```python
import pandas as pd

# 1. Load dataset from Jupyter home
df = pd.read_csv("olist_master_dataset.csv", encoding="latin1")

# 2. Check total null values in each column
print("🔍 Null values per column:")
print(df.isnull().sum())

# 3. Check null values as percentage of total rows
print("\n📊 Null values percentage:")
print((df.isnull().sum() / len(df)) * 100)

# 4. Create a summary dataframe for a clean view
null_summary = pd.DataFrame({
    "Null Count": df.isnull().sum(),
    "Null Percent": (df.isnull().sum() / len(df)) * 100
}).sort_values(by="Null Percent", ascending=False)

# Show first 10 columns with highest nulls
null_summary.head(10)
```

```
/var/folders/cr/11cjmzwx0xn_mvwywhv63tpw0000gn/T/
ipykernel_3219/2937764459.py:4: DtypeWarning: Columns
(1,2,3,4,5,6,7,8,10,11,13,17,18,19,22,31,32,33,35,36,37,38) have mixed
types. Specify dtype option on import or set low_memory=False.
  df = pd.read_csv("olist_master_dataset.csv", encoding="latin1")
```

```
🔍 Null values per column:
order_id                            26
customer_id                      20062
order_status                     20062
order_purchase_timestamp         20062
order_approved_at                20274
order_delivered_carrier_date     22186
order_delivered_customer_date    23521
order_estimated_delivery_date    20100
customer_unique_id               20100
customer_zip_code_prefix         20100
customer_city                    20100
customer_state                   20100
payment_sequential               20103
payment_type                     20103
payment_installments             20103
payment_value                    20103
order_item_id                    20933
product_id                       20933
seller_id                        20933
shipping_limit_date              20933
price                            20933
```

```
freight_value                  20933
product_category_name          22642
product_name_lenght            22642
product_description_lenght     22642
product_photos_qty             22642
product_weight_g               20953
product_length_cm              20953
product_height_cm              20953
product_width_cm               20953
seller_zip_code_prefix         20933
seller_city                    20933
seller_state                   20933
review_id                      21097
review_score                   21097
review_comment_title          138629
review_comment_message         89001
review_creation_date           44773
review_answer_timestamp        44773
delivery_time_days             47197
dtype: int64

 Null values percentage:
order_id                       0.015963
customer_id                   12.316968
order_status                  12.316968
order_purchase_timestamp      12.316968
order_approved_at             12.447124
order_delivered_carrier_date  13.620987
order_delivered_customer_date 14.440604
order_estimated_delivery_date 12.340298
customer_unique_id            12.340298
customer_zip_code_prefix      12.340298
customer_city                 12.340298
customer_state                12.340298
payment_sequential            12.342139
payment_type                  12.342139
payment_installments          12.342139
payment_value                 12.342139
order_item_id                 12.851714
product_id                    12.851714
seller_id                     12.851714
shipping_limit_date           12.851714
price                         12.851714
freight_value                 12.851714
product_category_name         13.900946
product_name_lenght           13.900946
product_description_lenght    13.900946
product_photos_qty            13.900946
product_weight_g              12.863993
```

```
product_length_cm                 12.863993
product_height_cm                 12.863993
product_width_cm                  12.863993
seller_zip_code_prefix            12.851714
seller_city                       12.851714
seller_state                      12.851714
review_id                         12.952401
review_score                      12.952401
review_comment_title              85.110602
review_comment_message            54.641732
review_creation_date              27.488166
review_answer_timestamp           27.488166
delivery_time_days                28.976369
dtype: float64
```

	Null Count	Null Percent
review_comment_title	138629	85.110602
review_comment_message	89001	54.641732
delivery_time_days	47197	28.976369
review_answer_timestamp	44773	27.488166
review_creation_date	44773	27.488166
order_delivered_customer_date	23521	14.440604
product_photos_qty	22642	13.900946
product_name_lenght	22642	13.900946
product_description_lenght	22642	13.900946
product_category_name	22642	13.900946

```python
import pandas as pd

# Load dataset
df = pd.read_csv("olist_master_dataset.csv", encoding="latin1")

# 1. Drop rows with missing critical IDs
df = df.dropna(subset=["order_id", "customer_id", "product_id",
"seller_id"])

# 2. Fill categorical nulls with "Unknown"
categorical_cols = ["product_category_name", "seller_state",
"seller_city", "customer_city", "customer_state"]
for col in categorical_cols:
    if col in df.columns:
        df[col] = df[col].fillna("Unknown")

# 3. Fill numerical nulls with median
numerical_cols = ["price", "freight_value", "product_weight_g",
"product_length_cm",
                  "product_height_cm", "product_width_cm"]
for col in numerical_cols:
    if col in df.columns:
        df[col] = df[col].fillna(df[col].median())
```

```python
# 4. Leave review text/date nulls as-is (don’t fill)

# 5. Quick check again
print("Remaining nulls after cleaning:")
print(df.isnull().sum()[df.isnull().sum() > 0])

# Save cleaned file for Tableau
df.to_csv("olist_cleaned.csv", index=False, encoding="utf-8")
```

```
/var/folders/cr/11cjmzwx0xn_mvwywhv63tpw0000gn/T/
ipykernel_3219/278287258.py:4: DtypeWarning: Columns
(1,2,3,4,5,6,7,8,10,11,13,17,18,19,22,31,32,33,35,36,37,38) have mixed
types. Specify dtype option on import or set low_memory=False.
  df = pd.read_csv("olist_master_dataset.csv", encoding="latin1")

Remaining nulls after cleaning:
order_approved_at                    15
order_delivered_carrier_date       1254
order_delivered_customer_date      2588
payment_sequential                    3
payment_type                          3
payment_installments                  3
payment_value                         3
product_name_lenght                1709
product_description_lenght         1709
product_photos_qty                 1709
review_id                           978
review_score                        978
review_comment_title             117793
review_comment_message            68631
review_creation_date              24654
review_answer_timestamp           24654
delivery_time_days                26264
dtype: int64
```