Table of Contents

Introduction

Problem Discussion

For my Data exploration project, I have tried to explore and analyze crimes against women in India from 2001-2015. Reports published by various news agencies like The Guardian and Washington Times gives India number 1 spot when it comes to crime against women. From 2001 to 2015 there have been 2,867,575 crimes against women. It means that on average there had been average 524 crimes per day. My objective in this project is to search and explore the relevant data for answering the below questions.

Questions

1. What are the top 10 states in terms of the number of crimes against women reported from 2001-2015?
2. India can be mainly divided into six regions, so which region had the most crime in 2014 and 2015. What was the overall percentage change from 2014 to 2015 and What was the regional percentage changes in crime from 2014 to 2015? What does the current trend indicate?
3. Which State had most crime in 2015? What was the crime which this state had most?
4. Which areas of state UP (Uttar Pradesh) are highly infected from kidnapping and Abduction? Can the cities be divided into some logical clusters? What relevant insight does this give?
5. What are the regions and states in them in which most Rape Crimes were committed against women in 2015?

Motivation

My motivation behind exploring crimes against women in India is to get deep insights into how safe is women in India and has the situation improved or worsen over time. Due to the lack of strong laws against women crime and how male dominating society has factored-in in different regions. To improve the situation, we need to strengthen our judicial, education and as well as society.

Data Wrangling

Data Sources

The records have been gathered with the aid of different datasets from the Open Government records platform (data.gov.in)[1] and Regions has been tagged to the states from LIC India[2]. The hyperlinks to these datasets are in references. The 4 datasets have been merged into a single dataset. Additional columns which were not frequent in all the datasets had been removed. After all the process Sates were tagged with the areas which they belong to. The final dataset generated after cleaning contained 10,712 rows and 12 columns. Range of attribute year is from 2001 to 2015. The information is about a total of 36 states/UT and 961 cities. The facts have the following attributes.

STATE/UT	DISTRICT
Year	Rape
Kidnapping and Abduction	Dowry Deaths
Assault on women with intent to outrage her modesty	Insult to the modesty of Women
Cruelty by Husband or his Relatives	Importation of Girls
Total Crime	Region

Data Cleaning and Checking

In the data cleaning phase, I have performed the below steps :

Column Checks

- The data for different years 2001-2012, 2013, 2014 and 2015 were imported and were stored in four data frames “csv_2001”, “csv_2013”, “csv_2014”, “csv_2015" respectively. Now a few columns which were not present in csv_2001 were removed from other data frames too.
- The columns had different names in different dataframes. For eg: in csv_2001 one column was named as " Kidnapping and Abduction" but in other, it was written as " Kidnapping & Abduction". Same columns in different dataframes were renamed so that all the there should be no discrepancy.

Merging Dataframes

After all the columns cleaning and checks, dataframes were merged into single dataframe “csv_total.”

Checking for Null Values

One null row was found when checked and we removed it because all the fields for that row were null.

Checking Numerical variable distribution

	Year	Rape	Kidnapping and Abduction	Dowry Deaths	Assault on women with intent to outrage her modesty	Insult to modesty of Women	Cruelty by Husband or his Relatives	Importation of Girls
count	11120.000000	11120.000000	11120.000000	11120.000000	11120.000000	11120.000000	11120.000000	11120.000000
mean	2008.301888	40.982824	55.439119	12.690108	83.762770	16.881475	143.937860	0.089299
std	4.353475	153.095095	270.054569	50.711981	354.740537	83.148256	596.625967	1.276022
min	2001.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	2005.000000	8.000000	6.000000	0.000000	10.000000	0.000000	11.000000	0.000000
50%	2008.000000	22.000000	20.000000	5.000000	34.000000	1.000000	50.000000	0.000000
75%	2012.000000	44.000000	49.000000	15.000000	84.000000	11.000000	139.000000	0.000000
max	2015.000000	5076.000000	10626.000000	2469.000000	11713.000000	4702.000000	23278.000000	60.000000

By looking at the standard deviation and difference between median and max, there surely looked some mistake with data. So I then checked the Categorical Variables

Checking Categorical Variables

By looking at categorical variables it got cleared that there is a huge difference between median and max is because of some Total rows in the dataframe which will have to be removed.

