Aviation Fatalities

Machine Learning in R

IST 707 DATA ANALYTICS

Dr Gates

- Editt
- Maria
- Bhavya
- Veasna

Agenda

- Introduction
- Problem Statement
- Objective, Analysis & Result
  - Clustering
  - Associate Rule Mining
  - Decision Tree
  - Support Vector Machine
- Conclusion

Introduction

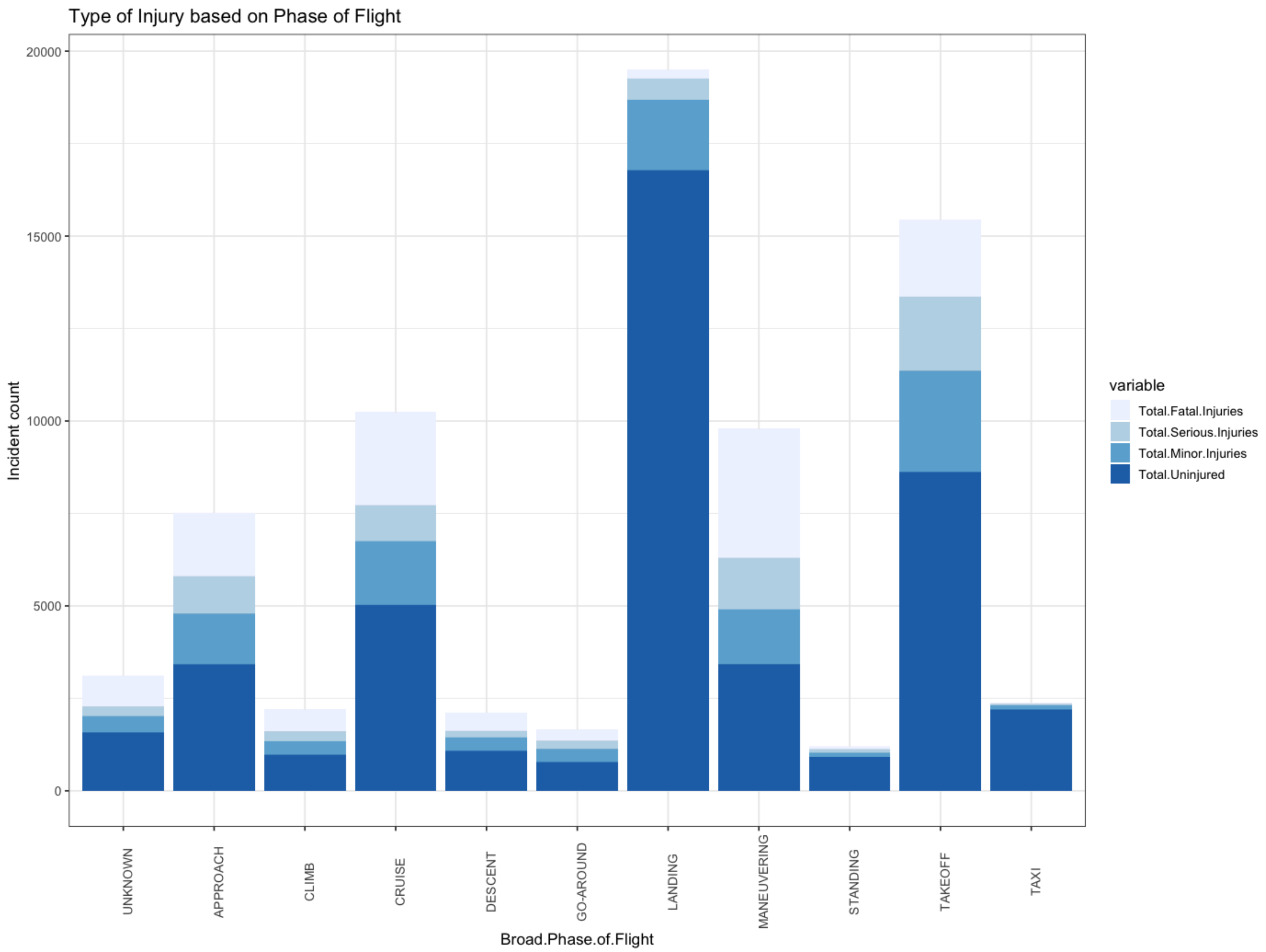
- Rising aviation accidents that involves faulty craftsmanship, pilot errors and systems errors led our individual curiosity
- Data is based on NTSB harnessing Data Science methodologies to investigate aviation crashes

Problem Statement

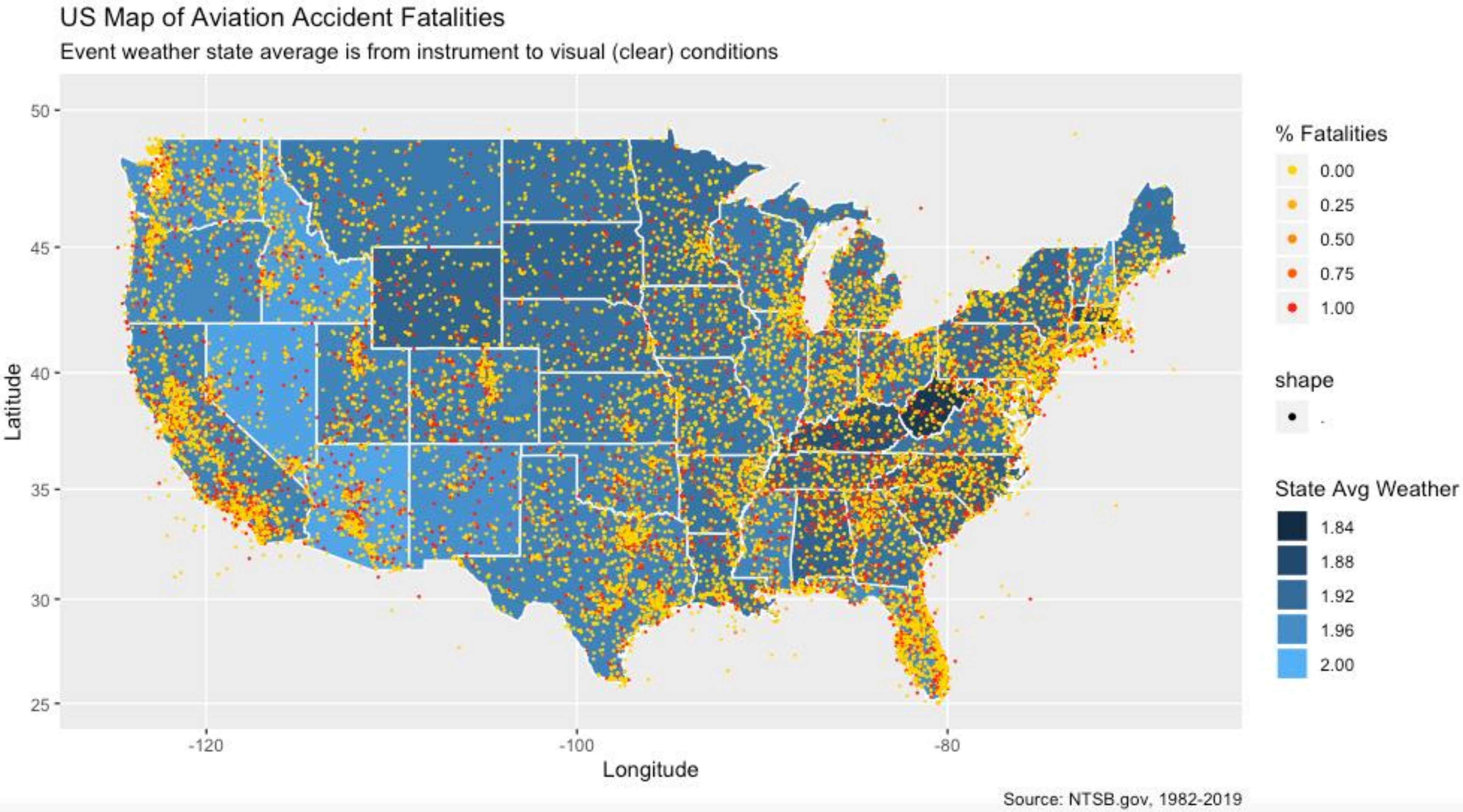
- Investigate commonalities among accident fatalities:

  Possible examples:
  - Bad weather conditions
  - Pilot Error
  - Mechanic Failure
  - Other causes

- NTSB data consists of approximately ~84k observations & 32 variables (numeric, nominal, ordinal)



Data Visualization

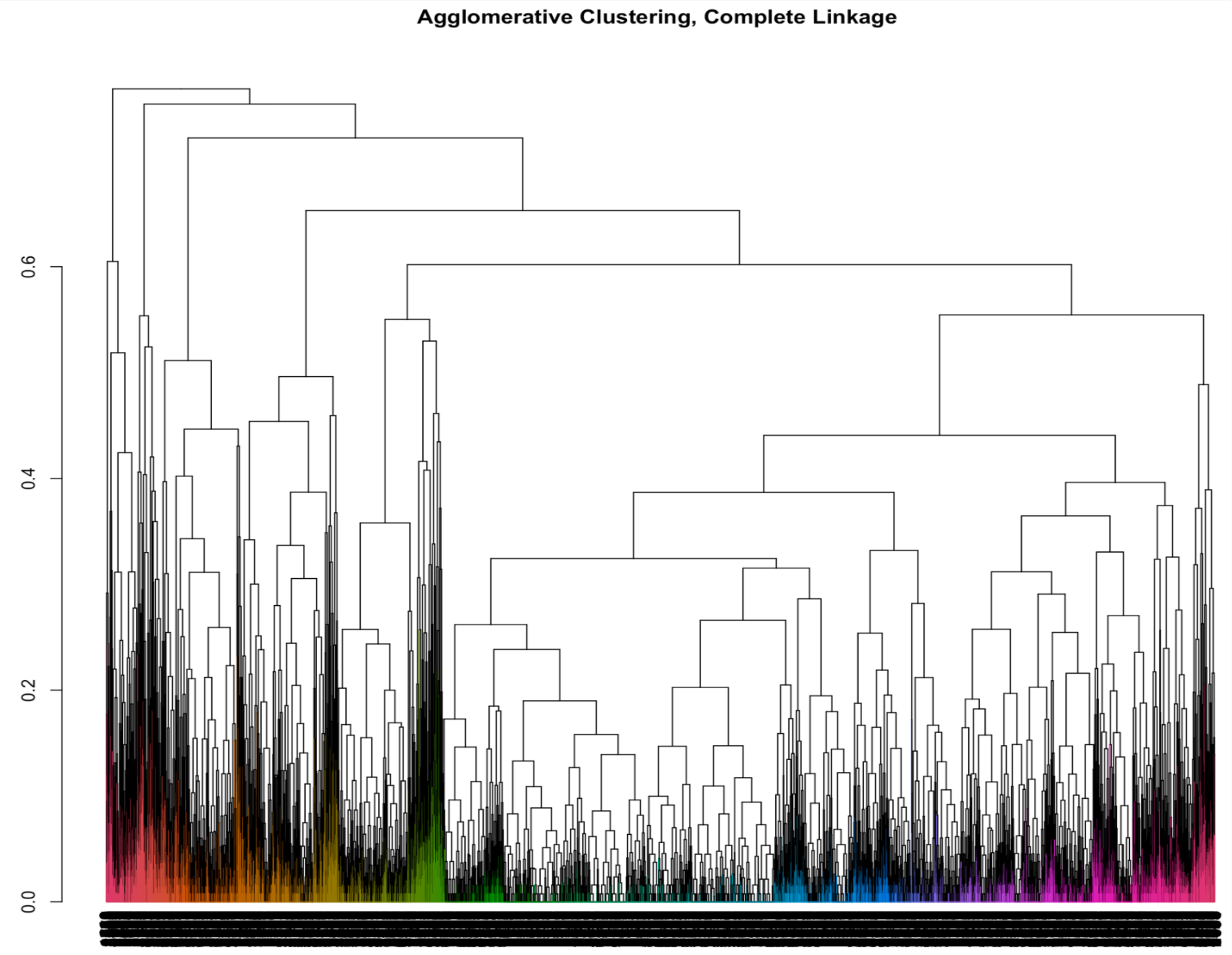


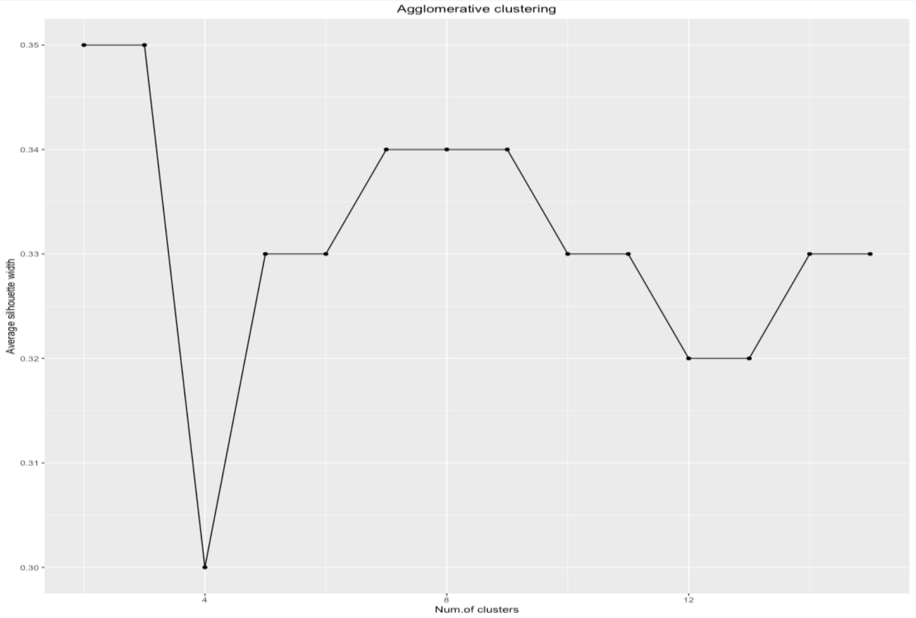
Clustering

ANALYSIS & MODEL

- Objective
  - To identify what characteristics are there for fatal accidents, and if we can identify distinct clusters in our data
- Process
  - Perform K-Means cluster analysis; Challenge: categorical attributes
- Tuning
  - Perform Principal Component Analysis to reduce dimensionality; perform hierarchical clustering to find more insights; tune number of clusters

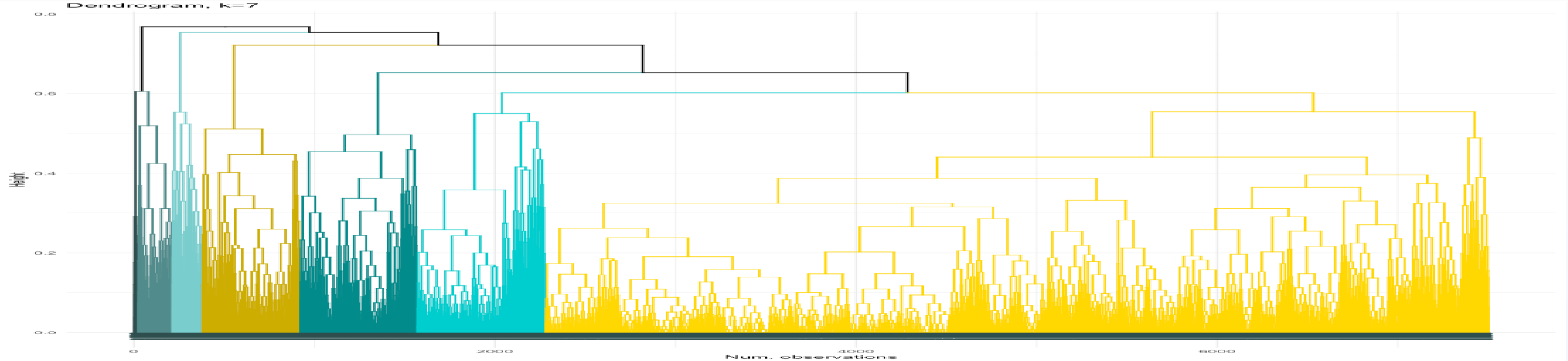
Visual



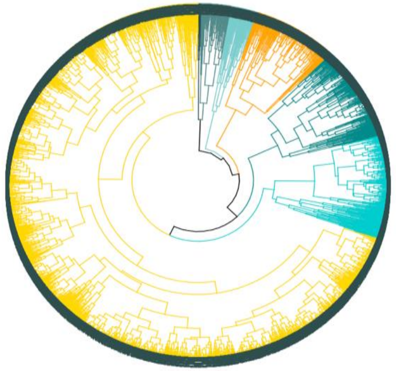


Hierarchical clustering recommends 3 or 7 clusters

Visual

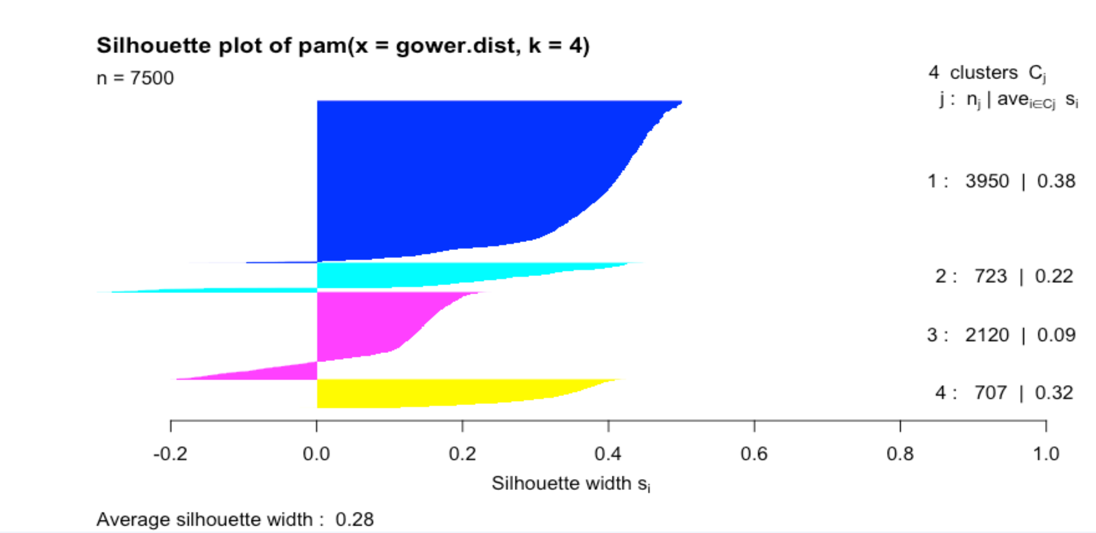


	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8	Test 9	Test 10	Test 11	Test 12	Test 13	Test 14
cluster.number	2.00	3.00	4.00	5.00	6.00	7.00	8.00	9.00	10.00	11.00	12.00	13.00	14.00	15.00
n	7500.00	7500.00	7500.00	7500.00	7500.00	7500.00	7500.00	7500.00	7500.00	7500.00	7500.00	7500.00	7500.00	7500.00
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



Visual

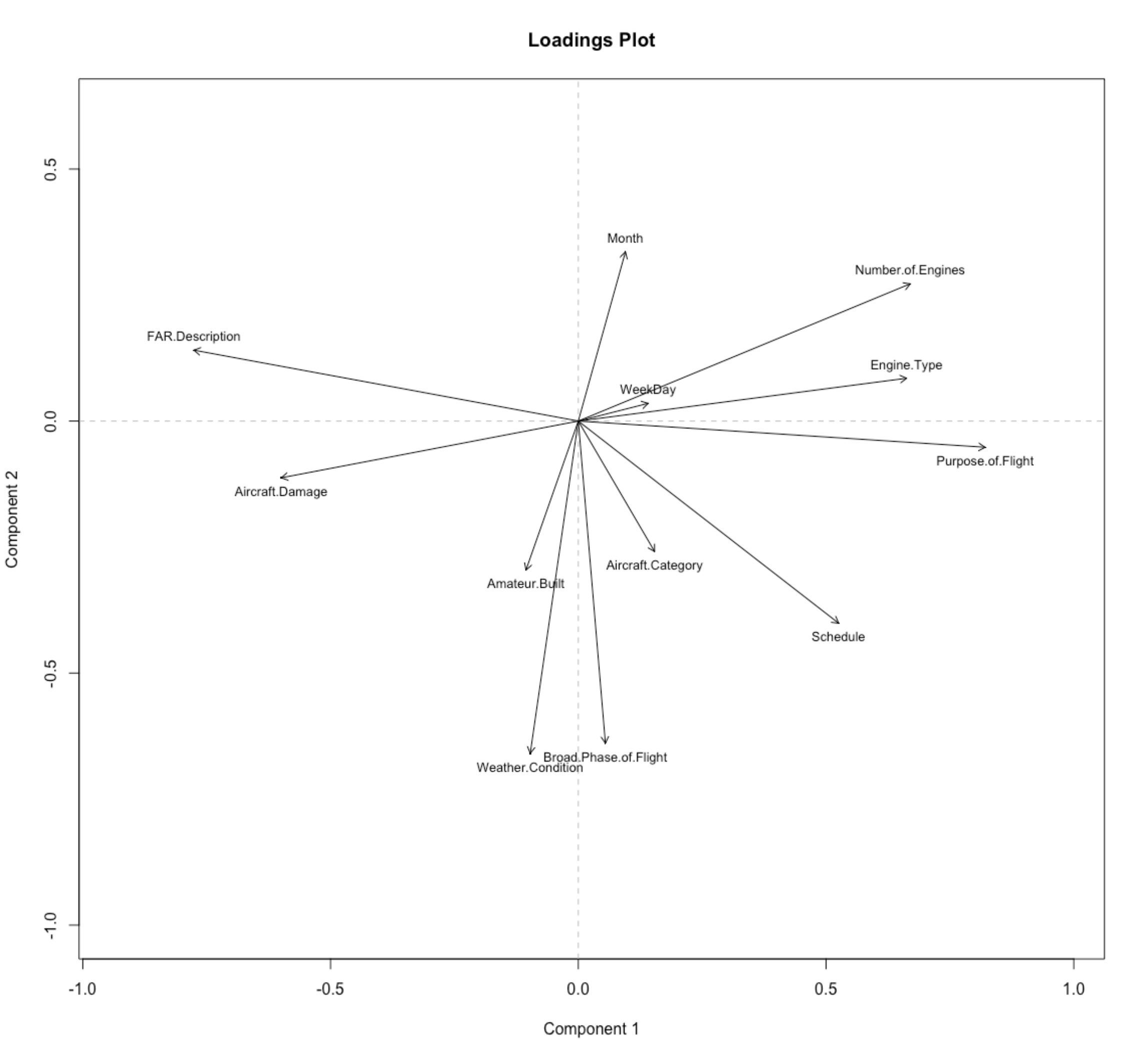
PCA helps explain what's in each component



	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



Association Rule Mining

ANALYSIS & MODEL

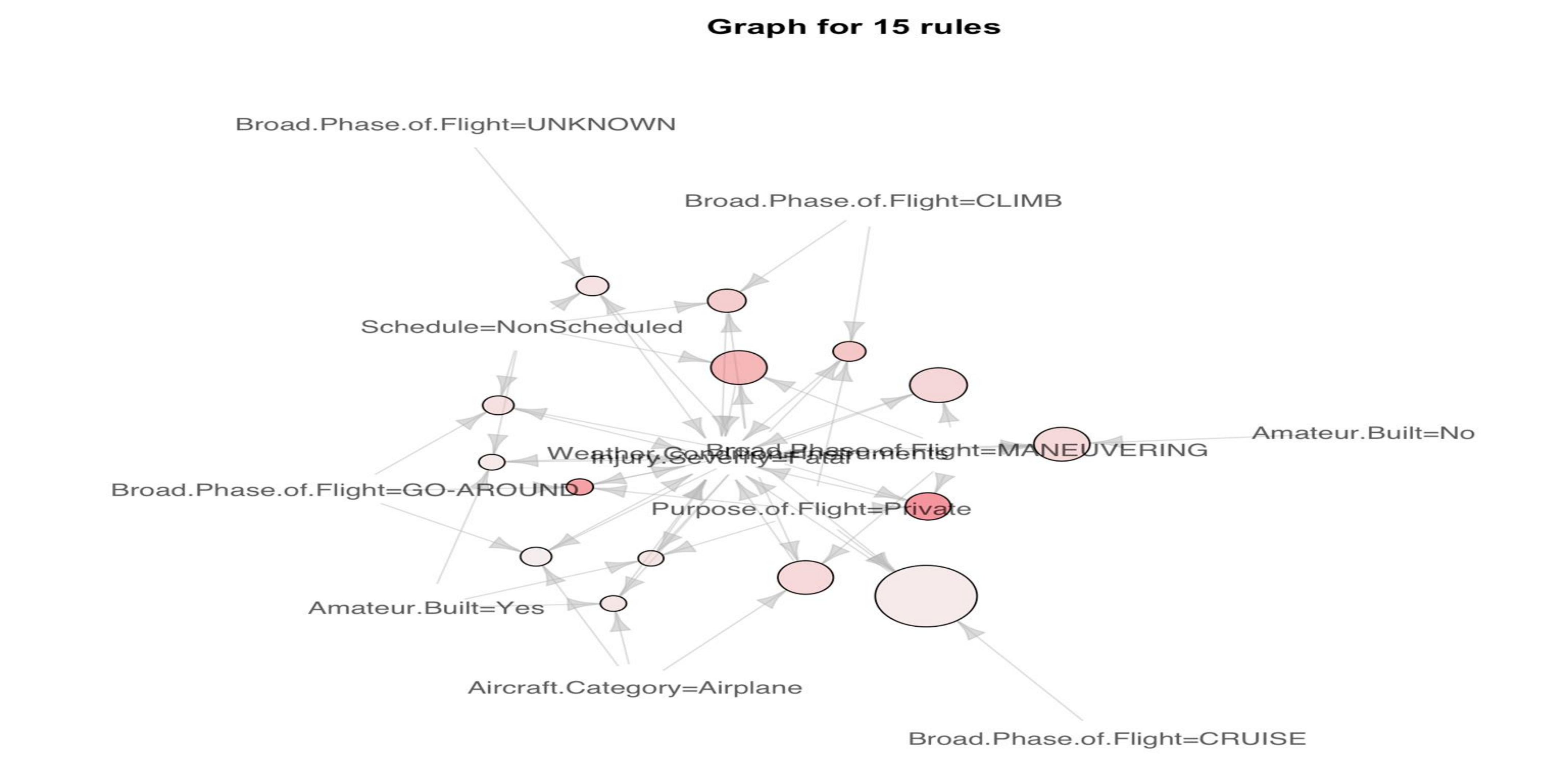
- Objective:
  - To find the probability of relationship between injury severity (Fatal) and various other attributes
- Process:
  - Two iterations with two sets of attributes
- Tuning:
  - Tuned support and confidence to arrive at meaningful rules

Association Rule Mining

ANALYSIS & MODEL

Iteration 1 (15 Rules)	Injury Severity = Fatal Lift 3.8 - 4.1
Injury Severity Aircraft Category Amateur Built Schedule Purpose of Flight Weather Condition Phase of Flight	Amateur Built = No Weather Condition = Instruments Phase of Flight = MANEUVERING 375 Aircraft Category = Airplane Weather Condition = Instruments Phase of Flight = MANEUVERING 371 Purpose of Flight = Private Weather Condition = Instruments Phase of Flight = CRUISE 814

Visual

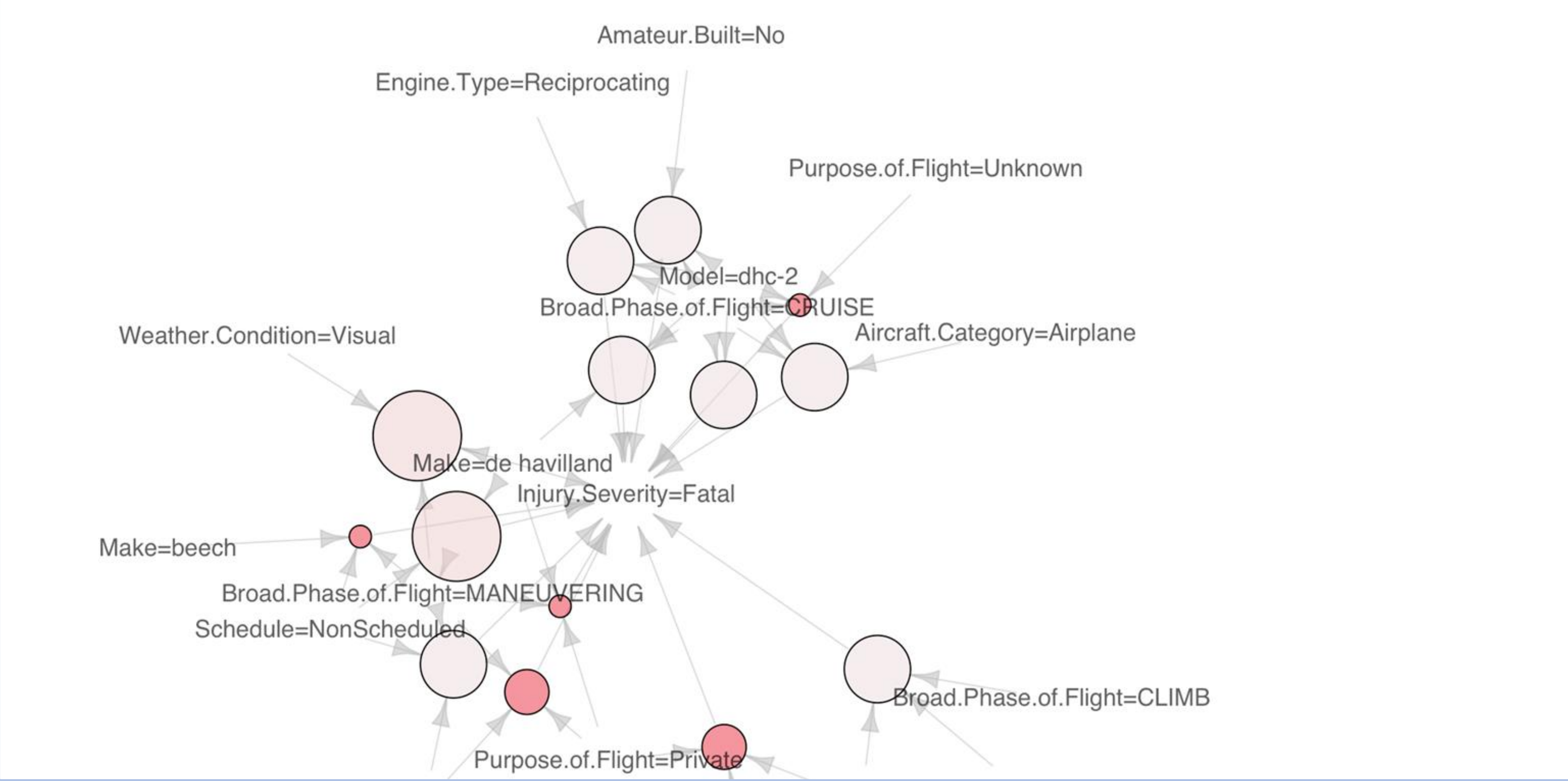

Graph for 15 rules
Broad.Phase.of.Flight=UNKNOWN
Broad.Phase.of.Flight=CLIMB
Schedule=NonScheduled
Amateur.Built=No
Broad.Phase.of.Flight=MANEUVERING
Injury.Severity=Fatal
Broad.Phase.of.Flight=GO-AROUND
Purpose.of.Flight=Private
Amateur.Built=Yes
Aircraft.Category=Airplane
Broad.Phase.of.Flight=CRUISE

Association Rule Mining

ANALYSIS & MODEL

Iteration 2 (14 Rules)	Injury Severity = Fatal Lift 5.8 - 6.1
Injury Severity	Model=dhc-2, Amateur Built=No Phase of Flight=CRUISE 6 Make=de havilland Weather Condition=Visual Phase.of.Flight=MANEUVERING 7
Aircraft Category	
Amateur Built	
Schedule	
Purpose of Flight	
Weather Condition	
Phase of Flight	
Make	
Model	
Air carrier	

Visual



Amateur.Built=No
Engine.Type=Reciprocating
Purpose.of.Flight=Unknown
Model=dhc-2
Broad.Phase.of.Flight=CRUISE
Weather.Condition=Visual
Aircraft.Category=Airplane
Make=de havilland
Injury.Severity=Fatal
Make=beech
Broad.Phase.of.Flight=MANEUVERING
Schedule=NonScheduled
Broad.Phase.of.Flight=CLIMB
Purpose.of.Flight=Private

