Gépjármű ütközések elemzése New York területén

Házi feladat specifikáció

Készítette: Harsányi Zsolt

Neptun: XORZVV

2024

Budapesti Műszaki és Gazdaságtudományi Egyetem
Méréstechnika és Információs Rendszerek Tanszék

1 A specifikáció célja

A specifikáció célja a tárgy házi feladatának megalapozása, a választott adatkészlet kiértékelése és kérdések megfogalmazása. A házi feladat feltételeit a vonatkozó előadás tartalmazza.

Javasolt, de nem kötelező a jelen minta használata. Egy megfelelő szöveges magyarázattal ellátott Jupyter Notebook a belőle generált HTML állománnyal együtt, vagy tetszőleges markdown állomány, esetleg kezdeti dashboard is elfogadható, ha a kötelező elemeket tartalmazza.

Metaadatok:Feladat címe (pl. *Fluxuskondenzátor hatékonysági adatainak vizsgálata*), hallgató neve, Neptun-kódja (ebben a sablonban a címlapon).

2 Kiválasztott adatkészlet

A választott adatkészlet az NYC OpenData egyik leghivatkozottabb adatkészlete, a New York területén történt gépjárműbalesetekről szóló adatkészlet, amelyet a NYPD (New York Police Department), vagyis New York-I rendőrség frissít a mai napit 2014.05.07 óta, tehát már egy 10 éves adatkészletről beszélünk. Emiatt az adatkészlet már hosszra is egészen mutatós 2.13 millió sorból áll, méretre pedig egy szinte pont fél GB méretű .csv fájlként lehet exportálni. Emellett az adatkészlet 29 sorból áll, amit utoljára 2021.05.19-én frissítettek, tehát várhatóan az azelőtti rekordokban hiány léphet fel. (Ezt az adatkészlet elemzésének hiányában merem állítani)

3 Adatkészlet alapvető statisztikai kiértékelése

Az adatkészlet rengeteg fontos oszlopot tartalmaz ami szükséges lesz az elemzéshez. A második kép egy úgynevezett Pandas.DataFrame.info() nevű függvény eredménye ami ki listázza az oszlopokat és az oszlopok típusát. Látszik hogy rengeteg oszlop object-ként inicializálódott. Első körben ezt érdemes szemügyre venni és amit csak lehet átkonvertálni egy specifikusabb formátummá.

Példának létezik olyan oszlop hogy CRASH TIME ami egy időpont 24 órás formátumban (óra:perc) amit a Pandas szintén object-ként tárol beolvasás után. Ezt is érdemes átkonvertálni valamilyen datetime formátumra mert utána sokkal könnyebb lesz vele dolgozni. Szerencsére sok változót helyesen konvertált lebegőpontos típusúvá, vagy egész szám típusúvá a Pandas, itt valószínűleg nem volt Nan, azaz üres érték. Ezen kívül a legtöbb változó valamilyen kategorikus változó amely nominális. Ilyen például a VEHICLE TYPE CODE, amelyből 4 van és mind a négy az ütközésben résztvevő gépjárművek valamilyen besorolása egy kategóriába. Természetesen rengeteg rekord van ahol esetleg csak 2 jármű vett részt az incidensben, ott a másik kettő oszlop Nan értéket vesz fel.

[3]:

	LATITUDE	LONGITUDE	NUMBER OF PERSONS INJURED	NUMBER OF PERSONS KILLED	NUMBER OF PEDESTRIANS INJURED	NUMBER OF PEDESTRIANS KILLED	NUMBER OF CYCLIST INJURED	NUMBER OF CYCLIST KILLED	NUMBER OF MOTORIST INJURED	NUMBER OF MOTORIST KILLED	COLLISION_ID
count	1.888458e+06	1.888458e+06	2.127499e+06	2.127486e+06	2.127517e+06	2.127517e+06	2.127517e+06	2.127517e+06	2.127517e+06	2.127517e+06	2.127517e+06
mean	4.062750e+01	-7.375173e+01	3.170652e-01	1.530915e-03	5.743597e-02	7.544006e-04	2.772293e-02	1.193880e-04	2.278139e-01	6.336025e-04	3.198273e+06
std	1.981879e+00	3.722336e+00	7.063615e-01	4.132949e-02	2.458192e-01	2.803202e-02	1.662858e-01	1.096877e-02	6.676160e-01	2.755304e-02	1.506493e+06
min	0.000000e+00	-2.013600e+02	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	2.200000e+01
25%	4.066762e+01	-7.397475e+01	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	3.167999e+06
50%	4.072062e+01	-7.392712e+01	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	3.700018e+06
75%	4.076963e+01	-7.386680e+01	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	4.232132e+06
max	4.334444e+01	0.000000e+00	4.300000e+01	8.000000e+00	2.700000e+01	6.000000e+00	4.000000e+00	2.000000e+00	4.300000e+01	5.000000e+00	4.764304e+06

```
 #   Column                         Dtype
---  ------                         -----
 0   CRASH DATE                     object
 1   CRASH TIME                     object
 2   BOROUGH                        object
 3   ZIP CODE                       object
 4   LATITUDE                       float64
 5   LONGITUDE                      float64
 6   LOCATION                       object
 7   ON STREET NAME                 object
 8   CROSS STREET NAME              object
 9   OFF STREET NAME                object
 10  NUMBER OF PERSONS INJURED      float64
 11  NUMBER OF PERSONS KILLED       float64
 12  NUMBER OF PEDESTRIANS INJURED  int64
 13  NUMBER OF PEDESTRIANS KILLED   int64
 14  NUMBER OF CYCLIST INJURED      int64
 15  NUMBER OF CYCLIST KILLED       int64
 16  NUMBER OF MOTORIST INJURED     int64
 17  NUMBER OF MOTORIST KILLED      int64
 18  CONTRIBUTING FACTOR VEHICLE 1  object
 19  CONTRIBUTING FACTOR VEHICLE 2  object
 20  CONTRIBUTING FACTOR VEHICLE 3  object
 21  CONTRIBUTING FACTOR VEHICLE 4  object
 22  CONTRIBUTING FACTOR VEHICLE 5  object
 23  COLLISION_ID                   int64
 24  VEHICLE TYPE CODE 1            object
 25  VEHICLE TYPE CODE 2            object
 26  VEHICLE TYPE CODE 3            object
 27  VEHICLE TYPE CODE 4            object
 28  VEHICLE TYPE CODE 5            object
```

4 Kérdések/célok

Szerencsére véleményem szerint egy olyan adatkészletet sikerült választanom aminek segítségével rengeteg kérdést meg lehet fogalmazni és válaszokat lehet keresni rá. Ilyen kérdések lehetnek esetleg:

- Melyik városrészben történt a legtöbb baleset az elmúlt 10 évben?
- Milyen típusú gépjárművekkel történt a legtöbb baleset?
- A nap melyik szakaszában történik a legtöbb baleset?
- Több baleset történik a nap reggeli, sietős szakaszában mint azután?

5 Elemzés és vizualizáció jellege, technológiája

Az elemzéshez szeretnék egy egyszerűbb Dashboard-ot készíteni dash+plotly segítségével aminek segítségével fogom megválaszolni a kérdéseket különböző chart-ok segítségével. Mivel a Python programozási nyelv áll hozzám a legközelebb ezért szeretnék releváns tapasztalatot szerezni a Dash+plotly-ban hogy utána más, hasonló elemzésekben is tudjam majd alkalmazni.