Data Analysis & Statistical Learning - Project: Incidenti Aerei dal 1908 al 2009

Davide Zicca

17/04/2021

Motivazione, Data Cleaning e Visualization

Il seguente link presenta il dataset originale oggetto di analisi in questo progetto: https://www.kaggle.com/saurograndi/airplane-crashes-since-1908. Si tratta di una raccolta di informazioni riguardante gli incidenti aerei rilevati dal 1908 fino al 2009. L'obiettivo di questo progetto è quello di presentare, illustrare il dataset e applicare alcune tecniche di clustering. Carico le librerie necessarie e il dataset in R:

```
library(tidyverse)
library(lubridate)
library(rapportools)
library(ggplot2)
library(repr)
library(RColorBrewer)
library(factoextra)
library(gridExtra)
library(cluster)
library(plyr)

AirCrash <- read.csv('Airplane_Crashes_and_Fatalities_Since_1908.csv')
```

Il dataset presenta 5268 osservazioni e 13 variabili. In particolare:

- Date: Data dell'incidente
- Time: Ora locale nel formato ore:minuti
- Location: Luogo dell'incidente
- Operator: Compagnia aerea o operatore del velivolo
- Flight: Numero del volo assegnato all'operatore del velivolo
- Route: Percorso completato prima dell'incidente
- Type: Tipo di velivolo
- Registration: Registrazione del velivolo nell'Organizzazione internazionale dell'aviazione civile
- cn/In: Numero seriale / Linea o numero di fusoliera
- Aboard: Totale persone a bordo
- Fatalities: Totale vittime a bordo
- Ground: Totale vittime al suolo
- Summary: Breve descrizione dell'incidente, e cause dello stesso se conosciute

Si procede dunque ad analizzare il dataset e a prepararlo per le fasi successive del progetto:

```
# Verifica dati duplicati
print(paste('Totale righe duplicate: ',nrow(AirCrash[duplicated(AirCrash),])))
```

```
## [1] "Totale righe duplicate:  0"
```

```
# Verifica valori 'NA'
print(paste("Totale valori 'NA': ",sum(is.na(AirCrash))))
```

```
## [1] "Totale valori 'NA':  56"
```

```
# Lista valori Null
Null_Values <- (sapply(AirCrash,function(x) sum(is.na(x))))
t(data.frame(Null_Values))
```

```
##             Date Time Location Operator Flight.. Route Type Registration cn.In
## Null_Values    0    0        0        0        0     0    0            0     0
##             Aboard Fatalities Ground Summary
## Null_Values     22         12     22       0
```

```
# Verifico i velivoli in cui non figurano valori in 'Aboard'
AirCrash[is.empty(AirCrash$Aboard),c(1,4,5,7,10,11,12)]
```

```
##            Date                              Operator Flight..
## 27   10/20/1919       Aircraft Transport and Travel
## 334  08/10/1934 China National Aviation Corporation
## 349  03/07/1935                           Deruluft
## 365  08/13/1935 China National Aviation Corporation
## 424  12/26/1936 China National Aviation Corporation
## 527  09/26/1939            KLM Royal Dutch Airlines
## 538  07/07/1940                         Air France
## 571  01/24/1942                              KNILM
## 572  01/26/1942                              KNILM
## 574  02/14/1942 China National Aviation Corporation
## 588  08/13/1942                         Air France
## 594  10/01/1942 China National Aviation Corporation
## 679  11/09/1944      Military - U.S. Army Air Corps
## 769  03/18/1946 China National Aviation Corporation
## 833  12/25/1946 China National Aviation Corporation
## 1480 04/20/1957                         Air France
## 3008 11/03/1977                              El Al
## 3308 09/22/1981        Military - Turkish Air Force
## 3324 12/16/1981                  Bristow Helicopters
## 3370 08/11/1982          Pan American World Airways      830
## 3612 03/27/1986         Military - French Air Force
## 3844 05/09/1989                         Aero Asahi
## 4081 02/20/1992               Aerolineas Argentinas      386
## 4706 03/22/2000       Military - EjÃ©rcito del Aire
##                              Type Aboard Fatalities Ground
## 27                 De Havilland DH-4     NA         NA     NA
## 334                   Sikorsky S-38B     NA         NA     NA
## 349                 Rochrbach Roland     NA          3      0
## 365                   Sikorsky S-38B     NA         NA     NA
## 424                     Douglas DC-2     NA         NA     NA
## 527                     Douglas DC-3     NA          1      0
## 538                  Dewoitine D-338     NA         NA     NA
```

```
## 571                     Douglas DC-3     NA         NA     NA
## 572               Grumman G-21 Goose     NA         NA     NA
## 574                     Douglas DC-2     NA         NA     NA
## 588  Liore et Olivier H-246 Air Boat     NA          4      0
## 594                     Douglas C-47     NA         NA     NA
## 679                                      NA         NA     NA
## 769                                      NA         NA     NA
## 833         Curtiss C-46, C-47, DC-3     NA         87      4
## 1480    Lockheed Super Constellation     NA          1      0
## 3008                    Boeing B-747     NA          1      0
## 3308                   Northrop F-5A      0          0     40
## 3324                Aerospatiale Puma     NA         12      0
## 3370                 Boeing B-747-121     NA          1      0
## 3612                 Sepecat Jaguar A      0          0     35
## 3844                         Bell 412     NA         10      0
## 4081                    Boeing B-747     NA          1      0
## 4706         CASA 212-DE Aviocar 200     NA         NA     NA
```

```
# Mantengo tutti i dati relarivi ai valori 'Aboard' che non sono 'NA'
AirCrash <- AirCrash[!is.empty(AirCrash$Aboard),]
# Converto i valori 'Ground NA' in 0
AirCrash$Ground[is.na(AirCrash$Ground)] <- 0

# Verifico spazi vuoti ""
Missing_Values <- (sapply(AirCrash,function(x) sum(x=="")))
(data.frame(Missing_Values))
```

```
##              Missing_Values
## Date                      0
## Time                   2196
## Location                 19
## Operator                 18
## Flight..               4177
## Route                  1687
## Type                     25
## Registration            330
## cn.In                  1215
## Aboard                    0
## Fatalities                0
## Ground                    0
## Summary                 383
```

```
# Converto i campi 'Date' in valori 'date'
AirCrash$Date <- as.Date(AirCrash$Date, format = "%m/%d/%Y")
# Converto i campti 'Time' in valori 'time'
AirCrash$LocalTime <- as.POSIXct(AirCrash$Time, format = "%H:%M")

# Aggiungo la colonna 'LocalHour' che è rappresentata in formato numerico
AirCrash$LocalHour <- as.numeric(format(AirCrash$LocalTime,"%H"))

# Sostituisco temporaneamente 'Local Hour NA's' con 25 per poter utilizzare la
# funzione cut
AirCrash$LocalHour <- ifelse(is.na(AirCrash$LocalHour), 25, AirCrash$LocalHour)
# Add discretized dayparts based on Local Hour
AirCrash$Daypart <- cut(AirCrash$LocalHour, breaks = c(-1,5,11,17,24,25),
```

