Análise Exploratória

Compreensão inicial dos dados, identificando variáveis, tipos e distribuições. Isso envolve carregar os dados, inspecionar as primeiras linhas, verificar tipos de dados, resumir estatísticas descritivas e lidar com valores ausentes.

Instalação das Bibliotecas

```
In [8]:  %pip install pandas matplotlib seaborn
```

```
Requirement already satisfied: pandas in c:\python312\lib\site-packages (2.2.2)No
te: you may need to restart the kernel to use updated packages.
```

[notice] A new release of pip is available: 24.0 -> 24.3.1
[notice] To update, run: python.exe -m pip install --upgrade pip

```
Requirement already satisfied: matplotlib in c:\python312\lib\site-packages (3.9.
2)
Requirement already satisfied: seaborn in c:\python312\lib\site-packages (0.13.2)
Requirement already satisfied: numpy>=1.26.0 in c:\python312\lib\site-packages (f
rom pandas) (1.26.4)
Requirement already satisfied: python-dateutil>=2.8.2 in c:\python312\lib\site-pa
ckages (from pandas) (2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in c:\python312\lib\site-packages (fr
om pandas) (2024.1)
Requirement already satisfied: tzdata>=2022.7 in c:\python312\lib\site-packages
(from pandas) (2024.1)
Requirement already satisfied: contourpy>=1.0.1 in c:\python312\lib\site-packages
(from matplotlib) (1.3.0)
Requirement already satisfied: cycler>=0.10 in c:\python312\lib\site-packages (fr
om matplotlib) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in c:\python312\lib\site-package
s (from matplotlib) (4.53.1)
Requirement already satisfied: kiwisolver>=1.3.1 in c:\python312\lib\site-package
s (from matplotlib) (1.4.7)
Requirement already satisfied: packaging>=20.0 in c:\users\salom\appdata\roaming
\python\python312\site-packages (from matplotlib) (24.0)
Requirement already satisfied: pillow>=8 in c:\python312\lib\site-packages (from
matplotlib) (10.3.0)
Requirement already satisfied: pyparsing>=2.3.1 in c:\python312\lib\site-packages
(from matplotlib) (3.1.4)
Requirement already satisfied: six>=1.5 in c:\python312\lib\site-packages (from p
ython-dateutil>=2.8.2->pandas) (1.16.0)
```

Importação das tabelas

```
In [9]:  import pandas as pd

         # Carregar os datasets
         customers = pd.read_csv("olist_customers_dataset.csv")
         geolocation = pd.read_csv("olist_geolocation_dataset.csv")
         order_items = pd.read_csv("olist_order_items_dataset.csv")
         order_payments = pd.read_csv("olist_order_payments_dataset.csv")
         order_reviews = pd.read_csv("olist_order_reviews_dataset.csv")
         orders = pd.read_csv("olist_orders_dataset.csv")
```

```python
products = pd.read_csv("olist_products_dataset.csv")
sellers = pd.read_csv("olist_sellers_dataset.csv")
product_category_translation = pd.read_csv("product_category_name_translation.cs

print("-----------------------------------------------------------------
print(customers.head())  # Primeiras 5 linhas
print(customers.info())   # Informações sobre as colunas
print(customers.describe()) # Estatísticas descritivas
print("-----------------------------------------------------------------

print("-----------------------------------------------------------------
print(geolocation.head())  # Primeiras 5 linhas
print(geolocation.info())   # Informações sobre as colunas
print(geolocation.describe()) # Estatísticas descritivas
print("-----------------------------------------------------------------

print("-----------------------------------------------------------------
print(order_items.head())  # Primeiras 5 linhas
print(order_items.info())   # Informações sobre as colunas
print(order_items.describe()) # Estatísticas descritivas
print("-----------------------------------------------------------------

print("-----------------------------------------------------------------
print(order_payments.head())  # Primeiras 5 linhas
print(order_payments.info())   # Informações sobre as colunas
print(order_payments.describe()) # Estatísticas descritivas
print("-----------------------------------------------------------------

print("-----------------------------------------------------------------
print(order_reviews.head())  # Primeiras 5 linhas
print(order_reviews.info())   # Informações sobre as colunas
print(order_reviews.describe()) # Estatísticas descritivas
print("-----------------------------------------------------------------

print("-----------------------------------------------------------------
print(orders.head())  # Primeiras 5 linhas
print(orders.info())   # Informações sobre as colunas
print(orders.describe()) # Estatísticas descritivas
print("-----------------------------------------------------------------

print("-----------------------------------------------------------------
print(products.head())  # Primeiras 5 linhas
print(products.info())   # Informações sobre as colunas
print(products.describe()) # Estatísticas descritivas
print("-----------------------------------------------------------------

print("-----------------------------------------------------------------
print(sellers.head())  # Primeiras 5 linhas
print(sellers.info())   # Informações sobre as colunas
print(sellers.describe()) # Estatísticas descritivas
print("-----------------------------------------------------------------

print("-----------------------------------------------------------------
print(product_category_translation.head())  # Primeiras 5 linhas
print(product_category_translation.info())   # Informações sobre as colunas
print(product_category_translation.describe()) # Estatísticas descritivas
print("-----------------------------------------------------------------
```

```
--------------------------------------------------------------------------
                     customer_id                customer_unique_id  \
0  06b8999e2fba1a1fbc88172c00ba8bc7  861eff4711a542e4b93843c6dd7febb0
1  18955e83d337fd6b2def6b18a428ac77  290c77bc529b7ac935b93aa66c333dc3
2  4e7b3e00288586ebd08712fdd0374a03  060e732b5b29e8181a18229c7b0b2b5e
3  b2b6027bc5c5109e529d4dc6358b12c3  259dac757896d24d7702b9acbbff3f3c
4  4f2d8ab171c80ec8364f7c12e35b23ad  345ecd01c38d18a9036ed96c73b8d066

   customer_zip_code_prefix          customer_city customer_state
0                     14409                 franca             SP
1                      9790  sao bernardo do campo             SP
2                      1151              sao paulo             SP
3                      8775        mogi das cruzes             SP
4                     13056               campinas             SP
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99441 entries, 0 to 99440
Data columns (total 5 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   customer_id               99441 non-null  object
 1   customer_unique_id        99441 non-null  object
 2   customer_zip_code_prefix  99441 non-null  int64
 3   customer_city             99441 non-null  object
 4   customer_state            99441 non-null  object
dtypes: int64(1), object(4)
memory usage: 3.8+ MB
None
       customer_zip_code_prefix
count              99441.000000
mean               35137.474583
std                29797.938996
min                 1003.000000
25%                11347.000000
50%                24416.000000
75%                58900.000000
max                99990.000000
--------------------------------------------------------------------------
--------------------------------------------------------------------------
   geolocation_zip_code_prefix  geolocation_lat  geolocation_lng  \
0                         1037       -23.545621       -46.639292
1                         1046       -23.546081       -46.644820
2                         1046       -23.546129       -46.642951
3                         1041       -23.544392       -46.639499
4                         1035       -23.541578       -46.641607

  geolocation_city geolocation_state
0        sao paulo                SP
1        sao paulo                SP
2        sao paulo                SP
3        sao paulo                SP
4        sao paulo                SP
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000163 entries, 0 to 1000162
Data columns (total 5 columns):
 #   Column                       Non-Null Count    Dtype
---  ------                       --------------    -----
 0   geolocation_zip_code_prefix  1000163 non-null  int64
 1   geolocation_lat              1000163 non-null  float64
 2   geolocation_lng              1000163 non-null  float64
 3   geolocation_city             1000163 non-null  object
```

```
  4   geolocation_state          1000163 non-null   object
dtypes: float64(2), int64(1), object(2)
memory usage: 38.2+ MB
None
       geolocation_zip_code_prefix  geolocation_lat  geolocation_lng
count                 1.000163e+06     1.000163e+06     1.000163e+06
mean                  3.657417e+04    -2.117615e+01    -4.639054e+01
std                   3.054934e+04     5.715866e+00     4.269748e+00
min                   1.001000e+03    -3.660537e+01    -1.014668e+02
25%                   1.107500e+04    -2.360355e+01    -4.857317e+01
50%                   2.653000e+04    -2.291938e+01    -4.663788e+01
75%                   6.350400e+04    -1.997962e+01    -4.376771e+01
max                   9.999000e+04     4.506593e+01     1.211054e+02
--------------------------------------------------------------------------
--------------------------------------------------------------------------
                          order_id  order_item_id  \
0  00010242fe8c5a6d1ba2dd792cb16214              1
1  00018f77f2f0320c557190d7a144bdd3              1
2  000229ec398224ef6ca0657da4fc703e              1
3  00024acbcdf0a6daa1e931b038114c75              1
4  00042b26cf59d7ce69dfabb4e55b4fd9              1


                         product_id                         seller_id  \
0  4244733e06e7ecb4970a6e2683c13e61  48436dade18ac8b2bce089ec2a041202
1  e5f2d52b802189ee658865ca93d83a8f  dd7ddc04e1b6c2c614352b383efe2d36
2  c777355d18b72b67abbeef9df44fd0fd  5b51032eddd242adc84c38acab88f23d
3  7634da152a4610f1595efa32f14722fc  9d7a1d34a5052409006425275ba1c2b4
4  ac6c3623068f30de03045865e4e10089  df560393f3a51e74553ab94004ba5c87


   shipping_limit_date    price  freight_value
0  2017-09-19 09:45:35    58.90          13.29
1  2017-05-03 11:05:13   239.90          19.93
2  2018-01-18 14:48:30   199.00          17.87
3  2018-08-15 10:10:18    12.99          12.79
4  2017-02-13 13:57:51   199.90          18.14
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 112650 entries, 0 to 112649
Data columns (total 7 columns):
 #   Column               Non-Null Count    Dtype
---  ------               --------------    -----
 0   order_id             112650 non-null   object
 1   order_item_id        112650 non-null   int64
 2   product_id           112650 non-null   object
 3   seller_id            112650 non-null   object
 4   shipping_limit_date  112650 non-null   object
 5   price                112650 non-null   float64
 6   freight_value        112650 non-null   float64
dtypes: float64(2), int64(1), object(4)
memory usage: 6.0+ MB
None
       order_item_id          price  freight_value
count  112650.000000  112650.000000  112650.000000
mean        1.197834     120.653739      19.990320
std         0.705124     183.633928      15.806405
min         1.000000       0.850000       0.000000
25%         1.000000      39.900000      13.080000
50%         1.000000      74.990000      16.260000
75%         1.000000     134.900000      21.150000
max        21.000000    6735.000000     409.680000
--------------------------------------------------------------------------
```

```
--------------------------------------------------------------------------
                       order_id  payment_sequential payment_type  \
0  b81ef226f3fe1789b1e8b2acac839d17                   1  credit_card
1  a9810da82917af2d9aefd1278f1dcfa0                   1  credit_card
2  25e8ea4e93396b6fa0d3dd708e76c1bd                   1  credit_card
3  ba78997921bbcdc1373bb41e913ab953                   1  credit_card
4  42fdf880ba16b47b59251dd489d4441a                   1  credit_card


   payment_installments  payment_value
0                     8          99.33
1                     1          24.39
2                     1          65.71
3                     8         107.78
4                     2         128.45
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 103886 entries, 0 to 103885
Data columns (total 5 columns):
 #   Column                Non-Null Count    Dtype
---  ------                --------------    -----
 0   order_id              103886 non-null   object
 1   payment_sequential    103886 non-null   int64
 2   payment_type          103886 non-null   object
 3   payment_installments  103886 non-null   int64
 4   payment_value         103886 non-null   float64
dtypes: float64(1), int64(2), object(2)
memory usage: 4.0+ MB
None
       payment_sequential  payment_installments  payment_value
count       103886.000000         103886.000000  103886.000000
mean             1.092679              2.853349     154.100380
std              0.706584              2.687051     217.494064
min              1.000000              0.000000       0.000000
25%              1.000000              1.000000      56.790000
50%              1.000000              1.000000     100.000000
75%              1.000000              4.000000     171.837500
max             29.000000             24.000000   13664.080000
--------------------------------------------------------------------------
--------------------------------------------------------------------------
                          review_id                          order_id  \
0  7bc2406110b926393aa56f80a40eba40  73fc7af87114b39712e6da79b0a377eb
1  80e641a11e56f04c1ad469d5645fdfde  a548910a1c6147796b98fdf73dbeba33
2  228ce5500dc1d8e020d8d1322874b6f0  f9e4b658b201a9f2ecdecbb34bed034b
3  e64fb393e7b32834bb789ff8bb30750e  658677c97b385a9be170737859d3511b
4  f7c4243c7fe1938f181bec41a392bdeb  8e6bfb81e283fa7e4f11123a3fb894f1


   review_score review_comment_title  \
0             4                  NaN
1             5                  NaN
2             5                  NaN
3             5                  NaN
4             5                  NaN


                           review_comment_message review_creation_date  \
0                                             NaN  2018-01-18 00:00:00
1                                             NaN  2018-03-10 00:00:00
2                                             NaN  2018-02-17 00:00:00
3            Recebi bem antes do prazo estipulado.  2017-04-21 00:00:00
4  Parabéns lojas lannister adorei comprar pela I...  2018-03-01 00:00:00

  review_answer_timestamp
```

```
0      2018-01-18 21:46:59
1      2018-03-11 03:05:13
2      2018-02-18 14:36:24
3      2017-04-21 22:02:06
4      2018-03-02 10:26:53
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99224 entries, 0 to 99223
Data columns (total 7 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   review_id              99224 non-null  object
 1   order_id               99224 non-null  object
 2   review_score           99224 non-null  int64
 3   review_comment_title   11568 non-null  object
 4   review_comment_message 40977 non-null  object
 5   review_creation_date   99224 non-null  object
 6   review_answer_timestamp 99224 non-null  object
dtypes: int64(1), object(6)
memory usage: 5.3+ MB
None
       review_score
count  99224.000000
mean       4.086421
std        1.347579
min        1.000000
25%        4.000000
50%        5.000000
75%        5.000000
max        5.000000
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
                           order_id                       customer_id  \
0  e481f51cbdc54678b7cc49136f2d6af7  9ef432eb6251297304e76186b10a928d
1  53cdb2fc8bc7dce0b6741e2150273451  b0830fb4747a6c6d20dea0b8c802d7ef
2  47770eb9100c2d0c44946d9cf07ec65d  41ce2a54c0b03bf3443c3d931a367089
3  949d5b44dbf5de918fe9c16f97b45f8a  f88197465ea7920adcdbec7375364d82
4  ad21c59c0840e6cb83a9ceb5573f8159  8ab97904e6daea8866dbdbc4fb7aad2c

  order_status order_purchase_timestamp     order_approved_at  \
0    delivered      2017-10-02 10:56:33   2017-10-02 11:07:15
1    delivered      2018-07-24 20:41:37   2018-07-26 03:24:27
2    delivered      2018-08-08 08:38:49   2018-08-08 08:55:23
3    delivered      2017-11-18 19:28:06   2017-11-18 19:45:59
4    delivered      2018-02-13 21:18:39   2018-02-13 22:20:29

  order_delivered_carrier_date order_delivered_customer_date  \
0          2017-10-04 19:55:00            2017-10-10 21:25:13
1          2018-07-26 14:31:00            2018-08-07 15:27:45
2          2018-08-08 13:50:00            2018-08-17 18:06:29
3          2017-11-22 13:39:59            2017-12-02 00:28:42
4          2018-02-14 19:46:34            2018-02-16 18:17:02

  order_estimated_delivery_date
0           2017-10-18 00:00:00
1           2018-08-13 00:00:00
2           2018-09-04 00:00:00
3           2017-12-15 00:00:00
4           2018-02-26 00:00:00
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99441 entries, 0 to 99440
```

```
Data columns (total 8 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   order_id                      99441 non-null  object
 1   customer_id                   99441 non-null  object
 2   order_status                  99441 non-null  object
 3   order_purchase_timestamp      99441 non-null  object
 4   order_approved_at             99281 non-null  object
 5   order_delivered_carrier_date  97658 non-null  object
 6   order_delivered_customer_date 96476 non-null  object
 7   order_estimated_delivery_date 99441 non-null  object
dtypes: object(8)
memory usage: 6.1+ MB
None
                            order_id                       customer_id  \
count                          99441                             99441
unique                         99441                             99441
top     e481f51cbdc54678b7cc49136f2d6af7  9ef432eb6251297304e76186b10a928d
freq                               1                                 1

       order_status order_purchase_timestamp    order_approved_at  \
count         99441                    99441                99281
unique            8                    98875                90733
top       delivered      2018-04-11 10:48:14  2018-02-27 04:31:10
freq          96478                        3                    9

       order_delivered_carrier_date order_delivered_customer_date  \
count                         97658                         96476
unique                        81018                         95664
top             2018-05-09 15:48:00           2018-05-08 23:38:46
freq                             47                             3

       order_estimated_delivery_date
count                          99441
unique                           459
top              2017-12-20 00:00:00
freq                             522
--------------------------------------------------------------------------
--------------------------------------------------------------------------
                         product_id  product_category_name  \
0  1e9e8ef04dbcff4541ed26657ea517e5             perfumaria
1  3aa071139cb16b67ca9e5dea641aaa2f                  artes
2  96bd76ec8810374ed1b65e291975717f          esporte_lazer
3  cef67bcfe19066a932b7673e239eb23d                  bebes
4  9dc1a7de274444849c219cff195d0b71  utilidades_domesticas

   product_name_lenght  product_description_lenght  product_photos_qty  \
0                 40.0                       287.0                 1.0
1                 44.0                       276.0                 1.0
2                 46.0                       250.0                 1.0
3                 27.0                       261.0                 1.0
4                 37.0                       402.0                 4.0

   product_weight_g  product_length_cm  product_height_cm  product_width_cm
0             225.0               16.0               10.0              14.0
1            1000.0               30.0               18.0              20.0
2             154.0               18.0                9.0              15.0
3             371.0               26.0                4.0              26.0
4             625.0               20.0               17.0              13.0
<class 'pandas.core.frame.DataFrame'>
```

```
RangeIndex: 32951 entries, 0 to 32950
Data columns (total 9 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   product_id                  32951 non-null  object
 1   product_category_name       32341 non-null  object
 2   product_name_lenght         32341 non-null  float64
 3   product_description_lenght  32341 non-null  float64
 4   product_photos_qty          32341 non-null  float64
 5   product_weight_g            32949 non-null  float64
 6   product_length_cm           32949 non-null  float64
 7   product_height_cm           32949 non-null  float64
 8   product_width_cm            32949 non-null  float64
dtypes: float64(7), object(2)
memory usage: 2.3+ MB
None
```

	product_name_lenght	product_description_lenght	product_photos_qty \
count	32341.000000	32341.000000	32341.000000
mean	48.476949	771.495285	2.188986
std	10.245741	635.115225	1.736766
min	5.000000	4.000000	1.000000
25%	42.000000	339.000000	1.000000
50%	51.000000	595.000000	1.000000
75%	57.000000	972.000000	3.000000
max	76.000000	3992.000000	20.000000

	product_weight_g	product_length_cm	product_height_cm \
count	32949.000000	32949.000000	32949.000000
mean	2276.472488	30.815078	16.937661
std	4282.038731	16.914458	13.637554
min	0.000000	7.000000	2.000000
25%	300.000000	18.000000	8.000000
50%	700.000000	25.000000	13.000000
75%	1900.000000	38.000000	21.000000
max	40425.000000	105.000000	105.000000

	product_width_cm
count	32949.000000
mean	23.196728
std	12.079047
min	6.000000
25%	15.000000
50%	20.000000
75%	30.000000
max	118.000000

```
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
```

	seller_id	seller_zip_code_prefix \
0	3442f8959a84dea7ee197c632cb2df15	13023
1	d1b65fc7debc3361ea86b5f14c68d2e2	13844
2	ce3ad9de960102d0677a81f5d0bb7b2d	20031
3	c0f3eea2e14555b6faeea3dd58c1b1c3	4195
4	51a04a8a6bdcb23deccc82b0b80742cf	12914

	seller_city	seller_state
0	campinas	SP
1	mogi guacu	SP
2	rio de janeiro	RJ
3	sao paulo	SP
4	braganca paulista	SP

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3095 entries, 0 to 3094
Data columns (total 4 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   seller_id              3095 non-null   object
 1   seller_zip_code_prefix 3095 non-null   int64
 2   seller_city            3095 non-null   object
 3   seller_state           3095 non-null   object
dtypes: int64(1), object(3)
memory usage: 96.8+ KB
None
       seller_zip_code_prefix
count            3095.000000
mean            32291.059451
std             32713.453830
min              1001.000000
25%              7093.500000
50%             14940.000000
75%             64552.500000
max             99730.000000
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
    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
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 71 entries, 0 to 70
Data columns (total 2 columns):
 #   Column                         Non-Null Count  Dtype
---  ------                         --------------  -----
 0   product_category_name          71 non-null     object
 1   product_category_name_english  71 non-null     object
dtypes: object(2)
memory usage: 1.2+ KB
None
       product_category_name product_category_name_english
count                     71                             71
unique                    71                             71
top             beleza_saude                  health_beauty
freq                       1                              1
--------------------------------------------------------------------------------
```

Exploração básica de algumas tabelas

```python
In [10]: import pandas as pd
         import matplotlib.pyplot as plt
         import seaborn as sns

         # Função para explorar um dataset
         def explore_data(df, name):
             print(f"\nExplorando o dataset {name}:")
             print(df.head())
             print(f"\nInformações sobre {name}:")
             print(df.info())
             print(f"\nEstatísticas descritivas de {name}:")
```

```python
    print(df.describe())
    print(f"\nValores ausentes em {name}:")
    print(df.isnull().sum())


# Explorar cada dataset
explore_data(customers, "Clientes")
explore_data(geolocation, "Geolocalização")
explore_data(order_items, "Itens do Pedido")
explore_data(order_payments, "Pagamentos")
explore_data(order_reviews, "Avaliações")
explore_data(orders, "Pedidos")
explore_data(products, "Produtos")
explore_data(sellers, "Vendedores")
explore_data(product_category_translation, "Tradução de Categorias")


# Visualizações iniciais (exemplos)
plt.figure(figsize=(10, 5))
sns.countplot(x='order_status', data=orders)
plt.title('Distribuição dos Status dos Pedidos')
plt.show()


plt.figure(figsize=(12, 6))
sns.histplot(order_items['price'], bins=50, kde=True)
plt.title('Distribuição do Preço dos Itens')
plt.show()


# Visualizações adicionais
plt.figure(figsize=(12, 6))
sns.boxplot(x='payment_type', y='payment_value', data=order_payments)
plt.title('Valor do Pagamento por Tipo de Pagamento')
plt.xticks(rotation=45, ha='right') # Rotaciona os labels do eixo x para melhor
plt.show()

plt.figure(figsize=(10, 5))
sns.countplot(x='payment_installments', data=order_payments)
plt.title('Número de Parcelas')
plt.show()


print("\nFim da Exploração Inicial.")
```

```
Explorando o dataset Clientes:
                       customer_id                  customer_unique_id  \
0  06b8999e2fba1a1fbc88172c00ba8bc7  861eff4711a542e4b93843c6dd7febb0
1  18955e83d337fd6b2def6b18a428ac77  290c77bc529b7ac935b93aa66c333dc3
2  4e7b3e00288586ebd08712fdd0374a03  060e732b5b29e8181a18229c7b0b2b5e
3  b2b6027bc5c5109e529d4dc6358b12c3  259dac757896d24d7702b9acbbff3f3c
4  4f2d8ab171c80ec8364f7c12e35b23ad  345ecd01c38d18a9036ed96c73b8d066

   customer_zip_code_prefix          customer_city customer_state
0                     14409                 franca             SP
1                      9790  sao bernardo do campo             SP
2                      1151              sao paulo             SP
3                      8775        mogi das cruzes             SP
4                     13056               campinas             SP

Informações sobre Clientes:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99441 entries, 0 to 99440
Data columns (total 5 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   customer_id               99441 non-null  object
 1   customer_unique_id        99441 non-null  object
 2   customer_zip_code_prefix  99441 non-null  int64
 3   customer_city             99441 non-null  object
 4   customer_state            99441 non-null  object
dtypes: int64(1), object(4)
memory usage: 3.8+ MB
None

Estatísticas descritivas de Clientes:
       customer_zip_code_prefix
count              99441.000000
mean               35137.474583
std                29797.938996
min                 1003.000000
25%                11347.000000
50%                24416.000000
75%                58900.000000
max                99990.000000

Valores ausentes em Clientes:
customer_id                 0
customer_unique_id          0
customer_zip_code_prefix    0
customer_city               0
customer_state              0
dtype: int64

Explorando o dataset Geolocalização:
   geolocation_zip_code_prefix  geolocation_lat  geolocation_lng  \
0                         1037       -23.545621       -46.639292
1                         1046       -23.546081       -46.644820
2                         1046       -23.546129       -46.642951
3                         1041       -23.544392       -46.639499
4                         1035       -23.541578       -46.641607

  geolocation_city geolocation_state
0        sao paulo                SP
1        sao paulo                SP
```

```
2        sao paulo                    SP
3        sao paulo                    SP
4        sao paulo                    SP

Informações sobre Geolocalização:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000163 entries, 0 to 1000162
Data columns (total 5 columns):
 #   Column                     Non-Null Count    Dtype
---  ------                     --------------    -----
 0   geolocation_zip_code_prefix  1000163 non-null  int64
 1   geolocation_lat              1000163 non-null  float64
 2   geolocation_lng              1000163 non-null  float64
 3   geolocation_city             1000163 non-null  object
 4   geolocation_state            1000163 non-null  object
dtypes: float64(2), int64(1), object(2)
memory usage: 38.2+ MB
None

Estatísticas descritivas de Geolocalização:
       geolocation_zip_code_prefix  geolocation_lat  geolocation_lng
count              1.000163e+06       1.000163e+06     1.000163e+06
mean               3.657417e+04      -2.117615e+01    -4.639054e+01
std                3.054934e+04       5.715866e+00     4.269748e+00
min                1.001000e+03      -3.660537e+01    -1.014668e+02
25%                1.107500e+04      -2.360355e+01    -4.857317e+01
50%                2.653000e+04      -2.291938e+01    -4.663788e+01
75%                6.350400e+04      -1.997962e+01    -4.376771e+01
max                9.999000e+04       4.506593e+01     1.211054e+02

Valores ausentes em Geolocalização:
geolocation_zip_code_prefix     0
geolocation_lat                 0
geolocation_lng                 0
geolocation_city                0
geolocation_state               0
dtype: int64

Explorando o dataset Itens do Pedido:
                          order_id  order_item_id  \
0  00010242fe8c5a6d1ba2dd792cb16214              1
1  00018f77f2f0320c557190d7a144bdd3              1
2  000229ec398224ef6ca0657da4fc703e              1
3  00024acbcdf0a6daa1e931b038114c75              1
4  00042b26cf59d7ce69dfabb4e55b4fd9              1

                     product_id                     seller_id  \
0  4244733e06e7ecb4970a6e2683c13e61  48436dade18ac8b2bce089ec2a041202
1  e5f2d52b802189ee658865ca93d83a8f  dd7ddc04e1b6c2c614352b383efe2d36
2  c777355d18b72b67abbeef9df44fd0fd  5b51032eddd242adc84c38acab88f23d
3  7634da152a4610f1595efa32f14722fc  9d7a1d34a5052409006425275ba1c2b4
4  ac6c3623068f30de03045865e4e10089  df560393f3a51e74553ab94004ba5c87

   shipping_limit_date   price  freight_value
0  2017-09-19 09:45:35   58.90          13.29
1  2017-05-03 11:05:13  239.90          19.93
2  2018-01-18 14:48:30  199.00          17.87
3  2018-08-15 10:10:18   12.99          12.79
4  2017-02-13 13:57:51  199.90          18.14
```

```
Informações sobre Itens do Pedido:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 112650 entries, 0 to 112649
Data columns (total 7 columns):
 #   Column              Non-Null Count    Dtype
---  ------              --------------    -----
 0   order_id            112650 non-null   object
 1   order_item_id       112650 non-null   int64
 2   product_id          112650 non-null   object
 3   seller_id           112650 non-null   object
 4   shipping_limit_date 112650 non-null   object
 5   price               112650 non-null   float64
 6   freight_value       112650 non-null   float64
dtypes: float64(2), int64(1), object(4)
memory usage: 6.0+ MB
None

Estatísticas descritivas de Itens do Pedido:
        order_item_id          price   freight_value
count   112650.000000  112650.000000   112650.000000
mean         1.197834     120.653739       19.990320
std          0.705124     183.633928       15.806405
min          1.000000       0.850000        0.000000
25%          1.000000      39.900000       13.080000
50%          1.000000      74.990000       16.260000
75%          1.000000     134.900000       21.150000
max         21.000000    6735.000000      409.680000

Valores ausentes em Itens do Pedido:
order_id              0
order_item_id        0
product_id           0
seller_id            0
shipping_limit_date  0
price                0
freight_value        0
dtype: int64

Explorando o dataset Pagamentos:
                           order_id  payment_sequential payment_type  \
0  b81ef226f3fe1789b1e8b2acac839d17                   1  credit_card
1  a9810da82917af2d9aefd1278f1dcfa0                   1  credit_card
2  25e8ea4e93396b6fa0d3dd708e76c1bd                   1  credit_card
3  ba78997921bbcdc1373bb41e913ab953                   1  credit_card
4  42fdf880ba16b47b59251dd489d4441a                   1  credit_card

   payment_installments  payment_value
0                     8          99.33
1                     1          24.39
2                     1          65.71
3                     8         107.78
4                     2         128.45

Informações sobre Pagamentos:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 103886 entries, 0 to 103885
Data columns (total 5 columns):
 #   Column      Non-Null Count    Dtype
---  ------      --------------    -----
 0   order_id    103886 non-null   object
```

