AIRCRAFT CRASH DATA ANALYSIS

Statistical Computing
Data Analysis Report

Student
Name
Meltem DEMİR

Project Title
Aircraft Crash Data Analysis

Apr 16, 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?"

Description of the problem 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. To do this, I will determine the factors according to types, specially operators and locations.

2. Description of Data

I found my data from someone's Github account who is Vincen Tarel Bundock. *This address* (1) includes plenty of R Datasets. My data frame is including with 5666 observations on the following 7 variables which are Date, location, operator, planeType, Dead, Aboard, Ground. Format of my dataset is in .CSV. If I give you detail about columns, here is it:

- Date - Date of Accident
- location - Location of accident
- operator - Aircraft operator
- planeType - Aircraft type
- Dead - Number of deaths
- Aboard - Number aboard
- Ground - Deaths on ground

3. Progress to Date

✓ Firstly, I read the airAccs.csv file to see in table format.

```
air.crash <- read.csv("/home/demirmeltem/Desktop/statictical-rapor1/airAccs.csv")
```

✓ To show the first six row, I use this command

```
head(air.crash)
```

✓ I changed the column names to be more descriptive.

```
colnames(air.crash) <- c("Number" ,"Accident.Date", "Accident.Location", "Aircraft.Operator", "Aircraft.Type", "Number.of.Dead", "Number.Aboard", "Deaths.on.Ground")
```

✓ I converted the variables into the correct format (numeric, integer, character, factor, ...)

```
air.crash[, 2] <- as.Date(air.crash[,2], format = "%Y-%m-%d")
air.crash[, 3] <- as.character(air.crash[, 3])
air.crash[, 4] <- as.character(air.crash[, 4])
air.crash[, 5] <- as.character(air.crash[, 5])
```

✔ I removed the unnecessary rows and columns in air.crash data frame.

```
air.crash <- air.crash[!is.na(air.crash[,2:8]),]
```

✔ I created a new data.frame which is air.crash.private. It is according to Operator types. I subset the private operator.

```
air.crash.private <- subset(air.crash, air.crash$Aircraft.Operator == "Private")
```

✔ I drawn a histogram to show number of deads in private operators.

```
ggplot(air.crash.private, aes(x=Number.of.Dead)) + geom_histogram()
```

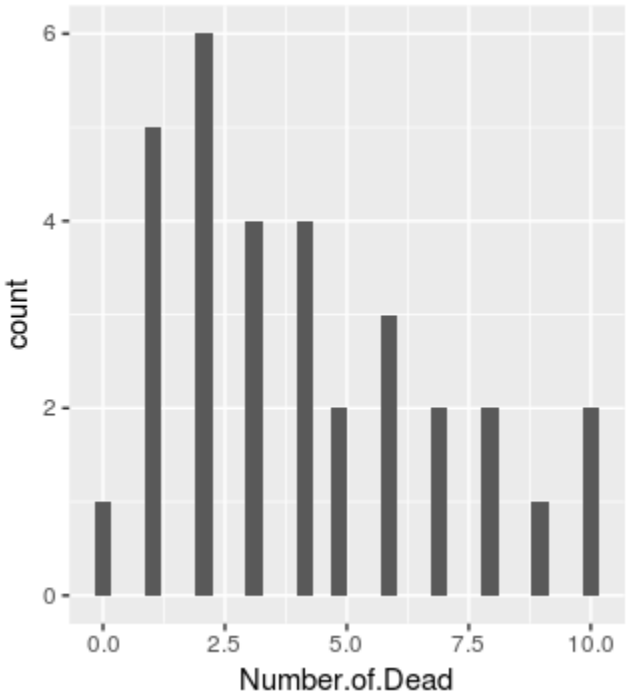

Figure 1. Histogram

✔ I drawn a scatter plot to show number of deads according to years.

```
ggplot(air.crash, aes(x=Accident.Date, y=Number.of.Dead)) + geom_point()
```

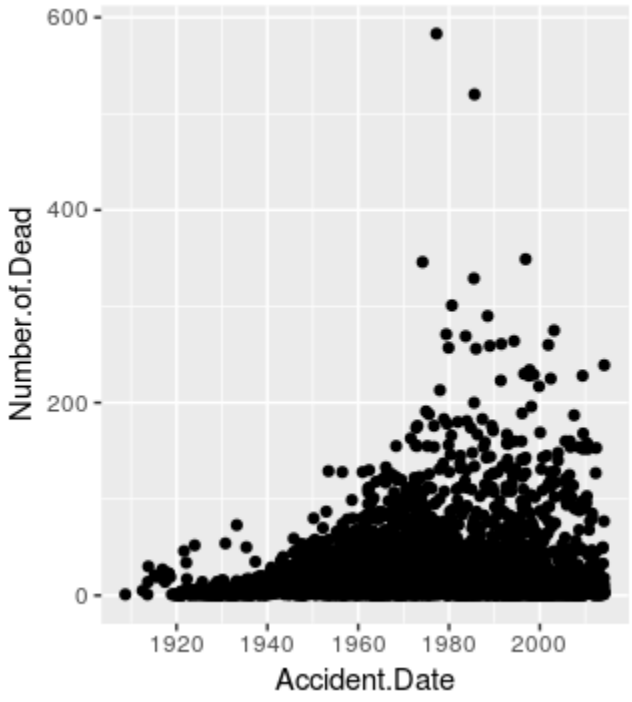

Figure 2. Scatter Plot

```
ggplot(air.crash, aes(x=Accident.Date, y=Number.Aboard)) + geom_point()
```

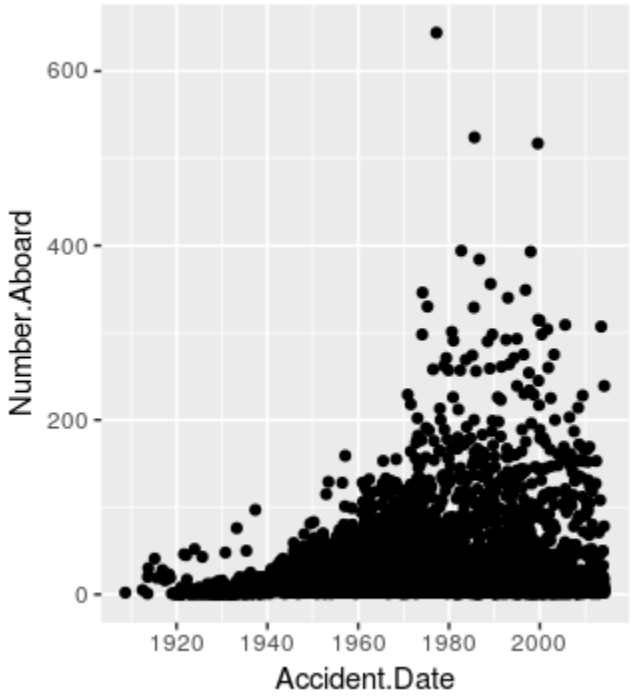

Figure 3. Scatter Plot

✔ I drawn a scatter plot to show number of deads according to aircraft type

```
ggplot(air.crash.private, aes(y=Aircraft.Type, x=Number.of.Dead)) + geom_point()
```

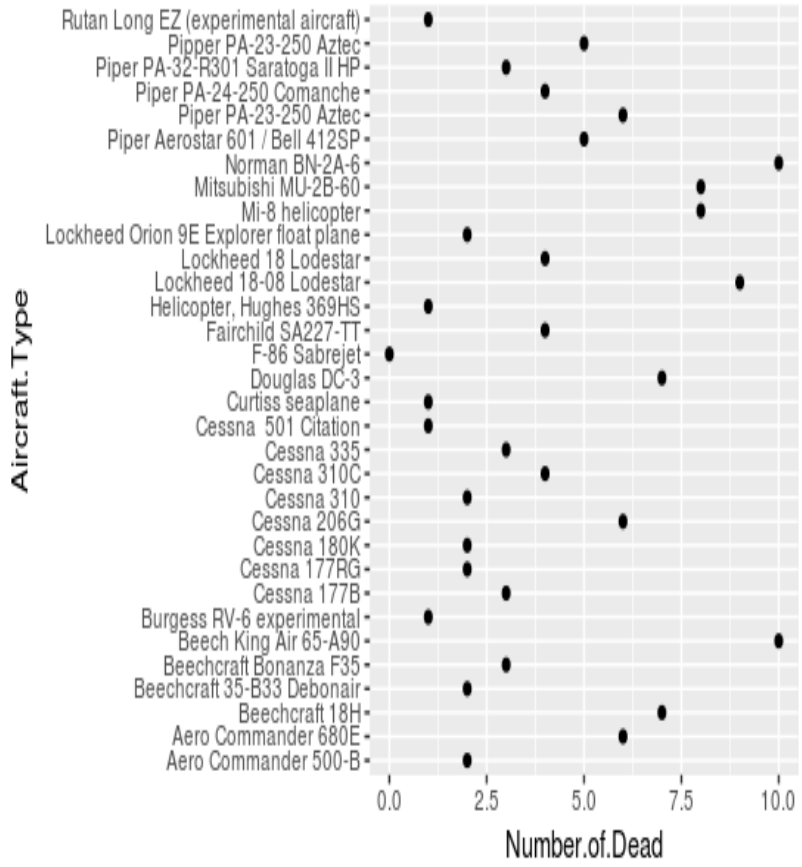

Figure 5. Scatter Plot

From now, I will do more ggplot with more detaile. These ggplot's is just for preview and to see what I can do more. For now, this is the progress that I made.

4. Conclusions

After this progress, I should do some experiments on data. Some methods that I will apply;

- Plotting (histogram, barplot) with detail.
- Sd(standart deviation), median, mean
- Correlation

This is it for now. I am in progress and soon I will have more solid results.

5. 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