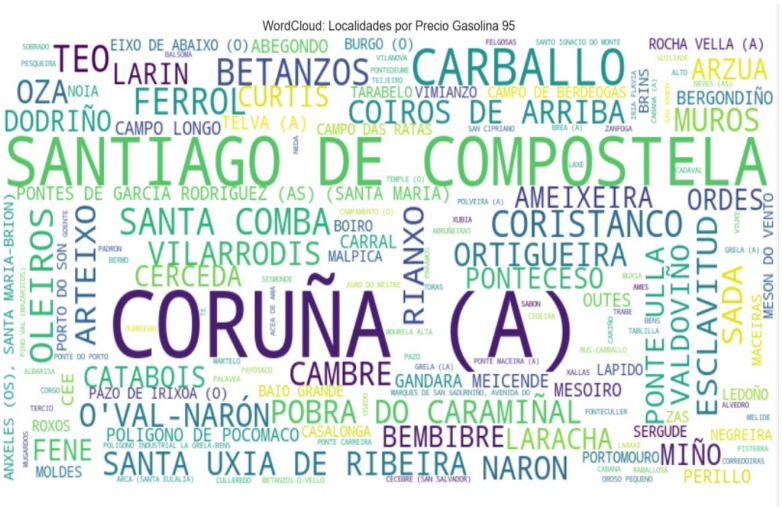


MEMORIA DE ACTIVIDAD: EXPLORACIÓN Y PREPROCESAMIENTO DE DATOS

DESCRIPCIÓN

Actividad centrada en la exploración y preprocesamiento de datos mediante el uso de Python, utilizando bibliotecas como Pandas, Matplotlib y AutoViz.

Carlos Gálvez Reguera

DESARROLLO DE LA ACTIVIDAD

El presente informe tiene como objetivo documentar el proceso de análisis exploratorio de datos (EDA) realizado sobre el conjunto de datos de precios de carburantes en estaciones de servicio de España. En concreto, se ha filtrado y trabajado con la información correspondiente a la provincia de A Coruña.

Carga, limpieza y transformación del dataset.

El dataset fue limpiado corrigiendo errores de codificación, convirtiendo precios a formato numérico y eliminando nulos y duplicados. Se aplicó normalización y discretización de precios. Todo el proceso fue automatizado para reutilizarlo con otras provincias.

Análisis exploratorio y representación gráfica: Histograma comparativo

La distribución de frecuencias de los precios de **Gasolina 95** y **Gasóleo A** en la provincia de **A Coruña**. Ambos combustibles presentan una concentración elevada de precios en el rango de **1,50 € a 1,60 €**, lo que coincide con los valores centrales observados en las estadísticas descriptivas (media y mediana).

Se observa una **asimetría leve hacia la izquierda** en la Gasolina 95, lo que indica una mayor concentración de precios altos. En cambio, el Gasóleo A presenta una **distribución más extendida hacia precios bajos**, con una mayor dispersión general, como se refleja en su **mayor desviación estándar (0,1025 frente a 0,0781)**.

Medidas de Dispersión:

Precio_gasolina_95:
- Rango: 0.32 €
- Varianza: 0.0061
- Desviación Estándar: 0.0781

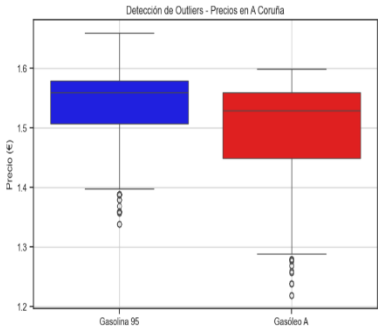
Precio_gasoleo_A:
- Rango: 0.38 €
- Varianza: 0.0105
- Desviación Estándar: 0.1025

Estadísticas Descriptivas:

	Precio_gasolina_95	Precio_gasoleo_A
count	276.000000	276.000000
mean	1.533130	1.485880
std	0.078061	0.102531
min	1.339000	1.219000
25%	1.506500	1.449000
50%	1.559000	1.529000
75%	1.579000	1.559000
max	1.659000	1.599000

Este histograma permite visualizar claramente cómo se agrupan los precios en torno a sus respectivas medianas y facilita la comparación directa entre ambos carburantes. La diferencia en la dispersión sugiere que el **mercado del gasóleo presenta más variabilidad**.

Boxplot comparativo



Ambos carburantes presentan *outliers* visibles como **puntos fuera del rango inferior**, lo que indica la existencia de estaciones que ofrecen precios **notablemente más bajos** que la media del conjunto. En el caso del Gasóleo A, los outliers son más pronunciados, extendiéndose incluso por debajo de los **1,25 €**, lo que refuerza la mayor dispersión observada en el histograma.

La **Gasolina 95** muestra una distribución más compacta, con una mediana ligeramente superior a la del gasóleo. La mayor parte de sus precios se agrupan entre **1,51 € y 1,58 €**, mientras que el Gasóleo A presenta un rango más amplio de precios válidos, entre **1,44 € y 1,57 €** aproximadamente.

Por lo que, aunque ambos carburantes siguen una distribución bastante simétrica, el **Gasóleo A muestra una mayor variabilidad y presencia de precios extremos** en el mercado local.

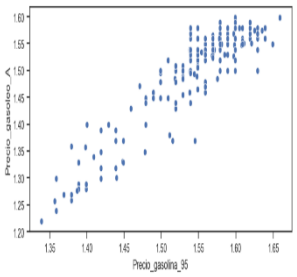
Informe generado Autoviz

El dataset analizado consta de **276 registros y 10 columnas**, clasificadas automáticamente por tipo.

Clasificación de Variables

Numéricas puras reales	*Ninguna identificada inicialmente debido al formato de texto en los precios*	*Se tuvieron que convertir manualmente los precios a tipo float64*
Categórica entera	1 columna (Código postal)	Representa códigos administrativos, no cantidades continuas
Categórica de tipo String	3 Columnas (Provincia,Municipio,Localidad)	Categorías geográficas, algunas con alta cardinalidad
Con valores de texto discreto	2 Columnas(Tipo_venta, Tipo_servicio)	Baja cardinalidad; adecuadas para codificación (una de ellas con valores faltantes)
Detectadas como NLP(texto libre o nombres únicos)	4 Columnas(Direccion, Rotulo, Municipio, Localidad)	Alta cardinalidad o nombres únicos; pueden requerir codificación hash o técnicas de embedding

Gráfico de dispersión entre precios

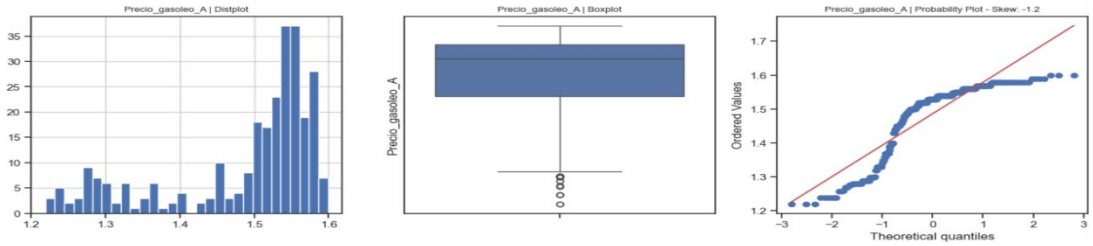


El gráfico muestra una **clara correlación positiva** entre los precios de gasolina 95 y gasóleo A, confirmando lo detectado por AutoViz. Este patrón sugiere que ambos están influenciados por factores comunes. Además, permite identificar **posibles agrupamientos**, útiles en futuros

análisis de segmentación o clustering. La visualización complementa el resto de gráficos y refuerza la calidad del análisis exploratorio realizado.

Visualización detallada de Precio del Gasóleo A

Histograma (Distplot): Muestra una distribución **asimétrica hacia la izquierda**, con

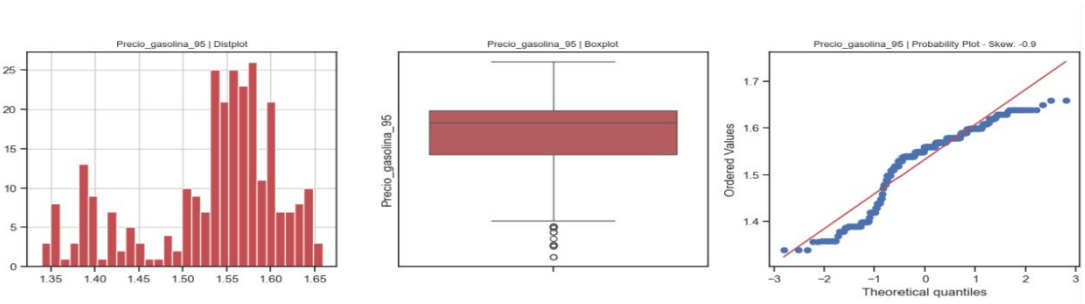


una alta concentración de precios en el intervalo entre **1,50 € y 1,60 €**.

Boxplot: Refuerza la presencia de **valores atípicos en el extremo inferior**, confirmando lo previamente detectado en la exploración inicial. La mayoría de los precios se concentran en torno a la mediana, entre aproximadamente **1,45 € y 1,56€**.

Q-Q Plot: El gráfico de cuantiles teóricos (Q-Q plot) indica una **desviación significativa respecto a la distribución normal**, especialmente en los valores más bajos. Esto sugiere que esta variable no sigue una distribución normal, lo cual es un aspecto relevante si se aplican métodos estadísticos que asumen normalidad.

Visualización detallada de Precio del Gasolina 95



Histograma (Distplot): Se observa una **alta frecuencia de precios entre 1,54 € y 1,60 €**, aunque con presencia de múltiples picos y cierto grado de dispersión hacia valores inferiores. La asimetría, aunque menos marcada que en el gasóleo, sigue presente.

Boxplot: Identifica **outliers en el extremo inferior** con valores por debajo de 1,40 €.

Q-Q Plot: La curva muestra una **asimetría negativa moderada** con una ligera desviación respecto a la línea teórica de normalidad, especialmente en los valores

más bajos. Aunque la distribución es algo más estable que la del gasóleo, sigue sin ajustarse completamente a una distribución normal.

Distribución normalizada de variables categóricas

Tipo_venta: Esta variable presenta **una única categoría dominante**, lo que indica **ausencia de variabilidad**. Como se había identificado previamente en el análisis de calidad de datos, esta columna carece de utilidad para análisis estadístico o modelado y puede ser descartada. **Tipo_servicio:** Muestra mayor variabilidad, aunque con una fuerte concentración en dos categorías principales (P y A), y presencia de valores nulos (nan). Esta variable podría tener cierto valor informativo en modelos predictivos si se gestiona adecuadamente la imputación de los valores faltantes y se codifican sus categorías.

Distribución normalizada de Municipio y Rótulo

Municipio: Destacan A Coruña (A), Santiago de Compostela y Arteixo como los municipios con más estaciones. La alta cantidad de categorías sugiere agrupar o codificar esta variable antes de modelar.

Rotulo: Se observa una fuerte concentración en marcas como REPSOL, CEPSA y GALP, seguidas por una larga cola de marcas minoritarias. Esta dispersión indica alta cardinalidad, lo que puede requerir codificación por frecuencia o agrupación.

Distribución Código Postal: Se observa que algunos códigos, como 15008, 15100 y 15145, concentran más estaciones.

Violin plot de variables Continuas: Ambas distribuciones son similares, aunque se observa una mayor concentración de valores centrales en el gasóleo. La forma achatada de las colas indica menor densidad en los extremos, mientras que el centro refleja la moda. Esta visualización refuerza lo detectado previamente: ambas variables presentan asimetría leve y una distribución no perfectamente normal

Heatmap de Correlación(Pearson)

El resultado más relevante es la **correlación muy alta (0.94)** entre Precio_gasolina_95 y Precio_gasoleo_A. Esta fuerte relación sugiere que **ambos carburantes se comportan de forma paralela en el mercado**, probablemente influenciados por factores comunes como costes logísticos, política fiscal y precios internacionales del crudo. Este nivel de correlación indica que, en un modelo predictivo, podría ser redundante incluir ambas variables simultáneamente, salvo que se busque comparar sus efectos o analizar su diferencia.

En cambio, Codigo_postal muestra **correlaciones bajas** con los precios (0.28 y 0.27), lo cual es coherente con su naturaleza geográfica discreta. Si bien puede tener cierta utilidad para segmentar datos o detectar variabilidad territorial, su capacidad predictiva directa sobre el precio es limitada.

Este heatmap consolida las conclusiones obtenidas en el análisis individual: **la estructura interna del conjunto de datos está altamente condicionada por la relación entre los dos tipos de carburante**, lo que ofrece una base sólida para aplicar técnicas de reducción de dimensionalidad (como PCA) o diseñar modelos con enfoque multivariante.

Conclusiones y comparación de precios

Los precios de gasolina 95 y gasóleo A presentan una distribución similar, con ligeros sesgos negativos y presencia de outliers. La gasolina tiende a tener precios algo más altos y estables, mientras que el gasóleo muestra mayor variabilidad. Existe una correlación muy alta entre ambos combustibles, lo que indica que responden a factores comunes del mercado.

Propuesta de problema de aprendizaje automático

A partir de este dataset, se podría plantear un problema de regresión supervisada para predecir el precio de un carburante (por ejemplo, gasolina 95) a partir de variables como municipio, rotulo, tipo de servicio y código postal. Otra opción sería un modelo de clasificación que, dadas las características de una estación, clasifique el precio como bajo, medio o alto.