**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Airplane Crashes and Fatalities Since 1908**

Marquette University

Data Science Project

Rafael Leite

04/11/2017

**Introduction**

Air travel has become more accessible to people in different parts of the world. Because of that, more people are using this type of transportation. This increase in number of passengers has been accompanied by an increase of safety measures. It is common knowledge that air travel is safer than other types of transportation, especially automobiles. In fact, commercial planes are the safer mean of transportation in the United States (What's the safest way to travel); however, there are a lot of people that still get nervous when they need to use them. The chances of a plane crash are very slim, but when an accident occurs, it usually leaves no survivors, and this is what scares people the most. The purpose of this project is to investigate if planes have indeed become safer by trying to answer questions like: Should we avoid airlines that have had crashes in the past? Have airplanes become more secure? And are we more likely to survive a plane crash now than we were in the last century?

**Data Set**

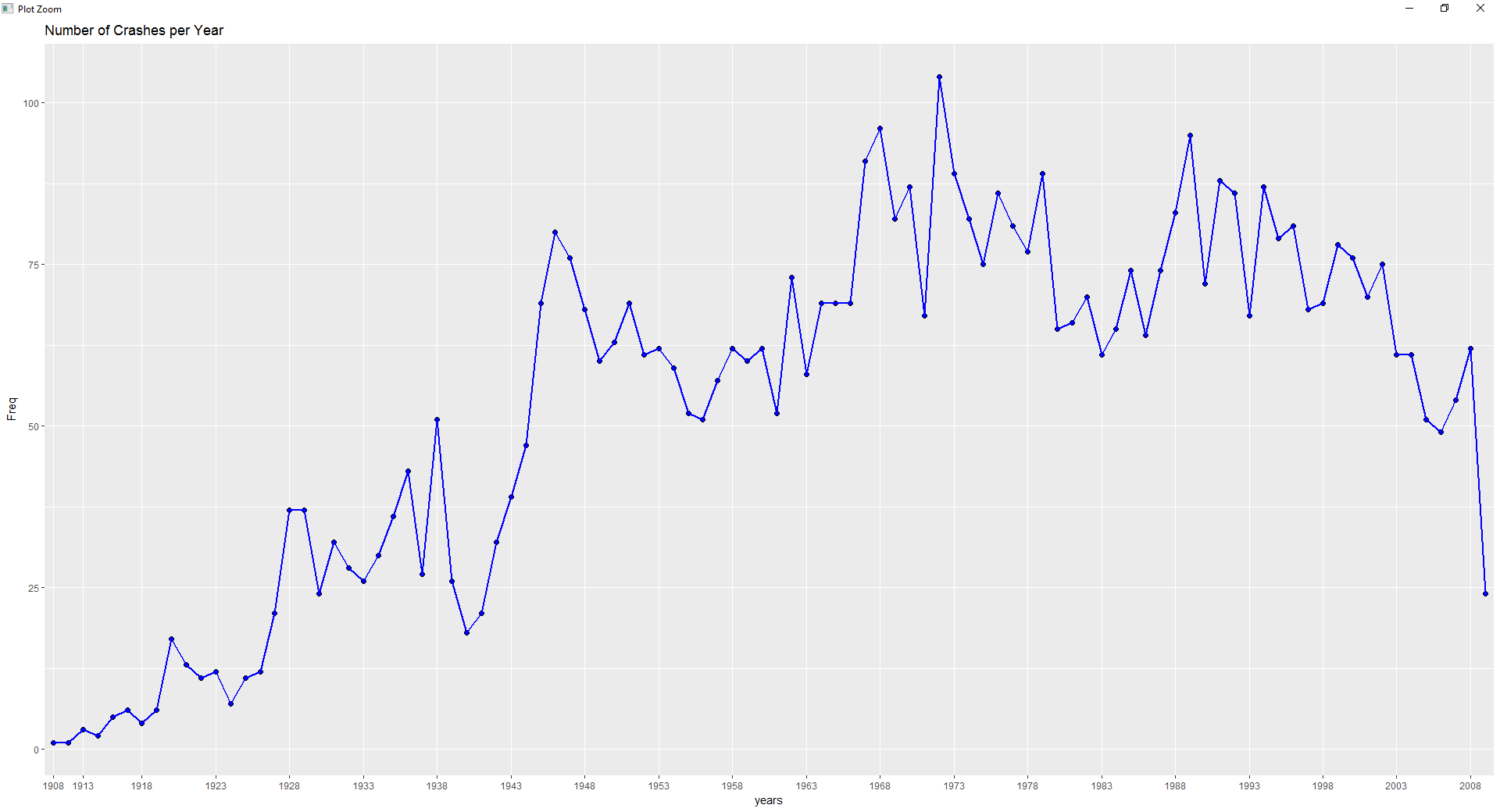
The data set used for this project is the “Airplane Crashes and Fatalities Since 1908” data set. It contains more than 5,000 records of plane accidents from 1908 to 2009. Its attributes are: Date, Time, Location, Operator, Flight #, Route, Type, Registration, cn/In, Aboard, Fatalities, Ground, and Summary.

