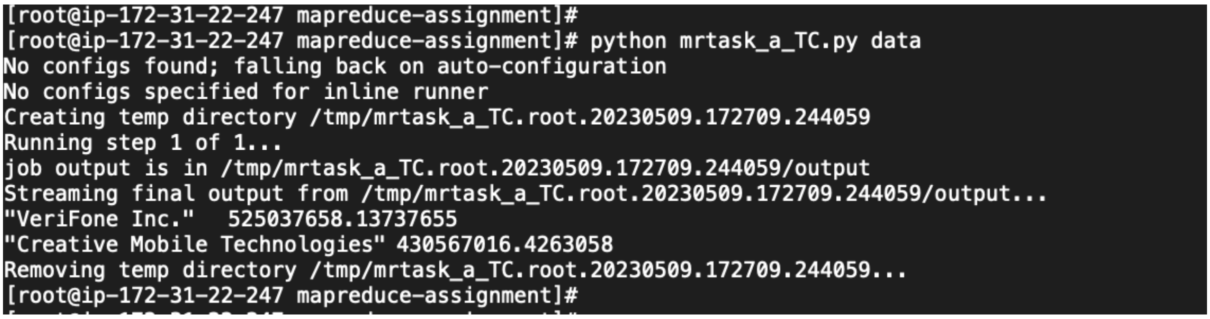
MapReduce tasks

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?

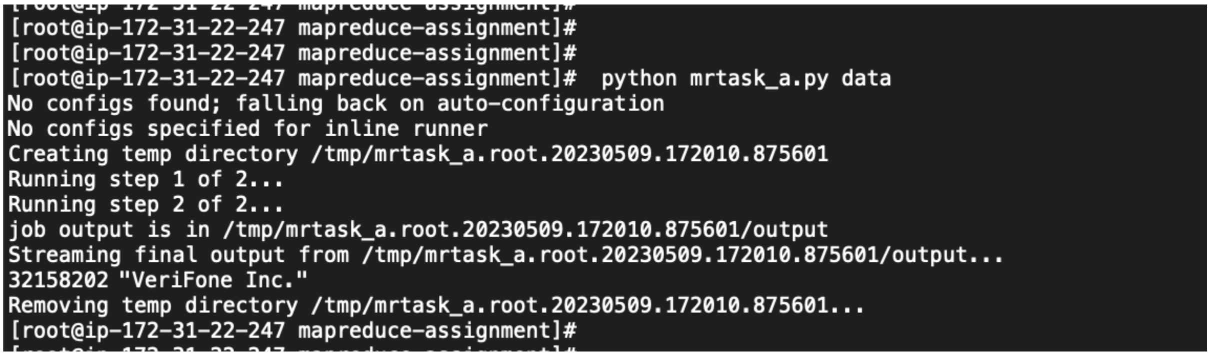
   Code:
   python mrtask_a_TC.py data



```
[root@ip-172-31-22-247 mapreduce-assignment]#
[root@ip-172-31-22-247 mapreduce-assignment]# python mrtask_a_TC.py data
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/mrtask_a_TC.root.20230509.172709.244059
Running step 1 of 1...
job output is in /tmp/mrtask_a_TC.root.20230509.172709.244059/output
Streaming final output from /tmp/mrtask_a_TC.root.20230509.172709.244059/output...
"VeriFone Inc."   525037658.13737655
"Creative Mobile Technologies" 430567016.4263058
Removing temp directory /tmp/mrtask_a_TC.root.20230509.172709.244059...
[root@ip-172-31-22-247 mapreduce-assignment]#
```

   Remarks:
   "VeriFone Inc." has total revenue generated: 525037658.137
   "Creative Mobile Technologies"   has total revenue generated 430567016.426

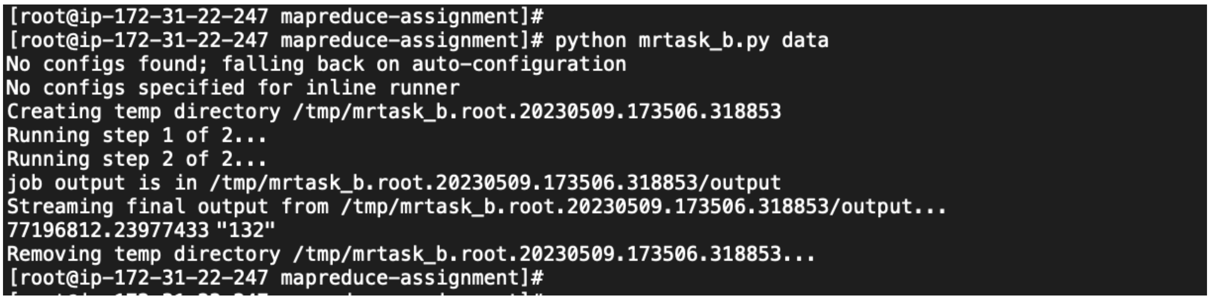
   Code:
   python mrtask_a.py data



```
[root@ip-172-31-22-247 mapreduce-assignment]#
[root@ip-172-31-22-247 mapreduce-assignment]#
[root@ip-172-31-22-247 mapreduce-assignment]#  python mrtask_a.py data
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/mrtask_a.root.20230509.172010.875601
Running step 1 of 2...
Running step 2 of 2...
job output is in /tmp/mrtask_a.root.20230509.172010.875601/output
Streaming final output from /tmp/mrtask_a.root.20230509.172010.875601/output...
32158202 "VeriFone Inc."
Removing temp directory /tmp/mrtask_a.root.20230509.172010.875601...
[root@ip-172-31-22-247 mapreduce-assignment]#
```

   Remarks:
   "VeriFone Inc." has most number of trips which is 32158202

b. Which pickup location generates the most revenue?
   Code:
   python mrtask_b.py data

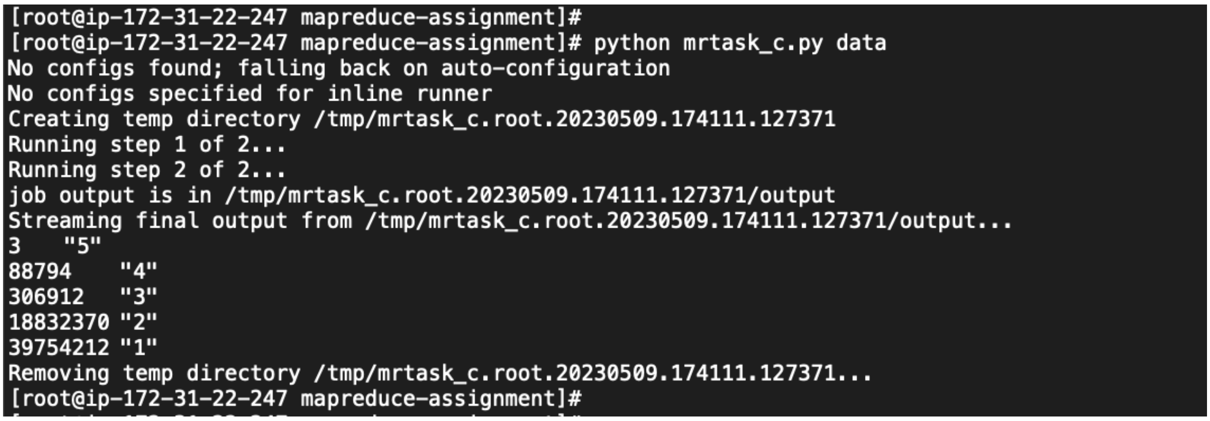


```
[root@ip-172-31-22-247 mapreduce-assignment]#
[root@ip-172-31-22-247 mapreduce-assignment]# python mrtask_b.py data
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/mrtask_b.root.20230509.173506.318853
Running step 1 of 2...
Running step 2 of 2...
job output is in /tmp/mrtask_b.root.20230509.173506.318853/output
Streaming final output from /tmp/mrtask_b.root.20230509.173506.318853/output...
77196812.23977433 "132"
Removing temp directory /tmp/mrtask_b.root.20230509.173506.318853...
[root@ip-172-31-22-247 mapreduce-assignment]#
```

Remarks:
Location 132 generated most revenue in the dataset.

