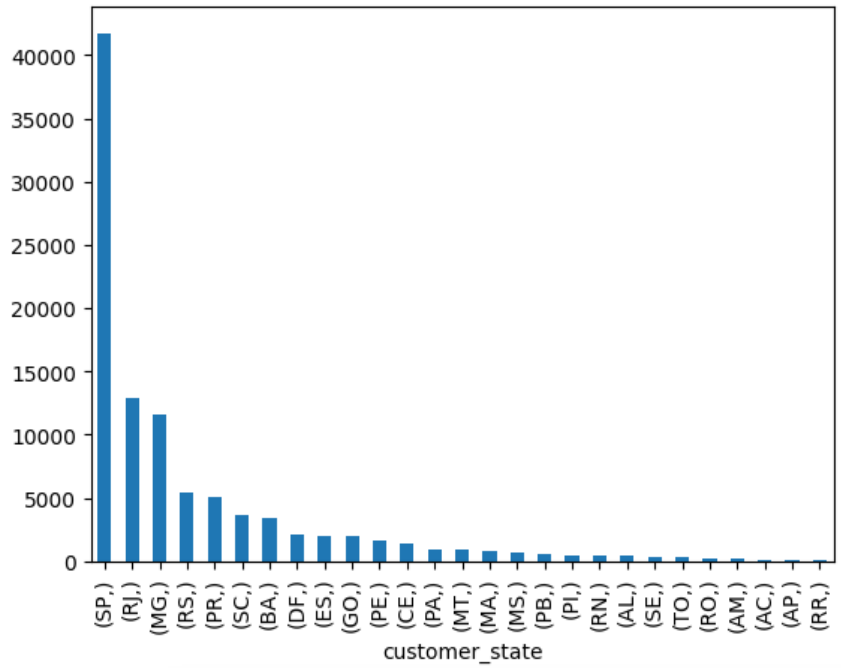
a. User data

The data is categorical data consisting of a list of customers that use the solution that comes up with state, city and zipcode. So we can use this data for knowing the customers distribution of every state, city, and zip code.

Total customer: 99.441 with around 3.345 possible duplicate users because of the redundancy in “customer unique id” and “customer id data. There should be only one id and id always unique.

State analytics:
There are 27 states that use the solution with mean equal to 3683 but the standard deviation is 8270 so that means the customer distribution is not spread evenly through the states.

customer_state	
SP	41746
RJ	12852
MG	11635
RS	5466
PR	5045
SC	3637
BA	3380
DF	2140
ES	2033
GO	2020
PE	1652
CE	1336
PA	975
MT	907
MA	747
MS	715
PB	536
PI	495
RN	485
AL	413
SE	350
TO	280
RO	253
AM	148
AC	81
AP	68
RR	46



count	27.000000
mean	3683.000000
std	8270.437664
min	46.000000
25%	381.500000
50%	907.000000
75%	2760.000000
max	41746.000000

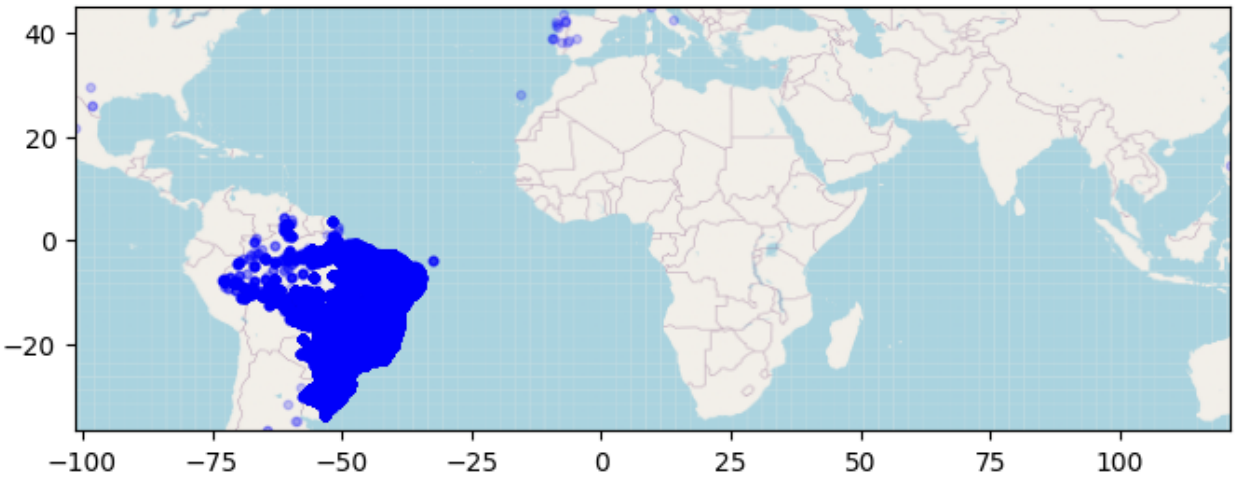
The most customers come from SP state and the least customer is from RR state. If needed we can also analyze our customer distribution by city. For example, the most of our customer comes from sao paulo with 15540 customer, and there are many city that only have 1 customer each such as Rio Espera, Logoao, etc

b. Geolocation data

This data consists of some information on zip code, city, state and geolocation latitude and longitude. Because there is no information on what location is in this data, we can assume this is our customer or office location data because there are some geolocations which are bound to the same zip code, city, and state.

Example of data:

	geolocation_zip_code_prefix	geolocation_lat	geolocation_lng	geolocation_city	geolocation_state
0	1037	-23.545621	-46.639292	sao paulo	SP
1	1046	-23.546081	-46.644820	sao paulo	SP
2	1046	-23.546129	-46.642951	sao paulo	SP
3	1041	-23.544392	-46.639499	sao paulo	SP
4	1035	-23.541578	-46.641607	sao paulo	SP
...	...	...	...	...	...
1000158	99950	-28.068639	-52.010705	tapejara	RS
1000159	99900	-27.877125	-52.224882	getulio vargas	RS
1000160	99950	-28.071855	-52.014716	tapejara	RS
1000161	99980	-28.388932	-51.846871	david canabarro	RS
1000162	99950	-28.070104	-52.018658	tapejara	RS

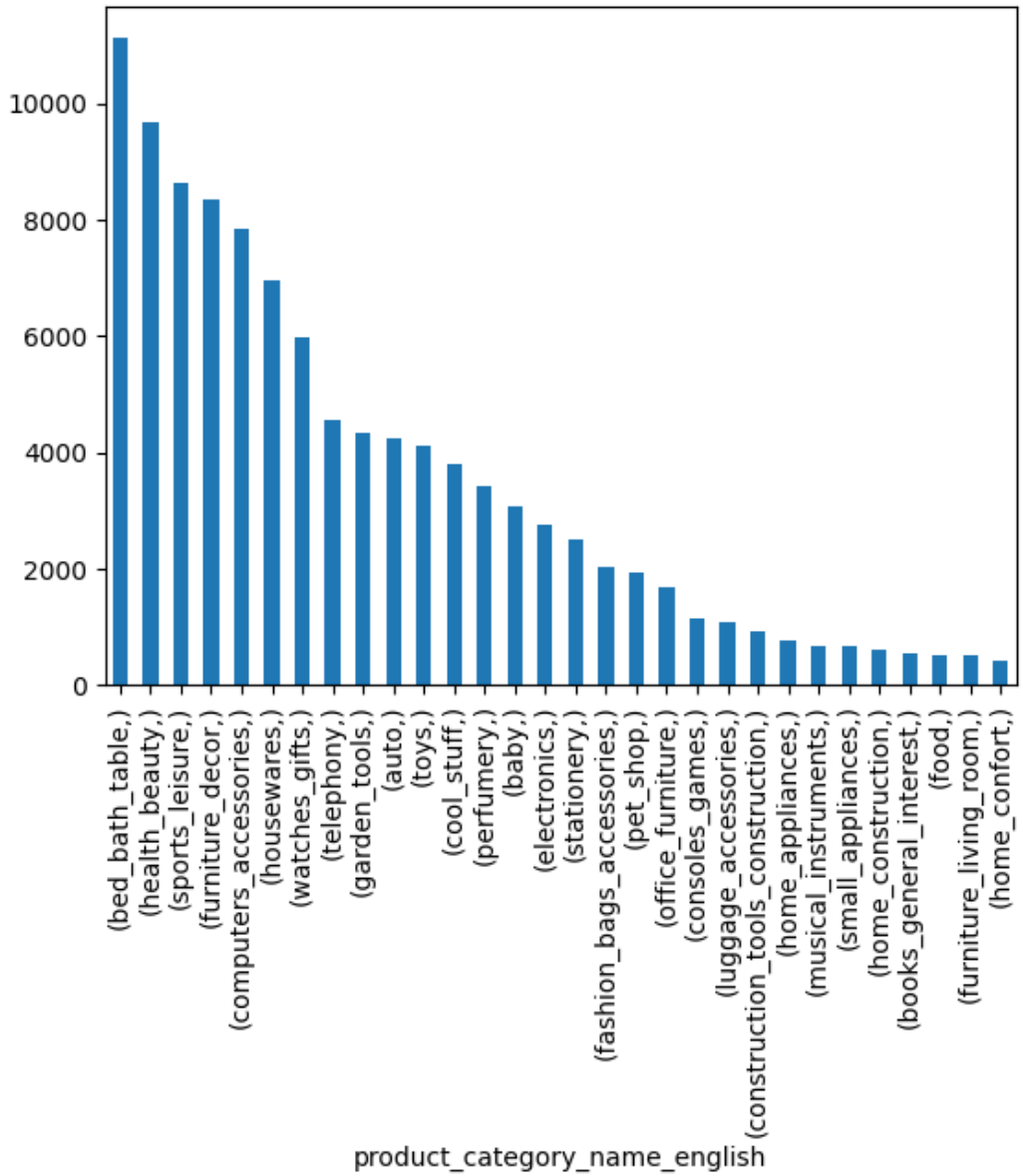


We can see that the majority of the data are in South America with few outliers in around Mexico,USA and around portugal.