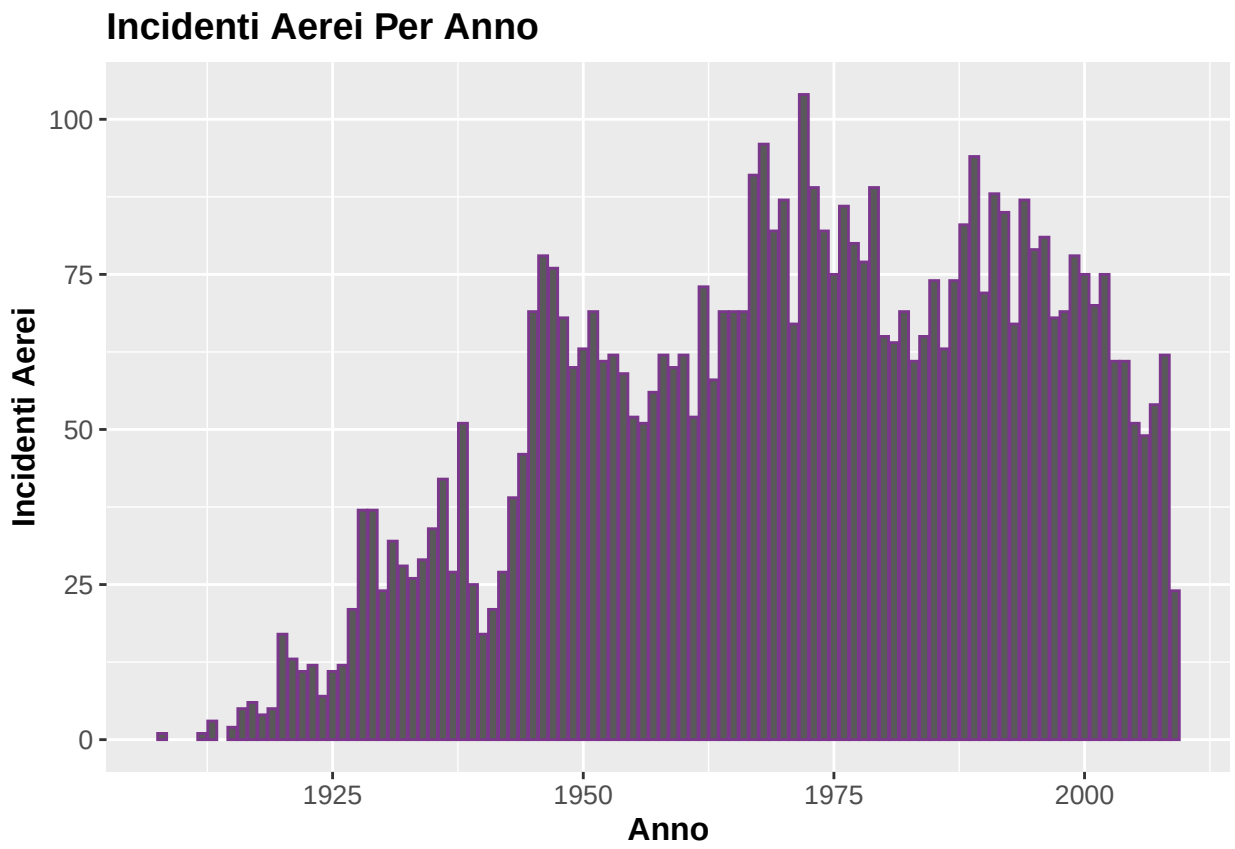
```
                        labels = c("Notturno", "Mattina","Pomeriggio","Sera",
                                   "Sconosciuto"))
# Reset 'NA's' in 'Local Hour'
AirCrash$LocalHour <- ifelse(AirCrash$LocalHour == 25,NA,AirCrash$LocalHour)

# Aggiungo le colonne Anno e Mese
# Il pacchetto 'lubridate' estrae 'Year' e 'Month' da 'Date' come colonne
AirCrash$Year <- (year(AirCrash$Date))
AirCrash$Month <- (month(ymd(AirCrash$Date), label = TRUE))

# Aggiungo due variabili
# Survivors: 'Aboard' - 'Fatalities'
# SurvivalRate: 'Survivors'/'Aboard'

AirCrash$Survivors <- AirCrash$Aboard - AirCrash$Fatalities
AirCrash$SurvivalRate <- AirCrash$Survivors/AirCrash$Aboard
summary(AirCrash)
```

```
##       Date                Time             Location           Operator
##  Min.   :1908-09-17   Length:5244        Length:5244        Length:5244
##  1st Qu.:1954-06-16   Class :character   Class :character   Class :character
##  Median :1973-03-15   Mode  :character   Mode  :character   Mode  :character
##  Mean   :1971-11-19
##  3rd Qu.:1990-07-28
##  Max.   :2009-06-08
##
##    Flight..            Route               Type           Registration
##  Length:5244        Length:5244        Length:5244        Length:5244
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     cn.In               Aboard         Fatalities         Ground
##  Length:5244        Min.   :  1.00   Min.   :  0.00   Min.   :   0.000
##  Class :character   1st Qu.:  5.00   1st Qu.:  3.00   1st Qu.:   0.000
##  Mode  :character   Median : 13.00   Median :  9.00   Median :   0.000
##                     Mean   : 27.57   Mean   : 20.09   Mean   :   1.594
##                     3rd Qu.: 30.00   3rd Qu.: 23.00   3rd Qu.:   0.000
##                     Max.   :644.00   Max.   :583.00   Max.   :2750.000
##
##    Summary            LocalTime                     LocalHour
##  Length:5244        Min.   :2021-04-17 00:00:00   Min.   : 0.00
##  Class :character   1st Qu.:2021-04-17 09:02:30   1st Qu.: 9.00
##  Mode  :character   Median :2021-04-17 13:30:00   Median :13.00
##                     Mean   :2021-04-17 13:16:39   Mean   :12.85
##                     3rd Qu.:2021-04-17 18:15:00   3rd Qu.:18.00
##                     Max.   :2021-04-17 23:59:00   Max.   :23.00
##                     NA's   :2209                  NA's   :2209
##         Daypart          Year          Month        Survivors
##  Notturno    : 382   Min.   :1908   dic    : 514   Min.   :  0.000
##  Mattina     : 879   1st Qu.:1954   gen    : 494   1st Qu.:  0.000
##  Pomeriggio : 969   Median :1973   ago    : 472   Median :  0.000
```

