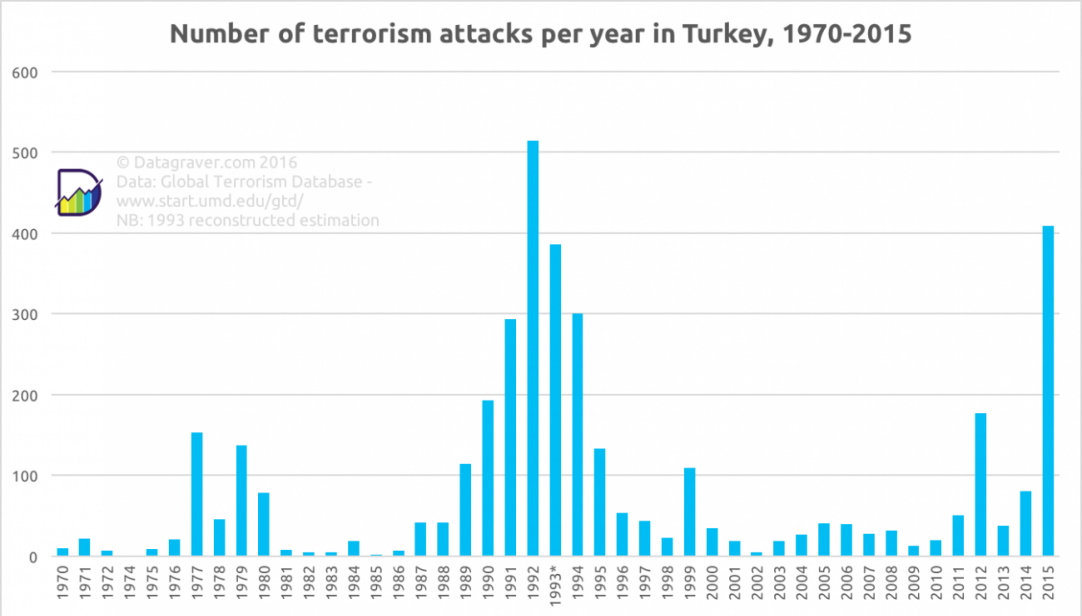
CS210 Final Proposal

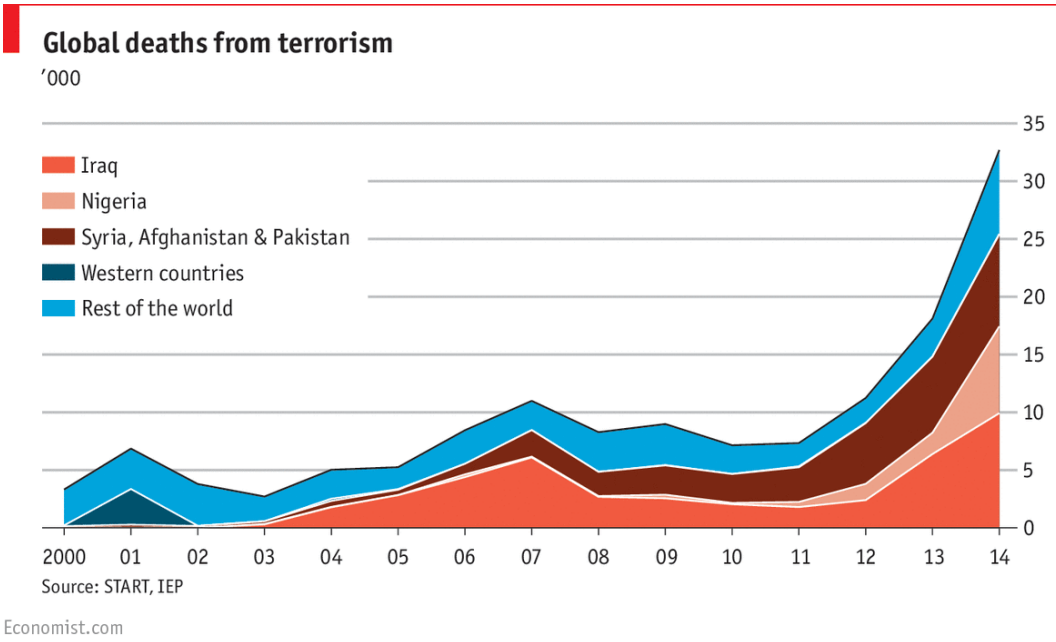
