Airplane Crashes

by Revathy Natarajan

Introduction

The project is about analysing, exploring and finding any interesting trends in airplane crashes from 1921 till 2016 (August). Over 95 years of data including military transport accidents, cargo flights, commercial accidents, private jet and helicopter accidents. This also includes world war-II accidents (1939 - 1945). After dropping some NA values and data mining and the below figures are considered for analysis

- 4941 crashes
- 99,857 fatalities
- 139,492 aboarded
- 39,635 survived
- 249 countries
- 42 Manufacturer Aircraft Types

Problem Statement:

Investigating Airplane crashes for across all years and see what is the trend for recent years(20/10)

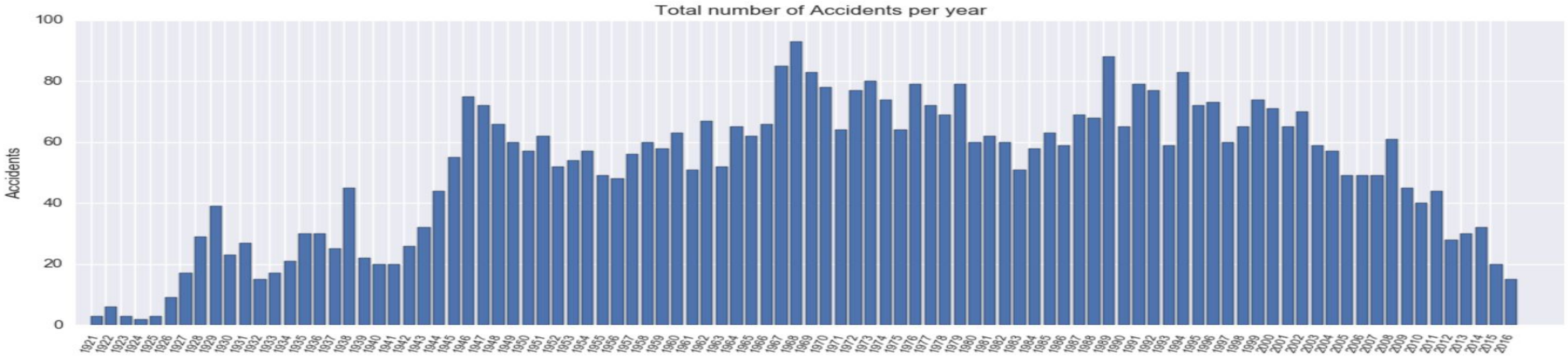
Goals

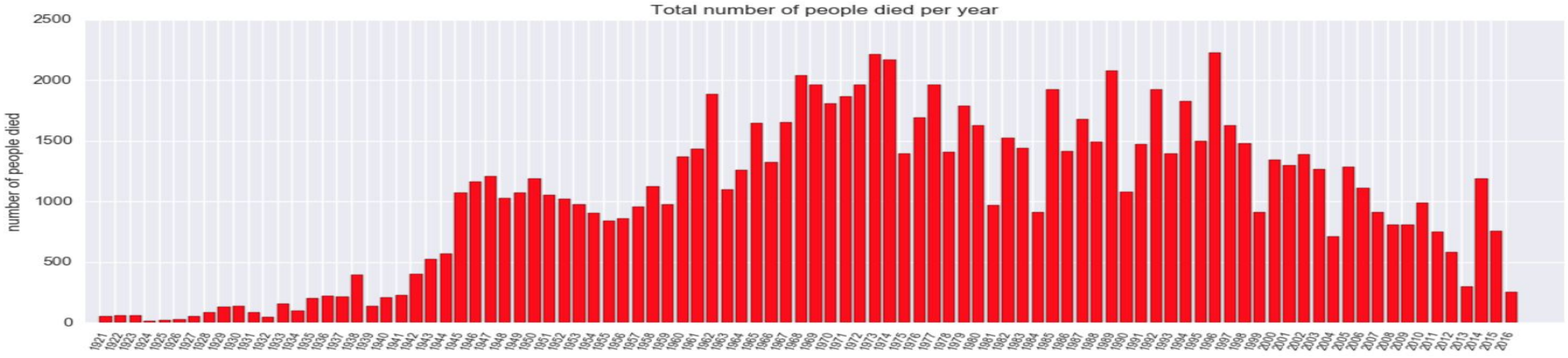
- Total number of Accidents per year
- Total number of Fatalities per year
- Total number of Survivals per year
- Countries with worst accidents

Predictive Modeling:

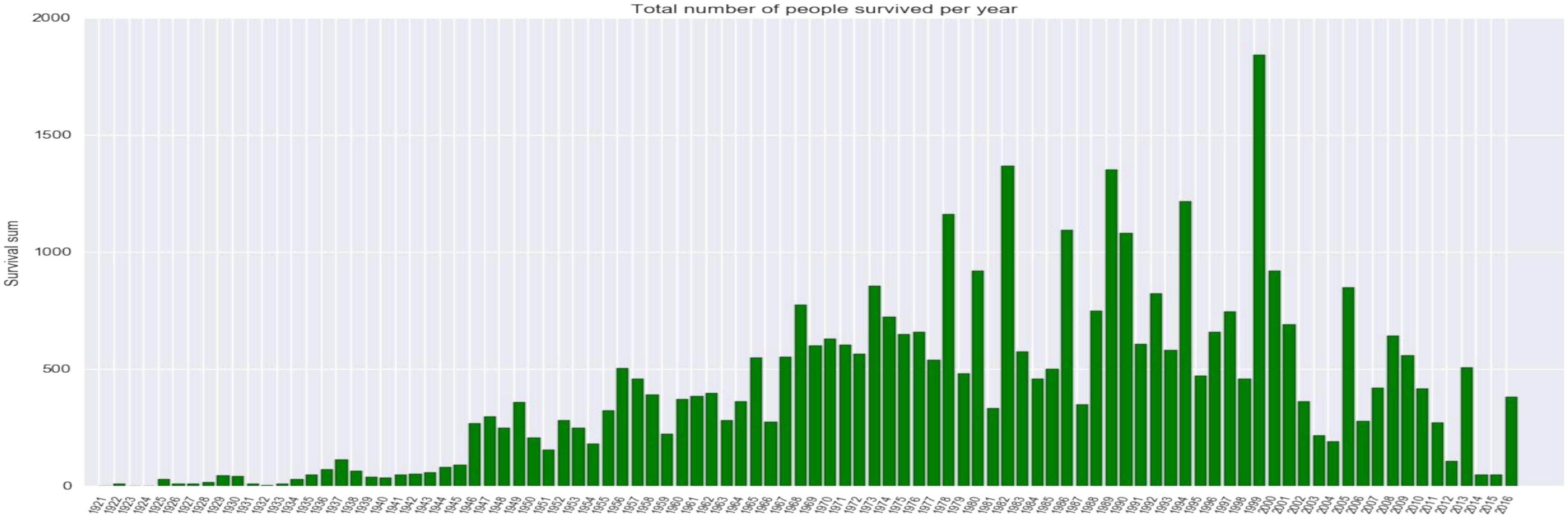
- Fatalities Range
- Survival Range
- Time Series for Fatalities per year

Fatalities, Accidents per year



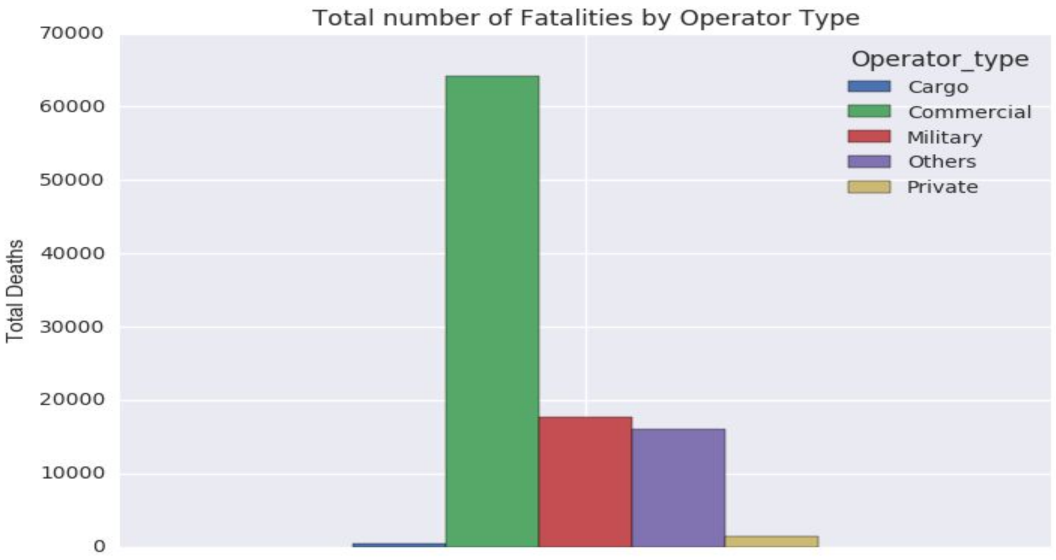


Survivals per year



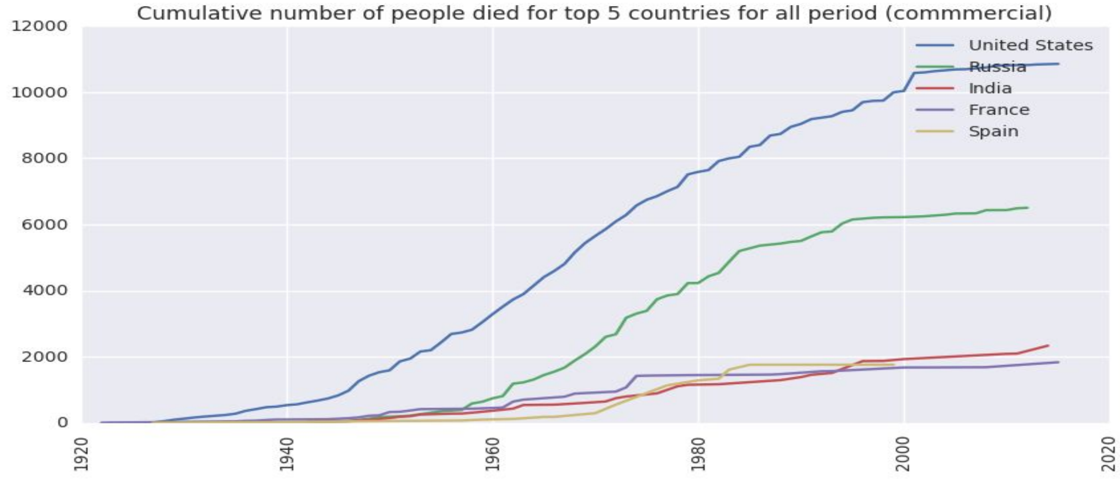
- 1999 survival is high - because the crashes were due to take-off, landing and the chance of survival is high.

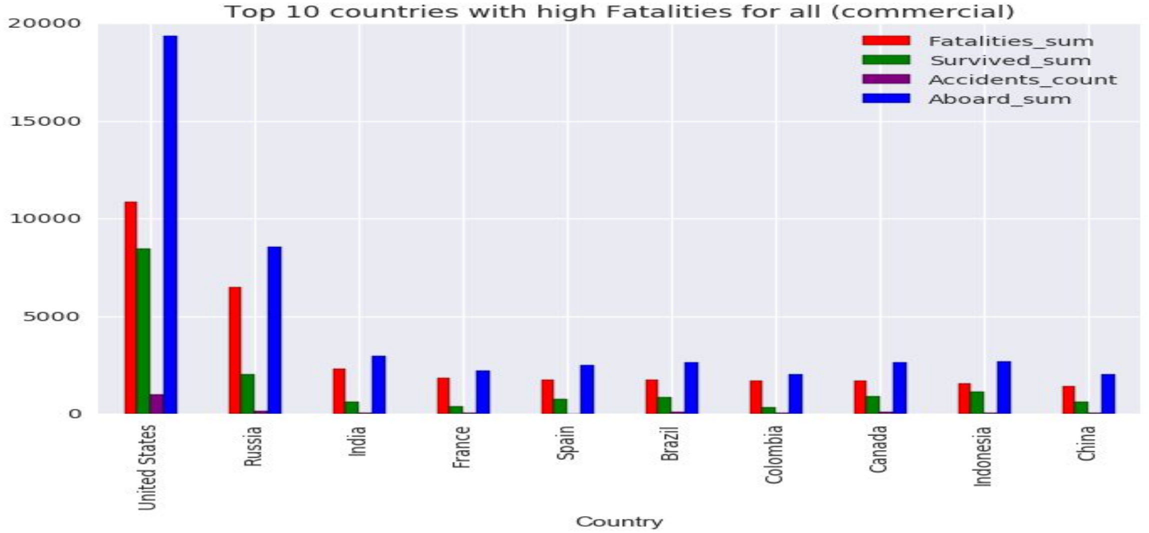
Operator Type



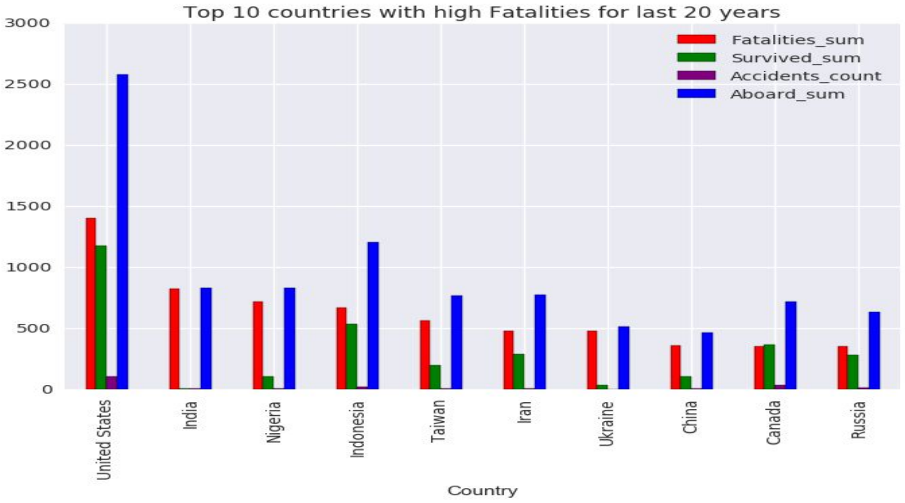
- Commercial flights only for further analysis
- Lot of accidents were Military planes, cargo and privates
- Military crashes include world war II, Iraq war and some training flights
- Something which are not in commercials classified as others
- Still commercial numbers are high

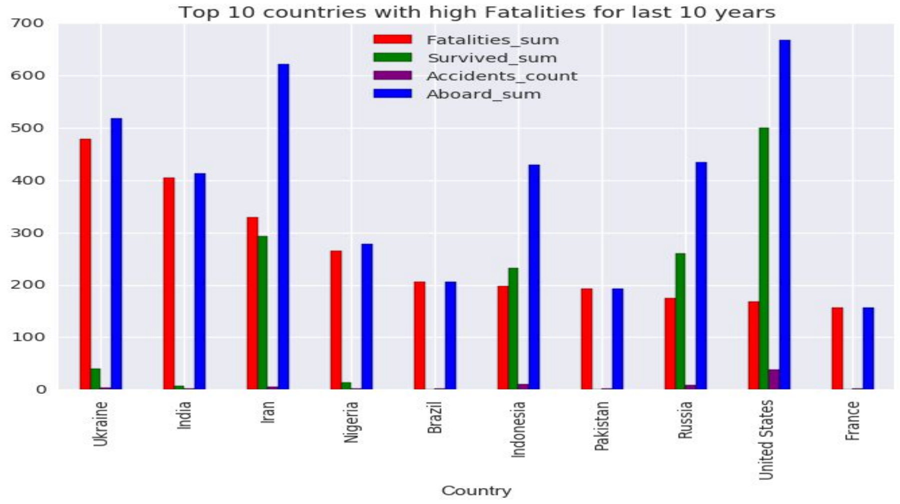
Countries (Commercial) with worst accidents





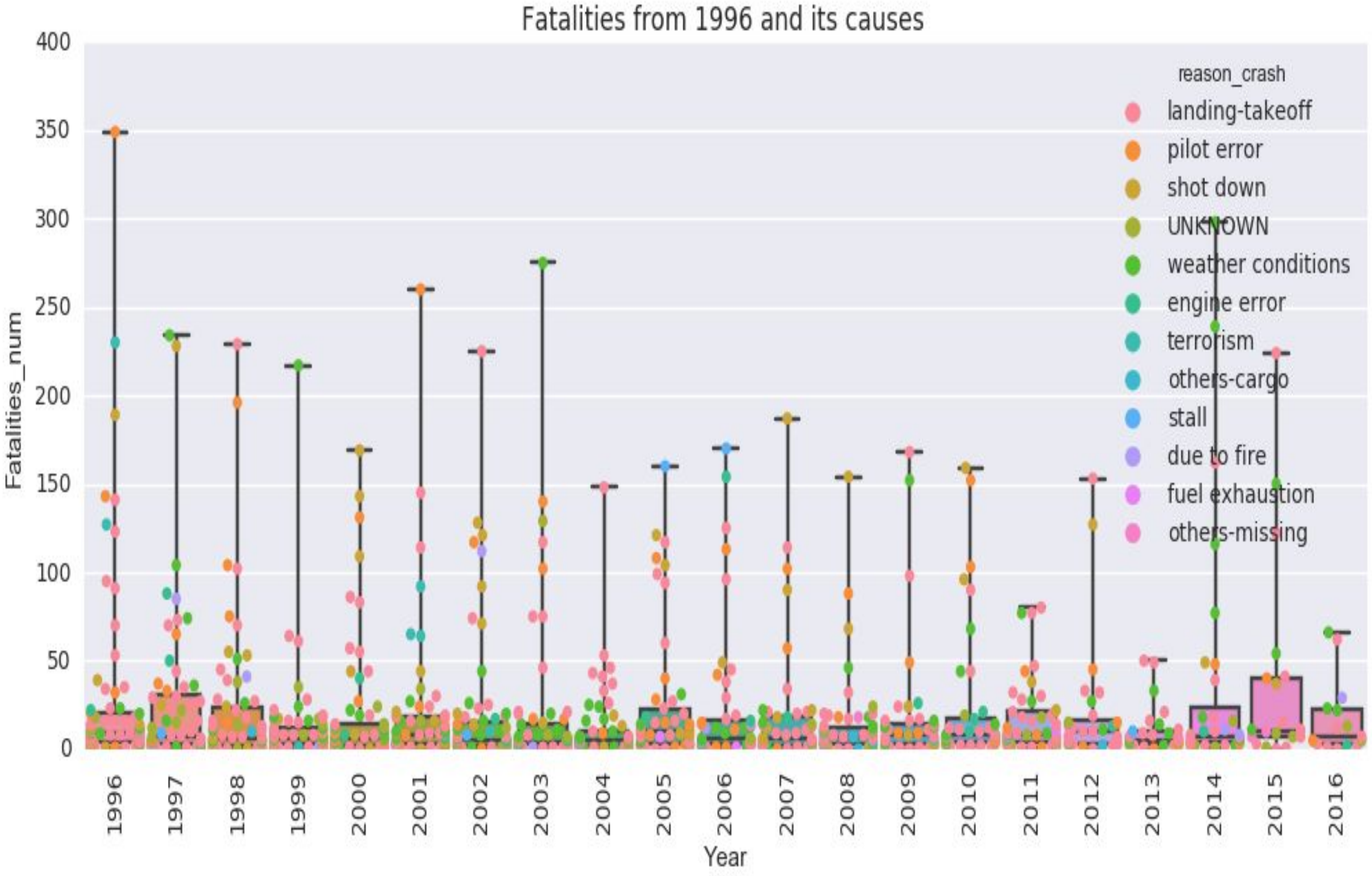
- If we look at the cumulative number of people died over the years for the worst places, upto year 2000 it was increasing rapidly.
- After 2000 it has become steady which means less number of people died due to crashes these recent years in commercial airline.
- It is due to improved skill set of pilots, maintenance, air crafts and technology improved in these countries.





