Start coding or generate with AI.

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
```

```
import pandas as pd
```

```
df = pd.read_csv("Pakistan Largest Ecommerce Dataset.csv", on_bad_lines='skip')
```

```
<ipython-input-37-a1fcc54b23f7>:1: DtypeWarning: Columns (7) have mixed types. Specify dtype option on import or set low_memory=False.
  df = pd.read_csv("Pakistan Largest Ecommerce Dataset.csv", on_bad_lines='skip')
```

```
df1 = df.copy()
```

```
df.head()
```

	item_id	status	created_at	sku	price	qty_ordered	grand_total	increment_id	category_name_1	sales_commission_code	...	Month	Customer Since	M
0	211131	complete	7/1/2016	kreations_YI 06-L	1950.0	1.0	1950.0	100147443	Women's Fashion	\N	...	7.0	2016-7	20
1	211133	canceled	7/1/2016	kcc_Buy 2 Frey Air Freshener & Get 1 Kasual Bo...	240.0	1.0	240.0	100147444	Beauty & Grooming	\N	...	7.0	2016-7	20
2	211134	canceled	7/1/2016	Ego_UP0017-999-MR0	2450.0	1.0	2450.0	100147445	Women's Fashion	\N	...	7.0	2016-7	20
3	211135	complete	7/1/2016	kcc_krone deal	360.0	1.0	60.0	100147446	Beauty & Grooming	R-FSD-52352	...	7.0	2016-7	20
4	211136	order_refunded	7/1/2016	BK7010400AG	555.0	2.0	1110.0	100147447	Soghaat	\N	...	7.0	2016-7	20

5 rows × 26 columns

```
df.columns
```

```
Index(['item_id', 'status', 'created_at', 'sku', 'price', 'qty_ordered',
       'grand_total', 'increment_id', 'category_name_1',
       'sales_commission_code', 'discount_amount', 'payment_method',
       'Working Date', 'BI Status', ' MV ', 'Year', 'Month', 'Customer Since',
       'M-Y', 'FY', 'Customer ID', 'Unnamed: 21', 'Unnamed: 22', 'Unnamed: 23',
       'Unnamed: 24', 'Unnamed: 25'],
      dtype='object')
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 268440 entries, 0 to 268439
Data columns (total 26 columns):
 #   Column                 Non-Null Count   Dtype  
---  ------                 --------------   -----  
 0   item_id                268440 non-null  int64  
 1   status                 268440 non-null  object 
 2   created_at             268440 non-null  object 
 3   sku                    268424 non-null  object 
 4   price                  268439 non-null  float64
 5   qty_ordered            268439 non-null  float64
 6   grand_total            268439 non-null  float64
 7   increment_id           268439 non-null  object 
 8   category_name_1        268439 non-null  object 
 9   sales_commission_code  268435 non-null  object 
 10  discount_amount        268439 non-null  float64
 11  payment_method         268439 non-null  object 
 12  Working Date           268439 non-null  object 
 13  BI Status              268439 non-null  object 
 14   MV                    268439 non-null  object 
 15  Year                   268439 non-null  float64
 16  Month                  268439 non-null  float64
 17  Customer Since         268439 non-null  object 
 18  M-Y                    268439 non-null  object 
 19  FY                     268439 non-null  object 
 20  Customer ID            268439 non-null  float64
 21  Unnamed: 21            0 non-null       float64
 22  Unnamed: 22            0 non-null       float64
 23  Unnamed: 23            0 non-null       float64
 24  Unnamed: 24            0 non-null       float64
 25  Unnamed: 25            0 non-null       float64
dtypes: float64(12), int64(1), object(13)
memory usage: 53.2+ MB
```

```
df.describe()
```

	item_id	price	qty_ordered	grand_total	discount_amount	Year	Month	Customer ID	Unnamed: 21	Unnamed: 22	Unnamed: 23	Unnamed: 24
count	268440.000000	268439.000000	268439.000000	2.684390e+05	268439.000000	268439.000000	268439.000000	268439.000000	0.0	0.0	0.0	0.0
mean	378519.340322	4364.935109	1.202754	6.166206e+03	254.113456	2016.501157	7.239291	20334.877633	NaN	NaN	NaN	NaN
std	96375.647995	11903.284028	5.128284	8.616263e+04	1195.627310	0.500000	3.395365	16761.244828	NaN	NaN	NaN	NaN
min	211131.000000	0.000000	1.000000	-1.000000e+01	-599.500000	2016.000000	1.000000	1.000000	NaN	NaN	NaN	NaN
25%	294893.500000	300.000000	1.000000	5.980000e+02	0.000000	2016.000000	5.000000	5756.500000	NaN	NaN	NaN	NaN
50%	379424.500000	700.000000	1.000000	1.282500e+03	0.000000	2017.000000	7.000000	15991.000000	NaN	NaN	NaN	NaN
75%	463247.250000	1950.000000	1.000000	3.711000e+03	49.820000	2017.000000	11.000000	34281.500000	NaN	NaN	NaN	NaN
max	540990.000000	479000.000000	1000.000000	1.788800e+07	90300.000000	2017.000000	12.000000	55451.000000	NaN	NaN	NaN	NaN

