# Análisis de patrones de consumo y factores socioeconómicos en una plataforma web de comercio electrónico en Brasil

**Introducción**

### El presente estudio tiene como propósito examinar el comportamiento de consumo de los clientes de comercio electrónico en Brasil, utilizando para ello datos transaccionales en conjunto con variables socioeconómicas.

### En particular, se pretende explorar las posibles relaciones entre indicadores como el Índice de Desarrollo Humano (IDH), los niveles de pobreza y la densidad poblacional, con distintas dimensiones del consumo, tales como las categorías de productos adquiridos, los métodos de pago empleados, la utilización de cuotas, los tiempos de entrega y el gasto destinado a envíos. La motivación principal de esta investigación radica en la identificación de patrones de consumo diferenciados en función de las condiciones socioeconómicas, aportando así evidencia empírica de relevancia tanto para el ámbito académico como para la formulación de estrategias empresariales en el sector del comercio electrónico.

### Hipótesis

El análisis se fundamenta en la premisa de que las decisiones de compra realizadas en plataformas en línea no dependen exclusivamente de factores individuales, sino que también están condicionadas por el contexto socioeconómico regional. En este marco, se plantean las siguientes hipótesis:

1. **En estados con mayor IDH y menores niveles de pobreza**: los consumidores realizan compras más frecuentes y diversificadas en categorías, aunque con tickets promedios más bajos, lo que refleja un consumo regular y masificado.
2. **En regiones con bajo IDH y alta pobreza:** las compras se concentran en menos categorías, con tickets más elevados y mayor uso de pagos en cuotas, lo que evidencia restricciones financieras.
3. **La elección de métodos de pago varía según el nivel socioeconómico**: predominando las tarjetas de crédito en los estados más desarrollados.
4. **Factores logísticos, como el costo del flete y los tiempos de entrega**: estos impactan de manera diferenciada en la decisión de compra según el perfil socioeconómico del consumidor.
5. **La densidad poblacional y las características demográficas**: se espera que presenten correlaciones positivas con el Índice de Desarrollo Humano (IDH) y negativas con los niveles de pobreza, reforzando así la influencia del entorno social y geográfico en la dinámica del comercio electrónico.

# Tratamiento de datos

**Integración de fuentes:** El análisis se construyó sobre una base integrada a partir de múltiples archivos con información de pedidos, pagos, productos, vendedores y variables socioeconómicas por estado. Se consolidó un conjunto de datos denominado *“Merge”*, compuesto por registros de transacciones en una plataforma de comercio electrónico.

**Calidad y pre-procesamiento:** En primera instancia se verificaron los tipos de datos, la presencia de valores nulos y la consistencia temporal entre las fechas de compra, pago y entrega. Luego se aplicaron procedimientos de limpieza, normalización y **escalamiento** para asegurar comparabilidad entre variables y evitar sesgos por diferencias de magnitud.

**Enriquecimiento externo:** Se incorporaron indicadores socioeconómicos a nivel estatal como subíndices de riqueza, salud y educación del IDH (2017), además de tasas de pobreza y densidad demográfica. Esto se realizó mediante uniones *“Merge”* por clave de estado.

**Segmentación (K-Means):** Una etapa clave fue la **clusterización** de los estados según sus valores promedio de IDH y, en paralelo, sus tasas de pobreza. Previa **estandarización** (con StandardScaler) se evaluaron distintas particiones usando la **inercia** y el **coeficiente de silhouette**. Aunque las métricas sugerían K=5K=5K=5, se optó por K=3K=3K=3 por su mayor interpretabilidad (grupos **bajo**, **medio** y **alto**) tanto para IDH como para pobreza.

**Tratamiento de atípicos:** Para las distribuciones de **ticket promedio** por categoría y por clúster se eliminaron outliers mediante el criterio del **rango intercuartílico (IQR)**, conservando observaciones representativas y evitando distorsiones en análisis posteriores.

**Correlaciones:** Se estudió la relación entre **densidad de población**, **pobreza** e **IDH** utilizando el coeficiente de **Spearman**, apropiado para dependencias monótonas y variables potencialmente no normales.

