Diccionario de tablas

Contamos con 9 archivos CSV diferentes y se creó una tabla a partir de ellos:

* olist\_customers\_dataset.csv



* olist\_geolocation\_dataset.csv



* olist\_order\_items\_dataset.csv



* olist\_order\_payments\_dataset.csv



* olist\_order\_reviews\_dataset.csv



* olist\_orders\_dataset.csv



* olist\_products\_dataset.csv



* olist\_sellers\_dataset.csv



* product\_category\_name\_translation.csv



* olist\_state\_location.csv



Calidad de los datos.

# Cargado de datos

closed\_deals = pd.read\_csv('..\data\olist\_closed\_deals\_dataset.csv', parse\_dates=["won\_date"])

customers = pd.read\_csv('..\data\olist\_customers\_dataset.csv')

geolocation = pd.read\_csv('..\data\olist\_geolocation\_dataset.csv')

marketing\_qualified\_leads = pd.read\_csv('..\data\olist\_marketing\_qualified\_leads\_dataset.csv', parse\_dates=["first\_contact\_date"])

order\_items = pd.read\_csv('..\data\olist\_order\_items\_dataset.csv', parse\_dates=["shipping\_limit\_date"])

order\_payments = pd.read\_csv('..\data\olist\_order\_payments\_dataset.csv')

order\_reviews = pd.read\_csv('..\data\olist\_order\_reviews\_dataset.csv', parse\_dates=["review\_creation\_date", 'review\_answer\_timestamp'])

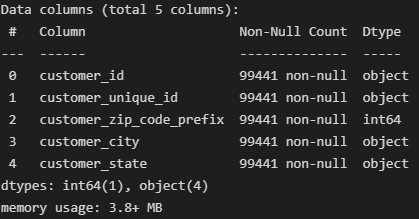
orders = pd.read\_csv('..\data\olist\_orders\_dataset.csv', parse\_dates=["order\_purchase\_timestamp", 'order\_approved\_at', 'order\_delivered\_carrier\_date', 'order\_delivered\_customer\_date', 'order\_estimated\_delivery\_date'])

products = pd.read\_csv('..\data\olist\_products\_dataset.csv')

sellers = pd.read\_csv('..\data\olist\_sellers\_dataset.csv')

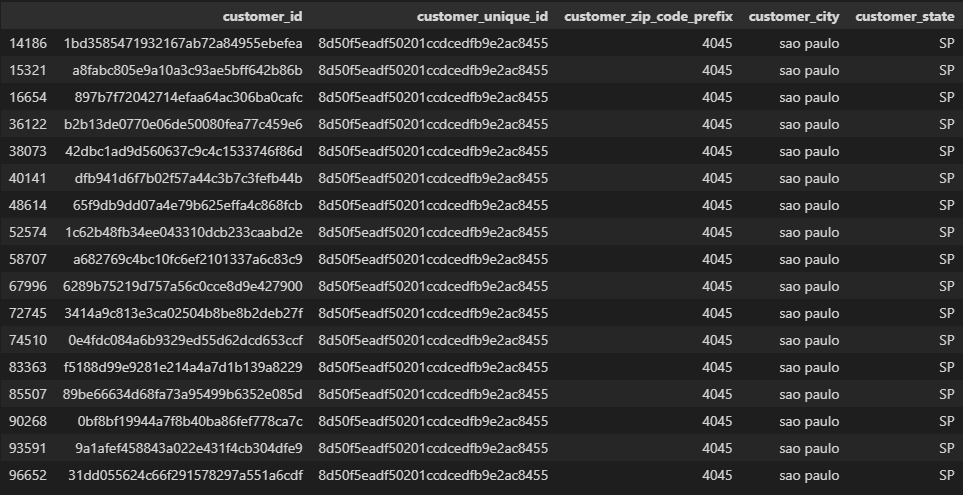
# Customers

customers = pd.read\_csv('data\olist\_customers\_dataset.csv')



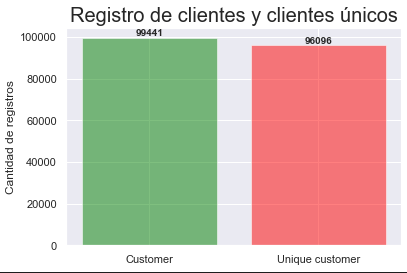
El total de registros de ‘customer\_id’ es 99441

Se observa que existen diferentes ‘customer\_id’ con un mismo ‘customer\_unique\_id’ (razón desconocida)



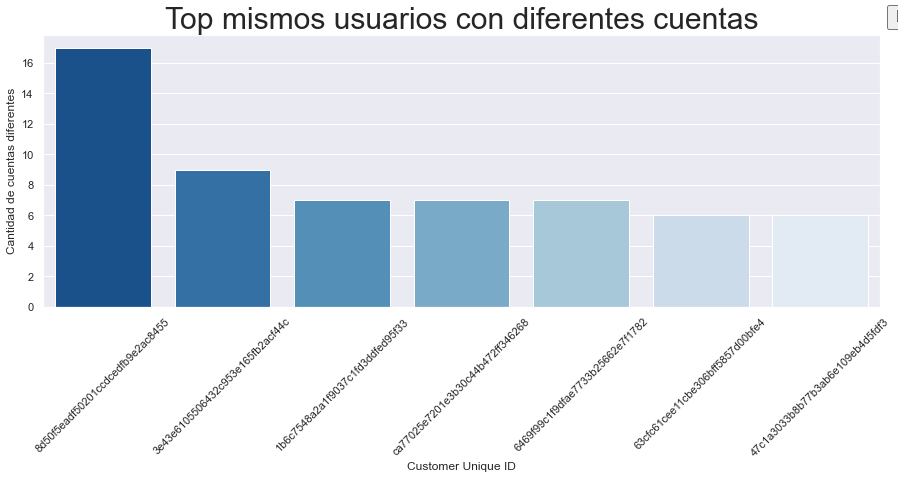
En algunos varía el ‘customer\_zip\_code\_prefix’; sin embargo, en general, se mantienen en el mismo estado.

El total de registros ‘customer\_unique\_id’ es 96096



Diferencia de 3345 que corresponde a que un 3.4% de los clientes tienen más de 1 cuenta

Las cuentas van desde 17 por usuario único hasta 1, como se muestran en la siguiente gráfica



La distribución de los clientes por estado queda de la siguiente manera

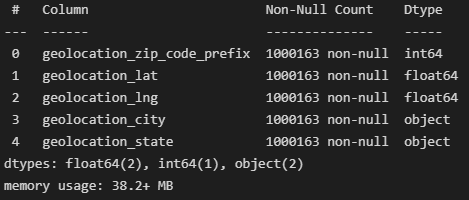




La región más importante de mayor potencial de ventas es el estado SP (Sa~o Paolo)

# Geolocation

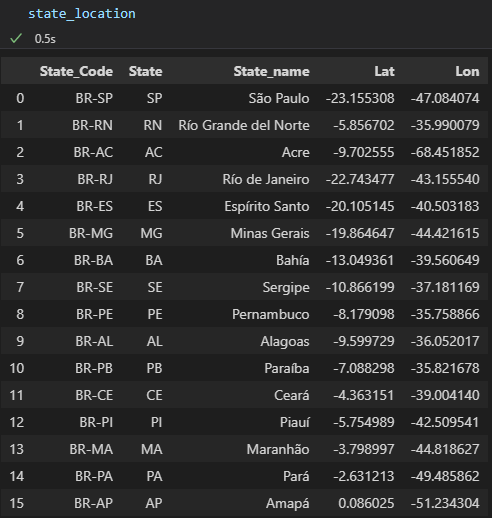
geolocation = pd.read\_csv('data\olist\_geolocation\_dataset.csv')



1000163 datos no nulos.

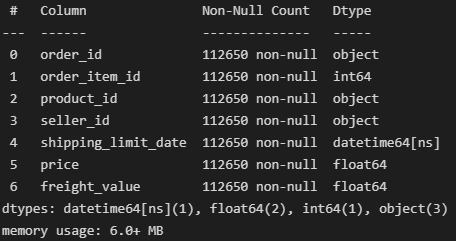
Se nos recomienda en el proyecto trabajar con una finura de estado.

Tomo una tabla de Wikipedia de la distribución geográfica estatal de Brasil y hago ‘join’ para tener un nuevo Dataframe que contenga los nombres de los estados (la tabla original no los contiene), además contiene la información de la latitud y longitud de cada uno utilizando el promedio de las latitudes y longitudes para cada estado (creo también un nuevo archivo csv con dicha información para utilizarlo más adelante) llamado ‘state\_location’.



# Order\_items

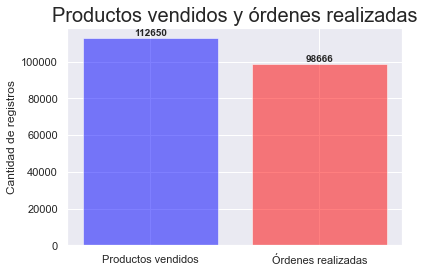
order\_items = pd.read\_csv('data\olist\_order\_items\_dataset.csv', parse\_dates=["shipping\_limit\_date"])



112650 datos no nulos.

De los cuales únicamente 98666 corresponden a pedidos diferentes, debido a que un mismo pedido puede contener uno o más productos.

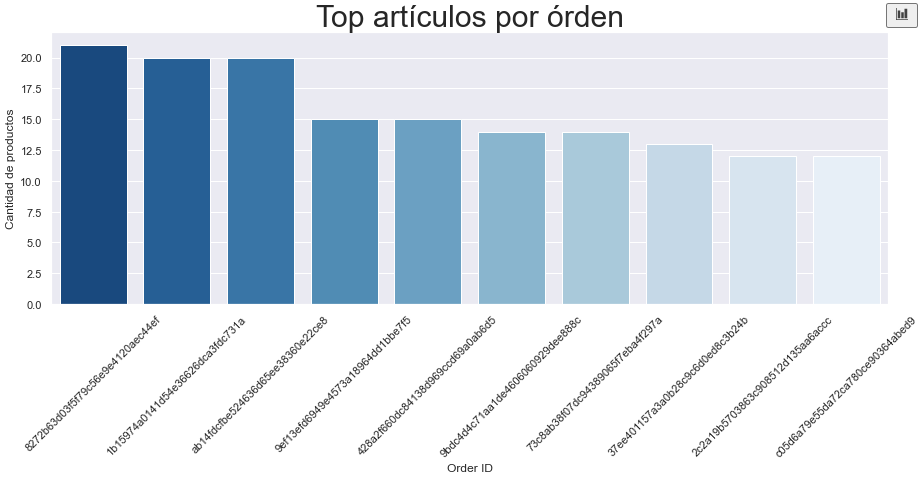
La distribución se encuentra en la siguiente gráfica:



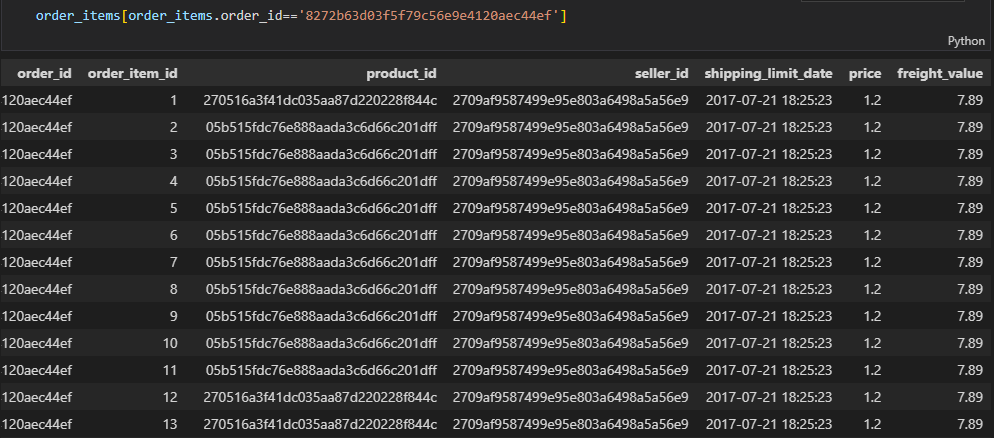
Siendo una diferencia de 13984

Lo que nos dice que el 12.4% de todas las órdenes contienen más de 2 artículos.

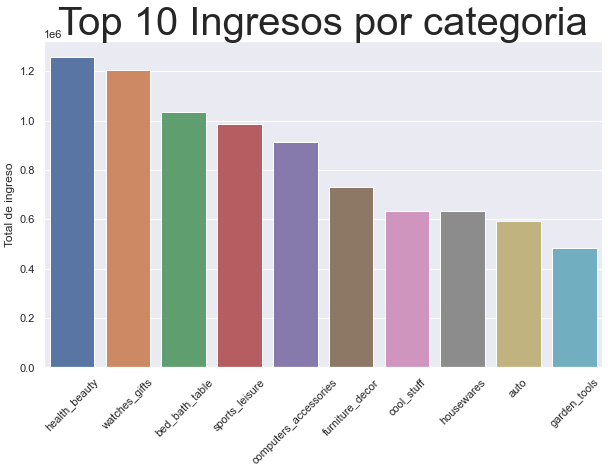
La distribución va de 21 a 1 producto por orden, el siguiente gráfico es el top de ellos



Se puede observar que, además, contamos con la información del precio del artículo y el precio del envío.

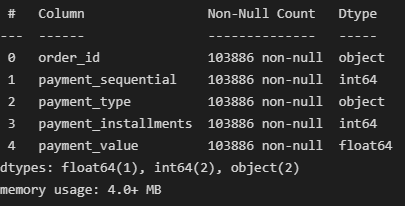


Top 10 ingresos por categoría:



# Order\_payments

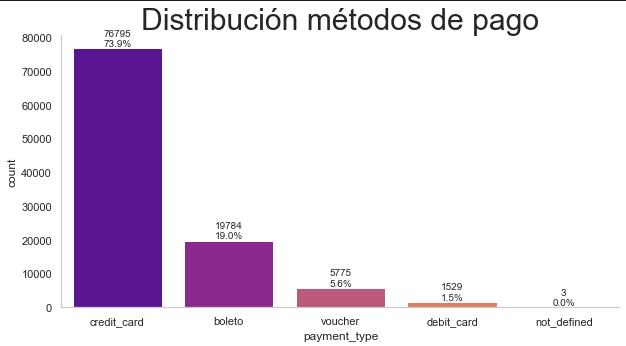
order\_payments = pd.read\_csv('data\olist\_order\_payments\_dataset.csv')



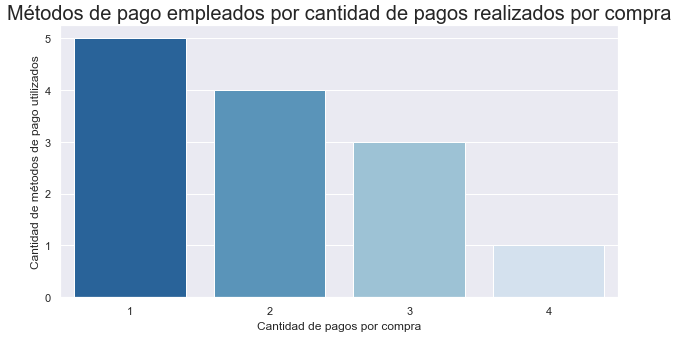
103886 datos no nulos.

Del total de órdenes, tenemos 99440 órdenes distintas realizadas, por lo que, una misma orden fue pagada con más de un método de pago

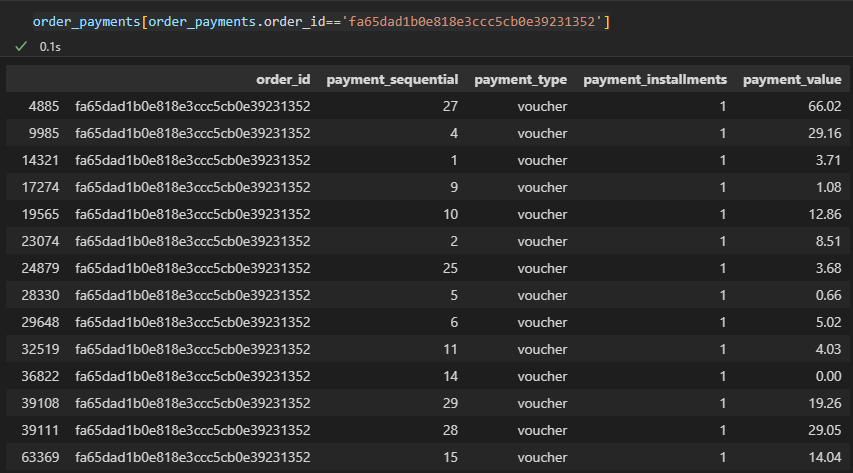
La distribución de los métodos de pago queda de la siguiente manera



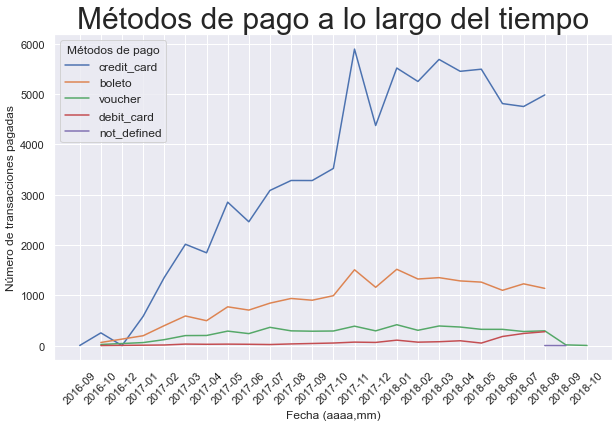
Entre más movimientos se necesitaron para realizar el pago, menos métodos de pago fueron empleados



A partir de 4 operaciones para realizar el pago de la orden se observa que sólo se utiliza un método de pago, el cual es el siguiente: ‘voucher’

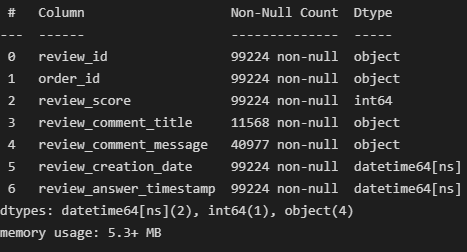


Históricamente la manera en cómo se han utilizado los métodos de pago se muestra en la siguiente gráfica:



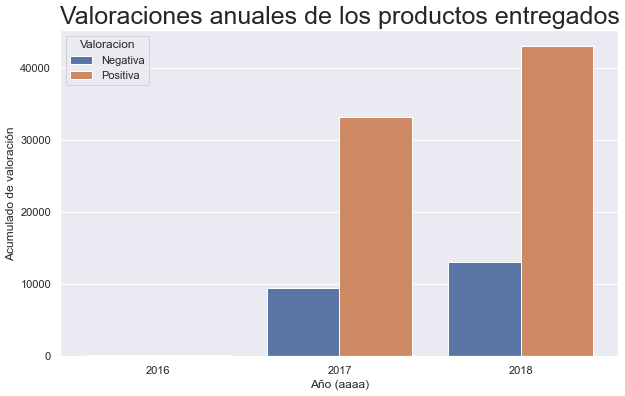
# Order\_reviews

order\_reviews = pd.read\_csv('data\olist\_order\_reviews\_dataset.csv', parse\_dates=["review\_creation\_date", 'review\_answer\_timestamp'])



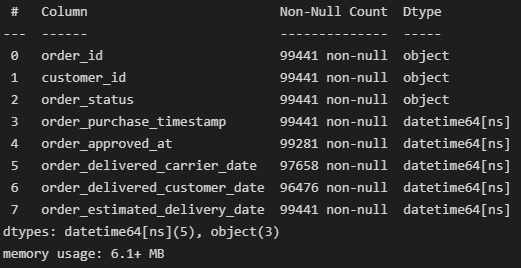
Se desconoce el motive de los datos faltantes

La comparativa entre calificaciones negativas y positivas se muestra a continuación

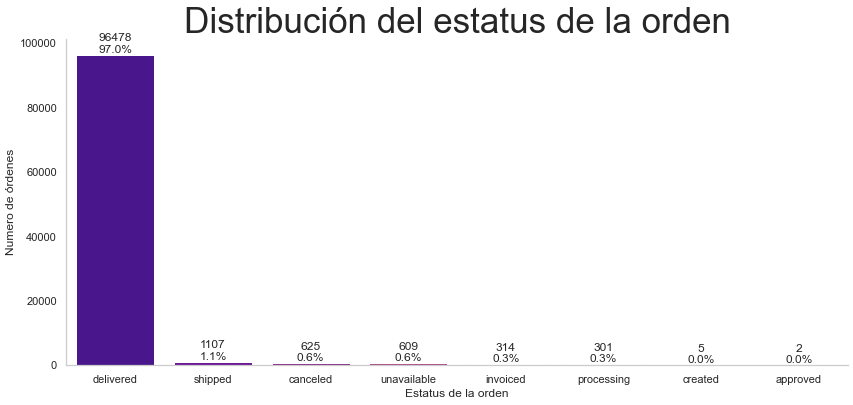


# Orders

orders = pd.read\_csv('data\olist\_orders\_dataset.csv', parse\_dates=["order\_purchase\_timestamp", 'order\_approved\_at', 'order\_delivered\_carrier\_date', 'order\_delivered\_customer\_date', 'order\_estimated\_delivery\_date'])

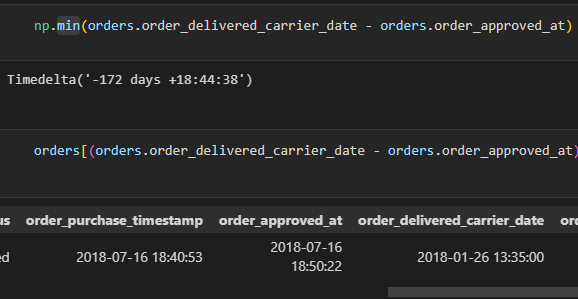


La distribución del estatus de las órdenes queda de la siguiente manera:



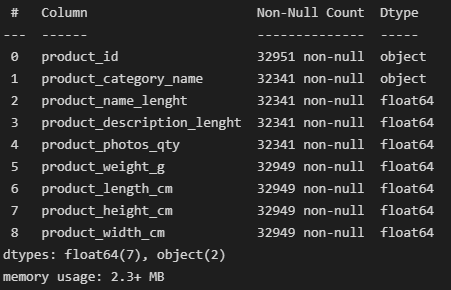
Nos queda que sólo el 3% de las ventas realizadas no fueron entregadas

