**Cleanup Process**

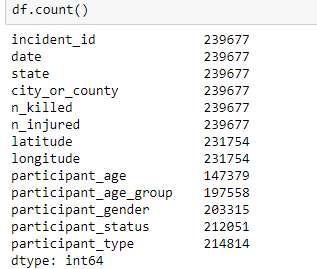
Data Set Characteristics

* Rows 240K rows
* Columns 29
* Size – 142 Mb

Fields provided

|  |
| --- |
| address |
| city\_or\_county |
| congressional\_district |
| date |
| gun\_stolen |
| gun\_type |
| incident\_characteristics |
| incident\_id |
| incident\_url |
| incident\_url\_fields\_missing |
| latitude |
| location\_description |
| longitude |
| n\_guns\_involved |
| n\_injured |
| n\_killed |
| notes |
| participant\_age |
| participant\_age\_group |
| participant\_gender |
| participant\_name |
| participant\_relationship |
| participant\_status |
| participant\_type |
| source\_url |
| sources |
| state |
| state\_house\_district |
| state\_senate\_district |

* Not all fields were required. At the same time a lot of them had null values. We ran tests to ensure we had enough data to ensure sufficient visualizations.



The top 8 fields in this had useable data, after these fields the null values increased and the we loose a lot of data. Hence we decided to split up the data into smaller files vertically. This allowed us to work separately with the specific fields that each person needed for their visualizations with out having to use such a big file.

Selection Set 1

fields=['incident\_id',

'date',

'state',

'city\_or\_county',

'n\_killed',

'n\_injured','latitude',

'longitude', 'participant\_age',

'participant\_age\_group',

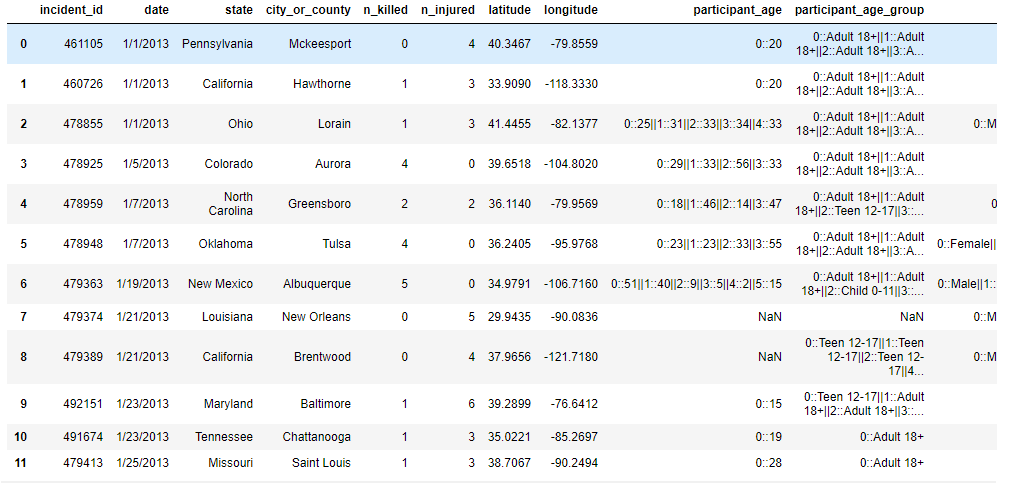
'participant\_gender',

'participant\_status',

'participant\_type']

df=pd.read\_csv(file,usecols=fields)

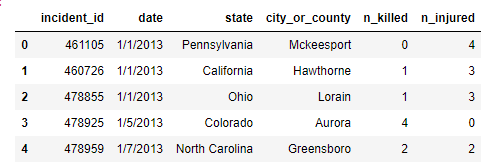
df.head(12)



