- pyspark-tutorial-gcloud

pyspark-tutorial-gcloud

Output from the script

```
(dsp)lin.yang@lin trip_analysis % python main.py
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
23/08/05 11:47:50 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java c
lasses where applicable
23/08/05 11:47:50 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
23/08/05 11:47:50 WARN Utils: Service 'SparkUI' could not bind on port 4041. Attempting port 4042.
+-----------+-----------------+------------+------------------+------------------+
|pickup_hour|pickup_day_of_week|pickup_month|      avg_duration|      avg_distance|
+-----------+-----------------+------------+------------------+------------------+
|          0|                1|           1| 888.9417608770127|3.2521391261294736|
|          0|                1|           2| 726.9726522187823|  2.89541996285979|
|          0|                1|           3| 888.4640534063677|2.8471578638497625|
|          0|                1|           4| 854.6067429406037|2.9782131248083417|
|          0|                1|           5| 883.9497659700705|2.8323654661521616|
|          0|                1|           6| 954.6639178045153| 2.721734596752614|
|          0|                1|           7|1011.6342042755344| 2.835302943232113|
|          0|                1|           8| 997.7019489609131|2.8669667749237213|
|          0|                1|           9|1057.5752172184677|2.5259332031492243|
|          0|                1|          10| 983.6769738118331|2.4983152044322683|
|          0|                1|          11| 935.9346515215258| 2.539192359159652|
|          0|                1|          12| 972.2170118615943| 3.150238748796069|
|          0|                2|           1| 915.7601380500431| 4.946866130558188|
|          0|                2|           2|1105.8451242829829| 4.494771421675113|
|          0|                2|           3|  820.387349953832| 5.216901033850259|
|          0|                2|           4| 960.7860179499291| 5.048313027017737|
|          0|                2|           5| 951.8014098690836| 4.508350586080599|
|          0|                2|           6| 971.4392452830189| 4.867711624592842|
|          0|                2|           7| 990.5213815789474|4.3910269512339335|
|          0|                2|           8|1029.8973904639174| 5.536793713681965|
+-----------+-----------------+------------+------------------+------------------+
only showing top 20 rows

Top 10 Pickup Locations:
+------------+-------+
|PULocationID|  count|
+------------+-------+
|         237|1553554|
|         236|1424614|
|         161|1091329|
|         132|1025063|
|         186|1019650|
|         142| 989927|
|         170| 967766|
|         162| 954917|
|         239| 932473|
|         141| 909845|
+------------+-------+
```

Top 10 Dropoff Locations:

DOLocationID	count
236	1434919
237	1356518
161	1001077
170	920433
141	902052
239	886837
142	854324
48	782803
238	779046
162	772823

Tip Analysis by Location:

PULocationID	DOLocationID	avg_tip_percentage	avg_distance
187	251	56.179775280898866	1.54
176	176	53.90625	0.32
96	236	48.85197850512946	11.31
109	172	46.948356807511736	2.09
251	161	46.200737170399776	19.58
120	151	43.01075268817204	0.73
118	214	41.3564929693962	4.16
172	214	39.96670910603205	3.293333333333333
208	114	38.55192080359299	8.975
34	1	38.41764929631039	15.11
82	253	38.28483920367534	2.48
112	214	37.67972235994051	16.03
96	177	37.522401433691755	4.699999999999999
98	253	36.231884057971016	5.0
34	236	36.050057963748074	8.965
214	172	35.85657881901123	4.154999999999999
118	172	34.72222222222222	3.47
175	28	34.692603978968656	4.966666666666666
98	191	34.63261668452944	2.286
202	191	34.56221198156682	16.99

only showing top 20 rows

Tip Analysis by Time:

pickup_hour	pickup_day_of_week	pickup_month	avg_tip_percentage	total_tip_amount
0	1	1	10.509465545793379	6285.209999999994
0	1	2	10.960206819588734	8032.64000000002
0	1	3	11.105551868685199	12657.859999999993
0	1	4	11.314341113424645	18623.729999999974
0	1	5	12.117369325323516	36288.720000000074
0	1	6	12.19923946046261	36665.33000000003
0	1	7	11.893981566402744	34842.08999999987
0	1	8	12.147081300758224	47726.230000000294
0	1	9	12.178405026405924	43128.000000000204
0	1	10	12.615973213402029	64931.850000001265
0	1	11	12.530480004739998	52851.220000000736
0	1	12	12.381016811103633	40848.40999999994
0	2	1	9.083634613534988	2937.700000000002
0	2	2	9.046358567573382	3625.010000000003
0	2	3	8.912576772587578	5795.150000000005
0	2	4	9.551659729865403	6010.509999999994
0	2	5	10.373283655908706	13920.53
0	2	6	10.56761617803041	16237.30999999995
0	2	7	10.498185701824484	17556.769999999986
0	2	8	10.404594988408133	20802.27999999999

only showing top 20 rows

Payment and Tip Analysis:

payment_type	avg_tip_percentage	avg_tip_amount	total_tip_amount
0	8.070807011379319	2.1700068168216093	3208778.2300000293
5	0.0	0.0	0.0
1	15.331362787552713	3.0755510306721945	6.913956203000465E7
3	0.145542231109957	-0.01167058060330...	-1799.0200000000004
2	-0.00155152365292...	4.108590704647675E-4	2740.4299999999994
4	-0.3495420614922692	0.022958282745690756	2779.72

Average Fare by Pickup & Drop Location:

PULocationID	DOLocationID	avg_fare
154	28	1164.0
234	189	843.4665424430641
1	247	420.0
83	136	378.5
5	74	306.0
54	265	275.5
29	264	213.75227272727273
2	265	200.25
6	265	192.25
123	265	177.35
235	115	170.0
221	265	160.0
253	208	160.0
112	109	155.0
204	265	152.375
44	138	151.5
118	265	151.25
55	1	150.0
10	1	148.21428571428572
221	1	148.0

only showing top 20 rows

Average Fare by Passenger Count:

passenger_count	avg_fare
null	25.50271663190539
0.0	12.251760226150822
1.0	12.709557736571188
2.0	13.77639929394148
3.0	13.555663818461737
4.0	14.284687986716264
5.0	12.664006463835 93
6.0	12.75109443706296
7.0	52.91679487179488

```
|             8.0| 49.14408163265307|
|             9.0|             61.35|
|            96.0|              11.5|
|           112.0|               9.0|
+----------------+------------------+

Correlation between Fare Amount and Trip Distance:
Correlation coefficient: 0.0008730862657094112
```