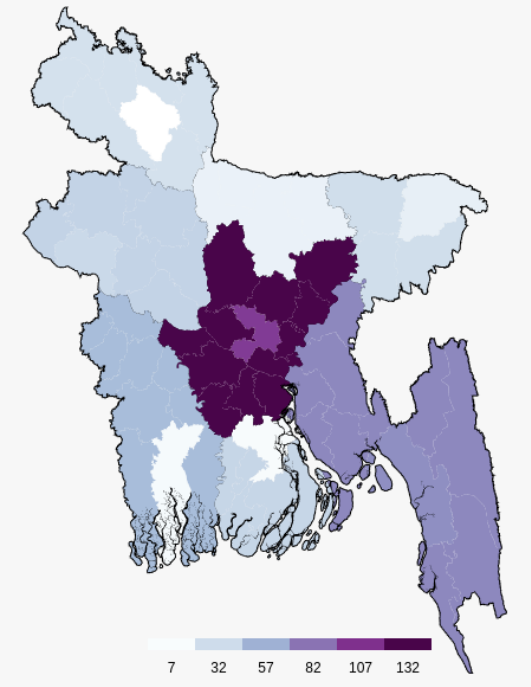
Visualizing Crime in Bangladesh: Region-Specific Insights (2010-2018)

Candidate Number 1053004

I declare that, except where otherwise indicated, this mini-project is entirely my own work, and that it has not been previously submitted and/or assessed and is not due to be submitted on its entirety or in part for any other course, module or assignment.

Contents



Robbery in Bangladesh (2017)

Overview

Bangladesh, the 8th most populous country in the world, is located in south Asia. It has the 11th highest population density in the world, and its capital city is Dhaka.

Bangladesh was ranked 20th in the 2023 world crime index. It has a hierarchical regional division system - the country (as of 2015) is divided into eight tier 1 divisions, which are themselves divided into multiple tier 2 divisions, and so on.

The intended audience for my visualization project is anyone interested in crime and safety in Bangladesh, including policymakers, law enforcement agencies, researchers, and the general public.

By visualizing crime statistics using an interactive choropleth, correlogram, and linked choropleth and heatmap view, I hope to help the audience gain a better understanding of the patterns and trends of crime in the country.

Policymakers could use the insights gained from my visualization to make informed decisions related to crime prevention strategies and allocation of re-

sources for law enforcement. Law enforcement agencies could use the visualization to identify crime hotspots and allocate resources accordingly. Researchers could use the data and insights presented in my visualization for further analysis and study.

Finally, the general public could use the visualization to gain a better understanding of the crime situation in Bangladesh and take necessary precautions to ensure their safety.

Data

Sources:
The crime data was taken from Kaggle
The administrative boundary data was taken from Github

Original Data: The administrative boundary data was in the GeoJSON format and required no preprocessing.
The crime data was in a CSV file with one line per area per year. There was one column per crime, and the crime rates were all integers.

It should be noted that the data is incomplete in the sense that the division system changed during this decade; in 2015 the Mymensingh division split off from the Dhaka division, and in 2017 the Rangpur division split off from the Rajshahi division.

Furthermore, the 2019 data is clearly erroneous and an order of magnitude smaller than the rest of the data, so I have chosen not to include it. I believe the data may be from early 2019, meaning the 2019 data may just reflect the first few months of the year.

Note that while no data preprocessing was required for the crime data, the column names contained underscores, trailing spaces and questionable capitalisation, so they are modified a bit before being displayed to the user.
Furthermore, since the word "dacoity" isn't too well known, I opted to render it as "gang robbery" instead.

