TU Dortmund

Advanced Multivariate analysis using R

Project : Airplane Crashes and Survival Rate

Lecturers:
Prof. Justyna Brzezińska
Dr. Philipp Doebler

Author: Nancy Bou Kamel

September 4, 2023

Contents

1. Introduction

The fastest and safest way to travel is by plane. However, when aviation accidents do happen, the results are often catastrophic. Aviation accident rates have gone down in recent years, but the growing popularity of private jet travel and helicopter flights may soon reverse that trend. Aviation accidents can be traced to a variety of causes, including pilot error, air traffic controller error, design and manufacturer defects etc... This project is a statistical analysis of a data set scrapped from planecrashinfo.com (Kebabjian, Copyright 1997-2022) with information on 5783 airline crashes between 1908 and 2018. The purpose of this project to group airline crashes by survival rate and passengers, investigate the fatalities over years and find the top 10 airline crashing operators for each cluster. First, silhouette coefficients and k means clustering were used to identify two clusters of airline crashes with different numbers of travelers and survival rate. The highest silhouette coefficient for two groups was 0.65. By using line charts, it was shown that small-sized airline passengers peaked in 1945, whereas mid-size airline passengers peaked in 1970. For mid- and small-size airlines, Aeroflot has the highest number of crashes.

2. Problem statement

planecrashinfo.com (Kebabjian, Copyright 1997-2022) is a real aviation crash web portal where all civil and military airship accidents from 1908 until 2022 are listed. This project involved scraping airplane accidents from 1908 to 2018. The dataset contains 5783 records and 13 variables. Data processing were performed for some variables: date was converted from character to date of format 'year-month-day'. A year and month variable has been added to the dataset. Additionally, aboard and fatalities variables were converted to numeric by extracting the first character from the string. ground was converted from character to numeric. The dataset contains 103 null values after data type transformation: 40 for aboard, 11 for fatalities, and 52 for ground. The null values of the ground variable are replaced by zero, and the rows containing NA values for the aboard column are dropped. Survivors and SurvivalRate are added to the dataset. Survivors were calculated by subtracting aboard from fatalities, while SurvivalRate was calculated by dividing the number of survivors by the number of passengers. The Final dataset consist of 5743 records and 17 variables. All the variables in the dataset are described in Table 1 from Appendix. In this project, k mean clustering is used to classify airline crashes by passengers and survival rate. The Silhouette coefficient was used to determine the optimal number of clusters. Moreover, the top ten airline crashes for both clusters were also obtained, as well as the fatalities of both clusters per year.

3. Statistical methods

3.1. Clustering:

Clustering aims to find groups such that, each item of a certain group share same properties. it is unsupervised task and it can be used in a wide variety of applications: customer segmentation, anomaly detection, image segmentation, search engine etc.. A good cluster quality can be identified by a high intra-class similarity between the items of the same cluster and low inter-class similarity between two or more cluster groups.

3.1.1. K Means

K mean (Geron, 2019, p.243-244) is a Non-hierarchial, unsupervised classification algorithm (unlabeled data). Its purpose is to find K groups based on their characteristics. There are 2 main steps of K means algorithm: First we choose K different cluster centers randomly and assign these to random points in the dataset. Then we go into the loop which consist of 2 steps until it converge (no more change in cluster assignment or centers): minimize the sum of the distance (euclidean or quadratic) between each object and the group of cluster centroid followed by recalculating the mean of each cluster to be taken as new centroids. K Mean requires a certain number of clusters before performing

the algorithm. (Geron, 2019, p.250) and didn't work well when the clusters have varying size, different densities, or non-spherical shapes.

The output of the K mean give the following elements (Ayyadevara, 2018, p.271-274): **cluster** center for each cluster is the mean (midpoint of each cluster) , **totss** - total sum of squares i.e. the sum of squared distance of all points to the data center. **withinss** is a vector of sum of squared distance of each point to its cluster center. **tot.withinss** is the sum of the squared distance between all points and their cluster center $tot.withinss = sum(withinss)$. **betweenss** is the between cluster sum of squares $betweenss = totss - tot.withinss$ and the **size** is the number of points in each cluster.

