Predicting General Aviation Accident Cause based on Pilot Profile

Springboard

Manuel Gomez

Problem

- Yearly aviation accident rates have remained steady at around 1.0 fatal accident(s) per 100,000 flight hours with some recent years slowing a slight increase
- As Improvements in automation and aircraft reliability continue, human factors are becoming the leading cause of aircraft accidents
- To further decrease the aviation accident rate, human factors need to be targeted

Stakeholders

Proposal

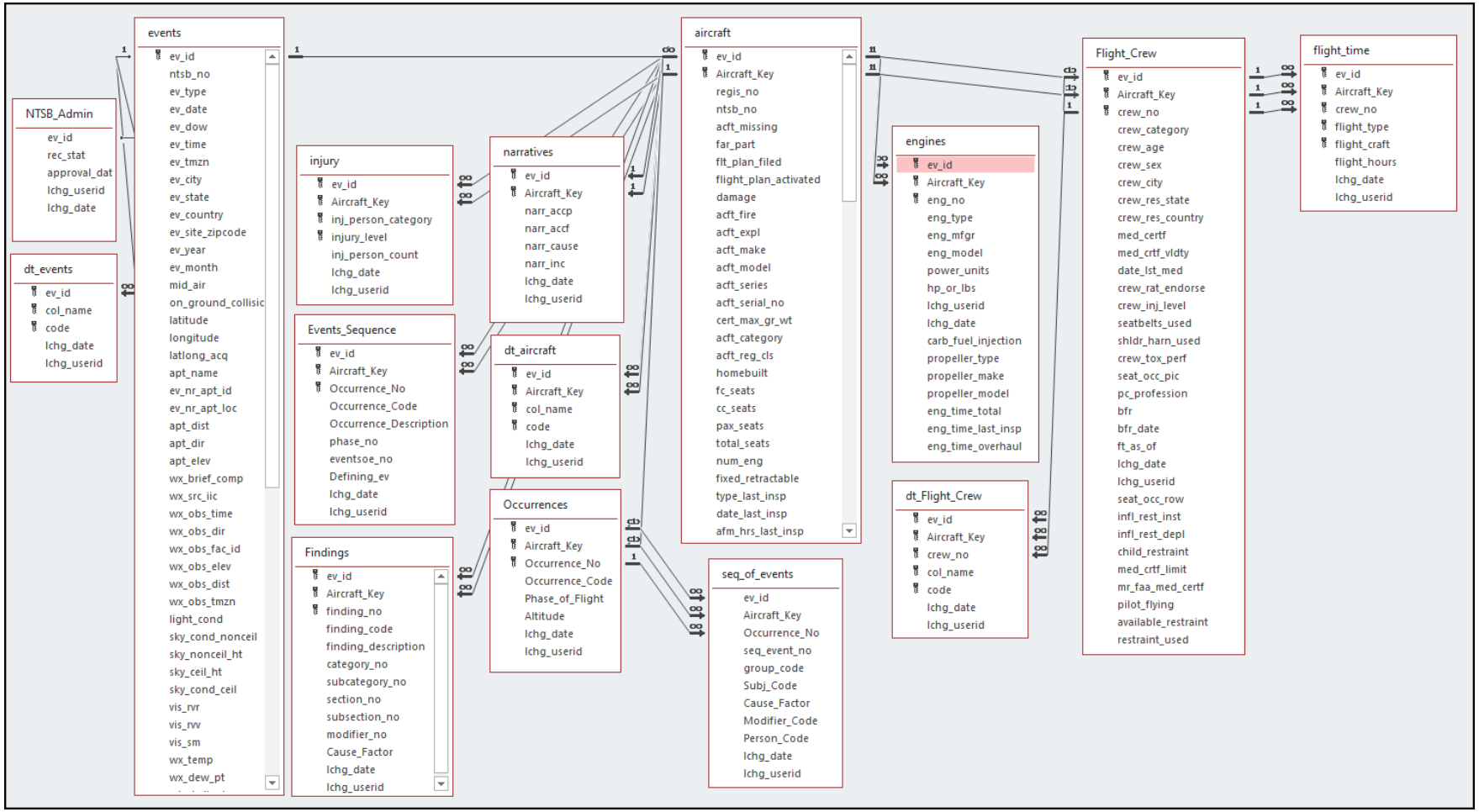
Determine Pilot Profile → Predict Likely Accident → Provide Awareness / Targeted Training To Pilot → 

NTSB Aviation Accident Database

Records of interest

General Aviation Operations
(Private, non-commercial flights)
Single Pilot Operations

34,900 Records
82 Features



Features

* Not all features represented

Numerical			Categorical		
Crew Age	Total Hours Instructor	Total Hours Instrument Flight	FAR Part	**Pilot Privilege**	Multi Engine Sea
Total Hours Flight Time	Total Hours in Accident Aircraft Type	Total Hours Night Instructor	Medical Certificate Type	Certificate Foreign	Airplane None
Total Hours Pilot in Command	Total Hours in Command of Accident Aircraft Type	Total Hours Multi Engine Instructor	**Medical Certificate Valid**	Certificate Military	Instructor Multi Engine
Total Hours Last 24 Hours	Total Hours in Instructor of Accident Aircraft Type	Total Hours Single Engine Instructor	Light Condition	Single Engine Land	Instructor Single Engine
Total Hours Last 30 Days	Total Hours Multi Engine	Total Hours Single Engine Instruction Received	Weather Condition	Multi Engine Land	Instructor Instrument
Total Hours Last 90 Days	Total Hours Night Instruction Received	Total Night Hours	Flight Type	Single Engine Sea	

High Incidence of Missing Values

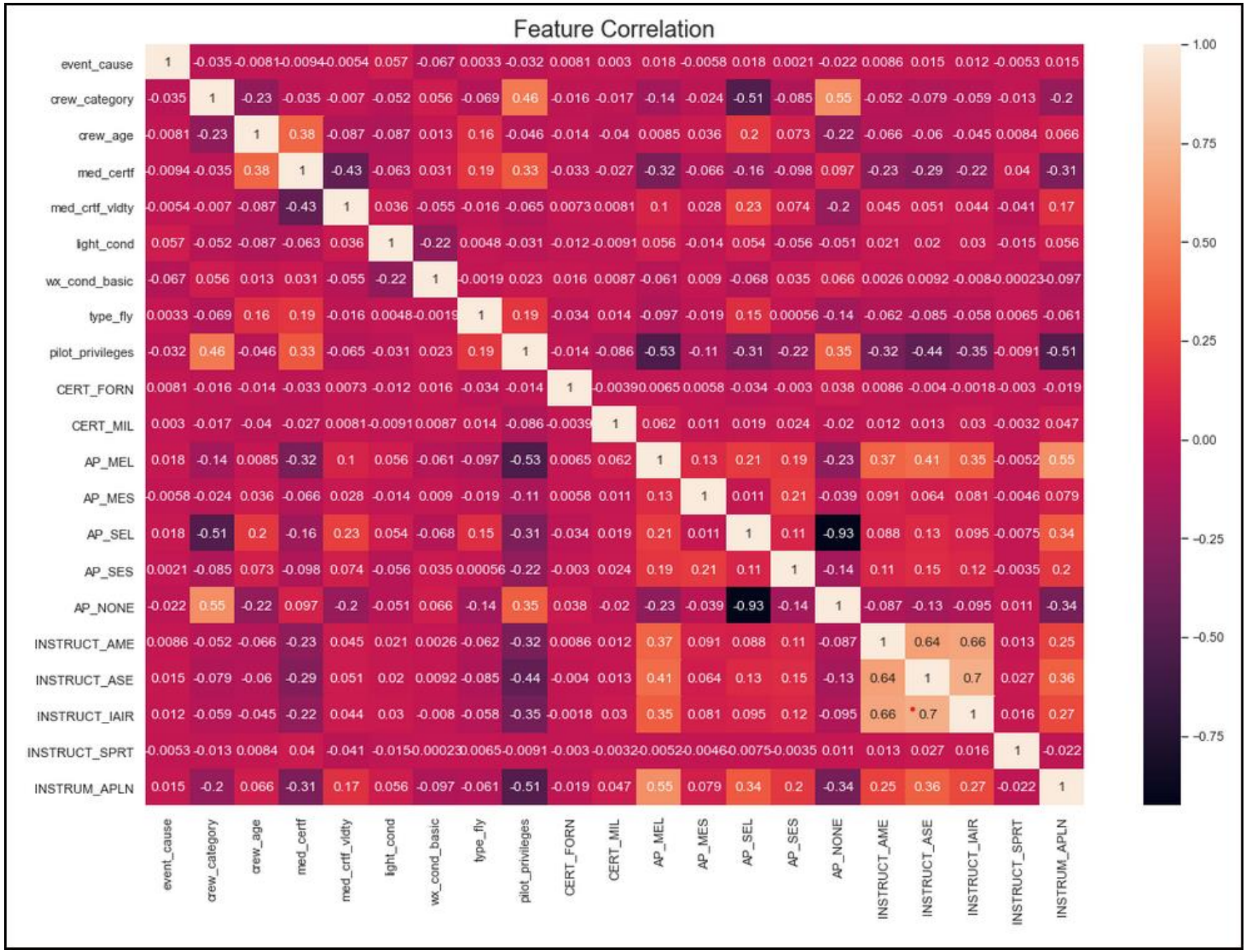
Engineered Feature

Data Wrangling

- Codified data transcribed to plain English to aid in analysis
- Missing Data
  - Age imputed based on overall pilot total hour – buckets
  - Some flight time imputed to minimum flight time required in category for type of certificate held
- Data Validation
  - Algorithms checked for errors such as incompatible pilot certificates being held at the same time
- Engineered Features
  - Medical Certificate Validity
    - Based on expiration date and date of accident
  - Pilot Privilege
    - Transcribed from One HOT Encode to Single Feature to aid analysis

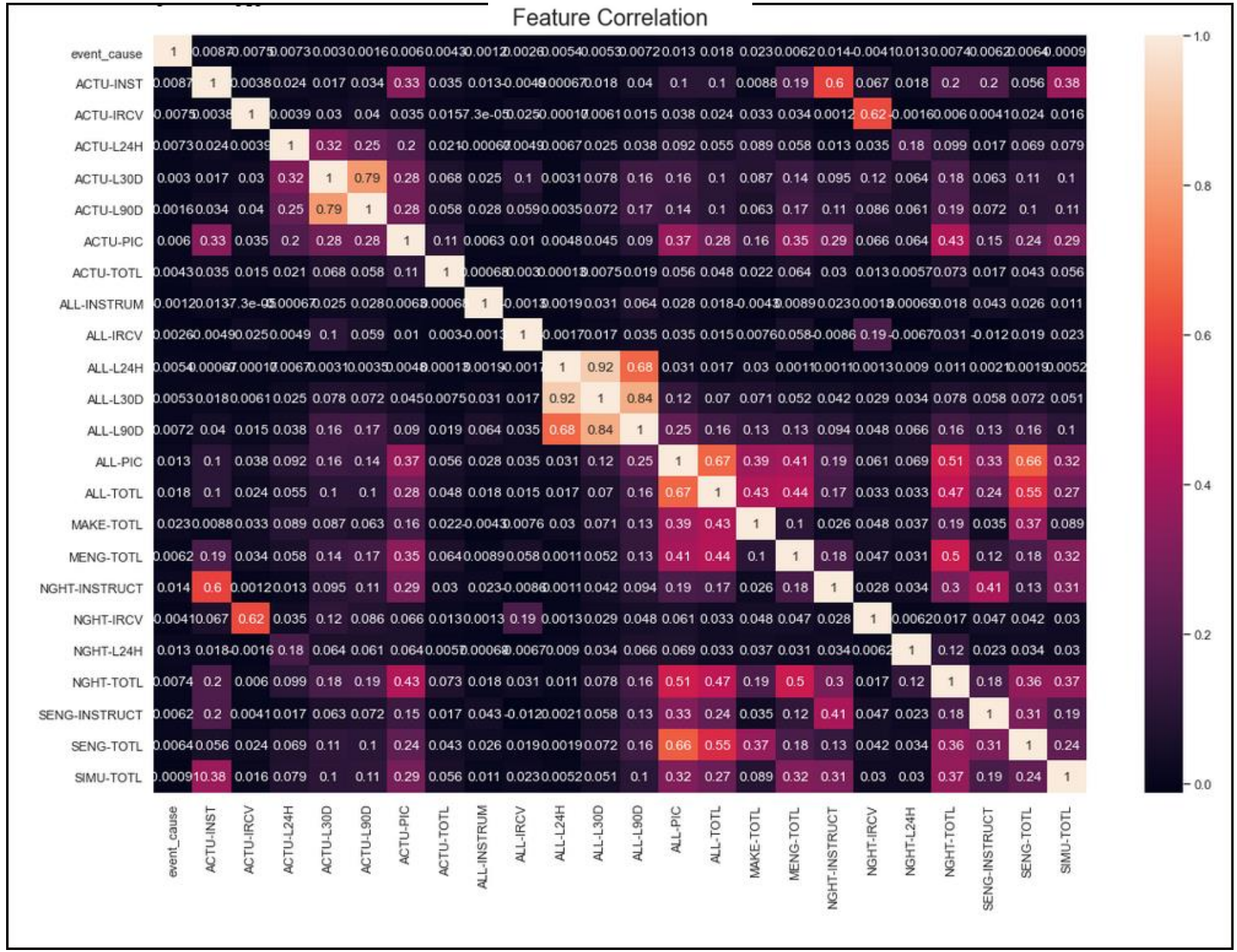
Data Exploration

- Questions
  - What is the overall correlation between features?
  - Is there a correlation between age and accident cause?
  - Is there a correlation between total flight hours and accident cause?
  - Is there a correlation between a pilot’s certificate privilege and accident cause?

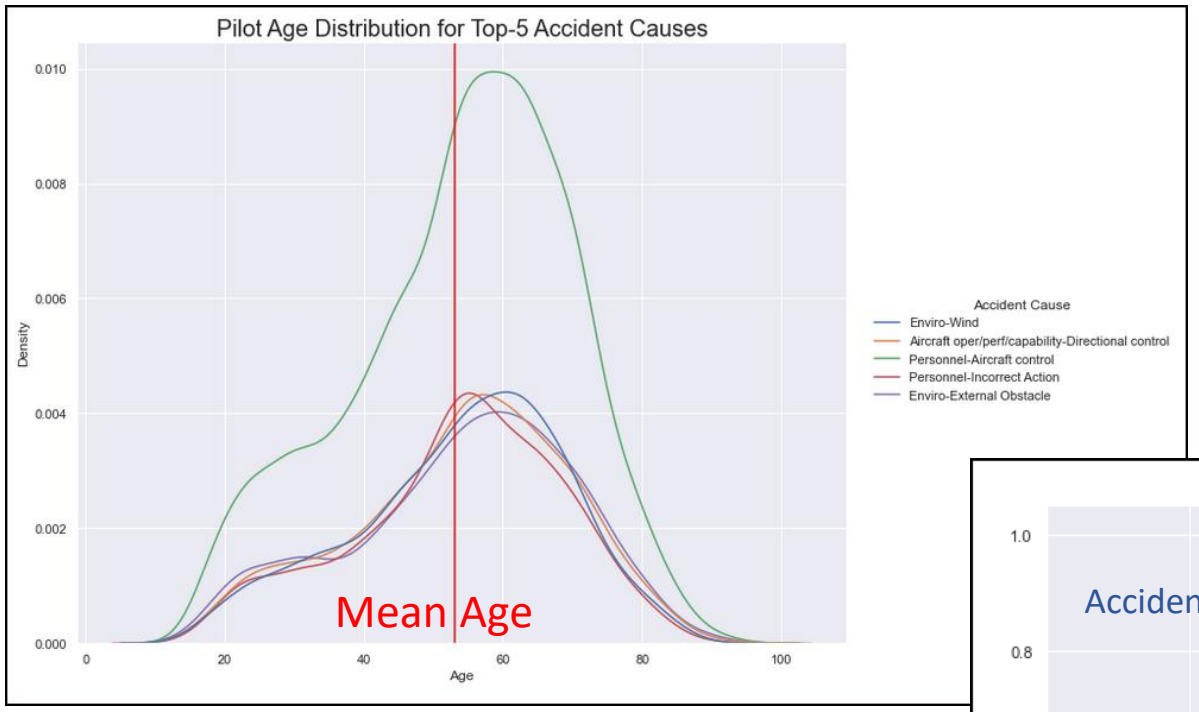


Correlation between Categorical Features and Event Cause

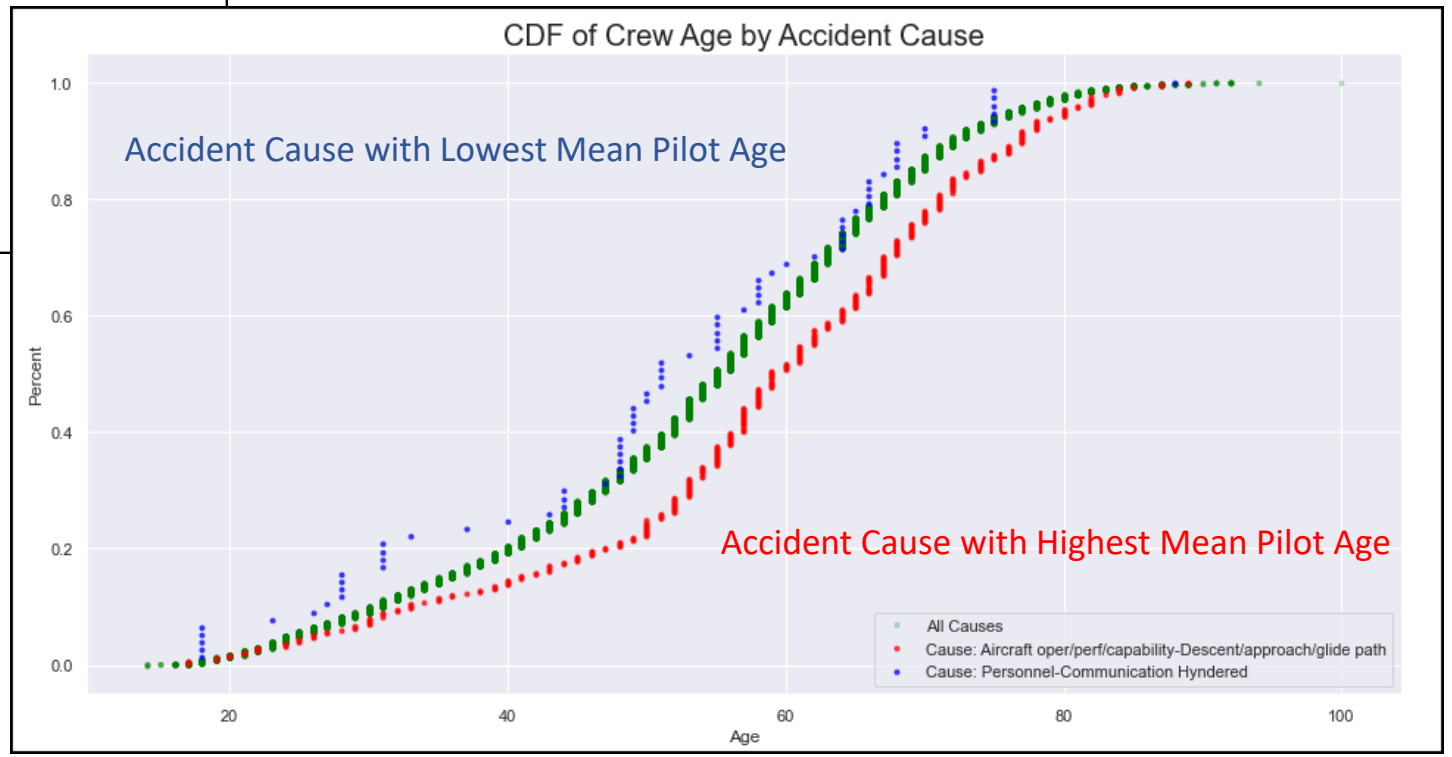
Correlation between Numerical Features and Event Cause



