0_preprocess_data

November 18, 2022

0.1 # 0_preprocessing the Crypto datasets

Group , October 19, 2022 1. Eduardo Garcia 2. Nari Kim 3. Thi Anh Ba Dang 4. Vishnu Prabhakar 5. VS Chaitanya Madduri 6. Yumeng Zhang

Description: merging the daily and minute crypto datasets to respective single files.

This mode

0.1.1 Pre requisites:

1. And add the shortcut of the drive link : https://drive.google.com/drive/folders/1KRMbTR4GNaDGlpBkRi3
   to your personal drive.

0.1.2 Input files:

No files

0.1.3 Output files:

crypto_data_hour_cleaned_v2.csv crypto_data_weekly_cleaned_v1.csv
crypto_data_minute_cleaned_v2.csv

DriveLink :https://drive.google.com/drive/folders/1F8P3UlqSE6lFpHyBidVArdsXJLenKi3O?usp=share_link

0.2 1. Import Required Packages

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

[ ]:
```
# Connecting to the google drive
from google.colab import drive
drive.mount('/content/drive')
from IPython.display import clear_output
```

```
Mounted at /content/drive
```

[ ]:
```
def fill_empty_dates(dataframe_crypto):
  '''
  The funtion will add the missing dates to the crypto coins.
  input:
```

```
  dataframe_crypto : The dataset where the missing dates to be added.
  Output:
  final_df: date filled dataset.

  '''
  start_date = dataframe_crypto['Open Time'].min()
  end_date = dataframe_crypto['Open Time'].max()
  final_list = pd.date_range(start_date,end_date,
               freq='min').strftime("%Y-%m-%d %H:%M:%S").tolist()
  date_df = pd.DataFrame(final_list, columns=['Open Time'])
  date_df['Open Time'] = pd.to_datetime(date_df['Open Time'])
  final_df = dataframe_crypto.merge(date_df, how='right', on='Open Time')
  final_df.ffill(inplace=True)
  return final_df
```

[ ]:
```
# last one yer records are made as test and rest of the records Train
def assign_train_test_tag(df):
  '''
  Assigning the test tag to the last 365 days entries and rest as Train
  to the crypto datasets
  input:
  df : The dataset where the train/test tag need to be added.
  Output:
  df: Data filled dataset.
  '''

  df['train_test'] = 'Train'
  df['train_test'][(df['Open Time']>max(df['Open Time']) -  timedelta(days=365)
 ↪) & (df['Open Time']<=max(df['Open Time']))] ='Test'
  df['train_test'][~((df['Open Time']>max(df['Open Time']) - 
 ↪timedelta(days=365) ) & (df['Open Time']<=max(df['Open Time'])))] ='Train'


  return df
```

[ ]:
```
def extract_hourly_data(crypto_data):
  '''
  The function processes the extarct the time series values from a given hourly
 ↪range.
  Input:
  crypto_data: The minute data of the time series coin/stock.
  Output:
  temp_df : the final dataset will all the columns.
  '''
  crypto_data.sort_values(['Open Time'], inplace=True)
  crypto_data['Open Time'] = pd.to_datetime(crypto_data['Open Time'])
```

```
    crypto_data['temp_col'] = crypto_data['Open Time'].dt.date.astype(str) + '_'␣
    ↪+ crypto_data['Open Time'].dt.hour.astype(str)
    temp_df = crypto_data.groupby('temp_col').agg({
        'Open Time':'min',
        'Open':'first',
        'High':'max',
        'Low':'min',
        'Close': 'last',
        'Volume': 'sum',

                                        })
    temp_df = temp_df.reset_index(drop=True).sort_values('Open Time')
    temp_df['Open Time'] = temp_df['Open Time'].dt.date.astype(str) + ' ' +␣
    ↪temp_df['Open Time'].dt.hour.astype(str) +':00:00'
    temp_df['Open Time'] = pd.to_datetime(temp_df['Open Time'])
    return temp_df
```

[ ]:
```
def extract_daily_data(crypto_data):
    '''
    The function processes the extarct the time series values from a given daily␣
    ↪range.
    Input:
    crypto_data: The minute/hourly data of the time series coin/stock.
    Output:
    temp_df : the final dataset will all the columns.
    '''
    crypto_data.sort_values(['Open Time'], inplace=True)
    crypto_data['Open Time'] = pd.to_datetime(crypto_data['Open Time'])
    crypto_data['temp_col'] = crypto_data['Open Time'].dt.date.astype(str) + '_'␣
    ↪+ crypto_data['Open Time'].dt.day.astype(str)
    temp_df = crypto_data.groupby('temp_col').agg({
        'Open Time':'min',
        'Open':'first',
        'High':'max',
        'Low':'min',
        'Close': 'last',
        'Volume': 'sum',
                                        })
    temp_df = temp_df.reset_index(drop=True).sort_values('Open Time')
    temp_df['Open Time'] = temp_df['Open Time'].dt.date.astype(str) + ' 00:00:00'
    temp_df['Open Time'] = pd.to_datetime(temp_df['Open Time'])
    return temp_df
```

[ ]:
```
def extract_weekly_data(crypto_data):
    '''
    The function processes the extarct the time series values from a given weekly␣
    ↪range.
```

```
  Input:
  crypto_data: The minute/daily data of the time series coin/stock.
  Output:
  temp_df : the final dataset will all the columns.
  '''
  crypto_data.sort_values(['Open Time'], inplace=True)
  crypto_data['Open Time'] = pd.to_datetime(crypto_data['Open Time'])
  crypto_data['year'] = crypto_data['Open Time'].dt.year
  crypto_data['week_number'] = crypto_data['Open Time'].dt.isocalendar().week
  crypto_data['temp_col'] = crypto_data['year'].astype(str) + '_' +␣
↪crypto_data['week_number'].astype(str)
  temp_df = crypto_data.groupby('temp_col').agg({
      'Open Time':'min',
      'year':'min',
      'week_number':'min',
      'Open':'min',
      'High':'max',
      'Low':'min',
      'Close': 'max',
      'Volume': 'sum',
                                        })
  temp_df.reset_index(drop=True, inplace=True)
  return temp_df
```

[ ]:
```
project_folder = "/content/drive/MyDrive/DL_2022_Assignment"
```

[ ]:
```
 # extracting the folder structure so that the dic will have all the csv␣
 ↪locations
folder_dic = {}
files = []
for main_path in os.listdir(project_folder):

  sub_folder = project_folder + '/' +  main_path
  folder_dic[sub_folder] = []
  for path in os.listdir(sub_folder):
      # check if current path is a file
      if os.path.isfile(os.path.join(sub_folder, path)):
          folder_dic[sub_folder].append(path)
```

[ ]:
```
folder_dic.keys()
```

[ ]:
```
dict_keys(['/content/drive/MyDrive/DL_2022_Assignment/Crypto_data_daily',
'/content/drive/MyDrive/DL_2022_Assignment/Crypto_data_minute'])
```

[ ]:
```
print("Number of the daily data files : {}".format(len(folder_dic['/content/
 ↪drive/MyDrive/DL_2022_Assignment/Crypto_data_daily'])))
```

```
Number of the daily data files : 106
```

```
[ ]: print("Number of the minute data files : {}".format(len(folder_dic['/content/
       ↪drive/MyDrive/DL_2022_Assignment/Crypto_data_minute'])))
```

```
Number of the minute data files : 10
```

```
[ ]: # making one dataframe out of the minute and the daily data
```

```
[ ]: folder_dic['/content/drive/MyDrive/DL_2022_Assignment/Crypto_data_minute'][2]
```

```
[ ]: 'XRP_1min.txt'
```

```
[ ]:
```