3.1.2. Silhouette Coefficient

Silhouette coefficient (Geron, 2019, p.248) is used to choose the best number of clusters but its computationally expensive. The silhouette coefficient can be defined as

$$S_i = \frac{c_i - d_i}{max(c_i, d_i)}$$

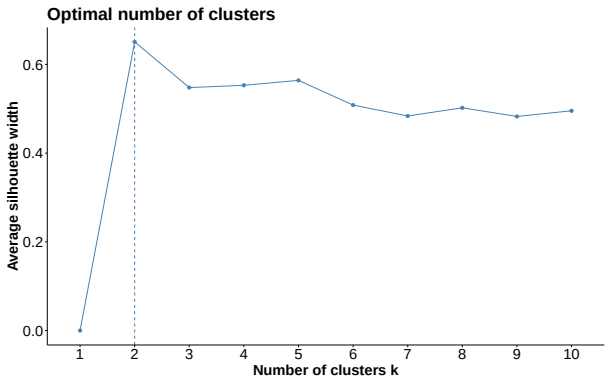
Where S_i represents the Silhouette, a_i is the intra-cluster distance which represents the average distance between instance i and all other instance in the same group and b_i is the mean nearest-cluster distance which is the distance between instance i and all other instances in the other nearest group. The Silhouette coefficient range from -1 to +1. +1 means the instances in a certain group are well apart from other groups, -1 mean that instance are assigned to the wrong cluster and 0 means that the distance between clusters is not significant.

3.2. Software tools

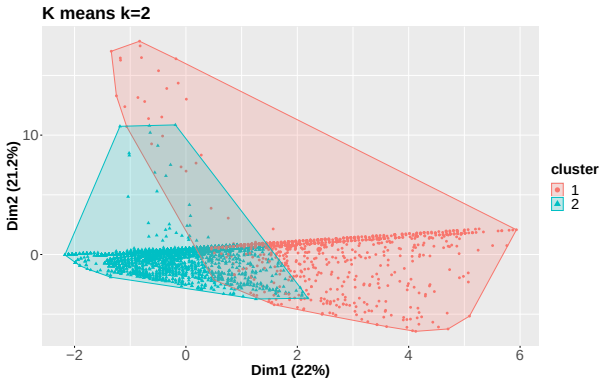
The statistical Software R (Team, 2013) version 4.0.3 was used for this analysis. **lubridate** library (Grolemund and Wickham, 2011) version 1.8.0 is used to convert date from character to date and extract the year and month from the date, **ggplot2** (Wickham, 2016) version number 3.3.2 to create visualizations to detect the airline operators that cause most crashes and analyze the fatalities over years, **cluster** (Struyf, 2022) version 2.1.3 to apply clustering algorithm and find groups of airline crashes, **factoextra** (Kassambara and Mundt, 2020-04-01) version 1.0.7 to determine and visualize the optimal number of cluster using silhouette methods, **readr** (Hester and Wickham, 2022) version 2.1.2 to read comma-separated csv file, **stringr** (Wickham, 2022) version 1.4.1 for data cleaning and preparation tasks, **dplyr** (Mueller and Wickham, 2022) version 1.0.10 for column separation and fixed interval of the year and **gridExtra** (Antonov and Auguie, 2017-09-09) version 2.3 for arranging multiple globs on a page.

