Statistical & Sentiment Analysis of Factors Affecting State-wise Tourism in India

(Data Mining and Warehousing Project)

Made By:

Akshay Nagpal
12-CSU-022

Akhil Atreya
12-CSU-018

Supervised By:

Dr. Latika Singh

CSE & IT Department

THE NORTHCAP UNIVERSITY (Formerly ITM University, Gurgaon)

CONTENTS

INTRODUCTION

The tourism industry of a country is a very prominent measure of its progress in the world. A higher tourism indicates the changing perception as well as acceptance of a country by other nations.

This project factors in the tourism and infrastructural data and relate it to the crime of each state of India. The aim of the project is to understand the relation between the 3 factors using data visualization as well as sentiment analysis to study the effect of the various factors on the tourism of the country.

The sentiment analysis done, aims to find out the reasons, if any for the changes in any of the parameters over the 10 year period that has been taken as a sample. The assumption behind the project being that whenever there is any change in one of the parameters, there is usually a change observed in any one or both the other parameters.

We aim to find out the reasons for changes in any one of the parameters, ranging from demographic, political, social, economic etc. and relate it to the other parameters.

DATA MINING OBJECTIVES

4 main objectives are:

1. To plot the graph of each state and union territory of India showing tourism, crime rate and number of hotel rooms between 2003-2013 and find the visual relationship between them.
2. To plot the above parameters for each state for every year from 2003-2013 and analyze how the states compare against each other for every year.
3. To perform regression analysis to quantify the dependence of crime and infrastructure on tourism.
4. To perform sentiment analysis of tweets for the state showing most positive trends.

PROJECT PLANNING AND EXECUTION

The planning and execution was done using CRISP-DM (Cross-Industry Standard Process for Data Mining).

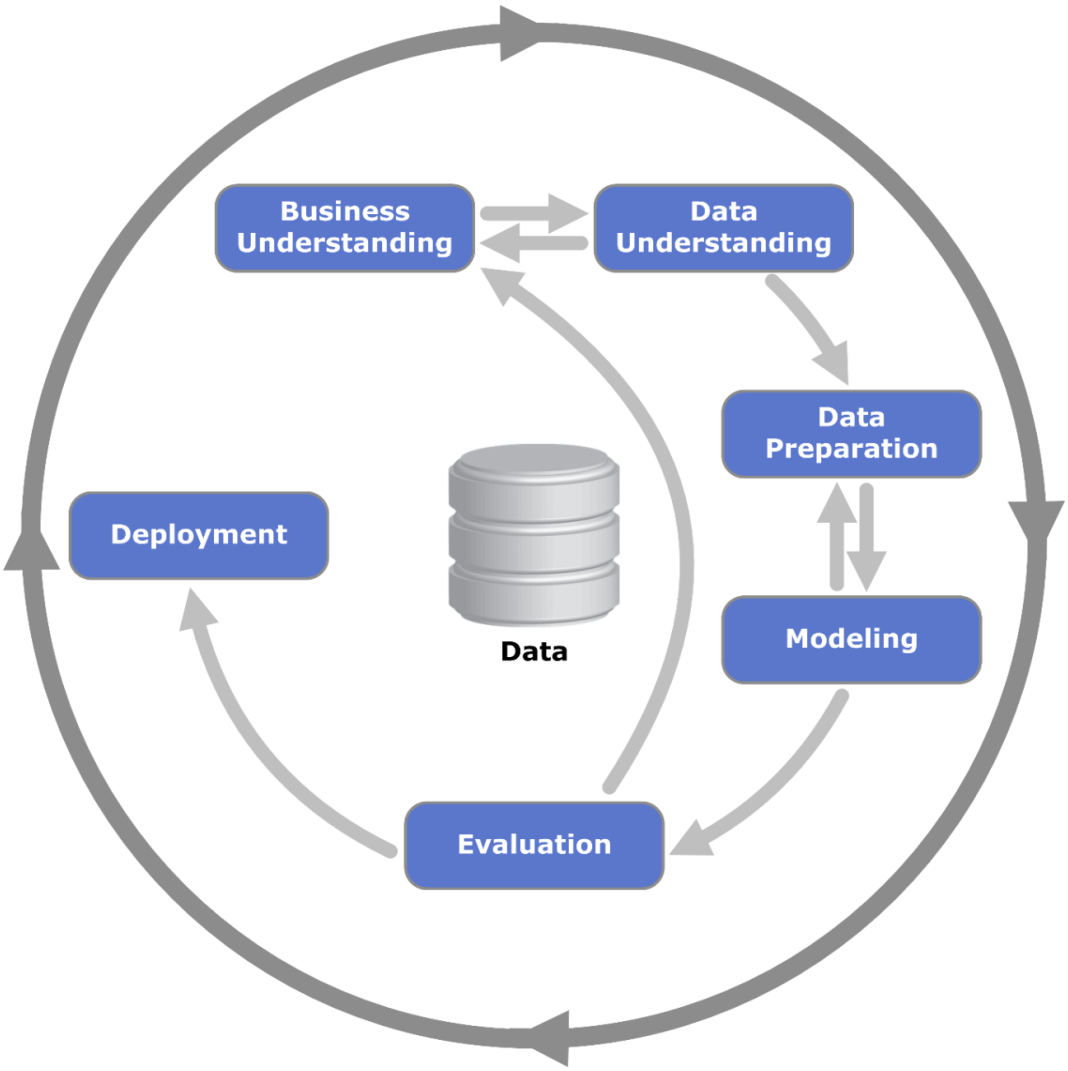


Figure 1 "CRISP-DM Process Diagram" by Kenneth Jensen. (Licensed under CC BY-SA 3.0 via Commons)

The corresponding project steps are given below:

CRISP-DM STEP	PROJECT STEP
BUSINESS UNDERSTANDING	The 4 project objectives were discussed and set. Project plan was made.
DATA UNDERSTANDING	Data was collected from 3 sources, explored and its quality verified.
DATA PREPARATION	The desired attributes from the datasets were selected and subsets of the original dataset was created. 3 subsets were integrated into a single set and values formatted.
MODELING	Plotting, color coding and regression was done in order to have the results for descriptive data analysis. Also, sentiment classifier was modelled for a single state or union territory based on manually classified training dataset of tweets.
EVALUATION	The results of previous step were verified by searching for trends on internet, related to rise and fall in the plots. Sentiment classifier was applied to the whole dataset and results were obtained successfully.
DEPLOYMENT	Final report was produced and project was reviewed by supervisor.

DATA UNDERSTANDING

Data was collected from 3 locations:

1. Open Government Data(OGD) Platform India (data.gov.in)
   From the above site we collected the crime statistics from 2003-2013. It had the following attributes:

STATE/UT
YEAR
CRIME HEAD
Male 16-18 years
Female 16-18 years
Total 16-18 years
Male 18-30 years
Female 18-30 years
Total 18-30 years

Male 30-50 years
Female 30-50 years
Total 30-50 years
Male Above 50 years
Female Above 50 years
Total Above 50 years
Total Male
Total Female
Grand Total

2. Ministry of Tourism (tourism.gov.in)
   From the above website we collected publications in PDF format titled 'Indian Tourism Statistics at a Glance' for the year 2003-2013. Statistics about state-wise domestic tourism, international tourism and number of govt. approved hotel rooms were obtained. The following attributes were there:

Tourism	Hotels
State/UT	State/UT
Year – Domestic Tourism	Rooms for each Type of Hotel
Year – Foreign Tourism	Total Rooms

3. Kimono API
   We used this API to collect tweets from 2009-2013 with hashtags relevant to our requirement. 100 tweets per year were collected from 2009-2013 i.e. 500 total tweets with the following phrases: 'weather', 'delhi'.

DATA PREPARATION

Selecting attributes

From the first dataset i.e. crime, we selected the following attributes as we were concerned with the quantitative side of the data.

STATE/UT
YEAR
Grand Total

From the second dataset, in the Tourism table, we decided to combine *'Year – Domestic Tourism'* and *'Year – Foreign Tourism'* together as both contributed to the tourism of a state. Hence the resulting attributes were:

State/UT
Year – Total Tourism (Domestic + Foreign)

In the Hotels data, we selected the following attributes as all kinds of hotels are occupied by various economic sections of the society.

State/UT
Total Rooms

After selection we integrated the above attributes using State/UT attribute and got the following structure:

Name of State/UT	**Tourism 2003**	**Hotel Rooms 2003**	**Crime Total 2003**	**Tourism 2004**	**Hotel Rooms 2004**	**Crime Total 2004**	**And so on for every year till 2013…**

Cleaning the Data

Cleaning of the dataset selected with above attributes was also carried out to get rid of the following problems:

1. The names of states were inconsistent over data collected from various sources. For e.g. Pondicherry and Puducherry, Andaman & Nicobar Islands and AN Islands, Orissa and Odisha etc. Such inconsistencies were resolved by deciding a single name by mutual discussion.

2. The values of attributes had different exponential ranges. Number of hotels were mostly in the range of 10^2 to 10^3, and crime count was in the range of 10^2 to 10^4, while the tourist count went up to the range of 10^8. Hence, we had to scale the data by dividing all the values by 10^3.

DESCRIPTIVE DATA ANALYSIS

There were 2 kinds of plots:

1. State-wise Tourism & Hotel Rooms & Crime v/s time (2003-2013)
2. Year-wise (2003-2013) plots for each state for each of the 3 attributes (tourism, crime, hotel rooms) on the map of India.

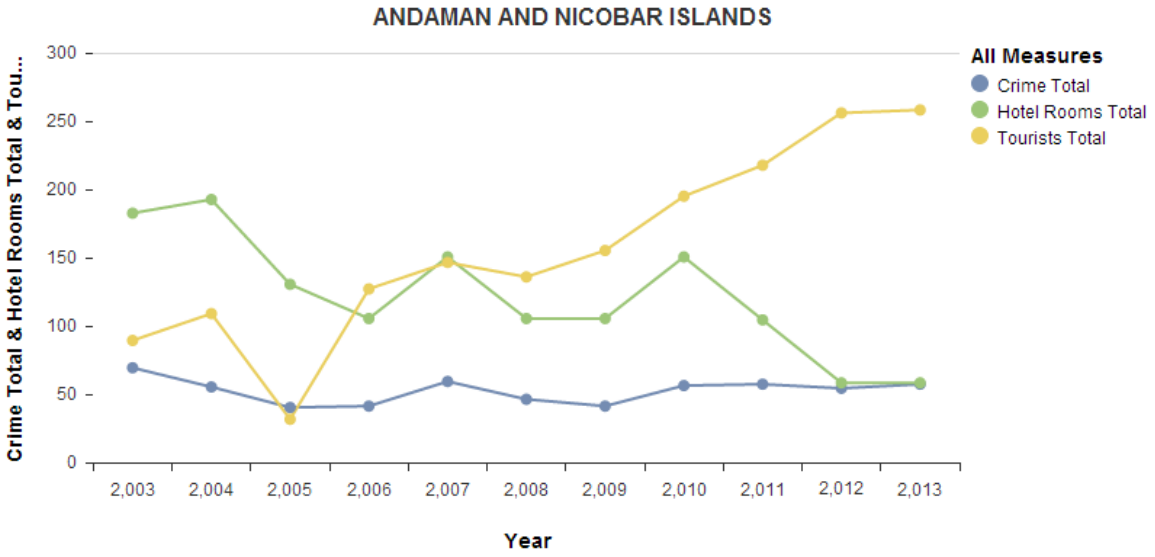
State-wise Tourism & Hotel Rooms & Crime v/s Time (2003-2013)

We plotted the 3 attributes with respect to time and found that in most of the plots, one or more attributes showed extensive rise and fall in certain time periods. We searched on the internet in order to find any events that occurred in the particular state within that period.

Regression was also done to quantify the effect of crime and hotel rooms on tourism. **Correlation coefficient (sqrt(R^2))** was obtained for every state, and only those values that indicated a strong relationship (**R > 0.7**) were recorded for the states and union territories.

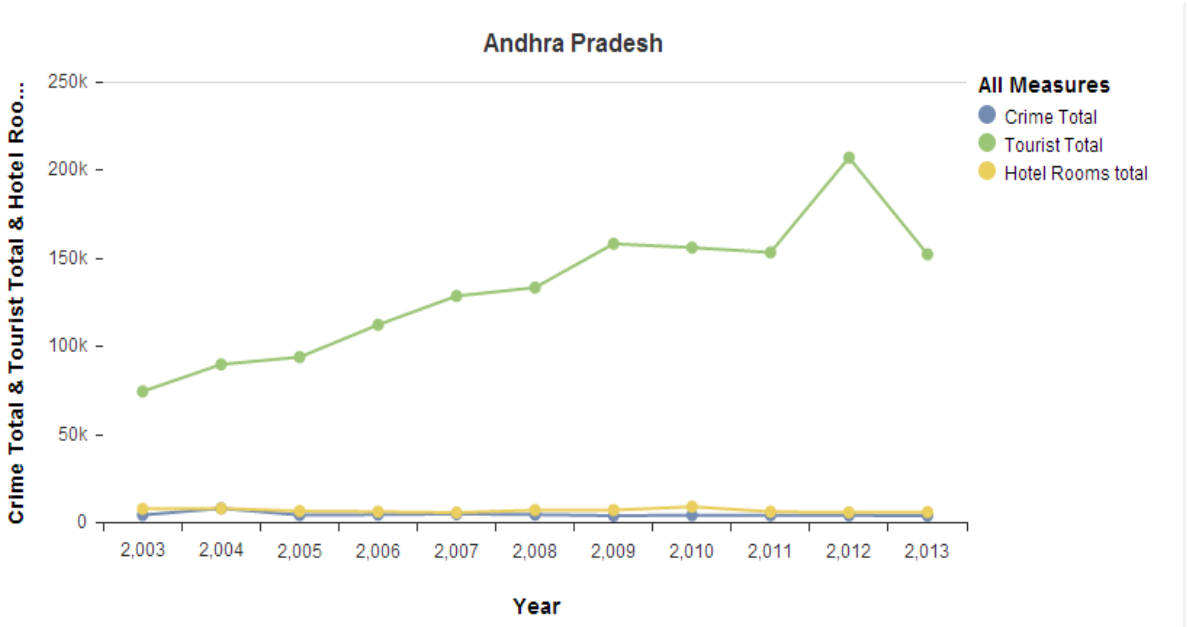
The state-wise plots along with the R value (**X = crimes and Y = tourism**) and related events are given below.

Andaman & Nicobar Islands



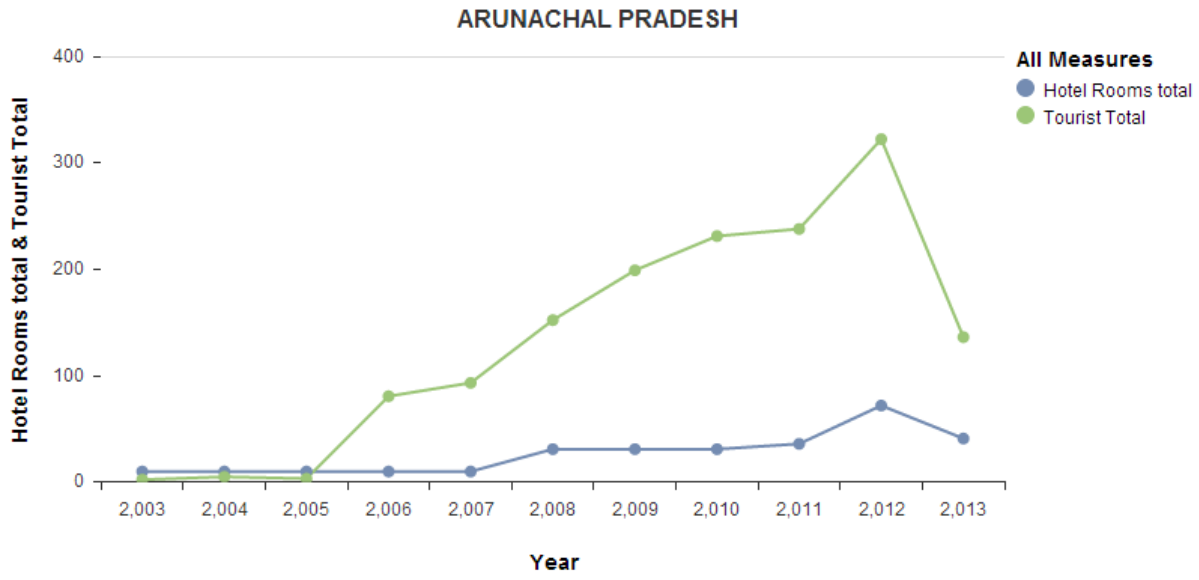
Tourism: Saw a sharp decline post 2004 in the wake of the Indian Ocean Tsunami.
Hotel: The number of hotel rooms also declined due to the destruction by the Tsunami. With major earthquakes in August 2009 and March 2010, the loss of property again resulted in a fall in the number of hotel rooms.
Crime: The crime rate has remained fairly stable throughout the years there with no spike in any criminal activities observed.

Andhra Pradesh



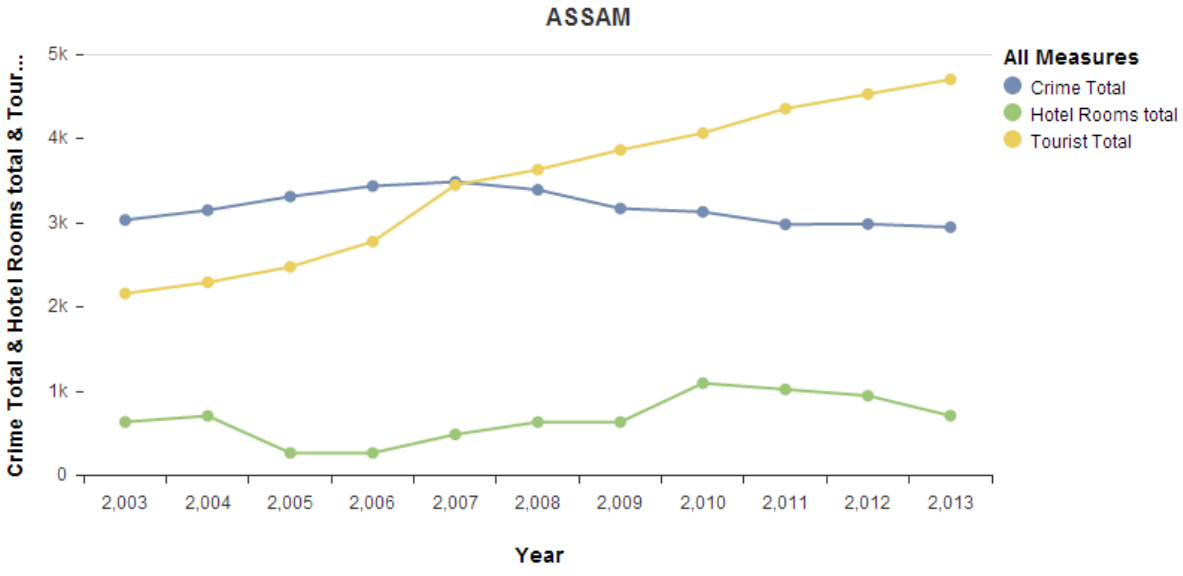
Tourism: The rate of tourism growth has remained steady throughout the 10 year period.
Hotel rooms and Crime: The number of hotel rooms have also stayed steady, with that also contributing to the low crime rate owing to better job opportunities.

Arunachal Pradesh



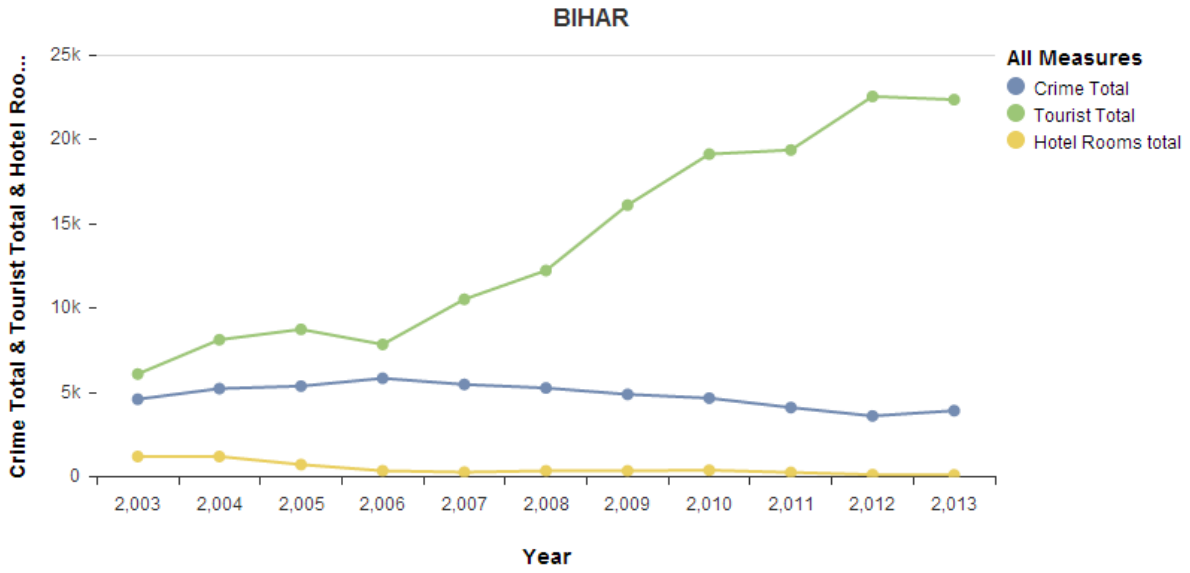
Tourism: has seen major growth from the year 2005 onwards. This is due to the rich biodiversity as well as the geographical location.
Hotel: The number of hotels in the state have seen a steady upward trend as well, keeping in line with the tourism increase as well.

Assam



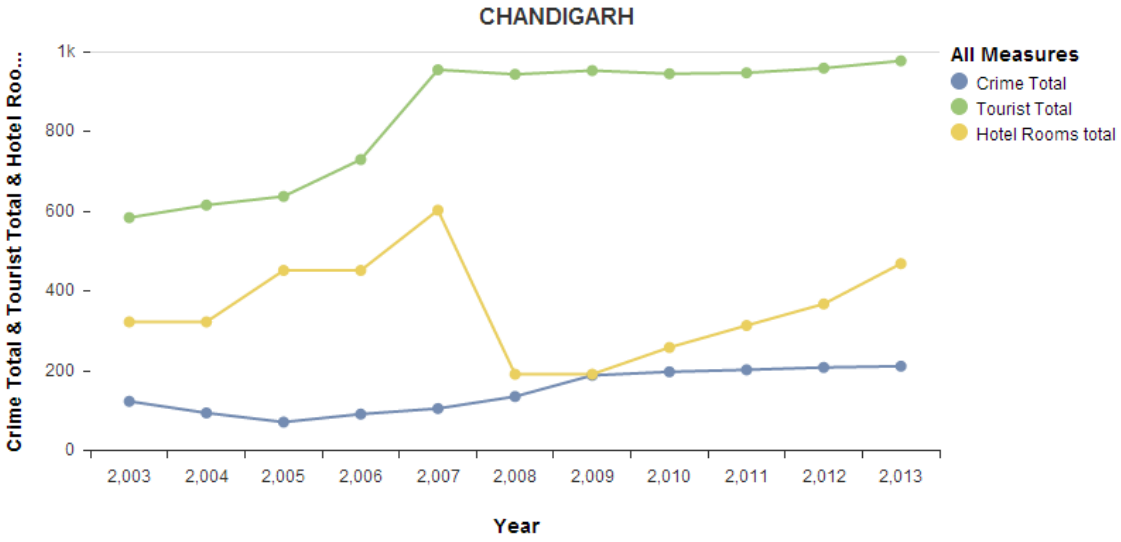
Tourism: Steady rise in the tourism rate throughout the 10 year period.
Crime: Fell with sharp increase in tourism in the year 2007, indicating there to be a relation between the two parameters.
Hotels: Varies around the average with no major spike or change in the graphs.

Bihar (R value = 0.809)



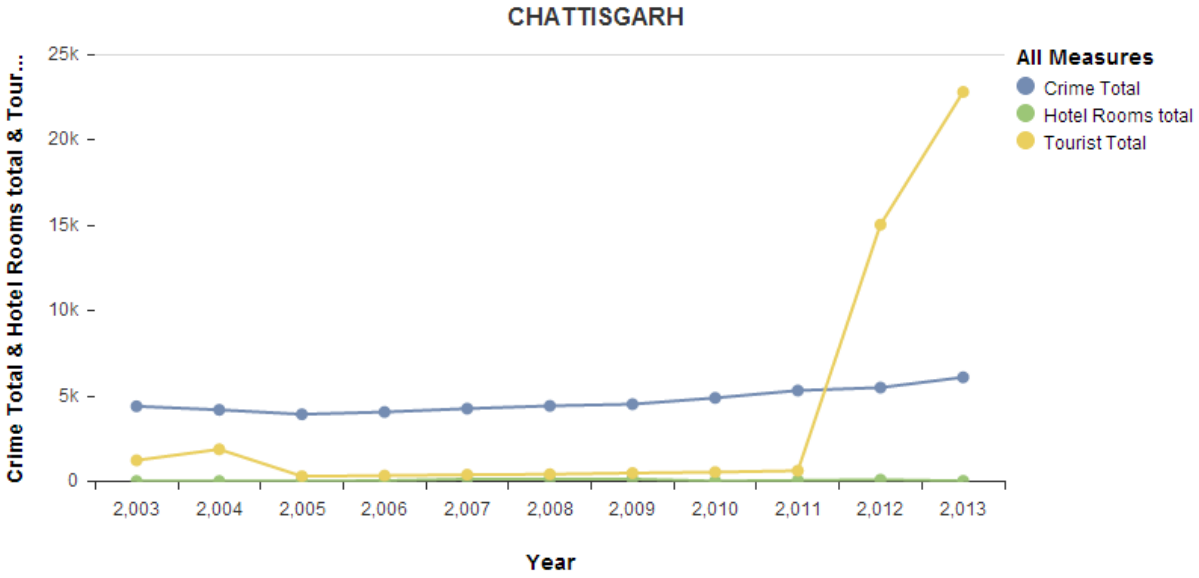
Crime and hotel rooms fairly stable around a particular mark.
The no of tourists has been on the rise since the year 2006 and we observe that from the same year there is a decrease in the total reported crimes as well.

Chandigarh (R value =0.748)



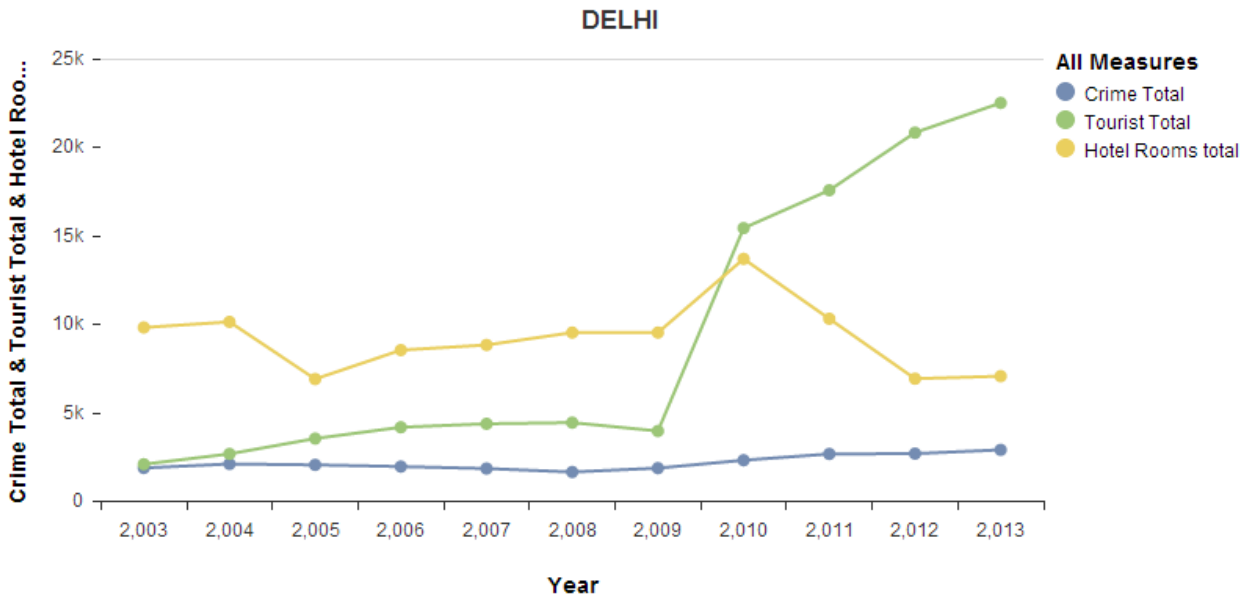
We observe that in mid-2007, there is a fall in the government approved hotels in the city, following which there is no increase in the tourist levels in the subsequent years. The crime rate has remained constant over the 10 years. With only a slight increase in the middle of 2009.

Chhattisgarh (R value = 0.822)



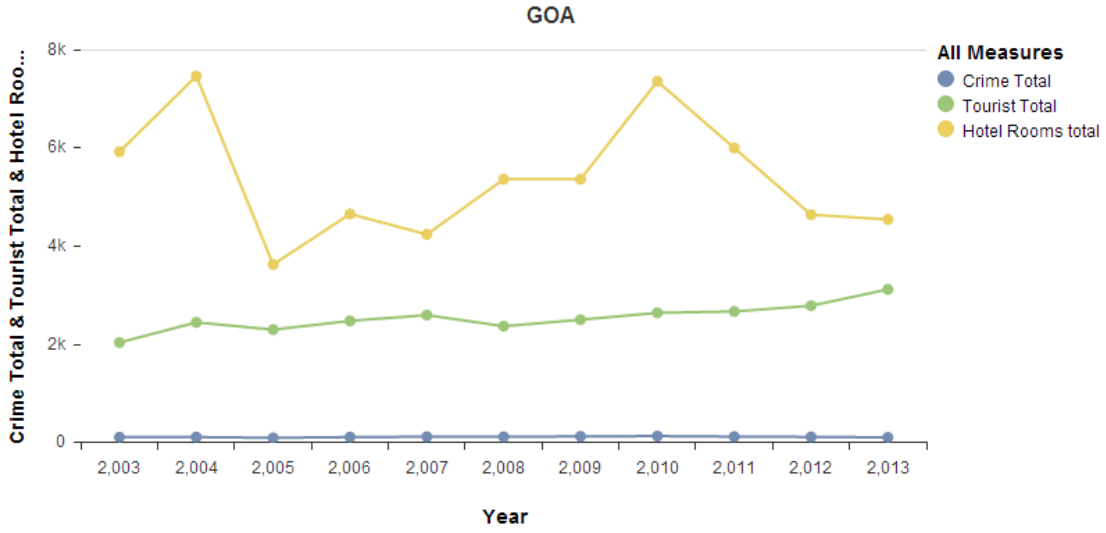
Increase in tourism due to the decrease in Naxalite attacks post 2011.
Crime rate has remained fairly constant mainly due to the low population.
Lack of any hot tourist destinations means that the number of hotels in the state are fairly low.

Delhi (R value = 0.930)



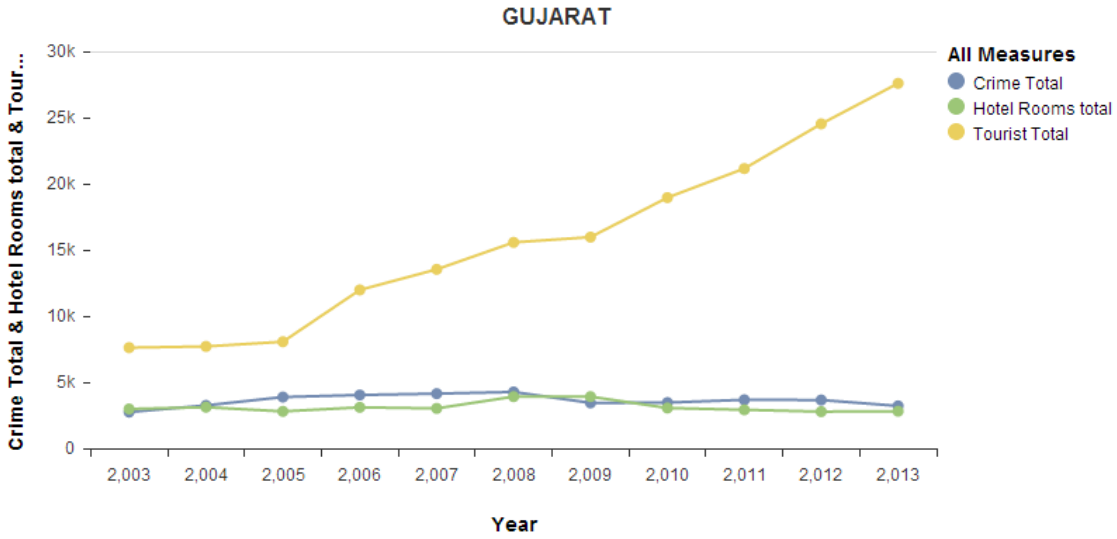
Steep Rise in tourism in 2010 due to the commonwealth games that were held in the city. The development and construction of hotels for the same led to rise in the number of hotel rooms in the city as well.

