Instituto Tecnológico y de Estudios Superiores de Monterrey.

Escuela de Ingenierías

Maestría en Inteligencia Artificial Aplicada – Proyecto Integrador Grupo 10, Equipo 47.

Profesores:

Dra. Grettel Barceló Alonso
Dr. Luis Eduardo Falcón Morales
Mtra. Verónica Sandra Guzmán de Valle
Dr. Gerardo Jesús Camacho González
Dr. Eusebio Vargas Estrada

Propuesta de proyecto y firma de convenios

Equipo #47

Erick Eduardo Betancourt Del Angel	A01795545
Lucero Guadalupe Contreras Hernández	A01794502
Erik Morales Hinojosa	A01795110

Fecha de entrega: 21 de Septiembre del 2025

Repositorio de GitHub:

https://github.com/erikmoralestec/proyecto-integrador-grafos-de-conocimiento-llm

I. Introducción

El análisis exploratorio de datos (EDA) constituye una etapa fundamental dentro del proceso de investigación, ya que permite comprender la estructura, calidad y características de los datos antes de su integración en modelos más complejos como los grafos de conocimiento. En este proyecto, se trabajó con tres conjuntos de datos principales: *Complaints*, *Investigations* y *Recalls*, cada uno con un propósito específico dentro del ciclo de gestión de la seguridad vehicular.

El conjunto *Complaints* recoge las quejas reportadas por los usuarios ante fallas o comportamientos anómalos en los vehículos. *Investigations* contiene información sobre las investigaciones oficiales realizadas por organismos reguladores, mientras que *Recalls* documenta las acciones de retiro o corrección emitidas por los fabricantes. En conjunto, estos tres conjuntos permiten analizar la trazabilidad de un problema desde su detección por el usuario hasta su resolución institucional.

El objetivo del EDA fue identificar patrones, relaciones y características comunes entre las variables de estos conjuntos para establecer una base sólida que permita construir representaciones semánticas consistentes. A través de análisis estadísticos, visualizaciones y procesamiento del lenguaje natural, se buscó detectar tendencias, valores atípicos, distribuciones relevantes y entidades clave que posteriormente podrán transformarse en nodos y relaciones dentro del grafo de conocimiento.

Este análisis constituye, por tanto, una fase preparatoria esencial para la construcción del modelo semántico del proyecto, al ofrecer una comprensión profunda tanto de los datos estructurados (fabricante, modelo, componente, año, etc.) como de la información no estructurada contenida en las descripciones textuales de las quejas e investigaciones.

II. Conjuntos de datos analizados

Para este avance trabajamos con **tres** fuentes públicas de la NHTSA (Office of Defects Investigation, ODI): **Consumer Complaints**, **Investigations** y **Recalls**. A continuación se describe su contenido y alcance.

Consumer Complaints

Conjunto de **quejas de consumidores** sobre fallas vehiculares. Para esta entrega se filtró a **2025**, quedando **59,548** registros y **30** variables tras la limpieza. Incluye campos estructurados (p. ej., **MAKE, MODEL, YEAR, COMPONENT, MILES, VEHSPEED,** indicadores de **CRASH, FIRE, INJURED, DEATHS**) y un campo textual largo (**CDESCR**) con la narrativa del incidente.

Investigations

153,501 registros y **11** columnas con el ciclo de vida de las **investigaciones formales**: identificadores (**ACTION_NUMBER, CAMP_NO**), taxonomía (**MAKE, MODEL, COMPONENT, MFG_NAME, YEAR**), fechas (**ODATE** apertura, **CDATE** cierre) y texto corto/largo (**SUBJECT**, **SUMMARY**). Se derivan **estado** (abierta/cerrada) y **duración** (días).

Recalls

Campañas de **retiro de mercado**. Tras consolidación a nivel campaña, se cuentan **14,290** campañas únicas (el archivo crudo tenía 221,835 filas por combinaciones marca–modelo–año). Variables clave: **CAMPNO** (ID), **MFGNAME, MAKETXT, MODELTXT, COMPNAME**, tipo (**RCLTYPECD**), temporalidad **(RCDATE, ODATE, YEARTXT**) e **impacto** (**POTAFF_num**). Incluye textos de **defecto**, **consecuencia** y **acción correctiva**.

IV. Análisis Exploratorio de Datos (EDA)

En este proyecto, el EDA se llevó a cabo de manera independiente para cada uno de los tres conjuntos de datos seleccionados: Complaints, Investigations y Recalls, siguiendo una metodología uniforme que facilita la comparación y el análisis cruzado entre ellos.

Cada sección aborda aspectos clave del conjunto de datos correspondiente, comenzando con una descripción estructural y una evaluación de los valores faltantes, seguida del análisis univariante, multivariante y de texto. Esta organización permite identificar patrones, tendencias y posibles relaciones entre variables, además de evaluar la consistencia y relevancia de la información contenida en cada dataset. De esta forma, el EDA sienta las bases para etapas posteriores orientadas a la construcción de representaciones semánticas y la integración del conocimiento derivado de los datos.

IV.1 Complaints

4.1.1 Estructura de datos

El conjunto de datos utilizado en este análisis corresponde a reportes de quejas relacionadas con vehículos registrados durante el año 2025. **Originalmente, el dataset contenía 79,376 registros y 49 columnas,** representando tanto datos técnicos del vehículo como información contextual y descripciones textuales de las fallas reportadas.

El análisis de completitud mostró que el **96.4% de los registros presenta al menos un valor faltante**, aunque la mayoría conserva información suficiente para el análisis textual y categórico.Tras un proceso exhaustivo de limpieza y filtrado, se redujo a **59,548 registros y 30 variables**, enfocándose exclusivamente en quejas de vehículos reales, eliminando productos accesorios o casos con información inconsistente.

Se identificaron **columnas con valores faltantes significativos**, particularmente en variables relacionadas con el distribuidor (DEALERCITY, DEALERNAME, DEALERSTATE) y datos del vehículo (VEHSPEED, MILES). Los primeros tres no aportan mucho valor a nuestro análisis, sin embargo se decidió mantenerlas; mientras que las dos últimas es posible encontrar esta información en la descripción del texto.

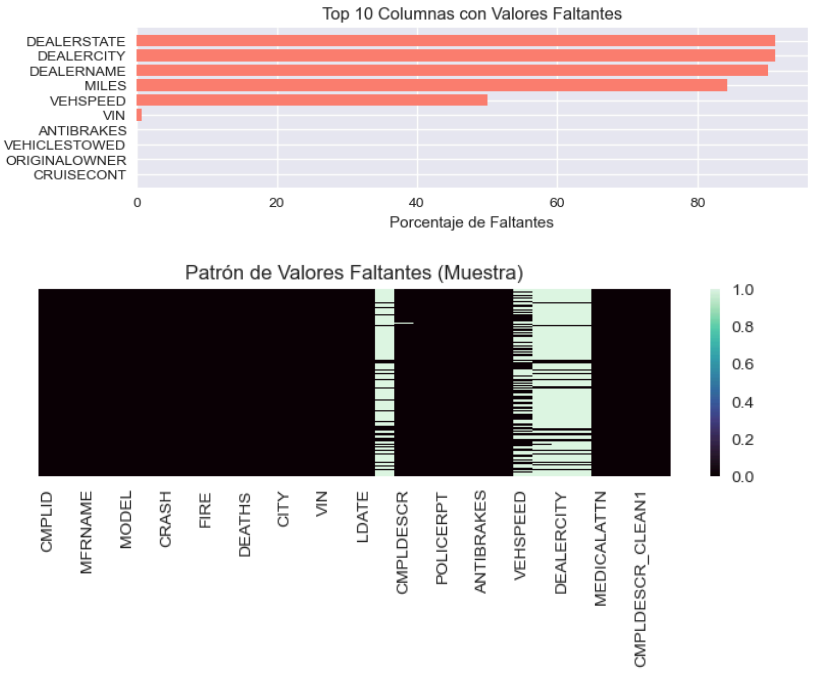


Figura 4.1.1 y 4.1.2 Estas visualizaciones muestran nuestras variables restantes que después de la limpieza han mantenido valores faltantes.

```
Tipos de datos:
object            20
int64              5
datetime64[ns]     3
float64            2
```

Tabla 4.1.1.1 Tipos de datos con los que se cuenta en este dataset

```
Estadísticas Resumidas Numéricas:
                YEAR       INJURED        DEATHS          MILES      VEHSPEED
count  59548.000000  59548.000000  59548.000000    9408.000000  29747.000000
mean    2019.106469      0.030580      0.001226   86651.384673     34.995966
std        4.259369      0.245171      0.110030   59157.986073     32.301967
min     1986.000000      0.000000      0.000000       0.000000      0.000000
25%     2017.000000      0.000000      0.000000   45000.000000      5.000000
50%     2020.000000      0.000000      0.000000   81000.000000     35.000000
75%     2022.000000      0.000000      0.000000  120000.000000     60.000000
max     2026.000000     10.000000     15.000000  763309.000000    999.000000
```

