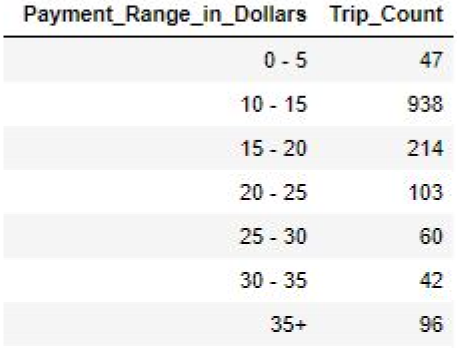
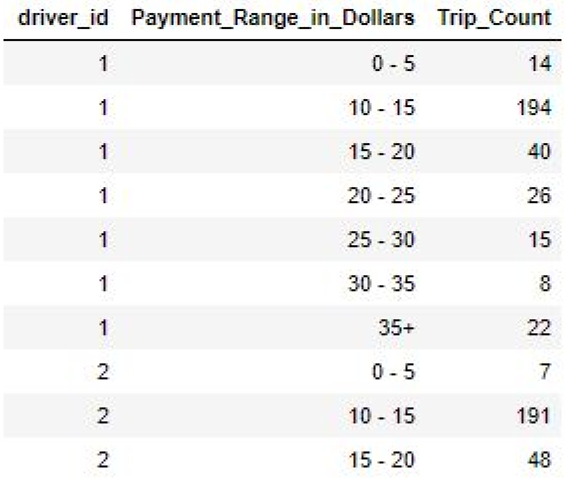
Use pickup information instead of drop-off, when necessary:

1. Most expensive trip (total amount).
2. Most expensive trip per mile (total amount/mile).
3. Most generous trip (highest tip).
4. Longest trip duration.
5. Mean tip by hour.
6. Median trip cost (This question is optional. You can search for “median” calculation if you want).
7. Average total trip by day of week (Fortunately, we have day of week information. Otherwise, we need to create a new date column without hours from date column. Then, we need to create "day of week" column, i.e Monday, Tuesday .. or 1, 2 .., from that new date column. Total trip count should be found for each day, lastly average total trip should be calculated for each day).
8. Count of trips by hour (Luckily, we have hour column. Otherwise, a new hour column should be created from date column, then count trips by hour).
9. Average passenger count per trip.
10. Average passenger count per trip by hour.
11. Which airport welcomes more passengers: JFK or EWR? Tip: check RateCodeID from data dictionary for the definition (2: JFK, 3: Newark).
12. How many nulls are there in Total_amount?
13. How many values are there in Trip_distance? (count of non-missing values)
14. How many nulls are there in Ehail_fee?
15. Find the trips of which trip distance is greater than 15 miles (included) or less than 0.1 mile (included). It is possible to write this with only one where statement. However, this time write two queries and "union" them. The purpose of this question is to use union function. You can consider this question as finding outliers in a quick and dirty way, which you would do in your professional life too often.
16. We would like to see the distribution (not like histogram) of Total_amount. Could you create buckets, or price range, for Total_amount and find how many trips there are in each buckets? Each range would be 5, until 35, i.e. 0-5, 5-10, 10-15 … 30-35, +35. The expected output would be as follows:



Payment_Range_in_Dollars	Trip_Count
0 - 5	47
10 - 15	938
15 - 20	214
20 - 25	103
25 - 30	60
30 - 35	42
35+	96

17. We also would like to analyze the performance of each driver's earning. Could you add driver_id to payment distribution table? The expected output would be as follows:



driver_id	Payment_Range_in_Dollars	Trip_Count
1	0 - 5	14
1	10 - 15	194
1	15 - 20	40
1	20 - 25	26
1	25 - 30	15
1	30 - 35	8
1	35+	22
2	0 - 5	7
2	10 - 15	191
2	15 - 20	48

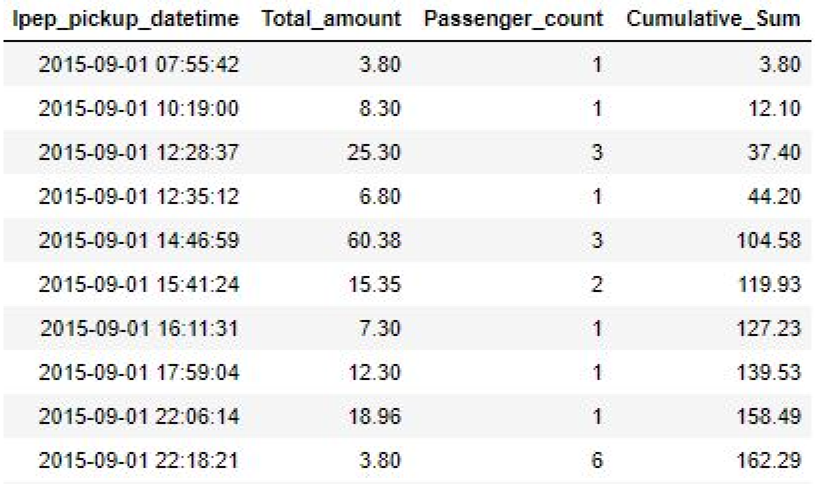
Note that there are more rows in this table.

18. Could you find the highest 3 Total_amount trips for each driver? Hint: Use "Window" functions.

19. Could you find the lowest 3 Total_amount trips for each driver? Hint: Use "Window" functions.

20. Could you find the lowest 10 Total_amount trips for driver_id 1? Do you see any anomaly in the rank? (same rank, missing rank etc). Could you "fix" that so that ranking would be 1, 2, 3, 4… (without any missing rank)? Note that 1 is the lowest Total_amount in this question. Also, same ranks would continue to exist since there might be the same Total_amount. Hint: dense_rank.

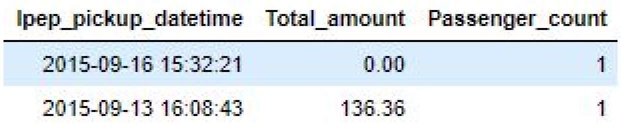
21. Our friend, driver_id 1, is very happy to see what we have done for her (Yes, it is "her". Her name is Gertrude Jeannette, https://en.wikipedia.org/wiki/Gertrude_Jeannette. That is why her id is 1). Could you do her a favor and track her earning after each trip? She would be very thankful if we can provide her with the following information:



Ipep_pickup_datetime	Total_amount	Passenger_count	Cumulative_Sum
2015-09-01 07:55:42	3.80	1	3.80
2015-09-01 10:19:00	8.30	1	12.10
2015-09-01 12:28:37	25.30	3	37.40
2015-09-01 12:35:12	6.80	1	44.20
2015-09-01 14:46:59	60.38	3	104.58
2015-09-01 15:41:24	15.35	2	119.93
2015-09-01 16:11:31	7.30	1	127.23
2015-09-01 17:59:04	12.30	1	139.53
2015-09-01 22:06:14	18.96	1	158.49
2015-09-01 22:18:21	3.80	6	162.29

Hint: Cumulative sum, running total

22. Gertrude is fascinated by your work and would like you to find max and min Total_amount. She is ok with the following:

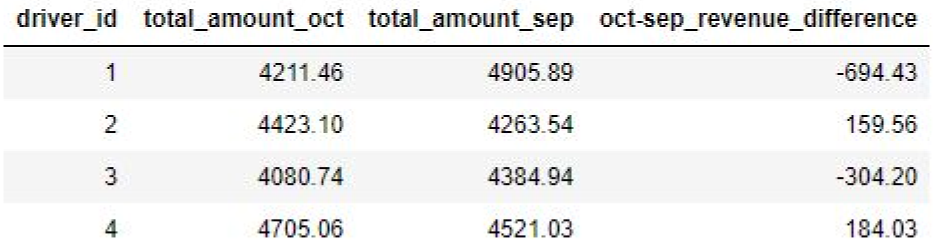


Ipep_pickup_datetime	Total_amount	Passenger_count
2015-09-16 15:32:21	0.00	1
2015-09-13 16:08:43	136.36	1

23. There is one thing that Gertrude could not understand. Min Total_amount is 0, however we did not show any 0 while we track her earning (in cumulative sum question). It seems we owe her an explanation. Why do you think this happened?

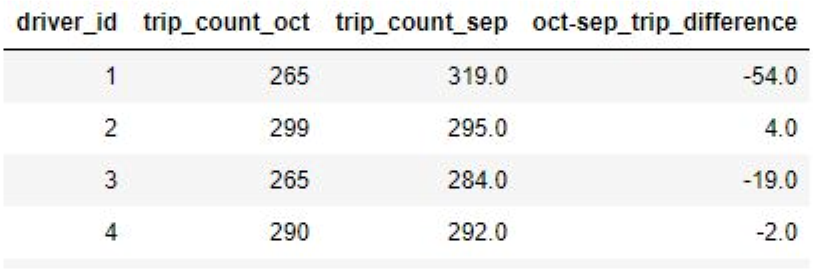
Use October data for Q24 - Q31:

24. Is there any new driver in October? Hint: Drivers existing in one table but not in another table.
25. Total amount difference between October and September.
26. Revenue of drivers each month.



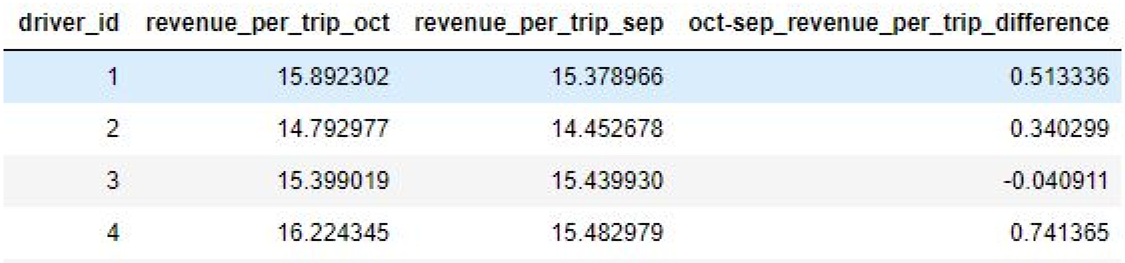
driver_id	total_amount_oct	total_amount_sep	oct-sep_revenue_difference
1	4211.46	4905.89	-694.43
2	4423.10	4263.54	159.56
3	4080.74	4384.94	-304.20
4	4705.06	4521.03	184.03

27. Trip count of drivers each month.



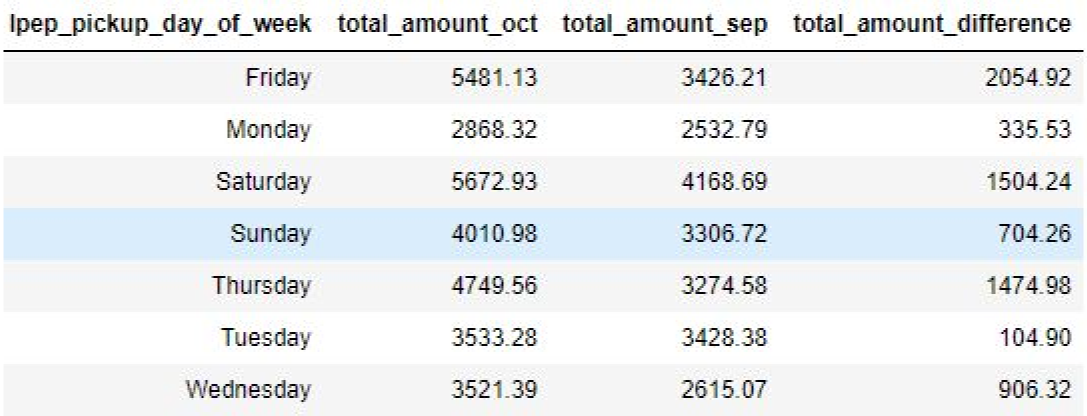
driver_id	trip_count_oct	trip_count_sep	oct-sep_trip_difference
1	265	319.0	-54.0
2	299	295.0	4.0
3	265	284.0	-19.0
4	290	292.0	-2.0

28. Revenue_per-trip of drivers each month.



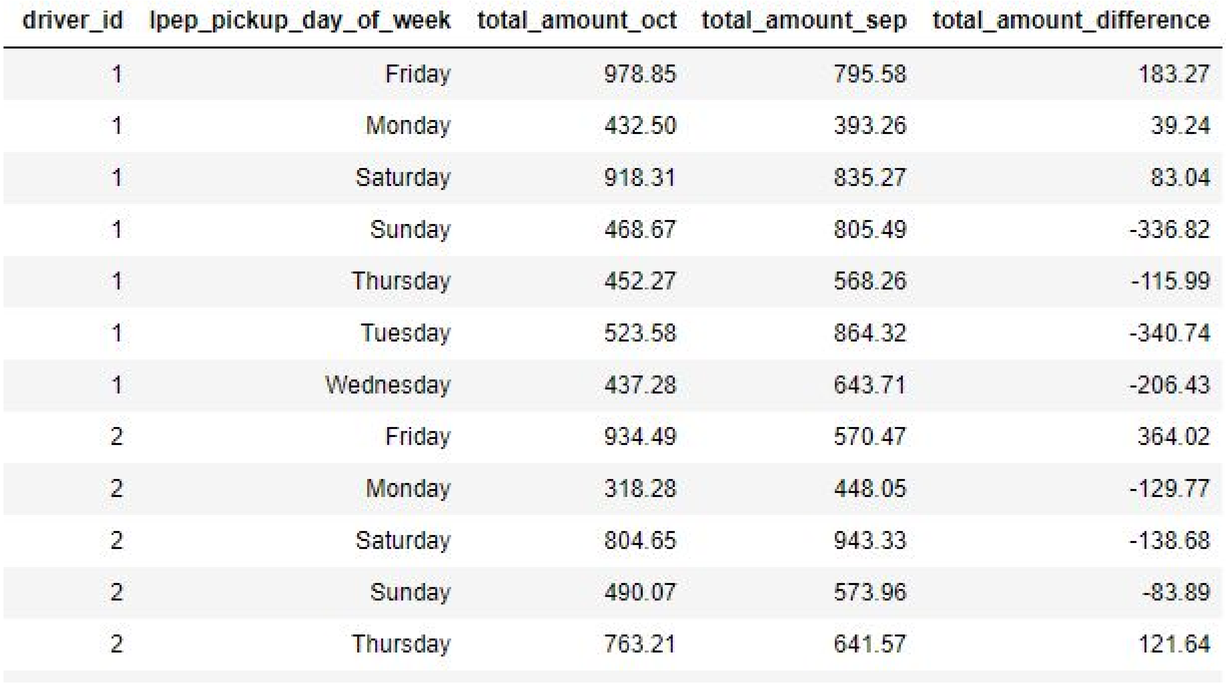
driver_id	revenue_per_trip_oct	revenue_per_trip_sep	oct-sep_revenue_per_trip_difference
1	15.892302	15.378966	0.513336
2	14.792977	14.452678	0.340299
3	15.399019	15.439930	-0.040911
4	16.224345	15.482979	0.741365

