Data Visualization Project

Idea 1: 15 Years of BITS Hyderabad

Members:
Ashutosh Wagh (2022H1030052H)
Shivam Rajput (2022H1030058H)
S Shashank (2022H1030067H)
Ajinkya Medhekar (2022H1030099H)

Goal of the project:

The primary aim of this data visualization project is to craft a captivating and comprehensive narrative that transcends mere statistics, providing a dynamic and immersive journey through the evolution and accomplishments of BITS Hyderabad over the past 15 years. Through the lens of data, we aspire to weave a tale of growth, resilience, and the indomitable spirit of excellence that defines the BITS Hyderabad community. The ultimate aspiration is to create an immersive and insightful experience that not only celebrates the quantitative achievements but also captures the ethos, culture, and human stories that have shaped BITS Hyderabad over the past 15 years. Through this data visualization project, we aim to foster a sense of pride among the BITS community, inspire current and future students, and showcase BITS Hyderabad as a beacon of academic and professional excellence.

A. **Student Data**

Data Acquisition:

Part 1:

1. **Set Parameters:** Define the college year and the number of students you want in the dataset.
2. **Prepare Information:** List various courses, degrees, and other details needed for student data.
3. **Get Ready to Store Data:** Create empty lists to hold information about each student.
4. **Decide Degrees for Students:** Use a distribution to decide how many students will pursue each degree. Create a list of degrees and shuffle it for randomness. Fill any remaining spots with the default degree.

5. **Generate Student Data:** For each student, randomly choose a course, degree, CGPA, and academic year. Create a unique student ID (roll number) based on this information.
6. **Organize Data:** Put all the generated information into a structured format.
7. **Save Data:** Save the structured data into a file (CSV) with a name indicating the college year.

Part 2:

1. **Setting Things Up:** Decide the college year and the number of students you want in the dataset. Make lists of different courses and degrees. Set some rules for specific degrees and courses.
2. **Preparing Student Numbers:** Decide how many students will study each degree. Add a bit of randomness to mix things up.
3. **Creating Student Profiles:** Assign courses, degrees, grades, and years to each student. Make a unique ID for each student.
4. **Checking Choices:** For some degrees, ensure students don't pick certain courses.
5. **Saving the Information:** Organize all this information into a neat table (DataFrame). Save this table as a file with a name indicating the college year.

Data Preparation:

1. **Folder Exploration:** Look into a specified folder for files that match a certain pattern.
2. **File Selection:** Pick files with names starting with "student_data_" and ending with ".csv"
3. **Data Extraction:** For each selected file, extract the year from its name. Combine the folder path and the filename to get the full file path.
4. **DataFrame Creation:** Read each CSV file into a Pandas DataFrame.
5. **Year Addition:** Add a new column to the DataFrame called 'Year' and assign the extracted year to it.
6. **Compilation:** Put all these DataFrames into a list.
7. **Result:** Return the list containing all the DataFrames

Data Visualisation:

Following are the results obtained in the graphs and analysis done based on them.

1. We plotted the graph for Number of Students each year from 2008 to 2023. Obviously, it shows an increase in the number of students each passing year.
2. Then, we analyzed the Number of Students each year from 2008 to 2023 according to their Degrees(i.e. BE, ME, PhD, M.Sc, Pharma).
3. Next, we plotted the graph for Number of Students till now in each degree. The maximum number of students are in BE degree and minimum number of students are in PhD.
4. We plotted a treemap showing the number of students in each degree for all the years.
5. We plotted circle packing showing the different courses for all the degrees.
6. Then, we plotted the graph for average CGPA for all the students based on their degrees for each year. This shows that, the average CGPA has been mostly around 7.5 throughout the years.
7. After that, we plotted the graph of CGPA distribution by Degree. The CGPA values mostly fall under the range between 7-8, also the average CGPA is around 7.5 for all the degrees.

B. Alumni Data

Data Acquisition:

1. **Read Student Data**: Load a CSV file containing student data into a Pandas DataFrame.

2. **Filter Fourth-Year Students**: Extract a subset of students who are in their fourth year.

3. **Define Placement Companies:** Create a dictionary representing placement companies for different courses. Each course has a list of companies with associated placement probabilities.

4. **Assign Placement for Each Fourth-Year Student**: For each fourth-year student: Based on the student's course, randomly select a placement company from the predefined list. The probability of selecting a company is weighted according to the

predefined placement probabilities. If the student is not placed (random chance less than 0.8), mark them as "Unplaced."

5. **Create Alumni Data**: Add a new column to the DataFrame to store the assigned placement for each fourth-year student. Save the DataFrame with the additional placement information to a new CSV file.

6. **Save Data**: Display a message indicating that the alumni data, including placement ratios and information for some students without placement, has been generated and saved to a CSV file.

Data Preparation:

1. **Folder and File Setup**:
   - Specify the path to the folder containing the alumni data CSV files.
   - Define a pattern for matching CSV file names.

2. **DataFrame List Initialization**:
   - Initialize an empty list to store individual DataFrames.

3. **Iterate Through Files**:
   - Loop through files in the specified folder.
   - Check if each file is a CSV file with a specific naming pattern.

4. **Extract Year and Read CSV**:
   - Extract the year information from the file name.
   - Read the CSV file into a pandas DataFrame.

5. **Add Year Column**:
   - Add a new column to the DataFrame to store the extracted year.

6. **Append to DataFrame List**:
   - Append the DataFrame to the list.

7. **Combine DataFrames**:
   - Use concatenation to combine all DataFrames into a single DataFrame.
   - Reset the index of the combined DataFrame.

9. **Final Result**:

- The final result is a combined DataFrame containing data from all the alumni data CSV files in the specified folder.
- Modify the combined data frame appropriately to perform the data visualization.

Data Visualisation:

Following are the results obtained in the graphs and analysis done based on them.

1. First, we plotted the graph for observing the number of placed and unplaced students over the years. The graph shows the overall placement of students has been good, nearly 80% people are getting placed from the college.
2. The next graph is regarding the top recruiters from each year. Top recruiters give an idea regarding the next placement seasons and also the number of students they may hire next year.
3. Then, the graph is plotted between the courses in the college i.e Computer Science, and Information Systems, Mechanical Engineering, etc. and the number of placed and unplaced students these courses have for each year. By this, we can infer the top branches for placement are Computer Science and Information Systems, Pharmacy, Chemical Engineering.
4. After that, a graph is plotted which shows CGPA ranges like 5-6,6-7,7-8,8-9,9-10 vs the number of placed and unplaced students in those CGPA ranges. This gives an idea that higher CGPA will mostly guarantee placement.
5. Lastly, the graph for average CGPA for placed and unplaced students is plotted. It shows variation in average CGPA for both types of students. Even though someone has a higher CGPA, it is not guaranteed that he/she will get placed. And, if someone has a lesser CGPA, it is not guaranteed that he/she will remain unplaced.

C. **Faculty Data**

Data Acquisition:

1. **HTML Content Retrieval:** Read the content of an HTML file.
2. **HTML Parsing:** Process the HTML content using a tool called BeautifulSoup.
3. **Faculty Item Extraction:** Identify and extract HTML sections related to faculty members.
4. **CSV File Setup:** Open a CSV file to store faculty information.
5. **Faculty Information Extraction Loop:** For each faculty item: Fetch the faculty's name, department, and position from the HTML structure. Find the link to the faculty's personal website.
6. **Web Scraping:** Access each faculty member's website and retrieve additional information (such as email) from there.

Data Preparation:

1. **Print and Write:** Display faculty information (name, department, position, and email) on the console. Save the same information to a CSV file.
2. **Making changes to the dataset:** Change the data in the Position Column by removing punctuation marks and also modifying some values for visualization purposes.

Data Visualization:

Following are the results obtained in the graphs and analysis done based on them.

1. We plotted Number of People in each Department of the college. The maximum number of faculties are in the Mechanical Engineering Department. And, the minimum number of faculties are in Pharmacy, Chemical Engineering and Humanities and Social Services Departments.
2. Then, the distribution of Faculty Positions was plotted across the college. It was observed that the maximum number of faculties present are Assistant Professors and minimum number of faculties are Visiting Assistant Professor.
3. After that, we plotted the graph for Faculty Positions for every Department in the college.

Sources of data:

1. Faculty data of BITS Pilani Hyderabad Campus acquired through : https://www.bits-pilani.ac.in/faculty/?campus=hyderabad
2. The student and alumni data has been generated randomly, as discussed in the Data Acquisition section of Part A and B.
3. Following is the link where all the datasets can be downloaded : https://drive.google.com/file/d/1-tJGkhWO9BzhUUiTG3a_NckKKBjenQE5/view?usp=sharing