Module 6: Data Manipulation Using Pandas

Assignment - II

edureka!

1. You are given a data set, which is present in the LMS, containing the number of hurricanes occurring in the United States along the coast of the Atlantic. Load the data from the dataset into your program and plot a Bar Graph of the data, taking the Year as the x-axis and the number of hurricanes occurring as the Y-axis.

2. The dataset given, records data of city temperatures over the years 2014 and 2015. Plot the histogram of the temperatures over this period for the cities of San Francisco and Moscow.

3. Plot a piechart of the number of models released by every manufacturer, recorded in the data provided. Also mention the name of the manufacturer with the largest releases.

4. Create csv file from the data below and read it in the pandas's dataframe.

Phase 1: Reading data

Phase 2: Describe the data

Describe the data on the **unit price.**

Phase 3: Filter the data

Create a new dataframe having columns 'name','net_price','date' and group all the records according to name.

Phase 4: Plotting graph

Plot the graph after calculating total sales by each customer. Customer name should be the on the x-axis and total sales on y-axis. Refer to sample-salesv2 to answer the following questions

5. Let the x-axis data points and y-axis data points are

X = [1,2,3,4]
y = [20, 21, 20.5, 20.8]

5.1: Draw a Simple plot

5.2: Configure the line and markers in a simple plot

5.3: configure the axes

5.4: Give the title of the graph and labels of the x-axis and y-axis

5.5: Give error bar if y_error = [0.12, 0.13, 0.2, 0.1]

5.6: define width, and height as figsize=(4,5) DPI and adjust plot dpi=100

5.7: Give a font size of 14

5.8: Draw a scatter graph of any 50 random values of the x and y-axis

5.9: Create a dataframe from the following data

'first_name': ['Jason', 'Molly', 'Tina', 'Jake', 'Amy'],
'last_name': ['Miller', 'Jacobson', 'Ali', 'Milner', 'Cooze'],
'female': [0, 1, 1, 0, 1],
'age': [42, 52, 36, 24, 73],
'preTestScore': [4, 24, 31, 2, 3],
'postTestScore': [25, 94, 57, 62, 70]

Draw a Scatterplot of preTestScore and postTestScore, with the size of each point determined by age.

5.10: Draw a Scatterplot from the data in question 9 of preTestScore and postTestScore with the size = 300 and the color determined by sex.

6. Case Study I

Domain – Education

Focus – Data analysis

Business challenge/requirement
You are a data analyst with the University of California. The University has data of Maths, Physics and Data Structure score of sophomore students. This data is stored in different files. The University has hired a data science company to do analysis of scores and find if there is any correlation of score with age, ethnicity, etc. Before the data is given to the company, you are required to do data wrangling.

Key issues

Ensure student identity is not revealed to the agency and only relevant data is shared.

Considerations

NONE

Data volume

Only around 1800 records are shared in files MathScoreTerm1.csv, DSScoreTerm1.csv, and PhysicsScoreTerm1.csv.

Additional information

NA

Business benefits

The University can enrol more students by improving its international ranking through a personalized course/curriculum for students

Approach to Solve

You have to use the fundamentals of NumPy and pandas covered in module 5 and 6.

1. Read the three CSV files, which contain the score of the same students in term1 for each Subject
2. Remove the name and ethnicity column (to ensure confidentiality)
3. Fill missing score data with zero
4. Merge the three files
5. Change Sex(M/F) Column to 1/2 for further analysis
6. Store the data in a new file – ScoreFinal.csv

Enhancements for code

You can try these enhancements in code

1. Convert ethnicity to a a numerical value
2. Fill the missing score for a student to the average of the class