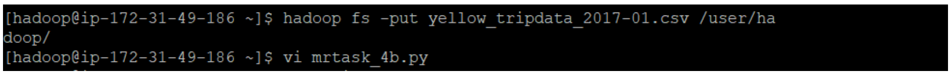
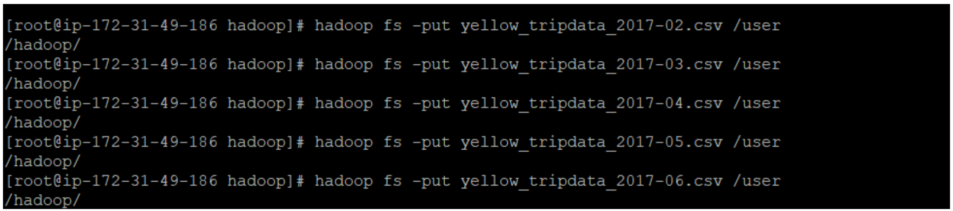
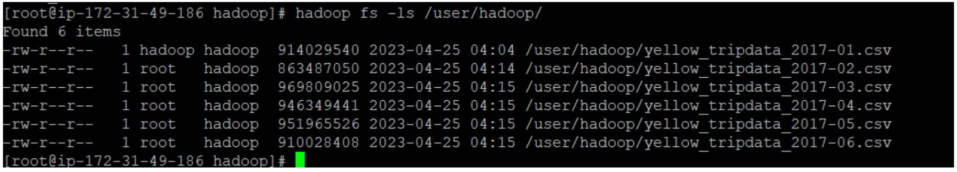
1. Adding input files to hdfs



```
[hadoop@ip-172-31-49-186 ~]$ hadoop fs -put yellow_tripdata_2017-01.csv /user/ha
doop/
[hadoop@ip-172-31-49-186 ~]$ vi mrtask_4b.py
```

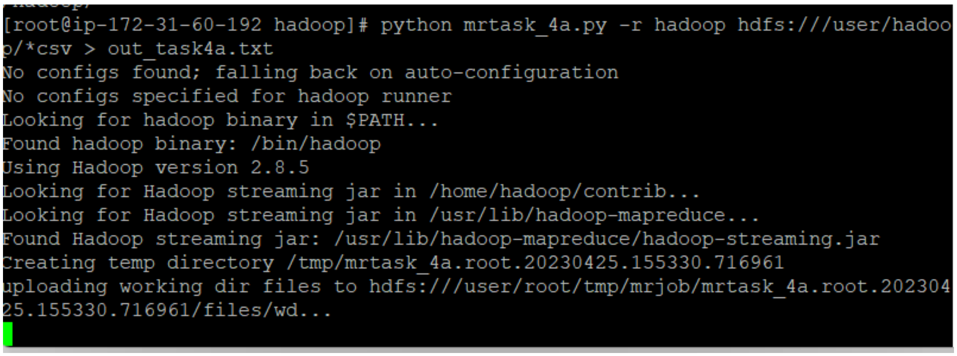


```
[root@ip-172-31-49-186 hadoop]# hadoop fs -put yellow_tripdata_2017-02.csv /user
/hadoop/
[root@ip-172-31-49-186 hadoop]# hadoop fs -put yellow_tripdata_2017-03.csv /user
/hadoop/
[root@ip-172-31-49-186 hadoop]# hadoop fs -put yellow_tripdata_2017-04.csv /user
/hadoop/
[root@ip-172-31-49-186 hadoop]# hadoop fs -put yellow_tripdata_2017-05.csv /user
/hadoop/
[root@ip-172-31-49-186 hadoop]# hadoop fs -put yellow_tripdata_2017-06.csv /user
/hadoop/
```

