MapReduce Tasks:

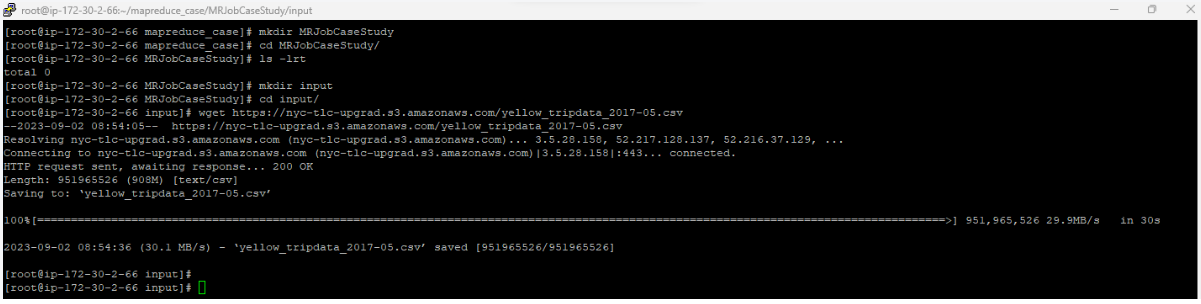
Task 4. Write MapReduce codes to perform the tasks using the files you have downloaded on your EMR Instance:

Solution:

Step 1: Launch EMR cluster

Step2: connect to cluster using putty

Step3: make a folder in your local EMR instance where you will import your data set and store your scripts for ease of use



```
[root@ip-172-30-2-66 mapreduce_case]# mkdir MRJobCaseStudy
[root@ip-172-30-2-66 mapreduce_case]# cd MRJobCaseStudy/
[root@ip-172-30-2-66 MRJobCaseStudy]# ls -lrt
total 0
[root@ip-172-30-2-66 MRJobCaseStudy]# mkdir input
[root@ip-172-30-2-66 MRJobCaseStudy]# cd input/
[root@ip-172-30-2-66 input]# wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-05.csv
--2023-09-02 08:54:05--  https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-05.csv
Resolving nyc-tlc-upgrad.s3.amazonaws.com (nyc-tlc-upgrad.s3.amazonaws.com)... 3.5.28.158, 52.217.128.137, 52.216.37.129, ...
Connecting to nyc-tlc-upgrad.s3.amazonaws.com (nyc-tlc-upgrad.s3.amazonaws.com)|3.5.28.158|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 951965526 (908M) [text/csv]
Saving to: ‘yellow_tripdata_2017-05.csv’

100%[=========================================================================================================>] 951,965,526 29.9MB/s   in 30s

2023-09-02 08:54:36 (30.1 MB/s) - ‘yellow_tripdata_2017-05.csv’ saved [951965526/951965526]

[root@ip-172-30-2-66 input]#
[root@ip-172-30-2-66 input]# 
```

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

Solution:

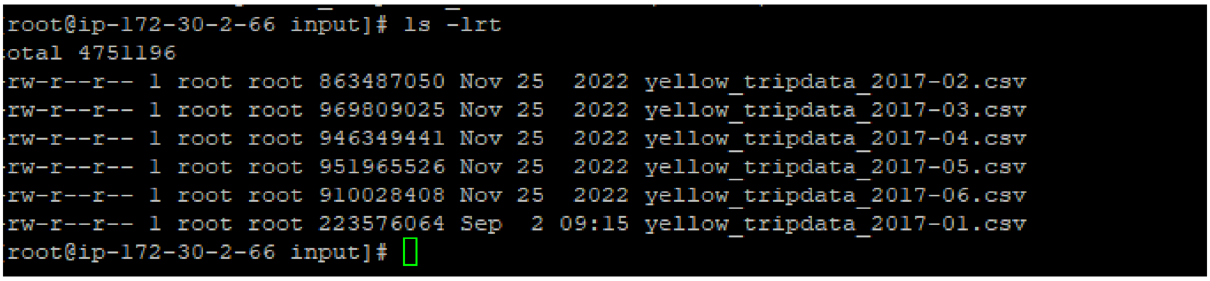
Steps1: Create mrtask_a.py under MRJobCaseStudy folder

Step2. Create output dir to save output file

Step: Run command to generate revenue in outputfile

python mrtask_a.py input > output/mrtaska_a.txt

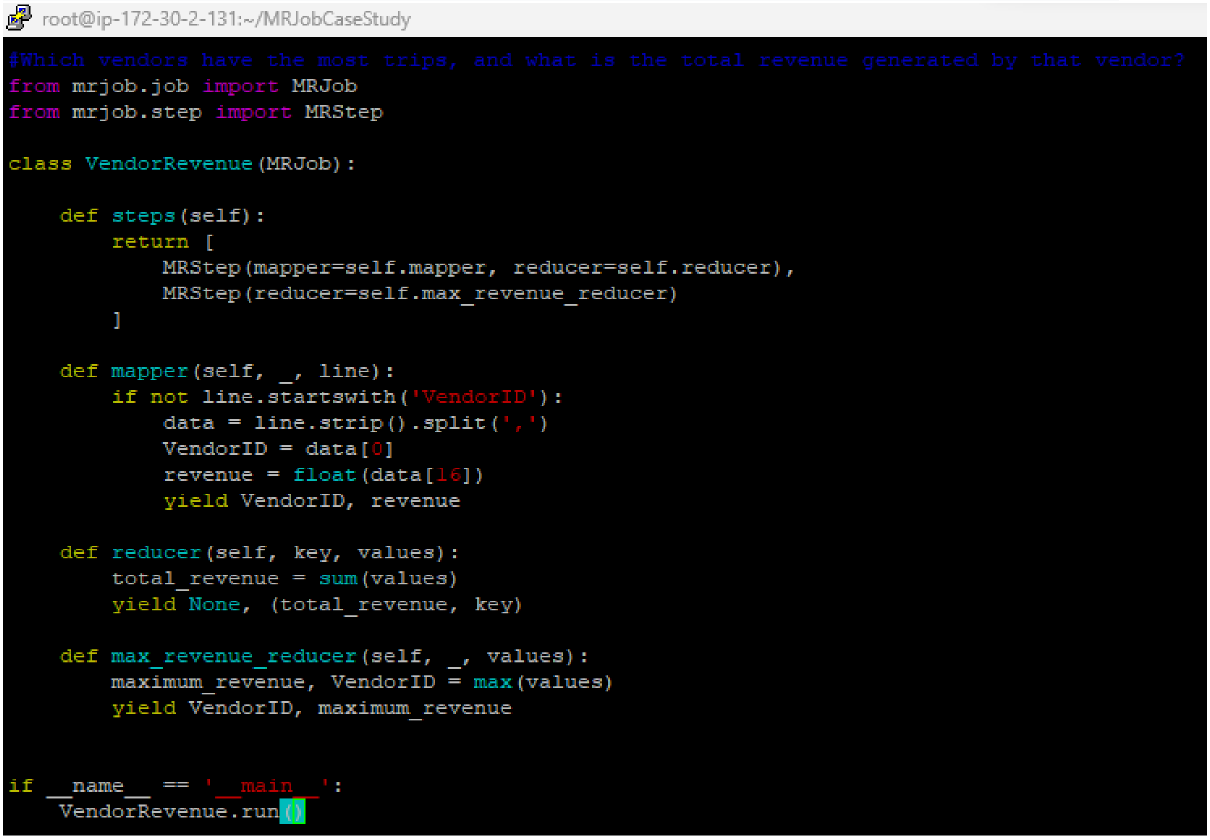
INPUT DATA:



```
root@ip-172-30-2-66 input]# ls -lrt
otal 4751196
rw-r--r-- 1 root root 863487050 Nov 25  2022 yellow_tripdata_2017-02.csv
rw-r--r-- 1 root root 969809025 Nov 25  2022 yellow_tripdata_2017-03.csv
rw-r--r-- 1 root root 946349441 Nov 25  2022 yellow_tripdata_2017-04.csv
rw-r--r-- 1 root root 951965526 Nov 25  2022 yellow_tripdata_2017-05.csv
rw-r--r-- 1 root root 910028408 Nov 25  2022 yellow_tripdata_2017-06.csv
rw-r--r-- 1 root root 223576064 Sep  2 09:15 yellow_tripdata_2017-01.csv
root@ip-172-30-2-66 input]# 
```

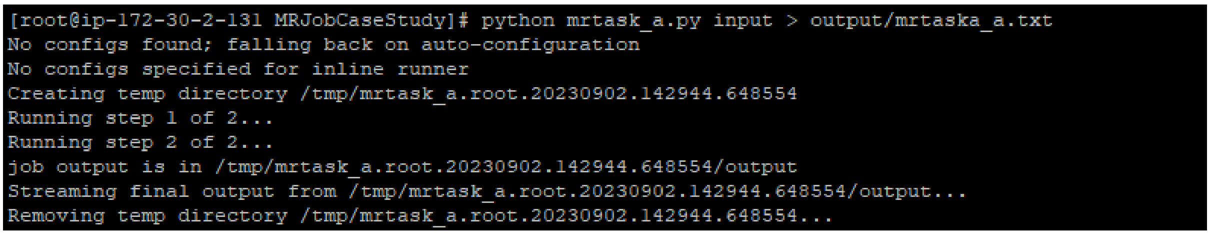
CODE: Revenue details are store in total_amount at 16 index location in input files.

Mapper will create dict with VendorID key and revenue as value, where reducer will sum up the total revenue for each vendor which will be finally reduce with max_revenue_reducer task



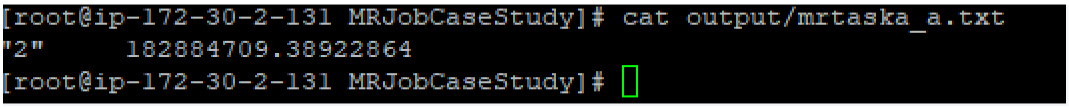
```
#Which vendors have the most trips, and what is the total revenue generated by that vendor?
from mrjob.job import MRJob
from mrjob.step import MRStep

class VendorRevenue(MRJob):

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

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

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

    def max_revenue_reducer(self, _, values):
        maximum_revenue, VendorID = max(values)
        yield VendorID, maximum_revenue


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

COMMAND: python mrtask_a.py input > output/mrtaska_a.txt



```
[root@ip-172-30-2-131 MRJobCaseStudy]# python mrtask_a.py input > output/mrtaska_a.txt
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/mrtask_a.root.20230902.142944.648554
Running step 1 of 2...
Running step 2 of 2...
job output is in /tmp/mrtask_a.root.20230902.142944.648554/output
Streaming final output from /tmp/mrtask_a.root.20230902.142944.648554/output...
Removing temp directory /tmp/mrtask_a.root.20230902.142944.648554...
```

OUTPUT: Vendor: 2: VeriFone Inc generate maximum revenue.

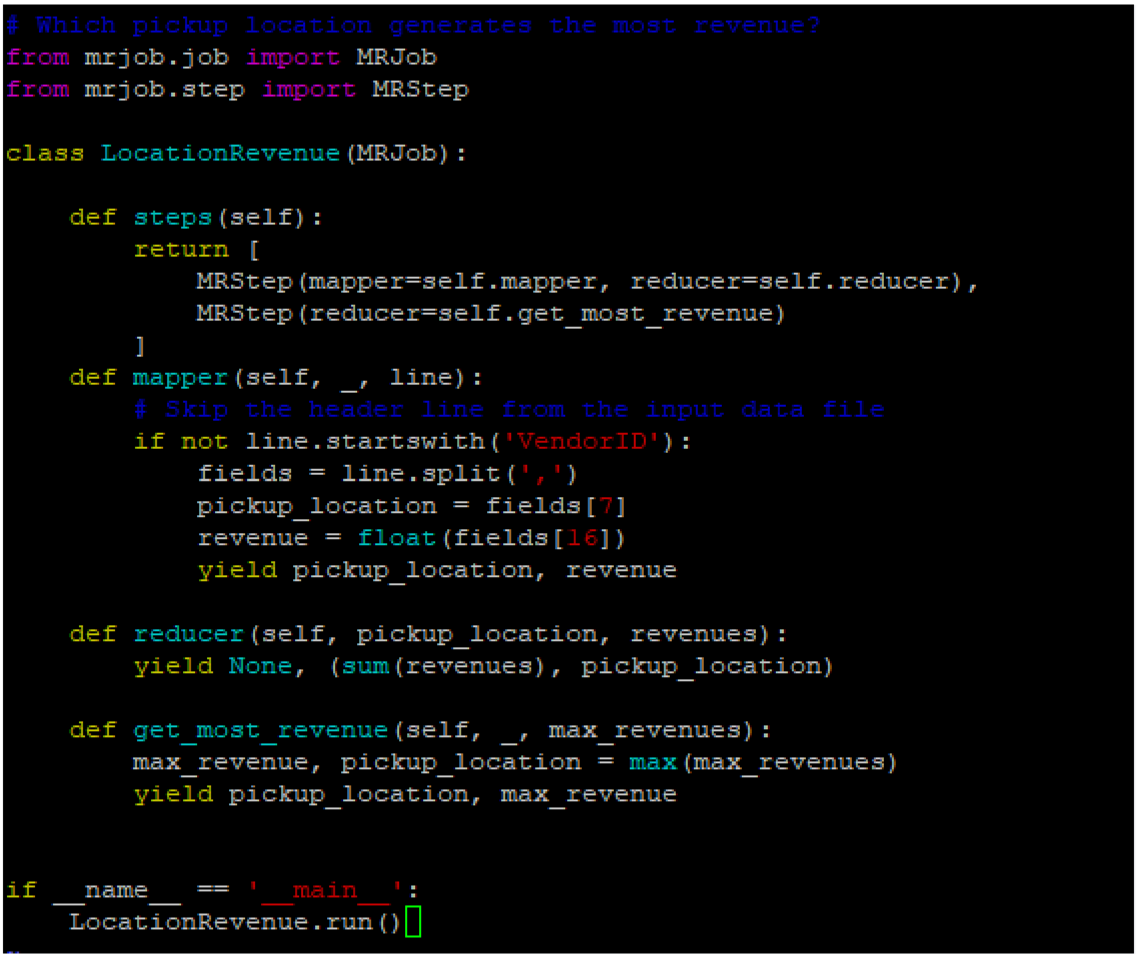


```
[root@ip-172-30-2-131 MRJobCaseStudy]# cat output/mrtaska_a.txt
"2"     182884709.38922864
[root@ip-172-30-2-131 MRJobCaseStudy]#
```

b) Which pickup location generates the most revenue?

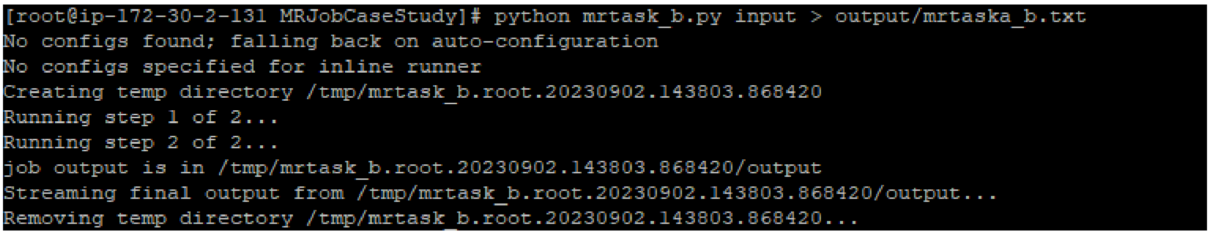
CODE: Pickup location details are retrieved from column PULocationID at index location 7 and revenue details from columns total_amount at 16th index location from input data.

Mapper will create the dict with pickup_location as key and revenue as value which will further reduce in reducer with sum of revenues for each location to get location with max revenue in the final output file.



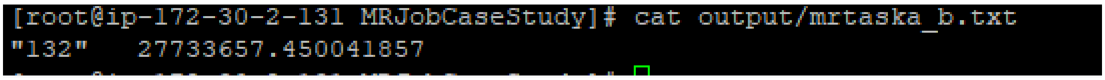
```
# Which pickup location generates the most revenue?
from mrjob.job import MRJob
from mrjob.step import MRStep

class LocationRevenue(MRJob):

    def steps(self):
        return [
            MRStep(mapper=self.mapper, reducer=self.reducer),
            MRStep(reducer=self.get_most_revenue)
        ]
    def mapper(self, _, line):
        # Skip the header line from the input data file
        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 get_most_revenue(self, _, max_revenues):
        max_revenue, pickup_location = max(max_revenues)
        yield pickup_location, max_revenue


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

COMMAND: python mrtask_b.py input > output/mrtaska_b.txt



```
[root@ip-172-30-2-131 MRJobCaseStudy]# python mrtask_b.py input > output/mrtaska_b.txt
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/mrtask_b.root.20230902.143803.868420
Running step 1 of 2...
Running step 2 of 2...
job output is in /tmp/mrtask_b.root.20230902.143803.868420/output
Streaming final output from /tmp/mrtask_b.root.20230902.143803.868420/output...
Removing temp directory /tmp/mrtask_b.root.20230902.143803.868420...
```

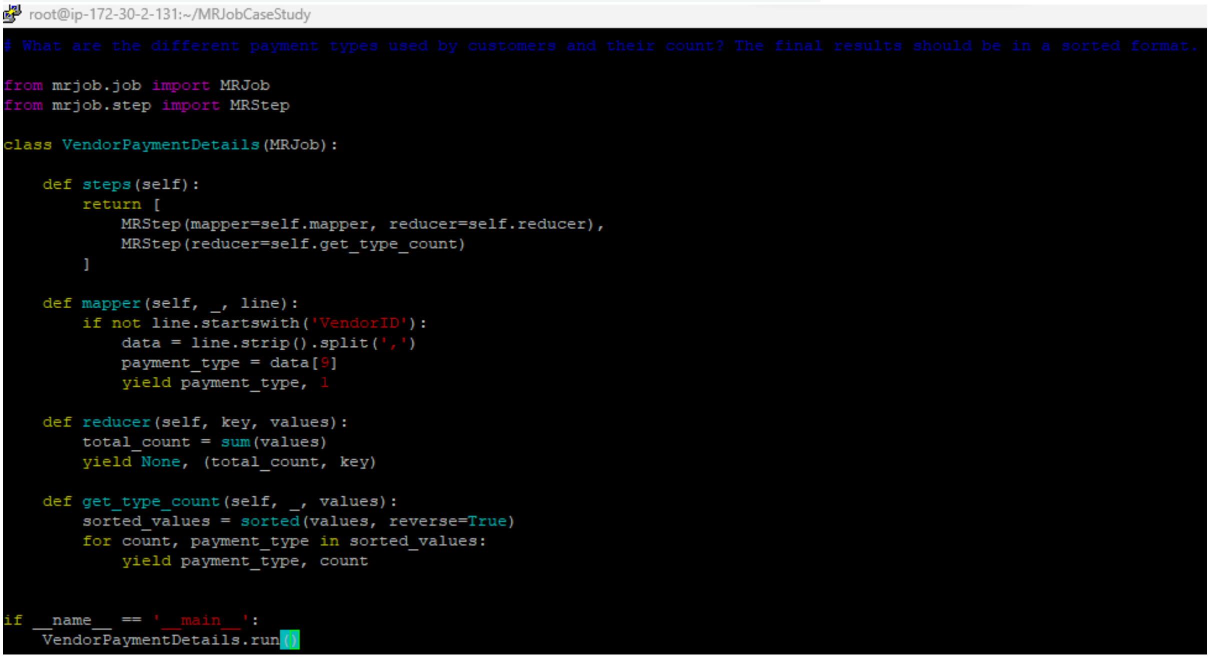
OUTPUT: Location ID 132 generates Maximum revenue.



```
[root@ip-172-30-2-131 MRJobCaseStudy]# cat output/mrtaska_b.txt
"132"    27733657.450041857
```

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

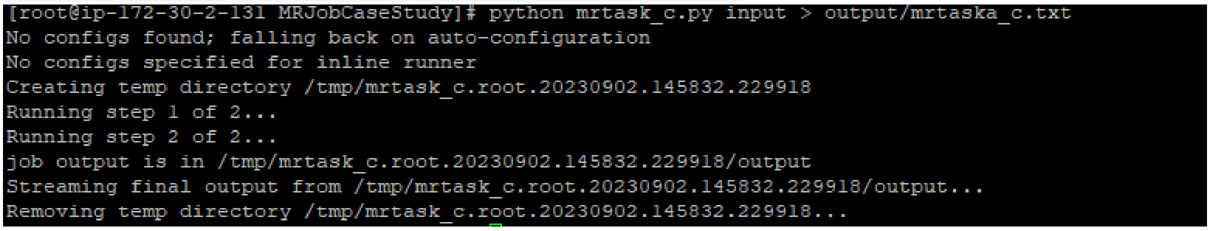
CODE Column payment_type help us to get the payment type details which will converted in dict in the mapper function and reducer will sum up the payment type with counter value 1 for each payment to get all payment type with overall count number.



root@ip-172-30-2-131:~/MRJobCaseStudy

```
# 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 VendorPaymentDetails(MRJob):

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

    def mapper(self, _, line):
        if not line.startswith('VendorID'):
            data = line.strip().split(',')
            payment_type = data[9]
            yield payment_type, 1

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

    def get_type_count(self, _, values):
        sorted_values = sorted(values, reverse=True)
        for count, payment_type in sorted_values:
            yield payment_type, count


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

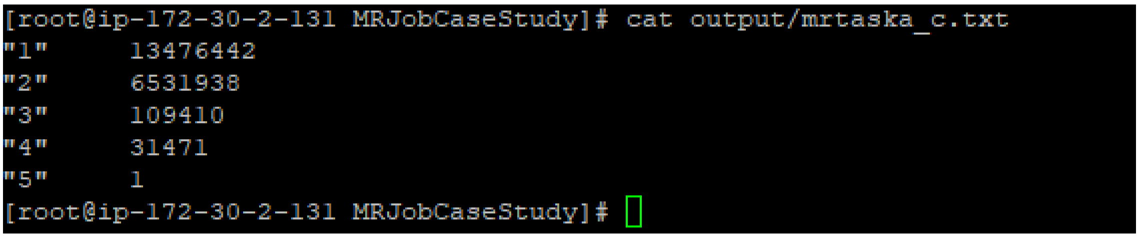
COMMAND: python mrtask_c.py input > output/mrtaska_c.txt



```
[root@ip-172-30-2-131 MRJobCaseStudy]# python mrtask_c.py input > output/mrtaska_c.txt
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/mrtask_c.root.20230902.145832.229918
Running step 1 of 2...
Running step 2 of 2...
job output is in /tmp/mrtask_c.root.20230902.145832.229918/output
Streaming final output from /tmp/mrtask_c.root.20230902.145832.229918/output...
Removing temp directory /tmp/mrtask_c.root.20230902.145832.229918...
```

OUTPUT Below Payment types describe by number id

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

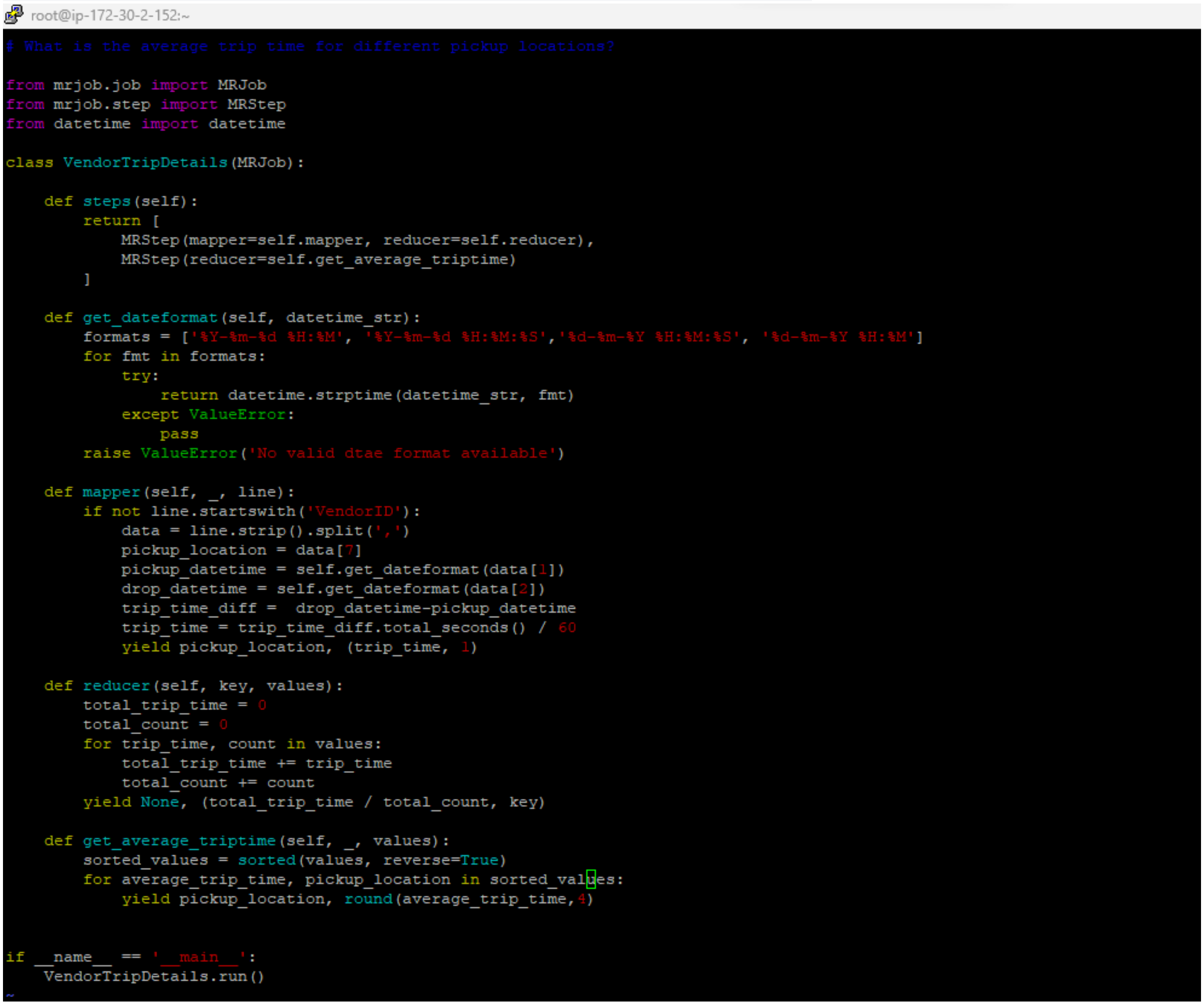


```
[root@ip-172-30-2-131 MRJobCaseStudy]# cat output/mrtaska_c.txt
"1"     13476442
"2"     6531938
"3"     109410
"4"     31471
"5"     1
[root@ip-172-30-2-131 MRJobCaseStudy]# 
```

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

CODE:

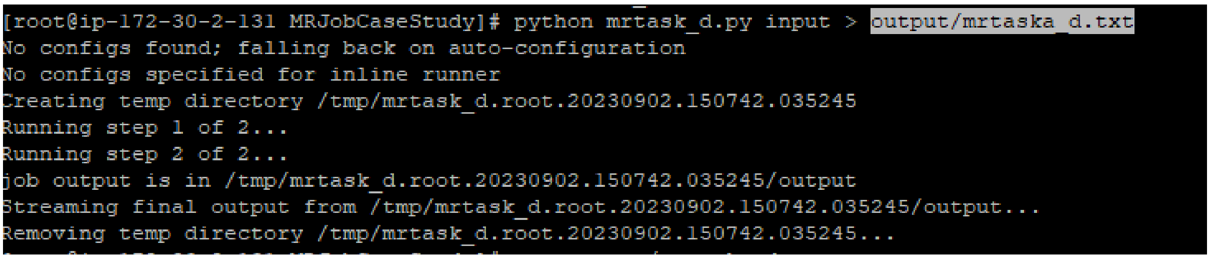
Explnation: olumn PULocationID at index location 7 help to get location detail and trip time can be calculated by using tpep_pickup_datetime and tpep_dropoff_datetime for each trip. Converting values into proper date format we get trip time for each trip in minutes and it will be passed to reducer to get total trip time for each location along with trip count which will be converted into average trip time in get_average_triptime method



root@ip-172-30-2-152:~

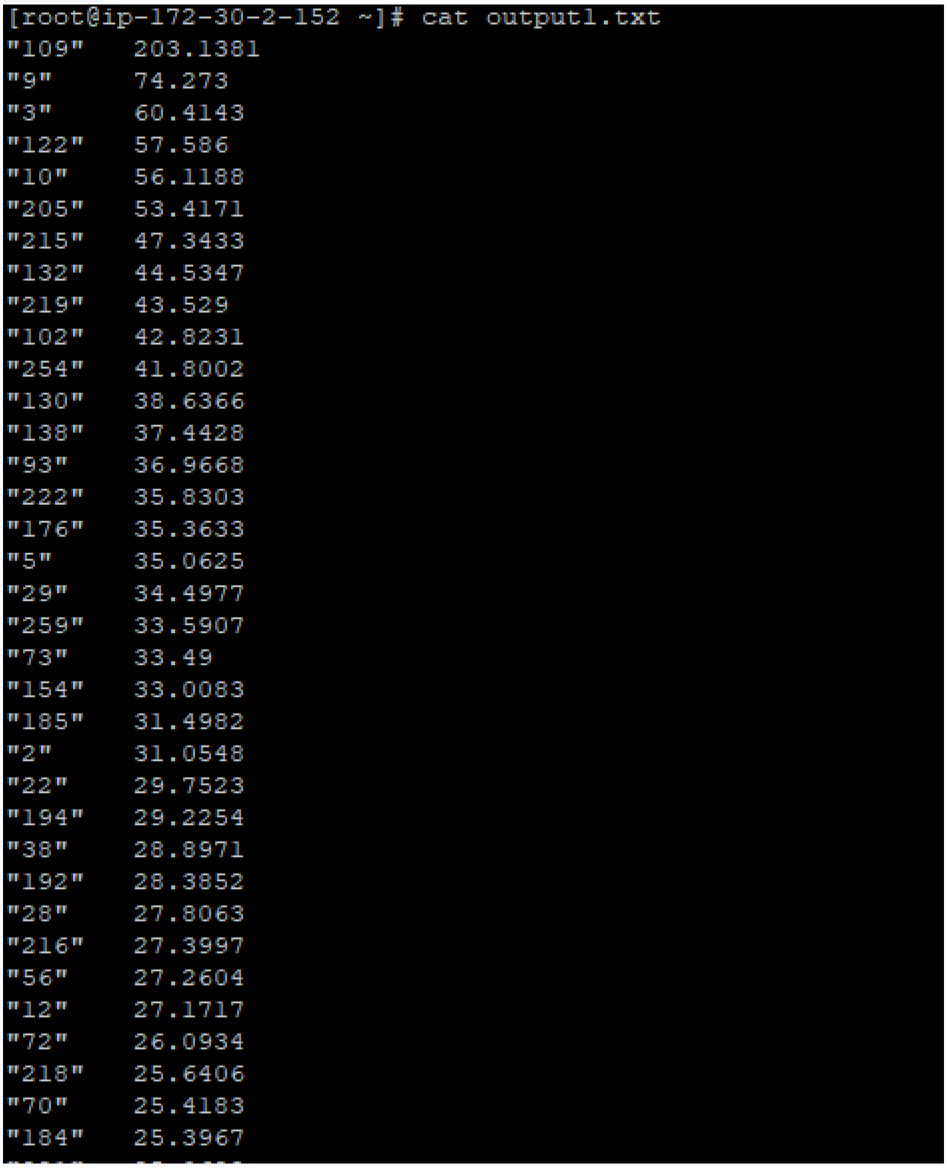
```
# What is the average trip time for different pickup locations?

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

class VendorTripDetails(MRJob):

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

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

    def mapper(self, _, line):
        if not line.startswith('VendorID'):
            data = line.strip().split(',')
            pickup_location = data[7]
            pickup_datetime = self.get_dateformat(data[1])
            drop_datetime = self.get_dateformat(data[2])
            trip_time_diff =  drop_datetime-pickup_datetime
            trip_time = trip_time_diff.total_seconds() / 60
            yield pickup_location, (trip_time, 1)

    def reducer(self, key, values):
        total_trip_time = 0
        total_count = 0
        for trip_time, count in values:
            total_trip_time += trip_time
            total_count += count
        yield None, (total_trip_time / total_count, key)

    def get_average_triptime(self, _, values):
        sorted_values = sorted(values, reverse=True)
        for average_trip_time, pickup_location in sorted_values:
            yield pickup_location, round(average_trip_time,4)


if __name__ == '__main__':
    VendorTripDetails.run()
~
```

COMMAND: python mrtask_d.py input > output/mrtaska_d.txt



```
[root@ip-172-30-2-131 MRJobCaseStudy]# python mrtask_d.py input > output/mrtaska_d.txt
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/mrtask_d.root.20230902.150742.035245
Running step 1 of 2...
Running step 2 of 2...
job output is in /tmp/mrtask_d.root.20230902.150742.035245/output
Streaming final output from /tmp/mrtask_d.root.20230902.150742.035245/output...
Removing temp directory /tmp/mrtask_d.root.20230902.150742.035245...
```

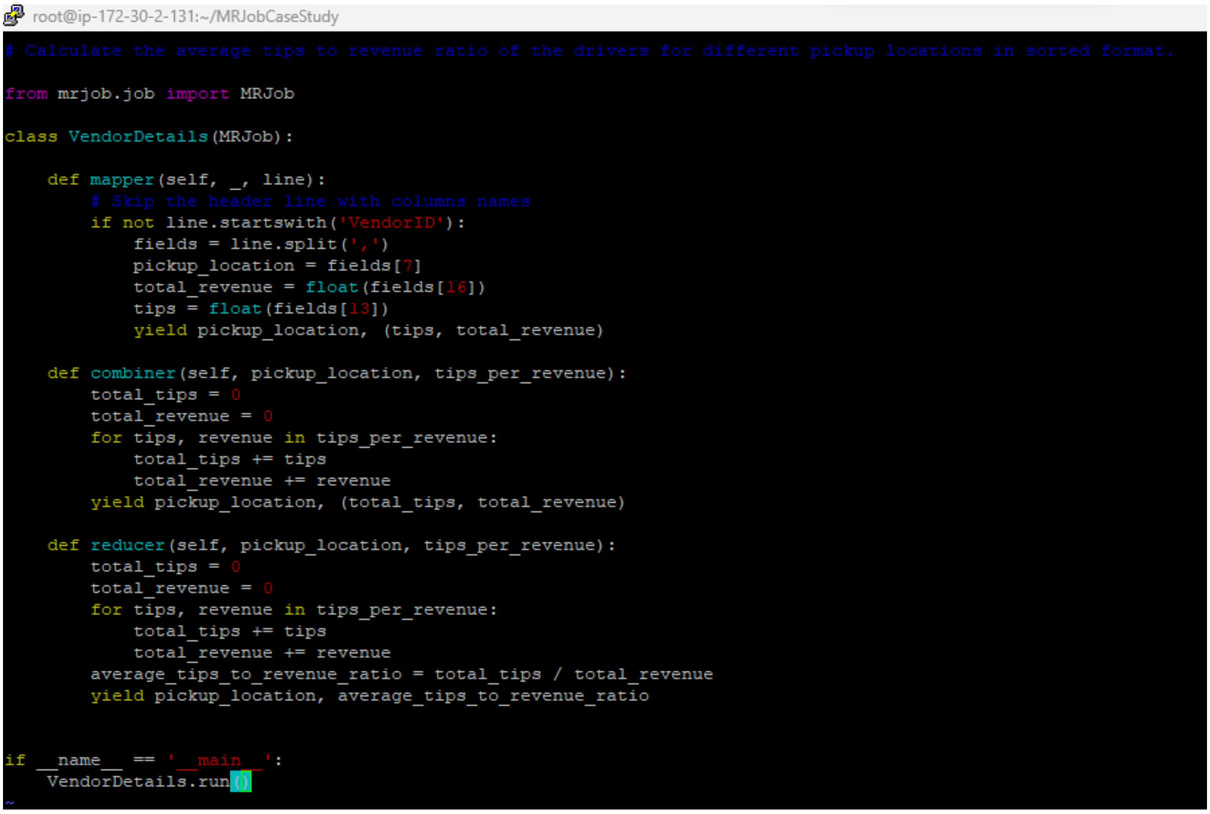
OUTPUT



```
[root@ip-172-30-2-152 ~]# cat output1.txt
"109"   203.1381
"9"     74.273
"3"     60.4143
"122"   57.586
"10"    56.1188
"205"   53.4171
"215"   47.3433
"132"   44.5347
"219"   43.529
"102"   42.8231
"254"   41.8002
"130"   38.6366
"138"   37.4428
"93"    36.9668
"222"   35.8303
"176"   35.3633
"5"     35.0625
"29"    34.4977
"259"   33.5907
"73"    33.49
"154"   33.0083
"185"   31.4982
"2"     31.0548
"22"    29.7523
"194"   29.2254
"38"    28.8971
"192"   28.3852
"28"    27.8063
"216"   27.3997
"56"    27.2604
"12"    27.1717
"72"    26.0934
"218"   25.6406
"70"    25.4183
"184"   25.3967
```

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

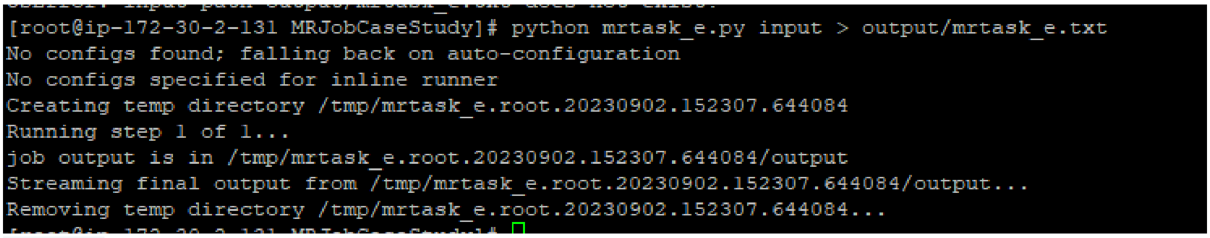
CODE PULocationID column helps us to get location details, Total_amount help us to get revenue details and tip_amount to get tip details for each trip. Mapper will help you with pickup_location to the tip and revenue details for each trip which will further combined by combiner to generate final tips to revenue ratio for each location and reducer will generate average ratio of tips to the revenue for each location.



root@ip-172-30-2-131:~/MRJobCaseStudy

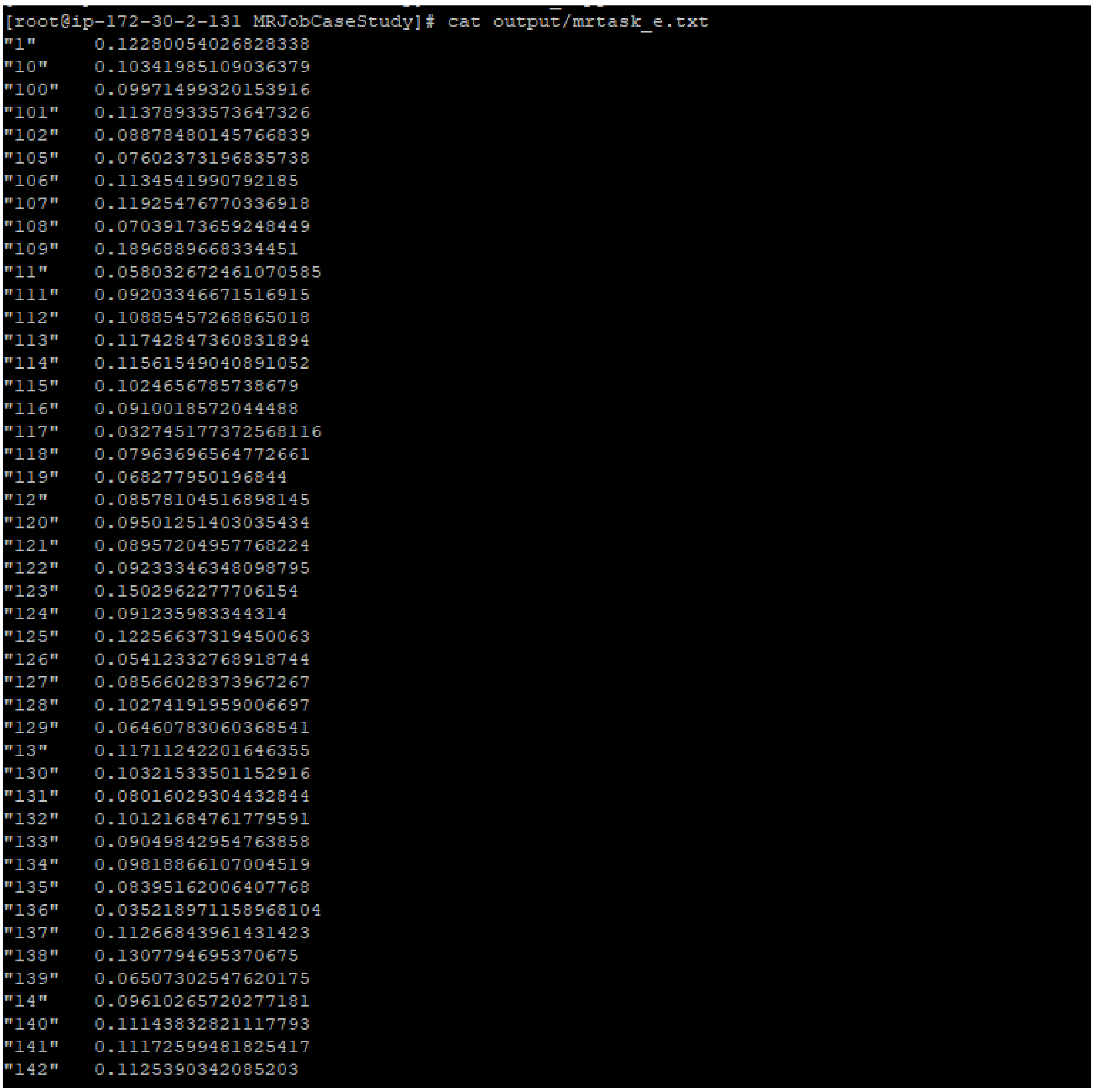
```
# Calculate the average tips to revenue ratio of the drivers for different pickup locations in sorted format.

from mrjob.job import MRJob

class VendorDetails(MRJob):

    def mapper(self, _, line):
        # Skip the header line with columns names
        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_per_revenue):
        total_tips = 0
        total_revenue = 0
        for tips, revenue in tips_per_revenue:
            total_tips += tips
            total_revenue += revenue
        yield pickup_location, (total_tips, total_revenue)

    def reducer(self, pickup_location, tips_per_revenue):
        total_tips = 0
        total_revenue = 0
        for tips, revenue in tips_per_revenue:
            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__':
    VendorDetails.run()
~
```

COMMAND: python mrtask_e.py input > output/mrtask_e.txt



```
[root@ip-172-30-2-131 MRJobCaseStudy]# python mrtask_e.py input > output/mrtask_e.txt
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/mrtask_e.root.20230902.152307.644084
Running step 1 of 1...
job output is in /tmp/mrtask_e.root.20230902.152307.644084/output
Streaming final output from /tmp/mrtask_e.root.20230902.152307.644084/output...
Removing temp directory /tmp/mrtask_e.root.20230902.152307.644084...
```

OUTPUT

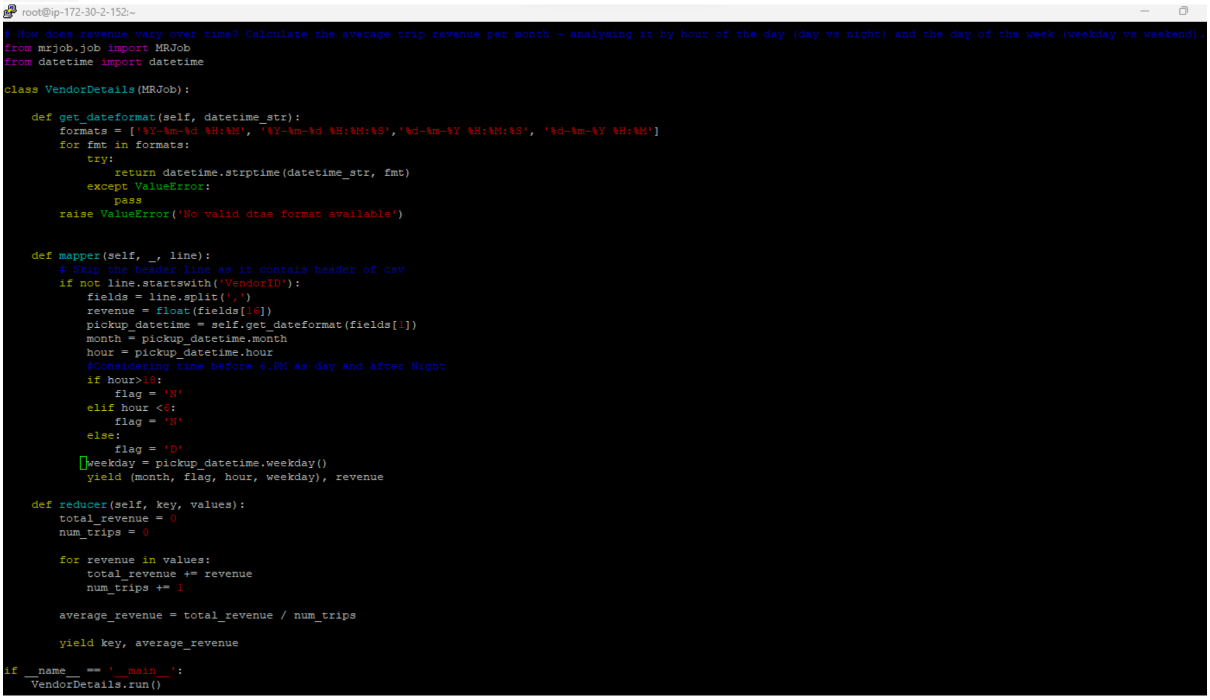


```
[root@ip-172-30-2-131 MRJobCaseStudy]# cat output/mrtask_e.txt
"1"     0.12280054026828338
"10"    0.10341985109036379
"100"   0.09971499320153916
"101"   0.11378933573647326
"102"   0.08878480145766839
"105"   0.07602373196835738
"106"   0.113454199079218 5
"107"   0.11925476770336918
"108"   0.07039173659248449
"109"   0.1896889668334451
"11"    0.058032672461070585
"111"   0.09203346671516915
"112"   0.10885457268865018
"113"   0.11742847360831894
"114"   0.11561549040891052
"115"   0.1024656785738679
"116"   0.0910018572044488
"117"   0.032745177372568116
"118"   0.07963696564772661
"119"   0.068277950196844
"12"    0.08578104516898145
"120"   0.09501251403035434
"121"   0.08957204957768224
"122"   0.09233346348098795
"123"   0.1502962277706154
"124"   0.091235983344314
"125"   0.12256637319450063
"126"   0.05412332768918744
"127"   0.08566028373967267
"128"   0.10274191959006697
"129"   0.06460783060368541
"13"    0.11711242201646355
"130"   0.10321533501152916
"131"   0.08016029304432844
"132"   0.10121684761779591
"133"   0.09049842954763858
"134"   0.09818866107004519
"135"   0.08395162006407768
"136"   0.035218971158968104
"137"   0.11266843961431423
"138"   0.1307794695370675
"139"   0.06507302547620175
"14"    0.09610265720277181
"140"   0.11143832821117793
"141"   0.11172599481825417
"142"   0.1125390342085203
```

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).

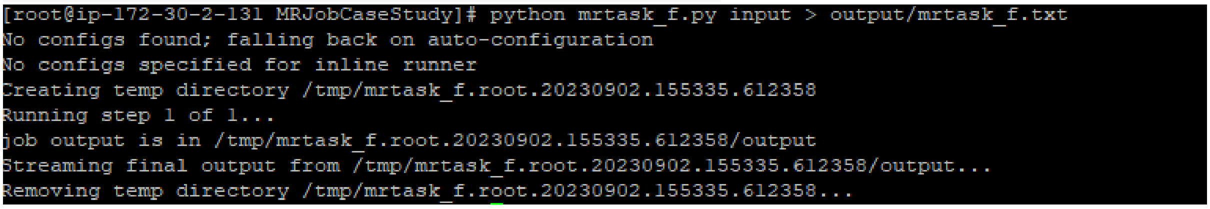
CODE: tpep_pickup_datetime help us with pickup location date but data in this columns is in different date types which we need to convert into proper date format and this has been done by parse_datetime function, columns total_amount halp us with revenue details.
But we need to get Hour, month, and weekday format so mapper will help us to achieve this and final result will further reduce by reducer.

Considering time before 6.PM and after 6.AM as day and after Night
CODE:



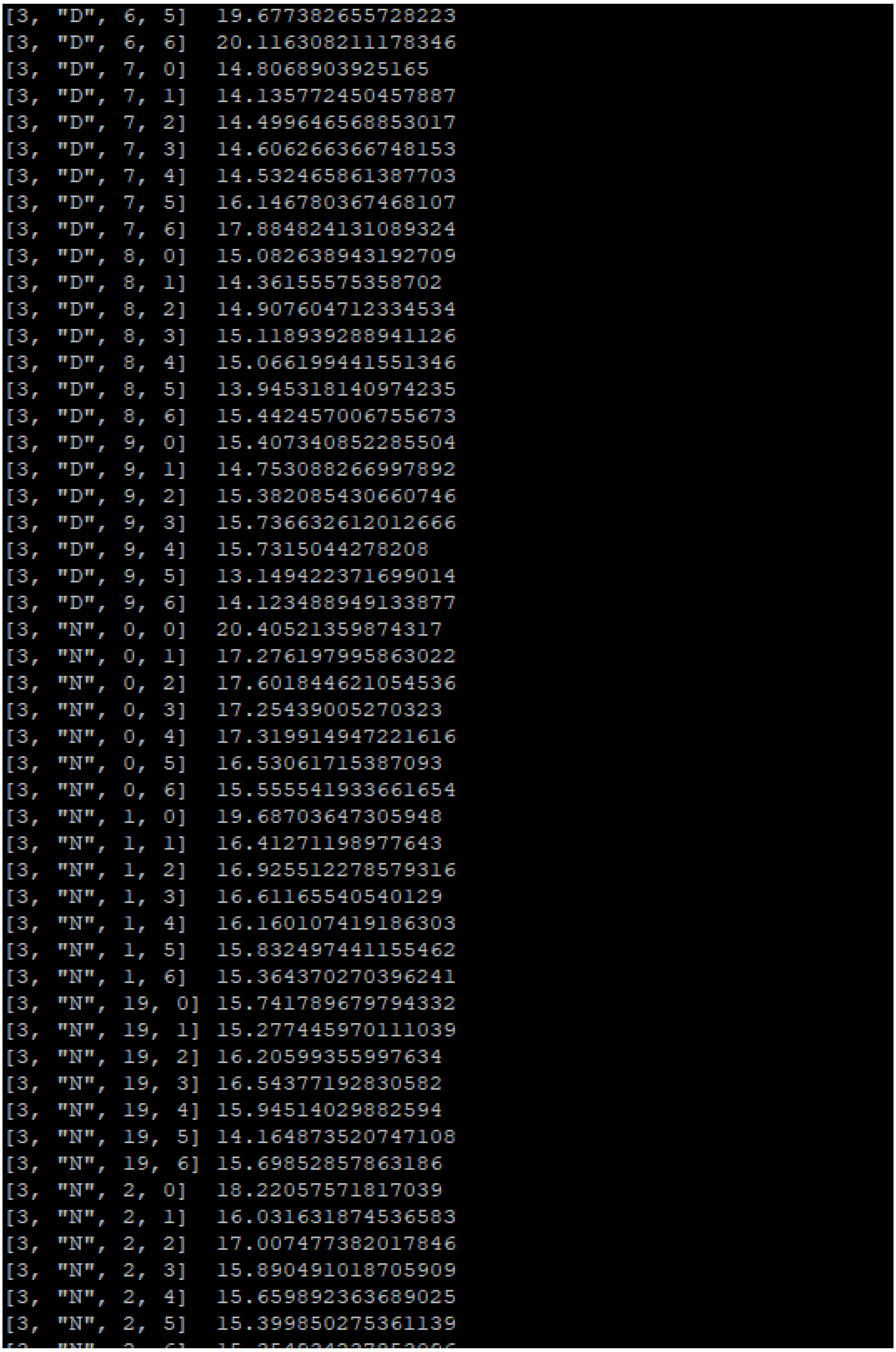
```
# How does revenue vary over time? Calculate the average trip revenue per month - analysing it by hour of the day (day vs night) and the day of the week (weekday vs weekend).
from mrjob.job import MRJob
from datetime import datetime

class VendorDetails(MRJob):

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


    def mapper(self, _, line):
        # Skip the header line as it contais header of csv
        if not line.startswith('VendorID'):
            fields = line.split(',')
            revenue = float(fields[16])
            pickup_datetime = self.get_dateformat(fields[1])
            month = pickup_datetime.month
            hour = pickup_datetime.hour
            #Considering time before 6.PM as day and after Night
            if hour>18:
                flag = 'N'
            elif hour <6:
                flag = 'N'
            else:
                flag = 'D'
            weekday = pickup_datetime.weekday()
            yield (month, flag, 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__':
    VendorDetails.run()
```

COMMAND: python mrtask_f.py input > output/mrtask_f.txt



```
[root@ip-172-30-2-131 MRJobCaseStudy]# python mrtask_f.py input > output/mrtask_f.txt
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/mrtask_f.root.20230902.155335.612358
Running step 1 of 1...
job output is in /tmp/mrtask_f.root.20230902.155335.612358/output
Streaming final output from /tmp/mrtask_f.root.20230902.155335.612358/output...
Removing temp directory /tmp/mrtask_f.root.20230902.155335.612358...
```

OUTPUT:



```
[3, "D", 6, 5]  19.677382655728223
[3, "D", 6, 6]  20.116308211178346
[3, "D", 7, 0]  14.8068903925165
[3, "D", 7, 1]  14.135772450457887
[3, "D", 7, 2]  14.499646568853017
[3, "D", 7, 3]  14.60626636674815З
[3, "D", 7, 4]  14.532465861387703
[3, "D", 7, 5]  16.146780367468107
[3, "D", 7, 6]  17.884824131089324
[3, "D", 8, 0]  15.082638943192709
[3, "D", 8, 1]  14.36155575358702
[3, "D", 8, 2]  14.907604712334534
[3, "D", 8, 3]  15.118939288941126
[3, "D", 8, 4]  15.066199441551346
[3, "D", 8, 5]  13.945318140974235
[3, "D", 8, 6]  15.442457006755673
[3, "D", 9, 0]  15.407340852285504
[3, "D", 9, 1]  14.753088266997892
[3, "D", 9, 2]  15.382085430660746
[3, "D", 9, 3]  15.736632612012666
[3, "D", 9, 4]  15.7315044278208
[3, "D", 9, 5]  13.149422371699014
[3, "D", 9, 6]  14.123488949133877
[3, "N", 0, 0]  20.40521359874317
[3, "N", 0, 1]  17.276197995863022
[3, "N", 0, 2]  17.601844621054536
[3, "N", 0, 3]  17.25439005270323
[3, "N", 0, 4]  17.319914947221616
[3, "N", 0, 5]  16.53061715387093
[3, "N", 0, 6]  15.555541933661654
[3, "N", 1, 0]  19.68703647305948
[3, "N", 1, 1]  16.41271198977643
[3, "N", 1, 2]  16.925512278579316
[3, "N", 1, 3]  16.61165540540129
[3, "N", 1, 4]  16.160107419186303
[3, "N", 1, 5]  15.832497441155462
[3, "N", 1, 6]  15.364370270396241
[3, "N", 19, 0] 15.741789679794332
[3, "N", 19, 1] 15.277445970111039
[3, "N", 19, 2] 16.20599355997634
[3, "N", 19, 3] 16.54377192830582
[3, "N", 19, 4] 15.94514029882594
[3, "N", 19, 5] 14.164873520747108
[3, "N", 19, 6] 15.69852857863186
[3, "N", 2, 0]  18.22057571817039
[3, "N", 2, 1]  16.031631874536583
[3, "N", 2, 2]  17.007477382017846
[3, "N", 2, 3]  15.890491018705909
[3, "N", 2, 4]  15.659892363689025
[3, "N", 2, 5]  15.399850275361139
```