USJH

I create two datasets based on USJH:

1. Item unit: a dataset with all purchased items (basically copy order information to each item row).

(85925, 41)

For item unit dataset, we should not care about order unit features like ‘Discount Amount’, ‘Refund Amount’, and ‘Total’

1. Order unit: a dataset only has order information without item information (group by order # and add 1 column recording total item number in each order)

(3746, 34)

For order unit dataset, we should not care about item unit features like ‘Lineitem name’, ‘Color’.

EDA Manipulation and conclusion

1. Data type convention

'Paid at', 'Created at', 'Cancelled at' from string to datetime type

1. Missing values

Missing not at random

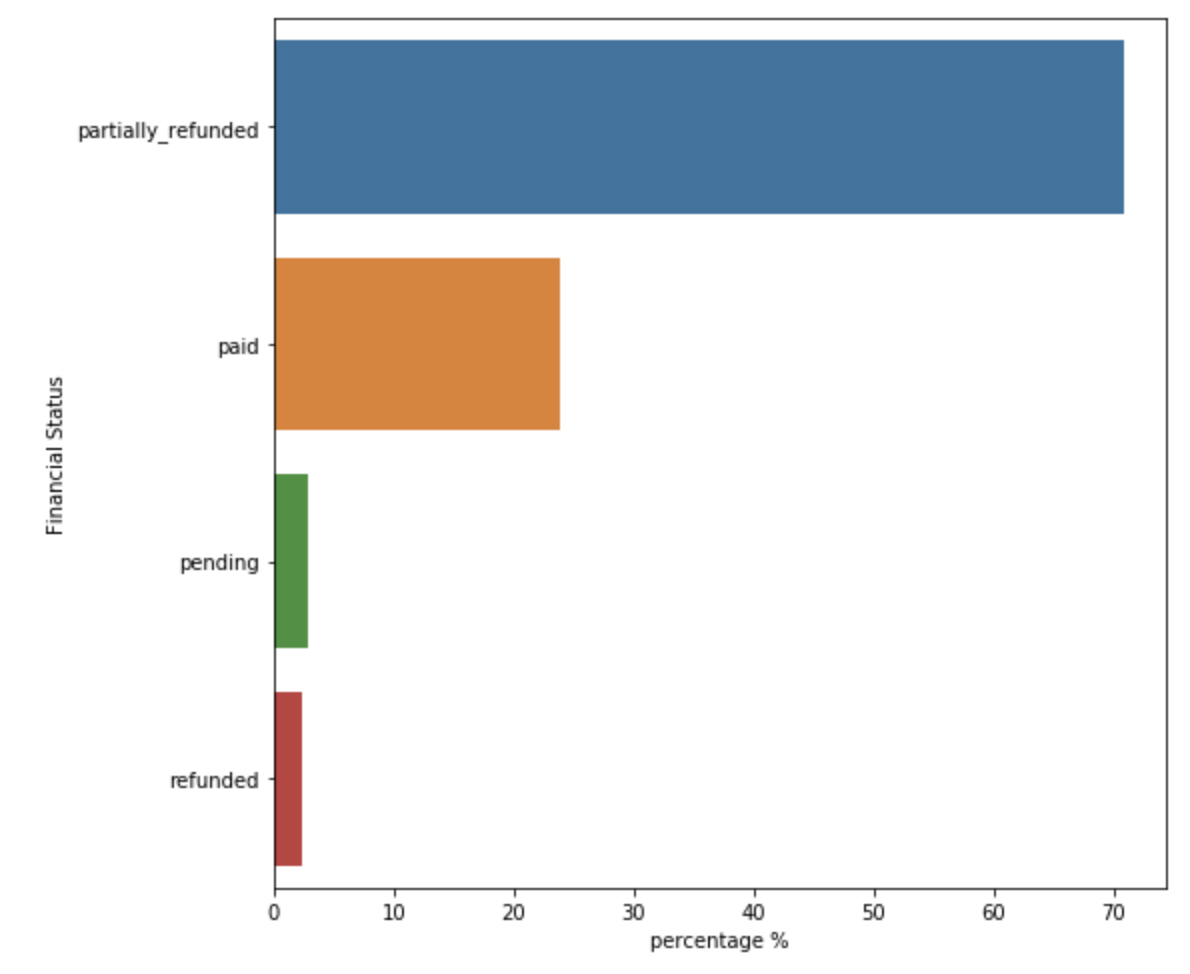
For item unit, imputed 'Note Attributes', 'Tags', 'Cancelled at', 'Notes', 'Discount Code' with ‘None’; 'Refunded Amount', 'Discount Amount' with 0.0

For order unit, imputed 'Note Attributes', 'Tags', 'Cancelled at', 'Notes', 'Discount Code' with ‘None

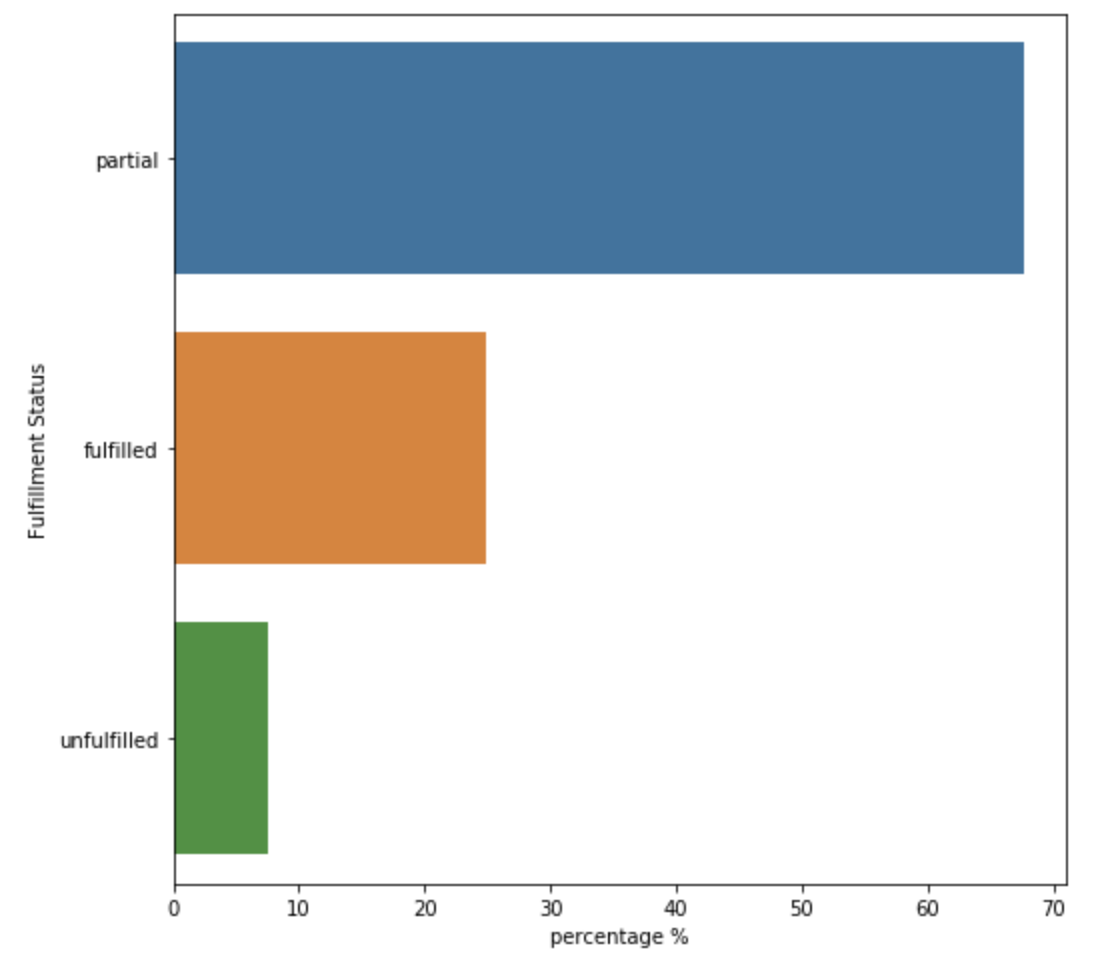
Categorical features

For USJH order unit dataset:

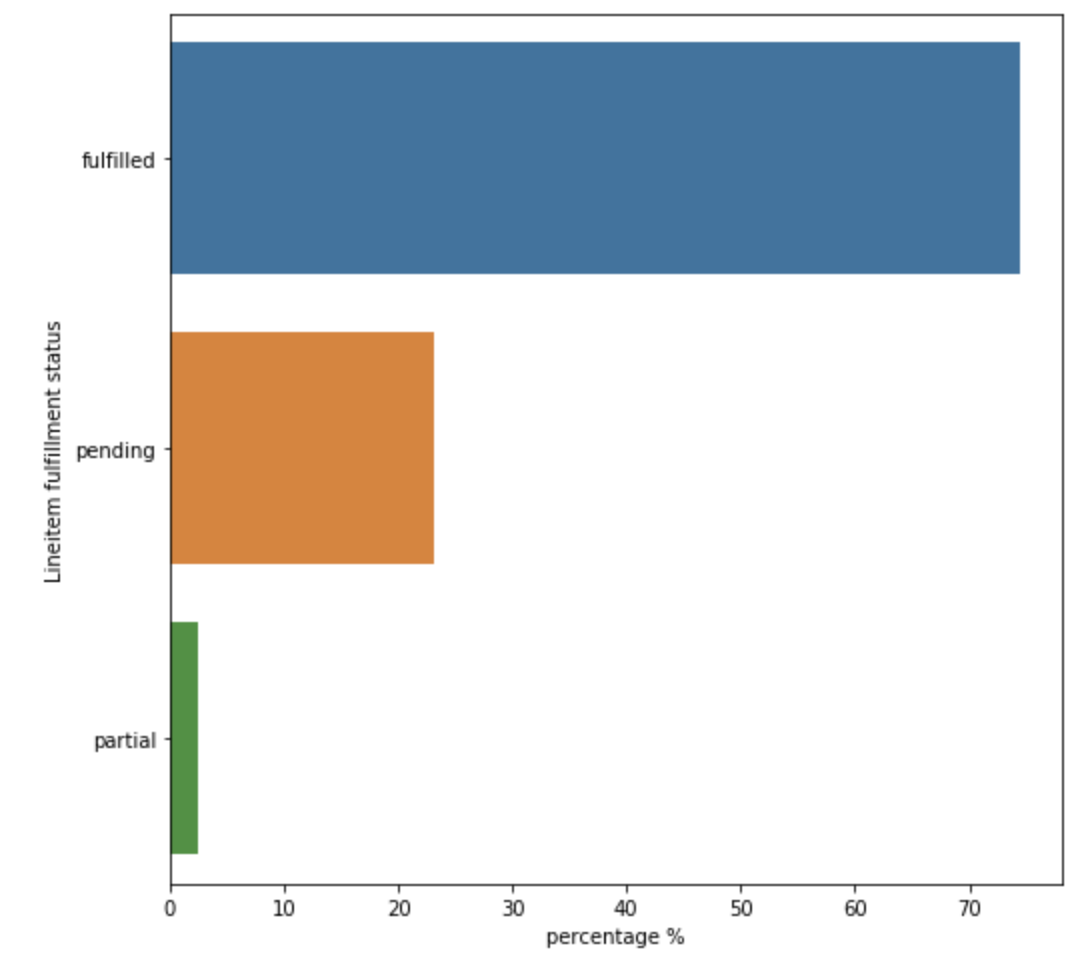
1. High partial return rate (>70%)



