Qlik Sense Analysis of Road Safety And Accident Patterns in India

-Varundeepak B A

1. Introduction
- 1.1 Overview
- 1.2 Purpose
- 1.3 Technical Architecture

2. Define Problem / Problem Understanding
- 2.1 Specify the Business Problem
- 2.2 Business Requirements
- 2.3 Literature Survey
- 2.4 Social Impact

3. Data Collection
- 3.1 Understanding the Data
- 3.2 Connect Data with Qlik Sense

4. Data Preparation

5. Data Visualization

6. Dashboards / Responsiveness

7. Storytelling / Report

8. Performance Testing
- 8.1 Amount of Data Rendered
- 8.2 Utilization of Data Filters

1. Introduction

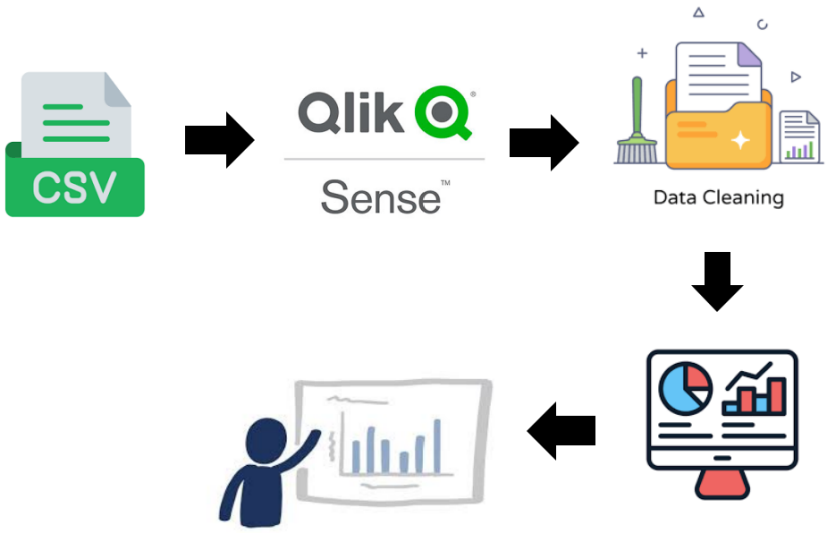
1.1 Overview

This project aims to perform an analysis of road safety and accident patterns in India using available data. By using advanced data visualization techniques in Qlik Sense, we try to find valuable insights into the causes, severity and such other cases of road accidents across the country. This analysis will look into the impact of accidents on various road users, examine the influence of factors like weather conditions and identify high-risk regions. Through interactive and visually attractive dashboards, the project aims to provide a data-driven understanding of the road safety situation in India.

1.2 Purpose

The main purpose of this project is to view the critical issue of road safety in India and contribute to the efforts in reducing the high number of accidents and fatalities in India. By using data analytics and visualization, the project tries to provide insights that can help make informed decision making. The visualizations and dashboards can help identify high risk areas, understand the root causes of accidents. The insights that we get from this analysis can contribute to the development of effective strategies to improve road safety and reducing accidents.

1.3 Technical Architecture



This project uses Qlik Sense, an advanced data analytics and visualization platform to process, analyze and visualize the road accident data. The datasets are in CSV format. These data are loaded into Qlik Sense, and then we do data preprocessing such as data cleaning, handling missing data, creating new data from existing ones if required etc. We do visualizations and dashboards with the extensive options available in Qlik Sense. Then, we do storytelling with the analysis, explaining what the data tells us through the charts and dashboards.

2. Define Problem / Problem Understanding

2.1 Specify the Business Problem

Road accidents are a major public health concern in India, resulting in a huge number of deaths, injuries and economic losses every year. The business problem address by this project is the urgent need to improve road safety and reduce the number of accidents in the country. We try to solve this problem by providing insights and recommendations to people involved in road safety management. By analyzing accident patterns, finding high risk regions and road users categories, and understanding the factors leading to accidents, the project tries to come up with solutions to these problems. The insights we obtain from this project will help making decisions and raise awareness related to accidents in the country and how to avoid them. Ultimately, this project aims to contribute to creating safer roads and reduce the number of accidents in the country.

2.2 Business Requirements

To address the road safety challenges in India, the project has identified some main business requirements. Firstly, there is a need to identify major causes of road accidents, such as over-speeding, drunk driving etc. This will help find related solutions to solve the main causes of accidents. Secondly, the project aims to analyze the regional variations in accident patterns and their severity across different states and union territories. This analysis will allow us to identify high risk areas and we can focus on those areas. Thirdly, understanding the impact of accidents on different road user categories such as pedestrians, two wheeler riders, car drivers etc. is important to identify and develop specific solutions. Fourthly, examining the relationship between accident severity and external factors like the weather will provide insights into how they affect the road users. Lastly, the project tries to use the insights into actions and visually attractive dashboards that can explain the findings to people involved and help in decision making.

2.3 Literature Survey

A comprehensive literature survey was done to review existing research works, reports and studies related to road safety and accident analysis in India. This was done to gather insights from various sources to understand the current state, identify main challenges and decide the approach for this project. This literature survey covered a wide range of souces including academic papers published in peer-reviewed journals, government reports and publications from the Ministry of Road Transport (MoRTH), reports and studies by international organizations such as World Health Organization (WHO) and the World Bank, and research conducted by road safety organizations. The literature survey focused on several key themes related to road safety and accident analysis in India such as identification of major accident causes, analysis of regional accident patterns, impact of accidents on different road users etc. Some of the key findings from the literature survey include:

- Over-speeding, drunk driving, driver distraction are consistently among major causes of road accidents in India.
- Pedestrians, bicyclists and two wheeler riders are among the most vulnerable road user groups.
- Inadequate road infrastructure, lack of safety features in vehicles and limited enforcement of traffic rules contribute to high number of accidents.
- Data-driven approaches like geospatial analysis and machine learning techniques have shown improvements in understanding these accident patterns.

By using the findings from the literature survey, this project tries to contribute to existing knowledge and provide actionable insights to make informed decision-making.

2.4 Social Impact

Road accidents can have strong impact on people, families and the community. Fatalities and injuries from road accidents can lead to permanent disabilities and can also lead to deaths, which would affect the families also. Th economic impact of road accidents is also very large, including costs related to medical treatments, property damage etc. Road accidents mostly affect vulnerable road users like pedestirans, cyclists and motorcyclists who do not have much safety features with them, and such users might often be from lower socio-economic backgrounds. This also shows the social inequalities and the need to have proper solutions for these groups. The social impact analysis in this project aims to show the broader society impact of road accidents. By understanding this, policy makers and the people involved in making decisions can look for solutions. The project's visualizations and insights related to the social impacts of road accidents will contribute to raising awareness and develop

focused solutions to improve the road conditions and safety for all types of people who use the roads in India.

