**A description of the problem and a discussion of the background. (part 1 of week 1)**

The work done in this project is analysis of the Accidents ( Collision Data)  from state of California USA. This report contains  findings and recommendations that help  control and reduce the number of collision Incidents. And a guidance to creation of a system that provide proactive notification and warning to  public on  roadsafety and driving conditions, and precautionary  steps to observe and follow guidance by the  State Transport Authority.

**Problem Statement** : The data from the source  indicates very large number of collision incidents, resulting in injuries  being significantly higher in numbers. Also the number of accidents are increasing year over year  resulting  in road safety concerns and damage to personal health and property.

This report provide you details of  conditions and factors responsible for  collision incidents and  methodology used in this research  to  prepare  the findings and  recommendations.

**From the data,** the cause of collision incidents cover range of samples  like  hitting pedestrians, animals crossing,hitting on the rear, hitting parked vehicles  etc.. reflects  road/weather conditions having an influence with less control on braking,  Light conditions.

Also  driver behavior ( like speeding, under the influence ) are also cause of accidents,  again are augmented by the weather/ lighting conditions.  Most of them occurring  during night when street lights are on.

While the existing traffic Law have defined guidelines for drivers to observe/follow rules  to prevent collision incidents ( ex: speeding, under influence etc),  preventing incidents as a result of weather/road condition requires new specific measures and guidelines.

The data analysis will provide insights and remediation suggestions

**A description of the data and how it will be used to solve the problem (part 2 of week1)**

**Research  Methodology.**

The data source  is the data file provided by Coursera  learning team.  Data cleansing is done as data being inconsistently updated for each incident and only relevant variables indicating cause  / influencing the incident are considered.

With Target variable being the severity of incidents ( from the observations),  data modelling is  considered  using  a  classification  method. Mostly a decision Tree Technique  and/or  SVM .

1.       Applied statistical/descriptive analysis on the data sent.  Based on it, variables considered for cleaned up data set :   Road ,weather and light condition, collision description,  Junction type

2.       Data cleansing will be done accordingly.

3.       A correlation map of the variables resulting/causing collision will e developed.

4.       Data modelling with Decision Tree  and SVM  built

5.       Prediction and model accuracy evaluation will be done

6.       Visualization of data  is presented.

7.       Based on specific observations  on cause of incidents  like  Parked vehicles,  speeding

The data contains information pertaining to each collision , however missing data is significantly high in many fields, are not considered ( not being cause of collision).    For the critical variables, missing data rows are removed. ( as the data set is significantly large and number of missing data in the final set are minimal).   Speed  and  UnderInf ( under influence) variables are not considered based in the ratio od data units (incidents).

From the statistical analysis we find, there are only 2 type of  result,  injury or property damage.  Hence this is a classic  case for categorical data analysis ( like the churn).

The data contains  location of accidents  in variables like  co-ordinates, Junction type( both as intersection  and  block), street /block address .  from the analysis we find there are few streets with a greater number of collisions. Given  each incident is equally important   **we will consider   Junction type as one of key variable in the data set.**

However, we will   get visibility to specific block or intersection with more incidents to make specific recommendation.( from Location data)

Data set also contains  conditions like road condition, weather and  lighting ( day or lights on) which play critical role in cause of collision.  **We will consider these as critical variables.**

Data set includes other possible causes like  Speeding,  or under the influence .However the  sighting of such is  relatively small in number

* Speeding 269 of ~7500 incidents
* under influence  394 cases  with 75% of  incidents occurring during dusk/ night with street lights on.

Assuming these are  possibly well-known causes already considered under traffic rules and law, will be ignored from the data set.

Another critical  variable is the collision description, where in  accidents with parked vehicles are quite handful. There are about ~ 25% ( 1791 cases)  involving  parked vehicles.  And also  accidents with  one car stopped, and other hitting at rear  is quite considerable ( 974 cases) 13%

* This would call in for a specific recommendation to consider remediation to  avoid these cases. Like review Parking guidelines/ space  on the streets.
* Many of incidents of later type are due to speeding,

There are other collision descriptions that reflect the cause possibly be due to driver behavior / breaking rules.  Also some losing control due to weather conditions. This would call in for a specific recommendation to consider remediation to  avoid these cases.

So Our data set will contain  limited fields/ columns **:   severity, junction type, road, weather, lighting conditions.** will work on the data model and code based on this descriptive analysis