Airline safety assessment based on Track Record

Pradeep Gurunathan

Introduction

The case study was done to assess the safety of airlines based on its track record, using the airline safety data hosted by FiveThirtyEight, which provides the safety records of major commercial airlines, over a span of 30 years. The context of this analysis is the outcome of certain academic studies that high-profile crashes can shift passenger demand away from the airlines involved in the disasters. The dataset divides the 30-year period into two halves and gives information about the number of crashes, fatal accidents and fatalities occurred over these years along with available seat kilometers (ASKs) of 56 different airlines. The prime objective of the case study was to check whether there was a relation between the crash rates of first time period and that of the second which would imply that the risk is persistent and is predictable based on its crash history.

Furthermore, other factors within the limit of the dataset were analyzed and type of ownership of the airline was identified as a potential predictor. ‘Are nationalized airlines safer than privatized ones?’ emerged as a sub question for the analysis.

The report documents the work undertaken to find out if the safety of airlines could be assessed based on its track record and type of ownership. The information pertaining to the crash history of airlines, the steps taken to convert this information into a risk score which indicates the risk of flying in that particular airline and the inferences upon completion of exploratory data analysis are discussed in length.

Analysis

Data Standardization:

Based on the dataset obtained we have categorized the airline into private and government airlines.

```
air <- read_csv("AirlineSafety-SAS.csv")
# Adding Type of airline.
air <- air %>%
  mutate(type=ifelse(grepl("*", airline, fixed = TRUE),"Government","Private"))
```

Data Visualization:

In-order to find the correlation of crash rate from one period to the next a scatterplot matrix was produced. The three variables: incident, fatal_accidents, fatalities are compared between two periods and how the distribution changes for the Government-owned and private-owned airlines.

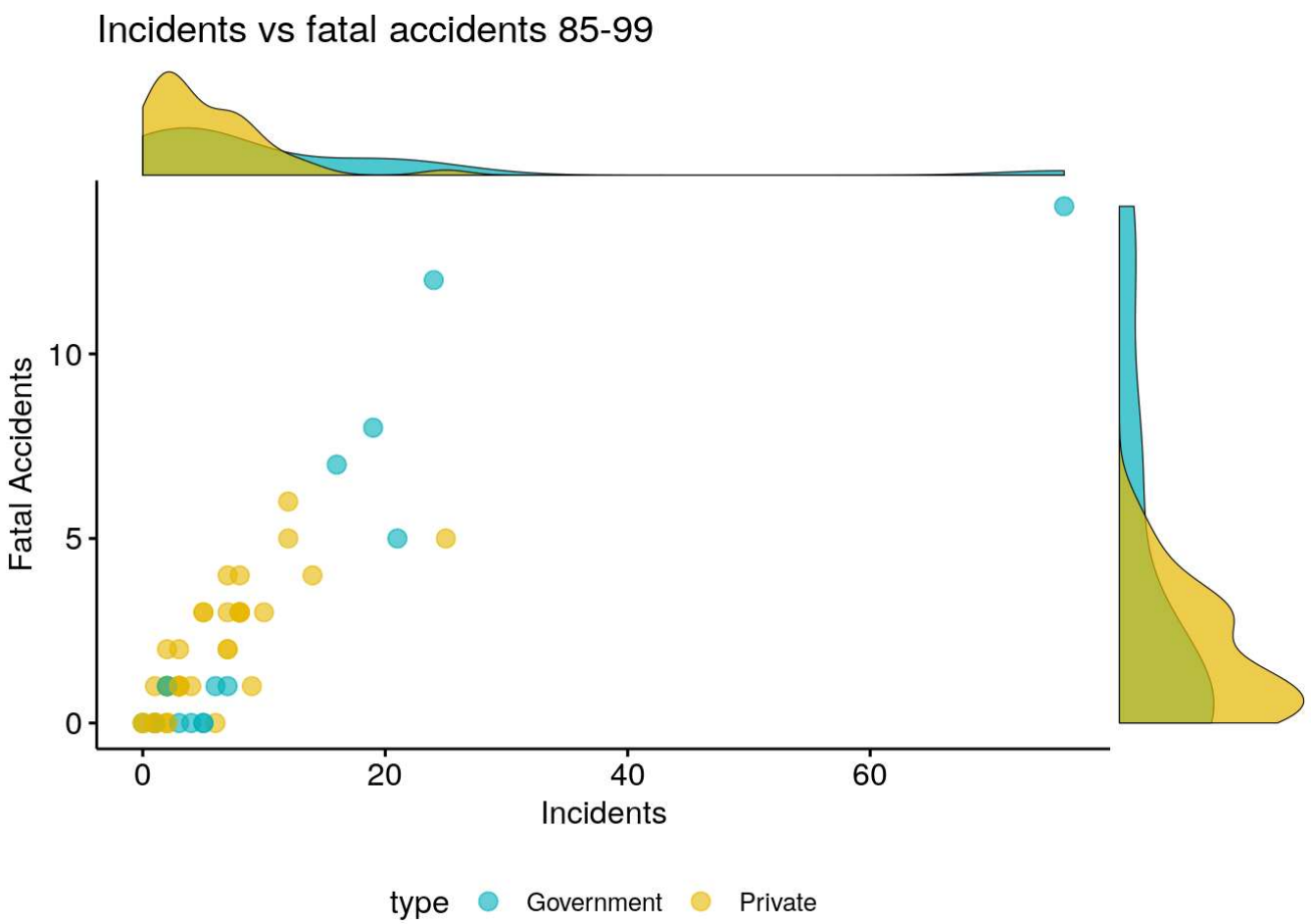
```
ggpairs(air,columns=2:8,progress = FALSE ,mapping=ggplot2::aes(colour = type),upper = list(
    continuous = wrap("cor", size = 4, alignPercent = 1)
  ) ) +  ggtitle("Airline safety Data ")
```

Overall, there is no relation between fatalities from one period to another but incidents ae slightly correlated.

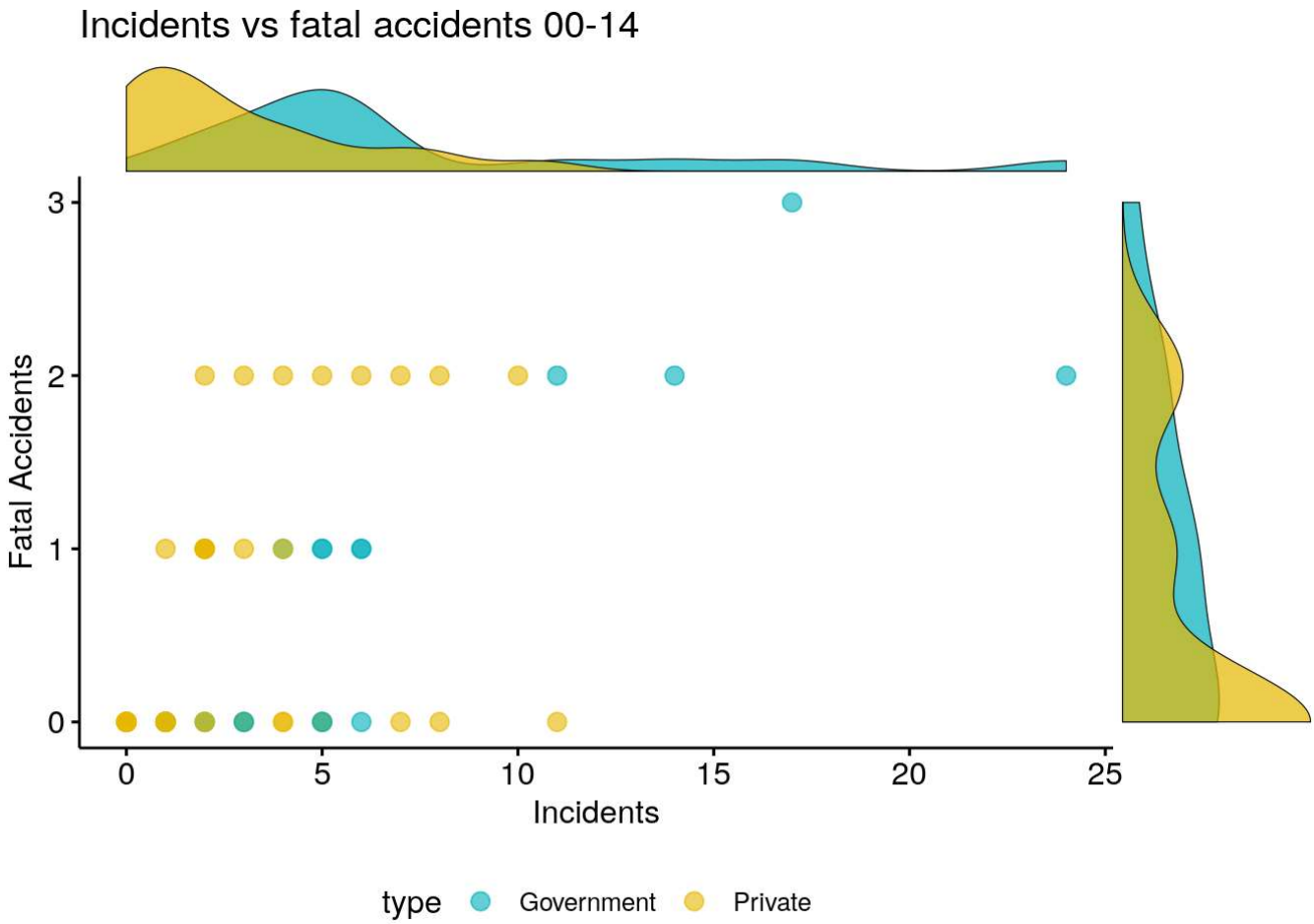
To gain detailed inferences about the data and the relationship among the variables, the following visualizations were produced:

The scatterplot below shows the incidents which resulted into fatal accidents for both the periods and the distribution of that between government and private owned airlines.

```
# Grouped Scatter plot with marginal density plots
ggscatterhist(
  air, x = "incidents_85_99", y = "fatal_accidents_85_99",
  color = "type", size = 3, alpha = 0.6,
  palette = c("#00AFBB", "#E7B800"),
  title = "Incidents vs fatal accidents 85-99", xlab = "Incidents", ylab = "Fatal Accidents",
legend = "bottom",
  margin.params = list(fill = "type", color = "black", size = 0.2))
```

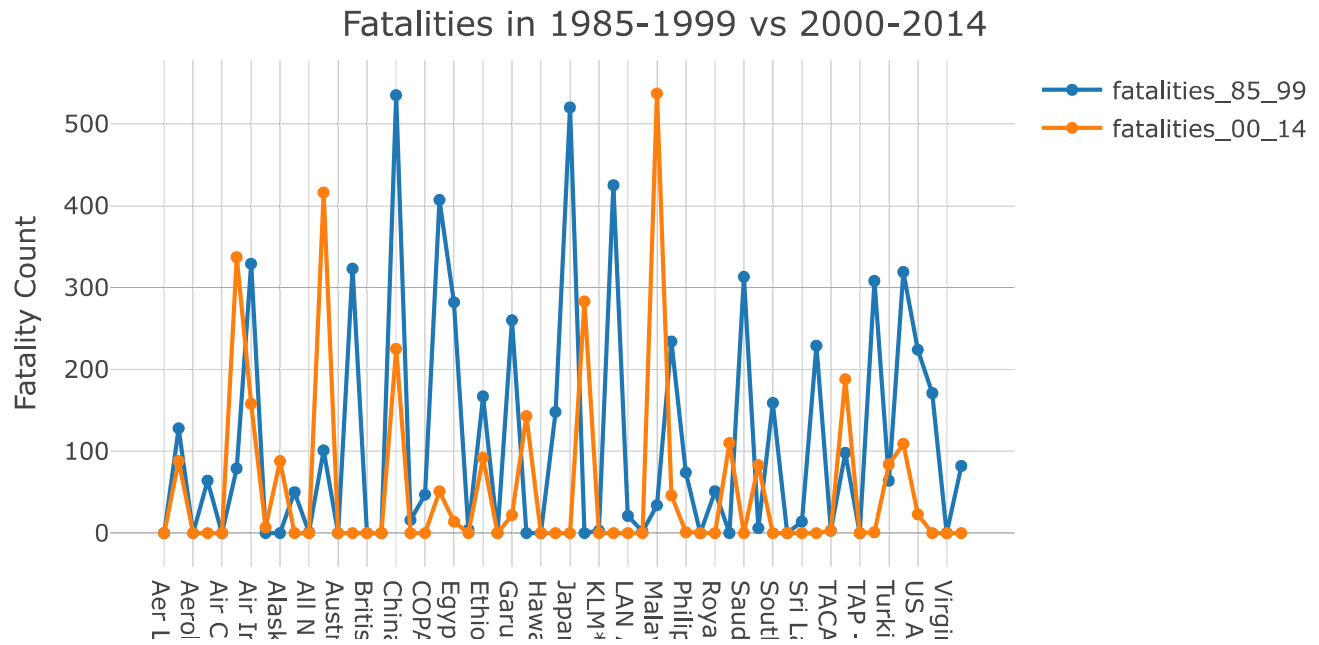


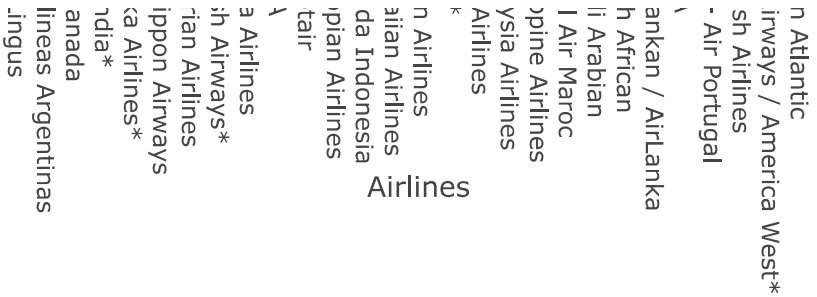

```
ggscatterhist(
  air, x = "incidents_00_14", y = "fatal_accidents_00_14",
  color = "type", size = 3, alpha = 0.6,
  palette = c("#00AFBB", "#E7B800"),
  title = "Incidents vs fatal accidents 00-14", xlab = "Incidents", ylab = "Fatal Accidents",
legend = "bottom",
  margin.params = list(fill = "type", color = "black", size = 0.2))
```



The above plots show that nationalized airlines tend to have more number crashes both in terms of incidents and fatal accidents and during both periods.