4. Statistical analysis

In order to obtain the optimal number of clusters, Silhouette coefficient method is used. The optimal number of cluster for airline crashes is two as show in Figure 1(a). The optimal Silhouette coefficient is equal to 0.65 which is close to 1 and represent a good separation between the two clusters. Figure 1(b) represents the two overlapped clusters. Inspecting the mean values by Aboard, Fatalities, Survivors and Survival Rates, gives a nice depiction on the distinctive features of each cluster. Essentially the clusters can be defined by the number of travelers aboard and their survival rates as shown in Table 2 in Appendix. Cluster 1 travelers range from 10 to 99 with an average of 53.35 aboard and a low mean survival rate 3%. Consequently, Cluster 2 has a range of passengers 0 to 47 with an average of 11.57 and a low survival rate of 14%. Cluster 2 has 4747 airline crashes comparing to 996 accidents in Cluster 1. The between cluster sum of squares is equal to 2476698 whereas, the within cluster sum of squares is equal to 2323224. So cluster 1 is Mid size crashes and cluster 2 is Small size crashes.

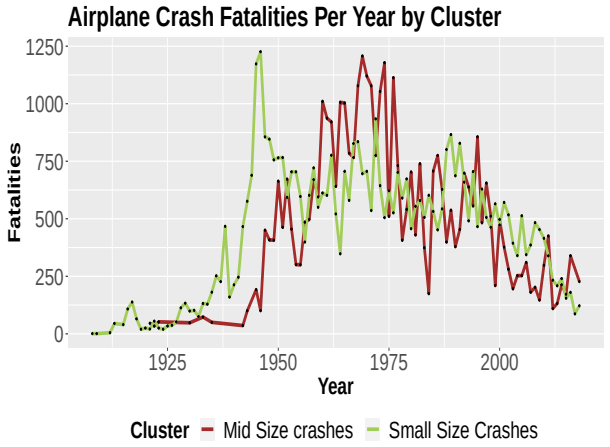


(a) Silhouette Plot

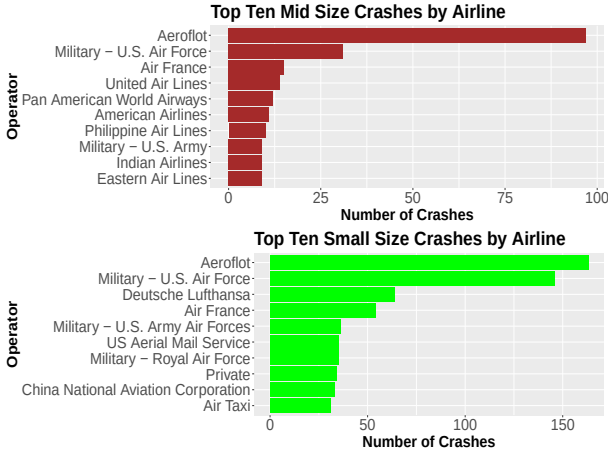


(b) k mean Clustering

Figure 1: Silhouette plot for the optimal number of clusters and K means clustering of airplane crashes into two groups



(a) Airline Fatalities per year by cluster



(b) Top Ten Mid and Small size airline Crashes

Figure 2: Mid and Small size airplane crashes by year and operator

Crash rates for small airlines peak around 1250 in 1945, then it decreases, while mid-size airline crashes peak around 1200 in 1970, then it drop starting in 1978 Figure 2(a). The number of small-size airplane crashes in 2018 is lower than the number of mid-size aviation crashes. A wide mix of airlines make up the top ten list of small and mid-size passenger plane crashes. On the top ten list of mid-size passenger plane crashes, Aeroflot has four times as many fatal crashes as any other operator Figure 2(b). In addition, Aeroflot and Military - U.S: AirFrance cause the most crashes among mid-size airlines.

