US CAR ACCIDENTS

A Countrywide Traffic Accident Dataset (2016 - 2019)

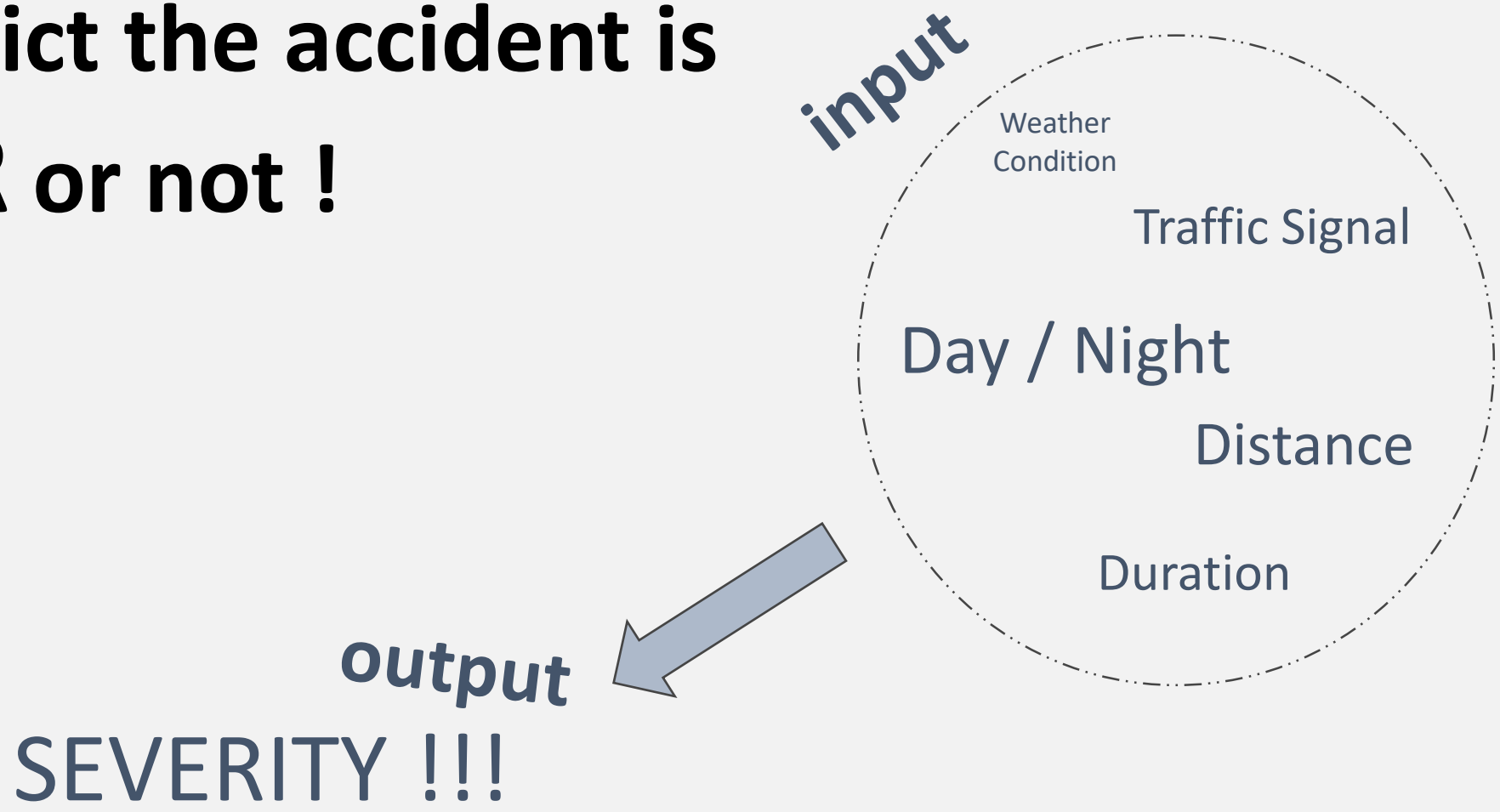
Group9
Eric Lee, Cindy Chan

Data source

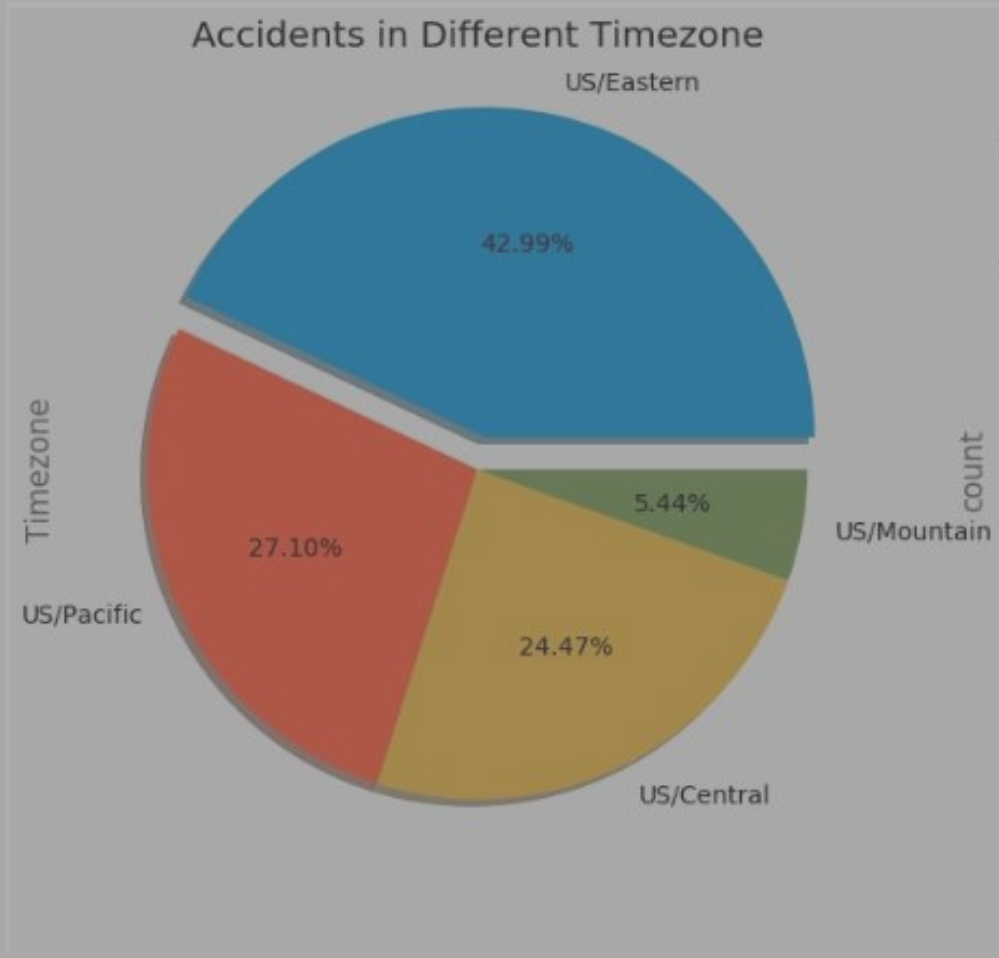
- from **Kaggle**
- **Labeled** Dataset
- Each Row is an **ACCIDENT**
- **2.24millions** rows x **49 columns**

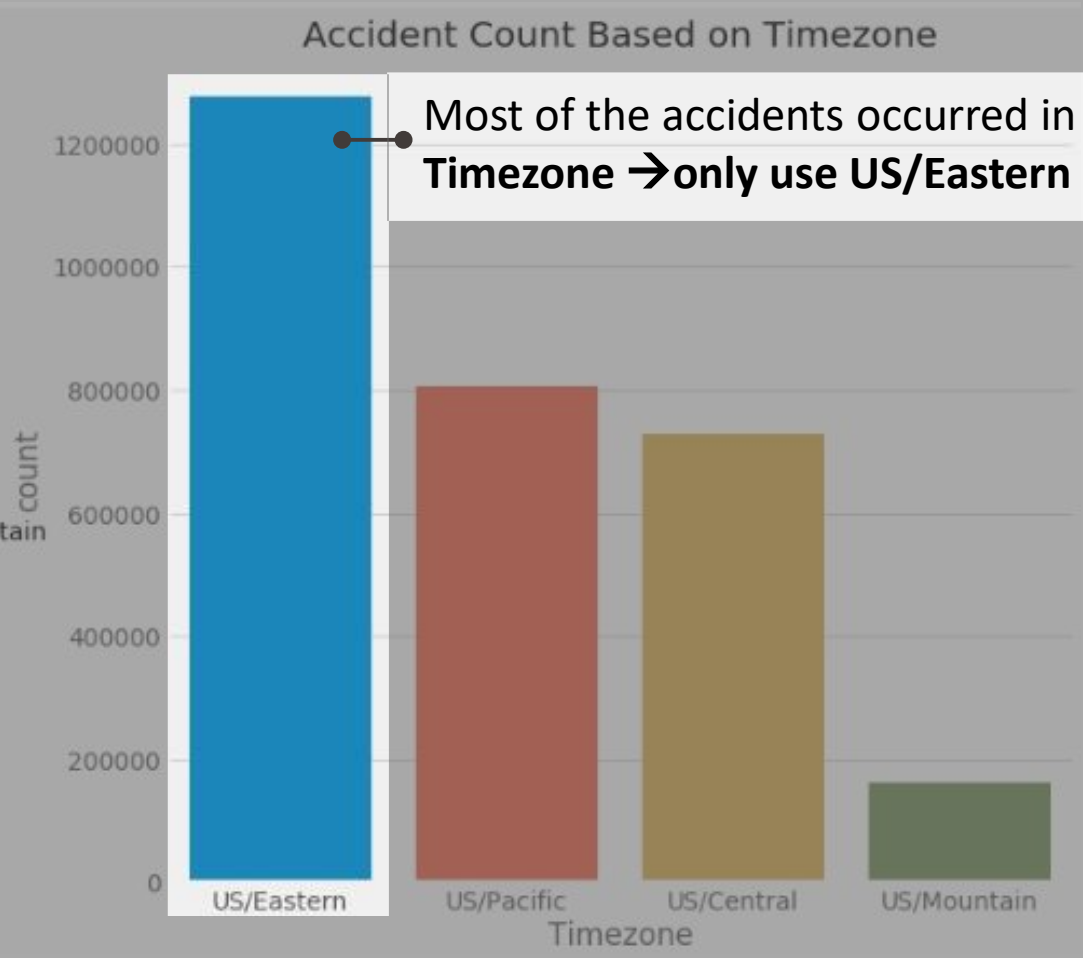
Goal

To predict the accident is **SEVER** or not !



Preprocessing Data





Most of the accidents occurred in US/Easten

Timezone →only use US/Eastern

Preprocessing Data

- Add columns

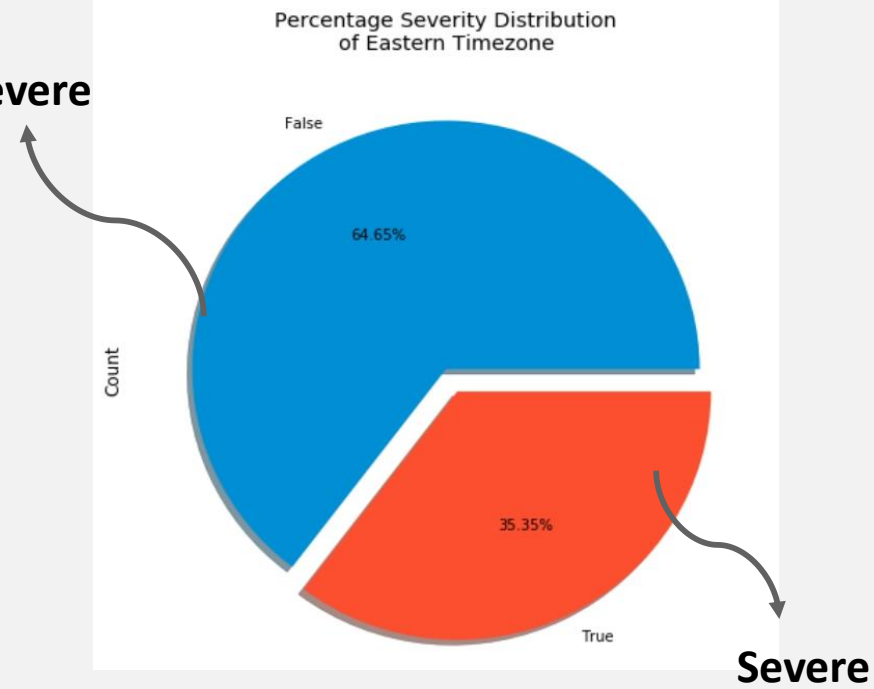
Start_Time	End_Time
2018-03-12 20:53:20+00:00	2018-03-12 21:21:43+00:00
2018-03-14 19:11:13+00:00	2018-03-14 19:55:05+00:00
2018-03-16 13:04:31+00:00	2018-03-16 13:48:33+00:00
2018-03-16 13:29:47+00:00	2018-03-16 14:29:12+00:00
2018-03-19 05:36:39+00:00	2018-03-19 06:20:09+00:00

→

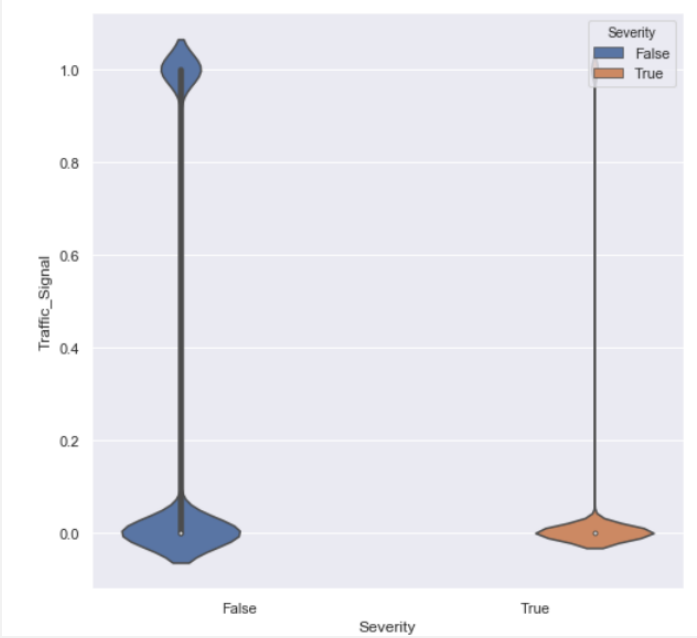
Duration
29.0
43.0
59.0
29.0
29.0

- Simplify Label

Sever (3,4) & Not Severe (1,2)



- Select Features
  - Two/Three-Dimension
  - Count plot
  - Violin plot



Remove Columns

Single value 11

- Country
- Turning_Loop
- Traffic_Calming
- Give_Way
- No_Exit
- Railway
- Roundabout
- Stop
- Bump
- Pressure(in)
- Precipitaion(in)

Irrelevant 17

- ID
- Source
- Description
- County
- State
- City
- Number
- Street
- End_Time
- Start_Lat
- Start_Lng
- Airport_Code
- Weather_Condition
- Wind_Direction
- Humidity
- Wind_Speed
- Wind_Chill _F

Alternative 7

- Timezone
- Start_Time
- Zipcode
- Weather_Timestamp
- Sunrise_Sunset
- Nautical_Twilight
- Astronomical_Twilight

Null 3

- TMC
- End_Lat
- End_Lng

Remove 38 Columns &

Drop all **the rows** that have **null value**

Label & Features

Features

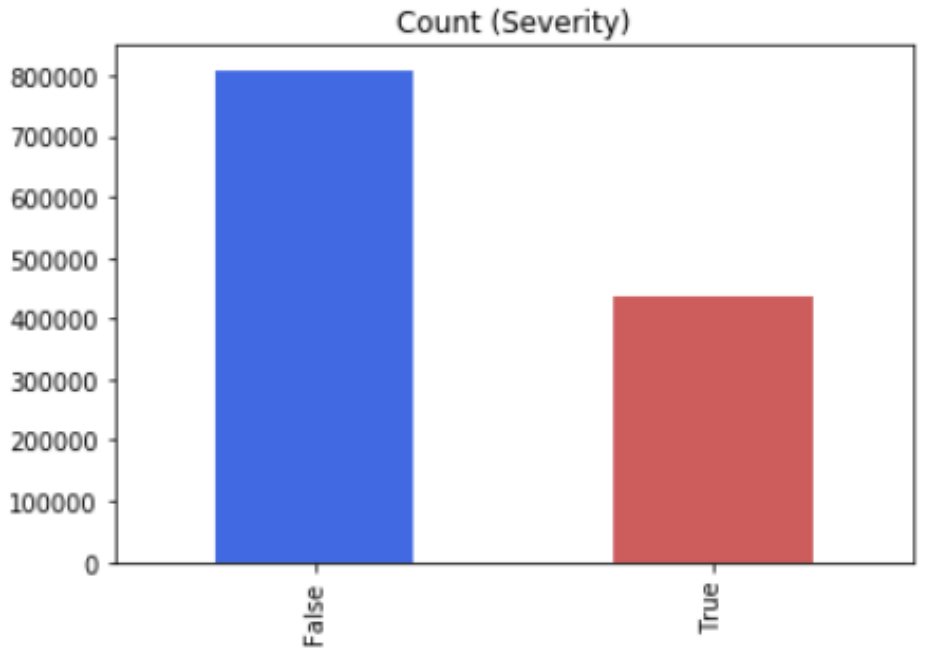
Label

	Severity	Duration	Distance(mi)	Side	Visibility(mi)	Amenity	Crossing	Junction	Station	Traffic_Signal	Twilight
0	True	314.0	0.010	1	10.0	False	False	False	False	False	False
1	False	30.0	0.010	0	10.0	False	False	False	False	False	False
2	False	30.0	0.010	1	10.0	False	False	False	False	True	False
3	True	30.0	0.010	1	9.0	False	False	False	False	False	True
4	False	30.0	0.010	1	6.0	False	False	False	False	True	True

Remain **11** Columns

Balance Data Down-sample Method

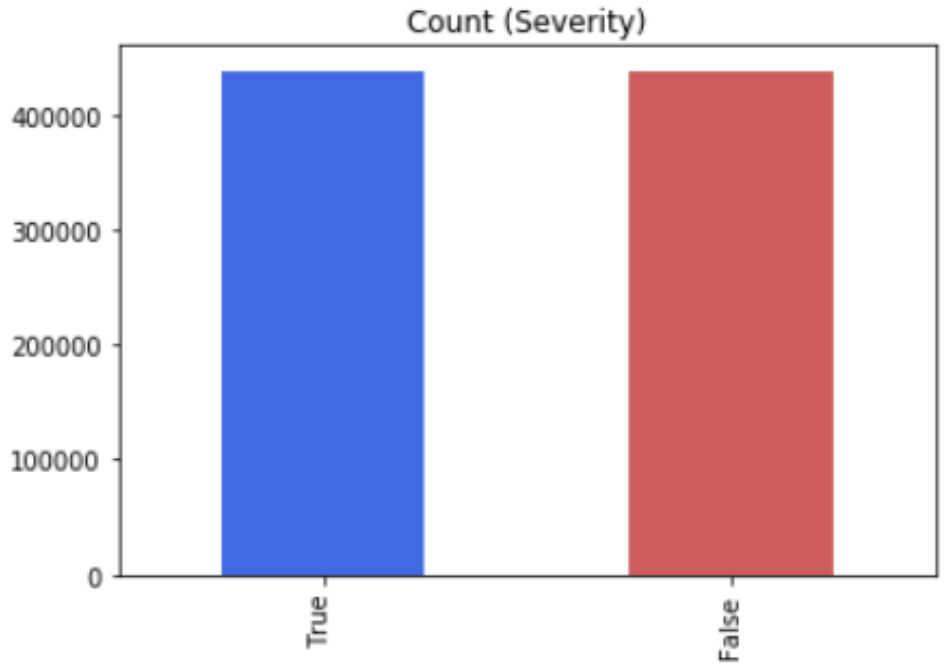
```
Class False: 809064
Class True: 438054
Proportion: 1.85 : 1

<matplotlib.axes._subplots.AxesSubplot at 0x11e6490>
```



Not Severe 65% , Severe 35%

→

```
Random under-sampling:
True     438054
False    438054
Name: Severity, dtype: int64
```



Not Sever , Severe 50%

Preprocessing Data

Original :

2.24 millions rows x **49** columns

↓

Remain:

0.88 millions rows x **11** columns

↓

Normalize!!!!

Method

Supervised Learning

MDC

KNN

Perceptron

UnSupervised Learning

K-Means

Supervised

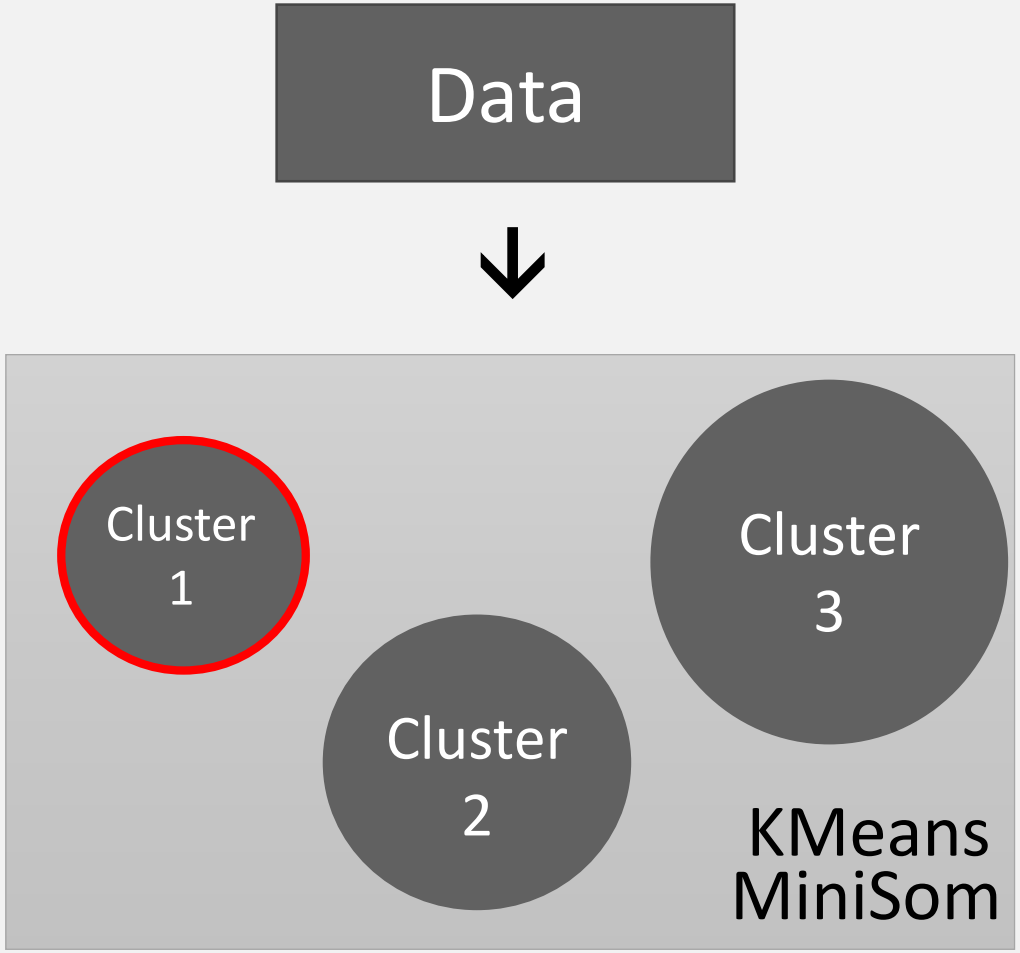
MDC

16%

Perceptron

56%

UnSupervised + Supervised

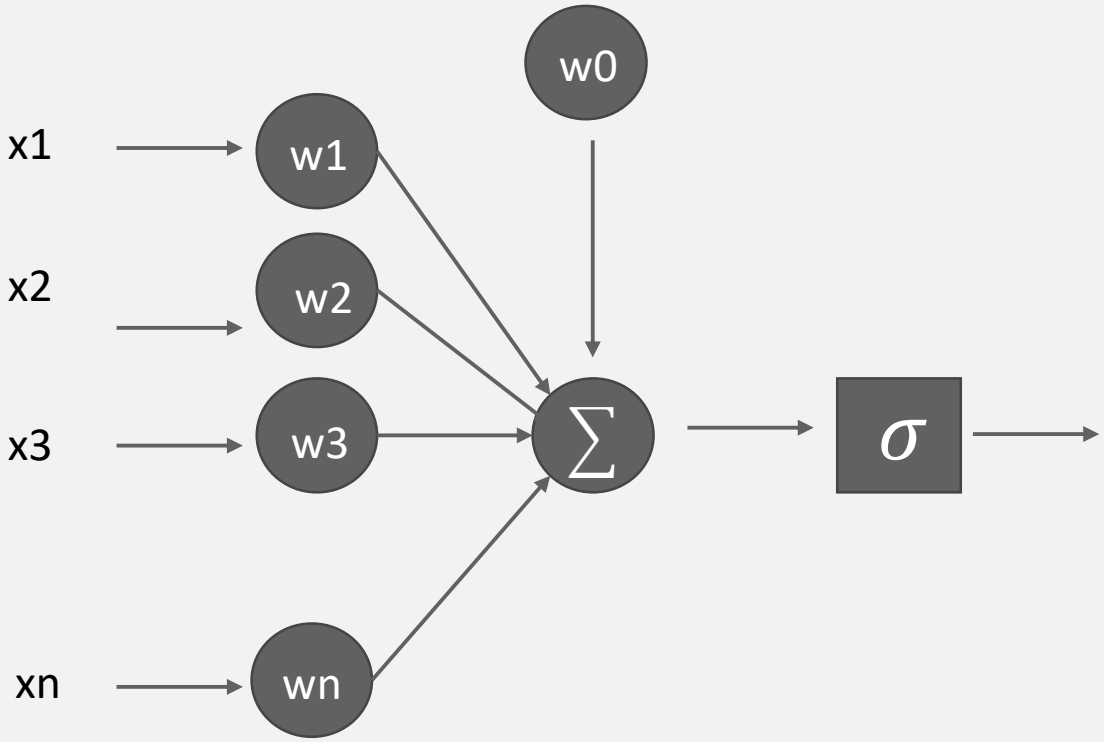


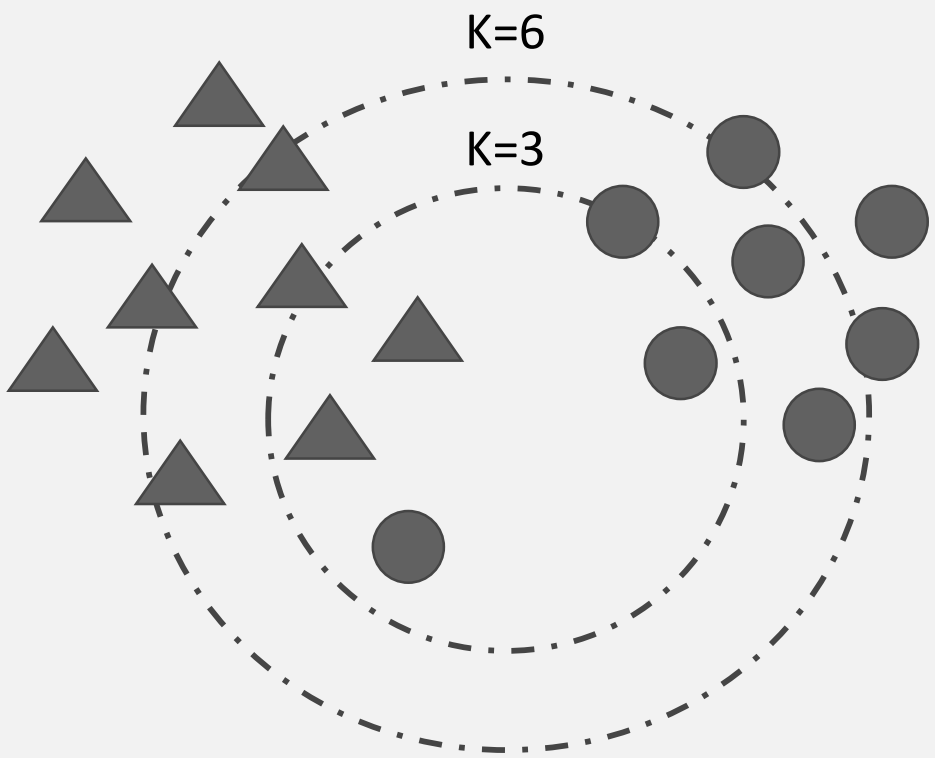
Cluster 1

Accuracy?

Perceptron

KNN

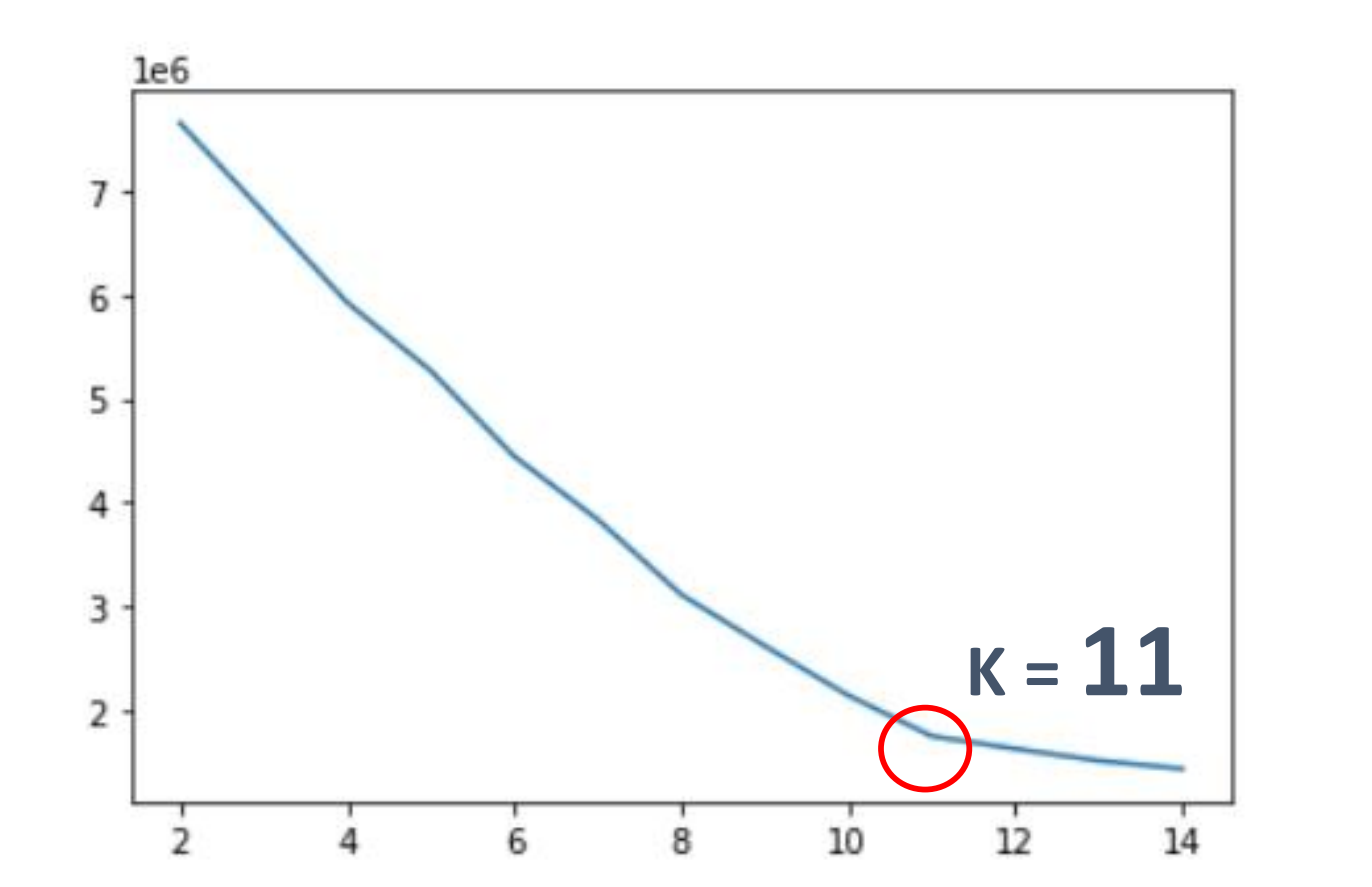


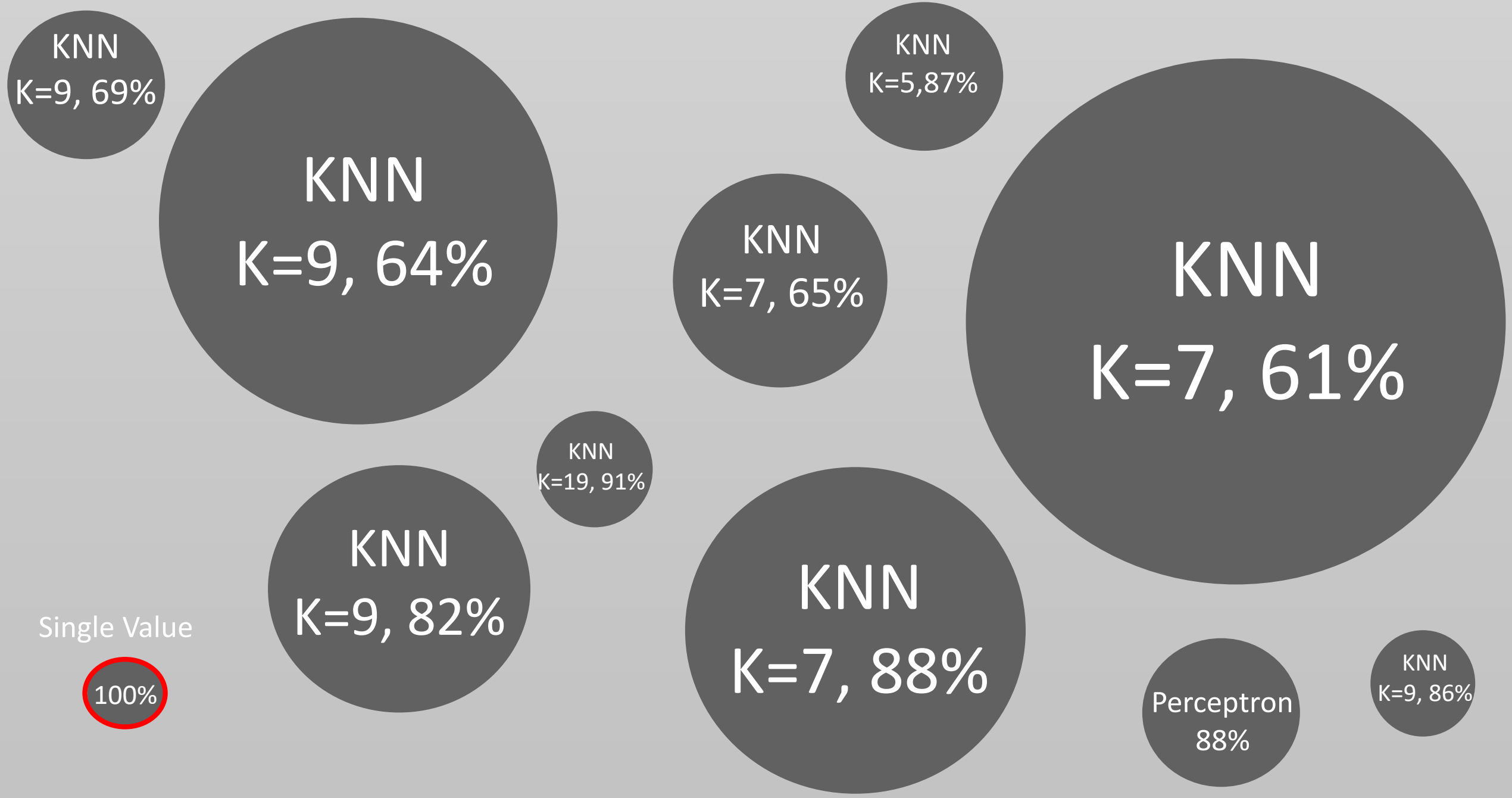


Unsupervised

Kmeans

Loss Function- Sum of Square Error




KNN
K=9, 69%
KNN
K=9, 64%
KNN
K=5,87%
KNN
K=7, 65%
KNN
K=7, 61%
KNN
K=19, 91%
KNN
K=9, 82%
KNN
K=7, 88%
Single Value
100%
Perceptron
88%
KNN
K=9, 86%

Overall

Accuracy

63%

Thanks for listening!

Reference

MainData Source: https://www.kaggle.com/sobhanmoosavi/us-accidents

US TimeZone Map : https://www.timetemperature.com/tzus/time_zone.shtml

Visualization: https://towardsdatascience.com/usa-accidents-data-analysis-d130843cde02

https://www.kaggle.com/biphili/road-accidents-in-us

https://arxiv.org/pdf/1909.09638.pdf