Terrorism.

Predicting Terrorist Groups and Events by Thomas Skowronek

Motivation

Why I chose this project.

Python:
Pandas, NumPy, Matplotlib, Seaborn, Folium, scikit-learn, Prophet, Anaconda, Jupyter

Terrorism:
Cowardly act inflicted on the innocent

GTD:
Global Terrorism Database maintained by START

Process Lifecycle:
Data cleansing, EDA, Visualizations, Feature Engineering, Machine Learning

The Data

170,350 observations and 135 attributes describing terrorist attacks across the world between 1970 through 2016

ATTRIBUTE	TYPE	DEFINITION
iyear	Numeric	The year in which the incident occurred
imonth	Numeric	The month in which the incident occurred. When the exact month of the incident is unknown, this will be recorded as "0".
iday	Numeric	The numeric day of the month on which the incident occurred. When the exact day of the incident is unknown, the field is recorded as "0".
extended	Categorical	The duration of an incident extended more than 24 hours. 1 = YES, 0 = NO
country_txt	Categorical	Identifies the country or location where the incident occurred. When incident occurred cannot be identified, it is coded as "Unknown.
region_txt	Categorical	Identifies the region in which the incident occurred, and divided into 1 of 12 categories
latitude	Numeric	The latitude (based on WGS1984 standards) of the city in which the event occurred
longitude	Numeric	The longitude (based on WGS1984 standards) of the city in which the event occurred.
specificity	Categorical	Identifies the geospatial resolution of the latitude and longitude fields. 1 to 5
vicinity	Categorical	1 = YES, The incident occurred in the immediate vicinity of the city in question. 0 = NO, The incident in the city itself.
crit1	Categorical	The violent act must be aimed at attaining a political, economic, religious, or social goal. 1 = YES, 0 = NO
crit2	Categorical	There must be evidence of an intention to coerce, intimidate, or convey some other message to a larger audience than the immediate victims. 1 = YES, 0 = NO
crit3	Categorical	The action is outside the context of legitimate warfare activities, insofar as it targets non-combatants. 1 = YES, 0 = NO
doubtterr	Categorical	There is doubt as to whether the incident is an act of terrorism. 1 = YES, 0 = NO
multiple	Categorical	Denote that the particular attack was part of a "multiple" incident. 1 = YES, 0 = NO
success	Categorical	A successful attack depends on the type of attack. The key question is whether or not the attack type took place. 1 = YES, 0 = NO
suicide	Categorical	Coded "Yes" in those cases where there is evidence that the perpetrator did not intend to escape from the attack alive. 1 = YES, 0 = NO
attacktype1_txt	Categorical	The general method of attack and often reflects the broad class of tactics used. 9 categories
targtype1_txt	Categorical	The general type of target/victim. 22 categories
targsubtype1_txt	Categorical	The more specific target category and provides the next level of designation for each target type. If a target subtype is not applicable this variable is left blank
natlty1_txt	Categorical	The nationality of the target that was attacked. For hijacking incidents, the nationality of the plane is recorded
gname	Text	The name of the group that carried out the attack
guncertain1	Categorical	Indicates whether or not the information reported about the Perpetrator Group Name(s) is based on speculation or dubious claims of responsibility. 1 = YES, 0 = NO
individual	Categorical	Indicates whether or not the attack was carried out by an individual or several individuals not known to be affiliated with a group or organization. 1 = YES, 0 = NO
nperpcap	Numeric	The number of perpetrators taken into custody. "-99" or "Unknown" appears when there is evidence of captured, but the number is not reported
claimed	Categorical	Indicates whether a group or person(s) claimed responsibility for the attack. 1 = YES, 0 = NO
weaptype1_txt	Categorical	Records the general type of weapon used in the incident. Up to four weapon types are recorded for each incident
weapsubtype1_txt	Categorical	A more specific value for most of the Weapon Types identified
nkill	Numeric	Total confirmed fatalities for the incident
nkillus	Numeric	The number of U.S. citizens who died as a result of the incident
nkillter	Numeric	Limited to only perpetrator fatalities
nwound	Numeric	The number of confirmed non-fatal injuries to both perpetrators and victims
nwoundus	Numeric	The number of confirmed non-fatal injuries to U.S. citizens, both perpetrators and victims
nwoundte	Numeric	Number of Perpetrators Injured
property	Categorical	There is evidence of property damage from the incident. 1 = YES, 0 = NO
ishostkid	Categorical	Whether or not the victims were taken hostage or kidnapped during an incident. 1 = YES, 0 = NO
INT_LOG	Categorical	It indicates whether a perpetrator group crossed a border to carry out an attack (logistically international). 1 = YES, 0 = NO, -9=UNKNOWN
INT_IDEO	Categorical	It indicates whether a perpetrator group attacked a target of a different nationality (ideologically international). 1 = YES, 0 = NO, -9=UNKNOWN
INT_MISC	Categorical	It indicates whether a perpetrator group attacked a target of a different nationality (not clear if logistically or ideologically international) 1 = YES, 0 = NO, -9=UNKNOWN
INT_ANY	Categorical	The attack was international on any of the dimensions. 1 = YES, 0 = NO, -9=UNKNOWN

Data Type Conversion

Categorical attributes

```
# List of attributes that are categorical
cat_attrs = ['extended_txt', 'country_txt', 'region_txt', 'specificity', 'vicinity_txt',
             'crit1_txt', 'crit2_txt', 'crit3_txt', 'doubtterr_txt', 'multiple_txt',
             'success_txt', 'suicide_txt', 'attacktype1_txt', 'targtype1_txt',
             'targsubtype1_txt', 'natlty1_txt', 'guncertain1_txt', 'individual_txt',
             'claimed_txt', 'weaptype1_txt', 'weapsubtype1_txt', 'property_txt',
             'ishostkid_txt', 'INT_LOG_txt', 'INT_IDEO_txt','INT_MISC_txt', 'INT_ANY_txt']

for cat in cat_attrs:
    gtd_df[cat] = gtd_df[cat].astype('category')
```

Imputation

Imputation using the mean and medium

Before Imputation

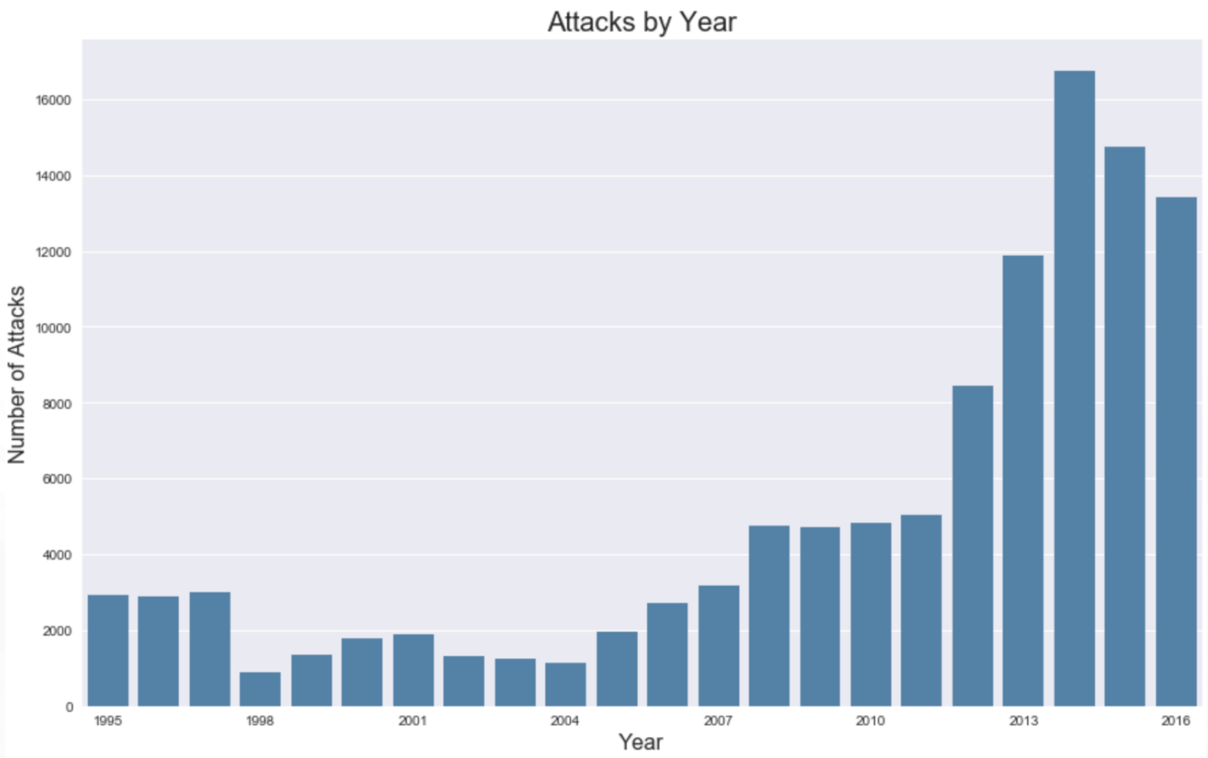
	count	mean	std	min	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	max
nperpcap	90348.0	0.117900	1.910558	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	406.0	406.0
nkill	90348.0	1.928288	6.840954	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	2.0	4.0	670.0	670.0
nkillus	90348.0	0.010404	0.284529	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	44.0	44.0
nkillter	90348.0	0.303803	2.559151	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	500.0	500.0
nwound	90348.0	3.435162	13.542081	0.0	0.0	0.0	0.0	0.0	0.0	1.0	2.0	4.0	8.0	1500.0	1500.0
nwoundus	90348.0	0.013614	0.658389	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	151.0	151.0
nwoundte	90348.0	0.103898	1.522910	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	200.0	200.0

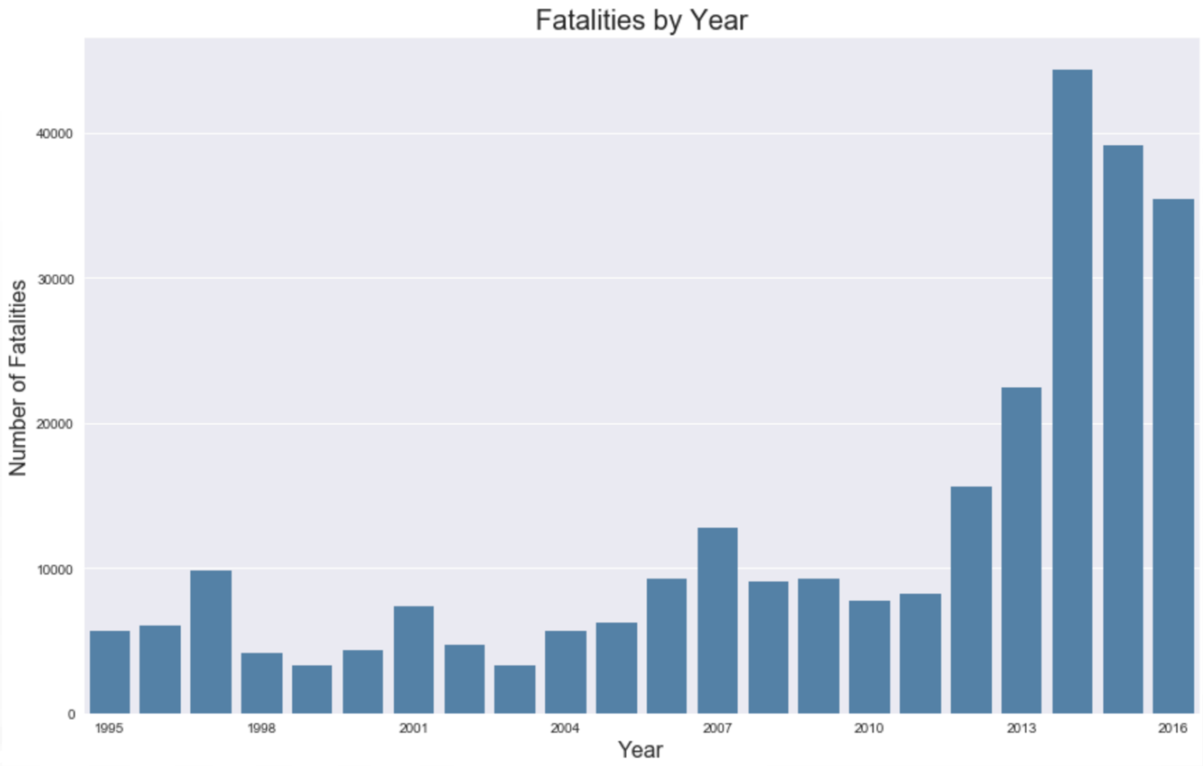
After Imputation

	count	mean	std	min	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	max
nperpcap	112251.0	0.111108	2.055072	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	406.0	406.0
nkill	112251.0	2.485938	11.820567	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	2.0	5.0	1500.0	1500.0
nkillus	112251.0	0.039109	5.723184	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1358.0	1358.0
nkillter	112251.0	0.428121	4.004935	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	500.0	500.0
nwound	112251.0	3.636662	40.069053	0.0	0.0	0.0	0.0	0.0	0.0	1.0	2.0	3.0	8.0	7366.0	7366.0
nwoundus	112251.0	0.023750	2.127698	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	650.0	650.0
nwoundte	112251.0	0.086538	1.378122	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	200.0	200.0

Attacks and Fatalities by Year

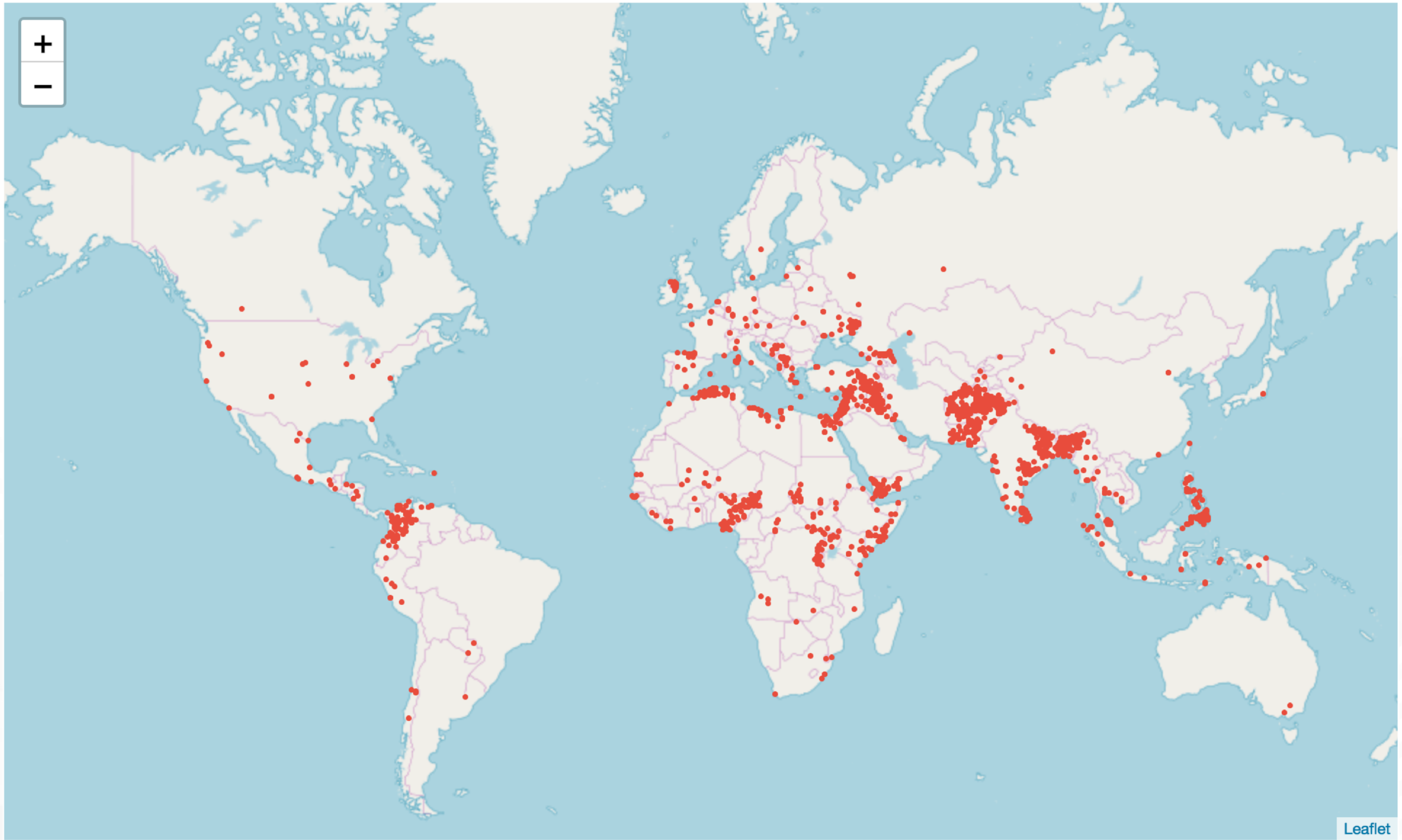
Total terrorist attacks and fatalities by year: 1995 - 2016





