**ABSTRACT** The motive is understanding the trends/patterns in road accidents in India based on time intervals and on monthly basis in order to understand the causes and hence reduce the accidents frequency. Parameters like average, maximum, the time period of maximum accidents can be used for the purpose. Hence, this report talks about Accident trends in India for the period 2001-2014.We try to find pattern based on these trends. We use descriptive and statistical analysis for the qualitative and quantitative proofs in the analysis.

***ROAD ACCIDENTS IN INDIA (2001-2014)***

Pavan Kumar A Srihari S

PES1201701489 PES1201701871

Dept. of CSE Dept. of CSE

PES University PES University

[pavankumara16.08@gmail.com](mailto:pavankumara16.08@gmail.com) Ssriharis1112@gmail.com

**INTRODUCTION** Increased usage of technologies and increased population requires more exchange of services amongst the people. Transport being the basic requirement has vastly developed in terms of efficiency and comfort, but as more of transports are being used there is one thing increasing inevitably, ACCIDENTS. We can observe increased interest in people for road trips and bike stunts. However, we are forgetting serious consequences it puts forward. Road accidents has become a common event nowadays and people are not bothered about it. People are not serious about the traffic rules and the only reason the put on a helmet is to avoid fine and not for their own safety. In such cases, apart from enforcing traffic rules, the government can play much important role in accident control by understanding the patterns in previous accidents and making it available to people so they can be on their own alert. This can be achieved using data science.Hence, we thought of trying out in analyzing the pattern in road accidents from past few consequent years. Now, the task ahead of us was to find a pattern or some data which would help predict future and hence be on our alert. How it works can be explained as below. Suppose a group of people want to go on a bike ride, if they know the pattern of accidents and traffic, it would help them greatly by allowing them to choose a proper route and time. Similarly, by understanding the peak time of accidents the governments can enforce more strict rules for particular time to make it safe and can give more freedom at other hours. Along with that, data science can also be used later i.e, by using comparative analysis they can decide whether their methods are optimal or not based on increase or decrease in accidents, compared to previous time period.

**DATABASE** The following analysis is done on two data sets, one of road accidents for period 2001-2014 for all India states and U.T’s, and the other is number of road accidents based on time-period in a days. The data sets are provided by the Ministry of Road Transport and Highways, Government of India. From first data set we can observe the columns YEAR, Months from Jan to Dec and then TOTAL number of accidents, row-wise it has States/UT’s and for each of the State/UT the accidents data is given for seven years(2001-14). The other data set is the same except that instead of giving data for every month, the data is classified on timings, like 0-3am, 3-6am, so on.. This gives us a vast opportunities for analyzing based on everyday time to monthly period. Most of the analysis is done on Total number of accidents v/s the parameter against which we want to analyse. The following data set is taken from Kaggle website and the link is <https://www.kaggle.com/manugupta/road-accidents-in-india#only_road_accidents_data3.csv>

**DATA-CLEANING** The data set used didn’t have any errors or missing values or wrong values, hence data cleaning process was not of much use. However there were some problems reading the csv in python, which were solved after few modification in arrangement of the table.

**DESCRIPTIVE ANALYSIS**

Table 1:

|  |  |
| --- | --- |
| Months | Average accidents |
| Jan-mar | 476044 |
| Apr-jun | 491600 |
| Jul-sep | 437972 |
| Oct-dec | 462019 |

The following table is just calculated from 1st data set by combining three months and then finding average of the total. It helps in finding average accidents in a time-period for a year. We observe it to be highest around the summer time of the year.

Table2:

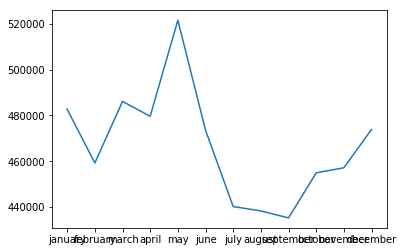
|  |  |
| --- | --- |
| Time period | Average accidents |
|
| 12-3 am | 390197 |
| 3-6am | 474926 |
| 6-9am | 671864 |
| 9-12am | 859444 |
| 12-3pm | 824089 |
| 3-6pm | 906639 |
| 6-9 pm | 873630 |
| 9-12 pm | 602117 |

The following table is just calculated from 2nd data set by considering the time periods in a day and then finding average of total accidents. All these setting up of table can be easily done using Microsoft Excel. It helps in finding the peak time-period in a day where the highest number of accidents occur. So, based on the table the traffic control can be highest at (3-9pm) .We observe it to be highest around the summer time of the year. Table 3:

The table below is the descriptive table of average accidents based on year as a categorical division. These type of tables can prove to be very useful in predicting pattern analysis for further actions. Again we use Excel in constructing the following table. The table is as follows,

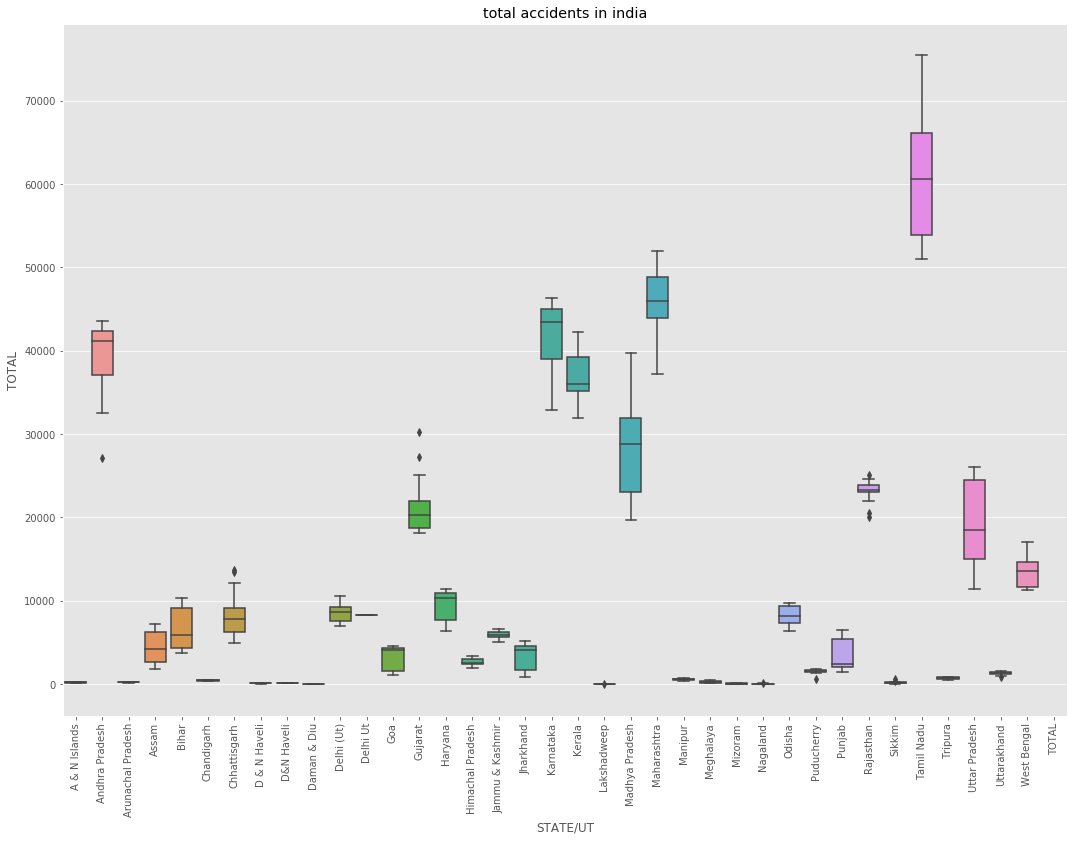
TABLE 3

|  |  |
| --- | --- |
| Year | Average accidents |
| 2001 | 4,06,726 |
| 2002 | 4,29,910 |
| 2003 | 4,39,255 |
| 2004 | 4,60,920 |
| 2005 | 4,79,216 |
| 2006 | 4,84,704 |
| 2007 | 4,86,384 |
| 2008 | 4,99,628 |
| 2009 | 4,97,686 |
| 2010 | 4,90,383 |
| 2011 | 4,86,476 |
| 2012 | 4,89,400 |
| 2013 | 5,01,423 |
| 2014 | 5,00,324 |