5. Summary:

The dataset for this project was scraped from the web portal planecrashinfo (Kebabjian, Copyright 1997-2022). It includes 5783 observations and 13 variables about real aviation crashes between 1908 and 2018. As a result of data processing, the survival and survival rate were calculated, the date variable was converted from character to date format 'year-month-year', the year and month variables were extracted from the date variable, the aboard and fatalities variables were converted from character to numeric after excluding the first character. Moreover, null values of ground variables were substituted by zero, whereas rows that contained NA values for aboard variable were removed. After processing, the dataset contains 5743 records and 17 variables. The airline crashes were clustered into two groups according to the number of passengers and survival rate by using K mean clustering and the Silhouette coefficient. The highest Silhouette coefficient for two group clusters was 0.65. Small size airline crashes peaked in 1745, whereas mid-size airline passengers in 1970. Moreover, Aeroflot cause the most crashes for the both clusters. For further investigations, it would be interesting to find the main cause of airline crashes using word clouds for the summary variable to obtain the most frequently appearing words for each cluster.

Bibliography

Anton Antonov and Baptiste Auguie. *Miscellaneous Functions for "Grid" Graphics*, 2017-09-09. https://cran.r-project.org/web/packages/gridExtra/index.html.

V Kishore Ayyadevara. *Pro Machine Learning Algorithms*. Apress, 2018. ISBN 9781484235645.

Aurelien Geron. *Hands-On Machine Learning with Scikit-Learn, Keras, and Tensorflow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, 2019.

Garrett Grolemund and Hadley Wickham. Dates and times made easy with lubridate. *Journal of Statistical Software*, 40(3):1–25, 2011. URL `https://www.jstatsoft.org/v40/i03/`.

Jennifer Bryan , Jim Hester and Hadley Wickham. *readr: Read Rectangular Text Data*, 2022. https://readr.tidyverse.org, https://github.com/tidyverse/readr.

Alboukadel Kassambara and Fabian Mundt. *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*, 2020-04-01. URL `http://www.sthda.com/english/rpkgs/factoextra`.

Richard Kebabjian. Plane crash information, Copyright 1997-2022. URL `http://www.planecrashinfo.com`.

Romain François , Lionel Henry , Kirill Mueller and Hadley Wickham. *dplyr: A Grammar of Data Manipulation*, 2022. https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr.

Kurt Hornik , Mia Hubert , Martin Maechler , Peter Rousseeuw , Anja Struyf. *cluster: Cluster Analysis Basics and Extensions*, 2022. URL `https://CRAN.R-project.org/package=cluster`. R package version 2.1.4 — For new features, see the 'Changelog' file (in the package source).

R Core Team. R: A language and environment for statistical computing, 2013. URL `http://www.R-project.org/`.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL `https://ggplot2.tidyverse.org`.

Hadley Wickham. *stringr: Simple, Consistent Wrappers for Common String Operations*, 2022. http://stringr.tidyverse.org.

A. Appendix

Variable	Data type	Description	Scale
date	categorical	The date of the incident	Nominal
time	categorical	Local time, in 24 hr. format	Nominal
location	categorical	The location of the incident	Nominal
operator	categorical	Airline or operator of the aircraft	Nominal
flight-no	categorical	The flight number of the aircraft	Nominal
route	categorical	The route of the aircraft	Nominal
ac-type	categorical	Aircraft type	Nominal
registration	categorical	ICAO registration of the aircraft	Nominal
cn-In	categorical	The construction or serial number	Nominal
aboard	Integer	Total aboard (passengers / crew)	Nominal
fatalities	Integer	Total fatalities aboard (passengers / crew)	Nominal
ground	Integer	Total killed on the ground	Nominal
summary	categorical	Brief description of the accident	Nominal
Year	Integer	year of the incident	Interval
Month	categorical	month of the incident	Nominal
Survivors	Integer	The number of survivors	Nominal
Survival Ratio	Integer	The survival ratio	Ratio

Table 1: Description of the dataset variables

Cluster	Plane Crashes	Max Aboard	Min Aboard	Mean Aboard
1	996	99	10	53.35
2	4747	47	0	11.57
Cluster	Mean Fatalities	Mean Survivors	Mean Survival Rate	
1	44.14	9.21	0.03	
2	9.27	2.30	0.14	

Table 2: Statistics of airplane crashes K mean clustering