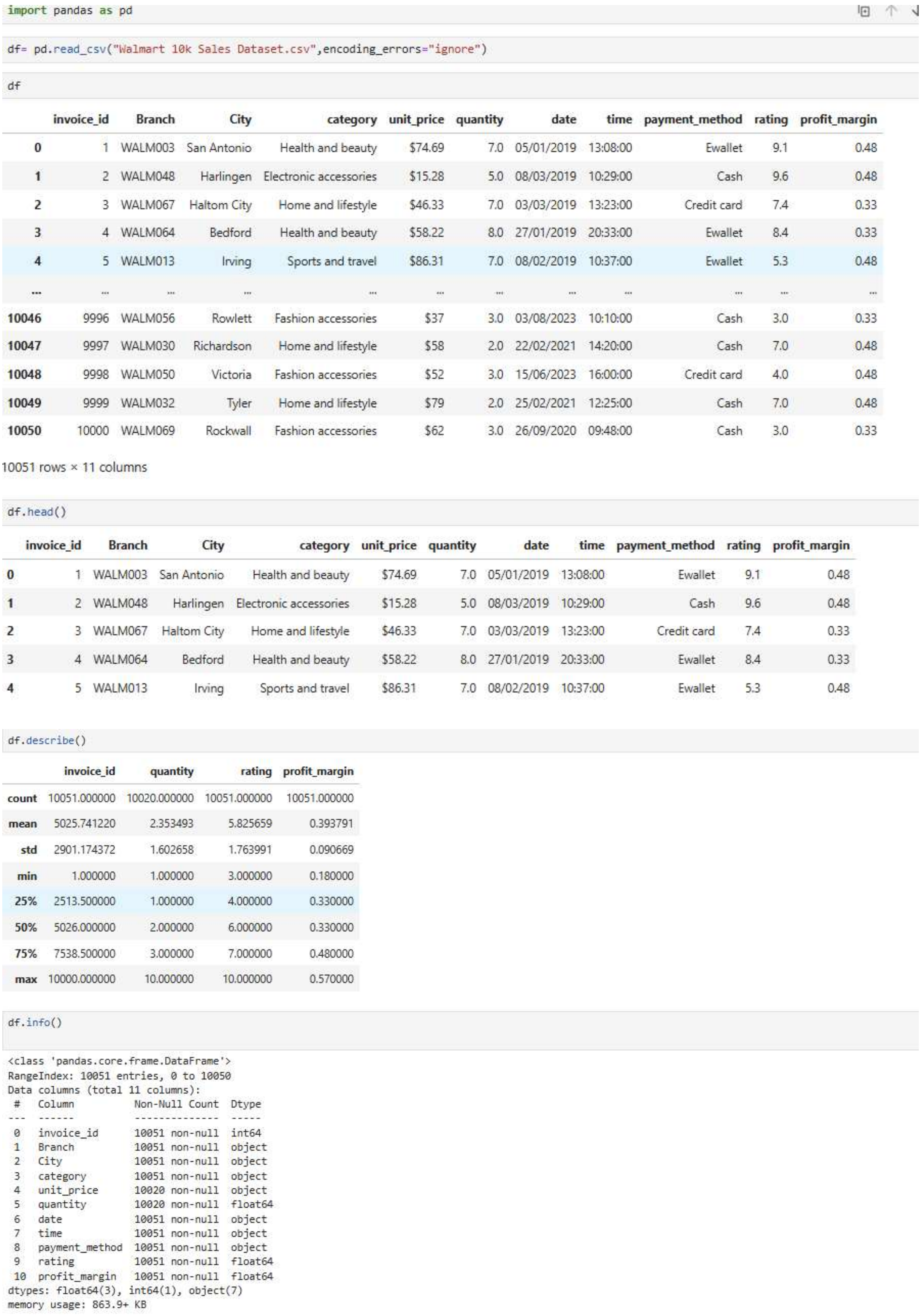


```
import pandas as pd
```

```
df= pd.read_csv("Walmart 10k Sales Dataset.csv",encoding_errors="ignore")
```

```
df
```

	invoice_id	Branch	City	category	unit_price	quantity	date	time	payment_method	rating	profit_margin
0	1	WALM003	San Antonio	Health and beauty	$74.69	7.0	05/01/2019	13:08:00	Ewallet	9.1	0.48
1	2	WALM048	Harlingen	Electronic accessories	$15.28	5.0	08/03/2019	10:29:00	Cash	9.6	0.48
2	3	WALM067	Haltom City	Home and lifestyle	$46.33	7.0	03/03/2019	13:23:00	Credit card	7.4	0.33
3	4	WALM064	Bedford	Health and beauty	$58.22	8.0	27/01/2019	20:33:00	Ewallet	8.4	0.33
4	5	WALM013	Irving	Sports and travel	$86.31	7.0	08/02/2019	10:37:00	Ewallet	5.3	0.48
...	...	...	...	...	...	...	...	...	...	...	...
10046	9996	WALM056	Rowlett	Fashion accessories	$37	3.0	03/08/2023	10:10:00	Cash	3.0	0.33
10047	9997	WALM030	Richardson	Home and lifestyle	$58	2.0	22/02/2021	14:20:00	Cash	7.0	0.48
10048	9998	WALM050	Victoria	Fashion accessories	$52	3.0	15/06/2023	16:00:00	Credit card	4.0	0.48
10049	9999	WALM032	Tyler	Home and lifestyle	$79	2.0	25/02/2021	12:25:00	Cash	7.0	0.48
10050	10000	WALM069	Rockwall	Fashion accessories	$62	3.0	26/09/2020	09:48:00	Cash	3.0	0.33

10051 rows × 11 columns

```
df.head()
```

	invoice_id	Branch	City	category	unit_price	quantity	date	time	payment_method	rating	profit_margin
0	1	WALM003	San Antonio	Health and beauty	$74.69	7.0	05/01/2019	13:08:00	Ewallet	9.1	0.48
1	2	WALM048	Harlingen	Electronic accessories	$15.28	5.0	08/03/2019	10:29:00	Cash	9.6	0.48
2	3	WALM067	Haltom City	Home and lifestyle	$46.33	7.0	03/03/2019	13:23:00	Credit card	7.4	0.33
3	4	WALM064	Bedford	Health and beauty	$58.22	8.0	27/01/2019	20:33:00	Ewallet	8.4	0.33
4	5	WALM013	Irving	Sports and travel	$86.31	7.0	08/02/2019	10:37:00	Ewallet	5.3	0.48

```
df.describe()
```

	invoice_id	quantity	rating	profit_margin
count	10051.000000	10020.000000	10051.000000	10051.000000
mean	5025.741220	2.353493	5.825659	0.393791
std	2901.174372	1.602658	1.763991	0.090669
min	1.000000	1.000000	3.000000	0.180000
25%	2513.500000	1.000000	4.000000	0.330000
50%	5026.000000	2.000000	6.000000	0.330000
75%	7538.500000	3.000000	7.000000	0.480000
max	10000.000000	10.000000	10.000000	0.570000

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10051 entries, 0 to 10050
Data columns (total 11 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   invoice_id      10051 non-null  int64  
 1   Branch          10051 non-null  object 
 2   City            10051 non-null  object 
 3   category        10051 non-null  object 
 4   unit_price      10020 non-null  object 
 5   quantity        10020 non-null  float64
 6   date            10051 non-null  object 
 7   time            10051 non-null  object 
 8   payment_method  10051 non-null  object 
 9   rating          10051 non-null  float64
 10  profit_margin   10051 non-null  float64
dtypes: float64(3), int64(1), object(7)
memory usage: 863.9+ KB
```

```
df.duplicated()
```

```
0        False
1        False
2        False
3        False
4        False
         ...  
10046     True
10047     True
10048     True
10049     True
10050     True
Length: 10051, dtype: bool
```

```
df.duplicated().sum()
```

```
np.int64(51)
```

```
df.isnull()
```

	invoice_id	Branch	City	category	unit_price	quantity	date	time	payment_method	rating	profit_margin
0	False	False	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...
10046	False	False	False	False	False	False	False	False	False	False	False
10047	False	False	False	False	False	False	False	False	False	False	False
10048	False	False	False	False	False	False	False	False	False	False	False
10049	False	False	False	False	False	False	False	False	False	False	False
10050	False	False	False	False	False	False	False	False	False	False	False

10051 rows × 11 columns

```
df.isnull().sum()
```

```
invoice_id         0
Branch             0
City               0
category           0
unit_price        31
quantity          31
date               0
time               0
payment_method     0
rating             0
profit_margin      0
dtype: int64
```

```
df.drop_duplicates(inplace=True)
```

```
df
```

	invoice_id	Branch	City	category	unit_price	quantity	date	time	payment_method	rating	profit_margin
0	1	WALM003	San Antonio	Health and beauty	$74.69	7.0	05/01/2019	13:08:00	Ewallet	9.1	0.48
1	2	WALM048	Harlingen	Electronic accessories	$15.28	5.0	08/03/2019	10:29:00	Cash	9.6	0.48
2	3	WALM067	Haltom City	Home and lifestyle	$46.33	7.0	03/03/2019	13:23:00	Credit card	7.4	0.33
3	4	WALM064	Bedford	Health and beauty	$58.22	8.0	27/01/2019	20:33:00	Ewallet	8.4	0.33
4	5	WALM013	Irving	Sports and travel	$86.31	7.0	08/02/2019	10:37:00	Ewallet	5.3	0.48
...	...	...	...	...	...	...	...	...	...	...	...
9995	9996	WALM056	Rowlett	Fashion accessories	$37	3.0	03/08/2023	10:10:00	Cash	3.0	0.33
9996	9997	WALM030	Richardson	Home and lifestyle	$58	2.0	22/02/2021	14:20:00	Cash	7.0	0.48
9997	9998	WALM050	Victoria	Fashion accessories	$52	3.0	15/06/2023	16:00:00	Credit card	4.0	0.48
9998	9999	WALM032	Tyler	Home and lifestyle	$79	2.0	25/02/2021	12:25:00	Cash	7.0	0.48
9999	10000	WALM069	Rockwall	Fashion accessories	$62	3.0	26/09/2020	09:48:00	Cash	3.0	0.33

10000 rows × 11 columns

```
df.shape
```

```
(10000, 11)
```

```
df.duplicated().sum()
```

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

```
df.dropna()
```

	invoice_id	Branch	City	category	unit_price	quantity	date	time	payment_method	rating	profit_margin
0	1	WALM003	San Antonio	Health and beauty	$74.69	7.0	05/01/2019	13:08:00	Ewallet	9.1	0.48
1	2	WALM048	Harlingen	Electronic accessories	$15.28	5.0	08/03/2019	10:29:00	Cash	9.6	0.48
2	3	WALM067	Haltom City	Home and lifestyle	$46.33	7.0	03/03/2019	13:23:00	Credit card	7.4	0.33
3	4	WALM064	Bedford	Health and beauty	$58.22	8.0	27/01/2019	20:33:00	Ewallet	8.4	0.33
4	5	WALM013	Irving	Sports and travel	$86.31	7.0	08/02/2019	10:37:00	Ewallet	5.3	0.48
...	...	...	...	...	...	...	...	...	...	...	...
9995	9996	WALM056	Rowlett	Fashion accessories	$37	3.0	03/08/2023	10:10:00	Cash	3.0	0.33
9996	9997	WALM030	Richardson	Home and lifestyle	$58	2.0	22/02/2021	14:20:00	Cash	7.0	0.48
9997	9998	WALM050	Victoria	Fashion accessories	$52	3.0	15/06/2023	16:00:00	Credit card	4.0	0.48
9998	9999	WALM032	Tyler	Home and lifestyle	$79	2.0	25/02/2021	12:25:00	Cash	7.0	0.48
9999	10000	WALM069	Rockwall	Fashion accessories	$62	3.0	26/09/2020	09:48:00	Cash	3.0	0.33

9969 rows × 11 columns

```
df.shape
```

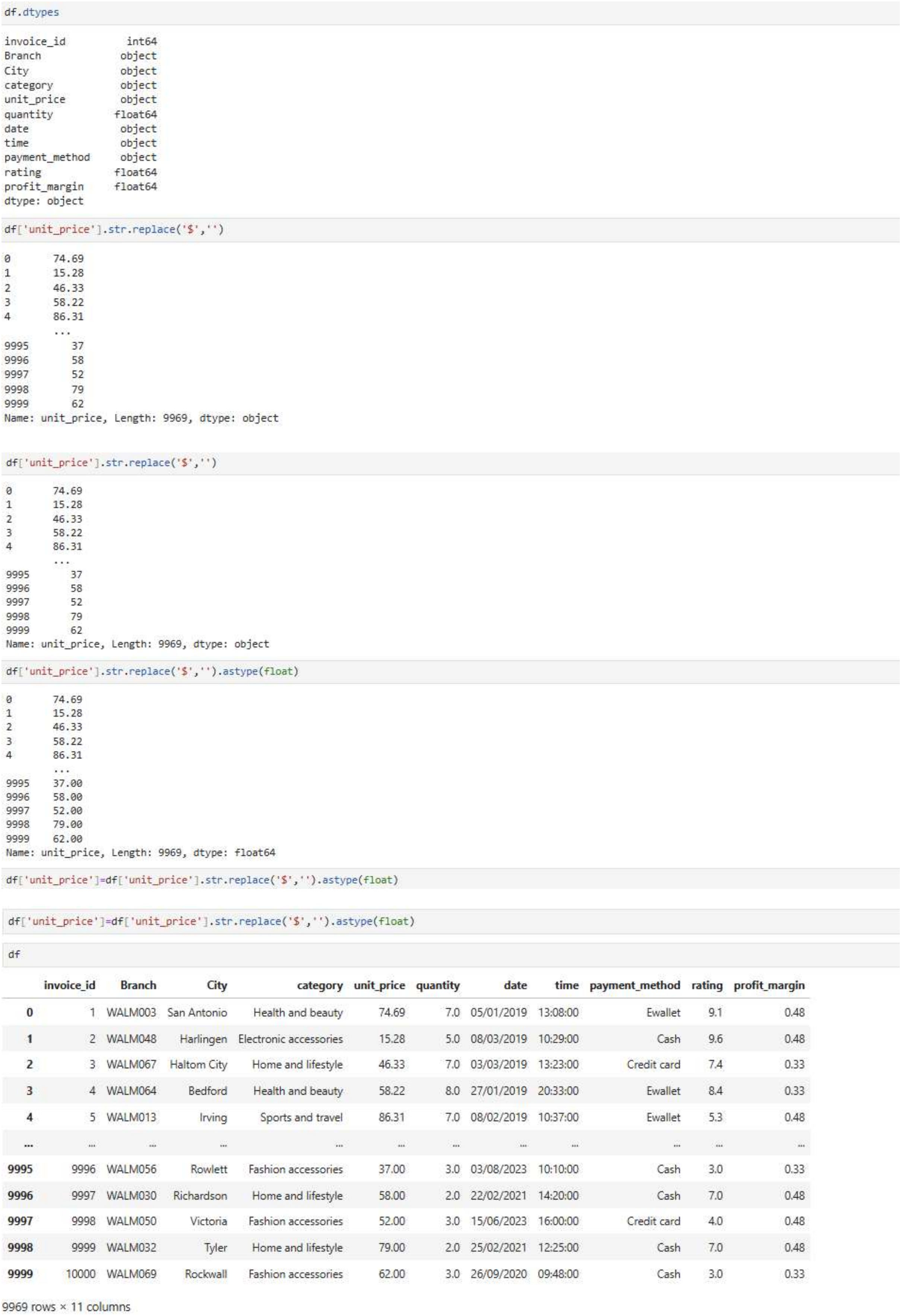
```
(10000, 11)
```

```
df.dropna(inplace=True)
```

```
df
```

	invoice_id	Branch	City	category	unit_price	quantity	date	time	payment_method	rating	profit_margin
0	1	WALM003	San Antonio	Health and beauty	$74.69	7.0	05/01/2019	13:08:00	Ewallet	9.1	0.48
1	2	WALM048	Harlingen	Electronic accessories	$15.28	5.0	08/03/2019	10:29:00	Cash	9.6	0.48
2	3	WALM067	Haltom City	Home and lifestyle	$46.33	7.0	03/03/2019	13:23:00	Credit card	7.4	0.33
3	4	WALM064	Bedford	Health and beauty	$58.22	8.0	27/01/2019	20:33:00	Ewallet	8.4	0.33
4	5	WALM013	Irving	Sports and travel	$86.31	7.0	08/02/2019	10:37:00	Ewallet	5.3	0.48
...	...	...	...	...	...	...	...	...	...	...	...
9995	9996	WALM056	Rowlett	Fashion accessories	$37	3.0	03/08/2023	10:10:00	Cash	3.0	0.33
9996	9997	WALM030	Richardson	Home and lifestyle	$58	2.0	22/02/2021	14:20:00	Cash	7.0	0.48
9997	9998	WALM050	Victoria	Fashion accessories	$52	3.0	15/06/2023	16:00:00	Credit card	4.0	0.48
9998	9999	WALM032	Tyler	Home and lifestyle	$79	2.0	25/02/2021	12:25:00	Cash	7.0	0.48
9999	10000	WALM069	Rockwall	Fashion accessories	$62	3.0	26/09/2020	09:48:00	Cash	3.0	0.33

9969 rows × 11 columns



```
df.dtypes
```

```
invoice_id          int64
Branch             object
City               object
category           object
unit_price         object
quantity          float64
date               object
time               object
payment_method     object
rating            float64
profit_margin     float64
dtype: object
```

```
df['unit_price'].str.replace('$','')
```

```
0       74.69
1       15.28
2       46.33
3       58.22
4       86.31
        ...
9995       37
9996       58
9997       52
9998       79
9999       62
Name: unit_price, Length: 9969, dtype: object
```

```
df['unit_price'].str.replace('$','')
```

```
0       74.69
1       15.28
2       46.33
3       58.22
4       86.31
        ...
9995       37
9996       58
9997       52
9998       79
9999       62
Name: unit_price, Length: 9969, dtype: object
```

```
df['unit_price'].str.replace('$','').astype(float)
```

```
0       74.69
1       15.28
2       46.33
3       58.22
4       86.31
        ...
9995    37.00
9996    58.00
9997    52.00
9998    79.00
9999    62.00
Name: unit_price, Length: 9969, dtype: float64
```

```
df['unit_price']=df['unit_price'].str.replace('$','').astype(float)
```

```
df['unit_price']=df['unit_price'].str.replace('$','').astype(float)
```

```
df
```

	invoice_id	Branch	City	category	unit_price	quantity	date	time	payment_method	rating	profit_margin
0	1	WALM003	San Antonio	Health and beauty	74.69	7.0	05/01/2019	13:08:00	Ewallet	9.1	0.48
1	2	WALM048	Harlingen	Electronic accessories	15.28	5.0	08/03/2019	10:29:00	Cash	9.6	0.48
2	3	WALM067	Haltom City	Home and lifestyle	46.33	7.0	03/03/2019	13:23:00	Credit card	7.4	0.33
3	4	WALM064	Bedford	Health and beauty	58.22	8.0	27/01/2019	20:33:00	Ewallet	8.4	0.33
4	5	WALM013	Irving	Sports and travel	86.31	7.0	08/02/2019	10:37:00	Ewallet	5.3	0.48
...	...	...	...	...	...	...	...	...	...	...	...
9995	9996	WALM056	Rowlett	Fashion accessories	37.00	3.0	03/08/2023	10:10:00	Cash	3.0	0.33
9996	9997	WALM030	Richardson	Home and lifestyle	58.00	2.0	22/02/2021	14:20:00	Cash	7.0	0.48
9997	9998	WALM050	Victoria	Fashion accessories	52.00	3.0	15/06/2023	16:00:00	Credit card	4.0	0.48
9998	9999	WALM032	Tyler	Home and lifestyle	79.00	2.0	25/02/2021	12:25:00	Cash	7.0	0.48
9999	10000	WALM069	Rockwall	Fashion accessories	62.00	3.0	26/09/2020	09:48:00	Cash	3.0	0.33

9969 rows × 11 columns

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 9969 entries, 0 to 9999
Data columns (total 11 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   invoice_id      9969 non-null   int64  
 1   Branch          9969 non-null   object 
 2   City            9969 non-null   object 
 3   category        9969 non-null   object 
 4   unit_price      9969 non-null   float64
 5   quantity        9969 non-null   float64
 6   date            9969 non-null   object 
 7   time            9969 non-null   object 
 8   payment_method  9969 non-null   object 
 9   rating          9969 non-null   float64
 10  profit_margin   9969 non-null   float64
dtypes: float64(4), int64(1), object(6)
memory usage: 934.6+ KB
```

```
df.columns
```

```
Index(['invoice_id', 'Branch', 'City', 'category', 'unit_price', 'quantity',
       'date', 'time', 'payment_method', 'rating', 'profit_margin'],
      dtype='object')
```

```
df['total'] = df['unit_price'] * df['quantity']
```

```
df['total']
```

```
0       522.83
1        76.40
2       324.31
3       465.76
4       604.17
         ...  
9995    111.00
9996    116.00
9997    156.00
9998    158.00
9999    186.00
Name: total, Length: 9969, dtype: float64
```

```
df
```

	invoice_id	Branch	City	category	unit_price	quantity	date	time	payment_method	rating	profit_margin	total
0	1	WALM003	San Antonio	Health and beauty	74.69	7.0	05/01/2019	13:08:00	Ewallet	9.1	0.48	522.83
1	2	WALM048	Harlingen	Electronic accessories	15.28	5.0	08/03/2019	10:29:00	Cash	9.6	0.48	76.40
2	3	WALM067	Haltom City	Home and lifestyle	46.33	7.0	03/03/2019	13:23:00	Credit card	7.4	0.33	324.31
3	4	WALM064	Bedford	Health and beauty	58.22	8.0	27/01/2019	20:33:00	Ewallet	8.4	0.33	465.76
4	5	WALM013	Irving	Sports and travel	86.31	7.0	08/02/2019	10:37:00	Ewallet	5.3	0.48	604.17
...	...	...	...	...	...	...	...	...	...	...	...	...
9995	9996	WALM056	Rowlett	Fashion accessories	37.00	3.0	03/08/2023	10:10:00	Cash	3.0	0.33	111.00
9996	9997	WALM030	Richardson	Home and lifestyle	58.00	2.0	22/02/2021	14:20:00	Cash	7.0	0.48	116.00
9997	9998	WALM050	Victoria	Fashion accessories	52.00	3.0	15/06/2023	16:00:00	Credit card	4.0	0.48	156.00
9998	9999	WALM032	Tyler	Home and lifestyle	79.00	2.0	25/02/2021	12:25:00	Cash	7.0	0.48	158.00
9999	10000	WALM069	Rockwall	Fashion accessories	62.00	3.0	26/09/2020	09:48:00	Cash	3.0	0.33	186.00

9969 rows × 12 columns

```
from sqlalchemy import create_engine
```

```
username = [redacted]
password =
host =
port = 3306
database= "walmart_db"
engine = create_engine(f"mysql+mysqlconnector://{username}:{password}@{host}:{port}/{database}")
```

```
df.to_sql(name = 'walmart', con =engine, if_exists = 'append', index = False)
```

```
9969
```