Part 3 - Preprocess liquidity data

November 25, 2020

Goal of this file: * Take output from R code and make sure it has quality for analysis

[328]:

```
import pandas as pd
import glob
import os
```

[104]:

```
folder = 'new_data/processed/liquidities'
liquidity_files = glob.glob(os.path.join(folder, '*_liq.csv'))
```

1 Read data

[317]:

```
raw = pd.concat([pd.read_csv(file, index_col=0) for file in liquidity_files])
raw = raw.rename({'effectiveSpread': 'effective_spread', 'realizedSpread':␣
 ↪'realized_spread',
                  'SYMBOL': 'stock_name'}, axis=1)
raw.index = pd.to_datetime(raw.index)
raw['date'] = raw.index.date
raw['hour'] = raw.index.time
```

2 Preprocess

Check for NAs etc

We are analyzing three different dates:

- Jun 19: option expiry date
- Nov 3: election day
- Nov 11: Veterans day

[318]:

```
aux = raw.isna().sum()
aux[aux >0]
```

[318]:

```
realized_spread               4500
proportionalRealizedSpread    4500
priceImpact                   4500
proportionalPriceImpact       4500
squaredLogReturn                15
absLogReturn                    15
```

```
quotedSlope                   604
logQSlope                     604
midQuoteSquaredReturn          15
midQuoteAbsReturn              15
dtype: int64
```

Inspect more closely the NAs in realized_spread

```
[321]: raw[raw['realized_spread'].isna()]['stock_name'].value_counts()
```

```
[321]: AMZN    900
       FB      900
       UAL     900
       TSLA    900
       AAPL    900
       Name: stock_name, dtype: int64
```

At what time do these happen? This is too regular, probably related to market close

```
[324]: aux = pd.Series(pd.to_datetime(raw[raw['realized_spread'].isna()].index)).
        ↪drop_duplicates()
```

```
[325]: for date in aux.dt.date.unique():
           print(date)
           print(aux[aux.dt.date == date].min())
           print(aux[aux.dt.date == date].max())
```

```
2020-11-03
2020-11-03 15:46:55
2020-11-03 16:00:00
2020-11-11
2020-11-11 15:48:08
2020-11-11 16:00:00
2020-06-19
2020-06-19 15:53:03
2020-06-19 16:00:00
```

It seems the empty `realizedSpread` values happen close to market close, for the last 900 ticks. Let's remove them.

```
[326]: raw = raw.dropna()
```

Save to Excel

```
[327]: raw[['stock_name', 'BID', 'BIDSIZ', 'OFR', 'OFRSIZ', 'PRICE','SIZE', 'date',␣
        ↪'hour',
           'midpoints', 'direction', 'effective_spread', 'realized_spread']].\
           to_excel(os.path.join(folder, 'joined_liquidity_date.xlsx'))
```