Map Reduce Tasks

a. Which vendors have the most trips, and what is the total revenue generated by that vendor?

```
[hadoop@ip-172-31-82-64 ~]$ python mrtask_a.py hadoop -r hdfs:///user/hbase/csv
usage: mrtask_a.py [options] [input files]
mrtask_a.py: error: argument -r/--runner: invalid choice: 'hdfs:///user/hbase/csv' (choose from
'dataproc', 'emr', 'hadoop', 'inline', 'local', 'spark')
[hadoop@ip-172-31-82-64 ~]$ python mrtask_a.py hadoop -r hadoop:///user/hbase/csv > out_a.txt
usage: mrtask_a.py [options] [input files]
mrtask_a.py: error: argument -r/--runner: invalid choice: 'hadoop:///user/hbase/csv' (choose from
'dataproc', 'emr', 'hadoop', 'inline', 'local', 'spark')
[hadoop@ip-172-31-82-64 ~]$ python mrtask_a.py hadoop -r hdfs:///user/hbase/csv > out_a.txt
usage: mrtask_a.py [options] [input files]
mrtask_a.py: error: argument -r/--runner: invalid choice: 'hdfs:///user/hbase/csv' (choose from
'dataproc', 'emr', 'hadoop', 'inline', 'local', 'spark')
[hadoop@ip-172-31-82-64 ~]$ python mrtask_a.py -r hadoop hdfs:///user/hbase/csv > out_a.txt
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 2.10.1
Looking for Hadoop streaming jar in /home/hadoop/contrib...
Looking for Hadoop streaming jar in /usr/lib/hadoop-mapreduce...
Found Hadoop streaming jar: /usr/lib/hadoop-mapreduce/hadoop-streaming.jar
Creating temp directory /tmp/mrtask_a.hadoop.20230116.173603.058827
uploading working dir files to
hdfs:///user/hadoop/tmp/mrjob/mrtask_a.hadoop.20230116.173603.058827/files/wd...
Copying other local files to
hdfs:///user/hadoop/tmp/mrjob/mrtask_a.hadoop.20230116.173603.058827/files/
Running step 1 of 2...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-2.10.1-amzn-4.jar]
/tmp/streamjob1820419840089542899.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-82-64.ec2.internal/172.31.82.64:8032
  Connecting to Application History server at ip-172-31-82-64.ec2.internal/172.31.82.64:10200
  Connecting to ResourceManager at ip-172-31-82-64.ec2.internal/172.31.82.64:8032
  Connecting to Application History server at ip-172-31-82-64.ec2.internal/172.31.82.64:10200
  Loaded native gpl library
```

```
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev
049362b7cf53ff5f739d6b1532457f2c6cd495e8]
  Total input files to process : 6
  Adding a new node: /default-rack/172.31.93.192:50010
  Adding a new node: /default-rack/172.31.80.131:50010
  number of splits:43
  Submitting tokens for job: job_1673889736878_0001
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
  Adding resource type - name = vcores, units = , type = COUNTABLE
  Submitted application application_1673889736878_0001
  The url to track the job: http://ip-172-31-82-
64.ec2.internal:20888/proxy/application_1673889736878_0001/
  Running job: job_1673889736878_0001
  Job job_1673889736878_0001 running in uber mode : false
   map 0% reduce 0%
   map 1% reduce 0%
   map 2% reduce 0%
   map 4% reduce 0%
   map 5% reduce 0%
   map 6% reduce 0%
   map 7% reduce 0%
   map 8% reduce 0%
   map 9% reduce 0%
   map 10% reduce 0%
   map 12% reduce 0%
   map 13% reduce 0%
   map 14% reduce 0%
   map 15% reduce 0%
   map 16% reduce 0%
   map 17% reduce 0%
   map 19% reduce 0%
   map 20% reduce 0%
   map 21% reduce 0%
   map 22% reduce 0%
   map 23% reduce 0%
```

map 24% reduce 0%
map 25% reduce 0%
map 27% reduce 0%
map 28% reduce 0%
map 29% reduce 0%
map 30% reduce 0%
map 31% reduce 0%
map 33% reduce 0%
map 34% reduce 0%
map 35% reduce 0%
map 37% reduce 0%
map 38% reduce 0%
map 40% reduce 0%
map 41% reduce 0%
map 42% reduce 0%
map 43% reduce 0%
map 44% reduce 0%
map 47% reduce 0%
map 48% reduce 0%
map 49% reduce 0%
map 52% reduce 0%
map 53% reduce 0%
map 54% reduce 0%
map 56% reduce 0%
map 58% reduce 0%
map 60% reduce 0%
map 62% reduce 0%
map 63% reduce 0%
map 64% reduce 0%
map 65% reduce 0%
map 66% reduce 0%
map 67% reduce 0%
map 69% reduce 0%
map 70% reduce 0%
map 71% reduce 0%
map 72% reduce 0%
map 73% reduce 0%

```
  map 74% reduce 0%
  map 76% reduce 0%
  map 77% reduce 0%
  map 79% reduce 0%
  map 80% reduce 0%
  map 81% reduce 0%
  map 83% reduce 0%
  map 84% reduce 0%
  map 86% reduce 0%
  map 87% reduce 0%
  map 89% reduce 0%
  map 90% reduce 0%
  map 91% reduce 0%
  map 94% reduce 0%
  map 95% reduce 0%
  map 97% reduce 0%
  map 99% reduce 0%
  map 100% reduce 0%
  map 100% reduce 44%
  map 100% reduce 56%
  map 100% reduce 89%
  map 100% reduce 100%
 Job job_1673889736878_0001 completed successfully
 Output directory: hdfs:///user/hadoop/tmp/mrjob/mrtask_a.hadoop.20230116.173603.058827/step-output/0000
Counters: 52
        File Input Format Counters
                Bytes Read=5558093822
        File Output Format Counters
                Bytes Written=54
        File System Counters
                FILE: Number of bytes read=223716658
                FILE: Number of bytes written=457791260
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=5558099627
```

HDFS: Number of bytes written=54
HDFS: Number of large read operations=0
HDFS: Number of read operations=138
HDFS: Number of write operations=6
Job Counters
Data-local map tasks=24
Killed map tasks=1
Killed reduce tasks=2
Launched map tasks=43
Launched reduce tasks=4
Rack-local map tasks=19
Total megabyte-milliseconds taken by all map tasks=2481404928
Total megabyte-milliseconds taken by all reduce tasks=1428830208
Total time spent by all map tasks (ms)=807749
Total time spent by all maps in occupied slots (ms)=77543904
Total time spent by all reduce tasks (ms)=232557
Total time spent by all reduces in occupied slots (ms)=44650944
Total vcore-milliseconds taken by all map tasks=807749
Total vcore-milliseconds taken by all reduce tasks=232557
Map-Reduce Framework
CPU time spent (ms)=907170
Combine input records=0
Combine output records=0
Failed Shuffles=0
GC time elapsed (ms)=5542
Input split bytes=5805
Map input records=58982297
Map output bytes=535844647
Map output materialized bytes=223718449
Map output records=58982291
Merged Map outputs=129
Physical memory (bytes) snapshot=25805451264
Reduce input groups=2
Reduce input records=58982291
Reduce output records=2
Reduce shuffle bytes=223718449
Shuffled Maps =129

```
        Spilled Records=117964582
        Total committed heap usage (bytes)=24298651648
        Virtual memory (bytes) snapshot=221611900928
    Shuffle Errors
        BAD_ID=0
        CONNECTION=0
        IO_ERROR=0
        WRONG_LENGTH=0
        WRONG_MAP=0
        WRONG_REDUCE=0
Running step 2 of 2...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-2.10.1-amzn-4.jar]
/tmp/streamjob8962804546827026061.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-82-64.ec2.internal/172.31.82.64:8032
  Connecting to Application History server at ip-172-31-82-64.ec2.internal/172.31.82.64:10200
  Connecting to ResourceManager at ip-172-31-82-64.ec2.internal/172.31.82.64:8032
  Connecting to Application History server at ip-172-31-82-64.ec2.internal/172.31.82.64:10200
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev
049362b7cf53ff5f739d6b1532457f2c6cd495e8]
  Total input files to process : 3
  number of splits:11
  Submitting tokens for job: job_1673889736878_0003
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
  Adding resource type - name = vcores, units = , type = COUNTABLE
  Submitted application application_1673889736878_0003
  The url to track the job: http://ip-172-31-82-
64.ec2.internal:20888/proxy/application_1673889736878_0003/
  Running job: job_1673889736878_0003
  Job job_1673889736878_0003 running in uber mode : false
   map 0% reduce 0%
   map 9% reduce 0%
   map 18% reduce 0%
   map 27% reduce 0%
   map 36% reduce 0%
```

map 45% reduce 0%
map 55% reduce 0%
map 64% reduce 0%
map 73% reduce 0%
map 82% reduce 0%
map 91% reduce 0%
map 100% reduce 0%
map 100% reduce 33%
map 100% reduce 67%
map 100% reduce 100%
Job job_1673889736878_0003 completed successfully
Output directory: hdfs:///user/hadoop/tmp/mrjob/mrtask_a.hadoop.20230116.173603.058827/output
Counters: 52
File Input Format Counters
Bytes Read=150
File Output Format Counters
Bytes Written=60
File System Counters
FILE: Number of bytes read=111
FILE: Number of bytes written=3152053
FILE: Number of large read operations=0
FILE: Number of read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=2152
HDFS: Number of bytes written=60
HDFS: Number of large read operations=0
HDFS: Number of read operations=42
HDFS: Number of write operations=6
Job Counters
Data-local map tasks=5
Killed reduce tasks=1
Launched map tasks=11
Launched reduce tasks=3
Other local map tasks=1
Rack-local map tasks=5
Total megabyte-milliseconds taken by all map tasks=112398336
Total megabyte-milliseconds taken by all reduce tasks=56260608

Total time spent by all map tasks (ms)=36588
Total time spent by all maps in occupied slots (ms)=3512448
Total time spent by all reduce tasks (ms)=9157
Total time spent by all reduces in occupied slots (ms)=1758144
Total vcore-milliseconds taken by all map tasks=36588
Total vcore-milliseconds taken by all reduce tasks=9157
Map-Reduce Framework
CPU time spent (ms)=10300
Combine input records=0
Combine output records=0
Failed Shuffles=0
GC time elapsed (ms)=1255
Input split bytes=2002
Map input records=2
Map output bytes=54
Map output materialized bytes=586
Map output records=2
Merged Map outputs=33
Physical memory (bytes) snapshot=6721949696
Reduce input groups=1
Reduce input records=2
Reduce output records=3
Reduce shuffle bytes=586
Shuffled Maps =33
Spilled Records=4
Total committed heap usage (bytes)=6189219840
Virtual memory (bytes) snapshot=73062064128
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
job output is in hdfs:///user/hadoop/tmp/mrjob/mrtask_a.hadoop.20230116.173603.058827/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/mrtask_a.hadoop.20230116.173603.058827/output...

