Full Stack Development with AI
Lab 7.2 – Pandas

Lab Overview

In the previous Lab 7.1, you have performed some useful numerical processing tasks with NumPy using a synthetic e-commerce dataset. The dataset contains only customers' purchase amounts, which are all floating-point numbers. Real-world datasets are rarely homogeneous in terms of data type. More often than not, real-world datasets contain mixed data types. For example, e-commerce datasets would likely also contain the demographic profile of customers such as age, gender and employment status as well as product information such as product category and brand. Clearly, some of these variables could be non-numerical or categorical in nature. In such cases, NumPy would not be able to handle them. This is where Pandas comes in useful.

In this lab, you will learn how to perform data preparation and data preprocessing using Pandas. Recall that Pandas is built on top of NumPy. Pandas's `Series` object extends from NumPy's `ndarray` to support labelled data but is still homogeneous. Pandas's `DataFrame` object further extends from `Series` to support two-dimensional dataset of mixed data types. That is, each column or `Series` in a `DataFrame` can be of different data types.

Similar to Lab 7.1, you are strongly encouraged to use the Jupyter Notebook format for this lab.

Exercise 1 – Basic Pandas Operations

This question is based on the occupation dataset hosted on GitHub – https://github.com/justmarkham/DAT8/blob/master/data/u.user. The actual data file is provided in data.zip as occupation.csv.

You may refer to the Pandas Documentation's API Reference here if you need any assistance – https://pandas.pydata.org/pandas-docs/stable/reference/index.html

Perform each of the following tasks and report the answer, if applicable. Note that there can be multiple different ways of performing any single task. You are encouraged to explore Pandas on your own.

a) Create a new Jupyter notebook with the file extension .ipynb and import the necessary libraries.

b) Import the occupation dataset from the occupation.csv file provided as a `DataFrame` with user_id as the index and assign it to a suitably named variable.

c) Print out the first 25 rows.

d) Print out the last 10 rows.

e) What is the number of observations in the dataset?

f) What is the number of columns in the dataset?

g) Print out the name of all the columns.

h) How is the dataset indexed?

i) What is the data type of each column?

j) Print out only the occupation column.

k) How many different occupations are there in this dataset?

l) Print out the list of users aged 50 and above. How many of such users are there?

m) Print out the descriptive statistics of the `DataFrame`.

n) Print out the descriptive statistics for all columns in the `DataFrame`.

o) Print out the descriptive statistics only for the occupation column.

p) What is the mean age of users?

q) What is the age with the least occurrence?

r) Add a new salary column to the `DataFrame` with an initial value of 0.

s) Set the salary of each user to 100 multiplied by the user's age. For example, the salary for observation 1 would be 24 * 100 = 2400.

   Print out the `DataFrame` to show the computed salary for all users.

Exercise 2 – Advanced Pandas Operations

The folder "forbes-global-2000-2008-2019" in data.zip contains a series of datasets in CSV (Comma-Separated Values) format representing the Forbes 2000 list of global companies for a 15-year period from year 2008 to 2022. Each dataset contains the following 8 variables:

Variable	Description
Rank_nr	The ranking of the company.
Company	The name of the company.
Industry	The industry of the company's primary business operations.
Country	The country the company is situated in.
Sales	The amount of sales of the company in thousand of million US dollars.
Profits	The profit of the company in thousand of million US dollars.
Assets	The assets of the company in thousand of million US dollars.
Market_Value	The market value of the company in thousand of million US dollars.

Observe that each individual dataset is essentially cross-sectional in nature, representing the financial data for multiple companies on the Forbes list in one year. An equivalent time series dataset would contain the financial data for only one company but across multiple years. Finally, a panel dataset would contain the financial data for multiple companies across multiple years.

Perform the following tasks with the Forbes datasets:

1. Construct a combined panel dataset of the Forbes 2000 list of companies from year 2008 to 2022. The dataset should include a new variable Year that denotes the calendar year of each observation in addition to the existing 8 variables. Note that each observation in the new panel dataset will represent the financial data of a company for a particular year.

   State the number of observations in your combined panel dataset. Is the number within your expectation? Why?

   Export the combined panel dataset to a file in CSV format.

```
df_panel
```
✓ 0.0s

	Year	Rank_nr	Company	Industry	Country	Sales	Profits	Assets	Market_Value
0	2008	1	HSBC Holdings	Banking	United Kingdom	146500.0	19130	2348980.0	180810.0
1	2008	2	General Electric	Conglomerates	United States	172740.0	22210	795340.0	330930.0
2	2008	3	Bank of America	Banking	United States	119190.0	14980	1715750.0	176530.0
3	2008	4	JPMorgan Chase	Banking	United States	116350.0	15370	1562150.0	136880.0
4	2008	5	ExxonMobil	Oil & Gas Operations	United States	358600.0	40610	242080.0	465510.0
...	...	...	...	...	...	...	...	...	...
29988	2022	1995	Shenzhen Feima International Supply Chain	Business Services & Supplies	China	37.0	1408.3	166.0	1136.0
29989	2022	1997	NMDC	Materials	India	3520.0	1406.4	5715.0	6401.0
29990	2022	1997	Sichuan Changhong Electric	Consumer Durables	China	15716.0	53.1	12105.0	1957.0
29991	2022	1999	Satellite Chemical	Chemicals	China	4413.0	931.3	7640.0	9521.0
29992	2022	2000	Sun Communities	Diversified Financials	United States	2273.0	375.7	13494.0	21714.0

2. Construct individual time series datasets of the Forbes 2000 list of companies from year 2008 to 2022, i.e., each dataset should only contain the financial data for one company across multiple years. Sort the observations in each dataset by Year in ascending order.

   How long did it take you to complete the time series datasets generation process?

   State the number of companies and for each company, the number of observations in its own time series dataset. Is it always the case that every company has 15 observations or equivalently 15 years of financial data from year 2008 to 2022? Explain your observations of the output.

```
Total number of companies: [redacted]
360 Security Technology 3
3M 15
3i Group 8
77 Bank 15
A2A 10
AAC Technologies Holdings 4
AB Sagax 2
ABB 15
ABK 1
ACC 1
ACE 8
ACE Aviation 2
```

3. Carefully inspect the individual time series datasets that you have generated in (2). Did you observe any problem? For example, were there any duplicate companies?

   Fixed the duplicate companies problem and report your observations of the new output. Do the figures make more sense now?

   Were you able to fix the duplicate companies completely? Why? Think about what additional steps you would need to take to resolve this problem completely.

   Do also take note that to perform time series analysis on a particular company, you would require the financial data for that company over a contiguous temporal period. For example, companies that had been consistently placed on the Forbes 2000 list throughout the 15-year period would fulfil this requirement. If there are missing periods, you would not be able to perform the time series analysis properly.

Exercise 3 – Data Analysis with Pandas

The tasks that you have performed in Exercise 2 are more of data preparation and data preprocessing. In this exercise, you will perform some data analysis tasks on the latest year 2022 Forbes data using Pandas.

1. Import the combined panel dataset that you have generated in Exercise 2 into a Pandas DataFrame.

   Filter the financial data for Forbes 2000 companies in year 2022. State the number of observations.

2. Generate a data quality report as shown in the figure below:

	Year	Rank_nr	Company	Industry	Country	Sales	Profits	Assets	Market_Value
count	2000.0	2000	2000	2000	2000	2000.0	2000.0	2000.0	2000.0
unique	NaN	1685	2000	26	58	NaN	NaN	NaN	NaN
top	NaN	1949	Berkshire Hathaway	Banking	United States	NaN	NaN	NaN	NaN
freq	NaN	4	1	290	584	NaN	NaN	NaN	NaN
mean	2022.0	NaN	NaN	NaN	NaN	23886.202	2500.6554	117114.2185	38238.4255
std	0.0	NaN	NaN	NaN	NaN	40942.544294	6159.244654	363745.753098	117207.269875
min	2022.0	NaN	NaN	NaN	NaN	0.0	-12052.3	166.0	144.0
25%	2022.0	NaN	NaN	NaN	NaN	5604.0	589.85	14790.75	7325.75
50%	2022.0	NaN	NaN	NaN	NaN	12167.5	1054.75	32007.0	16625.5
75%	2022.0	NaN	NaN	NaN	NaN	23742.5	2266.825	77870.25	34712.0
max	2022.0	NaN	NaN	NaN	NaN	572754.0	105363.0	5518508.0	2640316.0
Data Type	int64	object	object	object	object	float64	float64	float64	float64
Missing Values									
Present Values									

   Can you identify some differences between the output of the data quality report as compared to the output from the `describe()` *function?*

3. Are there any missing values in the dataset? If yes, drop all rows with missing values in the dataset and regenerate the data quality report.

4. Plot a histogram of Market_Value using the `hist()` function and describe the skewness of Market_Value. Calculate a suitable measure of skewness with Pandas and interpret the result. In particular, state whether the calculated measure is congruent with the histogram.

5. Transform the data with Pandas to resolve the skewness problem in (4). Thereafter, replot the histogram and recalculate the measure of skewness.

   ***Hint**: You need to use the NumPy library for the transformation.*

6. Suppose we want to predict Market_Value using linear regression analysis. Determine which other numerical variables (i.e., Sales, Profits and Assets) are useful independent variables for predicting Market_Value.

   Among the useful independent variable(s), which is likely to be the best predictor of Market_Value?

   ***Hint**: What statistical information would be useful to assess the relationships between the independent variables and the dependent variable?*

7. Suppose we want to convert the regression problem in (6) into a binary classification problem, i.e., predict whether Market_Value is low or high. Make the necessary changes to the dataset to enable this classification task.

   ***Hint**: We need to discretize the dependent variable into a categorical variable with two labels. There are different techniques to perform the required discretization, and you can use any technique.*

-- End of Lab --