## **MapReduce Tasks:**

**Task 4.** Using MapReduce to perform data analysis on files downloaded to the EMR instance

- Initially I set up the environment using commands sudo yum update, pip3 install mrjob etc,onces set up I proceeded
- I wrote python scripts, moved them to hadoop using WinSCP, gave python scripts execution permission using command chmod 700
- General code used for running the MapReduce jobs: Python 'mr\_script.py' 'input' > 'output file'
  - ❖ I used <u>vellow\_tripdata\_2017-05.csv</u> dataset for this task
  - **a.** Which vendors have the most trips, and what is the total revenue generated by that vendor?

Vendor 2 has the most trips, total revenue generated is 92896777.5452

```
[hadoop@ip-172-31-60-246 ~]$ python mrtask a.py yellow_tripdata_2017-05.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.20230906.144349.570870

Running step 1 of 2...

Running step 2 of 2...

glob output is in /tmm/mrtask_a.hadoop.20230906.144349.570870/output

Streaming final output from /tmp/mrtask_a.hadoop.20230906.144349.570870/output...

Removing temp directory /tmp/mrtask_a.hadoop.20230906.144349.570870...

[hadoop@ip-172-31-60-246 ~]$ ls

mrtask_a.py mrtask_a.txt mrtask_b.py mrtask_c.py mrtask_c.py mrtask_c.py mrtask_f.py yellow_tripdata_2017-05.csv

[hadoop@ip-172-31-60-246 ~]$ at mrtask_a.txt

"2" 92896777.5522054

[hadoop@ip-172-31-60-246 ~]$ at mrtask_a.txt

"2" 92896777.51522054

[hadoop@ip-172-31-60-246 ~]$ at mrtask_a.txt

"2" 92896777.5122054

[hadoop@ip-172-31-60-246 ~]$ at mrtask_a.txt
```

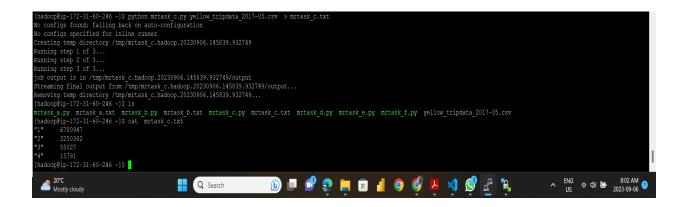
**b.** Which pickup location generates the most revenue?

Pickup location 132 generates the most revenue : 14040591.2200

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

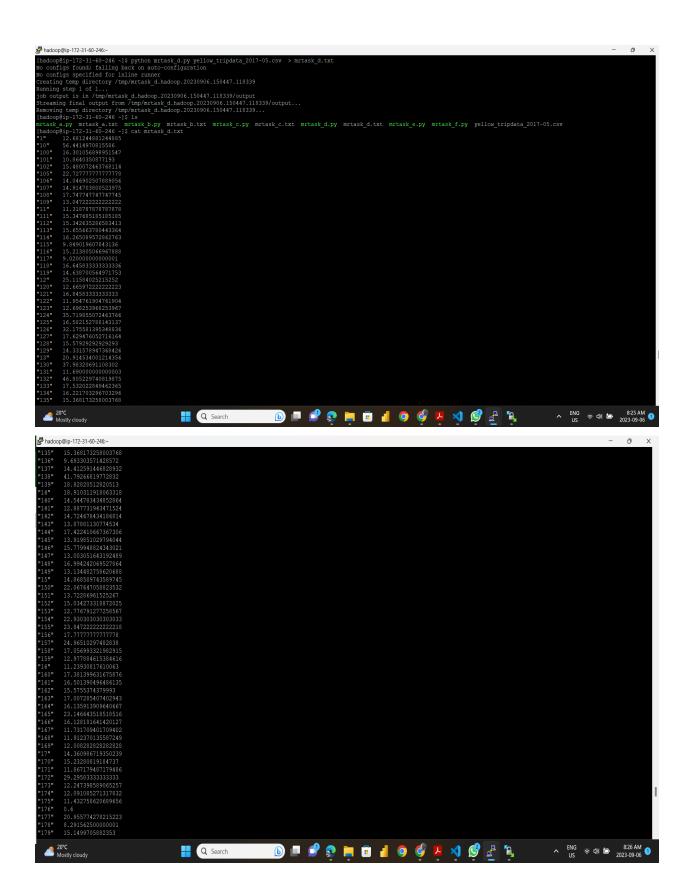
The different payment types used by customers and their count are shown below:

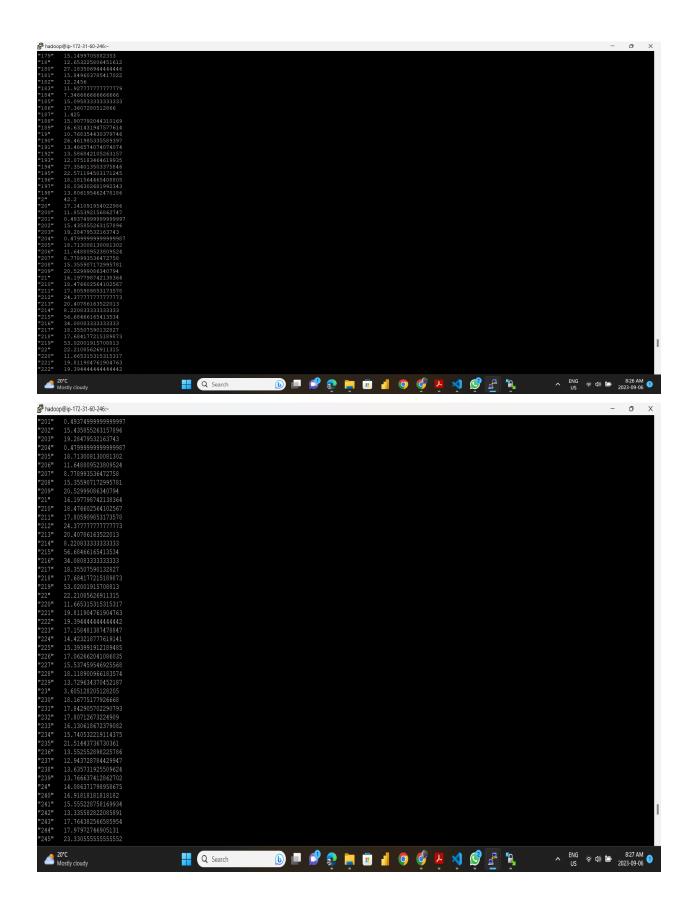
| Numeric code | Payment type | Count   |
|--------------|--------------|---------|
| 1            | Credit card  | 6780947 |
| 2            | Cash         | 3250362 |
| 3            | No charge    | 55027   |
| 4            | Dispute      | 15791   |

