5. stockPreProcessing

August 30, 2023

[ ]:
```python
import pandas as pd
from datetime import datetime, timedelta
import os
```

[ ]:
```python
# Load stock data

# Stocks :- AAPL, MSFT, AMZN, NVDA, TSLA, GOOGL
# Sector Indices :- SSINFT (^SP500-45)

ticker = "SSINFT"

df = pd.read_excel(f"RawStocks/{ticker}_stock_data.xlsx", engine='openpyxl')
```

[ ]:
```python
# Check for missing values in the DataFrame
df.isnull().sum().sum()
```

[ ]:
```python
# Check for missing values in the DataFrame
if df.isnull().sum().sum() > 0:
    # Use the forward fill method to handle missing values
    df.fillna(method='ffill', inplace=True)

    # If the first row has missing values even after forward fill, you can use back fill for those specific cases
    if df.iloc[0].isnull().sum() > 0:
        df.fillna(method='bfill', inplace=True)
```

[ ]:
```python
# Reformate date
df["Date"] = pd.to_datetime(df['Date'], format="%Y-%m-%d")
```

[ ]:
```python
# Add column indicating if price difference between open and close was positive or negative
df['Difference'] = df['Close'] - df['Open']
df['price_trend'] = df['Difference'].apply(lambda x: 'positive' if x > 0 else ('negative'))
```

[ ]:
```python
directory = "PreProcessedStocks"
```

```python
#If directory doesn't exist, create the directory
if not os.path.exists(directory):
    os.makedirs(directory)

filename = f"{directory}/{ticker}_stock_data.xlsx"
```

[ ]:
```python
df.to_excel(filename, index=False)
```

[ ]:
```python

```

[ ]:
```python

```