Show code

Análisis de E-Commerce Brasileño – Olist

Análisis Retail y de Clientes (2016–2018)

Emilio Nahuel Pattini
Buenos Aires, Argentina
1 de febrero de 2026

Tabla de Contenidos

Introducción

Objetivo del Proyecto
Análisis orientado al negocio: ventas, comportamiento del cliente, RFM, cohorts, CLV, pronóstico básico + recomendaciones accionables.

Dataset
Dataset Público de Olist – ~100k órdenes (Kaggle)

Tech Stack

- Python: pandas, seaborn, plotly, matplotlib, prophet
- Dashboard: Power BI

---

1. Comprensión del Negocio y Preguntas

Preguntas clave:

- ¿Cuáles son las categorías / productos top por ingresos?
- ¿Retención y recurrencia de clientes?
- ¿Segmentación RFM?
- ¿Retención de cohorts a lo largo del tiempo?
- ¿Performance de entrega por región?
- ¿Impacto de métodos de pago?
- ¿Oportunidades de cross-sell?
- ¿Pronóstico para categorías top?

---

2. Carga y Visión General de los Datos

Esta sección cubre la ingestión inicial de los archivos CSV crudos del dataset de Olist y proporciona una comprensión de alto nivel de la estructura, tamaño y contenido de cada tabla. El objetivo es confirmar la carga exitosa, identificar las relaciones clave entre tablas y

detectar cualquier señal inmediata de calidad de datos antes de proceder con la limpieza y el análisis.

2.1. Carga de Datos

Cargo las tablas más relevantes del dataset de Olist (orders, order_items, customers, payments, reviews) utilizando pandas.
En esta etapa solo se cargan las tablas esenciales para mantener bajo el uso de memoria y enfocarse en las entidades principales necesarias para el análisis de ventas, clientes y logística.

```
Execution environment initialized successfully.
• Pandas version: 2.3.3
```

2.2. Visión General de los Datos y Chequeos Iniciales

Realizo una inspección rápida de cada tabla cargada para entender:

- Número de filas y columnas
- Tipos de datos
- Presencia de valores faltantes
- Filas de muestra

Este paso ayuda a mapear el esquema del dataset y decidir las prioridades de limpieza siguientes.

```
=== ORDERS ===
Rows: 99,441
Columns: 8

Data types and missing values:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99441 entries, 0 to 99440
Columns: 8 entries, order_id to order_estimated_delivery_date
dtypes: object(8)
memory usage: 6.1+ MB
None

First three rows:
```

	order_id	customer_id	order_status	orde
0	e481f51cbdc54678b7cc49136f2d6af7	9ef432eb6251297304e76186b10a928d	delivered	
1	53cdb2fc8bc7dce0b6741e2150273451	b0830fb4747a6c6d20dea0b8c802d7ef	delivered	
2	47770eb9100c2d0c44946d9cf07ec65d	41ce2a54c0b03bf3443c3d931a367089	delivered	

```
=== ORDER ITEMS ===
Rows: 112,650
Columns: 7

Data types and missing values:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 112650 entries, 0 to 112649
Columns: 7 entries, order_id to freight_value
dtypes: float64(2), int64(1), object(4)
memory usage: 6.0+ MB
None

First three rows:
```

	order_id	order_item_id	product_id	
0	00010242fe8c5a6d1ba2dd792cb16214	1	4244733e06e7ecb4970a6e2683c13e61	484
1	00018f77f2f0320c557190d7a144bdd3	1	e5f2d52b802189ee658865ca93d83a8f	dd7
2	000229ec398224ef6ca0657da4fc703e	1	c777355d18b72b67abbeef9df44fd0fd	5b5

```
=== CUSTOMERS ===
Rows: 99,441
Columns: 5

Data types and missing values:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99441 entries, 0 to 99440
Columns: 5 entries, customer_id to customer_state
dtypes: int64(1), object(4)
memory usage: 3.8+ MB
None

First three rows:
```

	customer_id	customer_unique_id	customer_zip_cod
0	06b8999e2fba1a1fbc88172c00ba8bc7	861eff4711a542e4b93843c6dd7febb0	
1	18955e83d337fd6b2def6b18a428ac77	290c77bc529b7ac935b93aa66c333dc3	
2	4e7b3e00288586ebd08712fdd0374a03	060e732b5b29e8181a18229c7b0b2b5e	

```
=== PAYMENTS ===
Rows: 103,886
Columns: 5

Data types and missing values:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 103886 entries, 0 to 103885
Columns: 5 entries, order_id to payment_value
dtypes: float64(1), int64(2), object(2)
memory usage: 4.0+ MB
None

First three rows:
```

	order_id	payment_sequential	payment_type	payment_installme
0	b81ef226f3fe1789b1e8b2acac839d17	1	credit_card	
1	a9810da82917af2d9aefd1278f1dcfa0	1	credit_card	
2	25e8ea4e93396b6fa0d3dd708e76c1bd	1	credit_card	

```
=== REVIEWS ===
Rows: 99,224
Columns: 7

Data types and missing values:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99224 entries, 0 to 99223
Columns: 7 entries, review_id to review_answer_timestamp
dtypes: int64(1), object(6)
memory usage: 5.3+ MB
None

First three rows:
```

	review_id	order_id	review_score	revi
0	7bc2406110b926393aa56f80a40eba40	73fc7af87114b39712e6da79b0a377eb	4	
1	80e641a11e56f04c1ad469d5645fdfde	a548910a1c6147796b98fdf73dbeba33	5	
2	228ce5500dc1d8e020d8d1322874b6f0	f9e4b658b201a9f2ecdecbb34bed034b	5	

---

3. Limpieza y Preparación de Datos

3.1. Limpieza Inicial de Datos

En esta fase realizo chequeos fundamentales de calidad de datos: conversión de columnas de timestamp a formato datetime adecuado, verificación de unicidad de identificadores clave (order_id, customer_id), e inspección de valores faltantes en columnas críticas. El

objetivo es asegurar que los datos crudos sean confiables y estén listos para el análisis sin introducir errores en cálculos o joins.

Conversiones de tipo

Convierto las cadenas de timestamp a objetos datetime para poder realizar cálculos basados en tiempo (deltas, agrupaciones por mes, etc.) de forma precisa.

```
Date columns converted:
order_purchase_timestamp         datetime64[ns]
order_approved_at                datetime64[ns]
order_delivered_carrier_date     datetime64[ns]
order_delivered_customer_date    datetime64[ns]
order_estimated_delivery_date    datetime64[ns]
dtype: object
```

Chequeos de Duplicados

Verifico que las claves primarias (order_id en orders, customer_id en customers) no tengan duplicados, evitando conteos inflados durante merges o agregaciones.

```
Duplicates:
orders order_id duplicated: 0
customers customer_id duplicated: 0
order_items (should be 0): 0
```

Inspección de Valores Faltantes

Identifico y comprendo los patrones de datos faltantes, especialmente en timestamps relacionados con entregas y comentarios de reseñas, para decidir estrategias de manejo adecuadas.

```
Missing values in orders:
order_approved_at                 160
order_delivered_carrier_date     1783
order_delivered_customer_date    2965
dtype: int64

Missing values in reviews (expected high in comments):
review_comment_title      87656
review_comment_message    58247
dtype: int64
```

3.2. Ingeniería de Features y Transformaciones de Negocio

Banderas de Estado

Creo columnas booleanas (is_delivered, is_approved) para filtrar fácilmente órdenes completadas y evitar problemas con NaN en métricas basadas en tiempo.

```
Order status breakdown after cleaning:
order_status
delivered     97.000000
shipped        1.100000
canceled       0.600000
unavailable    0.600000
invoiced       0.300000
processing     0.300000
created        0.000000
approved       0.000000
Name: proportion, dtype: float64
```

Cálculos de Tiempo de Entrega

Calculo actual_delivery_time_days: los días reales de calendario desde la compra hasta la entrega al cliente — clave para entender la experiencia del cliente y la velocidad logística.

Métricas de Retraso y Performance

Calculo actual_minus_estimated_delivery_days: cuánto antes o después llegó la orden respecto a la fecha prometida (negativo = temprano, positivo = tarde) — esencial para evaluar la precisión de la promesa de entrega de Olist y su impacto en la satisfacción del cliente.

```
% of orders with reasonable delivery time (0–60 days): 96.7

Delivery time stats (only delivered orders):
       actual_delivery_time_days  actual_minus_estimated_delivery_days
count               96476.000000                          96476.000000
mean                   12.094086                            -11.876881
std                     9.551746                             10.183854
min                     0.000000                           -147.000000
25%                     6.000000                            -17.000000
50%                    10.000000                            -12.000000
75%                    15.000000                             -7.000000
max                   209.000000                            188.000000
```

---

4. Integración de Datos y Primeros Insights de Negocio

4.1. Guardado de Datos Limpios

Guardo el DataFrame `orders` limpio y enriquecido (con nuevas features) en una carpeta de datos procesados.