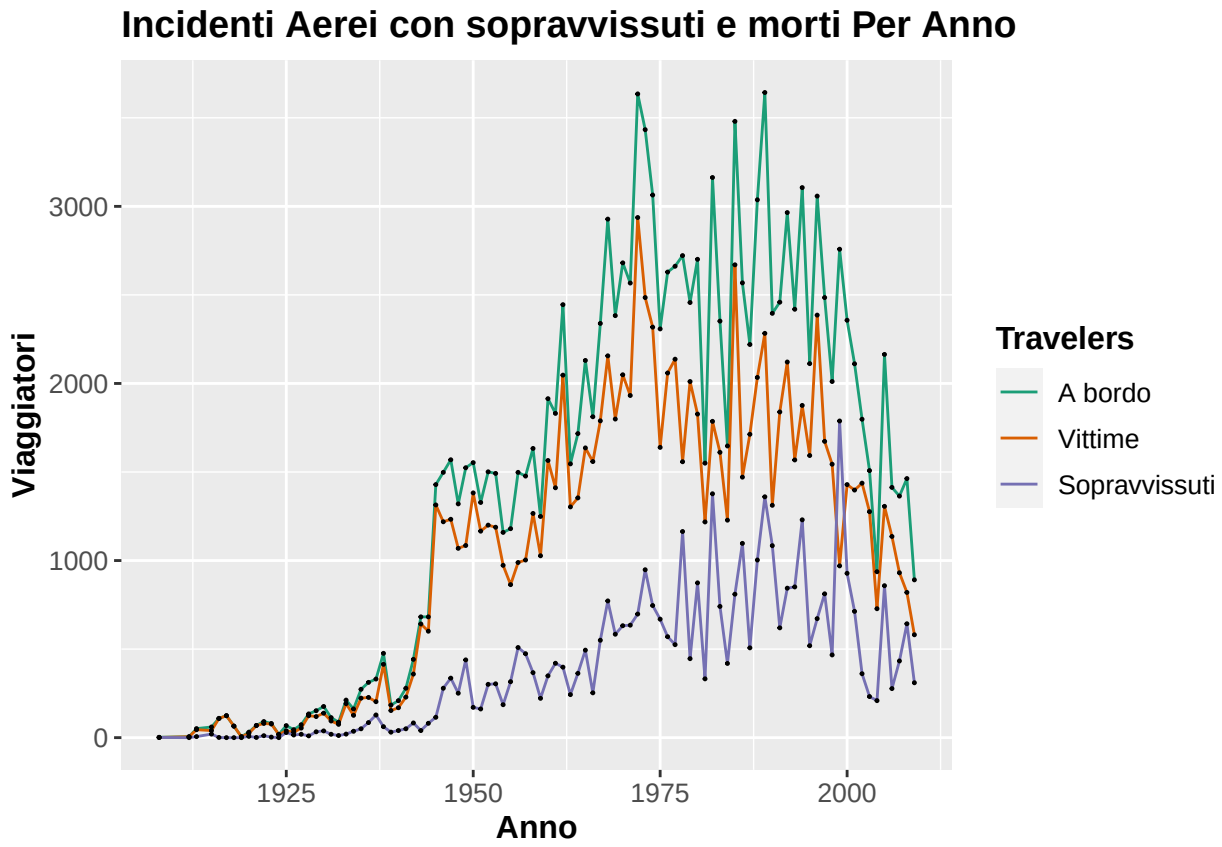
```
##  Sera       : 805   Mean   :1971   set    : 456   Mean   :  7.474
##  Sconosciuto:2209   3rd Qu.:1990   nov    : 453   3rd Qu.:  2.000
##                     Max.   :2009   ott    : 452   Max.   :516.000
##                                    (Other):2403
##   SurvivalRate
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.1651
##  3rd Qu.:0.2000
##  Max.   :1.0000
##
```

