Selección del conjunto de datos (Práctica)

Alberto Giménez Aragón - Máster en Ciencia de Datos (UOC)

Dataset: https://opendata-ajuntament.barcelona.cat/data/es/dataset/accidents-vehicles-gu-bcn/resource/f773b7eb-0e2f-467a-86c8-b90e9fd7d29a

1. **Justificad brevemente vuestra selección, ya sea por motivos personales o profesionales.**
   El dataset escogido recoge datos de los accidentes de tráfico gestionados por la Guardia Urbana de Barcelona durante el año 2019. He escogido este dataset porque cada día cruzo Barcelona en coche para ir a trabajar, y muchos de los días se producen largas caravanas debido a accidentes, por lo que siento bastante interés en conocer más detalles sobre estos accidentes, como por ejemplo las zonas y horarios en donde la probabilidad de producirse un accidente es superior.

2. **La relevancia del conjunto de datos en su contexto. ¿Son datos actuales? ¿Tratan un tema importante por algún colectivo concreto? ¿Se ha tenido en cuenta la perspectiva de género?**
   Los accidentes de tráfico es un tema muy importante en la actualidad, ya que ponen en grave peligro la seguridad y vida de las personas, tanto las que van en el vehículo como los peatones. Por tanto, con el fin de reducir los accidentes lo máximo posible, es fundamental detectar sus causas más comunes y características. Pese a que los datos más recientes disponibles son del año 2020, he escogido los de 2019 debido al confinamiento del 2020 por el COVID-19, que causó una gran reducción de la movilidad. La perspectiva de género no se tiene en cuenta, ya que en los datos no se especifica el género de la persona causante del accidente, por lo que no se puede discriminar a ningún género.

3. **La complejidad (medida, variables disponibles, tipos de datos, etc.). ¿Tiene del orden de centenares o miles de registros? ¿Tiene del orden de decenas de variables? ¿Combina datos categóricos y cuantitativos? ¿Incluye otros tipos de datos? Evitad los conjuntos excesivamente simples.**
   El dataset tiene está compuesto por 19037 filas con 28 variables, las cuales combinan diferentes tipos de datos, como cuantitativos, cualitativos o de coordenadas geoespaciales. A continuación, se especifican estas columnas con sus tipos.

Variable	Tipo
Numero_expedient	String
Codi_districte	Integer
Nom_districte	String
Codi_barri	Integer
Nom_barri	String
Codi_carrer	Integer
Nom_carrer	String
Num_postal	String
Descripcio_dia_setmana	String
Dia_setmana	String
Descripcio_tipus_dia	String
NK_Any	Integer
Mes_any	Integer
Nom_mes	String
Dia_mes	Integer
Hora_dia	Integer
Descripcio_torn	String
Descripcio_causa_vianant	String
Descripcio_tipus_vehicle	String
Descripcio_model	String
Descripcio_marca	String
Descripcio_color	String
Descripcio_carnet	String
Antiguitat_carnet	Integer
Coordenada_UTM_X	Geospatial (Float)
Coordenada_UTM_Y	Geospatial (Float)
Longitud	Geospatial (Float)
Latitud	Geospatial (Float)

4. **La originalidad. No repetid los conjuntos de datos clásicos. Podéis, por ejemplo, combinar o mejorar visualizaciones existentes. ¿Hay otras visualizaciones basadas en este conjunto de datos? ¿Es una evolución o actualización de un conjunto anterior? ¿Habéis enriquecido un conjunto de datos ya existente? (25%)**

   Visualizaciones sobre el conjunto de datos exacto no he encontrado, pero sí de datasets muy similares que no contemplan información del vehículo ni de la persona conductora. Por ejemplo, esta visualización muestra los accidentes en un mapa entre 2010 y 2018, así como una distribución de los accidentes por barrios, días, horas y causas. Esta otra visualización de *treemap* muestra las causas de los accidentes más comunes en las calles de cada barrio, el cual se puede seleccionar mediante un desplegable. Sin embargo, no considero que sea una visualización demasiado informativa ni útil, ya que en la mayoría de casos no hay ninguna causa explícita del accidente.
   El resto de visualizaciones que he encontrado, solo muestran las coordenadas de los accidentes sobre el mapa, como por ejemplo esta de 2014 que cambia el color de los puntos según el distrito o esta que muestra de forma dinámica los accidentes diarios en forma de burbujas sobre el mapa de la ciudad. En resumen, las visualizaciones que he encontrado son bastante simples al solo mostrar muy pocas variables, especialmente las de coordenadas sobre el mapa, sin mostrar valores numéricos que puedan darnos una idea de los comportamientos del resto de variables. La única visualización que me ha parecido más completa es la primera, ya que sí es más rica en cuanto a la información mostrada. No obstante, al disponer nosotros de un conjunto de datos más enriquecido, podemos realizar una visualización más completa y que permita el análisis de todas las variables que puedan ser relevantes y no solo de unas cuantas.

5. **Las cuestiones que responderéis con la visualización de datos, ¿Tienen en cuenta los puntos anteriores? ¿Están bien planteadas? ¿Son adecuadas por el conjunto de datos elegido? (30%)**

   Hay muchas preguntas que se intentarán responder con la visualización que pueden ayudar además a distintas personas y organismos a tomar ciertas decisiones para reducir el número de accidentes. Al tener un conjunto de datos bastante rico, las preguntas son bastante variadas e intentan aprovechar todas las variables relevantes que tiene el dataset, pudiendo extraer todo el conocimiento que hay en los datos. Las preguntas que se intentarán responder son las siguientes.

   - Días de la semana con más accidentes.
   - Horas del día con más accidentes
   - Días del mes con más accidentes
   - Culpabilidad del peatón en caso de que haya alguno implicado
   - Distritos con más accidentes
   - Causas más comunes de los accidentes
   - Tipo del vehículo involucrado (Coche, moto, camión, etc.)
   - Marcas de coche con más accidentes
   - Color de vehículos con más accidentes
   - Antigüedad de carnet de los conductores involucrados
   - Representación de los accidentes sobre un mapa