The ranges of each column are detailed below

Column	Min	Max
year	2010	2019
dacoity	0	184
robbery	0	294
murder	0	1395
riot	0	56
woman_child_Repression	0	5115
kidnapping	0	204
police_assault	0	336
burglary	0	686
theft	2	2240
other_cases	9	22429
speedy_trial	0	563
recovery_cases_arms_act	0	723
recovery_cases_explosive	0	387
recovery_cases_narcotics	44	22682
recovery_cases_smuggling	0	2509

There is also the categorical `area_name` column which has cardinality 17 (The eight regions each have a `metropolitan` and `division` data point, as well as one additional `railway range` division).

Derived Data:

Perason Correlation Coefficient: a standard statistical measure of correlation between two variables. This is used to help illustrate the fact that crime rates are not independent per crime, and that higher rates of one crime almost always suggest higher rates of every other crime.

Crime Severity Index: a key theme of this project is the observation that across all regions and periods, there are correlations between almost every pair of crimes. This leads to the natural question of asking how reliably we can predict crime rates based on other factors, or based on the number of occurrences of other crimes. I derive a 1 dimensional statistic per-region (or per-region-per-year) which I'm calling "crime severity index". This statistic is meant to be a holistic representation of general crime rates, which can be used to predict the number of occurrences of any crime with a surprising degree of accuracy. An interesting avenue of exploration would be to ask how this metric relates to population, poverty, etc, however I didn't have time to do that in this project.

Understanding the way we calculate this metric is not important for the user

The point I aim to communicate is that this 1d representation is still very predictive of the number of occurrences of each crime.

To calculate this metric, we first need to represent each region & year pair (or just region) as a vector ("crime vector"), with one dimension per crime. In the per-region-per-year case, each vector dimension represents the number of occurrences of one type of crime in that region in that year. In the per-region case, we average occurrences per crime across all years for each region.

Once we have our crime vectors, we use principal component analysis, to project these crime vectors down to 1d space. This gives a linear embedding of areas / areas and years. From this representation, we can then (imperfectly) project back up to the higher dimensional space, which will yield predictions for the number of occurrences of each crime. Comparing these predictions to the actual observed occurrences helps us spot outliers, e.g. the 2013 riots.

Note that for robustness purposes we normalise the matrix of vectors columnwise before performing PCA and denormalise afterwards. This ensures that each crime type holds an equal weight in the calculations.

The actual value of this statistic is unimportant since it's meant to be a relative

measure. However, for the reader's interest, the values are presented in the table below.

Area Name	Crime Severity Index (0 centred)
dhaka division	6.58
chittagong division	5.05
dhaka metropolitan	4.35
khulna division	2.07
rajshahi division	1.62
rangpur division	-0.02
sylhet division	-0.07
barisal division	-0.45
chittagong metropolitan	-0.83
mymensingh division	-1.28
rajshahi metropolitan	-1.98
sylhet metropolitan	-2.04
khulna metropolitan	-2.39
barisal metropolitan	-2.47
railway range	-2.68
gazipur metropolitan	-2.71
rangpur metropolitan	-2.75

Note that the projection back up to the high dimensional space is linear, i.e. for each crime there is a fixed α, β such that

$$\text{Predicted \# occurrences of crime in a region} = \alpha_{\text{crime}} \times \text{ crime index } + \beta_{\text{crime}}$$

E.g. for murder we can make predictions with

$$91.7 \times \text{ crime index } + 250$$

Goals and Tasks

1. Identify the years and regions which had the most of each crime

In domain specific terms, this means to identify which areas and in which years Bangladesh had the most occurrences of each crime.
In abstract terms, this means to identify which areas have the highest crime data points relative to other regions and years.

