**IBM Data Science Capstone Project**

**Car Accident Severity (week 1)**

**Oscar Vicente Remón**

**October 4th, 2020**

**Abstract**

Despite of the number of technological improvements, in a daily basis many car accidents still occur in the big urban areas. Seattle with 783,187 inhabitants has a relevant number of accidents. Through this project, the dataset provided by the Seattle Traffic Management Division and containing a detailed report of each of the accidents between 2004 and 2020 is going to be analysed. The first objective of the project will be to determine the influence of certain variables in the occurrence of an accident. As second objective a set of Machine Learning models will be simulated in order to foreseen under which variables conditions an accident could happen in the future. As last, a set of conclusions and recommendations are going to be draft to the stakeholders.

1. **Description of the problem and a discussion of the background.**

The seaport city of Seattle is the largest city in the state of Washington, as well as the largest in the Pacific Northwest.

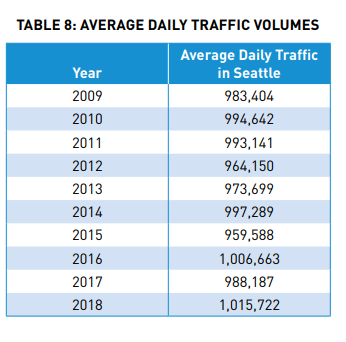
As of the 2020 census, there were 783,187 people living in Seattle.



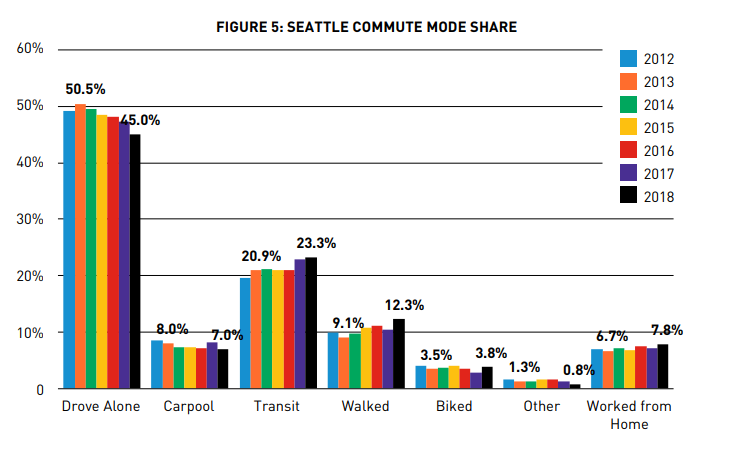
Source: https://worldpopulationreview.com/us-cities/seattle-wa-population

Seattle residents get around by car, trolley, streetcar, public bus, bicycle, on foot, and by rail. With such bustling streets, it is no surprise that Seattle sees car accidents every day.

In order to have an idea of the traffic volume order of magnitude and how is split per transportation type the tables and graphs below should give an overview to of the situation even if containing only data up to 2018.



Source: http://www.seattle.gov/Documents/Departments/SDOT/VisionZero/2019\_Traffic\_Report.pdf



Source: http://www.seattle.gov/Documents/Departments/SDOT/VisionZero/2019\_Traffic\_Report.pdf

The number of accidents according to the SDOT Traffic Management Division, Traffic Records Group dataset collisions shows a slight decrease in the last years.



Source: SDOT Traffic Management Division, Traffic Records Group

However, due to the amount of money these accidents cost to the taxpayers, the local authorities would like to understand better, which factors have an impact in the accidents and how the accidents could be predicted or better said, avoided, in the future.

With this objective, the project to be developed will:

* Evaluate the variables provided, extended in the full report version.
* Determine how they affect the accidents numbers and typology.
* Analyse the different prediction models applied to the set of selected variables.
* Proposed the best model for predicting the accidents
* Provide a lesson learnt and recommendations document to the stakeholders with the aim to improve the situation.

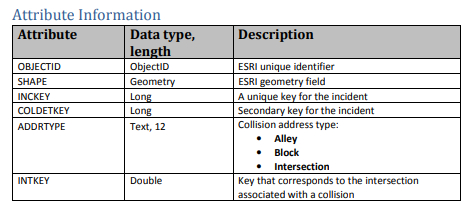
1. **A description of the data and how it will be used to solve the problem.**
   1. **Data understanding:**

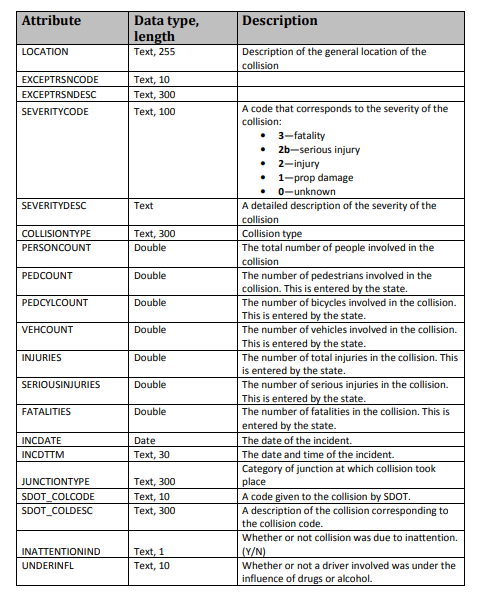
The dataset used for this project is based on car accidents occurring in the city of Seattle during the years 2004 to 2020**.**

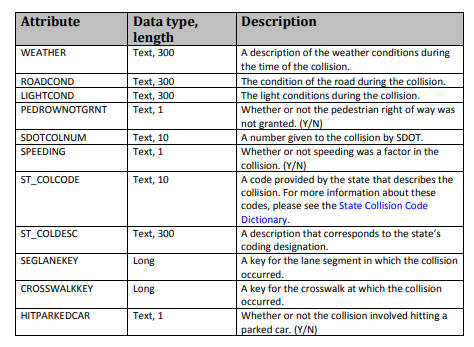
The [link](https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv) to download the dataset is provided in the Coursera Capstone project week 1 labs.

