lunes, 24 de octubre de 2022

MEMORIA EDA, Rosario Montalbán

EDA TERRORISMO

Para la realización del EDA he usado las siguientes herramientas:

- Visual Studio Code
- kaggle.com
- Excel
- Tableau

Los pasos a realizar fueron:

- Descarga de dataset desde la web de kaggle (API: kaggle kernels output ritwikdalmia/eda-terrorism-analysis -p /path/to/dest)
- Limpieza del dataset (desarrollado a continuación)
- Comprobación de relaciones entre los valores del dataset para posteriormente sacar conclusiones
- Comprobación de valores y consultas al dataset desde Visual Studio Code con código de programación Python
- Elaboración de gráficos básicos
- Creación de gráficos más complejos en Tableau para la creación de dashboards ordenados con los que contar una historia y sacar conclusiones

Para la limpieza del dataset (aproximadamente 2-3 días completos):

- El dataset inicial era un archivo de extension .csv con más de 181.000 entradas y 135 columnas, las cuales hubo que revisar para cambiar a nombres que se entendieran mejor aquellas que resultaban confusas y descartar las que no fueran relevantes para el desarrollo del EDA.
- Me encontré entonces con un dataset que contenía información bastante incompleta que me entorpecía el tratamiento de los datos. Así que el siguiente paso fue eliminar las columnas con valores Nan o que contenían grandes cantidades de este tipo de datos, pues de nuevo me “ensuciaban” las estadísticas y me llevarían a conclusiones erróneas.
- Traté de rellenar valores null en aquellos huecos donde no existía registro o se desconocía para las columnas que sí que necesitaba utilizar.
- Al ser un dataset tan amplio a nivel temporal (1970-2017) y espacial (muchos rincones del planeta), hubo datos que, bien por antiguos o bien por desinformación no estaban completos, por lo que se han tenido en cuenta a la hora de tratar la información. Hablo de fechas inexactas, ubicaciones poco o nada precisas, desconocimiento de los atacantes o de las características de las víctimas.
- Para el tratamiento de datos y consultas he utilizado principalmente las librerías Pandas, Numpy, Matplotlib, Seaborn y Plotly.

Para la representación gráfica y la presentación al publico usaré Tableau (dinámica) y Canva (estática y conclusiones).

Problemas principales en el proceso:

- El dataset tenia demasiados datos y costaba hacer comprobaciones de cada cambio para comprobar que no afectara a otros datos.
- Cuando pensaba que me había deshecho de todos los duplicados (más de 2.000), me di cuenta de que con todos los campos rellenos de los mismos datos, había más de 4.000 valores que variaban ligeramente en las cifras de muertes y heridos, lo que podía hacer contar los datos casi doble (especialmente importante para casos con gran cantidad de fallecidos como el 11-S) y dar mucho peso en las estadísticas. Este fue el mayor de los problemas porque ya había avanzado en consultas y gráficos (que finalmente descarté por razones obvias).
- Otro gran problema que me encontré que no estaba relacionado con los datos sino con el planteamiento del análisis en sí mismo fue la definición del objetivo e hipótesis del EDA. Me encontré con muchísima información que tratar (o ya tratada) y ningún hilo argumentativo que las guiase para poder sacar conclusiones claras. A pesar de esto, he logrado encaminar mi trabajo y estoy contenta con el resultado.
- Quizá lo que peor llevé fue la falta de tiempo y de ideas. Lo tendré en cuenta para futuros proyectos.

Ventajas con las que conté:

- Encontré un dataset muy completo (tras descartar la primera opción) y pude elegir el enfoque del análisis.
- Puedo darle la profundidad que quiera al análisis e incluso utilizarlo en el futuro para otros trabajos con la facilidad de conocer los datos con los que trabajo y tener objetivos más definidos.
- Me era un tema muy familiar y he podido ver la información más allá de los datos.

Lo que he aprendido:

- He ganado soltura con manejo de gran cantidad de datos.
- He ganado soltura con las bibliotecas más utilizadas.
- A usar herramientas un poco desconocidas como Tableau.

Lo que tengo que reforzar:

- La organización de un proyecto desde el planteamiento de objetivo e hipótesis y el manejo del tiempo y la frustración (que no ha sido poca :)).
- Conocimientos teóricos y prácticos de estadísticos para aplicar a mis datos y sacar información de gran valor.
- Uso de librerías estadísticas y gráficas para representar la información.