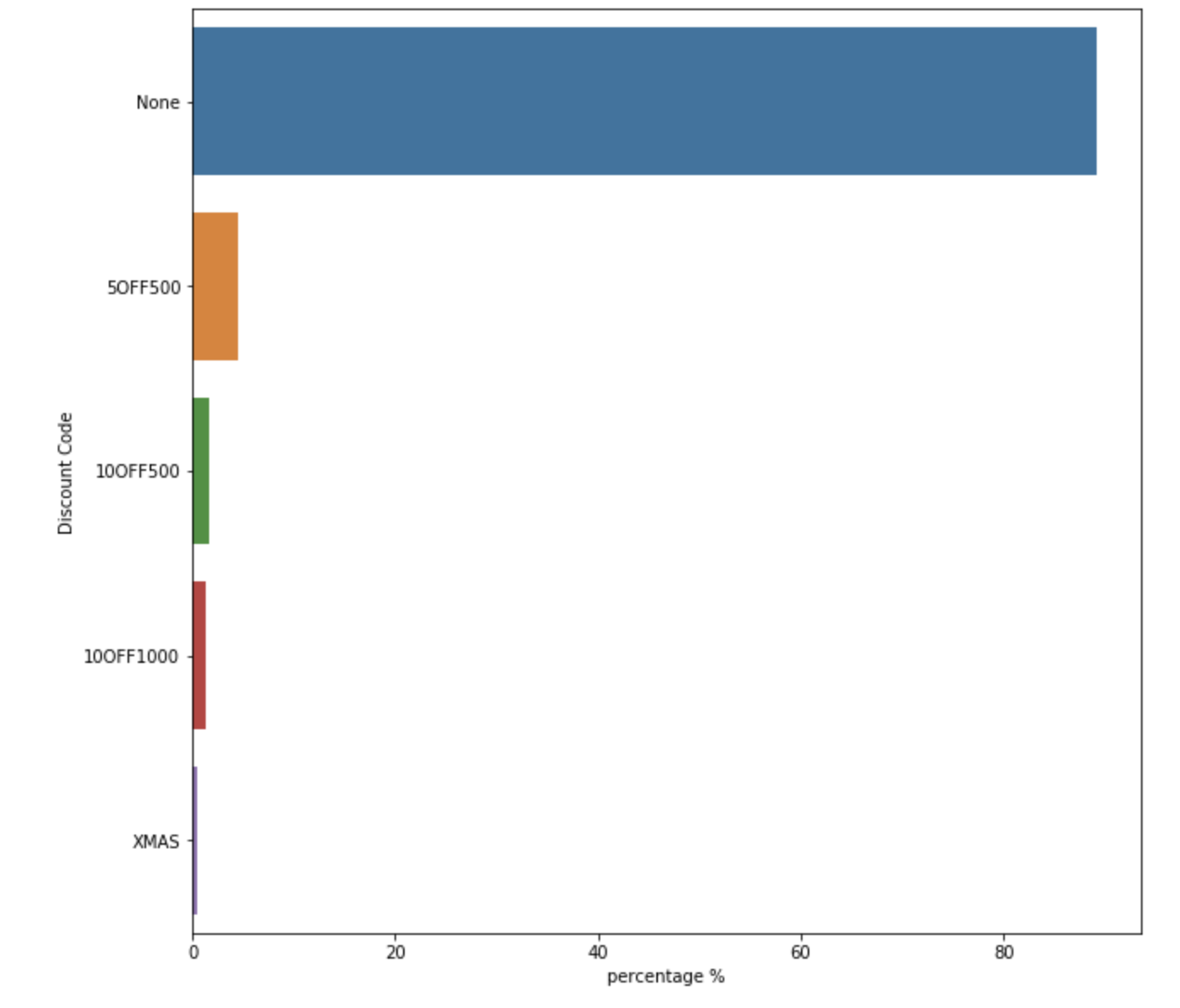
1. High partial fulfill rate (~68%)



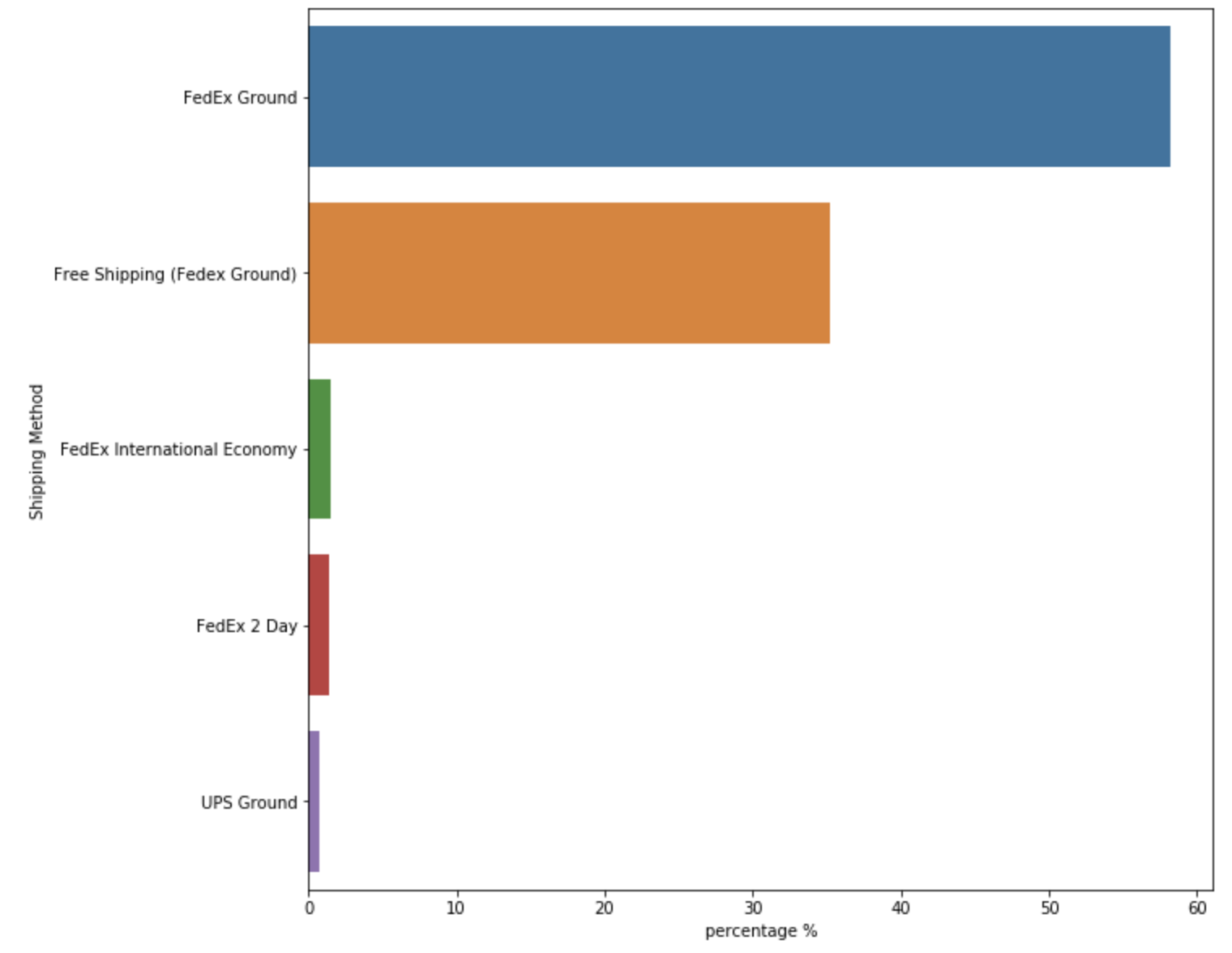
1. High lineitem fulfill rate



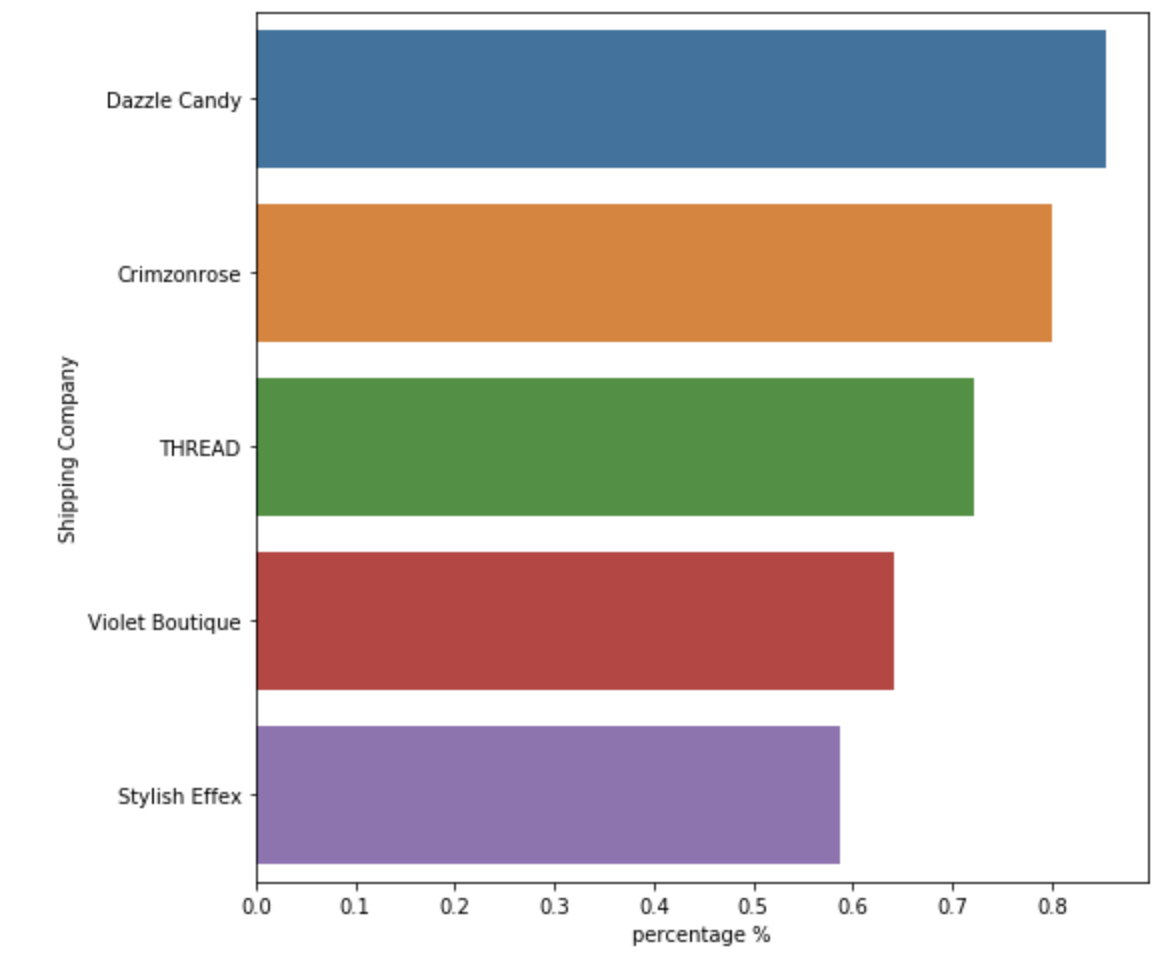
1. More than 90% customer accept marketing
2. More than 90% customer don’t apply discount code



1. FedEx Ground and Free shipping (FedEx Ground) are the most popular shipping methods



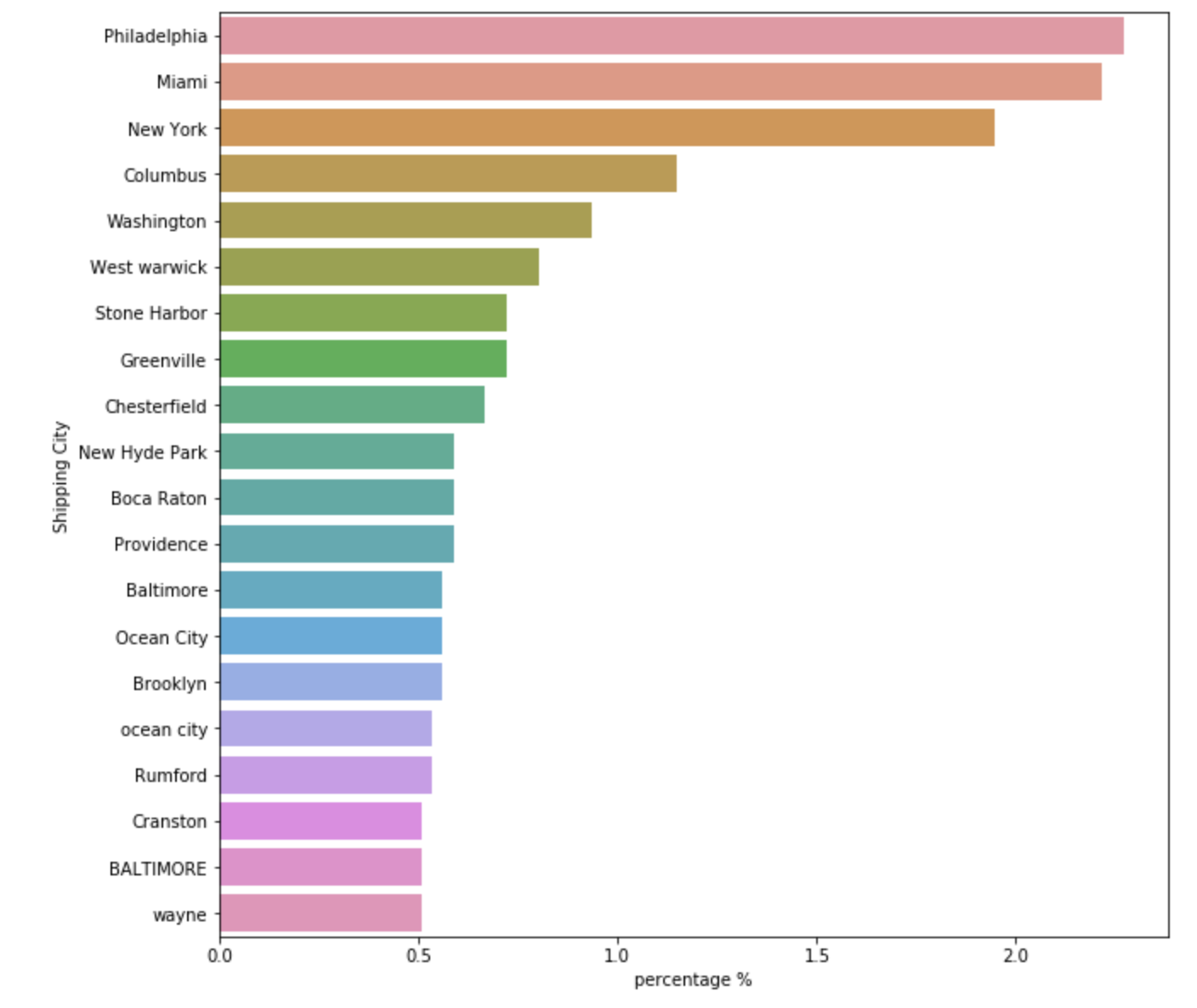
1. Top 5 shipping companies (there are 1484 companies in total)