The Seattle Traffic Management Division, in charge to produce the dataset, categorises the accidents in two types according to their severity: 1 - Standing for “Property Damage Only” and 2 - Injury Collision.

Attached to the severity for each accident record there is a set of attributes that classifies the accident record.

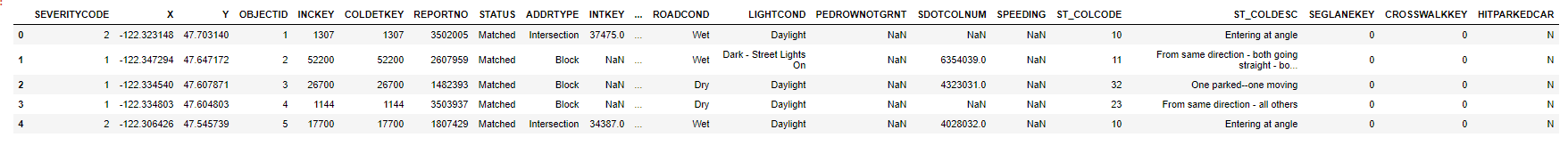






The dataset can be downloaded first in format csv and imported into pandas or directly read from the URL.

The contents can be displayed by using Head function.



A Jupiter notebook containing the final code will be submitted via Github during the next week assignment.

Out of the total number of variables available, we are going to focus in this project in:

Predicted and dependent variable: SEVERITYCODE

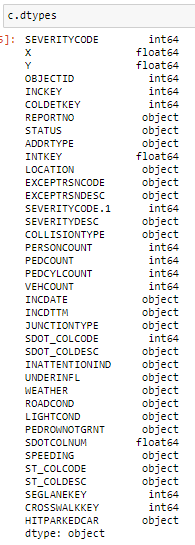
Independent variables:

Information about the location and characteristic of the address where the location took place: ADDRTYPE, JUNCTIONTYPE, X, Y, PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT, ROADCOND.

Environmental factors: WEATHER, LIGHTCOND

* 1. **Data Preparation:**

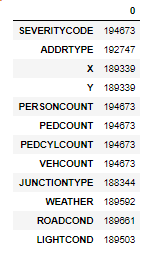
By applying the DTYPE instruction to the data frame read from the url containing the dataset, the technical structure of the dataset can be quickly displayed.



In the original format, the dataset is not suitable for quantitative analysis, because a number of categorical variables are identified, those identified as object type.

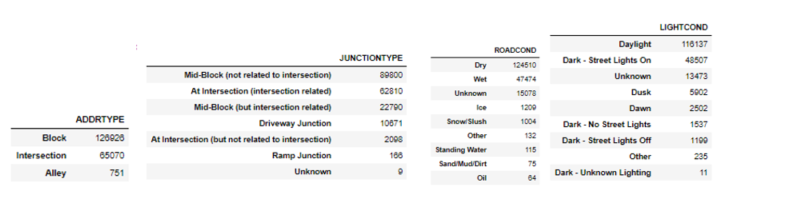
By looking further at the dataset description, some variables / columns will not provide too much information to the model, those are the variables associated to technical fields like OBJCTID, REPORTNO, INTKEY.

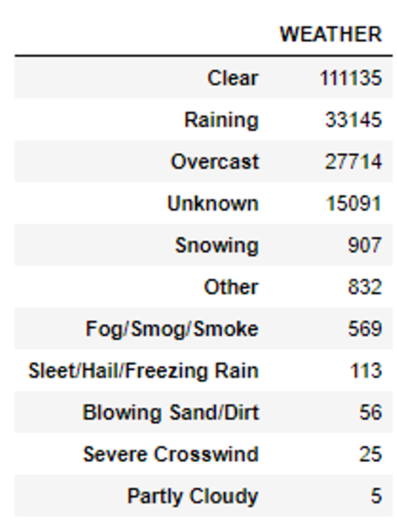
The total number of records contained in the dataset is 194,673. The drilldown in terms of populated records is displayed in the table below. Only the sort list of study variables concerning our project are shown.



This lead to the conclusion that not all variables are always populated. Some variables contain blank records.

Via a quick exploratory analysis to the sort list of categorical variables of study:





There are records associated to some of the variables showing: Unknown or Other values.

In order for this dataset to be used to feed the Machine Learning models, some further actions have to be carried out as part of the data cleaning process.

* Eliminate columns not included in the sort list of study.
* Eliminate all records with values adding potential noise into the model.
* Convert categorical variables to numerical.
* Down sampling the dataset in order to have a balanced dataset.

Reference Links:

https://worldpopulationreview.com/us-cities/seattle-wa-population

https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

http://www.seattle.gov/Documents/Departments/SDOT/VisionZero/2019\_Traffic\_Report.pdf

<https://www.colburnlaw.com/seattle-traffic-accidents/>

https://www.wsdot.wa.gov/mapsdata/crash/pdf/2015\_Annual\_Collision\_Summary.pdf