Análisis de Datos de Comercios Electrónicos en Pakistán entre Marzo de 2016 y Agosto de 2018

Elaborado por Camilo Herradora

Abstracto

Este informe presenta un análisis detallado de los registros de compras de comercios electrónicos en Pakistán entre 2016 y 2018. El objetivo principal detrás de este estudio y de tomar como referencia esta información es comprender acerca de los patrones de compras de los clientes, ver cómo se puede realizar una segmentación de clientes, así como identificar tendencias y mayores frecuencias entre un conjunto de datos determinado. Además de destacar el lenguaje de programación utilizado, python, así como las librerías relevantes más utilizadas (como pandas, matplotlib y sns), se ha de destacar la metodología de análisis RFM (Recency, Frequency, Monetary) utilizada, que oportunamente, se aprovecha de los datos históricos transaccionales, de tal forma que se ha podido instanciar el comportamiento de clientes hasta la fecha, permitiendo demostrar de manera gráfica una visión de las compras realizadas a manera de segmentos. Posteriormente hemos podido encontrar que en los segmentos que más desembolsan por compra, se destacan las categorías tecnológicas, como las tablets y teléfonos inteligentes, y también se ha identificado un pico de órdenes de este tipo a finales de cada año (noviembre - diciembre) de 2016 y 2017 respectivamente. Gracias a los detalles que nos brinda el análisis, en un futuro se podría determinar mejoras de cara a las estrategias de marketing para los comercios que oferten las categorías más deseadas por los clientes, o la manera en la que estos mismos se pueden acercar al segmento masivo, aquel que no desembolsa tantas cantidades pero que son clientes frecuentes.

Introducción

Gracias al creciente avance tecnológico que ha impactado lados que parecían remotos como es el caso de medio oriente, se ha podido evidenciar desde otras fuentes que efectivamente la cotidianidad detrás de actividades comerciales va a ir en aumento. Es precisamente por este conocimiento que se ha vuelto un desafío para comerciantes retailers saber cómo aproximarse de forma más eficiente y eficaz al cliente. Por otra parte, a medida que más usuarios optan por utilizar plataformas de comercio electrónico, las grandes empresas tecnológicas abordan una nueva preocupación, y es que al igual que los comercios, deben preocuparse por unificar un proceso propio de marketing, en el que mejoran la experiencia de usuario sabiendo por qué segmentos de clientes manejar el flujo de ventas y qué clase de ofertas deben mostrar.

El propósito de este informe es examinar los datos de muestra de un conjunto de transacciones de comercio electrónico, tratando de enfocar un análisis en las tendencias de compras realizadas, la segmentación de clientes y las correlaciones entre variables de los registros. Cabe destacar que no solo se buscan estas estadísticas en la información, sino que también sirven en un ejercicio de guía y estrategia para futuros comercios en la retención de clientes y adquisición de nuevos usuarios.

Revisión de Literaturas (Antecedentes)

En dos informes claves, podemos encontrar qué factores han beneficiado e impulsado el tráfico en línea en Pakistán, así como podemos ver qué factores perjudican este tráfico y el potencial de crecimiento para el comercio electrónico en Pakistán. ***Zeshan Muhammad (2023), E-commerce and its potential in Pakistan****,* es un documento de parte del instituto de desarrollo económico de Pakistán en el que se expone de manera muy clara el gran mercado que se está construyendo en el país asiático, así también contrastando los grandes retos que presentan, tales como la confianza del consumidor, la infraestructura digital en la que se construyen las plataformas de comercio electrónico, la logística deficiente y la nula o deficiente regulación y políticas de cara a los posibles delitos promovidos en internet. Se concluye de manera clara, que a pesar de las barreras detalladas en la población pakistaní, existen grandes oportunidades de crecimiento, evidenciadas por los 6.4 mil millones de dólares en ingresos para este rubro en 2023, así como la proyección anual esperada del 6.23%. Partiendo de las pequeñas empresas y negocios, aprovechando las redes sociales como plataformas que impulsen el comercio electrónico, así como la potenciación económica de aquellas empresas dispuestas a atender sujetos diásporas en el extranjero.

Por otro lado, tenemos al documento publicado por ***China Economic Net (Agosto 30, 2024), Pakistan a land bestowed with broad prospects for e-commerce development.*** En dicho informe se nos aborda las grandes bases actuales en las que se sostiene el crecimiento del comercio electrónico en el país de medio oriente. Primeramente se expone que *"en 2023, los ingresos del sector alcanzaron los 5.2 mil millones de dólares, y se espera que crezcan a una tasa anual compuesta de 5.92%, proyectando ingresos de hasta 6.711 mil millones de dólares para 2029."* También cabe destacar las estadísticas expuestas acerca del aproximamiento de la población a estas plataformas, y es que también se nos menciona *"el 80% de los usuarios de internet está accediendo a través de smartphones y con un 58% de las compras en línea realizadas mediante*

dispositivos móviles en 2023. La popularidad de las aplicaciones de compra en línea ha crecido, superando los 16.6 millones de usuarios activos mensuales en julio de 2024, reflejando la aceptación de los consumidores hacia las compras a través de apps."

El comercio electrónico en definitiva, muestra una perspectiva esperanzadora y de crecimiento. Con el alza del uso de dispositivos móviles, el apoyo gubernamental, y la influencia de las redes sociales, el sector se perfila como un actor importante en la economía digital global.

Ya que hemos visto qué fortalezas y debilidades existen de cara al crecimiento de este rubro, veremos qué tanto se refleja el interés en la tecnología y los picos de consumo existentes en portales de e-commerce por medio de este análisis.

Colección de Datos y Preprocesamiento

Los datos provienen en una estructura organizada “dataset” de kaggle publicada en 2020, Zeeshan-ul-hassan Usmani nos brinda el trasfondo, y es que recopila una colección de datos de órdenes históricos entre marzo de 2016 a agosto de 2018. Contiene 500 mil registros transaccionales recolectados de varios comerciantes en plataformas digitales. Abarca una cantidad total de 21 variables, incluyendo ID, estatus de la orden, fecha de la orden, SKU, precio, cantidad, cantidad total, categoría, método de pago y ID del cliente.

Primeramente, empezamos con un proceso de exploración de datos, visualizando su estructura:

```
[33]: df = pd.read_csv(file_name)
      df.head()
```

```
C:\Users\swamt\AppData\Local\Temp\ipykernel_2348\2994562364.py:1: DtypeWarning: Columns (1,2,3,7,8,9,11,12,13,14,17,18,19) have mixed types. Specify dtype option on import or set low_memory=False.
  df = pd.read_csv(file_name)
```

[33]:

	item_id	status	created_at	sku	price	qty_ordered	grand_total	increment_id	category_name_1	sales_commission_code	...	Month	Customer Since	M-Y	FY	Customer ID	Unnamed: 21
0	211131.0	complete	7/1/2016	kreations_YI 06-L	1950.0	1.0	1950.0	100147443	Women's Fashion	\N	...	7.0	2016-7	7-2016	FY17	1.0	NaN
1	211133.0	canceled	7/1/2016	kcc_Buy 2 Frey Air Freshener & Get 1 Kasual Bo...	240.0	1.0	240.0	100147444	Beauty & Grooming	\N	...	7.0	2016-7	7-2016	FY17	2.0	NaN
2	211134.0	canceled	7/1/2016	Ego_UP0017-999-MR0	2450.0	1.0	2450.0	100147445	Women's Fashion	\N	...	7.0	2016-7	7-2016	FY17	3.0	NaN
3	211135.0	complete	7/1/2016	kcc_krone deal	360.0	1.0	60.0	100147446	Beauty & Grooming	R-FSD-52352	...	7.0	2016-7	7-2016	FY17	4.0	NaN
4	211136.0	order_refunded	7/1/2016	BK7010400AG	555.0	2.0	1110.0	100147447	Soghaat	\N	...	7.0	2016-7	7-2016	FY17	5.0	NaN

5 rows × 26 columns

Encontramos una cantidad enorme de datos nulos, y los procedemos a eliminar:

```
[5]: #Dado que la exploración inicial con pandas muestra 5 columnas que en verdad no existen en el dataset original, procederemos a borr
     deleteCols =  ['Unnamed: 21', 'Unnamed: 22', 'Unnamed: 23', 'Unnamed: 24', 'Unnamed: 25']
     new_df = df.drop(columns = deleteCols)
```

Una vez realizada esta limpieza inicial, identificamos que existen pocos valores nulos por variables, no relevantes para el estudio:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1048575 entries, 0 to 1048574
Data columns (total 21 columns):
 #   Column                 Non-Null Count   Dtype
---  ------                 --------------   -----
 0   item_id                584524 non-null  float64
 1   status                 584509 non-null  object
 2   created_at             584524 non-null  object
 3   sku                    584504 non-null  object
 4   price                  584524 non-null  float64
 5   qty_ordered            584524 non-null  float64
 6   grand_total            584524 non-null  float64
 7   increment_id           584524 non-null  object
 8   category_name_1        584360 non-null  object
 9   sales_commission_code  447346 non-null  object
 10  discount_amount        584524 non-null  float64
 11  payment_method         584524 non-null  object
 12  Working Date           584524 non-null  object
 13  BI Status              584524 non-null  object
 14   MV                    584524 non-null  object
 15  Year                   584524 non-null  float64
 16  Month                  584524 non-null  float64
 17  Customer Since         584513 non-null  object
 18  M-Y                    584524 non-null  object
 19  FY                     584524 non-null  object
 20  Customer ID            584513 non-null  float64
dtypes: float64(8), object(13)
memory usage: 168.0+ MB


Cantidad de filas y columnas de datos
(1048575, 21)
```

new_df.info()

print('\n\nCantidad de filas y columnas de datos')
new_df.shape

Metodologías

Se ha elegido visualizar las tendencias en las compras para las categorías existentes en el dataset, las frecuencias de compras por mes en 2016, 2017 y 2018, las frecuencias de métodos de pago usados y los estados de órdenes más comunes (entre completados, cancelados, reembolsados.

Dichas elecciones de estudio no son casuales, sino que evocan a querer comprender el patrón de comportamiento que pueden tener los clientes en cuanto a las correlaciones entre estas variables se refiere, queremos saber qué es lo que más se consume en pakistán, qué tanto se desembolsa, cada cuánto se desembolsa y qué categoría es la que más se elige. De esta manera podemos saber qué clase de target es el que sobresale por encima del resto, y así construir segmentos de clientes básicos.

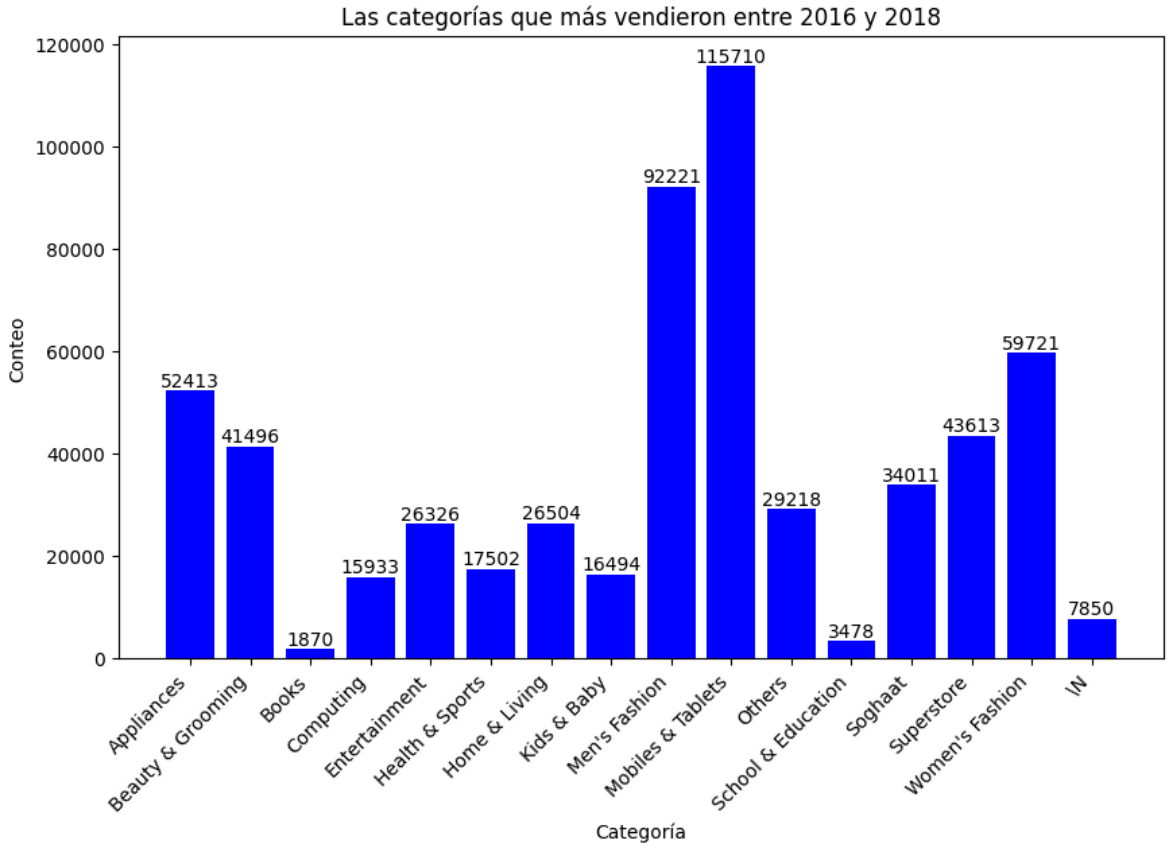
Una vez tenemos este conjunto organizado denominado “segmentos”, veremos cómo plantarlo a nivel de código de python. La idea inicial es que usando la Recencia, Frecuencia y Valor monetario, podemos tratar al dataset como instancias de comportamiento del cliente, y tratar únicamente los números de la compra (exactamente el total de la compra), posterior a esto, podemos definir una fecha de referencia para definir qué es reciente en comparación a las fechas dadas en el dataset. Una vez esto definimos indicadores para cada registro y a partir de ahí es que se conforman 3 segmentos base.

Resultados

Veremos tendencias de categorías en el historial de compras

Código

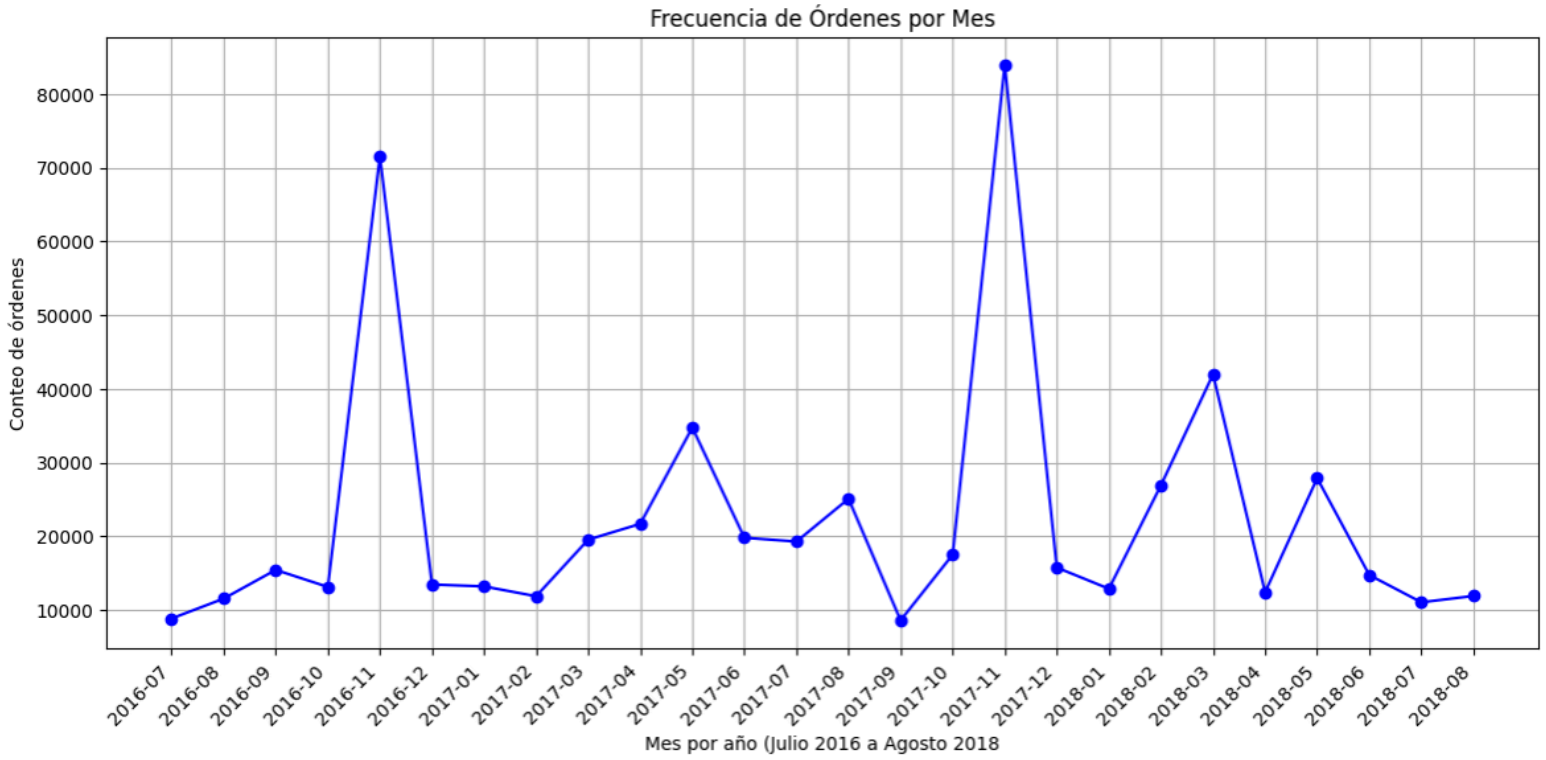
```
#Gráfico de barras con dimensiones 10 x 6
plt.figure(figsize=(10, 6))
catbar = plt.bar(category_counts.index, category_counts.values, color = 'blue')
plt.title('Las categorías que más vendieron entre 2016 y 2018')
plt.xlabel('Categoría')
plt.ylabel('Conteo')
plt.xticks(rotation=45, ha='right')
```



Veremos las tendencias de ventas en correspondencia con el mes de cada año

Código

```
#Usaremos la fecha original "Working Date" convirtiendola a fecha
new_df['Working Date'] = pd.to_datetime(new_df['Working Date'])

#Extraemos el mes
new_df['Month'] = new_df['Working Date'].dt.to_period('M')

#Frecuencia de registros de órdenes
bymonth_orders = new_df.groupby('Month').size()

plt.figure(figsize=(14, 6))
plt.plot(bymonth_orders.index.astype(str), bymonth_orders.values, marker='o',
linestyle='-', color='b')
plt.title('Frecuencia de Órdenes por Mes')
plt.xlabel('Mes por año (Julio 2016 a Agosto 2018')
plt.ylabel('Conteo de órdenes')
plt.xticks(rotation=45, ha='right')
plt.grid(True)
```



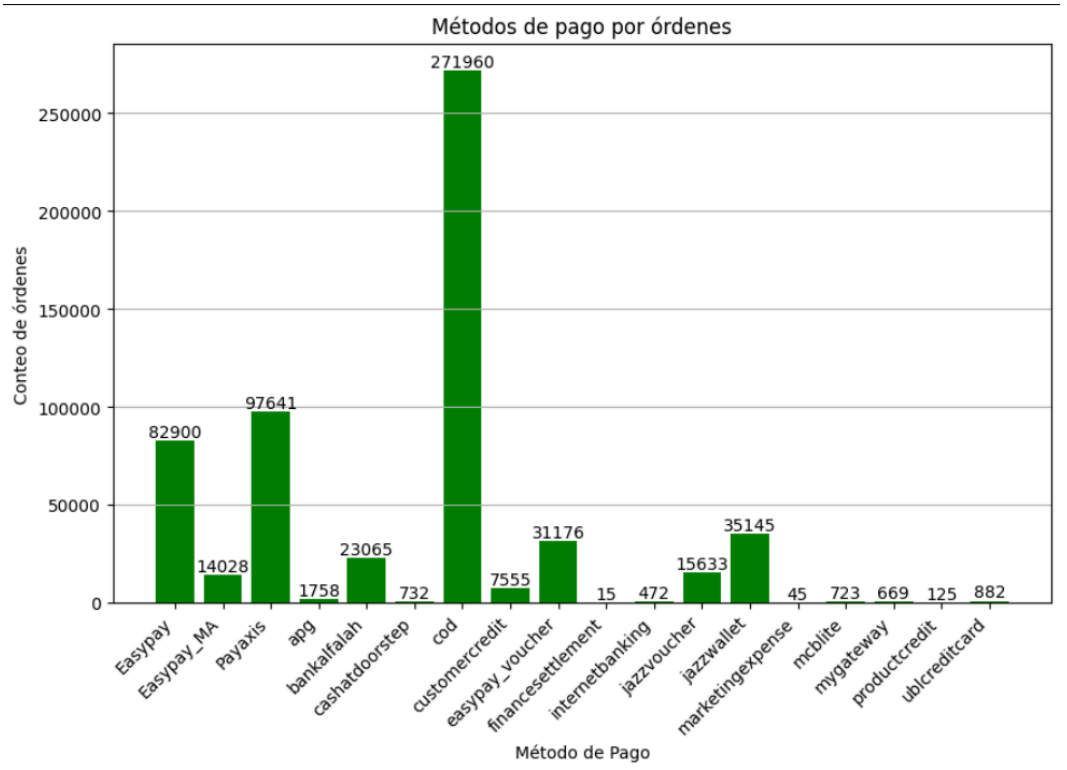
Ahora contabilizaremos los métodos de pago por órdenes

Código

#Para poder contabilizar métodos de pago por cada orden

payment_methods_count =
new_df['payment_method'].value_counts().sort_index()
payment_methods_count.head()

plt.figure(figsize=(10, 6))
payments_bar = plt.bar(payment_methods_count.index,
payment_methods_count.values, color='green')
plt.title('Métodos de pago por órdenes')
plt.xlabel('Método de Pago')
plt.ylabel('Conteo de órdenes')



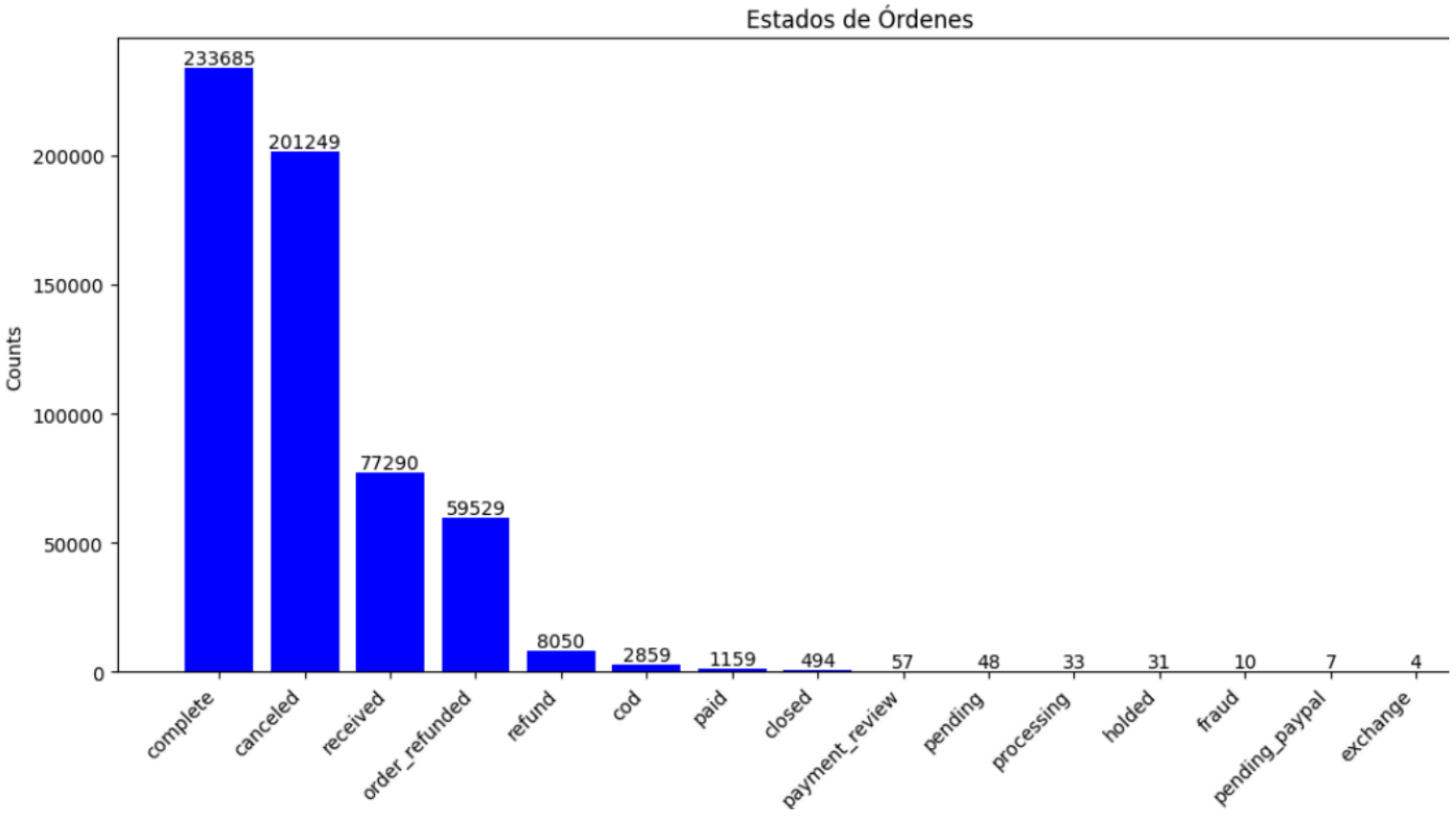
plt.xticks(rotation=45, ha='right')
plt.grid(axis='y')

get_yval(payments_bar)

Estatus de Órdenes

Código

```
#Conteo de los estatus de órdenes
status_counts = new_df['status'].value_counts()

plt.figure(figsize=(14, 6))
bargraph = plt.bar(status_counts.index, status_counts.values, color = 'blue')
plt.title('Estados de Órdenes')
plt.xlabel('Estado de Orden')
plt.ylabel('Counts')
plt.xticks(rotation=45, ha='right')

get_yval(bargraph)
```

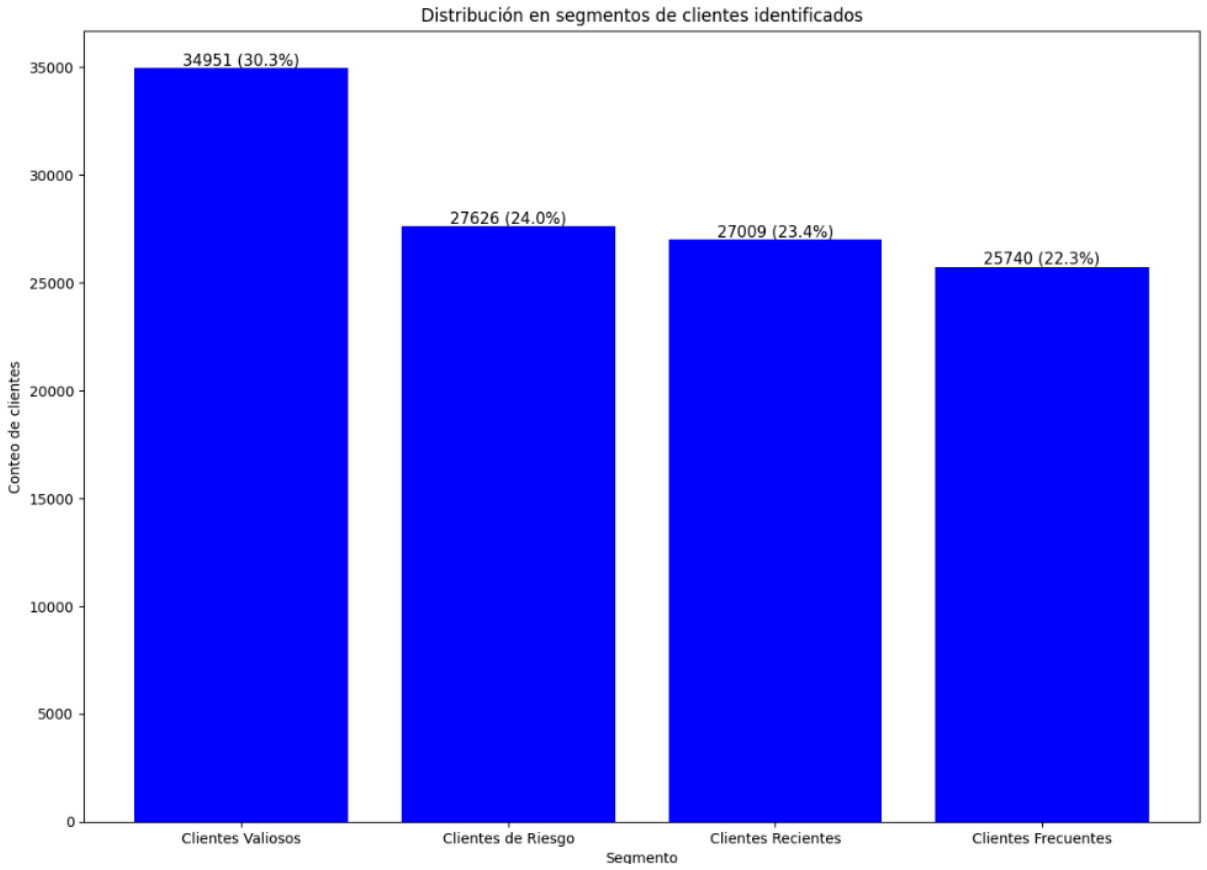


Segmentación de clientes

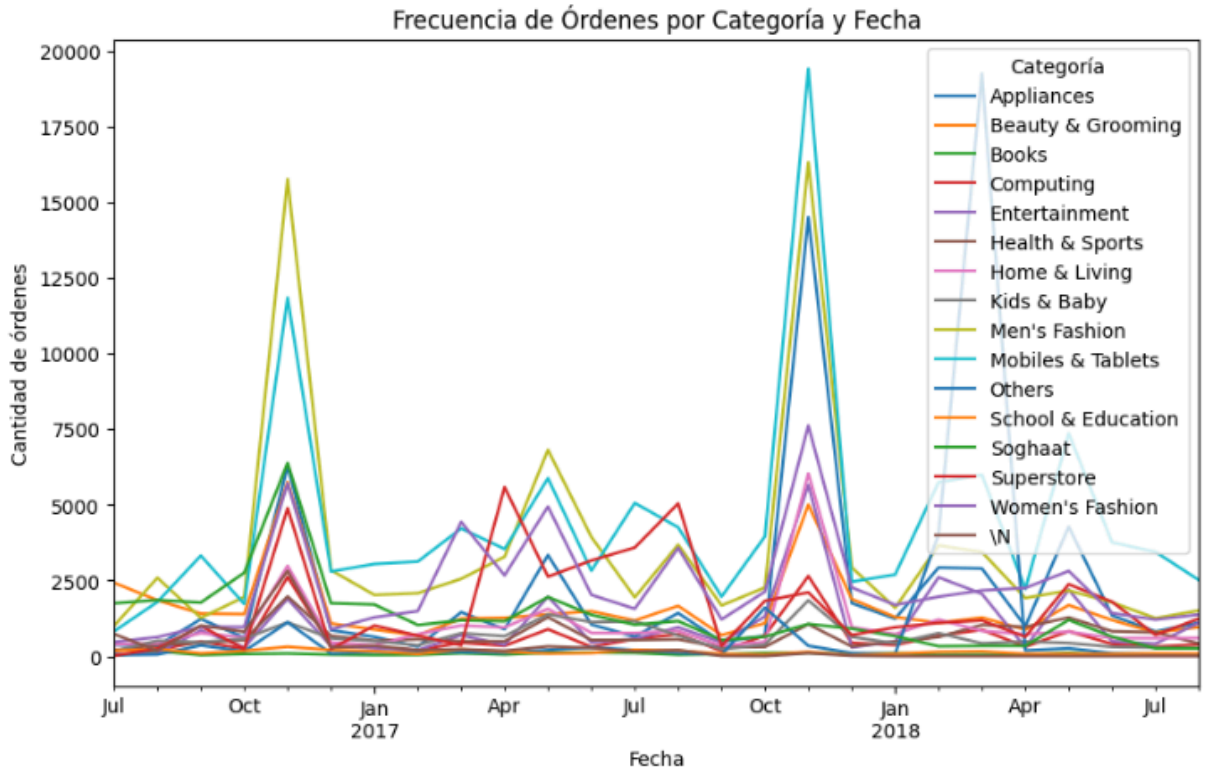
Código

```
#Para realizar dicho análisis tomaremos como referencia el ID del cliente, la fecha de creación de la orden y el monto total de dicha orden
new_df['created_at'] = pd.to_datetime(new_df['created_at'])

"""
Si requerimos de una fecha reciente como referencia, tomaremos en cuenta que estos datos abarcan hasta agosto de 2018, por lo tanto
Definiremos una recencia del 01-07-2018
"""

recency_reference = pd.to_datetime('2018-07-01')

rfm = new_df.groupby('Customer ID').agg({
   'created_at': lambda x: (recency_reference - x.max()).days,
   'grand_total': ['sum', 'count'] })

rfm.columns = ['Recency', 'Monetary', 'Frequency']
rfm = rfm.reset_index()

#Creamos los scores que definen y puntúan a cada tipo de cliente, del 1 al 4

rfm['R_score'] = pd.qcut(rfm['Recency'], 4, labels=[4, 3, 2, 1])
rfm['F_score'] = pd.qcut(rfm['Frequency'].rank(method="first"), 4, labels=[1, 2, 3, 4])
rfm['M_score'] = pd.qcut(rfm['Monetary'], 4, labels=[1, 2, 3, 4])

#Unificamos los 3 scores
rfm['RFM_score'] = rfm['R_score'].astype(str) + rfm['F_score'].astype(str) + rfm['M_score'].astype(str)
```

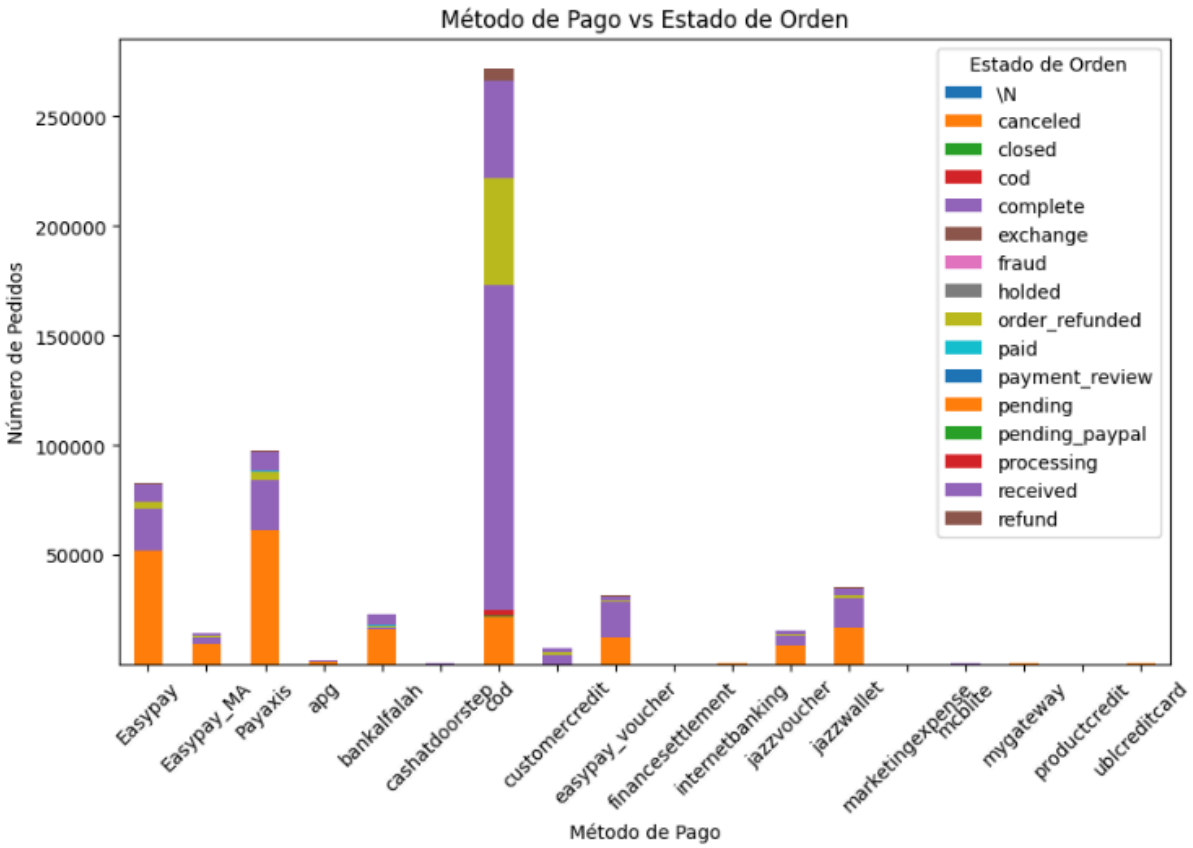
```
def segments(df):
    if df['RFM_score'] >= '344':
        return 'Clientes Valiosos'
    elif df['RFM_score'] >= '244':
        return 'Clientes Frecuentes'
    elif df['RFM_score'] >= '144':
        return 'Clientes Recientes'
    else:
        return 'Clientes de Riesgo'

rfm['Segment'] = rfm.apply(segments, axis = 1)

#Visualización de segmentos
segments_count = rfm['Segment'].value_counts()

segment_percentages = segments_count / segments_count.sum() * 100

plt.figure(figsize=(14,10))
segments_bar = plt.bar(segments_count.index, segments_count.values,
color='blue')
plt.title('Distribución en segmentos de clientes identificados')
plt.xlabel('Segmento')
plt.ylabel('Conteo de clientes')

for i, bar in enumerate(segments_bar):
    count = segments_count.values[i]
    percent = segment_percentages.values[i]
    plt.text(bar.get_x() + bar.get_width() / 2, bar.get_height() + 0.3,
          f'{count} ({percent:.1f}%)', ha='center', va='bottom', fontsize=11)
```




```
new_df['order_datetime'] = pd.to_datetime(new_df['created_at'])
category_date_counts =
new_df.groupby([new_df['order_datetime'].dt.to_period('M'),
'category_name_1']).size().unstack(fill_value=0)

category_date_counts.plot(kind='line', figsize=(10, 6))
plt.title('Frecuencia de Órdenes por Categoría y Fecha')
plt.xlabel('Fecha')
plt.ylabel('Cantidad de órdenes')
plt.legend(title='Categoría')
```




```
status_payment_counts = new_df.groupby(['payment_method',
'status']).size().unstack(fill_value=0)

plt.figure(figsize=(14, 12))
status_payment_counts.plot(kind='bar', stacked=True, figsize=(10, 6))
plt.title('Método de Pago vs Estado de Orden')
plt.xlabel('Método de Pago')
plt.ylabel('Número de Pedidos')
plt.legend(title='Estado de Orden')
plt.xticks(rotation=45)
```


Método de Pago vs Estado de Orden
Estado de Orden
\N
canceled
closed
cod
complete
exchange
fraud
holded
order_refunded
paid
payment_review
pending
pending_paypal
processing
received
refund
Número de Pedidos
250000
200000
150000
100000
50000
Easypay
Easypay_MA
Payaxis
apg
bankalfalah
cashatdoorstep
cod
customercredit
easypay_voucher
financesettlement
internetbanking
jazzvoucher
jazzwallet
marketingexpense
mcblite
mygateway
productcredit
ublcreditcard
Método de Pago

Discusión

Entre todos los datos identificados durante este estudio, es importante remarcar las tendencias de compras que existen en Pakistán, cada vez que se acerca la temporada de fin de año los picos de compra se disparan pero únicamente para todo ítem que se encuentre dentro de las categorías de tecnología.

Curiosamente, siendo otra de las categorías que más se definen en un pico de compras la moda varonil, pudiendo inferir que son los hombres quienes en medio oriente tienen una capacidad adquisitiva mayor al resto de la población.

También cabe destacar que a como lo mencionaban 2 estudios anteriormente acatados en este informe, el total de usuarios que frecuentan estas plataformas sigue siendo demasiado poco en contraste al total de la población de Pakistán, curiosamente los segmentos de clientes que más destacan en esta muestra son aquellos que gastan mayores cantidades y frecuentan mucho las compras en línea, el conjunto adinerado del país asiático. Apenas un 35% del total de 235 millones de habitantes posee habilidades digitales básicas y más de un 60% de pakistaníes desconfían totalmente de los pagos en línea.

Conclusiones

El análisis realizado en este informe ha permitido cumplir el objetivo de identificar y comprender cuáles son las tendencias de compras que existen, así como la segmentación de clientes y la examinación de correlación entre variables clave entre los registros de transacciones de comercio electrónico en Pakistán entre 2016 y 2018. Gracias a la segmentación de clientes y el análisis de estas tendencias, se logró presentar insights estratégicos para abordar las necesidades y preferencias de los perfiles de consumidores (predominando consumidores frecuentes y adinerados).

Estos hallazgos pueden ofrecer una base sólida y determinar estrategias de crecimiento para comercios retailers asiáticos que se alineen a la optimización y retención del cliente, así como su experiencia de usuario en interfaces digitales.

Referencias

- *Pakistan's largest E-Commerce dataset*. (2021, January 19). Kaggle. https://www.kaggle.com/datasets/zusmani/pakistans-largest-ecommerce-dataset/data
- Zeshan, M. Z. (2023). E-commerce and its potential in Pakistan. In *Pakistan Institute of Development Economics*. Pakistan Institute of Development Economics. Retrieved October 6, 2024, from https://pide.org.pk/research/e-commerce-and-its-potential-in-pakistan/
- Gràcia, A. (2023, January 2). Análisis RFM: ¿qué es y cómo puedes utilizarlo en retail? *Status2*. https://status2.com/analisis-rfm/#:~:text=El%20an%C3%A1lisis%20RFM%20(por%20sus,en%20el%20comportamiento%20hist%C3%B3rico%20transaccional.
- Bo, F. (n.d.). *Pakistan a land bestowed with broad prospects for e-commerce development--China Economic Net*. http://en.ce.cn/Insight/202408/30/t20240830_39122999.shtml#:~:text=Pakistan's%20e%2Dcommerce%20revenue%20reached,the%20end%20of%20this%20period