**Analyzing Historical Plane Crash Data with Python**

Mizanur Rahman

Data Science in Python

In this project, I hope to analyze the frequency of plane crashes and get an estimate about the likelihood of these accidents occurring. To do this, I will use a csv dataset of all recorded airplane accidents since 1908, which I obtained from Kaggle.com. This public dataset, Historical Plane Crashes, can be found at <https://www.kaggle.com/nguyenhoc/plane-crash#planecrashinfo_20181121001952.csv>. The author of the dataset gathered the information from <http://www.planecrashinfo.com> and published the information in December of 2018.

The dataset includes 5783 rows and 13 columns. The variables included in this data are the date, time, location, operator/airline, flight number, route, aircraft type, aircraft registration, serial number of the aircraft, total number of passengers aboard, total number of fatalities of passengers onboard, total number of fatalities of people on the ground, and a short description of the accident that took place and its cause (if known). A table of these variables with their variable name and description is below:

| Variable | Description |
| --- | --- |
| date | Date of accident, in the format Month DD, YYYY |
| time | Local time in 24hr format |
| location | Location information |
| operator/airline | Airline or operator of the aircraft |
| flight\_no | Flight number assigned by operator |
| route | Complete or partial route flown prior to accident |
| ac\_type | Aircraft type |
| registration | ICAO registration of the aircraft |
| cn\_ln | Construction or serial number |
| aboard | Total aboard, including passengers and crew |
| fatalities | Total fatalities aboard, including passengers and crew |
| ground | Total killed on the ground |
| Summary | Brief description and cause of the accident, if known |

In this data, I would be interested in looking at the following variables: date, time, location, operator/airline, ac\_type, aboard, fatalities, ground, and summary. I would choose to disregard the variables flight\_no, route, registration, and cn\_ln. I will analyze the plane crash data with Python 3 through Jupyter Notebook. I will represent our findings in python using pandas to read our csv data and matplotlib to create informative graphics which I will use to help visualize the results.

Below is a list of possible questions which I may want to use the data to answer:

1. How many crashes occurred each year?
2. What operator/airline had the most accidents?
3. How many fatalities due to plane crashes were there each year?
4. What percent of passengers died as a result of a plane crash?
5. What is the most frequent location of plane crashes?
6. What is the most frequent time of day that plane crashes occur?
7. Which type of aircraft is involved in the most crashes?
8. Which words come up the most in the description of the recorded accidents?

Considering the recent focus on aircraft safety standards and safe flight practices, I aim to analyze this data and make observations about the reasons that airplane crashes may occur. Our goal is to use our analysis to make recommendations about when and where it is safest to fly, given the data, and which airlines have the fewest records of accidents. Another topic I hope to learn more about is how likely a given person is to survive a plane crash. It would also be interesting to know what the most common causes of plane crashes are (i.e. weather, malfunction, take-off difficulties, hijacking, etc.)