```
df.dropna(axis=1, how='all', inplace=True)
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 268440 entries, 0 to 268439
Data columns (total 21 columns):
 #   Column                 Non-Null Count   Dtype  
---  ------                 --------------   -----  
 0   item_id                268440 non-null  int64  
 1   status                 268440 non-null  object 
 2   created_at             268440 non-null  object 
 3   sku                    268424 non-null  object 
 4   price                  268439 non-null  float64
 5   qty_ordered            268439 non-null  float64
 6   grand_total            268439 non-null  float64
 7   increment_id           268439 non-null  object 
 8   category_name_1        268439 non-null  object 
 9   sales_commission_code  268435 non-null  object 
 10  discount_amount        268439 non-null  float64
 11  payment_method         268439 non-null  object 
 12  Working Date           268439 non-null  object 
 13  BI Status              268439 non-null  object 
 14   MV                    268439 non-null  object 
 15  Year                   268439 non-null  float64
 16  Month                  268439 non-null  float64
 17  Customer Since         268439 non-null  object 
 18  M-Y                    268439 non-null  object 
 19  FY                     268439 non-null  object 
 20  Customer ID            268439 non-null  float64
dtypes: float64(7), int64(1), object(13)
memory usage: 43.0+ MB
```

```
df.shape
```

```
(268440, 21)
```

```
df.head()
```

	item_id	status	created_at	sku	price	qty_ordered	grand_total	increment_id	category_name_1	sales_commission_code	...	payment_method	Wor
0	211131	complete	7/1/2016	kreations_YI 06-L	1950.0	1.0	1950.0	100147443	Women's Fashion	\N	...	cod	7/1,
1	211133	canceled	7/1/2016	kcc_Buy 2 Frey Air Freshener & Get 1 Kasual Bo...	240.0	1.0	240.0	100147444	Beauty & Grooming	\N	...	cod	7/1,
2	211134	canceled	7/1/2016	Ego_UP0017-999-MR0	2450.0	1.0	2450.0	100147445	Women's Fashion	\N	...	cod	7/1,
3	211135	complete	7/1/2016	kcc_krone deal	360.0	1.0	60.0	100147446	Beauty & Grooming	R-FSD-52352	...	cod	7/1,
4	211136	order_refunded	7/1/2016	BK7010400AG	555.0	2.0	1110.0	100147447	Soghaat	\N	...	cod	7/1,

5 rows × 21 columns

```
df['item_id'].nunique()
```

```
268440
```

```
df.duplicated(['item_id']).sum()
```

```
np.int64(0)
```

```
df['status'].value_counts()
```

```
                   count
          status
        complete  134020
        canceled   80069
  order_refunded   30335
        received   19367
          refund    4112
          closed     318
            paid     192
           fraud      10
         holded        6
        exchange       4
              \N       4
  pending_paypal       3

dtype: int64
```

```
bad_values = ["\\N", "1", "3/26/2017", "Payaxis"]
df = df[~df["status"].isin(bad_values)]
```

```
df['status'].value_counts(dropna=False)
```

```
                  count
         status
       complete  134020
       canceled   80069
 order_refunded   30335
       received   19367
         refund    4112
         closed     318
           paid     192
          fraud      10
         holded       6
       exchange       4
 pending_paypal       3

dtype: int64
```

```
df["status"].isna().sum()
len(df)
```

```
268436
```

```
df['status'] = df['status'].replace('complete', 'Completed')
df['status'] = df['status'].replace('closed', 'Completed')
df['status'] = df['status'].replace('received', 'Completed')
df['status'] = df['status'].replace('paid', 'Completed')
df['status'] = df['status'].replace('cod', 'Completed')
df['status'] = df['status'].replace('order_refunded', 'Refund')
df['status'] = df['status'].replace('refund', 'Refund')
df['status'] = df['status'].replace('exchange', 'Refund')
df['status'] = df['status'].replace('pending', 'Pending')
df['status'] = df['status'].replace('payment_review', 'Pending')
df['status'] = df['status'].replace('processing', 'Pending')
df['status'] = df['status'].replace('holded', 'Pending')
df['status'] = df['status'].replace('pending_paypal', 'Pending')
df['status'] = df['status'].replace(r'\\N', 'Pending', regex=True)
df['status'] = df['status'].replace('fraud', 'Fraud')
```

```
df['status'] = df['status'].replace('canceled', 'Cancelled')
```

```
df['status'].value_counts()
```

```
                count
     status
  Completed    153897
  Cancelled     80069
     Refund     34451
      Fraud        10
    Pending         9

dtype: int64
```

```
df["status"].isna().sum()
```

```
np.int64(0)
```

```
df['status'].count()
```

```
np.int64(268436)
```

```
df['created_at'].head(10)
```

```
   created_at
0    7/1/2016
1    7/1/2016
2    7/1/2016
3    7/1/2016
4    7/1/2016
5    7/1/2016
6    7/1/2016
7    7/1/2016
8    7/1/2016
9    7/1/2016

dtype: object
```

```
print(df['created_at'].dtype)
```

```
object
```

```
df['created_at']=pd.to_datetime(df['created_at'])
```

```
df['created_at'].value_counts().head(10)
```

	count
created_at	
2016-11-25	15169
2017-05-19	11511
2016-11-23	8478
2016-11-24	8053
2016-11-19	5174
2016-11-27	5089
2016-11-26	4744
2016-11-22	4709
2017-05-20	4512
2017-05-22	3964

dtype: int64

```
print(df['created_at'])
```

```
0        2016-07-01
1        2016-07-01
2        2016-07-01
3        2016-07-01
4        2016-07-01
            ...
268435   2017-07-23
268436   2017-07-23
268437   2017-07-23
268438   2017-07-23
268439   2017-07-23
Name: created_at, Length: 268436, dtype: datetime64[ns]
```

```
print(df['created_at'].unique())
```

```
<DatetimeArray>
['2016-07-01 00:00:00', '2016-07-02 00:00:00', '2016-07-03 00:00:00',
 '2016-07-04 00:00:00', '2016-07-05 00:00:00', '2016-07-06 00:00:00',
 '2016-07-07 00:00:00', '2016-07-08 00:00:00', '2016-07-09 00:00:00',
 '2016-07-10 00:00:00',
 ...
 '2017-07-14 00:00:00', '2017-07-15 00:00:00', '2017-07-16 00:00:00',
 '2017-07-17 00:00:00', '2017-07-18 00:00:00', '2017-07-19 00:00:00',
 '2017-07-20 00:00:00', '2017-07-21 00:00:00', '2017-07-22 00:00:00',
 '2017-07-23 00:00:00']
Length: 388, dtype: datetime64[ns]
```

```
missing_values = df['created_at'].isna().sum()
print(f"number of missing value:{missing_values}")
```

```
number of missing value:0
```

```
print(df['created_at'].value_counts(dropna=False))
```

```
created_at
2016-11-25    15169
2017-05-19    11511
2016-11-23     8478
2016-11-24     8053
2016-11-19     5174
              ...  
2016-07-10       94
2016-09-14       83
2016-07-06       72
2016-09-13       52
2016-07-07       51
Name: count, Length: 388, dtype: int64
```

```
df=df.dropna(subset=['created_at'])
```

```
df['created_at'].count()
```

```
np.int64(268436)
```

```
print(df['created_at'].isna().sum())
```

```
0
```

```
print(len(df['created_at']))
```

```
268436
```

```
print(df['created_at'].nunique())
```

```
388
```

```
df.head()
```

	item_id	status	created_at	sku	price	qty_ordered	grand_total	increment_id	category_name_1	sales_commission_code	...	payment_method	Workin Dat
0	211131	Completed	2016-07-01	kreations_YI 06-L	1950.0	1.0	1950.0	100147443	Women's Fashion	\N	...	cod	7/1/201
1	211133	Cancelled	2016-07-01	kcc_Buy 2 Frey Air Freshener & Get 1 Kasual Bo...	240.0	1.0	240.0	100147444	Beauty & Grooming	\N	...	cod	7/1/201
2	211134	Cancelled	2016-07-01	Ego_UP0017-999-MR0	2450.0	1.0	2450.0	100147445	Women's Fashion	\N	...	cod	7/1/201
3	211135	Completed	2016-07-01	kcc_krone deal	360.0	1.0	60.0	100147446	Beauty & Grooming	R-FSD-52352	...	cod	7/1/201
4	211136	Refund	2016-07-01	BK7010400AG	555.0	2.0	1110.0	100147447	Soghaat	\N	...	cod	7/1/201

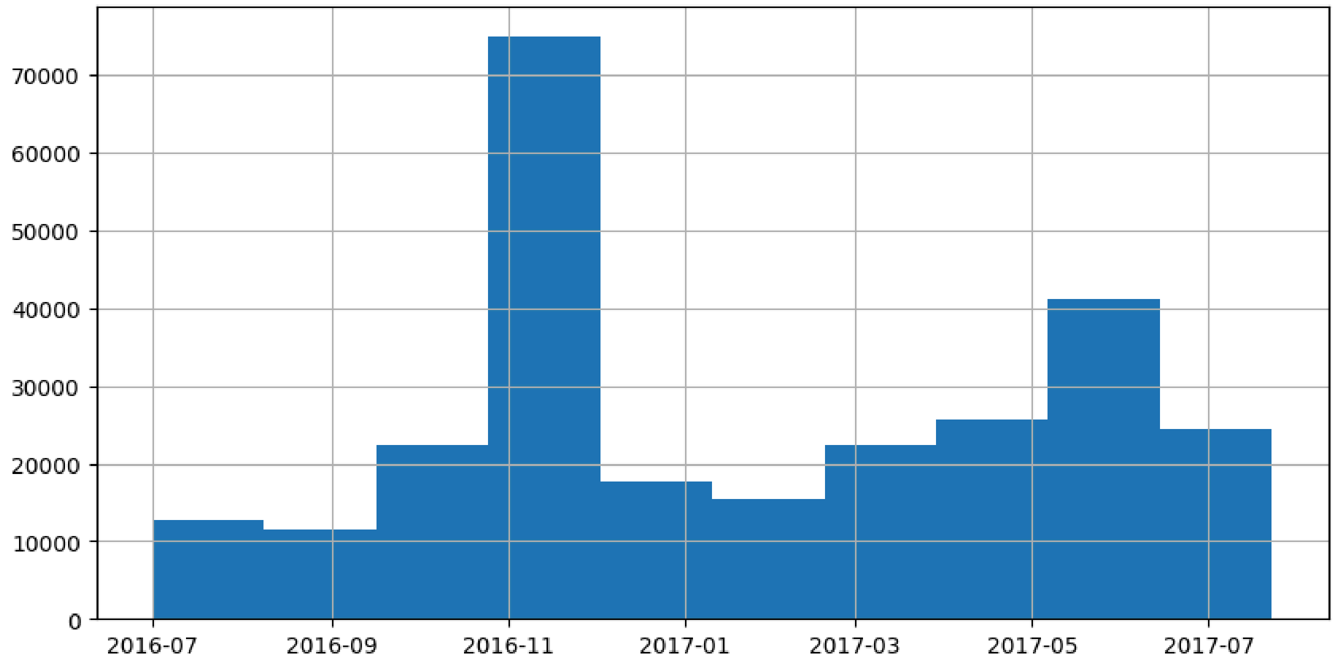
5 rows × 21 columns

```
print("first day", df['created_at'].min())
print("last day:", df['created_at'].max())
```

```
first day 2016-07-01 00:00:00
last day: 2017-07-23 00:00:00
```

```
df['created_at'].hist(figsize=(10,5))
```

```
<Axes: >
```



```
df['sku'].count()
```

```
np.int64(268420)
```

```
df['sku'].isna().sum()
```

```
np.int64(16)
```

```
df['sku'].nunique()
```

```
35066
```

```
print(df['sku'].unique())
```

```
['kreations_YI 06-L'
 'kcc_Buy 2 Frey Air Freshener & Get 1 Kasual Body Spray Free'
 'Ego_UP0017-999-MR0' ... 'BP_T049464-3' 'LS_8961014035926'
 'Mardaz_MA305FA0W5788NAFAMZ-']
```

```
df['sku'].value_counts()
```

	count
sku	
Al Muhafiz Sohan Halwa Almond	2241
emart_00-7	2023
kcc_krone deal	1894
infinix_Zero 4-Grey	1774
emart_00-1	1382
...	...
Echange-EP_01-42	1
Xarasoft_PES2201-BLACK-39	1
memsaab_36-C-off white-Free Size	1
SFEVER_HU379HB0845HGNAFAMZ	1
MD-DZ-09 Rose Gold	1

35066 rows × 1 columns

dtype: int64

```
print(df[df['sku'].isna()])
```

```
        item_id     status created_at  sku  price  qty_ordered  grand_total  \
14846    230008  Cancelled 2016-08-13  NaN    0.0          1.0          0.0   
20676    236830  Cancelled 2016-09-01  NaN    0.0          1.0          0.0   
39838    260006  Cancelled 2016-10-07  NaN    0.0          1.0          0.0   
39839    260007  Cancelled 2016-10-07  NaN    0.0          3.0          0.0   
39880    260061  Cancelled 2016-10-07  NaN    0.0          3.0          0.0   
124968   367292     Refund 2016-12-10  NaN    0.0          1.0          0.0   
125636   368122     Refund 2016-12-12  NaN    0.0          1.0          0.0   
125811   368362     Refund 2016-12-13  NaN    0.0          1.0          0.0   
149597   399798     Refund 2017-02-07  NaN    0.0          1.0          0.0   
170249   426105  Cancelled 2017-03-21  NaN    0.0          1.0       6952.0   
170401   426302  Cancelled 2017-03-22  NaN    0.0          1.0       1873.0   
173040   429392     Refund 2017-03-23  NaN    0.0          1.0          0.0   
173045   429393     Refund 2017-03-23  NaN    0.0          1.0          0.0   
173069   429421     Refund 2017-03-23  NaN    0.0          1.0          0.0   
175411   432276     Refund 2017-03-26  NaN    0.0          1.0          0.0   
238273   506324  Cancelled 2017-06-07  NaN    0.0          1.0          0.0   
```

```
          increment_id category_name_1 sales_commission_code  ...  \
14846        100160070              \N                    \N  ...   
20676        100164902              \N                    \N  ...   
39838        100181136              \N                    \N  ...   
39839        100181137              \N                    \N  ...   
39880        100181174              \N                    \N  ...   
124968       100247863              \N                    \N  ...   
125636       100248364              \N                    \N  ...   
125811       100248497              \N                    \N  ...   
149597       100267148              \N                    \N  ...   
170249       100281645              \N                    \N  ...   
170401       100281748              \N                    \N  ...   
173040       100283785              \N                    \N  ...   
173045       100283786              \N                    \N  ...   
173069       100283807              \N                    \N  ...   
175411       100285563              \N                    \N  ...   
238273       100322823              \N                    \N  ...   

       payment_method Working Date BI Status    MV     Year  Month  \
14846             cod    8/13/2016     Gross   -     2016.0    8.0   
20676             cod     9/1/2016     Gross   -     2016.0    9.0   
39838             cod    10/7/2016     Gross   -     2016.0   10.0   
39839             cod    10/7/2016     Gross   -     2016.0   10.0   
39880             cod    10/7/2016     Gross   -     2016.0   10.0   
124968            cod   12/10/2016     Valid   -     2016.0   12.0   
125636            cod   12/12/2016     Valid   -     2016.0   12.0   
125811            cod   12/13/2016     Valid   -     2016.0   12.0   
149597            cod     2/7/2017     Valid   -     2017.0    2.0   
170249            cod    3/21/2017     Gross   -     2017.0    3.0   
170401            cod    3/22/2017     Gross   -     2017.0    3.0   
173040            cod    3/23/2017     Valid   -     2017.0    3.0   
173045            cod    3/23/2017     Valid   -     2017.0    3.0   
173069            cod    3/23/2017     Valid   -     2017.0    3.0   
175411            cod    3/26/2017     Valid   -     2017.0    3.0   
238273            cod     6/7/2017     Gross   -     2017.0    6.0   

       Customer Since      M-Y    FY Customer ID  
14846          2016-8   8-2016  FY17      3468.0  
20676          2016-8   9-2016  FY17      4369.0  
39838          2016-7  10-2016  FY17       939.0  
```

```
df = df.dropna(subset=['sku'])
df = df.reset_index(drop=True)
```

```
print(df['sku'].isna().sum())
```

```
0
```

```
df['price'].value_counts()
```

```
           count
   price
   999.0    5306
   399.0    4385
 12599.0    3657
   799.0    3462
   499.0    3353
     ...     ...
  4475.0       1
 11280.0       1
  2272.0       1
  1520.5       1
  1784.0       1
4192 rows × 1 columns

dtype: int64
```

```
df['price'].isna().sum()
```

```
np.int64(1)
```

```
df['price'].value_counts().sort_index()
```

```
               count
     price
      0.00       582
      0.15         1
      1.00       698
      2.00       349
      2.64         1
       ...       ...
 265499.00         1
 289999.00         1
 300000.00         4
 330499.00         2
 479000.00         4

4192 rows × 1 columns

dtype: int64
```

```
print(df['price'].head(20))
```

```
0      1950.00
1       240.00
2      2450.00
3       360.00
4       555.00
5        80.00
6       360.00
7       170.00
8     96499.00
9     96499.00
10     5500.00
11      210.00
12      156.00
13      120.00
14      320.00
15     1550.00
16      420.00
17      360.00
18      490.00
```

```
19      899.25
Name: price, dtype: float64
```

```
print(df['grand_total'].isnull().sum())
```

```
1
```

```
df['new_grand_total'] = df['price'] * df['qty_ordered'] - df['discount_amount']
```

```
df['new_grand_total'].value_counts()
```

```
                   count
new_grand_total
         999.00     3800
         399.00     3785
         599.00     2698
       12599.00     2691
         799.00     2544
            ...      ...
         176.80        1
          13.60        1
         267.83        1
         744.30        1
        2137.37        1
19919 rows × 1 columns

dtype: int64
```

```
df['new_price']=df['new_grand_total']/df['qty_ordered']
```

```
df['new_price'].value_counts()
```

```
            count
new_price
   999.00    4035
   399.00    4012
 12599.00    2873
   599.00    2827
   799.00    2624
      ...     ...
   210.60       1
 21052.20       1
    15.23       1
  5785.00       1
    40.09       1
19218 rows × 1 columns

dtype: int64
```

```
df[df['new_price'] < 0]
```

	item_id	status	created_at	sku	price	qty_ordered	grand_total	increment_id	category_name_1	sales_commission_code	...	BI Status	MV	
155298	407448	Completed	2017-02-20	Nimcos_Mix-Nimco-200gm	110.0	1.0	604.25	100271403	Soghaat	\N	...	Net	110	2
155502	407718	Completed	2017-02-21	Aladdin_Hand Grip Pair - Black	269.0	1.0	4321.40	100271572	Health & Sports	HDD105640	...	Net	269	2
155639	407892	Completed	2017-02-21	sg_KajalPencil0.36g	100.0	1.0	683.29	100271675	Beauty & Grooming	\N	...	Net	100	2
158359	411348	Completed	2017-02-27	tram_TT23080083	55.0	2.0	637.40	100273658	Home & Living	\N	...	Net	110	2
159966	413387	Completed	2017-03-02	swi_LTLP	10.0	1.0	604.25	100274724	School & Education	C-PEW-31067	...	Net	10	2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
241730	510278	Cancelled	2017-06-11	LS_028400019903	55.0	1.0	124.29	100324476	Superstore	\N	...	Gross	55	2
252309	522917	Completed	2017-06-22	GMZV_White-Fidget-Spinner	145.0	1.0	141.25	100330103	Kids & Baby	\N	...	Net	145	2
252310	522918	Completed	2017-06-22	LS_5053990107278	80.0	1.0	141.25	100330103	Superstore	\N	...	Net	80	2
262276	534192	Cancelled	2017-07-13	RUB_Rubian Zipper	155.0	1.0	138.00	100335702.0	Mobiles & Tablets	\N	...	Gross	155	2
262277	534193	Cancelled	2017-07-13	BT_BT-263	125.0	1.0	138.00	100335702.0	Mobiles & Tablets	\N	...	Gross	125	2

87 rows × 23 columns

```
df['discount_amount'] = df['discount_amount'].abs()
```

```
print(df['discount_amount']<0)
```

```
0         False
1         False
2         False
3         False
4         False
          ...
268415    False
268416    False
```

```
268417    False
268418    False
268419    False
Name: discount_amount, Length: 268420, dtype: bool
```

```
print(df['new_grand_total']<0)
```

```
0         False
1         False
2         False
3         False
4         False
          ...  
268415    False
268416    False
268417    False
268418    False
268419    False
Name: new_grand_total, Length: 268420, dtype: bool
```

```
print(df['new_price']<0)
```

```
0         False
1         False
2         False
3         False
4         False
          ...  
268415    False
268416    False
268417    False
268418    False
268419    False
Name: new_price, Length: 268420, dtype: bool
```

```
df['new_price'].isna().sum()
```

```
np.int64(1)
```

```
df['new_grand_total'].isna().sum()
```

```
np.int64(1)
```

```
df= df.drop(['price', 'grand_total'],axis=1)
```

```
df.head()
```

	item_id	status	created_at	sku	qty_ordered	increment_id	category_name_1	sales_commission_code	discount_amount	payment_method	...	BI Status	
0	211131	Completed	2016-07-01	kreations_YI 06-L	1.0	100147443	Women's Fashion	\N	0.0	cod	...	#REF!	1,9
1	211133	Cancelled	2016-07-01	kcc_Buy 2 Frey Air Freshener & Get 1 Kasual Bo...	1.0	100147444	Beauty & Grooming	\N	0.0	cod	...	Gross	2
2	211134	Cancelled	2016-07-01	Ego_UP0017-999-MR0	1.0	100147445	Women's Fashion	\N	0.0	cod	...	Gross	2,4
3	211135	Completed	2016-07-01	kcc_krone deal	1.0	100147446	Beauty & Grooming	R-FSD-52352	300.0	cod	...	Net	3
4	211136	Refund	2016-07-01	BK7010400AG	2.0	100147447	Soghaat	\N	0.0	cod	...	Valid	1,

5 rows × 21 columns

```
columns = list(df.columns)

columns_to_move = ['new_price', 'qty_ordered', 'new_grand_total']

for col in columns_to_move:
    columns.remove(col)

columns = columns[:4] + columns_to_move + columns[4:]

df = df[columns]
```

```
df.head()
```

	item_id	status	created_at	sku	new_price	qty_ordered	new_grand_total	increment_id	category_name_1	sales_commission_code	...	payment_method
0	211131	Completed	2016-07-01	kreations_YI 06-L	1950.0	1.0	1950.0	100147443	Women's Fashion	\N	...	cod
1	211133	Cancelled	2016-07-01	kcc_Buy 2 Frey Air Freshener & Get 1 Kasual Bo...	240.0	1.0	240.0	100147444	Beauty & Grooming	\N	...	cod
2	211134	Cancelled	2016-07-01	Ego_UP0017-999-MR0	2450.0	1.0	2450.0	100147445	Women's Fashion	\N	...	cod
3	211135	Completed	2016-07-01	kcc_krone deal	60.0	1.0	60.0	100147446	Beauty & Grooming	R-FSD-52352	...	cod
4	211136	Refund	2016-07-01	BK7010400AG	555.0	2.0	1110.0	100147447	Soghaat	\N	...	cod

5 rows × 21 columns

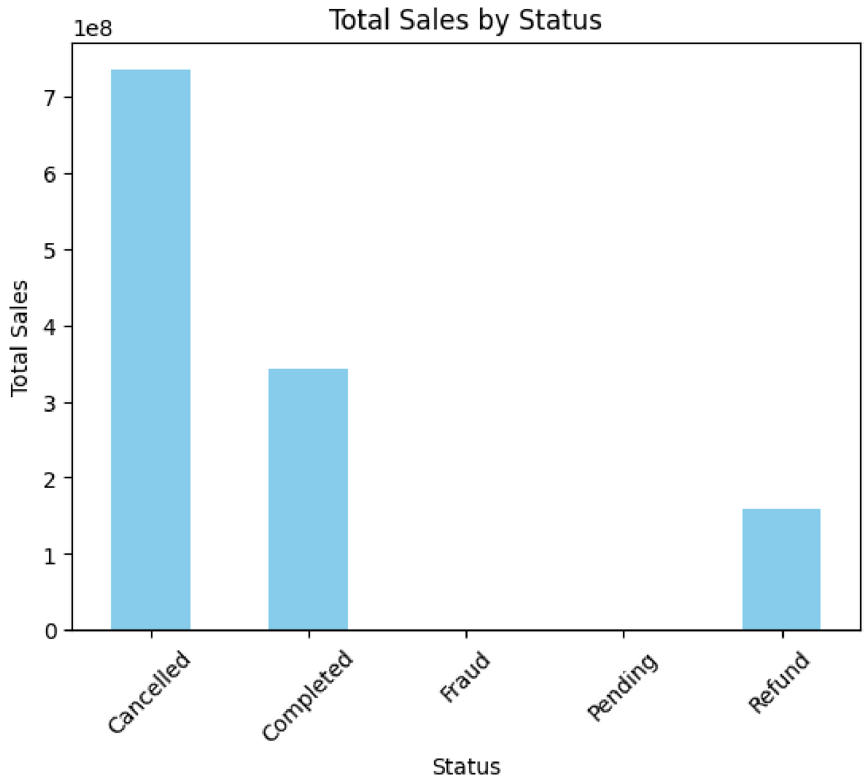
```
cols = list(df.columns)
cols.insert(6, cols.pop(cols.index('discount_amount')))  # מוסיף את discount_amount לעמודה ה-7 (אינדקס 6)
df = df[cols]
```

```
sales_by_status = df.groupby('status')['new_grand_total'].sum()
print(sales_by_status)
```

```
status
Cancelled    7.351451e+08
Completed    3.422965e+08
Fraud        6.269440e+05
Pending      6.510500e+03
Refund       1.598949e+08
Name: new_grand_total, dtype: float64
```

```
import matplotlib.pyplot as plt
```

```
sales_by_status.plot(kind='bar', color='skyblue')
plt.title('Total Sales by Status')
plt.xlabel('Status')
plt.ylabel('Total Sales')
plt.xticks(rotation=45)
plt.show()
```



```
df.head()
```

	item_id	status	created_at	sku	new_price	qty_ordered	discount_amount	new_grand_total	increment_id	category_name_1	...	payment_method	Worki Da
0	211131	Completed	2016-07-01	kreations_YI 06-L	1950.0	1.0	0.0	1950.0	100147443	Women's Fashion	...	cod	7/1/20
1	211133	Cancelled	2016-07-01	kcc_Buy 2 Frey Air Freshener & Get 1 Kasual Bo...	240.0	1.0	0.0	240.0	100147444	Beauty & Grooming	...	cod	7/1/20
2	211134	Cancelled	2016-07-01	Ego_UP0017-999-MR0	2450.0	1.0	0.0	2450.0	100147445	Women's Fashion	...	cod	7/1/20
3	211135	Completed	2016-07-01	kcc_krone deal	60.0	1.0	300.0	60.0	100147446	Beauty & Grooming	...	cod	7/1/20
4	211136	Refund	2016-07-01	BK7010400AG	555.0	2.0	0.0	1110.0	100147447	Soghaat	...	cod	7/1/20

5 rows × 21 columns

```
df['category_name_1'].value_counts()
```

	count
category_name_1	
Mobiles & Tablets	49196
Men's Fashion	47512
Women's Fashion	27667
Soghaat	25391
Beauty & Grooming	21736
Superstore	21429
Appliances	17850
Home & Living	11580
Kids & Baby	9183
Health & Sports	8437
Entertainment	7825
\N	7445
Computing	7056
Others	2977
School & Education	2156
Books	979

dtype: int64

```
df['category_name_1'].isna().sum()
```

```
np.int64(1)
```

```
df['category_name_1'] = df['category_name_1'].fillna(df['sku'])
```

```
df['category_name_1'].isna().sum()
```

```
np.int64(0)
```

```
df['category_name_1'].head(20)
```

```
     category_name_1
 0   Women's Fashion
 1   Beauty & Grooming
 2   Women's Fashion
 3   Beauty & Grooming
 4   Soghaat
 5   Soghaat
 6   Beauty & Grooming
 7   Soghaat
 8   Mobiles & Tablets
 9   Mobiles & Tablets
10   Appliances
11   Soghaat
12   Soghaat
13   Home & Living
14   Beauty & Grooming
15   Men's Fashion
16   Soghaat
17   Soghaat
18   Beauty & Grooming
19   Home & Living

dtype: object
```

```
df['category_name_1'].tail(20)
```

	category_name_1
268400	Men's Fashion
268401	Beauty & Grooming
268402	Women's Fashion
268403	Superstore
268404	Women's Fashion
268405	Mobiles & Tablets
268406	Mobiles & Tablets
268407	Kids & Baby
268408	Entertainment
268409	Women's Fashion
268410	Mobiles & Tablets
268411	Superstore
268412	Beauty & Grooming
268413	Beauty & Grooming
268414	Beauty & Grooming
268415	Women's Fashion
268416	Women's Fashion
268417	Men's Fashion
268418	Mobiles & Tablets
268419	Mardaz_MA305FA0W5788NAFAMZ-

dtype: object

```
print(df['Working Date'].dtype)
```

```
object
```

```
df['Working Date'].isna().sum()
```

```
np.int64(1)
```

```
df['Working Date'] = pd.to_datetime(df['Working Date'], errors='coerce')
df['Working Date'] = df['Working Date'].dt.strftime('%d/%m/%Y')
```

```
print(df['Working Date'].head(20))
```

```
0     01/07/2016
1     01/07/2016
2     01/07/2016
3     01/07/2016
4     01/07/2016
5     01/07/2016
6     01/07/2016
7     01/07/2016
8     01/07/2016
9     01/07/2016
10    01/07/2016
11    01/07/2016
12    01/07/2016
13    01/07/2016
14    01/07/2016
15    01/07/2016
16    01/07/2016
17    01/07/2016
18    01/07/2016
19    01/07/2016
Name: Working Date, dtype: object
```

```
print(df['Working Date'].dtype)
```

```
object
```

```
print(df['Working Date'].isnull().sum())
```

```
1
```

```
df['Working Date'] = pd.to_datetime(df['Working Date'], dayfirst=True, errors='coerce')
```

```
print(df['Working Date'].head())
print(df['Working Date'].dtype)
```

```
0   2016-07-01
1   2016-07-01
2   2016-07-01
3   2016-07-01
4   2016-07-01
Name: Working Date, dtype: datetime64[ns]
datetime64[ns]
```

```
print(df['payment_method'].unique())
```

```
['cod' 'ublcreditcard' 'mygateway' 'customercredit' 'cashatdoorstep'
 'mcblite' 'internetbanking' 'marketingexpense' 'productcredit'
 'financesettlement' 'Payaxis' 'jazzvoucher' 'jazzwallet' 'Easypay'
 'Easypay_MA' 'easypay_voucher' nan]
```

```
print(df['payment_method'].value_counts())
```

```
payment_method
cod                  153384
Payaxis               57804
jazzwallet            25651
Easypay               11662
jazzvoucher            9460
customercredit         3374
Easypay_MA             2340
easypay_voucher        1299
ublcreditcard           882
```