3. Data Collection

3.1 Understanding the Data

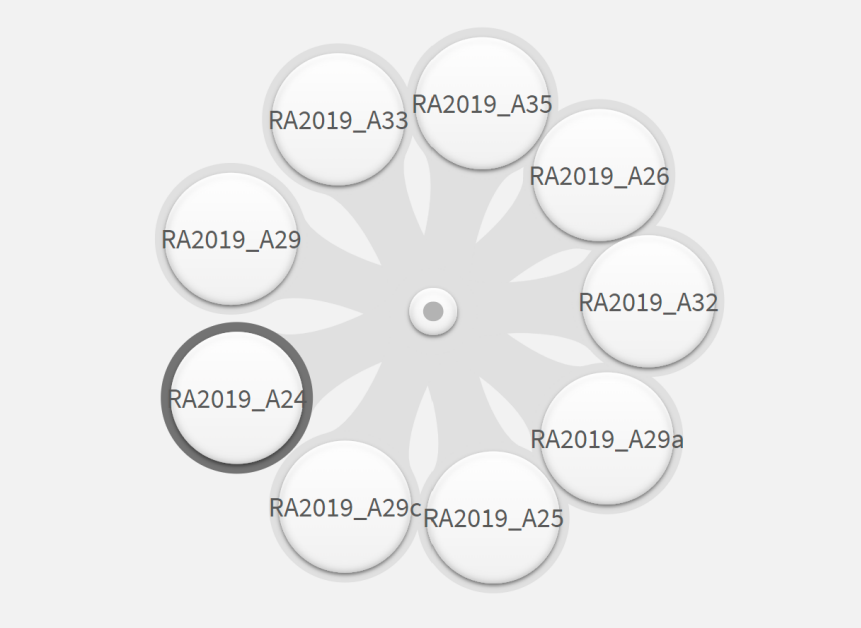
We have 9 CSV files that have different kinds of data related to accidents and road users like the gender, age groups, weather conditions etc. Let's have a look at each of the files and what data they have.

- RA2019_A32.csv - This file has the Pedestrian age group and gender across the states who were involved in an accident.
- RA2019_A33.csv - This file has the Pedestrain age group and gender across the states who were killed in an accident.
- RA2019_A29c.csv - This file contains information on pedestrians killed in accidents, classified by the type of impacting vehicles such as bicycles, two wheelers, cars etc.
- RA2019_A24.csv - This file contains the data on accidents classified by type of traffic control present at the accident location such as traffic light signals, police control etc. and they it includes the number of accidents, persons killed, and persons injured (grievously injured, minor injured, total injured) for each type of traffic control.
- RA2019_A25.csv - This file has the data on accidents classified by the weather conditions during the accident, and has information on the number of accidents, persons killed and person injured (grievously injured, minor injured and total injured) for different weather conditions like sunny/clear, rainy etc.
- RA2019_A29a.csv - This file contains data on two wheeler riders killed in accidents classified by impacting vehicles such as bicycles, two wheelers, cars etc.
- RA2019_A29.csv - This file has the data on road users killed in accidents classified by the gender and road user categories such as pedestrian-male, pedestrian-female, bus occupant-male, bus occupant-female etc.
- RA2019_A35.csv - This file has the data on accidents classified by the cause of the accident such as over-speeding, drunk driving etc.
- RA2019_A26.csv - This file has the data on accidents classified by severity of accident and type of vehicle involved such as persons killed, grievously injured, minorly injured for vechicle types such as bicycles, two wheelers, cars etc.

All these files have the States/UT as the common field through which we could link these files.

3.2 Connect Data with Qlik Sense

We first load the datasets to Qlik Sense to the Data Catalog so that we can use those data whenever needed.



Then, Qlik Sense reads the data and it is smart enough to figure out that all the tables have a common field (States/UT) and we can join the tables through that common field. So, Qlik Sense makes Associations between the tables through this common field. We can also make our own associations and break these if we want to, but since we want this, we leave it as is is.

RA2019_A29 — States/Uts — States-Uts-State-UT — States/UTs — RA2019_A24

States/Uts	Pedestrian - Male	Pedestrian - Female	Pedestrian - Total	Bicycles - Male	Bicycl
Andhra Pradesh	1331	392	1723	125	2
Arunachal Pradesh	6	2	8	2	0
Assam	593	125	718	99	7
Bihar	1000	259	1259	350	69

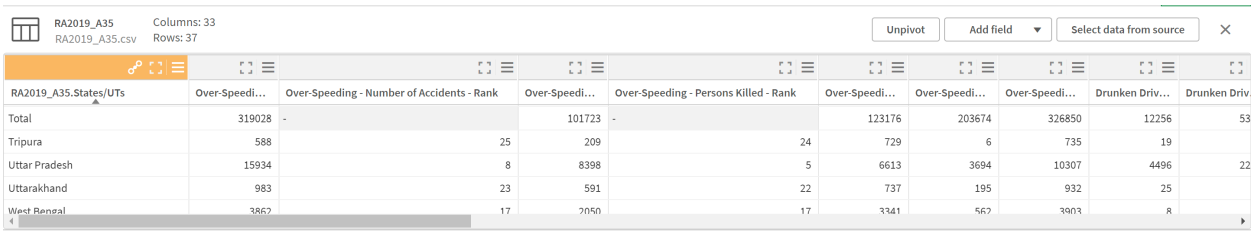
States/UTs	Traffic Light Signal - Total number of Accid.	Traffic Light Signal - Persons Killed	Traffic L
Andhra Pradesh	1197	329	484
Arunachal Pradesh	0	0	0
Assam	107	31	77
Bihar	0	0	0

Hide data preview

We see here that Qlik Sense has made an association between tables RA2019_A29 and RA2019_A24 by using States/UTs from both the tables.

4. Data Preparation

We now start with our data preprocessing section, in which we try to have a look at our data and try to clean the data, handle null values, create or delete data as required etc.



RA2019_A35.States/UTs	Over-Speedi...	Over-Speeding - Number of Accidents - Rank	Over-Speedi...	Over-Speeding - Persons Killed - Rank	Over-Speedi...	Over-Speedi...	Over-Speedi...	Drunken Driv...	Drunken Driv
Total	319028	-	101723	-	123176	203674	326850	12256	53
Tripura	588	25	209	24	729	6	735	19	
Uttar Pradesh	15934	8	8398	5	6613	3694	10307	4496	22
Uttarakhand	983	23	591	22	737	195	932	25	
West Bengal	3862	17	2050	17	3341	562	3903	8	

Let's take the table RA2019_A35 for this case. We see that the column 'Over-Speeding - Number of Accidents - Rank' shows the rank of each states based on no. of accidents caused by over-speeding. We have a record which has the value as 'Total' in the States/UT field, which usually has the sum of the numbers of all other states. Here in the Rank columns, we see that 'Total' has null value, because understandably, we don't have a state called Total, so it can't have a rank, and also, since it is always the sum of all the other states, it will always have the highest numbers, so if we give a rank to it, it will always be ranked as 1, so we don't give it a rank. In the initial dataset, it has the value of 'NA', and using Qlik Sense, we replace the 'NA's with the native null value. Since we are not going to use this rank column for the 'Total' value, we don't have to replace it with any other values.

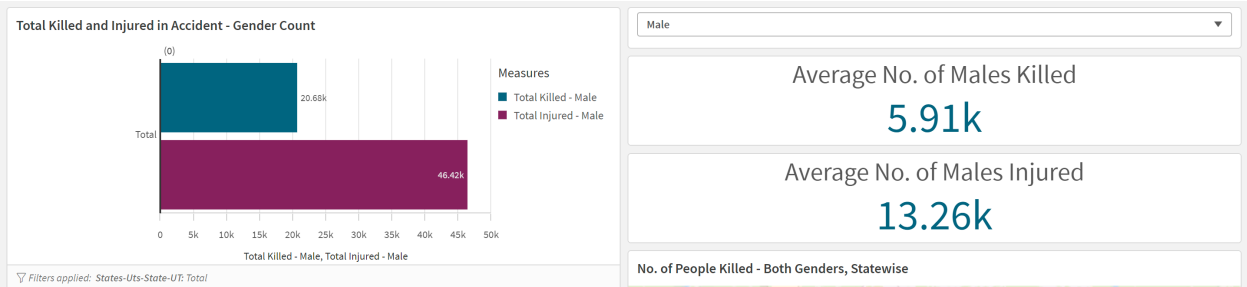
I am not creating any new fields here using Calculated fields because I do not want to modify the existing original dataset. So, if needed, I've created a Master Item and used that wherever needed instead of creating new fields in the dataset. I've also not deleted any fields from the dataset since everything here could be of value.

5. Data Visualization

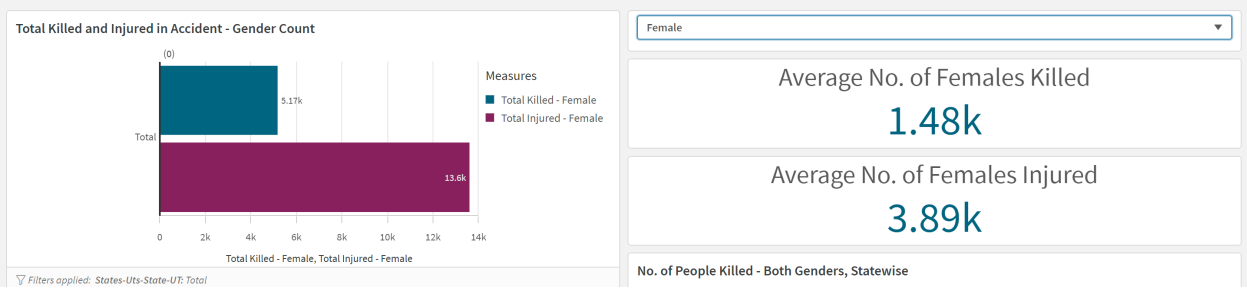
Now that we have finished the data preprocessing, we can start with the visualizations and dashboards. We can create sheets that can show charts, or we can make them interactive so that the charts and visualizations change as per the selections and our needs. Some of the common types of visualizations that can be used are bar charts, line charts, maps etc. Each types of chart has their own uses and have to be used accordingly.

First, we create a new sheet and in that, we make some charts and KPIs to view the gender and age distribution of the accidents. We also have an option to select between Male and Female and the charts change according to the gender selected, but that is not done using Filter Pane since Male and Female are included only in the column names and not as the column values themselves, we cannot use the gender directly in the filter pane since they are not actual dimensions present in the tables. We also don't use Filter Panes anywhere in any of the charts because States/UT is the only dimension available to us, and if we filter charts based on the States/UT, then in bar charts, we

would have only one bar, in line charts, we would have only one line etc. So, we do not use filter panes anywhere and we use other methods to filter data.



Here, we see the total and average number of males killed and injured in total across all states by using bar charts and KPIs. We can also see the numbers in the bar chart, showing us a more detailed view.



Now, we have changed the filter to Female and now the same bar chart and KPIs show the total and average number of females killed and injured in total across all states. As mentioned before, we do not use Filter Panes since we do not have the genders as dimensions in our data. So instead, we use Variables and we manipulate this variable to change the data in the charts.

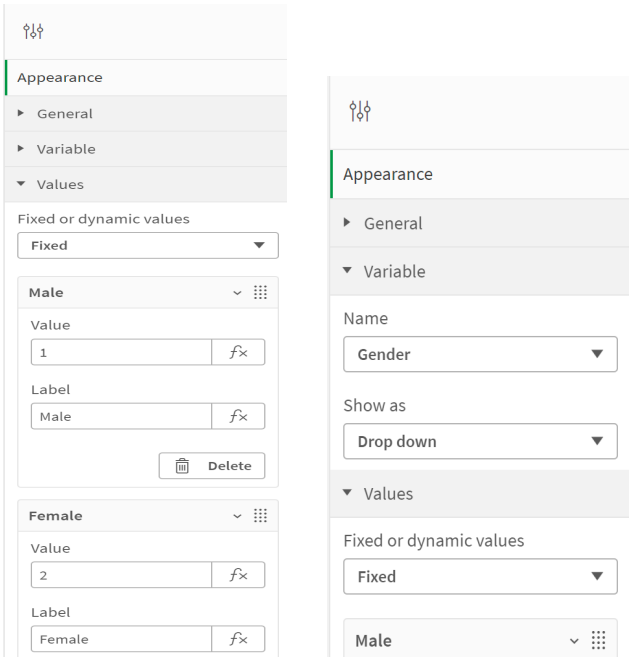
Also, we have created Master Items TotalKilledMale and TotalKilledFemale since we would be using them in multiple other places and I did not wish to modify the existing dataset.

Edit expression

```
Sum([18-25 Years - Killed - Male])+Sum([25-35 Years - Killed - Male])+Sum([35-45 Years - Killed - Male])+Sum([45-60 Years - Killed - Male])
+Sum([60 and Above - Killed - Male])+Sum([Age not known - Killed - Male])+Sum([Less than 18 years - Killed - Male])
```

Similarly, we can create other Master Items like TotalInjuredMale and TotalInjuredFemale and others as required. Now, I have created a Variable named 'Gender' and I use this variable to manipulate the measures in the bar chart adn KPIs.

The drop down box we see is actually a Custom Object called 'Variable Input' which allows the user to give their own inputs and change the charts as wanted. Here, we make that variable input as a drop-down box and we have the values Male and Female in them. We give the value for Male as 1 and Female as 2(no bias here, just randomly giving the values). We also associate the variable Gender to this variable input option, and we choose drop-down option so that it shows as a drop-down box for us.

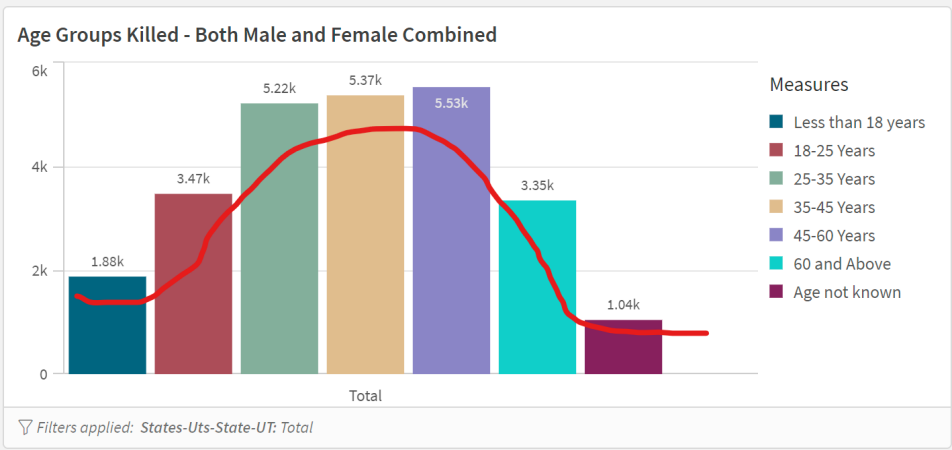


Now in the bar chart, in the measures section, we include the expression,

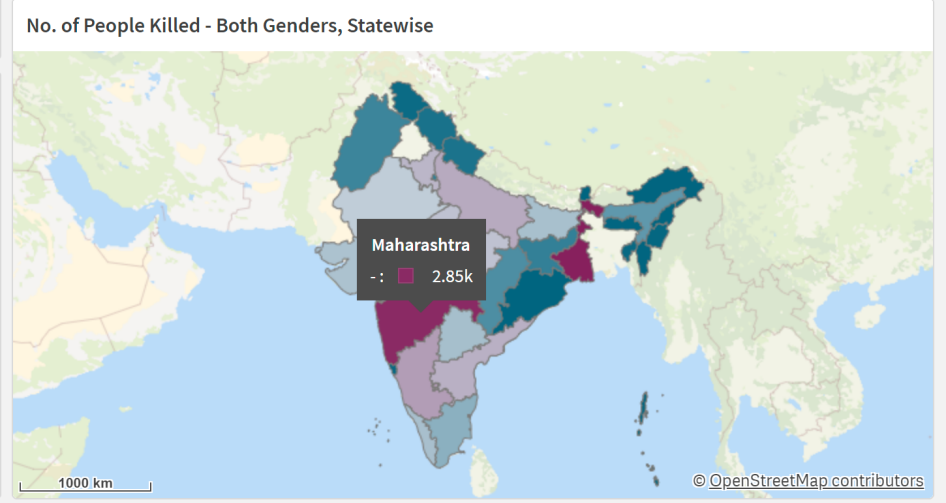
Edit expression

```
1  Pick(Gender, TotalKilledMale, TotalKilledFemale)
```

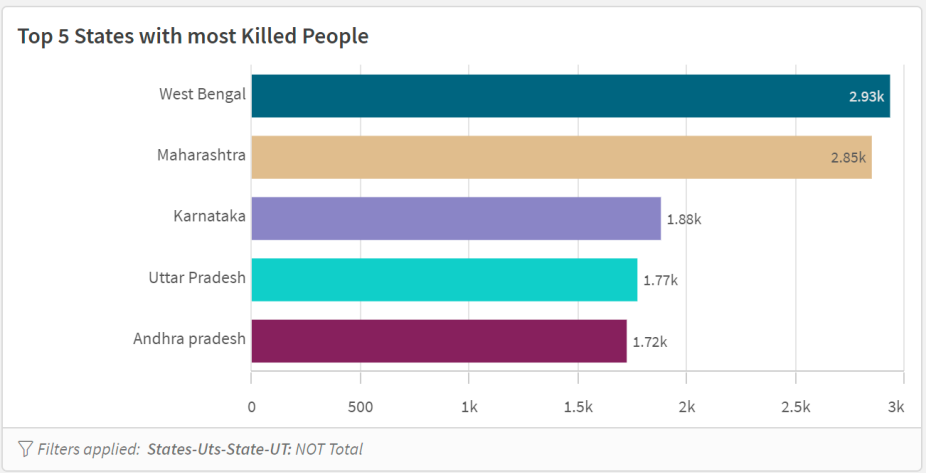
The Pick function takes the value of the first parameter (Gender here) and returns the value of the nth value in the following list of values. So here, if the value of Gender is 1 (Male is selected in the drop-down list), then the value in the 1st position in the list of values is selected (Here TotalKilledMale), so that value is showed in the bar as the measure. Here we have used the Master Items we created before, so that instead of having to type that long expression everywhere, we can just use this master item instead. This way, without using the Filter Pane, we can still do filters using variables.



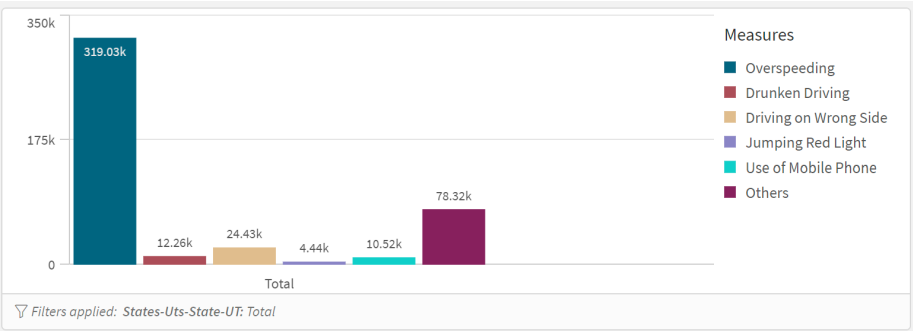
Here, we see the age distribution of the people killed in road accidents as a bar chart. We can see a bell-shaped (normal distribution) curve in these bars. So we can understand that most of the accidents happen around the data in the center, which is the 35-45 Years age group with 5.37K people killed, which is very close to the actual highest age group, which is the 45-60 Years which is 5.53K people killed.As with normal distributions, the age group at the center would be the mean and the age groups at one standard deviation away would be the majority of the people, and any age groups outside this would normally be considered outliers, but in this case, the counts for the remaining two groups are not that low, so we don't consider them as outliers.



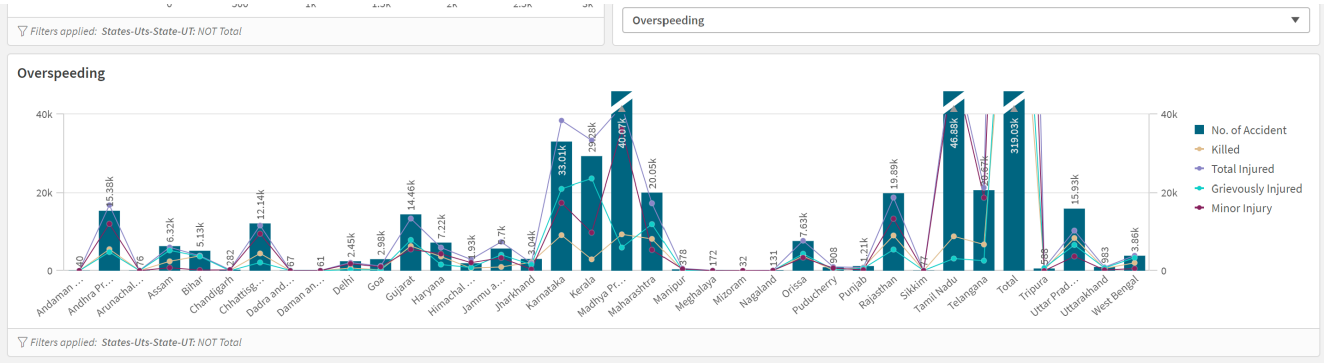
Next we have a Map of the country with total number of people killed across each states. By hovering the mouse over the states, we can see the count of the people killed in that state. Red areas show that there is a very high number of people killed there, while considerably cooler colors like a dark blue show that there is a very low number of deaths there.



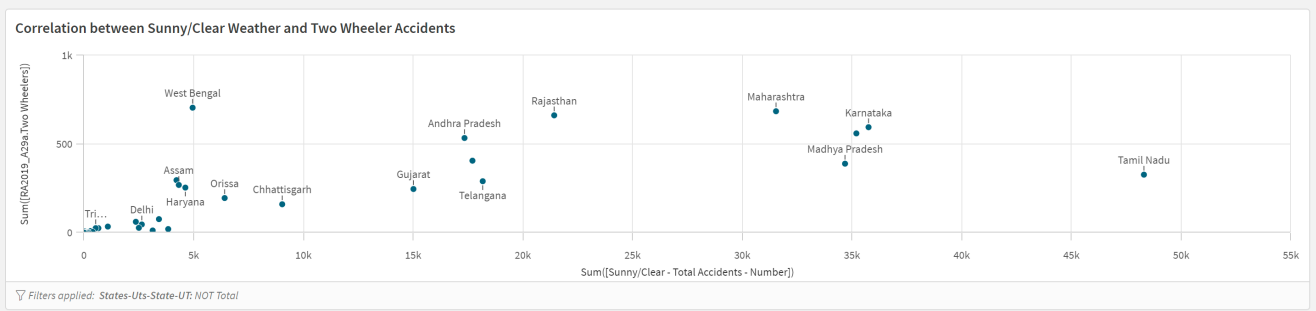
Next, we have the bar chart of top 5 states with the most number of people killed. The names of the states can be seen in this chart along with the count.



Here, we have the bar chart showing the causes of the accidents and their counts. We can see that Over-speeding is widely higher than the other reasons, almost 4 times higher than the next reason 'Others' which are reasons other than the ones mentioned here, and around 71% of the total deaths caused.