Goa (R value = 0.820)



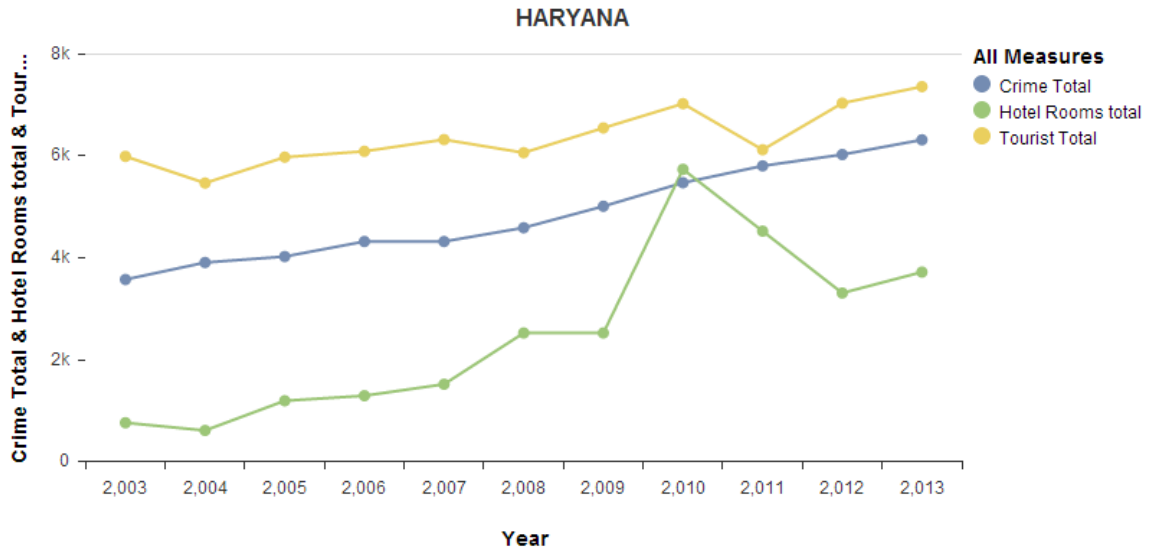
Crime rate is fairly low in Goa owing to the fact that it is a hot tourist destination and availability of sufficient jobs in the state. It has stayed constant in the 10 years of observation.
The fall in **tourism** mid-2004 was due to political turmoil which resulted in the government being dissolved in March 2005 and President's rule being declared in the state. The tourism picked up after that. The second fall in tourism was in mid-2010, largely due to worldwide economic slowdown as Goa has a large number of foreign tourists.

Gujarat



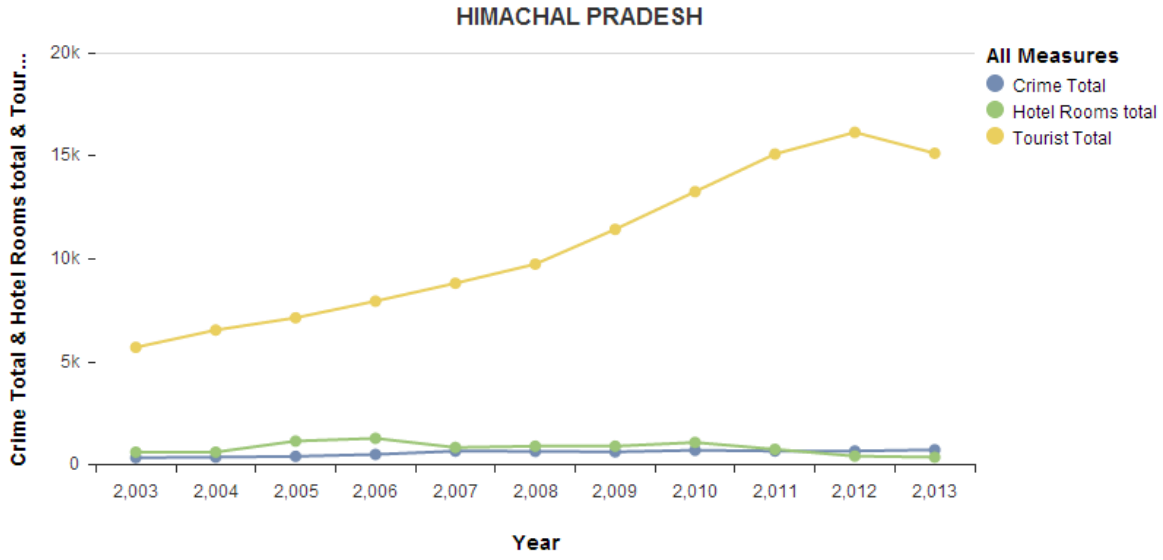
The upward trend in **tourism** was down to the development done in the tenure of Chief Minister Narendra Modi who came to power in 2002.
Crime rate has also been fairly low and there has been no spike in the crime rate throughout Modi's tenure.

Haryana



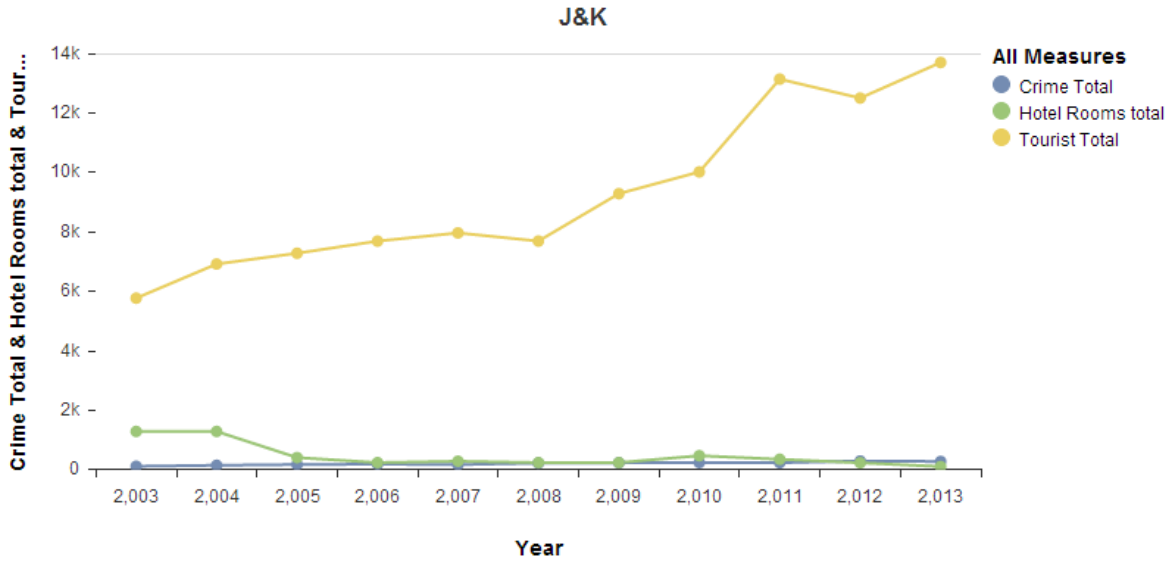
Crime rate has increased each year for the 10 year period, mainly due to lawlessness in this part of the country.
Tourism data is mainly built up by the domestic tourists and even that has progressively increased, along with the number of hotel rooms available.

Himachal Pradesh (R value = 0.843)



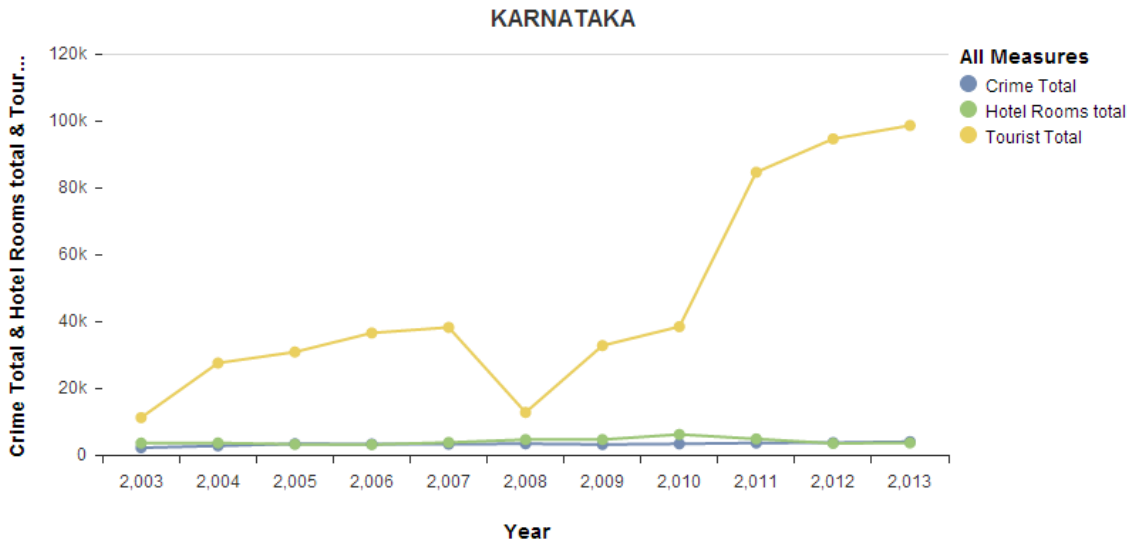
Tourism has seen a constant rise due to the state being a hot spot for tourists.
The difficult geography has resulted in not too many new **hotels** being built and approved by the government.

Jammu & Kashmir (R value = 0.883)



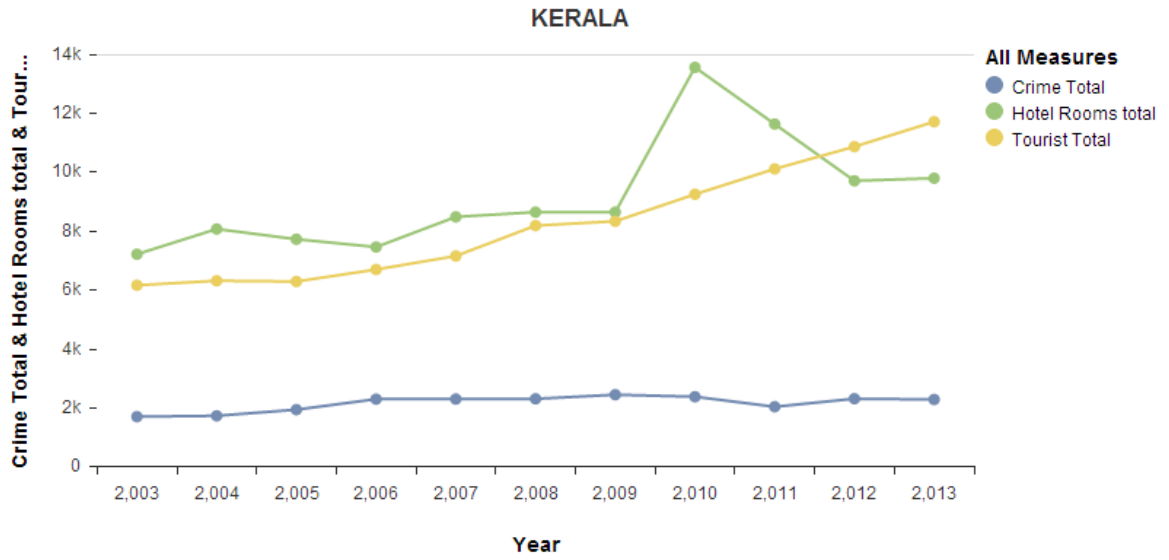
Has seen rise in **tourism** in all years except for a mild decline in mid-2011 to mid-2012, largely due to terrorist threats in the region.
The reported **crime** has been low and constant, as have been the number of **hotel rooms**, which is due to the fact that no person can own a property in the state if he is not a permanent resident of the state.

Karnataka (R value = 0.763)



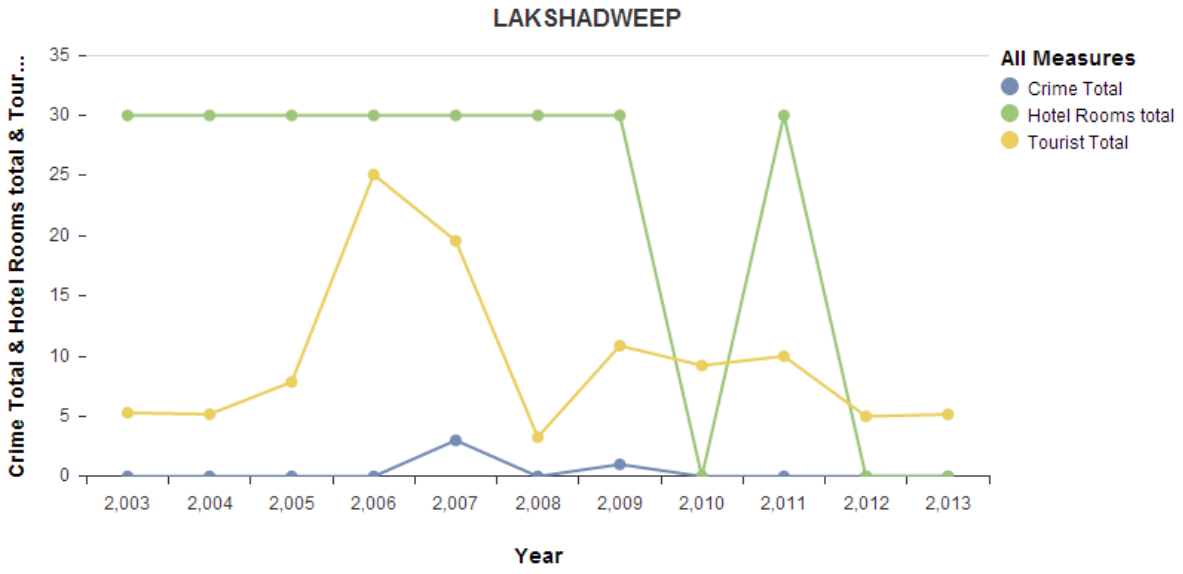
Tourism fall in mid-2007 due to global economic slowdown. That is the only anomaly. **Crime** and **hotel rooms** have stayed constant throughout. Highest literacy rate in the country has led to low crime numbers.

Kerala



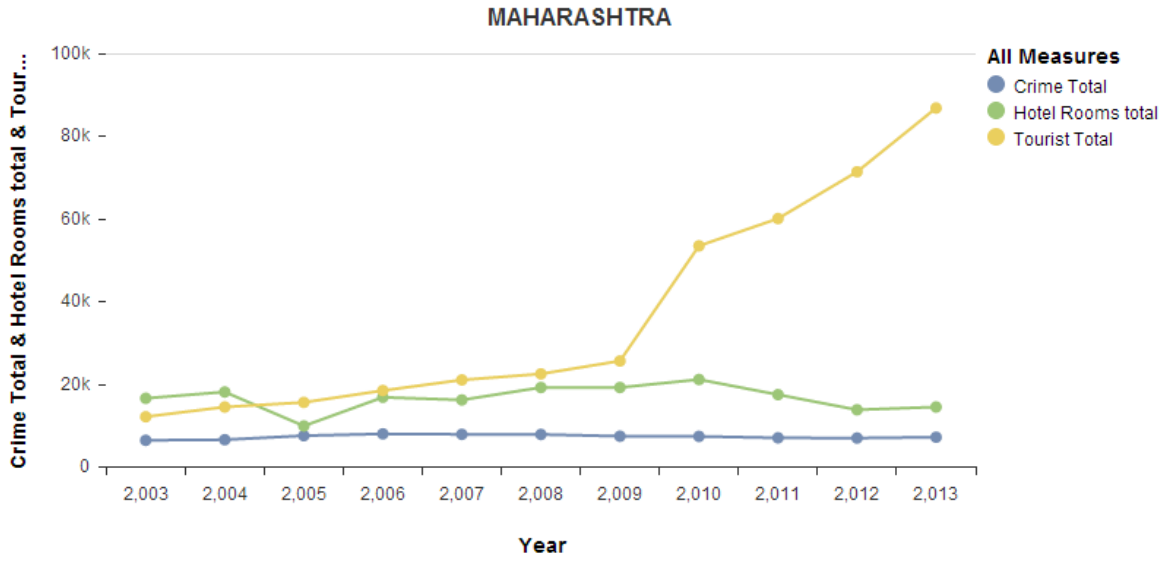
The high literacy rate again leads to low number of **crimes**, which have remained constant throughout.
The state is also a popular tourist destination, both amongst domestic and foreign, hence the constant rise in **tourism**.
The number of government approved **hotel** rooms spiked in mid-2009 after nearly 2 years of stagnation due to the economic slowdown.