**Analysis/Evaluation**

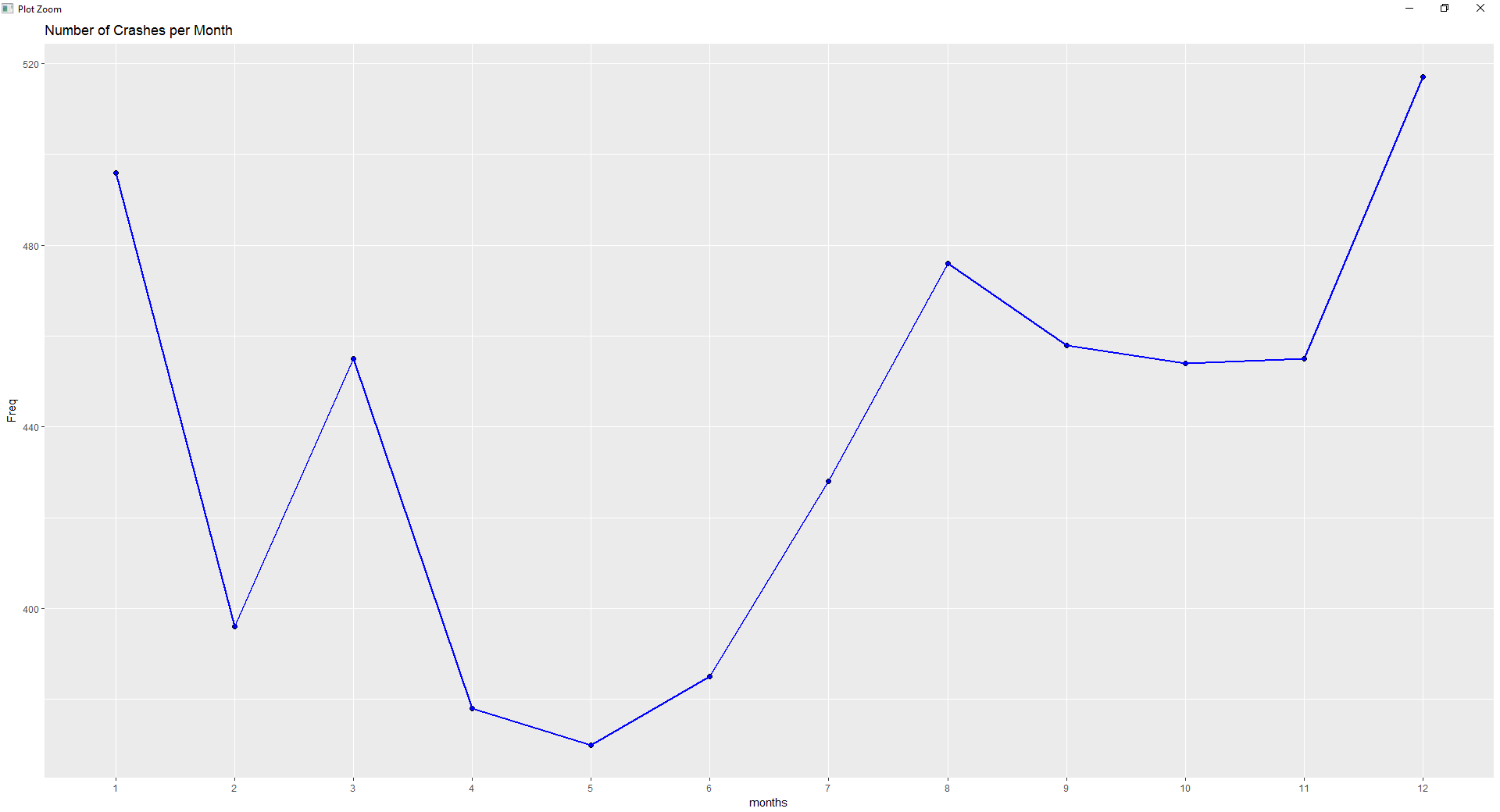
The link to the Github repository with the source code containing all the calculations and graphs can be seen below.

<https://github.com/rzufil/COSC4931_DataScience>

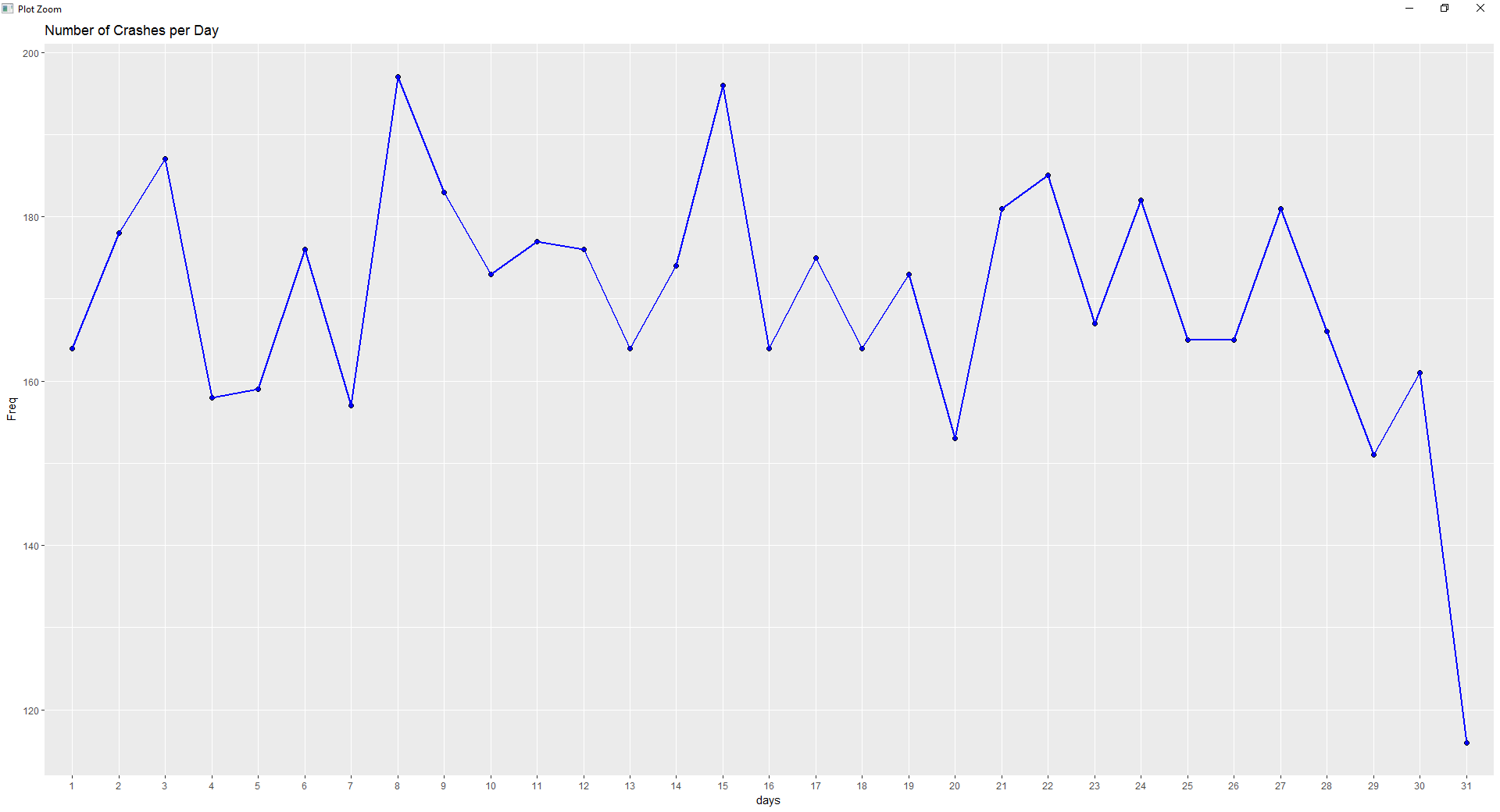
To have a better insight of the distribution of plane crashes throughout the years, exploratory analysis was conducted. RStudio was used to create the graphs below.



The first graph shows the frequency of plane crashes per year. A huge increase of plane crashes can be seen during the mid-40’s. The frequency stabilizes during the 50’s and 60’s. It spikes in the late 60’s and hits its all-time high in the early 70’s. It then returns to the frequency before the late 60’s spike and later proceeds to plummet after the turn of the millennium.

