Indra Ganesan
COLLEGE OF ENGINEERING

Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai
Accredited by NAAC with 'B+' Grade, 2(f) & 12B Status Institution by UGC
IG Valley, Madurai Main Road, Manikandam, Tiruchirappalli - 620012

NAAC DOCUMENTS

QUALITY INDICATOR FRAME WORK

CRITERION – 1

CURRICULAR ASPECTS

SUBMITTED BY

IQAC

INTERNAL QUALITY ASSURANCE CELL

INDRA GANESAN COLLEGE OF ENGINEERING

Criteria 1	Curricular Aspects	100

1.1 Curricular Planning and Implementation (20)

1.1.1 The Institution ensures effective curriculum planning and delivery through a well-planned and documented process including Academic calendar and conduct of continuous internal Assessment

Table of Content

S. No	Description
1.	Preface of the Course File
2.	Review of Course File
3.	Faculty Time Table
4.	Course Plan
5.	Course Committee Meeting
6.	Content Beyond Syllabus
7.	Rubrics Base Evaluation
8.	Academic Audit Form
9.	Student Feed Back on Faculty
10.	Internal Assessment Schedule
11.	Question Paper
12.	Answer Key
13.	Sample Answer Sheet
14.	Co Based Mark Entry
15.	Root Cause Analysis
16.	Retest Question Paper
17.	Retest Sample Answer Sheet
18.	Retest Co Based Mark Entry

INDRA GANESAN COLLEGE OF ENGINEERING

IG Valley, Manikandam, Tiruchirappalli, Tamil Nadu – 620 012, India
(Approved by AICTE, New Delhi, Affiliated to Anna University, Chennai-25)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

PREFACE OF THE COURSE FILE

Batch	: **2021-2025**
Academic Year	: **2022- 2023 / ODD**
Program	: **COMPUTER SCIENCE AND ENGINEERING**
Year & Semester	: **2nd Year / 3rd Semester**
Course Code	: **CS 3352** NBA Course Code: **C203**
Name of the Course	: **Foundations of Data Science**
Faculty in-charge	: **T. Sugashini AP / CSE**

Signature of the Faculty in-charge

HoD / CSE

Dr. G. Balakrishnan, M.E., Ph.D.,
Principal
Indra Ganesan College of Engineering
IG Valley, Madurai Main Road
Manikandam, Trichy-620 012.

INDRA GANESAN COLLEGE OF ENGINEERING

(Approved by AICTE, New Delhi and affiliated to Anna University, Chennai)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

REVIEW OF COURSE FILE

(to be pasted on the inner side of the file-backside).(#-State Yes/No.)

S.N	Details Date:	R-I-*	R-II-*&	R-III-*&	R-IV-*&$	R-V-*&$@
1.	Preface of the course file	Y				
2.	Vision, Mission, PEOs, POs, PSOs, Blooms taxonomy	Y				
3.	Subject handlers of yesteryears	Y				
4.	Timetable/Workload of the staff – Distribution of teaching load – Roles and Responsibilities	Y				
5.	Syllabus signed by staff & HoD	Y				
6.	Lecture Schedule signed by staff & HoD	Y				
7.	Course Committee meeting circular and minutes	Y				
8.	Identification of Curricular gap and Content Beyond the syllabus	Y				
9.	Self-study topics	Y				
10.	Previous AU Question papers	Y				
11.	Unit wise Q&A and Objective type questions	Y				
12.	Unit wise course material		Y	Y	Y	
13.	Assignment question paper with sample answer sheets and mark entry		Y	Y	Y	
14.	Tutorial question paper with key and mark entry		Y	Y	Y	
15.	Class test/IA test Q Paper with Key, sample answer papers and mark entry		Y	Y	Y	
16.	IA Test- result analysis-CAP-evidence-root cause analysis.		Y	Y	Y	
17.	Retest –Q paper-Attendance-marks		Y	Y	Y	
18.	AU Web portal entry sheet		Y	Y	Y	
19.	Very poor performance in first two tests-action taken.-communication to parents-evidence			Y	Y	
20.	Absence for two tests-action taken-communication to parents-evidence.			Y	Y	
21.	Indiscipline of student reported, if any					
22.	Special class/coaching class/remedial class/attendance-CAP		Y	Y	Y	
23.	Conduct of Seminar, Quizzes - proof					
24.	Content beyond the syllabus - proof					Y
25.	Student feedback on faculty					Y
26.	Course end survey					Y
27.	Internal Assessment sheet					Y
28.	AU question paper with students feedback					Y
29.	Discrepancy of the question paper and correspondence, if any					Y
30.	AU result analysis-Details of arrear students.					Y
31.	AU grade sheet					Y
32.	CO – PO & PSO attainment sheet					Y
	Signature of Course handling faculty					
	Signature of HoD					

Dr. G. Balakrishnan, M.E., Ph.D.,
Principal
Indra Ganesan College of Engineering
IG Valley, Madurai Main Road
Manikandam, Trichy-620 012.

INDRA GANESAN COLLEGE OF ENGINEERING

IG Valley, Manikandam, Tiruchirappalli, Tamil Nadu – 620 012, India
(Approved by AICTE, New Delhi, Affiliated to Anna University, Chennai-25)

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING

Faculty Time Table

T. Sugashini

Day Order	1	2	3	4	5	6	7	8
I		CS3352						
II								
III				CS3352				
IV		CS3352						
V				CS3352				

S.Code	Title	Year / Branch	Hours
CS3352	Foundation of Data Science	II / CSE	4
TOTAL - 4 hours			

Signature of the Faculty

HOD/CSE

Dr. G. Balakrishnan, M.E., Ph.D.,
Principal
Indra Ganesan College of Engineering
IG Valley, Madurai Main Road
Manikandam, Trichy-620 012.

CS3352	FOUNDATIONS OF DATA SCIENCE	L T P C 3 0 0 3

COURSE OBJECTIVES:

- To understand the data science fundamentals and process.
- To learn to describe the data for the data science process.
- To learn to describe the relationship between data.
- To utilize the Python libraries for Data Wrangling.
- To present and interpret data using visualization libraries in Python

UNIT I INTRODUCTION **9**

Data Science: Benefits and uses – facets of data - Data Science Process: Overview – Defining research goals – Retrieving data – Data preparation - Exploratory Data analysis – build the model– presenting findings and building applications - Data Mining - Data Warehousing – Basic Statistical descriptions of Data

UNIT II DESCRIBING DATA **9**

Types of Data - Types of Variables -Describing Data with Tables and Graphs –Describing Data with Averages - Describing Variability - Normal Distributions and Standard (z) Scores

UNIT III DESCRIBING RELATIONSHIPS **9**

Correlation –Scatter plots –correlation coefficient for quantitative data –computational formula for correlation coefficient – Regression –regression line –least squares regression line – Standard error of estimate – interpretation of r2 –multiple regression equations –regression towards the mean

UNIT IV PYTHON LIBRARIES FOR DATA WRANGLING **9**

Basics of Numpy arrays –aggregations –computations on arrays –comparisons, masks, boolean logic – fancy indexing – structured arrays – Data manipulation with Pandas – data indexing and selection – operating on data – missing data – Hierarchical indexing – combining datasets – aggregation and grouping – pivot tables

UNIT V DATA VISUALIZATION **9**

Importing Matplotlib – Line plots – Scatter plots – visualizing errors – density and contour plots – Histograms – legends – colors – subplots – text and annotation – customization – three dimensional plotting - Geographic Data with Basemap - Visualization with Seaborn.

COURSE OUTCOMES:

At the end of this course, the students will be able to:

CO1: Define the data science process
CO2: Understand different types of data description for data science process
CO3: Gain knowledge on relationships between data
CO4: Use the Python Libraries for Data Wrangling
CO5: Apply visualization Libraries in Python to interpret and explore data

TOTAL: 45 PERIODS

TEXT BOOKS

1. David Cielen, Arno D. B. Meysman, and Mohamed Ali, "Introducing Data Science", Manning Publications, 2016. (Unit I)
2. Robert S. Witte and John S. Witte, "Statistics", Eleventh Edition, Wiley Publications, 2017. (Units II and III)
3. Jake VanderPlas, "Python Data Science Handbook", O'Reilly, 2016. (Units IV and V)

REFERENCES:

1. Allen B. Downey, "Think Stats: Exploratory Data Analysis in Python", Green Tea Press, 2014.

HOD/CSE

Dr. G. Balakrishnan, M.E., Ph.D.,
Principal
Indra Ganesan College of Engineering
IG Valley, Madurai Main Road
Manikandam, Trichy-620 012.

PRINCIPAL

INDRA GANESAN COLLEGE OF ENGINEERING

IG Valley, Manikandam, Tiruchirappalli, Tamil Nadu – 620 012, India
(Approved by AICTE, New Delhi, Affiliated to Anna University, Chennai-25)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Lecture Schedule

Degree/Program: **B.E / CSE** Course code &Name: **CS3352** – Foundations of Data Science
Duration: **2022 – 2023 ODD SEM** Semester: **III** Faculty : T. Sugashini

AIM:

To encourage the students to solve real-world data-science problems and build applications in this field.

OBJECTIVES:

To impart knowledge on

(i) To study the data science fundamentals and familiarize with the data science process.
(ii) To familiarize describing data with tables, graphs, averages, and variability and converting the values from the normal distribution into z scores.
(iii) To study the data to describe the relationship by examining the form, direction, and strength of the association by quantitatively and qualitatively.
(iv) To apply the Python libraries for Data Analysis and Data Science, which involves data sorting or filtration and data grouping.
(v) To study visualization libraries in Python to create customized data along with its libraries, graphs, charts, and histogram

PREREQUISITES: Problem Solving and Python Programming, Problem Solving and Python Programming Laboratory, Statistics and Numerical Methods

COURSE OUTCOMES:

After the course, the student should be able to:

CO	Course Outcomes	POs	PSOs
C203.1	Explain the data science process and the basic concept of data science fundamentals	1,2,3,4,12	1,2
C203.2	Illustrate to convert the values from the normal distribution into z scores using data with tables, graphs, averages, and variability	1,2,3,4,12	1,2
C203.3	Examine the data to describe the relationship by examining the form, direction, and strength of the association by quantitatively and qualitatively.	1,2,3,4,12	1,2
C203.4	Examine the Numpy libraries to perform a wide variety of high-level mathematical functions that operate on the arrays and matrices.	1,2,3,4,12	1,2
C203.5	Examine the Pandas libraries for analyzing, cleaning, exploring, and manipulating data.	1,2,3,4,12	1,2
C203.6	Explain the visualization libraries in Python to identify patterns, trends, and outliers in large data sets along with its libraries, graphs, charts, and histogram	1,2,3,4,12	1,2

Dr. G. Balakrishnan, M.E., Ph.D.,
Principal
Indra Ganesan College of Engineering
IG Valley, Madurai Main Road
Manikandam, Trichy-620 012.

