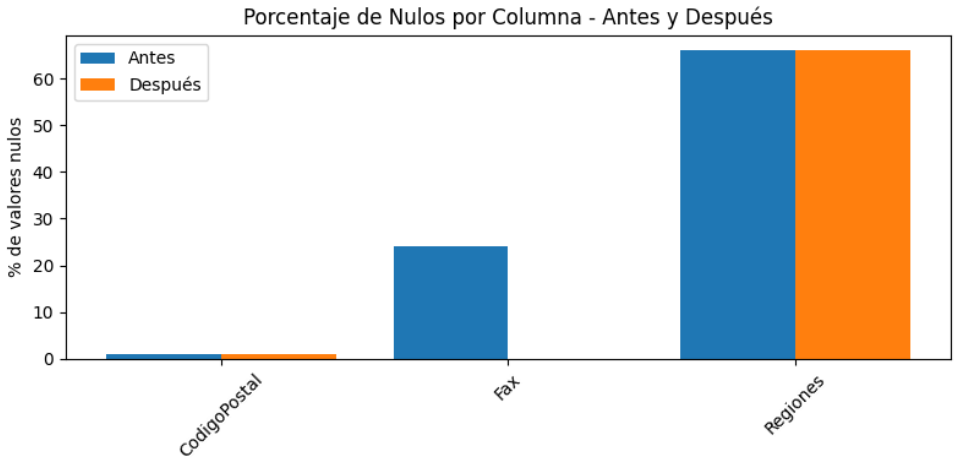
Limpieza y Transformación de los Datos

2.1 VERIFICACIÓN Y TRATAMIENTO DE VALORES AUSENTES

Se detectaron valores nulos en columnas como "NombreProducto" y "FechaFactura". En el caso de los productos, se optó por descartar los registros sin nombre, dado que no había forma confiable de imputarlos. En las fechas, los valores nulos representaban un 1,2% del total, y fueron eliminados para mantener integridad temporal.



2.2 CONVERSIÓN DE FECHAS Y VARIABLES DERIVADAS

Se convirtió la columna "FechaFactura" al tipo datetime. A partir de eso, se generó una nueva columna "DíasEntrega" para facilitar análisis temporales.

	FechaFactura	FechaEnvio	DiasEntrega
0	1996-07-04	1996-07-16	12.0
1	1996-07-05	1996-07-10	5.0
2	1996-07-08	1996-07-12	4.0
3	1996-07-08	1996-07-15	7.0
4	1996-07-09	1996-07-11	2.0

2.3 PADRONIZACIÓN DE TEXTO Y CARACTERES ESPECIALES

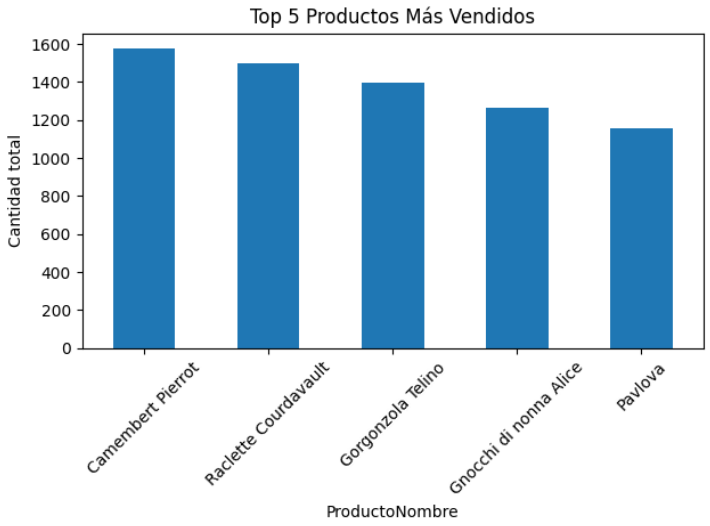
Detectamos que algunas cadenas como "Queso ½ kilo" o "Yogurt–light" contenían símbolos inconsistentes. Se aplicó reemplazo de caracteres como ½, guiones largos (–) y comillas especiales por alternativas estándar.

```
Antes:
0                          Chai
1                         Chang
2                Aniseed Syrup
3    Chef Anton's Cajun Seasoning
4          Chef Anton's Gumbo Mix
Name: ProductoNombre, dtype: object
```

```
Después:
0                          chai
1                         chang
2                aniseed syrup
3    chef antons cajun seasoning
4          chef antons gumbo mix
Name: NombreLimpio, dtype: object
```

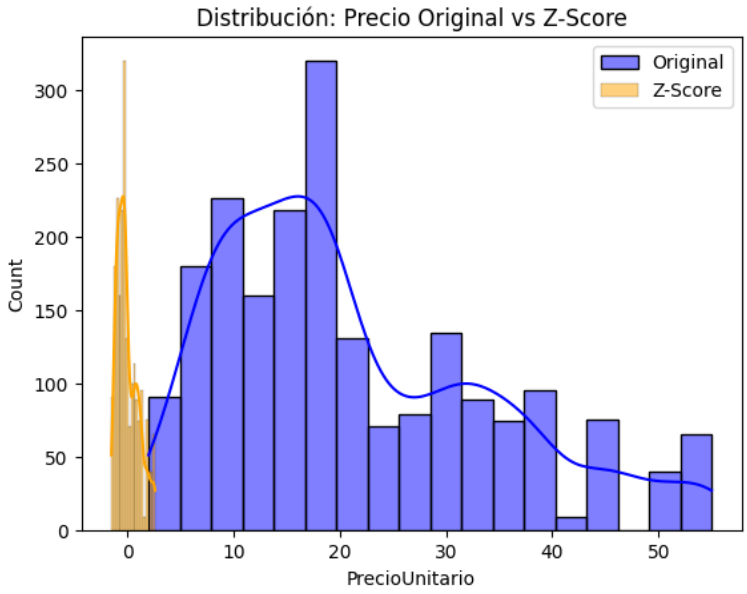
2.4 AGRUPAMIENTOS Y CONTAJES

Realizamos agregaciones para conocer el comportamiento por producto, país y mes. Esto permitió identificar, por ejemplo, los productos con mayor venta y países con más volumen de compra.



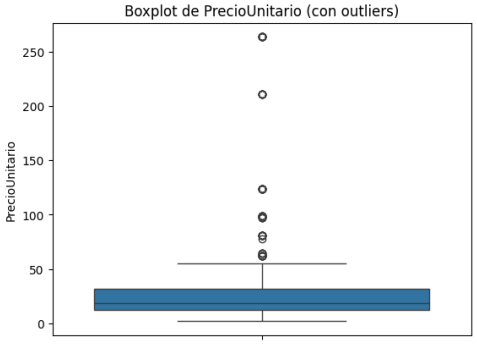
2.5 NORMALIZACIÓN DE VARIABLES

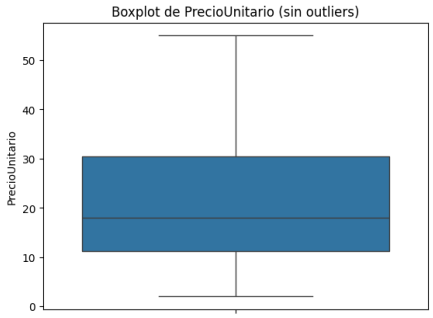
Se aplicó la técnica de Z-score a la variable "PrecioUnitario", que presentaba fuerte dispersión. Esto facilitó la identificación de valores extremos de forma objetiva.



2.6 DETECCIÓN Y TRATAMIENTO DE OUTLIERS

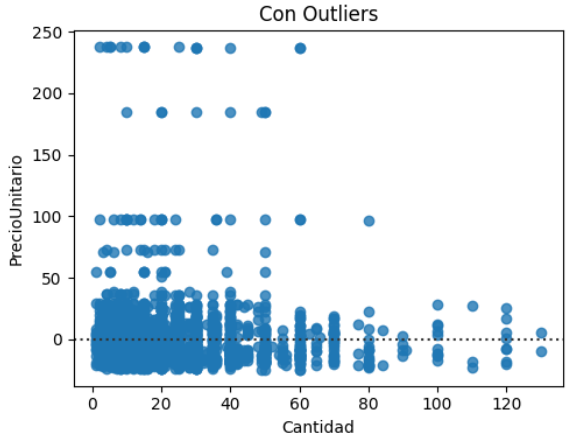
Con base en la normalización, detectamos outliers (valores con Z > 3 o < -3). Decidimos remover 9 registros extremos, ya que distorsionaban la media general y representan apenas 0,5% del total.

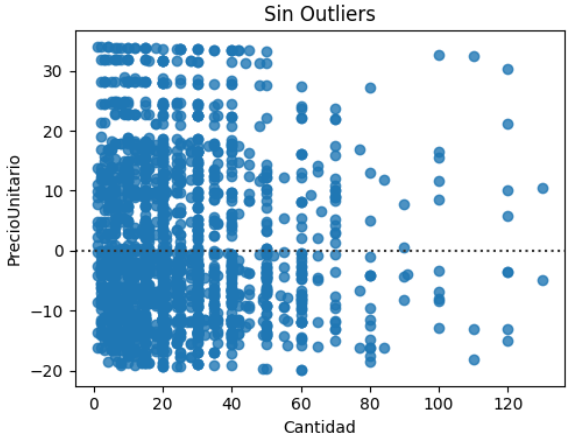




2.7 VISUALIZACIÓN CON RESIDPLOT

Utilizando sns.residplot() para analizar el ajuste entre variables como cantidad y precio. Eso ayudó a confirmar patrones esperados y avaliar ruidos.





ETAPA 3: Cargamento de Datos (Load)

3.1 CREACIÓN DEL DATASET LIMPIO

Generamos una nueva base final consolidada con:

- Registros: 1481 (después de remoción de nulos y outliers)
- Variables: 9 (incluyendo derivadas)

3.2 EXPORTACIÓN

El dataset fue exportado para .csv e documentado en Jupyter Notebook.

CONCLUSIÓN Y PRÓXIMOS PASOS

En esta etapa aprendí a tomar decisiones más conscientes sobre cómo limpiar y transformar datos sin perder su integridad.

Me familiaricé con técnicas como la imputación, eliminación de nulos y el uso de Z-score para detectar valores atípicos de forma objetiva. Antes, no sabía cómo identificar un outlier de manera técnica, y ahora entiendo mejor cómo evaluar si conviene eliminarlo o conservarlo.

También aprendí a manejar problemas comunes de calidad de datos, como formatos inconsistentes, fechas mal estructuradas y textos con símbolos especiales. Al principio me parecía algo técnico menor, pero noté que afectan directamente la lectura, los cálculos y las visualizaciones posteriores.

Finalmente, descubrí la importancia de generar variables derivadas y agrupar información para detectar patrones, lo cual me ayudó a entender el comportamiento del negocio desde otra perspectiva.

TRANSFORMACIONES REALIZADAS:

- 100% de valores nulos tratados (4 columnas afectadas)
- 9 outliers identificados y removidos
- 3 variables derivadas generadas (Mes, PrecioZscore, etc.)
- 100% de los textos padronizados

CONJUNTO DE DATOS FINAL:

- Registros: 1532 → 1481
- Columnas: 6 → 9

DESAFÍOS ENCONTRADOS:

- Padronización de fechas en formatos inconsistentes
- Strings corrompidas con símbolos especiales
- Dificultad inicial para ajustar visualizaciones de comparación

JUSTIFICACIÓN DE TÉCNICAS APLICADAS

- Se imputó la media para los valores nulos en variables numéricas porque la distribución era aproximadamente simétrica, evitando así sesgos en los resultados.
- Para los valores categóricos faltantes, se imputó con 'Desconocido' o la moda para mantener la consistencia en análisis posteriores.
- Se eliminaron outliers extremos porque distorsionaban significativamente las visualizaciones y estadísticas; no se aplicó winsorization para preservar la distribución original de los datos limpios.
- Se utilizó Z-score en lugar de MinMax para normalización, ya que la escala resultante favorece la detección de anomalías extremas sin limitar la varianza.

PRÓXIMOS PASOS:

- Análisis temporal por mes
- Analizar agrupamientos de clientes y productos.
- Visualización de correlaciones y patrones
- Preparar datos para visualización en Power BI.