# week 6

April 10, 2023

```
[7]: import pandas as pd
     df = pd.read_csv('Bitcoin_historical_data.csv')
```

```
[8]: df.head(5)
```

```
[8]:          Date        Open        High         Low       Close   Adj Close  \
     0  2014-09-17  465.864014  468.174011  452.421997  457.334015  457.334015
     1  2014-09-18  456.859985  456.859985  413.104004  424.440002  424.440002
     2  2014-09-19  424.102997  427.834991  384.532013  394.795990  394.795990
     3  2014-09-20  394.673004  423.295990  389.882996  408.903992  408.903992
     4  2014-09-21  408.084991  412.425995  393.181000  398.821014  398.821014

          Volume
     0  21056800
     1  34483200
     2  37919700
     3  36863600
     4  26580100
```

```
[2]: import dask.dataframe as dd
     df = dd.read_csv('Bitcoin_historical_data.csv')
```

```
[3]: df.head(5)
```

```
[3]:          Date        Open        High         Low       Close   Adj Close  \
     0  2014-09-17  465.864014  468.174011  452.421997  457.334015  457.334015
     1  2014-09-18  456.859985  456.859985  413.104004  424.440002  424.440002
     2  2014-09-19  424.102997  427.834991  384.532013  394.795990  394.795990
     3  2014-09-20  394.673004  423.295990  389.882996  408.903992  408.903992
     4  2014-09-21  408.084991  412.425995  393.181000  398.821014  398.821014

          Volume
     0  21056800
     1  34483200
     2  37919700
     3  36863600
     4  26580100
```

**Basic Validation on data columns:**

[5]:

```
df.columns = df.columns.str.replace('[^\w\s]', '').str.strip()
```

```
/var/folders/_0/nmpfpzw134n12j0c0z6jtrw80000gn/T/ipykernel_42306/4131416077.py:1
: FutureWarning: The default value of regex will change from True to False in a
future version.
  df.columns = df.columns.str.replace('[^\w\s]', '').str.strip()
```

**Create a YAML file:**

[9]:

```
import yaml
columns = df.columns.tolist()
with open('columns.yml', 'w') as file:
    documents = yaml.dump(columns, file)
```

**Validating the number of columns and column name of ingested file with YAML:**

[10]:

```
with open('columns.yml', 'r') as file:
    expected_columns = yaml.safe_load(file)
if set(expected_columns) != set(df.columns):
    raise ValueError('Column names do not match')
if len(expected_columns) != len(df.columns):
    raise ValueError('Incorrect number of columns')
```

**Writing the file in pipe-separated text file (|) in gz format:**

[11]:

```
df.to_csv('output_file.csv.gz', index=False, sep='|', compression='gzip')
```

**Creating a summary of the file:**

[14]:

```
import os

num_rows = len(df)
num_columns = len(df.columns)
file_size = os.path.getsize('Bitcoin_historical_data.csv')
```