```
import pandas as pd
from google.colab import files
files.upload()
df = pd.read_csv("retail_store_sales.csv")
print(df.head())
```

```
Choose Files retail_store_sales.csv
retail_store_sales.csv(text/csv) - 1192883 bytes, last modified: 1/8/2026 - 100% done
Saving retail_store_sales.csv to retail_store_sales (3).csv
  Transaction ID Customer ID       Category          Item  Price Per Unit  \
0    TXN_6867343     CUST_09     Patisserie   Item_10_PAT            18.5
1    TXN_3731986     CUST_22  Milk Products  Item_17_MILK            29.0
2    TXN_9303719     CUST_02       Butchers   Item_12_BUT            21.5
3    TXN_9458126     CUST_06      Beverages   Item_16_BEV            27.5
4    TXN_4575373     CUST_05           Food   Item_6_FOOD            12.5

   Quantity  Total Spent  Payment Method Location Transaction Date  \
0      10.0        185.0  Digital Wallet   Online       2024-04-08
1       9.0        261.0  Digital Wallet   Online       2023-07-23
2       2.0         43.0     Credit Card   Online       2022-10-05
3       9.0        247.5     Credit Card   Online       2022-05-07
4       7.0         87.5  Digital Wallet   Online       2022-10-02

  Discount Applied
0             True
1             True
2            False
3              NaN
4            False
```

```
df.head()
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12575 entries, 0 to 12574
Data columns (total 11 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Transaction ID    12575 non-null  object
 1   Customer ID       12575 non-null  object
 2   Category          12575 non-null  object
 3   Item              11362 non-null  object
 4   Price Per Unit    11966 non-null  float64
 5   Quantity          11971 non-null  float64
 6   Total Spent       11971 non-null  float64
 7   Payment Method    12575 non-null  object
 8   Location          12575 non-null  object
 9   Transaction Date  12575 non-null  object
 10  Discount Applied  8376 non-null   object
dtypes: float64(3), object(8)
memory usage: 1.1+ MB
```

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

	0
Transaction ID	0
Customer ID	0
Category	0
Item	1213
Price Per Unit	609
Quantity	604
Total Spent	604
Payment Method	0
Location	0
Transaction Date	0
Discount Applied	4199

dtype: int64

```
df
```

	Transaction ID	Customer ID	Category	Item	Price Per Unit	Quantity	Total Spent	Payment Method	Location	Transaction Date	Discount Applied
0	TXN_6867343	CUST_09	Patisserie	Item_10_PAT	18.5	10.0	185.0	Digital Wallet	Online	2024-04-08	True
1	TXN_3731986	CUST_22	Milk Products	Item_17_MILK	29.0	9.0	261.0	Digital Wallet	Online	2023-07-23	True
2	TXN_9303719	CUST_02	Butchers	Item_12_BUT	21.5	2.0	43.0	Credit Card	Online	2022-10-05	False
3	TXN_9458126	CUST_06	Beverages	Item_16_BEV	27.5	9.0	247.5	Credit Card	Online	2022-05-07	NaN
4	TXN_4575373	CUST_05	Food	Item_6_FOOD	12.5	7.0	87.5	Digital Wallet	Online	2022-10-02	False
...	...	...	...	...	...	...	...	...	...	...	...
12570	TXN_9347481	CUST_18	Patisserie	Item_23_PAT	38.0	4.0	152.0	Credit Card	In-store	2023-09-03	NaN
12571	TXN_4009414	CUST_03	Beverages	Item_2_BEV	6.5	9.0	58.5	Cash	Online	2022-08-12	False

Next steps: Generate code with df New interactive sheet

```
df.columns=df.iloc[0]
df=df.drop(index=0).reset_index(drop=True)
df
```

	TXN_6867343	CUST_09	Patisserie	Item_10_PAT	18.5	10.0	185.0	Digital Wallet	Online	2024-04-08	True
0	TXN_3731986	CUST_22	Milk Products	Item_17_MILK	29.0	9.0	261.0	Digital Wallet	Online	2023-07-23	True
1	TXN_9303719	CUST_02	Butchers	Item_12_BUT	21.5	2.0	43.0	Credit Card	Online	2022-10-05	False
2	TXN_9458126	CUST_06	Beverages	Item_16_BEV	27.5	9.0	247.5	Credit Card	Online	2022-05-07	NaN
3	TXN_4575373	CUST_05	Food	Item_6_FOOD	12.5	7.0	87.5	Digital Wallet	Online	2022-10-02	False
4	TXN_7482416	CUST_09	Patisserie	NaN	NaN	10.0	200.0	Credit Card	Online	2023-11-30	NaN
...	...	...	...	...	...	...	...	...	...	...	...
12569	TXN_9347481	CUST_18	Patisserie	Item_23_PAT	38.0	4.0	152.0	Credit Card	In-store	2023-09-03	NaN
12570	TXN_4009414	CUST_03	Beverages	Item_2_BEV	6.5	9.0	58.5	Cash	Online	2022-08-12	False
12571	TXN_5306010	CUST_11	Butchers	Item_7_BUT	14.0	10.0	140.0	Cash	Online	2024-08-24	NaN
12572	TXN_5167298	CUST_04	Furniture	Item_7_FUR	14.0	6.0	84.0	Cash	Online	2023-12-30	True
12573	TXN_2407494	CUST_23	Food	Item_9_FOOD	17.0	3.0	51.0	Cash	Online	2022-08-06	NaN

12574 rows × 11 columns

Next steps: Generate code with df New interactive sheet

```
df.columns.name=None
df.columns=df.columns.str.strip()
```

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12574 entries, 0 to 12573
Data columns (total 11 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   TXN_6867343     12574 non-null  object
 1   CUST_09         12574 non-null  object
 2   Patisserie      12574 non-null  object
 3   Item_10_PAT     11361 non-null  object
 4   nan             11965 non-null  float64
 5   nan             11970 non-null  float64
 6   nan             11970 non-null  float64
 7   Digital Wallet  12574 non-null  object
 8   Online          12574 non-null  object
 9   2024-04-08      12574 non-null  object
 10  nan             8375 non-null   object
dtypes: float64(3), object(8)
memory usage: 1.1+ MB
```

	0
TXN_6867343	0
CUST_09	0
Patisserie	0
Item_10_PAT	1213
NaN	609
NaN	604
NaN	604
Digital Wallet	0
Online	0
2024-04-08	0
NaN	4199

dtype: int64

```
df.duplicated().sum()
df = df.drop_duplicates()
```

```
df.dropna(subset=['Item_10_PAT'],inplace=True)
df.isnull().sum()
```

	0
TXN_6867343	0
CUST_09	0
Patisserie	0
Item_10_PAT	0
NaN	0
NaN	0
NaN	0
Digital Wallet	0
Online	0
2024-04-08	0
NaN	3783

dtype: int64

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

	0
TXN_6867343	0
CUST_09	0
Patisserie	0
Item_10_PAT	0
NaN	0
NaN	0
NaN	0
Digital Wallet	0
Online	0
2024-04-08	0
NaN	3783

dtype: int64

```
df.rename(columns={
    'Total Sales': 'total_sales',
    'Store Name': 'store_name'
}, inplace=True)
df
```

	TXN_6867343	CUST_09	Patisserie	Item_10_PAT	NaN	NaN	NaN	Digital Wallet	Online	2024-04-08	NaN
0	TXN_3731986	CUST_22	Milk Products	Item_17_MILK	29.0	9.0	261.0	Digital Wallet	Online	2023-07-23	True
1	TXN_9303719	CUST_02	Butchers	Item_12_BUT	21.5	2.0	43.0	Credit Card	Online	2022-10-05	False
2	TXN_9458126	CUST_06	Beverages	Item_16_BEV	27.5	9.0	247.5	Credit Card	Online	2022-05-07	NaN
3	TXN_4575373	CUST_05	Food	Item_6_FOOD	12.5	7.0	87.5	Digital Wallet	Online	2022-10-02	False
5	TXN_3652209	CUST_07	Food	Item_1_FOOD	5.0	8.0	40.0	Credit Card	In-store	2023-06-10	True
...	...	...	...	...	...	...	...	...	...	...	...
12569	TXN_9347481	CUST_18	Patisserie	Item_23_PAT	38.0	4.0	152.0	Credit Card	In-store	2023-09-03	NaN
12570	TXN_4009414	CUST_03	Beverages	Item_2_BEV	6.5	9.0	58.5	Cash	Online	2022-08-12	False
12571	TXN_5306010	CUST_11	Butchers	Item_7_BUT	14.0	10.0	140.0	Cash	Online	2024-08-24	NaN
12572	TXN_5167298	CUST_04	Furniture	Item_7_FUR	14.0	6.0	84.0	Cash	Online	2023-12-30	True
12573	TXN_2407494	CUST_23	Food	Item_9_FOOD	17.0	3.0	51.0	Cash	Online	2022-08-06	NaN

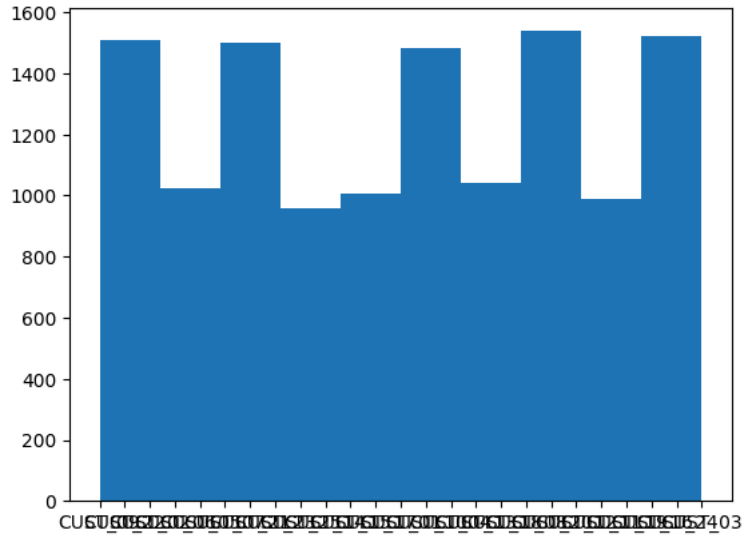
11361 rows × 11 columns


```
df.rename(columns={
    'Total Sales': 'total_sales',
    'Store Name': 'store_name'
}, inplace=True)
df
```

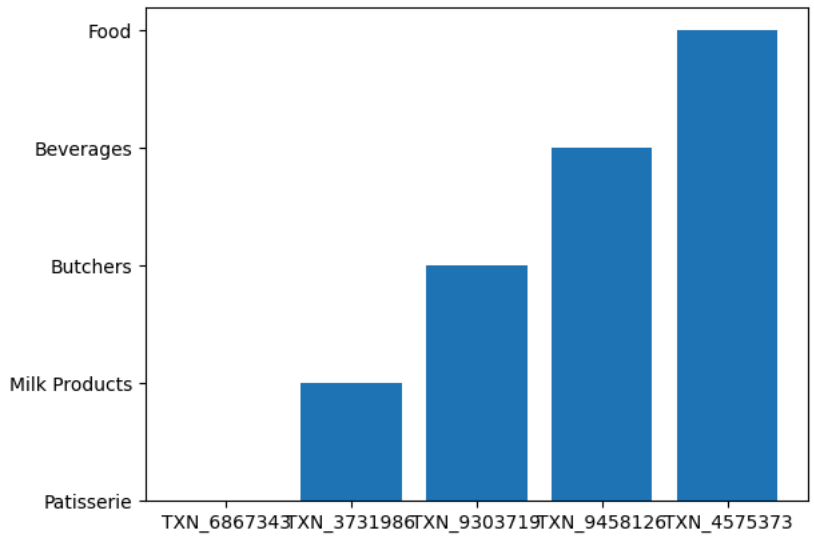
	TXN_6867343	CUST_09	Patisserie	Item_10_PAT	NaN	NaN	NaN	Digital Wallet	Online	2024-04-08	NaN
0	TXN_3731986	CUST_22	Milk Products	Item_17_MILK	29.0	9.0	261.0	Digital Wallet	Online	2023-07-23	True
1	TXN_9303719	CUST_02	Butchers	Item_12_BUT	21.5	2.0	43.0	Credit Card	Online	2022-10-05	False
2	TXN_9458126	CUST_06	Beverages	Item_16_BEV	27.5	9.0	247.5	Credit Card	Online	2022-05-07	NaN
3	TXN_4575373	CUST_05	Food	Item_6_FOOD	12.5	7.0	87.5	Digital Wallet	Online	2022-10-02	False
5	TXN_3652209	CUST_07	Food	Item_1_FOOD	5.0	8.0	40.0	Credit Card	In-store	2023-06-10	True
...	...	...	...	...	...	...	...	...	...	...	...
12569	TXN_9347481	CUST_18	Patisserie	Item_23_PAT	38.0	4.0	152.0	Credit Card	In-store	2023-09-03	NaN
12570	TXN_4009414	CUST_03	Beverages	Item_2_BEV	6.5	9.0	58.5	Cash	Online	2022-08-12	False
12571	TXN_5306010	CUST_11	Butchers	Item_7_BUT	14.0	10.0	140.0	Cash	Online	2024-08-24	NaN
12572	TXN_5167298	CUST_04	Furniture	Item_7_FUR	14.0	6.0	84.0	Cash	Online	2023-12-30	True
12573	TXN_2407494	CUST_23	Food	Item_9_FOOD	17.0	3.0	51.0	Cash	Online	2022-08-06	NaN

11361 rows × 11 columns

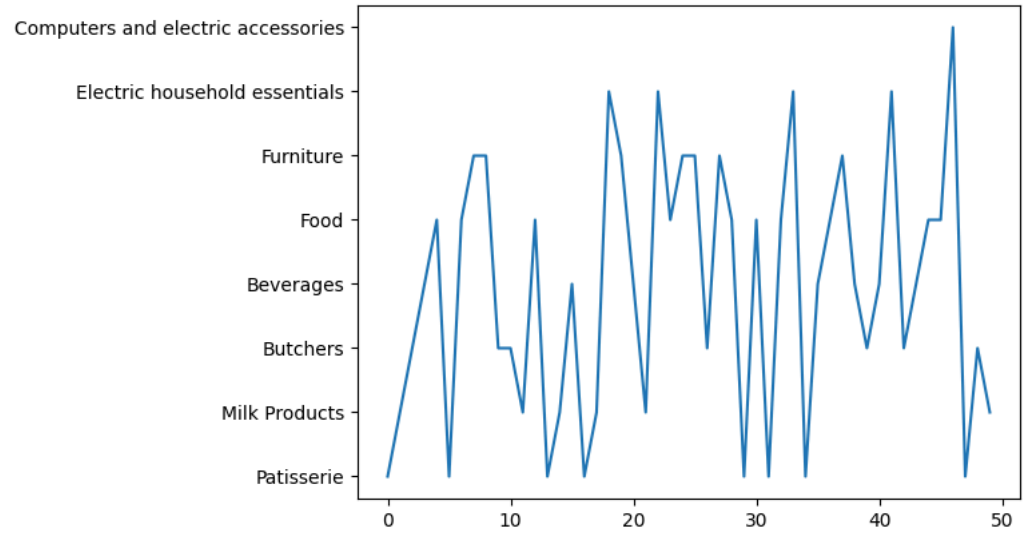
```
import pandas as pd
import matplotlib.pyplot as plt
df = pd.read_csv("retail_store_sales.csv")
plt.hist(df.iloc[:, 1], bins=10)
plt.show()
```




```
import pandas as pd
import matplotlib.pyplot as plt
df = pd.read_csv("retail_store_sales.csv")
plt.bar(df.iloc[:5, 0], df.iloc[:5, 2])
plt.show()
```




```
import pandas as pd
import matplotlib.pyplot as plt
df = pd.read_csv("retail_store_sales.csv")
plt.plot(df.iloc[:50, 2])
plt.show()
```




```
import pandas as pd
import matplotlib.pyplot as plt
df = pd.read_csv("retail_store_sales.csv")
plt.pie(df.select_dtypes(include='number').iloc[:5, 0])
plt.show()
```

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

	0
Transaction ID	0
Customer ID	0
Category	0
Item	1213
Price Per Unit	609
Quantity	604
Total Spent	604
Payment Method	0
Location	0
Transaction Date	0
Discount Applied	4199

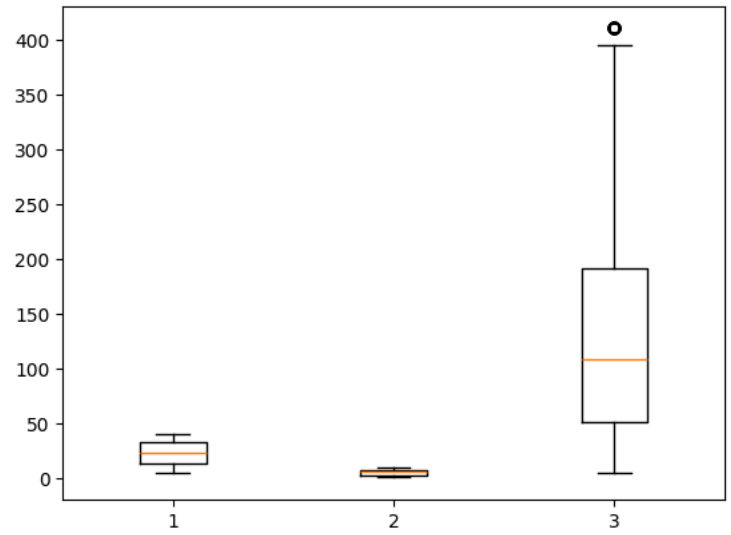
dtype: int64

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

```
df = pd.read_csv("retail_store_sales.csv")

sns.heatmap(df.select_dtypes(include='number').corr())
plt.show()
```


```
import pandas as pd
import matplotlib.pyplot as plt
df = pd.read_csv("retail_store_sales.csv")
plt.boxplot(df.select_dtypes(include='number').dropna())
plt.show()
```




```
import pandas as pd
```