## MapReduce Tasks:

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

All the tasks below are solved using file yellow_tripdata_2017-03.csv.
Pre-requisite for this task:

1. Install MRJob

pip install mrjob

```
[hadoop@ip-172-31-82-210 ~]$ pip install mrjob
Defaulting to user installation because normal site-packages is not writeable
Collecting mrjob
  Downloading mrjob-0.7.4-py2.py3-none-any.whl (439 kB)
     |                                | 439 kB 27.2 MB/s
Requirement already satisfied: PyYAML>=3.10 in /usr/local/lib64/python3.7/site-p
ackages (from mrjob) (5.4.1)
Installing collected packages: mrjob
  WARNING: The scripts mrjob, mrjob-3 and mrjob-3.7 are installed in '/home/hado
op/.local/bin' which is not on PATH.
  Consider adding this directory to PATH or, if you prefer to suppress this warn
ing, use --no-warn-script-location.
Successfully installed mrjob-0.7.4
```

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

```
[hadoop@ip-172-31-82-210 ~]$ python mrtask_a.py input_data > mrtask_a_out.txt
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/mrtask_a.hadoop.20231121.064136.621384
Running step 1 of 2...
Running step 2 of 2...
job output is in /tmp/mrtask_a.hadoop.20231121.064136.621384/output
Streaming final output from /tmp/mrtask_a.hadoop.20231121.064136.621384/output...
Removing temp directory /tmp/mrtask_a.hadoop.20231121.064136.621384...
[hadoop@ip-172-31-82-210 ~]$ cat mrtask_a_out.txt
"2"     [5583181, 91682368.33028187]
```

b. Which pickup location generates the most revenue?

```
[hadoop@ip-172-31-82-210 ~]$ python mrtask_b.py input_data > mrtask_b_out.txt
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/mrtask_b.hadoop.20231121.064618.748027
Running step 1 of 2...
Running step 2 of 2...
job output is in /tmp/mrtask_b.hadoop.20231121.064618.748027/output
Streaming final output from /tmp/mrtask_b.hadoop.20231121.064618.748027/output...
Removing temp directory /tmp/mrtask_b.hadoop.20231121.064618.748027...
[hadoop@ip-172-31-82-210 ~]$ cat mrtask_b_out.txt
"132"   13307409.48001407
```

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

```
[hadoop@ip-172-31-82-210 ~]$ python mrtask_c.py input_data > mrtask_c_out.txt
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/mrtask_c.hadoop.20231121.065150.858589
Running step 1 of 2...
Running step 2 of 2...
job output is in /tmp/mrtask_c.hadoop.20231121.065150.858589/output
Streaming final output from /tmp/mrtask_c.hadoop.20231121.065150.858589/output.
Removing temp directory /tmp/mrtask_c.hadoop.20231121.065150.858589...
[hadoop@ip-172-31-82-210 ~]$ cat mrtask_c_out.txt
"Credit card"   6994699
"Cash"  3231928
"No charge"     53815
"Dispute"       14999
[hadoop@ip-172-31-82-210 ~]$
```

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

```
[hadoop@ip-172-31-82-210 ~]$ cat mrtask_d_out.txt
"1"     6.105276981852914
"10"    56.11884170550665
"100"   15.604376259723034
"101"   12.203763440860214
"102"   42.82314814814816
"105"   18.59027777777778
"106"   15.196617436874702
"107"   14.326443236908414
"108"   18.1587962962963
"109"   203.13809523809525
"11"    10.917836257309942
"111"   15.513333333333332
"112"   14.553648561025495
"113"   15.095411481968693
"114"   16.23875211075014
"115"   11.043333333333335
"116"   15.671463848776575
"117"   9.8
"118"   2.986111111111111
"119"   13.447759103641456
"12"    27.171749165193486
"120"   13.396666666666665
"121"   13.307905982905982
"122"   57.585964912280694
"123"   12.791750841750844
"124"   16.08425287356322
"125"   16.281976174199638
"126"   16.042989417989418
"127"   15.29647435897436
"128"   14.925757575757574
"129"   13.91558404110307
"13"    19.149716258261186
"130"   38.636641382216595
"131"   10.526371308016879
"132"   44.534667046862694
"133"   16.76554997208264
"134"   18.376689381033508
"135"   14.40541922290389
"136"   12.584346846846849
"137"   14.060476623848091
"138"   37.442771056676094
"139"   23.40952380952381
"14"    15.749354243542435
"140"   14.278674909240369
"141"   12.391378608082539
"142"   13.910121859205608
"143"   13.206644272409084
"144"   16.929347185768396
"145"   12.501812450748616
"146"   15.172756058431132
"147"   16.578253968253968
"148"   16.584931917192876
"149"   10.220402298850576
"15"    11.010185185185186
"150"   16.546376811594204
"151"   13.512899896800828
"152"   15.19777807305047
"153"   12.657235142118862
"154"   33.00833333333334
"155"   22.139869281045748
"156"   4.492857142857142
"157"   23.146812196812
```

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

```
[hadoop@ip-172-31-82-210 ~]$ python mrtask_e.py input_data > mrtask_e_out.txt
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/mrtask_e.hadoop.20231121.075510.807812
Running step 1 of 2...
Running step 2 of 2...
job output is in /tmp/mrtask_e.hadoop.20231121.075510.807812/output
Streaming final output from /tmp/mrtask_e.hadoop.20231121.075510.807812/output...
Removing temp directory /tmp/mrtask_e.hadoop.20231121.075510.807812...
```

```
[hadoop@ip-172-31-82-210 ~]$ cat mrtask_e_out.txt
"86"    0.31385985951687506
"109"   0.2727411419199817
"201"   0.23398594719967258
"187"   0.23076923076923078
"58"    0.20677517842210683
"191"   0.1963891248937978
"172"   0.18060729407061474
"23"    0.17110724973453034
"84"    0.1663871184166387
"175"   0.16555962907051971
"108"   0.16459569486203868
"184"   0.16455893832943014
"176"   0.16006448641179177
"16"    0.15964166650420136
"96"    0.15366033684493052
"199"   0.14926695870154202
"252"   0.14370421057691493
"122"   0.1387652122291481
"71"    0.13638875572600404
"73"    0.1347386827278026
"9"     0.1327844471286185
"138"   0.1317185832146078
"52"    0.12949651566118686
"253"   0.12804966977102736
"240"   0.12758717660292465
"13"    0.12714951929426124
"206"   0.1270032305061126
"87"    0.12639901382300125
"125"   0.12518769144404782
"66"    0.12354358405949303
"194"   0.12353339789282343
"162"   0.12317702599821087
"234"   0.12285200245014952
"33"    0.12194910391730475
"251"   0.12126772106222404
"249"   0.12107964533332406
"180"   0.12101155295619763
"246"   0.12084865909975609
"1"     0.1207141821477579
"40"    0.12068726251898797
"107"   0.12060767877711343
"158"   0.12040536538332194
"231"   0.12007054766692572
"113"   0.11983561384760 97
"170"   0.11967256252276448
"90"    0.1193716516843036
"88"    0.11923831790881051
"195"   0.11921275351973504
"114"   0.11911520367510238
"161"   0.11910074531590943
"255"   0.11892329702334607
```

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-172-31-82-210 ~]$ cat mrtask_f_out.txt
[3, 0, 0]       20.40521359874317
[3, 0, 1]       17.276197995863022
[3, 0, 2]       17.601844621054536
[3, 0, 3]       17.25439005270323
[3, 0, 4]       17.319914947221616
[3, 0, 5]       16.53061715387093
[3, 0, 6]       15.555541933661654
[3, 1, 0]       19.68703647305948
[3, 1, 1]       16.41271198977643
[3, 1, 2]       16.925512278579316
[3, 1, 3]       16.61165540540129
[3, 1, 4]       16.160107419186303
[3, 1, 5]       15.832497441155462
[3, 1, 6]       15.364370270396241
[3, 10, 0]      15.792245838246089
[3, 10, 1]      15.261493329051431
[3, 10, 2]      15.93810925783485
[3, 10, 3]      16.30214075876192
[3, 10, 4]      16.080729421785076
[3, 10, 5]      13.143838603285518
[3, 10, 6]      13.905262614980442
[3, 11, 0]      16.17422854439143
[3, 11, 1]      15.433695314990839
[3, 11, 2]      16.34197199387861
[3, 11, 3]      16.615542982468714
[3, 11, 4]      16.38474991537654
[3, 11, 5]      13.390678702178493
[3, 11, 6]      13.940995981214432
[3, 12, 0]      16.282810720389826
[3, 12, 1]      15.430570743504797
[3, 12, 2]      16.544916132322516
[3, 12, 3]      17.02373108252729
[3, 12, 4]      16.61745941219783
[3, 12, 5]      13.700263177960306
[3, 12, 6]      14.484117559352311
[3, 13, 0]      16.38637314883534
[3, 13, 1]      15.864056889450703
[3, 13, 2]      17.298944478583774
[3, 13, 3]      17.851531498394657
[3, 13, 4]      17.00853579230275
[3, 13, 5]      14.254259457054172
[3, 13, 6]      15.15122902764985
[3, 14, 0]      16.62469914309496
[3, 14, 1]      16.017420900045288
[3, 14, 2]      17.232419968337283
[3, 14, 3]      18.193252459930395
[3, 14, 4]      17.451055929438706
[3, 14, 5]      14.67993677474162
[3, 14, 6]      16.21387120346392
[3, 15, 0]      16.592225876866596
[3, 15, 1]      15.852790671648188
[3, 15, 2]      17.06561720209403
[3, 15, 3]      18.295222864452015
[3, 15, 4]      17.562402247232335
[3, 15, 5]      14.855918342093966
[3, 15, 6]      16.797449883663646
[3, 16, 0]      18.042703708594157
[3, 16, 1]      17.33124934085766
[3, 16, 2]      18.864639905252556
[3, 16, 3]      18.768173255703436
```