1. STATING THE PROBLEM

My most recent freelance project involved **1,643 datasets** consisting of cryptocurrency trade data between 2022-2023. My tasks were to:

1. Check for dataset availability
2. Extract existing datasets
3. Merge them based on the primary key column
4. Handle missing data
5. Create features for missing datasets based on existing ones
6. Create a quality assurance algorithm to check the final master file
7. Automate the whole process for easy replication.

I came up with a tracking system (excerpt on the left) to always know where I left off, and for my client to see the progress.

It was also handy later in the project to verify any missing directories.

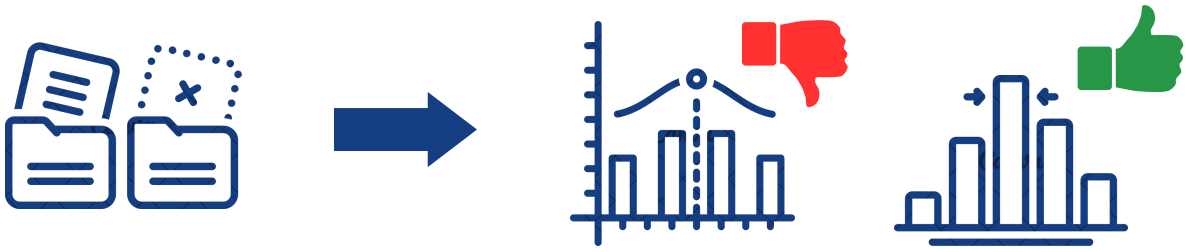
	2023					
Ticker Combination	Exist?	4hr	1hr	30min	5min	1min
ADABNB	YES	YES	YES	YES	YES	YES
ADABTC	YES	YES	YES	YES	YES	YES
ADADAI	NO	NO	NO	NO	NO	NO
ADADOT	NO	NO	NO	NO	NO	NO
ADAETH	YES	YES	YES	YES	YES	YES
ADALINK	NO	NO	NO	NO	NO	NO
ADALTC	NO	NO	NO	NO	NO	NO
ADAMATIC	NO	NO	NO	NO	NO	NO
ADASOL	NO	NO	NO	NO	NO	NO
ADATON	NO	NO	NO	NO	NO	NO
ADATRX	NO	NO	NO	NO	NO	NO
ADAUSDT	YES	YES	YES	YES	YES	YES
ADAVAX	NO	NO	NO	NO	NO	NO
ADAXLM	NO	NO	NO	NO	NO	NO
AVAXBNB	YES	YES	YES	YES	YES	YES
AVAXBTC	YES	YES	YES	YES	YES	YES
AVAXETH	YES	YES	YES	YES	YES	YES
AVAXLINK	NO	NO	NO	NO	NO	NO
AVAXUSDT	YES	YES	YES	YES	YES	YES
AVAXXLM	NO	NO	NO	NO	NO	NO
BNBBTC	YES	YES	YES	YES	YES	YES
BNBETH	YES	YES	YES	YES	YES	YES
BNBUSDT	YES	YES	YES	YES	YES	YES
BTCUSDT	YES	YES	YES	YES	YES	YES
DAIBNB	YES	NO	NO	NO	NO	NO

2. INTERMEDIARY STEPS

There are a couple of things I confirmed with my client to ensure minimum project revisions.