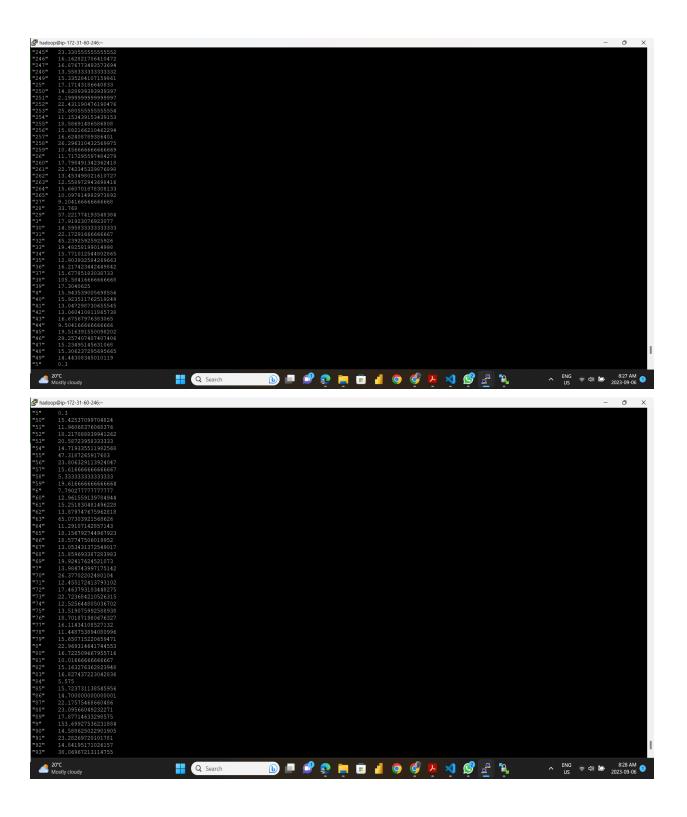


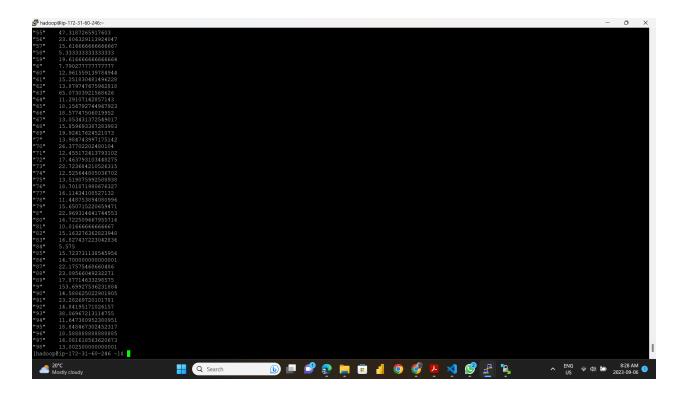
d. What is the average trip time for different pickup locations?

The average trip time for different pickup locations and code used is shown below:



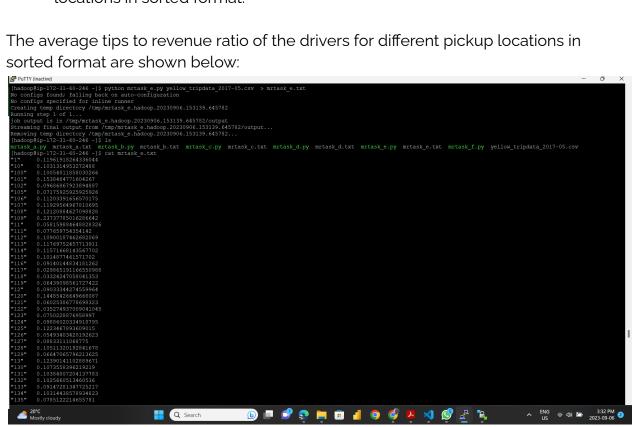


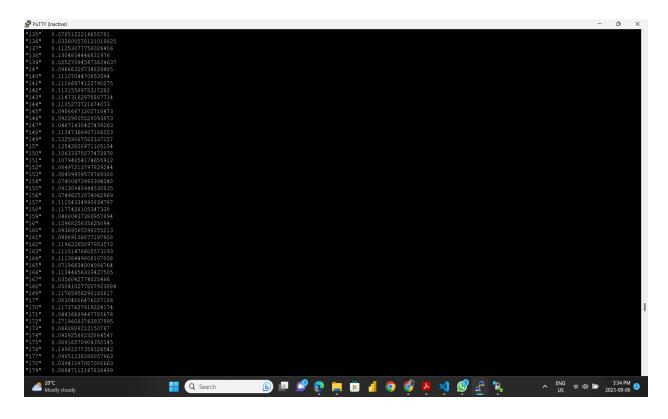


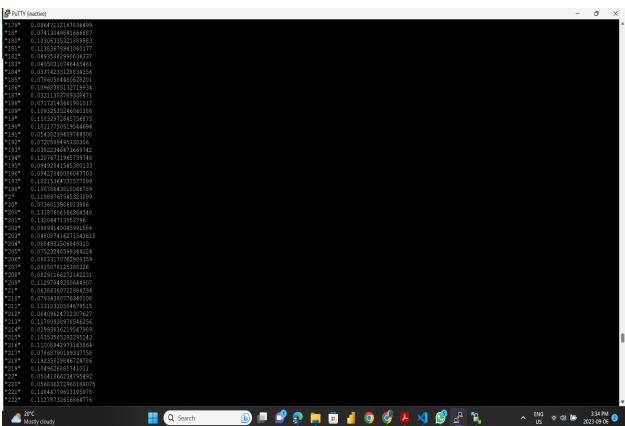


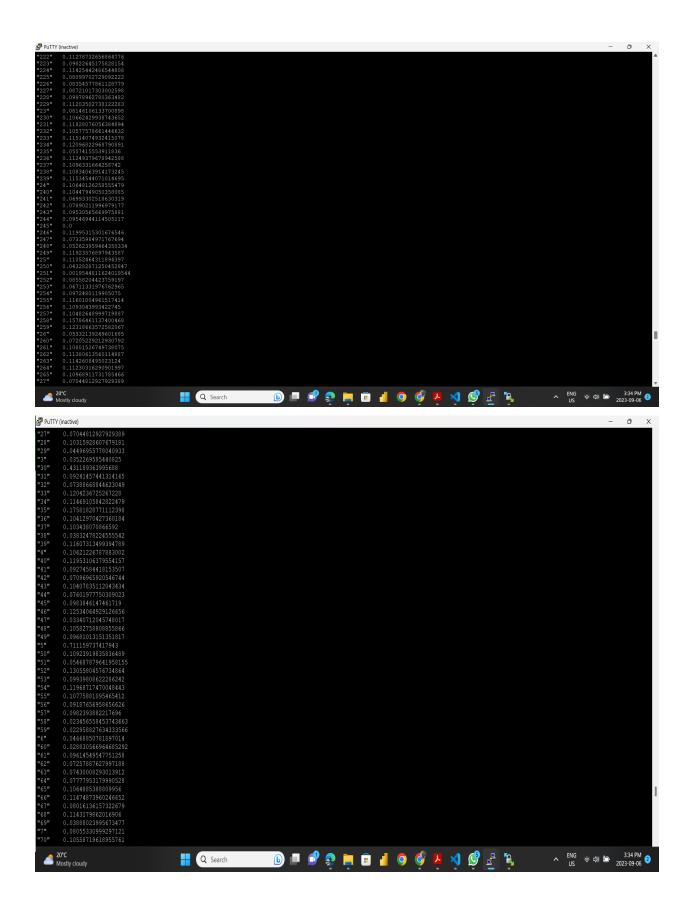
e. Calculate the average tips to revenue ratio of the drivers for different pickup locations in sorted format.

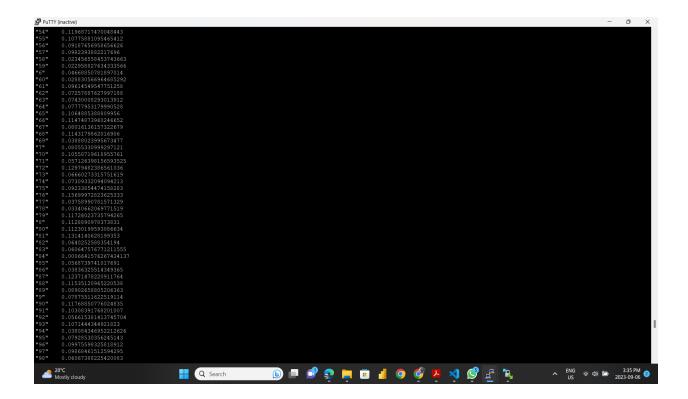
The average tips to revenue ratio of the drivers for different pickup locations in sorted format are shown below:











**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:

