AIRCRAFT CRASH DATA ANALYSIS

Statistical Computing
Data Analysis Report

Student
Name
Meltem DEMİR

Project Title
Aircraft Crash Data Analysis

May 28, 2017

Department of Computer Engineering
College of Engineering, Mugla Sitki Kocman University
Mugla, TR 48000

AIRCRAFT CRASH DATA ANALYSIS

Meltem Demir

Abstract

This data frame includes with 5666 observations on the following 7 variables which are Date, location, operator, planeType, Dead, Aboard, Ground. Format of this dataset is in .CSV format.

This exploratory data analysis of the airplane crash data analyzes the crash trend for over 100 years beginning from the year 1908 to 2014. It is particularly interesting to observe the trend of airplane crashes and the reasons behind them, as air travel is the one of the most common transport medium these days. It is also important to examine my progress in overcoming the crashes.

This analysis will be provide insights in observing the trend of air crash over the years. It shows the number of fatalities observed due to the crash. The analysis also will help in determining which airline operator and types are worst to fly with. I will also observe the top 10 countries which we should avoid to escape the crash. The analysis also will help in determining if it is increasing air crash year by year. In this manner, I can create a general judgment about this data frame. All these topics will be analyzed.

At the end of these analysis, this dataset will provide a general judgment about countries, operators and types.

Table of Contents

1. Description of Problem

The main question is "which factors that cause the air crash?"

Aircraft Crash Data Analysis is for over 100 years beginning from the year 1908 to 2008. Air travel is the one of the most common transport in these days. The main purpose is to find which factor is more effecting on air crash. There is always reason to happen these air crashes. In this data frame, I believe that I can find a factor to cause these situations. It is important to examine our progress in overcoming the crashes.

The data used for this analysis is a public dataset hosted by open Data by Vincen Tarel Bundock. Various data cleaning steps were performed to work on a tidy dataset. After calculations, graphs were plotted to visualize the results and come to a conclusion.

2. Description of Data

- **Original Data**

The data set I used is a public dataset: “AirAircraft Crash data” which is hosted by open Data by Vincent Tarel Bundock at:
https://vincentarelbundock.github.io/Rdatasets/doc/gamclass/airAccs.html

This data is including 5666 observations on the following 7 variables. All columns is at the below.
The first state of the data:

- Date - The date on which the flight crashed.
- Time - The time at which flight crashed.
- Location - Location of the crash
- Operator - The name of the flight operator
- Flight - Flight Number of the airplane that crashed
- Route - The Route of the flight
- Type - The type of flight carrier
- Registration - Description unavailable. This variable wouldn’t be used for analysis.
- cn.In - Description unavailable.
- Aboard - The number of passenger on board
- Fatalities - The number of deaths
- Ground - Description unavailable.
- Summary - Brief summary

Data Cleaning

All steps up into the end of the beginning is in Process.R. If I summarize the steps, it would be like this;

1. Importing Dataset (in csv format)
2. Evaluating missing values
3. Spliting "date" column in day month year
4. Spliting "location" column in Country and City.
5. Removing some unneccessary rows in data frame.
6. Giving new column names.
7. Converting some columns situation like factor to numeric etc..

Summary of Dataset

After cleaning dataset, here is the most important variables that are used for the analysis are as follows:

- Accident.Year : The dataset can be grouped by year to see the yearly trends.
- Accident.Month : The dataset can be grouped by monthly trends (if any).
- Accident.Day : The dataset can be grouped by daily trends (if any).
- Accident.Time : The dataset can be grouped by hourly trends (if any).
- Accident.Country : The dataset can be grouped by country names
- Accident.City : The dataset can be grouped by city names
- Aircraft.Operator : This Aircraft Operator column is used to understand which operator had maximum crashes.
- Aircraft.Type: This Aircraft Type column is used to analyze which type of aircraft caused maximum crashes.
- Number.Aboard: This column is used to determine the percent of deaths that occured every year.
- Number.of.Death: This is count of deaths occured. Helpful in determing the total loss.
- Deaths.on.Ground : This is count of deaths on ground.
- Summary: This has reasons for the crash. Text mining can be performed on this column to understand the most frequent causes of aircrash.

3. Progress to Date

After cleaning the dataset, I started to analyzing.

Data Analysis

✓ **Crash Situation**

✓ **By Years**

The Date column that was split into Month, Year, Day and time are used here. A trend line in the Total number of crashes per year shows that number of crashes are reducing from the decade 1968-1978. And it was the maximum in the decade 1968-1978.

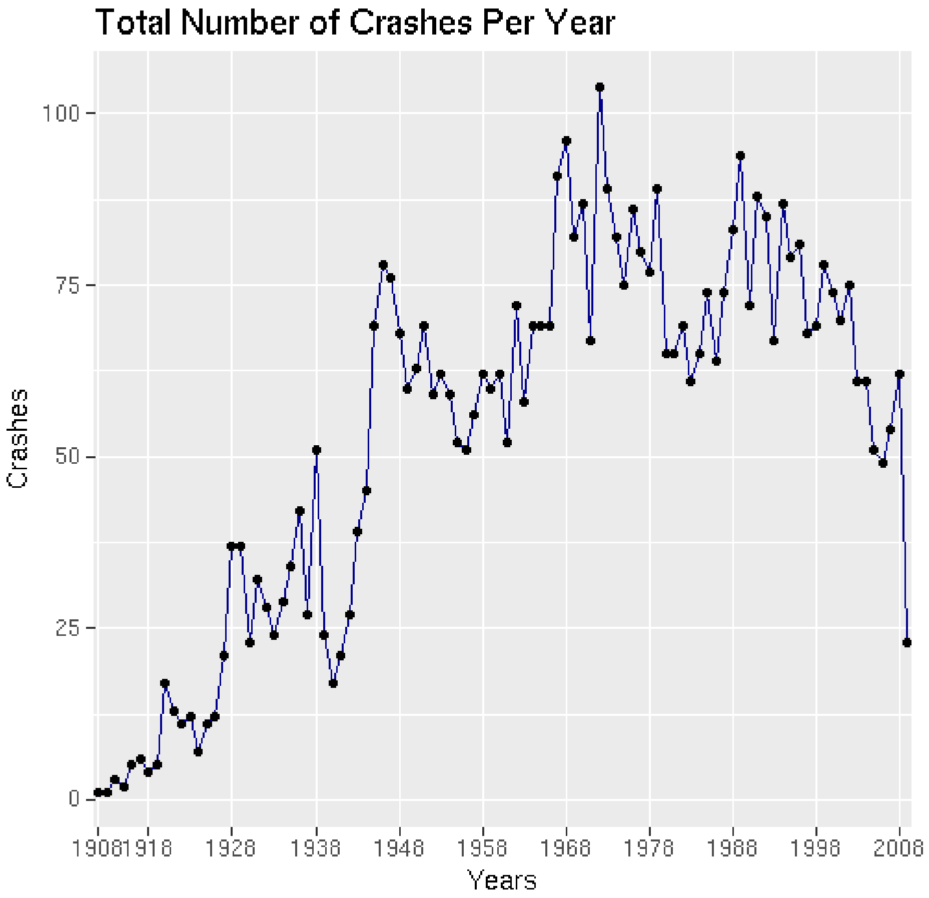


Figure 1. Line Graph for Total Number of Crashes Per Year

- ✓ **By Months**

The month analysis just gives a confirmation that crashes occur irrespective of the month. That means, the time of the year is not significant influencing parameter.

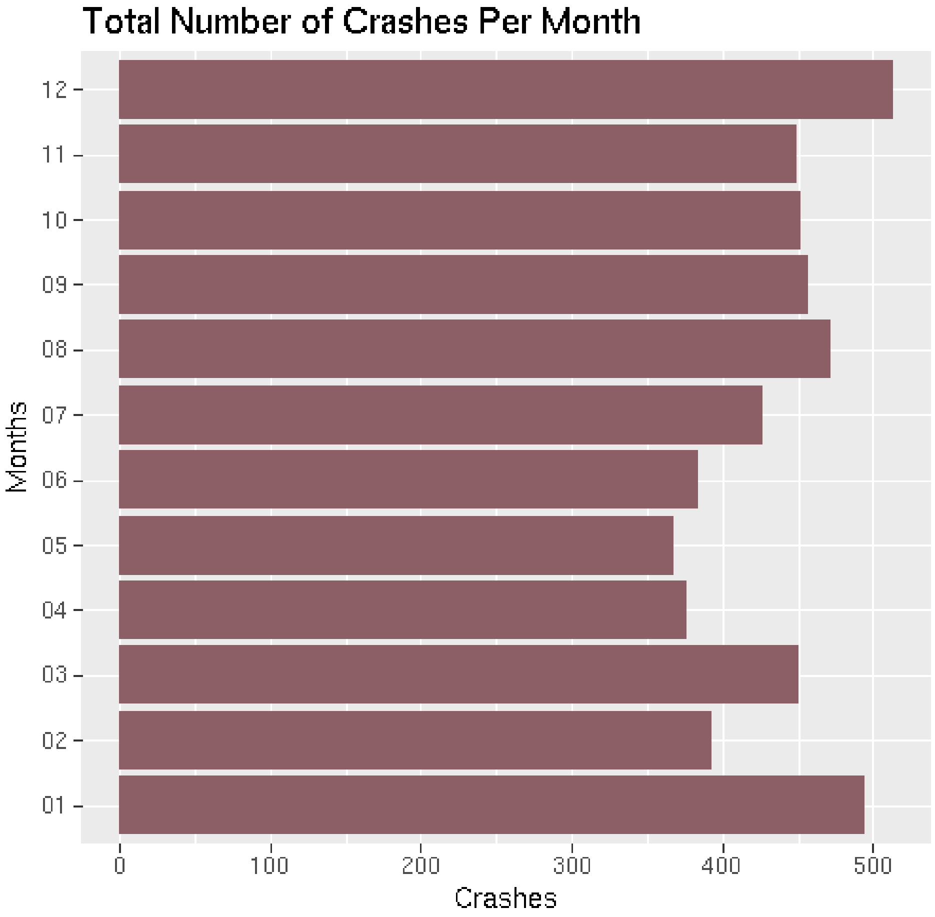


Figure 2. Bar Chart for Total Number of Crashes Per Month

- ✓ **By Days**

The day analysis just gives a confirmation that crashes occur irrespective of day. That means, the day of the month is not significant influencing parameter.

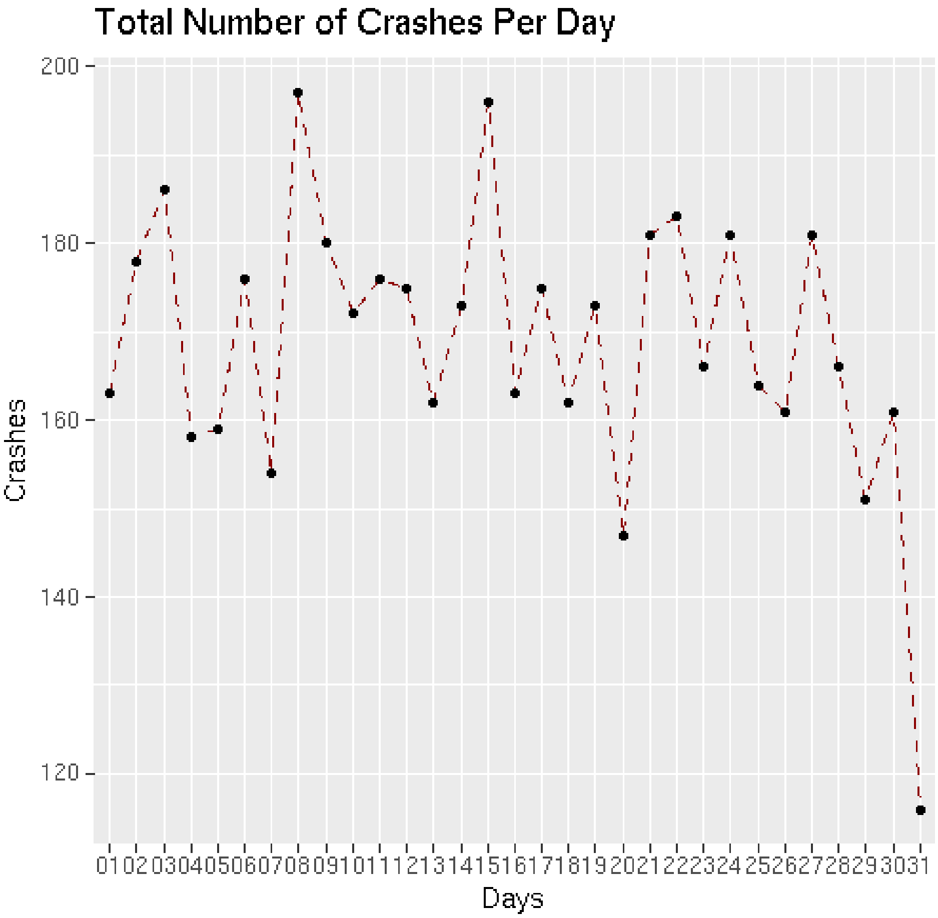


Figure 3. Line Graph for Total Number of Crashes Per Day

- ✔ **Death Percent**
  - ✔ **Death by Years**

First, the Deaths are grouped from the main Aircrash table grouped by the year. The total number of deaths is calculated by years. This information is used to plot the deaths over the year. Here we observe that the percent of deaths is increasing and decreasing with time. This analysis just gives a confirmation that crashes occur irrespective of the year. That means, Number of deaths by years is not a significant influencing parameter.

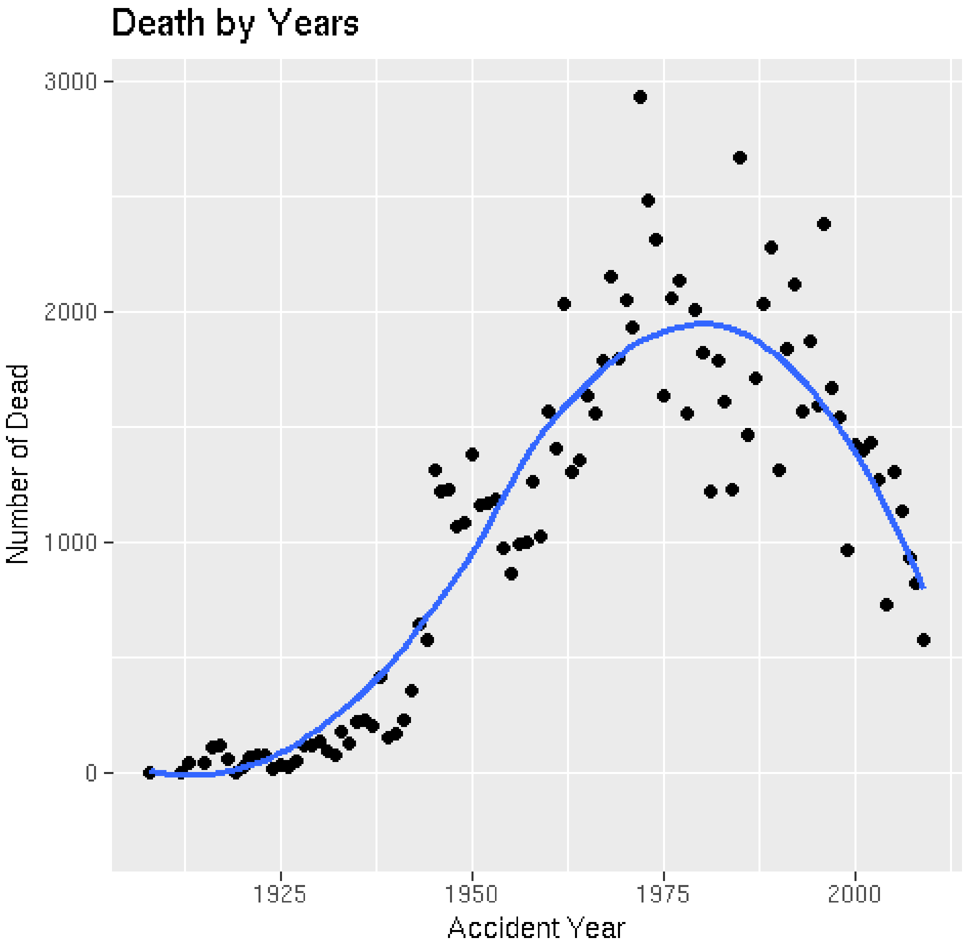


Figure 4. Point and Smooth Graph for Death by Years

✔ Percentage of Death by Years

First, the deaths are grouped from the main Aircrash table grouped by the year. The total number of deaths and number of passengers aboard is calculated. This information is used to plot the percent deaths over the year. Here we observe that the percent of deaths is decreasing with time.This should imply that the safety measures for the people onboard must have increased.

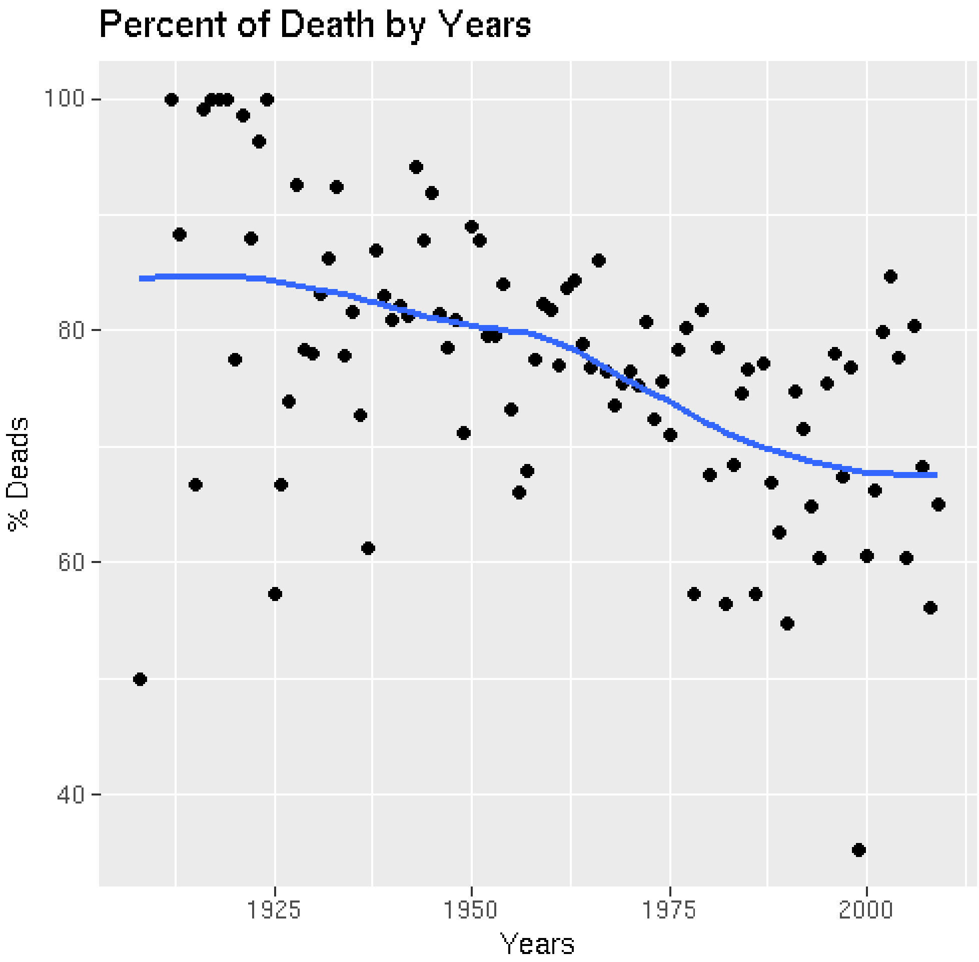


Figure 5. Point and Smooth Graph for Percentage of Death by Years

- **Crash Location**
  - **By Countries**

The Location column that was spilled into country and city is used here. The data is grouped by the country and the total deaths for each country is calculated. Here we plot a graph to observe the top 10 countries which encountered the aircrash. It is observed that Russia and Brazil has had the maximum crashes out of all the Countries.

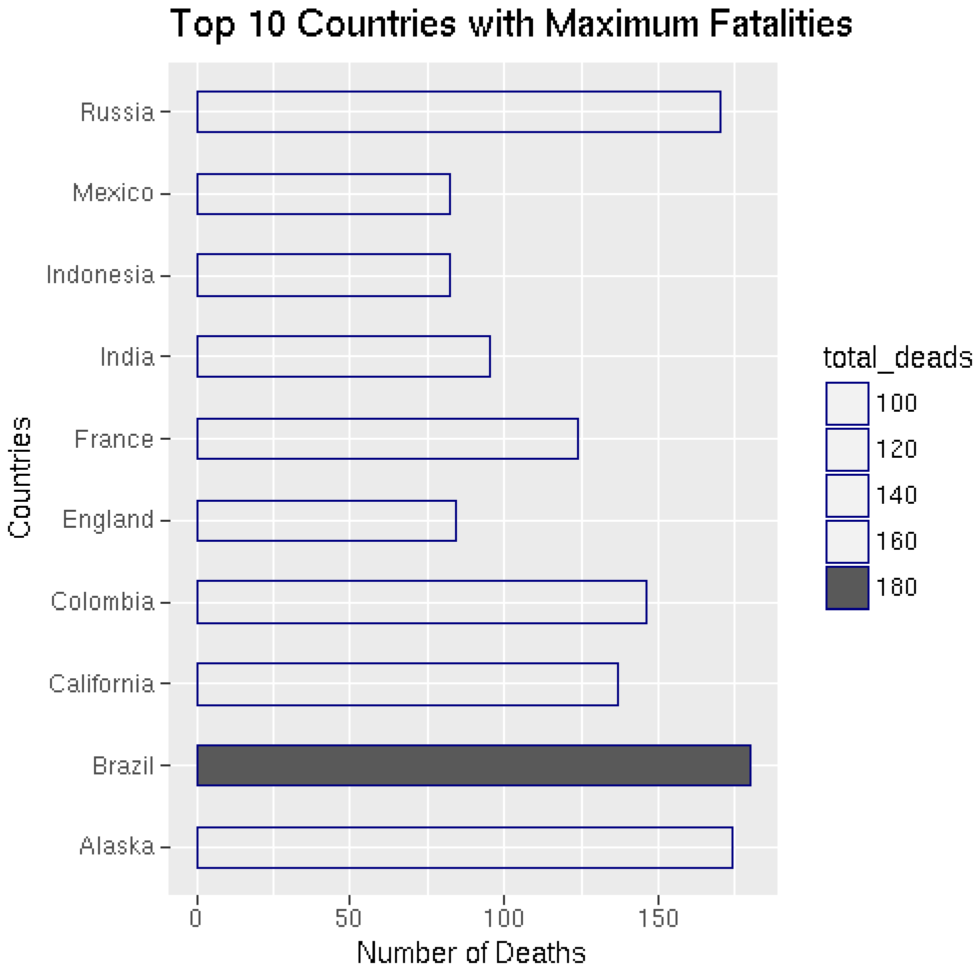


Figure 6. Bar Chart for Top 10 Countries with Maximum Death

- ✔ **By Cities**

The Location column that was spilled into country and city is used here. The data is grouped by the city and the total deaths for each city is calculated. Here we plot a graph to observe the top 10 cities which encountered the aircrash. It is observed that Moscow and San Paulo has had the maximum crashes out of all the Cities.

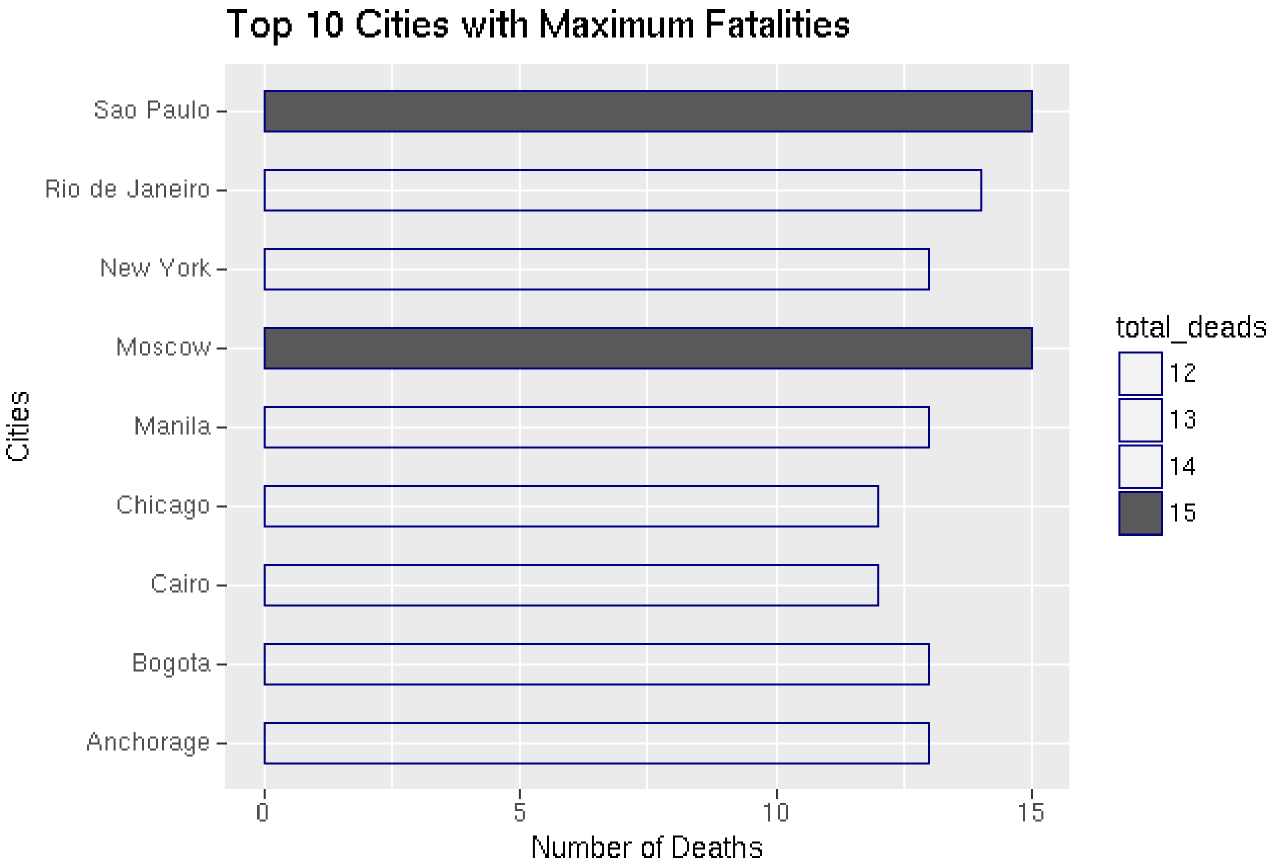


Figure 7. Bar Chart for Top 10 Cities with Maximum Deaths

- ### ✔ Aircraft Operator

To understand which Operators caused more crashes the data is grouped by the Operator and the total deaths for each operator. Here we plot a graph to observe the top 10 operators which encountered the aircrash. It is observed that Aeroflot has had the maximum crashes out of all the Operators

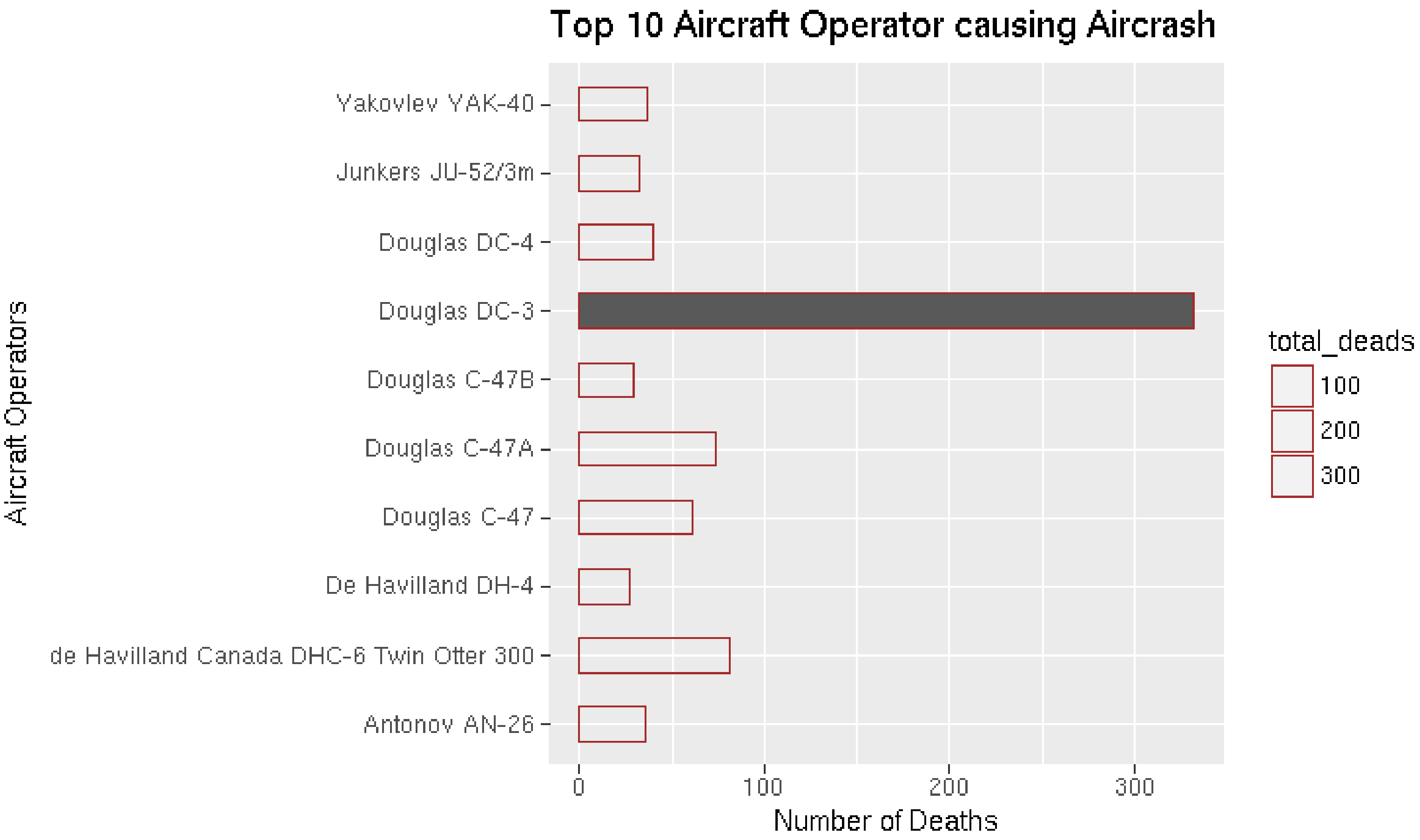


Figure 8. Bar Chart for Top 10 Operator causing Aircrash

- ### Aircraft Type

To understand which Airplane Type caused more crashes the data is grouped by the Type and the total deaths for each operator. Here we plot a graph to observe the top 10 Airplane Types which encountered the aircrash. It is observed that Douglas DC-3 has had the maximum crashes out of all the Types.

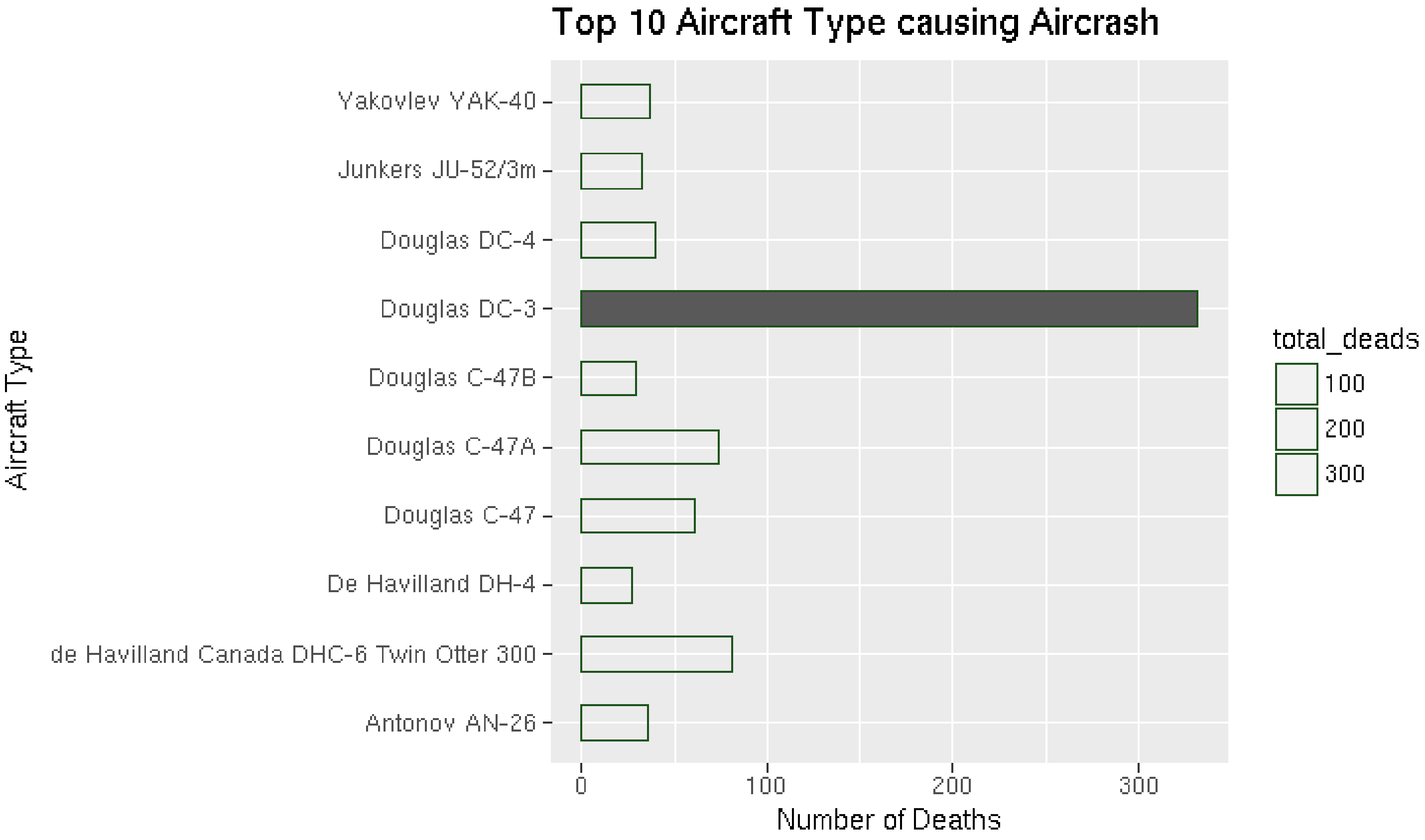


Figure 9. Bar Chart for Top 10 Airplane Type causing Aircrash

✔ **Word Cloud**

Here I have experimented with text mining in R on the summary column to form a word cloud that states the reasons for aircrash.

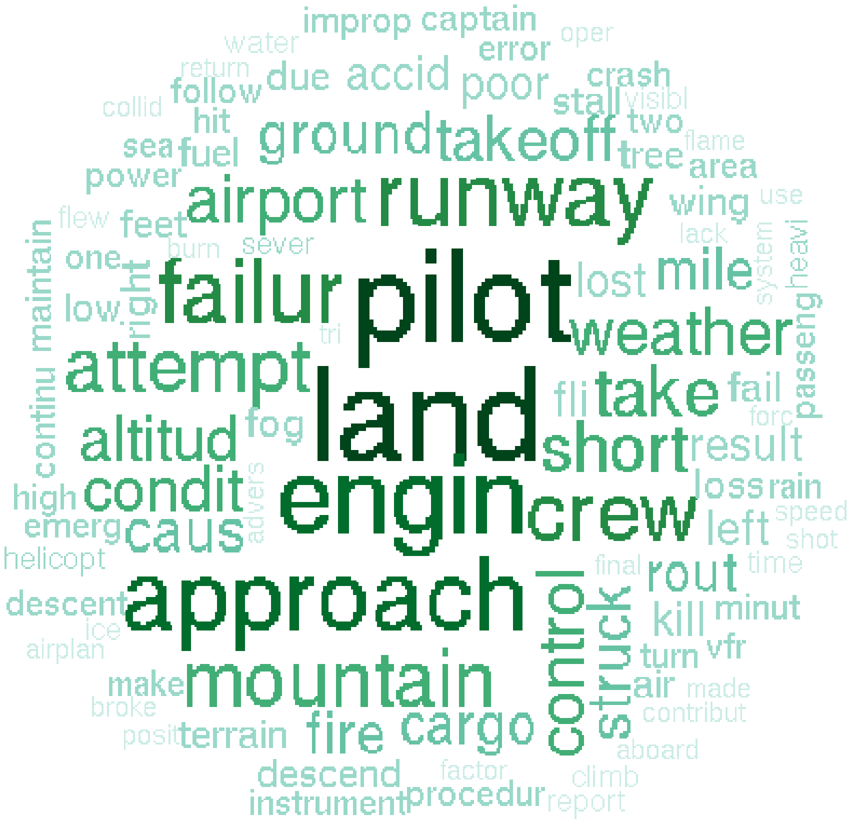


Figure 10. Word Cloud Graph for Summarizing Summary Column

4. One Sample Test

Let’s assume that I have two data series which are after 2000, number of death and before 2000 number of death.

Before ← c(1839, 2121, 1568, 1876, 1593, 2386, 1673,1544, 970)

After ← c(1398, 1437, 1276, 728, 1306, 1136, 931, 820, 577)

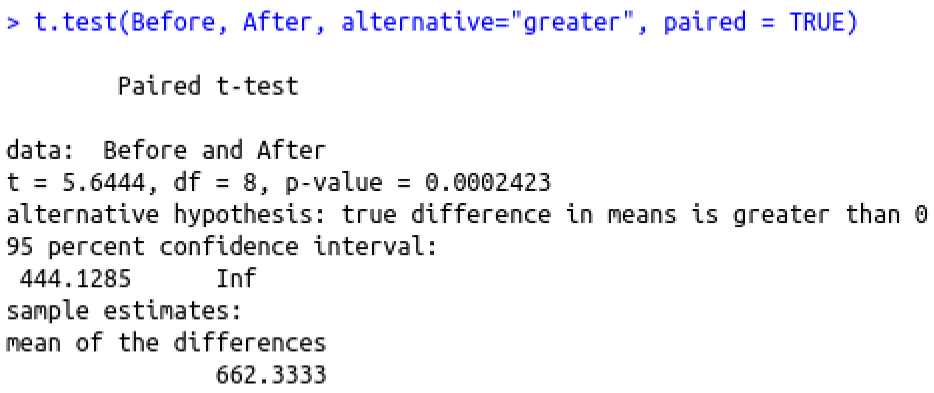
Logically, Security should increase in After dataset and number of death should decrease. Let’s test this hypotesis If it is true.

a) **Writing Hypothesis**

Ho: Mbefore = Mafter

Ha: Mbefore > Mafter

b) **Testing Hypothesis**



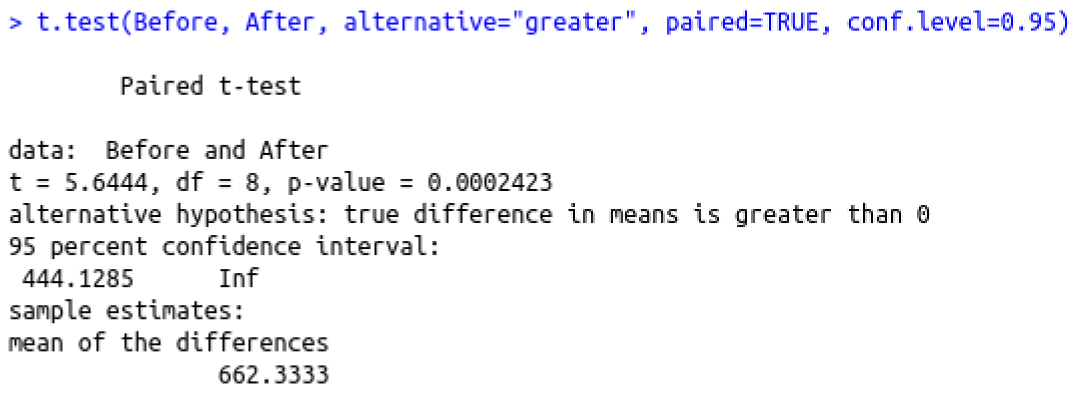
```
> t.test(Before, After, alternative="greater", paired = TRUE)

        Paired t-test

data:  Before and After
t = 5.6444, df = 8, p-value = 0.0002423
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 444.1285      Inf
sample estimates:
mean of the differences
               662.3333
```

p-value = 0.0002 <a =0.05   that means our hypothesis is not accept. Ho and Ha reject because p-value smaller than a(alpha).

c) Testing Hypothesis for 95% confidence level.



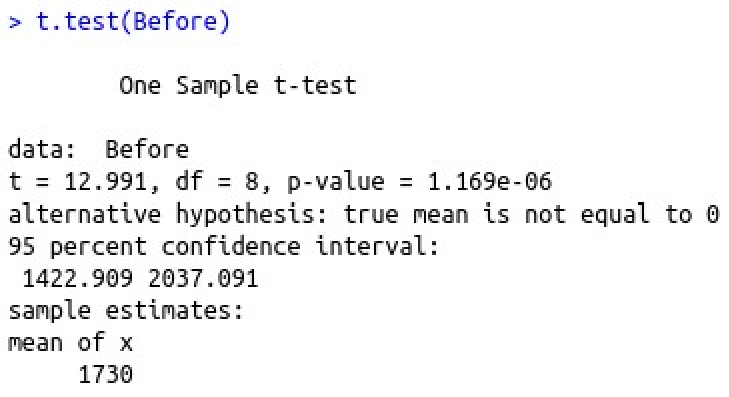
```
> t.test(Before, After, alternative="greater", paired=TRUE, conf.level=0.95)

        Paired t-test

data:  Before and After
t = 5.6444, df = 8, p-value = 0.0002423
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 444.1285      Inf
sample estimates:
mean of the differences
               662.3333
```

p-value = 0.0002 < a=0.05 that means our hypotesis is still so wrong. Ho and Ha reject because p-value smaller than our alpha value.

d) Finding confidence intervals of mean death numbers for these two datasets and comparing them (0.05)

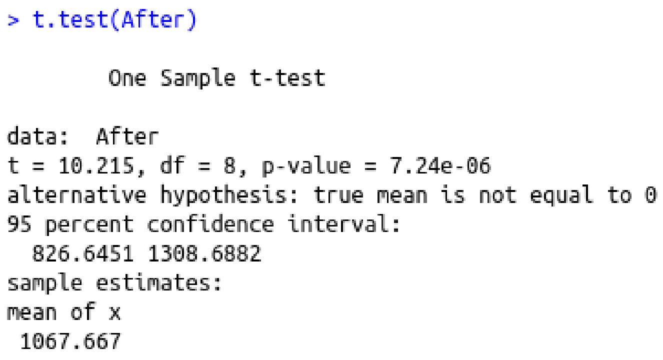


```
> t.test(Before)

        One Sample t-test

data:  Before
t = 12.991, df = 8, p-value = 1.169e-06
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 1422.909 2037.091
sample estimates:
mean of x
     1730
```

Confidence Intervals are different. There is not much difference between conf. Int.



```
> t.test(After)

        One Sample t-test

data:  After
t = 10.215, df = 8, p-value = 7.24e-06
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  826.6451 1308.6882
sample estimates:
mean of x
 1067.667
```

After data conf. İnt is lower than Before data. Years are not affecting properly.

5. Conclusions

This analysis provides insights in observing the trend of aircrash over the years. It shows the percent of fatalities observed due to the crash. The analysis also help in determining which airline operator and types are worst to fly with. We also observe the top 10 countries which we should avoid to esacpe the crash. All these topics will be addressed and analyzed.

1. Over the years the aircrash increased year until the decade 1968 - 1978. And then the number of crashes started reducing again, and it dropped considerably in the year 2008. The monthly crashes from 1908 were observed to check if any particular month was significantly responsible for the crash, but no such observation was made. Which implies that the crashes are well distributed through out the year.

2. It was observed that with time there is a decrease in the percent of fatalities. This might imply that constructive measures have been undertaken over the years for the safety of people on board.

3. A maximum number of aircrashes were observed as city in Moscow.

4. A maximum number of aircrashes were observed in Brazil. But There is not much difference between Brazil and Russia. Russia is making perfect sense why we should not trip there. Clearly because of the weather condition. The other countries that followed up were Russia, Colombia, USSR, France, India, China, Indonesia, Japan, Canada.

5. Aeroflot, Military - U.S. Air Force are worst operators as they have been responsible for maximum crashes.

6. Douglas Dc-3 types of aircraft are most prone to crashes.

7. Most crashes occured due to pilots, engine failures, approach, during take-off’s, on the runway, due to weather, mountains, land.

6. References

(1) Github, “vincentarelbundock/Rdatasets: An archive of datasets” Last Update March, 2016. https://github.com/vincentarelbundock/Rdatasets

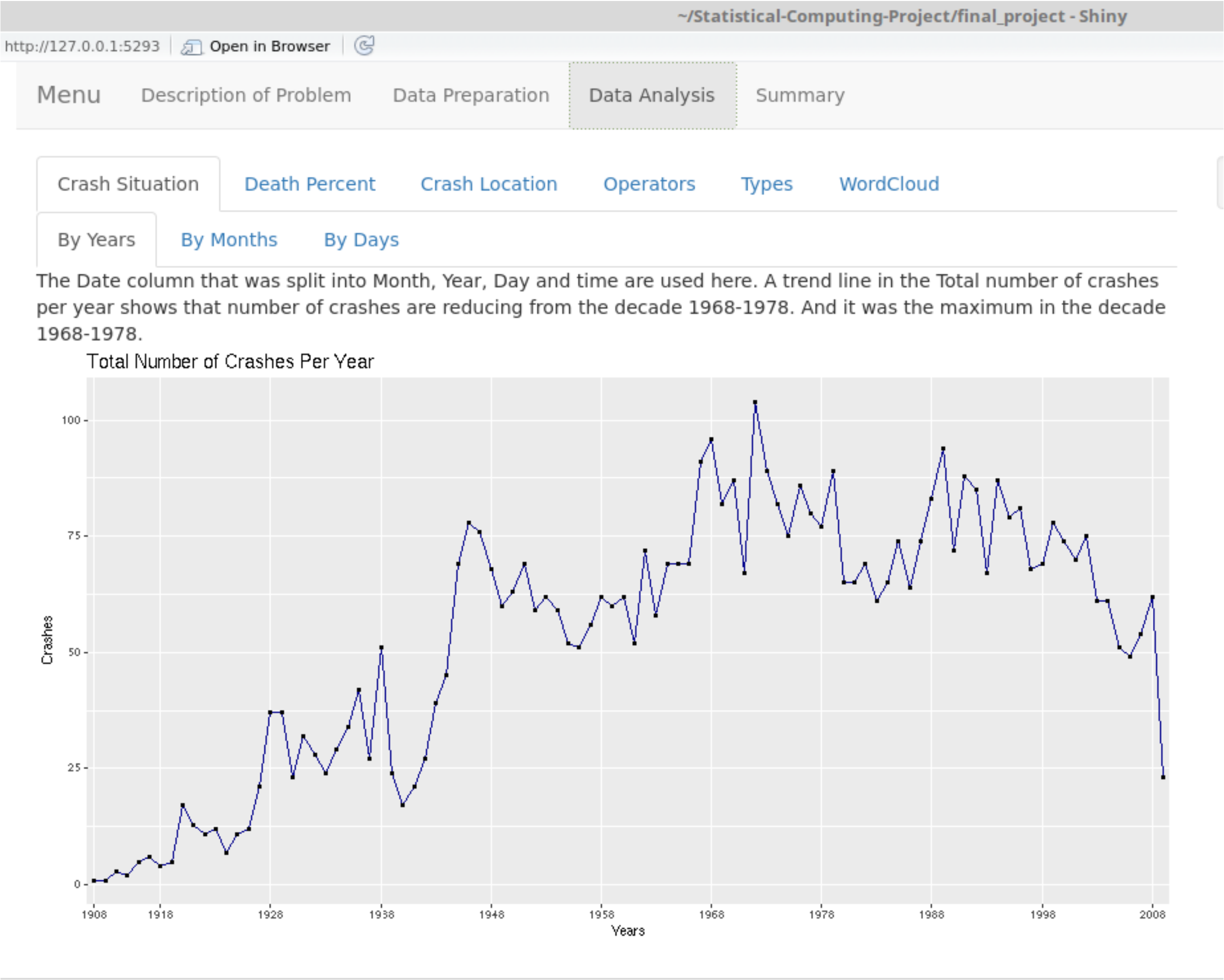
(2) Github, “vincentarelbundock/Rdatasets: Aircraft Crash Data” Last Update March, 2016. https://vincentarelbundock.github.io/Rdatasets/doc/gamclass/airAccs.html

(3) Github, “Plain Crash Info” Last Update March, 2016. http://www.planecrashinfo.com/reference.htm

Note : You can access all codes from my github account. It is shiny app.

Github: https://github.com/demirmeltem/Statistical-Computing-Project

Shiny Site look:



The Date column that was split into Month, Year, Day and time are used here. A trend line in the Total number of crashes per year shows that number of crashes are reducing from the decade 1968-1978. And it was the maximum in the decade 1968-1978.