29. Revenue per day of week comparison.



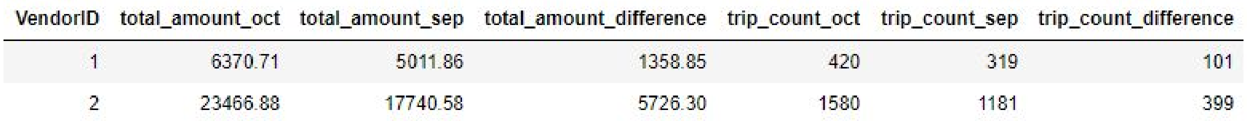
lpep_pickup_day_of_week	total_amount_oct	total_amount_sep	total_amount_difference
Friday	5481.13	3426.21	2054.92
Monday	2868.32	2532.79	335.53
Saturday	5672.93	4168.69	1504.24
Sunday	4010.98	3306.72	704.26
Thursday	4749.56	3274.58	1474.98
Tuesday	3533.28	3428.38	104.90
Wednesday	3521.39	2615.07	906.32

30. Revenue per day of week for each driver comparison



driver_id	lpep_pickup_day_of_week	total_amount_oct	total_amount_sep	total_amount_difference
1	Friday	978.85	795.58	183.27
1	Monday	432.50	393.26	39.24
1	Saturday	918.31	835.27	83.04
1	Sunday	468.67	805.49	-336.82
1	Thursday	452.27	568.26	-115.99
1	Tuesday	523.58	864.32	-340.74
1	Wednesday	437.28	643.71	-206.43
2	Friday	934.49	570.47	364.02
2	Monday	318.28	448.05	-129.77
2	Saturday	804.65	943.33	-138.68
2	Sunday	490.07	573.96	-83.89
2	Thursday	763.21	641.57	121.64

31. Revenue and trip count comparison of VendorID. You can also add passenger count, trip mile etc as a practice for yourself.



VendorID	total_amount_oct	total_amount_sep	total_amount_difference	trip_count_oct	trip_count_sep	trip_count_difference
1	6370.71	5011.86	1358.85	420	319	101
2	23466.88	17740.58	5726.30	1580	1181	399

Use September data for Q24 - Q31:

32 Find the trips that are longer than previous trip. Tip: Luckily, trips are sorted by date and trip IDs are consistent with date. So, we are ready to do "SELF JOIN". You should use trip id to join table with itself. The trick is each trip should be joined with the previous trip. That means join key would be table1.trip_id = table2.trip_id - 1 (or table1.trip_id = table2.trip_id + 1. This is another trick, think about it :)). Then, compare the duration of two trips.

33. For driver ID 1, find the trips that are shorter than the successor (next) trip?

34. Which drivers are having good days? :) (These are the drivers whose next trip is longer than previous trip. In other words, trip duration would increase by every trip for the driver).

35. Could you solve Q34 for total amount instead of trip duration.

Information about the data columns (e.g., the data dictionary) are provided in: https://data.cityofnewyork.us/api/views/hvrh-b6nb/files/65544d38-ab44-4187-a789-5701b114a754?download=true&filename=data_dictionary_trip_records_green.pdf

Note: All these questions are very applicable in business life. Questions after Q32 would look harder (Yes, it is!), however they are essential in some business domains. For example, Netflix and Amazon Prime Video measure customer engagement. The way to find out more engaged customers is comparing each session duration, i.e. any time you log in, with previous/next session. This business problem is a version of Q33.

How about measuring popularity of a movie? Calculate total minutes watched for each movie each day. Compare total minutes watched previous/next day for each movie. If a movie is attracting less viewer, i.e less minutes watched each day, then it is time to end the contract for this movie. This business question is a version of Q31, Q34 and Q35.