**Statistics Data Mining**



**Report on Analysis of Aircraft Crashes**

**TABLE OF CONTENTS**

PROBLEM STATEMENT……………………………………………………………………………………….. 3

DATASET DESCRIPTION………………………………………………………………………………………. 3

APPROACH…………………………………………………………………………………………………………. 4

DATA CLEANING………………………………………………………………………………………… 4

DATA VISAUALIZATION IN R AND TABLEAU……………………………………………… 4

NAÏVE BAYES CLASSIFICATION………………………………………………………………….10

FINDING TRENDS IN YEAR WISE FATALITIES RATE…………………………………….12

LINEAR REGRESSION……………………………………………………………………… 12

BOOT STRAPPING……………………………………………………………………………13

ANALYZING BY FINDING DIFFERENT CLUSTERS…………………………………………14

CONCLUSION……………………………………………………………………………………………………..22

**PROBLEM STATEMENT**

* Find a pattern related to the cause of aircraft crashes using clustering.
* Predict the risk categories (Low/High) for fatalities in a crash by considering aircraft type and number of incidents.
* Visualize the data and understand the past changes over the years for airline types.

**DATASET DESCRIPTION**

One Dataset is provided consist of data from last 100 years that is from 1908 to 2008. This data contains the data about the

* Date: Date on which Aircraft was crashed.
* Year: Year on which Aircraft was Crashed.
* Time: Time of Plane crash.
* Location: At which location, aircraft was crashed.
* Operator: who was the operator at the time of crash.
* Flight#: Flight number.
* Route: Route.
* Type: Aircraft Type which was crashed.
* Registration
* Aboard: Number of people in the plane at time of crashed.
* Fatalities
* Death Percentage
* Ground
* Summary: Reason of Plane crash

**APPROACH**

* Data Cleaning (huge task)
* Data Visualization in R and Tableau
* Naïve Bayes classification to predict the risk categories (Low/High) i.e., likelihood of people surviving in a crash given a particular aircraft type and other factors.
* Finding trends in year wise fatalities rate.
* Analyzing summary of the airplane crashes from the dataset by identifying different clusters.

**DATA CLEANING**

* Change the format of Date field in Dataset.
* Handle the missing values in the dataset.
* Using text to column option in Excel, we have separated country names from Location Field.
* Using text to column option in Excel, we have separated Airline type name from Type column.

**DATA VISUALIZATION**

* **Created the Word Cloud to extract the frequent occurring word to get the gist of cause of plane crash.**

**Below we can find the R code for the word cloud:**

## Word Cloud from summary column

aircrash<-read.csv(file.choose(), header = TRUE, sep=',')

data <- Corpus(VectorSource(aircrash$Summary))

data1 <- tm\_map(data, tolower)

data1 <- tm\_map(data1, removeNumbers)

data1 <- tm\_map(data1, removeWords, stopwords())

data1 <- tm\_map(data1, removePunctuation)

data1 <- tm\_map(data1, PlainTextDocument)

data1 <- tm\_map(data1, removeWords, "led")

data1 <- tm\_map(data1, removeWords, "flight")

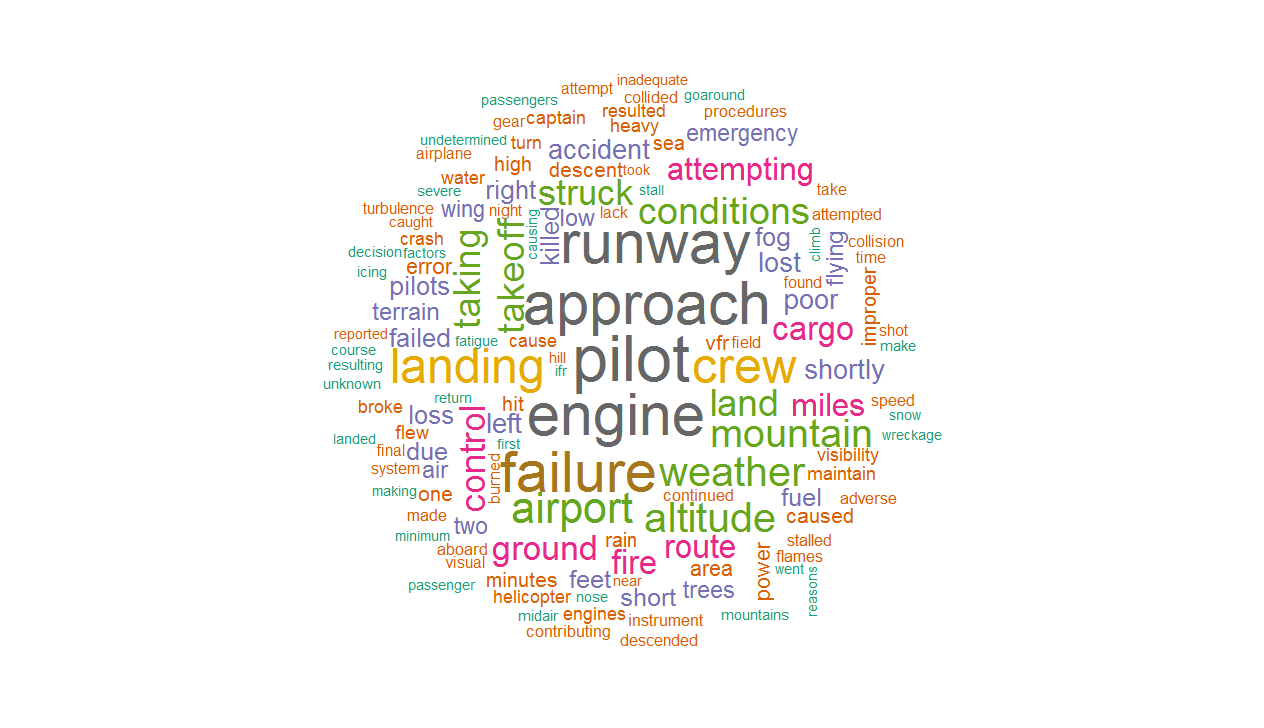
data1 <- tm\_map(data1, removeWords, "crashed")

data1 <- tm\_map(data1, removeWords, "plane")

data1<- tm\_map(data1, removeWords, "aircraft")

wordcloud(data1, min.freq = 100, random.order = FALSE, colors=brewer.pal(8, "Dark2"))

**Plot of Word Cloud:**



The above R code was used to create the word cloud. It contain words which have been used more than a 100 times . This word cloud gives the gist of summary coloumn, some obvious words and some other words such as mountain, fire, engine, shot, trees and more which can give the possible causes for the crash.

* **Data Visualization using R**

**Year wise, month wise and day wise incidents**

crashes <- read.csv(file.choose())

head(crashes)

crashes$date\_new<-as.Date(crashes$Date,"%d/%m/%Y")

crashes$Month <- month(crashes$date\_new)

head(crashes$Month)

crashes$Year <- year(crashes$date\_new)

head(crashes$Year)

crashes$Day <- day(crashes$date\_new)

head(crashes$Day)

## year wise plot containing number of crashes

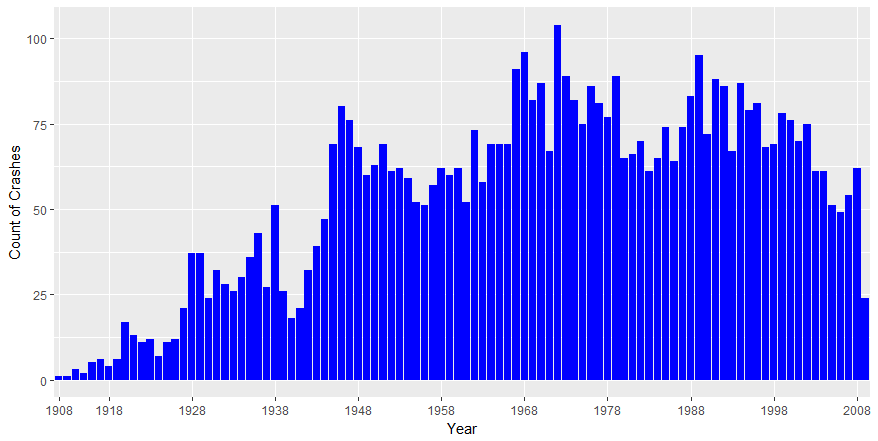
years <- as.data.frame(table(crashes$Year))

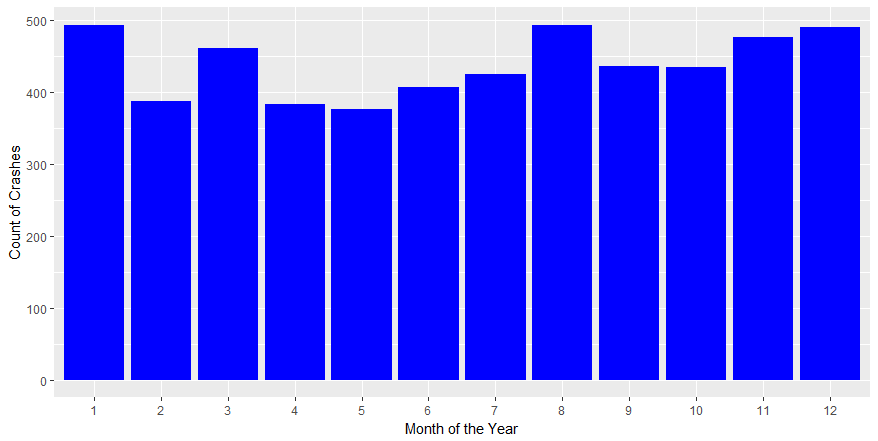
head(years)

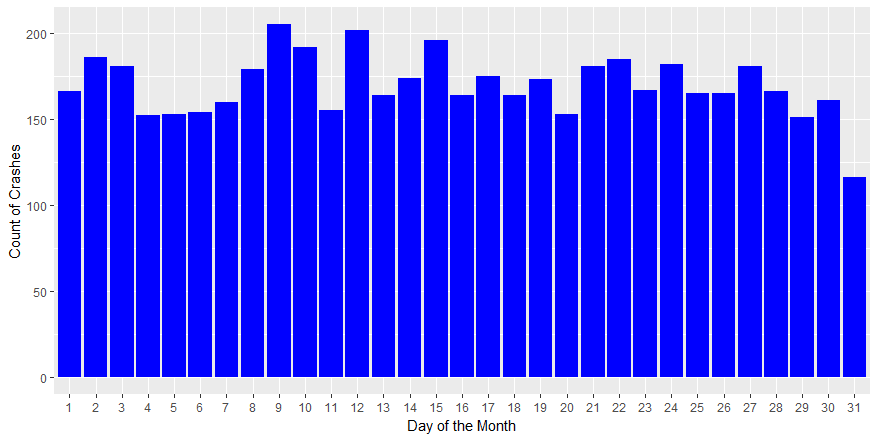
ggplot(years, aes(Var1, Freq)) + geom\_bar(stat = "identity", fill = "blue") + xlab("Year") +

ylab("Count of Crashes") + scale\_x\_discrete(breaks = seq(from = 1908, to = 2009, by = 10))

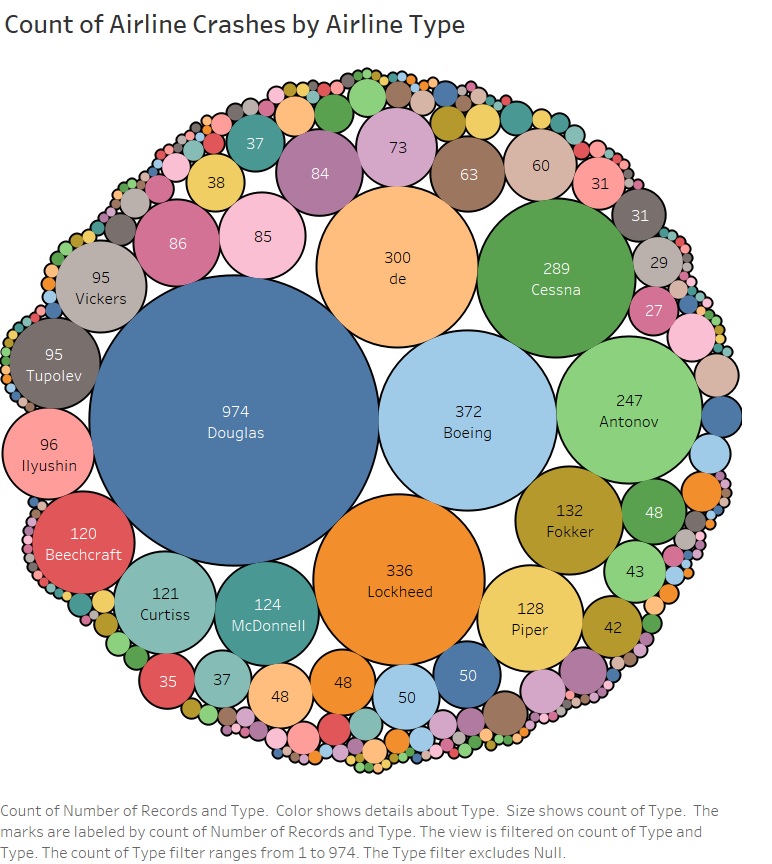
The following trend we can see:



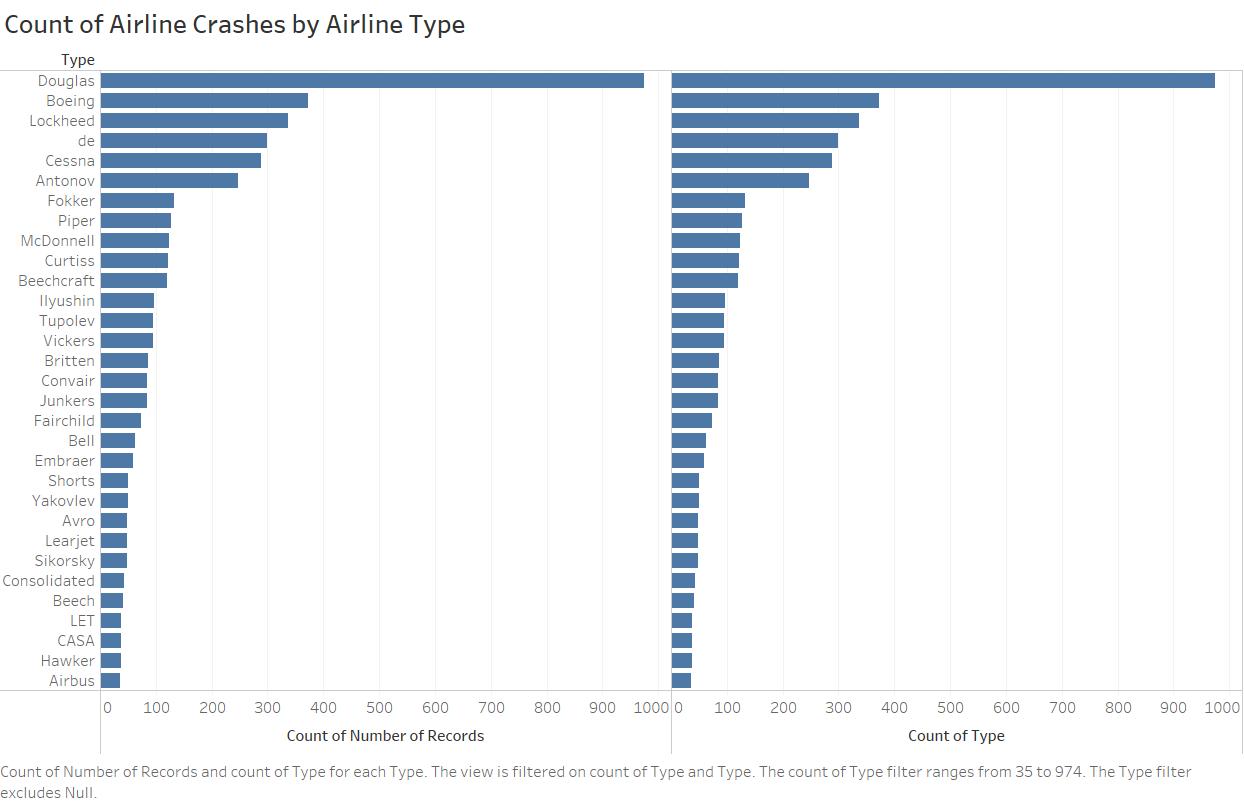




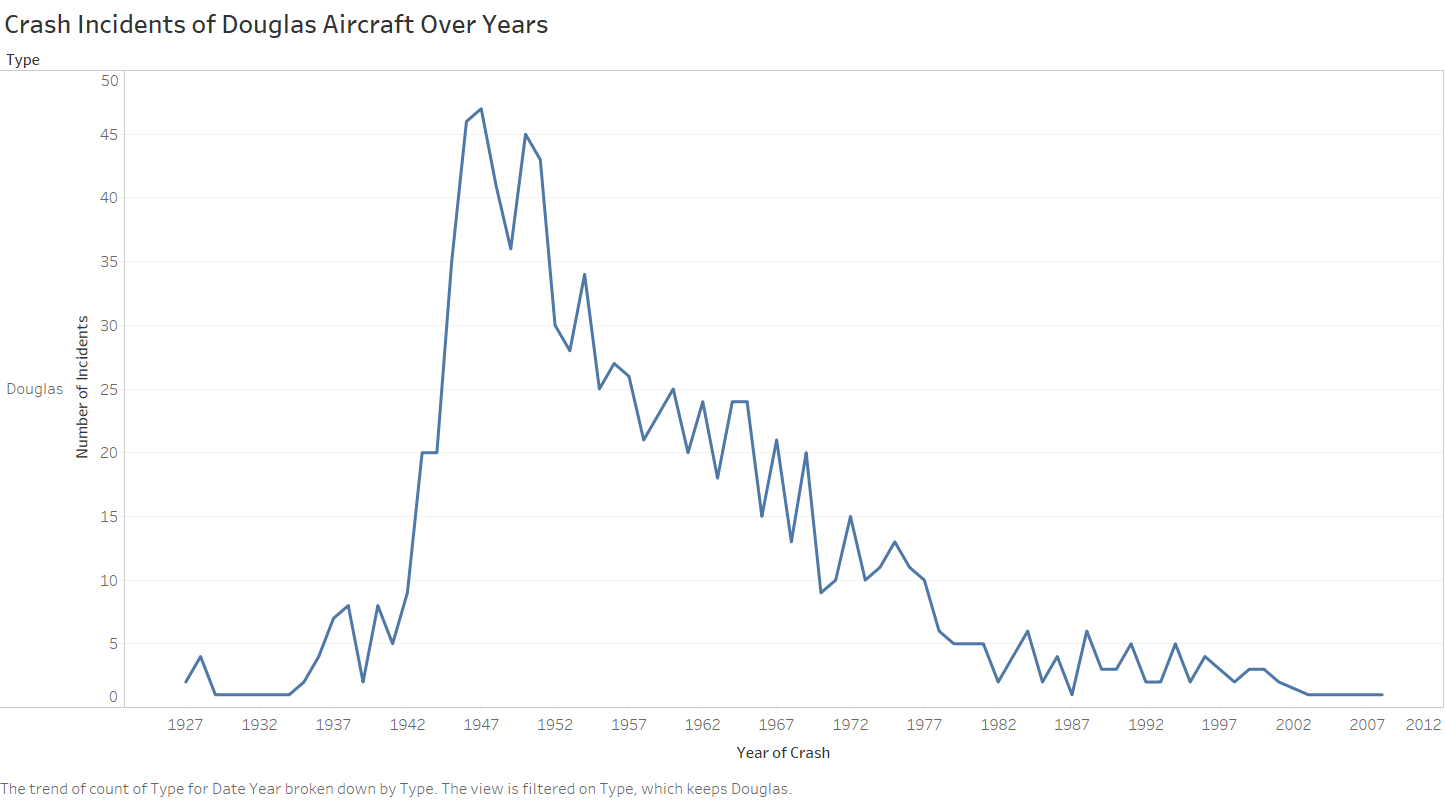
* **Data Visualization using Tableau**



**Count of Airline Crashes by Airline Type**



**Crash Incidents of Douglas Aircraft over years**



NAIVE BAYES CLASSIFICATION TO PREDICT THE RISK CATEGORIES

We are applying Bayesian Classification on Risk Category Datato predict the risk categories (Low/High) i.e., likelihood of people surviving in a crash given a particular aircraft type and other factors.

**R code**

library ("klaR")

library ("caret")

Risk\_Category= read.csv(file.choose())

Risk\_Category$Risk.Category<-as.factor(Risk\_Category$Risk.Category)

head(Risk\_Category)

ShuffledRisk\_Category <-Risk\_Category[sample(nrow(Risk\_Category)),]

train <- ShuffledRisk\_Category[1:200,]

test <- ShuffledRisk\_Category[201:267,]

model <- NaiveBayes(Risk.Category ~ ., data=train)

#test the model

predictions <- predict(model, test)

warnings()

confusionMatrix(test$Risk.Category, predictions$class)

**R Code Output**

> confusionMatrix(test$Risk.Category, predictions$class)

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 59 3

1 3 2

Accuracy : 0.9104

95% CI : (0.8152, 0.9664)

No Information Rate : 0.9254

P-Value [Acc > NIR] : 0.768

Kappa : 0.3516

Mcnemar's Test P-Value : 1.000

Sensitivity : 0.9516

Specificity : 0.4000

Pos Pred Value : 0.9516

Neg Pred Value : 0.4000

Prevalence : 0.9254

Detection Rate : 0.8806

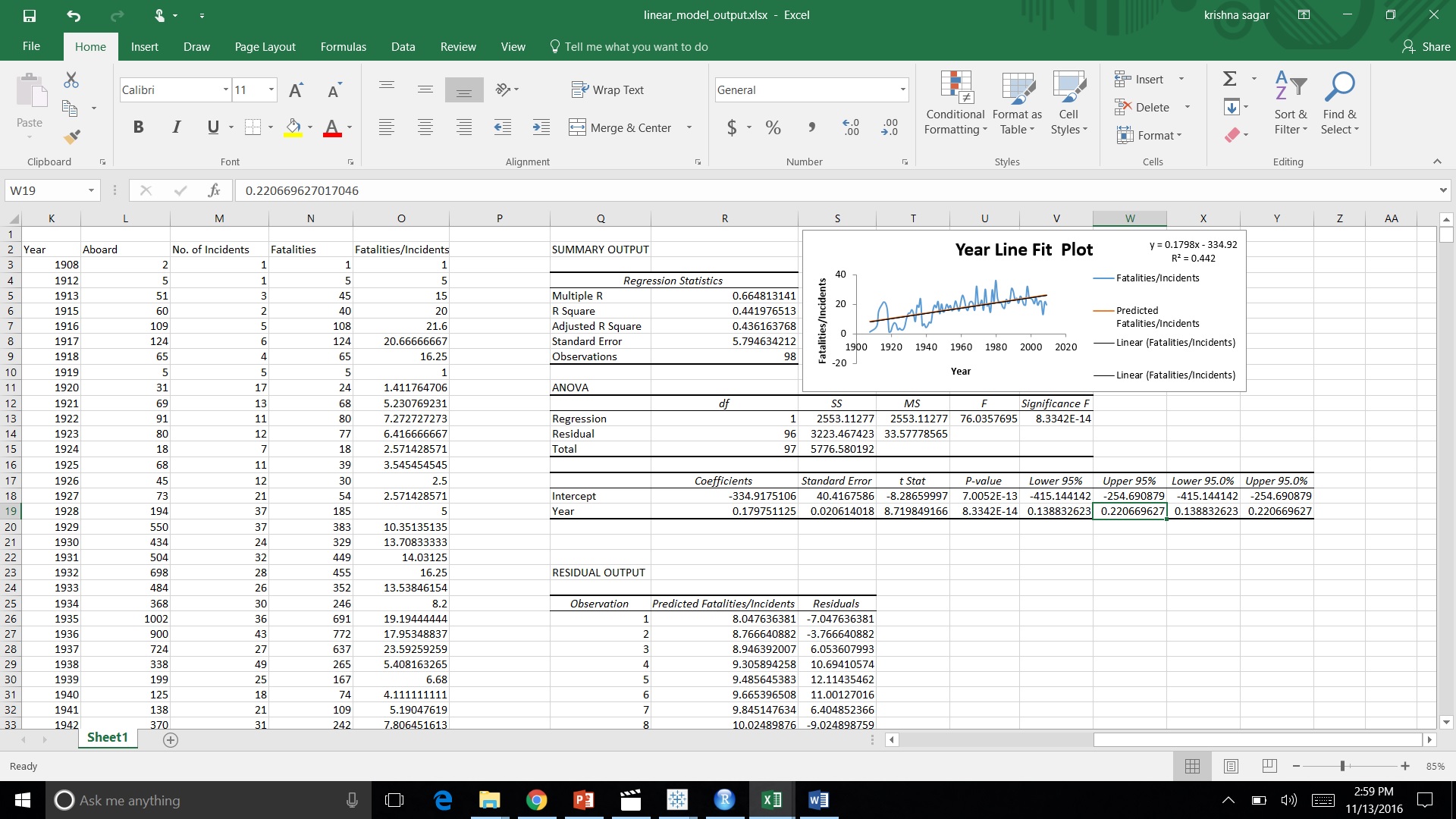
Detection Prevalence : 0.9254

Balanced Accuracy : 0.6758

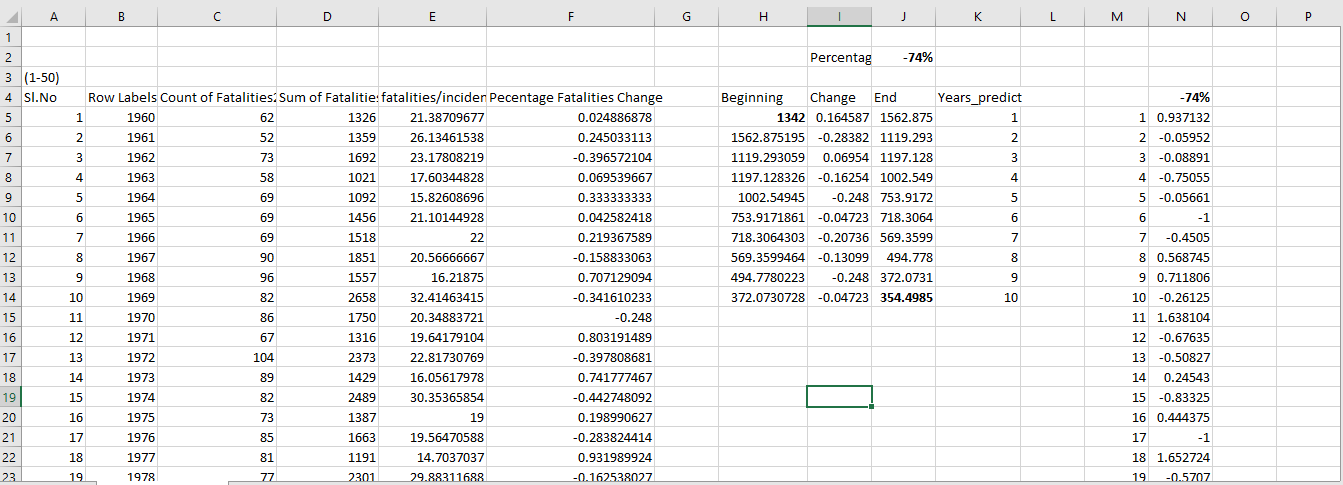
'Positive' Class : 0

FINDING TRENDS IN YEAR WISE FATALITIES RATE

* **We have used Linear Regression to identify any trend in the Fatalities with Year. Below we have computed the Linear Regression in Excel.**



* **BOOTSTRAPPING**



ANALYZING BY IDENTIFYING DIFFERENT CLUSTERS

Here we are loading libraries in R :

*# Load libraries*

**library**('tm')

**library**('dplyr')

**library**('stringr')

**library**('tidyr')

**library**('factoextra')

**library**('ggplot2')

* **Formatting Data**

We begin by loading the file and extract the text from the Summary column.

