IST707 DATA SCIENCE GROUP PROJECT

WINTER 2020

AVIATION ACCIDENTS AND FATALITIES ANALYSIS

Asiana Airlines flight 214 Boeing 777 after it crash landed on July 6, 2013, on final approach into San Francisco International Airport.
Image source: NTSB

Professor: Dr. Ami Gates

Group 4:

- Editt Gonen-Friedman egonenfr@syr.edu
- Maria Ng mng103@syr.edu
- Bhavya Madhavan bkmadhav@syr.edu
- Veasna Oum voum01@syr.edu

1. Introduction

On December 17, 1903, Wilbur and Orville Wright made their first controlled aircraft with the Wright Flyer at Kitty Hawk, North Carolina. The Wright brothers had made four brief flight attempts, which signified the beginning age of aviation. Brave men and women took to the skies in their flying machines while the government stayed out of the way allowing brilliant inventors to take risks. The aviation industry was thriving for innovation over the concern of safety. Until the first air accident took place on December 14 ,1920, where 2 passengers and 2 crew members were killed, and it was reported as the first recorded airliner crash in history.

During the 1920s, the first laws were passed in the USA to regulate civil aviation. The Air Commerce Act of 1926 required pilots and aircraft to be examined and licensed for accidents to be properly investigated, and for the establishment of safety rules and navigation aids, under the Aeronautics Branch of the United States Department of Commerce. Then other similar laws were passed throughout Europe and the United Kingdom. Aviation safety was no longer to be taken as a trivial matter and this was a concern that needed to be addressed.

Despite highly regulated guidelines on safety established now, there have been notable incidents being reported. In March 2019, a high-profile Boeing 737 Max crash in Ethiopia killed 157 people. On May 3rd, 2019, flight GL293 overran the end of the runway and came to a stop in the shallow waters of an adjacent river about 1,250 feet beyond the end of the runway. There were seven crew members and 136 passengers on board, and at least 21 of the occupants were injured. On December 26th, 2019, Eurocopter AS350 departed from Lihue, HI for a sightseeing flight over the island of Kauai. The aircraft crashed into a cliff in the northwest section of the island about a mile inland from the coast and killed everyone on board. On January 26th, 2020, Sikorsky S-76B crashed 40 minutes after it departed from John Wayne Airport in Orange County at 9:06 a.m. Amid the thick fog, the chopper hit the foothills of the Santa Monica mountains and caught fire, killing everyone on board, including the NBA star Kobe Bryant, and his daughter.

Interestingly, last year was "one of the safest years ever for commercial aviation", according to accident tracking website the Aviation Safety Network. There were 86 accidents involving large commercial planes, including eight fatal incidents, resulting in 257 fatalities. However, the study did not include small commuter planes, and some smaller turboprop aircraft. In 2018, NTSB officials reported that civil aviation fatalities rose from 347 in 2017 to 393 in 2018. The increase means that, on average, there was at least one aviation death per day in 2018. Alarmingly, there seemed to be more fatalities reported from civil aviation than the commercial one.

Were those accidents preventable? What were the causes of those recent aviation tragedies? Was it pilot training, mechanic failure, or weather? Aviation accidents seemed to become the norm in the news like a car accident reporting for the morning commuters. Was it the death of a basketball celebrity that triggered the awareness that this raising increase in aviation tragedies should not be not normal?

2. Analysis and Models

2.1 The Data

2.1.1 The Dataset

The American NTSB (National Transportation Safety Board) conducts investigations into every aviation accident and incident that happens in the U.S., and sometimes abroad. At the end of an investigation, NTSB my issue safety recommendations to a variety of stakeholders, such as the FAA (Federal Aviation Administration) regarding safer procedures for pilots, air traffic controllers, or aircraft manufacturers – regarding equipment on aircraft. All of this information is publicly available on https://ntsb.gov website. Event data is available for viewing and download as a .txt file, and that is our data set.

The data is composed of 84,301 accident report entries that have been gathered since 1948, described by 31 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 information. Refer to a data dictionary in section 6.2

2.1.2 The Variables

While each of the accidents becomes an observation, a lot of the information about them is categorical. For example, damage has categories such as 'destroyed' or 'substantial'; Location information includes country / state / city and coordinates. Some of the variables are numeric, such as number of engines, or number of fatalities. One thing is certain, the data is far from clean.

```
'data.frame':   84301 obs. of  32 variables:
 $ Event.Id              : Factor w/ 83074 levels "20001204X00000 ",..: 45209 45212 60530 45213 56804 79144 45211 49013 49012 49011 .
 $ Investigation.Type    : Factor w/ 3 levels "  "," Accident ",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ Accident.Number       : Factor w/ 84301 levels " ANC00FA018 ",..: 77639 58095 69362 58806 20240 69790 20241 76384 69793 61461 ...
 $ Event.Date            : Factor w/ 13731 levels "1/1/00","1/1/01",..: 1774 10638 12312 9498 11855 12918 11437 21 21 21 ...
 $ Location              : Factor w/ 26606 levels "  "," , Gabon ",..: 15745 2785 20954 7424 3562 2555 5280 19575 6697 11332 ...
 $ Country               : Factor w/ 180 levels "  "," Afghanistan ",..: 171 171 171 171 171 171 171 171 171 171 ...
 $ Latitude              : num  NA NA 36.9 NA NA ...
 $ Longitude             : num  NA NA -81.9 NA NA ...
 $ Airport.Code          : Factor w/ 10012 levels "","  "," - ",..: 2 2 2 2 2 2 2 2 7187 5152 ...
 $ Airport.Name          : Factor w/ 23911 levels "  "," --- "," (GRASS STRIP) ",..: 1 1 1 1 1 14833 1 1941 9030 10414 ...
 $ Injury.Severity       : Factor w/ 128 levels " Fatal(1) "," Fatal(10) ",..: 51 81 72 51 1 127 81 127 127 127 ...
 $ Aircraft.Damage       : Factor w/ 4 levels "  "," Destroyed ",..: 2 2 2 2 2 4 2 4 4 4 ...
 $ Aircraft.Category     : Factor w/ 14 levels "  "," Airplane ",..: 1 1 1 1 1 2 1 2 2 1 ...
 $ Registration.Number   : Factor w/ 72577 levels "  "," 00SLB ",..: 72434 34739 35445 2395 6188 139 33849 14160 57439 25177 ...
 $ Make                  : Factor w/ 8028 levels "  "," 107.5 Flying Corporation ",..: 6965 5592 1359 6113 1359 4758 1360 1360 1360 5
 $ Model                 : Factor w/ 11952 levels "  "," -269C ",..: 55 8283 262 11734 11879 4419 11768 11749 876 7580 ...
 $ Amateur.Built         : Factor w/ 3 levels "  "," No "," Yes ": 2 2 2 2 2 2 2 2 2 2 ...
 $ Number.of.Engines     : int  1 1 1 1 NA 2 1 1 2 1 ...
 $ Engine.Type           : Factor w/ 15 levels "  "," Electric ",..: 9 9 9 9 1 11 9 9 9 9 ...
 $ FAR.Description       : Factor w/ 19 levels "  "," Armed Forces ",..: 1 1 1 1 1 1 9 1 15 15 1 ...
 $ Schedule              : Factor w/ 4 levels "  "," NSCH ",..: 1 1 1 1 1 1 3 1 1 1 1 ...
 $ Purpose.of.Flight     : Factor w/ 23 levels "  "," Aerial Application ",..: 16 16 16 16 16 1 16 16 7 16 ...
 $ Air.Carrier           : Factor w/ 3035 levels "  "," (DBA: [EMS]) ",..: 1 1 1 1 1 359 1 1 1 1 ...
 $ Total.Fatal.Injuries  : int  2 4 3 2 1 NA 4 0 0 0 ...
 $ Total.Serious.Injuries: int  0 0 NA 0 2 NA 0 0 0 0 ...
 $ Total.Minor.Injuries  : int  0 0 NA 0 NA 1 0 0 0 3 ...
 $ Total.Uninjured       : int  0 0 NA 0 NA 44 0 2 2 0 ...
 $ Weather.Condition     : Factor w/ 4 levels "  "," IMC "," UNK ",..: 3 3 2 2 4 4 2 4 2 2 ...
 $ Broad.Phase.of.Flight : Factor w/ 13 levels "  "," APPROACH ",..: 4 13 4 4 2 3 13 11 7 4 ...
 $ Report.Status         : Factor w/ 4 levels " Factual "," Foreign ",..: 4 4 4 4 4 4 4 4 4 4 ...
 $ Publication.Date      : Factor w/ 3959 levels "  "," 01/01/1982 ",..: 1 2848 567 2764 1108 2856 3395 2 2 2 ...
 $ X                     : logi  NA NA NA NA NA NA ...
```

Figure 1: The dataset

2.1.3 Cleaning and Prep

(for explanation of these attributes check the data dictionary in section 6.2)

1. General inspection & variable elimination:

Eliminated the following variables that don't add value to the analysis: "Event.Id","Investigation.Type","Accident.Number","Registration.Number","Report.Status","Publication.Date", "X". Also eliminated 4.8K rows with foreign reports.

2. Structure:
   - Injury.Severity has 128 levels because it has parenthesis with # of fatalities in the values. removing the extra info so that we only have the categorical info left, with 3 levels. We later made this variable binary.
   - Cleaned empty spaces in the levels of many variables
   - Consolidated some categories of Aircraft.Category, Engine.Type, FAR.Description, Scheduled, Purpose.of.Flight
   - Event date is factor which needs to be converted to a date
3. Missing values:
   - for the following variables, converted missing values to zero: Total.Fatal.Injuries, Total.Serious.Injuries, Total.Minor.Injuries, Total.Uninjured.

The last one presented a problem: if uninjured is NA, we don't know if no one is injured, or if everyone is injured. This is important for normalization. It is also important because we have almost 12K NAs in this column.

Solution: if uninjured is NA convert it to zero. We're adding a 'total people' column for normalization. During normalization, when total people=0, each of the columns will get 0, but uninjured will get 1, because that means that 100% of people on board are uninjured.

   - Number.of.Engines -convert NAs to 1 if aircraft has more than 1 it would be reported. most privates have 1.
   - Purpose.of.Flight populated some NAs based on FAR.Description
4. Invalid values or values in bad format:
   - Weather.Condition – cleaned the factors, eliminated 4K rows with unknown weather
   - Amateur.Built – Anything "" converted to "No", that's the default and they do have make and model, so they are not built by amateurs.
5. Feature Generation:
   - Created total people variable for normalization of injury and fatality counts
   - Created year, month, and weekday attributes from event date to find more insights
   - Extracted state and city from Location
6. Normalization:

Absolute numbers can be misleading. Normalized fatalities and injuries as fraction of total people.

7. Transformation: Discretization

The variables of the numbers of injuries and fatalities in the data structure were discretized, for use in the models that prefer categorical data (Association Rule Mining and Decision Tree).

2.1.4 Visualization

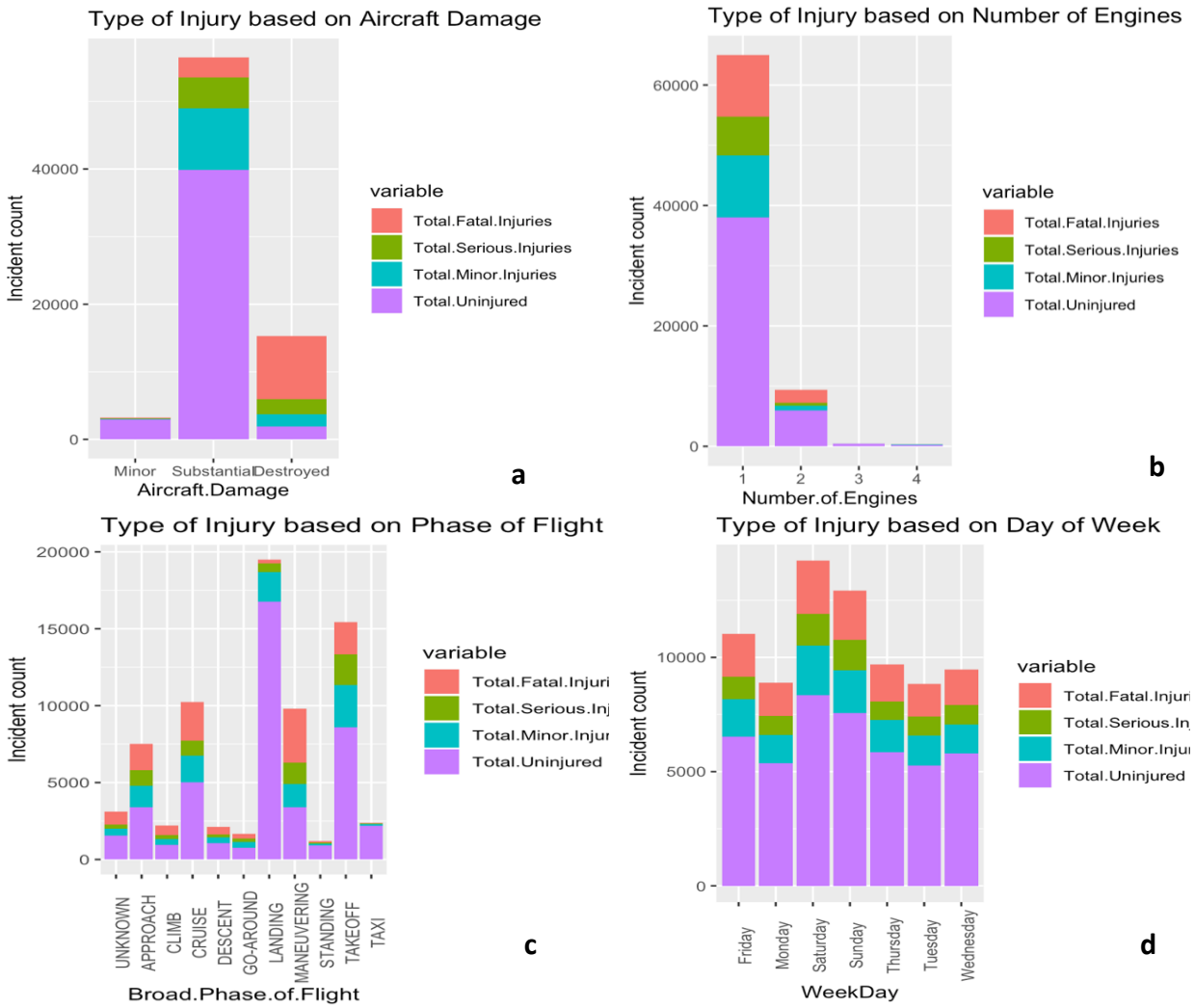


Figure 2. Incident count vs. Aircraft.Damage (a), Number.of.Engines (b), Broad.Phase.of.Flight (c), & WeekDay (d)

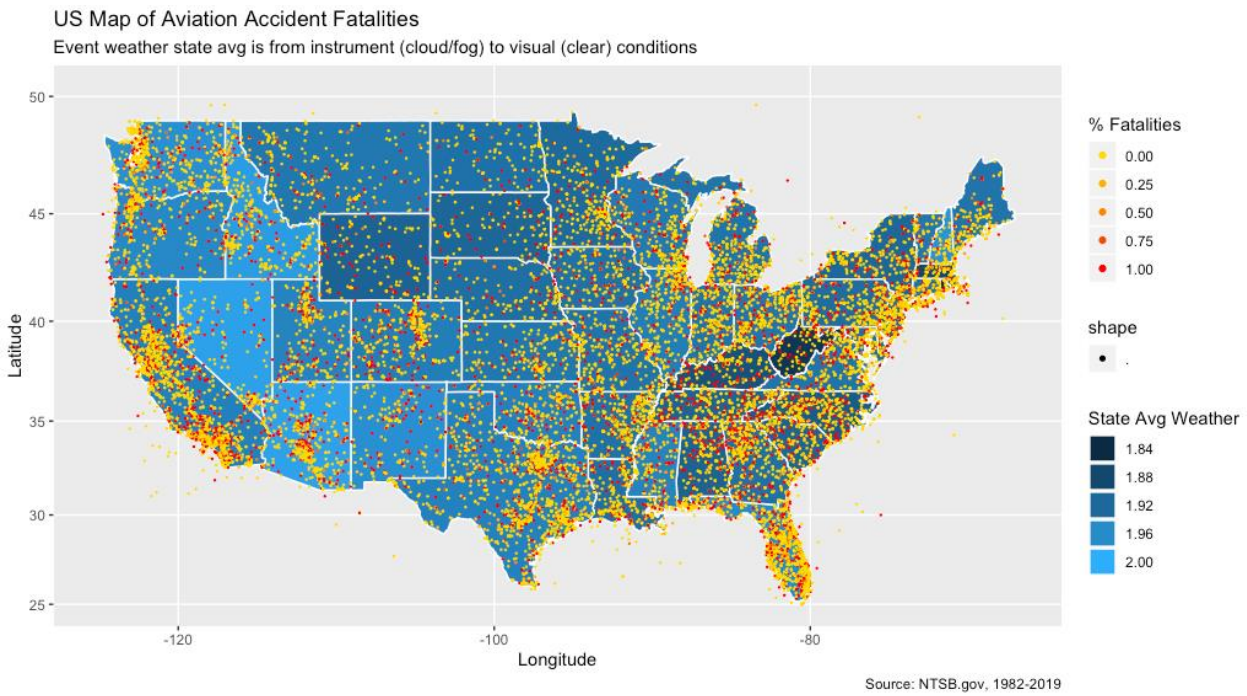


Figure 3. US Map of Aviation Accident Fatalities based on % of Fatalities (dots and shape size) and State Avg Weather (various shaded of blue)

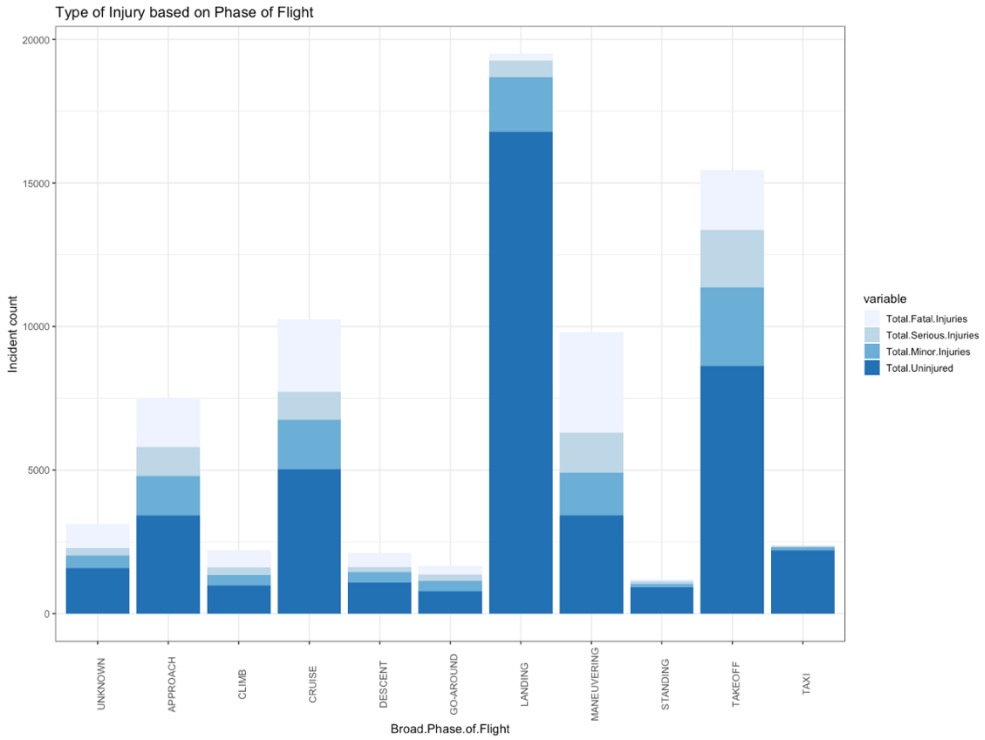


Figure 4: Injuries and Fatalities based on Phase of Flight

2.2 Model 1: Clustering

Clustering algorithms find groups in data. The goal for clustering is to find groupings that highlight important factors in the data.
Two different methods were used for this purpose:

1. Hierarchical Clustering
2. K-Means Clustering

2.2.1 Model Details: Hierarchical Agglomerative Clustering – Complete Linkage

Clustering algorithms prefer numeric data. The main reason for this is that distances need to be calculated between the data points, so that data close to each other are grouped together, and data far apart are not. However, for this dataset we need to work with mostly categorical data. There are ways to figure out distances by comparing categories, so that is still doable. The daisy distance function and the gower algorithm are tailored to work with nominal attributes.
Hierarchical clustering produces a set of nested clusters organized as a hierarchical tree that can be visualized as a dendrogram: a treelike diagram that records the sequences of merges or splits. Specifically used Agglomerative Clustering – that start with individual points, and group them together one data point at a time. Complete linkages is used in the calculation, which means a point is added to a cluster based on its distance from all members, not just the nearest one or the farthest one. This usually produces the most balanced results.
The next step after creating the clusters is to evaluate how many clusters there are. This is not a parameter for the same algorithm but a separate calculation, that can be done via a grid-search for the best k.

2.2.2 Parameter Values

Table 1: Parameter Values

Parameter	Default	Values used for this Analysis
1. method	complete linkage	complete linkage
2. K	none	2-15, ended with 3 or 7

2.2.3 Model Evaluation

This model is evaluated visually after running a grid search for the best k, and plotting the dendrogram, to see if groups are visible and see their make-up and important variables.

2.2.4 Model Details: K-Means (K-Medoid) Clustering

K-Means is a partitional approach to clustering, where centroids are selected, and the data points are organized based on how close they are to these centers. The centers are chosen at

random, and they are adjusted and readjusted to reflect the center of the group. When working with nominal data we can't calculate the mean, so medians are used instead. The pam algorithm is used instead of k-means as a k-medoid type, for nominal attributes.

2.2.5 Parameter Values

Table 2: Parameter Values

Parameter	Default	Values used for this Analysis
1. distance matrix method	none	gower – best suited for nominal attributes
2. K	none	2-6, ended with 4

2.2.6 Model Evaluation

This model is evaluated after running a grid search for the best k, by reviewing principal components to see the groups and their make-up and important variables.

2.3 Model 2: Association Rule Mining

In this unsupervised learning machine learning algorithm, the predicting attribute is the fatal injury severity. As stated above, the dataset consists for 32 attribute and most of the categorical attributes have thousands of levels in them. To achieve some meaningful rules, different iterations need to be employed.

In the first iteration, all categorical attributes with less than 10 levels are used namely Injury Severity, Aircraft Category, Amateur Built, Schedule, Purpose of Flight, Weather Condition and Phase of Flight. The support and confidence had to be tuned to arrive at the following rules. This algorithm generated 15 rules with the lift ranging from 3.8 to 4.1.

As seen from the generated rules generated below, it can be noted that the most common factor among most of the rules is the weather condition = instruments; meaning that fatal injuries were common when the weather was bad. This is a known fact and ARM didn't help much in generating interesting rules.

As it can be seen below, there is a strong correlation between the LHS (Fatal injury) and the RHS because of greater lift.

Table 3: *Association Rule First Iteration (LHS: Injury Severity = Fatal)*

RHS	Support	Confidence	Lift	Observations
Weather Condition = Instruments Phase of Flight = MANEUVERING	0.0051	0.725	3.986	388
Purpose of Flight=Private Weather Condition =Instruments Phase of Flight =GO-AROUND	0.0012	0.768	4.221	96
Schedule=Non-Scheduled Weather Condition=Instruments Phase of Flight=GO-AROUND	0.0018	0.715	3.930	138
Aircraft Category=Airplane Weather Condition=Instruments Phase of Flight=GO-AROUND	0.0018	0.700	3.850	138
Purpose of Flight=Private Weather Condition=Instruments Phase of Flight=CLIMB	0.0020	0.741	4.075	152
Schedule=Non-Scheduled Weather Condition=Instruments Phase of Flight=CLIMB	0.0027	0.735	4.044	206
Schedule=Non-Scheduled Weather Condition=Instruments Phase of Flight=UNKNOWN	0.0019	0.712	3.918	149
Amateur Built=Yes Purpose of Flight=Private Weather Condition=Instruments	0.0011	0.7094	3.899	83
Amateur Built=Yes Schedule=Non-Scheduled Weather Condition=Instruments	0.0011	0.7040	3.869	88
Aircraft Category=Airplane Amateur Built=Yes Weather Condition=Instruments	0.0011	0.7096	3.900	88
Purpose of Flight=Private Weather Condition=Instruments Phase of Flight=MANEUVERING	0.00362	0.7749	4.259	272
Amateur Built=No Weather Condition=Instruments Phase of Flight=MANEUVERING	0.0049	0.7225434	3.9716	375
Schedule= Non-Scheduled Weather Condition=Instruments Phase of Flight=MANEUVERING	0.0049	0.7530120	4.1391	375
Aircraft Category=Airplane Weather Condition=Instruments Phase of Flight=MANEUVERING	0.0049	0.7231969	3.9752	371
Purpose of Flight=Private Weather Condition=Instruments Phase of Flight=CRUISE	0.0108	0.7035	3.8672	814

For the second iteration, a few categorical with larger levels in the factors were added to the ARM algorithm. Therefore, the following attributes were present in the ARM algorithm: Injury Severity, Aircraft Category, Amateur Built, Schedule, Purpose of Flight, Weather Condition, Phase of Flight, Make, Model and Air carrier.

As seen from the rule set below, as the levels of the attributes increases, the number of observations satisfying each rule reduces. Once again, it can be seen that ARM is not a better model for this particular dataset.

Table 4: *Association Rule Second Iteration (LHS: Injury Severity = Fatal)*

RHS	Support	Confidence	Lift	Observations
Model=dhc-2 Phase of Flight=CRUISE	0.0016	0.857	5.840	6
Make=de havilland Model=dhc-2 Phase of Flight=CRUISE	0.0016	0.857	5.840	6
Model=dhc-2 Engine Type=Reciprocating Phase of Flight=CRUISE	0.0016	0.857	5.840	6
Model=dhc-2 Purpose of Flight=Unknown Phase of Flight=CRUISE	0.0010	1.000	6.814	4
Aircraft Category=Airplane Model=dhc-2 Phase of Flight=CRUISE	0.0016	0.857	5.840	6
Model=dhc-2 Amateur Built=No, Phase of Flight=CRUISE	0.0016	0.857	5.840	6
Make=de havilland Purpose of Flight=Private Phase of Flight=MANEUVERING	0.00107	1.000	6.814	4
Make=de havilland Schedule=Non-Scheduled Phase of Flight=MANEUVERING	0.00188	0.875	5.962	7
Make=de havilland Weather Condition=Visual Phase of Flight=MANEUVERING	0.00188	0.875	5.962	7
Make=bell Purpose of Flight=Private Weather Condition=Instruments	0.00134	1.000	6.814	5
Make=beech Schedule=Non-Scheduled Phase of Flight=MANEUVERING	0.00107	1.000	6.814	4
Engine Type=Turboprop	0.00134	1.000	6.814	5

Purpose of Flight=Private Phase of Flight=MANEUVERING				
Engine Type=Turboprop Schedule=Non-Scheduled Phase of Flight=MANEUVERING	0.00161	0.857	5.840	6
Make=cessna Weather Condition=Instruments Phase of Flight=CLIMB	0.00161	0.857	5.840	6

2.4 Model 3: Decision Tree

2.4.1 Model Details: methods used to analyze the data

Master.df was modified to 16 variables for decision tree analysis. Initial assessment of each variable to determine for model accuracy. The data was split into 2/3 for training data and 1/3 for testing data. Then, proceed to identify a model with a high accuracy for prediction of injury severity. Model accuracy with greater than 95% was selected for further tuning. The tuning process was consisted of cross reference with associate rules or combined a few factors with model accuracy > 95%.

2.4.2 Parameter Values

Table 5 showed the 16 variables with various levels for performing decision tree machine learning to determine model accuracy and confusion matrix.

Table 5: *Decision Tree Variable Table*

	Variable	Factor Levels
1	Injury.Severity	Fatal, No-Injury & Non-Fatal
2	Aircraft.Damage	Destroyed, Minor & Substantial
3	Aircraft.Category	Airplane Balloon, Gliders, Helicopter & LightSport
4	Amateur.Built	No & Yes
5	Number.of.Engines	1, 2, 3 & 4
6	Engine.Type	Reciprocating, Turbofan, Turbojet, Turboprop & Turboshaft
7	FAR.Description	Part 137: Agricultural, Part 91: General Aviation, Part 91F: SpecialOps, Part 135: Air Taxi & Commuter, Part 133: Rotorcraft Ext. Load, Part 121: Air Carrier, Part 103: Ultralight & Unknown
8	Schedule	Schedule & NonSchedule

9	Purpose.of.Flight	Private, Instructional, Agriculture, Skydiving, Business, Public_Aircraft, Aerial_observation, FlightTest, OtherWork, Ferry, ExternalLoad, AirShow, BannerTow, Unknown, CommercialAirline , Firefighting & GliderTow
10	Total.Fatal.Injuries	-1,0,1
11	Total.Serious.Injuries	-1,0,1
12	Total.Minor.Injuries	-1,0,1
13	Total.Uninjuries	-1,0,1
14	Weather.Condition	Instruments & Visual
15	Broad.Phase.of.Flight	APPROACH, CLIMB, CRUISE, DESCENT, GO-AROUND, LANDING, MANEUVERING STANDING TAKEOFF TAXI UNKNOWN
16	Month	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 & 12

2.4.3 Model Evaluation

Model accuracy was generated from each variable. The model accuracy ranged from 12.2% to 99.9%. Month had the lowest model accuracy and Total.Fatal.Injuries had the highest one.

Table 6: *Decision Tree Model Accuracy*

Variables	Model Accuracy
Injury.Severity	99.4%
Aircraft.Damage	86.3%
Aircraft.Category	97.2%
Amateur.Built	89.8%
Number.of.Engines	89.9%
Engine.Type	89.8%
FAR.Description	97%
Schedule	96.9%
Purpose.of.Flight	67.9%
Total.Fatal.Injuries	99.9%
Total.Serious.Injuries	93.8%
Total.Minor.Injuries	92.2%
Total.Uninjured	94.3%
Weather.Condition	92.8%
Broad.Phase.of.Flight	36.9%
Month	12.2%

2.5 Model 4: SUPPORT VECTOR MACHINE (SVM)

2.5.1 Model Details: methods used to analyze the data

Methods:

Machine learning is a science of collecting data and feeding it to an algorithm that can learn via pattern recognition and then outputs a probability matrix. The key to effective result is to collect large amount of cleaned data and then tune the algorithm parameters just like a radio station where the frequencies are clear enough.

Support vector machines so called as SVM is a *supervised learning algorithm* which can be used for classification and regression. SVM makes use of a hyperplane which acts like a decision boundary between the various classes

Analysis & Model

Data Munging

1. Perform Correlation and use *Principal Component Analysis* from Clustering Model to determine the best attributes for SVM
    a. Attributes:

Table 7: *SVM Attributes*

Injury Severity	Engine Type
Make	Purpose of Flight
Amateur Built	Weather Condition
Number of Engines	+ optional attribute

2. Combing any data that has a fatality of one or more into one feature: “Fatal”
3. Transform categorical data in numeric
4. Creating a “factor” for the Target Attribute: “Fatal” or “Nonfatal”
5. Create 2/3 Train data and 1/3 Test data
6. Train Control for SVM (Parameters Tuning - arbitrary)
7. Execute SVM
8. Fine tune parameters and rerun SVM with different kernels for better result

Preparing SVM Model by initializing the Train Control

RStudio - Setting parameters for train control in RStudio

CTRL ← trainControl (method = "repeatedcv",

repeats = 5,
summaryFunction = twoClassSummary,
classProbs = TRUE,
sampling = 'down')

Breakdown of Train Control in SVM

****

The parameter name is called "TrainControl"

Method = "repeatedcv" → The analysis will be done using repeat cross-validation
Repeats = 5 → This parameter can be any quantity
SummaryFunction = twoClassSummary → Points to the ROC analysis
ClassProbs = TRUE → **Default**
Sampling = 'down' → If there is an imbalance of the binary Target Attribute
(when one value is more than the other)

****

SVM Models

- SVM Linear Model
- **SVM Radial Model** (This kernel was selected based on the classification plot)
- SVM Polynomial Model

2.5.2 Parameter Values

Table 8: Parameter Setting Variations for Fine Tuning

KERNEL									
	Linear	*Cost*	0.25	0.5	0.75	1	1.25	1.5	**62%**
	Radial	*Cost*	0.25	0.5	0.75	1	1.25	1.5	**68%**
		Sigma	0.01		0.015		0.2		
	Polynomial	*Cost*	0.25	0.5	0.75	1	1.25	1.5	**67%**
		Degree	1		2		3		

3. Results

3.1 Results of Model Experiments: Clustering

3.1.1 Results of Model Experiments: Hierarchical Agglomerative Clustering

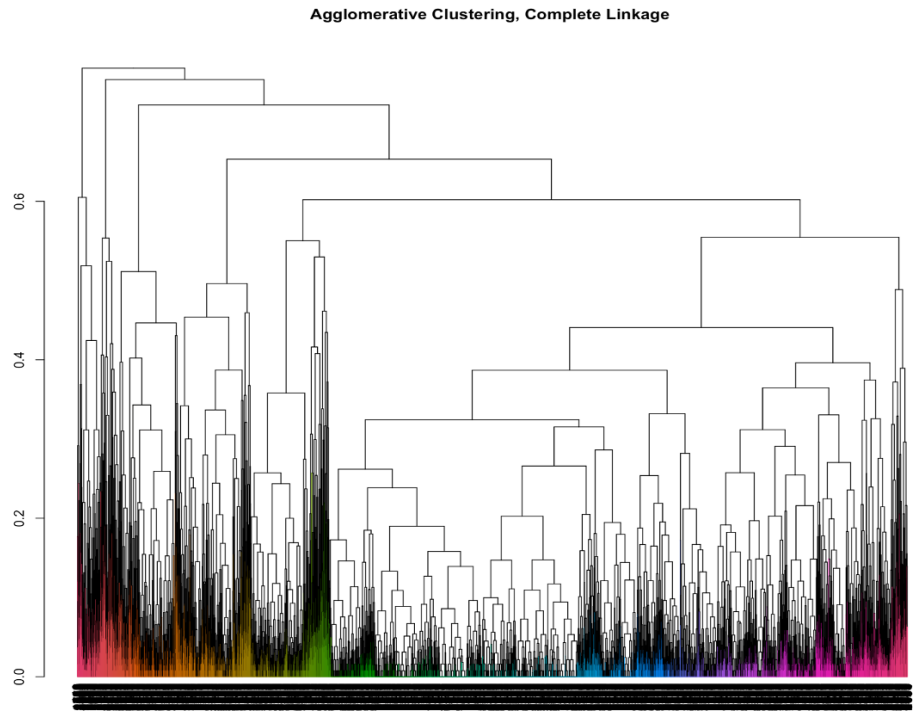


Figure 5: Hierarchical clustering

First, we can look at the hierarchies before we determine the number of clusters:

It's not easy to determine, and we need to use a grid search to help find out.