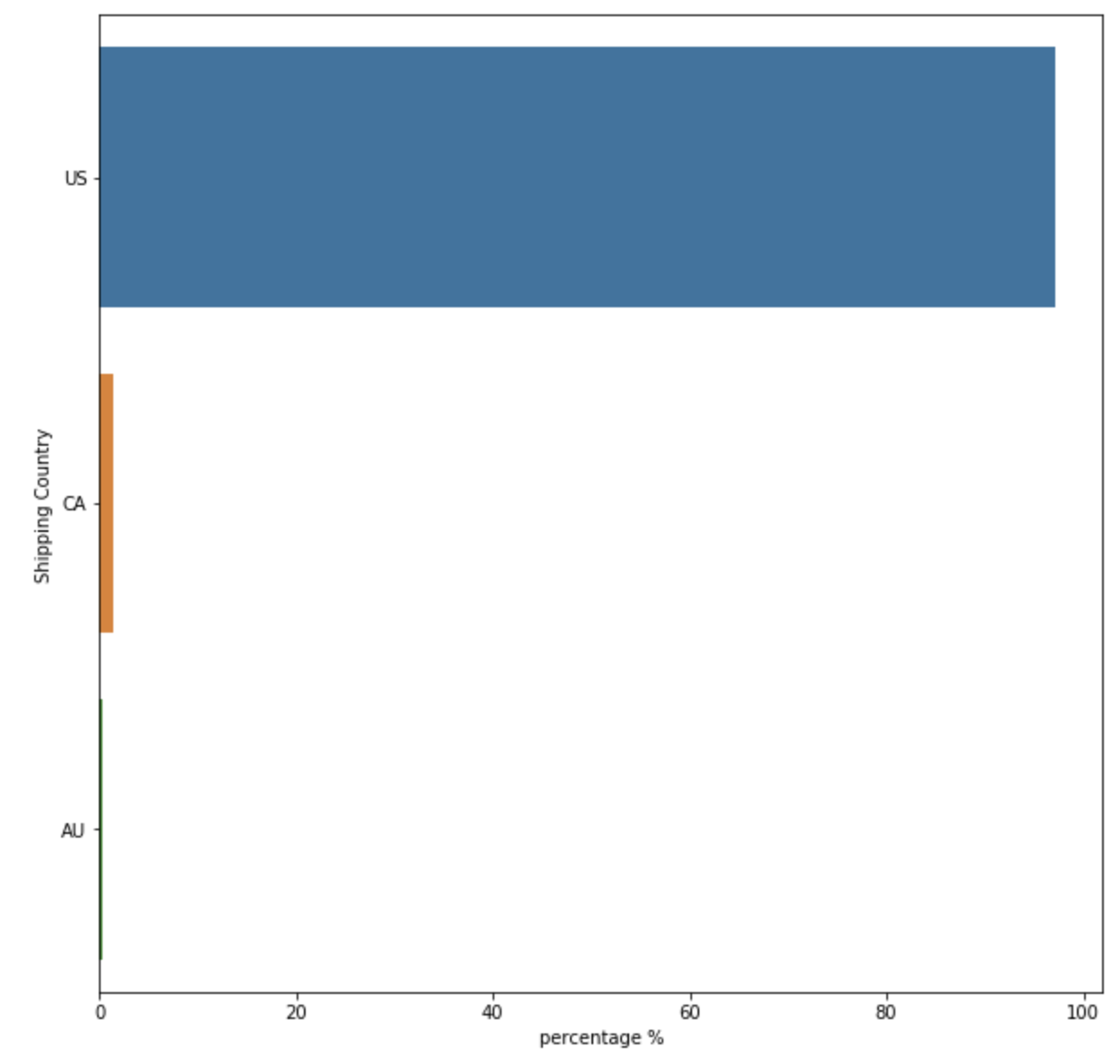
1. Most orders are shipped to NY, NJ, PA, but a lot to FL too.



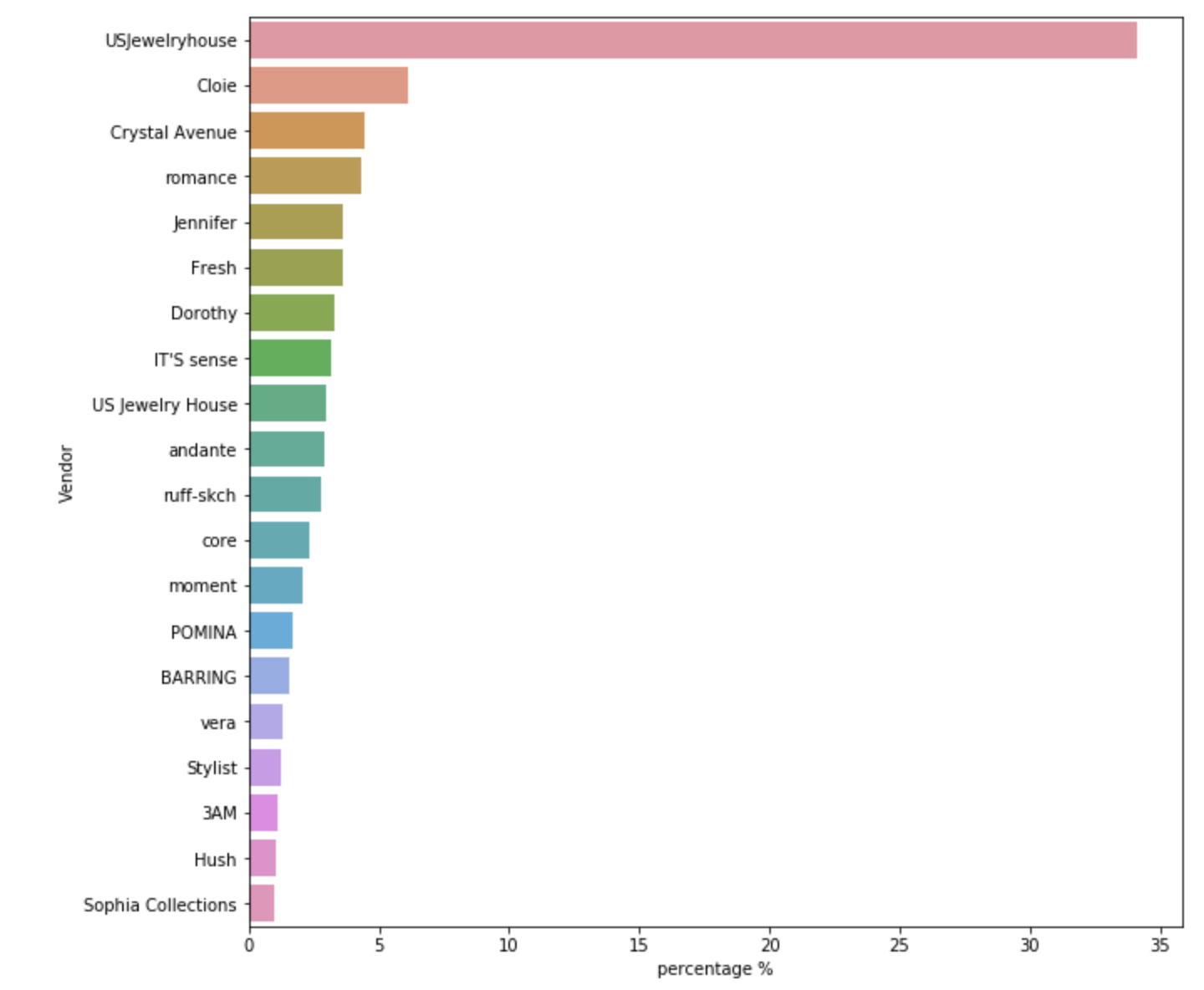
1. While most orders are from philly and miami