Visualizzazione del dataset con grafici che mostrano il numero di incidenti per anno:

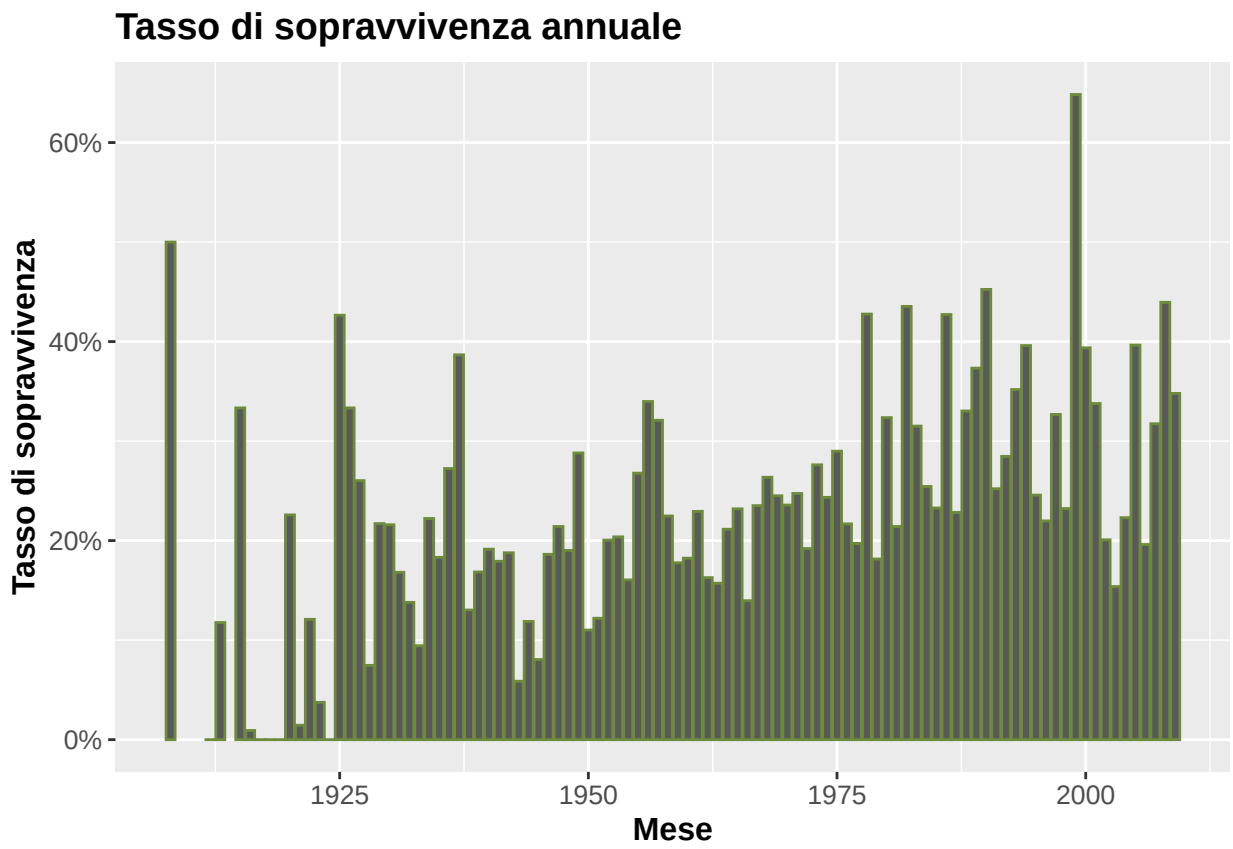
```
fig <- function(width, heigth){
  options(repr.plot.width = width, repr.plot.height = heigth)
}

ATheme <- theme(title = element_text(size = 12, face = 'bold'),
                axis.title = element_text(size = 12),
                axis.text = element_text(size = 10),
                legend.text = element_text(size = 10))

# Grafico che mostra il numero di incidenti per anno
fig(16,10)
CrashesPerYear = ggplot(AirCrash, aes(x=Year)) + geom_bar (colour = "mediumorchid4") +
  xlab("Anno") + ylab("Incidenti Aerei") + ggtitle("Incidenti Aerei Per Anno") + ATheme
CrashesPerYear
```




```
AC <- rbind(Survivors = aggregate(AirCrash$Survivors,by=list(AirCrash$Year),FUN =sum),
            Aboard = aggregate(AirCrash$Aboard,by=list(AirCrash$Year),FUN =sum),
            Fatalities = aggregate(AirCrash$Fatalities, by = list(AirCrash$Year),
                                   FUN = sum))


AC$Travelers <- rownames(AC)
AC$Travelers <- gsub("[.].*","",AC$Travelers)
AC <- AC %>% dplyr::rename(Year = Group.1, Count = x)

# Grafico che mostra il numero di viaggiatori a bordo, sopravvissuti e
# morti per anno
fig(16,10)
TravelersPerYear = ggplot(AC, aes(x=Year, y=Count, group = Travelers )) +
  geom_line(aes(colour=Travelers))  + geom_point(size = 0.2) +
  scale_colour_brewer (palette = "Dark2", labels = c("A bordo", "Vittime",
                                                     "Sopravvissuti"))+
  xlab("Anno") + ylab("Viaggiatori") + ATheme +
  ggtitle("Incidenti Aerei con sopravvissuti e morti Per Anno")

TravelersPerYear
```




```
# Il numero totale di viaggiatori a bordo di incidenti aerei è diminuito nell'
# ultimo ventennio. Nello stesso periodo, anche le vittime sono diminuite.

# Sopravvissuti per Anno
AC2 <- cbind(Survivors = aggregate(AirCrash$Survivors,by=list(AirCrash$Year),FUN =sum),
             Aboard = aggregate(AirCrash$Aboard,by=list(AirCrash$Year),FUN =sum))

fig(16,10)
SurvivorsPerYear = ggplot(AC2, aes(x=Survivors.Group.1, y=Survivors.x/Aboard.x)) +
  geom_col(colour = "darkolivegreen4") +
  xlab("Mese") + ylab("Tasso di sopravvivenza") +
  ggtitle("Tasso di sopravvivenza annuale") + ATheme +
  scale_y_continuous(labels = scales::percent)
SurvivorsPerYear
```



