Title:
In 2018, the Number of Women raped in India per each district & broken down by age group

Abstract:
Women being raped is an issue that is known, but it is important to see the numbers and analyze the data to understand the number of women that are becoming victims of this. Once we understand the mass impact, then the actions to stop this will follow suit.

This dataset provides the number of women that were raped in 2018 in India. The dataset provides each district's name and if that district is ruled by State or the Union Territories. There are a total of 29 states and 7 Union Territories. For each district, the number of cases of women raped is provided. Following that, the number of women who were raped is broken down into age categories. The age category is 0-5, 6-11, 12-15, 16-17, 18-29, 30-44, 45-59, and 60 plus.

When looking at the average number of cases and the number of victims, there are fewer cases reported than the number of victims. This shows that not all women report cases, yet given the number of victims, a total number of 262 girls (below age of 18) were raped, and an average of of 681 women (above age 18) were raped. In total, the average number of victims in 2018 in India was 943. This is a high number, especially since this data is only from a single year and in one country.

Introduction:
The dataset discussed in this report is about the number of women raped in India in each district in the year of 2018. The number of women raped have been broken down by age group, particularly to emphasize the number of girls under 18 and women 18 and above that are raped.The purpose of this analysis is to highlight the big issue of rape by showing how many little girls, middle age and elderly women are negatively impacted. Bringing light to the numbers, will give encouragement to help prevent the rape cases and also provide support for women who have been raped.

Data and Data Preparation:

I received this dataset from Kaggle, and was uploaded two years ago by Neha Prabhavalkar. She collected this data from the Open Government Data (OGD) Platform India. The attribute "State/UT" is a categorical- nominal data type that lists the districts in India. The Category attribute, is a categorical - symmetric data type, with a range of two values which are State and Union Territory. This tells the viewer if the district is ruled by the state or union. In total, there are a total of 27 states and 9 Union Territories, which sums up the 36 districts in India which are listed under the "State/UT" column. Following this, there is a column called "Cases Reported", this consists of numeric - ratio data type, where it tells the number of cases of rape that were reported. After this the number of rapes are broken down by age group, so they all are numeric. The first age group is "Child.Victims.of.Rape..Below.18.Yrs....Below.6.Years", which consists of the girls ages 0-5 years old that were raped. Secondly, "Child.Victims.of.Rape..Below.18.Yrs....6.Years...Above", this consists of girls ages 6-11 that were raped. Third, "Child.Victims.of.Rape.. Below.18.Yrs… 12.Years...Above.", which is of ages 12-15. Fourth,"Child.Victims.of.Rape..Below.18.Yrs…16.
Years..Above", which is of ages (16-17). Lastly, "Child.Victims.of.Rape..Below.18.Yrs…total.Girl Child.Victim", is the total number of victims that were children, so below the ages of 18 that were raped. In the same way, cases of women that were raped were broken down by age. So, 18-29,

30-44, 45-59, 60+, and then the total number of women (18+) that were raped. The last column, "Total Victims" constis of the total number of women that were raped, so ages 0-60+.

No preprocessing steps were done to the data prior to the analysis as all the data was filled, with no missing information.

Methods:

I did an Exploratory Data Analysis on the data set. I went ahead and ran a summary of all the variables by running a summary on the dataset, which looks like this in R ``summary(Victims_Rape)". This gave the minimum, 1st Quartile, median, mean, 3rd Quartile and maximum. Then I found the minimum,maximum, mean, and median for total victims, cases reported, and for the number of victims per each age group. An example of this in R was "median(Victims_Rape$Total.Victims). Then I found the variance and standard deviation for the total number of victims through the function var and sd. To understand the tendency of the distribution, I calculated the skewness of the total number of victims. I calculated this by multiplying the median of the total number of victims divided by the standard deviation of the total victims and subtracted this from three times the mean of the total number of victims. I also ran a summary of the categorical attribute "State/UT" to see how many states and how many Union Territories there are in India. I also visualized this through a dot plot, where the y-axis was the two variables, State and Union Territory, and the x-axis was the number of districts for each. To further this, in Figure 2, there is a comparison of total victims in the state vs. the union territory. In Figure 3, a box plot is made, to visulaize the total number of victims for each district. Lastly, I found the outliers in the data set using standard deviation. I established the upper and lower bounds , and if anything is above the upper and below the lower bound then that is an outlier.

Results:

Table 1 and 2, show the summary of the number of victims of rape for each age group. The summary consists of the minimum, 1st Quartile, median, mean, 3rd Quartile, and maximum. Also, the data distribution, with the variance and the standard deviation. Table 1 displays the minors that were raped and Table 2 shows the number of adults that were raped. Table 3 shows a similar summary, however, the number of cases is being looked at compared with the actual number of victims that were raped. Table 5 displays the outlier that was found using standard deviation, and what was the upper and lower bound.Figure 1, shows a dot plot, representing the number of districts that are ruled by the State which are 29 and the 7 districts are ruled by Union Territory. Figure 2, shows a density plot, representing the total number of victims in State and the Union Territory. There seems to be a left skweness for the total number of victims ruled by both State and Union Territoy. The total number of victims for State seems to be concentrated at 500, and 1,500 victims for the Union Territories.

Table 1: Summary of Girls, Ages 0-17, that were Raped

	Age: 0-5	Age: 6-11	Age: 12-15	Age: 16-17	Total Child Vicitms
Min	0	0	0	0	0
1st Quartile	0	0	0.75	0.75	2.75
Median	0	1	6	9.5	20
Mean	7.806	21.03	100.44	132.75	262.03
3rd Quartile	4.75	15.75	35.5	68.25	115.75
Max	67	174	1143	1502	2841
Variance	285.0754	1875.056	54690.77	90104.31	342208.3
Standard Deviation	16.88418	43.30192	233.8606	300.1738	584.9857

Table 2 : Summary of Women, Ages 18 & Above, that were Raped

	Age: 18-29	Age: 30-44	Age: 45-59	Age: 60+	Total Adult Vicitms
Min	0	0	0	0	0
1st Quartile	18.5	8.5	1	0	31.5
Median	312.5	88.5	6.5	1	397
Mean	489.9	169.7	20.19	2.028	681.8
3rd Quartile	751.5	236.2	18.25	2.25	921.8
Max	2263	912	129	14	3305
Variance	385755.5	52829.6	975.0754	11.91349	760204
Standard Deviation	621.0922	229.8469	31.2262	3.451593	871.8968

Table 3: Rape Cases Reported in 2018 vs. Total Number of Victims Raped

	Cases Reported	Total Victims
Min	0	0
1st Quartile	58.75	58.75
Median	522.5	525
Mean	926.56	943.81
3rd Quartile	1121.25	1127.75
Max	5433	5450
Variance	1669134	1749456
Standard Deviation	1291.95	1322.67
Skewdness	0.9382461	0.9499093

Table 4 : Outliers

	Total Victims - Outliers
Upper Bound	4911.816
Lower Bound	-3024.205
Number of Outliers	1
Outlier	5450

Figure 1: Number of Districts ruled by the State & Union Territories

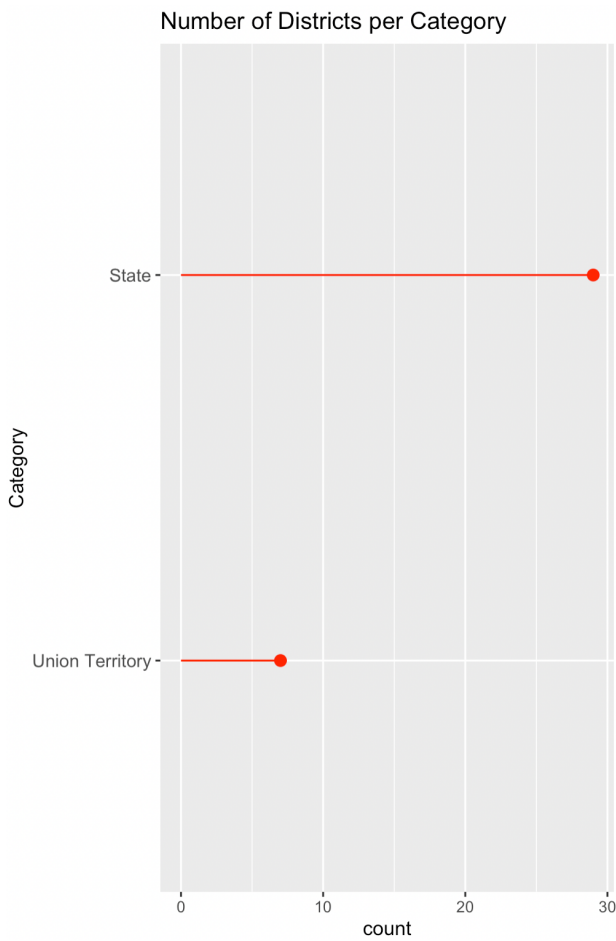Number of Districts per Category
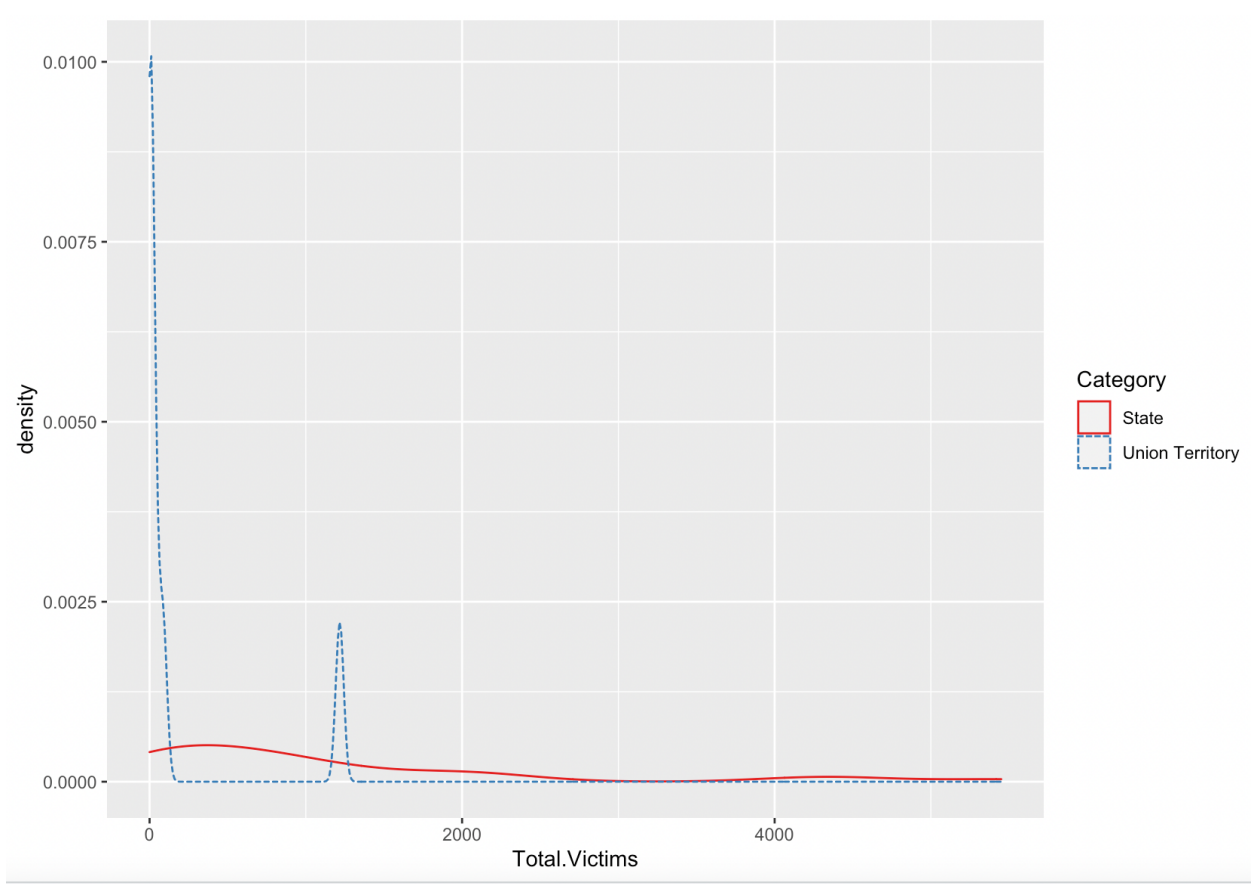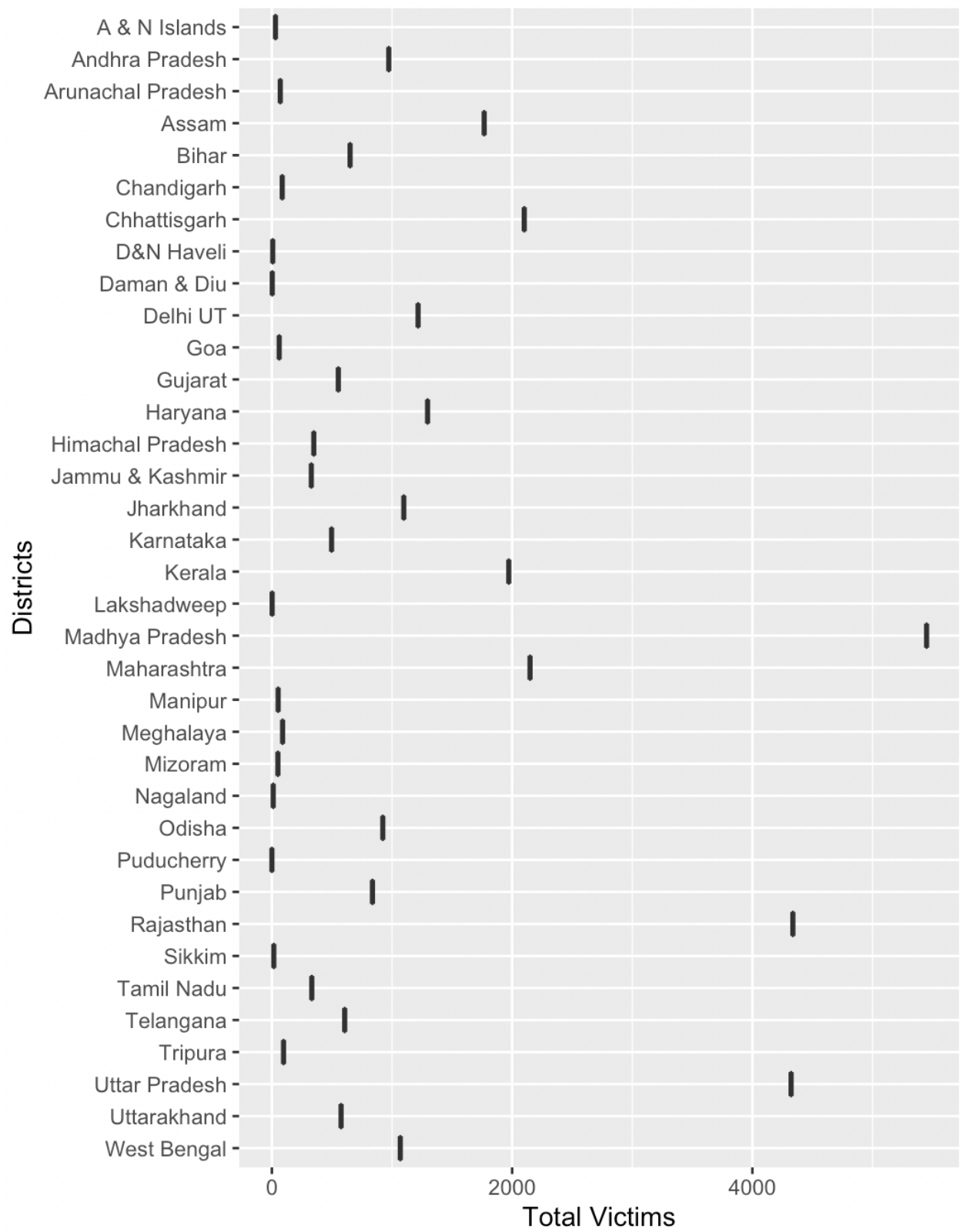
Figure 2: Total Victims in State vs. Union Territory

