2. Data

Our data from Seattle's transport department was downloaded from Seattle's website. It contains features on collisions in the city of Seattle, a seaport city on the West Coast of the United States of America from 2004 to 27 August 2020. This data is available here.

The dataset we are working with contains 221266 records and 40 columns.
SEVERITYCODE is going to be our label column.

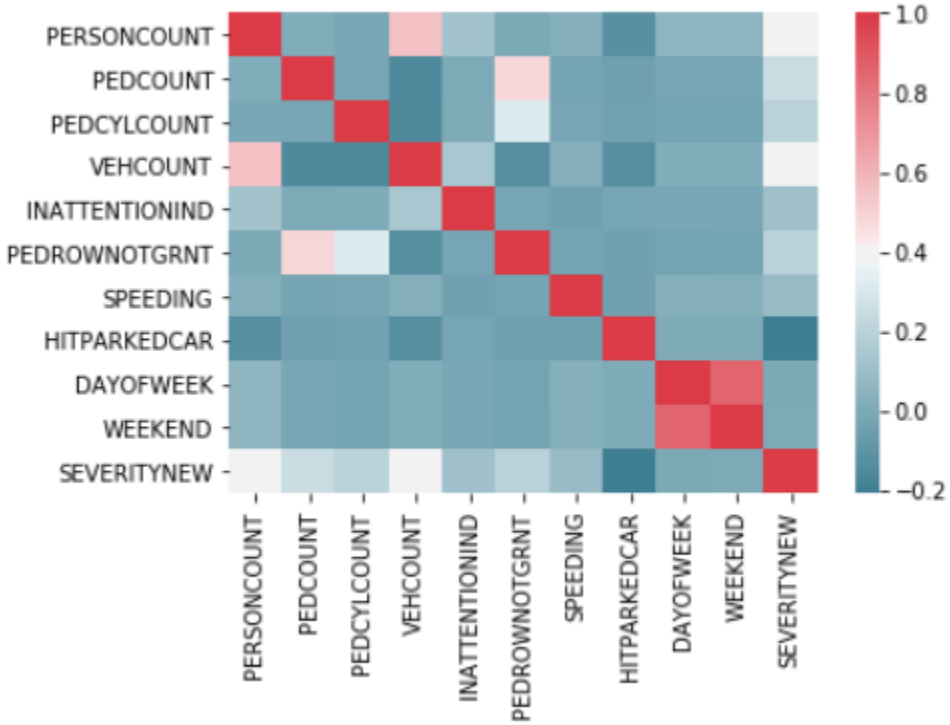
2.1 Removing features which are not required

The following features were seen to be not adding value when it comes to developing our predictive model and as such are removed.

- **OBJECTID, X, Y** : the ESRI unique identifier and location coordinates
- **INCKEY, REPORTNO, COLDETKEY, INTKEY** : a unique key for the incident, report number, secondary key for the incident and key referencing collision intersection respectively
- **STATUS** : it is not clear what information is stored in this column since it also doesn't appear in the metadata file of the dataset
- **LOCATION** : description of the location of the collision
- **EXCEPTRSNCODE,EXCEPTRSNDESC** :it is not clear what information is stored in these columns since no descriptions appear in the metadata file of the dataset
- **SEVERITYDESC** :this is simply a description of collision severity code (label) and will not be useful to keep it.
- **INJURIES, SERIOUSINJURIES, FATALITIES** : this information will not be available at the time of the collision at which point we want to predict how severe the collision is.
- **JUNCTIONTYPE** : this column is derived from the ADDRTYPE (collison address type). We will remove this column and retain the ADDRTYPE column.

- **SDOT_COLCODE, SDOT_COLDESC**: the code and description given to collision will only be available after the collision and hence these columns will not help us to predict the collision severity.
- **COLLISIONTYPE** : collision type is correlated with the collision severity that we want to predict.
- **SDOTCOLNUM** : collision number will not be able to assist in predicting the severity level of an accident.
- **ST_COLCODE, ST_COLDESC** : state collision code and description which will be available after collision has been attended to by state will not be useful to us to create a predictive model.
- **SEGLANEKEY, CROSSWALKKEY** : lane segment and crosswalk keys are not relevant for our predictions.

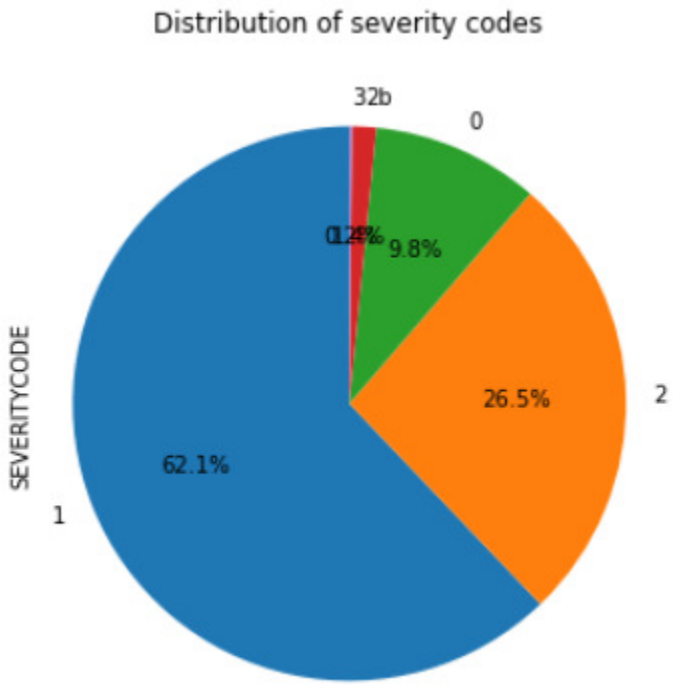
2.2 Correlation heatmap



Although the above heatmap drawn by the seaborn Heatmap function is not a perfect tool, we can have some better understanding at a glance that the vehicle count (VEHCOUNT) and the total number of people involved in a collision (PERSONCOUNT) has some positive correlation with the severity of the collision. A negative correlation exists between the severity of a collision and a collision which involves hitting a parked car.

2.3 Distribution of severity codes

62.1% of all the accident records are for severity level 1 (property damage) and the smallest percentage is 0.2% which is for severity level 3 (fatal).



```
1     62.1%
2     26.5%
0      9.8%
2b     1.4%
3      0.2%
Name: SEVERITYCODE, dtype: object
```

2.4 Null values and unknowns

Every column was checked for nulls and guided by the proportion of the null values within each particular column, the nulls were removed or treated. For example, null values in the WEATHER column constituted less than 5% of the rows and as such a decision to proceed without these rows was made.