Predictive Risk Modelling in Aviation Incident

The 10th Montreal Industry Problem Solving Workshop

Global Aviation Data Management, IATA

Hyuntae Jung
Andrea Mulone
Yuval Yakubov

Overview

1. Aviation Safety Data Management
2. Problem for the Workshop
3. Case Study
4. Data for the Workshop

IATA in brief

International Air Transport Association

- **Global** trade association for the world's **airlines** founded in 1945, representing **82%** of global traffic.
- To **represent, lead** and **serve** the airline industry.
- **Global Aviation Data Management** (GADM) is IATA Safety Data and Analytics program.

  GADM is a unique **global aviation safety database** with IATA serving as a custodian trusted by the industry to do this.

IATA ADX Accidents

IATA FDX Flight Data eXchange

IATA IDX Incident Data eXchange

IATA

Aviation, an engine for the economy

35%
of
global
trade

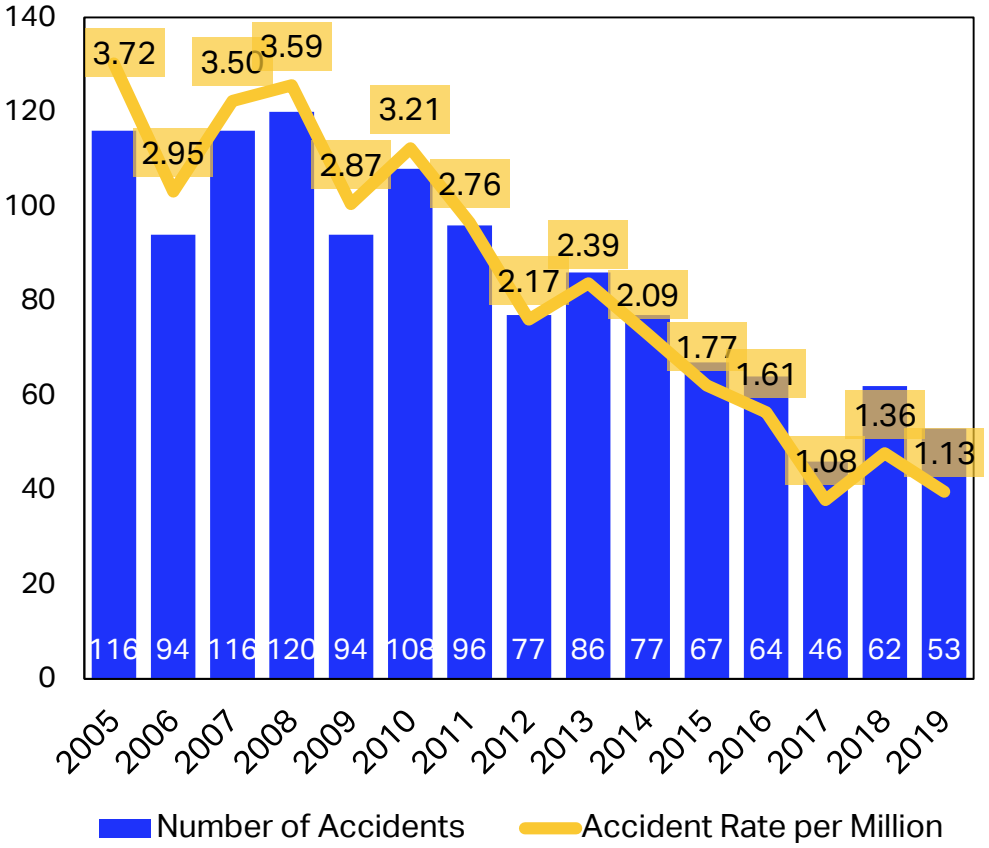
57%
of
world
tourists

65.5
million
jobs
supported

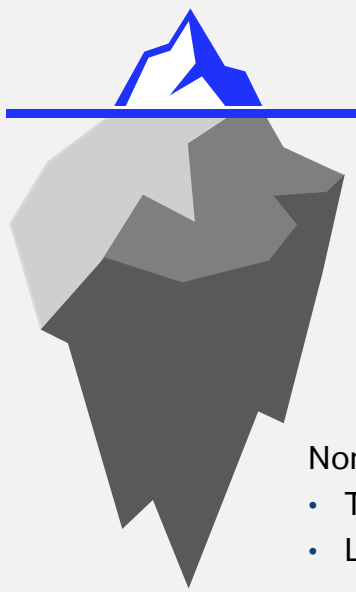
Facilitating
international
supply chain
and **business**

Flying is the safest mode of transportation and data is the key for effective safety risk management.

Accident rate in continuous decreasing trend, but we want to understand what is beneath the surface



Safety risk is a complex interaction of hidden factors under the surface.



1
Major Accidents

30
Incidents, near-miss

300
Hazardous Conditions

3,000
Unreported "Unsafe Acts"

Normal operations
- Threats and Errors
- Latent Conditions

Support decision making process by automating integrated data analysis.

Data Analytic Use Case in Aviation Safety and Security

Flight Data Analytic

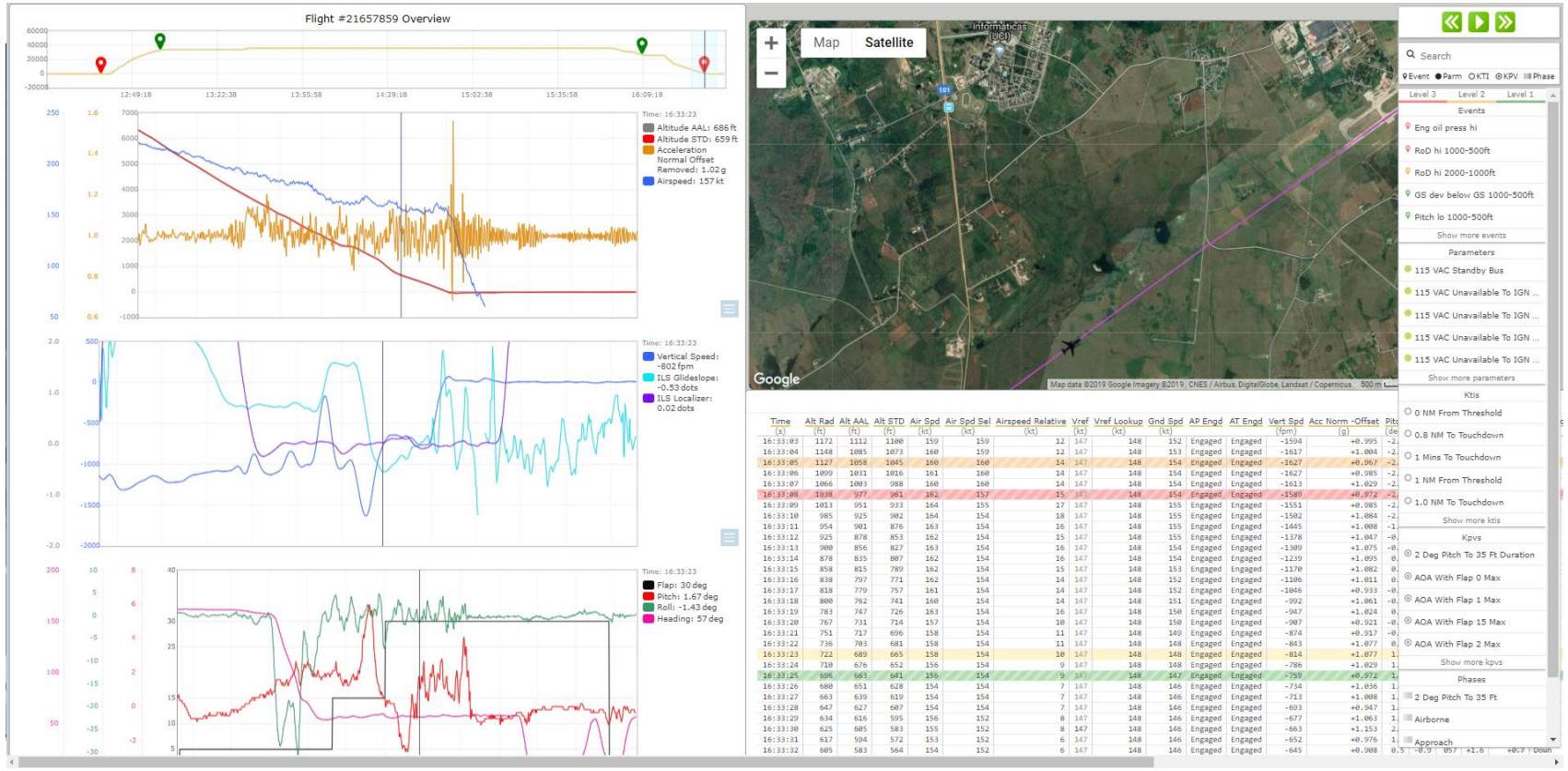
Support decision making process by automating integrated data analysis.