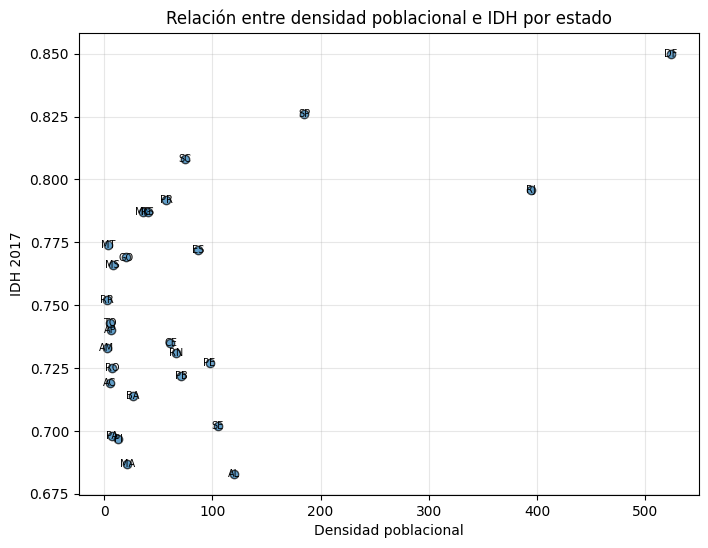
**Distribuciones de costos y precios:** Se evaluó a qué distribuciones se asemejaban los **costos de envío** y el **precio de producto**. Los datos mostraron un patrón **lognormal** (log-distribución), lo que permitiría estimar rangos esperados (intervalos centrales 95% mediante cuantiles 2.5–97.5), identificar valores extremos y anticipar **costos logísticos** a futuros.

## Análisis de datos

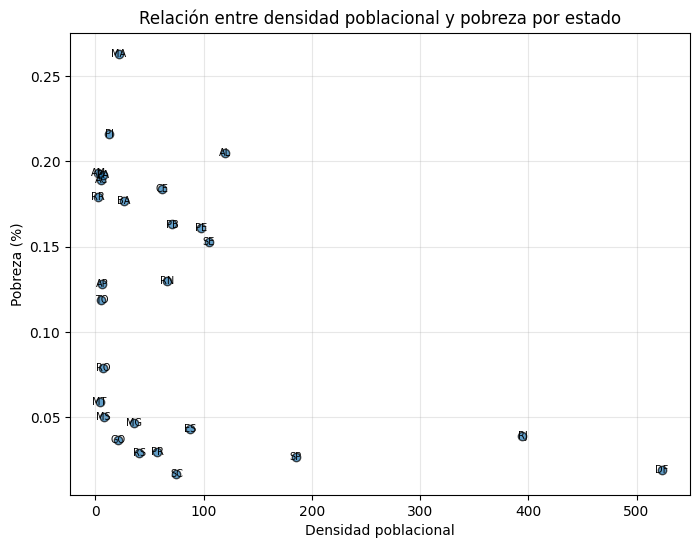
## El análisis se estructuró en diversas dimensiones con el propósito de ofrecer una comprensión integral del fenómeno del comercio electrónico en la plataforma.

## En primer término, se identificó una correlación negativa elevada entre el Índice de Desarrollo Humano (IDH) y la pobreza a nivel estatal (–0.88), lo que respalda la premisa inicial de que ambos indicadores operan como variables espejo. Asimismo, el cruce de los clústeres reveló que los estados con mayores niveles de IDH se concentraron en el grupo caracterizado por baja pobreza, mientras que aquellos con menores valores de IDH se ubicaron predominantemente en el grupo de alta pobreza, lo cual confirma la coherencia y solidez de las segmentaciones realizadas.

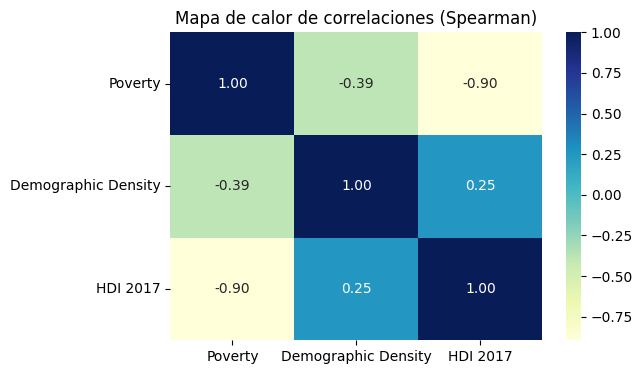
## Relación entre densidad de población e IDH/ pobreza



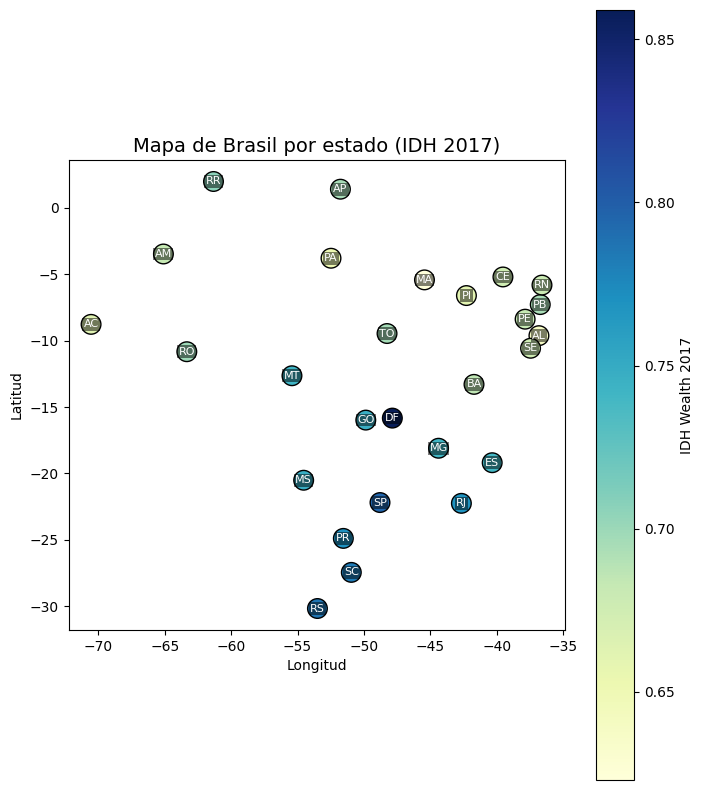
* Relación positiva entre la densidad poblacional e IDH, donde los estados con mayor densidad tienden a presentar un IDH más alto.



* Relación inversa entre pobreza y densidad poblacional, indicando que los estados más densamente poblados tienden a tener menores tasas de pobreza.



* Mapa de calor de correlaciones que confirma la fuerte relación negativa entre pobreza e IDH (-0.90) y una correlación moderada positiva entre densidad demográfica e IDH (0.25).



* Distribución geográfica del IDH de riqueza por estado en Brasil, evidenciando mayores niveles en regiones del sur y sureste.

## ****Análisis: Cluster por IDH****

### **Distribución del ticket promedio por cluster IDH**:****

| **Cluster** | **Ticket CON outliers** | **Ticket SIN outliers** |  |
| --- | --- | --- | --- |
| Alto | 151.98 | 146.79 |  |
| Medio | 202.41 | 178.04 |  |
| Bajo | 190.06 | 182.50 |  |

#### Observaciones:

* El clúster intermedio presenta el ticket promedio más elevado, comportamiento que se mantiene tanto al incluir la totalidad de los datos como tras la exclusión de outliers.
* El **cluster alto** tiene **el ticket más bajo**, a pesar de tener mayor desarrollo humano.
* La **eliminación de *outliers*** impacta fuertemente al cluster medio, lo que sugiere que había productos con precios muy elevados que distorsionaban el promedio.

**Top productos en IDH medio (sin *outliers*)**:

* Gifts, thers, electronics, health\_beauty y home.
* Ticket promedio en *gifts*: **246.4**
* Ticket promedio en *others*: **234.7**

Esto indica que en estados con IDH medio, hay consumo de productos más caros en ciertas categorías específicas.

## ****Análisis: Cluster por Pobreza****

### Se procedió a agrupar los estados según su nivel de pobreza con el fin de examinar las diferencias en el comportamiento de consumo, específicamente en términos de ticket promedio.

### ****Distribución del ticket promedio por cluster de pobreza:****

| **Cluster** | **Ticket CON outliers** | **Ticket SIN outliers** |  |
| --- | --- | --- | --- |
| Alto | 191.82 | 191.82 |  |
| Medio | 203.68 | 174.55 |  |
| Bajo | 152.14 | 146.92 |  |

#### Observaciones:

* El cluster **medio** nuevamente presenta el **ticket más alto**, pero se ve fuertemente afectado por los *outliers*.
* El cluster **bajo** que, representa mayor pobreza, tiene el ticket más bajo.
* El cluster **alto que,** en este caso, representa menor pobreza y tiene un ticket medio.

### ****Comparación IDH vs Pobreza****

Los análisis previamente mencionados evidencian un patrón no lineal, en el que el grupo “medio” (correspondiente a regiones ni muy pobres ni altamente desarrolladas) presenta el ticket promedio más elevado.   
Esta tendencia podría explicarse por un mayor equilibrio entre poder adquisitivo y deseo de consumo en dichas regiones, así como por la existencia de menores barreras para acceder a productos de mayor valor.

# Categorías con Outliers por Cluster IDH

Las siguientes categorías presentaron valores atípicos de ticket promedio dentro de cada clúster de IDH.

|  |  |  |  |
| --- | --- | --- | --- |
| Cluster IDH | Categoría | N° Outliers | Valor Total (Reales) |
| alto | construction | 1 | 320058.06 |
| alto | food | 1 | 82924.68 |
| alto | gifts | 1 | 1073841.63 |
| bajo | services | 1 | 13654.99 |
| medio | construction | 1 | 25737.61 |
| medio | services | 1 | 16061.26 |

**Observaciones :**

* La categoría 'gifts' en el cluster alto destaca con un valor total superior al millón de Reales.
* 'Construction' aparece como *outlier* en los tres niveles socioeconómicos, mostrando alta variabilidad de precios.
* 'Services' también presenta comportamientos atípicos en clústeres medio y bajo.

# 

# Categorías con Outliers por Cluster de Pobreza

A continuación, se presentan las categorías que mostraron comportamientos atípicos en cada grupo según el nivel de pobreza estatal.

|  |  |  |  |
| --- | --- | --- | --- |
| Cluster Pobreza | Categoría | N° Outliers | Valor Total (Reales) |
| bajo | construction | 1 | 322228.52 |
| bajo | food | 1 | 83051.13 |
| bajo | gifts | 1 | 1078851.81 |
| medio | construction | 1 | 16671.7 |
| medio | services | 1 | 14035.26 |

**Observaciones:**

* Al igual que en IDH, 'gifts' en el cluster de baja pobreza representa un valor extremo significativo.
* 'Construction' vuelve a figurar como categoría con fuerte dispersión de precios.
* La categoría 'food' sólo es outlier según el análisis de pobreza, no de IDH, lo que sugiere una particularidad

### Analisis de la distribucion de los metodos de pago

Los datos muestran que:

* **Las tarjetas de crédito** son el medio de pago dominante en todos los clusters, representando entre **74% y 78%** de las transacciones.
* **Los UPI** (pagos inmediatos digitales) aparece en segundo lugar, con una participación entre **16% y 19%**.
* **Los Vouchers** se ubican en torno al **2%–5%**, mientras que las **tarjetas de débito** apenas superan el **1%** de las transacciones.
* En términos de valor total, la tarjeta de crédito concentra aún más peso (**78%–80%**), lo cual indica que no solo es más frecuente, sino que también se asocia a compras de mayor importe promedio.

Existe **homogeneidad** en los métodos de pago predominantes (tarjeta de crédito y UPI) en todos los segmentos, lo que evidencia la estandarización de la plataforma de pago en la web.

Estos resultados sugieren que deben mantenerse optimizadas las opciones de **tarjeta de crédito** y **UPI**, ya que son los métodos clave.

### Distribución de cuotas por cluster de IDH y de pobreza

Se observa una **preferencia generalizada por el pago en una sola cuota**, especialmente en los clusters con mejores indicadores:

* En el cluster **alto (IDH)**, el **51%** de las transacciones se realizaron en **1 sola cuota**, seguido por un 22% en **2-3 cuotas**, y solo un 12% en **7 o más cuotas**.
* En el cluster **medio (IDH)**, el porcentaje de pagos en una sola cuota baja a **40%**, y el uso de **7+ cuotas sube a 17%**, indicando mayor uso del financiamiento.
* A nivel de **pobreza**, el patrón es similar: los grupos **con mayor pobreza** tienen **mayor uso de cuotas largas** (15% en 7+), frente a 11% en los clusters menos pobres.

Esto sugiere que **los consumidores en condiciones más vulnerables tienden a financiar sus compras en más cuotas**, probablemente como estrategia de acceso a productos de mayor valor o para distribuir el gasto.

### Productos con mayor utilización de financiamiento por número de cuotas:

Al analizar los **productos más comprados con 7 o más cuotas**, se encuentran patrones comunes en todos los clusters:

* **Home, Health-beauty y Gifts** son las categorías predominantes en pagos extensos, tanto en IDH como en pobreza.
* En el **cluster alto (IDH)**, por ejemplo, la categoría Home lidera con más de 4.000 compras en 7+ cuotas, mientras que Gifts alcanza los tickets promedio más altos (~$397).
* En contraste, con **1 cuota**, las categorías más compradas son Home*,* Electronics *y* Health-beauty*.* Reflejando compras de menor valor unitario o mayor accesibilidad.

