1.Introduction:

Traffic accidents are one of the daily incidents to a community that causes severe property damage, serious injuries, and high number of fatalities. According to Washington State Department of Transportation (WSDOT), there is a car accident for every 4 minutes while a person dies due to these accidents every 20 hours. Governments around the world invest heavily in making roads safer and reducing this threatening danger. To take protective measures against such incidents, individuals are advised to comply with certain laws and avoid some conditions when driving is dangerous. However, would not it be useful if one could predict the severity of an accident given some conditions are present when driving? For example, if it was raining and dark outside and you planned to visit a friend but you wanted to know whether driving with these conditions would result in an accident. Moreover, how severe that accident would be if the road was wet and it involved some intersections. This project, as part of Coursera Data Science Capstone, aims to create a machine learning model to predict the severity of a car accident that could occur given some conditions are present.

2.Business Problem

The project's objective is to predict the severity of car accident that could occur based on given features. These features include weather condition, driving speed, light condition, and others. The predicted severity levels are:

- 3—Fatality
- 2b—Serious injury
- 2—Injury
- 1—Property damage
- 0—Unknown

Since the predicted variable is categorical, a machine learning classification model is trained by using Seattle Geo Data.

3.Data Description

The data is retrieved from Seattle Geo Data website. It contains 40 features and over 220,000 accidents in Seattle. The features identify information about each accident (e.g. location and junction type) and they are explained in the metadata file given by Seattle Geo Data. The data timeframe is from 2004 to present date. Table 1 Shows an example of some of the accident's features along with the corresponding information from the data set. It is obvious that such data is not cleaned well to be used for training the ML model with some of the features being in wrong format or not useful to predict the severity level.

Table 1 Data set sample accident with features and information

Feature	Information	Feature	Information
X	-122.3867716	LOCATION	California ave sw and sw genesee st
Y	47.5647203	SEVERITYCODE	2
OBJECTID	1	SEVERITYDESC	Injury Collision
INCKEY	326234	COLLISIONTYPE	Pedestrian
COLDETKEY	327734	PERSONCOUNT	2
REPORTNO	E984735	PEDCOUNT	1
STATUS	Matched	PEDCYLCOUNT	0
ADDRTYPE	Intersection	VEHCOUNT	1
INTKEY	31893	INJURIES	1