AIRLINE SAFETY

Do past accidents predict future safety?

Applied Data Mining—Final Project

Q. Herron | 8 May 2017

THE QUESTION

- “Can we use data on an airline’s history of incidents and fatal accidents in one time period to predict its safety in the next?”

THE HYPOTHESIS

- H_0: There is no link between past incidents and current safety.
- H_1: Past incidents can predict current airline safety.

THE DATA

- Data source: **fivethirtyeight** (R package)—**airline_safety** dataset
- Observations on 56 airlines
  - Data for two 15-year periods (1985-1999 and 2000—2014)
    - Total incidents, fatal accidents, and number of fatalities
  - Available Seat-Kilometers (ASK), weekly average
    - Distance Flown times Seats Available
    - Indicator of airline size

DATA PREPARATION

- Objective—Make all airlines and their figures easily comparable
- Solution—Express the safety figures “per trillion ASK”

- Method—Adjustment factor = $\frac{1{,}000{,}000{,}000}{Avail.\ Seat\ KM}$. Then multiply each safety figure by the adjustment factor.

ADJUSTMENT: EXAMPLE

Original Data

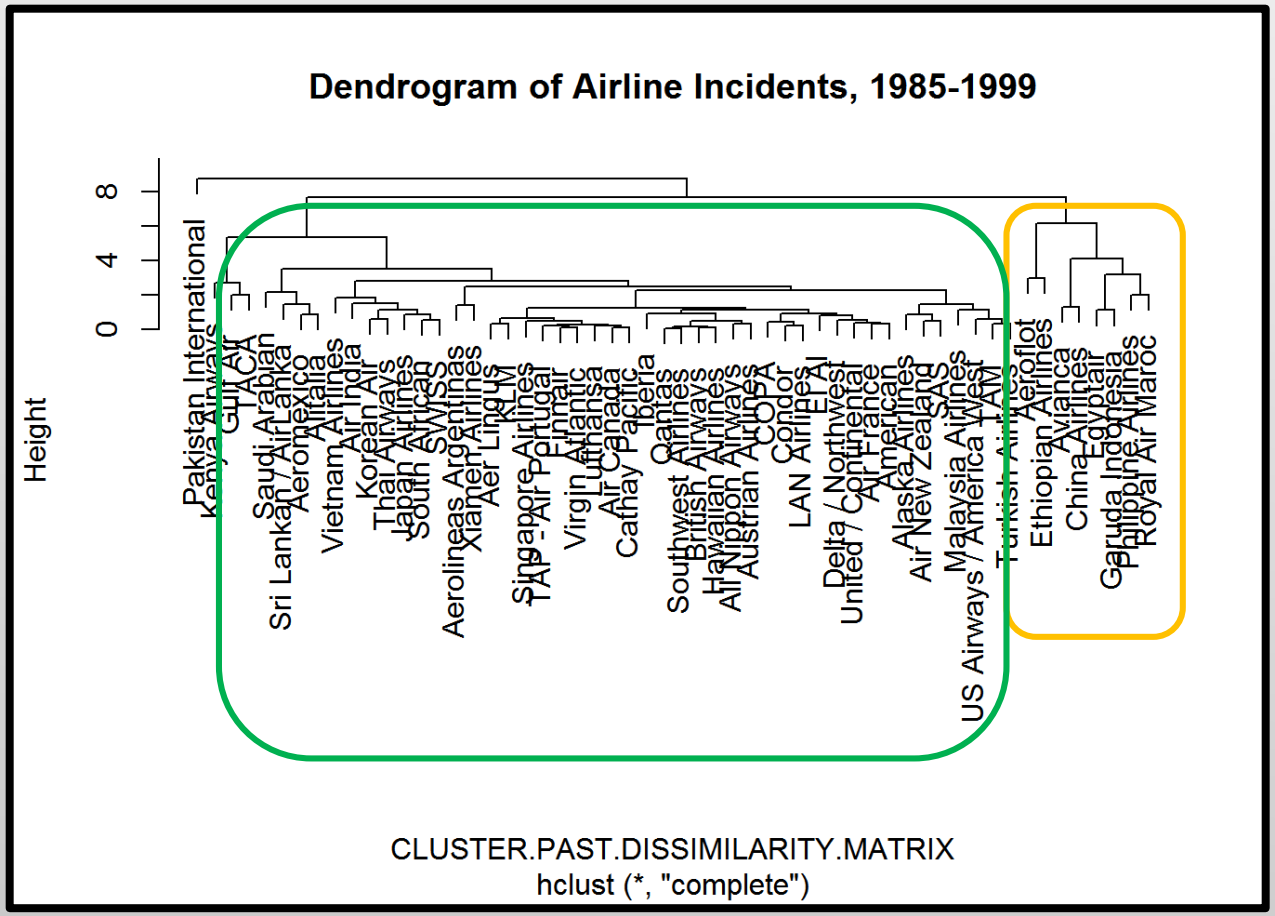
Airline	ASK	ASK Adjust	Incidents (85-99)	Fatal Accidents (85-99)	Fatalities (85-99)	Incidents (00-14)	Fatal Accidents (00-14)	Fatalities (00-14)
United / Continental	7,139,291,291		19	8	319	14	2	109

