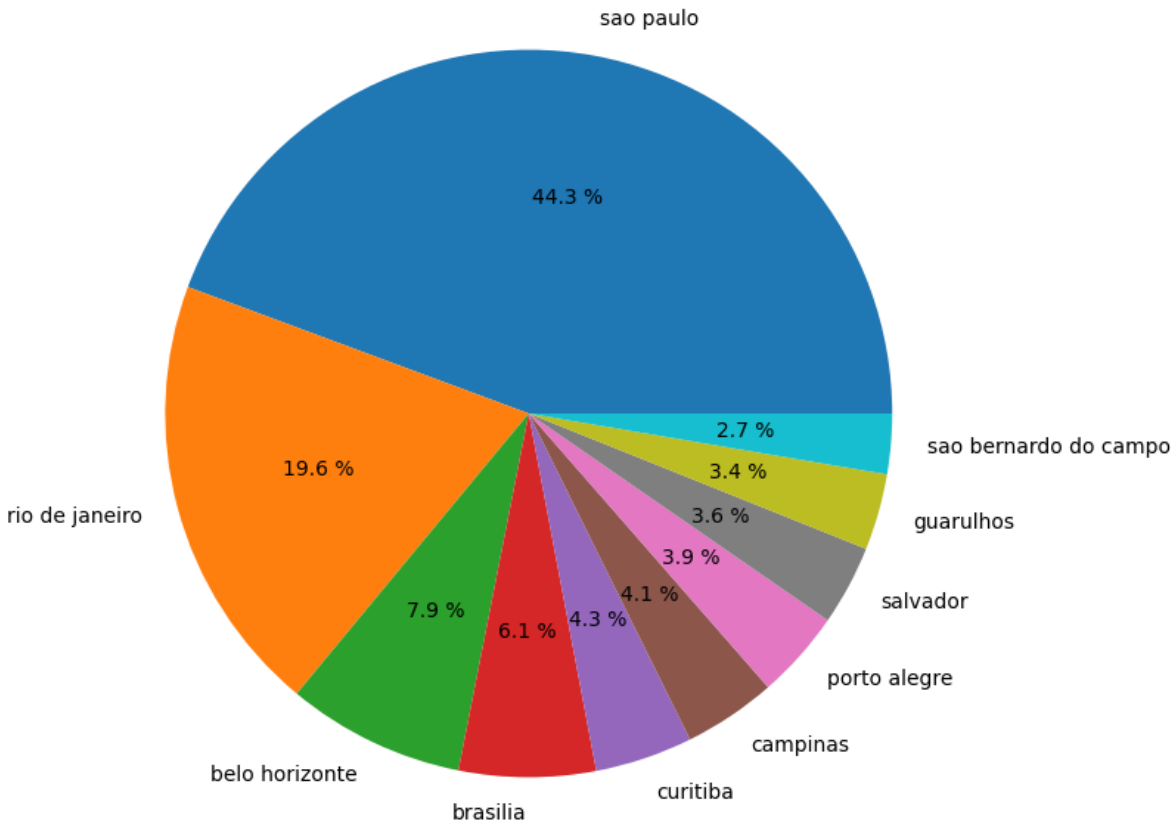
In [1]:

```
import pandas as pd
import matplotlib.pyplot as plt
```

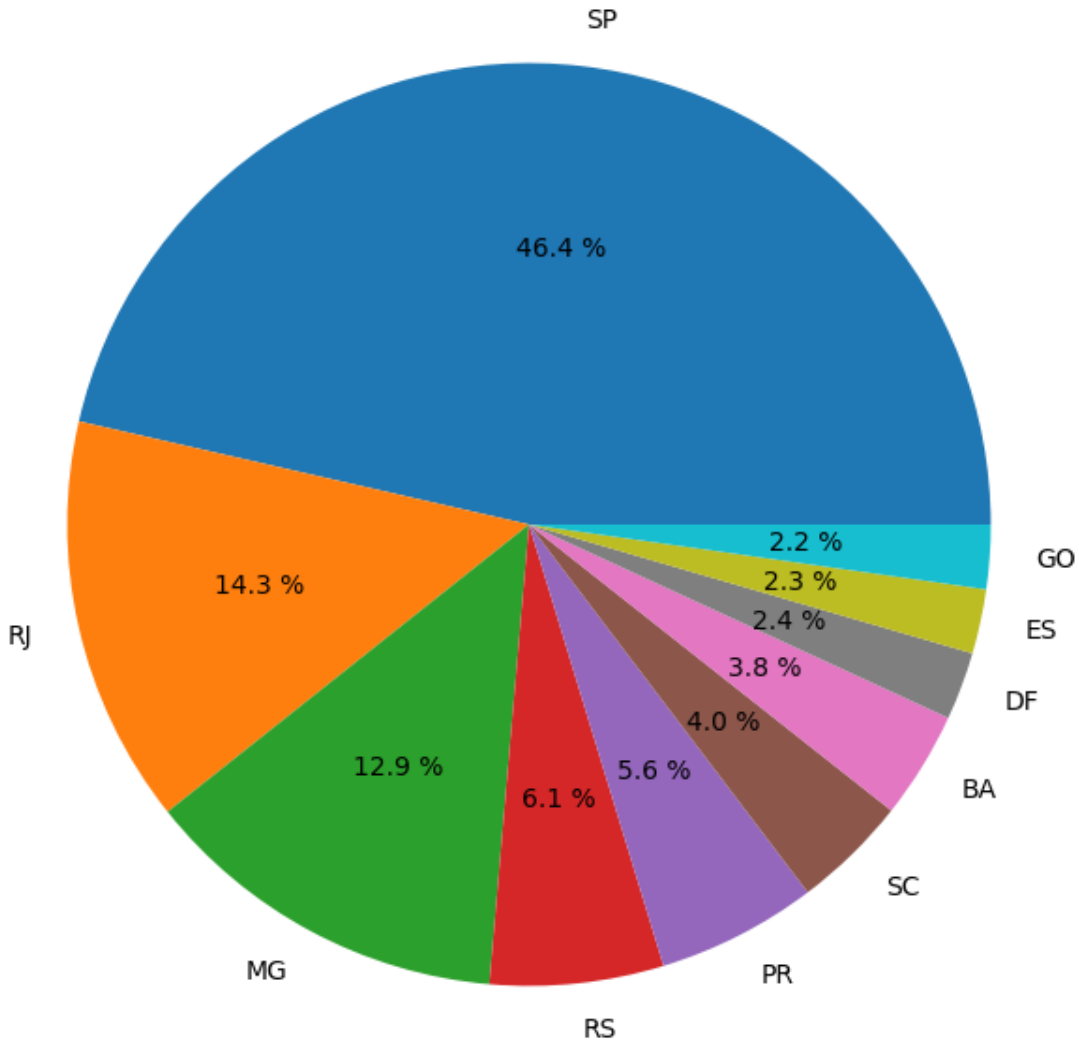
In [4]:

```
#1、customer_city饼图
olist_dataset=pd.read_csv('./data/olist_customers_dataset.csv')
city_type=olist_dataset['customer_city'].value_counts()[:10]
plt.figure(figsize=(12,8),dpi=100)
plt.pie(x=city_type,labels=city_type.index,autopct='%3.1f %%')
plt.show()
```