	STATE/UT	DISTRICT	Year
count	11120	11120	11120.0
unique	72	1618	15.0
top	UTTAR PRADESH	Total District(s)	2015.0
freq	866	72	852.0

India has a total of 29 states and 7 union territories. So the maximum number of unique "STATE/UT" should not be more than 36 but above this number is 72. We surely need to analyze it.

- In STATE/UT column it was found that some values were in small and some were in all caps. And some states had different names in different data sources so that got carry forward to our combined dataframe. All those errors were fixed and the total was brought to 36.

- We can observe that there is a field called "Total Districts(s)" in the District. This must be the total of each crime for all the districts in a particular state. We do not require it in this table.

- To remove fields containing Totals in the “District” columns I found the rows using below code. *csv_total[csv_total["DISTRICT"].str.contains('TOTAL')]*

- I dropped all the rows which I got from the above code.

Adding Region column to Dataframe.

As mentioned above under heading data sources, I found that India is mainly divided into six regions and states which are under them. I created a new dataframe manually which had states and regions associated with them. Below is the head of the dataframe “zone_df”.

	STATE/UT	Region
0	ANDHRA PRADESH	SOUTH CENTRAL REGION
1	ARUNACHAL PRADESH	EASTERN REGION
2	ASSAM	EASTERN REGION
3	BIHAR	EASTERN REGION
4	CHHATTISGARH	CENTRAL REGION

I joined dataframes “csv_total” and “zone_df” on column “STATE/UT” and updated “csv_total”.

Final touch

- I created a new column in "csv_total" to store a total of all the crimes for each row and stored it in " Total crime".
- After all the manipulation I rest the index of dataframe after dropping the old index. Now my data frame is ready to be exported to R for exploration. Below is the tail of my final dataframe.

	STATE/UT	DISTRICT	Year	Rape	Kidnapping and Abduction	Dowry Deaths	Assault on women with intent to outrage her modesty	Insult to modesty of Women	Cruelty by Husband or his Relatives	Importation of Girls	Total Crime	Region
10707	TELANGANA	NALGONDA	2015	123.0	40.0	21.0	444.0	209.0	611.0	0.0	1448.0	SOUTH CENTRAL REGION
10708	TELANGANA	RANGA REDDY	2015	31.0	10.0	7.0	60.0	21.0	131.0	0.0	260.0	SOUTH CENTRAL REGION
10709	TELANGANA	SECUNDERABAD RAILWAY	2015	2.0	0.0	0.0	8.0	3.0	2.0	0.0	15.0	SOUTH CENTRAL REGION
10710	TELANGANA	WARANGAL RURAL	2015	52.0	39.0	19.0	226.0	48.0	180.0	0.0	564.0	SOUTH CENTRAL REGION
10711	TELANGANA	WARANGAL CITY	2015	33.0	24.0	3.0	94.0	68.0	245.0	0.0	467.0	SOUTH CENTRAL REGION

Tools used for Wrangling

I have used Python 3.5 for major Data Wrangling and R for in between wrangling like subsetting dataframes or meting them to new ones.

Data Exploration and Visualisation

Question 1

What are the top 10 states in terms of the number of crimes against women reported from 2001-2015?

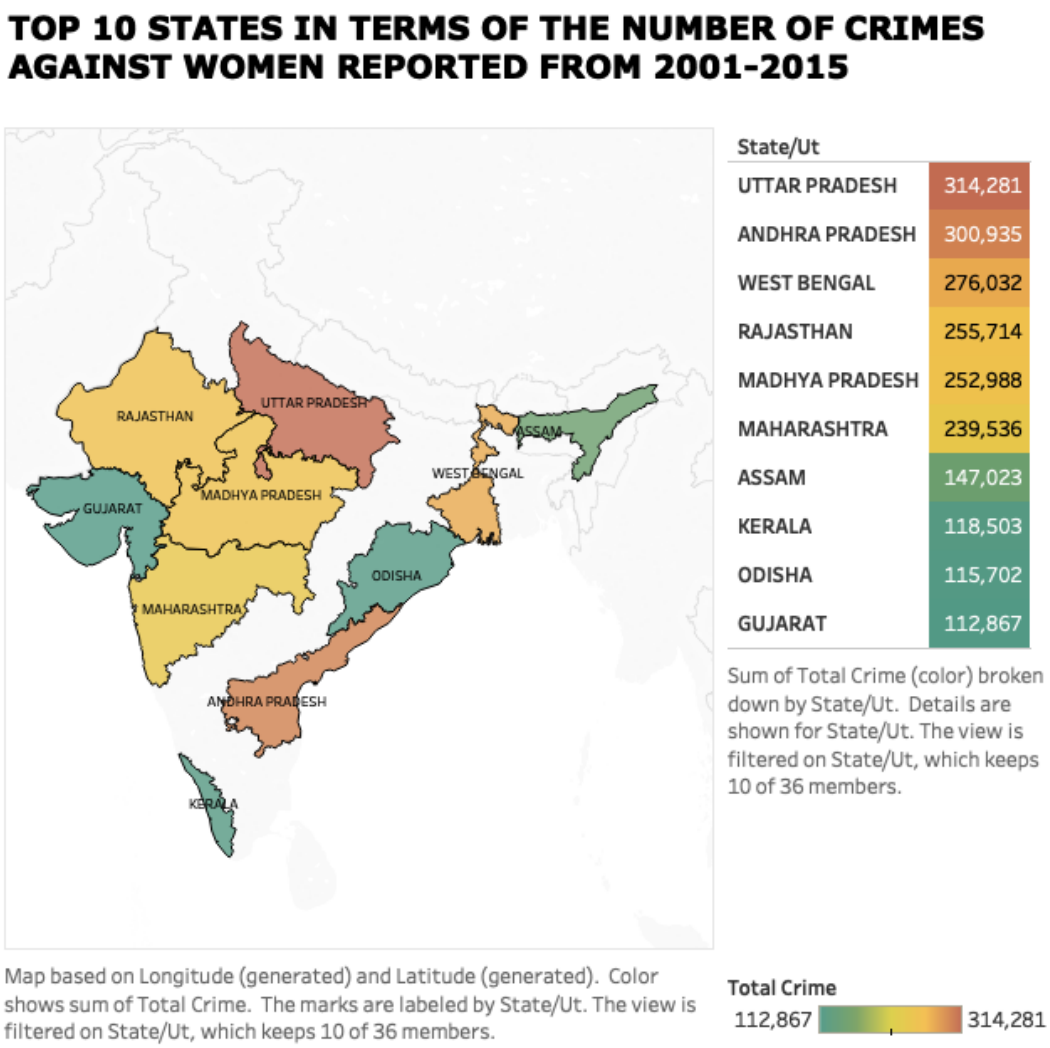


Figure 1 – Top 10 states in terms of the number of crimes

Above result shows that Uttar Pradesh was ranked first when total crimes of 10 years were accounted. There is a percentage difference of 178.45% between Uttar Pradesh which is on top and Gujarat which is on 10th.

Question 2

What was the overall percentage change from 2014 to 2015 and What was the regional percentage changes in crime from 2014 to 2015? What does the current trend indicate?

India has been divided into 6 regions shown by color below. Few Interesting things can be observed from the below visual as well as data exploration I did to get below visual. In 15 years crimes against women has been increased by 151% in India compared to 2001. But in 2015 there has been a decrease of 1.50% compared to crimes in 2014. The major decrease is in the Central region which has been decreased by 17.12%.

Northern Region, Central Region, and Southern Region comprises almost half India geographical and these regions have seen a reduction in crime. One more interesting correlation can be found from this is that India went to elections in 2014 and got the full majority government after 30 years. Can there be any correlation between crime and the authority and decisiveness the government holds when it is of the full majority or it is a collation on? This is something to look into. As it was out of the scope of this project and hence kept untouched. All the other statistics are in the infographic below.

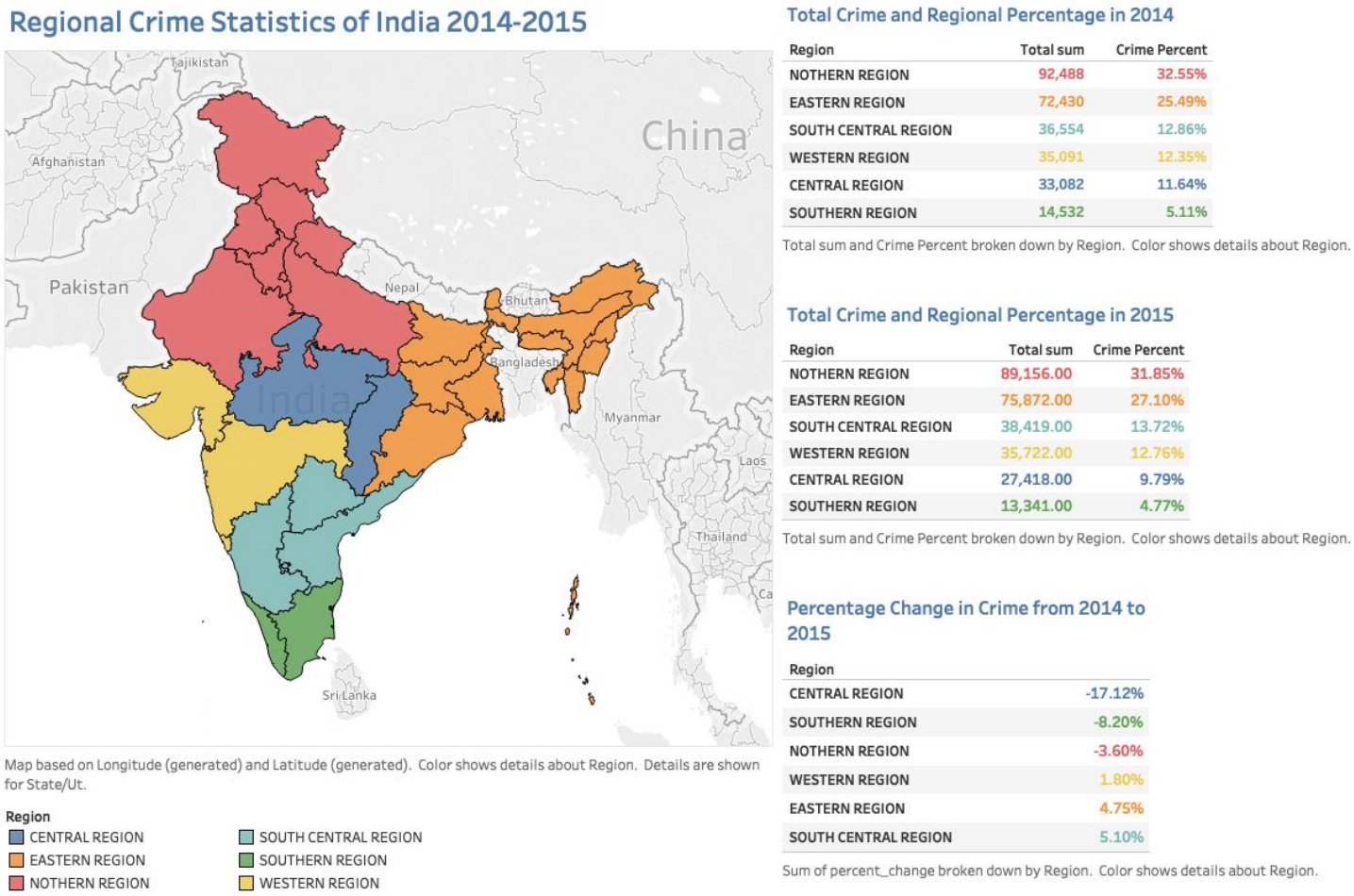
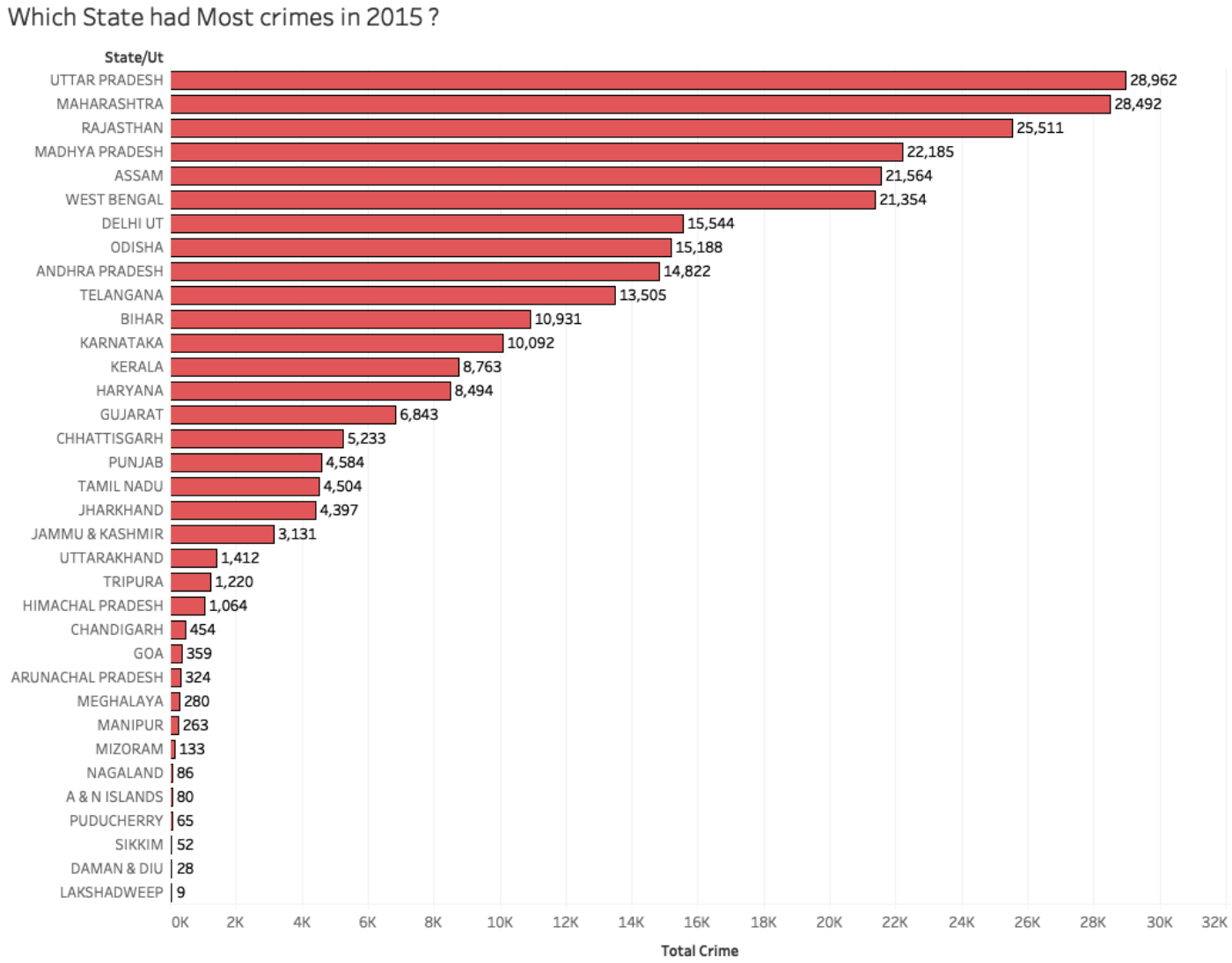


