0.1 Data definition

The database was download from (on 6.12.2018T8:00):

```
https://www.ntsb.gov/_layouts/ntsb.aviation/index.aspx
```

The NTSB aviation accident database contains information from 1962 and later about civil aviation accidents and selected incidents within the United States, its territories and possessions, and in international waters.

0.2 What/where/when

0.2.1 What

accident = defined precisely http://www.iprr.org/manuals/Annex13.html

incident = less serious

https://aviation.stackexchange.com/questions/14074/what-is-the-difference-between-aviation-accident-and-incident

Where = within the United States, its territories and possessions, and in international waters Not clear what does it precisely mean (within US?)

Moreover

```
awkF ’$6 !="United States" {print $0}’ AviationData.csv | wc -l
5088

awkF ’$6 =="Poland" {print $0}’ AviationData.csv | wc -l
23

awkF ’$6 =="Poland" {print $0}’ AviationData.csv
```

```
20180716X14029;Accident;CEN18WA268;07/11/2018;Domecko, Poland;Poland;50.624445;17.856944
...
http://klobuck.naszemiasto.pl/artykul/katastrofa-smiglowca-w-domecku-pod-opolem-wsrod-of

Internal flight, from Koszalin to Zibice. Relation to US is unclear
```

So we define our set: which occured in US (thus excluding outbound international flights and those crashed over ocean)

```
awkF ’$6 == "United States" {print $6}’ AviationData.csv > AviationData_US.csv
```

0.2.2 When

The file is dynamic and grows every day. The first observation is 1948

0.3 Data consistency

Size of dataset (how many cases

```
wc -l AviationData.txt
82575 AviationData.txt
```

Record is divided into FIELDs and the separation character is | I want to change ‘|’ into ‘;’ First I need to check if there are ‘;’ present

```
grep ’;’ AviationData.csv
20031209X02012 | Accident | ATL04FA045 | 12/04/2003 | Morlan, GA | United States | 33.29
20031008X01683 | Accident | IAD03CA071 | 08/19/2003 | CLEVELAND, OH | United States | 41
20001212X18693 | Accident | ATL99LA089 | 05/15/1999 | SMITH, AL | United States |  |  |
20001214X36822 | Accident | LAX85LA282 | 06/15/1985 | LAHAINA, HI | United States |  |
```

There are four ; and they need to be replaced with some other character before substituting ‘|’ to ‘;’

```
23832039 12-06 07:46 AviationData.txt
```

I rename the file to AviationData.csv

```
wc -l AviationData.csv
82575 AviationData.csv
```

Inspect 1st row (the header line)

```
Event Id | Investigation Type(2) | Accident Number(3) |
  Event Date(4) | Location(5) | Country(6) | Latitude(7) | Longitude(8) | Airport Code |
  Injury Severity | Aircraft Damage | Aircraft Category | Registration Number | Make | M
  Purpose of Flight | Air Carrier |
  Total Fatal Injuries(24) |
  Total Serious Injuries(25) | Total Minor Injuries | Total Uninjured | Weather Conditio
20181128X44044 | Accident | GAA19CA085 | 11/28/2018 | St. Petersburg, FL | United States
```

Check if crash site coordinates are always provided:

```
awkF ’$7 == "" {print $7}’ AviationData.csv  | wc -l
53894
```

How many occured in USA:

```
awkF ’$6 == "United States" {print $6}’ AviationData.csv | wc -l
77487
```

How many resulted in more than 100 fatalInjuries

```
awkF ’$24 > 100 {print $24}’ AviationData.csv | wc -l
61
```

How many resulted in at least one fatalInjury

```
awkF ’$24 > 0 {print $24}’ AviationData.csv | wc -l
16586
```

We create two files:AvData_USA_All.csv and AvData_USA_Fatal.csv

```
wc -l AvData*
      66 AvData.Readme
   77488 AvData_USA_All.csv
   14227 AvData_USA_Fatal.csv
   91781 razem
```

Our further analysis concerns AvData_USA_Fatal.csv

0.4 Data completness

Lack of coordinates:

```
awkF ’$7 == "" {print $7}’ AvData_USA_Fatal.csv  | wc -l
9228
```

Data consistency

```
awkF ’$4 !~ /[0-9][0-9]\/[0-9][0-9]\/[0-9][0-9]/ {print $4}’ AvData_USA_Fatal.csv
```

Event Date

```
time coverage:
awkF ’NR > 1 {split ($4, d, /\//); t=d[3] " " d[1] " " d[2] " 0 0 0"; if (d[3] < 1982) {p
5
```

Data starts from 1982 with just a few records from older accidents

```
awkF ’NR > 1 {split ($4, d, /\//); t=d[3] " " d[1] " " d[2] " 0 0 0"; if (d[3] > 1981) {p
"d[2] }}’ AvData_USA_Fatal.csv  > AvData_USA_1982A.csv
```

We limit our analysis further (when = 1982 and earlier)

```
awkF ’$24 > 0 && $6 == "United States" {print $24}’ AvData_USA_1982A.csv > AvData_USA_19
```

Last check

```
awkF ’NF!=36{print NF}’ AvData_USA_1982F_dow.csv
awkF ’NF!=36{print NF}’ AvData_USA_1982A_dow.csv
```