Task 4: Performing Data Analysis with MapReduce on EMR Instance

- To begin, I set up the environment on the EMR (Elastic MapReduce) instance by updating the system using the command 'sudo yum update' and installing the necessary Python packages such as mrjob using 'pip3 install mrjob'.

- Once the environment was configured, I proceeded to write Python scripts for performing MapReduce tasks. After writing the scripts, I transferred them to the EMR instance using WinSCP and granted execution permissions using the command 'chmod 700'.

- For running the MapReduce jobs, I utilized the following general code structure:

```
python mr_script.py input > output_file
```

- This command executed the Python script 'mr_script.py' on the specified input data and redirected the output to the designated output file.

- For this task, I used the 'yellow_tripdata_2017-06.csv' dataset to perform data analysis.

- Regarding the installation of mrjob, you can install it using pip3 with the following command:

```
pip3 install mrjob
```

- This command will install mrjob and its dependencies, allowing you to use it for MapReduce tasks in Python.

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

```
[hadoop@ip-172-31-5-239 ~]$ python mrtask_a.py yellow_tripdata_2017-06.csv > mrtask_a.txt
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/mrtask_a.hadoop.20240304.130023.511525
Running step 1 of 2...
Running step 2 of 2...
job output is in /tmp/mrtask_a.hadoop.20240304.130023.511525/output
Streaming final output from /tmp/mrtask_a.hadoop.20240304.130023.511525/output...
Removing temp directory /tmp/mrtask_a.hadoop.20240304.130023.511525...
[hadoop@ip-172-31-5-239 ~]$ cat mrtask_a.txt
"2"     88663878.19566278
```

Vendor 2 has the most trips,total revenue generated is 88663878.19566278

B: Which pickup location generates the most revenue?

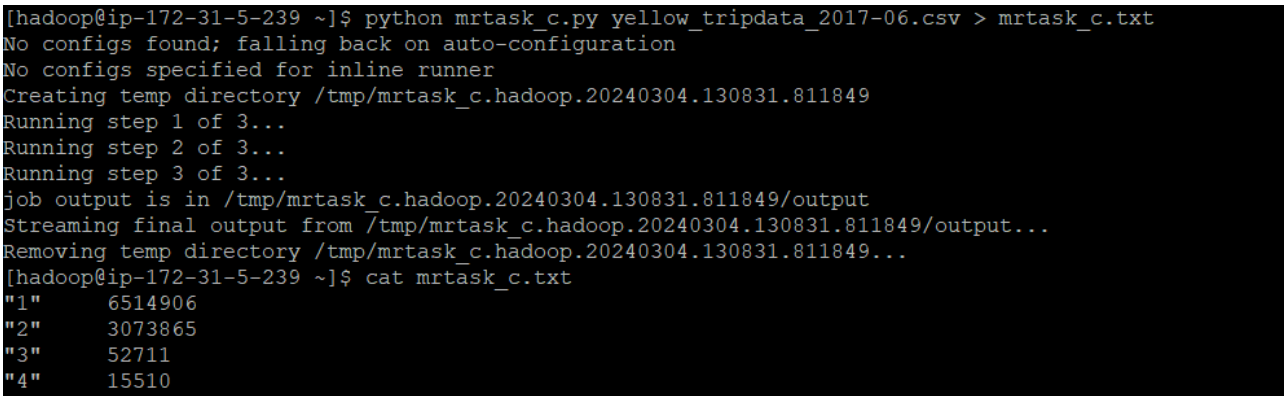
```
[hadoop@ip-172-31-5-239 ~]$ python mrtask_b.py yellow_tripdata_2017-06.csv > mrtask_b.txt
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/mrtask_b.hadoop.20240304.130425.213917
Running step 1 of 2...




Running step 2 of 2...
job output is in /tmp/mrtask_b.hadoop.20240304.130425.213917/output
Streaming final output from /tmp/mrtask_b.hadoop.20240304.130425.213917/output...
Removing temp directory /tmp/mrtask_b.hadoop.20240304.130425.213917...
[hadoop@ip-172-31-5-239 ~]$
[hadoop@ip-172-31-5-239 ~]$
[hadoop@ip-172-31-5-239 ~]$
[hadoop@ip-172-31-5-239 ~]$
[hadoop@ip-172-31-5-239 ~]$ cat mrtask_b.txt
"132"   13022071.03001022
```

Pickup location 132 generates the most revenue : 13022071.03001022

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

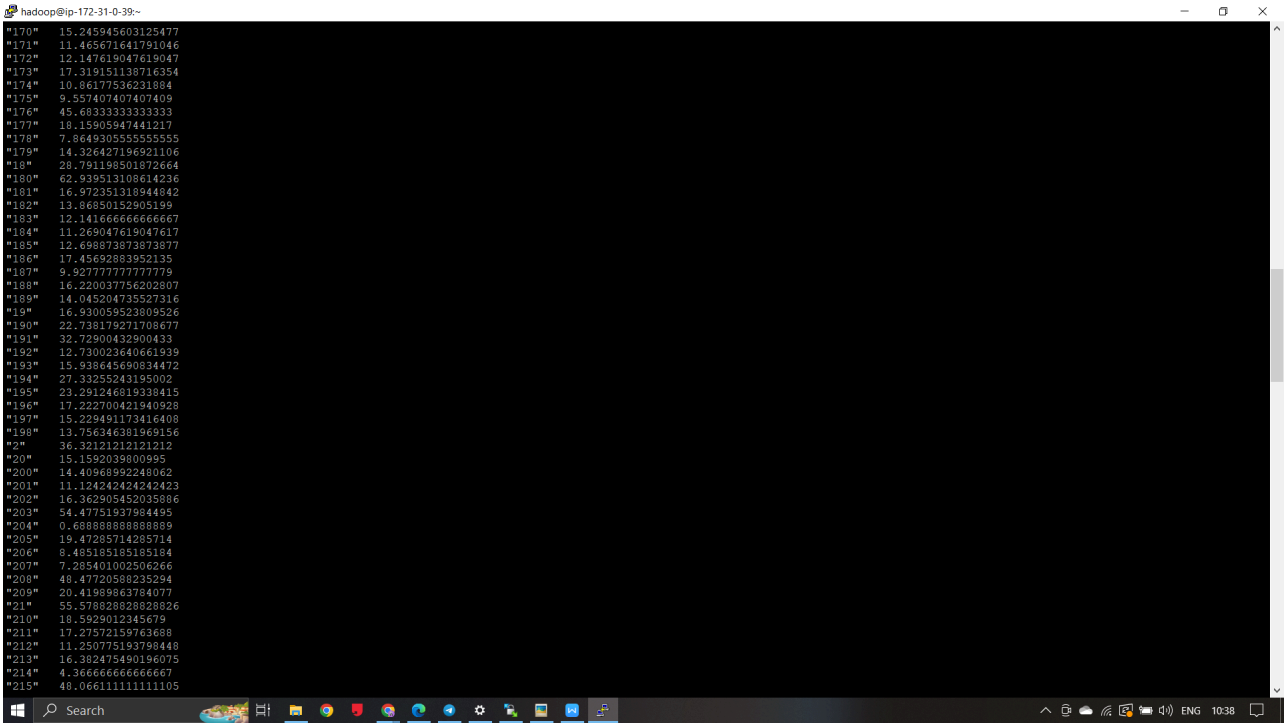
The different payment types used by customers and their count are shown below:

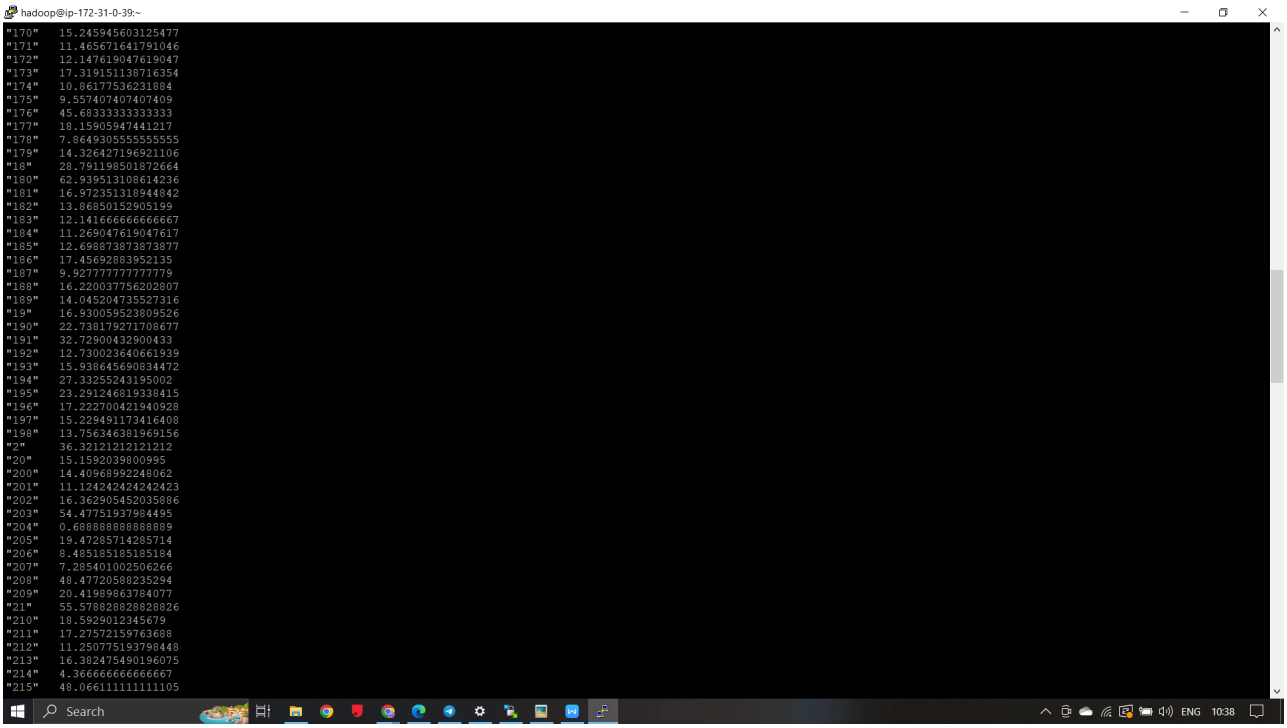
Numeric code	Payment type	Count
1	Credit card	6514906
2	Cash	3073865
3	No charge	52711
4	Dispute	15510



D:What is the average trip time for different pickup locations?

The average trip time for different pickup locations and code used is shown below:
command used : ‘ python mrtask_d.py yellow_tripdata_2017-06.csv > mrtask_d.txt’

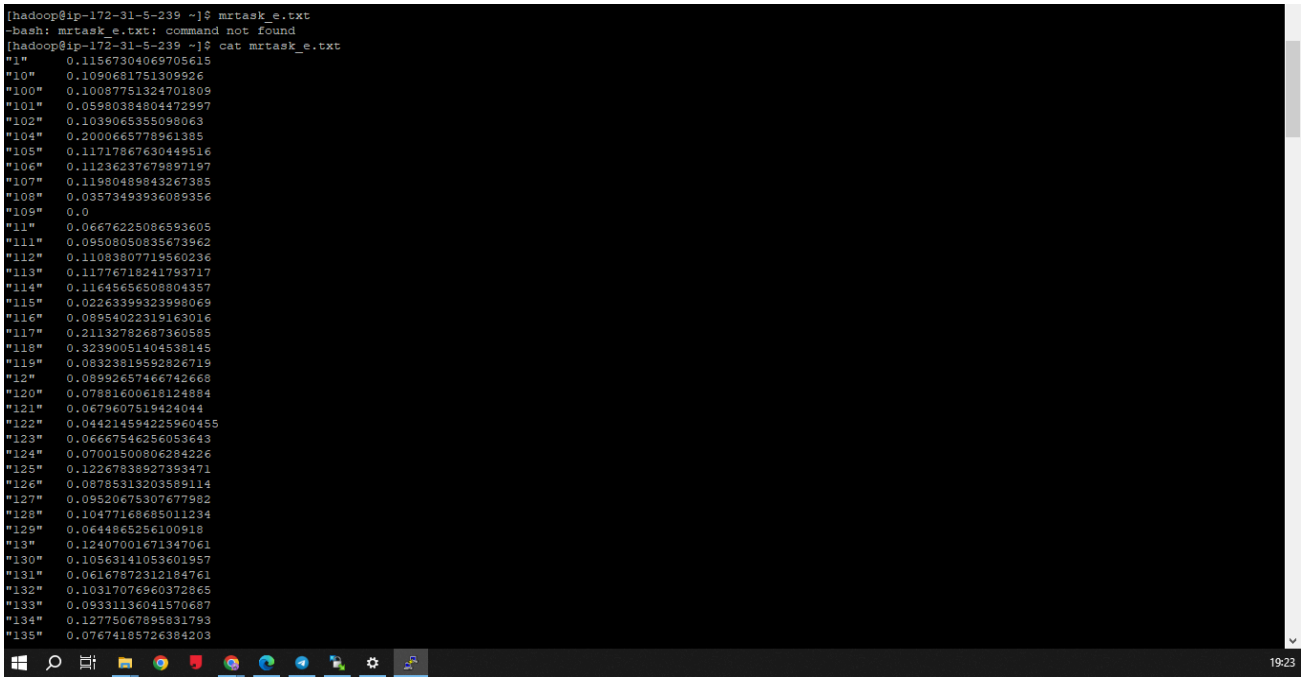

hadoop@ip-172-31-0-39:~
"170" 15.245945603125477
"171" 11.465671641791046
"172" 12.147619047619047
"173" 17.319151138716354
"174" 10.86177536231884
"175" 9.557407407407409
"176" 45.68333333333333
"177" 18.15905947441217
"178" 7.8649305555555555
"179" 14.326427196921106
"18" 28.791198501872664
"180" 62.939513108614236
"181" 16.972351318944842
"182" 13.86850152905199
"183" 12.141666666666667
"184" 11.269047619047617
"185" 12.698873873873877
"186" 17.45692883952135
"187" 9.927777777777779
"188" 16.220037756202807
"189" 14.045204735527316
"19" 16.930059523809526
"190" 22.738179271708677
"191" 32.72900432900433
"192" 12.730023640661939
"193" 15.938645690834472
"194" 27.33255243195002
"195" 23.291246819338415
"196" 17.222700421940928
"197" 15.229491173416408
"198" 13.756346381969156
"2" 36.32121212121212
"20" 15.1592039800995
"200" 14.40968992248062
"201" 11.124242424242423
"202" 16.362905452035886
"203" 54.47751937984495
"204" 0.688888888888889
"205" 19.472857142857142
"206" 8.485185185185184
"207" 7.285401002506266
"208" 48.47720588235294
"209" 20.41989863784077
"21" 55.578828828828826
"210" 18.5929012345679
"211" 17.27572159763688
"212" 11.250775193798448
"213" 16.382475490196075
"214" 4.366666666666667
"215" 48.066111111111105
Search
ENG 10:38


hadoop@ip-172-31-0-39:~
"170" 15.245945603125477
"171" 11.465671641791046
"172" 12.147619047619047
"173" 17.319151138716354
"174" 10.86177536231884
"175" 9.557407407407409
"176" 45.68333333333333
"177" 18.15905947441217
"178" 7.8649305555555555
"179" 14.326427196921106
"18" 28.791198501872664
"180" 62.939513108614236
"181" 16.972351318944842
"182" 13.86850152905199
"183" 12.141666666666667
"184" 11.269047619047617
"185" 12.698873873873877
"186" 17.45692883952135
"187" 9.927777777777779
"188" 16.220037756202807
"189" 14.045204735527316
"19" 16.930059523809526
"190" 22.738179271708677
"191" 32.72900432900433
"192" 12.730023640661939
"193" 15.938645690834472
"194" 27.33255243195002
"195" 23.291246819338415
"196" 17.222700421940928
"197" 15.229491173416408
"198" 13.756346381969156
"2" 36.32121212121212
"20" 15.1592039800995
"200" 14.40968992248062
"201" 11.124242424242423
"202" 16.362905452035886
"203" 54.47751937984495
"204" 0.688888888888889
"205" 19.472857142857142
"206" 8.485185185185184
"207" 7.285401002506266
"208" 48.47720588235294
"209" 20.41989863784077
"21" 55.578828828828826
"210" 18.5929012345679
"211" 17.27572159763688
"212" 11.250775193798448
"213" 16.382475490196075
"214" 4.366666666666667
"215" 48.066111111111105
Search
ENG 10:38

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

The average tips to revenue ratio of the drivers for different pickup locations in sorted format are shown below:

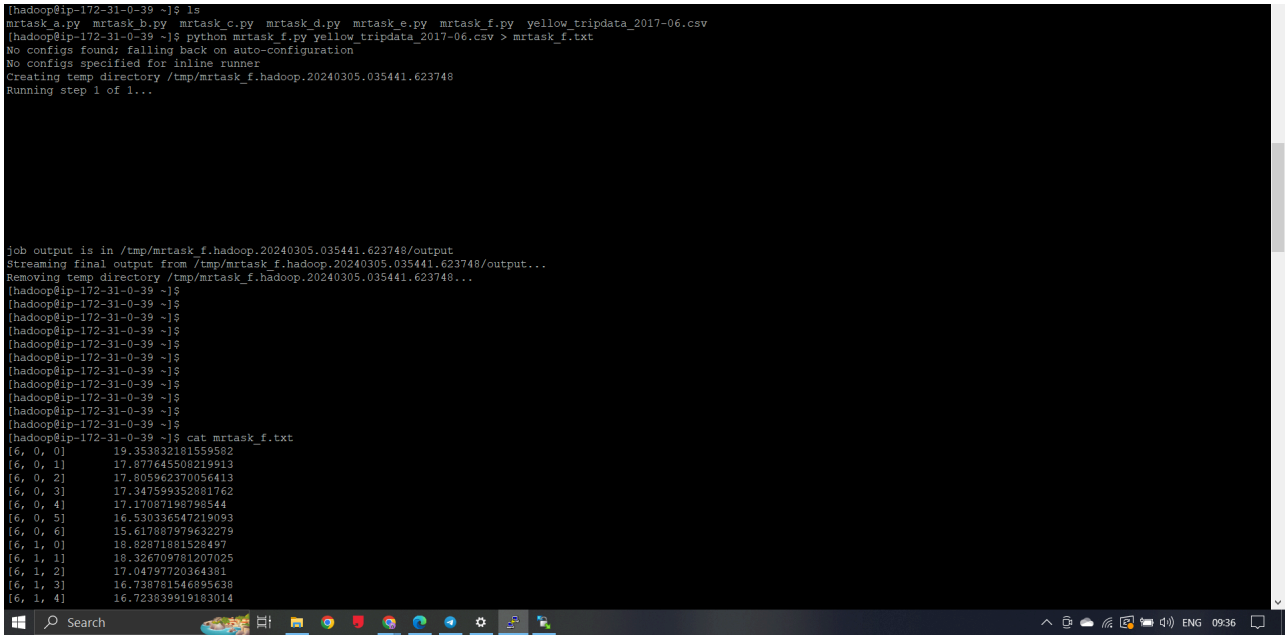
```
[hadoop@ip-172-31-5-239 ~]$ python mrtask_e.py yellow_tripdata_2017-06.csv > mrtask_e.txt
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/mrtask_e.hadoop.20240304.133614.271126
Running step 1 of 1...


job output is in /tmp/mrtask_e.hadoop.20240304.133614.271126/output
Streaming final output from /tmp/mrtask_e.hadoop.20240304.133614.271126/output...
Removing temp directory /tmp/mrtask_e.hadoop.20240304.133614.271126...
```

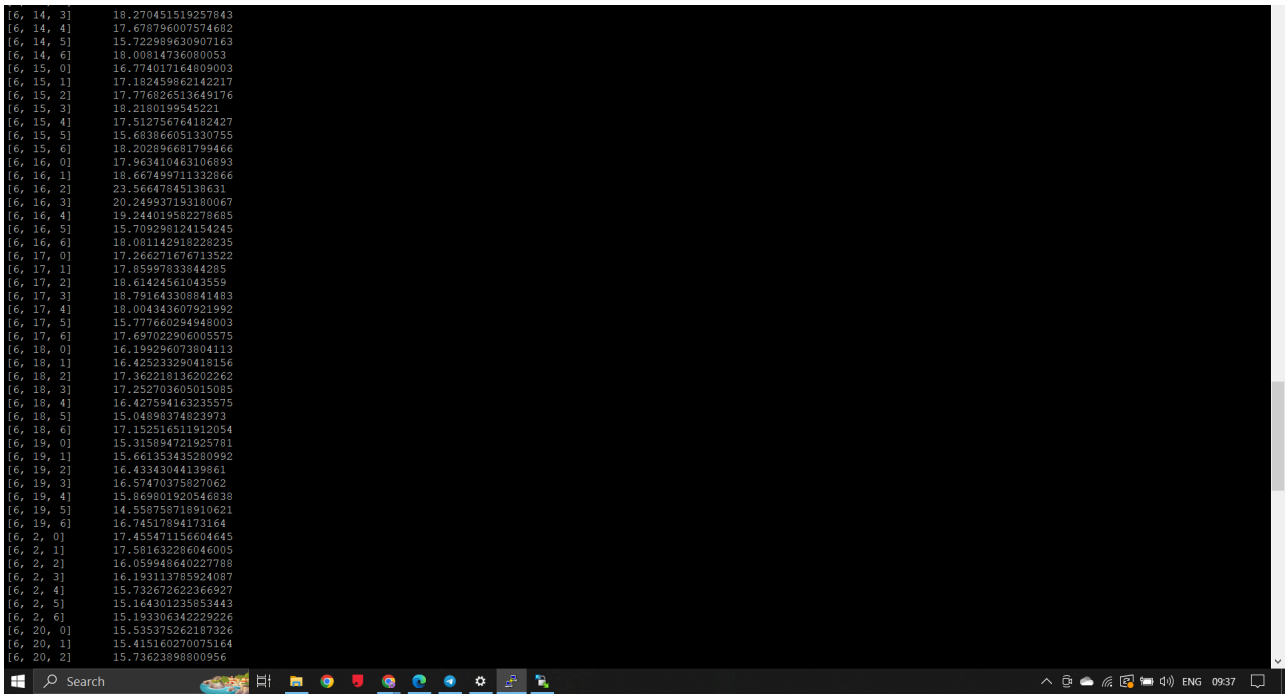


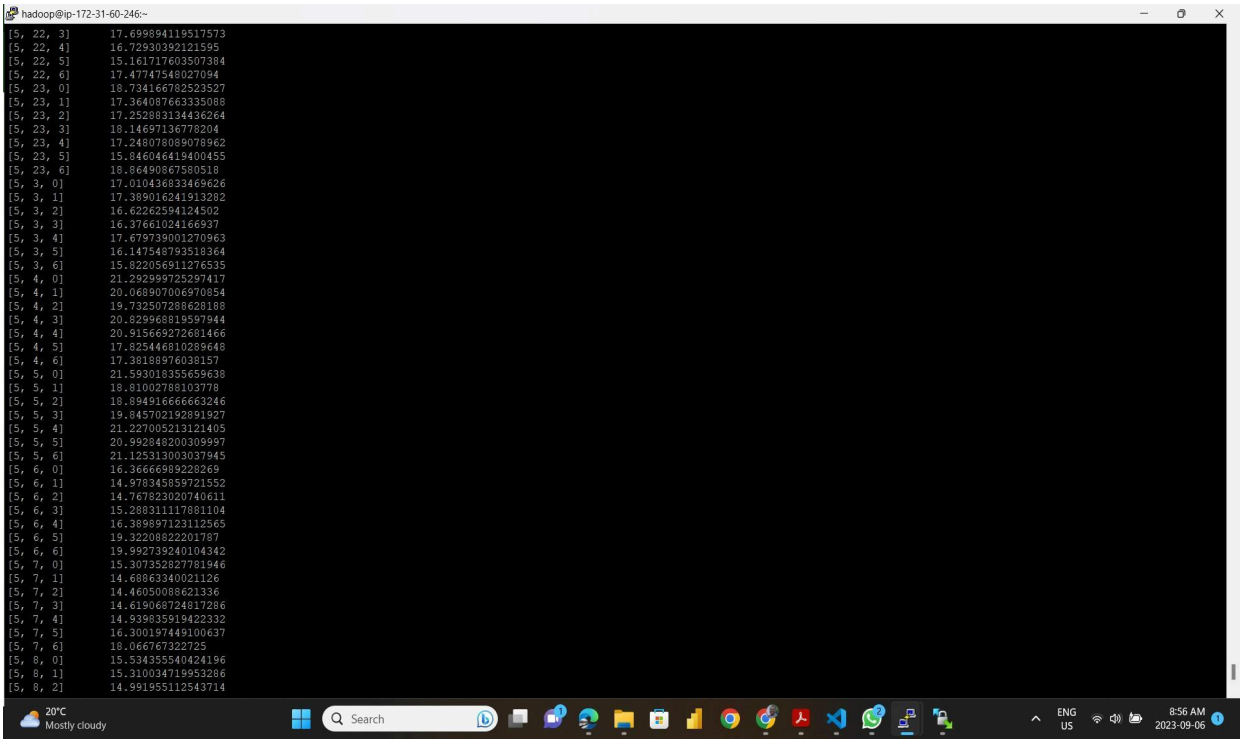
```
"262"   0.11469070998500533
"263"   0.11432379859521637
"264"   0.11273214918900741
"265"   0.11234048545585346
"27"    0.12553011026293465
"28"    0.11415863683577664
"29"    0.05187518100202723
"3"     0.1493922363089792
"30"    0.31531367828101575
"31"    0.13869858704023438
"32"    0.07976710334788936
"33"    0.12161817716977001
"34"    0.11006363779890946
"35"    0.0689372039683509
"36"    0.10223563014317112
"37"    0.10189901103758611
"38"    0.04514910110827603
"39"    0.15559622328692066
"4"     0.10647957295395655
"40"    0.12264788774925717
"41"    0.0903367201863291
"42"    0.07049437931291051
"43"    0.10479945401830335
"44"    0.057178752053603066
"45"    0.09805423323935449
"46"    0.039982864486648576
"47"    0.031771265145925716
"48"    0.10646502304065499
"49"    0.09678172999015307
"50"    0.1086678432721026
"51"    0.1425677131977872
"52"    0.1297757875679241
"53"    0.07457398346549686
"54"    0.11940485964418365
"55"    0.09180704479787068
"56"    0.09863684336077845
"57"    0.10562312517043905
"58"    0.1112571516932117
"59"    0.004178505766337957
"6"     0.07855632869654182
"60"    0.04222717209170866
"61"    0.0850538178631888
"62"    0.076784692781036
"63"    0.08255633940505075
```

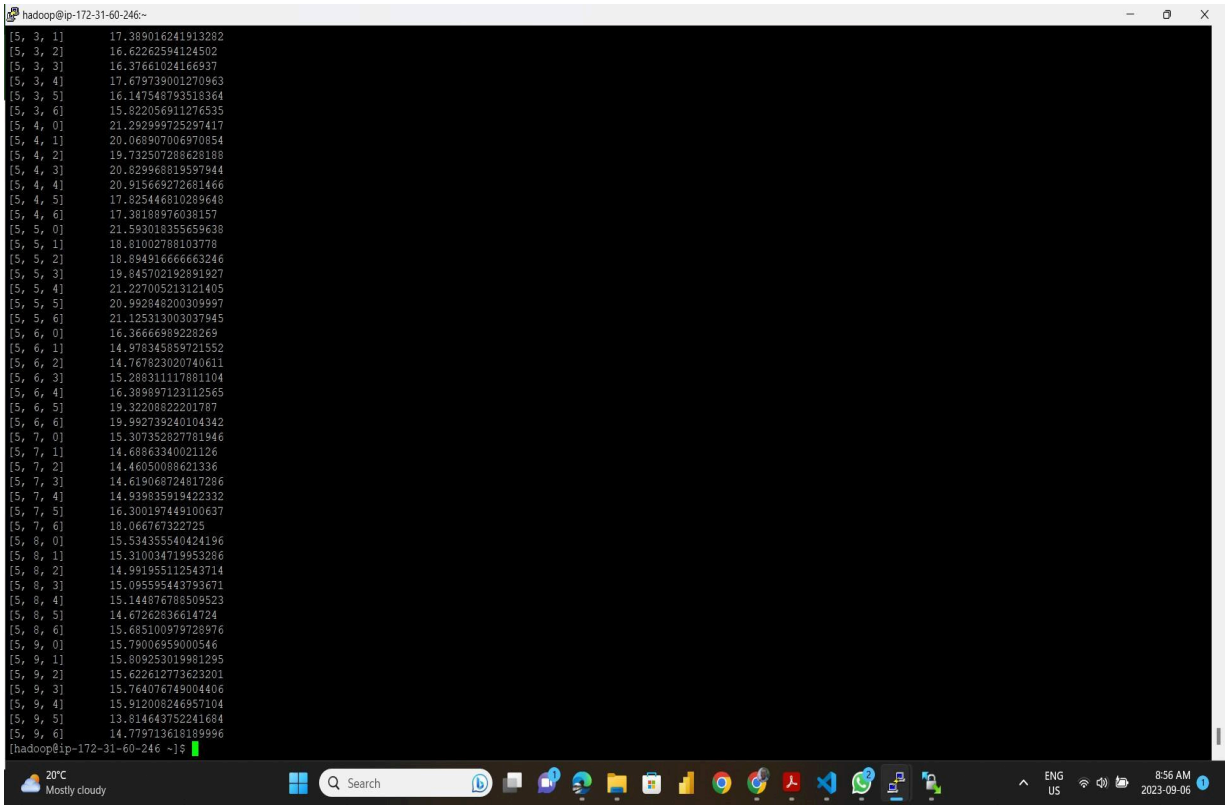
F: How does revenue vary over time? Calculate the average trip revenue per month - analyzing it by hour of the day (day vs night) and the day of the week (weekday vs weekend).

Variation in revenue over time is shown below including code used :