Removing HDFS temp directory

hdfs:///user/hadoop/tmp/mrjob/mrtask_a.hadoop.20230116.173603.058827...

Removing temp directory /tmp/mrtask_a.hadoop.20230116.173603.058827...

[hadoop@ip-172-31-82-64 ~]$

Output:

"Vendor Id" "Revenue"

"2" "525037658.14"

"1" "430567016.43"

Inference:

Vendor Id 2 have the most trips the total revenue generated by that vendor is 525037658.14

b. Which pickup location generates the most revenue?

```
[hadoop@ip-172-31-82-64 ~]$ python mrtask_b.py -r hadoop hdfs:///user/hbase/csv
> out_b.txt
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 2.10.1
Looking for Hadoop streaming jar in /home/hadoop/contrib...
Looking for Hadoop streaming jar in /usr/lib/hadoop-mapreduce...
Found Hadoop streaming jar: /usr/lib/hadoop-mapreduce/hadoop-streaming.jar
Creating temp directory /tmp/mrtask_b.hadoop.20230116.173800.855882
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/mrtask_b.hadoop.202
30116.173800.855882/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/mrtask_b.hadoop.20230
116.173800.855882/files/
Running step 1 of 2...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-2.10.1-amzn-4.jar] /tmp/st
reamjob3590602850508431303.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-82-64.ec2.internal/172.31.82.64:803
2
  Connecting to Application History server at ip-172-31-82-64.ec2.internal/172.3
1.82.64:10200
  Connecting to ResourceManager at ip-172-31-82-64.ec2.internal/172.31.82.64:803
2
  Connecting to Application History server at ip-172-31-82-64.ec2.internal/172.3
1.82.64:10200
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7c
f53ff5f739d6b1532457f2c6cd495e8]
  Total input files to process : 6
  Adding a new node: /default-rack/172.31.93.192:50010
  Adding a new node: /default-rack/172.31.80.131:50010
  number of splits:43
  Submitting tokens for job: job_1673889736878_0002
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
```

Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
Adding resource type - name = vcores, units = , type = COUNTABLE
Submitted application application_1673889736878_0002
The url to track the job: http://ip-172-31-82-64.ec2.internal:20888/proxy/appl
ication_1673889736878_0002/
Running job: job_1673889736878_0002
Job job_1673889736878_0002 running in uber mode : false

map 0% reduce 0%
map 1% reduce 0%
map 4% reduce 0%
map 5% reduce 0%
map 6% reduce 0%
map 7% reduce 0%
map 8% reduce 0%
map 9% reduce 0%
map 11% reduce 0%
map 12% reduce 0%
map 13% reduce 0%
map 14% reduce 0%
map 15% reduce 0%
map 16% reduce 0%
map 17% reduce 0%
map 19% reduce 0%
map 21% reduce 0%
map 22% reduce 0%
map 23% reduce 0%
map 24% reduce 0%
map 26% reduce 0%
map 27% reduce 0%
map 28% reduce 0%
map 29% reduce 0%
map 30% reduce 0%
map 32% reduce 0%
map 33% reduce 0%
map 34% reduce 0%
map 35% reduce 0%

map 36% reduce 0%
map 37% reduce 0%
map 40% reduce 0%
map 41% reduce 0%
map 42% reduce 0%
map 43% reduce 0%
map 44% reduce 0%
map 46% reduce 0%
map 47% reduce 0%
map 49% reduce 0%
map 52% reduce 0%
map 53% reduce 0%
map 54% reduce 0%
map 56% reduce 0%
map 58% reduce 0%
map 61% reduce 0%
map 62% reduce 0%
map 63% reduce 0%
map 65% reduce 0%
map 67% reduce 0%
map 67% reduce 7%
map 70% reduce 7%
map 72% reduce 7%
map 72% reduce 8%
map 73% reduce 8%
map 74% reduce 8%
map 75% reduce 8%
map 76% reduce 8%
map 77% reduce 8%
map 77% reduce 9%
map 79% reduce 9%
map 79% reduce 0%
map 81% reduce 0%
map 83% reduce 0%
map 86% reduce 0%
map 86% reduce 10%
map 87% reduce 10%

map 88% reduce 10%
map 90% reduce 10%
map 91% reduce 10%
map 92% reduce 10%
map 93% reduce 10%
map 94% reduce 10%
map 95% reduce 10%
map 95% reduce 11%
map 97% reduce 11%
map 99% reduce 11%
map 100% reduce 11%
map 100% reduce 20%
map 100% reduce 23%
map 100% reduce 46%
map 100% reduce 50%
map 100% reduce 52%
map 100% reduce 54%
map 100% reduce 57%
map 100% reduce 59%
map 100% reduce 61%
map 100% reduce 62%
map 100% reduce 63%
map 100% reduce 86%
map 100% reduce 87%
map 100% reduce 88%
map 100% reduce 89%
map 100% reduce 90%
map 100% reduce 92%
map 100% reduce 93%
map 100% reduce 94%
map 100% reduce 95%
map 100% reduce 96%
map 100% reduce 97%
map 100% reduce 98%
map 100% reduce 99%
map 100% reduce 100%
Job job_1673889736878_0002 completed successfully

Output directory: hdfs:///user/hadoop/tmp/mrjob/mrtask_b.hadoop.20230116.173800.855882/step-output/0000

Counters: 51

File Input Format Counters

Bytes Read=5558093822

File Output Format Counters

Bytes Written=8367

File System Counters

FILE: Number of bytes read=234843625

FILE: Number of bytes written=480964016

FILE: Number of large read operations=0

FILE: Number of read operations=0

FILE: Number of write operations=0

HDFS: Number of bytes read=5558099627

HDFS: Number of bytes written=8367

HDFS: Number of large read operations=0

HDFS: Number of read operations=138

HDFS: Number of write operations=6

Job Counters

Data-local map tasks=30

Killed reduce tasks=2

Launched map tasks=43

Launched reduce tasks=5

Rack-local map tasks=13

Total megabyte-milliseconds taken by all map tasks=2570090496

Total megabyte-milliseconds taken by all reduce tasks=2428231680

Total time spent by all map tasks (ms)=836618

Total time spent by all maps in occupied slots (ms)=80315328

Total time spent by all reduce tasks (ms)=395220

Total time spent by all reduces in occupied slots (ms)=75882240

Total vcore-milliseconds taken by all map tasks=836618

Total vcore-milliseconds taken by all reduce tasks=395220

Map-Reduce Framework

CPU time spent (ms)=945810

Combine input records=0

Combine output records=0

Failed Shuffles=0

GC time elapsed (ms)=5194
Input split bytes=5805
Map input records=58982297
Map output bytes=642695664
Map output materialized bytes=235764194
Map output records=58982291
Merged Map outputs=129
Physical memory (bytes) snapshot=26508668928
Reduce input groups=264
Reduce input records=58982291
Reduce output records=264
Reduce shuffle bytes=235764194
Shuffled Maps =129
Spilled Records=117964582
Total committed heap usage (bytes)=24881135616
Virtual memory (bytes) snapshot=221630914560
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
Running step 2 of 2...
packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-2.10.1-amzn-4.jar] /tmp/streamjob4388183140283375463.jar tmpDir=null
Connecting to ResourceManager at ip-172-31-82-64.ec2.internal/172.31.82.64:8032
Connecting to Application History server at ip-172-31-82-64.ec2.internal/172.31.82.64:10200
Connecting to ResourceManager at ip-172-31-82-64.ec2.internal/172.31.82.64:8032
Connecting to Application History server at ip-172-31-82-64.ec2.internal/172.31.82.64:10200
Loaded native gpl library
Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7cf53ff5f739d6b1532457f2c6cd495e8]
Total input files to process : 3
number of splits:9
Submitting tokens for job: job_1673889736878_0004
resource-types.xml not found

```
  Unable to find 'resource-types.xml'.
  Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
  Adding resource type - name = vcores, units = , type = COUNTABLE
  Submitted application application_1673889736878_0004
  The url to track the job: http://ip-172-31-82-
64.ec2.internal:20888/proxy/application_1673889736878_0004/
  Running job: job_1673889736878_0004
  Job job_1673889736878_0004 running in uber mode : false
   map 0% reduce 0%
   map 11% reduce 0%
   map 22% reduce 0%
   map 33% reduce 0%
   map 56% reduce 0%
   map 78% reduce 0%
   map 100% reduce 0%
   map 100% reduce 33%
   map 100% reduce 67%
   map 100% reduce 100%
  Job job_1673889736878_0004 completed successfully
  Output directory: hdfs:///user/hadoop/tmp/mrjob/mrtask_b.hadoop.20230116.173800.855882/output
Counters: 50
        File Input Format Counters
                Bytes Read=15696
        File Output Format Counters
                Bytes Written=6284
        File System Counters
                FILE: Number of bytes read=5147
                FILE: Number of bytes written=2712482
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=17334
                HDFS: Number of bytes written=6284
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=36
                HDFS: Number of write operations=6
        Job Counters
```

Data-local map tasks=2
Launched map tasks=9
Launched reduce tasks=3
Rack-local map tasks=7
Total megabyte-milliseconds taken by all map tasks=91413504
Total megabyte-milliseconds taken by all reduce tasks=50884608
Total time spent by all map tasks (ms)=29757
Total time spent by all maps in occupied slots (ms)=2856672
Total time spent by all reduce tasks (ms)=8282
Total time spent by all reduces in occupied slots (ms)=1590144
Total vcore-milliseconds taken by all map tasks=29757
Total vcore-milliseconds taken by all reduce tasks=8282
Map-Reduce Framework
CPU time spent (ms)=8660
Combine input records=0
Combine output records=0
Failed Shuffles=0
GC time elapsed (ms)=1148
Input split bytes=1638
Map input records=264
Map output bytes=8367
Map output materialized bytes=6213
Map output records=264
Merged Map outputs=27
Physical memory (bytes) snapshot=5519736832
Reduce input groups=1
Reduce input records=264
Reduce output records=265
Reduce shuffle bytes=6213
Shuffled Maps =27
Spilled Records=528
Total committed heap usage (bytes)=5307891712
Virtual memory (bytes) snapshot=63766257664
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0

WRONG_LENGTH=0

WRONG_MAP=0

WRONG_REDUCE=0

job output is in hdfs:///user/hadoop/tmp/mrjob/mrtask_b.hadoop.20230116.173800.855882/output

Streaming final output from hdfs:///user/hadoop/tmp/mrjob/mrtask_b.hadoop.20230116.173800.855882/output...

Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/mrtask_b.hadoop.20230116.173800.855882...

Removing temp directory /tmp/mrtask_b.hadoop.20230116.173800.855882...

[hadoop@ip-172-31-82-64 ~]$

Output:

"Pick up location"	**"Revenue"**
"132"	77196812.23975265
"138"	64480311.15005931
"161"	32910784.810575873
"230"	31638136.22054061
"186"	29804472.580555458
"162"	29439833.610487763
"237"	27341436.080601696
"234"	27237536.2204405
"170"	26918033.640432946
"236"	26541274.71052457
"48"	25973525.090475377
"79"	25537142.88041177
"163"	22988139.18035311
"142"	22624111.850368273
"164"	20490883.84027854
"239"	20043189.64028349
"68"	19738371.630267255
"107"	19263538.920251373
"231"	17639146.830172375
"249"	17626806.93020018
"141"	17144788.60023966
"100"	16463698.25019333
"264"	15677744.190126441
"90"	15403746.66013595
"238"	14806918.2701428

"229" 14795318.460143944
"140" 13868273.120113663
"148" 13854433.580092447
"113" 13455726.240094084
"263" 13312390.370103415
"246" 12548605.72006769
"114" 12088376.180055568
"233" 11696722.490039663
"13" 11128030.430018637
"158" 10734902.78001817
"43" 10537984.630015733
"137" 10126040.860005394
"144" 10118914.400004147
"87" 9539664.33999286
"262" 9394542.389982937
"50" 8199911.649957628
"143" 8174120.209955065
"211" 7332740.4999674745
"151" 6818072.629964825
"261" 6725657.549976835
"125" 5519954.489981784
"88" 5124168.38998802
"75" 5062266.769975233
"166" 4593718.649985568
"74" 2811993.9699967527
"41" 2809050.969996938
"232" 2498841.809998978
"24" 2459204.5900002234
"224" 2425451.7200003467
"265" 2367282.3500002604
"4" 2340230.860001019
"45" 2173274.2500020787
"209" 2146109.780001776
"226" 1847000.6100024034
"65" 1483344.7100013406
"7" 1474864.2800021258
"255" 1357749.1000011046

"42" 1280449.0200018405
"145" 1268926.2300012447
"33" 1251594.4900008093
"181" 1225742.0700009528
"244" 1144481.8300008944
"116" 1138515.860001026
"256" 1138370.6800008346
"25" 1035969.0600006635
"146" 980826.440000767
"97" 945635.4600005859
"152" 866576.5800005548
"223" 739696.0400001829
"10" 702650.1300000479
"52" 663017.1200000468
"40" 640130.2200000285
"112" 607837.5799999775
"80" 603902.479999973
"260" 591684.3699999474
"179" 568443.0799998947
"129" 549493.5199998369
"66" 476150.37999989407
"12" 475406.99999988236
"193" 404513.0799998566
"93" 389261.489999988
"1" 362708.2499999822
"49" 359258.15999989805
"70" 330961.7099999677
"189" 296205.6199999421
"82" 293450.6799999265
"37" 290241.4299999361
"17" 254361.23999994196
"215" 244459.11999999415
"243" 236280.1699999669
"61" 230727.33999996414
"130" 229602.90999999168
"216" 198947.619999995
"194" 195797.489999996

"247" 188856.1199999804
"219" 187297.65999999695
"95" 183741.38999998884
"106" 183462.35999998878
"168" 179693.53999998243
"28" 168439.5999999967
"36" 166126.8199999955
"188" 124676.90000001132
"228" 123028.41000000839
"83" 119217.8200000123
"127" 118998.76000000845
"196" 109646.22000000521
"225" 108150.92000000953
"89" 100567.79000000664
"157" 92362.13000000153
"134" 87805.12000000256
"217" 81012.75000000707
"195" 79454.85000000251
"197" 67154.58000000136
"92" 65687.66000000248
"14" 64979.03000000208
"56" 63385.270000000775
"54" 60264.300000001764
"133" 57427.91000000168
"190" 53633.6800000015
"202" 52506.59000000125
"159" 47432.56000000132
"257" 45502.640000000436
"62" 44878.52000000077
"76" 44802.12000000042
"69" 42375.26000000063
"124" 38977.710000000036
"198" 38220.07999999995
"173" 32922.46999999917
"34" 32739.119999999402
"119" 32443.699999999204
"207" 32265.139999998795

"177"	31473.639999999476
"63"	30322.529999999573
"39"	27212.48999999977
"85"	24830.16999999951
"160"	24656.94999999975
"220"	24585.139999999574
"8"	23939.629999999892
"26"	22539.549999999614
"22"	22311.19999999978
"252"	21569.429999999928
"135"	21443.88999999974
"235"	21153.91999999961
"167"	20969.07999999954
"91"	20090.87999999977
"169"	19638.899999999645
"258"	19244.8899999998
"213"	18481.97999999981
"165"	17120.769999999793
"29"	16466.339999999924
"123"	16222.019999999797
"242"	15471.599999999813
"227"	15428.92999999988
"180"	15353.159999999909
"35"	15300.889999999874
"78"	14972.049999999843
"208"	14400.57999999993
"72"	14296.059999999885
"126"	13985.809999999874
"55"	13708.599999999942
"212"	13590.899999999889
"192"	13509.799999999881
"250"	13304.63999999991
"155"	12579.90999999997
"71"	12442.78999999996
"241"	12362.919999999927
"136"	12221.429999999898
"108"	11996.85999999996

"174" 11976.629999999937
"102" 11913.969999999972
"182" 11867.149999999943
"11" 11790.819999999956
"191" 11738.289999999963
"203" 11604.419999999976
"218" 11565.019999999973
"18" 11149.239999999954
"47" 11119.669999999938
"178" 10730.819999999938
"19" 10479.769999999991
"200" 10450.139999999978
"121" 9987.990000000002
"147" 9907.149999999987
"120" 9887.32999999999
"77" 9755.569999999987
"153" 9636.129999999986
"210" 9448.98
"6" 9297.119999999994
"131" 9128.190000000022
"53" 8964.65000000002
"21" 8516.880000000017
"248" 8414.56000000004
"259" 8372.290000000034
"67" 8329.220000000027
"51" 8120.170000000027
"128" 7946.5100000000275
"154" 7811.950000000013
"253" 7794.860000000015
"16" 7782.970000000029
"254" 7761.240000000036
"20" 7623.3000000000375
"185" 7527.450000000042
"101" 7505.100000000022
"60" 7435.1600000000435
"156" 7219.190000000028
"31" 7148.2300000000205

"150" 7064.820000000013
"94" 7010.830000000043
"171" 6932.950000000034
"149" 6815.030000000022
"117" 6768.900000000008
"38" 6541.220000000013
"205" 6344.760000000015
"57" 5961.270000000018
"222" 5465.700000000008
"98" 5447.0200000000195
"64" 5368.090000000009
"221" 5344.380000000012
"23" 5186.300000000006
"240" 5021.290000000015
"9" 4874.190000000014
"81" 4615.250000000013
"175" 4539.900000000007
"115" 4435.710000000014
"32" 4268.070000000018
"15" 4137.850000000003
"251" 4108.370000000005
"3" 3761.3000000000043
"139" 3440.7800000000043
"122" 3260.360000000004
"73" 3256.2500000000077
"96" 3228.9800000000027
"118" 2725.3900000000003
"111" 2673.000000000001
"201" 2575.4500000000025
"2" 2383.5299999999997
"172" 2007.4099999999994
"86" 1998.4799999999984
"183" 1942.9699999999975
"184" 1869.0899999999988
"214" 1787.2599999999989
"206" 1743.3899999999992
"245" 1523.5499999999995

"204" 1404.2999999999995

"84" 1371.6499999999996

"46" 1296.2799999999995

"109" 1279.8699999999994

"58" 1118.76

"105" 1095.1499999999994

"176" 967.9699999999998

"5" 913.88

"30" 851.3999999999997

"44" 799.3699999999999

"187" 762.5599999999997

"27" 648.93

"59" 484.3200000000001

"199" 420.3300000000001

"99" 121.82

"104" 60.08

"110" 6.8

Inference:

Pickup location **132** generates most revenue.

c. What are the different payment types used by customers and their count? The final results should be in a sorted format.

```
[hadoop@ip-172-31-82-64 ~]$ python mrtask_c.py -r hadoop hdfs:///user/hbase/csv > out_b.txt
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 2.10.1
Looking for Hadoop streaming jar in /home/hadoop/contrib...
Looking for Hadoop streaming jar in /usr/lib/hadoop-mapreduce...
Found Hadoop streaming jar: /usr/lib/hadoop-mapreduce/hadoop-streaming.jar
Creating temp directory /tmp/mrtask_c.hadoop.20230116.175520.837150
uploading working dir files to
hdfs:///user/hadoop/tmp/mrjob/mrtask_c.hadoop.20230116.175520.837150/files/wd...
Copying other local files to
hdfs:///user/hadoop/tmp/mrjob/mrtask_c.hadoop.20230116.175520.837150/files/
Running step 1 of 2...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-2.10.1-amzn-4.jar]
/tmp/streamjob4189443226318416693.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-82-64.ec2.internal/172.31.82.64:8032
  Connecting to Application History server at ip-172-31-82-64.ec2.internal/172.31.82.64:10200
  Connecting to ResourceManager at ip-172-31-82-64.ec2.internal/172.31.82.64:8032
  Connecting to Application History server at ip-172-31-82-64.ec2.internal/172.31.82.64:10200
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev
049362b7cf53ff5f739d6b1532457f2c6cd495e8]
  Total input files to process : 6
  Adding a new node: /default-rack/172.31.80.224:50010
  Adding a new node: /default-rack/172.31.93.192:50010
  Adding a new node: /default-rack/172.31.86.34:50010
  Adding a new node: /default-rack/172.31.80.131:50010
  number of splits:43
  Submitting tokens for job: job_1673889736878_0006
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
```

