**SPRINT 10 -** PROYECTO FINAL

# ANÁLISIS DE LA DISTRIBUCIÓN DE ACCIDENTES DE TRÁNSITO EN BARCELONA

Laura Cuscurita Martínez
IT Academy
2024

Laura Cuscurita

## Introducción

Cada año, los incidentes en las carreteras urbanas generan pérdidas significativas, tanto en términos de vidas humanas como en daños materiales y costos sociales. Sin embargo, entender las causas y patrones de estos accidentes no es sencillo. Existen muchos factores en juego, desde el volumen de tráfico hasta el comportamiento de los conductores y las condiciones del entorno.

En este estudio, me centraré en explorar si los accidentes de tráfico en Barcelona están relacionados con ciertos patrones temporales, en particular con el horario del día y el día de la semana. Partimos de una idea común: las horas punta, caracterizadas por la congestión en las carreteras debido a los desplazamientos laborales, pueden constituir momentos de mayor riesgo.

Además de investigar la influencia de las horas punta, también analizo si existen variaciones a lo largo del mes y del año, y si ciertos días de la semana registran más accidentes que otros. Este tipo de análisis puede aportar datos valiosos para la planificación urbana y la implementación de políticas de seguridad vial, permitiendo a las autoridades locales concentrar sus esfuerzos de prevención en los momentos y lugares más críticos.

Este proyecto busca, entonces, no sólo confirmar o refutar esta hipótesis de la "hora punta", sino también descubrir otros patrones que puedan ayudar a las autoridades a entender mejor el comportamiento de los accidentes de tráfico y a tomar decisiones informadas para reducir su incidencia.

## Objetivo

El objetivo de este estudio es explorar patrones en los accidentes de tráfico y evaluar la hipótesis de que los accidentes son más frecuentes durante las horas punta en comparación con las horas no punta. Además, se pretende identificar posibles picos de accidentes en determinados meses o días de la semana, lo cual podría estar relacionado con factores estacionales o de comportamiento.

## Hipótesis

La hipótesis principal de este estudio es la siguiente:

**"**Los accidentes de tráfico son más frecuentes en horas punta que en horas no punta."

Además de esta hipótesis, el análisis exploratorio me ha permitido identificar tendencias mensuales que no me había planteado.

Laura Cuscurita

## Metodología

### 1. Obtención de Datos

Los datos utilizados en este estudio provienen de OpenData BCN. El dataset contiene información detallada sobre cada accidente, incluyendo la fecha, hora, ubicación, tipo de accidente, y otras variables relacionadas con el contexto de cada incidente. A continuación se detalla cada variable y su respectiva descripción:

| Camp | Descripción |
|---|---|
| **01.Numero_expedient** | Número identificatiu de l'expedient |
| **02.Codi_districte** | Codi del districte |
| **03.Nom_districte** | Nom del districte |
| **04.Codi_barri** | Codi del barri |
| **05.Nom_barri** | Nom del barri |
| **06.Codi_carrer** | Codi del carrer |
| **07.Nom_carrer** | Nom del carrer |
| **08.Num_postal** | Número postal |
| **09.Descripcio_dia_setmana** | Nom del dia de la setmana |
| **10.Dia_setmana** | Diminutiu del dia de la setmana (Fins l'any 2020) |
| **11.Descripcio_tipus_dia** | Tipus de dia (Fins l'any 2020) |
| **12.NK_Any** | Any |
| **13.Mes_any** | Mes de l'any |
| **14.Nom_mes** | Nom del mes |
| **15.Dia_mes** | Dia del mes |
| **16.Hora_dia** | Hora del dia |
| **17.Descripcio_torn** | Tipus del torn (matí, tarda, nit) |
| **24.Coordenada_UTM_X** | Coordenada X en format UTM |
| **25.Coordenada_UTM_Y** | Coordenada Y en format UTM |
| **24.Coordenada_UTM_X_ED50** | Coordenada X en format UTM (ED50) |
| **25.Coordenada_UTM_Y_ED50** | Coordenada Y en format UTM (ED50) |
| **26.Longitud_WGS84** | Longitud |
| **27.Latitud_WGS84** | Latitud |

## 2. Procesamiento de Datos

- **Valores nulls**: Utilizando un “ accidents.isnull().sum() “ se observó que había ciertos valores faltantes en algunas columnas del dataframe. En particular, algunas entradas para estos campos siguientes:
    - ‘Coordenada_UTM_X_ED50'
    - 'Num_postal_caption’
    - 'Longitud_WGS84’
    - 'Latitud_WGS84’
    - ‘Coordenada_UTM_Y_ED50’

Se observó que el motivo por el cual habia tantos nulls era debido a comoe estaban nombradas las columnas en los diferentes dataframes. Por ejemplo en el df2021 se llamaba a la columna de coordenada como ‘ Coordenada_UTM_X_ED50’, mientras que en el dataframe df2022 se la llamaba ‘ Coordenada_UTM_X’.

Para evitar los nulls se decidió renombrar todas las columnas que podian variar su titulo para poder homogeneizar las columnas.

- **Tipos de datos**: Se analizaron los diferentes tipos de variables y se optó por convertir algunos valores a int o a float, para evitar posibles errores de codigo futuros. l
- **Columnas y dataframes :** se optó por suprimir columnas de poca relevancia de la base de datos, como por ejemplo la columna 'Descripcio_torn' y 'Dia_setmana'. Además, para una mejor facilidad a la hora de tratar los datos relevantes, se creó un df filtrado únicamente con las variables que se querían analizar en este estudio.

## 3. Análisis Exploratorio de Datos (EDA)

Se llevaron a cabo distintos análisis para comprender la distribución y los patrones de accidentes, incluyendo:

- **Conteo de accidentes por años, por hora del día, por tipo de accidente, por distrito y por semana:** Para identificar posibles patrones temporales.
- **Conteo de accidentes por día de la semana**: Para ver cómo varía el número de accidentes según el día.
- **Distribución mensual de accidentes**: Para identificar picos de accidentes a lo largo del año, se realizó un análisis de tendencia por mes.
- **Distribución horaria de accidentes según hora punta vs. no punta**: Mediante gráficos de cajas y conteo, se miró en profundidad si los accidentes están más concentrados en horas punta.

**4. Visualización**

Se utilizaron diversas visualizaciones :

1. Gráfico de líneas
2. Gráfico de barras
3. Matriz de correlación
4. Gráfico de caja
5. Gráfico de barras apiladas

**Resultados**

- **Tendencia mensual**: Se observó que existen picos importantes en los meses clave : febrero, junio y septiembre. Esto posiblemente se debe deber a :
    - La vuelta al cole y la vuelta al trabajo
    - El inicio de las vacaciones : implica aumento de tráfico de visitantes y tráfico de coches que se van de la ciudad

  Por otro lado se observa una caída sustancial en abril y en agosto. Estas caídas se deben a dos motivos principales : la disminución de residentes de barcelona por vacaciones de verano y de primavera. Se puede concluir que una gran parte de los residentes de la ciudad utiliza su vehículo para viajar durante las vacaciones.

  Cabe destacar como se observa el impacto de la pandemia del covid en el gráfico, reduciendo en picado el nº de accidentes durante los meses de confinamiento. Confinamiento : 15 marzo de 2020 – 21 junio de 2020.

  En conclusión, estas tendencias sugieren una posible relación con factores estacionales o eventos específicos en esos meses clave.

- **Accidentes en horas punta vs. no punta**: El análisis reveló que la hipótesis inicial no fue respaldada por los datos, ya que la cantidad de accidentes en horas punta no es significativamente mayor que en horas no punta. Para ello, se aplicó el test de normalidad de Anderson-Darling y la prueba U de Mann-Whitney. No obstante, sí se observó una mayor concentración de accidentes en ciertas horas del día en las horas punta.
- **Día de la semana**: Se encontró que ciertos días tienen una frecuencia de accidentes más alta, aunque este resultado puede variar en función de otros factores como el tipo de día (laboral o festivo). Se observa que a medida que transcurre la semana, los accidentes se incrementan, llegando el pico máximo los viernes. No obstante, durante el fin de semana caen con bastante diferencia, por lo que podemos plantear una posible relación entre día laboral y accidentes.

- **Distritos** : En este análisis, el distrito del Eixample destaca como la zona con mayor incidencia de accidentes de tráfico. A continuación, se analiza en más profundidad las calles destacadas donde ocurren más incidentes [1]de tránsito:
    1. Carrer d'Aragó

       El Carrer d'Aragó es una de las principales arterias de Barcelona y atraviesa una gran parte de la ciudad, conectando áreas clave de transporte y comercio. Constituye una vía de gran afluencia, no solo para el tráfico local sino también para aquellos que transitan desde o hacia otras ciudades de Cataluña. Además, la falta de cruces peatonales suficientemente visibles en ciertos tramos y la velocidad a la que circulan los vehículos contribuyen a una mayor incidencia de accidentes.

       Posibles motivos del orígen de los accidentes:

       -Alta densidad de tráfico en todas las franjas horarias, especialmente en las horas punta.

       -Acceso limitado para peatones en varios tramos, lo que a menudo puede resultar en atropellos o colisiones laterales[2].

       -Velocidad de circulación elevada debido a la longitud y el diseño recto de la vía, que invita a muchos conductores a sobrepasar los límites de velocidad.

    2. Passeig de Gràcia

       El Passeig de Gràcia es otra arteria vital en Barcelona, que cuenta con un alto flujo de tráfico en ambos sentidos y es conocida por su importancia comercial y turística. Como centro de actividad, alberga tanto tráfico vehicular como un número altamente significativo de peatones. El diseño de doble sentido de esta vía aumenta la complejidad del flujo de tránsito, lo que puede contribuir a un incremento en las colisiones y accidentes de tráfico.Por último, la cercanía a tiendas, restaurantes y puntos de interés turístico constituye un aumento en la densidad peatonal.

       Posibles motivos del orígen de los accidentes:

       -Alta densidad peatonal debido al atractivo turístico de la zona.

       - Conducción en ambos sentidos, lo que aumenta la probabilidad de colisiones frontales o laterales en intersecciones o giros.

       - Intersecciones complejas con otras calles de alta circulación en el distrito.

[1] Para ver en qué calles hay más incidentes se ha creado un mapa de calor en código html : file:///C:/Users/lcusc/Documents/mapa_accidentes.html
[2] Es la causa principal de accidentes

- **Prueba de la hipótesis:**
    1. Test de normalidad

       Para poder hacer un test estadístico de correlaciones e hipotesis es necesario hacer un test de normalidad para saber si las variables siguen una distribución normal.

       Se realizó un test de Anderson-Darling sobre la variable *hora_dia* ya que la muestra contiene más de 30.000 registros.

       Resultados: estadístico de Anderson-Darling: 195.18402408390466.

       Valores críticos:( 0.576 = 15%),(0.656 = 10%),(0.787 5%),( 0.918 2.5%,),(1.092 1%). Como se observa, el estadístico resultante supera con creces cada valor crítico en diferentes niveles de significancia. Esto indica que la variable *hora_dia* no sigue una distribución normal.

    2. Test estadístico

       Como la variable que quiero estudiar no sigue una distribución normal, he utilizado la prueba estadística de Mann-Whitney U. El objetivo es determinar si existen diferencias significativas entre los accidentes ocurridos en horas punta y horas no punta.

       Hipótesis:

       $H_0$: No hay una diferencia significativa entre horas punta y no punta.

       $H_1$: Hay una diferencia significativa entre horas punta y no punta.

       Resultados del test : estadístico U: 96603391.0, p-value: 4.4363510020455395e-39

       Dado que el p-valor es inferior que el umbral de significancia de 0.05[3], se rechaza la hipótesis nula ($H_0$). Esto implica que hay una diferencia estadísticamente significativa entre el número de accidentes ocurridos en horas punta y horas no punta, lo que sugiere que el momento del día tiene un efecto relevante en la frecuencia de accidentes. Sin embargo, no quiere decir necesariamente que haya más accidentes en total en las horas punta, es decir, la distribución de las horas donde ocurren accidentes son diferentes.

[3] El p-valor es una medida utilizada en pruebas estadísticas para determinar la significancia de los resultados obtenidos. El 5% (0.05) es el umbral más comúnmente utilizado para decidir si un resultado es significativo o no

- **Estadísticas descriptivas**:

| | Tipus_dia | NK_Any | Mes_any | Hora_dia |
|---|---|---|---|---|
| count | 30.297.000.000 | 30.297.000.000 | 30.297.000.000 | 30.297.000.000 |
| mean | 3.770.604 | 2.021.566.525 | 6.591.346 | 13.878.767 |
| std | 1.868.708 | 1.093.801 | 3.487.460 | 5.297.474 |
| min | 1.000.000 | 2.020.000.000 | 1.000.000 | 0 |
| 25% | 2.000.000 | 2.021.000.000 | 3.000.000 | 10.000.000 |
| 50% | 4.000.000 | 2.022.000.000 | 7.000.000 | 14.000.000 |
| 75% | 5.000.000 | 2.023.000.000 | 10.000.000 | 18.000.000 |
| max | 7.000.000 | 2.023.000.000 | 12.000.000 | 23.000.000 |

El valor medio de Mes_any es 6.59, lo que sugiere que los datos se distribuyen principalmente en la mitad del año, con una ligera inclinación hacia los meses más cercanos al verano junio y julio.

El mínimo es 1 (enero) y el máximo es 12 (diciembre), lo que muestra que todos los meses están representados en los datos.

La media de Hora_dia es 13.88, lo que indica que los accidentes tienden a ocurrir durante la tarde, alrededor de las 14:00.

El rango de horas va desde 00:00 hasta las 23:00, con una desviación estándar de 5.30. Esto indica una distribución de accidentes de todo un día que tiene un mayor número de accidentes en las horas de la tarde. La media de accidentes está centrada en 13.88, que corresponde a las 14:00 horas. Esto parece coincidir con mi hipótesis de mayor accidentes en horas punta. Sin embargo, la desviación estándar relativamente alta (5.30) indica que hay una distribución amplia de accidentes a lo largo del día. Esto sugiere que, aunque hay picos durante ciertas horas, los accidentes también ocurren en otras horas fuera de las horas punta, lo que limita la conclusión de que las horas punta sean el único factor.

**Discusión**

Los resultados obtenidos ofrecen varias interpretaciones interesantes:

- **Sobre la hipótesis inicial**: La hipótesis de que los accidentes son más frecuentes en horas punta no fue respaldada por el análisis. Esto sugiere que los habitantes de la ciudad no son sensibles a los picos o atascos de tráfico.
- **Picos mensuales**: Los picos en meses específicos detallados anteriormente plantean preguntas sobre si hay factores estacionales, condiciones climáticas, o eventos locales que contribuyen a estos aumentos en los accidentes. Estos picos podrían investigarse más a fondo con datos adicionales, como las condiciones meteorológicas o la densidad de tráfico.

**Conclusiones**

En resumen, el análisis de los datos de accidentes de tráfico reveló patrones relevantes:

1. La distribución de accidentes si varía significativamente entre horas punta y no punta, pero no demuestra que haya más accidentes en horas punta.
2. Se identificaron tres picos anuales en los meses de febrero, junio y septiembre, lo cual podría merecer mayor investigación.
3. Los días de la semana también presentan diferencias en el número de accidentes, lo que sugiere una tendencia por parte de los habitantes de Barcelona, a tener más accidentes en los últimos días laborales. Esto se puede deber a factores psicológicos, físicos de las personas, o también a factores de movilidad, sugiriendo que más gente se desplaza a finales de los días laborales.

**Recomendaciones**

1. **Investigación Adicional**: Incluir datos sobre condiciones meteorológicas, entrada de vehículos a la ciudad y otros factores que podrían explicar los picos estacionales.
2. **Prevención Integral**: Dirigir campañas de concienciación a una variedad de horas y días, en lugar de enfocarse solo en las horas punta.
3. **Monitoreo Mensual**: Observar y analizar los accidentes en los meses con mayor frecuencia de incidentes para hacer un análisis más riguroso.

**Bibliografía**

-Open Data BCN | Servicio de datos abiertos del Ajuntament de Barcelona. (s. f.). https://opendata-ajuntament.barcelona.cat/es/

-Alberto, T. D. (2022, 3 octubre). Análisis de los accidentes de tráfico urbano con ArcGIS: El caso de Barcelona. https://docta.ucm.es/entities/publication/d28c17a2-c514-422a-a411-b344804ef25f/full

-María, O. A. (2021, 1 julio). Factores determinantes en la generación de accidentes de tráfico en la ciudad de Barcelona. https://docta.ucm.es/entities/publication/440f1c89-b301-4c95-bace-5e87c20fb949

-Hakkert, A. S., & Gitelman, V. (2007). *The effects of traffic enforcement on accident rates*. Accident Analysis & Prevention, 39(5), 989-999.

-17 Statistical hypothesis tests in Python (Cheat sheet). (2021, 7 noviembre). Machine Learning Mastery. Recuperado 10 de noviembre de 2024, de https://machinelearningmastery.com/statistical-hypothesis-tests-in-python-cheat-sheet/

-Evergreen, S. D. H. (2017). *Effective Data Visualization: The Right Chart for the Right Data*. SAGE Publications.