Show code

Analisi dell'E-Commerce Brasiliano – Olist

Analisi Retail e Clienti (2016–2018)

Emilio Nahuel Pattini
Buenos Aires, Argentina
1 febbraio 2026

Indice dei Contenuti

Introduzione

Obiettivo del Progetto
Analisi orientata al business: vendite, comportamento del cliente, RFM, cohorts, CLV, previsione di base + raccomandazioni actionable.

Dataset
Dataset Pubblico Olist – ~100k ordini (Kaggle)

Tech Stack

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

---

1. Comprensione del Business e Domande

Domande chiave:

- Quali sono le categorie / prodotti top per ricavi?
- Retention e recurrenza dei clienti?
- Segmentazione RFM?
- Retention dei cohorts nel tempo?
- Performance di consegna per regione?
- Impatto dei metodi di pagamento?
- Opportunità di cross-sell?
- Previsione per le categorie top?

---

2. Caricamento e Panoramica dei Dati

Questa sezione copre l'ingestione iniziale dei file CSV raw dal dataset Olist e fornisce una panoramica di alto livello sulla struttura, dimensione e contenuto di ciascuna tabella. L'obiettivo è confermare il caricamento riuscito, identificare le relazioni chiave tra le tabelle e

individuare eventuali segnali immediati di qualità dei dati prima di procedere con la pulizia e l'analisi.

2.1. Caricamento dei Dati

Carico le tabelle più rilevanti dal dataset Olist (orders, order_items, customers, payments, reviews) utilizzando pandas.
In questa fase vengono caricate solo le tabelle essenziali per mantenere basso l'uso della memoria e concentrarsi sulle entità principali necessarie per l'analisi di vendite, clienti e logistica.

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

2.2. Panoramica dei Dati e Controlli Iniziali

Eseguo un'ispezione rapida di ciascuna tabella caricata per comprendere:

- Numero di righe e colonne
- Tipi di dati
- Presenza di valori mancanti
- Righe di esempio

Questo passo aiuta a mappare lo schema del dataset e decidere le priorità di pulizia successive.

```
=== 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. Pulizia e Preparazione dei Dati

3.1. Pulizia Iniziale dei Dati

In questa fase eseguo controlli fondamentali di qualità dei dati: conversione delle colonne timestamp in formato datetime appropriato, verifica dell'unicità degli identificatori chiave (order_id, customer_id), e ispezione dei valori mancanti nelle colonne critiche. L'obiettivo è

garantire che i dati raw siano affidabili e pronti per l'analisi senza introdurre errori nei calcoli o nei join.

Conversioni di tipo

Converto le stringhe timestamp in oggetti datetime per poter eseguire calcoli basati sul tempo (delta, raggruppamenti per mese, ecc.) in modo accurato.

```
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
```

Controlli di Duplicati

Verifico che le chiavi primarie (order_id in orders, customer_id in customers) non abbiano duplicati, evitando conteggi gonfiati durante i merge o le aggregazioni.

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

Ispezione dei Valori Mancanti

Identifico e comprendo i pattern di dati mancanti, in particolare nei timestamp relativi alle consegne e nei commenti delle recensioni, per decidere strategie di gestione appropriate.

```
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. Feature Engineering e Trasformazioni di Business

Flag di Stato

Creo colonne booleane (is_delivered, is_approved) per filtrare facilmente ordini completati ed evitare problemi con NaN in metriche basate sul tempo.

```
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
```

Calcoli del Tempo di Consegna

Calcolo actual_delivery_time_days: i giorni di calendario reali dall'acquisto alla consegna al cliente — chiave per comprendere l'esperienza del cliente e la velocità logistica.

Metriche di Ritardo e Performance

Calcolo actual_minus_estimated_delivery_days: quanto prima o dopo è arrivata l'ordine rispetto alla data promessa (negativo = anticipato, positivo = in ritardo) — essenziale per valutare l'accuratezza della promessa di consegna di Olist e il suo impatto sulla soddisfazione 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. Integrazione dei Dati e Primi Insight di Business

4.1. Salvataggio dei Dati Puliti

Salvo il DataFrame `orders` pulito e arricchito (con nuove feature) in una cartella di dati processati.

Questo segue le best practice: non sovrascrivere mai i dati raw e creare file intermedi riproducibili.

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

4.2. Primo Merge – Creazione di una Tabella Base di Lavoro

Combino le tabelle principali (orders + customers + payments) in un unico DataFrame master.

Questo fornisce una tabella unica con posizione del cliente, valore totale di pagamento e dettagli dell'ordine — ideale per analisi dei ricavi, segmentazione clienti e insight geografici.

```
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	

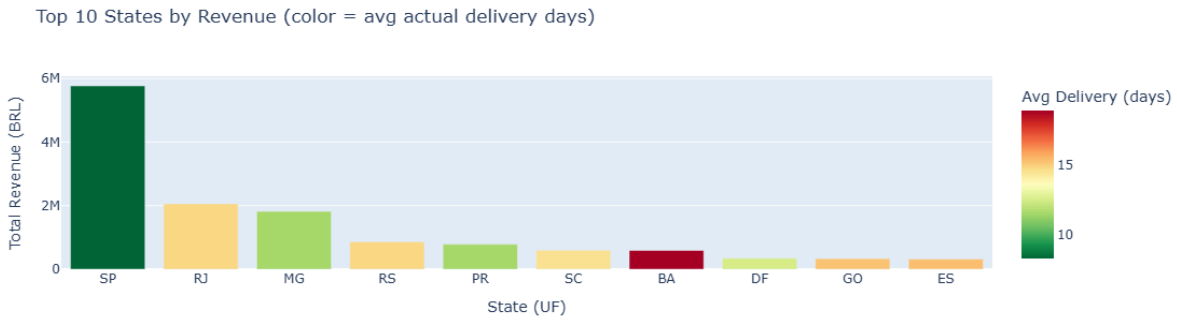
4.3. Primo Insight di Business: Ricavi e Performance di Consegna per Stato

Aggrego la tabella base per calcolare il ricavo totale e il tempo medio di consegna per stato del cliente.

Questo fornisce una visione iniziale delle performance geografiche — identificando regioni ad alto valore e potenziali colli di bottiglia logistici.

```
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	



Salvataggio del Riassunto per Stato

Esporto le metriche aggregate di performance per stato in un file processato per uso futuro (ad esempio, dashboarding in Power BI o visualizzazioni aggiuntive).

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

---

5. Analisi di Prodotti e Categorie

5.1. Caricamento di Tabelle Aggiuntive

Carico le tabelle relative ai prodotti per abilitare insight a livello di categoria.

- `products` : attributi dei prodotti (categoria, dimensioni)

- `product_category_name_translation` : traduzione in inglese dei nomi delle categorie in portoghese
- `order_items` : collega ordini ai prodotti (quantità, prezzo, spedizione) e già caricata precedentemente.

```
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 delle Informazioni sui Prodotti nella Tabella Base

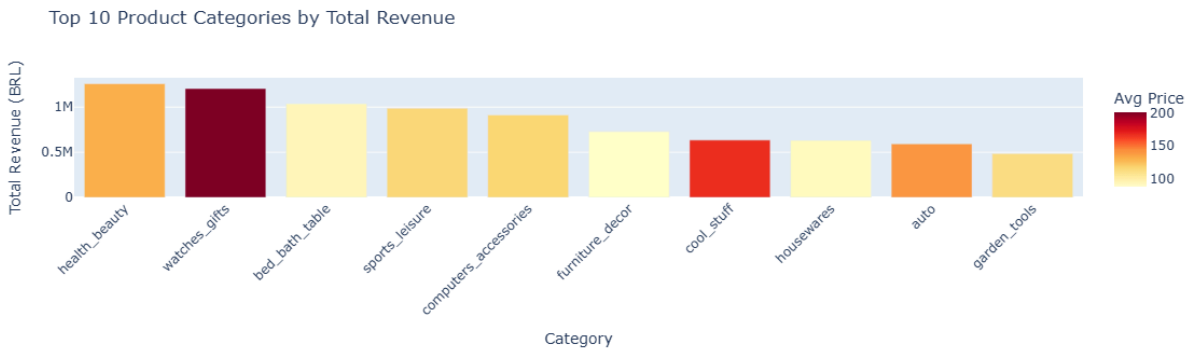
Unisco `order_items` con `products` e le traduzioni delle categorie, poi aggrego per ottenere metriche di performance per categoria (ricavi, conteggio ordini, prezzo medio, ecc.).

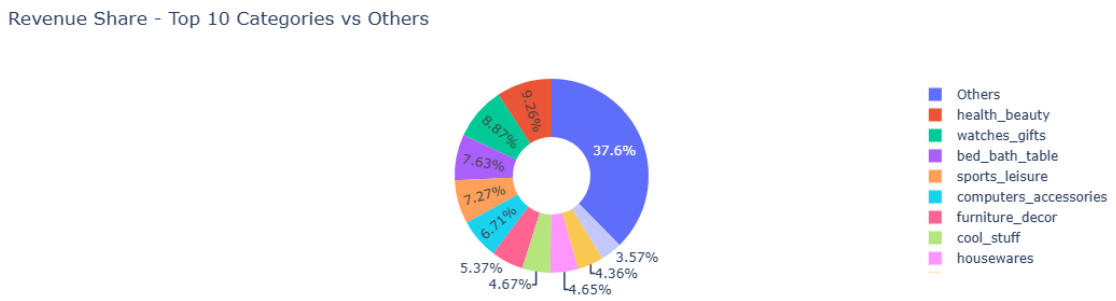
```
Missing English category names: 0

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. Visualizzazione della Performance per Categoria e Insight

Visualizzo le categorie top per ricavi e genero insight actionable su pricing, spedizione e opportunità.





Osservazioni Chiave e Raccomandazioni Iniziali:

- **Salute e Bellezza** guida con ~10 % del revenue totale: alto volume combinato con un prezzo medio solido → ideale per campagne di massa, promozioni per volume e sforzi di marketing ampi.
- **Letto, Bagno e Tavola** e **Arredamento e Decorazione** mostrano costi di spedizione significativamente alti rispetto al prezzo → opportunità per rivedere il pricing (aumentare i margini) o negoziare tariffe logistiche migliori con i corrieri.
- **Orologi e Regali** ha il ticket medio più alto (~201 BRL) → forte potenziale per upselling, bundle premium, raccomandazioni personalizzate e programmi fedeltà rivolti a clienti ad alto valore.
- Le top 5 categorie rappresentano più del 40 % del revenue totale → alto rischio di concentrazione; considerare di diversificare promuovendo categorie emergenti o sottoperformanti.
- 1.627 articoli rimangono non categorizzati (~1–2 % del revenue) → vale la pena una revisione manuale per creare nuove categorie, migliorare la discoverability dei prodotti e potenziare gli algoritmi di raccomandazione.

5.4. Performance per Categoria e Stato

Riunisco le informazioni sulle categorie con la tabella base dei clienti per analizzare quali categorie di prodotti performano meglio in ciascun stato brasiliano.

Questo aiuta a identificare preferenze regionali, opportunità localizzate e potenziali aggiustamenti logistici/pricing per regione.

```
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:

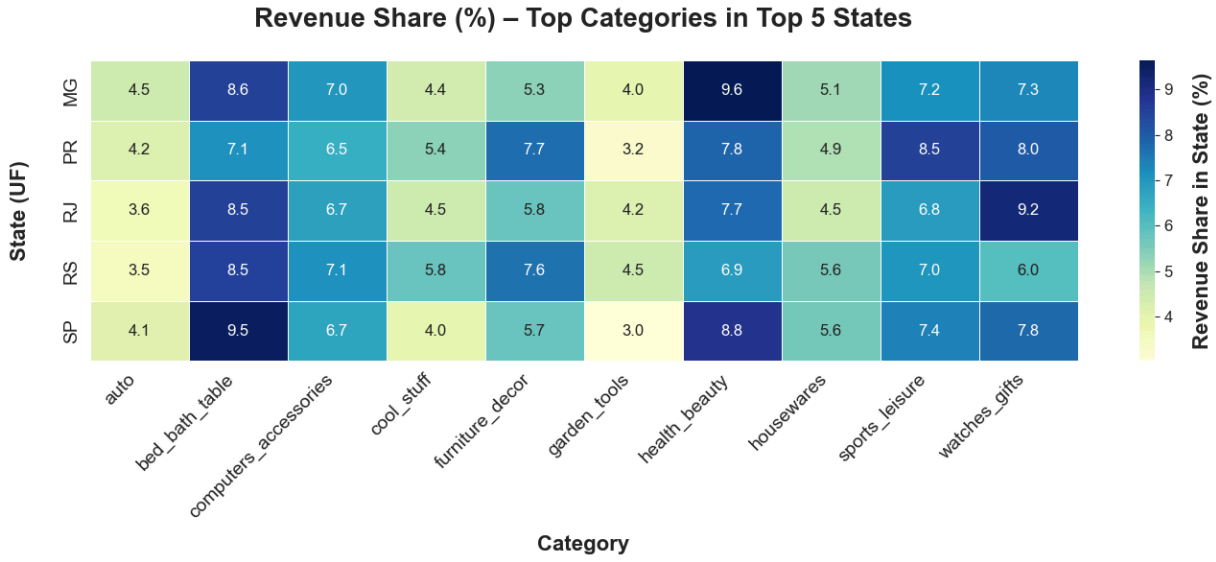
	customer_state	product_category_name_english	total_revenue	order_count	avg_ticket
1098	RS	bed_bath_table	73,416.22	532	138.00
1127	RS	furniture_decor	65,638.11	425	154.44
1106	RS	computers_accessories	61,275.72	385	159.16
1151	RS	sports_leisure	60,578.49	411	147.39
1130	RS	health_beauty	59,453.00	388	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. Visualizzazione: Partecipazione dei Ricavi per Categoria e Stato

Un heatmap mostra l'importanza relativa di ciascuna categoria all'interno degli stati principali, evidenziando preferenze regionali.



Insight Regionali per Categoria e Raccomandazioni:

- **São Paulo (SP)**: Dominato da letto_bagno_tavola (~9.5%), salute_bellezza (8.8%) e orologi_regali → mercato maturo con domanda diversificata; prioritizzare bundle su articoli per la casa + bellezza e annunci mirati in queste categorie.
- **Rio de Janeiro (RJ) e Minas Gerais (MG)**: Maggiore quota relativa in decorazione_mobili e articoli_casa → preferenza regionale per articoli domestici;

considerare promozioni di spedizione gratuita o pricing localizzato per compensare la sensibilità al costo di spedizione.

- **Stati del Sud (RS, PR)**: Più equilibrato verso sport_ricreazione, giocattoli e cose_cool → possibile influenza stagionale/culturale; esplorare campagne estive o promozioni focalizzate sui bambini.

- **Opportunità generale**: Personalizzare raccomandazioni di prodotti e marketing per stato (es. focus su bellezza in SP, mobili in MG/RJ) → potenziale aumento di conversione e valore medio dell'ordine.

---

6. Segmentazione Clienti – Analisi RFM

6.1. Calcolo RFM

Calcolo le metriche classiche RFM per ciascun cliente unico:

- **Recency**: Giorni dall'ultimo acquisto (minore = più recente)
- **Frequency**: Numero di ordini effettuati
- **Monetary**: Ricavo totale generato dal cliente

Questo costituisce la base per segmentare i clienti in gruppi (es. VIP, a rischio, nuovi, persi) e derivare strategie di retention e 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 e Segmentazione Clienti

Assegno punteggi (4 = migliore, 1 = peggiore) a Recency (minore = migliore), Frequency e Monetary.

- Recency e Monetary usano scoring basato sui quartili (`pd.qcut`)
- Frequency usa soglie personalizzate a causa dell'estrema skewness (97 % dei clienti acquista solo una volta)

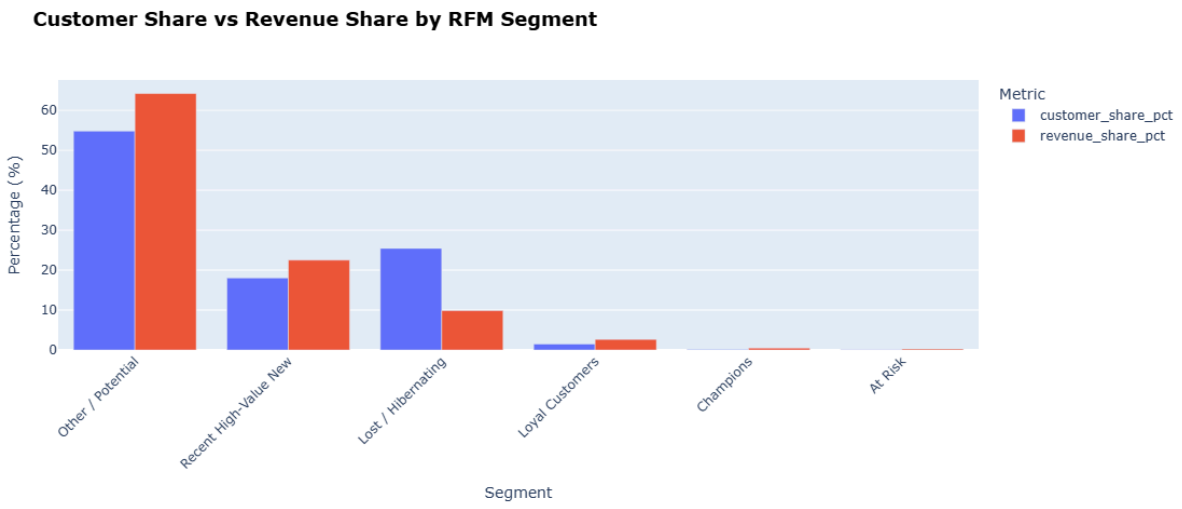
Poi combino i punteggi in segmenti clienti actionable per strategie di retention, re-engagement e upselling.

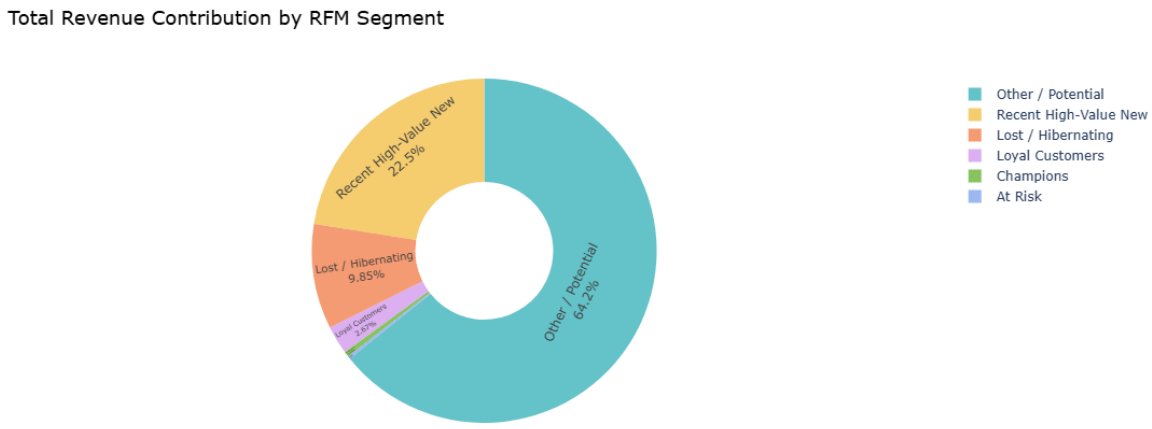
```
Frequency score distribution:
F_score
1   96.999657
2    2.756116
3    0.223874
4    0.020352
Name: proportion, dtype: float64

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. Visualizzazione dei Segmenti RFM e Raccomandazioni Azionabili

Visualizziamo la distribuzione di clienti e ricavi per segmento, poi deriviamo strategie concrete di business per ciascun gruppo (retention, re-engagement, upselling, aggiustamenti di pricing, ecc.).





Raccomandazioni Azionabili per Segmento:

- **Altri / Potenziale** (54.79 % clienti, 64.22 % ricavi)

  Gruppo più grande, motore principale dei ricavi ma con performance media.

  → Concentrarsi sulla conversione verso segmenti superiori: email personalizzate con sconti sul prossimo acquisto, raccomandazioni cross-sell (es. bundle salute_bellezza con letto_bagno_tavola).

- **Nuovi ad Alto Valore Recenti** (18.02 % clienti, 22.51 % ricavi)

  Clienti nuovi con primo acquisto di alto valore — molto preziosi!

  → Nutrimento immediato post-acquisto: email di ringraziamento + invito al programma fedeltà, suggerire prodotti complementari (upsell bundle), spedizione gratuita sul secondo ordine per incoraggiare la ripetizione.

- **Persi / Inattivi** (25.44 % clienti, 9.85 % ricavi)

  Grande gruppo inattivo con valore passato.

  → Campagne di riattivazione: email win-back con offerte a tempo limitato (es. 20 % off + spedizione gratuita), sondaggio per capire la ragione del churn, annunci mirati su categorie ad alto margine che hanno comprato in passato.

- **Clienti Leali** (1.51 % clienti, 2.67 % ricavi)

  Gruppo piccolo ma che ripete acquisti.

  → Benefici VIP: accesso anticipato alle vendite, bundle esclusivi, sistema di punti fedeltà per aumentare frequenza e ticket medio.

- **Champions** (0.15 % clienti, 0.50 % ricavi)

Gruppo élite — recenti, frequenti (per Olist), alto spend.

→ Trattamento premium: contatto personale, supporto dedicato, invito a beta/test di nuovi prodotti, programma referral con alte ricompense.

- **A Rischio** (0.09 % clienti, 0.25 % ricavi)

  Precedentemente buoni ma ora inattivi.

  → Riattivazione urgente: offerte personalizzate "ci manchi", sconti di alto valore a tempo limitato su categorie che hanno amato.

Opportunità Generale:

Con il 97 % di clienti one-time, il focus deve essere sull'aumento della frequenza in tutti i segmenti — bundle, abbonamenti (se possibile), programma fedeltà e consegne più rapide negli stati ad alto valore (SP/RJ) per migliorare soddisfazione e tasso di ripetizione.

6.4. Esportazione dei Risultati RFM

Salvo la tabella RFM completa (con punteggi e segmenti) e il riassunto per segmento nella cartella dei dati processati.
Questi file possono essere usati direttamente in Power BI per dashboard interattivi o report aggiuntivi.

```
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. Analisi dei Cohorts – Retention dei Clienti nel Tempo

7.1. Setup e Calcolo dei Cohorts

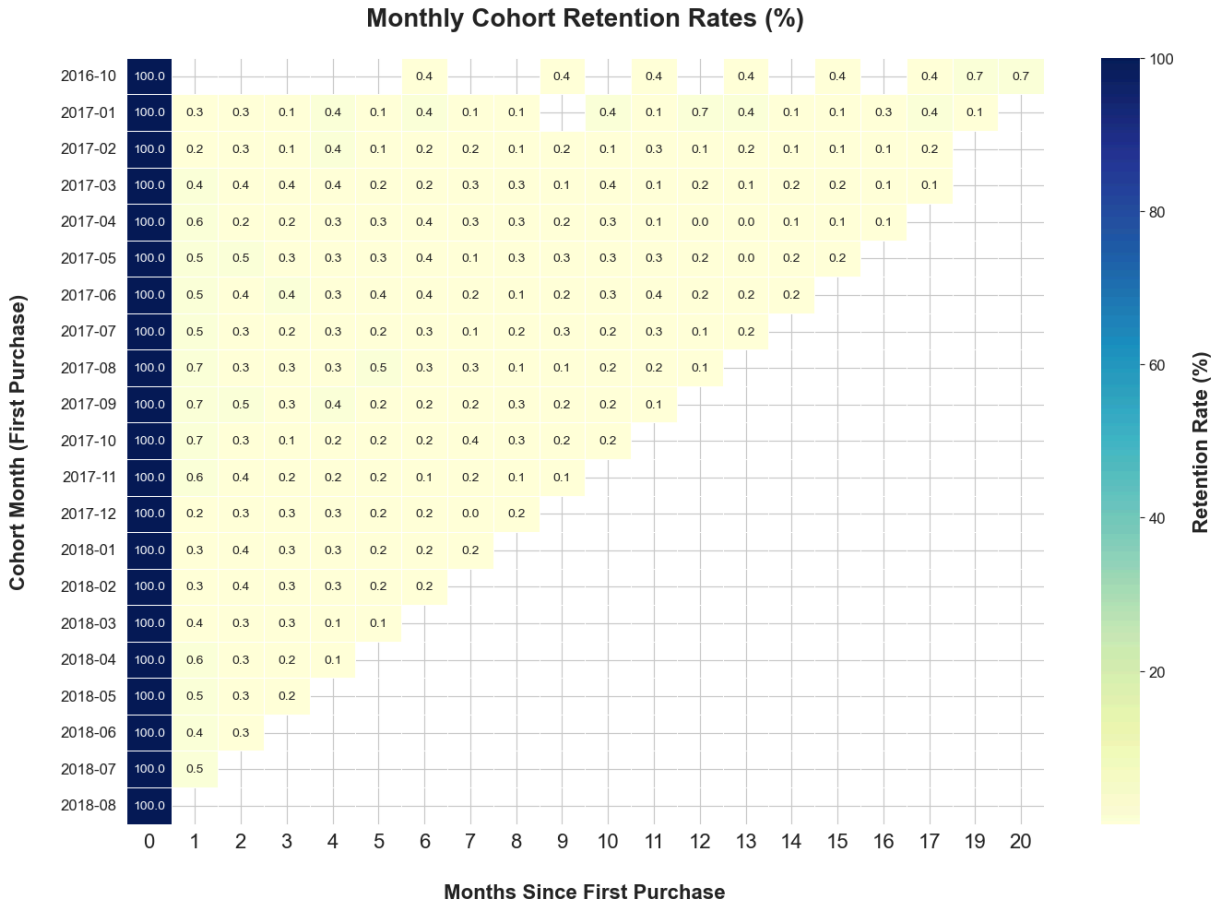
Definisco i cohorts in base al mese del **primo acquisto** di ciascun cliente.

Per ogni cohort, calcolo il **tasso di retention** — la percentuale di clienti che effettuano un acquisto ripetuto nei mesi successivi.

7.2. Tabella e Heatmap di Retention

Costruisco una matrice di retention dei cohorts e la visualizzo come heatmap.

Questo mostra come la retention evolve nel tempo per ciascun cohort di partenza (ad esempio, "clienti che hanno acquistato per la prima volta a gennaio 2017").

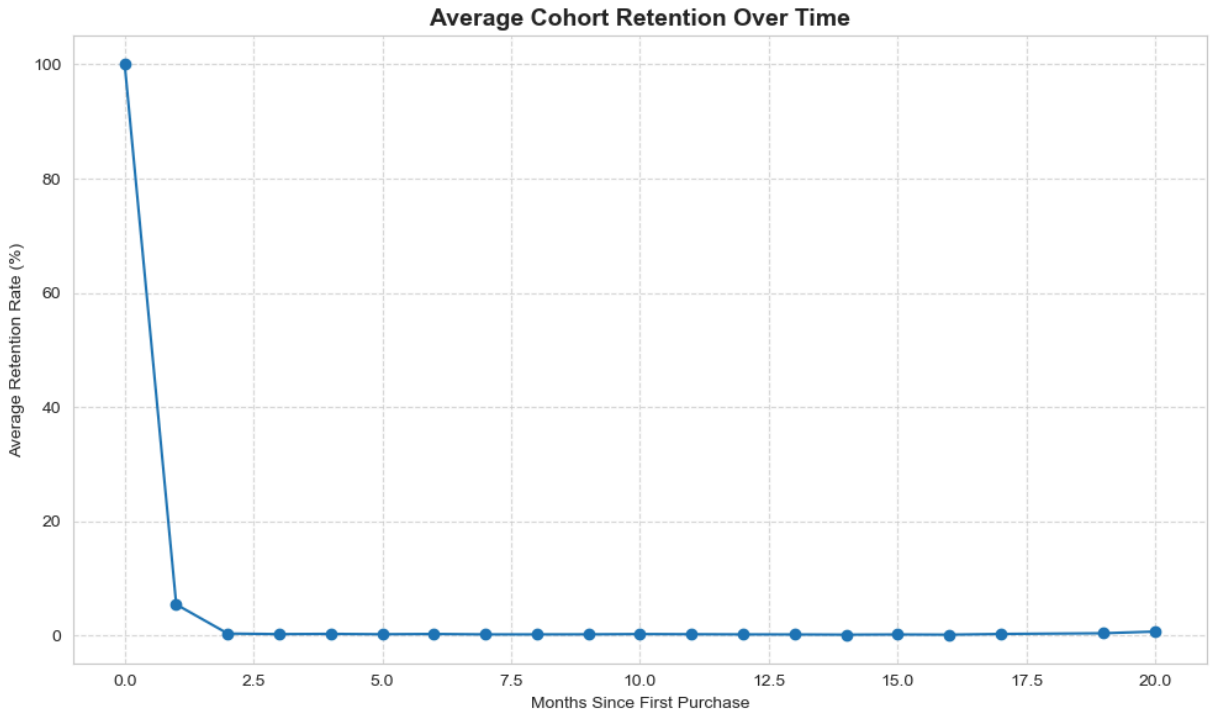



```
Cohort Retention Rates (%) - First 12 months:
```

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. Insight sui Cohorts e Raccomandazioni di Retention

Il heatmap rivela tassi di riacquisto molto bassi, tipici di un marketplace come Olist con alto comportamento di acquirenti one-time.
Riassumo i pattern chiave e propongo strategie actionable per migliorare la retention in tutti i cohorts.



Insight Chiave sui Cohorts:

- **La retention complessiva è estremamente bassa**: La retention nel mese 1 media ~3–6 % nella maggior parte dei cohorts, scendendo a <1 % entro il mese 6–12.
  → Questo conferma il precedente risultato RFM: ~97 % dei clienti acquista solo una volta. La sfida non è l'acquisizione, ma trasformare i compratori one-time in clienti ricorrenti.

- **I cohorts iniziali (2016–inizio 2017)** mostrano una retention a lungo termine leggermente migliore (fino a 1–2 % ancora attivi dopo 12+ mesi) rispetto a quelli successivi.
  → Possibili ragioni: più tempo per acquisti ripetuti, o i clienti iniziali erano più fedeli/coinvolti. I cohorts più recenti (2018) hanno meno mesi di dati, quindi i pattern a lungo termine sono incompleti.

- **I cohorts iniziali piccoli (es. 2016-09, 2016-10, 2016-12)** mostrano retention rumorosa/intermittente (100 % nel mese 1, poi valori sporadici).
  → Questi sono artefatti di dimensioni di campione molto piccole (spesso <10 clienti). Gli insight da queste righe non sono affidabili — concentrarsi sui cohorts più grandi (2017+ con ≥50–100 clienti).

- **Nessuna forte tendenza al rialzo nella retention nel tempo**: I cohorts successivi non trattengono meglio di quelli precedenti.
  → Suggerisce che non ci sono stati miglioramenti significativi nell'esperienza cliente, programmi fedeltà o engagement post-acquisto durante il 2017–2018.

Raccomandazioni Azionabili di Retention:

1. **Aumentare la Retention nel Mese 1 (prima ripetizione critica)**
    - Sequenza di email post-acquisto: ringraziamento + 10–20 % off sul prossimo ordine (valido 30 giorni)
    - Spedizione gratuita sul secondo acquisto o suggerimenti di bundle basati sulla categoria del primo ordine
    - Obiettivo: aumentare la retention nel mese 1 da ~4 % a 8–10 % → raddoppia i clienti ricorrenti
2. **Riattivare i Cohorts Inattivi (Lost/Hibernating da RFM)**
    - Campagne win-back: email personalizzate per clienti inattivi da 3–6 mesi (es. "Ci manchi! 25 % off sui tuoi preferiti")
    - Usare i dati dei cohorts per tempizzare le offerte: targettizzare i cohorts iniziali con valore a lungo termine provato
    - Testare SMS o notifiche push per categorie ad alto valore (es. bellezza, articoli per la casa)
3. **Aumentare la Frequenza nei Cohorts di Medio Termine**
    - Programma fedeltà: punti per ogni acquisto, riscattabili su categorie ad alto margine
    - Modelli tipo abbonamento per consumabili (bellezza, pet, prodotti per neonati)
    - Cross-sell bundle: "Completa il tuo set casa" per acquirenti di letto_bagno_tavola
4. **Focus su Prodotti e Categorie**
    - Prioritizzare la retention nelle categorie top per ricavi (salute_bellezza, letto_bagno_tavola, orologi_regali)
    - Offrire perks specifici per categoria: campioni gratuiti per bellezza, garanzia estesa per elettronica
5. **Misurazione e Iterazione**
    - Tracciare la retention dei cohorts mensilmente nel dashboard Power BI
    - Effettuare A/B test su tattiche di retention su nuovi cohorts → misurare il lift nella retention mese 1–3

Opportunità Complessiva:
Con tassi di ripetizione così bassi, anche un piccolo aumento nella frequenza (es. da 1.03 a 1.2 ordini medi per cliente) potrebbe aumentare i ricavi totali del 15–20 %.
Concentrarsi sull'esperienza post-acquisto, offerte personalizzate e meccanismi di fedeltà per trasformare i compratori one-time in ricorrenti.

---

8. Previsione Base – Predizione dei Ricavi Futuri con Prophet

8.1. Previsione con Prophet

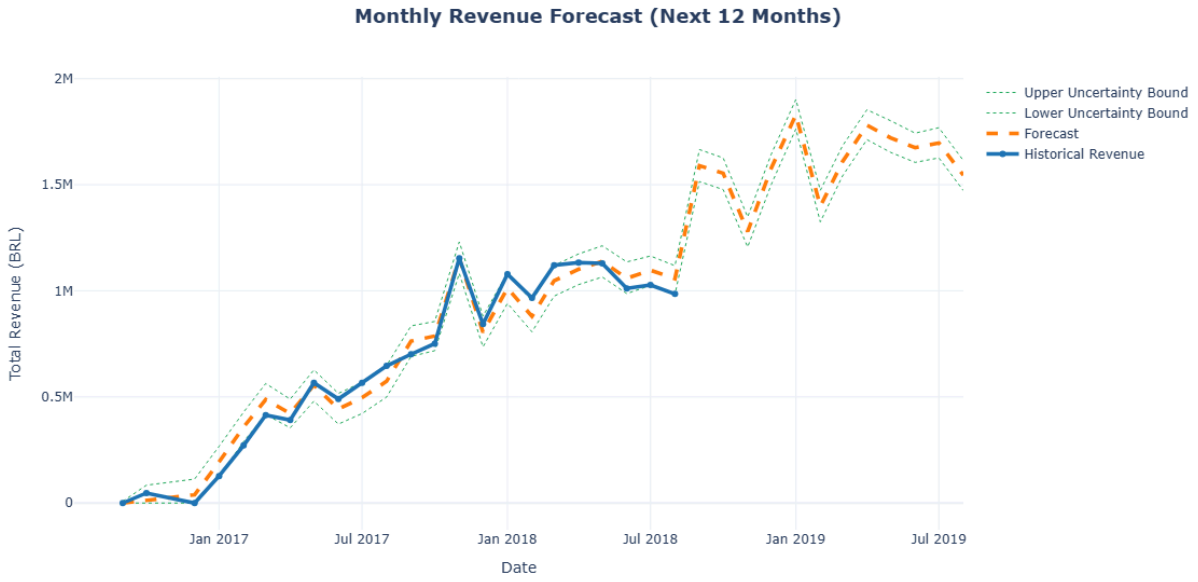
Utilizzo Facebook Prophet per prevedere i ricavi mensili futuri basati su trend storici e stagionalità.

Prophet è particolarmente adatto per dati di e-commerce con possibili pattern annuali (es. festività, domanda stagionale).

Per questa introduzione, mantengo il modello semplice (senza regressori esterni né festività personalizzate) per concentrarmi sulle capacità di base.

```
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
```

```
05:23:53 - cmdstanpy - INFO - Chain [1] start processing
05:23:53 - cmdstanpy - INFO - Chain [1] done processing
```




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

	ds	yhat	yhat_lower	yhat_upper
23	2018-09-01	1589425.680000	1514766.650000	1665541.430000
24	2018-10-01	1553858.560000	1477069.290000	1626246.180000
25	2018-11-01	1279389.770000	1206471.930000	1347997.920000
26	2018-12-01	1578930.870000	1501966.060000	1644733.610000
27	2019-01-01	1829107.500000	1762708.620000	1901570.950000
28	2019-02-01	1397814.710000	1324703.110000	1473307.300000

Nota: La linea tratteggiata della previsione e i limiti di incertezza si estendono indietro sul periodo storico per mostrare l'adattamento del modello in-sample.
La previsione futura effettiva inizia dopo l'ultimo punto dati storico (agosto 2018).

8.2. Insight di Previsione e Raccomandazioni di Business

Il modello Prophet (crescita lineare, stagionalità annuale) prevede una crescita moderata continua dei ricavi nei prossimi 12 mesi, con totali mensili probabili nel range 600k–900k BRL.

L'adattamento sui dati storici è forte, supportando fiducia nelle proiezioni a breve termine. Di seguito i principali insight e strategie actionable.

Insight Chiave:

- **Trend ascendente stabile** — I ricavi sono cresciuti costantemente dal 2017–2018, e la previsione estende questo pattern nel 2019 con lievi fluttuazioni stagionali (probabilmente picchi in Q4 per festività e spesa post-festiva in Q1).
- **Incertezza stretta a breve termine** — I primi 6–9 mesi mostrano intervalli stretti, indicando previsioni affidabili. Orizzonti più lunghi (12+ mesi) hanno intervalli più ampi — normale man mano che l'incertezza si accumula.
- **Stagionalità rilevata** — Cicli annuali sottili (es. maggiore in Q4/Q1) si allineano con i pattern dell'e-commerce (festività, ritorno a scuola, ecc.), sebbene meno pronunciati rispetto a dataset più grandi.
- **Validazione dell'adattamento del modello** — Le previsioni in-sample seguono da vicino i ricavi storici, confermando che il modello cattura bene trend e stagionalità.

Raccomandazioni Azionabili:

1. **Pianificazione Inventario e Logistica**

- Scalare lo stock per le categorie top (salute_bellezza, letto_bagno_tavola, orologi_regali) del 10–20 % sopra i livelli attuali per il 2019.
- Prioritizzare capacità in SP, RJ, MG — stati ad alto revenue con possibili picchi stagionali.

2. **Marketing e Promozioni**
   - Aumentare budget in Q4 (Black Friday, Natale) e Q1 (vendite post-festive) — target bundle su articoli per la casa, bellezza e regali per capitalizzare la stagionalità rilevata.
   - Lanciare campagne focalizzate sulla retention (es. "Sconto sul Secondo Acquisto") all'inizio del 2019 per aumentare i tassi di ripetizione e superare la previsione.
3. **Gestione del Rischio**
   - Usare i limiti inferiori come obiettivi di budget conservativi.
   - Monitorare reale vs previsione mensilmente — se sotto il limite inferiore, indagare churn (collegare al segmento "Lost / Hibernating" di RFM) o fattori esterni.
4. **Sinergia con Retention**
   - Combinare con cohort/RFM: focalizzarsi su "Recent High-Value New" e "At Risk" — un lift del 2–3 % nella retention mese 1 potrebbe spingere i ricavi 2019 del 15–20 % sopra il baseline previsto.

Opportunità Complessiva:

Il modello proietta una crescita solida assumendo che le tendenze attuali continuino.

Il vero upside sta nel migliorare la retention (attualmente ~3–6 % nel mese 1) — anche piccoli guadagni nelle ripetizioni supererebbero significativamente questo baseline.

---

9. Finalizzazione e Presentazione

9.1. Esportazione delle Tabelle Chiave

Tutte le tabelle processate vengono salvate in `./data/processed/` per un facile import in Power BI o altri tool.

Questo garantisce riproducibilità e abilita dashboard interattivi (ad esempio, ricavi per stato/categoria, segmenti RFM, retention dei cohorts, trend di previsione).

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

9.2. Riassunto del Progetto e Insight Chiave

Riassunto Principali Insight:

- **Concentrazione delle Vendite**: Le top 5 categorie rappresentano ~40–50 % dei ricavi; SP genera ~60 % delle vendite totali → alta dipendenza geografica e per categoria.
- **Maggioranza di Acquirenti One-Time**: ~97 % dei clienti acquista solo una volta (mediana frequency in RFM = 1.03) → enorme opportunità per aumentare il tasso di ripetizione.
- **Retention Molto Bassa**: Retention mese 1 ~3–6 %, scende sotto l'1 % dopo 6–12 mesi (analisi cohorts) → focus urgente sull'esperienza post-acquisto e riattivazione.
- **Performance di Consegna**: Media ~12 giorni prima del previsto, ma con outliers e variazione regionale → opportunità per ottimizzare la logistica negli stati più lenti.
- **Previsione 2019**: Crescita moderata attesa (~600k–900k BRL/mese), con picchi stagionali sottili → preparare inventario e campagne per Q4/Q1.

Impatto Potenziale: Migliorare la retention di appena 2–3 % (es. mese 1 da ~4 % a 8 %) potrebbe aumentare i ricavi totali del 15–25 % senza cambiare l'acquisizione.

Questo progetto dimostra competenze di analisi dati end-to-end: dai dati raw alle raccomandazioni di business, con forti capacità in Python/SQL/visualizzazione.

9.3. Dashboard in Power BI

Per rendere l'analisi più interattiva e pronta per l'uso business, ho creato un dashboard in Power BI utilizzando le tabelle esportate dalla cartella processed.

Caratteristiche Principali del Dashboard:

Pagina Overview ("Olist Overview")

Riassunto di alto livello con KPI globali e visual chiave:

- Card semplici per metriche core: Ricavi Totali, Clienti Unici Totali, % Share Ricavi, % Share Clienti, Spesa Media per Cliente, Acquisti Medi per Cliente.
- Grafico a linee: Ricavi Storici + Previsione 12 Mesi
- Matrice/Heatmap: Retention dei Cohorts (Mesi dalla prima acquisto)
- Grafico a barre raggruppate: % Share Ricavi vs % Share Clienti per Segmento RFM
- Mappa Bubble: Ricavi per Stato Brasiliano

- Grafico a barre orizzontali: Top 10 Categorie per Ricavi

Nessun slicer su questa pagina per mantenerla come snapshot globale pulito.

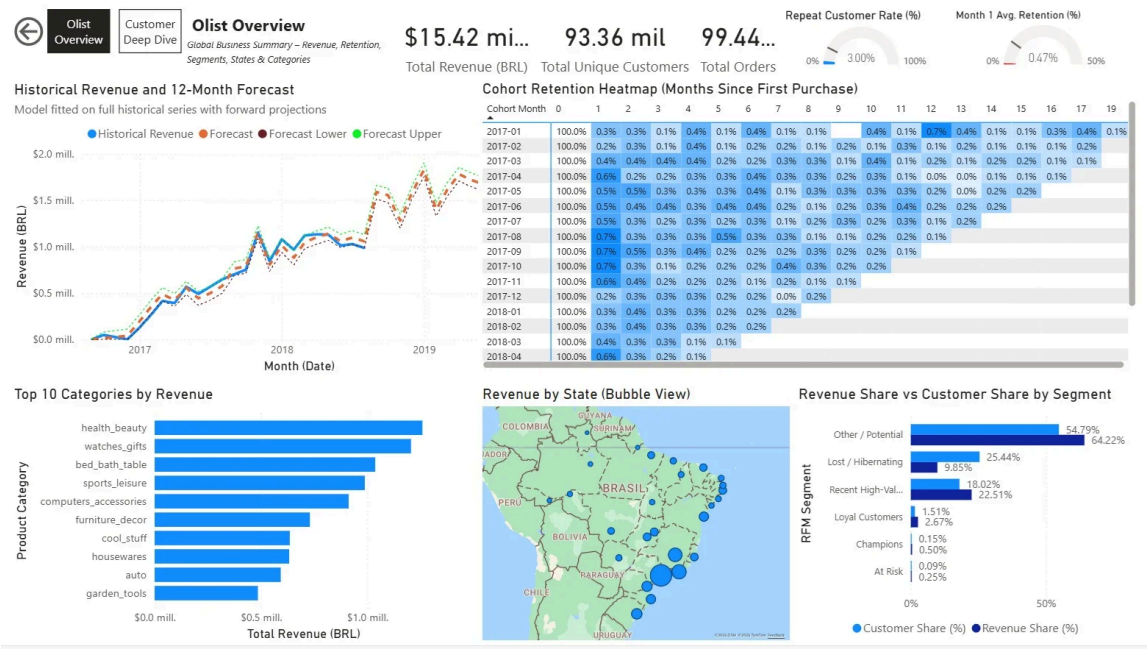
Pagina Customer Deep Dive

Focalizzata su analisi dettagliata dei clienti con interattività:

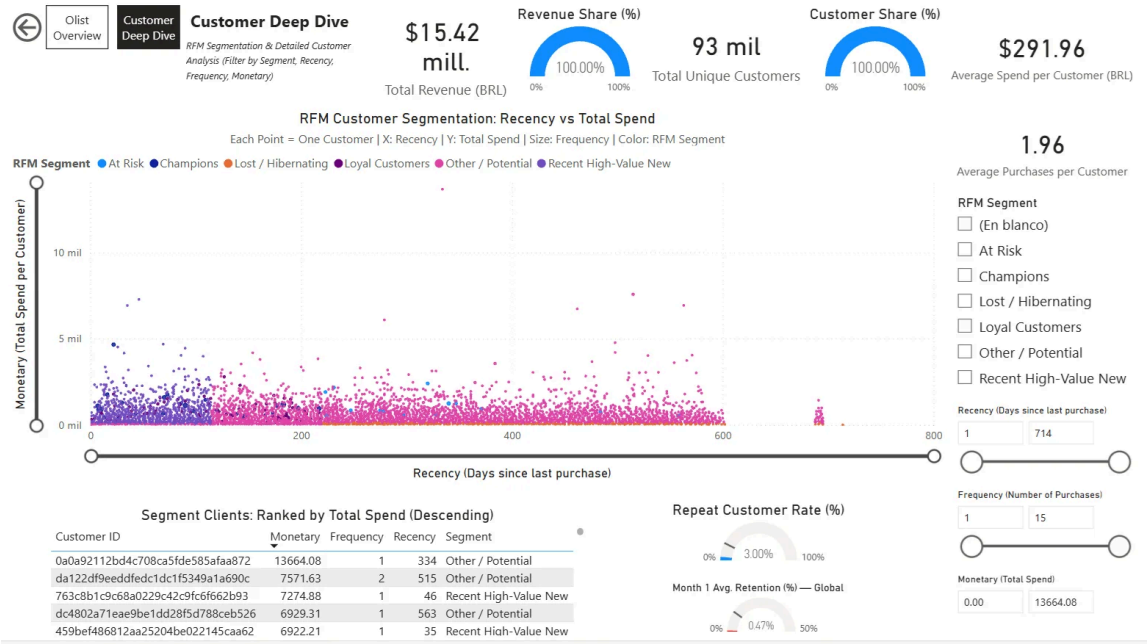
- Scatter Plot RFM: Recency vs Monetary (dimensione per Frequency, colore per Segmento)
- Tabella dinamica clienti: Filtrata e ordinata per segmento selezionato (monetary decrescente)
- Card semplici e Gauges per KPI specifici per segmento (es. % Repeat Customers, Spesa Media, ecc.)
- Multipli slicer: Segmento RFM, Range Recency, Range Frequency, Range Monetary — per filtraggio profondo e esplorazione di gruppi clienti.

Screenshots:

Pagina Olist Overview:



Pagina Customer Deep Dive:



Link al Dashboard:
Power BI Service – Olist Analytics Dashboard

Il dashboard è pubblicato su Power BI Service (account personale gratuito) e può essere condiviso via link per visualizzazione interattiva.

9.4. Conclusione del Progetto e Prossimi Passi

Riassunto dei Risultati:

- Caricato e pulito il dataset di E-Commerce Brasiliano Olist (~100k ordini, 9 tabelle).
- Eseguito EDA profondo: vendite per stato/categoria, preferenze regionali, performance di consegna.
- Consegnata segmentazione RFM con gruppi clienti actionable e raccomandazioni.
- Analizzata retention dei cohorts → evidenziate tassi di ripetizione molto bassi e strategie di miglioramento.
- Previsione dei ricavi futuri con Prophet → identificati trend di crescita e opportunità stagionali.
- Creati visual interattivi (Plotly) e esportate tabelle per dashboarding.

Lezioni Chiave Apprese:

- Transizione riuscita da R a Python: pandas per manipolazione dati, Prophet per forecasting, Plotly per visual interattivi.
- Realtà dell'e-commerce: tasso di riacquisto molto basso (~3 %), alta dipendenza da compratori one-time.
- Importanza del contesto business: ogni analisi ha portato a raccomandazioni concrete.
- Peculiarità di Power BI: relazioni e slicer necessitano setup attento per cross-filtering.

Miglioramenti Futuri e Idee:

- Incorporare dati esterni (es. festività brasiliane, indicatori economici) per forecasting più ricco.
- Esplorare impatto dei metodi di pagamento su segmenti e retention (boleto vs carta vs rate).
- Eseguire A/B test reali su tattiche di retention (es. email win-back, sconti) se dati live disponibili.
- Scalare a modelli più avanzati (es. deep learning per serie temporali, ML per customer lifetime value).

Questo progetto dimostra competenze di analisi dati end-to-end: dai dati raw alle raccomandazioni di business, con forti capacità in Python/SQL/visualizzazione.

Grazie per avermi seguito!

9.5 Report Pubblicato e Download

L'analisi interattiva completa è disponibile online sul mio sito web personale:

Visualizza Report Interattivo (HTML)
(Raccomandato – interattività completa, TOC cliccabile, celle codice espandibili, grafici Plotly)

Scarica Versione PDF:
Report Olist Analytics – PDF
(Export statico per lettura offline o stampa – generato dal notebook)

Entrambe le versioni si basano sullo stesso codice sorgente del Jupyter notebook disponibile nel mio repository.

Emilio Nahuel Pattini – Buenos Aires, 1 febbraio 2026
