Terrorism is Happening: What Are the Consequences?

Jaeyeon Won

Project Description

- **Goal**
  - To find the model that best predicts the casualties of a terrorist attack
- **Importance**
  - Design effective strategies to deal with the consequences of terrorist incidents ahead of time

Project Description

- **Machine Learning System**
  - Supervised multiple regression machine learning algorithm
- **Performance Measures**
  - Root Mean Squared Error (RMSE)
  - Mean Absolute Error (MAE)
  - R-squared
  - Standard Deviation

Project Description

- **Assumptions**
  - Complex relationship between the label and features (than a linear relationship)
  - Independence of features
  - Independence of errors with a normal distribution with a mean of 0.

Description of Data

- **Source of Data**
  - From the Global Terrorism Database (GTD)
  - 135 attributes over 190,000 terrorism cases between 1970 and 2018

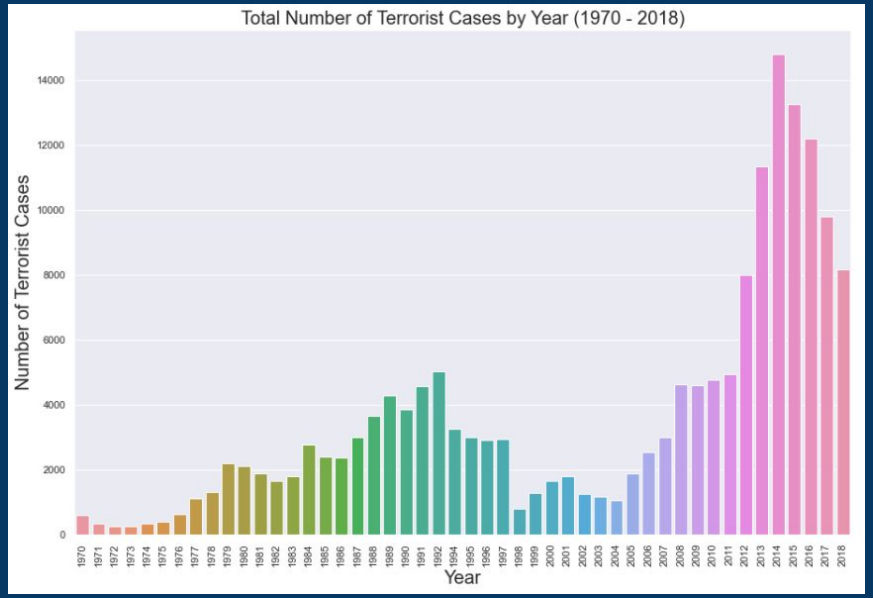
Description of Data

- **Data Cleaning**
  - Label: *nkill* + *nwould* = *casualty*
  - Initial feature selection: 12 attributes that are deemed to be related to casualties (includes only categorical attributes)
  - Rename the attributes & levels of categorical attributes
  - Drop missing values

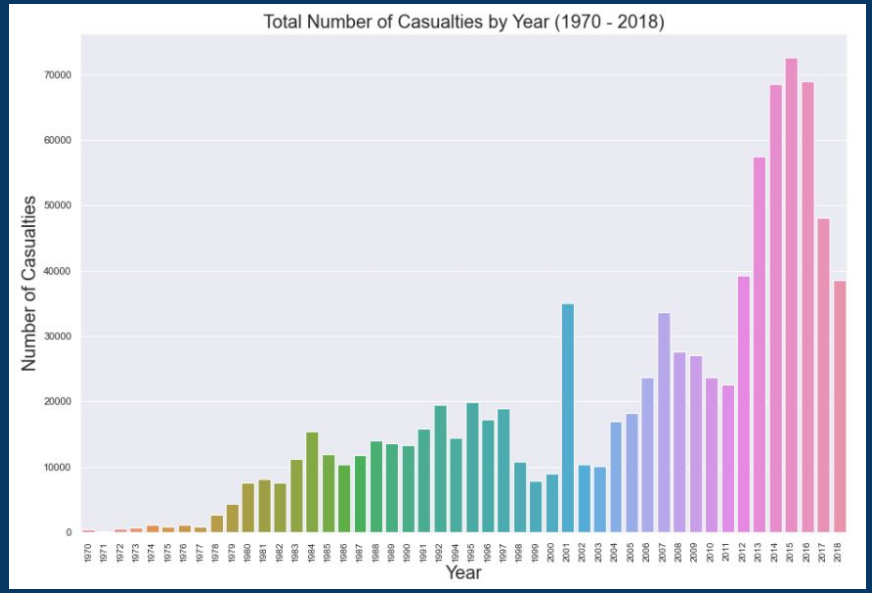
Exploration of Data

- **General Terrorist Activity Trends**

Total Number of Terrorist Cases



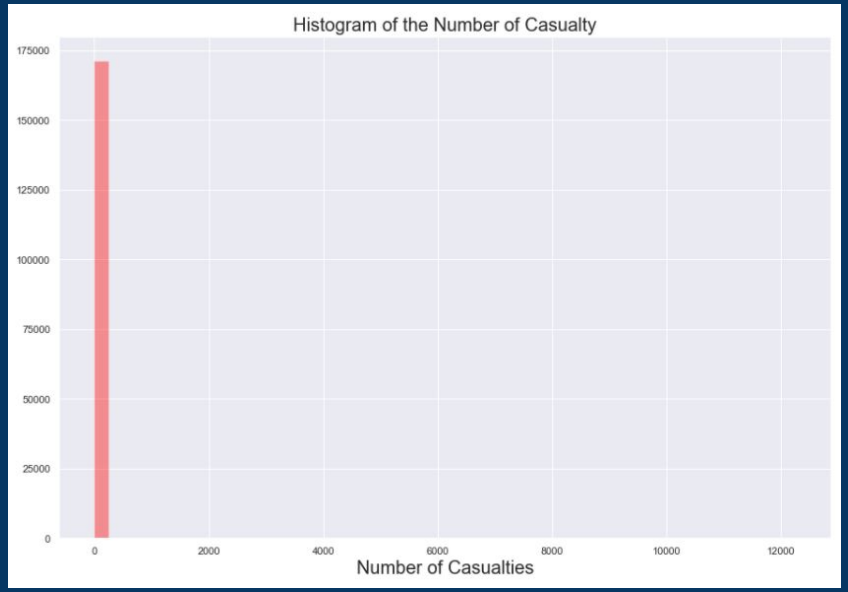
Total Number of Casualties



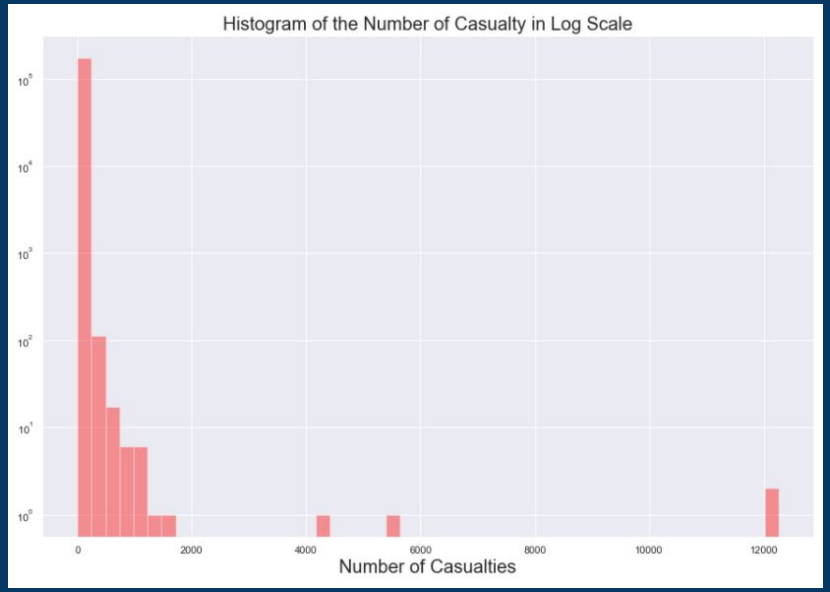