Tabla 4.1.1.2 Resumen estadístico de nuestras variables numéricas

```
=== Resumen de variables categóricas ===
            column   dtype  n_unique  missing  missing_pct
10        CMPLTYPE  object         2        0         0.00
3            CRASH  object         2        0         0.00
4             FIRE  object         2        0         0.00
18     MEDICALATTN  object         2        0         0.00
11       POLICERPT  object         2        0         0.00
13      ANTIBRAKES  object         2       17         0.03
14      CRUISECONT  object         2       17         0.03
12  ORIGINALOWNER  object         2       17         0.03
19  VEHICLESTOWED  object         2       17         0.03
17     DEALERSTATE  object        53    54269        91.13
7            STATE  object        58        0         0.00
0          MFRNAME  object       117        0         0.00
1             MAKE  object       147        0         0.00
5        COMPONENT  object       344        0         0.00
2            MODEL  object      1106        0         0.00
16      DEALERCITY  object      1873    54251        91.10
15      DEALERNAME  object      3909    53620        90.05
6             CITY  object      8160        8         0.01
8              VIN  object     29445      399         0.67
9        CMPLDESCR  object     40815       12         0.02
```

Tabla 4.1.1.3 Resumen estadístico de nuestras variables descriptivas

4.1.2 Análisis univariante

Las principales variables numéricas consideradas fueron YEAR, INJURED, DEATHS, MILES y VEHSPEED.

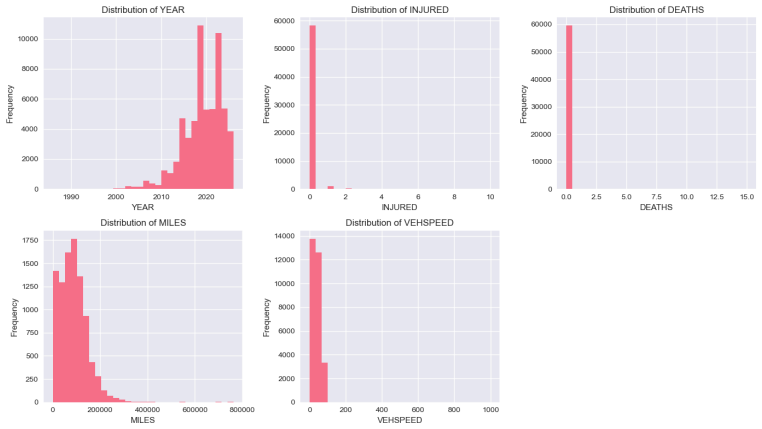


Figura 4.1.2.1 Histogramas de variables numéricas

Los histogramas evidencian distribuciones altamente asimétricas, con concentraciones notables en valores bajos en el caso de las variables INJURED, DEATHS y VEHSPEED, mientras que MILES presenta una distribución sesgada a la derecha (right-skewed), lo que sugiere que la mayoría de los vehículos reportan fallas con un kilometraje relativamente bajo, pero existen algunos casos con kilometrajes extremadamente altos. En el caso de YEAR, la distribución se debe posiblemente a que al ser una base del 2025, los vehículos van a tender a ser recientes.

Los diagramas de caja (boxplots) muestran la presencia de valores atípicos en todas las variables numéricas, aunque en diferentes magnitudes.

- En **MILES**, se observan varios puntos extremos que superan las 700,000 millas recorridas.
- **VEHSPEED** presenta valores superiores a los 600 km/h, posiblemente atribuibles a errores de captura.
- En **INJURED** y **DEATHS**, la mayoría de los registros se concentran en cero, con pocos casos que superan los 5 lesionados o 10 fallecidos por incidente.

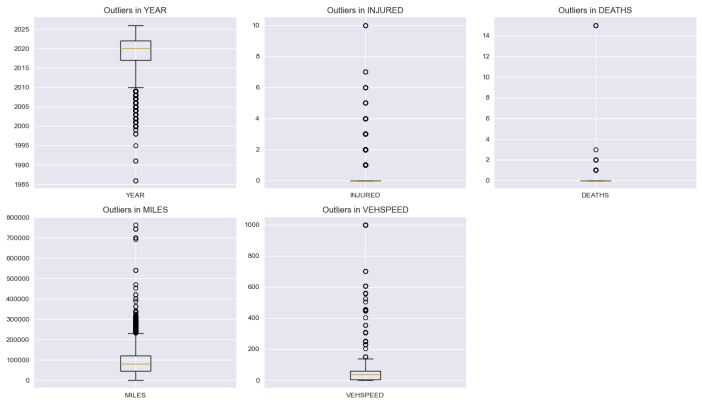


Figura 4.1.2.2 Boxplots de variables num[ericas

En general, estos valores atípicos no afectan la distribución central de los datos, pero representan casos excepcionales que deberán considerarse cuidadosamente durante fases posteriores del modelado o análisis causal.

4.1.3 Análisis bi/multivariante

Los reportes de 2025 muestran una tendencia estable entre enero a julio, con una disminución en agosto y septiembre, debido a que fueron las últimas fechas en las que se actualizaron los datos, por lo tanto no se detecta una tendencia.

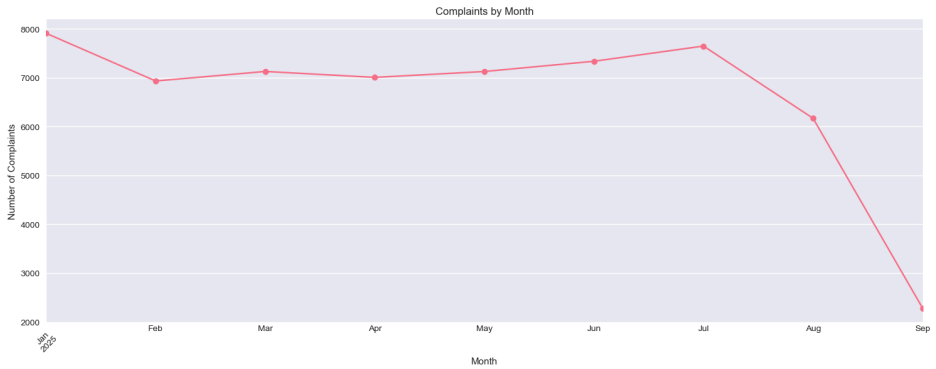


Figura 4.1.3.1 Linea del tiempo de quejas

La matriz de correlación que combina variables numéricas (YEAR, MILES, INJURED, DEATHS, VEHSPEED) y categóricas codificadas (MAKE, COMPONENT, CRASH, FIRE, etc.) muestra los siguientes hallazgos:

- Existe una **correlación positiva moderada (0.32)** entre **INJURED** y **DEATHS**, lo que refleja la coherencia esperada: a mayor número de lesionados, más probable es la presencia de muertes en un incidente.

- **YEAR** muestra una **correlación negativa (-0.59)** con **MILES**, lo que indica que los vehículos más recientes suelen reportar quejas con menor kilometraje.

- Variables categóricas como **CRASH** y **FIRE** presentan correlaciones positivas con **POLICERPT** y **MEDICALATTN**, lo que sugiere que la ocurrencia de un choque o incendio está más vinculada a la necesidad de atención médica o reporte policial.

- Se observa redundancia entre variables binarias relacionadas con seguridad (**ANTIBRAKES, CRUISECONT, ORIGINALOWNER**), lo que debe considerarse al modelar para evitar multicolinealidad.

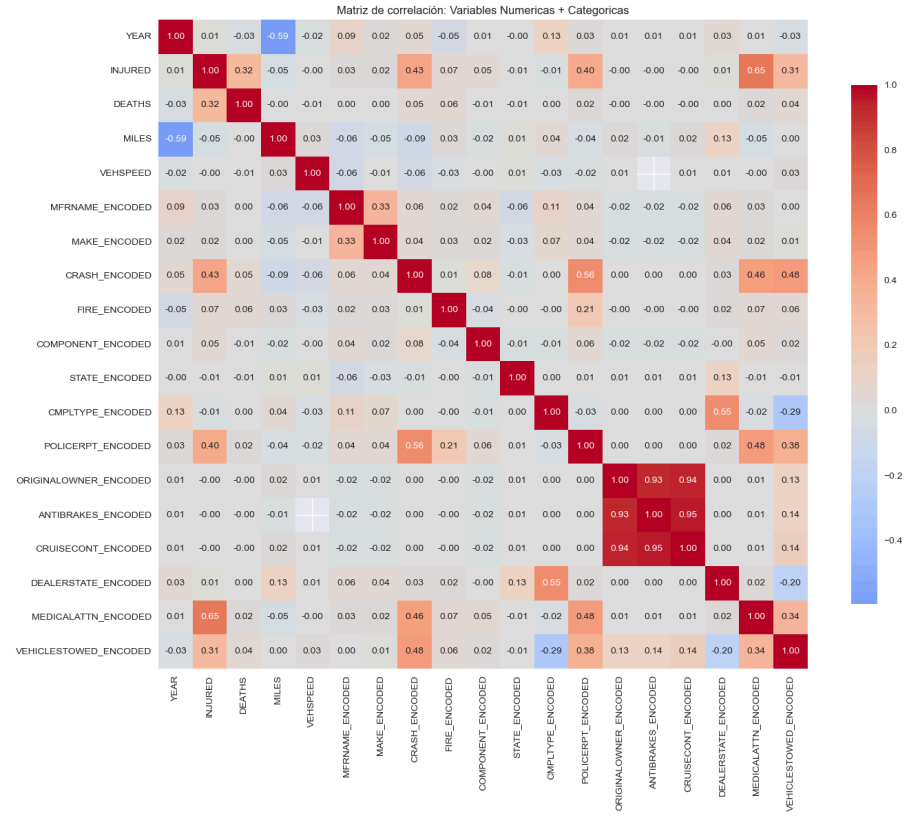


Figura 4.1.3.2 Matriz de correlación que combina variables numéricas

Los boxplots de longitud de texto muestran que:

- Las quejas asociadas a **Hyundai** y **Jeep** tienden a ser más extensas, posiblemente por la descripción detallada de problemas complejos o múltiples fallas reportadas.
- Los componentes **Engine** y **Electrical System** también presentan mayor longitud media en las quejas, lo que sugiere una mayor complejidad en la explicación del fallo.
- No se observan diferencias marcadas por año, aunque las quejas más recientes parecen ligeramente más largas.

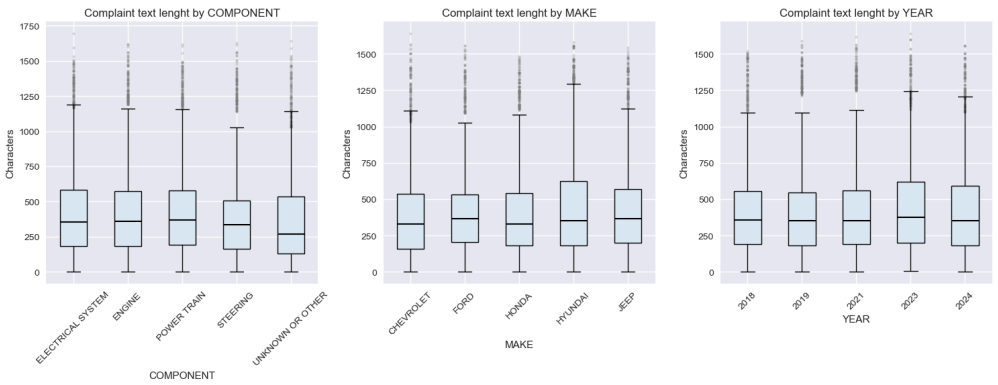


Figura 4.1.3.3 Boxplots de longitud de texto

4.1.4 Análisis textual y extracción de conocimiento

En todos los casos se observa una alta recurrencia de frases como *“contact stated”*, *“vehicle repaired”*, *“failure mileage”* y *“warning light”*, lo que refleja un patrón lingüístico estándar en las quejas formales dirigidas a la NHTSA. Este lenguaje indica que la mayoría de los reportes siguen una estructura descriptiva centrada en el contacto del propietario con el fabricante o distribuidor, y en la documentación de fallas mecánicas o eléctricas recurrentes.

Al segmentar las nubes por **marca**, se evidencian ligeras variaciones semánticas:

- En **Ford**, predominan términos como *“recall repair”* y *“engine light”*, asociados con incidencias de mantenimiento y notificaciones de retiro.
- En **Honda**, se destacan *“vehicle repaired”* y *“failure mileage”*, lo que sugiere una mayor proporción de reportes relacionados con fallas tras cierto uso acumulado.
- En **Chevrolet**, las palabras *“warning light”* y *“contact stated”* se repiten con mayor frecuencia, indicando un patrón similar al de Ford, posiblemente asociado a problemas en sistemas eléctricos o de sensores.

Por **componente**, las nubes muestran que *“engine”*, *“electrical system”* y *“power train”* son los ejes semánticos más recurrentes. Las frases acompañantes como *“light illuminated”*, *“vehicle taken”* o *“manufacturer notified”* sugieren que las quejas tienden a describir síntomas (luces de advertencia, pérdida de potencia, fallas de encendido) más que diagnósticos técnicos detallados. Esto evidencia el carácter subjetivo de las narrativas, pero también su valor como fuente de información contextual para análisis posteriores en el Knowledge Graph.

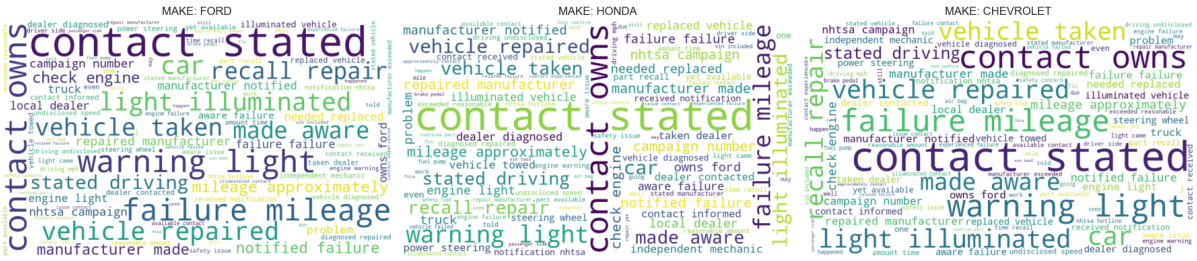


Figura 4.1.4.1 Nube de palabras por fabricante

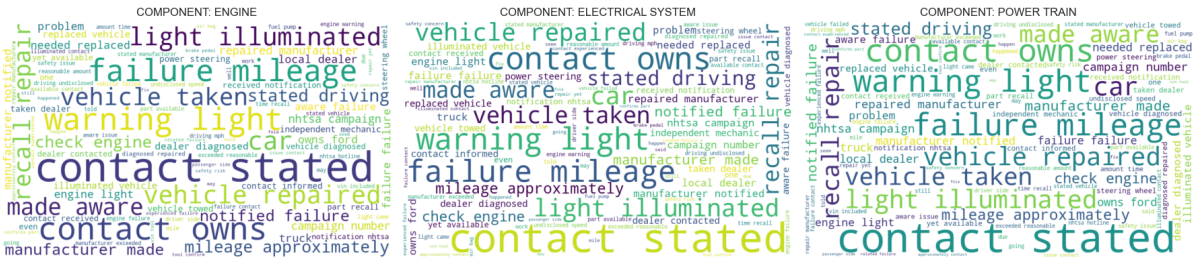


Figura 4.1.4.2 Nube de palabras por componentes

El mapa de calor **Topics × Components** evidencia patrones relevantes:

- **Tópico 1** y **Tópico 7** concentran quejas sobre **sistemas eléctricos y motores**, siendo estas áreas críticas en la percepción negativa del usuario.

- **Componentes como Power Train y Air Bags** aparecen asociados a múltiples tópicos, lo que indica su alta frecuencia en distintos tipos de incidentes.

- La dispersión de los tópicos sugiere que un mismo componente puede estar vinculado con distintos contextos narrativos (ej. fallas mecánicas vs. problemas electrónicos asociados al mismo subsistema).