  
 **ANALYSIS USING GRAPHS**

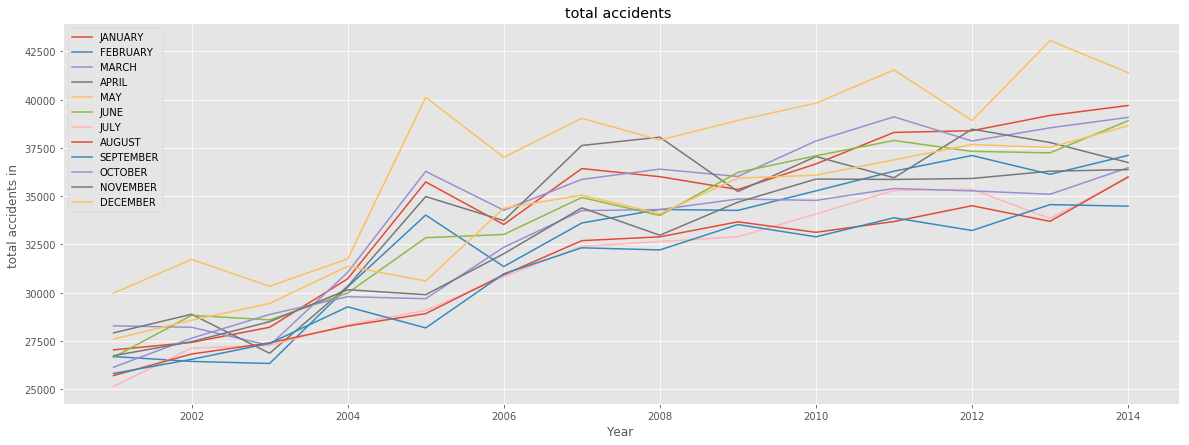
*GRAPH 1:* Line graph of month-wise accidents: ( x-axis->months y-axis->total accidents)

Again, the graph clearly says that highest number of accidents takes place around summer time, reason being most of people go travelling to enjoy the vacation. We must also have a look at minimum time of accidents i.e, around september and try to simulate similar kind of environment throughout the year. Like by observing traffic flow, travellers density etc..

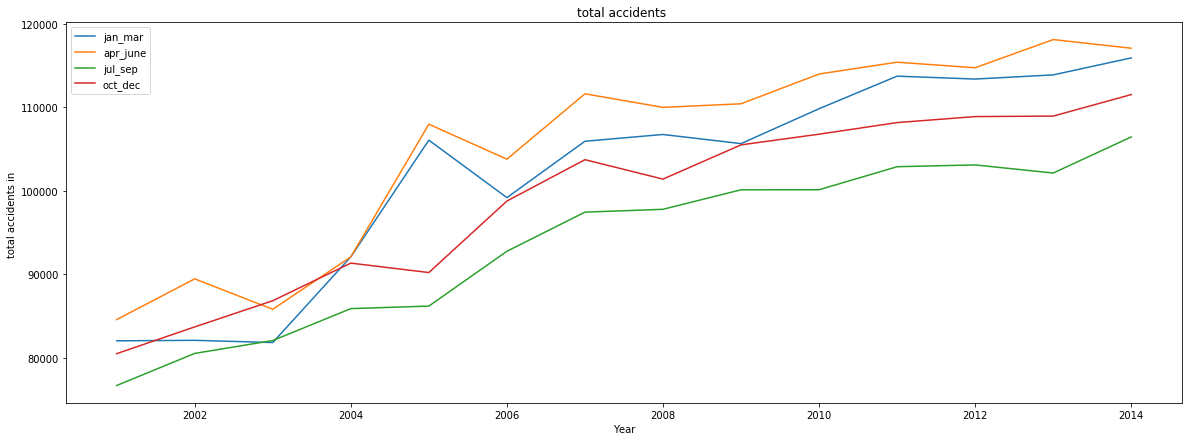
*GRAPH 2:* Box-plot of state wise accidents - **Outliers**: Andhra Pradesh-27,188, Chattisgarh-4,987, Gujarat-30,199 and 27,267, Puducherry – 1286, Rajasthan-25,114 and 19,999, Sikkhim-3517. This graph is a box-plot of total number of accidents according to state-wise distribution where the states are arranged on X-axis in ascending values of state names. We observe one unique set of values called outliers which basically are the values that are completely isolated from the rest of the values, hence causing abnormalities while finding a pattern. **Box plot is an efficient and easy way of finding and removing any outliers if present.**

Apart from that, on analyzing the graph we get to know that the number of accidents vary greatly for the different states and UT’s. The reason for this is quite obvious first being states with high population have higher accidents due to higher traffic, second states with uneven altitude(ghats/mountainous) have comparatively accidents despite less population density because of higher chances of accidents even for a little distractions. We also observe that only 9 states out of 36(including UT’S) have average accidents higher than national average of accidents. On giving a thought it seems that obvious reason might be because of very high accidents in these states, but, as mentioned earlier we must also make note of lower bounds in a graph. For example in above graph we observe that nearly 7 states/Ut’s have extremely less accidents rate. Hence all the states with high accident rate can be on extra alert and take responsibility in reducing accidents. Even just reducing unwanted usage of vehicles might make a very huge difference (it is even eco-friendly and economic). We see that Tamil Nadu has highest accidents occurring, even though it is not having very high population and not having uneven altitude. We might infer that this might be due to in-city accidents rather than accidents on highways, meaning, carelessness of people, poor enforcement of traffic rules, high population density.

*GRAPH 3:* A line graph of year v/s accidents for various months.

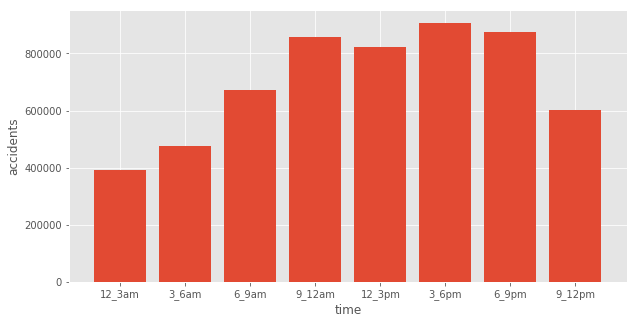


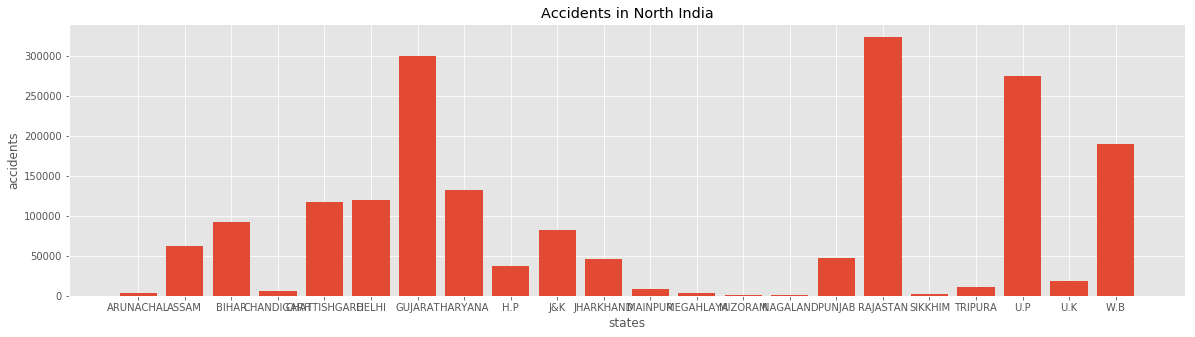
Again, the above graph plays an important role in prediction analysis. As expected of it, there is nearly 100% increase in accident rates over the past years. The graph is very simple to understand. It is basically a plot of how much the accidents have increased over the past few years per month.Since there is ovelap of colours we group every 3 months in to group of three to get the following graph,

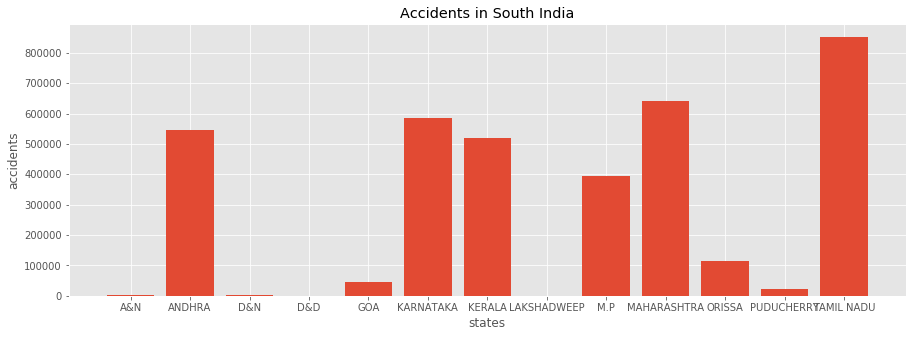


From this graph it is seen that highest accidents have always took place around April-June and least acccidents around July-September, which compares with the previous graph. We can conclude this drastic increae in rates might be due to increased traffic(population) with still old conventional traffic rules. The government must recognise this fact as in order to get a negative slope in the pattern or atleast stop the increase rate. Just simple measures can play an important role, for example just installing a railway gate at all rail-road junctions would completely remove rail-road accidents(provided society respects the purpose of it), which accounts for nearly 7% of Road accidents. Another measure that can be taken is by increasing road size and smoothness where there is high vehicle flow, uneveness is in the road is what gives us the “frustated Indian driver”. Small and uneven roads are expected to be the cause of nearly 60% of accidents. Then there comes careless driving which have to be solved at psycological level, or by imposing very high fines.