hadoop@ip-172-31-60-246:~
[5, 22, 3] 17.699894119517573
[5, 22, 4] 16.72930392121595
[5, 22, 5] 15.161717603507384
[5, 22, 6] 17.47747548027094
[5, 23, 0] 18.734166782523527
[5, 23, 1] 17.364087663335088
[5, 23, 2] 17.252883134436264
[5, 23, 3] 18.14697136778204
[5, 23, 4] 17.248078089078962
[5, 23, 5] 15.846046419400455
[5, 23, 6] 18.86490867580518
[5, 3, 0] 17.010436833469626
[5, 3, 1] 17.389016241913282
[5, 3, 2] 16.622625941245O2
[5, 3, 3] 16.37661024166937
[5, 3, 4] 17.679739001270963
[5, 3, 5] 16.147548793518364
[5, 3, 6] 15.8220569112765З5
[5, 4, 0] 21.292999725297417
[5, 4, 1] 20.068907006970854
[5, 4, 2] 19.732507288628188
[5, 4, 3] 20.829968819597944
[5, 4, 4] 20.915669272681466
[5, 4, 5] 17.825446810289648
[5, 4, 6] 17.38188976038157
[5, 5, 0] 21.593018355659638
[5, 5, 1] 18.81002788103778
[5, 5, 2] 18.894916666663246
[5, 5, 3] 19.845702192891927
[5, 5, 4] 21.227005213121405
[5, 5, 5] 20.992848200309997
[5, 5, 6] 21.125313003037945
[5, 6, 0] 16.36666989228269
[5, 6, 1] 14.978345859721552
[5, 6, 2] 14.767823020740611
[5, 6, 3] 15.288311117881104
[5, 6, 4] 16.389897123112565
[5, 6, 5] 19.32208822201787
[5, 6, 6] 19.992739240104342
[5, 7, 0] 15.307352827781946
[5, 7, 1] 14.68863340021126
[5, 7, 2] 14.46050088621336
[5, 7, 3] 14.619068724817286
[5, 7, 4] 14.939835919422332
[5, 7, 5] 16.300197449100637
[5, 7, 6] 18.066767322725
[5, 8, 0] 15.534355540424196
[5, 8, 1] 15.310034719953286
[5, 8, 2] 14.991955112543714
[5, 8, 3] 15.095595443793671
[5, 8, 4] 15.144876788509523
[5, 8, 5] 14.672628366147З4
[5, 8, 6] 15.685100979728976
[5, 9, 0] 15.790069590005З6
[5, 9, 1] 15.809253019981295
[5, 9, 2] 15.622612773623201
[5, 9, 3] 15.764076749004406
[5, 9, 4] 15.912008246957104
[5, 9, 5] 13.814643752241684
[5, 9, 6] 14.779713618189996
[hadoop@ip-172-31-60-246 ~]$
20°C
Mostly cloudy
Search
ENG
US
8:56 AM
2023-09-06


hadoop@ip-172-31-60-246:~
[5, 3, 1] 17.389016241913282
[5, 3, 2] 16.622625941245O2
[5, 3, 3] 16.37661024166937
[5, 3, 4] 17.679739001270963
[5, 3, 5] 16.147548793518364
[5, 3, 6] 15.8220569112765З5
[5, 4, 0] 21.292999725297417
[5, 4, 1] 20.068907006970854
[5, 4, 2] 19.732507288628188
[5, 4, 3] 20.829968819597944
[5, 4, 4] 20.915669272681466
[5, 4, 5] 17.825446810289648
[5, 4, 6] 17.38188976038157
[5, 5, 0] 21.593018355659638
[5, 5, 1] 18.81002788103778
[5, 5, 2] 18.894916666663246
[5, 5, 3] 19.845702192891927
[5, 5, 4] 21.227005213121405
[5, 5, 5] 20.992848200309997
[5, 5, 6] 21.125313003037945
[5, 6, 0] 16.36666989228269
[5, 6, 1] 14.978345859721552
[5, 6, 2] 14.767823020740611
[5, 6, 3] 15.288311117881104
[5, 6, 4] 16.389897123112565
[5, 6, 5] 19.32208822201787
[5, 6, 6] 19.992739240104342
[5, 7, 0] 15.307352827781946
[5, 7, 1] 14.68863340021126
[5, 7, 2] 14.46050088621336
[5, 7, 3] 14.619068724817286
[5, 7, 4] 14.939835919422332
[5, 7, 5] 16.300197449100637
[5, 7, 6] 18.066767322725
[5, 8, 0] 15.534355540424196
[5, 8, 1] 15.310034719953286
[5, 8, 2] 14.991955112543714
[5, 8, 3] 15.095595443793671
[5, 8, 4] 15.144876788509523
[5, 8, 5] 14.672628366147З4
[5, 8, 6] 15.685100979728976
[5, 9, 0] 15.790069590005З6
[5, 9, 1] 15.809253019981295
[5, 9, 2] 15.622612773623201
[5, 9, 3] 15.764076749004406
[5, 9, 4] 15.912008246957104
[5, 9, 5] 13.814643752241684
[5, 9, 6] 14.779713618189996
[hadoop@ip-172-31-60-246 ~]$
20°C
Mostly cloudy
Search
ENG
US
8:56 AM
2023-09-06