Final Project

Kaiwei Qian

2019/4/5

Data

The data used in this study is scraped from the Bureau of Aircraft Accidents Archives (http://www.baaa-acro.com/crash-archives) for the past 10 years (2009-2019). The Bureau of Aircraft Accidents Archives (B3A) was established in Geneva in 1990 for the purpose to deal with all information related to aviation accidentology.

Scraping involves two steps.

First, **read_html()** and **html_table()** is used to get the table data, whose row names are Date, A/C Type, Location, Fatalities and Registration. If you visit their website, you'll find that there is a plus sign at the end of each row of the table. Clicking on it leads you to the page that briefly summarizes the air crash. It provides us with the information such as Flight Phase and Circumstances. So, we also keep the hyperlinks to this page.

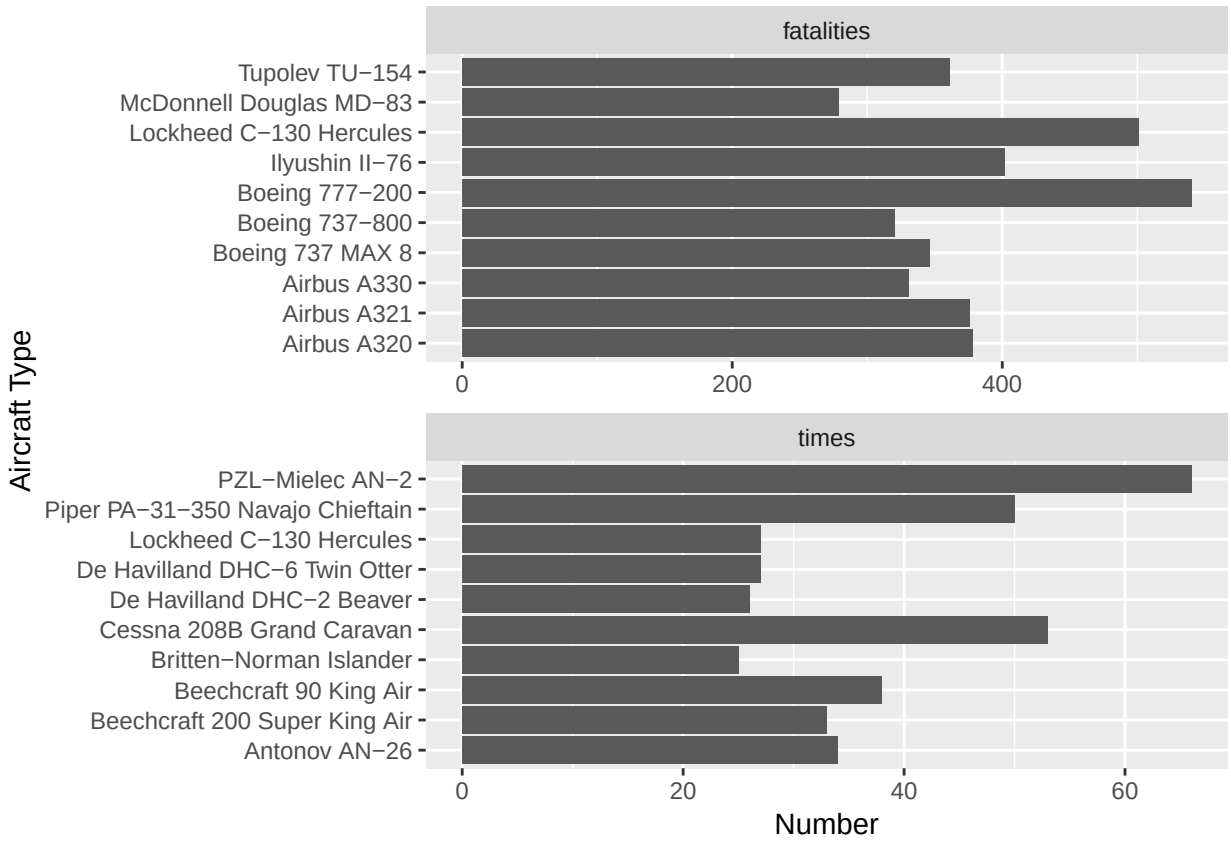
In the second step, our spider visits the page that contains the brief summary of air crashes. We choose to keep the records of Flight Phase and Circumstances for further study. The Circumstances are used mainly by the text mining part of this project.

Exploratory Analysis

```
## Observations: 1,368
## Variables: 7
## $ Date           <date> 2019-03-31, 2019-03-23, 2019-03-22, 2019-03-18...
## $ `A/C Type`     <chr> "Epic LT", "Beechcraft 200 Super King Air", "Ro...
## $ Location       <chr> "Egelsbach, Hesse", "Matsieng, Kgatleng Distric...
## $ Fatalities     <dbl> 3, 1, 0, 2, 1, 1, 157, 1, 14, 0, 0, 2, 3, 2, 1,...
## $ Registration   <chr> "RA-2151G", "A2-MBM", "N990PA", "N4MH", "N424TW...
## $ `Flight Phase` <chr> "Landing (descent or approach)", "Flight", "Lan...
## $ Circumstances  <chr> "While approaching Egelsbach runway 08 in good ...
```

