Data

The data to be used for this analysis comes from Kaggle.com (https://www.kaggle.com/tsiaras/uk-road-safety-accidents-and-vehicles). The original data is from the UK's Department for Transport. There are two data sources available:

- **Accident information,** which contains information on distinct traffic accidents from 2005 to 2017. This dataset contains attributes, include a target variable for accident severity, that can be used in machine learning models. This dataset contains 2,047,256 rows and 34 columns.
- **Vehicle information,** which contains information on the vehicles involved in the accident and passenger information from 2004 to 2016. This dataset contains 2,177,205 rows and 24 columns.

The features and target variable available in the datasets are shown below.

```
In [24]: accident_data.dtypes

Out[24]: Accident_Index                                 object
         1st_Road_Class                                 object
         1st_Road_Number                               float64
         2nd_Road_Class                                 object
         2nd_Road_Number                               float64
         Accident_Severity                              object
         Carriageway_Hazards                            object
         Date                                           object
         Day_of_Week                                    object
         Did_Police_Officer_Attend_Scene_of_Accident    float64
         Junction_Control                               object
         Junction_Detail                                object
         Latitude                                      float64
         Light_Conditions                               object
         Local_Authority_(District)                     object
         Local_Authority_(Highway)                      object
         Location_Easting_OSGR                         float64
         Location_Northing_OSGR                        float64
         Longitude                                     float64
         LSOA_of_Accident_Location                      object
         Number_of_Casualties                            int64
         Number_of_Vehicles                              int64
         Pedestrian_Crossing-Human_Control             float64
         Pedestrian_Crossing-Physical_Facilities       float64
         Police_Force                                   object
         Road_Surface_Conditions                        object
         Road_Type                                      object
         Special_Conditions_at_Site                     object
         Speed_limit                                   float64
         Time                                           object
         Urban_or_Rural_Area                            object
         Weather_Conditions                             object
         Year                                            int64
         InScotland                                     object
         dtype: object
```

```
In [31]: vehicle_data.dtypes

Out[31]: Accident_Index                          object
         Age_Band_of_Driver                      object
         Age_of_Vehicle                         float64
         Driver_Home_Area_Type                   object
         Driver_IMD_Decile                      float64
         Engine_Capacity_.CC.                   float64
         Hit_Object_in_Carriageway               object
         Hit_Object_off_Carriageway              object
         Journey_Purpose_of_Driver               object
         Junction_Location                       object
         make                                    object
         model                                   object
         Propulsion_Code                         object
         Sex_of_Driver                           object
         Skidding_and_Overturning                object
         Towing_and_Articulation                 object
         Vehicle_Leaving_Carriageway             object
         Vehicle_Location.Restricted_Lane       float64
         Vehicle_Manoeuvre                       object
         Vehicle_Reference                        int64
         Vehicle_Type                            object
         Was_Vehicle_Left_Hand_Drive             object
         X1st_Point_of_Impact                    object
         Year                                     int64
         dtype: object
```

Features for modelling

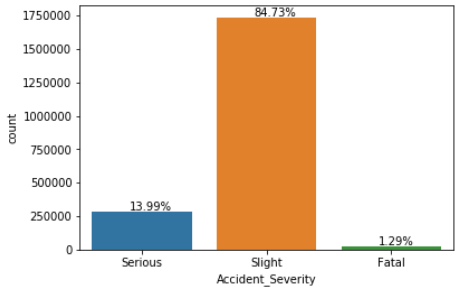
The two datasets can be joined together using a distinct ID (Accident_Index). I will join them into one dataset and then select the features to be used for modelling. The target variable will be *Accident_Severity.* Features for training the model will include the following:

- *Day_of_Week*
- *Junction_Detail*
- *Light_Conditions*
- *Road_Type*
- *Speed-Limit*
- *Urna_or_Rural_Area*
- *Weather_Conditions*
- *Age_of_Vehicle*
- *Engine_Capacity*
- *Sex_of_Driver*
- *Age_of_Driver*
- *Was_Vehicle_Left_Hand_Drive*
- *Vehicle_Type*

Exploring the target variable

The plot below shows the distribution of classes in the target variable. The distribution shows that the data is highly unbalanced, with the vast majority of accidents (85%) categorized as slightly severe. That poses issues for modelling and therefore will be addressed appropriately. In addition, given how small the proportion of fatal cases are, it will be grouped together with the serious category, making this a binary classification problem.

```
In [37]: ax = sns.countplot(x="Accident_Severity", data=accident_data)
         for p in ax.patches:
             ax.annotate('{:.2f}%'.format(100*p.get_height()/len(accident_data.Accident_Severity)), (p.get_x()+ 0.3, p.get_height()+10
```



Additional data preprocessing

Other data preprocessing steps that will be conducted includes standardizing all numeric variables, one hot encoding of all categorical variables, identifying appropriate methods for dealing with missing data.

The data from 2005 to 2014 will be used to train the machine learning classification models and then the models will be tested using the data from 2015 to 2017.