Proactive Safety/Security Risk Management

Risk Identification

Aviation Safety/Security Data

- Accident Report
- Flight Data
- Weather
- Airport & Airspace



Challenges for Global Aviation Safety Data Analysis

Support decision making process by automating integrated data analysis.

Proactive Safety/Security Risk Management

Risk Identification

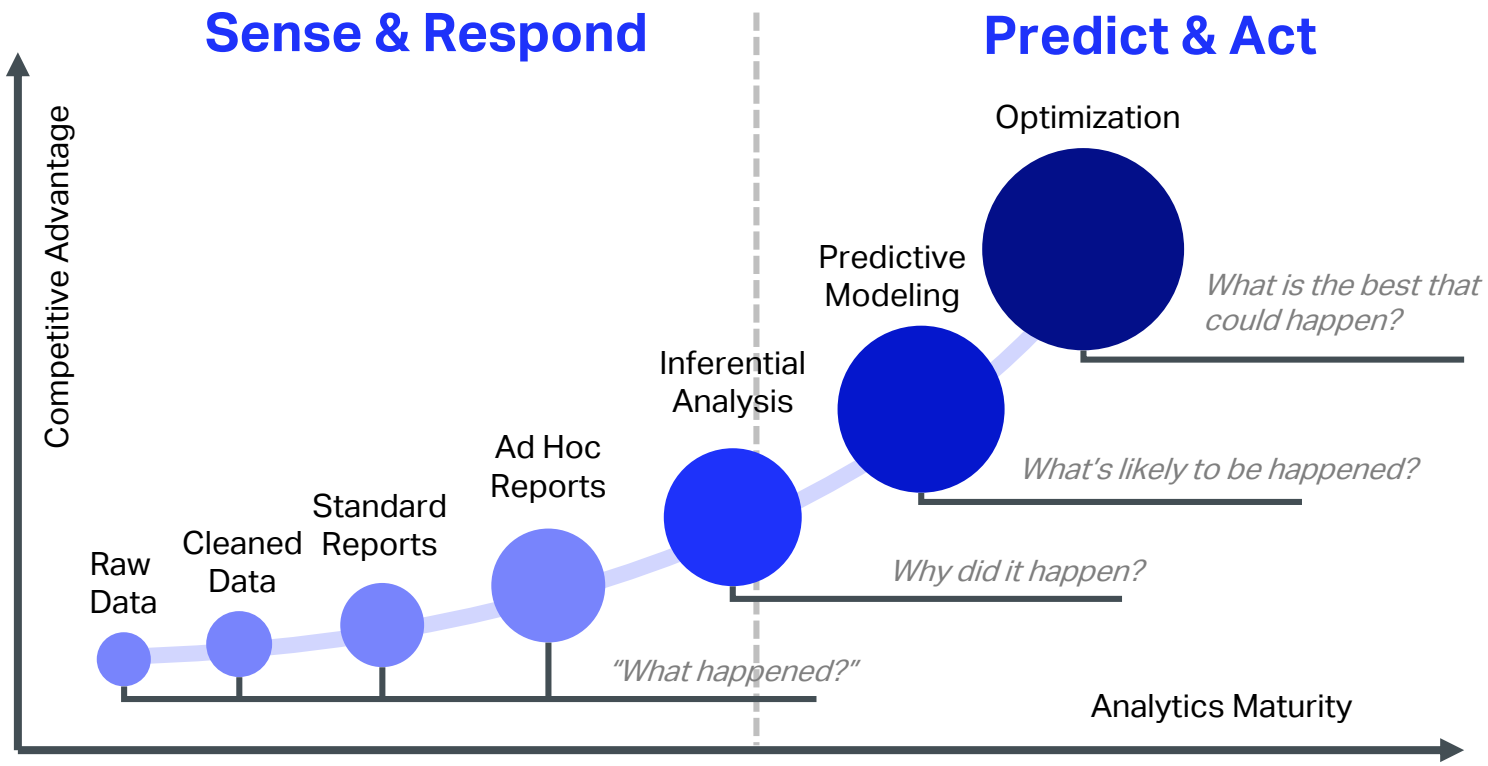
Aviation Safety/Security Data

- Accident Report
- Flight Data
- Weather
- Airport & Airspace

