olist

Teste Prático - Análise de Dados

Anne Glienke
github.com/anneglienke

Limpeza e tratamento

Etapas:

1. Análise exploratória dos datasets (Google Colab - Python): https://colab.research.google.com/drive/1J7PyXbgIKoiL_0h1GcV1H0OZMM9YtQt8?usp=sharing. O Jupyter Notebook está disponível na pasta "analise" do repositório.
2. Upload dos datasets no Power BI.
3. Alteração de tipos e formatos de dados (codificação, latitude/longitude, valores, datas) para se adequarem aos padrões do Power BI e criação de coluna para transformação da coluna seller_id em nomes mais "amigáveis" esteticamente para os dashes.
4. Relacionamento entre tabelas no Power BI.
5. Criação de métricas para verificação de valores e construção dos gráficos e tabelas.
6. Verificação de hipóteses (Google Colab - Python): https://colab.research.google.com/drive/1J7PyXbgIKoiL_0h1GcV1H0OZMM9YtQt8?usp=sharing. O Jupyter Notebook também está disponível na pasta "analise" do repositório.

Dicionário de dados: disponível na pasta “analise” do repositório.

Bases utilizadas

Dataset Olist:
olist_geolocation_dataset
olist_order_items_dataset
olist_payments_dataset
olist_order_reviews_dataset
olist_products_dataset
olist_sellers_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 alterá-las). No mundo todo, 8% dos homens e 0,5% das mulheres são daltônicos (fonte: colourblindawareness.org) e eu tive um colega de trabalho com esta condição. Por isso, acho relevante que os materiais que produzo sejam acessíveis.

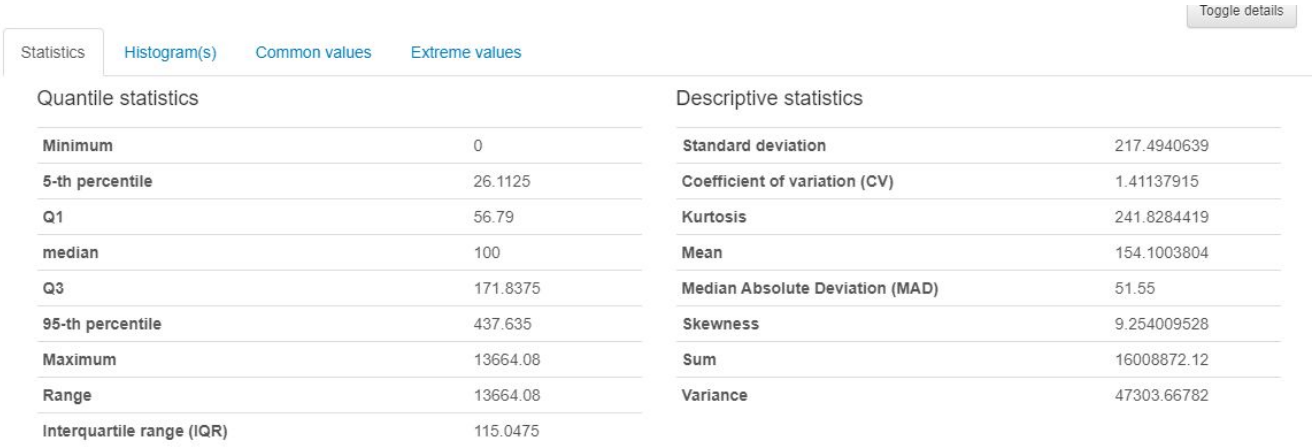
A Jhennifer mencionou, em nossa conversa, que o Olist trabalha com metas trimestrais. Por isso, todos os filtros de data possuem a opção de segmentação por trimestre.

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

Pela análise exploratória feita com o pandas-profiling notei que há poucos missing values, a maior parte diz respeito a entregas ainda não efetuadas e a detalhes de produtos, o que é esperado.

Pelos valores de mediana, máximo, variância e quartis, é possível perceber que os valores estão bem dispersos, porém com uma concentração maior entre valores mais baixos (vide Q3).



Alguns payment_values são muito baixos, fiquei na dúvida se isso é algum tipo de desconto que os clientes receberam.

Em relação ao Desafio, me propus a responder às seguintes perguntas:

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 demonstrados nos primeiros 2 dashes do pbix, conforme imagens abaixo):

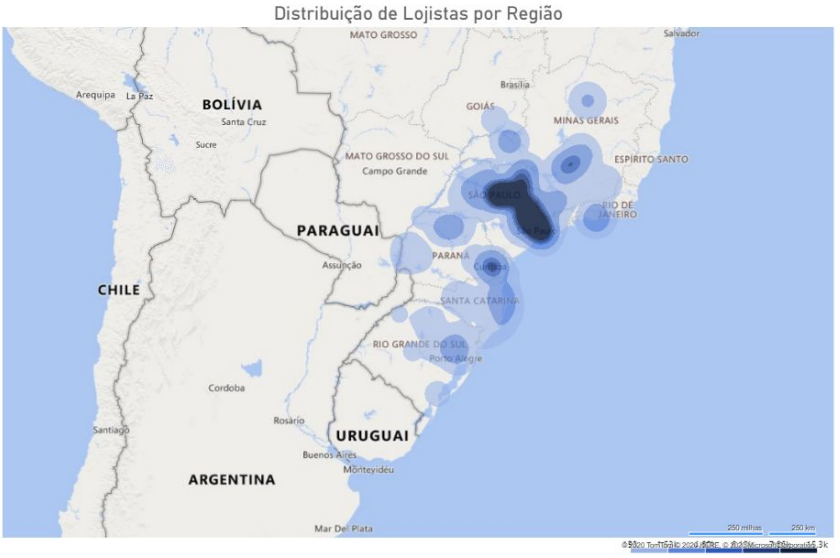
- Sazonalidade
- Carteiras/Lojistas
- Região

Não me senti segura para fazer inferências sobre a sazonalidade das vendas do Olist, especialmente pela aceleração constante do crescimento da receita. Talvez seja mais fácil observar este tipo de variação analisando produto a produto, individualmente.