Lakshadweep



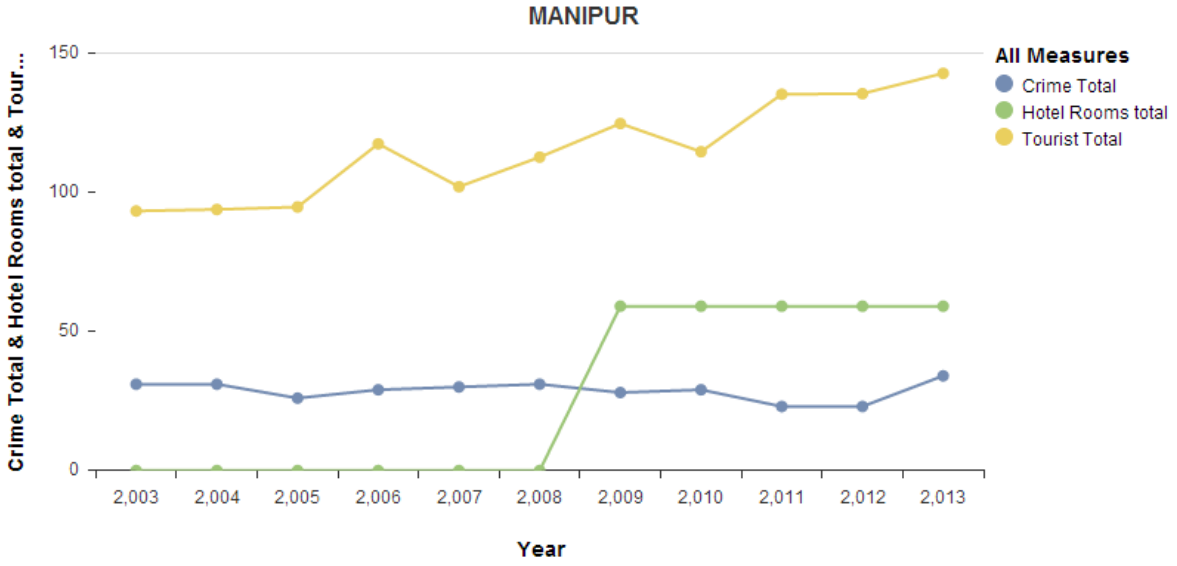
Low population and difficult demographics means that the island has a fairly low **crime** as well as **tourist** numbers. Also the high literacy rate of 92% explains the low crime rate.
The number of **hotels** did pick up post the tsunami of 2004, and they have varied constantly, but on a small scale.

Maharashtra



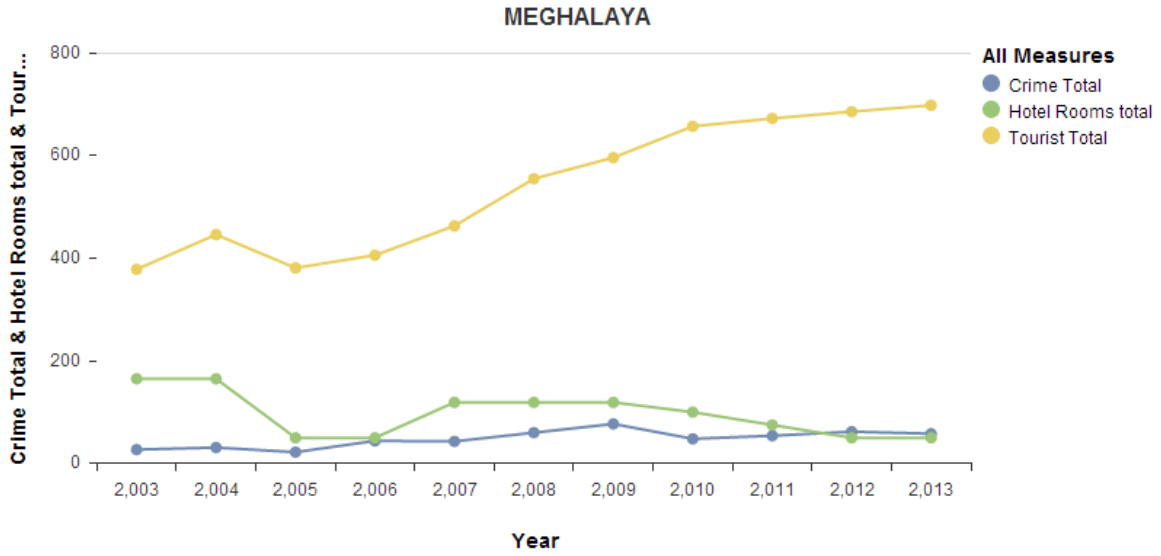
The **tourism** has been constantly on an upward trend mainly due to Mumbai and Pune, the two most famous destinations in the state.
Crime has averaged around 5,000 per month, which is one of the highest in the country. It can be attributed to the large area and population of the state as well.

Manipur



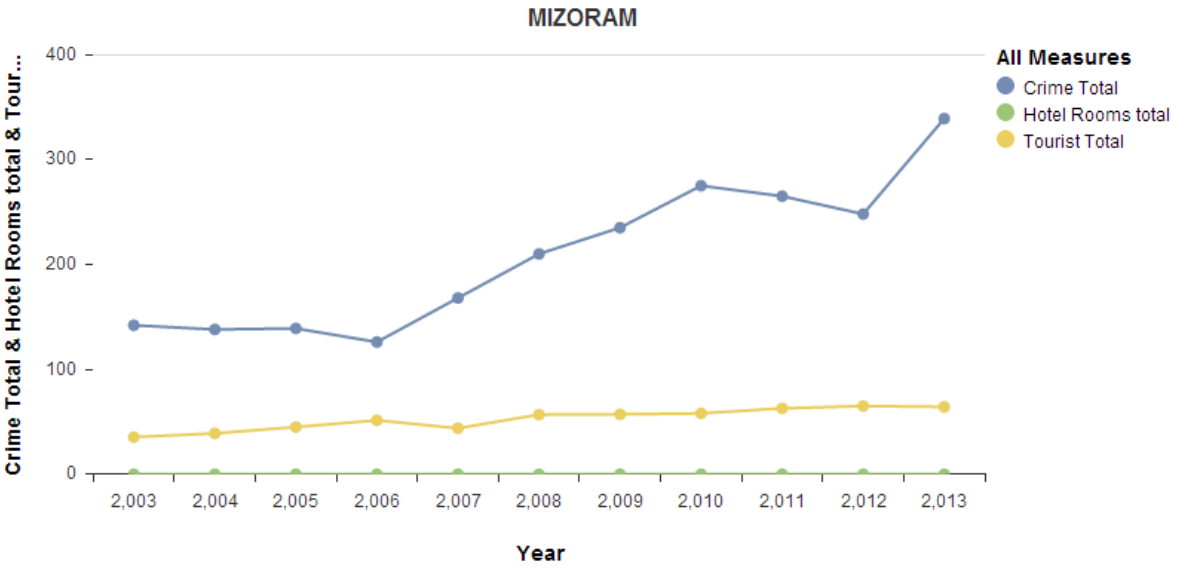
The harsh climate and demographics make it another state with fairly low numbers. **Tourism** keeps varying around a certain mark with no sharp rise or fall. The **hotel** room data also is not available for all the years, but is constant for the years it has been available for. **Crime** is low due to the low density of population, averaging less than 50 reported crimes a year.

Meghalaya (R value = 0.762)



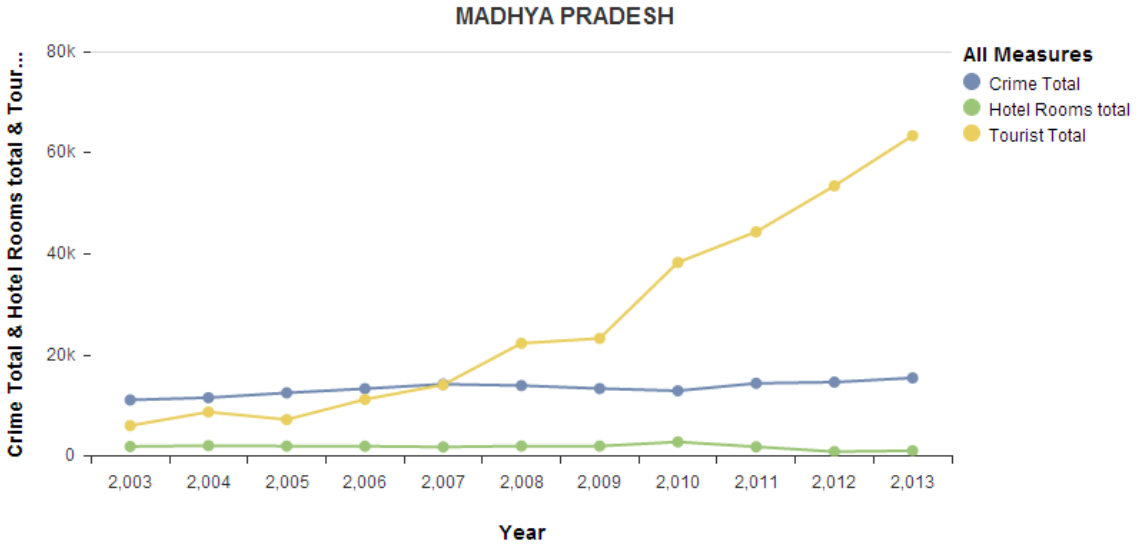
In line with the trend of other north eastern states, Meghalaya also has low population density. Also, due to the complex geography, it is not a major tourism destination. It also explains the low number of hotel rooms available in the state.

Mizoram (R value = 0.842)



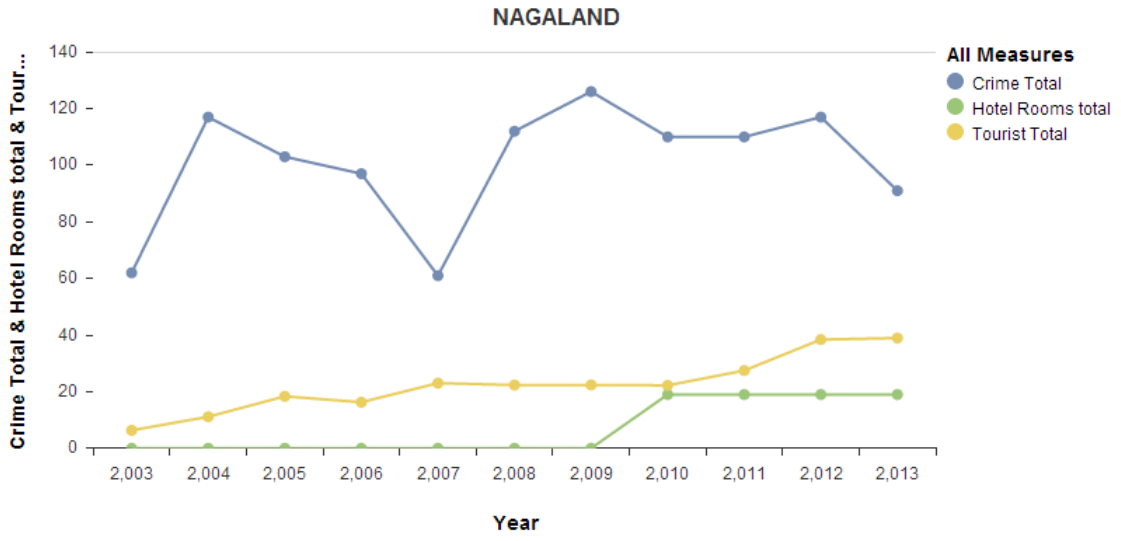
There are no government approved hotels listed in the state.
The **tourist** numbers are in the low 30's and 40's again due to the difficult geography. **Crime** rate in this instance has increased on a year-on-year basis mainly due to lesser availability of jobs and also because it shares its borders with neighboring countries as well.

Madhya Pradesh (R value = 0.777)



The state with the highest crime numbers in the country. It is mainly due to the fact that it is one of the least developed state in the country with an HDI (Human Development Index) value of 0.375. The majority of population is tribal which explains the **crime** rate of the state.
Tourism: The state is home to a lot of medieval and ancient history monuments, hence the high number of tourists visiting the state each year. In the years following 2011, there has been a sharp increase in the tourism numbers due to the promotional activities done by the Ministry of Tourism for the state.

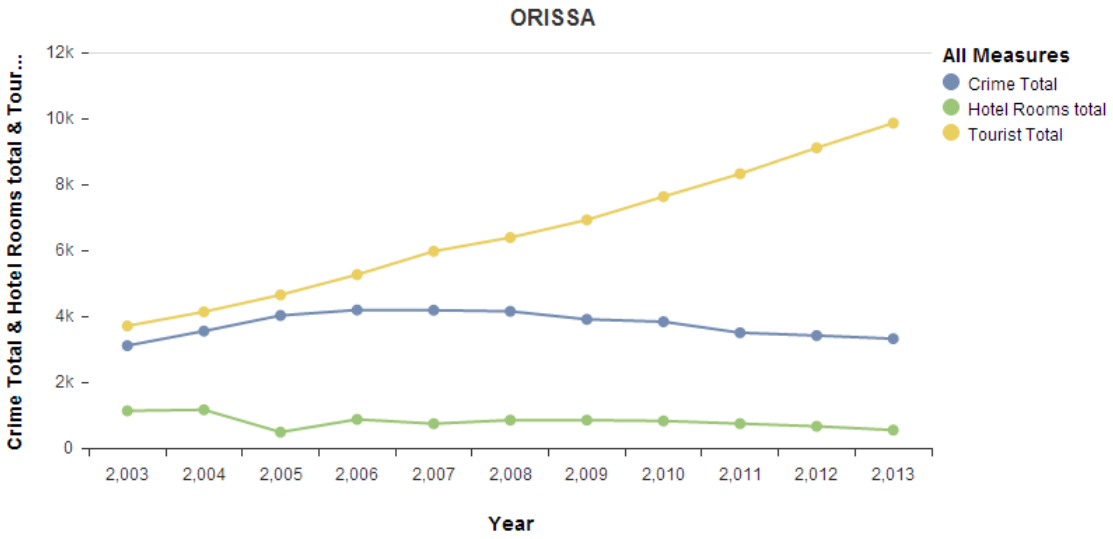
Nagaland



Tourism: difficult geography. Isolated region.
Crime: lack of tourism or proper jobs available leading to higher crime rate than normal for this size of population.
Increase in tourism from 2003 to 2007 lead to a decline in the crime numbers.

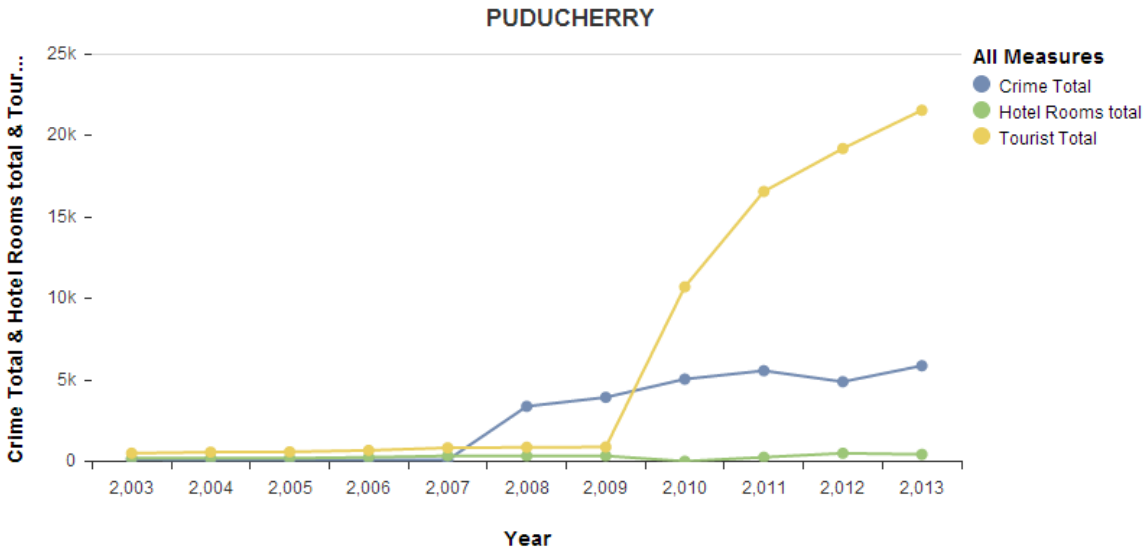
Hotels: There were no government approved hotel rooms available till 2009, but a few hotels have been added since which has corresponded to another fall in crime rate.