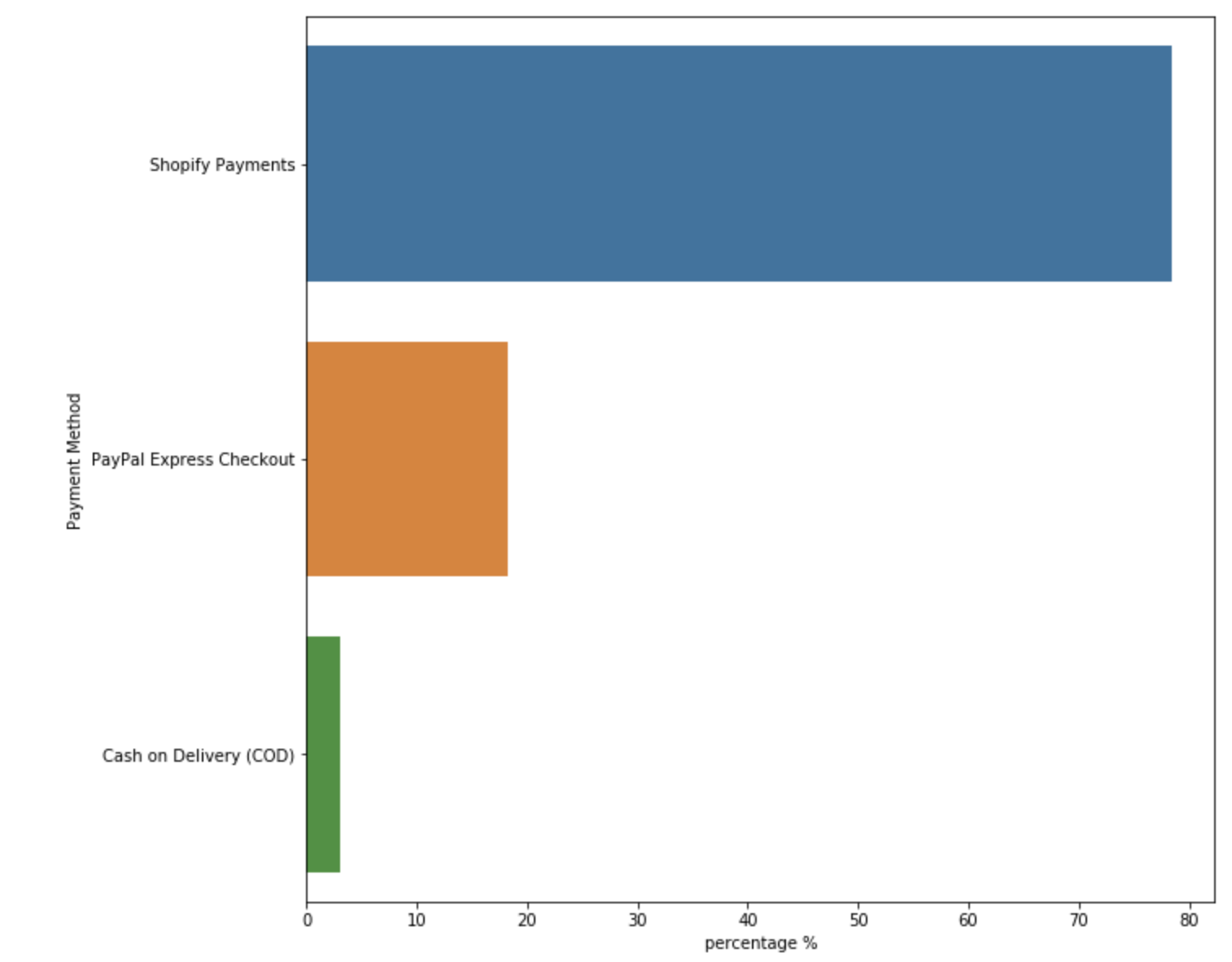
1. ~98% domestic



1. About 35% products are from USJH



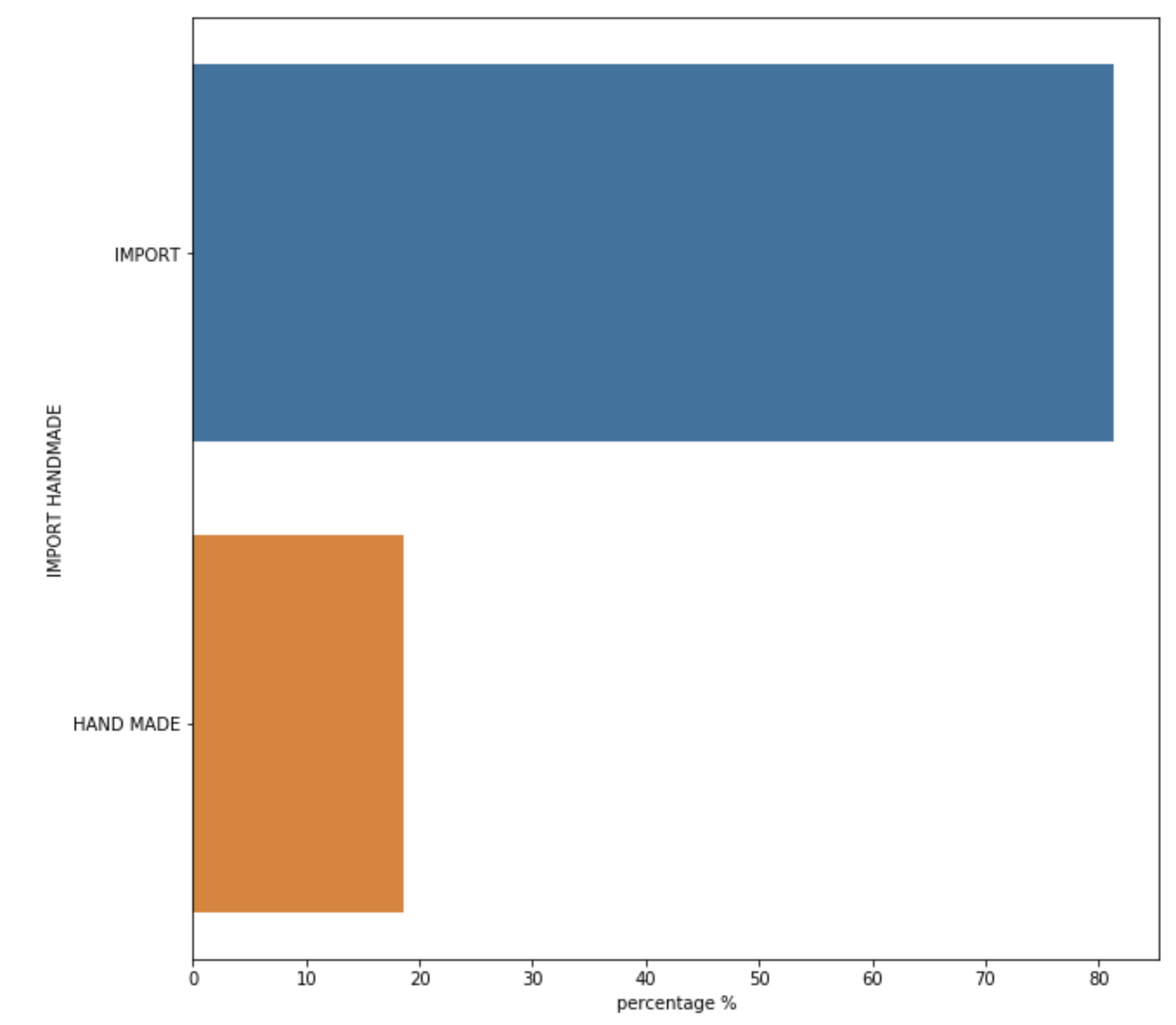
1. Shopify is the dominant payment method



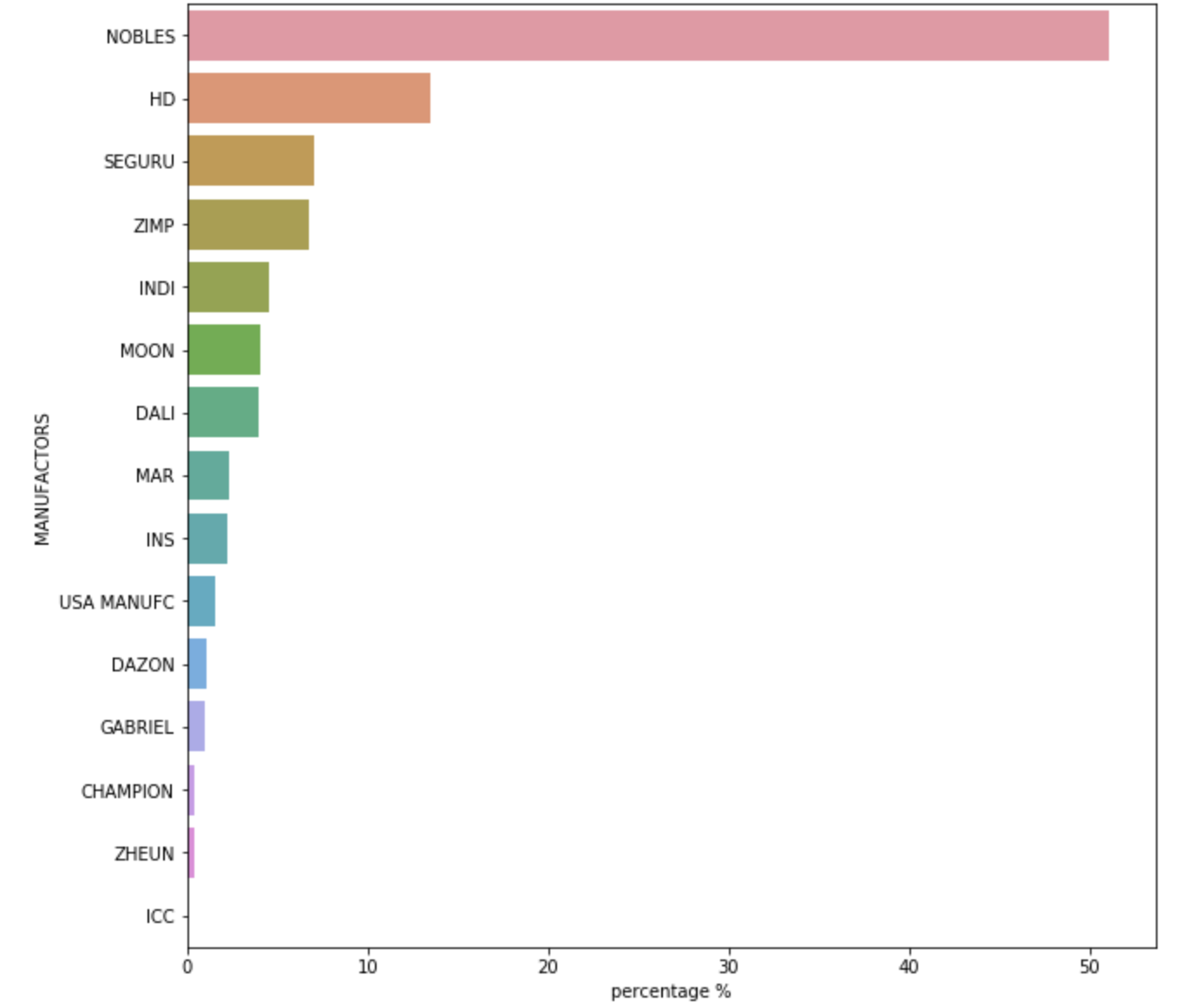
1. Very few people leave notes, most notes are about shipment
2. 98% source is web (means online shopping?)
3. Almost all orders are low risk level (99%)

For UNO Sale dataset

1. More than 80% products are imported



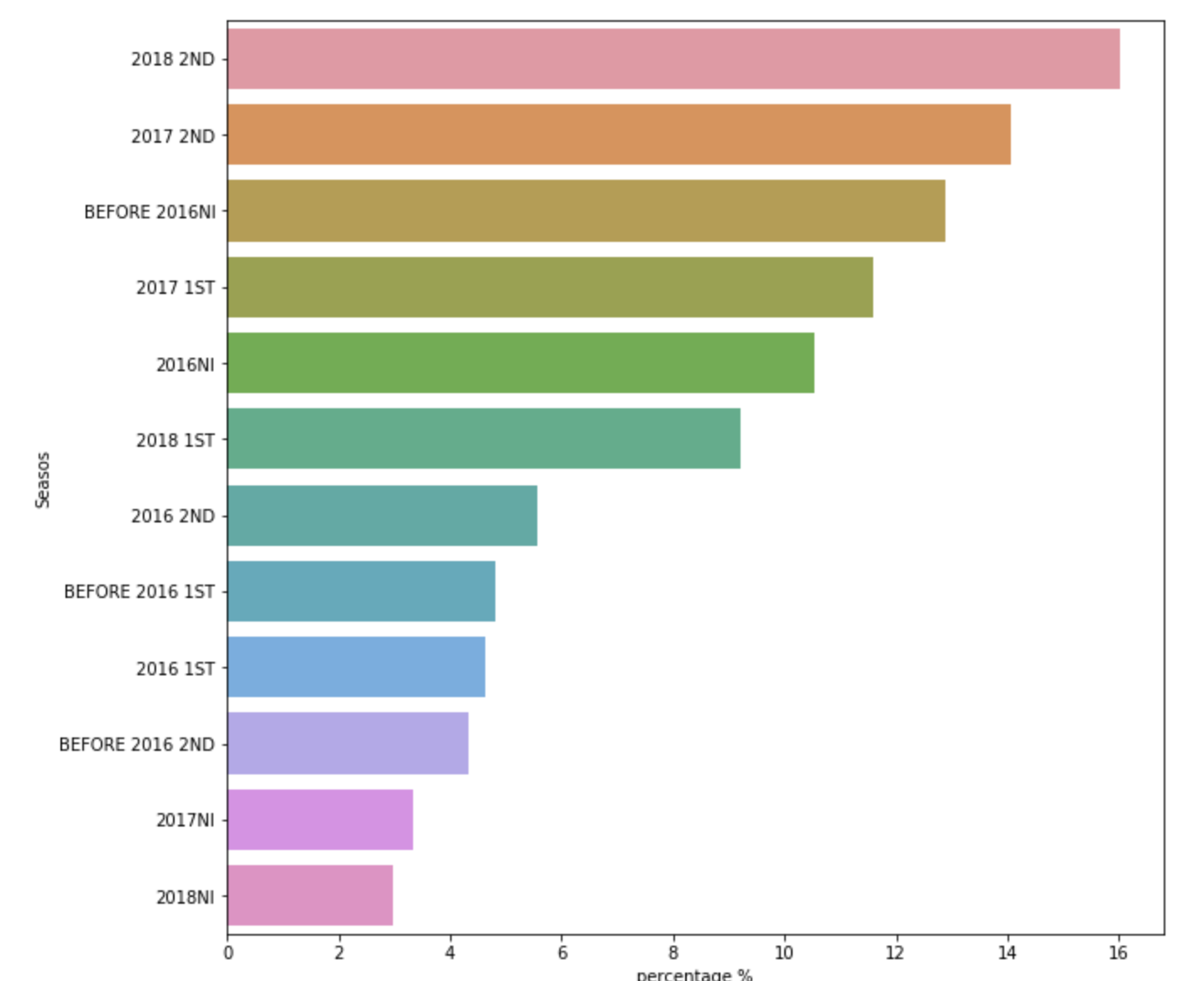
1. Manufactors, more than 50% products are from NOBLES



1. Most products shipping to CA, TX, and GA



1. Active season rank

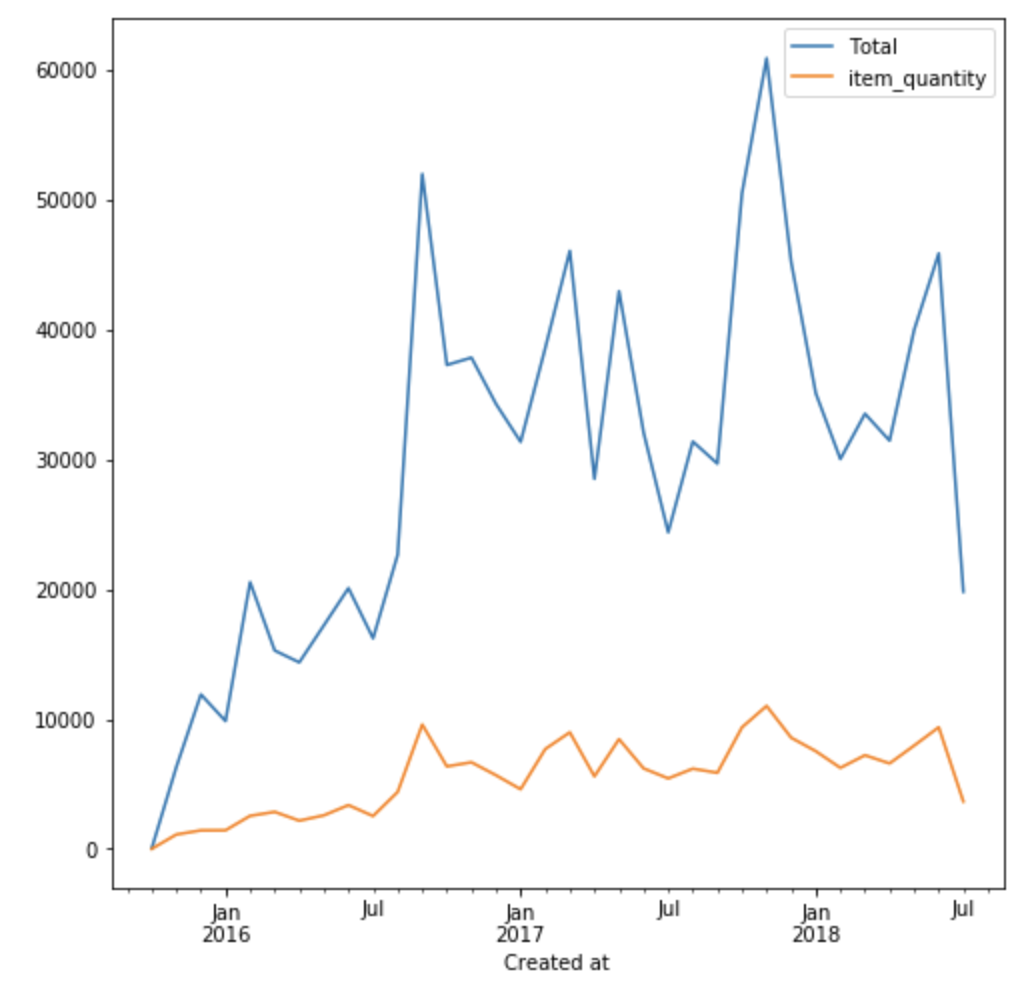
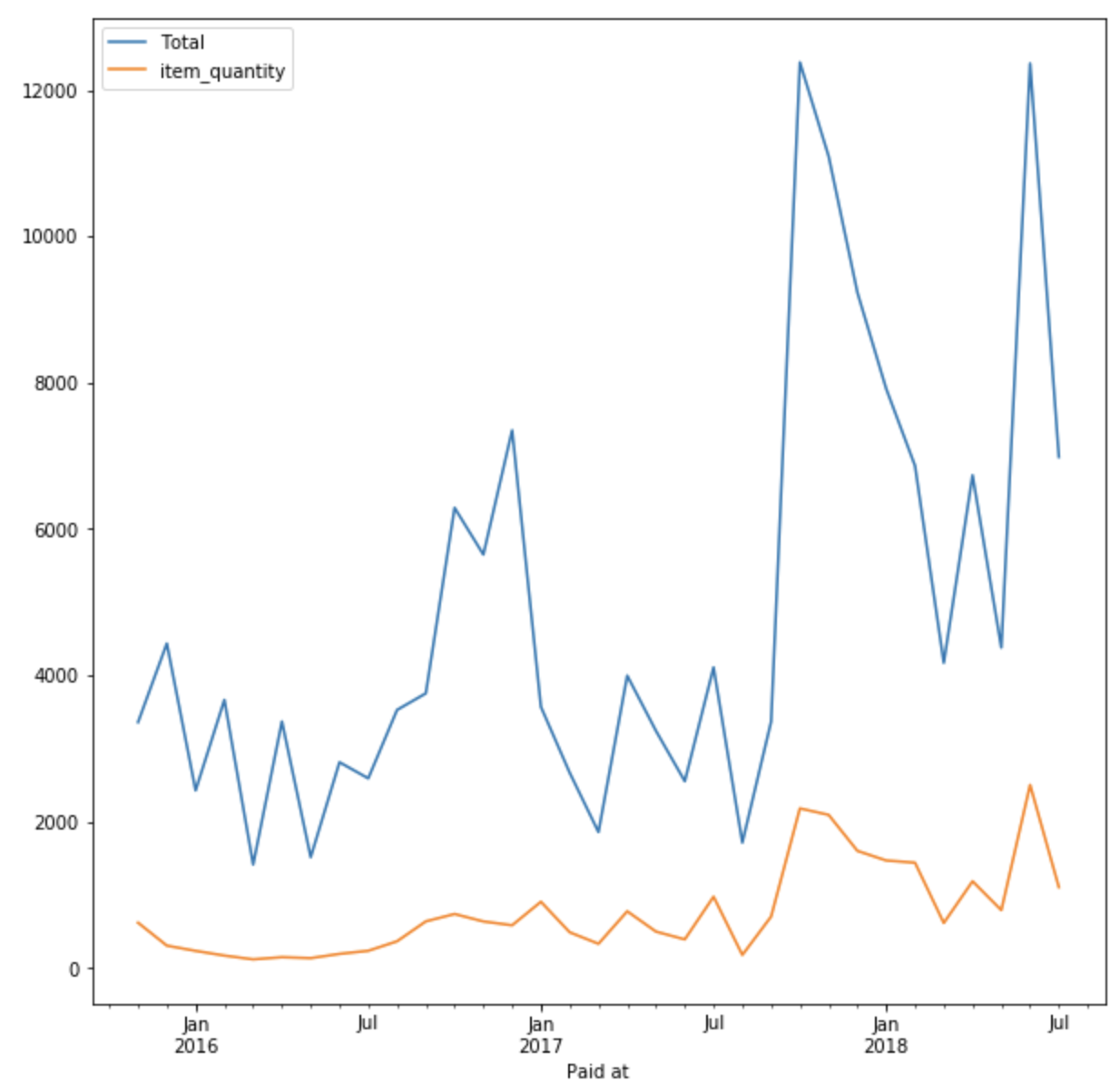


Time series Analysis

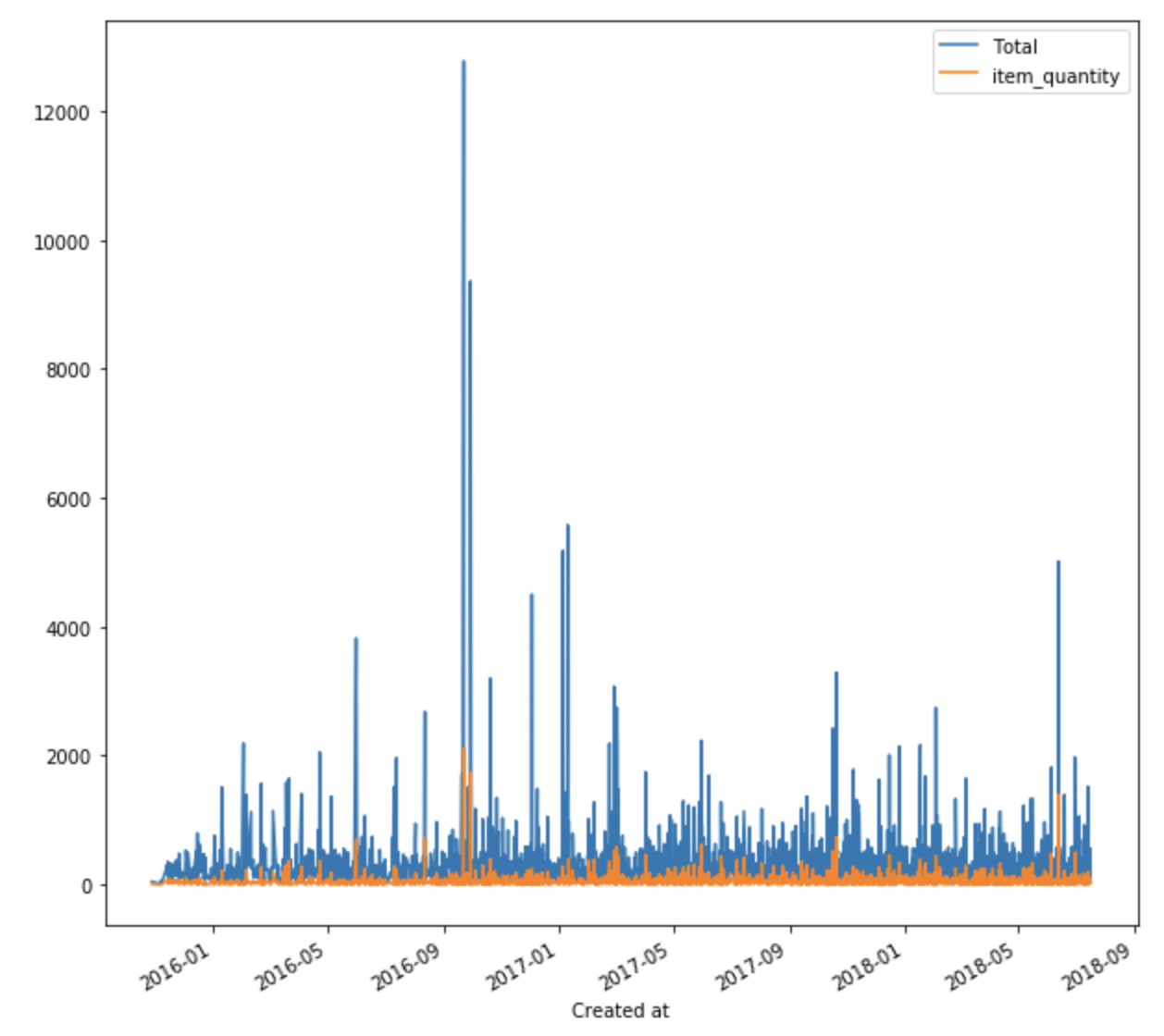
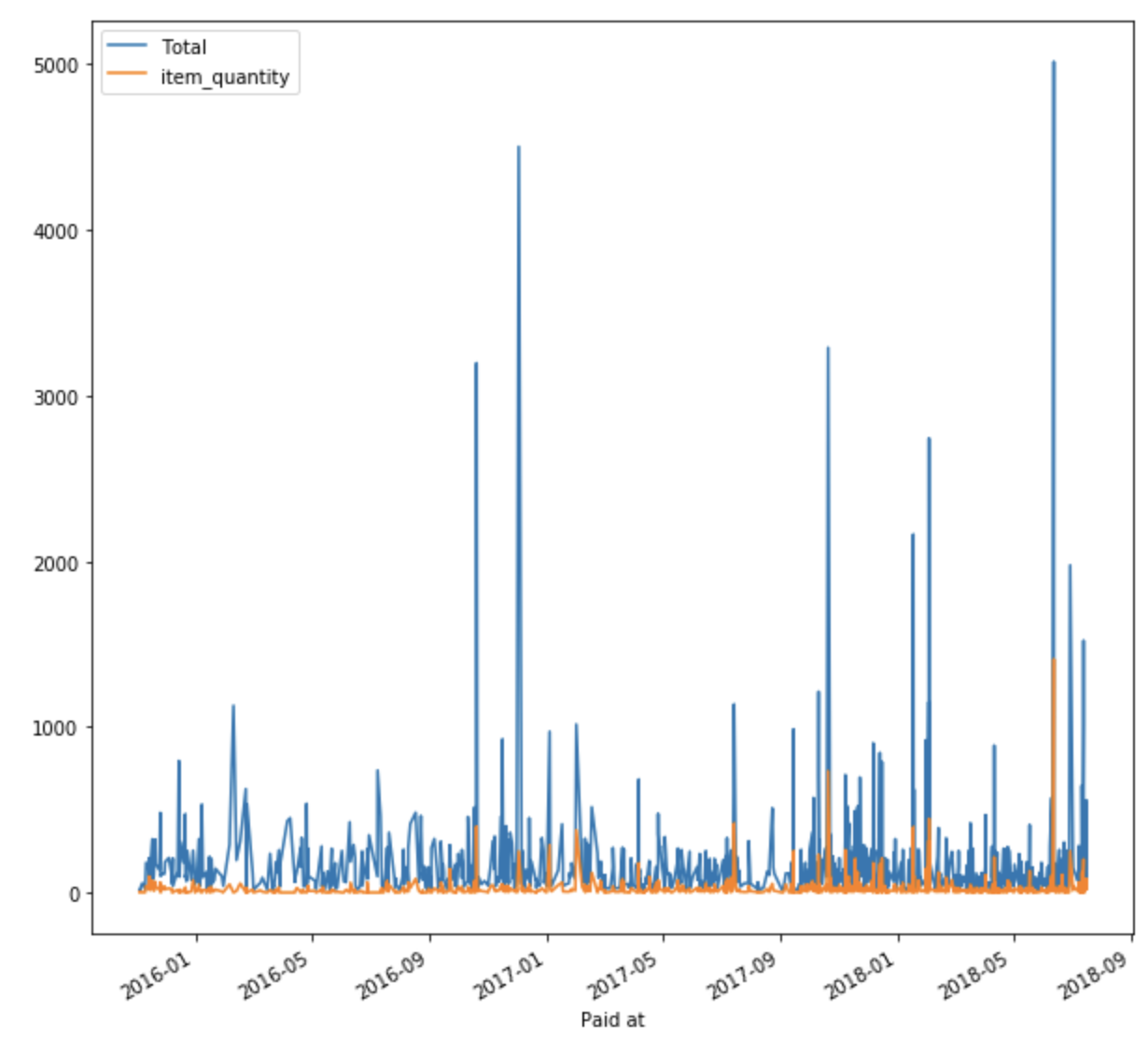
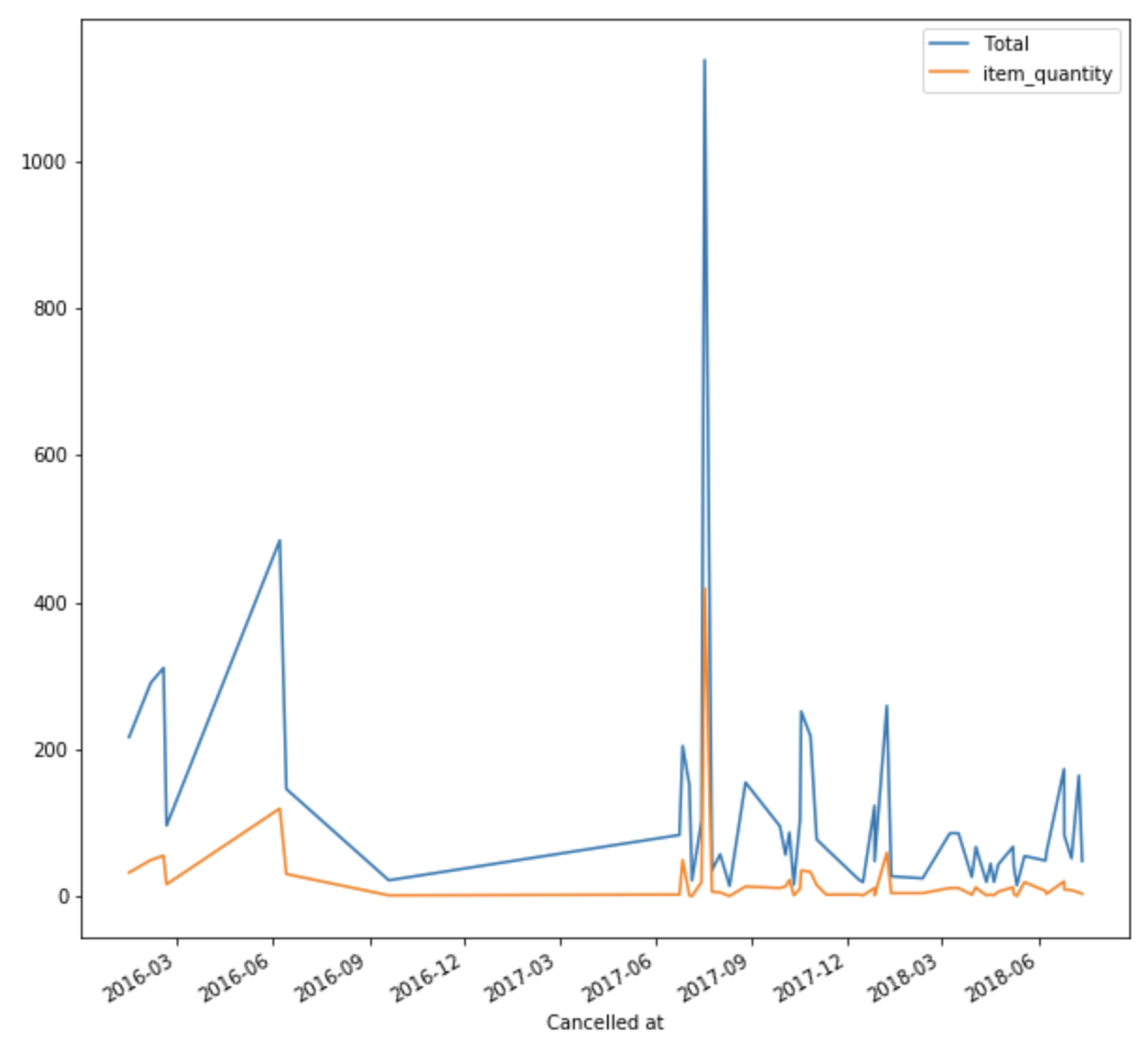
For USJH order unit

1. The peak season is October-November (the time order created, preparation for holiday), when sold most items and earn most money. The sold item quantity shape is basically consistent with gross shape.

The ‘paid at’ is not exactly consistent with ‘created at’ shape might due to refund?

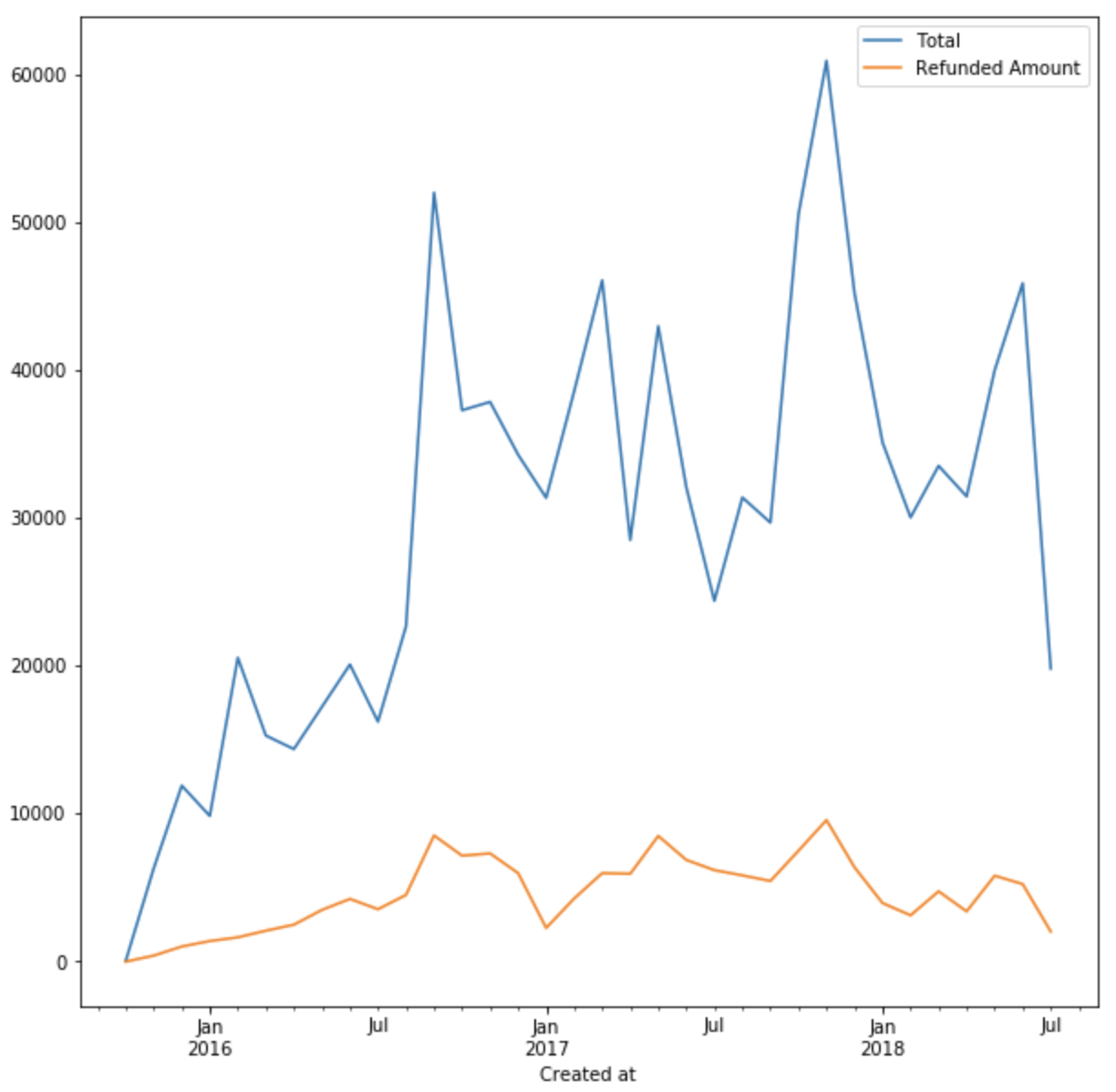
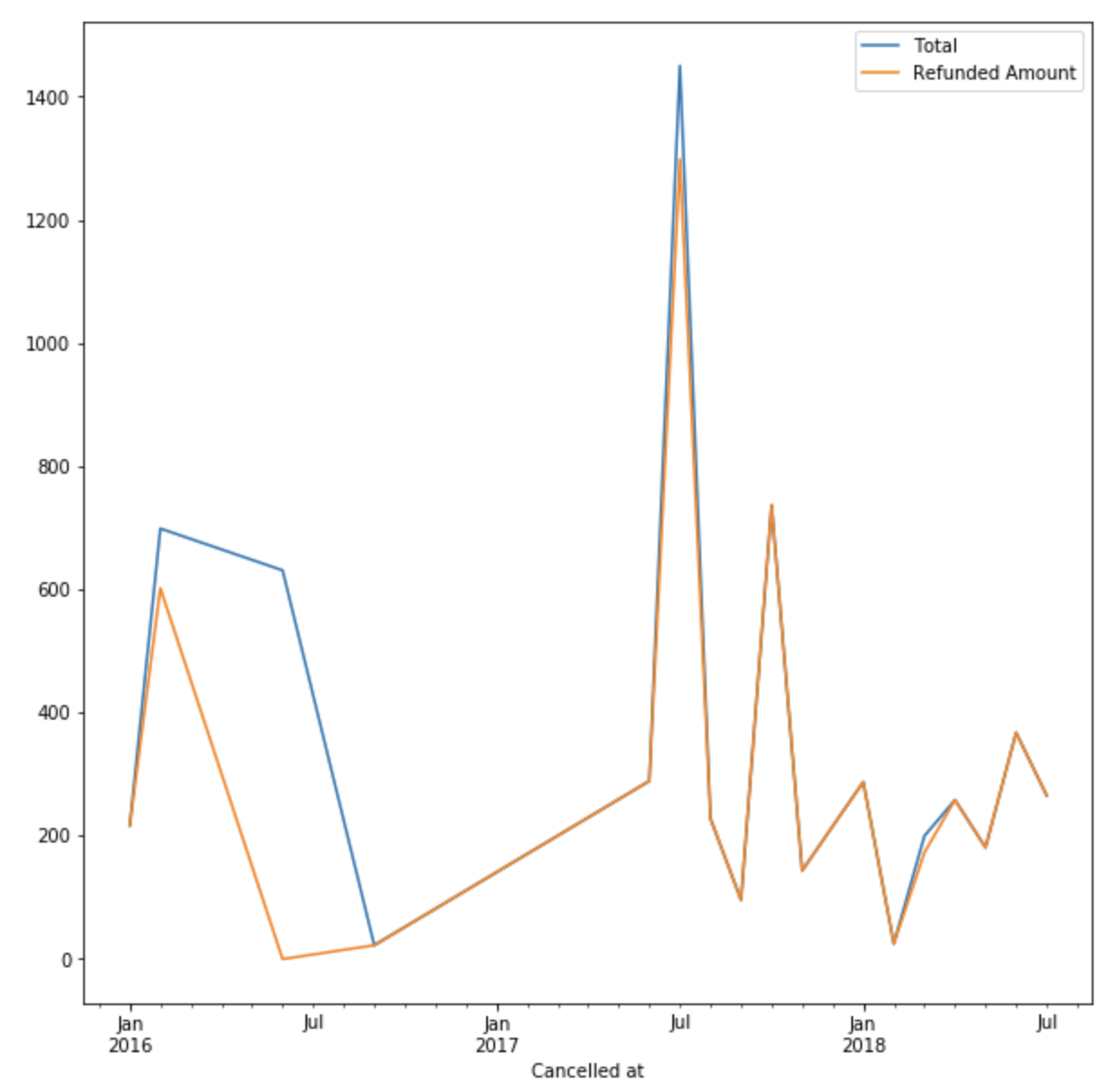
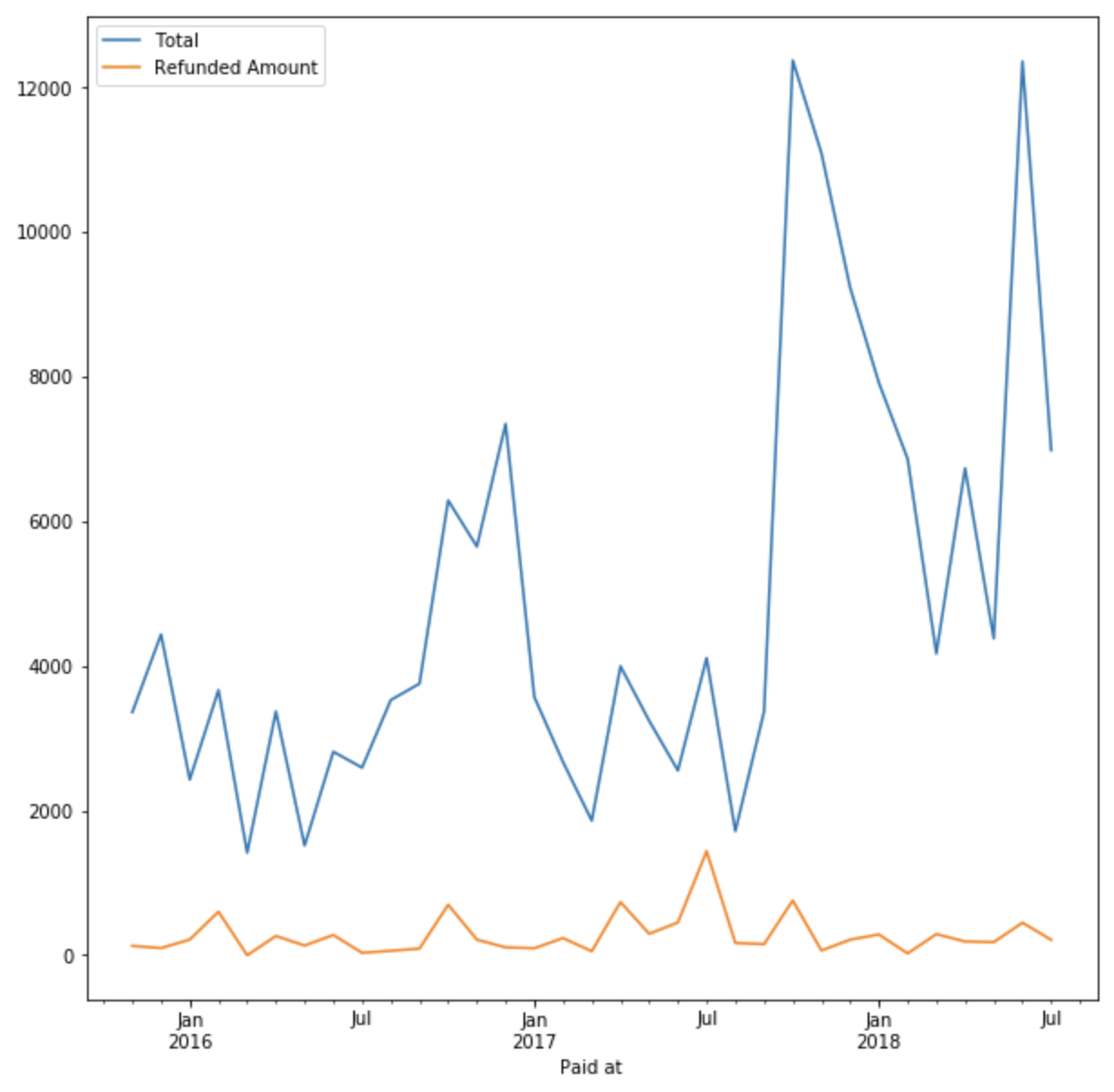
  

