Examining Homicides with World Bank data

DNSC 6211: Programming for Analytics

Lawrence Gadsden

12/9/2015

Abstract

This project is an exploration of the World Bank's intentional homicide data in relation to several other variables. The topic was chosen because reports of homicides are common in the United States media. At first, I wanted to know where the United States homicide rating is in relation to where the data tells us it should be. I also wanted to discover the main predictors for the intentional homicides. In the project, I was able to use decision trees to predict the United States homicide rate and create rules for predicting homicide rates for new countries.

Contents

1 Introduction

Here, in the United States, one can rarely go a day without seeing or reading a news report of a homicide, or watching a show/movie or playing a video game where a homicide occurs. We are, both, frightened and intrigued by homicide. Hence, when I saw that the World Bank had homicide data, I wanted to explore the United State's position in relation to other countries. I wanted to know where the United States is and where it should be (based on it's profile).

The World Bank defines the intentional homicides indicator as "estimates of unlawful homicides purposely inflicted as a result of domestic disputes, interpersonal violence, violent conflicts over land resources, intergang violence over turf or control, and predatory violence and killing by armed groups".

2 Background

I originally chose approximately thirty World Bank indicators (in addition to the homicides indicator). I noticed that most of the indicators had large amounts of variables missing, so I decided to use the mean of each indicator rather than the last year. I also noticed that many of the correlations between the indicators and homicide were weak. I ended up dropping many of the most incomplete and weak indicators. I then sought out more data from other sources. The two datasets that I ended up adding where data featuring the IMF's grouping of economies and data featuring the geographic regions for each country.

Before starting, I imagined that the project would highly visual. I thought that I would use maps and clustering to learn about the data and showcase my results. However, as the project went on, I got the idea to try a predictive method. I ended up using a deicison tree.

3 Method

The overall question that I answered is in what quartile for homicides should a new country be in based on it's profile (using a decision tree)?

The sub-question is, does the United States have more or less homicides (based on it's profile)?

3.1 Workflow

I started off by grabbing data from the World Bank indicators using the wbdata API. I calculated the means of each year to limit the the affect of missing data. I also used BeautifulSoup to scrape IMF and region data. After this, I uploaded the datasets to a mysql database. I then used R to download the datasets and convert them to data.frames. From here, I had to clean the data so that they could be combined by country. I then ran a plot to see the shape of the homicide data and noticed that the data is skewed to the right. I also examined correlation stats, correlation plots, and boxplots between homicides and the other variables. Next, I decided to split the data and create a decision tree from the training data. Lastly, I made predictions for the test data and the United States using the decision tree.

3.2 Project structure

1. The project contains 6 indicators from the World Bank: homicides, income for the highest 10 percent, water urban, internet users, population total, land area, and net migration. The World Bank Data set contained 214 countries (observations).

2. The second dataset is the IMF data. In the data, countries are classified as having "Advanced" or "Emerging and Developing" economies. The data contains 187 countries (observations).
3. The region data classifies countries by their geographical region. It contains 262 countries (observations).

All, of the datasets look at countries from a different lens.

3.3 Figures and Tables

Figure 2 and **Figure 3** show the distribution of homicides (the target). From the figure one can see the data is skewed to the right. Although the mean is 8.8, the median is only 4.7. This shows that several outliers are increasing the mean.

Figure 4 shows a correlation matrix between homicides and its most correlated continuous variables. Homicides tend to occur in countries where the richest 10 percent have a higher income share. On the other hand, they occur less as a country has more internet users (per 100 persons).

Figure 5 and **Figure 6** show box-plots of homicides by region and economic group. **Figure 5** shows that the number of homicides are much higher in Latin/South America and moderately higher in Africa. **Figure 6** shows that the homicide rates are typically much lower in advanced countries, than developing countries.

Figure 7 is the decision tree from the training data. It provides rules that help to place countries into a homicide quartile.

4 Discussion

This project was interesting. When I first got the homicide data, I saw that the United States only had 4.775 homicides (per 100,000). I thought this was an extremely low number compared to other countries. I also glanced at the mean (which is 8.8), so I figured that the United States would be average for its profile.

However, things did not go as smoothly. While the United States had a relatively low score overall, it was high when compared to similar countries.

4.1 Learnings

I learned that combining data can be very difficult. I had to do quite a bit of cleansing just to match the country names. Sadly, I imagine that matching country names is simple compared to other possible data combinations. After the project, I would like to check out GitHub or stackoverflow to view how other programmers are combining data.

I also re-learned that while R has many libraries, it could be extremely difficult to choose one. This made me appreciate the simplicity of python modules such as sklearn and matplotlib.