In [7]:

```
#2、customer_state饼图
olist_dataset
state_type=olist_dataset['customer_state'].value_counts()[:10]
plt.figure(figsize=(12,8),dpi=100)
plt.pie(x=state_type,labels=state_type.index,autopct='%3.1f %%')
plt.show()
```



In [16]:

```
#3、payment和order合并data，price和payment_value的差值
data_orders = pd.read_csv('./data/olist_order_items_dataset.csv')
data_products = pd.read_csv('./data/olist_order_payments_dataset.csv')
p = pd.merge(data_orders, data_products, on = "order_id")
p['payment_value-price']=p['payment_value']-p['price']
p=p.loc[(p['payment_value-price']>=0)&(p['payment_value-price']<=10000),:]
p
```

Out[16]:

	seller_id	shipping_limit_date	price	freight_value	payment_sequential	payr
	6dade18ac8b2bce089ec2a041202	2017-09-19 09:45:35	58.90	13.29	1	c
	dc04e1b6c2c614352b383efe2d36	2017-05-03 11:05:13	239.90	19.93	1	c
	032eddd242adc84c38acab88f23d	2018-01-18 14:48:30	199.00	17.87	1	c
	1d34a5052409006425275ba1c2b4	2018-08-15 10:10:18	12.99	12.79	1	c
	10393f3a51e74553ab94004ba5c87	2017-02-13 13:57:51	199.90	18.14	1	c
	...	...	...	...	...	
	c237ba3788b23da09c0f1f3a3288c	2018-05-02 04:11:01	299.99	43.41	1	
	8ab652836d21de61fb8314b69182	2018-07-20 04:31:48	350.00	36.53	1	
	dc648177fdbbbb35635a37472c53	2017-10-30 17:14:25	99.90	16.95	1	c
	4a2a3ea8e01938cabda2a3e5cc79	2017-08-21 00:04:32	55.99	8.72	1	c
	cf836d21b2fb1de37564105216cc1	2018-06-12 17:10:13	43.00	12.79	1	c

In [20]:

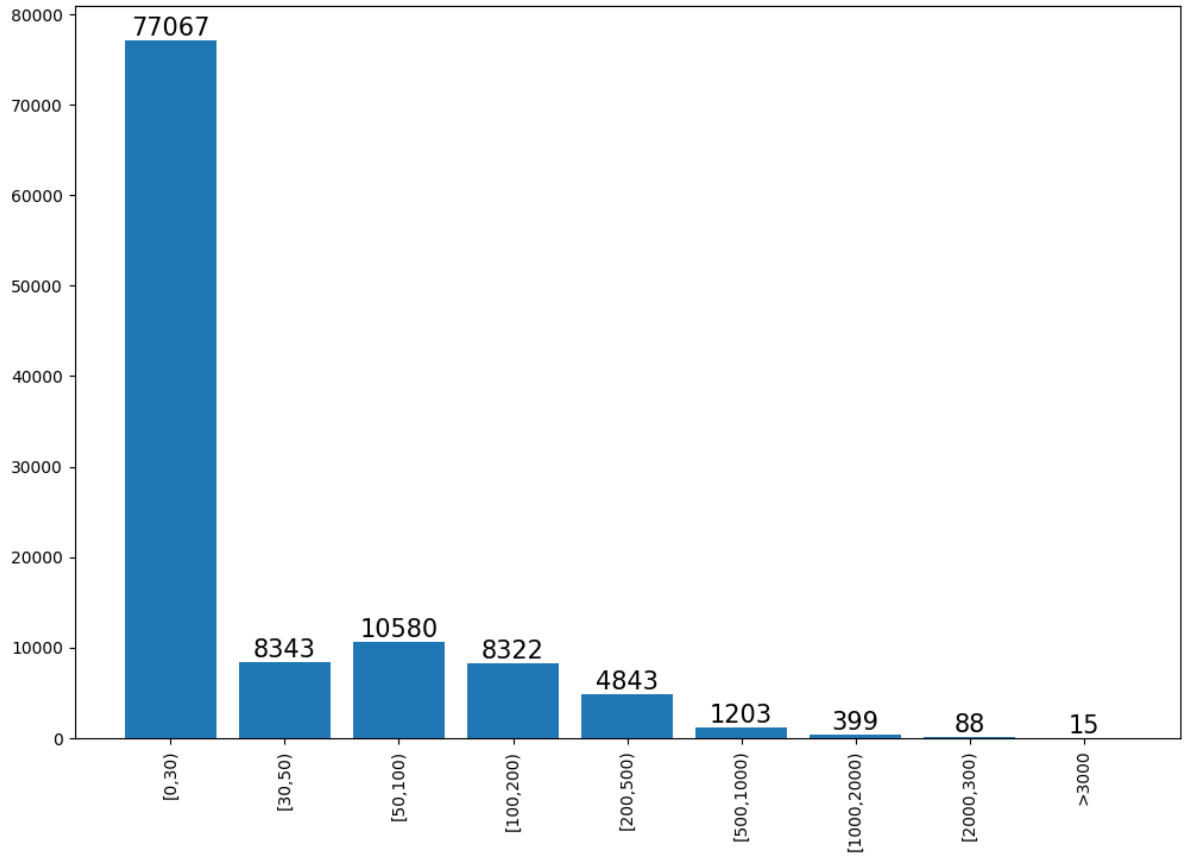
```
bins = [0,30,50,100,200,500,1000,2000,3000,15000]
labels = ['[0,30)','[30,50)','[50,100)','[100,200)',
       '[200,500)','[500,1000)','[1000,2000)','[2000,300)','>3000']
p['price_bins'] = pd.cut(p['payment_value-price'], bins, labels=labels)
bbb=p['price_bins'].value_counts()
ccc=bbb.sort_index()
ccc
```

Out[20]:

```
[0,30)          77067
[30,50)          8343
[50,100)        10580
[100,200)        8322
[200,500)        4843
[500,1000)       1203
[1000,2000)       399
[2000,300)         88
>3000              15
Name: price_bins, dtype: int64
```

In [23]:

```
def barNum(rects):
    for rect in rects:  #rects 是三根柱子的集合
        height = rect.get_height()
        plt.text(rect.get_x() + rect.get_width() / 2, height, str(height), size=15,
plt.figure(figsize=(12,8),dpi=100)
rects=plt.bar(ccc.index,ccc)
plt.xticks(rotation=90)
barNum(rects)
plt.show()
```



In [24]:

```
#4、每年每月每日分布
data_orders
```

Out[24]:

	order_id	order_item_id	product_id
0	00010242fe8c5a6d1ba2dd792cb16214	1	4244733e06e7ecb4970a6e2683c13e61
1	00018f77f2f0320c557190d7a144bdd3	1	e5f2d52b802189ee658865ca93d83a8f
2	000229ec398224ef6ca0657da4fc703e	1	c777355d18b72b67abbeef9df44fd0fd
3	00024acbcdf0a6daa1e931b038114c75	1	7634da152a4610f1595efa32f14722fc
4	00042b26cf59d7ce69dfabb4e55b4fd9	1	ac6c3623068f30de03045865e4e10089
...	...	...	...
112645	fffc94f6ce00a00581880bf54a75a037	1	4aa6014eceb682077f9dc4bffebc05b0
112646	fffcd46ef2263f404302a634eb57f7eb	1	32e07fd915822b0765e448c4dd74c828
112647	fffce4705a9662cd70adb13d4a31832d	1	72a30483855e2eafc67aee5dc2560482
112648	fffe18544ffabc95dfada21779c9644f	1	9c422a519119dcad7575db5af1ba540e
112649	fffe41c64501cc87c801fd61db3f6244	1	350688d9dc1e75ff97be326363655e01

112650 rows × 7 columns

In [35]:

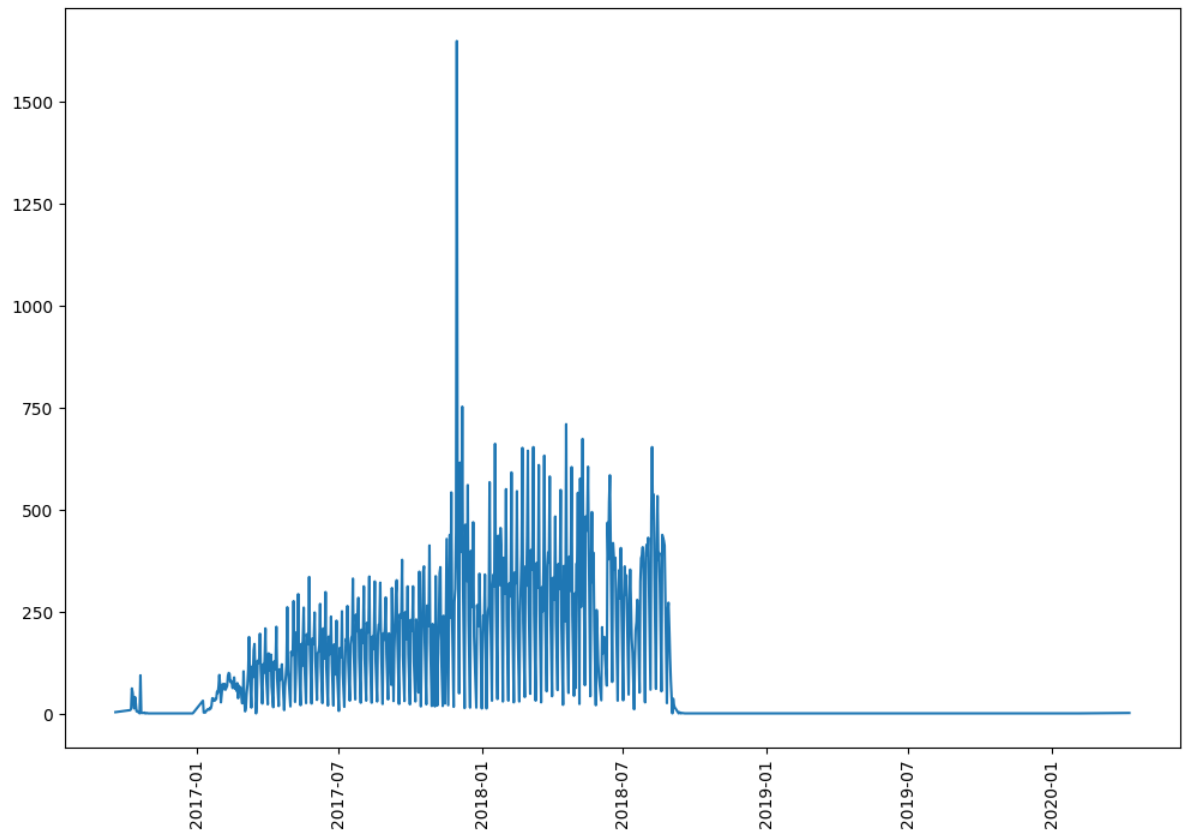
```
data_orders['shipping_limit_date']=pd.to_datetime(data_orders['shipping_limit_date']
data_orders=data_orders.sort_values(by='shipping_limit_date')
# data_orders['date']=list(map(lambda x:x.year,p['order_purchase_timestamp']))
year_counts=data_orders['shipping_limit_date'].value_counts()
year_counts=year_counts.sort_index()
year_counts
```

Out[35]:

```
2016-09-19      4
2016-10-08      9
2016-10-09     12
2016-10-10     62
2016-10-11     52
              ..
2018-09-14      2
2018-09-18      1
2020-02-03      1
2020-02-05      1
2020-04-09      2
Name: shipping_limit_date, Length: 555, dtype: int64
```

In [36]:

```
plt.figure(figsize=(12,8))
plt.plot(year_counts.index,year_counts)
plt.xticks(rotation=90)
plt.show()
```



In [37]:

```
#5、revoew_score的分布
review_data=pd.read_csv('./data/olist_order_reviews_dataset.csv')
review_data
```

Out[37]:

	review_id	order_id	review_score	re
0	7bc2406110b926393aa56f80a40eba40	73fc7af87114b39712e6da79b0a377eb	4	
1	80e641a11e56f04c1ad469d5645fdfde	a548910a1c6147796b98fdf73dbeba33	5	
2	228ce5500dc1d8e020d8d1322874b6f0	f9e4b658b201a9f2ecdecbb34bed034b	5	
3	e64fb393e7b32834bb789ff8bb30750e	658677c97b385a9be170737859d3511b	5	
4	f7c4243c7fe1938f181bec41a392bdeb	8e6bfb81e283fa7e4f11123a3fb894f1	5	
...	...	...	...	
99995	f3897127253a9592a73be9bdfdf4ed7a	22ec9f0669f784db00fa86d035cf8602	5	
99996	b3de70c89b1510c4cd3d0649fd302472	55d4004744368f5571d1f590031933e4	5	
99997	1adeb9d84d72fe4e337617733eb85149	7725825d039fc1f0ceb7635e3f7d9206	4	
99998	be360f18f5df1e0541061c87021e6d93	f8bd3f2000c28c5342fedeb5e50f2e75	1	
99999	efe49f1d6f951dd88b51e6ccd4cc548f	90531360ecb1eec2a1fbb265a0db0508	1	

100000 rows × 7 columns

In [40]:

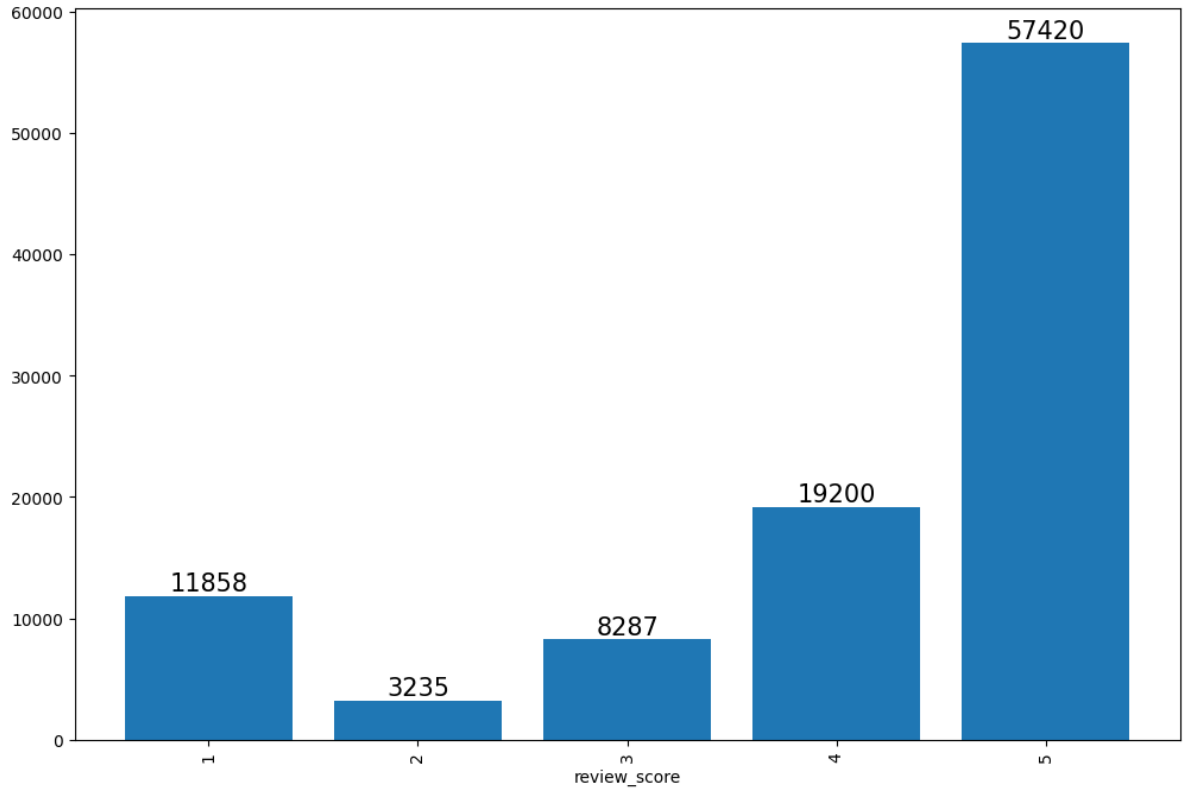
```
review=review_data['review_score'].value_counts()
review
```

