Chapter - 3
Data Handling using Pandas - II

Que 1. Write the statement to install the python connector to connect MySQL i.e. pymysql.

Ans. pip install pymysql

Que 2. Explain the difference between pivot() and pivot_ table() function?

Ans. pivot() :

The pivot function is used to reshape and create a new DataFrame from the original one.

pivot1=df.pivot(index='Store',columns='Year',va lues='Total_sales(Rs)')

Index specifies the columns that will be acting as an index in the pivot table, columns specifies the new columns for the pivoted data and values specifies columns whose values will be displayed.

pivot_table() :

It works like a pivot function, but aggregates the values from rows with duplicate entries for the specified columns. In other words, we can use aggregate functions like min, max, mean etc, wherever we have duplicate entries.

The default aggregate function is mean.

Syntax:
pandas.pivot_table(data, values=None, index=None, columns=None, aggfunc='mean')

The parameter aggfunc can have values among sum, max, min, len, np.mean, np.median.

pivot( ) vs pivot_table()

The pivot() and pivot_table() both perform data pivoting a data set.

If there are multiple entries for a column's value for the same values for index (row), then pivot() leads to an error.

The pivot_table() pivots the data by aggregating it, thus it can work with duplicate entries.

Que 3. What is sqlalchemy?

Ans. sqlalchemy is a library used to interact with the MySQL database by providing the required credentials. This library can be installed using the following command:

pip install sqlalchemy

Once it is installed, sqlalchemy provides a function create_engine() that enables this connection to be established.

Que 4. Can you sort a DataFrame with respect to multiple columns?

Ans. Yes
Sorting refers to the arrangement of data elements in a specified order, which can either be ascending or descending.

Pandas provide sort_values() function to sort the data values of a DataFrame.

The syntax of the function is as follows:

DataFrame.sort_values(by, axis=0, ascending=True)
Here, a column list (by), axis arguments (0 for rows and 1 for columns) and the order of sorting (ascending = False or True) are passed as arguments. By default, sorting is done on row indexes in ascending order.

Sorting on single column:
df.sort_values(by=['Name'], ascending = False)
Sorting on multiple columns:
dfUT3.sort_values(by=['Science','Hindi'])

Que 5. What are missing values? What are the strategies to handle them?

Ans. Missing values (data) means no information is provided for one or more items or for a whole unit. Pandas puts NaN in place of missing data in dataframes.

Missing data is a big problem in doing calculations, because NaN makes the whole calculation result is NaN. In Pandas, there are some functions, which helps in handling of missing data. These functions are – isnull(), fillna().

Que 6. Define the following terms: Median, Standard Deviation and variance.

Ans. (a) **Median** : Median is mid value in an ordered data set.

(b) **Standard Deviation**: It is a measure of dispersion of observations within dataset relative to their mean.lt is square root of the variance and denoted by Sigma (o).

(c) **Variance**: Variance is the numerical values that describe the variability of the observations from its arithmetic mean and denoted by sigma-

squared. **Variance** measures how far individuals in the group are spread out, in the set of data from the mean.

Que 7. What do you understand by the term MODE? Name the function which is used to calculate it.

Ans. Mode is the number which occurs most often in a data set.

The mode( ) is used to calculate it.

Que 8. Write the purpose of Data aggregation.

Ans. Data Aggregation is a process of producing a summary statistics from a dataset using statistical aggregation functions.

Que 9. Explain the concept of GROUP BY with help on an example.

Ans. The groupby( ) allows to create field wise group of values in a dataframe as per a specific aggregate function.

In pandas, DataFrame.GROUP BY() function is used to split the data into groups based on some criteria. Pandas objects like a DataFrame can be split on any of their axes.

The GROUP BY function works based on a split-apply-combine strategy which is shown below using a 3-step process:

Step 1: Split the data into groups by creating a GROUP BY object from the original DataFrame.
Step 2: Apply the required function.
Step 3: Combine the results to form a new DataFrame.

Que 10. Write the steps required to read data from a MySQL database to a DataFrame.

Ans. Do your self

Que 11. Explain the importance of reshaping of data with an example.

Ans. Do your self

Que 12. Why estimation is an important concept in data analysis?

Ans. Do your self

Que 13. Assuming the given table: Product. Write the python code for the following:

Item	Company	Rupees	USD
TV	LG	12000	700
TV	VIDEOCON	10000	650
TV	LG	15000	800
AC	SONY	14000	750

a) To create the data frame for the above table.

b) To add the new rows in the data frame.

c) To display the maximum price of LG TV.

d) To display the Sum of all products.

e) To display the median of the USD of Sony products.

f) To sort the data according to the Rupees and transfer the data to MySQL.

g) To transfer the new dataframe into the MySQL with new values.

Ans. Do your self

Que 14. Write the python statement for the following question on the basis of given dataset:

	Name	Degree	Score
0	Aparna	MBA	90.0
1	Pankaj	BCA	NaN
2	Ram	M.Tech	80.0
3	Ramesh	MBA	98.0
4	Naveen	NaN	97.0
5	Krrishnav	BCA	78.0
6	Bhawna	MBA	89.0

a) To create the above DataFrame.

b) To print the Degree and maximum marks in each stream.

c) To fill the NaN with 76.

d) To set the index to Name.

e) To display the name and degree wise average marks of each student.

f) To count the number of students in MBA.

g) To print the mode marks BCA.

Ans. Do your self