4.2 Challenges

I really wanted to use more detailed homicide and criminal data from unodc.org. They had nice stats on solved homicide rates, prison rates, objects used in homicides, and reason for homicide (domestic,drug,gang). Unfortunately, I could only get their data in a csv.

I also had planned on using clustering and regression. However, I found it difficult to preprocess the data, so I ended up deciding on a tree because of the small amount of reprocessing needed and their readability.

5 Conclusion

The decision tree shows that the United States is an outlier based on it's profile. The tree predicted that it would be in the lowest quartile, but it ended up being in the mid-high (third) quartile. – The median is 4.7 and the US homicide rate is 4.77.

While the United States low prediction will most likely stay consistent because of it's profile, the tree will be different and some other predictions will change with random training partitions. Interest, the current tree does not have a rule for predicting median-high countries, so it predicted all of the median-high countries wrong.

Altogether, I enjoyed learning how to build a decision tree in R. I intend on learning more about pre-processing, transformations, and normalization so that I can make use of other methods that expect more structured data.

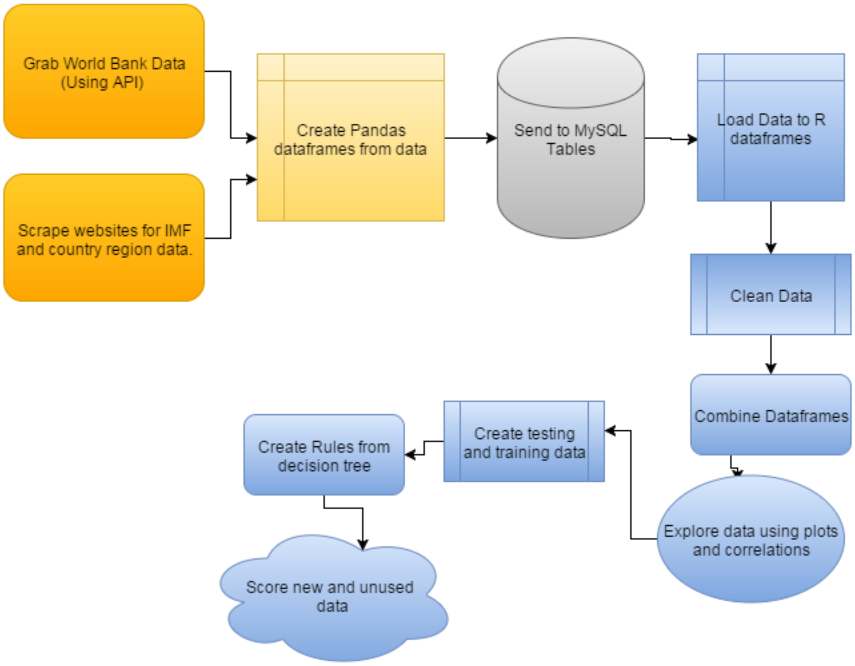
Grab World Bank Data (Using API)
Scrape websites for IMF and country region data.
Create Pandas dataframes from data
Send to MySQL Tables
Load Data to R dataframes
Clean Data
Combine Dataframes
Explore data using plots and correlations
Create testing and training data
Create Rules from decision tree
Score new and unused data

Figure 1: The project workflow

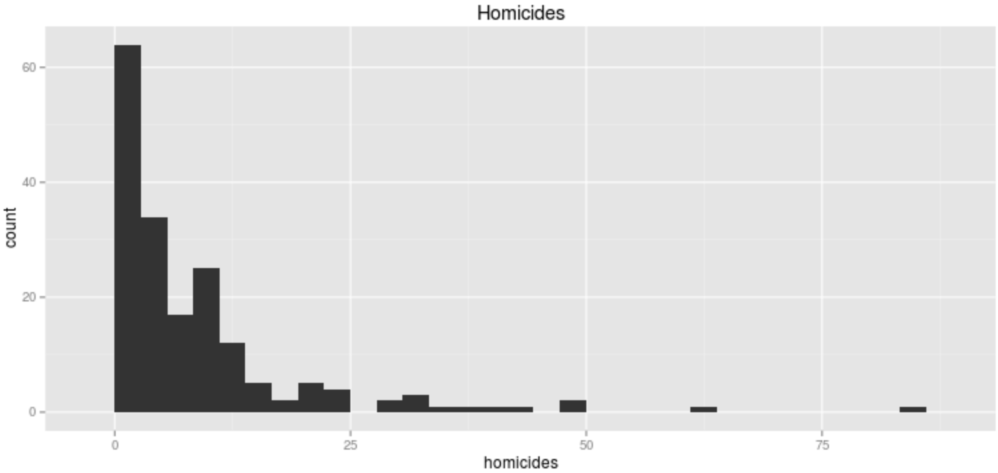

Figure 2: Homicides plot

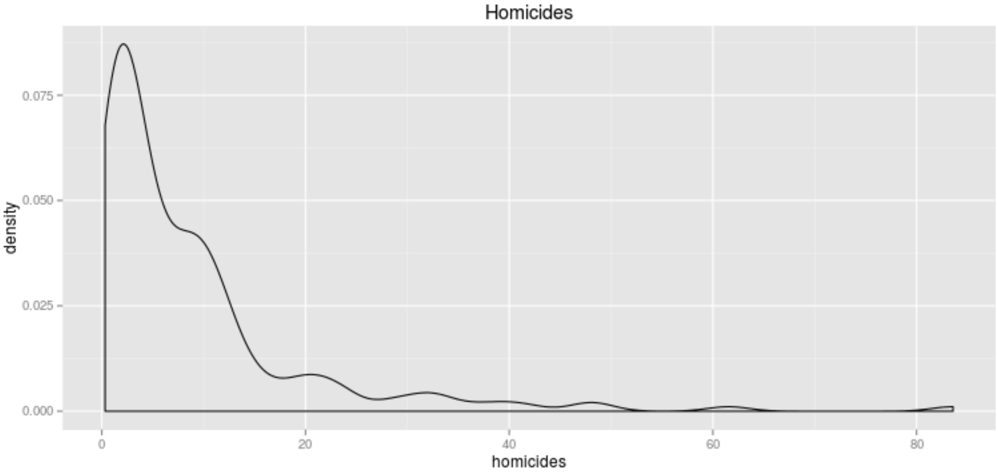

Figure 3: Homicides Density

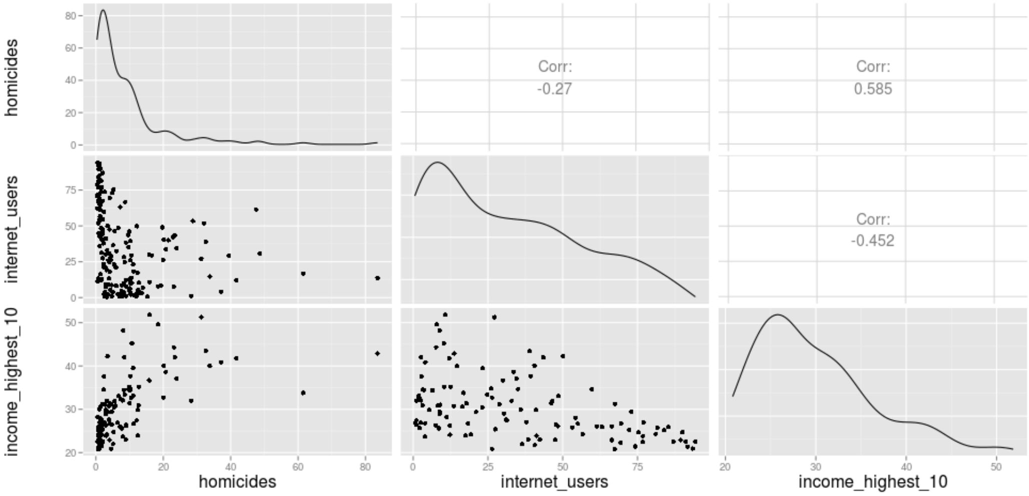

Figure 4: Correlation Matrix

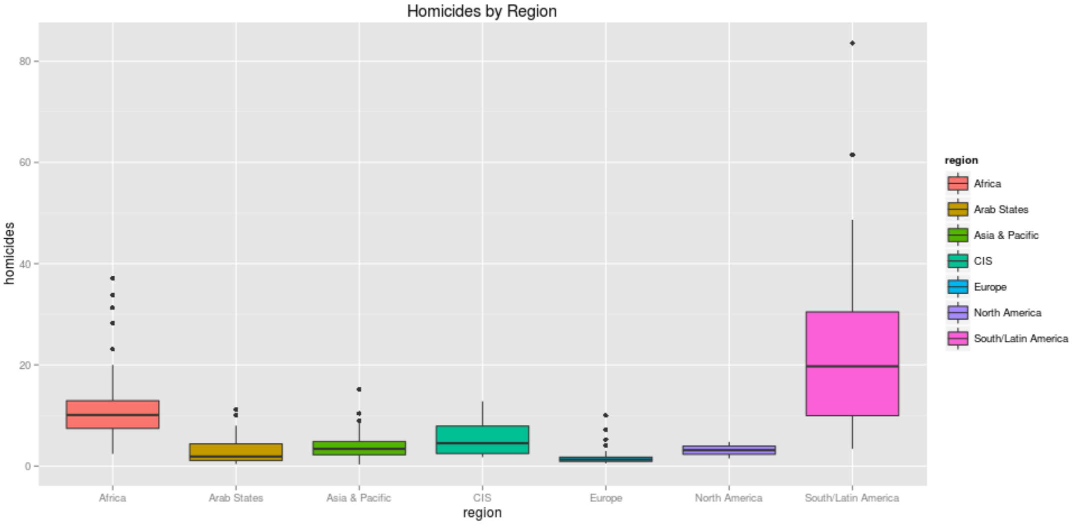

Figure 5: Homicides by Region

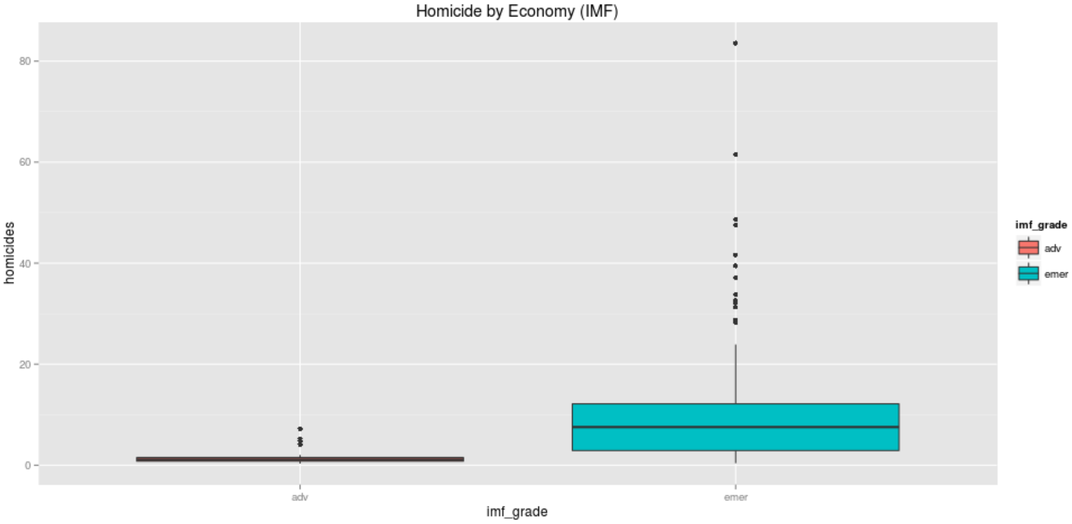

Figure 6: Homicides by Economic Classifier

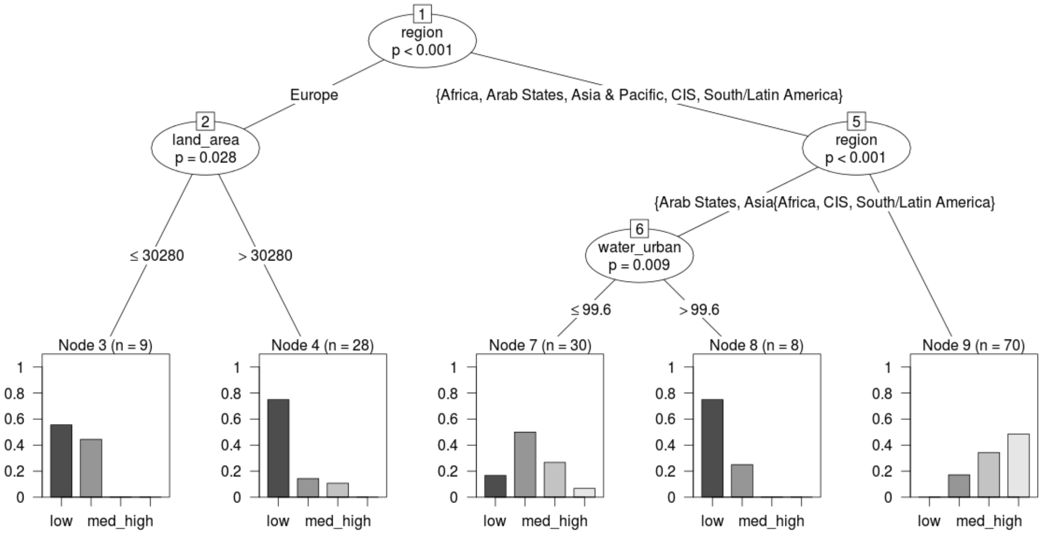

Figure 7: Decision Tree