S.No	Date	Period	Topics to be Covered	Book
UNIT I - INTRODUCTION				**Target periods :9**
1	22.08.22	2	Data Science: Benefits and uses – facets of data	T1
2	24.08.22	4	Data Science Process: Overview – Defining research goals	T1
3	25.08.22	2	Retrieving data	T1
4	26.08.22	4	Data preparation	T1
5	29.08.22	2	Exploratory Data analysis	T1
6	01.09.22	2	Build the model– presenting findings and building applications	T1
7	02.09.22	4	Data Mining	T1
8	05.09.22	2	Data Warehousing	T1
9	07.09.22	4	Basic Statistical descriptions of Data	T1
UNIT II - DESCRIBING DATA				**Target periods :9**
10	08.09.22	2	Types of Data	T2
11	09.09.22	4	Types of Variables	T2
12	12.09.22	2	Describing Data with Tables and Graphs	T2
13	14.09.22	4		
14	15.09.22	2	Describing Data with Averages	T2
15	16.09.22	4		
16	19.09.22	2	Describing Variability	T2
17	21.09.22	4	Normal Distributions	T2
18	23.09.22	2	Standard (z) Scores	T2
UNIT III - DESCRIBING RELATIONSHIPS				**Target Periods :9**
19	26.09.22	2	Correlation	T2
20	28.09.22	4	Scatter plots	T2
21	29.09.22	2	Correlation coefficient for quantitative data	T2
22	30.09.22	4	Computational formula for correlation coefficient	T2
23	3.10.22	2	Regression –regression line	T2
24	6.10.22	2	Least squares regression line	T2
25	7.10.22	4	Standard error of estimate	T2
26	12.10.22	4	Interpretation of r2	T2
27	13.10.22	2	Multiple regression equations – regression towards the mean	T2
UNIT IV PYTHON LIBRARIES FOR DATA WRANGLING				**Target Periods :9**
28	14.10.22	2	Basics of NumPy arrays –aggregations	T3
29	17.10.22	2	Computations on arrays	T3
30	19.10.22	4	Comparisons, masks, Boolean logic	T3
31	20.10.22	2	Fancy indexing – structured arrays	T3
32	21.10.22	4	Data manipulation with Pandas	T3
33	26.10.22	4	Data indexing and selection – operating on data – missing data	T3
34	27.10.22	2	Hierarchical indexing	T3
35	28.10.22	2	Combining datasets	T3
36	8.11.22	4	Aggregation and grouping – pivot tables	T3
UNIT V UNIT V DATA VISUALIZATION				**Target Periods:9**
37	10.11.22	4	Importing Matplotlib – Line plots	T1/BB
38	11.11.22	2	Scatter plots	R2/BB
39	14.11.22	4	Visualizing errors – density and contour plots	T1/BB
40	16.11.22	2	Histograms – legends	R3/BB

Dr. G. Balakrishnan, M.E., Ph.D.,
Principal
Indra Ganesan College of Engineering
IG Valley, Madurai Main Road
Manikandam, Trichy-620 012.

41	17-11-22	2	Colors – Subplots	R3/BB
42	18-11-22	4	Text and annotation	R2/BB
43	21-11-22	2	Customization – Three-dimensional plotting	T1/BB
44	23-11-22	4	Geographic Data with Basemap	T1/BB
45	24-11-22	2	Visualization with Seaborn	T1/BB
			Content Beyond the Syllabus	
46	25-11-22	2	Visual Aids for EDA	Material

Book Reference - Text Books

Sl.	Title of the Book	Author	Publisher	Year
1.	Introducing Data Science	David Cielen, Arno D. B. Meysman, and Mohamed Ali	Manning Publications	2016
2.	Statistics	Robert S. Witte and John S. Witte	Eleventh Edition, Wiley Publications	2017
3.	Python Data Science Handbook	Jake VanderPlas	O'Reilly	2016

Book Reference – References

Sl.	Title of the Book	Author	Publisher	Year
1.	Think Stats: Exploratory Data Analysis in Python	Allen B. Downey	Green Tea Press	2014

Website Reference:

https://nptel.ac.in/courses/106106179

https://www.udemy.com/course/the-data-science-course-complete-data-science-bootcamp/

Signature of the Faculty in-charge

HoD / CSE

Dr. G. Balakrishnan, M.E., Ph.D.,
Principal
Indra Ganesan College of Engineering
IG Valley, Madurai Main Road
Manikandam, Trichy-620 012.

INDRA GANESAN COLLEGE OF ENGINEERING

IG Valley, Manikandam, Tiruchirappalli, Tamil Nadu – 620 012, India
(Approved by AICTE, New Delhi, Affiliated to Anna University, Chennai-25)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Ref: SBECW/ CSE/ Course committee meeting / FDS-I/ 2021 – 2022 (ODD) DATE: 17.08.20222

COURSE COMMITTEE MEETING-CS3352 – FOUNDATIONS OF DATA SCIENCE

ACADEMIC YEAR: 2022 – 2023 (ODD) **SEM:** 03 **REGULATION:** 2021
PROGRAM: CSE **DATE OF MEETING:** 17.08.2022 **TIME:** 10.00 AM **Venue:** RDBMS LAB

Members Present

Table.1 Course committee members

S.No.	Name of the faculty & Designation, Program	Sem/Program	Signature
1.	T. Sugashini, AP/CSE - Course coordinator	III SEM / CSE	
2.	R. Nivethas AP/IT	III SEM / IT	

HOD welcomed all the members present

1. Content of syllabus, unit wise discussed. Nature of qualitative, quantitative, problematic, theoretical concepts etc. have been discussed
2. With reference to the R-2021 regulation, Number of periods per unit = 9, total number of periods = 45 periods.
3. Vision and mission of the college, department discussed. POs, PEOs, PSOs discussed.
4. Course outcomes defined for each units, considering learning outcomes.

Table.2 Course Outcomes

CO	Course Outcomes	POs	PSOs
C203.1	Explain the data science process and the basic concept of data science fundamentals	1,2,3,4,12	1,2
C203.2	Illustrate to convert the values from the normal distribution into z scores using data with tables, graphs, averages, and variability	1,2,3,4,12	1,2
C203.3	Examine the data to describe the relationship by examining the form, direction, and strength of the association by quantitatively and qualitatively.	1,2,3,4,12	1,2
C203.4	Examine the Numpy libraries to perform a wide variety of high-level mathematical functions that operate on the arrays and matrices.	1,2,3,4,12	1,2
C203.5	Examine the Pandas libraries for analyzing, cleaning, exploring, and manipulating data.	1,2,3,4,12	1,2
C203.6	Explain the visualization libraries in Python to identify patterns, trends, and outliers in large data sets along with its libraries, graphs, charts, and histogram	1,2,3,4,12	1,2

5. Mapping of COs with POs and PSOs is done with suitable correlation levels(1 for low, 2 for medium, 3 for high,"-" for no correlation, before content beyond syllabus)

Table.3 Mapping of COs, C, PSOs with POs- before CBS.

Course	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2
C203.1	3	2	1	1	-	-	-	-	-	-	-	1	2	2
C203.2	3	2	1	1	-	-	-	-	-	-	-	1	2	2
C203.3	3	2	1	1	-	-	-	-	-	-	-	1	2	2
C203.4	3	2	1	1	-	-	-	-	-	-	-	1	2	2
C203.5	3	2	1	1	-	-	-	-	-	-	-	1	2	2
C203.6	3	2	1	1	-	-	-	-	-	-	-	1	2	2
C203	3	2	1	1	-	-	-	-	-	-	-	1	2	2

6. Identification of content beyond syllabus- curricular gaps are identified considering industry needs, employers feedback, alumni feedback, government policy on industrialization, new investments by private/ public sectors, societal needs and level of correlation of COs with POs and PSOs. Accordingly the details of CBS added and its correlation is given below.

Dr. G. Balakrishnan, M.E., Ph.D.,
Principal
Indra Ganesan College of Engineering
IG Valley, Madurai Main Road
Manikandam, Trichy-620 012.

Table.4 Identification of content beyond syllabus

Content beyond syllabus added	POs strengthened/Vacant filled	CO/Unit
Visual Aids for EDA	PO5(2) Vacant filled	C203.1 & C203.3/ I & V

7. Mapping of COs with POs, PSOs- after CBS.

Table.5 Mapping of COs, C, PSOs with POs- after CBS.

Course	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2
C203.1	3	2	1	1	*2	-	-	-	-	2	1	1	2	2
C203.2	3	2	1	1	-	-	-	-	-	2	1	1	2	2
C203.3	3	2	1	1	-	-	-	-	-	2	1	1	2	2
C203.4	3	2	1	1	-	-	-	-	-	2	1	1	2	2
C203.5	3	2	1	1	-	-	-	-	-	2	1	1	2	2
C203.6	3	2	1	1	*2	-	-	-	-	2	1	1	2	2
C203	3	2	1	1	*2	-	-	-	-	2	1	1	2	2

8. Content beyond syllabus is thus identified based on the above. Plan for handling of CBS by internal/external resource person/ industrial visits are decided. This will be included in the class log book.
9. Lecture schedule should be prepared unit wise, as in the syllabus. Number of periods per unit and total number of periods planned should not be less than, periods allotted in the syllabus of Anna University.
10. Plan for additional Periods for CAT tests, CBS, NPTEL delivery, Seminar, Quiz etc are to be incorporated in the lecture schedule. These periods are added exclusive of number of periods prescribed in the syllabus.
11. Plan for at least three assignments (with level of correlation), seminar topic discussed.
12. Bright students and slow learners are to be identified, immediately after CAT test - I. such students may be counselled suitably and the evidence for counselling to be recorded in the attendance cum assessment record. (Sign of students with date and time of counselling, to be strictly recorded and to be attached in the course file). Such counselling may be conducted after college hours.
13. For those students secured less than 60% in the CAT Test, Retest should be conducted. Correspondingly root cause analysis for reasons of failure, corrective and preventive action, and follow up action taken should be filed properly.
14. Contents of course file to be reviewed periodically.
15. Lecture schedule, assignment questions, tutorial questions, course materials, AU questions (at least 5) should be supplied within one week after the commencement of classes.
16. Course material should be uploaded in the college website for student's reference.
17. Discrepancy in question paper, if any to be informed to the controller of examinations through web portal entry, after getting approval from the HoD & the Principal. Critically asked questions, if any to be discussed with the students of the next batch.
18. Immediately after the publication of the results, analysis are to be carried out and follow up action to be taken for the failures.
19. IA test question papers should be set as per the norms of the college, incorporating marks for learning outcomes and course outcomes. Common question papers should be set.
20. Certificate courses/Workshop/guest lectures may be planned inviting experts from industry/higher learning institutions.
21. CAT test papers, assignment papers or any other papers submitted by the students, should be returned to the students within 5 days after correction. Sample paper should be suitably filed.
22. Long absentees of students if any to be informed to the parents through class coordinator, if such students attendance less than 75%.

Course coordinator

HoD/CSE

Dr. G. Balakrishnan, M.E., Ph.D.,
Principal
Indra Ganesan College of Engineering
IG Valley, Madurai Main Road
Manikandam, Trichy-620 012.

INDRA GANESAN COLLEGE OF ENGINEERING

IG Valley, Manikandam, Tiruchirappalli, Tamil Nadu – 620 012, India
(Approved by AICTE, New Delhi, Affiliated to Anna University, Chennai-25)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Identification of Curricular Gap & Content Beyond Syllabus(CBS)

Name of the Faculty : T. Sugashini **Course Code & Name: CS3352 – Foundations of Data Science**

Degree & Program:B.E. /CSE **Semester : III /** **Academic Year: 2022 -2023 /ODD**

I.Mapping of Course Outcomes with POs & PSOs.(before CBS)

Table.1 Mapping of COs, C, PSOs with POs - before CBS.

Course	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2
C203.1	3	2	1	1	*2	-	-	-	-	2	1	1	2	2
C203.2	3	2	1	1	-	-	-	-	-	2	1	1	2	2
C203.3	3	2	1	1	-	-	-	-	-	2	1	1	2	2
C203.4	3	2	1	1	-	-	-	-	-	2	1	1	2	2
C203.5	3	2	1	1	-	-	-	-	-	2	1	1	2	2
C203.6	3	2	1	1	*2	-	-	-	-	2	1	1	2	2
C203	3	2	1	1	*2	-	-	-	-	2	1	1	2	2

II. Identification of content beyond syllabus.

Table.2 Identification of content beyond syllabus

Details of Content Beyond Syllabus(CBS) added	POs strengthened/ vacant filled	CO/Unit
Visual Aids for EDA	PO5(2) Vacant filled	C203.1 & C203.3/ I & V

III. Mapping of Course Outcomes with POs & PSOs. (After CBS)

Table.3 Mapping of COs, C, PSOs with POs- after CBS.

Course	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2
C203.1	3	2	1	1	*2	-	-	-	-	2	1	1	2	2
C203.2	3	2	1	1	-	-	-	-	-	2	1	1	2	2
C203.3	3	2	1	1	-	-	-	-	-	2	1	1	2	2
C203.4	3	2	1	1	-	-	-	-	-	2	1	1	2	2
C203.5	3	2	1	1	-	-	-	-	-	2	1	1	2	2
C203.6	3	2	1	1	*2	-	-	-	-	2	1	1	2	2
C203	3	2	1	1	*2	-	-	-	-	2	1	1	2	2

Signature of the Faculty

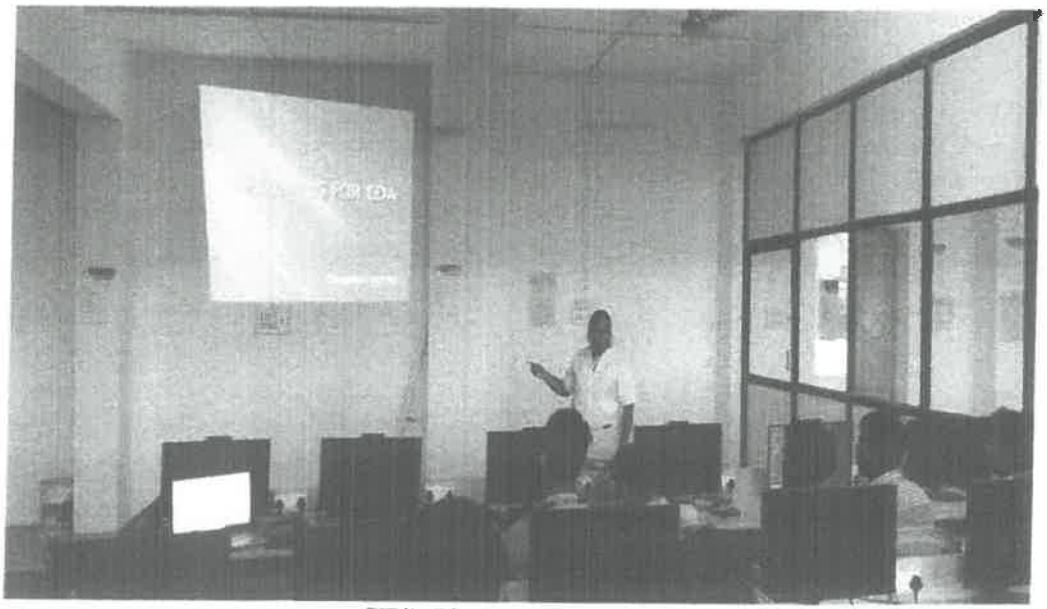
Dr. C. Balakrishnan, M.E., Ph.D.,
Principal
Indra Ganesan College of Engineering
IG Valley, Madurai Main Road
Manikandam, Trichy-620 012.

HoD/CSE

INDRA GANESAN COLLEGE OF ENGINEERING

IG Valley, Manikandam, Tiruchirappalli, Tamil Nadu – 620 012, India
(Approved by AICTE, New Delhi, Affiliated to Anna University, Chennai-25)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING



SESSION AT SD LAB

Dr. G. Balakrishnan, M.E., Ph.D.,
Principal
Indra Ganesan College of Engineering
IG Valley, Madurai Main Road
Manikandam, Trichy-620 012.

1.2 Visual Aids for EDA

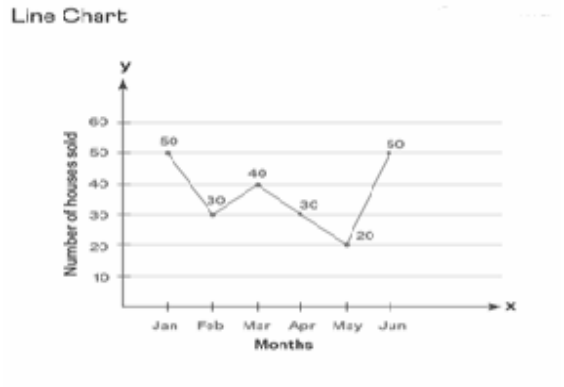
1.2.1 Introduction

The two important goals of data scientist would be to extract knowledge from the data and to present the data to stakeholders. Presenting results to stakeholders is very complex in the sense that the stakeholders may not have enough technical knowledge to understand programming terminologies and other technicalities. Hence, visual aids are very useful tools. The following are some of the visual aids for EDA.

- Line chart
- Bar chart
- Scatter plot
- Area plot and stacked plot
- Pie chart
- Table chart
- Polar chart
- Histogram
- Lollipop chart

1.2.2 Line chart

A line chart is a type of chart used to visualize the value of something over time. It is used to find trends in data over time. The chart consists of a horizontal x-axis and a vertical y-axis. Eg. The number of houses sold during various months of the year. The x-axis shows the time period whereas the y-axis shows the item that is being measured. A line chart clearly shows the increasing or decreasing trend of a particular item.



Simple Line Chart

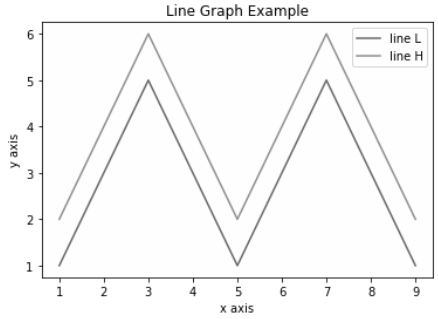
A simple line chart is plotted with only a single line that shows the relationship between two different variables

Multiple Line Chart

A multiple line chart is a line chart that is plotted with two or more lines. When we need to show data about two or more variables that have varying data points depending on the period of time, a multiple line chart can be used. This type of line chart is also helpful when we need to compare data like temperatures, prices, etc. The image below shows the comparison of prices of Mercedez-Benz among three cities.

Example:

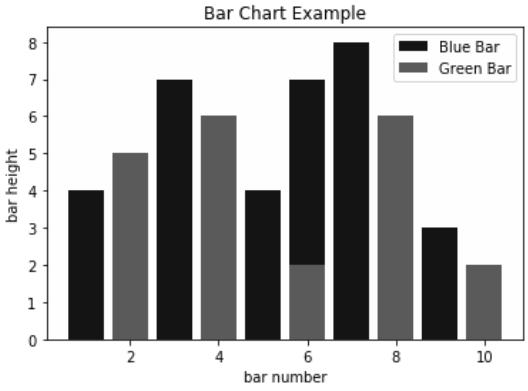
```
import matplotlib.pyplot as plt
x  = [1, 2, 3, 4, 5, 6, 7, 8, 9]
y1 = [1, 3, 5, 3, 1, 3, 5, 3, 1]
y2 = [2, 4, 6, 4, 2, 4, 6, 4, 2]
plt.plot(x, y1, label="line L")
plt.plot(x, y2, label="line H")
plt.plot()
plt.xlabel("x axis")
plt.ylabel("y axis")
plt.title("Line Graph Example")
plt.legend()
plt.show()
```



1.2.3 Bar chart

This is one of the most common types of visualization. Bars can be drawn horizontally or vertically to represent **categorical variables**. Bar charts are frequently used to distinguish objects between distinct collections in order to track variations over time. In most cases, bar charts are very convenient when the changes are large.

```
import matplotlib.pyplot as plt
# The index 4 and 6 demonstrate overlapping cases.
x1 = [1, 3, 4, 5, 6, 7, 9]
y1 = [4, 7, 2, 4, 7, 8, 3]
x2 = [2, 4, 6, 8, 10]
y2 = [5, 6, 2, 6, 2]
plt.bar(x1, y1, label="Blue Bar", color='b')
plt.bar(x2, y2, label="Green Bar", color='g')
plt.plot()
plt.xlabel("bar number")
plt.ylabel("bar height")
plt.title("Bar Chart Example")
plt.legend()
plt.show()
```



1.2.4 Scatter plot

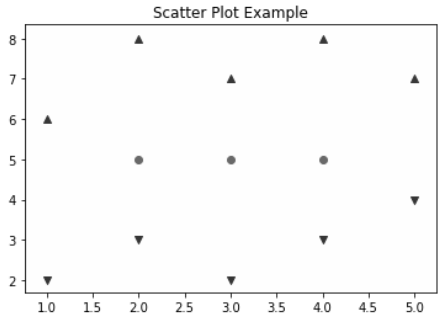
Scatter plots are also called scatter graphs, scatter charts, scattergrams, and scatter diagrams. They use a **Cartesian coordinates system** to display values of typically two variables for a set of data.

Scatter plots can be constructed in the following two situations:

- When one continuous variable is dependent on another variable, which is under the control of the observer.
- When both continuous variables are independent

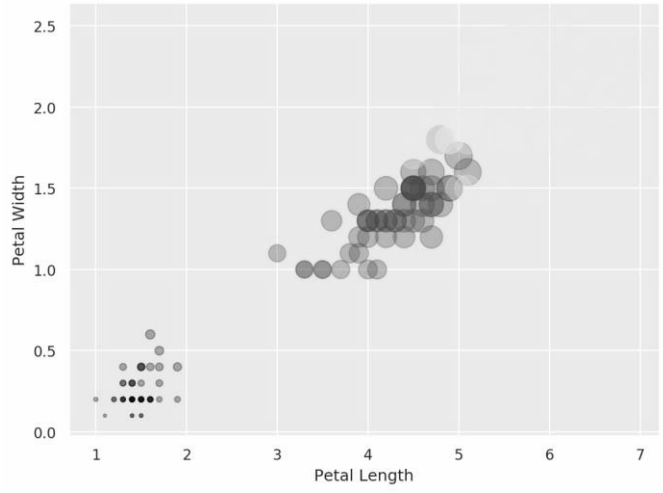
There are two important concepts—**independent variable** and **dependent variable**. In statistical modeling or mathematical modeling, the values of dependent variables rely on the values of independent variables. The dependent variable is the outcome variable being studied. The independent variables are also referred to as **regressors**. The scatter plots are used when we need to show the relationship between two variables, and hence are sometimes referred to as correlation plots.

```
import matplotlib.pyplot as plt
x1 = [2, 3, 4]
y1 = [5, 5, 5]
x2 = [1, 2, 3, 4, 5]
y2 = [2, 3, 2, 3, 4]
y3 = [6, 8, 7, 8, 7]
plt.scatter(x1, y1)
plt.scatter(x2, y2, marker='v', color='r')
plt.scatter(x2, y3, marker='^', color='m')
plt.title('Scatter Plot Example')
plt.show()
```



Bubble chart

A bubble plot is a scatter plot where each data point on the graph is shown as a bubble. Each bubble can be illustrated with a different color, size, and appearance.



1.2.5 Area plot and stacked plot

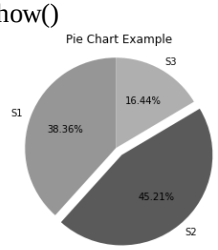
An area plot is a line plot that shows the area covered under the line by filling it with a color. Several such plots can be stacked on top of one another, giving the feeling of a stack and hence the name stacked plot. The stacked plot can be useful when we want to visualize the **cumulative effect** of multiple variables being plotted on the *y* axis.

```
import matplotlib.pyplot as plt
x = [1, 2, 3, 4, 5, 6, 7, 8, 9]
y1 = [23, 40, 28, 43, 8, 44, 43, 18, 17]
y2 = [17, 30, 22, 14, 17, 17, 29, 22, 30]
y3 = [15, 31, 18, 22, 18, 19, 13, 32, 39]
# Adding legend for stack plots is tricky.
plt.plot([], [], color='r', label = 'D 1')
plt.plot([], [], color='g', label = 'D 2')
plt.plot([], [], color='b', label = 'D 3')
plt.stackplot(x, y1, y2, y3, colors= ['r', 'g', 'b'])
plt.title('Stack Plot Example')
plt.legend()
plt.show()
```

1.2.6 Pie chart

A pie chart (or a circle chart) is a circular statistical graphic, which is divided into slices to illustrate numerical proportion. In a pie chart, the arc length of each slice and area is proportional to the quantity it represents. Pie charts are very widely used in the business world and the mass media. But, experts recommend avoiding them, as research has shown it is difficult to compare different sections of a given pie chart, or to compare data across different pie charts. Pie charts can be replaced in most cases by other plots such as the bar chart, box plot, dot plot, etc.

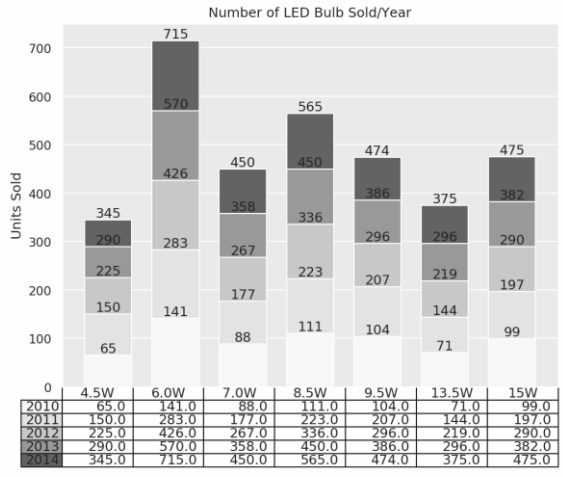
```
import matplotlib.pyplot as plt
labels = 'S1', 'S2', 'S3'
sections = [56, 66, 24]
colors = ['c', 'g', 'y']
plt.pie(sections, labels=labels, colors=colors,
        startangle=90,
        explode = (0, 0.1, 0),
        autopct = '%1.2f%%')
plt.axis('equal') # Try commenting this out.
plt.title('Pie Chart Example')
plt.show()
```



1.2.7 Table chart

A table chart combines a bar chart and a table. In order to understand the table chart, let's consider the following dataset. Consider standard LED bulbs that come in different wattages. The standard Philips LED bulb can be 4.5 Watts, 6 Watts, 7 Watts, 8.5 Watts, 9.5 Watts, 13.5 Watts, and 15 Watts. Let's assume there are two categorical variables, the year and the wattage, and a numeric variable, which is the number of units sold in a particular year.

```
import numpy as np
import matplotlib.pyplot as plt
# Years under consideration
years = ["2010", "2011", "2012", "2013", "2014"]
# Available watt
columns = ['4.5W', '6.0W', '7.0W','8.5W']
unitsSold = [
[65, 141, 88, 111],
[85, 142, 89, 112],
[75, 143, 90, 113],
[65, 144, 91, 114],
[55, 145, 92, 115],
]
# Define the range and scale for the y axis
values = np.arange(0, 600, 100)
colors = plt.cm.OrRd(np.linspace(0, 0.7, len(years)))
index = np.arange(len(columns)) + 0.3
bar_width = 0.7
y_offset = np.zeros(len(columns))
fig, ax = plt.subplots()
cell_text = []
n_rows = len(unitsSold)
for row in range(n_rows):
  plot = plt.bar(index, unitsSold[row], bar_width, bottom=y_offset,
  color=colors[row])
  y_offset = y_offset + unitsSold[row]
  cell_text.append(['%1.1f' % (x) for x in y_offset])
# Add a table to the bottom of the axes
the_table = plt.table(cellText=cell_text, rowLabels=years,
rowColours=colors, colLabels=columns, loc='bottom')
plt.ylabel("Units Sold")
plt.xticks([])
plt.title('Number of LED Bulb Sold/Year')
plt.show()
```



	4.5W	6.0W	7.0W	8.5W	9.5W	13.5W	15W
2010	65.0	141.0	88.0	111.0	104.0	71.0	99.0
2011	150.0	283.0	177.0	223.0	207.0	144.0	197.0
2012	225.0	426.0	267.0	336.0	296.0	219.0	290.0
2013	290.0	570.0	358.0	450.0	386.0	296.0	382.0
2014	345.0	715.0	450.0	565.0	474.0	375.0	475.0

1.2.8 Polar chart

A polar chart is a diagram that is plotted on a polar axis. Its coordinates are angle and radius, as opposed to the Cartesian system of x and y coordinates. Sometimes, it is also referred to as a spider web plot. Let's see how we can plot an example of a polar chart.

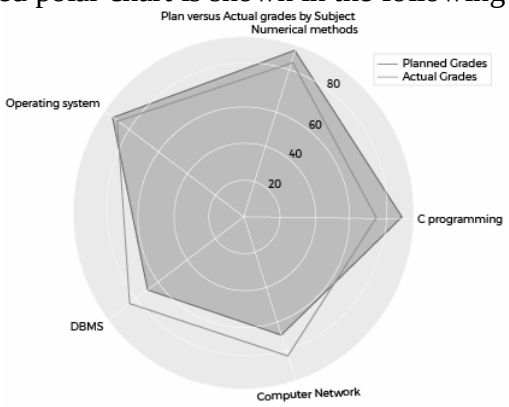
First, let's create the dataset:

1. Let's assume there are five courses in the academic year:
   subjects = ["C programming", "Numerical methods", "Operating system", "DBMS", "Computer Networks"]
2. And a student obtained the following grades in each subject:
   plannedGrade = [90, 95, 92, 68, 68, 90]
3. However, after final examination, these are the grades that the student got:
   actualGrade = [75, 89, 89, 80, 80, 75]

Now that the dataset is ready, let's try to create a polar chart. The first significant step is to initialize the spider plot. This can be done by setting the figure size and polar projection. Note that in the preceding dataset, the list of grades contains an extra entry. This is because it is a circular plot and we need to connect the first point and the last point together to form a circular flow. Hence, we copy the first entry from each list and append it to the list. In the preceding data, the entries 90 and 75 are the first entries of the list respectively. Let's look at each step:

1. Import the required libraries:
   ```
   import numpy as np
   import matplotlib.pyplot as plt
   ```
2. Prepare the dataset and set up theta:
   ```
   theta = np.linspace(0, 2 * np.pi, len(plannedGrade))
   ```
3. Initialize the plot with the figure size and polar projection:
   ```
   plt.figure(figsize = (10,6))
   plt.subplot(polar=True)
   ```
4. Get the grid lines to align with each of the subject names:
   ```
   (lines,labels) = plt.thetagrids(range(0,360,
   int(360/len(subjects))), (subjects))
   ```
5. Use the plt.plot method to plot the graph and fill the area under it:
   ```
   plt.plot(theta, plannedGrade)
   plt.fill(theta, plannedGrade, 'b', alpha=0.2)
   ```
6. Now, we plot the actual grades obtained:
   ```
   plt.plot(theta, actualGrade)
   ```
7. We add a legend and a nice comprehensible title to the plot:
   ```
   plt.legend(labels=('Planned Grades','Actual Grades'),loc=1)
   plt.title("Plan vs Actual grades by Subject")
   ```
8. Finally, we show the plot on the screen:
   ```
   plt.show()
   ```

The generated polar chart is shown in the following screenshot:



1.2.9 Histogram

A histogram is the graphical representation of data where data is grouped into continuous number ranges and each range corresponds to a vertical bar.

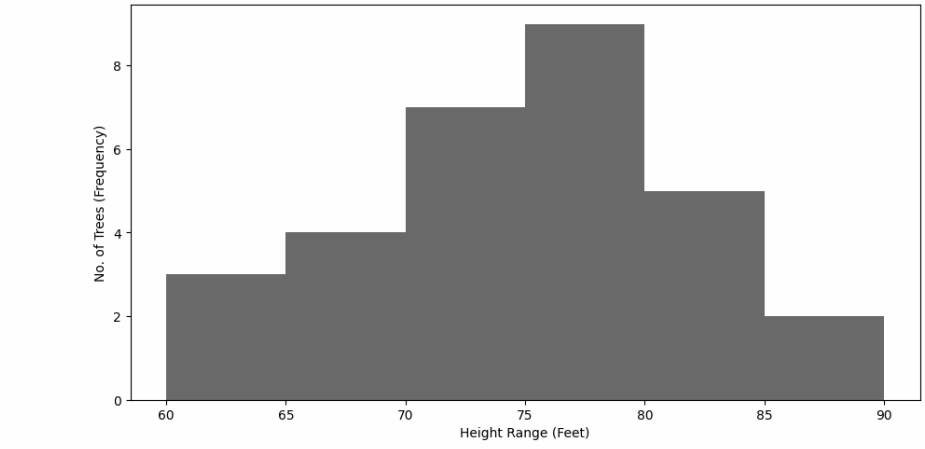
- The horizontal axis displays the number range.
- The vertical axis (frequency) represents the amount of data that is present in each range.

The number ranges depend upon the data that is being used.

Histogram is the easiest manner that can be used to visualize data distributions. Assume that a garden has 30 trees. Each tree is of a different height. The height of the trees (in inches): 61, 63, 64, 66, 68, 69, 69.5, 70, 72, 72.5, 73, 73.5, 74, 74.5, 76, 76.2, 76.5, 77, 77.5, 78, 78.5, 79, 79.2, 80, 81, 82, 83, 84, 85, 87. We can group the data as follows in a frequency distribution table by setting a range:

Height Range (ft)	**Number of Trees (Frequency)**
60 - 65	3
66 - 70	5
71 - 75	6
76 - 80	10
81 - 85	5
86 - 90	1

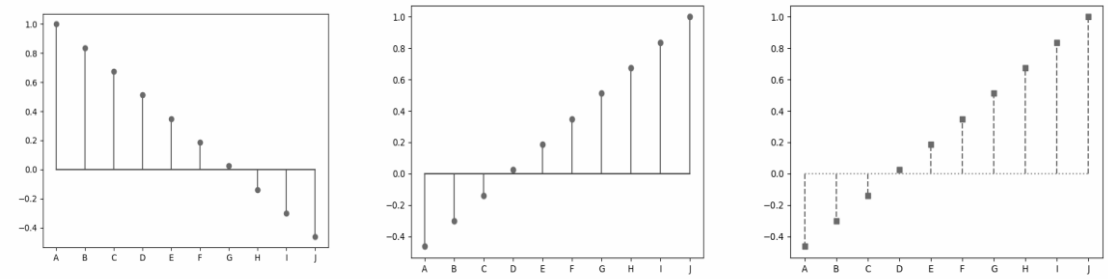
```
import matplotlib.pyplot as plt
import numpy as np
# Creating dataset
a = np.array([61, 63, 64, 66, 68, 69, 69.5, 70, 72,
                  72.5, 73, 73.5, 74, 74.5, 76, 76.2,
                  76.5, 77, 77.5, 78, 78.5, 79, 79.2,
                   80, 81, 82, 83, 84, 85, 87])
# Creating histogram
fig, ax = plt.subplots(figsize =(10, 7))
ax.hist(a, bins = [60,65,70,75,80,85,90])
# Show plot
plt.show()
```



1.2.10 Lollipop chart

A lollipop chart can be used to display ranking in the data. It is similar to an ordered bar chart. It is a variant of bar chart with a circle at the end to highlight the data value. Like bar chart lollipop chart is also used to compare categorical data. Let's consider the carDF dataset.

```
import matplotlib.pyplot as plt
import numpy as np
x = ['A', 'B', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'J']
y = list(np.linspace(1, (np.log(0.2 * np.pi)), 10))
plt.stem(x, y, use_line_collection = True)
plt.show()
y.sort()
plt.stem(x, y, use_line_collection = True)
plt.show()
plt.stem(x, y, markerfmt = 's', linefmt='--', basefmt = ':', use_line_collection=True)
plt.show()
```



1.2.11 Guidelines to choose the best chart

There is no standard that defines which chart we should choose to visualize the data. The guidelines to choose the best chart are:

a. It is important to understand what type of data we have.

b. If we have continuous variables, then a histogram would be a good choice.

c. If we want to show ranking, an ordered bar chart would be a good choice.

d. The chart that effectively conveys the right and relevant meaning of the data without actually distorting the facts must be chosen.

e. Simplicity is best. It is considered better to draw a simple chart that is comprehensible than to draw sophisticated ones that require several reports and texts in order to understand them.

f. Choose a diagram that does not overload the audience with information. Our purpose should be to illustrate abstract information in a clear way.

Purpose	Charts
Show correlation	Scatter plot, Correlogram, Pairwise plot, Jittering with strip plot, Counts plot, Marginal histogram, Scatter plot with a line of best fit, Bubble plot with circling
Show deviation	Area chart, Diverging bars, Diverging texts, Diverging dot plot, Diverging lollipop plot with markers
Show distribution	Histogram for continuous variable, Histogram for categorical variable, Density plot, Categorical plots, Density curves with histogram, Population pyramid, Violin plot, Joy plot, Distributed dot plot, Box plot
Show composition	Waffle chart, Pie chart, Treemap, Bar chart

INDRA GANESAN COLLEGE OF ENGINEERING

IG Valley, Manikandam, Tiruchirappalli, Tamil Nadu – 620 012, India
(Approved by AICTE, New Delhi, Affiliated to Anna University, Chennai-25)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Assignment Question Paper

Assignment – 01			Date of Issue:	13.09.2022	Marks	10
Course code	CS3352	Course Title	Foundation of Data Science			
Year	II	Semester/Section	III	Date of Submission:	23.09.22	

Q.No	Questions	CO
1	The IQ scores for a group of 35 high school dropouts are as follows (a) Construct a frequency distribution for grouped data. (b) Specify the real limits for the lowest class interval in this frequency distribution.	C203.2
2	GRE scores for a group of graduate school applicants are distributed as follows: (i) Convert to a relative frequency distribution. When calculating proportions, round numbers to two digits to the right of the decimal point. (ii) Convert to a cumulative frequency distribution. (iii)Convert to a cumulative percent frequency distribution.	C203.2

91	85	84	79	80	112	110
87	96	75	86	104	90	109
95	71	105	90	77	90	94
123	80	100	93	108	98	100
98	69	99	95	90	89	103

GRE	725-749	700-724	675-699	650-674	625-649	600-624	575-599	550-574	525-549	500-524	475-499	Total1
f	1	3	14	30	34	42	30	27	13	4	2	200

Name and Signature of the Faculty Incharge

HoD/CSE

Dr. G. Balakrishnan, M.E., Ph.D.,
Principal
Indra Ganesan College of Engineering
IG Valley, Madurai Main Road
Manikandam, Trichy-620 012.

INDRA GANESAN COLLEGE OF ENGINEERING

IG Valley, Manikandam, Tiruchirappalli, Tamil Nadu – 620 012, India
(Approved by AICTE, New Delhi, Affiliated to Anna University, Chennai-25)
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Assignment Answer Sheet

Name of the Student : T. Hema.

AU Register Number: 811221104011

Assignment – 01			Date of Issue: 13.09.2022	Marks	10
Course code	CS3352	Course Title	Foundation of Data Science		
Year	2022	Semester	II	Date of Submission:	23.09.22

Q.No	Questions	CO
1	The IQ scores for a group of 35 high school dropouts are as follows (a) Construct a frequency distribution for grouped data. (b) Specify the real limits for the lowest class interval in this frequency distribution.	C203.2
2	GRE scores for a group of graduate school applicants are distributed as follows: (i) Convert to a relative frequency distribution. When calculating proportions, round numbers to two digits to the right of the decimal point. (ii) Convert to a cumulative frequency distribution. (iii)Convert to a cumulative percent frequency distribution.	C203.2

91	85	84	79	80	112	110
87	96	75	86	104	90	109
95	71	105	90	77	90	94
123	80	100	93	108	98	100
98	69	99	95	90	89	103

GRE	725-749	700-724	675-699	650-674	625-649	600-624	575-599	550-574	525-549	500-524	475-499	Total1
f	1	3	14	30	34	42	30	27	13	4	2	200

Mark Allocation

Rubrics	Marks Allocated	Marks obtained
Content Quality	6	5
Presentation Quality	2	2
Timely submission	2	2
Total marks	10	9

T. Hema

Name and Signature of the Faculty Incharge

Dr. G. Balakrishnan, M.E., Ph.D.,
Principal
Indra Ganesan College of Engineering
IG Valley, Madurai Main Road
Manikandam, Trichy-620 012.

D. Ruba

HoD/CSE

INDRA GANESAN COLLEGE OF ENGINEERING

IG Valley, Manikandam, Tiruchirappalli, Tamil Nadu – 620 012, India
(Approved by AICTE, New Delhi, Affiliated to Anna University, Chennai-25)

IQAC Academic Audit Form

ACADEMIC YEAR: 2022 - 2023 ODD SEMESTER

Name of Department :	CSE	Year / Sem /	2 / III	No. of Students Registered :	36
Details of Examination :	IA Test -1 / IA Test -2 / IA Test -3 / Model Test				

S.No.	Course Code	List of Reg.No Verified	Course Log Book Verified (Y / N)	Course File Verified (Y / N)	No of students Attended	No of Absentees	No of Failures	Pass %	Remarks
1	CS3351	811221104005	Yes	Yes	35	01	2	94%	Presented well
2	CS3301	811221104024	Yes	Yes	36	-	3	91%	Big Question Points to be Improved.
3	CS3391	811221104011	Yes	Yes	35	01	2	94%	Well Answered.
4	CS3352	811221104006	Yes	Yes	36	-	3	91%	Answered all the points good
5	CS3354	811221104026	Yes	Yes	36	-	2	94%	Neatly Presented.

Verified by

External Member Name and Signature:	Sheebha
Internal Member Name and Signature:	K. Sathiyaraj

Overall Remarks:

D. Vinodh
HoD/ CSE

IQAC Co-ordinator

Principal

Dr. G. Balakrishnan, M.E., Ph.D.
Principal
Indra Ganesan College of Engineering
IG Valley, Madurai Main Road
Manikandam, Trichy-620 012.

INDRA GANESAN COLLEGE OF ENGINEERING

IG Valley, Manikandam, Tiruchirappalli, Tamil Nadu – 620 012, India
(Approved by AICTE, New Delhi, Affiliated to Anna University, Chennai-25)

STUDENT FEEDBACK ON FACULTY
THEORY COURSE

ACADEMIC YEAR: 2022-2023 ODD SEMESTER

Name of Department:	CSE	Year / Sem:	2 / III	Faculty Name	T. Sugashini
Subject Code & Name	CS3352 - Foundation of data science				

S.No.	QUESTIONS	Excellent	Very Good	good	Satisfactory	Somewhat Satisfactory	Not Satisfactory
		5	4	3	2	1	0
1.	Delivery of Lectures by Interactive Communication	✓					
2.	Use of Teaching Aids and ICT		✓				
3.	Level of Preparedness & Knowledge Level	✓					
4.	Involvement in mentoring and guiding	✓					
5.	Effective Time management	✓					
6.	Is the teacher completing syllabus as per lecture schedule?	✓					
7.	Is the teacher distributing answer scripts of students as per schedule?	✓					
8.	Is the teacher addressing grievances on answer scripts of IA while distributing?	✓					
9.	Is the teacher covering content beyond syllabus (CBS)?	✓					
10.	Is the teacher punctual to class?	✓					

HoD/CSE

IQAC Co-ordinator

Principal

Dr. G. Ba...krishna, M.E., Ph.D.,
Principal
Indra Ganesan College of Engineering
IG Valley, Madurai Main Road
Manikandam, Trichy-620 012.

Indra Ganesan
COLLEGE OF ENGINEERING

IGCE/EXAMCELL/IA/2022-23/ODD/UT/001

INTERNAL ASSESSMENT TEST – I

Test Time: (FN) 11.30 am to 1.00 pm

DATE	YEAR	10/10/2022	11/10/2022	12/10/2022	13/10/2022	14/10/2022	15/10/2022
BRANCH	SESSION	FN	FN	FN	FN	FN	FN
AI	II	MA3301	AI3301	AI3302	AI3303	ME3491	CE3351
AIDS	II	MA3354	CS3351	AD3301	AD3391	AD3351	AL3391
CSE	II	MA3354	CS3351	CS3352	CS3301	CS3391	
EEE	II	MA3303	EE3301	EE3302	EC3301	EE3303	CS3353
ECE	II	MA3355	EC3351	EC3352	EC3353	EC3354	CS3354
MECH	II	MA3351	ME3351	ME3391	CE3391	ME3392	ME3393
IT	II	MA3354	CS3351	CS3352	CD3291	CS3391	

EXAM CELL CO ORDINATOR

PRINCIPAL

COPY TO

1. The Director for favour of kind information
2. The Principal (file copy)
3. All HoDs :request to circulate among their faculty members
4. Exam cell file
5. Notice Board (Lab Block)

Vision of the Institution: *To evolve as a centre of excellence in Engineering, Technology and Management with distinctive research capabilities and to transform the students into knowledgeable, skilled professionals with high ethical values to cater the needs of the society*

Dr. G. Balakrishnan, M.E., Ph.D.,
Principal
Indra Ganesan College of Engineering
IG Valley, Madurai Main Road
Manikandam, Trichy-620 012.

Register Number: | | | | | | | | | | | | |

INDRA GANESAN COLLEGE OF ENGINEERING

IG Valley, Manikandam, Tiruchirappalli, Tamil Nadu – 620 012, India

(Approved by AICTE, New Delhi and affiliated to Anna University, Chennai)

Internal Assessment Exam - I					
Course code	CS3352	Course Title	Foundations of Data Science		
Regulation	2021	Duration	90 minutes	Academic Year	2022 - 2023
Year	II	Semester	III	Department	CSE

COURSE OUTCOMES

CO1:	Explain the data science process and the basic concept of data science fundamentals
CO2:	Illustrate to convert the values from the normal distribution into z scores using data with tables, graphs, averages, and variability
CO3:	Examine the data to describe the relationship by examining the form, direction, and strength of the association by quantitatively and qualitatively.
CO4:	Examine the NumPy libraries to perform a wide variety of high-level mathematical functions that operate on the arrays and matrices.
CO5:	Examine the Pandas libraries for analyzing, cleaning, exploring, and manipulating data.
CO6:	Explain the visualization libraries in Python to identify patterns, trends, and outliers in large data sets along with its libraries, graphs, charts, and histogram

Q.No.	Question	CO	BTS
	PART A (Answer all the Questions 10 x 2 = 20 Marks)		
1	Define data science?	1	1
2	Define streaming data	1	1
3	Define outliers?	1	1
4	Define Sanity Check?	1	1
5	List the disadvantage of combining data?	1	1
6	Define Key-Value stores	1	1
7	Define frequency distribution?	2	1
8	Define Percentile Ranks	2	1
9	Explain Histogram?	2	1
10	Define Mean, Median and Mode	2	2
	PART B (Answer all the Questions 2 x 10 = 20 Marks)		
11a	Describe the research goal, retrieving data and Data preparation process in Data Science	1	2
	OR		
11b	Explain the benefits, uses, and facets of data	1	2
12a	Describe the architecture of Data Warehouse	1	2
	OR		
12b	Explain the Data Exploration, data modelling, and presentation process in Data Science	1	2
	PART C (Answer all the Questions 1 x 10 = 10 Marks)		
13a	GRE scores for a group of graduate school applicants are distributed as follows: (i) Convert to a relative frequency distribution. When calculating proportions, round numbers to two digits to the right of the decimal point. (ii) Convert to a cumulative frequency distribution. (iii) Convert to a cumulative percent frequency distribution.	2	2

GRE	725-749	700-724	675-699	650-674	625-649	600-624	575-599	550-574	525-549	500-524	475-499	Total1
f	1	3	14	30	34	42	30	27	13	4	2	200

	OR		
13b	Explain the different types of data and variables with example	2	2

Course Faculty
(Name /Sign / Date)

Dr. G. Balakrishnan, M.E., Ph.D.,
Principal
Indra Ganesan College of Engineering
IG Valley, Madurai Main Road
Manikandam, Trichy-620 012.

HoD
(Name /Sign / Date)

CS3352 – Foundation of Data Science
Internal Assessment 1 Test
Question with Key
Part A

1. Define data science?
   Data science is an interdisciplinary field that seeks to extract knowledge or insights from various forms of data.
2. Define streaming data
   Streaming data is data that is generated continuously by thousands of data sources, which typically send in the data records simultaneously and in small sizes (order of Kilobytes).
3. Define outliers?
   An outlier is an observation that lies an abnormal distance from other values in a random sample from a population
4. Define Sanity Check?
   A sanity check or sanity test is a basic test to quickly evaluate whether a claim or the result of a calculation can possibly be true.
5. List the disadvantage of combining data?
   Data from different sources may be stored in different formats, making it difficult to create a seamless integration. This may require additional time and resources for data cleaning and validation.
6. Define Key-Value stores
   A key-value store, or key-value database is a simple database that uses an associative array (think of a map or dictionary) as the fundamental data model where each key is associated with one and only one value in a collection.
7. Define frequency distribution?
   Frequency distribution is a representation, either in a graphical or tabular format, that displays the number of observations within a given interval. The interval size depends on the data being analyzed and the goals of the analyst.
8. Define Percentile Ranks
   The percentile rank of a score is the percentage of scores in its frequency distribution that are equal to or lower than it
9. Explain Histogram?
   A histogram is a graphical representation of data points organized into user-specified ranges.
10. Define Mean, Median and Mode
    The arithmetic mean is found by adding the numbers and dividing the sum by the number of numbers in the list. This is what is most often meant by an average. The median is the middle value in a list ordered from smallest to largest. The mode is the most frequently occurring value on the list

Part B

11. Describe the research goal, retrieving data and Data preparation process in Data Science
    Defining research goals
    Spend time understanding the goals and context of your research
    Create a project charter
    Retrieving data
    Internal Data
    External Data
    Data Preparation (Cleansing, Integrating, Transforming Data)
    Cleansing data
    Overview of common errors
    Data Entry Errors
    Redundant Whitespace
    Fixing Capital Letter Mismatches
    Impossible Values and Sanity Checks
    Outliers
    Dealing with Missing Values
    Integrating data
12. Explain the benefits, uses, and facets of data
    Benefits and uses of data science
    Data science and big data are used almost everywhere in both commercial and noncommercial Settings

Dr. G. Balakrishnan, M.E., Ph.D.,
Principal
Indra Ganesan College of Engineering
IG Valley, Madurai Main Road
Manikandam, Trichy-620 012.

- Commercial companies in almost every industry use data science and big data to gain insights into their customers, processes, staff, completion, and products.
- Many companies use data science to offer customers a better user experience, as well as to cross-sell, up-sell, and personalize their offerings.
- Governmental organizations are also aware of data's value. Many governmental organizations not only rely on internal data scientists to discover valuable information, but also share their data with the public.
- Nongovernmental organizations (NGOs) use it to raise money and defend their causes.
- Universities use data science in their research but also to enhance the study experience of their students. The rise of massive open online courses (MOOC) produces a lot of data, which allows universities to study how this type of learning can complement traditional classes.

Facets of data

In data science and big data you'll come across many different types of data, and each of them tends to require different tools and techniques. The main categories of data are these:

- Structured
- Unstructured
- Natural language
- Machine-generated
- Graph-based
- Audio, video, and images
- Streaming

13. Describe the architecture of Data Warehouse

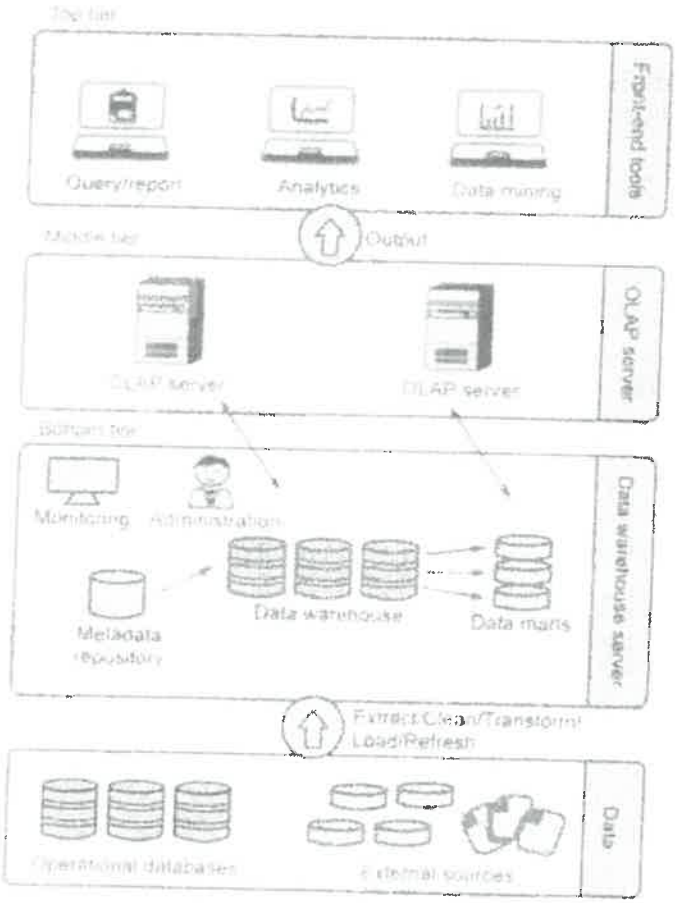


Fig. 1.11.1. Three tier architecture

14. Explain the Data Exploration, data modelling, and presentation process in Data Science

•The **visualization techniques** you use in this phase range from simple line graphs or histograms, to more complex diagrams such as Sankey and network graphs.

Data Modelling

- Selection of a modeling technique and variables to enter in the model
- Execution of the model
- Diagnosis and model comparison

Presenting findings and building applications

Dr. G. Balakrishnan, M.E., Ph.D.,
Principal
Indra Ganesan College of Engineering
IG Valley, Madurai Main Road
Manikandam, Trichy-620 012.

15. GRE scores for a group of graduate school applicants are distributed as follows:

(i) Convert to a relative frequency distribution. When calculating proportions, round numbers to two digits to the right of the decimal point.

GRE	RELATIVE f
725–749	.01
700–724	.02
675–699	.07
650–674	.15
625–649	.17
600–624	.21
575–599	.15
550–574	.14
525–549	.07*
500–524	.02
475–499	.01
	Totals 1.02

**From 13/200 = .065, which rounds to .07.*

(ii) Convert to a cumulative frequency distribution.

(iii) Convert to a cumulative percent frequency distribution.

GRE	(a) CUMULATIVE f	(b) CUMULATIVE PERCENT(%)
725–749	200	100
700–724	199	100
675–699	196	98
650–674	182	91
625–649	152	76
600–624	118	59
575–599	76	38
550–574	46	23
525–549	19	10
500–524	6	3
475–499	2	1

16. Explain the different types of data and variables with example

THREE TYPES OF DATA

Qualitative data

Ranked data.

Quantitative data

TYPES OF VARIABLES

Discrete and Continuous Variables

Independent and Dependent Variables

Signature of the Faculty

HOD/CSE

Dr. G. Balakrishnan, M.E., Ph.D.,
Principal
Indra Ganesan College of Engineering
IG Valley, Madurai Main Road
Manikandam, Trichy-620 012.

INDRA GANESAN COLLEGE OF ENGINEERING

IG Valley, Manikandam, Tiruchirappalli, Tamil Nadu – 622 012, India
(Approved by AICTE, New Delhi and affiliated to Anna University, Chennai)

Internal Assessment Test Answer Book

Name	Rosika A.V.R			**Year/ Semester/Section**	2022/II
Batch No.	8122110x1030	**Date/Session**	23.09.22	**Department**	CSE
Course code	CS3352	**Course Title**	Foundation of Data Science		
Internal Assessment Test	**IAT 1** ☑	**IAT 2** ☐	**IAT 3** ☐	**Model** ☐	
Name and Signature of the Invigilator with date	Richard Sathiresany.				

Instruction to the Student: Put tick mark to the question attended in the column against question.

Part A			Part B / Part C						Total Marks
Q. No.	✓	Marks	Q. NO.	✓	a Marks	✓	b Marks		
1	✓	2	11	✓	10				10
2	✓	2	12			✓	10		10
3	✓	2	13	✓	9				9
4	✓	2	14						
5	✓	1	15						
6	✓	2	16						
7	✓	2						Total	29
8	✓	2							
9	✓	2							
10	✓	2							
Total		19	Grand Total: Vg. good			Name and Signature of the Examiner with date: T. Santi			

To be filled by the examiner

Course Outcomes	1	2	3	4	5	6	Total
Marks allotted	32	18					50
Marks Obtained	31	17					48

IQAC Audit - Remarks

Name and Signature of the IQAC member

Dr. G. Balakrishnan, M.E., Ph.D.,
Principal
Indra Ganesan College of Engineering
IG Valley, Madurai Main Road
Manikandam, Trichy-620 012.

INDRA GANESAN COLLEGE OF ENGINEERING

IG VALLEY, MANIDANDAM, TIRUCHIRAPPALLI – 620 012

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ACADEMIC YEAR 2022 – 2023 (ODD SEMESTER)

STUDENTS MARK STATEMENT- CO BASED

INTERNAL ASSESSMENT TEST-I

SUBJECT CODE &TITLE: CS3352 & Foundations of Data Science

YEAR/SEM: II/III **MONTH & YEAR: 12.10.2022**

S.NO	REG NO	STUDENT NAME	CO203.1 (32)	CO203.2 (18)	TOTAL (50)	TOTAL (100)
1.	811221104001	AKSHAY K	26	10	36	72
2.	811221104002	BHARATHKUMAR S M	25	12	37	74
3.	811221104004	DHINESH C	50	5	20	40
4.	811221104005	EYARKAI KAMALI R	30	16	46	92
5.	811221104006	GOKULNATH P R	28	14	42	84
6.	811221104007	HARIHARASWAMY M	14	12	26	52
7.	811221104008	HARISH R	23	6	29	58
8.	811221104009	HARRISH M	15	10	25	50
9.	811221104011	HEMA T	29	14	43	86
10.	811221104012	JACOP ANTONY L	22	8	30	60
11.	811221104013	JEEVANANTHAM S	25	9	34	68
12.	811221104014	KATHIRVEL K	20	8	28	56
13.	811221104015	KEERTHANA J	24	11	35	70
14.	811221104018	MANIKANDAN N	29	14	46	86
15.	811221104020	MOHAMED THOUFIK U	26	12	38	76
16.	811221104023	NAVEENKUMAR S	20	9	29	58
17.	811221104024	NITHYA A	27	12	39	78
18.	811221104025	POORNIMA C	26	10	36	72
19.	811221104026	PRASANNA BALAJI C	29	13	42	84
20.	811221104028	RAJAPUSHPAM V	26	11	37	74
21.	811221104029	REETHIKA R	29	15	44	88
22.	811221104030	RESIKA A V R	31	17	48	96
23.	811221104031	SANTHOSH P	12	7	19	38
24.	811221104032	SARAVANAPERUMAL S	17	9	26	52
25.	811221104034	SELVALAKSHMI G	22	11	33	66
26.	811221104035	SIVAKUMAR P	20	11	31	62
27.	811221104036	SUDHAKARAN V	19	8	27	54
28.	811221104037	SUGAVANESHWARAN S	27	13	40	80

Dr. G. Balakrishnan, M.E., Ph.D.,
Principal
Indra Ganesan College of Engineering
IG Valley, Madurai Main Road
Manikandam, Trichy-620 012.

29.	811221104038	SUMAIYA BEGAM S	26	12	38	96
30.	811221104040	SURUTHI Y	28	14	42	84
31.	811221104041	SURYA D	20	9	29	58
32.	811221104043	SYED ANWAR S	25	10	35	70
33.	811221104045	VASANTHAVEL S	16	18	24	48
34.	811221104046	VENGADESWARI M	24	10	34	68
35.	811221104048	VISHWA S	23	9	32	64
36.	811221104049	YOGAPRIYA N	27	10	37	74

MARKS RANGE:

<20	20-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100
0	0	3	4	9	7	9	6	2

Total No.of Candidates Present	36
Total No.of Candidates Absent	0
Total No.of Students Pass	33
Total No. of Students Fail	3
Percentage of Pass	91%

STAFF INCHARGE **HoD/CSE** **PRINCIPAL**

Dr. G. Balakrishnan, M.E., Ph.D.,
Principal
Indra Ganesan College of Engineering
IG Valley, Madurai Main Road
Manikandam, Trichy-620 012.

INDRA GANESAN COLLEGE OF ENGINEERING

IG Valley, Manikandam, Tiruchirappalli, Tamil Nadu – 620 012, India

(Approved by AICTE, New Delhi, Affiliated to Anna University, Chennai-25)

DEPARTMENT OF COMPUTER SCEINCE AND ENGINEERING

ROOT CAUSE ANALYSIS

Name of the Faculty : T. SUGASHINI
Degree & Program : B.E CSE
IA Test : IA1
Target : 95 %

Course Code & Name : CS3352 & FOUNDATION OF DATA SCIENCE
Semester :III
Exam/Month & Year : 12.10.2022
Achieved : 91 %

S.NO	BATCH NO	NAME OF THE STUDENT	CAUSES FOR FAILURE	SIGNATURE OF THE STUDENT WITH DATE	CORRECTIVE ACTION TAKEN	PREVENTIVE ACTION TAKEN	FOLLOWUP STATUS	REMARKS OF THE HOD
1.	811221104004	DHINESH C.	Two marks not attended.	C.Dhi	Relest	Assignment		
2.	811221104031	SANTHOSH M	content not enough.	M.S	Relest	Assignment		
3.	811221104045	VASANTHAVEL S	Not attended all quest	S.V	Relest	Assignment		
4.								
5.								

Signature of the Faculty Member

Dr. C. Balakrishnan M.E., Ph.D.,
Principal
Indra Ganesan College of Engineering
IG Valley, Madurai Main Road
Manikandam, Trichy-620 012.

Signature of the HoD/CSE

Register Number:

INDRA GANESAN COLLEGE OF ENGINEERING

IG Valley, Manikandam, Tiruchirappalli, Tamil Nadu – 620 012, India

(Approved by AICTE, New Delhi and affiliated to Anna University, Chennai)

Internal Assessment Retest - I		**Date/Session**	20-10-2022	**Marks**	50
Course code	CS3352	**Course Title**	Foundations of Data Science		
Regulation	2021	**Duration**	90 minutes	**Academic Year**	2022 - 2023
Year	II	**Semester**	III	**Department**	CSE

COURSE OUTCOMES

CO1:	Explain the data science process and the basic concept of data science fundamentals
CO2:	Illustrate to convert the values from the normal distribution into z scores using data with tables, graphs, averages, and variability
CO3:	Examine the data to describe the relationship by examining the form, direction, and strength of the association by quantitatively and qualitatively.
CO4:	Examine the NumPy libraries to perform a wide variety of high-level mathematical functions that operate on the arrays and matrices.
CO5:	Examine the Pandas libraries for analyzing, cleaning, exploring, and manipulating data.
CO6:	Explain the visualization libraries in Python to identify patterns, trends, and outliers in large data sets along with its libraries, graphs, charts, and histogram

Q.No.	Question	CO	BTS
	PART A		
	(Answer all the Questions 10 x 2 = 20 Marks)		
1	Define data mining?	1	1
2	Define streaming data	1	1
3	Define outliers?	1	1
4	Differentiate structure data and unstructured data	1	1
5	List the disadvantage of combining data?	1	1
6	Define Key-Value stores	1	1
7	Define frequency distribution?	2	1
8	Define Percentile Ranks	2	1
9	Explain Histogram?	2	2
10	Define Mean, Median and Mode	2	1
	PART B		
	(Answer all the Questions 3 x 10 = 30 Marks)		
11a	Describe the research goal, retrieving data and Data preparation process in Data Science	1	2
	OR		
11b	Describe the architecture of Data Warehouse	1	2
12a	Explain the benefits, uses, and facets of data	1	2
	OR		
12b	Explain the Data Exploration, data modelling, and presentation process in Data Science	1	2
13a	The IQ scores for a group of 35 high school dropouts are as follows (a) Construct a frequency distribution for grouped data. (b) Specify the real limits for the lowest class interval in this frequency distribution.	2	2

91	85	84	79	80	112	110
87	96	75	86	104	90	109
95	71	105	90	77	90	94
123	80	100	93	108	98	100
98	69	99	95	90	89	103

	OR		
13b	Explain the different types of data and variables with example	2	2

Course Faculty
(Name /Sign / Date)

Dr. G. Balakrishnan, M.E., Ph.D.,
Principal
Indra Ganesan College of Engineering
IG Valley, Madurai Main Road
Manikandam, Trichy-620 012.