Preparazione Dataset per uso algoritmi di Clustering

```
AllAirCrash <- AirCrash

# Rimuovi incidenti in cui il 'Summary' corrispondente è vuoto
AirCrash <- AirCrash[!AirCrash$Summary == "",]

# Creo un data frame "AirClust" per contenere le variabili da usare nel
# k-means clustering

# Estraggo le variabili di interesse da 'AirCrash' in un data frame 'AirClust'
AirClustx <- AirCrash[,c(10,11,19,20)]
# Creo valori binomiali
AirScore <- data.frame(Year = AirCrash$Year)
AirScore$Y1908_Y1929 <- ifelse(AirScore$Year > 1929,0,1)
AirScore$Y1930_Y1949 <- ifelse(between(AirScore$Year,1930,1949), 1,0 )
AirScore$Y1950_Y1969 <- ifelse(between(AirScore$Year,1950,1969), 1,0 )
AirScore$Y1970_Y1989 <- ifelse(between(AirScore$Year,1970,1989), 1,0 )
AirScore$Y1990_Y2009 <- ifelse(AirScore$Year > 1989 ,1,0)
# Li unisco ad 'AirClust'
AirClust <- data.frame(AirClustx,AirScore[,-1])
head(AirClust)
```

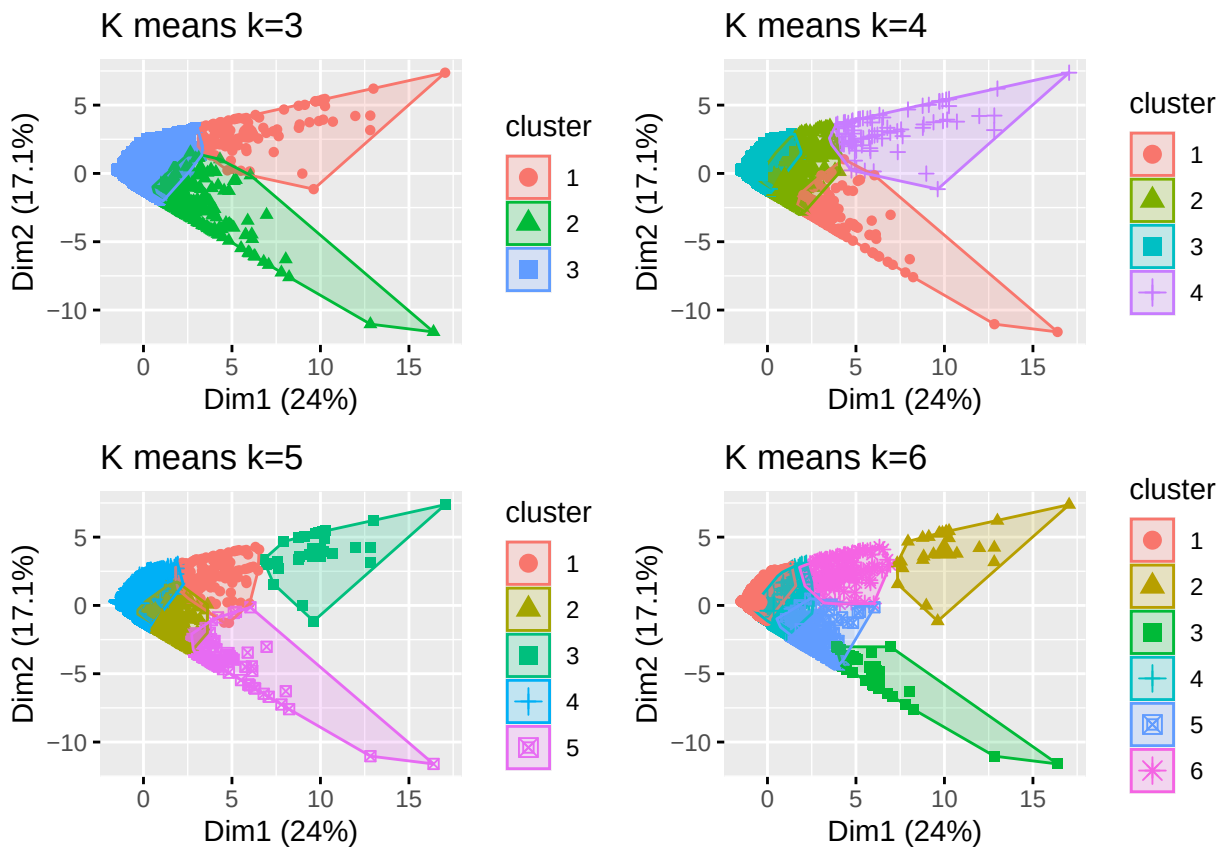
```
##   Aboard Fatalities Survivors SurvivalRate Y1908_Y1929 Y1930_Y1949 Y1950_Y1969
## 1      2          1         1    0.5000000           1           0           0
```

```
## 2      5          5         0    0.0000000           1           0           0
## 3      1          1         0    0.0000000           1           0           0
## 4     20         14         6    0.3000000           1           0           0
## 5     30         30         0    0.0000000           1           0           0
## 6     41         21        20    0.4878049           1           0           0
##   Y1970_Y1989 Y1990_Y2009
## 1           0           0
## 2           0           0
## 3           0           0
## 4           0           0
## 5           0           0
## 6           0           0
```

