Module 7: Data Wrangling with Pandas

CPE311 Computational Thinking with Python

Submitted by: Cuadra, Audrick Zander

Performed on: 03/20/2024

Submitted on: 03/20/2024

Submitted to: Engr. Roman M. Richard

7.1 Supplementary Activity

Using the datasets provided, perform the following exercises:

Exercise 1

We want to look at data for the Facebook, Apple, Amazon, Netflix, and Google (FAANG) stocks, but we were given each as seperate CSV file. Combine them into a single file and store the dataframe of the FAANG data as faang for the rest of the exercises:

1. Read each file in.
2. Add a column to each dataframe, called ticker, indicating the ticker symbol it is for (Apple's is AAPL for example). This is how you look yp a stock. Each file's name is also ticker symbol, so be sure to capitalize it.
3. Append them together into a single dataframe.
4. Save the result in CSV file called faang.csv.

```
# 1. Read each file in.
import pandas as p
import numpy as ny

facebook = p.read_csv('/content/fb.csv')
apple = p.read_csv('/content/aapl.csv')
amazon = p.read_csv('/content/amzn.csv')
netflix = p.read_csv('/content/nflx.csv')
google = p.read_csv('/content/goog.csv')
```

```
# 2. Add a column to each dataframe, called ticker, indicating the ticker symbol it
#    is for (Apple's is AAPL for example). This is how you look yp a stock. Each file's
#    name is also ticker symbol, so be sure to capitalize it.
facebook['Ticker'] = 'FB'
new_f = facebook
new_f
```

	date	open	high	low	close	volume	Ticker
0	2018-01-02	177.68	181.58	177.5500	181.42	18151903	FB
1	2018-01-03	181.88	184.78	181.3300	184.67	16886563	FB
2	2018-01-04	184.90	186.21	184.0996	184.33	13880896	FB
3	2018-01-05	185.59	186.90	184.9300	186.85	13574535	FB
4	2018-01-08	187.20	188.90	186.3300	188.28	17994726	FB
...	...	...	...	...	...	...	...
246	2018-12-24	123.10	129.74	123.0200	124.06	22066002	FB
247	2018-12-26	126.00	134.24	125.8900	134.18	39723370	FB
248	2018-12-27	132.44	134.99	129.6700	134.52	31202509	FB
249	2018-12-28	135.34	135.92	132.2000	133.20	22627569	FB
250	2018-12-31	134.45	134.64	129.9500	131.09	24625308	FB

251 rows × 7 columns

Next steps: View recommended plots

```
apple['Ticker'] = 'AAPL'
new_aa = apple
new_aa
```

	date	open	high	low	close	volume	Ticker
0	2018-01-02	166.9271	169.0264	166.0442	168.9872	25555934	AAPL
1	2018-01-03	169.2521	171.2337	168.6929	168.9578	29517899	AAPL
2	2018-01-04	169.2619	170.1742	168.8106	169.7426	22434597	AAPL
3	2018-01-05	170.1448	172.0381	169.7622	171.6751	23660018	AAPL
4	2018-01-08	171.0375	172.2736	170.6255	171.0375	20567766	AAPL
...	...	...	...	...	...	...	...
246	2018-12-24	147.5173	150.9027	145.9639	146.2029	37169232	AAPL
247	2018-12-26	147.6666	156.5585	146.0934	156.4987	58582544	AAPL
248	2018-12-27	155.1744	156.1004	149.4291	155.4831	53117065	AAPL
249	2018-12-28	156.8273	157.8430	153.8899	155.5627	42291424	AAPL
250	2018-12-31	157.8529	158.6794	155.8117	157.0663	35003466	AAPL

251 rows × 7 columns

```
amazon['Ticker'] = 'AMZN'
new_am = amazon
new_am
```

	date	open	high	low	close	volume	Ticker
0	2018-01-02	1172.00	1190.00	1170.51	1189.01	2694494	AMZN
1	2018-01-03	1188.30	1205.49	1188.30	1204.20	3108793	AMZN
2	2018-01-04	1205.00	1215.87	1204.66	1209.59	3022089	AMZN
3	2018-01-05	1217.51	1229.14	1210.00	1229.14	3544743	AMZN
4	2018-01-08	1236.00	1253.08	1232.03	1246.87	4279475	AMZN
...	...	...	...	...	...	...	...
246	2018-12-24	1346.00	1396.03	1307.00	1343.96	7219996	AMZN
247	2018-12-26	1368.89	1473.16	1363.01	1470.90	10411801	AMZN
248	2018-12-27	1454.20	1469.00	1390.31	1461.64	9722034	AMZN
249	2018-12-28	1473.35	1513.47	1449.00	1478.02	8828950	AMZN
250	2018-12-31	1510.80	1520.76	1487.00	1501.97	6954507	AMZN

251 rows × 7 columns

```
google['Ticker'] = 'GOOG'
new_g = google
new_g
```

	date	open	high	low	close	volume	Ticker
0	2018-01-02	1048.34	1066.94	1045.23	1065.00	1237564	GOOG
1	2018-01-03	1064.31	1086.29	1063.21	1082.48	1430170	GOOG
2	2018-01-04	1088.00	1093.57	1084.00	1086.40	1004605	GOOG
3	2018-01-05	1094.00	1104.25	1092.00	1102.23	1279123	GOOG
4	2018-01-08	1102.23	1111.27	1101.62	1106.94	1047603	GOOG
...	...	...	...	...	...	...	...
246	2018-12-24	973.90	1003.54	970.11	976.22	1590328	GOOG
247	2018-12-26	989.01	1040.00	983.00	1039.46	2373270	GOOG
248	2018-12-27	1017.15	1043.89	997.00	1043.88	2109777	GOOG
249	2018-12-28	1049.62	1055.56	1033.10	1037.08	1413772	GOOG
250	2018-12-31	1050.96	1052.70	1023.59	1035.61	1493722	GOOG

251 rows × 7 columns

```
netflix['Ticker'] = 'NFLX'
new_n = netflix
new_n
```

	date	open	high	low	close	volume	Ticker
0	2018-01-02	196.10	201.6500	195.4200	201.070	10966889	NFLX
1	2018-01-03	202.05	206.2100	201.5000	205.050	8591369	NFLX
2	2018-01-04	206.20	207.0500	204.0006	205.630	6029616	NFLX
3	2018-01-05	207.25	210.0200	205.5900	209.990	7033240	NFLX
4	2018-01-08	210.02	212.5000	208.4400	212.050	5580178	NFLX
...	...	...	...	...	...	...	...
246	2018-12-24	242.00	250.6500	233.6800	233.880	9547616	NFLX
247	2018-12-26	233.92	254.5000	231.2300	253.670	14402735	NFLX
248	2018-12-27	250.11	255.5900	240.1000	255.565	12235217	NFLX
249	2018-12-28	257.94	261.9144	249.8000	256.080	10987286	NFLX
250	2018-12-31	260.16	270.1001	260.0000	267.660	13508920	NFLX

251 rows × 7 columns

```
# 3. Append them together into a single dataframe
FAANG = new_f.append([new_aa, new_am, new_n, new_g])
FAANG
```

```
<ipython-input-40-50d17f77b9d0>:2: FutureWarning: The frame.append method is depreca
  FAANG = new_f.append([new_aa, new_am, new_n, new_g])
```

	date	open	high	low	close	volume	Ticker
0	2018-01-02	177.68	181.58	177.5500	181.42	18151903	FB
1	2018-01-03	181.88	184.78	181.3300	184.67	16886563	FB
2	2018-01-04	184.90	186.21	184.0996	184.33	13880896	FB
3	2018-01-05	185.59	186.90	184.9300	186.85	13574535	FB
4	2018-01-08	187.20	188.90	186.3300	188.28	17994726	FB
...	...	...	...	...	...	...	...
246	2018-12-24	973.90	1003.54	970.1100	976.22	1590328	GOOG
247	2018-12-26	989.01	1040.00	983.0000	1039.46	2373270	GOOG
248	2018-12-27	1017.15	1043.89	997.0000	1043.88	2109777	GOOG
249	2018-12-28	1049.62	1055.56	1033.1000	1037.08	1413772	GOOG
250	2018-12-31	1050.96	1052.70	1023.5900	1035.61	1493722	GOOG

1255 rows × 7 columns

Next steps: View recommended plots

```
FAANG.to_csv('/content/faang.csv',  index=False)
```

Exercise 2

- With faang, use type conversion to change the date column into a datetime and the volume column into integers. Then, sort by date and ticker.
- Find the seven rows with the highest value for volume.
- Right now, the data is somewhere between long and wide format. Use melt() to make it completely long format. Hint: date and ticker are our ID variables (they uniquely identify each row). We need to melt the rest so that we don't have

separate columns for open, high, low, close, and volume.

```
# With faang, use type conversion to change the date column into a datetime and
# the volume column into integers. Then, sort by date and ticker.
faang = p.read_csv('/content/faang.csv')
new_faang = faang.assign(
    date=p.to_datetime(faang.date),
    volume = faang['volume'].astype('int')
)
sorted_new_faang = new_faang.sort_values(by=['date', 'Ticker'])
sorted_new_faang
```

	date	open	high	low	close	volume	Ticker
251	2018-01-02	166.9271	169.0264	166.0442	168.9872	25555934	AAPL
502	2018-01-02	1172.0000	1190.0000	1170.5100	1189.0100	2694494	AMZN
0	2018-01-02	177.6800	181.5800	177.5500	181.4200	18151903	FB
1004	2018-01-02	1048.3400	1066.9400	1045.2300	1065.0000	1237564	GOOG
753	2018-01-02	196.1000	201.6500	195.4200	201.0700	10966889	NFLX
...	...	...	...	...	...	...	...
501	2018-12-31	157.8529	158.6794	155.8117	157.0663	35003466	AAPL
752	2018-12-31	1510.8000	1520.7600	1487.0000	1501.9700	6954507	AMZN
250	2018-12-31	134.4500	134.6400	129.9500	131.0900	24625308	FB
1254	2018-12-31	1050.9600	1052.7000	1023.5900	1035.6100	1493722	GOOG
1003	2018-12-31	260.1600	270.1001	260.0000	267.6600	13508920	NFLX

1255 rows × 7 columns

Next steps: View recommended plots

```
# Find the seven rows with the highest value for volume.
faang_max = faang.nlargest(n=7, columns='volume')
faang_max
```

	date	open	high	low	close	volume	Ticker
142	2018-07-26	174.8900	180.1300	173.7500	176.2600	169803668	FB
53	2018-03-20	167.4700	170.2000	161.9500	168.1500	129851768	FB
57	2018-03-26	160.8200	161.1000	149.0200	160.0600	126116634	FB
54	2018-03-21	164.8000	173.4000	163.3000	169.3900	106598834	FB
433	2018-09-21	219.0727	219.6482	215.6097	215.9768	96246748	AAPL
496	2018-12-21	156.1901	157.4845	148.9909	150.0862	95744384	AAPL
463	2018-11-02	207.9295	211.9978	203.8414	205.8755	91328654	AAPL

```
# Right now, the data is somewhere between long and wide format. Use melt() to make
# it completely long format. Hint: date and ticker are our ID variables (they uniquely identify each row).
# We need to melt the rest so that we don't have separate columns for open, high, low, close, and volume
melted_faang = new_faang.melt(
    id_vars = ['date', 'Ticker']
)
melted_faang
```

	date	Ticker	variable	value
0	2018-01-02	FB	open	177.68
1	2018-01-03	FB	open	181.88
2	2018-01-04	FB	open	184.90
3	2018-01-05	FB	open	185.59
4	2018-01-08	FB	open	187.20
...	...	...	...	...
6270	2018-12-24	GOOG	volume	1590328.00
6271	2018-12-26	GOOG	volume	2373270.00
6272	2018-12-27	GOOG	volume	2109777.00
6273	2018-12-28	GOOG	volume	1413772.00
6274	2018-12-31	GOOG	volume	1493722.00

6275 rows × 4 columns

Next steps: View recommended plots

Exercise 3

- Using web scraping, search for the list of the hospitals, their address and contact information. Save the list in a new csv file, hospitals.csv.
- Using the generated hospitals.csv, convert the csv file into pandas dataframe. Prepare the data using the necessary preprocessing techniques.

```
# Using web scraping, search for the list of the hospitals, their address and contact information.
# Save the list in a new csv file, hospitals.csv.
import requests
from bs4 import BeautifulSoup

url = 'https://en.wikipedia.org/wiki/List_of_hospitals_in_the_Philippines'

req = requests.get(url)
soup = BeautifulSoup(req.content, 'html.parser')
df = p.read_html(url)
dfs = df[0]
print(df)
```

```
   9                                     Region VII
   10                                   Region VIII
   11                                     Region IX
   12                                      Region X
   13                                     Region XI
   14                                    Region XII
   15                                   Region XIII

      .mw-parser-output .navbar{display:inline;font-size:88%;font-weight:normal}.mw-parser-output .navbar-coll
   0   Amang Rodriguez Memorial Medical Center East A...
   1   Baguio General Hospital and Medical Center Con...
   2   Ilocos Training and Regional Medical Center Ma...
   3   Batanes General Hospital Cagayan Valley Medica...
   4   Bataan General Hospital and Medical Center Dr....
   5   Batangas Medical Center Maria L. Eleazar Gener...
   6   Culion Sanitarium and General Hospital Ospital...
   7   Bicol Medical Center Bicol Region General Hosp...
   8   Corazon Locsin Montelibano Memorial Regional H...
   9   Don Emilio del Valle Memorial Hospital Eversle...
   10  Eastern Visayas Medical Center Governor Benjam...
   11  Basilan General Hospital Dr. Jose Rizal Memori...
   12  Amai Pakpak Medical Center Mayor Hilarion A. R...
   13  Davao Regional Medical Center Southern Philipp...
   14  Cotabato Regional and Medical Center Cotabato ...
   15  Adela Serra Ty Memorial Medical Center Caraga ...
   0                  Sovereign states
   1     States with limited recognition
   2  Dependencies and other territories
   3              Category  Asia portal

                        vteList of hospitals in Asia.1
   0  Afghanistan Armenia Azerbaijan Bahrain Banglad...
   1  Abkhazia Northern Cyprus Palestine South Osset...
   2  British Indian Ocean Territory Christmas Islan...
   3                             Category  Asia portal  ]
```

```python
dfs.to_csv('/content/hospitals.csv')
```

```python
# Using the generated hospitals.csv, convert the csv file into pandas dataframe.
# Prepare the data using the necessary preprocessing techniques.
hospitals = p.read_csv(
    '/content/hospitals.csv'
    )
hospitals.head()
```

	Unnamed: 0	Name of Hospital	Location	Class
0	0	Caloocan City Medical Center	450 A. Mabini St., Caloocan City	LGU
1	1	Ospital ng Malabon	F. Sevilla Boulevard, Tañong, Malabon City	LGU
2	2	San Lorenzo Ruiz General Hospital	O. Reyes St., Rosita Subdivision, Santulan, Ma...	DOH Retained

Next steps: View recommended plots

```python
hospitals.describe()
```

	Unnamed: 0
count	49.00000
mean	24.00000
std	14.28869
min	0.00000
25%	12.00000
50%	24.00000
75%	36.00000
max	48.00000

```
hospitals.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 49 entries, 0 to 48
Data columns (total 4 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   Unnamed: 0        49 non-null     int64 
 1   Name of Hospital  49 non-null     object
 2   Location          49 non-null     object
 3   Class             49 non-null     object
dtypes: int64(1), object(3)
memory usage: 1.7+ KB
```

```
hospitals.rename(
    columns={'Unnamed: 0' : 'Index'}, inplace=True
)
hospitals.head()
```

	Index	Name of Hospital	Location	Class
0	0	Caloocan City Medical Center	450 A. Mabini St., Caloocan City	LGU
1	1	Ospital ng Malabon	F. Sevilla Boulevard, Tañong, Malabon City	LGU
2	2	San Lorenzo Ruiz General Hospital	O. Reyes St., Rosita Subdivision, Santulan, Ma...	DOH Retained
3	3	Gat Andres Bonifacio Memorial Medical Center	8001 Delpan St., Tondo, Manila	LGU

Next steps: View recommended plots

```
contain_nulls = hospitals[
    hospitals.Index.isnull()
]
contain_nulls.shape[0]
```

```
0
```

```
hospitals_filled = hospitals.fillna(hospitals.mean)
hospitals_filled.head()
```

	Unnamed: 0	Name of Hospital	Location	Class
0	0	Caloocan City Medical Center	450 A. Mabini St., Caloocan City	LGU

Next steps: View recommended plots

7.2 Conclusion:

The accomplishment of this HoA has brought indispensable insight in the fundamentals of data analysis. The way that data could be collected in various methods albeit the difficulty of it depends on which methods you've used. I've only scratched the surface of web scraping so it was the most tedious part of the HoA. Overall, I think it's a great way to challenge what we've learned in the previous modules as well as an introduction in web scraping.