Análisis de Datos Financieros y Diseño de Indicadores – Examen 1

Instrucciones:

- Este examen utiliza el método del caso para evaluar tu capacidad de analizar datos, comprender quiénes son los clientes de Gabu, y predecir el valor del tiempo de vida (LTV) de los mismos.
- En cada pregunta, explica claramente tu razonamiento y muestra todos los cálculos necesarios.
- Puedes utilizar Python (Pandas, Matplotlib, Seaborn, statsmodels) para los cálculos y visualizaciones.
- Sé conciso y directo en tus explicaciones.

Estudio de Caso: Gabu

Gabu es una startup que supervisa las sesiones de juego de los niños mientras juegan en línea, proporcionando informes detallados a los padres sobre el tiempo de juego y el tipo de contenido con el que interactúan sus hijos. La empresa busca entender mejor a sus clientes y, sobre todo, predecir el valor de tiempo de vida de cada cliente (LTV) para optimizar sus estrategias de retención y adquisición.

La empresa ha recopilado datos por cliente que incluyen características demográficas y económicas como:

- **ID del Cliente**: Identificador único del cliente.
- **Edad del Niño**: Edad del niño para el que se supervisa el juego.
- **Sexo del Niño**: Hombre o mujer.
- **Nivel de Estudios del Padre**: El nivel de estudios del padre que registro al niño.
- **Duración de la Subscripción**: Tiempo en meses que el cliente ha estado suscrito al servicio.
- **Plan de Subscripción**: El tipo de plan (Básico, Premium, Familiar).
- **Ingresos del Hogar**: Ingresos anuales del hogar del cliente (USD).
- **Gasto Total en Subscripción**: Gasto acumulado del cliente en el servicio.
- **Número de Sesiones por Semana**: Número promedio de sesiones de juego supervisadas por semana.
- **Nivel de Satisfacción del Cliente**: Puntuación de satisfacción del cliente (1 a 5).
- **Churn**: Variable binaria que indica si el cliente ha cancelado el servicio (1 = canceló, 0 = activo).

- **Valor del Tiempo de Vida (LTV)**: Valor total proyectado que Gabu espera obtener de un cliente a lo largo de su relación.

Sección 1: Comprendiendo a los Clientes (20 puntos)

Pregunta 1: Exploración de los Datos de los Clientes
Con base en los datos proporcionados:

1. Segmenta a los clientes en función de sus características demográficas.
2. Analiza las características demográficas y económicas de los segmentos. Explica cómo estos segmentos podrían ayudar a Gabu a diseñar estrategias más efectivas de marketing o producto.

Sección 2: Selección y Cálculo de KPIs (20 puntos)

Pregunta 2: Selección del KPI Correcto
Gabu quiere identificar un KPI clave que les permita medir la retención de clientes y la capacidad de generar ingresos a largo plazo.

A continuación, se te presentan cuatro KPIs posibles. Solo uno de estos es el KPI correcto que Gabu debería usar para medir la retención de clientes y su impacto en los ingresos futuros. Selecciona el KPI correcto, calcúlalo usando los datos proporcionados, y justifica por qué es el más adecuado.

Opciones de KPI:

Opción 1: Tasa de Sesiones por Cliente

- *Definición*: Número promedio de sesiones supervisadas por cliente.
- *Fórmula*: $ \text{Tasa de Sesiones por Cliente} = \frac{\text{Número de Sesiones por Semana}}{\text{Clientes Existentes}}$

Opción 2: Tasa de Retorno de Inversión en Marketing (ROI de Marketing)

- *Definición*: Mide el retorno generado por los gastos de marketing en relación con los ingresos obtenidos.
- *Fórmula*: $\text{ROI de Marketing} = \frac{\text{Ingresos Totales} - \text{Gastos de Marketing}}{\text{Gastos de Marketing}}$

Opción 3: Tasa de Crecimiento de Nuevos Clientes

- *Definición*: El porcentaje de aumento o disminución en la cantidad de nuevos clientes adquiridos.
- *Fórmula*: $\text{Tasa de Crecimiento de Nuevos Clientes} = \frac{\text{Nuevos Clientes}_{\text{mes actual}} - \text{Nuevos Clientes}_{\text{mes anterior}}}{\text{Nuevos Clientes}_{\text{mes anterior}}}$

Opción 4: Tasa de Churn

- *Definición*: El porcentaje de clientes que cancelan el servicio durante un período de tiempo determinado.
- *Fórmula*: $\text{Tasa de Churn} = \frac{\text{Número de Clientes que Cancelaron}}{\text{Número Total de Clientes Activos al Inicio del Periodo}}$

Instrucciones:

1. Selecciona el KPI correcto que mejor mida la retención de clientes.
2. Explica por qué este KPI es el más adecuado para Gabu y cómo ayudará a medir

---

Sección 3: Predicción del Valor del Tiempo de Vida (LTV) (40 puntos)

Pregunta 3: Construcción de un Modelo de Regresión para Predecir LTV
El equipo de Gabu desea predecir el valor del tiempo de vida (LTV) de sus clientes usando las características demográficas y económicas del dataset.

1. **Preparación de los Datos**:
    - Usa las siguientes variables como independientes:
        - **Edad del Niño**
        - **Ingresos del Hogar**
        - **Plan de Subscripción**
        - **Número de Sesiones por Semana**
        - **Nivel de Satisfacción del Cliente**
        - **Segmento del cliente**
    - Usa el **Valor del Tiempo de Vida (LTV)** como la variable dependiente.
2. **Construcción del Modelo**:
    - Ajusta un modelo de regresión lineal utilizando las variables mencionadas.
    - Escribe la ecuación de la regresión (incluye los coeficientes de cada variable).
3. **Interpretación de los Coeficientes**:
    - Explica el significado de los coeficientes obtenidos en el modelo. ¿Cómo afecta cada variable independiente al LTV?