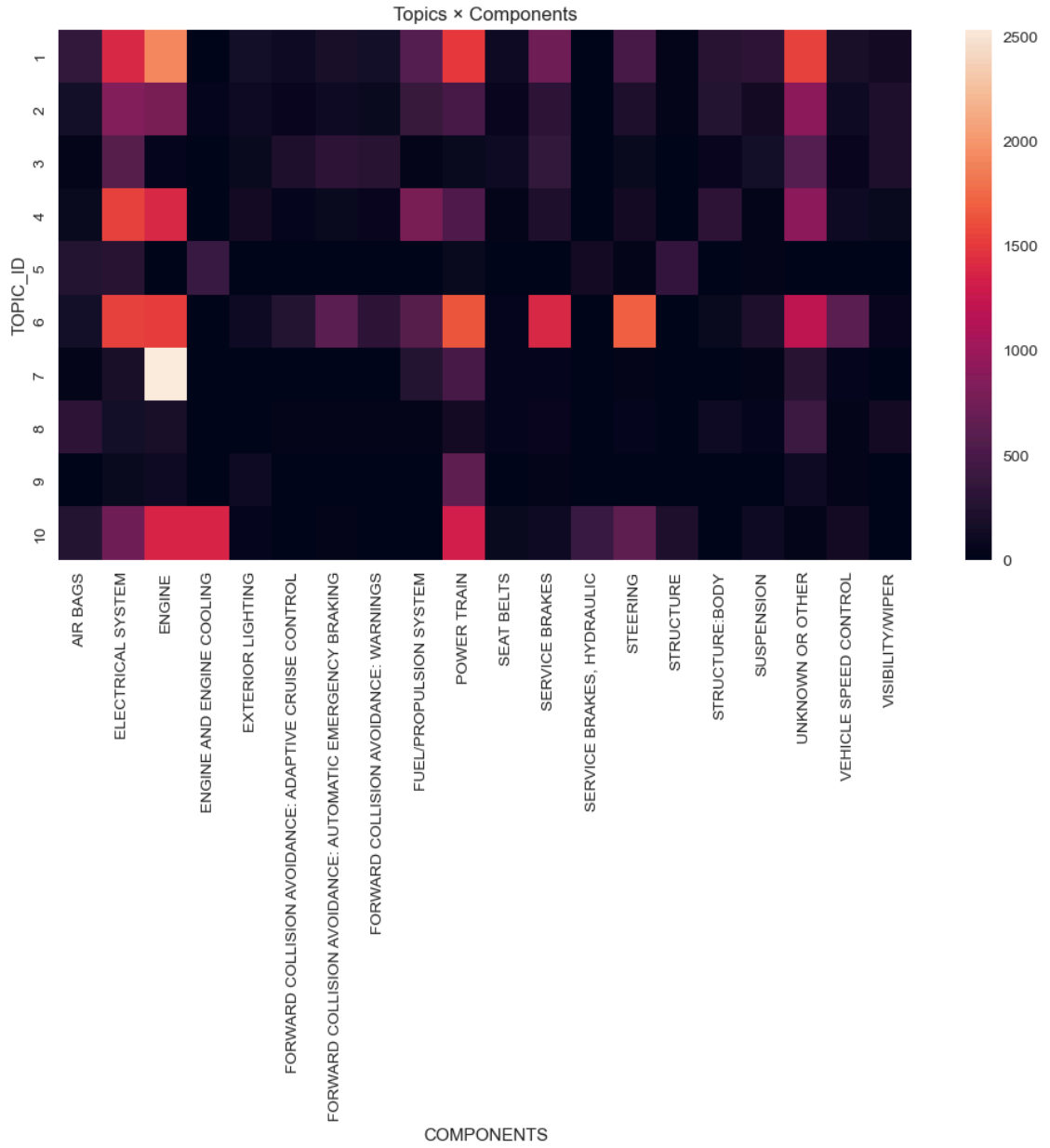


Figura 4.1.4.3 Mapa de calor Topics X Components

El modelado de tópicos se aplicó utilizando el algoritmo **Latent Dirichlet Allocation (LDA)** con el objetivo de identificar patrones semánticos recurrentes en las descripciones de quejas. Esta técnica permite descubrir **grupos latentes de palabras** que suelen aparecer juntas en el texto, representando posibles temas o tipos de fallas. La visualización generada mediante **pyLDAvis** (Figura X) muestra dos componentes principales:

El mapa de distancias intertópicas presentado a la izquierda muestra la distribución de los temas identificados por el modelo LDA, donde cada burbuja representa un tema y su tamaño es proporcional a su prevalencia dentro del corpus. La distancia entre las burbujas refleja la similitud semántica entre los temas: aquellas que se encuentran más próximas indican una relación conceptual más estrecha.

En este caso, se observa una adecuada separación entre los temas, lo que sugiere que el modelo logró capturar distintas dimensiones del contenido de las quejas, como fallas eléctricas, problemas en el motor, frenos o transmisiones. Por otro lado, la gráfica de barras situada a la derecha muestra los términos más relevantes asociados a cada tema, permitiendo interpretar su significado.

En particular, el Tema 1 concentra aproximadamente el 21.4 % de los tokens del corpus y está compuesto por términos como *contact*, *vehicle*, *failure*, *stated*, *dealer*, *repaired* y *manufacturer*. Este patrón sugiere que una gran proporción de las quejas se centra en situaciones donde los consumidores reportan fallas mecánicas o de seguridad tras comunicarse con el fabricante o distribuidor del vehículo, reflejando así el papel central del proceso de atención y reparación dentro de la narrativa de las quejas.

En conjunto, el modelo LDA proporciona una **visión estructurada de los tópicos dominantes** en las quejas, lo que permite identificar **las categorías más recurrentes de problemas reportados por los usuarios**. Este análisis servirá como base para la **posterior construcción del grafo de conocimiento**, donde los temas detectados se podrán vincular con fabricantes, componentes y años de producción. Se puede acceder a esta visualización interactiva desde nuestro repositorio en github: https://github.com/erikmoralestec/proyecto-integrador-grafos-de-conocimiento-llm/blob/main/notebooks/usercomplaints_lda_topics.html

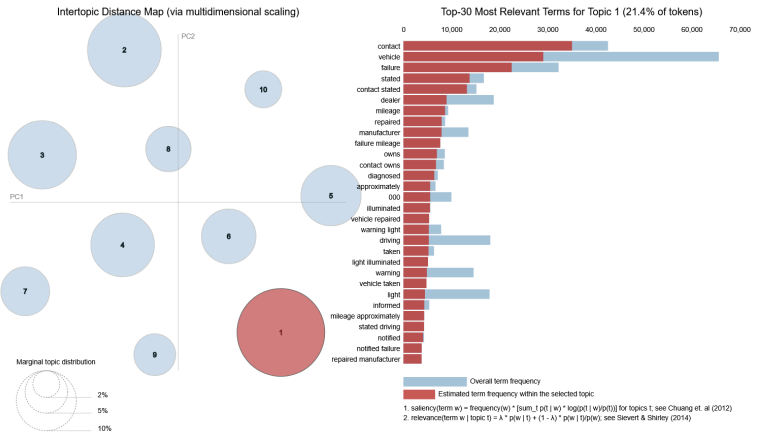


Figura 4.1.4.4 Modelado de tópicos pyLDAvis

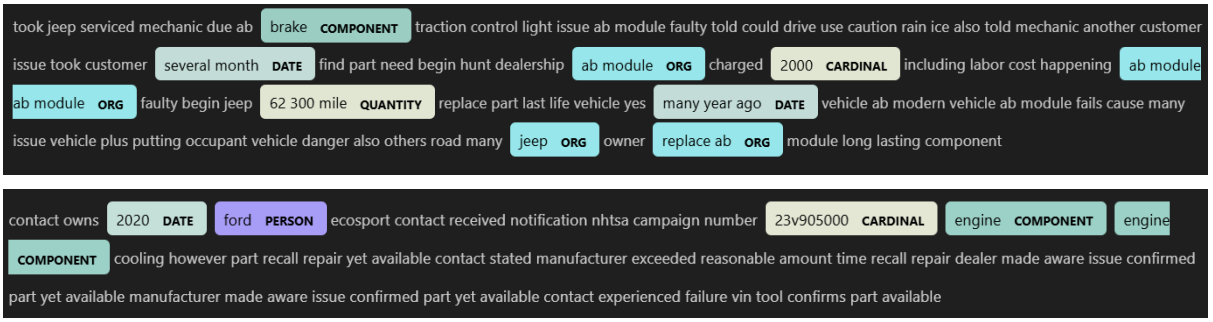
El modelo de reconocimiento de entidades nombradas (NER) permitió identificar categorías relevantes dentro del texto, como **COMPONENT**, **ORG**, **PERSON**, **DATE** y **QUANTITY**.

En general, el etiquetado fue adecuado para entidades como **componentes del vehículo** (*engine*, *brake*, *ab module*), **marcas** (*Ford*, *Jeep*), **fechas**, y **millas recorridas**. Sin embargo, se detectaron ciertos errores semánticos en la clasificación, principalmente en la asignación de etiquetas **WORK_OF_ART** a frases técnicas (*"Check Brake System"*, *"Engine Light"*). Estos casos indican que el modelo base, entrenado en contextos generales, interpreta términos como "light" o "system" fuera de su dominio automotriz.

A pesar de estas inconsistencias, el NER cumple un rol fundamental dentro del proceso analítico, ya que:

1. **Permite estructurar entidades clave** como *marca*, *modelo*, *componente* y *fecha*, que servirán como **nodos** en el futuro Knowledge Graph.
2. **Facilita la identificación de relaciones semánticas** entre conceptos, por ejemplo:
   (FORD) [has_issue_with]→ (ENGINE)
   (ENGINE) [failure_type]→ (COOLING)
   (COMPONENT: BRAKE) [reported_by]→ (OWNER)
3. **Contribuye a la normalización terminológica**, al agrupar menciones equivalentes de un mismo componente (ej. *"brake"*, *"braking system"*, *"service brake hydraulic"*).

En conjunto, las entidades extraídas y los patrones textuales visualizados proporcionan una base sólida para la construcción de un **Knowledge Graph semántico**, donde las relaciones entre fabricantes, componentes, síntomas y acciones reportadas podrán ser modeladas explícitamente.



Esto puede mejorarse utilizando como referencia algunos de los datos estructurados con los que se cuenta en el dataset para mejorar el nombramiento de los objetos, como por ejemplo la velocidad o la marca de los vehículos.

4.1.5 Preprocesamiento de los datos

Al realizar el filtrado del año, se detectaron 9 columnas con 100% de los valores nulos, por lo tanto se descartaron del análisis. Seguido de esto, se eliminaron otras 10 columnas que se eliminaron ya sea debido a que el 99.99% de los datos solamente contaban con 1 valor único, el 99.99% de los datos eran nulos o no aportan nada al análisis:s: **SEAT_TYPE, RESTRAINT_TYPE, MANUF_DATE, FUEL_TYPE, LOC_OF_TIRE, DOT, DRIVE_TRAIN, DEALER_TEL, DEALER_ZIP, PROD_TYPE, OCCURENCES, PURCHDATE, NUMCYLS, FUELSYS, TRANSTYPE, TIRESIZE, TIREFAILTYPE, ORIGINEQUIP** y **REPAIRED**

Durante la limpieza, se identificaron y eliminaron valores reservados o no válidos como:

- **YEAR = 9999**: utilizado como valor por defecto en **registros no vinculados a vehículos**. Se tomó la decisión de descartar estas filas.
- **Campos con** "**UNKNOWN**" o "UNSPECIFIED" en variables como COMPONENT o MAKE, sin embargo, esta información **es posible encontrarla en las descripciones de las quejas.**
- Descripciones genéricas o placeholders en texto ([XXX], "INFORMATION REDACTED PURSUANT TO THE FREEDOM OF INFORMATION ACT").

El proceso de limpieza textual fue una parte fundamental del preprocesamiento, especialmente para campos como **CDESCR** (descripción de la queja) y **MFRNAME**. Se aplicaron procedimientos de normalización que incluyeron la eliminación de espacios adicionales y caracteres no alfabéticos mediante expresiones regulares, así como la conversión del texto a minúsculas para unificar representaciones. Los valores nulos o vacíos se sustituyeron por cadenas vacías con el propósito de evitar errores en las operaciones de análisis lingüístico. Posteriormente, se realizaron pasos adicionales como la eliminación de *stopwords*, la tokenización y la lematización, con el fin de optimizar la calidad del texto para las tareas de modelado semántico.

En cuanto a las conversiones de tipo, se transformaron las variables con formato de fecha *YYYYMMDD* al tipo *datetime*, lo que permitió realizar análisis temporales precisos sobre la evolución de las quejas. Asimismo, las variables originalmente almacenadas como *object* pero con valores numéricos, como **MILES** o **YEAR**, fueron convertidas a tipos numéricos (*int o float*) para facilitar el cálculo de estadísticas descriptivas. Las variables categóricas o textuales, en cambio, se conservaron como cadenas de texto para mantener la semántica de las categorías.

Otro paso importante fue la eliminación de registros duplicados. Se detectaron y eliminaron aquellas filas que coincidían en las columnas **CDESCR**, **MAKE**, **MODEL** y **YEAR**, asegurando que cada observación representa un caso único. Esta depuración evitó redundancias comunes en bases de datos de reportes públicos, donde un mismo incidente puede ser reportado más de una vez por distintos usuarios o intermediarios.

IV.2 Investigations

4.2.1 Estructura de datos

El conjunto contiene 153,501 registros y 11 columnas. Incluye ocho campos categóricos o de texto (**ACTION_NUMBER, MAKE, MODEL, COMPONENT, MFG_NAME, CAMP_NO, SUBJECT, SUMMARY**), un numérico (**YEAR**) y dos fechas de proceso: **ODATE** (*opening date*, fecha de apertura) y **CDATE** (*closing date*, fecha de cierre), almacenadas como *YYYYMMDD* y convertibles a datetime.

```
Cargado CSV: INV_2025.csv con shape (153501, 11)
################################################################################
1) OVERVIEW
################################################################################
Shape (rows, cols): (153501, 11)
Memoria (MB): 434.21

Dtypes:
ACTION_NUMBER     object
MAKE              object
MODEL             object
YEAR             float64
COMPONENT         object
MFG_NAME          object
ODATE            float64
CDATE            float64
CAMP_NO           object
SUBJECT           object
SUMMARY           object
dtype: object
```

Figura 4.2.1: Resumen de datos en Investigations dataset.

La cobertura temporal es amplia: **ODATE** va de 1972-03-10 a 2025-09-15 (solo 0.1023% nula) y **CDATE** de 1972-05-30 a 2025-09-19 (49.2668% nula). No hay filas duplicadas. En faltantes, además de **CDATE** destaca **CAMP_NO** (18.3048%); **MAKE, MODEL** y **YEAR** rondan **0.8847%.** La **cardinalidad** es **alta** en **ACTION_NUMBER (5,307), SUBJECT (3,754), MODEL (3,667), ODATE (3,473), CDATE (3,299), CAMP_NO (2,789) y SUMMARY (2,665); MAKE (666), MFG_NAME (589), COMPONENT (415) y YEAR (63**) mantienen variedad útil para análisis.

Más adelante se sugiere usar llaves compuestas por entidad. Los textos muestran dos usos claros: **SUBJECT** es breve (mediana 24 caracteres) y funciona como título; **SUMMARY** es largo (mediana 3,176) y parece truncado en muchos casos (P90 = 3,176), lo que conviene documentar.

4.2.2 Análisis univariante

Los datos de **YEAR** se concentran entre 2003 y 2011 (P25=2003, mediana=2005, P75=2011). Con la regla IQR, los límites quedan cerca de 1991–2023 y aparecen 15,768 valores fuera de rango; además hay 6,956 inválidos/sentinela (p. ej., 9999 o fuera de 1900–2100), que deben normalizarse antes de cualquier métrica por año.

En frecuencias se observa el foco del dominio: en **MAKE** lideran *honda* (16,552), *dodge* (13,356), *kia* (12,796), *ford* (10,338) y *bmw* (9,879); en **COMPONENT** dominan *air bags* (37,845) y *air bags:frontal* (37,679), muy por encima del resto; en **MFG_NAME** destacan *chrysler (fca us, llc)* (20,746), *honda (american honda motor co.)* (20,067), *kia america, inc.* (12,810), *ford motor company* (11,732) y *general motors, llc* (10,806); y en **MODEL** sobresalen *element* (3,678), *soul* (3,616), *accord*(3,019), *sonata* (2,787) y *sonata hybrid* (2,640).

Para las fechas, **ODATE** está casi completa, mientras que **CDATE** presenta una ausencia estructural (~49%) coherente con casos abiertos o cierres no informados. En texto, la brevedad de **SUBJECT** y la extensión (con tope) de **SUMMARY** confirman su rol de título y narrativa, respectivamente.

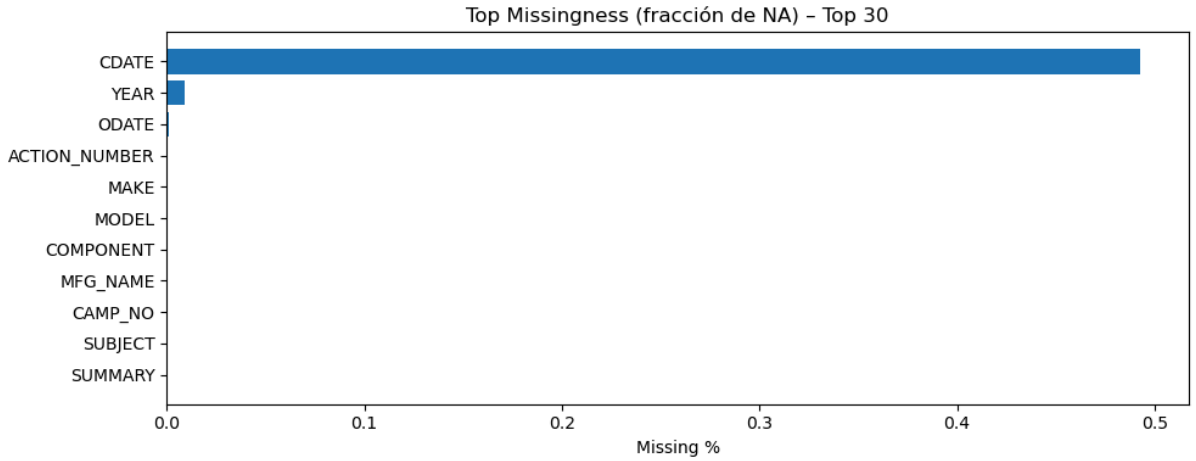


Figura 4.2.2: Reporte de datos faltantes

4.2.3 Análisis bi/multivariante

La relación temporal entre **ODATE** (opening) y **CDATE** (closing) es muy consistente: en los registros con ambas fechas, la correlación es Pearson=0.9955 y Spearman=0.9834 (n=77,875), lo que confirma el orden lógico apertura→cierre y permite calcular duraciones con confianza cuando existe cierre.

Entre **MAKE** y **MFG_NAME** hay una asociación casi determinista (Cramér's V=0.9465, muestra ≤50k; niveles **MAKE**=484, MFG=437); en la práctica, la marca queda bien definida por su fabricante, lo que simplifica normalizar catálogos y evita duplicados lógicos.

El peso de air bags en **COMPONENT** sugiere agrupamientos por marca/modelo/año y picos por periodos; para no sesgar los cruces, conviene depurar **YEAR** (outliers y sentinelas) y unificar la taxonomía de **COMPONENT** antes de comparar tasas o duraciones entre grupos.

4.2.4 Preprocesamiento de los datos

Se aplicó primero la normalización de texto en **MAKE**, **MODEL**, **COMPONENT**, **MFG_NAME**, **SUBJECT** y **SUMMARY** (recorte de espacios, conversión a minúsculas y remoción de acentos) para consolidar equivalentes. Se convirtieron ODATE (opening date, fecha de apertura) y **CDATE** (closing date, fecha de cierre) a datetime, se validó que ODATE ≤ CDATE, y se derivaron los campos de año/mes/trimestre; cuando CDATE fue nula, se etiquetó estado = abierto; cuando estuvo presente, estado = cerrado; y se calculó duration_days = CDATE − ODATE para análisis operativos.

En **YEAR**, se reemplazaron valores fuera de rango (p. ej., 9999) por NaN o "desconocido", conservando el valor crudo en un campo "_raw" para trazabilidad. Dado que no existió una clave única de una sola columna, se definieron llaves compuestas por entidad: por ejemplo, campaña con **CAMP_NO** (cuando resultó estable por jurisdicción) y acción/caso con (**ACTION_NUMBER**, **CAMP_NO**) o (**ACTION_NUMBER**, **ODATE**); para agregados de vehículo, se utilizó (MAKE, MODEL, YEAR).

Finalmente, se estandarizó la taxonomía de **COMPONENT** (p. ej., "air bags: frontal/lateral/cortina"), se mapearon sinónimos de fabricantes y marcas, y se documentaron todas las reglas de limpieza para asegurar consistencia y dejar la base lista para el diseño del Knowledge Graph.

IV.3 Recalls

4.3.1 Estructura de datos

El conjunto de datos de *recalls* contiene campañas de retiro de mercado y cumplimiento de seguridad vehicular en EE.UU. desde 1967 hasta 2025.
Después de la lectura y limpieza inicial, se consolidaron **14,290 campañas** distintas, con **18 variables clave** por cada una.
El archivo original incluía 221,835 filas debido a que una campaña puede asociarse con múltiples marcas, modelos o componentes, pero el preprocesamiento permitió reducirlo a nivel **único de campaña**.

Las variables más relevantes para el análisis son:

- **Identificación**: CAMPNO (id campaña), MFGNAME (fabricante), MAKETXT (marca), MODELTXT (modelo).
- **Producto:** COMPNAME (componente afectado), RCLTYPECD (tipo de retirada: vehículo, equipo, llanta, asiento infantil).
- **Impacto:** POTAFF_num (número potencial de unidades afectadas).
- **Temporalidad:** RCDATE (fecha de recepción), ODATE (fecha de notificación al consumidor), YEARTXT (año de modelo).
- **Documentación:** descripciones textuales de defecto, consecuencia, acción correctiva y notas.

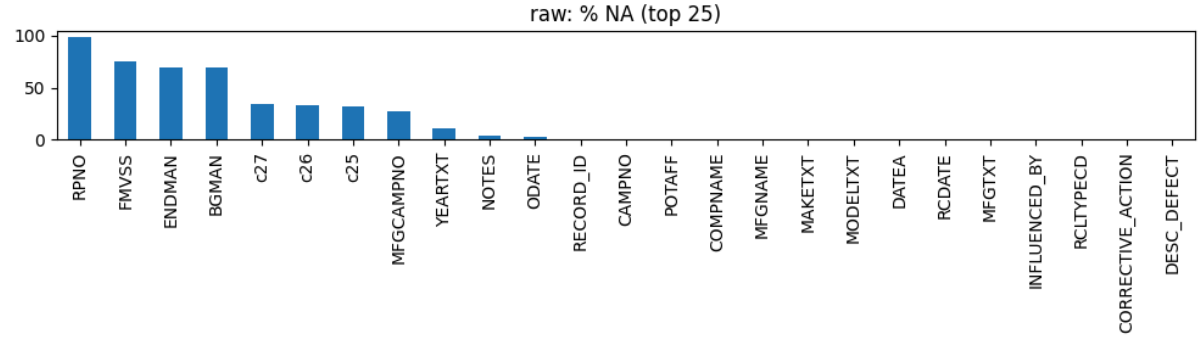
El análisis de completitud evidencia que algunas variables tienen valores faltantes significativos, por ejemplo:

- FMVSS (norma federal de seguridad vehicular), con ~79% de NA.
- BGMAN y ENDMAN (fechas de fabricación), con >40% de NA.
- YEARXT, con ~11% de NA.

A pesar de ello, la cobertura de variables críticas como identificador de campaña, fabricante, marca, modelo y componente es completa, lo que asegura la validez del análisis.

```
Forma: (221835, 29)
RECORD_ID                  object
CAMPNO                     object
MAKETXT                    object
MODELTXT                   object
YEARTXT                     int64
MFGCAMPNO                  object
COMPNAME                   object
MFGNAME                    object
BGMAN                      object
ENDMAN                     object
RCLTYPECD                  object
POTAFF                     object
ODATE                datetime64[ns]
INFLUENCED_BY              object
MFGTXT                     object
RCDATE               datetime64[ns]
DATEA                datetime64[ns]
RPNO                       object
FMVSS                      object
DESC_DEFECT                object
CONSEQUENCE_DEFECT         object
CORRECTIVE_ACTION          object
NOTES                      object
RCL_CMPT_ID                object
c24                        object
c25                        object
c26                        object
c27                        object
c28                        object
dtype: object

Numéricas: ['YEARTXT', 'POTAFF']
Fechas: ['ODATE', 'RCDATE', 'DATEA']
Categóricas: ['MAKETXT', 'MODELTXT', 'COMPNAME', 'MFGNAME', 'RCLTYPECD', 'FMVSS', 'INFLUENCED_BY']
```



La Tabla 4.3 y la gráfica de barras resumen los tipos de datos y el porcentaje de valores faltantes.

4.3.2 Análisis univariante

Las principales variables numéricas consideradas fueron **YEARTXT** (año de modelo) y **POTAFF_num** (unidades afectadas).

.

YEARTXT: La mediana se ubica en 2018, con un rango entre 1969 y 2027. Esto muestra que la mayoría de los recalls afectan a vehículos relativamente recientes, aunque existen registros atípicos hacia años futuros (probablemente estimaciones o errores de captura).

· **POTAFF_num**: La distribución es fuertemente asimétrica a la derecha. La mediana es de ~460 unidades, pero existen campañas que superan el millón de unidades afectadas, destacando el caso de los airbags Takata con más de 17 millones.

Categorías:

- RCLTYPECD: 89% de campañas corresponden a **vehículos**, 9% a **equipos**, 1.3% a **llantas** y 0.4% a **asientos infantiles**.
- COMPNAME: los componentes más recurrentes son **Equipment**, **Electrical System**, **Air Bags**, **Structure** y **Steering**.

Figura 4.3.1.a — POTAFF_num (escala lineal)

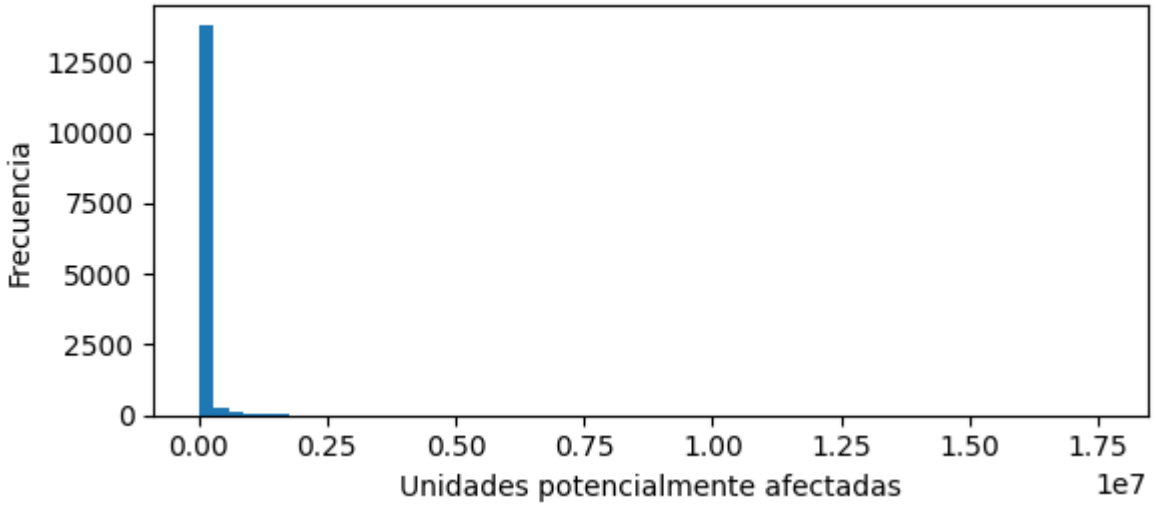
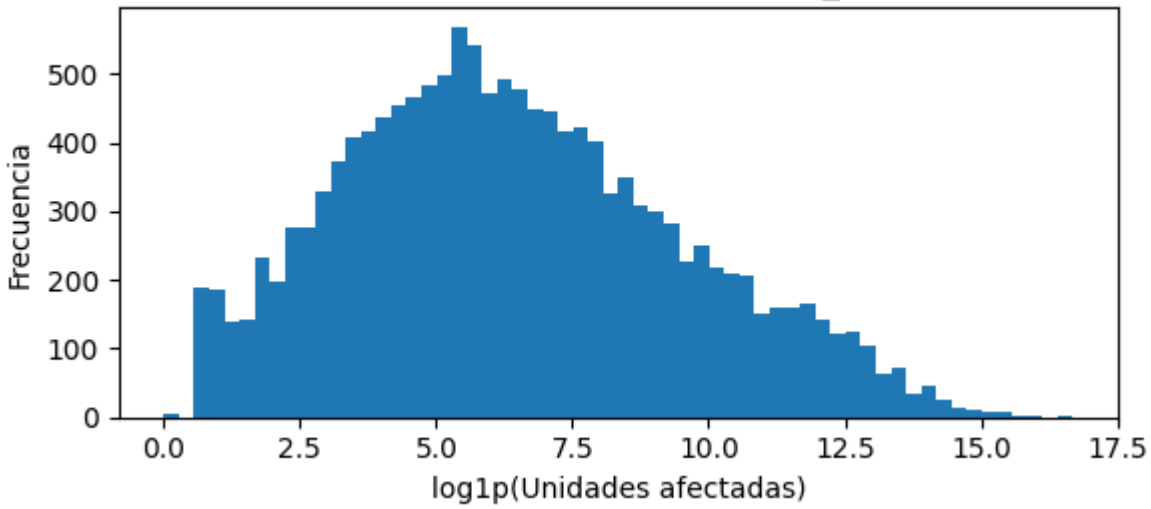
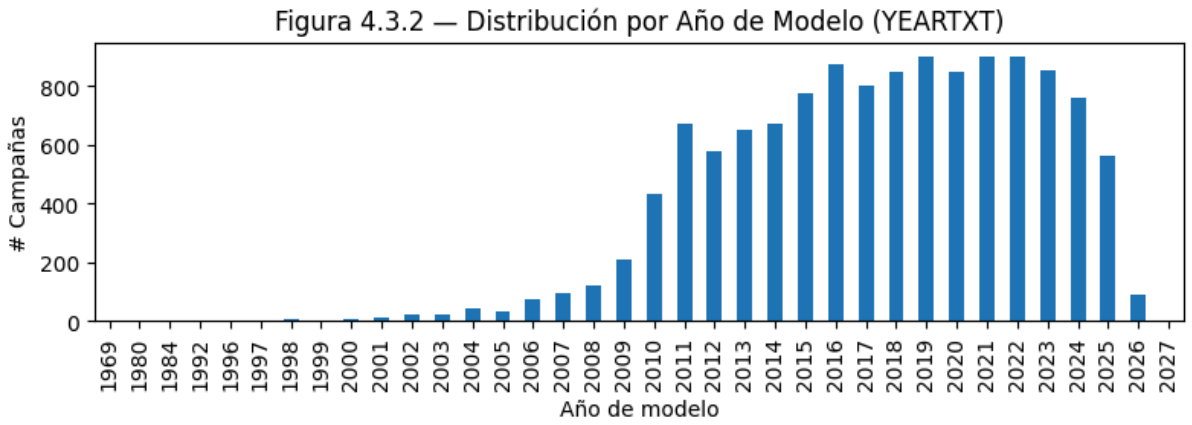


Figura 4.3.1.b — log1p(POTAFF_num)





La Figura 4.3.1a muestra la distribución de POTAFF_num en escala lineal y logarítmica, mientras que la Figura 4.3.2 refleja la concentración de campañas en modelos posteriores a 2000.

4.3.3 Análisis bi/multivariante

- **Evolución temporal:** Al agrupar por décadas, los recalls muestran un crecimiento sostenido desde 2010. En la década 2020–2025 se registran más de 5,700 campañas, indicando un aumento regulatorio o mayor vigilancia sobre fallas.

- **Fabricante–impacto:** Ford, Mercedes-Benz, Honda y GM concentran las campañas más numerosas, aunque en términos de unidades afectadas destacan casos como Oldsmobile y Kiekert, cuyos recalls tienen medianas muy superiores de vehículos afectados.

- **Componentes–impacto:** Los recalls de airbags, sistemas eléctricos y cierres/seguros de puertas aparecen entre los que más vehículos comprometen.

- **Correlaciones:** No existe relación fuerte entre año de modelo y magnitud del recall, pero sí se observa que recalls más grandes tienden a acompañarse de descripciones textuales más extensas del defecto, lo cual refleja la complejidad del problema.

Figura 4.3.3.1a — Campañas por mes (RCDATE)

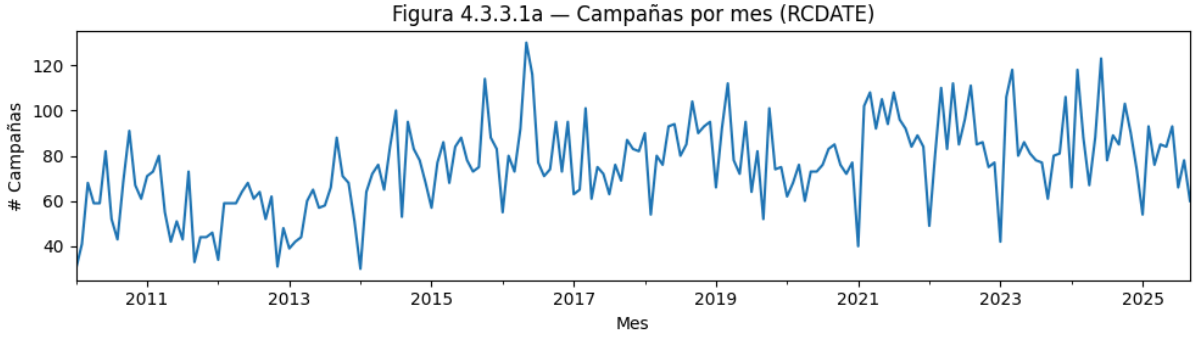


Figura 4.3.3.1b — Campañas por década (RCDATE)