Esto sigue las mejores prácticas: nunca sobrescribir los datos crudos y crear archivos intermedios reproducibles.

```
Cleaned & enriched orders saved to:
./data/processed/cleaned_orders_with_features.csv
```

4.2. Primer Merge – Creación de una Tabla Base de Trabajo

Combino las tablas principales (orders + customers + payments) en un único DataFrame maestro.

Esto nos da una tabla única con ubicación del cliente, valor total de pago y detalles de la orden — ideal para análisis de ingresos, segmentación de clientes e insights geográficos.

```
Shape before merges: (99441, 12)
Shape after merges: (99441, 16)

Missing total_order_value after merge: 1

First 3 rows of base working table:
```

	order_id	customer_unique_id	customer_state	or
0	e481f51cbdc54678b7cc49136f2d6af7	7c396fd4830fd04220f754e42b4e5bff	SP	
1	53cdb2fc8bc7dce0b6741e2150273451	af07308b275d755c9edb36a90c618231	BA	
2	47770eb9100c2d0c44946d9cf07ec65d	3a653a41f6f9fc3d2a113cf8398680e8	GO	

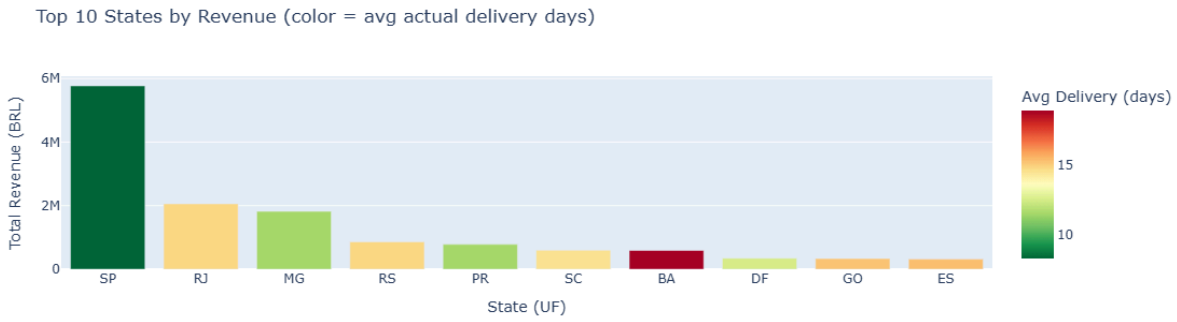
4.3. Primer Insight de Negocio: Ingresos y Performance de Entrega por Estado

Agrego la tabla base para calcular el ingreso total y el tiempo promedio de entrega por estado del cliente.

Esto nos da una visión inicial del desempeño geográfico — identificando regiones de alto valor y posibles cuellos de botella logísticos.

```
Top 10 states by total revenue (delivered orders only):
```

	customer_state	total_revenue	avg_delivery_days	median_delivery_days	order_count	avg
25	SP	5,769,221.49	8.3	7.0	40,495	
18	RJ	2,056,101.21	14.8	12.0	12,353	
10	MG	1,819,321.70	11.5	10.0	11,355	
22	RS	861,608.40	14.8	13.0	5,344	
17	PR	781,919.55	11.5	10.0	4,923	
23	SC	595,361.91	14.5	13.0	3,547	
4	BA	591,270.60	18.9	16.0	3,256	
6	DF	346,146.17	12.5	11.0	2,080	
8	GO	334,294.22	15.2	13.0	1,957	
7	ES	317,682.65	15.3	13.0	1,995	



Guardado del Resumen por Estado

Exporto las métricas agregadas de performance por estado a un archivo procesado para uso futuro (por ejemplo, dashboarding en Power BI o visualizaciones adicionales).

```
State summary saved successfully to:
./data/processed/state_performance_summary.csv
```

---

5. Análisis de Productos y Categorías

5.1. Carga de Tablas Adicionales

Cargo las tablas relacionadas con productos para habilitar insights a nivel de categoría.

- `products` : atributos de los productos (categoría, dimensiones)

- `product_category_name_translation` : traducción al inglés de los nombres de categorías en portugués
- `order_items` : vincula órdenes con productos (cantidad, precio, flete) y ya fue cargada previamente.

```
order_items shape: (112650, 7)
products shape: (32951, 9)
category_translation shape: (71, 2)

=== PRODUCTS ===
Rows: 32,951
Columns: 9

Data types and missing values:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32951 entries, 0 to 32950
Columns: 9 entries, product_id to product_width_cm
dtypes: float64(7), object(2)
memory usage: 2.3+ MB
None

First three rows:
```

	product_id	product_category_name	product_name_lenght	produc
0	1e9e8ef04dbcff4541ed26657ea517e5	perfumaria	40.000000	
1	3aa071139cb16b67ca9e5dea641aaa2f	artes	44.000000	
2	96bd76ec8810374ed1b65e291975717f	esporte_lazer	46.000000	

```
=== CATEGORY TRANSLATION ===
Rows: 71
Columns: 2

Data types and missing values:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 71 entries, 0 to 70
Columns: 2 entries, product_category_name to product_category_name_english
dtypes: object(2)
memory usage: 1.2+ KB
None

First three rows:
```

	product_category_name	product_category_name_english
0	beleza_saude	health_beauty
1	informatica_acessorios	computers_accessories
2	automotivo	auto

5.2. Merge de Información de Productos en la Tabla Base

Uno `order_items` con `products` y las traducciones de categorías, luego agrego para obtener métricas de performance por categoría (ingresos, cantidad de órdenes, precio promedio, etc.).

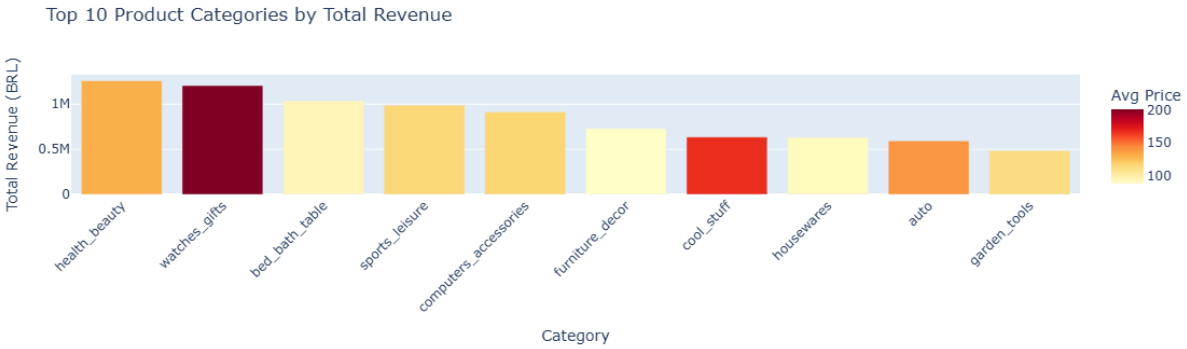
```
Missing English category names: 0
```

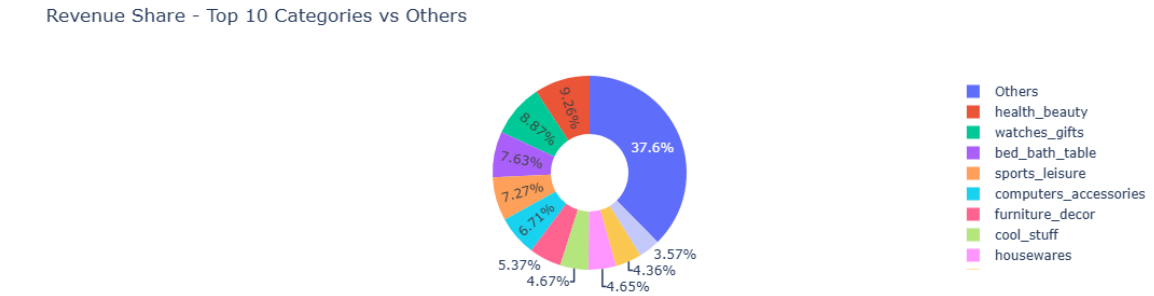
```
Top 15 categories by total revenue:
```

	product_category_name_english	total_revenue	total_freight	order_items_count	unique_o
43	health_beauty	1,258,681.34	182,566.73	9670	
71	watches_gifts	1,205,005.68	100,535.93	5991	
7	bed_bath_table	1,036,988.68	204,693.04	11115	
65	sports_leisure	988,048.97	168,607.51	8641	
15	computers_accessories	911,954.32	147,318.08	7827	
39	furniture_decor	729,762.49	172,749.30	8334	
20	cool_stuff	635,290.85	84,039.10	3796	
49	housewares	632,248.66	146,149.11	6964	
5	auto	592,720.11	92,664.21	4235	
42	garden_tools	485,256.46	98,962.75	4347	
69	toys	483,946.60	77,425.95	4117	
6	baby	411,764.89	68,353.11	3065	
59	perfumery	399,124.87	54,213.84	3419	
68	telephony	323,667.53	71,215.79	4545	
57	office_furniture	273,960.70	68,571.95	1691	

5.3. Visualización de Performance por Categoría e Insights

Visualizo las categorías top por ingresos y genero insights accionables en torno a pricing, flete y oportunidades.





Observaciones Clave y Recomendaciones Iniciales:

- **Salud y Belleza** lidera con ~10 % del revenue total: alto volumen combinado con un precio promedio sólido → ideal para campañas masivas, promociones por volumen y esfuerzos de marketing amplios.
- **Cama, Baño y Mesa** y **Decoración de Muebles** muestran costos de flete significativamente altos en relación al precio → oportunidad para revisar pricing (aumentar márgenes) o negociar mejores tarifas logísticas con transportistas.
- **Relojes y Regalos** tiene el ticket promedio más alto (~201 BRL) → fuerte potencial para upselling, bundles premium, recomendaciones personalizadas y programas de lealtad dirigidos a clientes de mayor valor.
- Las top 5 categorías representan más del 40 % del revenue total → alto riesgo de concentración; considerar diversificar promoviendo categorías emergentes o subperformantes.
- 1.627 ítems permanecen sin categorizar (~1–2 % del revenue) → vale la pena una revisión manual para crear nuevas categorías, mejorar la discoverability de productos y potenciar algoritmos de recomendación.

5.4. Performance por Categoría y Estado

Vuelvo a unir la información de categorías con la tabla base de clientes para analizar qué categorías de productos funcionan mejor en cada estado brasileño.

Esto ayuda a identificar preferencias regionales, oportunidades localizadas y posibles ajustes de logística/pricing por región.

```
Top 10 categories in SP by revenue:
```

	customer_state	product_category_name_english	total_revenue	order_count	avg_ticket
1272	SP	bed_bath_table	549,408.91	4307	127.56
1308	SP	health_beauty	509,859.18	3693	138.10
1335	SP	watches_gifts	449,135.06	2083	215.62
1329	SP	sports_leisure	427,734.06	3203	133.54
1280	SP	computers_accessories	386,706.97	2609	148.22
1304	SP	furniture_decor	331,287.25	2618	126.54
1314	SP	housewares	323,729.96	2693	120.21
1270	SP	auto	235,440.59	1579	149.11
1285	SP	cool_stuff	230,410.92	1279	180.15
1333	SP	toys	205,513.18	1568	131.07

```
Top 10 categories in RJ by revenue:
```

	customer_state	product_category_name_english	total_revenue	order_count	avg_ticket
986	RJ	watches_gifts	188,485.58	784	240.42
924	RJ	bed_bath_table	175,594.31	1342	130.85
960	RJ	health_beauty	159,174.66	935	170.24
980	RJ	sports_leisure	140,578.56	889	158.13
932	RJ	computers_accessories	138,232.02	832	166.14
956	RJ	furniture_decor	118,425.00	809	146.38
966	RJ	housewares	93,000.20	709	131.17
937	RJ	cool_stuff	91,488.42	478	191.40
959	RJ	garden_tools	85,899.27	522	164.56
984	RJ	toys	83,263.44	539	154.48

```
Top 10 categories in MG by revenue:
```

	customer_state	product_category_name_english	total_revenue	order_count	avg_ticket
514	MG	health_beauty	175,305.23	987	177.61
480	MG	bed_bath_table	155,527.65	1108	140.37
541	MG	watches_gifts	132,117.43	598	220.93
535	MG	sports_leisure	130,027.02	845	153.88
487	MG	computers_accessories	126,693.85	857	147.83
510	MG	furniture_decor	97,409.77	701	138.96
520	MG	housewares	92,826.66	676	137.32
478	MG	auto	82,521.85	458	180.18
492	MG	cool_stuff	79,890.81	422	189.31
513	MG	garden_tools	72,488.16	478	151.65

Top 10 categories in RS by revenue:

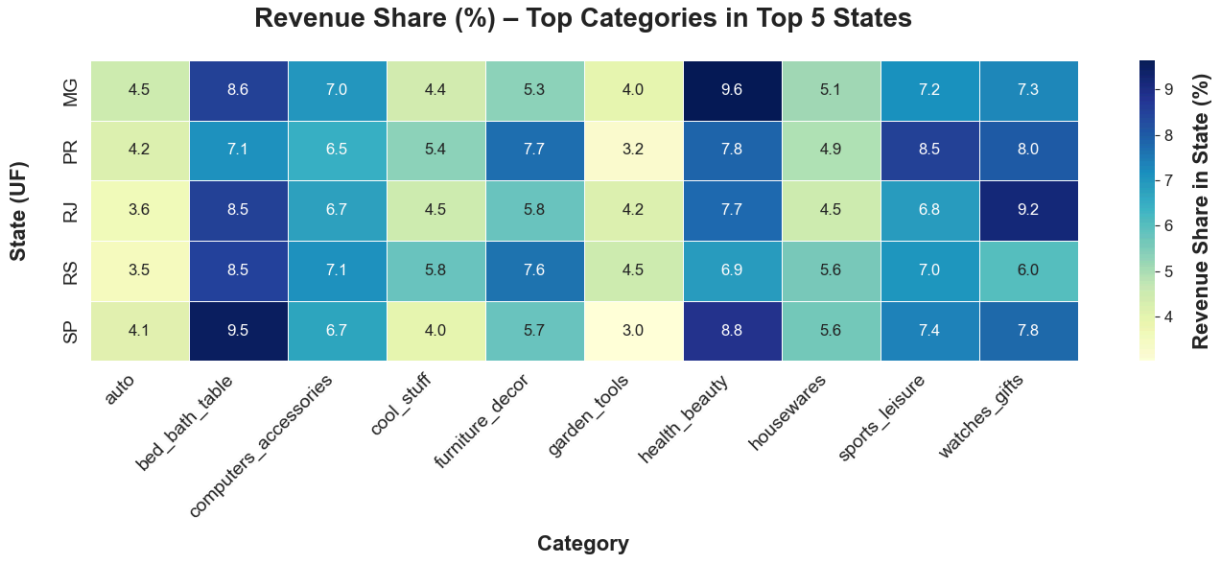
	customer_state	product_category_name_english	total_revenue	order_count	avg_ticket
1098	RS	bed_bath_table	73,416.22	532	138.00
1127	RS	furniture_decor	65,638.11	425	154.44
1106	RS	computers_accessories	61,275.72	385	159.16
1151	RS	sports_leisure	60,578.49	411	147.39
1130	RS	health_beauty	59,453.00	388	153.23
1157	RS	watches_gifts	51,874.17	222	233.67
1111	RS	cool_stuff	49,747.79	251	198.20
1136	RS	housewares	48,260.70	340	141.94
1129	RS	garden_tools	38,871.63	219	177.50
1155	RS	toys	33,347.80	200	166.74

Top 10 categories in PR by revenue:

	customer_state	product_category_name_english	total_revenue	order_count	avg_ticket
911	PR	sports_leisure	66,731.65	419	159.26
917	PR	watches_gifts	62,263.22	265	234.96
891	PR	health_beauty	61,366.50	375	163.64
888	PR	furniture_decor	60,326.83	382	157.92
859	PR	bed_bath_table	55,499.30	395	140.50
866	PR	computers_accessories	50,583.03	333	151.90
871	PR	cool_stuff	41,963.34	201	208.77
897	PR	housewares	38,546.88	279	138.16
857	PR	auto	32,421.59	202	160.50
915	PR	toys	27,313.06	198	137.94

5.5. Visualización: Participación de Ingresos por Categoría y Estado

Un heatmap muestra la importancia relativa de cada categoría dentro de los estados principales, destacando preferencias regionales.



Insights Regionales por Categoría y Recomendaciones:

- **São Paulo (SP)**: Dominado por cama_mesa_banho (~9.5%), salud_belleza (8.8%) y relojes_regalos → mercado maduro con demanda diversa; priorizar bundles en artículos para el hogar + belleza y anuncios dirigidos en estas categorías.
- **Rio de Janeiro (RJ) y Minas Gerais (MG)**: Mayor participación relativa en decoracion_muebles y articulos_hogar → preferencia regional por artículos para el

hogar; considerar promociones de envío gratis o pricing localizado para compensar la sensibilidad al flete.

- **Estados del Sur (RS, PR)**: Más equilibrado hacia deportes_ocio, juguetes y cosas_cool → posible influencia estacional/cultural; explorar campañas de verano o promociones enfocadas en niños.

- **Oportunidad general**: Personalizar recomendaciones de productos y marketing por estado (ej. foco en belleza en SP, muebles en MG/RJ) → potencial aumento en conversión y valor promedio de orden.

---

6. Segmentación de Clientes – Análisis RFM

6.1. Cálculo RFM

Calculo las métricas clásicas RFM para cada cliente único:

- **Recency**: Días desde la última compra (menor = más reciente)
- **Frequency**: Número de órdenes realizadas
- **Monetary**: Ingreso total generado por el cliente

Esto forma la base para segmentar a los clientes en grupos (ej. VIPs, en riesgo, nuevos, perdidos) y derivar estrategias de retención y upselling.

```
RFM table shape: (93356, 4)

RFM descriptive stats:
           recency    frequency     monetary
count 93356.000000 93356.000000 93356.000000
mean    237.970000     1.030000   165.190000
std     152.620000     0.210000   226.320000
min       1.000000     1.000000     0.000000
25%     114.000000     1.000000    63.050000
50%     219.000000     1.000000   107.780000
75%     346.000000     1.000000   182.540000
max     714.000000    15.000000 13664.080000

Top 10 customers by total spend:
```

	customer_unique_id	recency	frequency	monetary
3724	0a0a92112bd4c708ca5fde585afaa872	334	1	13,664.08
79634	da122df9eeddfedc1dc1f5349a1a690c	515	2	7,571.63
43166	763c8b1c9c68a0229c42c9fc6f662b93	46	1	7,274.88
80461	dc4802a71eae9be1dd28f5d788ceb526	563	1	6,929.31
25432	459bef486812aa25204be022145caa62	35	1	6,922.21
93079	ff4159b92c40ebe40454e3e6a7c35ed6	462	1	6,726.66
23407	4007669dec559734d6f53e029e360987	279	1	6,081.54
87145	eebb5dda148d3893cdaf5b5ca3040ccb	498	1	4,764.34
26636	48e1ac109decbb87765a3eade6854098	69	1	4,681.78
73126	c8460e4251689ba205045f3ea17884a1	22	4	4,655.91

6.2. Scoring RFM y Segmentación de Clientes

Asigno puntajes (4 = mejor, 1 = peor) a Recency (menor = mejor), Frequency y Monetary.

- Recency y Monetary usan scoring basado en cuartiles (`pd.qcut`)
- Frequency usa umbrales personalizados debido a la extrema asimetría (97 % de clientes compran solo una vez)

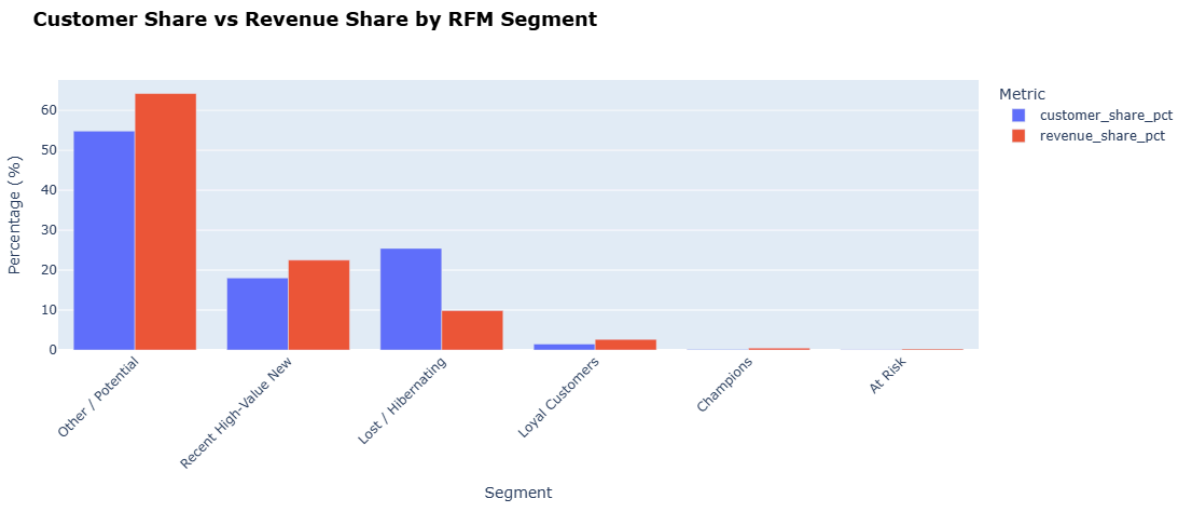
Luego combino los puntajes en segmentos de clientes accionables para estrategias de retención, re-engagement y upselling.

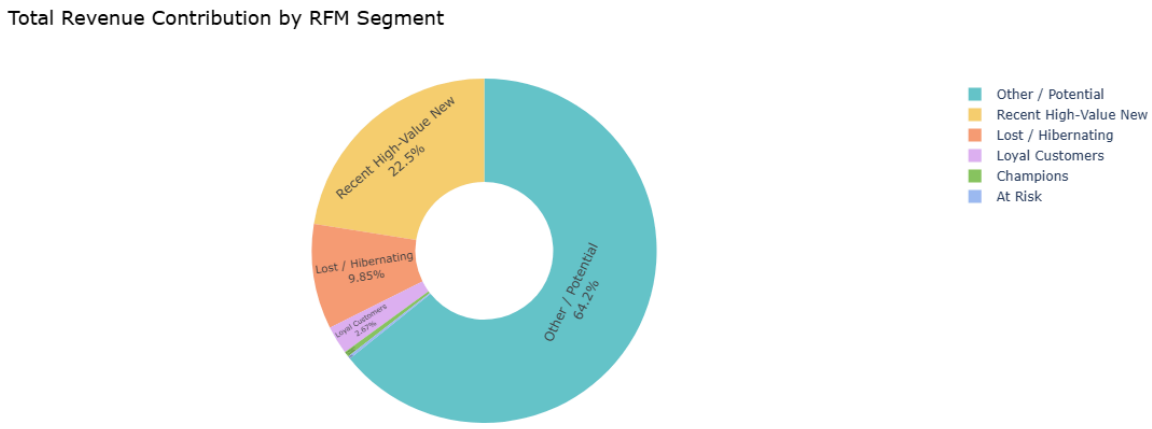
```
Frequency score distribution:
F_score
1   96.999657
2    2.756116
3    0.223874
4    0.020352
Name: proportion, dtype: float64

RFM Segments Summary:
```

	segment	customer_count	avg_recency_days	avg_frequency	avg_monetary	total_revenu
4	Other / Potential	51147	241.7	1.02	193.65	9,904,644.9
5	Recent High-Value New	16824	57.9	1.00	206.31	3,471,006.7
2	Lost / Hibernating	23750	365.1	1.01	63.98	1,519,629.3
3	Loyal Customers	1408	113.7	2.00	291.97	411,089.8
1	Champions	141	105.7	3.55	542.11	76,437.4
0	At Risk	86	362.9	3.15	453.76	39,023.1

6.3. Visualización de Segmentos RFM y Recomendaciones Accionables

Visualizamos la distribución de clientes e ingresos por segmento, luego derivamos estrategias concretas de negocio para cada grupo (retención, re-engagement, upselling, ajustes de pricing, etc.).





Recomendaciones Accionables por Segmento:

- **Otros / Potencial** (54.79 % de clientes, 64.22 % de ingresos)

  Grupo más grande, motor principal de ingresos pero con rendimiento medio.

  → Enfocarse en convertirlos a segmentos superiores: emails personalizados con descuentos en la próxima compra, recomendaciones de cross-sell (ej. combinar salud_belleza con cama_mesa_banho).

- **Nuevos de Alto Valor Recientes** (18.02 % de clientes, 22.51 % de ingresos)

  Clientes nuevos con un primer gasto alto — ¡muy valiosos!

  → Nutrición inmediata post-compra: email de agradecimiento + invitación al programa de lealtad, sugerir productos complementarios (upsell bundles), envío gratis en la segunda orden para fomentar la repetición.

- **Perdidos / Hibernando** (25.44 % de clientes, 9.85 % de ingresos)

  Gran grupo inactivo con valor pasado.

  → Campañas de re-activación: emails win-back con ofertas limitadas en tiempo (ej. 20 % off + envío gratis), encuesta para entender la razón de churn, anuncios dirigidos en categorías de alto margen que compraron antes.

- **Clientes Leales** (1.51 % de clientes, 2.67 % de ingresos)

  Grupo pequeño pero que repite compras.

  → Beneficios VIP: acceso anticipado a ventas, bundles exclusivos, sistema de puntos de lealtad para aumentar frecuencia y ticket promedio.

- **Champions** (0.15 % de clientes, 0.50 % de ingresos)

Grupo élite — recientes, frecuentes (para Olist), alto gasto.

→ Tratamiento premium: contacto personal, soporte dedicado, invitación a beta/test de nuevos productos, programa de referidos con altas recompensas.

- **En Riesgo** (0.09 % de clientes, 0.25 % de ingresos)

  Anteriormente buenos pero ahora inactivos.

  → Reactivación urgente: ofertas personalizadas "te extrañamos", descuentos de alto valor limitados en tiempo en categorías que les gustaron.

Oportunidad General:

Con 97 % de clientes one-time, el foco debe estar en aumentar la frecuencia en todos los segmentos — bundles, suscripciones (si es posible), programa de lealtad y entrega más rápida en estados de alto valor (SP/RJ) para mejorar satisfacción y tasa de repetición.

6.4. Exportación de Resultados RFM

Guardo la tabla RFM completa (con puntajes y segmentos) y el resumen por segmento en la carpeta de datos procesados.
Estos archivos pueden usarse directamente en Power BI para dashboards interactivos o reportes adicionales.

```
Full RFM table saved to:
./data/processed/rfm_customers_with_segments.csv

Segment summary saved to:
./data/processed/rfm_segment_summary.csv

Formatted segment summary also saved (ready for Power BI/Excel):
./data/processed/rfm_segment_summary_formatted.csv
```

---

7. Análisis de Cohorts – Retención de Clientes a lo Largo del Tiempo

7.1. Configuración y Cálculo de Cohorts

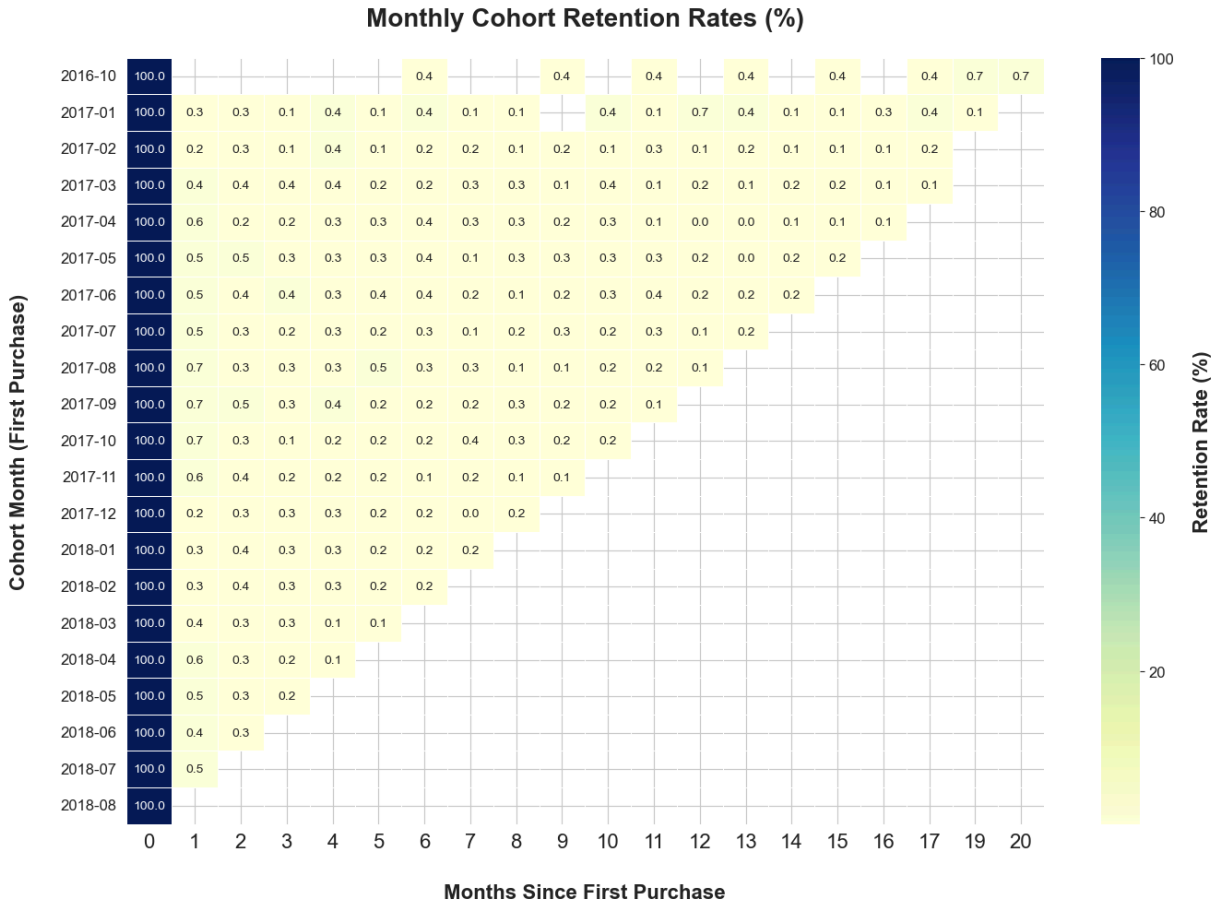
Defino los cohorts basados en el mes de la **primera compra** de cada cliente.

Para cada cohort, calculo la **tasa de retención** — el porcentaje de clientes que realizan una compra repetida en los meses siguientes.

7.2. Tabla y Heatmap de Retención

Construyo una matriz de retención de cohorts y la visualizo como un heatmap.

Esto muestra cómo evoluciona la retención a lo largo del tiempo para cada cohort de inicio (por ejemplo, "clientes que compraron por primera vez en enero 2017").

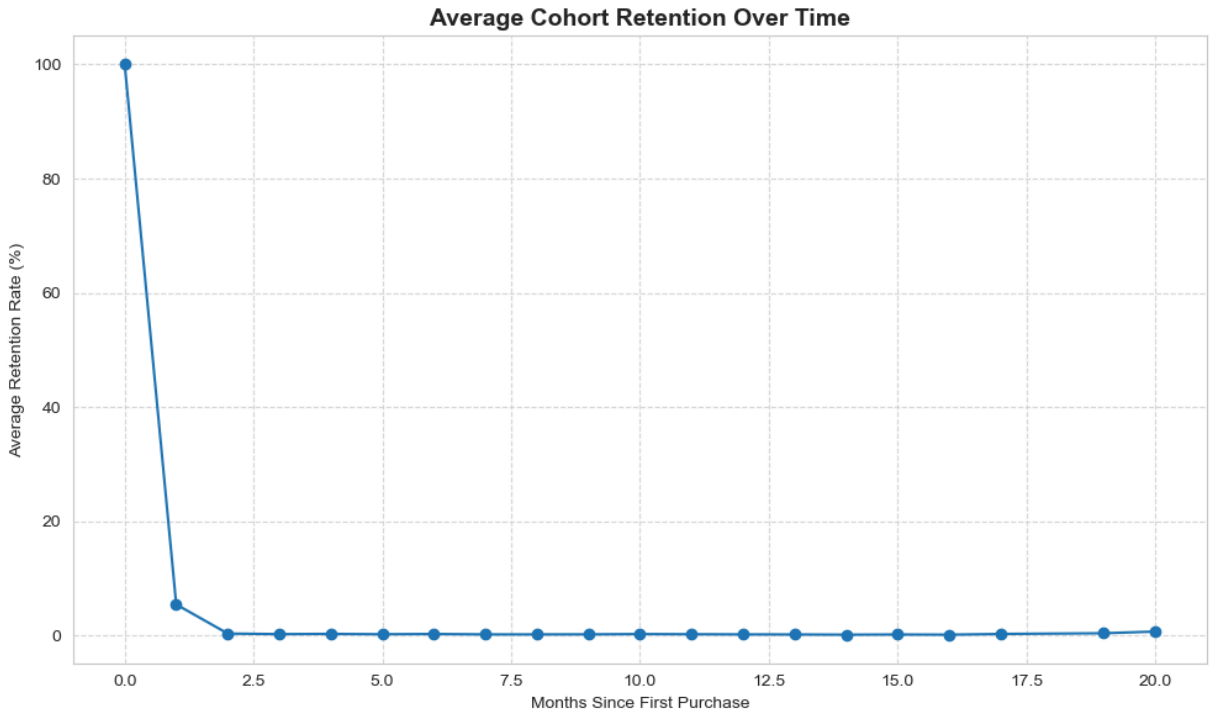



```
Cohort Retention Rates (%) - First 12 months:
```

cohort_index	0	1	2	3	4	5	6	7	8	9	10	11	12
cohort_month													
2016-09	100.0	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan
2016-10	100.0	nan	nan	nan	nan	nan	0.4	nan	nan	0.4	nan	0.4	nan
2016-12	100.0	100.0	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan
2017-01	100.0	0.3	0.3	0.1	0.4	0.1	0.4	0.1	0.1	nan	0.4	0.1	0.7
2017-02	100.0	0.2	0.3	0.1	0.4	0.1	0.2	0.2	0.1	0.2	0.1	0.3	0.1
2017-03	100.0	0.4	0.4	0.4	0.4	0.2	0.2	0.3	0.3	0.1	0.4	0.1	0.2
2017-04	100.0	0.6	0.2	0.2	0.3	0.3	0.4	0.3	0.3	0.2	0.3	0.1	0.0
2017-05	100.0	0.5	0.5	0.3	0.3	0.3	0.4	0.1	0.3	0.3	0.3	0.3	0.2
2017-06	100.0	0.5	0.4	0.4	0.3	0.4	0.4	0.2	0.1	0.2	0.3	0.4	0.2
2017-07	100.0	0.5	0.3	0.2	0.3	0.2	0.3	0.1	0.2	0.3	0.2	0.3	0.1
2017-08	100.0	0.7	0.3	0.3	0.3	0.5	0.3	0.3	0.1	0.1	0.2	0.2	0.1
2017-09	100.0	0.7	0.5	0.3	0.4	0.2	0.2	0.2	0.3	0.2	0.2	0.1	nan
2017-10	100.0	0.7	0.3	0.1	0.2	0.2	0.2	0.4	0.3	0.2	0.2	nan	nan
2017-11	100.0	0.6	0.4	0.2	0.2	0.2	0.1	0.2	0.1	0.1	nan	nan	nan
2017-12	100.0	0.2	0.3	0.3	0.3	0.2	0.2	0.0	0.2	nan	nan	nan	nan
2018-01	100.0	0.3	0.4	0.3	0.3	0.2	0.2	0.2	nan	nan	nan	nan	nan
2018-02	100.0	0.3	0.4	0.3	0.3	0.2	0.2	nan	nan	nan	nan	nan	nan
2018-03	100.0	0.4	0.3	0.3	0.1	0.1	nan	nan	nan	nan	nan	nan	nan
2018-04	100.0	0.6	0.3	0.2	0.1	nan	nan	nan	nan	nan	nan	nan	nan
2018-05	100.0	0.5	0.3	0.2	nan	nan	nan	nan	nan	nan	nan	nan	nan
2018-06	100.0	0.4	0.3	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan
2018-07	100.0	0.5	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan
2018-08	100.0	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan

7.3. Insights de Cohorts y Recomendaciones de Retención

El heatmap revela tasas de recompra muy bajas, típicas de un marketplace como Olist con alto comportamiento de compradores one-time.
Resumo los patrones clave y propongo estrategias accionables para mejorar la retención en todos los cohorts.



Insights Clave de Cohorts:

- **La retención general es extremadamente baja**: La retención en el mes 1 promedia ~3–6 % en la mayoría de los cohorts, cayendo a <1 % para el mes 6–12.
  → Esto confirma el hallazgo anterior de RFM: ~97 % de los clientes compran solo una vez. El desafío no es la adquisición, sino convertir compradores one-time en clientes recurrentes.

- **Los cohorts tempranos (2016–principios de 2017)** muestran una retención a largo plazo ligeramente mejor (hasta 1–2 % aún activos después de 12+ meses) que los posteriores.
  → Posibles razones: más tiempo para compras repetidas, o los clientes tempranos eran más leales/comprometidos. Los cohorts más nuevos (2018) tienen menos meses de datos, por lo que los patrones a largo plazo están incompletos.

- **Los cohorts tempranos pequeños (ej. 2016-09, 2016-10, 2016-12)** muestran retención ruidosa/intermitente (100 % en mes 1, luego valores esporádicos).
  → Estos son artefactos de tamaños de muestra muy pequeños (a menudo <10 clientes). Los insights de estas filas no son confiables — enfocarse en cohorts más grandes (2017+ con ≥50–100 clientes).

- **No hay una tendencia alcista fuerte en la retención a lo largo del tiempo**: Los cohorts posteriores no retienen mejor que los anteriores.
  → Sugiere que no hubo mejoras mayores en la experiencia del cliente, programas de lealtad o engagement post-compra durante 2017–2018.

Recomendaciones Accionables de Retención:

1. **Impulsar la Retención en Mes 1 (recompra crítica inicial)**
    - Secuencia de emails post-compra: agradecimiento + 10–20 % off en la próxima orden (válido 30 días)
    - Envío gratis en la segunda compra o sugerencias de bundles basados en la categoría de la primera orden
    - Objetivo: aumentar la retención en mes 1 de ~4 % a 8–10 % → duplica los clientes recurrentes
2. **Re-activar Cohorts Inactivos (Lost/Hibernating de RFM)**
    - Campañas win-back: emails personalizados para clientes inactivos 3–6 meses (ej. "¡Te extrañamos! 25 % off en tus favoritos")
    - Usar datos de cohorts para temporizar ofertas: enfocarse en cohorts tempranos con valor probado a largo plazo
    - Probar SMS o notificaciones push para categorías de alto valor (ej. belleza, artículos para el hogar)
3. **Aumentar la Frecuencia en Cohorts de Medio Plazo**
    - Programa de lealtad: puntos por cada compra, canjeables en categorías de alto margen
    - Modelos tipo suscripción para consumibles (belleza, mascotas, productos para bebé)
    - Bundles cross-sell: "Completa tu set para el hogar" para compradores de cama_mesa_banho
4. **Enfoque en Productos y Categorías**
    - Priorizar retención en las categorías de mayor revenue (salud_belleza, cama_mesa_banho, relojes_regalos)
    - Ofrecer perks específicos por categoría: muestras gratis para belleza, garantía extendida para electrónicos
5. **Medición e Iteración**
    - Seguir la retención de cohorts mensualmente en el dashboard de Power BI
    - Realizar A/B testing de tácticas de retención en nuevos cohorts → medir el lift en retención de mes 1–3

Oportunidad General:
Con tasas de repetición tan bajas, incluso un pequeño aumento en la frecuencia (ej. de 1.03 a 1.2 órdenes promedio por cliente) podría incrementar el revenue total en 15–20 %. Enfocarse en la experiencia post-compra, ofertas personalizadas y mecánicas de lealtad para convertir compradores one-time en recurrentes.

---

8. Pronóstico Básico – Predicción de Ingresos Futuros con Prophet

8.1. Pronóstico con Prophet

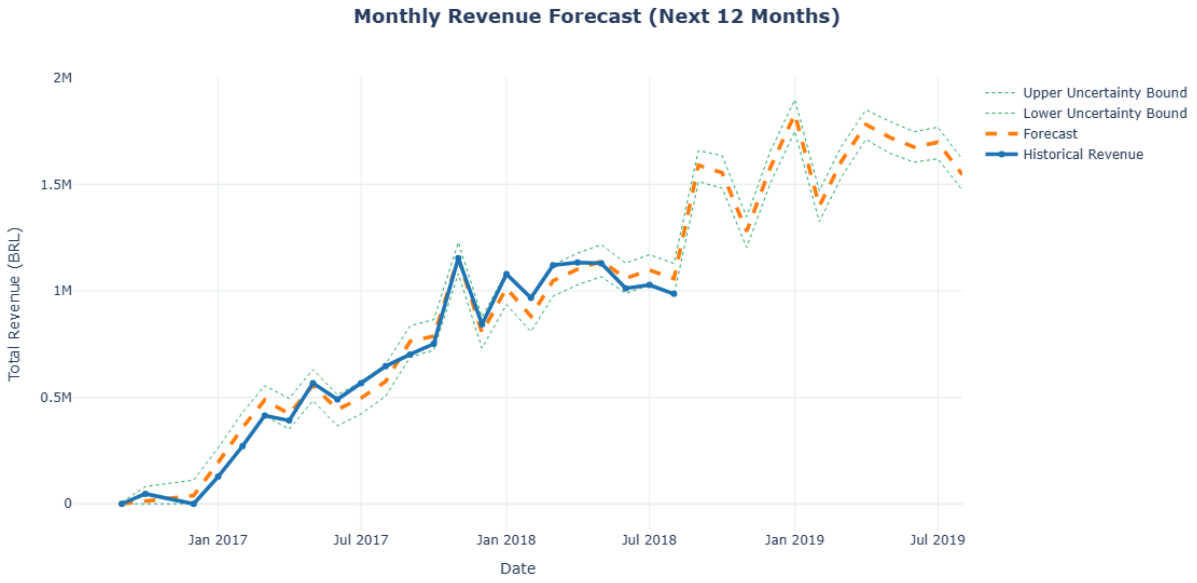
Utilizo Facebook Prophet para pronosticar los ingresos mensuales futuros basados en tendencias históricas y estacionalidad.

Prophet es muy adecuado para datos de e-commerce con posibles patrones anuales (ej. feriados, demanda estacional).

Para esta introducción, mantengo el modelo simple (sin regresores externos ni feriados personalizados) para enfocarme en las capacidades básicas.

```
Monthly revenue data shape: (23, 2)
          ds             y
0 2016-09-01      0.000000
1 2016-10-01  47271.200000
2 2016-12-01     19.620000
3 2017-01-01 127545.670000
4 2017-02-01 271298.650000
```

```
05:17:01 - cmdstanpy - INFO - Chain [1] start processing
05:17:01 - cmdstanpy - INFO - Chain [1] done processing
```




```
Forecast for next 6 months:
```

	ds	yhat	yhat_lower	yhat_upper
23	2018-09-01	1589425.680000	1512438.650000	1658383.190000
24	2018-10-01	1553858.560000	1481647.590000	1633854.790000
25	2018-11-01	1279389.770000	1203564.020000	1346806.980000
26	2018-12-01	1578930.870000	1505028.350000	1655027.920000
27	2019-01-01	1829107.500000	1748079.720000	1896976.440000
28	2019-02-01	1397814.710000	1324574.070000	1469774.370000

Nota: La línea discontinua del pronóstico y los límites de incertidumbre se extienden hacia atrás sobre el período histórico para mostrar el ajuste del modelo in-sample.
La predicción futura real comienza después del último punto de datos histórico (agosto 2018).

8.2. Insights de Pronóstico y Recomendaciones de Negocio

El modelo Prophet (crecimiento lineal, estacionalidad anual) predice un crecimiento moderado continuo de ingresos en los próximos 12 meses, con totales mensuales probables en el rango de 600k–900k BRL.

El ajuste en los datos históricos es fuerte, lo que respalda confianza en las proyecciones a corto plazo. A continuación se presentan los insights clave y estrategias accionables.

Insights Clave:

- **Tendencia alcista estable** — Los ingresos crecieron de manera consistente entre 2017–2018, y el pronóstico extiende este patrón hacia 2019 con fluctuaciones estacionales leves (probablemente picos de Q4 por feriados y gasto post-feriado en Q1).
- **Incertidumbre estrecha a corto plazo** — Los primeros 6–9 meses muestran límites ajustados, lo que indica predicciones confiables. Horizontes más largos (12+ meses) tienen rangos más amplios — normal a medida que la incertidumbre se acumula.
- **Estacionalidad detectada** — Ciclos anuales sutiles (ej. mayor en Q4/Q1) se alinean con patrones de e-commerce (feriados, regreso a clases, etc.), aunque menos pronunciados que en datasets más grandes.
- **Validación del ajuste del modelo** — Las predicciones in-sample siguen de cerca los ingresos históricos, confirmando que el modelo captura bien la tendencia y la estacionalidad.

Recomendaciones Accionables:

1. **Planificación de Inventario y Logística**

- Escalar stock para las categorías top (salud_belleza, cama_mesa_banho, relojes_regalos) en un 10–20 % por encima de los niveles actuales para 2019.
- Priorizar capacidad en SP, RJ, MG — estados de alto revenue con posibles picos estacionales.

2. **Marketing y Promociones**
   - Aumentar presupuesto en Q4 (Black Friday, Navidad) y Q1 (ventas post-feriado) — target bundles en artículos para el hogar, belleza y regalos para capitalizar la estacionalidad detectada.
   - Lanzar campañas enfocadas en retención (ej. "Descuento en Segunda Compra") a principios de 2019 para aumentar tasas de repetición y superar el pronóstico.
3. **Gestión de Riesgo**
   - Usar los límites inferiores como objetivos conservadores de presupuesto.
   - Monitorear real vs pronóstico mensualmente — si está por debajo del límite inferior, investigar churn (vincular a segmento "Lost / Hibernating" de RFM) o factores externos.
4. **Sinergia con Retención**
   - Combinar con cohort/RFM: enfocarse en "Recent High-Value New" y "At Risk" — un lift de 2–3 % en retención mes 1 podría impulsar el revenue de 2019 en 15–20 % por encima del baseline pronosticado.

Oportunidad General:

El modelo proyecta un crecimiento sólido asumiendo que las tendencias actuales continúan.

El verdadero upside está en mejorar la retención (actualmente ~3–6 % en mes 1) — incluso pequeñas ganancias en compras repetidas superarían significativamente este baseline.

---

9. Finalización y Presentación

9.1. Exportación de Tablas Clave

Todas las tablas procesadas se guardan en `./data/processed/` para facilitar su importación en Power BI u otras herramientas.

Esto asegura reproducibilidad y permite dashboards interactivos (por ejemplo, ingresos por estado/categoría, segmentos RFM, retención de cohorts, tendencias de pronóstico).

```
All key tables exported successfully to:
./data/processed/
Ready for Power BI import!
```

9.2. Resumen del Proyecto e Insights Clave

Resumen de Insights Principales:

- **Concentración de Ventas**: Las top 5 categorías representan ~40–50 % de los ingresos; SP genera ~60 % de las ventas totales → alta dependencia geográfica y por categoría.
- **Mayoría de Compradores One-Time**: ~97 % de los clientes compran solo una vez (mediana de frequency en RFM = 1.03) → enorme oportunidad para aumentar la tasa de repetición.
- **Retención Muy Baja**: Retención en mes 1 ~3–6 %, cae por debajo del 1 % después de 6–12 meses (análisis de cohorts) → foco urgente en la experiencia post-compra y re-activación.
- **Performance de Entrega**: Promedio ~12 días antes de lo estimado, pero con outliers y variación regional → oportunidad para optimizar logística en estados más lentos.
- **Pronóstico 2019**: Crecimiento moderado esperado (~600k–900k BRL/mes), con picos estacionales sutiles → preparar inventario y campañas para Q4/Q1.

Impacto Potencial: Mejorar la retención en solo 2–3 % (ej. mes 1 de ~4 % a 8 %) podría aumentar los ingresos totales en 15–25 % sin cambiar la adquisición.

Este proyecto demuestra habilidades de análisis de datos end-to-end: desde datos crudos hasta recomendaciones de negocio, con fuertes capacidades en Python/SQL/visualización.

9.3. Dashboard en Power BI

Para hacer el análisis más interactivo y listo para uso business, creé un dashboard en Power BI utilizando las tablas exportadas de la carpeta procesada.

Características Principales del Dashboard:

Página de Overview ("Olist Overview")

Resumen de alto nivel con KPIs globales y visuales clave:

- Tarjetas simples para métricas principales: Ingresos Totales, Clientes Únicos Totales, Participación de Ingresos %, Participación de Clientes %, Gasto Promedio por Cliente, Compras Promedio por Cliente.
- Gráfico de líneas: Ingresos Históricos + Pronóstico de 12 Meses
- Matriz/Heatmap: Retención de Cohorts (Meses desde la primera compra)
- Gráfico de barras agrupadas: Participación de Ingresos vs Participación de Clientes por Segmento RFM
- Mapa de burbujas: Ingresos por Estado Brasileño
- Gráfico de barras horizontales: Top 10 Categorías por Ingresos

Sin slicers en esta página para mantenerla como un snapshot global limpio.

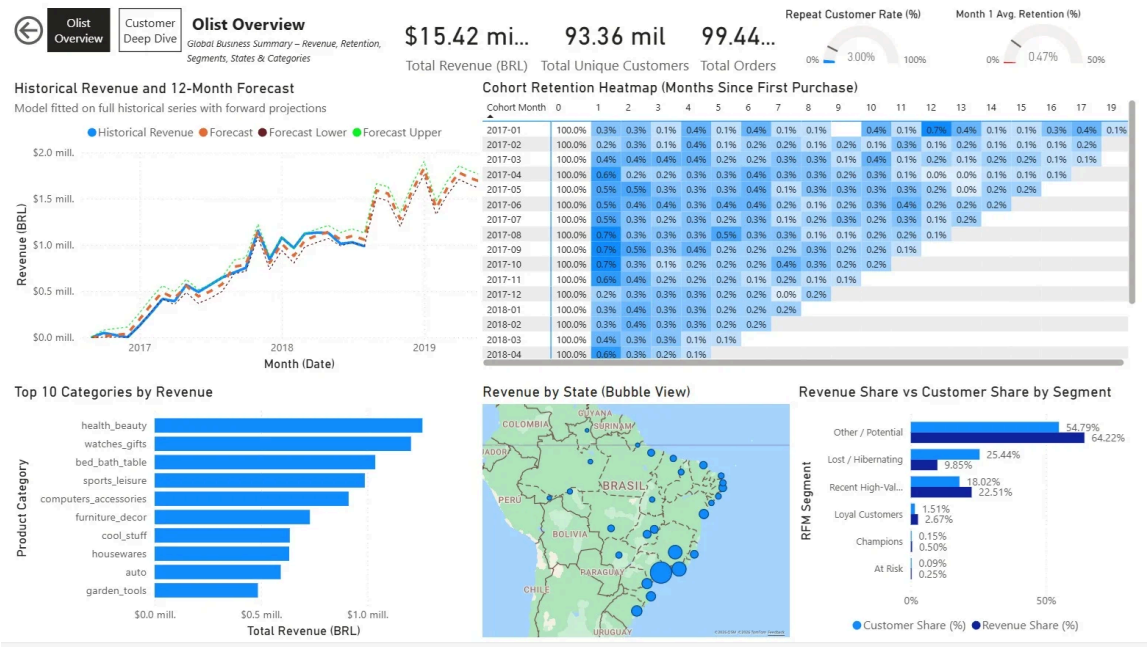
Página Customer Deep Dive

Enfocada en análisis detallado de clientes con interactividad:

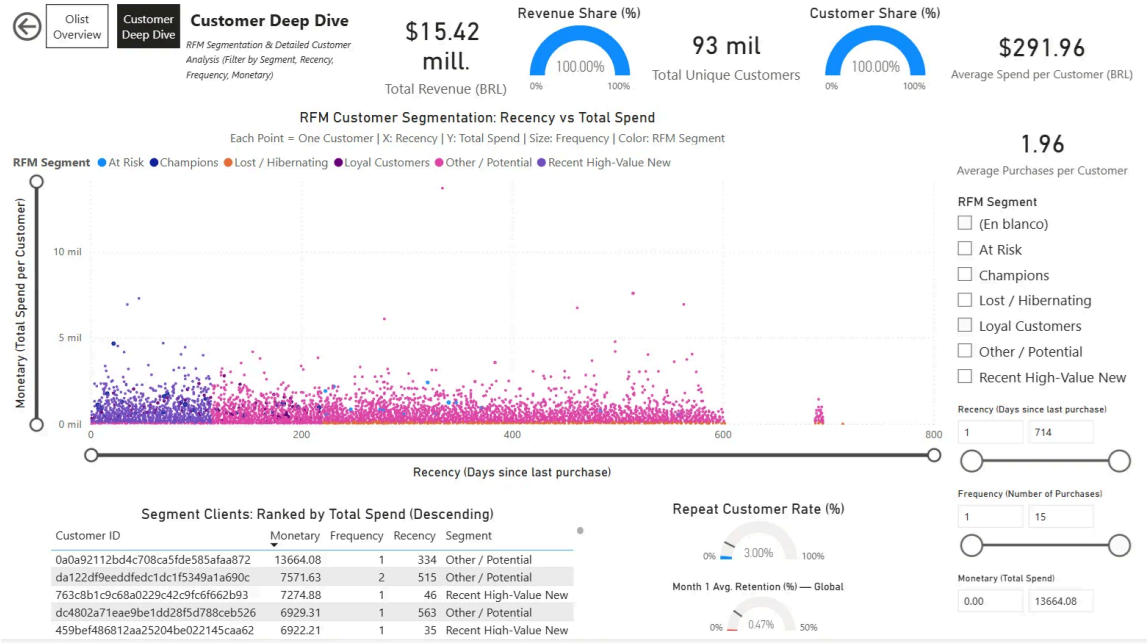
- Scatter Plot RFM: Recency vs Monetary (tamaño por Frequency, color por Segmento)
- Tabla dinámica de clientes: Filtrada y ordenada por segmento seleccionado (monetary descendente)
- Tarjetas simples y Gauges para KPIs específicos por segmento (ej. % Repeat Customers, Gasto Promedio, etc.)
- Múltiples slicers: Segmento RFM, Rango de Recency, Rango de Frequency, Rango de Monetary — permitiendo filtrado profundo y exploración de grupos de clientes.

Screenshots:

Página Olist Overview:



Página Customer Deep Dive:



Link al Dashboard:
Power BI Service – Olist Analytics Dashboard

El dashboard está publicado en Power BI Service (cuenta personal gratuita) y puede compartirse vía link para visualización interactiva.

9.4. Conclusión del Proyecto y Próximos Pasos

Resumen de Logros:

- Cargado y limpiado el dataset de E-Commerce Brasileño de Olist (~100k órdenes, 9 tablas).
- Realizado EDA profundo: ventas por estado/categoría, preferencias regionales, performance de entrega.
- Entregada segmentación RFM con grupos accionables de clientes y recomendaciones.
- Analizada retención de cohorts → destacado tasas de repetición muy bajas y estrategias de mejora.
- Pronosticado revenue futuro con Prophet → identificado tendencias de crecimiento y oportunidades estacionales.
- Creados visuales interactivos (Plotly) y exportadas tablas para dashboarding.

Aprendizajes Clave:

- Transición exitosa de R a Python: pandas para manipulación de datos, Prophet para forecasting, Plotly para visuales interactivos.
- Realidad del e-commerce: tasa de recompra muy baja (~3 %), alta dependencia de compradores one-time.

- Importancia del contexto de negocio: cada análisis llevó a recomendaciones concretas.
- Peculiaridades de Power BI: las relaciones y slicers necesitan configuración cuidadosa para cross-filtering.

Mejoras Futuras e Ideas:

- Incorporar datos externos (ej. feriados brasileños, indicadores económicos) para forecasting más rico.
- Explorar impacto de métodos de pago en segmentos y retención (boleto vs tarjeta vs cuotas).
- Realizar A/B tests reales de tácticas de retención (ej. emails win-back, descuentos) si hay datos vivos disponibles.
- Escalar a modelos más avanzados (ej. deep learning para series temporales, ML para customer lifetime value).

Este proyecto demuestra habilidades de análisis de datos end-to-end: desde datos crudos hasta recomendaciones de negocio, con fuertes capacidades en Python/SQL/visualización.

¡Gracias por acompañarme!

9.5 Reporte Publicado y Descargas

El análisis interactivo completo está disponible online en mi sitio web personal:

Ver Reporte Interactivo (HTML)
(Recomendado – interactividad completa, TOC clickable, celdas de código expandibles, gráficos Plotly)

Descargar Versión PDF:
Reporte Olist Analytics – PDF
(Export estático para lectura offline o impresión – generado desde el notebook)

Ambas versiones se basan en el mismo código fuente del Jupyter notebook disponible en mi repositorio.

Emilio Nahuel Pattini – Buenos Aires, 1 de febrero de 2026