HoD
(Name /Sign / Date)

CS3352 – Foundation of Data Science
Internal Assessment Retest 1
Question with Key
Part A

1. Define mining?
   Data mining is the process of sorting through large data sets to identify patterns and relationships that can help solve business problems through data analysis. Data mining techniques and tools enable enterprises to predict future trends and make more-informed business decisions.
2. Define streaming data
   Streaming data is data that is generated continuously by thousands of data sources, which typically send in the data records simultaneously and in small sizes (order of Kilobytes).
3. Define outliers?
   An outlier is an observation that lies an abnormal distance from other values in a random sample from a population
4. Differentiate structure data and unstructured data

Structured Data	Unstructured Data
In this type of data, the data is stored in processed form or containing labels in which searching of data is easy.	In this type, the data is stored in unprocessed form or raw form in which searching is complex.
This form of data is generally used to store quantitative data such as height, weight, age.	This form of data is used to store qualitative data such as invoices, records, and media-related data.
To store such types of data, data warehouses are used.	To store unstructured data, data lakes are used.
In this form, the data is stored in predefined formats support by underlying database architecture.	In this form, the data can be stored in different formats.
Several analytical tools are available for mining structured data.	There are no tools present currently for mining unstructured data.

5. List the disadvantage of combining data?
   Data from different sources may be stored in different formats, making it difficult to create a seamless integration. This may require additional time and resources for data cleaning and validation.
6. Define Key-Value stores
   A key-value store, or key-value database is a simple database that uses an associative array (think of a map or dictionary) as the fundamental data model where each key is associated with one and only one value in a collection.
7. Define frequency distribution?
   Frequency distribution is a representation, either in a graphical or tabular format, that displays the number of observations within a given interval. The interval size depends on the data being analyzed and the goals of the analyst.
8. Define Percentile Ranks
   The percentile rank of a score is the percentage of scores in its frequency distribution that are equal to or lower than it
9. Explain Histogram?
   A histogram is a graphical representation of data points organized into user-specified ranges.
10. Define Mean, Median and Mode
    The arithmetic mean is found by adding the numbers and dividing the sum by the number of numbers in the list. This is what is most often meant by an average. The median is the middle value in a list ordered from smallest to largest. The mode is the most frequently occurring value on the list

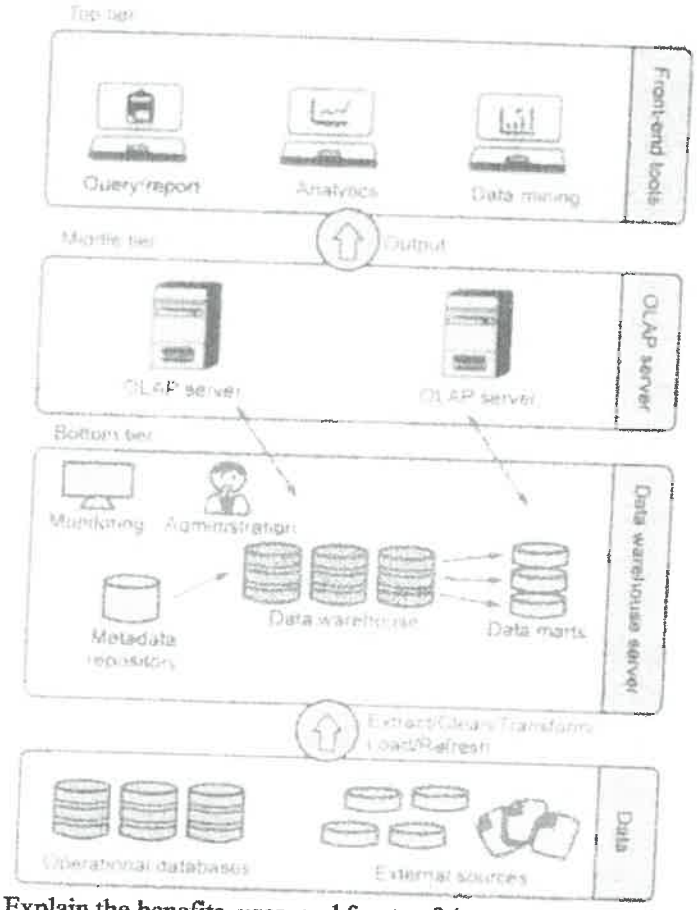
Part B

11. Describe the research goal, retrieving data and Data preparation process in Data Science
    Defining research goals
    Spend time understanding the goals and context of your research
    Create a project charter
    Retrieving data
    Internal Data
    External Data
    Data Preparation (Cleansing, Integrating, Transforming Data)
    Cleansing data
    Overview of common errors
    Data Entry Errors

Dr. G. Balakrishnan, M.E., Ph.D.,
Principal
Indra Ganesan College of Engineering
IG Valley, Madurai Main Road
Manikandam, Trichy-620 012.

Redundant Whitespace
Fixing Capital Letter Mismatches
Impossible Values and Sanity Checks
Outliers
Dealing with Missing Values
Integrating data

12. Describe the architecture of Data Warehouse



13. Explain the benefits, uses, and facets of data

Benefits and uses of data science

Data science and big data are used almost everywhere in both commercial and noncommercial Settings

- Commercial companies in almost every industry use data science and big data to gain insights into their customers, processes, staff, completion, and products.
- Many companies use data science to offer customers a better user experience, as well as to cross-sell, up-sell, and personalize their offerings.
- Governmental organizations are also aware of data's value. Many governmental organizations not only rely on internal data scientists to discover valuable information, but also share their data with the public.
- Nongovernmental organizations (NGOs) use it to raise money and defend their causes.
- Universities use data science in their research but also to enhance the study experience of their students. The rise of massive open online courses (MOOC) produces a lot of data, which allows universities to study how this type of learning can complement traditional classes.

Facets of data

In data science and big data you'll come across many different types of data, and each of them tends to require different tools and techniques. The main categories of data are these:

- Structured
- Unstructured
- Natural language
- Machine-generated
- Graph-based
- Audio, video, and images
- Streaming

14. Explain the Data Exploration, data modelling, and presentation process in Data Science

•The **visualization techniques** you use in this phase range from simple line graphs or histograms, to more complex diagrams such as Sankey and network graphs.

Data Modelling

- Selection of a modeling technique and variables to enter in the model
- Execution of the model
- Diagnosis and model comparison

Presenting findings and building applications

15. The IQ scores for a group of 35 high school dropouts are as follows

(a) Construct a frequency distribution for grouped data.

(a) Calculating the class width,

$$\frac{123-69}{10}=\frac{54}{10}=5.4$$

Round off to a convenient number, such as 5

IQ	TALLY*	f
120–124	/	1
115–119		0
110–114	//	2
105–109	///	3
100–104	////	4
95–99	//// /	6
90–94	//// //	7
85–89	////	4
80–84	///	3
75–79	///	3
70–74	/	1
65–69	/	1
	Total	35

(b) Specify the real limits for the lowest class interval in this frequency distribution

64.5–69.5

16. Explain the different types of data and variables with example

Three types of data

Qualitative data

Ranked data.

Quantitative data

Types of Variables

Discrete and Continuous Variables

Independent and Dependent Variables

Signature of the Faculty

HOD/CSE

Dr. G. Balakrishnan, M.E., Ph.D.,
Principal
Indra Ganesan College of Engineering
IG Valley, Madurai Main Road
Manikandam, Trichy-620 012.

INDRA GANESAN COLLEGE OF ENGINEERING

IG Valley, Manikandam, Tiruchirappalli, Tamil Nadu – 622 012, India
(Approved by AICTE, New Delhi and affiliated to Anna University, Chennai)

Internal Assessment Retest Answer Book

Name	S. Vasanthavel			**Year/ Semester/Section**	II / III
Batch No.	811221104045	**Date/Session**	22.10.2022	**Department**	CSE
Course code	CS3352	**Course Title**	Foundation of Data Science		
Internal Assessment Test	**IAT 1** Retest ☑	**IAT 2** ☐	**IAT 3** ☐	**Model** ☐	
Name and Signature of the Invigilator with date	G Paneeth.				

Instruction to the Student: Put tick mark to the question attended in the column against question.

Part A			Part B / Part C						Total Marks
Q. No.	✓	Marks	Q. NO.	✓	a Marks	✓	b Marks		
1		1	11		7				7
2		1	12				–		
3		2	13	✓	10				10
4		2	14						
5		2	15						
6		1	16						17
7		1					Total		
8		2							
9		1							
10		2							
Total		15							

32
Grand Total

T. Sasi
Name and Signature of the Examiner with date

To be filled by the examiner

Course Outcomes	1	2	3	4	5	6	Total
Marks allotted	32	18					50
Marks Obtained	22	10					32

IQAC Audit - Remarks

Dr. C. Balakrishnan, M.E., Ph.D.,
Principal
Indra Ganesan College of Engineering
IG Valley, Madurai Main Road
Manikandam, Trichy-620 012.

Name and Signature of the IQAC member

INDRA GANESAN COLLEGE OF ENGINEERING

IG VALLEY, MANIDANDAM, TIRUCHIRAPPALLI – 620 012

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ACADEMIC YEAR 2022 – 2023 (ODD SEMESTER)

STUDENTS MARK STATEMENT- CO BASED

INTERNAL ASSESSMENT RETEST-I

SUBJECT CODE &TITLE: CS3352 & Foundations of Data Science

YEAR/SEM: II/III **MONTH & YEAR:** 20.10.2022

S.NO	REG NO	STUDENT NAME	CO203.1 (32)	CO203.2 (18)	TOTAL (50)	TOTAL (100)
1.	811221104004	DHINESH C	23	6	29	58
2.	811221104031	SANTHOSH P	23	9	32	64
3.	811221104045	VASANTHAVEL S	22	10	32	64
4.						
5.						

MARKS RANGE:

<20	20-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100
0	0	0	0	1	2	0	0	0

Total No.of Candidates Present	3
Total No.of Candidates Absent	0
Total No.of Students Pass	3
Total No. of Students Fail	0
Percentage of Pass	100%

STAFF INCHARGE **HoD/CSE** **PRINCIPAL**

Dr. G. Balakrishnan, M.E., Ph.D.
Principal
Indra Ganesan College of Engineering
IG Valley, Madurai Main Road
Manikandam, Trichy-620 012.