Algoritmi di Clustering: modellazione e visualizzazione

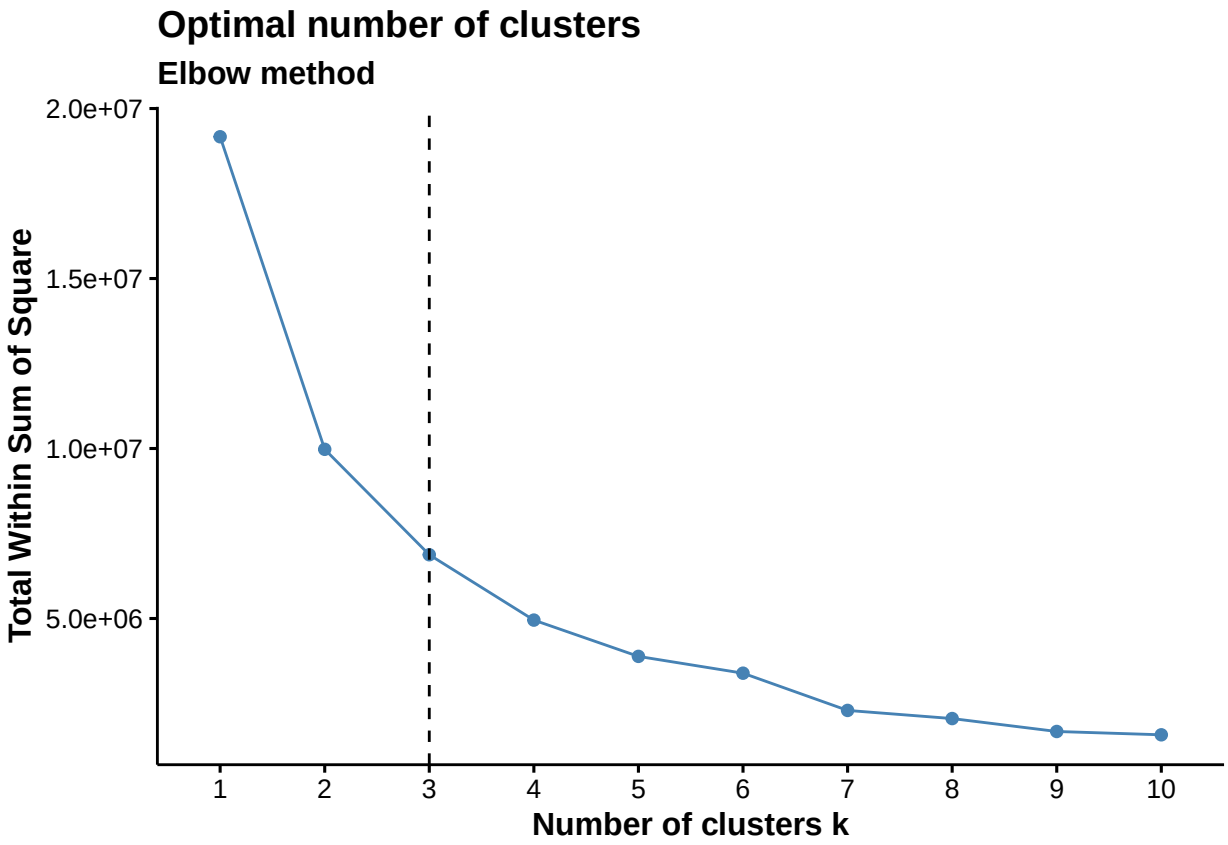
Applico al dataset creato alcuni algoritmi di clustering e procedo alla visualizzazione:

```r
# Comincio a testare il numero di cluster con k = (3, 4, 5, 6)
set.seed(23)
k1 <- kmeans(AirClust, centers = 3, nstart = 25)
k2 <- kmeans(AirClust, centers = 4, nstart = 25)
k3 <- kmeans(AirClust, centers = 5, nstart = 25)
k4 <- kmeans(AirClust, centers = 6, nstart = 25)

# Visualizzo i risultati dei cluster
p1 <- fviz_cluster(k1, geom = "point", data = AirClust) + ggtitle("K means k=3")
p2 <- fviz_cluster(k2, geom = "point", data = AirClust) + ggtitle("K means k=4")
p3 <- fviz_cluster(k3, geom = "point", data = AirClust) + ggtitle("K means k=5")
p4 <- fviz_cluster(k4, geom = "point", data = AirClust) + ggtitle("K means k=6")

grid.arrange(p1,p2,p3,p4)
```