Figure 2 – Regional Crime Statistics of India 2014-2015

Question 3

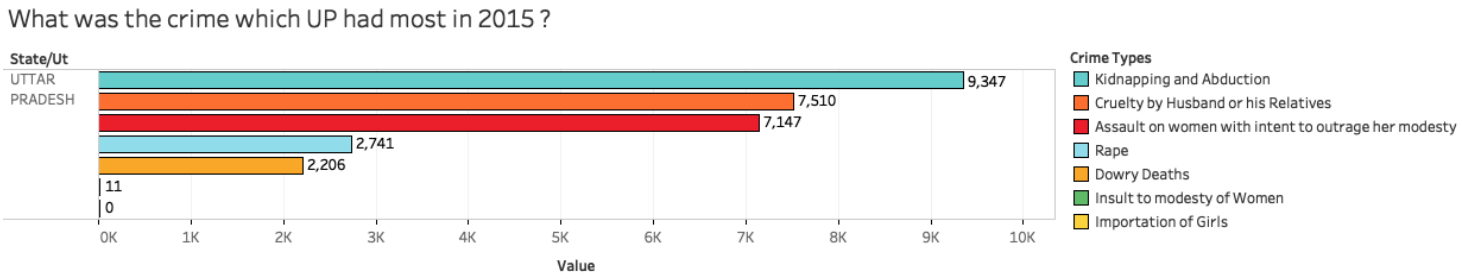
Which State had most crime in 2015? What was the crime which this state had most?

Exploring the question 1 in detail it was found that Uttar Pradesh(UP) had the most crimes in 2015. One interesting fact about UP is that it is the most populous state as well as one of the biggest state in India. There is surely a correlation between the size and population of the state and the crimes against women in them. This can be looked in detail in the visualization project. Lakshadweep, Daman & Diu, Puducherry are very small states compared to UP and crime in them is very very less compared to UP.



Sum of Total Crime for each State/Ut. The marks are labeled by sum of Total Crime. The data is filtered on Year, which ranges from 2015 to 2015.

Figure 3 – State Ranking of Total Crimes against Women in 2015

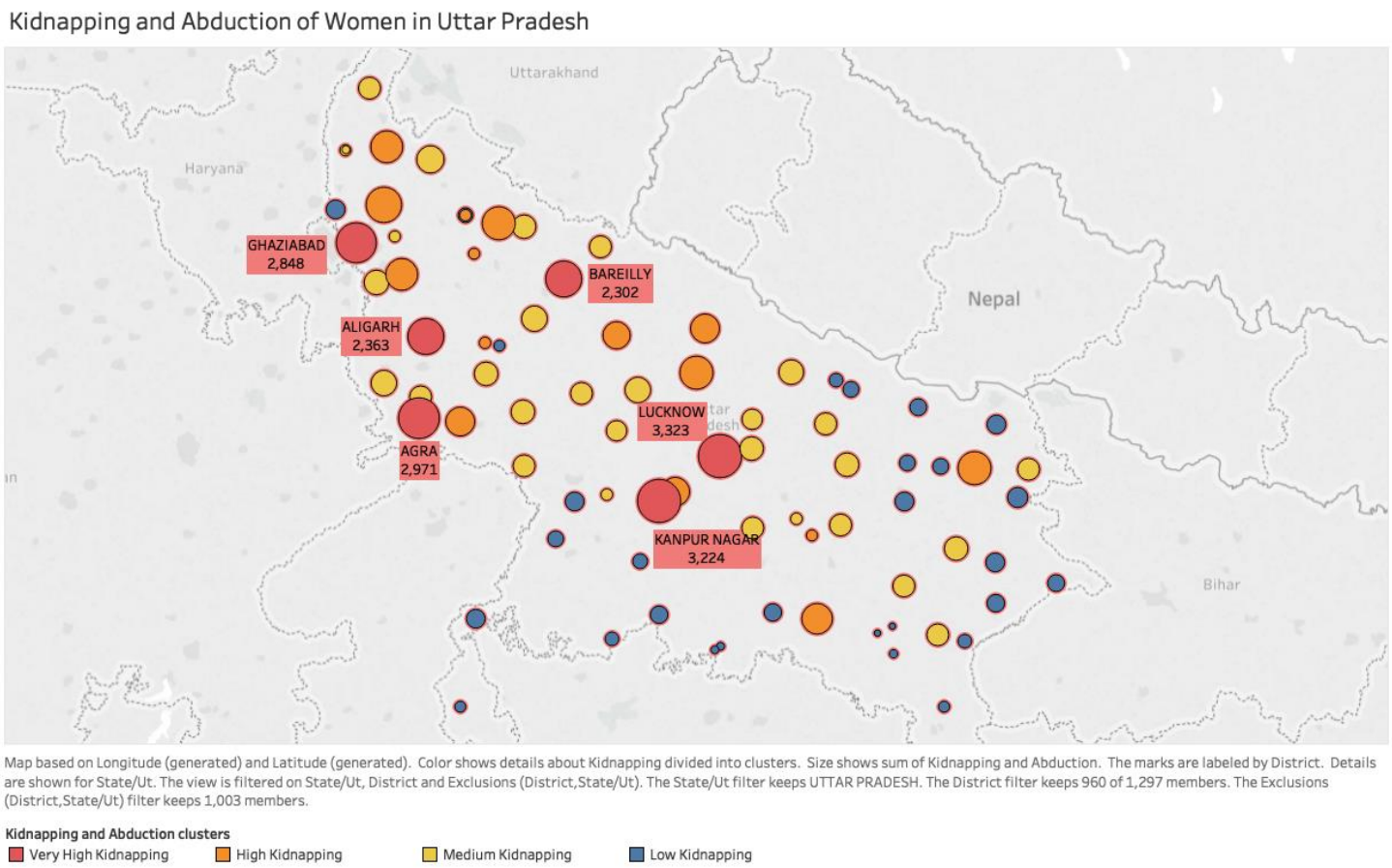


Kidnapping and Abduction, Cruelty by Husband or his Relatives, Assault on women with intent to outrage her modesty, Rape, Dowry Deaths, Insult to modesty of Women and Importation of Girls for each State/Ut. Color shows details about Kidnapping and Abduction, Cruelty by Husband or his Relatives, Assault on women with intent to outrage her modesty, Rape, Dowry Deaths, Insult to modesty of Women and Importation of Girls. The marks are labeled by Kidnapping and Abduction, Cruelty by Husband or his Relatives, Assault on women with intent to outrage her modesty, Rape, Dowry Deaths, Insult to modesty of Women and Importation of Girls. The data is filtered on Year, which ranges from 2015 to 2015. The view is filtered on State/Ut, which keeps UTTAR PRADESH.

Figure 4 – Crimes in UP against Women in 2015

Question 4

Which areas of state UP (Uttar Pradesh) are highly infected from kidnapping and Abduction? Can the cities be divided into some logical clusters? What relevant insight does this give?

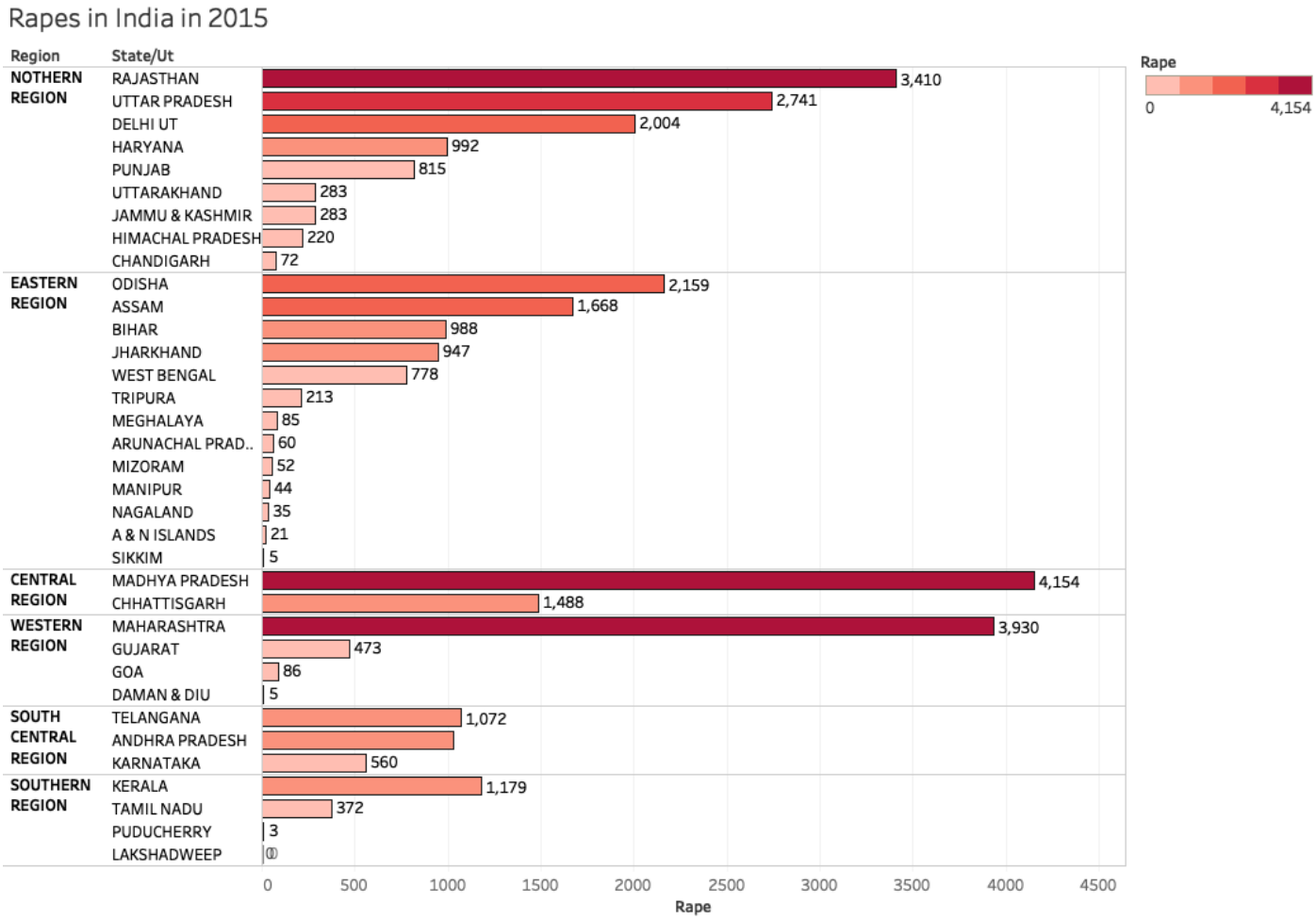


Extending the Questing 3 further the above map shows the detailed study of Kidnapping and Abduction crimes in UP. Here I have divided all the cities into the above-highlighted clusters. It can be seen from above that cities in red are the ones with the most kidnapping and Abduction cases. These cities are also the top 5 cities of UP in terms of development and richness.

Question 5

What are the regions and states in them in which most Rape Crimes were committed against women in 2015?

The below graph shows the state in each region which had highest Rape incidents in 2015. We can see that though there had been a reduction in crime in the Northern region by 3.60%, still it comes on first and Central region by 17.12%, still, it comes on the third.



Sum of Rape for each State/Ut broken down by Region. Color shows sum of Rape. The marks are labeled by sum of Rape. The data is filtered on Year, which ranges from 2015 to 2015.

Tools used for exploration and Visualization

I have used R and Tableau public for exploration and visualization. Majority of the exploration was performed in R and majority of the visualization was done in Tableau Public.

Conclusion

After analyzing and exploring the dataset, I have answered the questions I proposed. To conclude the analysis, I present the important points I have observed above.

1. Total crime in India has been increased by 151% in 15 years but there has been a decrease in crime by 1.50% from 2014 to 2015.
2. There is a difference of 178.45% in the 1st and 10th state in terms of total crime.
3. Uttar Pradesh had most crimes in 2015 and Kidnapping was on top of it.
4. Highly populous, rich and developed cities in UP had the highest kidnapping and Abduction crimes reported.
5. Out of total crimes in 2015, 39% of crimes were in the northern region and only 5% in the southern region. That means Northern region is more Criminal than southern or any other region in India.
6. Northern Region was also found to be top in the total number of reported Rape incidents in 2015 but overall Madhya Pradesh had the highest rape incidents in 2015.
7. Overall, it is not wrong when BBC[3], Washingtonpost[4], theguardian[5] says that India is a most dangerous place for women and needs to bring some hard and strong changes to improve the situation.

Reflection

During my work on this project, I have received publicity to one-of-a-kind techniques of no longer solely visualization and exploration records but also its cleansing and transformation. I have found that even after taking data from authenticated and reliable sources, it is very important to check data as it may require wrangling and cleansing processes. With this project, I have developed a better understanding of Tableau Public and R packages like ggplo2, gganimate, dplyr, ggmap.

References

1. https://data.gov.in/catalog/district-wise-crimes-committed-against-women
2. https://www.licindia.in/getattachment/df990094-a56e-4616-b468-12248816ebea/Empanelment-of-TPAs-for-providing-services-for-LIC
3. https://www.bbc.com/news/world-asia-india-42436817
4. https://www.washingtonpost.com/news/worldviews/wp/2018/06/27/india-ranked-worlds-most-dangerous-place-for-women-reigniting-debate-about-womens-safety/?utm_term=.0a625a0c0ed1
5. https://www.theguardian.com/global-development/2018/jun/28/poll-ranks-india-most-dangerous-country-for-women