Em relação à variação por carteira/lojistas, optei pelo 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 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 (em payment_values) de cada lojista em relação ao mesmo período do ano anterior na métrica % Crescimento. Assim, o setor comercial consegue acompanhar o desempenho dos lojistas e atuar rapidamente em caso de queda nas vendas.

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$ 8.317,94
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 do arquivo EmpresasPR_21.06.2020.xlxs, disponível neste repositório, na pasta Sugestao_Prospeccao_Lojistas, juntamente com os demais arquivos que vou citar adiante).

A partir destas listas é possível fazer uma limpeza de lojistas em potencial (como a do arquivo Enriquecimento_CNPJs.xlsx) 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 valer a pena 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 escolher entre utilizar payment_value ou order_id, já que as variáveis têm um comportamento similar. Em princípio, pensei em 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.). O mesmo tipo de viés também pode afetar order_items_id (por exemplo: me parece que um cliente que compra um celular está mais inclinado a comprar mais itens - capinha, película - do que um cliente que compra uma televisão.).

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

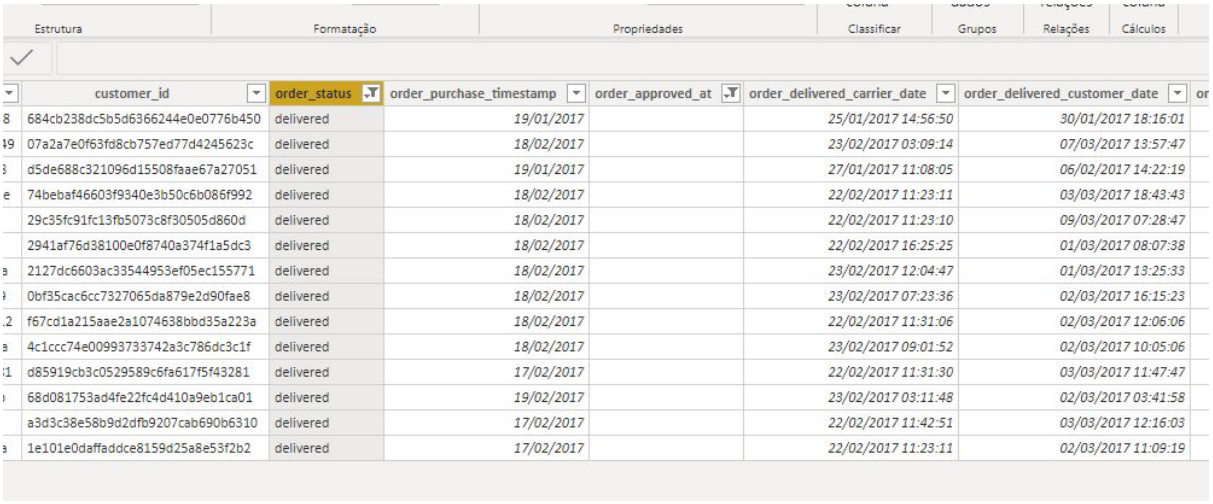
Como é possível observar pela nuvem de palavras, os 3 principais produtos comercializados pelos top10 lojistas são Relógios/Presentes, Móveis/Decoração e Cama/Mesa/Banho. Os 3 produtos estão entre os top10 produtos vendidos pelo Olist. Aqui é possível observar que alguns dos lojistas venderam menos que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

Melhorias

1. Caso eu automatizasse estes dashes, seria interessante adicionar à carga alguma informação sobre quando foi a última atualização do dataset.
2. Me falta conhecimento do negócio para definir um score que faça sentido para avaliar o desempenho dos lojistas, mas isso é algo que eu aprimoraria com mais tempo e mais informações.

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
684cb238dc5b5d6366244e0e0776b450	delivered	19/01/2017		25/01/2017 14:56:50	30/01/2017 18:16:01
07a2a7e0f63fd8cb757ed77d4245623c	delivered	18/02/2017		23/02/2017 03:09:14	07/03/2017 13:57:47
d5de688c321096d15508faae67a27051	delivered	19/01/2017		27/01/2017 11:08:05	06/02/2017 14:22:19
74bebaf46603f9340e3b50c6b086f992	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
2127dc6603ac33544953ef05ec155771	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
f67cd1a215aae2a1074638bbd35a223a	delivered	18/02/2017		22/02/2017 11:31:06	02/03/2017 12:06:06
4c1ccc74e00993733742a3c786dc3c1f	delivered	18/02/2017		23/02/2017 09:01:52	02/03/2017 10:05:06
d85919cb3c0529589c6fa617f5f43281	delivered	17/02/2017		22/02/2017 11:31:30	03/03/2017 11:47:47
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
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 invalidar parte das minhas conclusões. Qual é o campo correto?

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