MarketPlace -
Initial Data Inspection

Sellers Dataset

- There are 3,095 unique sellers
- Sellers are from ~600 different cities
- 50% of the sellers are located in 20 cities
- ~23% of the sellers are located in Sao Paulo

seller_city	
sao paulo	0.224233
curitiba	0.265267
rio de janeiro	0.296284
belo horizonte	0.318255
ribeirao preto	0.335057
guarulhos	0.351212
ibitinga	0.367044
santo andre	0.381583
campinas	0.394830
maringa	0.407754
sao jose do rio preto	0.418417
sorocaba	0.428756
sao bernardo do campo	0.439095
osasco	0.449435
porto alegre	0.458481
brasilia	0.467528
londrina	0.475929
goiania	0.483360
joinville	0.490468
blumenau	0.497254

Customers Dataset

- 99K customers (by `customer_id`)
- 96K customers (by `customer_unique_id`)
- Pareto
  - 6 states
  - 355 cities (~8.5% of total)
- `customer_city` values seems to be normalized, no characters outside the alphabet

customer_state	proportion	cumsum
SP	0.419807	0.419807
RJ	0.129242	0.549049
MG	0.117004	0.666053
RS	0.054967	0.721021
PR	0.050734	0.771754
SC	0.036574	0.808329
BA	0.033990	0.842319
DF	0.021520	0.863839
ES	0.020444	0.884283
GO	0.020314	0.904597
PE	0.016613	0.921210
CE	0.013435	0.934645
PA	0.009805	0.944449
MT	0.009121	0.953570
MA	0.007512	0.961082
MS	0.007190	0.968273
PB	0.005390	0.973663
PI	0.004978	0.978641
RN	0.004877	0.983518
AL	0.004153	0.987671
SE	0.003520	0.991191

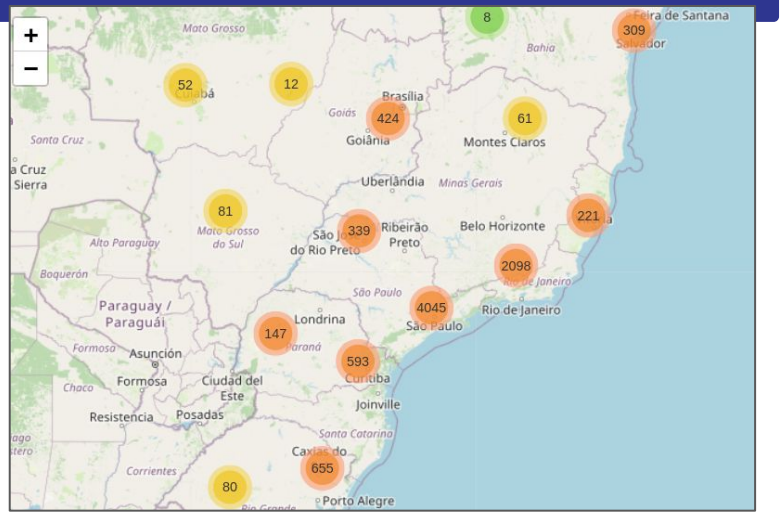
```
customer_cities.head(355)
```

✓ 0.0s

customer_city	proportion	cumsum
sao paulo	0.156274	0.156274
rio de janeiro	0.069207	0.225480
belo horizonte	0.027886	0.253366
brasilia	0.021430	0.274796
curitiba	0.015296	0.290092
...	...	...
caldas novas	0.000362	0.798866
pedreira	0.000362	0.799228
alegrete	0.000362	0.799590
campo bom	0.000362	0.799952
sao sebastiao do paraiso	0.000362	0.800314

Geo Locations Dataset

- **Reduce** data size by taking the mean (lat, lng) for each zip code
- 1M → 28K rows (97% reduction)
- `geolocation_city` values are **not normalized**:
  - Special chars
  - Non-alpha chars



geolocation_zip_code_prefix	geolocation_lat	geolocation_lng	geolocation_city	geolocation_state
99990	-28.329472	-51.769109	muliterno	RS
99990	-28.329718	-51.769615	muliterno	RS
99980	-28.386239	-51.847741	david canabarro	RS
99980	-28.386408	-51.844876	david canabarro	RS
99980	-28.386612	-51.846889	david canabarro	RS
...	...	...	...	...
1001	-23.550642	-46.634410	sao paulo	SP
1001	-23.551337	-46.634027	sao paulo	SP
1001	-23.551337	-46.634027	sao paulo	SP
1001	-23.551337	-46.634027	sao paulo	SP
1001	-23.551427	-46.634074	sao paulo	SP

