Data planning:

Now the main component of the data would be Data-Collision.csv from the Coursera course that you can find the Example of it by [clicking here](https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv). You can also find the Metadata by [Clicking here](https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf).

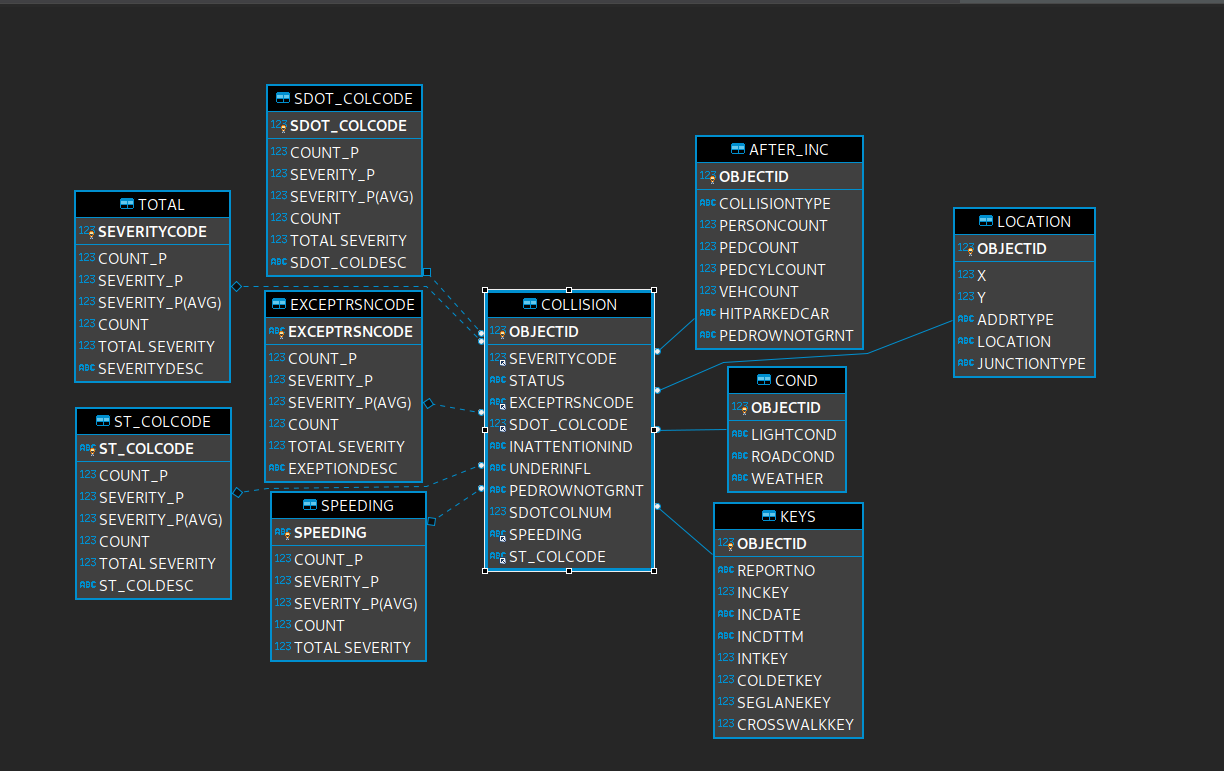
The first step is to group the features for ease of access. This can be done in multiple ways:

* With the power of the relational database, it is possible to create a connection with each other. Of course, it would be much easier if the relational database was been created and used in data itself
* Using pandas to differentiate different grouping styles

I choose the first method, as I like to work with sql.

So here is our division:

This is ERD(Entity Relationship Diagram) between 10 tables, which I created based on the main dataset.



There are 2 main types of grouping.

1. I group the similar features and extract it from the main dataset.

2. I group a feature and relevant details about it.

The first type of grouping is presented to the right of the main table(collision).

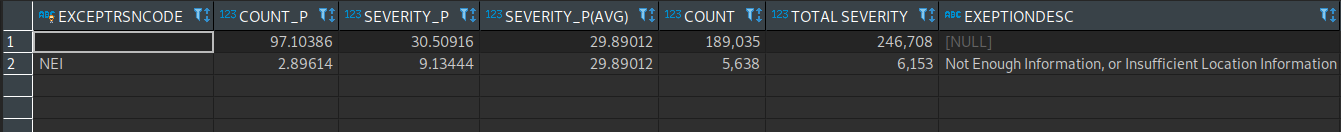
As you can see every table(to the right of collision) has OBJECTID, extracted from the main table. Then there are a group of features extracted from the table. For example, take a look at COND table shown below



As you can see some pattern emerged. I.e. if it is Raining it is certainly Wet on Road. This kind of division clearly expresses the relationship between certain variables of the table and decreases the variables from the main table(Condition). These type of groups will be processed and connections between variables will be spotted.

So we move to the second type of grouping: grouping according to 1 feature and details about it.

As an example of which is EXEPTRSNCODE table:

It groups according to EXEPTRSNCODE feature and counts what percent of it (in this case 2.89614) have exception code of NEI. If we look at the metadata we don’t find relevant details about NEI, but if we look at the EXEPTIONDESC we see that it describes NEI as Not Enough Information, or Insufficient Location Information. This gives us clue about dataset as a whole. It also tells us what percent of data have NEI(2.89614), how severe are accidents which have NEI (9.13%), what is average Severity(on average 29.89% of accidents involve injury) and how many accidents are classified as NEI (5638) and TOTAL SEVERITY (sum of SeverityCode which was classified as NEI), and so on. This is done to understand what the descriptions say(in a nutshell), because sometimes reading only descriptions will open load of information about data.

Data preparation: This will be ready after week 3 on coursera