Exploration of Data

- **Histogram of Casualty**

Original Histogram

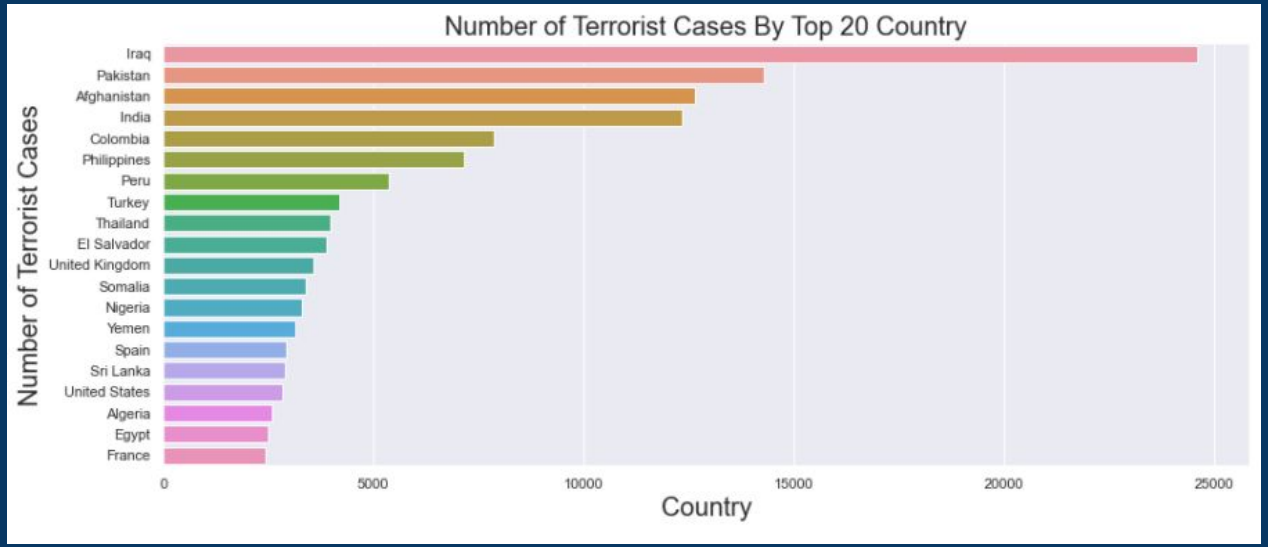


Log Scale on Y-axis



Exploration of Data

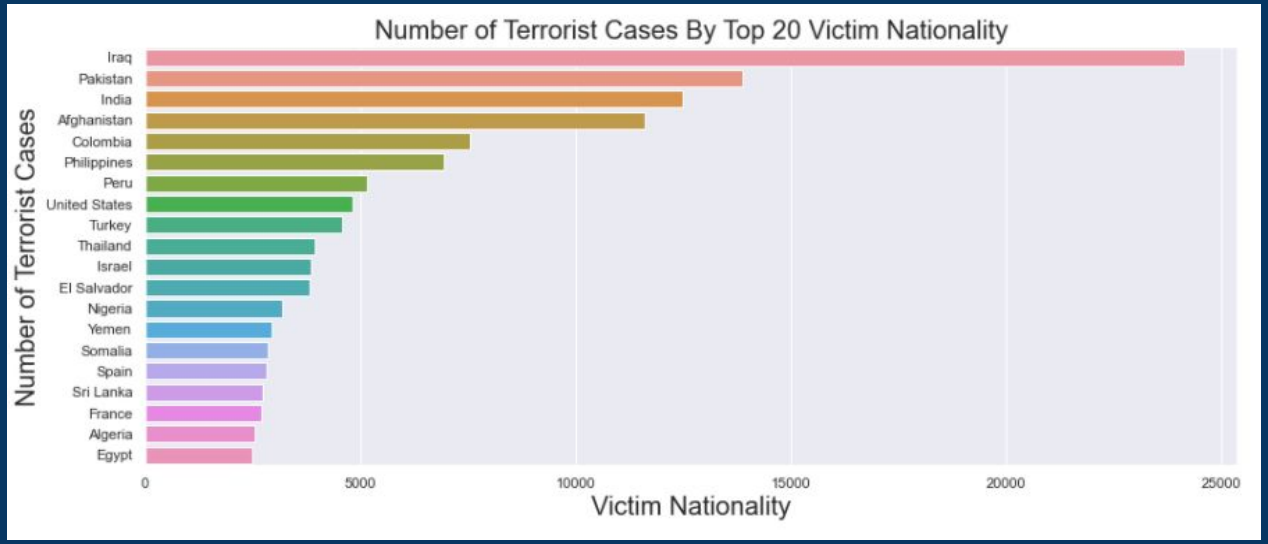
- Categorical Attributes with Multiple Categories