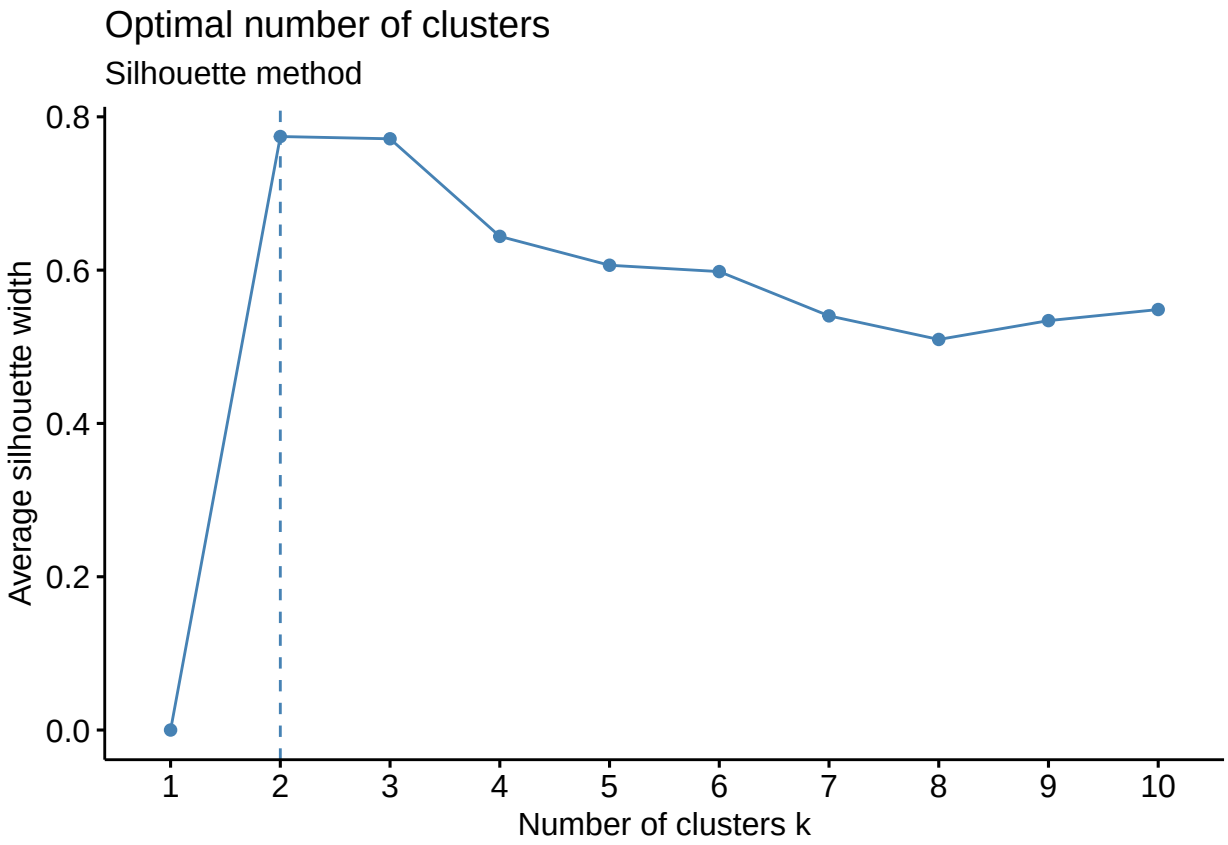


Per determinare il numero ottimale di cluster, utilizzo il metodo 'elbow'. La 'total within sum of squares' viene visualizzata con i cluster di diversa dimensione:

```
# utilizzo il kmeans e scelgo k con il 'elbow method'
fviz_nbclust(AirClust, kmeans, method = "wss") +labs(subtitle = "Elbow method")+
  geom_vline(xintercept = 3, linetype = 2) + ATheme
```




```
# Il numero ottimale di cluster è 3, secondo questo metodo

library(factoextra)
library(NbClust)

#non scalando ottengo k=4 con elbow e k=2 con silhouette
fviz_nbclust(AirClust, kmeans, method = "silhouette")+
  labs(subtitle = "Silhouette method")
```




```
# Il numero ottimale di cluster è 2, secondo questo metodo
```

Esamino i cluster:

```
AirCrash <- data.frame(AirCrash,Cluster = k1$cluster)

AboardClust <- aggregate(AirCrash$Aboard, by=list(Cluster = AirCrash$Cluster),
                         FUN = mean)
AboardClustx <- aggregate(AirCrash$Aboard, by=list(Cluster = AirCrash$Cluster),
                          FUN = max)
AboardClustm <- aggregate(AirCrash$Aboard, by=list(Cluster = AirCrash$Cluster),
                          FUN = min)
DeathClust <- aggregate(AirCrash$Fatalities, by=list(AirCrash$Cluster),
                        FUN = mean)
SurviveClust <- aggregate(AirCrash$Survivors, by = list(AirCrash$Cluster),
                          FUN = mean)
SRateClust <- aggregate(AirCrash$SurvivalRate, by = list(AirCrash$Cluster),
                        FUN = mean)

# Creo un data frame
PCluster <- data.frame(cbind(Cluster =AboardClust$Cluster,
                             Plane_Crashes = k1$size,Max_Aboard = AboardClustx$x,
                             Min_Aboard =AboardClustm$x, Mean_Aboard = AboardClust$x,
                             Mean_Fatalities = DeathClust$x,
                             Mean_Survivors = SurviveClust$x,
                             Mean_SurvivalRate = SRateClust$x))
```

```
PCluster
```

```
##   Cluster Plane_Crashes Max_Aboard Min_Aboard Mean_Aboard Mean_Fatalities
## 1       1           108        517         94   177.60185        13.02778
## 2       2           348        644         69   124.73563       116.98851
## 3       3          4405        101          1    17.76322        13.68536
##   Mean_Survivors Mean_SurvivalRate
## 1     164.574074        0.92870662
## 2       7.747126        0.06121528
## 3       4.077866        0.16134692
```

Metodi alternativi per determinare il numero di cluster:

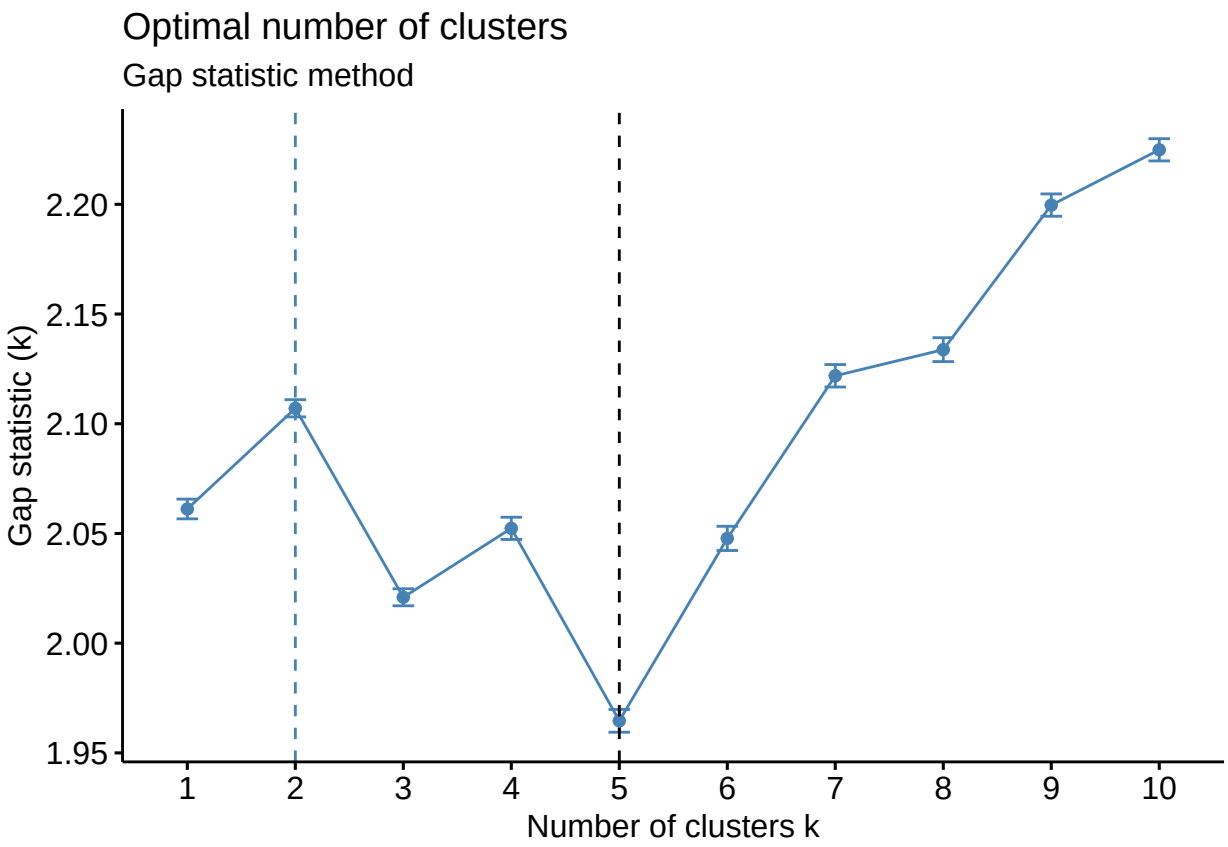
```
# uso la gap statistic
set.seed(123)
fviz_nbclust(AirClust, kmeans, nstart = 25, method = "gap_stat", nboot = 25)+
  geom_vline(xintercept = 5, linetype = 2)+
  labs(subtitle = "Gap statistic method")
```

```
## Warning: Quick-TRANSfer stage steps exceeded maximum (= 243050)
```

```
## Warning: did not converge in 10 iterations
```

```
## Warning: did not converge in 10 iterations
```

```
## Warning: did not converge in 10 iterations
```



```
# valore basso di nboot causa eccessivo carico computazionale
# k=5 usando gap statistic
```

Conclusione

Confrontando i risultati ottenuti, posso concludere che in questo dataset sono presenti 2, 3 o 5 cluster. Il metodo di base è il k-means ma i metodi decisionali sono stati 3: Silhouette, Elbow Method e Gap Statistic.