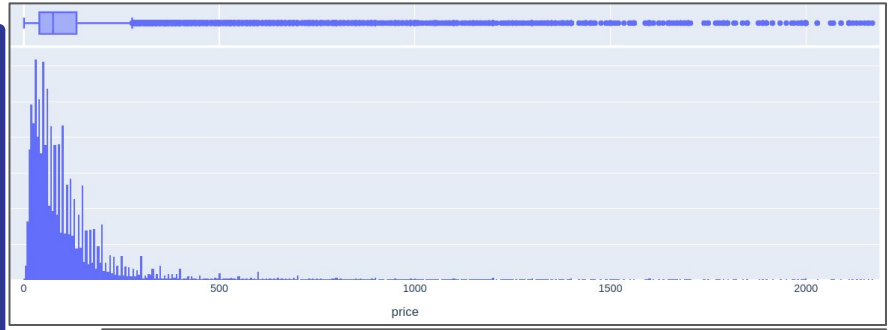
```
geolocation_clean_df[geolocation_clean_df['customer_zip_code_prefix'] == 17970]
```

✓ 0.0s

	customer_zip_code_prefix	geolocation_city	geolocation_lat	geolocation_lng
6083	17970	sao joao do pau d alho	-21.270781	-51.664380
6084	17970	sao joao do pau d%26apos%3balho	-21.269165	-51.668758
6085	17970	sao joao do pau d'alho	-21.263703	-51.667929
6086	17970	sao joao do pau dalho	-21.269116	-51.667028

Order Items Dataset

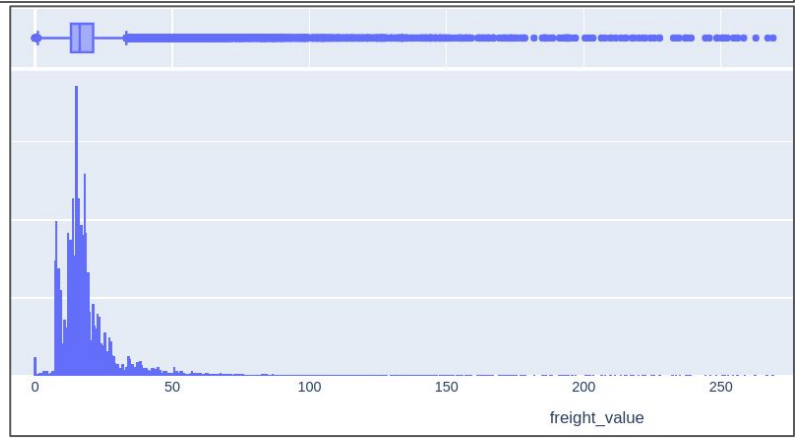
- **33K** different product IDs
- `price` and `freight_value` distribution are right-skewed.
- Freight charge:
  - Up to ~23% of product ½ of the time
  - Up to ~40% of product price ¾ of the time
  - Above product price ~4% of the time
- 3K different seller IDs, typically selling 10 or less products




```
order_items['freight_component'] = order_items.freight_value / order_items.price
order_items.freight_component.describe()
```

✓ 0.0s

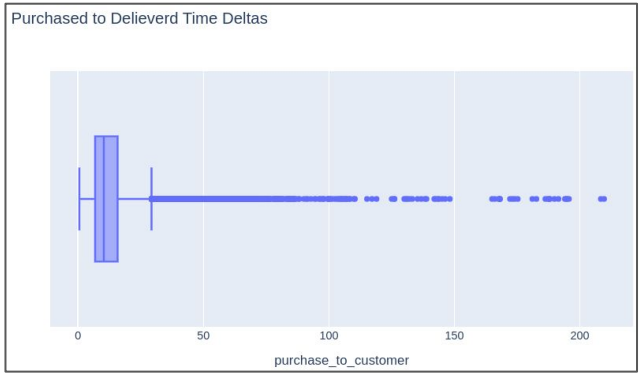
```
count    112650.000000
mean          0.320864
std           0.349894
min           0.000000
25%           0.134034
50%           0.231356
75%           0.393036
max          26.235294
Name: freight_component, dtype: float64
```

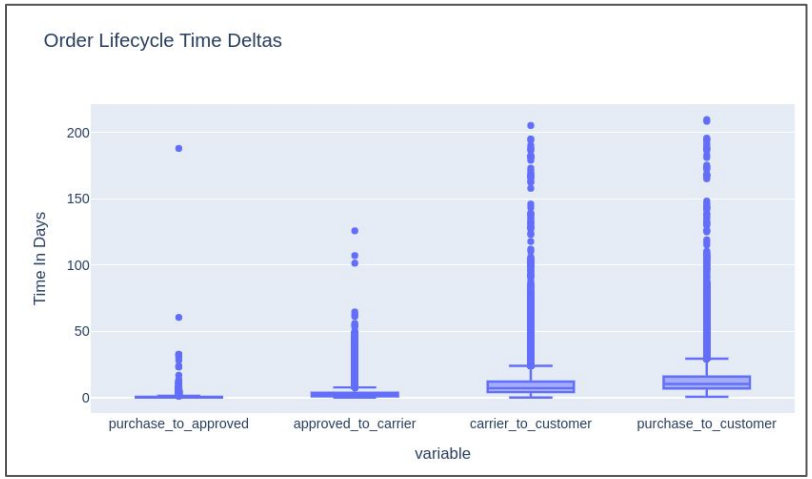


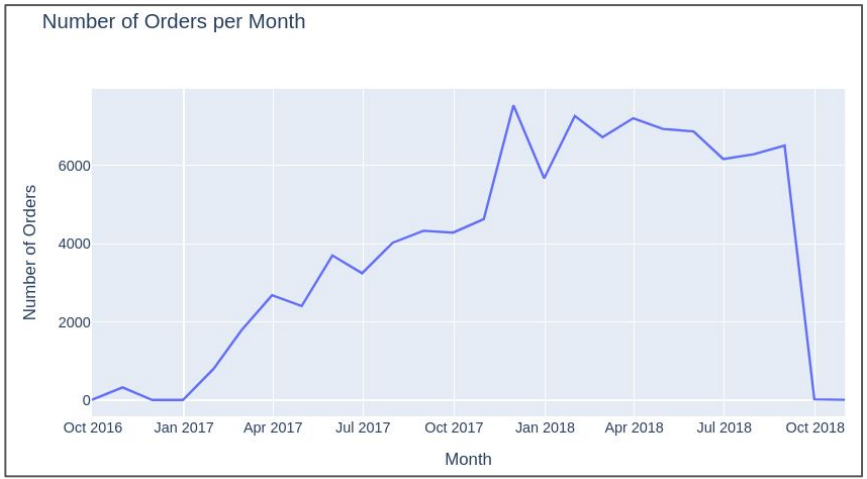
Orders Dataset

- ~2 Years of orders data (Oct 2016 - Oct 2018)
- ~1.4% of records might have **corrupted date** values
- **Very low cancellation** rate, ~0.6%
- Order cycle time is typically **2 weeks or less** (70% of the time), with some deliveries being exceptionally long.

order_status	
delivered	0.970203
shipped	0.011132
canceled	0.006285
unavailable	0.006124
invoiced	0.003158
processing	0.003027
created	0.000050
approved	0.000020







Products Dataset

- Typo in `product_name_lenght`
- 72 Product categories, 610 uncategorized product ids (~2%)
- Top 7 categories account for 52% of items.
- Product size (hXwXw) distribution is right skew.
- Product weight distribution is right skew.

product_category	
bed_table_bath	0.093658
sports_leisure	0.182307
furniture_decoration	0.264463
beauty_health	0.340033
household_utilities	0.412232
automotive	0.470981
computing_accessories	0.521660
toys	0.565289
watches_gifts	0.606382
telephony	0.641446