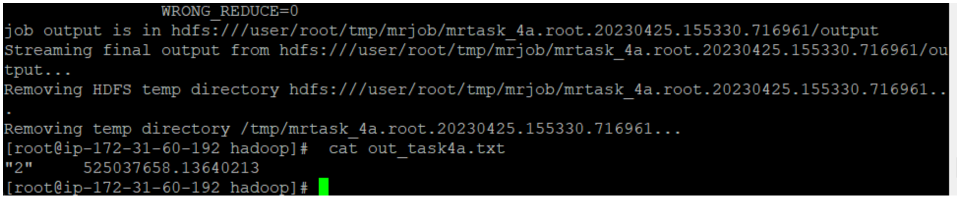


```
[root@ip-172-31-49-186 hadoop]# hadoop fs -ls /user/hadoop/
Found 6 items
-rw-r--r--   1 hadoop hadoop  914029540 2023-04-25 04:04 /user/hadoop/yellow_tripdata_2017-01.csv
-rw-r--r--   1 root   hadoop  863487050 2023-04-25 04:14 /user/hadoop/yellow_tripdata_2017-02.csv
-rw-r--r--   1 root   hadoop  969809025 2023-04-25 04:15 /user/hadoop/yellow_tripdata_2017-03.csv
-rw-r--r--   1 root   hadoop  946349441 2023-04-25 04:15 /user/hadoop/yellow_tripdata_2017-04.csv
-rw-r--r--   1 root   hadoop  951965526 2023-04-25 04:15 /user/hadoop/yellow_tripdata_2017-05.csv
-rw-r--r--   1 root   hadoop  910028408 2023-04-25 04:15 /user/hadoop/yellow_tripdata_2017-06.csv
[root@ip-172-31-49-186 hadoop]#
```

2. Task 4a



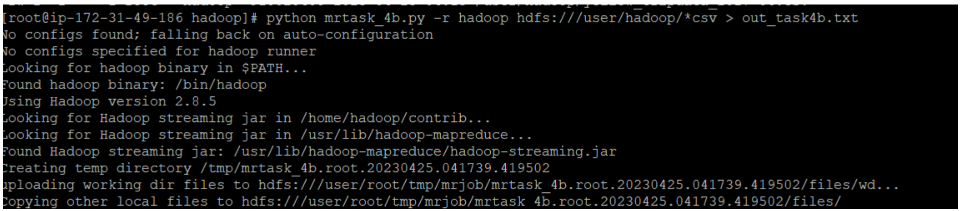
```
[root@ip-172-31-60-192 hadoop]# python mrtask_4a.py -r hadoop hdfs:///user/hadoo
p/*csv > out_task4a.txt
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /bin/hadoop
Using Hadoop version 2.8.5
Looking 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_4a.root.20230425.155330.716961
uploading working dir files to hdfs:///user/root/tmp/mrjob/mrtask_4a.root.202304
25.155330.716961/files/wd...
```



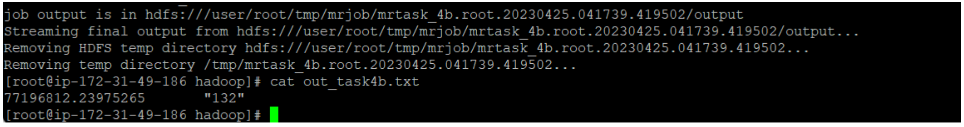
```
                WRONG_REDUCE=0
job output is in hdfs:///user/root/tmp/mrjob/mrtask_4a.root.20230425.155330.716961/output
Streaming final output from hdfs:///user/root/tmp/mrjob/mrtask_4a.root.20230425.155330.716961/ou
tput...
Removing HDFS temp directory hdfs:///user/root/tmp/mrjob/mrtask_4a.root.20230425.155330.716961..
.
Removing temp directory /tmp/mrtask_4a.root.20230425.155330.716961...
[root@ip-172-31-60-192 hadoop]#  cat out_task4a.txt
"2"     525037658.13640213
[root@ip-172-31-60-192 hadoop]#
```

Result: VendorId 2 has maximum trips with total revenue of 525037658.13640213.

3. Task 4b



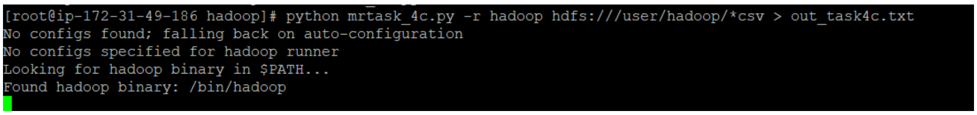
```
[root@ip-172-31-49-186 hadoop]# python mrtask_4b.py -r hadoop hdfs:///user/hadoop/*csv > out_task4b.txt
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /bin/hadoop
Using Hadoop version 2.8.5
Looking 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_4b.root.20230425.041739.419502
uploading working dir files to hdfs:///user/root/tmp/mrjob/mrtask_4b.root.20230425.041739.419502/files/wd...
Copying other local files to hdfs:///user/root/tmp/mrjob/mrtask_4b.root.20230425.041739.419502/files/
```



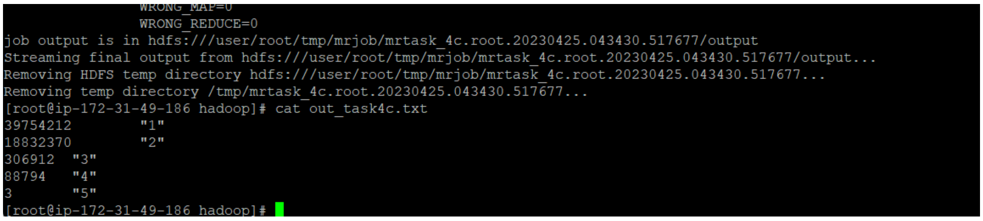
```
job output is in hdfs:///user/root/tmp/mrjob/mrtask_4b.root.20230425.041739.419502/output
Streaming final output from hdfs:///user/root/tmp/mrjob/mrtask_4b.root.20230425.041739.419502/output...
Removing HDFS temp directory hdfs:///user/root/tmp/mrjob/mrtask_4b.root.20230425.041739.419502...
Removing temp directory /tmp/mrtask_4b.root.20230425.041739.419502...
[root@ip-172-31-49-186 hadoop]# cat out_task4b.txt
77196812.23975265       "132"
[root@ip-172-31-49-186 hadoop]#
```

Result: Pickup location Id 132 has the highest total revenue of 77196812.23975265

4. Task 4c



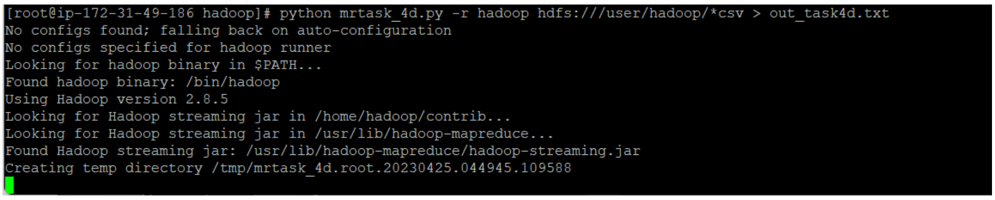
```
[root@ip-172-31-49-186 hadoop]# python mrtask_4c.py -r hadoop hdfs:///user/hadoop/*csv > out_task4c.txt
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /bin/hadoop
```



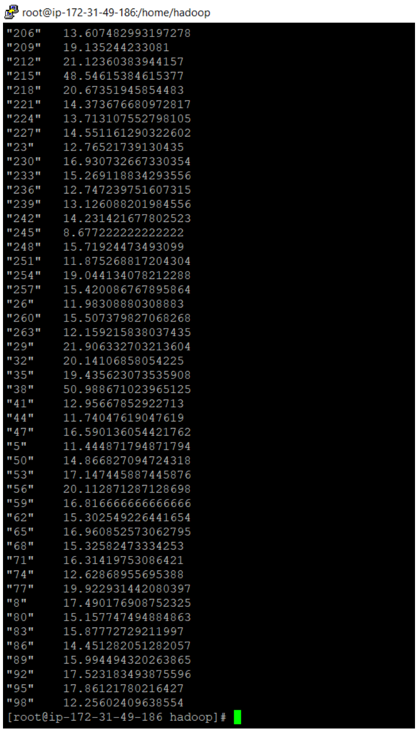
```
                WRONG_MAP=0
                WRONG_REDUCE=0
job output is in hdfs:///user/root/tmp/mrjob/mrtask_4c.root.20230425.043430.517677/output
Streaming final output from hdfs:///user/root/tmp/mrjob/mrtask_4c.root.20230425.043430.517677/output...
Removing HDFS temp directory hdfs:///user/root/tmp/mrjob/mrtask_4c.root.20230425.043430.517677...
Removing temp directory /tmp/mrtask_4c.root.20230425.043430.517677...
[root@ip-172-31-49-186 hadoop]# cat out_task4c.txt
39754212        "1"
18832370        "2"
306912  "3"
88794   "4"
3       "5"
[root@ip-172-31-49-186 hadoop]#
```

<u>Result:</u> Payment Type 1 (Credit Card) has the highest number of trips followed by 2 (Cash) and 3 (No Charge).

5. Task 4d



```
[root@ip-172-31-49-186 hadoop]# python mrtask_4d.py -r hadoop hdfs:///user/hadoop/*csv > out_task4d.txt
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /bin/hadoop
Using Hadoop version 2.8.5
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_4d.root.20230425.044945.109588
```



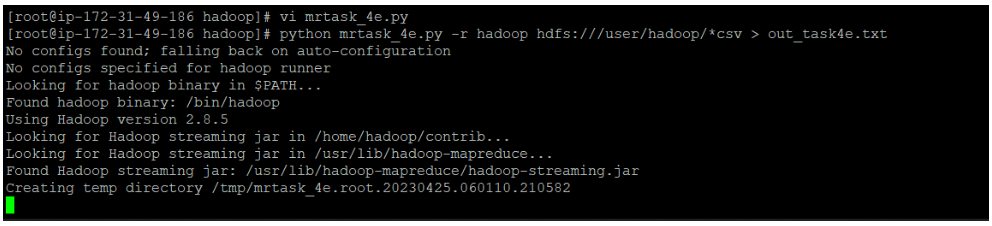
root@ip-172-31-49-186:/home/hadoop

```
"206"   13.607482993197278
"209"   19.135244233081
"212"   21.12360383944157
"215"   48.54615384615377
"218"   20.67351945854483
"221"   14.373676680972817
"224"   13.713107552798105
"227"   14.551161290322602
"23"    12.76521739130435
"230"   16.930732667330354
"233"   15.269118834293556
"236"   12.747239751607315
"239"   13.126088201984556
"242"   14.231421677802523
"245"   8.677222222222222
"248"   15.71924473493099
"251"   11.875268817204304
"254"   19.044134078212288
"257"   15.420086767895864
"26"    11.98308880308883
"260"   15.507379827068268
"263"   12.159215838037435
"29"    21.906332703213604
"32"    20.14106858054225
"35"    19.435623073535908
"38"    50.988671023965125
"41"    12.95667852922713
"44"    11.74047619047619
"47"    16.590136054421762
"5"     11.444871794871794
"50"    14.866827094724318
"53"    17.147445887445876
"56"    20.112871287128698
"59"    16.816666666666666
"62"    15.302549226441654
"65"    16.960852573062795
"68"    15.32582473334253
"71"    16.31419753086421
"74"    12.62868955695388
"77"    19.922931442080397
"8"     17.490176908752325
"80"    15.157747494884863
"83"    15.87772729211997
"86"    14.451282051282057
"89"    15.994494320263865
"92"    17.523183493875596
"95"    17.86121780216427
"98"    12.25602409638554
[root@ip-172-31-49-186 hadoop]#
```

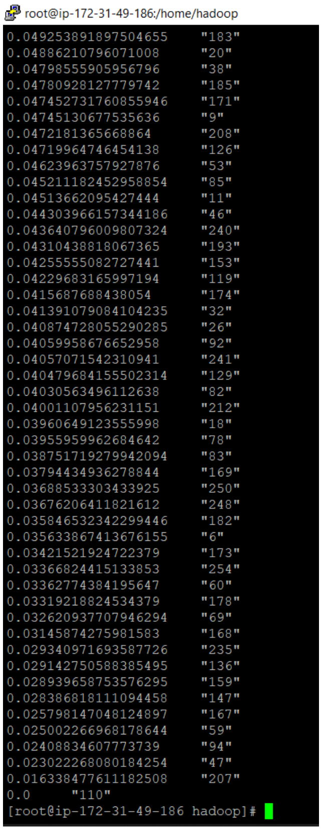
out_task4d.txt

<u>Result:</u> The average trip time is calculated in minutes for each pickup location.

6. Task 4e



```
[root@ip-172-31-49-186 hadoop]# vi mrtask_4e.py
[root@ip-172-31-49-186 hadoop]# python mrtask_4e.py -r hadoop hdfs:///user/hadoop/*csv > out_task4e.txt
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /bin/hadoop
Using Hadoop version 2.8.5
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_4e.root.20230425.060110.210582
```

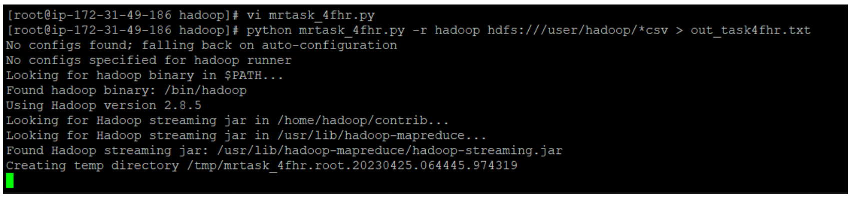


root@ip-172-31-49-186:/home/hadoop

```
0.049253891897504655    "183"
0.04886210796071008    "20"
0.04798555905956796    "38"
0.04780928127779742    "185"
0.047452731760855946    "171"
0.04745130677535636    "9"
0.0472181365668864    "208"
0.04719964746454138    "126"
0.04623963757927876    "53"
0.045211182452958854    "85"
0.04513662095427444    "11"
0.044303966157344186    "46"
0.043640796009807324    "240"
0.04310438818067365    "193"
0.04255555082727441    "153"
0.04229683165997194    "119"
0.0415687688438054    "174"
0.041391079084104235    "32"
0.040874728055290285    "26"
0.04059958676652958    "92"
0.04057071542310941    "241"
0.040479684155502314    "129"
0.04030563496112638    "82"
0.04001107956231151    "212"
0.03960649123555998    "18"
0.03955959962684642    "78"
0.038751719279942094    "83"
0.03794434936278844    "169"
0.03688533303433925    "250"
0.03676206411821612    "248"
0.035846532342299446    "182"
0.035633867413676155    "6"
0.03421521924722379    "173"
0.03366824415133853    "254"
0.03362774384195647    "60"
0.03319218824534379    "178"
0.032620937707946294    "69"
0.03145874275981583    "168"
0.029340971693587726    "235"
0.029142750588385495    "136"
0.028939658753576295    "159"
0.028386818111094458    "147"
0.025798147048124897    "167"
0.025002266968178644    "59"
0.02408834607773739    "94"
0.023022268080184254    "47"
0.016338477611182508    "207"
0.0     "110"
[root@ip-172-31-49-186 hadoop]#
```

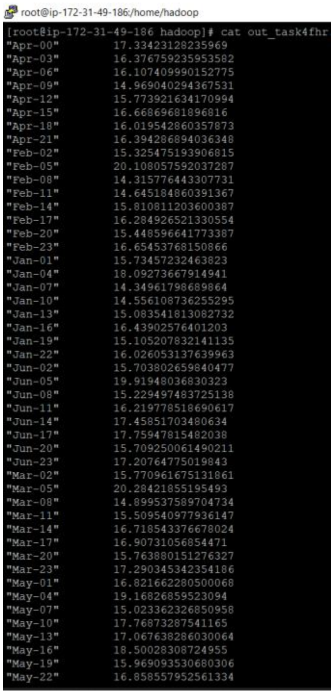
out_task4e.txt

7. Task 4f – Revenue per month analyzing hourly



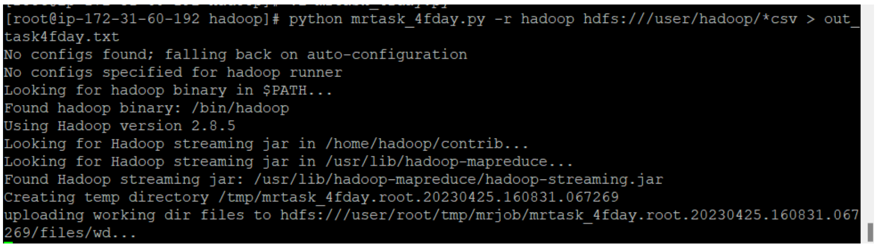
```
[root@ip-172-31-49-186 hadoop]# vi mrtask_4fhr.py
[root@ip-172-31-49-186 hadoop]# python mrtask_4fhr.py -r hadoop hdfs:///user/hadoop/*csv > out_task4fhr.txt
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /bin/hadoop
Using Hadoop version 2.8.5
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_4fhr.root.20230425.064445.974319
```

out_task4fhr.txt

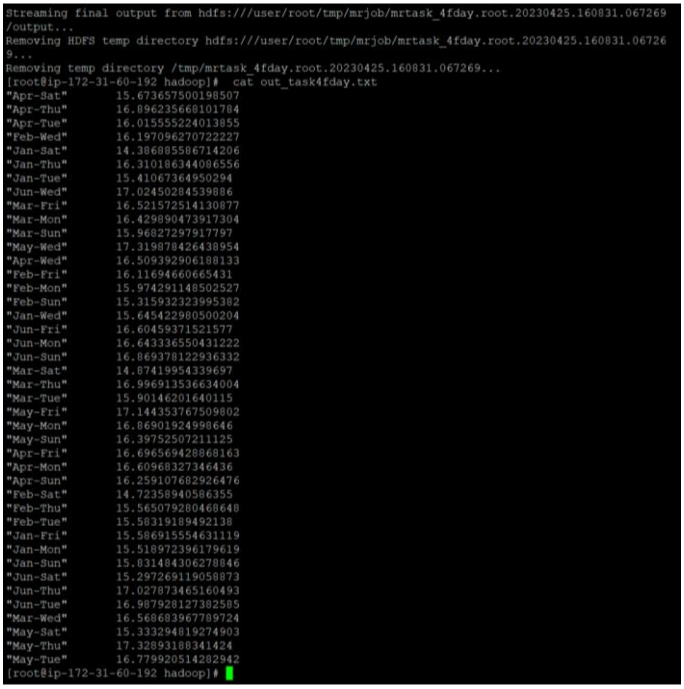


```
root@ip-172-31-49-186:/home/hadoop
[root@ip-172-31-49-186 hadoop]# cat out_task4fhr
"Apr-00"        17.33423128235969
"Apr-03"        16.376759235953582
"Apr-06"        16.107409990152775
"Apr-09"        14.969040294367531
"Apr-12"        15.773921634170994
"Apr-15"        16.66869681896816
"Apr-18"        16.019542860357873
"Apr-21"        16.394286894036348
"Feb-02"        15.325475193906815
"Feb-05"        20.108057592037287
"Feb-08"        14.315776443307731
"Feb-11"        14.645184860391367
"Feb-14"        15.810811203600387
"Feb-17"        16.284926521330554
"Feb-20"        15.448596641773387
"Feb-23"        16.65453768150866
"Jan-01"        15.73457232463823
"Jan-04"        18.09273667914941
"Jan-07"        14.34961798689864
"Jan-10"        14.556108736255295
"Jan-13"        15.083541813082732
"Jan-16"        16.43902576401203
"Jan-19"        15.105207832141135
"Jan-22"        16.026053137639963
"Jun-02"        15.703802659840477
"Jun-05"        19.91948036830323
"Jun-08"        15.229497483725138
"Jun-11"        16.219778518690617
"Jun-14"        17.45851703480634
"Jun-17"        17.75947815482038
"Jun-20"        15.709250061490211
"Jun-23"        17.20764775019843
"Mar-02"        15.770961675131861
"Mar-05"        20.28421855195493
"Mar-08"        14.899537589704734
"Mar-11"        15.509540977936147
"Mar-14"        16.718543376678024
"Mar-17"        16.90731056854471
"Mar-20"        15.763880151276327
"Mar-23"        17.290345342354186
"May-01"        16.821662280500068
"May-04"        19.16826859523094
"May-07"        15.023362326850958
"May-10"        17.76873287541165
"May-13"        17.067638286030064
"May-16"        18.50028308724955
"May-19"        15.969093530680306
"May-22"        16.858557952561334
```

8. Task 4f Revenue per month analyzing Daywise



```
[root@ip-172-31-60-192 hadoop]# python mrtask_4fday.py -r hadoop hdfs:///user/hadoop/*csv > out_
task4fday.txt
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /bin/hadoop
Using Hadoop version 2.8.5
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_4fday.root.20230425.160831.067269
uploading working dir files to hdfs:///user/root/tmp/mrjob/mrtask_4fday.root.20230425.160831.067
269/files/wd...
```



```
Streaming final output from hdfs:///user/root/tmp/mrjob/mrtask_4fday.root.20230425.160831.067269
/output...
Removing HDFS temp directory hdfs:///user/root/tmp/mrjob/mrtask_4fday.root.20230425.160831.06726
9...
Removing temp directory /tmp/mrtask_4fday.root.20230425.160831.067269...
[root@ip-172-31-60-192 hadoop]#  cat out_task4fday.txt
"Apr-Sat"       15.673657500198507
"Apr-Thu"       16.896235668101784
"Apr-Tue"       16.015555224013855
"Feb-Wed"       16.197096270722227
"Jan-Sat"       14.386885586714206
"Jan-Thu"       16.310186344086556
"Jan-Tue"       15.41067364950294
"Jun-Wed"       17.02450284539886
"Mar-Fri"       16.521572514130877
"Mar-Mon"       16.429890473917304
"Mar-Sun"       15.96827297917797
"May-Wed"       17.319878426438954
"Apr-Wed"       16.509392906188133
"Feb-Fri"       16.11694660665431
"Feb-Mon"       15.974291148502527
"Feb-Sun"       15.315932323995382
"Jan-Wed"       15.645422980500204
"Jun-Fri"       16.60459371521577
"Jun-Mon"       16.643336550431222
"Jun-Sun"       16.869378122936332
"Mar-Sat"       14.87419954339697
"Mar-Thu"       16.996913536634004
"Mar-Tue"       15.90146201640115
"May-Fri"       17.144353767509802
"May-Mon"       16.86901924998646
"May-Sun"       16.39752507211125
"Apr-Fri"       16.696569428868163
"Apr-Mon"       16.60968327346436
"Apr-Sun"       16.259107682926476
"Feb-Sat"       14.72358940586355
"Feb-Thu"       15.565079280468648
"Feb-Tue"       15.58319189492138
"Jan-Fri"       15.586915554631119
"Jan-Mon"       15.518972396179619
"Jan-Sun"       15.831484306278846
"Jun-Sat"       15.297269119058873
"Jun-Thu"       17.027873465160493
"Jun-Tue"       16.987928127382585
"Mar-Wed"       16.568683967789724
"May-Sat"       15.333294819274903
"May-Thu"       17.32893188341424
"May-Tue"       16.779920514282942
[root@ip-172-31-60-192 hadoop]#
```