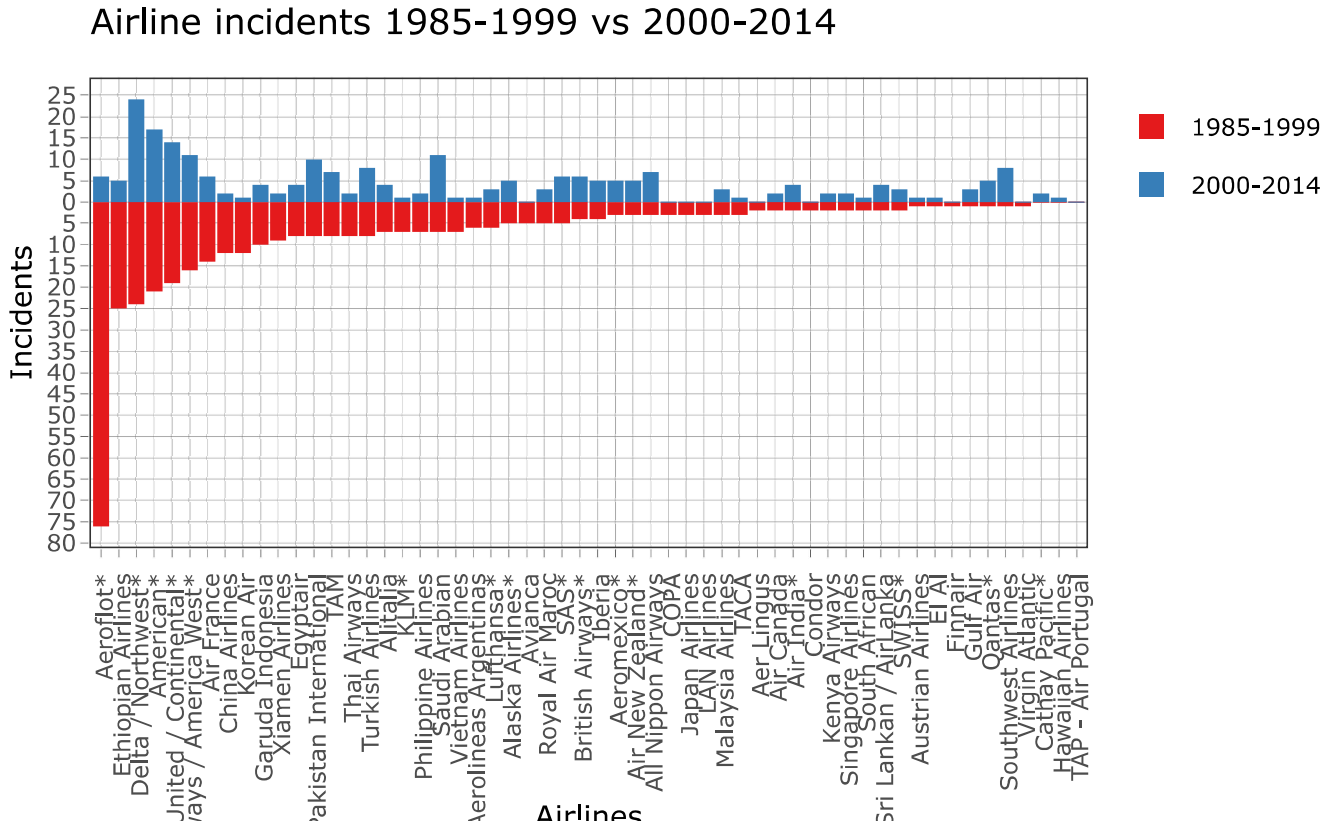
```
fig <- plot_ly(air, x = ~airline, y = ~fatalities_85_99, name = 'fatalities_85_99', type = 's
catter', mode = 'lines+markers')
fig <- fig %>% add_trace(y = ~fatalities_00_14, name = 'fatalities_00_14', mode = 'lines+mark
ers')
fig <- fig %>% layout(title = "Fatalities in 1985-1999 vs 2000-2014",
         xaxis = list(title = "Airlines"),
         yaxis = list (title = "Fatality Count"))

fig
```





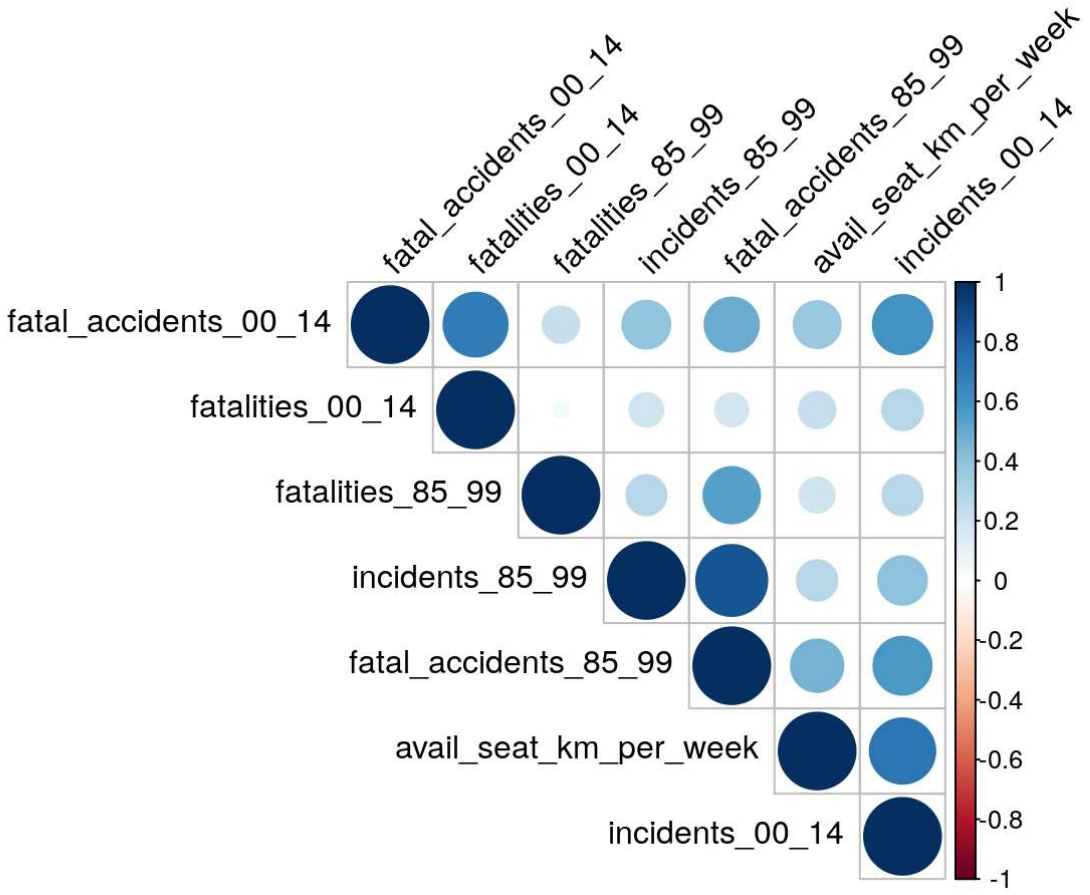
The above plot shows the distribution of fatalities over the two period. There are some airlines which contradicts our theory Eg: Malaysian Airlines which had a pretty good track record in the first period doomed to the most dangerous airline by the end of the second period whereas Japan Airlines which was apparently the most risky airlines in the first half improved significantly in the second period with 0 fatalities.

```
air_year <- air[,c("airline", "type", "incidents_85_99", "incidents_00_14")] %>%
  pivot_longer(-c("airline","type"), names_to = "year", values_to = "incidents") %>%
  mutate(year=ifelse(year == "incidents_85_99","1985-1999","2000-2014"))

p1<-ggplot(air_year, aes(x = reorder(airline,-incidents), fill = year)) +
  geom_bar(data = subset(air_year,year == "1985-1999"), aes(y=incidents*(-1)),    stat="ident
ity") +
  geom_bar(data = subset(air_year,year=="2000-2014"),aes(y=incidents),stat="identity") +
  scale_y_continuous(breaks=seq(-80,80,5),labels=abs(seq(-80,80,5))) +
  scale_fill_brewer(palette = "Set1") +
  theme_bw() +
  labs(y = "Incidents",
       x = "Airlines",
       title = "Airline incidents 1985-1999 vs 2000-2014") +
  labs(fill="") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

ggplotly(p1)
```