Geographical Attack Locations

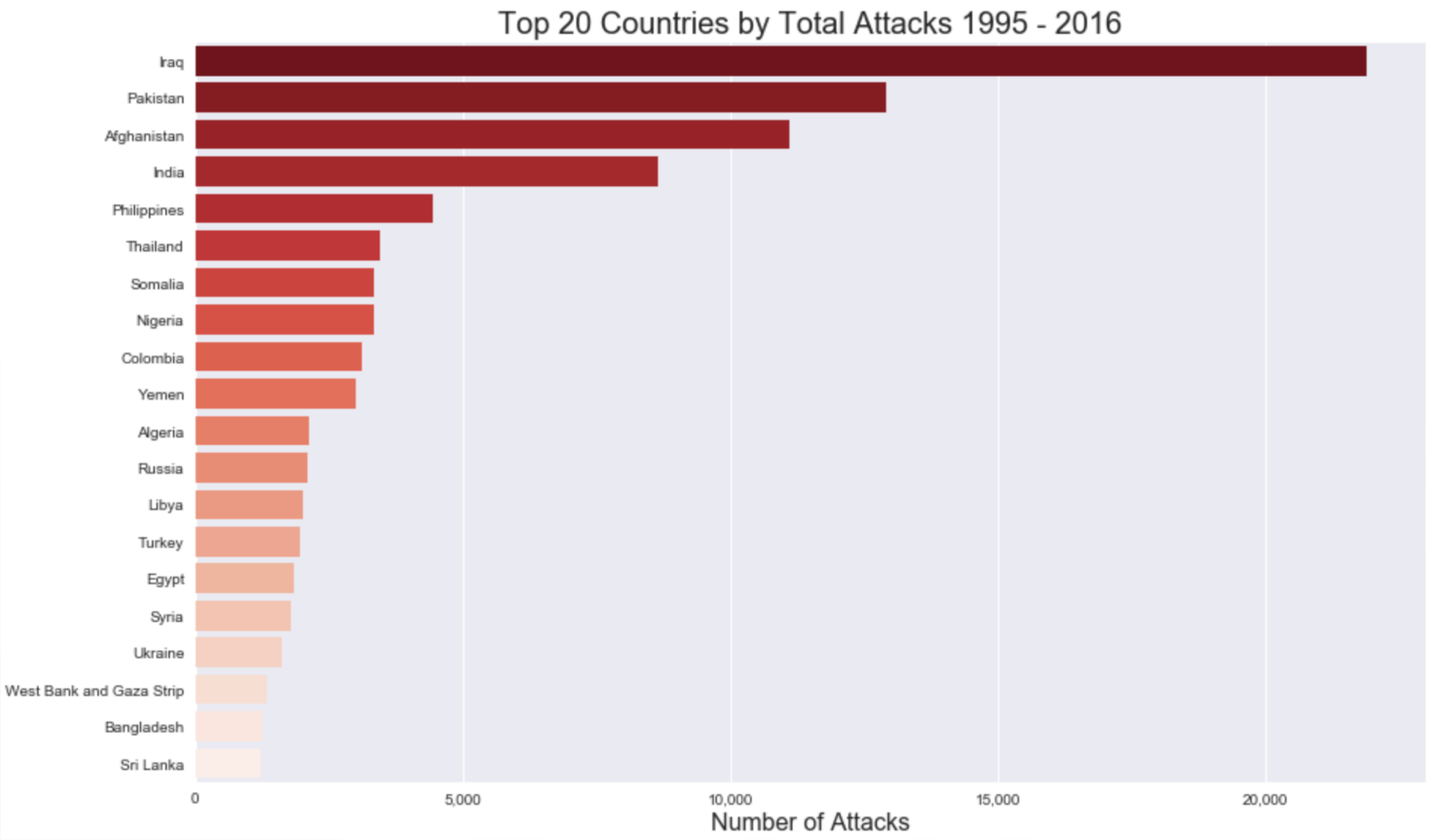
Random sample of latitude/longitude points and region counts



Region	Count
Middle East & North Africa	38723
South Asia	36158
Sub-Saharan Africa	11689
Southeast Asia	9062
Eastern Europe	4618
Western Europe	4382
South America	3892
North America	972
Central America & Caribbean	464
Central Asia	398
East Asia	369
Australasia & Oceania	117

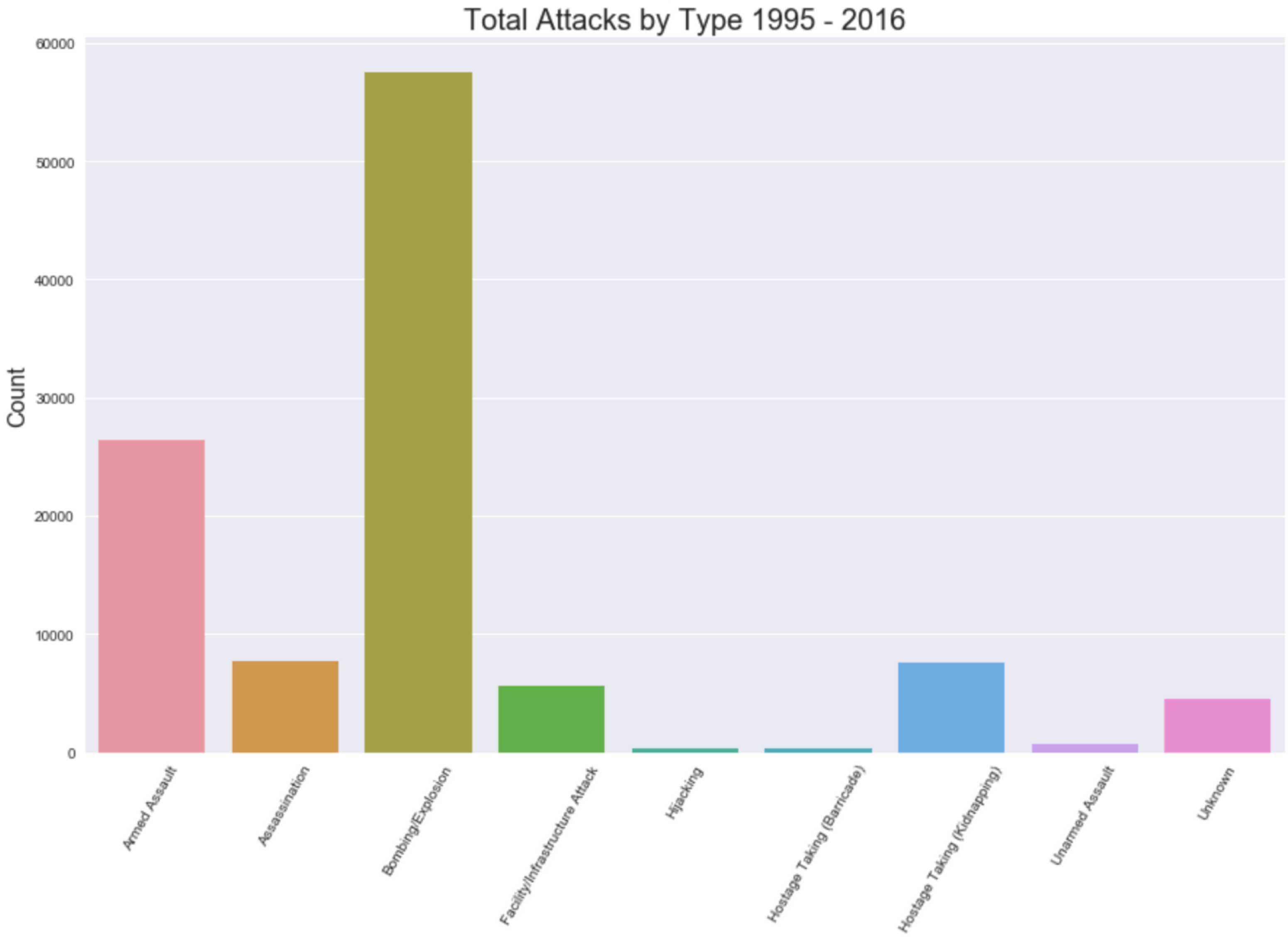
Attacks by Country

The total number of attacks by country spanning 1995 - 2016



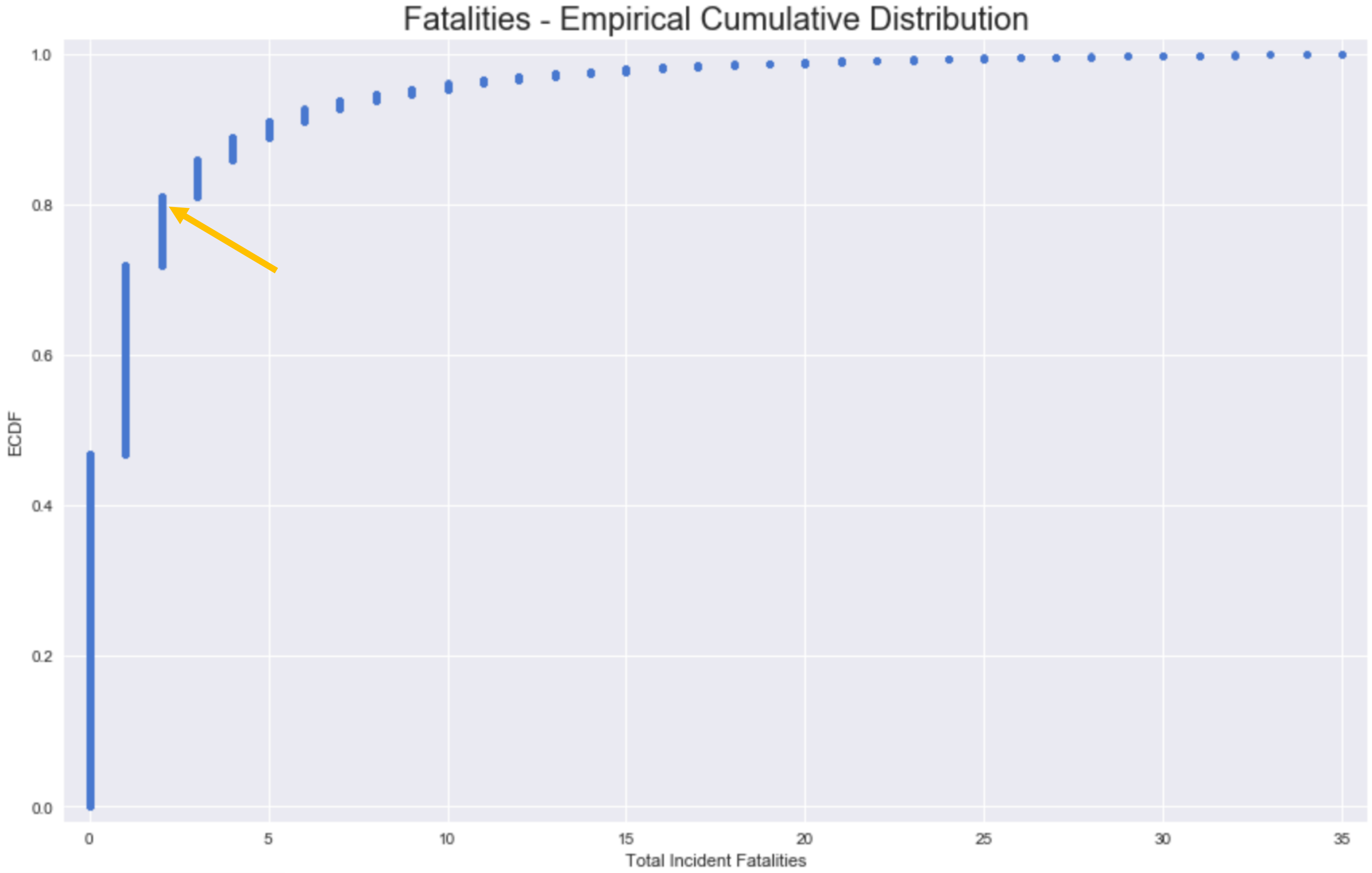
Attacks by Type

The total number of attacks by type spanning 1995 - 2016



Empirical Cumulative Distribution

Approximately 80% of attacks have less than three fatalities



Random Forest

Predict the group who conducted an attack

Major Groups
>= 20 Attacks

Standardization
Scale Numeric Attributes

Prepared Data
One Hot Encoding
+
Label Encoding

Predictive Model
Build, Tune & Evaluate

Predictive Insight
Unknown Attacks

Major Groups

Top 30 groups by number of attacks

Group	Count
Unknown	59531
Taliban	6558
Islamic State of Iraq and the Levant (ISIL)	4261
Al-Shabaab	2669
Boko Haram	2067
Communist Party of India - Maoist (CPI-Maoist)	1766
Revolutionary Armed Forces of Colombia (FARC)	1529
New People's Army (NPA)	1444
Maoists	1411
Kurdistan Workers' Party (PKK)	1255
Tehrik-i-Taliban Pakistan (TTP)	1250
Al-Qaida in the Arabian Peninsula (AQAP)	966
Liberation Tigers of Tamil Eelam (LTTE)	950
Houthi extremists (Ansar Allah)	862
Al-Qaida in Iraq	633
Donetsk People's Republic	613
National Liberation Army of Colombia (ELN)	569
Muslim extremists	534
Abu Sayyaf Group (ASG)	460
Separatists	454
Fulani extremists	431
Palestinian Extremists	384
Basque Fatherland and Freedom (ETA)	369
Algerian Islamic Extremists	366
Moro Islamic Liberation Front (MILF)	336
Hamas (Islamic Resistance Movement)	334
Tripoli Province of the Islamic State	327
Chechen Rebels	326
Sinai Province of the Islamic State	323
Bangsamoro Islamic Freedom Movement (BIFM)	320

Standardization

Robust Scalar

```python
scaler = preprocessing.RobustScaler()

# List of numeric attributes
scale_attrs = ['nperpcap', 'nkill', 'nkillus', 'nkillter', 'nwound', 'nwoundus', 'nwoundte']

# Standardize the attributes in place
major_groups[scale_attrs] = scaler.fit_transform(major_groups[scale_attrs])

# View the transformation
major_groups[scale_attrs].describe().transpose()
```

	count	mean	std	min	25%	50%	75%	max
nperpcap	106144.0	0.093185	1.766333	0.0	0.0	0.0	0.0	406.000000
nkill	106144.0	0.726518	5.881224	-0.5	-0.5	0.0	0.5	749.500000
nkillus	106144.0	0.039559	5.881791	0.0	0.0	0.0	0.0	1358.000000
nkillter	106144.0	0.438310	4.068101	0.0	0.0	0.0	0.0	500.000000
nwound	106144.0	1.181134	12.390895	0.0	0.0	0.0	1.0	2455.333333
nwoundus	106144.0	0.020416	2.125241	0.0	0.0	0.0	0.0	650.000000
nwoundte	106144.0	0.089256	1.410185	0.0	0.0	0.0	0.0	200.000000

Data Encoding

Label Encoding and One Hot Encoding

Target Variable Label Encoding

```
# Create the encoder
le = preprocessing.LabelEncoder()

# Fit the encoder to the target
le.fit(known_maj_groups['gname'])
```

Predictor Variables One Hot Encoding

```
# Seed for reproducible results
seed = 1009

# Predictor variables
X = pd.get_dummies(known_maj_groups.drop(['gname'], axis=1), drop_first=True)

# Labels
y = label_codes

# Create an 80/20 split for training and testing data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = seed, stratify = y)
```

Random Forest Model

Hyperparameters: 1,000 trees and min_samples_leaf = 2

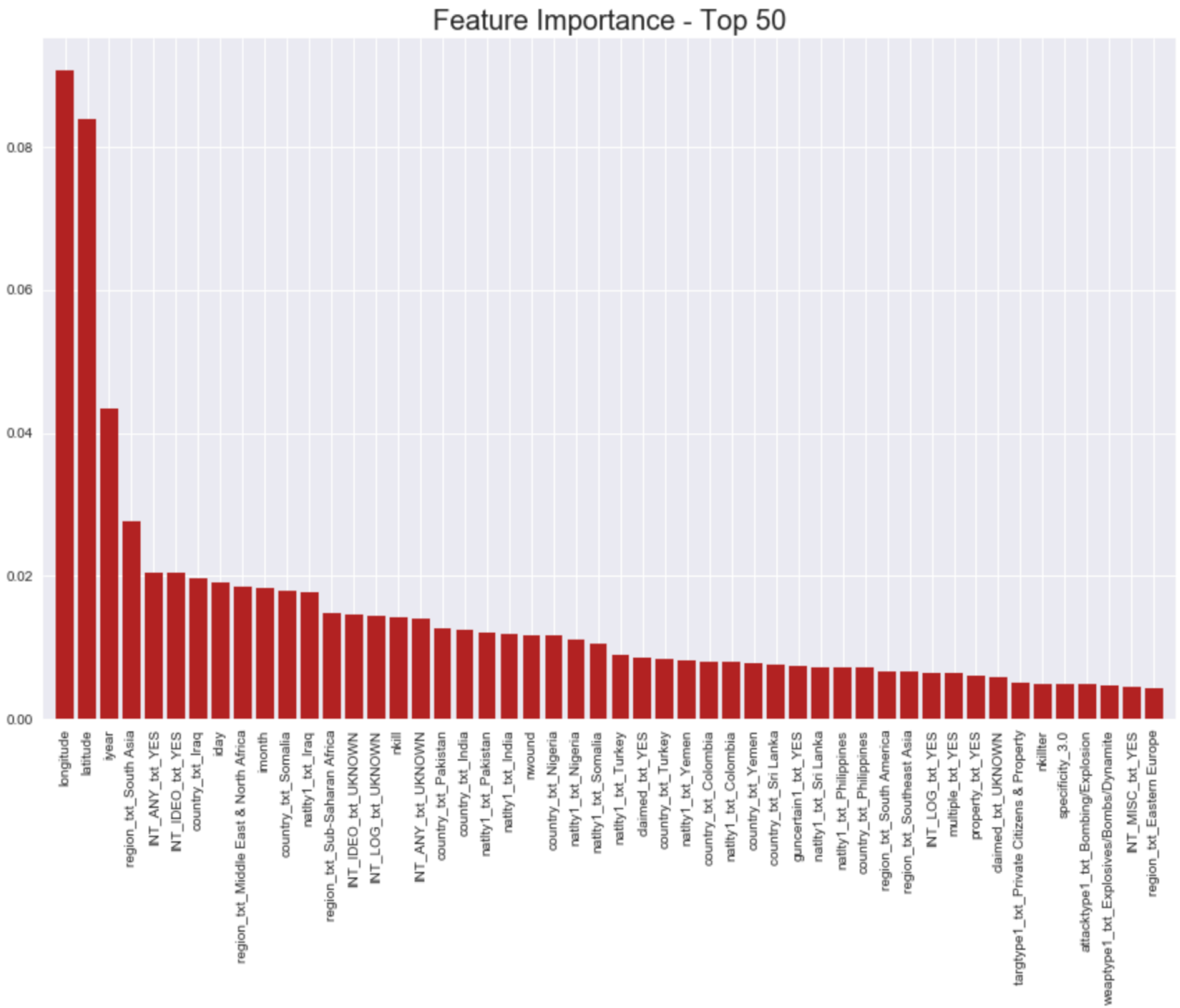
```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
            max_depth=None, max_features='auto', max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, n_estimators=1000, n_jobs=-1,
            oob_score=True, random_state=1009, verbose=0, warm_start=False)
```

Performance Measurements

- Accuracy: 0.9045
- Precision: 0.9026
- Recall: 0.9045
- F1: 0.8951

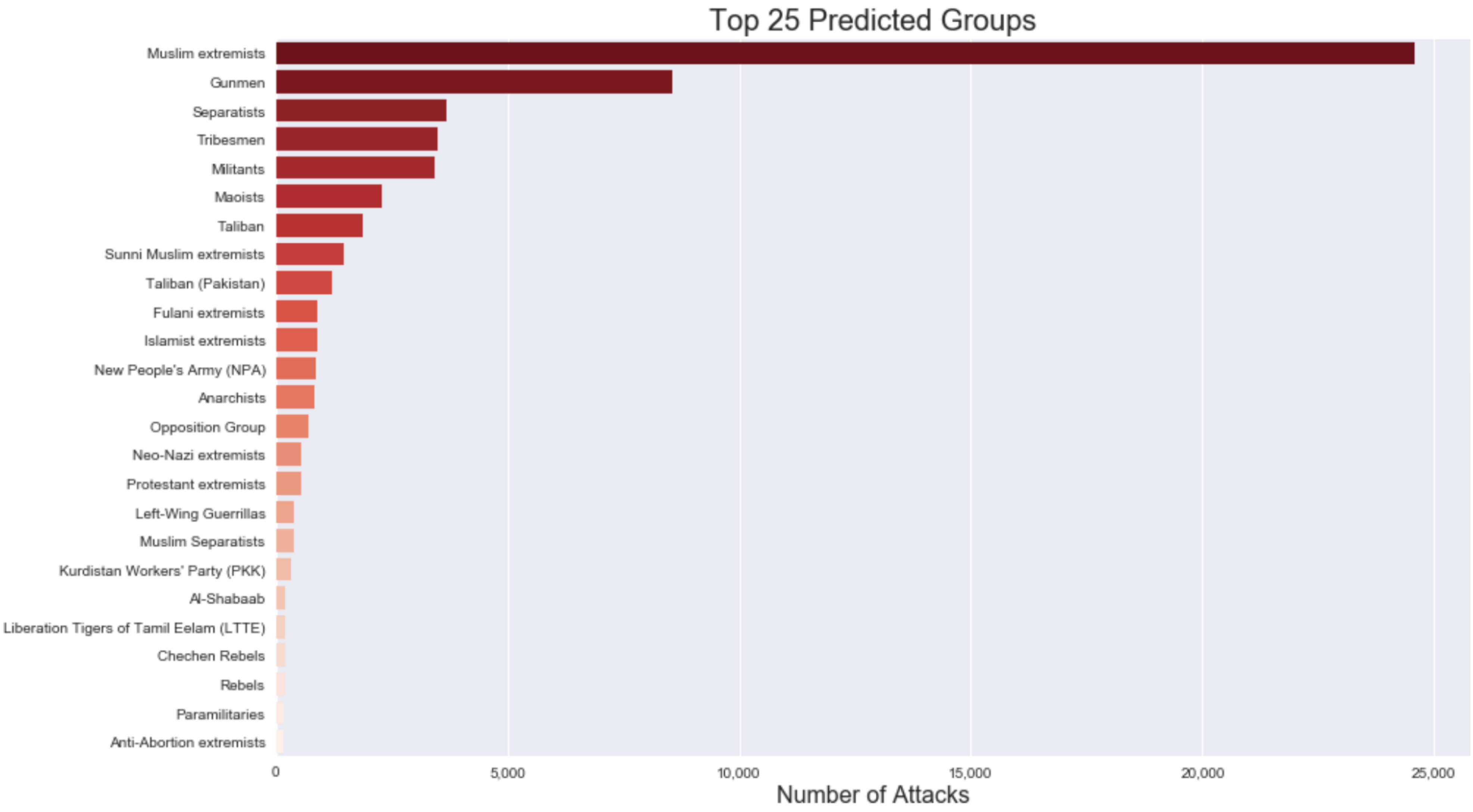
Feature Importance

Top 50 predictor variables



Predictive Insight

Unknown attacks



Iraq

Time series analysis and prediction.

Population

37,200,000

Land Area

438,317 km²

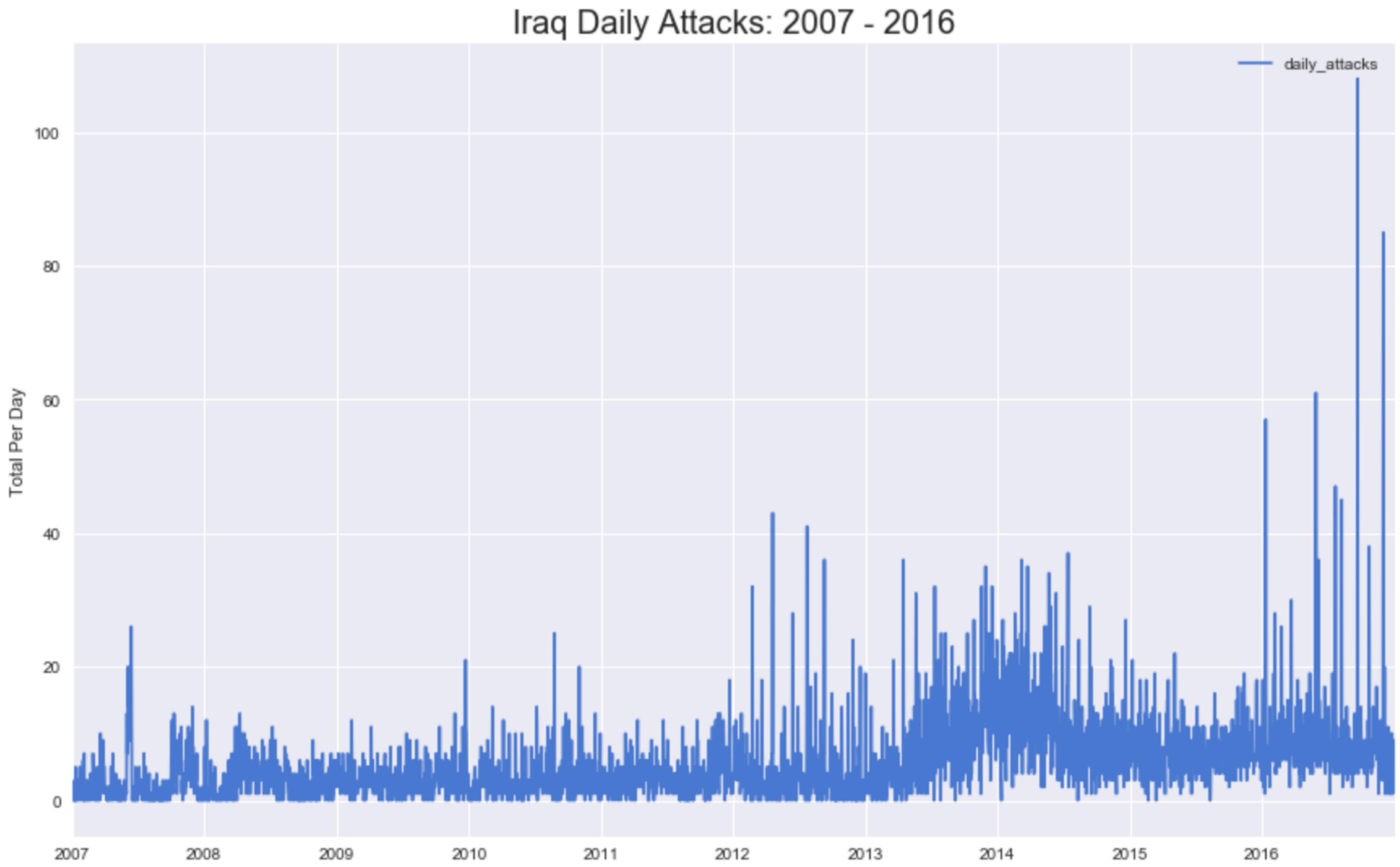
2007 – 2016

Attacks: 20,024

Fatalities: 60,257

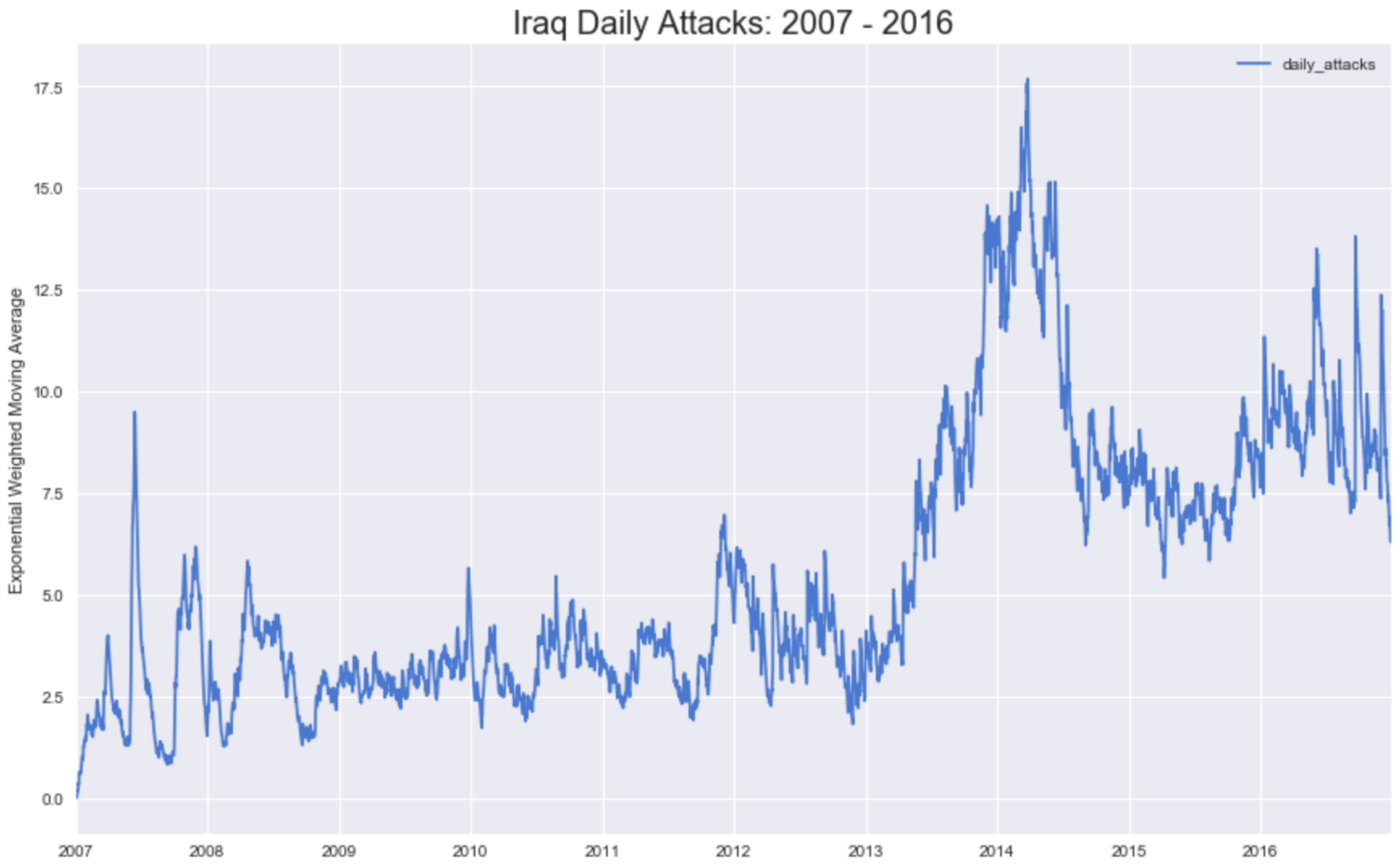
Iraq Daily Attacks

Last 10 years



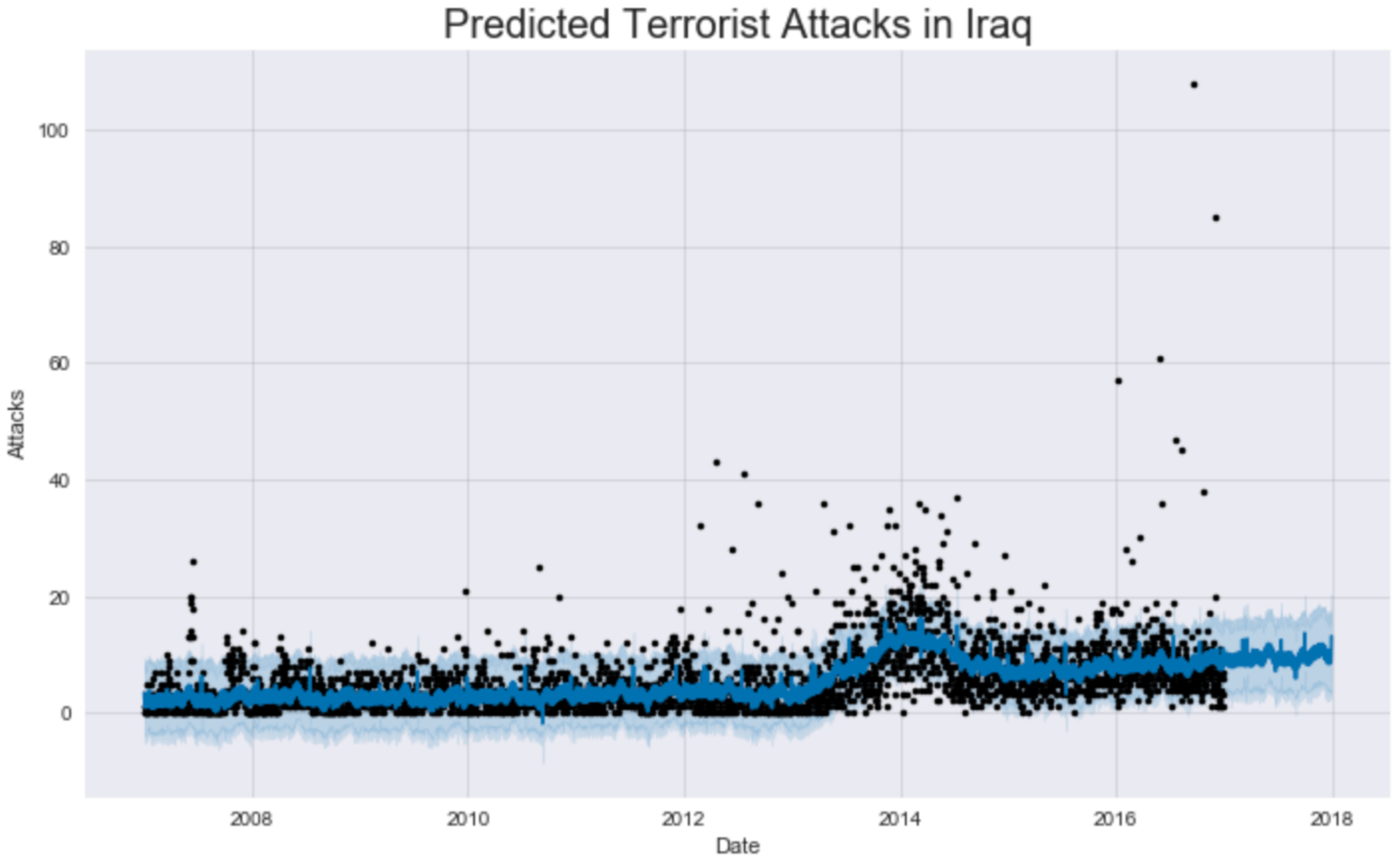
Exponential Weighted Moving Average

30 day window



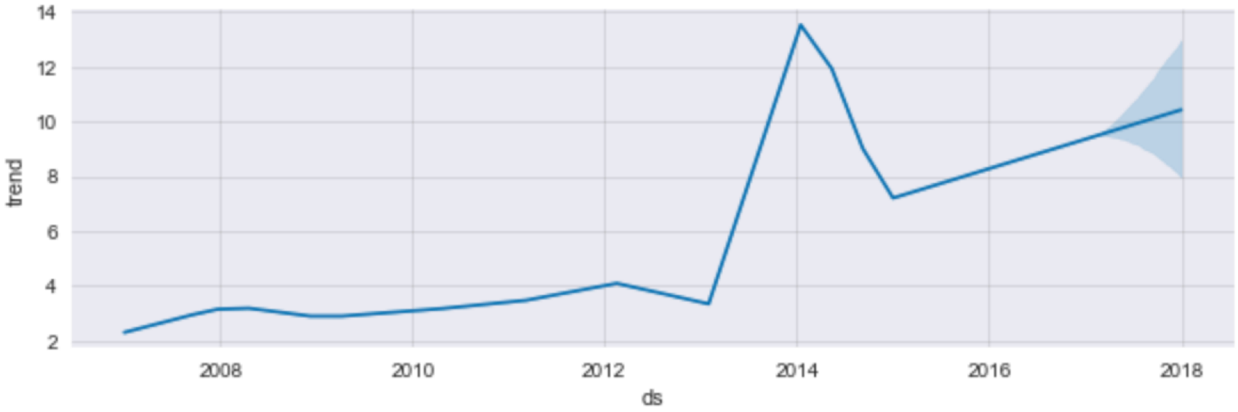
Facebook Prophet

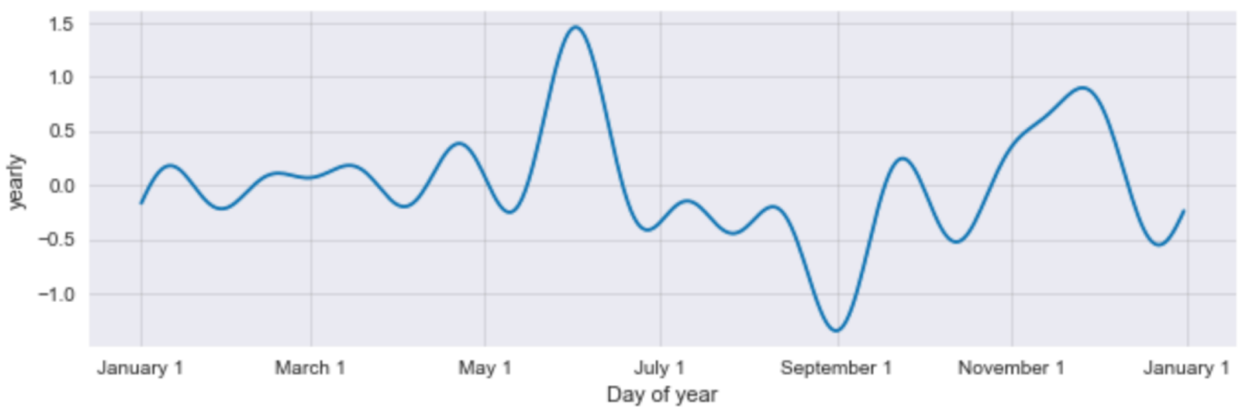
Predicting future attacks

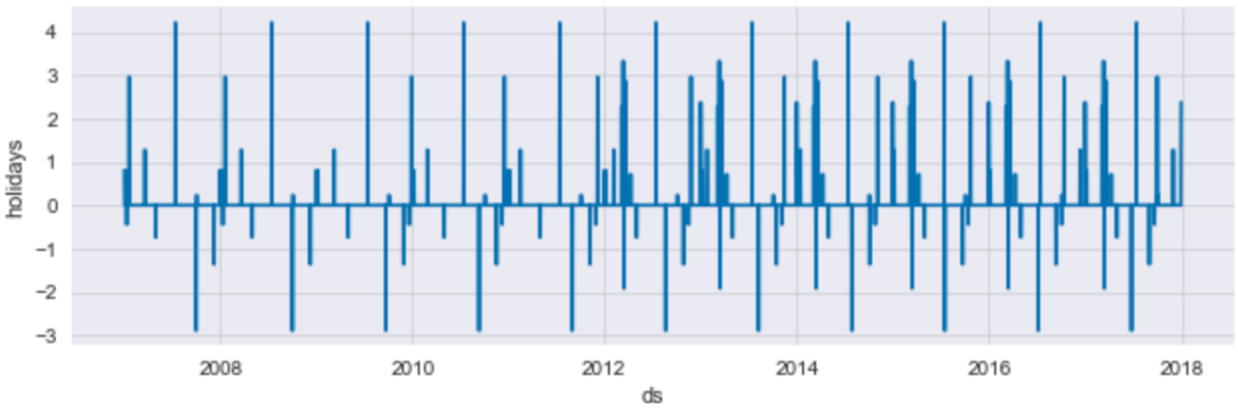


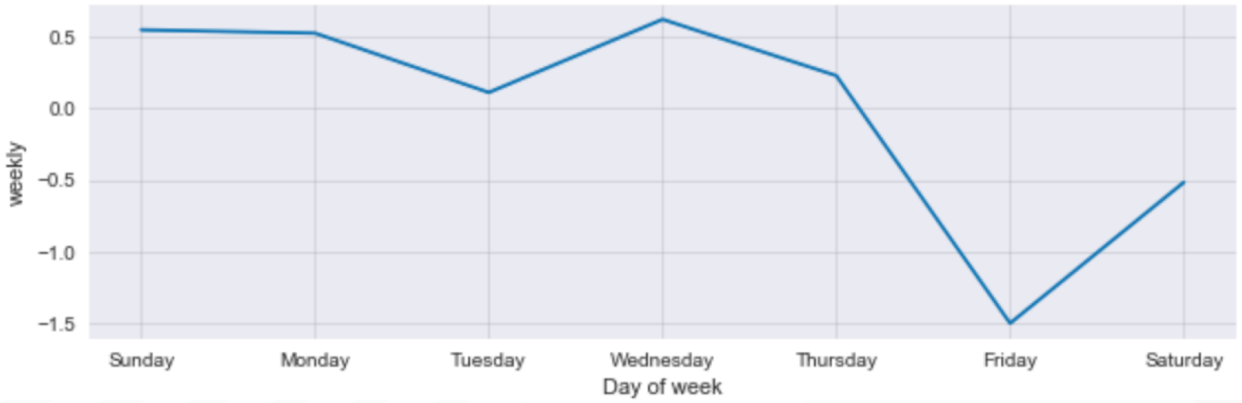
Attack Forecast

Seasonality and Trend Components



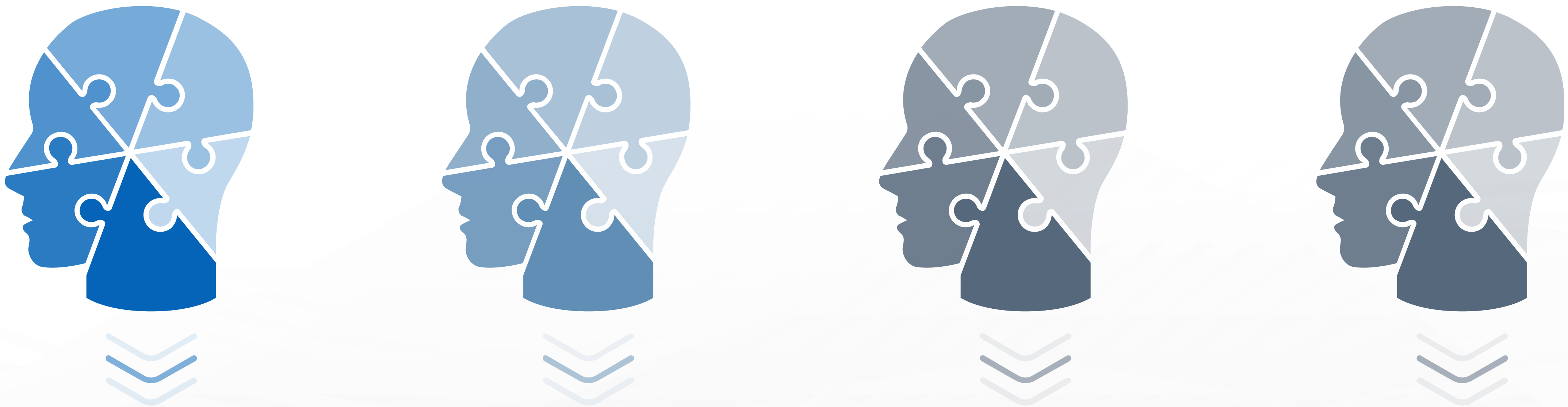






Conclusion

Project Takeaways.



Learning
Python

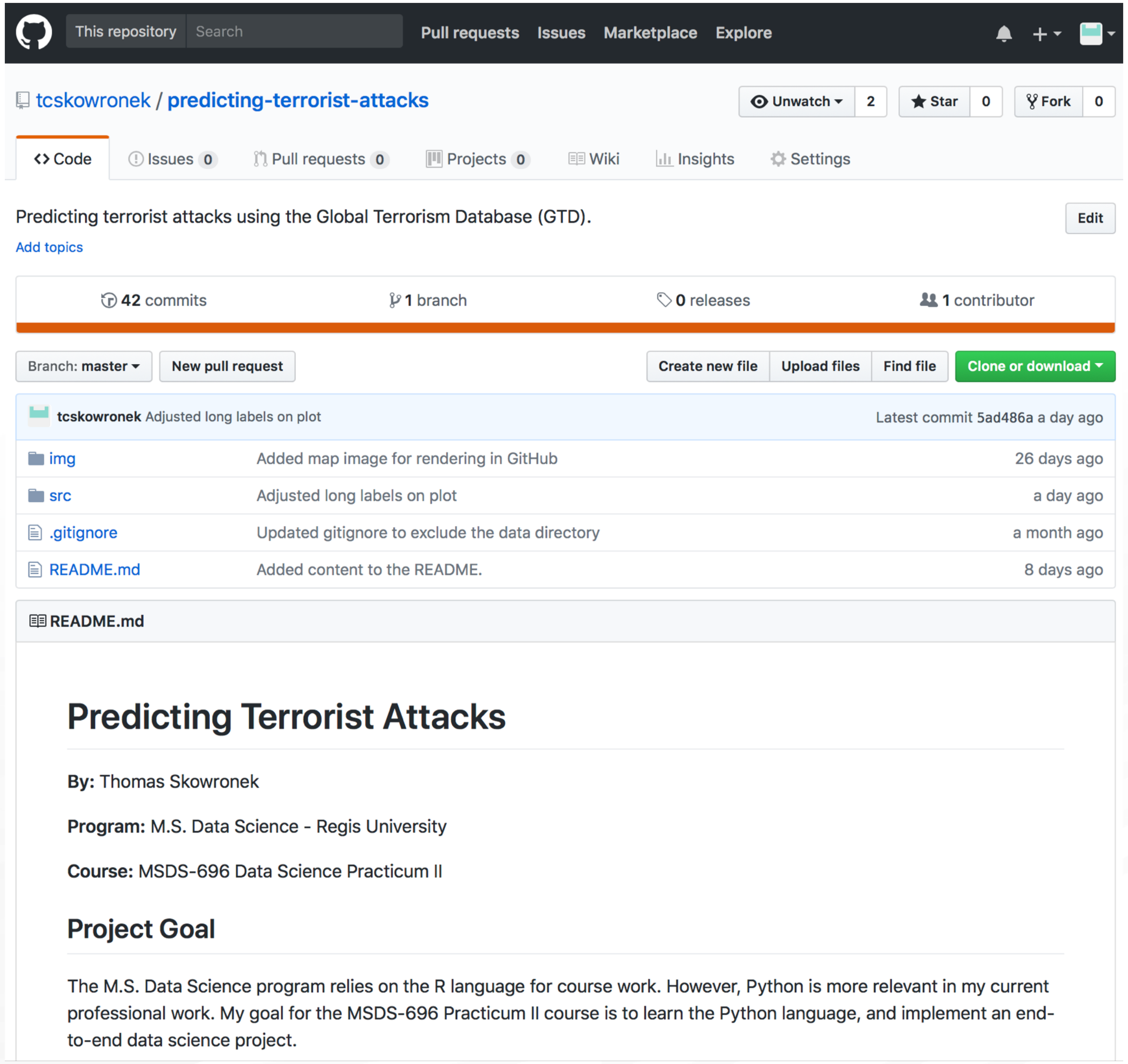
Data Cleansing
& EDA

Time Series
Models

Data
Augmentation

Project Repository

GitHub



GitHub Repository: https://github.com/tcskowronek/predicting-terrorist-attacks

References

Albon, C. (2017). Convert Pandas categorical data for scikit-learn Retrieved from https://chrisalbon.com/machine_learning/preprocessing_structured_data/convert_pandas_categorical_column_into_integers_for_scikit-learn/

Bois, J. (n.d.). Plotting all of your data: Empirical cumulative distribution functions Retrieved from https://campus.datacamp.com/courses/statistical-thinking-in-python-part-1/graphical-exploratory-data-analysis?ex=11

Brown, E. (2017). Forecasting time-series data with Prophet – Part 1. Retrieved from http://pythondata.com/forecasting-time-series-data-with-prophet-part-1/

Chen, D. (n.d.). Duplicate and missing data Retrieved from https://campus.datacamp.com/courses/cleaning-data-in-python/cleaning-data-for-analysis?ex=11

Facebook Open Source. (n.d.). Seasonality and holiday effects. Retrieved from http://holidayyear.com/holidays/Iraq/

Holidayyear. (2018). Iraqi public holidays. Retrieved from http://holidayyear.com/holidays/Iraq/

Keen, B. (2017). Feature scaling with scikit-learn Retrieved from http://benalexkeen.com/feature-scaling-with-scikit-learn/

Letham, B. (2018). Meaning of y-axis in component plot. Retrieved from https://github.com/facebook/prophet/issues/422

MapFight. (n.d.). Iraq California Comparison. Retrieved from https://mapfight.appspot.com/iq-vs-california/iraq-california-us-size-comparison

National Consortium for the Study of Terrorism and Responses to Terrorism. (2017). Global Terrorism Database Codebook: Inclusion criteria and variables. College Park, MD: University of Maryland.

Pandas. (n.d.). Categorical data. Retrieved from https://pandas.pydata.org/pandas-docs/stable/categorical.html

Saabas, A. (2014). Selecting good features – Part III: random forests Retrieved from http://blog.datadive.net/selecting-good-features-part-iii-random-forests/

scikit-learn. (n.d.). Feature importances with forests of trees Retrieved from http://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html#sphx-glr-auto-examples-ensemble-plot-forest-importances-py

The World Bank. (n.d.). Iraq. Retrieved from https://data.worldbank.org/country/iraq

Thank You For Watching.