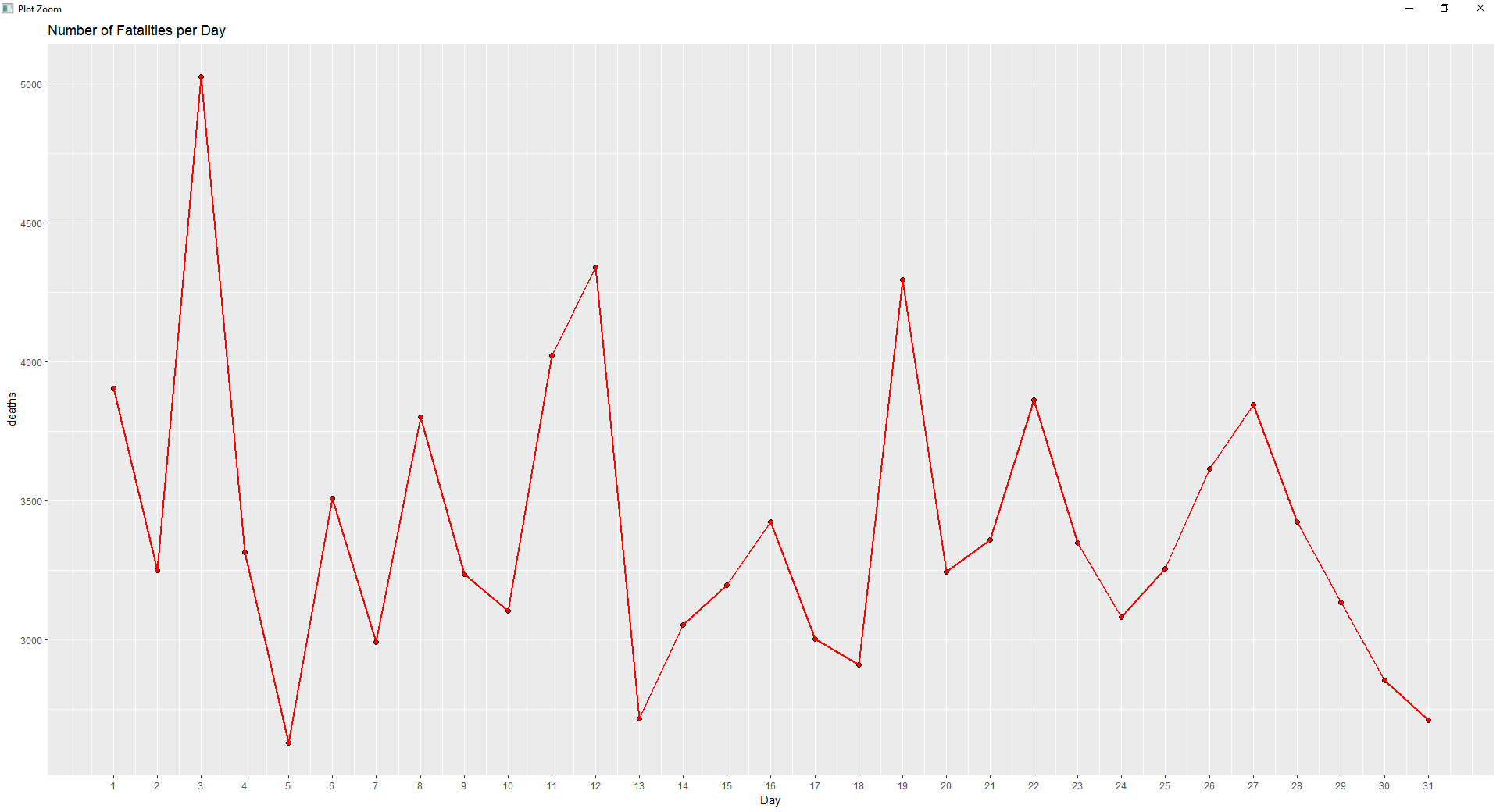


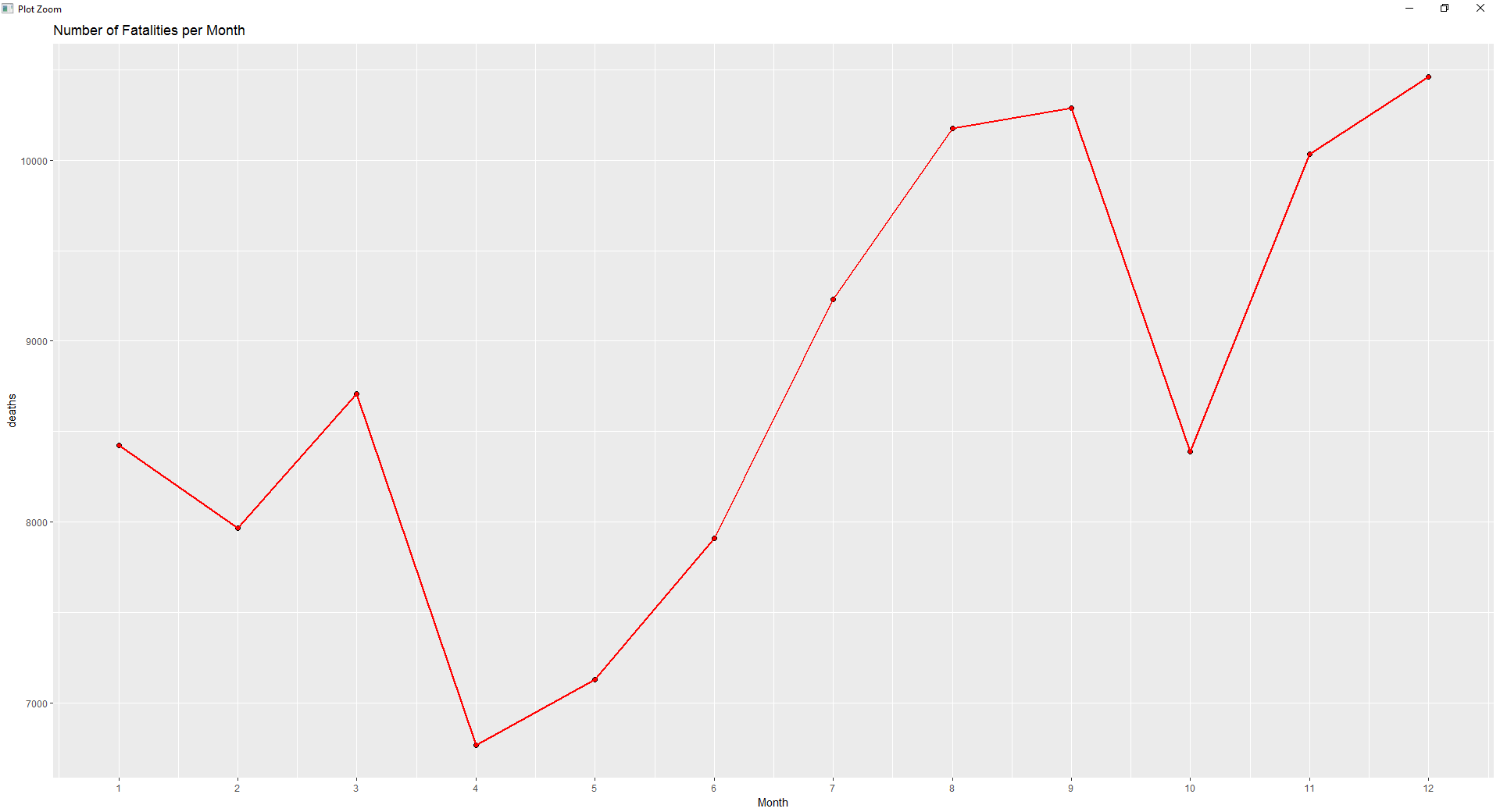
The frequency of accidents throughout the months of the year was also observed. The graph above shows a very high number of plane crashes during the last and the first month of the year. This finding was expected since there is an increase of travels during those months. May is the month with the lowest frequency of plane crashes, while December is the month with the highest frequency.

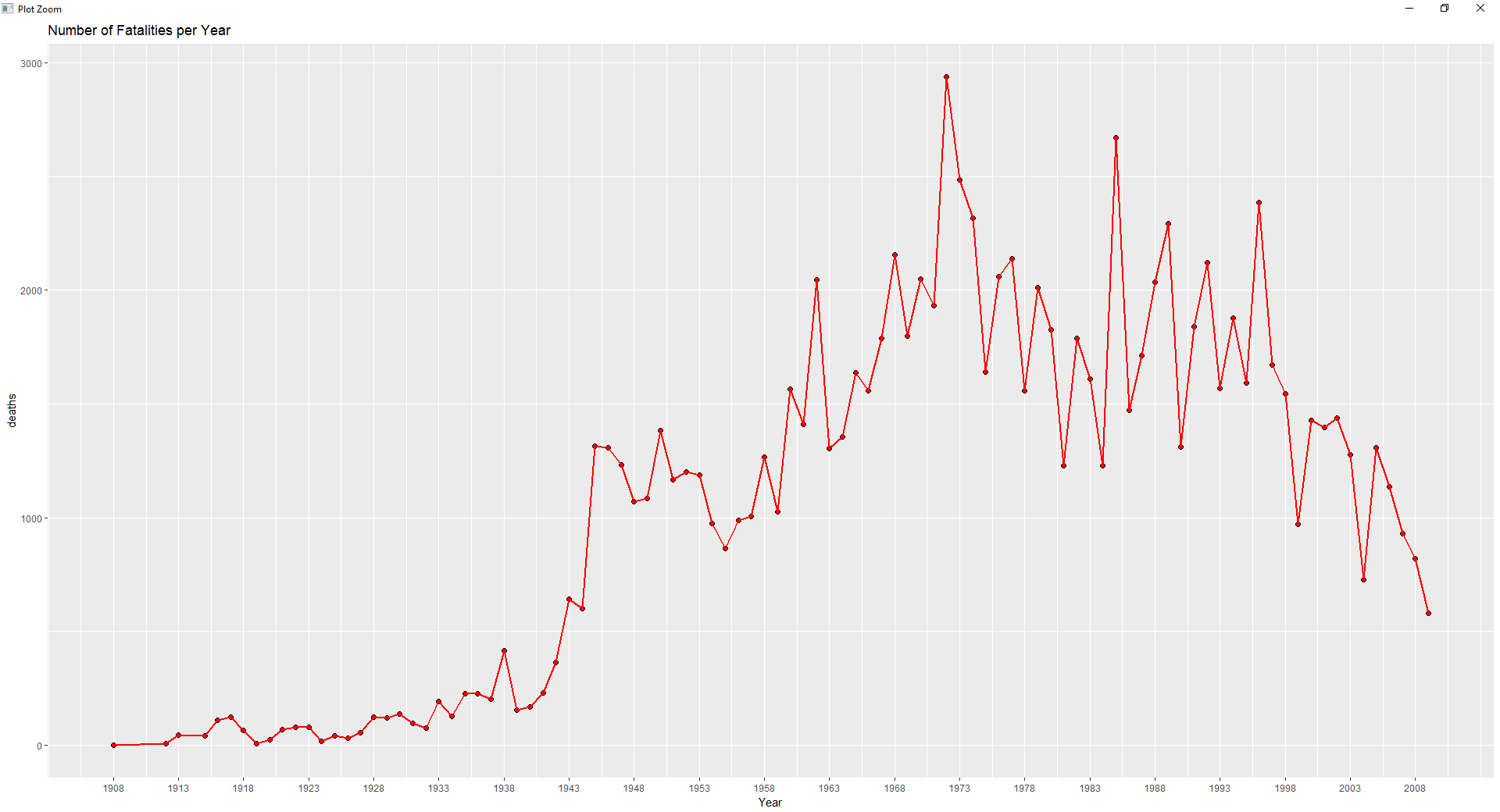


The third graph shows the frequency of plane crashes per day of the month. The graph does not show much variation throughout the month and is somewhat evenly distributed, except for the last day of the month, which exhibits a very low frequency of crashes compared to the other days of the month.

The number of deaths resulted from those plane crashes was also analyzed. The number of fatalities per day, month, and year somewhat followed the trend exhibited by the number of crashes per day, month, and year respectively. The graphs and results can be seen below.

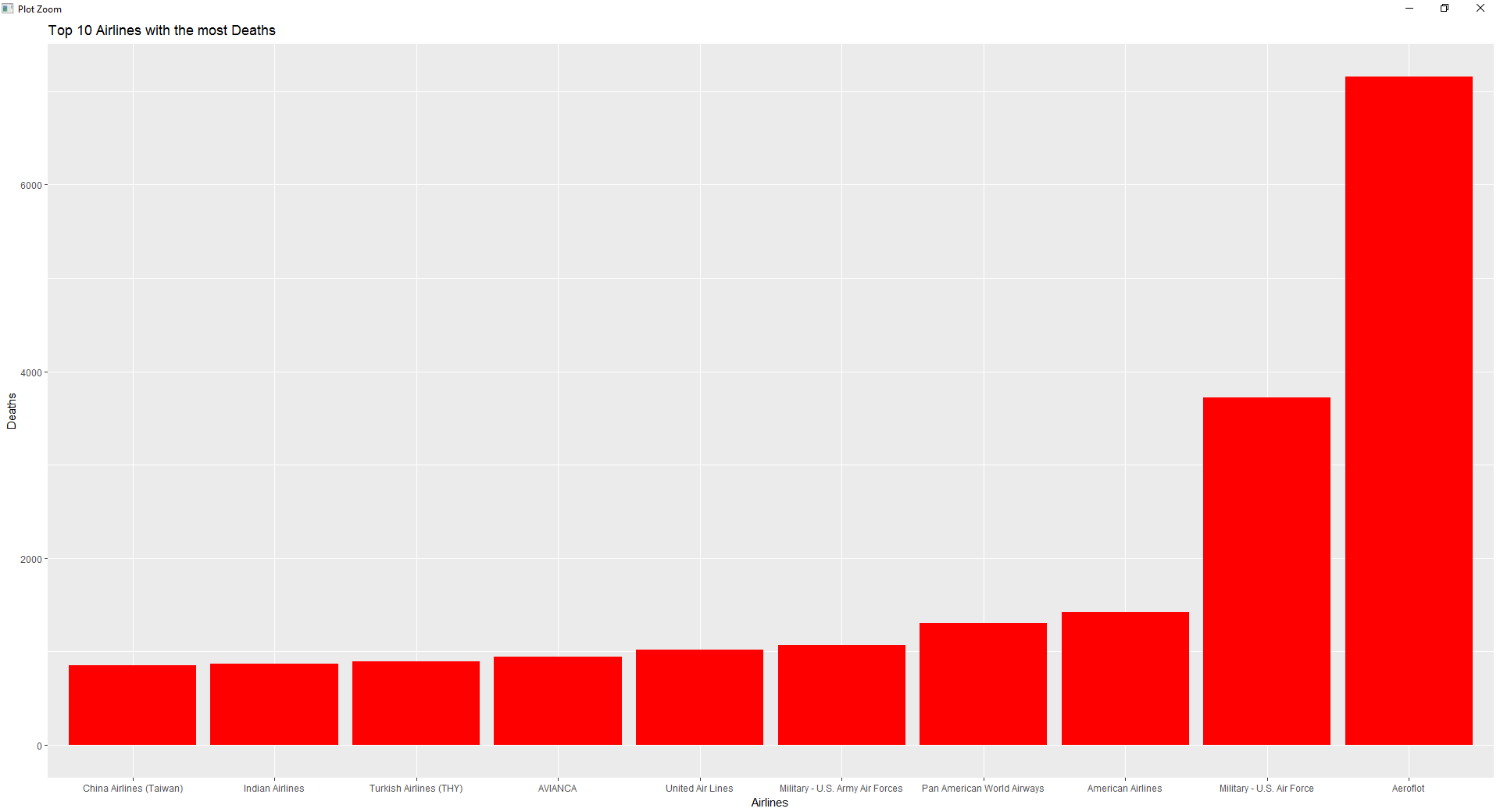


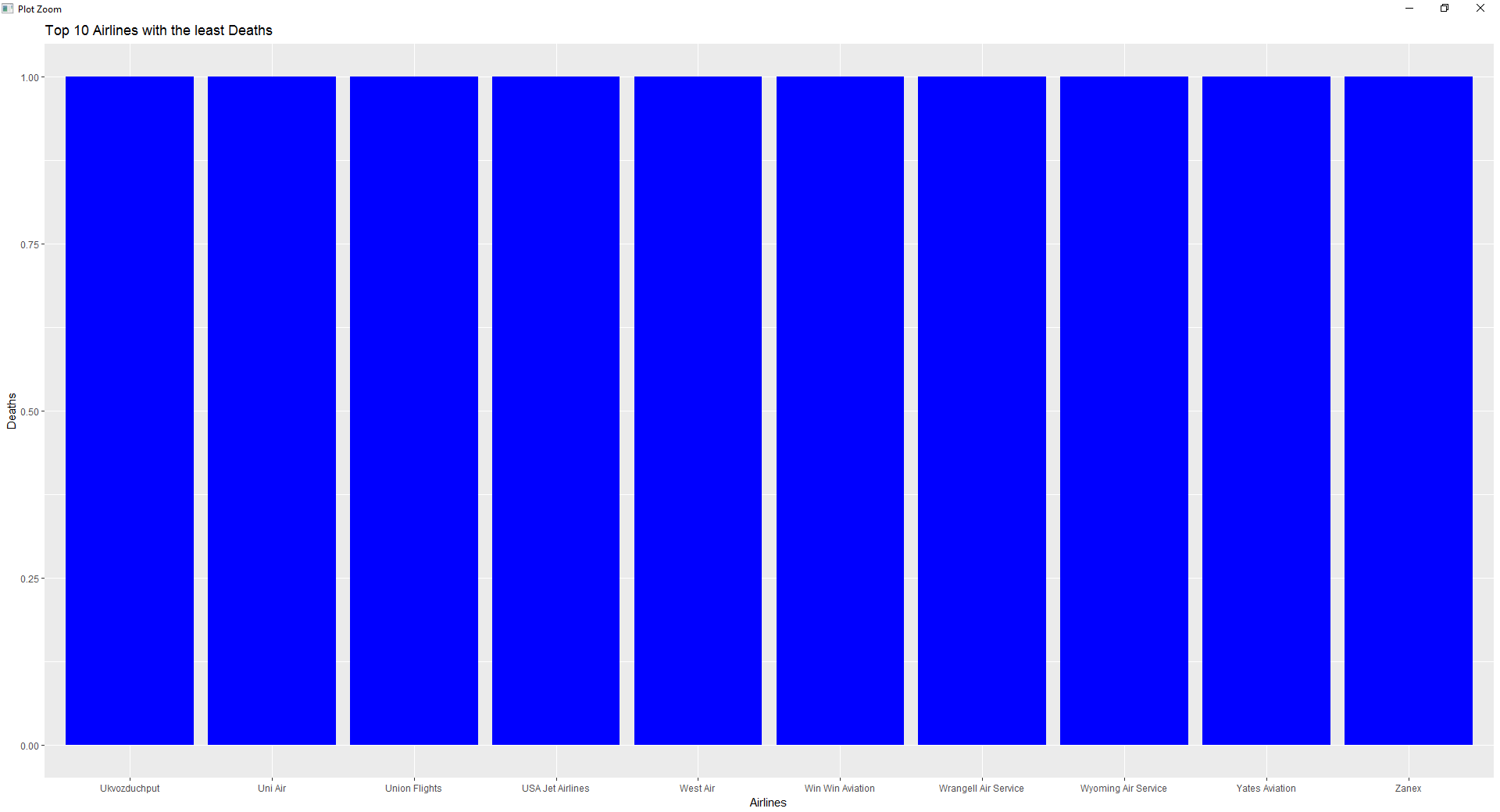




The results from the graphs hint that the chances of surviving a plane crash have not increased over the years. Otherwise the graphs would show a significantly higher decrease in the frequency of deaths in the past 20 years, something that was not observed.

Exploratory analysis was also conducted to investigate the top 10 airlines with the most accidents and the top 10 airlines with the least accidents.





The operators with the most fatalities caused by crashes during their flights are: Aeroflot, the U.S. Air Force, and American Airlines. Many airlines had only one occasion in which a person died during a flight that they were operating. However, further investigation will be done in order to confirm if the top 10 airlines with least fatalities are really safer than the others or if they are outliers.

To have a better understanding of the main causes of plane crashes, text analysis was also conducted. RStudio was also used for this analysis and to create the word cloud below.



word freq

pilot pilot 1015

approach approach 943

engine engine 925

runway runway 916

failure failure 880

crew crew 751

landing landing 734

airport airport 625

altitude altitude 608

weather weather 605

The list above shows the most frequent words in the “Summary” attribute of the data set. “Pilot” is the most cited word in the entire data set, with 1015 occurrences. Followed by “approach,” “engine,” “runway,” and “failure,” with 943, 925, 916, 880 occurrences respectively. This word cloud points that human error, made by the pilot or a crew member, is the most common cause of plane crashes. However, problems with the aircraft is also a common cause of accidents since engine was the third most mentioned word in the data set. It can also be inferred from the word cloud that most accidents happen when the aircraft is approaching the runway of the airport and not while mid-air. Further analysis will be done in order to find word clusters that could explain better the details of the most common accident causes.

**References**

Isidore, Chris. May 13, 2015. *What's the safest way to travel*. CNN Money. <http://money.cnn.com/2015/05/13/news/economy/train-plane-car-deaths/>

*Airplane Crashes and Fatalities Since 1908*. Socrata OpenData. <https://opendata.socrata.com/Government/Airplane-Crashes-and-Fatalities-Since-1908/q2te-8cvq>