**Data**

The data set in this project can be found in <https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv>

This data set contains driving conditions, the number of people and vehicles involved in the crash, and the severity of the crash. There are several questions about this data set. Some entries lack key data required by this algorithm. For example, in some columns, the number of vehicles or injured persons is "unknown". In order to solve this problem, I had to delete those rows that I thought were filled in with average or frequency. These rows could not accurately represent car accidents.

### Methodology

After cleaning the data, I only drew three charts. One is a simple scatter plot with vehcout and the number of personnel. It can be seen from the chart that when the vehicle speed is equal to 2, its personnel and property losses are the largest.

### A screenshot of a cell phone Description automatically generated

Another chart I made to help analyze the data set is collision code and frequency. It plots the number of specific conflicting codes. However, compared with the above code, the conflict code is more specific, pointing out that the conflict of code 10 or 11 is a conflict of "entering at a certain angle" or "all straight, moving, sliding", and the rest of the numbers.

A screenshot of a cell phone

Description automatically generated

The final graph is a scatter plot that only considers property losses involving the number of people and vehicles and collisions involving personal injury. From below, we found that property damage accidents mainly involved two or more vehicles and multiple people.

A screenshot of a cell phone

Description automatically generated

In order to predict the severity of the accident, I made KNN k the closest. The KNN implementation shows that the accuracy of the k value is between 1 and 9 (including 1 and 9) to see which value leads to a higher accuracy value.

As shown in the figure, I found that the highest accuracy is around 0.699 when the k value is 9. Then, I predicted the value of y, using the knn model, which produced 12 correct valuses out of 20 values. For predicting the severity of a car accident, I can obtain an accuracy of about 68.5%.

A close up of a map

Description automatically generated

### discussion

Based on the results, I believe that if I combine more variables to predict the target variable, the severity and accuracy will be higher. In addition, the project sparked thinking, what should we do if we are not predicting the severity of a car accident after the accident, what should we do if we use data from previous crashes, roads, weather, lighting conditions, and speeding to predict the likelihood ? Get into a car accident under these conditions.

### conclusion

In this study, I analyzed the relationship between the number of injured, the number of damaged vehicles, what kind of collision occurred, and the severity of the collision. I build a classification model, the k nearest neighbor model, to predict car accidents