

## MapReduce Tasks:

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

```
[hadoop@ip-172-31-19-74 ~]$ vi task4_a.py
[hadoop@ip-172-31-19-74 ~]$ python task4_a.py yellow_tripdata_2017-04.csv > task4_a_result.txt
Traceback (most recent call last):
  File "task4_a.py", line 3, in <module>
    from mrjob.job import MRJob
ModuleNotFoundError: No module named 'mrjob'
[hadoop@ip-172-31-19-74 ~]$ 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 34.6 MB/s
Requirement already satisfied: PyYAML>=3.10 in /usr/local/lib64/python3.7/site-packages (from mrjob) (5.4.1)
Installing collected packages: mrjob
  WARNING: The scripts mrjob, mrjob-3 and mrjob-3.7 are installed in '/home/hadoop/.local/bin' which is not on PATH.
  Consider adding this directory to PATH or, if you prefer to suppress this warning, use --no-warn-script-location.
Successfully installed mrjob-0.7.4
[hadoop@ip-172-31-19-74 ~]$
```

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

**Answer:**

**Data file:** Using the file '**yellow\_tripdata\_2017-04.csv**' for this exercise.

**Code snippet:**

```
[hadoop@ip-172-31-19-74 ~]$ cat task4_a.py
# Which vendors have the most trips, and what is the total revenue generated by that vendor?

from mrjob.job import MRJob
from mrjob.step import MRStep

class MostTripsTotalRevenue(MRJob):

    def steps(self):
        return [
            MRStep(mapper=self.mapper, reducer=self.reducer),
            MRStep(reducer=self.final_reducer)
        ]

    def mapper(self, _, line):
        if not line.startswith('VendorID'):
            data = line.split(',')
            vendor_id = data[0]
            revenue = float(data[16])
            yield vendor_id, revenue

    def reducer(self, key, values):
        yield None, (sum(values), key)

    def final_reducer(self, _, values):
        max_revenue, vendor_id = max(values)
        yield vendor_id, max_revenue

if __name__ == '__main__':
    MostTripsTotalRevenue.run()
[hadoop@ip-172-31-19-74 ~]$
```

## Result:

```
[hadoop@ip-172-31-19-74 ~]$ python task4_a.py yellow_tripdata_2017-04.csv > task4_a_result.txt
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/task4_a.hadoop.20231106.080240.083596
Running step 1 of 2...
Running step 2 of 2...
job output is in /tmp/task4_a.hadoop.20231106.080240.083596/output
Streaming final output from /tmp/task4_a.hadoop.20231106.080240.083596/output...
Removing temp directory /tmp/task4_a.hadoop.20231106.080240.083596...
[hadoop@ip-172-31-19-74 ~]$ ls
task2.py      task3.py      task4_a_result.txt  yellow_tripdata_2017-02.csv  yellow_tripdata_2017-04.csv
task3_hbase.py  task4_a.py  yellow_tripdata_2017-01.csv  yellow_tripdata_2017-03.csv
[hadoop@ip-172-31-19-74 ~]$ cat task4_a_result.txt
"2"      89987931.82539217
[hadoop@ip-172-31-19-74 ~]$
```

Based on the map reduce job result, in April 2017 Vendor Id = 2 "VeriFone Inc." had the most trips and generated the total revenue **89987931.83** (sum of Tolls\_amount 'Total amount of all tolls paid in trip.' field value).

b) Which pickup location generates the most revenue?

**Answer:**

**Data file:** Using the file 'yellow\_tripdata\_2017-04.csv' for this exercise.

**Code snippet:**

```
[hadoop@ip-172-31-19-74 ~]$ cat task4_b.py
# Which pickup location generates the most revenue?

from mrjob.job import MRJob
from mrjob.step import MRStep

class MostRevenuePickupLocation(MRJob):

    def steps(self):
        return [
            MRStep(mapper=self.mapper, reducer=self.reducer),
            MRStep(reducer=self.final_reducer)
        ]

    def mapper(self, _, line):
        # Skip the header line
        if not line.startswith('VendorID'):
            fields = line.split(',')
            pickup_location = fields[7]
            revenue = float(fields[16])
            yield pickup_location, revenue

    def reducer(self, pickup_location, revenues):
        yield None, (sum(revenues), pickup_location)

    def final_reducer(self, _, max_revenues):
        max_revenue, pickup_location = max(max_revenues)
        yield pickup_location, max_revenue

if __name__ == '__main__':
    MostRevenuePickupLocation.run()
[hadoop@ip-172-31-19-74 ~]$
```

**Answer:**

```

[hadoop@ip-172-31-19-74 ~]$ python task4_b.py yellow_tripdata_2017-04.csv > task4_b_result.txt
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/task4_b.hadoop.20231106.082115.547716
Running step 1 of 2...
Running step 2 of 2...
job output is in /tmp/task4_b.hadoop.20231106.082115.547716/output
Streaming final output from /tmp/task4_b.hadoop.20231106.082115.547716/output...
Removing temp directory /tmp/task4_b.hadoop.20231106.082115.547716...
[hadoop@ip-172-31-19-74 ~]$ ls -lart
total 3607168
-rw-rw-r-- 1 hadoop hadoop 914029540 Nov 25 2022 yellow_tripdata_2017-01.csv
-rw-rw-r-- 1 hadoop hadoop 863487050 Nov 25 2022 yellow_tripdata_2017-02.csv
-rw-rw-r-- 1 hadoop hadoop 969809025 Nov 25 2022 yellow_tripdata_2017-03.csv
-rw-rw-r-- 1 hadoop hadoop 946349441 Nov 25 2022 yellow_tripdata_2017-04.csv
-rw-r--r-- 1 hadoop hadoop 626 Feb 2 2023 .bashrc
-rw-r--r-- 1 hadoop hadoop 86 Feb 2 2023 .bash_profile
drwxr-xr-x 4 root root 36 Aug 15 01:51 ..
drwxr-xr-x 2 hadoop hadoop 20 Aug 15 01:53 .aws
drwx----- 2 hadoop hadoop 29 Nov 6 05:21 .ssh
-rw----- 1 hadoop hadoop 99 Nov 6 05:49 .mysql_history
-rwxr-xr-x 1 hadoop hadoop 1738 Nov 6 05:57 task3.py
-rwxr-xr-x 1 hadoop hadoop 1739 Nov 6 06:55 task2.py
-rw----- 1 hadoop hadoop 249 Nov 6 07:36 .bash_history
-rw-rw-r-- 1 hadoop hadoop 181 Nov 6 07:53 task3_hbase.py
-rw-rw-r-- 1 hadoop hadoop 837 Nov 6 07:57 task4_a.py
drwx----- 5 hadoop root 47 Nov 6 07:59 .cache
drwx----- 4 hadoop hadoop 28 Nov 6 07:59 .local
-rw-rw-r-- 1 hadoop hadoop 22 Nov 6 08:05 task4_a_result.txt
-rw----- 1 hadoop hadoop 7514 Nov 6 08:20 .viminfo
-rw-rw-r-- 1 hadoop hadoop 894 Nov 6 08:20 task4_b.py
drwxr-xr-x 6 hadoop hadoop 4096 Nov 6 08:21 .
-rw-rw-r-- 1 hadoop hadoop 25 Nov 6 08:24 task4_b_result.txt
[hadoop@ip-172-31-19-74 ~]$ cat task4_b_result.txt
"132" 13693066.230013764
[hadoop@ip-172-31-19-74 ~]$

```

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

**Data file:** Using the file 'yellow\_tripdata\_2017-04.csv' for this exercise.

**Code snippet:**

```
[hadoop@ip-172-31-19-74 ~]$ cat task4_c.py
# What are the different payment types used by customers and their count? The final results should be in a sorted format.

from mrjob.job import MRJob
from mrjob.step import MRStep

class PaymentTypesCount(MRJob):

    def mapper(self, _, line):
        # Skip the header line
        if not line.startswith('VendorID'):
            fields = line.split(',')
            payment_type = fields[9]
            yield payment_type, 1

    def combiner(self, payment_type, counts):
        yield payment_type, sum(counts)

    def reducer(self, payment_type, counts):
        yield payment_type, sum(counts)

    def reducer_sort_results(self, payment_type, counts):
        yield None, (sum(counts), payment_type)

    def reducer_output_result(self, _, sorted_results):
        for count, payment_type in sorted(sorted_results, reverse=True):
            yield payment_type, count

    def steps(self):
        return [
            MRStep(mapper=self.mapper, combiner=self.combiner, reducer=self.reducer),
            MRStep(reducer=self.reducer_sort_results),
            MRStep(reducer=self.reducer_output_result)
        ]

if __name__ == '__main__':
    PaymentTypesCount.run()
```

## Answer:

```
[hadoop@ip-172-31-19-74 ~]$ python task4_c.py yellow_tripdata_2017-04.csv > task4_c_result.txt
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/task4_c.hadoop.20231106.083805.049910
Running step 1 of 3...
Running step 2 of 3...
Running step 3 of 3...
job output is in /tmp/task4_c.hadoop.20231106.083805.049910/output
Streaming final output from /tmp/task4_c.hadoop.20231106.083805.049910/output...
Removing temp directory /tmp/task4_c.hadoop.20231106.083805.049910...
[hadoop@ip-172-31-19-74 ~]$ cat task4_c_result.txt
"1"      6695495
"2"      3281576
"3"       54383
"4"      15680
"5"         1
[hadoop@ip-172-31-19-74 ~]$
```

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

**Data file:** Using the file 'yellow\_tripdata\_2017-04.csv' for this exercise.

**Code snippet:**

hadoop@ip-172-31-19-74:~

```
[hadoop@ip-172-31-19-74 ~]$ cat task4_d.py
# What is the average trip time for different pickup locations?

from mrjob.job import MRJob
from datetime import datetime

class AverageTripTime(MRJob):

    def parse_datetime(self, datetime_str):
        formats = ['%d-%m-%Y %H:%M:%S', '%d-%m-%Y %H:%M', '%Y-%m-%d %H:%M', '%Y-%m-%d %H:%M:%S']
        for fmt in formats:
            try:
                return datetime.strptime(datetime_str, fmt)
            except ValueError:
                pass
        raise ValueError('no valid date format found')

    def mapper(self, _, line):
        # Skip the header line
        if not line.startswith('VendorID'):
            fields = line.split(',')
            pickup_location = fields[7]
            pickup_datetime = self.parse_datetime(fields[1])
            dropoff_datetime = self.parse_datetime(fields[2])
            trip_time = (dropoff_datetime - pickup_datetime).total_seconds() / 60.0
            yield pickup_location, (trip_time, 1)

    def combiner(self, pickup_location, trip_times):
        total_trip_time = 0
        total_count = 0
        for trip_time, count in trip_times:
            total_trip_time += trip_time
            total_count += count
        yield pickup_location, (total_trip_time, total_count)

    def reducer(self, pickup_location, trip_times):
        total_trip_time = 0
        total_count = 0
        for trip_time, count in trip_times:
            total_trip_time += trip_time
            total_count += count
        average_trip_time = total_trip_time / total_count
        yield pickup_location, average_trip_time
```

```
if __name__ == '__main__':
    AverageTripTime.run()
[hadoop@ip-172-31-19-74 ~]$
```

**Answer:**

hadoop@ip-172-31-19-74:~

```
[hadoop@ip-172-31-19-74 ~]$ python task4_d.py yellow_tripdata_2017-04.csv > task4_d_result.txt
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/task4_d.hadoop.20231106.084648.858319
Running step 1 of 1...
job output is in /tmp/task4_d.hadoop.20231106.084648.858319/output
Streaming final output from /tmp/task4_d.hadoop.20231106.084648.858319/output...
Removing temp directory /tmp/task4_d.hadoop.20231106.084648.858319...
[hadoop@ip-172-31-19-74 ~]$ cat task4_d_result.txt
"1"      8.69859496124031
"10"     46.56947758496023
"100"    15.86489867520792
"101"    57.85808080808081
"102"    27.02078853046595
"105"    12.205555555555556
"106"    14.100608029318675
"107"    14.39224446410785
"108"    17.885714285714286
"109"    11.458333333333334
"11"     34.16747311827957
"111"    8.989655172413793
"112"    13.645934065934068
"113"    15.704456568326293
"114"    16.46732887097284
"115"    14.095833333333333
"116"    15.483265311788555
"117"    15.862500000000002
"118"    13.819078947368425
"119"    13.650181159420288
"12"     24.355169683257913
"120"    14.37363184079602
"121"    12.520833333333334
"122"    17.093333333333334
"123"    10.440864197530864
"124"    25.828911564625848
"125"    16.83851913974292
"126"    15.35560606060606
"127"    15.639279437609842
"128"    13.14125683060109
"129"    15.438598146877146
"13"     19.94718083163505
"130"    36.34969437652811
"131"    22.668794326241137
"132"    44.921953837895835
```


hadoop@ip-172-31-19-74:~

```
"133" 21.909053030303028
"134" 15.840156599552573
"135" 17.89295774647887
"136" 10.338076923076924
"137" 13.8018368098946
"138" 36.88323042229314
"139" 157.42666666666668
"14" 20.473815676141257
"140" 13.672047369652681
"141" 12.363297893477288
"142" 13.870544062880667
"143" 13.358197694781422
"144" 17.50160171645993
"145" 12.264756711188392
"146" 15.16674741451597
"147" 10.31024096385542
"148" 17.078479882548983
"149" 15.395289855072464
"15" 20.501282051282047
"150" 25.636231884057974
"151" 13.162877263581485
"152" 13.215261915998116
"153" 19.398165137614676
"154" 25.57708333333333
"155" 15.302121212121213
"156" 19.22795698924731
"157" 19.558055555555555
"158" 17.37444961080658
"159" 12.52086956521739
"16" 48.17041666666667
"160" 16.990000000000002
"161" 15.938705808597048
"162" 15.440213051344449
"163" 16.33871373768244
"164" 15.68439179882676
"165" 13.980839002267574
"166" 14.601165429969408
"167" 16.15883620689655
"168" 17.07202848722986
"169" 19.148782343987822
"17" 15.841983191442914
"170" 14.771046389063152
"171" 14.547135416666668
"172" 20.51904761904762
```



 hadoop@ip-172-31-19-74:~

```
"172" 20.51904761904762
"173" 12.95126646403242
"174" 10.996551724137928
"175" 15.852380952380953
"176" 13.283333333333333
"177" 15.651353276353277
"178" 5.762028985507246
"179" 14.868812835338854
"18" 12.461633109619687
"180" 22.177864583333333
"181" 16.24970470018867
"182" 17.38532608695652
"183" 11.783333333333333
"184" 11.933333333333332
"185" 12.080845771144277
"186" 17.185830463292675
"187" 100.8125
"188" 16.357558859975214
"189" 15.417621450479867
"19" 11.64452736318408
"190" 17.003544061302684
"191" 10.666468253968253
"192" 30.41401515151515
"193" 13.281695316392629
"194" 22.627196752626553
"195" 22.6798213185459
"196" 15.241391457507707
"197" 13.34065460809647
"198" 12.963727959697733
"199" 8.716666666666667
"2" 43.466666666666666
"20" 13.596551724137932
"200" 14.57434210526316
"201" 13.017499999999999
"202" 18.053837719298244
"203" 18.139062499999998
"204" 1.1761904761904762
"205" 13.813475177304964
"206" 22.22
"207" 18.82079646017699
"208" 33.94600938967137
"209" 19.252090867417113
"21" 14.009615384615385
"210" 15.012318840579711
```

 hadoop@ip-172-31-19-74:~

```
"210" 15.012318840579711
"211" 17.432155984310008
"212" 26.1641975308642
"213" 12.697751322751321
"214" 10.016666666666667
"215" 54.3792299898683
"216" 29.7152841781874
"217" 14.53024553571429
"218" 16.003431372549024
"219" 45.942964352720445
"22" 18.867991169977923
"220" 12.347965571205009
"221" 17.824561403508774
"222" 22.13571428571429
"223" 15.366199500982107
"224" 13.95519143194468
"225" 15.451928480204344
"226" 16.852667132541338
"227" 14.087719298245615
"228" 16.69605959342801
"229" 13.303534208453824
"23" 8.895238095238097
"230" 17.327439098960472
"231" 17.271926437910555
"232" 17.511131392863025
"233" 15.092194117448754
"234" 15.399663529680868
"235" 11.131089743589744
"236" 12.81389881120063
"237" 12.475322071288154
"238" 13.083897433411947
"239" 13.42993881027248
"24" 13.499971240978894
"240" 14.314516129032258
"241" 13.014313725490199
"242" 10.567372881355933
"243" 17.29957368082368
"244" 16.058059332111974
"245" 2.6055555555555556
"246" 16.08435416319503
"247" 19.869504436631512
"248" 22.643771043771043
"249" 15.573829224607753
"25" 15.873454838509696
```

hadoop@ip-172-31-19-74:~

```
"25" 15.873454838509696
"250" 15.795327102803737
"251" 13.007407407407406
"252" 20.681560283687944
"253" 24.405797101449277
"254" 14.823924731182798
"255" 16.125786916655258
"256" 16.062604998578472
"257" 12.533091202582728
"258" 36.42598425196851
"259" 11.940960451977402
"26" 15.321031746031744
"260" 15.949111951415157
"261" 22.856443766671017
"262" 12.813506506295273
"263" 12.435918303995674
"264" 15.22916202082754
"265" 9.64018060157215
"27" 36.43333333333333
"28" 27.126793598234
"29" 16.22888888888889
"3" 15.43988095238095
"30" 5.054166666666666
"31" 21.14512195121951
"32" 10.317543859649122
"33" 18.06181397528261
"34" 14.182203389830509
"35" 26.8255905511811
"36" 14.360132720775908
"37" 17.069111747851
"38" 25.03888888888889
"39" 25.76678571428571
"4" 15.259576630181936
"40" 17.326805555555556
"41" 13.504857533335132
"42" 13.13328397401725
"43" 16.252798079753713
"45" 18.737312775987583
"46" 14.841666666666667
"47" 25.1599173553719
"48" 14.834210250314609
"49" 14.021429150453956
"5" 1.2499999999999998
"50" 14.889709471276278
```

hadoop@ip-172-31-19-74:~

```
"50" 14.889709471276278
"51" 11.613043478260874
"52" 17.119109663409336
"53" 14.749761904761906
"54" 14.592109634551496
"55" 31.638297872340427
"56" 17.991346153846155
"57" 16.201515151515153
"58" 5.794444444444444
"59" 32.33333333333333
"6" 6.461666666666667
"60" 11.427136752136752
"61" 15.892464646464646
"62" 17.689290364583336
"63" 15.502140672782875
"64" 15.840476190476192
"65" 17.54645821543105
"66" 18.457041299932296
"67" 12.864242424242423
"68" 15.613557365231687
"69" 15.813785557986874
"7" 13.390857052144032
"70" 25.083014659018485
"71" 17.10883190883191
"72" 15.213888888888889
"73" 14.358823529411765
"74" 12.686163376446553
"75" 12.997183555204359
"76" 17.55950570342205
"77" 14.869811320754717
"78" 10.655509641873278
"79" 15.701342585260225
"8" 17.285202492211837
"80" 15.716174113876319
"81" 8.749404761904763
"82" 14.267069649945867
"83" 16.50727321524423
"84" 19.933333333333334
"85" 18.134055118110236
"86" 2.5444444444444447
"87" 20.73834169510518
"88" 22.11798083389375
"89" 16.169079939668173
"9" 60.55438596491228
```

```

"90"    14.381375588561125
"91"    16.980269607843137
"92"    17.198686868686867
"93"    34.155569306930694
"94"    12.423666666666666
"95"    19.998493900921087
"96"    13.543750000000001
"97"    17.07624567013056
"98"    14.992424242424242
"99"    13.445833333333335
[hadoop@ip-172-31-19-74 ~]$

```

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

**Data file:** Using the file '**yellow\_tripdata\_2017-04.csv**' for this exercise.

**Code snippet:**

```

[hadoop@ip-172-31-16-177 ~]$ cat task4_e.py
# Calculate the average tips to revenue ratio of the drivers for different pickup locations in sorted format.

from mrjob.job import MRJob

class AverageTipsToRevenueRatio(MRJob):

    def mapper(self, _, line):
        # Skip the header line
        if not line.startswith('VendorID'):
            fields = line.split(',')
            pickup_location = fields[7]
            total_revenue = float(fields[16])
            tips = float(fields[13])
            yield pickup_location, (tips, total_revenue)

    def combiner(self, pickup_location, tips_revenues):
        total_tips = 0
        total_revenue = 0
        for tips, revenue in tips_revenues:
            total_tips += tips
            total_revenue += revenue
        yield pickup_location, (total_tips, total_revenue)

    def reducer(self, pickup_location, tips_revenues):
        total_tips = 0
        total_revenue = 0
        for tips, revenue in tips_revenues:
            total_tips += tips
            total_revenue += revenue
        average_tips_to_revenue_ratio = total_tips / total_revenue
        yield pickup_location, average_tips_to_revenue_ratio

if __name__ == '__main__':
    AverageTipsToRevenueRatio.run()
[hadoop@ip-172-31-16-177 ~]$

```

**Result:**

hadoop@ip-172-31-19-74:~

```
[hadoop@ip-172-31-19-74 ~]$ python task4_e.py yellow_tripdata_2017-04.csv > task4_e_result.txt
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/task4_e.hadoop.20231106.084922.169511
Running step 1 of 1...
Job output is in /tmp/task4_e.hadoop.20231106.084922.169511/output
Streaming final output from /tmp/task4_e.hadoop.20231106.084922.169511/output...
Removing temp directory /tmp/task4_e.hadoop.20231106.084922.169511...
[hadoop@ip-172-31-19-74 ~]$ cat task4_e_result.txt
"1" 0.12661883871869564
"10" 0.10374659165851471
"100" 0.09888719118269991
"101" 0.06552103121747421
"102" 0.07687790888132072
"105" 0.08322903629536921
"106" 0.11489709795296812
"107" 0.1192135776646746
"108" 0.02637088606517613
"109" 0.08011566592957985
"11" 0.0578386861252612
"111" 0.10726209670852266
"112" 0.10868076575531245
"113" 0.11716578188321991
"114" 0.1155210998395174
"115" 0.10324209081775185
"116" 0.09057735665574537
"117" 0.03951297955433034
"118" 0.09651874490956154
"119" 0.07177895461017873
"12" 0.08115904655840858
"120" 0.06227222584799847
"121" 0.10734300002513006
"122" 0.12171573584689652
"123" 0.21293053320531716
"124" 0.08303745158877048
"125" 0.12278634532282971
"126" 0.05357484060909746
"127" 0.08312914756062732
"128" 0.10027241104731568
"129" 0.06264643144946791
"13" 0.11087440981526232
"130" 0.09913268897650081
"131" 0.05726526279301529
"132" 0.09983340159448627
```

hadoop@ip-172-31-19-74:~

```
"133" 0.08930138256158082
"134" 0.09237952501729313
"135" 0.09157386240501104
"136" 0.03830919026534907
"137" 0.11280760458073265
"138" 0.13111897454914995
"139" 0.1050425143885744
"14" 0.09517203758994239
"140" 0.11183247702007675
"141" 0.11157929454908735
"142" 0.11186639091494581
"143" 0.11350632532597896
"144" 0.10969289921048324
"145" 0.0874989617604838
"146" 0.08729057089246053
"147" 0.03434268608866193
"148" 0.11382927194745346
"149" 0.06791303805624259
"15" 0.07532477891105593
"150" 0.07385954720056882
"151" 0.1069404640403521
"152" 0.0848723233014725
"153" 0.08461991277125197
"154" 0.07518712992961682
"155" 0.07553050183043379
"156" 0.09487707703730042
"157" 0.09225368652535867
"158" 0.1168768302721272
"159" 0.046382919512766914
"16" 0.0853732762719924
"160" 0.08581351670956266
"161" 0.11336147822631225
"162" 0.11872490124291064
"163" 0.10982363214379434
"164" 0.1098321013962342
"165" 0.03125672119223685
"166" 0.11370637145968245
"167" 0.05276061749381539
"168" 0.05439199467241606
"169" 0.06302230517474797
"17" 0.07926504789426335
"170" 0.11715397725778143
"171" 0.07690063692202972
"172" 0.11682592111989305
```

hadoop@ip-172-31-19-74:~

```
"173" 0.06382257977413355
"174" 0.06610759328065309
"175" 0.1581804619294949
"176" 0.1547458389563653
"177" 0.09064347395462567
"178" 0.051630944464337476
"179" 0.08394181171510168
"18" 0.06566376001242538
"180" 0.08771509464408833
"181" 0.11148196332052782
"182" 0.06129255369033105
"183" 0.1316594352372146
"184" 0.06065196656006284
"185" 0.07086932439108369
"186" 0.1080234867517429
"187" 0.08561956071280563
"188" 0.07871499072646206
"189" 0.11238088203943673
"19" 0.061219953640491535
"190" 0.10047272831150039
"191" 0.09819037611315835
"192" 0.05340498987745361
"193" 0.05572278812633908
"194" 0.11657972540696193
"195" 0.11636360375577366
"196" 0.07887413424058097
"197" 0.07257688335054813
"198" 0.09527226536139712
"199" 0.0
"2" 0.1677090809665402
"20" 0.1039202235347265
"200" 0.07504913809472488
"201" 0.12449299410029496
"202" 0.08443268758153132
"203" 0.11528323272794373
"204" 0.04293318002262569
"205" 0.10187177178600197
"206" 0.027821865372047425
"207" 0.08593152505624571
"208" 0.06781370217745851
"209" 0.11213804982509191
"21" 0.12055145332707326
"210" 0.08153520159966575
"211" 0.10975277039089627
```



hadoop@ip-172-31-19-74:~

```
"211" 0.10975277039089627
"212" 0.1081600393756685
"213" 0.08565432757904728
"214" 0.08754208754208755
"215" 0.0980709529411435
"216" 0.10005995183979269
"217" 0.07957201581403053
"218" 0.09503924297972174
"219" 0.09409349131303146
"22" 0.05099020873261107
"220" 0.06470576096657536
"221" 0.10664324928458259
"222" 0.08533611981887844
"223" 0.09609672282391263
"224" 0.11508568145121476
"225" 0.0810049772745501
"226" 0.08249922024958527
"227" 0.08006423056190354
"228" 0.09329682940945985
"229" 0.1119874592687155
"23" 0.11618037135278515
"230" 0.10378201617257417
"231" 0.11771662151875441
"232" 0.1066199488869949
"233" 0.11481151082800288
"234" 0.12013625469603266
"235" 0.05897125215303809
"236" 0.11231908289724975
"237" 0.10890396072449153
"238" 0.1128468161981967
"239" 0.1136117716601222
"24" 0.10599969308708145
"240" 0.013227863170983359
"241" 0.07945623083603622
"242" 0.04641100863215212
"243" 0.09866957779684171
"244" 0.0948790729669047
"245" 0.1465599390127692
"246" 0.11551907392804958
"247" 0.07545141978121578
"248" 0.037403915235594436
"249" 0.11835553004528475
"25" 0.10819873878001712
"250" 0.05342036200115804
```

hadoop@ip-172-31-19-74:~

```
"250" 0.05342036200115804
"251" 0.034833118544666704
"252" 0.09884044983114938
"253" 0.09880957363483628
"254" 0.07428890857084779
"255" 0.11535324461161518
"256" 0.10611641050481833
"257" 0.1271329145780954
"258" 0.2075796289715778
"259" 0.10366930371130974
"26" 0.08655640514326231
"260" 0.07108066569595983
"261" 0.10416763154155705
"262" 0.11345096467429362
"263" 0.11405024644036961
"264" 0.11101924291724302
"265" 0.1112947991800702
"27" 0.16664695446481373
"28" 0.08783540762834087
"29" 0.04191703467747774
"3" 0.043470924195223255
"30" 0.0
"31" 0.07660638278098919
"32" 0.07126422486163796
"33" 0.1176183517922327
"34" 0.10536520205426156
"35" 0.07483109044680833
"36" 0.10763254759932682
"37" 0.09775462518343546
"38" 0.06804233196260759
"39" 0.05624786919227044
"4" 0.10762468274888427
"40" 0.12070532904844701
"41" 0.09060433931239885
"42" 0.07044385617962277
"43" 0.10280777673360256
"45" 0.09666163030375485
"46" 0.08687837028160575
"47" 0.04250030091950533
"48" 0.10500967305120536
"49" 0.10322590753387893
"5" 0.04452415112386418
"50" 0.10651900644189898
"51" 0.0840065468740427
```

hadoop@ip-172-31-19-74:~

```
"60" 0.029510328227909815
"61" 0.09107979269569554
"62" 0.08119043979004235
"63" 0.10752448599463489
"64" 0.11408939986529756
"65" 0.10338479613875887
"66" 0.11178836485929464
"67" 0.05912722235049643
"68" 0.11271612813585354
"69" 0.05325839326396382
"7" 0.07697612913010916
"70" 0.10992329601863436
"71" 0.11343443657000002
"72" 0.07561321409450536
"73" 0.05208182213840128
"74" 0.07117463224568796
"75" 0.08913222843944525
"76" 0.0850366352185918
"77" 0.055469865205353464
"78" 0.09110575483561889
"79" 0.11740089370261075
"8" 0.10124318156299858
"80" 0.1110781201676926
"81" 0.05065504068374593
"82" 0.06664748039192138
"83" 0.05078383871574121
"84" 0.1306650677059911
"85" 0.07875213195641236
"86" 0.08083055246570264
"87" 0.12329279028457553
"88" 0.11258004298387361
"89" 0.08214199665133394
"9" 0.06032862084523987
"90" 0.09770660285405965
"91" 0.06943235194104917
"92" 0.06842726381139286
"93" 0.10684346474310821
"94" 0.03361085827488355
"95" 0.08219750024094305
"96" 0.06923837784371908
"97" 0.0997775408843665
"98" 0.06389513024038299
"99" 0.036276223776223776
[hadoop@ip-172-31-19-74 ~]$
```

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

**Data file:** Using the file 'yellow\_tripdata\_2017-04.csv' for this exercise.

## Code snippet:

```
hadoop@ip-172-31-16-177:~$ cat task4_f.py
# 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).

from mrjob.job import MRJob
from datetime import datetime

class AverageRevenueOverTime(MRJob):

    def parse_datetime(self, datetime_str):
        formats = ['%d-%m-%Y %H:%M:%S', '%d-%m-%Y %H:%M', '%Y-%m-%d %H:%M', '%Y-%m-%d %H:%M:%S']
        for fmt in formats:
            try:
                return datetime.strptime(datetime_str, fmt)
            except ValueError:
                pass
        raise ValueError('no valid date format found')

    def mapper(self, _, line):
        # Skip the header line
        if not line.startswith('VendorID'):
            fields = line.split(',')
            revenue = float(fields[16])
            pickup_datetime = self.parse_datetime(fields[1])
            month = pickup_datetime.month
            hour = pickup_datetime.hour
            weekday = pickup_datetime.weekday()
            yield (month, hour, weekday), revenue

    def reducer(self, key, values):
        total_revenue = 0
        num_trips = 0

        for revenue in values:
            total_revenue += revenue
            num_trips += 1

        average_revenue = total_revenue / num_trips

        yield key, average_revenue

if __name__ == '__main__':
    AverageRevenueOverTime.run()
```

## Result:

hadoop@ip-172-31-16-177:~

```
[hadoop@ip-172-31-16-177 ~]$ python task4_f.py yellow_tripdata_2017-04.csv > task4_f.txt
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/task4_f.hadoop.20231106.100916.126759
Running step 1 of 1...
job output is in /tmp/task4_f.hadoop.20231106.100916.126759/output
Streaming final output from /tmp/task4_f.hadoop.20231106.100916.126759/output...
Removing temp directory /tmp/task4_f.hadoop.20231106.100916.126759...
[hadoop@ip-172-31-16-177 ~]$ cat task4_f.txt
[4, 0, 0]      19.80647470053675
[4, 0, 1]      18.794717807035205
[4, 0, 2]      18.67812945973316
[4, 0, 3]      17.856533620807255
[4, 0, 4]      18.15557276678711
[4, 0, 5]      16.9124526059859
[4, 0, 6]      15.542700442549934
[4, 1, 0]      18.431992453922554
[4, 1, 1]      17.60355366026755
[4, 1, 2]      18.574224053254458
[4, 1, 3]      17.021952936812973
[4, 1, 4]      17.693392430957694
[4, 1, 5]      16.25433095878572
[4, 1, 6]      15.213919323210398
[4, 10, 0]     15.993256704988335
[4, 10, 1]     15.57174358566604
[4, 10, 2]     15.901819135946711
[4, 10, 3]     15.979665633014918
[4, 10, 4]     15.97471915931524
[4, 10, 5]     13.695509758832543
[4, 10, 6]     14.093601629098034
[4, 11, 0]     19.70811885113746
[4, 11, 1]     16.010772072533218
[4, 11, 2]     15.97254091588818
[4, 11, 3]     16.61980887765878
[4, 11, 4]     16.30626708776126
[4, 11, 5]     14.033743583021367
[4, 11, 6]     14.445806276315645
[4, 12, 0]     16.0949626513349
[4, 12, 1]     15.691658636989088
[4, 12, 2]     16.32499333063781
[4, 12, 3]     16.902250938566826
[4, 12, 4]     16.493451288847105
[4, 12, 5]     14.506856961056513
[4, 12, 6]     15.062507284243964
```

hadoop@ip-172-31-16-177:~

```
[4, 13, 0] 16.24414824647368
[4, 13, 1] 16.01739194415371
[4, 13, 2] 16.917057260757176
[4, 13, 3] 17.554169002460814
[4, 13, 4] 17.114006931654103
[4, 13, 5] 15.299302459182867
[4, 13, 6] 15.884092401300672
[4, 14, 0] 16.603916121846517
[4, 14, 1] 16.24374837336133
[4, 14, 2] 17.075083406883984
[4, 14, 3] 17.886031287751955
[4, 14, 4] 17.435698763995457
[4, 14, 5] 15.74869930565773
[4, 14, 6] 16.840753872341182
[4, 15, 0] 16.415363554903315
[4, 15, 1] 15.856924973245892
[4, 15, 2] 16.83182181191212
[4, 15, 3] 17.626056872448913
[4, 15, 4] 17.24548536361626
[4, 15, 5] 15.830416432397827
[4, 15, 6] 16.98076222352124
[4, 16, 0] 17.74133845166477
[4, 16, 1] 17.30359002032584
[4, 16, 2] 18.370869572157154
[4, 16, 3] 19.27780740375492
[4, 16, 4] 18.661179875981706
[4, 16, 5] 15.483719861959464
[4, 16, 6] 17.26051876421012
[4, 17, 0] 16.9353422344903
[4, 17, 1] 16.62974647221145
[4, 17, 2] 17.55105593631472
[4, 17, 3] 18.24797680251243
[4, 17, 4] 17.973244249944056
[4, 17, 5] 15.481917364996026
[4, 17, 6] 17.193937591282822
[4, 18, 0] 15.891023042168307
[4, 18, 1] 15.727883363375527
[4, 18, 2] 16.369772137047704
[4, 18, 3] 17.069782585925886
[4, 18, 4] 16.370225167639607
[4, 18, 5] 14.624873769592817
[4, 18, 6] 16.301217037129764
[4, 19, 0] 15.816425920002926
[4, 19, 1] 15.497339795567433
```

hadoop@ip-172-31-16-177:~

```
[4, 19, 2] 15.987707347852963
[4, 19, 3] 16.50774043785117
[4, 19, 4] 15.762378440794846
[4, 19, 5] 19.674569517438382
[4, 19, 6] 16.25635898224874
[4, 2, 0] 16.386729640930223
[4, 2, 1] 16.34083949787088
[4, 2, 2] 17.59978826484732
[4, 2, 3] 16.37205057846147
[4, 2, 4] 18.132232268254285
[4, 2, 5] 15.685473777728008
[4, 2, 6] 15.17695421262324
[4, 20, 0] 15.806433938201225
[4, 20, 1] 15.748832246881944
[4, 20, 2] 16.056880185873606
[4, 20, 3] 16.359462410280226
[4, 20, 4] 15.739578861645962
[4, 20, 5] 15.318073324437082
[4, 20, 6] 16.982370803329676
[4, 21, 0] 16.386294148043216
[4, 21, 1] 16.334466899807676
[4, 21, 2] 16.44989098881493
[4, 21, 3] 17.05383435867956
[4, 21, 4] 16.176966977489336
[4, 21, 5] 15.332053540601459
[4, 21, 6] 17.173090044917842
[4, 22, 0] 17.034962261919357
[4, 22, 1] 16.766731003338386
[4, 22, 2] 17.180006232602164
[4, 22, 3] 17.5076748235317
[4, 22, 4] 16.520808734827458
[4, 22, 5] 15.501073974969025
[4, 22, 6] 18.05421073846853
[4, 23, 0] 17.742434453538742
[4, 23, 1] 17.672756828321315
[4, 23, 2] 17.52890607368994
[4, 23, 3] 18.103497884552787
[4, 23, 4] 17.27365807245065
[4, 23, 5] 16.16109287687102
[4, 23, 6] 19.537749787659926
[4, 3, 0] 17.035402713379163
[4, 3, 1] 16.68720682663176
[4, 3, 2] 16.80449137931156
[4, 3, 3] 17.008209039548152
```

hadoop@ip-172-31-16-177:~

```
[4, 3, 4]      18.606886082748098
[4, 3, 5]      15.93510123368352
[4, 3, 6]      15.842982230339656
[4, 4, 0]      22.075331505482715
[4, 4, 1]      19.685773360418857
[4, 4, 2]      20.618848994142827
[4, 4, 3]      20.36250086795426
[4, 4, 4]      20.232893507327628
[4, 4, 5]      17.593235454788644
[4, 4, 6]      17.121296163282572
[4, 5, 0]      21.24480678670053
[4, 5, 1]      18.90079500756198
[4, 5, 2]      19.290775138158594
[4, 5, 3]      19.64950030551685
[4, 5, 4]      21.084620749778868
[4, 5, 5]      20.46585589308732
[4, 5, 6]      21.08362532981141
[4, 6, 0]      16.050704128041534
[4, 6, 1]      14.55793618895153
[4, 6, 2]      15.126502746810866
[4, 6, 3]      15.132386678529572
[4, 6, 4]      16.130625340167793
[4, 6, 5]      18.827777999099393
[4, 6, 6]      20.668970097719725
[4, 7, 0]      15.040514168915262
[4, 7, 1]      14.196078141746195
[4, 7, 2]      14.375317088653532
[4, 7, 3]      14.415348940380365
[4, 7, 4]      15.069849910325587
[4, 7, 5]      16.129460167015672
[4, 7, 6]      17.357560885301037
[4, 8, 0]      15.16209072532834
[4, 8, 1]      14.541790406913229
[4, 8, 2]      14.883921496467169
[4, 8, 3]      14.895804992617485
[4, 8, 4]      14.856661646321232
[4, 8, 5]      14.317182669008766
[4, 8, 6]      14.961029522661692
[4, 9, 0]      15.706627748472677
[4, 9, 1]      15.163883283185895
[4, 9, 2]      15.335301566626068
[4, 9, 3]      15.331071204911218
[4, 9, 4]      15.56764987159916
[4, 9, 5]      13.457522095582306
```

```
[4, 9, 6]      13.94283034104963
```

```
[hadoop@ip-172-31-16-177 ~]$
```