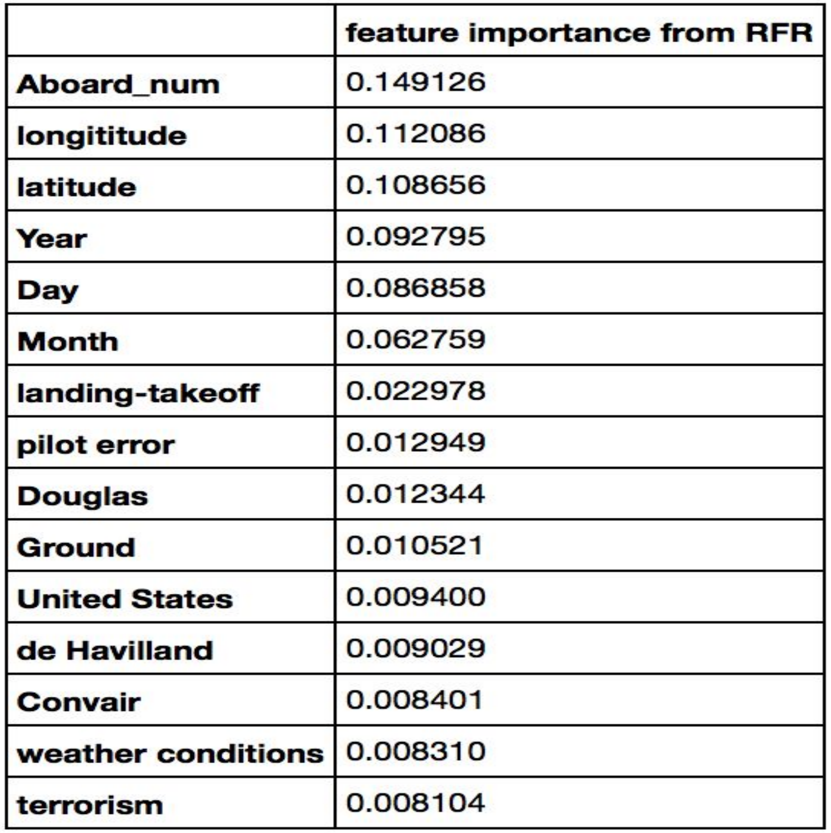
- Last 10 years data is interesting as total number of people died per countries ranges from 0 to under 400.
- Except first 4 countries Ukraine, India, Iran and Nigeria were ranges from 200 per year, which is far less compared to the recent usage of airlines for travel.
- Now lets see what is the trend or reasons for these crashes.

Reasons for crash for last 20 years



- Recent crashes were more due to landing-takeoff, weather conditions, pilot error and engine errors.
- As in last 10 years there were due to take-off and landing issues and hence the number of people survived per year is high.

Important Features



	feature importance from RFR
Aboard_num	0.149126
longititude	0.112086
latitude	0.108656
Year	0.092795
Day	0.086858
Month	0.062759
landing-takeoff	0.022978
pilot error	0.012949
Douglas	0.012344
Ground	0.010521
United States	0.009400
de Havilland	0.009029
Convair	0.008401
weather conditions	0.008310
terrorism	0.008104

- Datasets had features like location of crash, date, number of people on board, type of air crash, summary of crash.
- Categorical datasets were not well input and it was really good process of applying data cleaning, munging and wrangling techniques to extract useful information of features from the available data sets.
- Google Geocoding API were used to extract latitude and longitude of 5000+ places as it wasn’t properly (like 100m from Jersey Airport)
- After the full detailed feature extraction, the table shows the import features for predictions