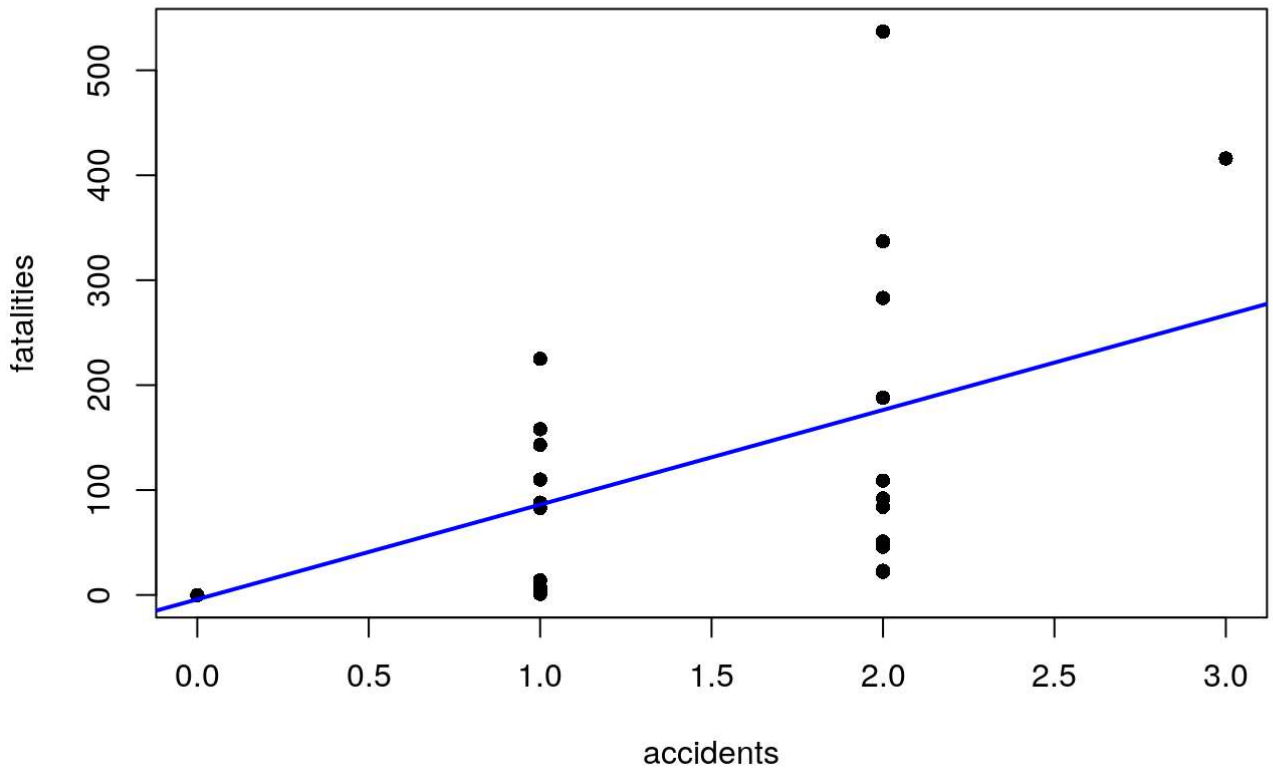
The distribution of incidents over the two time periods concur with our theory to an extent. Most of the airlines which had a good track record in terms of the number of incidents show the same trend in the second time period as well with few exceptions like Aeroflot.

```
res <- cor(air[,c(-1,-9)])
##round(res, 2)
corrplot(res, type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45)
```



We are interested to know whether there is a relationship between the various variables, a correlation coefficient can be calculated to answer this question. Creating a graphical display of a correlation matrix, highlighting the most correlated variables in a data table.In this plot, correlation coefficients are colored according to the value. Positive correlations are displayed in blue and negative correlations in red color. Color intensity and the size of the circle are proportional to the correlation coefficients. From the plot, we can observe fatalities_00_14 and fatal_accidents_00_14 are highly correlated with each other.

```
res2 <-cor.test(air$fatal_accidents_00_14, air$fatalities_00_14,  method = "kendall")
##res2
reduced<-lm(fatalities_00_14 ~fatal_accidents_00_14 ,
            data = air)

##summary(reduced)
plot(fatalities_00_14~fatal_accidents_00_14 ,data=air,pch=16,xlab = "accidents",ylab = "fatal
ities")
abline(reduced,
       col = "blue",
       lwd = 2)
```



Fitting a linear model with fatalities_00_14 as response. There are multiple fatalitites_00_14 observations at most of the fatal_accidents_00_14. Here is a plot of the data with the regression line shown, and the R Commands used to generate the test for lack of fit:

Model 1 is the usual linear regression model, which is the reduced model in this case; SSE(R) = 351615 Model 2 is telling R to consider accidents as a “Factor” instead of a continuous variable, thus treating it as categorical and fitting the mean at each fatalities.

SSE(F) = 325133 = SSE(PE) [F = Full Model, PE = Pure Error]

The Lack of Fit SSE is SSE(LF) = SSE(R) – SSE(F) = 351615 – 325133 = 26482

F-obs = (26482/2)/(325133/52) = 2.117 , Do not reject H0 because the p-value > 0.05 and conclude that using the linear regression is almost as good as using separate means at the 10 different accident values.

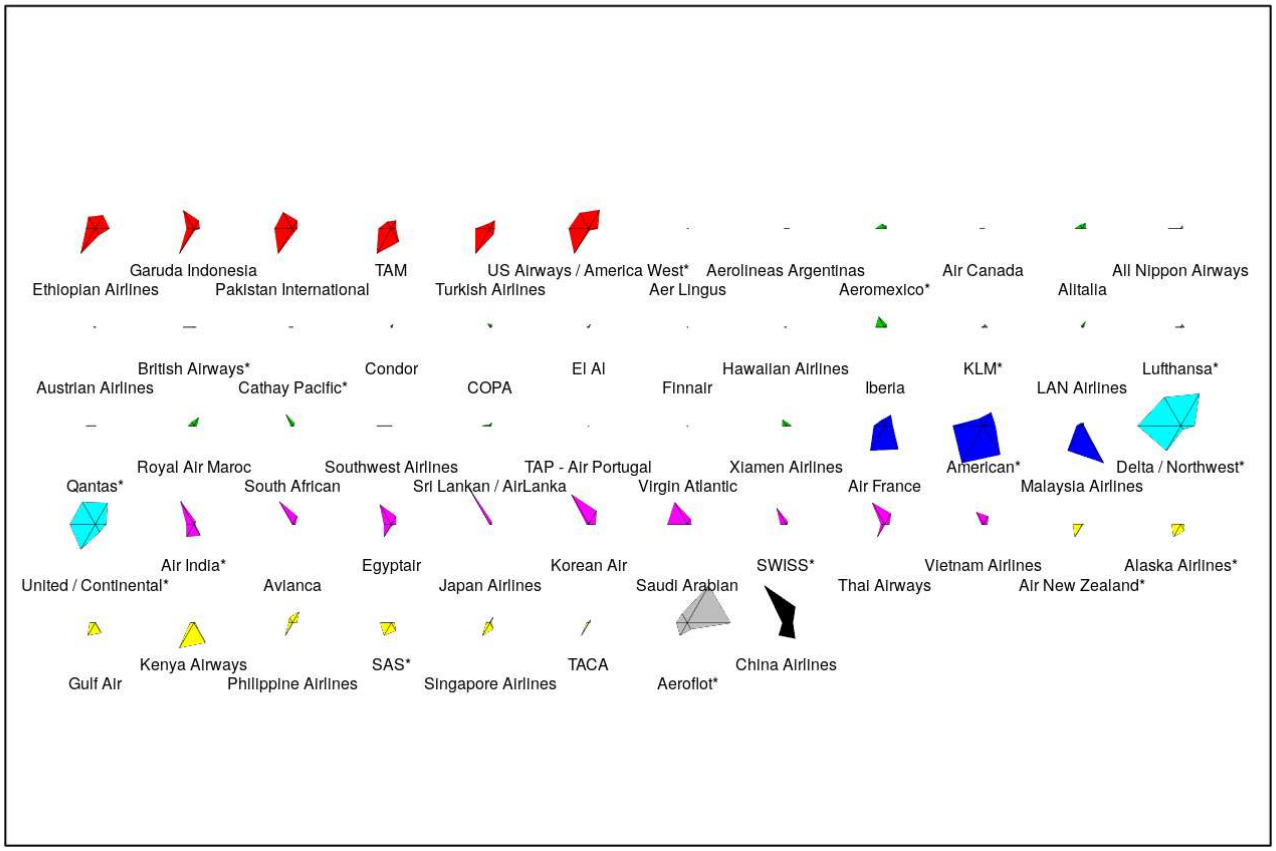
Advantage of linear regression is that we can predict even for accidents in between those measured.