Decision Tree

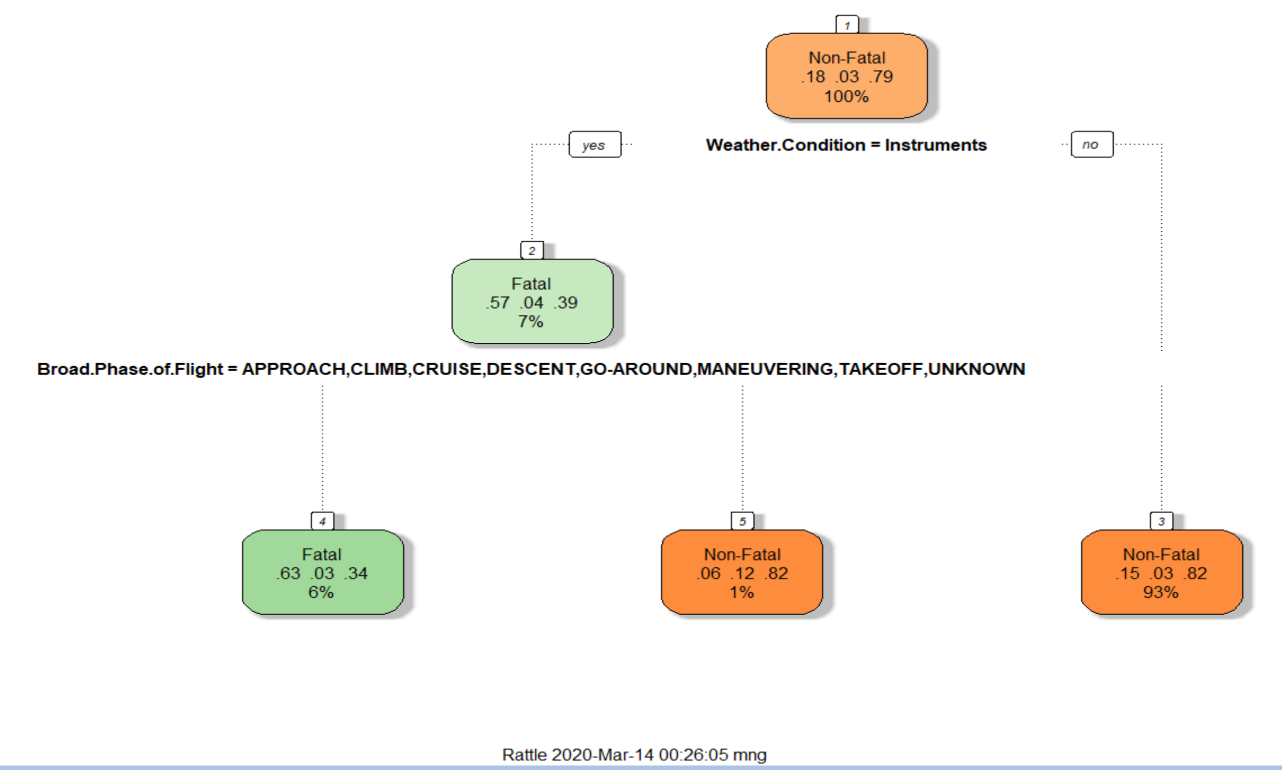
ANALYSIS & MODEL

- Purpose:
  - To identify a model with a high accuracy for prediction of injury severity
- Process:
  - Training data (2/3) & testing data (1/3)
  - Evaluate model accuracy by each factor
- Tuning
  - Cross reference with associate rules
  - Combine factors with model accuracy > 95%

Factor	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%

Visual - cross reference with association rule

- Association Rule #1
  - {Weather.condition = Instruments, Broad.Phase.of.Flight = Maneuvering} → {Injury.Severity = Fatal}



Confusion Matrix and Statistics

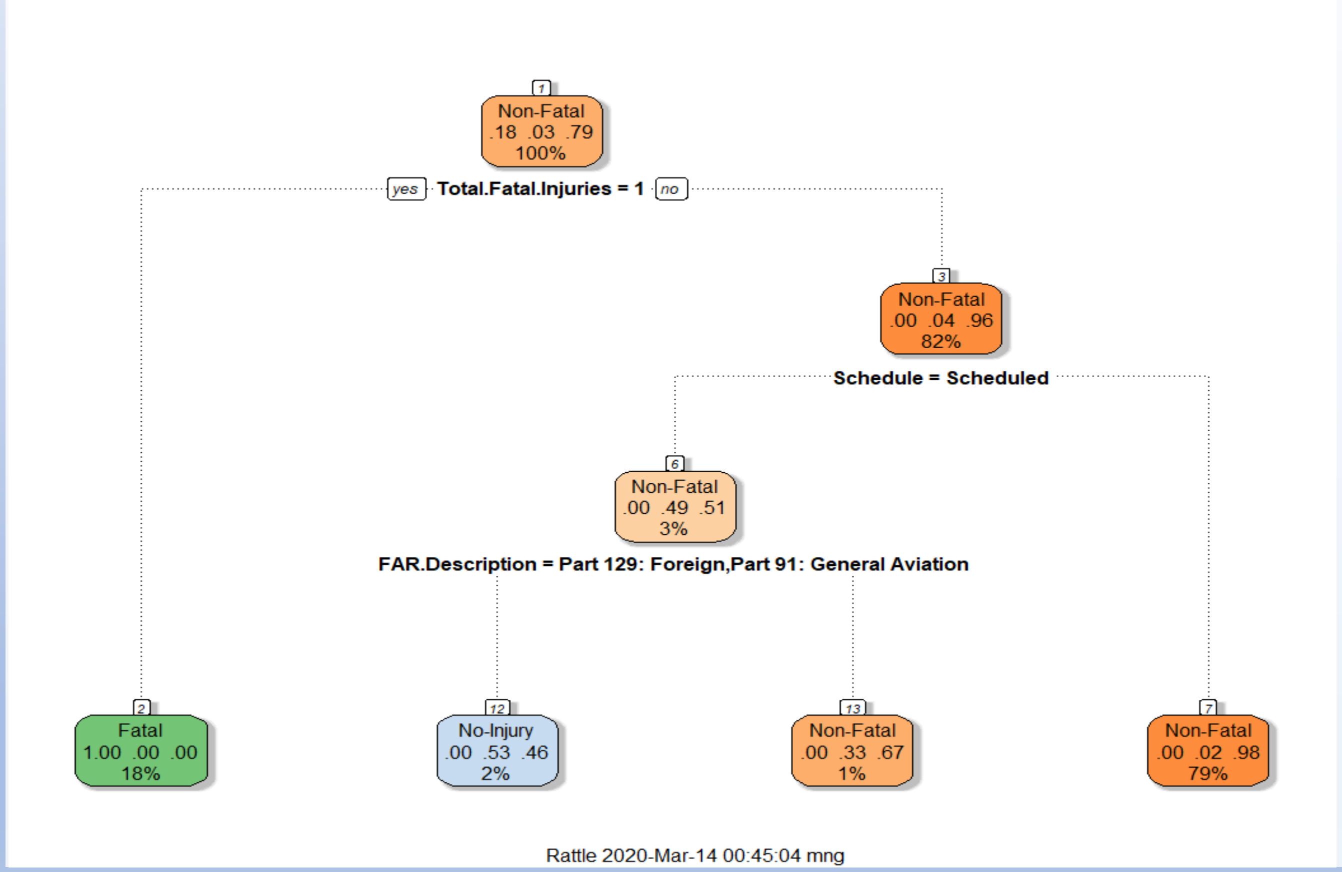
Prediction	Reference Fatal	No-Injury	Non-Fatal
Fatal	778	0	2787
No-Injury	36	0	482
Non-Fatal	390	0	14491

Model Accuracy: 80.5%

Visual - combined factors / tuning

Select factor with model accuracy > 95%:

Total.Fatal.Injuries, Schedule, Aircraft.Category, and FAR.Description




```
Confusion Matrix and Statistics

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

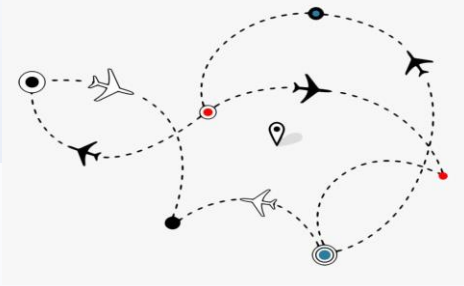
Model Accuracy: 97.2%

Support Vector Machine

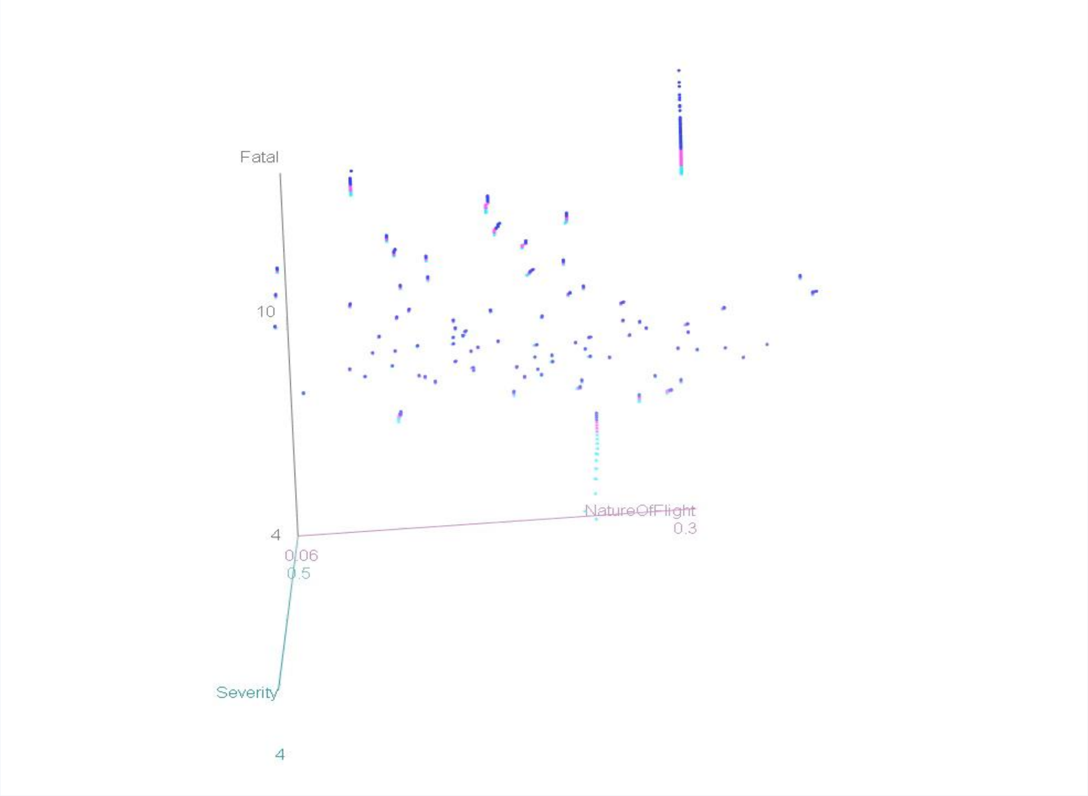
ANALYSIS & MODEL

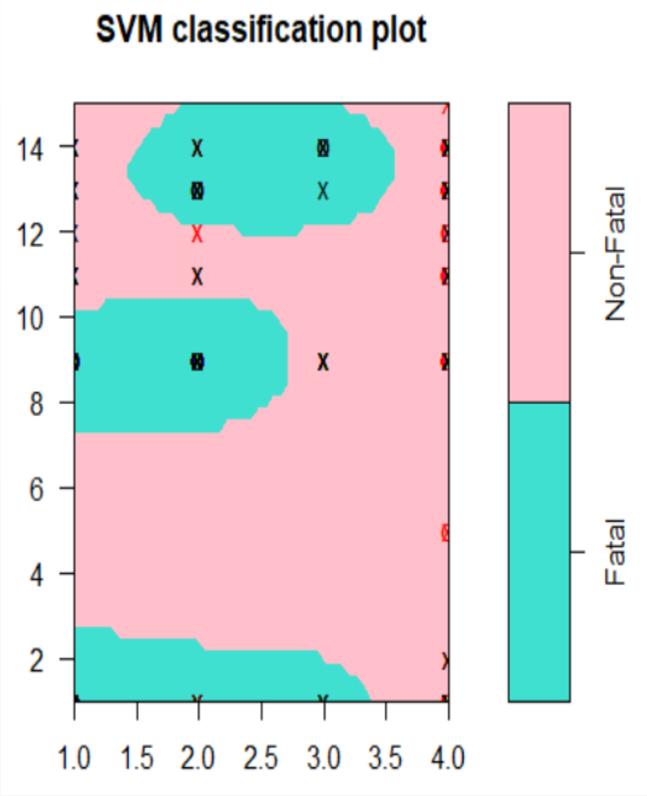
- Objective
  - Given a set of attributes to predict the target attribute: “Fatal” or “Non-Fatal”
- Process
  - Injury Severity, Built, Engine Type & Count, Flight Purpose, Weather Conditions, etc
  - Intuition and Correlation of choosing attributes
  - Transform nominal data to numeric for SVM algorithm
  - Compare different kernels
- Tuning
  - Using 10-fold Cross-validation, Kernel, Cost and Gamma


Professional
Amateur



Visual


Fatal
10
4
NatureOfFlight
0.3
0.06
0.5
Severity
4


SVM classification plot
14
12
10
8
6
4
2
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Non-Fatal
Fatal

Support Vector Machine

RESULT

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

```
               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	NonFatal
Fatal	185	388
NonFatal	145	949

1134

1667

0.680264

68.0%

Conclusion

Machine Learning		
Unsupervised	**Association Rule Mining**	{Amateur Built = No, Weather Condition = Instruments, Phase of Flight = MANEUVERING} → {Fatal Injury}
	Clustering	• Important Variables: FAR.Description, Weather Condition, Phase of Flight, Purpose of Flight
Supervised	**SVM**	• Model accuracy 68% (10-fold cross validation)
	Decision Tree	• Model accuracy 97% (3-fold cross validation)