# Final Report: Violence Against Women in India

**Data Visualization in Data Science: G0R72a**

**Academic year 2019/2020**

Arleen Lindenmeyer: 0785553
Kristýna Kacafírková: 0776599
Angelo Patane': 0793881
Krzysztof Sobkowiak: 0815539

## I. Introduction and background

Our project is focusing on one of the present social problems regarding gender inequality, which could be found all over the world - it is violence against women. We specifically choose India, because it is one of the most complex countries concerning its different social, religious, and economic background (Johnson and Johnson, 2001, p. 1053). Dissimilarities between particular states could be interesting for our analysis and can also show how important it is to approach the problem from a different perspective which suits the particular area and how crucial it is to be able to react to the issue with suitable means which could differ state to state, crime to crime.

The data, which we used for this analysis were collected by NCRB in India (National Crime Records Bureau), this data set shows the development of crime against women in 2001 - 2012 and it allows us to gain all information described above: violence classification, location of crime (state) and year in which was the situation recorded. In addition, we also used a data set from the Health Management Information System in India, which allowed us to compare the proportion of cases in India. The dataset contains data specifically for the year of beginning of recording crime (2001) and the final year of the research (2011)[1].

## II.Data description

The number of datapoints of our first data set (Crime Against Women in India) is 323, and it consists of 3 dimensions.

*Dimensions:* State/UT, Crime head, Year

*Types of dimensions:*

**categorical** = State/UT, Crime Head

**numeric** = Year

The number of datapoints of our second data set (Population of states/ut of India) is 35, and it consists of 15 dimensions.

*Dimensions:* State/UT, Population - 2001, Population - Males - 2001, Population - Females - 2001, Population - 2011, Population - Males - 2011, Population - Females - 2011, Sex ratio - 2001, Sex ratio - 2011, Density (per sq.km) - 2001, Density (per sq.km) - 2011, Decennial growth rate(%) - 1991-2001, Decennial growth rate(%) - 2001-2011, Average Annual Exponential Growth Rate (%) - 1991-2001, Average Annual Exponential Growth Rate (%) - 2001-2011.

[1] *As there were no population data available for 2012, we decided to focus on 2001-2011*

*Types of dimensions:*

**categorical**: State/UT

**numeric**: Population - 2001, Population - Males - 2001, Population - Females - 2001, Population - 2011, Population - Males - 2011, Population - Females - 2011, Sex ratio - 2001, Sex ratio - 2011, Density (per sq.km) - 2001, Density (per sq.km) - 2011

**percent**: Decennial growth rate(%) - 1991-2001, Decennial growth rate(%) - 2001-2011, Average Annual Exponential Growth Rate (%) - 1991-2001, Average Annual Exponential Growth Rate (%) - 2001-2011.

The most used dimensions are: 'State', which indicates the particular state of India, 'Crime', which represents the type of committed crime, 'Crime01' and 'Crime11' which stand for the number of committed crimes respectively in 2001 and 2011. The two numeric variables 'Crime01' and 'Crime11' are highly correlated with a correlation value of 0.985.

First, missing values are checked; in this case, none of the observations have missing values.

Then, it is noticed that among the levels of 'Crime' there is a particular type called "TOTAL CRIMES AGAINST WOMEN" which should account for the total number of crimes in a particular state either for 2001 or 2011; therefore, it is investigated whether the number of other crimes committed sum up correctly to this number. After a quick analysis, it can be noted that the sum of the crimes, both in 2001 and 2011, is not the same as the before mentioned values; as a result, a new level "OTHERS" is created, which accounts for the number of crimes not included in any of the other levels.

Finally, since the names of the types of crime are very long and informative, they are changed in something shorter and more compact (i.e., "CRUELTY BY HUSBAND OR RELATIVES" becomes "FAMILY CRUELTY").

Since the main interest in one of our visualizations is to see the change of the number of crimes over the two years, a new variable 'Percentage' is created, which stands for the percentage of a particular type of crime for a given year. Moreover, it is worth to be mentioned that a new version of the initial dataset was made[2] such that the dimensions now consist of: 'Crime', the same as before, 'Year', which represents the year of interest, and 'Total_Crimes', which indicates the number of the particular Crime in the given year, and, of course, 'Percentage'. Finally, in order to get an actual comparison over the two years, a new 'Difference' variable is created, representing the difference on the percentages between the two years for a given crime.

[2] *It fits better the making of the visualizations*.

To compare the years despite changed population size the total amount of crimes was multiplied by a million and divided by the population of the state in each year, to create a new variable: quantity of crimes per million citizens.

## III. Interactive visualizations

### Goal and task description

The goal of our project is to present the situation regarding crime against women in India in 2001 and 2011 to show particular differences in states in prevalence of violence and various types of crimes over the time and explain how this information can be relevant for following years. Precisely, we intended to discover detailed patterns which are crucial to answer our main research question:

*What action should be taken in India to prevent violence against women?*

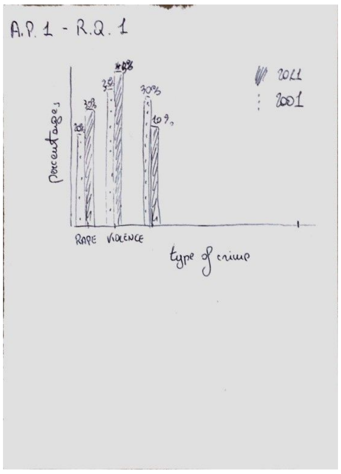
We try to identify the crime escalation over the time, the connection between types of crimes and significance of crimes in particular states which together with knowledge of local policies and current situation in states help us to answer our research question mentioned above. In fact, these insides could be enriching and essential in case of policymaking for further decades and can be helpful in case of developing more efficient and straightforward policies which react to the problem directly according to the needs of a particular area in India.

