IE 582 HW 4
Haizhou Wang

Q1.
a.

Total number of transactions T: 10
Supports:

$$Support\{e\} = \frac{\sigma(\{e\})}{T} = \frac{8}{10} = 0.8$$
$$Support\{b,d\} = \frac{2}{10} = 0.2$$
$$Support\{b,d,e\} = \frac{2}{10} = 0.2$$

b.

Confidence:

$$Confidence(\{b,d\} \rightarrow \{e\}) = \frac{\sigma(\{e,b,d\})}{\sigma(\{b,d\})} = \frac{2}{2} = 1$$
$$Confidence(\{e\} \rightarrow \{b,d\}) = \frac{\sigma(\{e,b,d\})}{\sigma(\{e\})} = \frac{2}{8} = 0.25$$

c.

Total number of transactions T: 5

Customer ID	Items
1	a,b,c,d,e
2	a,b,c,d,e
3	b,c,d,e
4	a,b,c,d
5	a,b,d,e

$$Support\{e\} = \frac{\sigma(\{e\})}{T} = \frac{4}{5} = 0.8$$
$$Support\{b,d\} = \frac{5}{5} = 1.0$$
$$Support\{b,d,e\} = \frac{4}{5} = 0.8$$

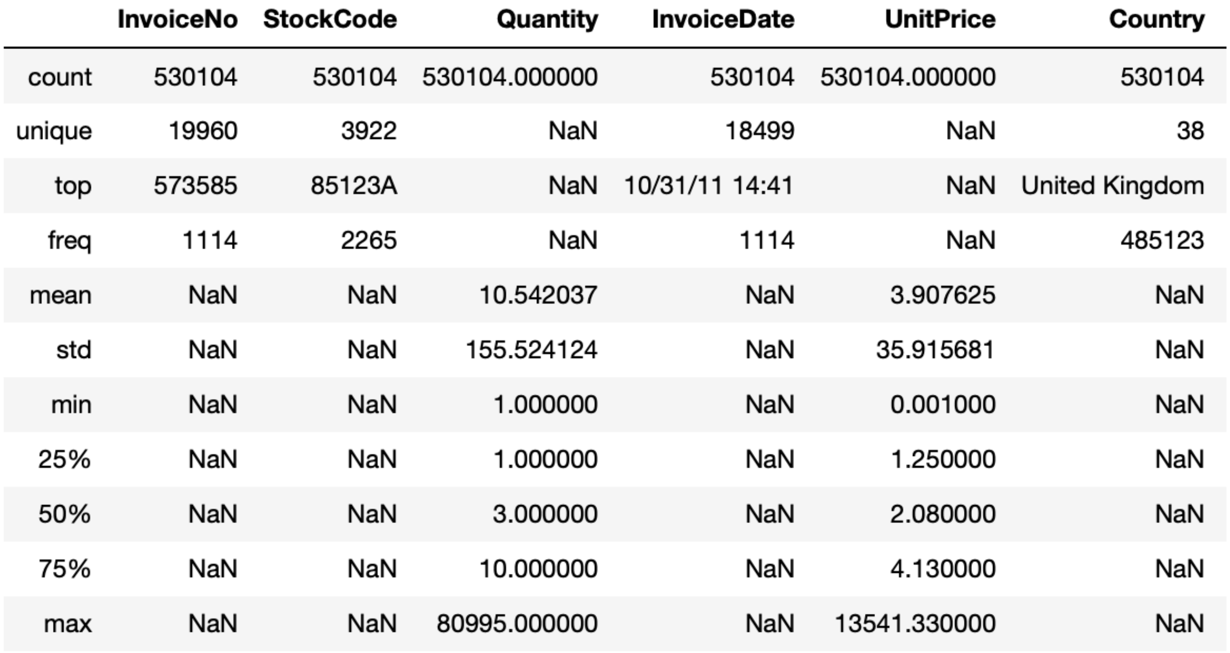
d.

$$Confidence(\{b,d\} \rightarrow \{e\}) = \frac{\sigma(\{e,b,d\})}{\sigma(\{b,d\})} = \frac{4}{5} = 0.8$$
$$Confidence(\{e\} \rightarrow \{b,d\}) = \frac{\sigma(\{e,b,d\})}{\sigma(\{e\})} = \frac{4}{4} = 1$$

e.

After changing the basket to per customer, the number of each item increased. Therefore, the support increased, making confidence increased as well.

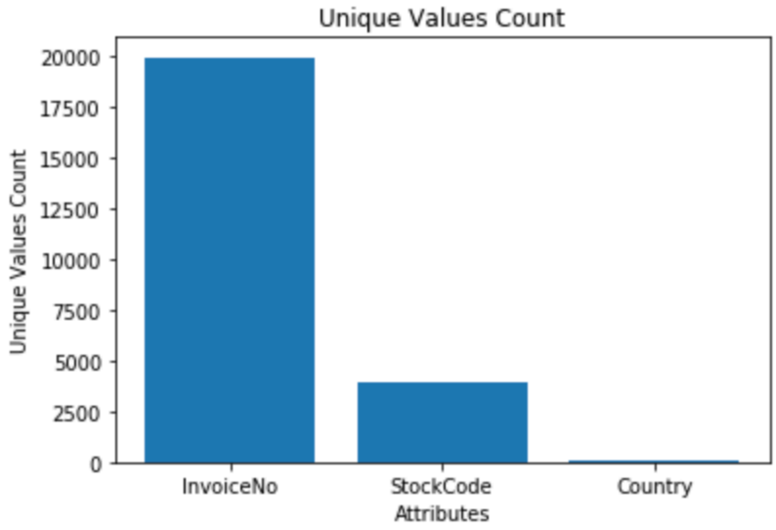
Q2.
a.



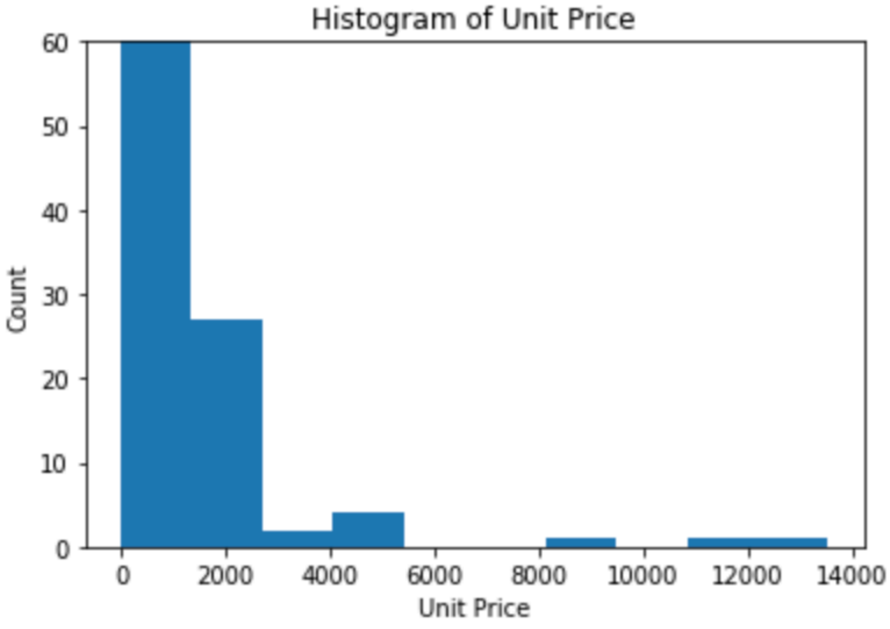
	InvoiceNo	StockCode	Quantity	InvoiceDate	UnitPrice	Country
count	530104	530104	530104.000000	530104	530104.000000	530104
unique	19960	3922	NaN	18499	NaN	38
top	573585	85123A	NaN	10/31/11 14:41	NaN	United Kingdom
freq	1114	2265	NaN	1114	NaN	485123
mean	NaN	NaN	10.542037	NaN	3.907625	NaN
std	NaN	NaN	155.524124	NaN	35.915681	NaN
min	NaN	NaN	1.000000	NaN	0.001000	NaN
25%	NaN	NaN	1.000000	NaN	1.250000	NaN
50%	NaN	NaN	3.000000	NaN	2.080000	NaN
75%	NaN	NaN	10.000000	NaN	4.130000	NaN
max	NaN	NaN	80995.000000	NaN	13541.330000	NaN

Description table of the data from pandas

```
Count of Unique Values:
('InvoiceNo', 19960) ('StockCode', 3922) ('Country', 38)
```

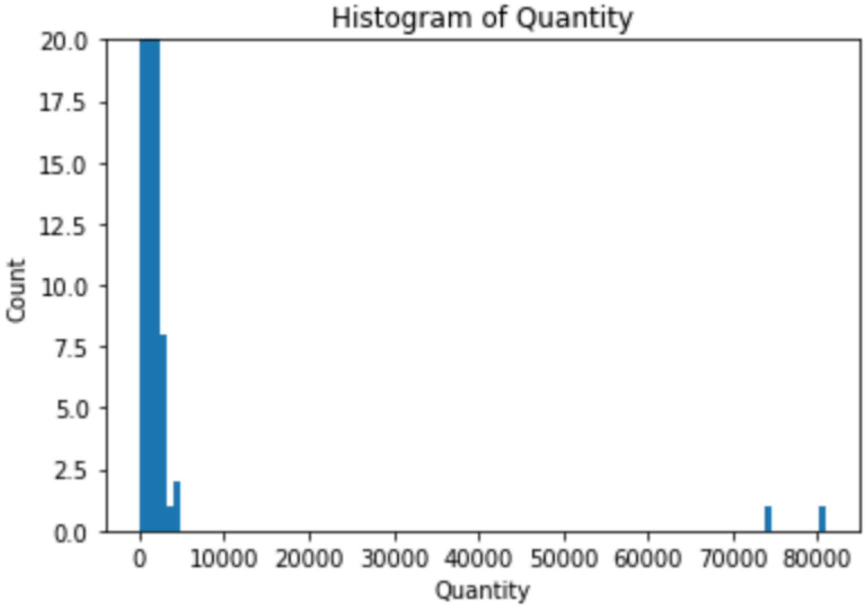


```
Unit Price Stats:
Mean:  3.90762524712132
std:  35.91568110425544
```



```
Remark: Y axis limited to 60
```

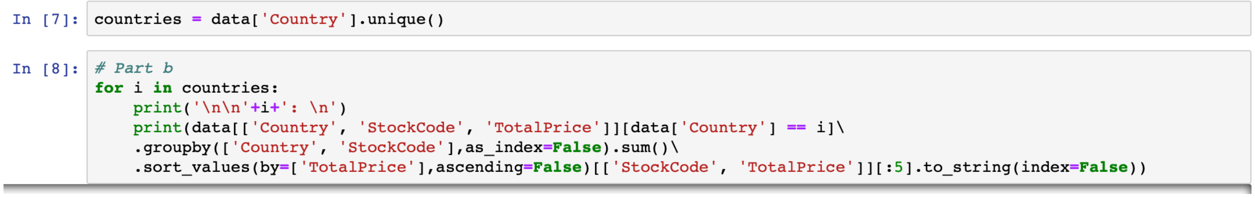
```
Quantity Stats:
Mean:  10.542037034242338
std:  155.52412351063626
```



```
Remark: Y axis limited to
```

b.

Code:



```
In [7]: countries = data['Country'].unique()

In [8]: # Part b
        for i in countries:
            print('\n\n'+i+': \n')
            print(data[['Country', 'StockCode', 'TotalPrice']][data['Country'] == i]\
            .groupby(['Country', 'StockCode'],as_index=False).sum()\
            .sort_values(by=['TotalPrice'],ascending=False)[['StockCode', 'TotalPrice']][:5].to_string(index=False))
```

Results:

```
United Kingdom:

StockCode  TotalPrice
      DOT   206248.77
    23843   168469.60
    22423   142273.29
   85123A    98723.75
    47566    93658.53


France:
```

```
StockCode  TotalPrice
     POST    15454.00
        M     9492.37
    23084     7277.20
    22423     2816.85
    21731     2169.75


Australia:

StockCode  TotalPrice
    23084     3375.84
    22722     2082.00
    21731     1987.20
    22720     1983.20
    22423     1978.20


Netherlands:

StockCode  TotalPrice
    23084     9568.48
    22326     7991.40
    22629     7485.60
    22630     6828.60
    22328     4039.20


Germany:

StockCode  TotalPrice
     POST    21001.00
    22423     9061.95
    22326     3598.95
        M     2296.25
    22328     1982.40


Norway:

StockCode  TotalPrice
     POST      2870.5
        M       840.3
    22693       538.8
    22635       527.4
    22634       487.6


EIRE:

StockCode  TotalPrice
    22423     7844.25
        M     7049.66
```

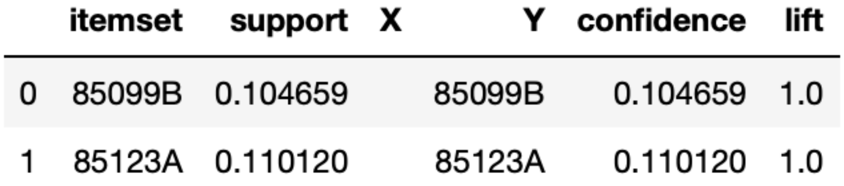
```
       C2     5240.00
    22838     4265.55
    22960     3097.50


Switzerland:

StockCode  TotalPrice
     POST     4002.00
    22326     1300.80
    22554      977.55
    22423      924.15
    22551      733.80


Spain:

StockCode  TotalPrice
     POST     5852.00
   84997D     3957.75
   84997C     3671.15
    22423     2049.00
   84997B     1044.76


Poland:

StockCode  TotalPrice
     POST      360.00
    21232      196.32
    37448      191.52
    22722      177.60
    22666      167.40


Portugal:

StockCode  TotalPrice
        M     4223.94
     POST     2508.00
    22139      463.35
    22411      387.40
    20725      354.75


Italy:

StockCode  TotalPrice
     POST     1663.00
    22720      252.45
    22847      247.20
    22139      222.75
    22960      217.50
```

```
Belgium:

StockCode  TotalPrice
     POST     4269.00
    22326     1181.40
    22630      643.80
    22629      643.80
    22423      599.25


Lithuania:

StockCode  TotalPrice
    20967       135.0
    22271       122.4
    22750       120.0
    22751       105.0
    22569        90.0


Japan:

StockCode  TotalPrice
    23084     6100.32
    22328     3812.10
    21218      858.00
    20750      762.00
    21217      751.80


Iceland:

StockCode  TotalPrice
   84558A      371.70
    23076      249.60
    22423      191.25
    23084      153.84
    22727      135.00


Channel Islands:

StockCode  TotalPrice
    22423       517.8
   85099B       460.7
    22720       408.0
   85099C       399.6
    23199       399.6


Denmark:

StockCode  TotalPrice
```

```
      POST      744.00
     22625      734.40
     22624      696.15
     22467      575.10
     22179      428.40


Cyprus:

StockCode  TotalPrice
    22827      580.00
   15056N      392.70
   85123A      386.40
    22423      382.50
        M      320.69


Sweden:

StockCode  TotalPrice
    22492     1895.40
    22720     1767.15
     POST     1509.00
    23297     1240.80
   85232B     1188.00


Finland:

StockCode  TotalPrice
     POST     3650.00
   84997D     2063.28
   84997C     1367.40
   84997A      919.60
        M      551.20


Austria:

StockCode  TotalPrice
     POST      1456.0
    22584       302.4
    22582       302.4
    20679       214.2
   15056N       214.2


Bahrain:

StockCode  TotalPrice
   72802B      231.24
    23076      120.00
    23077       75.00
    22890       59.70
    22649       39.60


Israel:

StockCode  TotalPrice
    22423      551.10
    23240      254.70
    22326      244.80
    22192      183.60
    23236      159.98


Greece:

StockCode  TotalPrice
     POST      335.00
    22423      175.20
   72760B      135.84
    22692      135.00
    48129      135.00


Hong Kong:

StockCode  TotalPrice
        M     5563.81
   84997D      488.40
   84997B      455.20
    22326      384.90
    22452      318.60


Singapore:

StockCode  TotalPrice
        M     12158.9
    48138       340.8
    22655       250.0
    22197       216.0
    20685       205.8


Lebanon:

StockCode  TotalPrice
    22423       153.0
    85066       102.0
    22606       102.0
    21906        81.0
    22842        71.4


United Arab Emirates:
```

```
StockCode  TotalPrice
    22423       153.0
    23007        89.7
    23008        89.7
    23009        89.7
   47590B        65.4


Saudi Arabia:

StockCode  TotalPrice
    22553        19.8
    22555        19.8
    22556        19.8
    22361        17.7
    22362        17.7


Czech Republic:

StockCode  TotalPrice
    22326        70.8
    84347        61.2
    22231        52.2
    21428        51.0
   47594B        46.8


Canada:

StockCode  TotalPrice
     POST      550.94
    37370      534.24
    20727       82.50
    84077       60.48
    22383       49.50


Unspecified:

StockCode  TotalPrice
    22960       70.50
    23236       69.36
    23234       69.36
    23076       60.00
    22138       54.45


Brazil:

StockCode  TotalPrice
    22423      175.20
    22722       82.80
    21430       81.36
    22366       67.50
    22699       61.20


USA:

StockCode  TotalPrice
    23328      162.72
    22423      114.75
    21121       90.00
    21122       90.00
    21123       90.00


European Community:

StockCode  TotalPrice
     POST       141.0
    22843        54.0
    22842        54.0
    22314        53.1
   85036B        51.0


Malta:

StockCode  TotalPrice
     POST      655.00
    72741      117.45
    22423       89.25
    23173       79.60
    22796       59.70


RSA:

StockCode  TotalPrice
    21340       38.25
    22605       29.90
    23298       29.70
    22526       25.50
    85066       25.50
```

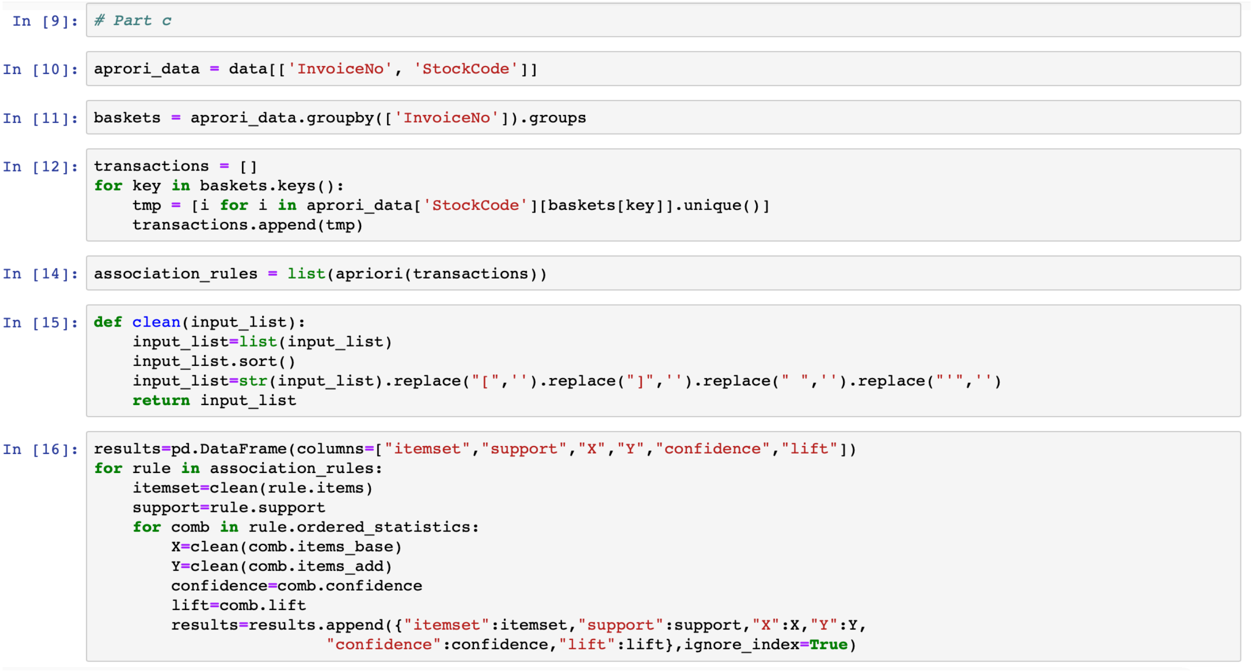
c.
Results:



	itemset	support	X	Y	confidence	lift
0	85099B	0.104659		85099B	0.104659	1.0
1	85123A	0.110120		85123A	0.110120	1.0

We only find two rules. The result implies that no matter what customer buys, we should always recommend item 85099B and 85123A.

Code:



```
In [9]:  # Part c

In [10]: aprori_data = data[['InvoiceNo', 'StockCode']]

In [11]: baskets = aprori_data.groupby(['InvoiceNo']).groups

In [12]: transactions = []
         for key in baskets.keys():
             tmp = [i for i in aprori_data['StockCode'][baskets[key]].unique()]
             transactions.append(tmp)

In [14]: association_rules = list(apriori(transactions))

In [15]: def clean(input_list):
             input_list=list(input_list)
             input_list.sort()
             input_list=str(input_list).replace("[",'').replace("]",'').replace(" ",'').replace("'",'')
             return input_list

In [16]: results=pd.DataFrame(columns=["itemset","support","X","Y","confidence","lift"])
         for rule in association_rules:
             itemset=clean(rule.items)
             support=rule.support
             for comb in rule.ordered_statistics:
                 X=clean(comb.items_base)
                 Y=clean(comb.items_add)
                 confidence=comb.confidence
                 lift=comb.lift
                 results=results.append({"itemset":itemset,"support":support,"X":X,"Y":Y,
                                 "confidence":confidence,"lift":lift},ignore_index=True)
```