

Name:Rhea Adhikari  
190905156  
Lab5  
DS Lab

Exercise 1 - Try the above word count program for the Heart Disease dataset, covid\_19\_data dataset, example dataset and German Credit dataset.

Mapper.py

```
import sys
cnt = 0
for line in sys.stdin:
    if cnt < 10:
        line = line.strip()
        words = line.split()
        for word in words:
            print("%s\t\t%s" %(word, 1))
        cnt+=1
    else :
        break
```

Reducer.py

```
from operator import itemgetter
import sys
current_word = None
current_count = 0
word = None
for line in sys.stdin:
    line = line.strip()
    word, count = line.split('\t', 1)
    try:
        count = int(count)
    except ValueError:
        continue
    if current_word == word:
        current_count += count
    else:
        if current_word:
            print('%s\t%s' % (current_word, current_count) )
        current_count = count
        current_word = word
    if current_word == word:
        print('%s\t%s' % (current_word, current_count))
```

```
Q Applications Wed, Apr 20 00:25
q1reducer.py - 190905156_DS - Visual Studio Code
File Edit Selection View Go Run Terminal Help

q1mapper.py x q1reducer.py x q1mapper.py
Lab5 > q1mapper.py > ...
1 import sys
2 cnt = 0
3 for line in sys.stdin:
4     if cnt < 10:

Lab5 > q1reducer.py > ...
1 from operator import itemgetter
2 import sys
3 current_word = None
4 current_count = 0

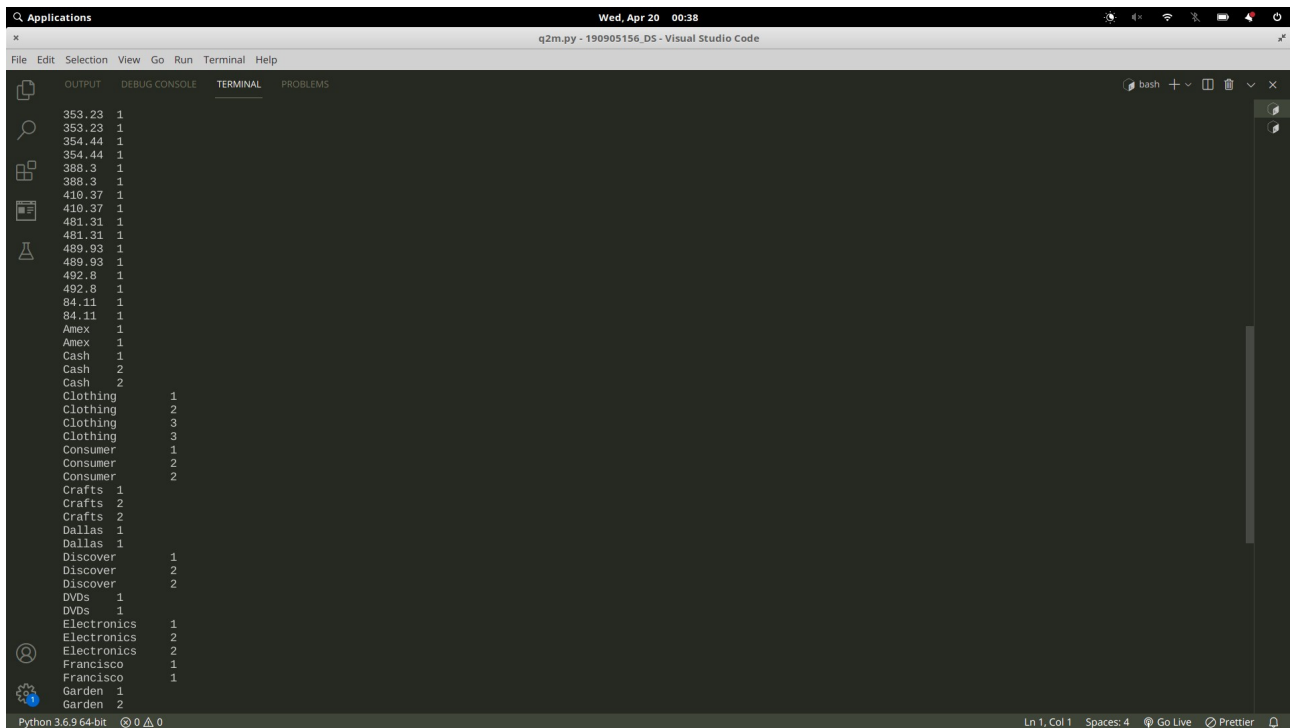
OUTPUT DEBUG CONSOLE TERMINAL PROBLEMS
covid_19_data.csv heart_disease_data.csv q1mapper.py q2freqmap2.py q3itemmap.py q4sepred.py q6mapper.py q8mapreduce.py
example.txt MapReduce_Lab5.docx q1reducer.py q2freqread1.py q3itemred.py q5map.py q6reducer.py
'German Credit.xlsx' MapReduce_Lab5.pptx q2freqmap1.py q2freqread2.py q4semap.py q5red.py q7mapreduce.py
rhea@rhea-Lenovo-IdeaPad-S145-15IWL:~/data/Lab/190905156_DS/Lab5$ cat heart_disease_data.csv | python3 q1mapper.py | sort |python3 q1reducer.py
File "q1reducer.py", line 6
for line in sys.stdin:
^
IndentationError: unexpected indent
rhea@rhea-Lenovo-IdeaPad-S145-15IWL:~/data/Lab/190905156_DS/Lab5$ cat heart_disease_data.csv | python3 q1mapper.py | sort |python3 q1reducer.py
37,1,2,130,250,0,1,187,0,3.5,0,0,2,Yes 1
37,1,2,130,250,0,1,187,0,3.5,0,0,2,Yes 1
41,0,1,130,204,0,0,172,0,1.4,2,0,2,Yes 1
41,0,1,130,204,0,0,172,0,1.4,2,0,2,Yes 1
44,1,1,120,263,0,1,173,0,0,2,0,3,Yes 1
44,1,1,120,263,0,1,173,0,0,2,0,3,Yes 1
52,1,2,172,199,1,1,162,0,0.5,2,0,3,Yes 1
52,1,2,172,199,1,1,162,0,0.5,2,0,3,Yes 1
56,0,1,140,294,0,0,153,0,1.3,1,0,2,Yes 1
56,0,1,140,294,0,0,153,0,1.3,1,0,2,Yes 1
56,1,1,120,236,0,1,178,0,0.8,2,0,2,Yes 1
56,1,1,120,236,0,1,178,0,0.8,2,0,2,Yes 1
57,0,0,120,354,0,1,163,1,0.6,2,0,2,Yes 1
57,0,0,120,354,0,1,163,1,0.6,2,0,2,Yes 1
57,1,0,140,192,0,1,148,0,0.4,1,0,1,Yes 1
57,1,0,140,192,0,1,148,0,0.4,1,0,1,Yes 1
63,1,3,145,233,1,0,150,0,2.3,0,0,1,Yes 1
63,1,3,145,233,1,0,150,0,2.3,0,0,1,Yes 1
age,sex,cp,trestbps,chol,fb,restecg,thalach,exang,oldpeak,slope,ca,thal,target 1
rhea@rhea-Lenovo-IdeaPad-S145-15IWL:~/data/Lab/190905156_DS/Lab5$
Python 3.6.9 64-bit 0 0 0 15 selections Spaces: 4 Go Live Prettier
```

## GermanCredit.csv

```
"1","1049","18" 1
"1","1049","18" 1
"1","1098","18" 1
"1","1098","18" 1
"1","1361","6" 1
"1","1361","6" 1
"1","2122","12" 1
"1","2122","12" 1
"1","2171","12" 1
"1","2171","12" 1
"1","2241","10" 1
"1","2241","10" 1
"1","2799","9" 1
"1","2799","9" 1
"1","3398","8" 1
"1","3398","8" 1
"1","841","12" 1
"1","841","12" 1
"Creditability","CreditAmount","DurationOfCreditInMonths"
```

```
Q Applications Wed, Apr 20 00:38
q2m.py - 190905156_DS - Visual Studio Code
File Edit Selection View Go Run Terminal Help

rhea@rhea-Lenovo-IdeaPad-S145-15IWL:~/data/Lab/190905156_DS/Lab5$ cat example.txt | python3 q1mapper.py | sort |python3 q1reducer.py
09:01 1
09:01 1
11:35 1
11:35 1
12:58 1
12:58 1
13:02 1
13:02 1
13:12 1
13:12 1
145,63 1
145,63 1
15:01 1
15:01 1
15:30 1
15:30 1
15:43 1
15:43 1
16:17 1
16:17 1
16:34 1
16:34 1
2012-04-22 1
2012-04-22 1
2012-06-11 1
2012-06-11 1
2012-07-13 1
2012-07-13 1
2012-07-16 1
2012-07-16 1
2012-08-05 1
2012-08-05 1
2012-09-07 1
2012-09-07 1
2012-10-17 1
2012-10-17 1
2012-10-19 1
2012-10-19 1
2012-10-25 1
2012-10-25 1
2012-11-06 1
2012-11-06 1
208,97 1
208,97 1
Python 3.6.9 64-bit 0 0 0 Ln 1, Col 1 Spaces: 4 Go Live Prettier
```

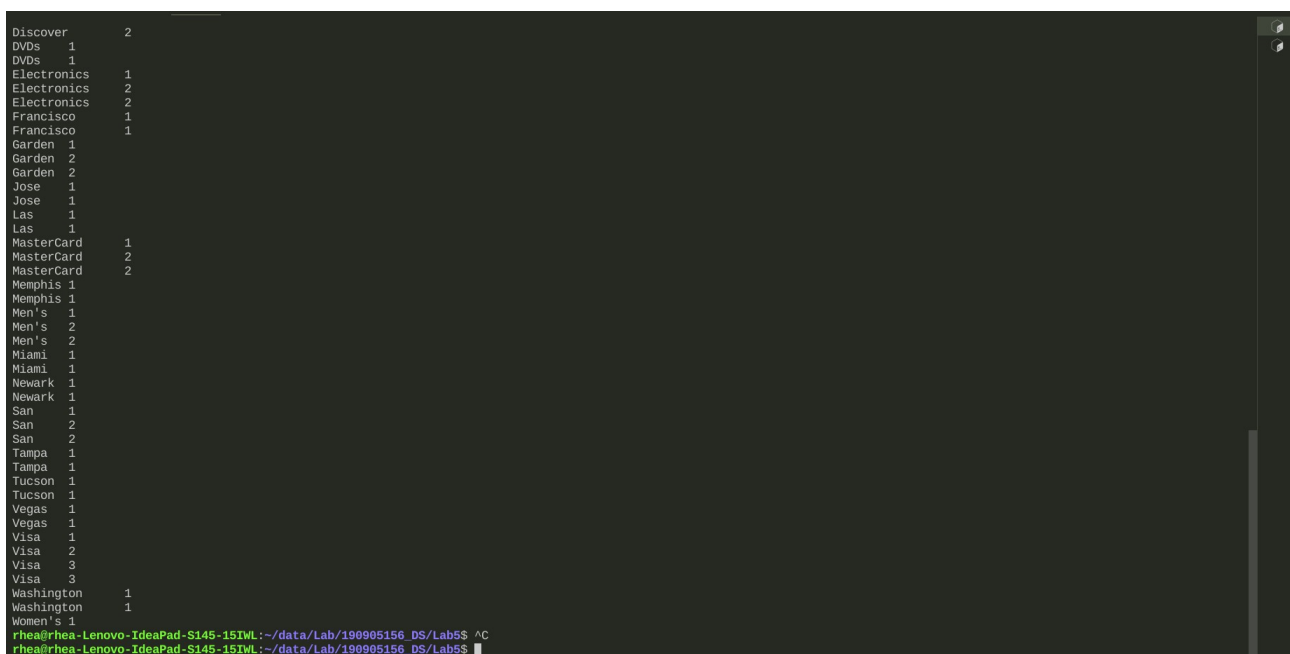


```
File Edit Selection View Go Run Terminal Help
q2m.py - 190905156_DS - Visual Studio Code

OUTPUT DEBUG CONSOLE TERMINAL PROBLEMS
bash

353.23 1
353.23 1
354.44 1
354.44 1
388.3 1
388.3 1
410.37 1
410.37 1
481.31 1
481.31 1
489.93 1
489.93 1
492.8 1
492.8 1
84.11 1
84.11 1
Amex 1
Amex 1
Cash 1
Cash 2
Cash 2
Clothing 1
Clothing 2
Clothing 3
Clothing 3
Consumer 1
Consumer 2
Consumer 2
Crafts 1
Crafts 2
Crafts 2
Dallas 1
Dallas 1
Discover 1
Discover 2
Discover 2
DVDs 1
DVDs 1
Electronics 1
Electronics 2
Electronics 2
Francisco 1
Francisco 1
Garden 1
Garden 2
Garden 2

Python 3.6.9 64-bit 0 0 0 Ln 1, Col 1 Spaces: 4 Go Live Prettier
```



```
Discover 2
DVDs 1
DVDs 1
Electronics 1
Electronics 2
Electronics 2
Francisco 1
Francisco 1
Garden 1
Garden 2
Garden 2
Jose 1
Jose 1
Las 1
Las 1
MasterCard 1
MasterCard 2
MasterCard 2
Memphis 1
Memphis 1
Men's 1
Men's 2
Men's 2
Miami 1
Miami 1
Newark 1
Newark 1
San 1
San 2
San 2
Tampa 1
Tampa 1
Tucson 1
Tucson 1
Vegas 1
Vegas 1
Visa 1
Visa 2
Visa 3
Visa 3
Washington 1
Washington 1
Women's 1
rhea@rhea-Lenovo-IdeaPad-S145-15IWL:~/data/Lab/190905156_DS/Lab$ ^C
rhea@rhea-Lenovo-IdeaPad-S145-15IWL:~/data/Lab/190905156_DS/Lab$
```

Exercise 2 : Try the above frequent word count program for the Heart Disease dataset, covid\_19\_data dataset, example dataset and German Credit data

```
from __future__ import print_function
import sys
for line in sys.stdin:
    L = [ (word.strip().lower(), 1 ) for word in line.strip().split() ]
```

```
for word, n in L:
print( '%s\t%d' % (word, n) )
```

```
#!/usr/bin/env python
# reducer.py
from __future__ import print_function
import sys
lastWord = None
sum = 0
for line in sys.stdin:
word, count = line.strip().split('\t', 1)
count = int(count)
if lastWord==None:
lastWord = word
sum = count
continue
if word==lastWord:
sum += count
else:
print( "%s\t%d" % ( lastWord, sum ) )
sum = count
lastWord = word
# output last word
if lastWord == word:
print( '%s\t%s' % (lastWord, sum ) )
```

```
from __future__ import print_function
import sys
# input comes from STDIN (standard input)
for line in sys.stdin:
word, count = line.strip().split('\t', 1)
count = int(count)
print( '%d\t%s' % (count, word) )
```

```
from __future__ import print_function
import sys
mostFreq = []
currentMax = -1
for line in sys.stdin:
count, word = line.strip().split('\t', 1)
count = int(count)
if count > currentMax:
currentMax = count
mostFreq = [ word ]
elif count == currentMax:
mostFreq.append( word )
# output mostFreq word(s)
for word in mostFreq:
print( '%s\t%s' % ( word, currentMax ) )
#combining all the four codes above we get the max frequency elements
```

```
Applications
Wed, Apr 20 00:53
freqred2.py - 190905156_DS - Visual Studio Code
File Edit Selection View Go Run Terminal Help
freqmap1.py freqred1.py freqmap2.py freqred2.py x
Lab5 > freqred2.py > ...
6 count, word = line.strip().split('\t', 1)
7 count = int(count)
8 if count > currentMax:
9     currentMax = count
10     mostFreq = [ word ]
11 elif count == currentMax:
12     mostFreq.append( word )
13 # output mostFreq word(s)
14 for word in mostFreq:
15     print( '%s\t%s' % ( word, currentMax ) )

OUTPUT DEBUG CONSOLE TERMINAL PROBLEMS
50,1,2,120,196,0,1,163,0,0,2,0,2,yes 1
50,1,2,140,233,0,1,163,0,0,6,1,1,3,no 1
50,1,2,140,233,0,1,163,0,0,6,1,1,3,no 1
50,1,2,140,233,0,1,163,0,0,6,1,1,3,no 1
51,0,0,130,305,0,1,142,1,1,2,1,0,3,no 1
51,0,0,130,305,0,1,142,1,1,2,1,0,3,no 1
51,0,0,130,305,0,1,142,1,1,2,1,0,3,no 1
51,0,2,120,295,0,0,157,0,0,6,2,0,2,yes 1
51,0,2,120,295,0,0,157,0,0,6,2,0,2,yes 1
51,0,2,130,256,0,0,140,0,0,5,2,0,2,yes 1
51,0,2,130,256,0,0,140,0,0,5,2,0,2,yes 1
51,0,2,130,256,0,0,140,0,0,5,2,0,2,yes 1
51,0,2,140,308,0,0,142,0,1,5,2,1,2,yes 1
51,0,2,140,308,0,0,142,0,1,5,2,1,2,yes 1
51,0,2,140,308,0,0,142,0,1,5,2,1,2,yes 1
51,1,0,140,261,0,0,186,1,0,2,0,2,yes 1
51,1,0,140,261,0,0,186,1,0,2,0,2,yes 1
51,1,0,140,298,0,1,122,1,4,2,1,3,3,no 1
51,1,0,140,298,0,1,122,1,4,2,1,3,3,no 1
51,1,0,140,298,0,1,122,1,4,2,1,3,3,no 1
51,1,0,140,299,0,1,173,1,1,6,2,0,3,no 1
51,1,0,140,299,0,1,173,1,1,6,2,0,3,no 1
51,1,0,140,299,0,1,173,1,1,6,2,0,3,no 1
51,1,2,100,222,0,1,143,1,1,2,1,0,2,yes 1
51,1,2,100,222,0,1,143,1,1,2,1,0,2,yes 1
51,1,2,110,175,0,1,123,0,0,6,2,0,2,yes 1
51,1,2,110,175,0,1,123,0,0,6,2,0,2,yes 1
51,1,2,110,175,0,1,123,0,0,6,2,0,2,yes 1
51,1,2,125,245,1,0,166,0,2,4,1,0,2,yes 1
51,1,2,125,245,1,0,166,0,2,4,1,0,2,yes 1
51,1,2,125,245,1,0,166,0,2,4,1,0,2,yes 1
51,1,2,94,227,0,1,154,1,0,2,1,3,yes 1
51,1,2,94,227,0,1,154,1,0,2,1,3,yes 1
51,1,3,125,213,0,0,125,1,1,4,2,1,2,yes 1
51,1,3,125,213,0,0,125,1,1,4,2,1,2,yes 1
52,0,2,136,196,0,0,169,0,0,1,1,0,2,yes 1
52,0,2,136,196,0,0,169,0,0,1,1,0,2,yes 1

Python 3.6.9 64-bit 0 0 0 Ln 15, Col 13 Spaces: 4 Go Live Prettier
```

```
"1","1258","24" 2
"1","1262","12" 2
"1","1374","6" 2
"1","1424","12" 2
"1","1478","15" 2
"1","2171","12" 2
"1","701","12" 2
```

```
rhea@rhea-Lenovo-IdeaPad-S145-15IWL:~/data/Lab/190905156_DS/Lab$ cat covid_19_data.csv | python3 freqmap1.py | sort | python3 freqred1.py | python3 freqmap2.py | sort | python3
freqred2.py
and 2675
and 2675
rhea@rhea-Lenovo-IdeaPad-S145-15IWL:~/data/Lab/190905156_DS/Lab$
```

```
OUTPUT    DEBUG CONSOLE    TERMINAL    PROBLEMS

09:04    1
09:23    1
09:23    1
09:25    1
09:25    1
09:57    1
09:01    1
09:02    1
09:02    1
09:04    1
09:04    1
09:23    1
09:23    1
09:25    1
09:25    1
09:57    1
09:57    1
amex     10
cash     10
amex     10
cash     10
cash     10
amex     10
cash     10
cash     10
discover 10
amex     10
cash     10
cash     10
discover 10
visa     10
amex     10
cash     10
cash     10
discover 10
visa     10
visa     10
amex     11
discover 11
amex     11
discover 11
discover 11
amex     13
amex     13
rhea@rhea-Lenovo-IdeaPad-S145-15IWL:~/data/Lab/190905156_DS/Lab5$
```

Exercise 3: Try the above 'Item explore and count' for the Heart Disease dataset, covid\_19\_data dataset, example dataset and German Credit data

```
import fileinput
for line in fileinput.input():
    data = line.strip().split("\t")
    if len(data) == 6:
        date, time, location, item, cost, payment = data
```

```

print("{0}\t{1}".format(location, cost))
import fileinput
transactions_count = 0
sales_total = 0
for line in fileinput.input():
    data = line.strip().split("\t")
    if len(data) != 2:
        continue
    current_key, current_value = data
    transactions_count += 1
    sales_total += float(current_value)
print(transactions_count, "\t", sales_total)

```

```

rhea@rhea-Lenovo-IdeaPad-S145-15IWL:~/data/Lab/190905156_DS/Lab5$ cat example.txt | python3 itemmap.py | sort | python3 itemred.py
1      82.38
rhea@rhea-Lenovo-IdeaPad-S145-15IWL:~/data/Lab/190905156_DS/Lab5$

```

```

290      242.76896551724138
291      242.99656357388315
292      243.2294520547945
293      243.47440273037543
294      243.71768707482994
295      243.96949152542373
296      244.22972972972974
297      244.4915824915825
298      244.76174496644296
299      245.0334448160535
300      245.31666666666666
301      245.61461794019934
302      245.91059602649005
303      246.26402640264027
303      246.26402640264027
rhea@rhea-Lenovo-IdeaPad-S145-15IWL:~/data/La

```

Exercise 4: Try to include separator using map reducing for the output of Heart Disease dataset, covid\_19\_data dataset, example dataset and German Credit data

```

import sys
def read_input(file):
    for line in file:
        yield line.split()

def main(separator='\t'):
    data = read_input(sys.stdin)
    for words in data:
        for word in words:
            print("%s%s%d" % (word, separator, 1))

if __name__ == "__main__":
    main()

```

```

from itertools import groupby
from operator import itemgetter
import sys
def read_mapper_output(file, separator='\t'):
    for line in file:
        yield line.rstrip().split(separator, 1)

def main(separator='\t'):
    data = read_mapper_output(sys.stdin, separator=separator)
    for current_word, group in groupby(data, itemgetter(0)):
        try:
            total_count = sum(int(count) for current_word, count in group)
            print ("%s%s%d" % (current_word, separator, total_count))
        except ValueError:
            pass

if __name__=="__main__":
    main()

```

```

64,1,3,110,211,0,0,144,1,1,8,1,0,2,Yes 1
64,1,3,170,227,0,0,155,0,0,6,1,0,3,Yes 1
65,0,0,150,225,0,0,114,0,1,1,3,3,No 1
65,0,2,140,417,1,0,157,0,0,8,2,1,2,Yes 1
65,0,2,155,269,0,1,148,0,0,8,2,0,2,Yes 1
65,0,2,160,360,0,0,151,0,0,8,2,0,2,Yes 1
65,1,0,110,248,0,0,158,0,0,6,2,2,1,No 1
65,1,0,120,177,0,1,140,0,0,4,2,0,3,Yes 1
65,1,0,135,254,0,0,127,0,2,8,1,1,3,No 1
65,1,3,130,282,1,0,174,0,1,4,1,1,2,No 1
66,0,0,170,228,1,1,165,1,1,1,2,3,No 1
66,0,2,146,278,0,0,152,0,0,1,1,2,Yes 1
66,0,3,150,226,0,1,114,0,2,6,0,0,2,Yes 1
66,1,0,112,212,0,0,132,1,0,1,2,1,2,No 1
66,1,0,120,302,0,0,151,0,0,4,1,0,2,Yes 1
66,1,0,160,228,0,0,138,0,2,3,2,0,1,Yes 1
66,1,1,160,246,0,1,120,1,0,1,3,1,No 1
67,0,0,106,223,0,1,142,0,0,3,2,2,2,Yes 1
67,0,2,115,564,0,0,160,0,1,6,1,0,3,Yes 1
67,0,2,152,277,0,1,172,0,0,2,1,2,Yes 1
67,1,0,100,290,0,0,125,1,0,0,1,2,2,No 1
67,1,0,120,229,0,0,120,1,2,6,1,2,3,No 1
67,1,0,120,237,0,1,71,0,1,1,0,2,No 1
67,1,0,125,254,1,1,163,0,0,2,1,2,3,No 1
67,1,0,160,286,0,0,108,1,1,5,1,3,2,No 1
67,1,2,152,212,0,0,150,0,0,0,1,0,3,No 1
68,0,2,120,211,0,0,115,0,1,5,1,0,2,Yes 1
68,1,0,144,193,1,1,141,0,3,4,1,2,3,No 1
68,1,2,118,277,0,1,151,0,1,2,1,3,Yes 1
68,1,2,180,274,1,0,150,1,1,6,1,0,3,No 1
69,0,3,140,239,0,1,151,0,1,8,2,2,2,Yes 1
69,1,2,140,254,0,0,146,0,2,1,3,3,No 1
69,1,3,160,234,1,0,131,0,0,1,1,2,Yes 1
70,1,0,130,322,0,0,100,0,2,4,1,3,2,No 1
70,1,0,145,174,0,1,125,1,2,6,0,0,3,No 1
70,1,1,156,245,0,0,143,0,0,2,0,2,Yes 1
70,1,2,160,269,0,1,112,1,2,9,1,1,3,No 1
71,0,0,112,149,0,1,125,0,1,6,1,0,2,Yes 1
71,0,1,160,302,0,1,162,0,0,4,2,2,2,Yes 1
71,0,2,110,265,1,0,130,0,0,2,1,2,Yes 1
74,0,1,120,269,0,0,121,1,0,2,2,1,2,Yes 1
76,0,2,140,197,0,2,116,0,1,1,1,0,2,Yes 1
77,1,0,125,304,0,0,162,1,0,2,3,2,No 1
age,sex,cp,trestbps,chol,fbs,restecg,thalach,exang,oldpeak,slope,ca,thal,target 1

```

Exercise 5: Try to apply finding max value using map reduce concept for the output of Heart Disease dataset, covid\_19\_data dataset, example dataset and German Credit data

```

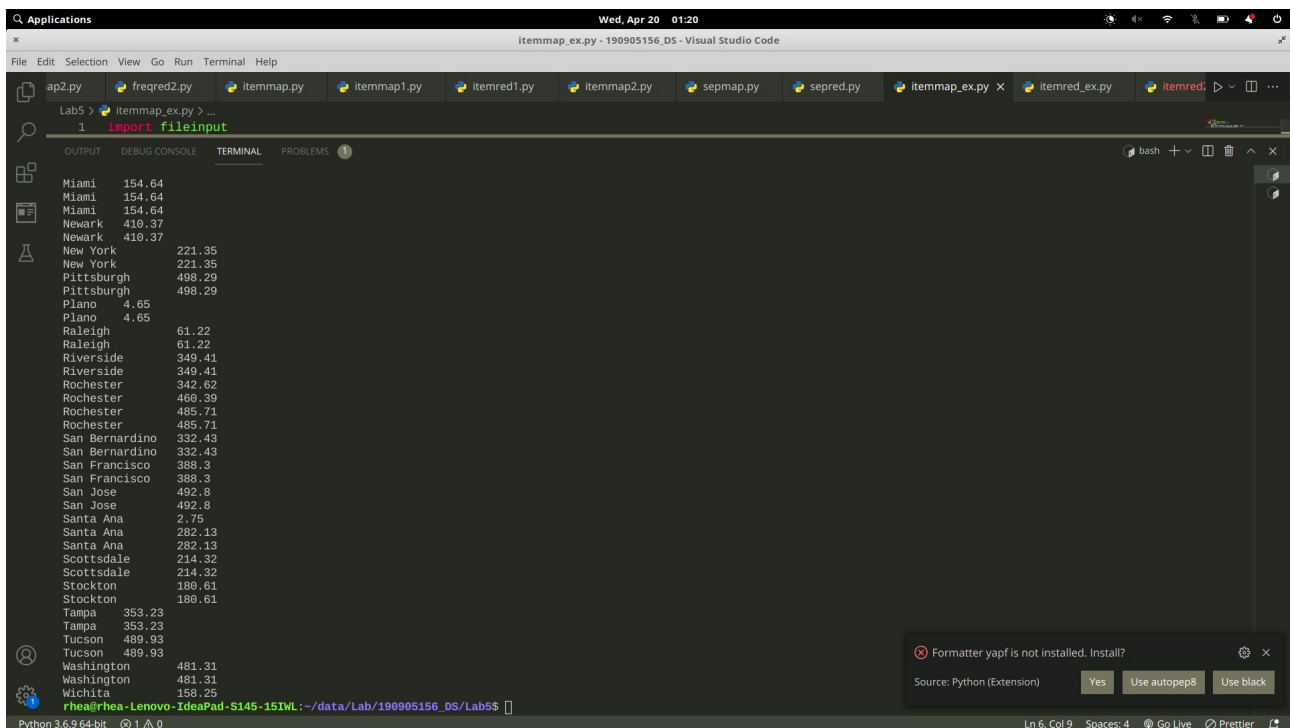
import fileinput
for line in fileinput.input():
    data = line.strip().split("\t")
    if len(data)==6:
        date, time, location, item, cost, payment = data

```



```
print("{0}\t{1}".format(location, cost))
```

```
import fileinput
max_value = 0
old_key = None
for line in fileinput.input():
    data = line.strip().split("\t")
    if len(data) != 2:
        continue
    current_key, current_value = data
    if old_key and old_key != current_key:
        print(old_key, "\t", max_value)
        max_value = 0
    if float(current_value) > float(max_value):
        max_value = float(current_value)
    old_key = current_key
if old_key != None:
    print (old_key, "\t", max_value)
```

The screenshot shows a Visual Studio Code window with a Python file named 'itemmap\_ex.py'. The code in the editor is a script that reads input from 'fileinput' and prints location and cost data. The output window shows the following data:

Location	Cost
Miami	154.64
Miami	154.64
Miami	154.64
Newark	410.37
Newark	410.37
New York	221.35
New York	221.35
Pittsburgh	498.29
Pittsburgh	498.29
Plano	4.65
Plano	4.65
Raleigh	61.22
Raleigh	61.22
Riverside	349.41
Riverside	349.41
Rochester	342.62
Rochester	460.39
Rochester	485.71
Rochester	485.71
San Bernardino	332.43
San Bernardino	332.43
San Francisco	388.3
San Francisco	388.3
San Jose	492.8
San Jose	492.8
Santa Ana	2.75
Santa Ana	282.13
Santa Ana	282.13
Scottsdale	214.32
Scottsdale	214.32
Stockton	180.61
Stockton	180.61
Tampa	353.23
Tampa	353.23
Tucson	489.93
Tucson	489.93
Washington	481.31
Washington	481.31
Wichita	158.25

Exercise 6: TOLD TO SKIP

Exercise 7: Write a map reduce program to count even or odd numbers in randomly generated natural numbers.

```
import fileinput
sum = 0
for line in fileinput.input():
    data = line.strip().split("\t")
    current_key, current_value = data
    sum += int(current_value)
print("Number of odd numbers is:", sum)
```

```

import sys
for line in sys.stdin:
    line = line.strip()
    words = line.split()
    for word in words:
        N = int(word)
        if N%2 == 1 and N > 0:
            print(N, '\t', 1)

```

```

import sys
for line in sys.stdin:
    line = line.strip()
    words = line.split()
    for word in words:
        N = int(word)
        if N%2 == 0 and N > 0:
            print(N, '\t', 1)

```

```

import fileinput
sum = 0
for line in fileinput.input():
    data = line.strip().split("\t")
    current_key, current_value = data
    sum += int(current_value)
print("Number of even numbers is:", sum)

```

```

rhea@rhea-Lenovo-IdeaPad-S145-15IWL:~/data/Lab/19090515
6_DS/Lab5$ echo "1 2 3 4 5" |python3 oddmap.py|sort|pyt
hon3 oddreduce.py
Number of odd numbers is: 3
rhea@rhea-Lenovo-IdeaPad-S145-15IWL:~/data/Lab/19090515
6_DS/Lab5$ █

```

```

6_DS/Lab5$ echo "11 12 321 12 124 21 111" |python3 even
map.py|sort|python3 evenreduce.py
Number of even numbers is: 3
rhea@rhea-Lenovo-IdeaPad-S145-15IWL:~/data/Lab/19090515
6_DS/Lab5$ █

```