

Olist Datasets Overview

About Dataset

Brazilian E-Commerce Public Dataset by Olist

Welcome! This is a Brazilian ecommerce public dataset of orders made at [Olist Store](#). The dataset has information of 100k orders from 2016 to 2018 made at multiple marketplaces in Brazil. Its features allows viewing an order from multiple dimensions: from order status, price, payment and freight performance to customer location, product attributes and finally reviews written by customers. We also released a geolocation dataset that relates Brazilian zip codes to lat/lng coordinates.

This is real commercial data, it has been anonymised, and references to the companies and partners in the review text have been replaced with the names of Game of Thrones great houses.

Join it With the Marketing Funnel by Olist

We have also released a [Marketing Funnel Dataset](#). You may join both datasets and see an order from Marketing perspective now!

Instructions on joining are available on this [Kernel](#).

Context

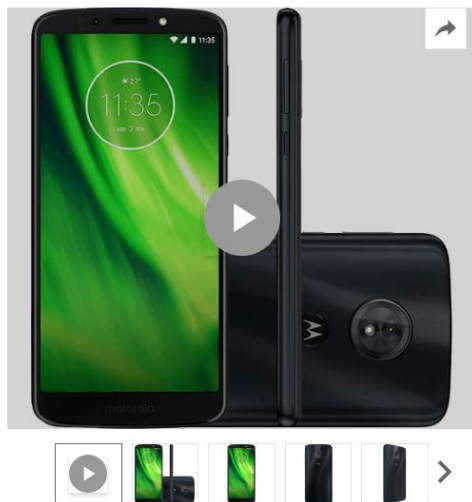
This dataset was generously provided by Olist, the largest department store in Brazilian marketplaces. Olist connects small businesses from all over Brazil to channels without hassle and with a single contract. Those merchants are able to sell their products through the Olist Store and ship them directly to the customers using Olist logistics partners. See more on our website: www.olist.com

After a customer purchases the product from Olist Store a seller gets notified to fulfill that order. Once the customer receives the product, or the estimated delivery date is due, the customer gets a satisfaction survey by email where he can give a note for the purchase experience and write down some comments.

Attention

1. An order might have multiple items.
2. Each item might be fulfilled by a distinct seller.
3. All text identifying stores and partners where replaced by the names of Game of Thrones great houses.

Example of a product listing on a marketplace



Smartphone Motorola Moto G6 Play Dual Chip Android Oreo - 8.0 Tela 5.7" Octa-Core 1.4 GHz 32GB 4G Câmera 13MP - Índigo

(Cód.133453169) ★★★★★ (215)

☐ Caixa de Som ANKER SoundCore Bluetooth 12W - Preta
+ R\$ 429,99

pegue na loja hoje!

Pegue na loja mais próxima, no mesmo dia :)
Sujeito à alteração de preço. [Saiba mais](#)

[ver lojas](#)

Escolha uma loja abaixo e compre

olist

R\$ 1.299,00
R\$ 26,04 - 7 a 10 dias úteis

on: ra

☐
R\$ 1.069,90
R\$ 38,32 - 7 a 10 dias úteis

mel: cê

☐
R\$ 975,00
R\$ 22,94 - 5 a 6 dias úteis

Mais opções deste produto a partir de **R\$ 959,00** >

vendido e entregue por **olist**

R\$ 1.299,00
10x de R\$ 129,90 s/ juros

Corra! Temos apenas 5 no estoque

☒ **R\$ 1.299,00** em até 12x de R\$ 108,25 s/ juros

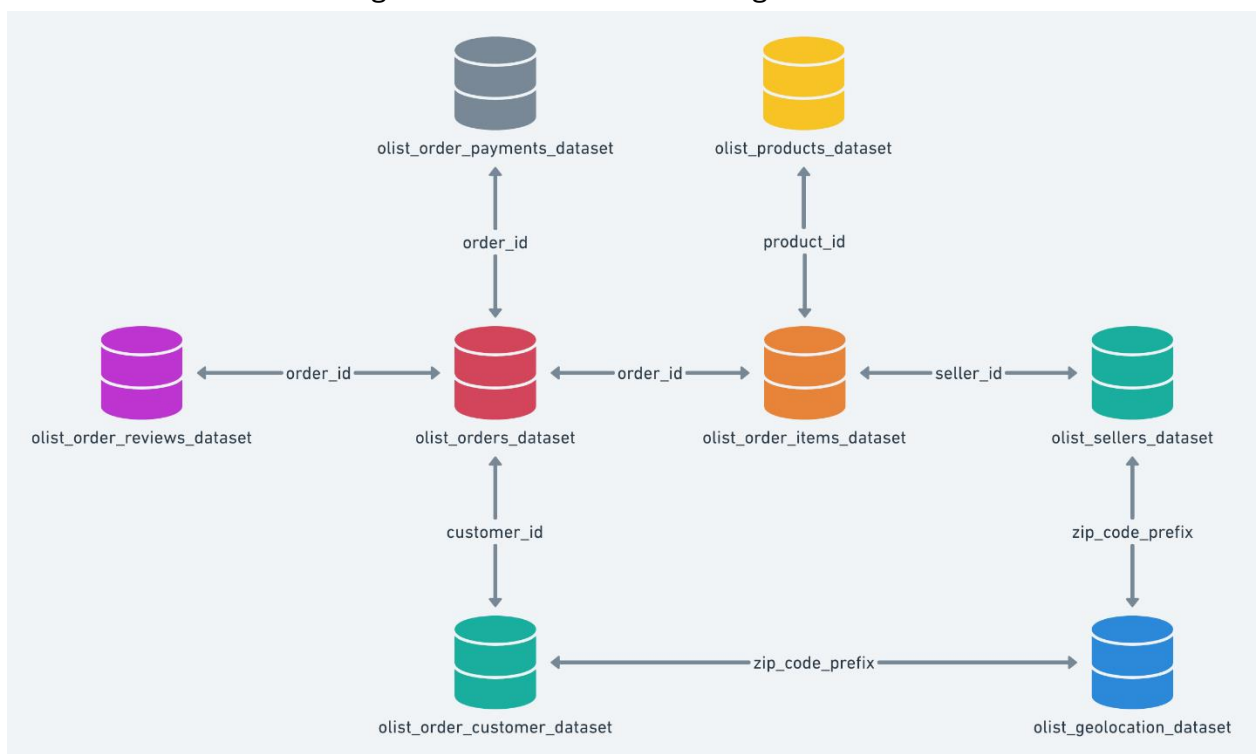
☐ **R\$ 1.299,00** no cartão : em até 24x de R\$ 54,12 s/ juros

[formas de parcelamento](#)

:) Este produto é vendido por uma loja parceira.

Data Schema

The data is divided in multiple datasets for better understanding and organization. Please refer to the following data schema when working with it:



Classified Dataset

We had previously released a classified dataset, but we removed it at *Version 6*. We intend to release it again as a new dataset with a new data schema. While we don't finish it, you may use the classified dataset available at the *Version 5* or previous.

Inspiration

Here are some inspiration for possible outcomes from this dataset.

NLP:

This dataset offers a supreme environment to parse out the reviews text through its multiple dimensions.

Clustering:

Some customers didn't write a review. But why are they happy or mad?

Sales Prediction:

With purchase date information you'll be able to predict future sales.

Delivery Performance:

You will also be able to work through delivery performance and find ways to optimize delivery times.

Product Quality:

Enjoy yourself discovering the products categories that are more prone to customer dissatisfaction.

Feature Engineering:

Create features from this rich dataset or attach some external public information to it.

Acknowledgements

Thanks to Olist for releasing this dataset

Olist Datasets Dictionary

Olist Customers Dataset

About this file

Suggest Edits

Customers Dataset

This dataset has information about the customer and its location. Use it to identify unique customers in the orders dataset and to find the orders delivery location.

At our system each order is assigned to a unique customer_id. This means that the same customer will get different ids for different orders. The purpose of having a customer_unique_id on the dataset is to allow you to identify customers that made repurchases at the store. Otherwise you would find that each order had a different customer associated with.

Please refer to the data schema:

customer_id	customer_unique...	# customer_zip_co...	A customer_city	A customer_state
key to the orders dataset. Each order has a unique customer_id.	unique identifier of a customer.	first five digits of customer zip code	customer city name	customer state
99441 unique values	96096 unique values	 1003100.0k	sao paulo16% rio de janeiro7% Other (77019)77%	SP42% RJ13% Other (44843)45%
06b8999e2fba1a1fbc88172c00ba8bc7	861eff4711a542e4b93843c6dd7febb0	14409	franca	SP
18955e83d337fd6b2def6b18a428ac77	290c77bc529b7ac935b93aa66c333dc3	09790	sao bernardo do campo	SP
4e7b3e00288586ebd08712fdd0374a03	060e732b5b29e8181a18229c7b0b2b5e	01151	sao paulo	SP
b2b6027bc5c5109e529d4dc6358b12c3	259dac757896d24d7702b9acbbff3f3c	08775	mogi das cruze	SP
4f2d8ab171c80ec8364f7c12e35b23ad	345ecd01c38d18a9036ed96c73b8d066	13056	campinas	SP
879864dab9bc3047522c92c82e1212b8	4c93744516667ad3b8f1fb645a3116a4	89254	jaragua do sul	SC

customer_id	key to the orders dataset. Each order has a unique customer_id.	 1003100.0k
99441 unique values		 Valid 99.4k100% Mismatched 00% Missing 00% Unique 99.4k Most Common 06b8999e...0%
customer_unique_id	unique identifier of a customer.	
96096 unique values		 Valid 99.4k100% Mismatched 00% Missing 00% Unique 96.1k Most Common 8d50f5ead...0%
# customer_zip_code_prefix	first five digits of customer zip code	
 1003100.0k		 Valid 99.4k100% Mismatched 00% Missing 00% Mean 35.1k Std. Deviation 29.8k Quantiles 1003Min 11.3k25% 24.4k50% 58.9k75% 100.0kMax

<div> <div></div> <div>customer_city</div> </div> <div>customer city name</div>				
sao paulo	16%	<div> <div></div> <div>Valid</div> <div></div> </div>		
		<div> <div></div> <div>Mismatched</div> <div></div> </div>		
rio de janeiro	7%	<div> <div></div> <div>Missing</div> <div></div> </div>		
Other (77019)	77%	<div> <div></div> <div>Unique</div> <div></div> </div>		
		<div> <div></div> <div>Most Common</div> <div></div> </div>		
			99.4k	100%
			0	0%
			0	0%
			4119	
			sao paulo	16%

<div> <div></div> <div>customer_state</div> </div> <div>customer state</div>				
SP	42%	<div> <div></div> <div>Valid</div> <div></div> </div>		
		<div> <div></div> <div>Mismatched</div> <div></div> </div>		
RJ	13%	<div> <div></div> <div>Missing</div> <div></div> </div>		
Other (44843)	45%	<div> <div></div> <div>Unique</div> <div></div> </div>		
		<div> <div></div> <div>Most Common</div> <div></div> </div>		
			99.4k	100%
			0	0%
			0	0%
			27	
			SP	42%

Olist Geolocation Dataset

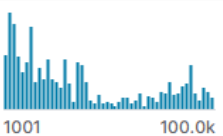
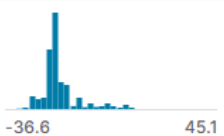

About this file

Suggest Edits

Geolocation Dataset

This dataset has information Brazilian zip codes and its lat/lng coordinates. Use it to plot maps and find distances between sellers and customers.

Please refer to the data schema:

# geolocation_zip_... first 5 digits of zip code	# geolocation_lat latitude	# geolocation_lng longitude	Δ geolocation_city city name	Δ geolocation_state state
			sao paulo 14%	SP 40%
			rio de janeiro 6%	MG 13%
			Other (802212) 80%	Other (469559) 47%
01037	-23.54562128115268	-46.63929204800168	sao paulo	SP
01046	-23.546081127035535	-46.64482029837157	sao paulo	SP
01046	-23.54612896641469	-46.64295148361138	sao paulo	SP

geolocation_zip_code_prefix

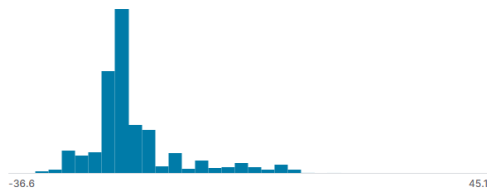
first 5 digits of zip code



Valid	1.00m	100%
Mismatched	0	0%
Missing	0	0%
Mean	36.6k	
Std. Deviation	30.5k	
Quantiles		
	1001	Min
	100.0k	Max

geolocation_lat

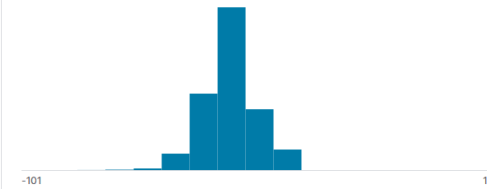
latitude



Valid	1.00m	100%
Mismatched	0	0%
Missing	0	0%
Mean	-21.2	
Std. Deviation	5.72	
Quantiles		
	-36.6	Min
	45.1	Max

geolocation_lng

longitude



Valid	1.00m	100%
Mismatched	0	0%
Missing	0	0%
Mean	-46.4	
Std. Deviation	4.27	
Quantiles		
	-101	Min
	121	Max

Δ geolocation_city

city name

sao paulo	14%
rio de janeiro	6%
Other (802212)	80%

Valid	1.00m	100%
Mismatched	0	0%
Missing	0	0%
Unique	8011	
Most Common	sao paulo	14%

Δ geolocation_state

state

SP	40%
MG	13%
Other (469559)	47%

Valid	1.00m	100%
Mismatched	0	0%
Missing	0	0%
Unique	27	
Most Common	SP	40%

Olist Orders Dataset









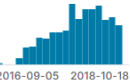
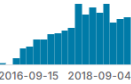
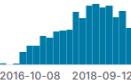
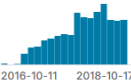
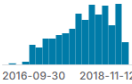
About this file


Suggest Edits

Order Dataset

This is the core dataset. From each order you might find all other information.

Please refer to the data schema:

 order_id	 customer_id	 order_status	 order_purchase_t...	 order_approved_at	 order_delivered_...	 order_delivered_...	 order_estimated_...
unique identifier of the order.	key to the customer dataset. Each order has a unique customer_id.	Reference to the order status (delivered, shipped, etc).	Shows the purchase timestamp.	Shows the payment approval timestamp.	Shows the order posting timestamp. When it was handled to the logistic partner.	Shows the actual order delivery date to the customer.	Shows the estimated delivery date that was informed to customer at the purchase moment.
99441 unique values	99441 unique values	delivered 97% shipped 1% Other (1856) 2%					
e481f51cbdc54678b7cc49136f2d6af7	9ef432eb6251297304e76186b10a928d	delivered	2017-10-02 10:56:33	2017-10-02 11:07:15	2017-10-04 19:55:00	2017-10-10 21:25:13	2017-10-18 00:00:00
53cdb2fc8bc7dce0b6741e2150273451	b0839fb4747a6c6d28dea0b8c802d7ef	delivered	2018-07-24 20:41:37	2018-07-26 03:24:27	2018-07-26 14:31:00	2018-08-07 15:27:45	2018-08-13 00:00:00
47770eb9100c2d0c44946d9cf07ec65d	41ce2a54c0b03bf3443c3d931a367089	delivered	2018-08-08 08:38:49	2018-08-08 08:55:23	2018-08-08 13:50:00	2018-08-17 18:06:29	2018-09-04 00:00:00

 order_id

unique identifier of the order.

99441
unique values

Valid

Mismatched

Missing

Unique

Most Common

99.4k

0

0

99.4k

e481f51cbd...


100%

0%

0%

0%

0%

 customer_id

key to the customer dataset. Each order has a unique customer_id.

99441
unique values

Valid

Mismatched

Missing

Unique

Most Common

99.4k

0

0

99.4k

9ef432eb6...


100%

0%

0%

0%

0%

 order_status

Reference to the order status (delivered, shipped, etc).

delivered

shipped

Other (1856)

97%

1%

2%

Valid

Mismatched

Missing

Unique

Most Common

99.4k

0

0

8

delivered


100%

0%

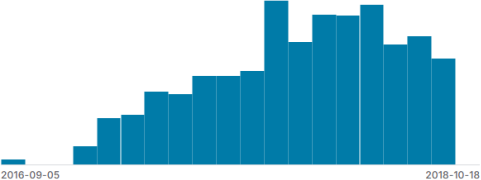
0%

0%

97%

 order_purchase_timestamp

Shows the purchase timestamp.



Valid

Mismatched

Missing

Minimum

Mean

Maximum

99.4k

0

0

5Sep16


31Dec17

18Oct18


100%

0%

0%

 order_approved_at

Shows the payment approval timestamp.



Valid

Mismatched

Missing

Minimum

Mean

Maximum

99.3k

0

160

15Sep16

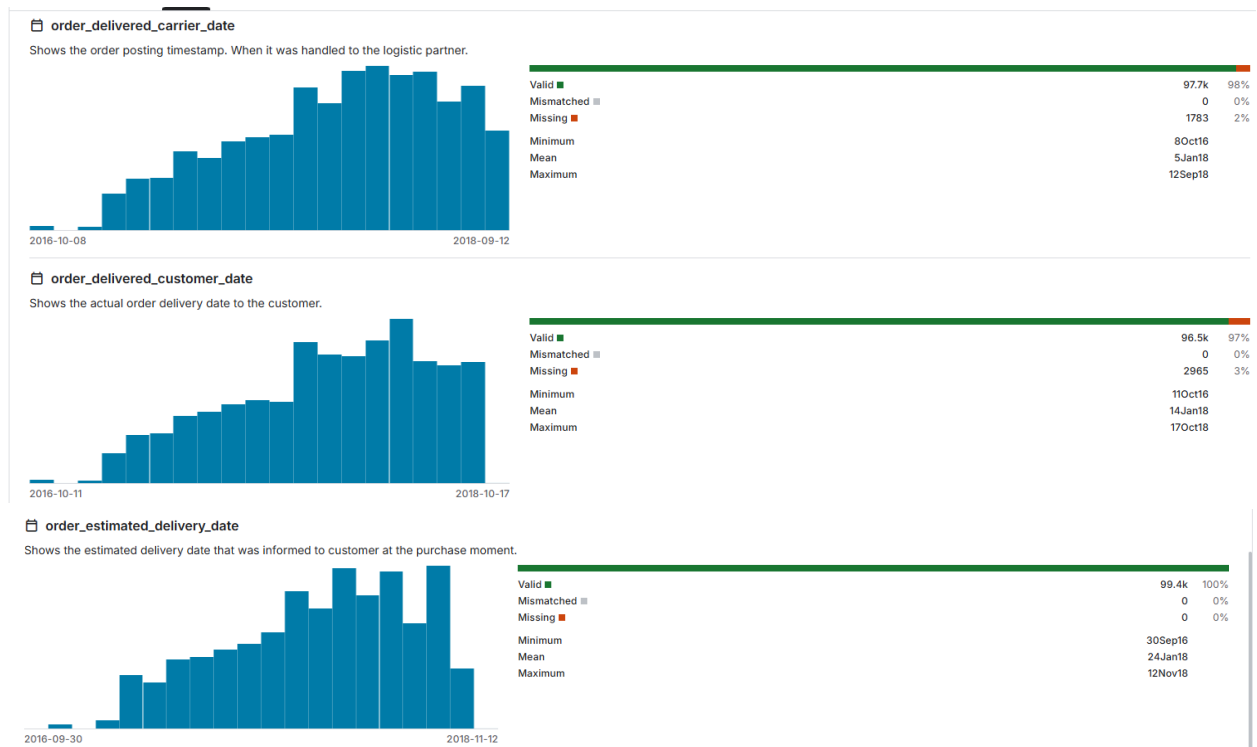
1Jan18

4Sep18

100%

0%

0%



Olist Order Items Dataset

About this file

Suggest Edits

Order Items Dataset

This dataset includes data about the items purchased within each order.

Example:

The order_id = 00143d0f86d6fbd9f9b38ab440ac16f5 has 3 items (same product). Each item has the freight calculated accordingly to its measures and weight. To get the total freight value for each order you just have to sum.

The total order_item value is: $21.33 * 3 = 63.99$

The total freight value is: $15.10 * 3 = 45.30$

The total order value (product + freight) is: $45.30 + 63.99 = 109.29$

 order_id  order unique identifier	 order_item_id  sequential number identifying number of items included in the same order.	 product_id  product unique identifier	 seller_id  seller unique identifier	 shipping_limit_date  Shows the seller shipping limit date for handling the order over to the logistic partner.	 # price item price
98666 unique values		32951 unique values	6560211a19b4799... 2% 4a3ca9315b744ce... 2% Other (108630) 96%		
00010242fe8c5a6d1ba2dd792cb16214	1	4244733e06e7ecb4970a6e2683c13e61	48436dade18ac8b2bce089ec2a041202	2017-09-19 09:45:35	58.90
00018f77f2f0320c557190d7a144bdd3	1	e5f2d52b802189ee658865ca93d83a8f	dd7ddc04e1b6c2c614352b383efe2d36	2017-05-03 11:05:13	239.90

 # price  item price	 # freight_value  item freight value item (if an order has more than one item the freight value is splitted between items)
	
0.85 6.74k	0 410
58.90	13.29
239.90	19.93
199.00	17.87











Olist Order Payments Dataset

About this file

Suggest Edits

Payments Dataset

This dataset includes data about the orders payment options.

 order_id 	# payment_sequen... 	 payment_type 	# payment_installm... 	# payment_value 
unique identifier of an order.	a customer may pay an order with more than one payment method. If he does so, a sequence will be created to	method of payment chosen by the customer.	number of installments chosen by the customer.	transaction value.
<div>99440</div> <div>unique values</div>		<div>credit_card74%</div> <div>boleto19%</div> <div>Other (7307)7%</div>		
b81ef226f3fe1789b1e8b2acac839d17	1	credit_card	8	99.33
a9810da82917af2d9aefd1278f1dcfa0	1	credit_card	1	24.39
25e8ea4e93396b6fa0d3dd708e76c1bd	1	credit_card	1	65.71

Olist Order Reviews Dataset

About this file

Suggest Edits

Order Reviews Dataset

This dataset includes data about the reviews made by the customers.

After a customer purchases the product from Olist Store a seller gets notified to fulfill that order. Once the customer receives the product, or the estimated delivery date is due, the customer gets a satisfaction survey by email where he can give a note for the purchase experience and write down some comments.

Detail		Compact	Column	7 of 7 columns	
<div><div><div><div><div></div><div></div></div></div><div><div>review_id</div><div>unique review identifier</div></div></div></div>	<div><div><div><div><div></div><div></div></div></div><div><div>order_id</div><div>unique order identifier</div></div></div></div>	<div><div><div><div><div></div><div></div></div></div><div><div># review_score</div><div>Note ranging from 1 to 5 given by the customer on a satisfaction survey.</div></div></div></div>	<div><div><div><div><div></div><div></div></div></div><div><div>review_comment...</div><div>Comment title from the review left by the customer, in Portuguese.</div></div></div></div>	<div><div><div><div><div></div><div></div></div></div><div><div>review_comment...</div><div>Comment message from the review left by the customer, in Portuguese.</div></div></div></div>	
<div>98410</div> <div>unique values</div>	<div>98673</div> <div>unique values</div>	<div><div><div><div><div></div><div></div></div></div><div><div>1</div><div>5</div></div></div></div>	<div>[null]88%</div> <div>Recomendo0%</div> <div>Other (11145)11%</div>	<div>[null]59%</div> <div>Muito bom0%</div> <div>Other (40747)41%</div>	
7bc2406110b926393aa56f80a40eba40	73fc7af87114b39712e6da79b0a377eb	4			
80e641a11e56f04c1ad469d5645fdfde	a548910a1c6147796b98fdf73dbeba33	5			
228ce5500dc1d8e020d8d1322874b6f0	f9e4b658b201a9f2ecdecbb34bed034b	5			
e64fb393e7b32834bb789ff8bb30750e	658677c97b385a9be170737859d3511b	5		Recebi bem antes do prazo estipulado.	
f7c4243c7fe1938f181bec41a392bdeb	8e6bfb81e283fa7e4f11123a3fb894f1	5		Parabéns lojas lannister adorei comprar pela Internet seguro e prático Parabéns a todos feliz Páscoa	

/ OT / columns ▾

<div>review_creation_... ▾</div> <div>Shows the date in which the satisfaction survey was sent to the customer.</div>	<div>review_answer_ti... ▾</div> <div>Shows satisfaction survey answer timestamp.</div>
<div> <div>6</div> <div>6</div> <div>6</div> <div>2016-10-02 2018-08-31</div> </div>	<div> <div>6</div> <div>6</div> <div>6</div> <div>2016-10-08 2018-10-29</div> </div>
2018-01-18 00:00:00	2018-01-18 21:46:59
2018-03-10 00:00:00	2018-03-11 03:05:13
2018-02-17 00:00:00	2018-02-18 14:36:24
2017-04-21 00:00:00	2017-04-21 22:02:06
2018-03-01 00:00:00	2018-03-02 10:26:53

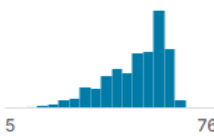
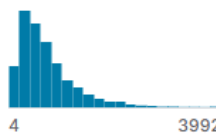
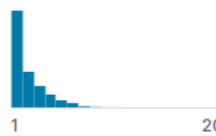
Olist Products Dataset

About this file


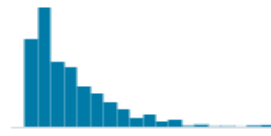
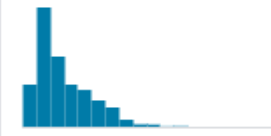
Suggest Edits

Products Dataset

This dataset includes data about the products sold by Olist.

Detail	Compact	Column	9 of 9 cc	
# product_id	product_category...	# product_name_le...	# product_descripti...	# product_photos_...
unique product identifier	root category of product, in Portuguese.	number of characters extracted from the product name.	number of characters extracted from the product description.	number of product published photos
32951 unique values	cama_mesa_banho 9% esporte_lazer 9% Other (27055) 82%			
1e9e8ef04dbcff4541ed26657ea517e5	perfumaria	40	287	1
3aa071139cb16b67ca9e5dea641aaa2f	artes	44	276	1
96bd76ec8810374ed1b65e291975717f	esporte_lazer	46	250	1

9 of 9 columns

# product_weight_g	# product_length_cm	# product_height_cm	# product_width_cm
product weight measured in grams.	product length measured in centimeters.	product height measured in centimeters.	product width measured in centimeters.
			
040.4k	7105	2105	6118
225	16	10	14
1000	30	18	20
154	18	9	15
371	26	4	26

Olist Sellers Dataset

About this file

Suggest Edits

Sellers Dataset

This dataset includes data about the sellers that fulfilled orders made at Olist. Use it to find the seller location and to identify which seller fulfilled each product.

Detail
Compact
Column

4 of 4 columns

<div> <div> seller_id </div> <div> seller unique identifier </div> </div>	<div> <div> # seller_zip_code_p... </div> <div> first 5 digits of seller zip code </div> </div>	<div> <div> seller_city </div> <div> seller city name </div> </div>	<div> <div> seller_state </div> <div> seller state </div> </div>
<div> <div>3095</div> <div>unique values</div> </div>	<div> <div> <div> <div>1001.00 - 10873.50</div> <div>Count: 1,027</div> </div> <div>100199.7k</div> </div> </div>	<div> <div> <div> sao paulo 22% </div> <div> curitiba 4% </div> <div> Other (2274) 73% </div> </div> </div>	<div> <div> <div> SP 60% </div> <div> PR 11% </div> <div> Other (897) 29% </div> </div> </div>
3442f8959a84dea7ee197c632cb2df15	13823	campinas	SP
d1b65fc7debc3361ea86b5f14c68d2e2	13844	mogi guacu	SP
ce3ad9de960102d0677a81f5d0bb7b2d	20031	rio de janeiro	RJ
c0f3eea2e14555b6faee a3dd58c1b1c3	04195	sao paulo	SP
51a04a8a6bdc23deccc82b0b80742cf	12914	braganca paulista	SP

Product Category Name Translation Dataset

About this file

Suggest Edits

Category Name Translation

Translates the product_category_name to english.

Detail
Compact
Column

2 of 2 columns

<div> <div> <div> product_category_name </div> <div> category name in Portuguese </div> </div> </div> <div> <div>71</div> <div>unique values</div> </div>	<div> <div> <div>Valid</div> <div>Mismatched</div> <div>Missing</div> </div> <div> <div>Unique</div> <div>Most Common</div> </div> </div> <div> <div>71</div> <div>100%</div> <div>0</div> <div>0%</div> <div>0</div> <div>0%</div> <div>71</div> <div>1%</div> </div> <div> <div>beleza_sau...</div> </div>
<div> <div> <div> product_category_name_english </div> <div> category name in English </div> </div> </div> <div> <div>71</div> <div>unique values</div> </div>	<div> <div> <div>Valid</div> <div>Mismatched</div> <div>Missing</div> </div> <div> <div>Unique</div> <div>Most Common</div> </div> </div> <div> <div>71</div> <div>100%</div> <div>0</div> <div>0%</div> <div>0</div> <div>0%</div> <div>71</div> <div>1%</div> </div> <div> <div>health_bea...</div> </div>