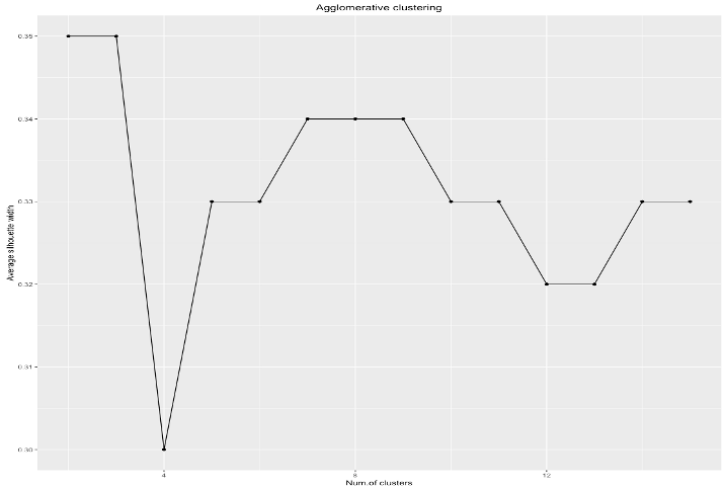
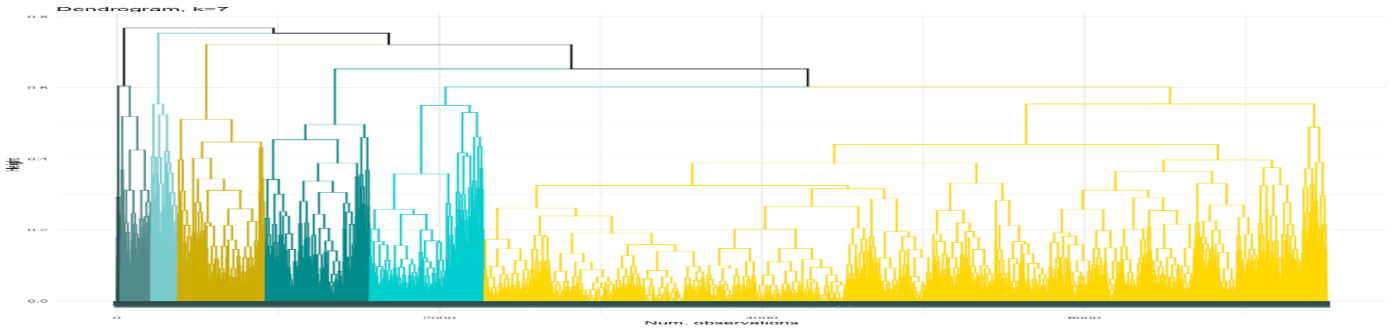


Figure 6: search for best k shows 3 or 7 are best for this type of clustering.



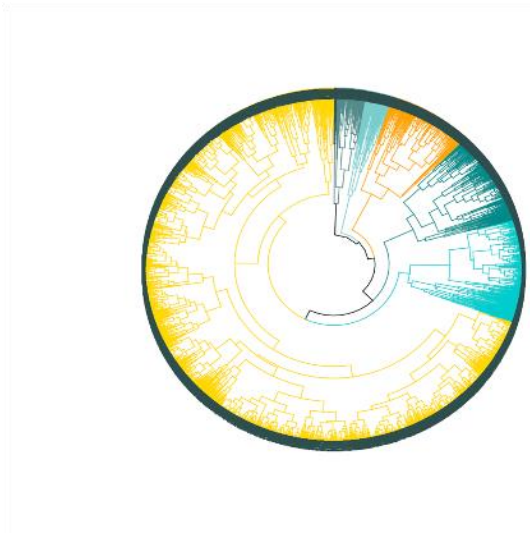


Figure 7: Hierarchical Clustering with 7 clusters

The problem: the clusters were not converging on meaningful cluster sizes. Only the small clusters were breaking further apart, leaving one huge cluster intact without ability to understand what's in it.

	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8	Test 9	Test 10	Test 11	Test 12	Test 13	Test 14
cluster.number	2.00	3.00	4.00	5.00	6.00	7.00	8.00	9.00	10.00	11.00	12.00	13.00	14.00	15.00
n	7500.00	7500.00	7500.00	7500.00	7500.00	7500.00	7500.00	7500.00	7500.00	7500.00	7500.00	7500.00	7500.00	7500.00
within.cluster.ss	176.30	163.90	145.28	121.73	120.17	105.22	100.81	99.55	94.09	92.34	91.64	90.62	88.62	87.41
average.within	0.20	0.19	0.18	0.16	0.16	0.15	0.15	0.15	0.14	0.14	0.14	0.14	0.14	0.14
average.between	0.30	0.32	0.28	0.27	0.27	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26
wb.ratio	0.66	0.59	0.63	0.60	0.60	0.58	0.57	0.56	0.55	0.55	0.55	0.54	0.54	0.53
dunn2	1.32	1.23	1.06	1.03	1.03	0.96	0.95	1.06	0.90	0.97	0.97	0.97	0.97	0.97
avg.silwidth	0.35	0.35	0.30	0.33	0.33	0.34	0.34	0.34	0.33	0.33	0.32	0.32	0.33	0.33
Cluster- 1 size	7287.00	7121.00	6577.00	5931.00	5931.00	5220.00	5067.00	5067.00	5067.00	5067.00	5067.00	5067.00	5067.00	5067.00
Cluster- 2 size	213.00	213.00	213.00	646.00	646.00	646.00	646.00	646.00	646.00	646.00	646.00	646.00	646.00	582.00
Cluster- 3 size	0.00	166.00	544.00	213.00	190.00	190.00	190.00	190.00	190.00	190.00	190.00	170.00	170.00	170.00
Cluster- 4 size	0.00	0.00	166.00	544.00	544.00	544.00	544.00	544.00	544.00	544.00	544.00	20.00	20.00	20.00
Cluster- 5 size	0.00	0.00	0.00	166.00	166.00	166.00	166.00	129.00	129.00	129.00	90.00	544.00	457.00	457.00
Cluster- 6 size	0.00	0.00	0.00	0.00	23.00	711.00	153.00	153.00	153.00	153.00	153.00	90.00	90.00	90.00
Cluster- 7 size	0.00	0.00	0.00	0.00	0.00	23.00	711.00	711.00	527.00	527.00	527.00	153.00	153.00	153.00
Cluster- 8 size	0.00	0.00	0.00	0.00	0.00	0.00	23.00	23.00	23.00	23.00	23.00	527.00	527.00	64.00
Cluster- 9 size	0.00	0.00	0.00	0.00	0.00	0.00	0.00	37.00	37.00	37.00	37.00	23.00	23.00	527.00
Cluster- 10 size	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	184.00	107.00	107.00	37.00	37.00	23.00
Cluster- 11 size	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	77.00	77.00	107.00	107.00	37.00
Cluster- 12 size	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	39.00	77.00	77.00	107.00
Cluster- 13 size	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	39.00	39.00	77.00
Cluster- 14 size	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	87.00	39.00
Cluster- 15 size	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	87.00

Figure 8: testing for different number of cluster, 2-15

A better picture was obtained with K-Means clustering.

3.1.2 Results of Model Experiments: K-Means Clustering

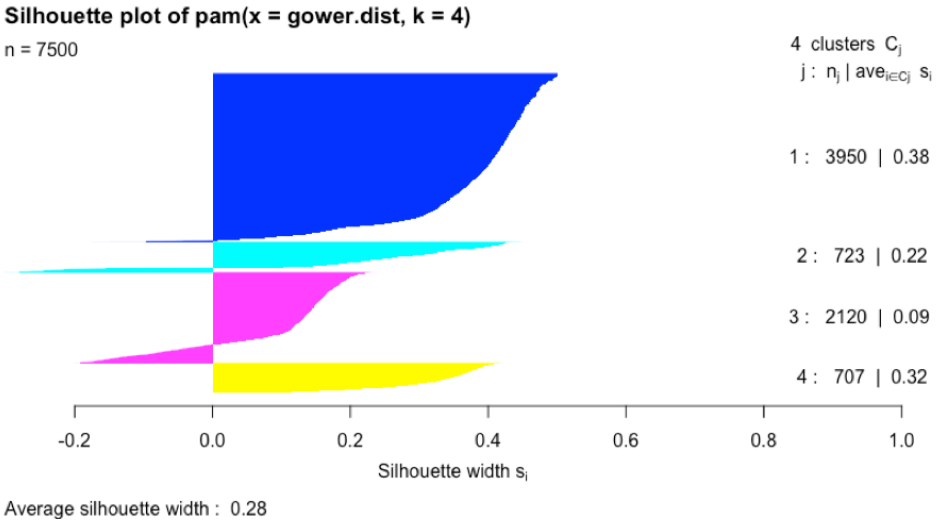


Figure 9: Using Silhouette analysis to find the best number of k for K-Means clustering. This method is specifically designed for working with categorical variables, so instead of looking at cluster size, it looks at the width of the distance between clusters. We want to choose the number that will maximize the distance between clusters, that how we found the best is k=4

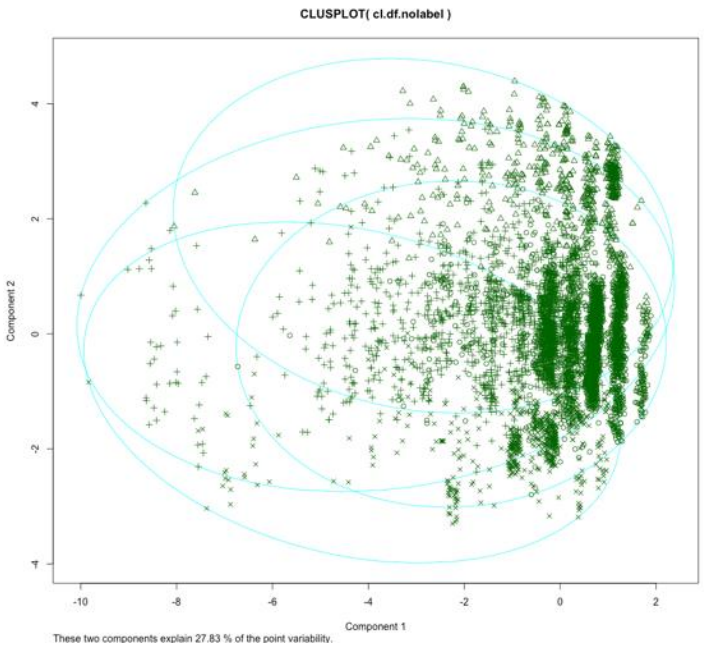


Figure 10: visualizing the clusters.

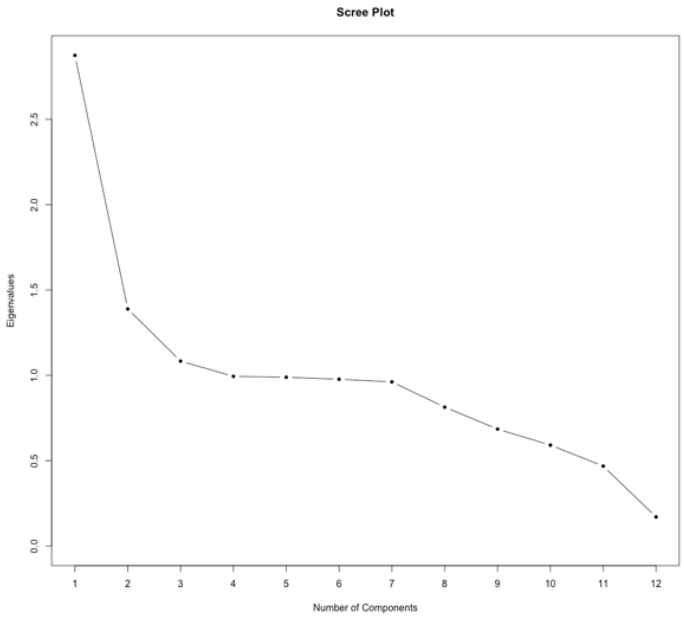


Figure 11: Principal Components Analysis to find number of PCs.

We want to analyze what's in the dimensions, because the data has too many dimensions. But how many components should we look at? The 4th point starts a line trend, where the principal components add very little to the variability. According to this plot, only the first 3-4 principal components are important to explain the variability in the data.

	Comp1	Comp2	Comp3	Comp4
Aircraft.Damage	-0.603		0.336	0.151
Number.of.Engines	0.672	0.149	-0.310	-0.158
Engine.Type	0.599		-0.415	-0.187
FAR.Description	-0.823		-0.522	-0.188
Purpose.of.Flight	0.825		0.519	0.187
Weather.Condition	-0.121	-0.695		0.160
Broad.Phase.of.Flight		-0.666	-0.206	0.187
Aircraft.Category		-0.150	0.262	-0.715
Amateur.Built		-0.321	0.285	-0.615
Schedule	0.411	-0.395	-0.364	
WeekDay	0.105		-0.206	
Month	0.114	0.328		

Importance (Variance Accounted For):

	Comp1	Comp2	Comp3	Comp4
Eigenvalues	2.7469	1.3608	1.2943	1.1279
VAF	22.8911	11.3401	10.7855	9.3993
Cumulative VAF	22.8900	34.2300	45.0200	54.4200

Figure 12: Variables that have great impact:

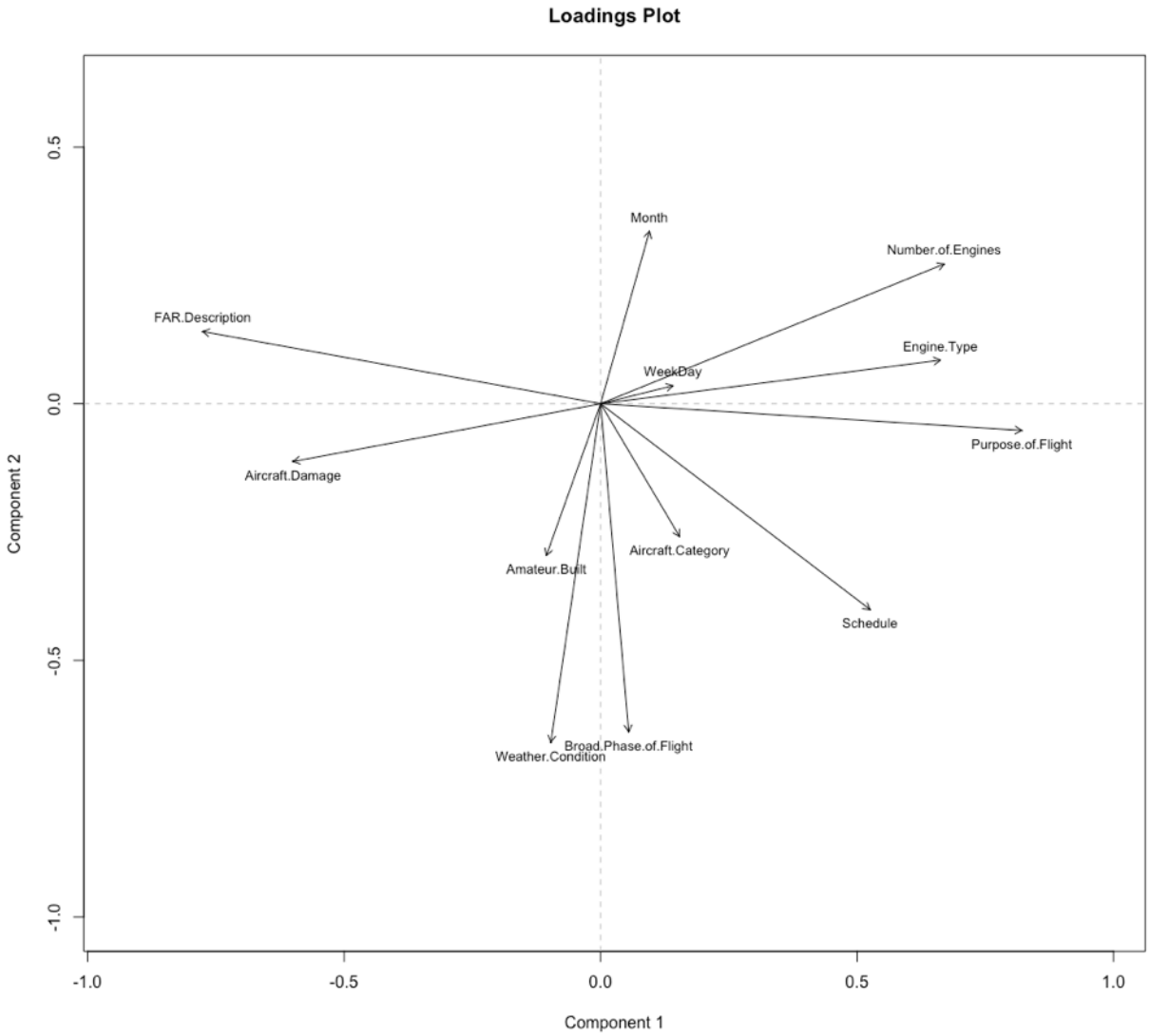


Figure 13: a loadings plot

Variables that have great impact:

On Component 1: Number.of.Engines, Engine.Type, Purpose.of.Flight, and Schedule have great impact, as well as Aircraft.Damage and FAR.Description. You can interpret this component to be reflecting attributes of the aircraft itself.

On Component 2: Broad.Phase.of.Flight and Weather.Condition have great impact, and Month, Aircraft.Category, Amateur.Built have medium impact. You can interpret this component to be more reflective of attributes of the flight.

Weekday has no significant impact on either.

3.2 Results of Model Experiments: Association Rule Mining

Association rule mining generated rules that included a known fact “bad weather causes more fatality”.

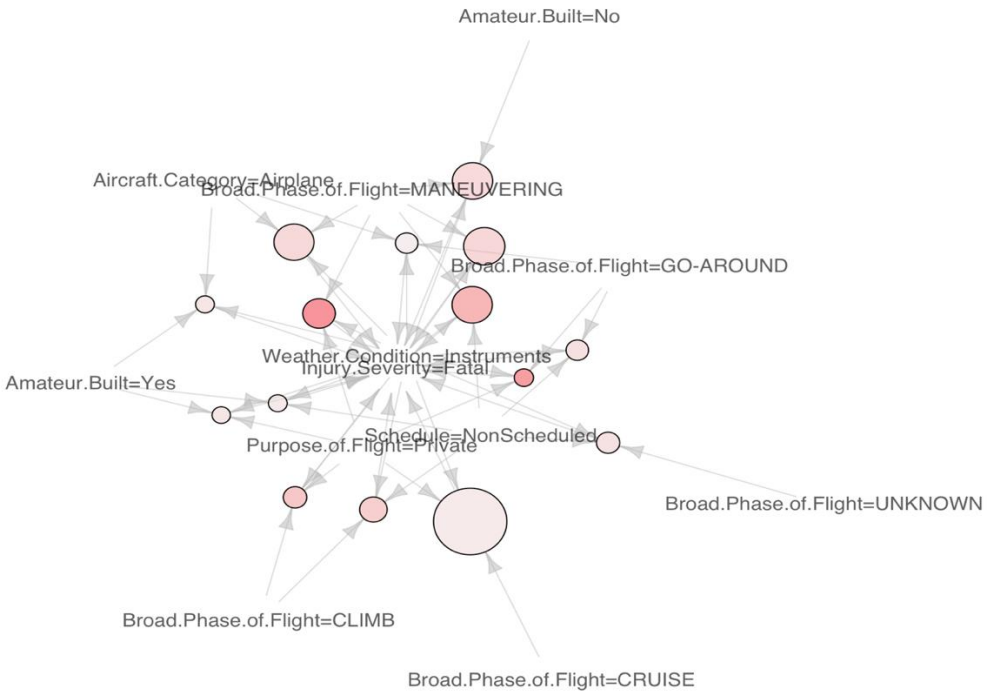


Figure #. Associate Rule First Iteration

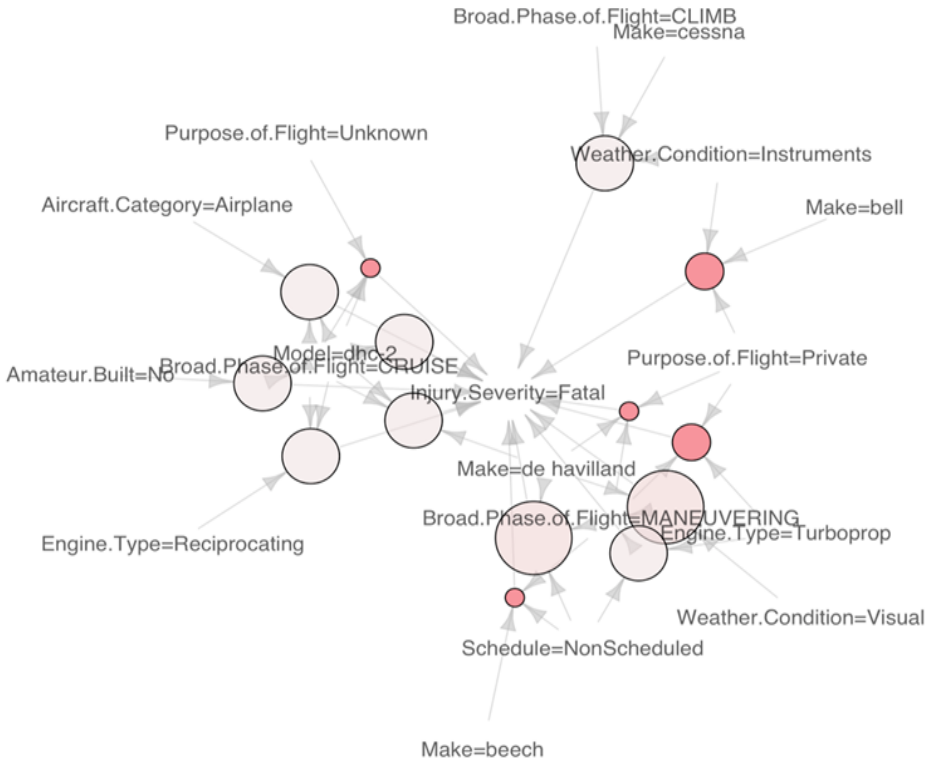


Figure #. Associate Rule second Iteration

3.3 Results of Model Experiments: Decision Tree

<u>Associate Rule Cross Reference</u>

Association Rule #1 :

{Weather.condition = Instructments, Broad.Phase.of.Flight = Maneuvering} =>
{Injury.Severity=Fatal}

Decision Tree Evaluation of Association Rule #1:

```
fit16 <- rpart(Injury.Severity ~ Weather.Condition + Broad.Phase.of.Flight,
        data=Aviations_train,
        method="class",
        control=rpart.control(minsplit=1, cp=0))
```

Table 8

Confusion Matrix and Statistics with Model Accuracy 80.5%

Prediction	Fatal	No-Injury	Non-Fatal
Fatal	813	0	2649
No-Injury	37	0	509
Non-Fatal	420	0	14536

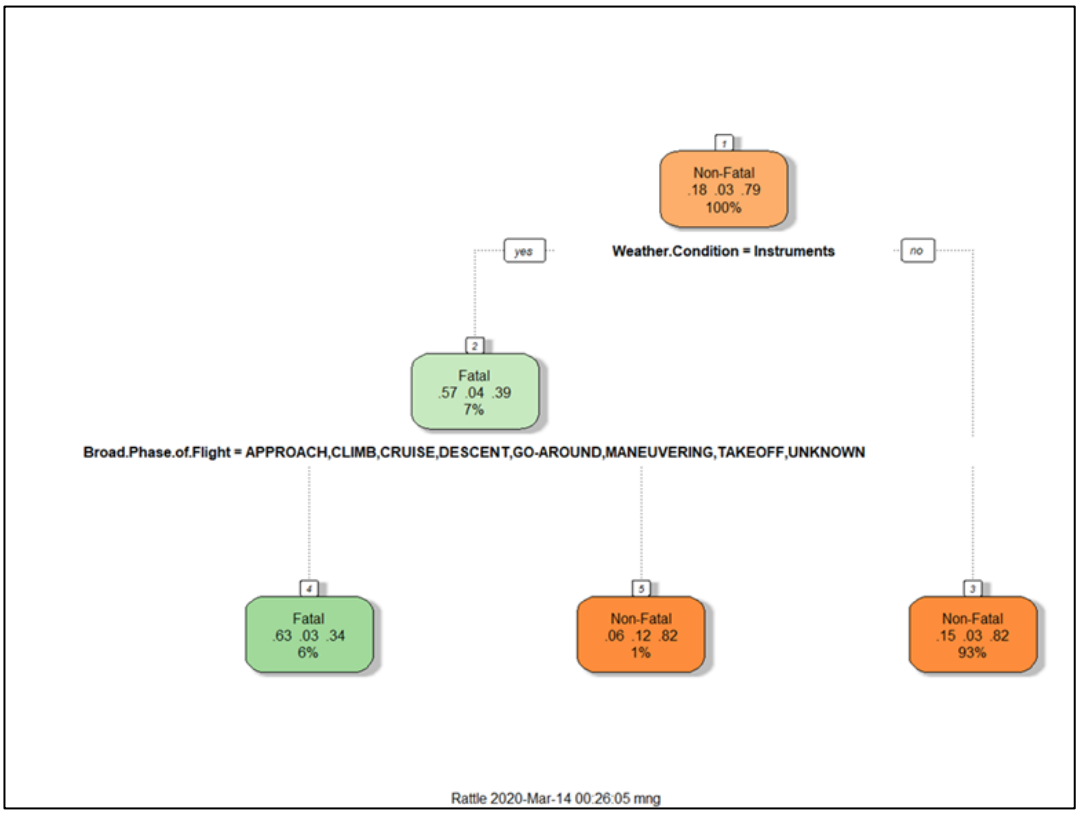


Figure #. Decision Tree based on Associate Rule #1

<u>Combined variables with model accuracy >95%</u>

Decision Tree with combined variables

```
fit19 <- rpart(Injury.Severity ~ Total.Fatal.Injuries + Schedule +  Aircraft.Category +
FAR.Description,
        data=Aviations_train,
        method="class",
        control=rpart.control(minsplit=2, cp=0.0001))
```

Table 9

Confusion Matrix and Statistics with Model Accuracy 97.2%

Prediction	Fatal	No-Injury	Non-Fatal
Fatal	3563	2	0
No-Injury	3	218	297
Non-Fatal	0	231	14650

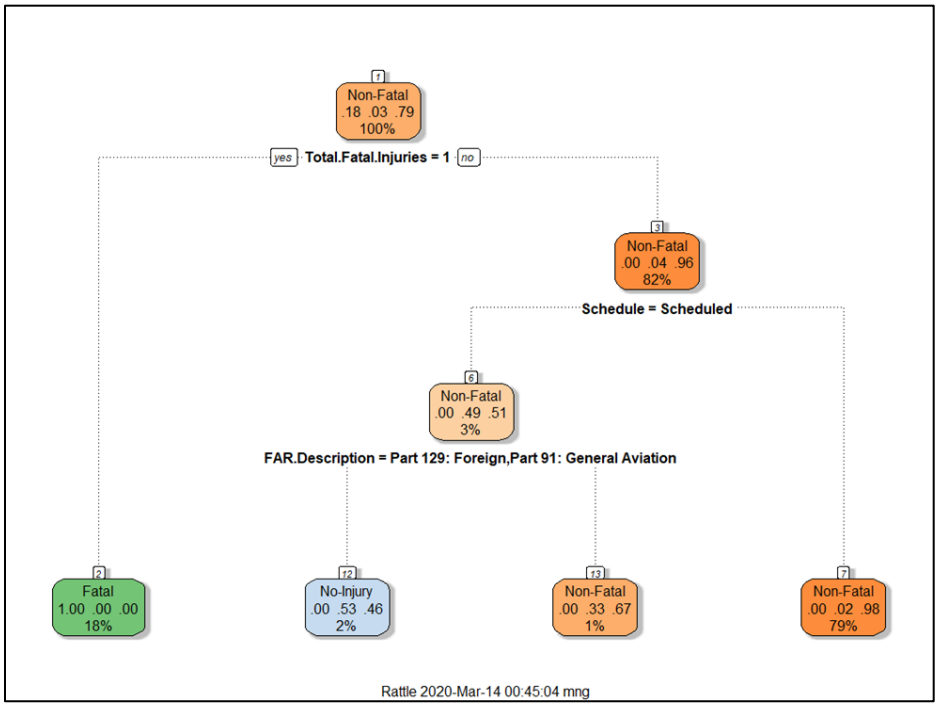


Figure #. Decision Tree based combined variable

Decision tree with combined variables of > 95% in model accuracy had the best accuracy performance.

3.4 Results of Model Experiments: Support Vector Machine

SVM Classification Plot

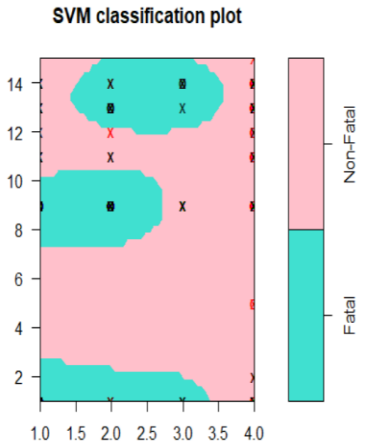


Figure #. SVM classification plot

From the figure above, this is a **radial model** as the diagram depicts a hyperplane observation

Table 9

Fine Tuning Results

KERNEL									
	Linear	*Cost*	0.25	0.5	0.75	1	1.25	1.5	**62%**
	Radial	*Cost*	0.25	0.5	0.75	1	1.25	1.5	**68%**
		Sigma	0.01		0.015		0.2		
	Polynomial	*Cost*	0.25	0.5	0.75	1	1.25	1.5	**67%**
		Degree	1		2		3		

```
               Accuracy : 0.6803
                 95% CI : (0.6573, 0.7026)
    No Information Rate : 0.802
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2117

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.7098
            Specificity : 0.5606
         Pos Pred Value : 0.8675
         Neg Pred Value : 0.3229
             Prevalence : 0.8020
         Detection Rate : 0.5693
   Detection Prevalence : 0.6563
      Balanced Accuracy : 0.6352

       'Positive' Class : NonFatal
```

```
                Reference
Prediction Fatal NonFatal
  Fatal      185      388
  NonFatal   145      949
```

Reference

Prediction	Fatal	NonFatal	
	Fatal	**NonFatal**	
Fatal	185	388	1134
NonFatal	145	949	1667
			0.680264
			68.0%

- SVM had best accuracy performance using the Radial kernel at a “cost” of 1.5 and “sigma” of 0.2. These numbers are the tuning parameters for Support Vector Machine.

3.5 Summary of experimentation and results

Table 10

Machine Learning		
Unsupervised	Associate rule mining	Amateur Built=No Weather Condition=Instruments Phase of Flight=Maneuvering Weather.condition = Instruments, Broad.Phase.of.Flight = Maneuvering} => {injury.Severity = Fatal}
	Clustering	FAR.Description, weather, phase of flight, purpose of flight
Supervised	SVM	Model accuracy 68% (10-fold cross validation)
	Decision tree	Model accuracy 97% (3-fold validation)

4. Conclusions

At the start of this project, there were crash incidents of 737 Max airplane or helicopter reporting. Although it has been said, statistically speaking, that flying is the safest way to travel, caution should be forewarn as observed through this project analysis. Factors such as weather conditions, mechanical failure, pilot error and other various causes can contribute to flight fatalities statistics and drastically impact the way on understanding aviation and its safety.

The project found that combination of various factors can predict possible fatality, aircraft damage, or serious injury. Interesting various combinations found from the project were when aircraft controlled by pilots that might be impaired in adverse conditions such as weather, equipment, training, or different flight stages. Weather seemed to play a significant role in determining accident severity. When a severe weather occurred, a pilot needed to use navigation instruments as the primary references rather than visual cues such as cloudy weather under the Instrument Meteorological Conditions (IMC) criteria for safer flight operation. However, personal aircraft such as helicopters and balloons did not have state of the art navigation instruments, and were relying on visual reference only. Thus, these explained one of the possible root causes to Kobe Bryant's helicopter accident.

A possible explanation is personal aircrafts were lacking navigation instrument and pilot support in comparison to the big commercial airliners commercial airliners were extremely concerning and dangerous under general aviation regulation. Based on this project, these are the general recommendations for pilots:

1. Improve access to real time weather reports, especially for personal aviation enthusiasts as well as pilots of smaller craft.
2. Increase training to practice on landings in a variety of weather conditions.
3. Need additional hours for pilot to acquire and master skillset on flying at various weather conditions before getting a private pilot's license.

The same research methodology applied for this project can be applied to other means of transportation to better understand different parameters that affect travel across the globe and how to mitigate the potential issues beforehand. Furthermore, the Covin-19 pandemic put a halt on the globe air traffic to curb the spread of the virus , which will be an interesting single factor for increasing infection rate globally and a new variable to be included in an updated NTSB dataset for future analysis.

5 References

5.1 Reference list

1. BBC.com website. *Plane crash fatalities fell more than 50% in 2019.* Updated 2 Jan 2020. [Online] Available from: https://www.bbc.com/news/business-50953712 [Accessed 24 Feb 2020].
2. GormAnalysis.com website. *Decision Trees in R using Rpart*. Updated 24 Aug 2014. [Online] Available from: https://www.gormanalysis.com/blog/decision-trees-in-r-using-rpart/ [Accessed 10 Feb 2020].
3. University of Cincinnati Business Analytics R Programming Guide website. *Regression Decision Trees.* No Date. [Online] Available from: *https://uc-r.github.io/regression_trees* [Accessed 14 Feb 2020].

5.2 Additional Resources

1. Individual aircraft accident investigation reports are available here: https://ntsb.gov/investigations/AccidentReports/Pages/aviation.aspx
2. A dataset on all investigations since 1962 (+1 from 1948) is available for download here: http://app.ntsb.gov/aviationquery/Download.ashx?type=csv
3. Safety recommendations are available here: https://ntsb.gov/safety/safety-recs/_layouts/ntsb.recsearch/RecTabs.aspx
4. The data dictionary was obtained here: https://www.ntsb.gov/_layouts/ntsb.aviation/index.aspx

6 Appendix

6.1 Glossary of Terms

Event – an event denotes an aviation event reportable to NTSB. A reportable event is an aviation accident or serious incident.

Aviation accident – The NTSB defines a reportable "accident" as "an occurrence associated with the operation of an aircraft that takes place between the time any person boards the aircraft with the intention of flight and all such persons have disembarked, and in which any person suffers death or serious injury, or in which the aircraft receives substantial damage."

Aviation serious incident – one of a specific list of events; for example: a complete loss of information from more than 50% of an aircraft's cockpit displays.

Aviation incident – any other occurrence that affects or could affect the safety of operations, or when an accident or serious incident nearly occurred (only marginally avoided).

Aircraft damage – there are 3 levels of damage possible: minor, substantial, or destroyed. An occurrence in which the aircraft only received minor damage and there are no injuries, is not reportable. It is reportable if the aircraft received substantial damage, or obviously, if the aircraft is destroyed.

Substantial damage – "damage or failure which adversely affects the structural strength, performance, or flight characteristics of the aircraft, and which would normally require major repair or replacement of the affected component."

Recommendation – a course of action that NTSB investigators believe can address a safety concern, resulting from an event. It is communicated as a letter to an assignee and is often issued before any investigation report; recommendations are listed in the final investigation report, once a report is published. The NTSB doesn't have any official authority to regulate the transportation industry, so adoption of these recommendations depends on their reputation as a respected and credible body that produced timely, well considered, professional recommendations.

6.2 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 using Global 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 Regulation	The applicable regulation part (14 CFR) or authority the aircraft was operating under at the time of the accident. The Federal Aviation Regulations (FARs) are rules prescribed by the Federal Aviation Administration (FAA) governing all aviation activities in the United States. The FARs are part of Title 14 of the Code of Federal Regulations (CFR).
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.