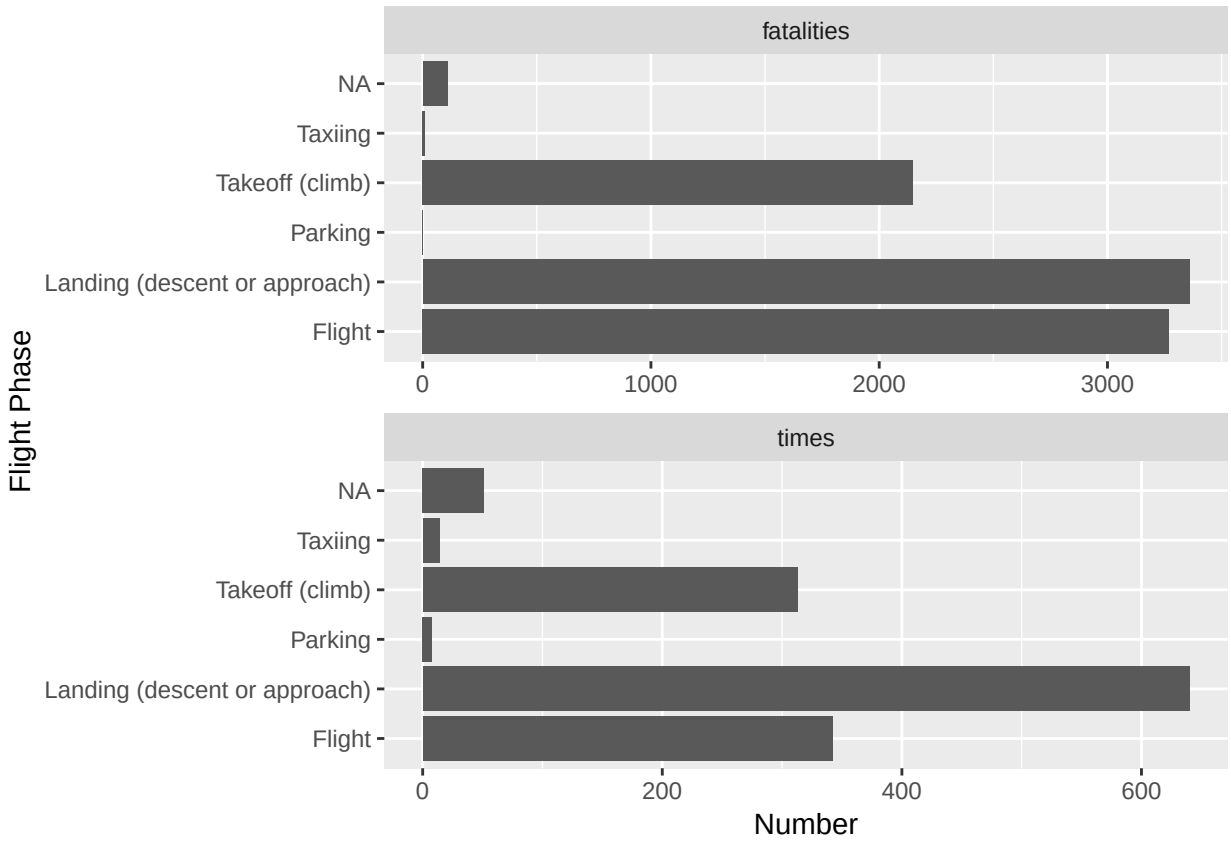
As shown the table above, our dataset contains 1368 observations and 7 variables. The 7 variables are Date, A/C Type, Location, Fatalities, Registration, Flight Phase, Circumstances.

Barplots are employed to see what types of aircraft is involved in most air crashes and what types of aircraft cause the most fatalities. Top 10 aircraft types are shown in the graphs below. **Boeing 777-200** have the most deaths and **PZL-Mielec AN-2** has been involved in the most accidents in the past 10 years Interestingly, the A/C Type which causes hundreds of deaths are not among the types which are most frequently involved in accidents, except for **Lockheed C-130 Hercules**. One possible reason for this phenomenon is that **Boeing** and **Airbus** produce commercial airplanes that can contain hundreds of people and a handful of aircrashes can lead to hundreds of fatalities.



From the barplots below, we can see that nearly half of the accidents that occur during the landing phase. Though only 1/4 of the accidents happen during the flight, the fatalities caused in this phase are as many as that those caused during the landing phase.

When the plane is landing, it has lower speed and lower altitude. So, even the airplane crashes during the landing or approaching phase, the possibility of having surviors must be higher than the accident in the sky. !!(Should cite someone from the profession)



Text Mining

Text mining techniques are employed to see the major causes of air crashes. Naive term frequency and TF-IDF are both tried in this part.

```
## # A tibble: 10,752 x 2
## # Groups:   word [10,752]
##    word          n
##    <chr>     <int>
##  1 aircraft   2789
##  2 flight     1564
##  3 pilot      1439
##  4 airplane   1354
##  5 runway     1318
##  6 airport    1116
##  7 crew       1010
##  8 engine      932
##  9 left        795
## 10 occupants   678
## # ... with 10,742 more rows
```

```
## # A tibble: 13,791 x 3
## # Groups:   is_fatal, word [13,791]
##    is_fatal word         n
##       <dbl> <chr>    <int>
##  1        0 aircraft  1212
##  2        0 runway     804
```

```
##  3        0 pilot      529
##  4        0 airplane   494
##  5        0 landing    493
##  6        0 flight     491
##  7        0 crew       425
##  8        0 airport    380
##  9        0 left       372
## 10        0 engine     364
## # ... with 13,781 more rows
```