Data Cleaning

The data-set we have chosen is the “Airplane Crashes since 1908” dataset. We got the dataset from Kaggle.com where it is provided by Sauro Grandi. The dataset is originally made available for public use on OpenData by Socrata and is available [here](https://opendata.socrata.com/Government/Airplane-Crashes-and-Fatalities-Since-1908/q2te-8cvq). The data-set contains 5268 records about the details of the airplane crashes that have occurred since the year 1908. It contains fields like Date of Crash, Time of Crash, Location, Operator of the airplane, the type or the model of the aircraft, the no. of people aboard, number of fatalities, the fatalities on the ground and the summary of the reason of the crash.

APA citation:

Sauro Grandi (September, 2016). Airplane Crashes Since 1908. Retrieved from <https://www.kaggle.com/saurograndi/airplane-crashes-since-1908>

The data-set is licensed under the [Open Database License](http://opendatacommons.org/licenses/odbl/1.0/). To summarize, a user is free to copy, distribute, use and produce works from the database as well as transform, modify and build upon the database if and only if the following conditions are taken care of:

* Attribution of any public use of the database (or works produced from it) in a manner specified by Open Database License. For any use/redistribution of the database (or works produced from it), the license of the database or any notices on it must be made clear to others.
* If a user is publicly making use of an adapted version of this database including the works produced from an adapted database, the user must also offer the adapted database under the Open Database License.
* If a user redistributes a database (or an adapted version of it), technological measures (such as Digital Rights Management) may be used to restrict the work if a version without such measures is also redistributed.

The information that helps us interpret and understand the data is basically just the column names which are self-explanatory. Such column names include:

* Date: The date on which the airplane crash happened.
* Time: The time when the crash happened.
* Location: Either the city name or other types of location (E.g. Over the North Sea) where the crash occurred.
* Operator: The airline company making use of the airplane. This could be either commercial carriers like Air Cameroon or the Army or the Navy of a country.
* Flight#: The flight number used to identify the flight.
* Route: The route on which the flight used to operate. For instance, New Orleans - Chicago.
* Type: The model or the type of the airplane. E.g. Boeing 247
* Registration: The registration # used to uniquely identify the aircraft. E.g. G-AADN
* Aboard: The number of people aboard the aircraft at the time of the crash.
* Fatalities: The number of people who died because of the airplane crash.
* Summary: A one or two sentence summaries describing the exact reason of the crash. E.g. “Shot down in flames by the British 39th Home Defense Squadron.”

There are 2 columns that we are unsure about.

* cn\ln : We have sent a mail to the creator of this dataset to confirm what the column denotes.
* Ground: We assume that this is the number of fatalities on the ground due to the air-crash but we have still sent a mail to the creator of this dataset to clarify whether our understanding is correct or not.

Our dataset contains several issues in different attributes. Listed below are the issues and the attributes associated with it.

1. Missing Values:

The following attributes have missing values in them.

* Time
* Operator
* Location
* Summary

1. Unstandardized content:

* Time: The time is represented in improper format in some fields. For example, c: 1:00 instead of 1:00 and 7’09 instead of 07:09.
* Summary: The summary attribute has a lot of text and excess data making it difficult to categorize the reasons for the air crash. For example, after developing engine trouble the pilot tried to parachute out of the plane but the parachute got caught on the plane and the pilot was dragged to his death. This field contains unwanted information and we will need to remove such information.

At some places, there is insufficient data provided due to which the reason for the crash cannot be understood. For example, Crashed on approach onto a golf course. This summary doesn’t explain the reason for the crash. We should categorize such fields into some understandable format.

1. Insufficient metadata:

Some attributes have unclear naming and explanation. For example, the attribute cn/ln is not self-explanatory and we are unable to understand what it represents.

We need to address the above-mentioned issues to extract useful information from our data. For this, we intend to make the following changes to our data:

* We will be ignoring the missing values in the Time of crash, Operator and Location columns. The percentage of missing data in these fields is low. In case we are interested in these details for any specific air crash/crashes, we will be researching more on those individual items to get these details. As far as our research questions are concerned, we would not be needing such details for air crashes at this point of time.
* There are very few instances where the time format used is different from the one used in majority of the entries. We would be identifying these issues manually, and making the changes in the dataset. For instance, there is an entry in which the time is entered as “6’30” instead of “6:30”. We have chosen to do this task manually since there are a very few entries with such irregularities and they can be identified easily.
* As pointed out before, the data regarding the reason for crashes in the summary column is not standardized. To proceed with our analysis, it is very important that we categorize this data into well-defined categories that reflect the reason for the crash. Categorizing this data into nominal variables would help us to more easily identify the reasons behind majority of the air crashes. The categories we have decided to use are given in a later section of this document. If the data is missing for this column, we will be categorizing such instances as “Reason Unknown”.
* We have decided to ignore the column “cn/ln” and “Ground” for our analysis, since we are not exactly sure what it means. As mentioned above, we have mailed the creator of the dataset regarding this and are awaiting a reply. However, we would not be needing this column for analysis regarding any of our research questions, so we will be ignoring this column for now.
* It is worth noting that in this dataset, it is very tough to fill in the missing values by guessing etc., as most of them contain data which is very specific, and hence cannot be filled in based on guesswork considering the other data related to the air crash. Since the percentage of data missing in the columns we are interested is very low, we have decided it is best to ignore these entries rather than filling them up.

Below is a step by step representation of the summary of the data cleaning process we have performed on our dataset. It is very important to note that the cleaning process is done majorly keeping in mind our research questions:

* Categorize the data in the summary column (reason for crash):

The intent of this categorization is to first understand the main reason of the crash by reading the contents of the ‘Summary’ column and then figure out the category it falls under. Following are the nominal variables(categories) we have assigned for the reason of the crashes, after going through the dataset:

-        Bad Weather

-        Pilot Error

-        Technical Failure

-        Shot down

-        Collision

-        Bombed/Hijacked

- Unknown

* Created a new column with the name “Reason for the Crash”, which will contain the nominal variable values mentioned in the previous step.
* Categorized records based on the reason for the crash and assigned a nominal variable to each such record. This will significantly help us in our analysis and answering our research questions including finding out the main reason for airplane crashes in each time-period.
* Changed the inconsistent entries in the ‘Time’ column to the standard format to get rid of inconsistencies in the results at a later stage**.**
* Created a new column for the time slot i.e. Day (1) or Night (0). For this we ignored the empty time fields. So, we got 3050 records for this subset.

R-script:

crash <- c("crash","engine",”failure”)

do <- c("Technical")

> categorize <- function(d, searchString, category) {

+     d$Reason <- "unknown"

+     for(i in seq(1, length(searchString), 1)) {

+         list <- grep(searchString[i], d$Summary, ignore.case=TRUE)

+         if (length(list) > 0) {

+             for(j in seq(1, length(list), 1)) {

+                 d$Reason[list[j]] <- category[1]

+             }

+         }

+     }

+     d

+ }

> cleanData <- categorize(d, crash, do)

> View(cleanData)