PROYECTO PÁGINA WEB INTERACTIVA PARA EL ANÁLISIS ESTADÍSTICO, SEGMENTACIÓN DE CLIENTES, PREDICCIÓN Y TOMA DE DECISIONES DE EMPRESA RETAIL CON COBERTURA GLOBAL

INTRODUCCIÓN

Somos The Data Guys, un grupo de entusiastas de los datos dedicados a colaborar con empresas de Retail, ayudándolos a generar valor a su negocio a través del análisis de datos, entre los servicios que aportamos:

- Estadística descriptiva
- Segmentación de clientes
- Análisis predictivos
- Recomendaciones para la toma de decisiones

Idea de Proyecto

Hemos elegido la idea del proyecto a partir de una necesidad del equipo de Marketing y ventas de la empresa de Retail En Línea. Nos aborda con una problemática: **muchos datos y pocas conclusiones**, a partir de esto hemos decidido emprender un análisis inicial para determinar necesidades y recomendaciones basados en la historia, a su vez, implementamos machine learning para generar predicciones futuras respecto a las ventas segmentadas por tipo de cliente, utilizando el modelo RFM (Recencia, Frecuencia, Valor Monetario).

Todo este análisis exploratorio culminará en la creación de una aplicación web interactiva que facilite a los líderes de la empresa la toma de decisiones. La aplicación ofrecerá herramientas para cargar datos, realizar análisis exploratorios y visualizar los resultados de la segmentación a través de gráficos interactivos, facilitando el filtrado de un universo de compradores a audiencias que realmente pueden convertirse en compradores y generar leads de ventas.

Estructura de Proyecto

El proyecto consta de 3 etapas, la primera de ellas consiste en la estadística descriptiva. En esta etapa se identifica en la base de datos una presencia muy fuerte (90% de los datos) de United Kingdom, por lo que los análisis realizados se dividen en dos: uno de ellos tratando la base de datos completa, incluyendo todos los países donde la empresa factura; y un análisis específico para United Kingdom.
Se obtuvieron las ventas totales por región, ventas totales por cluster, ventas totales por mes, el top de productos más vendidos (por cluster y región), además de histogramas de los ingresos totales por cliente y también de los pedidos realizados.

La segunda etapa consistió en la segmentación de clientes, se dividió la base de datos original en 5 clusters, de 0 a 4, donde 0 indica clientes muy poco activos, muy baja frecuencia de compra y bajo valor; y en contraste, 4 indica clientes muy activos, con alta frecuencia de compra y un alto valor.

La tercera etapa consistió en la visualización de los datos y el estructuramiento de la página de exposición.

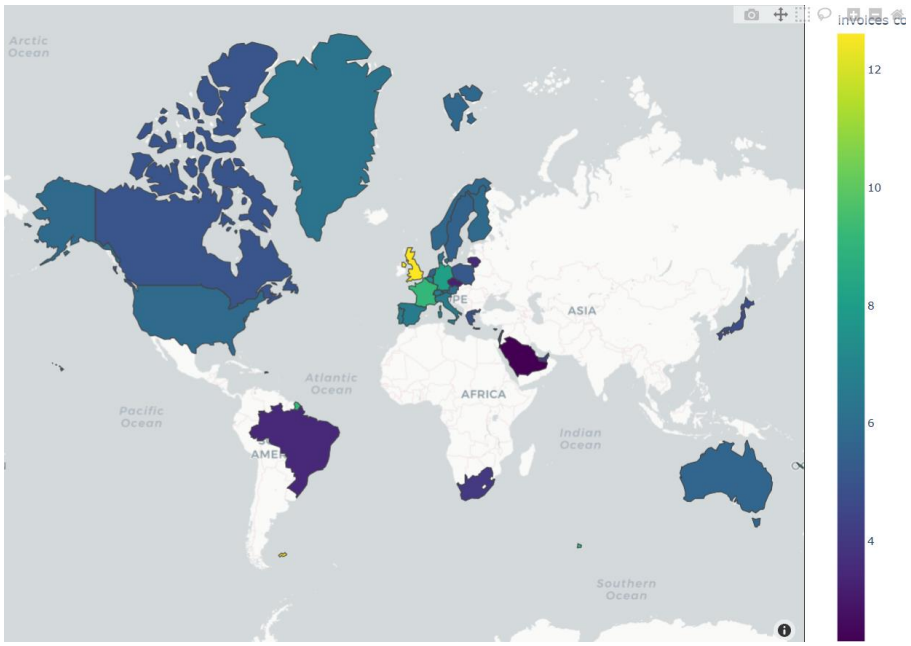
METODOLOGÍA

1. **Preprocesamiento y Manipulación de Datos:** Se le dio un primer vistazo a la base de datos, se eliminaron los duplicados y también se trató los valores nulos.
2. **Análisis Exploratorio de Datos (EDA):** Visualización de estadísticas descriptivas. Ofrece un resumen visual de las características clave de los datos, tanto para la base de datos completa como para la base de datos considerando solo UK.
3. **Estadística Descriptiva:** Cálculo de medidas de tendencia central y dispersión: Proporciona una visión general de la distribución de los datos.
4. **Aplicar Técnicas de Clustering:** Se implementa K-means para segmentar los clientes en grupos basados en similitudes. Se determina el número óptimo de clusters para asegurar una segmentación significativa y robusta. Finalmente, se visualizan los clusters, lo que ayuda a interpretar y validar los resultados del clustering.
5. **Evaluación de Modelos de Clustering:** Cálculo de inercia y score de silueta para proporcionar métricas objetivas y evaluar la calidad del clustering.
6. **Análisis Post-Clustering:** Se realizan estadísticas por cluster para caracterizar y entender cada segmento de clientes.
7. **Técnicas de Machine Learning Supervisado:** Se preparan los datos para modelado, se organiza los datos en el formato adecuado para el aprendizaje supervisado. Se dividen en conjuntos de entrenamiento y prueba, lo que permite una evaluación imparcial del rendimiento del modelo. Se implementan múltiples algoritmos para comparar diferentes enfoques de modelado para encontrar el más adecuado.
8. **Evaluación de Modelos de Clasificación:** Se calcula de métricas de rendimiento para proporcionar una evaluación cuantitativa del desempeño de los modelos. Generación de curvas ROC y AUC para evaluar el rendimiento de los modelos en diferentes umbrales de clasificación. También se realiza una comparación visual de modelos para facilitar la selección del mejor modelo para el problema en cuestión.
9. **Visualización Avanzada:** Se realizan gráficos 3D y heatmap para visualizar relaciones complejas entre múltiples variables, además de gráficos comparativos para comunicar eficazmente los resultados del análisis y modelado.
10. **Análisis de Series Temporales:** Análisis de datos temporales que permiten identificar patrones y tendencias a lo largo del tiempo. Análisis de estacionalidad, para identificar patrones cíclicos en el comportamiento de compra.
11. **Segmentación de Clientes:** Aplicación de técnicas RFM, para lograr una segmentación basada en el comportamiento reciente y el valor del cliente. Interpretación de segmentos para desarrollar estrategias personalizadas para diferentes grupos de clientes.
12. **Estadística Inferencial:** Cálculo de covarianzas para medir la relación entre variables. Análisis de asimetría y curtosis para obtener información sobre la forma de las distribuciones de datos.
13. **Manejo de Problemas Multiclase:** Adaptación de métricas y visualizaciones que permiten aplicar técnicas de clasificación a problemas con más de dos clases.

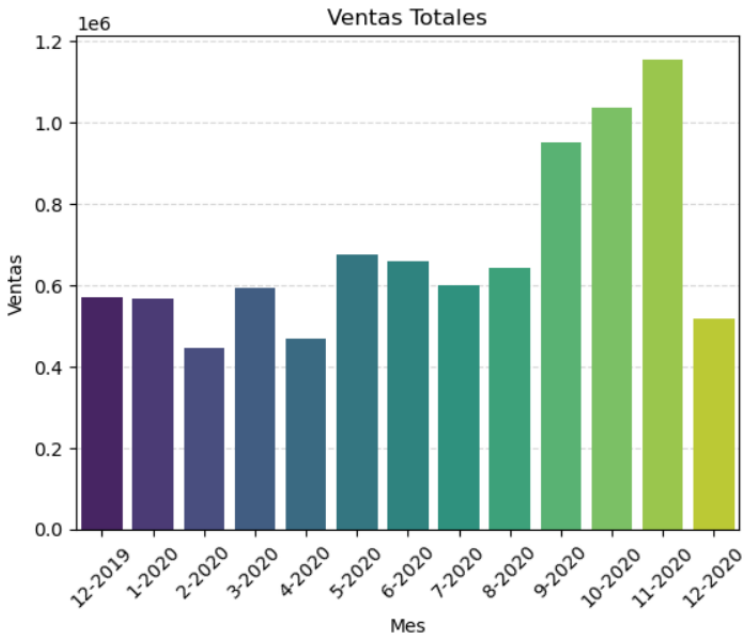
Esta estructura nos permite alcanzar un enfoque integral y sofisticado en el análisis de datos, abarcando desde la preparación inicial hasta la implementación de modelos avanzados, con un fuerte énfasis en la interpretación y visualización de resultados.

RESULTADOS

El “Retail En Linea” que tuvimos el gusto de analizar tiene presencia en más de 30 países, genera ventas de más de €700,000 euros al mes y tiene una basta cartera de clientes.

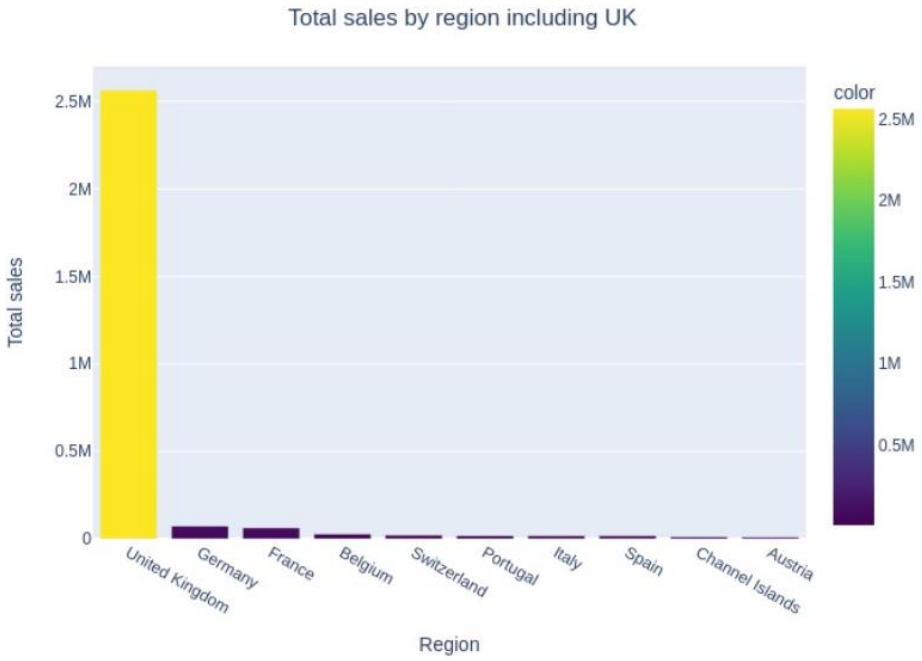


En el análisis nos dimos cuenta de que la mayor cantidad de ventas se genera en el Reino Unido, ya que esta región participa con más del 90% de las ventas. Además, podemos ver que la empresa tiene clientes no sólo en Europa, sino que también llega a América, África, Oceanía y Asia.



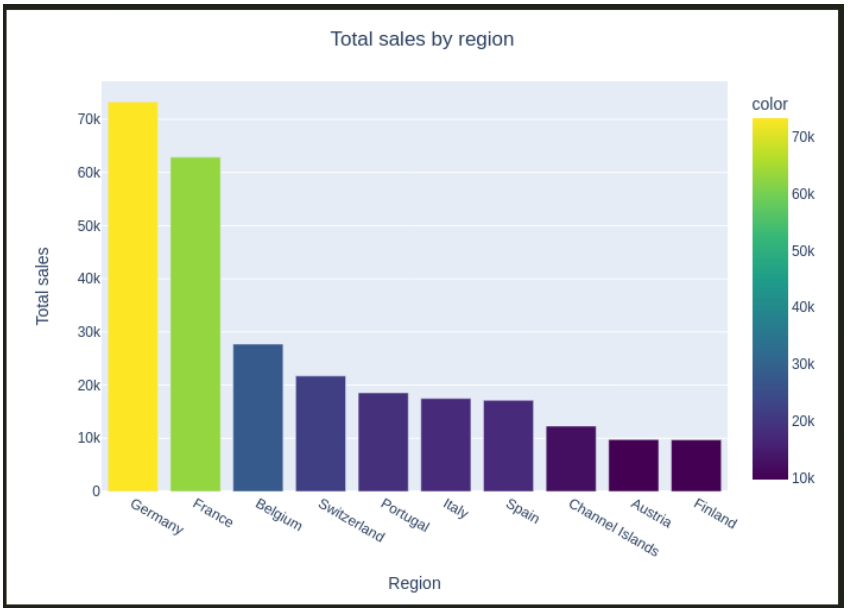
El grafico nos muestra que las ventas de la empresa van mejorando mes con mes, si observamos la venta de diciembre de 2019 podemos ver que fue de menos de €600,000 euros, si la comparamos con diciembre de 2020 el valor es muy similar sin embargo lamentablemente para diciembre de 2020 sólo tenemos datos del 1ro al 9 de diciembre, por lo que sin duda el cierre de diciembre del 2020 será mucho mejor que el del año pasado.

Ventas por Región:



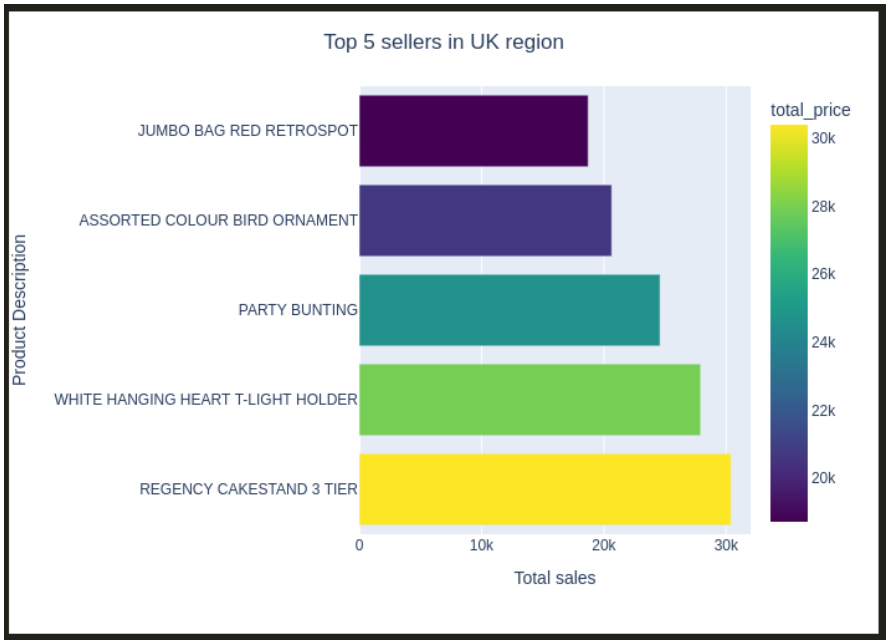
En esta gráfica podemos ver que el Reino Unido tiene la mayor cantidad de ventas, (este fue un determinante para dividir el análisis en 2 partes, la primera considerando las ventas de este país y la segunda analizando las ventas sin él. Además, podemos notar que el resto de los países con mayor venta son europeos. Lo que deja el resto de los continentes con una participación del 2%

Ventas Sin Reino Unido



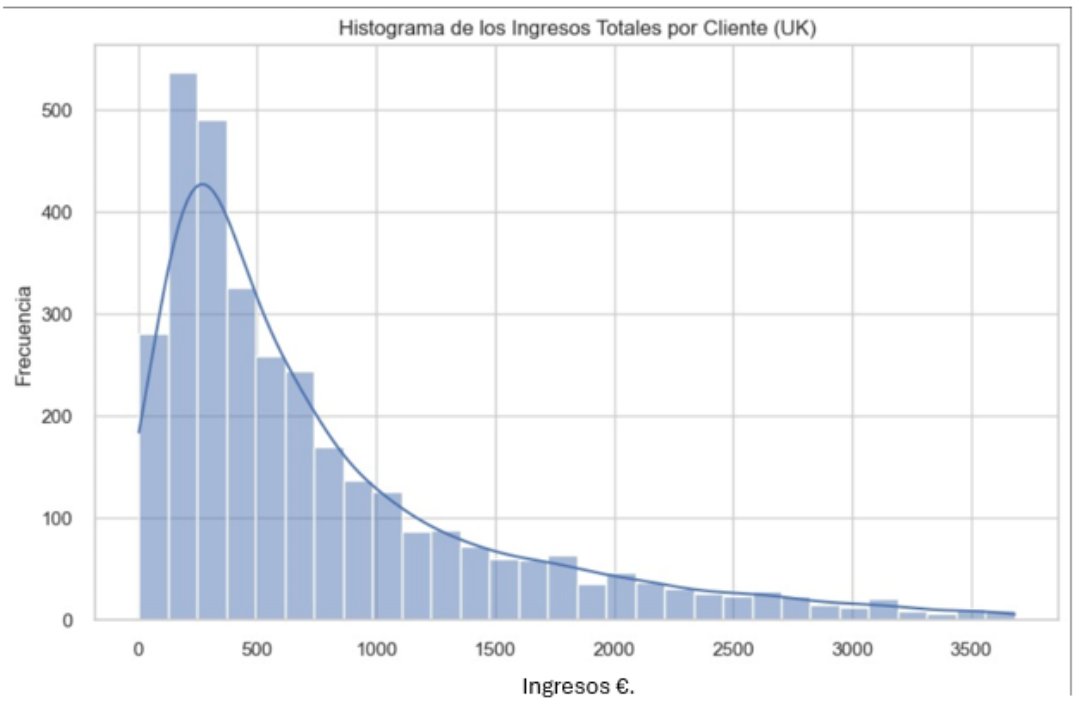
En este gráfico se puede apreciar mejor las ventas del top ten de los países (sin considerar al Reino Unido), como habíamos mencionado la mayor recaudación de la empresa proviene de países europeos, principalmente de Alemania, Francia quienes realizaron compras (en el periodo estudiado) por más de €70,000 y €60,000 euros, respectivamente

Top 5 productos más vendidos en el Reino Unido:



Este gráfico nos muestra los 5 productos más vendidos en el Reino Unido, siendo el artículo campeón (el más vendido) el “Regency Cakestand 3 tier”, el cual representa el 1.64% de la venta total de este país. Cabe mencionar que la empresa tiene en su stock desde tarjetas de felicitación hasta muebles.

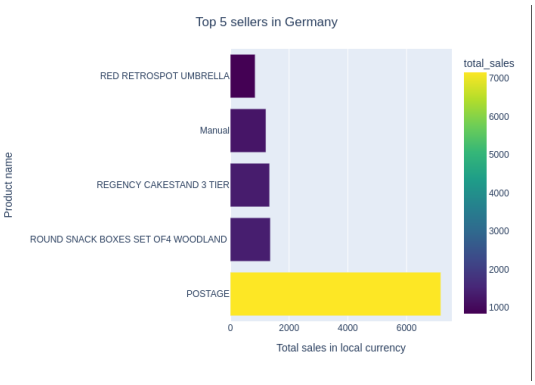
Precio promedio por Pedido en Reino Unido:



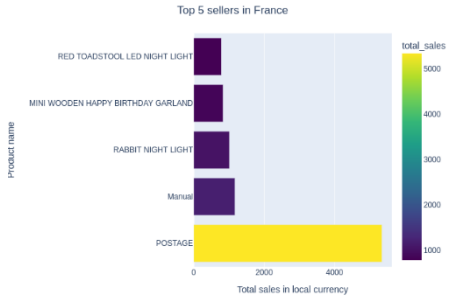
Con este gráfico queremos ilustrar cuanto están gastando nuestros clientes al adquirir cada pedido, como podemos ver, la mayoría de los pedidos cuestan en promedio €300 euros. Aunque tenemos clientes que hacen pedidos más grandes y llegan a gastar cantidades superiores a los €3,500 euros. Cabe mencionar que esta cantidad incluye el costo del envío.

Top 5 Por Región (País)

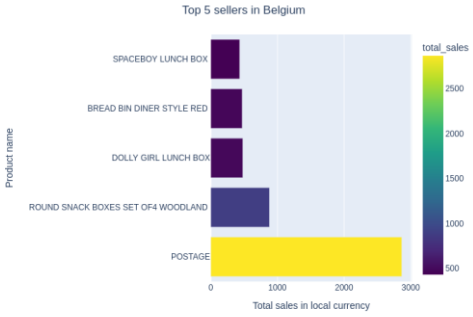
Alemania



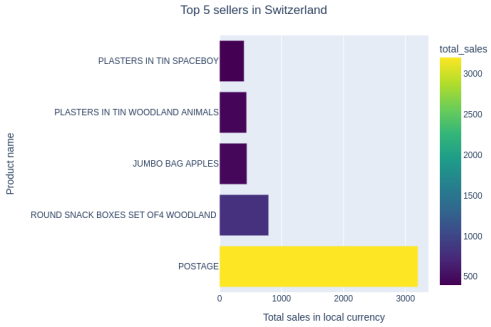
Francia



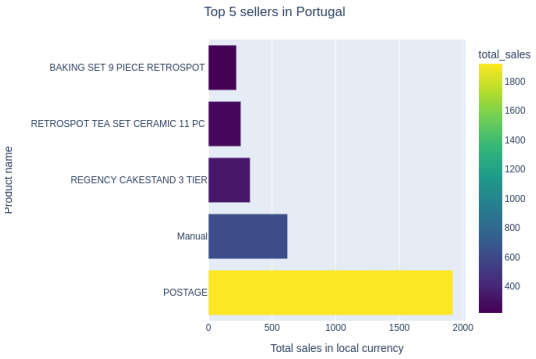
Belgica



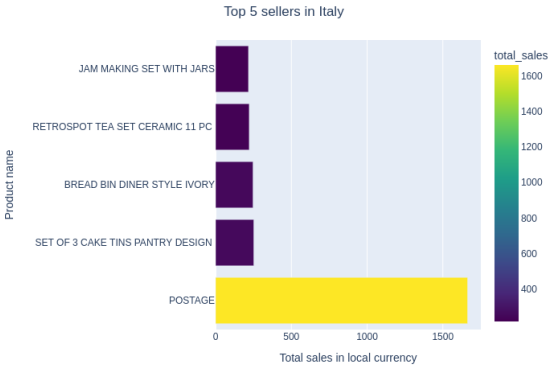
Suiza



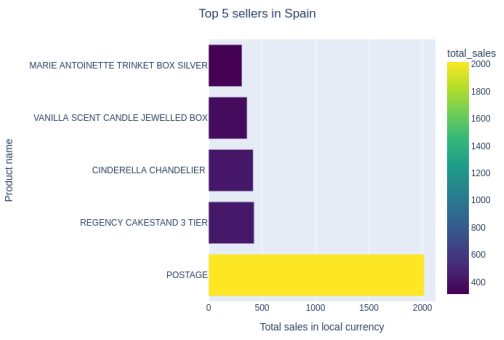
Portugal



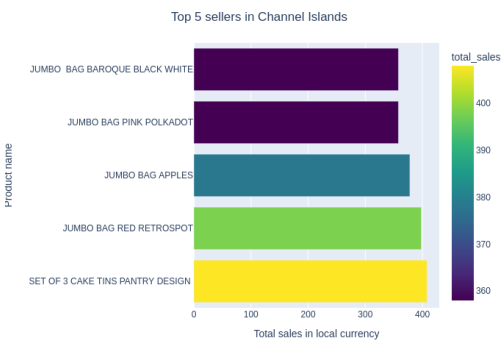
Italia



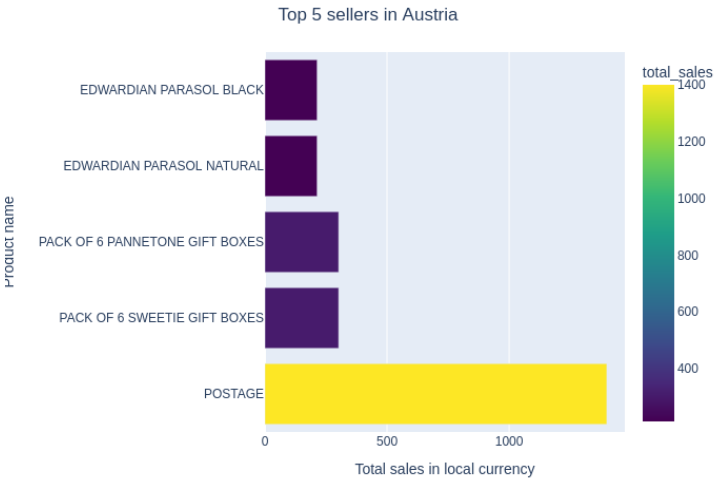
España



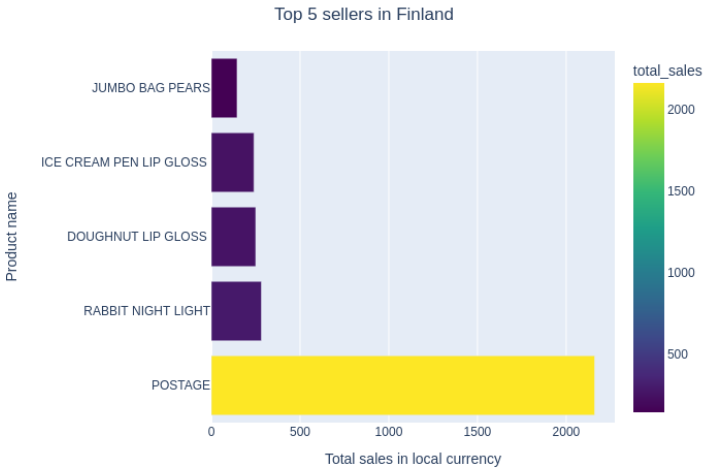
Islas Channel



Austria



Finlandia



De las gráficas anteriores podemos ver que en todos los países con excepción de las “Islas Channel” el mayor importe son los gastos de envío. Sin embargo podemos identificar que cada región tiene diferentes preferencias en los productos que solicita,

Por ejemplo, en Alemania, Bélgica, Portugal e Italia, la preferencia es sobre los artículos para almacenar o guardar alimentos.

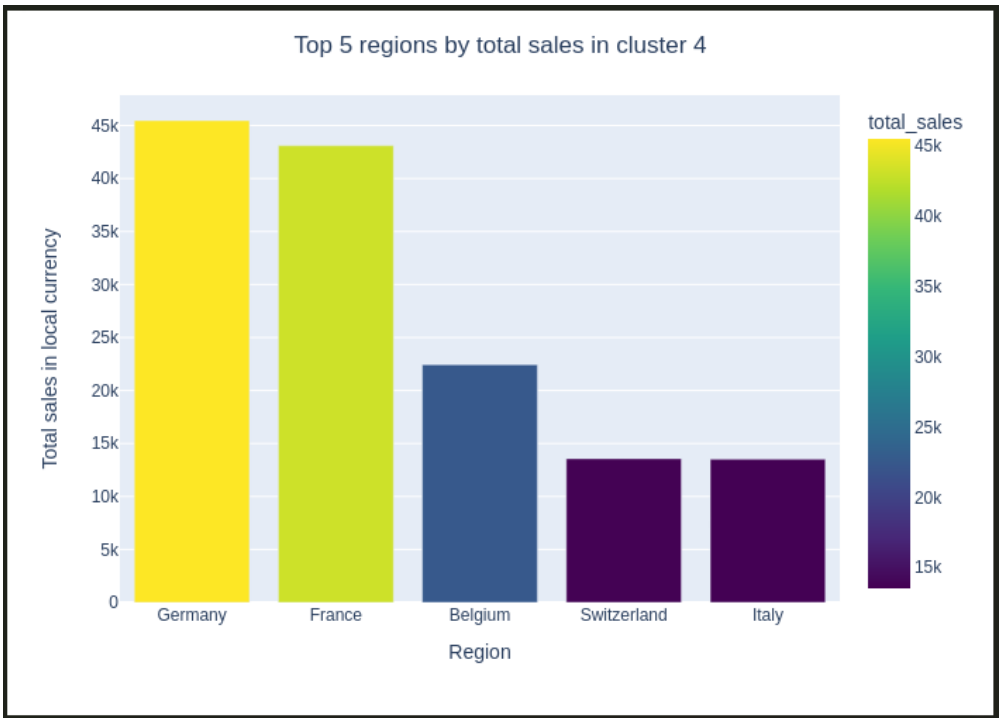
En Francia y Suiza, no se ve una tendencia clara a un tipo de producto, ya se que pidieron desde herramientas hasta artículos para decoración y juguetes.

España tiene mayor interés en artículos de decoración, mientras que en las Islas Channel pidieron mayormente bolsas jumbo de diferentes artículos.

En Austria la preferencia fue en artículo para protegerse del sol y finalmente en Finlandia el artículo más solicitado fue una lamparita de noche junto con labiales.

Cluster de Clientes:

Cluster 4: **Clientes de alto valor**



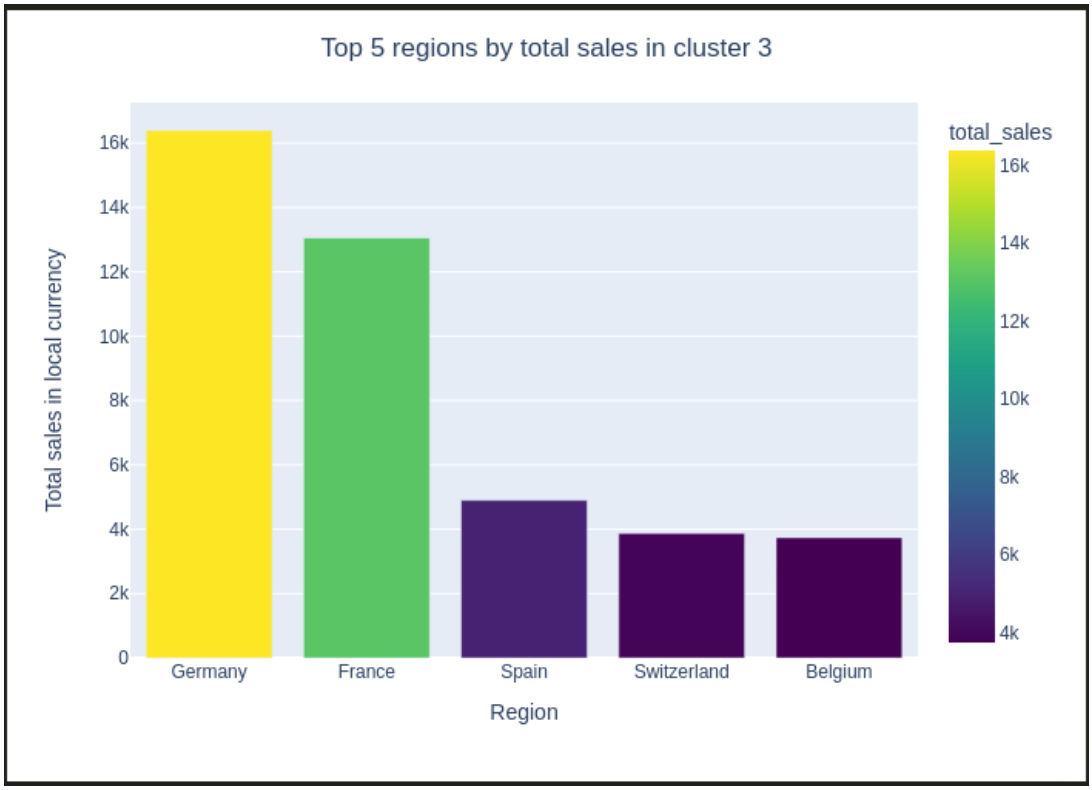
Características: Son clientes muy activos, con compras de alta frecuencia y pedidos de alto valor económico.

Estrategia: Se recomienda la Retención y maximización del valor de los pedidos que realizan.

Acciones: Se recomiendan estas 3 acciones:

- ✓ Implementar programa VIP exclusivo,
- ✓ Ofrecer acceso anticipado a nuevos productos,
- ✓ Proporcionar servicio al cliente dedicado.

Cluster 3: **Clientes activos de valor medio**



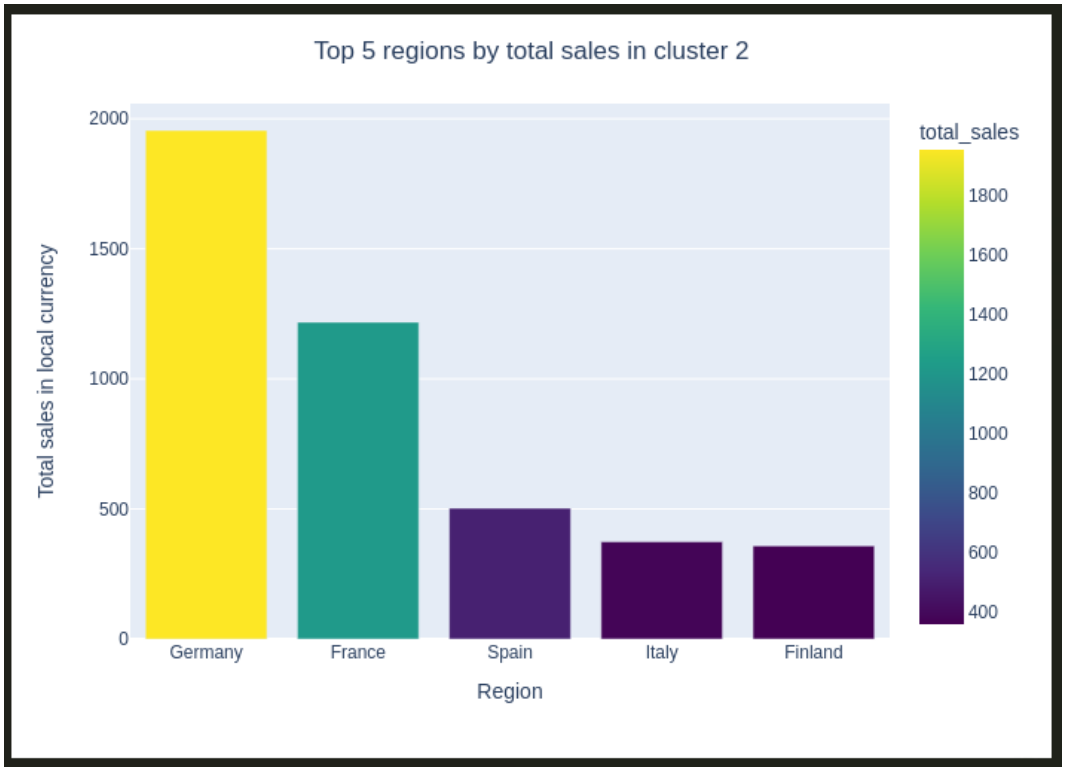
Características: Son clientes Activos, con compras de frecuencia media, y pedidos de valor económico medio.

Estrategia: Se recomienda incentivar un incremento de valor por compra

Acciones: Se recomiendan estas 3 acciones:

- ✓ Crear programa de escalado de compras,
- ✓ Ofrecer paquetes de productos para aumentar el valor por transacción,
- ✓ Desarrollar contenido educativo sobre productos premium.

Cluster 2: **Clientes recientes de bajo valor**



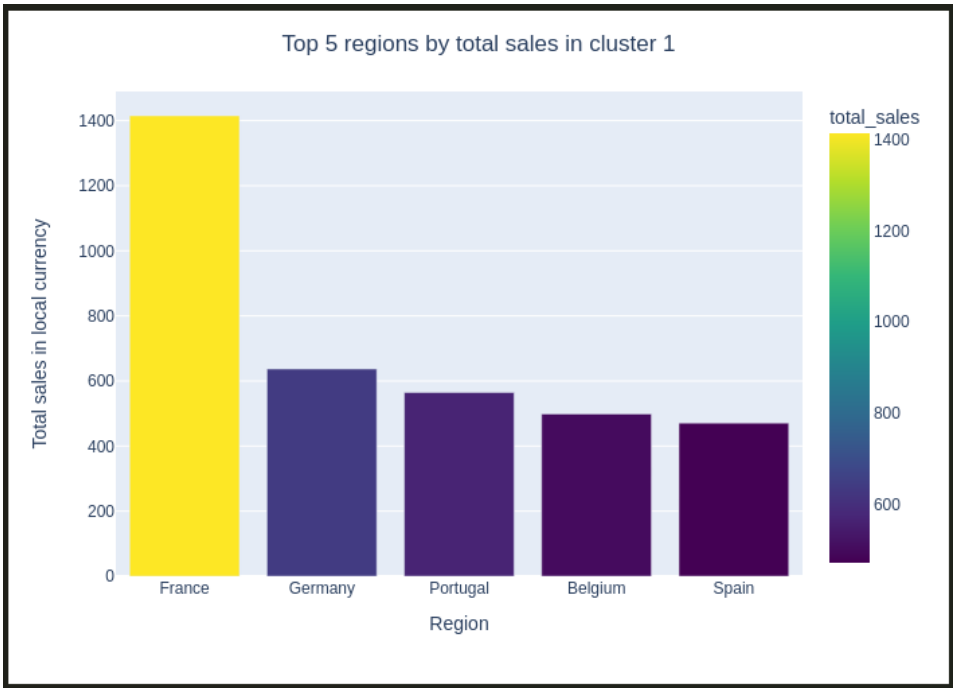
Características: Son clientes Recientes, con compras de baja frecuencia, y pedidos de valor económico bajo.

Estrategia: Se recomienda incentivar un aumento de valor y frecuencia en sus compras

Acciones: Se recomiendan estas 3 acciones:

- ✓ Lanzar campaña de "segunda compra" con incentivos,
- ✓ Implementar sistema de recomendaciones personalizadas,
- ✓ Introducir programa de fidelización básico.

Cluster 1: **Clientes de valor medio con baja recencia**



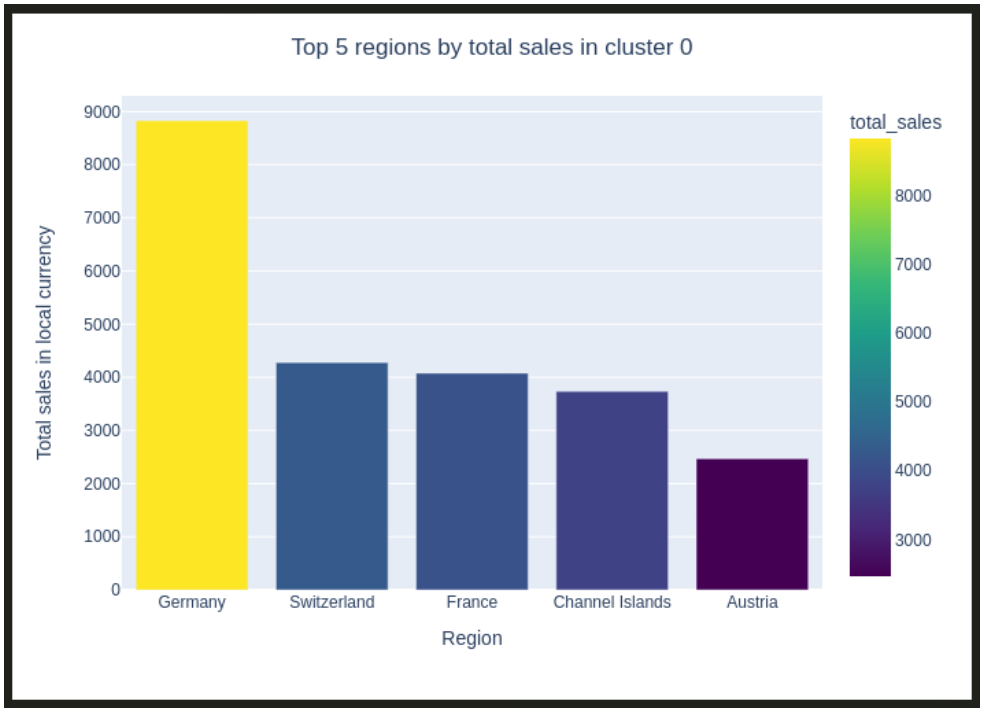
Características: Son clientes con baja “recencia”, con compras de frecuencia media, y pedidos de valor económico medio.

Estrategia: Se recomienda realizar una reactivación personalizada.

Acciones: Se recomiendan estas 3 acciones:

- ✓ Desarrollar campaña "te extrañamos" con ofertas especiales
- ✓ Enviar recordatorios de productos previamente comprados
- ✓ Realizar encuesta para entender razones de inactividad

Cluster 0: **Clientes inactivos de bajo valor**



Características: Son clientes con muy baja “recencia”, con compras de baja frecuencia, y pedidos de valor económico bajo.

Estrategia: Se recomienda realizar recuperación y reactivación de los clientes.

Acciones: Se recomiendan estas 3 acciones:

- ✓ Lanzar campaña de "última oportunidad" con descuentos significativos,
- ✓ Implementar programa de reactivación por etapas,
- ✓ Evaluar ROI de marketing para este grupo.

Desempeño de métricas para modelos de Machine Learning [%]

	Precisión	Sensibilidad	F1-Score	Accuracy	AUC-ROC	time_train	time_test
Random Forest	97.41	97.3	97.34	97.27	99.94	0.31	0.01
Gradient Boosting	96.95	96.55	96.72	96.73	99.89	1.71	0.01
Voting	96.07	95.43	95.72	95.5	99.79	2.73	0.12
Decision Tree	95.98	95.43	95.67	95.5	97.13	0.01	0
MLP	93.5	92.69	92.88	92.63	99.54	0.78	0
Naive Bayes	91.88	92.87	92.3	92.36	99.19	0	0
KNN	88.7	87.46	87.95	88.68	97.52	0	0.02
SVM	83.38	82.37	82.78	84.04	97.29	0.66	0.1

En esta imagen podemos observar el comparativo del desempeño de las métricas de los modelos probados, el modelo “Random Forest” fue el que tuvo el mejor desempeño superando el 97% de precisión para las predicciones, seguido de él, encontramos como alternativa el Gradient Boosting que nos ofrece métricas similares y tiene ventajas sobre tiempo de procesamiento.

CONCLUSIONES Y RECOMENDACIONES

Conclusiones.

La tienda tiene un alcance mundial, realizado ventas en países de los 5 continentes

Las ventas que se producen por mes en la tienda alcanzan hasta 700.000 euros al mes y en general, este aumenta a lo largo del año.

Se reconoce a Reino Unido (UK) como potencia, abarcando más del 90% de todas las ventas de la compañía presentes en la base de datos.

El continente Europeo representa el 98% de las ventas/ingresos logrados por la tienda. Solo un 2% queda representado por los demás países.

El producto más vendido dentro de UK corresponde a “Regency Cakestand 3 tier”, y dentro de los demás países europeos como Alemania, francia, Bélgica, Portugal, Italia, España, entre otros, corresponde a “Postage”.

Se reconoce 5 cluster de clientes (0-4), donde 0 indica a clientes menos frecuentes y de menor valor; y 4 indica clientes muy frecuentes y de gran valor. Los clusters 3 y 4 (clientes activos y de alto valor) representan la mayor parte de la base de clientes (57.8%). Hay una clara diferenciación entre los segmentos en términos de comportamiento de compra. La recencia parece ser un factor importante en la segmentación, con una clara distinción entre clientes activos e inactivos.

Recomendaciones

De manera general:

- Buscar la expansión de mercado hacia latinoamérica, considerando que ya se tienen embarcaciones que llevan producto tanto a Norteamérica como América del sur, se podría ingresar a mercados como México, Chile, Argentina, etc.
- Buscar a través de la transformación digital, mejorar las bases de datos para incluir información que sea útil para las métricas del negocio, como la categoría del producto, costos de envío, continentes, etc.
- Crear programas para incentivar y fidelizar a los clientes, considerando los tipos de cliente, las costumbres de consumo y la estacionalidad de las ventas.
- Incursionar en sistemas de calificación y reseñas de usuarios para implementar análisis de sentimientos y un soporte al cliente, buscando mejorar la satisfacción del cliente y su recomendación.
- Implementar estrategias de venta como upselling y cross selling para incentivar las ventas de productos no tan populares, pero a través del sistema de recomendaciones que concrete la venta.

Para clusters:

- Para cluster 4 se recomienda principalmente retencion implementando programas VIP y servicio al cliente dedicado.

- Para cluster 3 se recomienda principalmente crear un programa de escalado de compras ofreciendo paquetes de productos y contenido premium.

- Para cluster 2 se recomienda implementar campañas de ‘segunda compra’, recomendaciones personalizadas y programas de fidelización

- Para cluster 1 se recomienda desarrollar campaña ‘te extrañamos’, enviar recordatorios y encuestas.

- Para cluster 0 se recomienda campañas ‘última oportunidad’ y evaluar ROI.