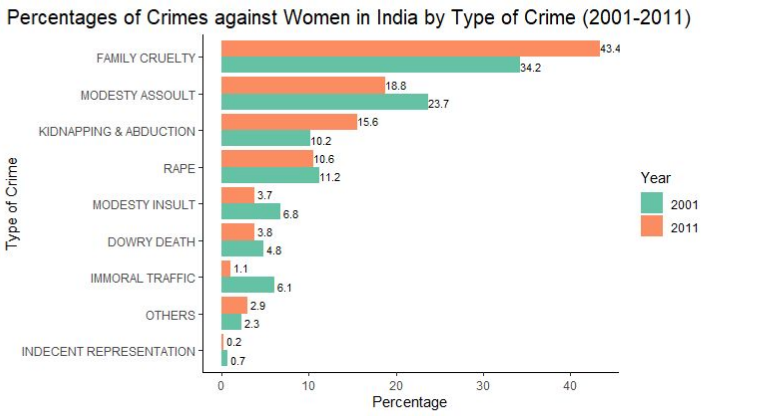
### Exploration of design space

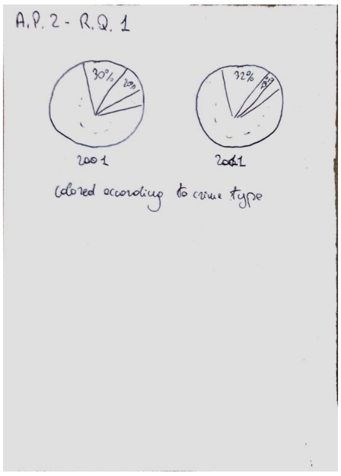
As we found out, most of the research dealing with crimes using simple line charts, bar charts or pie charts. However, many of them work also with maps to easily present the geographical role of prevalence, which inspired us to use it to show the differences between particular states. Besides a map we wanted to come up with more creative and interesting visualizations then mentioned above, therefore we chose to make several different graphs which overall make one story behind. Our designs then can be divided into three types: *A. general information of the situation in India*, *B. comparison between states*, *C. detailed information of one state*. In the following part, we are going to go through each graph's evolution including handwritten sketches.

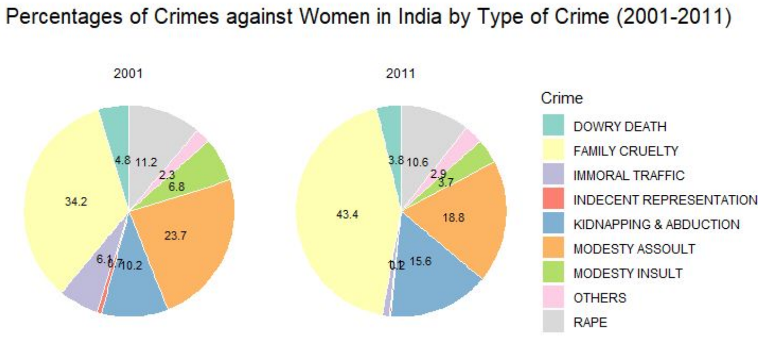
*Visualization type A*

First, in order to see how the number of crimes has changed over the decade 2001-2011, some static visualizations are proposed (check the pictures here);

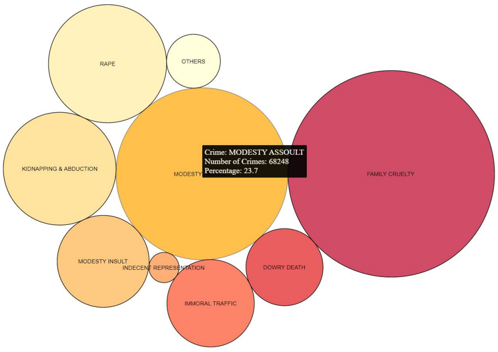


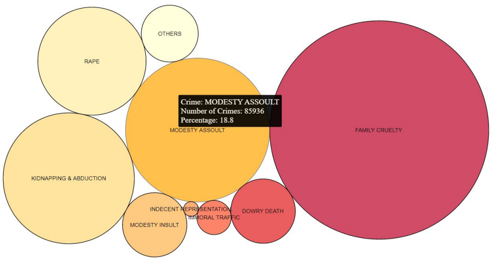




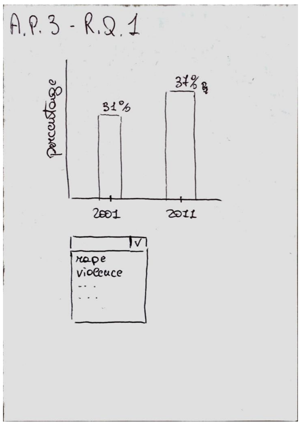


Then, the aim was to move into something more interactive; a shy approach to this can be found here for insights on 2001 and here on 2011.

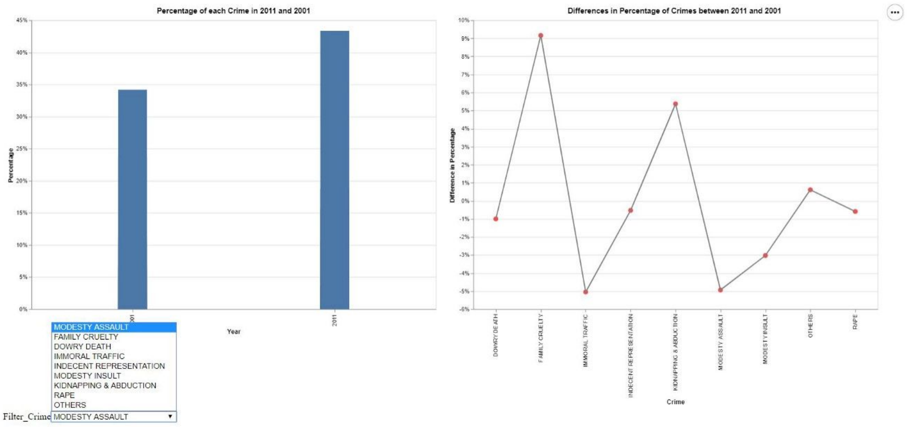




Afterwards, the last process of graph making here consisted of the usual initial sketching part;

