# Question Paper

Exam Date & Time: 11-Jan-2023 (02:00 PM - 05:00 PM)

## MANIPAL ACADEMY OF HIGHER EDUCATION

Manipal School of Information Sciences (MSIS), Manipal
First Semester Master of Engineering - ME (Artificial Intelligence & Machine Learning / Big Data Analytics) Degree Examination - January 2023

**Principles of Data Visualization (Elective -1) [BDA 5132]**

**Marks: 100** **Duration: 180 mins.**

**Wednesday, January 11, 2023**

**Answer all the questions.**

1) Illustrate the importance of crawler, scraper, and parser. List 3 types of parser. (TLO: 1.1) (6+4 marks) (10)

2) (10)

1. 
```
html_doc = """<html><head><title>Student Details</title></head>
<body>
<p class="title"><b> Student Details </b></p> <div>
<p class–"Details">contact information
<a hwef="http://example.com/Name" class="name" id="link">Ramesh</a>,
<a hwef–"http://example.com/Mobile" class–"mobile"
id="link2">9876543210</a> and
```

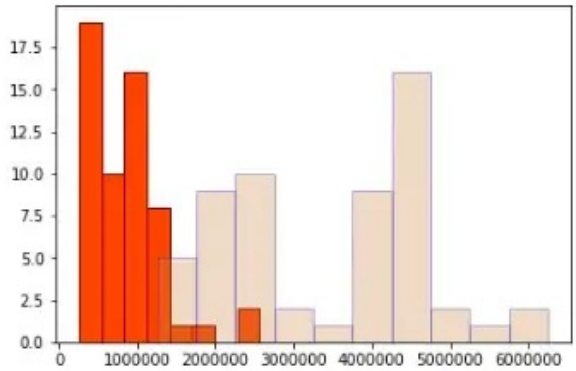
```
<a href="http://www.studentportal.com/email" class="email"
id="lnk3">Ramesh.m@xyz_com</a>;
and they lived at the bottom of a well.</p></div>
<p
"""
```

Using beautiful soup, write the script to extract following details
To display the title of the HTML file
To display the Name, mobile number and email
To display the email if the mail ID is valid(TLO: 1.2)

3) Using scrapy tool, write a spyder script to extract quotes, author and tags from the URL: http://quotes.toscrape.com having 10 web pages and store it in a csv file. (10)

4) Illustrate how fancy indexing is different from regular indexing with code snipet. (TLO: 2.1) (10)

5) Write a python script for multiple histogram as shown below (TLO 3.1) (10 Marks) (10)




6) Consider restaurant bill data set, consider the general split-apply-combine technique to calculate the following: (10)

1. sum of tips given by alcoholic and non-alcoholiccount
2. the bills generated for lunch and dinner (TLO 2.3)

7) Write a regular expression for matching URL under following consideration: (10)

1. Must start with http or https or ftp followed by ://
2. Must match a valid domain name
3. Could contain a port specification (http://www.sitepoint.com:80)Could contain digit, letter, dots, hyphens, forward slashes, multiple times (TLO: 2.3) (2.5+2.5+2.5+2.5 Marks)

| | | |
|---|---|---|
| 8) | Write a Pandas program to replace the missing values with the most frequent values present in each column of a given dataframe. Test data consists of following attributes: ord_no, purch_amt, sale_amt, ord_date, customer_id, salesman_id. (TLO 2.2) | (10) |
| 9) | Explain categories of explanatory visualizations based on the relationships between the three necessary players. (TLO 3.1) | (10) |
| 10) | Differentiate special indexing operator loc and iloc with example. (TLO 2.1) | (10) |

-----End-----