```
 1   payment_sequential    103886 non-null  int64
 2   payment_type          103886 non-null  object
 3   payment_installments  103886 non-null  int64
 4   payment_value         103886 non-null  float64
dtypes: float64(1), int64(2), object(2)
memory usage: 4.0+ MB
None
```

Estatísticas descritivas de Pagamentos:

	payment_sequential	payment_installments	payment_value
count	103886.000000	103886.000000	103886.000000
mean	1.092679	2.853349	154.100380
std	0.706584	2.687051	217.494064
min	1.000000	0.000000	0.000000
25%	1.000000	1.000000	56.790000
50%	1.000000	1.000000	100.000000
75%	1.000000	4.000000	171.837500
max	29.000000	24.000000	13664.080000

Valores ausentes em Pagamentos:
```
order_id              0
payment_sequential    0
payment_type          0
payment_installments  0
payment_value         0
dtype: int64
```

Explorando o dataset Avaliações:

	review_id	order_id	\
0	7bc2406110b926393aa56f80a40eba40	73fc7af87114b39712e6da79b0a377eb	
1	80e641a11e56f04c1ad469d5645fdfde	a548910a1c6147796b98fdf73dbeba33	
2	228ce5500dc1d8e020d8d1322874b6f0	f9e4b658b201a9f2ecdecbb34bed034b	
3	e64fb393e7b32834bb789ff8bb30750e	658677c97b385a9be170737859d3511b	
4	f7c4243c7fe1938f181bec41a392bdeb	8e6bfb81e283fa7e4f11123a3fb894f1	

	review_score	review_comment_title	\
0	4	NaN	
1	5	NaN	
2	5	NaN	
3	5	NaN	
4	5	NaN	

	review_comment_message	review_creation_date	\
0	NaN	2018-01-18 00:00:00	
1	NaN	2018-03-10 00:00:00	
2	NaN	2018-02-17 00:00:00	
3	Recebi bem antes do prazo estipulado.	2017-04-21 00:00:00	
4	Parabéns lojas lannister adorei comprar pela I...	2018-03-01 00:00:00	

	review_answer_timestamp
0	2018-01-18 21:46:59
1	2018-03-11 03:05:13
2	2018-02-18 14:36:24
3	2017-04-21 22:02:06
4	2018-03-02 10:26:53

Informações sobre Avaliações:
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99224 entries, 0 to 99223
Data columns (total 7 columns):
```

```
 #    Column                 Non-Null Count  Dtype
---   ------                 --------------  -----
 0    review_id              99224 non-null  object
 1    order_id               99224 non-null  object
 2    review_score           99224 non-null  int64
 3    review_comment_title   11568 non-null  object
 4    review_comment_message 40977 non-null  object
 5    review_creation_date   99224 non-null  object
 6    review_answer_timestamp 99224 non-null  object
dtypes: int64(1), object(6)
memory usage: 5.3+ MB
None

Estatísticas descritivas de Avaliações:
       review_score
count  99224.000000
mean       4.086421
std        1.347579
min        1.000000
25%        4.000000
50%        5.000000
75%        5.000000
max        5.000000

Valores ausentes em Avaliações:
review_id                      0
order_id                       0
review_score                   0
review_comment_title       87656
review_comment_message     58247
review_creation_date           0
review_answer_timestamp        0
dtype: int64

Explorando o dataset Pedidos:
                           order_id                        customer_id  \
0  e481f51cbdc54678b7cc49136f2d6af7  9ef432eb6251297304e76186b10a928d
1  53cdb2fc8bc7dce0b6741e2150273451  b0830fb4747a6c6d20dea0b8c802d7ef
2  47770eb9100c2d0c44946d9cf07ec65d  41ce2a54c0b03bf3443c3d931a367089
3  949d5b44dbf5de918fe9c16f97b45f8a  f88197465ea7920adcdbec7375364d82
4  ad21c59c0840e6cb83a9ceb5573f8159  8ab97904e6daea8866dbdbc4fb7aad2c

  order_status order_purchase_timestamp     order_approved_at  \
0    delivered      2017-10-02 10:56:33  2017-10-02 11:07:15
1    delivered      2018-07-24 20:41:37  2018-07-26 03:24:27
2    delivered      2018-08-08 08:38:49  2018-08-08 08:55:23
3    delivered      2017-11-18 19:28:06  2017-11-18 19:45:59
4    delivered      2018-02-13 21:18:39  2018-02-13 22:20:29

  order_delivered_carrier_date order_delivered_customer_date  \
0          2017-10-04 19:55:00           2017-10-10 21:25:13
1          2018-07-26 14:31:00           2018-08-07 15:27:45
2          2018-08-08 13:50:00           2018-08-17 18:06:29
3          2017-11-22 13:39:59           2017-12-02 00:28:42
4          2018-02-14 19:46:34           2018-02-16 18:17:02

  order_estimated_delivery_date
0           2017-10-18 00:00:00
1           2018-08-13 00:00:00
2           2018-09-04 00:00:00
```

```
3           2017-12-15 00:00:00
4           2018-02-26 00:00:00

Informações sobre Pedidos:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99441 entries, 0 to 99440
Data columns (total 8 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   order_id                      99441 non-null  object
 1   customer_id                   99441 non-null  object
 2   order_status                  99441 non-null  object
 3   order_purchase_timestamp      99441 non-null  object
 4   order_approved_at             99281 non-null  object
 5   order_delivered_carrier_date  97658 non-null  object
 6   order_delivered_customer_date 96476 non-null  object
 7   order_estimated_delivery_date 99441 non-null  object
dtypes: object(8)
memory usage: 6.1+ MB
None

Estatísticas descritivas de Pedidos:
                               order_id                       customer_id  \
count                             99441                             99441
unique                            99441                             99441
top     e481f51cbdc54678b7cc49136f2d6af7  9ef432eb6251297304e76186b10a928d
freq                                  1                                 1

       order_status order_purchase_timestamp     order_approved_at  \
count         99441                    99441                 99281
unique            8                    98875                 90733
top       delivered      2018-04-11 10:48:14   2018-02-27 04:31:10
freq          96478                        3                     9

       order_delivered_carrier_date order_delivered_customer_date  \
count                         97658                         96476
unique                        81018                         95664
top             2018-05-09 15:48:00           2018-05-08 23:38:46
freq                             47                             3

       order_estimated_delivery_date
count                          99441
unique                           459
top              2017-12-20 00:00:00
freq                             522

Valores ausentes em Pedidos:
order_id                          0
customer_id                       0
order_status                      0
order_purchase_timestamp          0
order_approved_at               160
order_delivered_carrier_date   1783
order_delivered_customer_date  2965
order_estimated_delivery_date     0
dtype: int64

Explorando o dataset Produtos:
                         product_id  product_category_name  \
0  1e9e8ef04dbcff4541ed26657ea517e5             perfumaria
```

```
1   3aa071139cb16b67ca9e5dea641aaa2f                   artes
2   96bd76ec8810374ed1b65e291975717f            esporte_lazer
3   cef67bcfe19066a932b7673e239eb23d                   bebes
4   9dc1a7de274444849c219cff195d0b71   utilidades_domesticas

    product_name_lenght  product_description_lenght  product_photos_qty  \
0                  40.0                       287.0                 1.0
1                  44.0                       276.0                 1.0
2                  46.0                       250.0                 1.0
3                  27.0                       261.0                 1.0
4                  37.0                       402.0                 4.0

    product_weight_g  product_length_cm  product_height_cm  product_width_cm
0              225.0               16.0               10.0              14.0
1             1000.0               30.0               18.0              20.0
2              154.0               18.0                9.0              15.0
3              371.0               26.0                4.0              26.0
4              625.0               20.0               17.0              13.0

Informações sobre Produtos:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32951 entries, 0 to 32950
Data columns (total 9 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   product_id                  32951 non-null  object
 1   product_category_name       32341 non-null  object
 2   product_name_lenght         32341 non-null  float64
 3   product_description_lenght  32341 non-null  float64
 4   product_photos_qty          32341 non-null  float64
 5   product_weight_g            32949 non-null  float64
 6   product_length_cm           32949 non-null  float64
 7   product_height_cm           32949 non-null  float64
 8   product_width_cm            32949 non-null  float64
dtypes: float64(7), object(2)
memory usage: 2.3+ MB
None

Estatísticas descritivas de Produtos:
       product_name_lenght  product_description_lenght  product_photos_qty  \
count         32341.000000                32341.000000        32341.000000
mean             48.476949                  771.495285            2.188986
std              10.245741                  635.115225            1.736766
min               5.000000                    4.000000            1.000000
25%              42.000000                  339.000000            1.000000
50%              51.000000                  595.000000            1.000000
75%              57.000000                  972.000000            3.000000
max              76.000000                 3992.000000           20.000000

       product_weight_g  product_length_cm  product_height_cm  \
count      32949.000000       32949.000000       32949.000000
mean        2276.472488          30.815078          16.937661
std         4282.038731          16.914458          13.637554
min            0.000000           7.000000           2.000000
25%          300.000000          18.000000           8.000000
50%          700.000000          25.000000          13.000000
75%         1900.000000          38.000000          21.000000
max        40425.000000         105.000000         105.000000

       product_width_cm
```

```
count        32949.000000
mean            23.196728
std             12.079047
min              6.000000
25%             15.000000
50%             20.000000
75%             30.000000
max            118.000000

Valores ausentes em Produtos:
product_id                   0
product_category_name      610
product_name_lenght        610
product_description_lenght  610
product_photos_qty         610
product_weight_g             2
product_length_cm            2
product_height_cm            2
product_width_cm             2
dtype: int64

Explorando o dataset Vendedores:
                          seller_id  seller_zip_code_prefix  \
0  3442f8959a84dea7ee197c632cb2df15                   13023
1  d1b65fc7debc3361ea86b5f14c68d2e2                   13844
2  ce3ad9de960102d0677a81f5d0bb7b2d                   20031
3  c0f3eea2e14555b6faeea3dd58c1b1c3                    4195
4  51a04a8a6bdcb23deccc82b0b80742cf                   12914


        seller_city seller_state
0          campinas           SP
1        mogi guacu           SP
2     rio de janeiro          RJ
3         sao paulo           SP
4  braganca paulista          SP

Informações sobre Vendedores:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3095 entries, 0 to 3094
Data columns (total 4 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   seller_id               3095 non-null   object
 1   seller_zip_code_prefix  3095 non-null   int64
 2   seller_city             3095 non-null   object
 3   seller_state            3095 non-null   object
dtypes: int64(1), object(3)
memory usage: 96.8+ KB
None

Estatísticas descritivas de Vendedores:
       seller_zip_code_prefix
count            3095.000000
mean            32291.059451
std             32713.453830
min              1001.000000
25%              7093.500000
50%             14940.000000
75%             64552.500000
max             99730.000000
```

```
Valores ausentes em Vendedores:
seller_id                0
seller_zip_code_prefix   0
seller_city              0
seller_state             0
dtype: int64

Explorando o dataset Tradução de Categorias:
     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

Informações sobre Tradução de Categorias:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 71 entries, 0 to 70
Data columns (total 2 columns):
 #   Column                         Non-Null Count  Dtype
---  ------                         --------------  -----
 0   product_category_name          71 non-null     object
 1   product_category_name_english  71 non-null     object
dtypes: object(2)
memory usage: 1.2+ KB
None

Estatísticas descritivas de Tradução de Categorias:
       product_category_name product_category_name_english
count                     71                             71
unique                    71                             71
top              beleza_saude                    health_beauty
freq                       1                              1

Valores ausentes em Tradução de Categorias:
product_category_name            0
product_category_name_english    0
dtype: int64
```
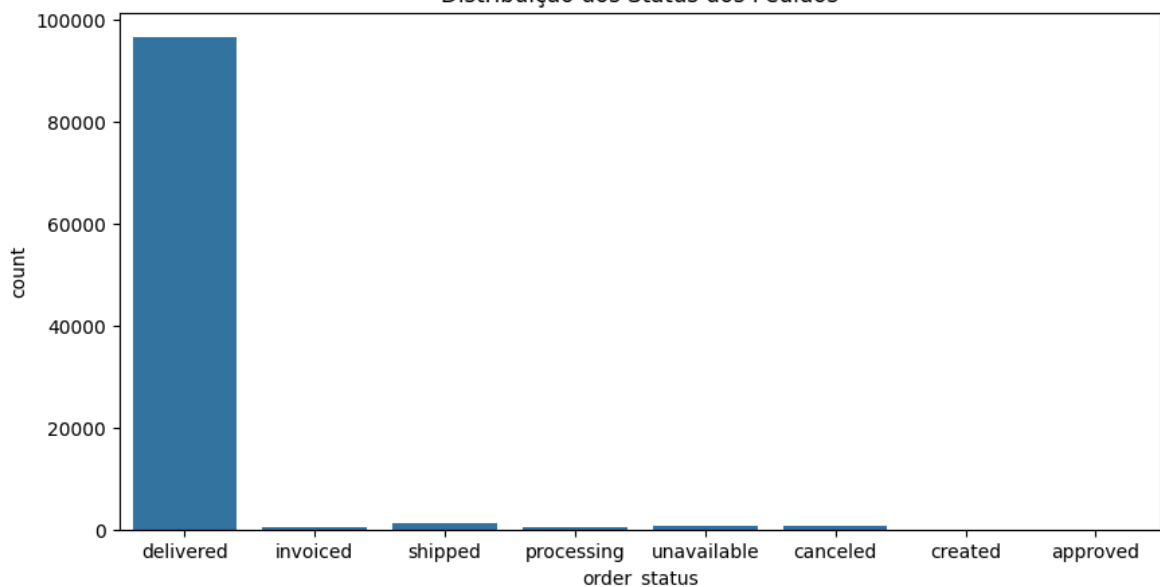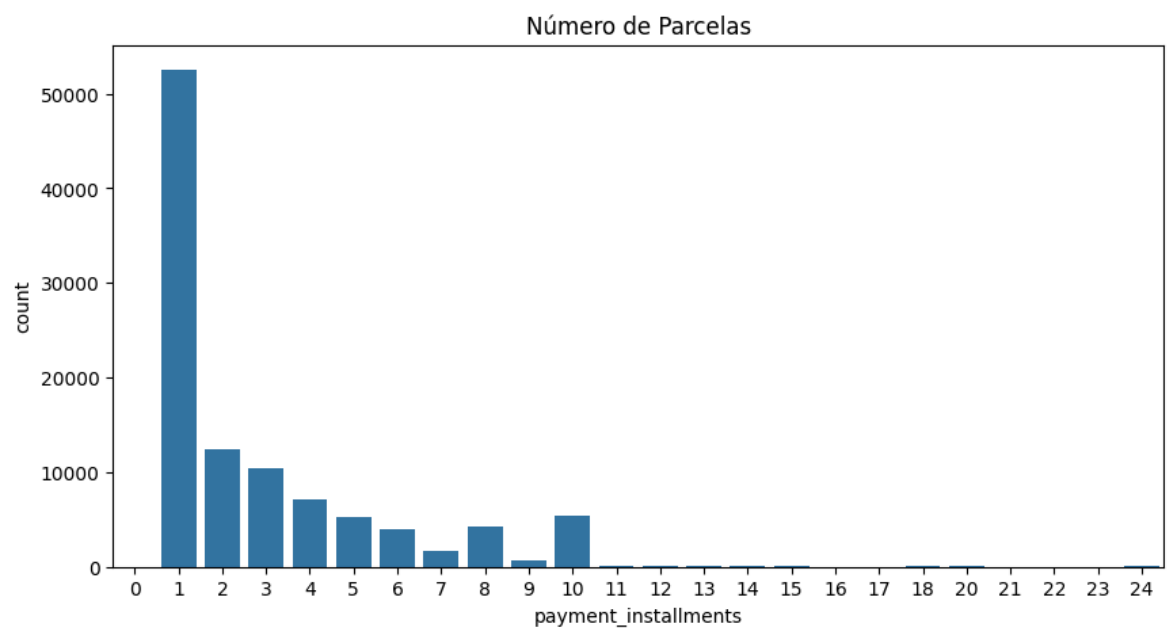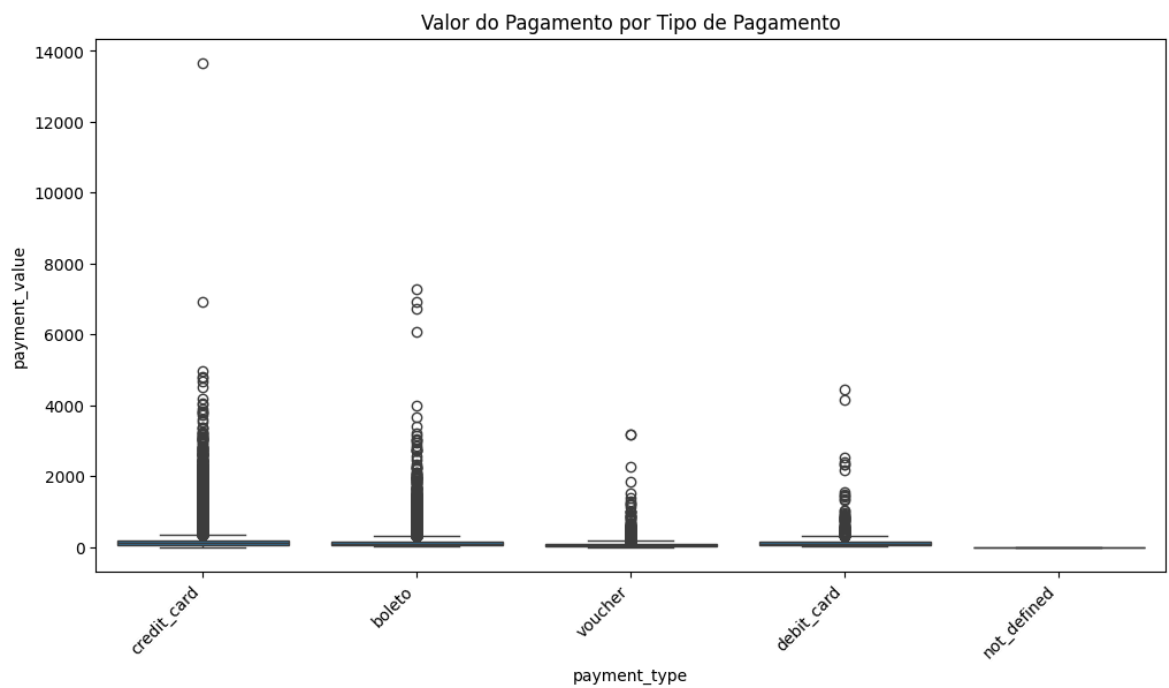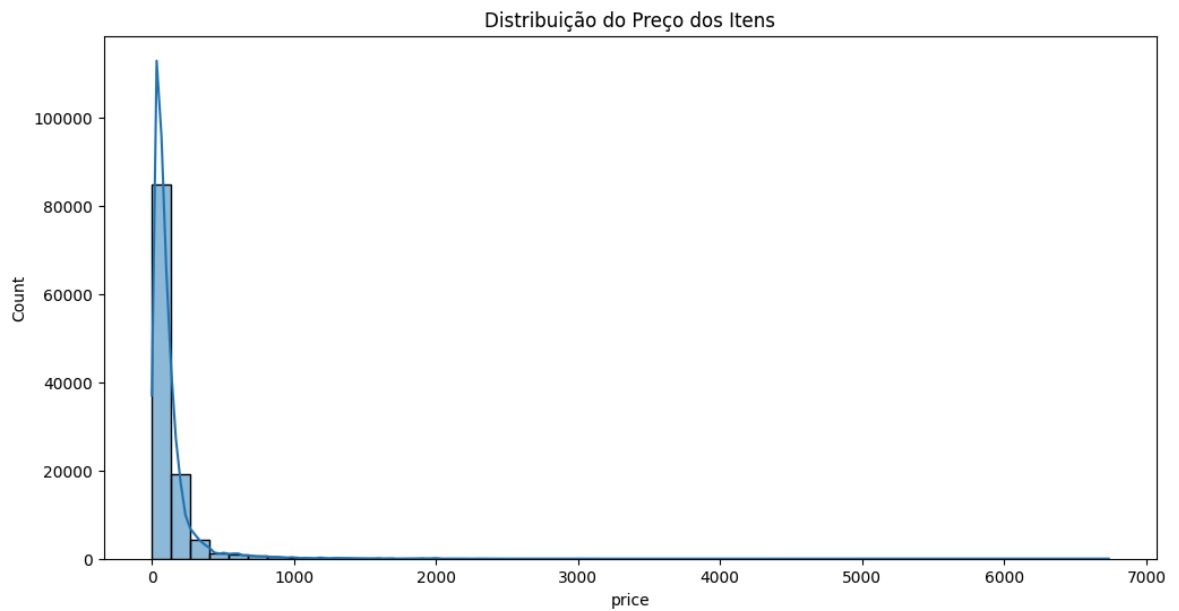

Distribuição dos Status dos Pedidos

Distribuição do Preço dos Itens

Valor do Pagamento por Tipo de Pagamento

Número de Parcelas

Fim da Exploração Inicial.