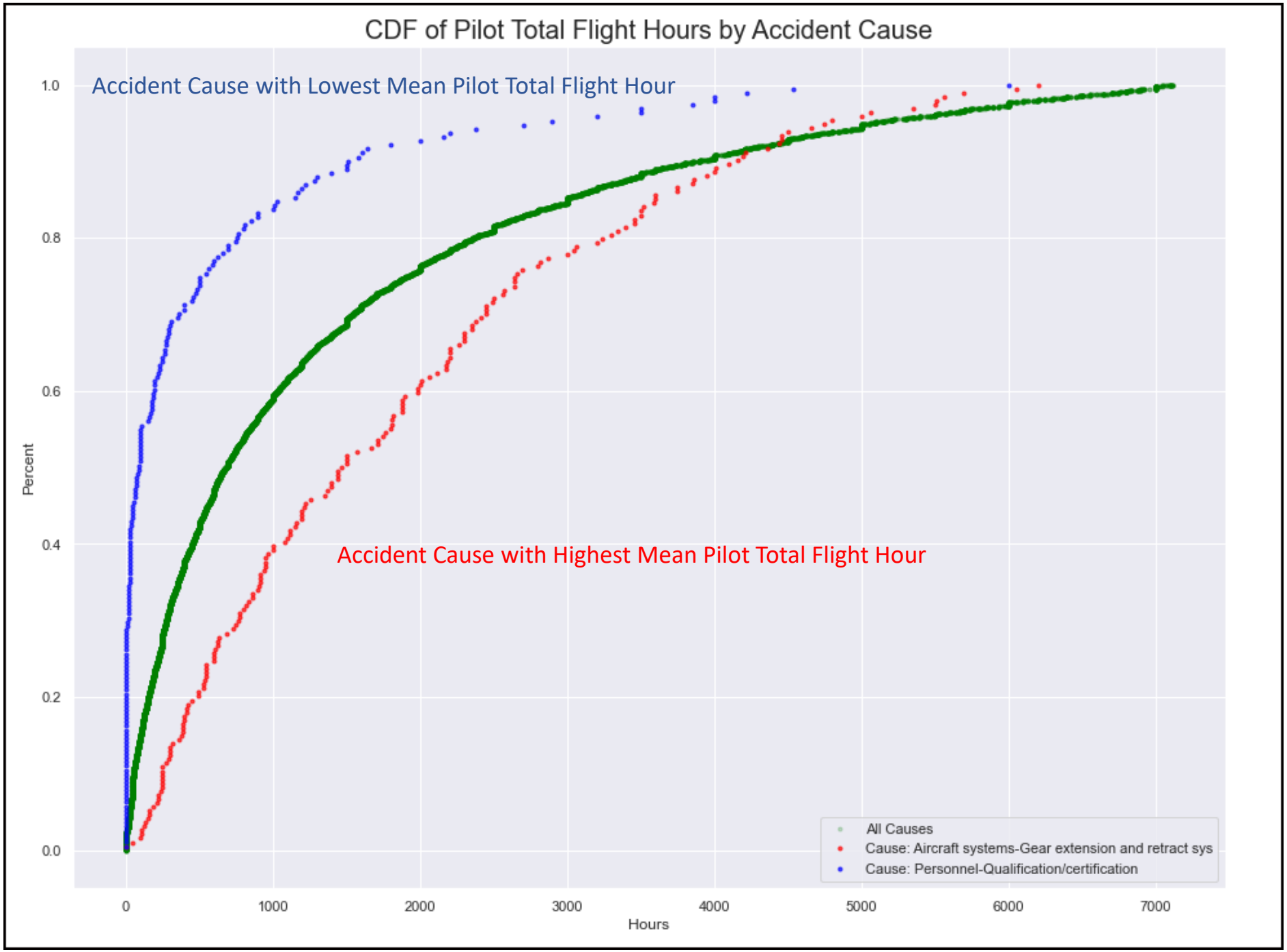
Age vs Accident Cause



No discernable difference in pilot mean age among the top 5 aircraft accident causes



T test showed there is no difference in mean pilot age between pilots involved in these accident causes.



T test showed there is little to no difference in the mean total flight time between pilots involved in these accident causes

Because these accident causes are at the extremes (lowest vs. highest) it follows that there is no statistical difference in pilot's mean total flight time between any accident cause

Does Pilot Certificate Privilege influence Accident Cause

Sample Proportions
Success / Sample Size

Population Proportions

```
Population PRIVATE proportion : 0.57
Population COMMERCIAL proportion : 0.24
Population STUDENT proportion : 0.08
Population NONE proportion : 0.03
Population ATP proportion : 0.04
Population SPRT/REC proportion : 0.03
```

	cause	privilege	sucess	sample_size	hypothesis	z_stat	p_val	significance
0	Personnel-Aircraft control	PRIVATE	2472	4390	0.57	-0.9220	0.3565	No Reject
1	Personnel-Aircraft control	COMMERCIAL	902	4390	0.24	-5.6629	0.0000	Reject
2	Personnel-Aircraft control	STUDENT	574	4390	0.08	9.9744	0.0000	Reject
3	Personnel-Aircraft control	SPRT/REC	142	4390	0.03	0.8787	0.3796	No Reject
4	Personnel-Aircraft control	ATP	158	4390	0.04	-1.4261	0.1538	No Reject
5	Personnel-Aircraft control	NONE	142	4390	0.03	0.8787	0.3796	No Reject
6	Aircraft oper/perf/capability-Directional control	PRIVATE	1105	1873	0.57	1.7566	0.0790	No Reject
7	Aircraft oper/perf/capability-Directional control	COMMERCIAL	345	1873	0.24	-6.2301	0.0000	Reject
8	Aircraft oper/perf/capability-Directional control	STUDENT	268	1873	0.08	7.7971	0.0000	Reject
9	Aircraft oper/perf/capability-Directional control	SPRT/REC	39	1873	0.03	-2.7817	0.0054	Reject

There are 232 instances out of 702 where the sample proportion does not match the population proportion

Pilot Certificate Privilege may help in classification in some cases

Conclusions on Data Analysis

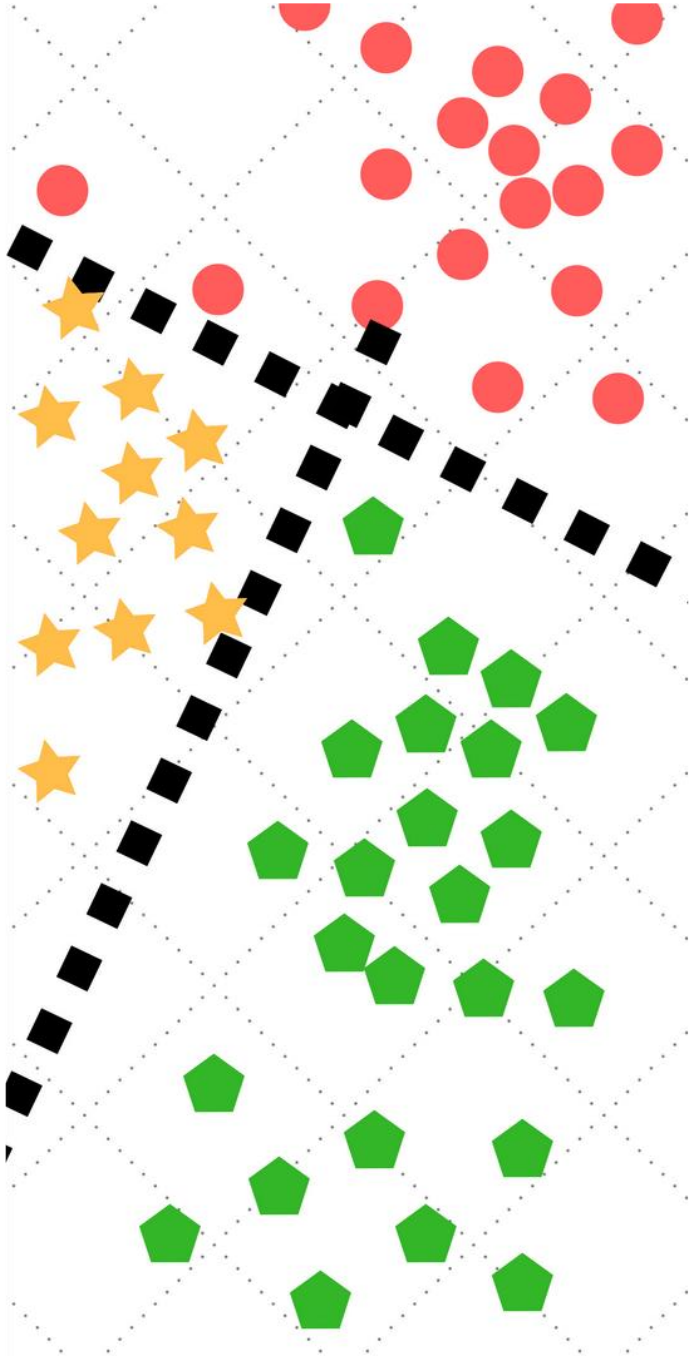
- There is no clear correlation between any one feature and accident cause
- Many numerical features have a significant number of missing values and these cannot be reasonably imputed
- Pilot certificate privilege has the potential to aid in classification but only in some cases

Modeling (Python SKLearn)

Model	n_estimators*	Accuracy
Decision Tree		0.13
Random Forrest	100	0.19
ADABoost Classifier	100	0.37
GradientBoost Classifier	100	0.37
ADABoost Classifier (PCA = 20)**	100	0.39

* determined via GridSearchCV

** PCA 20 selected to account for 80% of variance



Support Vector Machine (PCA = 20) Accuracy 0.39

Parameters	Options (RandomizedSearchCV)			
C	0.1	1	10	100
Gamma	1	0.1	0.001	0.0001
Kernel	Linear			
Decision Function Shape	OVO		OVR	

Modeling

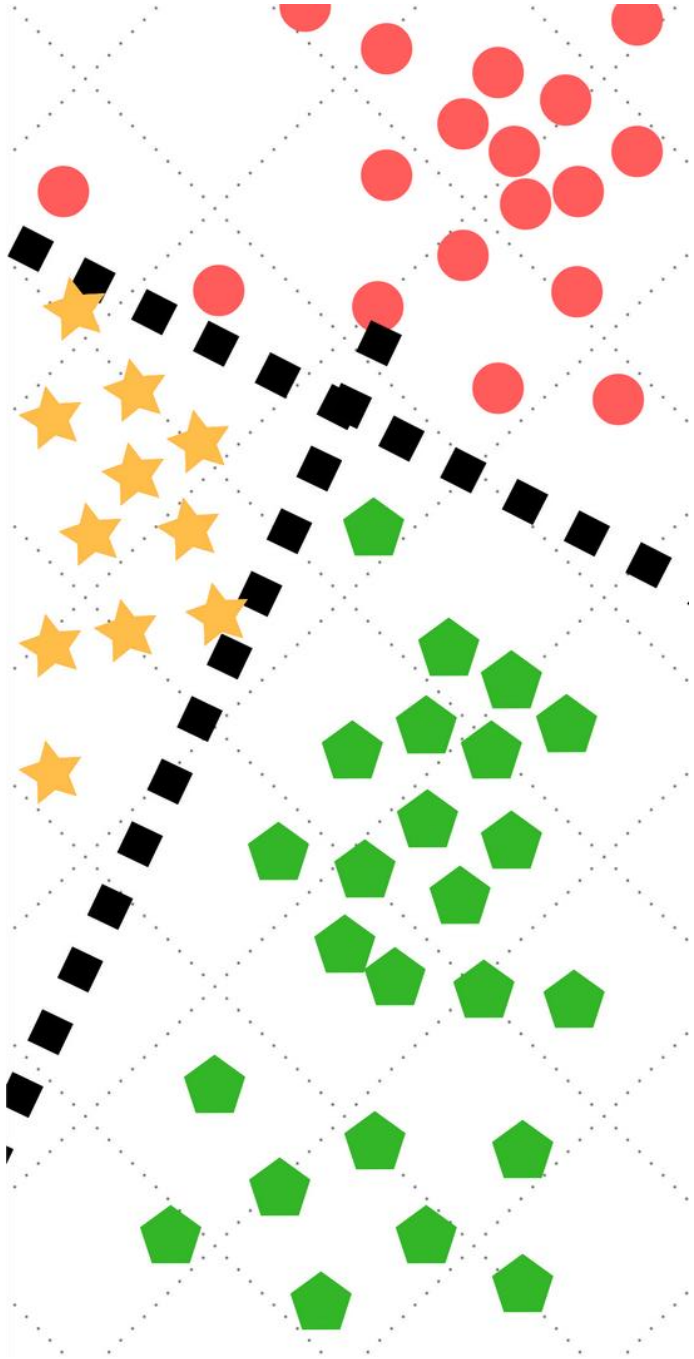
SVM Confusion Matrix					
	Cause 1	Cause 2	Cause 3	Cause 4	Cause 5
Cause 1	0	0	0	446	0
Cause 2	6	0	0	404	0
Cause 3	0	0	0	379	0
Cause 4	0	0	0	1059	0
Cause 5	0	11	0	387	0

ADABoost Classifier Confusion Matrix					
	Cause 1	Cause 2	Cause 3	Cause 4	Cause 5
Cause 1	7	1	3	440	9
Cause 2	6	10	4	378	6
Cause 3	9	2	4	358	6
Cause 4	18	10	5	1010	16
Cause 5	5	11	2	357	12

ADABoost Classifier

Final Model

Accuracy 0.39



Conclusion

- Effort failed to produce a deployable model
  - Issues with missing data (flight hours)
  - Not including aircraft Type/Model in analysis

- Future work
  - Improve NTSB accident data collection
    - Make data collection mandatory for all time-related fields
  - Reattempt analysis including aircraft Type/Model
    - Time commitment to clean data set will be considerable