Este patrón indica que **las categorías de bienes durables o de regalo tienden a ser financiadas en cuotas largas**, mientras que **productos de consumo cotidiano o de menor valor se pagan de contado.**

**Estimación de la probabilidad de pagos superiores a seis cuotas utilizando un modelo de Poisson:**

Se utilizó una distribución de Poisson para modelar la probabilidad de pagar en **más de 6 cuotas,** por categoría y cluster:

* En todos los clusters, las categorías con mayor probabilidad de uso extensivo de cuotas fueron G**ifts, Services y Health-beauty**, alcanzando probabilidades de hasta **18% en el cluster medio de IDH**.
* En el cluster **medio de pobreza** Services *y* Gifts muestran las probabilidades más altas: **12.5% y 17.6% respectivamente**.
* El clúster de alto IDH presenta, en general, menores probabilidades de financiar compras a través de cuotas. Por ejemplo, la probabilidad de optar por financiación en la categoría Gifts es solo del 7,2%, mientras que en la mayoría de las demás categorías se mantiene por debajo del 5%.

Estos datos refuerzan que los **clusters con menos recursos tienen una mayor dependencia del financiamiento en cuotas**, y que ciertos productos específicos tienden a requerir más cuotas para ser adquiridos.

### Distribución modal de cuotas según categoría y cluster:

El análisis de la **moda de cuotas** para pagos con tarjeta muestra patrones claros:

* En prácticamente **todas las categorías y clusters,** la **cuota más frecuente es 1**, especialmente en productos de menor valor o mayor rotación como en Food, Books, Fashion*.*
* En los clusters **altos y medios**, algunas categorías *como* Gifts*,* Services *o* Fashion presentan modas en **2 o 3 cuotas**, indicando más fraccionamiento.
* En el cluster **alto de pobreza**, Gifts destaca con una **moda de 8 cuotas**, lo que es altamente significativo. Aunque esta moda no es la mayoría (18.8%) evidencia una estrategia de financiamiento extendido para compras especiales.

Además, las **frecuencias relativas** muestran que en muchas categorías más del **40% de las compras se hacen en 1 cuota**, confirmando que, aunque se ofrece financiamiento, **la opción predilecta sigue siendo el pago al contado**, probablemente por razones de control financiero o falta de acceso a crédito extendido.

### Análisis de correlación: HDI vs ticket promedio:

Se evaluó la correlación lineal entre el **ticket promedio por estado** y los valores promedio estatales de cada uno de los tres subíndices del HDI.   
  
Los resultados son:

* **HDI Educación vs ticket promedio:** -0.624
* **HDI Salud vs ticket promedio:** -0.682
* **HDI Riqueza vs ticket promedio: -0.713**

Estos valores indican una **correlación negativa moderada a fuerte**: a mayor nivel de desarrollo (particularmente en riqueza), **menor es el ticket promedio**.   
Este hallazgo, aunque pueda parecer contra-intuitivo a primera vista, puede interpretarse bajo las siguientes consideraciones:

* En los estados con mayores niveles de riqueza, los consumidores tienden a fragmentar sus compras o a realizar transacciones más frecuentes, aunque de menor monto individual.
* En cambio, en estados más vulnerables, es posible que los consumidores **concentren su gasto en compras puntuales y de mayor valor**, recurriendo a financiamiento o ahorro previo.

**Correlación entre HDI y frecuencia de pagos en siete o más cuotas :**

Se calculó la proporción de transacciones realizadas en siete o más cuotas por estado y se correlacionó con los distintos subíndices del IDH:

* HDI Educación +7 cuotas**:** –0.585
* HDI Salud +7 cuotas: –0.301
* HDI Riqueza +7 cuotas: –0.457

La correlación negativa se mantiene, aunque con menor intensidad. El subíndice de educación muestra la relación más pronunciada, lo que sugiere que:

* A menor nivel educativo, existe una mayor propensión a utilizar financiamiento en cuotas prolongadas.
* El nivel educativo podría influir en las decisiones financieras de los consumidores, su capacidad de planificación y el acceso a información relativa a intereses y costos asociados a las cuotas.

**Análisis del valor del envío:**

Se creó una nueva variable denominada ratio\_flete\_producto, definida como el cociente entre el **valor del flete** y el **precio del producto**. Esta métrica permite comparar la **proporción relativa del costo de envío** respecto al precio de la compra.

Los resultados por clúster IDH muestran:

| Cluster idh | Gasto envio promedio | Precio promedio | Valor pagado promedio | Ratio flete-producto |
| --- | --- | --- | --- | --- |
| **alto** | 18.55 | 120.64 | 151.71 | **0.29** |
| **bajo** | 31.20 | 148.13 | 194.77 | **0.44** |
| **medio** | 35.25 | 161.09 | 206.35 | **0.46** |

**Interpretación:**

* Aunque los clústeres medio y bajo gastan más en promedio en productos y en valor total pagado, lo más relevante es que el **ratio del costo de envío respecto al precio del producto aumenta considerablemente** a medida que baja el nivel de IDH.
* En el clúster **alto**, el envío representa solo el **29%** del valor del producto, mientras que en el clúster **medio** supera el **46%**.   
  Esto indica que, para consumidores con menor IDH, **el costo del envío representa una carga proporcionalmente más significativa**, lo que podría condicionar su decisión de compra, especialmente para productos de bajo valor unitario.

### Análisis del tiempo de entrega

Se midió la variable tiempo\_entrega\_dias como la diferencia (en días) entre la fecha de compra y la fecha de entrega.   
Se agruparon los resultados por clúster:

| Cluster idh | Prom. dias | Mediana dias | Desv. estandar | Mínimo | Máximo | Nº pedidos |
| --- | --- | --- | --- | --- | --- | --- |
| **Alto** | 11.06 | 9 | 8.60 | 0 | 209 | 90,183 |
| **Bajo** | 20.20 | 18 | 12.57 | 0 | 195 | 6,480 |
| **Medio** | 19.46 | 17 | 12.09 | 1 | 187 | 4,961 |

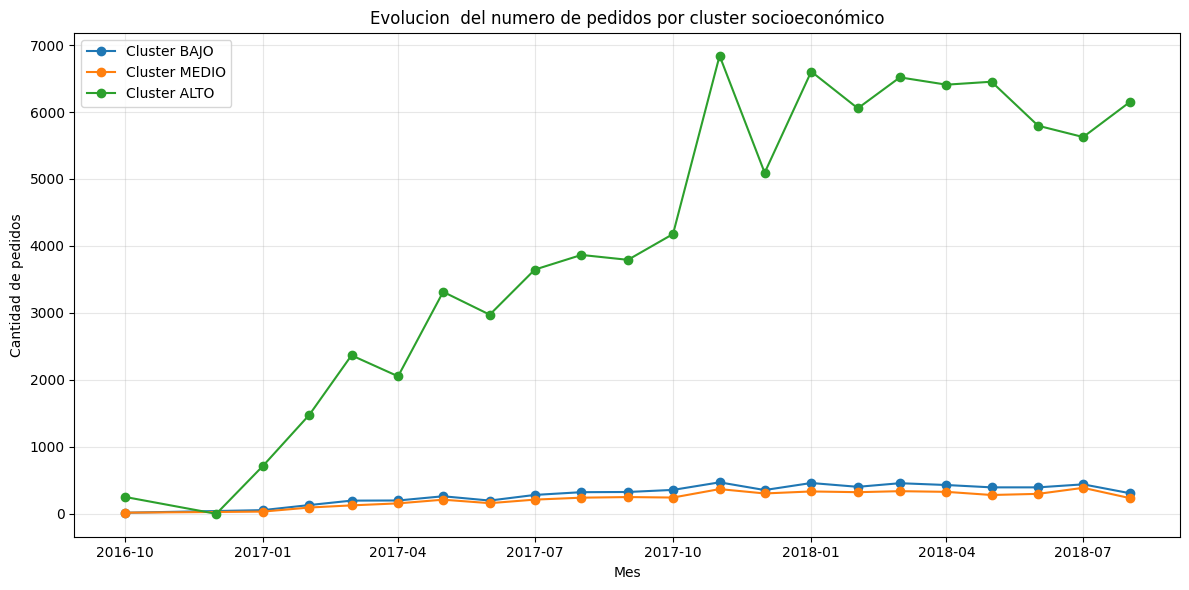
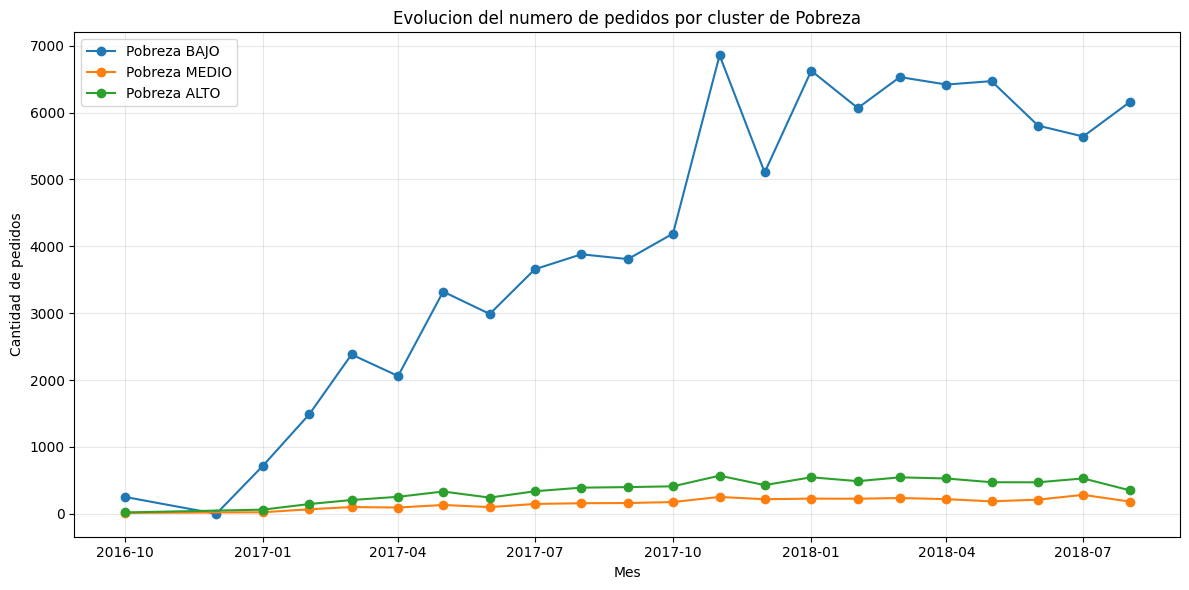
**Interpretación:**

 Los consumidores del clúster alto reciben sus productos casi el doble de rápido que los del clúster bajo, con un promedio de 11 días frente a 20 días.

 La mediana también refleja una diferencia significativa: 9 días frente a 18 días.

 Estas diferencias pueden atribuirse a factores logísticos, tales como la ubicación en zonas urbanas, la mejor infraestructura o la proximidad a los centros de distribución.

 Además, los clústeres bajo y medio no solo enfrentan mayores costos de envío, sino también tiempos de espera más prolongados, lo que constituye una doble penalización en términos de experiencia de compra.

**Análisis de la evolución temporal de la cantidad de pedidos por clúster**

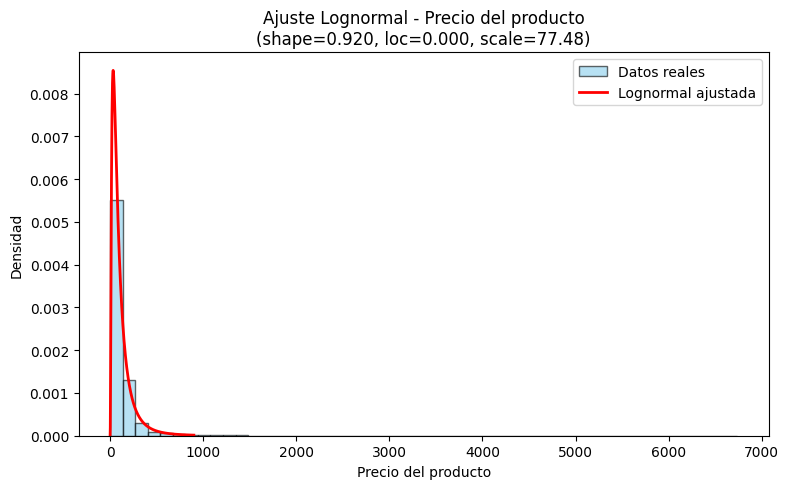
**Comparación entre IDH y Pobreza**

* Ambos enfoques muestran **la misma dinámica de fondo**:
* Donde hay **más recursos** (alto IDH / baja pobreza), hay **mucho más volumen de compras**, aunque con menor ticket promedio.
* Donde hay **menos recursos** (bajo IDH / alta pobreza), hay **menos volumen**, pero tickets más altos.
* Esto refleja diferentes **estrategias de consumo**:  
  + Los segmentos más favorecidos compran de forma **frecuente y diversificada**.
  + Los segmentos más vulnerables concentran sus recursos en **compras más esporádicas, pero de mayor valor unitario.**

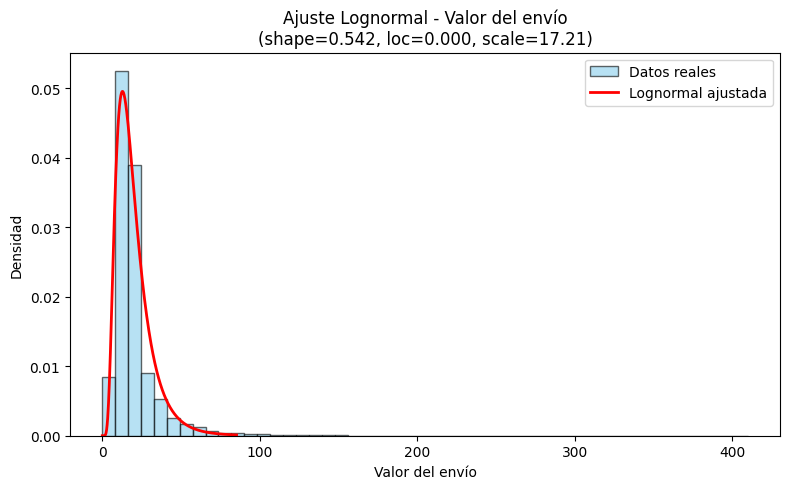
## Evaluación de Distribuciones:

Al aplicar el análisis a las variables 'Precio del producto' y 'Valor del envío', se determinó que la distribución lognormal es la que mejor se ajusta, en ambos casos, al presentar el valor más bajo de AIC entre las distribuciones consideradas. Esto sugiere que tanto el precio como el valor del envío son variables sesgadas a la derecha, con una mayoría de observaciones de bajo valor y una cola larga hacia valores altos.

**Ajuste lognormal del precio del producto**:



**Ajuste lognormal del valor del envío**:



## Interpretación de Resultados:

La gráfica evidencia el ajuste de una distribución lognormal sobre el precio de los productos, mostrando una marcada asimetría positiva, donde la mayoría de los precios se concentran en rangos bajos y una larga cola se extiende hacia valores más elevados. Este comportamiento es característico de plataformas de comercio electrónico, en las que predominan productos económicos, aunque también existen artículos de alto valor que representan casos menos frecuentes.

De manera similar, el valor del envío se ajusta satisfactoriamente a una distribución lognormal: la mayoría de los envíos presentan costos bajos, mientras que una proporción menor de pedidos incurre en costos significativamente mayores.   
La media teórica del valor del envío es de aproximadamente 19,93 unidades monetarias, superándose únicamente en el 39% de los casos, lo que refleja la asimetría de la distribución.

Además, el 95% de los valores de envío se sitúa dentro del rango de 5,95 a 49,79 unidades monetarias, información que resulta útil para establecer umbrales logísticos, definir políticas de subsidio de envío o identificar valores atípicos en el sistema.

## Conclusiones y consideraciones finales:

El estudio confirma que el **nivel socioeconómico regional** influye de manera directa y profunda en el comportamiento de compra dentro de la plataforma de comercio electrónico analizada.

Las hipótesis iniciales se ven respaldadas en gran medida por:

1. **Relación entre IDH y consumo:** Los consumidores de estados con alto IDH presentan hábitos de compra más frecuentes, con tickets moderados y mayor diversificación de categorías. En cambio, los consumidores de bajo IDH concentran su gasto en menos compras, con tickets más elevados en proporción a su poder adquisitivo.
2. **Financiamiento mediante cuotas:** Los clusters de menor desarrollo muestran una mayor propensión a financiar sus compras en cuotas extendidas, lo que denota dependencia del crédito para acceder a bienes de consumo duradero.
3. **Métodos de pago:** La tarjeta de crédito es el medio de pago dominante, aunque en estados con menor desarrollo socioeconómico crece la relevancia de alternativas como UPI o Vouchers.
4. **Logística desigual: El costo del envío es más determinante en consumidores de bajo IDH, no solo en valor absoluto, sino en su proporción respecto al producto. Esto podría limitar el acceso a ciertos bienes o incentivar estrategias de agrupación de compras para optimizar costos. El tiempo de entrega aumenta considerablemente en regiones o clústeres de menor IDH, lo que puede generar percepciones negativas del servicio o reducir la fidelización del cliente. Ambos factores, precio relativo del flete y demora en la entrega, reflejan una brecha estructural en el comercio electrónico según el nivel socioeconómico, lo que sugiere la necesidad de políticas logísticas diferenciadas o subsidios de envío para favorecer la inclusión digital.**
5. **Factores estructurales:** La densidad poblacional se asocia positivamente con un mayor IDH y mayor dinamismo comercial digital, lo cual sugiere que la infraestructura urbana y la concentración demográfica son elementos clave para el desarrollo del comercio electrónico.

Concluyendo, el análisis evidencia que el comercio electrónico en Brasil refleja las desigualdades socioeconómicas del país.   
Los consumidores de regiones más favorecidas acceden a una experiencia de compra más eficiente, económica y diversificada, mientras que los de regiones más vulnerables enfrentan mayores costos, mayores plazos de entrega y una dependencia del crédito que condiciona su acceso a determinados bienes.

Estos hallazgos resultan relevantes no solo para comprender el funcionamiento de la plataforma analizada, sino también para orientar estrategias empresariales y de política pública orientadas a **reducir las brechas en el acceso al comercio digital**, mejorando la inclusión financiera, la infraestructura logística y la equidad en la experiencia de consumo en línea.