Lab Exercise: Pandas Cleaning Basics

Overview

The objective of this Lab is to gain some experience with some of the pandas functions used to clean up messy data.

Examine and Load the Data

In the Jupyter file browser, examine the csv file USD_CNY.csv and notice how badly formatted it is.

Notice how it uses tab instead of comma as the separator.

Notice how the dates are listed “most recent first”.

Notice how the numerical values have currency symbol and percent signs, this will cause pandas to treat the entire column as a string rather than a number.

Now, create a new notebook.

Import the libraries you need to work with DataFrames

Load the csv file ‘Data/USD_CNY.csv’ into a new DataFrame

Set the ‘Date’ column as the index and parse_dates=True

Tell pandas that the separator is tab by includng the parameter sep=”\t” in read_csv

call sort_index to reorder in ascending order (e.g. df = df.sort_index())

Clean up the Data

Change the format of the Price, Open, High and Low columns so that they are numeric and also do not have any currency symbols e.g.:

```
df['Price'] =
pd.to_numeric(df['Price'].str.replace('$', ''))
```

Similarly, remove the % symbol from the Change column and make this numeric so that 3% is represented by the value 3.0 (not 0.03).

Check for nulls and replace each with the value preceding it.

```
df.fillna(method='ffill', inplace=True)
```

Test your changes by performing some simple checks.

Use dtypes to view the data type of each column.

Filter each column with isnull() and ensure nothing is returned.

```
df.isnull().sum()
```