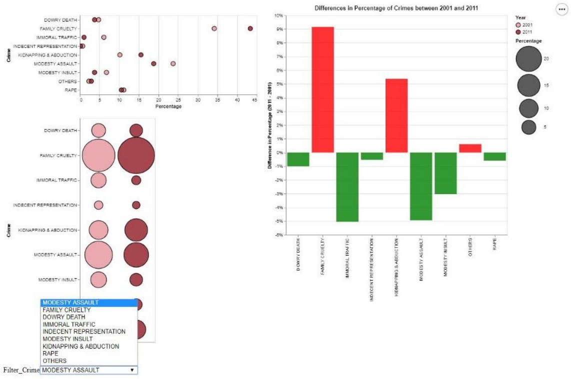


The idea behind this draft is being able to compare the amount of crimes between the two years with the least effort possible, just taking a quick look at the bar heights.
Then, a new graph is accompanied to the sketched one, once it is realized that it might be useful to have an actual quantitative difference for the comparison of the crimes.

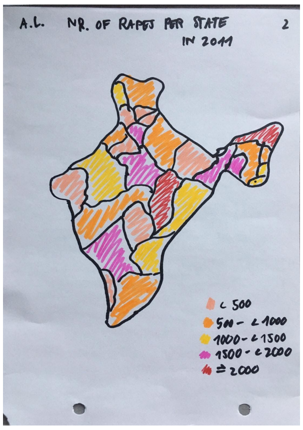


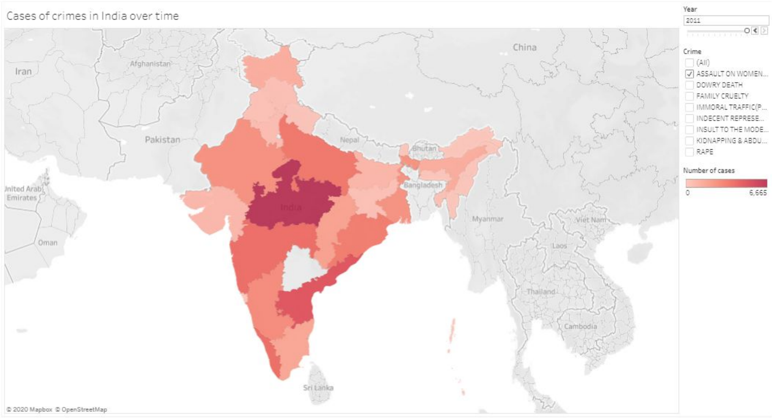
Take a look at it here.

Finally, the ultimate version of the interactive visualization is made and shown here, with the intent to make the previous version as more appealing and intuitive as possible.



In case of *visualization type B*, we decided for *Tableau* tool and started with rather simple one layer maps. The sketch and example of one of the first visualizations can be found below.



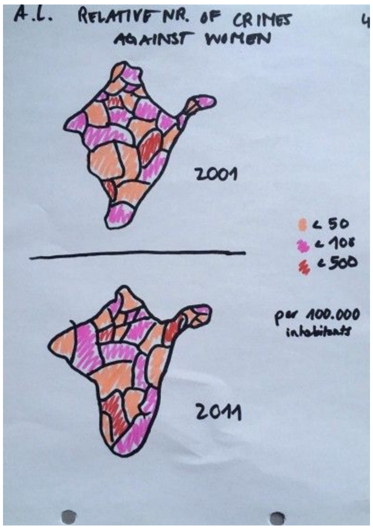

Cases of crimes in India over time

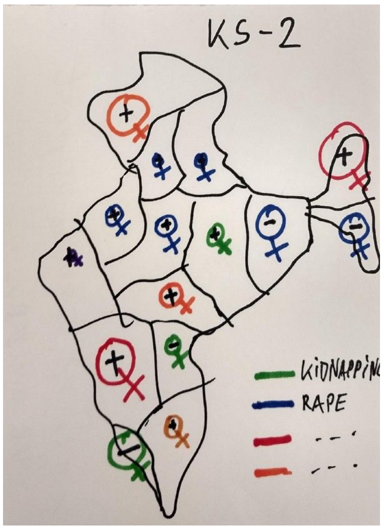
Year
2011

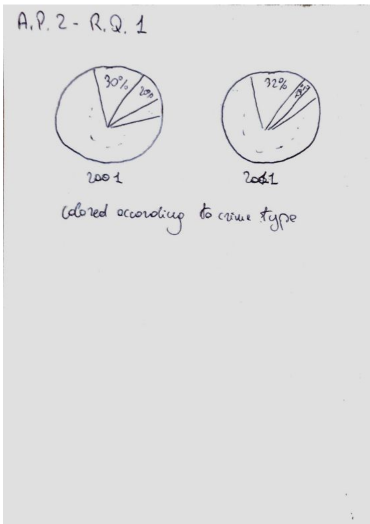
Crime
□ (All)
☑ ASSAULT ON WOMEN...
□ DOWRY DEATH
□ FAMILY CRUELTY
□ IMMORAL TRAFFIC(P...
□ INDECENT REPRESE...
□ INSULT TO THE MODE...
□ KIDNAPPING & ABDU...
□ RAPE

Number of cases
0 6,665

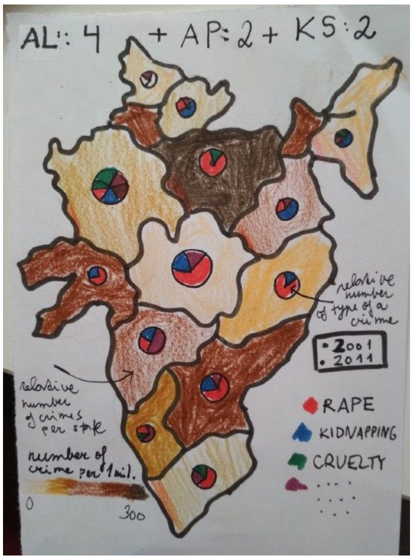
The following step of making the final visualization was to improve the former one in several areas: adding proportional data, allow to see all crimes at the same time and show only the map of India to make it more clear. For this was a new sketch made from following sketches merged together:

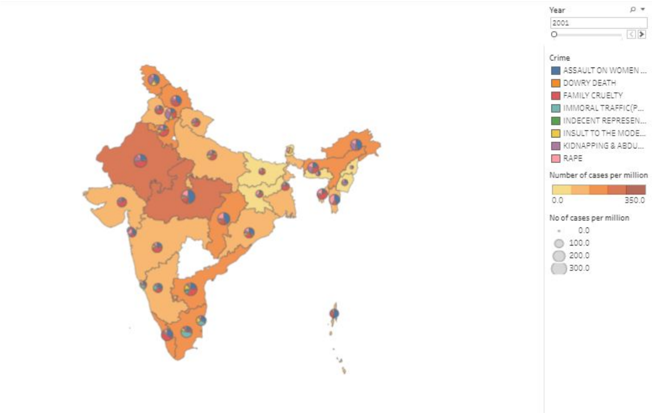






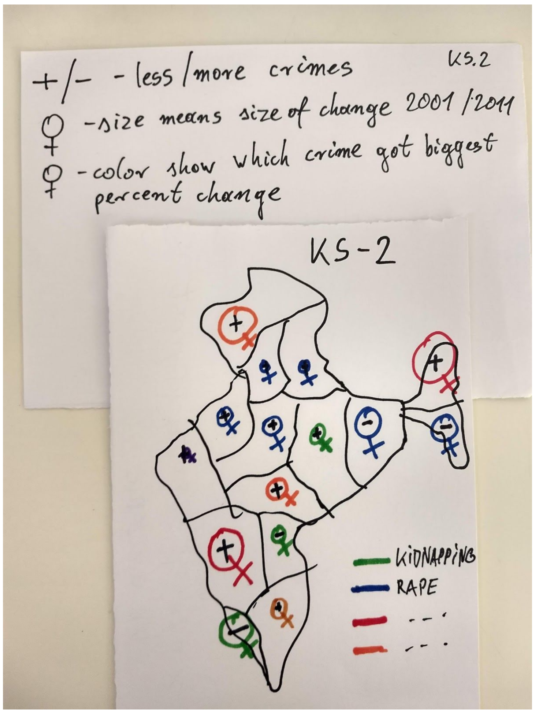
The idea for colours, the relative number of crimes and the change between 2001 and 2011 was used from the AL4 sketch. Number of inhabitants was modified to 1 million, as India is a very populated country and numbers for crimes would be then very small. Sketch KS2 was the inspiration for using a 2 layer map with different symbols. However, we wanted to use a symbol which can show the number of different types, that's why we decided to use a pie chart from the sketch AP2. The process was followed by creating the combination of sketches which can be seen in the next picture and the final result next to it.


AL: 4 + AP: 2 + KS: 2
relative number of type of a crime
2001
2011
relative number of crimes per state
number of crime per 1 mil.
0
300
RAPE
KIDNAPPING
CRUELTY

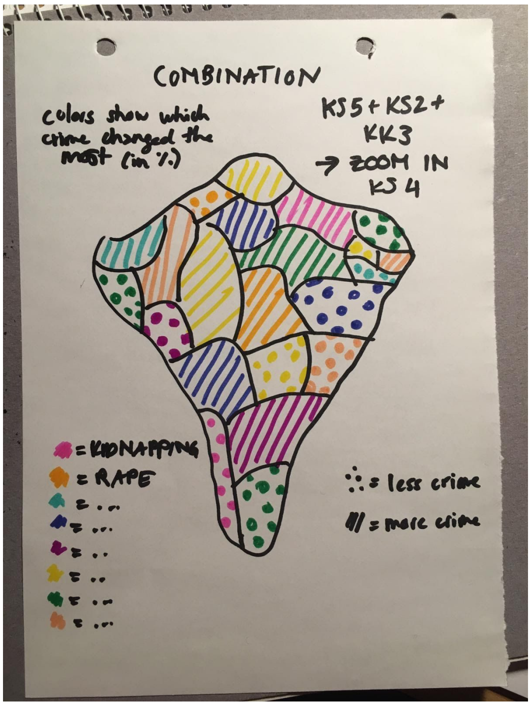


https://public.tableau.com/profile/kristyna.kacafirkov#!/vizhome/ViolenceinIndiaCasespermillion/Sheet1?publish=yes

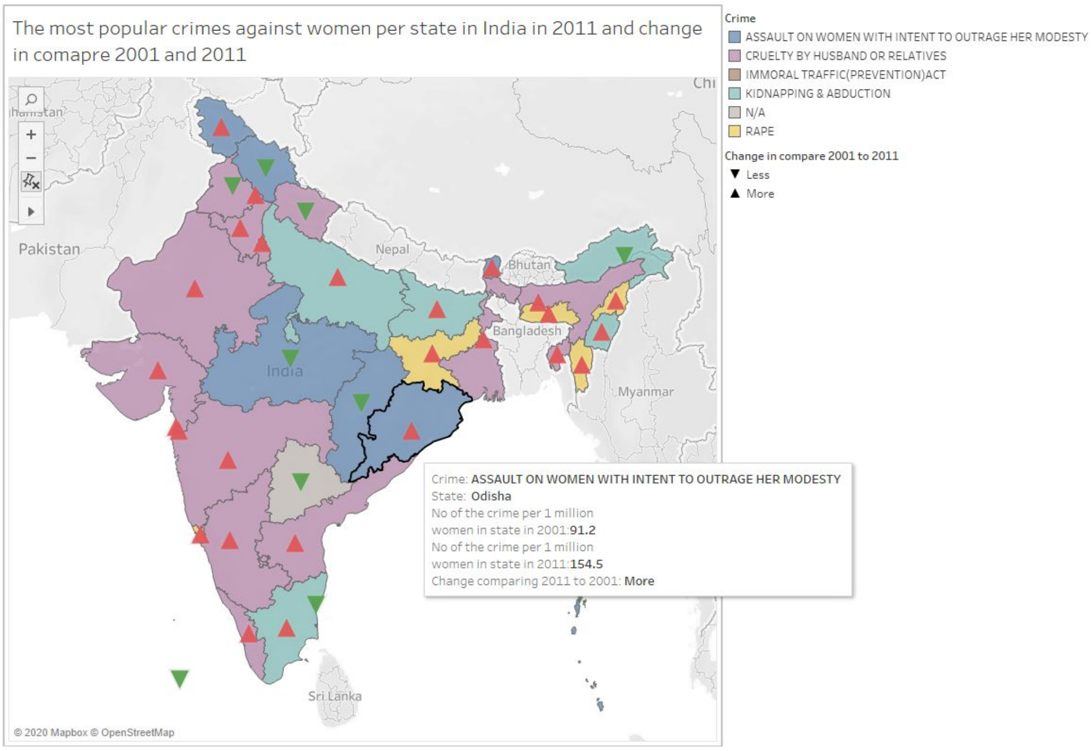
We also made one map showing the differences between years. Goal was to easily show what has changed in comparison 2001 to 2011. It is a little bit similar to the last one but it shows data in different ways.


+/- - less/more crimes
KS.2
♀ -size means size of change 2001/2011
♀ - color show which crime got biggest percent change
KS-2
KIDNAPPING
RAPE
- - -
- - -

+



First sketch was KS.2. Main feature there was that the size of the "women symbol" shows the size of the change in numbers of crimes between 2001 and 2011. Second thing was that the color of the symbol shows the type of crime. Problem was how to recognize the size of the symbol that there were less or more crimes. KS.2+KS.5+KK3 gave Combination1. Important changes were showing less/more by different marks and better visibility of type of crime by coloring the whole state.


The most popular crimes against women per state in India in 2011 and change in comapre 2001 and 2011

Crime
- 🟦 ASSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MODESTY
- 🟪 CRUELTY BY HUSBAND OR RELATIVES
- 🟫 IMMORAL TRAFFIC(PREVENTION)ACT
- 🟦 KIDNAPPING & ABDUCTION
- ⬜ N/A
- 🟨 RAPE

Change in compare 2001 to 2011
- ▼ Less
- ▲ More

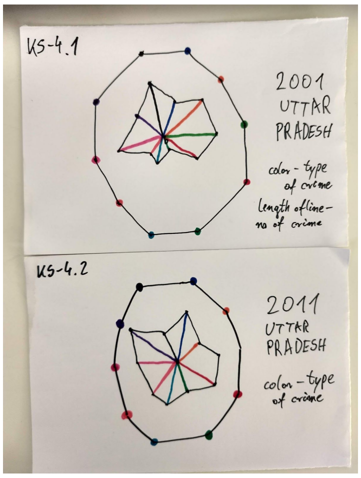
Crime: ASSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MODESTY
State: Odisha
No of the crime per 1 million women in state in 2001:91.2
No of the crime per 1 million women in state in 2011:154.5
Change comparing 2011 to 2001: More

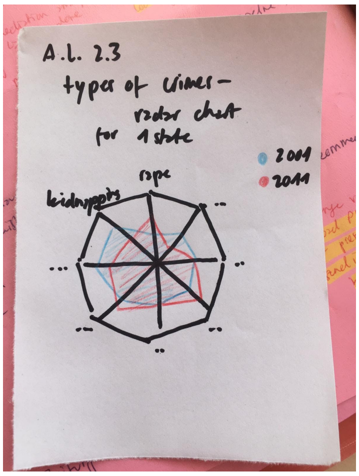
https://public.tableau.com/profile/krzysztof.sobkowiak#!/vizhome/indiaVIz/Dashboard5

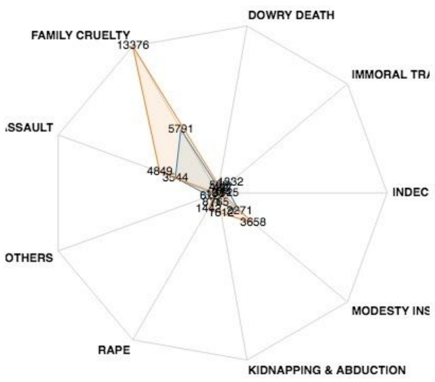
As you can see in the final visualization, colors of states still show type of crime. But the change in the number of crimes is shown by arrows. Red arrow shows that in 2011 there were more crimes than in 2001 and the green arrow shows the opposite. Also after clicking on state you can see details.

*Visualization type C* focuses on comparing the amount of crime in one state in each year. Each type of crime should be visible, and how it changed over the decade in one state.

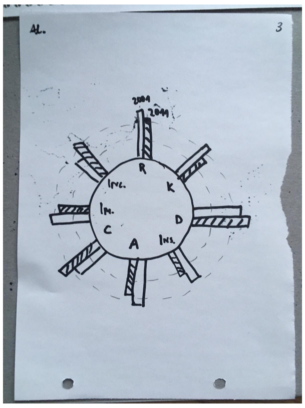
After some discussion, we came to the conclusion that a radar chart would be a suitable fit to represent our idea. In the sketches KS-4.1 and KS-4.2 each crime has its own color and each year has its own chart. To compare both years and the difference in crime more easily, in the sketch AL 2.3 we used color for years and the scales are marked with the crime they represent. But as you can see in the R altair visualization, the radar chart did not fit our data: because some crimes occur very often, but others not at all or are very seldom, and crimes that occur not often are barely visible.

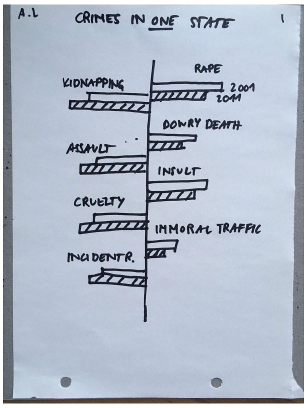






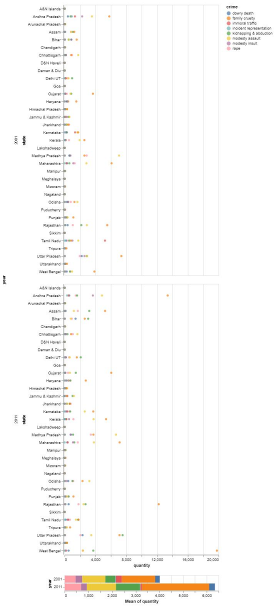
After this, we had to come up with other visualizations, for example in a bar chart you can still see small amounts, even when some of the numbers in the data are very large in comparison. The sketches below are some of the ideas how bar charts could be incorporated into the visualization.



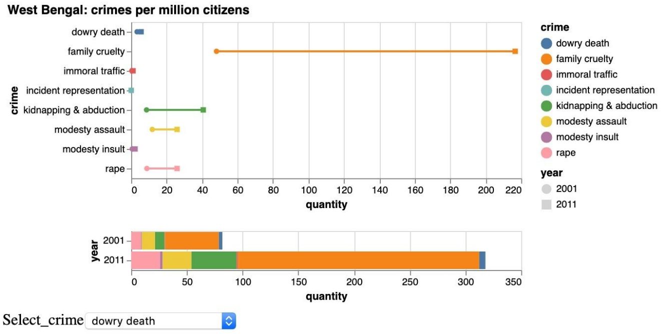


At first, all of the states were included the next R altair visualization. However, this makes it chaotic and long. Additionally, the points are hard to see, and because to years are separate graphs you cannot compare them directly easily. As this visualization is interactive, you can select the state (draw a quader over the point in the visual) and compare the years of one state, but you still cannot compare states to each other. Thus, it makes more sense to only show one state at a time, and compare how the crimes changed over the decade. It is important to see which crime changed more than others, and for this it is best if both years are represented within one plot.

https://arl-cloud.github.io/india_states/



In the final visual, we tried to improve these aspects. This visualization was also done using R altair. The graph gives an overview of all types of crimes in both years in one state. It has points (different shapes for years), which are connected by lines, so one can easily recognize if a (large) change in crime happened in the state. Further you can select a crime and compare how the amount of this specific crime changes easily with the bar chart. Lastly, we used the population data for the specific states to set the data in proportion.

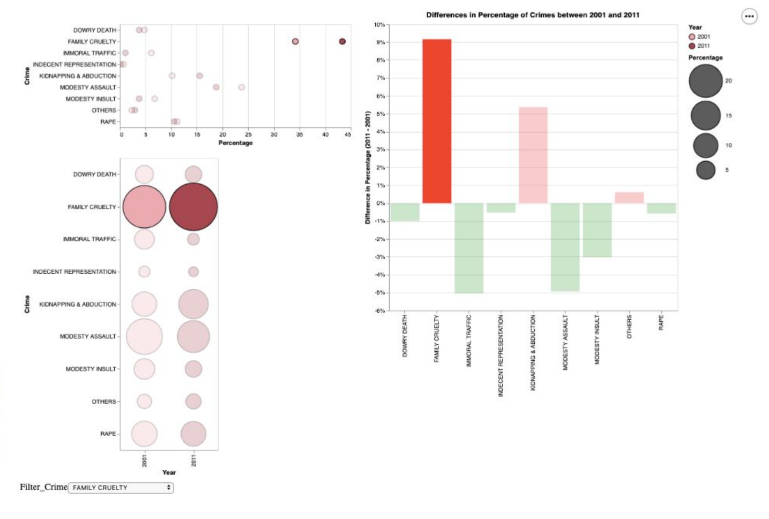


https://arl-cloud.github.io/india_West_Bengal/

## Final design with annotations

Our final visualization copies our design process, it is composed from three different visualizations which are connected together.

As we can see in our first graph the most significant problem to consider is family cruelty. It is also the one with the biggest increase during the decade. The second place of highest rise takes kidnapping and abduction. All other crimes surprisingly decreased.

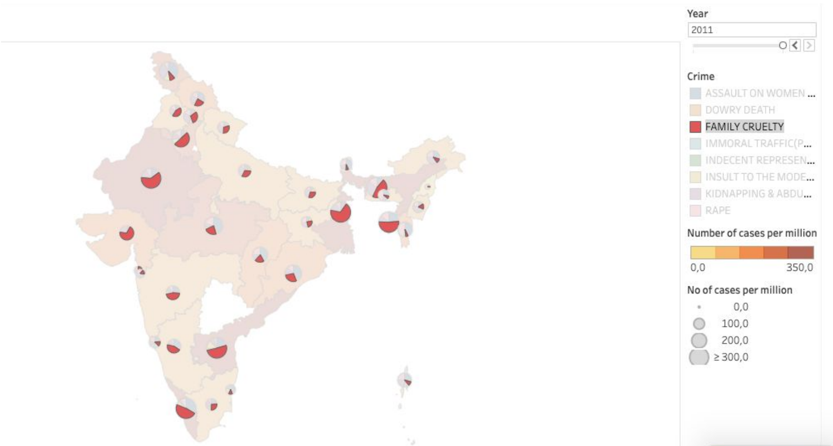


Filter_Crime FAMILY CRUELTY

https://patan3.github.io/interactive_2/

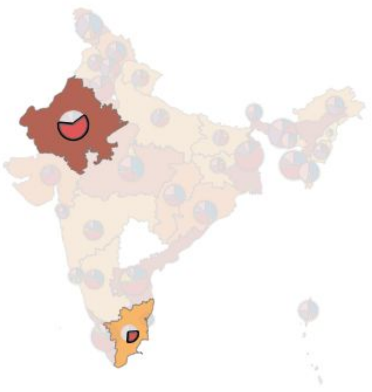
It is important to mention that despite the fact that that most of the crimes decreased, many of them are still substantial. For example modesty assault which recorded one of the highest decline, is at the same time the second highest crime in the country. Thus, it cannot be overlooked and must be also taken in consideration of prevention of violence.

However, as we mentioned above, family cruelty is persisting, which may be related to the upbringing and patterns which are handed generation to generation. It is a more complex problem which starts from early child age, when the perception of inequality begins with unequal access to education, unequal position in family etc. Moreover, it is very hard to reveal this type of crime, as it is committed by the closest persons to the victim. Therefore, it is more complicated to mitigate it, direct policy tools do not necessarily have to work very well as it is difficult to find out what is going on inside the family. To find out more about the situation of family cruelty, we will have a look at a detailed map.


Year
2011

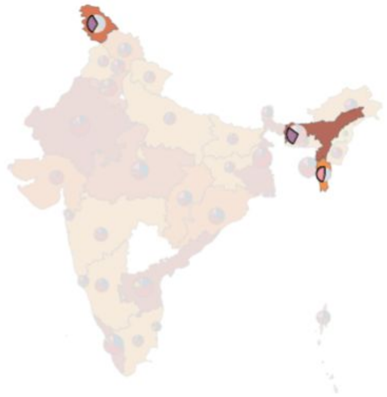
https://public.tableau.com/profile/kristyna.kacafirkov#!/vizhome/ViolenceinIndiaCasespermillion/Sheet1?publish=yes

As we can see, family cruelty is a persistent problem in almost all states of India, which is connected to the nature of the problem mentioned above. However we can see some slight differences.



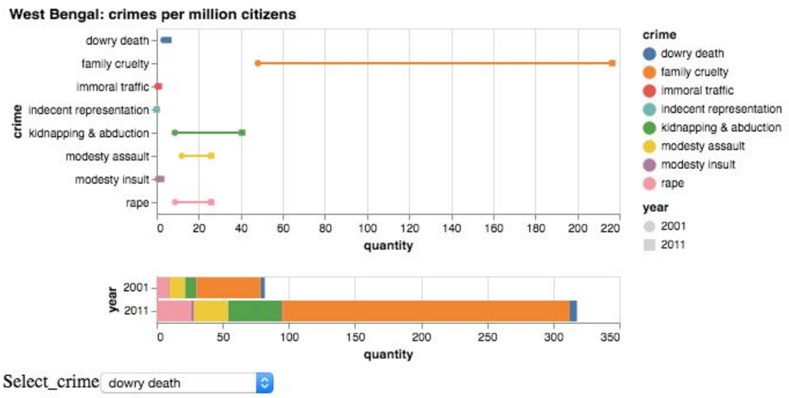
For instance in Rajasthan and Tamil Nadu. While Rajasthan recorded 178 cases per million in 2011, Tamil Nadu had only 25 cases/million.

This example shows how important background is in case of inclination to family cruelty. Tamil Nadu is one of the most urbanized states in India, whereas in Rajasthan ¾ residents live in rural settlements. Family violence public awareness is probably more easily spread in cities than in less connected villages. Therefore, it is important to focus on an information campaign which aims to address inhabitants from all various regions and do not forget rural areas and for example implement specific education in school regarding the topic as well.

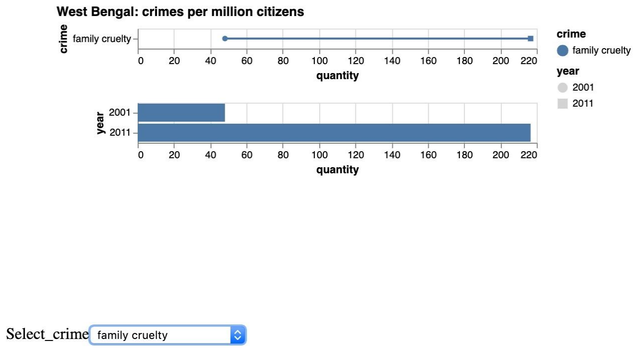


Nevertheless, it is also important to focus on states which were not affected by family cruelty but other crime was much more significant. Jammu and Kashmir together with Assam have a biggest problem with kidnapping (violet), Mizoram with a rape (pink). These crimes could be probably prevented by more strict punishment form of policies than in case of family cruelty.

To bring visualization A and B together, we took a closer look at two states in visualization C. We chose to visualize the state West Bengal, as this state has the largest increase in one type of crime (family cruelty) of all states.



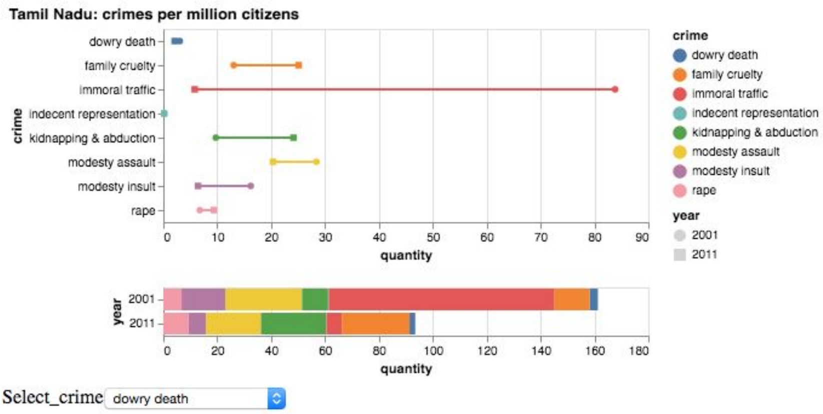
Select_crime dowry death

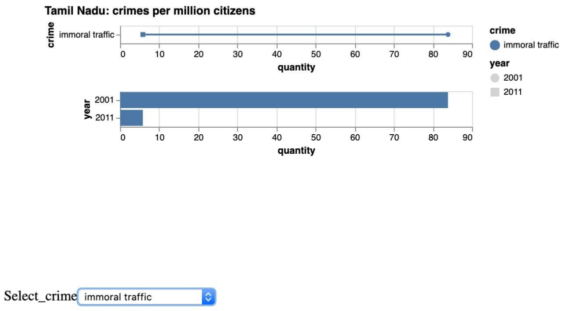


Select_crime family cruelty

https://arl-cloud.github.io/india_West_Bengal/

We chose to also visualize the state Tamil Nadu, as this state has the largest decrease of one type of crime (immoral traffic) of all the states. In the visualization below it is evident that immoral traffic, which was a huge issue in 2001 almost completely disappeared after the decade. Therefore, this state may work as an exemplary one for other states.





https://arl-cloud.github.io/india_Tamil_Nadu/

As you can see in visual A, of the types of crimes family cruelty had the biggest increase over the decade, and immoral traffic the biggest decrease. These visuals suggest that the two states had a big influence in the overall change of crimes in india. In both states, there is a surprisingly large change of crime over the years. To draw any conclusions it would be important to look closely at what could have caused these changes. Were there certain policies implemented in Tamil Nadu which caused immoral traffic to go down? What could have triggered the increase in family cruelty in West Bengal, for example a change in the socio-economic conditions of the state, political decisions? These insights will help to both implement effective interventions and to prevent an escalation in crime rates. Family cruelty is a persistent problem in India, and the rise in the occurrence of it is concerning. Especially when you take into account that violence from relatives is often not reported because of various reasons (e.g. shame, fear), it is clear that taking actions to reduce family cruelty is essential, and as most states have a problem with it, it may be helpful to implement policies on a county-wide level.

As we are in an unusual situation dealing with Covid-19 right now, it would be very interesting to see which effects this has on the amount of crime and types of crimes that occur. For example, many states report less crime (in total) occuring while lockdown measures are implemented, but others say crimes like domestic violence may increase during this period.

## Implementation

*Visualization type A:*

First, the static visualizations are made using R. In particular, the following libraries are used: *'ggplot2'*, for visualizations, *'RColorBrewer*', for choosing the color palettes and *'tydiverse'*, for data manipulation. The process here was straightforward, and there were no problems turning out the hand-drawn sketches into the actual visualizations.

Then, the bubble charts were still made in R, but this time using other libraries (i.e., 'packcircles', 'ggplot2', 'viridis', 'ggiraph') both for making the circles and for making them interactive.

Afterwards, after feeling a little bit constrained by the previous tools, it was chosen to still use R, but this time using the Altair interface to Vega-lite. After some data manipulation in the dataset, already described in the previous section, a first version of the interactive visualization was made, then a new and final update was ultimated with regards to our particular needs.

*Visualization type B:*

Tableau was used in case of maps. Overall Tableau is very intuitive and the potential problems can be discussed via Tableau Community Forums. However, in the case of two layers map, there was an issue with zooming (when you want to zoom in the map, the pie charts are getting smaller), for which I did not find a solution.

*Visualization type C:*

For all visuals of type C R altair was used. I personally like working with R, and I felt when using vega and vega lite, the code for the visualization easily escalates in size, making it hard to keep an overview of the different parts. Most problems I had were easily solvable and small(e.g. Typo in a variable so the visual is not working correctly). When being stuck (e.g. with doing the dropdown select), I found it very helpful to look for clues in the documentation and examples on the r altair website. However, I noticed a lot of useful documentation was written in python, which I am not familiar with. Sometimes I tried to change the documentation of python to r, but it did not always go well.

## Insights

*Working in R altair (type C):* In general, vega (in its various forms) is very useful especially to create interactive or more complex visuals. Particularly if you want to make specific or individualized changes, vega has more tools than other programs. I liked using r altair, and I still prefer it to using Vega or vega lite. It is hard to find useful information other than in the gallery or documentation, I do not think that many people are using this interface. After all, I think using vega in python is the best option if one is familiar with it, as the documentation for this is very good. This gives me more reason to try to learn how to use python. Also, I learned how to publish the interactive visualization on github, which is very useful.

*Working in Tableau (type B):* Even though Tableau is quite easy to use in comparison to Vega, Vega Lite you have restricted options of design to some extent. Working in R altair and vega allows you to be more creative in the sense of developing your very own visualization. However in the case of a map, where the form of the visualization was clear, it was useful to make it by an easier tool as then you can focus on small details without too much worry that if you change one little thing, you will completely destroy the others you have just created.

## IV. The division of work

Project idea: KK, AP, AL, (KS: joined us halfway into the project idea/data search)
Visual designs: KK, AP, AL, KS
Data preprocessing: KK, AP, AL, KS
Implementation: KK, AP, AL, KS
Report: KK (background, goal description, storyline + type B part), AP (data description + type A part), AL (type C part)
Video: KS

## V. Link to the video and code

Viz 1:
https://gist.github.com/patan3/071007c650f4cb5aeff395bc19e80a54
Viz 2:
https://public.tableau.com/profile/kristyna.kacafirkov#!/vizhome/ViolenceinIndiaCasespermillion/Sheet1?publish=yes
Viz 3:
https://arl-cloud.github.io/india_West_Bengal/
https://arl-cloud.github.io/india_Tamil_Nadu/

Video:

**https://www.youtube.com/watch?v=MYSLj-V7Ltk**

## Resources

Crime Against Women in India, Crime in India 2012, National Crime Records Bureau (NCRB), https://data.gov.in/resources/crime-against-women-during-2001-2012

Distribution of Population, Sex Ratio, Density and Growth Rate of Population - Census 2001 and 2011, Registrar General & Census Commissioner, India, https://data.gov.in/resources/distribution-population-sex-ratio-density-and-growth-rate-population-census-2001-and-2011

Johnson, P. & Johnson, J. (2001). The Oppression of Women in India. Violence Against Women. 7. 1051-1068. 10.1177/10778010122182893.