Capstone Project 1

Crude oil and CAD dollar correlation

The Data I chose for my first Capstone project is Time series Data from Quandl.com.
I chose two Data sets Crude oil and Canadian dollar currency. The idea behind the first Capstone is trying to find the correlation between the two Data sets in other words trying to find the correlation between the crude oil and the exchange rate of the exporting economy.
I imported the Data sets from Quandl.com using the API provided for crude oil and CAD.
The crude oil API is Crude_Oil and that of CAD is CAD
After importing the Data sets the first step was to create a time series data with the following steps:

A. df = pd.read_csv("Crude_Oi ",parse_dates=['Date'],index_col='Date')
B. df = pd.read_csv("CAD",parse_dates=['Date'],index_col='Date')

first step is to print the top 5 rows of the Data and the last five rows to check the Data
df.head() and *df.tail()*
The first findings was that the data has missing values as the following figures show

Date	Open	High	Low	Last	Change	Settle	Volume	Previous Day Open Interest
2019-05-28	NaN	56.39	NaN	56.39	0.81	56.63	4046.0	17260.0
2019-05-24	NaN	55.70	NaN	55.70	0.94	55.82	692.0	17969.0

:

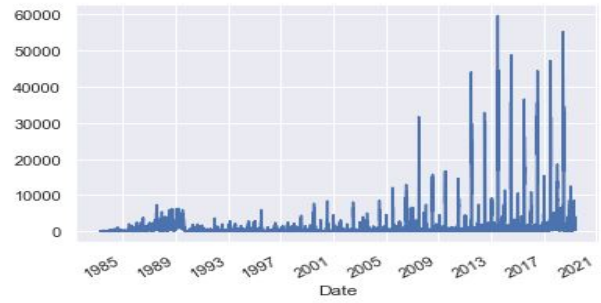
Date	Open	High	Low	Last	Change	Settle	Volume	Previous Day Open Interest
1983-04-06	29.10	29.20	29.10	29.20	NaN	29.20	5.0	19.0
1983-04-05	29.00	29.10	29.00	29.10	NaN	29.10	5.0	18.0

Then printing the shape of the Data df.shape the crude oil data has 9084 rows and 8 columns
The CAD data has 10112 rows and 8 columns
Then I called the .info() method provides important information about a DataFrame, such as the number of rows, number of columns,number of non-missing values in each column, and the data type stored in each column
df.info()

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 9084 entries, 2019-05-28 to 1983-03-3
Data columns (total 8 columns):
Open                            8628 non-null float64
High                            8786 non-null float64
Low                             8821 non-null float64
Last                            8935 non-null float64
Change                          1316 non-null float64
Settle                          9084 non-null float64
Volume                          9084 non-null float64
Previous Day Open Interest      9084 non-null float64
dtypes: float64(8)
```

As the above figure shows that we have a lot of missing data especially in the change column we have 7768 rows of missing data. Since the change is the difference between the settle price of two consecutive trading days.since the settle data has no missing values finding the changes is easy and would be accurate values.
Before dealing with the missing values I used the describe method on the data so I can capture any outliers and check the statistics of the data and it turned out that we have outliers in the volume column but this outliers as the plot shows is actual trading volumes in different time periods

df['Volume'].plot()



Dealing with missing values

As for High,Open,Low and Last applying the forward Method Since the Data has outliers and applying the mean would affect the analysis. The mean is $42 per barrel and some of the NaN values is at the prices level of 80 to 100 dollars per barrel so it is of more statistical sense to fill the data with the forward fill method. Same done to Canadian dollar data set

df_new=df.fillna(method='ffill')

Date	Open	High	Low	Last	Change	Settle	Volume	Previous Day Open Interest
2019-05-28	58.13	56.39	54.90	56.39	0.81	56.63	4046.0	17260.0
2019-05-24	58.13	55.70	54.90	55.70	0.94	55.82	692.0	17969.0
2019-05-23	58.13	59.01	54.90	54.90	2.77	54.88	794.0	17826.0

Date	Open	High	Low	Last	Change	Settle	Volume	Previous Day Open Interest
1983-04-06	29.10	29.20	29.10	29.20	NaN	29.20	5.0	19.0
1983-04-05	29.00	29.10	29.00	29.10	NaN	29.10	5.0	18.0
1983-04-04	28.95	28.95	28.95	28.95	NaN	28.95	0.0	13.0

As for the change calculation I used the shift method on the time series data I have and created a shifted period to push the settle column values into the future.

```
df_new['shifted']=df_new.Settle.shift()
```

After that I create a new column and named it Changes

```
df_new['Changes']=df_new['Settle']-df_new['shifted']
```

Now the changes are captured with the proper sign where the change is positive or negative and the data is ready for analysis and manipulation.The same process I followed with Canadian dollar data set

Date	Open	High	Low	Last	Change	Settle	Volume	Previous Day Open Interest	shifted	Changes
2019-05-28	58.13	56.39	54.90	56.39	0.81	56.63	4046.0	17260.0	55.82	0.81
2019-05-29	58.13	56.39	55.85	55.94	0.39	56.24	1304.0	19564.0	56.63	-0.39