Text Analysis of Aircraft Accident Reports

Introduction and Business Understanding:

The National Transportation Safety Board (NTSB) is a U.S. government investigative agency responsible for civil transportation accident investigations. In the case of aircraft accidents, the NTSB compiles reports to determine their cause. The structured data of the reports can be made into a crosstab to understand specific accidents. For instance, this method reveals that 45% of instructional flight accidents occur during landing (see Figure 1). Interestingly, the majority of accidents involving aircraft used for personal purposes also occur during the landing phase, accounting for 27%. However, the percentage of accidents related to landing for instructional sessions is nearly double that for personal sessions. One challenge is then figuring out more context that structured data alone cannot provide. Therefore, the NTSB can utilize their unstructured data, such as narratives, through text analysis to acquire an even greater level of detail about the accidents.

Data Understanding:

The Aircraft Incidents.jmp file contains 1,906 accident reports. The data set uses 27 variables, but only 3 variables will be of interest. These variables include a qualitative variable and two categorical variables, which are respectively labeled as “Narrative Cause,” “Purpose of Flight,” and “Broad Phase of Flight.” The “Purpose of Flight” variable was recoded to account for missing values. The empty cells were replaced with a “MISSING” label that was then defined as a missing value code. Tabulating the two categorical variables by setting “Purpose of Flight” as a row then the statistic “Column %” as a column shows the percent of accidents for each category as it relates to all reported accidents. Next, “Broad Phase of Flight” was added as a column with “Column %” being replaced by “Row %” to show the percent of accidents for different purposed flights at different flight phases, which Figure 1 depicts. From here the subgroups of interest, “Instructional” and “Landing,” continue with further analysis.

Analysis:

Text Frequency and Word Cloud:

The narrative causes were analyzed using Text Explorer to obtain a Pareto chart of the most frequent terms and phrases, known as a term and phrase list. Standard settings were used, including no stemming and regex tokenizing. A word cloud was then created. Thereafter, a local filter was added for “Purpose of Flight” and “Broad Phase of Flight.” The selection of the subgroups “Instructional” and “Landing” led to modifications in both the original term and phrase list and the word cloud. The results are shown in Figure 2 and Figure 3, respectively. The term “landing” was seen the most with a count of 118 followed by the terms “pilot’s,” “failure,” and “student.” These terms were also the largest in the corresponding word cloud, indicating their prominence within the narratives. Phrases with the highest counts were “directional control,” “student pilot’s,” and “failure to maintain.”

The aforementioned process was again utilized to acquire the same type of results with only one local filter rather than two, that being “Purpose of Flight” with the subgroup “Instructional” selected. Figure 4 is the resulting term and phrase list. The most frequent term is “landing” at a count of 166; following are the terms “failure,” “pilot’s,” “flight,” and “student.” The most frequent phrases accounted for are "failure to maintain," "student pilot's," and "pilot's failure." Similar to the previous word cloud, this word cloud shows the most frequent terms found in the corresponding term and phrase list, as seen in Figure 5.

Singular Value Decomposition (SVD):

Singular Value Decomposition (SVD) plots were created for the first two singular values for the subgroup of “landing” in “Instructional” flights using latent semantic analysis. One is a term plot that captures connections among different terms with similar meanings, as illustrated in Figure 6. The other is an SVD document plot, which features vectors representing documents with similar topics, as seen in Figure 8. The points of the term plot represent single words with the position of a term on the plot showing its association with different themes or concepts within a single dot within the document. Points that are grouped are usually used together in similar contexts. Similarly, the proximity of points to each other on the SVD document plot reveals that the documents are similar with clusters of points indicating groups of documents with similar themes.

Latent semantic analysis (LSA) also provides a list of singular values that help explain the amount of variance captured by each singular vector, as seen in Figure 9. For instance, the first five singular vectors account for approximately 25% of the variation. This means that 25% of the differences and variations in the reports can be understood and summarized by the topics that the LSA model identified.

Topic analysis of the vectors, particularly the term vectors, and their relationships reveals a topic list shown in Figure 7. This list organizes terms in topics/groups based on their meaningful relationships with one another that can then be used to gain a deeper insight into the thematic framework. For example, “Topic 1” lists a group of words that seem to focus on accidents where the aircraft was submerged or partially submerged as a result of landing. The use of terms such as "submerged," "partially," and "resulting" points to scenarios in which aircraft have been immersed in water or encountered similar conditions. The words "inspection" and "procedures" signal an emphasis on scrutinizing the aircraft after an incident, as well as the landing protocols that were either adhered to or neglected. "Due" and "total" are likely related to the root causes of the mishaps and the overall severity or scope of the damage. In essence, this subject probably pertains to instructional flight mishaps that involved aircraft becoming submerged, often attributable to procedural discrepancies or various other causes. In contrast, “Topic 2” appears to be centered around the decision-making process and environmental factors influencing landing accidents. Terms like "area," "selected," and "trees" suggest that the physical environment of the landing area, such as the presence of trees or terrain, is significant. "Decision," "delay," and "proper" point toward the decision-making process of the pilot, possibly indicating delayed or improper decisions impacting the landing. "Approach" and "landing" are directly related to the final phase of flight, and "factors" could imply various contributing elements to the accidents. This topic likely covers incidents where environmental challenges and decision-

making issues during the approach and landing phase led to accidents in instructional flights.

Conclusion:

Text analysis and quantifying texts are crucial in deriving actionable insights from data, such as flight accident reports. Text analysis provides qualitative insights, uncovering themes and patterns, while quantification allows for discoveries of connections and predictions not immediately apparent from purely qualitative analysis. The use of both methods for aircraft accident reports has produced profound insights for the NTSB.

Cross-tabulation in interpreting structured data reveals a significant percentage of flight accidents for instructional sessions occur during the aircraft's landing, thus requiring thorough exploration for context. The strength of this method is that it provides clear, quantifiable data that can guide safety improvements. However, its weakness is its limited depth beyond clear contexts.

Moreover, word clouds and text frequency analysis from the narratives offer an insightful view of common terms associated with these incidents, such as "landing," "pilot's," and "failure." These tools are less effective in delineating the complex dynamics of the multiple interrelated factors causing accidents. The SVD and LSA models capture deeper thematic structures and variabilities in the reports. SVD term and document plots illustrate associations between terms and document similarities, providing a conceptual map. LSA's strength is its ability to distill vast amounts of text into interpretable topic lists for better thematic pattern understanding. However, a notable limitation is that the first five singular vectors only account for 25% of the variation, leaving a significant portion of the data's complexity unsummarized.

Topic analysis offers thematic groupings that provide a deeper understanding of accident causality and characteristics. Topics like aircraft submersion during landing phases and decision-making in environmental challenges highlight nuances in instructional flight accidents. However, this method relies on the analyst's interpretative skills to derive meaningful conclusions, presenting a subjective challenge despite its ability to identify and organize latent themes.

In conclusion, applying text analysis methods to aircraft accident reports makes a strong case for their continued use and development. Each method has its strengths and weaknesses, but together they form a comprehensive toolkit for the NTSB, providing both broad trends and detailed insights. These findings reinforce the value of a multifaceted analytical approach, enabling the NTSB to enhance flight safety through informed, data-driven decisions.

Figure 1: Crosstab Purpose of Flight versus Broad Phase of Flight

Row %

Purpose of Flight	Broad Phase of Flight											
	APPROACH	CLIMB	CRUISE	DESCENT	GO-AROUND	LANDING	MANEUVERING	OTHER	STANDING	TAKEOFF	TAXI	UNKNOWN
Aerial Application	2.50%	0.00%	6.25%	0.00%	1.25%	3.75%	60.00%	0.00%	2.50%	21.25%	1.25%	1.25%
Aerial Observation	0.00%	0.00%	18.75%	0.00%	0.00%	31.25%	25.00%	0.00%	6.25%	12.50%	0.00%	6.25%
Air Drop	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Air Race/Show	0.00%	0.00%	20.00%	0.00%	0.00%	20.00%	40.00%	0.00%	0.00%	0.00%	20.00%	0.00%
Business	17.81%	1.37%	13.70%	4.11%	6.85%	15.07%	9.59%	0.00%	5.48%	23.29%	1.37%	1.37%
Executive/Corporate	12.50%	0.00%	12.50%	0.00%	12.50%	50.00%	12.50%	0.00%	0.00%	0.00%	0.00%	0.00%
Ferry	7.69%	0.00%	38.46%	0.00%	0.00%	0.00%	7.69%	0.00%	7.69%	30.77%	0.00%	7.69%
Flight Test	22.73%	4.55%	9.09%	4.55%	4.55%	4.55%	18.18%	4.55%	0.00%	22.73%	4.55%	0.00%
Instructional	8.46%	1.47%	5.88%	1.84%	4.78%	45.22%	10.66%	0.00%	1.10%	16.91%	3.68%	0.00%
Other Work Use	6.98%	2.33%	9.30%	4.65%	0.00%	13.95%	39.53%	4.65%	2.33%	13.95%	0.00%	2.33%
Personal	9.34%	3.75%	17.03%	3.02%	1.56%	27.29%	11.17%	0.18%	1.56%	20.97%	3.11%	1.01%
Positioning	17.19%	3.13%	18.75%	4.69%	3.13%	20.31%	6.25%	0.00%	4.69%	15.63%	6.25%	0.00%
Public Use	2.50%	0.00%	20.00%	2.50%	2.50%	25.00%	35.00%	0.00%	2.50%	7.50%	2.50%	0.00%
Skydiving	11.11%	0.00%	0.00%	66.67%	0.00%	0.00%	11.11%	0.00%	0.00%	11.11%	0.00%	0.00%
Unknown	11.11%	11.11%	11.11%	0.00%	0.00%	22.22%	0.00%	0.00%	0.00%	22.22%	0.00%	22.22%

Figure 2: Term & Phrase List for Instructional & Landing

Term	Count
landing	118
pilot's	73
failure	68
student	58
flight	56
control	53
maintain	44
inadequate	39
directional	38
resulted	33
factor	29
improper	29
instructor's	28
factors	27
airplane	26
pilot	26
supervision	22
crosswind	21
flare	21
contributing	20
accident	19
roll	19
hard	18
action	16
remedial	16

Phrase	Count	N
directional control	38	2
student pilot's	38	2
failure to maintain	33	3
maintain directional control	31	3
maintain directional	31	2
pilot's failure	29	2
flight instructor's	27	2
failure to maintain directional	26	4
pilot's failure to maintain	21	4
inadequate supervision	19	2
hard landing	18	2
landing roll	17	2
remedial action	16	2
student pilot's failure	15	3
flight instructor's inadequate	14	3
instructor's inadequate	14	2
landing gear	14	2
flight instructor's inadequate supervision	12	4
resulted in a hard	12	4
instructor's inadequate supervision	12	3
contributing factor	12	2
bounced landing	11	2
pilot's inadequate	11	2
pilot's lack	11	2
control during the landing	10	4
control during landing	10	3

Figure 3: Word Cloud with Landing and Instructions as Local Filter

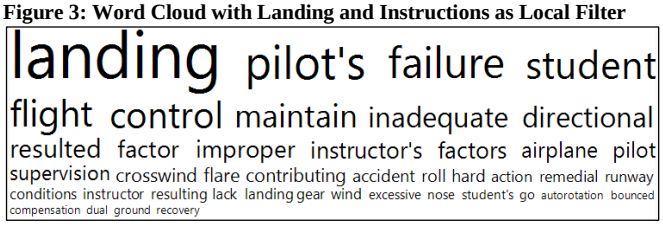

Figure 4: Term & Phrase List for Instructional excluding Landing

Term	Count
landing	166
failure	162
pilot's	125
flight	124
student	102
control	94
maintain	90
factor	77
inadequate	67
resulted	66
pilot	64
engine	55
airplane	54
factors	53
instructor's	53
loss	52
improper	50

Phrase	Count	N
failure to maintain	67	3
student pilot's	59	2
pilot's failure	56	2
directional control	49	2
flight instructor's	47	2
pilot's failure to maintain	39	4
maintain directional control	37	3
maintain directional	37	2
inadequate supervision	35	2
failure to maintain directional	31	4
contributing factor	30	2
engine power	28	2
loss of engine	27	3
loss of engine power	26	4
remedial action	26	2
student pilot's failure	25	3
hard landing	22	2

Figure 5: Word Cloud Instruction & No Landing as Local Filter

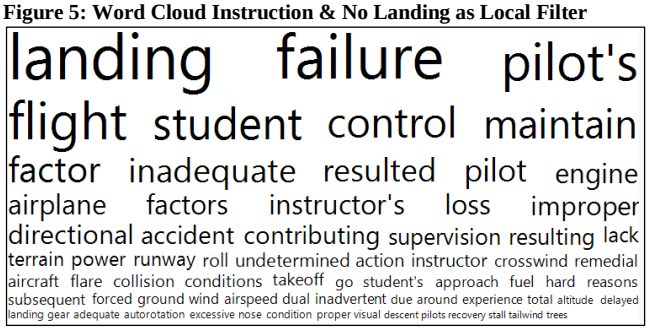

Figure 6: SVD Plot for Terms

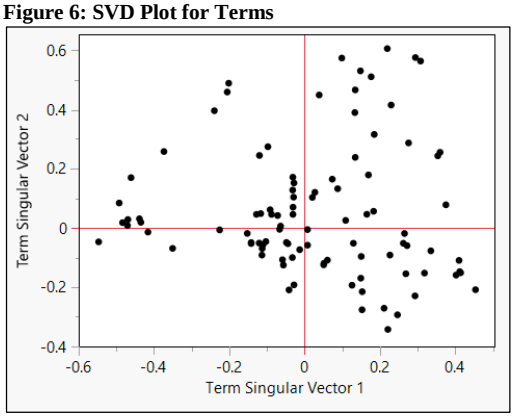

Figure 7: Topics

Topic 1		Topic 2		Topic 3		Topic 4		Topic 5	
Term	Loading	Term	Loading	Term	Loading	Term	Loading	Term	Loading
submerged	0.96150	area	0.76131	flight	0.79165	descent	0.8241	rotor	0.75327
inspection	0.94216	selected	0.75475	supervision	0.71252	pilots	0.7248	rpm	0.71175
partially	0.92436	trees	0.72141	inadequate	0.59324	rate	0.6890	resulted	0.63875
procedures	0.87079	decision	0.69350	instructor's	0.58700	maneuver	0.6554	autorotation	0.61919
airplane	0.79305	delay	0.50447	student's	0.58430	proper	0.6136	practice	0.50058
due	0.53091	included	0.47641	controls	0.50708	delay	0.4385	hard	0.46138
total	0.31089	subsequent	0.46046	use	0.48953	included	0.4031	ground	0.36867
resulting	0.29836	approach	0.44322	dual	0.41827	excessive	0.3737	main	0.36446
		proper	0.42171	instructor	0.30839	airspeed	0.3315	flare	0.30626
		factors	0.40071	improper	0.30682	controls	0.2976	touchdown	0.29259
		landing	0.38489	around	0.28181	control	-0.2851		

Topic 6		Topic 7		Topic 8		Topic 9		Topic 10	
Term	Loading	Term	Loading	Term	Loading	Term	Loading	Term	Loading
remedial	0.84954	recovery	0.6925	lack	0.7755	crosswind	0.6788	instruction	0.7047
action	0.84954	bounced	0.6530	experience	0.7656	compensation	0.6090	receiving	0.6729
delayed	0.70014	improper	0.5886	total	0.6760	conditions	0.5899	control	-0.4856
cfi's	0.52999	nose	0.5379	pilot's	0.6755	gusty	0.5462	pilot	0.4799
control	0.36542	landing	0.5024	reasons	-0.3902	go	-0.4945	maintain	-0.4476
rollout	0.33620	collapse	0.4976	undetermined	-0.3847	around	-0.4675	directional	-0.4380
maintain	0.29839	resulting	0.3513	adequate	0.3828	inadequate	0.4327	certified	0.3733
loss	0.27818	due	0.3289	student	0.3538	wind	0.3712	failure	-0.3672
directional	0.24835	flare	0.3234	inadvertent	0.3463	subsequent	0.3457	landing gear	0.3394
certified	0.24714	accident	-0.3049	collapse	-0.2914	aircraft	0.3347	supervision	0.3236
		landing gear	0.2959			condition	0.3095	contributing	0.2953
		hard	0.2900			roll	0.2886	approach	0.2798
		collision	-0.2898						

Figure 8: SVD Plot for Documents

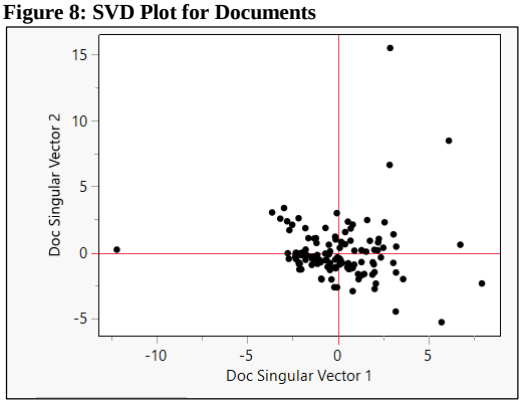

Figure 9: Singular Values

Number	Singular Value	Eigenvalue	Percent	Cum Percent
1	2.3486	5.5157	5.6283	5.6283
2	2.2518	5.0708	5.1743	10.8026
3	2.2156	4.9088	5.0089	15.8115
4	2.1258	4.5191	4.6113	20.4229
5	2.0158	4.0635	4.1465	24.5693
6	1.8991	3.6067	3.6804	28.2497