- Identifica los factores que parecen tener el mayor impacto en el valor de tiempo de vida y discute por qué podrían ser importantes para Gabu.

Sección 4: Insights y Estrategia de Retención (20 puntos)

Pregunta 4: Estrategias Basadas en el Análisis
Basándote en el análisis realizado y los resultados del modelo de regresión, responde a las siguientes preguntas:

1. ¿Qué acciones estratégicas recomendarías para mejorar el LTV de los clientes de Gabu?
2. Considera la tasa de churn. ¿Qué medidas concretas podría tomar Gabu para reducir esta tasa y aumentar la lealtad de los clientes?

Criterios de Evaluación:

- **Exploración de Datos** (20 puntos): Profundidad y claridad de los insights obtenidos de la segmentación y el análisis demográfico y económico de los clientes.
- **Identificación de KPIs** (20 puntos): Selección adecuada de KPIs estratégicos y justificación clara de su importancia.
- **Construcción del Modelo de Regresión** (40 puntos): Aplicación correcta de la regresión lineal, interpretación de los coeficientes y calidad de los análisis realizados.
- **Insights y Estrategia de Retención** (20 puntos): Calidad de las recomendaciones estratégicas basadas en el análisis de los datos y los resultados del modelo.

In [2]:
```
import pandas as pd
import numpy as np
from sklearn.cluster import KMeans, DBSCAN
from sklearn.preprocessing import StandardScaler
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn import set_config
import statsmodels.api as sm
set_config(working_memory=1024)
```

Sección 1

Pregunta 1: Exploración de los Datos de los Clientes

Con base en los datos proporcionados:

1. Segmenta a los clientes en función de sus características demográficas.

In [5]:
```
data = pd.read_csv('gabu_data.csv')
data.head()
```

Out[5]:

	ID del Cliente	Edad del Niño	Sexo del Niño	Nivel de Estudios del Padre	Duración de la Subscripción (meses)	Plan de Subscripción	Ingresos del Hogar (USD)	G Sul
0	1	10	Masculino	Undergraduate	22.0	Básico	60983.52	
1	2	7	Femenino	Undergraduate	26.0	Premium	46799.12	
2	3	16	Femenino	Undergraduate	11.0	Básico	59820.15	
3	4	14	Masculino	Undergraduate	21.0	Premium	42901.02	
4	5	11	Masculino	Undergraduate	24.0	Premium	85253.09	

In [6]:
```
# Separar variables numéricas y categóricas
numerical_features = ['Edad del Niño', 'Ingresos del Hogar (USD)']
categorical_features = ['Sexo del Niño', 'Nivel de Estudios del Padre']

# Hacer las categóricas dummies
encoded_features = pd.get_dummies(data[categorical_features],
               columns=categorical_features,
               drop_first=True)

# Estandarizar variables numéricas
data_to_model_standarized = StandardScaler().fit_transform(data[numerical_fe

# Hacer dataframe variables numéricas
data_to_model_df = pd.DataFrame(data_to_model_standarized,
                                columns=numerical_features).reset_index()

# Acomodar variables categóricas
encoded_features_df = encoded_features.reset_index()

# Juntar ambas variables
data_to_model = data_to_model_df.merge(encoded_features_df, on='index')
data_to_model = data_to_model.drop('index', axis=1)
```

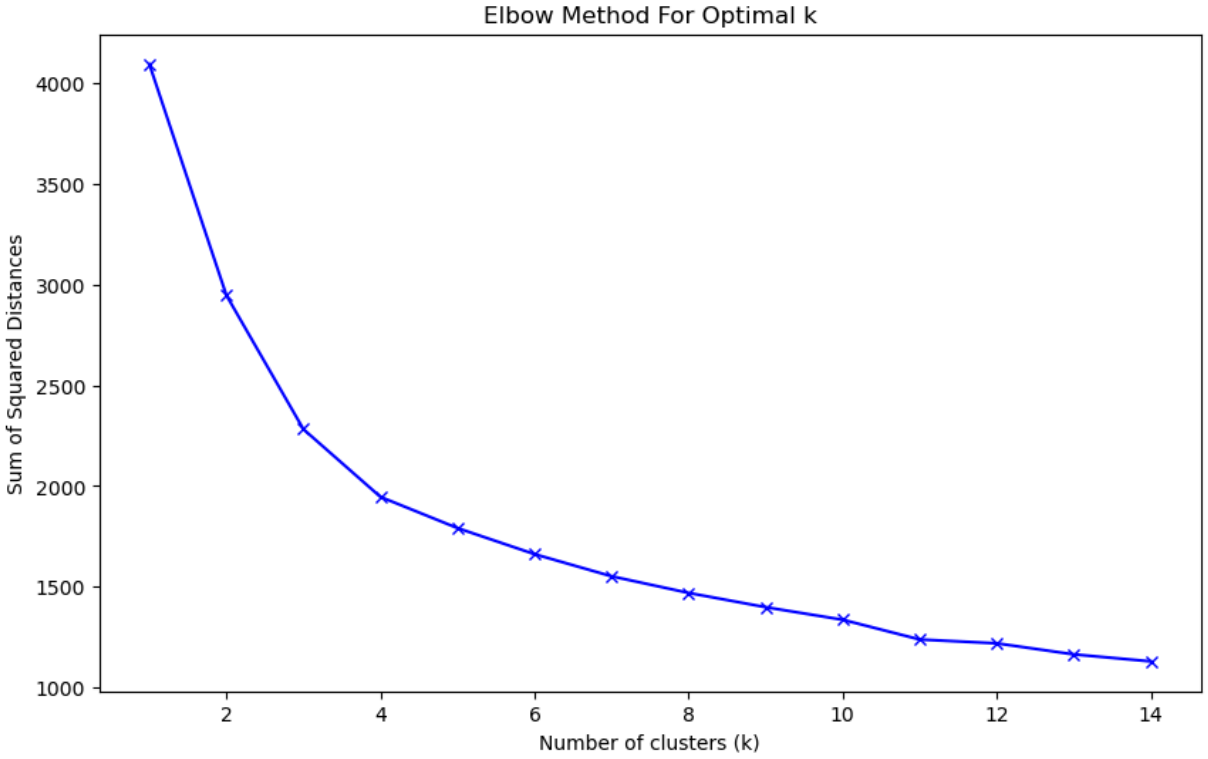
In [7]:
```
columns_from_education = data_to_model.loc[:, 'Sexo del Niño_Masculino':].se
data_to_model[columns_from_education] = data_to_model[columns_from_educatior
data_to_model
```

Out[7]:

	Edad del Niño	Ingresos del Hogar (USD)	Sexo del Niño_Masculino	Nivel de Estudios del Padre_High School	Nivel de Estudios del Padre_Undergraduate
0	-0.103677	0.755915	1	0	1
1	-0.837964	-0.221592	0	0	1
2	1.364896	0.675743	0	0	1
3	0.875371	-0.490227	1	0	1
4	0.141085	2.428435	1	0	1
...	...	...	...	...	...
1513	-0.348439	0.564749	0	1	0
1514	1.120134	1.409065	1	0	0
1515	-1.572250	-1.426295	1	0	1
1516	-1.327488	-0.204638	1	0	0
1517	1.609658	-0.632380	1	0	1

1518 rows × 5 columns

In [8]:
```
# Determinar el número óptimo de clusters usando el método del codo
sum_of_squared_distances = []
K = range(1, 15)  # Ajuste el rango según sea necesario
for k in K:
    km = KMeans(n_clusters=k, random_state=42)
    km = km.fit(data_to_model)
    sum_of_squared_distances.append(km.inertia_)

# Plot the Elbow curve
plt.figure(figsize=(10, 6))
plt.plot(K, sum_of_squared_distances, 'bx-')
plt.xlabel('Number of clusters (k)')
plt.ylabel('Sum of Squared Distances')
plt.title('Elbow Method For Optimal k')
plt.show()
```



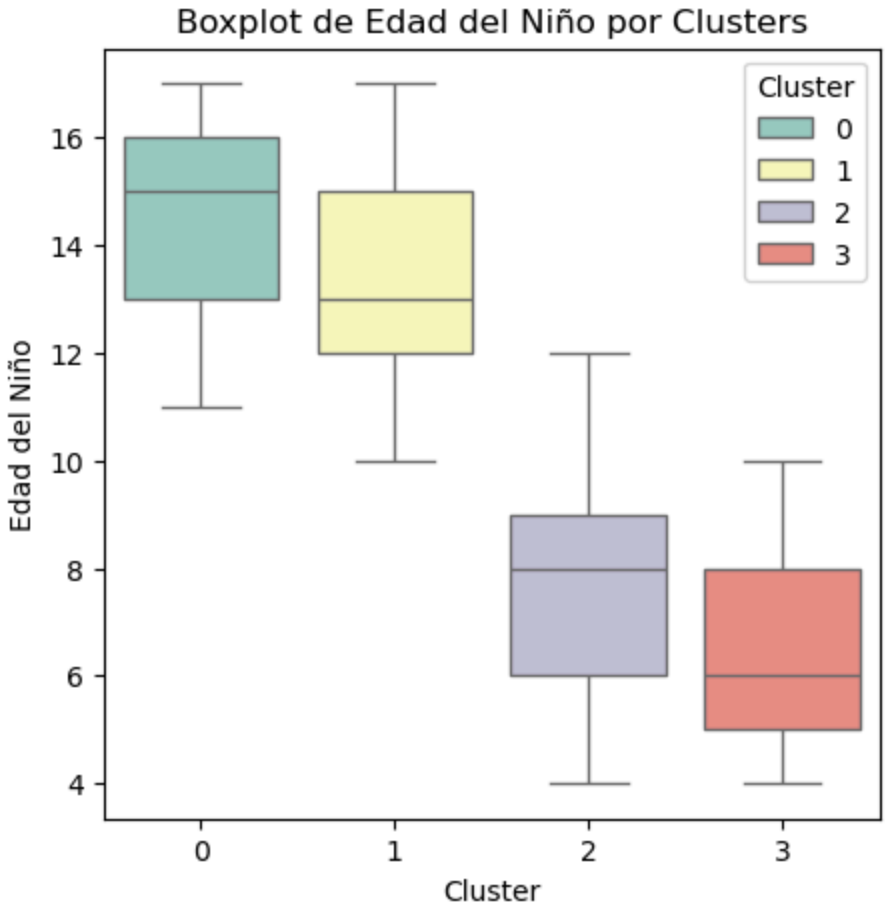
In [9]:
```
# Aplicar K-means clustering para identificar segmentos de clientes
kmeans = KMeans(n_clusters=4, random_state=42)
labels = kmeans.fit_predict(data_to_model)

# Agregar las etiquetas del cluster al DataFrame original para análisis
data_to_model['Cluster'] = labels
data['Cluster'] = labels
data.head()
```