0.3 Processing the Minute date extarcting Hourly, weekly , Minute data.

```
[ ]: col_names=['Open Time',  'Open', 'High', 'Low', 'Close', 'Volume']

master_hour_df = pd.DataFrame(columns=col_names)
master_daily_df = pd.DataFrame(columns=col_names)
master_week_df = pd.DataFrame(columns=col_names)


minute_path = "/content/drive/MyDrive/DL_2022_Assignment/Crypto_data_minute"
for file_name  in folder_dic[minute_path]:
  df = pd.read_csv(minute_path + '/' +  file_name, names=col_names)
  df['Open Time'] = pd.to_datetime(df['Open Time'])
  # df = fill_empty_dates(df.copy())



  daily_df = extract_daily_data(df.copy())
  hour_df = extract_hourly_data(df.copy())
  weekly_df = extract_weekly_data(df.copy())



  daily_df = assign_train_test_tag(daily_df.copy())
  hour_df = assign_train_test_tag(hour_df.copy())
  weekly_df = assign_train_test_tag(weekly_df.copy())


  daily_df['Crypto'] = file_name.split("_")[0]
  hour_df['Crypto'] = file_name.split("_")[0]
  weekly_df['Crypto'] = file_name.split("_")[0]

  master_hour_df = pd.concat([master_hour_df, hour_df])
  master_daily_df = pd.concat([master_daily_df, daily_df])
  master_week_df = pd.concat([master_week_df, weekly_df])
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:5:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  """
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:6:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:5:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  """
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:6:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:5:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  """
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:6:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:5:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
```

```
  """
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:6:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:5:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  """
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:6:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:5:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  """
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:6:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:5:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  """
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:6:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
```

```
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:5:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  """
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:6:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:5:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  """
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:6:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:5:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  """
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:6:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:5:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame
```

```
See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  """
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:6:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:5:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  """
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:6:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:5:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  """
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:6:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:5:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  """
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:6:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame
```

```
See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:5:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  """
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:6:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:5:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  """
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:6:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:5:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  """
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:6:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:5:
SettingWithCopyWarning:
```

```
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  """
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:6:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:5:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  """
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:6:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:5:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  """
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:6:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:5:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  """
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:6:
```

```
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:5:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  """
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:6:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:5:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  """
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:6:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:5:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  """
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:6:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:5:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  """
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:6:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:5:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  """
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:6:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:5:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  """
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:6:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:5:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
```

```
  """
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:6:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:5:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  """
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:6:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:5:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  """
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:6:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
```

0.4 New columns

0.4.1 Calculating the percentage change hourly data

[ ]:
```
# Calculating the percentage change

master_hour_df['pct_change_1hour'] = master_hour_df['Close'].
  ↪pct_change(periods=1)
master_hour_df['pct_change_2hour'] = master_hour_df['Close'].
  ↪pct_change(periods=2)
```

```
master_hour_df['pct_change_1day'] = master_hour_df['Close'].
 ↪pct_change(periods=24)
```

[ ]:
```
master_hour_df.head(2)
```

[ ]:
```
             Open Time    Open    High     Low   Close  Volume train_test  \
0 2013-04-01 00:00:00  93.155  93.155  93.155  93.155   12.25      Train
1 2013-04-01 01:00:00  93.700  93.790  93.700  93.790   54.12      Train

  Crypto  pct_change_1hour  pct_change_2hour  pct_change_1day
0    BTC               NaN               NaN              NaN
1    BTC          0.006817               NaN              NaN
```

0.4.2 Hourly Data Sample

[ ]:
```
master_hour_df[['train_test','Crypto']].value_counts()
```

[ ]:
```
train_test  Crypto
Train       BTC       74200
            ETH       48359
            LTC       41607
            ETC       41584
            XMR       41430
            XRP       40595
            XLM       34989
            TRX       34453
            ADA       30902
            LINK      19751
Test        BTC        8760
            ETH        8722
            ETC        8712
            ADA        8712
            LINK       8707
            LTC        8689
            XMR        8664
            XRP        8640
            TRX        8640
            XLM        8616
dtype: int64
```

[ ]:
```
# exporting the dataframe to csv
folder_path = '/content/drive/MyDrive/MADS_23_DL_final_project'
master_hour_df.to_csv(folder_path + "/data/crypto_data_hour_cleaned_v2.csv",␣
 ↪index=None)
```

0.4.3 Daily Data Sample

[ ]:
```
master_daily_df[['train_test','Crypto']].value_counts()
```

[ ]:
```
train_test  Crypto
Train       BTC       3098
            ETH       2023
            XMR       1734
            LTC       1734
            ETC       1734
            XRP       1703
            XLM       1511
            TRX       1455
            ADA       1310
            LINK       823
Test        BTC        365
            ETH        364
            LTC        363
            LINK       363
            ETC        363
            ADA        363
            XMR        361
            XRP        360
            TRX        360
            XLM        359
dtype: int64
```

[ ]:
```
# exporting the dataframe to csv
folder_path = '/content/drive/MyDrive/MADS_23_DL_final_project'
master_daily_df.to_csv(folder_path + "/data/crypto_data_daily_cleaned_v1.csv",␣
 ↪index=None)
```

0.4.4 Weekly data sample

[ ]:
```
master_week_df[['train_test','Crypto']].value_counts()
```

[ ]:
```
train_test  Crypto
Train       BTC       446
            ETH       291
            XMR       248
            LTC       248
            ETC       248
            XRP       244
            XLM       219
            TRX       209
            ADA       189
            LINK      119
Test        XRP        54
```

```
BTC         54
XMR         54
XLM         54
TRX         54
LTC         54
LINK        54
ETH         54
ETC         54
ADA         54
dtype: int64
```

[ ]:
```
# exporting the dataframe to csv
folder_path = '/content/drive/MyDrive/MADS_23_DL_final_project'
master_week_df.to_csv(folder_path + "/data/crypto_data_weekly_cleaned_v1.csv",␣
 ↪index=None)
```

0.4.5 combining all the datasets into one csv

[ ]:
```
# exporting the dataframe to csv
folder_path = '/content/drive/MyDrive/MADS_23_DL_final_project'
master_minute_df.to_csv(folder_path + "/data/crypto_data_minute_cleaned_v2.
 ↪csv", index=None)
```

[ ]:
```

```

[ ]:
```

```

0.5 Merging the Daily data

Discarded the code as we are not using the daily data for the analysis

[ ]:
```
col_names = ['Open Time', 'Open Time.1', 'Open', 'High', 'Low', 'Close',␣
 ↪'Volume',
       'Close Time', 'Quote Asset Volume', 'Number of Trades',
       'TB Base Volume', 'TB Quote Volume', 'Ignore']

master_daily_df = pd.DataFrame(columns=col_names)
minute_path = "/content/drive/MyDrive/DL_2022_Assignment/Crypto_data_daily"
for file_name  in folder_dic[minute_path]:
  df = pd.read_csv(minute_path + '/' +  file_name, names=col_names)
  df['filename_name'] = file_name
  master_daily_df = pd.concat([master_daily_df, df])
```

[ ]:
```
# exporting the dataframe to csv
folder_path = '/content/drive/MyDrive/MADS_23_DL_final_project'
master_daily_df.to_csv(folder_path + "/crypto_data_daily_all.csv", index=None)
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-3-d39bd88266ee> in <module>
      1 # exporting the dataframe to csv
      2 folder_path = '/content/drive/MyDrive/MADS_23_DL_final_project'
----> 3 master_daily_df.to_csv(folder_path + "/crypto_data_daily_all.csv",␣
 ↪index=None)

NameError: name 'master_daily_df' is not defined
```

[ ]:
```
master_daily_df.shape
```

[ ]: (88445, 14)

0.6 Loading the minute data

[ ]:
```
folder_path = '/content/drive/MyDrive/MADS_23_DL_final_project'
master_minute_df = pd.read_csv(folder_path + "/data/
 ↪crypto_data_minute_cleaned_v2.csv")
```

[ ]:
```
master_minute_df.shape
```

[ ]: (30210628, 13)

[ ]:
```
master_minute_df['Open Time'] = pd.to_datetime(master_minute_df['Open Time'])
master_minute_df['Open_time_minute'] = master_minute_df['Open Time'].dt.minute
final_df = master_minute_df[master_minute_df['Open_time_minute'] == 59]
```

[ ]:
```
final_df.drop(['Open_time_minute'], axis=1).to_csv(folder_path + "/data/
 ↪crypto_data_hour_cleaned_v1.csv", index=None)
```

[ ]:
```
final_df.shape
```

[ ]: (503505, 14)

[ ]:
```
master_minute_df.sample(10)
```

[ ]:

	Open Time	Open	High	Low	Close	\
5451669	2017-01-21 03:30:00	10.746	10.746	10.746	10.746	
15089835	2019-05-22 08:23:00	91.610	91.720	91.242	91.247	
5597379	2017-05-02 08:00:00	80.956	81.523	80.932	80.932	
21502321	2022-08-31 21:30:00	0.105	0.105	0.105	0.105	
15399204	2019-12-23 04:32:00	42.200	42.200	42.000	42.010	
12673342	2020-07-16 05:06:00	0.128	0.129	0.128	0.128	
14081044	2017-06-20 19:12:00	43.350	44.000	42.654	42.654	
7962660	2021-10-30 21:21:00	4298.020	4300.000	4297.710	4299.040	

```
25702076  2019-12-04 01:21:00    52.259    52.259    52.259    52.259
17222742  2020-03-10 12:51:00     4.218     4.221     4.212     4.220

              Volume  pct_change_1min  pct_change_30min  pct_change_1hour  \
5451669        1.025        -0.000372          0.002425          0.004675
15089835      84.960        -0.004289          0.000976         -0.006889
5597379      206.715         0.000396          0.007544         -0.009885
21502321   27948.000         0.000000          0.000000          0.000000
15399204      87.727        -0.004502         -0.008052         -0.008731
12673342  1045164.000        -0.007752         -0.007752         -0.015385
14081044    1031.803        -0.016736         -0.045558         -0.041548
7962660       72.962         0.000179         -0.004562         -0.008449
25702076       0.005         0.000843         -0.007784         -0.016597
17222742    4694.072         0.001424          0.013449          0.014423

          pct_change_2hour  pct_change_1day train/test filename_name
5451669           0.014156         0.013391      train           BTC
15089835         -0.007538         0.002494      train           BTC
5597379           0.033350        -0.044712      train           BTC
21502321          0.000000         0.019417       test           BTC
15399204         -0.005657         0.054362      train           BTC
12673342         -0.022901        -0.030303      train           BTC
14081044         -0.072134        -0.081327      train           BTC
7962660          -0.005611        -0.024105       test           BTC
25702076         -0.022356        -0.042875      train           BTC
17222742          0.029268         0.020309      train           BTC
```

0.7 Crypto Split

[ ]:
```
folder_path = '/content/drive/MyDrive/MADS_23_DL_final_project'
master_minute_df = pd.read_csv(folder_path + "/data/
 ↪crypto_data_daily_cleaned_v1.csv")
```

[ ]:
```
import plotly.express as px

df2 = master_minute_df.groupby(['Crypto'])['Crypto'].count().
 ↪reset_index(name='count')


fig = px.pie(df2 , values='count',
             names='Crypto',
             title='Crypto',
             width=600, height=500,
             color_discrete_sequence=px.colors.sequential.RdBu)
fig.update_layout( title_x=0.5)
fig.show()
```

[ ]:
```
import plotly.express as px


fig = px.bar(df2.sort_values(by='count') ,
             x='Crypto',
             y='count',
             title='Crypto',
             width=600, height=500,
            #  color_discrete_sequence=px.colors.sequential.RdBu
             )
fig.update_layout( title_x=0.5)
fig.show()
```

[ ]:

[ ]:
```
folder_path = '/content/drive/MyDrive/MADS_23_DL_final_project/data/old_files'
master_minute_df = pd.read_csv(folder_path + "/crypto_data_minute_cleaned_v2.
 ↪csv")
```

[ ]:
```
master_minute_df.shape
```

[ ]:
```
(30210628, 13)
```

[ ]:
```
# file path
folder_path = '/content/drive/MyDrive/MADS_23_DL_final_project'
hour = pd.read_csv(folder_path + '/data/crypto_data_daily_cleaned_v1.csv')
```

[ ]:
```
hour.shape
```

[ ]:
```
(20746, 8)
```

0.8 End of the Notebook

[ ]: