Sheet

Supermarket Sales

Assignment 1 done by: Mohamed Hatem El-Badry, Shahd Hasnoon

DATA SET = SUPERMARKET SALES Source: https://www.kaggle.com/aungpyaeap/supermarket-sales

Attribute information: Invoice id: Computer generated sales slip invoice identification number (Identifier) (Character string)

Branch: Branch of supercenter (3 branches are available identified by A, B and C). (Qualitative Categorical) (Factor)

City: Location of supercenters (Qualitative Categorical) (Factor)

Customer type: Type of customers, recorded by Members for customers using member card and Normal for without member card. (Qualitative Categorical) (Factor)

Gender: Gender type of customer (Qualitative Categorical) (Factor)

Product line: General item categorization groups - Electronic accessories, Fashion accessories, Food and beverages, Health and beauty, Home and lifestyle, Sports and travel (Qualitative Categorical) (Factor)

Unit price: Price of each product unit of measurement: in $(Numeric, Quantitative) Quantity : Number of products purchased by customer unit of measurement : items (Numeric, Quantitative) Tax : 5$ (Numeric, Quantitative)

Total: Total price including tax unit of measurement: in $(Numeric, Quantitative) Date : Date of purchase (Record available from January 2019 to March 2019) (Qualitative) (Factor) Time : Purchase time (10am to 9pm) (Qualitative) (Factor) Payment : Payment used by customer for purchase (3 methods are available – Cash, Credit card and Ewallet) (Qualitative Categorical) (Factor) COGS : Cost of goods sold unit of measurement : in$ (Numeric, Quantitative)

Gross margin percentage: Gross margin percentage unit of measurement: in % (Numeric, Quantitative)

Gross income: Gross income unit of measurement: in $ (Numeric, Quantitative)

Rating: Customer stratification rating on their overall shopping experience (On a scale of 1 to 10) unit of measurement: no unit of measurement (ordinal)(numeric)

Purpose: This dataset can be used for predictive data analytics purpose.

Population: All supermarkets

Sample(where): the 3 branches of the supermarket [A,B,C]

Observations(who): 1000 customers

Variables(what): Invoice id, Branch, City, Customer type, Gender, Product line, Unit price, Quantity, Tax, Total, Date, Time, Payment, COGS, Gross margin percentage, Gross income and Rating.

```
!pip install pandas
import pandas as pd
```

```
Requirement already satisfied: pandas in /opt/python/envs/default/lib/python3.8/site-packages (1.3.5)
Requirement already satisfied: pytz>=2017.3 in /opt/python/envs/default/lib/python3.8/site-packages (from pandas) (2021.3)
Requirement already satisfied: python-dateutil>=2.7.3 in /opt/python/envs/default/lib/python3.8/site-packages (from pandas) (2.8.2)
Requirement already satisfied: numpy>=1.17.3 in /opt/python/envs/default/lib/python3.8/site-packages (from pandas) (1.21.5)
Requirement already satisfied: six>=1.5 in /opt/python/envs/default/lib/python3.8/site-packages (from python-dateutil>=2.7.3->pandas)
WARNING: You are using pip version 21.3.1; however, version 22.0.3 is available.
You should consider upgrading via the '/opt/python/envs/default/bin/python -m pip install --upgrade pip' command.
```

```
!pip install openpyxl
import openpyxl
```

```
Collecting openpyxl
  Downloading openpyxl-3.0.9-py2.py3-none-any.whl (242 kB)
?25l     |▉                               | 10 kB 23.5 MB/s eta 0:00:01     |▉                                | 20 kB 28.8 MB/s eta 0:0(
?25hCollecting et-xmlfile
  Downloading et_xmlfile-1.1.0-py3-none-any.whl (4.7 kB)
Installing collected packages: et-xmlfile, openpyxl
Successfully installed et-xmlfile-1.1.0 openpyxl-3.0.9
WARNING: You are using pip version 21.3.1; however, version 22.0.3 is available.
You should consider upgrading via the '/opt/python/envs/default/bin/python -m pip install --upgrade pip' command.
```

```
x= pd.read_csv("data.csv")
x
```

	Invoice ID	Branch	City	Customer type	Gender	Product line	Unit price	Quantity	Tax 5%	Total	Date	Time	Payment	cogs	gross margin percentage	gross income	Rating
0	750-67-8428	A	Yangon	Member	Female	Health and beauty	74.69	7	26.1415	548.9715	1/5/2019	13:08	Ewallet	522.83	4.761905	26.1415	9.1
1	226-31-3081	C	Naypyitaw	Normal	Female	Electronic accessories	15.28	5	3.8200	80.2200	3/8/2019	10:29	Cash	76.40	4.761905	3.8200	9.6
2	631-41-3108	A	Yangon	Normal	Male	Home and lifestyle	46.33	7	16.2155	340.5255	3/3/2019	13:23	Credit card	324.31	4.761905	16.2155	7.4
3	123-19-1176	A	Yangon	Member	Male	Health and beauty	58.22	8	23.2880	489.0480	1/27/2019	20:33	Ewallet	465.76	4.761905	23.2880	8.4
4	373-73-7910	A	Yangon	Normal	Male	Sports and travel	86.31	7	30.2085	634.3785	2/8/2019	10:37	Ewallet	604.17	4.761905	30.2085	5.3
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
995	233-67-5758	C	Naypyitaw	Normal	Male	Health and beauty	40.35	1	2.0175	42.3675	1/29/2019	13:46	Ewallet	40.35	4.761905	2.0175	6.2
996	303-96-2227	B	Mandalay	Normal	Female	Home and lifestyle	97.38	10	48.6900	1022.4900	3/2/2019	17:16	Ewallet	973.80	4.761905	48.6900	4.4
997	727-02-1313	A	Yangon	Member	Male	Food and beverages	31.84	1	1.5920	33.4320	2/9/2019	13:22	Cash	31.84	4.761905	1.5920	7.7
998	347-56-2442	A	Yangon	Normal	Male	Home and lifestyle	65.82	1	3.2910	69.1110	2/22/2019	15:33	Cash	65.82	4.761905	3.2910	4.1
999	849-09-3807	A	Yangon	Member	Female	Fashion accessories	88.34	7	30.9190	649.2990	2/18/2019	13:28	Cash	618.38	4.761905	30.9190	6.6

1000 rows × 17 columns

```
m= x[[ 'Quantity' , 'Unit price' , 'Tax 5%', 'Rating', 'Total' , 'cogs' , 'gross income'  ]]
```

```
m
```

	Quantity	Unit price	Tax 5%	Rating	Total	cogs	gross income
0	7	74.69	26.1415	9.1	548.9715	522.83	26.1415
1	5	15.28	3.8200	9.6	80.2200	76.40	3.8200
2	7	46.33	16.2155	7.4	340.5255	324.31	16.2155
3	8	58.22	23.2880	8.4	489.0480	465.76	23.2880
4	7	86.31	30.2085	5.3	634.3785	604.17	30.2085
...	...	...	...	...	...	...	...
995	1	40.35	2.0175	6.2	42.3675	40.35	2.0175
996	10	97.38	48.6900	4.4	1022.4900	973.80	48.6900
997	1	31.84	1.5920	7.7	33.4320	31.84	1.5920
998	1	65.82	3.2910	4.1	69.1110	65.82	3.2910
999	7	88.34	30.9190	6.6	649.2990	618.38	30.9190

1000 rows × 7 columns

```
m.mean()
```

```
m.median()
```

```
!pip install numpy
import numpy as np
```

```
Requirement already satisfied: numpy in /opt/python/envs/default/lib/python3.8/site-packages (1.21.5)
WARNING: You are using pip version 21.3.1; however, version 22.0.3 is available.
You should consider upgrading via the '/opt/python/envs/default/bin/python -m pip install --upgrade pip' command.
```

```
a=np.array([x['Quantity'].quantile(0) ,
x['Quantity'].quantile(0.25) ,
x['Quantity'].quantile(0.5)  ,
x['Quantity'].quantile(0.75) ,
x['Quantity'].quantile(1) ,])
print(a)
```

```
[ 1.  3.  5.  8. 10.]
```

```
b=np.array([x['Tax 5%'].quantile(0) ,
x['Tax 5%'].quantile(0.25) ,
x['Tax 5%'].quantile(0.5)  ,
x['Tax 5%'].quantile(0.75) ,
x['Tax 5%'].quantile(1) ,])
print(b)
```

```
[ 0.5085    5.924875 12.088    22.44525  49.65    ]
```

```
c=np.array([x['Unit price'].quantile(0) ,
x['Unit price'].quantile(0.25) ,
x['Unit price'].quantile(0.5)  ,
x['Unit price'].quantile(0.75) ,
x['Unit price'].quantile(1) ,])
print(c)
```

```
[10.08  32.875 55.23  77.935 99.96 ]
```

```
d=np.array([x['Rating'].quantile(0) ,
x['Rating'].quantile(0.25) ,
x['Rating'].quantile(0.5)  ,
x['Rating'].quantile(0.75) ,
x['Rating'].quantile(1) ,])
print(d)
```

```
[ 4.   5.5  7.   8.5 10. ]
```

```
e=np.array([x['Total'].quantile(0) ,
x['Total'].quantile(0.25) ,
x['Total'].quantile(0.5)  ,
x['Total'].quantile(0.75) ,
x['Total'].quantile(1) ,])
print(e)
```

```
[  10.6785    124.422375  253.848     471.35025  1042.65    ]
```

```
f=np.array([x['cogs'].quantile(0) ,
x['cogs'].quantile(0.25) ,
x['cogs'].quantile(0.5)  ,
x['cogs'].quantile(0.75) ,
x['cogs'].quantile(1) ,])
print(f)
```

```
[ 10.17   118.4975 241.76   448.905  993.    ]
```

```
g=np.array([x['gross income'].quantile(0) ,
x['gross income'].quantile(0.25) ,
x['gross income'].quantile(0.5)  ,
x['gross income'].quantile(0.75) ,
x['gross income'].quantile(1) ,])
print(g)
```

```
[ 0.5085    5.924875 12.088    22.44525  49.65    ]
```

```
m.var()
```

```
m.std()
```

```
x[['Tax 5%', 'Unit price']].corr(method='pearson')
```

	Tax 5%	Unit price
Tax 5%	1.000000	0.633962
Unit price	0.633962	1.000000

the correlation coefficient is close to 1 which shows a strong relation betweem the unit price and the tax; which is what was expected

```
x[['Quantity', 'Unit price']].corr(method='pearson')
```

	Quantity	Unit price
Quantity	1.000000	0.010778
Unit price	0.010778	1.000000

since the coefficient correlation is close to zero, therefore there is no linear relationship between Quantity and Unit price

```
x['Quantity'].value_counts()
```

```
x['Quantity'].value_counts(normalize=True)
```

the discrete variable Quantity doesnt have a certain relation/shape that it follows. the most quantity of items bought is 10 with a percentage of 11.9 and the least is 8 with a percentage of 8.5

```
x['Gender'].value_counts()
```

```
x['Gender'].value_counts(normalize=True)
```

the discrete variable gender shows an approximatly half-half relationship of males and females going to the 3 branches of supermarket

```
x['Branch'].value_counts()
```

```
x['Branch'].value_counts(normalize=True)
```

the discrete variable branch shows a nearly equal observations from all the 3 branches

```
q=pd.crosstab(x['Gender'], x['Branch'])
q
```

Branch	A	B	C
Gender			
Female	161	162	178
Male	179	170	150

```
!pip install statsmodels
```

```
Requirement already satisfied: statsmodels in /opt/python/envs/default/lib/python3.8/site-packages (0.13.1)
Requirement already satisfied: pandas>=0.25 in /opt/python/envs/default/lib/python3.8/site-packages (from statsmodels) (1.3.5)
Requirement already satisfied: patsy>=0.5.2 in /opt/python/envs/default/lib/python3.8/site-packages (from statsmodels) (0.5.2)
Requirement already satisfied: scipy>=1.3 in /opt/python/envs/default/lib/python3.8/site-packages (from statsmodels) (1.7.3)
Requirement already satisfied: numpy>=1.17 in /opt/python/envs/default/lib/python3.8/site-packages (from statsmodels) (1.21.5)
Requirement already satisfied: python-dateutil>=2.7.3 in /opt/python/envs/default/lib/python3.8/site-packages (from pandas>=0.25->stats
```

```
Requirement already satisfied: pytz>=2017.3 in /opt/python/envs/default/lib/python3.8/site-packages (from pandas>=0.25->statsmodels) (2
Requirement already satisfied: six in /opt/python/envs/default/lib/python3.8/site-packages (from patsy>=0.5.2->statsmodels) (1.16.0)
WARNING: You are using pip version 21.3.1; however, version 22.0.3 is available.
You should consider upgrading via the '/opt/python/envs/default/bin/python -m pip install --upgrade pip' command.
```

```
import statsmodels.api as sm
```

```
tab = pd.crosstab(x['Branch'], x['Gender'])
```

```
 tab.loc[:, ['Female'] ]
```

Gender	Female
Branch	
A	161
B	162
C	178

```
table = sm.stats.Table(tab)
```

```
table.resid_pearson
```

Gender	Female	Male
Branch		
A	-0.715630	0.717063
B	-0.335893	0.336565
C	1.066538	-1.068673

since the pearson risidual is close to zero and has a modulus of less than 2, therefore the variables are independant

```
!pip install seaborn
```

```
Requirement already satisfied: seaborn in /opt/python/envs/default/lib/python3.8/site-packages (0.11.2)
Requirement already satisfied: numpy>=1.15 in /opt/python/envs/default/lib/python3.8/site-packages (from seaborn) (1.21.5)
Requirement already satisfied: scipy>=1.0 in /opt/python/envs/default/lib/python3.8/site-packages (from seaborn) (1.7.3)
Requirement already satisfied: matplotlib>=2.2 in /opt/python/envs/default/lib/python3.8/site-packages (from seaborn) (3.5.1)
Requirement already satisfied: pandas>=0.23 in /opt/python/envs/default/lib/python3.8/site-packages (from seaborn) (1.3.5)
Requirement already satisfied: packaging>=20.0 in /opt/python/envs/default/lib/python3.8/site-packages (from matplotlib>=2.2->seaborn)
Requirement already satisfied: kiwisolver>=1.0.1 in /opt/python/envs/default/lib/python3.8/site-packages (from matplotlib>=2.2->seabor
Requirement already satisfied: fonttools>=4.22.0 in /opt/python/envs/default/lib/python3.8/site-packages (from matplotlib>=2.2->seabor
Requirement already satisfied: pillow>=6.2.0 in /opt/python/envs/default/lib/python3.8/site-packages (from matplotlib>=2.2->seaborn) (
Requirement already satisfied: cycler>=0.10 in /opt/python/envs/default/lib/python3.8/site-packages (from matplotlib>=2.2->seaborn) (0
Requirement already satisfied: python-dateutil>=2.7 in /opt/python/envs/default/lib/python3.8/site-packages (from matplotlib>=2.2->seab
Requirement already satisfied: pyparsing>=2.2.1 in /opt/python/envs/default/lib/python3.8/site-packages (from matplotlib>=2.2->seaborn)
Requirement already satisfied: pytz>=2017.3 in /opt/python/envs/default/lib/python3.8/site-packages (from pandas>=0.23->seaborn) (2021
Requirement already satisfied: six>=1.5 in /opt/python/envs/default/lib/python3.8/site-packages (from python-dateutil>=2.7->matplotlib
WARNING: You are using pip version 21.3.1; however, version 22.0.3 is available.
You should consider upgrading via the '/opt/python/envs/default/bin/python -m pip install --upgrade pip' command.
```

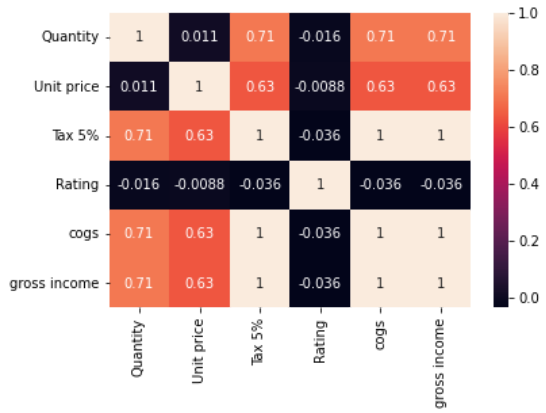
```
import seaborn as sns
```

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

```
cormat= x[['Quantity', 'Unit price' , 'Tax 5%', 'Rating' , 'cogs' , 'gross income']].corr(method='pearson')
```

```
sns.heatmap(cormat, annot=True)
```

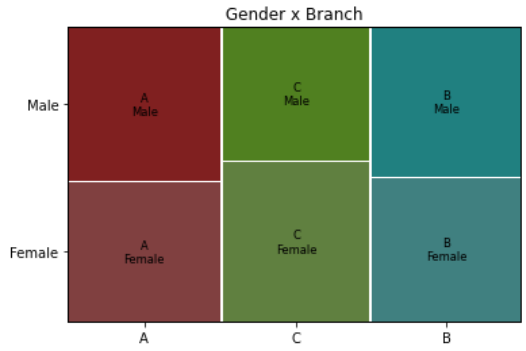
```
<AxesSubplot:>
```




```
l=x['Branch'].value_counts()
l
```

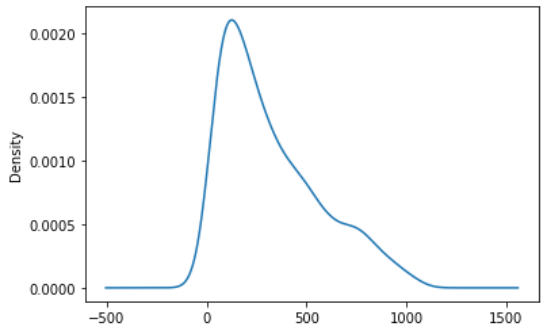
```
import matplotlib.pyplot as plt
from statsmodels.graphics.mosaicplot import mosaic
```

```
mosaic(x, ['Branch', 'Gender'], title=' Gender x Branch ')
plt.show()
```



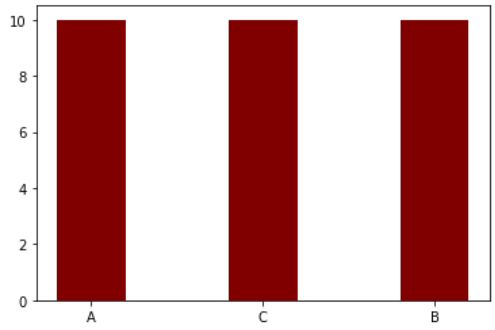

```
x.Total.plot.density()
```

```
<AxesSubplot:ylabel='Density'>
```



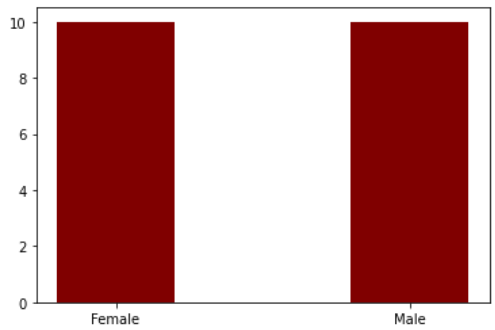

```
plt.bar(x['Branch'], x['Quantity'], color ='maroon',
        width = 0.4)
```

```
<BarContainer object of 1000 artists>
```




```
plt.bar(x['Gender'], x['Quantity'], color ='maroon',
        width = 0.4)
```

```
<BarContainer object of 1000 artists>
```




```
v=np.arange(8)
v
```