1. Data

The downloaded data is loaded into a Pandas data frame. The dataframe contains 194673 rows × 38 columns. A new dataframe containing relevant information for this study has been constructed which contains the following columns:

Column name	Description	Values
'SEVERITYCODE'	A code that indicates the severity of the collision	• 3—fatality • 2b—serious injury • 2—injury • prop damage • 0—unknown
'Y'	Latitude	Latitude in deg
'X'	Longitude	Longitude in deg
'ADDRTYPE'	Collision address type	Three values possible: Alley Block Intersection
'SEVERITYDESC'	A detailed description of the severity of the collision	Text describing what happened
'COLLISIONTYPE'	Collision type	Text describing the type of collision
'INCDATE'	Date of the incident	
'SDTO_COLCODE'	State collision code	Numerical code described in metadata
'INNATENTIONIND'	Whether the accident is due to inattention	(Y/N) data will be changed to y=1, N=0 and missing data NaN
'UNDERINFL'	Whether the driver was under the influence of alcohol or drugs	(Y/N/1/0) data will be uniformed Y=1, N=0 and NaN for missing data
'WEATHER'	Description of weather conditions	Values: Blowing Sand/Dirt Clear Fog/Smog/Smoke Other Overcast Partly Cloudy Raining Severe Crosswind Sleet/Hail/Freezing Rain Snowing Unknown
'ROADCOND'	Description of the conditions of the road	Dry Ice Oil Other Sand/Mud/Dirt Snow/Slush Standing water Unknown Wet

'LIGHTCOND'	Description of light conditions.	Dark-No Street lights Dark-Street lights off Dark-Street lights on Dark-Unknown lighting Dawn Daylight Dusk Other Unknown

In all cases, the values coded as ‘UNKNOWN’ or ‘NAN’ will be removed from the dataframe.

All of the values will be hot-encoded for further analysis.

-The severity of the collision will be related to weather and road conditions. Furthermore, the relation between light conditions and severity of condition will be studied.

-Latitude and Longitude data will be used to map severe collisions and study if there are areas of the city where these kinds of collisions happen often.