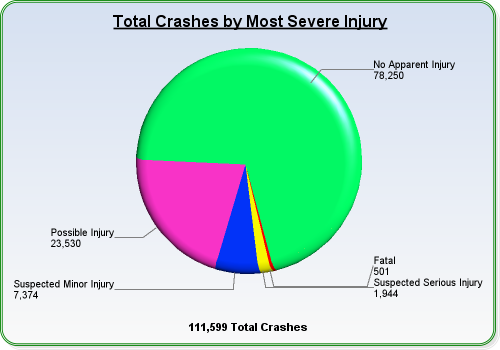
**Background**

Car accidents are a common occurrence on US roads, with some of them resulting in serious injuries or death.

The [Washington State Department of Transportation Crash Data Portal](https://remoteapps.wsdot.wa.gov/HighwaySafety/Collision/Data/Portal/Public/) provides crash information for accidents that occurred statewide. According to the 2019 data, there were 111,599 crashes. The number has increased from 101,876 in 2010.



City of Seattle accounted for almost 10% of these crashes in 2019.

The major types of car crashes include – a) Distracted driver, b) Teen driver, c) Pedestrian, d) Bicycle, e) Drunk driver, f) Motorcycle, g) Bus and h) Large Truck.

As the number of crashes continue to be on rise, it will be useful to analyze the crash/accident data and find a potential solution.

Prediction of severity of car accidents along with a detailed data analysis can be useful for policymakers to find a solution to this problem.

For the purpose of this exercise, the focus will be on the accidents data associated with city of Seattle. We are going to apply data science techniques to explore the accident severity of the past incidents and use it to build a model that can predict the severity for future based on certain conditions.

**Data Section**

1. **Data Description**

The data used for the purpose of this analysis comes from the [sample data](https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv) provided as part of the IBM Data Science Coursera course. The .csv file contains 194,673 collision records for Seattle. There are 38 columns covering different aspects of the data. [Metadata](https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf) provides data set summary along with the attribute information and other useful details.

The objective is to predict severity of car accidents and therefore we are going to predict “SEVERITYCODE” (target variable) based on other variables.

1. **Data cleaning**

In its original form, this data is not fit for analysis because it has many columns which will not be used for this model. Also, most of the features are of type object, when they should be numerical type.

1. **Feature selection**

In order to select the relevant features, I do an exploratory analysis of the data to identify relevant attributes to predict the target variable. Based on the analysis, I selected ADDRTYPE, WEATHER, ROADCOND, LIGHTCOND as the features.

1. **Balancing the dataset**

The target variable SEVERITYCODE is not perfectly balanced. The severitycode in class 1 is nearly

three times the size of class 2. This can be fixed by downsampling the majority class.