Then we do some cleaning on the text, and create a Document-Term Matrix, from which we remove the most generic terms like ‘aircraft’ or ‘crash’, that would not add any value to the analysis of the causes.

As the full Document-Term Matrix is very sparse, we reduce it by removing the sparser terms, reducing its size from 9852 few frequent terms.

*### Read file*

airline\_crash<-read.csv(file.choose())

### Load text as corpus

airline\_crash\_summary <- VCorpus(VectorSource(airline\_crash$Summary))

### Cleaning

airline\_crash\_summary <- tm\_map(airline\_crash\_summary, removePunctuation)

airline\_crash\_summary <- tm\_map(airline\_crash\_summary, content\_transformer(tolower))

airline\_crash\_summary <- tm\_map(airline\_crash\_summary, removeWords, stopwords("english"))

airline\_crash\_summary <- tm\_map(airline\_crash\_summary, stripWhitespace)

### Create document-terms matrix, removing generic terms

doc\_term\_matrix <- DocumentTermMatrix(airline\_crash\_summary, control = list(stopwords = c("aircraft", "plane", "crashed", "crash", "flight", "flew", "killed", "due", "resulted", "cause", "caused", "one", "two")))

doc\_term\_matrix

### Remove sparse terms from document-terms matrix

doc\_term\_matrix\_nosparse <- removeSparseTerms(doc\_term\_matrix, 0.97)

doc\_term\_matrix\_nosparse

* **Clustering**

To find frequently associated words, we first compute a distance matrix based on our reduced Document-Term Matrix, then apply K-means clustering.

*### Compute distance matrix*

d <- dist(t(dtms), method = "euclidian")

*### Compute K-means*

km <- kmeans(d, 6, iter.max = 50, nstart = 10)

We can first print the groups generated by the clustering, which gives the following:

*### Display results as a list*

grouplist = **function**(input) {

output <- list()

**for** (i **in** 1:max(input)) {

output[[i]] <- names(input[input == i])

}

**for** (i **in** output) {

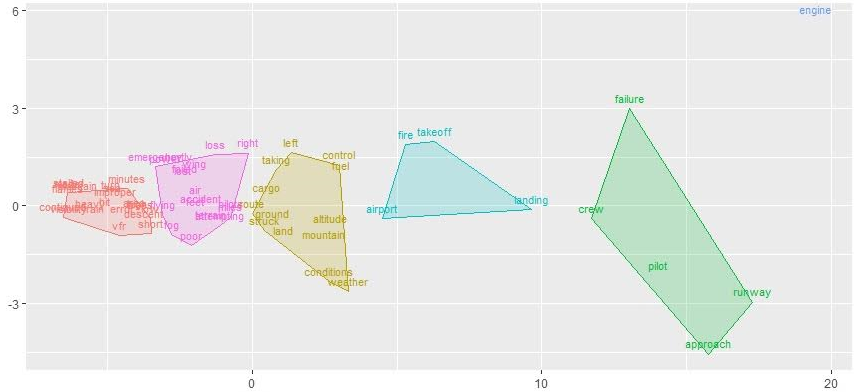
cat('\* ')

cat(i, sep = ", ")

cat("\n")

}

}



**Output of group list**

> grouplist(k\_means$cluster)

\* engine

\* airport, altitude, cargo, conditions, control, fuel, ground, land, left, mountain, struck, taking, weather

\* approach, crew, failure, pilot, runway

\* accident, air, attempting, emergency, failed, feet, flying, fog, loss, lost, miles, pilots, poor, power, right, route, shortly, terrain, wing

\* Fire, landing, takeoff

\* area, continued, descent, error, flames, heavy, high, hit, improper, low, made, maintain, minutes, rain, sea, short, stalled, trees, turn, vfr, visibility

We can go further than just printing the list of clusters, by plotting the clusters using the factoextra package. The axes will show us how far clusters are from each others.

fviz\_cluster(km, data = d, geom = "text", show.clust.cent = **FALSE**, repel = **TRUE**, labelsize = 3) +

theme(legend.position = "none") +

labs(title = "", x = "", y = "")

With this first part of the analysis, we can spot a few groups of words that are likely causes or situations of the crashes:

* pilot
* engine
* weather conditions, and notably poor visibility
* incorrect altitude
* approach of the runway
* **Frequent Terms**

To complete this first semantic analysis, we can look at the most frequent terms, and their correlation with other terms.

We begin by plotting the 20 most frequent terms. All of them are obviously included in the above cluster analysis, but here we get a sense of their frequency relatively to each others.

