# Predicting the severity of car accident

# Predicting the severity of car accident is valuable for drivers, government and insurers

- **Background**

Car accident causes damages to both properties and injuries, even death. Everyone wants to avoid it.

- **Business Problem**

To predict the severity of having a car accident by taking into consideration multiple factors that could lead to a car accident.

- **Interest**

- **Drivers:** they would be interested in it for their own health and happiness
- **Government:** the prediction could help to decrease local the injury or fatality rate.
- **Insurers:** by preventing the car accident from happening, the prediction can help to decrease the insurance policy claim cost

# Data acquisition and cleaning

- **Data source**

- Given example data set by Coursera
- The Data has 38 types of car accident relevant factors and include 1 946 734 rows of records

- **Feature selection**

7 features and most are categorial attributes:

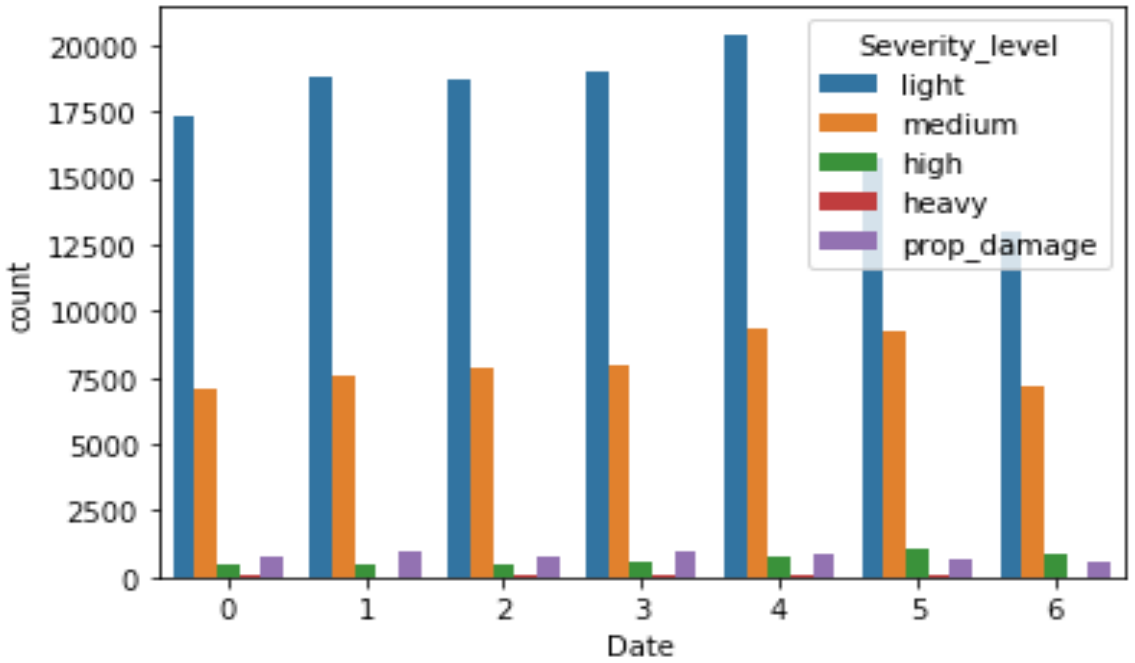
- INCDTTM: The date and time of the incident.
- INATTENTIONIND: Whether or not collision was due to inattention. (Y/N)
- UNDERINFL: Whether or not a driver involved was under the influence of drugs or alcohol
- WEATHER: A description of the weather conditions during the time of the collision
- ROADCOND: The condition of the road during the collision
- LIGHTCOND: The light conditions during the collision

- **Data cleasing:**

- Handling with missing values: assumption, drop lines and fill with means
- Convert from text to int for further modeling

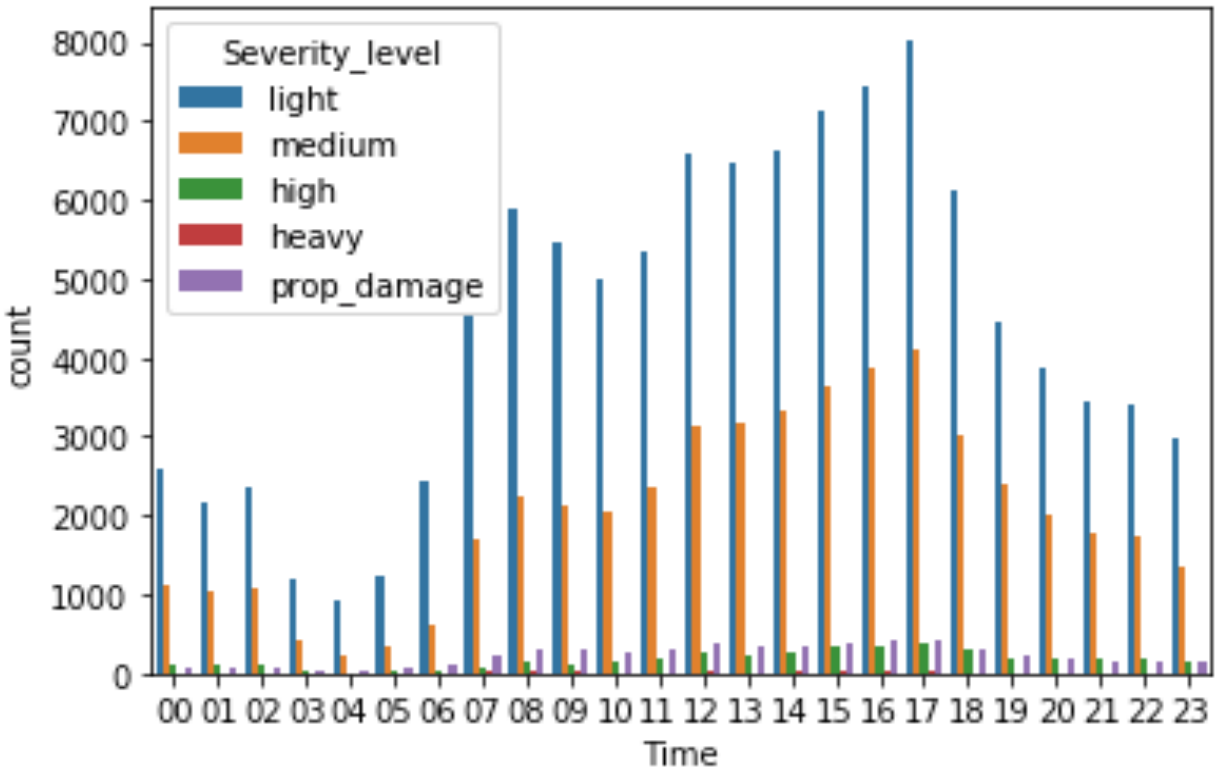
# Relationship between severity and weekday

The frequency of light car accident is significantly lower during weekend (number 5 and 6). Besides, Friday is more likely to have car accident no matter of the severity of the accident.
Please note that based on the code rule of pandas, 0 stands for Monday, 5 stands for Saturday and 6 represents Sunday.



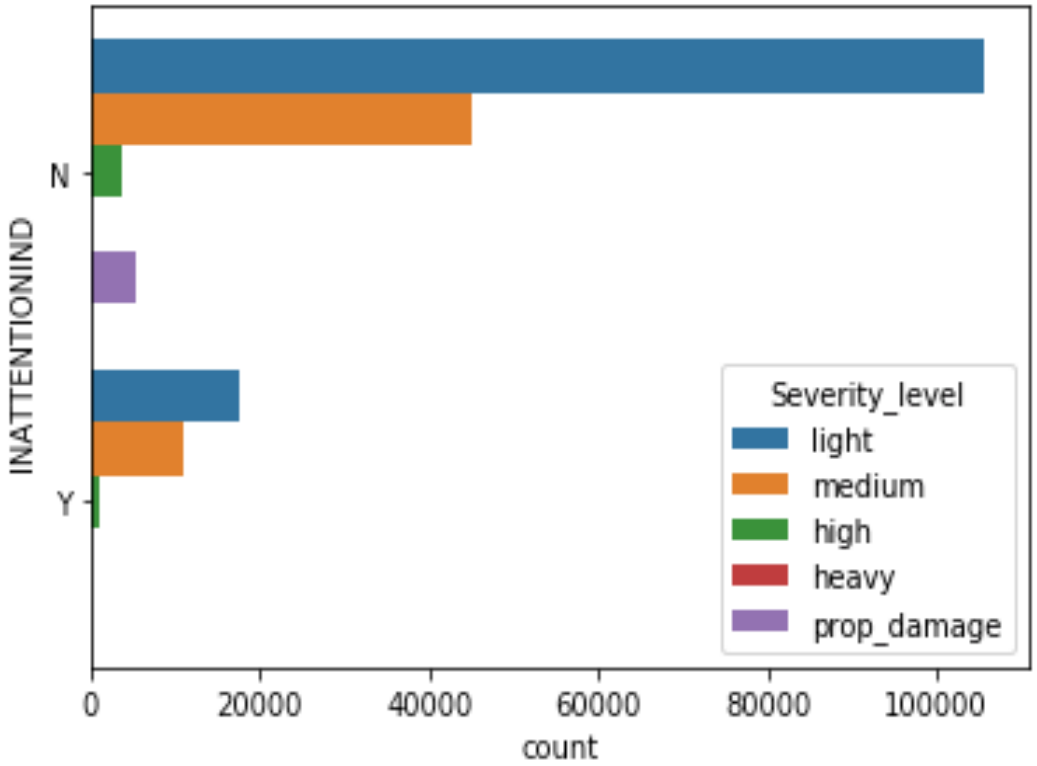

# Relationship between severity and day time

The frequency of car accident starts to increase after noon and reaches a peak at around 17h no matter of the car accident severity level. It could be that in the afternoon, drivers tend to be fatigue and is less focused.



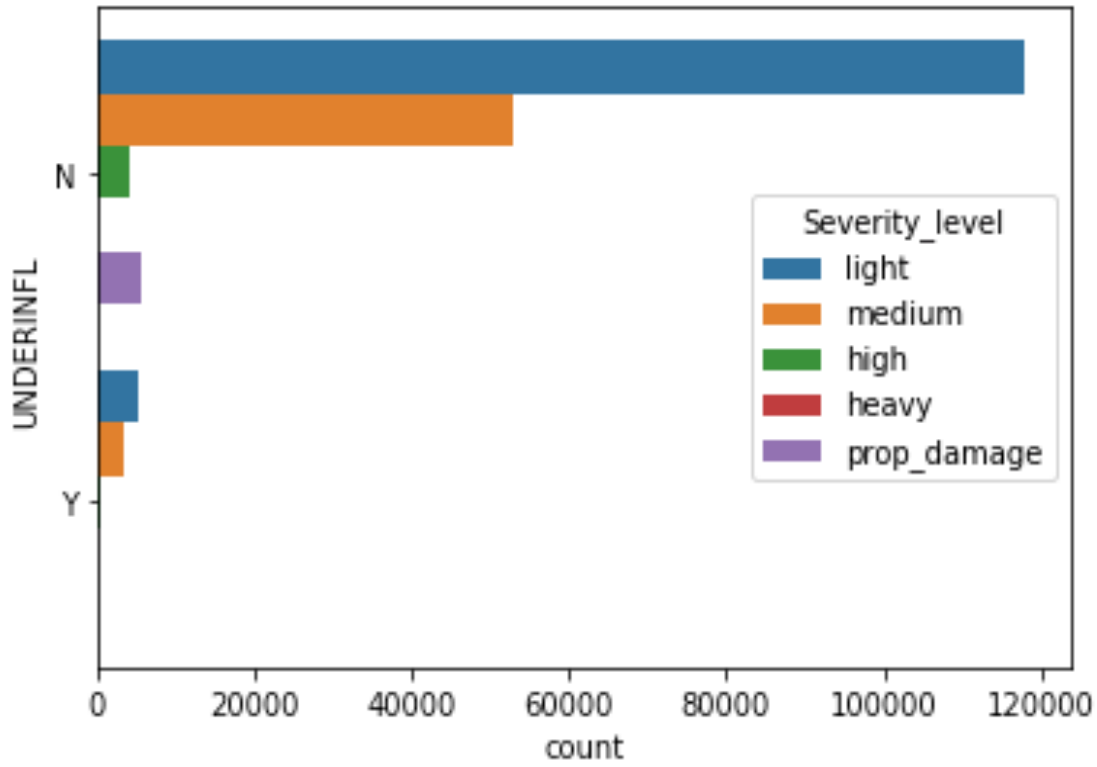

# Relationship between severity and inattentioning

A certain number of car accidents happens due to lacking of attentions. More accidents happen for others reasons no matter of the severity level of the car accident.



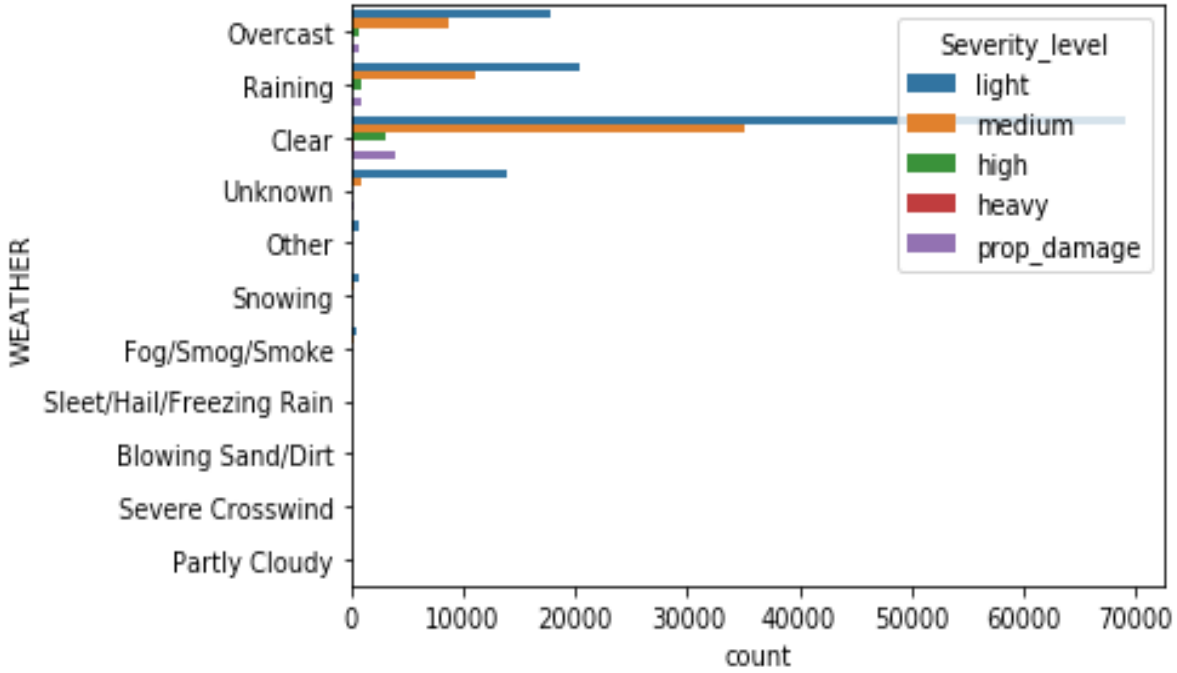

# Relationship between severity and UNDERINFL (impact of alcohol pr drugs)

A certain number of car accidents happens due to alcohol impact. More accidents happen for others reasons no matter of the severity level of the car accident.



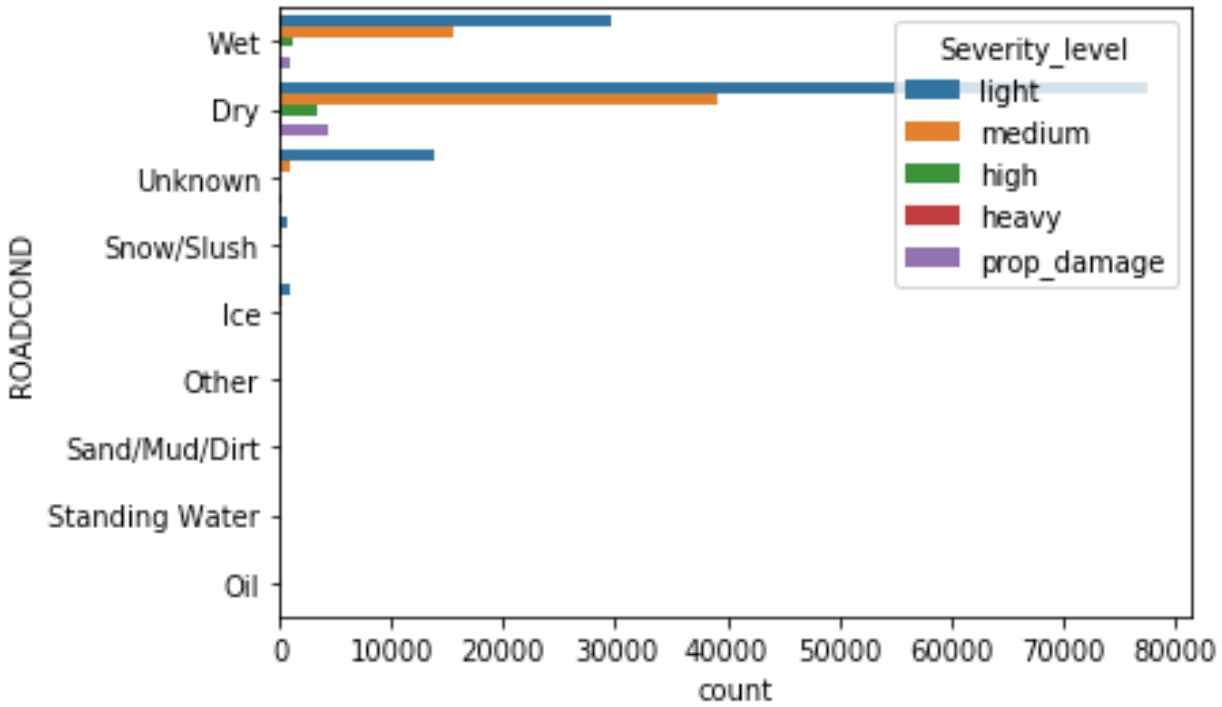

# Relationship between severity and weather

It's more likely to have car accident when the weather is clear no matter of the severity of the accident. Based on the generated graph below, Clear, Raining and Overcast are top three weathers that have the most cases of car accident.



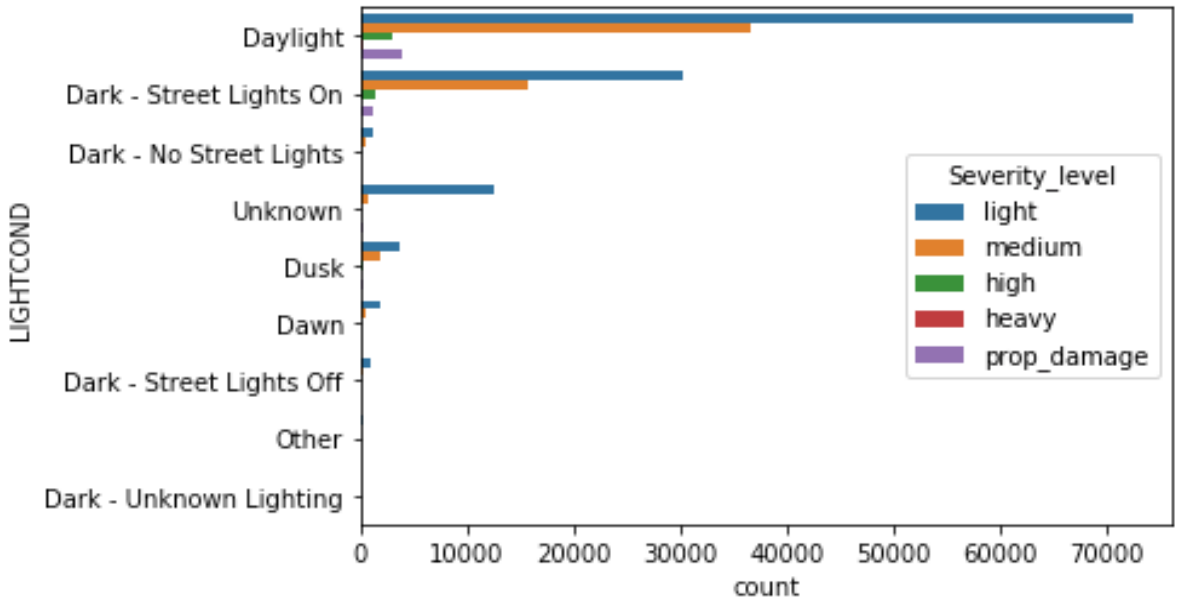

# Relationship between severity and road condition

Most car accidents happen when the road condition is dry, the next high one is wet.



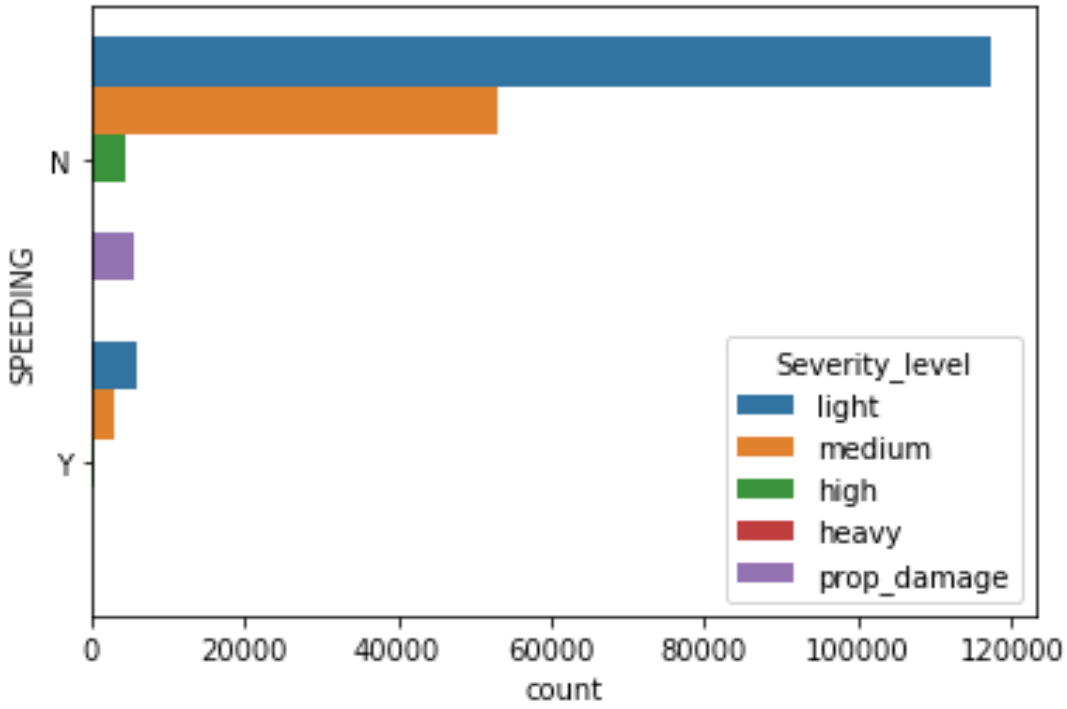

# Relationship between severity and light condition

Most car accidents happen when there is daylight, the next high one is when it is dark but the street lights on. It seems that the more normal the light condition, the more likely that driver get diverted and lead to car accident.




# Relationship between severity and speeding

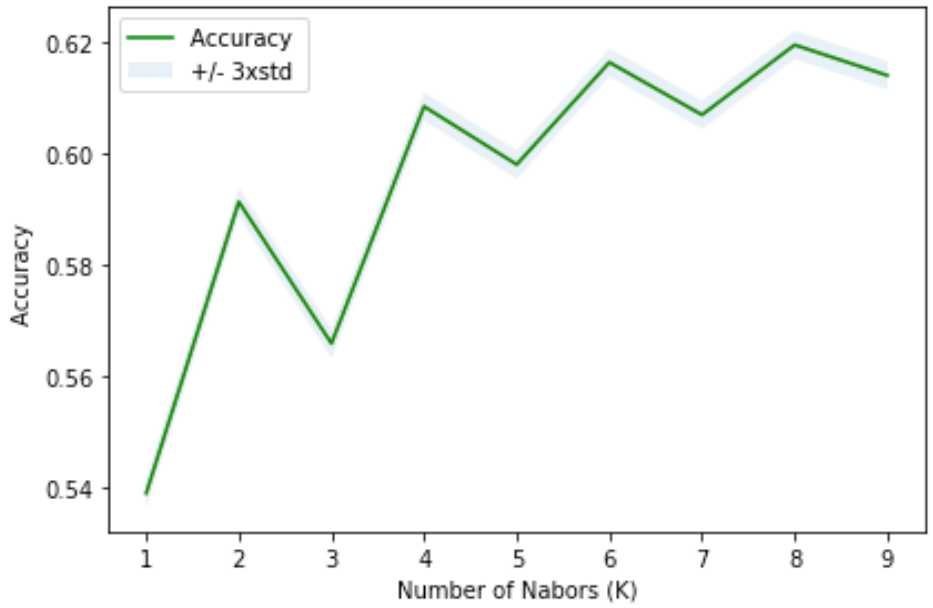
A certain number of car accidents happens due to over speed. More accidents happen for others reasons no matter of the severity level of the car accident.




# Classification Models

- **K-nearest Neighbors**

The best accuracy was with 0.62 with k= 8:




- **Decision Tree**

The decision tree's accuracy is 0.64.

- **Support Vector Machine (SVM)**

I use the train set to fit the model and test set to generate yhat.

- **Regression**

I use the train set to fit the model and test set to generate yhat probability.

# Evaluation Method: F-1 score, Jaccard, Log loss

| | F1-score | Jaccard | Log loss |
|---|---|---|---|
| **K-nearest Neighbors** | 0.55 | 0.61 | na |
| **Decision Tree** | 0.51 | 0.65 | na |
| **Support Vector Machine** | 0.51 | 0.65 | na |
| **Regression** | 0.51 | 0.65 | 0.82 |

End

# Conclusion & Future Direction

- **Conclusion**

Built useful models to predict the severity of the car accident. ~64% accuracy was achieved in the classification model.

- **Future Direction**

- **The example data set has some missing values:** I made some assumptions to fill the missing values. If we could get more qualified data or data from other data source, the prediction could be more accurate.
- **The type of attributes is not diversified:** As most of the available attributes are categorial, it could be hard to train other types of model such as non-linear regression.
- **The intercorrelation of attribute is not clear and could potentially impact the accuracy of the prediction.** For example, the weather could impact the road condition and the impact of alcohol may have correlation with inattentioning.