Análisis de consumo

In [ ]:
```
# Cargar las librerias
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from scipy.stats import shapiro
from scipy.stats import f_oneway
from sklearn.preprocessing import OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
from sklearn.preprocessing import StandardScaler
```

In [ ]:
```
# Subir la base de datos
consumo = pd.read_csv('Consumo.csv')
```

In [ ]:
```
# Verificar la base de datos
consumo.isnull().sum()
consumo.dtypes
consumo.head()
consumo.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 440 entries, 0 to 439
Data columns (total 8 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Channel           440 non-null    int64
 1   Region            440 non-null    int64
 2   Fresh             440 non-null    int64
 3   Milk              440 non-null    int64
 4   Grocery           440 non-null    int64
 5   Frozen            440 non-null    int64
 6   Detergents_Paper  440 non-null    int64
 7   Delicassen        440 non-null    int64
dtypes: int64(8)
memory usage: 27.6 KB
```

Análisis descriptivo de los datos

In [ ]:
```
consumo.describe()
```

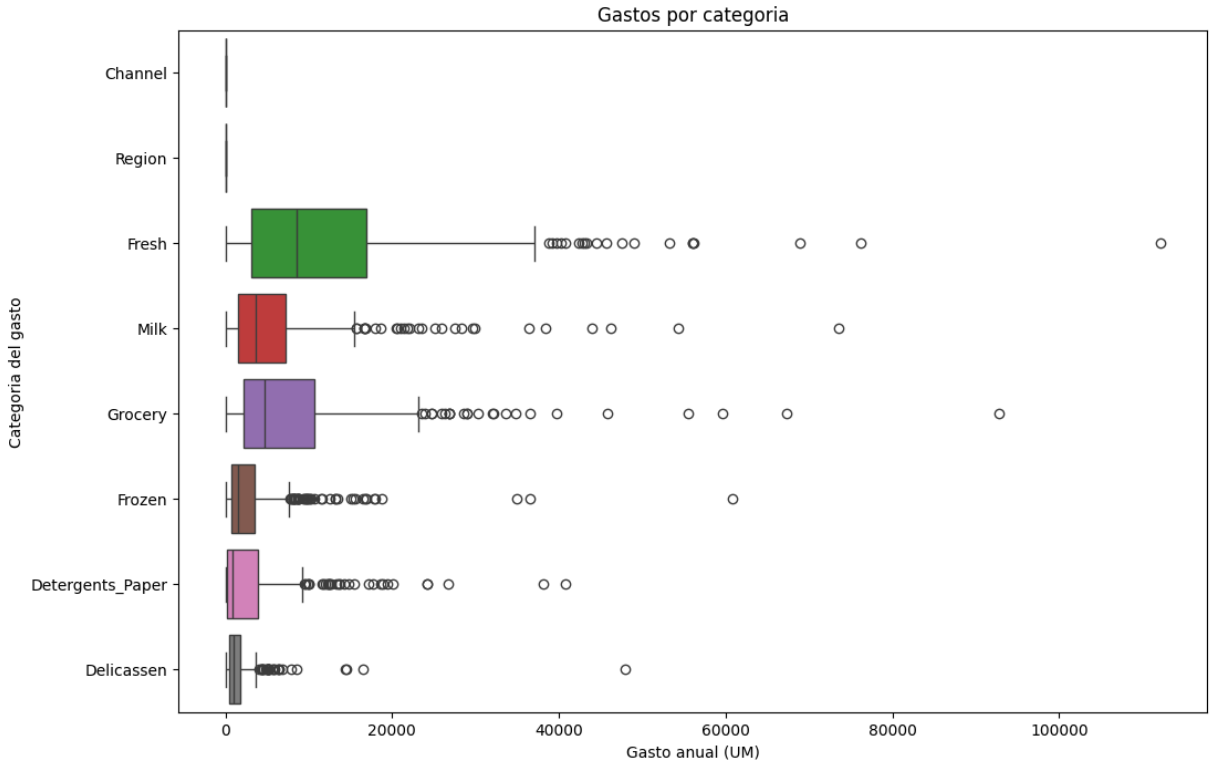
Out[ ]:

	Channel	Region	Fresh	Milk	Grocery	Frozen	D
count	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000	
mean	1.322727	2.543182	12000.297727	5796.265909	7951.277273	3071.931818	
std	0.468052	0.774272	12647.328865	7380.377175	9503.162829	4854.673333	
min	1.000000	1.000000	3.000000	55.000000	3.000000	25.000000	
25%	1.000000	2.000000	3127.750000	1533.000000	2153.000000	742.250000	
50%	1.000000	3.000000	8504.000000	3627.000000	4755.500000	1526.000000	
75%	2.000000	3.000000	16933.750000	7190.250000	10655.750000	3554.250000	
max	2.000000	3.000000	112151.000000	73498.000000	92780.000000	60869.000000	

En el análisis descriptivo de los datos, ya hemos encontrado información interesante. En promedio, los productos más consumidos son los productos frescos, con una media de 12.000 UM. Además, el valor máximo registrado alcanzó los 112.151 UM. La categoría más cercana a los productos frescos son los productos comestibles, los cuales también presentan cifras significativas. Por otro lado, los productos con menor consumo promedio son los delicatessen, probablemente debido a su exclusividad. Un aspecto común en todos los datos es su alta variabilidad; para cada una de las variables, la desviación estándar está por encima de la media, lo que indica que los datos están bastante dispersos. Esta alta variabilidad puede complicar algunos análisis predictivos.

In [ ]:
```
plt.figure(figsize=(12, 8))
sns.boxplot(data=consumo, orient="h")
plt.title("Gastos por categoria")
plt.xlabel("Gasto anual (UM)")
plt.ylabel("Categoria del gasto")
```

Out[ ]:
```
Text(0, 0.5, 'Categoria del gasto')
```

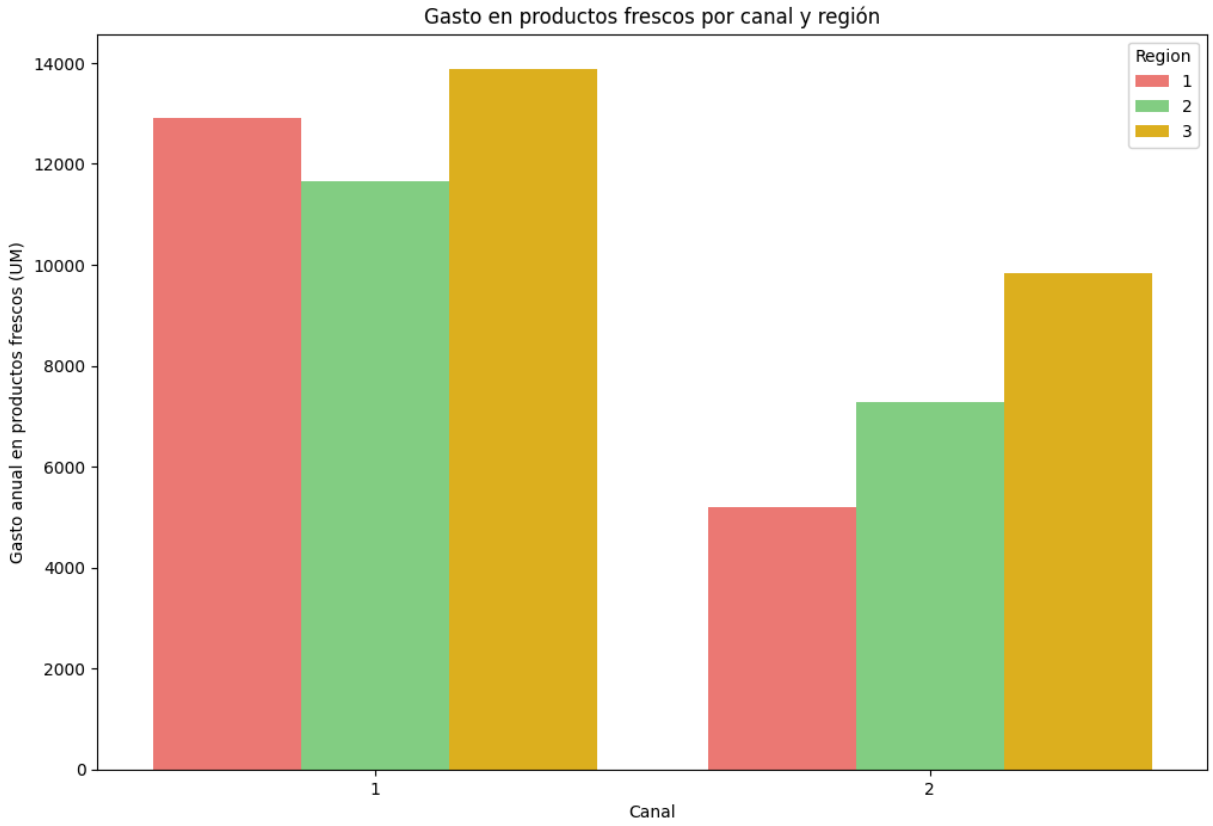


En el boxplot que hemos creado, podemos observar la distribución gráfica de los datos. Notamos que la mayoría de los datos para todas las categorías se encuentran entre 3 UM y 18.000 UM aproximadamente, con una dispersión más alta en los productos frescos y comestibles. También podemos observar que fuera de los cuartiles hay una gran cantidad de valores atípicos, los cuales no son suficientes para entrar en los cuartiles, pero aún así deben ser tenidos en cuenta.

Análisis por canal y región

In [ ]:
```
paleta = ["#FF6961","#77DD77","#FFC300"]
plt.figure(figsize=(12, 8))
sns.barplot(data=consumo, x='Channel', y='Fresh', hue='Region', errorbar=None, pale
plt.title("Gasto en productos frescos por canal y región")
plt.xlabel("Canal")
plt.ylabel("Gasto anual en productos frescos (UM)")
```

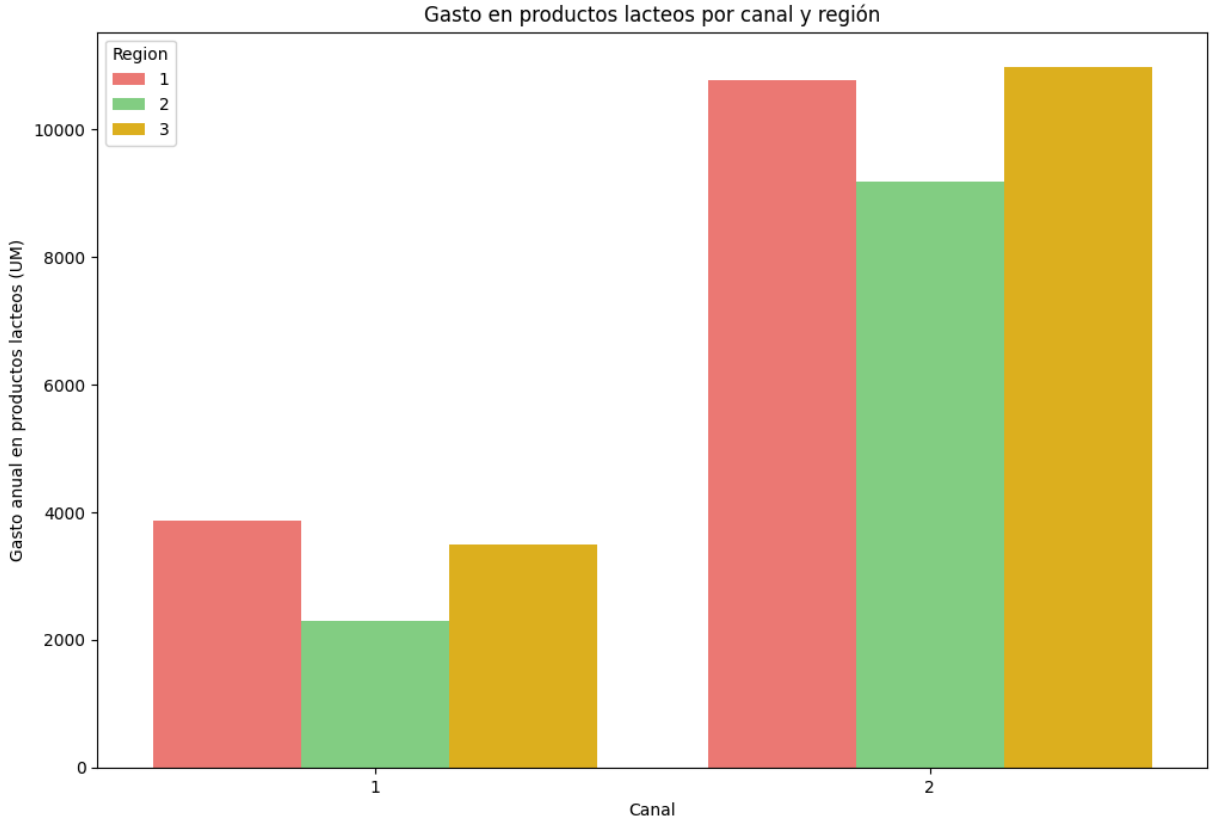
Out[ ]:
```
Text(0, 0.5, 'Gasto anual en productos frescos (UM)')
```



En el siguiente gráfico de barras, podemos observar varias cosas. Para los productos frescos, separados por región y categoría, se observa que tienen una mayor representación en el mercado Horeca en cada una de las regiones. Es necesario recalcar que la región 'Otro' tiene una gran participación en este mercado.