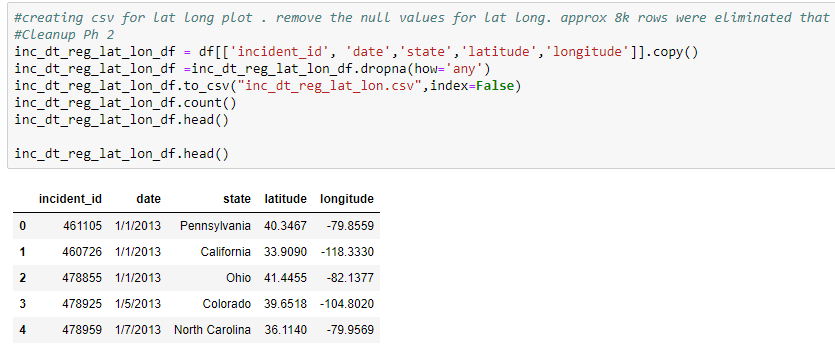
Selection set 2

inc\_dt\_region\_df = df[['incident\_id', 'date','state','city\_or\_county','n\_killed','n\_injured']].copy()

inc\_dt\_region\_df.head()



Selection set 3

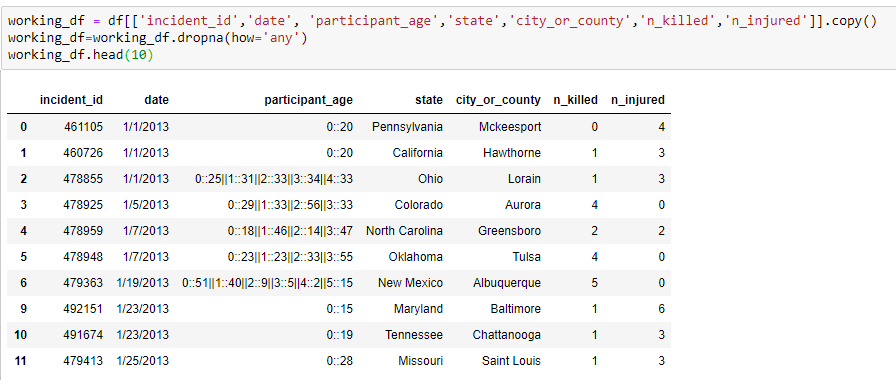


Three Sets of files allowed all of the members to get working without interfering with the original dataset

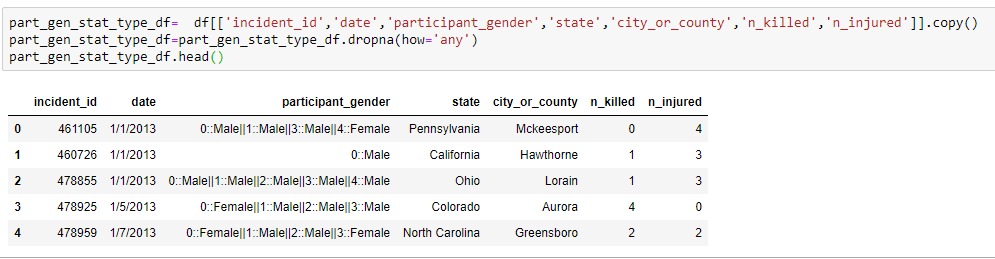
Age Cleanup and creating rows

The age column had multiple people ages in one row as it was based on per incident

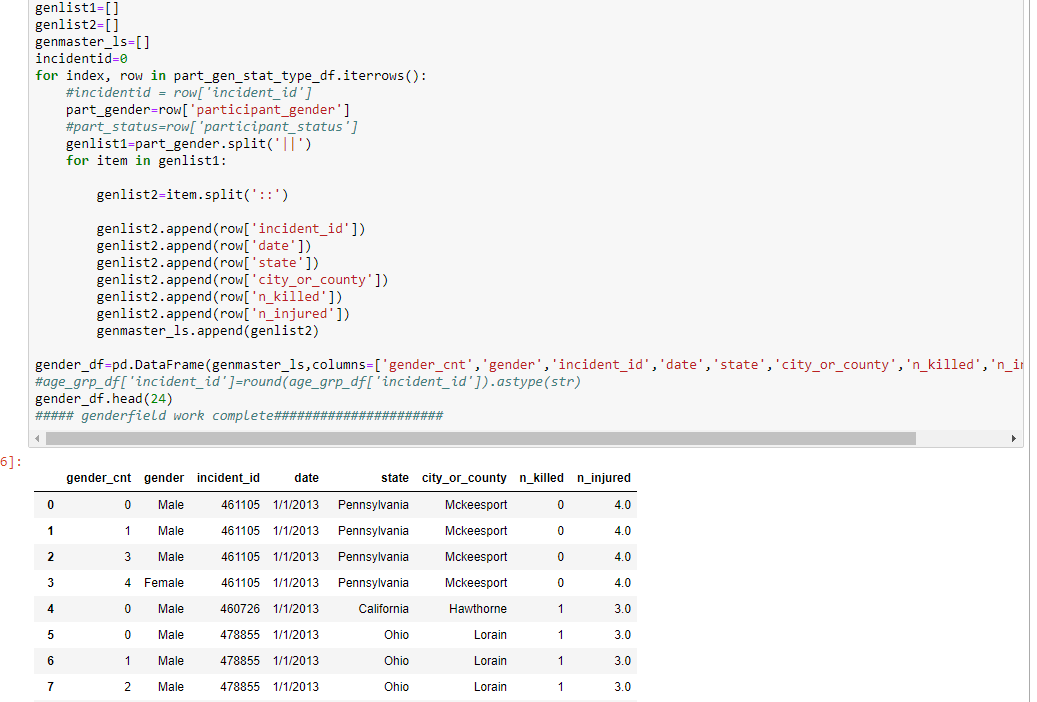
In order to break this to multiple rows and get each age for that incident in one row, code had to be written



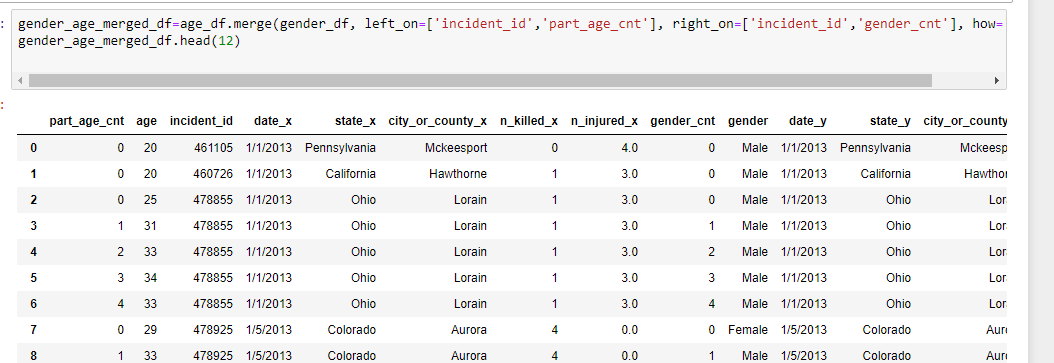
We had the same issue with Gender



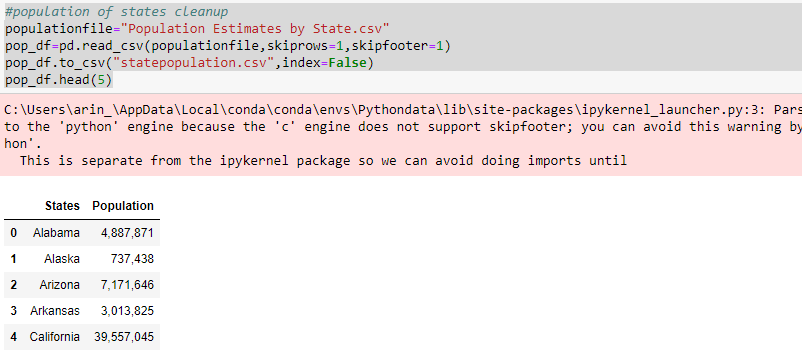
A cleanup of break up was required and this is how it looked after the cleanup for the gender



Gender and the age data-frames were merged to form a working data-frame



The next file was the state population that had to be cleaned up as it had an extra header and extra footer



The third and final file also had extra rows that needed removal from top and bottom

