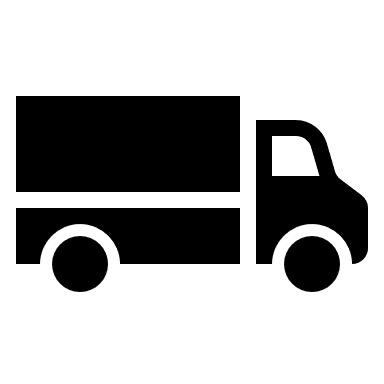
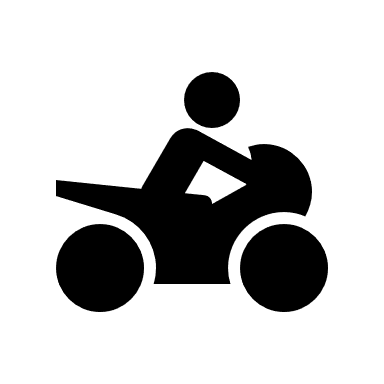
## 

**** ****

**SEATTLE TRAFFICK DATA ANALYSIS AND RISK PREDICTION**

**IBM Final Capstone Project**

**Puja Misra**

**Introduction**

This project is our Final submission to IBM Data Science Professional Certificate course on Coursera. The goal of the project is to detail and use Data Science tool set for Predictive analysis.

We will be working on a real-life problem and demonstrate how Machine Learning can help us predict and process the value by applying the learned skills.

**Business Understanding**

* **Background:**

According to 2017 WSDOT data, a car accident occurs every 4 minutes and a person dies due to a car crash every 20 hours. Fatal crashes went from 508 in 2016 to 525 in 2017, resulting in the death of 555 people. This number has stayed relatively steady for the past decade. According to 2017 WSDOT data, a car accident occurs every 4 minutes and a person dies due to a car crash every 20 hours. Fatal crashes went from 508 in 2016 to 525 in 2017, resulting in the death of 555 people. This number has stayed relatively steady for the past decade.

* **Problem Statement:**

As we see in the background statement above, the numbers of fatal crashes are having an upward trend or has been steady for past decade for Seattle as per WSDOT .

A car parked on a beach

Description automatically generated

* **Objective of the Project:**
* The purpose of the project is to gather the data and determine what causes the accident and the attributes that leads to the severity.
* Through data visualization and machine learning algorithm we will be analyzing a significant range of attributes, including weather conditions, road condition, speeding, special events, roadworks, traffic jams among others and we will try to predict what are the conditions that can contribute to high severity accidents which may cause loss of life or loss of property .WSOT can use the model to take precaution to minimize the loss of property and life.
* Reducing the insurance cost and Preventing fatalities
* **Stakeholders:**
* Government Officials
* Emergency Responders (911 dispatchers)
* Common People
* Insurance Companies

**Data understanding**

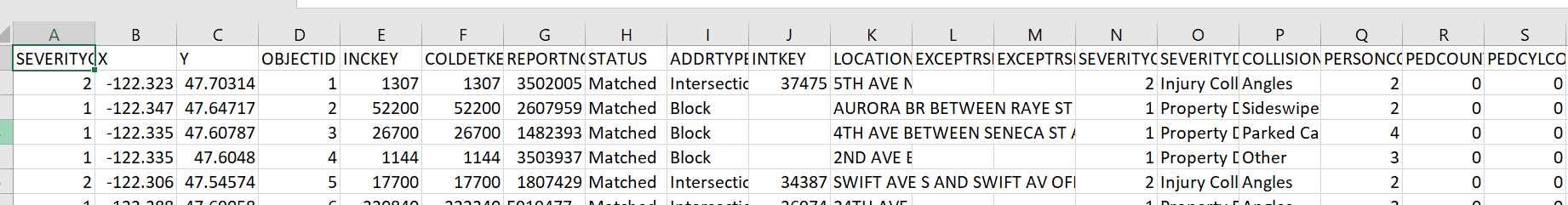
We chose the public data from open source available with labeled columns and attributes and observations data to help us do our analysis better.

Sample Data Below

Link to the Data

<https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv>

https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions\_OD.pdf



The data consists of 37 independent variables and 194,673 rows. The dependent variable, “SEVERITYCODE”, contains numbers that correspond to different levels of severity caused by an accident from 0 to 4.

Severity codes are as follows:

0: Unknown

1: Property Damage

2: Injury

2b: Serious Injury

3: Fatality

**Data Preparation**

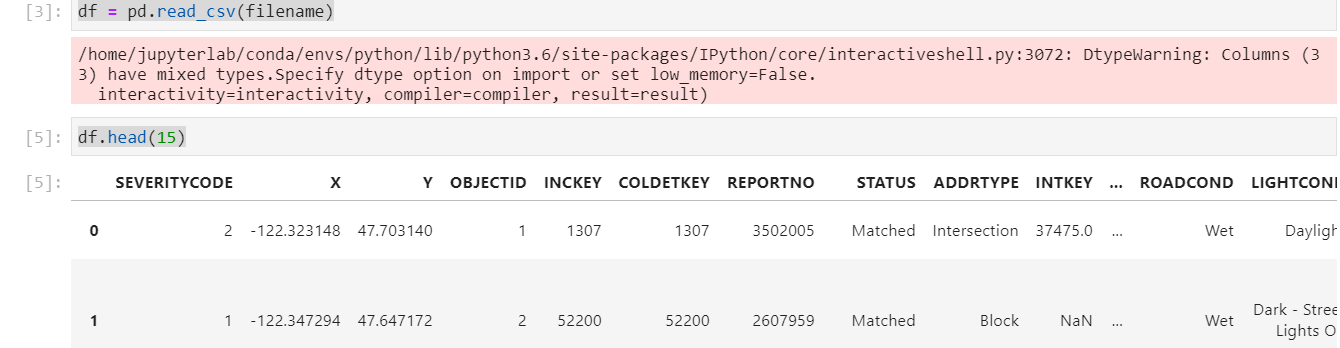
# Public Traffic data for Seattle city USA is available from the Open source(Link Mentioned above)

# After the data has been extracted, keeping the columns required in the data frame.

# Excluding the rows with null values.

# Transform the data type for analysis.

* Load the Data to data frame.

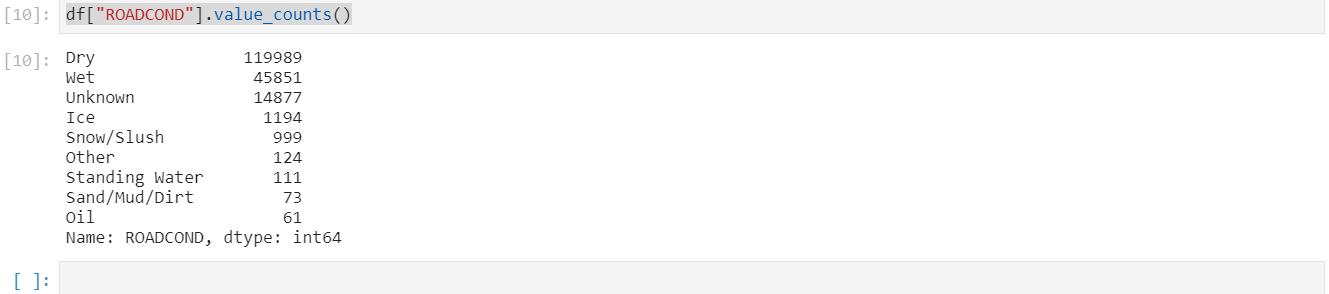


We will run a value count on road (‘ROADCOND’) and weather condition (‘WEATHER’) to get ideas of the different road and weather conditions. We will also check the value count on light condition (’LIGHTCOND’), to see the breakdowns of accidents occurring during the different light conditions. The results will then be used for data modeling.

Count Based on the Severity code



Count Based on the Road Condition



Count Based on the Speeding



We will create an Incident dataset. Following features will be used: "Incident Date", "Incident Time", "CollisionDescription”,”Weather”,”Road condition”,”Light Condition” indicating the Incident Detail.

**Modeling**

* Using Numpy, Scarlar, Linear regression on the clean transformed data (Shown Above)

After importing necessary packages and splitting preprocessed data into test and train sets, for each machine learning model, we will build and evaluated the model with the techniques as follow:

Machine Learning Models

* K Nearest Neighbour (KNN)
* Decision Tree
* Linear Regression

Based on the dataset and the ML methods above we can predict impact on whether travel could result in property damage (class 1) or injury (class 2) and the road conditions hazards and safety.