Whom Should you fly with?

Aviation Accidents Analysis– Text Mining & Time Series

Bidisha Datta
Mengmeng Zhai
Ming Zhu
Sukanya Ghosh
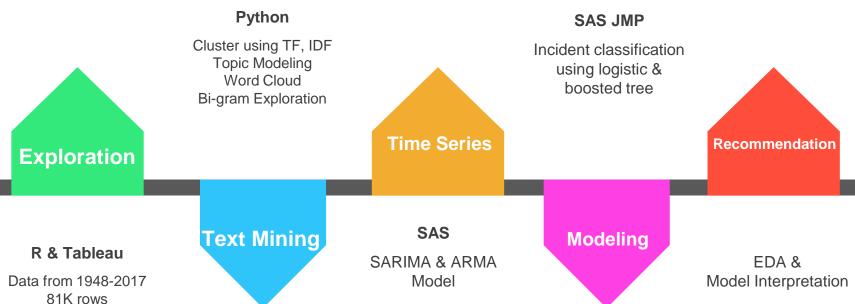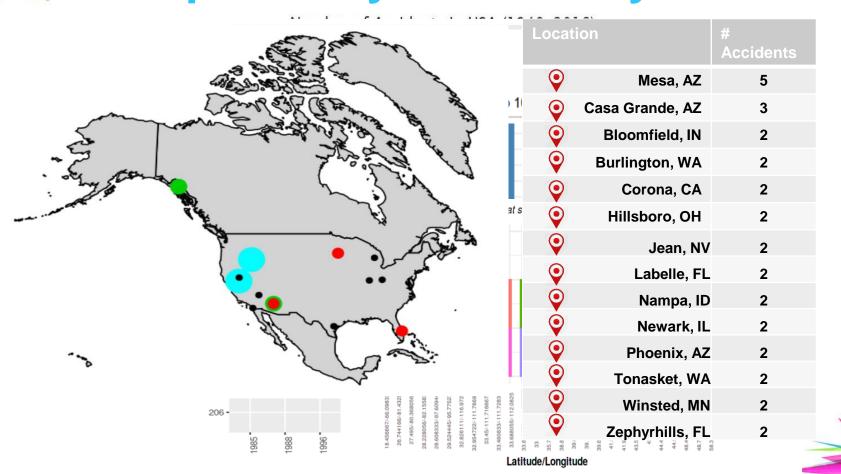
Team 10 | OPIM5671

fppt.com

Our Story

Past Accidents

Which Location?

Failure Reasons

Pilot
Blade Crash
Runway Error
Failure
Fan
Engine

Which Carrier?

SOUTHWEST

UNITED AIRLINES

NTSB: Engine in deadly Southwest jet incident missing a fan blade

By: Staff

Posted: Apr 18, 2018 3:47 AM EST

FACEBOOK   TWITTER   @ EMAIL

Analysis Approach

Python

Cluster using TF, IDF
Topic Modeling
Word Cloud
Bi-gram Exploration

SAS JMP

Incident classification
using logistic &
boosted tree

Exploration

Time Series

Recommendation

R & Tableau

Data from 1948-2017
81K rows
178 Countries
274 Carriers

Text Mining

SAS

SARIMA & ARMA
Model

Modeling

EDA &
Model Interpretation

Exploratory Data Analysis

Location	# Accidents
Mesa, AZ	5
Casa Grande, AZ	3
Bloomfield, IN	2
Burlington, WA	2
Corona, CA	2
Hillsboro, OH	2
Jean, NV	2
Labelle, FL	2
Nampa, ID	2
Newark, IL	2
Phoenix, AZ	2
Tonasket, WA	2
Winsted, MN	2
Zephyrhills, FL	2

Latitude/Longitude

Text Mining Steps

1 K-Means
Text Analysis (plain text, remove punctuation, to lower, removing stop words etc.). Created TFIDF matrix and used caret and tm packages in R.

2 Wordcloud
Used TermDocumentMatrix on the Corpus and used min frequency and max frequency to calculate the word cloud.
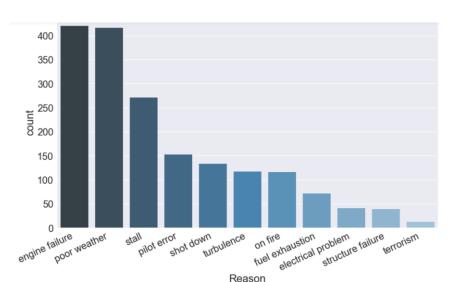
3 Topic Modeling
To discover hidden semantic structure, annotating documents acc. to