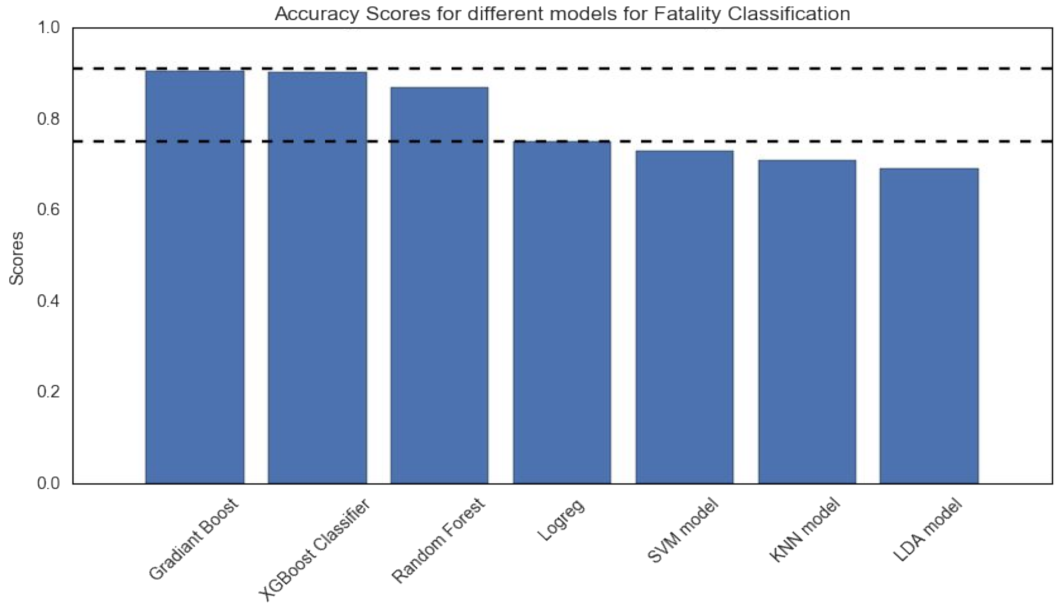
Predictive Modeling

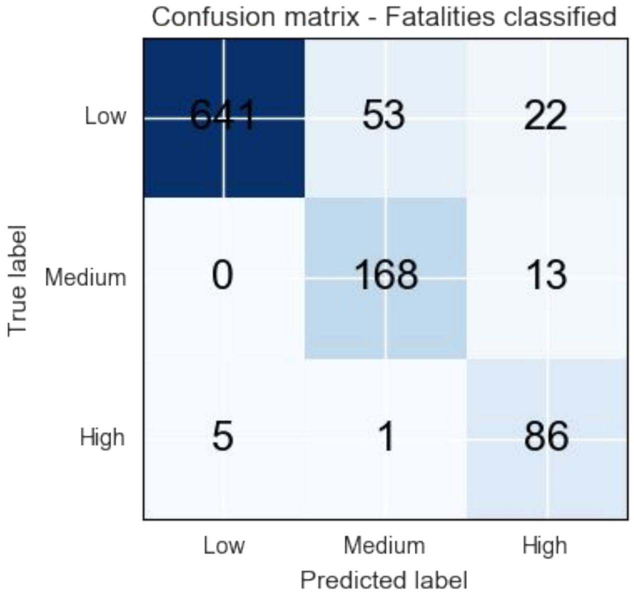
- As predicting the fatality ratio with available data set is very difficult and it hasn't resulted in good results. The best was 12% performance
- So converted into classification problem (Fatality Class)
  - Low Fatality (fatalities per accident is less than mean fatality (<=20)
  - Medium Fatality (fatalities per accident is between 21 and 50)
  - High Fatality (fatalities per accident is higher than 51)
- Also predicting the survival class based on the available data
  - All Survived
  - None Survived
  - Some of them Survived

Fatality Classification Model

- Baseline Accuracy:

Low	0.719895
Medium	0.188828
High	0.091277

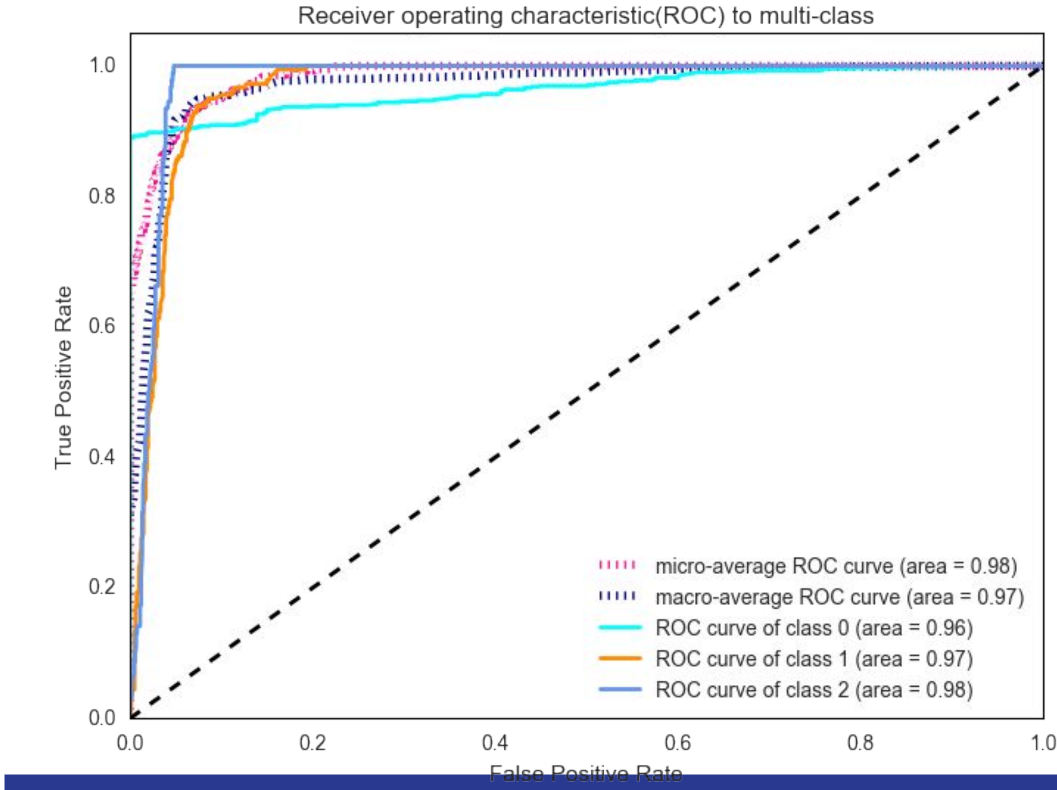


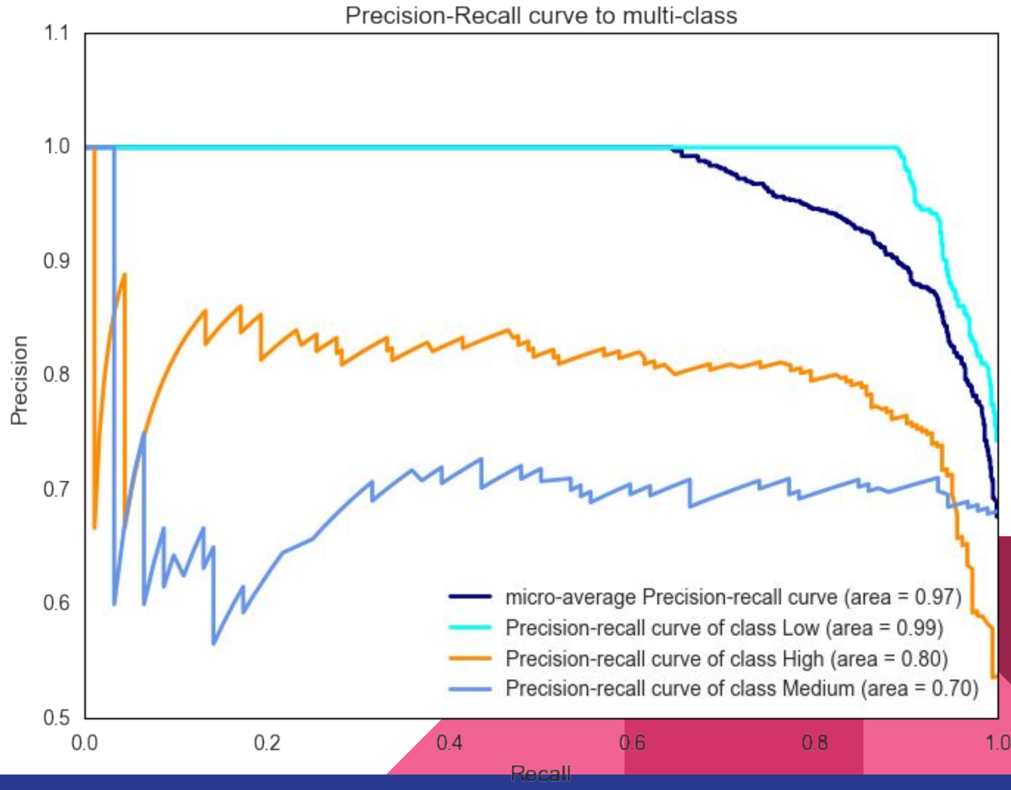

Confusion matrix - Fatalities classified

True label \ Predicted label	Low	Medium	High
Low	641	53	22
Medium	0	168	13
High	5	1	86

Fatality Classification Model ......

- Best Model: Gradient Boost Classifier (0.90 accuracy)

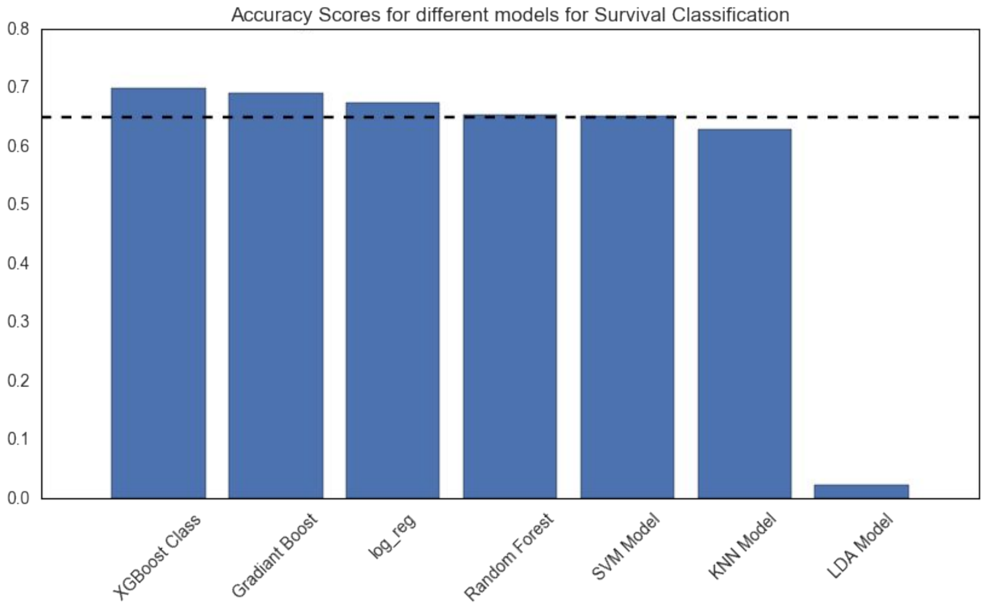


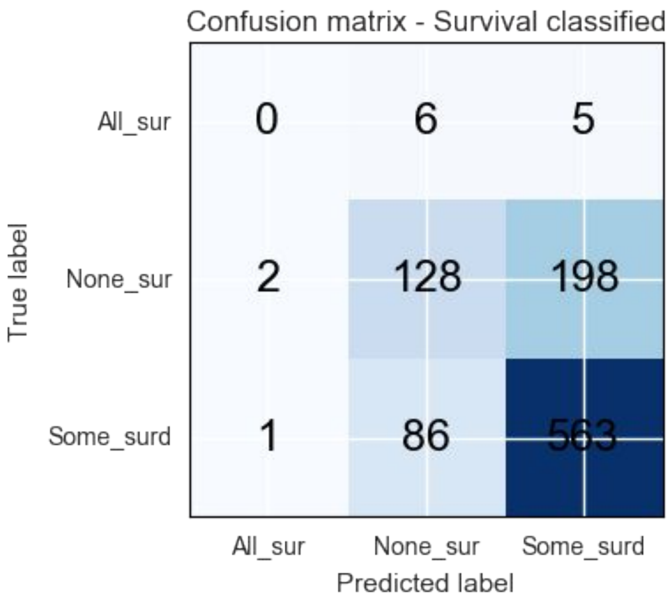


Survival Classification Model

- Baseline Accuracy:

Some_survived	0.652297
None_survived	0.334750
All_survived	0.012953

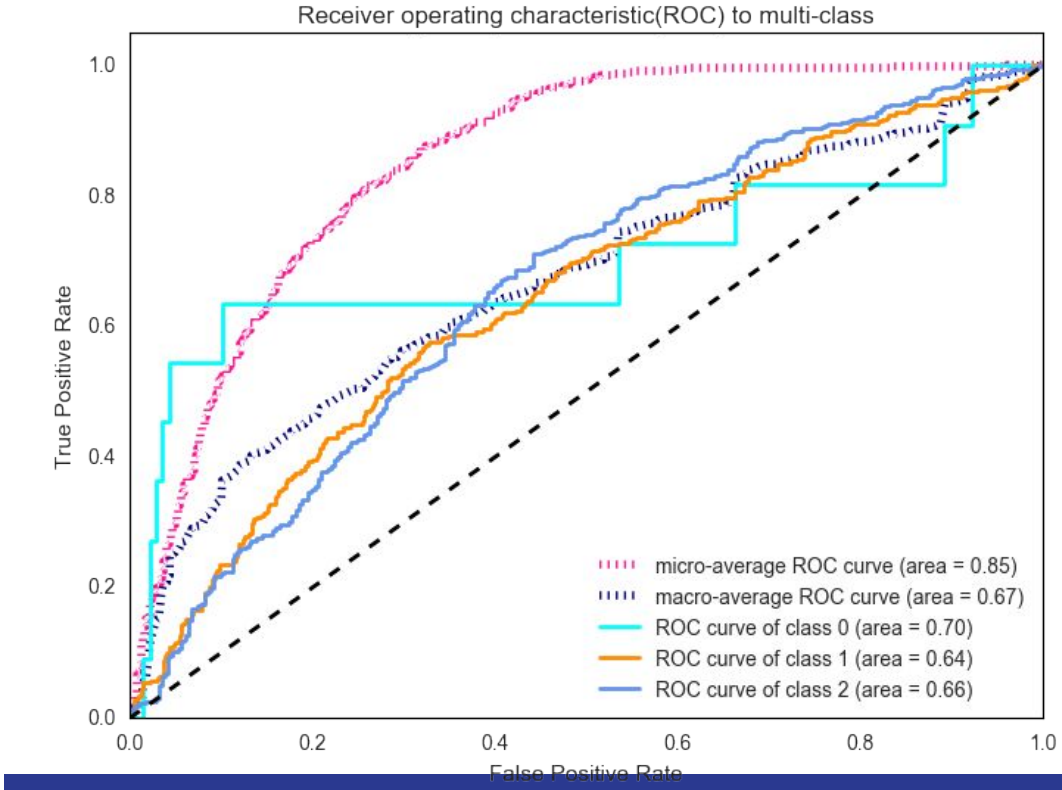


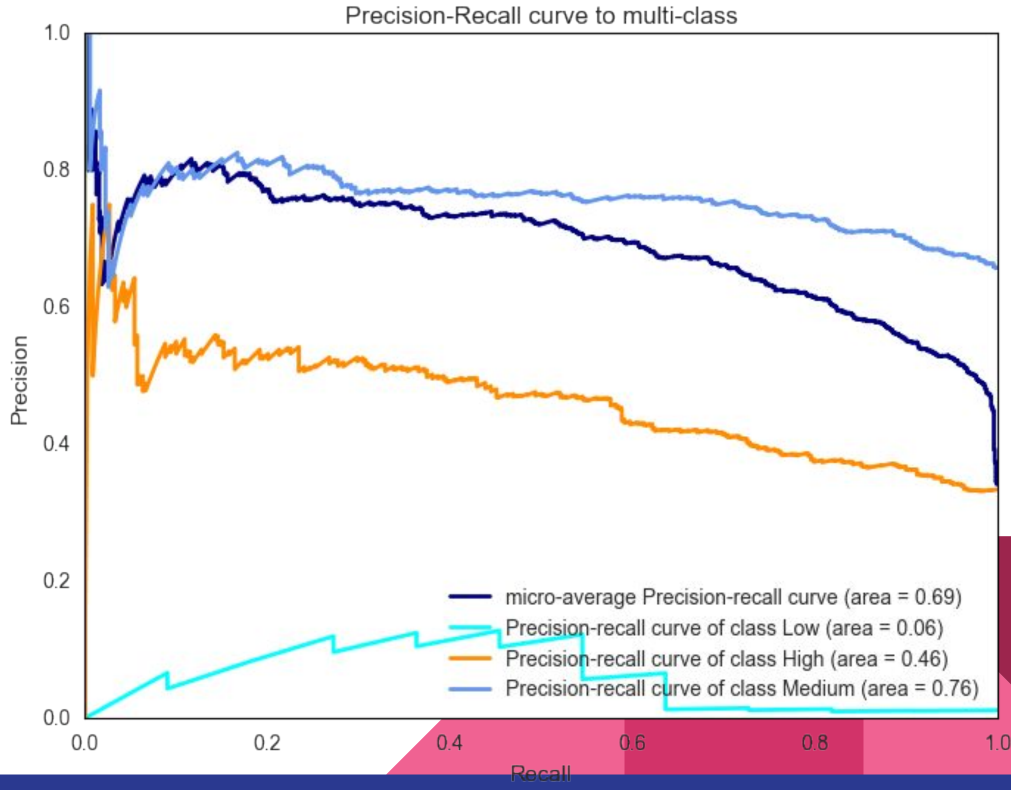

Confusion matrix - Survival classified

True label \ Predicted label	All_sur	None_sur	Some_surd
All_sur	0	6	5
None_sur	2	128	198
Some_surd	1	86	563

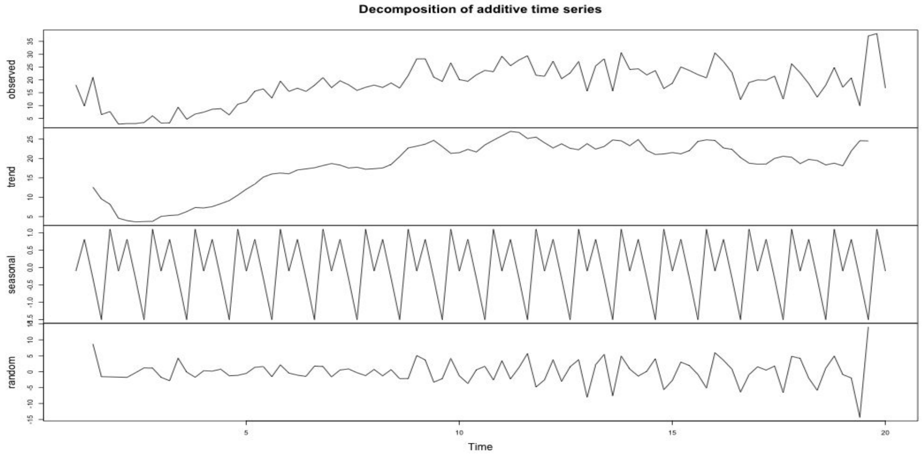
Survival Classification Model ……

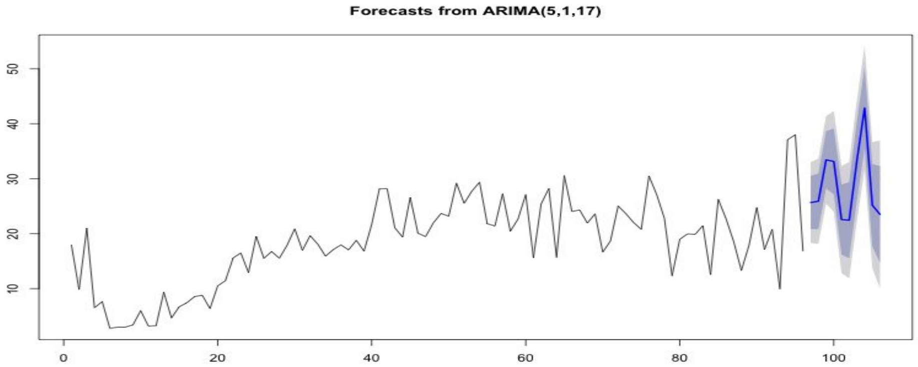
- Best Model: XGBoost Classifier (0.69 accuracy)





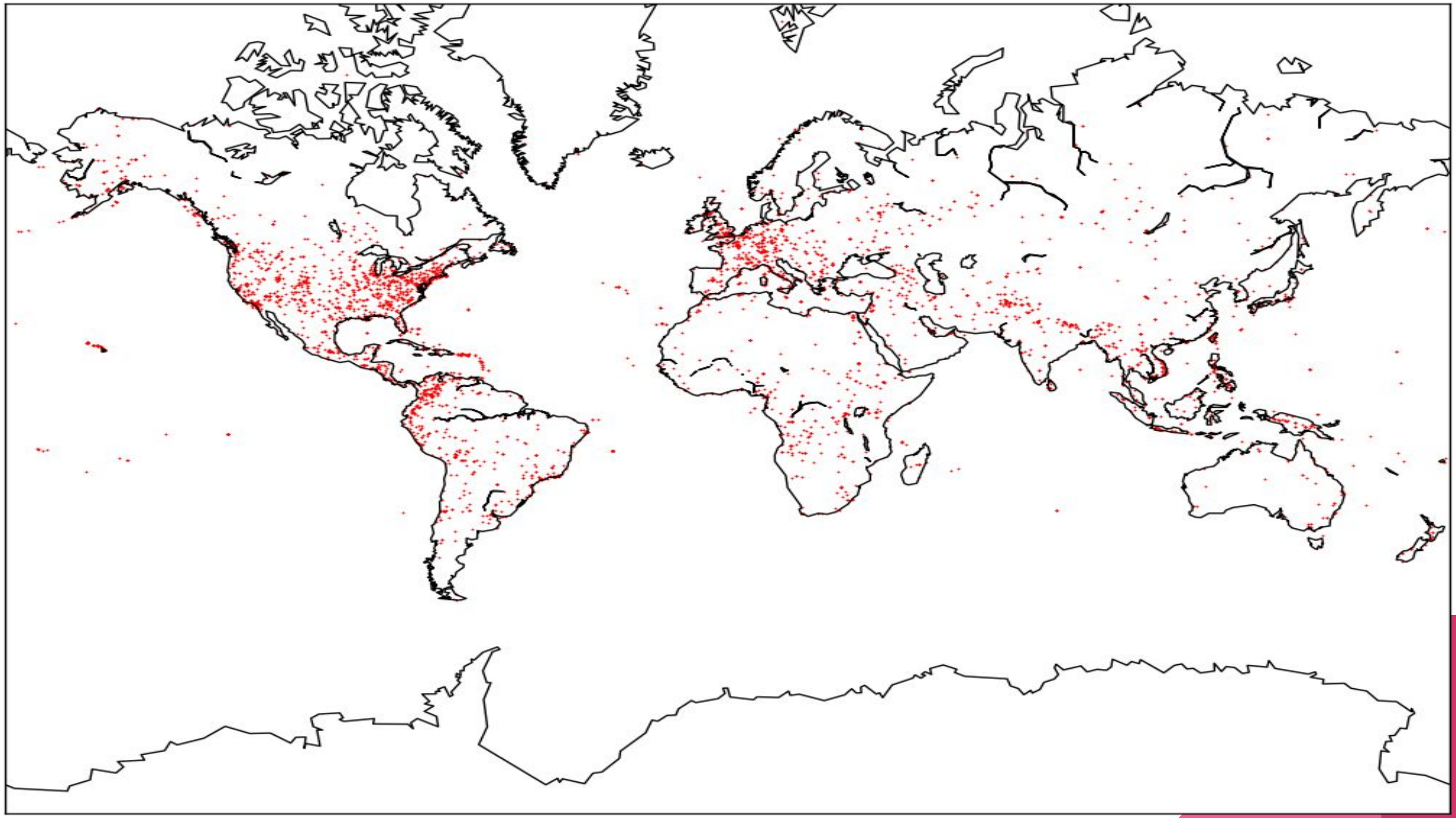
Time series - Fatalities





- As its yearly data and would like to try the time series method for predicting the mean fatalities for next 5 years.
- As there is some trend and not much of seasonality.
- Time series prediction was implemented using R package in python.
- ARIMA model with (5,1,17) has provided much similar result of past.

Crashes across Globe



Crashes across Globe for last 10 years