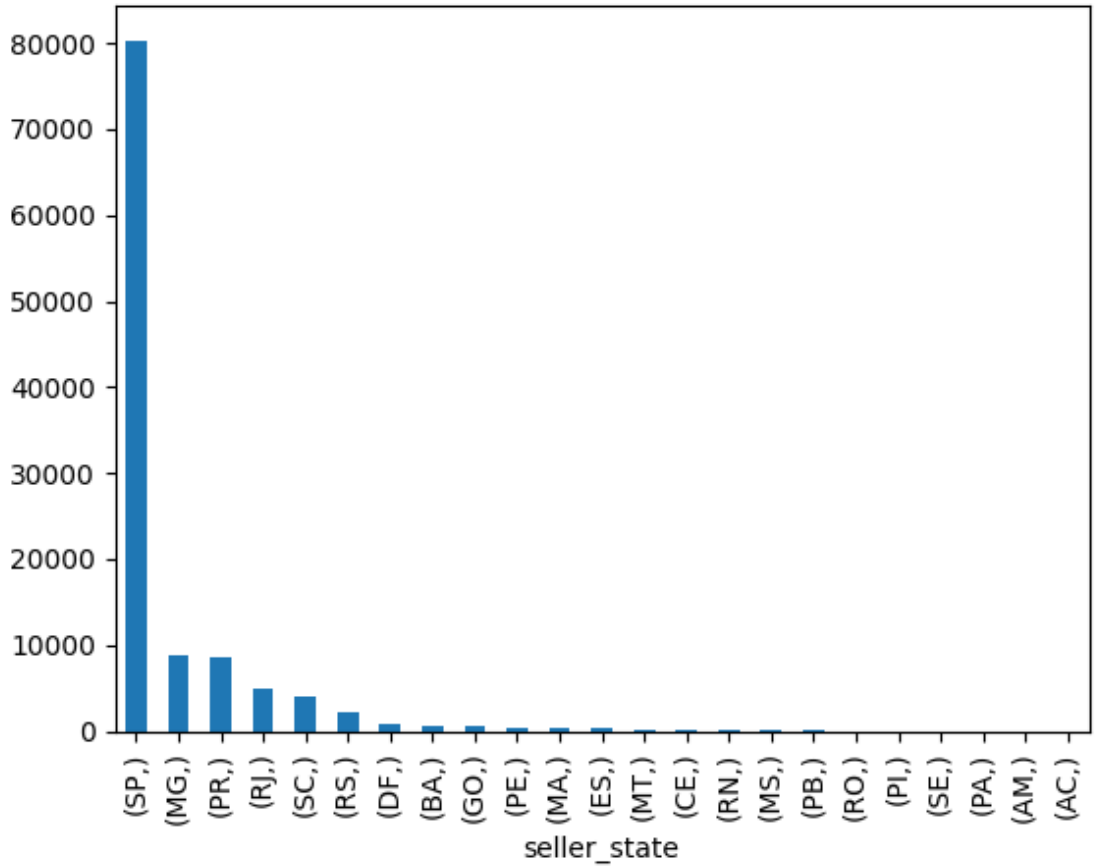
c. Order data

On order data we can see few insights on our company order.

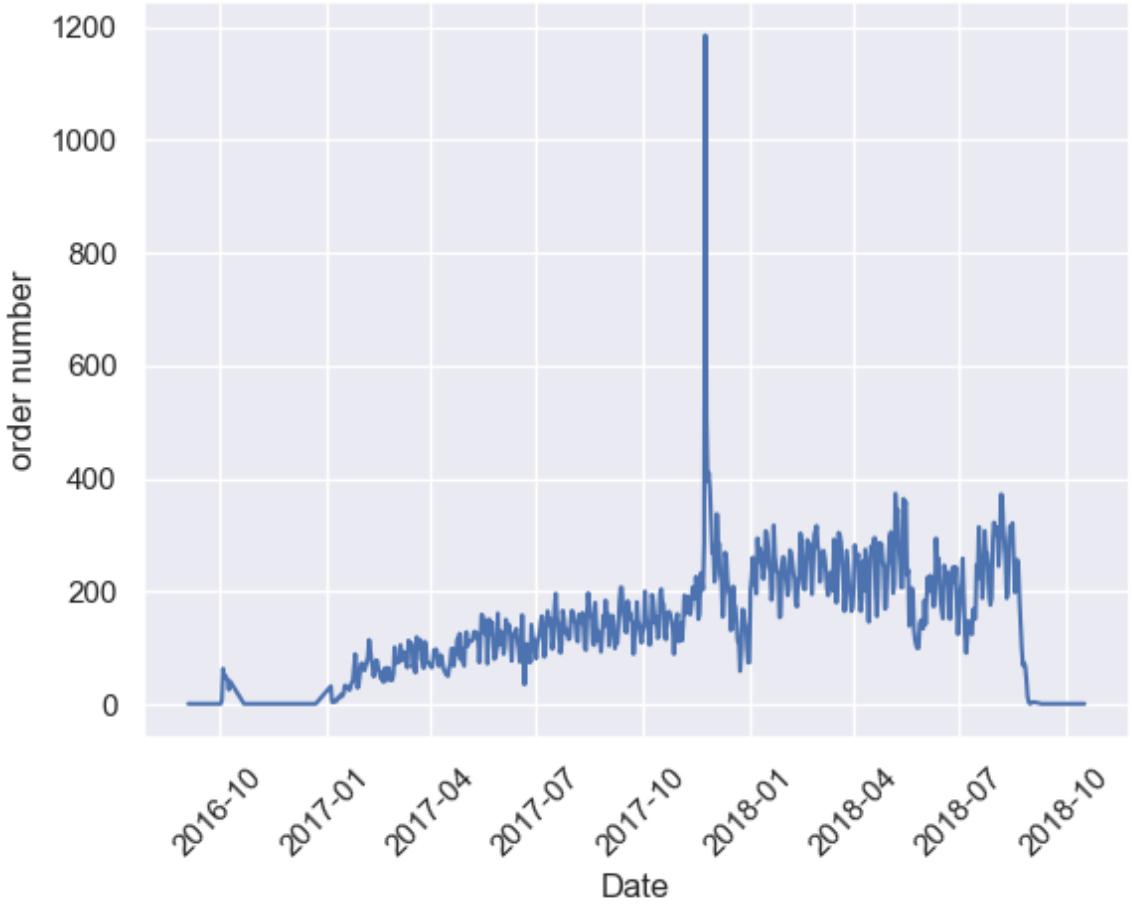
1. First we can merge order data with the product data so we can get all the information on every order. So for product ordered frequency, we can get which product has the most ordered.