Adding resource type - name = vcores, units = , type = COUNTABLE
Submitted application application_1673889736878_0006
The url to track the job: http://ip-172-31-82-64.ec2.internal:20888/proxy/application_1673889736878_0006/
Running job: job_1673889736878_0006
Job job_1673889736878_0006 running in uber mode : false
map 0% reduce 0%
map 1% reduce 0%
map 2% reduce 0%
map 4% reduce 0%
map 5% reduce 0%
map 6% reduce 0%
map 7% reduce 0%
map 8% reduce 0%
map 11% reduce 0%
map 12% reduce 0%
map 14% reduce 0%
map 15% reduce 0%
map 16% reduce 0%
map 18% reduce 0%
map 19% reduce 0%
map 20% reduce 0%
map 21% reduce 0%
map 22% reduce 0%
map 23% reduce 0%
map 24% reduce 0%
map 26% reduce 0%
map 27% reduce 0%
map 29% reduce 0%
map 30% reduce 0%
map 31% reduce 0%
map 34% reduce 0%
map 35% reduce 0%
map 36% reduce 0%
map 37% reduce 0%
map 38% reduce 0%
map 40% reduce 0%

map 41% reduce 0%
map 42% reduce 0%
map 43% reduce 0%
map 44% reduce 0%
map 45% reduce 0%
map 48% reduce 0%
map 49% reduce 0%
map 50% reduce 0%
map 51% reduce 0%
map 52% reduce 0%
map 53% reduce 0%
map 54% reduce 0%
map 55% reduce 0%
map 56% reduce 0%
map 57% reduce 0%
map 58% reduce 0%
map 59% reduce 0%
map 60% reduce 0%
map 62% reduce 0%
map 65% reduce 0%
map 67% reduce 0%
map 70% reduce 0%
map 70% reduce 8%
map 71% reduce 8%
map 73% reduce 8%
map 74% reduce 8%
map 74% reduce 0%
map 75% reduce 0%
map 77% reduce 0%
map 78% reduce 0%
map 79% reduce 0%
map 81% reduce 0%
map 83% reduce 0%
map 84% reduce 0%
map 84% reduce 9%
map 85% reduce 9%
map 86% reduce 9%

map 87% reduce 9%
map 88% reduce 9%
map 88% reduce 10%
map 90% reduce 10%
map 91% reduce 10%
map 92% reduce 10%
map 93% reduce 10%
map 98% reduce 10%
map 98% reduce 11%
map 100% reduce 11%
map 100% reduce 18%
map 100% reduce 22%
map 100% reduce 44%
map 100% reduce 56%
map 100% reduce 89%
map 100% reduce 100%
Job job_1673889736878_0006 completed successfully
Output directory: hdfs:///user/hadoop/tmp/mrjob/mrtask_c.hadoop.20230116.175520.837150/step-output/0000
Counters: 51
File Input Format Counters
Bytes Read=5558093822
File Output Format Counters
Bytes Written=93
File System Counters
FILE: Number of bytes read=22221057
FILE: Number of bytes written=54800192
FILE: Number of large read operations=0
FILE: Number of read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=5558099627
HDFS: Number of bytes written=93
HDFS: Number of large read operations=0
HDFS: Number of read operations=138
HDFS: Number of write operations=6
Job Counters
Data-local map tasks=21

Killed reduce tasks=3
Launched map tasks=43
Launched reduce tasks=5
Rack-local map tasks=22
Total megabyte-milliseconds taken by all map tasks=2263265280
Total megabyte-milliseconds taken by all reduce tasks=2152968192
Total time spent by all map tasks (ms)=736740
Total time spent by all maps in occupied slots (ms)=70727040
Total time spent by all reduce tasks (ms)=350418
Total time spent by all reduces in occupied slots (ms)=67280256
Total vcore-milliseconds taken by all map tasks=736740
Total vcore-milliseconds taken by all reduce tasks=350418
Map-Reduce Framework
CPU time spent (ms)=870490
Combine input records=0
Combine output records=0
Failed Shuffles=0
GC time elapsed (ms)=5053
Input split bytes=5805
Map input records=58982297
Map output bytes=353893746
Map output materialized bytes=22223135
Map output records=58982291
Merged Map outputs=129
Physical memory (bytes) snapshot=26299174912
Reduce input groups=5
Reduce input records=58982291
Reduce output records=5
Reduce shuffle bytes=22223135
Shuffled Maps =129
Spilled Records=117964582
Total committed heap usage (bytes)=24307040256
Virtual memory (bytes) snapshot=221550362624
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0

```
		WRONG_LENGTH=0
		WRONG_MAP=0
		WRONG_REDUCE=0
Running step 2 of 2...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-2.10.1-amzn-4.jar]
/tmp/streamjob3961424177553196943.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-82-64.ec2.internal/172.31.82.64:8032
  Connecting to Application History server at ip-172-31-82-64.ec2.internal/172.31.82.64:10200
  Connecting to ResourceManager at ip-172-31-82-64.ec2.internal/172.31.82.64:8032
  Connecting to Application History server at ip-172-31-82-64.ec2.internal/172.31.82.64:10200
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev
049362b7cf53ff5f739d6b1532457f2c6cd495e8]
  Total input files to process : 3
  number of splits:10
  Submitting tokens for job: job_1673889736878_0008
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
  Adding resource type - name = vcores, units = , type = COUNTABLE
  Submitted application application_1673889736878_0008
  The url to track the job: http://ip-172-31-82-
64.ec2.internal:20888/proxy/application_1673889736878_0008/
  Running job: job_1673889736878_0008
  Job job_1673889736878_0008 running in uber mode : false
   map 0% reduce 0%
   map 30% reduce 0%
   map 50% reduce 0%
   map 60% reduce 0%
   map 80% reduce 0%
   map 100% reduce 0%
   map 100% reduce 33%
   map 100% reduce 67%
   map 100% reduce 100%
  Job job_1673889736878_0008 completed successfully
  Output directory: hdfs:///user/hadoop/tmp/mrjob/mrtask_c.hadoop.20230116.175520.837150/output
Counters: 51
```

File Input Format Counters

Bytes Read=191

File Output Format Counters

Bytes Written=76

File System Counters

FILE: Number of bytes read=146

FILE: Number of bytes written=2926949

FILE: Number of large read operations=0

FILE: Number of read operations=0

FILE: Number of write operations=0

HDFS: Number of bytes read=2011

HDFS: Number of bytes written=76

HDFS: Number of large read operations=0

HDFS: Number of read operations=39

HDFS: Number of write operations=6

Job Counters

Data-local map tasks=6

Killed reduce tasks=1

Launched map tasks=10

Launched reduce tasks=3

Rack-local map tasks=4

Total megabyte-milliseconds taken by all map tasks=109317120

Total megabyte-milliseconds taken by all reduce tasks=55019520

Total time spent by all map tasks (ms)=35585

Total time spent by all maps in occupied slots (ms)=3416160

Total time spent by all reduce tasks (ms)=8955

Total time spent by all reduces in occupied slots (ms)=1719360

Total vcore-milliseconds taken by all map tasks=35585

Total vcore-milliseconds taken by all reduce tasks=8955

Map-Reduce Framework

CPU time spent (ms)=9780

Combine input records=0

Combine output records=0

Failed Shuffles=0

GC time elapsed (ms)=1343

Input split bytes=1820

Map input records=5

Map output bytes=93

Map output materialized bytes=583

Map output records=5

Merged Map outputs=30

Physical memory (bytes) snapshot=6011453440

Reduce input groups=1

Reduce input records=5

Reduce output records=6

Reduce shuffle bytes=583

Shuffled Maps =30

Spilled Records=10

Total committed heap usage (bytes)=5680660480

Virtual memory (bytes) snapshot=68400787456

Shuffle Errors

BAD_ID=0

CONNECTION=0

IO_ERROR=0

WRONG_LENGTH=0

WRONG_MAP=0

WRONG_REDUCE=0

job output is in hdfs:///user/hadoop/tmp/mrjob/mrtask_c.hadoop.20230116.175520.837150/output

Streaming final output from

hdfs:///user/hadoop/tmp/mrjob/mrtask_c.hadoop.20230116.175520.837150/output...

Removing HDFS temp directory

hdfs:///user/hadoop/tmp/mrjob/mrtask_c.hadoop.20230116.175520.837150...

Removing temp directory /tmp/mrtask_c.hadoop.20230116.175520.837150...

[hadoop@ip-172-31-82-64 ~]$

Output:

"Payment_type""Count"

"1" 39754212

"2" 18832370

"3" 306912

"4" 88794

"5" 3

Inference:

The above output shows payment type and the number of customer using those payment types.

d. What is the average trip time for different pickup locations?

```
[hadoop@ip-172-31-82-64 ~]$ python mrtask_d.py -r hadoop hdfs:///user/hbase/csv > out_b.txt
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 2.10.1
Looking for Hadoop streaming jar in /home/hadoop/contrib...
Looking for Hadoop streaming jar in /usr/lib/hadoop-mapreduce...
Found Hadoop streaming jar: /usr/lib/hadoop-mapreduce/hadoop-streaming.jar
Creating temp directory /tmp/mrtask_d.hadoop.20230116.175503.404786
uploading working dir files to
hdfs:///user/hadoop/tmp/mrjob/mrtask_d.hadoop.20230116.175503.404786/files/wd...
Copying other local files to
hdfs:///user/hadoop/tmp/mrjob/mrtask_d.hadoop.20230116.175503.404786/files/
Running step 1 of 2...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-2.10.1-amzn-4.jar]
/tmp/streamjob8409000395018116325.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-82-64.ec2.internal/172.31.82.64:8032
  Connecting to Application History server at ip-172-31-82-64.ec2.internal/172.31.82.64:10200
  Connecting to ResourceManager at ip-172-31-82-64.ec2.internal/172.31.82.64:8032
  Connecting to Application History server at ip-172-31-82-64.ec2.internal/172.31.82.64:10200
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev
049362b7cf53ff5f739d6b1532457f2c6cd495e8]
  Total input files to process : 6
  Adding a new node: /default-rack/172.31.80.224:50010
  Adding a new node: /default-rack/172.31.93.192:50010
  Adding a new node: /default-rack/172.31.86.34:50010
  Adding a new node: /default-rack/172.31.80.131:50010
  number of splits:43
  Submitting tokens for job: job_1673889736878_0005
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
  Adding resource type - name = vcores, units = , type = COUNTABLE
```

Submitted application application_1673889736878_0005
The url to track the job: http://ip-172-31-82-64.ec2.internal:20888/proxy/application_1673889736878_0005/
Running job: job_1673889736878_0005
Job job_1673889736878_0005 running in uber mode : false
map 0% reduce 0%
map 2% reduce 0%
map 3% reduce 0%
map 5% reduce 0%
map 6% reduce 0%
map 7% reduce 0%
map 8% reduce 0%
map 9% reduce 0%
map 11% reduce 0%
map 12% reduce 0%
map 14% reduce 0%
map 15% reduce 0%
map 16% reduce 0%
map 17% reduce 0%
map 19% reduce 0%
map 20% reduce 0%
map 21% reduce 0%
map 22% reduce 0%
map 23% reduce 0%
map 25% reduce 0%
map 26% reduce 0%
map 28% reduce 0%
map 29% reduce 0%
map 30% reduce 0%
map 31% reduce 0%
map 33% reduce 0%
map 34% reduce 0%
map 36% reduce 0%
map 37% reduce 0%
map 38% reduce 0%
map 39% reduce 0%
map 40% reduce 0%

map 41% reduce 0%
map 42% reduce 0%
map 43% reduce 0%
map 44% reduce 0%
map 45% reduce 0%
map 46% reduce 0%
map 47% reduce 0%
map 48% reduce 0%
map 49% reduce 0%
map 50% reduce 0%
map 51% reduce 0%
map 52% reduce 0%
map 53% reduce 0%
map 54% reduce 0%
map 55% reduce 0%
map 57% reduce 0%
map 58% reduce 0%
map 59% reduce 0%
map 60% reduce 0%
map 61% reduce 0%
map 62% reduce 0%
map 64% reduce 0%
map 65% reduce 0%
map 66% reduce 0%
map 67% reduce 0%
map 69% reduce 0%
map 70% reduce 0%
map 71% reduce 0%
map 72% reduce 0%
map 73% reduce 0%
map 74% reduce 0%
map 75% reduce 0%
map 76% reduce 0%
map 77% reduce 0%
map 78% reduce 0%
map 79% reduce 0%
map 80% reduce 0%

map 81% reduce 0%
map 82% reduce 0%
map 83% reduce 0%
map 84% reduce 0%
map 85% reduce 0%
map 86% reduce 0%
map 87% reduce 0%
map 88% reduce 0%
map 89% reduce 0%
map 90% reduce 0%
map 91% reduce 0%
map 92% reduce 0%
map 93% reduce 0%
map 94% reduce 0%
map 95% reduce 0%
map 97% reduce 0%
map 98% reduce 0%
map 99% reduce 0%
map 100% reduce 0%
map 100% reduce 24%
map 100% reduce 47%
map 100% reduce 49%
map 100% reduce 50%
map 100% reduce 51%
map 100% reduce 53%
map 100% reduce 54%
map 100% reduce 55%
map 100% reduce 56%
map 100% reduce 57%
map 100% reduce 58%
map 100% reduce 59%
map 100% reduce 60%
map 100% reduce 62%
map 100% reduce 63%
map 100% reduce 86%
map 100% reduce 87%
map 100% reduce 89%

map 100% reduce 91%
map 100% reduce 92%
map 100% reduce 93%
map 100% reduce 94%
map 100% reduce 95%
map 100% reduce 96%
map 100% reduce 97%
map 100% reduce 98%
map 100% reduce 99%
map 100% reduce 100%
Job job_1673889736878_0005 completed successfully
Output directory: hdfs:///user/hadoop/tmp/mrjob/mrtask_d.hadoop.20230116.175503.404786/step-output/0000
Counters: 52
File Input Format Counters
Bytes Read=5558093822
File Output Format Counters
Bytes Written=4894
File System Counters
FILE: Number of bytes read=197776310
FILE: Number of bytes written=406209693
FILE: Number of large read operations=0
FILE: Number of read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=5558099627
HDFS: Number of bytes written=4894
HDFS: Number of large read operations=0
HDFS: Number of read operations=138
HDFS: Number of write operations=6
Job Counters
Data-local map tasks=23
Killed map tasks=1
Killed reduce tasks=1
Launched map tasks=43
Launched reduce tasks=4
Rack-local map tasks=20
Total megabyte-milliseconds taken by all map tasks=3628664832

Total megabyte-milliseconds taken by all reduce tasks=1935058944
Total time spent by all map tasks (ms)=1181206
Total time spent by all maps in occupied slots (ms)=113395776
Total time spent by all reduce tasks (ms)=314951
Total time spent by all reduces in occupied slots (ms)=60470592
Total vcore-milliseconds taken by all map tasks=1181206
Total vcore-milliseconds taken by all reduce tasks=314951
Map-Reduce Framework
CPU time spent (ms)=1358120
Combine input records=0
Combine output records=0
Failed Shuffles=0
GC time elapsed (ms)=5928
Input split bytes=5805
Map input records=58982297
Map output bytes=789126820
Map output materialized bytes=198077365
Map output records=58982291
Merged Map outputs=129
Physical memory (bytes) snapshot=26851635200
Reduce input groups=264
Reduce input records=58982291
Reduce output records=264
Reduce shuffle bytes=198077365
Shuffled Maps =129
Spilled Records=117964582
Total committed heap usage (bytes)=25726287872
Virtual memory (bytes) snapshot=221655871488
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
Running step 2 of 2...

```
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-2.10.1-amzn-4.jar]
/tmp/streamjob3707548099514650183.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-82-64.ec2.internal/172.31.82.64:8032
  Connecting to Application History server at ip-172-31-82-64.ec2.internal/172.31.82.64:10200
  Connecting to ResourceManager at ip-172-31-82-64.ec2.internal/172.31.82.64:8032
  Connecting to Application History server at ip-172-31-82-64.ec2.internal/172.31.82.64:10200
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev
049362b7cf53ff5f739d6b1532457f2c6cd495e8]
  Total input files to process : 3
  number of splits:9
  Submitting tokens for job: job_1673889736878_0007
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
  Adding resource type - name = vcores, units = , type = COUNTABLE
  Submitted application application_1673889736878_0007
  The url to track the job: http://ip-172-31-82-
64.ec2.internal:20888/proxy/application_1673889736878_0007/
  Running job: job_1673889736878_0007
  Job job_1673889736878_0007 running in uber mode : false
   map 0% reduce 0%
   map 11% reduce 0%
   map 22% reduce 0%
   map 33% reduce 0%
   map 44% reduce 0%
   map 56% reduce 0%
   map 67% reduce 0%
   map 78% reduce 0%
   map 89% reduce 0%
   map 100% reduce 0%
   map 100% reduce 33%
   map 100% reduce 67%
   map 100% reduce 100%
  Job job_1673889736878_0007 completed successfully
  Output directory: hdfs:///user/hadoop/tmp/mrjob/mrtask_d.hadoop.20230116.175503.404786/output
Counters: 50
```

File Input Format Counters
    Bytes Read=9183
File Output Format Counters
    Bytes Written=2812
File System Counters
    FILE: Number of bytes read=1801
    FILE: Number of bytes written=2705598
    FILE: Number of large read operations=0
    FILE: Number of read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=10821
    HDFS: Number of bytes written=2812
    HDFS: Number of large read operations=0
    HDFS: Number of read operations=36
    HDFS: Number of write operations=6
Job Counters
    Data-local map tasks=6
    Launched map tasks=9
    Launched reduce tasks=3
    Rack-local map tasks=3
    Total megabyte-milliseconds taken by all map tasks=98282496
    Total megabyte-milliseconds taken by all reduce tasks=45477888
    Total time spent by all map tasks (ms)=31993
    Total time spent by all maps in occupied slots (ms)=3071328
    Total time spent by all reduce tasks (ms)=7402
    Total time spent by all reduces in occupied slots (ms)=1421184
    Total vcore-milliseconds taken by all map tasks=31993
    Total vcore-milliseconds taken by all reduce tasks=7402
Map-Reduce Framework
    CPU time spent (ms)=9270
    Combine input records=0
    Combine output records=0
    Failed Shuffles=0
    GC time elapsed (ms)=1091
    Input split bytes=1638
    Map input records=264
    Map output bytes=4894

Map output materialized bytes=2708
Map output records=264
Merged Map outputs=27
Physical memory (bytes) snapshot=5657284608
Reduce input groups=1
Reduce input records=264
Reduce output records=265
Reduce shuffle bytes=2708
Shuffled Maps =27
Spilled Records=528
Total committed heap usage (bytes)=5626658816
Virtual memory (bytes) snapshot=63815311360
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
job output is in hdfs:///user/hadoop/tmp/mrjob/mrtask_d.hadoop.20230116.175503.404786/output
Streaming final output from hdfs:///user/hadoop/tmp/mrjob/mrtask_d.hadoop.20230116.175503.404786/output...
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/mrtask_d.hadoop.20230116.175503.404786...
Removing temp directory /tmp/mrtask_d.hadoop.20230116.175503.404786...
[hadoop@ip-172-31-82-64 ~]$ python mrtask_f.py -r hadoop hdfs:///user/hbase/csv > out_f.txt
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 2.10.1
Looking for Hadoop streaming jar in /home/hadoop/contrib...
Looking for Hadoop streaming jar in /usr/lib/hadoop-mapreduce...
Found Hadoop streaming jar: /usr/lib/hadoop-mapreduce/hadoop-streaming.jar
Creating temp directory /tmp/mrtask_f.hadoop.20230116.181733.820703
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/mrtask_f.hadoop.20230116.181733.820703/files/wd...

Copying other local files to
hdfs:///user/hadoop/tmp/mrjob/mrtask_f.hadoop.20230116.181733.820703/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-2.10.1-amzn-4.jar]
/tmp/streamjob4689407670852389096.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-82-64.ec2.internal/172.31.82.64:8032
  Connecting to Application History server at ip-172-31-82-64.ec2.internal/172.31.82.64:10200
  Connecting to ResourceManager at ip-172-31-82-64.ec2.internal/172.31.82.64:8032
  Connecting to Application History server at ip-172-31-82-64.ec2.internal/172.31.82.64:10200
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev
049362b7cf53ff5f739d6b1532457f2c6cd495e8]
  Total input files to process : 6
  Adding a new node: /default-rack/172.31.80.224:50010
  Adding a new node: /default-rack/172.31.86.34:50010
  number of splits:43
  Submitting tokens for job: job_1673889736878_0010
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
  Adding resource type - name = vcores, units = , type = COUNTABLE
  Submitted application application_1673889736878_0010
  The url to track the job: http://ip-172-31-82-
64.ec2.internal:20888/proxy/application_1673889736878_0010/

Output:

"location"	"Trip Time (Mins)"
"1"	"3"
"10"	"46"
"100"	"13"
"101"	"12"
"102"	"14"
"104"	"23"
"105"	"20"
"106"	"12"
"107"	"12"
"108"	"14"
"109"	"7"

"11" "13"
"110" "3"
"111" "11"
"112" "12"
"113" "13"
"114" "13"
"115" "14"
"116" "13"
"117" "19"
"118" "12"
"119" "13"
"12" "20"
"120" "14"
"121" "14"
"122" "26"
"123" "14"
"124" "21"
"125" "14"
"126" "14"
"127" "13"
"128" "15"
"129" "12"
"13" "17"
"130" "32"
"131" "14"
"132" "40"
"133" "14"
"134" "15"
"135" "13"
"136" "11"
"137" "12"
"138" "34"
"139" "18"
"14" "14"
"140" "12"
"141" "11"
"142" "12"

"143" "12"
"144" "14"
"145" "11"
"146" "13"
"147" "13"
"148" "14"
"149" "12"
"15" "15"
"150" "18"
"151" "12"
"152" "12"
"153" "12"
"154" "24"
"155" "19"
"156" "15"
"157" "18"
"158" "14"
"159" "12"
"16" "12"
"160" "15"
"161" "13"
"162" "13"
"163" "14"
"164" "13"
"165" "14"
"166" "13"
"167" "12"
"168" "11"
"169" "13"
"17" "11"
"170" "13"
"171" "13"
"172" "16"
"173" "12"
"174" "13"
"175" "13"
"176" "31"

"177" "16"
"178" "7"
"179" "12"
"18" "12"
"180" "21"
"181" "13"
"182" "13"
"183" "12"
"184" "17"
"185" "12"
"186" "14"
"187" "8"
"188" "14"
"189" "13"
"19" "12"
"190" "15"
"191" "13"
"192" "16"
"193" "10"
"194" "21"
"195" "18"
"196" "16"
"197" "13"
"198" "13"
"199" "17"
"2" "38"
"20" "14"
"200" "13"
"201" "9"
"202" "14"
"203" "21"
"204" "4"
"205" "17"
"206" "13"
"207" "5"
"208" "14"
"209" "16"

"21" "16"
"210" "15"
"211" "14"
"212" "12"
"213" "15"
"214" "9"
"215" "44"
"216" "26"
"217" "11"
"218" "19"
"219" "40"
"22" "20"
"220" "12"
"221" "14"
"222" "24"
"223" "13"
"224" "12"
"225" "12"
"226" "15"
"227" "14"
"228" "14"
"229" "12"
"23" "12"
"230" "14"
"231" "15"
"232" "14"
"233" "13"
"234" "13"
"235" "12"
"236" "11"
"237" "11"
"238" "11"
"239" "11"
"24" "12"
"240" "14"
"241" "13"
"242" "12"

"243" "13"
"244" "14"
"245" "9"
"246" "14"
"247" "14"
"248" "13"
"249" "13"
"25" "13"
"250" "14"
"251" "12"
"252" "19"
"253" "21"
"254" "13"
"255" "13"
"256" "13"
"257" "13"
"258" "17"
"259" "10"
"26" "11"
"260" "13"
"261" "18"
"262" "11"
"263" "11"
"264" "13"
"265" "5"
"27" "9"
"28" "26"
"29" "14"
"3" "13"
"30" "25"
"31" "19"
"32" "13"
"33" "16"
"34" "14"
"35" "15"
"36" "13"
"37" "12"

"38" "28"
"39" "16"
"4" "13"
"40" "15"
"41" "11"
"42" "11"
"43" "13"
"44" "12"
"45" "16"
"46" "17"
"47" "13"
"48" "12"
"49" "12"
"5" "11"
"50" "13"
"51" "16"
"52" "16"
"53" "17"
"54" "14"
"55" "19"
"56" "18"
"57" "14"
"58" "6"
"59" "17"
"6" "7"
"60" "12"
"61" "13"
"62" "13"
"63" "15"
"64" "13"
"65" "15"
"66" "15"
"67" "12"
"68" "13"
"69" "12"
"7" "11"
"70" "23"

"71" "15"
"72" "16"
"73" "18"
"74" "11"
"75" "11"
"76" "19"
"77" "16"
"78" "12"
"79" "13"
"8" "17"
"80" "12"
"81" "13"
"82" "13"
"83" "12"
"84" "12"
"85" "15"
"86" "12"
"87" "18"
"88" "19"
"89" "14"
"9" "62"
"90" "12"
"91" "21"
"92" "14"
"93" "30"
"94" "12"
"95" "16"
"96" "16"
"97" "13"
"98" "12"
"99" "13"

Inference:

The above output shows pickup location id and average trip time.

e. Calculate the average tips to revenue ratio of the drivers for different locations in sorted format.

```
[hadoop@ip-172-31-82-64 ~]$ python mrtask_e.py -r hadoop hdfs:///user/hbase/csv >
out_e.txt
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 2.10.1
Looking for Hadoop streaming jar in /home/hadoop/contrib...
Looking for Hadoop streaming jar in /usr/lib/hadoop-mapreduce...
Found Hadoop streaming jar: /usr/lib/hadoop-mapreduce/hadoop-streaming.jar
Creating temp directory /tmp/mrtask_e.hadoop.20230116.181717.807571
uploading working dir files to
hdfs:///user/hadoop/tmp/mrjob/mrtask_e.hadoop.20230116.181717.807571/files/wd...
Copying other local files to
hdfs:///user/hadoop/tmp/mrjob/mrtask_e.hadoop.20230116.181717.807571/files/
Running step 1 of 2...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-2.10.1-amzn-4.jar]
/tmp/streamjob3866179702134088411.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-82-64.ec2.internal/172.31.82.64:8032
  Connecting to Application History server at ip-172-31-82-
64.ec2.internal/172.31.82.64:10200
  Connecting to ResourceManager at ip-172-31-82-64.ec2.internal/172.31.82.64:8032
  Connecting to Application History server at ip-172-31-82-
64.ec2.internal/172.31.82.64:10200
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev
049362b7cf53ff5f739d6b1532457f2c6cd495e8]
  Total input files to process : 6
  Adding a new node: /default-rack/172.31.80.224:50010
  Adding a new node: /default-rack/172.31.86.34:50010
  number of splits:43
  Submitting tokens for job: job_1673889736878_0009
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
```

Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
Adding resource type - name = vcores, units = , type = COUNTABLE
Submitted application application_1673889736878_0009
The url to track the job: http://ip-172-31-82-64.ec2.internal:20888/proxy/application_1673889736878_0009/
Running job: job_1673889736878_0009
Job job_1673889736878_0009 running in uber mode : false
map 0% reduce 0%
map 2% reduce 0%
map 3% reduce 0%
map 4% reduce 0%
map 5% reduce 0%
map 7% reduce 0%
map 9% reduce 0%
map 11% reduce 0%
map 14% reduce 0%
map 15% reduce 0%
map 16% reduce 0%
map 17% reduce 0%
map 18% reduce 0%
map 19% reduce 0%
map 20% reduce 0%
map 21% reduce 0%
map 22% reduce 0%
map 23% reduce 0%
map 24% reduce 0%
map 25% reduce 0%
map 26% reduce 0%
map 27% reduce 0%
map 28% reduce 0%
map 29% reduce 0%
map 30% reduce 0%
map 31% reduce 0%
map 32% reduce 0%
map 33% reduce 0%
map 34% reduce 0%
map 35% reduce 0%

map 36% reduce 0%
map 37% reduce 0%
map 38% reduce 0%
map 39% reduce 0%
map 40% reduce 0%
map 41% reduce 0%
map 42% reduce 0%
map 43% reduce 0%
map 44% reduce 0%
map 46% reduce 0%
map 47% reduce 0%
map 48% reduce 0%
map 49% reduce 0%
map 50% reduce 0%
map 51% reduce 0%
map 51% reduce 6%
map 52% reduce 6%
map 53% reduce 6%
map 54% reduce 6%
map 55% reduce 6%
map 56% reduce 6%
map 57% reduce 6%
map 58% reduce 6%
map 58% reduce 0%
map 60% reduce 0%
map 61% reduce 0%
map 63% reduce 0%
map 65% reduce 0%
map 66% reduce 0%
map 67% reduce 0%
map 69% reduce 0%
map 70% reduce 0%
map 71% reduce 0%
map 72% reduce 0%
map 73% reduce 0%
map 74% reduce 0%
map 75% reduce 0%

map 76% reduce 0%
map 77% reduce 0%
map 78% reduce 0%
map 79% reduce 0%
map 80% reduce 0%
map 81% reduce 0%
map 82% reduce 0%
map 83% reduce 0%
map 84% reduce 0%
map 85% reduce 0%
map 86% reduce 0%
map 87% reduce 0%
map 88% reduce 0%
map 89% reduce 0%
map 90% reduce 0%
map 91% reduce 0%
map 92% reduce 0%
map 93% reduce 0%
map 94% reduce 0%
map 95% reduce 0%
map 97% reduce 0%
map 98% reduce 0%
map 100% reduce 0%
map 100% reduce 22%
map 100% reduce 24%
map 100% reduce 46%
map 100% reduce 47%
map 100% reduce 49%
map 100% reduce 50%
map 100% reduce 51%
map 100% reduce 52%
map 100% reduce 53%
map 100% reduce 54%
map 100% reduce 55%
map 100% reduce 56%
map 100% reduce 57%
map 100% reduce 58%

map 100% reduce 59%
map 100% reduce 60%
map 100% reduce 61%
map 100% reduce 62%
map 100% reduce 63%
map 100% reduce 86%
map 100% reduce 87%
map 100% reduce 88%
map 100% reduce 89%
map 100% reduce 90%
map 100% reduce 91%
map 100% reduce 92%
map 100% reduce 93%
map 100% reduce 94%
map 100% reduce 95%
map 100% reduce 96%
map 100% reduce 97%
map 100% reduce 98%
map 100% reduce 99%
map 100% reduce 100%
Job job_1673889736878_0009 completed successfully
Output directory:
hdfs:///user/hadoop/tmp/mrjob/mrtask_e.hadoop.20230116.181717.807571/step-output/0000
Counters: 51
File Input Format Counters
Bytes Read=5558093822
File Output Format Counters
Bytes Written=4866
File System Counters
FILE: Number of bytes read=339836727
FILE: Number of bytes written=691057536
FILE: Number of large read operations=0
FILE: Number of read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=5558099627
HDFS: Number of bytes written=4866
HDFS: Number of large read operations=0

HDFS: Number of read operations=138

HDFS: Number of write operations=6

Job Counters

Data-local map tasks=23

Killed reduce tasks=2

Launched map tasks=43

Launched reduce tasks=5

Rack-local map tasks=20

Total megabyte-milliseconds taken by all map tasks=2876138496

Total megabyte-milliseconds taken by all reduce tasks=2613651456

Total time spent by all map tasks (ms)=936243

Total time spent by all maps in occupied slots (ms)=89879328

Total time spent by all reduce tasks (ms)=425399

Total time spent by all reduces in occupied slots (ms)=81676608

Total vcore-milliseconds taken by all map tasks=936243

Total vcore-milliseconds taken by all reduce tasks=425399

Map-Reduce Framework

CPU time spent (ms)=1146450

Combine input records=0

Combine output records=0

Failed Shuffles=0

GC time elapsed (ms)=5848

Input split bytes=5805

Map input records=58982297

Map output bytes=1449089423

Map output materialized bytes=340864745

Map output records=58982291

Merged Map outputs=129

Physical memory (bytes) snapshot=27487907840

Reduce input groups=264

Reduce input records=58982291

Reduce output records=264

Reduce shuffle bytes=340864745

Shuffled Maps =129

Spilled Records=117964582

Total committed heap usage (bytes)=26182418432

Virtual memory (bytes) snapshot=221685092352

```
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
Running step 2 of 2...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-2.10.1-amzn-4.jar]
/tmp/streamjob7345052595375882850.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-82-64.ec2.internal/172.31.82.64:8032
  Connecting to Application History server at ip-172-31-82-
64.ec2.internal/172.31.82.64:10200
  Connecting to ResourceManager at ip-172-31-82-64.ec2.internal/172.31.82.64:8032
  Connecting to Application History server at ip-172-31-82-
64.ec2.internal/172.31.82.64:10200
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev
049362b7cf53ff5f739d6b1532457f2c6cd495e8]
  Total input files to process : 3
  number of splits:9
  Submitting tokens for job: job_1673889736878_0011
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
  Adding resource type - name = vcores, units = , type = COUNTABLE
  Submitted application application_1673889736878_0011
  The url to track the job: http://ip-172-31-82-
64.ec2.internal:20888/proxy/application_1673889736878_0011/
  Running job: job_1673889736878_0011
  Job job_1673889736878_0011 running in uber mode : false
   map 0% reduce 0%
   map 11% reduce 0%
   map 22% reduce 0%
   map 33% reduce 0%
   map 44% reduce 0%
   map 67% reduce 0%
```

map 100% reduce 0%
map 100% reduce 67%
map 100% reduce 100%
Job job_1673889736878_0011 completed successfully
Output directory:
hdfs:///user/hadoop/tmp/mrjob/mrtask_e.hadoop.20230116.181717.807571/output
Counters: 51
File Input Format Counters
Bytes Read=9126
File Output Format Counters
Bytes Written=2800
File System Counters
FILE: Number of bytes read=1744
FILE: Number of bytes written=2705553
FILE: Number of large read operations=0
FILE: Number of read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=10764
HDFS: Number of bytes written=2800
HDFS: Number of large read operations=0
HDFS: Number of read operations=36
HDFS: Number of write operations=6
Job Counters
Data-local map tasks=2
Killed map tasks=1
Launched map tasks=9
Launched reduce tasks=3
Rack-local map tasks=7
Total megabyte-milliseconds taken by all map tasks=90083328
Total megabyte-milliseconds taken by all reduce tasks=56150016
Total time spent by all map tasks (ms)=29324
Total time spent by all maps in occupied slots (ms)=2815104
Total time spent by all reduce tasks (ms)=9139
Total time spent by all reduces in occupied slots (ms)=1754688
Total vcore-milliseconds taken by all map tasks=29324
Total vcore-milliseconds taken by all reduce tasks=9139
Map-Reduce Framework

```
                CPU time spent (ms)=8280
                Combine input records=0
                Combine output records=0
                Failed Shuffles=0
                GC time elapsed (ms)=1117
                Input split bytes=1638
                Map input records=264
                Map output bytes=4866
                Map output materialized bytes=2729
                Map output records=264
                Merged Map outputs=27
                Physical memory (bytes) snapshot=5585465344
                Reduce input groups=1
                Reduce input records=264
                Reduce output records=265
                Reduce shuffle bytes=2729
                Shuffled Maps =27
                Spilled Records=528
                Total committed heap usage (bytes)=5338824704
                Virtual memory (bytes) snapshot=63787819008
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
    job output is in
    hdfs:///user/hadoop/tmp/mrjob/mrtask_e.hadoop.20230116.181717.807571/output
    Streaming final output from
    hdfs:///user/hadoop/tmp/mrjob/mrtask_e.hadoop.20230116.181717.807571/output...
    Removing HDFS temp directory
    hdfs:///user/hadoop/tmp/mrjob/mrtask_e.hadoop.20230116.181717.807571...
    Removing temp directory /tmp/mrtask_e.hadoop.20230116.181717.807571...
[hadoop@ip-172-31-82-64 ~]$
```

Output:

"location Id" "Average tips to revenue ratio"

"1" 0.1
"10" 0.1
"100" 0.09
"101" 0.06
"102" 0.06
"104" 0.2
"105" 0.09
"106" 0.09
"107" 0.11
"108" 0.07
"109" 0.13
"11" 0.05
"110" 0.0
"111" 0.09
"112" 0.1
"113" 0.11
"114" 0.11
"115" 0.09
"116" 0.07
"117" 0.11
"118" 0.09
"119" 0.04
"12" 0.08
"120" 0.06
"121" 0.06
"122" 0.07
"123" 0.07
"124" 0.09
"125" 0.12
"126" 0.05
"127" 0.06
"128" 0.08
"129" 0.04
"13" 0.12
"130" 0.08
"131" 0.06
"132" 0.09

"133" 0.08
"134" 0.07
"135" 0.06
"136" 0.03
"137" 0.11
"138" 0.12
"139" 0.07
"14" 0.07
"140" 0.1
"141" 0.1
"142" 0.1
"143" 0.11
"144" 0.11
"145" 0.07
"146" 0.07
"147" 0.03
"148" 0.11
"149" 0.07
"15" 0.09
"150" 0.07
"151" 0.1
"152" 0.07
"153" 0.04
"154" 0.08
"155" 0.06
"156" 0.09
"157" 0.07
"158" 0.11
"159" 0.03
"16" 0.09
"160" 0.07
"161" 0.11
"162" 0.11
"163" 0.1
"164" 0.1
"165" 0.05
"166" 0.1

"167" 0.03
"168" 0.03
"169" 0.04
"17" 0.07
"170" 0.11
"171" 0.05
"172" 0.13
"173" 0.03
"174" 0.04
"175" 0.1
"176" 0.13
"177" 0.05
"178" 0.03
"179" 0.07
"18" 0.04
"180" 0.07
"181" 0.1
"182" 0.04
"183" 0.05
"184" 0.06
"185" 0.05
"186" 0.1
"187" 0.1
"188" 0.06
"189" 0.1
"19" 0.05
"190" 0.09
"191" 0.06
"192" 0.06
"193" 0.04
"194" 0.11
"195" 0.09
"196" 0.06
"197" 0.06
"198" 0.07
"199" 0.12
"2" 0.09

"20" 0.05
"200" 0.07
"201" 0.09
"202" 0.08
"203" 0.07
"204" 0.07
"205" 0.06
"206" 0.06
"207" 0.02
"208" 0.05
"209" 0.1
"21" 0.06
"210" 0.08
"211" 0.11
"212" 0.04
"213" 0.05
"214" 0.07
"215" 0.1
"216" 0.08
"217" 0.06
"218" 0.07
"219" 0.08
"22" 0.05
"220" 0.05
"221" 0.1
"222" 0.07
"223" 0.08
"224" 0.11
"225" 0.07
"226" 0.07
"227" 0.07
"228" 0.08
"229" 0.1
"23" 0.08
"230" 0.1
"231" 0.11
"232" 0.1

"233" 0.11
"234" 0.11
"235" 0.03
"236" 0.1
"237" 0.1
"238" 0.1
"239" 0.11
"24" 0.09
"240" 0.04
"241" 0.04
"242" 0.05
"243" 0.07
"244" 0.07
"245" 0.07
"246" 0.11
"247" 0.06
"248" 0.04
"249" 0.11
"25" 0.1
"250" 0.04
"251" 0.09
"252" 0.09
"253" 0.06
"254" 0.03
"255" 0.11
"256" 0.1
"257" 0.09
"258" 0.06
"259" 0.05
"26" 0.04
"260" 0.05
"261" 0.1
"262" 0.1
"263" 0.1
"264" 0.1
"265" 0.09
"27" 0.1

"28"	0.08
"29"	0.07
"3"	0.05
"30"	0.16
"31"	0.07
"32"	0.04
"33"	0.11
"34"	0.1
"35"	0.06
"36"	0.09
"37"	0.09
"38"	0.05
"39"	0.07
"4"	0.1
"40"	0.11
"41"	0.08
"42"	0.05
"43"	0.1
"44"	0.06
"45"	0.09
"46"	0.04
"47"	0.02
"48"	0.1
"49"	0.08
"5"	0.08
"50"	0.1
"51"	0.05
"52"	0.12
"53"	0.05
"54"	0.11
"55"	0.08
"56"	0.06
"57"	0.07
"58"	0.08
"59"	0.03
"6"	0.04
"60"	0.03

"61"	0.07
"62"	0.06
"63"	0.09
"64"	0.08
"65"	0.1
"66"	0.11
"67"	0.07
"68"	0.11
"69"	0.03
"7"	0.06
"70"	0.07
"71"	0.07
"72"	0.06
"73"	0.05
"74"	0.06
"75"	0.08
"76"	0.07
"77"	0.08
"78"	0.04
"79"	0.11
"8"	0.1
"80"	0.1
"81"	0.05
"82"	0.04
"83"	0.04
"84"	0.08
"85"	0.05
"86"	0.06
"87"	0.12
"88"	0.11
"89"	0.06
"9"	0.05
"90"	0.11
"91"	0.05
"92"	0.04
"93"	0.1
"94"	0.02

"95" 0.06

"96" 0.08

"97" 0.09

"98" 0.06

"99" 0.09

Inference:

The above output shows different memory locations and average ratio of tips to total revenue.

f. How does revenue vary over time? Calculate the average trip revenue per month - analysing it by hour of the day (day vs night) and the day of the week (weekday vs weekend).

```
[hadoop@ip-172-31-82-64 ~]$ python mrtask_f.py -r hadoop hdfs:///user/hbase/csv >
out_f.txt
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 2.10.1
Looking for Hadoop streaming jar in /home/hadoop/contrib...
Looking for Hadoop streaming jar in /usr/lib/hadoop-mapreduce...
Found Hadoop streaming jar: /usr/lib/hadoop-mapreduce/hadoop-streaming.jar
Creating temp directory /tmp/mrtask_f.hadoop.20230116.181733.820703
uploading working dir files to
hdfs:///user/hadoop/tmp/mrjob/mrtask_f.hadoop.20230116.181733.820703/files/wd...
Copying other local files to
hdfs:///user/hadoop/tmp/mrjob/mrtask_f.hadoop.20230116.181733.820703/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-2.10.1-amzn-4.jar]
/tmp/streamjob4689407670852389096.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-82-64.ec2.internal/172.31.82.64:8032
  Connecting to Application History server at ip-172-31-82-
64.ec2.internal/172.31.82.64:10200
  Connecting to ResourceManager at ip-172-31-82-64.ec2.internal/172.31.82.64:8032
  Connecting to Application History server at ip-172-31-82-
64.ec2.internal/172.31.82.64:10200
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev
049362b7cf53ff5f739d6b1532457f2c6cd495e8]
  Total input files to process : 6
  Adding a new node: /default-rack/172.31.80.224:50010
  Adding a new node: /default-rack/172.31.86.34:50010
  number of splits:43
  Submitting tokens for job: job_1673889736878_0010
  resource-types.xml not found
```

Unable to find 'resource-types.xml'.
Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
Adding resource type - name = vcores, units = , type = COUNTABLE
Submitted application application_1673889736878_0010
The url to track the job: http://ip-172-31-82-64.ec2.internal:20888/proxy/application_1673889736878_0010/
Running job: job_1673889736878_0010
Job job_1673889736878_0010 running in uber mode : false
map 0% reduce 0%
map 1% reduce 0%
map 2% reduce 0%
map 3% reduce 0%
map 4% reduce 0%
map 5% reduce 0%
map 6% reduce 0%
map 7% reduce 0%
map 8% reduce 0%
map 9% reduce 0%
map 10% reduce 0%
map 11% reduce 0%
map 12% reduce 0%
map 13% reduce 0%
map 14% reduce 0%
map 15% reduce 0%
map 16% reduce 0%
map 17% reduce 0%
map 19% reduce 0%
map 20% reduce 0%
map 21% reduce 0%
map 22% reduce 0%
map 24% reduce 0%
map 25% reduce 0%
map 26% reduce 0%
map 27% reduce 0%
map 28% reduce 0%
map 29% reduce 0%
map 30% reduce 0%

map 31% reduce 0%
map 32% reduce 0%
map 33% reduce 0%
map 34% reduce 0%
map 35% reduce 0%
map 36% reduce 0%
map 37% reduce 0%
map 38% reduce 0%
map 39% reduce 0%
map 40% reduce 0%
map 41% reduce 0%
map 42% reduce 0%
map 43% reduce 0%
map 45% reduce 0%
map 48% reduce 0%
map 49% reduce 0%
map 50% reduce 0%
map 51% reduce 0%
map 51% reduce 6%
map 52% reduce 6%
map 53% reduce 6%
map 54% reduce 6%
map 56% reduce 6%
map 57% reduce 6%
map 58% reduce 6%
map 59% reduce 6%
map 60% reduce 6%
map 60% reduce 7%
map 62% reduce 7%
map 63% reduce 7%
map 65% reduce 7%
map 67% reduce 7%
map 68% reduce 7%
map 70% reduce 7%
map 70% reduce 8%
map 71% reduce 8%
map 72% reduce 8%

map 74% reduce 8%
map 77% reduce 8%
map 77% reduce 9%
map 78% reduce 9%
map 79% reduce 9%
map 80% reduce 9%
map 81% reduce 9%
map 82% reduce 9%
map 83% reduce 9%
map 84% reduce 9%
map 85% reduce 9%
map 86% reduce 9%
map 87% reduce 10%
map 88% reduce 10%
map 89% reduce 10%
map 90% reduce 10%
map 91% reduce 10%
map 92% reduce 10%
map 93% reduce 10%
map 94% reduce 10%
map 95% reduce 10%
map 96% reduce 10%
map 96% reduce 11%
map 97% reduce 11%
map 98% reduce 11%
map 100% reduce 11%
map 100% reduce 18%
map 100% reduce 22%
map 100% reduce 44%
map 100% reduce 50%
map 100% reduce 55%
map 100% reduce 61%
map 100% reduce 67%
map 100% reduce 89%
map 100% reduce 94%
map 100% reduce 100%
Job job_1673889736878_0010 completed successfully

Output directory:
hdfs:///user/hadoop/tmp/mrjob/mrtask_f.hadoop.20230116.181733.820703/output
Counters: 52
File Input Format Counters
Bytes Read=5558093822
File Output Format Counters
Bytes Written=1208
File System Counters
FILE: Number of bytes read=297567638
FILE: Number of bytes written=605490794
FILE: Number of large read operations=0
FILE: Number of read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=5558099627
HDFS: Number of bytes written=1208
HDFS: Number of large read operations=0
HDFS: Number of read operations=138
HDFS: Number of write operations=6
Job Counters
Data-local map tasks=25
Killed map tasks=1
Killed reduce tasks=1
Launched map tasks=43
Launched reduce tasks=4
Rack-local map tasks=18
Total megabyte-milliseconds taken by all map tasks=3673015296
Total megabyte-milliseconds taken by all reduce tasks=3911823360
Total time spent by all map tasks (ms)=1195643
Total time spent by all maps in occupied slots (ms)=114781728
Total time spent by all reduce tasks (ms)=636690
Total time spent by all reduces in occupied slots (ms)=122244480
Total vcore-milliseconds taken by all map tasks=1195643
Total vcore-milliseconds taken by all reduce tasks=636690
Map-Reduce Framework
CPU time spent (ms)=1410230
Combine input records=0
Combine output records=0

```
                Failed Shuffles=0
                GC time elapsed (ms)=5940
                Input split bytes=5805
                Map input records=58982297
                Map output bytes=1066685266
                Map output materialized bytes=297567555
                Map output records=58982291
                Merged Map outputs=129
                Physical memory (bytes) snapshot=27529572352
                Reduce input groups=6
                Reduce input records=58982291
                Reduce output records=24
                Reduce shuffle bytes=297567555
                Shuffled Maps =129
                Spilled Records=117964582
                Total committed heap usage (bytes)=26003111936
                Virtual memory (bytes) snapshot=221721579520
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
    job output is in
    hdfs:///user/hadoop/tmp/mrjob/mrtask_f.hadoop.20230116.181733.820703/output
    Streaming final output from
    hdfs:///user/hadoop/tmp/mrjob/mrtask_f.hadoop.20230116.181733.820703/output...
    Removing HDFS temp directory
    hdfs:///user/hadoop/tmp/mrjob/mrtask_f.hadoop.20230116.181733.820703...
    Removing temp directory /tmp/mrtask_f.hadoop.20230116.181733.820703...
[hadoop@ip-172-31-82-64 ~]$
```

Output

```
    "JUN"   ["Week Day, Day Time ", 70551188.3473795]
    "JUN"   ["Week Day, Night Time ", 42252581.76976587]
    "JUN"   ["Week End, Day Time ", 27994198.28043421]
    "JUN"   ["Week End, Night Time ", 19955446.610244542]
```

"MAY" ["Week Day, Day Time ", 79118571.85658517]
"MAY" ["Week Day, Night Time ", 45380213.64955819]
"MAY" ["Week End, Day Time ", 25898374.800398573]
"MAY" ["Week End, Night Time ", 18994202.510225344]
"FEB" ["Week Day, Day Time ", 61468276.18807749]
"FEB" ["Week Day, Night Time ", 37696635.63019301]
"FEB" ["Week End, Day Time ", 25100766.910404745]
"FEB" ["Week End, Night Time ", 19078854.68022781]
"JAN" ["Week Day, Day Time ", 68701153.1271197]
"JAN" ["Week Day, Night Time ", 41031570.42990915]
"JAN" ["Week End, Day Time ", 23453641.330353357]
"JAN" ["Week End, Night Time ", 17589860.340188645]
"APR" ["Week Day, Day Time ", 69988742.72731388]
"APR" ["Week Day, Night Time ", 42582584.20977221]
"APR" ["Week End, Day Time ", 28912111.430452816]
"APR" ["Week End, Night Time ", 22825932.500301514]
"MAR" ["Week Day, Day Time ", 71360108.18715025]
"MAR" ["Week Day, Night Time ", 43099636.98974699]
"MAR" ["Week End, Day Time ", 29846359.800507683]
"MAR" ["Week End, Night Time ", 22723662.000311993]

Inference:

From the output it can be inferred that average trip revenue is highest in the month of May and lowest in the month of February.

It is also inferred that during week days, day time average trip revenue is highest.

It is also inferred that during week ends, night time average trip revenue is lowest.