

# Pandas Part II and matplotlib Worksheet

March 8, 2022

Before you attempt these problems, please ensure that you have worked through the **NumPy worksheet** from Week 4, the **pandas Notes and worksheet from Week 5**, and have read and understood the **pandas Part II and matplotlib Notes** on the course website. Throughout, you may need to consult the pandas and matplotlib online documentation available at <https://pandas.pydata.org> and <https://matplotlib.org>, respectively.

For question 1, you will need to download the `wineData.csv` file from the course website and for question 3 you will need to download the `energyData.csv` file.

## 1. Some more pandas operations

- (i) From the `wineData` database, extract all the wines that cost less than \$500. Plot the ten most numerous wines in this subset on a bar plot.
- (ii) How many countries are in the database, and which country has the most wines in the database? Using the `.groupby()` function, find the mean rating (points) for the wines of each country in the database.

## 2. Merging DataFrames

*Merging* is a really useful tool for data analysis when we wish to combine two DataFrames containing different information that have an common column or common indices.

- (i) First create the following DataFrames (for the High Score column in `df2` you can use any numbers you like):

```
[34]: df1
```

```
[34]:
```

|   | User | Forename  | Surname   | Nationality   |
|---|------|-----------|-----------|---------------|
| 0 | 001  | Alfie     | Alfredson | Swedish       |
| 1 | 002  | Brian     | Boggs     | American      |
| 2 | 003  | Christine | Cattrell  | Canadian      |
| 3 | 004  | Diana     | Dash      | English       |
| 4 | 005  | Euan      | Everglade | Scottish      |
| 5 | 006  | Felicity  | Furnish   | English       |
| 6 | 007  | George    | Gallop    | South African |
| 7 | 008  | Henry     | Hoover    | Ireland       |

```
[36]: df2
```

```
[36]:
```

|    | User | Year | High Score |
|----|------|------|------------|
| 0  | 001  | 2021 | 1729       |
| 1  | 001  | 2021 | 1110       |
| 2  | 002  | 2021 | 1968       |
| 3  | 002  | 2021 | 1853       |
| 4  | 003  | 2021 | 1544       |
| 5  | 003  | 2021 | 1562       |
| 6  | 004  | 2021 | 1246       |
| 7  | 004  | 2022 | 1581       |
| 8  | 005  | 2022 | 1146       |
| 9  | 005  | 2022 | 1239       |
| 10 | 006  | 2022 | 1087       |
| 11 | 006  | 2022 | 1502       |
| 12 | 007  | 2022 | 1604       |
| 13 | 007  | 2022 | 1584       |

- (ii) As we can see, we have two DataFrames that both contain the column `User`. We can use these to `merge` the DataFrames into one. Read the pandas `.merge()` documentation to merge `df1` and `df2` on the `User` column. What do you notice with regards to the original two DataFrames?
- (iii) Use the option `how = 'left'` in the `merge` function to include *all* the values from both `df1` and `df2`. What happens if we do `how = 'right'`? And finally, what about `how = 'outer'` and `how = 'inner'`?

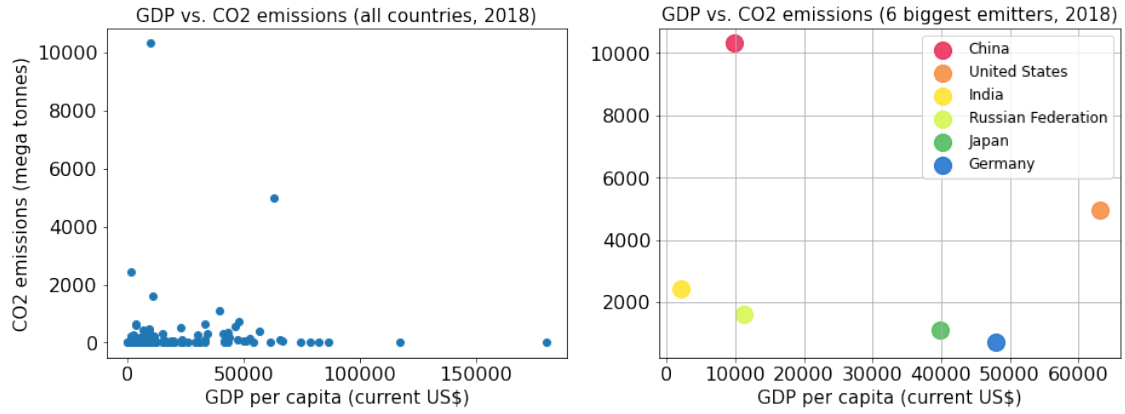
### 3. Relationship between GDP and carbon footprint

In this question, you will produce an informative set of figures, detailing information regarding the relationship between GDP and CO2 emissions per capita.

- (i) First, explore the data. For example, how many countries are in the data set? What does the `head` and `tail` of the DataFrame look like? Are there `NaN` values that can be safely removed? Any other strange looking values? Can these be replaced? Before you move on, delete the `Series Code` column.
- (ii) Create a new DataFrame that contains the columns `Country Name`, `CO2 emissions 2018` (`CO2 emissions (kt)`), and `GDP per capita 2018` (`GDP per capita (current US$)`), where the data type of `CO2 emissions 2018` and `GDP per capita 2018` is `float64`. (*Hint:* You may find the `.to_numeric()` or `.astype()` functions useful.)

Use your new data frame to construct two scatter plots, similar to the ones shown below. Save your figure as a `.png` file.

[265]:



- (iii) Can you modify your the second subplot from part (ii) so that the *marker size* for each data point scales with the amount of *Fossil fuel energy consumption*? Export this new figure as another *.png* file.