Pandas Foundation

November-15-17 7:12 PM

https://campus.datacamp.com/courses/pandas-foundations/data-ingestion-inspection?ex=4

#Ch 1 Data ingestion & inspection

#NumPy and pandas working together
# Import numpy
import numpy as np

# Create array of DataFrame values: np_vals
np_vals = df.values

# Create new array of base 10 logarithm values: np_vals_log10
np_vals_log10 = np.log10 (np_vals)

# Create array of new DataFrame by passing df to np.log10(): df_log10
df_log10 = np.log10 (df)

# Print original and new data containers
print(type(np_vals), type(np_vals_log10))
print(type(df), type(df_log10))

#Zip lists to build a DataFrame
# Zip the 2 lists together into one list of (key,value) tuples: zipped
zipped = list (zip(list_keys, list_values))

# Inspect the list using print()
print(zipped)

# Build a dictionary with the zipped list: data
data = dict (zipped)

# Build and inspect a DataFrame from the dictionary: df
df = pd.DataFrame (data)
print(df)

#Labeling your data
# Build a list of labels: list_labels
list_labels = ['year', 'artist', 'song', 'chart weeks']

# Assign the list of labels to the columns attribute: df.columns
df.columns = list_labels

#Building DataFrames with broadcasting
# Make a string with the value 'PA': state
state = 'PA'

# Construct a dictionary: data
data = {'state': state, 'city':cities}

# Construct a DataFrame from dictionary data: df
df = pd.DataFrame (data)

# Print the DataFrame
print(df)

#Reading a flat file
# Read in the file: df1
df1 = pd.read_csv ('world_population.csv')

# Create a list of the new column labels: new_labels
new_labels = ['year', 'population']

# Read in the file, specifying the header and names parameters: df2
df2 = pd.read_csv ('world_population.csv', header=0, names=new_labels)

# Print both the DataFrames
print(df1)
print(df2)

#Delimiters, headers, and extensions
# Read the raw file as-is: df1
df1 = pd.read_csv(file_messy)

# Print the output of df1.head()
print(df1.head())

# Read in the file with the correct parameters: df2
df2 = pd.read_csv(file_messy, delimiter=' ', header=3, comment='#')

# Print the output of df2.head()
print(df2.head())

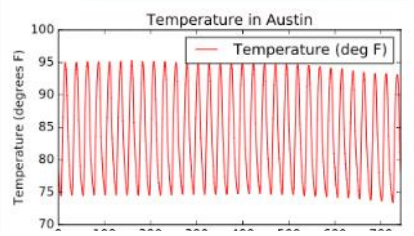
# Save the cleaned up DataFrame to a CSV file without the index
df2.to_csv(file_clean, index=False)

# Save the cleaned up DataFrame to an excel file without the index
df2.to_excel('file_clean.xlsx', index=False)

#Plotting series using pandas
Create a plot with color='red'
df.plot (color = 'red')

Add a title
plt.title ('Temperature in Austin')

Specify the x-axis label
plt.xlabel ('Hours since midnight August 1, 2010')

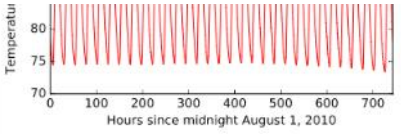


```
# Add a title
plt.title ('Temperature in Austin')

# Specify the x-axis label
plt.xlabel ('Hours since midnight August 1, 2010')

# Specify the y-axis label
plt.ylabel ('Temperature (degrees F)')

# Display the plot
plt.show ()
```




```
#Plotting DataFrames
# Plot all columns (default)
df.plot ()
plt.show()

# Plot all columns as subplots
df.plot (subplots = True)
plt.show()

# Plot just the Dew Point data
column_list1 = ['Dew Point (deg F)']
df[column_list1].plot ()
plt.show()

# Plot the Dew Point and Temperature data, but not the Pressure data
column_list2 = ['Temperature (deg F)','Dew Point (deg F)']
df[column_list2].plot ()
plt.show()
```