**VISVESVARAYA TECHNOLOGICAL UNIVERSITY, BELGAUM 590014**

****

Data ScienceReport on

**“Air Crash Analysis”**

By

DIVIJ ASIJA (1BM14CS037)

HEMANTH P (1BM14CS121)

DARSHAN R P (1BM14CS035)

Under the Guidance of

**Prof. Nagarathna N**

Assistant Professor, Department of CSE

BMS College of Engineering

Data Science carried out at

****

Department of Computer Science and Engineering

BMS College of Engineering

(Autonomous college under VTU)

P.O. Box No.: 1908, Bull Temple Road, Bangalore-560 019

2017-2018

**BMS College of EngineerinG**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

****

***CERTIFICATE***

This is to certify that the data science assignment titled “**Air Crash Analysis**” has been carried out by DIVIJ ASIJA(1BM14CS037), DARSHAN R P (1BM14CS035), HEMANTH P (1BM14CS121) during the academic year 2017-2018.

Signature of the guide Signature of the external **Prof. Nagarathna N**  invigilator

Assistant Professor

Department of Computer Science and Engineering

BMS College of Engineering, Bangalore

**BMS College of EngineerinG**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

****

***DECALARATION***

We, DIVIJ ASIJA (1BM14CS036), DARSHAN R P (1BM14CS035), HEMANTH P(1BM14CS121) , students of 7th Semester, B.E, Department of Computer Science and Engineering, BMS College of Engineering, Bangalore, hereby declare that, this Data Science analysis work entitled " **Air Crash Analysis** " has been carried out by us under the guidance of Prof. Nagarathna N, Assistant Professor, Department of CSE, BMS College of Engineering, Bangalore during the academic semester Jan-April 2017.

We also declare that to the best of our knowledge and belief, the development reported here is not from part of any other report by any other students.

Signature

DIVIJ ASIJA (1BM14CS037)

DARSHAN R P (1BM14CS035)

HEMANTH P (1BM14CS121)

**CONTENTS**

**1. Introduction 1**

**2. Design 2**

**3. Implementation 3**

**4. Screenshots and results 5**

**5. Conclusions 12**

**6. References 13**

1. **Introduction**

The main goal of our assignment is to analyze the air crashes that took place during the 20th century. We want to compare the different operators and plane types and determine which was the most unsafe airplane amongst all of them. We have used text analysis and linear regression to try and identify the dangerous air crafts. We also have compared the number of crashes that occurred as time passed and technology improved.

**SYSTEM REQUIREMENTS**

* Windows 8 or above
* R Studio 3.4.2 or above
* Dataset URL (dating.csv)

*https://vincentarelbundock.github.io/Rdatasets/datasets.html*

**KEY TO DATASET**

id : The id of the crash

Date: the date on which the accident took place

Location : The cities between which the airplane was travelling

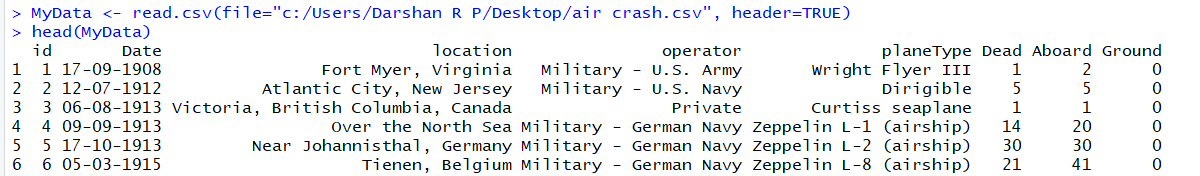
Operator : The operator of the plane

PlaneType: The type of plane that was flying during the accident

1. **Design**

The number of people that died in accidents of a particular plane type was analyzed and the most dangerous plane type was identified. Similarly, the most dangerous operator was identified. The frequency of plane crashes based on the variables such as location, date and plane type were analyzed using linear regression and plots between these variables. We have mainly used text analysis to identify the fields we want and have checked if they satisfy some conditions such as maximum number of deaths caused to identify the dangerous air crafts. Plots have been drawn to represent the same.

**Sample data set:**

****

**Exploring Co-relations between columns**

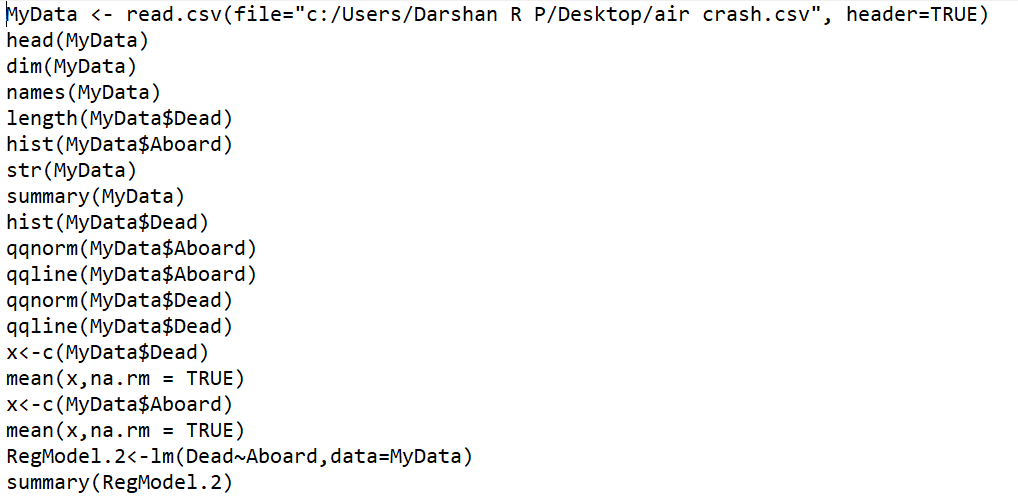
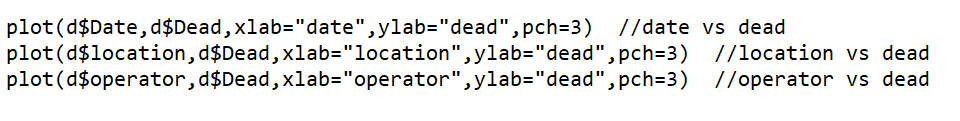
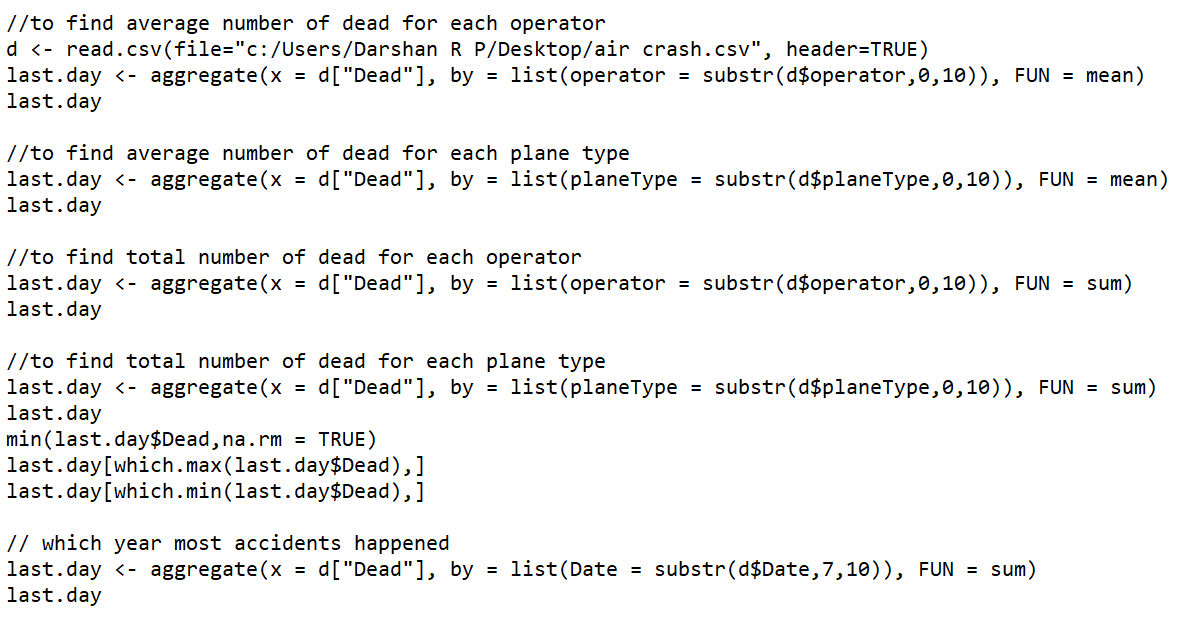
Library ‘car’ was used to generate a cumulative display of dependencies within columns. Plots for the same have been provided later.

**Specific Queries**

The main aim was to see if the improvement in technology had an effect on the number of air crashes. Also we had to determine the most dangerous airplane based on the variables available.

**IMPLEMENTATION:**

We have mainly used descriptive analysis, which is a form of text analysis to answer a few questions. The following R code was executed to get the results and plots.



install.packages ("car", dep=T)

library(car)

//vif(RegModel.2)

outlierTest(RegModel.2) //removes out of bounds values

//MyData2<-MyData[-c(369,372,373),]

//RegModel.2<-lm(MV~B+CHAS+CRIM+DIS+LSTAT+NOX+PT+RAD+RM+ZN,data=MyData2)

//summary(RegModel.2)

RegModel.3<-lm(Dead~Aboard+Ground,data=MyData)

summary(RegModel.3)

vif(RegModel.3) //variance

mean(RegModel.3$residuals)

hist(RegModel.3$residuals,xlab="residuals",main="Histogram of Residuals")

qqnorm(RegModel.3$residuals,main="Normal Probability Plot",pch=19)

qqline(RegModel.3$residuals)

z<-c(RegModel.3$residuals)

z1<-z[0:5666]

plot(MyData$Dead,z1,main="Residuals vs. Predictor",xlab="Dead", ylab="Residuals",pch=19)

abline(h=0)

plot(MyData$Aboard,z1,main="Residuals vs. Predictor",xlab="Average number of people aboard", ylab="Residuals",pch=19)

abline(h=0)

a<-c(MyData$Dead)

b<-c(a,na.rm=TRUE)

var(b,y=NULL,na.rm=TRUE)

//plot(MyData2$PT,RegModel.3$residuals,main="Residuals vs. Predictor",xlab="Pupil/Teacher Ratio", ylab="Residuals",pch=19)

//abline(h=0)

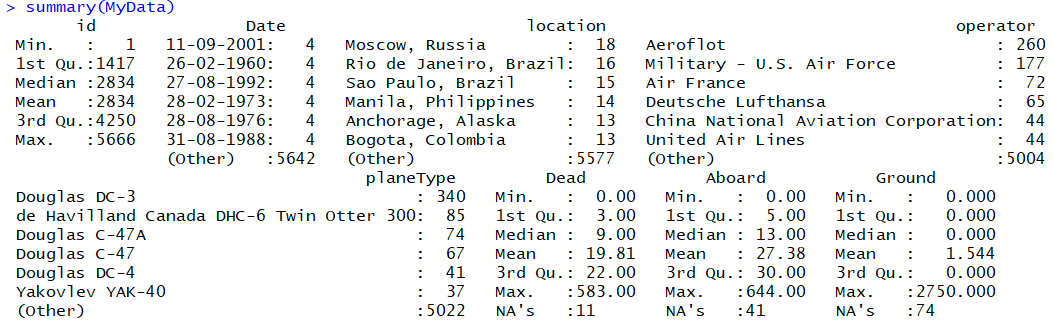
//plot(RegModel.3$fitted.values,RegModel.3$residuals,main="Residuals vs. Fitted",xlab="Fitted Vaues", ylab="Residuals",pch=19)

//abline(h=0)

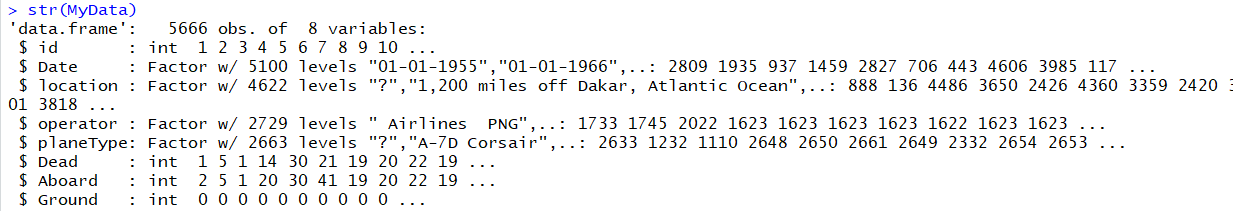
anova(RegModel.2,RegModel.3) //to compare 2 reg models

**SCREENSHOTS OF PLOTS AND RESULTS:**

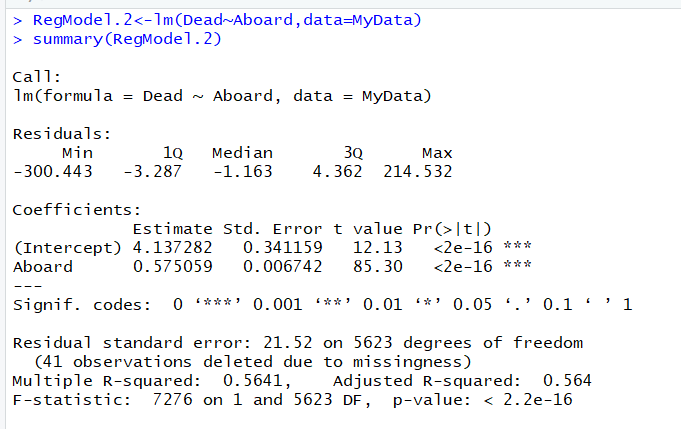
**Summary of the given data:**

****

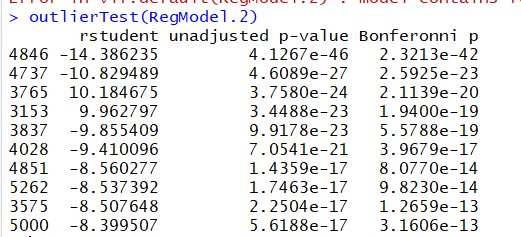
**To view the dataset as a string**

****

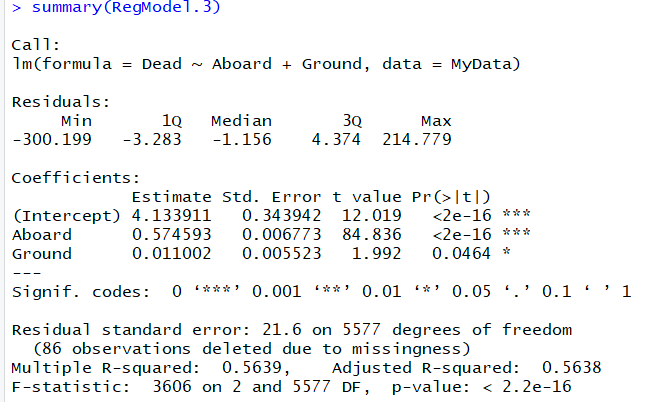
**To create a linear regression between the number of people dead and the number of people aboard the airplane:**

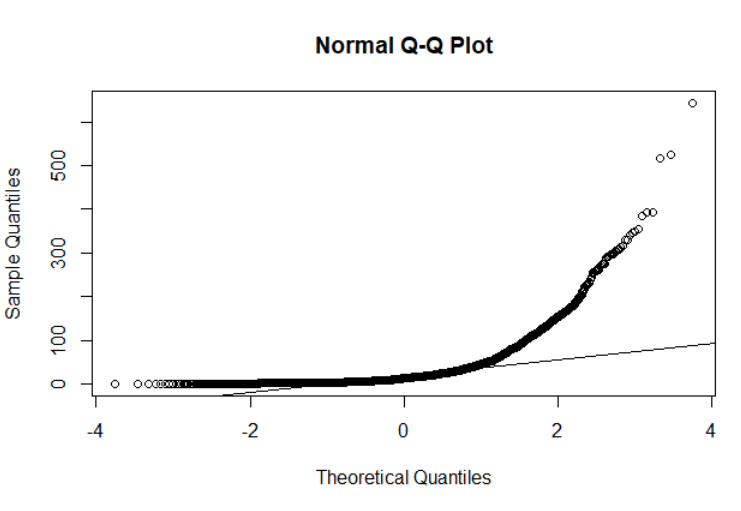
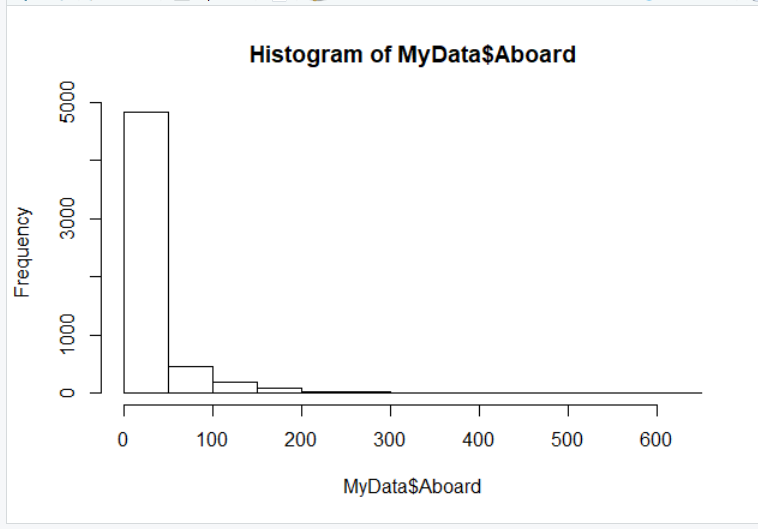
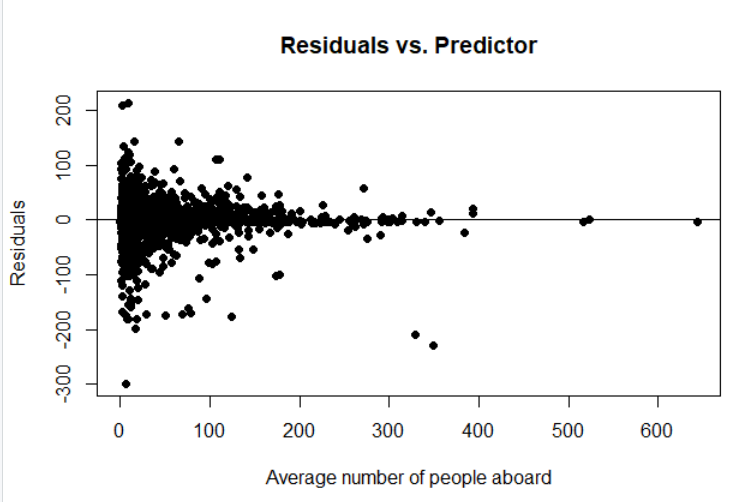
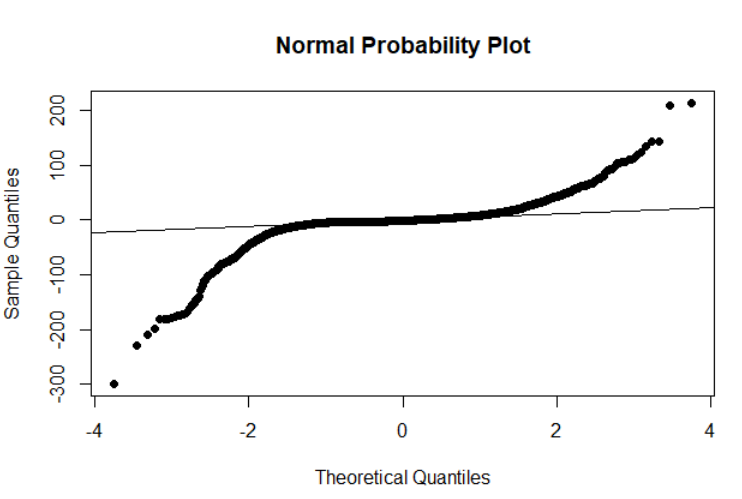
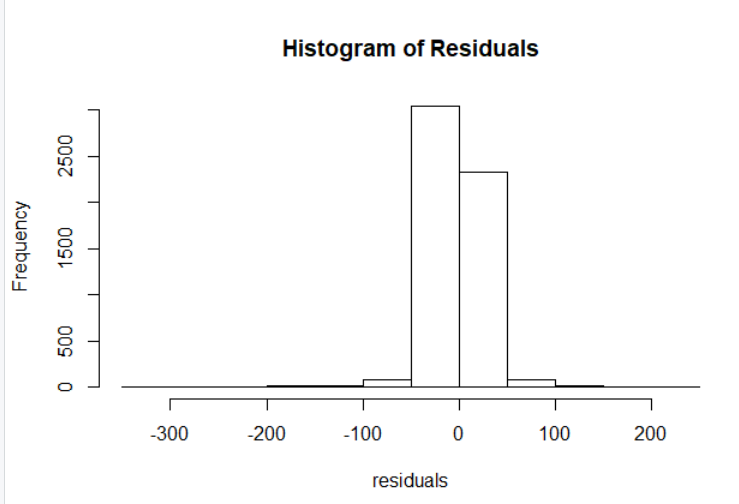
****

**To view the outliers:**

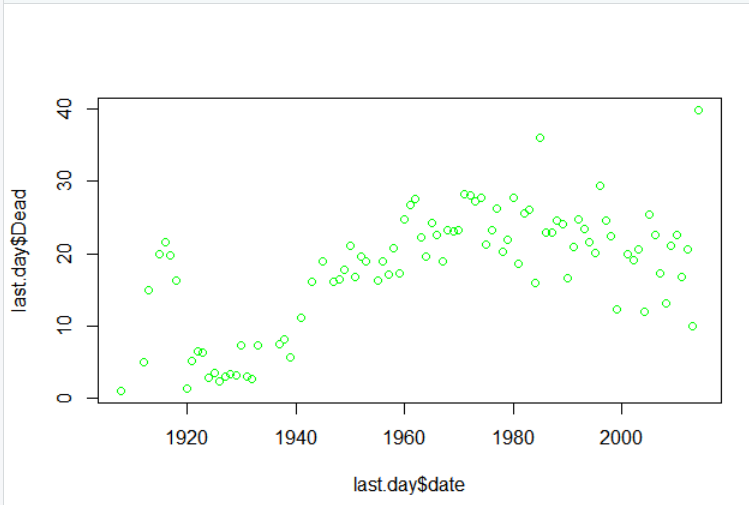
****

**Summary of regression model:**

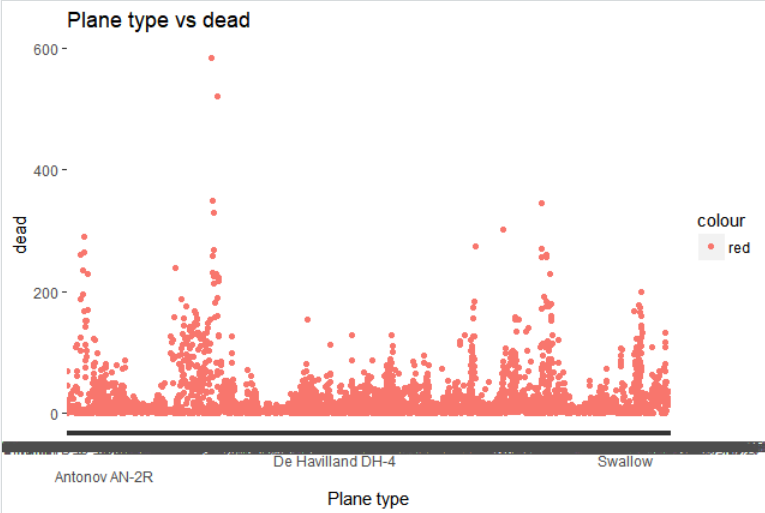
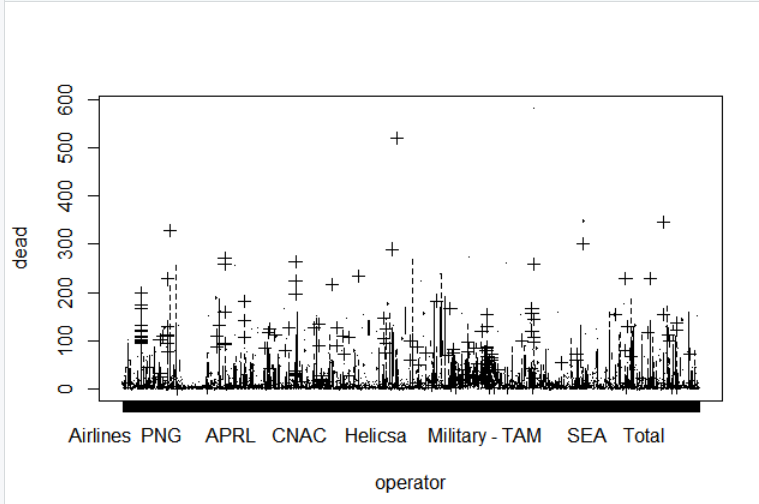
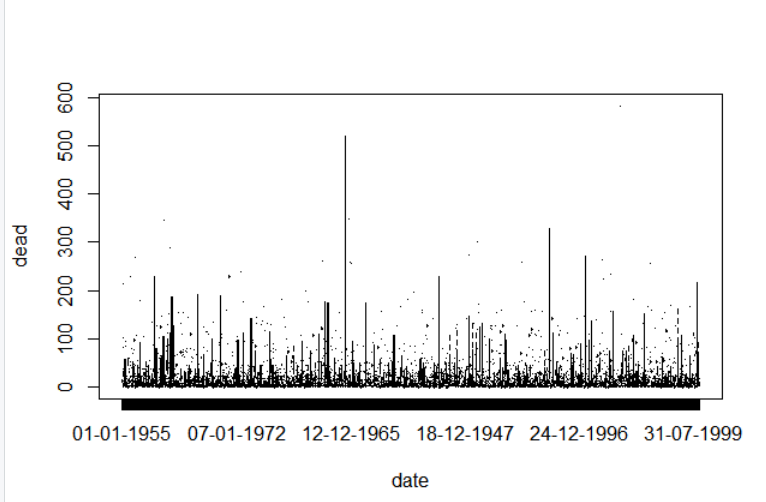
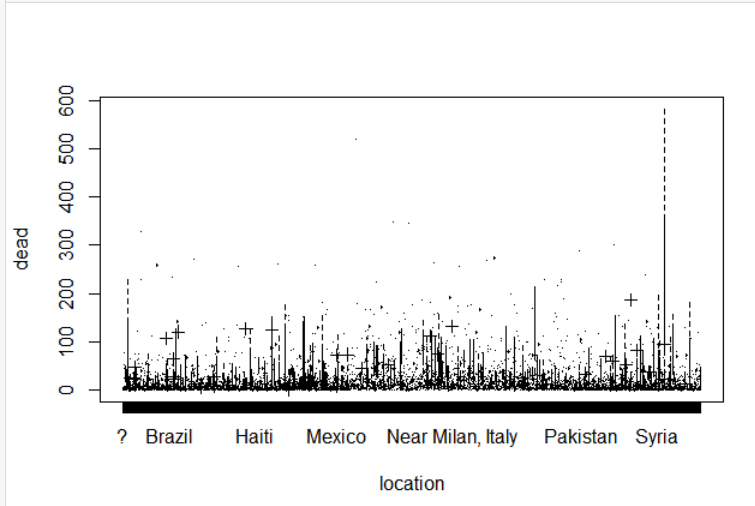
****

****

**Plot to see average number of deaths per year when accidents happen**

****

**To determine the most dangerous location and year:**

****

**CONCLUSIONS:**

* The most dangerous plane type and operator was determined.
* It was observed that Syria was the place where the most number of accidents occurred and a possible reason for this could have been the fact that there was a lot of firing in that particular region.
* We would assume that the number of deaths in accidents would reduce from 20th to 21st century but this is not true based on our analysis. A possible reason for this could be that the accidents that occur these days are less probable but too severe and causes more deaths.
* Text analysis proved to be efficient in performing analysis where numbers and strings were concerned.
* It is easier to depict results using graphs since they combine a lot of information and provide it in a simple and understandable manner.
* There are usually a lot of outliers that occur, and if not removed can affect the calculations.

**REFERENCES**

1.Stack Overflow - https://stackoverflow.com/questions/17888764/r-finding-rows-of-a-data-frame-where-certain-columns-match-those-of-another

2.Kaggle - https://www.kaggle.com/

3.R documents - https://www.r-project.org/other-docs.html

5.Google -https://www.google.co.in/?gfe\_rd=cr&dcr=0&ei=hncXWtCcE4Wh8wfKm4i4Cg

6. Tutorial Point - https://www.tutorialspoint.com/r/r\_decision\_tree.htm