Not Achieving Proactive Safety Risk Management	»	Limited capability of delivering the "right message at the right time" – challenge for proactive risk management.
No Automation Support in Continuous Monitoring	»	Lack of automated anomaly detection makes data-driven risk identification challenging especially in a global scale.
Data Quality and Standardization	»	Time-consuming and expensive integration of unstructured data – collected from voluntary reporting programs.

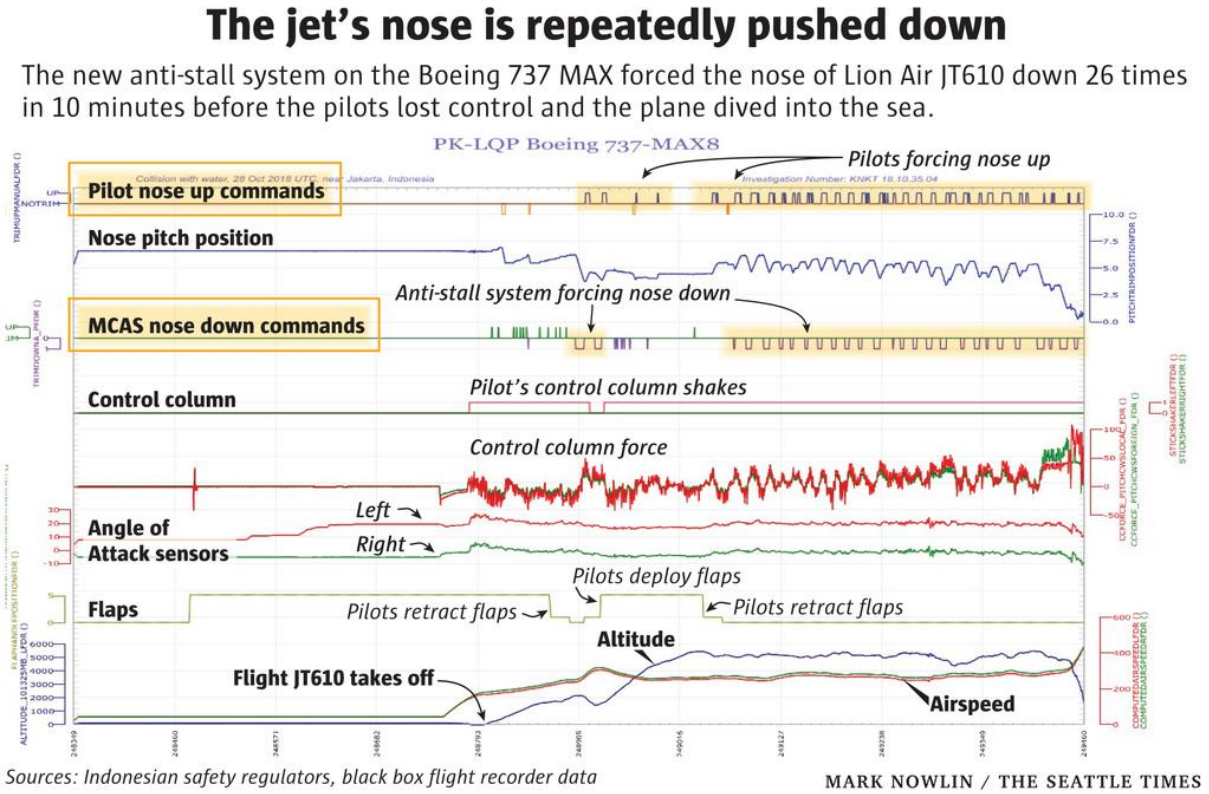
Challenges for Global Aviation Safety Data Analysis

Support decision making process by automating integrated data analysis.



Challenges for Global Aviation Safety Data Analysis

“We don’t know what can be the problem until we face the problem.”



Challenges for Global Aviation Safety Data Analysis

"COVID-19 is brining up the new risk area – we want to detect anomalies to closely monitor the situation"

Bloomberg

Business

FAA Warns of Tail Strikes, Off-Course Flying by Near-Empty Jets

By Alan Levin
June 12, 2020 4:00 AM GMT-4

- Incidents in pandemic show risks of light planes, parked craft
- Regulator, industry have increased vigilance to reduce hazards

EASA published Review of Aviation Safety Issues Arising from the COVID-19 Pandemic

12 Jun 2020 POPULAR

News category: Coronavirus CoVID-19

The European Union Aviation Safety Agency (EASA) is working closely with Member State regulators and industry partners to identify the new or emerging safety issues

Downloads

IATA's Problem for the 10th Montreal IPSW

Predictive Risk Modelling in Aviation Incidents

IPSW Challenge Target

The goal for this challenge is to define the safety risk model, which can detect the potential precursors or anomalies compared to the previous trends by drilling down into multiple subsets of attributes.

1. Anomaly Detection

First, we want to be able to assign risk for the certain attribute or the combination of multiple attributes.

For example, whether a certain aircraft type or airport may have significantly different risk pattern than the others. This is to raise a flag to human analyst to have a deeper look.

2. Predictive Analysis

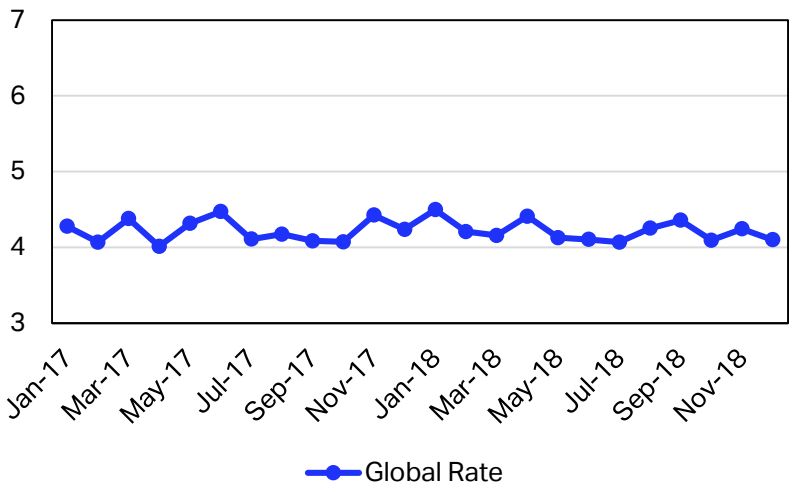
Also, we want a way to be able to predict how much events will happen in a month for a certain descriptor. The goal of this is not to just provide a point estimate but also a confidence interval.

For example, If the dataset contains a certain number of incidents for February which falls outside this confidence interval, then this should be flagged.

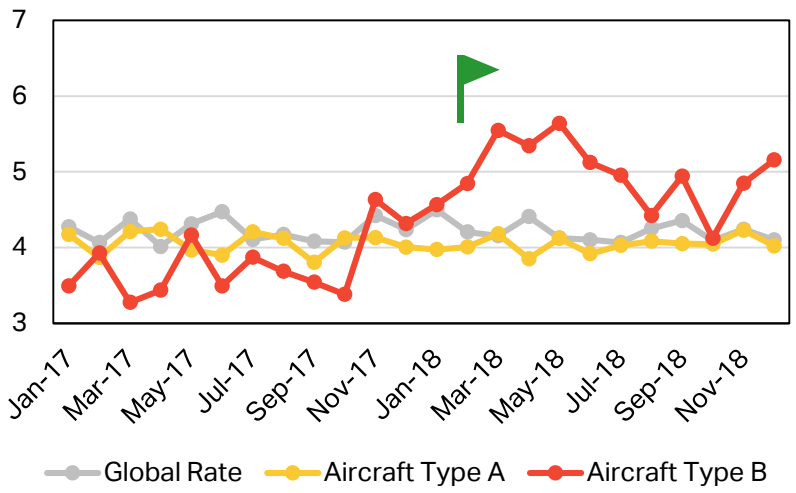
IPSW Challenge Target – (1) Anomaly Detection

Model to give hints to safety analysts where to look before querying every criteria one-by-one.

Example: **Global Aggregated** Hard Landing Rate



Example: Hard Landing Rate **by Aircraft Type**



The model examines the set of incidents by drilling down into specific aircraft type, finding:

- Aircraft Type A does not show significant difference to the global rate
- Aircraft Type B shows a spike over the global rate (green flag), which may indicate prominent safety risk.

Once the model automatically identify such "anomaly" with statistical evidence, the flag will be raised, so that human safety analysts can perform deeper investigation.

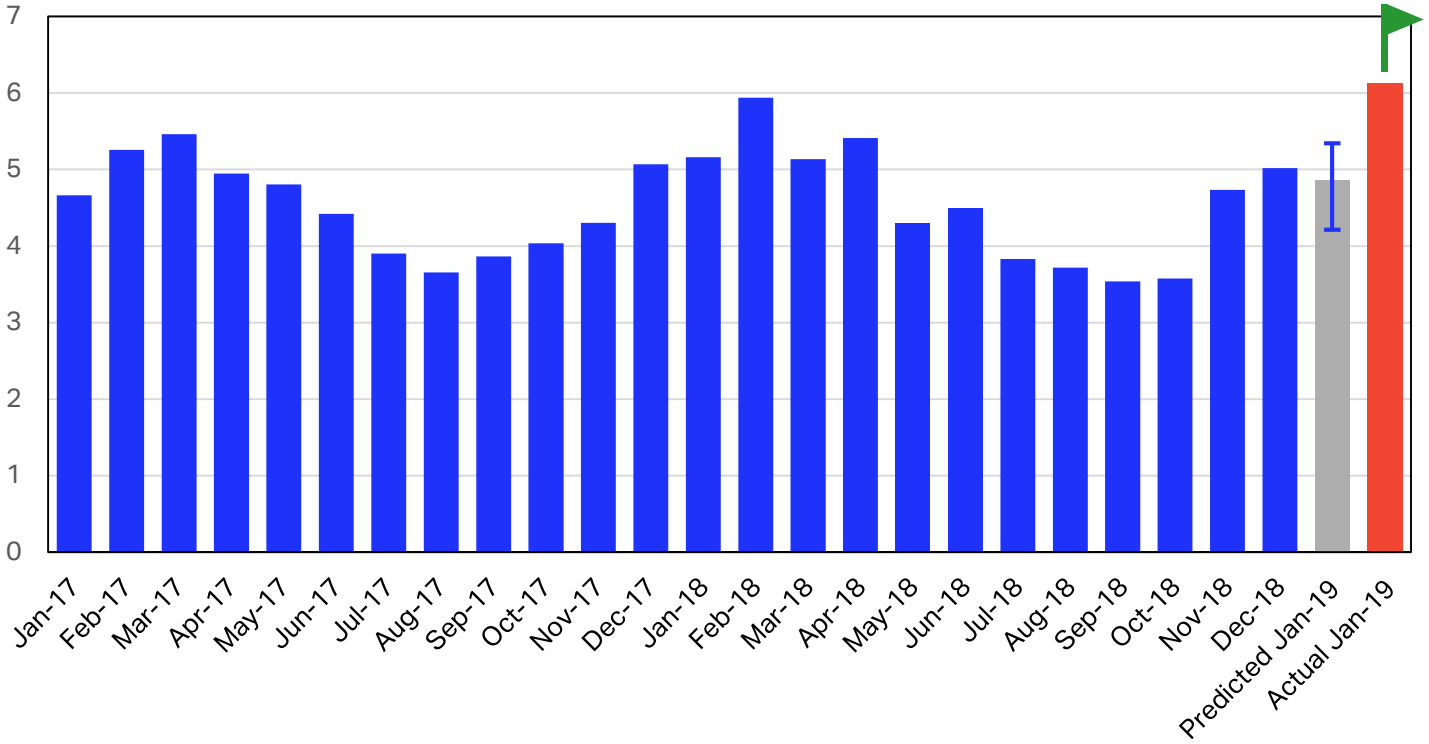
Slicing the data with combinations of multiple attributes

- Events
  (e.g. Hard Landing, Engine Overheat, etc.)
- Date of Occurrence
  (e.g. Seasonal factors)
- Geographic
  (Region, Country, Airport Level)
- Aircraft Type
- Flight Phase

IPSW Challenge Target – (2) Predictive Analysis

Model to predict event rate based on historical records, and flag if the actual rate is exceptional.

Example: Monthly rate of Event A (with the seasonal pattern)



In this example, the historical data for Event A shows seasonal pattern – higher rate in the winter season and lower rate in the summer season.

After trained by the 2 years of historical incident data, the model makes a prediction for January 2019, with given interval of confidence. However, the actual data from January 2019 was out of the boundary.

If the actual rate of the certain incident is out of the boundary, this may mean there might be significant change in the risk profile. The model will flag this finding, so that human safety analysts can perform deeper analysis.

Data – Incident Reports

We prepared 621K reports from 2013 Q1 to 2018 Q2 from IATA aviation safety incident database exported in Excel format. For deidentification purpose, we removed and masked sensitive columns.

Each column represents the following information:

- **Report ID**: unique # of report
- **Year**
- **Month**
- **Fleet Family**: (deidentified) the reported aircraft type.
- **Location**: (deidentified) the reported airport.
- **Location Country**: (deidentified) the country of the airport reported.
- **Phase**: The phase of flight of the aircraft when the event occurred.
- **Event**: the reported type of incident (632 different events)

For example, one report contains the following values:

- **Report ID**: 7723515
- **Year**: 2017
- **Month**: October
- **Fleet Family**: ACType1
- **Location**: Airport162
- **Location Country**: Country256
- **Phase**: Approach
- **Event**: Weather - Windshear

Data – Sector (Flight Segments)

Sector data is the number of flight segments flown of the contributors to incident dataset by quarter, aircraft type, departure and arrival airport. This sector data is used in IATA to normalize the incident rate in a global, country and airport level as below:

STEADES Sectors Calculation

Let J_c be the set of airports in the State c and $I_{j,q}$ be the set of operators operated in the airport $j \in J_c$ in the quarter q. $\delta(i,q) = 1$ if the airline i submitted one report or more than one reports successfully to STEADES database in the quarter q and $\delta(i,q) = 0$ else. Then the **STEADES sector** $S_{c,q}$ of the State c in the quarter q is calculated as below:

$$S_{c,q} = \sum_j \sum_i \left[\left(0.5 \cdot \left(dep_{i,j} + arr_{i,j}\right)\right) \cdot \delta(i,q)\right], \forall i \in I_{j,q}, \forall j \in J_c$$

Based on the STEADES sectors, **the rate per 1,000 STEADES flights** $r_{c,q}$ of the State c in the quarter q is calculated as below:

$$r_{c,q} = \frac{1000 \cdot A(x)}{S_{c,q}}, A(x) = \{x \in X | x_j \in J_c \wedge x_q = q\}$$

Where X denotes the set of queried STEADES reports, x_j and x_q represents the airport of occurrence and the quarter of occurrence in report x respectively.

For example, one records from Sector contains values:

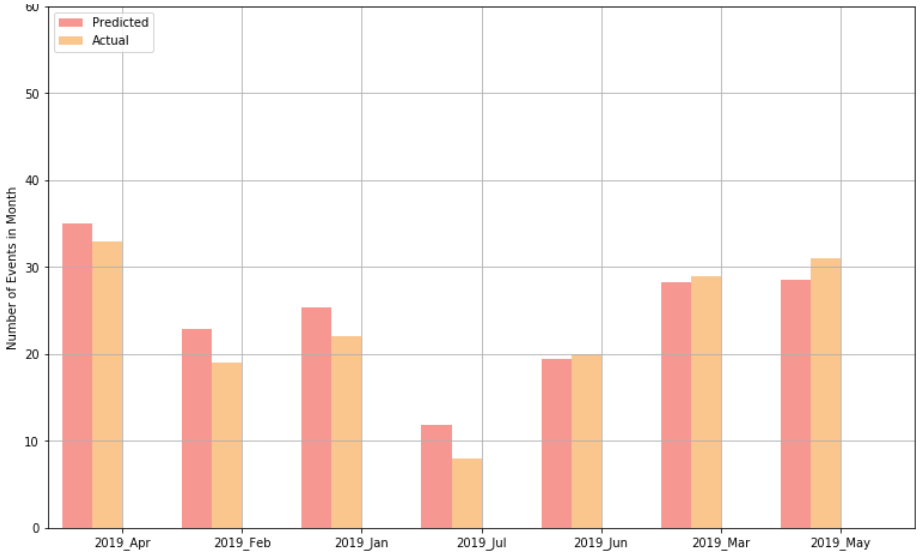
- **Quarter**: 2018 Q2
- **Fleet Family**: ACType5
- **Departure**: Airport162
- **Departure Country**: Country256
- **Arrival**: Airport359
- **Arrival Country**: Country26
- **Sectors**: 3,631

Case Study – Incident Rate Prediction

Generalized Linear Model (GLM)

- As an internal experiment, we focused primarily on GLM models because incidence rate ratio and odds ratio can be extracted from these models, and then used for risk modelling.
- Poisson regression was not a feasible model: Cameron – Trivedi dispersion test failed.
- Failure to reject the null hypothesis in the C-T test indicates that an assumption is violated, and our model is either under dispersed or over dispersed.

Negative Binomial Model



The final model that we used was the Negative Binomial model, this model passed the chai squared test as well as the Cameron Trivedi dispersion test.

Observation & Suggestion

Given the previous experiments that we did, some of our observations are:

1) Find risk metrics for attributes in our data that aren't odds ratio
2) Find better times series predication models. GLM's can provide times series predictions but in most of the cases will not perform as well as seasonal time series, or deep learning models.

[1] F. Bati and L. Withington. 2019. "Application of Machine Learning for Aviation Safety Risk Metric." Paper presented at the *17th IEEE International Conference on Dependable, Automatic and Secure Computing (DASC 2019), Fukuoka, 2019*

[2] F. Famoye, J. T. Wulu and K. P. Singh. 2004 "On the Generalized Poisson regression model with an application to accident data." *Journal of Data Science* 2: 287-295.

Question and Answer