En el proceso de envío, las fechas aumentan desde que se realizo la compra hasta el estimado de entrega del producto; sin embargo, existen fechas de procesos anteriores superiores al proceso siguiente, por ejemplo:

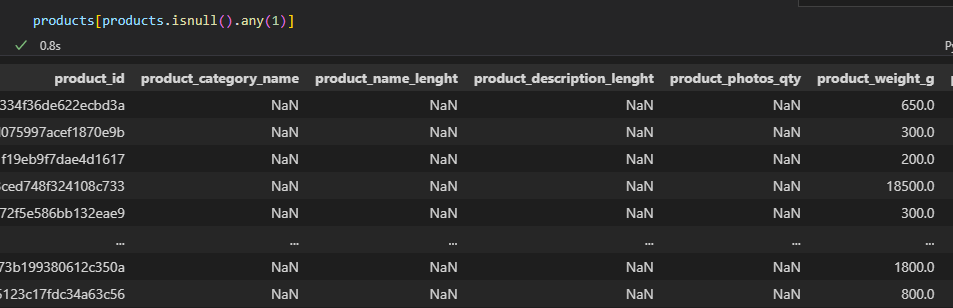


# Products

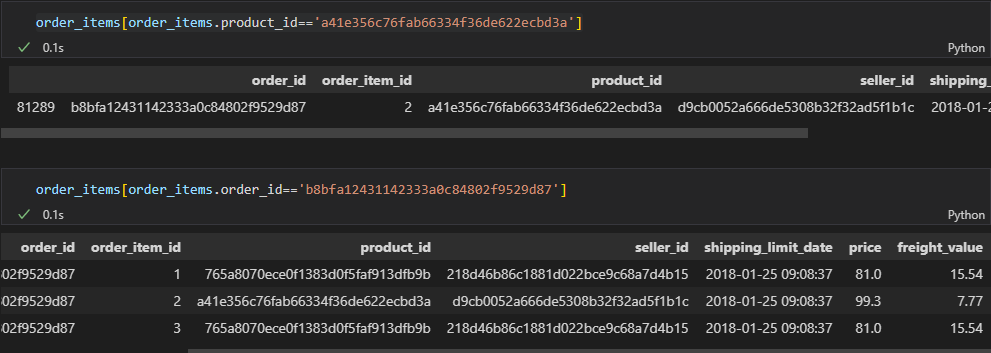
products = pd.read\_csv('data\olist\_products\_dataset.csv')



Existen productos que no tienen ningún tipo de descripción más que los valores de sus dimensiones

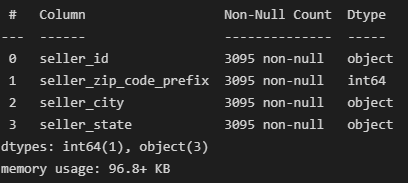


Sin embargo, sí han sido comprados



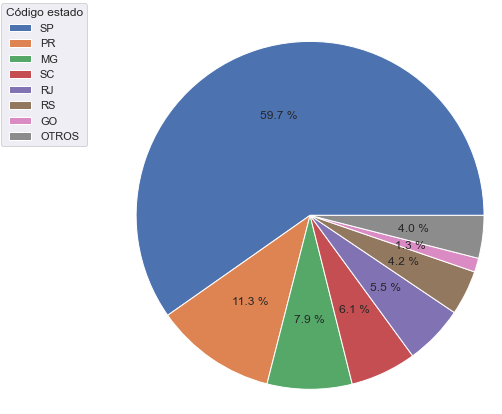
# Sellers

sellers = pd.read\_csv('data\olist\_sellers\_dataset.csv')



Distribución de los vendedores por estado





La mayor cantidad de vendedores se encuentra en SP (Sao Paolo)

Los Top 10 vendedores por categoría son