Adjusted (per trillion ASK)

Airline	ASK	ASK Adjust	Incidents (85-99)	Fatal Accidents (85-99)	Fatalities (85-99)	Incidents (00-14)	Fatal Accidents (00-14)	Fatalities (00-14)
United / Continental	7,139,291,291	0.14	2.66	1.12	44.68	1.96	0.28	15.27

CLUSTER ANALYSIS

- Objective—Sort airlines into clusters based on safety records
- Solution—Use k-Medoids to find best-fit clusters

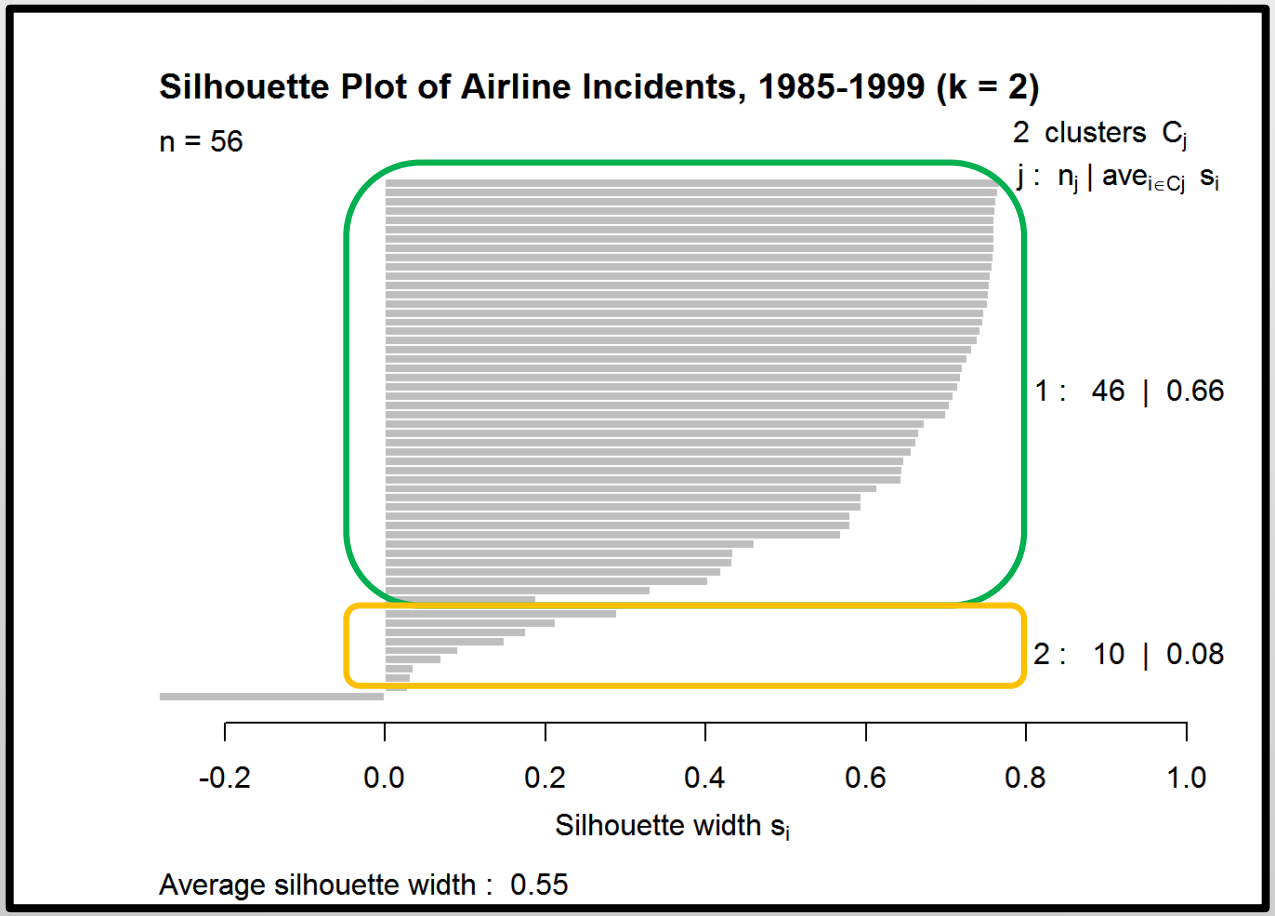


CLUSTER ANALYSIS

Dendrogram—Shows how clusters are formed.

Left side—"safer" airlines

Right side—"riskier" airlines



CLUSTER ANALYSIS

Silhouette plot—Shows each observation as a line. The width of the line shows how well an observation fits into its cluster.

Top cluster—"safer" airlines

Bottom cluster—"riskier" airlines

CLUSTER ANALYSIS

- 1985—1999
  - **Aeroflot** (Russia)
  - **Avianca** (Colombia)
  - **China Airlines**
  - **Egyptair**
  - **Ethiopian Airlines**
  - **Garuda Indonesia**
  - **Pakistan International**
  - **Philippine Airlines**
  - **Royal Air Maroc** (Morocco)
  - Saudi Arabian

- 2000-2014
  - **Aeroflot** (Russia)
  - **Avianca** (Colombia)
  - **China Airlines**
  - **Egyptair**
  - **Ethiopian Airlines**
  - **Garuda Indonesia**
  - **Pakistan International**
  - **Philippine Airlines**
  - **Royal Air Maroc** (Morocco)
  - TACA (Avianca El Salvador)
  - Vietnam Airlines
  - Xiamen Airlines

CLUSTER ANALYSIS: RESULTS

- "Risky" clusters are very similar between time periods
- However, these were two independent clusters
  - Doesn't necessarily imply a correlation

LOGISTIC REGRESSION

- Objective—Predict “risky” airlines today based on past performance
- Solution—Define a “risky” airline, and fit a logistic regression model

- Create a target variable—We will define "risky" as follows:
  - Ten or more fatalities per trillion ASK, between 2000 and 2014
  - This can be user-specified as desired
  - Should not be too restrictive (Need 25-50% positive class)

In the real world, this "risk" is about on par with your odds of winning the Powerball jackpot.

LOGISTIC REGRESSION

LOGISTIC REGRESSION: RESULTS

- No Significance—Past variables were not good predictors
  - p-values from 0.12 to 0.45 ($p < 0.05$ is significant)
- Confusion Matrix—Risk = TRUE
  - Correctly predicted all "non-risky" airlines
  - But only 33% effective for picking "risky"

		Actual	
		TRUE	FALSE
Predicted	TRUE	2	4
	FALSE	0	10

BAGGING

- Objective—Prove whether or not past data can predict risk
- Solution—Use bootstrap aggregating (“bagging”), a complex ensemble model approach

BAGGING

- Method—Run random forest models on multiple subsets of the data. "Majority vote" determines each prediction.

- Out-of-Bag Error—An estimate of how often the model will fail to predict correctly
  - OOB Error estimate is 40%—not looking good

BAGGING: RESULTS

- Confusion Matrix—Risk = TRUE
  - Sensitivity (true prediction) has improved (67%)
  - Specificity (false prediction) dropped to 50%
  - Overall Accuracy is 56%

		Actual	
		TRUE	FALSE
Predicted	TRUE	4	5
	FALSE	2	5

- Kappa Statistic
  - Kappa adjusts Accuracy to account for predictions due to chance.
  - Value of 0.15 suggests that 85% of correct prediction is the result of luck.

CONCLUSION

- Clustering suggested a potential link
  - Most of the same airlines in both “risky” clusters
  - Airlines of poorer countries more dangerous? Look at GDP per capita.
- Regression failed to support a significant link
- Bagging proved no substantial link was present in the data
  - Kappa—most of the correct predictions were due to random chance.

- H_o: There is no link between past incidents and current safety.
- We fail to reject the null hypothesis.
  - Not enough evidence to prove any significant correlation.
  - Inaccurate and unfair to arbitrarily label some airlines—especially "third-world" airlines—as more risky.
- **An airline's current safety cannot be reasonably predicted by past incidents and accidents.**

CONCLUSION

Source: https://fivethirtyeight.com/features/should-travelers-avoid-flying-airlines-that-have-had-crashes-in-the-past/