Orissa



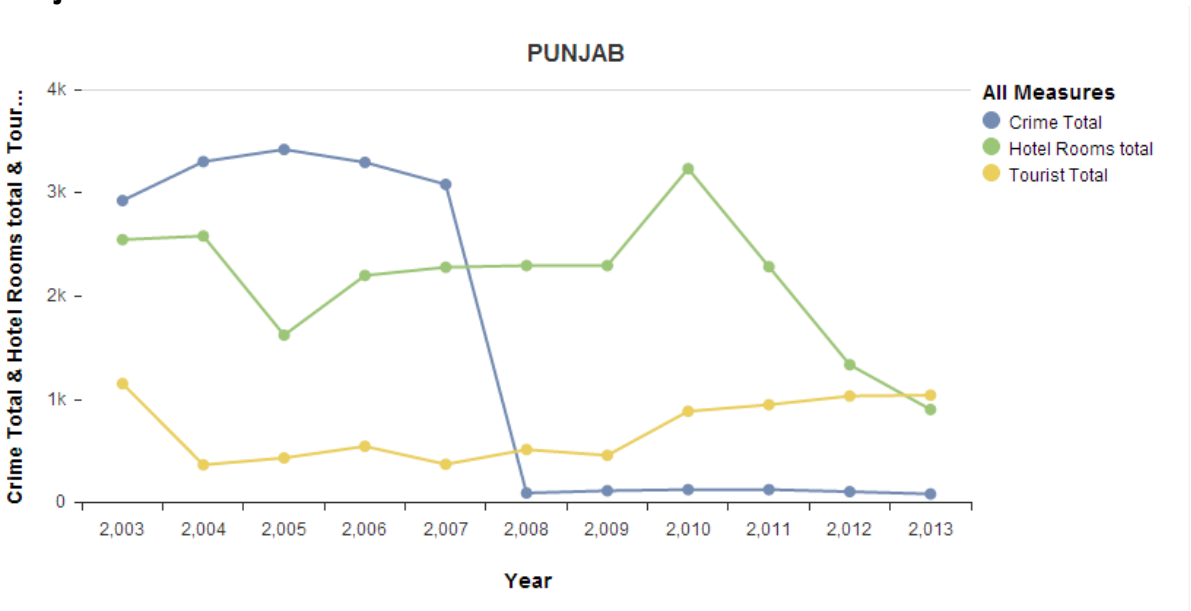
Tourism: Linear increase in tourism.
Crime: Decline in crime rate with the increase in the number of tourists in the state post 2006.
Hotels: The crime rate increased slightly from 2004 to 2006 when there was a decline in the number of hotel rooms that were available.

Puducherry (R value = 0.823)



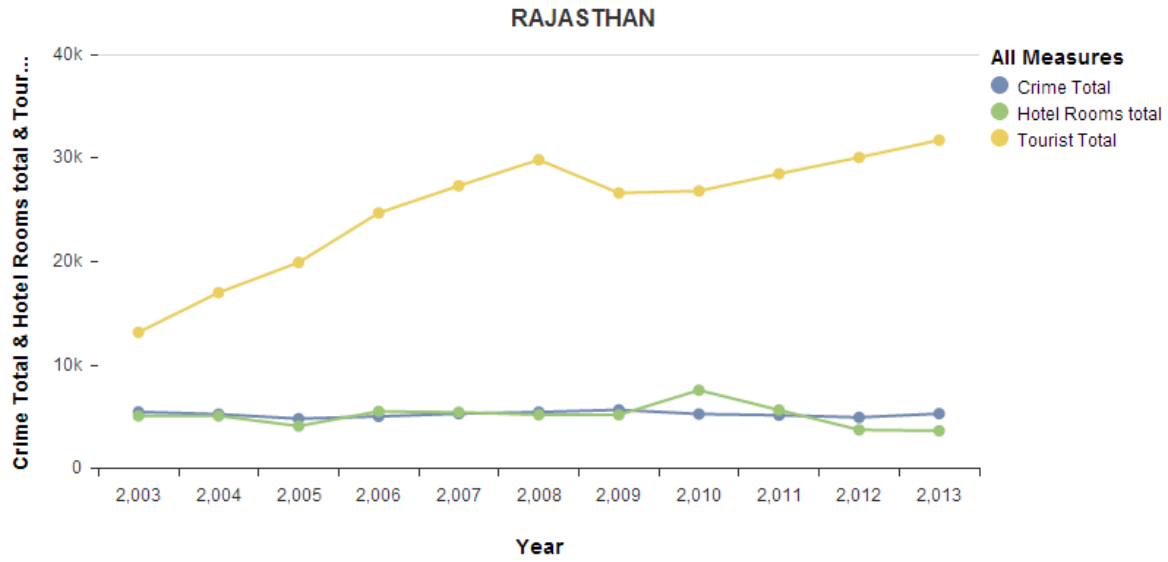
Crime rate increased in the years post 2007 due to the near negligible number of **hotel** rooms available or tourists visiting the state. Small geographical area is the cause of the low numbers.

Punjab



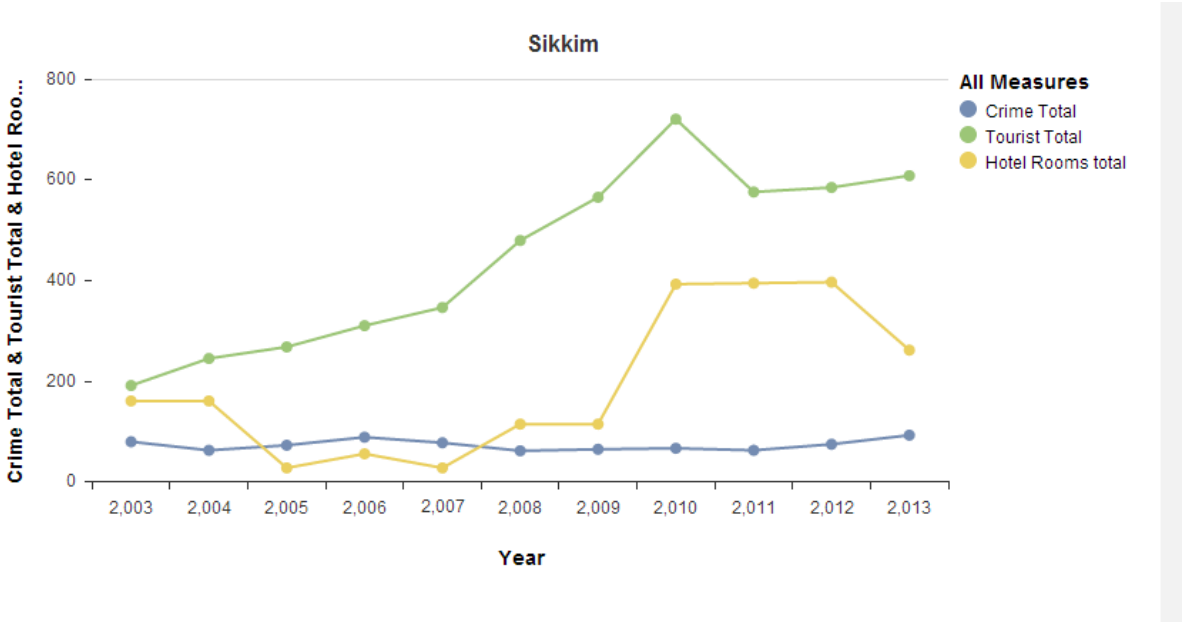
With increase in number of **hotel** rooms in the state along with the stabilization of the number of **tourists**, the **crime** rate fell sharply in mid-2008.

Rajasthan



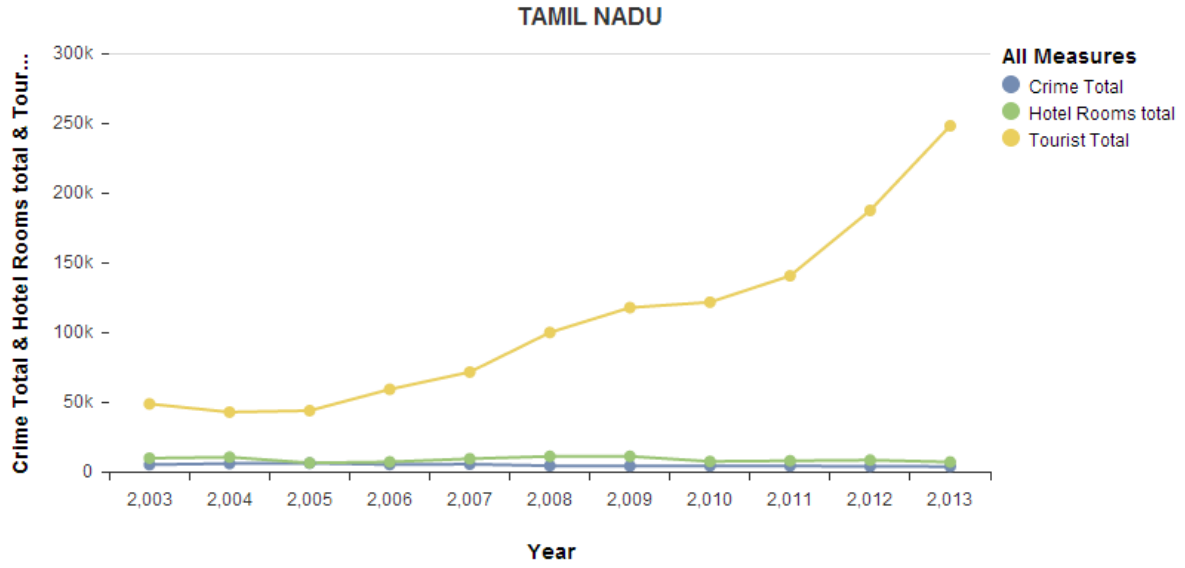
Crime has remained constant over the 10 year period, with the only effect observed in the year 2010, where a slight increase in the number of **hotel** rooms available led to a slight fall in the crime rate.
The number of **tourists** has increased each year except for 2008, again attributed to the economic slowdown.

Sikkim



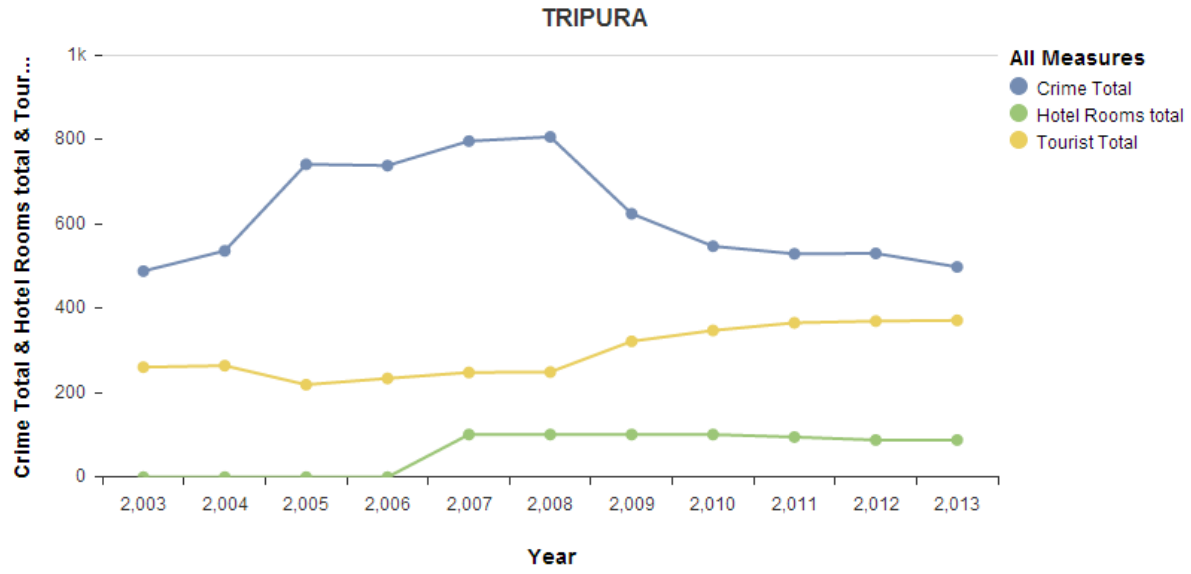
With the tourism peaking in the year 2010, the number of hotel rooms also saw a sharp rise.
The crime rate decreased in 2008 when the number of hotel rooms increased and has remained constant ever since.

Tamil Nadu (R value = 0.858)



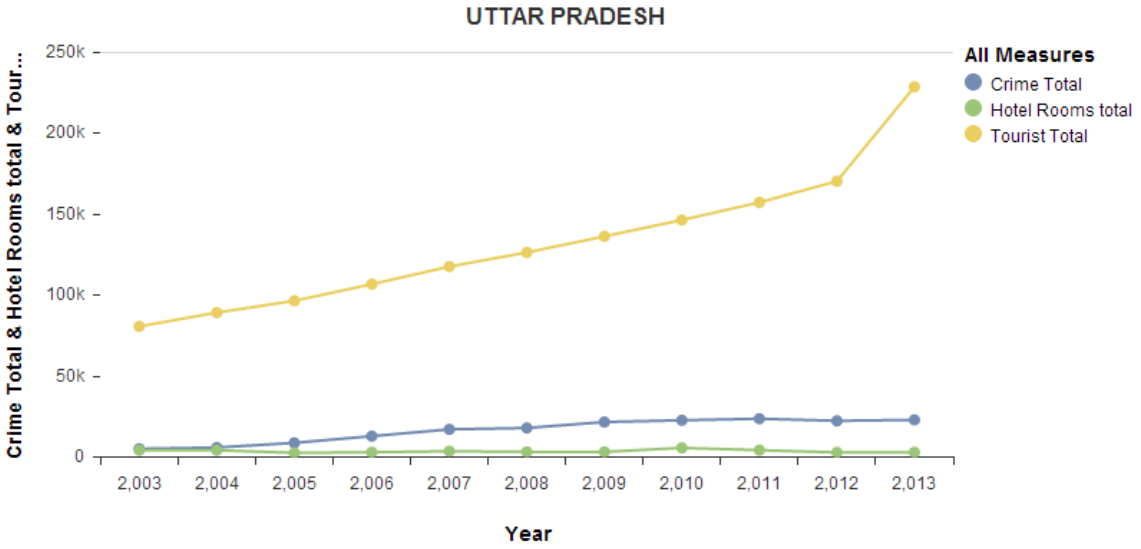
The high tourism rate has kept the crime rate very low in the state. In 2008 and 09 when the hotel rooms increased, the crime numbers went even lower.

Tripura (R value = 0.701)



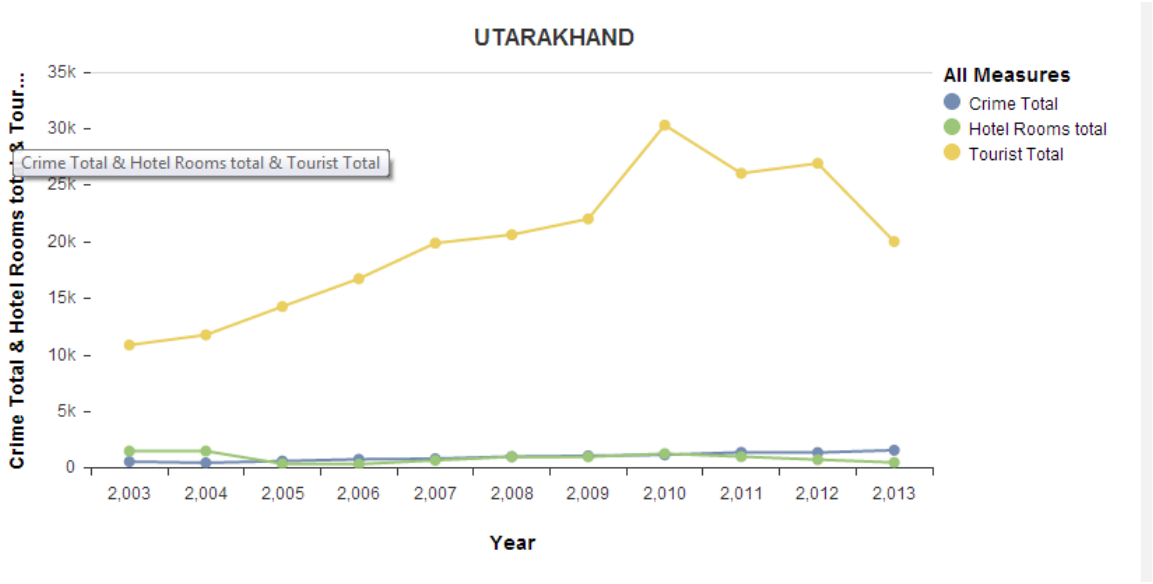
With the construction of hotels in 2007 and 08, the crime rate started declining, which again led to an increase in the tourism of the state.

Uttar Pradesh (R value = 0.828)



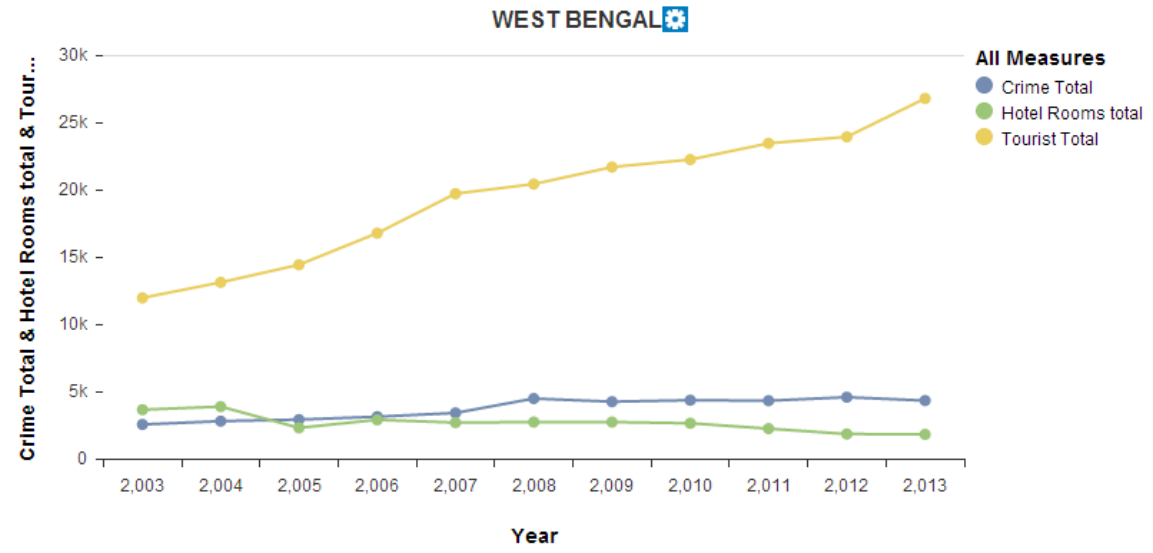
The data of the state is an outlier with the crime rate unaffected by the increase in tourism or even vice versa. The majority of tourists are for Taj Mahal while the prevalent lawlessness in this part of the country means that there is no decrease in crime rate observed.

Uttarakhand (R value = 0.789)



In 2010, when the number of **tourists** declined, the **crime** rate also increased, with no increase in the number of **hotel** rooms.

West Bengal (R value = 0.922)

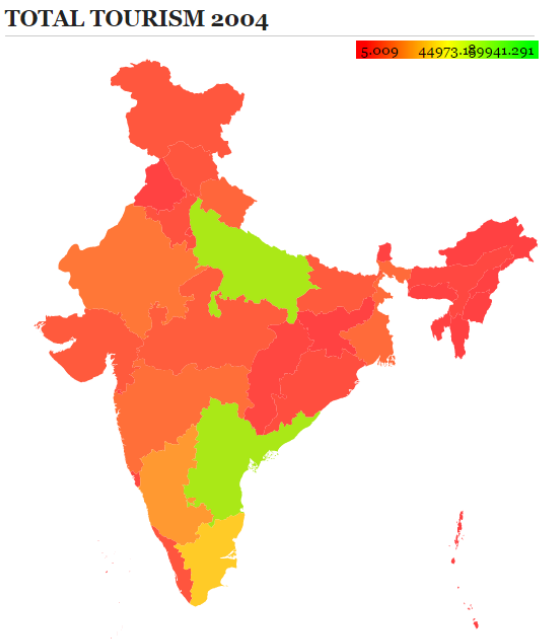


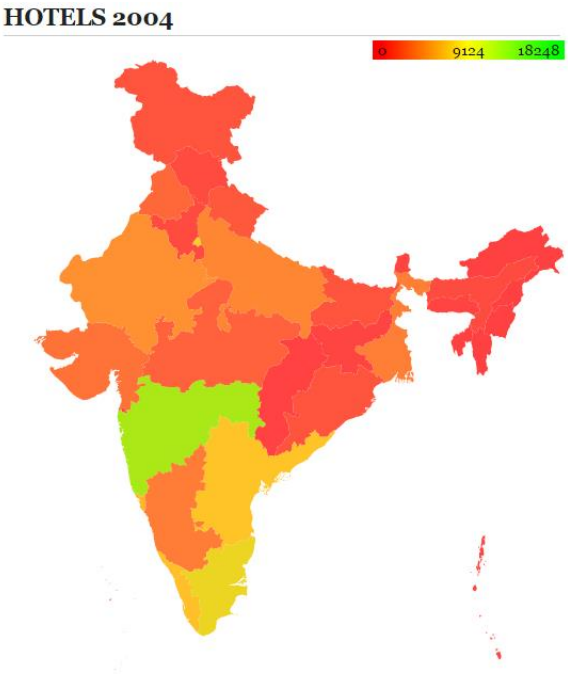
After being nearly equal in 2006, 2007 saw a fall in the no of **hotel** rooms available, with it leading to a subsequent increase in the **crime** rate. Although it did not deter the **tourist** rate in the state at all.

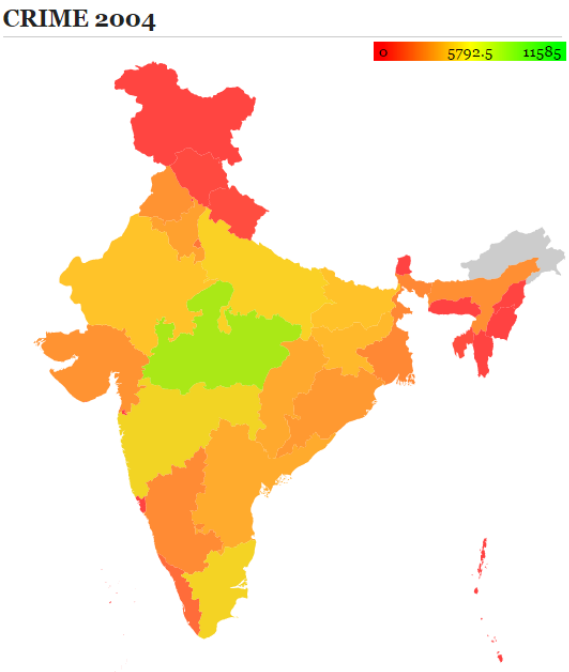
Year-wise (2003-2013) color coding for each of the 3 attributes (tourism, crime & hotel rooms) on the map of India

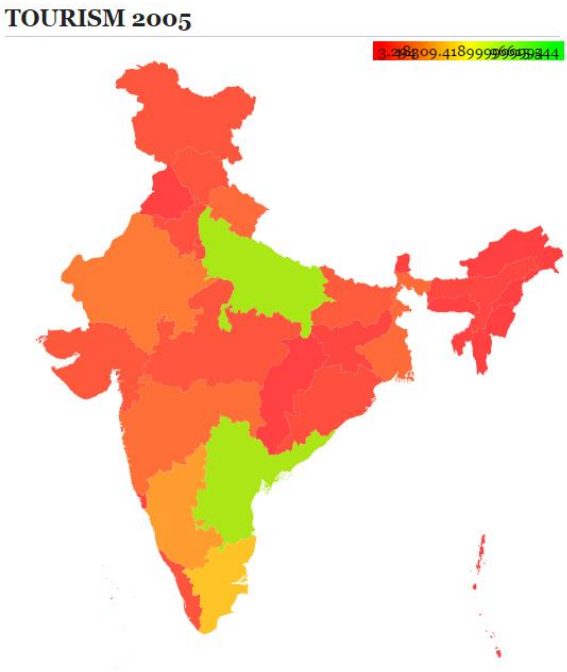
We also used state-wise color coding in order to show the variation and the maxima and minima for the 3 attributes in a visually appealing manner. Each color coded

map is shown below, with green, yellow and red showing the maxima, median and minima respectively for each attribute in each tear from 2003-2013.

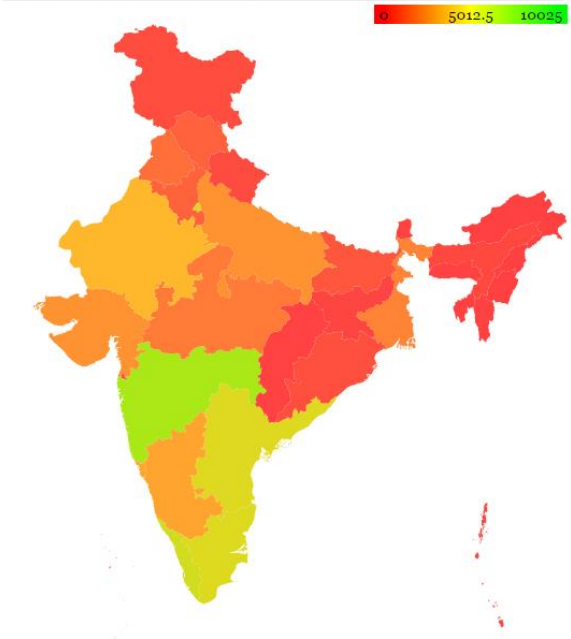




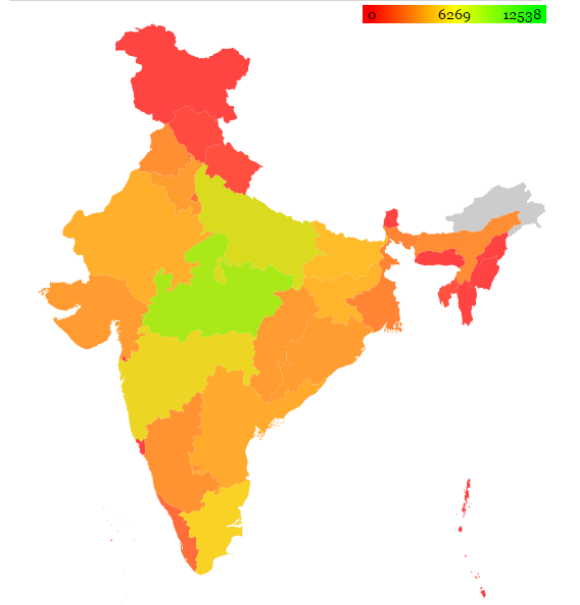




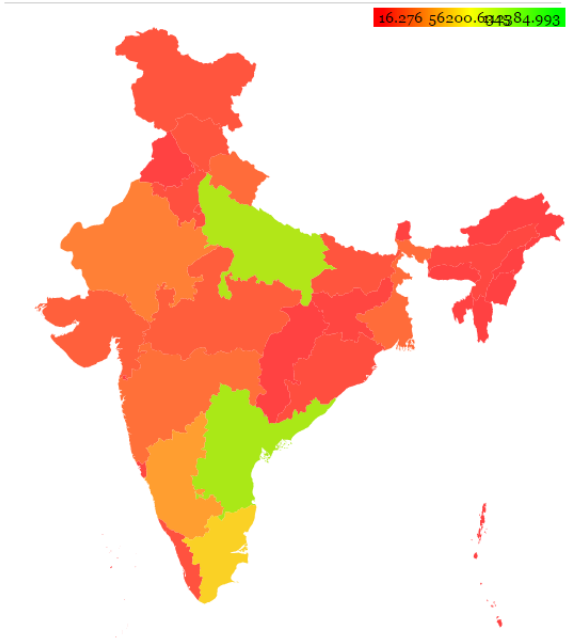
HOTEL ROOMS 2005



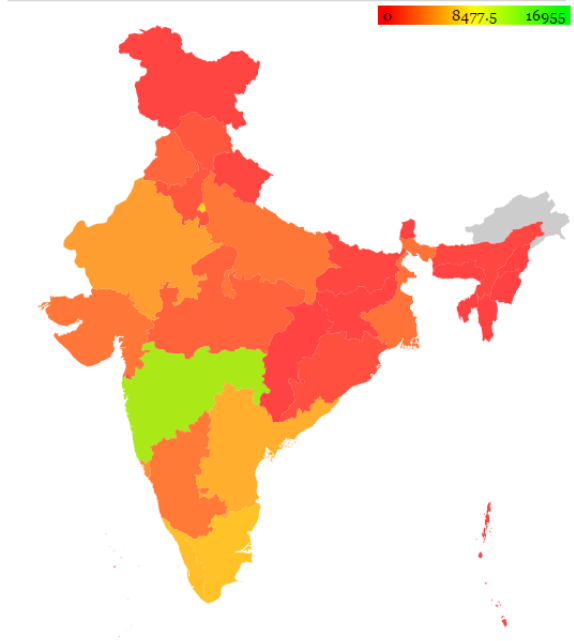
CRIMES 2005



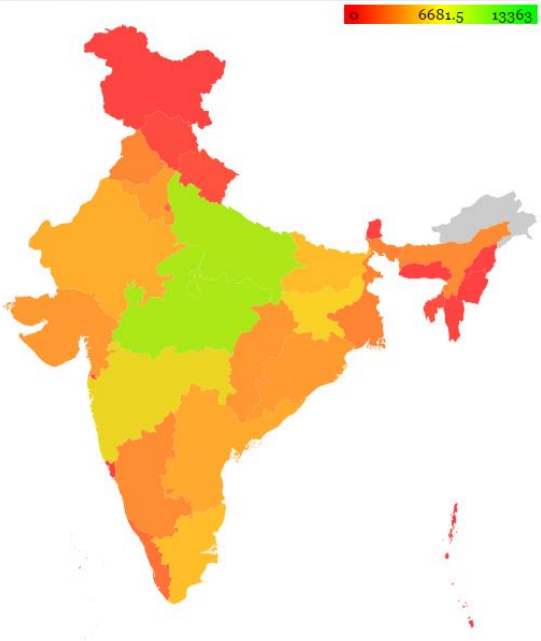
TOURISM 2006



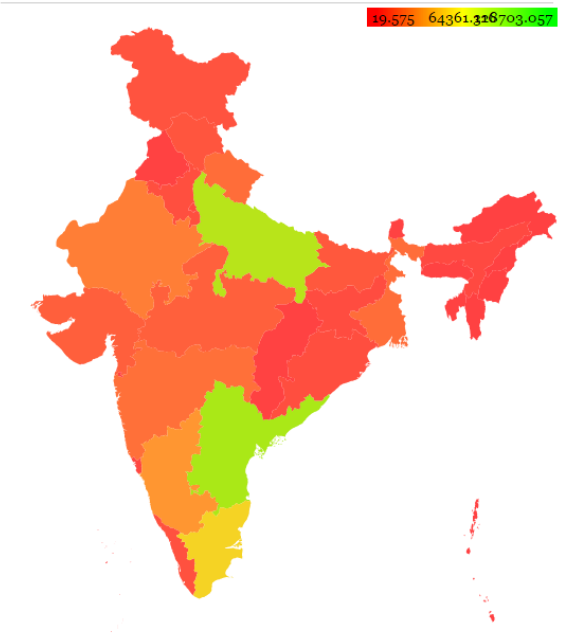
HOTEL ROOMS 2006



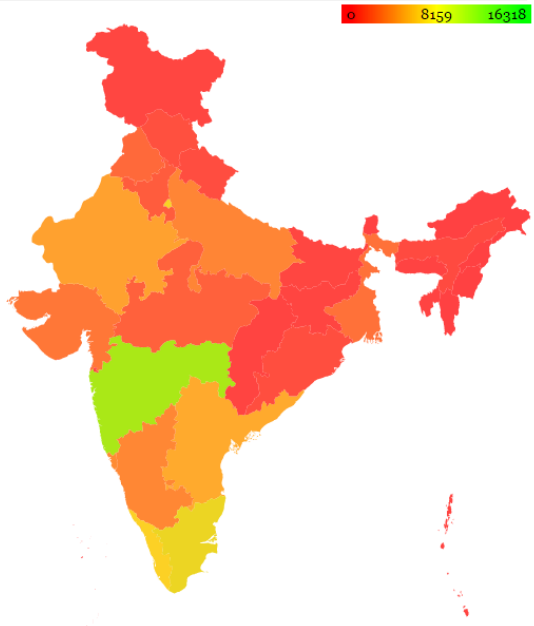
CRIMES 2006



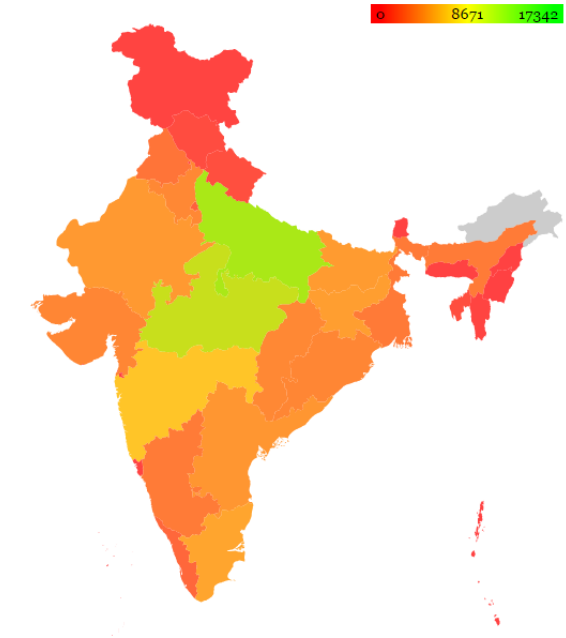
TOURISM 2007



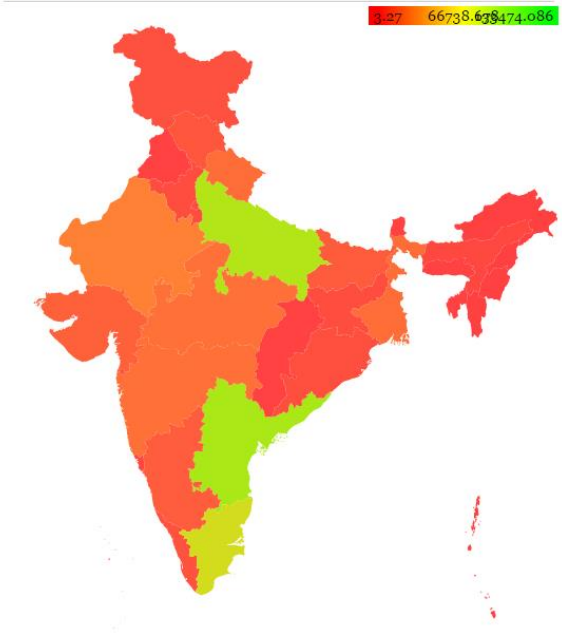
HOTELS 2007



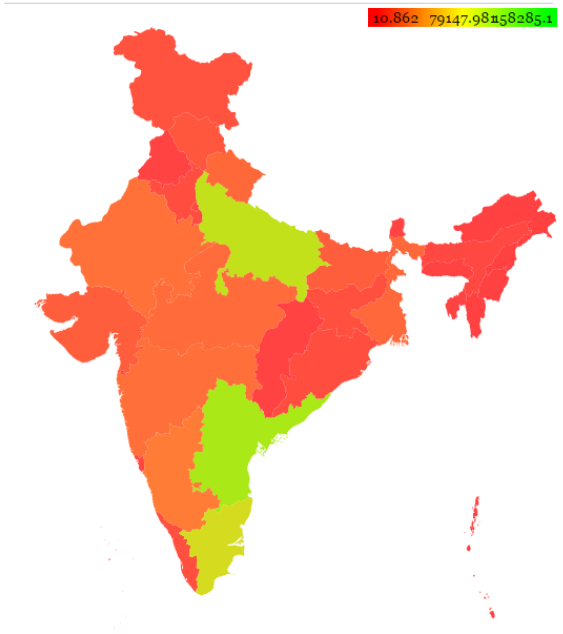
CRIMES 2007



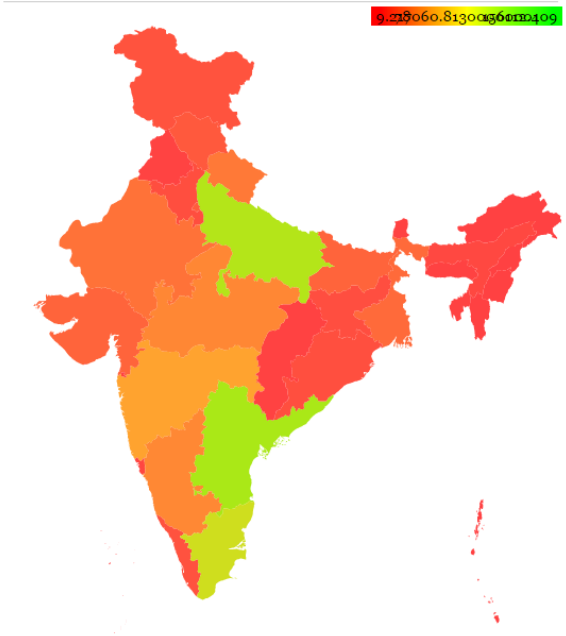
TOURISM 2008



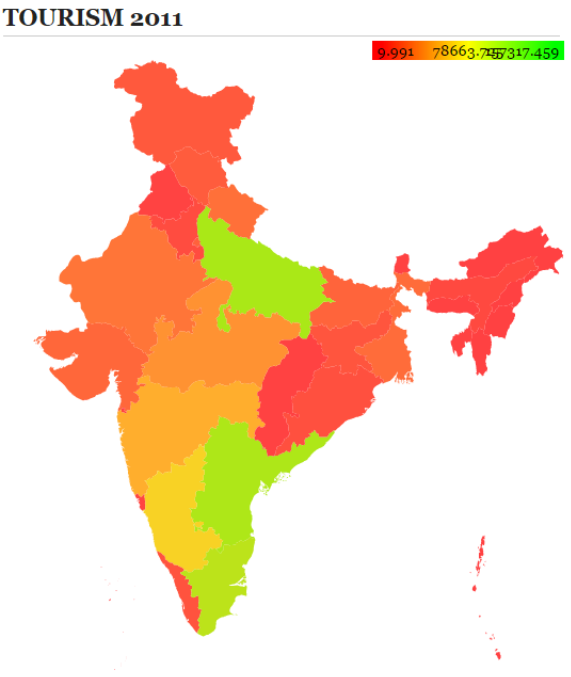
TOURISM 2009



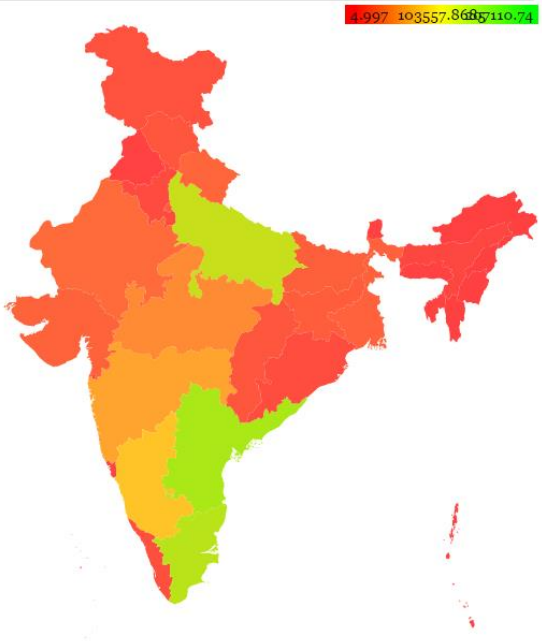
TOURISM 2010



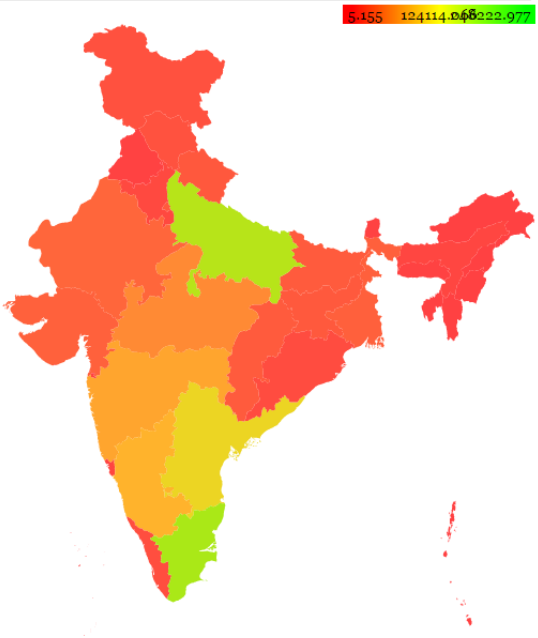
TOURISM 2011



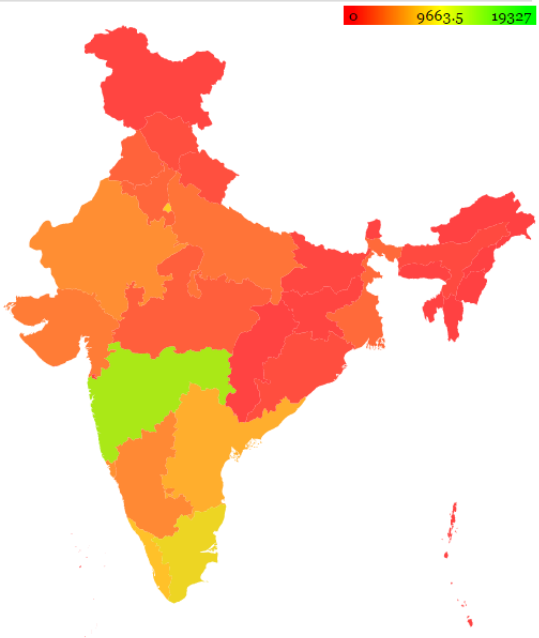
TOURISM 2012



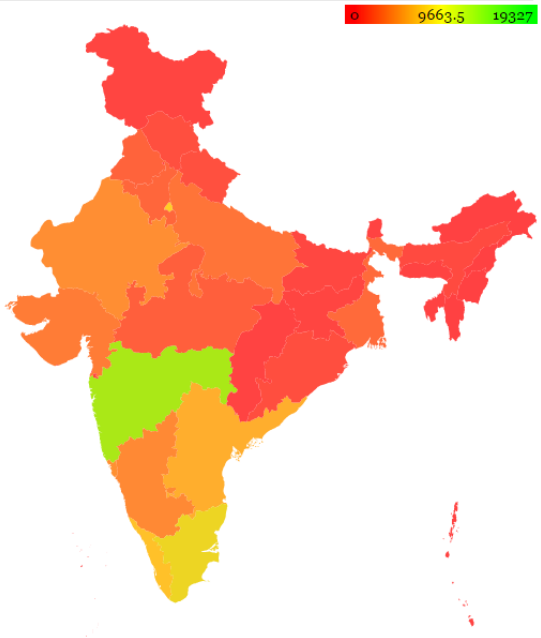
TOURISM 2013



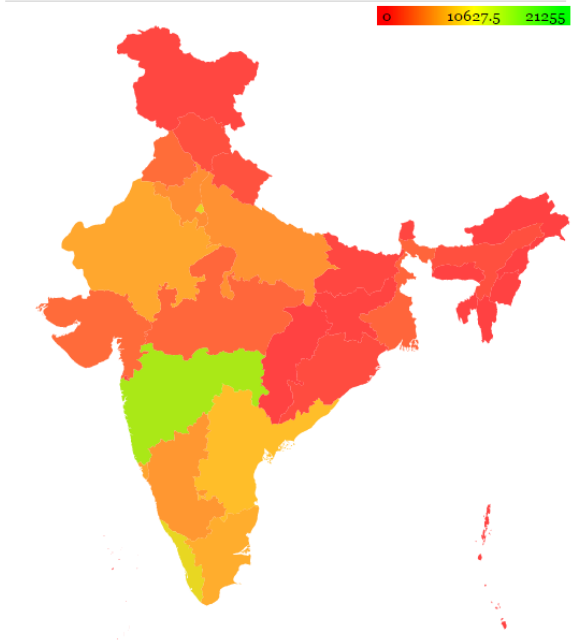
HOTEL ROOMS 2008



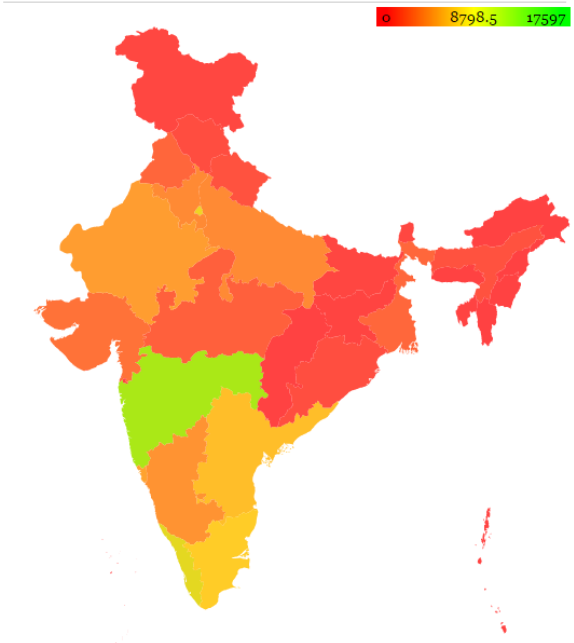
HOTEL ROOMS 2009



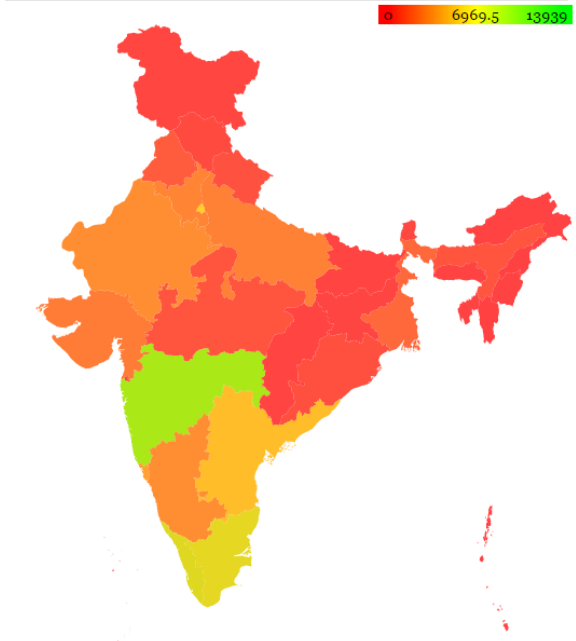
HOTEL ROOMS 2010



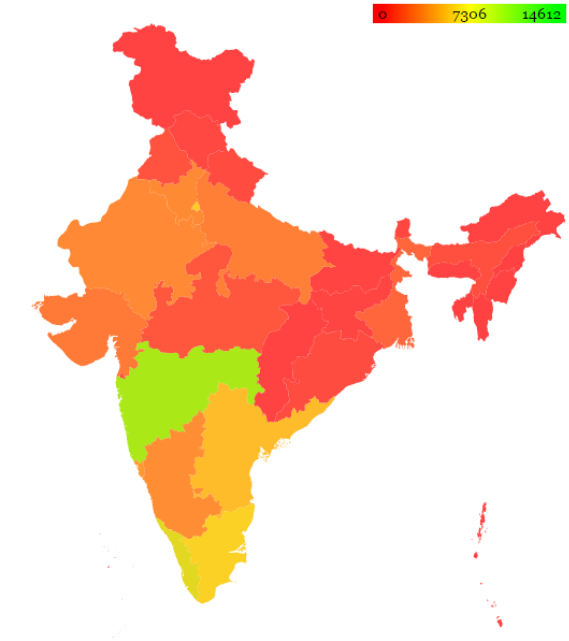
HOTEL ROOMS 2011



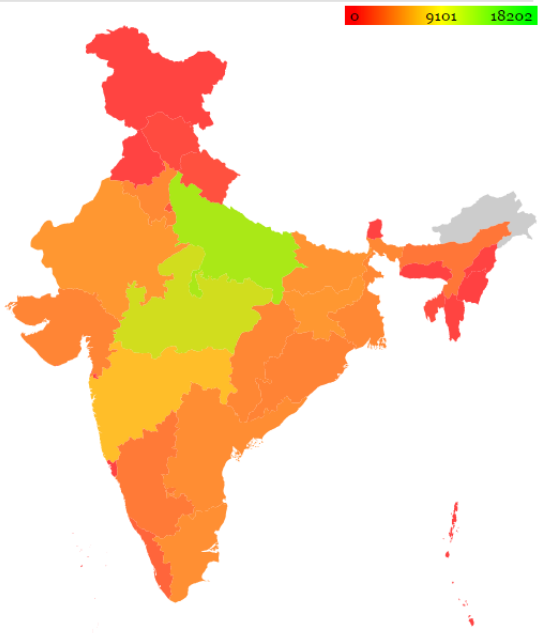
HOTEL ROOMS 2012



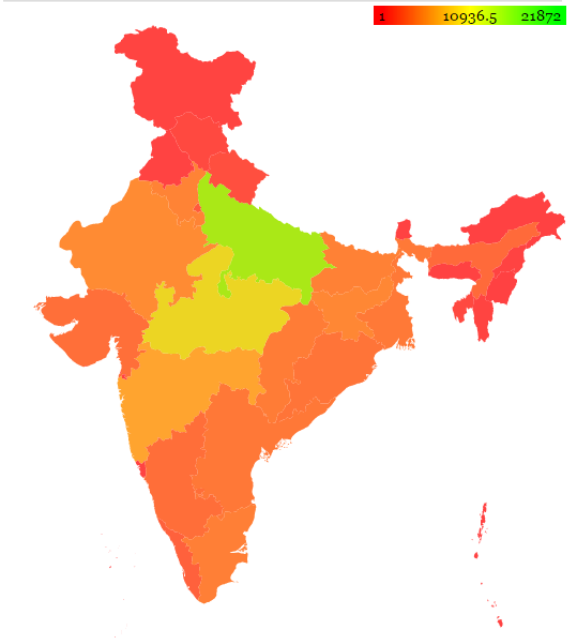
HOTEL ROOMS 2013



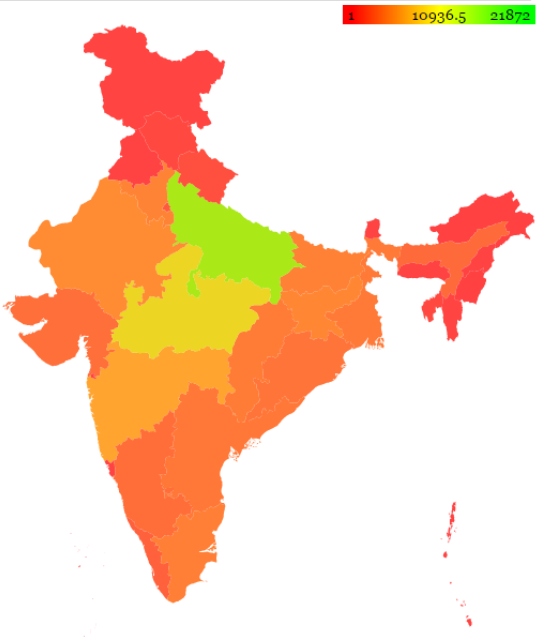
CRIME 2008



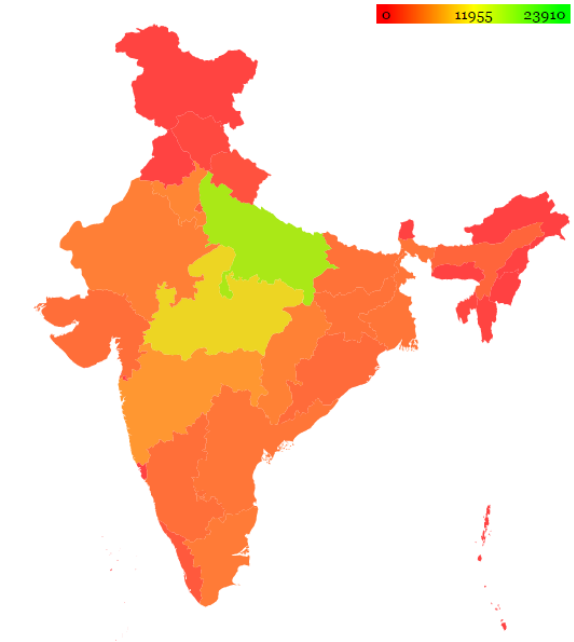
CRIME 2009



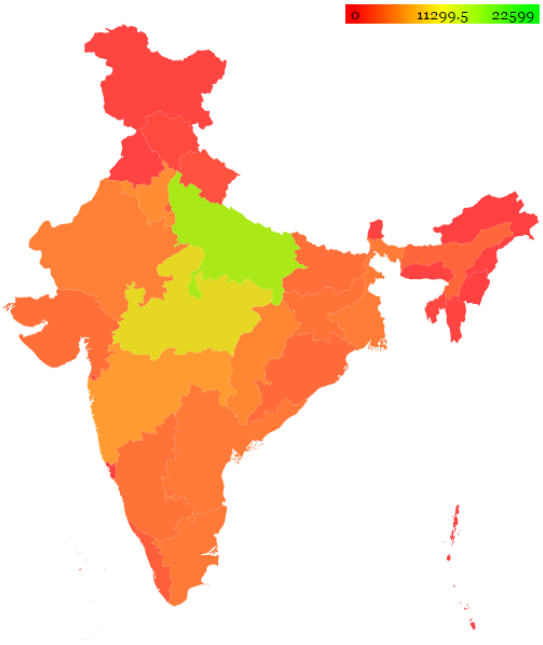
CRIME 2010



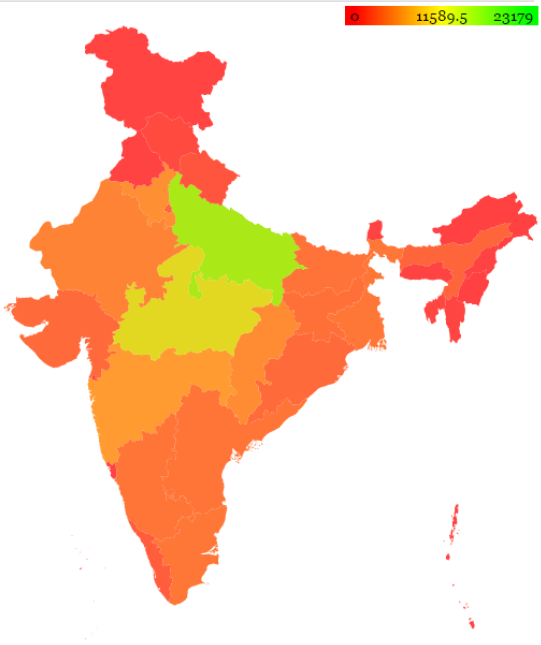
CRIME 2011



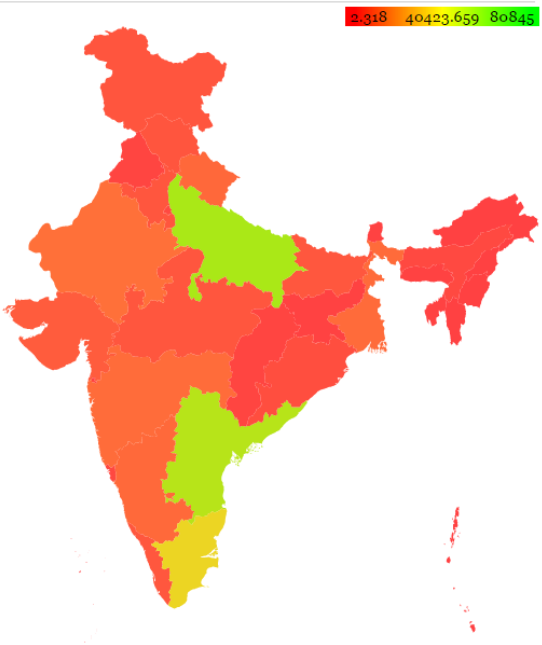
CRIME 2012



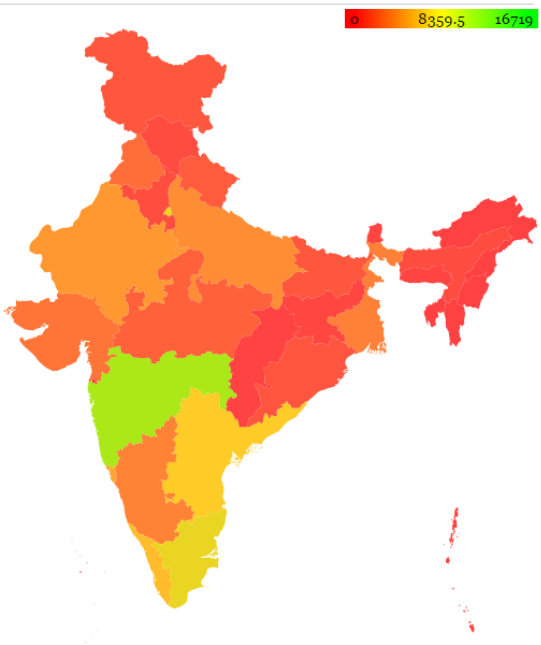
CRIME 2013

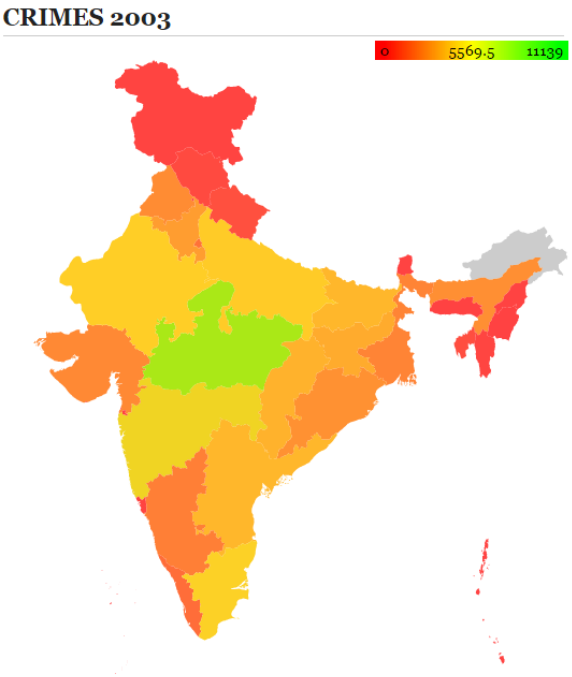


TOTAL TOURISM 2003



HOTEL ROOMS 2003





SENTIMENT ANALYSIS OF TWITTER DATA

There was a phenomenal rise observed in the number of tourists in Delhi after 2009. We also knew that people generally like the weather of Delhi and hence were curious to know if there is a relation between the weather and tourism. Twitter is a popular medium where people express their feelings. We searched for tweets on the twitter API using the phrases 'delhi' and 'weather' and found numerous tweets. But, due to the restriction of Twitter API, we had to set a limit to the tweets per year that we would be tracking from 2009 to 2013. Hence, we obtained 100 tweets per year for performing sentiment analysis.

We used Python along with Natural Language Toolkit (NLTK) 3.0 to model our classifier and perform sentiment analysis. First, 55 random tweets were manually classified into positive and negative as shown below.

```
pos_tweets = [

('Loving the weather in Delhi today...Its just amazing!','positive'),

('Awesome weather in delhi. Just got back from a walk','positive'),

('Really great weather in delhi! Going home with the window rolled down and the
hair haywire \m/','positive'),

……

('Yep. Lovely weather in Delhi region today.','positive'),

('Wonderful weather in Delhi today. Nice & cool, with a slight breeze. Zero
humidity. Outside, its much worse though.','positive'),

('Its such a *beautiful* weather in Delhi. BEAUTIFUL!','positive')

]
```

```
neg_tweets = [

('bad weather in delhi. Internet dead. Glad i\'m home on time.','negative'),

('Landed safely in terrible weather in Delhi Now stuck in bad jam http://bit.ly/dcYXu http://yfrog.com/ccj02j','negative'),

('So now the flight lands in Nagpur..bad weather in Delhi ....c\'est la vie','negative'),

……

 ('@PritishNandy Never really liked weather\'s of Delhi, Mumbai and Kolkata..everything is way to extreme !!','negative'),

('@Sharanya yeah peach tree. bt the weather in delhi is nt conducive to its growth :\\','negative')

]


#combining the tweets and keeping only words that are longer than 2 letters.

tweets = []

for (words,sentiment) in pos_tweets + neg_tweets:

      words_filtered = [e.lower() for e in words.split() if len(e)>=3]

      tweets.append((words_filtered,sentiment))
```

Then the features of the training set were applied to the classifier using **`classify.apply_features`** function of NLTK. Finally, the classifier was trained using the **`NaiveBaiyesClassifier.train`** function of the NLTK as shown below.

```
training_set = nltk.classify.apply_features(extract_features,tweets)

classifier = nltk.NaiveBayesClassifier.train(training_set)

print classifier.show_most_informative_features(32)

test_tweets = []

with open('TWEETS.csv','rb') as csvfile:

        linereader = csv.reader(csvfile)

        for row in linereader:

                test_tweets.append(row)


result_tweets = []

for tweet in test_tweets:

        print tweet,

classifier.classify (extract_features(tweet[0].split()))
```

```
        result_tweets.append((tweet,
classifier.classify(extract_features(tweet[0].split()))))


with open('sentiment_results.csv','wb') as csvfile:
     csvwriter = csv.writer(csvfile)
     for (tweet,sentiment) in result_tweets:
          csvwriter.writerow([tweet,sentiment])
```

The complete code along with comments can be found in Bibliography.

A snapshot of the file sentiment_results.csv is given below.

Table 1 Snapshot of file sentiment_results.csv

['What weather in Delhi NCR! Pitch dark at 5pm']	positive
['Amazing/sleepy weather in Delhi today. pic.twitter.com/TE4BEVJt']	positive
["Great weather in Delhi. Hope it remains like this. I don't feel thirsty during roza yo."]	positive
['Cousin: Sucha gloomy weather in Delhi! Me: Whatta sexy weather!!! :O You call that gloomy? Go die!']	positive
['For a change pleasant weather in Delhi today evening']	positive
['but i like this weather RT @_imonlyindian The weather in Delhi now is like fu**ing shit :-(']	positive
['Hope weather in Delhi is not freezing anymore.']	positive
['Lovely weather in Delhi. Can feel the chill in the air on the way to the airport! #fb']	positive
['Pleasant weather in Delhi but no rains.']	positive
['Wonderful weather in Delhi! Wowowowow']	positive
['The weather in Delhi has been most unbearable over the past few days. Hot and muggy...']	negative
['Several tweets about good weather in Delhi. What timing! #fb']	positive
['I am waiting for the day when someone responds with "Link?" to a tweet like "Lovely weather in Delhi"']	positive
['Beautiful weather in Delhi. Time for love.']	positive
['Lovely weather in Delhi. Mild chill. Sat on a bench in Connaught Place, had chai from hawker. Who knows, he might be prime minister someday']	positive
['The weather in Delhi is so cold, Shashi Tharoor said something stupid just to be in hot water again.']	positive
["Awesome weather in Delhi = When a ballast of super hot air doesn't hit your face when u roll down your Car window."]	positive
['Wearing an Anita Dongre outfit for the Pasta Party for ADHM 2013!Beautiful weather in Delhi:) pic.twitter.com/KQ0qH4Z0AA']	positive
['Aisi bhi baatein hoti hain....lovely weather in Delhi']	positive
['Can we have Bangalore/Hyderabad weather in Delhi, please? :(:(']	positive
["Indigo announces delay due to 'bad weather in Delhi.' Jet announces boarding for Delhi. Simultaneously!"]	negative
["Perfect weather in Delhi for a long drive. Just don't feel like going home!! Even though am almost drenched. It's lush green out on streets."]	positive
['Lovely weather in Delhi. Should share some songs to go with it. Goes with the festive mood']	positive

Out of total 497 tweets, 487 were labelled as positive by the classifier.

Hence percentage of positive tweets = (positive tweets / total tweets) * 100

= (487/497) * 100

= 97.98 %

CONCLUSION

The main motivation behind the project was to understand the intangible as well as tangible factors that can affect the tourism of a particular state. The tangible parameters and its effect on the tourism were identified firstly by using Line graph representation and then by applying the principle of Regression.

Regression tells us the degree of dependence between certain factors and hence a high R-value implied that there is a strong underlying relationship between the two parameters.

The sentiment analysis, which was done for the capital city, Delhi, was one of the intangible parameters that could affect the tourism in the city. The project was important to help us relate 3 completely different parameters and judge their effects on each other. This project has led to an understanding such that we can also further provide inputs to various states on how to improve their tourism to a certain extent.

We were also able to identify the states where the factors affected each other the most and proper measures are needed to make sure that these states are prone to such variations in future.

Also, it was evident that in some of the states, there was no effect of the crime rate on the tourism in that state, and these were rightly identified to be the outliers of the states that we picked for the study.

BIBLIOGRAPHY

1. Twitter sentiment analysis using Python and NLTK- http://www.laurentluce.com/posts/twitter-sentiment-analysis-using-python-and-nltk/

2. India Map - https://gramener.com/indiamap/

3. Complete Code - https://github.com/akshaynagpal/data_mining_tourism