Aviation Data Report

Soham Parikh

This report describes my findings and analysis on the Aviation Accidents Dataset provided by the National Transportation Safety Board. All the findings mentioned in this report can be reproduced using the IPython Notebook provided in the GitHub repository.

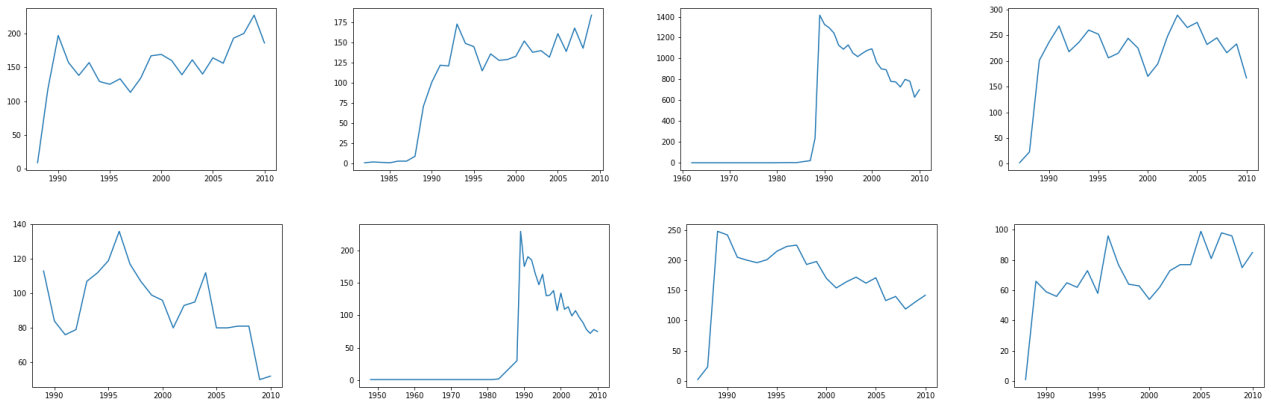
Dataset: In this dataset, each accident, marked by a unique ID, is provided with different structured attributes like the date of the accident, geographic coordinates, phase of flight, make of the aircraft, airport, etc. along with a narrative regarding the accident and a description of the probable cause.

Aim The main aim of my data exploration and analysis was to try to determine the broad reasons and causes for the accidents from the unstructured/unlabeled descriptions of the narrative and the cause of accident.

Clustering Causes: Using words from probable causes to group the events into clusters reveals traits and the different kinds of reasons and causes for the accident. This is better understood from the different clusters given below in the figures. We can see that each cluster contains words which indicate the cause of the failure *e.g.,* accidents belonging to the first cluster probably happened because the pilot lost directional control, accidents belonging to the seventh cluster (row 2, column 3) probably happened because of fuel exhaustion in the vehicle.

Moreover, on analyzing the collocations from each of these clusters, we obtain more information about the type of cause *e.g.,* in the probable cause description of the first cluster, we frequently observe the occurrence of the phrase "failure to maintain directional control".

Yearwise Trends in Causes This section indicates the yearwise trends for each of the clusters identified previously. Each plot shows the number of accidents vs the year for the corresponding cluster. This indicates reasons which have become more(less) frequent in the recent years and helps the aviation industry gain more insights about these accidents.



Use Case: We observe that the reasons for the accidents, corresponding to each cluster, are repeated frequently. Identifying the reasons from the probable causes can help foresee and prevent these accidents in the future. This information can be used by the aviation industry as a whole, in terms of what kind of improvements/changes are needed to curb the number of accidents. Moreover, this can help in organizing the database of events based on the cause of the accident by adding a new attribute along with the other structured attributes. Automated extraction of this attribute would aid the annotation of the events as well. Further, treating these as classes, we can use textual features from narratives to classify events with missing probable cause fields (30% of the dataset) into one of these clusters.