*GRAPH-4:* Three basic histograms with frequency of some important keys







Here we have 3- right skewed graph telling us frequency of total accidents in a state, average accidents in India and Total accidents in India. We see that most of the states have less (around 20) have accidents of less than 1,00,000. Hence just by keeping a check on 6-7 states would reduce the national average of accidents. The height of the two graphs is same because average number of accidents add up to total accidents. We have plotted both so as to get an estimate of count of accidents because the distribution of accidents might be skewed. Average accidents mean accidents occurring for a year in one state, which is around 2000 per year for majority of states.

**STATISTICAL ANALYSIS:**

From descriptive analysis we get to know the relation between two or more parameters. But to calculate the relation/predict the trend we need a quanitative term on which we can make conclusions. There are many statistical analysis methods like Chi-squared test, correlation analysis, paired t-test, Regression analysis, Wilcoxon test. We carried out few related tests for quantitative analysis.

**STATISTICAL-HYPOTHESIS** We consider the following observations, modern times have caused more awareness and increased safety features in vehicles, hence there has been significant decrease in number of accidents in recent periods. Total population being 489 for both cases,

Null Hypothesis: There is increase/no change in number of accidents during recent times. Alternate hypothesis: There is a atleast some decrease in number of accidents over the recent years. H0: no. of accidents in year1-no. of accidents in year2 < 0 (year1>year2) Ha: no. of accidents in year1- no. of accidents in year2 > 0 (year1>year2) Average accidents in year 2:457192 Standard deviation: 1269.392 n: 489 Average accidents in year 1: 473905 Standard deviation: 1337.868 n: 489 NULL distribution (0,1269.392^2/489+1337.868^2/489) //z=2.40

P-value=0.0082(<0.05).Therefore, Null hypothesis can be rejected and we can consider there is a decrease in no. of accidents. The test method used above is hypothesis test. But we must also remember that it gives on a whole nation basis. This just means that there are more cases with decrease in rate than compared to cases with increase. We must make sure that none of the states have increase in the accidents rate , if we want to call ourselves a developed state, because this is basic factor in determining nations state.

**PEARSON CORRELATION TEST**

To find the relation between time-interval of a day and total number of accidents occurring in that time period. This might be very useful because this is basic step in future pattern prediction. We know, Pearson correlation coefficient is given by, r=∑((x’-x)(y’-y))/∑(x’-x)^2 . ∑(y’-y)^2 where, x’(12.5) is average time, x is time periods, y’(70,000) is average accidents y is accidents in a time period. ∑((x’-x)(y’-y))= 6218894 ∑(x’-x)^2 . ∑(y’-y)^2=1.03E+14 r=0.621889

Hence, we observe that there is some relation between these two contents. i.e., there is increase in number of accidents in evening and night period as compared to accidents before 12pm. Keeping ourselves alert during these peak hours and even the government can do its part by increasing traffic signal activeness, providing sufficient public transport at right time. We observe r value to be little more than a half meaning that it is not that strongly increasing but increasing however. Again in in a country like India it is impossible to pin-point what will happen at what time. It is good to be at our alert all the time.

**SUMMARY**

Our Data analysis was on Road Accidents in India (2001-2014). Fortunately the data was consistent and uniform except for the arrangement, which was done using MS EXCEL. The purpose of selecting this dataset was in order to understand this pattern in accidents so as to understand causes of accidents in non-technical terms. We have done enough analysis based on the data set but to know the complete theory we need to analyse several datasets and then only we can come to a proper conclusion. We must know that in developed countries the bus timings are all set after studying the flow pattern of the people. It wouldn’t have been necessary in previous years to analyse everything before enforcing a rule. But in today’s world it is a must because of increased population and every citizen riding his/her own vehicle. To conclude the analysis I would like to say that the best way to get a negative slope in year v/s accidents graph is by combining the data analytics and law enforcements.