INSTITUTO SUPERIOR POLITÉCNICO DE CÓRDOBA

TECNICATURA SUPERIOR EN CIENCIA DE DATOS E INTELIGENCIA ARTIFICIAL

Módulo de Práctica Profesionalizante

Entrega 2: Análisis Exploratorio de Datos de Flota de Autos

16/09/2024

Integrantes:

- Galeano, Agustín
- López, Erick
- Nüesch, Christian
- Zurita Rojo, Debora

Índice

Resumen Ejecutivo

En esta fase del proyecto, se llevó a cabo un **Análisis Exploratorio de Datos (EDA)** sobre un dataset de una flota de automóviles con el objetivo de entender la estructura y características de los datos, identificar patrones iniciales y detectar problemas como valores atípicos o datos faltantes. Los análisis proporcionaron insights preliminares que serán útiles para la toma de decisiones informadas en futuras etapas del proyecto.

Introducción

El **objetivo del EDA** es explorar los datos, generar visualizaciones descriptivas y descubrir patrones que permitan mejorar la comprensión de las variables. Esta fase es esencial para el análisis posterior, ya que establece las bases para modelado predictivo o recomendaciones estratégicas para la flota automotriz.

El dataset incluye información detallada sobre automóviles, como el precio, marca, modelo, año de fabricación, tipo de combustible, kilometraje acumulado, y otras características relevantes.

Metodología

Proceso de Análisis

El análisis se llevó a cabo utilizando **librerías de Python** como **Pandas** y **Seaborn** para la manipulación y visualización de datos. Los pasos realizados fueron:

1. **Carga y limpieza del dataset**: Identificación y tratamiento de valores faltantes o inconsistentes.
2. **Análisis univariado**: Exploración de las variables individuales mediante estadísticas descriptivas.
3. **Visualización**: Histogramas, boxplots y gráficos de barras para representar distribuciones y relaciones.
4. **Análisis bivariado**: Exploración de relaciones entre pares de variables, utilizando gráficos de dispersión y mapas de calor de correlaciones.

Datos Utilizados

El dataset con el que se trabajó en este análisis contiene múltiples características relacionadas con automóviles de una flota. A continuación, se describen las columnas del dataset:

- **Precio**: precio del auto, expresado en la moneda especificada en la columna "Moneda". **Tipo de dato**: Entero.
- **Marca**: marca del auto. **Tipo de dato**: Cadena de caracteres.
- **Modelo**: modelo específico del auto. **Tipo de dato**: Cadena de caracteres.
- **Año**: año de fabricación del auto. **Tipo de dato**: Entero.
- **Color**: color del exterior del auto. **Tipo de dato**: Cadena de caracteres.
- **Combustible**: tipo de combustible que utiliza el vehículo. **Tipo de dato**: Cadena de caracteres.
- **Puertas**: cantidad de puertas que tiene el vehículo. **Tipo de dato**: Entero.
- **Caja**: tipo de caja de cambios del vehículo. **Tipo de dato**: Cadena de caracteres.
- **Motor**: tamaño del motor del auto,expresado en litros. **Tipo de dato**:Flotante con un decimal.
- **Carrocería**: tipo de carrocería del vehículo. **Tipo de dato**: Cadena de caracteres.
- **Kilómetros**: kilometraje acumulado por el auto.**Tipo de dato**: Entero.
- **Moneda**: moneda en la que se cotiza el precio del auto. **Tipo de dato**: Cadena de caracteres.

Hallazgos Clave

Análisis Descriptivo

Una vez limpiado nuestros datos, con el método “describe( )”, analizando sobre nuestras variables de interés , obtenemos la siguiente información:

De lo cual, podemos inferir lo siguiente:

	Precio	Año	Puertas	Motor	Kilómetros
count	500.00	500.00	500.00	500.00	500.00
mean	3989186.54	2016.27	4.47	1.88	74732.35
std	2892938.07	3.71	0.76	0.70	46804.16
min	8000.00	1995.00	2.00	1.00	500.00
25%	2380000.00	2014.00	4.00	1.60	43750.00
50%	3489950.00	2017.00	5.00	1.60	65750.00
75%	5259975.00	2019.00	5.00	2.00	99100.00
max	14299000.00	2022.00	5.00	6.40	335000.00

1. **Precio**
   - El precio promedio es de $3.989.186, con una alta variabilidad que refleja la diversidad de la flota.
   - El 25% de los autos tiene un precio inferior a $2.380.000, mientras que el 75% está por debajo de $5.259.975, lo que muestra una distribución sesgada hacia precios más bajos.
2. **Año**
   - La mayoría de los autos tienen menos de 10 años. El promedio de fabricación es 2016, lo que sugiere que la flota es relativamente moderna.
3. **Kilómetros**
   - El kilometraje promedio es de 74.732 km. La mayoría de los autos han recorrido distancias moderadas, con un 75% por debajo de 99.100 km, indicando que la flota está en buen estado.
4. **Puertas**
   - La mayoría de los autos tiene entre **4 y 5 puertas**, con un promedio de **4.47 puertas**. Esto sugiere que la flota está compuesta principalmente por **vehículos familiares o utilitarios**, que suelen tener más puertas para comodidad de los pasajeros o para uso comercial.
5. **Motor**
   - El **motor promedio** es de **1.88** litros, lo que indica que la flota está compuesta principalmente por **autos de cilindrada moderada**, comúnmente usados para tareas estándar.
   - Hay autos con motores pequeños (**1.0 litros**), y el valor máximo de **6.4 litros** probablemente corresponda a un auto de alta gama o de características muy específicas.
   -

Resumen del análisis:

- La **flota** está compuesta por autos **relativamente nuevos**, con una mayoría que tiene entre **4 y 5 puertas**, motores moderados y kilometrajes moderados a bajos.
- Hay una **gran variabilidad en los precios**, que puede estar influida por el estado, año y características específicas de los vehículos.
- Se podrían realizar análisis adicionales para explorar la relación entre **precio y kilometraje**, o entre **motor y precio**, para determinar qué factores influyen más en el valor de los autos en la flota.

- Tendencias y Patrones: Descripción de cualquier tendencia significativa o patrones encontrados.
- Análisis de Correlación: Relación entre diferentes variables (por ejemplo, cómo el precio se correlaciona con el año y los kilómetros).

Visualizaciones y Tendencias

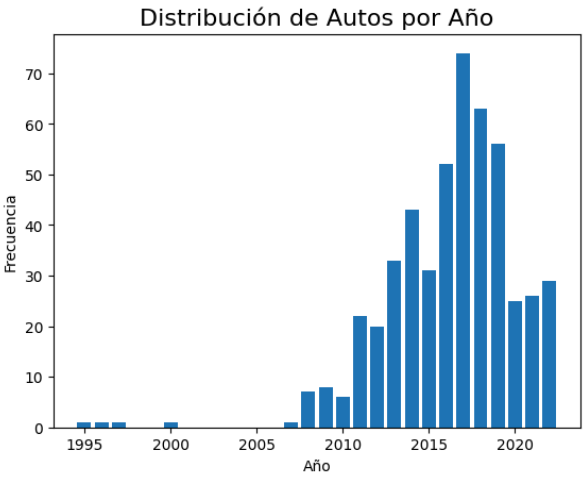
Incluye **gráficos y tablas claves** que respalden tus hallazgos. Asegúrate de que sean claros y relevantes para los puntos que estás tratando de comunicar.

Análisis Univariado

Gráficas

Realizamos un análisis de las variables individuales para conocer mejor nuestro dataset y entender su distribución. Aquí se han utilizado histogramas, boxplots y estadísticas descriptivas (media, mediana, desviación estándar, etc.) para las variables numéricas. Para las variables categóricas, se pueden calcular frecuencias y proporciones, entre otras.

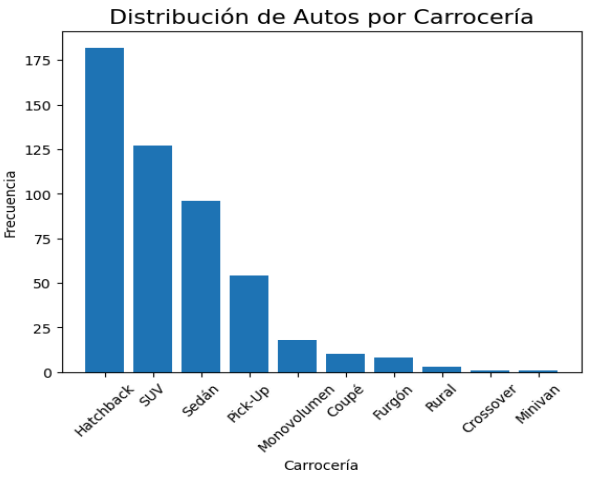
Distribución de Autos por Año



Al observar el gráfico de la **distribución de autos de la flota por año**, podemos observar

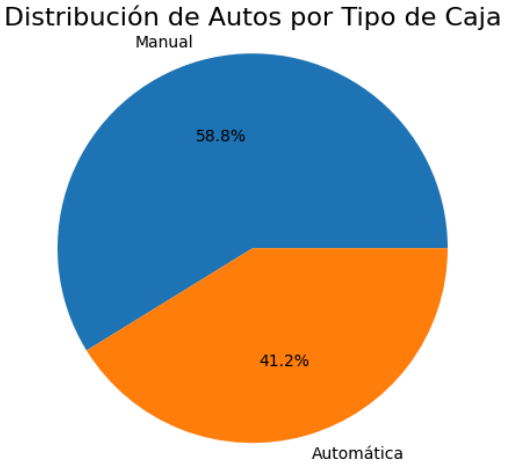
- **Concentración temporal**: Observamos que la mayoría de los autos en la flota fueron adquiridos entre **2012 y 2020**, con un pico notable alrededor de **2015-2016**. Esto sugiere que la empresa realizó compras significativas de vehículos durante este período, probablemente debido a la necesidad de renovar la flota o expandir sus operaciones.
- **Flota reciente**: Hay muy pocos autos anteriores a **2010**, lo que podría indicar que los autos más antiguos han sido dados de baja o vendidos. Esto puede también señalar que la flota se mantiene moderna, con vehículos relativamente nuevos, que suelen tener menores costos de mantenimiento y mayores eficiencias.
- **Renovación y reemplazo**: El descenso gradual en la cantidad de autos desde el **2016** sugiere que la flota se ha estabilizado o se ha reducido en años más recientes. Esto podría estar relacionado con cambios en las políticas de la empresa o en las necesidades operativas.
- **Recuperación en 2021**: Hay un leve aumento en la adquisición de autos en **2021**, lo que podría indicar un repunte en la compra de vehículos después de una desaceleración o estabilización en los años anteriores.

Distribución de Autos por Carrocería



- **Dominancia de Hatchbacks y SUVs:** Claramente, los vehículos tipo hatchback y SUV son los más numerosos en esta flota. Esto sugiere una preferencia por vehículos más compactos y versátiles, que se adaptan a diversos usos y estilos de vida.
- **Menor presencia de Coupés, Furgones, Rurales, Crossovers y Minivans:** Estos tipos de carrocería tienen una representación mucho menor en la flota. Esto podría indicar que la flota está orientada hacia un público que prioriza la funcionalidad y la capacidad de carga sobre el diseño deportivo o la capacidad de pasajeros de una minivan.
- **Sedanes y Pick-ups con representación intermedia:** Los sedanes, vehículos tradicionales y familiares, y las pick-ups, más orientadas a trabajos livianos o uso recreativo, ocupan una posición intermedia en la distribución. Esto sugiere un equilibrio entre la preferencia por vehículos más convencionales y aquellos con capacidades específicas.

Distribución de Autos por Carrocería



el gráfico de torta sobre la **distribución de autos por tipo de caja** :

- **Predominio de caja manual** : El **58.8%** de los autos de la flota tienen caja manual, lo que indica una preferencia o disponibilidad mayor de este tipo de vehículos. Esto podría deberse a que los autos manuales suelen ser más económicos tanto en compra como en mantenimiento, lo que podría reducir costos operativos.
- **Fuerte presencia de autos automáticos** : El **41.2%** de la flota es automática, una proporción significativa. Esto podría reflejar una tendencia hacia la comodidad y la eficiencia en ciertas tareas o para mejorar la experiencia de conducción, particularmente si los autos automáticos son utilizados en zonas urbanas o en operaciones que requieren menos esfuerzo para el conductor.

En resumen, la flota tiene una **ligera preferencia por vehículos con caja manual** , lo que podría estar impulsado por consideraciones de costo, aunque la **proporción significativa de autos automáticos** sugiere una inclinación hacia la comodidad en ciertas áreas.

Distribución de Autos por Tipo de Moneda

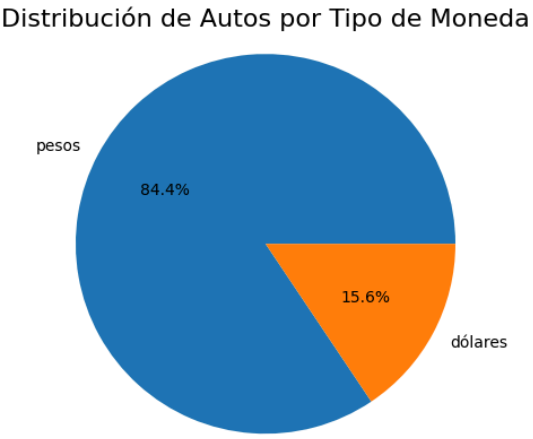


gráfico de torta que muestra la **distribución de autos por tipo de moneda**:

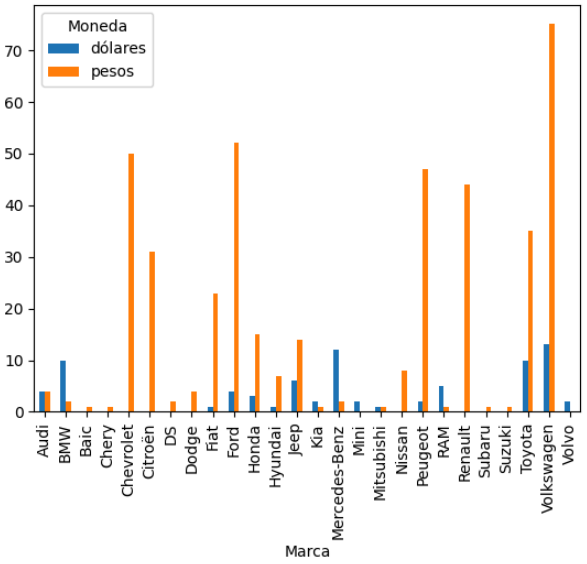
- **Mayoría de autos valuados en pesos**: El **84.4%** de los autos están valuados en pesos, lo que refleja que la gran parte de la flota está orientada al mercado local, en el que las transacciones en moneda local son más comunes. Esto también puede estar relacionado con la preferencia de la empresa por evitar la volatilidad del dólar.
- **Autos valuados en dólares**: El **15.6%** de los autos están valuados en dólares, lo que puede indicar que estos vehículos son modelos importados, de gama más alta, o que la empresa los adquirió a través de concesionarios que operan en dólares. Los autos en dólares probablemente sean más costosos, lo que podría implicar un segmento más exclusivo dentro de la flota.

Análisis Bivariado

Relaciones entre Variables

En esta última parte de los análisis, procedemos a analizar la relación entre pares de variables para **identificar patrones o correlaciones** si es que los hubiera. Para ello nos ayudamos de gráficos de dispersión (scatter plots) y mapas de calor de correlación (heatmaps) para evaluar la relación entre variables numéricas.

Distribución de monedas (dólares y pesos) según marca de vehículo

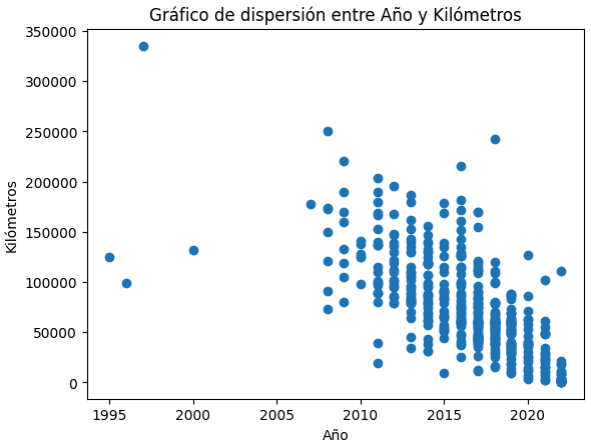


El gráfico que presentas es un **gráfico de barras agrupadas**. Este tipo de gráfico se utiliza para comparar la frecuencia o cantidad de una variable (en este caso, la cantidad de monedas, tanto dólares como pesos) entre diferentes categorías (las distintas marcas de autos).

Interpretación General:

- **Comparación por Marca:** Cada barra representa una marca de auto. La altura de cada barra indica la cantidad de monedas (dólares o pesos) asociada a esa marca.
- **Dos Monedas:** Se están comparando dos tipos de monedas: dólares y pesos. Cada marca tiene dos barras, una para cada tipo de moneda.
- **Variabilidad entre Marcas:** Se observa una gran variabilidad en la cantidad de monedas entre las diferentes marcas. Algunas marcas tienen una cantidad significativamente mayor de dólares o pesos que otras.

Análisis de la relación entre el año de fabricación y el kilometraje de vehículos



Un gráfico de dispersión nos permite visualizar la relación entre dos variables numéricas. En este caso, estamos comparando el **año** (en el eje horizontal) con los **kilómetros** recorridos (en el eje vertical).

Interpretación del Gráfico:

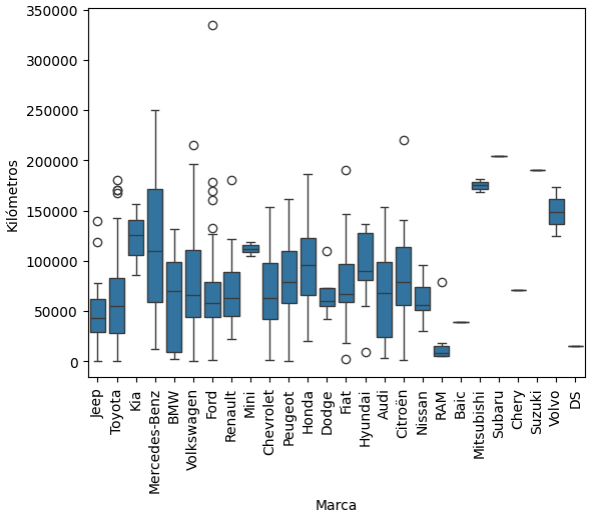
1. **Distribución de los datos:** Los puntos en el gráfico representan vehículos individuales. La mayoría de los vehículos se concentran en los años más recientes (a partir de 2005 aproximadamente), lo que sugiere que la mayor parte de la data corresponde a vehículos más modernos.
2. **Relación entre las variables:**
    - **No hay una relación lineal clara:** No observamos una tendencia ascendente o descendente definida a medida que aumenta el año. Es decir, no podemos afirmar que a mayor año, mayor cantidad de kilómetros recorridos, o viceversa.
    - **Mayor dispersión en años recientes:** En los años más recientes, los datos se encuentran más dispersos, lo que indica una mayor variabilidad en la cantidad de kilómetros recorridos por vehículos del mismo año. Esto podría deberse a diferentes factores como el uso que se le da a cada vehículo, la marca, el modelo, etc.
    - **Algunos valores atípicos:** Se observan algunos puntos aislados que representan vehículos con un kilometraje inusualmente alto para su año. Estos podrían ser casos especiales, como vehículos utilizados para fines comerciales o de transporte de larga distancia.

Conclusiones preliminares:

- **No hay una relación directa entre el año de fabricación y los kilómetros recorridos.** Otros factores, como el uso individual del vehículo, influyen más en el kilometraje.
- **La mayoría de los datos corresponden a vehículos más recientes.**
- **Existe una gran variabilidad en el kilometraje, incluso entre vehículos del mismo año.**

Comparaciones Categóricas

Distribución de kilometrajes por marca



este gráfico puede aportar información muy interesante. Un análisis del mismo sería el siguiente:

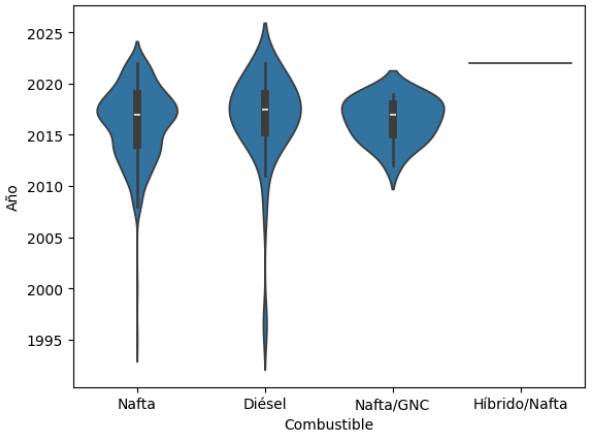
1. Rango de los kilómetros recorridos:
    - Las marcas como **BMW**, **Renault**, **Jeep**, y **Volkswagen** presentan una mayor **variabilidad** en los kilómetros recorridos, ya que sus cajas (el rango intercuartílico) son más amplias, lo que indica que hay más dispersión en los datos.
    - Por otro lado, marcas como **DS**, **Suzuki**, **Volvo**, y **Chery** tienen rangos de kilómetros mucho más pequeños, lo que sugiere que los vehículos de estas marcas tienden a recorrer distancias más similares entre sí.
2. Posición de la mediana:

- La mediana (línea en el centro de la caja) es más alta en marcas como **BMW** y **Renault**, lo que podría sugerir que los vehículos de estas marcas suelen tener un mayor **kilometraje medio** comparado con otras marcas.
- Marcas como **DS**, **Chery**, y **Suzuki** tienen una mediana significativamente más baja, lo que indica que en general recorren menos kilómetros.

3. Presencia de outliers:
   - Hay **outliers** (puntos fuera de los "bigotes" del boxplot) en varias marcas como **Volkswagen**, **Mercedes-Benz**, **Renault**, y **Chevrolet**, lo que sugiere que hay algunos vehículos que han recorrido una cantidad inusualmente alta de kilómetros en comparación con el resto de los vehículos de la misma marca.
   - Por otro lado, marcas como **DS**, **Chery**, y **Suzuki** no tienen outliers significativos, lo que indica que sus datos son más consistentes y no tienen vehículos que se desvíen drásticamente en términos de kilómetros recorridos.
4. Marcas con menos kilómetros recorridos:
   - Las marcas como **DS**, **Suzuki**, y **Chery** parecen tener, en general, **vehículos con menos kilómetros recorridos**, ya que tanto la mediana como los bigotes están en la parte inferior del gráfico.
   - Esto podría sugerir que los vehículos de estas marcas son usados con menos frecuencia o que son vehículos más nuevos.
5. Marcas con mayor kilometraje:
   - En comparación, marcas como **BMW**, **Renault**, y **Volkswagen** tienden a tener **mayor kilometraje** en promedio. Esto podría indicar que los autos de estas marcas son más usados o que tienen una mayor vida útil en términos de kilómetros recorridos.
6. Distribución en algunas marcas es más compacta:
   - Las marcas como **Suzuki** y **DS** tienen una distribución de kilómetros más compacta, lo que puede indicar que los autos de estas marcas suelen recorrer distancias más consistentes y cercanas entre sí.

En general, este boxplot sugiere que aunque hay algunas marcas cuyos vehículos recorren más kilómetros que otras, existe una amplia variabilidad entre las marcas en cuanto al uso (kilómetros recorridos). Algunas marcas tienden a tener autos que recorren más kilómetros en promedio (**BMW**, **Renault**, **Volkswagen**), mientras que otras tienden a tener vehículos que recorren menos kilómetros o más consistentemente (**Suzuki**, **DS**, **Chery**).

Distribución de Año por Tipo de Combustible



Conclusiones a partir del gráfico:

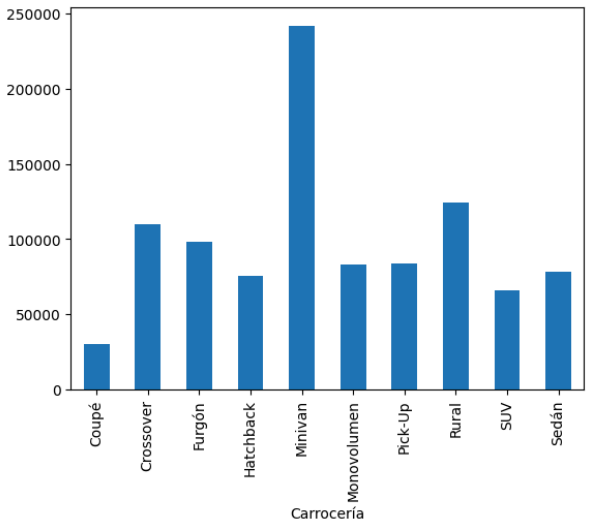
- **Nafta y Diesel:** Presentan distribuciones similares, con una mediana alrededor de 2015 y una amplia gama de años, desde modelos más antiguos hasta más recientes.
- **Nafta/GNC:** También muestra una distribución similar a la nafta y el diésel, con una mediana ligeramente más baja, lo que podría indicar que los vehículos que utilizan este tipo de combustible tienden a ser un poco más antiguos.
- **Híbrido/Nafta:** Se observa una distribución distinta, con una mediana más reciente y una menor dispersión de los datos. Esto sugiere que los vehículos híbridos tienden a ser modelos más nuevos.

- **Presencia de outliers:** La longitud de los "bigotes" y la presencia de puntos individuales fuera de ellos indican la existencia de algunos vehículos con años de fabricación inusualmente altos o bajos para cada tipo de combustible.

Interpretación general:

1. **Variedad de años:** Existe una amplia variedad de años de fabricación en todos los tipos de combustible, lo que indica que el conjunto de datos incluye vehículos de diferentes generaciones.
2. **Tendencia a modelos más recientes en híbridos:** Los vehículos híbridos tienden a ser más nuevos en comparación con los que utilizan otros tipos de combustible.
3. **Similitudes entre nafta y diésel:** Las distribuciones de nafta y diésel son bastante similares, lo que sugiere que no hay una diferencia significativa en términos de antigüedad entre estos dos tipos de vehículos.

Kilometraje y carrocería: Una relación estrecha



De este gráfico de barras también se pueden extraer conclusiones interesantes:

1. La categoría **Minivan** destaca claramente como la que tiene el **kilometraje más alto** en comparación con las demás. Esto podría deberse a que las minivanes son vehículos que se utilizan para transporte familiar o de pasajeros a largas distancias, por lo que es lógico que acumulen más kilómetros.
2. Por otro lado, los vehículos tipo **Coupé** tienen el **kilometraje más bajo** de todas las categorías. Esto podría indicar que estos vehículos, que suelen ser deportivos o de dos puertas, se utilizan con menos frecuencia o para trayectos más cortos.
3. Las carrocerías como **Crossover**, **Rural**, y **SUV** presentan un **kilometraje moderado**. Estos vehículos, al ser multifuncionales y con mayor capacidad de carga o espacio, también suelen recorrer distancias considerables, aunque no tanto como las minivanes.
4. En contraste, **Furgón**, **Pick-Up**, y **Hatchback** parecen tener un kilometraje algo menor, posiblemente porque se usan más en contextos urbanos o para tareas específicas que no requieren largos desplazamientos.
5. Se pueden hacer ciertas inferencias sobre el **tipo de uso** de los vehículos según la carrocería. Por ejemplo:
    - **Minivan**: Uso familiar o de transporte de pasajeros a largas distancias.
    - **Coupé**: Vehículos deportivos o para uso recreativo, probablemente no usados con frecuencia.
    - **SUV y Crossover**: Vehículos que, aunque no tienen el kilometraje tan alto como las minivanes, son utilizados para viajes de mediana distancia o en terrenos mixtos (ciudad/campo).
    - **Furgón y Pick-Up**: Podrían usarse para tareas más específicas (transporte de carga, trabajos rurales, etc.), y por eso tienen un kilometraje más moderado.
6. Los vehículos tipo **Sedán** y **Rural** tienen un **kilometraje similar**, lo que podría reflejar que ambos se utilizan tanto para uso diario como para viajes de media distancia, aunque no tan intensivamente como los vehículos tipo Minivan o SUV.

Este gráfico sugiere que el **kilometraje acumulado de los vehículos varía bastante según la carrocería**, lo cual está directamente relacionado con el **uso típico** de cada tipo de vehículo. Las minivanes, por su naturaleza de vehículo familiar o de transporte de pasajeros, tienden a recorrer más kilómetros, mientras que los Coupé, probablemente deportivos o de lujo, recorren menos kilómetros. Los SUVs y Crossover también presentan un uso considerable, probablemente debido a su versatilidad.

Conclusiones y Recomendaciones

- Conclusiones: Resumen de los hallazgos más importantes.
- Recomendaciones: Sugerencias basadas en los hallazgos, que pueden incluir cambios en estrategias, áreas de mejora, o futuras investigaciones.
-