MEMORIA TFM
“FIREMAN INTERVENTIONS”

BEATRIZ ARÉVALO SÁNCHEZ

MASTER DATA SCIENCE

El primer propósito de este TFM era conocer donde se produciría la siguiente intervención de los bomberos en la ciudad de Nueva York, en este caso ese dato nos lo proporcionaría la columna *FIRE_BOXX*. Tras limpiar el dataset, no fui capaz de hacerlo, con lo cual cambié mi objetivo y decidí que iba a ser conocer el tiempo de duración de una intervención, la columna *TOTAL_INCIDENT_DURATION.*

Las columnas que escogí para ello son: *CO_DETECTOR_PRESENT_DESC*., *UNITS_ONSCENE, ZIP_CODE, CODE_PROPERTY* y *CODE_INCIDENT*. Las dos últimas columnas son variables categóricas, las hice lineales con *get_dummies*. Tras esto ejecuté modelos de Regresión Lineal, KNN y Random Forest, en los tres me daba errores. No he sido capaz de solucionarlos. En el notebook 4.TFM_ML, podéis ver este proceso.

Por esta razón mi TFM se limita a un análisis visual de las diferentes variables que tenemos y ver como se comportan unas con otras.

1. **TFM_primeros_pasos**

En este notebook lo que se realiza es la transformación de alguna columna de los dos dataset para poder unirlos en uno solo. El dataset resultante es el llamado "df_nulls".

En el dataset principal, *"Incidents_Response_to_by_Fire_Companies"*, está compuesto por 2.518.758 filas y 24 columnas.

1. IM_INCIDENT_KEY: Identificador único para cada incidente que sirve como clave primaria
2. FIRE_BOX: Identificador del área de la caja de alarma contra incendios en la que ocurrió el incidente, único por municipio
3. INCIDENT_TYPE_DESC: El código y la descripción del tipo de categoría de incidente
4. INCIDENT_DATE_TIME: La fecha y hora en que se registró el incidente en el sistema de Despacho Asistido por la Computadora
5. ARRIVAL_DATE_TIME: La fecha y hora en que la primera unidad llegó a escena
6. UNITS_ONSCENE: Número total de unidades que llegaron a la escena.
7. LAST_UNIT_CLEARED_DATETIME: La fecha y hora en que se completó el incidente y la última unidad despejó la escena.
8. HIGHEST_LEVEL_DESC: El nivel de alarma más alto que recibió el incidente.
9. TOTAL_INCIDENT_DURATION: El número total de segundos desde que se creó el incidente hasta que se cerró
10. ACTION_TAKEN1_DESC: El código y la descripción de la primera acción tomada.
11. ACTION_TAKEN2_DESC: El código y la descripción de la segunda acción tomada.
12. ACTION_TAKEN3_DESC: El código y la descripción de la tercera acción tomada.
13. PROPERTY_USE_DESC: El código y la descripción del tipo de calle o edificio donde ocurrió el incidente
14. STREET_HIGHWAY: El nombre de la calle donde ocurrió el incidente.
15. ZIP_CODE: El código postal donde ocurrió el incidente.

16. BOROUGH_DESC: El distrito donde ocurrió el incidente.
17. FLOOR: El piso del edificio donde ocurrió el incidente.
18. CO_DETECTOR_PRESENT_DESC: Indicador de cuándo estaba presente un detector de CO^2
19. FIRE_ORIGIN_BELOW_GRADE_FLAG: Indicador de cuando el incendio se originó por debajo del grado
20. STORY_FIRE_ORIGIN_COUNT: Historia en la que se originó el fuego.
21. FIRE_SPREAD_DESC: Que lejos se extendió el fuego del objeto de origen
22. DETECTOR_PRESENCE_DESC: Indicador de cuándo estaba presente un detector
23. AES_PRESENCE_DESC: Indicador de cuando un sistema de extinción automática está presente
24. STANDPIPE_SYS_PRESENT_FLAG: Indicador de cuando una tubería vertical estaba presente en el área de origen del incendio

El segundo dataset, *"FIRE_BOX",* tiene 16.286 filas y 4 columnas:

1. Long: Punto longitudinal donde esta ubicada la caja de alarma
2. Lat: Punto latitudinal donde esta ubicada la caja de alarma
3. FIRE_BOX: Identificador del área de la caja de alarma contra incendios en la que ocurrió el incidente, único por municipio
4. Adress: Dirección del Fire Box

El merge de ambos se ha realizado por la columna FIRE_BOX, con el inconveniente de que al no tener en el segundo dataset todas las cajas de alarmas que hay, se ha perdido gran cantidad de los datos.

2. TFM_Clean_Nulls

En este paso lo que se ha realizado ha sido la limpieza de los valores NaN, se han eliminado columnas que no se iban a utilizar. Además, muchas de ellas estaban únicamente con valores nulos. En el notebook esta explicado, paso por paso, lo que he ido realizando.

3. TFM_Data_Analysis

En este notebook lo que he realizado un análisis de las diferentes columnas, comparando unas con otras para sacar algunas conclusiones.

Lo primero que he hecho ha sido un *count*, para ver cual es el incidente al que más y menos acuden los bomberos. La tabla resultante es esta:

	INCIDENT_TYPE	count
0	Rescue, EMS incident, other	767993
1	Smoke scare, odor of smoke	132131
2	Removal of victim(s) from stalled elevator	111622
3	Malicious, mischievous false call, other	105360
4	Water or steam leak	103622

La mayoría de los incidentes, un 37%, son rescates o emergencias sanitarias. Y por el que menos llamadas se reciben es por autocarava o vehículos recrativo.

Para ver las actuaciones por distrito he realizado otro *count*. Aquí se puede ver como Brooklyn es el que encabeza la lista. Esto puede ser debido a que, aunque no es el más extenso en cuanto a superficie si es donde se concentra la mayor población por metro cuadrado en Nueva York y con ello donde hay el mayor número de viviendas.

	BOROUGH_DESC	count
0	BROOKLYN	628774
1	MANHATTAN	532546
2	BRONX	417300
3	QUEENS	416502
4	STATEN ISLAND	80713

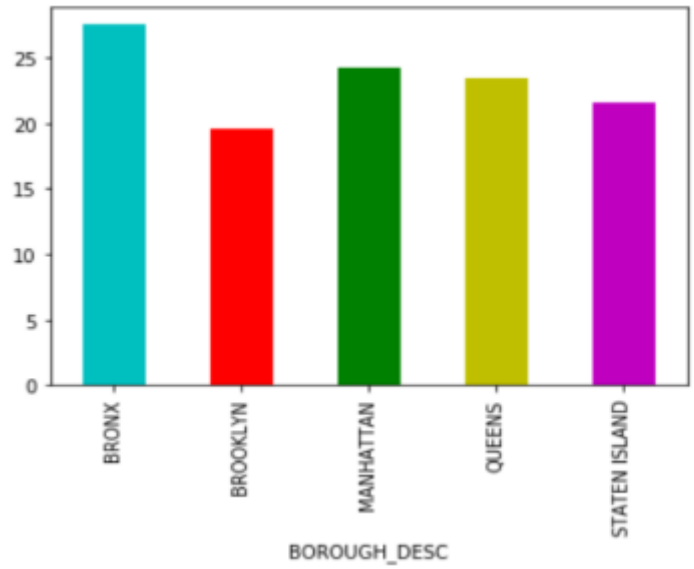
Para ver la evolución de las intervenciones entre los años 2013 y 2018 en cada distrito, he creado una variable para cada año a través de una selección en el índice. Después he hecho un *groupby* y por último un *count.*

Durante este periodo de tiempo el distrito en el que más intervenciones se producen es Brooklyn y en el que menos Staten Island.

Por otro lado, en el año 2018 solo tenemos datos hasta julio, esto puede deberse al *merge* de los dos dataset dónde se perdieron bastantes datos.

En los años 2014, 2015, 2017 y 2018 se ven valles en las gráficas, esto seguramente se deba a que, en los meses que aparecen la bajada, no hay suficientes datos. Puede que también se perdieran en el primer notebook.

El último análisis que he realizado es el tiempo medio que tardan en resolver un incidente. En este caso un *groupby* y *mean* de las columnas BOROUGH_DESC y TOTAL_INCIDENT_DURATION



El distrito donde antes se resuelve un incidente es en Brooklyn con una media de 20 minutos, y en el que más se tarda es en Bronx con una media de alrededor de 30 minutos.

4. TFM_ML

En este notebook, lo primero que he hecho ha sido un *groupby*, *count* y *mean* de las distintas variables con mi variable objetivo.

He realizado una matriz de correlación para poder ver si hay relación entre las variables y el resultado es este:

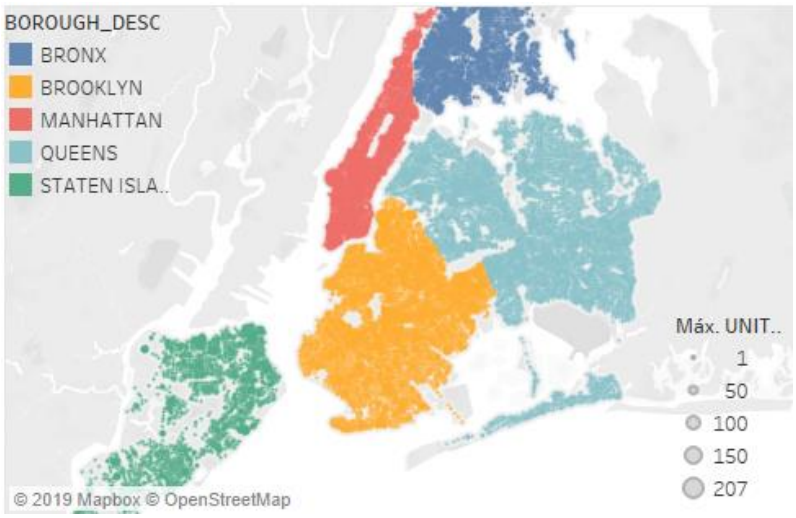
Se puede observar que no hay relación entre variables.

He transformado las columnas CODE_INCIDENT y CODE_PROPERTY de variables categóricas a lineales para poder usar en regresión. Esto lo he hecho con *get_dummies.*

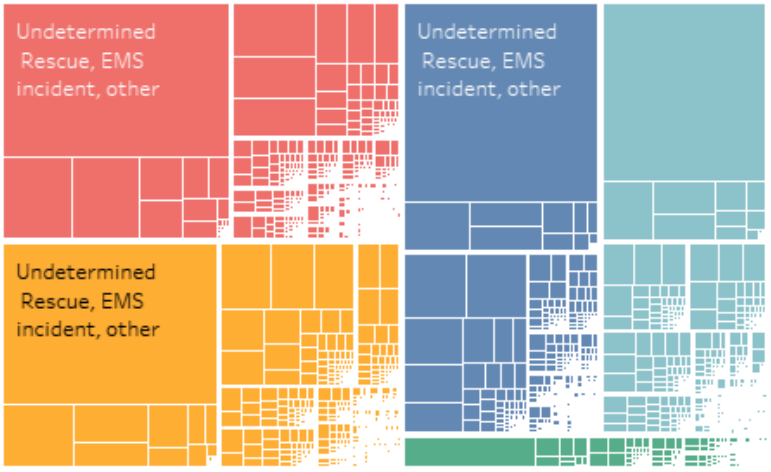
Tras esto he querido hacer un modelo de Regresión lineal, uno de KNN y un Random Forest, no he sido capaz de que funcionase ninguno. He probado de varias formas y ninguna ha dado resultado.

5. GRÁFICOS TABLEAU

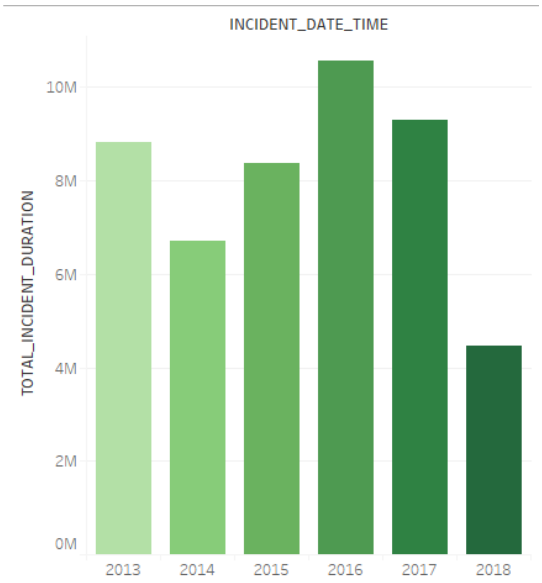
El primero de los gráficos es un mapa, donde se puede ver los incidentes que han ocurrido, los puntos, en cada distrito representado por los colores y las unidades que se han necesitado en casa caso, que esto es el tamaño de los puntos.



En el segundo gráfico es una descripción visual del total de incidentes por tipo de edificio y tipo de incidente.



Este gráfico representa la duración total de los incidentes en cada año. Cada barra representa un año y la altura de estas es la duración total en minutos.



Este gráfico es para ver las unidades que acuden por tipo de incidente. Es una variación del mapa que vimos al principio. Cada punto es un tipo de incidente, el color son los distritos y el tamaño son las unidades que acuden.

Por distritos:

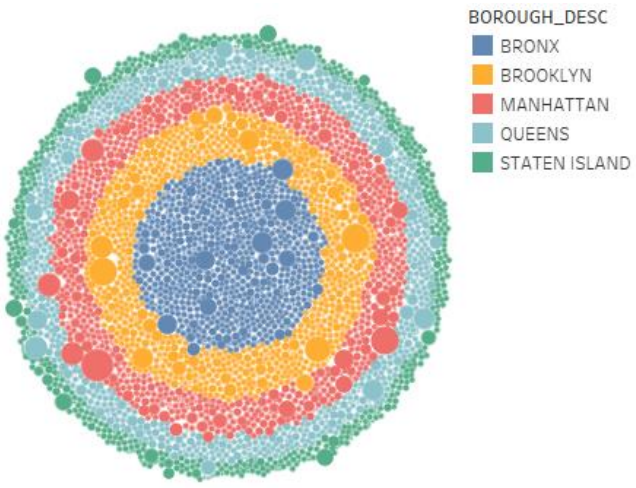
Manhattan: 207 unidades en el año 2015 por un incendio en un edificio

Brooklyn: año 2015, 168 unidades por un fuego

Bronx: 2018 y 2015, en ambos años 86 unidades por un fuego en un edificio

Queens: 90 unidades en los años 2016, 2017 y 2018 por fuegos

Staten Island: 56 unidades en el año 2018 por un fuego



De aquí se puede deducir que como era de esperar los incendios son lo que más unidades requieren.

Por último, esta es una versión del gráfico de barras donde el color de cada burbuja representa el distrito y la altura a la que están cuál fue la duración máxima.