**CS352                                     Data Analysis Lab**

**Course Objectives:**

The main objectives of the course are to:

1. Introduce Python libraries used for data manipulation and visualization
2. Create awareness on data cleaning, wrangling and various operations on data
3. Impart knowledge on visualizing the data using various plots

**Course Outcomes:**

On successful completion of the course, students will be able to:

1. Perform operations on data using basic concepts of Numpy and Pandas
2. Perform Data cleaning and Data wrangling operations
3. Visualize data using the tool  Matplotlib
4. Perform operations on aggregations and time series data

**Course Content:**

**UNIT-I**

**NumPy Basics: Arrays and Vectorized Computation**The NumPyndarray, Universal Functions, Array-Oriented Programming with Arrays, File Input and Output with Arrays, Linear Algebra, Pseudorandom Number Generation, Example: Random Walks

**Pandas Data Structure:** Introduction to pandas Data Structure, Essential Functionality, Summarizing and Computing Descriptive Statistics

**UNIT-II**

**Data Loading, Storage, and File Formats:** Reading and Writing Data in Text Format, Binary Data Formats, Interacting with Web APIs, Interacting with Databases.

**Data Cleaning and Preparation:** Handling Missing Data, Data Transformation, String Manipulation.

**UNIT-III**

**Data Wrangling: Join, Combine, and Reshape:** Hierarchical Indexing, Combining and Merging Datasets, Reshaping and Pivoting

**Plotting and Visualization:** A Brief matplotlib API Primer, Plotting with pandas and seaborn.

**UNIT-IV**

**Data Aggregation and Group Operations:** Group By Mechanics, Data Aggregation, Apply: General split-apply-combine, Pivot Tables and Cross-Tabulation.

**Time Series:** Date and Time Data Types and Tools, Time Series Basics, Date Ranges, Frequencies, and Shifting, Time Zone Handling, Periods and Period Arithmetic, Resampling and Frequency Conversion, Moving Window Functions

**Learning Resources:**

**Textbook(s):**

1. Wes McKinney, Python for Data Analysis - Data Wrangling with Pandas, NumPy, and IPython 2nd Edition. O’Reilly/SPD

**References:**

1. Jake VanderPlas, Python Data Science Handbook Essential Tools for Working with Data. O’Reilly/SPD
2. David Taieb ,”Data Analysis with Python: A Modern Approach “ 1st Edition, Packt Publishing

**List of Experiments:**

1. Numpy Array operations
2. Iris Dataset
3. Pandas Series
4. Pandas Dataframes
5. Canada Pizza Price Prediction
6. Mobile Phone Price Data set
7. National Universities Rankings.
8. Adidas Sales Dataset
9. Movies Dataset.
10. Avocado Prices

**1.Numpy Array operations**

Write a Python program to do the following operations: Library: NumPy

1. Create a one-dimensional array and perform all operations on it.
2. Create multi-dimensional arrays and find its shape and dimension
3. Create a matrix full of zeros and ones
4. Reshape and flatten data in the array
5. Perform arithmetic operations on multi-dimensional arrays
6. Append data vertically and horizontally
7. Apply indexing and slicing on array
8. Use statistical functions on array – Min, Max, Mean, Median and Standard Deviation
9. Dot matrix product of two arrays
10. Compute the Eigen values of a matrix
11. Solve a linear matrix equation such as 3 \* x0 + x1 = 9, x0 + 2 \* x1 = 8
12. Compute the multiplicative inverse of a matrix
13. Compute the rank of a matrix
14. Compute the determinant of an array
15. Perform transpose and change of axes operations on arrays.
16. Perform splitting operations on arrays.

**2. Fisher’s Iris Dataset**

**Description:**

This famous (Fisher’s or Anderson’s) iris data set gives the measurements in ormalizes of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris ormal, versicolor, and virginica.

Format

iris is a data frame with 150 cases (rows) and 5 variables (columns) named Sepal.Length, Sepal.Width, Petal.Length, Petal.Width, and Species.

The header is : sepal length, sepal width, petal length, petal width, iris, Species No. It has value 1 for Iris setosa, 2 for Iris virginica and 3 for Iris versicolor.

**Questions:**

1. Load the data in the file Iris.txt in a 2-D array called iris.
2. Drop column whose index=4 from the array iris.
3. Display the shape, dimensions and size of iris.
4. Split iris into three 2-D arrays, each array for a different species.callthem iris1, iris2, and iris3.
5. Print the three arrays iris1,iris2,iris3
6. Create a 1-D array header having elements “sepal length”,” sepalwidth”, “petallength”, “petalwidth”,” species No” in that order.
7. Display the array header.
8. Find the max, min, mean, and standard deviation for the columns of the iris and store the results in the arrays iris\_max, iris\_min, iris\_avg, iris\_std,iris\_varresp.The results must be rounded to not more than two decimal places.
9. Similarly find the max, min, mean, and standard deviation for the columns of the iris1, iris2, iris3 and store the results in the arrays with appropriate names.
10. Check the minimum value for sepal length, sepal width , petal length, petal width of the three species in comparison to the minimum value of sepal length,sepalwidth,petallength,petal width for the data set as awhole and fill the table below with True if the species value is greater than the dataset value and False otherwise.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Iris setosa | Iris virginica | Iris versicolor |
| Sepal length |  |  |  |
| Sepal width |  |  |  |
| Petal length |  |  |  |
| Petal width |  |  |  |

1. Compare Iris setosa’s average sepal width to that of Iris virginica.
2. Compare Iris setosa’s average petallength to that of Iris virginica.
3. Compare Iris setosa’s average petal width to that of Iris virginica.
4. Save the array iris\_avg in a comma separated file named IrisMeanValues.txt on the hard disk.
5. Save the arrays irisw\_max, iris\_avg, iris\_min in a comma separated file named IrisStat.txt on the hard disk.

**3. Pandas Series Programs**

**Write a Python program to do the following operations: Library: PandasSeries**

1. To add, subtract, multiple and divide two pandas Series.
2. To convert all the string values to upper, lower cases in a given pandas series. Also find the length of the string values.
3. To remove whitespaces, left sided whitespaces and right sided whitespaces of the string values of a given pandas series.
4. To create a series from a list, numpy array and dict
5. To calculate the number of characters in each word in a series.
6. To compare the elements of the two Pandas Series.

Sample Series: [2, 4, 6, 8, 10], [1, 3, 5, 7, 10]

1. To convert a Panda module Series to Python list and it’s type.
2. To create a series from a list, numpy array and dict.
3. To Combine many series to form a dataframe.
4. to stack two series vertically and horizontally
5. To create and display a DataFrame from a specified dictionary data which has the index labels.
6. Identify frequency counts of unique items of a series.

## To get the items of series A not present in series B?

1. To convert a numpy array to a dataframe of given shape.

**4. Pandas DataFrames Programs**

**Python program to do the following operations: Library: Pandas DataFrames**

1. **import and read a CSV file**
2. **To Generate a basic understanding of a given data.**
3. Print First 5 rows, last 5 rows of data
4. Check the basic information of the data
5. Extract the shape of the data
6. Print the unique values of the marital status field based on the column
7. To make it consistent data( widow and widowed )are two different naming for the same category on the column values
8. Check for duplicates and null values in the whole dataset
9. **select and filter data based on conditions**
10. Select a subset of the data points(Birthdate, Education, and Income of every customer) from the data frame.
11. Using loc() and iloc() methods to retrieve the first seven data points.
12. Filter data using the loc() and isin() methods.(Note:we choose the variable of interest and we select the categories )
13. In our data, that satisfies two conditions such as choosing the customers with an income higher than 75,000 and with a master’s degree(using python operators) and display the output.
14. **apply various data operation tools such as creating new variables or changing data types**

**We can apply different operations on the dataset using Pandas such as**

1. setting a new index with the variable of our interest using the .set\_index() method
2. sorting the data frame by one of the variable using .sort\_values() with ascending or descending order;
3. creating a new variable which could be the result of a mathematical operation such as sum of other variables
4. changing the datatype of variables into datetime or integer types
5. determining the age based on year of birth
6. creating the week date (calendar week and year) from the purchase date
7. **perform data aggregation using group by and pivot table methods**

After we created new variables, we can further aggregate  and to analyze data by groups,

1. To apply groupby()method to find the mean of income ,recency,number of web and store purchases by educational group.
2. To apply pivot\_table()method to find the aggregated sum of purchases and mean of recency per education and marital status group.

**5. Canada Pizza Price Prediction**

**Columns:**

company, price\_cad, diameter, topping, variant, size, extra\_sauce, extra\_cheese, extra\_mushrooms

**Questions:**

1. Count the number of null values in the pizza dataset and replace null values with average of the concerned columns.
2. Calculate average price of pizza prepared by each company.
3. Find the companies, who prepared pizzas with different variants with same diameter.
4. Which company has more pizzas? Show the result with graph.
5. Check whether the pizza data set contains null value or not. /Count the no. of null values in the pizza dataset./ Find the number of missing data points per column.
6. Rename the column price\_cad as price.
7. Identify the number of companies in each category
8. Identify which type of pizza is more expensive.
9. Find diameter of jumbo size pizza.
10. Any jumbo pizza with diameter less than 16 exists, remove such rows.
11. Calculate average price of a pizza prepared by company A.
12. Find the mean of the diameter and average price of pizzas prepared by company C.
13. Find the companies, who prepared pizzas with different variants with same diameter.
14. Find the pizza variant with extra\_mushrooms and topping with chicken.
15. What is the most expensive pizza in each company?
16. Which company has more pizzas on the menu? Show the result with graph.
17. What is the average price of pizza in each company?

**6.Mobile Phone Price Data set:**

* **Columns:**
  + **Brand:** the manufacturer of the phone
  + **Model:** the name of the phone model
  + **Storage (GB):** the amount of storage space (in gigabytes) available on the phone
  + **RAM (GB):** the amount of RAM (in gigabytes) available on the phone
  + **Screen Size (inches):** the size of the phone's display screen in inches
  + **Camera (MP):** the megapixel count of the phone's rear camera(s)
  + **Battery Capacity (mAh):** the capacity of the phone's battery in milliampere hours
  + **Price ($):** the retail price of the phone in US dollars

**Questions:**

1. Identify the models & the price released by each brand.
2. Identify the correlation between Battery Capacity and price.
3. Find how many models are there per each Battery capacity with same price.
4. Count the number of models in each brand with highest storage. Draw the graph.
5. Identify how many models are released by each brand.
6. Find the RAM capacity of all models of every brand.
7. Identify the correlation between Battery Capacity and price.
8. Find how many models are there per each Battery capacity.
9. Calculate average price of each brand.
10. Find which mobile brand has highest price.
11. Identify any missing values are there in mobile phone price dataset.
12. Display all models associated with apple brand.
13. Find the mobile prices based on Camera (MP).
14. List the models along with brands which have highest storage.
15. How many models in each brand having RAM>6.
16. List the models having price >600 and Storage between 100 and 200.

**7.National Universities Rankings**

**Columns:**

* Name – institution name,
* Location – City and state where located,
* Rank – Ranking according to U.S News & World Report ,
* Description – Snippet of text overview from U.S News ,
* Tuition and fees – Combined tuition and fees for out–of–state students ,
* In–state – Tuition and fees for in–state students ,
* Undergraduate Enrollment – Number of enrolled undergraduate students .

**Questions:**

1. Find the universities along with state whose fee is in between 25,000$ to 30,000$
2. Find university where undergraduate enrollement is morethan 25000 containing in-state students.
3. Find the states where universities are located in three or more cities.
4. Find max & min tuition fee in each state.
5. Find the city & state where maximum tuition fee difference in the universities in that city is greater than 5,000.
6. Print the names of universities having no. of branches along with the names of the branches.
7. Print university name and where it is located.
8. Find cities having more than 2 universities along with state.
9. Find the no. of states and cities locating top 100 universities.
10. Draw the plot to show undergraduate nrolment of each university.
11. Draw the plot to show university name and its corresponding tuition fee.
12. Plot the no. of universities in each state having ranks>100.

**8.Adidas Sales Dataset**

**Columns:**

* Retailer ID, Invoice Date,
* Region, State,
* City, Product,
* Price per Unit, Units Sold,
* Total Sales, Operating Profit,
* Operating Margin, Sales Method

**Questions:**

1. List all the products sold in every region.
2. Find the Cities & the retailers who sold womens related products.
3. Find the total sales of each womens product in in-store method.
4. For each product, find region wise total sales & units sold.
5. For men’s & women’s products, find state wise units sold & total sales.
6. Find states where women’s products sold were more than men’s products.
7. Find region wise units sold for each product
8. Find region wise profit for every retailer.
9. Find the states along with units sold where products sold in more than one city in the state.
10. Draw plot to show monthly sales in 2020 in every region
11. Draw the plot to show year wise sales in every region.
12. Draw plots to show Region wise sales in every year.

**9. Movies dataset**

**Columns:**

* Title, US Gross,
* Worldwide Gross, US DVD Sales,
* Production Budget, Release Date,
* MPAA Rating, Running Time (min),
* Distributor, Source,
* Major Genre, Creative Type,
* Director, Rotten Tomatoes Rating,
* IMDB Rating, IMDB Votes

**Questions:**

1. Find number of movies released under each genre in each year.
2. Find movies with loss every year for each distributor.
3. Find the Directors who directed for each creative type with IMDB rating above 6.
4. Draw the plot to compare the number of movies released till now by each director.
5. Find the genres of the movies released in each year in the ascending order.
6. Find the budgets of the movies released by each distributor along with movie names.
7. Find the movies with the same IMDD rating but with different no.of IMDD rating.
8. Write a Pandas program to get those movies whose revenue more than 2 million and spent less than 1 million.
9. Find the no. of movies in each genre under each source.
10. Find the no. of movies released in each decade.
11. Draw the plot showing the no. of movies released in each genre.
12. Show the no.of movies not rated under each genre in each fiction.
13. **Avocado Prices**

Historical data on avocado prices and sales volume in multiple US markets

**Some relevant columns in the dataset**:

* Date - The date of the observation
* AveragePrice - the average price of a single avocado
* type - conventional or organic
* year - the year
* Region - the city or region of the observation
* Total Volume - Total number of avocados sold
* 4046 - Total number of avocados with PLU 4046 sold
* 4225 - Total number of avocados with PLU 4225 sold
* 4770 - Total number of avocados with PLU 4770 sold

**About this file**

Numerical column names refer to price lookup codes.

1. small Hass
2. large Hass
3. extra large Hass

**Questions:**

1. How to identify the unique values in the region column.
2. What is the maximum price for an avocado in the dataset.
3. Identify the type distribution and take a single avocado in the dataset and find out the median price ,mean, and standard deviation.
4. Find the highest, lowest price for conventional avocado’s in year with location.
5. Draw the plots of the distribution of average price for different types of Avocados
6. Find the correlation matrix to measure the strength of the correlation between variables.
7. Find out the volume of avocado sales has increased in the last 5 years.