```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt


id = "1_8Tx-yFlcA_4PZDU2LWxiugRwwK8PvJe"
print("https://drive.google.com/uc?export=download&id=" + id)
```

```
https://drive.google.com/uc?export=download&id=1_8Tx-yFlcA_4PZDU2LWxiugRwwK8Pv
```

```
!wget "https://drive.google.com/uc?export=download&id=1_8Tx-yFlcA_4PZDU2LWxiugRwwK8
```

```
--2022-04-23 16:44:39--  https://drive.google.com/uc?export=download&id=1_8Tx-
Resolving drive.google.com (drive.google.com)... 173.194.215.101, 173.194.215.
Connecting to drive.google.com (drive.google.com)|173.194.215.101|:443... conn
HTTP request sent, awaiting response... 303 See Other
Location: https://doc-0k-14-docs.googleusercontent.com/docs/securesc/ha0ro937g
Warning: wildcards not supported in HTTP.
--2022-04-23 16:44:40--  https://doc-0k-14-docs.googleusercontent.com/docs/sec
Resolving doc-0k-14-docs.googleusercontent.com (doc-0k-14-docs.googleuserconte
Connecting to doc-0k-14-docs.googleusercontent.com (doc-0k-14-docs.googleuserc
HTTP request sent, awaiting response... 200 OK
Length: 227054 (222K) [text/csv]
Saving to: 'marketing_data.csv'

marketing_data.csv  100%[===================>] 221.73K  --.-KB/s    in 0.002s

2022-04-23 16:44:40 (122 MB/s) - 'marketing_data.csv' saved [227054/227054]
```

```
df = pd.read_csv('./marketing_data.csv')
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2240 entries, 0 to 2239
Data columns (total 28 columns):
 #   Column               Non-Null Count  Dtype 
---  ------               --------------  ----- 
 0   ID                   2240 non-null   int64 
 1   Year_Birth           2240 non-null   int64 
 2   Education            2240 non-null   object
 3   Marital_Status       2240 non-null   object
 4    Income              2216 non-null   object
 5   Kidhome              2240 non-null   int64 
 6   Teenhome             2240 non-null   int64 
 7   Dt_Customer          2240 non-null   object
 8   Recency              2240 non-null   int64 
 9   MntWines             2240 non-null   int64 
 10  MntFruits            2240 non-null   int64 
 11  MntMeatProducts      2240 non-null   int64 
 12  MntFishProducts      2240 non-null   int64 
 13  MntSweetProducts     2240 non-null   int64 
 14  MntGoldProds         2240 non-null   int64 
```

```
 15  NumDealsPurchases    2240 non-null   int64
 16  NumWebPurchases      2240 non-null   int64
 17  NumCatalogPurchases  2240 non-null   int64
 18  NumStorePurchases    2240 non-null   int64
 19  NumWebVisitsMonth    2240 non-null   int64
 20  AcceptedCmp3         2240 non-null   int64
 21  AcceptedCmp4         2240 non-null   int64
 22  AcceptedCmp5         2240 non-null   int64
 23  AcceptedCmp1         2240 non-null   int64
 24  AcceptedCmp2         2240 non-null   int64
 25  Response             2240 non-null   int64
 26  Complain             2240 non-null   int64
 27  Country              2240 non-null   object
dtypes: int64(23), object(5)
memory usage: 490.1+ KB
```

```
df.head()
```

```
df.shape
```

```
(2240, 28)
```

```
df['MntGoldProds'].mean()
```

```
44.021875
```

```
df['MntFruits'].mean()
```

```
26.302232142857143
```

```
df['MntSweetProducts'].mean()
```

```
27.06294642857143
```

```
df['MntWines'].mean()
```

```
303.9357142857143
```

```
df["MntMeatProducts"].mean()
```

```
166.95
```

```
df["MntFishProducts"].mean()
```

```
37.52544642857143
```

```
df['MntWines']
```

```
0       189
1       464
2       134
3        10
4         6
       ...
2235    372
2236      5
2237    185
2238    267
2239    169
Name: MntWines, Length: 2240, dtype: int64
```

```
print("Gold:",df['MntGoldProds'].max(), df['MntGoldProds'].min())
print("Fruits:",df['MntFruits'].max(), df['MntFruits'].min())
print("Sweets:",df['MntSweetProducts'].max(), df['MntSweetProducts'].min())
print("Wine:",df['MntWines'].max(), df['MntWines'].min())
print("Meat:",df['MntMeatProducts'].max(), df['MntMeatProducts'].min())
print("Fish:",df['MntFishProducts'].max(), df['MntFishProducts'].min())
```

```
Gold: 362 0
Fruits: 199 0
Sweets: 263 0
Wine: 1493 0
Meat: 1725 0
Fish: 259 0
```

```
print("Gold:",df['MntGoldProds'].mean(), df['MntGoldProds'].median())
print("Fruits:",df['MntFruits'].mean(), df['MntFruits'].median())
print("Sweets:",df['MntSweetProducts'].mean(), df['MntSweetProducts'].median())
print("Wine:",df['MntWines'].mean(), df['MntWines'].median())
print("Meat:",df['MntMeatProducts'].mean(), df['MntMeatProducts'].median())
print("Fish:",df['MntFishProducts'].mean(), df['MntFishProducts'].median())
```

```
Gold: 44.021875 24.0
Fruits: 26.302232142857143 8.0
Sweets: 27.06294642857143 8.0
Wine: 303.9357142857143 173.5
Meat: 166.95 67.0
Fish: 37.52544642857143 12.0
```

```
#mode
df["Education"].value_counts()
```

```
Graduation    1127
PhD            486
Master         370
2n Cycle       203
Basic           54
Name: Education, dtype: int64
```

```
#Std
print("Gold:",df['MntGoldProds'].std())
print("Fruits:",df['MntFruits'].std())
print("Sweets:",df['MntSweetProducts'].std())
print("Wine:",df['MntWines'].std())
print("Meat:",df['MntMeatProducts'].std())
print("Fish:",df['MntFishProducts'].std())
```

```
Gold: 52.167438914997064
Fruits: 39.77343376457871
Sweets: 41.2804984878548
Wine: 336.5973926053717
Meat: 225.71537251175445
Fish: 54.62897940287769
```

```
from scipy import stats
print(stats.median_absolute_deviation(df['MntGoldProds']))
```

```
26.686799999999998
```

```
stats.median_absolute_deviation(df['MntFruits'])
```

```
11.8608
```

```
stats.median_absolute_deviation(df['MntSweetProducts'])
```

```
11.8608
```

```
stats.median_absolute_deviation(df['MntWines'])
```

```
243.8877
```

```
stats.median_absolute_deviation(df['MntMeatProducts'])
```

```
87.4734
```

```
stats.median_absolute_deviation(df['MntFishProducts'])
```

```
17.7912
```

```
# 30% of the customers spending less than $X
np.percentile(df['MntWines'], 30)

# X=$34.0
```

```
34.0
```

```
#IQR
stats.iqr(df['MntWines'])
```

```
480.5
```

```
stats.iqr(df['MntFruits'])
```

```
32.0
```

```
stats.iqr(df['MntGoldProds'])
```

```
47.0
```

```
#outliers using IQR range
r = 1.5*stats.iqr(df['MntWines'])
lb = np.percentile(df['MntWines'], 25)-r
ub = np.percentile(df['MntWines'], 75)+r
print(lb)
print(ub)
```

```
-697.0
1225.0
```

```
df['MntWines'].max()
```

```
1493
```

```
sum(df['MntWines'] > ub)
```

```
35
```

```
df['MntWines'].plot.hist()
#chooses nbins=10
```

```
df['MntWines'].plot.hist(bins=100)
```

```
import seaborn as sns
sns.kdeplot(data=df['MntWines'])
```

```
df[' Income ']
```

```
0       $84,835.00
1       $57,091.00
2       $67,267.00
3       $32,474.00
4       $21,474.00
           ...
2235    $66,476.00
2236    $31,056.00
2237    $46,310.00
2238    $65,819.00
2239    $94,871.00
Name:  Income , Length: 2240, dtype: object
```

```
def get_clean_income(income):
    if isinstance(income, str):
        return float(income[1:].replace(',', ''))
    return income
```

```
df[' Income '] =  df[' Income '].map(lambda x : get_clean_income(x) )
```

```
df[' Income '].plot.hist(bins=100)
# few extreme values to the right.
```

```
#box plot

ax = df[' Income '].plot.box()
ax.set_ylabel("Income in Dollars")
```

```
#box plot
ax = df['MntWines'].plot.box()
ax.set_ylabel("Amount spend in Dollars")
```

```
#bar plot
df['Education'].value_counts().plot.bar()
```

```
df['Country'].value_counts().plot.bar()
```

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

```
Married     864
Together    580
Single      480
Divorced    232
Widow        77
Alone         3
YOLO          2
Absurd        2
Name: Marital_Status, dtype: int64
```

```
#scatter plot
df.plot.scatter(x=' Income ', y='MntWines',  figsize = (10,10), c='green')
```

```
plt.figure(figsize=(8,8))
plt.subplot(2,2,1)
plt.scatter(x=df[' Income '], y=df['MntFruits'], c='green')

plt.subplot(2,2,2)
plt.scatter(x=df[' Income '], y=df['MntGoldProds'], c='green')

plt.subplot(2,2,3)
plt.scatter(x=df[' Income '], y=df['MntSweetProducts'], c='green')

plt.subplot(2,2,4)
plt.scatter(x=df[' Income '], y=df['MntWines'], c='green')
```

```
# Have they responded to the marketing survey by our company
df['Response']
```

```
0       1
1       1
```

```
2       0
3       0
4       1
       ..
2235    0
2236    0
2237    0
2238    0
2239    1
Name: Response, Length: 2240, dtype: int64
```

```
df.boxplot(by='Response', column=' Income ', figsize=(10,6))
plt.show()
```

✓ 0s completed at 22:14