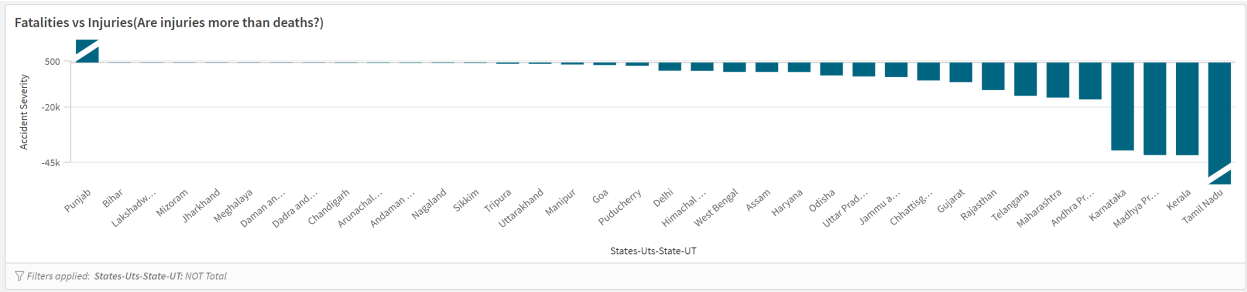
Now, we see a Combo chart which shows the severity of the accidents across the states for each of the causes for the accidents. Here, we see for over-speeding, and here too, we use a drop-down box to choose the type of cause we would want to see for.



This is a Scatter Plot showing if there are any correlation between Sunny/Clear weather and Two wheeler accidents.We say there is a positive correlation if the total number of accidents increase as the accidents due to sunny/clear weather increases, and we say there is a negative correlation between them if the total number of accidents decrease as the accidents due to sunny/clear weather increases.

It seems like there is a positive correlation between these two, even if we have some states like West Bengal and Tamil Nadu as outliers. The general trend we see here is that, as the no. of accidents due to sunny/clear weather increases, the total no. of two wheeler accidents also increases.

What we can understand from this and the previous chart with over-speeding as the major cause for accident is that, if the weather is sunny/clear, then people tend to over-speed, causing accidents. Most people would not over-speed during rain or hail, so during those times, we might not have as many accidents due to over-speeding, but accidents due to any other causes are still possible.



Now, we see a diverging bar chart which shows the rate of fatalities vs injuries. It shows the difference between the number of persons killed and the sum of persons grievously injured and minorly injured for each state. A positive value shows that there are more killed than injured, and a negative value shows that there are more injured than killed.

Edit expression

```
1  ={<[States-Uts-State-UT]-={'Total'}>}[Total - Number of Persons - Killed] -
2  ({<[States-Uts-State-UT]-={'Total'}>}[Total - Number of Persons - Greviously Injured] +
3  {<[States-Uts-State-UT]-={'Total'}>}[Total - Number of Persons - Minor Injured])
```

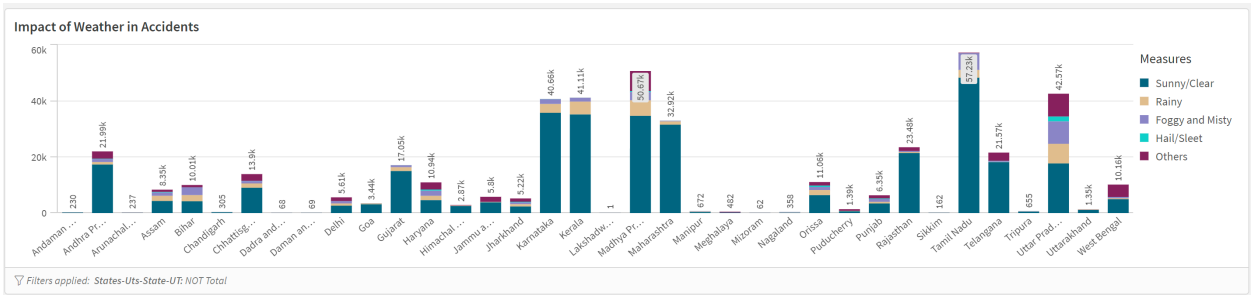
We have also used a Set Expression here that excludes the States/UT record which has the value as 'Total' in them.

Now, we have a Sankey chart of the accident severity of each states. Here, we see that the states 'flow' into one of the accident severity, and based on that, we can say which type of accident severity is the most common in each state. For example, we can see that in Karnataka, the most common accident severity is 'Grievously Injured' while in Uttar Pradesh, the most common accident severity is 'Killed'.

Edit expression

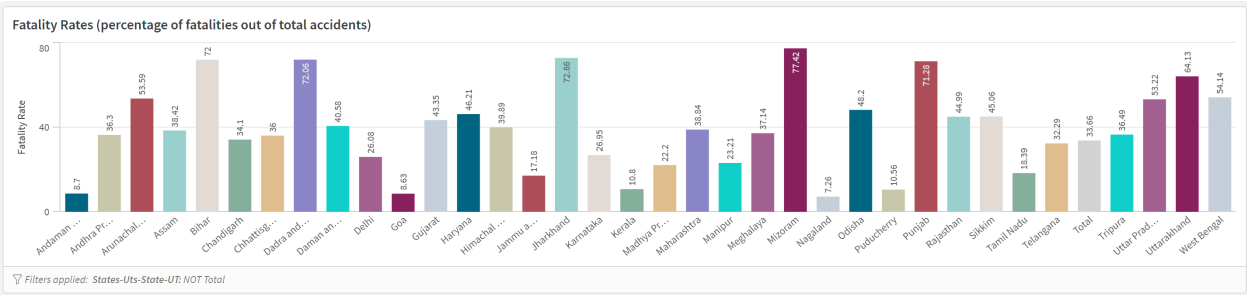
```
=if([Total - Number of Persons - Killed] >= [Total - Number of Persons - Greviously Injured] and [Total - Number of Persons - Killed]
>= [Total - Number of Persons - Minor Injured], 'Killed',
  if([Total - Number of Persons - Greviously Injured] >= [Total - Number of Persons - Killed] and
  [Total - Number of Persons - Greviously Injured] >= [Total - Number of Persons - Minor Injured], 'Grievously Injured',
    if([Total - Number of Persons - Minor Injured] >= [Total - Number of Persons - Killed] and
    [Total - Number of Persons - Minor Injured] >= [Total - Number of Persons - Greviously Injured], 'Minor Injured', 'Unknown'))
```

This was achieved by using a series of nested if expression and comparing each of the accident severity with the other two and choosing the one with the maximum count.

