Approach

# Step-1 (Data Retrieval)

First, I had a look at the websites mentioned in the resources given by my mentor. Initially it was difficult to find relevant un-visualised data available on the website but upon further research I found each state had published an annual report of the number of deaths caused by the black spots present in each state. They had published pdfs enlisting each black spot with its chainage and nearby police station along with the district in which it is present. There were a few states for which state government websites were non-existent or non-functional, however data.gov.in came in helpful with a consolidated data set of remote areas including UTs and Northeastern States.

# Step-2 (Processing)

Now to perform any manipulation or geocoding, the file must be in CSV format and all the published reports were in PDF format. Hence online tools and extensions like smallpdf.com were used to convert the pdfs into excel and then saved as CSV format. There were minor changes needed to be done before saving it into CSV format like changing the headers and footers that automatically crept in due to formatting and remove any coloured headings and extra text. Note, no changes to the data were made.

# Step-3 (Working on the data)

Once we have the final file ready, we first run a python command to convert the CSV file to a data-frame. Now for processing the data, we first check that there exist no null values in the main ‘Location of black spots’ column. Generally, there were no null values apart from places which have a different regional language name. Now knowing that it is not possible to map every location with the exact specifications on the map we need to set up a substitute array that can fetch results in case the exact location fails to yield its geocode. Then these arrays are inputted in a function that converts them into viable query names and removes white space and escape sequences. Although this step is not required it gets slightly more precise results. Now we add each individual name from the arrays into the specific URL format to fetch results from Google’s Geocoding API and return the latitude and longitude for each place. In case the main location is inaccessible we query the substitute array(decided for each data set) and if the substitute fails too then we append null value. After retrieving all the values we export it into the required file to out desired destination.

# Step-4 (Additional Visualisations for Local Area)

For my local state Maharashtra, I used Google’s Places API to first query in all the accident black spots and then get all nearby hospitals and police stations. I implemented a condition under which the business status is ‘OPERATIONAL’ with the radius as 5000 and then further used GeoCoding API to get the latitude and longitude. For local additional visualizations I studied the data from two data sets included in the repo acquired from data.gov.in. The data set helped me generate more insights for the accidents including some helpful graphs of which some are included in the report.

# Step-5 (QGIS Plotting)

For QGIS visualisation I inserted a delimited layer for each state and then imported the concerned CSV file for the state. After the points were plotted, the layer was then converted to a ESRI shapefile and then exported. This process was repeated to plot the entirety of the country.

For further insights refer the project report.