Out[40]:

```
5    57420
4    19200
1    11858
3     8287
2     3235
Name: review_score, dtype: int64
```

In [43]:

```
plt.figure(figsize=(12,8),dpi=100)
rects=plt.bar(review.index,review)
plt.xticks(rotation=90)
barNum(rects)
plt.xlabel('review_score')
plt.show()
```



In [45]:

```
#product_category_name分布
product=pd.read_csv('./data/product_category_name_translation.csv')
product
```

Out[45]:

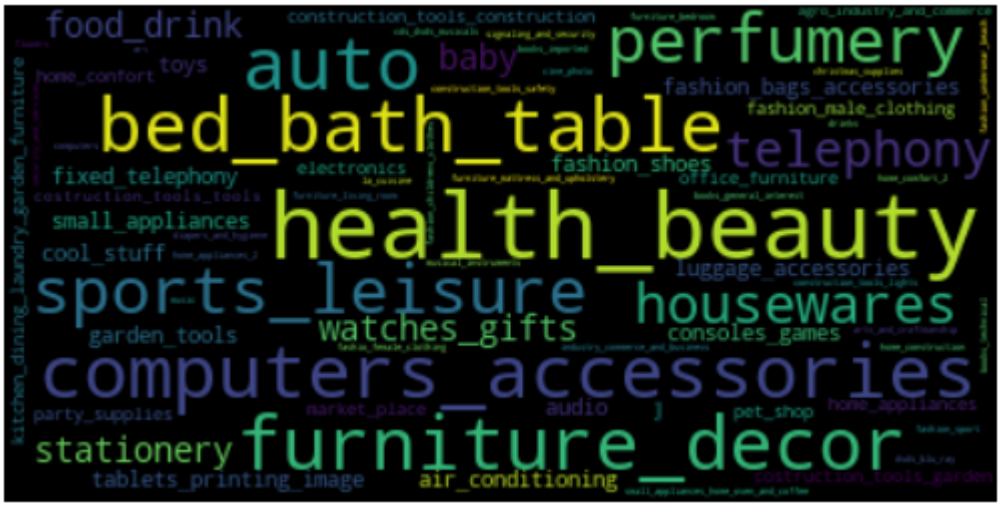
	product_category_name	product_category_name_english
0	beleza_saude	health_beauty
1	informatica_acessorios	computers_accessories
2	automotivo	auto
3	cama_mesa_banho	bed_bath_table
4	moveis_decoracao	furniture_decor
...	...	...
66	flores	flowers
67	artes_e_artesanato	arts_and_craftmanship
68	fraldas_higiene	diapers_and_hygiene
69	fashion_roupa_infanto_juvenil	fashion_childrens_clothes
70	seguros_e_servicos	security_and_services

71 rows × 2 columns

In [48]:

```
# 导入词云制作第三方库wordcloud
import wordcloud

# 创建词云对象，赋值给w，现在w就表示了一个词云对象
w = wordcloud.WordCloud()
text=' '.join(product['product_category_name_english'])
wordcloud=w.generate(text)
plt.figure(dpi=100)
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis("off")
plt.show()
```



In [49]:

```
#7、weights的分布
weights_data=pd.read_csv('./data/olist_products_dataset.csv')
weights_data
```

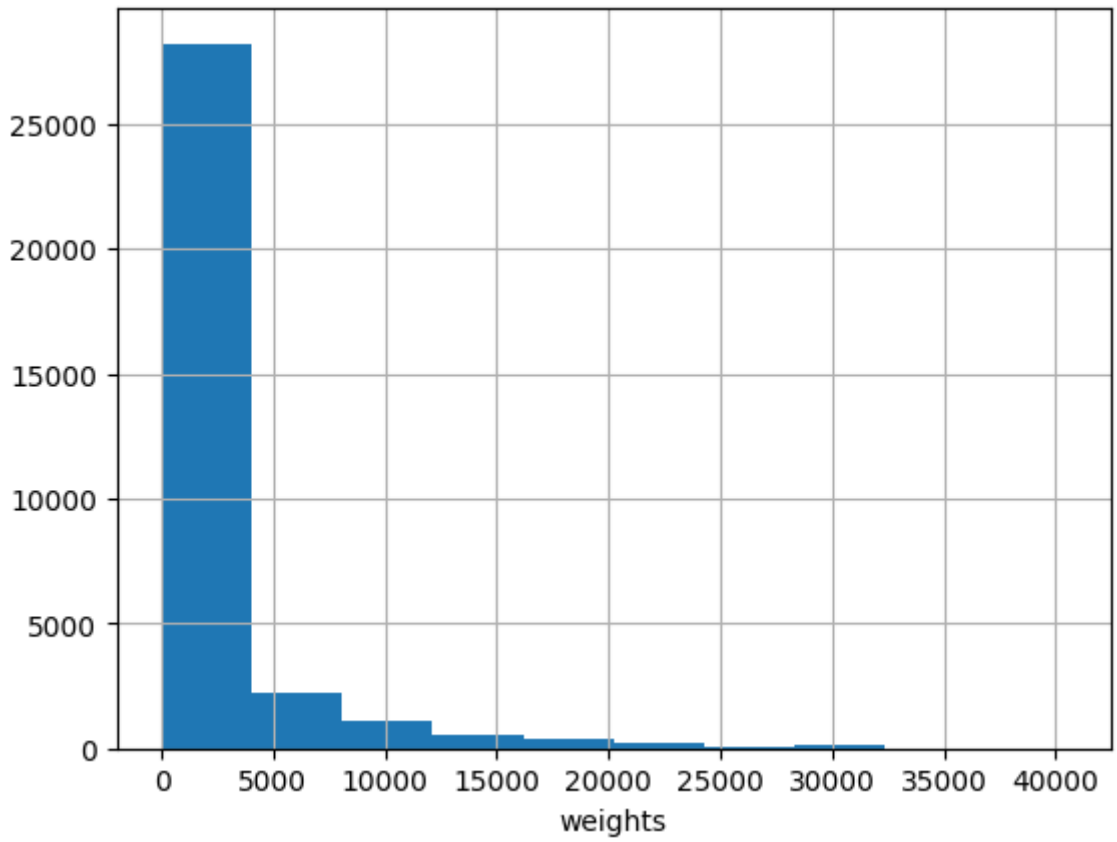
Out[49]:

	product_id	product_category_name	product_name_lengh
0	1e9e8ef04dbcff4541ed26657ea517e5	perfumaria	40.
1	3aa071139cb16b67ca9e5dea641aaa2f	artes	44.
2	96bd76ec8810374ed1b65e291975717f	esporte_lazer	46.
3	cef67bcfe19066a932b7673e239eb23d	bebes	27.
4	9dc1a7de274444849c219cff195d0b71	utilidades_domesticas	37.
...	...	...	.
32946	a0b7d5a992ccda646f2d34e418fff5a0	moveis_decoracao	45.
32947	bf4538d88321d0fd4412a93c974510e6	construcao_ferramentas_iluminacao	41.
32948	9a7c6041fa9592d9d9ef6cfe62a71f8c	cama_mesa_banho	50.
32949	83808703fc0706a22e264b9d75f04a2e	informatica_acessorios	60.
32950	106392145fca363410d287a815be6de4	cama_mesa_banho	58.

32951 rows × 9 columns

In [53]:

```
weights_data['product_weight_g'].hist()
plt.xlabel('weights')
plt.show()
```



In [57]:

```
#8、seller-state的分布
order=pd.read_csv('./data/olist_orders_dataset.csv')
order
```

Out[57]:

	order_id	customer_id	order_status	orc
0	e481f51cbdc54678b7cc49136f2d6af7	9ef432eb6251297304e76186b10a928d	delivered	
1	53cdb2fc8bc7dce0b6741e2150273451	b0830fb4747a6c6d20dea0b8c802d7ef	delivered	
2	47770eb9100c2d0c44946d9cf07ec65d	41ce2a54c0b03bf3443c3d931a367089	delivered	
3	949d5b44dbf5de918fe9c16f97b45f8a	f88197465ea7920adcdbec7375364d82	delivered	
4	ad21c59c0840e6cb83a9ceb5573f8159	8ab97904e6daea8866dbdbc4fb7aad2c	delivered	
...	...	...	...	
99436	9c5dedf39a927c1b2549525ed64a053c	39bd1228ee8140590ac3aca26f2dfe00	delivered	
99437	63943bddc261676b46f01ca7ac2f7bd8	1fca14ff2861355f6e5f14306ff977a7	delivered	
99438	83c1379a015df1e13d02aae0204711ab	1aa71eb042121263aafbe80c1b562c9c	delivered	
99439	11c177c8e97725db2631073c19f07b62	b331b74b18dc79bcdf6532d51e1637c1	delivered	
99440	66dea50a8b16d9b4dee7af250b4be1a5	edb027a75a1449115f6b43211ae02a24	delivered	

99441 rows × 8 columns

In [58]:

```
items=pd.read_csv('./data/olist_order_items_dataset.csv')
items
```

Out[58]:

	order_id	order_item_id	product_id
0	00010242fe8c5a6d1ba2dd792cb16214	1	4244733e06e7ecb4970a6e2683c13e61
1	00018f77f2f0320c557190d7a144bdd3	1	e5f2d52b802189ee658865ca93d83a8f
2	000229ec398224ef6ca0657da4fc703e	1	c777355d18b72b67abbeef9df44fd0fd
3	00024acbcdf0a6daa1e931b038114c75	1	7634da152a4610f1595efa32f14722fc
4	00042b26cf59d7ce69dfabb4e55b4fd9	1	ac6c3623068f30de03045865e4e10089
...	...	...	...
112645	fffc94f6ce00a00581880bf54a75a037	1	4aa6014eceb682077f9dc4bffebc05b0
112646	fffcd46ef2263f404302a634eb57f7eb	1	32e07fd915822b0765e448c4dd74c828
112647	fffce4705a9662cd70adb13d4a31832d	1	72a30483855e2eafc67aee5dc2560482
112648	fffe18544ffabc95dfada21779c9644f	1	9c422a519119dcad7575db5af1ba540e
112649	fffe41c64501cc87c801fd61db3f6244	1	350688d9dc1e75ff97be326363655e01

112650 rows × 7 columns

In [59]:

```
concat_items_order=pd.merge(order,items,on=['order_id'])
concat_items_order
```

Out[59]:

	order_id	customer_id	order_status	o
0	e481f51cbdc54678b7cc49136f2d6af7	9ef432eb6251297304e76186b10a928d	delivered	
1	53cdb2fc8bc7dce0b6741e2150273451	b0830fb4747a6c6d20dea0b8c802d7ef	delivered	
2	47770eb9100c2d0c44946d9cf07ec65d	41ce2a54c0b03bf3443c3d931a367089	delivered	
3	949d5b44dbf5de918fe9c16f97b45f8a	f88197465ea7920adcdbec7375364d82	delivered	
4	ad21c59c0840e6cb83a9ceb5573f8159	8ab97904e6daea8866dbdbc4fb7aad2c	delivered	
...	...	...	...	
112645	63943bddc261676b46f01ca7ac2f7bd8	1fca14ff2861355f6e5f14306ff977a7	delivered	
112646	83c1379a015df1e13d02aae0204711ab	1aa71eb042121263aafbe80c1b562c9c	delivered	
112647	11c177c8e97725db2631073c19f07b62	b331b74b18dc79bcdf6532d51e1637c1	delivered	
112648	11c177c8e97725db2631073c19f07b62	b331b74b18dc79bcdf6532d51e1637c1	delivered	
112649	66dea50a8b16d9b4dee7af250b4be1a5	edb027a75a1449115f6b43211ae02a24	delivered	

