ANNA UNIVERSITY, CHENNAI UNIVERSITY DEPARTMENTS REGULATIONS – 2023 FACULTY OF INFORMATION & COMMUNICATION ENGINEERING RAMANUJAN COMPUTING CENTRE MINOR DEGREE ON 'DATA SCIENCE'

Courses for minor degree on 'Data Science'

SL. NO	COURSE CODE	COURSE TITLE	CONTACT PERIODS	L	т	Ρ	С
1.	CSM507	Foundations of Data Science with Python	3	3	0	0	3
2.	CSM508	Machine Learning for Data Science	3	3	0	0	3
3.	CSM509	Data Visualization	3	3	0	0	3
4.	CSM510	Data Security & Privacy	3	3	0	0	3
5.	CSM511	Big data analytics	3	3	0	0	3
6.	CSM512	Exploratory Analysis	3	3	0	0	3
		Total Credits					18

CS	M507	FOUNDATIONS OF DATASCIENCE	L	Т	Р	С		
		WITH PYTHON	3	0	0	3		
Cour	se Obiec	tives:		<u> </u>		<u> </u>		
•	To unde	erstand fundamentals and the process of data	scie	ence				
•	To com	prehend different types and representation of	data	a and	d ana	lvze	ther	n.
•	To apply	v inferential techniques to extrapolate informat	tion	from	n the a	availa	able	•
	data.							
•	To utiliz	e the Python libraries for Data Wrangling.						
•	To inter	pret data and present it using visualization libr	arie	s in	Pytho	on.		
Unit I	INT	RODUCTION						9
Data	Science:	Benefits and uses - facets of data - Data Sc	iena	ce P	roces	s: O	verv	view –
Defini	ing resea	rch goals – Retrieving data – Data preparation	– Ex	xploi	ratory	Data	a an	alysis
– buil	d the mo	del – presenting findings and building applicati	ions	;				
Unit I	I DES	SCRIBING DATA						9
Types	s of Data	- Types of Variables - Basic Statistical descri	ptio	ns o	f Data	a – C	Desc	ribing
Data	with Tabl	es and Graphs –Describing Data with Average	es –	Des	scribir	ng Va	ariat	oility –
Norm	al Distrib	utions and Standard (z) Scores						
Unit I	II PR	OBABILITY & STATISTICS						9
Proba	ability Rev	view, Joint & Conditional. Review on statistics	- Po	pula	ations	and	Sar	nples
– Sar	npling Di	stribution of the mean – Hypothesis testing -	- Z	Tes	t – O	ne-T	aile	d and
Two-	Tailed Te	sts – Estimation – t-Test for one Sample – A	naly	/sis (of Va	rianc	e fo	or one
factor	-Chi-Sc	luare lest						
Unit		HON LIBRARIES FOR DATA WRANGLING						9
Basic	s of Num	py arrays – aggregations –computations on ar	rays	S – CC	mpai	rison	s, m	iasks,
boole	an logic ·	 fancy indexing – structured arrays – Data n 	nani	Ipula	ition \	Nith	Pan	das –
data i	ndexing a	and selection – operating on data – missing o		– H	ierarc	nica	i ina	lexing
	ibiriiriy ua	alasets – aggregation and grouping – pivot tac	ne2					
Unit								•
Unit								9
Impor		plotilib – Line plots – Scatter plots – visualizing e	erro	rs – (densi	ty an		ontour
piols -	– HISLOGI	ams – legends – colors – subplots – text and	ann	onal		Cusi	omi	
- une Seab	orn	Isional plotting - Geographic Data with Bas	sem	ap -	v 150	aliza		
Ceab	0111.				Total	: 45	Per	iods
Refer	ences			<u> </u>				
1.	David C	ielen, Arno D. B. Meysman, and Mohamed Ali,	"Intr	rodu	cing [Data	Scie	ence",
	Manning	Publications, 2016.						
2.	Robert	S. Witte and John S. Witte, "Statistics",	El	levei	nth E	ditio	on,	Wiley
2	Publicat	ions, 2017. ndor Diage "Dython Data Science Llandheat" (חיר		2040			
ა.	јаке va		J K	emy,	2016).		
4.	Avrim E	Blum, John Hopcroft, Ravindran Kannan, "Fou	Inda	ation	s of E	Data	Scie	ence",
	Cambrid	lge Press, 2020						

Course Outcomes:

Upon completion of the course, the students will be able to

- Understand data science fundamental and follow the correct process for applying data science.
- Represent and understand data in different formats and analyse it.
- Infer new information from the data using different analysis techniques.
- Gather, collect, and transform raw data into useful formats with Python libraries.
- Apply Python libraries to visualize and study data.

CSM50	08	вл А		NI		1.1			D) N			2		=		Г		\т					т	Ţ	D	Ţ	C		•••••		•••••
COMIS	00							A		11			J		-0	Jr	\ L	יונ	11	H			-	•	-	<u> </u>	-	<u> </u>				
		30)	U		U		3	<u> </u>			
	Objec	tives	:	- I-							1		_	,			L															
• 10	o unde	erstar		a c	bas Lui	SIC	0 C		no 	CE	epi	tS	0 		m	ac	nır	າຍ	le	al a	rnı	ng.										
• 10) unde	erstar	id an		ושמ הוים	illa	J S	su	ib.	er "L	rvi:	Se	эа		e2	arn 2	inę	gr mh	nc		eis	ó. A af	~		.: -							
• 10 • To		erstar	id ne	an a	ar bui	ne bliu	eเง สาม	.wc			ai no				an A		:01 201	ni. Vin	лп а	m			C	ass	511	iers	5					
• To	dosi	an an	d an	ıuı əlv	Jucio	iiiu ie r	of	un f n	151 210	րոր Ծե	he Vəf	;i v hil	i5 lic	eti/	u cí	ar	ann amh	ni i bic	iy Səl	l n	00	dala										
• 10	ucsi	grian	u an	ary	y 31.	3 (01	1 P	510		Jai		13	our	U į	gra	api	inc			10	ucic	,									
Unit I	INT	ROD	UCTI	0	N 1	ТС	0	Μ	IA	١C	H	IN	ĪE	EL	E	EA	R١	111	NG	3											9	
Machine	Lear	ning	– Ba	asi	ic	CC	on	nc	ce	pt	ts	_		Τy	yp	es	5 C	of	Μ	lad	chi	ne	le	arr	nir	ng ·	_	Ex	am	ple	es	&
Applicatio	ons -	Data	Pre-p	orc	oce	es	ssii	inç	g	-	No	ois	se) F	Re	m	ov	al	_	N	orr	nali	za	tio	n -	–Bi	ias	s &	Va	ria	anc	e,
Review o	n Pro	babil	ity – (Co	onc	diti	tio	<u>)na</u>	al	l p	oro	bb	at	oil	lity	/ -	·B	ay	/es	sia	an	cor	di	tior	าอ	l pr	ob	bab	ility	<u> </u>		
	. <u>.</u>																															
Unit II	SUF	PERV	ISED) L	.E/	AF	RN	NII	N	G	ì																				9	
Linear R	egres	sion	Mod	lels	s:	Μ	Лu	ılti	ipl	le	e r	.eć	gr	es	SS	ior	ר ר	_	Lo	рg	ist	ic r	eą	gres	SS	ion	,	Naï	ve	В	aye	es
classifier,	, Nea	rest	Neig	hb	DOU	ur	а	ane	hd	ł	K٨	١N	I	A	lg	or	ith	m,	, I	De	eci	sior	ו י	Tre	e	s, S	Su	ippo	ort	V	ecto	or
Machines	s, Ker	nel fu	nctio	ns	3																											
	·																															
Unit III	NE	URA		TV	NC	CR	RK	٢S	5,	E	EN	S	Ξľ	MI	BI	LE	Т	EC	CH	IN	IIQ	UE	S								9	
Artificial	Neura	al Ne	twork	k(/	٩Ņ	IN)	I),	p	be	erc	cep	pt	ro	n	,	m	ulti	ila	ye	er	pe	erce	pt	ron	,	Ba	ck	pr	opo	bg	atic	n
network(E	BPN)	acti	vatio	n n	tl an	un			on	ns ch	5, h:m	g	jra	ac	3IE	ent	- t	de	esc	ce	nt) ol	اtC ا	miz	za	tior), 	er	ror	~ ~	bac	K.
tuning ba	ion, l atch n	Jnil :	lizati	ion	n n	ι (rοc	(Và (Nà	ula ula	nı: ori	SI iz:	nın əti	ig ior	n n	yn F	ac In	שוג ספ	m	ր հե	סופ אר		en hte	n) - Ode	- -		-C	, i ain	iyf a	Bo	pai ost	an ind	neu r	31
turning, be			inzati		, ,	υų	gu			120	au		',			30				vic	201	003		DC	<u>'9</u>	gin	<u>y,</u>	00	031		<u>.</u>	
Unit IV	UNS	SUPE	RVIS	SO)R`	Y	&	k F	RE	EII	NF	FC)F	20	CE	EM	EN	νT	۰ L	.E	AF	RNI	N	3							9	
Clustering	a –	Dista	ance	F	Fur	nc	ctic	or	n.		Mi	ini	im	าน	Im	۱.	m	ax	kin	nu	Im	&		ave	era	ae		С	onr	ied	ctio	n.
Hierarchi	cal C	luste	ring,	а	igg	jlo	om	ne	era	ati	ive	э.	_	ł	<	M	ea	เทร	5 (clu	ust	terir	ng	, S	e	f-o	rga	aniz	zing	g	Ma	p,
Reinforce	ement	t Lear	ning	O١	ver	rvi	iev	W																								
Unit V	GR/	APHI	CAL	Μ	0	DE	EL	LS	5 8	&	D	IM	1E	IN	IS	IC	N	R	EC	DL	JC	TIO	Ν								9	
Directed	Grap	hical	Mo	de	els,	, F	Ba	ay	ye	esi	iar	n	Ν	le	etw	VO	rks	5,	Μ	lar	rkc	ov I	Мo	ode	ls	, ⊦	lic	lde	n l	Ma	arko	V
Models,	Inf	erend	ce-	Le	arı	nii	ing	g	(G	en	ne	ra	liz	za	tic	n,			Di	im	ens	io	n	re	duo	cti	on-	Cu	rse	e (of
Dimensin	nality,	PCA																							-	- 1		45	-			
																									1	ota	11:	45	26	eri	ods	5
COURS		ITCO																														_l
COURS			viE9;																													
At the	end o	of thi	s coi	urs	se,	, t	the	e	s	tu	ıde	en	nts	S	W	ill	be	e a	ıbl	le	to	:										
		• 4							. 1		- (_	. 1			1.0															

CO1: Explain the basic concepts of machine learning.

CO2: Construct supervised learning models.

CO3: Construct unsupervised learning algorithms.

CO4: Evaluate and compare different models

CO5: Design of experiments using machine learning

Reference	ces:
1.	Ethem Alpaydin, "Introduction to Machine Learning", MIT Press, Fourth Edition, 2020.
2.	Stephen Marsland, "Machine Learning: An Algorithmic Perspective, "Second Edition", CRC Press, 2014
3.	Sridhar S & Vijayalakshmi M, "Machine Learning", Oxford University Press, 2021
4.	Tom Mitchell, "Machine Learning", McGraw Hill, 3rd Edition, 1997.
5.	Mehryar Mohri, Afshin Rostamizadeh, Ameet Talwalkar, "Foundations of Machine Learning", Second Edition, MIT Press, . 2018.
6.	Ian Goodfellow, Yoshua Bengio, Aaron Courville, "Deep Learning", MIT Press, 2016
7.	Sebastain Raschka, Vahid Mirjalili , "Python Machine Learning", Packt publishing, 3rd Edition, 2019.
8.	Francois Chollet, "Deep Learning with Python", Second Edition, Manning Publications, 2021.

			DAT	A VI	SUA	ALIZ	ATI	ION			L	T	Ρ	С	С	redits
											3	0	0	3	<u>+</u>	3
Course O • To t • To t • To t • To t • To t	bject unde know acqui visua gain	tives: erstand to the wo ire know alize the skill in d	he func rking p vledge Data u lesignir	dame rincip abou ising ng rea	entals bles d it the tools al tim	s of of va e iss s Ta ne ir	data ariou sues ablea	a vis us in in d au activ	ualiz form ata r e inf	atior atior epre orma	n. n visi sent ation	ualiz atioi visu	atior n. ıaliza	dep	th tc syste	ools. em.
Unit I	INTF	RODUC	TION													9
Context of Key Facto Data repredata visua Scatterplo	f data ors – esent alizat t Map	a visualiz Purpose tation, E ion tool os - Tree	zation - e, visua Data Pr ls. Maj es, Hie	– Def alizati resen pping rarch	finitio ion f ntatio g - T nies, a	on, f func on, s Time and	Meth stion Seve e Se I Re	hodo anc en s eries curs	logy I ton tage s - (ion -	, Vis e, vis s of Conn Netv	ualiz suali data ectio work	atio zatio a vis ons s an	n des on de ualiz and d Gra	sign o sign ation Corr aphs	objeo opt , wi relat	ctives. ions – dgets, ions -
Unit II	VISU	JALIZAT	ION TE	CHN	IQUE	ES F	OR	TIM	E-SE	RIES	, TR	EES	& GF	RAPH	IS	9
Mapping - Scatter ch Hierarchie graph-Mat	Time arts, es and trix re	e series Scatter d Recurs present	- Conn maps sion - N ation fo	ectioi - Tre letwo or gra	ns ar e ma orks a aphs-	nd c aps, and s- Inf	corre , Sp Gra fo gr	elatio ace phs- raph	ons – filling Disp ics	- Indi g and Iayin	cato d noi ig Ar	r-Ar n-sp bitra	ea ch ace f iry Gi	art-F illing aphs	Pivot me s-no	table- thods- de link
Unit III	TEX	T AND	DOCU	MEN	T VIS	SU/	ALIZ	ZATI	ON							9
Acquiring of Files for U Files in a Levels of Expression Binary Dat	data, se wi Fold Effc ns, G ta Fo	- Where ith Proce ler ,Asyr ort, Toc Grammal ormats, A	e to Fina essing, nchrona ols for rs and Advanc	d Dat Loac ous I Gat BNF ed Do	ta, To ding Imag therir Nota etect	ools Tex ge D ng atior tive	s for tt Da Dowr Clu n, C Wo	Acq ata, I nload es, comp ork.	uiring Deali ds, V Tex ress	g Dat ng w Veb t Ma ed D	a frc ith F Tech arkup ata,	om th iles nniq D La Vec	ne Int and I ues, angua tors	erne Folde Pars ages and	t, Lo ers, I ing , R Geol	cating Listing data - egular metry,
Unit IV	INTE	ERACTI	VE DA	TA V	/ISU/	ALI	ZAT	ΓΙΟΝ								9
Drawing w Layouts – Dashboard	vith d Geor ds	lata – S mapping	cales - g – Exp	– Axe ortine	es – g, Fr	Up rame	date ewo	es, T ork –	rans D3.j	ition s, Ta	and blea	Mo u	tion ·	– Inte	erac	tivity -
Unit V	SEC			ΓΑ VI	ISUA	ALIZ	ΖΑΤΙ	ION								9
Port scan visualizatio systems –	visu on - Ii · Crea	ualizatio ntrusion ating sec	n - Vi detect cured v	ulnera ion Ic visual	abilit og vis lizatio	ty a sual on s	asse lizat syste	essm ion - em	ent Atta	and cking	exp j anc	loita I def	tion endir	- Fi ng vis	rewa suali	all log zation

Total: 45 Periods

Course Outcomes:

Upon completion of the course, the students will be able to

- Apply mathematics and basic science knowledge for designing information visualizing System.
- Collect data ethically and solve engineering problem in visualizing the information.
- Implement algorithms and techniques for interactive information visualization.
- Conduct experiments by applying various modern visualization tool and solve the
- space layout problem.
- Analyze and design system to visualize multidisciplinary multivariate Data individually or in teams.

Develop a cost effective and a scalable information visualization system.

Refer	ences
1	Robert Spence, "Information Visualization An Introduction", Third Edition, Pearson
	Education, 2014.
2	Colin Ware, "Information Visualization Perception for Design", Third edition, Margon Kaufmann Publishers, 2012.
3	Robert Spence, "Information Visualization Design for Interaction", Second Edition, Pearson Education, 2006.
4	Benjamin B. Bederson and Ben shneiderman, "The Craft of Information Visualization", Morgan Kaufmann Publishers, 2003.
5	Thomas strothotte, "Computational Visualization: Graphics, Abstraction and Interactivity", Springer, 1998.
6	Matthew O. Ward, George Grinstein, Daniel Keim, "Interactive Data Visualization: Foundation, Techniques and Applications", Second Edition, A. K. Peters/CRC Press, 2015.
7	Joerg Osarek, "Virtual Reality Analytics", Gordon's Arcade, 2016.

												•			
llnit l	Δ	ΤΤΔΟ								3	0	0	3	 	3 9
Attacks about Scienc Archite	s: Ana attack e: Pr ecting	alysing ks – D ivacy privac	i comr ata se preser y in Da	non a ecurity ving t ata an	attack / mitiga technio nd ML	vector ations. ques i oroject	rs – D . Priva in ML ts	ata Se acy awa - Ope	ecurit are N en-so	y – Macł ource	Prot nine ibi	abilis learn aries	stic 1 ing for	reas and PP	onin Dat ML
Unit II	E	NCRY	PTED	COM	IPUTA	TION									9
Encryp compu interse Learnii	tation ction ng. PS	compu – Hor – Priv SI and	itation nomor ate joi Moose	– T phic e in and e	Types encrypt d com	of ei tion. R pute –	ncrypt eal-w - Secu	ed con orld end ure Agg	mput crypt grega	tatio ed c ation	n: S omp – E	Secur utatic Encry	e N on: F oted	/lulti- riva Ma	-par te s chir
Unit II	D		OVEF	RNAN				Y APPI	ROA	CHE	S				9
mecha Anony	nism mity –	– Gau - Build	ssian r ng Priv	noise vacy i	for diff nto Da	ferentia ata Pip	al priv elines	acy – S	Sens	iti∨ity	/ and	d Priv	acy	unit	s – I
Unit IV	/ F	EDER	ATED	LEAF	RNING	AND	DATA	A SCIEI	NCE						9
systerr Unit V GDPR		Den-s Den-s EGAL	OURCE I	edera federa - PRI V	Optimiz ated lib VACY GPD - F	zation praries PIPL-	- Fede – Feo	al polic		a scie	contr	acts	– Ac	dheri	9
system Unit V GDPR contrac govern	L – CC ct ag	EGAL PPA – PA – reeme 2.0 - 1	ITY OF HIPAA nts ar ndian [Federa Federa F PRIV - LG nd la Data F	Optimizated lib VACY GPD - F Iw – Protect	PIPL- Interpr	- Fede – Feo Interna reting amew	al polic Data vork - U	ies a prot	and o ectic	contron re	acts egula ysis Fotal	– Ac tion: : 45	dheri s - Peri	9 ing t Dat
System Unit V GDPR contrac govern Course Gair goal Mole critic Able Know with Reference	Ling - CC CC ag ance Cuto Anov s and wledg e to ur wledg offen ces	EGAL PA – Ireeme 2.0 - I comes wledge I frame ge on t fferent rastruc ndersta ge on I ses.	iTY OF HIPAA nts ar ndian I on the on the work ne lanc iate an cture and pol T Act, a	e natu dicies f and a	Optimizated lib VACY SPD - F Protect ure of a e of has egrate a to mitig mendr	PIPL- I Interpr tion Fr attacks acking stratec gate da ments,	- Fede - Fede - Fede International reting cameway a and tagies for ata secopy	al polic Data vork - U chreats lefense or data s ecurity b rights,	and e med pread	and o ectionase secu secu chan rity a ching and	isms cybe	acts egula ysis Fotal mana protec	- Actions : 45 agementing	dheri s - Peri nent	9 ing t Dat
System Unit V GDPR contrac govern Course Gair goal More Able critic Able Know with Reference	Ling - CC Ling - CC ct ag ance for know s and wledg e to dif cal infi e to ur wledg offen ces Kath	EGAL PA – reeme 2.0 - 1 comes wledge I frame ge on t fferent rastruc ndersta ge on l ses	iTY OF HIPAA nts ar ndian I on the on the work ne lanc iate an cture and pol T Act, a armul,	e natu dicies f and a Practi	Optimizated lib VACY OPD - F OPD - F OPD -	PIPL- I Interpr tion Fr attacks acking stratec gate da nents,	- Fede - Fede - Fede International reting camewa s and t and d gies for ata sec copy 	al polic Data vork - U chreats lefense or data s ecurity b rights, Reilly M	and secu preac and secu preac IPR	a scie and o ectio ase secu chan rity a ching and	isms cybe	acts egula ysis Fotal mana protec	- Actions : 45 gem cting	dheri s - Peri hent	9 ing t Dat
System Unit V GDPR contrac govern Course Gair goal Kno Able critic Able Kno with Referen 1. 2.	Davi Market	EGAL PA – reeme 2.0 - I comes wledge I frame ge on I fferent rastruc ndersta ge on I ses harine c i-Party s://secu	ITY OF HIPAA nts ar ndian I on the work ne land iate an cture and pol T Act, a armul, s, Vlad Cor irecomp	F PRIN F PRIN - LG nd la Data F e natu dscap d inte licies f and an Praction putation	Optimizated lib Ated lib VACY OPD - I Protect ure of a e of ha egrate to mitig mendr cal Dat olesnik tion, on.org/)	PIPL- Interpr tion Fr attacks acking strateg gate da nents, a Priva	- Fede - Fede - Fede Internation reting ramew and the and the and the gies for ata see copy acy, O' I Mike Pu	erated I derated al polic Data vork - U threats lefense or data s ecurity b rights, Reilly M Rosulek ublishers	and e mec secu preac IPR Media, (, A P s,	and of ectional sectional	isms and p 2023	acts egula ysis Fotal mana protec er law 3 Introc (Free	- Actions : 45 igem cting to c	dheri s - Peri nent leal	9 ing 1 Dat iods Sec
System Unit V GDPR contrac govern Course Gair goal Mole critic Able critic Able Know with Ceferen 1. 2.	Li CC Ct ag ance Outc t ag ance Outc t ag ance t ag t	EGAL PA – Ireeme 2.0 - I comes wledge I frame ge on t fferent rastruc ndersta ge on I ses harine c i-Party s://secu am Sta on, Pe	interiore interi	e natu dicies f and a pata F e natu discap di inte licies f and a Practiv imir Ka putatio Crypto 2017	Optimiz ated lib VACY OPD - I Protect ure of a e of ha egrate s to mitig mendr cal Dat tolesnik tion, on.org/) ography	PIPL- Interpr tion Fr attacks acking strateg gate da nents, a Priva ov and NOW	- Fede - Fede - Fede Internation reting amew and the and the	erated I derated al polic Data vork - U chreats lefense or data s ecurity b rights, Reilly M Rosulek ublishers < Securi	and ies a prot <u>lse ca</u> and secu preac IPR fedia, c, A P s, ity - F	and o ectionase secu chan rity a ching and Tinc, Pragn 2022	isms and p 2023 ples	acts egula ysis Fotal mana oroteo er law 3 Introc (Free and F	- Actions : 45 gem cting to c	Peri Peri hent leal	9 ing t Dat iods Sec

BIG DATA ANALYTICS	L	Т	Ρ	С	Credits
	3	0	2	3	3
Course Objectives:					
To understand big data.					
• To learn and use NoSQL big data management.					
 To learn mapreduce analytics using Hadoop and relat 	ed t	ools	•		
To work with map reduce applications					
 I o understand the usage of Hadoop related tools for I 	Big I	Jata	Ana	ytics	5.
Unit I UNDERSTANDING BIG DATA					5
Introduction to big data – convergence of key trends – ur	nstru	ictur	ed d	ata -	 industry
examples of big data – web analytics – big data application	s – I	big c	data t	echr	ologies –
introduction to Hadoop – open source technologies – clo	ud	and	big (data	 mobile
business intelligence – Crowd sourcing analytics – inter and	tran	s fir	ewall	ana	ytics.
Unit II NOSQL DATA MANAGEMENT					7
Introduction to NoSQL – aggregate data models – key-value	and	doc	umer	nt da	ta models
 relationships – graph databases – schemaless databas 	es -	– m	ateria	lizec	l views –
distribution models – master-slave replication – consistency	– C	ass	andra	1 – C	assandra
data model – Cassandra examples – Cassandra clients.					
					6
Man Rebool AIT Elocationo	ndl	0001	toete		ontomy of
MapReduce iob run – classic Map-reduce – YARN – failures	ind	lass	ic Ma	an-re	duce and
YARN – job scheduling – shuffle and sort – task execution -	- Ma	anRe		tvo	es – input
formats – output formats.	ivic	, price	Judoc		oo mpat
Unit IV BASICS OF HADOOP					6
Data format – analyzing data with Hadoop – scaling out – H	ado	op s	tream	ning ·	- Haddop
pipes design of Haddop distributed file system (HDFS) – HDF	S co	nce	pts –	Java	a interface
- data flow - Hadoop I/O - data integrity - compression -	ser	ializ	ation	– A	vro – file-
based data structures – Cassandra – Hadoop integration.					
Unit V HADOOP RELATED TOOLS					6
Hbase – data model and implementations – Hbase clients –	· Hb	ase	exan	ples	s – praxis.
Pig – Grunt – pig data model – Pig Latin – developing and te	stin	g Pi	g Lati	n sc	ripts.
Hive – data types and file formats – HiveQL data definition –	- Hiv	/eQL	_ data	a ma	nipulation
– HiveQL queries.					
			lota	: 30	Periods

Course Outcomes: After the completion of this course, students will be able to:

- Describe the big data and use cases from selected business domains.
- Explain NoSQL big data management.
- Install, configure and run Hadoop and HDFS.
- Perform map-reduce analytics using Hadoop.
- Use Hadoop-related tools such as HBase, Cassandra, Pig and Hive for big data analytics.

Refer	ences
1	Michael Minelli, Michael Chambers, and AmbigDhiraj, "Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses", Wiley, 2013.
2	Eric Sammer, "Hadoop Operations", O'Reilley, 2012.
3	Sadalage, Pramod J. "NoSQL distilled", 2013
4	E. Capriolo, D. Wrampler, and J. Rutherglen, "Programming Hive", O'Reilley, 2012.
5	Lars George, "HBase: The Definitive Guide: O'Reilley, 2011.
6	Eben Hewitt, "Cassandra: The Definitive Guide: O'Reilley, 2010.
7	Alan Gates, "Programming Pig", O'Reilley, 2011.

		EXPLORATORY [DATA ANALY	(SIS	L	Т	Ρ	С	Cred	its
					3	0	0	3	3	
Course (Object	ves:				4	-	.	L	
• To	o outlir	e an overview of explo	ratory data an	nalysis.						
• To	o imple	ment data visualization	using Matplo	otlib.						
• To	o perfo	m univariate data expl	oration and a	nalysis.						
• To	o apply	bivariate data explorat	ion and analy	/sis.						
• To) use [ata exploration and vis	sualization tec	chniques	for r	nulti	varia	te ar	nd time	
Se Linit I		TA.	Vele							
Unit I	EXP	URATURT DATA ANAI	1313						9	
and pivot	ting, T	ansformation technique	es.							
Unit II Data Mar Operating datasets	EDA nipula g on – Con	USING PYTHON on using Pandas – P Data – Handling Miss at, Append, Merge an	andas Object ing Data – I d Join – Aggre	ts – Data Hierarch egation a	a Ind ical and g	dexir Inde grouj	ng ar xing ping ·	nd So - C - Piv	9 electior combini vot Tab	ר – ing les
Unit II Data Mar Operating datasets – Vectori	EDA nipula g on – Con zed St	USING PYTHON on using Pandas – P Data – Handling Miss at, Append, Merge an ing Operations.	andas Object ing Data – I d Join – Aggre	ts – Data Hierarch egation a	a Ind ical and (dexir Inde grou	ng ar xing bing	nd So – C – Piv	9 electior Combini vot Tab	ר – ing les
Unit II Data Mar Operating datasets – Vectori Unit III	EDA nipula g on – Con zed St	USING PYTHON on using Pandas – P Data – Handling Miss cat, Append, Merge and ing Operations.	andas Object ing Data – I d Join – Aggro	ts – Data Hierarch egation a	a Ind ical and (dexir Inde grou	ng ar xing bing	nd So – C – Piv	9 electior combini vot Tab	ר – ing les
Unit II Data Mar Operating datasets – Vectori Unit III Introducti and Spre	EDA nipula g on – Con zed St UNI ion to ead – S	USING PYTHON on using Pandas – P Data – Handling Miss cat, Append, Merge and ing Operations. ARIATE ANALYSIS Single Variable: Distrib caling and Standardizi	andas Object ing Data – I d Join – Aggro ution Variable ng – Inequalit	ts – Data Hierarch egation a es – Nur	a Ind ical and (dexir Inde grou	ng ar xing oing · umm	nd So – C – Piv	9 electior combini vot Tab 9 s of Le	n – ing les vel
Unit II Data Mar Operating datasets – Vectori Unit III Introducti and Spre	EDA nipula g on – Con zed St ion to ead – St BIV	USING PYTHON on using Pandas – P Data – Handling Miss cat, Append, Merge and ing Operations. ARIATE ANALYSIS Dingle Variable: Distrib caling and Standardizi	andas Object ing Data – I d Join – Aggro ution Variable ng – Inequalit	ts – Data Hierarch egation a es – Nur ty.	a Ind ical and (meric	dexir Inde grou	ng ar xing bing umm	nd So – C – Piv	9 electior combini vot Tab 9 s of Le 9	n – ing les vel
Unit II Data Mar Operating datasets – Vectoria Unit III Introducti and Spre Unit IV Relations	EDA nipula g on – Con zed St ion to ead – St BIVA	USING PYTHON on using Pandas – P Data – Handling Miss at, Append, Merge and ing Operations. ARIATE ANALYSIS Single Variable: Distrib caling and Standardizi RIATE ANALYSIS etween Two Variables	andas Object ing Data – I d Join – Aggro ution Variable ng – Inequalit s – Percenta	ts – Data Hierarch egation a es – Nur ty.	a Ind ical and g meric	dexir Inde grou cal S	ng ar xing oing - umm alysis	nd So – C – Piv naries	9 electior combini vot Tab 9 s of Le 9 ntinger	n – ing les vel
Unit II Data Mar Operating datasets – Vectori Unit III Introducti and Spre Unit IV Relations Tables –	EDA nipula g on – Con zed St ion to ead – St BIVA ships I Hand	USING PYTHON on using Pandas – P Data – Handling Miss at, Append, Merge and ing Operations. ARIATE ANALYSIS Single Variable: Distrib caling and Standardizi RIATE ANALYSIS etween Two Variables ng Several Batches – S	andas Object ing Data – I d Join – Aggro ution Variable ng – Inequalit s – Percenta Scatterplots a	ts – Data Hierarch egation a es – Nur ty. age Table and Resis	a Ind ical and g meric es - stant	dexir Inde group cal S	ig ar xing oing umm alysis	nd So – C – Piv	9 electior combini vot Tab 9 s of Le 9 ntinger	n – ing les vel
Unit II Data Mar Operating datasets – Vectori Unit III Introducti and Spre Unit IV Relations Tables – Unit V	EDA nipula g on – Con zed St ion to ad – St BIV/ ships I Hand	USING PYTHON on using Pandas – P Data – Handling Miss cat, Append, Merge and ing Operations. ARIATE ANALYSIS Dingle Variable: Distrib caling and Standardizi RIATE ANALYSIS etween Two Variables ng Several Batches –	andas Object ing Data – I d Join – Aggro ution Variable ng – Inequalit s – Percenta Scatterplots a	ts – Data Hierarch egation a es – Nur es – Nur ge Table ind Resis	a Ind ical and g meric es - stant	dexir Inde group cal S	ng ar xing oing umm alysis	nd So – C – Piv	9 electior combini vot Tab 9 s of Le 9 ntinger	n – ing les vel
Unit II Data Mar Operating datasets – Vectoria Unit III Introducti and Spre Unit IV Relations Tables – Unit V Introducir	EDA nipula g on – Con zed St ion to ead – S BIVA ships I Handl MUL ng a	USING PYTHON on using Pandas – P Data – Handling Miss at, Append, Merge and ing Operations. ARIATE ANALYSIS Single Variable: Distrib caling and Standardizi RIATE ANALYSIS etween Two Variables ng Several Batches – S TIVARIATE AND TIME hird Variable – Caus	andas Object ing Data – I d Join – Aggro oution Variable ng – Inequalit s – Percenta Scatterplots a E SERIES AN sal Explanatio	ts – Data Hierarch egation a es – Nur ty. age Table and Resis IALYSIS ons – T	a Inc ical and g meric es – stant	dexir Inde grou cal S cal S	ig ar xing bing umm alysis as.	nd So – C – Piv naries	9 electior combini vot Tab 9 s of Le 9 ntinger 9 ntinger	n – ing les vel
Unit II Data Mar Operating datasets – Vectori Unit III Introducti and Spre Unit IV Relations Tables – Unit V Introducin Tables an	EDA nipula g on – Con zed St ion to ad – S BIV/ ships I Handl MUL ng a nd Be	USING PYTHON on using Pandas – P Data – Handling Miss at, Append, Merge and ing Operations. ARIATE ANALYSIS Single Variable: Distrib caling and Standardizi RIATE ANALYSIS etween Two Variables ng Several Batches – S FIVARIATE AND TIME Third Variable – Caus ond – Fundamentals of	andas Object ing Data – I d Join – Aggro oution Variable ng – Inequalit s – Percenta Scatterplots a E SERIES AN sal Explanatio of TSA – Chra	ts – Data Hierarch egation a es – Nur es – Nur y. ige Table and Resis IALYSIS ons – T acteristics	a Ind ical and g meric es - stant hree s of	dexir Inde group cal S cal S	iable serie	aries	9 electior combini vot Tab 9 s of Le 9 ntinger ta – Da	n – ing les vel ncy ata
Unit II Data Mar Operating datasets – Vectoria Unit III Introducti and Spre Unit IV Relations Tables – Unit V Introducin Tables an Cleaning	EDA nipula g on – Con zed St ion to ead – S BIV/ ships I Handl Handl MUL ng a nd Bey – Tim	USING PYTHON on using Pandas – P Data – Handling Miss at, Append, Merge and ing Operations. ARIATE ANALYSIS Single Variable: Distrib caling and Standardizi RIATE ANALYSIS etween Two Variables of Several Batches – S FIVARIATE AND TIME hird Variable – Caus ond – Fundamentals of b-based indexing – Vis	andas Object ing Data – I d Join – Aggro ution Variable ng – Inequalit s – Percenta Scatterplots a E SERIES AN sal Explanatio of TSA – Chra ualizing – Gro	ts – Data Hierarch egation a es – Nur es – Nur y. age Table and Resis IALYSIS ons – T acteristics ouping –	a Inc ical and g neric es – stant hree s of	dexir Inde grou cal S cal S cal S	ig ar xing bing umm alysis as.	aries	9 electior combini vot Tab 9 s of Le 9 ntinger ta – Da	n – ing iles vel ncy ata

COUR	RSE OUTCOMES:							
At the end of this course, the students will be able to:								
CO1:	Understand the fundamentals of exploratory data analysis.							
CO2:	Implement the data Visualization using Matplotlib.							
CO3:	Perform univariate data exploration and analysis.							
CO4:	Apply bivariate data exploration and analysis.							
CO5:	Use Data exploration and visualization techniques for multivariate and time series data.							

Total: 60 Periods

REFERE	NCES:
1	Suresh Kumar Mukhiya, Usman Ahmed, "Hands-On Exploratory Data Analysis with Python", Packt Publishing, 2020. (Unit 1)
2	Jake Vander Plas, "Python Data Science Handbook: Essential Tools for Working with Data". First Edition, O Reilly, 2017. (Unit 2)
3	Catherine Mars, Jane Elliott, "Exploring Data: An Introduction to Data Analysis for Social Scientists", Wiley Publications, 2 nd Edition, 2008. (Unit 3,4,5)
4	Eric Pimpler, Data Visualization and Exploration with R, GeoSpatial Training service, 2017.
5	Claus O. Wilke, "Fundamentals of Data Visualization", O'reilly Publications, 2019.
6	Matthew O. Ward, Georges Grinstein, Daniel Keim, "Interactive Data Visualization: Foundations, Techniques, and Applications", 2 nd Edition, CRC press, 2015.