Data Analysis and modelling:

K-Means Clustering

To find groups within the data based on the available features we used K-means clustering algorithm. The data was divided into 8 clusters based on feature similarity.

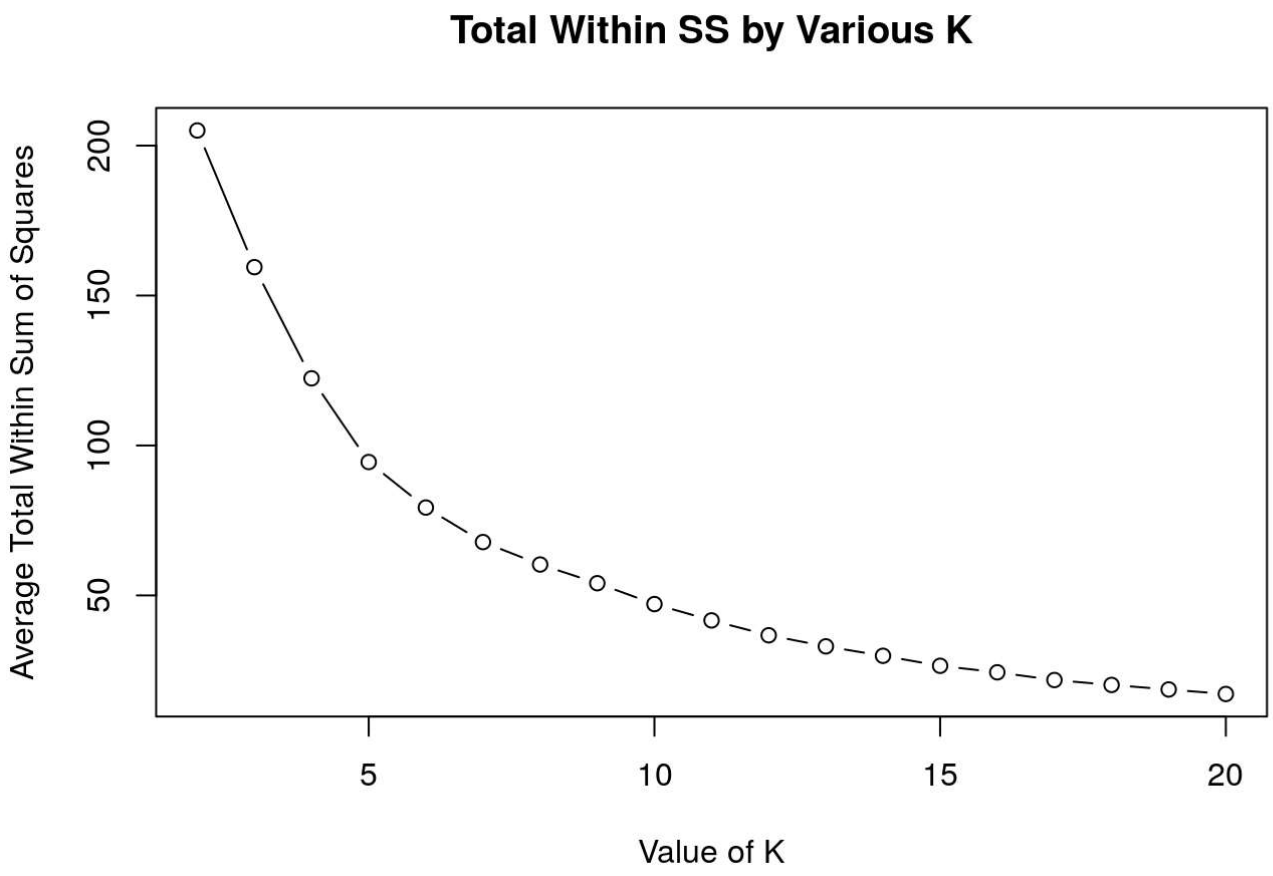
```
airline<-read.csv("AirlineSafety-SAS.csv")
airline2<-scale(airline[,c(-1,-2,-9)])
rownames(airline2)<-airline$airline

set.seed(2)
km<-kmeans(airline2,8,nstart = 10)
clusk<-km$cluster
o <- order(clusk)

stars(airline2[o,],nrow=9,ncol=12, col.stars=clusk[o]+1,frame.plot = TRUE,cex = .5)
```




```
km$tot.withinss
```

```
## [1] 60.21786
```

```
rng<-2:20 #K from 2 to 20
tries <-100 #Run the K Means algorithm 100 times
avg.totw.ss <-integer(length(rng)) #Set up an empty vector to hold all of points
for(v in rng){ # For each value of the range variable
  v.totw.ss <-integer(tries) #Set up an empty vector to hold the 100 tries
  for(i in 1:tries){
    k.temp <-kmeans(airline2,centers=v,nstart = 10) #Run kmeans
    v.totw.ss[i] <-k.temp$tot.withinss#Store the total withinss
  }
  avg.totw.ss[v-1] <-mean(v.totw.ss) #Average the 100 total withinss
}
plot(rng,avg.totw.ss,type="b", main="Total Within SS by Various K",
     ylab="Average Total Within Sum of Squares",
     xlab="Value of K")
```




```
airline$clusters<-as.factor(km$cluster)
airlinenew2<-airline%>%
  group_by(clusters)
airline[airline$clusters==1,]
```

```
##                       airline avail_seat_km_per_week incidents_85_99
## 23         Ethiopian Airlines              488560643              25
## 25           Garuda Indonesia              613356665              10
## 36     Pakistan International              348563137               8
## 48                       TAM             1509195646               8
## 51           Turkish Airlines             1946098294               8
## 53 US Airways / America West*             2455687887              16
##    fatal_accidents_85_99 fatalities_85_99 incidents_00_14 fatal_accidents_00_14
## 23                     5              167               5                     2
## 25                     3              260               4                     2
## 36                     3              234              10                     2
## 48                     3               98               7                     2
## 51                     3               64               8                     2
## 53                     7              224              11                     2
##    fatalities_00_14 clusters
## 23               92        1
## 25               22        1
## 36               46        1
## 48              188        1
## 51               84        1
## 53               23        1
```

```
airline[airline$clusters==2,]
```

```
##                  airline avail_seat_km_per_week incidents_85_99
## 1             Aer Lingus              320906734               2
## 3  Aerolineas Argentinas              385803648               6
## 4            Aeromexico*              596871813               3
## 5             Air Canada             1865253802               2
## 10              Alitalia              698012498               7
## 11    All Nippon Airways             1841234177               3
## 13     Austrian Airlines              358239823               1
## 15      British Airways*             3179760952               4
## 16       Cathay Pacific*             2582459303               0
## 18                Condor              417982610               2
## 19                  COPA              550491507               3
## 22                El Al              335448023               1
## 24               Finnair              506464950               1
## 27     Hawaiian Airlines              493877795               0
## 28               Iberia             1173203126               4
## 31                 KLM*             1874561773               7
## 33          LAN Airlines             1001965891               3
## 34            Lufthansa*             3426529504               6
## 38               Qantas*             1917428984               1
## 39       Royal Air Maroc              295705339               5
## 43         South African              651502442               2
## 44    Southwest Airlines             3276525770               1
## 45 Sri Lankan / AirLanka              325582976               2
## 49    TAP - Air Portugal              619130754               0
## 55       Virgin Atlantic             1005248585               1
## 56       Xiamen Airlines              430462962               9
##    fatal_accidents_85_99 fatalities_85_99 incidents_00_14 fatal_accidents_00_14
## 1                      0                0               0                      0
## 3                      0                0               1                      0
## 4                      1               64               5                      0
## 5                      0                0               2                      0
## 10                     2               50               4                      0
## 11                     1                1               7                      0
## 13                     0                0               1                      0
## 15                     0                0               6                      0
## 16                     0                0               2                      0
## 18                     1               16               0                      0
## 19                     1               47               0                      0
## 22                     1                4               1                      0
## 24                     0                0               0                      0
## 27                     0                0               1                      0
## 28                     1              148               5                      0
## 31                     1                3               1                      0
## 33                     2               21               0                      0
## 34                     1                2               3                      0
## 38                     0                0               5                      0
## 39                     3               51               3                      0
## 43                     1              159               1                      0
## 44                     0                0               8                      0
## 45                     1               14               4                      0
## 49                     0                0               0                      0
## 55                     0                0               0                      0
## 56                     1               82               2                      0
##    fatalities_00_14 clusters
## 1                 0        2
## 3                 0        2
```

```
## 4                 0        2
## 5                 0        2
## 10                0        2
## 11                0        2
## 13                0        2
## 15                0        2
## 16                0        2
## 18                0        2
## 19                0        2
## 22                0        2
## 24                0        2
## 27                0        2
## 28                0        2
## 31                0        2
## 33                0        2
## 34                0        2
## 38                0        2
## 39                0        2
## 43                0        2
## 44                0        2
## 45                0        2
## 49                0        2
## 55                0        2
## 56                0        2
```

```
airline[airline$clusters==3,]
```

```
##              airline avail_seat_km_per_week incidents_85_99
## 6         Air France             3004002661              14
## 12         American*             5228357340              21
## 35 Malaysia Airlines             1039171244               3
##    fatal_accidents_85_99 fatalities_85_99 incidents_00_14 fatal_accidents_00_14
## 6                      4               79               6                     2
## 12                     5              101              17                     3
## 35                     1               34               3                     2
##    fatalities_00_14 clusters
## 6               337        3
## 12              416        3
## 35              537        3
```

```
airline[airline$clusters==4,]
```

```
##                  airline avail_seat_km_per_week incidents_85_99
## 20    Delta / Northwest*             6525658894              24
## 52 United / Continental*             7139291291              19
##    fatal_accidents_85_99 fatalities_85_99 incidents_00_14 fatal_accidents_00_14
## 20                    12              407              24                     2
## 52                     8              319              14                     2
##    fatalities_00_14 clusters
## 20               51        4
## 52              109        4
```

```
airline[airline$clusters==5,]
```

```
##             airline avail_seat_km_per_week incidents_85_99
## 7        Air India*              869253552               2
## 14          Avianca              396922563               5
## 21         Egyptair              557699891               8
## 29   Japan Airlines             1574217531               3
## 32       Korean Air             1734522605              12
## 41    Saudi Arabian              859673901               7
## 46           SWISS*              792601299               2
## 50     Thai Airways             1702802250               8
## 54 Vietnam Airlines              625084918               7
##    fatal_accidents_85_99 fatalities_85_99 incidents_00_14 fatal_accidents_00_14
## 7                      1              329               4                      1
## 14                     3              323               0                      0
## 21                     3              282               4                      1
## 29                     1              520               0                      0
## 32                     5              425               1                      0
## 41                     2              313              11                      0
## 46                     1              229               3                      0
## 50                     4              308               2                      1
## 54                     3              171               1                      0
##    fatalities_00_14 clusters
## 7               158        5
## 14                0        5
## 21               14        5
## 29                0        5
## 32                0        5
## 41                0        5
## 46                0        5
## 50                1        5
## 54                0        5
```

```
airline[airline$clusters==6,]
```

```
##               airline avail_seat_km_per_week incidents_85_99
## 8     Air New Zealand*              710174817               3
## 9     Alaska Airlines*              965346773               5
## 26            Gulf Air              301379762               1
## 30       Kenya Airways              277414794               2
## 37 Philippine Airlines              413007158               7
## 40                SAS*              682971852               5
## 42  Singapore Airlines             2376857805               2
## 47                TACA              259373346               3
##    fatal_accidents_85_99 fatalities_85_99 incidents_00_14 fatal_accidents_00_14
## 8                      0                0               5                     1
## 9                      0                0               5                     1
## 26                     0                0               3                     1
## 30                     0                0               2                     2
## 37                     4               74               2                     1
## 40                     0                0               6                     1
## 42                     2                6               2                     1
## 47                     1                3               1                     1
##    fatalities_00_14 clusters
## 8                 7        6
## 9                88        6
## 26              143        6
## 30              283        6
## 37                1        6
## 40              110        6
## 42               83        6
## 47                3        6
```

```
airline[airline$clusters==7,]
```

```
##     airline avail_seat_km_per_week incidents_85_99 fatal_accidents_85_99
## 2 Aeroflot*             1197672318              76                    14
##   fatalities_85_99 incidents_00_14 fatal_accidents_00_14 fatalities_00_14
## 2              128               6                     1               88
##   clusters
## 2        7
```

```
airline[airline$clusters==8,]
```

```
##           airline avail_seat_km_per_week incidents_85_99 fatal_accidents_85_99
## 17 China Airlines              813216487              12                     6
##    fatalities_85_99 incidents_00_14 fatal_accidents_00_14 fatalities_00_14
## 17              535               2                     1              225
##    clusters
## 17        8
```

Cluster 1: Airlines which have Highest Fatal Accidents among all the airlines (excluding American airlines) in Phase 2, with the number of fatalities significantly reduced in Phase 2, when compared to Phase 1.

Cluster 2: Airlines with significant number of Incidents and Fatalities in Phase 1, with incidents and fatalities highly reduced (nearly zero) in Phase 2.

Cluster 3: Airlines with highest fatality rates in Phase 2. Probably had met with huge accidents and hence one of the most dangerous groups.

Cluster 4: Airlines with high fatality rates in Phase 1, which significantly reduced in Phase 2. (These airlines can be trusted, as they have probably ensured to take safety measures and improve their overall flight safety).

Cluster 5: Airlines with all the incident, fatal accidents and fatality rates reduced in phase 2 from Phase 1. Supposedly the safest group.

Cluster 6: Airlines with very low incident rates in Phase 1 but have significantly higher rates in Phase 2. (except for Philippines airlines, which seems to be an outlier for the group where the accident rates have reduced).

Cluster 7: Aeroflot: Highest number of accidents in Phase 1. Have improved in their safety in Phase 2. Seems to have improved.

Cluster 8: China airlines has the number of highest fatalities over the period of 1985 to 1999.

Risk Score Calculation:

To quantitatively measure the risk of airlines and compare between we calculated a risk score by combining the 3 variables after scaling and giving appropriate weightages. The score for both periods was then combined to generate an overall score. The score was calculated through the following steps:

1. Calculating Harmless_Incidents: Incidents-Fatal Accidents
2. Standardizing Harmless_Incidents, Fatal_Accidents and Fatalities
3. Degree of harmfulness can be determined in the order: Fatalities > Fatal_Accidents > Harmless_Incident. So providing following weights to the columns:

Harmless_Incidents-25%

Fatal_Accidents-35%

Fatalities - 40%

4. Calculating Risk Score : avg(Harmless_Incidents, Fatal_Accidents , Fatalities) Positive Risk Score indicate a bad track record meaning the airline is unsafe and negative Risk score tells that the airline is relatively safe.

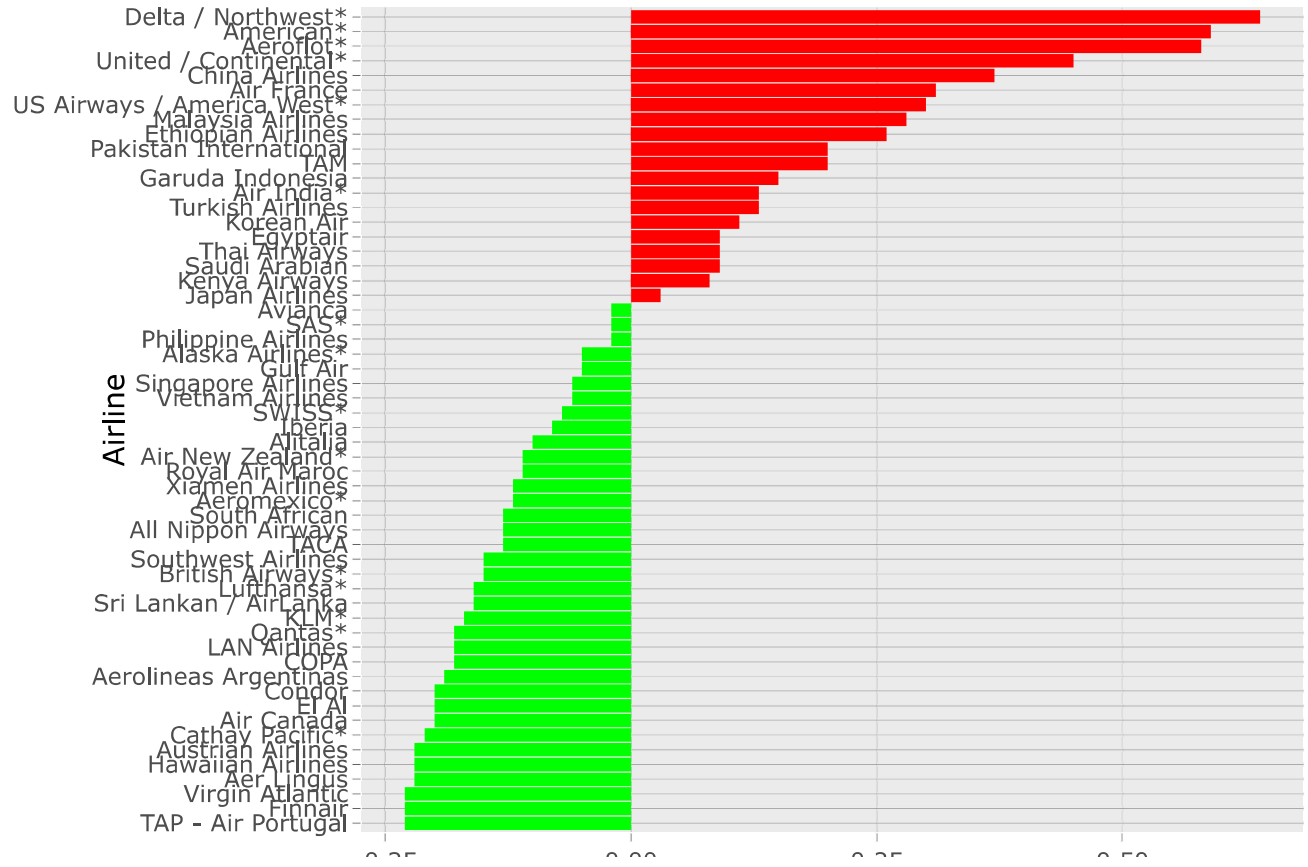
```
airline$type<-air$type
airline <- airline %>%
  mutate(Harmless_Inci_85_99 = incidents_85_99 - fatal_accidents_85_99,
         Harmless_Inci_00_14 = incidents_00_14 - fatal_accidents_00_14)
```

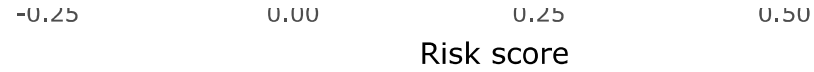
```
airline_std <- airline %>% rename(rowname=airline) %>% select(rowname, Harmless_Inci_85_99, f
atal_accidents_85_99, fatalities_85_99,Harmless_Inci_00_14, fatal_accidents_00_14,fatalities_
00_14) %>% column_to_rownames() %>% scale() %>% data.frame() %>% rownames_to_column() %>% ren
ame(airline=rowname)

airline_std$airline <- as.factor(airline_std$airline)

airline_std <- airline_std %>%
  inner_join(airline[,c("airline","avail_seat_km_per_week", "type")], by = "airline")
```

```
airline_std <- airline_std %>%
  mutate(Harmless_Inci_85_99 = 0.25 * Harmless_Inci_85_99,
         fatal_accidents_85_99 = 0.35 * fatal_accidents_85_99,
         fatalities_85_99 = 0.4 * fatalities_85_99,
         Harmless_Inci_00_14 = 0.25 * Harmless_Inci_00_14,
         fatal_accidents_00_14 = 0.35 * fatal_accidents_00_14,
         fatalities_00_14 = 0.4 * fatalities_00_14)
```

```
airline_risk_score <- airline_std %>%
  mutate(risk_score_85_99 = (Harmless_Inci_85_99 + fatal_accidents_85_99 + fatalities_85_99)
/ 3,
         risk_score_00_14 = (Harmless_Inci_00_14 + fatal_accidents_00_14 + fatalities_00_14)
 / 3) %>%
  select(airline,type,risk_score_85_99,risk_score_00_14) %>%
  mutate(risk_score = (risk_score_85_99 + risk_score_00_14)/2) %>%
  arrange(desc(risk_score))
```

```
airline_risk_score$Airline <- factor(airline_risk_score$airline, levels=airline_risk_score$ai
rline[order(airline_risk_score$risk_score)], ordered=TRUE)
airline_risk_score$Risk_Score <- round(airline_risk_score$risk_score,2)

gg_risk_score <- ggplot(data = airline_risk_score, aes(x = Airline, y=Risk_Score, fill=ifelse
(risk_score > 0,"red","green"))) +
  geom_col() +
  scale_fill_manual(values = c("green", "red")) +
  xlab("Airline") + ylab("Risk score") +
  theme(legend.position = "none") +
  coord_flip()

ggplotly(gg_risk_score, tooltip = c("x","y"))
```



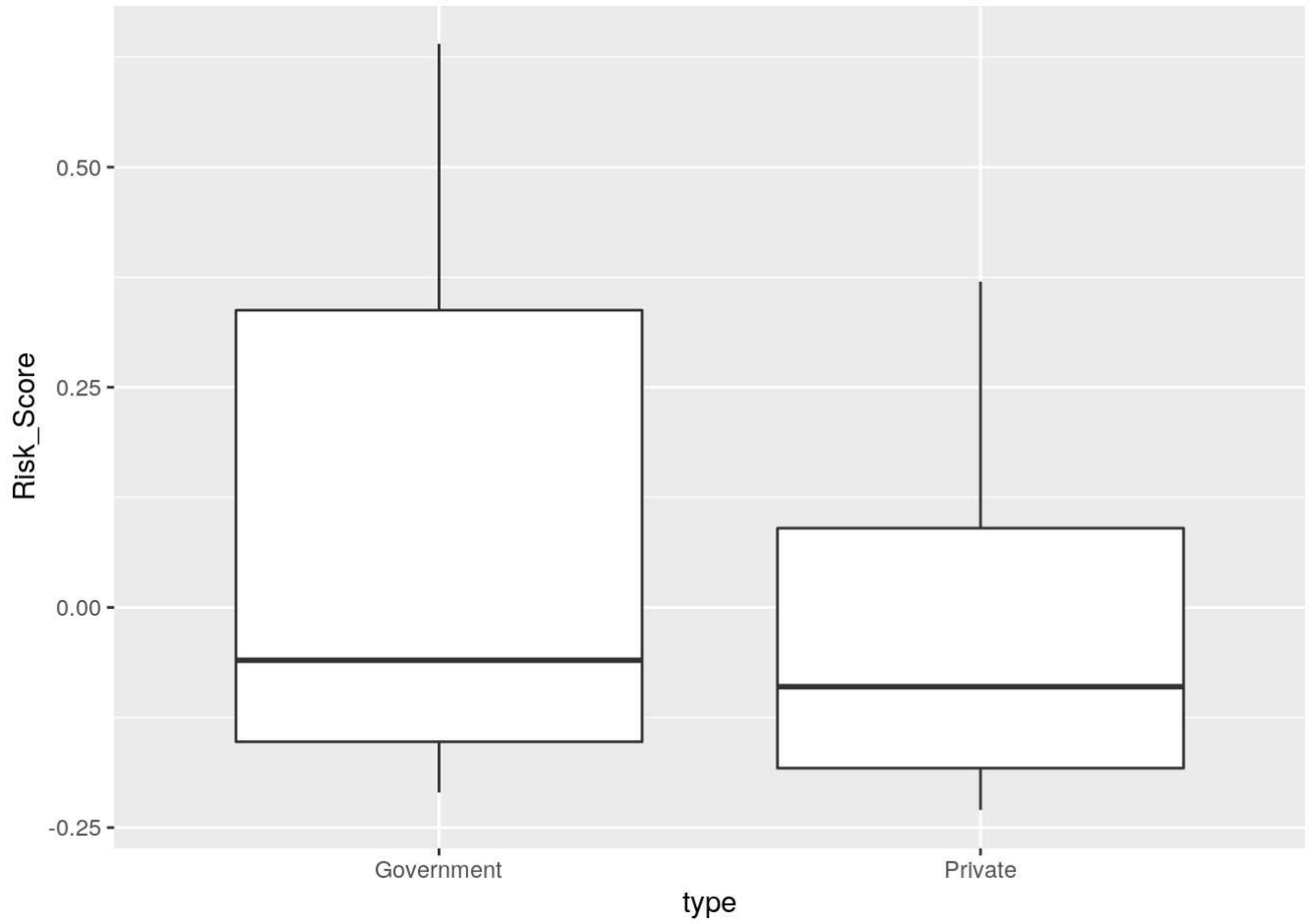


Inference:

if you conduct a regression on an airline's safety score from 2000-2014 on the basis of its safety score from 1985-1999 we can infer that only a small proportion is predictable based on the track record and can be ignored.

However, plotting the overall Score against the nationalized and Privatized airlines the mean risk score of privatized airlines is relatively small.

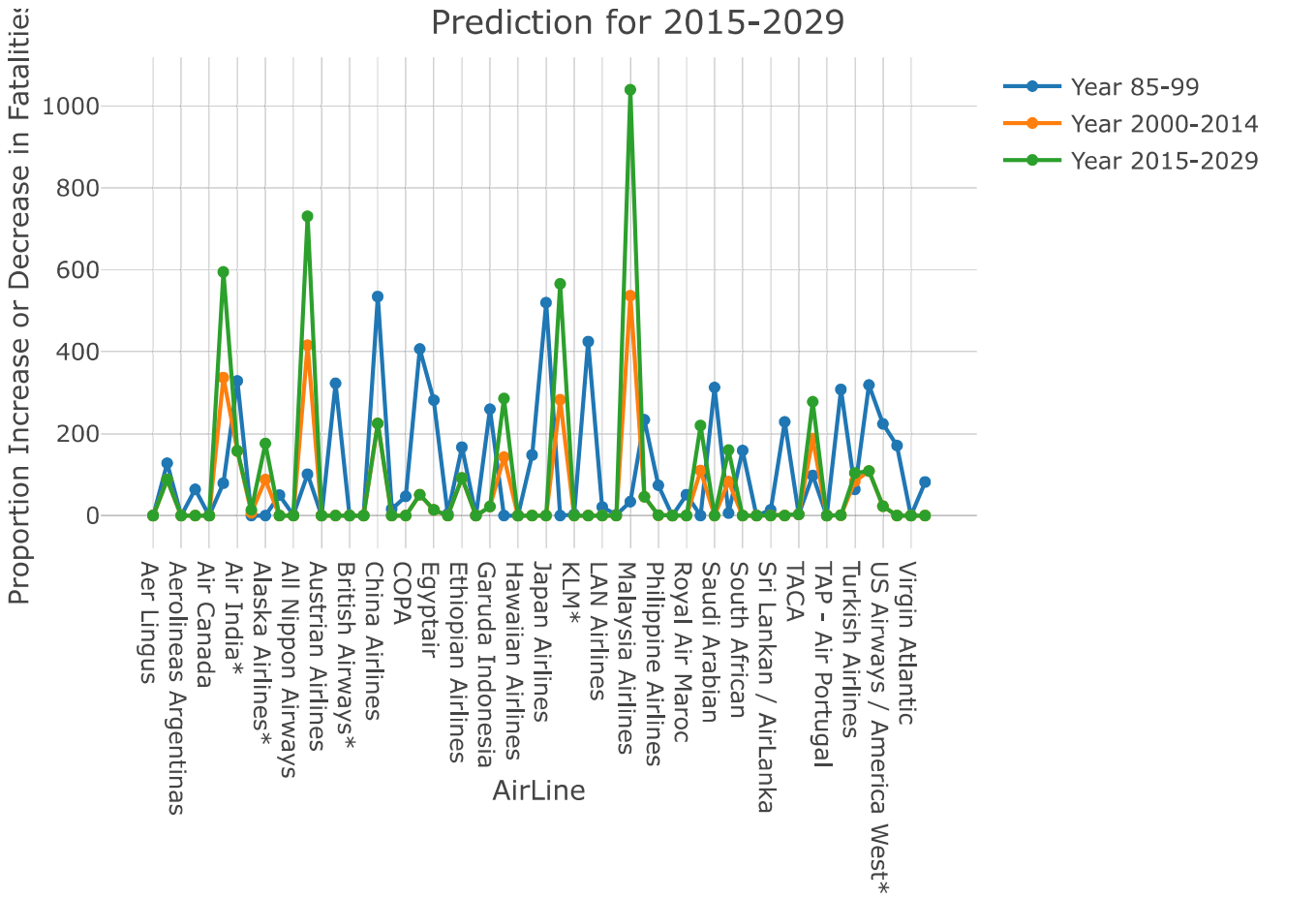
```
ggplot(airline_risk_score, aes(x=type , y=Risk_Score)) + geom_boxplot()
```



Prediction:

Below plot shows the prediction of the fatalities for 2015-2029 time period.

```
air85_14<-air[,c(1,5,8)]
air85_14<-air85_14%>%
  mutate(prop=fatalities_00_14-fatalities_85_99)
air85_14<-air85_14%>%
  mutate(g=ifelse(prop<0,0,prop))%>%
  mutate(fatalities_15_29=fatalities_00_14+g)

fig1 <- plot_ly(air85_14, x = ~airline, y = ~fatalities_85_99, name = 'Year 85-99', type = 's
catter', mode = 'lines+markers')
fig1 <- fig1 %>% add_trace(y = ~fatalities_00_14, name = 'Year 2000-2014', mode = 'lines+mark
ers')
fig1 <- fig1 %>% add_trace(y = ~fatalities_15_29, name = 'Year 2015-2029', mode = 'lines+mark
ers')
fig1 <- fig1 %>% layout(title = "Prediction for 2015-2029", xaxis = list(title = "AirLine"),
 yaxis = list (title = "Proportion Increase or Decrease in Fatalities"))
fig1
```



Malaysian airline has the high possible fatalities based on the previous data.

Conclusion

Based on the analysis done it can be said that Private airlines are much safer than Government owned airlines. Malaysian Airlines which had a pretty good track record but ended up to be a most dangerous airline by the end of the second period. Whereas, Japan Airlines which was apparently the riskiest airlines in the first half improved significantly in the second period with no fatalities.

From Risk score generated we can say that TAP-Air Portugal is the safest airline, while the Delta/Northwest airlines is most un-safe airline.

Finally, for further precise conclusion over “Which Airlines is the safest to travel?” we have to take into account on other useful details like Total Passengers travelled, Number of Journey, Total Journey Hours, Type of accident (i.e. Human Error, Mechanical, Terrorist, etc.), Airplane type (i.e. Jumbo, mid-size, small, etc.) and Service Status of the Airlines (i.e. Active, Shutdown).

Reference

[1] Nate Silver. ” Should Travelers Avoid Flying Airlines That Have Had Crashes in the Past?”. https://fivethirtyeight.com/features/should-travelers-avoid-flying-airlines-that-have-had-crashes-in-the-past/ (https://fivethirtyeight.com/features/should-travelers-avoid-flying-airlines-that-have-had-crashes-in-the-past/)

[2] Catherine Hurley. ” K-means Clustering”. Maynooth university.