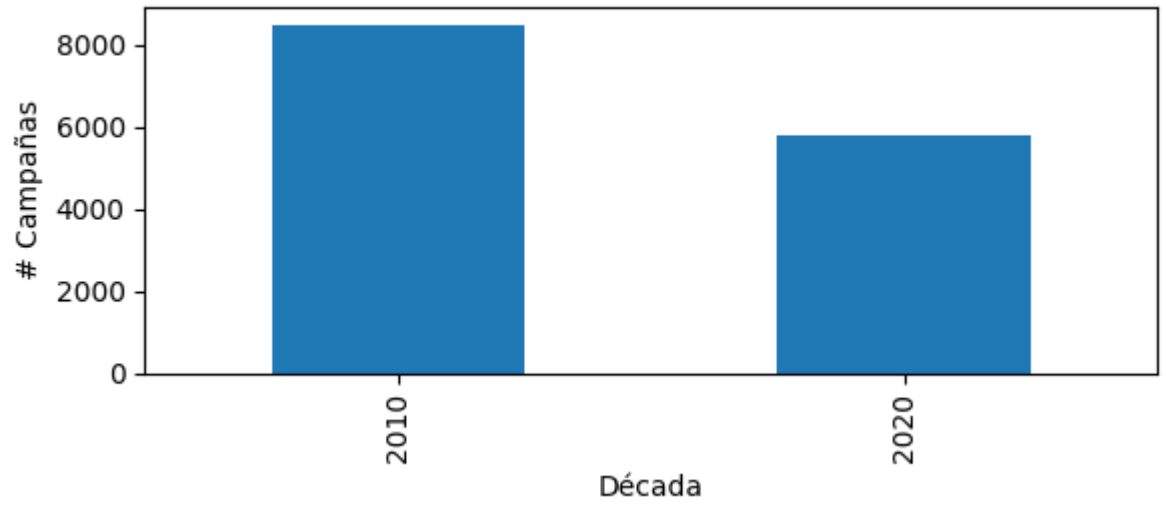


Figura 4.3.4 — Top-15 Componentes por # campañas

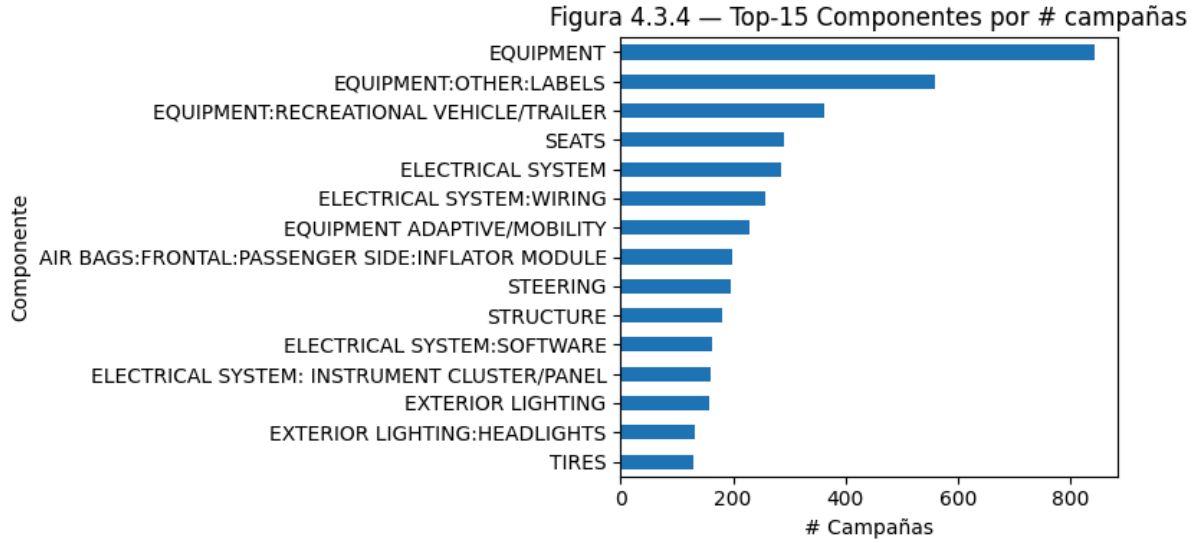
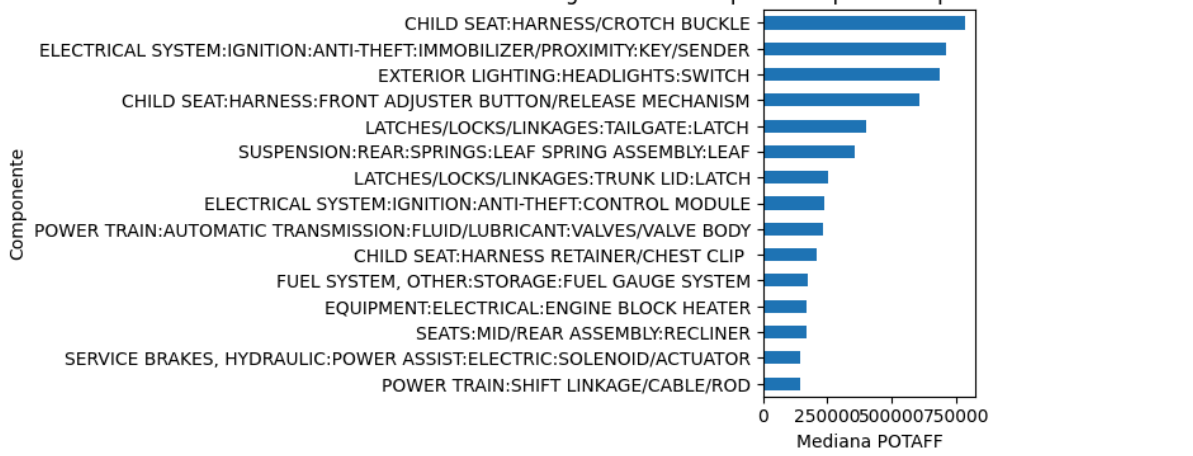


Figura 4.3.5 — Top-15 Componentes por mediana de POTAFF



4.3.4 Preprocesamiento de los datos

Las principales acciones de limpieza y normalización fueron:

1. Conversión de formatos: fechas (RCDATE, ODATE, DATEA) a formato estándar, y POTAFF a numérico.
2. Tratamiento de valores faltantes: YEARTXT=9999 se recodificó como NA; columnas con >70% de NA se documentaron como incompletas pero se mantuvieron.
3. Normalización categórica:
    a. RCLTYPECD se mapeó a categorías legibles: Vehicle, Equipment, Tire, Child Seat.
    b. Fabricantes (MFGNAME) y marcas (MAKETXT) se transformaron a mayúsculas limpias.
    c. Componentes (COMPNAME) se redujeron de 589 etiquetas a 356 categorías canónicas, mediante consolidación jerárquica (sistema, subsistema, parte).
4. Consistencia campaña–impacto: se verificó que cada CAMPNO tuviera un único valor de POTAFF_num y fechas consistentes.
5. Exportación para grafo: se generaron tablas de nodos y aristas (campañas, fabricantes, marcas, modelos, componentes, años y normas FMVSS) listas para su integración en Neo4j o NetworkX.

V. Conclusiones del EDA

El EDA evidenció que los tres conjuntos pueden ser integrados bajo un marco semántico común, ya que comparten atributos fundamentales como fabricante, modelo, año, componente y descripción del problema. Estas variables actúan como puntos de enlace naturales para representar relaciones en un grafo de conocimiento. Por ejemplo, los vínculos entre *modelo*, *componente* y *tipo de falla* observados en las quejas pueden conectarse con los resultados de las investigaciones y los registros de *recalls* asociados al mismo componente o fabricante.

Desde una perspectiva analítica, los datos presentaron distribuciones sesgadas en variables numéricas como millas recorridas o velocidad del vehículo, así como una concentración significativa de reportes en los últimos diez años. Este comportamiento sugiere una evolución en la frecuencia y el tipo de incidentes reportados, lo cual puede ser modelado semánticamente mediante relaciones temporales que representen la aparición, persistencia o resolución de problemas específicos.

El análisis textual, que incluyó la extracción de entidades y la identificación de temas latentes, demostró que las descripciones de quejas y reportes contienen información semántica valiosa. Se reconocieron entidades correspondientes a fabricantes, componentes, fechas y números de campaña, lo que permite vincular los textos no estructurados con las variables estructuradas del dataset. Estos hallazgos respaldan la posibilidad de representar tanto conocimiento explícito (campos estructurados) como conocimiento implícito (información textual) dentro del mismo grafo.

En conjunto, los resultados del EDA permiten definir una estructura conceptual inicial compuesta por entidades como *Fabricante*, *Modelo*, *Componente*, *Vehículo*, *Queja*, *Investigación* y *Recall*, unidas por relaciones como *produce*, *incorpora*, *presenta falla*, *es investigado* o *es retirado del mercado*. Este esquema ontológico servirá como punto de partida para el diseño del grafo de conocimiento, que buscará integrar información estructurada y no estructurada con el objetivo de facilitar el análisis semántico de fallas vehiculares, tendencias y riesgos de seguridad.

.Las relaciones entre ellas se derivan de vínculos explícitos e implícitos observados en los datos, tales como:

- *Fabricante produce Modelo*
- *Modelo presenta Componente defectuoso*
- *Componente está asociado con Queja*
- *Queja origina Investigación*
- *Investigación deriva en Recall*

VI. Referencias

National Highway Traffic Safety Administration. (2008). *Recall data file: Flat file format and data dictionary* (Appendix A). U.S. Department of Transportation. https://www-odi.nhtsa.dot.gov/downloads/

National Highway Traffic Safety Administration. (2025). *NHTSA vehicle recall dataset (1967–2025)*. U.S. Department of Transportation, Office of Defects Investigation. https://www.nhtsa.gov/recalls

NHTSA. (s. f.). Investigations datasets & APIs. Recuperado de https://www.nhtsa.gov/nhtsa-datasets-and-apis#investigations

NHTSA. (s. f.). Complaints datasets & APIs. Recuperado de https://www.nhtsa.gov/nhtsa-datasets-and-apis#complaints

SpaCy. (2023). Industrial-strength Natural Language Processing in Python. Explosion AI. https://spacy.io

PyData. (2023). *Matplotlib Documentation*. PyData Development Team. https://matplotlib.org/stable/index.html

Galli, S. (2022). *Python Feature Engineering Cookbook* (2nd ed.). Packt Publishing. https://learning.oreilly.com/library/view/python-feature-engineering/9781804611302/

Pyle, D. (1999). *Data preparation for data mining*. Morgan Kaufmann.

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.

Wickham, H., & Grolemund, G. (2017). *R for data science: Import, tidy, transform, visualize, and model data*. O'Reilly Media.