Exploración y preprocesamiento de datos

Introducción

El objetivo de esta memoria es presentar un análisis exploratorio de datos (EDA) realizado sobre un conjunto de datos que contiene información sobre estaciones de servicio y precios de carburantes en España. Para este análisis, se ha utilizado Pandas, Matplotlib y AutoViz, permitiendo visualizar, limpiar y transformar los datos con el fin de obtener conclusiones relevantes.

Este informe detalla:

- La estructura y características del dataset.
- Las transformaciones realizadas sobre los datos.
- El análisis de las variables principales.
- Las visualizaciones obtenidas y la interpretación de los resultados.

Descripción de los Datos

El dataset proporcionado contiene 11,865 registros y 10 columnas, cada una representando una característica de las estaciones de servicio y los precios de los carburantes.

Transformaciones de Datos Realizadas

Para mejorar la calidad del análisis, se realizaron las siguientes transformaciones y limpiezas de datos:

1. conversión de Precios a Valores Numéricos

Los valores de Precio gasolina 95 E5 y Precio gasóleo A estaban representados como texto con formato europeo (1,529 en lugar de 1.529). Se realizó la conversión a formato numérico.

2. Manejo de Valores Faltantes

Se identificaron valores faltantes en los precios de los carburantes y en `Tipo servicio`. Se rellenaron los valores nulos con la media o con 'Desconocido' en el caso de valores categóricos.

3. Eliminación de duplicados

Se detectaron 122 filas duplicadas en el dataset, que fueron eliminadas para evitar sesgos en el análisis.

Exploración de Variables y Análisis Estadístico

Se generaron histogramas de distribución de precios y medidas estadísticas como media, mediana y rango.

Análisis de Correlación entre Variables

Para identificar posibles relaciones entre las variables numéricas del dataset, se generó una matriz de correlación utilizando el coeficiente de Pearson. Dado que en el conjunto de datos solo tenemos dos variables numéricas (`Precio gasolina 95 E5` y `Precio gasóleo A`), se analizó su correlación para evaluar si existe alguna dependencia entre ellas.

Interpretación de los Resultados

- **Distribución de precios**

El análisis con **Matplotlib** ha permitido visualizar la distribución de precios de la gasolina 95 E5 y del gasóleo A en la provincia de Madrid mediante histogramas. Ambas variables presentan una **distribución unimodal** y relativamente simétrica, con un claro pico en el rango entre **1,50 € y 1,60 €**, lo que indica que la mayoría de las estaciones fijan precios en torno a esos valores. Esta concentración sugiere una **estandarización del mercado**, posiblemente debida a la regulación o a la presión competitiva entre marcas.

Aunque la mayoría de precios se mantienen dentro de un rango ajustado, se han detectado **algunos valores atípicos** en ambos extremos de la distribución: precios más bajos en algunas estaciones automáticas o low cost, y precios más elevados en estaciones de servicio convencionales, posiblemente por ofrecer servicios adicionales o por su localización estratégica.

- **Correlación entre carburantes**

El análisis estadístico llevado a cabo mediante AutoViz revela una **correlación positiva significativa** entre el precio de la gasolina 95 E5 y el gasóleo A. Esto implica que cuando sube o baja el precio de uno, es probable que el otro siga la misma tendencia. Este comportamiento es esperable, ya que ambos carburantes están influenciados por los mismos factores estructurales: coste del crudo, impuestos, gastos de logística y refino, así como políticas económicas generales.

- **Impacto del tipo de estación**

AutoViz también permitió analizar cómo influyen las variables categóricas en el precio, en especial **el tipo de venta** y **el tipo de servicio**. Se observó que las estaciones con tipo de venta **particular (P)** y servicio **automático (A)** presentan **precios más competitivos**, probablemente debido a sus **menores costes operativos**. Por el contrario, las estaciones con servicio asistido o de grandes marcas tienden a tener precios ligeramente superiores, lo que refleja su modelo de negocio más tradicional y enfocado al cliente.

Esta diferenciación sugiere que los **modelos de gestión y la automatización** juegan un papel clave en la política de precios, permitiendo a las estaciones automáticas captar una parte importante del mercado gracias a tarifas más ajustadas.

- **Calidad de los datos**

Durante el preprocesamiento se detectaron **valores nulos** y **filas duplicadas**, que fueron correctamente tratados antes del análisis. La conversión de los precios a formato numérico estándar (de coma a punto decimal) fue un paso clave para garantizar la integridad de las visualizaciones y los cálculos estadísticos posteriores.

Posibles Mejoras y Trabajo Futuro

Existen múltiples maneras de enriquecer este análisis para futuras investigaciones. Algunas de las mejoras que podrían implementarse incluyen:

- Incorporación de datos temporales: Analizar la variación de precios a lo largo del tiempo permitiría detectar tendencias estacionales o eventos que influyan en los costos de los carburantes.
- Análisis geoespacial: Visualizar la distribución de estaciones de servicio en un mapa ayudaría a identificar patrones en los precios según la ubicación.
- Modelos predictivos: Utilizar técnicas de Machine Learning para predecir la evolución de los precios basados en datos históricos y variables macroeconómicas.
- Comparación entre provincias: Ampliar el análisis para incluir diferentes regiones permitiría identificar diferencias en la competencia y en la regulación del mercado de carburantes.

Conclusiones

En este estudio se ha llevado a cabo un análisis exhaustivo de los datos de precios de carburantes en estaciones de servicio utilizando técnicas de análisis exploratorio de datos (EDA). Se han aplicado diversas transformaciones para limpiar y normalizar la información, asegurando la calidad de los datos y permitiendo una exploración más precisa de las tendencias de precios.

Los resultados obtenidos muestran que los precios de la gasolina y el gasóleo presentan una distribución relativamente homogénea, con una variación limitada entre estaciones dentro de una misma provincia. Sin embargo, se han identificado algunas estaciones con precios significativamente más altos o más bajos, lo que sugiere la posible influencia de factores como:

- La competencia local entre estaciones.
- La ubicación estratégica de la estación de servicio.
- Diferencias en la política de precios entre marcas comerciales.

El análisis de correlación entre los precios de la gasolina 95 y el gasóleo A indica que existe una fuerte relación entre ambos, lo que implica que cualquier variación en uno de estos combustibles tiende a reflejarse en el otro. Este comportamiento es esperado, dado que ambos carburantes están influenciados por los mismos factores macroeconómicos, como:

- El precio del crudo en los mercados internacionales.
- Los costos de refinación y transporte.
- Los impuestos y regulaciones gubernamentales.

El uso de la herramienta AutoViz ha sido fundamental en este análisis, ya que ha permitido generar un informe detallado sobre la estructura y calidad de los datos de forma rápida y eficiente. Gracias a esta herramienta, fue posible:

- Detectar problemas como valores nulos y duplicados.
- Identificar la necesidad de convertir variables categóricas a numéricas.
- Explorar gráficamente la distribución de los precios y las relaciones entre variables.

En conclusión, este estudio proporciona un primer acercamiento al análisis de precios de carburantes en España, con una metodología reproducible que puede aplicarse a otras regiones o expandirse con datos temporales para detectar patrones de variación en el tiempo.

Las futuras mejoras pueden incluir:

- Modelos predictivos para anticipar cambios en los precios del combustible.
- Análisis geoespacial para entender la distribución de estaciones y su impacto en la competencia de precios.
- Comparación de precios entre diferentes provincias para detectar diferencias regionales en la política de precios.
- Incorporación de datos sobre demanda y consumo para evaluar cómo influyen estos factores en la fijación de precios.

Este análisis representa un punto de partida clave para futuras investigaciones en el sector energético y comercialización de combustibles.