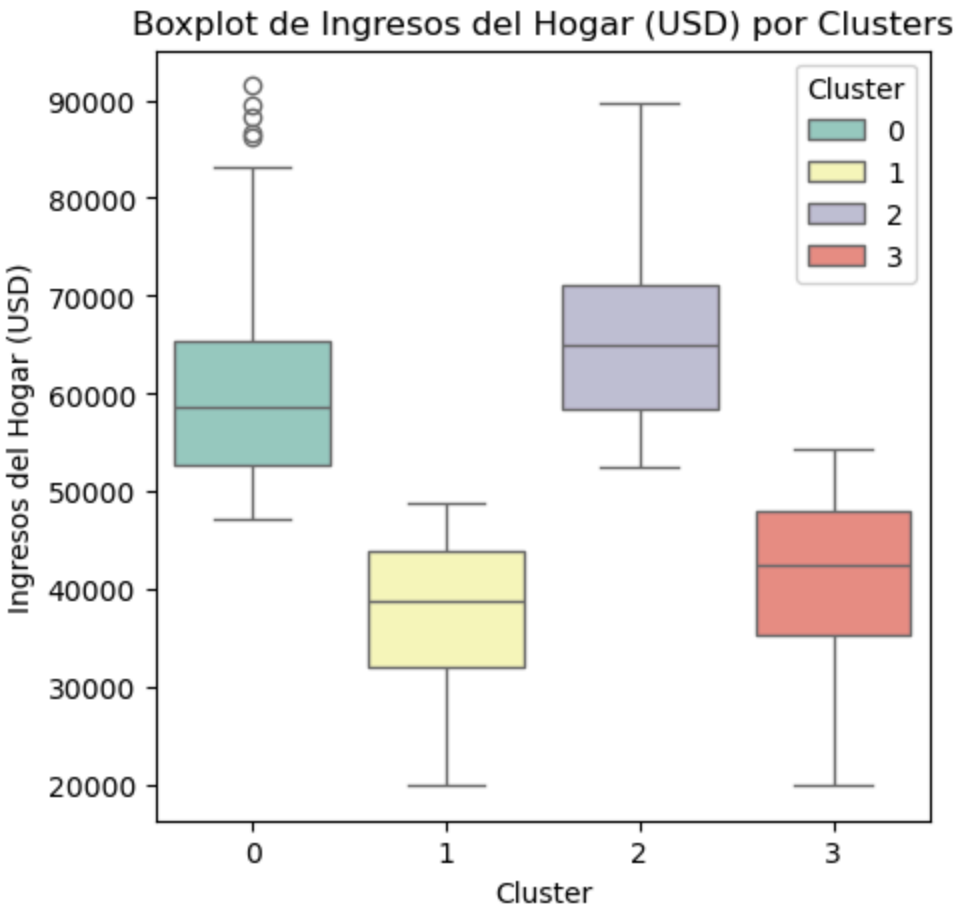
Out[9]:

	ID del Cliente	Edad del Niño	Sexo del Niño	Nivel de Estudios del Padre	Duración de la Subscripción (meses)	Plan de Subscripción	Ingresos del Hogar (USD)	G Sul
0	1	10	Masculino	Undergraduate	22.0	Básico	60983.52	
1	2	7	Femenino	Undergraduate	26.0	Premium	46799.12	
2	3	16	Femenino	Undergraduate	11.0	Básico	59820.15	
3	4	14	Masculino	Undergraduate	21.0	Premium	42901.02	
4	5	11	Masculino	Undergraduate	24.0	Premium	85253.09	

In [10]:
```
def boxplot_clusters(columns):
    for col in columns:
        plt.figure(figsize=(5, 5))
        sns.boxplot(x='Cluster', y=col, data=data, hue='Cluster', palette='S
        plt.title(f'Boxplot de {col} por Clusters')
        plt.show()
```

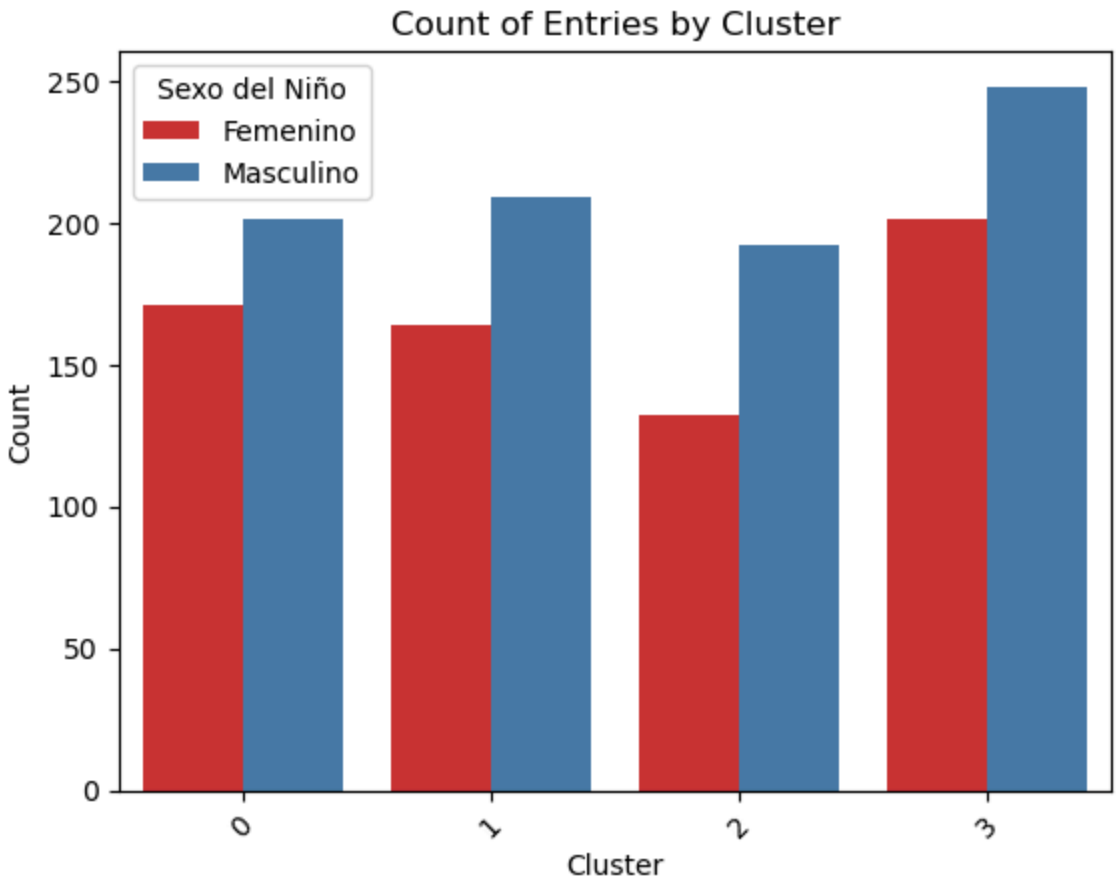
In [11]:
```
boxplot_clusters(numerical_features)
```

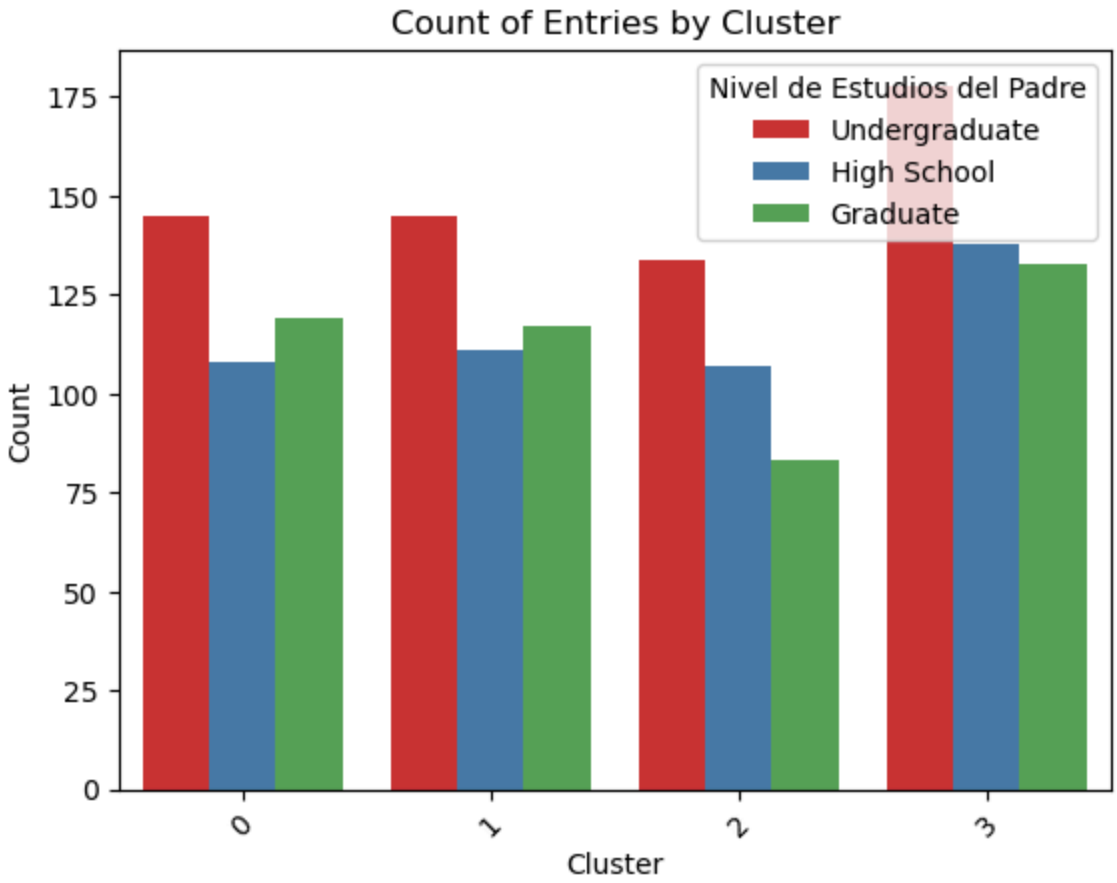




In [12]:
```
def barplot_clusters(columns):
    for col in columns:
        sns.countplot(x='Cluster', hue=col, data=data, palette='Set1')
        plt.title('Count of Entries by Cluster')
        plt.xlabel('Cluster')
        plt.ylabel('Count')
        plt.xticks(rotation=45)
        plt.show()
```

In [13]:
```
barplot_clusters(categorical_features)
```


Count of Entries by Cluster
Sexo del Niño
Femenino
Masculino
Count
Cluster
0
50
100
150
200
250
0
1
2
3


Count of Entries by Cluster
Nivel de Estudios del Padre
Undergraduate
High School
Graduate
Count
Cluster
0
25
50
75
100
125
150
175
0
1
2
3

2. Analiza las características demográficas y económicas de los segmentos. Explica cómo estos segmentos podrían ayudar a Gabu a diseñar estrategias más efectivas de marketing o producto.

Los clusters 0 y 1 se encuentran los niños de mayor edad, y los clusters 0 y 2 están las familias con un mayor ingreso. Además en todos los clusters hay más hombres que mujeres por una ligera diferencia. De igual forma en todos los clusters la mayoría tiene padres con nivel de estudios Undergraduate.

Identificar los clientes en los clusters 0 y 1 y hacer anuncios enfocados a niños de entre 12 y 16 años para que conozcan la empresa. Lo más importante para gabu es dar con los padres pues son quienes van a contratar los servicios, por lo que es importante analizar padres con hijos hombres pues la mayoría de sus clientes son hombres, además analizar que tipo de contenido y redes sociales ven padres con niveles de estudio de Undergraduate pues son la mayoría y así hacer que conozcan la empresa y se preocupen por la seguridad de sus hijos y decidan contratar a la empresa.

Sección 2

Pregunta 2: Selección del KPI Correcto
Gabu quiere identificar un KPI clave que les permita medir la retención de clientes y la capacidad de generar ingresos a largo plazo.

A continuación, se te presentan cuatro KPIs posibles. Solo uno de estos es el KPI correcto que Gabu debería usar para medir la retención de clientes y su impacto en los ingresos futuros. Selecciona el KPI correcto, calcúlalo usando los datos proporcionados, y justifica por qué es el más adecuado.

Instrucciones:

1. Selecciona el KPI correcto que mejor mida la retención de clientes.
2. Explica por qué este KPI es el más adecuado para Gabu y cómo ayudará a medir

Opción 4: Tasa de Churn

- *Definición*: El porcentaje de clientes que cancelan el servicio durante un período de tiempo determinado.
- *Fórmula*: $\text{Tasa de Churn} = \frac{\text{Número de Clientes que Cancelaron}}{\text{Número Total de Clientes Activos al Inicio del Periodo}}$

La tasa de churn es un indicador directamente relacionado con la retención de clientes, pues nos da el porcentaje de clientes que se van, lo cual a su vez nos dice cuantos se quedan, esto nos dice directamente cuantos clientes estamos reteniendo mes a mes.

Además al saber tu retención de clientes puedes hacer buena estimación del impacto en tus ingresos a futuro pues sabes el tipo de suscripción y cuanto pagan, entonces si sabes cuantos se van mes a mes es fácil calcular cuanto pueden reducirse tus ingresos a futuro (sin contar posibles nuevos clientes).

Sección 3

Pregunta 3: Construcción de un Modelo de Regresión para Predecir LTV
El equipo de Gabu desea predecir el valor del tiempo de vida (LTV) de sus clientes usando las características demográficas y económicas del dataset.

1. **Preparación de los Datos**:
   - Usa las siguientes variables como independientes:
     - **Edad del Niño**
     - **Ingresos del Hogar**
     - **Plan de Subscripción**
     - **Número de Sesiones por Semana**
     - **Nivel de Satisfacción del Cliente**
     - **Segmento del cliente**
   - Usa el **Valor del Tiempo de Vida (LTV)** como la variable dependiente.
2. **Construcción del Modelo**:
   - Ajusta un modelo de regresión lineal utilizando las variables mencionadas.
   - Escribe la ecuación de la regresión (incluye los coeficientes de cada variable).

In [20]:
```
# Separar variables numéricas y categóricas
numerical_features = ['Edad del Niño','Ingresos del Hogar (USD)','Número de
categorical_features = ['Plan de Subscripción','Cluster']

# Hacer las categóricas dummies
encoded_features = pd.get_dummies(data[categorical_features],
               columns=categorical_features)

# Estandarizar variables numéricas
data_to_model_standarized = StandardScaler().fit_transform(data[numerical_fe

# Hacer dataframe variables numéricas
data_to_model_df = pd.DataFrame(data_to_model_standarized,
                                columns=numerical_features).reset_index()

# Acomodar variables categóricas
encoded_features_df = encoded_features.reset_index()

# Juntar ambas variables
data_to_model = data_to_model_df.merge(encoded_features_df, on='index')
data_to_model = data_to_model.drop('index', axis=1)
data_to_model.head()
```

Out[20]:

	Edad del Niño	Ingresos del Hogar (USD)	Número de Sesiones por Semana	Nivel de Satisfacción del Cliente	Plan de Subscripción_Básico	Subscripción_I
0	-0.103677	0.755915	-1.423380	-0.886681	True	
1	-0.837964	-0.221592	0.475985	1.058627	False	
2	1.364896	0.675743	-0.232816	-1.415919	True	
3	0.875371	-0.490227	1.262311	1.616473	False	
4	0.141085	2.428435	0.088359	0.343440	False	

In [21]:
```
columns_from_education = data_to_model.loc[:, 'Plan de Subscripción_Básico':
data_to_model[columns_from_education] = data_to_model[columns_from_educatior
data_to_model.head()
```

Out[21]:

	Edad del Niño	Ingresos del Hogar (USD)	Número de Sesiones por Semana	Nivel de Satisfacción del Cliente	Plan de Subscripción_Básico	Subscripción_I
0	-0.103677	0.755915	-1.423380	-0.886681	1	
1	-0.837964	-0.221592	0.475985	1.058627	0	
2	1.364896	0.675743	-0.232816	-1.415919	1	
3	0.875371	-0.490227	1.262311	1.616473	0	
4	0.141085	2.428435	0.088359	0.343440	0	

In [22]:
```
X = data_to_model[['Edad del Niño','Ingresos del Hogar (USD)','Número de Ses
                   'Plan de Subscripción_Básico','Plan de Subscripción_Famil
                   'Cluster_0','Cluster_1','Cluster_2','Cluster_3']]
y = data['Valor del Tiempo de Vida (LTV)']
```

In [23]:
```
# Agregar la constante (intercepto) a las variables independientes
X = sm.add_constant(X)

# Ajustar el modelo usando OLS
model = sm.OLS(y, X)
results = model.fit()

# Mostrar el resumen del modelo
print(results.summary())
```

```
                                 OLS Regression Results
==============================================================================
==============
Dep. Variable:     Valor del Tiempo de Vida (LTV)   R-squared:
0.730
Model:                                        OLS   Adj. R-squared:
0.728
Method:                             Least Squares   F-statistic:
452.6
Date:                            Tue, 08 Oct 2024   Prob (F-statistic):
0.00
Time:                                    19:54:35   Log-Likelihood:
-10039.
No. Observations:                            1518   AIC:
2.010e+04
Df Residuals:                                1508   BIC:
2.015e+04
Df Model:                                       9
Covariance Type:                        nonrobust
==============================================================================
==========================
                                      coef    std err          t      P>|t
|      [0.025      0.975]
------------------------------------------------------------------------------
-------------------------
const                               771.4674     12.277     62.841      0.00
0     747.386     795.548
Edad del Niño                        -7.2439      9.377     -0.772      0.44
0     -25.638      11.150
Ingresos del Hogar (USD)            160.6266      8.425     19.067      0.00
0     144.102     177.152
Número de Sesiones por Semana        29.6188      7.479      3.960      0.00
0      14.948      44.289
Nivel de Satisfacción del Cliente    52.0913      7.139      7.297      0.00
0      38.088      66.095
Plan de Subscripción_Básico         -74.8023     17.726     -4.220      0.00
0    -109.572     -40.033
Plan de Subscripción_Familiar       635.4292     40.182     15.814      0.00
0     556.610     714.248
Plan de Subscripción_Premium        210.8406     15.375     13.713      0.00
0     180.682     240.999
Cluster_0                           193.5224     13.816     14.007      0.00
0     166.421     220.624
Cluster_1                           201.2711     13.227     15.216      0.00
0     175.325     227.217
Cluster_2                           189.3368     14.411     13.139      0.00
0     161.070     217.604
Cluster_3                           187.3372     12.997     14.414      0.00
0     161.844     212.831
==============================================================================
==
Omnibus:                       34.479   Durbin-Watson:                   1.9
94
Prob(Omnibus):                  0.000   Jarque-Bera (JB):                48.5
10
Skew:                           0.244   Prob(JB):                     2.92e-
```

```
11
Kurtosis:                       3.727   Cond. No.                     6.53e+
15
==============================================================================
==

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is corre
ctly specified.
[2] The smallest eigenvalue is 7.88e-29. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

In [24]:
```
X1 = data_to_model[['Ingresos del Hogar (USD)','Número de Sesiones por Semar
                    'Plan de Subscripción_Básico','Plan de Subscripción_Famil
                    'Cluster_0','Cluster_1','Cluster_2','Cluster_3']]
y1 = data['Valor del Tiempo de Vida (LTV)']
```

In [25]:
```
# Agregar la constante (intercepto) a las variables independientes
X1 = sm.add_constant(X1)

# Ajustar el modelo usando OLS
model1 = sm.OLS(y1, X1)
results1 = model1.fit()

# Mostrar el resumen del modelo
print(results1.summary())
```

```
                                 OLS Regression Results
==============================================================================
==============
Dep. Variable:     Valor del Tiempo de Vida (LTV)   R-squared:
0.730
Model:                                        OLS   Adj. R-squared:
0.728
Method:                             Least Squares   F-statistic:
509.3
Date:                            Tue, 08 Oct 2024   Prob (F-statistic):
0.00
Time:                                    19:54:35   Log-Likelihood:
-10039.
No. Observations:                            1518   AIC:
2.010e+04
Df Residuals:                                1509   BIC:
2.014e+04
Df Model:                                       8
Covariance Type:                        nonrobust
==============================================================================
==========================
                                      coef    std err          t      P>|t
|      [0.025      0.975]
------------------------------------------------------------------------------
-------------------------
const                               771.5807     12.274     62.863      0.00
0     747.505     795.657
Ingresos del Hogar (USD)            160.5632      8.423     19.062      0.00
0     144.041     177.085
Número de Sesiones por Semana        29.2821      7.465      3.922      0.00
0      14.639      43.926
Nivel de Satisfacción del Cliente    52.0476      7.138      7.292      0.00
0      38.046      66.049
Plan de Subscripción_Básico         -75.2685     17.713     -4.249      0.00
0    -110.013     -40.524
Plan de Subscripción_Familiar       635.9695     40.171     15.832      0.00
0     557.173     714.766
Plan de Subscripción_Premium        210.8797     15.373     13.718      0.00
0     180.725     241.034
Cluster_0                           186.6418     10.561     17.673      0.00
0     165.926     207.357
Cluster_1                           195.9369     11.280     17.370      0.00
0     173.810     218.064
Cluster_2                           194.7702     12.576     15.488      0.00
0     170.102     219.438
Cluster_3                           194.2317      9.446     20.561      0.00
0     175.702     212.761
==============================================================================
==
Omnibus:                       34.909   Durbin-Watson:                   1.9
96
Prob(Omnibus):                  0.000   Jarque-Bera (JB):                49.2
45
Skew:                           0.246   Prob(JB):                     2.03e-
11
Kurtosis:                       3.732   Cond. No.                     6.47e+
```

```
15
================================================================================
==

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is corre
ctly specified.
[2] The smallest eigenvalue is 7.95e-29. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

771.58 + 160.5632 * Ingresos + 29.28 * Número de sesiones + 52.04 * Satisfacción del cliente - 75.26 * Plan Básico + 635.96 * Plan Familiar + 210.87 * Plan Premium + 186.64 * Cluster 0 + 195.93 * Cluster 1 + 194.77 * Cluster 2 + 194.23 * Cluster 3

3. **Interpretación de los Coeficientes**:
   - Explica el significado de los coeficientes obtenidos en el modelo. ¿Cómo afecta cada variable independiente al LTV?
   - Identifica los factores que parecen tener el mayor impacto en el valor de tiempo de vida y discute por qué podrían ser importantes para Gabu.

Los coeficientes multiplican al camnio en la variable y eso afecta al LTV, por cada unidad de aumento en ingresos del hogar el LTV sube 160.56, por cada unidad de aumento en Plan Básico el LTV decrece 75.26, por cada unidad de aumento en el Plan Premium el LTV sube 210, para los clusters significa que si es del cluster 0 sube 186 al LTV, si es del 1 sube 195 el LTV, si es del 2 sube 194 igual que si es del 3.

De acuerdo a los resultados del modelo los factores que tienen mayor impacto en el LTV son el Plan familiar y el Plan Premium pues son los que tienen los coeficientes más altos. Esto es importante pues saben que lo que más afecta al LTV es el tipo de plan al que se suscriben por lo que es útil impulsar más la contratación del Plan Familiar y Plan Premium pues de acuerdo al modelo les de más LTV.

Sección 4

Pregunta 4: Estrategias Basadas en el Análisis
Basándote en el análisis realizado y los resultados del modelo de regresión, responde a las siguientes preguntas:

1. ¿Qué acciones estratégicas recomendarías para mejorar el LTV de los clientes de Gabu?
2. Considera la tasa de churn. ¿Qué medidas concretas podría tomar Gabu para reducir esta tasa y aumentar la lealtad de los clientes?

In [31]:
```
churn_rate = data.Churn.mean()
churn_rate
```

Out[31]: 0.7628458498023716

Como recomendación sería impulsar la contratación del plan familiar pues es lo que más aporta al LTV, además buscar padres con hijos hombres pues son la mayoría de sus clientes y ver que sean Undergraduate pues usan bastante el servicio.

El churn rate es bastante alto por lo que es de mucha importancia atender el problema, se puede analizar el nivel de satisfacción de quienes se van y quienes se quedan, para ver que es lo que hace que los clientes se queden, además de ver que segemntos siguen en la empresa para dar prioridad a ellos y replicar los servicios que haya con ellos.

Sin estandarizar

In [33]:
```
# Separar variables numéricas y categóricas
numerical_features = ['Edad del Niño','Ingresos del Hogar (USD)','Número de
categorical_features = ['Plan de Subscripción','Cluster']

# Hacer las categóricas dummies
encoded_features = pd.get_dummies(data[categorical_features],
               columns=categorical_features)

# Estandarizar variables numéricas
data_to_model_standarized = data[numerical_features]

# Hacer dataframe variables numéricas
data_to_model_df = pd.DataFrame(data_to_model_standarized,
                                columns=numerical_features).reset_index()

# Acomodar variables categóricas
encoded_features_df = encoded_features.reset_index()

# Juntar ambas variables
data_to_model = data_to_model_df.merge(encoded_features_df, on='index')
data_to_model = data_to_model.drop('index', axis=1)
data_to_model.head()
```

Out[33]:

	Edad del Niño	Ingresos del Hogar (USD)	Número de Sesiones por Semana	Nivel de Satisfacción del Cliente	Plan de Subscripción_Básico	Plan de Subscripción_Familia
0	10	60983.52	1.00	3.25	True	Fals
1	7	46799.12	4.43	4.61	False	Fals
2	16	59820.15	3.15	2.88	True	Fals
3	14	42901.02	5.85	5.00	False	Fals
4	11	85253.09	3.73	4.11	False	Fals

In [34]:
```
columns_from_education = data_to_model.loc[:, 'Plan de Subscripción_Básico':
data_to_model[columns_from_education] = data_to_model[columns_from_educatior
data_to_model.head()
```

Out[34]:

	Edad del Niño	Ingresos del Hogar (USD)	Número de Sesiones por Semana	Nivel de Satisfacción del Cliente	Plan de Subscripción_Básico	Plan d Subscripción_Familia
0	10	60983.52	1.00	3.25	1	
1	7	46799.12	4.43	4.61	0	
2	16	59820.15	3.15	2.88	1	
3	14	42901.02	5.85	5.00	0	
4	11	85253.09	3.73	4.11	0	

In [35]:
```
X = data_to_model[['Edad del Niño','Ingresos del Hogar (USD)','Número de Ses
                   'Plan de Subscripción_Básico','Plan de Subscripción_Famil
                   'Cluster_0','Cluster_1','Cluster_2','Cluster_3']]
y = data['Valor del Tiempo de Vida (LTV)']
```

In [36]:
```
# Agregar la constante (intercepto) a las variables independientes
X = sm.add_constant(X)

# Ajustar el modelo usando OLS
model = sm.OLS(y, X)
results = model.fit()

# Mostrar el resumen del modelo
print(results.summary())
```

```
                                 OLS Regression Results
==============================================================================
==============
Dep. Variable:     Valor del Tiempo de Vida (LTV)   R-squared:
0.730
Model:                                        OLS   Adj. R-squared:
0.728
Method:                             Least Squares   F-statistic:
452.6
Date:                            Tue, 08 Oct 2024   Prob (F-statistic):
0.00
Time:                                    19:54:35   Log-Likelihood:
-10039.
No. Observations:                            1518   AIC:
2.010e+04
Df Residuals:                                1508   BIC:
2.015e+04
Df Model:                                       9
Covariance Type:                        nonrobust
==============================================================================
==========================
                                        coef    std err          t      P>|t
|      [0.025      0.975]
------------------------------------------------------------------------------
-------------------------
const                               214.3772     38.852      5.518      0.00
0     138.167     290.587
Edad del Niño                        -1.7730      2.295     -0.772      0.44
0      -6.275       2.729
Ingresos del Hogar (USD)              0.0111      0.001     19.067      0.00
0       0.010       0.012
Número de Sesiones por Semana        16.4014      4.142      3.960      0.00
0       8.278      24.525
Nivel de Satisfacción del Cliente    74.5100     10.212      7.297      0.00
0      54.480      94.540
Plan de Subscripción_Básico        -260.4991     16.639    -15.656      0.00
0    -293.137    -227.861
Plan de Subscripción_Familiar       449.7324     44.216     10.171      0.00
0     363.000     536.465
Plan de Subscripción_Premium         25.1438     19.651      1.280      0.20
1     -13.403      63.691
Cluster_0                            54.2498     19.791      2.741      0.00
6      15.429      93.070
Cluster_1                            61.9985     15.501      4.000      0.00
0      31.592      92.405
Cluster_2                            50.0642     17.699      2.829      0.00
5      15.348      84.781
Cluster_3                            48.0647     11.702      4.107      0.00
0      25.111      71.018
==============================================================================
==
Omnibus:                       34.479   Durbin-Watson:                   1.9
94
Prob(Omnibus):                  0.000   Jarque-Bera (JB):                48.5
10
Skew:                           0.244   Prob(JB):                     2.92e-
```

```
11
Kurtosis:                       3.727   Cond. No.                     1.52e+
20
==============================================================================
==

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is corre
ctly specified.
[2] The smallest eigenvalue is 1.79e-28. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

afecta más Plan Familiar al LTV