1. Name and Uni:
   I. PING-FENG LIN, pl2730
   II. Weihan Chen, wc2681
2. List of all the files:
   I. proj3.tar.gz
   II. INTEGRATED-DATASET.csv
   III. README.pdf
   IV. Example-run.txt
3. How to run the program:
   I. Uncompress the proj3.tar.gz to obtain Project3.py
   II. Put Project3.py and INTEGRATED-DATASET.csv under the same folder
   III. At that folder, input in the command line:

```
python3 Project3.py INTEGRATED-DATASET.csv min_sup min_conf
```

   IV. There will be a file named output.txt generated in the folder
4. Description of the dataset:
   I. Which NYC Open Data data set: NYPD Complaint Data Historic, https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i
   II. What procedure needs to make INTEGRATED-DATASET.csv:

In python3, under the folder where NYPD_Complaint_Data_Historic.csv is at, input the following commands:

```
import pandas
# load dataset, and only take the 11 columns that we want
data = pandas.read_csv("./NYPD_Complaint_Data_Historic.csv")
data = data[['CMPLNT_FR_DT', 'CMPLNT_FR_TM', 'OFNS_DESC', 'BORO_NM',
'PREM_TYP_DESC', 'SUSP_AGE_GROUP', 'SUSP_RACE', 'SUSP_SEX',
'VIC_AGE_GROUP', 'VIC_RACE', 'VIC_SEX']]
# drop rows with missing value and UNKNOWN
data = data.dropna()
data = data[data['SUSP_AGE_GROUP'] != 'UNKNOWN']
data = data[data['SUSP_RACE'] != 'UNKNOWN']
data = data[data['SUSP_SEX'] != 'U']
data = data[data['VIC_AGE_GROUP'] != 'UNKNOWN']
data = data[data['VIC_RACE'] != 'UNKNOWN']
data = data[data['VIC_SEX'] != 'U']
# change the date to only year, change the time to 3 period (morning, afternoon,
night)
```

```
data['CMPLNT_FR_DT'] = pandas.to_datetime(data['CMPLNT_FR_DT'],
errors='coerce')
data['CMPLNT_FR_DT'] = data['CMPLNT_FR_DT'].dt.year
data['CMPLNT_FR_TM'] = pandas.to_datetime(data['CMPLNT_FR_TM'],
errors='coerce')
data['CMPLNT_FR_TM'] = data['CMPLNT_FR_TM'].dt.hour
data['CMPLNT_FR_TM'] = (data['CMPLNT_FR_TM']%24+10)//8
data['CMPLNT_FR_TM'].replace({1:'Night', 2:'Morning', 3:'Afternoon', 4:'Night'},
inplace=True)
# change notation to discriminate SUSP attributes and VIC attributes
data['SUSP_AGE_GROUP'] = 'SUSP_A_' + data['SUSP_AGE_GROUP']
data['SUSP_RACE'] = 'SUSP_R_' + data['SUSP_RACE']
data['SUSP_SEX'] = 'SUSP_S_' + data['SUSP_SEX']
data['VIC_AGE_GROUP'] = 'VIC_A_' + data['VIC_AGE_GROUP']
data['VIC_RACE'] = 'VIC_R_' + data['VIC_RACE']
data['VIC_SEX'] = 'VIC_S_' + data['VIC_SEX']
# reduce the dataset size to around 2M rows by random sampling
data = data.sample(frac=0.2, replace=True, random_state=1)
# save the file
data.to_csv('INTEGRATED-DATASET.csv', index=False)
```

III. what makes your choice: we believe by investigating these data, we can discovery useful information that can facilitate NYPD’s work on tracking down criminals. Also, by knowing the insight of the data, we can avoid ourselves from being in danger.

5. Description of the internal design:

Our implementation follows the algorithm shown in Section 2.1.1 of the Agrawal and Srikant paper. Starting with single item, discard item that has support value less than min_sup. Next, pair two items together from the previous set of items, discard itemsets that have support values less than min_sup. For the further higher dimension itemset, we first pick two itemsets from the previous level collection, check that do they differ from each other only by one element, and test their support. Discard the itemsets that do not meet the criterion, keep doing until encounter empty collection, finally form the Frequent itemsets.

High-confidence association rules: take itemsets from the Frequent itemsets and calculate confidence for relation [itemset(remove one element)] => element_removed by (supp. of itemset / supp. of itemset(remove one element)). Discard rules with confidence less than min_conf.

6. The command line specification of a compelling sample run:

```
python3 Project3.py INTEGRATED-DATASET.csv 0.05 0.8
```

It returns nearly 800 Frequent itemsets and 50 High-confidence association rule. And we found it produced a sufficient result indicating about relations between suspect and victim, places where the criminal happened and time, races and gender, and so on.