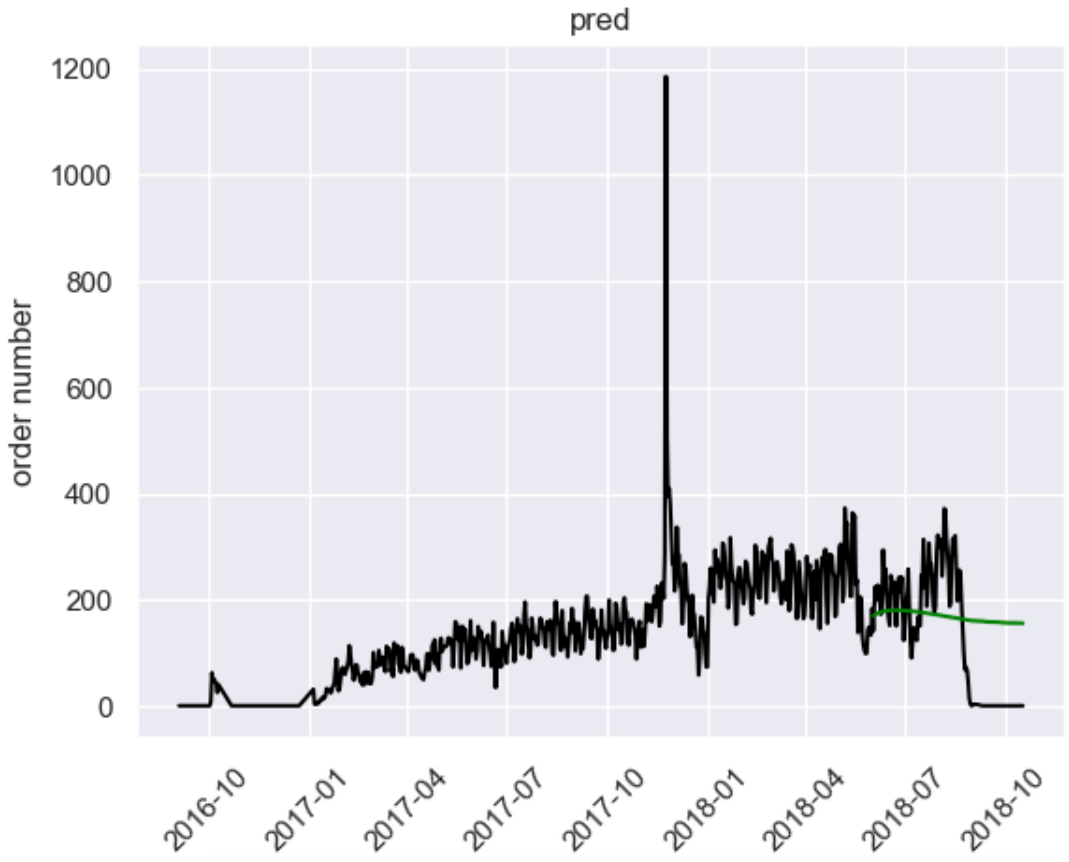
2. We also can use the seller data to get another insight. Where is the state and city that the seller comes from?d



3. We can also get another insight by combining our order data with payments and order review, the first insight that we can see how many order graphs on our data.



We can see in our graph we have peak order in 2017-11-24 with 1185 orders in with lowest order just 1 order. We can also see that our order has slightly increased from 2016 to 2018. With this data we can also forecast our order number for our solution preparation. Here is a simple forecasting model prediction:
I use https://builtin.com/data-science/time-series-forecasting-python as a reference because the distribution of the data is quietly the same. The result:


pred
order number
1200
1000
800
600
400
200
0
2016-10
2017-01
2017-04
2017-07
2017-10
2018-01
2018-04
2018-07
2018-10