Programming for Artificial Intelligence (Python)
Homework 8
Due: June 7 before class

This homework focuses on working with the `pandas` library.

1. China's GDP by province

The National Bureau of Statistics of China releases information about the gross domestic product (GDP) on their official website (stats.gov.cn). The file `GDP_by_province.csv` contains the GDP of 31 China's provinces (Hong Kong, Macao, and Taiwan not included). Read in the data set and do the following analysis.

1.1 The GDP data

In econometrics, we typically say a data is cross-sectional when all observations of the data set are taken in the same time. We say a data is a time series when the data is taken from the same person over time. We say a data is a panel data, when the data is taken from multiple time points, and in each time point, the data set contains sereral observations.

Read in `GDP_by_province.csv`. What is the type of this data? **Hint: Do not forget to go through the data carefully and make necessary changes.**

1.2 Wide to long

Convert the GDP data to a long format so that each province, combined with each year, forms one observation. For example, the following is the desired format. Make sure to give the long table appropriate column names.

	Province	Year	GDP
0	Beijing	2022	41540.9
1	Tianjing	2022	16132.2
2	Hebei	2022	41988.0
3	Shanxi	2022	25583.9
4	Neimenggu	2022	23388.9
...	...	...	...
283	Gansu	2014	6518.4
284	Qinghai	201	
285	Ni		

1.3 The `shift` method of `pandas`

We introduced several methods in class, e.g., `mean`, `sum`, `agg`, etc. There is another useful method of `pandas`: `shift`. The following example shows you how it works:

```
age_data = pd.DataFrame({
  "id" : [1, 1, 1, 1, 1],
  "age" : [23, 24, 25, 26, 27, ]
})
age_data["last_age"] = age_data.age.shift(periods=1)
```

Note that the `shift` method takes the "periods" parameter. Read more about this parameter from `https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.shift.html`.

When using the `shift` method, we might encounter two traps that need our caution:

1. Trap 1: Suppose our data now changed. Would "last_age" be meaningful?

```
age_data1 = pd.DataFrame({
  "id" : [1, 1, 1, 1, 1],
  "year": [2002, 2004, 2003, 2005, 2006],
  "age" : [23, 25, 24, 26, 27, ]
})
age_data1["last_age"] = age_data1.age.shift(1)
```

2. Trap 2: Suppose we have one more person, whould "last_age" be correct always?

```
age_data2 = pd.DataFrame({
  "id" : [1, 1, 1, 1, 1, 2, 2, 2, 2, 2],
  "year": [2002, 2004, 2003, 2005, 2006, 2002, 2004,
    2003, 2005, 2006],
  "age" : [23, 25, 24, 26, 27, 17, 19, 18, 20, 21]
})
age_data2["last_age"] = age_data2.age.shift(1)
```

Please answer the following two questions:

1. Make changes before using `shift` in `age_data1` so that `shift` can work properly. **Hint: when we first introduced `pandas` in `pandas1.ipynb`, there is a method called `sort_values`.**

2. Make changes before using `shift` in `age_data2` so that `shift` can work properly.

1.4 The average increments in GDP

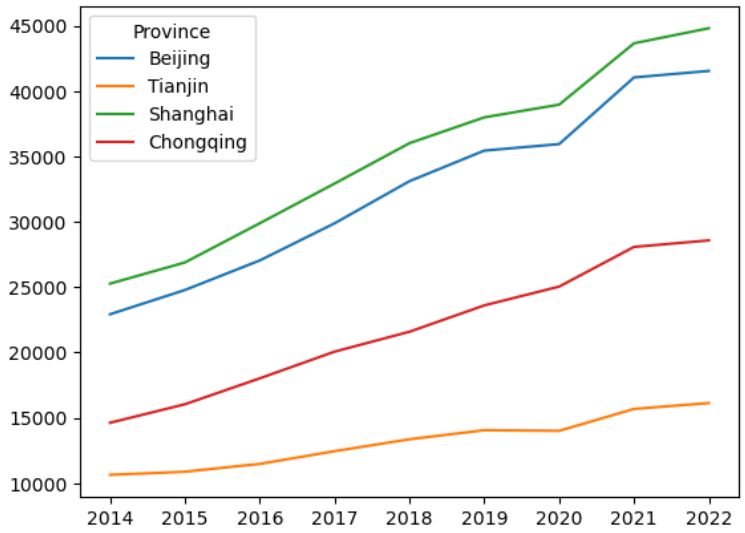
Now it's time to compute the GDP of last year. You can make a new column "last_GDP." Then compute the difference between "GDP" and "last_GDP" to get the increment.

Then compute the average increments of GDP for each province.

1.5 Plotting the GDP of four cities

In one figure, plot the GDP of the four Municipality directly under the Central Government (Beijing, Chongqing, Shanghai, Tianjin).

Hint: This question may be more difficult than it seems. Do whatever you can do as long as you can make the plot as in the following.



1.6 GDP and consumption

The file "consumption_by_province.csv" contains the annual consumption of all provinces from stats.gov.cn. Read in the file. Combine this data with the table you've made in Question 1.4.

The consumptions data contain many missing values, remove them from the combined data set before moving on.

Will last year's GDP influence the consumption of this year? analyze the influence by year. **Hint: regression (OLS) is enough.**

1.7 GDP by region

The file "region.csv" contains information about China's geographical regions. Make a two-way table to show the average GDP of each region in each year. Check your answer with the following code

```
contribution.loc[:, ["Region", "Year", "percentage"]].
   groupby(["Region", "Year"]).sum()
```

1.8 Provinces' contributions

Lastly, we want to know in each year, the contributions of the provinces to the region's GDP. For example, in 2014, Anhui contributed about 9.23% in GDP to the East part of China.

1. Make a table to summarise the contributions by year and region.
2. Make a picture to show the bar charts of contributions for the East China region in years 2020, 2021, and 2022, as follows: