Project 2: NYC taxi analytics

Rene Puusepp

Preprocessing

For this task I used the 27gb trip_data folder, as I had no access to the correct data.

Preprocessing was divided into two parts, sampling and cleaning.

Sampling was done in a separate Notebook (see sampling.ipynb) and consisted of taking random samples of 12 .csv files and unioning them. I sampled 2gb from 27gb.

Cleaning is in the main project notebook and consist mainly of removing rows with invalid, illogical data like nulls and zeros where there should be something. Overall 98.3% of data was correctly presented and moved on

Query 1

First block of code was for timestamp filtering, in this sample of data 20 trips happened in the last 30 minutes, which is sadly little to work with later, but correct answer nevertheless

```
earliest timestamp: 2014-01-01 00:19:00
latest timestamp: 2014-01-01 00:49:00
completed within last 30 min: 20
```

Next block was for the grid cells, now I was too hasty to actually read the whole task till the end of part 2 and defined what a “grid cell” is myself. The comments on code explain more but essentially grid is a Tuple consisting of longitude latitude rounded ([60.44,40.02],[17.59,56.20]). The 0.01 degree is roughly 1km, so sensible aswers are expected.

I first selected 30 min range rows and then applied grid system. Just to be frugal with computing power(laptop tired). Later it came back to bite me in the cheek, wrong order for upcoming tasks.

Below is the result of query 1, counts have varied when I sampled at first, it’s a random thing(now with seed), but this dataset was too scarce to get any meaningful counts (max 2)

[14]:

	start_cell	end_cell	count
0	(-73.98, 40.77)	(-73.99, 40.75)	1
1	(-73.78, 40.65)	(-73.8, 40.78)	1
2	(-73.98, 40.77)	(-73.97, 40.71)	1
3	(-73.98, 40.76)	(-73.99, 40.76)	1
4	(-73.99, 40.76)	(-73.98, 40.73)	1
5	(-73.97, 40.8)	(-73.95, 40.83)	1
6	(-73.98, 40.78)	(-73.78, 40.75)	1
7	(-73.87, 40.77)	(-73.98, 40.67)	1
8	(-73.99, 40.72)	(-73.95, 40.72)	1
9	(-74.01, 40.72)	(-73.92, 40.7)	1

Second part of the query is incomplete, as I largely couldn't see the missing pieces in a template of code I met online and let the AI try to realize and fit it to my solution, the online material and AI both expected I know a little more than I do, prime mistake of using sources outside of academia. But the idea of a solution exists in code. Kafka was confusing.

Query 2

First block calculates median profits and checks drop-offs in last 15min.

Sadly since I used wrong dataset 6 much-needed columns were missing and the rest of the code is thus untested, but carefully examined. In real situations test-driven development can solve this issue, I'm too lazy for that.

Next block is the empty taxi detector. Simply polls all who dropped of and haven't picked a new customer up and the block after that divides median income in grid cells with empty taxis and adding it as "profitability" column to the dataframe. Last block simply presents the sadly rhetorical answer.

Even if the columns would be there, it would only honestly help me debug issues. It's a wrong dataset so comparison is mostly useless.

Query 2 part 2

Doesn't exist, I ran out of time and have to study for next 2 exams, sorry.

Conclusion

Partitioning, sampling, preprocessing, queries themselves, easy stuff, but streaming and kafka kept bullying me and I did not take the time to go through all the practice sessions to complete it, so I hacked through problems with StackOverflow, some ChatGPT and whatever I have learned from other database courses.

Biggest mistake was not getting your courses provided material (found it timecostly to go through) which sure as hell seemed simpler than what the great internet recommended me to do.