Keshav Mahavidyalaya

B.Sc. (H) Computer Science - III Sem (Section A+B)

DSE-I Data Analysis and Visualization (DAV)

Assignment 2

Instructions

- *Note the deadline for submission. Last date of submission is* ***20 Nov 2023****.Late submissions may incur penalties.*
- *Ensure your assignment is submitted in PDF format , with the questions and answers presented as source code accompanied by proper comments. Additionally, include the output in the document. Moreover, submit the source code for each question as a separate '.ipynb' file, named* ***'question_no_1.ipynb'****.*
- *Upload pdf and source code in google class room and also submit hard copy of pdf in lecture class.*
- *Ensure that variable names adhere to the proper naming convention, specifically using snake cases. For example, use* ***variable*** *names like* ***'total_number_of_questions'*** *that clearly convey the meaning of the variable. Additionally, provide proper comments for each function and line of code.*
- *Carefully read and understand the assignment prompt or guidelines provided.*
- *Ensure that your work is original and properly cited. Plagiarism is a serious offence in academia.*
- *Maintain a professional tone and approach in your writing.*
- *Organise your assignment with a clear introduction, body, and conclusion. Use headings and subheadings if necessary.*

Q1. Use a dataset of your choice from Open Data Portal (https:// data.gov.in/, UCI repository) .
Load a Pandas dataframe with a selected dataset. Identify and count the missing values in a dataframe. Clean the data after removing noise as follows

a) Drop duplicate rows.
b) Detect the outliers and remove the rows having outliers
c) Identify the most correlated positively correlated attributes and negatively correlated attributes

Q2. Given below is a dictionary having two keys 'Boys' and 'Girls' and having two lists of heights of five Boys and Five Girls respectively as values associated with these keys Original dictionary of lists:
{'Boys': [72, 68, 70, 69, 74], 'Girls': [63, 65, 69, 62, 61]}
From the given dictionary of lists create the following list of dictionaries:
[{'Boys': 72, 'Girls': 63}, {'Boys': 68, 'Girls': 65}, {'Boys': 70, 'Girls': 69}, {'Boys': 69, 'Girls': 62}, {'Boys':74, 'Girls':61] . What are multiple ways to do it? Give at least 3 methods to achieve it. Explain each method as the comment of your code.

Q3. Create a dataframe having at least 5 columns and 100 rows to store numeric data generated using a random function. Replace 25% of the values by null values whose index positions are generated using random function. Do the following:

a. Identify and count missing values in a dataframe.
b. Drop the column having more than 5 null values.
c. Identify the row label having maximum of the sum of all values in a row and drop that row.
d. Sort the data frame on the basis of the first column.
e. Remove all duplicates from the first column.
f. Find the correlation between first and second column and covariance between second and third column.
g. Detect the outliers and remove the rows having outliers.
h. Discretize second column and create 5 bins

Q4. Consider two excel files having attendance of a workshop's participants for two days. Each file has three fields 'Name', 'Time of joining', duration (in minutes) where names are unique within a file. Note that duration may take one of three values (30, 40, 50) only. Import the data into two dataframes and do the following:

a. Perform merging of the two dataframes to find the names of students who had attended the workshop on both days.

b. Find names of all students who have attended workshop on either of the days.

c. Merge two data frames row-wise and find the total number of records in the data frame.

d. Merge two data frames and use two columns names and duration as multi-row indexes. Generate descriptive statistics for this multi-index.

Q5. Consider a data frame containing data about students i.e. name, gender and passing division:

	Name	Birth_Month	Gender	Pass_Division
0	Mudit Chauhan	December	M	III
1	Seema Chopra	January	F	II
2	Rani Gupta	March	F	I
3	Aditya Narayan	October	M	I
4	Sanjeev Sahni	February	M	II
5	Prakash Kumar	December	M	III
6	Ritu Agarwal	September	F	I
7	Akshay Goel	August	M	I
8	Meeta Kulkarni	July	F	II
9	Preeti Ahuja	November	F	II
10	Sunil Das Gupta	April	M	III
11	Sonali Sapre	January	F	I
12	Rashmi Talwar	June	F	III
13	Ashish Dubey	May	M	II
14	Kiran Sharma	February	F	II
15	Sameer Bansal	October	M	I

a. Perform one hot encoding of the last two columns of categorical data using the get_dummies() function.

b. Sort this data frame on the "Birth Month" column (i.e. January to December). (Hint: Convert Month to Categorical.)

Q6. Consider the following data frame containing a family name, gender of the family member and her/his monthly income in each record.

FamilyName	Gender	MonthlyIncome (Rs.)
Shah	Male	44000.00
Vats	Male	65000.00
Vats	Female	43150.00
Kumar	Female	66500.00
Vats	Female	255000.00
Kumar	Male	103000.00
Shah	Male	55000.00
Shah	Female	112400.00
Kumar	Female	81030.00
Vats	Male	71900.00

Write a program in Python using Pandas to perform the following:

A. Calculate and display familywise gross monthly income.

B. Display the highest and lowest monthly income for each family name.

C. Calculate and display monthly income of all members earning income less than Rs. 80000.00.

D. Calculate and display the average monthly income of the female members in the Shah family.
E. Calculate and display monthly income of all members with income greater than Rs. 60000.00.
F. Display total number of females along with their average monthly income.
G. Delete rows with Monthly income less than the average income of all members

Q7. Using the **parsed.csv** file, complete the following exercises to practise your pandas skills:

a. Find the 95th percentile of earthquake magnitude in Japan using the magType of 'mb'.
b. Find the percentage of earthquakes in Indonesia that were coupled with tsunamis.
c. Get summary statistics for earthquakes in Nevada.
d. Add a column to the dataframe indicating whether or not the earthquake happened in a country or US state that is on the Ring of Fire. Use Bolivia, Chile, Ecuador, Peru, Costa Rica, Guatemala, Mexico (be careful not to select New Mexico), Japan, Philippines, Indonesia, New Zealand, Antarctica (look for Antarctic), Canada, Fiji, Alaska, Washington, California, Russia, Taiwan, Tonga, and Kermadec Islands.
e. Calculate the number of earthquakes in the Ring of Fire locations and the number outside them.
f. Find the tsunami count along the Ring of Fire.

Q8. Using the CSV files in the **earthquakes.csv** folder, Write a program in Python using Pandas to perform the following: :

a. With the **earthquakes.csv** file, select all the earthquakes in Japan with a magType of mb and a magnitude of 4.9 or greater.
b. Create bins for each full number of magnitude (for example, the first bin is 0-1, the second is 1-2, and so on) with a **magType** of **ml** and count how many are in each bin.
c. Build a crosstab with the earthquake data between the **tsunami column** and the **magType column**. Rather than showing the **frequency** count, show the **maximum magnitude** that was observed for each combination. Put the **magType** along the columns.

Q9. Using the **faang.csv** file, group by the ticker and resample to monthly frequency. Make the following aggregations:

a. Mean of the opening price
b. Maximum of the high price
c. Minimum of the low price
d. Mean of the closing price
e. Sum of the volume traded