Final Project Blog

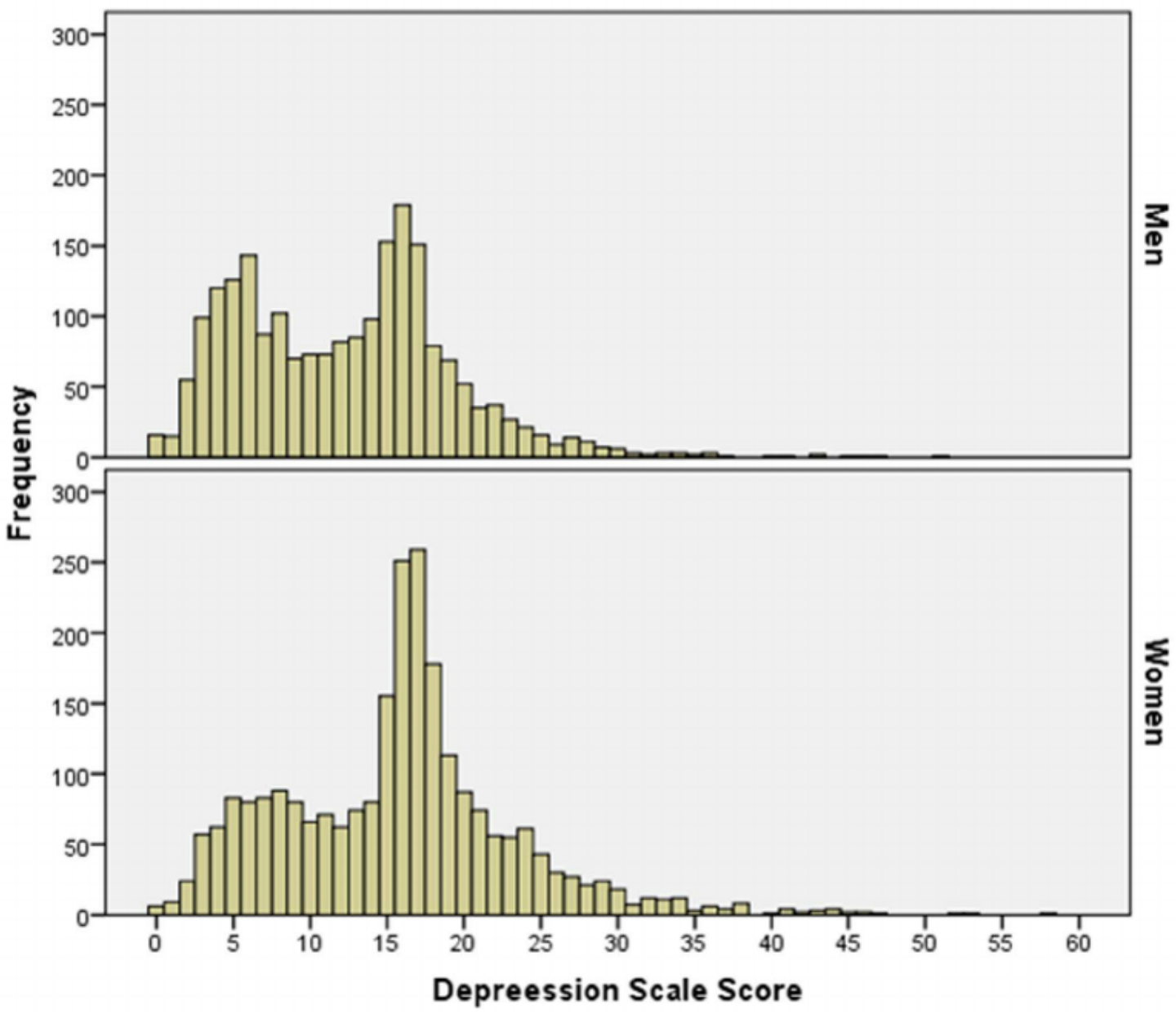
Terrorism Attacks and Happiness Indexes of the Countries

Project Members
Berkay Ersever, Bilge Bahadır Berber, Egemen Yiğit Kömürcü, Ali Yasin Akalın

First data set we will use in this project is about terrorist attacks among different countries. This data set includes information about date, country, city, man killed and injured and the description including the type of the attack such as bomb, pistol, etc.. Moreover, the happiness index data has information about GDP of the country per year, HDI (human development index) per year. Also, it has a column named generation which indicates the category of that group such as Generation X, Boomers, Millenials etc. Every country has several lines of information according to the years, age groups and sexes. In case of need, we can group the data by country to analyze the differences among countries. Other than that, we can analyze how much the civilized country factor affects the type of the terrorist attacks.

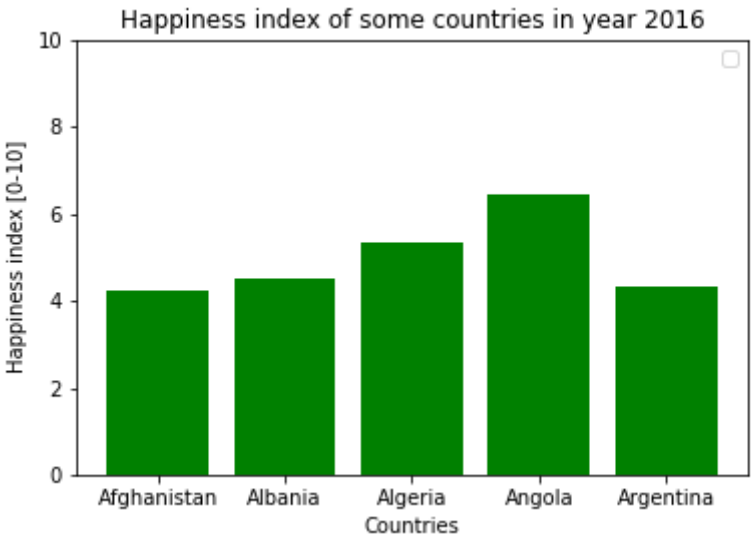




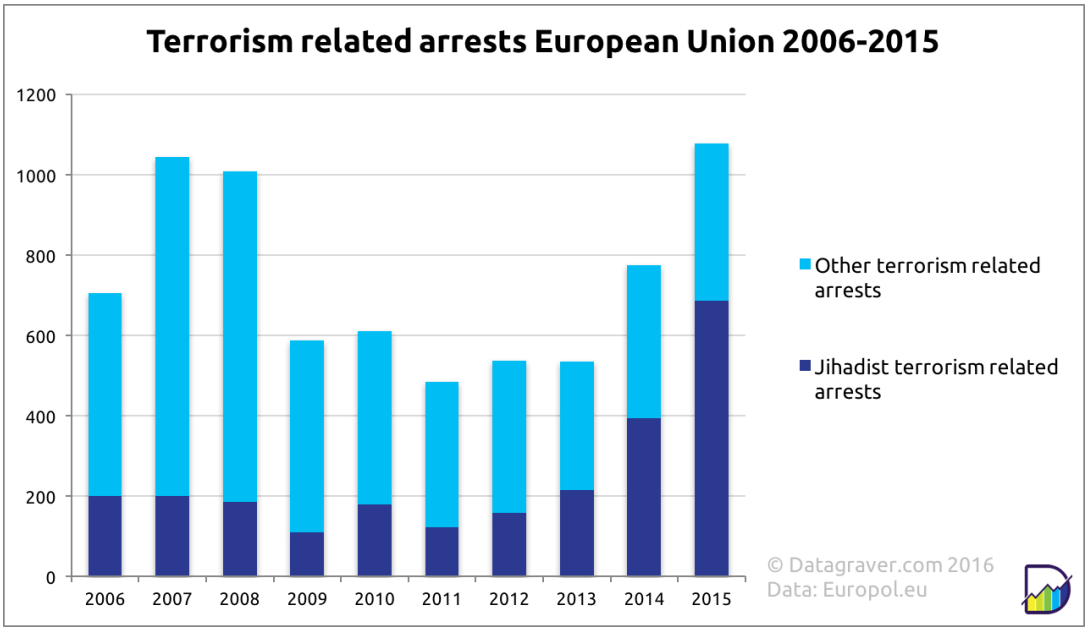


This histogram examines the depressions scales vs. frequency of suicides.

Second data set that we will use is about happiness indexes of the countries. This index consists of different aspects such as life expectancy, freedom, trust etc. There is also information about happiness index standard error for the countries which gives information about the distribution of this happiness index.



Happiness index of 5 countries in 2016



The reason why we choose these data sets is because there might be a relation between the happiness of the country and terrorism attacks in that country. Arguably, countries who have higher happiness indexes should have lower terrorist attacks or can be analysed otherwise. We will analyze these two different data sets throughout the course and test our hypothesis.

Data Description

Datasets in this project are acquired from Kaggle. One of our datasets is about terrorism and the other one is happiness index of the countries. Terrorism dataset includes information about terrorist attacks in the world. It consists of columns like date, country, number of killed and injured people and

description of the attack. Happiness dataset includes columns like country, happiness index, freedom index, trust index etc.

Hypothesis

Our main hypothesis is, terrorist attacks in 2016 did not affect the happiness index of 2017 of the countries. As a result of our hypothesis, we found out that terrorist attacks in 2016 do not affect much the happiness index of 2017 in the related countries.

Linear Regression

In regression part, we tried to estimate the GDP of the countries by considering the total killed people in that country because of the terrorist attacks. As a result of our prediction Syria has 0.72 GDP in regression whilst actual values indicates that it has 0.86 GDP which has an acceptable error.

Machine Learning

In the machine learning part, we implemented decision tree and random forest models. Before creating the model, in preprocessing part we categorized the terrorist attacks by searching specific words in the description columns such as gun, bomb, ambush etc. After categorizing the attacks, we chose it as target feature and tried to predict it. We merged two datasets and used the happiness and other indexes of the countries. Also, we had information about the killed and injured people in that attack. With these features, we tried to estimate the type of the attack.

Project Evaluation

– What were the difficulties you encountered during the project?

As a group, most difficult part of the project was finding appropriate datasets. There were thousands of datasets on the internet but most of them were difficult to handle. Also, it was difficult to find two matching datasets. After finding two suitable datasets, proposing a hypothesis was not that difficult. Linear regression and ML part were not that difficult too. Even though the results of ML techniques were not quite good, we learnt the idea behind ML and implementation of it.

– If you were given sufficient amount of resources, what additional datasets would you utilize?

About terrorism, we could have found plenty of datasets if we had enough resources. For example, number of people joining the terrorist groups in a country per year might work for our project. Also, number of marginal/radical groups in a country can help us to build a better model as well.

– Compare the machine learning algorithms you used, in terms of performance and applicability to your dataset.

We have used two different machine learning techniques; decision tree and random forest. Even though random forest performed little bit better in terms of accuracy, they don't have that much difference. They both performed accuracy around %65 which is poor for us.

– What improvements could have been done in your project?

We could search for different reasons behind the terrorist attack other than happiness and gdp of the countries. If we would able to find and implement some other datasets which are related with terrorism, our model could have been better in terms of performance. Also, our datasets had a 'description' feature originally. Since we would not be able to use it, we turned it into a new feature called 'type of attacks'. We searched specific strings such as 'gun' or 'bomb' and categorized that column in preprocessing phase. In the end, it was our target feature and we were trying to predict the type of terrorist attack. If we could be able to find a better categorized data, maybe it may have worked better.

information (https://cs210proposal.home.blog/category/information/) / news (https://cs210proposal.home.blog/category/news/)

ekomurcu
April 2, 2019May 19, 2019

Blog at WordPress.com.