Getting the dataset from the web

```
mydat <-read.csv('https://raw.githubusercontent.com/fivethirtyeight/data/master/airline-safety/airline-
```

Create totals for incidents, fatal accidents and fatalities

```
mydat <- transform (mydat,
                    total_incidents = incidents_85_99 + incidents_00_14,
                    total_fatal_accidents = fatal_accidents_85_99 + fatal_accidents_00_14,
                    total_fatalities = fatalities_85_99 + fatalities_00_14)
```

Airline_data_analysis.R

The first thing that this file does is pulling the data from FiveThirtyEight. This dataset contains the following variables:

Variable	Explanation
airline	Name of the airline
avail_seat_km_per_week	Available seat kilometers flown every week
incidents_85_99	Total number of incidents, 1985-1999
fatal_Accidents_85_99	Total number of fatal accidents, 1985-1999
fatalities_85_99	Total numer of fatalities, 1985-1999
incidents_00_14	Total numer of incidents, 2000-2014
fatal_accidents_00_14	Total numer of fatal accidents 2000-2014
fatalities_00_14	Total numer of fatalities, 2000-2014

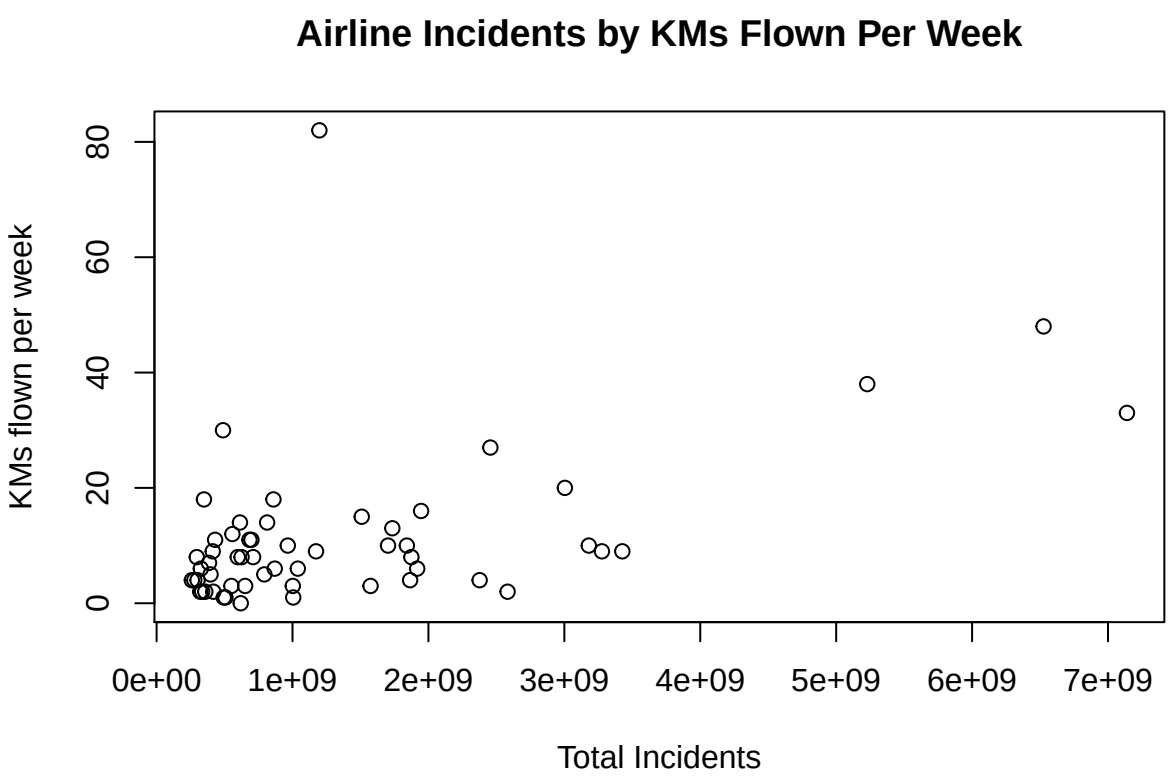
Then it calculates means. This exercise is useful to compare the mean number of incidents in the period of time 1985-1999 with the mean number of accidents in the period of time 2000-2014. It then does the same exercise for the number of fatal accidents and the number of fatalities. The mean number of indicents, fatal accidents and fatalities decreased from the period 1985-1999 to the period 2000-2014.

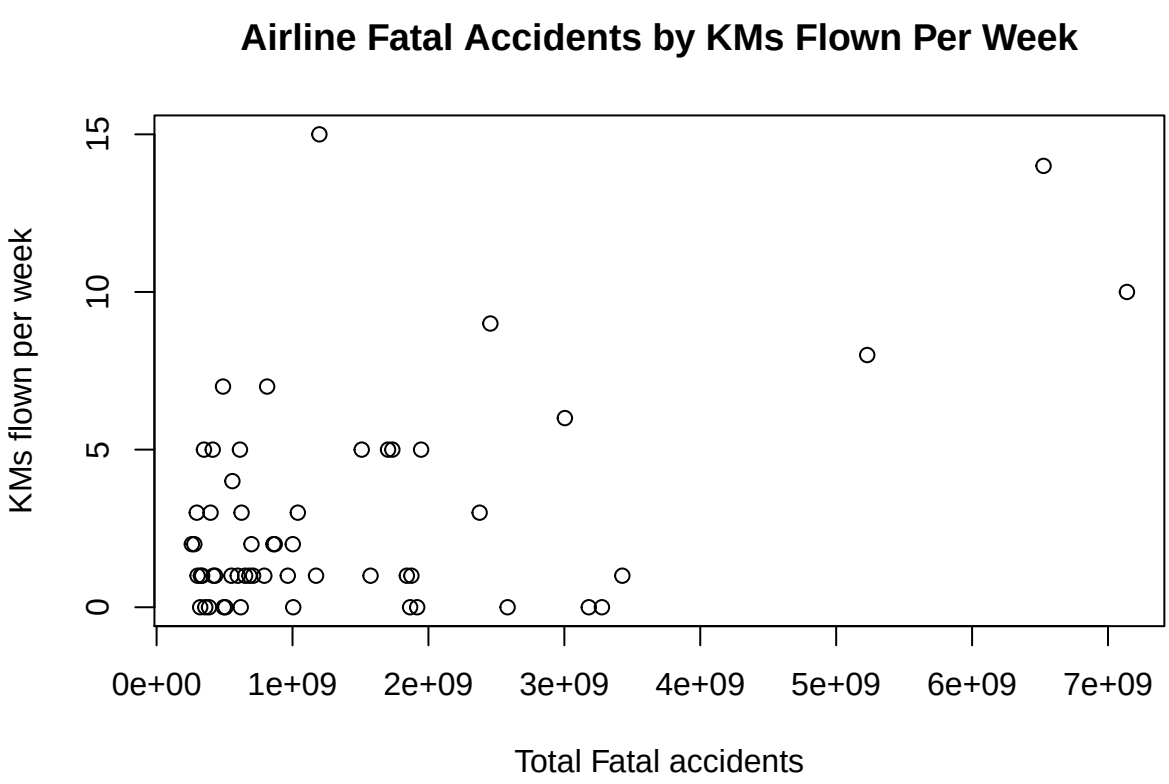
The file also creates three variables that are added to the existing dataframe. The variables are *total_incidents*, *total_fatal_accidents* and *total_fatalities*. These variables are just the totals for the entire time period we have available data: 1985-2014.

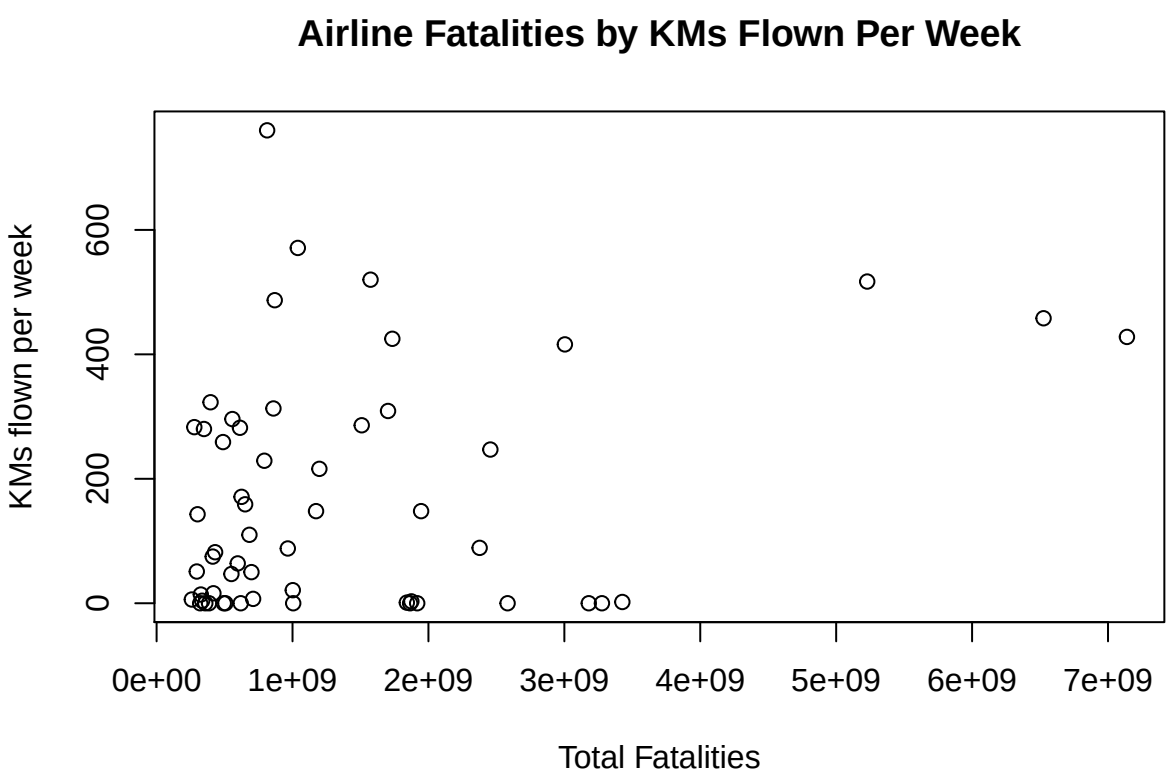
These variables were useful to create three indexes that represent the amount of incidents, fatal accidents and fatalities per number of available seat kilometers per week and multiplied by 1000000000. This is useful to learn about which airlines had more incidents, fatal accidents and fatalities taking into account that the more (less) they fly the bigger (smaller) is the probability of this events to happen.The indexes were added as new variables in the dataframe.

After this three different dataframes were created each of them sorting by index (smallest to largest) to check which ones were the most *dangerous* airlines (please note that I am aware that this is a quite strong afirmation with the data limitations I have but still this is what I can do in this case). The bigger the index, the more likely is the airline to suffer from indicents, fatal accidents and fatalities respectively.

Lastly, I created three graphs using incidents, fatal accidents and fatalities against number of available seat kilometers per week. In this way we can clearly see that some airlines are outliers and that are out of the logic that the more the airline fly, the more incidents, fatal accidents or fatalities would suffer from. Therefore, the outlier airlines are more likely to suffer this kind of accidents.


Airline Incidents by KMs Flown Per Week
KMs flown per week
0
20
40
60
80
0e+00
1e+09
2e+09
3e+09
4e+09
5e+09
6e+09
7e+09
Total Incidents


Airline Fatal Accidents by KMs Flown Per Week
KMs flown per week
0
5
10
15
0e+00
1e+09
2e+09
3e+09
4e+09
5e+09
6e+09
7e+09
Total Fatal accidents


Airline Fatalities by KMs Flown Per Week
KMs flown per week
0
200
400
600
0e+00
1e+09
2e+09
3e+09
4e+09
5e+09
6e+09
7e+09
Total Fatalities