- Significant decrease after top 7 countries

Exploration of Data

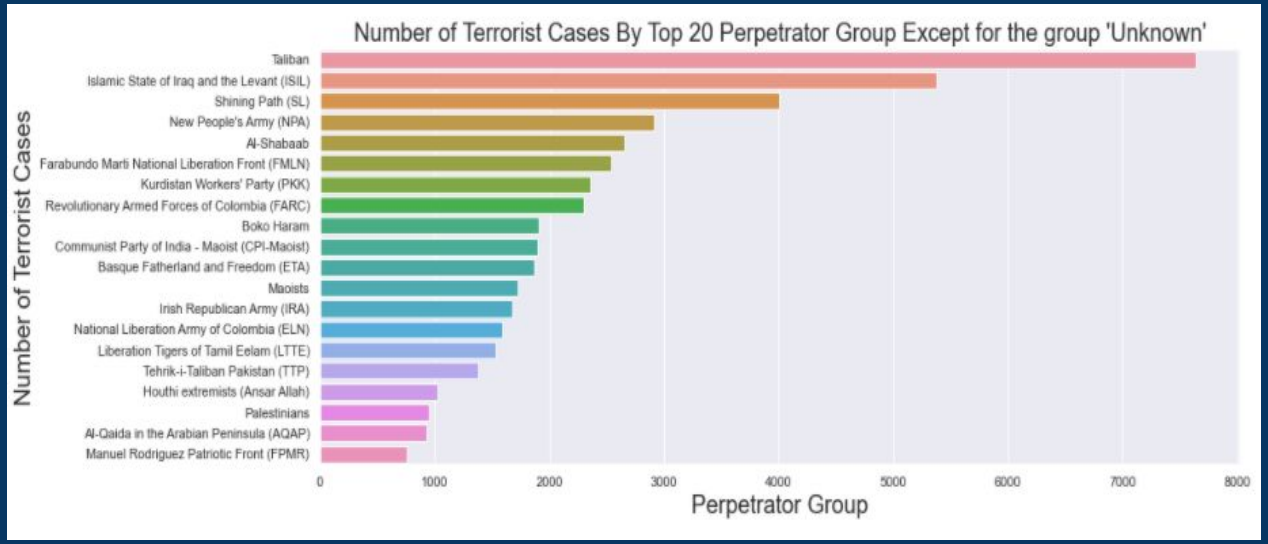
- Categorical Attributes with Multiple Categories



- Significant decrease after top 6 nationalities

Exploration of Data

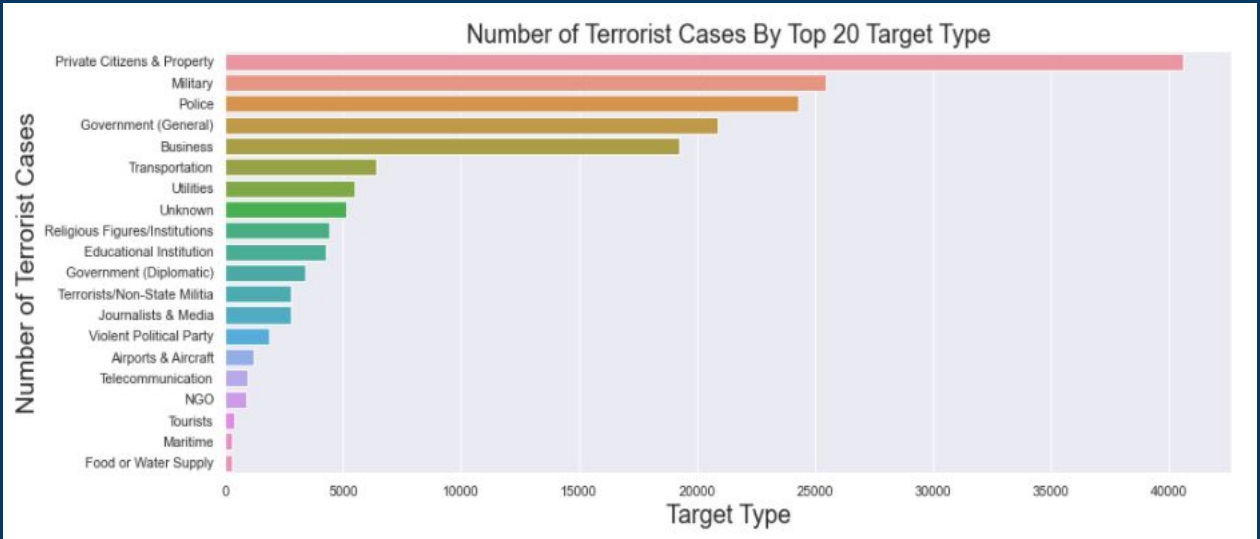
- Categorical Attributes with Multiple Categories



- Significant decrease after top 3 groups

Exploration of Data

- Categorical Attributes with Multiple Categories



- Significant decrease after top 5 target types

Exploration of Data

- **Correlation**
  - Levels with less frequencies grouped into ‘Other’ category based on the previous explorations
  - Highest correlation was around 0.1
  - Variables with a correlation greater than 0.1: *suicide, weaponType*

Preparation of Data

- **Label**
  - Original *casualty* attribute with no transformation
- **Features**
  - *suicide* and *weaponType*
  - One-hot encoding
- **Train Set and Test Set**
  - 80% training set, 20% testing set, random state of 42

Exploration of Different Models

- Cross-Validation
  - 3 folds

	Polynomial Regression	Ridge Regression	Lasso Regression	Elastic Net	SVM Regression	Decision Tree Regressor
RMSE	53.4026	53.4013	53.5647	53.6604	53.8653	53.7124
MAE	6.3488	6.3488	6.3385	6.3386	4.9374	6.3619
R-Squared	0.0181	0.0181	0.0121	0.0086	0.0010	0.0067
Standard Deviation	8.4779	8.4287	5.8548	3.1146	2.3281	36.5305

Fine-Tuning the Model

- Hyperparameter Tuning & Ensemble

	Hyperparameter Tuning				Ensemble	
	Polynomial Regression with Degree 3	Polynomial Regression with Degree 5	Ridge Regression with Alpha 50	Ridge Regression with Alpha 100	Polynomial Regression Adaboost	Ridge Regression Adaboost
RMSE	53.4026	53.4026	53.3946	53.4174	85.0294	100.5837
MAE	6.3488	6.3488	6.3343	6.3267	25.1447	32.3648
R-Squared	0.0181	0.0181	0.0184	0.0176	-1.4893	-2.4834
Standard Deviation	8.4779	8.4779	7.0379	6.4328	81.6044	99.0716

Evaluation of Final System on the Test Set

- Ridge Regression with alpha = 50

RMSE	MAE	R-Squared	Standard Deviation
18.4634	6.1581	0.0046	7.2722

Presenting the Solution

- **What worked & What did not**
  - **Exploration of Different Models**
  - **Worked**: Polynomial Regression, Ridge Regression
  - **Not Worked**: Lasso Regression, Elastic Net, SVM Regression, Decision Tree Regressor
  - **Fine-Tuning the Models**
  - **Worked**: Hyperparameter Tuning on Ridge Regression
  - **Not Worked**: Hyperparameter Tuning on Polynomial Regression, Adaboost

Presenting the Solution

- **Limitations of the system**
  - Not much variance of the label is explained by the features
  - May not generalize to the overall data