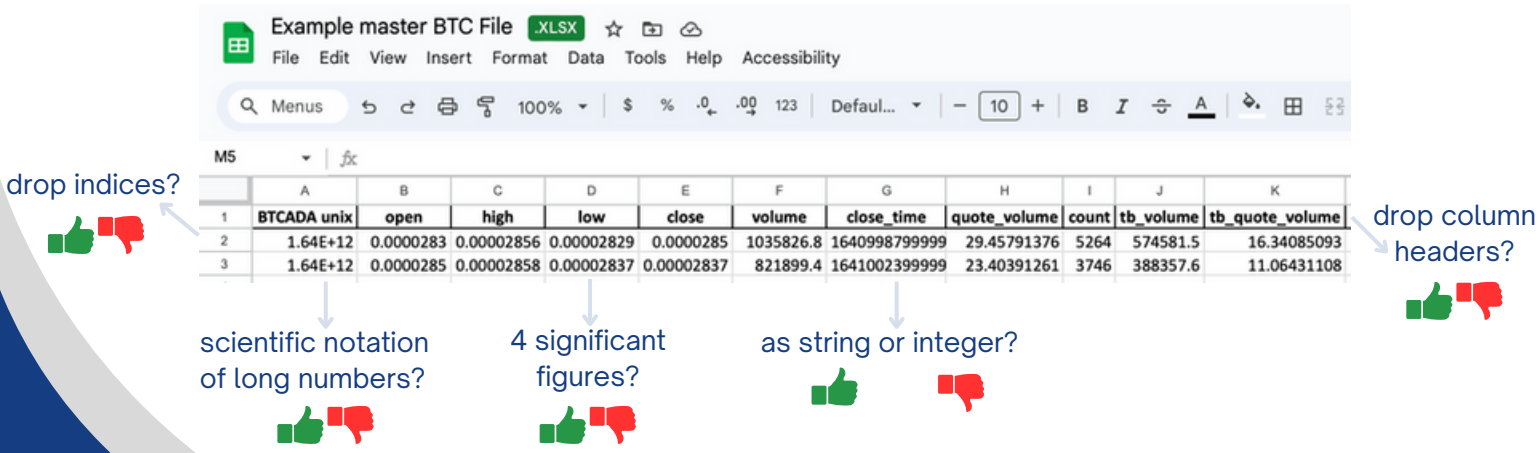
Missing data handling

I suggested the best way to handle missing data in their specific case. However, if you have a preference for whether missed data should be filled with nulls, mediums, averages, or a specific value - let me know!



Example final file

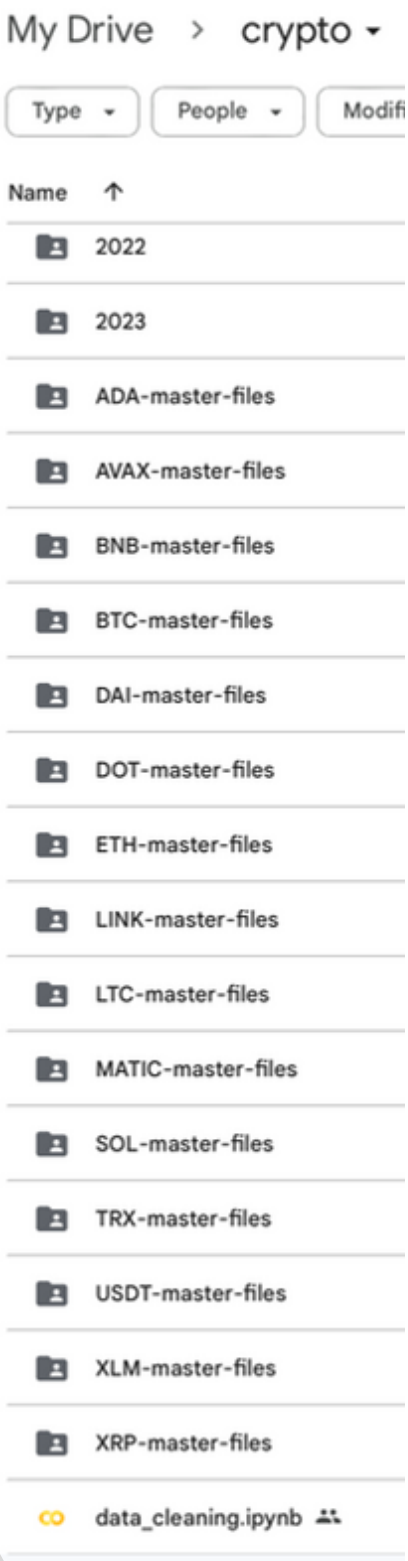
Before diving into automation, I double-checked whether the first master file looked good to my client. It mostly includes column naming, datatypes, row indexing, and file format (csv, xlsx, parquet etc.)



	A	B	C	D	E	F	G	H	I	J	K
1	BTCADA unix	open	high	low	close	volume	close_time	quote_volume	count	tb_volume	tb_quote_volume
2	1.64E+12	0.0000283	0.00002856	0.00002829	0.0000285	1035826.8	1640998799999	29.45791376	5264	574581.5	16.34085093
3	1.64E+12	0.0000285	0.00002858	0.00002837	0.00002837	821899.4	1641002399999	23.40391261	3746	388357.6	11.06431108

Note: all example final files will have up to first 5 rows to not be overwhelmed by its contents.

3. DELIVERABLES & AFTERCARE



My final deliverables included **75 master files** (5 for each crypto coin) and **a Python script** to recreate master files based on the desired coin, dates, and timestamp interval (4 hr, 1 hr, 30 m, 5m, 1m).

The script outputs all errors occurring during the final quality check (if any), and indicates the folder path, in which the master file is stored. Each step in the source code is commented with **instructions to easily replicate** data transformation.

```
_____2023_____

Starting check for ADA-2023-4hr...
Note: path ADAAVAX is not available. 11 columns are created and filled with nulls
Note: path ADADAI is not available. 11 columns are created and filled with nulls
Note: path ADADOT is not available. 11 columns are created and filled with nulls
Note: path ADALINK is not available. 11 columns are created and filled with nulls
Note: path ADALTC is not available. 11 columns are created and filled with nulls
Note: path ADAMATIC is not available. 11 columns are created and filled with nulls
Note: path ADASOL is not available. 11 columns are created and filled with nulls
Note: path ADATRX is not available. 11 columns are created and filled with nulls
Note: path ADAXLM is not available. 11 columns are created and filled with nulls
Note: path ADAXRP is not available. 11 columns are created and filled with nulls
All unix timestamps match!
Master file for ADA-2023-4hr is downloaded to /content/drive/MyDrive/crypto/ADA-master-files/ADA-2023-4hr.csv

Starting check for AVAX-2023-4hr...
Note: path AVAXADA is not available. 11 columns are created and filled with nulls
Note: path AVAXDAI is not available. 11 columns are created and filled with nulls
Note: path AVAXDOT is not available. 11 columns are created and filled with nulls
Note: path AVAXLINK is not available. 11 columns are created and filled with nulls
Note: path AVAXLTC is not available. 11 columns are created and filled with nulls
Note: path AVAXMATIC is not available. 11 columns are created and filled with nulls
Note: path AVAXSOL is not available. 11 columns are created and filled with nulls
Note: path AVAXTRX is not available. 11 columns are created and filled with nulls
Note: path AVAXXLM is not available. 11 columns are created and filled with nulls
Note: path AVAXXRP is not available. 11 columns are created and filled with nulls
All unix timestamps match!
Master file for AVAX-2023-4hr is downloaded to /content/drive/MyDrive/crypto/AVAX-master-files/AVAX-2023-4hr.csv

Starting check for BNB-2023-4hr...
Note: path BNBDAI is not available. 11 columns are created and filled with nulls
All unix timestamps match!
Master file for BNB-2023-4hr is downloaded to /content/drive/MyDrive/crypto/BNB-master-files/BNB-2023-4hr.csv

Starting check for BTC-2023-4hr...
Note: path BTCDAI is not available. 11 columns are created and filled with nulls
All unix timestamps match!
```

NOTES

- The project was done off Fiverr via reference.
- The project was delivered in 8 working days.
- The project had a total of 3 revisions.
- The project’s budget was £500.