Diffusion Labs

Assignment Report

Priyanshu Chaurasiya
Final Year UG, IIT KANPUR (208016)

Problem Statement

Fetch historical Uniswap v3 pool data (ETH/USDC at 0.3% fee) and compare it against OHLC data from CoinMarketCap for the past one month. The goal is to convert the raw ETH/USDC price data (fetched from Google BigQuery) into hourly OHLC (Open, High, Low, Close) format and then compare this dataset with the corresponding hourly OHLC data from CoinMarketCap. Finally, calculate and report the deviation between the two data sources.

Approach

Well, firstly i followed the given material which was a GitHub page which walkthrough the environment setup, and how to fetch the raw data from uniswap what i got is looks like this

	hour	block_number	block_timestamp	log_index	transaction_index	amount0	amount1	sqrtPrice	liquidity	tick	price
0	2025-01-07 15:00:00+00:00	21573549	2025-01-07 15:40:23+00:00	140	6	508.350990	-0.146335	16991	0x0000000000000000000000000000000000000000000000000...	194819	3463.708608
1	2025-01-07 15:00:00+00:00	21573549	2025-01-07 15:40:23+00:00	199	6	1999.649107	-0.575564	16990	0x0000000000000000000000000000000000000000000000000...	194818	3464.054979
2	2025-01-07 15:00:00+00:00	21573549	2025-01-07 15:40:23+00:00	168	6	2633.044499	-0.757923	16991	0x0000000000000000000000000000000000000000000000000...	194818	3464.054979
3	2025-01-07 15:00:00+00:00	21573552	2025-01-07 15:40:59+00:00	363	118	42420.329300	-12.20272	16981	0x0000000000000000000000000000000000000000000000000...	194807	3467.867345
4	2025-01-07 15:00:00+00:00	21573560	2025-01-07 15:42:35+00:00	184	7	-393670.342994	114.488758	17075	0x0000000000000000000000000000000000000000000000000...	194917	3429.931729

The data we got was in timestamp with no proper intervals as usually we get through api such as yfinance like hourly, days, etc.
By the way this is the final raw data i got following the github page, to extract OHLC data from it i group the time stamp on hourly basis and assigned as OHLC as follows:

First transactions in the group ----> Open
Last transaction in the group -----> Close
Minimum in the group ----> Low

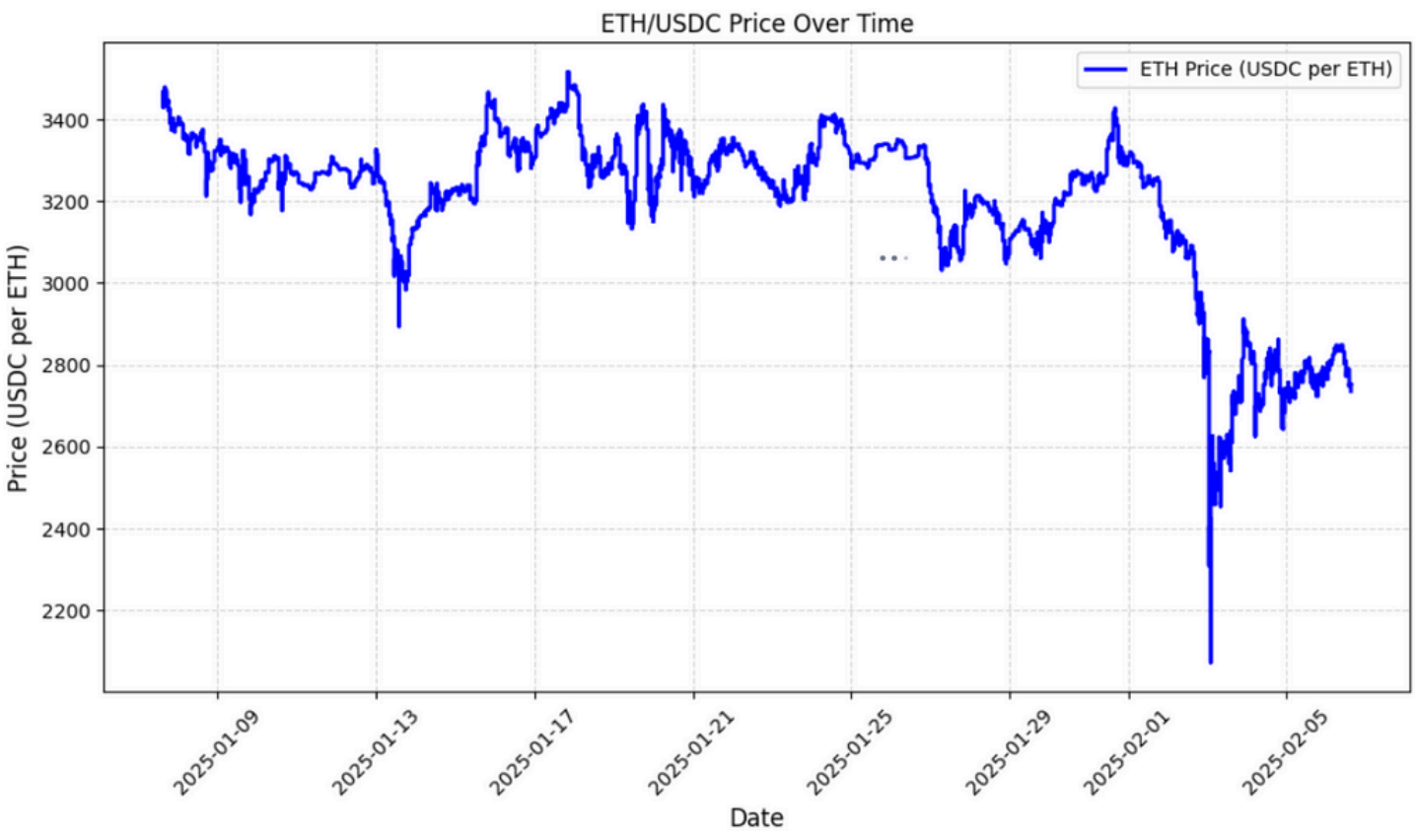
Maximum in the group ---> High

```
# Group by hour and calculate OHLC for the price column
ohlc_data = df.groupby('hour')['price'].agg(['first', 'max', 'min', 'last']).reset_index()
```

Note: Here hour is the timestamp.
Finally my data looks like this

	hour	open	high	low	close
0	2025-01-07 15:00:00+00:00	3463.708608	3467.867345	3429.931729	3465.440809
1	2025-01-07 16:00:00+00:00	3461.631110	3478.633875	3461.284981	3469.601625
2	2025-01-07 17:00:00+00:00	3469.601625	3469.601625	3431.990203	3437.829256
3	2025-01-07 18:00:00+00:00	3440.924540	3446.089542	3421.368037	3421.368037
4	2025-01-07 19:00:00+00:00	3425.475938	3425.818485	3390.376364	3390.376364

This is the Price Plot i got from Uniswap raw data



Now i move fowards to fetch the similar data from CoinMarketCap. I made an account got the api_key and followed the API docs, a standard code for me easily available on internet. I got

the OHLC data which looks like this

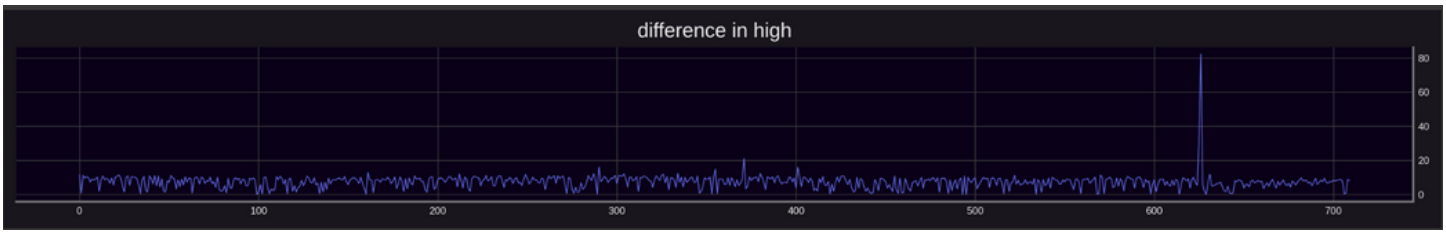
Timestamp	Open	High	Low	Close	Volume	price
2025-01-07 16:00:00+00:00	3466.575689	3490.055257	3451.996226	3460.731914	2.674535e+10	3460.731914
2025-01-07 17:00:00+00:00	3459.813687	3468.764547	3419.708557	3443.779064	2.734356e+10	3443.779064
2025-01-07 18:00:00+00:00	3444.022848	3457.332444	3413.339004	3420.428592	2.783882e+10	3420.428592
2025-01-07 19:00:00+00:00	3420.132912	3435.383417	3379.925000	3382.254017	2.947766e+10	3382.254017
2025-01-07 20:00:00+00:00	3384.304372	3397.462390	3366.275918	3381.126416	3.047348e+10	3381.126416

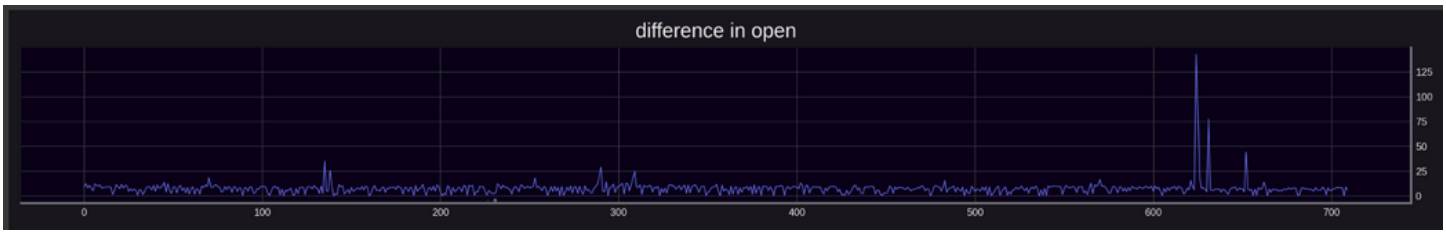
Now i renamed the columns as per their source and merged both the dataset on Timestamp column. And to calculate the deviation in the two source i taken the absolute difference b/w uniswap and CMC data for each Column. Finally it looks like this

	Timestamp	uniswap_open	uniswap_high	uniswap_low	uniswap_close	cmc_open	cmc_high	cmc_low	cmc_close
0	2025-01-07 16:00:00+00:00	3461.631110	3478.633875	3461.284981	3469.601625	3466.575689	3490.055257	3451.996226	3460.731914
1	2025-01-07 17:00:00+00:00	3469.601625	3469.601625	3431.990203	3437.829256	3459.813687	3468.764547	3419.708557	3443.779064
2	2025-01-07 18:00:00+00:00	3440.924540	3446.089542	3421.368037	3421.368037	3444.022848	3457.332444	3413.339004	3420.428592
3	2025-01-07 19:00:00+00:00	3425.475938	3425.818485	3390.376364	3390.376364	3420.132912	3435.383417	3379.925000	3382.254017
4	2025-01-07 20:00:00+00:00	3386.649186	3386.987851	3374.142362	3384.279480	3384.304372	3397.462390	3366.275918	3381.126416

swap_open	uniswap_high	uniswap_low	uniswap_close	cmc_open	cmc_high	cmc_low	cmc_close	diff_open	diff_high	diff_low	diff_close
61.631110	3478.633875	3461.284981	3469.601625	3466.575689	3490.055257	3451.996226	3460.731914	4.944579	11.421382	9.288755	8.869712
69.601625	3469.601625	3431.990203	3437.829256	3459.813687	3468.764547	3419.708557	3443.779064	9.787938	0.837079	12.281646	5.949808
40.924540	3446.089542	3421.368037	3421.368037	3444.022848	3457.332444	3413.339004	3420.428592	3.098308	11.242903	8.029033	0.939445
25.475938	3425.818485	3390.376364	3390.376364	3420.132912	3435.383417	3379.925000	3382.254017	5.343025	9.564931	10.451364	8.122346
86.649186	3386.987851	3374.142362	3384.279480	3384.304372	3397.462390	3366.275918	3381.126416	2.344815	10.474538	7.866443	3.153064

Here diff_open = abs(Uniswap_open - CMC_open) and likewise





Looking at the graphs it doesn't seem there is any high deviation in these two sources except on some timestamp where i can see a huge peak this might be due to missing value cause i have not handled it. sorry for that, but it's rare in the data so we can ignore such peak and lets figure out the statistics of these biases.

Discrepancy Statistics i got from these two sources:

	diff_open	diff_high	diff_close	diff_low
count	710.000000	710.000000	710.000000	710.000000
mean	4.485647	7.403663	4.752714	7.529811
std	4.507731	4.280340	5.140303	7.527539
min	0.000070	0.009366	0.001893	0.012514
25%	2.095366	5.479967	2.189886	4.820275
50%	3.952718	8.107354	4.502583	7.682395
75%	6.385848	9.473812	6.986997	9.203153
max	96.362798	82.257831	120.502898	142.484107

From above here are some interpretations i made

1. Looking at Median and Mean i can say on an average there are more biasness in high and low value in these two data sources compared to open and close.
2. Maximum variance is in low value.
3. The max value is far away from the 3rd quartile value , suggest there are no huge deviation in the price but we have to look at the % value for more clear understanding.
4. Avg. price of Eth in the past one month is 3200 USD, and consider the maximum of mean deviation which is 7.52 comes out to be 0.23%.
5. Mean deviations suggest that, on average, Uniswap prices are higher than CMC prices across all metrics (open, high, low, and close).

To figure out are there hours with very high deviations, i find out the timestamp for which

difference > mean + 2* std. (as a threshold)

Here are the results

Hours with high deviations:

	Timestamp	diff_open	diff_close	diff_high	diff_low
45	2025-01-09 14:00:00+00:00	15.395499	7.129614	4.519440	13.522278
135	2025-01-13 11:00:00+00:00	6.057558	4.410262	7.263556	34.935636
138	2025-01-13 14:00:00+00:00	2.686169	2.107540	2.836236	25.656561
290	2025-01-19 22:00:00+00:00	13.731893	4.652510	16.069579	29.024053
297	2025-01-20 05:00:00+00:00	13.715097	0.842991	9.541387	11.940963
299	2025-01-20 07:00:00+00:00	16.301691	1.509859	11.138788	9.049988
309	2025-01-20 17:00:00+00:00	2.306460	4.615081	7.963658	24.989408
371	2025-01-23 07:00:00+00:00	6.946016	2.601314	21.076850	7.699531
570	2025-01-31 19:00:00+00:00	21.039882	6.111024	11.408392	16.581520
622	2025-02-02 23:00:00+00:00	13.835863	3.851554	10.010814	9.083166
624	2025-02-03 01:00:00+00:00	0.485264	120.502898	5.453375	142.484107
625	2025-02-03 02:00:00+00:00	96.362798	11.082154	35.710404	86.724674
626	2025-02-03 03:00:00+00:00	7.788007	0.362325	82.257831	19.725580
631	2025-02-03 08:00:00+00:00	10.664158	0.645803	11.717598	77.312746
652	2025-02-04 05:00:00+00:00	0.903281	6.391746	5.969034	43.851371

From the results i conclude there are no any specific hours for which deviation is high.
The discrepancy in the prices from these two data might be due to following reasons:

Liquidity Differences – Uniswap is a decentralized exchange (DEX) with liquidity pools, whereas CMC aggregates data from various centralized exchanges.

Latency & API Differences – Data refresh rates or update delays may lead to temporary mismatches.

Arbitrage Effects – Traders exploiting price differences between Uniswap and other platforms can momentarily create divergences.

Market Conditions – High deviations may coincide with volatile periods (e.g., news events, major trades).

References

---

https://github.com/panoptic-labs/research/blob/main/_research_bites/_tutorials/gettingData.ipynb
https://web3-ethereum-defi.readthedocs.io/tutorials/uniswap-v3-price-analysis.htm
https://coinmarketcap.com/api/documentation/v1/