*### Order terms by frequency*

freq <- colSums(as.matrix(dtm))

freq <-

freq %>%

data.frame(term = names(freq), frequency = freq) %>%

select(term, frequency) %>%

arrange(desc(frequency))

*### Plot most frequent terms*

ggplot(freq[1:20, ], aes(x = frequency, y = reorder(term, frequency))) +

geom\_point(colour = "#2b83ba") +

geom\_segment(aes(xend = 0, yend = term), size = 1, colour = "#2b83ba") +

geom\_text(aes(label = term, vjust = "middle", hjust = "left"), nudge\_x = 10, size = 3.5) +

theme(panel.background = element\_rect(fill = "white"),

panel.grid.major.x = element\_line(colour = "#f7f7f7"),

panel.grid.major.y = element\_blank(),

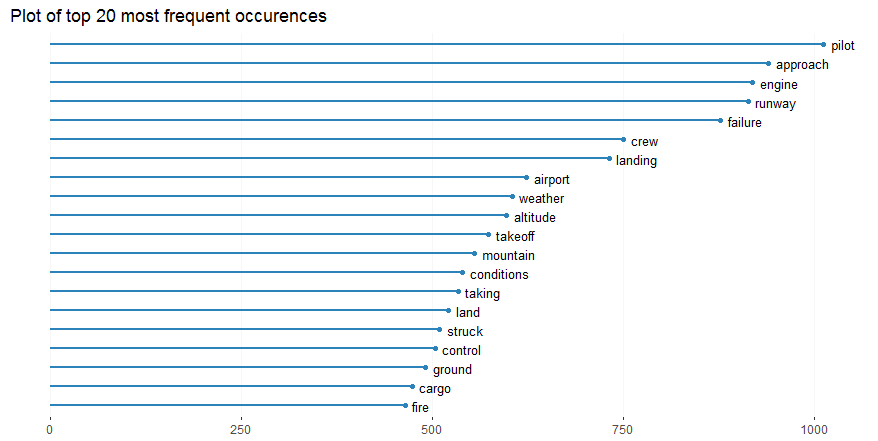
panel.grid.minor = element\_blank(),

axis.text.y = element\_blank(),

axis.title = element\_blank(),

axis.ticks.y = element\_blank()) +

labs(title = "Occurences of top 20 most frequent terms")



Some hypothesis we can do on the top 5 terms:

* **Pilot**: is it only because this is a generic term, or indicating that pilot is the cause ?
* **Approach**: this suggest that accidents often happen in the runway approach phase
* **Engine**: probably one of the most common causes
* **Runway**: relates to the approach phase
* **Failure**: this is too generic to draw conclusions, we'll some more context

To add more context to the list, we have to look at which terms are most correlated with these 20 frequent terms. For each of them, let's plot the top 5 terms that have a correlation higher than 0.17.

* **Terms Correlation**

assocs <- findAssocs(dtm, as.character(freq[1:20, 1]), corlimit = 0.16)

print(assocs)

$pilot

error  turn

 0.18  0.18

$approach

       final          ils   instrument      descent        short       missed       visual    procedure      minimum      monitor

        0.34         0.30         0.28         0.23         0.23         0.22         0.21         0.19         0.18         0.18

  stabilized nonprecision   procedures

        0.18         0.17         0.17

$engine

       power        right         shut         left    emergency         loss       failed    feathered    propeller      engines

        0.26         0.26         0.26         0.25         0.23         0.22         0.20         0.20         0.19         0.18

     trouble      cowling       return    sunshines     throwing undocumented  maintenance necessitated

        0.18         0.17         0.17         0.17         0.17         0.17         0.16         0.16

$runway

       short          end      overran      taxiway      require    threshold          05r    180degree          aug  contentious

        0.41         0.33         0.30         0.30         0.25         0.25         0.23         0.23         0.23         0.23

         dim     entrance          nov         sane          sep      tarmaat        xiang       lights       length          137

        0.23         0.23         0.23         0.23         0.23         0.23         0.23         0.22         0.21         0.20

        5191       affect        befoe       cleard       closed    crossings      guayana        horse     negative nonpertinent

        0.20         0.20         0.20         0.20         0.20         0.20         0.20         0.20         0.20         0.20

  specifiair        fence         past         take       beyond        notam    perimeter         stop      touched construction

        0.20         0.19         0.19         0.19         0.18         0.18         0.18         0.18         0.18         0.17

        feet          ran         slid

        0.17         0.17         0.17

$failure

contributing     maintain       pilots     accident   procedures     airplane       system     adequate  maintenance      provide

        0.29         0.28         0.24         0.22         0.21         0.20         0.20         0.19         0.19         0.19

  structural   adequately     properly          use    airplanes      monitor       follow

        0.19         0.18         0.18         0.18         0.17         0.17         0.16

$crew

            atc         realize             853     reactionary             eal       situation             420             707

           0.27            0.24            0.23            0.23            0.21            0.21            0.20            0.20

  compensations        dragging hydrauliproblem         mirabel       montreals            ovht           peril    peterborough

           0.20            0.20            0.20            0.20            0.20            0.20            0.20            0.20

         repeat           seals        speaking        squarely      wingengine         yorkthe   constellation    coordination

           0.20            0.20            0.20            0.20            0.20            0.20            0.19            0.19

      different         upslope         declare         evasive        extended     overheating         advised          member

           0.19            0.19            0.18            0.18            0.18            0.18            0.17            0.17

    misjudgment         optical         warning   communicating      components          dorval            shut            term

           0.17            0.17            0.17            0.16            0.16            0.16            0.16            0.16

          three        weakened

           0.16            0.16

$landing

     gear emergency   attempt      make    forced attempted      made

     0.46      0.31      0.26      0.25      0.19      0.17      0.16

$airport

international         miles        return     municipal

         0.29          0.22          0.22          0.17

$weather

      adverse          poor           vfr     continued deteriorating           bad

         0.53          0.42          0.33          0.25          0.24          0.20

$altitude

  minimum      gain clearance      feet  maintain       low   descent      safe descended

     0.33      0.25      0.23      0.22      0.21      0.20      0.19      0.19      0.16

$takeoff

       aborted          abort        overran administration           code          comet       computed        federal         modify

          0.23           0.18           0.18           0.17           0.17           0.17           0.17           0.17           0.17

    nonrevenue     nonroutine         partly    permissable       rearward         strict        shortly

          0.17           0.17           0.17           0.17           0.17           0.17           0.16

$mountain

numeric(0)

$conditions

       adverse            vfr meteorological      continued          icing     instrument           poor        terrain         pilots

          0.48           0.42           0.28           0.25           0.24           0.21           0.20           0.20           0.19

      decision  deteriorating        factors          night       maintain       whiteout

          0.18           0.18           0.18           0.18           0.17           0.17

$taking

shortly minutes

   0.44    0.24

$land

attempting   prepared

      0.53       0.16

$struck

trees

 0.17

$control

       loss        lost    elevator      system directional    inflight  stabilizer

       0.47        0.24        0.20        0.19        0.17        0.16        0.16

$ground

high

0.18

$cargo

shifted latched    hold

   0.20    0.18    0.17

$fire

         caught             420        dragging hydrauliproblem         mirabel       montreals     overheating            ovht

           0.42            0.29            0.29            0.29            0.29            0.29            0.29            0.29

   peterborough           seals      wingengine      components         advised   hydraulifluid          dorval          origin

           0.29            0.29            0.29            0.26            0.25            0.24            0.23            0.22

           burn           cabin        weakened          leaked          source  uncontrollable         upwards             hot

           0.21            0.21            0.21            0.20            0.20            0.20            0.20            0.19

          smoke       attending           brake       emergency        festival            folk         helping      inhalation

           0.19            0.18            0.18            0.18            0.18            0.18            0.18            0.18

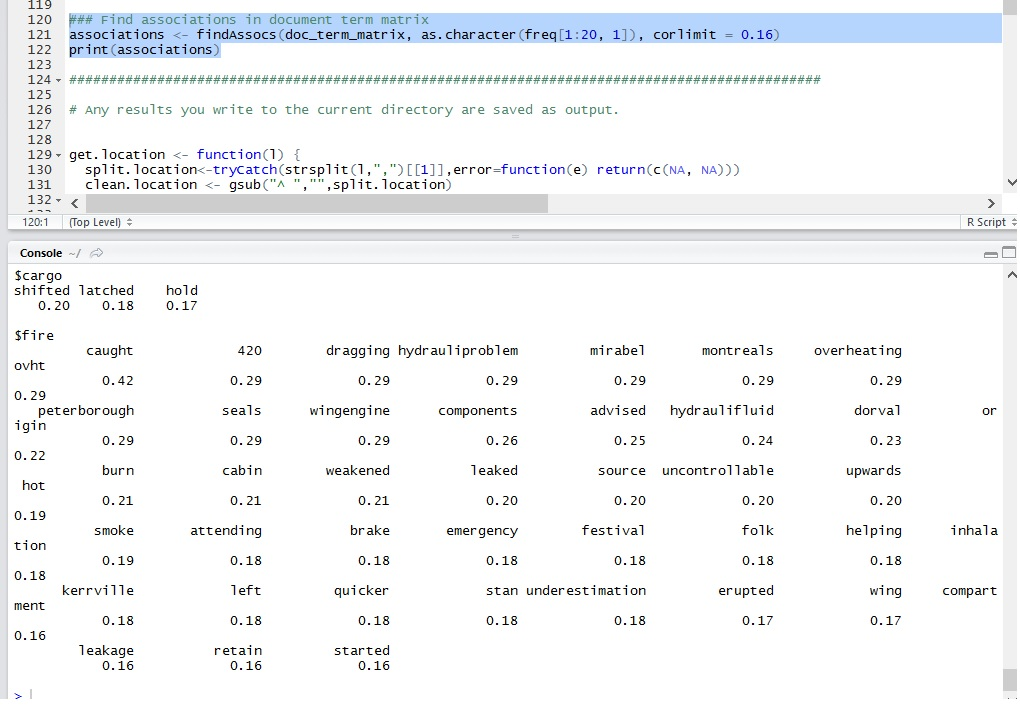
      kerrville            left         quicker            stan underestimation         erupted            wing     compartment

           0.18            0.18            0.18            0.18            0.18            0.17            0.17            0.16

        leakage          retain         started

           0.16            0.16            0.16

https://ssl.gstatic.com/ui/v1/icons/mail/images/cleardot.gif



This is quite enlightening. Let's look at some of the terms associations:

* **Pilot**: 'error' is one of the most correlated words, which is consistent with the fact that 60% of crashes are due to pilot errors
* **Approach**: the accidents in final approach phase seem to be often caused by confusion in reading instruments and low visibility ('ils', 'instruments', 'visual', 'missed')
* **Engine** seems related to shutdown of engine and/or loss of power
* **Runway** is associated with 'short', 'end' and 'overran', that could be as well in takeoff or landing phases
* **Failure**: we have more context here, suggesting that it can be pilot, maintainance, procedure or system failures
* **Landing**: this shows that it is not necessarily about the standard landing phase, but rather about landing gears, or emergency landings
* **Weather** and **Conditions** suggest that *visibility* is one of the most important crashes factors in bad weather

CONCLUSION

* Word cloud gave the overview of the causes of plane crash by showing the most frequent words used in the summary.
* Data visualization in Tableau gave the pattern showing the aircraft type ‘Douglas’ is being crashed most of the time compared to others.
* Data Visualization in R gave the pattern showing the large number of aircraft crash is occurred during the year 1972 in the month of January, August and September.
* We have risk of the event in 2 categories: high and low. So, we have applied Bayesian Classification on Risk Category Data and predicted the accuracy of the event considering all the factors as 88.5%.