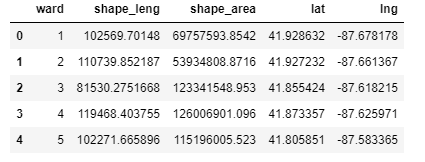
**Raw Data include:**

**Chicago geo data:**

- Geodata is acquired from data.cityofchicago.org in json format, ward number and coordinates are extracted.

- Use ward instead of neighborhood name as ward is more “official” and better defined.



**Public transportation -- CTA L train and Divvy bike share data:**

- Both data are aquired in json format, station/stop names and coordinates are extracted into data frame.

- Distance between each ward and each station/stop is calculated use coordinates, then total count of stations/stops within a defined “distance” to each ward is summed up, and added to the data set.

def distance(lat1, lon1, lat2, lon2):

radius = 6371 #--km

dlat = math.radians(lat2-lat1)

dlon = math.radians(lon2-lon1)

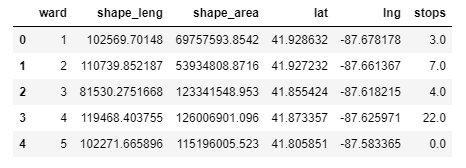
a = math.sin(dlat/2)\*math.sin(dlat/2)+math.cos(math.radians(lat1))\*math.cos(math.radians(lat2))\*math.sin(dlon/2)\*math.sin(dlon/2)

c = 2 \* math.atan2(math.sqrt(a), math.sqrt(1-a))

d = radius \* c

return d

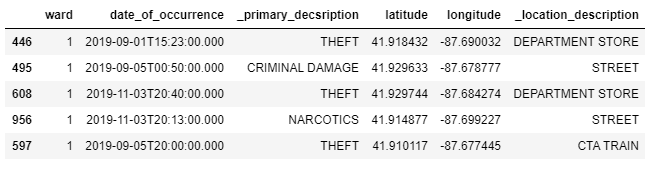


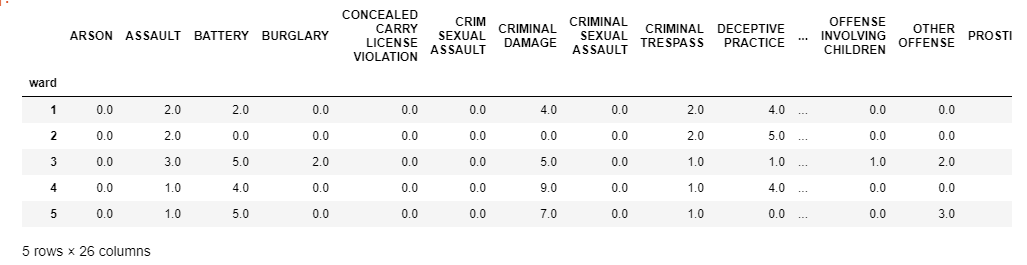


**Crime data:**

- Data is aquired in json format, crime type, ward, coordinates are put into a data frame.

- Group by ward and crime type, then pivot to calculate a total crime count for each ward, before added to dataset.





**foursquare dataset:**

- First all venue data are gathered using explore, radius set to 10km (~6mi)

- category of venue is investigated, and divide into three type:

1. enviroment: such as park, beach, garden, museum, stadium
2. unrelevant: such as bank, apartment
3. retail: all stores provide goods/services, such as restaurant, cafe, spa

- Frequency data is generated based on retail category only, then top venue of each ward is calculated use this frequency

- Venue count is further grouped by category, into:

1. cafe: cafe, coffee shops, game cafe
2. other drinks: tea house, juice bar
3. snack: bakery, donut shop, ice cream shop, froyo shop

