Documentación de la semana 2

En esta segunda semana del proyecto comenzamos con una selección y limpieza profunda de los datos que nos aclaró la información que podíamos sacar de los mismos y decidir mejor los KPI.

Las fuentes de datos que elegimos fueron:

* Choques de vehículos a motor:

(<https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95>)

Esta fue la fuente principal de nuestros datos. De este dataset extrajimos la ubicación en el que ocurrió el accidente(por coordenadas, barrio o dirección), hora y fecha, quienes estuvieron involucrados en el accidente(bicicletas, autos, personas) y el daño que les fue causado, y factores que contribuyeron al accidente.\*

En cuanto a la limpieza de este dataset, para empezar tuvimos que descartar varias columnas que, o no tenían relevancia para nuestros KPI o nos mostraban información que podría llegar a ser más confusa que útil. Cuando definimos esas columnas, las descartamos y finalmente conservamos la información mencionada en el párrafo anterior.

* Choques a personas:

(<https://data.world/city-of-ny/f55k-p6yu>)

Este dataset nos brindó información más detallada sobre las personas que fueron afectadas por cada accidente como el rol (peatón, conductor, pasajero), la edad, el sexo, el daño que recibió y la posición en el vehiculo (en caso de estarlo).

En este dataset la principal limpieza que se realizó fue en las edades ya que en el original había valores negativos. Además del descarte de algunas columnas y completar los valores nulos.

* Choques entre vehículos:

(<https://data.world/city-of-ny/bm4k-52h4>)

Similar al dataset anterior, este dataset contiene detalles de los vehículos involucrados en los accidentes. De aquí sacamos datos como el estado en el que está registrado el vehículo, el estado que emitió la licencia y si hubo daño a la propiedad pública.

En la limpieza de este dataset, principalmente hubo que corregir muchos valores ingresados con errores de tipeo o variaciones de mayúsculas y minúsculas en las columnas de los factores que contribuyeron al accidente, aparte de completar los valores nulos.

####

Faltan ingresar a la base de datos

* Volumen de tráfico:

(<https://data.world/city-of-ny/btm5-ppia>)

Este dataset contiene el flujo de trafico por hora

* Circulación de bicicletas:

(<https://data.world/city-of-ny/uczf-rk3c>,

<https://data.cityofnewyork.us/Transportation/Bicycle-Counters/smn3-rzf9>)

* Rutas de camiones:

(<https://data.world/city-of-ny/jjja-shxy>)

* Historial del clima:

(<https://openweathermap.org/api>)

####

Ingesta de datos:

Para ingresar y procesar los datos, en principio utilizamos los csv que nos permite descargar cada página de las apis para trabajar localmente. Luego intentamos llamar la información directamente de las APIs pero era mucho más ineficiente si teníamos que llamar la totalidad de los datos cada vez que llamaríamos un dataset. Entonces decidimos usar los datasets en forma de csv y actualizarlos periódicamente con la api.

Servidor y base de datos y API:

Los scripts mencionados anteriormente, encargados de la limpieza, el proceso, la actualización y la manipulación de los datasets, de momento están almacenados localmente pero van a terminar corriendo en un servidor. Para esto usamos una máquina virtual con sistema operativo Debian en la nube, brindada por **DonWeb**. Esto nos permite tanto mantener abierta la base de datos como desplegar la API con **Django REST**.

Diccionario de la API:

([http://vps-2671696-x.dattaweb.com:8000/collision/](http://vps-2671696-x.dattaweb.com:8000/collision/collisions_for_day/01%2F01%2F2014/))

Ejemplo api colisiones por fecha (01/01/2014 = 01%2F01%2F2014):

<http://vps-2671696-x.dattaweb.com:8000/collision/collisions_for_day/01%2F01%2F2014/>

* API colisiones por tiempo:

<http://vps-2671696-x.dattaweb.com:8000/collision/time_crash/>

* API colisiones por zona:

<http://vps-2671696-x.dattaweb.com:8000/collision/zone_crash/>

* Causa de accidente por género:

<http://vps-2671696-x.dattaweb.com:8000/collision/factor_genere/>

* Cantidad de accidentes por género:

<http://vps-2671696-x.dattaweb.com:8000/collision/genere/>

* Accidentes vs estado que otorgó la licencia:

<http://vps-2671696-x.dattaweb.com:8000/collision/licence_state/>

* Cantidad de ocasiones que se produjeron daños a la propiedad pública:

<http://vps-2671696-x.dattaweb.com:8000/collision/public_damage/>

* Estado de la persona al momento de tener el accidente:

<http://vps-2671696-x.dattaweb.com:8000/collision/status/>

* Accidente por rol de personas en accidentes:

<http://vps-2671696-x.dattaweb.com:8000/collision/type/>

\* Columnas que conservamos: