

### Limpeza e tratamento

Etapas:

- Exploração do dataset (Colab - Python)
- Upload do dataset no PowerBI
- Alteração de tipos e formatos de dados (codificação, latitude/longitude, valores, datas) para se adequarem aos padrões do PowerBI e criação de coluna para transformação da coluna lojistas\_id em nomes mais 'amigáveis' esteticamente para os dashes.
- Relacionamento entre tabelas
- Criação de métricas para verificação de valores e construção dos gráficos e tabelas.
- Verificação de hipóteses com Python

### Bases utilizadas

Dataset Olist:

olist\_geolocation\_dataset  
olist\_order\_items\_dataset  
olist\_payments\_dataset  
olist\_order\_reviews\_dataset  
olist\_products\_dataset  
olist\_lojistas\_dataset

Outros Datasets:

Sintegra PR - Contribuintes Ativos no ICMS  
Receita Federal (consultas via API)

### Observações

Defini como escopo a observação dos lojistas e sugestões para aprimoramento desta área.

Optei por usar como filtro de data a coluna order\_approved, em vez de order\_purchased, porque acredito que seja mais importante, neste caso, observar o fluxo de pagamentos do que a tendência

de intenção de compra. Caso este fosse um estudo sobre produtos, talvez considerasse `order_purchased` mais relevante.

Optei, também, por fazer todas as análises a partir do ano de 2018, por ser o mais recente disponível e por me permitir observar o crescimento dos lojistas em relação a períodos do ano anterior.

No arquivo de visualização, busquei utilizar apenas cores que podem ser facilmente identificadas por daltônicos (com exceção das cores das setas em % Crescimento, porque não tenho opção de alterar as cores). No mundo todo, 8% dos homens e 0,5% das mulheres são daltônicos (fonte: [colourblindawareness.org](http://colourblindawareness.org)) e eu tive um colega de trabalho com esta condição. Por isso, acho relevante que os materiais que produzo sejam acessíveis.

A Jhennifer mencionou, em nossa conversa, que o Olist trabalha com metas trimestrais. Por isso, todos os filtros de data possuem a opção de segmentação por trimestre.

Os dados em azul nos texto (ex: “No período selecionado, o Olist possuía 3087 lojistas.”) são dinâmicos e alteram conforme a seleção de filtros.

## **Análise**

As perguntas que me propus a responder foram:

1. Como a receita está distribuída?
2. Quem são os principais lojistas? O que eles vendem?

1. Busquei entender a distribuição da receita a partir de 3 aspectos (que estão sinalizados nas 2 primeiras abas do pbix, conforme imagem abaixo):

- Sazonalidade
- Carteiras/Lojistas
- Região

Data: 2018  
Localização: Todos

No período selecionado, o Olist possuía **3087** lojistas.

As vendas contabilizavam **R\$ 8,69M**, e a variação com relação ao mesmo período no ano anterior era de **20,2%**.



#### Distribuição de R\$ entre Lojistas

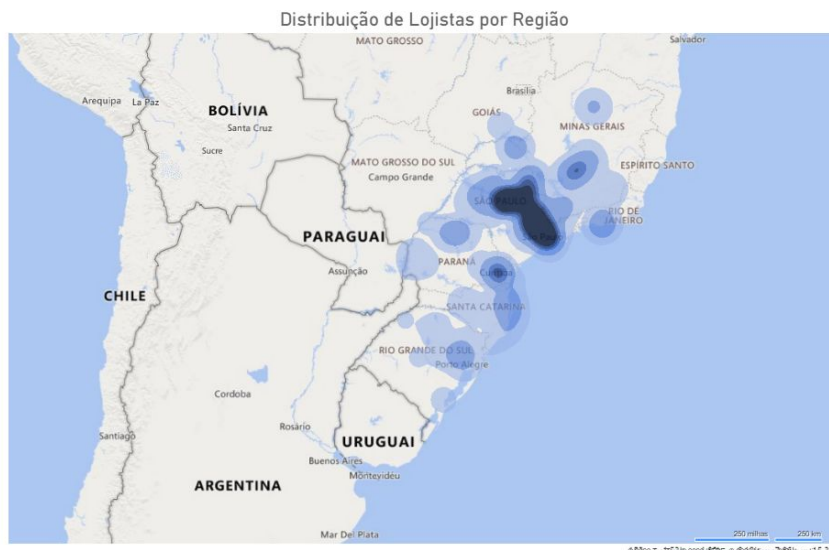
Lojistas	Vendas	Curva ABC	% do Total	% Acum	% Crescimento
Lojista2617	R\$ 153.365	A	1,8%	1,8%	53,4%
Lojista390	R\$ 140.398	A	1,6%	3,4%	549,6%
Lojista2720	R\$ 134.878	A	1,6%	4,9%	215,6%
Lojista242	R\$ 124.914	A	1,4%	6,4%	0,0%
Lojista1182	R\$ 122.031	A	1,4%	7,8%	2,4%
Lojista1873	R\$ 106.590	A	1,2%	9,0%	27,3%
Lojista557	R\$ 101.920	A	1,2%	10,2%	-1,9%
Lojista2463	R\$ 93.035	A	1,1%	11,3%	-40,3%
Lojista797	R\$ 84.967	A	1,0%	12,2%	23,0%
Lojista2746	R\$ 81.091	A	0,9%	13,2%	50,4%
Lojista474	R\$ 66.315	A	0,8%	13,9%	-19,1%
Lojista1438	R\$ 65.383	A	0,8%	14,7%	0,0%
Lojista2241	R\$ 64.794	A	0,7%	15,4%	247,3%
Lojista481	R\$ 60.444	A	0,7%	16,1%	357,5%
Lojista423	R\$ 59.280	A	0,7%	16,8%	0,0%
Lojista2340	R\$ 58.039	A	0,7%	17,5%	-15,2%
Lojista1893	R\$ 56.673	A	0,7%	18,1%	-14,9%
Lojista2191	R\$ 54.600	A	0,6%	18,8%	536,0%
Lojista1126	R\$ 53.914	A	0,6%	19,4%	0,0%
Lojista2207	R\$ 53.516	A	0,6%	20,0%	-52,6%
Lojista923	R\$ 53.443	A	0,6%	20,6%	0,0%
Lojista2345	R\$ 52.456	A	0,6%	21,2%	-21,7%
Lojista1930	R\$ 48.632	A	0,6%	21,8%	0,0%
Lojista901	R\$ 46.314	A	0,5%	22,3%	-75,6%
Lojista1111	R\$ 46.003	A	0,5%	22,8%	0,0%
Total	R\$ 8.685.500		100,0%	100,0%	20,2%

Achei muito difícil fazer alguma inferência sobre a sazonalidade das vendas do Olist como um todo, até pelo crescimento contínuo da receita. Talvez seja mais fácil observar este tipo de variação quando analisamos produtos, individualmente.

Em relação à variação por carteira/lojistas, optei pelo o conceito de curva ABC para identificar os lojistas que trazem mais receita para a empresa. Notei que 123 lojistas fazem parte da curva A no ano selecionado, o que indica uma boa distribuição do risco. Acrescentei a esta tabela uma métrica de % do Total, para identificar o impacto de cada lojista específico no montante da empresa, e também uma métrica de % Acum para conseguir checar rapidamente quantos lojistas correspondem a 10%, ou 25%, ou 80% da receita da empresa, por exemplo. Além disso, calculei o crescimento de cada lojista em relação ao mesmo período do ano passado na métrica % Crescimento, para que o setor comercial consiga acompanhar o desempenho dos lojistas e atuar rapidamente em caso de queda de receita.

Data: 2018 Localização: Todos

Estado	Nº Lojistas	Pedidos	Ped x Lojis	Vendas
SP	1814	38323	21	R\$ 5.622.039,5
PR	359	4517	13	R\$ 868.897,41
MG	249	3795	15	R\$ 636.506,11
SC	197	1868	9	R\$ 404.292,49
RJ	176	2599	15	R\$ 560.867,52
RS	131	1011	8	R\$ 239.965,66
GO	39	256	7	R\$ 43.724,79
DF	29	413	14	R\$ 54.602,31
ES	24	118	5	R\$ 26.584,18
BA	20	200	10	R\$ 84.366,23
CE	13	48	4	R\$ 16.403,85
PE	9	367	41	R\$ 94.459,13
PB	6	19	3	R\$ 6.317,94
MS	5	16	3	R\$ 3.486,96
RN	5	21	4	R\$ 2.363,73
MT	4	74	19	R\$ 12.071,91
RO	2	3	2	R\$ 359,62
SE	2	3	2	R\$ 761,09
AC	1			
AM	1			
MA	1	392	392	R\$ 48.631,88
PI	1	12	12	R\$ 3.184,5
Total	3087	53832	17	R\$ 8.685.499,54



Já no que diz respeito à distribuição regional, notei que a maior parte dos pedidos feitos no Brasil em 2018 se concentram em São Paulo, Paraná e Minas Gerais. E que o Estado com a maior com o menor número de lojistas por pedidos é o Maranhão, seguido do Pernambuco. Acredito que seria interessante promover uma ação para expandir o número de lojistas nestes Estados. Minha sugestão é a seguinte:

As Secretarias da Fazenda de cada Estado geralmente disponibilizam listas de empresas contribuintes de ICMS (como a EmpresasPR\_21.06.2020.xlsx, disponível neste repositório).

A partir destas listas é possível fazer uma limpeza de lojistas em potencial (como a do arquivo Enriquecimento\_CNPJs.xlsx, disponível neste repositório) e fazer uma consulta (via API) à Receita Federal (com o script Enriquecimento\_CNPJs.ipynb), para obter endereços, telefones, e-mails, quadros societários das empresas etc (como no modelo CNPJs\_Enriquecidos.xlsx).

Atualmente estou trabalhando em um ETL dos arquivos completos de CNPJs ativos na Receita Federal (disponíveis aqui: <https://receita.economia.gov.br/orientacao/tributaria/cadastros/cadastro-nacional-de-pessoas-juridicas-cnpj/dados-publicos-cnpj>).

Dessa forma, o setor comercial consegue mapear seu potencial de alcance e pode direcionar o time de forma mais acurada na busca por novos lojistas.

2. Inicialmente, tentei definir os principais lojistas a partir de um score (1 a 5) que eu criaria utilizando as seguintes variáveis e pesos diferentes, dependendo da relevância de cada uma:

- payment\_values (soma)
- order\_item\_id (mediana)
- order\_id (contagem distinta)
- review\_score (média)

Para isso, rodei um teste de correlação para determinar se havia variáveis que fossem parecidas demais para serem contadas 2 vezes no cálculo do score.

	payment_value	order_item_id	order_id	review_score
payment_value	1.000000	-0.008614	0.813578	0.003024
order_item_id	-0.008614	1.000000	-0.052199	-0.111426
order_id	0.813578	-0.052199	1.000000	0.022728
review_score	0.003024	-0.111426	0.022728	1.000000

Percebi que seria mais interessante utilizar ou payment\_value ou order\_id já que têm um comportamento similar. Em princípio achei que seria interessante usar payment\_value e review\_score, mas fiquei insegura quando percebi que o review\_score muito provavelmente é diferente para cada produto (por exemplo: acredito ser muito mais regular que um cliente que compra calçados dê um review do que um cliente que compra grampos de roupas. E, pelo que entendi do Kaggle, inferindo que as bases sejam as mesmas, nesta amostra apenas constam os dados de compras que possuem review, o que gera um viés enorme em qualquer análise que eu possa fazer neste sentido.).

Por conta disso, optei por seguir com a sugestão dada no próprio desafio de estudar os top10 lojistas em vendas.

Data: 2018 Localização: Todos

Neste período, os Top10 Lojistas do Olist venderam **R\$ 1,14M**, e a variação em relação ao mesmo período do ano anterior foi **52,2%**. Neste grupo, a média de score é **4,0**.



Top 10 Lojistas (por R\$)

Lojistas	Vendas	Produtos	Méd Review	% Crescimento
<b>Lojista2617</b>	<b>R\$ 153.365</b>	<b>755</b>	<b>4,1</b>	<b>53,4%</b>
relogios_presentes	R\$ 131.586	641	4,1	45,3% ↑
cool_stuff	R\$ 7.950	40	4,2	168,4% ↑
audio	R\$ 7.843	46	3,8	204,0% ↑
beleza_saude	R\$ 2.906	13	4,1	65,7% ↑
eletronicos	R\$ 1.891	10	4,8	0,0%
telefonla	R\$ 1.103	3	4,0	0,0%
esporte_lazer	R\$ 188	1	5,0	0,0%
automotivo	R\$ 179	1	5,0	0,0%
consoles_games				-100,0% ↓
informatica_acessorios				-100,0% ↓
<b>Lojista390</b>	<b>R\$ 140.398</b>	<b>1557</b>	<b>4,1</b>	<b>549,6%</b>
moveis_decoracao	R\$ 35.394	495	4,1	224,0% ↑
utilidades_domesticas	R\$ 30.279	315	3,9	429,7% ↑
esporte_lazer	R\$ 17.244	134	4,1	1165,8% ↑
beleza_saude	R\$ 14.927	123	4,1	14424,7% ↑
ferramentas_jardim	R\$ 11.136	63	4,2	4645,7% ↑
cool_stuff	R\$ 5.513	113	4,3	1233,5% ↑
industria_comercio_e_negocios	R\$ 4.277	35	4,2	0,0%
construcao_ferramentas_iluminacao	R\$ 3.342	90	4,1	0,0%
papelaria	R\$ 2.926	45	4,2	649,5% ↑
moveis_escritorio	R\$ 2.463	6	4,2	0,0%
informatica_acessorios	R\$ 2.435	22	4,1	1157,7% ↑
instrumentos_musicais	R\$ 2.374	24	4,3	1122,2% ↑
<b>Total</b>	<b>R\$ 1.139.683</b>	<b>9601</b>	<b>4,0</b>	<b>52,2%</b>

Como é possível observar pela nuvem de palavras, os 3 principais produtos comercializados pelos top10 lojistas são Relógios/Presentes, Móveis/Decoração e Cama/Mesa/Banho. Os 3 produtos estão entre os top10 produtos vendidos pelo Olist. Aqui é possível observar que alguns dos lojistas venderam menos que no ano anterior e quais foram os produtos que apresentaram queda.

## Melhorias

1. Caso eu decidisse automatizar, seria interessante adicionar à carga alguma informação sobre quando foi a última atualização do dataset.

## Dúvidas sobre o dataset

1. Alguns produtos foram entregues sem data de aprovação. Isso está correto?

Estrutura		Formatação		Propriedades		Classificar	Grupos	Relações	Cálculos				
✓													
▼	customer_id	▼	order_status	▼	order_purchase_timestamp	▼	order_approved_at	▼	order_delivered_carrier_date	▼	order_delivered_customer_date	▼	or
8	684cb238dc5b5d6366244e0e0776b450		delivered		19/01/2017				25/01/2017 14:56:50		30/01/2017 18:16:01		
19	07a2a7e0f63fd8cb757ed77d4245623c		delivered		18/02/2017				23/02/2017 03:09:14		07/03/2017 13:57:47		
3	d5de688c321096d15508faae67a27051		delivered		19/01/2017				27/01/2017 11:08:05		06/02/2017 14:22:19		
e	74beba46603f9340e3b50c6b086f992		delivered		18/02/2017				22/02/2017 11:23:11		03/03/2017 18:43:43		
	29c35fc91fc13fb5073c8f30505d860d		delivered		18/02/2017				22/02/2017 11:23:10		09/03/2017 07:28:47		
	2941af76d38100e0f8740a374f1a5dc3		delivered		18/02/2017				22/02/2017 16:25:25		01/03/2017 08:07:38		
3	2127dc6603ac33544953e05ec155771		delivered		18/02/2017				23/02/2017 12:04:47		01/03/2017 13:25:33		
	0bf35cac6cc7327065da879e2d90fae8		delivered		18/02/2017				23/02/2017 07:23:36		02/03/2017 16:15:23		
.2	f67cd1a215aae2a1074638bbd35a223a		delivered		18/02/2017				22/02/2017 11:31:06		02/03/2017 12:06:06		
3	4c1ccc74e00993733742a3c786dc3c1f		delivered		18/02/2017				23/02/2017 09:01:52		02/03/2017 10:05:06		
11	d85919cb3c0529589c6fa617f543281		delivered		17/02/2017				22/02/2017 11:31:30		03/03/2017 11:47:47		
i	68d081753ad4fe22fc4d410a9eb1ca01		delivered		19/02/2017				23/02/2017 03:11:48		02/03/2017 03:41:58		
	a3d3c38e58b9d2dfb9207cab690b6310		delivered		17/02/2017				22/02/2017 11:42:51		03/03/2017 12:16:03		
3	1e101e0daffaddce8159d25a8e53f2b2		delivered		17/02/2017				22/02/2017 11:23:11		02/03/2017 11:09:19		

2. Algumas cidades parecem estar atribuídas ao Estado errado, o que pode gerar uma falha nas minhas conclusões:

geolocation_zip_code_prefix	geolocation_lat	geolocation_lng	geolocation_city	geolocation_state
23056	-22,9191643110341	-43,6110969448295	rio de janeiro	AC
21550	-22,8578614257603	-43,3526126904033	rio de janeiro	AC