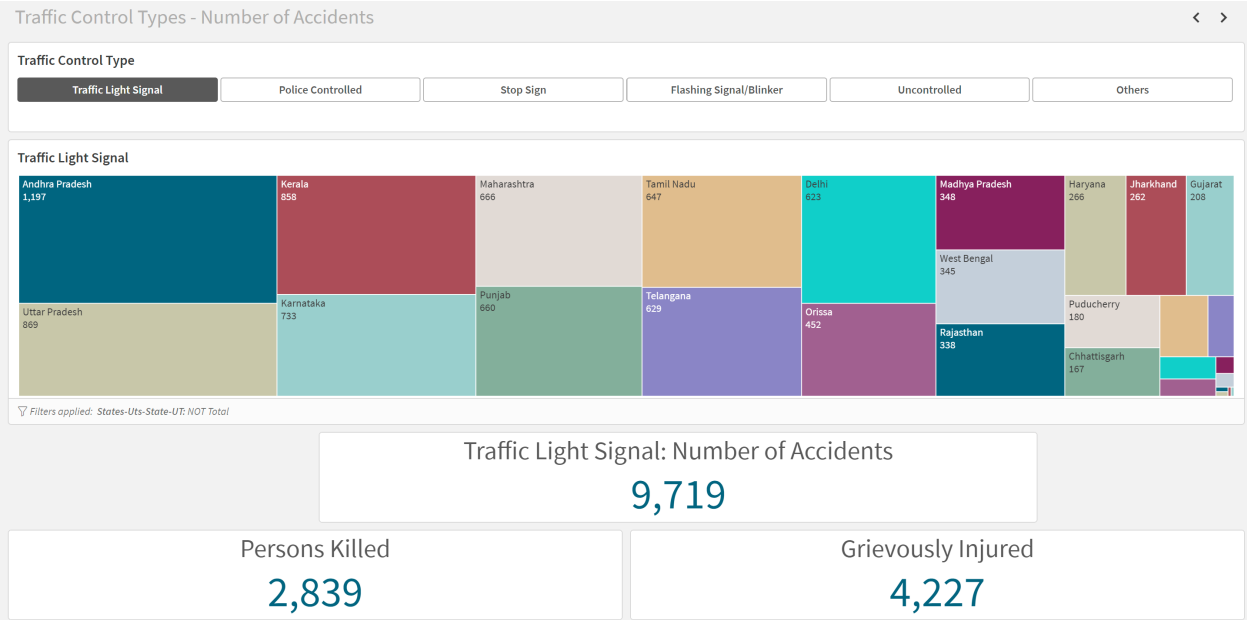


Next, we have a stacked bar chart showing the impact of weather conditions in the

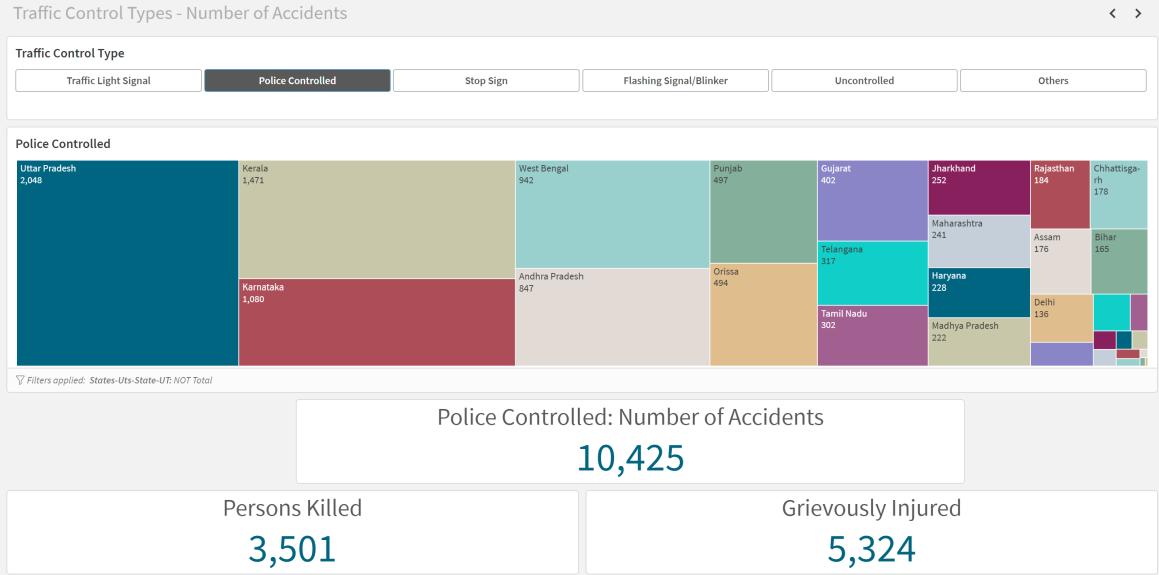
accidents. We can see that Sunny/Clear weather has caused the most number of accidents across the states, while also seeing the contribution of the other weather conditions at the same time.



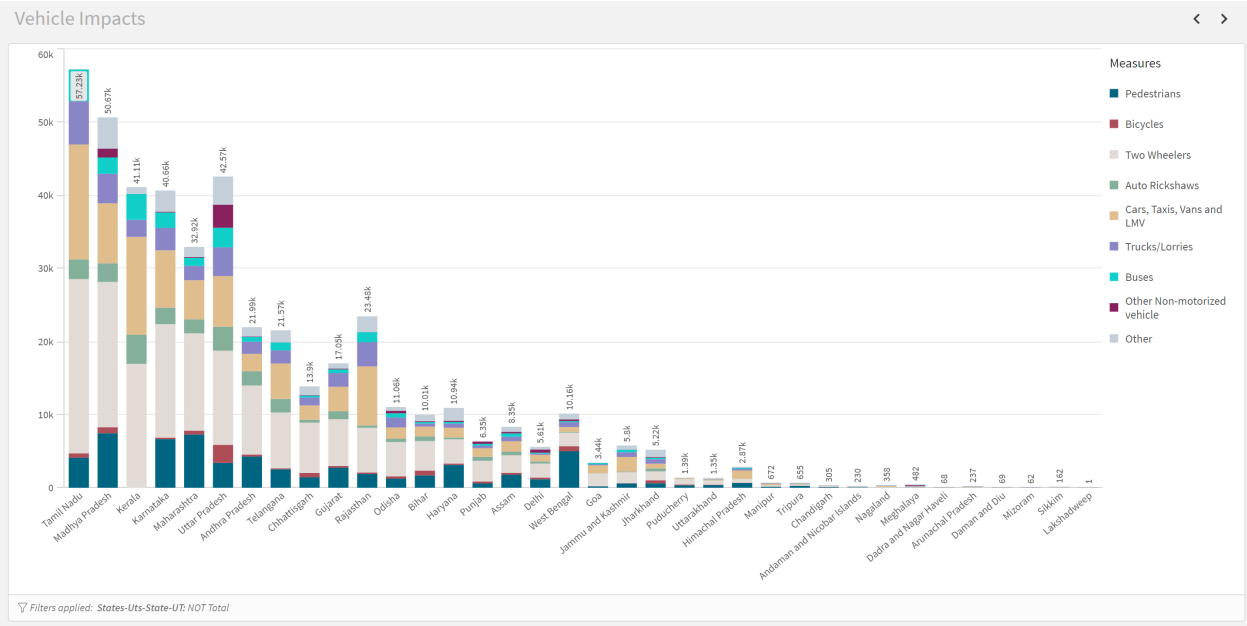
Here we have the Fatality rates across each states. This shows what percentage of the accidents ended up as deaths. We see that the highest Fatality Rate is in Mizoram, with 77.42% of the accidents resulting in deaths.



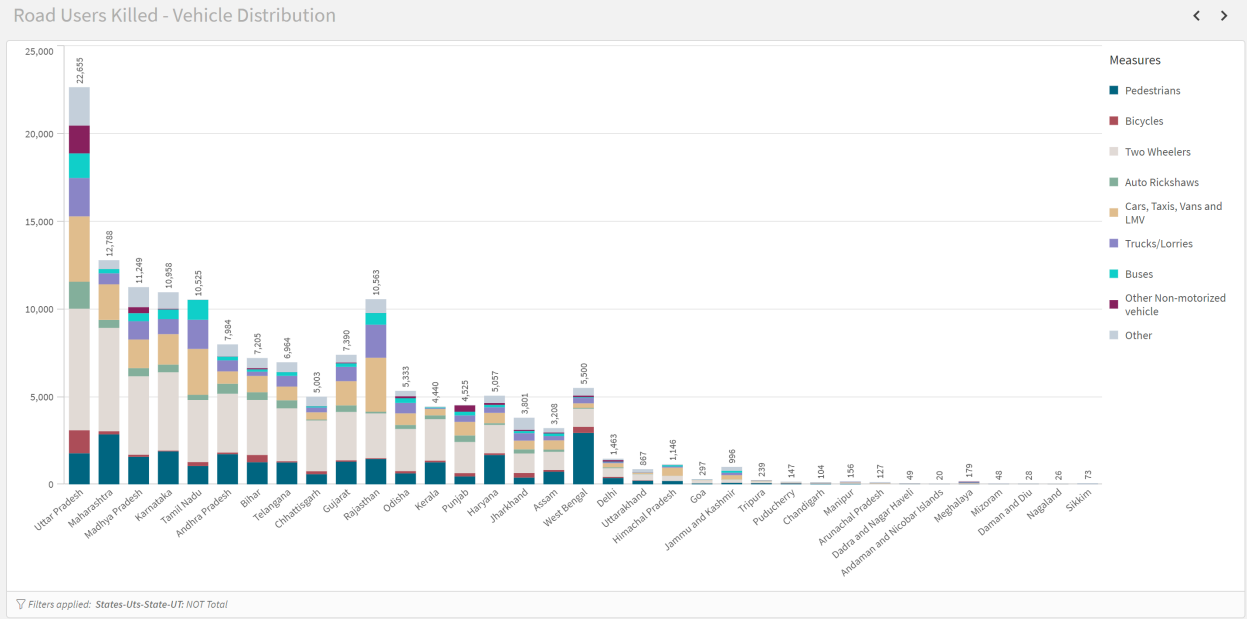
Here, we have a Treemap and KPIs working together with another variable input selector, this time as buttons to show the number of accidents across states involving different types of traffic controls. We can see that for each control type, we can see the states with the highest accident count having larger boxes and the next highest state with the next biggest box and so on, and the KPIs show the count of number of accidents, people killed and grievously injured for that traffic control type. If we change the traffic control type, we can see that the treemap and the KPIs change according to that.



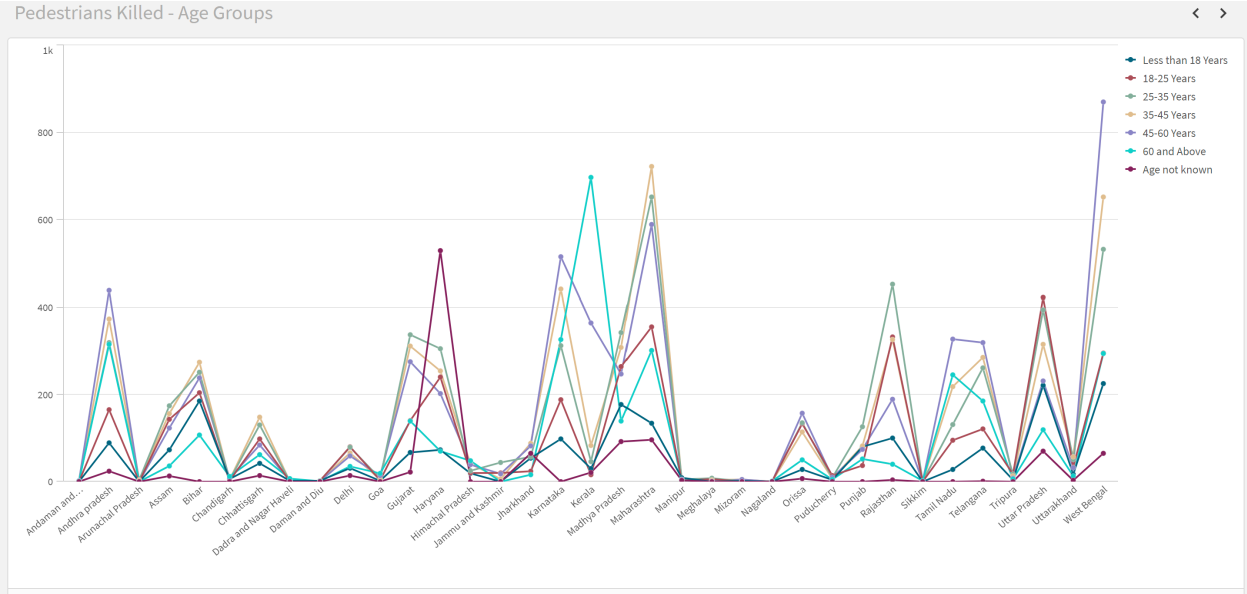
Now, after clicking Police Controlled type, we see that the map and the KPIs have changed.



In this stacked bar chart, we see the vehicle contribution towards total accidents. We can tell that the vehicle which contributed the most was two-wheelers. Given that most of the road users in India are two wheelers, and now a days people are driving recklessly with two wheelers, we can easily say that two wheelers are contributing the most to accidents these days.

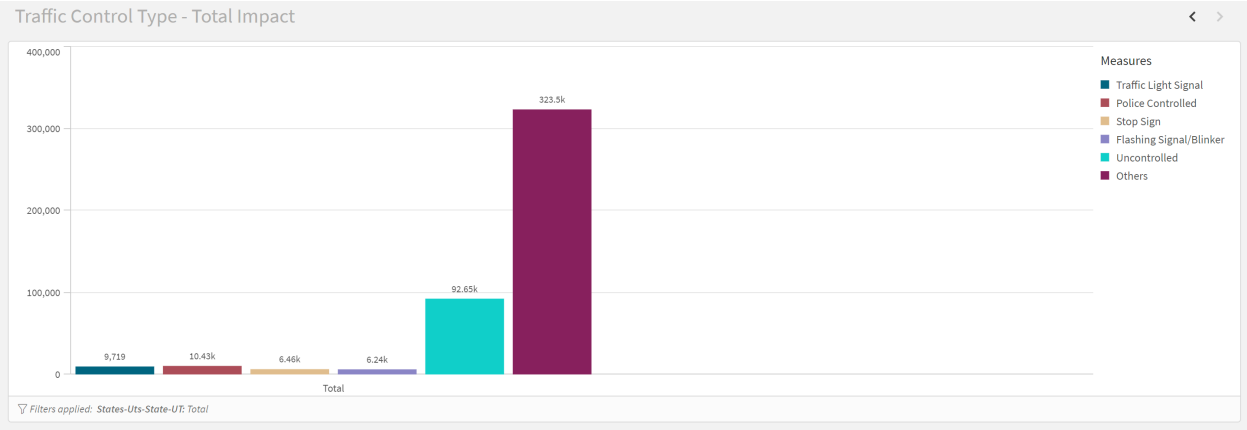


In this stacked bar chart, we see the vehicle distribution of road users killed based on the vehicle they used. Here also, the leading vehicle users that were killed on road are the two wheeler users. Just as before, since most of the road users are two wheeler users, they are the ones who get killed the most in the roads.



In this Line chart, we see the pedestrian age groups that were killed across each of the state. Previously we saw a bar chart of the total people in each age group that were killed, but now, we see it across each state. While line charts are usually used to show the trends across the data, since we don't have any temporal data here, we just use a line chart instead of a combo chart to show some variety here. Now too we see that the

most killed age group across the states would be 45-60 Years, closly followed by 35-45 Years and then 25-35 Years, confirming our previous bar chart showing the people killed by their age groups.

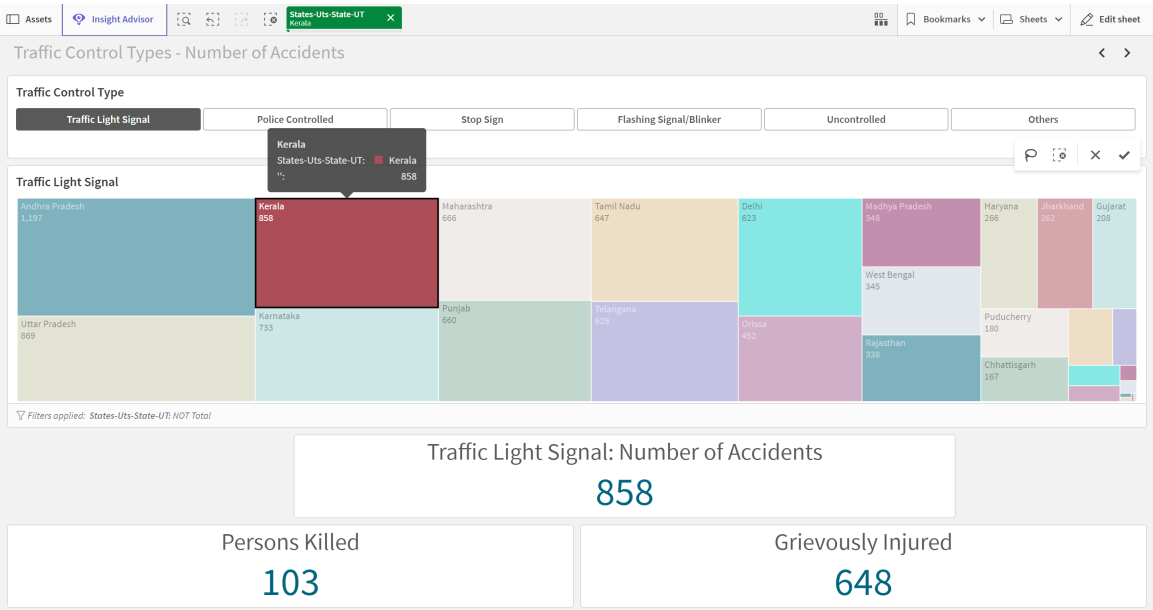


Finally, we have this bar chart which shows the total count of accidents happened around each of the traffic control types. We see that 'Others' has the most count, meaning that there are many other reasons contributing to the traffic control type other than the ones mentioned.

6. Dashboards / Responsiveness

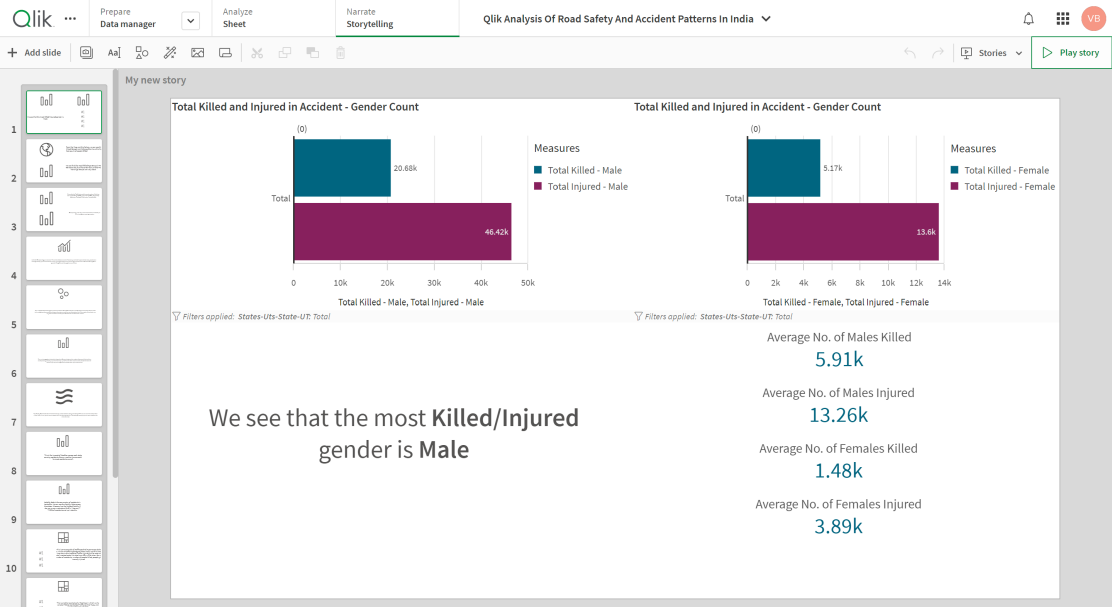
Instead of having just static charts, we also have interactive dashboards as seen before that changes the charts based on the option selected, providing a better way to understand more data with few charts. This also gives us some control over the kind of data we would want to see from the visualizations we already have. This would be useful if we embed these dashboards in a webpage so that any users that visit the webpage can interact with these dashboards and even possibly work with real-time data.

As seen before, we have used drop-down boxes and buttons to change the data in the charts as required. Now, in the Treemap, we can also see that by clicking the name of the state, we can see the data of only that state in those KPIs. The best thing about this is, Qlik Sense automatically does this for us, so we don't have to modify any of the settings in the charts to do this. We see that after clicking on the state 'Kerala', only the data related to that state is now displayed in the KPIs.



7. Storytelling / Report

Storytelling is very much similar to the presentations we do with slides, images etc. As the name suggests, storytelling is used to tell 'stories' with the data, explaining it visually to people who might not be very well versed with the numbers and technical jargons being used at many places. Our visualizations and dashboards should be in such a way that even peopl who do not know much about things should be able to make out what is being said to them easily. Similarly, we use Qlik Sense's storytelling feature to make our visualizations into slides that explain what the charts are.

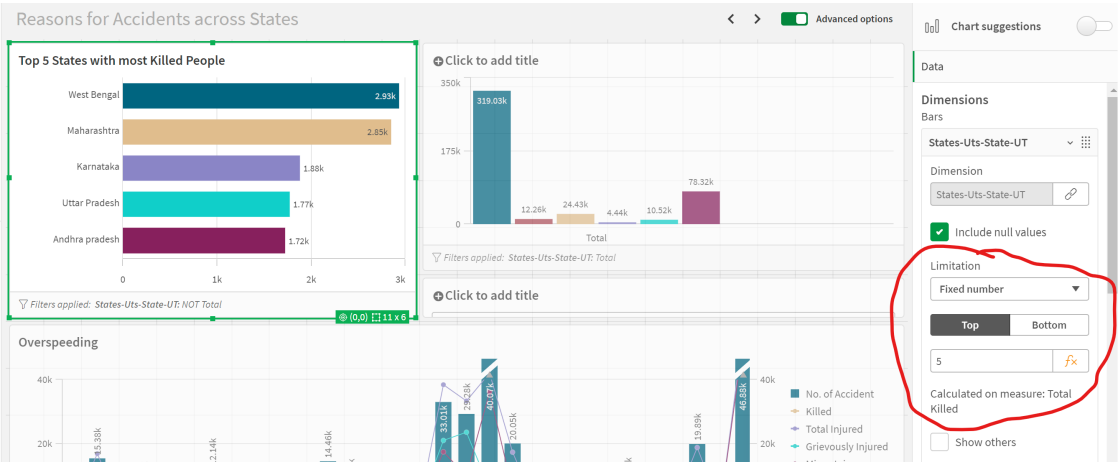


8. Performance Testing

8.1 Amount of Data Rendered

We have uploaded all 9 CSV files in the dataset to Qlik Sense and used all the datasets in our visualizations. We have not removed or added any fields to our datasets. In that case, the data amounts to total of only around 36.1 KB, which is very lightweight and hence, it would be easier for Qlik Sense to work with this smaller size of data. For example, in case we wanted to join 2 tables or perform any cross products between the 2 tables, if the size of the tables and the datasets are very large and contains huge number of records, then it would be difficult to effeciently perform cross products or joins between the 2 tables. So in that way, we have efficiently performed all our visualizations and dashboards, even without adding or removing any data.

8.2 Utilization of Data Filters



We have applied data filters wherever required to limit the amount of data shown on the charts. This is done if there are too many data and we only want to see only a specific number of them, or if we want to filter out any unwanted data. We also apply filters to some charts where we don't want to see the 'Total' field because it doesn't make sense to have it in those charts.