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

Code:
python mrtask_c.py data

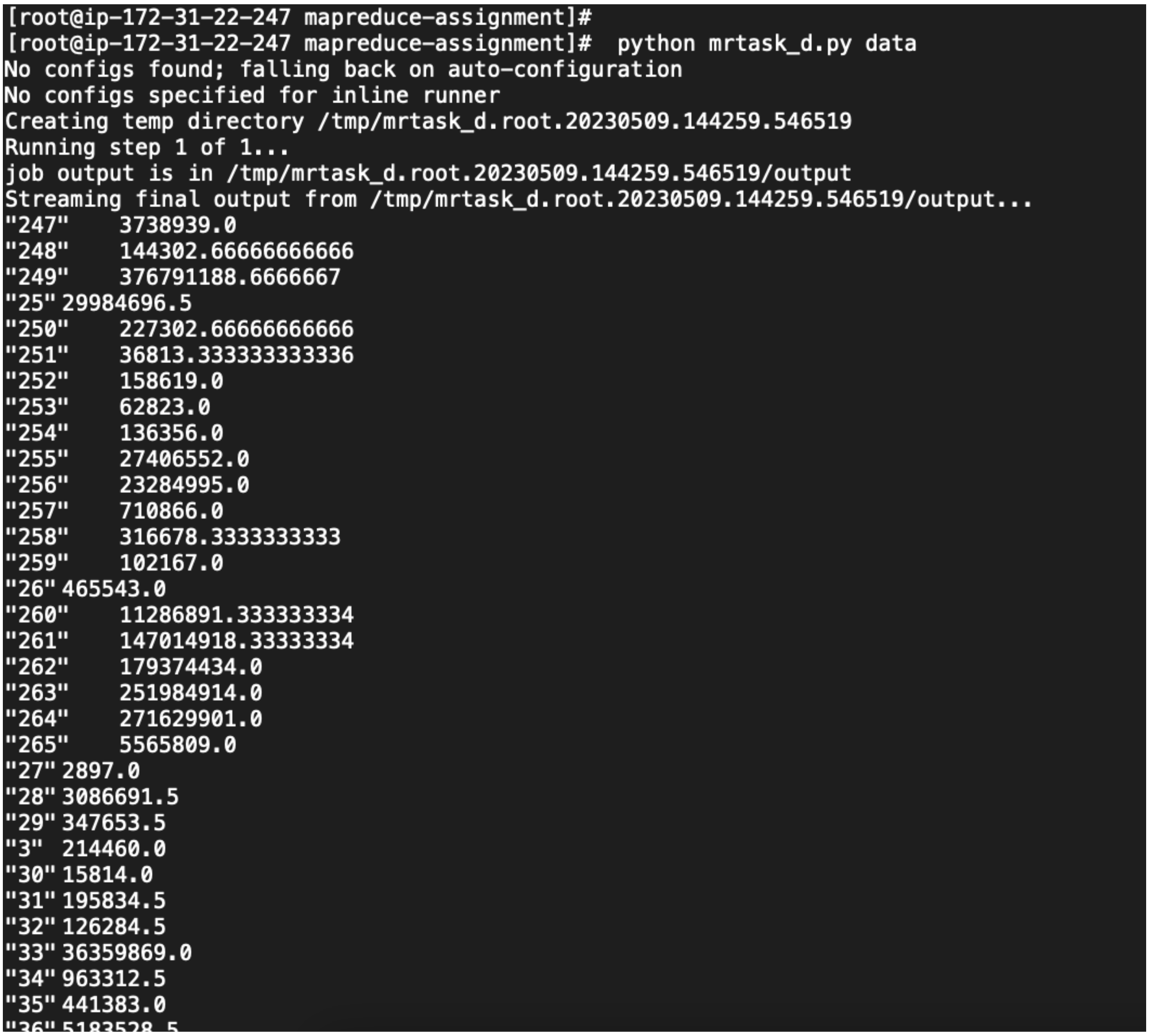


```
[root@ip-172-31-22-247 mapreduce-assignment]#
[root@ip-172-31-22-247 mapreduce-assignment]# python mrtask_c.py data
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/mrtask_c.root.20230509.174111.127371
Running step 1 of 2...
Running step 2 of 2...
job output is in /tmp/mrtask_c.root.20230509.174111.127371/output
Streaming final output from /tmp/mrtask_c.root.20230509.174111.127371/output...
3   "5"
88794    "4"
306912   "3"
18832370 "2"
39754212 "1"
Removing temp directory /tmp/mrtask_c.root.20230509.174111.127371...
[root@ip-172-31-22-247 mapreduce-assignment]#
```

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:
python mrtask_d.py data

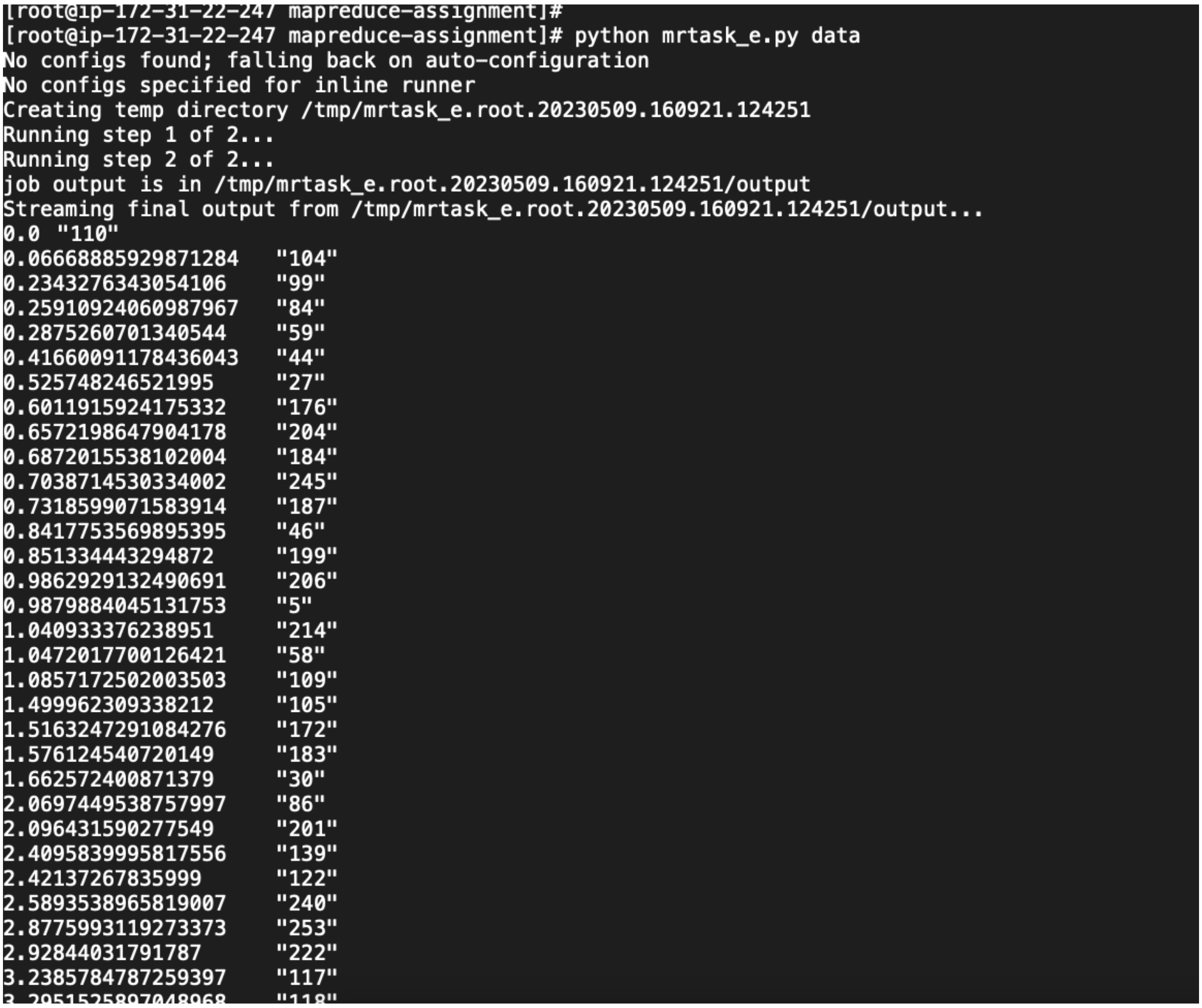


```
[root@ip-172-31-22-247 mapreduce-assignment]#
[root@ip-172-31-22-247 mapreduce-assignment]#  python mrtask_d.py data
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/mrtask_d.root.20230509.144259.546519
Running step 1 of 1...
job output is in /tmp/mrtask_d.root.20230509.144259.546519/output
Streaming final output from /tmp/mrtask_d.root.20230509.144259.546519/output...
"247"    3738939.0
"248"    144302.66666666666
"249"    376791188.6666667
"25" 29984696.5
"250"    227302.66666666666
"251"    36813.333333333336
"252"    158619.0
"253"    62823.0
"254"    136356.0
"255"    27406552.0
"256"    23284995.0
"257"    710866.0
"258"    316678.3333333333
"259"    102167.0
"26" 465543.0
"260"    11286891.333333334
"261"    147014918.33333334
"262"    179374434.0
"263"    251984914.0
"264"    271629901.0
"265"    5565809.0
"27" 2897.0
"28" 3086691.5
"29" 347653.5
"3"  214460.0
"30" 15814.0
"31" 195834.5
"32" 126284.5
"33" 36359869.0
"34" 963312.5
"35" 441383.0
"36" 5183528.5
```

Remarks:
Above screenshot has average trip time in seconds for each location.

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

Code:
python mrtask_e.py data

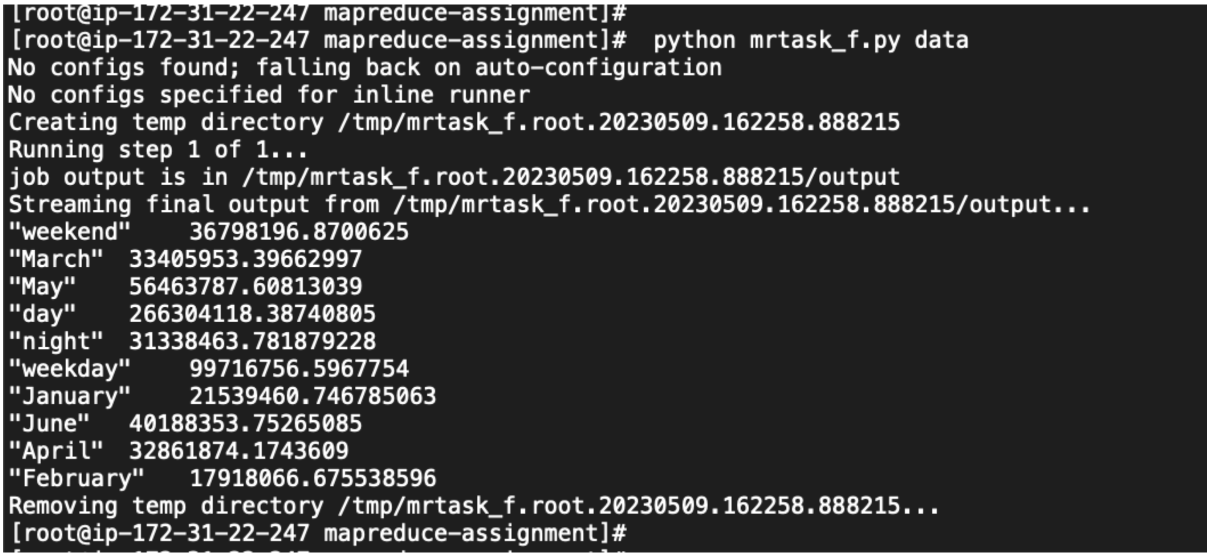


```
[root@ip-172-31-22-247 mapreduce-assignment]#
[root@ip-172-31-22-247 mapreduce-assignment]# python mrtask_e.py data
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/mrtask_e.root.20230509.160921.124251
Running step 1 of 2...
Running step 2 of 2...
job output is in /tmp/mrtask_e.root.20230509.160921.124251/output
Streaming final output from /tmp/mrtask_e.root.20230509.160921.124251/output...
0.0 "110"
0.06668885929871284   "104"
0.2343276343054106    "99"
0.25910924060987967   "84"
0.2875260701340544    "59"
0.41660091178436043   "44"
0.525748246521995     "27"
0.6011915924175332    "176"
0.6572198647904178    "204"
0.6872015538102004    "184"
0.7038714530334002    "245"
0.7318599071583914    "187"
0.8417753569895395    "46"
0.851334443294872     "199"
0.9862929132490691    "206"
0.9879884045131753    "5"
1.040933376238951     "214"
1.0472017700126421    "58"
1.0857172502003503    "109"
1.499962309338212     "105"
1.5163247291084276    "172"
1.576124540720149     "183"
1.662572400871379     "30"
2.0697449538757997    "86"
2.096431590277549     "201"
2.4095839995817556    "139"
2.42137267835999      "122"
2.5893538965819007    "240"
2.8775993119273373    "253"
2.92844031791787      "222"
3.2385784787259397    "117"
3.2951525897048968    "118"
```

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:
python mrtask_f.py data



```
[root@ip-172-31-22-247 mapreduce-assignment]#
[root@ip-172-31-22-247 mapreduce-assignment]#  python mrtask_f.py data
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/mrtask_f.root.20230509.162258.888215
Running step 1 of 1...
job output is in /tmp/mrtask_f.root.20230509.162258.888215/output
Streaming final output from /tmp/mrtask_f.root.20230509.162258.888215/output...
"weekend"	36798196.8700625
"March"	33405953.39662997
"May"	56463787.60813039
"day"	266304118.38740805
"night"	31338463.781879228
"weekday"	99716756.5967754
"January"	21539460.746785063
"June"	40188353.75265085
"April"	32861874.1743609
"February"	17918066.675538596
Removing temp directory /tmp/mrtask_f.root.20230509.162258.888215...
[root@ip-172-31-22-247 mapreduce-assignment]#
```

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

1. Average revenue in March month
2. Average trip revenue in the day is higher than night.
3. Average trip revenue in the weekday is higher than weekend