Figure 3: Total Victims per District

Conclusion and Discussion:

From Table 1&2, we can see that on average 262 girls (below age 18) were raped. While, an average of 681 women (above age 18) were raped. In total, the average number of victims in 2018 in India was 943. When looking at each age group, the highest number of victims for minors was in the age group 16-17. For adult women, age group 18-29 had the highest number of raped victims.

From Table 3, the skewness value is about 0.95 which is positive meaning there is a right skewness. We can also support this by looking at the mean which is greater than the median which supports the right skewness in data. Also, Table 3 shows that the average number of cases reported is less than the number of victims.The difference is about 20 cases, meaning that not all cases of rape are being reported. As for the outliers in the data, there was one found. As seen in Table 4, the value of the outlier was 5450, which is above the upper bound of 4991.86. The difference is not too major, and should be considered but also investigated. The value of 5450 is the number of victims raped in Madhya Pradesh. Figure 3, shows a break down of the total number of vicitms per district and as mentioned, Madya Pradesh has the highest number of victims. Also, in Figure 1 we see that there are less districts ruled by Union Terriory vs. the State. However, in Figure 2, we see that there is a higher concentration of victims in the Union Terrioty (1,500) vs. districts in State (500). This needs further investigation, as in why are districts ruled by Union Teriiory have a higher number of vicitms, is there a different governing protocol or standards that are causing this?

Overall, the numbers of victims raped is high, whether it is minors or adults. Especially since this is data of just one year and one country. The highest number was in Madhya Pradesh, and this should be investigated further. Why are there the highest number of cases there? Is it because it is in the middle of India? Does the location have something to do with the number? In Figure 3, we see that districts such as D&N Haveli and Daman& Diu have lower number of victims. Why is that and what are they doing differently? Is it because of population, security, awareness, or another reason? Also, for further investigation, variables such as time the rape occurred and where. Such as in school, transportation, traveling, etc. When looking at these factors, changes such as increasing security can be done to lower the number of cases. Regardless, with the high number of cases, there is a strong need of awareness and action. This is just in one country, for more knowledge, one can look at the number of cases worldwide.

References:

Prabhavalkar, N. (2021). *Crimes in India-2018 Victims of Rape.* Kaggle.