreation is something to keep noted. This makes sense, again, given the lack of technology and support pilots of these airplanes get compared the big commercial airliners. Again, we recommend that more hours are needed and in a variety of conditions before someone can get a Private Pilot’s License (PPL)

Final comment

It’s hard to separate and isolate the different determinants. When we looked at variables one by one, some of the final findings were not obvious. The algorithm might be able to find patterns in the data that we are not able to find by looking at the variables one by one.

##### **References**

The main source of information for this project is NTSB – National Transportation Safety Board. The NTSB investigates all US aircraft accidents (and some that occur abroad)

**1.** [**https://www.ntsb.gov/investigations/data/pages/data\_stats.aspx**](https://www.ntsb.gov/investigations/data/pages/data_stats.aspx)

This is the main page, home page for data and stats for all safety investigations.

* From here ‘Aviation Accidents’ links to item 3 - the database
* ‘Annual review for aircraft accident data’ links to the following page containing annual accident stats reviews

**2.** [**https://www.ntsb.gov/investigations/data/Pages/aviation\_stats.aspx**](https://www.ntsb.gov/investigations/data/Pages/aviation_stats.aspx)

Between 2007-2011 NTSB produced a PDF file for annual reviews. For example: Aircraft Accident Data Review 2010.PDF - this file has summaries for 2010 and comparisons 2001-2010. After that, the annual reviews are linked as a dynamic web page. The annual review pages can be useful as references, but not as a dataset.

**3.** [**https://ntsb.gov/\_layouts/ntsb.aviation/index.aspx**](https://ntsb.gov/_layouts/ntsb.aviation/index.aspx)

This is the database from where we downloaded our data

4. Additional data and supplemental information was found on Kaggle:

[**https://www.kaggle.com/saurograndi/airplane-crashes-since-1908**](https://www.kaggle.com/saurograndi/airplane-crashes-since-1908)

##### **Appendices**

Data Dictionary

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Short Description** | **Meaning** |
| EventId | Unique Identification for Each Event | Each event is assigned a unique 14-character alphanumeric code in the database. This code, used in conjunction with other primary keys (if applicable), are used to reference all database records. All database queries using a relational database (e.g., MS Access) should link tables using the ev\_id variable. |
| InvestigationType | Type of Event | Refers to a regulatory definition of the event severity. The severity of a general aviation accident or incident is classified as the combination of the highest level of injury sustained by the personnel involved (that is, fatal, serious, minor, or none) and level of damage to the aircraft involved (that is, destroyed, substantial, minor, or none). The |
| AccidentNumber | NTSB Number | Each accident/incident is assigned a unique case number by the NTSB. This number is used as a reference in all documents referring to the event. The first 3 characters are a letter abbreviation of the NTSB office that filed the report. The next 2 numbers represent the fiscal year in which the accident occurred. The next two letters indicate the investigation category (Major, Limited, etc) and mode (Aviation, Marine, etc). The next three digits indicate the chronological sequence in which the case was created within the given fiscal year. And a final letter (A, B, C, etc) may exist if the event involved multiple aircraft |
| EventDate | Event Date | The date of the event. Dates are to be entered in the format: MM/DD/YYYY |
| Location | Event Location Nearest City | The city or place location closest to the site of the event. |
| Country | Event Country | The country in which the event took place. |
| Latitude | Event Location Latitude | Latitude and longitude are entered for the event site in degrees and decimal degrees. If the event occurred on an airport, the published coordinates for that airport can be entered. If the event was not on an airport, position coordinates may be obtained usingGlobal Positioning System equipment or nearest known reading. |
| Longitude | Event Location Longitude |  |
| AirportCode | Event Location Nearest Airport ID | Airport code if the event took place within 3 miles of an airport, or the involved aircraft was taking off from, or on approach to an airport. |
| AirportName | Event Location Airport | Airport name if the event took place within 3 miles of an airport, or the involved aircraft was taking off from, or on approach to, an airport. |
| InjurySeverity | Event Highest Injury | Indicate the highest level of injury among all injuries sustained as a result of the event. |
| AircraftDamage | Damage | Indicate the severity of damage to the accident aircraft. For the purposes of this variable, aircraft damage categories are defined in 49 CFR 830.2. |
| AircraftCategory | Aircraft Category | The category of the involved aircraft. In this case, the definition of aircraft category is the same as that used with respect to the certification, ratings, privileges, and limitations of airmen. Also note that there is some overlap of category and class in the available choices. |
| RegistrationNumber | Aircraft Registration Number | The full registration (tail) number of the involved aircraft, including the International Civil Aviation Organization (ICAO) country prefix. Note: the prefix for US registered aircraft is "N." |
| Make | Aircraft Manufacturer's Full Name | Name of the manufacturer of the involved aircraft. |
| Model | Aircraft Model | The full alphanumeric aircraft model code, including any applicable series or derivative identifiers. For example, a 200 series Boeing 737 is entered as 737-200. |
| AmateurBuilt | Aircraft is a homebuilt (Y/N). |  |
| NumberOfEngines | Number of Engines | The total number of engines on the accident aircraft. |
| EngineType | Engine Type | Type of engine(s) on the involved aircraft. |
| FARDescription | Federal Aviation Reg. Part | The applicable regulation part (14 CFR) or authority the aircraft was operating under at the time of the accident. |
| Schedule | Indicates whether an air carrier operation  was scheduled or not | If the accident aircraft was conducting air carrier operations under 14 CFR 121, 125, 129, or 135, indicate whether it was operating as a "scheduled or commuter" air carrier or as a "non-scheduled or air taxi" carrier. |
| PurposeOfFlight | Type of Flying (Per\_Bus / Primary) | If the accident aircraft was operating under 14 CFR part 91,103,133, or 137, this was the primary purpose of flight. |
| AirCarrier | Operator Name& Operator Is Doing Business As | The full name of the operator of the accident aircraft. This typically refers to an organization or group (e.g., airline or corporation) rather than the pilot; contaminated with the carrier, business, or code share name if the accident aircraft was operated by a business, air carrier, or as part of a code share agreement. |
| TotalFatalInjuries | Injury Total Fatal | The total number of fatal injuries from an event. |
| TotalSeriousInjuries | Injury Total Serious | The total number of serious injuries from an event. |
| TotalMinorInjuries | Injury Total Minor | The total number of minor injuries from an event. |
| TotalUninjured | Non-Injury Total | The total number of non-injuries from an event. |
| WeatherCondition | Basic weather conditions | The basic weather conditions at the time of the event. |
| BroadPhaseOfFlight | Phase of Flight | All occurrences include information about the phase of flight in which the occurrence took place. Phase of flight refers to the point in the aircraft operation profile in which the event occurred. |
| ReportStatus | Latest Report Level | The furthest level to which a report has been completed |
| PublicationDate | Publication data of the Latest Report Level | The date on which the previous column was published to the web. |