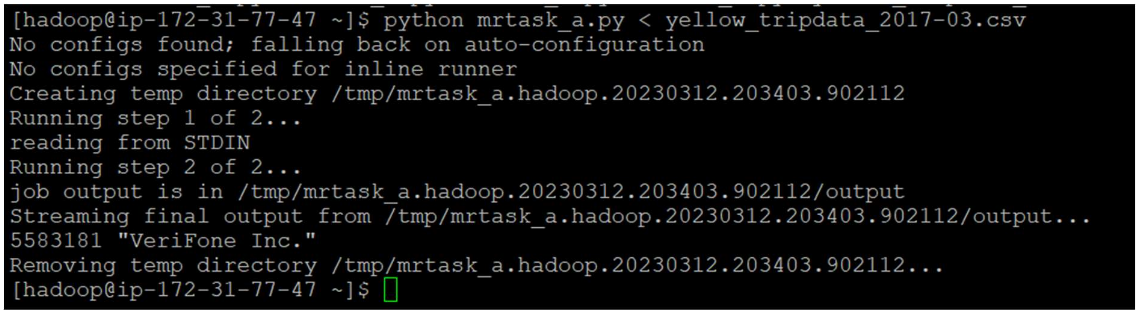
Task 4. Write MapReduce codes to perform the tasks using the files you've downloaded on your EMR Instance:

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

Code File reference: mrtask_a.py

Code execution screenshot:

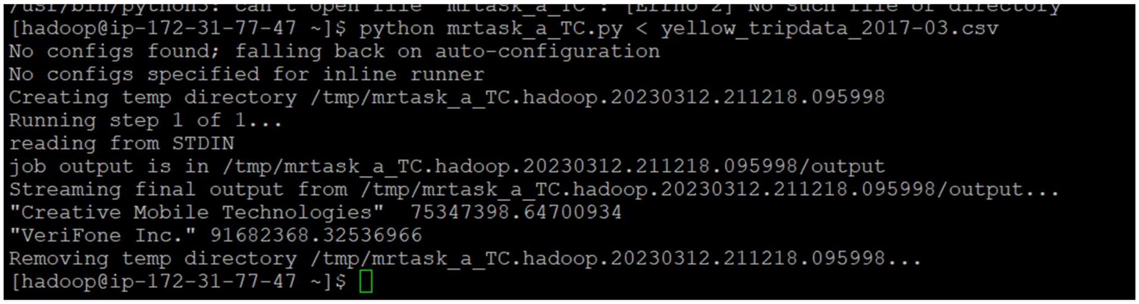


```
[hadoop@ip-172-31-77-47 ~]$ python mrtask_a.py < yellow_tripdata_2017-03.csv
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/mrtask_a.hadoop.20230312.203403.902112
Running step 1 of 2...
reading from STDIN
Running step 2 of 2...
job output is in /tmp/mrtask_a.hadoop.20230312.203403.902112/output
Streaming final output from /tmp/mrtask_a.hadoop.20230312.203403.902112/output...
5583181 "VeriFone Inc."
Removing temp directory /tmp/mrtask_a.hadoop.20230312.203403.902112...
[hadoop@ip-172-31-77-47 ~]$ 
```

Remarks: In yellow_tripdata_2017-03.csv, VeriFone Inc. has most number of trips which is 5583181

Code File reference: mrtask_a_TC.py

Code execution screenshot:



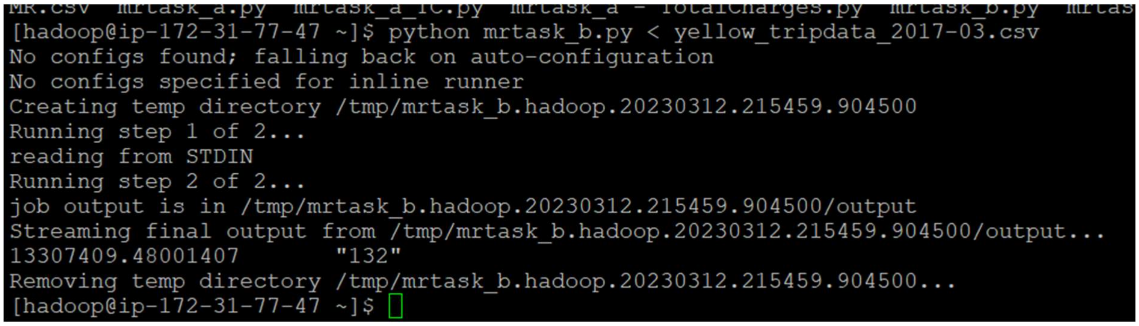
```
/usr/bin/python3: can't open file 'mrtask_a_TC': [Errno 2] No such file or directory
[hadoop@ip-172-31-77-47 ~]$ python mrtask_a_TC.py < yellow_tripdata_2017-03.csv
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/mrtask_a_TC.hadoop.20230312.211218.095998
Running step 1 of 1...
reading from STDIN
job output is in /tmp/mrtask_a_TC.hadoop.20230312.211218.095998/output
Streaming final output from /tmp/mrtask_a_TC.hadoop.20230312.211218.095998/output...
"Creative Mobile Technologies"  75347398.64700934
"VeriFone Inc." 91682368.32536966
Removing temp directory /tmp/mrtask_a_TC.hadoop.20230312.211218.095998...
[hadoop@ip-172-31-77-47 ~]$ 
```

Remarks: VeriFone Inc. has to revenue generated: 91682368.32

b) Which pickup location generates the most revenue?

Code File reference: mrtask_b.py

Code execution screenshot:



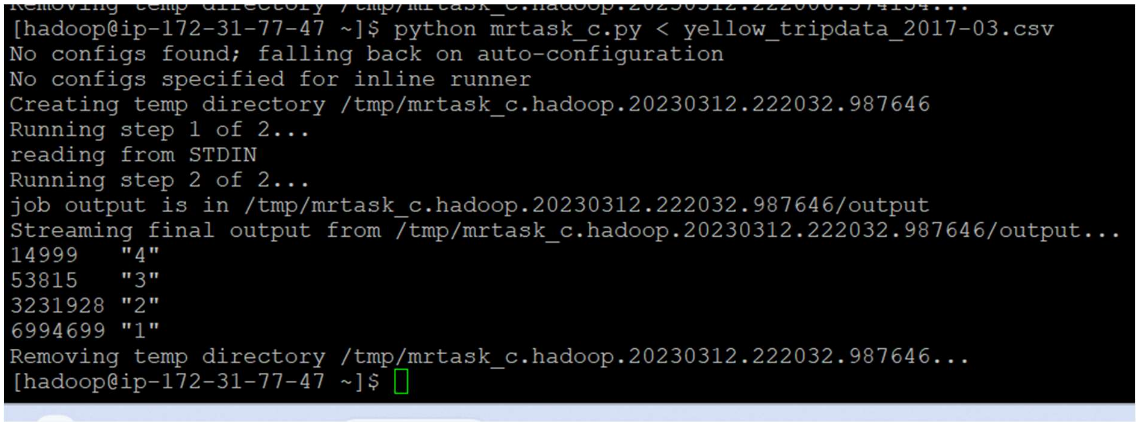
```
MR.csv  mrtask_a.py  mrtask_a_TC.py  mrtask_a - TotalCharges.py  mrtask_b.py  mrtas
[hadoop@ip-172-31-77-47 ~]$ python mrtask_b.py < yellow_tripdata_2017-03.csv
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/mrtask_b.hadoop.20230312.215459.904500
Running step 1 of 2...
reading from STDIN
Running step 2 of 2...
job output is in /tmp/mrtask_b.hadoop.20230312.215459.904500/output
Streaming final output from /tmp/mrtask_b.hadoop.20230312.215459.904500/output...
13307409.48001407       "132"
Removing temp directory /tmp/mrtask_b.hadoop.20230312.215459.904500...
[hadoop@ip-172-31-77-47 ~]$ 
```

Remarks: Location 132 generated most revenue in the file tested.

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

Code File reference: mrtask_c.py

Code execution screenshot:



```
Removing temp directory /tmp/mrtask_c.hadoop.20230312.222000.974134...
[hadoop@ip-172-31-77-47 ~]$ python mrtask_c.py < yellow_tripdata_2017-03.csv
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/mrtask_c.hadoop.20230312.222032.987646
Running step 1 of 2...
reading from STDIN
Running step 2 of 2...
job output is in /tmp/mrtask_c.hadoop.20230312.222032.987646/output
Streaming final output from /tmp/mrtask_c.hadoop.20230312.222032.987646/output...
14999   "4"
53815   "3"
3231928 "2"
6994699 "1"
Removing temp directory /tmp/mrtask_c.hadoop.20230312.222032.987646...
[hadoop@ip-172-31-77-47 ~]$ 
```

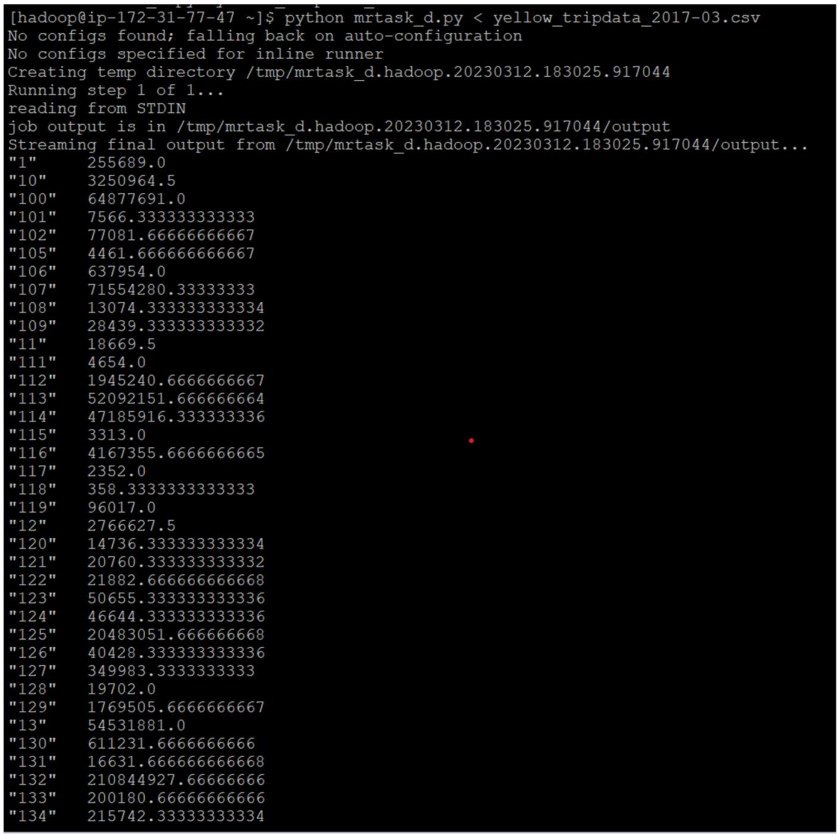
Remark: Credit card is mostly used followed by cash, then no charge and lowest one observed is dispute

1= Credit card, 2= Cash, 3= No charge,,4= Dispute, 5= Unknown, 6= Voided trip

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

Code File reference: mrtask_d.py

Code execution screenshot:

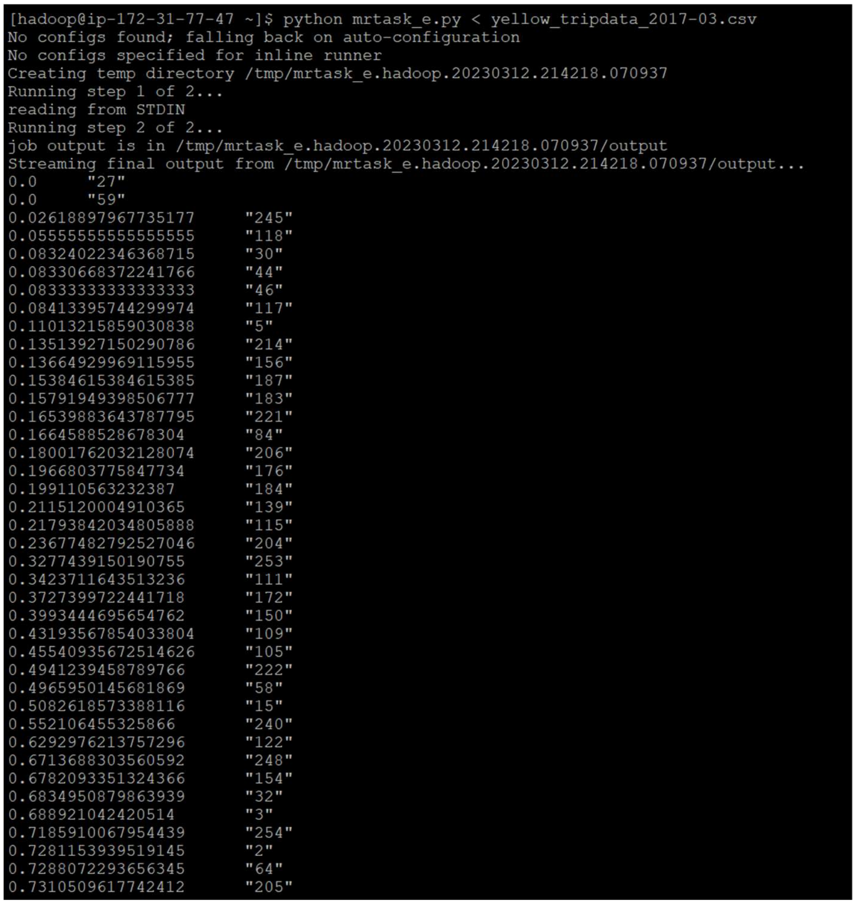


```
[hadoop@ip-172-31-77-47 ~]$ python mrtask_d.py < yellow_tripdata_2017-03.csv
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/mrtask_d.hadoop.20230312.183025.917044
Running step 1 of 1...
reading from STDIN
job output is in /tmp/mrtask_d.hadoop.20230312.183025.917044/output
Streaming final output from /tmp/mrtask_d.hadoop.20230312.183025.917044/output...
"1"     255689.0
"10"    3250964.5
"100"   64877691.0
"101"   7566.333333333333
"102"   77081.66666666667
"105"   4461.666666666667
"106"   637954.0
"107"   71554280.33333333
"108"   13074.333333333334
"109"   28439.333333333332
"11"    18669.5
"111"   4654.0
"112"   1945240.6666666667
"113"   52092151.666666664
"114"   47185916.333333336
"115"   3313.0
"116"   4167355.6666666665
"117"   2352.0
"118"   358.3333333333333
"119"   96017.0
"12"    2766627.5
"120"   14736.333333333334
"121"   20760.333333333332
"122"   21882.666666666668
"123"   50655.333333333336
"124"   46644.333333333336
"125"   20483051.666666668
"126"   40428.333333333336
"127"   349983.3333333333
"128"   19702.0
"129"   1769505.6666666667
"13"    54531881.0
"130"   611231.6666666666
"131"   16631.666666666668
"132"   210844927.66666666
"133"   200180.66666666666
"134"   215742.33333333334
```

Remarks: Running map reduce job on yellow_tripdata_2017-03.csv. Above screenshot has average trip time in seconds for each location.

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

Code File reference: mrtask_e.py



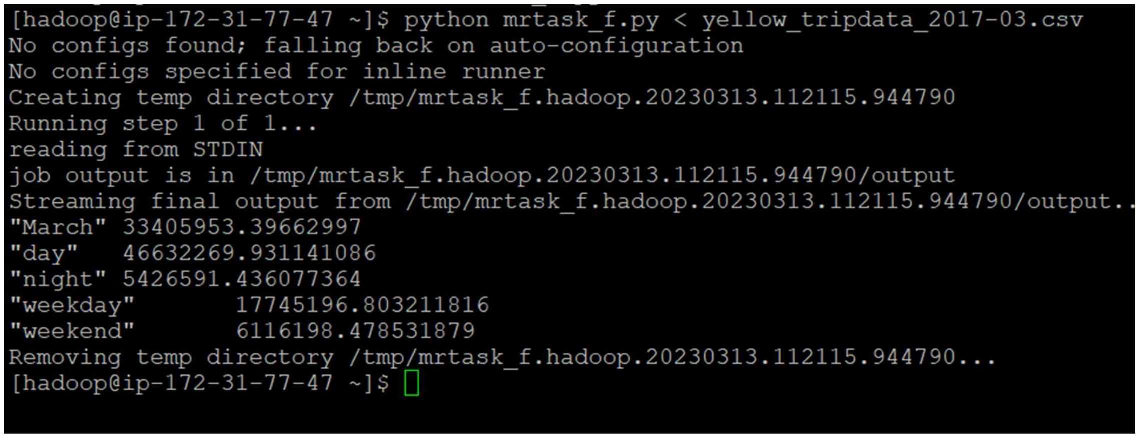
```
[hadoop@ip-172-31-77-47 ~]$ python mrtask_e.py < yellow_tripdata_2017-03.csv
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/mrtask_e.hadoop.20230312.214218.070937
Running step 1 of 2...
reading from STDIN
Running step 2 of 2...
job output is in /tmp/mrtask_e.hadoop.20230312.214218.070937/output
Streaming final output from /tmp/mrtask_e.hadoop.20230312.214218.070937/output...
0.0     "27"
0.0     "59"
0.02618897967735177     "245"
0.05555555555555555     "118"
0.08324022346368715     "30"
0.08330668372241766     "44"
0.08333333333333333     "46"
0.08413395744299974     "117"
0.11013215859030838     "5"
0.13513927150290786     "214"
0.13664929969115955     "156"
0.15384615384615385     "187"
0.15791949398506777     "183"
0.16539883643787795     "221"
0.1664588528678304      "84"
0.18001762032128074     "206"
0.1966803775847734      "176"
0.199110563232387       "184"
0.2115120004910365      "139"
0.21793842034805888     "115"
0.23677482792527046     "204"
0.3277439150190755      "253"
0.3423711643513236      "111"
0.3727399722441718      "172"
0.3993444695654762      "150"
0.43193567854033804     "109"
0.45540935672514626     "105"
0.4941239458789766      "222"
0.4965950145681869      "58"
0.5082618573388116      "15"
0.552106455325866       "240"
0.6292976213757296      "122"
0.6713688303560592      "248"
0.6782093351324366      "154"
0.6834950879863939      "32"
0.688921042420514       "3"
0.7185910067954439      "254"
0.7281153939519145      "2"
0.7288072293656345      "64"
0.7310509617742412      "205"
```

Remarks: Above screenshot is captured after testing code for average tips to revenue ratio for the drivers of different locations

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).

Code File reference: mrtask_f.py

Code execution screenshot:



```
[hadoop@ip-172-31-77-47 ~]$ python mrtask_f.py < yellow_tripdata_2017-03.csv
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/mrtask_f.hadoop.20230313.112115.944790
Running step 1 of 1...
reading from STDIN
job output is in /tmp/mrtask_f.hadoop.20230313.112115.944790/output
Streaming final output from /tmp/mrtask_f.hadoop.20230313.112115.944790/output..
"March" 33405953.39662997
"day"   46632269.931141086
"night" 5426591.436077364
"weekday"       17745196.803211816
"weekend"       6116198.478531879
Removing temp directory /tmp/mrtask_f.hadoop.20230313.112115.944790...
[hadoop@ip-172-31-77-47 ~]$
```

Remarks/Insights: This file is mostly data for March month.

- Average revenue in March month
- Average trip revenue in the day is higher than night. (Assuming night time starts from 11 PM to 5 AM morning.)
- Average trip revenue in the weekday is higher than weekend (assuming Saturday and Sunday are weekends)

Assignment Submitted by:

Susil Patro, Vybhava P, Vivek Agrawal