1. Something happened someday in 2016-10

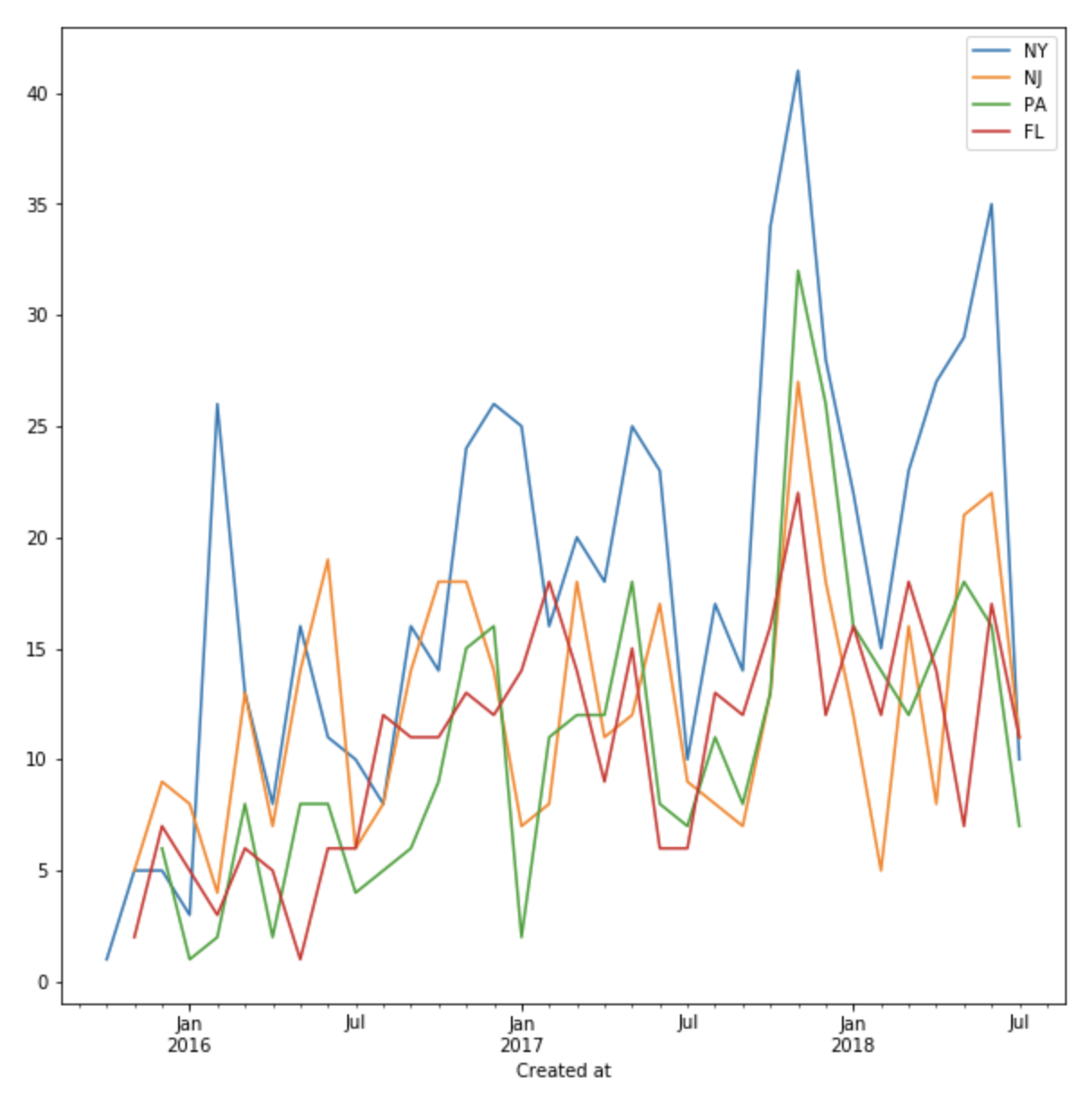
  

1. Basically, more sale more refund too

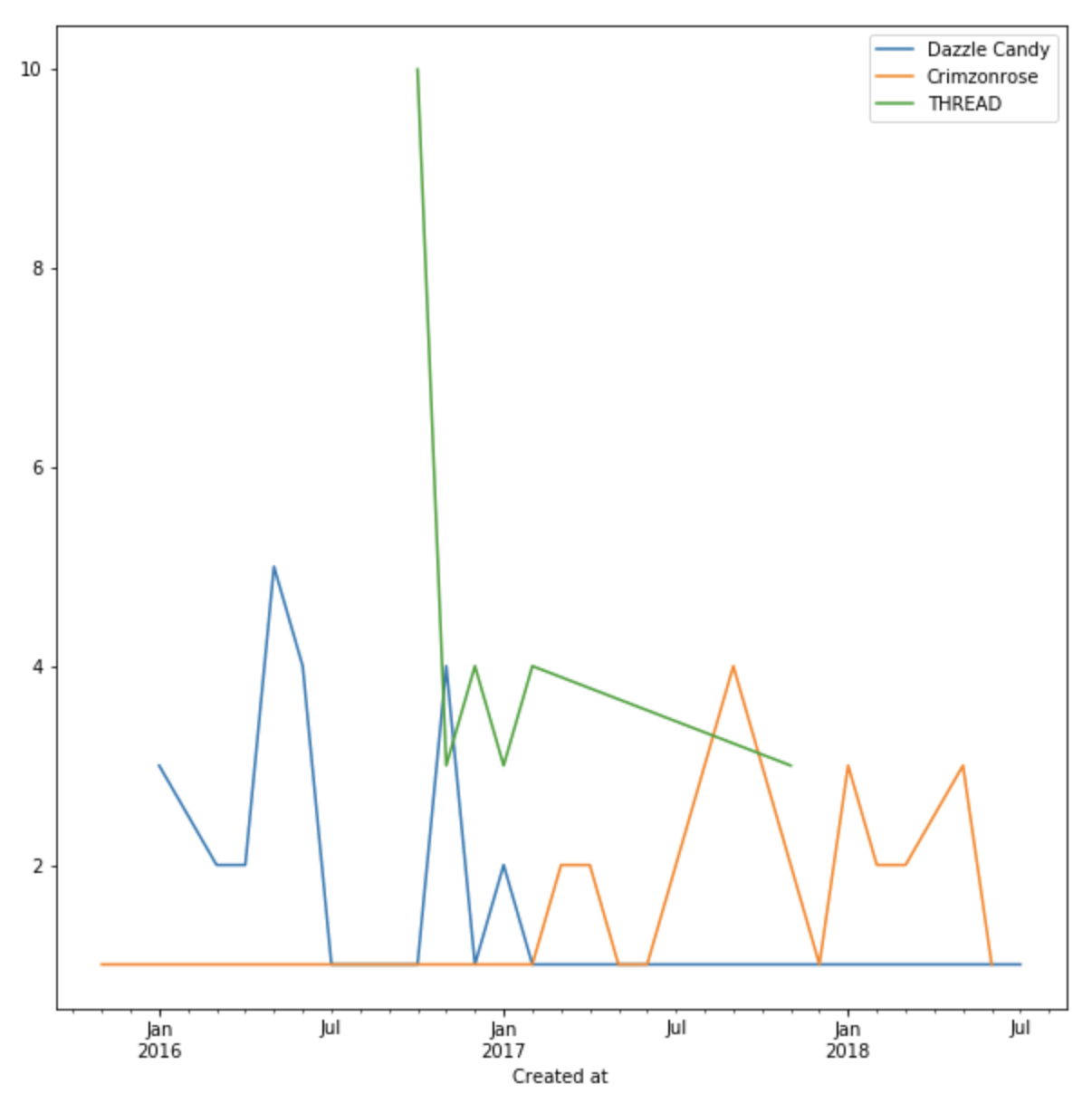
See the difference between partial refund and fully refund in the third plot

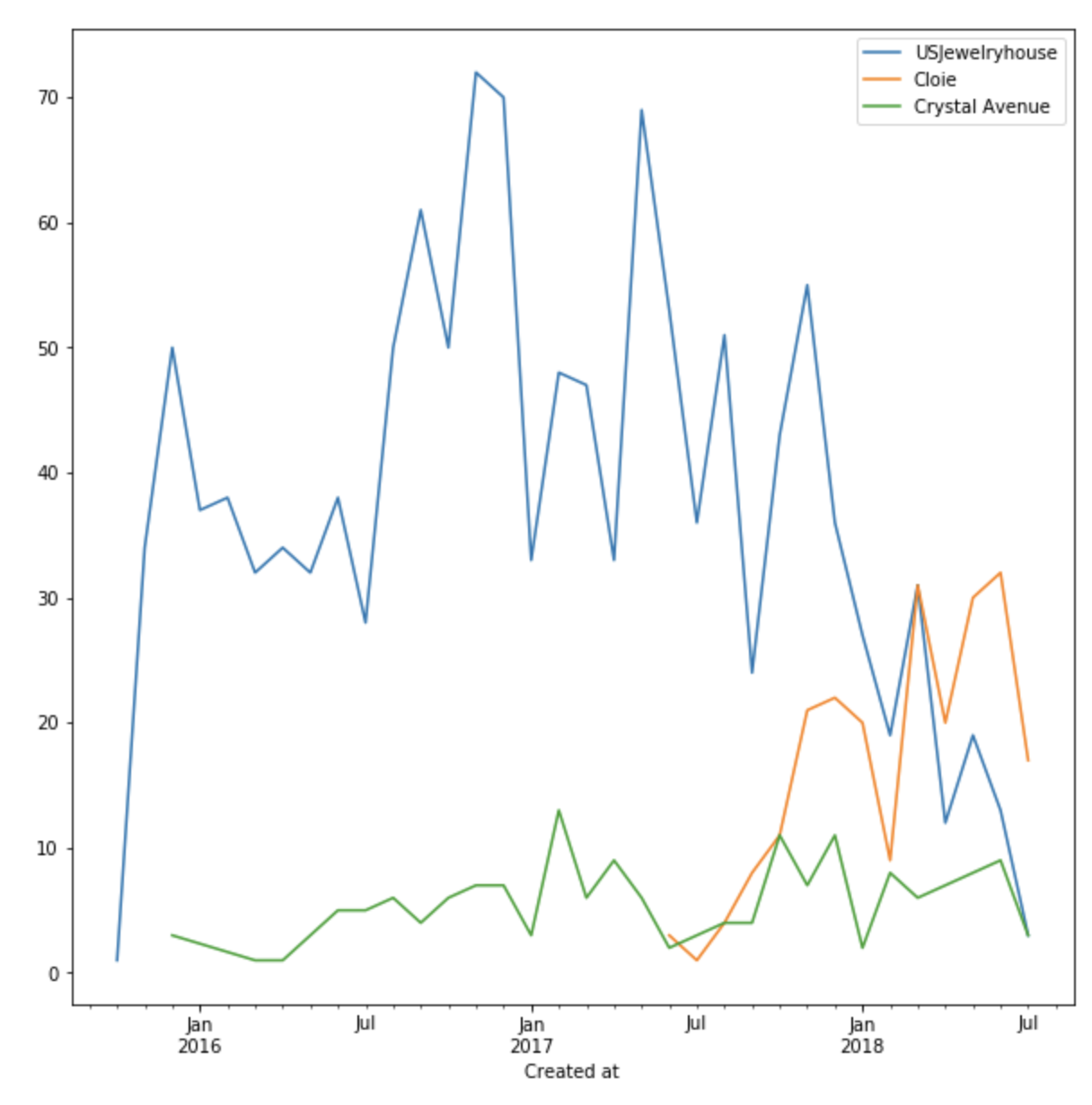
1. Orders to top states, basically consistent



1. Top 3 companies

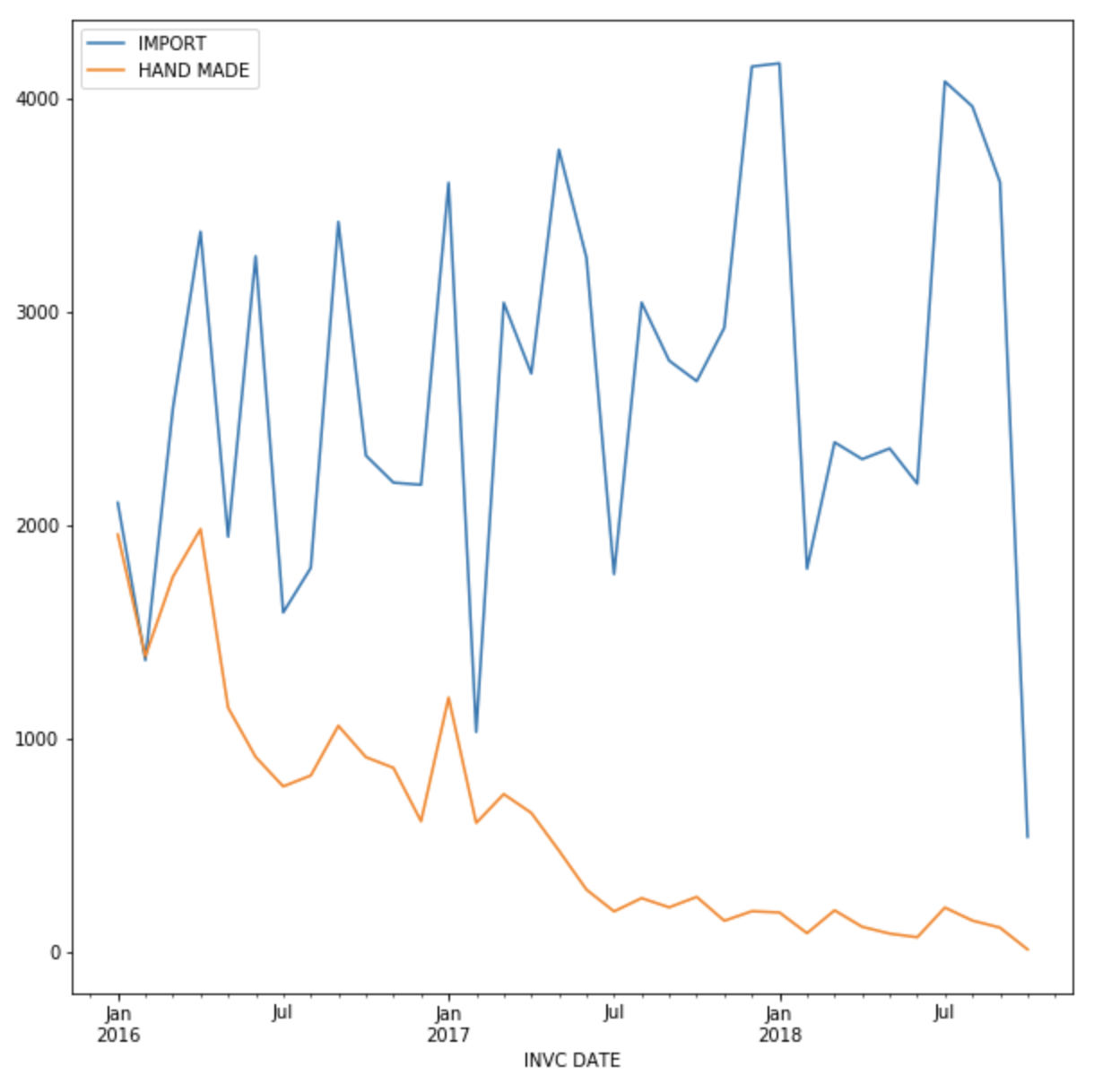


1. Top venders. Are we losing USJH

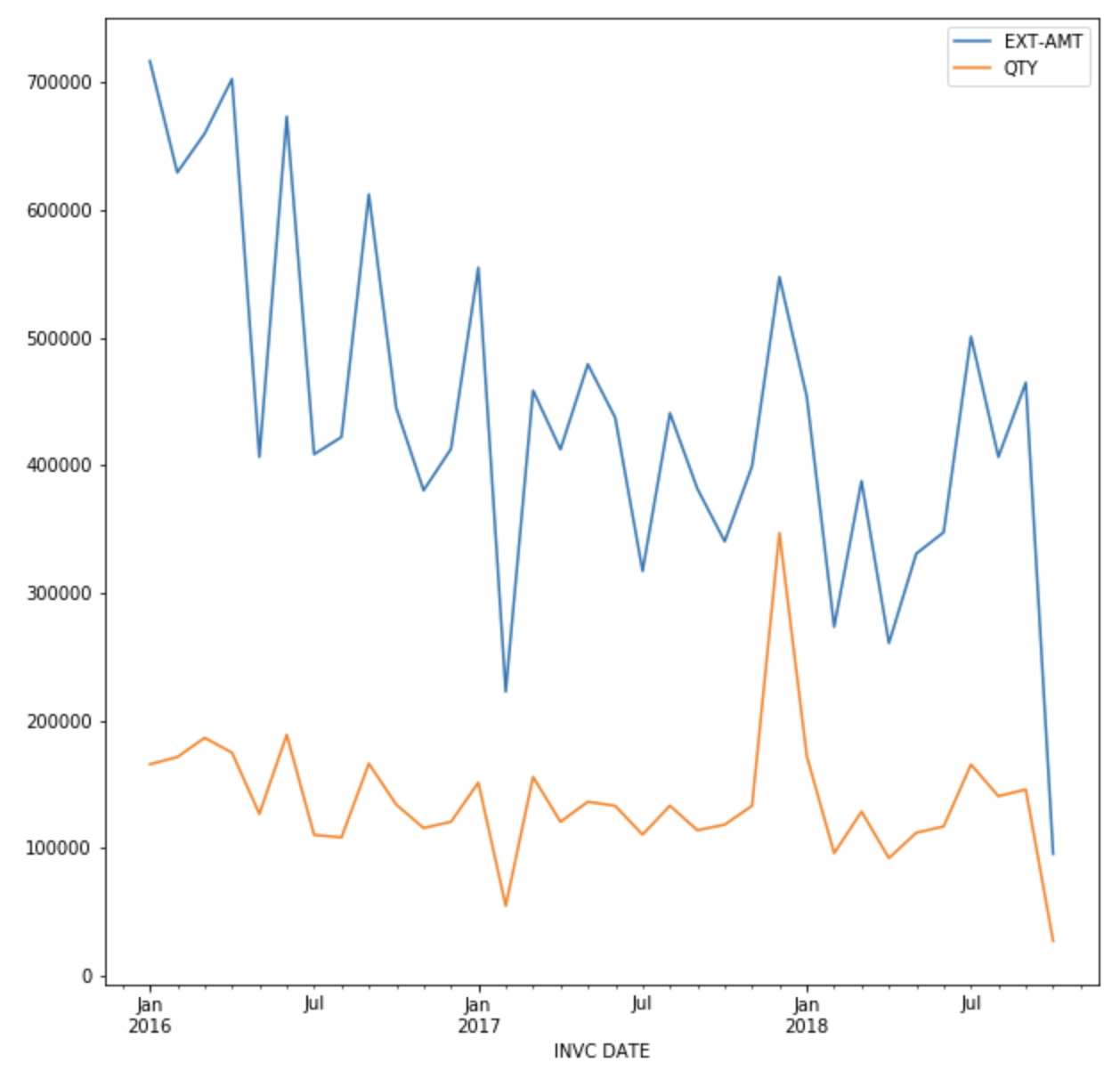


UNO sale

1. Less and less hand made



1. Larger qty than I expected



1. Top states