2. Recognise that crime rates are correlated

In domain specific terms, this means to understand that generally speaking, regions that are high in any crime are high in every crime, and vice versa.
In abstract terms, this means to recognise that regardless of which areas or period we consider, there is generally a large Pearson correlation coefficient between the number of occurrences of any pair of crimes

3. Identify which regions have the highest and lowest general crime rates

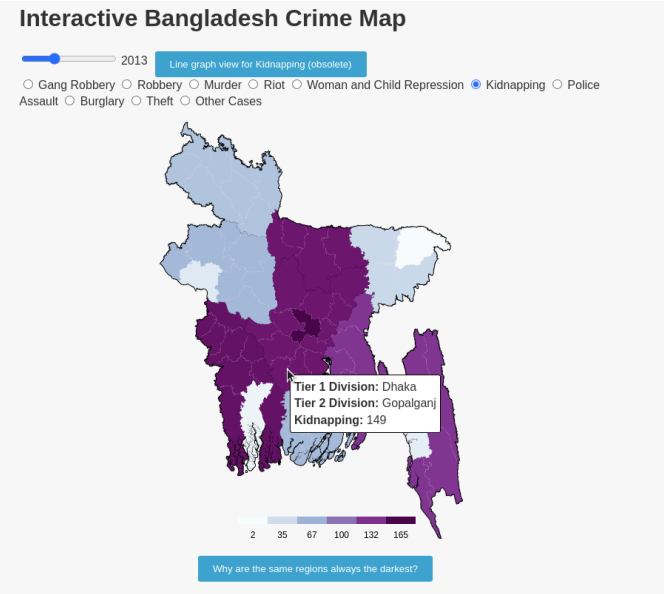
In domain specific terms, this means to have a holistic understanding of which regions of Bangladesh have the highest crime rates, and which have the lowest.
In abstract terms, this means to understand and digest the crime severity index metric and understand what it represents.

4. Identify which regions had *relatively* high and low rates of crimes in which years

In domain specific terms, this means to identify instances where a particular crime occurred an unprecedentedly high or low number of times in a region
In abstract terms, this means to identify data points that have a significant deviation from the predictions yielded by the crime severity index metric.

Visualisation

Interactive Choropleth View:



The first interface is an interactive choropleth view. This view allows the user to switch between different crimes, and provides a slider allowing the user to see how the data evolves from year to year.

Each region is represented as a shape on the map, and the number of occurrences of the selected crime in the selected year determine the fill colour of the region

The code to draw the map is quite fast, meaning that dragging the slider forwards and backwards in time leads to a smooth, reactive transition of the data throughout the years.

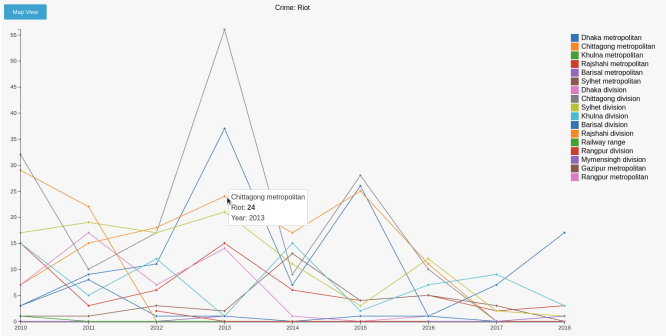
There is a tooltip which, upon hovering over a point on a map, will display the names of the tier 1 and tier 2 region that point is in, as well as the relevant data point.

Interacting with this view for a while leads to a very noticeable pattern, namely that the same regions (mainly Dhaka, Khulna and Chittagong) usually have the highest crime rates regardless of crime or year.

This view is primarily intended to contribute to the first and third goals. The user should get a sense of the general regional crime rates and will likely notice that some regions generally have higher crime rates than others, contributing

to the second goal.

Line Graph View:



The choropleth view recalculates the colour scale and redraws the legend each time the year or crime selection is changed. As a result, the trends over time can sometimes not be obvious, e.g. a region could have the same number of murders in two consecutive years, but be drawn a different colour in each rendering. As a result, I have included a line graph view that makes the absolute values more apparent.

Each region is represented as a line. The x axis represents time, and the y value of each point represents the number of occurrences of the selected crime in the corresponding year

I have labelled this view "obsolete" since the line graph doesn't lead to many obvious conclusions - the patterns are more apparent in the "Absolute vs Relative Crime Rates View" detailed later.

Furthermore, since there is a large deviation between the general crime rates between regions, a line graph doesn't do the data justice. The most dangerous regions dominate the view, and the regions with lower crime rates could have dramatic trends that don't look very dramatic on the line graph.

This view is primarily intended to contribute to the first and third goals, and is intended to support the choropleth view, since they are quite similar in the sense that, for the selected crime, they show the absolute number of occurrences of said crime for each region.

Correlation View:

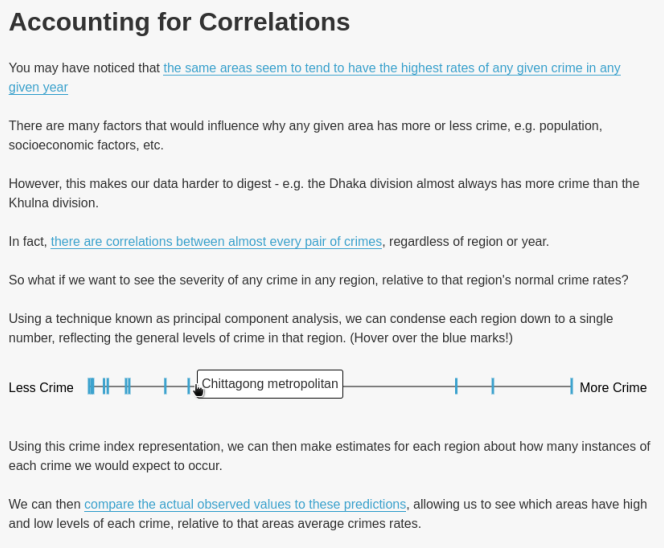
As mentioned earlier, interacting with the choropleth leads one to notice that the same regions tend to have higher and lower crime rates. I have implemented a correlation view allowing the user to explore the extent of this effect.

There is one mini scatter plot for each pair of crimes, coloured depending on the correlation coefficient.

This view allows the user to select any subset of the regions, and a range slider that allows the user to choose a range of years. In almost all cases, regardless of the regions or years selected, there will be a correlation between any pair of crimes.

This view is intended to hit home the second goal, convincing the user that crime rates are not independent of one another.

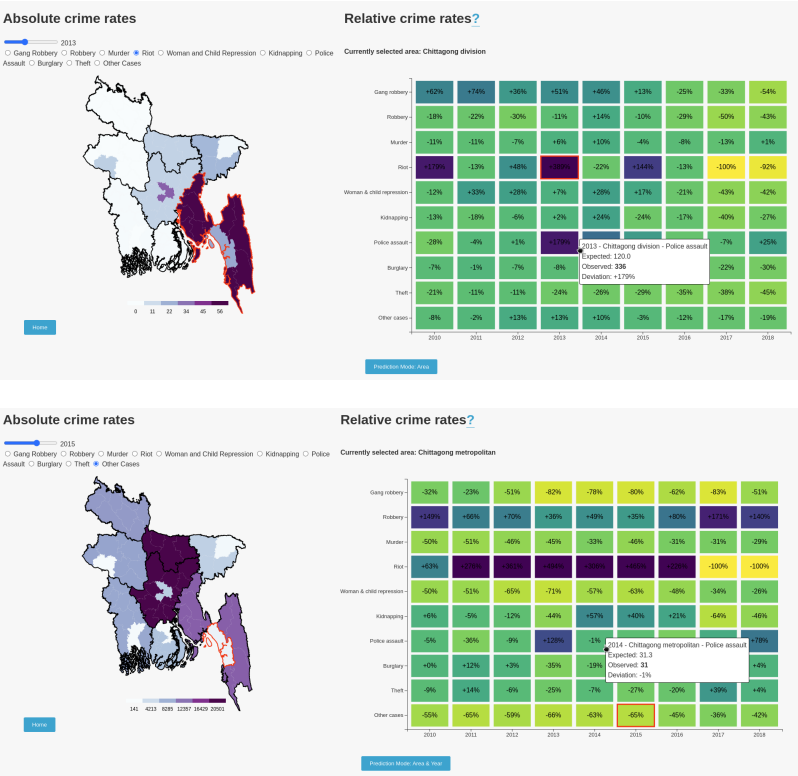
Crime Severity Index Explanation:



At this point, the user should have observed that the same regions tend to have more/less of any crime, and that there are pairwise correlations between the rates of almost every crime. This page introduces the crime severity index metric. Since this metric is 1 dimensional, I have represented it as a line with one mark per region.

This view contributes to the third goal, however it doesn't convey much information. It's more of a transition page between the other more involved views.

Absolute vs Relative Crime Rates View (innovative):



This view features a choropleth alongside a heatmap.

There is one heatmap view per region, and clicking on a region in the choropleth will select that region and update the heatmap to show that region's data.

Above the heatmap, the currently selected region is displayed. The heatmap communicates the selected region's **relative** crime rates; for a given region, we can predict the number of occurrences of each crime in each year from the crime severity index as explained earlier. By comparing the actual observed value to this predicted value, we can get a sense of whether this crime occurred a disproportionately high or low number of times in the given region in that year.

In the first image above, it should be clear that the Chittagong division experienced an exceptionally high number of riots in 2010 and 2013, and that gang robbery was initially very prevalent in the region at the start of the decade, and much less of a problem by the end of it.

The second image should highlight the hugely disproportional number of riots

in the Chittagong metropolitan between 2011 and 2016, as well as the proportionally high number of robberies and low number of other cases.

The rectangles are coloured based on the % deviation from the predicted value, and hovering over each rectangle will display the raw predicted / observed values. This view will allow users to easily spot instances where a crime happened a relatively high or low number of times.

There is a button beneath the heatmap which allows the user to toggle between per-region predictions and per-region-per-year predictions.

The choropleth is almost identical to the choropleth from the first view, other than the fact it has the currently selected region highlighted.

Furthermore, clicking on a heatmap rectangle will highlight that rectangle and update the choropleth view to display the crime rates for that crime in that year.

This view allows the user to efficiently digest the crime data, and get a sense of the crime rates in different regions. For any region, even at a glance it's quick to see which crimes were more and less prevalent in which years. If a particular heatmap entry is of interest, the user can click it and immediately see a choropleth displaying the occurrences of that crime in that year across all regions.

For the choropleth, there is one mark per region. The channel is colour which represents # occurrences of the selected crime in the selected year.

For the heatmap, there is one mark per crime per year. The main channel is colour which represents the % deviation from predicted value to observed value, however the raw predicted and observed values are themselves also communicated via the tooltip.

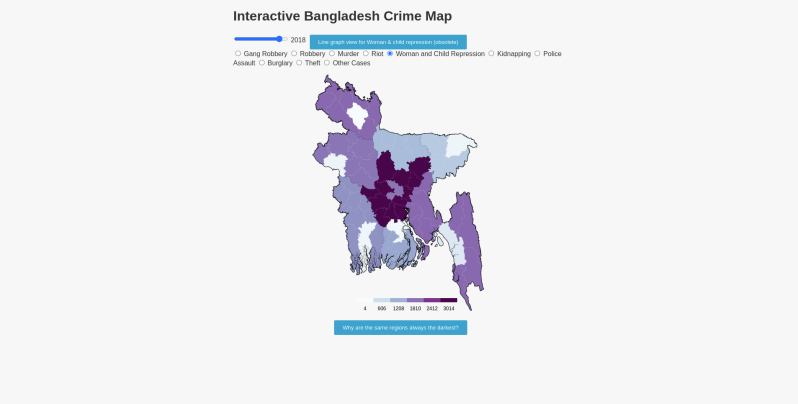
The colour scheme for the heatmap, viridis, is colour-blind friendly, in order to facilitate spotting outliers.

This view is intended to achieve the fourth goal. The scale and locality of the 2013 Bangladesh riots become very apparent, and it's clear to see which crimes were most prevalent in which regions.

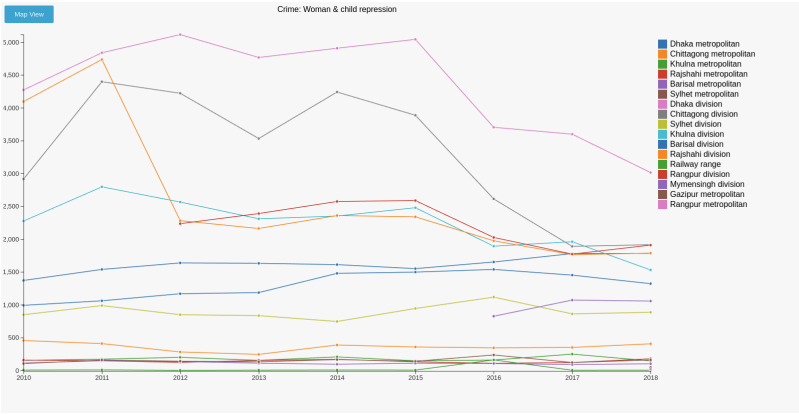
Usage Scenario

The user is first displayed the landing page

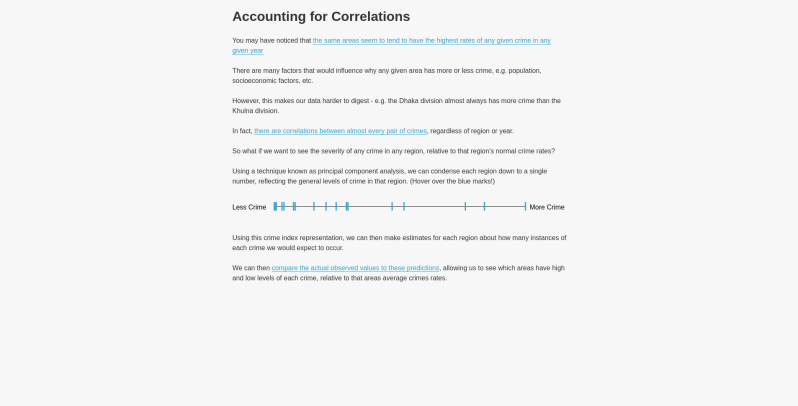
Clicking the button will lead the user to the choropleth view.



After interacting with it a bit, they may click the line graph view button



After clicking back, they may click the "why are the same regions always the darkest" button, which will bring them to the PCA explanation page



If they click the first link, it will lead them back to the choropleth page. If they click the second link, they will be met with the correlogram view

Once they've played with the correlogram view, clicking the back button will take them back to the PCA explanation page.

After reading the page, they should have a general understanding of the crime severity index metric. When they do, they should click the final link, leading them to the linked correlogram & heatmap view

On this page, the user should now understand what the heatmap conveys. If they do not, the ? link will take them back to the PCA explanation page.

On this page, the user should be able to efficiently recognise anomalous data points, and digest which crimes are/were most prevalent in which regions.

Afterwards, they may click the home button, which will take them back to the landing page, where they are free to navigate through the views at their leisure.

Credits

• The correlogram view has a range slider, that differs from a normal slider in that it allows the user to choose a start date and an end date. I took the code for this from StackOverflow

• The correlogram also calculates pairwise correlations between different crimes. I took the code for this from StackOverflow

• Lastly, the crime severity index attribute is calculated using principal component analysis. I used this library to calculate the direction of the first component (i.e. first eigenvector of $X^T X$). The rest of the implementation was my own code.