In [ ]:
```
paleta = ["#FF6961","#77DD77","#FFC300"]
plt.figure(figsize=(12, 8))
sns.barplot(data=consumo, x='Channel', y='Milk', hue='Region', errorbar=None, palet
plt.title("Gasto en productos lacteos por canal y región")
plt.xlabel("Canal")
plt.ylabel("Gasto anual en productos lacteos (UM)")
```

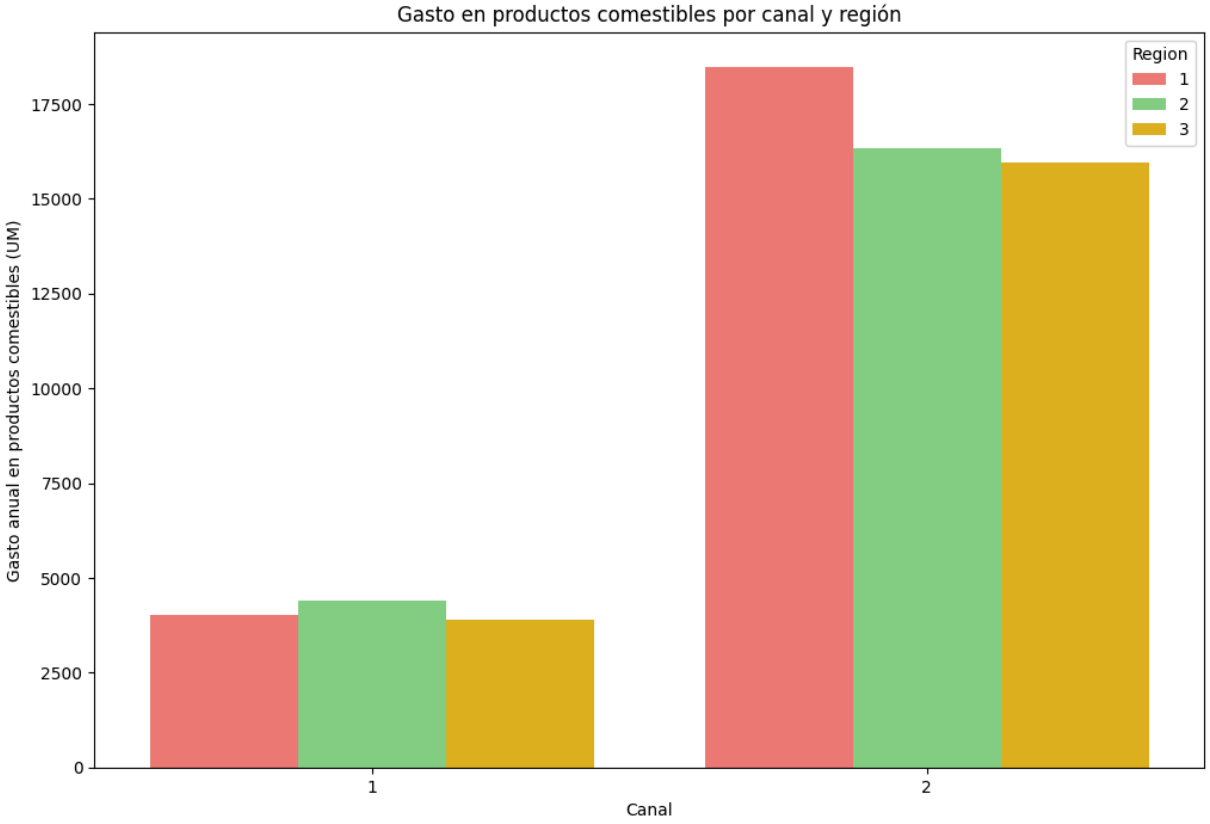
Out[ ]:
```
Text(0, 0.5, 'Gasto anual en productos lacteos (UM)')
```



Para los productos lácteos, se observa el caso contrario a los frescos: la mayor participación se encuentra en el canal minorista. Posiblemente, este cambio se deba a los patrones de consumo y costumbres de los hogares, o al problema del almacenamiento.

In [ ]:
```
paleta = ["#FF6961","#77DD77","#FFC300"]
plt.figure(figsize=(12, 8))
sns.barplot(data=consumo, x='Channel', y='Grocery', hue='Region', errorbar=None, pa
plt.title("Gasto en productos comestibles por canal y región")
plt.xlabel("Canal")
plt.ylabel("Gasto anual en productos comestibles (UM)")
```

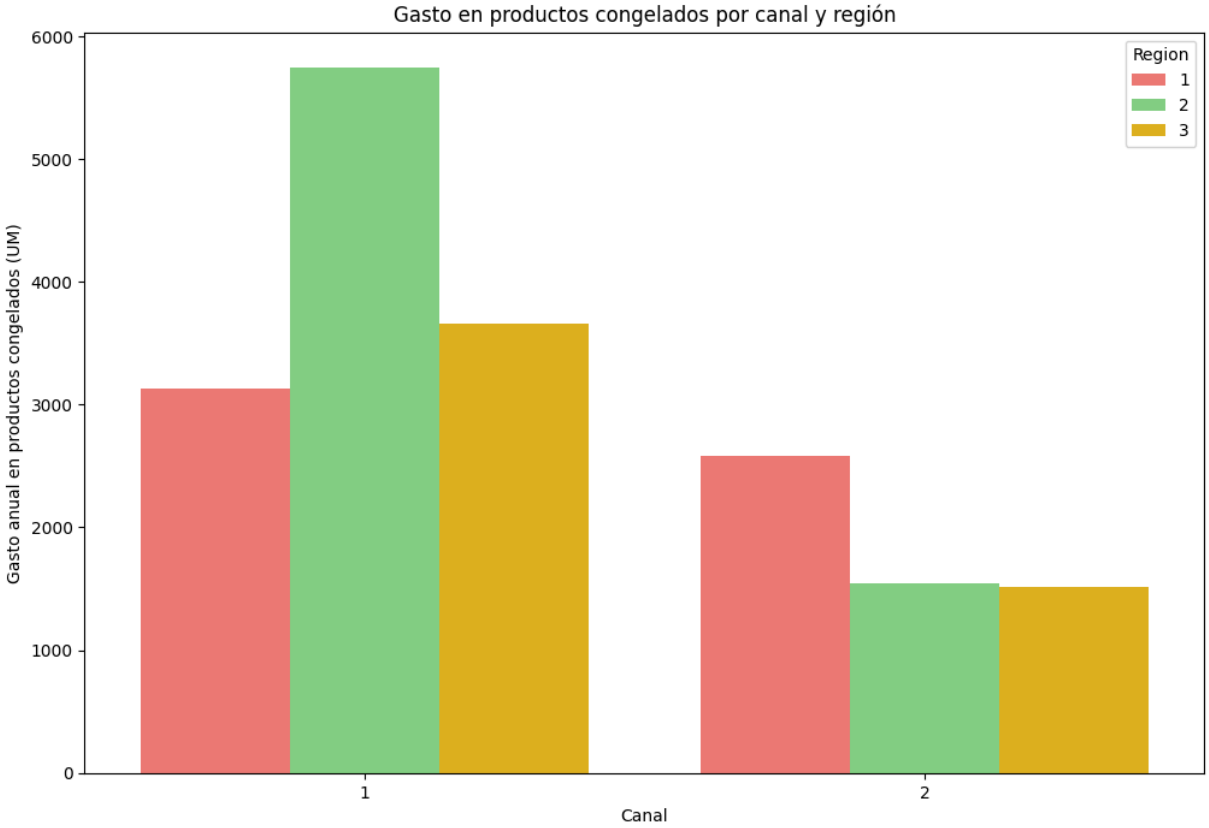
Out[ ]:
```
Text(0, 0.5, 'Gasto anual en productos comestibles (UM)')
```



Para la categoría de comestibles, nuevamente se observa una mayor representación del canal doméstico, lo cual tiene sentido ya que estos productos son más específicos y dirigidos a los hogares en particular.

In [ ]:
```
paleta = ["#FF6961","#77DD77","#FFC300"]
plt.figure(figsize=(12, 8))
sns.barplot(data=consumo, x='Channel', y='Frozen', hue='Region', errorbar=None, pal
plt.title("Gasto en productos congelados por canal y región")
plt.xlabel("Canal")
plt.ylabel("Gasto anual en productos congelados (UM)")
```

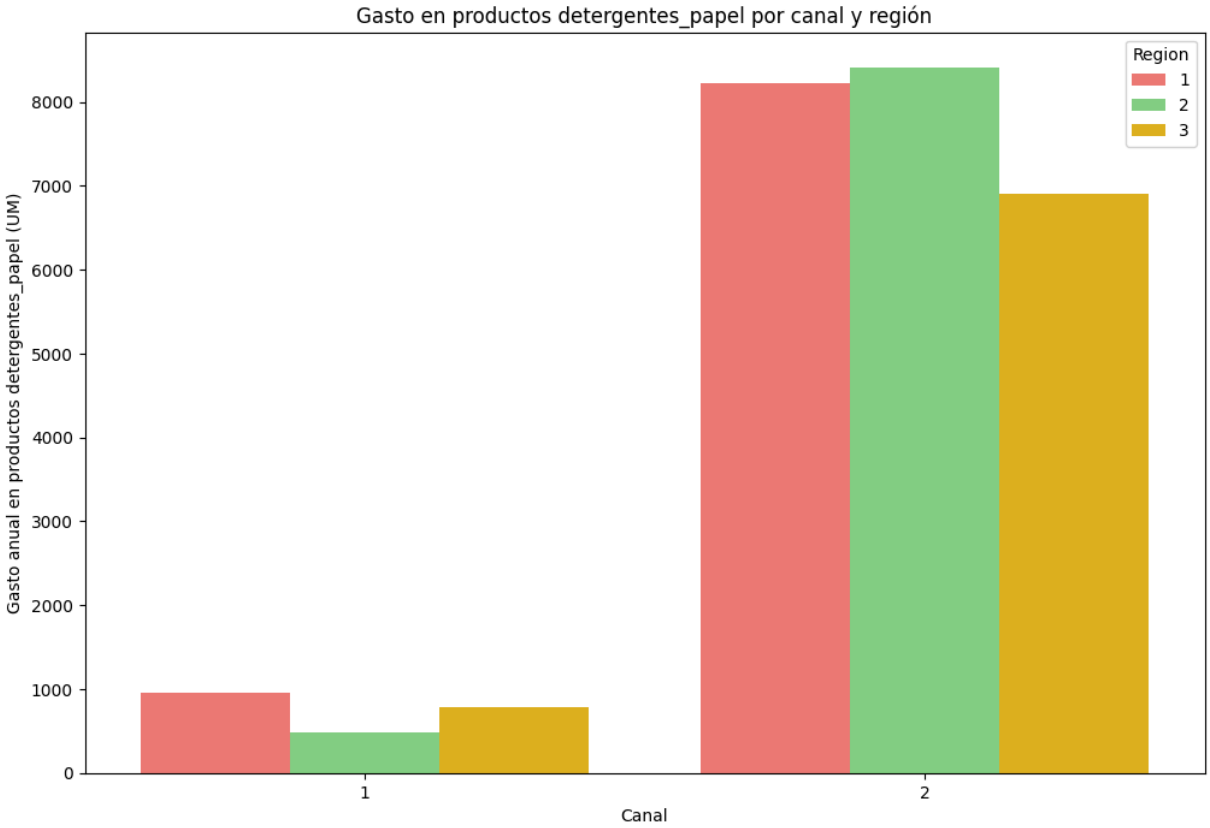
Out[ ]:
```
Text(0, 0.5, 'Gasto anual en productos congelados (UM)')
```



Las gráficas coinciden con lo esperado. Para el canal Horeca, se espera una mayor participación, ya que los productos congelados se utilizan mucho más en este sector, ya sean carnes, verduras o frutas. En el caso de los productos congelados, sería interesante analizar el mercado de Oporto, ya que muestra un alto gasto en estos productos para el sector Horeca. Se necesita más información para encontrar la razón.

In [ ]:
```
paleta = ["#FF6961","#77DD77","#FFC300"]
plt.figure(figsize=(12, 8))
sns.barplot(data=consumo, x='Channel', y='Detergents_Paper', hue='Region', errorbar
plt.title("Gasto en productos detergentes_papel por canal y región")
plt.xlabel("Canal")
plt.ylabel("Gasto anual en productos detergentes_papel (UM)")
```

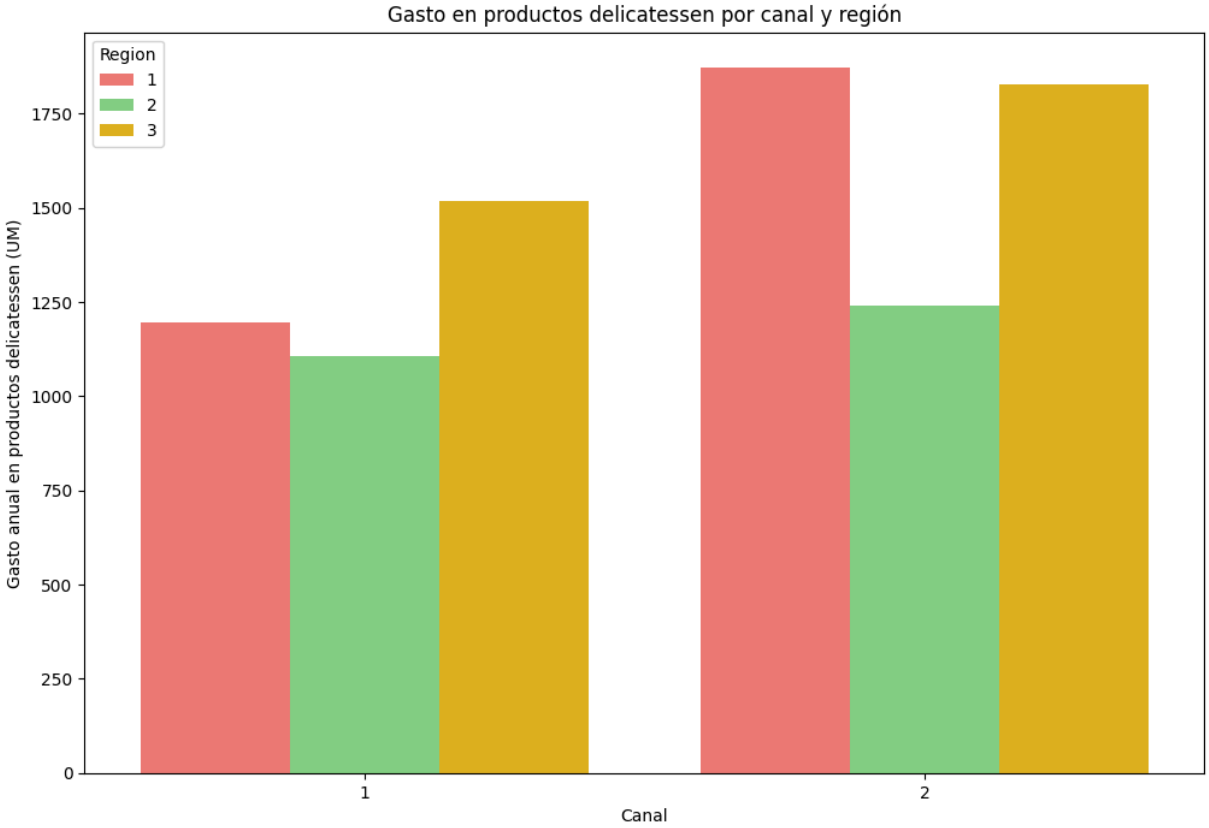
Out[ ]:
```
Text(0, 0.5, 'Gasto anual en productos detergentes_papel (UM)')
```



Como era de esperarse, el canal minorista es el mayor consumidor de este tipo de productos, con un gasto de más de 8.000 unidades monetarias anuales para las regiones de Lisboa y Oporto, y un poco menos para la región 'Otro'.

In [ ]:
```
paleta = ["#FF6961","#77DD77","#FFC300"]
plt.figure(figsize=(12, 8))
sns.barplot(data=consumo, x='Channel', y='Delicassen', hue='Region', errorbar=None,
plt.title("Gasto en productos delicatessen por canal y región")
plt.xlabel("Canal")
plt.ylabel("Gasto anual en productos delicatessen (UM)")
```

Out[ ]:
```
Text(0, 0.5, 'Gasto anual en productos delicatessen (UM)')
```



El caso de los productos delicatessen es muy interesante. Para la región de Oporto, se registran consumos similares en ambos canales, lo cual es curioso. Esto nos indica que las demandas de productos delicatessen no varían demasiado entre estos canales. Sin embargo, se puede notar que Lisboa y la región 'Otro' tienen posiblemente un nivel socioeconómico un poco más elevado que Oporto. Aunque estas son conclusiones preliminares.

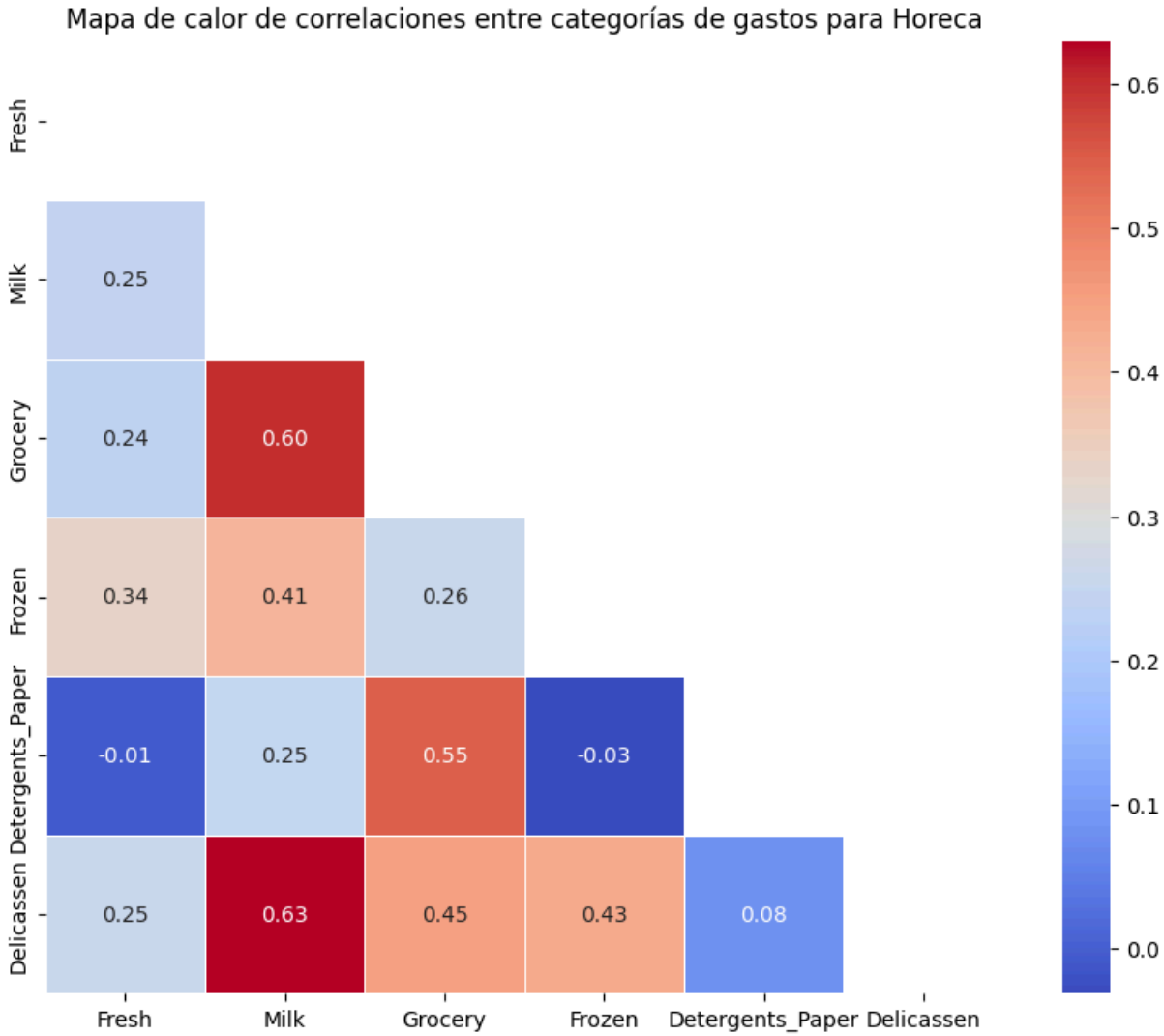
Análisis de correlación

In [ ]:
```
data_horeca = consumo[consumo['Channel'] == 1]

# Calcular la matriz de correlación
correlaciones_horeca = data_horeca[['Fresh', 'Milk', 'Grocery', 'Frozen', 'Detergen

# Visualizar el mapa de calor de la matriz de correlación para Horeca
mask = np.triu(np.ones_like(correlaciones_horeca, dtype=bool))
plt.figure(figsize=(10, 8))
sns.heatmap(correlaciones_horeca, annot=True, cmap='coolwarm', fmt=".2f", linewidth
plt.title("Mapa de calor de correlaciones entre categorías de gastos para Horeca")
```

Out[ ]:
```
Text(0.5, 1.0, 'Mapa de calor de correlaciones entre categorías de gastos para Hor
eca')
```



En este análisis de correlación para el canal Horeca, podemos observar una interesante relación positiva entre los alimentos lácteos y los delicatessen. Esto puede deberse a algunos tipos de recetas en común o a que algunos de los productos delicatessen más comprados sean a la vez lácteos, lo que nos permite generar alguna campaña publicitaria específica.

Otra alta relación se encuentra entre los lácteos y los productos comestibles. Esto puede ser nuevamente una relación por categoría, ya que los lácteos también son comestibles, o puede ser un tema de complementariedad, ya que a la hora de comprar los productos de consumo en una misma canasta estén incluidos los productos lácteos para la preparación de alguna receta en el mercado Horeca, dejando una alta correlación en el consumo. Algo similar sucede con los productos comestibles y los detergentes y productos de papel, pero en menor medida.

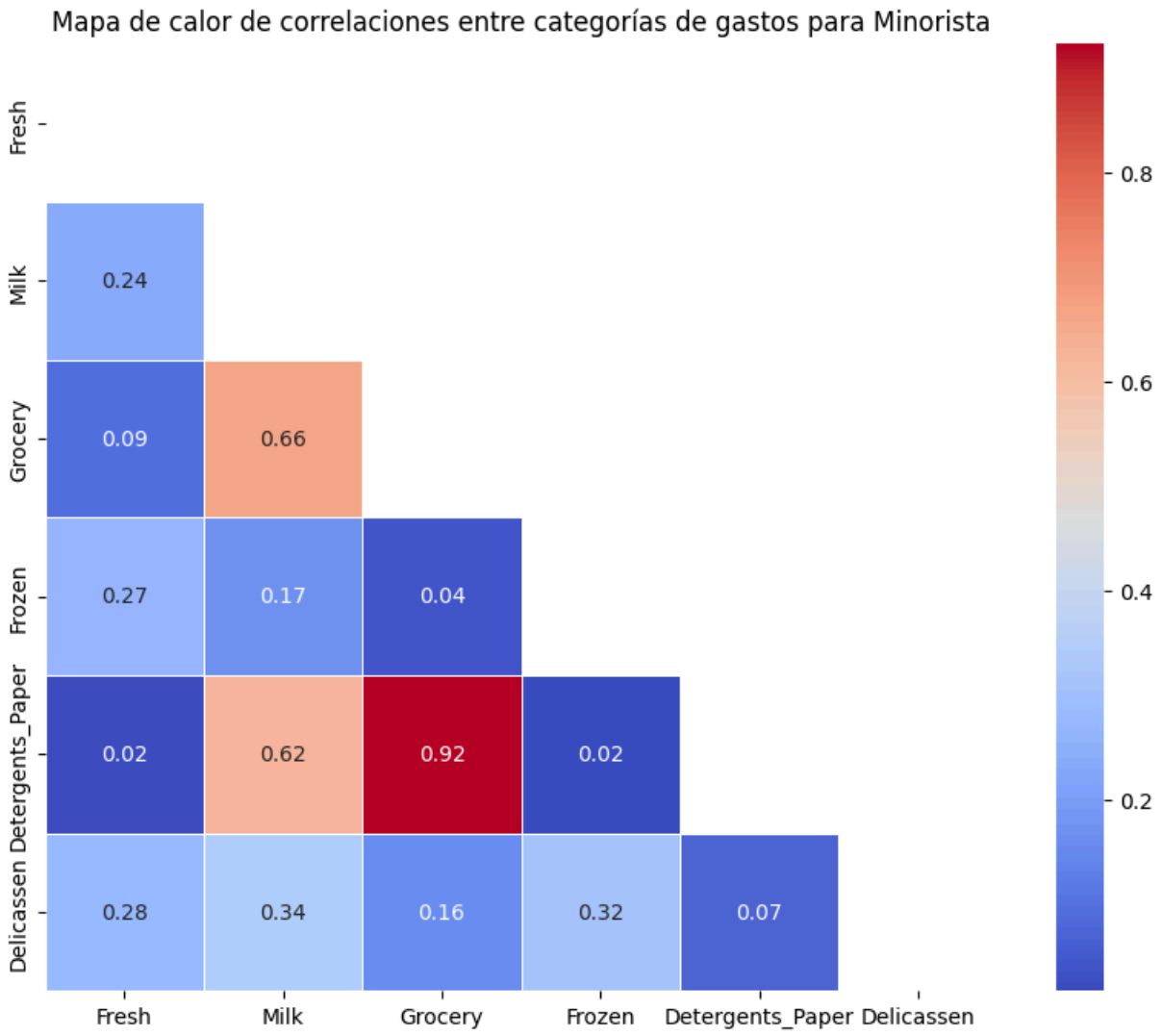
In [ ]:
```
data_minorista = consumo[consumo['Channel'] == 2]

# Calcular la matriz de correlación
correlaciones_minorista = data_minorista[['Fresh', 'Milk', 'Grocery', 'Frozen', 'De

# Visualizar el mapa de calor de la matriz de correlación
```

```
mask = np.triu(np.ones_like(correlaciones_minorista, dtype=bool))
plt.figure(figsize=(10, 8))
sns.heatmap(correlaciones_minorista, annot=True, cmap='coolwarm', fmt=".2f", linewi
plt.title("Mapa de calor de correlaciones entre categorías de gastos para Minorista
```

Out[ ]:
```
Text(0.5, 1.0, 'Mapa de calor de correlaciones entre categorías de gastos para Min
orista')
```



Para el canal minorista, encontramos varios datos significativamente altos. Se trata de los productos comestibles, detergentes y productos de papel, lácteos y detergentes y productos de papel, y comestibles con lácteos. Esto era de esperarse, ya que en la cesta de la compra de una familia, estos productos son comunes, por lo que no sorprende la similitud en el gasto entre ellos.

Análisis estadisticos

In [ ]:
```
# Prueba de normalidad
variables = ['Fresh', 'Milk', 'Grocery', 'Frozen', 'Detergents_Paper', 'Delicassen'

# Realizar el test de normalidad de Shapiro-Wilk para cada columna numérica
for columna in variables:
```

```
    stat, p_valor = shapiro(consumo[columna])
    alpha = 0.05
    print("Variable:", columna)
    print("Estadístico de prueba:", stat)
    print("Valor p:", p_valor)
    if p_valor > alpha:
        print("Los datos de", columna, "parecen provenir de una distribución normal
    else:
        print("Los datos de", columna, "no parecen provenir de una distribución nor
    print()
```

```
Variable: Fresh
Estadístico de prueba: 0.7814360027084953
Valor p: 7.91834381615532e-24
Los datos de Fresh no parecen provenir de una distribución normal (se rechaza H0)

Variable: Milk
Estadístico de prueba: 0.6283341286386022
Valor p: 9.762266807066569e-30
Los datos de Milk no parecen provenir de una distribución normal (se rechaza H0)

Variable: Grocery
Estadístico de prueba: 0.6762299716941169
Valor p: 3.906111904318252e-28
Los datos de Grocery no parecen provenir de una distribución normal (se rechaza H0)

Variable: Frozen
Estadístico de prueba: 0.5282971036489638
Valor p: 1.291356071720309e-32
Los datos de Frozen no parecen provenir de una distribución normal (se rechaza H0)

Variable: Detergents_Paper
Estadístico de prueba: 0.6054819736760775
Valor p: 1.9145818979673575e-30
Los datos de Detergents_Paper no parecen provenir de una distribución normal (se rec
haza H0)

Variable: Delicassen
Estadístico de prueba: 0.36106769717477094
Valor p: 1.7534157228595292e-36
Los datos de Delicassen no parecen provenir de una distribución normal (se rechaza H
0)
```

Al realizar el test de normalidad de Shapiro-Wilk para cada una de las variables, nos damos cuenta de que no cumplen con el supuesto de normalidad. Por lo tanto, realizar pruebas estadísticas paramétricas no sería adecuado. Por ello, intentamos realizar pruebas no paramétricas, como la prueba de Wilcoxon.

In [ ]:
```
#Segmentar las categorias
canal1 = consumo[consumo['Channel'] == 1]
lisboa1 = canal1[canal1['Region'] == 1]
oporto1 = canal1[canal1['Region'] == 2]
otro1 = canal1[canal1['Region'] == 3]
canal2 = consumo[consumo['Channel'] == 2]
```

```
lisboa2 = canal2[canal2['Region'] == 1]
oporto2 = canal2[canal2['Region'] == 2]
otro2 = canal2[canal2['Region'] == 3]
len(canal1), len(canal2)
```

Out[ ]: (298, 142)

A la hora de intentar la prueba de Wilcoxon, se requiere la misma cantidad de datos para cada variable a analizar, lo cual no es posible de realizar sin borrar datos. Por lo tanto, se procede a realizar las pruebas paramétricas normales, aunque no tengan mucha validez estadística.

In [ ]:
```
# Agrupar las variables a análizar
variables_analisis = ['Fresh', 'Milk', 'Grocery', 'Frozen', 'Detergents_Paper', 'De

# Realiza el test de ANOVA para cada variable
for variable in variables_analisis:
    muestras = [canal1[variable], canal2[variable]]

    # Realiza el test de ANOVA
    stat, p_valor = f_oneway(*muestras)

    alpha = 0.05
    print("Variable:", variable)
    print("Estadístico de prueba:", stat)
    print("Valor p:", p_valor)
    if p_valor > alpha:
        print("No hay diferencias significativas entre los canales para la variable
    else:
        print("Hay diferencias significativas entre los canales para la variable",
    print()
```

```
Variable: Fresh
Estadístico de prueba: 12.904516823782437
Valor p: 0.00036495302669371045
Hay diferencias significativas entre los canales para la variable Fresh

Variable: Milk
Estadístico de prueba: 118.02326252660595
Valor p: 1.664946203516875e-24
Hay diferencias significativas entre los canales para la variable Milk

Variable: Grocery
Estadístico de prueba: 257.93181532119576
Valor p: 5.695744747791638e-46
Hay diferencias significativas entre los canales para la variable Grocery

Variable: Frozen
Estadístico de prueba: 18.641269794739
Valor p: 1.9528460419233e-05
Hay diferencias significativas entre los canales para la variable Frozen

Variable: Detergents_Paper
Estadístico de prueba: 297.5528596874019
Valor p: 2.95411666578313e-51
Hay diferencias significativas entre los canales para la variable Detergents_Paper

Variable: Delicassen
Estadístico de prueba: 1.378453213262366
Valor p: 0.24100277082820962
No hay diferencias significativas entre los canales para la variable Delicassen
```

Estos resultados son congruentes con las gráficas que se presentaron al inicio del análisis, ya que para todas las variables hay diferencias significativas entre los canales Horeca y minorista, excepto para los productos delicatessen, para los cuales los datos eran curiosamente similares en ambos canales.

Análisis predictivo

In [ ]:
```python
for variable_objetivo in variables_analisis:
    # Seleccionar la variable objetivo actual
    y = consumo[variable_objetivo]

    # Seleccionar las características (X) correspondientes
    X = consumo.drop(['Channel', 'Region'], axis=1)

    # Dividir los datos en conjuntos de entrenamiento y prueba
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random

    # Inicializar el modelo de regresión lineal
    model = LinearRegression()

    # Entrenar el modelo
    model.fit(X_train, y_train)
```

```
    # Hacer predicciones en el conjunto de prueba
    y_pred = model.predict(X_test)

    # Calcular el error cuadrático medio (MSE)
    mse = mean_squared_error(y_test, y_pred)
    print("Error cuadrático medio para", variable_objetivo, ":", mse)

    # Imprimir los coeficientes del modelo para ver su importancia
    print("Coeficientes del modelo para", variable_objetivo, ":", model.coef_)
    print()
```

```
Error cuadrático medio para Fresh : 1.015582306171917e-22
Coeficientes del modelo para Fresh : [ 1.00000000e+00 -8.55938781e-17  2.28457172e-1
7 -1.57239784e-16
 -3.81118365e-17 -2.23191082e-17]

Error cuadrático medio para Milk : 3.630137084869404e-23
Coeficientes del modelo para Milk : [-3.39469203e-17  1.00000000e+00  1.11022302e-16
3.33066907e-16
 -2.22044605e-16  2.22044605e-16]

Error cuadrático medio para Grocery : 2.5406723267595534e-23
Coeficientes del modelo para Grocery : [ 6.78938406e-17  1.66533454e-16  1.00000000e
+00  3.05311332e-16
 -4.99600361e-16 -2.77555756e-17]

Error cuadrático medio para Frozen : 9.273183290032719e-24
Coeficientes del modelo para Frozen : [-1.17753380e-16 -1.11022302e-16  4.99600361e-
16  1.00000000e+00
 -1.66533454e-16  1.66533454e-16]

Error cuadrático medio para Detergents_Paper : 2.5088541851072917e-23
Coeficientes del modelo para Detergents_Paper : [ 1.01840761e-16 -1.66533454e-16 -4.
57966998e-16 -2.77555756e-16
  1.00000000e+00  2.77555756e-16]

Error cuadrático medio para Delicassen : 3.0596771978535157e-24
Coeficientes del modelo para Delicassen : [-5.94071105e-17  5.55111512e-17 -1.942890
29e-16  0.00000000e+00
  4.71844785e-16  1.00000000e+00]
```

Para realizar el análisis predictivo, se escogió un modelo de regresión lineal, ya que es el más común y, al no tener datos temporales, es el más acertado.

El modelo intenta explicar una variable en función de las demás variables presentes en el modelo. La idea es encontrar relaciones que nos permitan vincular la variable objetivo con las demás.

En este caso, no fue posible, ya que ninguno de los valores que arroja el modelo para cada una de las variables correspondientes resultó significativo, excepto para la variable en sí misma, pero ese resultado no nos dice mucho. Esto se puede explicar porque la mayoría de

las variables tienen una alta desviación estándar, lo que hace difícil llegar a conclusiones válidas.

Al momento de realizar las regresiones, se segmentaron los datos por canal y región para cada una de las variables en particular, pero en todos los casos los valores eran similares y no significativos, por lo que se dejó el dataframe completo en el resultado final.

Conclusiones

Los datos proporcionados nos muestran los patrones de consumo anuales para ciertos productos en particular, del análisis podemos concluir.

1. Los productos más consumidos son los productos frescos. Sería necesario realizar un análisis más profundo para decidir si los demás productos requieren una campaña publicitaria o para entender por qué los productos frescos están tan bien posicionados.
2. Los datos muestran una dispersión considerable, con desviaciones estándar más altas que la media, lo que complica la validez de los modelos predictivos.
3. Para ambos canales, los productos delicatessen tienen consumos similares en las tres regiones. Sería interesante analizar este segmento más a fondo para encontrar la razón en particular.
4. En el caso del canal Horeca, la matriz de correlación muestra una fuerte relación entre los lácteos y los productos delicatessen, así como también con los productos comestibles. Por lo tanto, organizar estos productos en los establecimientos de tal forma que los lácteos estén cerca de estos dos productos sería una buena estrategia para incentivar las ventas.
5. Para el canal minorista, existe una estrecha relación entre los productos de detergentes y papel y los comestibles. Sin embargo, colocarlos muy cerca no es necesario debido a la alta relación. Sería más efectivo realizar campañas de promoción entre lácteos y comestibles, ya que tendrían un buen impacto en las ventas de ambos.
6. Los datos no siguen una distribución normal y algunos tests no son muy confiables.