112650 rows × 14 columns

In [62]:

```
customer_state=pd.read_csv('./data/olist_customers_dataset.csv')
df1=pd.merge(concat_items_order,customer_state,on=['customer_id'])
df1
```

Out[62]:

	order_id	customer_id	order_status	o
0	e481f51cbdc54678b7cc49136f2d6af7	9ef432eb6251297304e76186b10a928d	delivered	
1	53cdb2fc8bc7dce0b6741e2150273451	b0830fb4747a6c6d20dea0b8c802d7ef	delivered	
2	47770eb9100c2d0c44946d9cf07ec65d	41ce2a54c0b03bf3443c3d931a367089	delivered	
3	949d5b44dbf5de918fe9c16f97b45f8a	f88197465ea7920adcdbec7375364d82	delivered	
4	ad21c59c0840e6cb83a9ceb5573f8159	8ab97904e6daea8866dbdbc4fb7aad2c	delivered	
...	...	...	...	
112645	63943bddc261676b46f01ca7ac2f7bd8	1fca14ff2861355f6e5f14306ff977a7	delivered	
112646	83c1379a015df1e13d02aae0204711ab	1aa71eb042121263aafbe80c1b562c9c	delivered	
112647	11c177c8e97725db2631073c19f07b62	b331b74b18dc79bcdf6532d51e1637c1	delivered	
112648	11c177c8e97725db2631073c19f07b62	b331b74b18dc79bcdf6532d51e1637c1	delivered	
112649	66dea50a8b16d9b4dee7af250b4be1a5	edb027a75a1449115f6b43211ae02a24	delivered	

112650 rows × 18 columns

In [63]:

```
seller_state=pd.read_csv('./data/olist_sellers_dataset.csv')
df2=pd.merge(df1,seller_state,on=['seller_id'])
df2
```

Out[63]:

	order_id	customer_id	order_status	o
0	e481f51cbdc54678b7cc49136f2d6af7	9ef432eb6251297304e76186b10a928d	delivered	
1	8736140c61ea584cb4250074756d8f3b	ab8844663ae049fda8baf15fc928f47f	delivered	
2	a0151737f2f0c6c0a5fd69d45f66ceea	fc2697314ab7fbeda62bb6f1afa4efcd	delivered	
3	a3bf941183211246f0d42ad757cba127	3718e1873d5dc3e8d96c0ab783278b02	delivered	
4	1462290799412b71be32dd880eaf4e1b	220e4b027f0294fd79d2869ef67e7db6	delivered	
...	...	...	...	
112645	1ab38815794efa43d269d62b98dae815	a0b67404d84a70ef420a7f99ad6b190a	delivered	
112646	b159d0ce7cd881052da94fa165617b05	e0c3bc5ce0836b975d6b2a8ce7bb0e3e	canceled	
112647	735dce2d574afe8eb87e80a3d6229c48	d531d01affc2c55769f6b9ed410d8d3c	delivered	
112648	25d2bfa43663a23586afd12f15b542e7	9d8c06734fde9823ace11a4b5929b5a7	delivered	
112649	1565f22aa9452ff278638e87cc895678	56772dfbcbe7df908a284ff0d53adf7d	delivered	

112650 rows × 21 columns

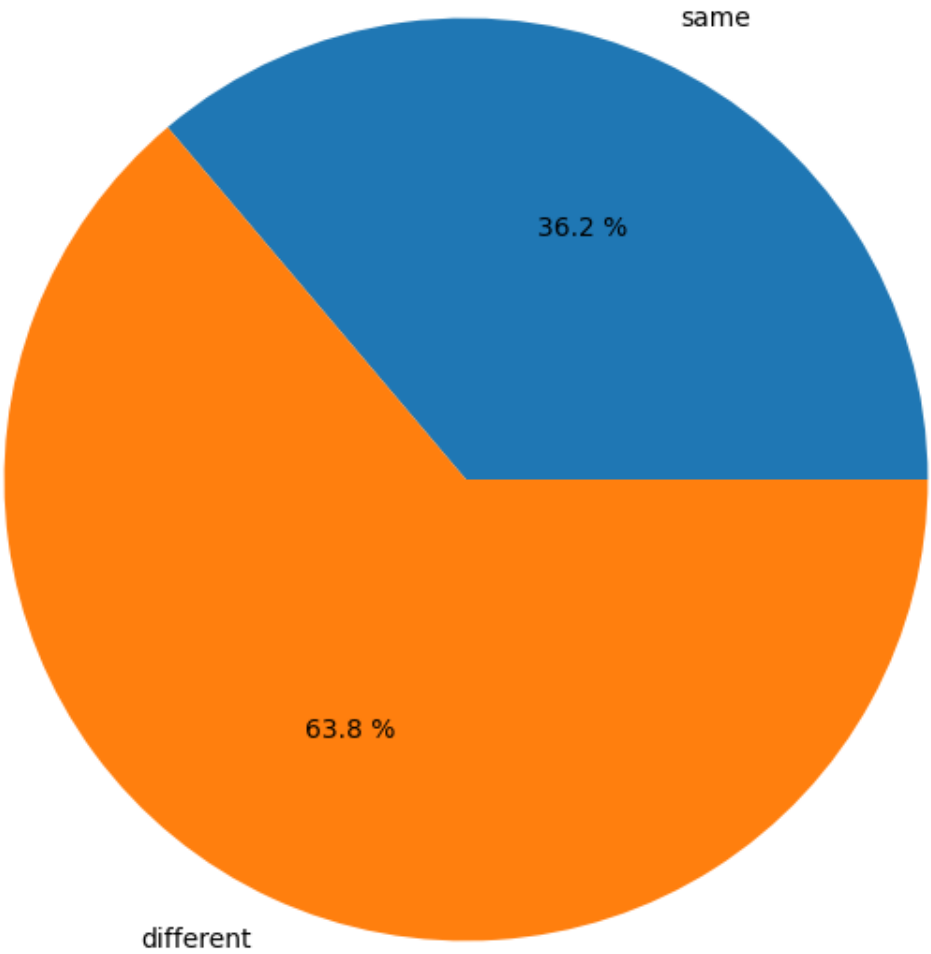
In [70]:

```
##customer_id和seller_id在一个state的比例
same_index=[i for i in range(df2.shape[0]) if df2.loc[i,'customer_state']==df2.loc[i
print('same counter:',len(same_index))
```

```
same counter: 40756
```

In [71]:

```
labels=['same','different']
num=[len(same_index),df2.shape[0]-len(same_index)]
plt.figure(figsize=(12,8),dpi=100)
plt.pie(x=num,labels=labels,autopct='%3.1f %%')
plt.show()
```



In [93]:

```
#9、用order_approved_at和order_purchase_timestamp算一个平均的，下单到approve的时间
order=order.dropna()
order['order_purchase_timestamp']=pd.to_datetime(order['order_purchase_timestamp'],f
order['order_approved_at']=pd.to_datetime(order['order_approved_at'],format='%Y-%m-%
order
```

Out[93]:

	order_id	customer_id	order_status	orc
0	e481f51cbdc54678b7cc49136f2d6af7	9ef432eb6251297304e76186b10a928d	delivered	
1	53cdb2fc8bc7dce0b6741e2150273451	b0830fb4747a6c6d20dea0b8c802d7ef	delivered	
2	47770eb9100c2d0c44946d9cf07ec65d	41ce2a54c0b03bf3443c3d931a367089	delivered	
3	949d5b44dbf5de918fe9c16f97b45f8a	f88197465ea7920adcdbec7375364d82	delivered	
4	ad21c59c0840e6cb83a9ceb5573f8159	8ab97904e6daea8866dbdbc4fb7aad2c	delivered	
...	...	...	...	
99436	9c5dedf39a927c1b2549525ed64a053c	39bd1228ee8140590ac3aca26f2dfe00	delivered	
99437	63943bddc261676b46f01ca7ac2f7bd8	1fca14ff2861355f6e5f14306ff977a7	delivered	
99438	83c1379a015df1e13d02aae0204711ab	1aa71eb042121263aafbe80c1b562c9c	delivered	
99439	11c177c8e97725db2631073c19f07b62	b331b74b18dc79bcdf6532d51e1637c1	delivered	
99440	66dea50a8b16d9b4dee7af250b4be1a5	edb027a75a1449115f6b43211ae02a24	delivered	

96461 rows × 9 columns

In [91]:

```
order.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 96461 entries, 0 to 99440
Data columns (total 9 columns):
 #   Column                         Non-Null Count  Dtype
---  ------                         --------------  -----
 0   order_id                       96461 non-null  object
 1   customer_id                    96461 non-null  object
 2   order_status                   96461 non-null  object
 3   order_purchase_timestamp       96461 non-null  datetime64[ns]
 4   order_approved_at              96461 non-null  object
 5   order_delivered_carrier_date   96461 non-null  object
 6   order_delivered_customer_date  96461 non-null  object
 7   order_estimated_delivery_date  96461 non-null  object
 8   order_approved_a               96461 non-null  datetime64[ns]
dtypes: datetime64[ns](2), object(7)
memory usage: 7.4+ MB
```

In [95]:

```
order['time_']=order['order_approved_at']-order['order_purchase_timestamp']
order
```

Out[95]:

	order_id	customer_id	order_status	orc
0	e481f51cbdc54678b7cc49136f2d6af7	9ef432eb6251297304e76186b10a928d	delivered	
1	53cdb2fc8bc7dce0b6741e2150273451	b0830fb4747a6c6d20dea0b8c802d7ef	delivered	
2	47770eb9100c2d0c44946d9cf07ec65d	41ce2a54c0b03bf3443c3d931a367089	delivered	
3	949d5b44dbf5de918fe9c16f97b45f8a	f88197465ea7920adcdbec7375364d82	delivered	
4	ad21c59c0840e6cb83a9ceb5573f8159	8ab97904e6daea8866dbdbc4fb7aad2c	delivered	
...	...	...	...	
99436	9c5dedf39a927c1b2549525ed64a053c	39bd1228ee8140590ac3aca26f2dfe00	delivered	
99437	63943bddc261676b46f01ca7ac2f7bd8	1fca14ff2861355f6e5f14306ff977a7	delivered	
99438	83c1379a015df1e13d02aae0204711ab	1aa71eb042121263aafbe80c1b562c9c	delivered	
99439	11c177c8e97725db2631073c19f07b62	b331b74b18dc79bcdf6532d51e1637c1	delivered	
99440	66dea50a8b16d9b4dee7af250b4be1a5	edb027a75a1449115f6b43211ae02a24	delivered	

96461 rows × 10 columns

In [98]:

```
s = pd.to_timedelta(order['time_']).astype('timedelta64[s]').astype(int)##转换为秒钟
s
```

Out[98]:

```
0            642
1         110570
2            994
3           1073
4           3710
           ...
99436          0
99437        699
99438       1053
99439        474
99440      51778
Name: time_, Length: 96461, dtype: int64
```

In [102]:

```
##平均多少秒:
import numpy as np
print('mean seconds:',np.mean(s))
```

```
mean seconds: 36999.83789303449
```

In [ ]: