In [1]:
```
import matplotlib.pyplot as plt; plt.rcdefaults()
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import warnings
warnings.filterwarnings('ignore')
```

In [2]:
```
df = pd.read_csv('Workshop-5-dataset.zip', sep='\t',dtype=np.str)
```

Q1 answer

In [3]:
```
df.head()
```

Out[3]:

	transaction_ID	Date	Time	item_0	item_1	item_2	item_3	item_4	item_5	item_6	...	item_31	item_3
0	536365	01/12/2010	08:26	WHITE HANGING HEART T-LIGHT HOLDER	WHITE METAL LANTERN	CREAM CUPID HEARTS COAT HANGER	KNITTED UNION FLAG HOT WATER BOTTLE	RED WOOLLY HOTTIE WHITE HEART	SET 7 BABUSHKA NESTING BOXES	GLASS STAR FROSTED T-LIGHT HOLDER	...	NaN	Nal
1	536366	01/12/2010	08:28	HAND WARMER UNION JACK	HAND WARMER RED POLKA DOT	NaN	NaN	NaN	NaN	NaN	...	NaN	Nal
2	536367	01/12/2010	08:34	ASSORTED COLOUR BIRD ORNAMENT	POPPY'S PLAYHOUSE BEDROOM	POPPY'S PLAYHOUSE KITCHEN	FELTCRAFT PRINCESS CHARLOTTE DOLL	IVORY KNITTED MUG COSY	BOX OF 6 ASSORTED COLOUR TEASPOONS	BOX OF VINTAGE JIGSAW BLOCKS	...	NaN	Nal
3	536368	01/12/2010	08:34	JAM MAKING SET WITH JARS	RED COAT RACK PARIS FASHION	YELLOW COAT RACK PARIS FASHION	BLUE COAT RACK PARIS FASHION	NaN	NaN	NaN	...	NaN	Nal
4	536369	01/12/2010	08:35	BATH BUILDING BLOCK WORD	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	Nal

5 rows × 44 columns

Q2 answer

In [4]:
```
df.shape
```

Out[4]:
```
(31941, 44)
```

The dataset has 44 columns and 31941 rows

Genearting your unique Dataset

In [5]:
```
STUDENT_NAME = 'PreciousAdaugoReginald'
STUDENT_NO = '5671'
```

In [6]:
```
np.random.seed(int(STUDENT_NO))
unique_id = int('2' + STUDENT_NO)
rows = np.random.choice(df.index.values, unique_id)
data = df.loc[rows]
```

In [7]:
```
file_name = STUDENT_NAME + "_" + STUDENT_NO + ".csv"
data.to_csv(file_name)
```

Basic data analysis

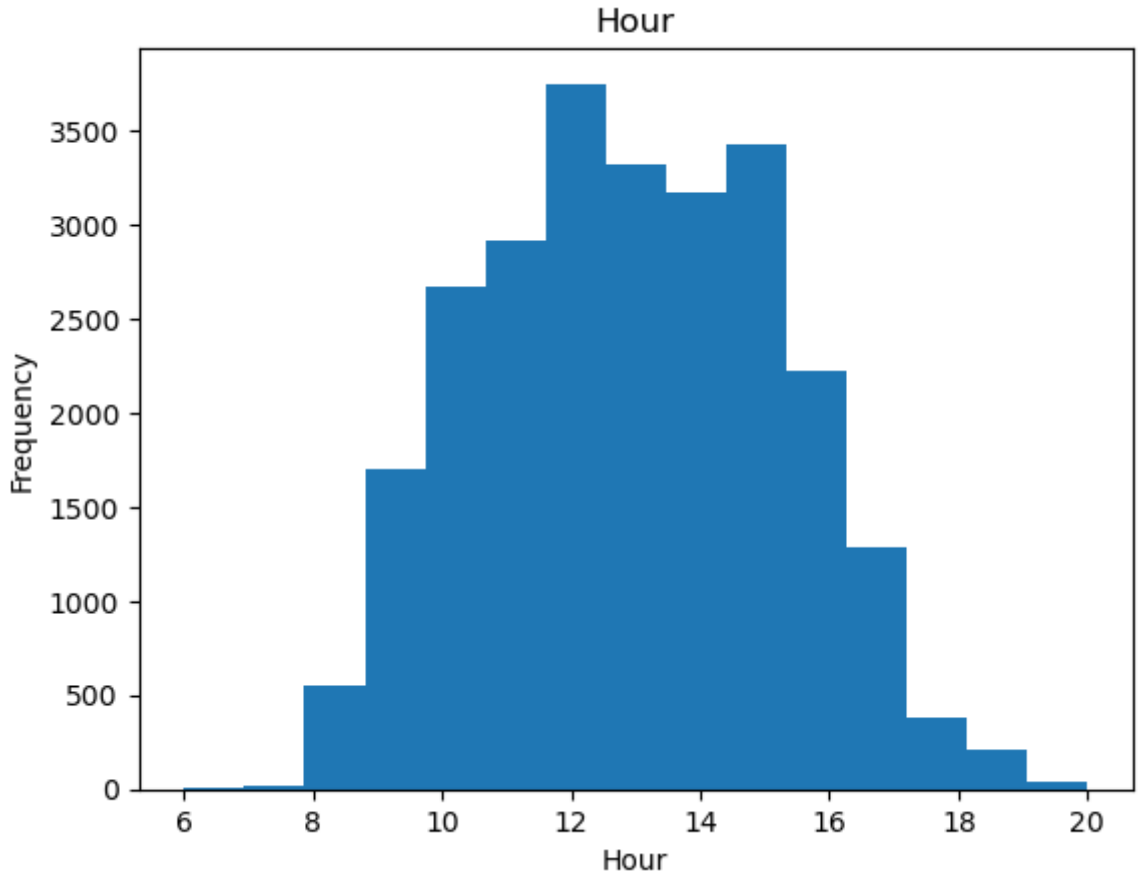
Q3 answer

In [ ]:
```

```

In [8]:
```
data['Hour'] = pd.to_datetime(data['Time'], format='%H:%M').dt.hour
```

In [9]:
```
hour_hist = data.hist(column="Hour", bins=15, grid=False)
for ax in hour_hist.flatten():
 ax.set_xlabel("Hour")
 ax.set_ylabel("Frequency")
```



Apriori algorithm

In [10]:
```
# import apyori
from apyori import apriori
```

Data preprocessing

In [11]:
```
data.head(1)
```

Out[11]:

	transaction_ID	Date	Time	item_0	item_1	item_2	item_3	item_4	item_5	item_6	...	item_32	item_33	item_34	item_35	item_
7561	547225	21/03/2011	15:20	CUPID DESIGN SCENTED CANDLES	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	Na

1 rows × 45 columns

In [12]:
```
items_df=data[data.columns[3:44]]
```

In [13]:
```
items_df.head()
```

Out[13]:

	item_0	item_1	item_2	item_3	item_4	item_5	item_6	item_7	item_8	iten
7561	CUPID DESIGN SCENTED CANDLES	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
5554	PINK 3 PIECE POLKADOT CUTLERY SET	BLUE 3 PIECE POLKADOT CUTLERY SET	RED 3 PIECE RETROSPOT CUTLERY SET	GREEN 3 PIECE POLKADOT CUTLERY SET	SET/3 RED GINGHAM ROSE STORAGE BOX	CHILLI LIGHTS	JUMBO BAG PINK POLKADOT	STRAWBERRY LUNCH BOX WITH CUTLERY	LUNCH BOX WITH CUTLERY FAIRY CAKES	LUNCH B W CUTLE RETROSP
30392	BLUE DINER WALL CLOCK	HENRIETTA HEN MUG	RED EGG SPOON	PAPER CHAIN KIT RETROSPOT	RED RETROSPOT SMALL MILK JUG	BOX OF 6 ASSORTED COLOUR TEASPOONS	COFFEE MUG DOG + BALL DESIGN	COFFEE MUG CAT + BIRD DESIGN	CHERRY BLOSSOM DECORATIVE FLASK	VINTA R ENAM TRIM J
31789	SET OF 3 WOODEN TREE DECORATIONS	COFFEE MUG DOG + BALL DESIGN	SMALL GLASS HEART TRINKET POT	BREAD BIN DINER STYLE IVORY	GLITTER CHRISTMAS STAR	SET/5 RED RETROSPOT LID GLASS BOWLS	WHITE ROCKING HORSE HAND PAINTED	MILK MAIDS MUG	HENRIETTA HEN MUG	ZI FOLKA SLEI BEL
13396	GREEN METAL BOX ARMY SUPPLIES	RED METAL BOX TOP SECRET	SET 6 PAPER TABLE LANTERN STARS	LED TEA LIGHTS	PAPER LANTERN 7 POINT SNOW STAR	PAPER LANTERN 9 POINT SNOW STAR	GARDENERS KNEELING PAD CUP OF TEA	IVORY CAFE HANGING LAMP	MIRROR MOSAIC VOTIVE HOLDER	ICE CRE SUNDAE GLC

5 rows × 41 columns

In [14]:
```
baskets = items_df.T.apply(lambda x: x.dropna().tolist()).tolist()
```

In [15]:
```
for i in baskets[:5]:
 print(i)
```

```
['CUPID DESIGN SCENTED CANDLES']
['PINK 3 PIECE POLKADOT CUTLERY SET', 'BLUE 3 PIECE POLKADOT CUTLERY SET', 'RED 3 PIECE RETROSPOT CUTLERY SET', 'GRE
EN 3 PIECE POLKADOT CUTLERY SET', 'SET/3 RED GINGHAM ROSE STORAGE BOX', 'CHILLI LIGHTS', 'JUMBO BAG PINK POLKADOT',
'STRAWBERRY LUNCH BOX WITH CUTLERY', 'LUNCH BOX WITH CUTLERY FAIRY CAKES', 'LUNCH BOX WITH CUTLERY RETROSPOT', 'WOOD
LAND CHARLOTTE BAG', 'CHARLOTTE BAG PINK POLKADOT', 'RED RETROSPOT CHARLOTTE BAG', 'RED TOADSTOOL LED NIGHT LIGHT',
'SET/2 RED RETROSPOT TEA TOWELS', 'CHARLOTTE BAG SUKI DESIGN', 'LUNCH BAG SUKI  DESIGN', 'JUMBO STORAGE BAG SUKI',
'JUMBO STORAGE BAG SKULLS', 'JUMBO BAG WOODLAND ANIMALS', 'JUMBO BAG OWLS', 'JUMBO BAG RED RETROSPOT']
['BLUE DINER WALL CLOCK', 'HENRIETTA HEN MUG', 'RED  EGG  SPOON', 'PAPER CHAIN KIT RETROSPOT', 'RED RETROSPOT SMALL
MILK JUG', 'BOX OF 6 ASSORTED COLOUR TEASPOONS', 'COFFEE MUG DOG + BALL DESIGN', 'COFFEE MUG CAT + BIRD DESIGN', 'CH
ERRY BLOSSOM  DECORATIVE FLASK', 'VINTAGE RED ENAMEL TRIM JUG', "BISCUIT TIN 50'S CHRISTMAS", 'ILLUSTRATED CAT BOW
L', 'DOG BOWL CHASING BALL DESIGN', 'RED FLOWER CROCHET FOOD COVER', 'PSYCHEDELIC TILE COASTER', 'CHILLI LIGHTS', 'M
IRRORED DISCO BALL', 'CARD BILLBOARD FONT', 'BANQUET BIRTHDAY  CARD']
['SET OF 3 WOODEN TREE DECORATIONS', 'COFFEE MUG DOG + BALL DESIGN', 'SMALL GLASS HEART TRINKET POT', 'BREAD BIN DIN
ER STYLE IVORY', 'GLITTER CHRISTMAS STAR', 'SET/5 RED RETROSPOT LID GLASS BOWLS', 'WHITE ROCKING HORSE HAND PAINTE
D', 'MILK MAIDS MUG', 'HENRIETTA HEN MUG', 'ZINC FOLKART SLEIGH BELLS', 'SWALLOW WOODEN CHRISTMAS DECORATION', 'HEAR
T WOODEN CHRISTMAS DECORATION', 'PACK OF 12 VINTAGE DOILY TISSUES', 'PACK OF 12 VINTAGE LEAF TISSUES', 'PACK OF 12 S
KULL TISSUES', 'PACK OF 12 SUKI TISSUES', 'PACK OF 12 SPACEBOY TISSUES', "JUMBO BAG 50'S CHRISTMAS", 'JUMBO BAG VINT
AGE CHRISTMAS', 'VINTAGE CHRISTMAS BUNTING', 'CLASSIC CHROME BICYCLE BELL', 'BAKING MOULD HEART MILK CHOCOLATE', 'JI
NGLE BELL HEART DECORATION', 'BULL DOG BOTTLE OPENER', 'HEART WREATH DECORATION WITH BELL', 'RED HANGING HEART T-LIG
HT HOLDER', 'CLASSIC BICYCLE CLIPS', 'BICYCLE PUNCTURE REPAIR KIT', 'PLASTERS IN TIN WOODLAND ANIMALS', 'PLASTERS IN
TIN SPACEBOY', 'DOG BOWL VINTAGE CREAM', 'FRYING PAN RED RETROSPOT', 'FRYING PAN PINK POLKADOT', 'FRYING PAN BLUE PO
LKADOT', 'FRYING PAN UNION FLAG', 'MILK PAN RED RETROSPOT', 'PLASTERS IN TIN CIRCUS PARADE', 'FRYING PAN RED RETROSP
OT', 'MILK PAN BLUE POLKADOT', 'CHILDRENS CUTLERY SPACEBOY', 'CHILDRENS CUTLERY CIRCUS PARADE']
['GREEN METAL BOX ARMY SUPPLIES', 'RED METAL BOX TOP SECRET', 'SET 6 PAPER TABLE LANTERN STARS', 'LED TEA LIGHTS',
'PAPER LANTERN 7 POINT SNOW STAR', 'PAPER LANTERN 9 POINT SNOW STAR', 'GARDENERS KNEELING PAD CUP OF TEA', 'IVORY CA
FE HANGING LAMP', 'MIRROR MOSAIC VOTIVE HOLDER', 'ICE CREAM SUNDAE LIP GLOSS', 'DOUGHNUT LIP GLOSS', 'STAR T-LIGHT H
OLDER WILLIE WINKIE', 'BOX OF 24 COCKTAIL PARASOLS', 'SET OF 10 LANTERNS FAIRY LIGHT STAR', 'SPOTTY BUNTING', 'MIRRO
R MOSAIC T-LIGHT HOLDER', 'POTTERING IN THE SHED METAL SIGN', 'PARTY METAL SIGN', 'SET 12 LAVENDER  BOTANICAL T-LIGH
TS', 'SMALL MEDINA STAMPED METAL BOWL', 'TRAVEL SEWING KIT', 'T-LIGHT GLASS FLUTED ANTIQUE', 'ZINC T-LIGHT HOLDER ST
AR LARGE', 'ALUMINIUM STAMPED HEART', 'ASSORTED COLOUR BIRD ORNAMENT', 'HANGING HEART JAR T-LIGHT HOLDER', 'SILVER P
LATE CANDLE BOWL SMALL', 'S/2 ZINC HEART DESIGN PLANTERS', 'LUNCH BAG DOILEY PATTERN', 'LUNCH BAG PINK POLKADOT', 'J
UMBO STORAGE BAG SKULLS', 'JUMBO BAG PEARS', 'JUMBO BAG APPLES', 'JUMBO BAG DOILEY PATTERNS', 'JUMBO SHOPPER VINTAGE
RED PAISLEY', 'JUMBO BAG PINK POLKADOT', 'JUMBO BAG STRAWBERRY', 'JUMBO  BAG BAROQUE BLACK WHITE', 'STRAWBERRY  PIC
NIC BAG', 'RED RETROSPOT PICNIC BAG', 'SCANDINAVIAN PAISLEY PICNIC BAG']
```

Algorithm Parameters

In [16]:
```
association_rules = apriori(baskets, min_support=0.01, min_confidence=0.2,
 min_lift=3, min_length=2)
association_results = list(association_rules)
```

Generation of association rules

In [17]:
```
print('Rules generated: ', len(association_results))
```

```
Rules generated:  91
```

Overvew of how the association rules looks like for the algorithm. The first rule for the algorithm will be printed below

In [18]:
```
print(association_results[0])
```

```
RelationRecord(items=frozenset({'60 TEATIME FAIRY CAKE CASES', 'PACK OF 72 RETROSPOT CAKE CASES'}), support=0.010673
522652019788, ordered_statistics=[OrderedStatistic(items_base=frozenset({'60 TEATIME FAIRY CAKE CASES'}), items_add=
frozenset({'PACK OF 72 RETROSPOT CAKE CASES'}), confidence=0.408955223880597, lift=9.640302619135728), OrderedStatis
tic(items_base=frozenset({'PACK OF 72 RETROSPOT CAKE CASES'}), items_add=frozenset({'60 TEATIME FAIRY CAKE CASES'}),
confidence=0.2516069788797062, lift=9.640302619135728)])
```

A different rule can be shown below. For that, the index value [ ] for association_results[ ] has ben changed to 9

In [19]:
```
print(association_results[9])
```

```
RelationRecord(items=frozenset({'DOLLY GIRL LUNCH BOX', 'SPACEBOY LUNCH BOX'}), support=0.015854466129095086, ordere
d_statistics=[OrderedStatistic(items_base=frozenset({'DOLLY GIRL LUNCH BOX'}), items_add=frozenset({'SPACEBOY LUNCH
BOX'}), confidence=0.5708274894810659, lift=19.643046223148044), OrderedStatistic(items_base=frozenset({'SPACEBOY LU
NCH BOX'}), items_add=frozenset({'DOLLY GIRL LUNCH BOX'}), confidence=0.5455764075067023, lift=19.643046223148044)])
```

Analysing the results

In [20]:
```
def display_rules(association_results):
    for item in association_results:
        pair = item[0]
        items = [x for x in pair]
        print("Rule: " + items[0] + " -> " + items[1])
        print("Support: " + str(item[1]))
        print("Confidence: " + str(item[2][0][2]))
        print("Lift: " + str(item[2][0][3]))
        print("====================================")
```

In [21]:
```
display_rules(association_results[:5])
```

```
Rule: 60 TEATIME FAIRY CAKE CASES -> PACK OF 72 RETROSPOT CAKE CASES
Support: 0.010673522652019788
Confidence: 0.408955223880597
Lift: 9.640302619135728
====================================
Rule: ALARM CLOCK BAKELIKE PINK -> ALARM CLOCK BAKELIKE GREEN
Support: 0.013789879630711698
Confidence: 0.4154929577464789
Lift: 15.369048585460893
====================================
Rule: ALARM CLOCK BAKELIKE GREEN -> ALARM CLOCK BAKELIKE RED
Support: 0.020295274823731058
Confidence: 0.6115023474178404
Lift: 17.838496318822024
====================================
Rule: ALARM CLOCK BAKELIKE RED -> ALARM CLOCK BAKELIKE IVORY
Support: 0.011179930661057224
Confidence: 0.6198704103671706
Lift: 18.082606027881404
====================================
Rule: ALARM CLOCK BAKELIKE PINK -> ALARM CLOCK BAKELIKE RED
Support: 0.015659693817926843
Confidence: 0.5792507204610952
Lift: 16.897665051087245
====================================
```

In [22]:
```
from collections import Counter

counter = Counter(baskets[0])
for i in baskets[1:]:
    if i != 'nan':
        counter.update(i)

del counter['nan']
counter.most_common(10)
```

Out[22]:
```
[('WHITE HANGING HEART T-LIGHT HOLDER', 1848),
 ('JUMBO BAG RED RETROSPOT', 1838),
 ('REGENCY CAKESTAND 3 TIER', 1821),
 ('PARTY BUNTING', 1380),
 ('LUNCH BAG RED RETROSPOT', 1357),
 ('SET OF 3 CAKE TINS PANTRY DESIGN', 1172),
 ('ASSORTED COLOUR BIRD ORNAMENT', 1135),
 ('PACK OF 72 RETROSPOT CAKE CASES', 1113),
 ('LUNCH BAG  BLACK SKULL', 1061),
 ('JUMBO BAG PINK POLKADOT', 1045)]
```

Q4 (a) answer

In the rules I have just dispalyed I can only find 1 item which is pack of 72 retrospot cases

Q4 (b) answer

Not all the 10 top items are included. This is because for certian rules the minimum confidence and minimum support is below the the given values of 0.2 and 0.01 respectively

Report

Q5 Setting 1

Min Support = 0.015, Min Confidence = 0.7, Min Lift = 3

In [23]:
```
association_rules = apriori(baskets, min_support=0.015, min_confidence=0.7,
 min_lift=3, min_length=2)
association_results = list(association_rules)
```

In [24]:
```
print('Rules generated: ', len(association_results))
```

```
Rules generated:  5
```

Q5 setting 2

Min Support = 0.009, Min Confidence = 0.5, Min Lift = 3

In [25]:
```
association_rules = apriori(baskets, min_support=0.009, min_confidence=0.5,
 min_lift=3, min_length=2)
association_results = list(association_rules)
```

In [26]:
```
print('Rules generated: ', len(association_results))
```

```
Rules generated:  55
```

Q5 setting 3

Min Support = 0.015, Min Confidence = 0.5, Min Lift = 9

In [27]:
```
association_rules = apriori(baskets, min_support=0.015, min_confidence=0.5,
 min_lift=9, min_length=2)
association_results = list(association_rules)
```

In [28]:
```
print('Rules generated: ', len(association_results))
```

```
Rules generated:  12
```

after having changed they support minimum confidence and lift it can be seen but the number of rules generated changes each time for each setting an example is that's when minimum support east 0.015 and the confidence is 0.7 the rules generated are less due to both values having gone up compared to day 0.01 and 0.2 of the respective minimum support and minimum confidence

Q6 answer

In [ ]:
```
```

Report

Association rule mining involves the use of apriory algorithm which is a machine learning model used to analyse data to look on patterns in a database'

In this particular data set the transactions that are used for association rule mining are the ones that have minimum confidence and minimum support equal to or above the threshold given.

With this specific data, a personalised data set was produced Then just before the apriory algorithm was applied some data pre-processing steps we are carried out. This involved the removal of order unnecessary columns and by leaving out only the description columns and invoice numbers and also reducing the data frame into only 41 columns because the algorithm can only be used for a small data set.

By setting specific values for minimum support and minimum confidence and is and is a specified lift a number of rules which in this case is 91 is generated. In addition a frequent item sets can be found by using a so-called candidate generation. To conlcude when it comes to rules generated the number 91 represents the value for the rules that have minimum support and minimum confidence equal to or above the threshold. An example is found for the first rule in cell 22, where support is above 0.01 and confidence is above 0.04.