topics, use annotations to summarize texts. We have used Latent Dirichlet Allocation (LDA) model.
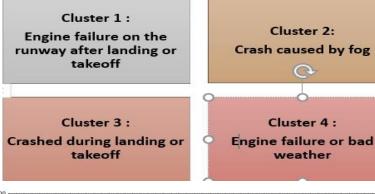
4 Regular Expression
Created a lookup table with keywords pertaining to each category like pilot error, engine failure, bad weather etc. and used RE to match the summary text.
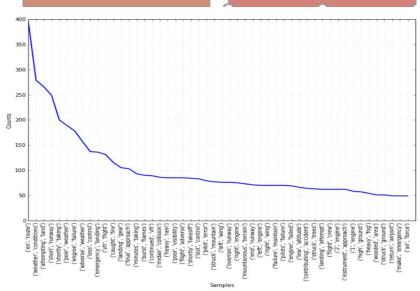
5 Bigrams
we have investigated bigrams to understand the consecutive occurring words.

Text Mining Results

Cluster 1: Engine failure on the runway after landing or takeoff

Cluster 2: Crash caused by fog

Cluster 3: Crashed during landing or takeoff

Cluster 4: Engine failure or bad weather

SARIMA- Sum of Fatality (Include Severe Injury)

Step1: Preliminary Transformation

Step2: Identification

Step3: Estimation of model

Step4: Diagnostic Checking

SARIMA - Sum of Fatality (Includes Severe Injury)

Actual 2017

Time Series Forecast 2017

Prediction Error White Noise/Stationarity Test Probabilities
SUM_OF_INJURED
Log ARIMA(0,1,1)s NOINT

Parameter Estimates
SUM_OF_INJURED
Log ARIMA(0,1,1)s NOINT

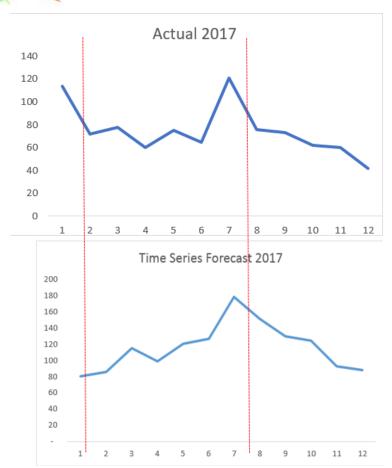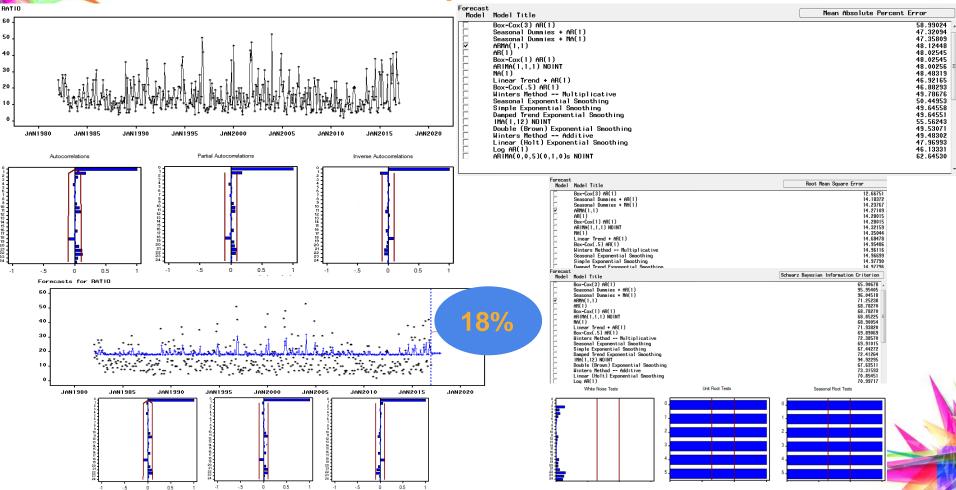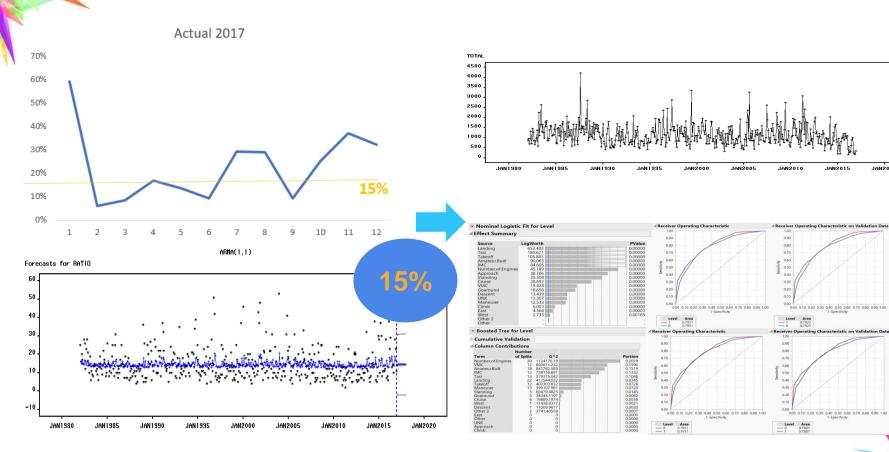| Model Parameter	Estimate	Std. Error	T	Prob>|T|
Seasonal Moving Average, Lag 12	0.88761	0.0342	25.9444	<.0001
Model Variance (sigma squared)	0.20896	.	.	.

- ❖ Pattern captured and 'Log ARIMA(0,1,1)s Model' justified as forecast
- ❖ Analysis shows the summer is the key season to focus
- ❖ Overall, the forecast shows the decline trend , which is business sensible

ARMA- Ratio of Fatality (Includes Severe Injury)

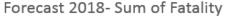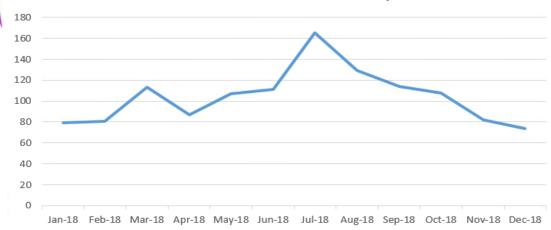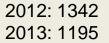| Forecast Model	Model Title	Mean Absolute Percent Error
	Box-Cox(3) AR(1)	58.99024
	Seasonal Dummies + AR(1)	47.32094
	Seasonal Dummies + MA(1)	47.35809
✓	ARMA(1,1)	48.12448
	AR(1)	48.02545
	Box-Cox(1) AR(1)	48.02545
	ARIMA(1,1,1) NOINT	48.00256
	MA(1)	48.48319
	Linear Trend + AR(1)	46.92165
	Box-Cox(.5) AR(1)	46.88293
	Winters Method -- Multiplicative	49.78676
	Seasonal Exponential Smoothing	50.44953
	Simple Exponential Smoothing	49.64558
	Damped Trend Exponential Smoothing	49.64551
	IMA(1,12) NOINT	55.56243
	Double (Brown) Exponential Smoothing	49.53071
	Winters Method -- Additive	49.48302
	Linear (Holt) Exponential Smoothing	47.96993
	Log AR(1)	46.13331
	ARIMA(0,0,5)(0,1,0)s NOINT	62.64530

Forecast Model	Model Title	Root Mean Square Error
	Box-Cox(3) AR(1)	12.66751
	Seasonal Dummies + AR(1)	14.18372
	Seasonal Dummies + MA(1)	14.23767
✓	ARMA(1,1)	14.27109
	AR(1)	14.28015
	Box-Cox(1) AR(1)	14.28015
	ARIMA(1,1,1) NOINT	14.32159
	MA(1)	14.35044
	Linear Trend + AR(1)	14.60478
	Box-Cox(.5) AR(1)	14.95486
	Winters Method -- Multiplicative	14.96116
	Seasonal Exponential Smoothing	14.96699
	Simple Exponential Smoothing	14.97790
	Damped Trend Exponential Smoothing	14.97796

Forecast Model	Model Title	Schwarz Bayesian Information Criterion
	Box-Cox(3) AR(1)	65.90678
	Seasonal Dummies + AR(1)	95.95405
	Seasonal Dummies + MA(1)	96.04518
✓	ARMA(1,1)	71.25238
	AR(1)	68.78270
	Box-Cox(1) AR(1)	68.78270
	ARIMA(1,1,1) NOINT	68.85225
	MA(1)	68.90054
	Linear Trend + AR(1)	71.93820
	Box-Cox(.5) AR(1)	69.89069
	Winters Method -- Multiplicative	72.38570
	Seasonal Exponential Smoothing	69.91015
	Simple Exponential Smoothing	67.42272
	Damped Trend Exponential Smoothing	72.41264
	IMA(1,12) NOINT	94.92295
	Double (Brown) Exponential Smoothing	67.63511
	Winters Method -- Additive	73.31593
	Linear (Holt) Exponential Smoothing	70.85451
	Log AR(1)	70.99717

Autocorrelations | Partial Autocorrelations | Inverse Autocorrelations

Forecasts for RATIO

18%

White Noise Tests | Unit Root Tests | Seasonal Root Tests

Justification Ratio- Business Value

Business Forecast - 2018

Forecast 2018- Sum of Fatality

2012: 1342
2013: 1195
2014: 1759
2015 :1462
2016: 1128
2017 : 898
2018 : 1250

Sum of Fatality

15%

Forecasts for RATIO

RATIO
ARMA(1,1)

Recommendations